B2.536

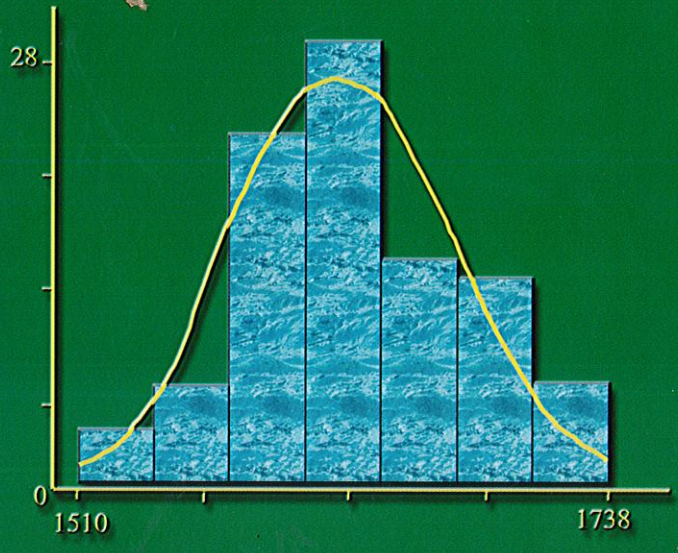
TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN
KHOA TOÁN KINH TẾ

PGS.TS. Nguyễn Cao Văn (Chủ biên)
PGS.TS. Ngô Văn Thứ
TS. Trần Thái Ninh

Giáo trình

LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ TOÁN

(Tái bản lần thứ 6)



TEM CHỐNG HÀNG GIẢ

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

NHÀ XUẤT BẢN ĐẠI HỌC KINH TẾ QUỐC DÂN
2018

TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN
KHOA TOÁN KINH TẾ

PGS.TS. NGUYỄN CAO VĂN (Chủ biên)
PGS.TS. NGÔ VĂN THỨ
TS. TRẦN THÁI NINH

Giáo trình

LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ TOÁN

(Tái bản lần thứ 6)

TRƯỜNG Đ.H KINH DOANH VÀ CÔNG NGHỆ HÀ NỘI
THƯ VIỆN

NHÀ XUẤT BẢN ĐẠI HỌC KINH TẾ QUỐC DÂN
2018

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

THƯ VIỆN
HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

MỤC LỤC

LỜI NÓI ĐẦU 1

Phần I. LÝ THUYẾT XÁC SUẤT 3

Chương 1. BIẾN CỐ NGẪU NHIÊN VÀ XÁC SUẤT 5

1. Mở đầu 5
2. Phép thử và các loại biến cố 5
3. Xác suất của biến cố 7
4. Định nghĩa cổ điển về xác suất 7
   4.1. Thí dụ 7
   4.2. Định nghĩa 8
   4.3. Các tính chất của xác suất 8
   4.4. Các phương pháp tính xác suất bằng định nghĩa cổ điển 9
   4.5. Ưu điểm và hạn chế của định nghĩa cổ điển về xác suất 17
5. Định nghĩa thống kê về xác suất 18
   5.1. Định nghĩa 1 18
   5.2. Định nghĩa 2 19
   5.3. Ưu điểm và hạn chế của định nghĩa thống kê về xác suất 19
6. Một số định nghĩa khác về xác suất 20
   6.1. Định nghĩa hình học về xác suất 20
   6.2. Xác suất chủ quan 21
   6.3. Định nghĩa tiên đề về xác suất 21
7. Nguyên lý xác suất lớn và xác suất nhỏ 22
8. Định lý cộng xác suất 25
   8.1. Định lý 26
   8.2. Hệ quả 27
9. Định lý nhân xác suất 31
   9.1. Định lý 1 35
   9.2. Định lý 2 36
   9.3. Định lý 3 41

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.4. Định nghĩa 3 ..... *180*
8.5. Công thức tính xác suất để biến ngẫu nhiên phân phối chuẩn nhận giá trị trong khoảng (a, b) *181*
8.6. Xác suất của sự sai lệch giữa biến ngẫu nhiên và kỳ vọng toán của nó *183*
8.7. Quy tắc hai xích ma và ba xích ma *184*
8.8. Phân phối xác suất của tổng các biến ngẫu nhiên độc lập tuân theo cùng một quy luật *186*
8.9. Sự hội tụ của quy luật nhị thức và quy luật Poisson về quy luật chuẩn *186*
8.10. Ứng dụng của quy luật chuẩn *188*
9. Quy luật khi bình phương $\chi^2(n)$ 191
10. Quy luật Student - T(n) 193
11. Quy luật Fisher-Snedecor - $F(n_1, n_2)$ 195
12. Tóm tắt chương 3 197
Các ký hiệu và công thức cơ bản *197*
Các thuật ngữ cơ bản *200*
Câu hỏi ôn tập *201*
Bài tập tổng hợp chương 3 ..... *204*
Tài liệu tham khảo *208*
Chương 4. BIẾN NGẪU NHIÊN HAI CHIỀU. HÀM CÁC BIẾN NGẪU NHIÊN **209**
1. Mở đầu 209
2. Khái niệm về biến ngẫu nhiều chiều 210
3. Bảng phân phối xác suất của biến ngẫu nhiên hai chiều 210
4. Hàm phân bố xác suất của biến ngẫu nhiên hai chiều 214
5. Hàm mật độ xác suất của biến ngẫu nhiên hai chiều 219
6. Quy luật phân phối xác suất có điều kiện của các thành phần của hệ hai biến ngẫu nhiên 225
7. Các tham số đặc trưng của hệ hai biến ngẫu nhiên 230
8. Kỳ vọng toán có điều kiện - Hàm hồi quy 238
9. Phân phối chuẩn hai chiều 241
10. Quy luật phân phối xác suất của hàm các biến ngẫu nhiên 242

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

10.1. Quy luật phân phối xác suất của hàm một biến ngẫu nhiên 242
10.2. Quy luật phân phối xác suất của hàm hai biến ngẫu nhiên 245
10.3. Các tham số đặc trưng của hàm các biến ngẫu nhiên 247
11. Tóm tắt chương 4 252
Các ký hiệu và công thức cơ bản 253
Các thuật ngữ cơ bản 257
Câu hỏi ôn tập 258
Bài tập tổng hợp chương 4 261
Tài liệu tham khảo 270
Chương 5. CÁC ĐỊNH LÝ GIỚI HẠN 271
1. Mở đầu 271
2. Bất đẳng thức Trêbusép 271
3. Định lý Trêbusép 275
4. Định lý Bernoulli 278
5. Định lý giới hạn trung tâm 279
5.1. Hàm đặc trưng 280
5.2. Định lý Lindenberg - Lewi 282
6. Tóm tắt chương V 285
Các ký hiệu và công thức cơ bản 286
Các thuật ngữ cơ bản 287
Câu hỏi ôn tập 288
Bài tập tổng hợp chương 5 289
Tài liệu tham khảo 292
Phần II. THỐNG KÊ TOÁN 293
Chương 6. CƠ SỞ LÝ THUYẾT MẪU 295
1. Mở đầu 295
2. Khái niệm về phương pháp mẫu 295
3. Tổng thể nghiên cứu 296
3.1. Định nghĩa 296
3.2. Các phương pháp mô tả tổng thể 297
3.3. Các tham số đặc trưng của tổng thể 298

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4. Mẫu ngẫu nhiên 306
4.1. Định nghĩa mẫu ngẫu nhiên 306
4.2. Các phương pháp chọn mẫu 308
4.3. Thang đo các giá trị mẫu 311
4.4. Các phương pháp mô tả số liệu mẫu 312
5. Thống kê 324
5.1. Định nghĩa 324
5.2. Một số thống kê đặc trưng của mẫu ngẫu nhiên 324
5.3. Đồ thị hình hộp 339
6. Mẫu ngẫu nhiên hai chiều 348
6.1. Khái niệm 348
6.2. Phương pháp mô tả mẫu ngẫu nhiên hai chiều 349
6.3. Một số thống kê đặc trưng của mẫu ngẫu nhiên hai chiều 350
7. Quy luật phân phối xác suất của một số thống kê đặc trưng mẫu 354
7.1. Trường hợp biến ngẫu nhiên gốc phân phối theo quy luật chuẩn 354
7.2. Trường hợp có hai biến ngẫu nhiên gốc cùng phân phối theo quy luật chuẩn 356
7.3. Trường hợp biến ngẫu nhiên gốc phân phối theo quy luật không - một 359
7.4. Trường hợp có hai biến ngẫu nhiên gốc cùng phân phối theo quy luật không - một 360
8. Suy diễn thống kê 361
8.1. Suy diễn về mẫu ngẫu nhiên rút ra từ tổng thể phân phối chuẩn 361
8.2. Suy diễn về hai mẫu ngẫu nhiên rút ra từ hai tổng thể phân phối chuẩn 364
8.3. Suy diễn về mẫu ngẫu nhiên rút ra từ tổng thể phân phối không - một 367
8.4. Suy diễn về hai mẫu ngẫu nhiên rút ra từ hai tổng thể phân phối không - một 370
9. Tóm tắt chương 6 374
Các ký hiệu và công thức cơ bản 375

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản 380
Câu hỏi ôn tập 382
Bài tập tổng hợp chương 6 384
Tài liệu tham khảo 388
Chương 7. ƯỚC LƯỢNG CÁC THAM SỐ CỦA BIẾN NGẪU NHIÊN **389**
1. Mở đầu 389
2. Phương pháp ước lượng điểm 390
2.1. Phương pháp hàm ước lượng 390
2.2. Phương pháp ước lượng hợp lý tối đa 399
3. Phương pháp ước lượng bằng khoảng tin cậy 403
3.1. Khái niệm 403
3.2. Ước lượng kỳ vọng toán của biến ngẫu nhiên phân phối theo quy luật chuẩn 405
3.3. Ước lượng hiệu hai kỳ vọng toán của hai biến ngẫu nhiên phân phối chuẩn 421
3.4. Ước lượng xác suất p của biến ngẫu nhiên phân phối theo quy luật không – một 431
3.5. Ước lượng hiệu hai tham số p của hai biến ngẫu nhiên phân phối không - một 438
3.6. Ước lượng phương sai của biến ngẫu nhiên phân phối theo quy luật chuẩn 440
3.7. Ước lượng tỷ số của hai phương sai của hai biến ngẫu nhiên phân phối chuẩn 446
3.8. Ước lượng trung vị của tổng thể nghiên cứu 449
4. Tóm tắt chương 7 453
Các ký hiệu và công thức cơ bản 454
Các thuật ngữ cơ bản 457
Câu hỏi ôn tập 458
Bài tập tổng hợp chương 7 459
Tài liệu tham khảo 464
Chương 8. KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ **465**
1. Mở đầu 465

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2. Khái niệm chung 465
2.1. Giả thuyết thống kê 465
2.2. Tiêu chuẩn kiểm định giả thuyết thống kê 467
2.3. Miền bác bỏ giả thuyết 467
2.4. Giá trị quan sát của tiêu chuẩn kiểm định 468
2.5. Quy tắc kiểm định giả thuyết thống kê 468
2.6. Sai lầm loại một và sai lầm loại hai 468
2.7. Thủ tục kiểm định giả thuyết thống kê 470
3. Kiểm định tham số 471
3.1. Kiểm định giả thuyết về kỳ vọng toán của biến ngẫu nhiên khi đã biết phương sai 471
3.2. Kiểm định giả thuyết về kỳ vọng toán của biến ngẫu nhiên phân phối chuẩn khi chưa biết phương sai 482
3.3. Kiểm định giả thuyết về hai kỳ vọng toán của hai biến ngẫu nhiên phân phối chuẩn 491
3.4. Kiểm định giả thuyết về tham số p của biến ngẫu nhiên phân phối không – một 512
3.5. Kiểm định giả thuyết về hai tham số p của hai biến ngẫu nhiên phân phối không – một 519
3.6. Kiểm định giả thuyết về phương sai của biến ngẫu nhiên phân phối chuẩn 524
3.7. Kiểm định giả thuyết về sự bằng nhau của hai phương sai của hai biến ngẫu nhiên phân phối chuẩn 527
3.8. Kiểm định k phương sai của k biến ngẫu nhiên phân phối chuẩn 531
4. Tóm tắt chương 8 536
Các ký hiệu và công thức cơ bản 537
Các thuật ngữ cơ bản 541
Câu hỏi ôn tập 542
Bài tập tổng hợp chương 8 543
Tài liệu tham khảo 549
Chương 9. KIỂM ĐỊNH PHI THAM SỐ 551
1. Mở đầu 551
2. Kiểm định khi bình phương 552

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.1. Kiểm định giả thuyết về tính độc lập của hai dấu hiệu định tính 552
2.2. Kiểm định giả thuyết về k tham số p của k biến ngẫu nhiên phân phối không một 557
2.3. Kiểm định giả thuyết về quy luật phân phối xác suất của biến ngẫu nhiên 562
3. Một số kiểm định khác về quy luật phân phối xác suất 569
3.1. Tiêu chuẩn phù hợp Kolmogorov 569
3.2. Kiểm định Lilliefors về dạng phân phối chuẩn 573
3.3. Kiểm định Jarque - Bera về dạng phân phối chuẩn 575
4. Kiểm định theo dấu 580
4.1. Kiểm định về một trung vị của tổng thể 580
4.2. Kiểm định về hai trung vị của hai tổng thể khi hai mẫu gồm các giá trị theo cặp 582
5. Kiểm định tổng hạng Wilcoxon 584
6. Kiểm định tổng hạng theo dấu Wilcoxon. 591
7. Kiểm định Kruskal - Wallis 596
8. Kiểm định đoạn mạch 600
8.1. Kiểm định một mẫu 600
8.2. Kiểm định hai mẫu (Kiểm định Wald – Wolfowitz) 603
9. Tóm tắt chương 9. 606
Các ký hiệu và công thức cơ bản 607
Các thuật ngữ cơ bản 609
Câu hỏi ôn tập 610
Bài tập tổng hợp chương 9 611
Tài liệu tham khảo 614
Chương 10. PHÂN TÍCH PHƯƠNG SAI **615**
1. Mở đầu 615
2. Mô hình phân tích phương sai một nhân tố 617
3. Mô hình phân tích phương sai hai nhân tố 626
3.1. Mô hình hai nhân tố tác động riêng rẽ 626
3.2. Mô hình phân tích phương sai hai nhân tố tác động tổng hợp 631
4. Tóm tắt chương 10 640

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các ký hiệu và công thức cơ bản *641*
Các thuật ngữ cơ bản *643*
Câu hỏi ôn tập *643*
Bài tập tổng hợp chương 10 ..... *644*
Tài liệu tham khảo *650*
Chương 11. PHÂN TÍCH TƯƠNG QUAN VÀ HỒI QUY **651**
1. Mở đầu 651
2. Phân tích tương quan 652
2.1. Phân tích tương quan bảng số liệu định lượng *652*
2.2. Phân tích tương quan bảng số liệu định tính *660*
3. Phân tích hồi quy 669
3.1. Hàm hồi quy *669*
3.2. Mô hình hồi quy tuyến tính đơn *670*
3.3. Mô hình hồi quy tuyến tính bội *681*
3.4. Một số dạng hàm hồi quy phi tuyến có thể đưa về dạng tuyến tính *690*
4. Tóm tắt chương 11 696
Các ký hiệu và công thức cơ bản *697*
Các thuật ngữ cơ bản *702*
Câu hỏi ôn tập *703*
Bài tập tổng hợp chương 11 ..... *704*
Tài liệu tham khảo *714*
Chương 12. PHÂN TÍCH CHUỖI THỜI GIAN **715**
1. Mở đầu 715
2. Phân tích bằng hàm xu thế 719
2.1. Phân tích hồi quy chuỗi thời gian *719*
2.2. Phương pháp bình quân di động (MA) *720*
2.3. Phương pháp san mũ *722*
3. Phân tích biến động mùa vụ 727
4. Phương pháp san mũ Holt – Winters 734
4.1. Phương pháp san mũ Holt - Winters không mùa vụ *734*

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.2. Phương pháp san mũ Holt - Winters có mùa vụ 735
5. Phân tích biến động chu kỳ 739
6. Tóm tắt chương 12 743
Các ký hiệu và công thức cơ bản 744
Các thuật ngữ cơ bản 746
Câu hỏi ôn tập 746
Bài tập tổng hợp chương 12 747
Tài liệu tham khảo 752
Phần III. PHÂN TÍCH NHÂN TỐ 753
Chương 13. PHƯƠNG PHÁP THÀNH PHẦN CHÍNH 755
1. Mở đầu 755
2. Bài toán phân tích nhân tố 756
3. Các ý tưởng cơ bản 757
3.1. Một thí dụ không chuẩn mực 757
3.2. Bài toán phân tích nhân tố trong không gian 2 và 3 chiều 760
3.3. Bài toán trong không gian p chiều 761
4. Mô tả, tóm tắt số liệu 763
4.1. Số liệu và các đặc trưng 763
4.2. Không gian các cá thể 768
4.3. Không gian các biến 770
5. Tạo biến và phép chiếu trong không gian tuyến tính 771
5.1. Tạo biến mới 771
5.2. Phép chiếu 772
6. Phân tích thành phần chính 774
6.1. Phép chiếu lên không gian con 774
6.2. Trục chính, thành phần chính và nhân tố chính 776
6.3. Xác định λ và chọn số thành phần chính 779
6.4. Tái hiện dữ liệu 780
6.5. Phân tích thành phần chính với ma trận hệ số tương quan 781
7. Phân tích và đánh giá kết quả của việc áp dụng phương pháp thành phần chính trong phân tích nhân tố 783

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7.1. Các kiểm định chung *783*
7.2. Tương quan của các thành phần và các biến ban đầu *784*
7.3. Các hệ số phản ánh liên hệ của các cá thể và các thành phần chính *786*
7.4. Thí dụ bằng số *787*
8. Tỷ lệ giải thích trên các siêu phẳng chiếu 794
8.1. Mức giải thích chung tổng cộng *794*
8.2. Tỷ lệ bộ phận *794*
9. Tiêu chuẩn chọn số thành phần chính cho một phân tích 795
9.1. Tiêu chuẩn lý thuyết *795*
9.2. Tiêu chuẩn Kaiser *795*
10. Mô hình hồi qui thành phần chính 796
10.1. Đa cộng tuyến và hồi qui thành phần chính *796*
10.2. Tác động của các biến độc lập đến biến phụ thuộc *797*
10.3. Một số thí dụ thực nghiệm *798*
11. SPSS với phân tích thành phần chính 804
12. Vai trò của các biến và cách thức lựa chọn biến 822
12.1. Xác định mục đích phân tích và lựa chọn biến ban đầu *823*
12.2. Phân tích tương quan và lọc biến ban đầu *823*
12.3. Lựa chọn mô hình phân tích *824*
12.4. Phân tích nhân tố bằng phương pháp thành phần chính *824*
13. Tóm tắt chương 13 837
Các ký hiệu và công thức cơ bản *838*
Các thuật ngữ cơ bản *840*
Câu hỏi ôn tập *841*
Bài tập tổng hợp chương 13 ..... *842*
Tài liệu tham khảo *844*
Chương 14. PHÂN TÍCH TƯƠNG ỨNG **845**
1. Mở đầu 846
2. Bảng tiếp liên và các độ đo đối với đám mây số liệu 846
2.1. Mô tả hình học một bảng tiếp liên *846*

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.2. Mô tả ma trận và cấu trúc bài toán 847
3. Độ đo khoảng cách khi bình phương 851
4. Phân tích thành phần chính trong các không gian chiếu 855
4.1. Phân tích thành phần chính không trung tâm hóa và nhân tố tầm thường theo số liệu sơ cấp đối với các dòng và các cột 856
4.2. Phân tích thành phần chính hình chiếu của các đám mây 859
4.3. Công thức chuyển đổi 862
4.4. Vết ma trận và việc tái hiện dữ liệu 864
5. Biểu diễn đồng thời 865
5.1. Hình chiếu của các dòng và các cột trên cùng một siêu phẳng 865
5.2. Mô tả đồng thời mức đóng góp của các điểm dòng, cột đối với các nhân tố chính 867
6. Áp dụng phân tích tương ứng cho phân tích tương quan chính tắc đối với hai biến định tính 868
6.1. Mô tả định lượng các biến định tính 868
6.2. Mã hóa các biến định tính và phân tích tương ứng 870
6.3. Phân tích chính tắc với hai nhóm chỉ số 872
6.4. Biểu diễn đồng thời tối ưu các dấu hiệu của các cá thể 874
6.5. Mô tả đồng thời các dấu hiệu 877
6.6. Mô tả đồng thời theo hai độ đo 879
7. Phân tích tương ứng trên SPSS 880
8. Tóm tắt chương 14 887
Các ký hiệu và công thức cơ bản 888
Các thuật ngữ cơ bản 890
Câu hỏi ôn tập 891
Bài tập tổng hợp chương 14 892
Tài liệu tham khảo 893
Chương 15. CÁC PHƯƠNG PHÁP PHÂN LỚP 895
1. Mở đầu 895
2. Một số khái niệm cơ sở 896
2.1. Khoảng cách và độ khác biệt 896
2.2. Vấn đề tổ hợp trong phân nhóm 900

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3. Các phương pháp chia lớp 900
3.1. Các phương pháp kiểu đám mây động *900*
3.2. Phân lớp với các biến nhị phân *904*
4. Phân lớp theo thứ bậc 907
4.1. Thứ bậc trong phép phân lớp *908*
4.2. Các độ đo kết hợp dựa trên sự phát tán (khác biệt) *909*
4.3. Tiêu chuẩn quán tính và phương pháp Ward *910*
4.4. Trường hợp biến định tính *916*
5. Phân lớp đối với các biến 916
5.1. Phân lớp các biến - tách tổng thể theo giác độ nghiên cứu *917*
5.2. Độ đo khoảng cách đối với các biến định lượng *917*
5.3. Độ đo khoảng cách đối với các biến định tính *918*
5.4. Tiếp cận Lerman và thuật toán dựa trên tính đúng đắn của các liên hệ *918*
6. Bài toán phân lớp trên SPSS 918
6.1. Thủ tục K - Means Cluster *918*
6.2. Phân lớp có thứ bậc các cá thể *921*
7. Tóm tắt chương 15 931
Các ký hiệu và công thức cơ bản *932*
Các thuật ngữ cơ bản *933*
Câu hỏi ôn tập *933*
Bài tập tổng hợp chương 15 ..... *934*
Tài liệu tham khảo *937*
PHẦN PHỤ LỤC **938**

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

LỜI NÓI ĐẦU

"Giáo trình Lý thuyết xác suất và thống kê toán" được biên soạn cho sinh viên kinh tế và quản trị kinh doanh sau khi đã được trang bị các kiến thức cơ bản về toán cao cấp bao gồm giải tích cổ điển và đại số tuyến tính.

Mục đích của giáo trình là trang bị cho các nhà kinh tế tương lai phần đảm bảo về toán học cho quá trình thu nhập và xử lý thông tin kinh tế - xã hội sẽ được tiếp tục nghiên cứu trong các môn học khác như Lý thuyết thống kê, Dự báo kinh tế, Dân số học, Marketing, Lý thuyết đầu tư, Lý thuyết bảo hiểm, Lý thuyết tài chính... Nó cũng chuẩn bị các kiến thức cho sinh viên tiếp thu các môn học toán kinh tế sẽ nghiên cứu ở các năm sau như Kinh tế lượng, Lý thuyết phục vụ công cộng, Lý thuyết quản lý dự trữ...

Ra đời từ thế kỷ 17, lý thuyết xác suất nghiên cứu quy luật của các hiện tượng ngẫu nhiên. Dựa vào các thành tựu của lý thuyết xác suất, thống kê toán xây dựng các phương pháp ra quyết định trong điều kiện thông tin không đầy đủ. Hơn 300 năm phát triển, đến nay nội dung và các phương pháp xác suất và thống kê toán rất phong phú được ứng dụng rộng rãi trong nhiều lĩnh vực tự nhiên và xã hội khác nhau. Do khuôn khổ có hạn, giáo trình chỉ đề cập những nội dung cơ bản nhất mà nhà kinh tế hoặc kinh doanh không thể thiếu trong hành trang của mình. Những vấn đề không được đề cập như lý thuyết các quá trình ngẫu nhiên, phân tích hiệp phương sai, lý thuyết ra quyết định Bayes... bạn đọc có thể tiếp tục nghiên cứu ở các tài liệu đầy đủ hơn về xác suất và thống kê toán.

Giáo trình được viết theo quan điểm thực hành, chú trọng việc áp dụng các phương pháp của xác suất, thống kê toán trong nghiên cứu kinh tế hơn là trình bày dưới dạng thuần tuý toán học. Mỗi khái niệm, vấn đề hay phương pháp đều được minh hoạ bằng nhiều thí dụ trong những lĩnh vực thực tế khác nhau nhằm giới thiệu khả năng ứng dụng rộng rãi của các phương pháp đó, đồng thời chứng tỏ ưu thế của việc sử dụng các phương pháp toán nói chung và xác suất thống kê nói riêng trong việc giải quyết các vấn đề thực tiễn. Riêng đối với sinh viên kinh tế và quản trị kinh doanh thì điều này lại càng có ý nghĩa, nhất là khi nước ta đang chuyển mạnh sang nền kinh tế thị trường.

Để tạo điều kiện cho các nhà kinh tế tương lai sử dụng các phương pháp thống kê toán thuận lợi trong điều kiện được trang bị các phương tiện xử lý thông tin hiện đại, trong giáo trình đã đưa thêm minh họa về đảm bảo chương trình cho các phương pháp được xét. Các phần mềm này được viết bằng ngôn

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ngữ thông dụng và có thể sử dụng được ở mọi loại máy vi tính đang phổ biến ở Việt Nam hiện nay.

Mặc dù đối tượng phục vụ của giáo trình là sinh viên kinh tế, nó vẫn có thể có ích cho nhiều người quan tâm sử dụng trong công việc hoặc trong nghiên cứu phải tiến hành thu nhập và xử lý một khối lượng lớn thông tin, số liệu.

Tương ứng với giáo trình này, chúng tôi đã xuất bản một cuốn hướng dẫn giải các bài tập. Việc phân công biên soạn giáo trình này như sau:

- *PGS. TS. NGUYỄN CAO VĂN: Chủ biên và viết các chương 1, 2, 3, 4, 5, 6, 7, 8 và 9.*
- *TS. TRẦN THÁI NINH: Viết các chương 10, 11 và 12.*
- *PGS.TS. NGÔ VĂN THỨ: Viết các chương 13, 14 và 15.*

Giáo trình này có thể sử dụng như sau:

Với môn học gồm ba tín chỉ có thể dùng các chương từ chương 1 đến chương 4 và từ chương 6 đến chương 8.

Với môn học gồm bốn tín chỉ có thể dùng các chương từ chương 1 đến chương 9.

Với môn học gồm sáu tín chỉ có thể dùng các chương từ chương 1 đến chương 12.

Phần ba của giáo trình có thể sử dụng như một chuyên đề độc lập cho sinh viên năm cuối của các chuyên ngành toán kinh tế, toán tài chính, thống kê hoặc chuyên đề cho học viên cao học kinh tế và quản trị kinh doanh.

Trong lần xuất bản này, chúng tôi đã nhận được nhiều ý kiến đóng góp quý báu của các đồng nghiệp ở Bộ môn Toán kinh tế - trường Đại học Kinh tế Quốc dân và ở nhiều trường đại học khác. Chúng tôi chân thành cảm ơn tất cả những đóng góp đó. Tuy vậy chắc chắn giáo trình không tránh khỏi những hạn chế và thiếu sót. Chúng tôi mong tiếp tục nhận được các ý kiến nhận xét, phê bình của bạn đọc để tiếp tục hoàn thiện nội dung của giáo trình. Mọi đóng góp xin gửi về theo địa chỉ:

nguyencaovanktqd@yahoo.com, ninhktqd@yahoo.com.vn

và thunvtkt@neu.edu.vn

CÁC TÁC GIẢ

THƯ VIỆN HUBT

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phần thứ nhất
LÝ THUYẾT XÁC SUẤT

Lý thuyết xác suất là bộ môn toán học xác lập những quy luật tất nhiên ẩn giấu sau những hiện tượng mang tính ngẫu nhiên khi nghiên cứu một số lớn lần lặp lại cùng các hiện tượng ấy. Việc nắm bắt các quy luật này sẽ cho phép dự báo các hiện tượng ngẫu nhiên đó sẽ xảy ra như thế nào.

Để xác lập tính quy luật của các hiện tượng ngẫu nhiên, lý thuyết xác suất sử dụng phương pháp suy diễn bằng cách thiết lập các mô hình lý thuyết mang tính khái quát và xem xét các điều kiện để các quy luật đó được bộc lộ trên các hiện tượng cụ thể.

Các phương pháp của lý thuyết xác suất được ứng dụng rộng rãi trong việc giải quyết các bài toán thuộc nhiều lĩnh vực khác nhau của khoa học tự nhiên, kỹ thuật và kinh tế - xã hội.

Phần này sẽ nghiên cứu những nội dung cơ bản nhất của xác suất.

Chương 1 nghiên cứu các phương pháp tính xác suất của một biến cố.

Chương 2 nghiên cứu biến ngẫu nhiên và phương pháp xác lập phân phối xác suất của một biến ngẫu nhiên.

Chương 3 nghiên cứu một số quy luật phân phối xác suất thường được sử dụng trong phân tích kinh tế xã hội.

Chương 4 nghiên cứu phương pháp xác lập quy luật phân phối xác suất của các biến ngẫu nhiên nhiều chiều.

Chương 5 được xem là phần kết thúc của xác suất, nghiên cứu cơ sở lý thuyết của xác suất thông qua các định lý hội tụ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ


THƯ VIỆN
HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 1
BIẾN CỐ NGẪU NHIÊN VÀ XÁC SUẤT

1. MỞ ĐẦU

Chương này giới thiệu các khái niệm nền tảng của xác suất là các loại biến cố, xác suất của biến cố và các phương pháp tìm xác suất của một biến cố vì có thể nói rằng mọi hiện tượng mà ta gặp trong tự nhiên đều là các biến cố ngẫu nhiên. Vì thế việc đánh giá được một cách chính xác khả năng xảy ra của chúng sẽ là cơ sở để dự báo khả năng xảy ra của các biến cố này trong tương lai. Trong kinh tế và xã hội thì việc tìm được xác suất của một biến cố sẽ là cơ sở cho việc ra các quyết định trong quản lý và kinh doanh. Trong thực tế có thể gặp nhiều loại biến cố khác nhau, vì vậy tuỳ theo bản chất và tính chất của từng loại biến cố mà sử dụng các công cụ khác nhau để tìm xác suất của chúng.

Mục 2 và 3 đưa ra định nghĩa về biến cố, phân loại biến cố và khái niệm về xác suất của biến cố. Từ mục 4 đến mục 6 trình bày các định nghĩa khác nhau dùng để tìm xác suất của các biến cố giản đơn. Mục 7 đề cập đến hai nguyên lý cơ bản của xác suất là nguyên lý xác suất lớn và nguyên lý xác suất nhỏ. Từ mục 8 đến mục 10 trình bày các định lý xác suất dùng để tìm xác suất của các biến cố phức hợp.

2. PHÉP THỬ VÀ CÁC LOẠI BIẾN CỐ

Trong tự nhiên và xã hội, mỗi hiện tượng đều gắn liền với một nhóm các điều kiện cơ bản và các hiện tượng đó chỉ có thể xảy ra khi nhóm các điều kiện cơ bản gắn liền với nó được thực hiện. Do đó, khi muốn nghiên cứu một hiện tượng, ta cần thực hiện nhóm các điều kiện cơ bản ấy. Chẳng hạn, nếu muốn quan sát việc xuất hiện mặt sấp hay mặt ngửa của một đồng xu, ta phải gieo đồng xu xuống đất; còn để xét xem viên đạn trúng bia hay trượt, ta phải bắn các viên đạn; khi muốn nghiên cứu chất lượng của một lô sản phẩm, ta phải lấy ngẫu nhiên một hoặc một số sản phẩm của lô sản phẩm đó v.v...

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Việc thực hiện một nhóm các điều kiện cơ bản để quan sát một hiện tượng nào đó có thể xảy ra hay không được gọi là thực hiện một phép thử, còn hiện tượng có thể xảy ra trong kết quả của phép thử đó được gọi là biến cố.

Thí dụ 1. Gieo một con xúc xắc xuống đất là một phép thử, còn việc lật lên một mặt nào đó là biến cố.

Thí dụ 2. Bắn một phát súng vào bia. Việc bắn súng là phép thử, còn việc trúng vào một miền nào đó của bia là biến cố.

Thí dụ 3. Từ một lô sản phẩm gồm chính phẩm và phế phẩm lấy ngẫu nhiên một sản phẩm. Việc lấy sản phẩm là phép thử, còn việc lấy được chính phẩm hay phế phẩm là biến cố.

Như vậy, ta thấy rằng một biến cố chỉ có thể xảy ra khi một phép thử gắn liền với nó được thực hiện. Trong thực tế có thể xảy ra các loại biến cố sau đây:

+ *Biến cố chắc chắn*: Là biến cố nhất định sẽ xảy ra khi thực hiện một phép thử. Biến cố chắc chắn được ký hiệu là U.

Thí dụ 4. Thực hiện phép thử là gieo một con xúc xắc. Gọi U là biến cố "Xuất hiện mặt có số chấm nhỏ hơn hoặc bằng 6", U là biến cố chắc chắn.

+ *Biến cố không thể có*: Là biến cố nhất định không xảy ra khi thực hiện phép thử. Biến cố không thể có được ký hiệu là V.

Thí dụ 5. Gieo một con xúc xắc, gọi V là biến cố "Xuất hiện mặt có 7 chấm", V là biến cố không thể có.

+ *Biến cố ngẫu nhiên*: Là biến cố có thể xảy ra hoặc không xảy ra khi thực hiện một phép thử. Các biến cố ngẫu nhiên được ký hiệu là A, B, C... hoặc A_1, A_2,... A_n, B_1, B_2,...B_m.

Thí dụ 6. Gieo một con xúc xắc, gọi A là biến cố "Xuất hiện mặt 1 chấm", A là biến cố ngẫu nhiên.

Thí dụ 7. Bắn một phát đạn vào bia, gọi B là biến cố "Trúng vòng 10", B là biến cố ngẫu nhiên.

Tất cả các biến cố mà chúng ta gặp trong thực tế đều thuộc về một trong ba loại biến cố kể trên. Tuy nhiên, các biến cố ngẫu nhiên là các biến cố thường gặp hơn cả.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3. XÁC SUẤT CỦA BIẾN CỐ

Như trên đã thấy, việc biến cố ngẫu nhiên xảy ra hay không xảy ra trong kết quả của phép thử là điều không thể đoán trước được. Tuy nhiên, bằng trực quan ta có thể nhận thấy các biến cố ngẫu nhiên khác nhau có những khả năng xảy ra khác nhau. Chẳng hạn biến cố "Xuất hiện mặt sấp" khi gieo một đồng xu sẽ có khả năng xảy ra lớn hơn nhiều so với biến cố "Xuất hiện mặt một chấm" khi gieo một con xúc xắc. Hơn nữa, khi lặp đi lặp lại nhiều lần cùng một phép thử trong những điều kiện như nhau, người ta thấy tính chất ngẫu nhiên của biến cố mất dần đi và khả năng xảy ra của biến cố sẽ được thể hiện theo những quy luật nhất định. Từ đó ta thấy có khả năng định lượng (đo lường), khả năng khách quan xuất hiện một biến cố nào đó.

Xác suất của một biến cố là một con số đặc trưng khả năng khách quan xuất hiện biến cố đó khi thực hiện phép thử.

Ta chú ý rằng đây là khả năng khách quan, do những điều kiện xảy ra của phép thử quy định chứ không tùy thuộc vào ý muốn chủ quan của con người.

Như vậy, bản chất xác suất của một biến cố là một con số xác định. Để tính xác suất của một biến cố, người ta xây dựng các định nghĩa và định lý xác suất. Đối với các biến cố giản đơn thì áp dụng các định nghĩa xác suất, còn đối với các biến cố phức hợp thì dùng các định lý xác suất.

4. ĐỊNH NGHĨA CỔ ĐIỂN VỀ XÁC SUẤT

4.1. Thí dụ

Giả sử thực hiện một phép thử là gieo một con xúc xắc đều đặn và đồng chất. Gọi A là biến cố "xuất hiện mặt chẵn chấm". Ta phải xác định xác suất của biến cố A.

Khi gieo một con xúc xắc đều đặn và đồng chất, ta thấy có 6 trường hợp có thể xảy ra là: xuất hiện các mặt 1 chấm, 2 chấm,... 6 chấm. Những trường hợp này thoả mãn hai điều kiện: Trước hết chúng duy nhất, tức là trong kết quả của phép thử xảy ra một và chỉ một trong các trường hợp đó. Sau nữa đây là những trường hợp có khả năng xảy ra như nhau. Các trường hợp thoả mãn hai điều kiện nói trên được gọi là các trường hợp (kết cục) duy nhất đồng khả năng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trong số 6 kết cục duy nhất đồng khả năng đó ta thấy chỉ có 3 kết cục mà nếu xảy ra thì biến cố A sẽ xảy ra, đó là những kết cục được mặt 2 chấm. 4 chấm và 6 chấm. Những kết cục làm cho biến cố xảy ra được gọi là kết cục thuận lợi cho biến cố.

Như vậy đứng về mặt trực quan ta thấy khả năng xảy ra của biến cố A là 3 phần 6, tức là 1 phần 2. Đó chính là cách xác định xác suất của biến cố theo quan điểm cổ điển.

4.2. Định nghĩa

Xác suất xuất hiện biến cố A trong một phép thử là tỉ số giữa số kết cục thuận lợi cho A và tổng số các kết cục duy nhất đồng khả năng có thể xảy ra khi thực hiện phép thử đó.

Nếu ký hiệu P (A) là xác suất của biến cố A, m là số kết cục thuận lợi cho biến cố A, n là số kết cục duy nhất đồng khả năng của phép thử, ta có công thức sau:

$$P(A) = \frac{m}{n} \qquad (1.1)$$

trong đó: m, n là các số nguyên không âm

4.3. Các tính chất của xác suất

Từ định nghĩa cổ điển về xác suất ta có thể suy ra các tính chất sau đây:

a. Xác suất của biến cố ngẫu nhiên là một số dương nằm trong khoảng giữa 0 và 1.

$$0 < P(A) < 1$$

Thật vậy, vì số kết cục thuận lợi cho một biến cố ngẫu nhiên luôn thoả mãn $0 < m < n$ do đó:

$$0 < \frac{m}{n} < 1$$

từ đó: $0 < P(A) < 1$

b. Xác suất của biến cố chắc chắn bằng một.

$$P(U) = 1$$

Thật vậy, nếu U là biến cố chắc chắn thì tất cả các kết cục duy nhất đồng khả năng có thể xảy ra trong phép thử đều thuận lợi cho biến cố xảy ra. Do đó $m = n$ và ta có:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P(U) = \frac{m}{n} = 1$$

c. Xác suất của biến cố không thể có bằng không.

$$P(V) = 0$$

Thật vậy, nếu V là biến cố không thể có thì trong số các kết cục duy nhất đồng khả năng có thể xảy ra trong phép thử không có kết cục nào thuận lợi cho biến cố xảy ra. Như vậy $m = 0$, do đó:

$$P(V) = \frac{m}{n} = 0$$

Như vậy, xác suất của một biến cố bất kỳ luôn thoả mãn điều kiện:

$$0 \le P(A) \le 1 \qquad (1.2)$$

Đối với các tính chất trên ta chú ý rằng mệnh đề đảo của hai tính chất b và c chưa chắc đã đúng. Tức là, nếu một biến cố có xác suất bằng 1 thì chưa chắc đã là biến cố chắc chắn và nếu một biến cố có xác suất bằng 0 thì chưa chắc đã là biến cố không thể có.

4.4. Các phương pháp tính xác suất bằng định nghĩa cổ điển

1. Phương pháp suy luận trực tiếp: Nếu số các kết cục trong phép thử là khá nhỏ và việc suy đoán là khá đơn giản thì có thể sử dụng phương pháp suy luận trực tiếp.

Thí dụ 1. Trong bình có a quả cầu trắng và b quả cầu đen (a, b nguyên không âm). Lấy ngẫu nhiên một quả cầu. Tìm xác suất để lấy được quả cầu trắng.

Giải. Gọi A là biến cố "lấy được quả cầu trắng". Khi lấy ngẫu nhiên từ bình ra một quả cầu, ta có thể lấy được bất kỳ quả nào trong số a + b quả cầu có trong bình. Như vậy số kết cục duy nhất đồng khả năng có thể xảy ra trong phép thử $n = a + b$.

Biến cố A sẽ xảy ra khi ta lấy được một trong số a quả cầu trắng. Như vậy số kết cục thuận lợi $m = a$. Từ đó theo định nghĩa cổ điển về xác suất, ta có:

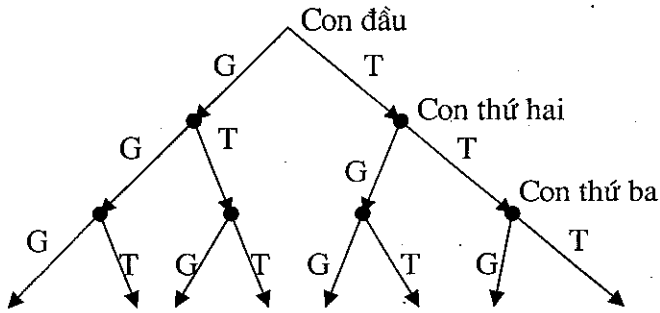
$$P(A) = \frac{m}{n} = \frac{a}{a+b}$$

2. Phương pháp dùng sơ đồ Venn: Khi số kết cục là khá lớn và việc suy đoán phức tạp hơn thì có thể dùng sơ đồ Venn, tức là mô tả các kết cục của phép thử dưới dạng sơ đồ để dễ nhận biết hơn. Trong thực tế có thể

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

dùng các loại sơ đồ sau:

a. Sơ đồ hình cây



Hình 1.1. Sơ đồ hình cây

Thí dụ 2. Giả sử xác suất sinh con trai và con gái là như nhau và bằng 0,5. Một gia đình có 3 con. Tìm xác suất để gia đình đó có 2 con gái.

Giải. Gọi A là biến cố "Gia đình đó có 2 con gái". Số kết cục đồng khả năng có thể suy ra từ sơ đồ trên hình 1.1.

Như vậy, tổng số ta có $n = 8$ kết cục đồng khả năng là GGG, GGT, GTG, GTT, TGG, TGT, TTG và TTT. Trong đó có 3 kết cục thuận lợi để có 2 con gái.

Vậy $P(A) = \frac{3}{8}$.

b. Sơ đồ dạng bảng

Thí dụ 3. Gieo một con xúc xắc hai lần. Tìm xác suất để trong đó có một lần được 6 chấm.

Giải. Gọi A là biến cố "Trong 2 lần gieo con xúc xắc có đúng 1 lần được mặt 6 chấm". Số kết cục đồng khả năng của phép thử có thể mô tả dưới dạng bảng sau (hình 1.2):

Như vậy ta có $n = 36$ kết cục đồng khả năng. Trong đó có $m = 10$ kết cục thuận lợi. Vậy:

$$P(A) = \frac{m}{n} = \frac{10}{36} = \frac{5}{18}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

I \ II	1	2	3	4	5	6
1	11	12	13	14	15	16
2	21	22	23	24	25	26
3	31	32	33	34	35	36
4	41	42	43	44	45	46
5	51	52	53	54	55	56
6	61	62	63	64	65	66

***Hình 1.2.* Sơ đồ dạng bảng**

c. Sơ đồ dạng tập hợp

Thí dụ 4. Trong một lớp 50 học sinh có:

20 người chơi bóng đá

15 người chơi bóng chuyền

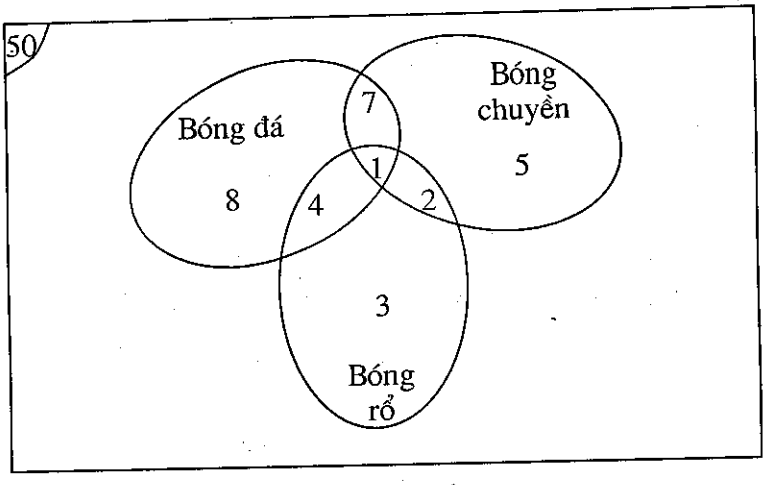
10 người chơi bóng rổ

8 người chơi bóng đá và bóng chuyền

5 người chơi bóng đá và bóng rổ

3 người chơi bóng chuyền và bóng rổ

1 người chơi bóng đá, bóng chuyền và bóng rổ.



***Hình 1.3.* Sơ đồ dạng tập hợp**

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lấy ngẫu nhiên 1 học sinh. Tìm xác suất để người đó chơi ít nhất 1 trong các môn bóng nói trên.

Giải. Gọi A là biến cố "Lấy ngẫu nhiên một học sinh thì người đó chơi ít nhất một môn bóng". Số kết cục đồng khả năng có thể mô tả dưới dạng tập hợp như trên hình 1.3.

Vậy trong $n = 50$ kết cục đồng khả năng thì số kết cục thuận lợi $m = 8+5+3+7+4+2+1 = 30$.

Vậy $P(A) = \frac{m}{n} = \frac{30}{50} = 0,6$.

3. Phương pháp dùng các công thức của giải tích tổ hợp

Nếu số kết cục của phép thử là rất lớn mà không thể suy đoán trực tiếp được thì có thể dùng các công thức của giải tích tổ hợp, chủ yếu là các công thức chỉnh hợp, chỉnh hợp lặp, hoán vị và tổ hợp để tính toán.

Thí dụ 5. Một người khi gọi điện thoại quên mất hai số cuối của số điện thoại và chỉ nhớ được rằng chúng khác nhau. Tìm xác suất để quay ngẫu nhiên một lần được đúng số cần gọi.

Giải. Gọi B là biến cố "Quay ngẫu nhiên một lần được đúng số cần gọi". Số kết cục đồng khả năng là tất cả các phương thức để lập nên một cặp hai số khác nhau từ 10 số tự nhiên đầu tiên. Nó bằng số chỉnh hợp chập 2 từ 10. Như vậy $n = A_{10}^2 = 10.9 = 90$. Còn số kết cục thuận lợi cho biến cố B xảy ra chỉ có một kết cục. Do đó theo định nghĩa cổ điển:

$$P(B) = \frac{m}{n} = \frac{1}{90}$$

Thí dụ 6. Trong bình có 6 quả cầu giống nhau được đánh số, lấy ngẫu nhiên lần lượt từng quả cầu. Tìm xác suất để số của quả cầu được lấy ra trùng với số thứ tự của lần lấy.

Giải. Gọi A là biến cố "Số của các quả cầu trùng với số thứ tự của lần lấy". Số kết cục đồng khả năng trong phép thử này là tất cả các phương thức để lần lượt lấy được 6 quả cầu ra khỏi bình. Nó bằng số hoán vị của 6 phần tử. Do đó $n = P_6 = 6! = 720$. Trong đó chỉ có một kết cục thuận lợi cho biến cố A xảy ra là lấy được các quả cầu theo trình tự các số 1, 2, 3, 4, 5, 6.

Như vậy: $P(A) = \frac{m}{n} = \frac{1}{720}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 7. Một hộp có 10 sản phẩm, trong đó có 6 chính phẩm và 4 phế phẩm. Lấy ngẫu nhiên từ hộp đó 3 sản phẩm. Tìm xác suất để:

a) Cả 3 sản phẩm lấy ra đều là chính phẩm.

b) Trong ba sản phẩm lấy ra có đúng 2 chính phẩm.

Giải. a) Gọi A là biến cố "Lấy được 3 chính phẩm". Số kết cục đồng khả năng trong phép thử bằng số tổ hợp chập 3 từ 10 phần tử. Như vậy $n = C_{10}^3 = 120$. Số kết cục thuận lợi cho A xảy ra bằng số tổ hợp chập 3 (chính phẩm) từ 6 (chính phẩm) cho trước. Vậy $m = C_6^3 = 20$.

Do đó:

$$P(A) = \frac{m}{n} = \frac{20}{120} = \frac{1}{6}.$$

b) Gọi B là biến cố "Trong 3 sản phẩm lấy ra có đúng 2 chính phẩm". Số kết cục thuận lợi cho B xảy ra bằng số tổ hợp chập 2 (chính phẩm) từ 6 (chính phẩm) cho trước, bằng C_6^2. Ngoài ra sản phẩm thứ ba phải là phế phẩm. Ta có C_4^1 cách lấy được 1 phế phẩm. Như vậy số kết cục thuận lợi cho B xảy ra bằng $m = C_6^2.C_4^1$

Do đó: $P(B) = \frac{m}{n} = \frac{C_6^2.C_4^1}{C_{10}^3} = \frac{1}{2}$

Thí dụ 8: Biết rằng trong 3 tháng cuối năm có 5 máy đã bị hỏng. Tìm xác suất để không có ngày nào có quá 1 máy bị hỏng.

Giải. Gọi A là biến cố "Không có ngày nào có quá 1 máy bị hỏng". Số kết cục đồng khả năng là số chỉnh hợp lặp chập 5 từ 92 phần tử $(n = \overline{A}_{92}^5 = 92^5)$.

Số kết cục thuận lợi là số chỉnh hợp chập 5 từ 92 phần tử $(m = A_{92}^5 = 88.89.90.91.92)$.

Vậy: $P(A) = \frac{m}{n} = \frac{88.89.90.91.92}{92^5} = 0,8954.$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

1.1. Gieo một con xúc xắc đối xứng và đồng chất.

Tìm xác suất để được:

a. Mặt sáu chấm xuất hiện.

b. Mặt có số chẵn chấm xuất hiện.

ĐS: $P_a = 1/6 \quad P_b = 0,5$.

1.2. Có 100 tấm bìa hình vuông như nhau được đánh số từ 1 đến 100. Ta lấy ngẫu nhiên một tấm bìa. Tìm xác suất:

a. Được một tấm bìa có số không có chữ số 5.

b. Được một tấm bìa có số chia hết cho 2 hoặc cho 5 hoặc cả cho 2 và cho 5.

ĐS: $P_a = 0.81 \quad P_b = 0,6$.

1.3. Một hộp có a quả cầu trắng và b quả cầu đen (a, b nguyên, không âm) 1). Lấy ngẫu nhiên lần lượt hai quả cầu

a. Tìm xác suất để quả cầu thứ nhất là trắng.

b. Tìm xác suất để quả cầu thứ hai là trắng nếu biết rằng quả thứ nhất là trắng.

c. Tìm xác suất để quả cầu thứ nhất là trắng nếu biết rằng quả thứ hai là trắng.

ĐS: $P_a = \frac{a}{a+b}$: $P_b = \frac{a-1}{a+b-1}$; $P_c = \frac{a-1}{a+b-1}$

1.4. Một hộp có a quả cầu trắng và b quả cầu đen (a, b nguyên, không âm) 1). Lấy ngẫu nhiên ra lần lượt từng quả cầu.

Tìm xác suất để:

a. Quả cầu thứ hai là trắng.

b. Quả cầu cuối cùng là trắng.

ĐS: $P_a = \frac{a}{a+b}$: $P_b = \frac{a}{a+b}$

1.5. Gieo đồng thời hai đồng xu. Tìm xác suất để được:

a. Hai mặt cùng sấp xuất hiện.

b. Một sấp một ngửa.

c. Có ít nhất một mặt sấp.

ĐS: $P_a = 0,25$; $P_b = 0,5$, $P_c = 0,75$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1.6. Gieo đồng thời hai con xúc xắc. Tìm xác suất để được hai mặt.

a. Có tổng số chấm bằng 7

b. Có tổng số chấm nhỏ hơn 8

c. Có ít nhất một mặt 6 chấm

ĐS: $P_a = \frac{1}{6}$; $P_b = \frac{7}{12}$; $P_c = \frac{11}{36}$

1.7. Ba người khách cuối cùng ra khỏi nhà bỏ quên mũ. Chủ nhà không biết rõ chủ của những chiếc mũ đó nên gửi trả cho họ một cách ngẫu nhiên.

Tìm xác suất để:

a. Cả 3 người cùng được trả sai mũ.

b. Có đúng một người được trả đúng mũ.

c. Có đúng hai người được trả đúng mũ.

d. Cả ba người đều được trả đúng mũ.

ĐS: $P_a = \frac{1}{3}$; $P_b = \frac{1}{2}$; $P_c = 0$; $P_d = \frac{1}{6}$

1.8. Một lớp sinh viên có 50% học tiếng Anh, 40% học tiếng Pháp, 30% học tiếng Đức, 20% học tiếng Anh và tiếng Pháp, 15% học tiếng Anh và tiếng Đức, 10% học tiếng Pháp và tiếng Đức, 5% học cả ba thứ tiếng Anh, Pháp, Đức. Tìm xác suất để khi lấy ngẫu nhiên 1 sinh viên lớp này thì người đó:

a. Học ít nhất 1 trong 3 thứ ngoại ngữ kể trên.

b. Chỉ học tiếng Anh và tiếng Đức

c. Chỉ học tiếng Pháp

d. Học tiếng Pháp biết rằng người đó học tiếng Anh

ĐS: $P_a = 0,8$; $P_b = 0,1$; $P_c = 0,15$; $P_d = 0,4$

1.9. Một người gọi điện thoại cho bạn nhưng quên mất 3 chữ số cuối và chỉ nhớ rằng chúng khác nhau. Tìm xác suất để người đó quay số một lần được đúng số điện thoại của bạn.

ĐS: $P = \frac{1}{720}$

1.10. Trong một hòm đựng 10 chi tiết đạt tiêu chuẩn và 5 chi tiết không đạt

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

tiêu chuẩn. Lấy đồng thời 3 chi tiết. Tính xác suất:

a. Cả 3 chi tiết lấy ra thuộc loại đạt tiêu chuẩn.

b. Trong số 3 chi tiết lấy ra có 2 chi tiết đạt tiêu chuẩn.

ĐS: $P_a = 0{,}264$; $P_b = 0{,}495$.

1.11. Một nhi đồng tập xếp chữ. Em có các chữ N, Ê, H, G, H, N. Tìm xác suất để em đó trong khi sắp ngẫu nhiên được chữ NGHÊNH.

ĐS: $P = \frac{1}{180}$

1.12. Thang máy của một toà nhà 7 tầng xuất phát từ tầng một với 3 khách. Tìm xác suất để:

a. Tất cả cùng ra ở tầng bốn.

b. Tất cả cùng ra ở một tầng.

c. Mỗi người ra ở một tầng khác nhau.

ĐS:

$$P_a = \frac{1}{216} \quad P_b = \frac{1}{36} \quad P_c = \frac{5}{9}$$

1.13. Trên giá sách có xếp ngẫu nhiên một tuyển tập của tác giả X gồm 12 cuốn. Tìm xác suất để các tập được xếp theo thứ tự tăng dần hoặc giảm dần.

ĐS: $P = \frac{2}{479001600}$

1.14. Lấy ngẫu nhiên ba quân bài từ một cỗ bài gồm 52 quân. Tìm xác suất để:

a. Được 3 quân Át.

b. Được 1 quân Át.

ĐS: $P_a = \frac{1}{5525}$; $P_b = \frac{1128}{5525}$

1.15. Một lô hàng gồm 6 chính phẩm và 4 phế phẩm được chia ngẫu nhiên thành 2 phần có số sản phẩm bằng nhau. Tìm xác suất để mỗi phần đều có số chính phẩm như nhau.

ĐS: $P = \frac{10}{21}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1.16. Mỗi vé xổ số có 5 chữ số. Tìm xác suất để một người mua một vé được vé:

a. Có 5 chữ số khác nhau.

b. Có 5 chữ số đều là lẻ.

ĐS: $P_a = 0,3024$; $P_b = 0,03125$.

1.17. Năm người A, B, C, D, E ngồi một cách ngẫu nhiên vào một chiếc ghế dài. Tìm xác suất để:

a. C ngồi chính giữa

b. A và B ngồi ở hai đầu ghế.

ĐS: $P_a = 0,2$; $P_b = 0,1$.

1.18. Trong một chiếc hộp có n (n>1) quả cầu được đánh số từ 1 tới n. Một người lấy ngẫu nhiên cùng một lúc ra hai quả. Tính xác suất để người đó lấy được một quả có số hiệu nhỏ hơn k và một quả có số hiệu lớn hơn k ($1 < k < n$).

1.19. Gieo n con xúc xắc đối xứng và đồng chất. Tìm xác suất để được tổng số chấm là n + 1.

4.5. Ưu điểm và hạn chế của định nghĩa cổ điển về xác suất

Định nghĩa cổ điển về xác suất có một ưu điểm cơ bản là để tìm xác suất của biến cố, ta không cần phải tiến hành phép thử (phép thử chỉ tiến hành một cách giả định). Ngoài ra nếu đáp ứng đầy đủ các yêu cầu của định nghĩa thì nó cho phép ta tìm được một cách chính xác giá trị của xác suất.

Tuy nhiên định nghĩa cổ điển về xác suất cũng có những hạn chế đáng kể. Nó đòi hỏi là số kết cục duy nhất đồng khả năng có thể xảy ra trong phép thử phải là hữu hạn. Trong thực tế có nhiều phép thử mà trong đó số kết cục có thể là vô hạn. Trong những trường hợp này, định nghĩa cổ điển về xác suất không áp dụng được. Chỉ riêng điều đó đã hạn chế khả năng áp dụng của định nghĩa cổ điển. Thật ra hạn chế này có thể khắc phục được bằng cách mở rộng định nghĩa cổ điển.

Hạn chế lớn nhất của định nghĩa cổ điển là trong thực tế nhiều khi không thể biểu diễn kết quả của phép thử dưới dạng tập hợp các kết cục duy nhất và đồng khả năng. Thường thì tính đồng khả năng của các kết cục được suy ra từ tính đối xứng. Chẳng hạn khi tung một con xúc xắc ta giả thiết

THƯ VIỆN HUBT
TRƯỜNG Đ.H KINH DOANH VÀ CÔNG NGHỆ HÀ NỘI
THƯ VIỆN
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

rằng nó đều đặn và đồng nhất. Tuy nhiên những bài toán trong đó ta có thể đưa ra các giả thiết về tính đối xứng rất hiếm khi gặp trong thực tế.

Vì lý do đó mà ngoài định nghĩa cổ điển về xác suất, trong thực tế người ta còn sử dụng định nghĩa xác suất theo quan điểm thống kê sau đây.

5. ĐỊNH NGHĨA THỐNG KÊ VỀ XÁC SUẤT

5.1. Định nghĩa 1

Tần suất xuất hiện biến cố trong n phép thử là tỷ số giữa số phép thử trong đó biến cố xuất hiện và tổng số phép thử được thực hiện.

Như vậy, nếu ký hiệu số phép thử là n, số lần xuất hiện biến cố A là k, tần suất xuất hiện biến cố A là f(A) thì:

$$f(A) = \frac{k}{n} \tag{1.3}$$

Cùng với khái niệm xác suất, khái niệm tần suất là một trong những khái niệm cơ bản của lý thuyết xác suất.

Thí dụ 1. Khi kiểm tra ngẫu nhiên 80 sản phẩm do một máy sản xuất, người ta phát hiện ra 3 phế phẩm. Gọi A là biến cố "Xuất hiện phế phẩm". Vậy tần suất xuất hiện phế phẩm bằng:

$$f(A) = \frac{3}{80}$$

Thí dụ 2. Bắn 50 phát đạn vào bia thấy có 47 phát trúng. Gọi A là "Biến cố bắn trúng bia". Tần suất của việc bắn trúng bia bằng:

$$f(A) = \frac{47}{50}$$

Người ta nhận thấy nếu tiến hành các thí nghiệm trong những điều kiện như nhau và số phép thử khá lớn thì tần suất thể hiện tính ổn định của nó khá rõ ràng. Tính chất này thể hiện ở chỗ là *nếu tiến hành một số khá lớn cùng một phép thử thì tần suất giao động rất ít xung quanh một giá trị nào đó*. Ta sẽ thấy rõ điều đó qua thí dụ sau.

Thí dụ 3. Để nghiên cứu khả năng xuất hiện mặt sấp khi tung một đồng xu, người ta tiến hành tung một đồng xu nhiều lần và thu được kết quả sau đây:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Người làm thí nghiệm	Số lần tung (n)	Số lần được mặt sấp (k)	Tần suất $f(A)=\frac{k}{n}$
Buffon	4040	2048	0,5069
Pearson	12000	6019	0,5016
Pearson	24000	12012	0,5005

Qua thí dụ trên ta thấy khi số phép thử tăng lên thì tần suất xuất hiện mặt sấp sẽ dao động ngày càng ít hơn xung quanh giá trị không đổi là 0,5. Điều đó cho phép hy vọng rằng khi số phép thử tăng lên vô hạn, tần suất sẽ hội tụ về giá trị 0,5.

Tính ổn định của tần suất là cơ sở để đưa ra định nghĩa thống kê về xác suất.

5.2. Định nghĩa 2

Xác suất xuất hiện biến cố A trong một phép thử là một số p không đổi mà tần suất f xuất hiện biến cố đó trong n phép thử sẽ dao động rất ít xung quanh nó khi số phép thử tăng lên vô hạn.

Như vậy về mặt thực tế với số phép thử đủ lớn ta có thể lấy:

$$P(A) \approx f(A)$$

Cơ sở của cách lấy xấp xỉ này sẽ được trình bày ở chương V.

5.3. Ưu điểm và hạn chế của định nghĩa thống kê về xác suất

Định nghĩa thống kê về xác suất có ưu điểm lớn là nó không đòi hỏi những điều kiện áp dụng như đối với định nghĩa cổ điển. Nó hoàn toàn dựa trên các quan sát thực tế để làm cơ sở kết luận về xác suất xảy ra của một biến cố.

Tuy nhiên, định nghĩa thống kê về xác suất chỉ áp dụng được đối với các hiện tượng ngẫu nhiên mà tần suất của nó có tính ổn định. Hơn nữa, để xác định một cách tương đối chính xác giá trị của xác suất ta phải tiến hành trên thực tế một số đủ lớn các phép thử. Nói cách khác xác suất theo quan điểm thống kê là xác suất được tính sau khi phép thử đã thực hiện. Trong nhiều bài toán thực tế rất khó hoặc không thể tiến hành nhiều phép thử để dựa vào đó mà tính xác suất của một biến cố.

Trong nhiều trường hợp nếu không cần thiết phải thực sự tiến hành phép thử thực tế, để khắc phục hạn chế trên, người ta có thể mô phỏng kết quả của các phép thử bằng cách sử dụng bảng số ngẫu nhiên (Phụ lục 10).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Để làm điều đó, người ta chọn ngẫu nhiên một dòng của bảng số ngẫu nhiên và dùng các chữ số của dòng đó để thay thế cho kết quả của phép thử. Chẳng hạn, nếu tiến hành gieo một con xúc xắc trong 10 lần thì có thể mô tả kết quả gieo bằng cách chọn ngẫu nhiên một dòng của bảng số ngẫu nhiên. Giả sử ta chọn dòng thứ nhất và thu được dãy số sau:

$$\underset{\text{bỏ}}{1}\,55\,\underset{\text{bỏ}}{9}\,\underset{\text{bỏ}}{90}\,6\,8\,9\,2\,\underset{\text{bỏ}}{90}\,\underset{\text{bỏ}}{8}\,3\,\underset{\text{bỏ}}{0}\,\underset{\text{bỏ}}{3}\,\underset{\text{bỏ}}{8}\,5\,\underset{\text{bỏ}}{08}\,\underset{\text{bỏ}}{89}\,54$$

Vậy kết quả của phép thử có thể mô phỏng như sau: Lần đầu được 1 điểm, lần 2 được 5 điểm, lần 3 được 5 điểm, lần 4 được 6 điểm... và dựa vào kết quả đó để xác định tần suất.

Bài tập

1.20. Tần suất xuất hiện biến cố viên đạn trúng đích của một xạ thủ là 0,85. Tìm số viên đạn trúng đích của xạ thủ đó nếu người bắn 200 viên đạn.

ĐS: 170.

1.21. Có thể xem xác suất sinh được con trai là bao nhiêu nếu theo dõi 88200 trẻ sơ sinh ở một vùng thấy có 45.600 con trai.

ĐS: $P \approx 0{,}517$.

1.22. Dùng bảng số ngẫu nhiên để mô phỏng kết quả của 50 lần gieo một con xúc xắc. Từ đó tìm tần suất xuất hiện các mặt 1, 2, ..., 6 chấm và mô tả bằng đồ thị. Đồ thị các tần suất này sẽ như thế nào khi gieo 1 triệu lần.

6. MỘT SỐ ĐỊNH NGHĨA KHÁC VỀ XÁC SUẤT

Trong thực tế ngoài định nghĩa cổ điển và định nghĩa thống kê về xác suất người ta còn sử dụng một số định nghĩa sau về xác suất.

6.1. Định nghĩa hình học về xác suất

Định nghĩa hình học về xác suất có thể sử dụng khi xác suất để một điểm ngẫu nhiên rơi vào một phần nào đó của một miền cho trước tỷ lệ với

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

độ đo của miền đó (độ dài, diện tích, thể tích v.v...) và không phụ thuộc vào vị trí và dạng thức của miền đó.

Nếu độ đo hình học của toàn bộ miền cho trước là S, còn độ đo hình học của một phần A nào đó của nó là S_A thì xác suất để điểm ngẫu nhiên rơi vào phần A sẽ bằng:

$$p = \frac{S_A}{S}$$

trong đó S và S_A có thể có độ đo bất kỳ.

Như vậy, có thể xem định nghĩa hình học về xác suất là sự mở rộng tương ứng của định nghĩa cổ điển về xác suất.

6.2. Xác suất chủ quan

Xác suất chủ quan được định nghĩa như sự đánh giá chủ quan của một cá nhân nào đó về khả năng xảy ra của biến cố. Sự đánh giá này chủ yếu dựa vào những nhận xét cá nhân, thông tin ngoại lai, trực giác hoặc các kinh nghiệm tích luỹ được của mỗi cá nhân liên quan đến hiện tượng được xem xét. Như vậy, với cùng một hiện tượng thì xác suất chủ quan của người này có thể khác biệt rất nhiều so với xác suất chủ quan của người khác, vì vậy, nó còn được gọi là xác suất của cá nhân.

Cách tiếp cận này chủ yếu được sử dụng khi không thể áp dụng các phương pháp tính xác suất một cách khách quan, chẳng hạn tình huống không thể quan niệm được hết các kết cục có thể có của một phép thử hay không thể lặp lại nhiều lần một phép thử để xác định tần suất xuất hiện biến cố.

6.3. Định nghĩa tiên đề về xác suất

Vào những năm 30 của thế kỷ 20, nhà toán học người Nga là Kolmogorov đã xây dựng hệ tiên đề làm cơ sở cho việc định nghĩa một cách hoàn chỉnh khái niệm xác suất về mặt lý thuyết. Hệ tiên đề được xây dựng trên cơ sở khái niệm về không gian các biến cố sơ cấp E_1, E_2...E_n, thực tế là tập hợp mọi kết cục có thể có của phép thử. Lúc đó mỗi biến cố A có thể quan niệm như một tập hợp con của không gian đó. Từ đó ta có các tiên đề sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tiên đề 1: Với mọi biến cố A đều có $P(A) \geq 0$.

Tiên đề 2: Nếu $E_1E_2...E_n$ tạo nên không gian các biến cố sơ cấp thì:

$$P(E_1) + P(E_2) + ... + P(E_n) = 1$$

Tiên đề 3: Nếu các biến cố $A_1A_2...A_n...$ là các tập hợp con không giao nhau của các biến cố sơ cấp thì:

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Từ các tiên đề trên có thể xây dựng các định lý cơ bản của xác suất.

7. NGUYÊN LÝ XÁC SUẤT LỚN VÀ XÁC SUẤT NHỎ

Trong nhiều bài toán thực tế ta thường gặp các biến cố có xác suất rất nhỏ, tức là gần bằng 0. Trong trường hợp đó liệu có thể cho rằng những biến cố có xác suất rất nhỏ sẽ không xảy ra khi thực hiện một phép thử? Tất nhiên là không thể kết luận như vậy, vì như trên đã nêu, thậm chí một biến cố có xác suất bằng không vẫn chưa chắc đã là biến cố không thể có, tức là vẫn có thể xảy ra.

Tuy nhiên qua nhiều lần quan sát, người ta thấy rằng các biến cố có xác suất nhỏ gần như sẽ không xảy ra khi tiến hành một phép thử. Trên cơ sở đó có thể đưa ra "*Nguyên lý thực tế không thể có của các biến cố có xác suất nhỏ*" sau đây: *Nếu một biến cố có xác suất rất nhỏ thì thực tế có thể cho rằng trong một phép thử biến cố đó sẽ không xảy ra.*

Hiển nhiên là việc quy định một mức xác suất được coi là rất nhỏ sẽ tuỳ thuộc vào từng bài toán cụ thể. Chẳng hạn, nếu xác suất để dù không mở khi sử dụng bằng 0,01 thì xác suất đó chưa thể coi là nhỏ và chắc chắn là không thể sử dụng loại dù đó. Song nếu xác suất để một chuyến tàu đường dài đến ga chậm bằng 0,01 thì thực tế lại có thể cho rằng tàu sẽ đến ga đúng giờ.

Một xác suất khá nhỏ mà với nó có thể cho rằng biến cố thực tế sẽ không xảy ra được gọi là *mức ý nghĩa*. Tuỳ thuộc vào từng bài toán thực tế, mức ý nghĩa này có thể được lấy trong khoảng từ 0,01 đến 0,05. Nguyên lý xác suất nhỏ sẽ là cơ sở để xây dựng một phương pháp quan trọng của thống kê là lý thuyết kiểm định các giả thuyết thống kê.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tương tự như vậy ta có thể đưa ra "*Nguyên lý thực tế chắc chắn xảy ra của các biến cố có xác suất lớn*" như sau: *Nếu biến cố ngẫu nhiên có xác suất gần bằng 1 thì thực tế có thể cho rằng biến cố đó sẽ xảy ra trong một phép thử.* Hiển nhiên là, cũng như ở trên, việc quy định một mức xác suất đủ coi là lớn tuỳ thuộc vào từng bài toán cụ thể. Nguyên lý xác suất lớn sẽ là cơ sở để xây dựng một phương pháp khác của thống kê là lý thuyết ước lượng.

Bài tập tổng hợp

1.23. Người ta chuyên chở một hòm gồm a chính phẩm và b phế phẩm (a, b ≥ 1) vào kho. Trên đường đi, người ta đánh rơi 1 sản phẩm. Đến kho kiểm tra ngẫu nhiên 1 sản phẩm thì được chính phẩm. Tìm xác suất để sản phẩm đánh rơi là chính phẩm.

ĐS: $P = \dfrac{a-1}{a+b-1}$

1.24. Số lượng nhân viên của công ty A được phân loại theo lứa tuổi và giới tính như sau:

Tuổi \ Giới tính	Nam	Nữ
Dưới 30	120	170
Từ 30- 40	260	420
Trên 40	400	230

Tìm xác suất để lấy ngẫu nhiên một người của công ty đó thì được:

a. Một nhân viên từ 40 tuổi trở xuống.

b. Một nam nhân viên trên 40.

c. Một nữ nhân viên từ 40 tuổi trở xuống.

ĐS: $P_a \approx 0,61$; $P_b = 0,25$; $P_c = 0,37$

1.25. Một cửa hàng đồ điện nhập lô bóng điện đóng thành từng hộp, mỗi hộp 12 chiếc. Chủ cửa hàng kiểm tra chất lượng bằng cách lấy ngẫu nhiên 3 bóng để thử và nếu cả 3 bóng cùng tốt thì hộp bóng điện đó được chấp nhận. Tìm xác suất để một hộp bóng điện được chấp nhận nếu trong hộp đó có 4 bóng bị hỏng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1.26. Giả sử xác suất sinh con trai và con gái là bằng nhau. Một gia đình có 3 con. Tìm xác suất để gia đình đó có:

a. Hai con gái.

b. Ít nhất hai con gái.

c. Hai con gái biết rằng đứa con đầu lòng là gái.

d. Ít nhất hai con gái biết rằng gia đình đó có ít nhất 1 con gái.

ĐS: $P_a = \frac{3}{8}$; $P_b = \frac{4}{8}$; $P_c = \frac{1}{2}$; $P_d = \frac{4}{7}$

1.27. Tìm xác suất để gặp ngẫu nhiên ba người không quen biết nhau ở ngoài đường thì họ:

a. Có ngày sinh nhật khác nhau.

b. Có ngày sinh nhật trùng nhau.

ĐS: $P_a = 0{,}992$

1.28. Một lô sản phẩm gồm 100 chiếc ấm sứ trong đó có 20 chiếc vỡ nắp, 15 chiếc sứt vòi, 10 chiếc mẻ miệng, 7 chiếc vừa vỡ nắp vừa sứt vòi, 5 chiếc vừa vỡ nắp vừa mẻ miệng, 3 chiếc vừa sứt vòi vừa mẻ miệng, 1 chiếc vừa vỡ nắp vừa sứt vòi vừa mẻ miệng. Lấy ngẫu nhiên 1 sản phẩm để kiểm tra. Tìm xác suất:

a. Sản phẩm đó có khuyết tật.

b. Sản phẩm đó chỉ bị sứt vòi.

c. Sản phẩm đó bị sứt vòi biết rằng nó bị vỡ nắp.

1.29. Biết rằng tại xí nghiệp trong ba tháng cuối năm đã có 6 vụ tai nạn lao động. Tìm xác suất để không có ngày nào có quá một vụ tai nạn lao động.

1.30. Có n người trong đó có m người trùng tên xếp hàng một cách ngẫu nhiên. Tìm xác suất để m người trùng tên đứng cạnh nhau nếu:

a. Họ xếp hàng ngang.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b. Họ xếp vòng tròn.

1.31. Ba nữ nhân viên phục vụ A, B, C, thay nhau rửa đĩa chén và giả thiết ba người này đều "khéo léo" như nhau. Trong một tháng có 4 chén bị vỡ. Tìm xác suất:

a. Chị A đánh vỡ 3 chén và chị B đánh vỡ 1 chén.

b. Một trong 3 người đánh vỡ 3 chén.

c. Một trong 3 người đánh vỡ cả 4 chén.

ĐS: $P_a = \frac{4}{81}$; $P_b = \frac{8}{27}$; $P_c = \frac{3}{81}$

1.32. Có 10 khách bước ngẫu nhiên vào một cửa hàng có ba quầy. Tìm xác suất để có 3 người đến quầy số 1.

8. ĐỊNH LÝ CỘNG XÁC SUẤT

Ở các mục trước, chúng ta đã nghiên cứu các phương pháp tính trực tiếp xác suất của các biến cố bằng các định nghĩa xác suất. Song những cách tính trực tiếp này không phải là cơ bản trong lý thuyết xác suất. Việc áp dụng chúng không phải lúc nào cũng tiện lợi và có thể dùng được.

Vì vậy, để xác định xác suất của biến cố, người ta thường không áp dụng các phương pháp tính trực tiếp mà áp dụng phương pháp gián tiếp, cho phép tính xác suất của một biến cố dựa vào xác suất đã biết của các biến cố khác có liên quan với nó thông qua các định lý xác suất, thường được gọi là định lý cộng và định lý nhân xác suất.

Định nghĩa 1. Biến cố C được gọi là tổng của hai biến cố A và B, ký hiệu $C = A + B$ *nếu C chỉ xảy ra khi có ít nhất một trong hai biến cố A và B xảy ra.*

Thí dụ 1. Hai người cùng bắn vào một bia. Gọi A là biến cố "người thứ nhất bắn trúng", B là biến cố "người thứ hai bắn trúng", C là biến cố "bia bị trúng đạn". Rõ ràng là biến cố C sẽ xảy ra khi có ít nhất một trong hai biến cố A và B xảy ra. Vậy $C = A + B$.

Thí dụ 2. Gieo một con xúc xắc. Gọi A là biến cố "xuất hiện mặt 6 chấm", B là biến cố "xuất hiện mặt 5 chấm", C là biến cố "được ít nhất 5 chấm". Biến cố C xảy ra khi hoặc A hoặc B xảy ra. Vậy $C = A + B$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Định nghĩa 2. Biến cố A được gọi là tổng của n biến cố $A_1A_2...A_n$ *nếu A xảy ra khi có ít nhất một trong n biến cố ấy xảy ra.*

Ký hiệu $A = \sum_{i=1}^{n} A_i$.

Gắn liền với khái niệm tổng các biến cố là khái niệm về sự xung khắc của các biến cố.

Định nghĩa 3. Hai biến cố A và B gọi là xung khắc với nhau nếu chúng không thể đồng thời xảy ra trong một phép thử. Trường hợp ngược lại, nếu hai biến cố có thể cùng xảy ra trong một phép thử thì được gọi là không xung khắc.

Thí dụ 3. Trong thí dụ 1, khi thực hiện phép thử là cho hai người cùng bắn vào một bia và gọi A và B tương ứng là các biến cố người thứ nhất và thứ hai bắn trúng bia thì hiển nhiên A và B là không xung khắc. Mặt khác, trong thí dụ 2, khi gieo một con xúc xắc và gọi A và B tương ứng là các biến cố được 6 chấm và 5 chấm thì A và B xung khắc với nhau.

Thí dụ 4. Một bình có 3 loại cầu là cầu trắng, cầu xanh, cầu đỏ. Lấy ngẫu nhiên từ bình đó một quả cầu. Gọi A là biến cố "Lấy được cầu trắng", B là biến cố "Lấy được cầu xanh", A và B là hai biến cố xung khắc với nhau.

Khi áp dụng khái niệm xung khắc cho nhóm gồm n biến cố, ta có khái niệm xung khắc từng đôi.

Định nghĩa 4. Nhóm n biến cố $A_1A_2...A_n$ *được gọi là xung khắc từng đôi nếu bất kỳ hai biến cố nào trong nhóm này cũng xung khắc với nhau.*

Chú ý rằng việc nhận xét tính chất xung khắc hay không xung khắc của các biến cố chủ yếu dựa vào trực giác.

Các khái niệm trên cho phép chúng ta phát biểu định lý cộng xác suất sau đây.

8.1. Định lý

Xác suất của tổng hai biến cố xung khắc bằng tổng xác suất của các biến cố đó.

Như vậy, nếu A và B là hai biến cố xung khắc với nhau thì

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P(A+B) = P(A) + P(B) \tag{1.4}$$

Chứng minh: Ta ký hiệu:

n- Số kết cục đồng khả năng có thể xảy ra khi phép thử được thực hiện.

m_1 - Số kết cục thuận lợi cho biến cố A xảy ra.

m_2 - Số kết cục thuận lợi cho biến cố B xảy ra.

Do A và B xung khắc nhau do đó không thể có các kết cục thuận lợi cho cả A và B cùng đồng thời xảy ra. Vậy số kết cục thuận lợi cho A hoặc cho B xảy ra bằng $m_1 + m_2$. Vì thế:

$$P(A+B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n}$$

Song ta lại có:

$$\frac{m_1}{n} = P(A); \frac{m_2}{n} = P(B)$$

Do đó:

$$P(A+B) = P(A) + P(B)$$

Chú ý rằng (1.4) chỉ là điều kiện cần chứ không phải là điều kiện đủ để A và B xung khắc và mặc dù nó được chứng minh bằng định nghĩa cổ điển nhưng định lý đúng cho mọi trường hợp.

Từ định lý trên có thể suy ra một số hệ quả sau đây.

8.2. Hệ quả

Hệ quả 1. Xác suất của tổng các biến cố xung khắc từng đôi $A_1A_2...A_n$ bằng tổng xác suất của các biến cố đó:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \tag{1.5}$$

Chứng minh: Việc chứng minh được tiến hành theo phương pháp quy nạp toán học. Trong trường hợp có hai biến cố, hệ quả đúng như đã chứng minh trong định lý. Giả sử hệ quả đúng với $n-1$ biến cố, tức là:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P\left(\sum_{i=1}^{n-1} A_i\right) = \sum_{i=1}^{n-1} P(A_i)$$

Cần chứng minh rằng nó cũng đúng với n biến cố.

Thật vậy nếu ký hiệu $\sum_{i=1}^{n-1} A_i = B$ thì:

$$P\left(\sum_{i=1}^{n} A_i\right) = P(B + A_n) = P(B) + P(A_n)$$

Song theo giả thiết ở trên

$$P(B) = P\left(\sum_{i=1}^{n-1} A_i\right) = \sum_{i=1}^{n-1} P(A_i)$$

do đó $$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n-1} P(A_i) + P(A_n) = \sum_{i=1}^{n} P(A_i)$$

Chú ý rằng hệ quả trên có thể mở rộng cho một tổng vô hạn các biến cố.

Thí dụ 5. Xác suất để một xạ thủ bắn bia trúng điểm 10 là 0,1; trúng điểm 9 là 0,2; trúng điểm 8 là 0,25 và ít hơn 8 điểm là 0,45. Xạ thủ ấy bắn một viên đạn. Tìm xác suất để xạ thủ được ít nhất 9 điểm.

Giải. Gọi A_1 là biến cố "Xạ thủ bắn trúng điểm 10", A_2 là biến cố "Xạ thủ bắn trúng điểm 9", A là biến cố "Xạ thủ được ít nhất 9 điểm". Vậy $A = A_1 + A_2$.

Vì A_1 và A_2 xung khắc nhau do đó theo định lý cộng xác suất ta có:

$$P(A) = P(A_1 + A_2) = P(A_1) + P(A_2) = 0,1 + 0,2 = 0,3$$

Trước khi phát biểu hệ quả tiếp theo ta đưa ra các khái niệm về nhóm đầy đủ các biến cố.

Định nghĩa 5. Các biến cố $A_1, A_2, ..., A_n$ được gọi là một nhóm đầy đủ các biến cố nếu trong kết quả của một phép thử sẽ xảy ra một và chỉ một trong các biến cố đó.

Nói cách khác, các biến cố nói trên sẽ tạo nên một nhóm đầy đủ các biến cố nếu chúng xung khắc từng đôi với nhau và tổng của chúng là một

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

biến cố chắc chắn.

Thí dụ 6. Gieo một con xúc xắc, gọi $A_i (i = \overline{1,6})$ là biến cố "Xuất hiện mặt i chấm" thì các biến cố $A_1, A_2, A_3, A_4, A_5, A_6$ tạo nên một nhóm đầy đủ các biến cố.

Đối với nhóm đầy đủ các biến cố ta có hệ quả sau.

Hệ quả 2. Nếu các biến cố $A_1, A_2, ..., A_n$ tạo nên một nhóm đầy đủ các biến cố thì tổng xác suất của chúng bằng 1.

$$\sum_{i=1}^{n} P(A_i) = 1 \qquad (1.6)$$

Chứng minh. Vì các biến cố $A_1, A_2, ..., A_n$ tạo nên một nhóm đầy đủ các biến cố do đó việc xảy ra của ít nhất một trong các biến cố đó là một biến cố chắc chắn. Vì vậy:

$$P\left(\sum_{i=1}^{n} A_i\right) = P(U) = 1$$

Mặt khác, các biến cố này xung khắc từng đôi với nhau, do đó theo hệ quả 1 ta có:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

Từ đó $$\sum_{i=1}^{n} P(A_i) = 1$$

Trường hợp riêng của nhóm đầy đủ các biến cố là các biến cố đối lập.

Định nghĩa 6. Hai biến cố A và $\overline{A}$ gọi là đối lập với nhau nếu chúng tạo nên một nhóm đầy đủ các biến cố.

Thí dụ 7. Bắn một viên đạn vào bia. Gọi A là biến cố "Bắn trúng bia", $\overline{A}$ là biến cố "Bắn trượt bia". A và $\overline{A}$ là hai biến cố đối lập nhau.

Thí dụ 8. Một hòm có a chính phẩm và b phế phẩm. Lấy ngẫu nhiên từ hòm đó n sản phẩm. Gọi A là biến cố "Trong n sản phẩm lấy ra có ít nhất một chính phẩm", $\overline{A}$ là biến cố "Trong n sản phẩm lấy ra không có chính phẩm nào (toàn phế phẩm)" thì A và $\overline{A}$ là các biến cố đối lập nhau.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hệ quả 3. Tổng xác suất của hai biến cố đối lập nhau bằng 1.

$$P(A)+P(\overline{A})=1 \qquad (1.7)$$

Chứng minh. Các biến cố đối lập tạo nên một nhóm đầy đủ các biến cố mà theo hệ quả 2 thì tổng xác suất của các biến cố tạo nên một nhóm đầy đủ các biến cố bằng 1.

Trong thực tế có nhiều trường hợp xác định xác suất của biến cố đối lập $\overline{A}$ đơn giản hơn nhiều so với việc xác định xác suất của biến cố A. Lúc đó người ta thường xác định $P(\overline{A})$ trước, sau đó xác định P(A) theo công thức $P(A)=1-P(\overline{A})$.

Thí dụ 9. Xác suất để sản phẩm sản xuất ra là chính phẩm bằng 0,9. Tìm xác suất để sản phẩm sản xuất ra là phế phẩm.

Giải. Gọi A là biến cố "Sản phẩm là phế phẩm" và $\overline{A}$ là biến cố "Sản phẩm là chính phẩm. A và $\overline{A}$ đối lập nhau, do đó:

$$P(A)=1-P(\overline{A})=1-0,9=0,1.$$

Thí dụ 10. Trong hòm có n sản phẩm, trong đó có m chính phẩm. Lấy ngẫu nhiên k sản phẩm ($k \le n$). Tìm xác suất để trong đó có ít nhất một chính phẩm.

Giải.

Gọi A là biến cố "Trong k sản phẩm lấy ra có ít nhất một chính phẩm" thì biến cố đối lập $\overline{A}$ sẽ là "k sản phẩm lấy ra đều là phế phẩm".

Do đó $$P(A)=1-P(\overline{A})$$

* Khi $k > n - m$ thì $P(A) = 1$

* Xét $k \le n - m$. Ta đi tìm $P(\overline{A})$. Tổng số kết cục duy nhất đồng khả năng là tổ hợp chập k từ n phần tử bằng C_n^k. Số phế phẩm có trong hòm là $n-m$. Do đó số kết cục thuận lợi cho biến cố $\overline{A}$ xảy ra là số tổ hợp chập k (k phế phẩm) từ $n-m$ phế phẩm bằng C_{n-m}^k. Vậy xác suất để cả k sản phẩm lấy ra đều là phế phẩm bằng:

$$P(\overline{A})=\frac{C_{n-m}^k}{C_n^k}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Do đó $$P(A) = 1 - P(\overline{A}) = 1 - \frac{C_{n-m}^{k}}{C_{n}^{k}}$$

Thí dụ 11. Trong hòm có 10 chi tiết, trong đó có 2 chi tiết hỏng. Tìm xác suất để khi lấy ngẫu nhiên ra 6 chi tiết thì có không quá một chi tiết hỏng.

Giải. Gọi A_0 là biến cố "Trong 6 chi tiết lấy ra không có chi tiết nào hỏng". Gọi A_1 là biến cố "Trong 6 chi tiết lấy ra có 1 chi tiết hỏng". Gọi A là biến cố "Trong 6 chi tiết lấy ra có không quá 1 chi tiết hỏng".

Vậy $A = A_0 + A_1$

Vì A_0 và A_1 xung khắc nhau do đó:

$$P(A) = P(A_0 + A_1) = P(A_0) + P(A_1)$$

Dùng định nghĩa cổ điển về xác suất ta tính được:

$$P(A_0) = \frac{C_8^6}{C_{10}^6} = \frac{2}{15}$$

$$P(A_1) = \frac{C_2^1 . C_8^5}{C_{10}^6} = \frac{8}{15}$$

Vậy $$P(A) = \frac{2}{15} + \frac{8}{15} = \frac{2}{3}$$

9. ĐỊNH LÝ NHÂN XÁC SUẤT

Bây giờ ta chuyển sang nghiên cứu trường hợp khi một biến cố có thể xem như tích của các biến cố khác.

Định nghĩa 1. Biến cố C được gọi là tích của hai biến cố A và B nếu C xảy ra khi và chỉ khi cả hai biến cố A và B cùng đồng thời xảy ra. Ký hiệu C=A.B.

Thí dụ 1. Một mạch điện gồm hai bóng đèn mắc song song. Gọi A là biến cố "Bóng thứ nhất bị cháy khi điện quá tải", B là biến cố "Bóng thứ hai bị cháy khi điện quá tải", C là biến cố "Mạch điện bị ngắt khi điện quá tải". Rõ ràng là biến cố C chỉ xảy ra khi cả hai biến cố A và B cùng đồng

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

thời xảy ra. Vậy C = A.B.

Thí dụ 2. Có hai hộp, mỗi hộp đều đựng một số cầu trắng và cầu đen. Lấy ngẫu nhiên từ mỗi hộp một quả cầu. Gọi A là biến cố "Lấy được cầu trắng ở hộp thứ nhất", B là biến cố "Lấy được cầu trắng ở hộp thứ hai", C là biến cố "Lấy được hai quả cầu trắng". Vậy C = A.B.

Định nghĩa 2. Biến cố A được gọi là tích của n biến cố $A_1, A_2, ..., A_n$ *nếu A xảy ra khi và chỉ khi cả n biến cố nói trên cùng đồng thời xảy ra.*

Ký hiệu $$A = \prod_{i=1}^{n} A_i$$

Gắn liền với khái niệm về tích các biến cố là khái niệm về sự độc lập và phụ thuộc vào các biến cố đó.

Định nghĩa 3. Hai biến cố A và B gọi là độc lập với nhau nếu việc xảy ra hay không xảy ra của biến cố này không làm thay đổi xác suất xảy ra của biến cố kia và ngược lại. Trong trường hợp việc biến cố này xảy ra hay không xảy ra làm cho xác suất xảy ra của biến cố kia thay đổi thì hai biến cố đó gọi là phụ thuộc nhau.

Thí dụ 3. Trong bình có 3 cầu trắng và 2 cầu đen. Lấy ngẫu nhiên 1 quả cầu. Gọi A là biến cố "Lấy được cầu trắng". Hiển nhiên là $P(A) = \frac{3}{5}$. Quả cầu được bỏ trở lại bình và tiếp tục lấy 1 quả cầu. Gọi B là biến cố "Lần thứ hai cũng được cầu trắng". Cũng như trước $P(B) = \frac{3}{5}$ và không phụ thuộc gì vào kết quả lấy được của lần trước (biến cố A). Cũng như vậy xác suất lấy được cầu trắng lần thứ nhất (biến cố A) cũng không phụ thuộc gì vào kết quả lấy của lần thứ hai (biến cố B). Vậy hai biến cố A và B độc lập với nhau.

Thí dụ 4. Nếu ở thí dụ trên lần lượt lấy ra hai quả cầu theo phương thức không hoàn lại và gọi A là biến cố "Lần thứ nhất lấy được cầu trắng" thì $P(A) = \frac{3}{5}$. Song biến cố "Lần thứ hai lấy được cầu trắng" (biến cố B) sẽ phụ thuộc vào kết quả lấy của lần thứ nhất. Nếu lần thứ nhất lấy được cầu trắng (biến cố A xảy ra) thì $P(A) = \frac{1}{2}$, còn nếu lần thứ nhất lấy được cầu đen

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

(biến cố A không xảy ra) thì $P(B)=\frac{3}{4}$. Vậy A và B phụ thuộc nhau.

Ta chú ý rằng tính chất độc lập của biến cố có tính tương hỗ theo nghĩa là nếu A và B độc lập với nhau thì A và $\overline{B}$, $\overline{A}$ và B, $\overline{A}$ và $\overline{B}$ cũng độc lập với nhau.

Trong thực tế việc nhận xét tính độc lập hay phụ thuộc của các biến cố chủ yếu dựa vào trực giác.

Việc mở rộng khái niệm độc lập cho nhiều biến cố sẽ dẫn đến hai khái niệm khác nhau là sự độc lập từng đôi và sự độc lập toàn phần.

Định nghĩa 4. Các biến cố $A_1, A_2, ..., A_n$ gọi là độc lập từng đôi với nhau nếu mỗi cặp hai trong n biến cố đó độc lập với nhau.

Chẳng hạn ba biến cố A_1, A_2, A_3 sẽ độc lập từng đôi với nhau nếu A_1 độc lập với A_2, A_1 độc lập với A_3 và A_2 độc lập với A_3.

Thí dụ 5. Gieo một đồng xu 3 lần, gọi $A_i (i=\overline{1,3})$ là biến cố "Được mặt sấp ở lần gieo thứ i". Rõ ràng là mỗi cặp hai trong ba biến cố có độc lập với nhau.

Vậy A_1, A_2, A_3 độc lập từng đôi với nhau.

Định nghĩa 5. Các biến cố $A_1, A_2, ..., A_n$ gọi là độc lập toàn phần với nhau nếu mỗi biến cố độc lập với một tổ hợp bất kỳ của các biến cố còn lại.

Chẳng hạn ba biến cố A_1, A_2, A_3 sẽ độc lập toàn phần với nhau nếu A_1 độc lập với A_2, A_1 độc lập với A_3, A_2 độc lập với A_3, A_1 độc lập với tích A_2A_3, A_2 độc lập với tích A_1A_3, A_3 độc lập với tích A_1A_2.

Chú ý rằng nếu các biến cố độc lập từng đôi với nhau thì từ đó chưa thể suy ra rằng chúng độc lập toàn phần với nhau. Xét theo nghĩa đó thì điều kiện độc lập toàn phần mạnh hơn điều kiện độc lập từng đôi.

Ta sẽ minh họa điều đó qua *thí dụ* sau đây. Giả sử trong bình có 4 quả cầu, 1 quả màu đỏ, 1 quả màu xanh, 1 quả màu vàng, 1 quả sơn cả 3 màu đó. Nếu gọi A là biến cố lấy ngẫu nhiên từ bình đó được cầu có màu đỏ thì $P(A)=\frac{1}{2}$. Tương tự như vậy, xác suất để lấy được quả cầu có màu xanh là

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$P(B)=\frac{1}{2}$, quả cầu có màu vàng là $P(C)=\frac{1}{2}$.

Bây giờ ta giả sử là quả cầu được lấy ra có màu xanh, tức là biến cố B đã xảy ra. Vậy lúc đó xác suất của biến cố A có thay đổi không? Trong hai quả cầu có màu xanh có một quả được sơn cả màu đỏ, do đó vẫn như trước đây $P(A)=\frac{1}{2}$. Vậy A và B độc lập với nhau.

Tương tự ta có thể chứng tỏ rằng A và C, B và C cũng độc lập với nhau, do đó A, B và C độc lập từng đôi với nhau.

Song các biến cố trên có độc lập toàn phần với nhau không? Có thể chứng tỏ rằng không. Thật vậy, giả sử quả cầu lấy ra đã có hai màu xanh và vàng, tức là các biến cố B và C đã xảy ra. Vậy xác suất $P(A)$ để quả cầu đó có màu đỏ là bao nhiêu? Vì trong các quả cầu chỉ có một quả được sơn cả ba màu do đó chắc chắn là quả cầu đó cũng có màu đỏ. Như vậy với điều kiện B và C đã xảy ra thì xác suất $P(A)=1$. Do đó các biến cố A, B, C chỉ độc lập từng đôi chứ không độc lập toàn phần với nhau.

Khi giải nhiều bài toán thực tế nhiều lúc ta phải biểu diễn các biến cố phức hợp dưới dạng kết hợp của các biến cố đơn giản hơn bằng cách sử dụng phép nhân các biến cố.

Chẳng hạn một máy sản xuất ra ba sản phẩm. Ta xét các biến cố đơn giản sau:

A_1 - Sản phẩm thứ nhất là chính phẩm

$\overline{A}_1$ - Sản phẩm thứ nhất là phế phẩm

A_2 - Sản phẩm thứ hai là chính phẩm

$\overline{A}_2$ - Sản phẩm thứ hai là phế phẩm

A_3 - Sản phẩm thứ ba là chính phẩm

$\overline{A}_3$ - Sản phẩm thứ ba là phế phẩm

Lúc đó nếu gọi B là biến cố trong ba sản phẩm sản xuất ra có đúng một chính phẩm thì B có thể biểu diễn qua các biến cố giản đơn như sau:

$$B=A_1\overline{A}_2\overline{A}_3+\overline{A}_1A_2\overline{A}_3+\overline{A}_1\overline{A}_2A_3$$

Gọi C là biến cố trong ba sản phẩm đó có ít nhất hai chính phẩm thì C được biểu diễn như sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$C = A_1A_2\overline{A}_3 + A_1\overline{A}_2A_3 + \overline{A}_1A_2A_3 + A_1A_2A_3$$

Những thủ thuật như vậy để biểu diễn các biến cố phức hợp thông qua các biến cố giản đơn thường được sử dụng trong lý thuyết xác suất.

Với những khái niệm đó ta có thể phát biểu định lý sau đây.

9.1. Định lý 1

Xác suất của tích hai biến cố độc lập bằng tích các xác suất thành phần:

$$P(A.B) = P(A).P(B) \qquad (1.8)$$

Chứng minh. Giả sử n_1 và n_2 tương ứng là số kết cục đồng khả năng cho biến cố A và B xảy ra; m_1 và m_2 tương ứng là số kết cục thuận lợi cho A và B xảy ra. Do A và B độc lập nên số kết cục đồng khả năng cho tích AB xảy ra sẽ là $n_1.n_2$ và số kết cục thuận lợi cho tích xảy ra là m_1m_2. Vì vậy:

$$P(A.B) = \frac{m_1.m_2}{n_1.n_2} = \frac{m_1}{n_1}.\frac{m_2}{n_2} = P(A).P(B)$$

Chú ý rằng (1.8) vừa là điều kiện cần, vừa là điều kiện đủ để A và B độc lập.

Từ định lý trên có thể suy ra một số hệ quả sau:

Hệ quả 1. Nếu A và B độc lập thì:

$$P(A) = \frac{P(A.B)}{P(B)} \text{ và } P(B) = \frac{P(A.B)}{P(A)} \qquad (1.9)$$

Khi $P(B) > 0$ và $P(A) > 0$.

Hệ quả 2. Xác suất của tích n biến cố độc lập toàn phần bằng tích các xác suất thành phần.

$$P\left(\prod_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i) \qquad (1.10)$$

Bạn đọc có thể tự chứng minh hệ quả trên bằng phương pháp quy nạp toán học.

Thí dụ 6. Có hai hộp đựng chi tiết. Hộp thứ nhất đựng 10 cái ốc, trong đó có 6 cái tốt. Hộp thứ hai đựng 15 cái vít, trong đó có 9 cái tốt. Lấy ngẫu

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

nhiên từ mỗi hộp một chi tiết. Tìm xác suất để lấy được một bộ ốc vít tốt.

Giải. Gọi A_1 là biến cố "Lấy được cái ốc tốt ở hộp thứ nhất", A_2 là biến cố "Lấy được cái vít tốt ở hộp thứ hai". Gọi A là biến cố "Lấy được một bộ ốc vít tốt".

Vậy: $A = A_1.A_2$

Vì các biến cố A_1 và A_2 độc lập với nhau, do đó:

$$P(A) = P(A_1.A_2) = P(A_1).P(A_{2)}) = \frac{6}{10}.\frac{9}{15} = \frac{9}{25}$$

Để xem xét trường hợp các biến cố phụ thuộc trước hết ta xét khái niệm xác suất có điều kiện.

Định nghĩa 6. Xác suất của biến cố A được tính với điều kiện biến cố B đã xảy ra gọi là xác suất có điều kiện của A và ký hiệu là P(A/B).

Thí dụ 7. Trong bình có 5 cầu trắng và 3 cầu đen. Lấy ngẫu nhiên lần lượt hai quả cầu. Tìm xác suất để lần thứ hai lấy được cầu trắng (biến cố B) nếu biết rằng lần thứ nhất đã lấy được cầu trắng (biến cố A).

Giải. Sau khi lấy lần thứ nhất (biến cố A đã xảy ra) thì trong bình chỉ còn lại 7 quả cầu, trong đó có 4 quả cầu trắng.

Do đó xác suất có điều kiện của B bằng:

$$P(B/A) = \frac{4}{7}$$

Bây giờ ta có thể phát biểu định lý nhân xác suất đối với một tích các biến cố phụ thuộc.

9.2. Định lý 2

Xác suất của tích hai biến cố phụ thuộc A và B bằng tích xác suất của một trong hai biến cố đó với xác suất có điều kiện của biến cố còn lại.

$$P(A.B) = P(A).P(B/A) = P(B).P(A/B) \qquad (1.11)$$

Chứng minh. Giả sử n là số kết cục đồng khả năng có thể xảy ra trong phép thử, m_1 là số kết cục thuận lợi cho biến cố A xảy ra, m_2 là số kết cục thuận lợi cho biến cố B xảy ra. Vì ta không giả thiết A và B xung khắc do đó nói chung sẽ có k kết cục thuận lợi cho cả A và B cùng đồng thời xảy ra. Lúc đó:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P(A.B) = \frac{k}{n};\ P(A) = \frac{m_1}{n}$$

Ta đi tìm P(B/A). Với điều kiện biến cố A đã xảy ra thì số kết cục duy nhất đồng khả năng cho phép thử đối với biến cố B là m_1 trong đó có k kết cục thuận lợi cho B xảy ra. Do đó:

$$P(B/A) = \frac{k}{m_1}$$

Như vậy: $P(A.B) = \frac{k}{n} = \frac{m_1}{n}.\frac{k}{m_1} = P(A).P(B/A)$

Hiển nhiên là khi áp dụng định lý nhân xác suất thì không nhất thiết phải phân biệt xem biến cố nào trong hai biến cố A và B là thứ nhất và thứ hai. Do đó có thể chứng minh được dạng thứ hai là

$$P(A.B) = P(B).P(A/B).$$

Từ định lý trên ta có thể suy ra các hệ quả sau đây:

Hệ quả 1. Nếu P(B)>0 thì xác suất của biến cố A với điều kiện biến cố B đã xảy ra được tính bằng công thức:

$$P(A/B) = \frac{P(AB)}{P(B)} \qquad (1.12)$$

Còn nếu P(B) = 0 thì xác suất trên không xác định.

Tương tự nếu P(A)>0 thì ta có:

$$P(B/A) = \frac{P(AB)}{P(A)} \qquad (1.13)$$

Các công thức trên được suy ra trực tiếp từ định lý vừa xét.

Hệ quả 2. Xác suất của tích n biến cố phụ thuộc bằng tích xác suất của n biến cố, trong đó xác suất của mỗi biến cố tiếp sau đều được tính với điều kiện tất cả các biến cố xét trước đó đã xảy ra.

$$P(A_1A_2...A_n) = P(A_1)P(A_2/A_1)...P(A_n/A_1...A_{n-1}) \qquad (1.14)$$

Bạn đọc có thể tự chứng minh hệ quả này bằng phương pháp quy nạp toán học.

Hệ quả 3. Nếu A và B độc lập thì:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P(A/B) = P(A) \text{ và } P(B/A) = P(B) \qquad (1.15)$$

Thật vậy, từ định lý vừa xét ta có:

$$P(A/B) = \frac{P(AB)}{P(B)} = \frac{P(A).P(B)}{P(B)} = P(A)$$

$$P(B/A) = \frac{P(AB)}{P(A)} = \frac{P(A).P(B)}{P(A)} = P(B)$$

Thí dụ 8. Trong hòm có 7 chính phẩm và 3 phế phẩm. Lấy ngẫu nhiên lần lượt hai sản phẩm. Tìm xác suất để cả hai sản phẩm lấy ra đều là chính phẩm.

Giải. Gọi A_1 là biến cố "Sản phẩm thứ nhất lấy ra là chính phẩm", A_2 là biến cố "Sản phẩm thứ hai lấy ra là chính phẩm", A là biến cố "Cả hai sản phẩm lấy ra đều là chính phẩm"

Vậy: $A = A_1.A_2$

Theo định lý nhân xác suất đối với hai biến cố phụ thuộc ta có:

$$P(A) = P(A_1.A_2) = P(A_1).P(A_2/A_1) = \frac{7}{10}.\frac{6}{9} = \frac{7}{15}$$

Trong thực tế ta thường gặp các bài toán trong đó phải áp dụng cùng một lúc cả định lý cộng và định lý nhân xác suất, tức là các biến cố mà ta phải xác định xác suất được biểu diễn dưới dạng biến cố đối lập hoặc dưới dạng tổng của một số biến cố xung khắc từng đôi với nhau và mỗi biến cố như vậy lại là tích của một số biến cố khác.

Thí dụ 9. Một thiết bị gồm 3 bộ phận. Trong khoảng thời gian T, việc các bộ phận đó bị hỏng là độc lập với nhau và với các xác suất tương ứng là 0,1; 0,2; 0,3. Cả thiết bị sẽ bị hỏng nếu có ít nhất một bộ phận hỏng. Tìm xác suất hoạt động tốt của thiết bị đó.

Giải. Gọi $A_i (i = \overline{1,3})$ là biến cố "Bộ phận thứ i của thiết bị hoạt động tốt trong khoảng thời gian T". Gọi A là biến cố "Thiết bị hoạt động tốt trong khoảng thời gian T".

Vậy $A = A_1.A_2.A_3$

Vì A_1, A_2, A_3 độc lập toàn phần với nhau do đó:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P(A) = P(A_1)P(A_2)P(A_3)$$

Các biến cố "Bộ phận thứ nhất hoạt động tốt" và "Bộ phận thứ nhất bị hỏng" là đối lập với nhau, do đó:

$$P(A_1) = 1 - 0,1 = 0,9$$

Tương tự

$$P(A_2) = 1 - 0,2 = 0,8$$

$$P(A_3) = 1 - 0,3 = 0,7$$

$$P(A) = 0,9.0,8.0,7 = 0,504$$

Thí dụ 10. Một xí nghiệp có 3 ô tô hoạt động độc lập. Xác suất để trong một ngày các ô tô bị hỏng tương ứng là 0,1; 0,2; 0,15. Tìm xác suất để trong một ngày có đúng 1 ô tô bị hỏng.

Giải. Gọi A_i là biến cố "Ô tô thứ i bị hỏng trong ngày" $(i = \overline{1,3})$; A là biến cố "Trong ngày có đúng 1 ô tô bị hỏng".

Vậy $$A = A_1\overline{A}_2\overline{A}_3 + \overline{A}_1A_2\overline{A}_3 + \overline{A}_1\overline{A}_2A_3$$

Vì các nhóm biến cố $A_1\overline{A}_2\overline{A}_3, \overline{A}_1A_2\overline{A}_3$ và $\overline{A}_1\overline{A}_2A_3$ là xung khắc từng đôi và trong mỗi nhóm các biến cố lại độc lập toàn phần với nhau, do đó:

$$P(A) = P(A_1)P(\overline{A}_2)P(\overline{A}_3) + P(\overline{A}_1)P(A_2)P(\overline{A}_3) +$$

$$+P(\overline{A}_1)P(\overline{A}_2)P(A_3)$$

Vì $P(A_1) = 0,1; P(A_2) = 0,2; P(A_3) = 0,15$

Do đó: $P(\overline{A}_1) = 0,9; P(\overline{A}_2) = 0,8; P(\overline{A}_3) = 0,85$

Và ta có:

$$P(A) = 0,1.0,8.0,85 + 0,9.0,2.0,85 + 0,9.0,8.0,15 = 0,329$$

Thí dụ 11. Với các điều kiện của thí dụ trước, tìm xác suất để trong một ngày có ít nhất một ô tô bị hỏng.

Giải. Gọi B là biến cố "Trong 1 ngày có ít nhất 1 ô tô bị hỏng". Áp dụng thủ thuật như trên có thể biểu diễn biến cố B như sau:

$$B = A_1\overline{A}_2\overline{A}_3 + \overline{A}_1A_2\overline{A}_3 + \overline{A}_1\overline{A}_2A_3 +$$

$$+A_1A_2\overline{A}_3 + A_1\overline{A}_2A_3 + \overline{A}_1A_2A_3 + A_1A_2A_3$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Xác suất của biến cố B có thể tìm bằng cách áp dụng định lý cộng và định lý nhân như đã làm ở thí dụ trước. Tuy nhiên cách giải như vậy rất dài dòng và phức tạp. Ở đây nên chuyển từ biến cố B sang biến cố đối lập với nó. Hiển nhiên $\overline{B}$ sẽ là biến cố cả 3 ô tô cùng hoạt động tốt:

$$\overline{B} = \overline{A}_1 \overline{A}_2 \overline{A}_3$$

Theo định lý nhân xác suất:

$$P(\overline{B}) = P(\overline{A}_1).P(\overline{A}_2).P(\overline{A}_3) = 0,9.0,8.0,85 = 0,612$$

Do đó: $P(B) = 1 - P(\overline{B}) = 1 - 0,612 = 0,388$

Thí dụ 12. Gieo một đồng xu 6 lần. Tìm xác suất để số lần được mặt sấp nhiều hơn số lần được mặt ngửa.

Giải. Để tìm xác suất của biến cố được số mặt sấp nhiều hơn số mặt ngửa (biến cố A) ta có thể phân nó ra thành các trường hợp sau đây:

A_1 - Được 6 mặt sấp, không có lần nào được mặt ngửa;

A_2 - Được 5 mặt sấp, 1 mặt ngửa...

Song bài toán này có thể giải bằng cách đơn giản hơn. Ta liệt kê các trường hợp có thể xảy ra như sau:

A - Số mặt sấp nhiều hơn số mặt ngửa

B - Số mặt ngửa nhiều hơn số mặt sấp

C - Số mặt sấp và ngửa bằng nhau.

Các biến cố A, B và C tạo nên một nhóm đầy đủ các biến cố. Vì vậy:

$$P(A) + P(B) + P(C) = 1$$

Vì mặt sấp và mặt ngửa đối xứng nhau, do đó P(A)=P(B)

Từ đó:

$$2P(A) + P(C) = 1$$

$$P(A) = \frac{1 - P(C)}{2}$$

Ta đi tìm P(C) là xác suất để trong 6 lần tung có 3 lần được sấp, 3 lần được ngửa. Xác suất của mỗi phương thức xảy ra biến cố C (chẳng hạn theo trình

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

tự SSSNNN) cùng bằng $\left(\frac{1}{2}\right)^6$. Số những phương thức như vậy là $C_6^3 = 20$.

Vì vậy:

$$P(C) = \frac{20}{64} = \frac{5}{16}$$

Từ đó:

$$P(A) = \frac{1 - \frac{5}{16}}{2} = \frac{11}{32}$$

Bây giờ ta chuyển sang nghiên cứu trường hợp xác suất của tổng hai biến cố không xung khắc.

9.3. Định lý 3

Xác suất của tổng hai biến cố không xung khắc bằng tổng xác suất các biến cố đó trừ đi xác suất của tích các biến cố đó.

$$P(A+B) = P(A) + P(B) - P(AB) \qquad (1.16)$$

Chứng minh. Giả sử n là số kết cục đồng khả năng có thể xảy ra trong phép thử, m_1 là số kết cục thuận lợi cho A xảy ra, m_2 là số kết cục thuận lợi cho B xảy ra. Vì A và B không xung khắc nhau, do đó sẽ có k kết cục nào đó thuận lợi cho tích AB xảy ra. Lúc đó số kết cục thuận lợi cho ít nhất một trong hai biến cố A và B xảy ra bằng $m_1 + m_2 - k$. Như vậy:

$$P(A+B) = \frac{m_1 + m_2 - k}{n} = \frac{m_1}{n} + \frac{m_2}{n} - \frac{k}{n}$$

$$= P(A) + P(B) - P(AB).$$

Nếu biến cố A và B độc lập thì công thức có dạng:

$$P(A+B) = P(A) + P(B) - P(A).P(B)$$

Còn đối với các biến cố phụ thuộc thì công thức có dạng:

$$P(A+B) = P(A) + P(B) - P(A)P(B/A)$$

Nếu A và B xung khắc nhau thì lúc đó biến cố tích AB sẽ là biến cố không thể có, do đó P(AB)=0. Ta lại thu được công thức cộng xác suất đã xét ở phần trước.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Khi mở rộng định lý trên cho trường hợp tổng của n biến cố không xung khắc, ta thu được các hệ quả sau đây.

9.4. Hệ quả

Hệ quả 1. Xác suất của tổng n biến cố không xung khắc được xác định bằng công thức:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) - \ldots + (-1)^{n-1} P(A_1 A_2, \ldots A_n) \quad (1.17)$$

Hệ quả trên có thể chứng minh bằng phương pháp quy nạp toán học. Cũng bằng phương pháp quy nạp toán học có thể chứng minh được hệ quả sau đây, cho phép biểu diễn xác suất của tích n biến cố thông qua xác suất của tổng các biến cố đó.

Hệ quả 2. Xác suất của tích n biến cố được xác định bằng công thức:

$$P\left(\prod_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i + A_j) + \sum_{i<j<k} P(A_i + A_j + A_k) - \ldots + (-1)^{n-1} P(A_1 + A_2 + \ldots + A_n) \quad (1.18)$$

Thực tế các hệ quả đã xét ở trên được sử dụng để biến đổi các biểu thức chứa xác suất của tổng và tích các biến cố. Tuỳ thuộc vào đặc điểm của từng loại bài toán, trong một số trường hợp chỉ nên dùng tổng các biến cố, trong một vài trường hợp khác lại chỉ nên dùng tích các biến cố. Các công thức trên được sử dụng để thực hiện các phép biến đổi qua lại đó.

Tuy nhiên trong thực tế giải các bài toán của lý thuyết xác suất việc vận dụng các hệ quả trên là khá phức tạp. Trong một số trường hợp có thể chuyển qua biến cố đối lập để việc giải quyết trở nên đơn giản hơn. Ta xét một trường hợp thường gặp trong thực tế sau đây.

9.5. Định lý 4

Xác suất của tổng n biến cố không xung khắc và độc lập toàn phần với nhau bằng một trừ đi tích xác suất của các biến cố đối lập với các biến cố đó.

$$P\left(\sum_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n} P(\overline{A}_i) \quad (1.19)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chứng minh. Nếu ký hiệu $A = \sum_{i=1}^{n} A_i$ thì lúc đó $\overline{A} = \prod_{i=1}^{n} \overline{A}_i$, do đó:

$$P\left(\sum_{i=1}^{n} A_i\right) = P(A) = 1 - P(\overline{A}) = 1 - P\left(\prod_{i=1}^{n} \overline{A}_i\right)$$

Theo giả thiết các biến cố A_i độc lập toàn phần, do đó các biến cố $\overline{A}_i$ cũng độc lập toàn phần với nhau.

Vì thế:

$$P\left(\prod_{i=1}^{n} \overline{A}_i\right) = \prod_{i=1}^{n} P(\overline{A}_i)$$

Từ đó: $$P\left(\sum_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n} P(\overline{A}_i)$$

Trong trường hợp riêng nếu $P(A_1) = P(A_2) = ... = P(A_n) = p$ thì công thức trên có dạng sau đây:

$$P\left(\sum_{i=1}^{n} A_i\right) = 1 - (1-p)^n \qquad (1.20)$$

Thí dụ 13. Hai máy bay ném bom một mục tiêu, mỗi máy bay ném 1 quả với xác suất trúng mục tiêu tương ứng là 0,7 và 0,8. Tìm xác suất để mục tiêu bị trúng bom.

Giải. Gọi A_1 và A_2 tương ứng là biến cố quả bom thứ nhất và thứ hai trúng mục tiêu. Gọi A là biến cố mục tiêu bị trúng bom. Vậy $A = A_1 + A_2$.

Vì A_1 và A_2 không xung khắc nhau, do đó:

$$P(A) = P(A_1) + P(A_2) - P(A_1A_2)$$

A_1 và A_2 lại độc lập với nhau, do đó:

$$P(A) = P(A_1) + P(A_2) - P(A_1)P(A_2) =$$

$$= 0,7 + 0,8 - 0,7.0,8 = 0,94$$

Bài toán trên cũng có thể giải bằng cách sử dụng định lý đã xét ở trên. Do A_1 và A_2 là không xung khắc và độc lập nên:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P(A) = 1 - P(\overline{A}_1)P(\overline{A}_2) = 1 - 0,3.0,2 = 0,94$$

Ta cũng thu được kết quả tương tự.

Thí dụ 14. Phải gieo một con xúc xắc tối thiểu bao nhiêu lần để với xác suất không nhỏ hơn 0,5 có thể hy vọng rằng trong đó có ít nhất một lần được mặt 6 chấm.

Giải. Giả sử ta gieo con xúc xắc n lần.

Gọi A_i là biến cố "Gieo lần thứ i được mặt 6 chấm", $i = \overline{1,n}$. Gọi A là biến cố "Trong n lần gieo có ít nhất một lần được mặt 6 chấm". Vậy $A = \sum_{i=1}^{n} A_i$.

Các biến cố A_i là không xung khắc và độc lập toàn phần với nhau, do đó:

$$P(A) = 1 - \prod_{i=1}^{n} P(\overline{A}_i)$$

Vì $P(A_1) = P(A_2) = ... = P(A_n) = \frac{1}{6}$ (xác suất để mỗi lần gieo được mặt 6 chấm) do đó:

$$P(\overline{A}_1) = P(\overline{A}_2) = ... = P(\overline{A}_n) = \frac{5}{6}$$

Nên ta có:

$$P(A) = 1 - \left(\frac{5}{6}\right)^n$$

Theo giả thiết xác suất của biến cố A không nhỏ hơn 0,5 do đó ta thu được bất phương trình sau:

$$1 - \left(\frac{5}{6}\right)^n \geq 0,5$$

Từ đó:

$$\left(\frac{5}{6}\right)^n \leq 0,5$$

Lấy Lôgarit của cả hai vế ta có:

$$n \lg\left(\frac{5}{6}\right) \leq \lg 0,5$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chú ý rằng $\lg\left(\frac{5}{6}\right) \le 0$, ta có:

$$n \ge \frac{\lg 0,5}{\lg\left(\frac{5}{6}\right)} = 3,8$$

Như vậy $n \ge 4$, tức là phải gieo ít nhất 4 lần.

Thí dụ 15. Xác suất để động cơ thứ nhất của máy bay bị trúng đạn là 0,2; để động cơ thứ hai của máy bay bị trúng đạn là 0,3 còn xác suất bị trúng đạn của phi công là 0,1. Tìm xác suất để máy bay rơi, biết rằng máy bay rơi khi hoặc cả hai động cơ bị trúng đạn, hoặc phi công bị trúng đạn.

Giải. Gọi A_1 và A_2 tương ứng là biến cố "Động cơ thứ nhất và thứ hai của máy bay bị trúng đạn", Gọi A_3 là biến cố "Phi công bị trúng đạn". Gọi A là biến cố "Máy bay rơi". Lúc đó:

$$A = A_1A_2 + A_3$$

Vì các biến cố A_1A_2 và A_3 không xung khắc với nhau, do đó:

$$P(A) = P(A_1A_2) + P(A_3) - P(A_1A_2A_3)$$

Mặt khác các biến cố $A_1A_2A_3$ độc lập toàn phần với nhau, do đó:

$$P(A) = P(A_1)P(A_2) + P(A_3) - P(A_1)P(A_2)P(A_3) =$$

$$= 0,2.0,3 + 0,1 - 0,2.0,3.0,1 = 0,154$$

Thí dụ 16. Một người viết n lá thư, bỏ ngẫu nhiên vào n phong bì có đề sẵn địa chỉ. Tìm xác suất để có ít nhất một lá thư bỏ đúng địa chỉ.

Giải. Gọi A_i là biến cố "Lá thư thứ i bỏ đúng địa chỉ", $i = \overline{1,n}$, A là biến cố "Trong n lá thư đó có ít nhất một lá thư bỏ đúng địa chỉ", $A = \sum_{i=1}^{n} A_i$.

Vì các biến cố A_i không xung khắc với nhau, do đó:

$$P(A) = \sum_i P(A_i) - \sum_{i<j} P(A_iA_j) + \sum_{i<j<k} P(A_iA_jA_k) - ... +$$

$$+ ... + (-1)^{n-1} P(A_1A_2...A_n)$$

Ta đi tìm xác suất thành phần.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì trong n lá thư chỉ có một lá thư của phong bì thứ i, do đó xác suất để lá thư thứ i bỏ đúng địa chỉ bằng:

$$P(A_i) = \frac{1}{n} \quad (\forall i)$$

Tất cả có n lá thư, do đó:

$$\sum_i P(A_i) = n.\frac{1}{n} = 1$$

Ta đi tìm $P(A_iA_j) \forall ij$. Vì các biến cố A_i và A_j phụ thuộc nhau do đó:

$$P(A_iA_j) = P(A_i)P(A_j / A_i)$$

Như ở trên đã xác định $P(A_i) = \frac{1}{n}, \forall i$. Với điều kiện biến cố A_i đã xảy ra thì xác suất bỏ đúng địa chỉ của lá thư j sẽ bằng $\frac{1}{n-1}$, do đó:

$$P(A_iA_j) = \frac{1}{n}.\frac{1}{n-1}$$

Số phương thức để thành lập những nhóm hai biến cố như vậy từ n biến cố là C_n^2, do đó:

$$\sum_{i=1} P(A_iA_j) = C_n^2 \frac{1}{n}.\frac{1}{n-1} = \frac{1}{2!}$$

Tương tự có thể tìm được:

$$\sum_{i<j<k} P(A_iA_jA_k) = \frac{1}{3!}$$

Và $$P(A_1A_2...A_n) = \frac{1}{n!}$$

Do đó:

$$P(A) = 1 - \frac{1}{2!} + \frac{1}{3!} - ... + (-1)^{n-1}\frac{1}{n!} = \sum_{x=1}^{n} (-1)^{x-1}\frac{1}{x!}$$

Với n khá lớn $P(A) \approx 1 - e^{-1}$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

1.33. Một chi tiết được lấy ngẫu nhiên có thể là chi tiết loại 1 (ký hiệu là A) hoặc chi tiết loại 2 (ký hiệu là B) hoặc chi tiết loại 3 (ký hiệu là C). Hãy mô tả các biến cố sau đây:

a. $A + B$

b. $AB + C$

c. $\overline{A+B}$

d. AC.

1.34. Ba người cùng bắn vào một mục tiêu. Gọi A_k là biến cố người thứ k bắn trúng mục tiêu ($k = \overline{1,3}$). Hãy viết bằng ký hiệu các biến cố biểu thị rằng:

a. Chỉ có người thứ nhất bắn trúng mục tiêu.

b. Chỉ có một người bắn trúng mục tiêu.

c Chỉ có hai người bắn trúng mục tiêu.

d. Có người bắn trúng mục tiêu.

1.35. Ta kiểm tra theo thứ tự một lô hàng có 10 sản phẩm. Các sản phẩm đều thuộc một trong hai loại: Tốt hoặc xấu. Ký hiệu $A_k = (k = \overline{1,10})$ là biến cố chỉ sản phẩm kiểm tra thứ k thuộc loại xấu. Viết bằng ký hiệu các biến cố sau:

a. Cả 10 sản phẩm đều xấu.

b. Có ít nhất một sản phẩm xấu.

c. Có 6 sản phẩm đầu kiểm tra là tốt, còn các sản phẩm còn lại là xấu.

d. Các sản phẩm kiểm tra theo thứ tự chẵn là tốt, còn các sản phẩm kiểm tra theo thứ tự lẻ là xấu.

1.36. Gọi A là biến cố sinh con gái và B là biến cố sinh con có trọng lượng hơn 3 kg. Hãy mô tả tổng và tích của hai biến cố trên.

1.37. Một công ty tham gia đấu thầu 2 dự án. Gọi A và B tương ứng là biến cố công ty thắng thầu dự án thứ nhất và thứ hai. Hãy mô tả tổng và tích của A và B.

1.38. Cơ cấu chất lượng sản phẩm của một nhà máy như sau:

Sản phẩm loại 1: 40%, sản phẩm loại 2: 50%, còn lại là phế phẩm.

Lấy ngẫu nhiên một sản phẩm của nhà máy. Tính xác suất sản phẩm lấy ra thuộc loại 1 hoặc loại 2.

ĐS: $P = 0,9$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1.39. Để nhập kho, sản phẩm của nhà máy phải qua 3 phòng kiểm tra chất lượng, xác suất phát hiện ra phế phẩm ở các phòng theo thứ tự là 0,8; 0,9; và 0,99. Tính xác suất phế phẩm được nhập kho (các phòng kiểm tra hoạt động độc lập).

ĐS: P = 0,0002.

1.40. Xác suất để khi đo một đại lượng vật lý phạm sai số vượt quá tiêu chuẩn cho phép là 0,4. Thực hiện 3 lần đo độc lập. Tìm xác suất sao cho có đúng một lần đo sai số vượt quá tiêu chuẩn cho phép.

1.41. Một hộp chứa 3 bi trắng, 7 bi đỏ và 15 bi xanh. Một hộp khác chứa 10 bi trắng, 6 bi đỏ và 9 bi xanh. Lấy ngẫu nhiên từ mỗi hộp một bi. Tìm xác suất để hai bi lấy ra có cùng màu.

ĐS: $P = \frac{207}{625}$

1.42. Hai người cùng bắn vào một mục tiêu. Khả năng bắn trúng của từng người là 0,8 và 0,9. Tìm xác suất:

a. Chỉ có một người bắn trúng.

b. Có người bắn trúng mục tiêu.

c. Cả hai người bắn trượt.

ĐS: $P_a = 0,26$; $P_b = 0,98$; $P_c = 0,02$

1.43. Chi tiết được gia công qua k công đoạn nối tiếp nhau và chất lượng chi tiết chỉ được kiểm tra sau khi đã được gia công xong. Xác suất gây ra khuyết tật cho chi tiết ở công đoạn thứ i là P_i (i =1,k). Tìm xác suất để sau khi gia công xong chi tiết có khuyết tật.

1.44. Trong hộp có n quả bóng bàn mới. Người ta lấy ra k quả để chơi (k ≤ n/2) sau đó lại bỏ vào hộp. Tìm xác suất để lần sau lấy ra k quả để chơi thì lấy được toàn bóng mới.

ĐS: $P = \frac{[(n-k)!]^2}{n!(n-2k)!}$

1.45. Tín hiệu thông tin được phát 3 lần với xác suất thu được của mỗi lần là 0,4.

a. Tìm xác suất để nguồn thu nhận được thông tin đó.

b. Nếu muốn xác suất thu được thông tin lên đến 0,9 thì phải phát ít nhất bao nhiêu lần.

ĐS: a. $P_a = 0,784$; b. Ít nhất 5 lần

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1.46. Hai người cùng bắn vào một mục tiêu, khả năng chỉ có một người bắn trúng là 0,38. Tìm xác suất bắn trúng của người thứ nhất, biết rằng khả năng bắn trúng của người thứ hai là 0,8.

1.47. Trong 10 sản phẩm có 2 phế phẩm. Lấy ngẫu nhiên 2 sản phẩm. Tính xác suất để cả 2 sản phẩm đều là phế phẩm trong trường hợp.

a. Lấy hoàn lại.

b. Lấy không hoàn lại.

ĐS: $P_a = 0,04$; $P_b = 0,022$.

1.48. Một nồi hơi được lắp van bảo hiểm với xác suất hỏng của các van tương ứng là 0,1 và 0,2. Nồi hơi sẽ hoạt động an toàn khi có van không hỏng. Tìm xác suất để nồi hơi hoạt động.

a. An toàn

b. Mất an toàn.

ĐS: $P_a = 0,98$; $P_b = 0,02$.

1.49. Bắn liên tiếp vào một mục tiêu cho đến khi viên đạn đầu tiên trúng mục tiêu thì dừng. Tìm xác suất sao cho phải bắn đến viên thứ 6, biết rằng xác suất trúng mục tiêu của mỗi viên đạn là 0,2 và các lần bắn độc lập nhau.

ĐS: $P = 0,065536$.

1.50. Một thủ kho có chùm chìa khoá gồm 9 chiếc trong đó chỉ có một chiếc mở cửa kho. Anh ta thử ngẫu nhiên từng chìa khoá một, chiếc nào đã được thử thì không thử lại. Tính xác suất anh ta mở được đã ở lần thử thứ 4.

ĐS: $P \approx 0,11$.

1.51. Công ty sử dụng hai hình thức quảng cáo là đài phát thanh và vô tuyến truyền hình. Giả sử 25% khách hàng nắm được thông tin này qua vô tuyến truyền hình, 34% khách hàng nắm được thông tin qua đài phát thanh và 10% khách hàng nắm được thông tin này qua cả hai hình thức quảng cáo. Tìm xác suất để chọn ngẫu nhiên 1 khách hàng thì người đó nắm được thông tin về sản phẩm của công ty.

ĐS: 0,49.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1.52. Gieo hai con xúc xắc đối xứng và đồng chất. Gọi A là biến cố xuất hiện khi tổng số chấm thu được là lẻ, B là biến cố được ít nhất một mặt một chấm. Tính P (AB), P(A+B); P($\overline{AB}$).

ĐS: $P(AB) = \frac{1}{6}$; $P(A+B) = \frac{23}{36}$; $P(\overline{AB}) = \frac{5}{6}$

1.53. Có hai bóng đèn điện với xác suất hỏng tương ứng là 0,1 và 0,2 và việc chúng hỏng là độc lập với nhau. Tính xác suất để mạch không có điện do bóng hỏng nếu chúng mắc:

a. Nối tiếp.

b. Song song

1.54. Có hai lô hàng.

Lô 1: Có 90 chính phẩm và 10 phế phẩm

Lô 2: Có 80 chính phẩm và 20 phế phẩm.

Lấy ngẫu nhiên từ mỗi lô hàng một sản phẩm.

Tính xác suất để:

a. Lấy được một chính phẩm.

b. Lấy được ít nhất một chính phẩm.

ĐS: $P_a = 0,26$; $P_b = 0,98$.

1.55. Một lô hàng có 9 sản phẩm. Mỗi lần kiểm tra chất lượng lấy ngẫu nhiên 3 sản phẩm. Sau khi kiểm tra xong lại trả vào lô hàng. Tính xác suất để sau 3 lần kiểm tra lô hàng, tất cả các sản phẩm đều được kiểm tra.

ĐS: p = 0,0028.

1.56. Xác suất để bắn một viên đạn trúng đích là 0,8. Hỏi phải bắn bao nhiêu viên đạn để với xác suất nhỏ hơn 0,4 có thể hy vọng rằng không có viên nào trượt.

ĐS: n = 5 viên.

1.57. Phải gieo cùng một lúc hai con xúc xắc ít nhất trong bao nhiêu lần để với xác suất lớn hơn 0,5 có thể hy vọng rằng có ít nhất 1 lần được hai mặt 6 chấm.

ĐS: Gieo ít nhất 25 lần.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

10. CÁC HỆ QUẢ CỦA ĐỊNH LÝ CỘNG VÀ ĐỊNH LÝ NHÂN XÁC SUẤT

10.1. Công thức Bernoulli

Trong nhiều bài toán thực tế ta thường gặp trường hợp cùng một phép thử được lặp lại nhiều lần. Trong kết quả của mỗi phép thử có thể xảy ra hoặc không thể xảy ra một biến cố A nào đó và ta không quan tâm đến kết quả của từng phép thử mà quan tâm đến tổng số lần xảy ra của biến cố A trong cả dãy phép thử đó. Chẳng hạn nếu tiến hành sản xuất hàng loạt một loại chi tiết nào đó thì ta thường quan tâm đến tổng số chi tiết đạt tiêu chuẩn của cả quá trình sản xuất. Trong những bài toán như vậy cần phải biết cách xác định xác suất để biến cố A xảy ra một số lần nhất định trong kết quả của cả một dãy phép thử. Bài toán này sẽ được giải quyết khá dễ dàng nếu các phép thử là độc lập với nhau.

Các phép thử được gọi là độc lập với nhau nếu xác suất để xảy ra một biến cố nào đó trong từng phép thử sẽ không phụ thuộc vào việc biến cố đó có xảy ra ở các phép thử khác hay không. Chẳng hạn tung nhiều lần một đồng xu sẽ tạo nên các phép thử độc lập, lấy nhiều lần sản phẩm từ một lô sản phẩm theo phương thức có hoàn lại cũng tạo nên các phép thử độc lập v.v... Giả sử ta tiến hành n phép thử độc lập. Trong mỗi phép thử chỉ có hai trường hợp: Hoặc biến cố A xảy ra, hoặc biến cố A không xảy ra, xác suất xảy ra của biến cố A trong mỗi phép thử đều bằng $q = 1 - p$. Những bài toán thoả mãn cả ba điều giả thiết trên được gọi là tuân theo lược đồ Bernoulli. Lúc đó xác suất để trong n phép thử độc lập nói trên, biến cố A xuất hiện đúng x lần, ký hiệu là $P_n(x)$, được tính bằng công thức Bernoulli sau đây:

$$P_n(x) = C_n^x p^x q^{n-x} \tag{1.21}$$

Trong đó $x = 0, 1, 2, ..., n$.

Chứng minh. Gọi A_i là biến cố "Xảy ra biến cố A trong phép thử thứ i", $i = \overline{1,n}$. Như vậy, $\overline{A}_i$ là biến cố "Không xảy ra biến cố A trong phép thử thứ i", $i = \overline{1,n}$.

Gọi B là biến cố "Trong n phép thử biến cố A xảy ra đúng x lần". Biến cố này có thể xảy ra theo nhiều phương thức khác nhau, trong đó việc xảy ra của A đúng x lần và không xảy ra đúng n-x lần có thể diễn ra theo những

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

trình tự khác nhau.Do đó:

$$B = A_1A_2...A_x\overline{A}_{x+1}...\overline{A}_n + ... + \overline{A}_1\overline{A}_2...\overline{A}_{n-x}A_{n-x+1}...A_n$$

Tổng số các tích biến cố như vậy trong biểu thức trên là C_n^x tức là số cách chọn ra x phép thử trong đó biến cố A xảy ra từ n phép thử. Đối với mỗi tích ta thấy biến cố A xảy ra x lần, còn $\overline{A}$ xảy ra n-x lần, do đó xác suất của mỗi biến cố tích đều bằng p^xq^{n-x}. Vì các tích biến cố đó là xung khắc từng đôi với nhau, do đó:

$$P_n(x) = P(B) = \underbrace{p^xq^{n-x} + ... + p^xq^{n-x}}_{C_n^x \text{lần}} = C_n^xp^xq^{n-x}$$

Việc sử dụng công thức Bernoulli cho phép trong nhiều trường hợp giải bài toán ngắn gọn hơn nhiều so với sử dụng định lý cộng và định lý nhân xác suất.

Thí dụ 1. Trong phân xưởng có 5 máy hoạt động, xác suất để trong ca mỗi máy bị hỏng đều bằng 0,1. Tìm xác suất để trong ca đó đúng 2 máy hỏng.

Giải. Nếu coi sự hoạt động của mỗi máy là một phép thử, ta có 5 phép thử độc lập. Trong mỗi phép thử chỉ có hai trường hợp: hoặc máy hỏng, hoặc máy chạy tốt. Xác suất hỏng của mỗi máy đều bằng 0,1. Như vậy, bài toán thoả mãn lược đồ Bernoulli. Vì thế xác suất để trong ca có đúng 2 máy hỏng được tính bằng công thức Bernoulli như sau:

$$P_2(2) = C_5^2(0,1)^2(0,9)^3 = 0,0729$$

Hiển nhiên bài toán trên cũng có thể giải bằng cách sử dụng định lý cộng và định lý nhân xác suất, song sẽ dài dòng hơn nhiều.

Thí dụ 2. Bắn 6 viên đạn vào bia. Xác suất trúng đích của mỗi viên là 0,7.Tìm xác suất để có 3 viên trúng bia.

Giải. Bài toán thỏa mãn lược đồ Bernoulli. Do đó xác suất để trong 6 viên có 3 viên trúng bia bằng:

$$P_6(3) = C_6^3(0,7)^3(0,3)^3 = 0,18522$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

1.58. Một nhân viên bán hàng mỗi ngày đi chào hàng ở 10 nơi với xác suất bán được hàng mỗi nơi là 0,2.

a. Tìm xác suất để người đó bán được hàng ở 2 nơi.

b. Tìm xác suất để người đó bán được hàng ở ít nhất 1 nơi.

ĐS: P = 0,8926

1.59. Tỷ lệ phế phẩm của một máy là 5%. Tìm xác suất để trong 12 sản phẩm do máy đó sản xuất ra có

a. 2 phế phẩm

b. Không quá 2 phế phẩm.

1.60. Đề thi trắc nghiệm có 10 câu hỏi, mỗi câu có 5 cách trả lời, trong đó chỉ có một cách trả lời đúng. Một thí sinh chọn cách trả lời một cách hoàn toàn hú hoạ. Tìm xác suất để người đó thi đỗ, biết rằng để đỗ phải trả lời đúng ít nhất 8 câu.

ĐS: P = 0,000078.

1.61. Một siêu thị lắp 4 chuông báo cháy hoạt động độc lập nhau. Xác suất để khi có cháy mỗi chuông kêu là 0,95. Tìm xác suất để có chuông kêu khi có cháy.

10.2. Công thức xác suất đầy đủ

Giả sử biến cố A có thể xảy ra đồng thời với một trong các biến cố $H_1, H_2..., H_n$. Nhóm $H_1, H_2..., H_n$ là nhóm đầy đủ các biến cố. Lúc đó xác suất của biến cố A được tính bằng công thức sau đây.

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A/H_i) \qquad (1.22)$$

Các biến cố $H_1, H_2..., H_n$ thường được gọi là các giả thuyết.

Chứng minh. Vì các biến cố $H_1, H_2..., H_n$ tạo nên một nhóm đầy đủ các biến cố nên biến cố A chỉ có thể xảy ra đồng thời với một trong các biến cố đó:

$$A = H_1A + H_2A + ... + H_nA$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì các biến cố $H_1, H_2, ..., H_n$ xung khắc từng đôi, do đó các tích biến cố $H_1A, H_2A, ...H_nA$ cũng xung khắc từng đôi.

Do vậy:

$$P(A) = P(H_1A) + P(H_2A) + ... + P(H_nA)$$

Áp dụng định lý nhân xác suất đối với các tích H_iA ta có:

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A/H_i)$$

Thí dụ 3. Có 3 hộp giống nhau. Hộp thứ nhất đựng 10 sản phẩm, trong đó có 6 chính phẩm, hộp thứ hai đựng 15 sản phẩm, trong đó có 10 chính phẩm, hộp thứ ba đựng 20 sản phẩm, trong đó có 15 chính phẩm. Lấy ngẫu nhiên một hộp và từ đó lấy ngẫu nhiên một sản phẩm. Tìm xác suất để lấy được chính phẩm.

Giải. Gọi A là biến cố "Lấy được chính phẩm". Biến cố A có thể xảy ra đồng thời với một trong 3 biến cố sau đây tạo nên một nhóm đầy đủ các biến cố:

H_1 - Sản phẩm lấy ra thuộc hộp I.

H_2 - Sản phẩm lấy ra thuộc hộp II.

H_3 - Sản phẩm lấy ra thuộc hộp III.

Vì theo giả thiết của bài toán, các biến cố H_1, H_2 và H_3 là đồng khả năng, do đó:

$$P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}$$

Xác suất có điều kiện của biến cố A khi các biến cố H_1, H_2 và H_3 xảy ra bằng

$$P(A/H_1) = \frac{6}{10}; P(A/H_2) = \frac{10}{15}; P(A/H_3) = \frac{15}{20}$$

Do đó:

$$P(A) = P(H_1)P(A/H_1) + P(H_2)P(A/H_2) + P(H_3)P(A/H_3) =$$

$$= \frac{1}{3}.\frac{6}{10} + \frac{1}{3}.\frac{10}{15} + \frac{1}{3}.\frac{15}{20} = \frac{124}{180} = \frac{31}{45}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 4. Có 2 hộp đựng sản phẩm. Hộp thứ nhất có 10 sản phẩm trong đó có 9 chính phẩm. Hộp thứ hai có 20 sản phẩm trong đó có 18 chính phẩm. Từ hộp thứ nhất lấy ngẫu nhiên một sản phẩm bỏ sang hộp thứ hai. Tìm xác suất để lấy ngẫu nhiên một sản phẩm từ hộp thứ hai được chính phẩm.

Giải. Gọi A là biến cố "Lấy được chính phẩm từ hộp thứ hai". Biến cố A có thể xảy ra đồng thời với một trong hai biến cố sau đây tạo nên một nhóm đầy đủ các biến cố:

H_1 - Sản phẩm bỏ từ hộp thứ nhất sang hộp thứ hai là chính phẩm.

H_2 - Sản phẩm bỏ từ hộp thứ nhất sang hộp thứ hai là phế phẩm.

Xác suất để từ hộp một bỏ sang hộp hai chính phẩm bằng $P(H_1)=\frac{9}{10}$.

Xác suất để từ hộp một bỏ sang hộp hai phế phẩm bằng $P(H_2)=\frac{1}{10}$.

Xác suất có điều kiện để từ hộp thứ hai lấy được chính phẩm khi các giả thuyết H_1 và H_2 xảy ra là:

$$P(A/H_1)=\frac{19}{21}; P(A/H_2)=\frac{18}{21}=\frac{6}{7}$$

Do đó:

$$P(A)=P(H_1)P(A/H_1)+P(H_2)P(A/H_2)$$

$$=\frac{9}{10}.\frac{19}{21}+\frac{1}{10}.\frac{18}{21}=0,9$$

10.3. Công thức Bayes

Giả sử biến cố A có thể xảy ra đồng thời với một trong n biến cố $H_1, H_2, ..., H_n$ tạo nên một nhóm đầy đủ các biến cố.

Lúc đó:

$$P(H_i/A)=\frac{P(H_i)P(A/H_i)}{\sum_{i=1}^{n}P(H_i)P(A/H_i)} \qquad (i=\overline{1,n}) \qquad (1.23)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chứng minh. Theo định lý nhân xác suất ta có:

$$P(AH_i) = P(A)P(H_i / A) = P(H_i)P(A / H_i) \qquad (i = \overline{1,n})$$

Từ đó:

$$P(A)P(H_i / A) = P(H_i)P(A / H_i) \qquad (i = \overline{1,n})$$

Hay: $$P(H_i / A) = \frac{P(H_i)P(A / H_i)}{P(A)} \qquad (i = \overline{1,n})$$

Thay P(A) bằng công thức xác suất đầy đủ ta có:

$$P(H_i / A) = \frac{P(H_i)P(A / H_i)}{\sum_{i=1}^{n} P(H_i)P(A / H_i)} \qquad (i = \overline{1,n})$$

Như trên đã nói, các biến cố $H_1, H_2, ..., H_n$ thường được gọi là *các giả thuyết*. Các xác suất $P(H_1), P(H_2), ..., P(H_n)$ được xác định trước khi phép thử được tiến hành, do đó thường được gọi là các *xác suất tiên nghiệm*. Còn các xác suất $P(H_1 / A), P(H_2 / A), ...,)P(H_n / A)$ được xác định sau khi phép thử đã tiến hành và biến cố A đã xảy ra, do đó được gọi là *các xác suất hậu nghiệm*. Như vậy, công thức Bayes cho phép đánh giá lại xác suất xảy ra các giả thuyết sau khi đã biết kết quả của phép thử là biến cố A đã xảy ra.

Thí dụ 5. Dây chuyền lắp ráp nhận được các chi tiết do hai máy sản xuất. Trung bình máy thứ nhất cung cấp 60% chi tiết, máy thứ hai cung cấp 40% chi tiết. Khoảng 90% chi tiết do máy thứ nhất sản xuất là đạt tiêu chuẩn, còn 85% chi tiết do máy thứ hai sản xuất là đạt tiêu chuẩn. Lấy ngẫu nhiên từ dây chuyền một sản phẩm, thấy nó đạt tiêu chuẩn. Tìm xác suất để sản phẩm đó do máy thứ nhất sản xuất.

Giải. Gọi A là biến cố "Chi tiết lấy từ dây chuyền đạt tiêu chuẩn". Biến cố A có thể xảy ra đồng thời với một trong hai biến cố sau đây tạo nên một nhóm đầy đủ các biến cố:

H_1 - Chi tiết do máy một sản xuất;

H_2 - Chi tiết do máy hai sản xuất;

Như vậy, xác suất để chi tiết do máy một sản xuất bằng:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P(H_1/A)=\frac{P(H_1)P(A/H_1)}{P(H_1)P(A/H_1)+P(H_2)P(A/H_2)}.$$

Theo điều kiện bài toán:

$$P(H_1)=0,6 \qquad P(A/H_1)=0,9$$

$$P(H_2)=0,4 \qquad P(A/H_2)=0,85$$

Do đó:

$$P(H_1/A)=\frac{0,6.0,9}{0,6.0,9+,04.0,85}=0,614$$

Như ta thấy, trước khi phép thử được thực hiện, xác suất của giả thuyết H_1 bằng 0,6. Còn sau khi đã biết kết quả của phép thử thì xác suất đó thay đổi và bằng 0,614.

Thí dụ 6. Trước khi đưa sản phẩm ra thị trường người ta đã phỏng vấn ngẫu nhiên 200 khách hàng về sản phẩm đó và thấy có 34 người trả lời "sẽ mua", 96 người trả lời "Có thể sẽ mua" và 70 người trả lời "Không mua". Kinh nghiệm cho thấy tỷ lệ khách hàng thực sự sẽ mua sản phẩm tương ứng với những cách trả lời trên là 40%, 20% và 1%.

a) Hãy đánh giá thị trường tiềm năng của sản phẩm đó.

b) Trong số khách hàng thực sự mua sản phẩm thì có bao nhiêu phần trăm trả lời "sẽ mua" ?

Giải. a) Thị trường tiềm năng của sản phẩm chính là tỷ lệ khách hàng thực sự sẽ mua sản phẩm đó. Vì vậy, gọi A là biến cố "lấy ngẫu nhiên một khách hàng thì người đó thực sự sẽ mua sản phẩm". Có 3 giả thuyết đối với người khách hàng đó:

H_1 - Người đó trả lời "Sẽ mua".

H_2 - Người đó trả lời "Có thể mua".

H_3 - người đó trả lời "Không mua".

Theo công thức xác suất đầy đủ:

$$P(A)=P(H_1).P(A/H_1)+P(H_2).P(A/H_2)+P(H_3).P(A/H_3)=$$

$$=\frac{34}{200}.0,4+\frac{96}{200}.0,2+\frac{70}{200}.0,01=0,1675$$

Vậy thị trường tiềm năng của sản phẩm đó là 16,75%

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b) Theo công thức Bayes

$$P(H_1/A) = \frac{P(H_1).P(A/H_1)}{P(A)} = \frac{0,17.0,4}{0,1675}$$

$$= 0,40597 = 40,597\%$$

Thí dụ 7. Có hai lô sản phẩm, lô thứ nhất có tỷ lệ chính phẩm là $\frac{3}{4}$, còn lô thứ hai có tỷ lệ chính phẩm là $\frac{2}{3}$. Lấy ngẫu nhiên một lô và từ đó lấy ngẫu nhiên một sản phẩm thấy nó là chính phẩm. Sản phẩm được bỏ trở lại và từ lô đó lấy tiếp một sản phẩm. Tìm xác suất để lần thứ hai cũng lấy được chính phẩm.

Giải. Gọi A là biến cố "Sản phẩm lấy lần đầu là chính phẩm". Biến cố A có thể xảy ra với một trong hai giả thuyết sau:

H_1 - Sản phẩm được lấy ra từ lô I;

H_2 - Sản phẩm được lấy ra từ lô II.

Theo công thức xác suất đầy đủ, ta có:

$$P(A) = P(H_1)P(A/H_1) + P(H_2)P(A/H_2)$$

Theo điều kiện đầu bài

$$P(H_1) = P(H_2) = \frac{1}{2}$$

$$P(A/H_1) = \frac{3}{4}; P(A/H_2) = \frac{2}{3}$$

Do đó:

$$P(A) = \frac{1}{2}.\frac{3}{4} + \frac{1}{2}.\frac{2}{3} = \frac{17}{24}$$

Sau khi biến cố A đã xảy ra. xác suất của biến cố H_1 và H_2 thay đổi theo công thức Bayes như sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P(H_1/A)=\frac{P(H_1).P(A/H_1)}{P(A)}=\frac{3}{8}:\frac{17}{24}=\frac{9}{17}$$

$$P(H_2/A)=\frac{P(H_2).P(A/H_2)}{P(A)}=\frac{1}{3}:\frac{17}{24}=\frac{8}{17}$$

Gọi B là biến cố "Sản phẩm lấy lần thứ hai là chính phẩm". B vẫn có thể xảy ra với một trong hai giả thuyết H_1 và H_2, do đó theo công thức xác suất đầy đủ:

$$P(B)=P(H_1/A)P(B/H_1A)+P(H_2/A)P(B/H_2A)$$

Vì sản phẩm thứ nhất được bỏ trở lại lô, do đó tỷ lệ chính phẩm ở các lô đó vẫn không thay đổi.Vì thế:

$$P(B/H_1A)=\frac{3}{4};P(B/H_2A)=\frac{2}{3}$$

$$P(B)=\frac{9}{17}.\frac{3}{4}+\frac{8}{17}.\frac{2}{3}=\frac{145}{204}=0,71$$

Bài tập

1.62. Hai máy cùng sản xuất một loại sản phẩm. Tỷ lệ phế phẩm của máy I là 3% của máy II là 2%. Từ một kho gồm 2/3 sản phẩm của máy I và 1/3 sản phẩm của máy II ta lấy ra một sản phẩm. Tính xác suất để sản phẩm đó là tốt.

ĐS: P= 0,9733.

1.63. Có hai xạ thủ loại I và tám xạ thủ loại II, xác suất bắn trúng đích của các loại xạ thủ theo thứ tự là 0,9 và 0,8.

a. Lấy ngẫu nhiên ra một xạ thủ và xạ thủ đó bắn một viên đạn. Tìm xác suất để viên đạn đó trúng đích.

b. Nếu lấy ra hai xạ thủ và mỗi người bắn một viên thì khả năng để cả hai viên đều trúng đích là bao nhiêu?

ĐS: $P_a = 0,82$; $P_b = 0,6722$.

1.64. Có 2 lô sản phẩm

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lô 1: Gồm toàn chính phẩm

Lô 2: Có tỷ lệ phế phẩm và chính phẩm là 1/4

Chọn ngẫu nhiên một lô, từ lô này lấy ngẫu nhiên 1 sản phẩm, thấy nó là chính phẩm. Hỏi rằng nếu lấy ngẫu nhiên (cũng từ lô đã chọn) một sản phẩm nữa thì xác suất để sản phẩm này là phế phẩm là bao nhiêu?

ĐS: P= 4/45

1.65. Bắn 3 phát đạn vào một máy bay với xác suất trúng tương ứng là 0,4; 0,5; và 0,7. Nếu trúng một phát thì xác suất rơi máy bay là 0,2; nếu trúng 2 phát thì xác suất rơi máy bay là 0,6 còn nếu trúng cả 3 phát thì chắc chắn máy bay rơi. Tìm xác suất để máy bay rơi.

ĐS: P = 0,458

1.66. Có hai lô hàng

Lô 1: Có 7 chính phẩm và 3 phế phẩm.

Lô 2: Có 8 chính phẩm và 2 phế phẩm.

Từ lô thứ nhất lấy ra 2 sản phẩm, từ lô thứ hai lấy ra 3 sản phẩm rồi trong số sản phẩm được lấy ra lại lấy tiếp ngẫu nhiên 2 sản phẩm. Tính xác suất để trong 2 sản phẩm đó có ít nhất một chính phẩm.

ĐS: P = 0,951.

1.67. Có hai lô sản phẩm

Lô 1: Có a chính phẩm và b phế phẩm.

Lô 2: Có c chính phẩm và d phế phẩm.

Từ lô thứ nhất bỏ sang lô thứ hai một sản phẩm, sau đó từ lô thứ hai bỏ sang lô thứ nhất một sản phẩm, sau đó từ lô thứ nhất lấy ra một sản phẩm. Tìm xác suất để lấy được chính phẩm.

$$\textit{ĐS:}\ P = \frac{a}{a+b} + \frac{bc - ad}{(a+b)^2(c+d+1)}$$

Kết quả cho thấy rằng nếu tỷ lệ chính phẩm và phế phẩm trong cả hai lô là như nhau

$$\frac{b}{a} = \frac{d}{c} \Rightarrow bc - ad = 0$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thì xác suất để lấy được chính phẩm sẽ không thay đổi.

1.68. Tỷ lệ người dân nghiện thuốc lá ở một vùng là 30%. Biết rằng tỷ lệ người bị viêm họng trong số người nghiện thuốc là 60%, còn tỷ lệ người bị viêm họng trong số người không hút thuốc lá là 40%.

a. Lấy ngẫu nhiên một người, biết rằng người đó viêm họng. Tính xác suất người đó nghiện thuốc.

b. Nếu người đó không bị viêm họng, tính xác suất để người đó là nghiện thuốc.

ĐS: $P_a = 0,3913$; $P_b = 0,222$.

1.69. Một cỗ máy có 3 bộ phận 1, 2, 3. Xác suất hỏng của các bộ phận trong thời gian làm việc theo thứ tự là 0,2; 0,4; và 0,3. Cuối ngày làm việc được biết rằng có 2 bộ phận bị hỏng. Tính xác suất để hai bộ phận hỏng đó là bộ phận 1 và 2.

ĐS: $P = 0,298$.

1.70. Trong một bệnh viên, tỷ lệ bệnh nhân các tỉnh như sau: tỉnh A: 25%, tỉnh B: 35%, tỉnh C: 40%. Biết rằng tỷ lệ bệnh nhân là kỹ sư của các tỉnh là: tỉnh A: 2%, tỉnh B: 3%, tỉnh C: 3,5%. Chọn ngẫu nhiên một bệnh nhân. Tính xác suất để bệnh nhân đó là kỹ sư.

1.71. Một người có 3 chỗ ưa thích như nhau để câu cá. Xác suất để câu được cá ở những chỗ đó tương ứng là: 0,6; 0,7 và 0,8. Biết rằng ở một chỗ người đó đã thả câu 3 lần và chỉ câu được một con cá. Tìm xác suất để cá được câu ở chỗ thứ nhất.

ĐS: $P = 0,502$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

11. TÓM TẮT CHƯƠNG 1

1. Trong thực tế mọi hiện tượng đều là các biến cố ngẫu nhiên.
2. Xác suất của một biến cố là con số đặc trưng cho khả năng xuất hiện khách quan của biến cố đó, đồng thời nó cũng đặc trưng cho cơ cấu của một tập hợp.
3. Đối với các biến cố giản đơn có thể tìm xác suất bằng định nghĩa cổ điển với điều kiện là các kết cục của phép thử phải là hữu hạn và đồng khả năng.
4. Khi các kết cục của phép thử là không đồng khả năng thì có thể dùng định nghĩa thống kê về xác suất nếu có thể tiến hành được một số lớn phép thử.
5. Biến cố phức hợp thường được biểu diễn thông qua tổng hoặc tích của các biến cố giản đơn. Khái niệm tổng các biến cố luôn gắn liền với khái niệm về tính xung khắc hay không xung khắc của các biến cố đó, còn khái niệm tích các biến cố luôn gắn liền với tính độc lập hay phụ thuộc của chúng.
6. Với tổng các biến cố có thể dùng định lý cộng xác suất, còn với tích các biến cố có thể dùng định lý nhân xác suất để tìm xác suất của chúng. Mặt khác còn có thể dùng các hệ quả của chúng để tính.
7. Khái niệm tổng và tích các biến cố có thể biểu diễn tương đương theo nhiều cách khác nhau nhằm cho phép dùng những cách thức khác nhau để tìm xác suất.
8. Khi trong phép thử xuất hiện một nhóm các giả thuyết tạo nên một nhóm đầy đủ các biến cố thì dẫn đến việc áp dụng công thức xác suất đầy đủ và công thức Bayes.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các ký hiệu và công thức cơ bản

* A, B, C $A_1, A_2, ..., A_n$ - Biến cố ngẫu nhiên
* U - Biến cố chắc chắn
* V - Biến cố không thể có
* $\overline{A}$ - Biến cố đối lập với biến cố A
* P(A) - Xác suất của biến cố A
* Định nghĩa cổ điển về xác suất

$P(A) = \frac{m}{n}$, trong đó: m- Số kết cục thuận lợi

n- Số kết cục duy nhất đồng khả năng

* Định nghĩa thống kê về xác suất

$P(A) \approx f = \frac{k}{n}$, trong đó: k- Số lần xuất hiện biến cố

n-Số phép thử được thực hiện

* Định lý cộng xác suất

$P(A+B) = P(A) + P(B)$ nếu A và B xung khắc

$P(A+B) = P(A) + P(B) - P(AB)$ nếu A và B không xung khắc

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \text{ nếu } A_1, ..., A_n \text{ xung khắc từng đôi}$$

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) - ... +$$

$$+(-1)^{n-1}.P(A_1 ... A_n)$$

Nếu $A_1, ..., A_n$ không xung khắc.

*Định lý nhân xác suất

$P(A.B) = P(A).P(B)$ nếu A và B độc lập

$P(A.B) = P(A).P(B/A) = P(B).P(A/B)$ nếu A và B phụ thuộc

$$P\left(\prod_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i) \text{ nếu } A_1, ..., A_n \text{ độc lập toàn phần}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P\left(\prod_{i=1}^{n} A_i\right) = P(A_1).P(A_2/A_1)...P(A_n/A_1...A_{n-1})$$

nếu $A_1,...,A_n$ phụ thuộc

$$P\left(\prod_{i=1}^{n} A_i\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i + A_j) + \sum_{i<j<k} P(A_i + A_j + A_k) -$$
$$-... + (-1)^{n-1} P(A_1 + A_2 + ... + A_n)$$

* $P(A) = 1 - P(\overline{A})$

* Xác suất của tổng các biến cố không xung khắc và độc lập toàn phần

$$P\left(\sum_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n} P(\overline{A}_i)$$

* Công thức Bernoulli

$$P_n(x) = C_n^x p^x (1-p)^{n-x}$$

* Công thức xác suất đầy đủ

$$P(A) = \sum_{i=1}^{n} P(H_i).P(A/H_i)$$

* Công thức Bayes

$$P(H_i/A) = \frac{P(H_i).P(A/H_i)}{\sum_{i=1}^{n} P(H_i).P(A/H_i)}$$

* Nhị thức Newton

$$(a+b)^n = \sum_{n=0}^{n} C_n^x a^x b^{n-x}$$

* Các công thức giải tích tổ hợp

Số chỉnh hợp: $A_n^k = \frac{n!}{(n-k)!}$

Số hoán vị: $P_n = n!$

Số tổ hợp: $C_n^k = \frac{n!}{k!(n-k)!}$

Số chỉnh hợp lặp: $\overline{A}_n^k = n^k$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Tiếng Việt	Tiếng Anh
Phép thử	Experiment
Biến cố	Event
Biến cố chắc chắn	Certain event
Biến cố không thể có	Imposible event
Biến cố ngẫu nhiên	Random event
Xác suất	Probability
Kết cục đồng khả năng	Equaly likely possible outcomes
Kết cục thuận lợi	Favorable outcomes
Định nghĩa cổ điển về xác suất	Clasical definition of probability
Sơ đồ Venn	Venn diagram
Tần suất của biến cố	Relative-frequency of event
Định nghĩa thống kê về xác suất	Statistical definition of probability
Bảng số ngẫu nhiên	Random number table
Định nghĩa hình học về xác suất	Geometrical definition of probability
Định nghĩa tiên đề về xác suất	Axiomical definition of probability
Định nghĩa xác suất chủ quan	Subjective definition of probability
Nguyên lý xác suất lớn	Rule of the grand probability
Nguyên lý xác suất nhỏ	Rule of the small probability
Tổng của các biến cố	Union of events
Biến cố xung khắc	Mutually exclusive events
Biến cố không xung khắc	Non-exclusive events
Biến cố xung khắc từng đôi	Paired mutually exclusive events
Nhóm đầy đủ các biến cố	Universal set of events
Biến cố đối lập	Complement of event

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tích của các biến cố	Intersection of events
Biến cố độc lập	Independent events
Biến cố phụ thuộc	Dependent events
Biến cố độc lập từng đôi	Paired independent events
Biến cố độc lập toàn phần	Universal independent events
Xác suất có điều kiện	Conditional probability
Định lý cộng xác suất	Theorem of Addition
Định lý nhân xác suất	Theorem of Multiplication
Công thức Bernoully	Bernoully' theorem
Công thức xác suất đầy đủ	Law of total probability
Công thức Bayes	Bayes' theorem
Xác suất tiên nghiệm	A priori probability
Xác suất hậu nghiệm	A posteri probability

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

1.72. Hãy cho biết với các biến cố sau đây thì phải dùng định nghĩa xác suất nào? Giải thích?

A- Ngày mùng một tết năm tới trời sẽ mưa

B- Trúng thưởng khi mua một vé số

C- Bị tai nạn khi đi ô tô khách

D- Phải bảo hành khi mua một TV Sony

1.73. Tỷ lệ cá cược giữa đội bóng A và B là 1/5. Vậy xác suất để A thắng B bằng bao nhiêu?

1.74. Các mệnh đề sau đây đúng hay sai? Hãy giải thích.

a- Nếu A và B xung khắc nhau thì chúng độc lập nhau.

b- Nếu A và B là các biến cố đối lập nhau thì chúng:

+ Xung khắc với nhau.

+ Độc lập nhau.

c- Các biến cố độc lập từng đôi thì cũng độc lập toàn phần với nhau.

d- Nếu $P(A+B)=P(A)+P(B)$ thì A và B xung khắc.

1.75. Gieo hai con xúc xắc. Nếu ký hiệu:

A- Chỉ một con xúc xắc xuất hiện mặt 6 chấm

B- Cả hai con xúc xắc đều xuất hiện mặt 6 chấm

A và B có tạo nên một nhóm đầy đủ các biến cố hay không?

1.76. Hãy chứng minh rằng với 3 biến cố bất kỳ A, B, C luôn thoả mãn điều kiện:

$$P(A+B+C) \le P(A)+P(B)+P(C)$$

1.77. Hãy chứng minh rằng các biến cố $A, \overline{A}B$ và $\overline{A+B}$ lập thành một nhóm đầy đủ các biến cố.

1.78. Hãy chứng minh rằng với mọi A và B

+ $P(\overline{A+B}) = P(\overline{A}.\overline{B})$

+ $P(\overline{A.B}) = P(\overline{A}+\overline{B})$

1.79. Bất đẳng thức sau đây là đúng hay sai? Tại sao?

$P(A+B) \ge P(A.B)$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1.80. Hãy chứng minh rằng nếu $P(A) \neq 0; P(B) \neq 0$ và $P(A/B) > P(A)$ thì $P(B/A) > P(B)$.

1.81. Cho biết phát biểu sau đây đúng hay sai? Tại sao?

a- Nếu $P(A+B) = 0$ và P(A)=0; P(B)=0 thì A và B đều là các biến cố không thể có.

b- Nếu A và B không xung khắc thì luôn luôn độc lập nhau.

1.82. Cho $P(A) = p_1; P(B) = p_2$ và $P(AB) = p_3$. Hãy xác định xác suất của biến cố $\overline{A}.\overline{B}$.

1.83. Chứng minh rằng nếu hai biến cố A và B độc lập thì các cặp biến cố dưới đây cũng độc lập nhau:

a- A và $\overline{B}$

b- $\overline{A}$ và B

c- $\overline{A}$ và $\overline{B}$

1.84. Gieo một con xúc xắc đối xứng với đồng chất. Gọi A là biến cố được mặt chẵn chấm. B là biến cố được mặt có số chấm là bội số của 3.

a- Hai biến cố trên có xung khắc không? Tại sao?

b- Hai biến cố trên có độc lập không? Tại sao?

1.85. Cho hai biến cố ngẫu nhiên A và B có:

$$P(A) = \frac{1}{4}; P(B) = \frac{1}{3}; P(A+B) = \frac{23}{60}$$

Hãy tính: $P(A/B)$; $P(\overline{A}/B)$; $P(AB/B)$; $P(A.\overline{B}/B)$;

$P(A+B/A.\overline{B}); P(\overline{A}.B/\overline{A}+B); P(A\overline{B}/\overline{B})$

1.86. Có 3 người muốn đi xem đá bóng mà chỉ có 2 vé nên phải dùng phương pháp bắt thăm với hai thăm "có" và một thăm "không". Chứng minh rằng phương pháp này là công bằng.

1.87. Hãy cho biết hai ý nghĩa có thể khai thác từ giá trị xác suất của một biến cố. Cho ví dụ minh hoạ.

1.88. Điều kiện cần và đủ để hai biến cố A và B độc lập là gì?

1.89. Mục đích của việc sử dụng công thức Bayes là gì? Cho ví dụ minh hoạ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập tổng hợp chương 1

1.90. Xác suất của biến cố A là 0,7. Hãy cho biết con số đó có ý nghĩa gì?

1.91. Có 30 sản phẩm trong đó có 3 phế phẩm được bỏ ngẫu nhiên vào 3 hộp với số lượng bằng nhau. Tìm xác suất để có một hộp nào đó có một phế phẩm.

ĐS: 0,911

1.92. Ra khỏi phòng khách N người cùng số giày xỏ ngẫu nhiên vào một đôi giày trong bóng tối. Mỗi người chỉ có thể phân biệt chiếc giày phải với chiếc giày trái, còn không thể phân biệt được giày của mình với giày của người khác. Tìm xác suất để:

a. Mỗi người khách xỏ vào đúng đôi giày của mình.

b. Mỗi người khách xỏ vào đúng hai chiếc giày của cùng một đôi giày nào đó.

ĐS: $P_a = \frac{1}{(N!)^2}$; $P_b = \frac{1}{N!}$

1.93. Tỷ lệ phế phẩm của một máy là 5%. Người ta dùng một thiết bị kiểm tra tự động đạt được độ chính xác khá cao song vẫn có sai sót. Tỷ lệ sai sót đối với chính phẩm là 4% còn đối với phế phẩm là 1%. Nếu sản phẩm bị kết luận là phế phẩm thì bị loại.

a. Tìm tỷ lệ sản phẩm được kết luận là chính phẩm mà thực ra là phế phẩm.

b. Tìm tỷ lệ sản phẩm bị kết luận là phế phẩm mà thực ra là chính phẩm.

c. Tìm tỷ lệ sản phẩm bị thiết bị kiểm tra đó kết luận nhầm.

1.94. Thống kê 2000 sinh viên một khoá của trường kinh tế theo giới tính và ngành học thu được các số liệu sau:

	Nam	Nữ
Học kinh tế	400	500
Học quản trị kinh doanh	800	300

Lấy ngẫu nhiên một sinh viên khoá đó. Tìm xác suất để được:

a. Nam sinh viên

b. Sinh viên học kinh tế

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

c. Hoặc nam sinh viên, hoặc học kinh tế.

d. Nam sinh viên và học kinh tế.

e. Nếu đã chọn được một nam sinh viên thì xác suất để người đó học kinh tế bằng bao nhiêu?

ĐS: $P_a = 0,6$ $P_b = 0,45$ $P_c = 0,85$

$P_d = 0,2$ $P_e = 1/3$

1.95. Lô hàng xuất khẩu có 100 kiện hàng, trong đó có 60 kiện hàng của xí nghiệp A và 40 kiện hàng của xí nghiệp B. Tỷ lệ phế phẩm của xí nghiệp A và B tương ứng là 30% và 10%. Người ta lấy ngẫu nhiên một kiện hàng để kiểm tra.

a. Trước khi mở kiện hàng để kiểm tra thì xác suất để kiện hàng đó là kiện hàng của xí nghiệp A là bao nhiêu?

b. Giả sử mở kiện hàng và lấy ngẫu nhiên ra một sản phẩm thì được phế phẩm. Vậy xác suất để đó là kiện hàng của xí nghiệp A là bao nhiêu?

c. Giả sử lấy tiếp sản phẩm thứ hai từ kiện hàng đó thì cũng lấy được phế phẩm. Vậy xác suất để đó là kiện hàng của xí nghiệp A là bao nhiêu?

1.96. Điều tra sở thích xem TV của các cặp vợ chồng cho thấy 30% các bà vợ thường xem chương trình thể thao, 50% ông chồng thường xem chương trình thể thao, song nếu thấy vợ xem thì tỷ lệ chồng xem cùng là 60%. Lấy ngẫu nhiên một cặp vợ chồng. Tìm xác suất để:

a. Cả hai cùng thường xem chương trình thể thao

b. Có ít nhất một người thường xem

c. Không có ai thường xem

d. Nếu chồng xem thì vợ xem cùng.

e. Nếu chồng không xem thì vợ vẫn xem.

ĐS: $P_a = 0,18$ $P_b = 0,62$ $P_c = 0,38$

$P_d = 0,36$ $P_e = 0,24$

1.97. Một nhân viên bán hàng mỗi năm đến bán ở công ty A ba lần. Xác suất để lần đầu bán được hàng là 0,8. Nếu lần trước bán được hàng thì xác suất để lần sau bán được hàng là 0,9 còn nếu lần trước không bán được hàng thì xác suất để lần sau bán được hàng chỉ là 0,4. Tìm xác suất để:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Cả ba lần đều bán được hàng.

b. Có đúng hai lần bán được hàng.

1.98. Người ta biết một cặp trẻ sinh đôi có thể là một cặp sinh đôi thật do cùng một trứng sinh ra (E_1). Trong trường hợp đó chúng bao giờ cũng có cùng giới tính. Nếu chúng do các trứng khác nhau sinh ra (E_2) thì xác suất để chúng có cùng giới tính là 1/2. Bây giờ nếu cặp trẻ sinh đôi có cùng giới tính thì xác suất để chúng là cặp sinh đôi thật là bao nhiêu?

Giải: Gọi P là xác suất để một cặp sinh đôi do cùng một trứng sinh ra $P(E_1) = P$; E_2 là biến cố đối lập của E_1. Ta có $P(E_2) = 1 - P$. Gọi A là biến cố cặp sinh đôi có cùng giới tính. Phải tìm $P(E_1/A)$. Ta có $P(E_1) = P$;

$$P(A/E_1) = 1 \; ; \; P(A/E_2) = \frac{1}{2}$$

$$\text{Do đó } P(E_1/A) = \frac{1.P}{1.P + \frac{(1-P)}{2}} = \frac{2P}{P+1}$$

1.99. Một hộp kín đựng 3 quả cầu đỏ và 6 quả cầu xanh. Tính xác suất để khi chia hộp cầu đó một cách ngẫu nhiên thành 3 phần bằng nhau thì:

a. Cả 3 quả cầu đỏ ở trong một phần.

b. Mỗi một phần có một quả cầu đỏ.

1.100. Tại một siêu thị, hệ thống phun nước tự động được lắp liên kết với một hệ thống báo động hoả hoạn. Khả năng hệ thống phun nước bị hỏng là 0,1. Khả năng để hệ thống báo động bị hỏng là 0,2. Khả năng để hai hệ thống này cùng hỏng là 0,04. Hãy tính xác suất:

a. Có ít nhất một hệ thống hoạt động bình thường

b. Cả hai hệ thống đều hoạt động bình thường.

ĐS: a. 0,96;b. 0,74

1.101. Trong một kho rượu số lượng chai rượu loại A và loại B là bằng nhau. Người ta lấy ngẫu nhiên 1 chai rượu trong kho và đưa cho 4 người sành rượu nếm thử để xác định xem đây là loại rượu nào. Giả sử mỗi người có khả năng đoán đúng là 80%. Có ba người kết luận chai rượu thuộc loại A và một người kết luận chai rượu thuộc loại B. Vậy

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

chai rượu được chọn thuộc loại A với xác suất bằng bao nhiêu?

1.102. Có 2 hộp đựng các mẫu hàng xuất khẩu. Hộp thứ nhất đựng 10 mẫu trong đó có 6 mẫu loại A và 4 mẫu loại B. Hộp thứ hai đựng 10 mẫu trong đó có 3 mẫu loại A và 7 mẫu loại B.

a. Từ mỗi hộp lấy ngẫu nhiên một mẫu hàng. Tính xác suất để 2 mẫu lấy ra đều cùng loại.

b. Giả sử xác suất lựa chọn các hộp lần lượt là 0,45 và 0,55.

* Chọn ngẫu nhiên một hộp và từ đó lấy ngẫu nhiên một mẫu. Tính xác suất để mẫu lấy ra là loại B.
* Chọn ngẫu nhiên một hộp và từ đó lấy ngẫu nhiên một mẫu thì được mẫu loại A. Hỏi mẫu đó có khả năng thuộc hộp nào là nhiều hơn?

1.103. Qua kinh nghiệm, người quản lý của một cửa hàng bán giầy thể thao biết rằng xác suất để một đôi đế cao su của một hãng nào đó có 0 hoặc 1 hoặc 2 chiếc bị hỏng tương ứng là: 0,90; 0,08 và 0,02.
Anh ta lấy ngẫu nhiên một đôi giày loại đó từ tủ trưng bày và sau đó lấy ngẫu nhiên 1 chiếc thì nó bị hỏng. Hỏi xác suất để chiếc kia cũng bị hỏng là bao nhiêu:
ĐS: 1/3

1.104. Hai cửa hàng A và B cung cấp các hộp đĩa mềm máy tính cho một trung tâm tin học với tỷ lệ 3/2. Tỷ lệ đĩa bị lỗi của các cửa hàng tương ứng là 1% và 2%. Một sinh viên đến thực tập tại trung tâm chọn ngẫu nhiên một hộp đĩa gồm 20 chiếc và từ đó rút ngẫu nhiên ra một đĩa.
a. Tính xác suất để sinh viên đó rút phải đĩa bị lỗi.
b. Sau khi khởi động máy, sinh viên đó nhận thấy quả thật đĩa bị lỗi. Tính xác suất để đĩa này thuộc cửa hàng A.

1.105. Tỷ lệ phế phẩm của máy I là 1%, của máy II là 2%. Một lô sản phẩm gồm 40% sản phẩm của máy I và 60% sản phẩm của máy II. Người ta lấy ngẫu nhiên ra 2 sản phẩm để kiểm tra.
a. Tìm xác suất trong 2 sản phẩm lấy ra có ít nhất một sản phẩm tốt?
b. Giả sử hai sản phẩm kiểm tra đều là tốt thì khả năng lấy tiếp được hai sản phẩm tốt nữa là bao nhiêu?

1.106. Một công nhân đi làm ở thành phố khi trở về nhà có hai cách: đi

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

theo đường ngầm hoặc đi qua cầu. Biết rằng anh ta đi lối đường ngầm trong 1/3 trường hợp; còn lại đi lối cầu. Nếu đi lối đường ngầm 75% trường hợp anh ta về nhà trước 6 giờ; còn lối đi cầu chỉ có 70% trường hợp. Tìm xác suất để công nhân đó đã đi lối cầu, biết rằng anh ta về nhà sau 6 giờ.

ĐS: 0,7059.

1.107. Ba công nhân cùng sản xuất một loại sản phẩm, xác suất để người thứ nhất và thứ hai làm ra chính phẩm bằng 0,9. Còn xác suất để người thứ 3 làm ra chính phẩm bằng 0,8. Một người trong số đó làm ra 8 sản phẩm, thấy có 2 phế phẩm. Tìm xác suất để trong 8 sản phẩm tiếp theo cũng do người đó sản xuất sẽ có 6 chính phẩm.

ĐS: $P \approx 0,23$.

1.108. Một lô hàng có 8 sản phẩm cùng loại. Kiểm tra ngẫu nhiên 4 sản phẩm thấy có 3 chính phẩm và 1 phế phẩm. Tìm xác suất để khi kiểm tra tiếp 3 sản phẩm nữa sẽ có một chính phẩm và 2 phế phẩm.

ĐS: $P = 0,214$

1.109. Một hộp có n sản phẩm, bỏ vào đó một sản phẩm tốt sau đó lấy ngẫu nhiên ra 1 sản phẩm. Tìm xác suất để sản phẩm lấy ra là tốt nếu mọi giả thiết về trạng thái cấu thành ban đầu của hộp là đồng xác suất.

ĐS: $P = \dfrac{n+2}{2(n+1)}$

1.110. A chơi cờ với B với xác suất thắng mỗi ván bằng P. Tìm xác suất P để A thắng chung cuộc trong 2 ván dễ hơn thắng chung cuộc trong 4 ván, biết rằng để thắng chung cuộc A phải thắng được ít nhất một nửa tổng số ván chơi.

ĐS: $0 < p < \dfrac{2}{3}$

1.111. Một xí nghiệp có hai dây chuyền cùng lắp ráp một loại sản phẩm với tỷ lệ phế phẩm tương ứng là 2% và 3%. Tìm xác suất để một khách hàng mua 2 sản phẩm của xí nghiệp đó thì mua phải một phế phẩm.

ĐS: $P = 0,04875$

1.112. Hai cầu thủ bóng rổ, mỗi người ném bóng 2 lần, xác suất ném trúng

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

đích của mỗi cầu thủ theo thứ tự là 0,6 và 0,7.

Tính xác suất:

a. Số lần ném trúng rổ của cầu thủ thứ nhất nhiều hơn số lần ném trúng rổ của cầu thủ thứ hai.

b. Số lần ném trúng rổ của hai người như nhau.

ĐS: $P_a = 0,2268$; $P_b = 0,3924$.

1.113. Trong bình cầu có a cầu trắng và b cầu đen. Hai người lần lượt lấy ra từng quả cầu theo phương thức có hoàn lại. Tính xác suất để người thứ nhất lấy được cầu trắng trước.

ĐS: $P = \dfrac{a+b}{a+2b}$

1.114* Trong hộp có n sản phẩm, trong đó mỗi sản phẩm đều có thể là chính phẩm hoặc phế phẩm với xác suất như nhau. Lấy ngẫu nhiên lần lượt k sản phẩm theo phương thức có hoàn lại thì được toàn chính phẩm. Tính xác suất để hộp có chứa toàn chính phẩm.

Hướng dẫn: Gọi A là biến cố k sản phẩm lấy ra đều là chính phẩm; $H_i (i = 0,n)$ là biến cố hộp đó có i chính phẩm.

$$P(H_n/A) = \frac{n^k}{n + \dfrac{n(n-1)}{2!}2^k + \dfrac{n(n-1)(n-2)}{3!}3^k + ... + n^k}$$

1.115*. Trong rạp có n chỗ được đánh số, n người có vé vào ngồi một cách ngẫu nhiên. Tìm xác suất để có m người ngồi đúng chỗ.

Giải: Xác suất để có m người ngồi đúng chỗ $P(A) = \dfrac{C_n^m}{A_n^m} = \dfrac{1}{m!}$

Xác suất để n-m người còn lại ngồi sai chỗ $P(B) = \sum_{k=0}^{n-m} \dfrac{(-1)^k}{k!}$

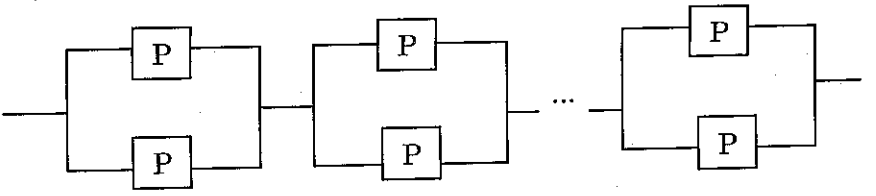
Từ đó $P = P(A).P(B) = \dfrac{1}{m!}\sum_{k=0}^{n-m} \dfrac{(-1)^k}{k!}$

1.116*. Một hệ thống kỹ thuật gồm n bộ phận với xác suất hoạt động tốt

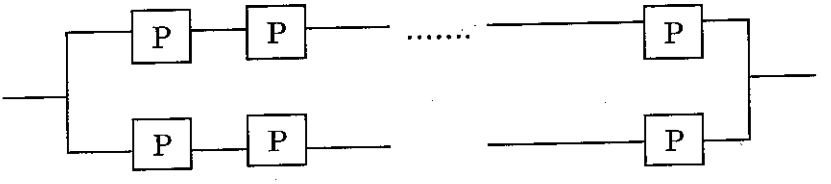
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

của mỗi bộ phận là p. Hệ thống sẽ ngừng hoạt động khi có ít nhất một bộ phận bị hỏng. Để nâng cao độ tin cậy của hệ thống người ta dự trữ thêm n bộ phận nữa theo hai phương thức sau:

a)



b)



Hỏi phương thức dự trữ nào mang lại độ tin cậy cao hơn cho cả hệ thống.

Giải: Độ tin cậy của hệ thống được dự trữ theo phương thức a là:

$P_a = [1 - (1 - p)^2]^n$

và theo phương thức b là

$P_b = 1 - (1 - p^n)^2$

Ta sẽ chứng tỏ rằng $P_a > P_b$ với mọi $n > 1$ và $0 < P < 1$.

Vì $P_a = p^n (2 - p^n)$

nên chỉ cần chứng minh rằng $(2 - p)^n > 2 - p^n$

Đặt $q = 1 - p$ $(q > 0)$ ta có

$(2 - 1 + q)^n > 2 - (1 - q)^n$ hay $(1 + q)^n > 2$

Thật vậy theo nhị thức Newton ta có:

$$(1 + q)^n + (1 - q)^n = 1 + nq + \frac{n(n-1)}{2} q^2 + ... + ...$$

$$+ 1 - nq + \frac{n(n-1)}{2} q^2 + ... = 2 + n(n - 1)q^2 + ... > 2.$$

Đó là điều phải chứng minh.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1.117*. Hai người chơi cờ thoả thuận với nhau là ai thắng trước một số ván nhất định thì sẽ thắng cuộc. Trận đấu bị gián đoạn khi người thứ nhất còn thiếu m ván thắng, người thứ hai còn thiếu n ván thắng. Vậy phải phân chia phần tiền đặt như thế nào là hợp lý nếu xác suất thắng mỗi ván của mỗi người đều bằng 0,5.

ĐS: Phần tiền đặt phải chia theo tỷ lệ: P_1/P_2 trong đó P_1 và P_2 là xác suất thắng cuộc của hai người nếu trận đấu được tiếp tục đến cùng.

$$P_1 = \frac{1}{2^m}\left(1+\frac{1}{2}C_m^1+\frac{1}{2^2}C_{m-1}^2+...+\frac{1}{2^{n-1}}C_{m+n-2}^{n-1}\right)$$

$$P_2 = \frac{1}{2^n}\left(1+\frac{1}{2}C_n^1+\frac{1}{2^2}C_{n-1}^2+...+\frac{1}{2^{m-1}}C_{m+n-2}^{m-1}\right)$$

1.118* Một người bỏ hai bao diêm vào túi, mỗi bao có n que diêm. Mỗi khi hút thuốc người đó rút ngẫu nhiên một bao và đánh một que. Tìm xác suất để khi người đó phát hiện một bao đã hết diêm thì bao kia còn lại đúng r que diêm.

ĐS: $P = \dfrac{C_{2n-r+1}^r}{2^{2n-r+1}}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Lê Văn Phong. *Giáo trình lý thuyết xác suất*. ĐH Kinh tế Kế hoạch. Hà Nội. 1972.
2. Nguyễn Đình Cử, Nguyễn Cao Văn. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Thống kê. Hà Nội. 1991.
3. Nguyễn Cao Văn, Trần Thái Ninh. Giáo *trình lý thuyết xác suất và thống kê toán*. NXB ĐH Kinh tế quốc dân. Hà nội. 2008.
4. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà nội. 2009.
5. Гмурман В.Е. Теория вероятностей и математическая статистика. Москва. 1972.
6. Гнеденко Б. В. Теория вероятностей. Москва. 1969.
7. Baillageon G et Rainville J. Introduction à la Statistique appliquée. Québec. 1976
8. Giard V. Statistique appliquée à la gestion. Economica. Paris. 1992
9. Amir D Aczel. Complete Business Statistics. Irwin Boston 1993.
10. De groot M. Probability and Statistics. New york. 1989
11. OTT L.R. And introduction to statistical methods and data analysis. California. 1993
12. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons. New york. 1990

THƯ VIỆN HUBT

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

THƯ VIỆN
HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 2

BIẾN NGẪU NHIÊN VÀ QUY LUẬT PHÂN PHỐI XÁC SUẤT

1. MỞ ĐẦU

Ở chương 1 ta đã nghiên cứu các loại biến cố và phương pháp tính xác suất xảy ra của biến cố đó. Nó cho phép ta chuyển sang nghiên cứu khái niệm trung tâm của lý thuyết xác suất, đó là khái niệm về biến ngẫu nhiên vì mọi đại lượng mà ta phải nghiên cứu trong thực tế về thực chất đều là các biến ngẫu nhiên. Trong kinh tế và xã hội thì việc nắm rõ bản chất cũng như quy luật thay đổi của các biến ngẫu nhiên sẽ là cơ sở cho việc ra các quyết định quản lý và kinh doanh.

Mục 2 sẽ tiến hành định nghĩa và phân loại các biến ngẫu nhiên thường gặp trong thực tế. Mục 3 sẽ trình bày các phương pháp cơ bản để mô tả quy luật phân phối xác suất của biến ngẫu nhiên rời rạc và liên tục. Mục 4 sẽ nghiên cứu các tham số đặc trưng của biến ngẫu nhiên, cách tính toán cũng như ý nghĩa của chúng trong thực tế, đặc biệt là trong kinh tế và xã hội.

2. ĐỊNH NGHĨA VÀ PHÂN LOẠI BIẾN NGẪU NHIÊN

2.1. Định nghĩa

Một biến số được gọi là biến ngẫu nhiên nếu trong kết quả của phép thử nó sẽ nhận một và chỉ một trong các giá trị có thể có của nó tùy thuộc vào sự tác động của các nhân tố ngẫu nhiên.

Các biến ngẫu nhiên được ký hiệu là X, Y, Z hoặc $X_1, X_2..., X_n$; $Y_1, Y_2..., Y_m, ...$, còn các giá trị có thể có của chúng được ký hiệu là $x_1, x_2..., x_n, x, y_1, y_2..., y_m, y...$

Ta chú ý rằng sở dĩ biến X nào đó gọi là biến ngẫu nhiên vì trước khi tiến hành phép thử ta chưa có thể nói một cách chắc chắn nó sẽ nhận giá trị bằng bao

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

nhiêu, mà chỉ có thể dự đoán điều đó với một xác suất nhất định. Nói cách khác, việc X nhận một giá trị nào đó $(X = x_1)$ hoặc $(X = x_2), ..., (X = x_n)$ về thực chất là các biến cố ngẫu nhiên. Hơn nữa vì trong kết quả của phép thử biến X nhất định sẽ nhận một và chỉ một trong các giá trị có thể có của nó, do đó các biến cố $(X = x_1)$, $(X = x_2), ..., (X = x_n)$ tạo nên một nhóm đầy đủ các biến cố.

Thí dụ 1. Gieo một con xúc xắc. Gọi X là "Số chấm xuất hiện". X là biến ngẫu nhiên vì trong kết quả của phép thử nó sẽ nhận 1 trong 6 giá trị có thể có là 1, 2, 3, 4, 5, 6.

Thí dụ 2. Gọi X là "Số con trai trong 100 trẻ sắp được sinh ra tại một nhà hộ sinh". X cũng là một biến ngẫu nhiên.

Thí dụ 3. Gọi Y là "Khoảng cách từ điểm chạm của viên đạn đến tâm bia". Y là biến ngẫu nhiên.

2.2. Phân loại biến ngẫu nhiên

Biến ngẫu nhiên có thể là rời rạc hoặc liên tục.

Biến ngẫu nhiên gọi là rời rạc nếu các giá trị có thể có của nó lập nên một tập hợp hữu hạn hoặc đếm được.

Nói cách khác, biến ngẫu nhiên sẽ là rời rạc nếu ta có thể liệt kê được tất cả các giá trị có thể có của nó.

Thí dụ 4. Trong phép thử về gieo con xúc xắc, nếu ta gọi X là "Số điểm thu được" thì X là biến ngẫu nhiên rời rạc vì các giá trị có thể có của nó là một tập hợp hữu hạn.

Thí dụ 5. Gọi Y là "Số người vào mua hàng tại một siêu thị trong một ngày". Y là một biến ngẫu nhiên rời rạc vì các giá trị có thể có của nó lập nên một tập hợp đếm được $Y = 0, 1, 2, ...$

Thí dụ 6. Một phân xưởng có 5 máy hoạt động. Gọi X là "Số máy hỏng trong một ca". X là biến ngẫu nhiên rời rạc với các giá trị có thể có là $X = 0, 1, 2, 3, 4, 5$.

Biến ngẫu nhiên gọi là liên tục nếu các giá trị có thể có của nó lấp đầy một khoảng trên trục số.

Đối với biến ngẫu nhiên *liên tục* ta không thể liệt kê được tất cả các giá trị có thể có của nó.

Thí dụ 7. Phép thử là bắn một phát súng vào bia. Nếu gọi X là "Khoảng

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

cách từ điểm chạm của viên đạn đến tâm bia" thì X là biến ngẫu nhiên liên tục vì ta không thể kể ra được tất cả các giá trị có thể có của nó. Ta chỉ có thể nói rằng các giá trị có thể có của X nằm trong khoảng (a,b) nào đó.

Thí dụ 8. Gọi Y là "Sai số khi đo lường một đại lượng vật lý", Y là biến ngẫu nhiên liên tục.

Thí dụ 9. Gọi X là "Kích thước của chi tiết do một máy sản xuất ra", X là biến ngẫu nhiên liên tục.

Thí dụ 10. Gọi X là "Năng suất lúa vụ mùa của một tỉnh", X là biến ngẫu nhiên liên tục.

Có thể nói rằng gần như tất cả các đại lượng mà ta gặp trong thực tế đều là các biến ngẫu nhiên và chúng sẽ thuộc về một trong hai loại biến ngẫu nhiên đã kể ở trên.

3. QUY LUẬT PHÂN PHỐI XÁC SUẤT CỦA BIẾN NGẪU NHIÊN

Ta có thể nghĩ rằng chỉ cần xác định các giá trị có thể có của một biến ngẫu nhiên là đủ để xác định biến ngẫu nhiên ấy. Tuy nhiên điều này chưa đủ. Trong thực tế có những đại lượng rất khác nhau mà các giá trị có thể có của chúng lại giống nhau. Hơn nữa việc các biến ngẫu nhiên nhận một giá trị nào đó trong kết quả của phép thử chỉ là một biến cố ngẫu nhiên, do đó nếu chỉ mới biết được các giá trị có thể có của nó thì ta mới nắm được rất ít thông tin về biến ngẫu nhiên ấy. Vì vậy ta còn phải xác định các xác suất tương ứng với các giá trị có thể có của biến ngẫu nhiên để hoàn toàn xác định nó. Từ đó ta có định nghĩa sau đây.

3.1. Định nghĩa

Quy luật phân phối xác suất của biến ngẫu nhiên là sự tương ứng giữa các giá trị có thể có của nó và các xác suất tương ứng với các giá trị đó.

Trong thực tế người ta thường sử dụng ba phương pháp để mô tả quy luật phân phối xác suất của biến ngẫu nhiên là: bảng phân phối xác suất, hàm phân bố xác suất và hàm mật độ xác suất. Ta sẽ lần lượt nghiên cứu các phương pháp đó.

3.2. Bảng phân phối xác suất

Bảng phân phối xác suất chỉ dùng để mô tả quy luật phân phối xác suất của các biến ngẫu nhiên rời rạc.

Giả sử biến ngẫu nhiên rời rạc X có thể nhận một trong các giá trị có

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

thể có là $x_1, x_2, ..., x_n$ với các xác suất tương ứng là $p_1, p_2, ..., p_n$. Bảng phân phối xác suất của biến ngẫu nhiên rời rạc X có dạng:

X	x_1	x_2	...	x_i	...	x_n
P	p_1	p_2	...	p_i	...	p_n

Ta chú ý rằng để tạo nên một quy luật phân phối xác suất thì các xác suất p_i phải thoả mãn điều kiện:

$$\begin{cases} 0 \le p_i \le 1 \quad \forall i \\ \sum_{i=1}^{n} p_i = 1 \end{cases} \tag{2.1}$$

Điều kiện thứ nhất là hiển nhiên theo tính chất của xác suất, còn điều kiện thứ hai suy ra từ định nghĩa của biến ngẫu nhiên. Do các biến cố $(X = x_1), (X = x_2), ..., (X = x_n)$ tạo nên một nhóm đầy đủ các biến cố nên tổng các xác suất của chúng bằng một.

Thí dụ 1. Gieo một con xúc xắc. Gọi X là "Số chấm xuất hiện". Hãy tìm quy luật phân phối xác suất của X.

Giải. Vì X là biến ngẫu nhiên rời rạc với các giá trị có thể có $X = 1, 2, 3, 4, 5, 6$ với các xác suất tương ứng đều bằng 1/6, do đó bảng phân phối xác suất của X có dạng:

X	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Kiểm tra: $\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=1$

Thí dụ 2. Trong hộp có 10 sản phẩm trong đó có 6 chính phẩm. Lấy ngẫu nhiên 2 sản phẩm. Tìm quy luật phân phối xác suất của số chính phẩm được lấy ra.

Giải. Gọi Y là "Số chính phẩm được lấy ra trong 2 sản phẩm". Y là biến ngẫu nhiên rời rạc với các giá trị có thể có $Y = 0, 1, 2$.

Xác suất $P(Y = 0)$ chính là xác suất để trong 2 sản phẩm lấy ra không có chính phẩm nào (được 2 phế phẩm). Theo định nghĩa cổ điển về xác suất

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ta có:

$$P(Y=0)=\frac{C_4^2}{C_{10}^2}=\frac{6}{45}=\frac{2}{15}$$

Tương tự:

$$P(Y=1)=\frac{C_6^1.C_4^1}{C_{10}^2}=\frac{24}{45}=\frac{8}{15}$$

$$P(Y=2)=\frac{C_6^2}{C_{10}^2}=\frac{15}{45}=\frac{5}{15}$$

Như vậy quy luật phân phối xác suất của X có dạng:

X	0	1	2
P	$\frac{2}{15}$	$\frac{8}{15}$	$\frac{5}{15}$

Kiểm tra: $\frac{2}{15}+\frac{8}{15}+\frac{5}{15}=1$

Thí dụ 3. Xác suất để xạ thủ bắn trúng bia là 0,8. Xạ thủ được phát từng viên đạn để bắn cho đến khi trúng bia. Tìm quy luật phân phối xác suất của số viên đạn được phát.

Giải. Gọi X là "Số viên đạn xạ thủ được phát" X là biến ngẫu nhiên rời rạc với các giá trị có thể có $X=1,2,...,k...$

Ta tìm các xác suất tương ứng.

Xác suất $P(X=1)$ là xác suất để số viên đạn được phát bằng 1. Muốn xảy ra biến cố đó thì ngay phát đạn đầu tiên xạ thủ phải bắn trúng bia. Do đó:

$$P(X=1)=0,8$$

Xác suất $P(X=2)$ là xác suất để người ấy được phát 2 viên đạn. Muốn vậy phải xảy ra đồng thời hai biến cố: Phát thứ nhất bắn trượt và phát thứ hai bắn trúng. Theo định lý nhân xác suất ta có:

$$P(X=2)=0,2.0,8$$

Ta tìm xác suất tổng quát: $P(X=k)=(0,2)^{k-1}.0,8$

Như vậy bảng phân phối xác suất của X có dạng:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

X	1	2	...	k	...
P	0,8	$0,2.0,8$	...	$0,2^{k-1}.0,8$	...

Kiểm tra: Ta sẽ chứng tỏ rằng các xác suất vừa tìm được tạo nên một quy luật phân phối xác suất. Thật vậy, điều kiện thứ nhất hiển nhiên thỏa mãn. Còn đối với điều kiện thứ hai ta phải tìm tổng:

$$0,8+0,2.0,8+...+(0,2)^{k-1}.0,8+...$$

Song đó chính là tổng của một cấp số nhân lùi vô hạn với công bội $q=0,2$, do đó:

$$\sum_{k=1}^{\infty}(0,2)^{k-1}0,8=\frac{0,8}{1-0,2}=1$$

Hạn chế của bảng phân phối xác suất là nó chỉ mô tả được quy luật phân phối xác suất của biến ngẫu nhiên rời rạc mà thôi.

Bài tập

2.1. Một hộp có 3 quả cầu trắng và 2 quả cầu đen. Lấy ngẫu nhiên lần lượt từng quả cầu cho đến khi lấy được cầu trắng. Tìm quy luật phân phối xác suất của số cầu được lấy ra.

X	1	2	3
P	0,6	0,3	0,1

2.2. Xác suất để một người bắn trúng bia là 0,8. Người ấy được phát từng viên đạn để bắn cho đến khi trúng bia. Tìm quy luật phân phối xác suất của viên đạn bắn trượt.

ĐS: Gọi X là số viên đạn bắn trượt.

X	0	1	2	...	n ...
P	0,8	$0,2.0,8$	$(0,2)^2 0,8$	...	$(0,2)^n 0,8$ ...

2.3. Hai cầu thủ bóng rổ lần lượt ném bóng vào rổ cho đến khi có người ném trúng với xác suất ném trúng của từng người tương ứng là 0,3 và 0,4. Người thứ nhất ném trước.

a. Tìm quy luật phân phối xác suất của số lần ném rổ cho mỗi người.

b. Tìm quy luật phân phối xác suất của tổng số lần ném rổ của cả hai người.

ĐS: Gọi X_1 là số lần ném rổ của người thứ nhất

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vậy X_1 có quy luật phân phối xác suất như sau.

X_1	1	2	...	n	...
P	0,58	0,42.0,58	...	$(0,42)^{n-1}$ 0,58	...

Quy luật phân phối xác suất của số lần ném rổ của người thứ hai:

X_2	0	1	2	...	n	...
P	0,3	0,406	0,42.0,406	...	$(0,42)^{n-1}$ 0,406	...

Quy luật phân phối xác suất của cả hai người

X	1	2	...	2n - 1	2n	...
P	0,3	0,28	...	$(0,42)^{n-1}$.0,3	$(0,42)^{n-1}$ 0,28	

Độc giả có thể dễ dàng chứng minh rằng các xác suất trên tạo nên một quy luật phân phối xác suất.

3.3. Hàm phân bố xác suất

Khái niệm hàm phân bố xác suất áp dụng được đối với cả biến ngẫu nhiên rời rạc và liên tục. Giả sử X là biến ngẫu nhiên bất kỳ, x là một số thực nào đó. Xét biến cố "Biến ngẫu nhiên X nhận giá trị nhỏ hơn x", ký hiệu $(X < x)$. Hiển nhiên là x thay đổi thì xác suất $P(X < x)$ cũng thay đổi theo. Như vậy, xác suất này là một hàm số của x.

1. Định nghĩa. *Hàm phân bố xác suất của biến ngẫu nhiên X, ký hiệu là F(x), là xác suất để biến ngẫu nhiên X nhận giá trị nhỏ hơn x, với x là một số thực bất kỳ.*

$$F(x) = P(X < x) \tag{2.2}$$

Ta chú ý rằng đây chỉ là định nghĩa tổng quát của hàm phân bố xác suất. Đối với từng loại biến ngẫu nhiên hàm phân bố xác suất được tính theo những công thức riêng. Chẳng hạn nếu X là biến ngẫu nhiên rời rạc thì hàm phân bố xác suất được xác định bằng công thức:

$$F(x) = \sum_{x_i < x} P_i$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 4. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	1	3	4
P	0,1	0,5	0,4

Hãy tìm hàm phân bố của X và vẽ đồ thị.

Giải. Nếu $x \le 1$ biến cố (X<x) là biến cố không thể có, do đó:

$$F(x) = 0.$$

Nếu $1 < x \le 3$ biến cố (X<x) chỉ xảy ra khi (X=1), do đó:

$$F(x) = 0,1.$$

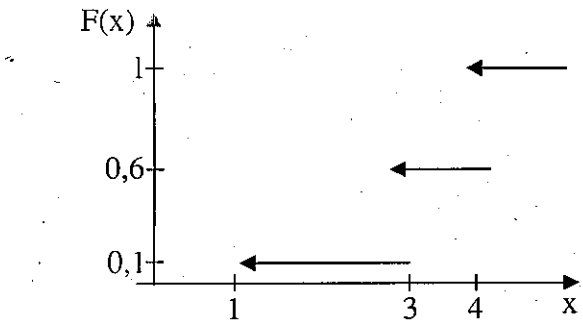
Nếu $3 < x \le 4$ biến cố (X<x) sẽ xảy ra hoặc khi (X=1) hoặc khi (X=3) do đó: $F(x) = 0,1+0,5 = 0,6$

Nếu $x > 4$ biến cố (X<x) sẽ xảy ra hoặc khi (X=1) hoặc khi (X=3) hoặc khi (X = 4), do đó: $F(x) = 0,1+0,5+0,4 = 1$

Vậy hàm phân bố xác suất của X có dạng:

$$F(x) = \begin{cases} 0 & \text{với } x \le 1 \\ 0,1 & \text{với } 1 < x \le 3 \\ 0,6 & \text{với } 3 < x \le 4 \\ 1 & \text{với } x > 4 \end{cases}$$

Đồ thị của hàm F(x) có dạng (hình 2.1).



Hình 2.1. **Đồ thị hàm** F(x)

Như vậy đồ thị của hàm phân bố xác suất của biến ngẫu nhiên rời rạc có dạng bậc thang với số điểm gián đoạn chính bằng số giá trị có thể có của X.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2. Các tính chất của hàm phân bố xác suất

Tính chất 1. Hàm phân bố xác suất luôn nhận giá trị trong đoạn [0, 1]:

$$0 \le F(x) \le 1 \tag{2.3}$$

Tính chất này trực tiếp suy ra từ định nghĩa của hàm phân bố xác suất, vì nó là một xác suất nên giá trị của nó luôn nằm trong đoạn [0, 1]:

Tính chất 2. Hàm phân bố xác suất là hàm không giảm, tức là với $x_2 > x_1$ *thì:*

$$F(x_2) \ge F(x_1)$$

Chứng minh. Giả sử $x_2 > x_1$. Xét biến cố $(X < x_2)$. Biến cố này có thể phân tích thành tổng của hai biến cố xung khắc là $(X < x_1)$ và $(x_1 \le X < x_2)$. Theo định lý cộng xác suất ta có:

$$P(X < x_2) = P(X < x_1) + P(x_1 \le X < x_2)$$

Từ đó $P(X < x_2) - P(X < x_1) = P(x_1 \le X < x_2)$

hay $F(x_2) - F(x_1) = P(x_1 \le X < x_2)$

Song vế phải là một xác suất, nó luôn không âm do đó ta có:

$$F(x_2) - F(x_1) \ge 0 \text{ từ đó } F(x_2) \ge F(x_1)$$

Từ tính chất thứ hai có thể suy ra một số hệ quả sau đây:

Hệ quả 1. Xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng [a, b) bằng hiệu số của hàm phân bố xác suất tại hai đầu khoảng đó:

$$P(a \le X < b) = F(b) - F(a) \tag{2.4}$$

Hệ quả này suy ra trực tiếp từ quá trình chứng minh tính chất.

Hệ quả 2. Xác suất để biến ngẫu nhiên liên tục X nhận một giá trị xác định bằng 0:

$$P(X = x) = 0 \tag{2.5}$$

Thật vậy, nếu ta đặt $a = x$ và $b = x + \Delta x$ ta có:

$$P(x \le X < x + \Delta x) = F(x + \Delta x) - F(x)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lấy giới hạn của cả hai vế khi $\Delta x \to 0$.

$$\lim_{\Delta x \to 0} P(x \le X < x + \Delta x) = \lim_{\Delta x \to \infty} F(x + \Delta x) - F(x)$$

Vì X là biến ngẫu nhiên liên tục, do đó tại điểm x hàm phân bố xác suất cũng liên tục. Vì vậy $\lim_{x \to \infty} F(x + \Delta x) = F(x)$

Từ đó: $P(X = x) = F(x) - F(x) = 0$.

Hệ quả 3. Đối với biến ngẫu nhiên liên tục X ta có các đẳng thức sau đây:

$$P(a \le x \le b) = P(a \le x < b) = P(a < x \le b)$$

$$= P(a < x < b) \qquad (2.6)$$

Chẳng hạn đẳng thức $P(a \le x < b) = P(a < x < b)$ có thể chứng minh như sau:

$$P(a \le x < b) = P(X = a) + P(a < X < b)$$

$$= P(a < X < b)$$

Như vậy việc xét xác suất để biến ngẫu nhiên liên tục X nhận một giá trị xác định là không có ý nghĩa, song việc tìm xác suất để nó nhận giá trị trong một khoảng, dù là rất nhỏ lại có ý nghĩa.

Tính chất 3. Ta có biểu thức giới hạn sau:

$$F(-\infty) = 0; F(+\infty) = 1 \qquad (2.7)$$

Thật vậy:

$$F(-\infty) = P(X < -\infty) = P(V) = 0$$

$$F(+\infty) = P(X < +\infty) = P(U) = 1$$

Từ tính chất trên có thể suy ra hệ quả sau:

Hệ quả. Nếu biến ngẫu nhiên X chỉ nhận giá trị trong đoạn [a,b] thì với $x \le a, F(x) = 0$ và với $x > b, F(x) = 1$.

Thật vậy, với $x \le a$ biến cố (X<x) là biến cố không thể có, do đó xác suất của nó bằng 0. Còn với x>b thì biến cố (X<x) là biến cố chắc chắn, do đó xác suất của nó bằng 1.

Chú ý rằng, nếu X là biến ngẫu nhiên rời rạc thì hàm phân bố xác suất chỉ liên tục về phía trái tại mỗi giá trị có thể có của nó, còn về phía phải thì nó bị gián đoạn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3. Ý nghĩa của hàm phân bố xác suất

Từ định nghĩa của hàm phân bố xác suất $F(x) = P(X < x)$ ta thấy hàm phân bố xác suất phản ánh mức độ tập trung xác suất ở về phía bên trái một số thực x nào đó. Như đã biết toàn bộ xác suất của biến ngẫu nhiên bằng một, do đó giá trị của hàm phân bố xác suất tại mỗi điểm x cho biết có bao nhiêu phần của một đơn vị xác suất phân bổ trong đoạn $(-\infty, x)$.

Thí dụ 5. Biến ngẫu nhiên X có hàm phân bố xác suất như sau:

$$F(x) = \begin{cases} 0 & \text{với } x \le -1 \\ \frac{3}{4}x + \frac{3}{4} & \text{với } -1 < x \le \frac{1}{3} \\ 1 & \text{với } x > \frac{1}{3} \end{cases}$$

Tìm xác suất để trong kết quả của phép thử, X nhận giá trị trong khoảng $[0, \frac{1}{3})$..

Theo tính chất của hàm phân bố xác suất:

$$P(0 \le X < \frac{1}{3}) = F(\frac{1}{3}) - F(0)$$

Vì trong khoảng $[0, \frac{1}{3})$ giá trị của hàm phân bố xác suất bằng:

$$F(x) = \frac{3}{4}x + \frac{3}{4}$$

Do đó: $F\left(\frac{1}{3}\right) - F(0) = \left[\frac{3}{4}.\frac{1}{3} + \frac{3}{4}\right] - \left[\frac{3}{4}.0 + \frac{3}{4}\right] = \frac{1}{4}$

Như vậy: $P\left(0 \le X < \frac{1}{3}\right) = \frac{1}{4}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

2.4. Một xí nghiệp có hai ô tô vận tải hoạt động. Xác suất trong ngày làm việc các ô tô bị hỏng tương ứng bằng 0,1 và 0,2. Gọi X là số ô tô bị hỏng trong thời gian làm việc.

a. Tìm quy luật phân phối xác suất của X.

b. Thiết lập hàm phân bố xác suất của X và vẽ đồ thị của nó.

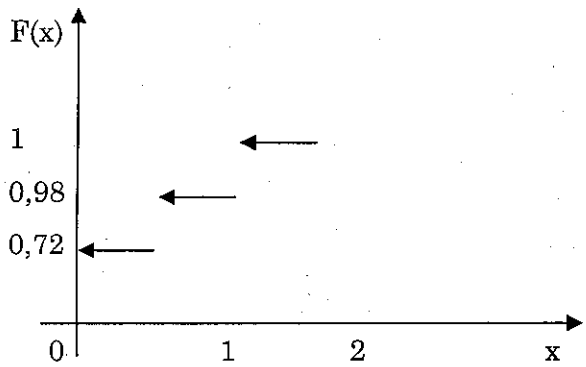
ĐS: a. Gọi X là số ô tô bị hỏng trong thời gian làm việc

X	0	1	2
P	0,72	0,26	0,02

b. Theo định nghĩa hàm phân bố xác suất ta có:

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ 0,72 & \text{với } 0 < x \le 1 \\ 0,98 & \text{với } 1 < x \le 2 \\ 1 & \text{với } x > 2 \end{cases}$$

Vậy đồ thị của F(x) có dạng (Hình vẽ)



2.5. Một thiết bị gồm 3 bộ phận hoạt động độc lập với nhau. Xác suất trong thời gian t các bộ phận bị hỏng tương ứng là 0,4; 0,2 và 0,3.

a. Tìm quy luật phân phối xác suất của số bộ phận bị hỏng X.

b. Thiết lập hàm phân bố xác suất của X.

c. Tính xác suất trong thời gian t có không quá hai bộ phận bị hỏng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ĐS: a.

X	0	1	2	3
P	0,336	0,452	0,188	0,024

b.

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ 0{,}336 & \text{với } 0 < x \le 1 \\ 0{,}788 & \text{với } 1 < x \le 2 \\ 0{,}976 & \text{với } 2 < x \le 3 \\ 1 & \text{với } x > 3 \end{cases}$$

c. $P(X \le 2) = 0{,}976$

2.6. Có hai lô sản phẩm.

Lô 1: Có 8 chính phẩm và 2 phế phẩm

Lô 2: Có 7 chính phẩm và 3 phế phẩm

Từ lô thứ nhất lấy ngẫu nhiên 2 sản phẩm bỏ sang lô thứ hai, sau đó từ lô thứ hai lấy ra 2 sản phẩm.

a. Tìm quy luật phân phối xác suất của số chính phẩm được lấy ra.

b. Xây dựng hàm phân bố xác suất của số chính phẩm được lấy ra và vẽ đồ thị.

ĐS: *a.*

X	0	1	2
P	0,06397	0,43872	0,49731

b.

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ 0{,}06397 & \text{với } 0 < x \le 1 \\ 0{,}50269 & \text{với } 1 < x \le 2 \\ 1 & \text{với } x > 2 \end{cases}$$

2.7. Gieo cùng một lúc hai con xúc xắc. Gọi X là tổng số chấm thu được.

a. Tìm bảng phân phối xác suất của X.

b. Tìm hàm phân bố xác suất của X

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.4. Hàm mật độ xác suất

Đối với biến ngẫu nhiên liên tục X có thể dùng hàm phân bố xác suất để mô tả quy luật phân phối xác suất của nó. Tuy nhiên phương pháp này cũng có hạn chế. Hàm phân bố xác suất không thể đặc trưng được xác suất để biến ngẫu nhiên liên tục X nhận một giá trị xác định. Vì thế, đối với các biến ngẫu nhiên liên tục, người ta thường dùng hàm mật độ xác suất để mô tả quy luật phân phối xác suất của nó.

1. Định nghĩa. *Hàm mật độ xác suất của biến ngẫu nhiên liên tục X (ký hiệu là f(x)) là đạo hàm bậc nhất của hàm phân bố xác suất của biến ngẫu nhiên đó.*

$$f(x) = F'(x) \tag{2.8}$$

Chú ý rằng, khái niệm hàm mật độ xác suất chỉ áp dụng được đối với các biến ngẫu nhiên liên tục mà không áp dụng được đối với biến ngẫu nhiên rời rạc vì muốn $F'(x)$ tồn tại thì tối thiểu $F(x)$ phải liên tục và khả vi, do đó X phải là biến ngẫu nhiên liên tục.

2. Các tính chất của hàm mật độ xác suất

Tính chất 1. Hàm mật độ xác suất luôn không âm:

$$f(x) \geq 0 \quad \forall x \tag{2.9}$$

Chứng minh. Hàm phân bố xác suất $F(x)$ là một hàm không giảm, do đó đạo hàm của nó $F'(x) = f(x)$ là một hàm không âm. Về mặt hình học điều đó có nghĩa là đồ thị của hàm $f(x)$ không nằm thấp hơn trục 0x.

Tính chất 2. Xác suất để biến ngẫu nhiên liên tục X nhận giá trị trong khoảng (a,b) bằng tích phân xác định của hàm mật độ xác suất trong khoảng đó:

$$P(a < X < b) = \int_a^b f(x)dx \tag{2.10}$$

Chứng minh. Theo tính chất của hàm phân bố xác suất ta có:

$$P(a \leq X < b) = F(b) - F(a)$$

Theo công thức Newton-Laibnitz

$$F(b) - F(a) = \int_a^b F'(x)dx = \int_a^b f(x)dx$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Như vậy: $P(a \le X < b) = \int_a^b f(x)dx$

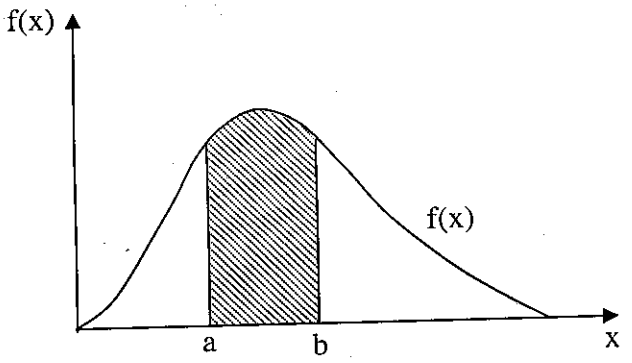
Song vì X là biến ngẫu nhiên liên tục nên

$$P(a \le X < b) = P(a < X < b)$$

Từ đó ta có $P(a < X < b) = \int_a^b f(x)dx$

Về mặt hình học, kết quả trên có thể minh hoạ như sau: Xác suất để biến ngẫu nhiên liên tục X nhận giá trị trong khoảng (a,b) bằng diện tích của hình giới hạn bởi trục 0x, đường cong f(x) và các đường thẳng $x = a$ và $x = b$

(xem hình 2.2).



***Hình 2.2.* Đồ thị hàm f(x)**

Tính chất 3. Hàm phân bố xác suất F(x) của biến ngẫu nhiên liên tục X bằng tích phân suy rộng của hàm mật độ xác suất trong khoảng $(-\infty, x)$ *:*

$$F(x) = \int_{-\infty}^{x} f(x)dx \qquad (2.11)$$

Chứng minh. Theo định nghĩa của hàm phân bố xác suất, ta có:

$$F(x) = P(X < x) = P(-\infty < X < x)$$

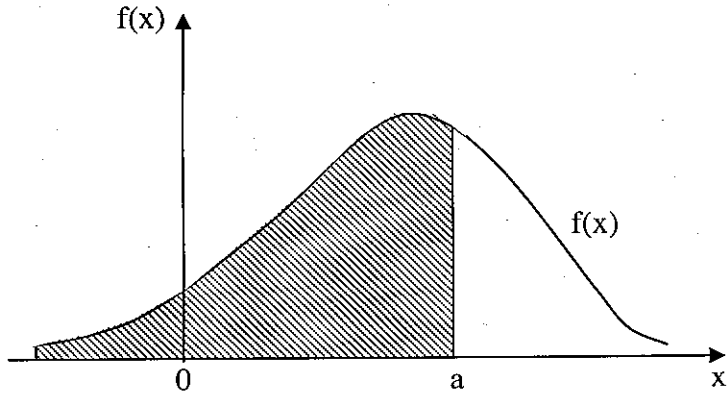
Theo tính chất 2, đặt $a = -\infty$ và $b = x$, ta có:

$$P(-\infty < X < x) = \int_{-\infty}^{x} f(x)dx$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Công thức trên cho phép tìm hàm phân bố xác suất của biến ngẫu nhiên liên tục khi đã biết hàm mật độ xác suất của nó.

Về mặt hình học, công thức trên cho thấy giá trị của hàm phân bố xác suất F(x) tại điểm a bằng diện tích giới hạn bởi trục 0x, đường cong f(x) và đường thẳng x=a (Xem hình 2.3)



***Hình 2.3.* Giá trị của hàm F(x)**

Tính chất 4. Tích phân suy rộng trong khoảng $(-\infty,+\infty)$ *của hàm mật độ xác suất bằng 1.*

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \qquad (2.12)$$

Nếu theo tính chất 2 ta đặt $a=-\infty$ và $b=+\infty$, ta có:

$$P(-\infty < x < +\infty) = \int_{-\infty}^{+\infty} f(x)dx$$

Song biến cố $(-\infty < X < +\infty)$ là biến cố chắc chắn, do đó:

$$\int_{-\infty}^{+\infty} f(x)dx = P(U) = 1$$

Về mặt hình học, điều đó có nghĩa là toàn bộ diện tích giới hạn bởi đường cong f(x) và trục 0x bằng 1.

Ta chú ý rằng để một hàm số f(x) có thể là hàm mật độ xác suất của một biến ngẫu nhiên liên tục nào đó thì nó phải thoả mãn hai tính chất cơ bản là tính chất 1 và tính chất 4, tức là:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$\begin{cases} f(x) \geq 0 \quad \forall x \\ \int\limits_{-\infty}^{+\infty} f(x)dx = 1 \end{cases} \tag{2.13}$$

Thí dụ 6. Hàm phân bố xác suất của biến ngẫu nhiên liên tục X có dạng:

$$F(x) = \begin{cases} 0 & \text{với} \quad x \leq 0 \\ ax^2 & \text{với} \quad 0 < x \leq 1 \\ 1 & \text{với} \quad x > 1 \end{cases}$$

a) Tìm hệ số a;

b) Tìm hàm mật độ xác suất f(x);

c) Tìm xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng (0,25; 0,75).

Giải. a) Vì hàm phân bố xác suất F(x) là liên tục, do đó tại $x = 1.ax^2 = 1$, từ đó a=1.

b) Theo định nghĩa của hàm mật độ xác suất:

$$f(x) = F'(x) = \begin{cases} 0 & \text{với} \quad x \leq 0 \\ 2x & \text{với} \quad 0 < x \leq 1 \\ 0 & \text{với} \quad x > 1 \end{cases}$$

d) Theo tính chất của hàm phân bố xác suất

$$P(0,25 < X < 0,75) = F(0,75) - F(0,25)$$

$$= (0,75)^2 - (0,25)^2 = 0,5$$

Thí dụ 7. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} a\cos x & \text{với } x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \\ 0 & \text{với } x \notin \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \end{cases}$$

a) Tìm hệ số a;

b) Tìm hàm phân bố xác suất F(x);

c) Tìm xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng $(0, \pi/4)$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. a) Theo tính chất của hàm mật độ xác suất:

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\pi/2}^{\pi/2} a\cos x dx = 2a = 1$$

Từ đó: $a = \frac{1}{2}$

b) Để tìm hàm phân bố xác suất, ta sử dụng tính chất của hàm mật độ xác suất:

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Với $x \le -\frac{\pi}{2}$: $F(x) = \int_{-\infty}^{x} 0dx = 0$

Với $-\frac{\pi}{2} < x \le \frac{\pi}{2}$

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{-\pi/2} 0dx + \int_{-\pi/2}^{x} \frac{1}{2}\cos x dx$$

$$= \frac{1}{2}(\sin x + 1)$$

Với $x > \frac{\pi}{2}$

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{-\pi/2} 0dx + \int_{-\pi/2}^{\pi/2} \frac{1}{2}\cos x dx + \int_{\pi/2}^{x} 0dx = 1$$

Vậy hàm phân bố xác suất của X có dạng:

$$F(x) = \begin{cases} 0 & \text{với } x \le -\frac{\pi}{2} \\ \frac{1}{2}(\sin x + 1) & \text{với } -\frac{\pi}{2} < x \le \frac{\pi}{2} \\ 1 & \text{với } x > \frac{\pi}{2} \end{cases}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

c) Theo tính chất của hàm phân bố xác suất:

$$P\left(0 < x < \frac{\pi}{4}\right) = F\left(\frac{\pi}{4}\right) - F(0) =$$

$$= \frac{1}{2}\left(\sin\frac{\pi}{4} + 1\right) - \frac{1}{2}(\sin 0 + 1) = \frac{\sqrt{2}}{4}$$

Thí dụ 8. Hàm mật độ xác suất của biến ngẫu nhiên X có dạng:

$$f(x) = \frac{1}{\pi(1+x^2)} \quad \forall x$$

Tìm xác suất để tiến hành 3 phép thử độc lập có 2 lần X nhận giá trị trong khoảng $(-1;+1)$.

Giải. Trước hết ta tìm xác suất để tiến hành một phép thử biến ngẫu nhiên X nhận giá trị trong khoảng $(-1;+1)$.

Theo tính chất của hàm mật độ xác suất:

$$P(-1 < X < 1) = \int_{-1}^{1} \frac{dx}{\pi(1+x^2)} = \frac{1}{\pi}\text{arctgx}\Big|_{-1}^{1} = \frac{1}{2}$$

Theo điều kiện của bài toán, ta tiến hành ba phép thử độc lập, trong mỗi phép thử chỉ có hai khả năng: hoặc X nhận giá trị trong khoảng $(-1,1)$ (biến cố A), hoặc không nhận giá trị trong khoảng đó (biến cố $\overline{A}$). Xác suất để trong mỗi phép thử biến ngẫu nhiên X nhận giá trị trong khoảng $(-1,1)$, như đã xác định ở trên, đều bằng 0,5. Do đó theo công thức Bernoulli, xác suất để trong 3 phép thử có 2 lần X nhận giá trị trong khoảng đó bằng:

$$P_3(2) = C_3^2(0,5)^2(1-0,5) = 0,375$$

Thí dụ 9. Tuổi thọ của một loại sản phẩm là biến ngẫu nhiên liên tục có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \dfrac{a}{x^2} & \text{với } x \geq 400 \text{(giờ)} \\ 0 & \text{với } x < 400 \text{(giờ)} \end{cases}$$

a) Tìm a.

b) Tìm xác suất để lấy ngẫu nhiên 1 sản phẩm thì tuổi thọ của nó kéo dài ít nhất là 600 giờ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. a) Theo điều kiện của hàm mật độ xác suất

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{400} 0dx + a\int_{400}^{+\infty} \frac{dx}{x^2} = -a.\frac{1}{x}\Big|_{400}^{+\infty} = \frac{a}{400}$$

$$\Rightarrow a = 400$$

$$\text{b) } P(X > 600) = \int_{600}^{+\infty} f(x)dx = 400\int_{600}^{+\infty} \frac{dx}{x^2} = -400.\frac{1}{x}\Big|_{600}^{+\infty} = \frac{2}{3}$$

3. Ý nghĩa của hàm mật độ xác suất

Hàm mật độ xác suất của biến ngẫu nhiên X tại mỗi điểm x cho biết mức độ tập trung xác suất tại điểm đó. Thật vậy, theo định nghĩa của hàm mật độ xác suất, ta có:

$$f(x) = F'(x) = \lim_{\Delta x \to \infty} \frac{F(x+\Delta x) - F(x)}{\Delta x} =$$

$$= \lim_{\Delta x \to \infty} \frac{P(x \le X < x + \Delta x)}{\Delta x}$$

Như vậy hàm mật độ xác suất tại điểm x chính là giới hạn của xác suất để biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(x; x + \Delta x)$ chia cho độ dài của khoảng đó (khi $\Delta x \to 0$), tức là giới hạn của mật độ xác suất trung bình trên đoạn $(x; x + \Delta x)$ khi $\Delta x \to 0$.

Bài tập

2.8. Cho X là một biến ngẫu nhiên liên tục trong khoảng (- ∞; +∞) với hàm mật độ xác suất là f(x). Hãy tính giá trị của $\int_{-\infty}^{1} f(x)dx$ biết rằng

$P(X \ge 1) = 0{,}3$.

2.9. Cho hàm số

$$f(x) = \begin{cases} x^2/9 & \text{với } x \in (0, 3) \\ 0 & \text{với } x \notin (0, 3) \end{cases}$$

a. Hàm số trên có phải là hàm mật độ xác suất không.

b. Nếu có thì tìm xác suất để trong 3 phép thử độc lập có ít nhất một lần X nhận giá trị trong khoảng (1, 2).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.10. Cho hàm phân bố xác suất của biến ngẫu nhiên liên tục

$$F(x) = c + b.\text{arctg}\ \frac{x}{a} \quad \text{với } (-\infty < x < +\infty); \qquad a > 0$$

a. Tìm các hệ số c và b

b. Tìm hàm mật độ xác suất f(x)

ĐS: a. $c = \frac{1}{2}$; $b = \frac{1}{\pi}$

b. $f(x) = \frac{a}{\pi(x^2 + a^2)}$

2.11. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất.

$$f(x) = \begin{cases} C \sin 2x & \text{với } x \in (0, \pi/2) \\ 0 & \text{với } x \notin (0, \pi/2) \end{cases}$$

Tìm hệ số C.

ĐS: C = 1

2.12. Biến ngẫu nhiên liên tục X có hàm phân bố xác suất như sau:

$$F(x) = \begin{cases} 0 & \text{với } x \le 2 \\ (1/2)x - 1 & \text{với } 2 < x \le 4 \\ 1 & \text{với } x > 4 \end{cases}$$

Tìm xác suất để trong kết quả của phép thử X nhận giá trị:

a. Nhỏ hơn 3.

b. Trong khoảng [2,3)

ĐS: $P_a = 0,5$; $P_b = 0,5$

2.13. Hàm phân bố xác suất của biến ngẫu nhiên liên tục X có dạng

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ \sin 2x & \text{với } 0 < x \le \pi/4 \\ 1 & \text{với } x > \pi/4 \end{cases}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tìm hàm mật độ xác suất f(x).

ĐS: $f(x) = \begin{cases} 2\cos 2x & \text{với } x \in (0, \pi/4) \\ 0 & \text{với } x \notin (0, \pi/4) \end{cases}$

2.14. Cho X là biến ngẫu nhiên liên tục có hàm phân bố xác suất sau đây:

$$F(x) = \frac{1}{2} + \frac{1}{\pi} arctgx$$

a. Tìm P (0 < x < 1)

b. Tìm hàm mật độ xác suất của X

ĐS: a. $P = \frac{1}{4}$ b. $f(x) = \frac{1}{\pi(1 + x^2)}$

2.15. Biến ngẫu nhiên X có hàm phân bố xác suất như sau:

$$F(x) = \begin{cases} 1 - e^{-x/\theta} & \text{với } x > 0; \theta > 0 \\ 0 & \text{với } \le 0 \end{cases}$$

a. Tìm hàm mật độ xác suất của X.

b. Tìm P (0 < x < θ)

ĐS: a. $f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & \text{với } x > 0 \\ 0 & \text{Với } x \le 0 \end{cases}$

b. $P(0 < X < \theta) = \frac{e - 1}{e}$

2.16. Biến ngẫu nhiên X liên tục trên toàn trục số và có hàm phân bố xác suất

$$F(x) = \frac{1}{2} + \frac{1}{\pi} arctgx \frac{x}{2}$$

Tìm giá trị có thể có x_1 thoả mãn điều kiện $P(X > x_1) = 1/4$

ĐS: $x_1 = 2$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.17. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \frac{2}{\pi}\cos^2 x & \text{với } x \in \left[-\frac{\pi}{2}; \frac{\pi}{2}\right] \\ 0 & \text{với } x \notin \left[-\frac{\pi}{2}; \frac{\pi}{2}\right] \end{cases}$$

Tìm xác suất để trong 3 phép thử độc lập có 2 lần X nhận giá trị trong khoảng $(0, \pi/4)$

ĐS: 0,296

2.18. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất.

$$f(x) = \frac{k}{e^x + e^{-x}} \quad (-\infty < x < +\infty)$$

Tìm k

ĐS: $k = \frac{2}{\pi}$

4. CÁC THAM SỐ ĐẶC TRƯNG CỦA BIẾN NGẪU NHIÊN

Như ta đã thấy ở trên, quy luật phân phối xác suất của biến ngẫu nhiên (dưới dạng bảng phân phối xác suất, hàm phân bố xác suất hay hàm mật độ xác suất) hoàn toàn xác định biến ngẫu nhiên.

Như vậy, khi ta đã xác định được quy luật phân phối xác suất của một biến ngẫu nhiên thì ta đã nắm được toàn bộ thông tin về biến ngẫu nhiên đó. Tuy nhiên, trong thực tế ta không chỉ cần đến những thông tin đó mà còn phải quan tâm đến những thông tin cô đọng phản ánh tổng hợp những đặc trưng quan trọng nhất của biến ngẫu nhiên được nghiên cứu. Những thông tin cô đọng phản ánh từng phần về biến ngẫu nhiên được gọi là *các tham số đặc trưng*.

Các tham số đặc trưng của biến ngẫu nhiên được chia thành ba loại như sau:

- Các tham số đặc trưng cho xu hướng trung tâm của biến ngẫu nhiên như: kỳ vọng toán, trung vị, mốt v.v...

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Các tham số đặc trưng cho độ phân tán của biến ngẫu nhiên như: phương sai, độ lệch chuẩn, hệ số biến thiên v.v...

- Các tham số đặc trưng cho dạng phân phối xác suất.

Ở phần này chỉ hạn chế ở việc xem xét một số tham số đặc trưng quan trọng nhất. Sau này trong một số trường hợp cụ thể các tham số khác cũng sẽ được đề cập.

4.1. Kỳ vọng toán

Kỳ vọng toán có những định nghĩa riêng đối với biến ngẫu nhiên rời rạc và biến ngẫu nhiên liên tục.

1.Định nghĩa. *Giả sử biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có* $x_1, x_2, ..., x_n$ *với các xác suất tương ứng* $p_1, p_2, ..., p_n$. *Kỳ vọng toán của biến ngẫu nhiên rời rạc X, ký hiệu E(X) là tổng các tích giữa các giá trị có thể có của biến ngẫu nhiên với các xác suất tương ứng:*

$$E(X) = \sum_{i=1}^{n} x_i p_i \qquad (2.14)$$

Nếu X là biến ngẫu nhiên liên tục với hàm mật độ xác suất f(x) thì kỳ vọng toán E(X) được xác định bằng biểu thức:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \qquad (2.15)$$

Thí dụ 1. Tìm kỳ vọng toán của biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	1	3	4
P	0,1	0,5	0,4

Giải. Theo định nghĩa kỳ vọng toán của biến ngẫu nhiên rời rạc ta có:

$$E(X) = 1.0,1 + 3.0,5 + 4.0,4 = 3,2$$

Thí dụ 2. Tìm kỳ vọng toán của biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \frac{3}{4}(x^2 + 2x) & \text{với } x \in (0.1) \\ 0 & \text{với } x \notin (0.1) \end{cases}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Theo định nghĩa kỳ vọng toán của biến ngẫu nhiên liên tục ta có:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \frac{3}{4}\int_0^1 x(x^2+2x)dx =$$

$$= \frac{3}{4}\int_0^1 (x^3+2x^2)dx = \frac{3}{4}\left(\frac{x^4}{4}+\frac{2x^3}{3}\right)\Big|_0^1 = \frac{11}{16}$$

Ta chú ý rằng kỳ vọng toán của biến ngẫu nhiên là một số xác định.

2. Các tính chất của kỳ vọng toán

Sau đây sẽ phát biểu và chứng minh một số tính chất cơ bản của kỳ vọng toán. Các tính chất được chứng minh đối với các biến ngẫu nhiên rời rạc. Việc chứng minh trong trường hợp các biến ngẫu nhiên là liên tục cũng tiến hành tương tự.

Tính chất 1. Kỳ vọng toán của một hằng số bằng chính hằng số đó. Như vậy, nếu C là hằng số thì E(C)=C.

Thật vậy, có thể coi C như một biến ngẫu nhiên rời rạc đặc biệt, với một giá trị có thể có bằng C và xác suất tương ứng bằng 1. Lúc đó theo định nghĩa của kỳ vọng toán ta có:

$$E(C) = C.1 = C$$

Tính chất 2. Kỳ vọng toán của tích giữa một hằng số và một biến ngẫu nhiên bằng tích giữa hằng số đó và kỳ vọng toán của biến ngẫu nhiên ấy.

$$E(CX) = C.E(X) \qquad (2.16)$$

Thật vậy, giả sử biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

Lúc đó tích CX sẽ là một biến ngẫu nhiên rời rạc mà các giá trị có thể có của nó bằng tích giữa hằng số C và các giá trị có thể có của X. Mặt khác, do đó là các biến cố tương đương nhau (có khả năng xảy ra như nhau) nên các xác suất tương ứng bằng nhau. Vậy bảng phân phối xác suất của biến ngẫu nhiên CX có dạng:

CX	Cx_1	Cx_2	...	Cx_n
P	p_1	p_2	...	p_n

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lúc đó theo định nghĩa kỳ vọng toán ta có:

$$E(CX) = Cx_1p_1 + Cx_2p_2 + ... + Cx_np_n$$

$$= C(x_1p_1 + x_2p_2 + ... + x_np_n) = C.E(X)$$

Trước khi phát biểu tính chất tiếp theo ta xét khái niệm độc lập của các biến ngẫu nhiên.

Hai biến ngẫu nhiên gọi là *độc lập* với nhau nếu quy luật phân phối xác suất của biến ngẫu nhiên này không phụ thuộc gì vào việc biến ngẫu nhiên kia nhận giá trị bằng bao nhiêu. Tương tự các biến ngẫu nhiên gọi là độc lập lẫn nhau, nếu các quy luật phân phối xác suất của một số bất kỳ các biến ngẫu nhiên nào đó không phụ thuộc vào việc các biến ngẫu nhiên còn lại nhận giá trị bằng bao nhiêu.

Tổng của hai biến ngẫu nhiên X và Y là biến ngẫu nhiên $X + Y$ mà các giá trị có thể của nó là tổng của mỗi giá trị có thể có của X và mỗi giá trị có thể có của Y. Khi X và Y độc lập nhau thì các xác suất tương ứng sẽ bằng tích các xác suất thành phần. Còn khi X và Y phụ thuộc nhau thì các xác suất tương ứng sẽ bằng xác suất của thành phần này nhân với xác xuất có điều kiện của thành phần kia.

Tính chất sau đây đúng với cả các biến ngẫu nhiên độc lập và phụ thuộc.

Tính chất 3. Kỳ vọng toán của tổng hai biến ngẫu nhiên bằng tổng các kỳ vọng toán thành phần.

$$E(X + Y) = E(X) + E(Y) \qquad (2.17)$$

Thật vậy giả sử các biến ngẫu nhiên rời rạc X và Y có các quy luật phân phối xác suất như sau:

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

Y	y_1	y_1	...	y_m
P	q_1	q_2	...	q_m

Lúc đó ta có quy luật phân phối xác suất của tổng $X + Y$ như sau:

$X + Y$	$x_1 + y_1$	$x_1 + y_2$	...	$x_n + y_m$
P	p_{11}	p_{12}	...	p_{nm}

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trong đó ta ký hiệu p_{ij} là xác suất để tổng $X+Y$ nhận giá trị bằng $x_i + y_j$.

Theo định nghĩa kỳ vọng toán ta có:

$$E(X+Y) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i + y_j)p_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m}x_i p_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m}y_j p_{ij}$$

$$= \sum_{i=1}^{n}x_i\sum_{j=1}^{m}p_{ij} + \sum_{j=1}^{m}y_j\sum_{i=1}^{n}p_{ij}$$

Ta sẽ chứng minh rằng $\sum_{j=1}^{m}p_{ij} = p_i$

Thật vậy, biến cố $X = x_i$ sẽ xảy ra khi tổng $X+Y$ nhận giá trị $x_i + y_i$ hoặc $x_i + y_2 ...$ hoặc $x_i + y_m$. Do đó theo định lý cộng xác suất:

$$P(X = x_i) = P[(X+Y) = (x_i + y_1)] + ... + P[(X+Y) = (x_i + y_m)] \text{ hay}$$

$$p_i = p_{i1} + p_{i2} + ... + p_{im} = \sum_{j=1}^{m}p_{ij} = p_i$$

Tương tự như vậy có thể chứng minh được rằng:

$$\sum_{i=1}^{n}P_{ij} = q_j$$

Từ đó ta có:

$$E(X+Y) = \sum_{i=1}^{n}x_i p_i + \sum_{j=1}^{m}y_j q_j = E(X) + E(Y)$$

Bằng phương pháp quy nạp toán học ta có thể chứng minh được hệ quả sau đây.

Hệ quả. Kỳ vọng toán của tổng n biến ngẫu nhiên $X_1, X_2 ... X_n$ bằng tổng các kỳ vọng toán thành phần:

$$E\left(\sum_{i=1}^{n}X_i\right) = \sum_{i=1}^{n}E(X_i) \qquad (2.18)$$

Tích của hai biến ngẫu nhiên độc lập X và Y là biến ngẫu nhiên XY mà

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

các giá trị có thể có của nó là tích giữa mỗi giá trị có thể có của X và mỗi giá trị có thể có của Y. Các xác suất tương ứng là tích của các xác suất thành phần.

Tính chất 4. Kỳ vọng toán của tích hai biến ngẫu nhiên độc lập bằng tích các kỳ vọng toán thành phần.

$$E(X.Y) = E(X).E(Y) \tag{2.19}$$

Thật vậy, giả sử các biến ngẫu nhiên X và Y độc lập, có các quy luật phân phối xác suất như sau:

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

Y	y_1	y_1	...	y_m
P	q_1	q_2	...	q_m

Lúc đó tích X.Y có quy luật phân phối xác suất như sau:

X.Y	x_1y_1	x_1y_2	...	x_ny_m
P	p_1q_1	p_1q_2	...	p_nq_m

Ta có:

$$E(X.Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j p_i q_j = \sum_{i=1}^{n} x_i p_i \sum_{j=1}^{m} y_j q_j = E(X).E(Y)$$

Bằng phương pháp quy nạp toán học có thể chứng minh được hệ quả sau đây:

Hệ quả. Kỳ vọng toán của tích n biến ngẫu nhiên $X_1, X_2, ..., X_n$ độc lập lẫn nhau bằng tích các kỳ vọng toán thành phần.

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i) \tag{2.20}$$

3. Bản chất và ý nghĩa của kỳ vọng toán

Giả sử đối với biến ngẫu nhiên X tiến hành n phép thử trong đó có n_1 lần X nhận giá trị x_1, n_2 lần X nhận giá trị $x_2, ..., n_k$ lần X nhận giá trị x_k $\left(\sum_{i=1}^{k} n_i = n\right)$. Giá trị trung bình của biến ngẫu nhiên X trong n phép thử này là:

$$\bar{x} = \frac{x_1n_1 + x_2n_2 + ... + x_kn_k}{n} = x_1\frac{n_1}{n} + x_2\frac{n_2}{n} + ... + x_k\frac{n_k}{n}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ta chú ý rằng $\frac{n_1}{n}, \frac{n_2}{n}, ..., \frac{n_k}{n}$ chính là tần suất xuất hiện các giá trị $x_1, x_2, ..., x_k$ trong n phép thử trên, do đó:

$$\bar{x} = x_1 f_1 + x_2 f_2 + ... + x_k f_k$$

Theo định nghĩa thống kê về xác suất khi $n \to \infty$ các tần suất sẽ hội tụ theo xác suất về các xác suất tương ứng, do đó với n đủ lớn ta có thể viết:

$$\bar{x} \approx x_1 p_i + x_2 p_2 + ... + x_k p_k = E(X)$$

Như vậy ta thu được kết quả:

$$E(X) \approx \bar{x} \qquad (2.21)$$

Vậy kỳ vọng toán của biến ngẫu nhiên gần bằng trung bình số học của các giá trị quan sát của biến ngẫu nhiên. Nó phản ánh giá trị trung tâm của phân phối xác suất của biến ngẫu nhiên.

Thí dụ 3. Gieo con xúc xắc n lần. Tìm kỳ vọng toán của tổng số chấm thu được.

Giải. Gọi $X_i (i = \overline{1,n})$ là số chấm thu được ở lần gieo thứ i và gọi X là tổng số chấm thu được trong n lần tung.

Như vậy $X = \sum_{i=1}^{n} X_i$. Theo tính chất của kỳ vọng toán

$$E(X) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$$

Mỗi biến ngẫu nhiên X_i đều có bảng phân phối xác suất như sau:

X_i	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Do đó:

$$E(X_i) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2} \quad \forall_i$$

Do đó:

$$E(X) = \frac{7}{2}n$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 4. Thời gian xếp hàng chờ mua hàng của khách hàng là biến ngẫu nhiên liên tục T có hàm mật độ xác suất như sau (đơn vị: phút):

$$f(t)=\begin{cases}\dfrac{4}{81}t^3 & \text{với } t\in(0,3)\\ 0 & \text{với } t\notin(0,3)\end{cases}$$

Tìm thời gian xếp hàng trung bình của khách hàng.

Giải. Thời gian xếp hàng trung bình chính là kỳ vọng toán. Theo định nghĩa ta có:

$$E(T)=\int_{+\infty}^{-\infty} t.f(t).dt=\frac{4}{81}\int_0^3 t^4dt=\frac{4}{405}t^5\Big|_0^3=2,4\,\text{phút}$$

Thí dụ 5. Một dự án xây dựng được viện thiết kế C soạn thảo cho cả 2 bên A và B xét duyệt một cách độc lập. Xác suất để A và B chấp nhận dự án khi xét duyệt là 0,7 và 0,8. Nếu chấp nhận dự án thì A phải trả cho C là 4 triệu đồng, còn ngược lại thì phải trả 1 triệu. Với B, nếu chấp nhận dự án thì phải trả cho C là 10 triệu đồng, ngược lại phải trả 3 triệu. Chi phí cho thiết kế là 10 triệu và thuế 10% doanh thu. Hỏi C có nên nhận thiết kế hay không?

Giải. Để quyết định xem có nên nhận thiết kế hay không thì C phải tính số lãi kỳ vọng mà C có thể nhận được. Nếu gọi X là số lãi mà C có thể nhận được sau khi trừ mọi chi phí thì X có bảng phân phối xác suất như sau:

X	−6,4	−3,7	- 0,1	2,6
P	0,06	0,14	0,24	0,56

Từ đó $E(X)=0,53>0$. Vậy C vẫn có thể nhận thiết kế.

Thí dụ 6. Xác suất để một nhà máy sản xuất ra phế phẩm bằng p. Máy sẽ được sửa chữa ngay sau khi làm ra phế phẩm.

Tìm số sản phẩm trung bình được sản xuất giữa hai lần sửa chữa.

Giải: Gọi X là số sản phẩm được sản xuất ra giữa hai lần sửa chữa. X có bảng phân phối xác suất như sau:

X	1	2	3	...	k	...
P	p	qp	q^2p	...	$q^{k-1}p$	...

Trong đó q là xác suất để sản phẩm sản xuất ra là chính phẩm, $q=1-p$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Số sản phẩm trung bình được sản xuất ra chính là kỳ vọng toán E(X) như đã phân tích trong phần bản chất của kỳ vọng toán.

Do đó:

$$E(X) = \sum_{k=1}^{\infty} kq^{k-1}p = p\sum_{k-1}^{\infty} kq^{k-1}$$

Song tổng $\sum_{k=1}^{\infty} kq^{k-1}$ có thể biểu diễn dưới dạng.

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{d}{dq}\sum_{k=0}^{\infty} q^k$$

Mà tổng $\sum_{k=0}^{\infty} q^k$ là tổng của cấp số nhân lùi vô hạn, do đó:

$$\frac{d}{dq}\sum_{k=0}^{\infty} q^k = \frac{d}{dq}\left(\frac{1}{1-q}\right) = \frac{1}{(1-q)^2} = \frac{1}{p^2}$$

Do đó: $E(X) = p\frac{1}{p^2} = \frac{1}{p}$

4. Ứng dụng thực tế của kỳ vọng toán

Khái niệm kỳ vọng toán lúc đầu xuất hiện trong các trò chơi may rủi để tính giá trị mà người chơi mong đợi sẽ nhận được. Hiện nay khái niệm này được áp dụng rộng rãi trong nhiều lĩnh vực kinh doanh và quản lý như một tiêu chuẩn để ra quyết định trong tình huống cần lựa chọn giữa nhiều chiến lược khác nhau. Tiêu chuẩn này thường được biểu diễn dưới dạng lợi nhuận kỳ vọng hay doanh số kỳ vọng để làm căn cứ lựa chọn chiến lược kinh doanh. Ta sẽ xét một bài toán như vậy làm ví dụ.

Giả sử một cửa hàng sách dự định nhập vào một số cuốn niên giám thống kê. Nhu cầu hàng năm về loại sách này được cho trong bảng phân phối xác suất sau đây:

Nhu cầu j (cuốn)	20	21	22	23	24	25
Xác suất P_j	0,3	0,25	0,18	0,14	0,1	0,03

Cửa hàng này mua vào với giá 7 USD/cuốn và bán ra với giá 10 USD/cuốn, song đến cuối năm thì phải bán hạ giá còn 4 USD/cuốn trước khi

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

niên giám thống kê của năm tới được xuất bản. Cửa hàng muốn xác định số lượng nhập vào sao cho lợi nhuận kỳ vọng là lớn nhất.

Bài toán được giải như sau: Gọi i là số lượng sách cần nhập và j là nhu cầu. Hiển nhiên là lợi nhuận sẽ phụ thuộc vào số lượng sách cần nhập và nhu cầu thực tế về loại sách đó. Từ đó có thể xây dựng một bảng liệt kê những kết quả khác nhau có thể thu được từ những chiến lược nhập hàng khác nhau. Ta gọi nó là *Bảng lợi nhuận có điều kiện.*

Gọi: R là giá bán 1 cuốn sách

C là giá mua

V là giá bán vào cuối năm.

Lúc đó lợi nhuận có điều kiện được xác định bằng biểu thức:

$$P_{ij} = \begin{cases} R.j - C.i + V(i-j) & \text{với } j \le i \\ R.i - C.i & \text{với } j > i \end{cases} \qquad (2.22)$$

Với các số liệu đã cho ta có:

$$P_{ij} = \begin{cases} 10.j - 7.i + 4(i-j) & \text{với } j \le i \\ 10i - 7i = 3i & \text{với } j > i \end{cases}$$

Ta có bảng lợi nhuận có điều kiện sau đây (bảng 2.1)

Bảng 2.1

Nhu cầu	P_j / j	0,3 20	0,25 21	0,18 22	0,14 23	0,1 24	0,03 25
Lượng hàng nhập	20	60	60	60	60	60	60
	21	57	63	63	63	63	63
	22	54	60	66	66	66	66
	23	51	57	63	69	69	69
	24	48	54	60	66	72	72
	25	45	51	57	63	69	75

Chiến lược của cửa hàng là phải chọn số lượng sách cần nhập i để cực đại lợi nhuận kỳ vọng. Với mỗi số lượng nhập i lợi nhuận kỳ vọng được tính bằng công thức:

$$PE_i = \sum_j P_j . P_{ij} \qquad (2.23)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đó ta có các giá trị lợi nhuận kỳ vọng như sau tuỳ thuộc vào số lượng nhập.

Số lượng nhập i	Lợi nhuận kỳ vọng PE_i
20	60,00
21	61,20
22	60,90
23	59,52
24	57,30
25	54,48

Vậy chiến lược mang lại lợi nhuận kỳ vọng tối đa là nhập 21 cuốn sách.

4.2. Trung vị

Trung vị, ký hiệu là m_d là giá trị nằm ở chính giữa phân phối xác suất của biến ngẫu nhiên. Nói cách khác đó là giá trị chia phân phối của biến ngẫu nhiên thành hai phần bằng nhau.

Nếu X là biến ngẫu nhiên rời rạc thì giá trị X_i sẽ là trung vị m_d nếu thoả mãn điều kiện:

$$F(X_i) \le 0,5 < F(X_{i+1}) \tag{2.24}$$

Còn nếu X là biến ngẫu nhiên liên tục thì trung vị m_d là giá trị thoả mãn điều kiện:

$$\int_{-\infty}^{m_d} f(x)dx = 0,5 \tag{2.25}$$

4.3. Mốt

Mốt, ký hiệu là m_0, là giá trị của biến ngẫu nhiên tương ứng với:

1. Xác suất lớn nhất nếu là biến ngẫu nhiên rời rạc
2. Cực đại của hàm mật độ xác suất nếu là biến ngẫu nhiên liên tục.

Trong thực tế có thể gặp biến ngẫu nhiên không có giá trị mốt hoặc ngược lại nhiều giá trị mốt cùng một lúc.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 7. Tìm trung vị và mốt của biến ngẫu nhiên có bảng phân phối xác suất sau:

X	20	21	22	23	24	25
P	0,3	0,25	0,18	0,14	0,1	0,03

Giải. Để tìm trung vị trước hết ta xây dựng hàm phân bố xác suất của X.

$$F(x)=\begin{cases}0 \text{ với } x \le 20\\ 0,3 \text{ với } 20<x \le 21\\ 0,55 \text{ với } 21<x \le 22\\ 0,73 \text{ với } 22<x \le 23\\ 0,87 \text{ với } 23<x \le 24\\ 0,97 \text{ với } 24<x \le 25\\ 1 \text{ với } x>25\end{cases}$$

Từ đó $m_d=21$

Dễ thấy rằng $m_0=20$.

Thí dụ 8. Thu nhập của dân cư tại một vùng là biến ngẫu nhiên liên tục có hàm phân bố xác suất như sau:

$$F(x)=\begin{cases}1-\left(\frac{x_0}{x}\right)^{\alpha} & \text{Với } x \ge x_0\\ 0 & \text{Với } x<x_0\end{cases} \qquad (\alpha>0)$$

Hãy tìm một mức thu nhập sao cho một nửa số dân của vùng đó có thu nhập cao hơn mức nói trên.

Giải. Mức thu nhập cần tìm chính là m_d.

Ta có:

$$f(x)=\begin{cases}\alpha x_0^{\alpha} x^{-\alpha-1} & \text{với } x \ge x_0\\ 0 & \text{với } x>x_0\end{cases}$$

Từ đó:

$$\int_{x_0}^{m_d} f(x)dx=0,5 \rightarrow m_d=x_0.2^{1/\alpha}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

2.19. Tỷ lệ khách hàng phản ứng tích cực đối với một chiến dịch quảng cáo là biến ngẫu nhiên có bảng phân phối xác suất như sau:

X (%)	0	10	20	30	40	50
P_x	0,1	0,2	0,35	0,2	0,1	0,05

a. Hãy chứng tỏ các xác suất P_x tạo nên một bảng phân phối xác suất.

b. Tìm tỷ lệ khách hàng bình quân phản ứng tích cực đối với chiến dịch quảng cáo đó.

c. Tìm xác suất để có trên 20% khách hàng phản ứng tích cực đối với chiến dịch quảng cáo.

2.20. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	1	4	8
P	0,3	0,1	0,6

Tìm $P(|X - E(X)| < 4)$

ĐS: $P = 0,7$.

2.21. Biến ngẫu nhiên rời rạc X nhận ba giá trị có thể có là $x_1 = 4$ với xác suất $P_1 = 0,5$; $x_2 = 0,6$ với xác suất $P_2 = 0,3$ và x_3 với xác suất P_3. Tìm x_3 và P_3 biết $E(X) = 8$.

2.22. Một người đi từ nhà đến cơ quan phải qua 3 ngã tư, xác suất để người đó gặp đèn đỏ ở các ngã tư tương ứng là 0,2; 0,4 và 0,5. Hỏi thời gian trung bình phải ngừng trên đường là bao nhiêu, biết rằng mỗi khi gặp đèn đỏ người ấy phải đợi khoảng 3 phút.

ĐS: Phải chờ khoảng 3,3 phút.

2.23. Có 5000 người xét nghiệm máu để tìm ký sinh trùng sốt rét. Tỷ lệ mắc bệnh ở địa phương theo thống kê là 10%. Có thể làm xét nghiệm theo 2 phương pháp:

Phương pháp 1: Xét nghiệm từng người.

Phương pháp 2: Lấy máu 10 người một trộn lẫn làm một xét nghiệm. Nếu kết quả xét nghiệm là âm tính (vô trùng) thì thông qua 10 người không ai mắc bệnh. Nếu kết quả xét nghiệm là dương tính (có trùng) thì chứng tỏ trong 10 người đó có ít nhất một người mắc bệnh. Lúc đó phải làm thêm 10 xét nghiệm lẻ để phát hiện các con bệnh cụ thể. Hỏi làm theo cách nào lợi hơn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải: Theo phương pháp 1 thì phải làm 5000 xét nghiệm.
Theo phương pháp 2 thì gọi X là số xét nghiệm phải làm đối với từng loạt 10 người.
X = 1 (nếu kết quả xét nghiệm là âm tính) và X = 11 (nếu kết quả là dương tính)
$P_1 = P(X=1) = (1 - 0,1)^{10} = (0,9)^{10}$
$P_2 = P(X=11) = 1 — P_1 = 1 - (0,9)^{10}$

Từ đó $E(X) \approx 7,5$ tức là trung bình phải làm 7,5 xét nghiệm.

Vậy tổng số xét nghiệm phải làm là: $\frac{5000}{10} \times 7,5 = 3750$ xét nghiệm.

Như vậy làm theo phương pháp 2 lợi hơn phương pháp 1 là 25%.

2.24. Nhu cầu hàng năm về loại hàng A là biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau (đơn vị: ngàn sản phẩm).

$$f(x) = \begin{cases} k(30 - x) & x \in (0, 30) \\ 0 & x \notin (0, 30) \end{cases}$$

a. Tìm k.

b. Tìm xác suất để nhu cầu về loại hàng đó không vượt quá 12.000 sản phẩm trong một năm.

c. Tìm nhu cầu trung bình hàng năm về loại hàng đó.

ĐS: a. k = 1/450; b. $P_b = 0,64$; c. E(X) = 10

2.25. Thời gian xếp hàng chờ mua hàng của khách là biến ngẫu nhiên liên tục với hàm phân bố xác suất như sau (đơn vị phút)

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ ax^3 - 3x^2 + 2x & \text{với } 2 < x \le 1 \\ 1 & \text{với } x > 1 \end{cases}$$

a. Tìm hệ số a

b. Tìm thời gian xếp hàng trung bình.

c. Tìm xác suất để trong 3 người xếp hàng thì có không quá 2 người phải chờ quá 0,5 phút.

ĐS: a = 2; *E(X) = 0,5; Pc = 0,875*

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.26. Biến ngẫu nhiên liên tục X có hàm phân bố xác suất như sau:

$$F(x) = \begin{cases} 0 & \text{với } x \le 2 \\ Cx - 1 & \text{với } 2 < x \le 4 \\ 1 & \text{với } x > 4 \end{cases}$$

a. Tìm hằng số C

b. Tìm E(X)

ĐS: a. $C = \frac{1}{2}$

b. $E(X) = 3$

2.27. Biến ngẫu nhiên X có hàm mật độ xác suất

$$f(x) = \begin{cases} a\cos x & \text{với } x \in (-\pi/2, \pi/2) \\ 0 & \text{với } x \notin (\pi/2, \pi/2) \end{cases}$$

a. Tìm hệ số a.

b. Tìm $P(0 \le X < \pi/4)$

c. Tìm E(X)

ĐS: a. $a = \frac{1}{2}$; b. $P(0 \le X < \frac{\pi}{4}) = \frac{\sqrt{2}}{4}$; c. $E(X) = 0$

2.28. Biến ngẫu nhiên liên tục X có hàm phân bố xác suất

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ 1/2 - k\cos x & \text{với } 0 < x \le \pi \\ 1 & \text{với } x > \pi \end{cases}$$

a. Tìm hệ số k

b. Tìm $P(0 < X < \pi/2)$

c. Tìm E(X)

ĐS: a. $k = \frac{1}{2}$; b. $P(0 < X < \pi/2) = \frac{1}{2}$ c. $E(X) = \frac{\pi}{2}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.4. Phương sai

Trong thực tế nhiều khi chỉ xác định kỳ vọng toán của biến ngẫu nhiên thì chưa đủ để xác định biến ngẫu nhiên đó. Ta còn phải xác định mức độ phân tán của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình của nó nữa. Chẳng hạn, khi nghiên cứu biến ngẫu nhiên là năng suất lúa của một địa phương nào đó, thì năng suất lúa trung bình (kỳ vọng toán) mới chỉ phản ánh được một khía cạnh của đại lượng đó mà thôi. Mức độ biến động về năng suất của các thửa ruộng khác nhau xung quanh giá trị trung bình cũng là một khía cạnh quan trọng cần nghiên cứu.

Ta có thể nghĩ rằng để đặc trưng cho mức độ phân tán thì đơn giản nhất là tìm tất cả các sai lệch của các giá trị của biến ngẫu nhiên so với kỳ vọng toán của nó và lấy trung bình số học của các sai lệch đó. Song cách làm này không mang lại kết quả vì có thể dễ dàng chứng minh được rằng với mọi biến ngẫu nhiên thì $E\left[X - E(X)\right] = 0$. Sở dĩ có điều đó vì các sai lệch dương và sai lệch âm xung quanh giá trị kỳ vọng toán bao giờ cũng bù trừ cho nhau, do đó giá trị trung bình của các sai lệch sẽ bằng không. Để khắc phục điều đó, người ta không tính trực tiếp trung bình của các sai lệch mà tính trung bình của các giá trị tuyệt đối hoặc bình phương của các sai lệch. Song đơn giản hơn là tìm trung bình của bình phương các sai lệch.

Từ đó ta có khái niệm phương sai.

1. Định nghĩa. *Phương sai của biến ngẫu nhiên X, ký hiệu là V(X) là kỳ vọng toán của bình phương sai lệch của biến ngẫu nhiên so với kỳ vọng toán của nó.*

$$V(X) = E\left[X - E(X)\right]^2 \tag{2.26}$$

Như vậy, Nếu X là biến ngẫu nhiên rời rạc thì phương sai sẽ được xác định bằng công thức:

$$V(X) = \sum_{i=1}^{n} \left[x_i - E(X)\right]^2 p_i \tag{2.27}$$

Còn nếu X là biến ngẫu nhiên liên tục thì phương sai được xác định bằng công thức:

$$V(X) = \int_{-\infty}^{+\alpha} \left[x - E(X)\right]^2 f(x)dx \tag{2.28}$$

Trong thực tế việc tính phương sai bằng các công thức định nghĩa trên

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

có thể gặp khó khăn. Người ta thường tính phương sai bằng công thức sau đây:

$$V(X) = E(X^2) - [E(X)]^2 \qquad (2.29)$$

Thật vậy, theo định nghĩa của phương sai ta có:

$$V(X) = E[X - E(X)]^2 = E\{\{X^2 - 2XE(X) + [E(X)]^2\} =$$

$$= E(X)^2 + E[-2XE(X)] + E[E(X)]^2$$

Song ta biết rằng kỳ vọng toán E(X) là một số xác định, do đó theo tính chất của kỳ vọng toán, ta có:

$$V(X) = E(X)^2 - 2E(X).E(X) + [E(X)]^2 = E(X)^2 - [E(X)]^2$$

Như vậy, đối với biến ngẫu nhiên rời rạc, ta có thể tính phương sai bằng công thức:

$$V(X) = \sum_{i=1}^{n} x_i^2 p_i - E[(X)]^2 \qquad (2.30)$$

Còn đối với biến ngẫu nhiên liên tục:

$$V(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - [E(X)]^2 \qquad (2.31)$$

Thí dụ 9. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất sau:

X	1	3	4
P	0,1	0,5	0,4

Tìm phương sai V(X).

Trước hết ta tìm kỳ vọng toán E(X):

$$E(X) = 1.0,1 + 3.0,5 + 4.0,4 = 3,2$$

Ta tìm kỳ vọng toán $E(X^2)$:

$$E(X^2) = 1^2.0,1 + 3^2.0,5 + 4^2.0,4 = 11$$

Như vậy: $V(X) = 11 - (3,2)^2 = 0,76$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 10. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x)=\begin{cases}2x & \text{với } x\in(0,1)\\ 0 & \text{với } x\notin(0,1)\end{cases}$$

Tìm phương sai V(X).

Ta tìm kỳ vọng toán E(X)

$$E(X)=\int_{+\infty}^{-\infty} xf(x)dx=\int_0^1 x.2x.dx=\frac{2x^3}{3}\Big|_0^1=\frac{2}{3}$$

Tìm kỳ vọng toán $E(X^2)$:

$$E(X^2)=\int_{+\infty}^{-\infty} x^2f(x)dx=\int_0^1 x^2.2x.dx=\frac{x^4}{2}\Big|_0^1=\frac{1}{2}$$

Vậy:

$$V(X)=\frac{1}{2}-\left(\frac{2}{3}\right)^2=\frac{1}{18}$$

Ta thấy rằng phương sai của biến ngẫu nhiên là một giá trị xác định không âm.

2. Các tính chất của phương sai

Tính chất 1. Phương sai của một hằng số bằng không:

$$V(C)=0 \tag{2.32}$$

Thật vậy, theo định nghĩa của phương sai:

$$V(C)=E\left[C-E(C)\right]^2=E\left[C-C\right]^2=E(0)=0$$

Tính chất 2. Phương sai của tích giữa một hằng số và một biến ngẫu nhiên bằng tích giữa bình phương hằng số đó và phương sai của biến ngẫu nhiên ấy.

$$V(CX)=C^2V(X) \tag{2.33}$$

Thật vậy, theo định nghĩa của phương sai

$$V(CX)=E\left[CX-E(CX)\right]^2=E\left[CX-CE(X)\right]^2=$$

$$=EC^2\left[X-E(X)\right]^2=C^2E\left[X-E(X)\right]^2=C^2V(X)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tính chất 3. Phương sai của tổng hai biến ngẫu nhiên độc lập bằng tổng các phương sai thành phần:

$$V(X+Y)=V(X)+V(Y) \tag{2.34}$$

Thật vậy, theo công thức tính phương sai:

$$V(X+Y)=\left[E(X+Y)^2\right]-E\left[(X+Y)\right]^2=$$

$$=E\left[X^2+2XY+Y^2\right]-\left[E(X)+E(Y)\right]^2=$$

$$=E(X^2)+2E(X).E(Y)+E(Y^2)-\left[E(X)\right]^2-2E(X).E(Y)-\left[E(Y)\right]^2$$

$$=E(X^2)-\left[E(X)\right]^2+E(Y^2)-\left[E(Y)\right]^2=$$

$$=V(X)+V(Y)$$

Bằng phương pháp quy nạp có thể chứng minh được hệ quả sau.

Hệ quả 1. Phương sai của tổng n biến ngẫu nhiên độc lập với nhau $X_1, X_2, ..., X_n$ bằng tổng các phương sai thành phần:

$$V\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}V(X_i) \tag{2.35}$$

Ngoài ra từ các tính chất nêu trên có thể chứng minh được các hệ quả sau đây:

Hệ quả 2. Phương sai của tổng một hằng số với một biến ngẫu nhiên bằng phương sai của chính biến ngẫu nhiên đó:

$$V(C+X)=V(X) \tag{2.36}$$

Hệ quả 3. Phương sai của hiệu hai biến ngẫu nhiên độc lập bằng tổng các phương sai thành phần:

$$V(X-Y)=V(X)+V(Y) \tag{2.37}$$

Một số tính chất khác của phương sai sẽ được đề cập tiếp ở chương IV.

3. Bản chất và ý nghĩa của phương sai

Xuất phát từ định nghĩa của phương sai, ta thấy phương sai chính là trung bình số học của bình phương các sai lệch giữa các giá trị có thể có của biến ngẫu nhiên so với giá trị trung bình của các giá trị đó. Do đó nó phản ánh mức độ phân tán của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình của nó là kỳ vọng toán.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 1. Gieo n con xúc xắc. Tìm phương sai của tổng số điểm thu được.

Giải. Gọi $X_i (1 = \overline{1,n})$ là số điểm thu được ở con xúc xắc thứ i. Gọi X là tổng số điểm thu được ở cả n con xúc xắc. Vậy:

$$X = \sum_{i=1}^{n} X_i$$

Vì các $X_i (1 = \overline{1,n})$ độc lập với nhau, do đó theo tính chất của phương sai ta có:

$$V(X) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$$

Ta tìm $V(X_1)$ theo công thức:

$$V(X_1) = E(X_1^2) - [E(X_1)]^2$$

Mỗi biến ngẫu nhiên $X_i (i = \overline{1,n})$ đều có bảng phân phối xác suất như sau:

X_i	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Do đó: $E(X_i) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$

Và $E(X_1^2) = \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) = \frac{91}{6}$

Do đó: $V(X_i) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12} \quad (i = \overline{1,n})$

Vì vậy: $V(X) = \sum_{i=1}^{n} V(X_i) = \frac{35}{12} n$

Thí dụ 12. Xác suất để một máy sản xuất ra phế phẩm bằng p. Máy sẽ được sửa chữa ngay sau khi làm ra phế phẩm. Tìm phương sai của số sản phẩm được sản xuất ra giữa hai lần sửa chữa.

Giải. Ở phần trình bày về kỳ vọng toán ta đã xây dựng được bảng phân phối xác suất của X (số sản phẩm sản xuất ra giữa hai lần sửa chữa) như sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

X	1	2	3	...	k	...
P	p	qp	q^2p	...	$q^{k-1}p$	...

Và đã tìm được $E(X) = \frac{1}{p}$

Ta tìm $E(X^2)$

$$E(X^2) = \sum_{k=1}^{\infty} k^2 q^{k-1} p = p\sum_{k=1}^{\infty} k^2 q^{k-1}$$

Ở trên ta đã chứng minh được rằng:

$$\sum_{k=1}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2}$$

Nhân cả hai vế với q và sau đó lấy đạo hàm theo q, ta có:

$$\frac{d}{dq}\sum_{k=1}^{\infty} k q^{k} = \sum_{k=1}^{\infty} k^2 q^{k-1}$$

Và $\left(\frac{q}{(1-q)^2}\right)' = \frac{(1+q)}{(1-q)^3} = \frac{2-p}{p^3}$

Như vậy: $\sum_{k=1}^{\infty} k^2 q^{k-1} = \frac{2-p}{p^3}$

Do đó: $E(X^2) = \frac{2-p}{p^2}$

Và $V(X) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$

4. Ứng dụng thực tế của phương sai

Cùng với kỳ vọng toán, phương sai có những ứng dụng to lớn trong nhiều lĩnh vực thực tiễn. Nếu như trong kỹ thuật phương sai đặc trưng cho mức độ phân tán của các chi tiết gia công hay sai số của thiết bị thì trong quản lý và kinh doanh nó đặc trưng cho mức độ rủi ro của các quyết định. Ta sẽ minh hoạ điều đó qua ví dụ sau: Một nhà đầu tư đang cân nhắc giữa việc đầu tư vào hai dự án A và B trong hai lĩnh vực độc lập nhau. Khả năng

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

thu hồi vốn sau 2 năm (tính bằng %) của hai dự án là các biến ngẫu nhiên có bảng phân phối xác suất như sau:

Dự án A

X_A	65	67	68	69	70	71	73
P	0,04	0,12	0,16	0,28	0,24	0,08	0,08

Dự án B

X_B	66	68	69	70	71
P	0,12	0,28	0,32	0,20	0,08

Từ các bảng phân phối xác suất trên tìm được:

$$E(X_A) = 69,16\%; V(X_A) = 3,0944;$$
$$E(X_B) = 68,72\%; V(X_B) = 1,8016;$$

Như vậy nếu cần chọn phương án đầu tư sao cho tỉ lệ thu hồi vốn kỳ vọng cao hơn thì nên chọn dự án A, song nếu cần chọn phương án đầu tư sao cho độ rủi ro của tỉ lệ thu hồi vốn thấp hơn tức là khả năng thu hồi vốn ổn định hơn thì lại nên chọn dự án B.

4.5. Độ lệch chuẩn

Độ lệch chuẩn của biến ngẫu nhiên X, ký hiệu σ_x, là căn bậc hai của phương sai:

$$\sigma_x = \sqrt{V(X)} \qquad (2.38)$$

Ta thấy rằng đơn vị đo của phương sai bằng bình phương đơn vị đo của biến ngẫu nhiên. Vì vậy khi cần phải đánh giá mức độ phân tán của biến ngẫu nhiên theo đơn vị đo của nó người ta thường tính độ lệch chuẩn chứ không phải là phương sai vì độ lệch chuẩn có cùng đơn vị đo với biến ngẫu nhiên cần nghiên cứu.

4.6. Hệ số biến thiên

Để đo lường mức độ quan trọng tương đối của sự phân tán của một phân phối người ta thường dùng *hệ số biến thiên* ký hiệu là CV và được xác định bằng biểu thức

$$CV = \left|\frac{\sigma_x}{E(X)}\right| \times 100\% \text{ nếu } E(X) \neq 0 \qquad (2.39)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

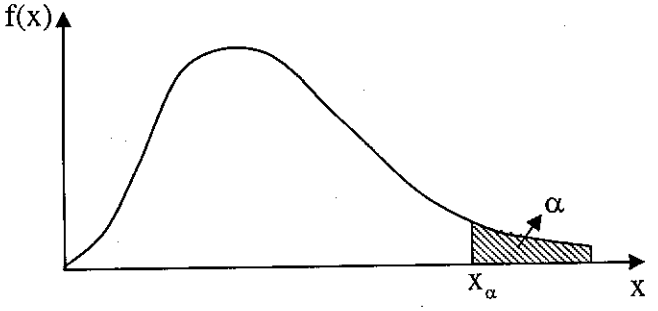
Hệ số biến thiên thường được dùng để đo mức độ thuần nhất của một phân phối. Giá trị của nó càng nhỏ thì mức độ thuần nhất càng lớn. Ngoài ra nó còn dùng để so sánh mức độ phân tán của hai phân phối mà kỳ vọng toán và độ lệch chuẩn của chúng không nhất thiết phải như nhau.

4.7. Giá trị tới hạn

Đối với biến ngẫu nhiên liên tục X trong một số trường hợp người ta còn tìm một loại giá trị gọi là *giá trị tới hạn*. Giá trị tới hạn mức α của biến ngẫu nhiên X, ký hiệu là x_α, là giá trị của X thoả mãn điều kiện

$$P(X > x_\alpha) = \alpha \qquad (2.40)$$

Trên đồ thị giá trị tới hạn x_α có thể mô tả như sau:



Như vậy giá trị tới hạn x_α là giá trị sao cho diện tích giới hạn bởi trục hoành, đường cong hàm mật độ xác suất và đường thẳng $x = x_\alpha$ bằng α.

Bài tập

2.29. Cho X_1 X_2 X_3 là các biến ngẫu nhiên độc lập và có bảng phân phối xác suất của chúng như sau:

X_1	0	1
P	0,6	0,4

X_2	1	2
P	0,4	0,6

X_3	0	2
P	0,8	0,2

Lập $\overline{X} = \dfrac{X_1 + X_2 + X_3}{3}$ Tính $E(\overline{X})$ và $V(\overline{X})$

ĐS: $E(\overline{X}) = 0,8$; $V(\overline{X}) = 0,12$

2.30. Bảng phân phối xác suất của biến ngẫu nhiên X như sau

X	-5	2	3	4
P	0,4	0,3	0,1	0,2

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Tính E(X); V(X) và σ_x.

b. Tìm giá trị mốt m_0.

ĐS: E(X) = - 0,3 ; V(X) = 15,21 ; σ_x = 3,9.

M_0 = - 5

2.31. Tại một cửa hàng bán xe máy Honda người ta thống kê được số xe máy bán ra hàng tuần (X) với bảng phân phối xác suất như sau:

X	0	1	2	3	4	5	6	7	8	9	10	11
P_x	0,05	0,12	0,17	0,08	0,12	0,20	0,07	0,02	0,07	0,02	0,03	0,05

a. Tìm số xe trung bình bán được mỗi tuần.

b. Tìm phương sai và độ lệch chuẩn của số xe bán được mỗi tuần và giải thích ý nghĩa của kết quả nhận được.

2.32. Trở lại bài 2.31. Nếu giá bình quân của mỗi xe máy bán ra tại cửa hàng đó là 12 triệu đồng thì doanh thu bình quân hàng tuần của cửa hàng đó là bao nhiêu?

2.33. Thống kê tại tất cả 52 cửa hàng bán sản phẩm của công ty trên toàn quốc thu được các số liệu sau:

Số nhân viên bán hàng ở cửa hàng	2	3	4	5
Số cửa hàng tương ứng	10	12	16	14

a. Xây dựng bảng phân phối xác suất và hàm phân bố xác suất của số nhân viên bán hàng tại mỗi cửa hàng.

b. Tìm số nhân viên trung bình ở mỗi cửa hàng và phương sai tương ứng.

2.34. Cho X và Y là 2 biến ngẫu nhiên độc lập với

E(X) = V(X) = 3; E(Y) = V(Y) = 2

a. Tính E(Z) và V(Z) với Z = (3X — 2Y) / 5

b. Tính E(T) với $T = \dfrac{Z - E(Z)}{V(Z)}$

2.35. Thực hiện ba lần bắn bia với xác suất trúng bia tương ứng là 0,3; 0,4; 0,6. Tìm kỳ vọng toán và phương sai của số lần bắn trúng bia.

ĐS: E(X) = 1,3; V(X) = 0,69

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.36. Thống kê số khách trên 1 ô tô buýt tại một tuyến giao thông thu được các số liệu sau:

Số khách trên 1 chuyến	20	25	30	35	40
Tần suất tương ứng	0,2	0,3	0,15	0,1	0,25

Tìm kỳ vọng toán và phương sai của số khách đi mỗi chuyến và giải thích ý nghĩa của kết quả thu được.

2.37. Với các số liệu của bài 2.36, giả sử chi phí cho mỗi chuyến xe là 200 ngàn đồng không phụ thuộc vào số khách đi trên xe thì để công ty xe buýt có thể thu được lãi bình quân cho mỗi chuyến xe là 100 ngàn đồng thì phải quy định giá vé là bao nhiêu?

ĐS: 10, 17 ngàn.

2.38. Biến ngẫu nhiên rời rạc X có các giá trị có thể có là $x_1 = -1$; $x_2 = 0$; $x_3 = +1$. Tìm các xác suất tương ứng P_1; P_2 và P_3 biết rằng $E(X) = 0,1$ và $E(X^2) = 0,9$.

ĐS: $P_1 = 0,4$; $P_2 = 0,1$ $P_3 = 0,5$.

2.39. Biến ngẫu nhiên rời rạc X có quy luật phân phối xác suất như sau:

X	x_1	x_2
P	P_1	0,7

Tìm x_1, x_2 và P_1 biết $E(X) = 2,7$ và $V(X) = 0,21$ ($x_2 > x_1$)

ĐS: $x_1 = 2$, $x_2 = 3$, $P_1 = 0,3$

2.40. Có 5 sản phẩm trong đó có 4 chính phẩm và một phế phẩm. Người ta lấy ra lần lượt 2 sản phẩm (lấy không hoàn lại).

a. Gọi X là "số phế phẩm có thể gặp phải". Lập bảng phân phối xác suất của X.

b. Tính E(X) và V(X)

c. Gọi Y là "số chính phẩm có thể gặp phải". Lập hệ thức cho biết mối quan hệ giữa Y và X.

d. Tính E(Y) và V(Y).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.41. Kinh nghiệm cho thấy là số lượng một loại sản phẩm mà một khách hàng mua có bảng phân phối xác suất như sau:

Số lượng sản phẩm	0	1	2	3
Xác suất tương ứng	0,5	0,1	0,2	0,2

a. Nếu mỗi sản phẩm được bán với giá 110 ngàn đồng và nhân viên bán hàng được hưởng 10% trên số sản phẩm bán được thì số tiền hoa hồng bình quân mà nhân viên bán hàng được hưởng từ mỗi khách hàng là bao nhiêu.

b. Tìm phương sai của số tiền hoa hồng đó và nêu ý nghĩa của kết quả thu được.

2.42. Qua theo dõi trong nhiều năm kết hợp với sự đánh giá của các chuyên gia tài chính thì lãi suất đầu tư vào một công ty là biến ngẫu nhiên có bảng phân phối xác suất như sau:

X (%)	9	10	11	12	13	14	15
P_x	0,05	0,15	0,3	0,2	0,15	0,1	0,05

a. Tìm xác suất để khi đầu tư vào công ty đó thì sẽ đạt được lãi suất ít nhất là 12%.

b. Tìm lãi suất có thể hy vọng khi đầu tư vào công ty đó.

c. Mức độ rủi ro khi đầu tư vào công ty đó có thể đánh giá bằng cách nào?

2.43. Tỷ lệ mắc một loại bệnh trong một vùng dân cư là biến ngẫu nhiên liên tục có hàm mật độ

$$f(x) = \begin{cases} \dfrac{1}{20} & \text{với } x \in (5, 25) \\ 0 & \text{với } x \notin (5, 25) \end{cases}$$

a. Tính $P(|X - 10| > 2,5)$

b. Tính tỷ lệ mắc bệnh trung bình và phương sai.

ĐS: a. $P = 0,75$

b. $E(X) = 15$; $V(X) = 33,3$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.44. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất

$$f(x) = \begin{cases} k & \text{với } x \in (a, b) \\ 0 & \text{với } x \notin (a, b) \end{cases}$$

a. Tìm hệ số k
b. Tìm E(X) và V(X)
c. Vẽ đồ thị hàm f(x) và F(x)

ĐS: a. $k = \dfrac{1}{b - a}$

b. $E(X) = \dfrac{a + b}{2}$; $V(X) = \dfrac{(b - a)^2}{12}$

$$F(x) = \begin{cases} 0 & \text{với } x \le a \\ \dfrac{x - a}{b - a} & \text{với } a < x \le b \\ 1 & \text{với } x > b \end{cases}$$

4.8. Một vài tham số đặc trưng dạng phân phối xác suất

1. Hệ số bất đối xứng

Mức độ đối xứng của một phân phối có thể quan sát qua đồ thị của nó song để đo lường mức độ bất đối xứng người ta dùng *hệ số bất đối xứng được xác định bằng biểu thức.*

$$\alpha_3 = \frac{\mu_3}{\sigma_3} \qquad (2.41)$$

Trong đó $\mu_3 = E[X - E(X)]^3$ và σ^3 là lập phương của độ lệch chuẩn.

Tuỳ thuộc vào giá trị của α_3 mà ta có thể đưa ra các kết luận sau:

- Nếu $\alpha_3<0$ thì phân phối là bất đối xứng và đồ thị sẽ xuôi về bên trái nhiều hơn.

- Nếu $\alpha_3=0$ thì phân phối là đối xứng.

- Nếu $\alpha_3>0$ thì phân phối là bất đối xứng và đồ thị sẽ xuôi về bên phải nhiều hơn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Qua so sánh giá trị của trung vị và kỳ vọng toán cũng có thể biết được dấu của hệ số bất đối xứng như sau:

- Nếu $m_d > E(X)$ thì $\alpha_3 < 0$
- Nếu $m_d = E(X)$ thì $\alpha_3 = 0$
- Nếu $m_d < E(X)$ thì $\alpha_3 > 0$

2. Hệ số nhọn

Hệ số nhọn cho phép nhận xét về dạng của một phân phối và bổ sung thêm thông tin về phương sai. Phương sai của biến ngẫu nhiên có thể được xem là nhỏ, lớn hay trung bình. Lúc đó đồ thị của phân phối hoặc sẽ rất tập trung, ít tập trung hay tập trung ở mức bình thường. Trường hợp cuối cùng này thường là mốc để so sánh.

Hệ số nhọn được xác định bằng công thức sau:

$$\alpha_4 = \frac{\mu_4}{\sigma_4}$$

Trong đó $\mu_4 = E[X - E(X)]^4$ và σ^4 là bình phương của phương sai.

Khi phân phối xác suất được tập trung ở mức bình thường thì $\sigma_4 = 3$. Nếu phân phối tập trung ở mức độ cao hơn thì $\sigma_4 > 3$ còn ngược lại nếu phân phối tập trung ở mức thấp hơn thì $\sigma_4 < 3$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5. TÓM TẮT CHƯƠNG 2

1. Trong thực tế mỗi đại lượng mà ta thường phải nghiên cứu trong khoa học tự nhiên, kinh tế xã hội hay khoa học kỹ thuật về thực chất đều là *các biến ngẫu nhiên*, tức là trong kết quả của một phép thử, chúng sẽ nhận giá trị bằng bao nhiêu là điều không thể khẳng định chắc chắn được do ảnh hưởng của các tác động ngẫu nhiên lên các đại lượng đó.
2. Quy luật biến động của một biến ngẫu nhiên có thể mô tả một cách đầy đủ nhất thông qua quy luật phân phối xác suất của nó.
3. Đối với biến ngẫu nhiên rời rạc thì hình thức mô tả quy luật phân phối xác suất tiện dụng nhất là bảng phân phối xác suất còn đối với biến ngẫu nhiên liên tục thì hình thức mô tả quy luật phân phối xác suất tiện dụng nhất là hàm mật độ xác suất, còn hàm phân bố xác suất có thể dung chung cho cả hai loại biến ngẫu nhiên.
4. Trong số các tham số đặc trưng về mặt vị trí của biến ngẫu nhiên thì kỳ vọng toán mang nhiều ý nghĩa thực tiễn hơn cả, còn trung vị và mốt mang các thông tin bổ trợ cho kỳ vọng toán. Trong kinh tế nếu xét trong một số lớn phép thử tương tự thì kỳ vọng toán phản ánh giá trị trung bình, song nếu xét trong một phép thử đơn lẻ thì nó lại phản ánh giá trị mong đợi.
5. Trong số các tham số phản ánh mức độ phân tán thì phương sai ít có ý nghĩa vì đơn vị đo của nó bằng bình phương đơn vị đo của biến ngẫu nhiên. Trong hai tham số có ý nghĩa phản ánh hơn cả thì độ lệch chuẩn phản ánh mức độ biến động tuyệt đối còn hệ số biến thiên phản ánh mức độ biến động tương đối của các giá trị của biến ngẫu nhiên so với giá trị trung bình.
6. Trong kinh tế thì ngoài ý nghĩa là thước đo mức độ phân tán, mức độ bất ổn định, mức độ mất chính xác giống như các lĩnh vực khác, Phương sai, độ lệch chuẩn và hệ số biến thiên còn được dùng để đo mức độ rủi ro của các hoạt động kinh tế.
7. Để nhận xét về dạng phân phối xác suất có thể dùng hệ số bất đối xứng và hệ số nhọn. Các tham số này còn được sử dụng làm căn cứ cho một số kiểm định về dạng phân phối xác suất.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các ký hiệu và công thức cơ bản

- $X, Y, Z, X_1, X_2, ..., X_n$ - Biến ngẫu nhiên
- Bảng phân phối xác suất của biến ngẫu nhiên rời rạc X

X	x_1	x_2	...	x_n
P_x	p_1	p_2	...	p_n

Điều kiện cơ bản:

$$\begin{cases} p_i \geq 0 \quad \forall i \\ \sum_{i=1}^{n} p_i = 1 \end{cases}$$

- Hàm phân bố xác suất $F(x) = P(X < x)$

Với biến ngẫu nhiên X rời rạc: $F(x) = \sum_{x_i < x} P_i$

Với biến ngẫu nhiên X liên tục $F(x) = \int_{-\infty}^{x} f(x)dx$

- Hàm mật độ xác suất $f(x) = F'(x)$

Điều kiện cơ bản: $\begin{cases} f(x) \geq 0 \quad \forall x \\ \int_{-\infty}^{+\infty} f(x)dx = 1 \end{cases}$

- Các công thức tính xác suất

– Nếu X là biến ngẫu nhiên rời rạc

$$P(a \leq X < b) = F(b) - F(a) = \sum_{a < x_i < b} P(X = x_i)$$

– Nếu X là biến ngẫu nhiên liên tục

$$P(a < X < b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Các tham số đặc trưng cơ bản

+ Kỳ vọng toán:

$$E(X) = \sum_{i=1}^{n} x_i p_i \text{ (X rời rạc)}$$

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \text{ (X liên tục)}$$

+ Phương sai:

$$V(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

$$V(X) = \sum_{i=1}^{n} x_i^2 p_i - \left(\sum_{i=1}^{n} x_i p_i \right)^2 \text{ (X rời rạc)}$$

$$V(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - \left(\int_{-\infty}^{+\infty} xf(x)dx \right)^2 \text{ (X liên tục)}$$

+ Độ lệch chuẩn: $\sigma_x = \sqrt{V(X)}$

+ Trung vị: m_d

+ Mốt: m_0

+ Hệ số biến thiên:

$$CV_x = \left| \frac{\sigma_x}{E(X)} \right| .100(\%)$$

+ Hệ số bất đối xứng: $\alpha_3 = \frac{\mu_3}{\sigma_X^3}$

+ Hệ số nhọn: $\alpha_4 = \frac{\mu_4}{\sigma_X^4}$

+ Giá trị tới hạn mức α: x_α thoả mãn $P(X > x_\alpha) = \alpha$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Biến ngẫu nhiên	Random variable
Biến ngẫu nhiên rời rạc	Discrete random variable
Biến ngẫu nhiên liên tục	Continuos random variable
Quy luật phân phối xác suất	Probability distribution
Bảng phân phối xác suất	Probability distribution table
Hàm phân bố xác suất	Cumulative distribution function
Hàm mật độ xác suất	Density probability function
Kỳ vọng toán	Expected value, Expectation
Trung vị	Median
Mốt	Mode
Phương sai	Variance
Độ lệch chuẩn	Standard deviation
Hệ số biến thiên	Coefficient of variation
Giá trị tới hạn	Critical value
Hệ số bất đối xứng	Skewness
Hệ số nhọn	Kurtosis

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

2.45. Gieo một con xúc xắc. Lập bảng phân phối xác suất của số lần xuất hiện mặt 6 chấm.

2.46. Trong một hộp có 6 quả cầu trắng, 4 quả cầu đỏ. Lấy ngẫu nhiên 1 quả cầu. Tìm quy luật phân phối xác suất của số cầu đỏ được lấy ra.

2.47. Một người được phát 3 viên đạn để bắn lần lượt vào bia cho đến khi trúng với xác suất bắn trúng mỗi viên là 0,8. Tìm quy luật phân phối xác suất của số viên đạn được bắn ra.

2.48. Cho hàm số

$$f(x)=\begin{cases}0 \text{ với } x \notin (10, 20)\\ \dfrac{1}{20} \text{ với } x \in (10, 20)\end{cases}$$

$f(x)$ có phải là hàm mật độ xác suất không?

2.49. Chứng tỏ rằng hàm số

$$F(x)=\begin{cases}0 & \text{với } x \le 1\\ 2x-2 & \text{với } 1< x \le 2\\ 1 & \text{với } x > 2\end{cases}$$

Không phải là hàm phân bố xác suất.

2.50. Hãy cho biết các mệnh đề sau đây là đúng hay sai? Tại sao?

a. Kỳ vọng toán của tổng một số hữu hạn các biến ngẫu nhiên bằng tổng các kỳ vọng toán thành phần.

b. Kỳ vọng toán của tích một số hữu hạn các biến ngẫu nhiên bằng tích các kỳ vọng toán thành phần.

c. Phương sai của hiệu hai biến ngẫu nhiên bằng hiệu các phương sai thành phần.

2.51. Cho X và Y là hai biến ngẫu nhiên độc lập có các bảng phân phối xác suất như sau:

X	2	3	5
P	0,3	0,5	0,2

Y	1	4
P	0,2	0,8

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Tìm các bảng phân phối xác suất của $X+Y$ và $X.Y$.

b. Tìm $E(X+Y), E(X.Y), V(X+Y)$ và $V(X.Y)$ bằng tất cả các phương pháp có thể.

2.52. Cho hai biến ngẫu nhiên X và Y độc lập. Tính V(Z) biết:

a. $Z = 2X + 3Y$

b. $Z = -3X$

2.53. Biến ngẫu nhiên X có kỳ vọng toán $E(X) = a$ và phương sai $V(X) = b^2$. Tìm kỳ vọng toán và phương sai của biến ngẫu nhiên

$$U = \frac{X - a}{b}$$

2.54. Cho X là biến ngẫu nhiên. Chứng minh rằng:

$$E[X - E(X)] = 0$$

2.55. Biến ngẫu nhiên X chỉ nhận hai giá trị có thể có là c và -c với xác suất như nhau. Tìm V(X).

2.56. Chứng minh rằng:

a. $E(Y) = a.E(X) + b$ và $V(Y) = a^2.V(x)$ nếu $Y = aX + b$

b. $E(Y) = \sum_{i=1}^{n} a_i E(X_i) + b$ và $V(Y) = \sum_{i=1}^{n} a_i^2 V(X_i)$ nếu

$Y = \sum_{i=1}^{n} a_i X_i + b$ X_i là các biến ngẫu nhiên độc lập.

2.57. Chứng minh rằng với mọi biến ngẫu nhiên X thì

$$E(X^2) \geq \left[E(X)\right]^2$$

2.58. Chứng minh rằng kỳ vọng toán của biến ngẫu nhiên rời rạc nằm trong khoảng giá trị bé nhất và giá trị lớn nhất của biến ngẫu nhiên đó.

2.59. Biến ngẫu nhiên rời rạc X chỉ nhận hai giá trị có thể có là x_1 và x_2 với các xác suất tương ứng bằng nhau. Chứng minh rằng

$$V(X) = \frac{(x_2 - x_1)^2}{4}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.60. Chứng minh rằng phương sai của số lần xuất hiện biến cố trong một phép thử không vượt quá giá trị 1/4.

2.61. Cho hàm số $f(x)=\dfrac{a}{1+x^2} \quad -\infty < x < +\infty$

Hãy tìm mọi giá trị của a để hàm số trên không thể là hàm mật độ xác suất của biến ngẫu nhiên.

2.62. Cho hàm số

$$f(x)=\begin{cases}0 & \text{với } x \le -1\\ x+1 & \text{với } -1<x \le 0\\ -x+1 & \text{với } 0<x \le 1\\ 0 & \text{với } x>1\end{cases}$$

a. Chứng minh rằng f(x) là hàm mật độ xác suất.

b. Tìm hàm phân bố xác suất F(x) tương ứng.

c. Tìm E(X) và V(X).

d. Với $k > 0$ chứng tỏ rằng

$$P(|X| > k) \le \frac{1}{6k^2}$$

e. Tìm $m_d, m_0, x_{0,125}$. Vẽ đồ thị của hàm f(x) và mô tả các giá trị nói trên bằng hình vẽ.

2.63. Trong kinh tế và kinh doanh, ý nghĩa của kỳ vọng toán và phương sai là gì?

2.64. Tại sao trong thực tế dùng độ lệch chuẩn lại có ý nghĩa hơn dùng phương sai?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập tổng hợp chương 2

2.65. Theo thống kê dân số thì một người ở độ tuổi 40 sẽ sống thêm 1 năm nữa là 0,995. Một công ty bảo hiểm nhân thọ bán bảo hiểm một năm cho những người ở độ tuổi đó với giá là 10 ngàn và trong trường hợp người mua bảo hiểm bị chết thì số tiền bồi thường là 1 triệu. Hỏi lợi nhuận trung bình của công ty khi bán mỗi thẻ bảo hiểm loại này là bao nhiêu?

ĐS: 5 ngàn

2.66. Giá hàng ngày trên thị trường thế giới về đường (đơn vị: USD/fao) có bảng phân phối suất như sau:

X	0,78	0,79	0,80	0,81	0,82	0,83
P_x	0,05	0,10	0,25	0,40	0,15	0,05

a. Tìm xác suất để giá đường vào một ngày nào đó sẽ đạt ít nhất là 0,8 USD/fao.

b. Tìm xác suất để giá đường vào một ngày nào đó sẽ thấp hơn 0,82 USD/fao.

c. Giả sử giá hàng ngày của đường là độc lập nhau, tìm xác suất để trong hai ngày liên tiếp giá đường đều cao hơn 0,8 USD/fao.

2.67. Lợi nhuận X thu được khi đầu tư 50 triệu đồng vào một dự án có bảng phân phối xác suất như sau (đơn vị: triệu đồng).

X	-2	-1	0	1	2	3
P_x	0,1	0,1	0,2	0,2	0,3	0,1

a. Tìm mức lợi nhuận có khả năng nhiều nhất khi đầu tư vào dự án đó.

b. Việc đầu tư vào dự án này có hiệu quả hay không? Tại sao?

c. Làm thế nào để đo được mức độ rủi ro của vụ đầu tư này? Hãy tìm mức độ rủi ro đó.

2.68. Lợi nhuận thu được từ 1 triệu đồng đầu tư vào công ty A (X_A) và công ty B (X_B) có các bảng phân phối xác suất như sau:

(đơn vị: ngàn đồng)

X_A	-500	-100	100	500	700
PX_A	0,2	0,3	0,2	0,2	0,1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

X_B	-200	50	100
Px_B	0,1	0,6	0,3

a. Nếu dự định đầu tư 10 triệu đồng thì lợi nhuận kỳ vọng khi đầu tư vào công ty A và B là bao nhiêu.

b. Nếu dùng hệ số biến thiên như độ đo mức độ rủi ro của đầu tư thì việc đầu tư vào công ty nào rủi ro hơn?

2.69. Một công ty thuê một luật sư trong một vụ kiện với hai phương án trả công như sau:

Phương án 1: Trả 5 triệu đồng bất kể thắng hay thua kiện.

Phương án 2: Trả 100 ngàn đồng nếu thua kiện và 15 triệu đồng nếu thắng kiện.

Luật sư đã chọn phương án 2. Vậy theo đánh giá của luật sư thì khả năng thắng kiện của công ty tối thiểu là bao nhiêu.
ĐS: 0,329

2.70. Trên một tuyến bay, người ta thống kê được rằng có 0,5% hành khách bị mất hành lý và giá trị trung bình mà khách đòi bồi thường cho số hành lý bị mất là 600 ngàn đồng. Công ty hàng không muốn tăng thêm giá vé để bù đắp cho số tiền phải bồi thường cho số hành lý bị mất. Vậy công ty nên tăng thêm giá vé là bao nhiêu? Tại sao?

2.71. Số lượng thuyền gỗ X mà một xưởng đóng thuyền có thể làm được trong một tháng có bảng phân phối xác suất như sau:

X	2	3	4	5	6	7	8
P_X	0,2	0,2	0,3	0,1	0,1	0,05	0,05

a. Tìm xác suất để trong tháng tới xưởng đó sẽ đóng được từ 4 đến 7 con thuyền.

b. Tìm hàm phân bố xác suất của X.

c. Dùng hàm phân bố xác suất, hãy tính xác suất để trong tháng tới xưởng đó sẽ đóng được không quá 6 con thuyền.

d. Số thuyền có khả năng nhiều nhất mà xưởng đó có thể đóng được trong tháng tới là bao nhiêu?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

e. Giả sử việc đóng thuyền có chi phí cố định hàng tháng là 25 triệu đồng và chi phí bổ sung cho mỗi con thuyền là 5 triệu đồng. Hãy tìm chi phí bình quân hàng tháng của xưởng đó.

2.72. Số lượng sản phẩm hỏng mà một công nhân có thể làm ra trong một tháng có bảng phân phối xác suất như sau:

Số SP hỏng	0	1	2	3	4	5	6
Xác suất tương ứng	0,01	0,09	0,30	0,20	0,20	0,10	0,10

a.Tìm xác suất để trong một tháng người công nhân đó làm ra không quá 4 sản phẩm hỏng.

b. Giả sử số sản phẩm mà người công nhân đó phải làm bù bằng bình phương số sản phẩm hỏng mà người đó đã làm trong tháng. Tìm số sản phẩm phải làm bù bình quân mỗi tháng của người công nhân đó.

2.73. Số lượng xe ô tô TOYOTA mà một đại lý bán được trong một tuần có bảng phân phối xác suất như sau:

Số xe bán được	0	1	2	3	4	5
Xác suất tương ứng	0,1	0,1	0,2	0,2	0,3	0,1

a. Tìm xác suất để đại lý đó bán được ít nhất 4 xe trong một tuần.

b. Giả sử chi phí cho hoạt động của đại lý bằng căn bậc hai của số xe bán được nhân với 3 triệu. Tìm chi phí trung bình cho hoạt động của đại lý mỗi tuần.

2.74. Qua kinh nghiệm, một cửa hàng bán bánh trung thu biết rằng dịp tết trung thu số bánh có thể bán được có phân phối xác suất như sau

Số bánh bán được (x)	400	500	600	700	800	900
Xác suất	0,05	0,15	0,41	0,34	0,04	0,01

a. Tìm trung bình và độ lệch chuẩn của số bánh bán được.

b. Nếu cửa hàng đặt mua 600 chiếc thì xác suất bán hết bánh là bao nhiêu, xác suất còn thừa lại là bao nhiêu.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

c. Để có thể chắc chắn đến 95% là sẽ đủ bánh bán thì cửa hàng cần đặt mua bao nhiêu chiếc bánh.

2.75. Trong 900000 vé số phát hành thì có 20 giải trị giá 50 triệu, 150 giải trị giá 5 triệu và 1600 giải trị giá 1 triệu. Tìm số tiền lãi kỳ vọng của một người khi mua một vé, biết giá vé là 5 ngàn đồng.

2.76. Một nhà kinh doanh muốn đầu tư 10 triệu đồng vào một công ty mà nếu trong năm tới công ty làm ăn thuận lợi có thể sẽ mang lại lãi suất đến 14% còn nếu gặp khó khăn thì lãi suất có thể giảm đến mức 4%. Trong khi đó nếu gửi tiền vào ngân hàng thì lãi suất đảm bảo sau một năm là 8%. Vậy nếu dùng tiền để đầu tư thì khả năng có lãi hơn gửi ngân hàng là bao nhiêu.

ĐS: 0,6

2.77. Nhu cầu hàng ngày về một loại thực phẩm tươi sống có phân phối xác suất như sau:

Nhu cầu (kg)	30	31	32	33	34	35
Xác suất	0,15	0,2	0,35	0,15	0,1	0,15

Mỗi kg thực phẩm mua vào với giá 2,5 ngàn và bán ra với giá 4 ngàn. Nếu bị ế đến cuối ngày phải bán hạ giá còn 1,5 ngàn mới bán được hết. Vậy phải đặt mua hàng ngày bao nhiêu kg thực phẩm để có lãi nhất.

ĐS: 32kg

2.78. Hàm mật độ xác suất của biến ngẫu nhiên liên tục X có dạng

$$f(x) = \begin{cases} x^m \dfrac{e^{-x}}{m!} & \text{với } x \geq 0 \\ 0 & \text{với } x < 0 \end{cases}$$

Tìm E(X) và V(X).

Hướng dẫn: Sử dụng hàm Gam- ma

$$\int_0^{\infty} x^m e^{-x} dx = m!$$

E(X) = m + 1

V(X) = m + 1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.79. Biến ngẫu nhiên liên tục X có hàm phân bố xác suất

$$F(x) = \begin{cases} 1 - \dfrac{x^3}{x_0^3} & \text{với } x \geq x_0 \\ 0 & \text{với } x < x_0 \end{cases} \qquad (x_0 > 0)$$

Tìm E(X) và V(X)

ĐS: $E(X) = \dfrac{3}{2} x_0$ $\qquad V(X) = \dfrac{3}{4} x_0^2$

2.80. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất

$$f(x) = \begin{cases} (\sin x)/2 & \text{với } x \in (0, \pi) \\ 0 & \text{với } x \notin (0, \pi) \end{cases}$$

a. Tìm hàm phân bố xác suất F(x).

b. Tìm $P(0 < X < \pi/4)$

c. Tìm E(X)

ĐS: a

$$F(x) = \begin{cases} 0 & \text{với } x \leq 0 \\ \dfrac{1 - \cos x}{2} & \text{với } 0 < x \leq \pi \\ 1 & \text{với } x > \pi \end{cases}$$

b. $P(0 < X < \dfrac{\pi}{4}) = \dfrac{1}{2} - \dfrac{\sqrt{2}}{4}$

c. $E(X) = \dfrac{\pi}{2}$

2.81. Một công ty cung cấp nguyên vật liệu gửi 5 giấy đòi nợ tới một xí nghiệp yêu cầu thanh toán tiền cho 5 đợt giao hàng vừa qua với số lượng hàng của các đợt không khác nhau nhiều lắm. Trong số 5 giấy đòi nợ này (mỗi giấy viết riêng cho từng đợt) có 2 giấy ghi sai số tiền phải thanh toán. Do đến hạn phải trả nợ ngân hàng, công ty yêu cầu xí nghiệp thanh toán ngay cho ba đợt bất kỳ trong 5 đợt giao hàng ngày. Kế toán viên của xí nghiệp lấy ngẫu nhiên cùng một lúc ra 3 giấy để kiểm tra và làm các phiếu chi.

Tính xác suất để trong số 3 giấy lấy ra đó có ít nhất một giấy ghi sai số tiền phải thanh toán.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.82. Xí nghiệp và công ty nói ở bài trên thoả thuận với nhau rằng nếu kế toàn viên của xí nghiệp phát hiện thấy có giấy đòi nợ nào trong số 3 tờ giấy được lấy ra mà ghi sai số tiền thì xí nghiệp có quyền hoãn trả số nợ của đợt giao hàng đó. Mỗi giấy bị hoãn trả sẽ làm thiệt cho công ty khoảng 5 triệu đồng do phải trả lãi nợ quá hạn cho ngân hàng.

Hãy xác định số tiền thiệt hại trung bình có thể xảy ra đối với công ty do phải trả lãi nợ quá hạn.

2.83. Tuổi thọ (tính theo giờ) của một loại van điện lắp trong một thiết bị là biến ngẫu nhiên có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} 0 & \text{nếu } x \le 0 \\ 100/x^2 & \text{nếu } x > 100 \end{cases}$$

Tìm xác xuất để có 2 trong số 5 van điện này bị thay thế trong 150 giờ hoạt động đầu tiên biết rằng việc hỏng của các van điện là độc lập với nhau.

2.84. Tuổi thọ (tính theo giờ) của một trò chơi điện tử bấm tay là một biến ngẫu nhiên có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} ke^{-x/100} & \text{với } x \ge 0 \\ 0 & \text{với } x < 0 \end{cases}$$

Trong đó k là hằng số

Tính xác suất.

a. Tuổi thọ của trò chơi này nằm trong khoảng từ 50 đến 150 giờ

b. Tuổi thọ của trò chơi này ít hơn 100 giờ.

2.85. Tuỳ theo tình hình kinh tế trong nước mà trong năm tới một công ty thu được mức lãi (tính theo triệu đơn vị tiền tệ của nước này) khi đầu tư vào hai ngành A và B sẽ như sau:

Tình hình kinh tế / Mức lãi	Kém phát triển	Ổn định	Phát triển
Ngành A	20	80	120
Ngành B	-30	100	140

Theo dự báo thì xác suất để nền kinh tế của nước đang xét trong năm tới sẽ rơi vào các tình trạng tương ứng nêu trên là 0,3; 0,5 và 0,2.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vậy công ty nên đầu tư vào ngành nào để cho:

a. Mức lãi kỳ vọng là cao hơn.

b. Độ rủi ro (phương sai của mức lãi) là ít hơn?

2.86. Trong một cuộc thi, người ta có hai hình thức sau:

* Hình thức thứ nhất là mỗi người phải trả lời hai câu hỏi, mỗi câu trả lời đúng được 5 điểm.

* Hình thức thứ hai là nếu trả lời đúng câu thứ nhất mới được trả lời câu thứ hai, nếu không thì dừng. Trả lời đúng câu thứ nhất được 5 điểm, trả lời đúng câu thứ hai được 10 điểm.

Trong cả hai hình thức thi, các câu trả lời sai đều không được điểm. Giả sử xác suất trả lời đúng mỗi câu đều là 0,75; việc trả lời đúng mỗi câu là độc lập với nhau.

Theo bạn, nên chọn hình thức nào để số điểm trung bình đạt được nhiều hơn.

2.87. Chủ một cửa hàng sửa chữa điện dân dụng thuê 5 thợ sửa chữa làm việc 40 giờ một tuần với lương là 800 ngàn/tuần. Do nhu cầu sửa chữa tăng lên nên nhiều hợp đồng phải từ chối. Để xét xem có cần thuê thêm một thợ nữa không, ông chủ đã khảo sát nhu cầu X và thu được các số liệu sau:

Nhu cầu X	Xác suất P(x)
180 - 190	0,03
190 - 200	0,09
200 - 210	0,12
210 - 220	0,15
220 - 230	0,22
230 - 240	0,21
240 - 250	0,13
250 - 260	0,05

Nếu mỗi giờ sửa chữa chủ cửa hàng thu được 30 ngàn thì có nên thuê thêm 1 công nhân nữa không nếu:

a. Năm người công nhân cũ chỉ đồng ý làm đúng 40 giờ/tuần.

b. Năm người công nhân cũ đồng ý làm thêm tối đa mỗi người 5 giờ mỗi tuần với tiền công là 25 ngàn/giờ làm thêm.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.88. Biến ngẫu nhiên X có hàm mật độ xác suất

$$f(x) = \begin{cases} \dfrac{1}{\pi\sqrt{a^2 - x^2}} & \text{với } x \in (-a, a) \\ 0 & \text{với } x \notin (-a, a) \end{cases}$$

Tìm E(X)

ĐS: E(X) = 0

2.89. Biến ngẫu nhiên X có hàm phân bố xác suất

$$F(x) = \begin{cases} 0 & \text{với } x \le -2 \\ \dfrac{1}{2} + \dfrac{1}{\pi} \arcsin \dfrac{x}{2} & \text{với } -2 < x \le 2 \\ 1 & \text{với } x > 2 \end{cases}$$

a. Tìm P (-1 < X < 1)

b. Tìm hàm mật độ xác suất f(x)

ĐS: a. P = 1/3

b.

$$f(x) = \begin{cases} \dfrac{1}{\pi\sqrt{4 - x^2}} & \text{với } x \in (-2, 2) \\ 0 & \text{với } x \notin (-2, 2) \end{cases}$$

2.90. Trong lô xổ số phát hành có m_1 giải trị giá k_1, ..., m_n giải trị giá k_n. Tổng số có N vé. Vậy phải quy định giá vé là bao nhiêu để giải thưởng trung bình cho mỗi vé chỉ bằng nửa giá vé.

2.91*. Biến ngẫu nhiên X có hàm mật độ xác suất

$f(x) = ax^2e^{-kx}$ $(k > 0, 0 \le x < +\infty)$

a. Tìm hệ số a

b. Tìm hàm phân bố xác suất F(x)

ĐS: $a = \dfrac{k^3}{2}$

b. $F(x) = 1 - \dfrac{k^2x^2 + 2kx + 2}{2} e^{-kx}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.92*. Một ôtô khách chạy trên đoạn đường AB khoảng cách L theo nguyên tắc dừng bánh ở bất kỳ chỗ nào khách có yêu cầu. Biết rằng trên quãng đường đó đã có 1 khách lên xe và sau đó một đoạn đường đã xuống xe. Mật độ xác suất của việc lên xe ở điểm x ($0 \le x \le L$) tỷ lệ với giá trị $x(L - x)^2$ còn mật độ xác suất của việc xuống xe tại điểm y với điều kiện là khách đã lên xe tại điểm x ($x \le y < L$) tỷ lệ với giá trị $(y - x)^h$. $h \ge 0$. Tìm xác suất để:

a. Khách lên xe trước điểm Z.

b. Khách đã lên xe ở điểm x, sẽ xuống xe sau điểm Z.

ĐS: a. $P = \frac{Z^2}{L^4}(6L^2 - 8LZ + 3Z^2)$

b. $P = 1 - \left(\frac{Z - x}{L - x}\right)^{h+1}$

2.93*. Một khẩu súng được dùng để bắn thử với cùng một loạt đạn. Xác suất trúng đích của mỗi viên đạn là p. Phép thử kết thúc khi viên đạn đầu tiên trúng đích. Gọi Y là số viên đạn cần bắn (số lần bắn không hạn định).

a. Viết quy luật phân phối xác suất của Y.

b. Tìm hàm phân bố xác suất của Y.

ĐS: a.

Y	1	2	3	...	n	...
P	p	qp	q^2p		$q^{n-1}p$	...

$$F(y) = \begin{cases} 0 & \text{với } y \le 1 \\ p & \text{với } 1 < y \le 2 \\ p(1+q) & \text{với } 2 < y \le 3 \\ \cdots\cdots\cdots & \\ p\dfrac{1-q^n}{1-q} & \text{với } n < y \le n+ \\ 1 & \end{cases}$$

2.94*. Trong bình có a quả cầu trắng và b quả cầu đen. Một người lấy lần lượt từng quả cầu theo phương thức có hoàn lại cho đến khi lấy được cầu trắng. Tìm kỳ vọng toán và phương sai của số cầu đen được lấy ra.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ĐS: $E(X) = \frac{b}{a}$ $V(X) = \frac{b(a+b)}{a^2}$

2.95*. Tìm kỳ vọng toán và phương sai của số sản phẩm được sản xuất ra giữa hai lần sửa chữa của một máy nếu xác suất làm ra phế phẩm của máy đó bằng p và máy sẽ được sửa chữa khi làm ra k phế phẩm.

Hướng dẫn: Sử dụng kết quả bài 2.74.

$$E(X) = \frac{k}{p}$$

$$V(X) = \frac{k(1-p)}{p^2}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Lê Văn Phong. *Giáo trình lý thuyết xác suất*. NXB ĐH Kinh tế quốc dân. Hà nội. 2007.
2. Nguyễn Đình Cử, Nguyễn Cao Văn. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Thống kê. Hà Nội. 1991.
3. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Khoa học và Kỹ thuật. Hà nội.1999.
4. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà nội.2009.
5. Гмурман В.Е. Теория вероятностей и математическая статистика. Москва. 1972.
6. Гнеденко Б. В. Теория вероятностей. Москва. 1969.
7. Baillageon G et Rainville J. Introduction à la Statistique appliquée. Québec. 1976
8. Giard V. Statistique appliquée à la gestion. Economica. Paris. 1992
9. Amir D Aczel. Complete Business Statistics. Irwin Boston 1993.
10. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992
11. De groot M. Probability and Statistics. New york. 1989

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 3

MỘT SỐ QUY LUẬT PHÂN PHỐI XÁC SUẤT THÔNG DỤNG

1. MỞ ĐẦU

Trong thực tế, mỗi biến ngẫu nhiên đều tuân theo một quy luật phân phối xác suất nào đó và ngược lại mỗi quy luật phân phối xác suất lại tương ứng với một lớp các biến ngẫu nhiên. Vì vậy việc phân loại các biến ngẫu nhiên theo từng quy luật phân phối xác suất sẽ giúp cho việc sử dụng chúng được dễ dàng hơn. Ở chương này ta sẽ nghiên cứu một số quy luật phân phối xác suất thông dụng nhất với các biến ngẫu nhiên rời rạc và liên tục. Mỗi quy luật phân phối xác suất sẽ được giới thiệu từ dạng phân phối xác suất lý thuyết của nó cho đến cách tính các tham số đặc trưng cơ bản và đặc điểm áp dụng của nó trong các tình huống thực tế. Từ mục 2 đến mục 5 sẽ nghiên cứu các quy luật phân phối xác suất của các biến ngẫu nhiên rời rạc còn từ mục 6 đến mục 11 sẽ nghiên cứu các quy luật phân phối xác suất của các biến ngẫu nhiên liên tục. Đây là các quy luật phân phối xác suất được sử dụng nhiều trong kinh tế và xã hội.

Để làm rõ những đặc điểm cơ bản của mỗi quy luật phân phối xác suất, ta sẽ xuất phát từ các thí dụ có tính điển hình cho mỗi quy luật để làm cơ sở xây dựng những lược đồ khác nhau, từ đó đi đến các quy luật phân phối xác suất tương ứng với mỗi lược đồ.

Giả sử trong bình có N quả cầu trong đó có M quả cầu trắng và N-M quả cầu đen. Mỗi phép thử là việc lấy ngẫu nhiên từ bình ra một quả cầu. Theo những cách lấy khác nhau sẽ dẫn đến những lược đồ khác nhau và các quy luật phân phối xác suất khác nhau.

2. QUY LUẬT KHÔNG – MỘT (Quy luật Bernoulli)-A(p)

Giả sử từ bình lấy ngẫu nhiên một quả cầu. Như vậy trong phép thử này chỉ có 2 biến cố có thể xảy ra: Hoặc lấy được cầu trắng (biến cố A) hoặc lấy được cầu đen (biến cố A không xảy ra tức biến cố $\overline{A}$ xảy ra). Xác suất để A

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

xảy ra (lấy được cầu trắng) bằng $p = \frac{M}{N}$. Như vậy, xác suất để $\overline{A}$ xảy ra (lấy được cầu đen) bằng $q = \frac{N-M}{N} = 1 - \frac{M}{N} = 1 - p$.

Một cách tổng quát, giả sử ta tiến hành một phép thử, trong đó biến cố A có thể xảy ra với xác suất bằng p. Gọi X là số lần xuất hiện biến cố A trong phép thử đó. Như vậy X là biến ngẫu nhiên rời rạc với hai giá trị có thể có bằng 0 (nếu biến cố không xuất hiện) hoặc bằng 1 (nếu biến cố xuất hiện). Hiển nhiên là xác suất để biến ngẫu nhiên X nhận một trong hai giá trị có thể có nói trên có thể biểu thị bằng công thức:

$$P_x = p^x q^{1-x} \quad \text{với } x=0;1 \qquad (3.1)$$

2.1. Định nghĩa

Biến ngẫu nhiên rời rạc X nhận một trong hai giá trị có thể có X= 0;1 với các xác suất tương ứng được tính bằng công thức (3.1) gọi là phân phối theo quy luật không- một (Quy luật Bernoulli) với tham số là p.

Quy luật không - một được ký hiệu là A(p)

Như vậy, bảng phân phối xác suất của biến ngẫu nhiên X phân phối theo quy luật không- một có dạng.

X	0	1
P	q	p

$(q = 1 - p)$

2.2. Các tham số đặc trưng của quy luật không - một

Theo bảng phân phối xác suất của X ta có:

$$E(X) = 0.q + 1.p = p$$

Như vậy $E(X) = p$.

Để tìm phương sai, trước hết ta tìm $E(X^2)$:

$$E(X^2) = 0^2.q + 1^2.p = p$$

Từ đó $$V(X) = p - p^2 = p(1-p) = pq$$

Như vậy:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$V(X) = pq$$

Suy ra độ lệch chuẩn

$$\sigma_x = \sqrt{V(X)} = \sqrt{pq}$$

Trong thực tế quy luật không - một thường được dùng để đặc trưng cho các dấu hiệu nghiên cứu định tính có hai phạm trù luân phiên. Chẳng hạn, khi muốn nghiên cứu giới tính của khách hàng ta có thể đặc trưng cho giới tính bằng biến ngẫu nhiên với 2 giá trị bằng 0 (Nam) và bằng 1 (Nữ). Lúc đó, xác suất p sẽ đặc trưng cho tỉ lệ khách hàng nữ trong tập hợp khách hàng. Nếu dấu hiệu định tính có nhiều hơn hai phạm trù thì có thể dùng nhiều biến ngẫu nhiên phân phối "không - một" cùng một lúc.

Về mặt lý thuyết quy luật không - một có thể được dùng làm cơ sở để tìm quy luật phân phối xác suất của các biến ngẫu nhiên khác.

Bài tập

3.1. Gieo một con xúc xắc đối xứng và đồng chất. Tìm quy luật phân phối xác suất và các tham số đặc trưng của số lần xuất hiện mặt 6 chấm.

ĐS: Gọi X là số lần xuất hiện mặt 6 chấm khi gieo con xúc xắc.

$$X \sim A(p = \frac{1}{6}) \Rightarrow E(X) = \frac{1}{6}; V(X) = \frac{5}{36}$$

3.2. Một lô hàng có 80% chính phẩm và 20% phế phẩm. Lấy ngẫu nhiên một sản phẩm. Tìm quy luật phân phối xác suất và các tham số đặc trưng của số chính phẩm được lấy ra.

ĐS: Gọi X là số chính phẩm được lấy ra khi lấy ngẫu nhiên một sản phẩm.

$$X \sim A(p = 0,8) \Rightarrow E(X) = 0,8; V(X) = 0,16$$

3.3. Điều tra ý kiến của khách hàng đối với sản phẩm của doanh nghiệp thì thấy có 60% khách hàng thích sản phẩm đó. Tìm quy luật phân phối xác suất và các tham số đặc trưng của thái độ ưa thích của khách hàng đối với sản phẩm.

3.4. Điều tra ý kiến của người dân đối với một quyết định của chính phủ thì thấy có 70% số người được điều tra ủng hộ quyết định đó. Tìm quy luật phân phối xác suất và các tham số đặc trưng của thái độ ủng hộ của dân chúng đối với quyết định nói trên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3. QUY LUẬT NHỊ THỨC - B(n,p)

Ta quay trở lại thí dụ về lô cầu.

Giả sử từ lô cầu gồm M cầu trắng và N - M cầu đen, lấy lần lượt ra n quả cầu theo phương thức hoàn lại. Nếu lấy theo phương thức này thì n phép thử nói trên sẽ độc lập với nhau, vì việc lấy được cầu trắng hoặc cầu đen trong mỗi lần lấy không ảnh hưởng đến khả năng lấy được cầu trắng hoặc đen trong các lần lấy khác. Trong mỗi lần lấy chỉ có hai trường hợp đối lập xảy ra: Hoặc lấy được cầu trắng (biến cố A), hoặc lấy được cầu đen (biến cố $\overline{A}$). Xác suất lấy được cầu trắng mỗi lần đều bằng $p = \frac{M}{N}$ và xác suất lấy được cầu đen mỗi lần cũng đều bằng $\frac{N-M}{N} = 1 - \frac{M}{N} = 1 - p = q$. Như ở chương I đã trình bày, những điều kiện đó sẽ dẫn đến một lược đồ gọi là lược đồ Bernoulli.

Một cách tổng quát, giả sử ta có một lược đồ Bernoulli, tức là tiến hành n phép thử độc lập, trong mỗi phép thử chỉ có hai trường hợp, hoặc biến cố A xuất hiện, hoặc biến cố A không xuất hiện, xác suất xuất hiện biến cố A trong mỗi phép thử đều bằng p như vậy xác suất A không xuất hiện trong mỗi phép thử đều bằng $q = 1 - p$.

Gọi X là "Số lần xuất hiện biến cố A trong n phép thử độc lập" nói trên thì X là biến ngẫu nhiên rời rạc với các giá trị có thể có $X = 0, 1, ..., n$. Như đã chứng minh ở chương I, xác suất để X nhận các giá trị tương ứng được tính bằng công thức Bernoulli:

$$P_x = C_n^x p^x q^{n-x} \quad x = 0, 1, ..., n \tag{3.2}$$

3.1. Định nghĩa

Biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có $X = 0, 1, ..., n$ với các xác suất tương ứng được tính bằng công thức (3.2) gọi là phân phối theo quy luật nhị thức với các tham số là n và p.

Quy luật nhị thức được ký hiệu là B(n,p).

Như vậy, bảng phân phối xác suất của biến ngẫu nhiên X phân phối theo quy luật nhị thức có dạng:

X	0	1	...	x	...	n
P	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	...	$C_n^x p^x q^{n-x}$	...	$C_n^n p^n q^0$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trong thực tế đôi khi ta phải tính xác suất để biến ngẫu nhiên X phân phối theo quy luật nhị thức nhận giá trị trong một khoảng $[x, x+h]$ trong đó h là một số nguyên dương $(h \le n-x)$. Lúc đó ta có thể tính xác suất này theo công thức:

$$P(x \le X \le x+h) = p_x + p_{x+1} + ... + p_{x+h} \qquad (3.3)$$

Trong đó mỗi xác suất thành phần được tính bằng công thức (3.2). Thật vậy, biến cố $(x \le X \le x+h)$ có thể tách ra thành tổng của $h+1$ biến cố xung khắc từng đôi là $(X = x)$, $(X = x+1), ..., (X = x+h)$, do đó áp dụng định lý cộng xác suất với các biến cố đó ta có:

$$P(x \le X \le x+h) = P(X = x) + P(X = x+1) + ... +$$

$$+ ... + P(X = x+h) = p_x + p_{x+1} + ... + p_{x+h}$$

Giữa các xác suất P_x và P_{x-1} có mối liên hệ truy chứng sau đây:

$$P_x = \frac{p(n-x+1)}{qx} P_{x-1} \qquad (3.4)$$

Thật vậy, xét tỷ số:

$$\frac{P_x}{P_{x-1}} = \frac{\frac{n!}{x!(n-x)!} p^x q^{n-x}}{\frac{n!}{(x-1)!(n-x+1)!} p^{x-1} q^{n-x+1}}$$

Sau khi giản ước ta thu được (3.4). Quan hệ truy chứng này có thể làm cho việc tính toán theo công thức (3.3) được dễ dàng hơn.

Các xác suất theo công thức (3.2) được tính sẵn thành bảng (Phụ lục 1).

Thí dụ 1. Một phân xưởng có 5 máy hoạt động độc lập. Xác suất để trong một ngày mỗi máy bị hỏng đều bằng 0,1. Tìm xác suất để:

a. Trong một ngày có 2 máy hỏng.

b. Trong một ngày có không quá 2 máy hỏng.

Giải. Nếu coi sự hoạt động của mỗi máy là một phép thử, ta có 5 phép thử độc lập. Trong mỗi phép thử chỉ có hai trường hợp: hoặc máy hỏng, hoặc máy tốt. Xác suất hỏng của mỗi máy đều bằng 0,1. Như vậy, ta có một lược đồ Bernoulli. Gọi X là số máy hỏng trong ngày, X phân phối theo quy luật nhị thức với các tham số n = 5 và p = 0,1.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Do đó xác suất để trong một ngày có 2 máy hỏng chính là xác suất để $X = 2$.

Theo công thức (3.2) ta có:

$$P_2 = C_5^2(0,1)^2.(0,9)^3 = 0,0729$$

Xác suất để trong một ngày có không quá 2 máy hỏng là xác suất để X nhận giá trị trong khoảng [0;2]. Theo công thức (3.3) ta có:

$$P(0 \le X \le 2) = p_0 + p_1 + p_2$$

$$P_0 = C_5^0(0,1)^0.(0,9)^5 = 0,59049$$

$$P_1 = C_5^1(0,1)^1.(0,9)^4 = 0,32805$$

Vậy: $P(0 \le X \le 2) = 0,59049 + 0,32805 + 0,0729 = 0,99144$

3.2. Các tham số đặc trưng của quy luật nhị thức

Ta sẽ chứng minh rằng nếu biến ngẫu nhiên X phân phối theo quy luật nhị thức thì kỳ vọng toán:

$$E(X) = np \qquad (3.5)$$

và phương sai:

$$V(X) = npq \qquad (3.6)$$

Thật vậy, gọi $X_i (i = \overline{1,n})$ là số lần xuất hiện biến cố A trong phép thử thứ i. Lúc đó, do các phép thử tiến hành độc lập, các biến ngẫu nhiên X_i độc lập với nhau và mỗi X_i đều phân phối theo quy luật không - một với tham số là p. Như vậy, số lần xuất hiện biến cố A trong n phép thử X bằng:

$$X = \sum_{i=1}^{n} X_i$$

Theo tính chất của kỳ vọng toán và phương sai ta có:

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

Và $$V(X) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$$

Vì $X_i (i = \overline{1,n})$ cùng phân phối theo quy luật không - một với tham số là p, do đó:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$E(X_i) = p \qquad i = \overline{1,n}$$

và $$V(X_i) = pq \qquad i = \overline{1,n}$$

Từ đó:

$$E(X) = \sum_{i=1}^{n} E(X_i) = np$$

và $$V(X) = \sum_{i=1}^{n} V(X_i) = npq$$

Như vậy, độ lệch chuẩn $\sigma_x = \sqrt{V(X)} = \sqrt{npq}$.

Thí dụ 2. Một nhân viên bán hàng mỗi ngày đi chào hàng ở 10 nơi với xác suất bán được hàng mỗi nơi là 0,2. Vậy nếu một năm người đó đi bán hàng 300 ngày thì trung bình sẽ có khoảng bao nhiêu ngày người đó bán được hàng.

Giải. Dễ thấy xét trong một ngày nào đó ta có một lược đồ Bernoulli và gọi X là số lần bán được hàng trong ngày của người đó thì X tuân theo quy luật nhị thức với $n = 10$ và $p = 0,2$. Vậy xác suất để người đó bán được hàng trong một ngày bằng:

$$p = P(X \geq 1) = 1 - P(X = 0) = 1 - 0,8^{10} = 0,8926$$

Tương tự nếu xét trong một năm ta cũng có một lược đồ Bernoulli và gọi Y là số ngày người ấy bán được hàng trong một năm thì Y tuân theo quy luật nhị thức với n = 300 và p = 0,8926. Vậy số ngày trung bình trong một năm mà người ấy bán được hàng chính là kỳ vọng toán:

$$E(X) = np = 300.0,8926 = 267,78 \text{ ngày}$$

Ngoài kỳ vọng toán, phương sai và độ lệch chuẩn, trong quy luật nhị thức tham số mốt cũng hay được dùng.

Nếu X phân phối theo quy luật nhị thức thì Mốt m_0 có thể tìm trực tiếp từ bảng phân phối xác suất bằng cách tìm trong số các giá trị có thể có của X giá trị tương ứng với xác suất lớn nhất. Tuy nhiên, có thể tìm mốt mà không cần phải xây dựng bảng phân phối xác suất. Nó được xác định bằng công thức sau:

$$np - q \leq m_0 \leq np + p \qquad (3.7)$$

Thật vậy, vì mốt là giá trị có xác suất lớn nhất trong phân phối, do đó xác

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

suất tương ứng với giá trị m_0 phải không nhỏ hơn xác suất tương ứng với các giá trị cạnh nó là $m_0 - 1$ và $m_0 + 1$. Do đó ta có các bất đẳng thức sau:

$$P_{m_0} \geq P_{m_0-1}$$

Và

$$P_{m_0} \geq P_{m_0+1}$$

Thay các biểu thức của xác suất vào các bất đẳng thức trên theo công thức Bernoulli:

$$C_n^{m_0} p^{m_0} q^{n-m_0} \geq C_n^{m_0-1} p^{m_0-1} q^{n-m_0+1}$$

hay:

$$\frac{n!}{m_0!(n-m_0)!}.p^{m_0}q^{n-m_0} \geq \frac{n!}{(m_0-1)!(n-m_0+1)!}.p^{m_0-1}q^{n-m_0+1}$$

Từ đó sau một vài phép biến đổi đơn giản, ta có:

$$m_0 \leq np + p$$

Tương tự ta có:

$$C_n^{m_0} p^{m_0} q^{n-m_0} \geq C_n^{m_0+1} p^{m_0+1} q^{n-m_0-1}$$

hay:

$$\frac{n!}{m_0!(n-m_0)!}.p^{m_0}q^{n-m_0} \geq \frac{n!}{(m_0+1)!(n-m_0-1)!}.p^{m_0+1}q^{n-m_0-1}$$

Từ đó ta có: $m_0 \geq np - q$

Kết hợp hai kết quả vừa thu được ta có công thức xác định mốt m_0.

Ta chú ý rằng, vì trong quy luật nhị thức mốt phải là một giá trị nguyên, do đó có thể xảy ra hai trường hợp: Nếu $np + p$ là một số nguyên thì $np - q$ cũng là một số nguyên, lúc đó mốt sẽ cùng một lúc nhận hai giá trị $m_0 = np + p$ và $m_0 = np - p$. Còn nếu $np + p$ là một số thập phân thì mốt sẽ là giá trị nguyên nằm trong khoảng hai số thập phân là $np + p$ và $np - q$.

Thí dụ 3. Xác suất để mỗi con lợn khi tiêm phòng bằng một loại văcxin được miễn dịch là 0,9. Có 50 con lợn được tiêm phòng. Tìm số lợn được miễn dịch có khả năng nhiều nhất.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Bài toán thoả mãn lược đồ Bernoulli, do đó nếu gọi X là số lợn được miễn dịch thì X là biến ngẫu nhiên phân phối theo quy luật nhị thức.

Vậy số lợn được miễn dịch có khả năng xảy ra nhiều nhất chính là giá trị mốt. Theo công thức mốt ta có:

$$np - q \le m_0 \le np + p$$

với $n = 50, p = 0,9, q = 1 - 0,9 = 0,1$

$$50.0,9 - 0,1 \le m_0 \le 50.0,9 + 0,9$$

$$44,9 \le m_0 \le 45,9$$

Vậy $m_0 = 45$ tức là số lợn miễn dịch có khả năng nhiều nhất là 45 con.

Như đã trình bày ở trên, mối liên hệ giữa quy luật nhị thức và quy luật không - một được thể hiện như sau: Nếu $X_1X_2...X_n$ là các biến ngẫu nhiên độc lập lẫn nhau và cùng phân phối theo quy luật không - một với tham số là p thì tổng của các biến ngẫu nhiên đó sẽ là biến ngẫu nhiên tuân theo quy luật nhị thức với tham số là n và p.

Mặt khác, nếu X_1 và X_2 là các biến ngẫu nhiên độc lập và cùng phân phối nhị thức với tham số tương ứng là n_1, p và n_2, p thì tổng $X = X_1 + X_2$ cũng sẽ phân phối nhị thức với các tham số là $n_1 + n_2$ và p (xem thêm mục 10 chương IV).

3.3. Quy luật phân phối xác suất của tần suất

Trong thực tế nhiều khi người ta quan tâm đến tỷ lệ xuất hiện biến cố A trong lược đồ Bernoulli hơn là bản thân số lần xuất hiện biến cố đó. Để làm điều đó có thể biến đổi biến ngẫu nhiên X thành tần suất xuất hiện biến cố A trong n phép thử độc lập qua phép chia:

$$f = \frac{X}{n}$$

Chú ý rằng việc nhân biến ngẫu nhiên cho một hằng số không làm thay đổi phân phối xác suất của biến ngẫu nhiên đó mà chỉ dẫn đến sự thay đổi của các tham số đặc trưng mà thôi. Vì vậy, tần suất f vẫn phân phối theo quy luật nhị thức với tham số là n và p. Lúc đó, bảng phân phối xác suất của f có dạng:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

f	0	$\frac{1}{n}$	...	$\frac{x}{n}$	...	1
P	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$		$C_n^x p^x q^{n-x}$		$C_n^n p^n q^0$

Lúc đó, các tham số đặc trưng của biến ngẫu nhiên f như sau:

$$E(f) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{np}{n} = p$$

Vậy: $E(f) = p$ (3.8)

Và $$V(f) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{npq}{n^2} = \frac{pq}{n} \quad (3.9)$$

Từ đó: $$\sigma_f = \frac{\sqrt{pq}}{\sqrt{n}} \quad (3.10)$$

Quy luật phân phối xác suất của tần suất thường được gọi là quy luật nhị thức theo tỷ lệ.

Bài tập

3.5. Bắn 5 viên đạn vào mục tiêu. Xác suất trúng đích của mỗi lần bắn như nhau và bằng 0,2. Muốn phá huỷ mục tiêu phải có ít nhất 3 viên trúng mục tiêu. Tìm xác suất mục tiêu bị phá huỷ.

ĐS: p = 0,0579

3.6. Một gia đình có 5 con. Tìm xác suất sao cho trong số đó có:

a. 2 con trai

b. Không quá 2 con trai.

Giả thiết xác suất sinh con trai là 0,51.

3.7. Thống kê cho thấy cứ chào hàng 3 lần thì có 1 lần bán được hàng. Nếu chào hàng 12 lần và gọi X là số lần bán được hàng thì X tuân theo quy luật gì? Tại sao?

3.8. Một nữ công nhân quản lý 12 máy dệt. Xác suất để mỗi máy trong khoảng thời gian t cần đến sự chăm sóc của nữ công nhân là 1/3. Tính xác suất:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Trong khoảng thời gian t có 4 máy cần đến sự chăm sóc của nữ công nhân.

b. Trong khoảng thời gian t có từ 3 đến 6 máy cần đến sự chăm sóc của nữ công nhân.

ĐS: a. P = 0,238

b. P = 0,751

3.9. Việc sản xuất ra các sản phẩm được tiến hành độc lập. Hỏi phải sản xuất mỗi đợt bao nhiêu sản phẩm để trung bình có được 10 sản phẩm đạt tiêu chuẩn biết rằng xác suất được sản phẩm đạt tiêu chuẩn là 0,8.

ĐS: n = 13 sản phẩm

3.10. Trong một lô hàng có 800 sản phẩm loại 1 và 200 sản phẩm loại 2. Lấy ngẫu nhiên ra 5 sản phẩm theo phương thức có hoàn lại. Gọi X là số sản phẩm loại 1 lấy được.

a. X tuân theo quy luật gì? Viết biểu thức xác suất tổng quát của quy luật.

b. Tìm E(X) và V(X)

c. Tìm số sản phẩm loại 1 trung bình được lấy ra và tính khả năng để xảy ra điều đó.

ĐS: a. X tuân theo quy luật nhị thức với các tham số n = 5 và p = 0,8.

$P_x = C_5^x (0,8)^x (0,2)^{5-x}$; $x = \overline{0,5}$

b. E(X) = 4; V(X) = 0,8

c. $\overline{X} \approx 4$; $P(X = \overline{X}) = 0,4096$

3.11. Xác suất để sản phẩm sản xuất ra bị hỏng bằng 0,1.

a. Tìm xác suất để trong 5 sản phẩm sản xuất ra có không quá 2 sản phẩm hỏng.

b. Tìm số sản phẩm hỏng trung bình trong 5 sản phẩm đó.

c. Tìm số sản phẩm hỏng có khả năng xảy ra nhiều nhất.

3.12. Gieo 10.000 hạt giống với xác suất để mỗi hạt giống nảy mầm là 0,85. Gọi X là số hạt nảy mầm.

a. X tuân theo quy luật gì?

b. Tìm E(X) và V(X).

3.13. Xác suất để mỗi hành khách chậm tàu là 0,02. Tìm số khách chậm

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

có khả năng xảy ra nhiều nhất trong 855 hành khách.

ĐS: 17 hành khách.

3.14. Xác suất để khỏi bệnh khi dùng loại thuốc A là 3/4. Có 5 người mắc bệnh B dùng thuốc A. Tìm xác suất:

a. Có 3 người khỏi bệnh.

b. Có ít nhất 1 người khỏi bệnh.

c. Có nhiều nhất 2 người khỏi bệnh.

ĐS: a. P = 0,263; b. P = 0,999902; c. P = 0,1035

3.15. Tìm phương sai của biến ngẫu nhiên X là số lần xuất hiện biến cố A trong hai phép thử độc lập và P(A) = p trong mỗi phép thử, biết rằng E(X) = 1,2.

ĐS: V(X) = 0,48.

3.16. Tiến hành các phép thử độc lập với xác suất xuất hiện biến cố trong mỗi phép thử đều bằng p. Tìm p nếu phương sai của số lần xuất hiện biến cố trong ba phép thử độc lập là 0,63.

3.17. Một kho hàng chuyên cung cấp hàng cho 12 cửa hàng. Xác suất để mỗi cửa hàng đặt hàng cho kho đó trong ngày là 0,3. Tìm số đơn đặt hàng có khả năng nhiều nhất cho một ngày và xác suất tương ứng với nó.

ĐS: 3 đơn đặt hàng; P = 0,2397.

3.18. Bia bắn được chia làm 2 vòng, xác suất bắn trúng vòng trong là 0,7 còn trúng vòng ngoài là 0,3. Tìm xác suất sao cho bắn 3 viên đạn thì được ít ra là 29 điểm biết rằng bắn trúng vòng trong thì được 10 điểm, trúng vòng ngoài thì được 9 điểm.

ĐS: P = 0,784.

3.19. Một nghiên cứu cho thấy 70% công chức cho rằng việc nghỉ làm hai ngày một tuần sẽ nâng cao được hiệu suất công tác. Nếu chọn ngẫu nhiên 15 công chức ở một bộ để phỏng vấn thì xác suất để có ít nhất 10 người đồng ý với ý kiến trên là bao nhiêu?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4. QUY LUẬT POISSON - P(λ)

Giả sử tiến hành n phép thử độc lập, trong mỗi phép thử xác suất để biến cố A xảy ra đều bằng p và không xảy ra đều bằng $q = 1 - p$. Lúc đó, nếu gọi X là số lần xuất hiện biến cố A trong n phép thử đó thì X phân phối theo quy luật nhị thức và xác suất để X nhận một trong các giá trị có thể có của nó được tính bằng công thức Bernoulli. Tuy nhiên, nếu số phép thử n quá lớn mà xác suất p lại quá nhỏ thì việc tính toán sẽ gặp nhiều khó khăn. Vì vậy, trong trường hợp này (n lớn, p nhỏ) người ta sử dụng công thức xấp xỉ Poisson.

Như vậy, trong một số rất lớn phép thử độc lập mà xác suất xuất hiện biến cố A trong mỗi phép thử lại rất nhỏ ta phải tìm xác suất để biến cố A xuất hiện đúng x lần.

Giả sử tích np luôn luôn bằng một giá trị không đổi $np = \lambda$, lúc đó công thức Bernoulli có thể viết như sau:

$$p_x = \frac{n(n-1)(n-2)...[n-(n-1)]}{x!} p^x q^{n-x}$$

Vì $np = \lambda$ nên $p = \frac{\lambda}{n}$ do đó:

$$p_x = \frac{n(n-1)(n-2)...[n-(x-1)]}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$=\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{x-1}{n}\right)\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

Vì n lớn, do đó thay cho p_x ta tìm $\lim\limits_{n\to\infty} p_x$. Lúc đó, mỗi thừa số $\left(1-\frac{1}{n}\right)$ với $i = \overline{1, x-1}$ đều tiến tới 1, còn giới hạn

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-x} = e^{-\lambda}$$

Do đó:
$$\lim_{n\to\infty} p_x = \frac{\lambda^x}{x!} e^{-\lambda}$$

Như vậy là trong trường hợp số phép thử n rất lớn, xác suất p rất nhỏ và

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

tích $np = \lambda$ không đổi, các xác suất p_x của công thức Bernoulli có thể thay thế bằng công thức xấp xỉ Poisson sau đây:

$$P_x = \frac{\lambda^x}{x!} e^{-\lambda} \quad x = 0,1,2,... \tag{3.11}$$

Trong thực tế, công thức Poisson có thể dùng thay cho công thức Bernoulli nếu thoả mãn điều kiện $n \geq 20$ và $p \leq 0,1$ tức là $np \approx npq$.

Một cách tổng quát quy luật Poisson được định nghĩa như sau:

4.1. Định nghĩa

Biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có X = 0, 1, ... với các xác suất tương ứng được tính bằng công thức (3.11) gọi là phân phối theo quy luật Poisson với tham số là λ.

Quy luật Poisson được ký hiệu là $P(\lambda)$.

Như vậy, bảng phân phối xác suất của biến ngẫu nhiên X phân phối theo quy luật Poisson có dạng:

X	0	1	...	x	...
P	$e^{-\lambda}.\frac{\lambda^0}{0!}$	$e^{-\lambda}.\frac{\lambda^1}{1!}$	...	$e^{-\lambda}.\frac{\lambda^x}{x!}$	...

Nếu phải tìm xác suất để trong n phép thử biến ngẫu nhiên X phân phối theo quy luật Poisson nhận giá trị trong khoảng [x, x + h] trong đó h là số nguyên dương tuỳ ý thì xác suất này được tính bằng công thức:

$$P(x \leq X \leq x+h) = P_x + P_{x+1} + ... + P_{x+h} \tag{3.12}$$

Trong đó mỗi xác suất thành phần được tính bằng công thức (3.11).

Giữa các xác suất P_x và P_{x-1} có mối quan hệ truy chứng sau:

$$P_x = \frac{\lambda}{x} P_{x-1} \tag{3.13}$$

Thật vậy, xét tỷ số:

$$\frac{P_x}{P_{x-1}} = \frac{\frac{e^{-\lambda}\lambda^x}{x!}}{e^{-\lambda}\frac{\lambda^{x-1}}{(x-1)!}} = \frac{\lambda}{x}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đó suy ra (3.13).

Các xác suất P_x theo công thức (3.11) được tính sẵn thành bảng (Phụ lục 2).

Thí dụ 1. Một máy dệt có 5000 ống sợi, xác suất để trong một phút mỗi ống sợi bị đứt bằng 0,0002. Tìm xác suất để trong một phút có không quá 2 ống sợi bị đứt.

Giải. Bài toán thỏa mãn lược đồ Bernoulli song vì n=5000 rất lớn, p=0,0002 quá nhỏ và tích $np = 5000 \times 0,0002 = 1 \approx npq$ không đổi, do đó nếu gọi X là số ống sợi bị đứt trong một phút thì X là biến ngẫu nhiên rời rạc và có thể coi như phân phối theo quy luật Poisson. Xác suất để số ống sợi bị đứt trong một phút không quá 2 chính là xác suất để X nhận giá trị trong khoảng $[0,2]$. Theo công thức (3.12) ta có:

$$P(0 \le X \le 2) = P_0 + P_1 + P_2$$

$$P_0 = e^{-\lambda}\frac{\lambda^0}{0!} = (2,71)^{-1}\frac{1^0}{0!} = (2,71)^{-1}$$

$$P_1 = \frac{1}{1}P_0 = (2,71)^{-1}$$

$$P_2 = \frac{1}{2}P_1 = \frac{1}{2}(2,71)^{-1}$$

Do đó:

$$P(0 \le X \le 2) = (2,71)^{-1}\left[1+1+\frac{1}{2}\right] = \frac{2,5}{1,71} = 0,9225$$

4.2. Các tham số đặc trưng của quy luật Poisson

Giả sử X phân phối theo quy luật Poisson. Ta sẽ chứng minh rằng:

$$E(X) = \lambda \qquad (3.14)$$

Thật vậy, theo định nghĩa của kỳ vọng toán, ta có:

$$E(X) = \sum_{x=0}^{+\infty} xP_x = \sum_{x=0}^{+\infty} xe^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=1}^{+\infty}\frac{\lambda^x}{(x-1)!} = \lambda e^{-\lambda}\sum_{x=1}^{+\infty}\frac{\lambda^{x-1}}{(x-1)!}$$

Song ta lại có $\sum_{x=1}^{+\infty}\frac{\lambda^{x-1}}{(x-1)!} = e^{\lambda}$ do đó $E(X) = e^{-\lambda}.e^{\lambda}.\lambda = \lambda$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bằng cách tính tương tự có thể tìm được $E(X^2) = \lambda^2 + \lambda$

Do đó: $V(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$

Vậy: $V(X) = \lambda$ (3.15)

Như vậy là trong quy luật Poisson cả kỳ vọng toán và phương sai đều bằng λ. Đó là tính chất đặc biệt của quy luật Poisson.

Bằng cách so sánh các xác suất như đã làm trong quy luật nhị thức có thể chứng minh rằng nếu X phân phối theo quy luật Poisson thì mốt được xác định bằng công thức:

$$\lambda - 1 \le m_0 \le \lambda$$

Ở đây cũng có thể xảy ra hai trường hợp: Nếu λ là một số nguyên thì mốt sẽ cùng một lúc nhận hai giá trị nguyên là $m_0 = \lambda - 1$ và $m_0 = \lambda$. Còn nếu λ là một số thập phân thì mốt sẽ là giá trị nguyên nằm trong khoảng của hai số thập phân là λ và $\lambda - 1$.

Thí dụ 2. Xác suất để trong khi vận chuyển mỗi chai rượu bị vỡ là 0,001. Người ta tiến hành vận chuyển 2000 chai rượu đến cửa hàng.

a. Tìm số chai vỡ trung bình khi vận chuyển;

b. Tìm số chai vỡ có khả năng nhiều nhất khi vận chuyển.

Giải. Bài toán thoả mãn lược đồ Bernoulli. Vì n = 2000 khá lớn, p = 0,001 khá nhỏ và tích $np = 2000 \times 0,001 = 2$ không đổi do đó nếu gọi X là số chai rượu bị vỡ khi vận chuyển thì X là biến ngẫu nhiên phân phối theo quy luật Poisson. Số chai vỡ trung bình chính là kỳ vọng toán của X. Ta có:

$$E(X) = \lambda = 2 \text{ chai.}$$

Số chai vỡ có khả năng xảy ra nhiều nhất là giá trị mốt $m_0 \Rightarrow \lambda - 1 \le m_0 \le \lambda$. Vì $\lambda = 2$ do đó mốt sẽ nhận hai giá trị là $m_0 = 2$ và $m_0 = 1$. Như vậy số chai vỡ có khả năng nhiều nhất là 1 và 2 chai.

Chú ý rằng nếu X_1 và X_2 là các biến ngẫu nhiên độc lập, X_1 phân phối $P(\lambda_1)$ còn X_2 phân phối $P(\lambda_2)$ thì lúc đó tổng của chúng là biến ngẫu nhiên $X = X_1 + X_2$ cũng sẽ phân phối Poisson với tham số là $\lambda_1 + \lambda_2$ (xem thêm mục 10 chương IV).

Quy luật Poisson có ứng dụng rộng rãi trong nhiều lĩnh vực thực tế như

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

kiểm tra chất lượng sản phẩm, lý thuyết phục vụ công cộng, lý thuyết quản lý dự trữ v.v... Trong lý thuyết phục vụ công cộng, người ta đề cập đến các hệ thống phục vụ dòng các yêu cầu đến như dòng người vào cửa hàng mậu dịch, dòng các con tàu đến cảng chờ bốc xếp, dòng xe ô tô vào một xưởng sửa chữa, dòng khách vào một cửa hàng cắt, sấy tóc v.v...

Trong nhiều trường hợp dòng các yêu cầu đó thường là dòng tối giản tức là thoả mãn các điều kiện dừng, không hậu quả và đơn nhất. Trong những dòng tối giản như vậy số yêu cầu đến hệ thống trong một khoảng thời gian nào đó thường phân phối theo quy luật Poisson, do đó ta có thể đánh giá được các tính chất của những dòng yêu cầu này, từ đó đưa ra những phương thức tổ chức hệ thống để đạt hiệu quả kinh tế và kỹ thuật cao nhất.

Bài tập

3.20. Một xe tải vận chuyển 1000 sản phẩm vào kho. Xác suất để khi vận chuyển mỗi sản phẩm bị hỏng là 0,004. Tìm xác suất để sau khi vận chuyển có 5 sản phẩm bị hỏng.

ĐS: P = 0,1562

3.21. Tổng đài điện thoại phục vụ 100 máy điện thoại. Xác suất để trong mỗi phút mỗi máy gọi đến tổng đài là 0,02. Tìm số máy gọi đến tổng đài trung bình trong 1 phút.

ĐS: 2 máy

3.22. Số khách hàng vào một cửa hàng bách hoá trong một giờ là biến ngẫu nhiên tuân theo quy luật Poisson với mật độ (số khách trung bình) là 8 khách hàng trong 1 giờ. Tìm xác suất để trong một giờ nào đó có hơn 4 khách vào.

ĐS: P = 0,9

3.23. Một lô hàng có tỷ lệ phế phẩm là 4%. Người ta kiểm tra 150 sản phẩm của lô hàng đó và nếu trong đó có không quá 2 phế phẩm thì lô hàng được chấp nhận. Tìm xác suất để lô hàng được chấp nhận.

3.24. Tại sân bay cứ 15 phút có một chuyến ô tô buýt loại 6 chỗ ngồi phục vụ chở khách vào trung tâm thành phố. Biết rằng số khách chờ đi ô tô phân

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

phối Poisson với mật độ trung bình là 8 người một giờ. Tìm xác suất để trong chuyến xe tiếp theo:

a. Không có khách nào chờ đi xe.

b. Xe sẽ chật khách.

c. Người ta sẽ tăng thêm một xe chở khách nữa nếu xác suất để có hơn một người phải chờ chuyến xe sau lớn hơn 0,1. Vậy có tăng thêm xe chở khách không?

3.25. Cứ 5000 con cá biển đánh bắt được thì có 1 con bị nhiễm khuẩn có hại cho sức khoẻ con người.

Tìm xác suất để trong một lô cá gồm 1800 mới đánh bắt về có không quá 2 con bị nhiễm khuẩn.

5. QUY LUẬT SIÊU BỘI - M(N,n)

Trong thực tế ta thường gặp trường hợp tiến hành n phép thử và trong mỗi phép thử cũng chỉ có hai trường hợp: hoặc A xảy ra hoặc A không xảy ra. Tuy nhiên, các phép thử này không tiến hành một cách độc lập với nhau. Do đó, xác suất để biến cố A xảy ra hoặc không xảy ra trong mỗi phép thử sẽ không bằng nhau nữa mà thay đổi từ phép thử này qua phép thử khác. Trong những lược đồ như vậy số lần xuất hiện biến cố A trong n phép thử sẽ không phân phối theo quy luật nhị thức hoặc quy luật Poisson, vì vậy không thể dùng công thức Bernoulli hoặc công thức Poisson để tìm xác suất để biến cố A xuất hiện x lần trong n phép thử nói trên. Một trong những lược đồ như vậy dẫn đến quy luật siêu bội.

Ta trở lại thí dụ về lô cầu. Giả sử trong bình có N quả cầu trong đó có M quả cầu trắng và $N-M$ quả cầu đen. Lấy ngẫu nhiên lần lượt ra n quả cầu theo phương thức không hoàn lại. Lúc đó các phép thử sẽ không độc lập với nhau nữa và xác suất để lấy được cầu trắng ở mỗi lần lấy sẽ thay đổi. Theo định nghĩa cổ điển về xác suất, xác suất để trong n quả cầu lấy ra có x quả cầu trắng được tính bằng công thức:

$$P_x = \frac{C_M^x . C_{N-M}^{n-x}}{C_N^n} \qquad (3.16)$$

Trong đó x có thể bằng 0, 1, ..., n. Từ đó ta có định nghĩa sau đây:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5.1. Định nghĩa

Biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có X = 0, 1, 2, ..., n với các xác suất tương ứng được tính bằng công thức (3.16) gọi là phân phối theo quy luật siêu bội với các tham số là N và n.

Quy luật siêu bội được ký hiệu là M(N, n).

Bảng phân phối xác suất của biến ngẫu nhiên X phân phối theo quy luật siêu bội có dạng:

X	0	1	...	x	...	n
P	$\frac{C_M^0.C_{N-M}^n}{C_N^n}$	$\frac{C_M^1.C_{N-M}^{n-1}}{C_N^n}$	...	$\frac{C_M^x.C_{N-M}^{n-x}}{C_N^n}$	...	$\frac{C_M^n.C_{N-M}^0}{C_N^n}$

Thí dụ 6. Trong cửa hàng có bán 100 bóng đèn trong đó có 5 bóng hỏng mà không kiểm tra thì không thể xác định được. Một người khách chọn ngẫu nhiên 2 bóng. Tìm xác suất để người đó mua được cả 2 bóng đều tốt.

Giải. Gọi X là số bóng tốt mà người đó có thể mua được. X phân phối theo quy luật siêu bội với N=100, M=95 và n=2. Xác suất để mua được cả 2 bóng tốt là xác suất để X=2. Theo công thức (3.16) ta có:

$$P_2 = \frac{C_{95}^2.C_5^0}{C_{100}^2} = \frac{95.94}{100.99} \approx 0,9$$

5.2. Các tham số đặc trưng của quy luật siêu bội

Có thể chứng minh được rằng nếu X phân phối theo quy luật siêu bội thì kỳ vọng toán:

$$E(X) = n\frac{M}{N} = np \qquad (3.17)$$

Và phương sai:

$$V(X) = n\frac{M}{N}.\frac{N-M}{N}.\frac{N-n}{N-1} = npq.\frac{N-n}{N-1} \qquad (3.18)$$

Giá trị $\frac{N-n}{N-1}$ được gọi là *hệ số hiệu chỉnh.*

Trong thực tế quy luật siêu bội được áp dụng để tính xác suất xuất hiện

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

x lần biến cố khi lấy một cách ngẫu nhiên n đơn vị từ một tập hợp nào đó theo phương thức *không hoàn lại*. Chẳng hạn, để kiểm tra chất lượng của một lô sản phẩm được sản xuất ra, người ta thường lấy ngẫu nhiên từ lô đó n sản phẩm theo phương thức không hoàn lại và đánh giá xác suất để trong đó có x phế phẩm hoặc chính phẩm.

Ta chú ý rằng khi số phép thử n là rất bé so với số N thì phân phối siêu bội thực tế không khác biệt so với phân phối nhị thức vì:

$$\lim_{\substack{n\\N}\to 0} \frac{C_M^x . C_{N-M}^{n-x}}{C_N^n} = C_n^x . \left(\frac{M}{N}\right)^x . \left(1-\frac{M}{N}\right)^{n-x}$$

Điều đó có nghĩa là khi tỷ số $\frac{n}{N}$ là rất nhỏ (tức là $\frac{n}{N} < 0,1$) thì phương pháp lấy không hoàn lại và có hoàn lại gần như không khác nhau. Do đó có thể lấy theo phương thức không hoàn lại song vẫn có thể tính toán như trong trường hợp lấy có hoàn lại cho đơn giản.

Bài tập

3.26. Trong kho có 10 cái lốp xe, trong đó có 3 cái hỏng. Lấy ngẫu nhiên 4 cái lốp để lắp cho một xe. Nếu gọi X là số lốp xe bị hỏng có thể được lấy ra thì X tuân theo quy luật nào? Hãy giải thích?

3.27. Trong số 20 công nhân của một công ty có 12 người có tay nghề khá. Tìm xác suất để kiểm tra ngẫu nhiên tay nghề của 5 công nhân thì có ít nhất 3 người có tay nghề khá.

3.28. Để thanh toán 1 triệu đồng tiền hàng, một khách hàng gian lận đã xếp lẫn 5 tờ 50 ngàn tiền giả với 15 tờ tiền thật. Chủ cửa hàng rút ngẫu nhiên 3 tờ giấy bạc đem đi kiểm tra và giao hẹn nếu phát hiện có bạc giả thì cứ mỗi tờ giả khách hàng phải đền hai tờ thật. Tìm số tiền phạt mà khách có thể phải trả.

ĐS: 75 ngàn.

3.29. Trong 20 giấy thông báo thuế thu nhập có 3 giấy mắc sai sót. Nhân viên kiểm tra lấy ngẫu nhiên 5 giấy thông báo để kiểm tra.

a. Thiết lập phân phối xác suất của số giấy thông báo có lỗi.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b. Tìm trung bình và phương sai của số giấy thông báo có lỗi được kiểm tra.

3.30. Trong 100 bóng đèn có 40 bóng hỏng. Tìm xác suất để lấy được 3 bóng hỏng trong 5 bóng được kiểm tra ngẫu nhiên.

Trên đây ta xét một số quy luật phân phối xác suất thông dụng nhất của các *biến ngẫu nhiên rời rạc*. Sau đây ta sẽ xét một số quy luật phân phối xác suất cơ bản của *các biến ngẫu nhiên liên tục* vì nhiều đại lượng cần nghiên cứu trong thực tế chính là các biến ngẫu nhiên liên tục. Do đó, việc hiểu biết các quy luật phân phối xác suất của chúng cho phép tiến hành phân tích một cách sâu sắc, cụ thể và chính xác hơn các hiện tượng này.

6. QUY LUẬT PHÂN PHỐI ĐỀU-U(a,b)

Quy luật phân phối đều là quy luật phân phối xác suất đơn giản nhất trong các quy luật phân phối xác suất của biến ngẫu nhiên liên tục. Nếu biến ngẫu nhiên X có thể nhận bất kỳ giá trị nào trên khoảng (a,b) với a và b là các số thực và ứng với mỗi giá trị là một mật độ xác suất như nhau thì biến X sẽ có phân phối đều.

Như vậy, trong khoảng (a,b) hàm mật độ xác suất của biến ngẫu nhiên phải bằng một giá trị xác định tức là f(x)=c trong khoảng (a,b). Từ đó, theo tính chất của hàm mật độ xác suất ta có:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{a}^{b} cdx = 1$$

Từ đó: cb-ca=1 suy ra $c = \frac{1}{b-a}$. Như vậy, ta có định nghĩa sau:

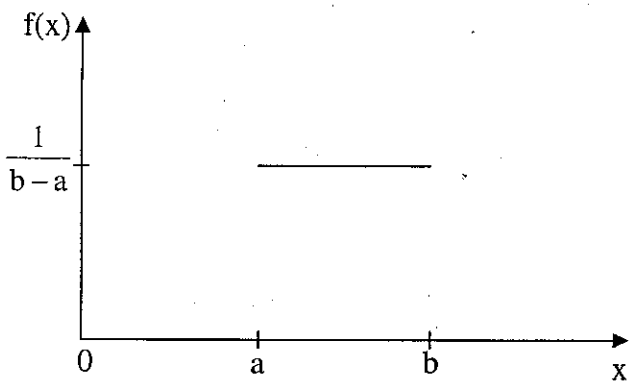
6.1. Định nghĩa

Biến ngẫu nhiên liên tục X gọi là phân phối theo quy luật đều trong khoảng (a,b) nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{với } x \in (a,b) \\ 0 & \text{với } x \notin (a,b) \end{cases} \qquad (3.19)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Đồ thị hàm f(x) có dạng như ở hình 3.1.



Hình 3.1. **Đồ thị hàm f(x) của quy luật phân phối đều**

6.2. Các tham số đặc trưng của quy luật phân phối đều

Giả sử X phân phối đều, lúc đó theo định nghĩa kỳ vọng toán của biến ngẫu nhiên liên tục, ta có:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{a}^{b} x\frac{1}{b-a}dx =$$

$$\frac{1}{b-a}\cdot\frac{x^2}{2}\Big|_a^b = \frac{1}{b-a}\cdot\frac{b^2-a^2}{2} = \frac{a+b}{2}$$

Như vậy: $$E(X) = \frac{a+b}{2} \qquad (3.20)$$

Để tìm phương sai ta tìm $E(X^2)$:

$$E(X^2) = \int_{a}^{b} x^2\frac{1}{b-a}dx = \frac{1}{b-a}\frac{x^3}{3}\Big|_a^b =$$

$$= \frac{1}{b-a}\cdot\frac{b^3-a^3}{3} = \frac{b^2+ab+a^2}{3}$$

Do đó:

$$V(X) = E(X^2) - [E(X)]^2 = \frac{(b-a)^2}{12}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Như vậy: $$V(X) = \frac{(b-a)^2}{12} \quad (3.21)$$

Quy luật phân phối đều có ứng dụng rộng trong thống kê toán. Nó có ý nghĩa to lớn trong các phương pháp phi tham số. Khái niệm phân phối đều đôi khi còn được sử dụng trong lý thuyết các ước lượng thống kê. Trong một số lý thuyết kết luận thống kê, người ta thường xuất phát từ quy tắc sau đây: *Nếu ta không biết gì về giá trị của tham số cần ước lượng thì mỗi giá trị có thể có của tham số đó là đồng khả năng.* Điều đó dẫn đến việc quan niệm tham số cần ước lượng như một biến ngẫu nhiên tuân theo quy luật phân phối đều.

Thí dụ: Khi thâm nhập vào thị trường mới, doanh nghiệp không thể khẳng định được một cách chắc chắn doanh số hàng tháng có thể đạt được sẽ là bao nhiêu mà chỉ dự kiến được rằng doanh số tối thiểu sẽ là 20 triệu đồng/tháng và tối đa là 40 triệu đồng/ tháng. Tìm xác suất để doanh nghiệp đạt được doanh số tối thiểu là 35 triệu đồng/tháng.

Giải. Gọi X là doanh số hàng tháng mà doanh nghiệp có thể đạt được ở thị trường đó. Do không có thông tin gì hơn nên có thể xem X là biến ngẫu nhiên liên tục phân phối đều trên khoảng (20; 40).

Vậy X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \dfrac{1}{40-20} = 0,05 & \text{với } x \in (20; 40) \\ 0 & \text{với } x \notin (20; 40) \end{cases}$$

Từ đó xác suất để doanh nghiệp đạt được doanh số tối thiểu là 35 triệu đồng/tháng được tìm theo tính chất của hàm mật độ xác suất như sau:

$$P(X > 35) = \int_{35}^{+\infty} f(x)dx = \int_{35}^{40} 0,05dx = 0,05x\Big|_{35}^{40} = 0,25$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

3.31. Nhu cầu về một loại hàng hoá phân phối đều trong khoảng [30;50] (tấn/tháng).

a. Xác định hàm mật độ xác suất.

b. Xác định hàm phân bố xác suất.

c. Tìm kỳ vọng toán và phương sai.

d. Vẽ đồ thị hàm mật độ xác suất và hàm phân bố xác suất.

e. Tìm xác suất để nhu cầu không vượt quá 45 tấn.

ĐS: c. 40; 33,333; e. p = 0,75

3.32. Khi thâm nhập một thị trường mới, doanh nghiệp chỉ dự kiến được rằng có thể đạt được doanh số trung bình là 30 triệu đồng/tháng và độ lệch chuẩn là 5 triệu. Tìm xác suất để khi thâm nhập vào thị trường đó doanh nghiệp sẽ đạt được doanh số ít nhất là 32 triệu đồng/tháng.

7. QUY LUẬT PHÂN PHỐI LUỸ THỪA - E(λ)

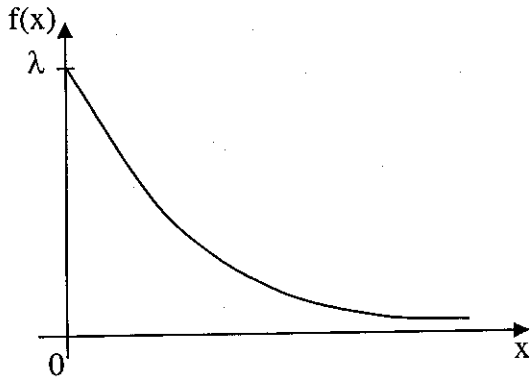
7.1. Định nghĩa

Biến ngẫu nhiên liên tục X gọi là phân phối theo quy luật luỹ thừa (quy luật mũ) nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \begin{cases} 0 & \text{với } x < 0) \\ \lambda e^{-\lambda x} & \text{với } x \geq 0 \end{cases} \qquad (3.22)$$

Trong đó λ là một hằng số dương.

Đồ thị của hàm f(x) có dạng như ở hình 3.2.



Hình 3.2. **Đồ thị hàm f(x) của quy luật luỹ thừa**

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ta tìm hàm phân bố xác suất của quy luật lũy thừa. Theo tính chất của hàm mật độ xác suất, ta có:

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{0}^{x} \lambda e^{-\lambda x}dx = 1 - e^{-\lambda x} \qquad (3.23)$$

7.2. Các tham số đặc trưng của quy luật luỹ thừa

Giả sử biến ngẫu nhiên X phân phối theo quy luật lũy thừa. Lúc đó kỳ vọng toán

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{0}^{+\infty} x\lambda e^{-\lambda x}dx$$

Lấy tích phân từng phần, ta được:

$$E(X) = \frac{1}{\lambda} \qquad (3.24)$$

Còn phương sai:

$$V(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - [E(X)]^2 = \int_{0}^{+\infty} \lambda x^2 e^{-\lambda x}dx - \frac{1}{\lambda^2}$$

Lấy tích phân từng phần ta được:

$$\lambda \int_{-\infty}^{\infty} x^2 e^{-\lambda x}dx = \frac{2}{\lambda^2}$$

Do đó:
$$V(X) = \frac{1}{\lambda^2} \qquad (3.25)$$

Suy ra độ lệch chuẩn:

$$\sigma_x = \sqrt{V(X)} = \frac{1}{\lambda} \qquad (3.26)$$

Như vậy, trong quy luật luỹ thừa kỳ vọng toán và độ lệch chuẩn đều bằng $\frac{1}{\lambda}$. Đây chính là tính chất đặc biệt của quy luật luỹ thừa. Nó có thể được sử dụng để kiểm tra xem một biến ngẫu nhiên mà ta nghiên cứu trong thực tế có phân phối theo quy luật luỹ thừa hay không.

Ta tìm xác suất để biến ngẫu nhiên X phân phối theo quy luật luỹ thừa nhận giá trị trong khoảng (a, b). Theo tính chất của hàm phân bố xác suất:

$$P(a \le X < b) = F(b) - F(a) = [1 - e^{-b\lambda}] - [1 - e^{-a\lambda}]$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vậy:

$$P(a \le X < b) = e^{-a\lambda} - e^{-b\lambda} \qquad (3.27)$$

Giá trị của hàm e^{-x} được tính sẵn thành bảng (Phụ lục 3).

Thí dụ. Biến ngẫu nhiên liên tục X phân phối theo quy luật lũy thừa với hàm mật độ xác suất.

$$f(x) = \begin{cases} 2e^{-2x} \text{ với } x \ge 0) \\ 0 \quad \text{với } x < 0 \end{cases}$$

a. Viết hàm F(x)

b. Tìm xác suất để trong kết quả của phép thử X nhận giá trị trong khoảng [0,3; 1].

c. Tìm kỳ vọng toán và phương sai của X.

Giải. a. Hàm phân bố xác suất có dạng:

$$F(x) = \begin{cases} 1 - e^{-2x} \text{ với } x \ge 0) \\ 0 \text{ với } x < 0 \end{cases}$$

b. Theo công thức (3.27) ta có:

$$P(0,3 < X < 1) = e^{-2.0,3} - e^{-2.1} = e^{-0,6} - e^{-2} =$$

$$= 0,54881 - 0,13534 \approx 0,41$$

c. Theo công thức tính kỳ vọng toán và phương sai:

$$E(X) = \frac{1}{\lambda} = \frac{1}{2} = 0,5$$

$$V(X) = \frac{1}{\lambda^2} = \frac{1}{4} = 0,25$$

$$\sigma_x = \frac{1}{\lambda} = \frac{1}{2} = 0,5$$

Quy luật phân phối luỹ thừa có ứng dụng trong nhiều lĩnh vực khác nhau. Người ta chứng minh được rằng thời gian giữa hai lần xuất hiện yêu cầu của một dòng yêu cầu tối giản trong các hệ thống phục vụ công cộng phân phối theo quy luật lũy thừa. Trong các hệ thống kỹ thuật, thời gian làm việc liên tục của máy móc thiết bị giữa hai lần sửa chữa cũng thường phân phối theo quy luật lũy thừa.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Khi áp dụng quy luật lũy thừa để giải quyết các bài toán nảy sinh trong thực tế, ngoài ưu điểm là đơn giản (nó chỉ phụ thuộc vào một tham số là λ) nó còn có một tính chất rất quan trọng sau đây: *Xác suất hoạt động liên tục của thiết bị trong khoảng thời gian t không phụ thuộc vào quãng thời gian hoạt động trước đó mà chỉ phụ thuộc vào độ dài của khoảng thời gian t mà thôi.*

Thật vậy, gọi A là biến cố thiết bị hoạt động tốt trong khoảng thời gian $(0,\ t_0)$, gọi B là biến cố thiết bị hoạt động tốt trong khoảng thời gian (t_0, t_0+t). Lúc đó tích AB là biến cố thiết bị hoạt động tốt trong khoảng thời gian $(0, t_0+t)$ có độ dài (t_0+t), lúc đó P(A) là xác suất để thời gian hoạt động tốt T của thiết bị không nhỏ hơn t_0 và P(B) là xác suất để thời gian hoạt động tốt của thiết bị không nhỏ hơn t.

$$P(A) = P(T \geq t_0) = 1 - P(T < t_0) = 1 - F(t_0)$$

$$P(B) = P(T \geq t) = 1 - P(T < t) = 1 - F(t)$$

Vì T phân phối theo quy luật lũy thừa do đó theo công thức (3.23) ta có

$$P(A) = 1 - [1 - e^{-\lambda t_0}] = e^{-\lambda t_0}$$

$$P(B) = 1 - [1 - e^{-\lambda t}] = e^{-\lambda t}$$

$$P(AB) = P(T \geq t_0 + t) = 1 - P(T < t_0 + t) =$$

$$= 1 - F(t_0 + t) = 1 - [1 - e^{-\lambda(t_0+t)}]$$

Vậy $P(AB) = e^{-\lambda(t_0+t)}$

Bây giờ ta tìm xác suất để thiết bị sẽ hoạt động tốt trong khoảng (t_0, t_0+t) với điều kiện nó đã hoạt động tốt trong khoảng $(0, t_0)$. Từ định lý nhân xác suất ta có:

$$P(B/A) = \frac{P(AB)}{P(A)} = \frac{e^{-\lambda t_0}.e^{-\lambda t}}{e^{-\lambda t_0}} = e^{-\lambda t}$$

Như vậy, P(B/A) = P(B) tức là A và B độc lập với nhau và trong biểu thức của P(B) chỉ có tham số t chứ không có t_0. Điều đó chứng tỏ xác suất để thiết bị hoạt động tốt trong khoảng thời gian t sẽ không phụ thuộc gì vào khoảng thời gian hoạt động tốt đã diễn ra trước đó, mà chỉ phụ thuộc vào độ dài của khoảng thời gian đang xét mà thôi.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Người ta chứng minh được rằng chỉ có quy luật phân phối lũy thừa mới có tính chất trên. Đó cũng là tiêu chuẩn để nhận biết quy luật này trong thực tế.

Bài tập

3.33. Tuổi thọ của một loại bóng đèn là biến ngẫu nhiên tuân theo quy luật luỹ thừa với hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} 0 & \text{với } x < 0 \\ \dfrac{1}{1500} e^{-x\,1500} & \text{với } x \geq 0 \end{cases}$$

a. Hãy xây dựng hàm phân bố xác suất.

b. Tìm tỷ lệ các bóng đèn có tuổi thọ dưới 1500 giờ.

ĐS: P = 0,632.

3.34. Thời gian phục vụ mỗi khách hàng tại một cửa hàng mậu dịch là biến ngẫu nhiên X tuân theo quy luật luỹ thừa với hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} 5e^{-5x} & \text{với } x > 0 \\ 0 & \text{với } x < 0 \end{cases}$$

Với x được tính bằng phút/khách hàng.

a. Tìm xác suất để thời gian phục vụ khách hàng nào đó sẽ nằm trong khoảng từ 0,4 đến 1 phút.

b. Tìm thời gian trung bình để phục vụ một khách hàng.

ĐS: a. P = 0,13; b. E(X) = 0,2

3.35. Thời gian chờ bốc xếp của các con tàu tại một bến cảng là biến ngẫu nhiên tuân theo quy luật luỹ thừa với hàm mật độ xác suất là:

$$f(x) = \begin{cases} 2e^{-2x} & \text{với } x > 0 \\ 0 & \text{với } x < 0 \end{cases}$$

Với x được tính bằng tháng. Tìm thời gian chờ đợi trung bình của mỗi con tàu để được bốc xếp.

3.36. Khoảng cách thời gian mà hai khách hàng kế tiếp nhau đến ngân hàng là biến ngẫu nhiên phân phối luỹ thừa với trung bình là 3 phút. Giả sử vừa có 1 khách đến. Tìm xác suất để trong vòng ít nhất 2 phút nữa mới có người khách tiếp theo đến ngân hàng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8. QUY LUẬT PHÂN PHỐI CHUẨN - $N(\mu, \sigma^2)$

8.1. Định nghĩa 1

Biến ngẫu nhiên liên tục X nhận các giá trị trong khoảng $(-\infty, +\infty)$ gọi là phân phối theo quy luật chuẩn với các tham số μ và σ^2, nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (3.28)$$

Nếu tiến hành khảo sát hàm số trên và vẽ đồ thị của nó ta sẽ thu được các kết luận sau đây:

a. Hàm số xác định trên toàn trục Ox

b. Với mọi giá trị của x hàm số luôn luôn dương, như vậy, đồ thị của nó luôn nằm cao hơn trục Ox.

c. Khi $x \to \pm\infty$ thì $f(x) \to 0$ tức là trục Ox là đường tiệm cận ngang.

d. Ta tìm đạo hàm bậc nhất

$$f'(x) = -\frac{x-\mu}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Dễ dàng thấy rằng $f'(x) = 0$ khi $x = \mu, f'(x) > 0$ khi $x < \mu, f'(x) < 0$ khi $x > \mu$.

Như vậy khi $x = \mu$ hàm số có cực đại bằng $\frac{1}{\sigma\sqrt{2\pi}}$

e. Hiệu $x - \mu$ trong biểu thức của hàm f(x) nằm trong dạng bình phương, tức là hàm số đối xứng qua đường thẳng $x = \mu$.

g. Ta tìm điểm uốn của hàm. Đạo hàm bậc hai

$$f''(x) = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[1 - \frac{(x-\mu)^2}{\sigma^2}\right]$$

Dễ dàng thấy rằng khi $x = \mu + \sigma$ và $x = \mu - \sigma$ đạo hàm bậc hai bằng

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

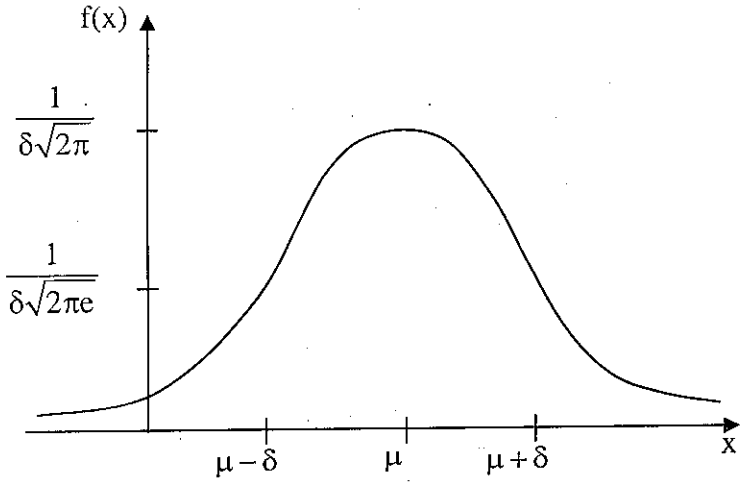
không và đi qua hai điểm đó nó đổi dấu (tại cả hai điểm đó hàm số đều bằng $\frac{1}{\sigma\sqrt{2\pi e}}$. Như vậy các điểm:

$$\left(\mu-\sigma;\frac{1}{\sigma\sqrt{2\pi e}}\right) \text{và} \left(\mu+\sigma;\frac{1}{\sigma\sqrt{2\pi e}}\right)$$

là các điểm uốn.

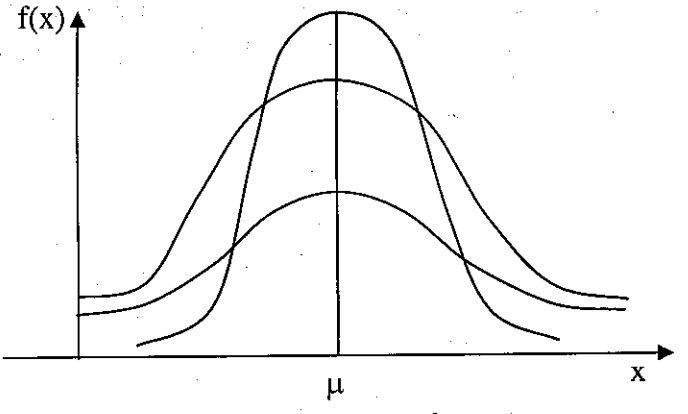
Vậy đồ thị của hàm mật độ xác suất của phân phối chuẩn có dạng như sau (hình 3.3) .

Hai tham số μ và σ có ý nghĩa rất quan trọng trong phân phối chuẩn (bản chất của nó sẽ được trình bày về sau). Khi μ và σ thay đổi, dạng đồ thị của hàm mật độ xác suất f(x) cũng thay đổi như sau: Khi μ thay đổi thì dạng của đường cong f(x) không thay đổi song nó sẽ chuyển dịch sang phải hoặc sang trái theo trục Ox. Khi μ tăng lên đồ thị sẽ dịch sang phải, còn khi μ giảm đồ thị sẽ dịch sang trái. Trái lại, khi σ thay đổi thì dạng của đồ thị sẽ thay đổi theo. Nếu σ tăng lên thì đồ thị sẽ thấp xuống và phình ra, còn khi σ giảm đi thì đồ thị sẽ cao lên và nhọn thêm. Trên hình 3.4 ta minh họa đồ thị của f(x) với ba giá trị khác nhau của σ.



Hình 3.3. **Đồ thị hàm f(x) phân phối chuẩn**

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Hình 3.4. **Sự thay đổi của f(x) theo σ**

Theo tính chất của hàm mật độ xác suất, ta có hàm phân bố xác suất của biến ngẫu nhiên X phân phối theo quy luật chuẩn được xác định bằng biểu thức:

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

8.2. Các tham số đặc trưng của quy luật chuẩn

Ta sẽ chứng minh rằng trong quy luật chuẩn thì μ chính là kỳ vọng toán còn σ chính là độ lệch chuẩn của X. Thật vậy, theo định nghĩa kỳ vọng toán của biến ngẫu nhiên liên tục, ta có:

$$E(X)=\int_{-\infty}^{+\infty} xf(x)dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

Ta thực hiện phép đổi biến số: $Z=\frac{x-\mu}{\sigma}$

Từ đó $x=\sigma Z+\mu, dx=\sigma dZ$. Chú ý rằng khi đổi biến các cận lấy tích phân không thay đổi, ta có:

$$E(X)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}(\sigma Z+\mu)e^{-\frac{Z^2}{2}}dZ=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\sigma Z e^{-\frac{Z^2}{2}}dZ+\frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{Z^2}{2}}dZ$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tích phân thứ nhất bằng không do hàm dưới dấu tích phân là hàm lẻ mà cận lấy tích phân lại đối xứng qua gốc toạ độ. Còn tích phân thứ hai bằng:

$$\int_{-\infty}^{+\infty} e^{-\frac{Z^2}{2}} dZ = \sqrt{2\pi} \text{ (tích phân Poisson)}$$

Do đó: $E(X)=\mu$ (3.29)

Theo định nghĩa phương sai của biến ngẫu nhiên liên tục và do $E(X)=\mu$ ta có:

$$V(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Ta thực hiện phép đổi biến số $Z = \frac{x-\mu}{\sigma}$, từ đó $x - \mu = Z\sigma$, $dx = \sigma dZ$.

Chú ý rằng cận lấy tích phân không thay đổi ta có:

$$V(X) = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} Z^2 e^{-\frac{Z^2}{2}} dZ$$

Lấy tích phân từng phần bằng cách đặt $U = Z$, $dv = Ze^{-\frac{Z^2}{2}} dZ$ ta tìm được $V(X) = \sigma^2$

Do đó: $\sigma_x = \sqrt{V(X)} = \sigma$ (3.30)

Như vậy, kỳ vọng toán của biến ngẫu nhiên X phân phối chuẩn là $E(X)=\mu$ và phương sai là: $V(x) = \sigma^2$. Phân phối chuẩn được ký hiệu là $N(\mu, \sigma^2)$.

Có liên quan mật thiết với phân phối chuẩn là một phân phối khác gọi là phân phối chuẩn hoá.

Giả sử biến ngẫu nhiên X phân phối chuẩn với kỳ vọng toán bằng μ và độ lệch chuẩn σ. Xét biến ngẫu nhiên (trong thống kê thường gọi là phép chuẩn hoá biến ngẫu nhiên X):

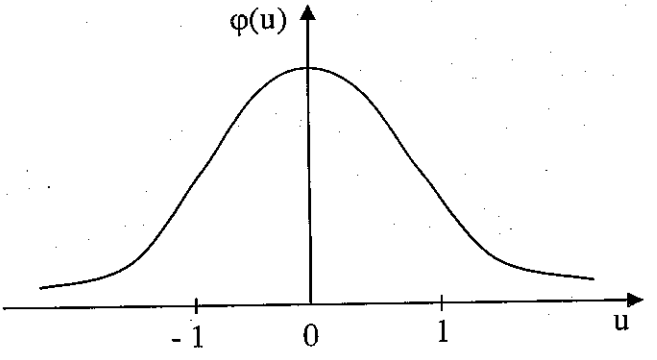
$$U = \frac{X-\mu}{\sigma}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.3. Định nghĩa 2

Biến ngẫu nhiên U nhận các giá trị trong khoảng $(-\infty,+\infty)$ gọi là tuân theo quy luật phân phối chuẩn hoá nếu hàm mật độ xác suất của nó có dạng:

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}} \qquad (3.31)$$



***Hình 3.5.* Đồ thị hàm $\varphi(u)$**

Đồ thị của hàm $\varphi(u)$ có dạng như hình vẽ.

Đặc điểm của đồ thị này là nó lấy trục tung làm trục đối xứng. Các giá trị của hàm $\varphi(u)$ được tính sẵn thành bảng (Phụ lục 4).

Hàm phân bố xác suất của biến ngẫu nhiên U phân phối chuẩn hoá có dạng:

$$\phi(u)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}e^{-\frac{u^2}{2}}du$$

Các giá trị của hàm $\phi(u)$ cũng được tính sẵn thành bảng (Phụ lục 5).

Ta tìm các tham số đặc trưng của biến ngẫu nhiên U phân phối chuẩn hoá:

$$E(U)=E\left(\frac{X-\mu}{\sigma}\right)$$

Theo tính chất của kỳ vọng toán ta có:

$$E(U)=\frac{1}{\sigma}E(X-\mu)=\frac{1}{\sigma}[E(X)-\mu]$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Song $E(X)=\mu$, do đó $E(U)=0$

$$V(U)=V\left(\frac{X-\mu}{\sigma}\right)$$

Cũng theo tính chất của phương sai ta có:

$$V(U)=\frac{1}{\sigma^2}V(X-\mu)=\frac{1}{\sigma^2}V(X)$$

Song $V(X)=\sigma^2$, do đó $V(U)=1$

Phân phối chuẩn hoá được ký hiệu là N(0; 1).

Ngoài các tham số đặc trưng là kỳ vọng toán μ và phương sai σ^2, trong phân phối chuẩn có một tham số khác có nhiều ứng dụng thực tế, đó là giá trị tới hạn chuẩn.

8.4. Định nghĩa 3

Giá trị tới hạn chuẩn mức α (ký hiệu là u_α) là giá trị của biến ngẫu nhiên U có phân phối chuẩn hoá thoả mãn điều kiện $P(U>u_\alpha)=\alpha$.

Vì U chuẩn hoá nên theo (3.31) hàm mật độ của U là:

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$$

Theo tính chất hàm mật độ thì:

$$P(U>u_\alpha)=\int_{u_\alpha}^{+\infty}\varphi(u)du$$

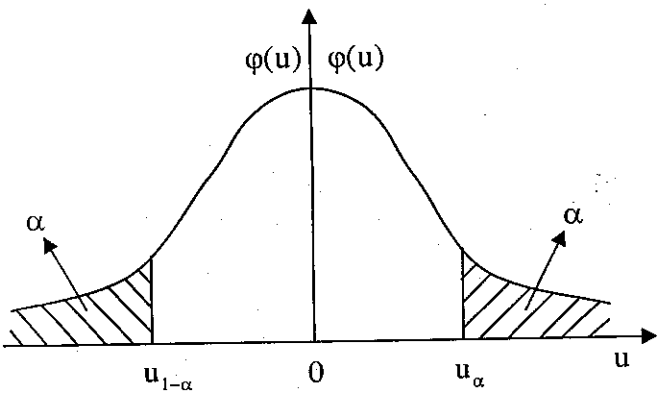
Do đó: $P(U>u_\alpha)=\frac{1}{\sqrt{2\pi}}\int_{u_\alpha}^{+\infty}e^{-\frac{u^2}{2}}=\alpha$

Cho trước α, dựa vào biểu thức trên người ta tính được u_α và ngược lại.

Các giá trị của u_α được tính sẵn thành bảng (Phụ lục 6).

Trên đồ thị giá trị tới hạn chuẩn u_α là giá trị sao cho diện tích giới hạn bởi đường cong phân phối chuẩn hóa, trục 0U và đường thẳng $U = u_\alpha$ bằng α.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Hình 3.6. **Giá trị tới hạn chuẩn u_α**

Từ hình vẽ ta thấy ngay giá trị tới hạn chuẩn có tính chất sau đây:

$$u_{1-\alpha} = -u_\alpha$$

Sau đây ta sẽ xây dựng một số công thức có nhiều ứng dụng trong việc giải các bài toán thực tế.

8.5. Công thức tính xác suất để biến ngẫu nhiên X phân phối chuẩn nhận giá trị trong khoảng (a, b)

Ta biết rằng nếu biến ngẫu nhiên liên tục X có hàm mật độ xác suất là f(x) thì xác suất để X nhận giá trị trong khoảng (a, b) sẽ được tính bằng công thức:

$$P(a < X < b) = \int_a^b f(x)dx$$

Giả sử X phân phối chuẩn. Lúc đó:

$$P(a < X < b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Ta thay biến mới $Z = \frac{x-\mu}{\sigma}$, lúc đó $x = \sigma Z + \mu$ và $dx = \sigma dZ$.

Ta tìm cận lấy tích phân khi đổi biến. Khi $x = a$ thì $Z = \frac{a-\mu}{\sigma}$ và khi $x = b$ thì $Z = \frac{b-\mu}{\sigma}$. Như vậy ta có:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P(a<X<b)=\frac{1}{\sqrt{2\pi}}\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-\frac{Z^2}{2}}dZ=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{\frac{a-\mu}{\sigma}}^{0} e^{-\frac{Z^2}{2}}dZ+\frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{b-\mu}{\sigma}} e^{-\frac{Z^2}{2}}dZ=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{b-\mu}{\sigma}} e^{\frac{Z^2}{2}}dZ-\frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{a-\mu}{\sigma}} e^{\frac{Z^2}{2}}dZ=$$

$$=\Phi_0\left(\frac{b-\mu}{\sigma}\right)-\Phi_0\left(\frac{a-\mu}{\sigma}\right)$$

Trong đó:

$$\Phi_0(u)=\frac{1}{\sqrt{2\pi}}\int_{0}^{u} e^{-\frac{Z^2}{2}}dZ$$

Giá trị của hàm $\Phi_0(u)$ được tính sẵn thành bảng (Phụ lục 5).

Chú ý rằng hàm $\Phi_0(u)$ có các tính chất sau:

+ $\Phi_0(-u)=-\Phi_0(u)$

+ Với mọi $u>5$ thì $\Phi_0(u)\approx\Phi_0(5)=0,5$.

Các tính chất trên được vận dụng khi tra bảng giá trị hàm $\Phi_0(u)$.

Như vậy ta thu được công thức:

$$P(a<X<b)=\Phi_0\left(\frac{b-\mu}{\sigma}\right)-\Phi_0\left(\frac{a-\mu}{\sigma}\right) \qquad (3.32)$$

Thí dụ 1. Một nhà sản xuất cần mua một loại gioăng cao su có độ dày từ 0,118cm đến 0,122cm. Có hai cửa hàng cùng bán loại gioăng này với độ dày là các biến ngẫu nhiên phân phối chuẩn với các đặc trưng được cho trong bảng sau:

	Độ dày trung bình	Độ lệch chuẩn	Giá bán
Cửa hàng A	0,12	0,001	3 USD/hộp/1000 c
Cửa hàng B	0,12	0,0015	2,6D/hộp/1000 c

Hỏi nhà sản xuất nên mua gioăng của cửa hàng nào?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Trước hết cần xác định tỉ lệ gioăng đáp ứng được yêu cầu của nhà sản xuất trong mỗi hộp sản phẩm của hai cửa hàng. Gọi X_A và X_B tương ứng là độ dày của gioăng do cửa hàng A và B bán. Theo giả thiết X_A và X_B đều phân phối chuẩn. Vậy, tỉ lệ gioăng dùng được của hai cửa hàng tương ứng là:

$$P(0,118 < X_A < 0,122) = \Phi_0\left(\frac{0,122-0,12}{0,001}\right) - \Phi_0\left(\frac{0,118-0,12}{0,001}\right)$$
$$= 2\Phi_0(2) = 0,9544$$

$$P(0,118 < X_B < 0,122) = \Phi_0\left(\frac{0,122-0,12}{0,0015}\right) - \Phi_0\left(\frac{0,118-0,12}{0,0015}\right)$$
$$= 2\Phi_0(1,33) = 0,8164$$

Vậy giá bán đối với mỗi chiếc gioăng dùng được của cửa hàng A là:

$$\frac{3}{954,4} = 0,00314\text{USD}$$

Và của cửa hàng B là: $\frac{2,6}{816,4} = 0,00318\text{USD}$

Vậy nhà sản xuất nên mua gioăng của cửa hàng A.

8.6. Xác suất của sự sai lệch giữa biến ngẫu nhiên và kỳ vọng toán của nó

Trong thực tế nhiều khi ta phải tính xác suất để biến ngẫu nhiên X phân phối chuẩn nhận giá trị sai lệch so với kỳ vọng toán của nó về giá trị tuyệt đối nhỏ hơn một số dương cho trước, tức là ta phải tìm xác suất để xảy ra bất đẳng thức $|X-\mu| < \varepsilon$.

Ta thay bất đẳng thức trên bằng bất đẳng thức kép tương đương với nó.

$$\mu - \varepsilon < X < \mu + \varepsilon$$

Sử dụng công thức (3.32) ta có:

$$P(|X-\mu| < \varepsilon) = P(\mu-\varepsilon < X < \mu+\varepsilon)$$
$$= \Phi_0\left(\frac{\mu+\varepsilon-\mu}{\sigma}\right) - \Phi_0\left(\frac{\mu-\varepsilon-\mu}{\sigma}\right)$$
$$= \Phi_0\left(\frac{\varepsilon}{\sigma}\right) - \Phi_0\left(-\frac{\varepsilon}{\sigma}\right) = \Phi_0\left(\frac{\varepsilon}{\sigma}\right) + \Phi_0\left(\frac{\varepsilon}{\sigma}\right) = 2\Phi_0\left(\frac{\varepsilon}{\sigma}\right)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Như vậy ta thu được công thức:

$$P(|X-\mu|<\varepsilon)=2\Phi_0\left(\frac{\varepsilon}{\sigma}\right) \qquad (3.33)$$

Thí dụ 2. Các vòng bi do một máy tự động sản xuất ra được coi là đạt tiêu chuẩn nếu đường kính của nó sai lệch so với đường kính thiết kế không quá 0,7mm. Biết rằng sai lệch này là biến ngẫu nhiên phân phối chuẩn với $\mu=0$ và $\sigma=0,4$ mm. Tìm tỷ lệ vòng bi đạt tiêu chuẩn của máy đó.

Giải. Ta thấy rằng tỷ lệ vòng bi đạt tiêu chuẩn chính là xác suất để lấy ngẫu nhiên một vòng bi được vòng bi đạt tiêu chuẩn. Nếu gọi X là sai lệch giữa đường kính của vòng bi được sản xuất ra so với đường kính thiết kế thì xác suất này chính là xác suất để xảy ra bất đẳng thức $|X-\mu|<0,7$. Theo giả thiết X phân phối chuẩn với $\mu=0$ và $\sigma=0,4$ do đó theo công thức (3.33) ta có:

$$P(|X-\mu|<0,07)=P(|X|<0,07)=2\Phi_0\left(\frac{0,7}{0,4}\right)=2\Phi_0(1,75)$$

Tra bảng giá trị hàm $\Phi_0(U)$ ta có: $\Phi_0(1,75)=0,4599$

Do đó $P(|X|<0,7)=2.0,4599=0,9198$.

Vậy tỷ lệ vòng bi đạt tiêu chuẩn của máy đó gần bằng 92%.

Ta chú ý rằng do tính chất $\int_{-\infty}^{+\infty}\varphi(u)du=1$ và do hàm $\varphi(u)$ đối xứng qua gốc toạ độ do đó $\int_{-\infty}^{0}\varphi(u)du=0,5$ và như vậy: $P(-\infty<U<0)=0,5$

Từ đó $\Phi(u)=P(-\infty<U<0)+P(0<U<u)=0,5+\Phi_0(u)$

Như vậy ta có mối liên hệ sau $\Phi(u)=0,5+\Phi_0(u)$

8.7. Quy tắc hai xích ma và ba xích ma

Nếu trong công thức (3.33) ta đặt $\varepsilon=2\sigma$ tức là bằng hai lần độ lệch chuẩn của X thì ta có:

$$P(|X-\mu|<2\sigma)=P(\mu-2\sigma<X<\mu+2\sigma)$$

$$=2\phi_0(2)=0,9544 \qquad (3.34)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

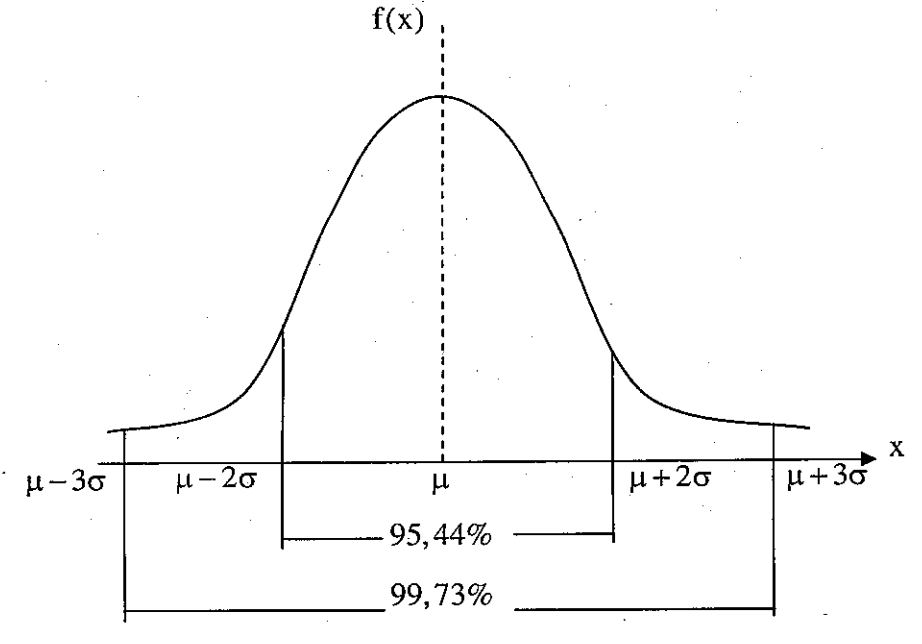
Biểu thức (3.34) được gọi là quy tắc hai xích ma. Tương tự nếu đặt $\varepsilon = 3\sigma$ thì ta sẽ thu được biểu thức tương ứng.

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 2\phi_0(3) \approx 0,9973 \qquad (3.35)$$

Các quy tắc trên cho thấy xác suất để biến ngẫu nhiên phân phối chuẩn nhận giá trị trong khoảng $(\mu - 2\sigma < X < \mu + 2\sigma)$ là 0,9544 hay 95,44% các giá trị của X sẽ nằm trong khoảng nói trên. Còn theo quy tắc ba xích ma thì hầu hết (tới 99,73%) các giá trị của X phân phối chuẩn sẽ nằm trong khoảng $(\mu - 3\sigma; \mu + 3\sigma)$. Điều này được thể hiện trên đồ thị như sau (hình 3.7).

Trong thực tế quy tắc hai xích ma và quy tắc ba xích ma được áp dụng như sau: Nếu quy luật phân phối xác suất của biến ngẫu nhiên được nghiên cứu chưa biết, song nó thoả mãn điều kiện của quy tắc hai xích ma hoặc ba xích ma thì có thể xem như biến ngẫu nhiên đó phân phối chuẩn.

Mặt khác nếu biến ngẫu nhiên phân phối chuẩn thì 95,44% các giá trị của nó sẽ nằm trong khoảng $(\mu - 2\sigma; \mu + 2\sigma)$ còn hầu như toàn bộ các giá trị của nó sẽ nằm trong khoảng $(\mu - 3\sigma; \mu + 3\sigma)$.



Hình 3.7. **Quy tắc hai xích ma và ba xích ma**

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.8. Phân phối xác suất của tổng các biến ngẫu nhiên độc lập tuân theo cùng một quy luật

Giả sử X_1 và X_2 là hai biến ngẫu nhiên độc lập, X_1 tuân theo quy luật chuẩn với kỳ vọng toán μ_1 và phương sai σ_1^2 còn X_2 cũng tuân theo quy luật chuẩn với kỳ vọng toán μ_2 và phương sai σ_2^2. Lúc đó tổng của chúng là biến ngẫu nhiên $X = X_1 + X_2$ cũng phân phối theo quy luật chuẩn với kỳ vọng toán là $\mu_1 + \mu_2$ và phương sai là $\sigma_1^2 + \sigma_2^2$ (xem thêm mục 10 chương IV). Tính chất trên cũng có thể mở rộng cho một số bất kỳ các biến ngẫu nhiên độc lập lẫn nhau và cùng phân phối chuẩn.

Mặt khác nếu $X_1X_2...X_n$ là n biến ngẫu nhiên độc lập lẫn nhau và cùng tuân theo một quy luật phân phối xác suất nào đó (không nhất thiết là quy luật chuẩn) với các kỳ vọng toán $E(X_1), E(X_2)...E(X_n)$ và các phương sai $V(X_1), V(X_2)...V(X_n)$ đã biết thì biến ngẫu nhiên $X = \sum_{i=1}^{n} X_i$ sẽ phân phối xấp xỉ chuẩn với $E(X) = \sum_{i=1}^{n} E(x_i)$ và $V(X) = \sum_{i=1}^{n} V(x_i)$ khi n khá lớn (n > 30). Tính chất trên thường được gọi là định lý giới hạn trung tâm của Liapunốp (xem thêm mục 5 Chương V).

8.9. Sự hội tụ của quy luật nhị thức và quy luật Poisson về quy luật chuẩn

Khi sử dụng quy luật nhị thức, nếu n khá lớn thì việc tính toán theo công thức Bernoulli sẽ gặp khó khăn. Lúc đó nếu p nhỏ đến mức $np \approx npq$ thì có thể dùng quy luật Poisson thay thế cho quy luật nhị thức. Song nếu p lại không nhỏ (p > 0,1) thì không thể dùng quy luật Poisson để thay thế được. Lúc đó có thể dùng quy luật chuẩn để thay thế cho quy luật nhị thức.

Trong thực tế quy luật chuẩn có thể thay thế cho quy luật nhị thức nếu thoả mãn đồng thời hai điều kiện là:

$$n > 5 \text{ và } \left| \sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}} \right| \frac{1}{\sqrt{n}} < 0,3$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lúc đó thì biến ngẫu nhiên X phân phối nhị thức có thể coi như phân phối xấp xỉ chuẩn với kỳ vọng toán $\mu = np$ và phương sai $\sigma^2 = npq$.

Từ đó:

$$P(X = x) = C_n^x p^x q^{n-x} \approx \frac{1}{\sqrt{npq}} \varphi\left(\frac{x - np}{\sqrt{npq}}\right) \qquad (3.36)$$

Công thức (3.36) được gọi là định lý địa phương Laplace.

Mặt khác:

$$P(x \le X \le x + h) = P_x + P_{x+1} + ... + P_{x+h} \approx$$

$$\approx \Phi_0\left(\frac{x + h - np}{\sqrt{npq}}\right) - \Phi_0\left(\frac{x - np}{\sqrt{npq}}\right) \qquad (3.37)$$

Công thức (3.37) được gọi là định lý tích phân Laplace (xem thêm mục 5 chương V).

Thí dụ 3. Xác suất để sản phẩm sau khi sản xuất không được kiểm tra chất lượng bằng 0,2. Tìm xác suất để trong 400 sản phẩm được sản xuất ra có:

a. 80 sản phẩm không được kiểm tra chất lượng.

b. Có từ 70 đến 100 sản phẩm không được kiểm tra chất lượng.

Giải. Bài toán thoả mãn lược đồ Bernoulli do đó nếu gọi X là số sản phẩm không được kiểm tra chất lượng thì $X \sim B(n = 400; p = 0,2)$. Song vì $n = 400 > 5$ và

$$\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \frac{1}{\sqrt{n}} = \left|\frac{\sqrt{0,2}}{\sqrt{0,8}} - \frac{\sqrt{0,8}}{0,2}\right| \frac{1}{\sqrt{400}} = 0,075 < 0,3$$

Nên có thể coi $X \sim N(\mu = np = 80; \sigma^2 = npq = 64)$

a. Từ đó $P(X = 80) \approx \frac{1}{\sqrt{400.0,2.0,8}} \varphi(u) = \frac{1}{8} \varphi(u)$

Với $u = \frac{80 - 400.0,2}{\sqrt{400.0,2.0,8}} = 0$

$$\rightarrow P(X = 80) = \frac{1}{8} \varphi(0) = \frac{1}{8}.0,3989 = 0,04986$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b. $P(70 \le X \le 100) \approx \Phi_0\left(\frac{100-400.0,2}{\sqrt{400.0,2.0,8}}\right) - \Phi_0\left(\frac{70-400.0,2}{\sqrt{400.0,2.0,8}}\right)$

$= \Phi_0(2,5) - \Phi_0(-1,25)$

$= \Phi_0(2,5) + \Phi_0(1,25) = 0,8882$

Đối với quy luật Poisson thì quá trình hội tụ của nó về quy luật chuẩn sẽ diễn ra khi λ trở nên lớn hơn 20. Vì vậy, nếu X phân phối Poisson song $\lambda > 20$ thì có thể xem là X phân phối xấp xỉ chuẩn với $\mu = \lambda$ và $\sigma^2 = \lambda$.

8.10. Ứng dụng của quy luật chuẩn

Quy luật phân phối chuẩn là quy luật phân phối xác suất được áp dụng rất rộng rãi trong thực tế. Trong nhiều lĩnh vực của khoa học và đời sống ta đều gặp các biến ngẫu nhiên phân phối chuẩn. Lý do của sự phổ biến đó không những đã được giải thích trong định lý giới hạn trung tâm như đã xét ở trên mà còn từ hệ quả của định lý đó: *Nếu biến ngẫu nhiên X là tổng của một số lớn các biến ngẫu nhiên độc lập và giá trị của mỗi biến chỉ chiếm vị trí nhỏ trong tổng đó thì X sẽ có phân phối xấp xỉ chuẩn.* Trong thực tế ta gặp chính các biến ngẫu nhiên như vậy. Chẳng hạn, trong công nghiệp, người ta đã xác định được rằng kích thước của các chi tiết do các nhà máy sản xuất ra sẽ phân phối chuẩn nếu quá trình sản xuất diễn ra bình thường. Trong nông nghiệp, năng suất của cùng một loại cây trồng tại các thửa ruộng khác nhau cũng phân phối chuẩn. Năng suất lao động của các công nhân có cùng tay nghề và làm cùng một công việc như nhau cũng phân phối chuẩn. Nhu cầu về các loại hàng hoá khác nhau cũng phân phối chuẩn v.v... Người ta ghi nhận rằng các năng lực về trí tuệ và thể lực con người cũng phân phối theo quy luật chuẩn. Thậm chí cả một số chỉ tiêu về sinh lý của những người cùng giới (chẳng hạn chiều cao, vòng ngực, chiều dài cánh tay v.v...) cũng phân phối theo quy luật chuẩn. Sự nhận biết này cho phép lập kế hoạch sản xuất quần áo may sẵn sản xuất hàng loạt sao cho đáp ứng một cách hợp lý nhất kích cỡ của người mua, tránh tình trạng thừa, thiếu do không vừa kích cỡ v.v... Tóm lại, khó có thể liệt kê được hết các hiện tượng và lĩnh vực trong đó có thể áp dụng quy luật phân phối chuẩn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

3.37. Tìm xác suất để biến ngẫu nhiên phân phối chuẩn hoá nhận giá trị:

a. Trong khoảng (-2,33; 2,33)

b. Trong khoảng (-2; 1)

c. Trong khoảng (-0,89; 2,5)

d. Lớn hơn 3,02

e. Nhỏ hơn 2,5

3.38. Biến ngẫu nhiên X tuân theo quy luật chuẩn với $\mu = 10$; $\delta = 2$. Tính xác suất để X nhận giá trị trong khoảng (8, 12).

ĐS: P = 0,6826

3.39. Trọng lượng sản phẩm X do một máy tự động sản xuất là biến ngẫu nhiên tuân theo quy luật chuẩn với E(X) = 100 gam và độ lệch chuẩn 1 gam. Sản phẩm được coi là đạt tiêu chuẩn kỹ thuật nếu trọng lượng của nó đạt từ 98 đến 102 gam.

a. Tìm tỷ lệ sản phẩm đạt tiêu chuẩn kỹ thuật của nhà máy.

b. Tìm tỷ lệ phế phẩm của nhà máy.

c. Giải thích bằng đồ thị kết quả tìm được ở phần (a).

ĐS: a. 95,44%; b. 4,56%.

3.40. Trong hệ thống tỷ giá hối đoái thả nổi, sự biến động của tỷ giá hối đoái chịu sự tác động của rất nhiều nhân tố và có thể xem như biến ngẫu nhiên phân phối chuẩn. Giả sử ở một giai đoạn nào đó tỷ giá của USD với VND có trung bình là 15000đ và độ lệch chuẩn là 500đ. Tìm xác suất để trong một ngày nào đó.

a. Tỷ giá sẽ cao hơn 16000đ

b. Tỷ giá sẽ thấp hơn 14500đ

c. Tỷ giá sẽ nằm trong khoảng từ 14500đ đến 16500đ

3.41. Việc tiêu dùng điện hàng tháng của các hộ gia đình ở Hà Nội là biến ngẫu nhiên phân phối chuẩn với trung bình là 200 KWh và độ lệch chuẩn là 40 KWh. Tìm xác suất để chọn ngẫu nhiên một hộ gia đình thì hộ đó:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Có mức tiêu dùng điện hàng tháng trên 250 KWh.

b. Có mức tiêu dùng điện hàng tháng dưới 180 KWh

3.42. Chiều cao của nam giới khi trưởng thành ở một vùng dân cư là biến ngẫu nhiên phân phối chuẩn với $\mu = 160$cm và $\sigma = 6$cm. Một thanh niên bị coi là lùn nếu có chiều cao nhỏ hơn 155cm.

a. Tìm tỷ lệ thanh niên lùn của vùng đó.

b. Tìm xác suất để lấy ngẫu nhiên 4 người thì có ít nhất 1 người không bị lùn.

3.43. Kích thước chi tiết là biến ngẫu nhiên phân phối chuẩn với $\mu = 50$cm. Kích thước thực tế của các chi tiết không nhỏ hơn 32cm và không lớn hơn 68cm. Tìm xác suất để lấy ngẫu nhiên một chi tiết có kích thước.

a. Lớn hơn 55cm.

b. Nhỏ hơn 40cm.

ĐS: a. P = 0,0823; b. P = 0,0027

3.44. Năng suất lúa của một vùng là biến ngẫu nhiên phân phối chuẩn với μ = 50 tạ/ha và $\sigma = 3,6$ tạ/ha. Tìm xác suất để gặt ngẫu nhiên 3 thửa ruộng của vùng đó thì có 2 thửa ruộng có năng suất sai lệch so với năng suất trung bình không quá 0,5 tạ/ha.

ĐS: P = 0,033

3.45. Cho X_i^2 (i = 1,n) là các biến ngẫu nhiên độc lập cùng tuân theo quy luật chuẩn với

$E(X_1) = E(X_2) = ... = E(X_n) = m.$

$V(X_1) = V(X_2) = ... = V(X_n) = \sigma^2$

Lập công thức tính

$P(|\overline{X} - m| < \varepsilon)$ biết rằng $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Và cũng tuân theo quy luật chuẩn; $\varepsilon > 0$ tuỳ ý.

ĐS: $P(|\overline{X} - m| < \varepsilon) = 2\,\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.46. Tiến hành 1000 phép thử độc lập. P(A) = 0,75 trong mỗi phép thử. Tìm xác suất của sự sai lệch giữa tần suất với xác suất không vượt quá 0,02.

ĐS: P = 0,8558.

3.47. Tiến hành kiểm tra chất lượng 900 chi tiết. Xác suất được chi tiết đạt tiêu chuẩn là 0,9. Hãy tìm với xác suất 0,9544 xem số chi tiết đạt tiêu chuẩn nằm trong khoảng nào xung quanh số chi tiết đạt tiêu chuẩn trung bình?

ĐS: Từ 792 đến 828 sản phẩm.

3.48. Xác suất xuất hiện biến cố A trong mỗi phép thử độc lập là 0,5. Tìm số phép thử n để với xác suất 0,7698 có thể khẳng định rằng tần suất sai lệch so với xác suất không vượt quá 0,02.

3.49. Việc kiểm tra các viên bi tiến hành như sau: nếu viên bi không lọt qua lỗ có đường kính d_1 song lọt qua lỗ có đường kính d_2 thì viên bi được coi là đạt tiêu chuẩn, nếu không thì viên bi bị loại. Biết rằng đường kính các viên bi sản xuất ra là biến ngẫu nhiên phân phối chuẩn với

$$\mu = \frac{d_1 + d_2}{2} \qquad \text{và } \sigma = \frac{d_2 - d_1}{4}$$

Tìm xác suất để viên bi bị loại

ĐS: p = 0,0456

9. QUY LUẬT KHI BÌNH PHƯƠNG - $\chi^2(n)$

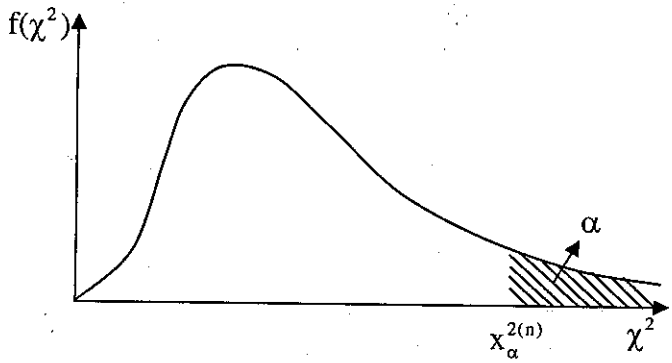
Biến ngẫu nhiên liên tục χ^2 gọi là phân phối theo quy luật khi bình phương với n bậc tự do nếu hàm mật độ xác suất của nó được xác định bằng biểu thức sau:

$$f(x) = \begin{cases} 0 & \text{với } x \le 0 \\ \dfrac{1}{2^{\frac{n}{2}}.\Gamma\left(\frac{n}{2}\right)} e^{\frac{x}{2}}.x^{\frac{n}{2}-1} & \text{với } x > 0 \end{cases} \qquad (3.38)$$

Trong đó $\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt$ là hàm Gamma. Nếu n là một số nguyên thì

$\Gamma(n+1) = n!$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Hình 3.8. **Đồ thị hàm $f(\chi^2)$ của quy luật "khi bình phương"**

Đồ thị của hàm $f(\chi)$ có dạng như ở hình 3.8. Có thể chứng minh được rằng nếu biến ngẫu nhiên χ^2 phân phối theo quy luật khi bình phương với n bậc tự do thì kỳ vọng toán:

$$E(\chi^2) = n$$

Và phương sai

$$V(\chi^2) = 2n$$

Ngoài ra trong quy luật "khi bình phương" giá trị tới hạn χ^2 cũng là tham số được sử dụng nhiều. Giá trị tới hạn "khi bình phương" mức α ký hiệu là $\chi_\alpha^{2(n)}$ là giá trị của biến ngẫu nhiên χ^2 tuân theo quy luật phân phối "khi bình phương" với n bậc tự do thoả mãn điều kiện:

$$P(\chi^2 > \chi_\alpha^{2(n)}) = \alpha$$

Các giá trị tới hạn $\chi_\alpha^{2(n)}$ được tính sẵn thành bảng (Phụ lục 7).

Ý nghĩa của nó được thấy rõ ở hình 3.8.

Khi số bậc tự do n tăng lên, quy luật "khi bình phương" sẽ xấp xỉ với quy luật chuẩn.

Quy luật khi bình phương có tính chất sau đây: Nếu χ_1^2 và χ_2^2 là các biến ngẫu nhiên độc lập cùng phân phối theo quy luật "khi bình phương" với số bậc tự do tương ứng là n_1 và n_2 thì tổng của chúng là biến ngẫu nhiên:

$$\chi^2 = \chi_1^2 + \chi_2^2$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Cũng phân phối theo quy luật khi bình phương với số bậc tự do là $n = n_1 + n_2$.

Trong thực tế quy luật khi bình phương thường được sử dụng trong trường hợp sau đây: Giả sử có các biến ngẫu nhiên $x_i (i = \overline{1,n})$ độc lập, cùng phân phối theo quy luật chuẩn hoá tức là có kỳ vọng toán bằng không và độ lệch chuẩn bằng một. Nếu xét tổng bình phương của các biến ngẫu nhiên nói trên ta có:

$$\chi^2 = \sum_{i=1}^{n} x_i^2$$

Biến ngẫu nhiên χ^2 sẽ phân phối theo quy luật khi bình phương với n bậc tự do.

Như vậy, nếu biến ngẫu nhiên X phân phối chuẩn hoá thì biến ngẫu nhiên X^2 sẽ phân phối theo quy luật χ^2 với 1 bậc tự do.

10. QUY LUẬT STUDENT - T(n)

Biến ngẫu nhiên liên tục T gọi là phân phối theo quy luật Student với n bậc tự do nếu hàm mật độ xác suất của nó được xác định bằng biểu thức sau:

$$f(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\Gamma\left(\frac{n-1}{2}\right)}\left[1+\frac{t^2}{n-1}\right]^{-\frac{n}{2}} \forall t \qquad (3.39)$$

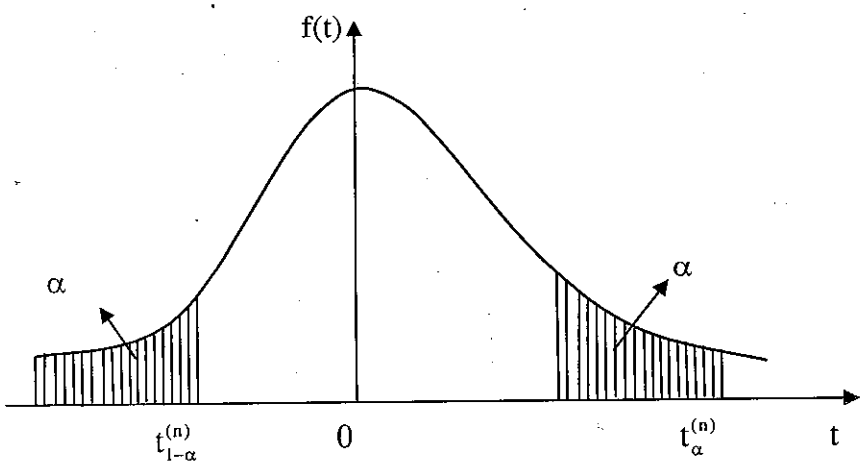
Trong đó $\Gamma(x)$ là hàm Gamma.

Đồ thị của hàm f(t) có dạng như ở hình 3.9.

Có thể chứng minh được rằng nếu biến ngẫu nhiên T phân phối theo quy luật Student với n bậc tự do thì kỳ vọng toán E(T)=0 và phương sai

$$V(T) = \frac{n}{n-2}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



***Hình 3.9.* Đồ thị hàm f(t) của quy luật Student**

Giá trị tới hạn Student, ký hiệu là $t_{\alpha}^{(n)}$ là giá trị của biến ngẫu nhiên T phân phối theo quy luật Student với n bậc tự do, thoả mãn điều kiện:

$$P(T > t_{\alpha}^{(n)}) = \alpha$$

Các giá trị tới hạn $t_{\alpha}^{(n)}$ được tính sẵn thành bảng (Phụ lục 8).

Giá trị tới hạn Student có tính chất sau đây:

$$t_{1-\alpha}^{(n)} = -t_{\alpha}^{(n)}$$

Ý nghĩa của nó được thể hiện trên hình 3.9.

Khi số bậc tự do n tăng lên, phân phối Student sẽ hội tụ rất nhanh về phân phối chuẩn hoá. Do đó nếu n khá lớn (n > 30) có thể dùng phân phối chuẩn hoá thay thế cho phân phối Student.

Tuy nhiên cần phải nhấn mạnh rằng với số bậc tự do nhỏ ($n \leq 30$) việc thay thế quy luật Student bằng quy luật chuẩn có thể dẫn đến những sai số rất lớn. Chẳng hạn với $n = 4$ và $\alpha = 0,05$ thì giá trị tới hạn Student $t_{0,05}^{(4)} = 4,6$ trong khi đó $U_{0,05} = 2,58$, tức là sai lệch nhau tới $4,6 - 2,58 = 2,02$.

Trong thực tế quy luật Student thường được sử dụng trong trường hợp sau đây: Giả sử có U là biến ngẫu nhiên phân phối chuẩn hoá N(0, 1) và biến ngẫu nhiên V độc lập với U, phân phối theo quy luật khi bình phương với n bậc tự do.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nếu xét biến ngẫu nhiên

$$T = \frac{U}{\sqrt{\frac{V}{n}}}$$

Thì biến ngẫu nhiên T sẽ phân phối theo quy luật Student với n bậc tự do.

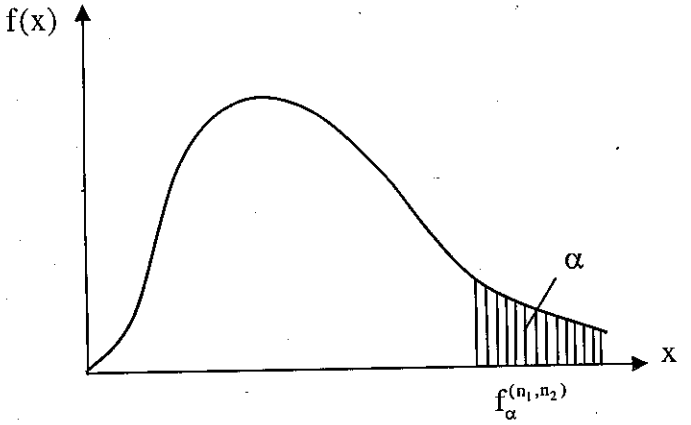
11. QUY LUẬT FISHER - SNEDECOR - $F(n_1, n_2)$

Biến ngẫu nhiên liên tục F gọi là phân phối theo quy luật Fisher - Snedecor với n_1 và n_2 bậc tự do nếu hàm mật độ xác suất của nó được xác định bằng biểu thức sau:

$$f(x) = \begin{cases} 0 & \text{với } x \le 0 \\ C\dfrac{x^{\frac{n_1-n_2}{2}}}{(n_2+n_1.x)^{\frac{(n_1-n_2)}{2}}} & \text{với } x > 0 \end{cases} \qquad (3.40)$$

Với $$C = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right).n_1^{\frac{n_1}{2}}.n_2^{\frac{n_2}{2}}}{\Gamma\left(\frac{n_1}{2}\right).\Gamma\left(\frac{n_2}{2}\right)}$$

Đồ thị của hàm f(x) có dạng sau (xem hình 3.10)



Hình 3.10. **Đồ thị của hàm f(x) của phân phối Fisher - Snedecor**

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Có thể chứng minh được rằng trong quy luật Fisher - Snedecor thì:

$$E(F) = \frac{n_2}{n_2 - 2}$$

Và

$$V(F) = \frac{2n_2^2(n_1 + n_2^2 \div 2)}{n_1(n_2 - 2)^2(n_2 - 4)}$$

Giá trị tới hạn Fisher - Snedecor ký hiệu là $f_\alpha^{(n_1,n_2)}$, là giá trị của biến ngẫu nhiên F phân phối theo quy luật Fisher - Snedecor với n_1 và n_2 bậc tự do, thoả mãn điều kiện:

$$P(F > f_\alpha^{(n_1,n_2)}) = \alpha$$

Giá trị $f_\alpha^{(n_1,n_2)}$ có tính chất sau đây:

$$f_{1-\alpha}^{(n_1,n_2)} = \frac{1}{f_\alpha^{(n_2,n_1)}}$$

Ý nghĩa của nó được thể hiện trên hình 3.10.

Các giá trị tới hạn $f_\alpha^{(n_1,n_2)}$ được tính sẵn thành bảng (Phụ lục 9).

Trong thực tế quy luật Fisher - Snedecor thường được sử dụng trong trường hợp sau: Giả sử có các biến ngẫu nhiên U và V độc lập với nhau và cùng phân phối theo quy luật khi bình phương với các bậc tự do tương ứng là n_1 và n_2. Lúc đó nếu xét biến ngẫu nhiên:

$$F = \frac{\frac{U}{n_1}}{\frac{V}{n_2}}$$

thì F sẽ phân phối theo quy luật Fisher - Snedecor với n_1 và n_2 bậc tự do.

Mặt khác, có thể chứng minh được rằng nếu biến ngẫu nhiên T phân phối Student với n bậc tự do thì biến ngẫu nhiên T^2 sẽ phân phối Fisher-Snedecor với số bậc tự do là 1 và n.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

12. TÓM TẮT CHƯƠNG 3

1. Mỗi biến ngẫu nhiên trong thực tế đều tuân theo một quy luật phân phối xác suất nào đó với các tham số đặc trưng tương ứng.
2. Quy luật không - một thường được dùng để mô tả các hiện tượng kinh tế xã hội định tính với hai phạm trù luân phiên.
3. Sự khác biệt cơ bản của các lược đồ dẫn đến quy luật nhị thức và quy luật siêu bội là trong quy luật nhị thức thì các phép thử là độc lập còn trong quy luật siêu bội thì các phép thử là phụ thuộc.
4. Quy luật Poisson thường được dùng để thay thế cho quy luật nhị thức khi số phép thử là lớn và xác suất p lại nhỏ.
5. Quy luật phân phối đều thường được dùng khi có rất ít thông tin về biến ngẫu nhiên, vì vậy có thể xem mọi trạng thái của biến ngẫu nhiên là đồng xác suất.
6. Quy luật chuẩn được xem là quy luật giới hạn của tất cả các quy luật phân phối xác suất khác.
7. Quy luật chuẩn và quy luật Student là các quy luật đối xứng còn quy luật Khi bình phương và quy luật Fisher là các quy luật bất đối xứng.

Các ký hiệu và công thức cơ bản

* Quy luật không - một - A(p)

$$P(x) = p^x(1-p)^{1-x} \qquad x = \overline{0,1}$$

$$E(X) = p; V(X) = p(1-p)$$

* Quy luật nhị thức - B(n,p)

$$P_x = C_n^x p^x (1-p)^{n-x} \qquad x = \overline{0,n}$$

$$E(X) = np; V(X) = np(1-p)$$

$$np + p - 1 \le m_0 \le np + p$$

* Quy luật Poisson - P(λ)

$$P_x = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0,1,2,...$$

$$E(X) = \lambda; V(X) = \lambda$$

$$\lambda - 1 \le m_0 \le \lambda$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

* Quy luật siêu bội - M(N, n)

$$P_x = \frac{C_M^x . C_{N-M}^{n-x}}{C_N^n} \qquad x = \overline{0,n}$$

$$E(X) = n\frac{M}{N} = np; V(X) = np(1-p)\frac{N-n}{N-1}$$

* Quy luật phân phối đều - U(a, b)

$$f(x) = \begin{cases} 0 & x \notin (a,b) \\ \dfrac{1}{b-a} & x \in (a,b) \end{cases}$$

$$E(X) = \frac{a+b}{2}, V(X) = \frac{(b-a)^2}{12}$$

* Quy luật phân phối luỹ thừa - E(λ)

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

$$E(X) = \frac{1}{\lambda}, V(X) = \frac{1}{\lambda^2}$$

* Quy luật phân phối chuẩn - $N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < +\infty)$$

$$E(X) = \mu; V(X) = \sigma^2$$

Nếu $U = \dfrac{X-\mu}{\sigma}$ thì U phân phối chuẩn hoá N(0, 1)

Với $\varphi(u) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$

Và $E(U) = 0; V(U) = 1$

$$P(a < X < b) = \Phi_0\left(\frac{b-\mu}{\sigma}\right) - \Phi_0\left(\frac{a-\mu}{\sigma}\right)$$

$$P(|X-\mu| < \varepsilon = 2\Phi_0\left(\frac{\varepsilon}{\sigma}\right)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Quy tắc 2σ và 3σ

$P(|X-\mu|<2\sigma)=2\Phi_0(2)=0,9544$

$P(|X-\mu|<3\sigma)=2\Phi_0(3)=0,9973$

Nếu $n>5$ và $\left|\sqrt{\frac{p}{1-p}}-\sqrt{\frac{1-p}{p}}\right|\times\frac{1}{\sqrt{n}}<0,3$ thì quy luật nhị thức hội tụ về quy luật chuẩn $(\mu=np;\sigma^2=np(1-p))$. Nếu $\lambda>20$ thì quy luật Poisson hội tụ về quy luật chuẩn $N(\mu=\lambda;\sigma^2=\lambda)$.

* Quy luật khi bình phương - $\chi^2(n)$

Nếu $X_1,X_2,...,X_n$ là các biến ngẫu nhiên độc nhau và cùng tuân theo quy luật chuẩn hoá N(0, 1) thì:

$$\chi^2=\sum_{i=1}^{n}X_i^2\sim\chi^2(n)$$

* Quy luật Student - T(n)

Nếu U và V độc lập, U~N(0,1) và V~χ^2(n) thì $T=\frac{U}{\sqrt{\frac{V}{n}}}\sim T(n)$

* Quy luật Fisher — Snedecor $F(n_1,n_2)$

Nếu U và V độc lập $U\sim\chi^2(n_1)$ và $V\sim\chi^2(n_2)$ thì:

$$F=\frac{\frac{U}{n_1}}{\frac{V}{n_2}}\sim F(n_1,n_2)$$

Nếu $T\sim T(n)$ thì $T^2\sim F(1,n)$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Quy luật không – một	Alternative (Bernoully) distribution
Dấu hiệu nghiên cứu định tính	Qualitative variable
Quy luật nhị thức	Binomial distribution
Quy luật nhị thức theo tỷ lệ	Binomial distribution of proportion
Quy luật Poisson	Poisson distribution
Chuỗi Poisson	Poisson series
Lý thuyết phục vụ công cộng	The queue theory
Lý thuyết quản lý dự trữ	Inventory managment theory
Quy luật siêu bội	Hypergeometric distribution
Hệ số hiệu chỉnh	Adjustment coefficient
Quy luật phân phối đều	Uniform distribution
Quy luật phân phối lũy thừa	Exponential distribution
Quy luật phân phối chuẩn	Normal distribution
Quy luật phân phối chuẩn hóa	Stadard normal distribution
Phép chuẩn hóa biến ngẫu nhiên	Standardisation of random variable
Tích phân Poisson	Poisson integral
Hàm Laplace	Laplace function
Quy tắc hai xích ma	Two sigma rule
Quy tắc ba xích ma	Three sigma rule
Định lý địa phương Laplace	Local laplace theorem
Định lý tích phân Laplace	Integral laplace theorem
Quy luật khi bình phương	Chi squares distribution
Số bậc tự do	Degree of freedom
Quy luật Student	T-Student distribution
Quy luật Fisher - Snedecor	Fisher-Snedecor distribution

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

3.50. Cho một thí dụ trong kinh tế hoặc kinh doanh về:

a. Một biến ngẫu nhiên phân phối A(p)

b. Một biến ngẫu nhiên phân phối B(n, p)

c. Một biến ngẫu nhiên phân phối $P(\lambda)$

d. Một biến ngẫu nhiên phân phối M(N, n)

3.51. Hai kiện tướng bóng bàn ngang sức thi đấu với nhau. Hỏi thắng 2 trong 4 ván dễ hơn hay thắng 3 trong 6 ván dễ hơn?

3.52. Một cầu thủ nổi tiếng về đá phạt đền với xác suất đá vào gôn là 4/5. Có người cho rằng vậy thì cứ sút 5 quả thì chắc chắn có 4 quả vào gôn. Điều khẳng định đó có đúng không? Tại sao?.

5.53. Hãy cho biết các mệnh đề sau đây là đúng hay sai? Tại sao?.

a. Tổng của một số hữu hạn n các biến ngẫu nhiên phân phối A(p) sẽ phân phối B(n, p).

b. Nếu $X_1 \sim B(n_1, p)$ và $X_2 \sim B(n_2, p)$ thì:

$$X_1 + X_2 \sim B(n_1 + n_2, p).$$

3.54. Với những điều kiện nào thì quy luật nhị thức B(n, p) sẽ hội tụ về quy luật Poisson $P(\lambda)$?

3.55. Với những điều kiện nào thì quy luật siêu bội M(N, n) xấp xỉ quy luật nhị thức B(n, p).

3.56. Cho một thí dụ trong kinh tế hoặc kinh doanh về:

a. Một biến ngẫu nhiên phân phối U(a, b)

b. Một biến ngẫu nhiên phân phối $N(\mu, \sigma^2)$.

3.57. Phân phối chuẩn là rời rạc hay liên tục? Tại sao?

3.58. a. Vẽ phác qua đồ thị hàm mật độ xác suất của phân phối chuẩn hoá.

b. Cho u_α là một giá trị thoả mãn $\Phi(u_\alpha) = \alpha$. Hãy minh hoạ bằng đồ thị các giá trị $\alpha, u_\alpha, \Phi_0(u_\alpha), \varphi(u_\alpha), P(U > u_\alpha)$ và $P(a < U < b)$ với a và b là hai giá trị nào đó U.

3.59. Cho hai biến ngẫu nhiên cùng phân phối chuẩn với kỳ vọng toán bằng nhau song phương sai khác nhau. Nếu cùng lấy ở mức 68,8% các giá trị của hai biến ngẫu nhiên ấy thì khoảng giá trị của biến ngẫu nhiên nào sẽ rộng hơn. Hãy mô tả kết luận bằng đồ thị.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.60. Biến ngẫu nhiên X tuân theo quy luật chuẩn với kỳ vọng toán bằng 10. Xác suất để X nhận giá trị trong khoảng (10;20) là 0,3. Không cần tính toán, hãy cho biết xác suất để X nhận giá trị trong khoảng (0;10) bằng bao nhiêu ?

3.61. Biến ngẫu nhiên liên tục X tuân theo quy luật chuẩn với kỳ vọng toán $\mu = 3$ và độ lệch chuẩn $\sigma = 2$. Viết hàm mật độ xác suất của biến ngẫu nhiên đó.

3.62. Biến ngẫu nhiên liên tục X có hàm phân bố xác suất như sau:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-\mu)^2/2\sigma^2} dx$$

a. Tìm hàm mật độ xác suất của X, từ đó cho biết X phân phối theo quy luật nào?

b. Dùng bảng tính $P(0 \le X \le 0,92)$ khi $\mu = 0$ và $\sigma^2 = 1$

3.63. Hãy sử dụng bảng giá trị tới hạn chuẩn để tìm các xác suất sau:

a. P(U > 1,96) b. P(U > 1,64)

c. P(U < -1,64) d. P(U < 1,64)

e. P(1 < U < 1,5) f. P(-1< U < 2)

3.64. Hãy dùng bảng giá trị hàm $\Phi_0(x)$ và bảng giá trị tới hạn chuẩn để tìm các xác suất sau nếu cho biến ngẫu nhiên X phân phối chuẩn với $\mu = 10$ và $\sigma = 5$. So sánh kết quả thu được bằng hai phương pháp:

a. P(X > 20) b. P(20 < X < 25)

c. P(X < 10) d. P(12 < X < 24)

3.65. Cho các biến ngẫu nhiên X_1, X_2, X_3 độc lập nhau, cùng phân phối chuẩn với các kỳ vọng toán tương ứng là 10; 15; 20 và các phương sai tương ứng là 3; 6; 9. Tìm quy luật phân phối xác suất và các tham số đặc trưng của biến ngẫu nhiên $X = X_1 + X_2 + X_3$.

3.66. Người ta kiểm tra chất lượng 900 chi tiết. Xác suất để chi tiết đạt tiêu chuẩn là 0,9. Gọi X là số chi tiết đạt tiêu chuẩn trong số 900 chi tiết được kiểm tra thì có thể coi X phân phối theo quy luật chuẩn được không ? Vì sao?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.67. Dùng bảng giá trị tới hạn χ^2 để tìm các xác suất sau đây:

a. $P(\chi^{2(9)} > 2,7)$ b. $P(\chi^{2(15)} > -27,49)$

c. $P(\chi^{2(30)} < 18,49)$ d. $P(\chi^{2(25)} > -44,31)$

3.68. Cho hai biến ngẫu nhiên độc lập X_1 và X_2 cùng phân phối chuẩn hoá. Tìm xác xuất để biến ngẫu nhiên $Z = X_1^2 + X_2^2$ nhận giá trị lớn hơn 9,21.

3.69. Cho biến ngẫu nhiên X phân phối chuẩn hoá. Tìm xác suất để X^2 nhận giá trị nhỏ hơn 5,024.

3.70. Cho hai biến ngẫu nhiên độc lập và cùng phân phối theo quy luật khi bình phương với số bậc tự do tương ứng là 6 và 8. Tìm xác suất để tổng của hai biến ngẫu nhiên đó lớn hơn 21,06.

3.71. Dùng bảng giá trị tới hạn Student để tìm các xác suất sau:

a. P(T(15) > 2,602) b. P(T(25) > - 2,06)

c. P(T(30) < 2,75) d. P(T(10) < - 2,228)

3.72. Cho hai biến ngẫu nhiên độc lập $U \sim N(0,1)$ và $V \sim \chi(5)$. Tìm xác suất để biến ngẫu nhiên $T = \dfrac{U}{\sqrt{\dfrac{V}{5}}}$ nhận giá trị lớn hơn 3,365.

3.73. Dùng bảng giá trị tới hạn Fisher để tìm các xác suất sau:

a. P(F(3,7) > 5,89)

b. P(F(4,9) < 4,72)

3.74. Dùng bảng giá trị giới hạn Fisher để tìm các giá trị sau:

a. $F_{0,05}^{(5,7)}$

b. $F_{0,025}^{(3,32)}$

c. $F_{0,95}^{(38,6)}$

3.75. Cho hai biến ngẫu nhiên độc lập U và V cùng phân phối theo quy luật khi bình phương với số bậc tự do tương ứng là 3 và 5. Tìm xác suất để biến ngẫu nhiên $F = \dfrac{V/5}{U/3}$ lớn hơn 5,31.

3.76. Cho biến ngẫu nhiên $T \sim T(7)$. Tìm xác suất để $T^2 > 5,59$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập tổng hợp chương 3

3.77. Xác suất để máy bị hỏng trong một ngày hoạt động là 0,01. Mỗi lần máy hỏng chi phí sửa chữa hết khoảng 1 triệu đồng. Vậy có nên ký một hợp đồng bảo dưỡng là 120 ngàn đồng một tháng để giảm xác suất hỏng của máy đi một nửa hay không và nếu ký thì hiệu quả mang lại là bao nhiêu. Giả sử máy hoạt động liên tục cả năm.

ĐS: Nên ký vì sẽ tiết kiệm được khoảng 385 ngàn đồng một năm.

3.78. Thời gian bảo hành sản phẩm được quy định là 3 năm. Nếu bán được một sản phẩm thì cửa hàng lãi 150 ngàn song nếu sản phẩm bị hỏng trong thời gian bảo hành thì cửa hàng phải chi phí 500 ngàn cho việc bảo hành. Biết rằng tuổi thọ của sản phẩm là biến ngẫu nhiên phân phối chuẩn với tuổi thọ trung bình là 4,2 năm và độ lệch chuẩn là 1,8 năm.

a. Tìm số tiền lãi mà cửa hàng hy vọng thu được khi bán mỗi sản phẩm.

b. Nếu muốn số tiền lãi cho mỗi sản phẩm bán ra là 50 ngàn thì phải quy định thời gian bảo hành là bao nhiêu.

ĐS: a. $E(X) = 24,3$ ngàn; b. Thời gian bảo hành 2,688 năm.

3.79. Ở một trạm bơm xăng bình quân mỗi giờ có 12 xe máy đến tiếp xăng. Tìm xác suất để trong một giờ nào đó có.

a. Hơn 8 xe đến tiếp xăng.

b. Hơn 15 xe đến tiếp xăng.

c. Dưới 10 xe đến tiếp xăng.

ĐS: $P_a = 0,845$; $P_b = 0,1556$; $P_c = 0,2424$

3.80. Tuổi thọ của một loại sản phẩm là biến ngẫu nhiên phân phối chuẩn với trung bình là 11 năm và độ lệch chuẩn là 2 năm.

a. Nếu quy định thời gian bảo hành là 10 năm thì tỷ lệ sản phẩm phải bảo hành là bao nhiêu.

b. Nếu muốn tỷ lệ sản phẩm phải bảo hành là 10% thì phải quy định thời gian bảo hành là bao nhiêu.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.81. Độ dài chi tiết (tính bằng cm) do một máy tự động sản xuất là biến ngẫu nhiên phân phối chuẩn với độ lệch chuẩn 9cm.

Nếu được biết 84,13% chi tiết do máy sản xuất có độ dài không vượt quá 84cm thì xác suất để lấy ngẫu nhiên 3 chi tiết được ít nhất 1 chi tiết có độ dài không dưới 80cm là bao nhiêu?

3.82. Hàng sáng đi tàu đến nơi làm việc, một người mê trò chơi ô chữ có đúng 30 phút để điền các ô chữ. Qua kinh nghiệm đối với ô chữ in trong báo A thì trung bình phải mất 25,2 phút để điền xong với độ lệch chuẩn là 3,9 phút. Với ô chữ in trong báo B thì trung bình cũng cần 25,2 phút để điền xong với độ lệch chuẩn là 1,9 phút. Biết thời gian điền xong ô chữ là biến ngẫu nhiên phân phối chuẩn. Vậy người này nên mua báo nào nhằm đảm bảo khả năng điền xong ô chữ cao hơn.

3.83. Cho Z là biến ngẫu nhiên tuân theo quy luật chuẩn hoá N (0, 1) và

$X = \alpha + \beta Z + \delta Z^2$

Trong đó α, β, δ là các hằng số.

Hãy tính E(X) và V(X).

3.84. Một người cân nhắc giữa việc mua nhà bây giờ hay gửi tiền vào tiết kiệm với lãi suất 12% một năm để chờ một năm sau mới mua. Biết mức tăng giá nhà là biến ngẫu nhiên phân phối chuẩn với kỳ vọng toán là 8% một năm và độ lệch chuẩn là 10% một năm. Tìm khả năng rủi ro của người đó nếu gửi tiền vào tiết kiệm và chờ một năm.

3.85. Thời gian cần thiết để hoàn thành một công việc là biến ngẫu nhiên phân phối đều U [5;9] (phút).

a. Viết biểu thức của hàm mật độ xác suất.

b. Tìm xác suất để công việc được hoàn thành dưới 8 phút.

c. Thời gian trung bình để hoàn thành công việc đó bằng bao nhiêu?

3.86. Một hãng quảng cáo tuyên bố rằng cứ 5 bác sĩ thì có hai người chỉ định cho bệnh nhân dùng loại thuốc do hãng đó quảng cáo. Chọn ngẫu nhiên 20 bác sĩ thì thấy chỉ có 2 người chỉ định cho bệnh nhân dùng loại thuốc đó.

a. Giả sử lời tuyên bố của hãng quảng cáo là đúng, tìm xác suất của biến cố trên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b. Giả sử lời tuyên bố của hãng quảng cáo là đúng, tìm xác suất để có không quá 2 người chỉ định dùng loại thuốc trên.

c. Với kết quả chọn ngẫu nhiên như trên, liệu có thể tin là lời tuyên bố đúng sự thật được không? Tại sao?

3.87. Lãi suất đầu tư vào một công ty là biến ngẫu nhiên phân phối chuẩn. Biết xác suất để đạt được lãi suất trên 20% một năm là 0,2 và dưới 10% một năm là 0,1. Tìm xác suất để khi đầu tư vào công ty đó sẽ được lãi suất ít nhất là 14% năm.

3.88. Chiều dài X và chiều rộng Y của một chi tiết được gia công một cách độc lập và là các biến ngẫu nhiên phân phối chuẩn với $\mu_x = 8cm$; $\mu_y = 4cm$; $\delta_x = 0,3$ và $\delta_y = 0,2cm$. Chi tiết được coi là đạt tiêu chuẩn nếu các kích thước của nó sai lệch so với kích thước trung bình không quá 0,1cm.

a. Tìm xác suất để chi tiết đạt tiêu chuẩn.

b. Tìm xác suất để khi gia công 3 chỉ tiết có ít nhất một chi tiết đạt tiêu chuẩn.

3.89. Thời gian hoạt động tốt (không phải sửa chữa) của một loại tivi là biến ngẫu nhiên phân phối chuẩn với $\mu = 4.300$ giờ và $\delta = 250$ giờ. Giả thiết mỗi ngày người ta dùng trung bình là 10 giờ và thời hạn bảo hành miễn phí là 1 năm (360 ngày).

a. Tìm tỷ lệ sản phẩm phải bảo hành.

b. Phải nâng chất lượng sản phẩm bằng cách tăng thời gian hoạt động tốt trung bình của sản phẩm lên bao nhiêu để tỷ lệ bảo hành vẫn như trên song có thể nâng thời gian bảo hành lên thành 2 năm?

3.90. Hội đồng quản trị của doanh nghiệp có 7 thành viên và mọi vấn đề đều được giải quyết theo đa số. Ông chủ tịch muốn thông qua một đề án kinh doanh do ông soạn thảo. Giả sử khả năng ủng hộ hay phản đối của mỗi thành viên hội đồng là như nhau.

a. Tìm xác suất để đề án được thông qua.

b. Giả sử trong hội đồng quản trị ngoài ông chủ tịch ra còn có 2 thành viên khác là đảng viên lập thành một chi bộ và chi bộ họp trù bị để thông qua đề án cũng theo nguyên tắc đa số. Sau đó ra cuộc họp chung thì mọi đảng viên phải tuân theo quyết định của cuộc họp trù bị. Vậy lúc đó xác suất để đề án được thông qua là bao nhiêu.

c. Từ đó có thể rút ra kết luận gì về việc áp dụng nguyên tắc tập trung dân chủ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.91. Máy bay bay dọc theo cầu dài 8m, rộng 4m và ném hai quả bom. Biết rằng khoảng cách từ điểm rơi của quả bom đến trục đối xứng theo chiều dọc và ngang của cầu là các biến ngẫu nhiên độc lập, cùng phân phối chuẩn với kỳ vọng toán bằng 0 và các độ lệch chuẩn tương ứng là 4m và 6m. Tìm xác suất để cầu bị trúng bom.

ĐS: p = 0,3445.

3.92. Trên một tuyến bay người ta dùng loại máy bay ATR 72 có 72 chỗ ngồi. Thực tế cho thấy đến giờ chót vẫn có khách bỏ chuyến bay. Để tận dụng hết chỗ ngồi bằng cách bán thêm vé dự phòng người ta đã thống kê 20 chuyến bay và thu được các số liệu sau.

Số khách bỏ chuyến bay	2	3	4	5	6	7	8	9	10	11
Số chuyến tương ứng	1	4	0	4	2	5	1	1	0	2

Hãy ước lượng xác suất để trong một chuyến bay nào đó có:

a. 1 khách bỏ chuyến bay đó.

b. 2 khách bỏ chuyến bay đó.

c. Tìm số khách bỏ chuyến trung bình và số khách bỏ chuyến có khả năng xảy ra nhiều nhất ở mỗi chuyến bay.

3.93. Một người cân nhắc giữa việc mua cổ phiếu của công ty A và công ty B hoạt động trong hai lĩnh vực độc lập nhau. Biết lãi suất cổ phiếu của hai công ty là các biến ngẫu nhiên phân phối chuẩn với các tham số đặc trưng như sau:

	Kỳ vọng toán (%)	Độ lệch chuẩn (%)
Công ty A	11	4
Công ty B	10,4	2,6

a. Vậy nếu người đó muốn đạt được lãi suất tối thiểu là 10% thì nên mua cổ phiếu của công ty nào.

b. Nếu người đó muốn hạn chế rủi ro bằng cách mua cổ phiếu của cả hai công ty thì nên mua theo tỷ lệ bao nhiêu để mức độ rủi ro về lãi suất là nhỏ nhất.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như. *Thống kê toán học*. NXB Đại học và THCN. Hà Nội. 1984
2. Lê Văn Phong. *Giáo trình lý thuyết xác suất*. NXB ĐH Kinh tế quốc dân. Hà Nội. 2007.
3. Nguyễn Đình Cử, Nguyễn Cao Văn. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Thống kê. Hà Nội. 1991.
4. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội. 1999.
5. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB ĐH Kinh tế quốc dân. Hà Nội. 2008.
6. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà Nội. 2009.
7. Harald Cramer. *Phương pháp toán học trong thống kê*. NXB Khoa học và Kỹ thuật. Hà Nội. 1970
8. Гмурман В.Е. Теория вероятностей и математическая статистика. Москва. 1972.
9. Гнеденко Б. В. Теория вероятностей. Москва. 1969.
10. Paul Newbold: Statistics for Business & Economics (Fourth Edition). Pretice-Hall International, Inc.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 4

BIẾN NGẪU NHIÊN HAI CHIỀU HÀM CÁC BIẾN NGẪU NHIÊN

1. MỞ ĐẦU

Ở các chương trước, ta đã xét các biến ngẫu nhiên mà các giá trị có thể có của chúng được biểu diễn bằng một số. Đó là các biến ngẫu nhiên một chiều.

Ngoài các biến ngẫu nhiên một chiều, trong thực tế ta còn gặp các biến số mà các giá trị có thể có của chúng được xác định bằng hai, ba,..., n số. Những biến số này được gọi một cách tương ứng là các biến ngẫu nhiên hai chiều, ba chiều,..., n chiều.

Việc nghiên cứu các biến ngẫu nhiên nhiều chiều không những cho phép xác định tính quy luật của từng thành phần của nó mà còn phân tích được sự ảnh hưởng qua lại của các thành phần đó với nhau. Trong kinh tế và xã hội thì điều này đặc biệt quan trọng. Chẳng hạn nếu ta chỉ nghiên cứu một cách riêng rẽ các biến ngẫu nhiên một chiều là thu nhập và tiêu dùng thì chỉ xác lập được tính quy luật của bản thân các biến ngẫu nhiên đó mà thôi, song nếu muốn đánh giá ảnh hưởng của thu nhập đến tiêu dùng như thế nào thì lại phải đặt chúng trong một biến ngẫu nhiên hai chiều là thu nhập và tiêu dùng thì mới phân tích được.

Mục 2 sẽ đề cập đến khái niệm biến ngẫu nhiên nhiều chiều. Từ mục 3 đến mục 6 sẽ trình bày các phương pháp mô tả quy luật phân phối xác suất của biến ngẫu nhiên hai chiều. Mục 7 và 8 sẽ trình bày các tham số đặc trưng của biến ngẫu nhiên hai chiều. Mục 9 sẽ xem xét một trường hợp có nhiều ứng dụng quan trọng trong thực tế là phân phối chuẩn hai chiều. Mục 10 sẽ đề cập đến một vài nội dung cơ bản nhất của hàm các biến ngẫu nhiên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2. KHÁI NIỆM VỀ BIẾN NGẪU NHIÊN NHIỀU CHIỀU

Ta sẽ ký hiệu biến ngẫu nhiên hai chiều là (X, Y) trong đó X và Y được gọi là các thành phần của biến ngẫu nhiên hai chiều mà thực chất mỗi thành phần lại là một biến ngẫu nhiên một chiều. Như vậy, biến ngẫu nhiên hai chiều thực chất là hệ hai biến ngẫu nhiên X và Y được xét một cách đồng thời. Tương tự như vậy, biến ngẫu nhiên n chiều có thể xem xét như hệ của n biến ngẫu nhiên một chiều.

Thí dụ1. Một máy sản xuất một loại sản phẩm. Nếu kích thước của sản phẩm được đo bằng chiều dài X và chiều rộng Y thì ta có biến ngẫu nhiên hai chiều, còn nếu tính thêm cả chiều cao Z nữa thì ta có biến ngẫu nhiên ba chiều.

Thí dụ 2. Gọi X là thu nhập và Y là tiêu dùng của dân cư một nước. Khi kết hợp X và Y thành một hệ thống thì ta có biến ngẫu nhiên hai chiều (X,Y).

Thí dụ 3. Gọi X là sản lượng, Y là vốn và Z là số lao động của một doanh nghiệp. Khi kết hợp X, Y, Z thành một hệ thống thì ta có biến ngẫu nhiên ba chiều (X,Y,Z)

Trong thực tế, người ta cũng phân biệt các biến ngẫu nhiên nhiều chiều thành hai loại: rời rạc và liên tục. Các biến ngẫu nhiên nhiều chiều gọi là rời rạc nếu các thành phần của nó là rời rạc và gọi là liên tục nếu các thành phần của nó là liên tục.

Sau đây ta sẽ chỉ xét các biến ngẫu nhiên hai chiều.

3. BẢNG PHÂN PHỐI XÁC SUẤT CỦA BIẾN NGẪU NHIÊN HAI CHIỀU

Đối với các biến ngẫu nhiên hai chiều người ta cũng dùng bảng phân phối xác suất, hàm phân bố xác suất và hàm mật độ xác suất để mô tả quy luật phân phối xác suất của chúng.

Bảng phân phối xác suất đồng thời của biến ngẫu nhiên hai chiều rời rạc liệt kê các giá trị có thể có của nó và các xác suất tương ứng. Nó có dạng sau đây:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

X / Y	x_1	x_2	...	x_i	...	x_n
y_1	$P(x_1,y_1)$	$P(x_2,y_1)$	...	$P(x_i,y_1)$	...	$P(x_n,y_1)$
y_2	$P(x_1,y_2)$	$P(x_2,y_2)$	...	$P(x_i,y_2)$	...	$P(x_n,y_2)$
...	...	...	...	...	...	...
y_j	$P(x_1,y_j)$	$P(x_2,y_j)$	...	$P(x_i,y_j)$	...	$P(x_n,y_j)$
...	...	...	...	...	...	...
y_m	$P(x_1,y_m)$	$P(x_2y_m)$	...	$P(x_i,y_m)$		$P(x_n,y_m)$

Trong đó $x_i(\overline{1,n})$ là các giá trị có thể có của thành phần $X; y_j(j=\overline{1,m})$ là các giá trị có thể có của thành phần Y; còn $P(x_i,y_j)$ là xác suất đồng thời để biến ngẫu nhiên hai chiều (X,Y) nhận giá trị (x_i,y_j). Ta chú ý rằng để tạo nên một quy luật phân phối xác suất thì các xác suất đồng thời $P(x_i,y_j)$ phải thoả mãn điều kiện:

$$\begin{cases} P(x_i,y_j) \geq 0 \\ \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j) = 1 \end{cases} \quad \text{với } i=\overline{1,n};\ j=\overline{1,m} \qquad (4.1)$$

Biết bảng phân phối xác suất đồng thời của biến ngẫu nhiên hai chiều bao giờ cũng có thể tìm được bảng phân phối xác suất biên của mỗi thành phần. Bảng phân phối xác suất biên của thành phần X có dạng:

X	x_1	x_2	...	x_i	...	x_n
P	$P(x_1)$	$P(x_2)$	...	$P(x_i)$	...	$P(x_n)$

Trong đó:

$$P(x_i) = \sum_{j=1}^{m} P(x_i,y_j) \qquad \text{với } i=\overline{1;n} \qquad (4.2)$$

được gọi là xác suất biên của thành phần X.

Rõ ràng là : $\sum_{i=1}^{n} P(x_i) = 1$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bảng phân phối xác suất biên của thành phần Y có dạng:

Y	y_1	y_2	...	y_j	...	y_m
P	$P(y_1)$	$P(y_2)$	...	$P(y_j)$	...	$P(y_m)$

Trong đó:

$$P(y_j) = \sum_{i=1}^{n} P(x_i, y_j) \qquad \text{với } j = \overline{1;m} \qquad (4.3)$$

được gọi là xác suất biên của thành phần Y.

Và $\sum_{j=1}^{m} P(y_j) = 1$

Thí dụ. Thu nhập hàng năm của các cặp vợ chồng có bảng phân phối xác suất đồng thời như sau:

Bảng 4.1

X \ Y	10	20	30	40
10	0,2	0,04	0,01	0
20	0,1	0,36	0,09	0
30	0	0,05	0,1	0
40	0	0	0	0,05

Trong đó X là thu nhập của chồng (triệu đồng/năm)

Y là thu nhập của vợ (triệu đồng/năm)

Tìm phân phối biên của mỗi thành phần.

Giải. Cộng các xác suất theo dòng ta thu được các xác suất tương ứng của các giá trị của thành phần X.

$$P(X_1) = 0,2 + 0,04 + 0,01 = 0,25$$

$$P(X_2) = 0,1 + 0,36 + 0,09 = 0,55$$

$$P(X_3) = 0.05 + 0,1 = 0,15$$

$$P(X_4) = 0,05$$

Vậy phân phối biên của thu nhập của chồng như sau:

X	10	20	30	40
Y	0,25	0,55	0,15	0,05

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tiến hành tương tự đối với từng cột của bảng 4.1 ta thu được phân phối biên của thu nhập của vợ như sau:

Y	10	20	30	40
P	0,3	0,45	0,2	0,05

Bài tập

4.1. Cho bảng phân phối xác suất đồng thời của biến ngẫu nhiên hai chiều rời rạc.

Y \ X	26	30	41	50
2,3	0,05	0,08	0,12	0,04
2,7	0,09	0,30	0,11	0,21

Tìm bảng phân phối xác suất biên của các thành phần X và Y.

ĐS:

X	26	30	41	50
P	0,14	0,38	0,23	0,25

Y	2,3	2,7
P	0,29	0,71

4.2. Thống kê dân số của một nước theo trình độ học vấn X và lứa tuổi Y cho kết quả sau:

Học vấn X \ Lứa tuổi Y		(25 - 35) 30	(35 - 55) 45	(55 - 100) 70
Thất học	0	0,01	0,02	0,05
Tiểu học	1	0,03	0,06	0,10
Trung học	2	0,18	0,21	0,15
Đại học	3	0,07	0,08	0,04

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Tìm xác suất để lấy ngẫu nhiên 1 người thì người đó ở độ tuổi 30 và tốt nghiệp trung học.

b. Xây dựng bảng phân phối xác suất biên của học vấn và của lứa tuổi.

ĐS: a. 0,18

b. p(x) = 0,08; 0,19; 0,54; 0,19.

P(y) = 0,29; 0,37; 0,34.

4.3. Điều tra thu nhập hàng năm (đơn vị triệu đồng) của các cặp vợ chồng đang làm việc thu được kết quả sau (X - Thu nhập của chồng; Y - thu nhập của vợ).

Y \ X	10	20	30	40
10	0,20	0,04	0,01	0
20	0,10	0,36	0,09	0
30	0	0,05	0,10	0
40	0	0	0	0,05

Tìm phân phối biên của thu nhập của chồng và của vợ.

ĐS: P(x) = 0,25; 0,55; 0,15; 0,05

P(y) = 0,30; 0,45; 0,20; 0,05

4. HÀM PHÂN BỐ XÁC SUẤT CỦA BIẾN NGẪU NHIÊN HAI CHIỀU

Xét biến ngẫu nhiên hai chiều (X, Y) có thể là rời rạc hoặc liên tục.

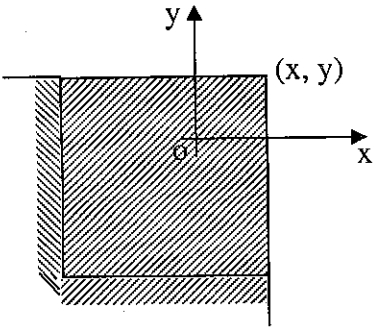
Giả sử x và y là một cặp số thực bất kỳ. Xét biến cố (X<x;Y<y) là biến cố để X nhận giá trị nhỏ hơn x và Y nhận giá trị nhỏ hơn y. Hiển nhiên là khi x và y thay đổi thì xác suất của biến cố trên cũng thay đổi theo, tức là nó là một hàm số của x, y.

Hàm phân bố xác suất đồng thời của biến ngẫu nhiên hai chiều (X,Y), ký hiệu là F(x,y) là xác suất để thành phần X nhận giá trị nhỏ hơn x và thành phần Y nhận giá trị nhỏ hơn y với x,y là các số thực tuỳ ý:

$$F(x,y) = P(X < x, Y < y) \qquad (4.4)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Về mặt hình học giá trị của hàm phân bố xác suất đồng thời tại mỗi điểm (x,y) là xác suất để biến ngẫu nhiên (X,Y) nhận giá trị tại một góc phẳng có đỉnh là (x,y) và nằm ở bên dưới và bên trái đỉnh đó.



***Hình 4.1.* Đồ thị hàm F(x, y)**

Thí dụ 1. Tìm xác suất để trong kết quả của phép thử thành phần X của biến ngẫu nhiên hai chiều (X, Y) nhận giá trị $X < 2$ và thành phần Y nhận giá trị $Y < 3$ nếu biết hàm phân bố xác suất của nó có dạng:

$$F(x,y) = \left(\frac{1}{\pi}\operatorname{arc\,tg}\frac{x}{2} + \frac{1}{2}\right) \times \left(\frac{1}{\pi}\operatorname{arc\,tg}\frac{y}{3} + \frac{1}{2}\right)$$

Giải. Theo định nghĩa hàm phân bố xác suất của biến ngẫu nhiên hai chiều, ta có:

$$P(X < 2, Y < 3) = F(2,3) = \left(\frac{1}{\pi}\operatorname{arc\,tg}\frac{2}{2} + \frac{1}{2}\right) \times \left(\frac{1}{\pi}\operatorname{arc\,tg}\frac{3}{3} + \frac{1}{2}\right)$$

$$= \left(\frac{1}{\pi}\frac{\pi}{4} + \frac{1}{2}\right) \times \left(\frac{1}{\pi}\frac{\pi}{4} + \frac{1}{2}\right) = \frac{3}{4}\cdot\frac{3}{4} = \frac{9}{16}$$

Hàm phân bố xác suất đồng thời của biến ngẫu nhiên hai chiều có các tính chất sau đây:

Tính chất 1. Giá trị của hàm phân bố xác suất đồng thời luôn nằm trong đoạn [0,1].

Tính chất 2. Hàm phân bố xác suất đồng thời là hàm không giảm theo từng đối số, tức là:

$$F(x_2, y) \geq F(x_1, y) \text{ nếu } x_2 > x_1$$

$$F(x, y_2) \geq F(x, y_1) \text{ nếu } y_2 > y_1$$

Ta sẽ chứng minh rằng F(x,y) là hàm không giảm theo x, còn việc chứng minh nó là hàm không giảm theo y tiến hành tương tự. Chúng ta xét

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

biến cố $(X < x_2, Y < y)$. Biến cố này có thể tách ra thành tổng của hai biến cố xung khắc là $(X < x_1, Y < y)$ và $(x_1 \le X < x_2, Y < y)$. Do đó theo định lý cộng xác suất ta có:

$$P(X < x_2, Y < y) = P(X < x_1, Y < y) + P(x_1 \le X < x_2, Y < y).$$

Từ đó:

$$P(X < x_2, Y < y) - P(X < x_1, Y < y) = P(x_1 \le X < x_2, Y < y).$$

Hay

$$F(x_2, y) - F(x_1, y) = P(x_1 \le X < x_2, Y < y) \ge 0$$

Do đó: $F(x_2, y) \ge F(x_1, y)$.

Tính chất 3. Ta có biểu thức giới hạn sau:

$$F(-\infty, y) = 0;\ F(x, -\infty) = 0$$

$$F(-\infty, -\infty) = 0;\ F(+\infty, +\infty) = 1$$

Thật vậy, $F(-\infty, y)$ là xác suất của biến cố $(X < -\infty).(Y < y)$, song biến cố này là biến cố không thể có, do đó xác suất của nó bằng không. Tương tự như vậy:

$F(x, -\infty) = 0$; và $F(-\infty, \infty) = 0$

Biến cố $(X < +\infty).(Y < +\infty)$ là biến cố chắc chắn, do đó xác suất của nó bằng một: $F(+\infty, +\infty) = 1$.

Tính chất 4. Khi $y = +\infty$ *hàm phân bố xác suất đồng thời của hệ hai biến ngẫu nhiên trở thành hàm phân bố xác suất biên của riêng thành phần X:*

$$F(x, +\infty) = F_1(x).$$

Và khi $x = +\infty$ *hàm phân bố xác suất đồng thời của hai biến ngẫu nhiên trở thành hàm phân bố xác suất biên của riêng thành phần Y.*

$$F(+\infty, y) = F_2(y)$$

Thật vậy, vì biến cố $(Y < +\infty)$ là biến cố chắc chắn do đó $F(x, +\infty)$ chỉ xác định xác suất của biến cố $X < x$ tức là nó trở thành hàm phân bố xác suất của thành phần X.

Tương tự, biến cố $(X < +\infty)$ là biến cố chắc chắn do đó $F(+\infty, y)$ chỉ còn là xác suất để $Y < y$ và nó trở thành hàm phân bố xác suất của riêng thành phần Y.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tính chất trên cho phép ta luôn luôn có thể tìm được hàm phân bố xác suất biên của mỗi thành phần của biến ngẫu nhiên hai chiều khi biết hàm phân bố xác suất đồng thời của hệ hai biến ngẫu nhiên.

Từ các tính chất trên ta có thể suy ra một vài công thức sau đây:

- Xác suất để biến ngẫu nhiên (X,Y) nhận giá trị trong một dải $(x_1 < X < x_2, Y < y)$ bằng

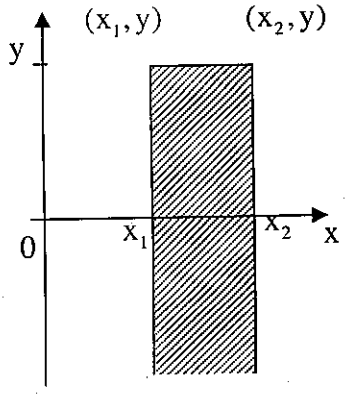
$$P(x_1 < X < x_2, Y < y) = F(x_2,y) - F(x_1,y) \quad \text{(hình 4.2)}$$

và trong dải $(X < x, y_1 < Y < y_2)$ bằng:

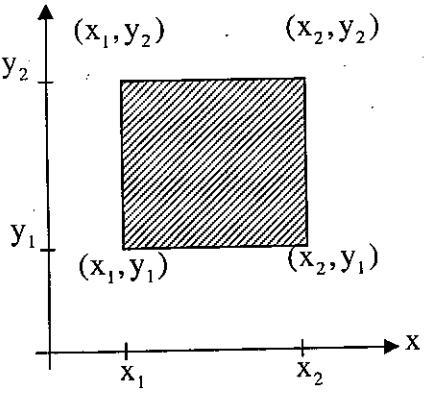
$$P(X < x, y_1 < Y < y_2) = F(x,y_2) - F(x,y_1) \quad \text{(hình 4.3).}$$

- Xác suất để biến ngẫu nhiên (X,Y) nhận giá trị trong hình chữ nhật $(x_1 < X < x_2, y_1 < Y < y_2)$ bằng.

$$P(x_1 < X < x_2, y_1 < Y < y_2) = F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1)$$



Hình 4.2



Hình 4.3

Thí dụ 2. Tìm xác suất để biến ngẫu nhiên hai chiều (X,Y) nhận giá trị trong hình chữ nhật giới hạn bởi các đường thẳng.

$$x_1 = \frac{\pi}{6}, x_2 = \frac{\pi}{2}, y_1 = \frac{\pi}{4}, y_2 = \frac{\pi}{3}$$

nếu biết hàm phân bố xác suất đồng thời:

$$F(x, y) = \sin x . \sin y \ (0 \le x \le \frac{\pi}{2}, 0 \le y \le \frac{\pi}{2})$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Theo công thức tính xác suất để (X, Y) nhận giá trị trong hình chữ nhật, ta có:

$$P(\frac{\pi}{6} < X < \frac{\pi}{2}, \frac{\pi}{4} < Y < \frac{\pi}{3}) = F(\frac{\pi}{2}, \frac{\pi}{3}) - F(\frac{\pi}{2}, \frac{\pi}{4}) - F(\frac{\pi}{3}, \frac{\pi}{6}) + F(\frac{\pi}{6}, \frac{\pi}{4})$$

$$= \sin\frac{\pi}{2}.\sin\frac{\pi}{3} - \sin\frac{\pi}{2}.\sin\frac{\pi}{4} - \sin\frac{\pi}{6}.\sin\frac{\pi}{3} + \sin\frac{\pi}{6}.\sin\frac{\pi}{4}$$

$$= \frac{\sqrt{3}}{2} - \frac{\sqrt{2}}{2} - \frac{1}{2}.\frac{\sqrt{3}}{2} + \frac{1}{2}.\frac{\sqrt{2}}{2} = \frac{\sqrt{3}-\sqrt{2}}{4} = 0{,}08$$

Bài tập

4.4. Tìm xác suất để biến ngẫu nhiên (X, Y) nhận giá trị trong hình chữ nhật giới hạn bởi các đường thẳng x = 1; x = 2; y = 3; y = 5 nếu biết hàm phân bố xác suất

$$F(x) = \begin{cases} 1 - 2^{-x} - 2^{-y} + 2^{-x-y} & \text{với } x \geq 0;\ y \geq 0 \\ 0 & \text{với } x < 0;\ y < 0 \end{cases}$$

ĐS: 0,0234

4.5. Tìm xác suất để thành phần X của biến ngẫu nhiên hai chiều nhận giá trị nhỏ hơn 1/2 và thành phần Y nhận giá trị nhỏ hơn 1/3 nếu biết hàm phân bố xác suất.

$$F(x,y) = ((1/\pi)\,\text{arctg}2x + 1/2)\,((1/\pi)\text{arctg}3y + 1/2)$$

ĐS: $P(X < 1/2,\ Y < 1/3) = 9/16$

4.6. Hai người bắn vào hai bia với xác suất trúng tương ứng là p_1 và p_2. Mỗi người bắn một viên đạn. Hãy xây dựng hàm phân bố xác suất của biến ngẫu nhiên (X, Y) trong đó X và Y tương ứng là số viên đạn trúng bia của từng người.

ĐS: Giá trị của F(x,y) được cho trong bảng sau:

Y \ X	$x \leq 0$	$0 < x \leq 1$	$x > 1$
$y \leq 0$	0	0	0
$0 < y \leq 1$	0	P_1P_2	P_2
$y > 1$	0	P_1	1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5. HÀM MẬT ĐỘ XÁC SUẤT CỦA BIẾN NGẪU NHIÊN HAI CHIỀU

Đối với biến ngẫu nhiên liên tục (X,Y) ngoài hàm phân bố xác suất ra còn có thể dùng hàm mật độ xác suất để biểu diễn quy luật phân phối xác suất của nó. Ta sẽ giả thiết rằng với biến ngẫu nhiên liên tục (X,Y) hàm phân bố xác suất luôn liên tục và có đạo hàm riêng hỗn hợp bậc hai ở mọi đường cong (có thể trừ một số đường cong nhất định).

Hàm mật độ xác suất đồng thời của biến ngẫu nhiên hai chiều liên tục (X, Y), ký hiệu là f(x, y), là đạo hàm riêng hỗn hợp bậc 2 của hàm phân bố xác suất đồng thời:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x.\partial y} \tag{4.5}$$

Về mặt hình học, hàm f(x, y) có thể xem như một mặt, được gọi là mặt phân phối xác suất.

Thí dụ 1. Tìm hàm mật độ xác suất đồng thời của biến ngẫu nhiên hai chiều liên tục (X, Y) nếu biết hàm phân bố xác suất đồng thời của nó:

$$F(x, y) = \sin x.\ \sin y\ (0 \le x \le \frac{\pi}{2}, 0 \le y \le \frac{\pi}{2})$$

Giải. Theo định nghĩa hàm mật độ xác suất đồng thời, trước hết ta tìm đạo hàm riêng của hàm phân bố xác suất đồng thời theo x:

$$\frac{\partial F(x,y)}{\partial x} = \cos x.\sin y$$

Từ kết quả thu được, ta tìm tiếp đạo hàm riêng của nó theo y, ta được:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x.\partial y} = \cos x.\cos y$$

với $(0 \le x \le \frac{\pi}{2}, 0 \le y \le \frac{\pi}{2})$

Hàm mật độ xác suất đồng thời của biến ngẫu nhiên hai chiều liên tục có các tính chất sau đây:

Tính chất 1. Hàm mật độ xác suất đồng thời luôn không âm

$$f(x, y) \ge 0$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tính chất này được suy ra trực tiếp từ chỗ nó là đạo hàm của hàm không giảm F(x, y) theo các thành phần của nó.

Tính chất 2. Xác suất để biến ngẫu nhiên hai chiều liên tục (X, Y) nhận giá trị trong một miền D được xác định bằng công thức:

$$P[(X,Y) \in D] = \iint_D f(x,y)dxdy$$

Thật vậy, theo tính chất của hàm phân bố xác suất đồng thời, ta có xác suất để biến ngẫu nhiên (X, Y) nhận giá trị trong một hình chữ nhật bằng:

$$P(x_1 < X < x_2, y_1 < Y < y_2)$$
$$= [F(x_2,y_2) - F(x_1,y_2)] - [F(x_2,y_1) - F(x_1,y_1)]$$

Gọi vế trái là P_{ABCD} và áp dụng định lý Lagrange đối với vế phải ta có:

$$P_{ABCD} = F''_{xy}(\xi,\eta)\Delta x\Delta y$$

Song theo định nghĩa của hàm mật độ xác suất thì:

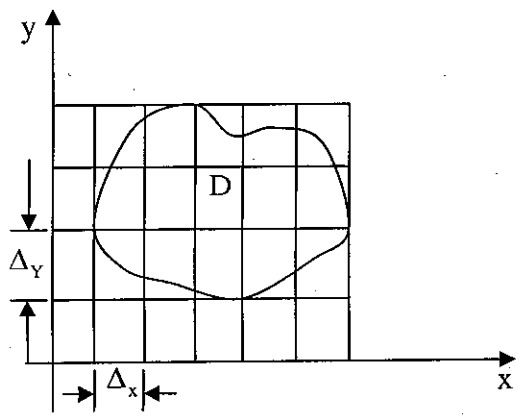
$$F''_{xy}(\xi,\eta) = f(\xi,\eta)$$

Do đó ta có:

$$P_{ABCD} = f(\xi,\eta).\Delta y\Delta y$$

Như vậy tích $f(\xi,\eta).\Delta_x\Delta_y$ chính là xác suất để biến ngẫu nhiên (X, Y) nhận giá trị trong hình chữ nhật ABCD với các cạnh là Δ_x và Δ_y.

Chia D thành các miền bằng các đường thẳng song song với trục 0Y cách nhau một đoạn Δ_x và các đường thẳng song song với trục 0X cách nhau một đoạn Δ_y (Hình 4.4)



Hình 4.4

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì các biến cố để biến ngẫu nhiên (X, Y) nhận giá trị trong n miền nói trên là xung khắc từng đôi nên xác suất để (X, Y) nhận giá trị trong miền D có thể tính xấp xỉ như sau:

$$P[(X,Y) \in D] = \sum_{i=1}^{n} f(\xi_i, \eta_i)\Delta x.\Delta y$$

Lấy giới hạn khi $\Delta_x \to 0$ và $\Delta_y \to 0$ ta có:

$$P[(X,Y) \in D] = \iint_D f(x,y)dxdy$$

Về mặt hình học, biểu thức trên có thể xem như xác suất để biến ngẫu nhiên (X, Y) nhận giá trị trong miền D bằng thể tích của khối giới hạn bởi mặt xác suất f(x, y) mà đáy của nó là hình chiếu của mặt đó trên mặt phẳng X0Y (hình 4.4).

Tính chất 3. Hàm phân bố xác suất đồng thời được xác định thông qua hàm mật độ xác suất đồng thời bằng biểu thức sau:

$$F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x, y)dxdy$$

Tính chất này suy ra trực tiếp từ định nghĩa của hàm mật độ xác suất đồng thời.

Tính chất 4. Tích phân suy rộng hai lớp của hàm mật độ xác suất đồng thời bằng một:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)dxdy = 1$$

Tích phân suy rộng từ - ∞ đến + ∞ cho thấy các miền lấy tích phân là toàn bộ mặt phẳng x0y. Mà biến cố để biến ngẫu nhiên (X, Y) nhận giá trị trên toàn mặt phẳng x0y là biến cố chắc chắn, nên xác suất của nó bằng một.

Thí dụ 2. Tìm hàm phân bố xác suất đồng thời của biến ngẫu nhiên (X, Y) theo hàm mật độ xác suất đồng thời sau đây:

$$f(x, y) = \frac{1}{\pi^2(1+x^2)(1+y^2)}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Theo tính chất của hàm mật độ xác suất đồng thời, ta có:

$$F(x, y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(x. y)dxdy = \frac{1}{\pi^2}\int_{-\infty}^{y}\frac{1}{1+y^2}\int_{-\infty}^{x}\frac{dx}{1+x^2}dy$$

$$= \frac{1}{\pi^2}\int_{-\infty}^{y}\frac{1}{1+y^2}\left(\text{arctgx}+\frac{\pi}{2}\right)dy = \left(\frac{1}{\pi}\text{arctgx}+\frac{1}{2}\right)\frac{1}{\pi}\int_{-\infty}^{y}\frac{dy}{1+y^2} =$$

$$= \left(\frac{1}{\pi}\text{arctgx}+\frac{1}{2}\right)\left(\frac{1}{\pi}\text{arctgy}+\frac{1}{2}\right)$$

Thí dụ 3. Với hàm mật độ xác suất đồng thời đã cho ở thí dụ trước, tìm xác suất để (X, Y) nhận giá trị trong hình chữ nhật với các đỉnh A(1, 1), B($\sqrt{3}$,1), C(1, 0) và D($\sqrt{3}$,0).

Giải. Theo tính chất của hàm mật độ xác suất đồng thời:

$$P[(X,Y)\in D] = \iint_D f(x,y)dxdy = \frac{1}{\pi^2}\int_0^1\left(\frac{1}{1+y^2}\int_1^{\sqrt{3}}\frac{dx}{1+x^2}\right)dy$$

$$= \frac{1}{\pi^2}\text{arctgx}\left|_1^{\sqrt{3}}\int_0^1\frac{dy}{1+y^2} = \frac{1}{\pi^2}\left(\frac{\pi}{3}-\frac{\pi}{4}\right)\text{arctgy}\right|_0^1 = \frac{1}{\pi^2}\cdot\frac{\pi}{12}\cdot\frac{\pi}{4} = \frac{1}{48}$$

Khi đã biết hàm mật độ xác suất đồng thời của hệ hai biến ngẫu nhiên, bao giờ ta cũng có thể xác định được hàm mật độ xác suất biên của từng thành phần của nó. Thực vậy, gọi hàm mật độ xác suất biên của thành phần X là $f_1(x)$ thì theo định nghĩa hàm mật độ xác suất, ta có:

$$f_1(x) = \frac{dF_1(x)}{dx} = \frac{dF(x,\infty)}{dx} = \frac{d}{dx}\int_{-\infty}^{x} f(x,y)dxdy = \int_{-\infty}^{+\infty} f(x,y)dy$$

Như vậy:
$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy \qquad (4.6)$$

Tương tự, ta có thể chứng minh được rằng:

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx \qquad (4.7)$$

Như vậy, hàm mật độ xác suất biên của một thành phần nào đó bằng tích phân suy rộng của hàm mật độ xác suất đồng thời của hệ trong đó biến lấy tích phân là thành phần kia của hệ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 4. Biến ngẫu nhiên hai chiều (X, Y) có hàm mật độ xác suất đồng thời như sau:

$$f(x,y)=\begin{cases}\dfrac{1}{6\pi} \text{ với} \dfrac{x^2}{9}+\dfrac{y^2}{4}<1\\ 0 \quad \text{với} \dfrac{x^2}{9}+\dfrac{y^2}{4}\geq 1\end{cases}$$

Tìm hàm mật độ xác suất biên của thành phần X và Y.

Giải. Theo công thức ta có:

$$f_1(x)=\int_{-\infty}^{+\infty} f(x,y)dy=\frac{1}{6\pi}\int_{-2\sqrt{1-\frac{x^2}{9}}}^{2\sqrt{1-\frac{x^2}{9}}} dy=\frac{2}{6\pi}\int_{0}^{2\sqrt{1-\frac{x^2}{9}}} dy=\frac{2}{9\pi}\sqrt{9-x^2}$$ Như vậy:

$$f_1(x)=\begin{cases}\dfrac{2}{9\pi}\sqrt{9-x^2} \text{ với} |x|<3\\ 0 \qquad\qquad \text{với} |x|\geq 3\end{cases}$$

Tương tự, ta tìm được

$$f_2(y)=\begin{cases}\dfrac{1}{2\pi}\sqrt{4-y^2} \text{ với} |y|<2\\ 0 \qquad\qquad \text{với} |y|\geq 2\end{cases}$$

Bài tập

4.7. Tìm hàm mật độ xác suất f(x,y) của biến ngẫu nhiên hai chiều liên tục (X, Y) nếu biết hàm phân bố xác suất của nó có dạng:

$$F(x,y)=\begin{cases}(1-e^{-2x})(1-e^{-3y}) & \text{với } x\geq 0;\ y\geq 0\\ 0 & \text{với } x<0;\ y<0\end{cases}$$

ĐS: $f(x,y)=\begin{cases}6e^{-(2x+3y)} & \text{với } x\geq 0;\ y\geq 0\\ 0 & \text{với } x<0;\ y<0\end{cases}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.8. Tìm hàm phân bố xác suất của biến ngẫu nhiên hai chiều (X, Y) nếu biết hàm mật độ xác suất của nó có dạng:

$$f(x,y) = \begin{cases} (1/2)\sin(x+y) & \text{với } x,y \in (0;\ \pi/2) \\ 0 & \text{với } x,y \notin (0;\ \pi/2) \end{cases}$$

ĐS: $F(x,y) = \begin{cases} \dfrac{1}{2}\left[\sin x + \sin y - \sin(x+y)\right] & \text{với } x,y \in (0;\ \pi/2) \\ 0 & \text{với } y \notin (0;\ \pi/2) \end{cases}$

4.9. Hệ hai biến ngẫu nhiên phân phối đều trong hình chữ nhật giới hạn bởi các đường thẳng x = 4; x = 6; y = 10; y = 15. Trong hình chữ nhật đó hàm mật độ xác suất bằng một giá trị cố định, còn ngoài hình chữ đó giá trị của nó bằng không. Tìm hàm mật độ xác suất và hàm phân bố xác suất của hệ trên.

ĐS: $f(x,y) = \begin{cases} 0{,}1 & \text{trong hình chữ nhật} \\ 0 & \text{ngoài hình chữ nhật} \end{cases}$

$$F(x,y) = \begin{cases} \dfrac{(x-4)(y-10)}{10} \end{cases}$$

4.10. Cho hàm mật độ xác suất của biến ngẫu nhiên hai chiều liên tục.

$f(x, y) = (1/\pi)e^{(-1/2)(x^2+2y+5y^2)}$

Tìm hàm mật độ xác suất của các thành phần.

ĐS: $f_1(x) = \sqrt{\dfrac{2}{5\pi}e^{-0,4x^2}}$

$f_2(y) = \sqrt{\dfrac{2}{\pi}e^{-y^2}}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6. QUY LUẬT PHÂN PHỐI XÁC SUẤT CÓ ĐIỀU KIỆN CỦA CÁC THÀNH PHẦN CỦA HỆ HAI BIẾN NGẪU NHIÊN

Ở các phần trên ta đã thấy quá trình phân tích hệ hai biến ngẫu nhiên thành các thành phần của nó luôn luôn có thể thực hiện được, tức là biết quy luật phân phối xác suất đồng thời của hệ hai biến ngẫu nhiên, ta luôn tìm được quy luật phân phối xác suất biên của từng thành phần của nó.

Vấn đề tổng hợp hai biến ngẫu nhiên một chiều thành hệ hai biến ngẫu nhiên được tiến hành phức tạp hơn. Để làm điều đó ta phải sử dụng khái niệm *phân phối xác suất có điều kiện*.

Xét biến ngẫu nhiên rời rạc (X, Y) trong đó các giá trị có thể có của thành phần X là $x_1, x_2, ..., x_n$, còn các giá trị có thể có của thành phần Y là $y_1, y_2, ..., y_m$. Gọi $P(x_i / y_j)$ $(i = \overline{1,n}; j = \overline{1,m})$ là xác suất có điều kiện để thành phần X nhận giá trị bằng x_i với điều kiện thành phần Y nhận giá trị bằng y_j.

Bảng phân phối xác suất có điều kiện của thành phần X với điều kiện $Y = y_j$ có dạng:

X/y_j	x_1	x_2	...	x_i	...	x_n
P	$P(x_1/y_j)$	$P(x_2/y_j)$	...	$P(x_i/y_j)$	...	$P(x_n/y_j)$

Trong đó các xác suất có điều kiện được tính bằng công thức:

$$P(x_i/y_j) = \frac{P(x_i, y_j)}{P(y_j)} \qquad i = \overline{1,n}; j = \overline{1,m} \tag{4.8}$$

Ta chú ý rằng:

$$\sum_{i=1}^{n} P(x_i/y_j) = \sum_{i=1}^{n} \frac{P(x_i, y_j)}{P(y_j)} = \frac{P(y_j)}{P(y_j)} = 1 \quad j = \overline{1,m}$$

Tức là các xác suất có điều kiện $P(x_i/y_j)$ cũng phải thoả mãn các yêu cầu của một quy luật phân phối xác suất.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tương tự, bảng phân phối xác suất có điều kiện của thành phần Y với điều kiện $X = x_i$ có dạng:

Y/x_i	y_1	y_2	...	y_j	...	y_m
P	$P(y_1/x_i)$	$P(y_2/x_i)$	...	$P(y_j/x_i)$	...	$P(y_m/x_i)$

Trong đó:

$$P(y_j/x_i) = \frac{P(x_i, y_j)}{P(x_i)} \quad i = \overline{1,n}; j = \overline{1,m} \tag{4.9}$$

Và $\sum_{j=1}^{m} P(y_j/x_i) = \sum_{j=1}^{m} \frac{P(x_i, y_j)}{P(x_i)} = \frac{P(x_i)}{P(x_i)} = 1 \quad i = \overline{1,n}$

Thí dụ 1. Phân phối xác suất của lương tháng Y (triệu đồng) và giới tính X của công nhân một công ty như sau:

Bảng 4.2

X \ Y	0,5	1	1,5
Nữ : 0	0,1	0,3	0.2
Nam : 1	0.06	0,18	0,16

Tìm phân phối xác suất của lương tháng của nữ công nhân.

Giải. Trước hết ta tìm $P(x_1) = 0,1 + 0,3 + 0,2 = 0,6$

Từ đó:

$$P(y_1/x_1) = \frac{P(x_1, y_1)}{P(x_1)} = \frac{0,10}{0,60} = \frac{1}{6}$$

$$P(y_2/x_1) = \frac{P(x_1, y_2)}{P(x_1)} = \frac{0,30}{0,60} = \frac{1}{2}$$

$$P(y_3/x_1) = \frac{P(x_1, y_3)}{P(x_1)} = \frac{0,20}{0,60} = \frac{1}{3}$$

Vậy bảng phân phối xác suất của lương tháng của nữ công nhân là:

$Y/X = x_1$	0,5	1	1,5
P	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giả sử (X, Y) là biến ngẫu nhiên hai chiều liên tục. Hàm mật độ xác suất có điều kiện của thành phần X với Y=y, ký hiệu f(x/y) là biểu thức:

$$f(x/y) = \frac{f(x,y)}{f_2(y)} = \frac{f(x,y)}{\int_{-\infty}^{\infty} f(x,y)dx} \qquad (4.10)$$

Tương tự, hàm mật độ xác suất có điều kiện của thành phần Y với X = x, ký hiệu f(y/x) là biểu thức:

$$f(y/x) = \frac{f(x,y)}{f_1(x)} = \frac{f(x,y)}{\int_{-\infty}^{\infty} f(x,y)dy} \qquad (4.11)$$

Ta chú ý rằng cũng như mọi hàm mật độ xác suất khác, các hàm mật độ xác suất có điều kiện cũng phải thoả mãn các điều kiện:

$$f(x/y) \geq 0; \int_{-\infty}^{+\infty} f(x/y)dx = 1$$

$$f(y/x) \geq 0; \int_{-\infty}^{+\infty} f(y/x)dy = 1$$

Thí dụ 2. Biến ngẫu nhiên liên tục (X, Y) có hàm mật độ xác suất đồng thời như sau:

$$f(x,y) = \begin{cases} \frac{1}{\pi r^2} & \text{với } x^2+y^2 < r^2 \\ 0 & \text{với } x^2+y^2 > r^2 \end{cases}$$

Tìm các hàm mật độ xác suất có điều kiện của các thành phần.

Giải. Theo công thức:

$$f(x/y) = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dx} = \frac{\frac{1}{\pi r^2}}{\frac{1}{\pi r^2}\int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} 1dr} = \frac{1}{2\sqrt{r^2-y^2}}$$

với $|x| < \sqrt{r^2-y^2}$

vì $f(x,y) = 0$ với $x^2+y^2 > r^2$ nên $f(x,y) = 0$

với $|x| > \sqrt{r^2-y^2}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tương tự, ta tìm được:

$$f(y/x)=\begin{cases}\dfrac{1}{2\sqrt{r^2-x^2}} & \text{với } |y|<\sqrt{r^2-x^2}\\ 0 & \text{với } |y|>\sqrt{r^2-x^2}\end{cases}$$

Trên cơ sở các phân phối xác suất có điều kiện, ta có các công thức tổng hợp hệ hai biến ngẫu nhiên theo phân phối xác suất của các thành phần như sau:

- Nếu (X, Y) là biến ngẫu nhiên hai chiều rời rạc thì:

$$P(x_i,y_j)=P(x_i)P(y_j/x_i)=P(y_j)P(x_i/y_j)\quad i=\overline{1,n}, j=\overline{1,m} \qquad (4.12)$$

- Còn nếu (X, Y) là biến ngẫu nhiên hai chiều liên tục thì:

$$f(x,y)=f_1(x)f(y/x)=f_2(y)f(x/y) \qquad (4.13)$$

Như vậy phân phối xác suất của hệ hai biến ngẫu nhiên bằng tích giữa phân phối xác suất biên của một thành phần với phân phối xác suất có điều kiện của thành phần còn lại.

- Nếu hai thành phần X và Y độc lập với nhau thì phân phối xác suất có điều kiện cũng bằng phân phối xác suất không điều kiện, lúc đó ta có công thức:

$$P(x_i,y_j)=P(x_i).P(y_j)\qquad i=\overline{1,n}, j=\overline{1,m} \qquad (4.14)$$

nếu (X, Y) là biến ngẫu nhiên rời rạc và

$$f(x,y)=f_1(x).f_2(y) \qquad (4.15)$$

nếu (X, Y) là biến ngẫu nhiên liên tục.

Chú ý rằng các hệ thức (4.14) và (4.15) là điều kiện cần và đủ để X và Y độc lập.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

4.11. Với số liệu của bài 4.1 hãy tìm bảng phân phối xác suất có điều kiện của Y khi X = 26 và của X khi Y = 2,7.

ĐS: a.

X	26	30	41	50
P	0,14	0,38	0,23	0,25

X	2,3	2,7
P	0,29	0,71

b.

Y/X = 26	2,3	2,7
P	0,357	0,643

X/Y = 2,7	26	30	41	50
P	0,1268	0,4225	0,1549	0,2958

4.12. Với số liệu của bài 4.2 hãy xây dựng phân phối xác suất của học vấn của những người ở tuổi 30.

ĐS: p(x/y = 30) = 1/29; 3/29; 18/29; 7/29

4.13. Với số liệu của bài 4.3 hãy tìm phân phối thu nhập của các bà vợ có chồng thu nhập 20 triệu/năm.

ĐS: P(y/x = 20) = 10/55; 36/55; 9/55

4.14. Với số liệu của bài 4.9 hãy tìm hàm mật độ xác suất có điều kiện của các thành phần.

ĐS:

$$f(x/y) = \frac{1}{\sqrt{2\pi}} e^{-(x+y)^2} \qquad f(y/x) = \frac{5}{\sqrt{2\pi}} e^{-0,1(x+5y)^2}$$

4.15. Hàm mật độ xác suất của biến ngẫu nhiên hai chiều có dạng

$$f(x,y) = \frac{3\sqrt{3}}{\pi} e^{-4x^2-6xy-9y^2}$$

Tìm hàm mật độ xác suất có điều kiện của các thành phần.

ĐS: $f(x/y) = \frac{2}{\sqrt{\pi}} e^{-(2x+\frac{3}{2}y)^2}$ $\qquad f(y/x) = \frac{3}{\sqrt{\pi}} e^{-(x+3y)^2}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7. CÁC THAM SỐ ĐẶC TRƯNG CỦA HỆ HAI BIẾN NGẪU NHIÊN

Đối với hệ hai biến ngẫu nhiên, các tham số đặc trưng cơ bản của nó trước hết là các kỳ vọng toán và phương sai của các thành phần. Các kỳ vọng toán được xác định bằng các công thức sau đây:

$$E(X) = \sum_{i=1}^{n} x_i P(x_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i P(x_i, y_j) \qquad (4.16)$$

$$E(Y) = \sum_{j=1}^{m} y_j P(y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} y_j P(x_i, y_j) \qquad (4.17)$$

nếu (X, Y) là biến ngẫu nhiên rời rạc, và

$$E(X) = \int_{-\infty}^{+\infty} x f_1(x)dx = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x f(x,y)dxdy \qquad (4.18)$$

$$E(Y) = \int_{-\infty}^{+\infty} y f_2(y)dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y f(x,y)dxdy \qquad (4.19)$$

nếu (X, Y) là biến ngẫu nhiên liên tục. Các phương sai được xác định bằng công thức:

$$V(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 P(x_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i^2 p(x_i, y_j) - [E(X)]^2 \qquad (4.20)$$

$$V(Y) = \sum_{j=1}^{m} [y_j - E(Y)]^2 P(y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} y_j^2 p(x_i, y_j) - [E(Y)]^2 \qquad (4.21)$$

nếu biến ngẫu nhiên (X, Y) rời rạc, và

$$V(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 f_1(x)dx = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^2 f(x,y)dxdy - [E(X)]^2 \qquad (4.2)$$

$$V(Y) = \int_{-\infty}^{+\infty} [y - E(Y)]^2 f_2(y)dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y^2 f(x,y)dxdy - [E(Y)]^2 \qquad (4.23)$$

nếu (X, Y) liên tục.

Ngoài các tham số trên, người ta còn thường xác định các tham số quan trọng khác là hiệp phương sai và hệ số tương quan.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hiệp phương sai, ký hiệu Cov(X, Y) của các biến ngẫu nhiên X và Y là kỳ vọng toán của tích các sai lệch của các biến ngẫu nhiên đó với kỳ vọng toán của chúng.

$$\mathrm{Cov}(X,Y) = E\{[X - E(X)][Y - E(Y)]\} \tag{4.24}$$

Để tìm hiệp phương sai của các biến ngẫu nhiên rời rạc, người ta thường dùng công thức sau:

$$\mathrm{Cov}(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(x_i, y_j) - E(X).E(Y) \tag{4.25}$$

và đối với các biến ngẫu nhiên liên tục:

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x.y f(x,y)dxdy - E(X).E(Y) \tag{4.26}$$

Từ định nghĩa trên, ta thấy hiệp phương sai có đơn vị đo lường bằng tích đơn vị đo lường của các biến ngẫu nhiên X và Y. Do đó, hiệp phương sai sẽ có các giá trị khác nhau tuỳ thuộc vào đơn vị đo lường của các biến đó.

Để khắc phục hạn chế này người ta đưa ra một tham số khác là hệ số tương quan.

Hệ số tương quan, ký hiệu ρ_{xy} là tỷ số giữa hiệp phương sai và tích các độ lệch chuẩn của các biến ngẫu nhiên đó.

$$\rho_{xy} = \frac{\mathrm{Cov}(X,Y)}{\sigma_x.\sigma_y} \tag{4.27}$$

Hệ số tương quan không có đơn vị đo và có các tính chất cơ bản sau đây:

1. $\rho_{xy} = \rho_{yx}$

2. $-1 \le \rho_{xy} \le 1$

3. Nếu $\rho_{xy} > 0$ thì X và Y là đồng biến còn $\rho_{xy} < 0$ thì X và Y là nghịch biến.

4. Nếu X và Y độc lập thì $\rho_{xy} = 0$

5. Nếu $\rho_{xy} = \pm 1$ thì X và Y phụ thuộc hàm số tuyến tính với nhau.

Hiệp phương sai và hệ số tương quan được dùng để đặc trưng cho mức độ chặt chẽ của mối liên hệ phụ thuộc giữa các biến ngẫu nhiên X và Y.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chú ý rằng nếu X và Y độc lập thì $\rho_{xy} = 0$ song điều ngược lại chưa chắc đã đúng, tức là nếu $\rho_{xy} = 0$ thì X và Y có thể độc lập hoặc phụ thuộc ở một dạng thức nào đó. Trong thực tế trường hợp được quan tâm hơn cả là sự phụ thuộc tương quan.

Hai biến ngẫu nhiên gọi là tương quan với nhau nếu hiệp phương sai (cũng tức là hệ số tương quan) khác không và hai biến nói trên gọi là không tương quan nếu hiệp phương sai (cũng tức là hệ số tương quan) bằng không.

Ta chú ý rằng nếu hai biến ngẫu nhiên tương quan với nhau thì cũng phụ thuộc nhau, song điều ngược lại chưa chắc đã đúng, tức là nếu các biến ngẫu nhiên phụ thuộc thì chúng có thể tương quan nhưng cũng có thể không tương quan với nhau.

Thí dụ 1. Với các số liệu của thí dụ 1 (bảng 4.2) hãy cho biết lương tháng của công nhân có tương quan với giới tính của công nhân hay không và mức độ tương quan chặt chẽ đến đâu.

Giải. Dễ thấy rằng

$$P(x_1, y_1) = 0,1 \neq 0,2.0,06 = P(x_1).P(y_1)$$

do đó X và Y phụ thuộc nhau.

Ta tìm hiệp phương sai giữa X và Y (bảng 4.3)

$$\sum_i \sum_j x_i y_j P(x_i, y_j) = 0,03 + 0,18 + 0,24 = 0,45$$

$$E(X) = 0.0,6 + 1.0,4 = 0,4$$

$$E(Y) = 0,5.0,16 + 1.0,48 + 1,5.0,36 = 1,1$$

Bảng 4.3

X \ Y	0,5	1	1,5	P(x)
0	0,1 0	0,3 0	0,2 0	0,6
1	0,06 0,03	0,18 0,18	0,16 0,24	0,4
P(y)	0,16	0,48	0,36	1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đó $Cov(X,Y) = 0,45 - 0,4.1,1 = 0,01$

$V(X) = 0^2.0,6 + 1^2.0,4 - (0,4)^2 = 0,24$

$\rightarrow \sigma_x = 0,4899$

$V(Y) = 0,5^2.0,16 + 1^2.0,48 + 1,5^2.0,36 - (1,1)^2 = 0,12$

$\rightarrow \sigma_y = 0,3464$

Vậy $\rho_{xy} = \dfrac{0,01}{0,4899.0,3464} = 0,0589$

Kết quả cho thấy Y và X có tương quan với nhau song mức độ phụ huộc tương quan không chặt chẽ.

Thí dụ 2. Biến ngẫu nhiên hai chiều (X, Y) có hàm mật độ xác suất đồng thời như sau:

$$f(x,y) = \begin{cases} \dfrac{1}{6\pi} \text{ nếu } \dfrac{x^2}{9} + \dfrac{y^4}{4} < 1 \\ 0 \quad \text{nếu } \dfrac{x^2}{9} + \dfrac{y^4}{4} > 1 \end{cases}$$

Hãy chứng tỏ rằng X và Y phụ thuộc và không tương quan với nhau.

Giải. Theo hàm mật độ xác suất đồng thời đã cho có thể tìm được các hàm mật độ xác suất biên của các thành phần như sau:

$$f_1(x) = \begin{cases} \dfrac{2}{9\pi}\sqrt{9-x^2} \text{ nếu } |x| < 3 \\ 0 \qquad\qquad \text{với } |x| > 3 \end{cases}$$

và

$$f_2(y) = \begin{cases} \dfrac{1}{2\pi}\sqrt{4-y^2} \text{ nếu } |y| < 2 \\ 0 \qquad\qquad \text{với } |y| > 2 \end{cases}$$

Vì $f(x,y) \neq f_1(x).f_2(y)$ nên X và Y phụ thuộc nhau. Còn để chứng minh rằng X và Y không tương quan với nhau, ta chỉ cần chứng minh rằng $Cov(X,Y) = 0$.

Thật vậy theo công thức ta có:

$$Cov(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dx.dy - E(X).E(Y)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì các hàm mật độ xác suất biên $f_1(x)$ và $f_2(x)$ đối xứng qua các trục toạ độ, do đó E(X) = 0 và E(Y) = 0, vì thế:

$$Cov(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dx.dy =$$

$$= \frac{1}{6\pi}\int_{-\frac{2}{3}\sqrt{9-x^2}}^{\frac{2}{3}\sqrt{9-x^2}} y\left(\int_{-\frac{3}{2}\sqrt{4-y^2}}^{\frac{3}{2}\sqrt{4-y^2}} xdx\right)dy$$

Tích phân trong ngoặc bằng không (hàm dưới dấu tích phân là hàm lẻ, cận lấy tích phân đối xứng qua gốc toạ độ), do đó Cov(X, Y) = 0 tức là X và Y không tương quan với nhau.

Với khái niệm hiệp phương sai ta có thể xét thêm tích chất của phương sai của tổng hai biến ngẫu nhiên phụ thuộc. Nếu X và Y là hai biến ngẫu nhiên phụ thuộc thì phương sai của tổng hoặc hiệu các biến ngẫu nhiên đó được xác định bằng biểu thức sau:

$$V(X \pm Y) = V(X) + V(Y) \pm 2Cov(X,Y) \qquad (4.28)$$

Trong trường hợp tổ hợp tuyến tính của các biến đó ta có công thức

$$V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2abCov(X,Y) \qquad (4.29)$$

Nếu X và Y là hai biến ngẫu nhiên độc lập thì phương sai của tích được xác định bằng biểu thức

$$V(X.Y) = [E(Y)]^2.V(X) + [E(X)]^2.V(Y) + V(X).V(Y) \qquad (4.30)$$

Thí dụ 3. Lãi suất hàng năm của trái phiếu T và cổ phiếu S của một công ty có bảng phân phối xác suất như sau (bảng 4.4):

Bảng 4.4

T \ S	-10%	0	10%	20%	P(T)
6%	0	0	0,1	0,1	0,2
8%	0	0,1	0,3	0,2	0,6
10%	0,1	0,1	0	0	0,2
P(S)	0,1	0,2	0,4	0,3	1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nếu muốn đầu tư tiền vào cả trái phiếu và cổ phiếu thì nên đầu tư theo tỉ lệ bao nhiêu để:

a - Lãi suất kỳ vọng thu được là lớn nhất

b - Độ rủi ro về lãi suất là nhỏ nhất.

Giải. Theo bảng phân phối xác suất đồng thời ta tìm được:

$$E(T) = 6.0,2 + 8.0,6 + 10.0,2 = 8\%$$

$$V(T) = 6^2.0,2 + 8^2.0,6 + 10^2 0,2 - 8^2 = 1,6$$

$$\sigma_T = 1,2649\%$$

$$E(S) = -10.0,1 + 0.0,2 + 10.0,4 + 20.0,3 = 9\%$$

$$V(S) = (-10)^2.0,1 + 0^2.0,2 + 10^2.0,4 + 20^2.0,3 - 9^2 = 89$$

$$\sigma_S = 9,4339\%$$

Từ đó Cov(S, T) = - 8

a. Gọi p là tỷ lệ đầu tư cho trái phiếu $(0 \le p \le 1)$ và gọi X là lãi suất thu được khi đầu tư cho cả trái phiếu và cổ phiếu thì

$$X = pT + (1-p)S$$

Từ đó

$$E(X) = p.E(T) + (1-p).E(S) = 8p + 9(1-p) = -p + 9$$

Hiển nhiên E(X) sẽ đạt cựu đại khi p = 0 tức là khi đầu tư toàn bộ tiền cho cổ phiếu.

b. Độ rủi ro được đặc trưng bởi phương sai hoặc độ lệch chuẩn của X.

Ta có $V(X) = p^2 V(T) + (1-p)^2 V(S) + 2p(1-p).Cov(S,T)$

$$= 106,6p^2 - 194p + 89$$

V(X) sẽ đạt cực tiểu khi p = 0,9099%.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

4.16. Biến ngẫu nhiên rời rạc hai chiều có bảng phân phối xác suất như sau:

Y \ X	1	3	4	8
3	0,15	0,06	0,25	0,04
6	0,30	0,10	0,03	0,07

a. Tìm kỳ vọng toán có điều kiện của Y với X = 1

b. Tìm các kỳ vọng toán E(X), E(Y) và các phương sai V(X), V(Y).

c. Tìm hiệp phương sai và hệ số tương quan giữa X và Y.

ĐS: a. E(X/Y = 1) = 5

b. E(X) = 2,93; E(Y) = 4,5

V(X) = 4,83; V(Y) = 2,25

c. Cov(X,Y) = - 0,555; ρ_{xy} = - 0,168.

4.17. Với số liệu của bài tập 4.2 hãy xác định.

a. Học vấn trung bình của dân nước đó.

b. Lứa tuổi trung bình của dân nước đó.

c. Học vấn và lứa tuổi có phụ thuộc nhau không và nếu có thì mức độ tương quan chặt chẽ đến đâu.

4.18. Với các số liệu của bài tập 4.3 hãy xác định

a. Thu nhập trung bình hàng năm của chồng và của vợ trong các cặp vợ chồng đang đi làm.

b. Thu nhập của chồng và của vợ có phụ thuộc nhau không và nếu có thì phụ thuộc chặt chẽ đến đâu.

4.19. Cho hàm mật độ xác suất của biến ngẫu nhiên hai chiều liên tục.

$$f(x,y) = \begin{cases} 4xye^{-x^2-y^2} & \text{với } x > 0;\ y > 0 \\ 0 & \text{với } x < 0;\ y < 0 \end{cases}$$

Tìm các kỳ vọng toán E(X), E(Y) và các phương sai V(X), V(Y).

ĐS: $E(X) = E(Y) = \sqrt{\frac{\pi}{2}}$; $V(X) = V(Y) = 1 - \frac{\pi}{4}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.20. Cho hàm mật độ xác suất của biến ngẫu nhiên hai chiều liên tục

$$f(x,y) = \frac{1}{4} \text{SinxSiny} \qquad \text{với } x, y \in (0, \pi)$$

a. Tìm E(X), E(Y), V(X), V(Y).

b. Tìm hiệp phương sai và hệ số tương quan.

ĐS: $E(X) = E(Y) = \pi/2$; $V(X) = V(Y) = \pi^2 - 4$

4.21*. Hàm mật độ xác suất f(x,y) của biến ngẫu nhiên liên tục (X,Y) là khối trụ tròn có tâm đáy trùng với gốc tọa độ và chiều cao bằng h. Tìm bán kính hình trụ r, tìm $f_1(x)$ $f_2(y)$, f(x/y), f(y/x), Cov(x,y).

Giải: Bán kính đáy hình trụ được xác định từ điều kiện là thể tích của hình trụ đó bằng 1, từ đó $r = \sqrt{1/\pi h}$

Hàm mật độ xác suất

$$f(x,y) = \begin{cases} h \text{ nếu } x^2 + y^2 < r^2 \\ 0 \text{ nếu } x^2 + y^2 > r^2 \end{cases}$$

Vì vậy

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \begin{cases} 2\sqrt{r^2 - x^2}\, h & \text{với } |x| < r \\ 0 & \text{với } |x| > r \end{cases}$$

Tương tự

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx = \begin{cases} 2\sqrt{r^2 - y^2}\, h & \text{với } |y| < r \\ 0 & \text{với } |y| > r \end{cases}$$

Với $|y| < r$

$$f(x/y)\ \frac{f(x,y)}{f_2(y)} = \begin{cases} \dfrac{1}{2\sqrt{r^2 - x^2}} & \text{với } |x| < \sqrt{r^2 - y^2} \\ 0 & \text{với } |x| > \sqrt{r^2 - y^2} \end{cases}$$

tương tự $|x| < r$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$f(y/x)\ \frac{f(x,y)}{f_1(x)} = \begin{cases} \dfrac{1}{2\sqrt{r^2-y^2}} & \text{với } |y| < \sqrt{r^2-x^2} \\ 0 & \text{với } |y| > \sqrt{r^2-x^2} \end{cases}$$

$E(X) = E(Y) = 0$

Vì hàm f(x,y) là hàm chẵn theo x và y nên.

$$V(X) = 2h \int_{-r}^{r} x^2 \sqrt{r^2-x^2}\, dx = \frac{\pi}{4} hr^4 = \frac{r^2}{4}$$

$$V(Y) = \frac{r^2}{4} \Rightarrow Cov(x,y) = 0$$

8. KỲ VỌNG TOÁN CÓ ĐIỀU KIỆN - HÀM HỒI QUY

Giá trị của hệ số tương quan chỉ cho ta biết mức độ chặt chẽ của sự phụ thuộc tương quan tuyến tính giữa X và Y. Để biểu diễn sự phụ thuộc tương quan này người ta sử dụng các hàm hồi quy.

Trước hết ta xét khái niệm kỳ vọng toán có điều kiện. *Kỳ vọng toán có điều kiện* của biến ngẫu nhiên rời rạc Y với X = x (x là một giá trị xác định của X) là tổng các tích giữa các giá trị có thể có của Y với các xác suất có điều kiện tương ứng:

$$E(Y/X=x) = \sum_{j=1}^{m} y_j P(y_j/x) \qquad (4.31)$$

Đối với các biến ngẫu nhiên liên tục, kỳ vọng toán có điều kiện được xác định bằng công thức.

$$E(Y/X=x) = \int_{-\infty}^{+\infty} y f(y/x)dy \qquad (4.32)$$

Trong đó f(y/x) là hàm mật độ xác suất có điều kiện của Y với X = x. Tương tự, ta có định nghĩa kỳ vọng toán có điều kiện của X khi Y = y

$$E(X/Y=y) = \sum_{i=1}^{n} x_i P(x_i/y) \qquad (4.33)$$

đối với biến ngẫu nhiên rời rạc và

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$E(X/Y=y)=\int_{-\infty}^{+\infty} xf(x/y)dx$$

đối với biến ngẫu nhiên liên tục. Hàm hồi quy của Y đối với X là kỳ vọng toán có điều kiện của Y đối với X:

$$f(x)=E(Y/x) \tag{4.34}$$

Tương tự, hàm hồi quy của X đối với Y là kỳ vọng toán có điều kiện của X đối với Y:

$$f(y)=E(X/y) \tag{4.35}$$

Các hàm hồi quy cho biết giá trị trung bình của biến ngẫu nhiên này phụ thuộc vào biến kia như thế nào.

Thí dụ. Thống kê dân số của một nước ở độ tuổi trưởng thành theo trình độ học vấn X và lứa tuổi Y thu được kết quả sau (bảng 4.5)

Tìm học vấn trung bình theo lứa tuổi.

Bảng 4.5

Y / X	25-35 30	35-55 45	55-100 70
Thất học: 0	0,01	0,02	0,05
Tiểu học: 1	0,03	0,06	0,10
Trung học: 2	0,18	0,21	0,15
Đại học: 3	0,07	0,08	0,04

Giải. Học vấn trung bình theo lứa tuổi là kỳ vọng toán có điều kiện của X theo Y. Với Y = 30 ta có bảng phân phối xác suất có điều kiện sau:

X/Y = 30	0	1	2	3
P	$\frac{0,01}{0,29}$	$\frac{0,03}{0,29}$	$\frac{0,18}{0,29}$	$\frac{0,07}{0,29}$

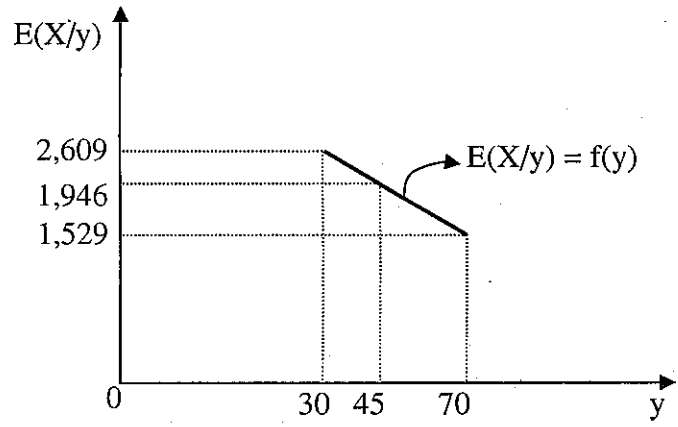
Từ đó $E(X/Y=30)=2,069$

Tương tự $E(X/Y=45)=1,946$

$E(X/Y=70)=1,529$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hàm hồi quy E(X/y) = f(y) có thể mô tả trên đồ thị sau (hình 4.5).



Hình 4.5

Bài tập

4.22. Biến ngẫu nhiên rời rạc hai chiều có bảng phân phối xác suất như sau:

Y \ X	1	3	4	8
3	0,15	0,06	0,25	0,04
6	0,30	0,10	0,03	0,07

a. Tìm kỳ vọng toán có điều kiện của Y với X = 1

b. Tìm các kỳ vọng toán E(X), E(Y) và các phương sai V(X), V(Y).

c. Tìm hiệp phương sai và hệ số tương quan giữa X và Y.

ĐS: a. E(X/Y = 1) = 5

b. E(X) = 2,93; E(Y) = 4,5

V(X) = 4,83; V(Y) = 2,25

c. Cov(X,Y) = - 0,555; ρ_{xy} = - 0,168.

4.23. Với các số liệu của bài tập 4.3 hãy xác định thu nhập trung bình của các bà vợ có chồng ở mức thu nhập 20 triệu/năm.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9. PHÂN PHỐI CHUẨN HAI CHIỀU

Đối với biến *ngẫu nhiên hai chiều* thì trường hợp phổ biến trong thực tế là phân phối theo quy luật chuẩn.

Biến ngẫu nhiên (X, Y) gọi là *phân phối theo quy luật chuẩn hai chiều* nếu hàm mật độ xác suất đồng thời của nó có dạng

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}.\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}-\right.\right.$$
$$\left.\left.-\frac{2\rho(x-\mu_x)((y-\mu_y)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\} \qquad (4.36)$$

Ta thấy rằng quy luật chuẩn hai chiều có 5 tham số đặc trưng là $\mu_x,\mu_y,\sigma_x,\sigma_y$ và ρ. Có thể chứng minh được rằng μ_x và μ_y chính là kỳ vọng toán của các thành phần X và Y. Còn σ_x và σ_y tương ứng là các độ lệch chuẩn, ρ là hệ số tương quan giữa X và Y.

Đặc điểm của biến ngẫu nhiên phân phối theo quy luật chuẩn hai chiều là nếu các thành phần của nó không tương quan thì chúng cũng độc lập với nhau. Thật vậy, giả sử X và Y không tương quan với nhau, lúc đó $\rho=0$ và biểu thức của hàm mật độ xác suất có dạng:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}.e^{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_y^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right]}$$
$$=\frac{1}{\sigma_x\sqrt{2\pi}}e^{\frac{(x-\mu_x)^2}{2\sigma_y^2}}.\frac{1}{\sigma_y\sqrt{2\pi}}e^{\frac{(y-\mu_y)^2}{2\sigma_y^2}}=f_1(x).f_2(y) \qquad (4.37)$$

Như vậy, nếu các thành phần của biến ngẫu nhiên hai chiều phân phối chuẩn mà không tương quan với nhau thì hàm mật độ xác suất đồng thời của hệ hai biến ngẫu nhiên ấy bằng tích các hàm mật độ xác suất biên, từ đó suy ra tính độc lập của các thành phần đó. Điều ngược lại cũng đúng như vậy.

Như vậy là trong quy luật chuẩn hai chiều khái niệm độc lập và không tương quan tương đương nhau.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

10. QUY LUẬT PHÂN PHỐI XÁC SUẤT CỦA HÀM CÁC BIẾN NGẪU NHIÊN

Trong thực tế người ta thường gặp trường hợp một biến ngẫu nhiên là hàm số của một hoặc nhiều biến ngẫu nhiên khác. Lúc đó, khi biết quy luật phân phối xác suất của các đối số, ta có thể tìm được quy luật phân phối xác suất của hàm số tương ứng.

10.1. Quy luật phân phối xác suất của hàm một biến ngẫu nhiên

Nếu mỗi giá trị có thể có của biến ngẫu nhiên X tương ứng với một giá trị có thể có của biến ngẫu nhiên Y thì Y được gọi là hàm của biến ngẫu nhiên X:

$$Y = \varphi(X) \tag{4.38}$$

Giả sử biến ngẫu nhiên là rời rạc thì ứng với các giá trị khác nhau của X ta có các giá trị khác nhau của Y và xác suất tương ứng với các giá trị đó bằng nhau.

Thí dụ 1. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	2	3
P	0,6	0,4

Tìm quy luật phân phối xác suất của $Y = X^2$.

Giải. Ta tìm các giá trị có thể có của Y:

$$y_1 = 2^2 = 4; \qquad y_2 = 3^2 = 9$$

Vậy bảng phân phối xác suất của Y có dạng:

Y	4	9
P	0,6	0,4

Nếu trong số các giá trị có thể có của Y có các giá trị giống nhau thì phải cộng các xác suất tương ứng lại.

Thí dụ 2. Biến ngẫu nhiên X có bảng phân phối xác suất như sau:

Y	- 2	2	3
P	0,4	0,5	0,1

Tìm quy luật phân phối xác suất của $Y = X^2$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Xác suất tương ứng với giá trị $y_1 = 4$ bằng tổng xác suất của các biến cố xung khắc x = - 2 và x = 2 tức là bằng $0,4 + 0,5 = 0,9$; xác suất để $y_2 = 9$ bằng 0,1 do đó Y có bảng phân phối xác suất như sau:

Y	4	9
P	0,9	0,1

Giả sử biến ngẫu nhiên X là liên tục với hàm mật độ xác suất f(x) đã biết và giả sử $Y = \varphi(X)$. Có thể chứng minh được rằng nếu $Y = \varphi(X)$ là khả vi, đơn điệu tăng hoặc đơn điệu giảm, có hàm ngược là $X = \psi(Y)$ thì hàm mật độ xác suất g(y) của biến ngẫu nhiên Y được xác định bằng biểu thức:

$$g(y) = f[\psi(y)]\left|\psi'(y)\right| \tag{4.39}$$

Thí dụ 3. Biến ngẫu nhiên X phân phối chuẩn với kỳ vọng toán $\mu = 0$. Tìm quy luật phân phối xác suất của biến ngẫu nhiên $Y = X^3$.

Giải. Vì hàm $y = X^3$ là khả vi và đơn điệu tăng, do đó có thể áp dụng công thức:

$$g(y) = f[\psi(y)]\left|\psi'(y)\right|$$

Ta tìm hàm ngược của hàm $y = x^3$

$$\psi(y) = x = y^{1/3}$$

Ta tìm $f[\psi(y)]$. Theo điều kiện

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

Do đó:

$$f[\psi(y)] = f\left[y^{1/3}\right] = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-y^{2/3}}{2\sigma^2}}$$

Ta tìm đạo hàm của hàm ngược $\psi(y)$

$$\left|\psi'(y)\right| = \left|(y^{1/3})\right| = \left|\frac{1}{3y^{2/3}}\right|$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đó ta có hàm mật độ xác suất của biến ngẫu nhiên Y như sau:

$$g(y)=\frac{1}{\left|3\sigma y^{2/3}\right|\sqrt{2\pi}}e^{\frac{-y^{2/3}}{2\sigma^2}}$$

Thí dụ 4. Cho biến ngẫu nhiên X phân phối chuẩn với kỳ vọng μ. Tìm quy luật phân phối xác suất của biến ngẫu nhiên: $Y=aX+b$ trong đó a và b là các hệ số.

Giải. Vì hàm số $Y=aX+b$ là khả vi và đơn điệu, do đó có thể áp dụng công thức:

$$g(y)=f[\psi(y)]\left|\psi'(y)\right|$$

Hàm ngược $\psi(y)$ có dạng:

$$\psi(y)=x=\frac{y-b}{a}$$

Vậy $$\left|\psi'(y)\right|=\left|\frac{1}{a}\right|=\frac{1}{|a|}$$

$$f[\psi(y)]=f\left[\frac{y-b}{a}\right];g(y)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-\left(\frac{y-b}{a}-\mu\right)^2}{2(\sigma)^2}\right]\cdot\frac{1}{|a|}$$

$$=\frac{1}{\sigma|a|\sqrt{2\pi}}\exp\left\{\frac{-[y-(a\mu+b)]^2}{2(a\sigma)^2}\right\}$$

Như vậy hàm mật độ xác suất g(y) có dạng:

$$g(y)=\frac{1}{\sigma|a|\sqrt{2\pi}}\exp\left\{\frac{-[y-(a\mu+b)]^2}{2(a\sigma)^2}\right\}$$

Đó lại chính là hàm mật độ xác suất của quy luật chuẩn với các tham số đặc trưng là $E(Y)=a\mu+b$ và $V(Y)=(|a|\sigma)^2$.

Như vậy, ta đã chứng tỏ được rằng nếu biến ngẫu nhiên X phân phối chuẩn thì một hàm tuyến tính bất kỳ của nó cũng phân phối theo quy luật chuẩn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

10.2. Quy luật phân phối xác suất của hàm hai biến ngẫu nhiên

Nếu ứng với mỗi cặp giá trị có thể có của biến ngẫu nhiên X và Y có một giá trị có thể có của Z thì Z được gọi là hàm của hai biến ngẫu nhiên $Z = \varphi(X,Y)$.

Để tìm quy luật phân phối xác suất của Z ta xét một trường hợp cụ thể là $Z = X + Y$ khi đã biết các quy luật phân phối xác suất của X và Y.

Giả sử X và Y là các biến ngẫu nhiên rời rạc và độc lập với nhau. Để xây dựng quy luật phân phối xác suất của $Z = X + Y$ ta phải tìm tất cả các giá trị có thể có của Z và các xác suất tương ứng.

Thí dụ 5. Các biến ngẫu nhiên X và Y rời rạc độc lập với nhau và có các bảng phân phối xác suất như sau:

X	1	2
P	0,4	0,6

X	3	4
Y	0,2	0,8

Tìm quy luật phân phối xác suất của $Z = X + Y$.

Giải. Các giá trị có thể có của Z là tổng của mỗi giá trị có thể có của X với mỗi giá trị có thể có của Y:

$Z_1 = 1 + 3 = 4$; $Z_2 = 1 + 4 = 5$;

$Z_3 = 2 + 3 = 5$; $Z_4 = 2 + 4 = 6$;

Ta đi tìm các xác suất tương ứng. Xác suất để $Z_1 = 4$ là xác suất để X nhận giá trị bằng 1 và Y nhận giá trị bằng 3. Vì X và Y độc lập, do đó theo định lý nhân xác suất, ta có:

$$P(Z_1 = 4) = P(x = 1).P(y = 3) = 0,4.0,2 = 0,08$$

Tương tự, ta được:

$$P(Z_2 = 5) = P(x = 1).P(y = 4) = 0,4.0,8 = 0,32$$

$$P(Z_3 = 5) = P(x = 2).P(y = 3) = 0,6.0,2 = 0,12$$

$$P(Z_4 = 6) = P(x = 2).P(y = 4) = 0,6.0,8 = 0,48$$

Vì hai giá trị Z_2 và Z_3 đều bằng 5 do đó xác suất tương ứng của chúng phải được cộng lại:

$$P(Z = 5) = 0,32 + 0,12 = 0,44$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vậy Z có bảng phân phối xác suất như sau:

Z	4	5	6
P	0,08	0,44	0,48

Giả sử X và Y là các biến ngẫu nhiên liên tục. Có thể chứng minh được rằng khi X và Y độc lập thì hàm mật độ xác suất g(z) của tổng $Z = X + Y$ (với điều kiện là khi mật độ xác suất của ít nhất một trong hai đối số xác định trong khoảng $(-\infty; +\infty)$ bằng một biểu thức) có thể tìm được theo công thức:

$$g(z) = \int_{-\infty}^{z} f_1(x) f_2(z-x)dx \qquad (4.40)$$

Hoặc
$$g(z) = \int_{-\infty}^{z} f_1(z-y) f_2(y)dy \qquad (4.41)$$

Trong đó f_1 và f_2 tương ứng là các hàm mật độ xác suất biên của X và Y.

Thí dụ 6. Các biến ngẫu nhiên X và Y độc lập, có các hàm mật độ xác suất như sau:

$$f_1(x) = \frac{1}{3}e^{-x/3} \text{ với } x \geq 0$$

$$f_2(y) = \frac{1}{4}e^{-y/4} \text{ với } y \geq 0$$

Tìm quy luật phân phối xác suất của $Z = X + Y$.

Giải. Vì các giá trị có thể có của X và Y không âm, do đó hàm mật độ xác suất g(z) được xác định bằng biểu thức:

$$g(z) = \int_{-\infty}^{z} f_1(x) f_2(z-x)dx = \int_{0}^{z} \left[\frac{1}{3}e^{-\frac{x}{3}}\right]\left[\frac{1}{4}e^{-\frac{z-x}{4}}\right]dx$$

$$= \frac{1}{12}e^{-\frac{z}{4}} \int_{0}^{z} e^{\frac{x}{12}}dx = e^{-\frac{z}{4}}\left(1 - e^{-\frac{z}{12}}\right)$$

Ta chú ý rằng ở đây $Z \geq 0$ vì $Z = X + Y$ mà theo đầu bài thì X và Y đều không âm.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

10.3. Các tham số đặc trưng của hàm các biến ngẫu nhiên

Ở trên ta đã xét các phương pháp để tìm quy luật phân phối xác suất của hàm các biến ngẫu nhiên. Khi ta đã biết được quy luật phân phối xác suất dưới dạng bảng phân phối xác suất hoặc hàm mật độ xác suất thì việc tìm các tham số đặc trưng không còn trở ngại gì. Song như đã thấy việc tìm quy luật phân phối xác suất của các biến ngẫu nhiên nhiều khi khá phức tạp. Mặt khác, trong thực tế có nhiều trường hợp ta chỉ quan tâm đến các tham số đặc trưng của hàm các biến ngẫu nhiên chứ không quan tâm đến bản thân quy luật phân phối xác suất của nó. Trong mục này ta xét cách xác định các tham số đặc trưng của hàm các biến ngẫu nhiên mà không cần xác định trước quy luật phân phối xác suất của nó. Giả sử có biến ngẫu nhiên X rời rạc với bảng phân phối xác suất đã biết:

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

Ta phải tìm kỳ vọng toán và phương sai của biến ngẫu nhiên $Y = \varphi(X)$. Các tham số đặc trưng này được xác định bằng các công thức sau đây:

$$E(Y) = E[\varphi(X)] = \sum_{i=1}^{n} \varphi(x_i)p_i \tag{4.42}$$

$$V(Y) = V[\varphi(X)] = \sum_{i=1}^{n} \{\varphi(x_i) - E[\varphi(X)]\}^2 p_i$$

$$= \sum_{i=1}^{n} \varphi^2(x_i)p_i - \{E[\varphi(X)]\}^2 \tag{4.43}$$

Thí dụ 7. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	1	3	5
P	0,2	0,5	0,3

Tìm kỳ vọng toán và phương sai của biến ngẫu nhiên

$Y = \varphi(X) = X^2 + 1$

Giải. Các giá trị có thể có của Y là:

$\varphi(1) = 1^2 + 1 = 2; \varphi(3) = 3^2 + 1 = 10;$

$\varphi(5) = 5^2 + 1 = 26.$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Do đó:

$$E(Y) = E(X^2 + 1) = 2.0,2 + 10.0,5 + 26.0,3 = 13,2$$

$$V(Y) = 2^2.0,2 + 10^2.0,5 + 26^2.0,3 - [13,2]^2 = 79,36$$

Nếu X là biến ngẫu nhiên liên tục với hàm mật độ xác suất là f(x) thì kỳ vọng toán và phương sai của biến ngẫu nhiên $Y = \varphi(X)$ được xác định bằng các công thức:

$$E(Y) = E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x).f(x)dx \tag{4.44}$$

$$V(Y) = V[\varphi(X)] = \int_{-\infty}^{+\infty} \{\varphi(x) - E[\varphi(X)]\}^2 f(x)dx$$

$$= \int_{-\infty}^{+\infty} \varphi^2(x)dx - \{E[\varphi(X)]\}^2 \tag{4.45}$$

Thí dụ 8. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \sin x & x \in \left(0, \dfrac{\pi}{2}\right) \\ 0 & x \notin \left(0, \dfrac{\pi}{2}\right) \end{cases}$$

Tìm kỳ vọng toán và phương sai của hàm $Y = X^2$.

Giải. Vì X là biến ngẫu nhiên liên tục có hàm mật độ xác suất khác không trong khoảng $\left(0, \dfrac{\pi}{2}\right)$ do đó:

$$E(Y) = E(X^2) = \int_0^{\frac{\pi}{2}} x^2 \sin x dx$$

Lấy tích phân từng phần ta được

$$E(Y) = \pi - 2$$

$$V(Y) = V(X^2) = \int_0^{\frac{\pi}{2}} x^4 \sin x dx - (\pi - 2)^2 = \frac{\pi^3}{2} - \pi^2 - 8\pi + 20$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tương tự, có thể xây dựng các công thức tìm các tham số đặc trưng của hàm hai biến ngẫu nhiên Z = φ(X,Y). Chẳng hạn kỳ vọng toán E(Z) được xác định bằng các công thức sau:

Nếu X và Y là các biến ngẫu nhiên rời rạc thì:

$$E(Z) = E[\varphi(X.Y)] = \sum_{i=1}^{n}\sum_{j=1}^{m}\varphi(x_i, y_j)P_{ij} \qquad (4.46)$$

Còn nếu X và Y là các biến ngẫu nhiên liên tục thì:

$$E(Z) = E[\varphi(X.Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\varphi(x.y)f(x,y)dxdy \qquad (4.47)$$

Bài tập

4.24. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

a.

X	1	2	3
P	0,2	0,1	0,7

b.

X	-1	1	2
P	0,1	0,2	0,7

Tìm quy luật phân phối xác suất của biến ngẫu nhiên $Y = X^4$

ĐS:

a.

Y	1	16	81
P	0,2	0,1	0,7

b.

Y	1	16
P	0,3	0,7

4.25. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất f(x). Tìm hàm mật độ xác suất g(y) của biến ngẫu nhiên Y nếu

a. $Y = X + 1$ $(-\infty < x < +\infty)$

b. $Y = 2X$ $(-a < x < +a)$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ĐS: a. $g(y) = f(y-1)$ $(-\infty < y < +\infty)$

b. $g(y) = (1/2)f(y/2)$ $(-2a < y < 2a)$

4.26. Các biến ngẫu nhiên X và Y là rời rạc và có các bảng phân phối xác suất như sau:

X	2	3	5
P	0,3	0,5	0,2

Y	1	4
P	0,2	0,8

Tìm các quy luật phân phối xác suất của các hàm

a. $Z = X + Y$

b. $Z = X.\ Y$

ĐS: a.

Z	3	4	6	7	9
P	0,06	0,10	0,28	0,40	0,16

b.

Z	2	3	5	8	12	20
P	0,06	0,10	0,04	0,24	0,40	,016

4.27. Các biến ngẫu nhiên X và Y độc lập và có các hàm mật độ xác suất như sau:

$$f_1(x) = \frac{1}{3} e^{-x3} \qquad (0 \le x \le +\infty)$$

$$f_2(y) = \frac{1}{5} e^{-y5} \qquad (0 \le y \le +\infty)$$

Tìm hàm mật độ xác suất của biến ngẫu nhiên $Z = X + Y$

ĐS:

$$g(Z) = \begin{cases} 2x \text{ với } x \in (0, 1) \\ 0 \text{ với } x \notin (0, 1) \end{cases}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.28. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất

X	-1	0	1	2
P	0,2	0,1	0,3	0,4

Tìm kỳ vọng toán và phương sai của biến ngẫu nhiên $Y = 2^X$.

ĐS: E(Y) = 2,4; V(Y) = 1,99

4.29. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất

$$f(x) = \begin{cases} 2x & \text{với } x \in (0, 1) \\ 0 & \text{với } x \notin (0, 1) \end{cases}$$

Tìm kỳ vọng toán và phương sai của biến ngẫu nhiên $Y = X^2$

ĐS: E(Y) = 0,5; V(Y) = 0,083

4.30. Biến ngẫu nhiên liên tục X phân phối theo quy luật luỹ thừa.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{với } x > 0 \\ 0 & \text{với } x < 0 \end{cases}$$

Hãy xác định điều kiện tồn tại và tìm kỳ vọng toán và phương sai của biến ngẫu nhiên $Y = e^x$

ĐS: $E(Y) = \dfrac{\lambda}{\lambda - 1}$ với $\lambda > 0$

$$V(Y) = \frac{\lambda}{(\lambda - 1)(\lambda - 2) - 2} \quad \text{với } \lambda > 2$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

11. TÓM TẮT CHƯƠNG 4

1. Trong kinh tế, mục đích chính của việc nghiên cứu biến ngẫu nhiên nhiều chiều là đánh giá ảnh hưởng qua lại của các hiện tượng kinh tế.
2. Để mô tả quy luật phân phối xác suất của biến ngẫu nhiên hai chiều rời rạc chủ yếu là dùng bảng phân phối xác suất đồng thời, bản phân phối xác suất biên và bảng phân phối xác suất có điều kiện của mỗi thành phần, còn để mô tả quy luật phân phối xác suất của biến ngẫu nhiên hai chiều liên tục thì chủ yếu dùng hàm mật độ xác suất đồng thời, hàm mật độ xác suất biên và hàm mật độ xác suất có điều kiện của mỗi thành phần.
3. Trong kinh tế và xã hội thường chỉ gặp sự phụ thuộc lỏng lẻo hơn so với sự phụ thuộc hàm số, đó là sự phụ thuộc tương quan.
4. Các tham số chủ yếu dùng để đo mức độ chặt chẽ của sự phụ thuộc tương quan giữa hai thành phần của biến ngẫu nhiên hai chiều là hiệp phương sai và hệ số tương quan.
5. Hình thức cơ bản để biểu diễn sự phụ thuộc tương quan giữa các thành phần của biến ngẫu nhiên hai chiều là Kỳ vọng toán có điều kiện và Hàm hồi quy.
6. Trong kinh tế các biến số kinh tế thường là hàm số của các biến số kinh tế khác, vì thế về bản chất chúng là hàm của các biến ngẫu nhiên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các ký hiệu và công thức cơ bản

* Xác suất đồng thời của biến ngẫu nhiên (X, Y) rời rạc

$$p(x_i, y_j) = P[(X = x_i)(Y = y_i)] \quad i = \overline{1,n};\ j = \overline{1,m}$$

* Xác suất biên của thành phần X

$$P(x_i) = \sum_{j=1}^{m} P(x_i, y_i) \qquad i = \overline{1,n}$$

* Xác suất biên của thành phần Y

$$P(y_j) = \sum_{i=1}^{n} P(x_i, y_i) \qquad j = \overline{1,m}$$

* Hàm phân bố xác suất đồng thời (X,Y)

$$F(x, y) = P(X < x, Y < y)$$

* Hàm phân bố xác suất biên của X

$$F_1(x) = F(x, +\infty)$$

* Hàm phân bố xác suất biên của Y

$$F_2(y) = F(+\infty, y)$$

* $$P(x_1 < X < x_2, y_1 < Y < y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$$

* Hàm mật độ xác suất đồng thời của biến ngẫu nhiên (X, Y) liên tục

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x . \partial y} \Rightarrow \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)dxdy = 1$$

* $$P[(X, Y) \in D] \iint_D f(x, y)dxdy$$

* Hàm mật độ xác suất biên của X

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y)dy$$

* Hàm mật độ xác suất biên của Y

$$f_2(y) = \int_{-\infty}^{+\infty} f(x, y)dx$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

* Xác suất có điều kiện của (X, Y) rời rạc

$$P(x_i / y_j) = \frac{P(x_i, y_j)}{P(y_j)} \qquad i = \overline{1,n};\ j = \overline{1,m}$$

$$P(y_j / x_i) = \frac{P(x_i, y_j)}{P(x_i)} \qquad i = \overline{1,n};\ j = \overline{1,m}$$

* Hàm mật độ xác suất có điều kiện của (X, Y) liên tục

$$f(x/y) = \frac{f(x,y)}{f_2(y)} \quad \text{và} \quad f(y/x) = \frac{f(x,y)}{f_1(x)}$$

* Các tham số đặc trưng của biến ngẫu nhiên hai chiều

\+ Kỳ vọng toán

$$E(X) = \sum_{i=1}^{n} x_i P(x_i)$$

$$E(Y) = \sum_{j=1}^{m} y_i P(y_j) \quad \text{nếu (X, Y) rời rạc}$$

$$E(X) = \int_{-\infty}^{+\infty} x f_1(x) dx$$

$$E(Y) = \int_{-\infty}^{+\infty} x f_2(y) dy \quad \text{nếu (X, Y) liên tục}$$

\+ Phương sai

$$V(X) = \sum_{i=1}^{n} x_i^2 P(x_i) - [E(X)]^2$$

$$V(Y) = \sum_{j=1}^{m} y_j^2 P(y_i) - [E(Y)]^2 \quad \text{nếu (X, Y) rời rạc}$$

$$V(X) = \int_{-\infty}^{+\infty} x^2 f(x) - [E(X)]^2$$

$$V(Y) = \int_{-\infty}^{+\infty} y^2 f_2(y) - [E(Y)]^2 \quad \text{nếu (X, Y) liên tục}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Hiệp phương sai

$$Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$
$$= E(X.Y) - E(X).E(Y)$$

Trong đó $E(X.Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(x_i y_j)$

nếu (X, Y) rời rạc và $E(X.Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x.y.f(x,y)dxdy$

nếu (X, Y) liên tục.

+ Hệ số tương quan

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X.\sigma_Y}$$

* Kỳ vọng toán có điều kiện

$$E(Y/x_i) = \sum_{j=1}^{m} y_j P(y_j/x_i)$$

$$E(X/y_j) = \sum_{i=1}^{n} x_i P(x_i/y_j)$$ nếu (X, Y) rời rạc

$$E(Y/x) = \int_{-\infty}^{+\infty} yf(y/x)dy$$

$$E(X/y) = \int_{-\infty}^{+\infty} xf(x/y)dx$$ nếu (X, Y) liên tục

* Hàm hồi quy của Y đối với X

$$E(Y/x) = f(x)$$

* Hàm hồi quy của X đối với Y

$$E(X/y) = g(y)$$

* Phân phối chuẩn hai chiều

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - \right.\right.$$
$$\left.\left. - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X.\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

* Hàm các biến ngẫu nhiên

+ Nếu X là biến ngẫu nhiên rời rạc và $Y = \varphi(X)$ thì bảng phân bố xác suất của Y là:

Y	$\varphi(x_1)$	$\varphi(x_2)$	$\varphi(x_n)$
P	p_1	p_2	p_n

+ Nếu X là biến ngẫu nhiên liên tục, $Y = \varphi(X)$ và hàm ngược là $X = \psi(Y)$ thì hàm mật độ xác suất

$$g(y) = f[\psi(y)].|\psi'(y)|$$

+ Các tham số đặc trưng

$$E(Y) = E[\varphi(X)] = \sum_{i=1}^{n} \varphi(x_i)P_i$$

$$V(Y) = V[\varphi(X)] = \sum_{i=1}^{n} \varphi^2(x_i)P_i - \left\{E\left[\varphi(X)\right]\right\}^2$$

* Các tính chất của phương sai

+ $V(aX \pm bY) = a^2.V(X) + b^2.V(Y) \pm 2abCov(X,Y)$

+ Nếu X và Y độc lập thì

$$V(X.Y) = [E(Y)]^2.V(X) + [E(X)]^2.V(Y) + V(X).V(Y)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Biến ngẫu nhiên hai chiều	Two random variable
Bảng phân phối xác suất đồng thời	Joint distribution table
Bảng phân phối xác suất biên	Marginal distribution table
Bảng phân phối xác suất có điều kiện	Conditional distribution table
Hàm phân bố xác suất đồng thời	Joint cumulative distribution function
Hàm phân bố xác suất biên	Marginal cumulative function
Hàm phân bố xác suất có điều kiện	Conditional cumulative function
Hàm mật độ xác suất đồng thời	Joint density probability function
Hàm mật độ xác suất biên	Marginal density function
Hàm mật độ xác suất có điều kiện	Conditional density function
Hiệp phương sai	Covariance
Hệ số tương quan	Correlation coefficient
Kỳ vọng toán có điều kiện	Conditional expected value
Hàm hồi quy	Regession function
Hàm các biến ngẫu nhiên	Fuction of random variables

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

4.31. Cho một ví dụ thực tế trong kinh tế hoặc kinh doanh về:

a. Một biến ngẫu nhiên nhiều chiều rời rạc.

b. Một biến ngẫu nhiên nhiều chiều liên tục.

4.32. Cho bảng phân phối xác suất đồng thời của biến ngẫu nhiên rời rạc (X, Y) như sau:

Y \ X	-1	0	1	Σ
0		0,25	0,15	
1	0,15	0,2		0,45
Σ				

a. Hãy điền vào các giá trị còn thiếu trong bảng.

b. Tìm các phân phối biên của X và Y

c. Tìm phân phối có điều kiện của X khi Y = 1

4.33. Cho biến ngẫu nhiên hai chiều rời rạc (X, Y). Nếu X và Y độc lập thì các dòng của bảng phân phối xác suất đồng thời có tỉ lệ với nhau không? Câu hỏi tương tự đối với các cột của bảng.

4.34. Cho hai biến ngẫu nhiên rời rạc X và Y độc lập nhau và có các bảng phân phối xác suất như sau:

X	1	2	3
P_X	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{5}{12}$

X	-2	-1
P_Y	$\frac{1}{3}$	$\frac{2}{3}$

Hãy tìm bảng phân phối xác suất đồng thời của (X, Y).

4.35. Hãy cho biết các mệnh đề sau đây là đúng hay sai?

Giải thích:

a. Nếu $P(x_i, y_j) = P(x_i).P(y_j) \forall i, j$ thì X và Y độc lập.

b. Nếu X và Y độc lập thì $\rho_{XY} = 0$

c. Nếu $\rho_{XY} = 0$ thì X và Y độc lập.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.36. Có hai hộp. Hộp thứ nhất có 3 bi đỏ và 2 bi xanh. Hộp thứ hai có 2 bi đỏ và 3 bi xanh. Lấy ngẫu nhiên 1 bi từ hộp thứ nhất bỏ sang hộp thứ hai, sau đó từ hộp thứ hai lấy ra 1 bi. Gọi X và Y tương ứng là số bi đỏ được lấy ra từ hộp thứ nhất và thứ hai. Lập bảng phân phối xác suất đồng thời của (X, Y).

4.37. Thống kê về giá thành sản phẩm (Y)và sản lượng (X) của một ngành sản xuất thu được bảng phân phối xác suất như sau:

Y \ X	30	50	80	100
6	0,05	0,06	0,08	0,11
7	0,06	0,15	0,04	0,08
8	0,07	0,09	0,1	0,11

a. Tìm giá thành sản phẩm trung bình và mức độ phân tán của nó.

b. Tìm sản lượng trung bình khi giá thành bằng 8.

c. X và Y có độc lập không?

d. X và Y có tương quan với nhau không?

4.38. Biến ngẫu nhiên rời rạc (X, Y) có bảng phân phối xác suất đồng thời như sau:

Y \ X	1	3	4	8
3	0,15	0,06	0,25	0,04
6	0,3	0,1	0,03	0,07

Tìm hàm phân bố xác suất đồng thời F(x, y).

4.39. Tìm hàm mật độ xác suất đồng thời của biến ngẫu nhiên hai chiều liên tục (X, Y) nếu hàm phân bố xác suất đồng thời có dạng:

$$F(x,y) = (1-e^{-ax})(1-e^{-by}) \qquad x \geq 0, y \geq 0$$

4.40. Biến ngẫu nhiên hai chiều liên tục (X, Y) có hàm mật độ xác suất đồng thời như sau:

$$f(x,y) = \frac{A}{\pi^2(16+x^2)(25+y^2)}$$

a. Tìm giá trị A.

b. Tìm hàm phân bố xác suất đồng thời F(x, y).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.41. Hàm mật độ xác suất đồng thời của biến ngẫu nhiên liên tục (X, Y) có dạng:

$$f(x,y) = Ae^{-ax^2+bxy-cy^2} \qquad a>0, b>0, c>0$$

a. Tìm các hàm mật độ xác suất biên $f_1(x), f_2(y)$

b. Với điều kiện nào thì X và Y độc lập?

4.42. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	20	30	40	50
P_X	0,1	0,3	0,4	0,2

và biến ngẫu nhiên Y = 0,2X - 2.

a. Lập bảng phân phối xác suất của Y.

b. Tìm E(Y) và V(Y) theo hai cách:

+ Dùng bảng phân phối xác suất của Y

+ Dùng tính chất của kỳ vọng toán và phương sai.

4.43. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	0	1	2	3
P_X	0,14	0,39	0,36	0,11

và biến ngẫu nhiên $Y = -100X^2 + 300X + 500$.

Tìm E(Y) và V(Y) bằng hai cách:

+ Dùng bảng phân phối xác suất của Y

+ Dùng tính chất của kỳ vọng toán và phương sai.

4.44. Cho X là biến ngẫu nhiên và $Y = \varphi(X)$. Để tìm kỳ vọng toán E(Y) và phương sai V(Y) thì dùng cách nào dễ hơn.

a. Nếu $\varphi(X)$ là hàm tuyến tính

b. Nếu $\varphi(X)$ là hàm phi tuyến.

4.45. Cho $Z = X^2 + Y^2$ với X và Y có bảng phân phối xác suất đồng thời như sau:

Y \ X	0	2	4
0	0,1	0,1	0
2	0,1	0,4	0,1
4	0	0,1	0,1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Tìm E(Z) bằng cách trực tiếp dùng bảng phân phối xác suất đồng thời.

b. Tìm các kỳ vọng toán sau đây bằng cách đơn giản nhất.

+ $E[(X - 2)(Y - 2)]$

+ $E[(X - 2)^2]$

+ $E(4X - 2Y)$

Bài tập tổng hợp chương IV

4.46. Một nhân viên bán hàng mỗi ngày đi chào hàng ở ba nơi với xác suất bán được hàng mỗi nơi là 0,2. Nếu bán được hàng ở nơi thứ nhất và thứ hai thì tiền lãi mỗi nơi là 100 USD còn nếu bán được ở nơi thứ ba thì do hàng có chất lượng cao hơn nên lãi được 200 USD.

a. Thiết lập bảng phân phối xác suất đồng thời của số lần bán được hàng X và tổng số tiền lãi Y.

b. Tìm số lần bán được hàng trung bình và số tiền lãi trung bình của mỗi ngày bán hàng.

4.47. Cho bảng phân phối xác xuất đồng thời của biến ngẫu nhiên rời rạc hai chiều (X, Y) như sau:

Y \ X	1	2	3	P(x)
-2	0,05			0,4
-1		0,2		
P(y)	0,2	0,5		

a. Hãy điền vào các giá trị còn thiếu của bảng.

b. Tìm E(X + Y)

c. Tìm V(X- Y)

4.48. Cho biến ngẫu nhiên rời rạc hai chiều có xác suất đồng thời được xác định bằng biểu thức sau:

$$P(x_i, y_j) = \begin{cases} C(x_i + 2y_j) & \text{với } x_i = 1; 2; 3 \text{ và } y_j = 0; 1; 2 \\ 0 & \text{với các giá trị khác của } x_i; y_j \end{cases}$$

Hãy thiết lập bảng phân phối xác suất đồng thời của (X,Y)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.49. Thống kê những người tuổi trưởng thành ở một nước theo hai chỉ tiêu sau:

X - thu nhập sau thuế (ngàn USD/năm)

Y - mức độ hạnh phúc của người đó

Thu được bảng số liệu sau:

X \ Y	Bất hạnh	Hạnh phúc vừa phải	Rất hạnh phúc
2,5	0,03	0,12	0,07
7,5	0,02	0,13	0,11
12,5	0,01	0,13	0,14
17,5	0,01	0,09	0,14

a. Tìm mức độ hạnh phúc trung bình tương ứng với mỗi mức thu nhập. Vẽ đồ thị và cho nhận xét.

b. Tìm hệ số tương quan và cho nhận xét.

ĐS: a. E(Y/X) = 2,5) = 1,118; 1,346 v.v...

b. $\rho_{XY} \approx 0,21$

4.50. Một chủ cửa hàng điện tử dự kiến doanh thu của cửa hàng trong tháng có phân phối xác suất như sau. Cột cuối cùng là mức thuế doanh thu tương ứng.

Doanh thu x	Xác suất P(x)	Thuế T(x)
20	0,10	4
30	0,30	6
40	0,40	9
50	0,20	13

a. Tìm doanh thu kỳ vọng

b. Tìm mức thuế kỳ vọng.

c. Tìm doanh thu sau thuế kỳ vọng.

ĐS: a. 37; b. 8,4; c. 28,6

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.51. Thống kê dân số của một nước theo hai chỉ tiêu học vấn X và giới tính Z thu được kết quả sau

X \ Z	Nam 0	Nữ 1
Thất học 0	0,04	0,05
Tiểu học 1	0,10	0,12
Trung học 2	0,23	0,29
Đại học 3	0,10	0,07

a. Lập bảng phân phối xác suất biên của học vấn.
b. Tìm xác suất để lấy ngẫu nhiên một người thì người đó tốt nghiệp từ trung học trở lên.

c. Tìm trình độ học vấn trung bình.

d. Lập bảng phân phối xác suất biên của Z và tính E(Z). Cho nhận xét.

ĐS: b. P = 0,69

c. E(X) = 1,77

d. E(Z) = 0,53

4.52. Tiếp tục bài 4.51.

a. Lập bảng phân phối xác suất về học vấn của nam.

b. Học vấn có độc lập với giới tính không.

c. Tìm trình độ học vấn trung bình của nam.

d. Tìm trình độ học vấn trung bình của nữ và so sánh với trình độ học vấn trung bình của nam.

e. Hãy so sánh E(X) tìm được ở bài 4.30 với hai kỳ vọng toán có điều kiện tìm được ở trên.

ĐS: a. p(x/z) = 0.085; 0.213; 0,489; 0,213

b. X phụ thuộc vào Z

c. 1,83 d. 1,72

e. E(X) nằm trong khoảng hai kỳ vọng toán có điều kiện

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.53. Tiếp tục bài 4.51. Giả sử tiền lương S của mỗi người phụ thuộc vào học vấn và giới tính như sau:

$$S = 10 + 2X + 4Z$$

a. Lập bảng phân phối xác suất của S từ đó tìm lương kỳ vọng E(S).

b. E(S) có thể tìm bằng cách nào khác?

ĐS: a. E(S) = 15,66

b. E(S) = 10 + 2E(X) + 4E(Z)

4.54. Tiếp tục bài 4.53 và tìm lương kỳ vọng E(S) nếu tiền lương S phụ thuộc vào học vấn và giới tính như sau:

a. $S = 10 + 10XZ$

b. $S = 10 + 3X - 2Z$

c. $S = (2X + 40)/5$

d. $S = 10 + X^2$

ĐS: a. 19,1 b. 14,25 c. 8,7 d. 13,83

4.55. Cho bảng phân phối xác suất đồng thời của biến ngẫu nhiên hai chiều (X,Y):

X \ Y	0	1	2	3	Tổng số
1	0,08	0,12	0,16	0,04	0,4
2	0,06	0,09	0,12	0,03	0,3
3	0,06	0,09	0,12	0,03	0,3
Tổng số	0,2	0,3	0,4	0,1	1

a. X và Y có độc lập với nhau không? Tại sao? Tính Cov (X, Y), E(X), E(Y), V(X - Y).

b. Lập bảng phân phối xác suất của (X - Y). Tính xác suất để (X + Y) bé hơn 2.

4.56. Một hộp đựng 5 sản phẩm, trong đó có 3 phế phẩm. Các sản phẩm được lấy ra để kiểm tra cho tới khi phát hiện thấy phế phẩm.

Ký hiệu X_1 là số lần kiểm tra tới khi phế phẩm đầu tiên bị phát hiện, X_2 là số lần kiểm tra thêm cho tới khi phế phẩm thứ hai bị phát hiện.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Lập bảng phân phối xác suất đồng thời của (X_1, X_2)

b. Tính hiệp phương sai giữa X_1 và X_2

4.57. Cho (X,Y) là biến ngẫu nhiên hai chiều có bảng phân phối xác suất như sau (trong đó y, a, b là các số chưa biết).

X \ Y	y	0	1
0	1/4	a	1/8
1	1/5	b	1/10

a. Xác định a và b sao cho X và Y là độc lập với nhau.

b. Quy luật phân bố xác suất có điều kiện của X đối với các giá trị khác nhau của Y trong trường hợp này như thế nào?

c. Giả thiết rằng a = 1/5, xác định giá trị y sao cho hiệp phương sai của (X,Y) bằng không. X và Y trong trường hợp này có độc lập với nhau không?

4.58. Một người tham gia đấu thầu 6 dự án nhỏ với xác suất thắng thầu mỗi dự án là 0,4. Nếu thắng thầu mỗi dự án người đó thu được 200USD. Chi phí để chuẩn bị cả 6 dự án là 300 USD.

a. Số dự án mà người đó hy vọng sẽ thắng thầu là bao nhiêu?

b. Lợi nhuận kỳ vọng E(P) là bao nhiêu?

c. Tìm xác suất để người đó có lãi khi dự thầu.

4.59. Tiếp tục bài 4.58 và giả sử càng thắng thầu nhiều thì càng phải thuê thêm công nhân làm ngoài giờ. Vì vậy lợi nhuận P phụ thuộc vào số dự án thắng thầu X như sau:

$P = 200X - 300 - 20X^2$

Lúc đó thì lợi nhuận kỳ vọng là bao nhiêu và xác suất có lãi khi dự thầu bằng bao nhiêu.

ĐS: E(P) = 36; P = 0,767

4.60. Tiếp tục bài 4.59 có người khuyên là không nên đấu thầu dự án thứ 6 vì dù có thắng thầu được cả 6 dự án thì chi phí thuê công nhân làm ngoài giờ sẽ làm cho dự án thứ 6 bị lỗ. Vậy có nên nghe theo lời khuyên đó không (biết rằng người đó đã bỏ ra 300 USD để chuẩn bị dự thầu cả 6 dự án).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.61. Thống kê dân số của một nước theo ba chỉ tiêu học vấn X, lứa tuổi Y và giới tính Z thu được bảng số liệu sau

Học vấn X	Giới tính Z					
	Nam 0			Nữ 1		
	Lứa tuổi Y			Lứa tuổi Y		
	30	45	70	30	45	70
Thất học 0	0,004	0,012	0,027	0,004	0,011	0,033
Tiểu học 1	0,016	0,033	0,047	0,017	0,036	0,063
Trung học 2	0,084	0,091	0,056	0,093	0,115	0,086
Đại học 3	0,037	0,042	0,021	0,030	0,028	0,014

a. Từ bảng số liệu trên hãy xây dựng bảng phân phối xác suất đồng thời của học vấn X và giới tính Z.

b. Hãy xây dựng bảng phân phối xác suất đồng thời của học vấn X và lứa tuổi Y.

c. Hãy so sánh trình độ học vấn trung bình của nam và nữ ở độ tuổi 30.

d. Tiến hành tương tự với độ tuổi 45 và 70.

ĐS: c. $E(X/Z = 0) = 2,09$; $E(X/Z = 1) = 2,03$

d. Với tuổi 45: 1,92 và 1,84;

với tuổi 70: 1,47 và 1,41

1.62. Điều tra thu nhập của của 10 cặp vợ chồng (tính bằng triệu đồng/năm) thu được kết quả sau:

Thu nhập của chồng (X): 20 30 30 20 20 30 40 30 40 40

Thu nhập của vợ (Y): 15 35 25 25 25 15 25 25 35 25

a. Lập bảng phân phối xác suất đồng thời của (X, Y)

b. Lập bảng phân phối biên của X, từ đó tìm E(X) và V(X)

c. Tiến hành tương tự đối với Y.

d. Tìm hiệp phương sai Cov(X,Y). X và Y có độc lập với nhau không.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

e. Tính kỳ vọng toán và phương sai của tổng thu nhập S của mỗi cặp vợ chồng.

ĐS: b. E(X) = 30; V(X) = 60; c. E(Y) = 25; V(Y) = 40

d. Cov(X,Y) = 20; e. E(S) = 55; V(S) = 140

4.63. Tiếp tục bài 4.62.

a. Giả sử thu nhập sau thuế W của các cặp vợ chồng được xác định bằng biểu thức

$$W = 0,6X + 0,8Y$$

Tìm E(W) và V(W)

b. Giả sử tiền đóng bảo hiểm xã hội Z của các cặp vợ chồng được xác định bằng biểu thức.

$$Z = 0,1X + 0,2Y$$

Tìm E(Z) và V(Z)

c. Để đo mức độ phân biệt giới tính các nhà xã hội học thường dùng công thức.

$$D = X - Y$$

Tìm E(D) và V(D) và cho nhận xét

d. Có ba cách đánh thuế thu nhập T như sau:

α. T = 0,20S

β. T = 0,40 (S - 15)

γ. Thu thuế theo bảng thuế luỹ tiến sau

S:	30	35	40	45	50	55	60	65	70	75
T:	4	5	6	7	9	11	13	16	19	22

Hỏi cách tính thuế nào có lợi nhất cho chính phủ.

ĐS: a. E(W) = 38; V(W) = 66,4 d. E(D) = 5 V(D) = 60

4.64. Giá cổ phiếu A và B là các biến ngẫu nhiên X_A và X_B (Đơn vị: ngàn đồng) có bảng phân phối xác suất đồng thời như sau:

X_B \ X_A	15	16	17
15	0,15	0,2	0,25
17	0,05	0,2	0,15

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Tính giá trị trung bình của các cổ phiếu trên.

b. X_A và X_B có độc lập với nhau hay không?

c. Khả năng để giá cổ phiếu B cao hơn giá trung bình của cổ phiếu A là bao nhiêu?

d. Cổ phiếu nào có mức độ rủi ro cao hơn.

ĐS: a. $E(X_A) = 16,2$

$E(X_B) = 15,8$

b.Vì $p(x_i, x_j) \neq p(x_i)p(x_j)$ nên X_A và X_B phụ thuộc

c. $P(X_B > 16,2) = P(X_B = 17) = 0,4$

d. $V(X_A) = 0,56$

$V(X_B) = 0,96$

Vậy cổ phiếu B có mức độ rủi ro cao hơn.

4.65. Khi đầu tư vào hai thị trường A và B lãi suất X_A và X_B (Đơn vị: %) là các biến ngẫu nhiên có bảng phân phối xác suất đồng thời như sau:

X_B \ X_A	- 1	3	5
- 2	0,05	0,1	0,15
3	0,1	0,2	0,15
6	0,1	0,1	0,05

a. Muốn có lãi suất hơn 3% thì nên lựa chọn phương án nào trong số các phương án sau:

+ Phương án 1: Đầu tư toàn bộ vào A.

+ Phương án 2: Đầu tư toàn bộ vào B.

+ Phương án 3: Chia đều vốn đầu tư vào A và B.

b. Tìm lãi suất trung bình khi đầu tư vào thị trường B, biết năm đó lãi suất thị trường A là 3%.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

c. Nếu muốn giảm rủi ro đầu tư xuống mức tối thiểu thì nên đầu tư như thế nào?

ĐS:

a. + Phương án 1: Đầu tư toàn bộ vào A.

$P(X_A > 3) = 0,35$

+ Phương án 2: Đầu tư toàn bộ vào B.

$P(X_B > 3) = 0,25$

+ Phương án 3: chia đều vốn vào cả hai thị trường.

$P(X_A + X_B > 3) = 0,3$

Vậy nên chọn phương án 1.

b. $E(X_B/X_A=3) = 2,5$

c. Để giảm rủi ro đầu tư xuống mức tối thiểu thì dành 60,9% vốn cho thị trường A và 39,1% vốn cho thị trường B.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Lê Văn Phong. *Giáo trình lý thuyết xác suất*. ĐH Kinh tế Kế hoạch. Hà Nội. 1972.
2. Lê Văn Phong. *Giáo trình lý thuyết xác suất*. NXB ĐH Kinh tế quốc dân. Hà Nội. 2007.
3. Nguyễn Đình Cử, Nguyễn Cao Văn. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Thống kê. Hà Nội. 1991.
4. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB ĐH Kinh tế quốc dân. Hà Nội. 2008.
5. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà nội. 2009.
6. Harald Cramer. *Phương pháp toán học trong thống kê*. NXB Khoa học và Kỹ thuật. Hà Nội. 1970
7. Гмурман В.Е. Теория вероятностей и математическая статистика. Москва. 1972.
8. Гнеденко Б. В. Теория вероятностей. Москва. 1969.
9. Уйпкс. Математическая статистика. Москва. 1982.
10. Многомерный статитистический анализ. Москва. 1982.
11. Giard V. Statistique appliquée à la gestion. Economica. Paris. 1992
12. Amir D Aczel. Complete Business Statistics. Irwin Boston 1993.
13. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992
14. De groot M. Probability and Statistics. New york. 1989

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 5
CÁC ĐỊNH LÝ GIỚI HẠN

1. MỞ ĐẦU

Như ta đã thấy ở các chương trước, không thể dự đoán trước được một cách chắc chắn xem biến ngẫu nhiên sẽ nhận giá trị nào trong các giá trị có thể có của nó khi thực hiện phép thử. Điều đó phụ thuộc vào rất nhiều nhân tố mà ta không thể tính hết được. Tuy nhiên vấn đề sẽ khác đi khi ta xét cùng một lúc một số lớn các biến ngẫu nhiên. Với một số điều kiện khá rộng rãi, hành vi tổng thể của một số lớn các biến ngẫu nhiên lại gần như mất đi tính ngẫu nhiên và trở nên có quy luật. Nói cách khác, khi ta tổng hợp một số lượng lớn các biến ngẫu nhiên thì tính ngẫu nhiên của hiện tượng mất đi và quy luật tất nhiên của nó được bộc lộ.

Đối với thực tiễn thì điều quan trọng là phải xác định các điều kiện trong đó tác động đồng thời của rất nhiều nguyên nhân ngẫu nhiên sẽ dẫn đến kết quả gần như không phụ thuộc gì vào các yếu tố ngẫu nhiên nữa và lúc đó ta có thể dự đoán được tiến trình của hiện tượng. Các điều kiện này được chỉ ra trong các định lý có tên là các định lý giới hạn mà một vài kết luận của chúng đã được đề cập ở các chương trước. Ở đây ta sẽ chỉ xét một số định lý có nhiều ứng dụng hơn cả trong thực tế.

Mục 2 sẽ xét một công cụ bổ trợ là bất đẳng thức Trêbusép. Mục 3 và 4 sẽ xét hai định lý của luật số lớn về sự hội tụ của tham số có ứng dụng nhiều trong kinh tế là định lý Trêbusep và định lý Bernoully. Mục 5 sẽ đề cập đến định lý giới hạn trung tâm, đây là định lý khái quát về sự hội tụ của các quy luật phân phối xác suất.

2. BẤT ĐẲNG THỨC TRÊBUSÉP

Nếu X là biến ngẫu nhiên có kỳ vọng toán và phương sai hữu hạn thì với mọi số dương ε *tuỳ ý ta đều có:*

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2} \qquad (5.1)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chứng minh. Ta sẽ chứng minh cho trường hợp X là biến ngẫu nhiên rời rạc. Việc chứng minh cho trường hợp X là biến ngẫu nhiên liên tục cũng tiến hành tương tự.

Giả sử X là biến ngẫu nhiên rời rạc với các giá trị có thể là $x_1, x_2, ..., x_n$ với các xác suất tương ứng $p_1, p_2, ..., p_n$. Ta giả thiết thêm là k giá trị đầu tiên của X thoả mãn điều kiện $|x_i - E(X)| < \varepsilon$, còn n - k giá trị còn lại thoả mãn điều kiện $|x_i - E(X)| \geq \varepsilon$. Vì các biến cố để thực hiện các bất đẳng thức $|X - E(X)| < \varepsilon$ và $|X - E(X)| \geq \varepsilon$ đối lập với nhau, do đó:

$$P(|(X - E(X)| < \varepsilon) = 1 - P(|(X - E(X)| \geq \varepsilon) \quad (*)$$

Theo định nghĩa của phương sai của biến ngẫu nhiên rời rạc ta có:

$$V(X) = [x_1 - E(X)]^2 p_1 + ... + [x_K - E(X)]^2 p_K +$$
$$+ [x_{K+1} - E(X)]^2 p_{K+1} + ... + [x_n - E(X)]^2 p_n$$

Rõ ràng là tất cả các số hạng của tổng trên đều không âm. Nếu ta bỏ bớt đi k số hạng đầu của tổng trên thì tổng chỉ có thể giảm đi. Do đó:

$$V(X) \geq [x_{k+1} - E(X)]^2 p_{k+1} + ... + [x_n - E(X)]^2 p_n$$

Chú ý rằng đối với các giá trị $x_i; (i = \overline{k+1, n})$ theo giả thiết ta đều có $|x_i - E(X)| \geq \varepsilon; (i = \overline{k+1, n})$ do đó $[x_i - E(X)]^2 \geq \varepsilon_i^2$; nên nếu trong mỗi số hạng ta thay một thừa số $[x_i - E(X)]^2$ bằng ε^2; thì bất đẳng thức đã cho tiếp tục mạnh thêm. Ta có:

$$V(X) \geq (p_{k+1} + ... + p_n)\varepsilon^2$$

Theo định lý cộng xác suất tổng $p_{k+1} + ... + p_n$ là xác suất để biến ngẫu nhiên X nhận một trong các giá trị $x_{k+1} + ... + x_n$ song mọi giá trị nói trên đều thoả mãn bất đẳng thức $(|x_i - E(X)| \geq \varepsilon)$. Từ đó suy ra là tổng $p_{k+1} + ... + p_n$ chính là xác suất

$$P(|X - E(X)|) \geq \varepsilon$$

Từ đó ta có:

$$V(X) \geq P(|X - E(X)| \geq \varepsilon)\varepsilon^2$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hay $$P(|X - E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2} \quad (5.2)$$

Thay (5.2) vào (*) ta thu được:

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}$$

Trong thực tế biểu thức (5.2) cũng được sử dụng như một dạng khác của bất đẳng thức Trêbusép.

Về mặt thực tiễn bất đẳng thức Trêbusép chỉ cho phép đánh giá cận trên hoặc cận dưới xác suất để biến ngẫu nhiên X nhận giá trị sai lệch so với kỳ vọng toán của nó lớn hơn hoặc nhỏ hơn ε. Đôi khi sự đánh giá đó là hiển nhiên và không có ý nghĩa. Chẳng hạn nếu $V(X) \geq \varepsilon^2$ thì bất đẳng thức Trêbusép cho kết quả hiển nhiên. Song nó lại có ưu điểm là áp dụng được đối với mọi biến ngẫu nhiên mà không cần biết quy luật phân phối xác suất của nó. Về mặt lý thuyết, bất đẳng thức Trêbusép có ý nghĩa rất to lớn. Nó được sử dụng để chứng minh các định lý của luật số lớn.

Thí dụ. Thu nhập trung bình hàng năm của dân cư một vùng là 700USD và độ lệch chuẩn là 120USD. Hãy xác định một khoảng thu nhập hàng năm xung quanh giá trị trung bình của ít nhất 95% dân cư vùng đó.

Giải. Gọi X là thu nhập hàng năm của dân cư vùng đó thì X là biến ngẫu nhiên với quy luật phân phối xác suất chưa biết song có kỳ vọng toán E(X) = 700 và độ lệch chuẩn $\sigma_x = 120$. Do đó theo bất đẳng thức Trêbusép.

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}$$

$$\rightarrow P(|X - 700| < \varepsilon) \geq 1 - \frac{120^2}{\varepsilon^2} = 0,95$$

Từ đó $\varepsilon = 536,656$

Vậy ít nhất 95% dân cư vùng đó có thu nhập hàng năm nằm trong khoảng (700 - 536,656; 700 + 536,656) tức là khoảng (163,344; 1236,656).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

5.1. Có 10 máy hoạt động độc lập với nhau. Xác suất để trong ca làm việc mỗi ngày máy bị hỏng là 0,05. Dựa vào bất đẳng thức Trebusep hãy đánh giá xác suất của sự sai lệch giữa số máy hỏng và số máy hỏng trung bình.

a. Nhỏ hơn 2.

b. Lớn hơn 2.

ĐS: ***a.*** $P \geq 0,88$

b. $P \leq 0,12$

5.2. Xác suất xuất hiện sản phẩm loại 1 (biến cố A) khi kiểm tra một sản phẩm là 0,5. Sử dụng bất đẳng thức Trêbusep để đánh giá xác suất của X là số lần xuất hiện biến cố A nằm trong khoảng từ 40 đến 60 nếu tiến hành kiểm tra 100 sản phẩm (kiểm tra từng sản phẩm một).

ĐS: $P \geq 0,75$

5.3. Bảng phân phối xác suất của biến ngẫu nhiên X như sau:

X	0,1	0,4	0,6
P	0,2	0,3	0,5

Sử dụng bất đẳng thức Trêbusep hãy đánh giá xác suất của biến cố

$|X - E(X)| < \sqrt{0,4}$

ĐS: $P > 0,909$

5.4. Dùng bất đẳng thức Trêbusep hãy đánh giá xác suất để biến ngẫu nhiên phân phối chuẩn nhận giá trị sai lệch so với kỳ vọng toán lớn hơn ba lần độ lệch chuẩn.

5.5. Xác suất chậm tàu của mỗi hành khách là 0,007. Dùng bất đẳng thức Trêbusep hãy đánh giá xác suất để trong 20.000 hành khách có từ 100 - 180 người chậm tàu.

ĐS: $P \geq 0,9131$

5.6. Phải kiểm tra bao nhiêu chi tiết để với xác suất không nhỏ hơn 0,98 có thể hy vọng rằng sai lệch giữa tần suất xuất hiện chi tiết tốt và xác suất để chi tiết là tốt bằng 0,95 sẽ không vượt quá 0,01.

ĐS: Kiểm tra 23 750 chi tiết

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5.7. Xác suất để chi tiết sản xuất ra đạt tiêu chuẩn là 0,8. Hãy dùng bất đẳng thức Trêbusep để đánh giá xác suất tỷ lệ chi tiết đạt tiêu chuẩn trong 4000 sản phẩm sẽ nằm trong khoảng từ 78% đến 83%.

3. ĐỊNH LÝ TRÊBUSÉP

Nếu các biến ngẫu nhiên $X_1, X_2, ..., X_n ...$ *độc lập từng đôi, có các kỳ vọng toán hữu hạn và các phương sai đều bị chặn trên bởi hằng số* $C\ (V(X_i) \le C; i=\overline{1,n})$ *thì với mọi* ε *dương bé tuỳ ý ta luôn có:*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+...+X_n}{n} - \frac{E(X_1)+E(X_2)+...+E(X_n)}{n}\right| < \varepsilon\right) = 1 \quad (5.3)$$

Chứng minh. Xét biến ngẫu nhiên $\overline{X}$ là trung bình số học của các biến ngẫu nhiên nói trên:

$$\overline{X} = \frac{X_1+X_2+...+X_n}{n}$$

Tìm kỳ vọng toán và phương sai của $\overline{X}$:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$

Áp dụng bất đẳng thức Trêbusép đối với biến ngẫu nhiên $\overline{X}$:

$$P(\left|\overline{X} - E(\overline{X})\right| < \varepsilon) \ge 1 - \frac{V(\overline{X})}{\varepsilon^2} = 1 - \frac{\sum_{i=1}^{n} V(X_i)}{n^2\varepsilon^2}$$

Theo giả thiết $V(X_i) \le C; i = \overline{1,n}$ do đó trong biểu thức trên nếu ta thay mỗi $V(X_i) \le C; i = \overline{1,n}$ bằng C thì bất đẳng thức sẽ chỉ mạnh thêm:

$$P(\left|\overline{X} - E(\overline{X})\right| < \varepsilon) \ge 1 - \frac{nC}{n^2\varepsilon^2} = 1 - \frac{C}{n\varepsilon^2}$$

Lấy giới hạn của cả hai vế khi $n \to \infty$, ta có:

$$\lim_{n\to\infty} P(\left|\overline{X} - E(\overline{X})\right| < \varepsilon) \ge \lim_{n\to\infty}\left(1 - \frac{C}{n\varepsilon^2}\right) = 1$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sau nữa ta chú ý rằng xác suất của một biến cố không thể lớn hơn một, do đó $\lim\limits_{n\to\infty} P(|\overline{X} - E(\overline{X})| < \varepsilon) = 1$

Ở trên, ta giả thiết là các biến ngẫu nhiên $X_1, X_2, ..., X_n$ có các kỳ vọng toán khác nhau. Trong thực tế thường gặp trường hợp các biến ngẫu nhiên có cùng một kỳ vọng toán. Lúc đó có thể áp dụng trường hợp riêng của định lý Trêbusép như sau:

Nếu $X_1, X_2, ..., X_n$ *là các biến ngẫu nhiên độc lập từng đôi, có cùng kỳ vọng toán* $(E(X_i) = m; i = \overline{1,n})$ *và các phương sai cùng bị chặn trên* $(V(X_i) \le C; i = \overline{1,n})$ *thì với mọi* ε *dương bé tuỳ ý ta luôn có:*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X + ... + X_n}{n} - m\right| < \varepsilon\right) = 1 \qquad (5.4)$$

Định lý trên còn còn được gọi là *luật số lớn của Trêbusép*.

Bản chất của định lý Trêbusép là nó chứng minh sự hội tụ theo xác suất của trung bình số học của một số lớn các biến ngẫu nhiên về trung bình số học của các kỳ vọng toán tương ứng. Nói cách khác nó chứng tỏ sự ổn định của trung bình số học của một số lớn các biến ngẫu nhiên xung quanh trung bình số học của các kỳ vọng toán của các biến ngẫu nhiên ấy.

Như vậy, mặc dù từng biến ngẫu nhiên độc lập có thể nhận giá trị khác nhiều so với kỳ vọng toán của chúng, song trung bình số học của một số lớn các biến ngẫu nhiên lại nhận giá trị gần bằng trung bình số học của các kỳ vọng toán của chúng với xác suất rất lớn. Điều đó cho phép dự đoán giá trị của của trung bình số học của các biến ngẫu nhiên.

Trong thực tế định lý Trêbusép có ứng dụng rộng rãi trong nhiều lĩnh vực, chẳng hạn trường hợp riêng của nó chính là cơ sở cho phương pháp đo lường trong vật lý. Để xác định giá trị của một đại lượng vật lý nào đó người ta thường tiến hành đo n lần và lấy trung bình số học của các kết quả đo làm giá trị thực của đại lượng cần đo. Thật vậy, giả sử xem kết quả của n lần đo là các biến ngẫu nhiên $X_1, X_2, ..., X_n$. Ta thấy rằng đối với các biến ngẫu nhiên này có thể áp dụng được trường hợp riêng của định lý Trêbusép vì chúng độc lập với nhau (từ đó chúng cũng độc lập từng đôi một với nhau), có cùng kỳ vọng toán vì nếu không có sai số hệ thống thì kỳ vọng toán của các biến ngẫu nhiên ấy chính bằng giá trị thực của đại lượng vật lý. Cuối

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

cùng các phương sai của chúng đều bị chặn trên bằng chính độ chính xác của thiết bị đo. Do đó theo định lý Trêbusép ta có thể chứng tỏ rằng trung bình số học của các kết quả đo sẽ sai lệch rất ít so với giá trị thực của đại lượng vật lý và điều đó xảy ra với xác suất gần như bằng một.

Định lý Trêbusép còn là cơ sở cho một phương pháp được áp dụng rộng rãi trong thống kê là phương pháp mẫu mà thực chất của nó là dựa vào một mẫu ngẫu nhiên khá nhỏ có thể kết luận về toàn bộ tập hợp tổng quát của các đối tượng được nghiên cứu.

Chẳng hạn để đánh giá năng suất cây trồng của một vùng nào đó người ta không cần phải điều tra trên toàn bộ diện tích của vùng đó mà chỉ cần dựa vào kết quả thu hoạch được của một mẫu ngẫu nhiên khá nhỏ mà vẫn đưa ra được các kết luận đủ chính xác về năng suất cây trồng của vùng đó.

Qua vài ví dụ như vậy có thể thấy được ý nghĩa to lớn của định lý Trêbusép đối với thực tiễn.

Bài tập

5.8. Dãy các biến ngẫu nhiên độc lập X_i ($i = \overline{1, n}$) có quy luật phân phối xác suất như sau:

X	-ia	0	ia
P	$1/2i^2$	$1 - 1/i^2$	$1/2i^2$

Hỏi có thể áp dụng định lý Trêbusep đối với dãy các biến ngẫu nhiên nói trên không? Tại sao.

Hướng dẫn: Kiểm tra điều kiện của định lý Trêbusep.

5.9. Có thể áp dụng được định lý Trêbusep cho dãy các biến ngẫu nhiên X_i ($i = \overline{1, n}$) có các phân phối xác suất sau đây hay không.

a.

X_i	$-\sqrt{3}$	0	$\sqrt{3}$
P	1/3	1/3	1/3

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b.

X_i	a	-a
P	$\frac{i}{2i+1}$	$\frac{i+1}{2i+1}$

ĐS: a. Có b. Có

4. ĐỊNH LÝ BERNOULLI

Nếu f là tần suất xuất hiện biến cố A trong n phép thử độc lập và p là xác suất xuất hiện biến cố đó trong mỗi phép thử thì với mọi ε dương bé tuỳ ý ta luôn có

$$\lim_{n\to\infty} P\left(|f-p|<\varepsilon\right)=1 \qquad (5.5)$$

Chứng minh. Ta xét biến ngẫu nhiên $f=\frac{X}{n}$ là tần suất xuất hiện biến cố A trong n phép thử độc lập.

Ta tìm kỳ vọng toán và phương sai của nó:

$$E(f)=E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)$$

$$V(f)=V\left(\frac{X}{n}\right)=\frac{1}{n^2}V(X)$$

Song X là số lần xuất hiện biến cố A trong n phép thử độc lập, mà theo giả thiết xác suất xuất hiện biến cố trong mỗi phép thử bằng p. Như vậy X phân phối theo quy luật nhị thức với các tham số là n và p. Do đó ta có E(X)=np và V(X)=np(1-p), từ đó:

$$E(f)=\frac{1}{n}np=p$$

$$V(f)=\frac{1}{n^2}np(1-p)=\frac{p(1-p)}{n}$$

Áp dụng bất đẳng thức Trêbusép với biến ngẫu nhiên f ta có:

$$P\left(|f-p|<\varepsilon\right)\geq 1-\frac{p(1-p)}{n\varepsilon^2}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lấy giới hạn của hai vế khi $n \to \infty$:

$$\lim_{n\to\infty} P\left(|f-p|<\varepsilon\right) \geq \lim_{n\to\infty}\left(1-\frac{p(1-p)}{n\varepsilon^2}\right)=1$$

Mặt khác, vì xác suất không thể lớn hơn một, do đó:

$$\lim_{n\to\infty} P\left(|f-p|<\varepsilon\right)=1$$

Định lý trên còn được gọi là *luật số lớn của Bernoulli*.

Định lý Bernoulli chứng minh sự hội tụ theo xác suất của tần suất xuất hiện biến cố trong n phép thử độc lập về xác suất xuất hiện biến cố đó trong mỗi phép thử khi số phép thử tăng lên vô hạn. Nó chứng tỏ sự ổn định của tần suất xung quanh giá trị xác suất của biến cố đó.

Định lý Bernoulli là cơ sở lý thuyết của định nghĩa thống kê về xác suất, do đó nó cũng là cơ sở cho mọi áp dụng của định nghĩa thống kê về xác suất trong thực tế.

Chú ý rằng trong các định lý của luật số lớn ta chỉ đề cập đến sự hội tụ theo xác suất chứ không phải sự hội tụ theo nghĩa thông thường của giải tích toán học. Chẳng hạn theo định lý Bernoulli không thể kết luận rằng $\lim_{n\to\infty} f = p$ tức là không thể kết luận khi n đủ lớn thì f sẽ luôn luôn sai lệch không đáng kể so với p. Sự hội tụ theo xác suất chỉ có nghĩa là khi n đủ lớn thì việc f và p sai lệch nhau không đáng kể sẽ có thể xem như có xác suất bằng 1. Như vậy thì với từng giá trị riêng biệt của n, bất đẳng thức có thể vẫn không thoả mãn, tức là f và p vẫn có thể sai lệch nhau đáng kể. Vì vậy định lý Bernoulli có thể viết ngắn gọn như sau:

$$f \xrightarrow[n\to\infty]{\text{Hội tụ theo xác suất}} p$$

5. ĐỊNH LÝ GIỚI HẠN TRUNG TÂM

Định lý giới hạn tổng quát hơn cả là định lý giới hạn trung tâm của Liapunov (xem sử dụng của nó ở mục 8.8 chương III). Ở ta chỉ xét một trường hợp của nó được sử dụng nhiều trong thống kê. Trước hết ta xét khái niệm hàm đặc trưng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5.1. Hàm đặc trưng

1. Định nghĩa. *Hàm đặc trưng của biến ngẫu nhiên X là kỳ vọng toán của biến ngẫu nhiên e^{itx} và được ký hiệu là $\varphi_X(t)$. Tức là:*

$$\varphi_X(t) = E\left[e^{itx}\right] = E(\cos tX) + iE(\sin tX)$$

Như vậy nếu X là biến ngẫu nhiên rời rạc thì:

$$\varphi_X(t) = \sum_j e^{itx_j} P_j$$

Còn nếu X là biến ngẫu nhiên liên tục thì:

$$\varphi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f(x)dx$$

2. Các tính chất của hàm đặc trưng

a. $|\varphi_X(t)| \le 1$

Thật vậy vì $|e^{itx}| = 1$ nên:

$$\left|\int_{-\infty}^{+\infty} e^{itx} f(x)dx\right| \le \int_{-\infty}^{+\infty} |e^{itx}| f(x)dx = 1$$

b. Nếu $Y = ax + b$ thì $\varphi_Y(t) = e^{ibt}\varphi_X(at)$

Thật vậy, theo định nghĩa:

$$\varphi_Y(t) = E\left[e^{itY}\right] = E\left[e^{it(ax+b)}\right] = e^{itb}.E\left[e^{itax}\right] = e^{itb}\varphi_X(at)$$

c. F(x) xác định một cách duy nhất hàm đặc trưng $\varphi_X(t)$

d. Nếu $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên độc lập thì

$$\varphi_{X_1+...+X_n}(t) = \prod_{k=1}^{n} \varphi_{X_k}(t)$$

Thật vậy, dễ dàng thấy rằng vì các biến ngẫu nhiên $X_1, X_2, ..., X_n$ độc lập nên theo tính chất của kỳ vọng toán ta có:

$$\varphi_{X_1+...+X_n}(t) = E\left[e^{it(X_1+X_2+...+X_n)}\right]$$

$$= E\left[e^{itX_1}.e^{itX_2}e^{itX_n}\right]$$

$$= \prod_{k=1}^{n} E\left[e^{itX_k}\right] = \prod_{k=1}^{n} \varphi_{X_k}(t)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

e. Nếu tồn tại $E|X|^k$ thì hàm đặc trưng $\varphi_X(t)$ cũng tồn tại đạo hàm đến bậc k tại mọi điểm t (khả vi đến bậc k tại mọi điểm t).

Hệ quả: Nếu tồn tại $E|X|^k$ thì $\varphi_X(t)$ sẽ có khai triển Taylor như sau:

$$\varphi_X(t) = 1 + m_1 it + m_2 \frac{(it)^2}{2!} + ... + m_k \frac{(it)^k}{k!} + 0(t^k)$$

trong đó $m_\ell = E\left[X^\ell\right] \quad (\ell = \overline{1,k})$.

f. Nếu $\{F_n(x)\}$ là dãy hàm phân bố xác suất và $\{\varphi_n(t)\}$ là dãy các hàm đặc trưng tương ứng thì điều kiện cần và đủ để $\{F_n(x)\}$ hội tụ yếu (tức là hội tụ tại các điểm $F_n(x)$ liên tục) tới hàm phân bố xác suất $F(x)$ là $\{\varphi_n(t)\}$ hội tụ tại mọi t đến hàm đặc trưng $\varphi(t)$ tương ứng với $F(x)$.

Thí dụ 1. Cho biến ngẫu nhiên $X \sim A(p)$. Tìm hàm đặc trưng $\varphi_X(t)$.

Giải. Theo định nghĩa của quy luật A(p)

$$P_x = p^x(1-p)^{1-x} \text{ với } x = \overline{0,1}$$

nên

$$\varphi_X(t) = E\left[e^{itx}\right] = e^{it0}(1-p) + e^{it1}p = p.e^{it} + (1-p)$$

Thí dụ 2. Cho biến ngẫu nhiên $X \sim B(n,p)$. Tìm $\varphi_X(t)$.

Giải. Theo định nghĩa của quy luật B(n,p)

$$P_x = C_n^x p^x (1-p)^{n-x} \quad x = \overline{0,n}$$

nên

$$\varphi_X(t) = E\left[e^{itx}\right] = \sum_{x=0}^{n} e^{itx} C_n^x p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} C_n^x (pe^{it})^x (1-p)^{n-x} = [pe^{ti} + (1-p)]^n$$

Qua đó chứng minh mối liên hệ giữa quy luật nhị thức và quy luật không - một.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 3. Cho biến ngẫu nhiên $X \sim P(\lambda)$. Tìm $\varphi_X(t)$.

Giải. Theo định nghĩa của quy luật Poisson

$$P_x = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0,1...$$

nên

$$\varphi_X(t) = \sum_{x=0}^{\infty} e^{itx}e^{-\lambda}\frac{\lambda^x}{x!} = \sum_{x=0}^{\infty} e^{-\lambda}\frac{(\lambda e^{it})^x}{x!} = e^{-\lambda}.e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}$$

Ngoài ra có thể chứng minh được rằng nếu biến ngẫu nhiên $X \sim N(\mu, \sigma^2)$ thì hàm đặc trưng bằng:

$$\varphi_X(t) = e^{i\mu t\frac{\sigma^2 t^2}{2}}$$

và đối với phân phối N(0, 1) thì:

$$\varphi_U(t) = e^{-\frac{t^2}{2}}$$

Với khái niệm hàm đặc trưng ta sẽ xét định lý giới hạn trung tâm sau đây:

5.2. Định lý Lindenberg - Lewi

Nếu $X_1, X_2, ..., X_n$ *... là một dãy các biến ngẫu nhiên độc lập cùng tuân theo một quy luật phân phối xác suất nào đó với kỳ vọng toán và phương sai hữu hạn:*

$$E(X_K) = a, V(X_K) = \sigma^2; \forall k$$

Thì quy luật phân phối xác suất của biến ngẫu nhiên

$$U_n^c = \frac{U_n - E(U_n)}{\sqrt{V(U_n)}} \quad \text{với } U_n = \sum_{K=1}^{n} X_K$$

sẽ hội tụ khi $n \to \infty$ *tới quy luật chuẩn hoá N(0,1). Tức là:*

$$P(U_n^c < x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$$

Chứng minh. Theo tính chất của hàm đặc trưng, ta chỉ cần chỉ ra rằng hàm đặc trưng $\varphi_{U_n^c}(t)$ hội tụ đến hàm đặc trưng của phân phối chuẩn hoá là

$\varphi_U(t) = e^{-\frac{t^2}{2}}$ khi $n \to \infty$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Xét biến ngẫu nhiên

$$U_n - E(U_n) = (X_1 - E(X_1)) + (X_2 - E(X_2)) + ... + ((X_n - E(X_n))$$

nếu ký hiệu $Y_K = X_K - E(X_K)$ ta có:

$$U_n - E(U_n) = Y_1 + Y_2 + ... + Y_n$$

Mặt khác ta lại có

$$E(Y_K) = E[X_K - E(X_K)] = 0 \quad \forall K$$

$$V(Y_K) = E[Y_K^2] = \sigma^2 = V(X_K) \quad \forall K$$

Do đó nếu đặt $\varphi(t) = \varphi_{YK}(t)$ thì theo hệ quả trên ta có:

$$\varphi(t) = 1 + \frac{(it)^2}{2!}\sigma^2 + 0(t^2)$$

Hơn nữa, do $U_n - E(U_n) = \sum_{K=1}^{n} Y_K$ nên

$$\varphi_{U_n - E(U_n)^t} = [\varphi(t)]^n$$

Từ đó theo tính chất của hàm đặc trưng ta thu được hàm đặc trưng của U_n^c là:

$$\varphi_{U_n^c}(t) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n = \left[1 - \frac{\sigma^2}{2}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^2 + 0\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^2\right)\Bigg]^n =$$

$$= \left[1 - \frac{t^2}{2n} + 0\left(\frac{t^2}{\sigma^2 n}\right)\right]^n$$

do đó:

$$\lim_{n\to\infty} \varphi_{U_n^c}(t) = \lim_{n\to\infty}\left[1 - \frac{t^2}{2n} + 0\left(\frac{t^2}{\sigma^2 n}\right)\right]^n$$

$$= e^{\lim_{n\to\infty} n\left(\frac{-t^2}{2n} + 0\left(\frac{t^2}{\sigma^2 n}\right)\right)} = e^{-\frac{t^2}{2}}$$

Về mặt thực hành người ta có thể ứng dụng định lý giới hạn trung tâm như sau: Với n đủ lớn ta có thể cho rằng:

$$P(U_n^C < x) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \phi(x)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

hoặc

$$P(a < \bigcup n < b) \approx \phi_0\left(\frac{b - E(U_n)}{\sqrt{V(U_n)}}\right) - \phi_0\left(\frac{a - E(U_n)}{\sqrt{V(U_n)}}\right)$$

Thí dụ 4. Chọn ngẫu nhiên 192 số trên đoạn [0,1]. Tìm xác suất để tổng số điểm thu được (X) nằm trong khoảng (88,104).

Giải. Ta có thể coi như $X = \sum_{i=1}^{192} X_i$, trong đó mọi biến ngẫu nhiên X_i độc lập và cùng tuân theo quy luật phân phối đều U(0,1). Từ đó:

$$E(X_i) = \frac{1}{2} \text{ và } V(X_i) = \frac{1}{12} \quad \forall i$$

Từ đó $E(X) = nE(X_i) = 96$ và

$V(X) = nV(X_i) = 16$

$\rightarrow \sqrt{V(X)} = 4 = \sigma_X$

Vì vậy:

$$P(88 < X < 104) \approx \phi_0\left(\frac{104 - 96}{4}\right) - \phi_0\left(\frac{88 - 96}{4}\right)$$

$$= 2\phi_0(2) = 0,9544$$

Thí dụ 5. Cho biến ngẫu nhiên X~B(n=1000,p=0,02). Tìm xác suất để X nhận giá trị trong khoảng (40;50).

Giải. Có thể coi $X = \sum_{i=1}^{1000} X_i$ trong đó các X_i độc lập và cùng phân phối A(p=0,02). Từ đó theo định lý giới hạn trung tâm suy ra:

$$X \sim N(\mu = np = 20; \sigma = np(1-p) = 19,6)$$

do đó

$$P(40 < X < 50) \approx \phi_0\left(\frac{50 - 20}{\sqrt{19,6}}\right) - \phi_0\left(\frac{40 - 20}{\sqrt{19,6}}\right)$$

$$= \phi_0(6,77) - \phi_0(4,51) =$$

$$= 0,5 - 4,9999 = 0,0001$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6. TÓM TẮT CHƯƠNG 5

1. Các định lý giới hạn được xem là cơ sở để lý thuyết xác suất trở thành một khoa học. Nó chứng minh rằng khi xét một số lớn các hiện tượng tương tự thì quy luật của các hiện tượng ngẫu nhiên được bộc lộ ra. Nói cách khác, các định lý giới hạn chứng minh mối liên hệ giữa ngẫu nhiên và tất nhiên.
2. Các định lý của luật số lớn chứng minh sự hội tụ của các tham số. Định lý Trêbusep chứng minh sự hội tụ theo xác suất của giá trị trung bình còn định lý Bernoully chứng minh sự hội tụ theo xác suất của giá trị tần suất.
3. Các định lý giới hạn trung tâm chứng minh sự hội tụ của các quy luật phân phối xác suất. Chúng chứng minh rằng với một số điều kiện nhất định thì mọi quy luật phân phối xác suất đều hội tụ về quy luật chuẩn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các ký hiệu và công thức cơ bản

* Bất đẳng thức Trêbusép:

$$P(|X-E(X)|<\varepsilon)\geq 1-\frac{V(X)}{\varepsilon^2}$$

* Định lý Trêbusép

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+...+X_n}{n}-\frac{E(X_1)+E(X_2)+...+E(X_n)}{n}\right|<\varepsilon\right)=1$$

* Định lý Bernoulli

$$\lim_{n\to\infty} P(|f-p|<\varepsilon)=1$$

* Hàm đặc trưng $\varphi_X(t)=E\left[e^{itX}\right]$

Nếu $U\sim N(0,1)\Rightarrow \varphi_U(t)=e^{-\frac{t^2}{2}}$

* Định lý giới hạn trung tâm

Nếu $X_1,X_2,...,X_n$ ... độc lập và cùng tuân theo một quy luật phân phối xác suất với kỳ vọng toán và phương sai hữu hạn thì:

$$\lim_{n\to\infty} P(U_n^c<x)=\lim_{n\to\infty} F_{U_n^c}(x)=\phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$$

Trong đó: $U_n=\sum_{i=1}^{n}X_i$; $U_n^c=\frac{U_n-E(U_n)}{\sqrt{V(U_n)}}$

+ Với n đủ lớn ta có:

$$P(a<U_n<b)=\phi_0\left(\frac{b-E(U_n)}{\sqrt{V(U_n)}}\right)-\phi_0\left(\frac{a-E(U_n)}{\sqrt{V(U_n)}}\right)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Luật số lớn	Law of grand numbers
Bất đẳng thức Trêbusep	Chebyshev' inegality
Định lý Trêbusep	Chebyshev' theorem
Định lý Bernoully	Bernoully' theorem
Định lý giới hạn trung tâm	Central limit theorem
Định lý Lindenberg - Lewi	Lindenberg- Lewi' theorem
Hàm đặc trưng	Caracteristic function

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

5.10. Khi nào thì có thể áp dụng được bất đẳng thức Trêbusép đối với biến ngẫu nhiên X?

5.11. Khi nào thì việc áp dụng bất đẳng thức Trêbusép sẽ không có ý nghĩa?

5.12. Dùng bất đẳng thức Trêbusép hãy xây dựng quy tắc 3σ cho một biến ngẫu nhiên bất kỳ.

5.13. Cho biến ngẫu nhiên X với kỳ vọng toán bằng 20 và phương sai bằng 2. Dùng bất đẳng thức Trêbusép hãy đánh giá xác suất để X nhận giá trị trong khoảng (17;24).

5.14. Với những điều kiện nào thì có thể áp dụng Định lý Trêbusép đối với các biến ngẫu nhiên $X_1, X_2, ..., X_n$ ...?

5.15. Khái niệm hội tụ theo xác suất khác với khái niệm hội tụ trong giải tích ở điểm nào?

5.16. Dãy các biến ngẫu nhiên độc lập X_i (i=1,2,...) có bảng phân phối xác suất như sau:

X_i	$\sqrt{2}$	0	$-\sqrt{2}$
P_{X_i}	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Hỏi có thể áp dụng được luật số lớn của Trêbusép đối với dãy các biến ngẫu nhiên nói trên hay không?

5.17. Gọi $X_K (K = 1,2,...,n)$ là số mặt sấp ở lần tung thứ k của một đồng xu và $S_n = X_1 + X_2 + ... + X_K$. Hãy chứng tỏ rằng $\forall \varepsilon > 0$ bé tuỳ ý thì:

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - 0,5\right| \geq \varepsilon\right) = 0$$

5.18. Tìm hàm đặc trưng của biến ngẫu nhiên liên tục X phân phối đều trên đoạn [a,b].

5.19. Tìm hàm đặc trưng của biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} e^{-x} & \text{với } x \geq 0 \\ 0 & \text{với } x<0 \end{cases}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5.20. Cho dãy các biến ngẫu nhiên độc lập $X_1, X_2, ..., X_n$ cùng phân phối theo một quy luật nào đó với kỳ vọng toán đều bằng a và phương sai đều bằng b^2. Hãy áp dụng định lý giới hạn trung tâm của Lindenberg-Lewi để tìm quy luật phân phối xác suất của biến ngẫu nhiên

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

khi n đủ lớn.

Bài tập tổng hợp chương 5

5.21. Biến ngẫu nhiên rời rạc X có hàm phân bố xác suất như sau:

$$F(x) = \begin{cases} 0 & \text{với } x \le -2 \\ 1/8 & \text{với } -2 < x \le -1 \\ 3/8 & \text{với } -1 < x \le 0 \\ 4/8 & \text{với } 0 < x \le 1 \\ 1 & \text{với } x > 1 \end{cases}$$

a. Viết bảng phân phối xác suất của X.

b. Áp dụng bất đẳng thức Trêbusep đối với X nếu lấy $\varepsilon = 2$

5.22. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} 2K(10 - x) & x \in (0, 10) \\ 0 & x \notin (0, 10) \end{cases}$$

Áp dụng bất đẳng thức Trêbusep đối với X nếu lấy $\varepsilon = 2$

5.23. Thu nhập trung bình hàng năm của các giám đốc doanh nghiệp vừa và nhỏ là 18 triệu và độ lệch chuẩn là 1,6 triệu đồng. Hãy xác định khoảng thu nhập của 96% giám đốc.

ĐS: (10 triệu; 26 triệu)

5.24. Có thể sử dụng giá trị

$$S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - a)^2$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Làm giá trị đúng cho phương sai của sai số đo lường hay không nếu X_1, X_2, …, X_n,… là các kết quả đo độc lập của đại lượng a, có trùng hàm phân bố xác suất và các phương sai hữu hạn.

5.25. Cho biến ngẫu nhiên X_K chỉ nhận một trong hai giá trị là k^s hoặc $-k^s$ với xác suất như nhau. Với giá trị nào của s thì có thể áp dụng được luật số lớn của Trêbusep đối với trung bình số học của dãy các biến ngẫu nhiên độc lập X_1, X_2, … X_k...

5.26. Cho X_1, X_2, ...,X_n,... là n biến ngẫu nhiên độc lập cùng phân phối chuẩn với $E(X_i) = \mu$ và $V(X_i) = \sigma^2$ ($i = \overline{1,n}$). Tìm cận dưới của xác suất để trung bình số học của các biến ngẫu nhiên nói trên sai lệch so với kỳ vọng toán của nó nhỏ hơn ε ($\varepsilon > 0$).

ĐS: $P(\left|\overline{X} - \mu\right| < \varepsilon) \geq 1 - \dfrac{\sigma^2}{\varepsilon^2 . n}$

5.27. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \frac{1}{2} e^{-|x|}$$

Tìm xác suất tối thiểu để biến ngẫu nhiên X nhận giá trị sai lệch so với kỳ vọng toán không quá 2.

5.28. Cho X_1, X_2,… X_n là biến ngẫu nhiên độc lập, đều có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} 1/2 & \text{với } x - 1 \leq x \leq 1 \\ 0 & \text{với } x \notin [-1,1] \end{cases}$$

Đặt $S_n = X_1 + X_2 + ... + X_n$. Hãy chứng minh rằng

$$P\left\{\left|\frac{S_n}{n}\right| \geq \varepsilon\right\} \leq \frac{1}{3n\varepsilon^2}; \quad \forall \varepsilon > 0$$

5.29. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \dfrac{x^m}{m!} e^{-x} & \text{với } x \geq 0 \\ 0 & \text{với } x < 0 \end{cases}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hãy chứng minh rằng

$$P\,[\,0 < X < 2\,(m+1)] \geq \frac{m}{m+1}$$

5.30. Tung một con xúc xắc 3000 lần.

a. Tìm xác suất tối thiểu để số lần xuất hiện mặt sáu chấm nằm trong khoảng từ 450 đến 550.

b. Gọi $\overline{X}$ là số chấm trung bình xuất hiện trong 3000 lần tung ấy. Tìm xác suất của biến cố $|\overline{X} - 3{,}5| < 0{,}1$

ĐS: a. $P_{min} = \frac{5}{6}$ b. $P \geq 0{,}9$

5.31*. Phải tung một con xúc xắc bao nhiêu lần để với xác suất không nhỏ hơn 0,99 số chấm xuất hiện trung bình sai lệch so vo với 3,5 không vượt quá 1%.

ĐS: $n = 2{,}91.10^6$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Lê Văn Phong. *Giáo trình lý thuyết xác suất*. NXB ĐH Kinh tế quốc dân. Hà Nội. 2007.
2. Nguyễn Đình Cử, Nguyễn Cao Văn. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Thống kê. Hà Nội. 1991.
3. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà Nội. 2009.
4. Гмурман В.Е. Теория вероятностей и математическая статистика. Москва. 1972.
5. Гнеденко Б. В. Теория вероятностей. Москва. 1969.
6. De groot M. Probability and Statistics. New york. 1989

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phần thứ hai
THỐNG KÊ TOÁN

Thống kê toán là bộ môn toán học nghiên cứu quy luật của các hiện tượng ngẫu nhiên có tính chất số lớn trên cơ sở thu thập và xử lý các số liệu thống kê - các kết quả quan sát. Như vậy, nội dung chủ yếu của thống kê toán là xây dựng các phương pháp thu thập và xử lý các số liệu thống kê nhằm rút ra các kết luận khoa học và thực tiễn.

Các phương pháp thống kê toán là công cụ để giải quyết nhiều vấn đề khoa học và thực tiễn nảy sinh trong các lĩnh vực khác nhau của tự nhiên và kinh tế - xã hội.

Sử dụng phương pháp quy nạp, thống kê toán dùng thông tin của các hiện tượng có tính riêng lẻ để khái quát hoá thành quy luật của các hiện tượng ngẫu nhiên.

Phần này sẽ nghiên cứu một số nội dung cơ bản mang tính cơ sở của thống kê toán.

Chương 6 sẽ nghiên cứu khái niệm về mẫu ngẫu nhiên và phương pháp suy diễn thống kê.

Chương 7 sẽ nghiên cứu phương pháp ước lượng tham số của tổng thể.

Chương 8 sẽ đề cập đến một số nội dung cơ bản của lý thuyết kiểm định giả thuyết thống kê.

Chương 9 sẽ giới thiệu một số kiểm định phi tham số thường dùng trong thực tế.

Chương 10 sẽ đề cập đến một phương pháp thống kê quan trọng là phân tích phương sai.

Chương 11 sẽ nghiên cứu một số nội dung mở đầu của phương pháp phân tích hồi quy và tương quan.

Chương 12 sẽ giới thiệu một số phương pháp đơn giản nhất trong phân tích chuỗi thời gian.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

THƯ VIỆN
HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 6
CƠ SỞ LÝ THUYẾT MẪU

1. MỞ ĐẦU

Chương này sẽ giới thiệu một phương pháp được sử dụng rộng rãi trong thực tế là phương pháp mẫu, tức là dùng thông tin đã biết của một bộ phận của hệ thống để suy đoán về toàn bộ hệ thống đó. Quá trình suy đoán thống kê đòi hỏi phải sử dụng các mô hình xác suất đã được thiết lập ở phần trước song nó lại cho phép khắc phục hầu hết các hạn chế của phương pháp điều tra toàn bộ.

Mục 2 sẽ giới thiệu khái quát về phương pháp mẫu. Mục 3 sẽ giới thiệu khái niệm tổng thể và các tham số cơ bản của tổng thể cần xác định. Từ mục 4 đến mục 7 sẽ nghiên cứu các nội dung cơ bản của một mẫu ngẫu nhiên. Mục 7 sẽ đề cập đến một ứng dụng quan trọng của mẫu ngẫu nhiên là quá trình suy diễn thống kê.

2. KHÁI NIỆM VỀ PHƯƠNG PHÁP MẪU

Trong thực tế thường phải nghiên cứu một tập hợp các phần tử đồng nhất theo một hay nhiều dấu hiệu định tính hoặc định lượng đặc trưng cho các phần tử đó. Chẳng hạn một doanh nghiệp phải nghiên cứu tập hợp các khách hàng của nó thì dấu hiệu định tính có thể là mức độ hài lòng của khách hàng đối với sản phẩm hoặc dịch vụ của doanh nghiệp, còn dấu hiệu định lượng là nhu cầu của khách hàng về số lượng sản phẩm của doanh nghiệp.

Để nghiên cứu tập hợp các phần tử này theo một dấu hiệu nhất định, đôi khi người ta sử dụng phương pháp nghiên cứu toàn bộ, tức là thống kê toàn bộ tập hợp đó và phân tích từng phần tử của nó theo dấu hiệu nghiên cứu. Chẳng hạn để nghiên cứu dân số của một nước theo các dấu hiệu như tuổi tác, trình độ văn hoá, địa bàn cư trú, cơ cấu nghề nghiệp... có thể tiến hành tổng điều tra dân số và phân tích từng người theo các dấu hiệu trên từ

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

đó tổng hợp thành dấu hiệu chung cho toàn bộ dân số của nước đó. Tuy nhiên trong thực tế việc áp dụng phương pháp này gặp phải những khó khăn chủ yếu sau:

- Nếu qui mô của tập hợp quá lớn thì việc nghiên cứu toàn bộ sẽ đòi hỏi nhiều chi phí vật chất và thời gian.

- Nhiều khi cũng do qui mô của tập hợp quá lớn nên có thể xảy ra trường hợp tính trùng hoặc bỏ sót các phần tử của nó.

- Do qui mô nghiên cứu lớn mà trình độ tổ chức nghiên cứu lại hạn chế dẫn đến các sai sót trong quá trình thu thập thông tin ban đầu, hạn chế độ chính xác của kết quả phân tích.

- Trong nhiều trường hợp không thể nắm được toàn bộ các phần tử của tập hợp cần nghiên cứu, do đó không thể tiến hành nghiên cứu toàn bộ được.

- Nếu các phần tử của tập hợp lại bị phá huỷ trong quá trình nghiên cứu, thì phương pháp nghiên cứu toàn bộ trở thành vô nghĩa.

Vì thế trong thực tế phương pháp nghiên cứu toàn bộ thường chỉ được áp dụng với các tập hợp có qui mô nhỏ, còn chủ yếu người ta áp dụng phương pháp nghiên cứu không toàn bộ, đặc biệt là phương pháp nghiên cứu chọn mẫu. Phương pháp này chủ trương từ tập hợp cần nghiên cứu chọn ra một số phần tử (gọi là mẫu), phân tích các phần tử này và dựa vào đó mà suy ra các kết luận về tập hợp cần nghiên cứu. Nếu mẫu được chọn ra một cách ngẫu nhiên và xử lí bằng các phương pháp xác suất thì vừa thu được các kết luận một cách nhanh chóng, đỡ tốn kém mà vẫn đảm bảo độ chính xác cần thiết.

Việc thu thập, sắp xếp và trình bầy các số liệu của tổng thể hoặc một mẫu gọi là thống kê mô tả. Còn việc sử dụng thông tin của mẫu để tiến hành các suy đoán, kết luận về tổng thể gọi là thống kê suy diễn.

3. TỔNG THỂ NGHIÊN CỨU

3.1. Định nghĩa. *Toàn bộ tập hợp các phần tử đồng nhất theo một dấu hiệu nghiên cứu định tính hoặc định lượng nào đó được gọi là tổng thể nghiên cứu hay tổng thể.*

Số lượng các phần tử của tổng thể được gọi là kích thước của tổng thể, ký hiệu là N. Thường thì kích thước N của tổng thể là hữu hạn, song nếu tổng thể quá lớn hoặc không thể nắm được toàn bộ các phần tử của tổng thể ta có thể giả thiết rằng kích thước của tổng thể là vô hạn. Điều giả thiết này

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

dựa trên cơ sở là khi tăng kích thước của tổng thể lên khá lớn thì thực tế không ảnh hưởng gì đến kết quả tính toán trên số liệu của từng bộ phận rút ra từ tổng thể đó.

Với mỗi tổng thể ta không nghiên cứu trực tiếp tổng thể đó mà thông qua một hay nhiều dấu hiệu đặc trưng cho tổng thể đó. Chúng được gọi là dấu hiệu nghiên cứu, ký hiệu là χ. Các dấu hiệu này là định tính hoặc định lượng.

3.2. Các phương pháp mô tả tổng thể

1. Giả sử trong tổng thể dấu hiệu nghiên cứu định lượng χ nhận các giá trị $X_1, X_2, ..., X_k$ với các tần số tương ứng $N_1, N_2, ..., N_k$ (N_i là số phần tử của tổng thể có chung giá trị x_i). Lúc đó tổng thể có thể mô tả bằng bảng phân phối tần số sau:

Giá trị của χ	x_1	x_2	...	x_i	...	x_k
Tần số	N_1	N_2	...	N_i	...	N_k

Hiển nhiên

$$\begin{cases} 0 \le N_i \le N \quad \forall_i \\ \sum_{i=1}^{k} N_i = N \end{cases} \tag{6.1}$$

2. Nếu ký hiệu $p_i (i = \overline{1,k})$ là tần suất của x_i, tức là tỷ số giữa tần số của x_i và kích thước của tổng thể thì:

$$p_i = \frac{N_i}{N} \qquad i = \overline{1,n} \tag{6.2}$$

và lúc đó tổng thể còn có thể mô tả bằng bảng phân phối tần suất sau:

Giá trị của χ	x_1	x_2	...	x_i	...	x_k
Tần suất	p_1	p_2	...	p_i	...	p_k

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ (6.1) và (6.2) suy ra

$$\begin{cases} 0 \le p_i \le 1 & \forall_i \\ \sum_{i=1}^{k} p_i = 1 \end{cases}$$

Về mặt hình thức, bảng phân phối tần suất của tổng thể tương tự như bảng phân phối xác suất của biến ngẫu nhiên rời rạc. Nó phản ánh cơ cấu của tổng thể theo dấu hiệu χ.

3. Nếu ký hiệu $w_i (i = \overline{1,k})$ là tần số tích luỹ của x_i, tức là tổng số các phần tử có giá trị nhỏ hơn x_i thì.

$$w_i = \sum_{x_j < x_i} N_j \qquad (6.3)$$

và $F(x_i)$ $((i = \overline{1,k})$ là tần suất tích luỹ của x_i, tức là tỷ số giữa tần số tích luỹ của nó và kích thước của tổng thể thì:

$$F(x_i) = \frac{w_i}{N} = \sum_{x_j < x_i} \frac{N_j}{N} \qquad (6.4)$$

Tần suất tích luỹ là một hàm của x_i, có tính chất giống như hàm phân bố xác suất của biến ngẫu nhiên rời rạc.

Việc mô tả tổng thể theo dấu hiệu nghiên cứu χ bằng bảng phân phối tần số, bảng phân phối tần suất và tần suất tích luỹ cho thấy dấu hiệu định lượng χ hoàn toàn có thể mô hình hoá bằng một biến ngẫu nhiên rời rạc X. Điều này cũng đúng đối với các tổng thể mang dấu hiệu χ phân phối liên tục. Biến ngẫu nhiên X dùng để mô hình hoá dấu hiệu nghiên cứu χ được gọi là *biến ngẫu nhiên gốc*, còn quy luật phân phối xác suất của nó gọi là *quy luật phân phối gốc*. Việc mô hình hoá dấu hiệu nghiên cứu χ bằng biến ngẫu nhiên X cho phép áp dụng các công cụ xác suất đã xét ở phần trước trong việc nghiên cứu tổng thể.

3.3. Các tham số đặc trưng của tổng thể

Việc mô tả tổng thể theo dấu hiệu χ bằng bảng phân phối tần số, bảng phân phối tần suất và tần suất tích luỹ mới chỉ cung cấp những thông tin

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

chung nhất về tổng thể đó. Trong thực tế nhiều khi người nghiên cứu cần quan tâm đến những thông tin tổng hợp phản ánh những khía cạnh quan trọng nhất của tổng thể theo dấu hiệu nghiên cứu đó. Những thông tin này được biểu hiện qua các tham số đặc trưng chủ yếu sau đây của tổng thể.

1. Trung bình tổng thể

Giả sử trong tổng thể kích thước N dấu hiệu định lượng χ nhận các giá trị $x_1, x_2, ..., x_N$. Trung bình tổng thể, ký hiệu là m là trung bình số học của các giá trị của dấu hiệu trong tổng thể.

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (6.5)$$

Nếu trong tổng thể dấu hiệu χ chỉ nhận các giá trị $x_1, x_2, ..., x_k$ với các tần số tương ứng $N_1, N_2, ..., N_k$ thì trung bình tổng thể được xác định bằng biểu thức

$$m = \frac{1}{N}\sum_{i=1}^{k} x_i N_i \qquad (6.6)$$

Bản chất của trung bình tổng thể có thể làm rõ như sau: Giả sử tổng thể kích thước N bao gồm các phần tử mang các giá trị khác nhau của dấu hiệu nghiên cứu χ là $x_1, x_2, ..., x_N$. Giả sử từ tập hợp này lấy ngẫu nhiên ra một phần tử thì xác suất để lấy được phần tử mang giá trị x_i hiển nhiên là $\frac{1}{N}(i = \overline{1, N})$. Như vậy giá trị của dấu hiệu χ có thể xem như một biến ngẫu nhiên X với các giá trị có thể có $x_1, x_2, ..., x_N$ và các xác suất tương ứng đều bằng $\frac{1}{N}$. Từ đó

$$E(X) = x_1\frac{1}{N} + x_2\frac{1}{N} + ... + x_N\frac{1}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i = m$$

Như vậy m = E(X).

Mở rộng kết quả thu được cho tổng thể với dấu hiệu nghiên cứu liên tục ta thu được kết quả là nếu xem dấu hiệu nghiên cứu như biến ngẫu nhiên X thì trung bình tổng thể chính là kỳ vọng toán của biến ngẫu nhiên đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 1. Tổng thể nghiên cứu là một xí nghiệp có N = 40 công nhân với dấu hiệu nghiên cứu là năng suất lao động (sản phẩm/đơn vị thời gian). Số liệu của tổng thể theo dấu hiệu nghiên cứu được cho trong bảng sau (bảng 6.1).

Bảng 6.1

Năng suất lao động x_i	Số công nhân N_i	$N_i\ x_i$
50	3	150
55	5	275
60	10	600
65	12	780
70	7	490
75	3	225
	N = 40	$\sum N_i x_i = 2520$

Tìm năng suất lao động trung bình của mỗi công nhân.

Giải. Theo công thức (6.6) ta có:

$$m = \frac{1}{N}\sum_{i=1}^{k} x_i N_i = \frac{2520}{40} = 63$$

Ngoài trung bình tổng thể m (mà thực chất là trung bình số học) là loại trung bình được sử dụng nhiều nhất, trong thực tế, tuỳ thuộc và từng trường hợp người ta còn tính các loại trung bình sau.

a. Trung bình điều hoà. Tương tự như trung bình số học, trong đó thay cho các giá trị của dấu hiệu nghiên cứu người ta dùng giá trị nghịch đảo của chúng. Như vậy nếu ký hiệu trung bình điều hoà là m_h thì ta có công thức:

$$m_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \qquad (6.7)$$

Nếu dấu hiệu của tổng thể nhận các giá trị $x_1, x_2, ..., x_k$ với tần số tương ứng $N_1, N_2, ..., N_k$ thì:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$m_h = \frac{N}{\sum_{i=1}^{k} \frac{N_i}{x_i}} \qquad (6.8)$$

Thí dụ 2. Một xí nghiệp có hai phân xưởng cùng lắp ráp một loại sản phẩm. Phân xưởng thứ nhất lắp ráp một sản phẩm hết 15 phút, phân xưởng thứ hai hết 20 phút. Nếu trong một ngày mỗi phân xưởng làm việc 8 giờ thì hãy tìm thời gian trung bình để lắp ráp 1 sản phẩm.

Giải. Trong trường hợp này muốn tính thời gian trung bình để lắp ráp 1 sản phẩm không thể dùng công thức trung bình số học vì thời gian trung bình để lắp ráp 1 sản phẩm phải bằng tổng số thời gian sản xuất chia cho tổng số sản phẩm được lắp ráp trong thời gian đó. Vì vậy để tính thời gian trung bình phải tìm được tổng số sản phẩm mà hai phân xưởng lắp ráp được trong một ngày làm việc. Theo công thức (6.8) ta có:

$$m_h = \frac{N}{\sum_{i=1}^{k} \frac{N_i}{x_i}} = \frac{60x8 + 60x8}{\frac{60x8}{15} + \frac{60x8}{20}} = \frac{960}{56} = 17,14 \text{ phút}$$

Chú ý rằng cũng giống như trung bình số học, trung bình điều hoà chỉ được áp dụng khi các giá trị của dấu hiệu nghiên cứu có quan hệ tổng.

b. Trung bình nhân: Là căn bậc N của tích các giá trị của dấu hiệu trong tổng thể. Vậy nếu ký hiệu trung bình nhân là m_g thì

$$m_g = \sqrt[N]{x_1.x_2...x_N} \qquad (6.9)$$

hoặc

$$m_g = \sqrt[N]{x_1^{N_1}.x_2^{N_2}...x_k^{N_k}} \qquad (6.10)$$

nếu dấu hiệu chỉ nhận các giá trị $x_1,...,x_k$ với các tần số tương ứng $N_1,...,N_k$.

Thí dụ 3. Trong khoảng thời gian 10 năm, tốc độ tăng giá trị sản lượng của một xí nghiệp như sau: Có 5 năm tốc độ tăng so với năm trước là 110%; có 2 năm tốc độ tăng là 125% và có 3 năm tốc độ tăng là 115%. Tìm tốc độ tăng trưởng trung bình hàng năm của xí nghiệp trong 10 năm đó.

Giải. Ở đây tốc độ tăng trưởng giá trị sản lượng không có quan hệ tổng (vì gốc so sánh khác nhau) mà có quan hệ tích. Do đó để tính tốc độ tăng trưởng bình quân hàng năm so với năm trước đó ta phải tính số trung bình nhân. Theo công thức (6.10) ta có:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$m_g = \sqrt[N]{x_1^{N_1}.x_2^{N_2}...x_k^{N_k}} = \sqrt[10]{(1,1)^5.(1,25)^2.(1,15)^3}$$

$$\rightarrow \lg m_g = \frac{1}{10}(5.\lg 1,1 + 2.\lg 1,25 + 3.\lg 1,15) = 0,0583$$

$$\rightarrow m_g = 1,144 \text{ hay } 114,4\%.$$

Trong kinh tế và xã hội, trung bình nhân thường chỉ dùng để tính tốc độ tăng trưởng bình quân.

2. Phương sai tổng thể

Phương sai tổng thể, ký hiệu là σ^2, là trung bình số học của bình phương các sai lệch giữa các giá trị của dấu hiệu trong tổng thể và trung bình tổng thể.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2 \qquad (6.11)$$

Nếu các giá trị $x_1, x_2, ..., x_k$ của dấu hiệu có các tần số tương ứng là $N_1, N_2, ..., N_k$ với $N_1 + N_2 + ... + N_k = N$ thì

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}N_i(x_i - m)^2 \qquad (6.12)$$

Do có thể viết

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}N_i(x_i - m)^2 = \sum_{i=1}^{k}\frac{N_i}{N}(x_i - m)^2$$

$$= \sum_{i=1}^{k}(x_i - m)^2 p_i$$

nên về thực chất phương sai tổng thể chính là phương sai của biến ngẫu nhiên trong tổng thể đó. Nó phản ánh mức độ phân tán của các giá trị của dấu hiệu χ xung quanh giá trị trung bình tổng thể.

Trong thực tế, để tiện cho việc tính toán, phương sai tổng thể thường được tính bằng công thức:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}N_i x_i^2 - m^2 \qquad (6.13)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thật vậy theo định nghĩa

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} N_i(x_i - m)^2 = \frac{1}{N}\sum_{i=1}^{k} N_i(x_i^2 - 2mx_i + m^2) =$$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i x_i^2 - 2m.\frac{1}{N}\sum_{i=1}^{k} N_i x_i + \frac{1}{N}\sum_{i=1}^{k} N_i m^2$$

song do $\frac{1}{N}\sum_{i=1}^{k} N_i x_i = m$

và $$\frac{1}{N}\sum_{i=1}^{k} N_i m^2 = \frac{Nm^2}{N} = m^2$$

nên $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} N_i x_i^2 - 2m^2 + m^2 = \frac{1}{N}\sum_{i=1}^{k} N_i x_i^2 - m^2$$

Nếu lấy căn bậc hai của phương sai ta sẽ thu được *độ lệch chuẩn tổng thể*:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2}$$

Thí dụ 4. Với các số liệu cho trong bảng (6.1) hãy tìm phương sai và độ lệch chuẩn của năng suất lao động của công nhân xí nghiệp đó.

Giải. Theo các số liệu của bảng (6.1) ta tìm ngay được

$$\frac{1}{N}\sum_{i=1}^{k} N_i x_i^2 = 160500$$

Theo công thức (6.13)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} N_i x_i^2 - m^2 = \frac{160500}{40} - 63^2 = 43,5$$

$$\sigma = 6,6$$

3. Một loại tổng thể thường được nghiên cứu trong thực tế là tổng thể kích thước N, trong đó M phần tử mang dấu hiệu nghiên cứu, còn N - M phần tử còn lại không mang dấu hiệu đó. Lúc đó *tần suất của tổng thể* là tỷ số giữa số phần tử mang dấu hiệu nghiên cứu và kích thước của tổng thể:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$p = \frac{M}{N} \qquad (6.14)$$

Ta thấy rằng p chính là xác suất để lấy ngẫu nhiên một phần tử thì phần tử đó mang dấu hiệu nghiên cứu. Như vậy, nếu xem dấu hiệu nghiên cứu là biến ngẫu nhiên X phân phối theo quy luật không - một thì p chính là kỳ vọng toán, tức là thực chất tần suất p là trường hợp riêng của trung bình tổng thể m và phản ánh cơ cấu của tổng thể theo dấu hiệu nghiên cứu χ.

Vì có thể đặc trưng dấu hiệu nghiên cứu trong tổng thể bằng một biến ngẫu nhiên X, vì thế các tham số đặc trưng khác của X như Mốt, Trung vị, Hệ số biến thiên, Hệ số bất đối xứng, Hệ số nhọn v.v... cũng đều là các tham số đặc trưng của tổng thể. Biểu thức tính các tham số này giống như đã trình bày ở mục 4 chương 2.

Bài tập

6.1. Một công ty có 25 nhân viên. Lương tháng của họ được cho trong bảng phân phối tần số sau đây (đơn vị triệu đồng):

Lương tháng (x_i)	3	3,5	3,8	4,4	4,5
Số nhân viên (n_i)	2	6	9	7	1

a. Hãy tìm các tham số đặc trưng xu hướng trung tâm là trung bình, mốt, trung vị.

b. Hãy tìm các tham số đặc trưng độ phân tán là phương sai, độ lệch chuẩn, hệ số biến thiên.

ĐS: a. $m = 3,86$; $m_0 = 3,8$; $m_d = 3,8$

b. $\sigma^2 = 0,1896$; $\sigma = 0,4354$; $CV = 11,28\%$

6.2. Xí nghiệp có 50 công nhân. Thời gian hoàn thành một sản phẩm của họ được cho trong bảng phân phối tần số sau (đơn vị: phút)

Thời gian (x_i)	12-14	14-16	16-18	18-20	20-22
Số công nhân (n_i)	1	4	10	14	12
	22-24	24-26	26-28		
	6	2	1		

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Tìm trung bình số học

b. Tìm phương sai, độ lệch chuẩn, hệ số biến thiên.

ĐS: a. m = 19,52

b. $\sigma^2 = 8,1296$; $\sigma = 2,8512$; CV = 14,61%

6.3. Thống kê dân số của một nước năm 2010 theo học vấn và giới tính cho kết quả sau đây.

Học vấn Giới tính	Thất học	Tiểu học	Trung học	Đại học
Nam	0,04	0,10	0,23	0,10
Nữ	0,05	0,12	0,29	0,07

a. Tìm tỷ lệ nam và nữ trong cơ cấu dân số của nước đó vào năm 2010.

b. Tìm tỷ lệ dân số của nước đó theo các trình độ học vấn vào năm 2010.

ĐS: a. $P_{nam} = 0,47$; $P_{nữ} = 0,53$

b. $P_{Thất\ học} = 0,09$

$P_{Tiểu\ học} = 0,22$

$P_{Trung\ học} = 0,52$

$P_{Đại\ học} = 0,17$

c. 0,36

6.4. Có hai công nhân cùng làm việc mỗi ngày 8 giờ để sản xuất một sản phẩm. Thời gian làm xong một sản phẩm của họ tương ứng là 2 phút và 6 phút. Dùng công thức trung bình điều hoà, hãy tính thời gian trung bình để sản xuất 1 sản phẩm của 2 công nhân đó.

ĐS: 3 phút

6.5. Tốc độ tăng trưởng của một ngành sản xuất như sau:

- Năm 2005 so với 2004 bằng 116%
- Năm 2006 so với 2005 bằng 111%
- Năm 2007 so với 2006 bằng 112%
- Năm 2008 so với 2007 bằng 113%
- Năm 2009 so với 2008 bằng 112%

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Năm 2010 so với 2009 bằng 111%

Hãy dùng công thức trung bình nhân để tính tốc độ tăng trưởng trung bình của ngành đó từ 2004 đến 2010.

ĐS: 1,125

4. MẪU NGẪU NHIÊN

4.1. Định nghĩa mẫu ngẫu nhiên

Các tham số đặc trưng của tổng thể có thể xác định được một cách trực tiếp nếu áp dụng phương pháp nghiên cứu toàn bộ tổng thể. Song do những hạn chế như đã xét ở mục 1, chẳng hạn quy mô quá lớn của tổng thể hay mức độ kém tin cậy của số liệu điều tra nên việc tính toán vừa khó khăn, tốn kém mà vẫn không thu được kết quả chính xác. Đặc biệt, khi không thể nắm được kích thước của tổng thể (và phải coi N là vô hạn) thì thực tế là không thể nghiên cứu trực tiếp tổng thể được. Vì vậy, thường người ta áp dụng phương pháp mẫu bằng cách chọn ra từ tổng thể n phần tử và chỉ tập trung nghiên cứu các phần tử đó mà thôi. Tập hợp n phần tử này được gọi là *mẫu kích thước n.*

Để có thể căn cứ vào thông tin của mẫu đưa ra những kết luận đủ chính xác về dấu hiệu nghiên cứu trong tổng thể thì trước hết mẫu được chọn phải mang tính đại diện cho tổng thể, tức là phản ánh đúng đặc điểm của tổng thể theo dấu hiệu nghiên cứu đó. Để đảm bảo tính đại diện của mẫu và tiện cho việc mô hình hoá, mẫu được tạo lập với những giả thiết sau:

- Lấy lần lượt từng phần tử vào mẫu. Phương pháp này gọi là phương pháp đơn giản để phân biệt với cách lấy cùng một lúc nhiều phần tử vào mẫu.

- Mỗi phần tử được lấy vào mẫu một cách hoàn toàn ngẫu nhiên, tức là mọi phần tử của tổng thể đều được lấy vào mẫu với khả năng như nhau.

- Các phần tử được lấy vào mẫu theo phương thức hoàn lại, tức là trước khi lấy phần tử thứ k thì trả lại tổng thể phần tử thứ (k-1) mà ta đã nghiên cứu xong $(k = \overline{2,n})$.

Trong thực tế nếu kích thước của tổng thể khá lớn còn mẫu chỉ chiếm một phần rất nhỏ của tổng thể thì phương thức lấy mẫu hoàn lại và không hoàn lại cho ta các kết quả sai lệch không đáng kể. Đặc biệt khi kích thước

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

của tổng thể là vô hạn còn kích thước của mẫu lại là hữu hạn thì không còn sự khác biệt giữa hai phương thức lấy mẫu nói trên nữa. Lúc đó có thể chọn mẫu theo phương thức không hoàn lại và vẫn có thể giả thiết mẫu được chọn theo phương thức hoàn lại.

Giả sử theo một phương thức nào đó từ tổng thể lấy ra n phần tử tạo nên mẫu kích thước n. Vì mẫu được lấy ra theo nguyên tắc đơn giản, ngẫu nhiên và hoàn lại nên có thể mô hình hoá mẫu được chọn như sau:

Gọi $X_i\ (i=\overline{1,n})$ là giá trị của dấu hiệu χ đo lường được trên phần tử thứ i của mẫu. Vì có thể mô hình hoá dấu hiệu χ bằng một biến ngẫu nhiên X với một quy luật phân phối xác suất nào đó nên việc chọn mẫu kích thước n theo nguyên tắc trên có thể xem như tiến hành n phép thử độc lập đối với X và lúc đó các giá trị X_i của dấu hiệu thu được trên mẫu có thể xem như các biến ngẫu nhiên thu được qua việc tiến hành n phép thử độc lập đối với biến ngẫu nhiên X. Từ đó ta có định nghĩa sau:

Định nghĩa. Mẫu ngẫu nhiên kích thước n là tập hợp của n biến ngẫu nhiên độc lập $X_1, X_2, ..., X_n$ được thành lập từ biến ngẫu nhiên X trong tổng thể nghiên cứu và có cùng quy luật phân phối xác suất với X.

Mẫu ngẫu nhiên thường được ký hiệu là:

$$W = (X_1, X_2, ..., X_n)$$

Chú ý rằng với cách xây dựng mẫu ngẫu nhiên như vậy thì các biến ngẫu nhiên $X_1, X_2, ..., X_n$ của mẫu không những có cùng dạng phân phối xác suất với biến ngẫu nhiên gốc X, tức là có cùng hàm phân bố xác suất F(x) mà các tham số đặc trưng của chúng cũng bằng các tham số đặc trưng của X, tức là:

$$E(X_1) = E(X_2) = ... = E(X_n) = E(X) = m \qquad (6.15)$$

$$V(X_1) = V(X_2) = ... = V(X_n) = V(X) = \sigma^2 \qquad (6.16)$$

Chẳng hạn gọi X là số chấm thu được khi gieo một con xúc xắc, X là biến ngẫu nhiên với bảng phân phối xác suất như sau:

X	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nếu gieo con xúc xắc 3 lần và gọi $X_i (i = \overline{1,3})$ là số chấm xuất hiện ở lần gieo thứ i thì ta có 3 biến ngẫu nhiên độc lập tạo nên mẫu ngẫu nhiên kích thước n = 3.

$$W = (X_1, X_2, X_3)$$

Hơn nữa, mỗi biến ngẫu nhiên X_i trong mẫu đều có bảng phân phối xác suất giống như bảng phân phối xác suất của biến ngẫu nhiên X, do đó:

$$E(X_i) = \frac{1}{6}(1+2+...+6) = \frac{21}{6} = E(X) \quad \forall i$$

$$V(X_i) = \frac{1}{6}(1^2+2^2+...+6^2) - \left(\frac{21}{6}\right)^2 = \frac{35}{12} = V(X) \quad \forall i$$

Lúc đó việc thực hiện một phép thử đối với mẫu ngẫu nhiên W chính là thực hiện một phép thử đối với mỗi thành phần của mẫu. Giả sử X_1 nhận giá trị x_1, X_2 nhận giá trị x_2,..., X_n nhận giá trị x_n. Tập hợp n giá trị $x_1, x_2, ..., x_n$ tạo thành một giá trị của mẫu ngẫu nhiên, hay còn gọi là một mẫu cụ thể, ký hiệu:

$$w = (x_1, x_2, ..., x_n)$$

Chẳng hạn, trong thí dụ trên nếu tiến hành một phép thử đối với mẫu ngẫu nhiên bằng cách tung cụ thể ba lần, lần đầu được 6 chấm, lần thứ hai được 5 chấm, lần thứ ba được 1 chấm thì ta thu được một mẫu cụ thể $w = (6,5,1)$. Nếu thực hiện một phép thử khác đối với W lại thu được một mẫu cụ thể khác, chẳng hạn $w = (2,1,4)$...

Như vậy, mẫu ngẫu nhiên là tập hợp của n biến ngẫu nhiên, còn mẫu cụ thể lại là tập hợp của n giá trị cụ thể quan sát được khi thực hiện một phép thử đối với mẫu ngẫu nhiên.

4.2. Các phương pháp chọn mẫu

Như đã trình bày ở trên, với những cách thức tiến hành phép thử khác nhau, ta sẽ thu được những mẫu cụ thể khác nhau từ cùng một mẫu ngẫu nhiên song phải đảm bảo yêu cầu là mẫu phải đại diện cho tổng thể nghiên cứu.

Tuỳ thuộc vào đặc điểm của từng tổng thể nghiên cứu mà mẫu có thể

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

được chọn theo nhiều phương pháp khác nhau để đảm bảo yêu cầu về tính đại diện của mẫu. Sau đây là một số phương pháp chọn mẫu chủ yếu thường được sử dụng để nghiên cứu các tổng thể kinh tế - xã hội.

1. Mẫu ngẫu nhiên giản đơn: Là loại mẫu được chọn trực tiếp từ danh sách đã đánh số của tổng thể. Từ tổng thể kích thước N người ta dùng cách rút thăm đơn giản ra n phần tử của mẫu theo một bảng số ngẫu nhiên nào đó.

Các bảng số ngẫu nhiên có thể sử dụng là:

- Các bảng của Tippett gồm các số có 4 chữ số;
- Các bảng của Fisher và Yates;
- Các bảng của Kendall và Babington Smith gồm các số có 5 chữ số;
- Các bảng của Burke Haton;
- Các bảng của công ty Rand... (xem phụ lục 10)

Phương pháp này có ưu điểm là cho phép thu được một mẫu có tính đại diện cao, cho phép suy rộng các kết quả của mẫu cho tổng thể với một sai số xác định, song để vận dụng phải có được toàn bộ danh sách của tổng thể nghiên cứu. Mặt khác chi phí chọn mẫu sẽ khá lớn.

2. Mẫu ngẫu nhiên hệ thống: Là loại mẫu đã được đơn giản hoá trong cách chọn, trong đó chỉ có phần tử đầu tiên được chọn một cách ngẫu nhiên, sau đó dựa trên danh sách đã được đánh số của tổng thể để chọn ra các phần tử tiếp theo vào mẫu theo một thủ tục nào đó. Chẳng hạn trên một danh sách N khách hàng cần chọn ra một mẫu kích thước n thì chia danh sách đó ra n phần bằng nhau, ở phần thứ nhất gồm $\frac{N}{n}$ phần tử chọn ngẫu nhiên ra một phần tử, sau đó theo danh sách cứ cách $\frac{N}{n}$ phần tử ta lấy ra một phần tử vào mẫu cho đến khi đủ n phần tử.

Nhược điểm chính của phương pháp này là dễ mắc sai số hệ thống khi danh sách của tổng thể không được sắp xếp một cách ngẫu nhiên mà lại theo một trật tự chủ quan nào đó. Tuy vậy, do cách thức đơn giản của nó, mẫu ngẫu nhiên hệ thống hay được dùng ở cấp chọn mẫu cuối cùng và khi tổng thể tương đối thuần nhất.

3. Mẫu chùm: Trong một số trường hợp, để tiện cho việc nghiên cứu người ta muốn quy diện nghiên cứu gọn về một khu vực nhất định chứ không để cho các phần tử của mẫu phân tán quá rộng, chẳng hạn tập trung

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

nghiên cứu khách hàng tại một địa phương nào đó. Lúc đó mẫu được chọn theo chùm. Chẳng hạn chùm có thể là một hộ gia đình có nhiều người, một làng có nhiều hộ gia đình... Theo phương pháp này, trước tiên tổng thể điều tra được phân chia thành nhiều chùm theo nguyên tắc:

- Mỗi phần tử của tổng thể chỉ được phân vào một chùm.

- Mỗi chùm cố gắng chứa nhiều phần tử khác nhau về dấu hiệu nghiên cứu, sao cho nó có độ phân tán cao như của tổng thể.

- Phân chia sao cho các chùm tương đối đồng đều nhau về quy mô.

Các chùm được chọn một cách ngẫu nhiên và tất cả các phần tử của chùm đó đều được chọn vào mẫu.

Trong phương pháp chọn mẫu chùm do sự đồng đều cao hơn giữa các chùm nên dễ dẫn đến khả năng lặp lại thông tin. Như vậy sai số chọn mẫu có thể cao hơn phương pháp chọn ngẫu nhiên đơn với cùng kích thước mẫu, song nó vẫn được sử dụng cho đỡ tốn kém chi phí và thích hợp với việc nghiên cứu theo nhiều dấu hiệu cùng một lúc.

4. Mẫu phân tổ: Trong chọn mẫu phân tổ, trước hết người ta phân chia tổng thể ra thành các tổ có độ thuần nhất cao để chọn ra các phần tử đại diện cho từng tổ. Việc phân tổ có hiệu quả khi tổng thể nghiên cứu không thuần nhất theo dấu hiệu nghiên cứu. Sau khi đã phân tổ thì kích thước mẫu được phân bổ cho mỗi tổ theo một quy tắc nào đó, chẳng hạn tỷ lệ thuận với kích thước mỗi tổ.

5. Mẫu nhiều cấp: Nếu các phần tử của tổng thể phân tán quá rộng và thiếu thông tin về chúng, người ta thường chọn mẫu theo nhiều cấp. Khi chọn nhiều cấp, ta có nhiều loại đơn vị chọn mẫu ở mỗi cấp, thường được gọi là đơn vị chọn mẫu cấp 1, cấp 2,... Để chọn mẫu ở mỗi cấp chỉ cần có thông tin về phân bố của dấu hiệu ở cấp ấy là đủ. Chẳng hạn, để điều tra ý kiến của tổng thể khách hàng trong cả nước về sản phẩm của doanh nghiệp có thể chọn mẫu nhiều cấp như sau:

- Đơn vị mẫu cấp 1: Chọn ra các tỉnh, thành phố đại diện.

- Đơn vị mẫu cấp 2: Trong các tỉnh, thành phố đã chọn, chọn ra một số quận, huyện đại diện.

- Đơn vị mẫu cấp 3: Trong các quận, huyện đã chọn, chọn ra một số phường, xã đại diện...

Việc chọn mẫu ở mỗi cấp có thể tiến hành theo phương pháp mẫu ngẫu nhiên đơn, mẫu ngẫu nhiên hệ thống, mẫu chùm hay mẫu phân tổ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.3. Thang đo các giá trị mẫu

Biến ngẫu nhiên trong tổng thể nghiên cứu có thể là định tính hoặc định lượng, do đó mẫu rút ra từ tổng thể cũng gồm các giá trị định tính hoặc định lượng. Vì vậy, để biểu diễn các giá trị của dấu hiệu nghiên cứu trong tổng thể cũng như của mẫu phải dùng các thang đo khác nhau nhằm mục đích lượng hoá dấu hiệu nghiên cứu đó.

Trong kinh tế xã hội các thang đo được sử dụng là:

1. Thang định danh: Là việc đánh số những tính chất hoặc phạm trù cùng loại. Chẳng hạn, thang giới tính gồm hai phạm trù là nam [0] và nữ [1]. Màu sắc sản phẩm có thể có nhiều hơn hai phạm trù là xanh [1] đỏ [2] trắng [3] vàng [4]... Giữa các con số ở đây không có quan hệ hơn kém do đó không thể thực hiện các phép tính số học đối với chúng. Thang định danh thường chỉ dùng để đếm tần số của các hiện tượng xảy ra.

2. Thang thứ bậc: Là loại thang định danh mà giữa các phạm trù đã có quan hệ thứ bậc hơn kém. Chẳng hạn, để đặc trưng học vấn có thể dùng thang thứ bậc thất học [0], tiểu học [1], trung học [2], đại học trở lên [3]; để đặc trưng thái độ của khách hàng đối với giá sản phẩm có thể dùng thang đo: rẻ [1], vừa phải [2], đắt [3]. Đương nhiên sự sai khác giữa các phạm trù không bắt buộc phải đều nhau. Xét về mặt toán học, tập hợp các phạm trù đó đã được sắp xếp nhưng chưa có một metric.

3. Thang đo khoảng: Là thang đo thứ bậc có các khoảng cách đều nhau giữa các bậc. Một thang đo như vậy đã có kết cấu metric, có thể đánh giá sự khác biệt giữa các phạm trù bằng loại thang đo này mặc dù điểm gốc ở đây chỉ là tương đối. Với thang đo khoảng việc cộng và trừ các số đo mới bắt đầu có ý nghĩa, trên cơ sở đó có thể tính được các tham số đặc trưng như trung bình, phương sai v.v... Yêu cầu có khoảng cách đều nhau là đặt ra đối với thang đo, còn đối với các lớp hiện tượng được đo bằng thang này thì không bắt buộc phải đều nhau. Chẳng hạn, để đặc trưng lứa tuổi có thể dùng thang đo khoảng: trẻ (dưới 35 tuổi) [30], trung niên (từ 36 tuổi đến 60 tuổi) [50], già (từ 60 tuổi trở lên) [70]. Để thu được thang đo khoảng có thể bắt đầu bằng thang thứ bậc sau đó chuẩn hoá sao cho các quãng cách đều nhau để được một thang đo khoảng, sao cho việc tính toán các trị số đo trở nên có ý nghĩa.

Các thang đo định danh, thứ bậc và thang đo khoảng dùng để đặc trưng các giá trị của dấu hiệu nghiên cứu định tính.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4. Thang đo tỉ lệ: Là thang đo khoảng với một điểm gốc tuyệt đối. Chỉ với thang đo tỷ lệ ta mới có thể đo lường các hiện tượng như các đơn vị đo lường vật lý thông thường và mới có thể thực hiện được tất cả các phép toán với các trị số đo theo nghĩa là lượng thông tin sẽ tăng lên cùng với trị số đo.

Thang đo tỷ lệ được dùng để đặc trưng các giá trị của dấu hiệu nghiên cứu định lượng.

Mỗi thang đo cấp cao hơn có thể chuyển xuống một thang đo cấp thấp hơn, chẳng hạn chuyển thang đo tỷ lệ thành thang đo thứ bậc nhưng ngược lại thì không được.

Khi nghiên cứu một dấu hiệu của tổng thể hay rút ra một mẫu từ tổng thể thì việc đầu tiên là phải xác định đúng thang đo cho các giá trị điều tra của tổng thể hay của mẫu làm cơ sở cho quá trình xử lý tiếp theo.

4.4. Các phương pháp mô tả số liệu mẫu

1. Giả sử từ tổng thể với biến ngẫu nhiên gốc X rút ra một mẫu cụ thể kích thước n, trong đó giá trị x_1 xuất hiện với tần số n_1, x_2 xuất hiện với tần số n_2,..., x_k xuất hiện với tần số n_k (trong đó các giá trị của X có thể được đo bằng các thang khác nhau tuỳ thuộc vào việc χ là định tính hay định lượng). Lúc đó, sau khi các x_i đã được sắp xếp theo trình tự tăng dần, giá trị cụ thể của mẫu w có thể mô tả bằng bảng phân phối tần số thực nghiệm sau:

x_i	x_1	x_2	...	x_i	...	x_k
n_i	n_1	n_2	...	n_i	...	n_k

Hiển nhiên là $n_1 + n_2 + ... + n_k = n$

2. Nếu ký hiệu

$$f_i = \frac{n_i}{n} \qquad (6.17)$$

là tần suất xuất hiện giá trị x_i trong mẫu thì lúc đó giá trị của mẫu w còn có thể mô tả bằng bảng phân phối tần suất thực nghiệm sau:

x_i	x_1	x_2	...	x_i	...	x_k
f_i	f_1	f_2	...	f_i	...	f_k

Từ (6.17) suy ra $f_1 + f_2 + ... + f_k = 1$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 1. Gặt ngẫu nhiên 365 điểm trồng lúa của một huyện thu được các số liệu được sắp xếp thành bảng sau:

Năng suất (tạ/ha)	25	30	33	34	35	36	37	39	40
Số điểm gặt tương ứng	6	13	38	74	106	85	30	10	3

Như vậy, giá trị của mẫu đã được mô tả dưới dạng bảng phân phối tần số thực nghiệm. Còn bảng phân phối tần suất thực nghiệm có dạng:

x_i	25	30	33	34	35	36	37	39	40
f_i	0,016	0,036	0,104	0,203	0,290	0,233	0,082	0,027	0,009

3. Nếu ký hiệu $\omega_i (i = \overline{1,k})$ là tần số tích luỹ của x_i

$$\omega_i = \sum_{x_j < x_i} n_j \tag{6.18}$$

và $F^*(x_i)$ là tần suất tích luỹ của x_i

$$F^*(x_i) = \frac{\omega_i}{n} = \sum_{x_j < x_i} \frac{n_j}{n} \tag{6.19}$$

thì $F^*(x_i)$ là một hàm của x_i và gọi là hàm phân bố thực nghiệm của mẫu.

Khác với hàm phân bố xác suất F(x) của biến ngẫu nhiên gốc X trong tổng thể xác định xác suất của biến cố $X < x$, hàm phân bố thực nghiệm $F^*(x)$ xác định tần suất của biến cố đó. Theo định lý Bernoulli, khi kích thước mẫu đủ lớn tần suất của biến cố $X < x$ tức là $F^*(x)$ sẽ hội tụ theo xác suất về xác suất F(x) của biến cố đó, tức là $F^*(x)$ và F(x) sai khác nhau không đáng kể. Từ đó có thể dùng hàm phân bố thực nghiệm của mẫu để biểu diễn một cách gần đúng quy luật phân phối gốc F(x) của tổng thể.

4. Để mô tả số liệu mẫu một cách rõ ràng hơn cho phép đưa ra những nhận xét sơ bộ ban đầu về tổng thể, người ta còn xây dựng các loại đồ thị khác nhau của phân phối thực nghiệm.

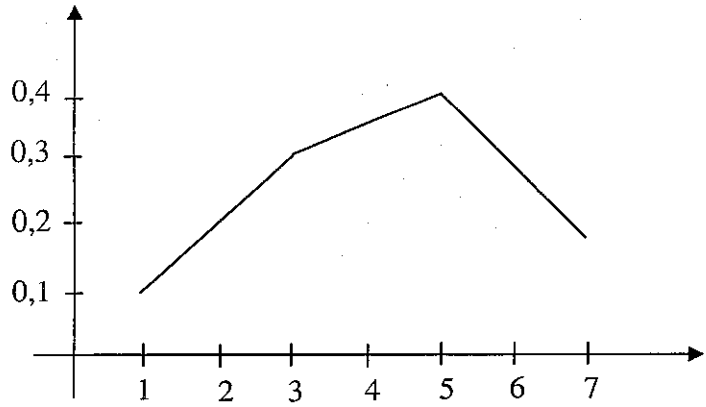
- *Đa giác tần số* là một đường gãy khúc mà các đoạn thẳng của nó nối các điểm $(x_1, n_1), (x_2, n_2) \ldots (x_k, n_k)$ trên mặt phẳng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- *Đa giác tần suất* là một đường gãy khúc mà các đoạn thẳng của nó nối các điểm $(x_1, f_1), (x_2, f_2) ... (x_k, f_k)$ trên mặt phẳng.

Thí dụ 2. Vẽ đa giác tần suất của phân phối thực nghiệm sau:

x_i	1	3	5	7
f_i	0,1	0,3	0,4	0,2



***Hình 6.1.* Đa giác tần suất**

Đa giác tần suất thường được dùng để mô tả các số liệu mẫu theo thời gian.

- Khi dấu hiệu nghiên cứu có phân phối liên tục thì nên xây dựng *biểu đồ tần số* hoặc *biểu đồ tần suất*. Để làm điều đó khoảng chứa tất cả các giá trị quan sát được của mẫu được chia ra thành một số đoạn có chiều dài bằng h và tại mỗi đoạn đưa vào các tần số hoặc tần suất tương ứng với đoạn đó. Như vậy, biểu đồ tần số sẽ là một hình bậc thang tạo nên bởi nhiều hình chữ nhật có đáy bằng h và chiều cao bằng $\frac{n_i}{h}$. Lúc đó diện tích của hình chữ nhật thứ i bằng $h.\frac{n_i}{h} = n_i$ là tổng tần số tương ứng với đoạn thứ i, vì vậy diện tích của tất cả các hình chữ nhật sẽ bằng kích thước mẫu n.

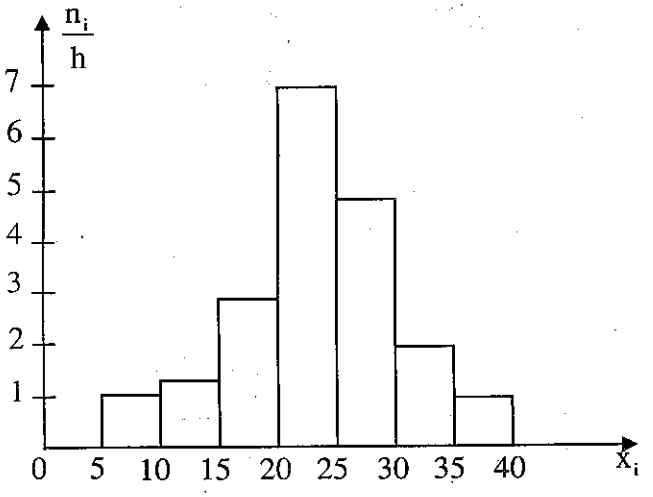
Tương tự biểu đồ tần suất là một hình bậc thang tạo nên bởi nhiều hình chữ nhật có đáy bằng h và chiều cao bằng $\frac{f_i}{h}$. Lúc đó diện tích của hình chữ nhật thứ i bằng $h.\frac{f_i}{h} = f_i$ và diện tích toàn bộ hình bậc thang đó sẽ bằng 1.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 3. Vẽ biểu đồ tần số của phân phối thực nghiệm cho ở bảng sau:

Bảng 6.2

Đoạn giá trị chiều dài h = 5	Tổng các tần số tương ứng n_i	$\frac{n_i}{h}$
5 - 10	4	0,8
10 - 15	6	1,2
15 - 20	16	3,2
20 - 25	36	7,2
25 - 30	24	4,8
30 - 35	10	2,0
35 - 40	4	0,8



***Hình 6.2.* Biểu đồ tần số**

Với các dấu hiệu nghiên cứu là định tính thì người ta thường mô tả các số liệu mẫu bằng biểu đồ hình bánh xe. Đó là một hình tròn được chia ra thành nhiều bộ phận tương ứng với cơ cấu của phạm trù xuất hiện trong mẫu.

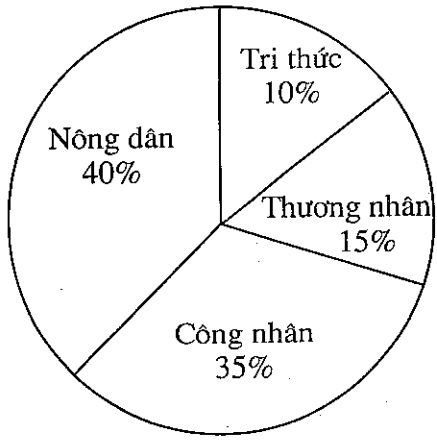
Thí dụ 4. Điều tra ngẫu nhiên 100 khách hàng của doanh nghiệp thì thấy các khách hàng được phân theo tỉ lệ sau đây về tầng lớp xã hội (bảng 6.3).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bảng 6.3

Tầng lớp xã hội	Số khách hàng	Tỷ lệ
Công nhân	35	0,35
Nông dân	40	0,40
Thương nhân	15	0,15
Trí thức	10	0,10
Tổng số	100	1

Vẽ biểu đồ hình bánh xe về cơ cấu của 100 khách hàng được điều tra (hình 6.3).



***Hình 6.3.* Biểu đồ hình bánh xe**

Đồ thị của phân phối mẫu có thể vẽ dễ dàng nhờ việc sử dụng các phần mềm cho máy vi tính. Mọi phần mềm thống kê như Excel, SPSS, MFIT, Stata đều có chương trình cho phép vẽ các loại đồ thị trên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ A. Từ ba vùng người ta điều tra ngẫu nhiên thu nhập hàng năm (tính bằng USD) của 300 người (mỗi vùng 100 người) đang làm việc tại các công ty tư nhân và thu được các số liệu sau:

Vùng 1		Vùng 2		Vùng 3	
Thu nhập	Số người	Thu nhập	Số người	Thu nhập	Số người
1547	1	1557	1	1510	1
1553	1	1563	1	1525	1
1573	2	1576	1	1527	1
1575	1	1582	1	1530	1
1580	1	1586	2	1544	1
1582	1	1588	1	1547	1
1586	1	1590	2	1550	1
1592	1	1595	1	1560	1
1595	3	1596	1	1573	2
1597	4	1597	1	1578	1
1598	1	1598	2	1582	2
1599	2	1600	1	1583	2
1602	1	1603	1	1584	1
1607	2	1605	2	1585	1
1608	2	1610	2	1586	1
1610	1	1612	2	1588	2
1617	1	1613	1	1590	2
1620	3	1620	1	1592	1
1621	1	1624	1	1593	2

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

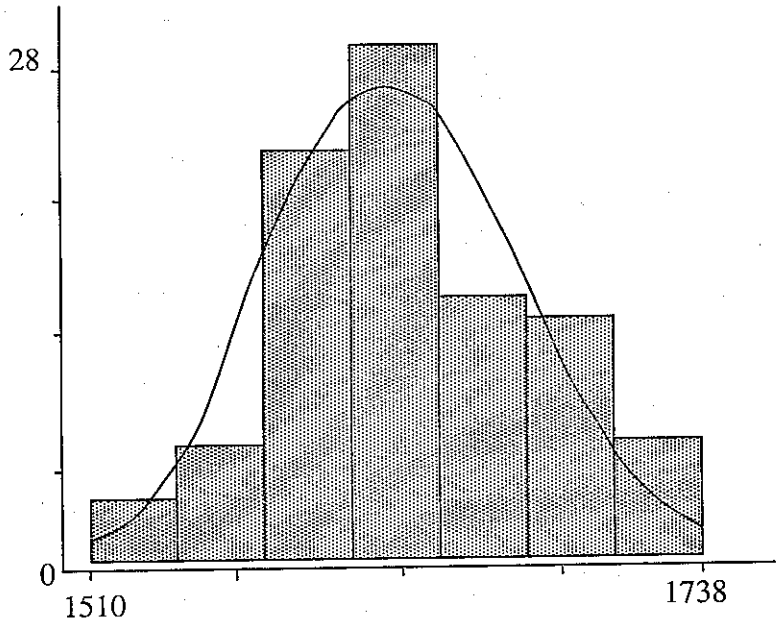
Vùng 1		Vùng 2		Vùng 3	
Thu nhập	Số người	Thu nhập	Số người	Thu nhập	Số người
1623	2	1625	2	1594	1
1624	1	1628	3	1597	2
1625	1	1629	1	1601	1
1626	1	1630	1	1602	2
1627	2	1632	1	1603	2
1630	3	1633	1	1607	1
1632	1	1637	2	1608	1
1633	1	1638	1	1610	3
1636	1	1640	2	1613	1
1637	3	1642	1	1614	1
1638	1	1643	1	1615	1
1640	2	1645	1	1616	2
1642	1	1648	2	1620	3
1644	1	1650	2	1622	1
1645	1	1652	1	1623	3
		1656	1	1625	4
1654	1	1657	2	1627	1
1655	3	1658	4	1630	1
1658	1	1660	2	1632	3
1660	1	1662	3	1633	2
1662	1	1663	1	1640	1
1667	1	1665	1	1642	2
1670	3	1666	2	1645	2
1672	1	1670	3	1648	1
1674	1	1672	1	1650	4
1675	2	1673	1	1652	1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vùng 1		Vùng 2		Vùng 3	
Thu nhập	Số người	Thu nhập	Số người	Thu nhập	Số người
1678	1	1675	2	1653	1
1680	3	1677	1	1654	1
1682	3	1680	1	1655	1
1683	2	1685	2	1660	2
1685	4	1692	1	1670	1
1687	1	1698	1	1674	1
1688	2	1705	2	1675	3
1690	1	1710	1	1678	1
1692	1	1715	1	1682	4
1693	1	1720	1	1692	1
1706	1	1726	3	1697	2
1708	1	1729	1	1703	1
1710	1	1732	2	1709	1
1712	1	1737	1	1715	1
1718	1	1740	1	1719	1
1720	2	1744	1	1724	1
1724	1	1747	1	1727	1
1737	1	1750	1	1732	1
1740	1	1755	1	1738	1
1747	2	1760	2	1743	1
1750	1	1765	2	1750	1
1755	1	1773	1	1760	1

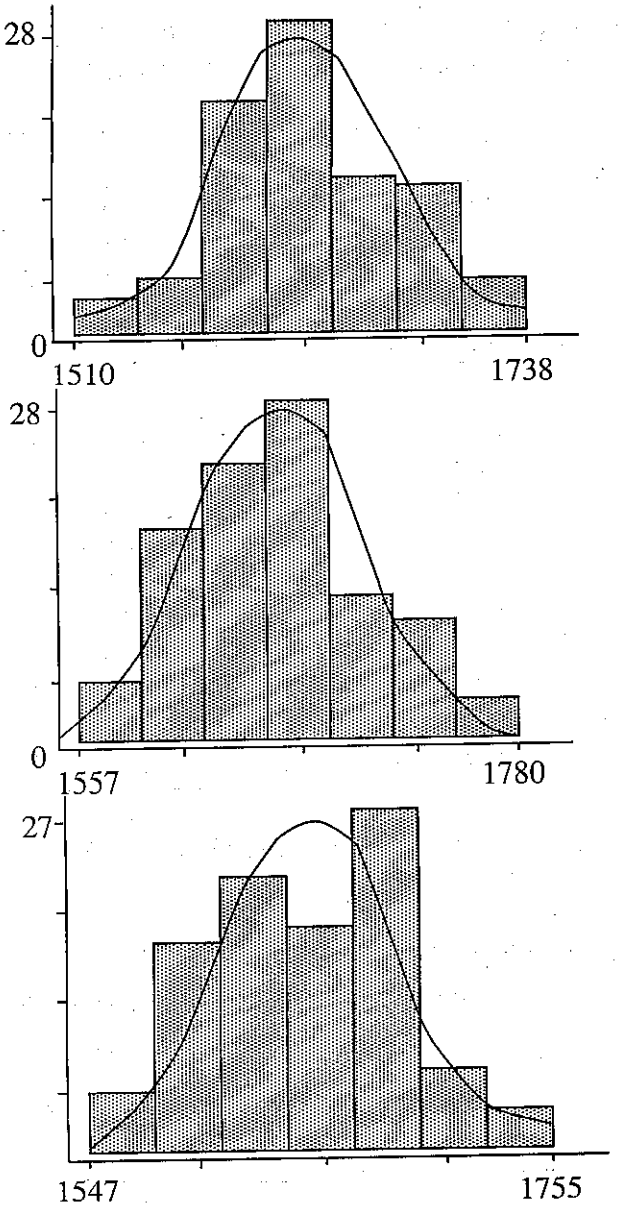
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Dùng phần mềm Excel ta vẽ được biểu đồ hình cột như sau đây cho các số liệu của vùng 3 (được ghép lại thành 7 lớp thu nhập). Các vùng khác cũng vẽ tương tự.



Để có thể so sánh sơ bộ phân phối thu thập của 3 vùng ta có thể đặt các biểu đồ hình cột cạnh nhau.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

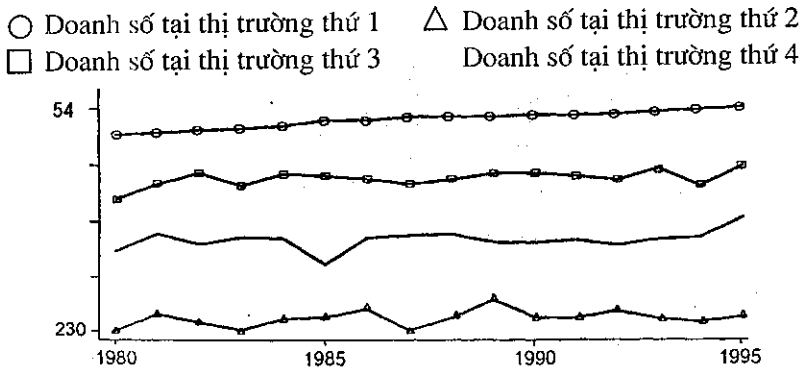


THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ B. Một công ty hoạt động ở 4 thị trường. Thống kê doanh số X_1, X_2, X_3, X_4 của công ty tại 4 thị trường nói trên trong 16 năm thu được kết quả sau:

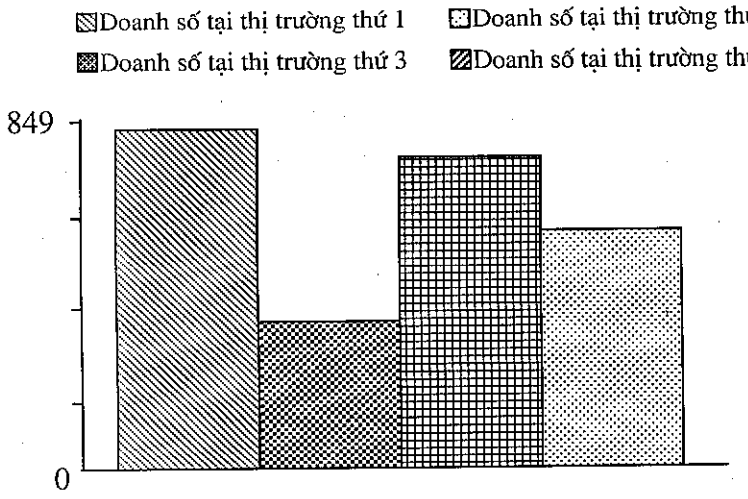
	X_1	X_2	X_3	X_4	Năm
1	510	240	428	345	1980
2	515	250	445	365	1981
3	520	240	460	355	1982
4	520	230	440	370	1983
5	522	245	460	370	1984
6	530	245	455	342	1985
7	530	255	455	370	1986
8	530	230	450	375	1987
9	535	245	460	375	1988
10	535	270	465	365	1989
11	540	245	465	365	1990
12	540	245	463	370	1991
13	540	255	460	360	1992
14	542	240	470	370	1993
15	543	235	455	373	1994
16	545	240	470	405	1995

Dùng Excel vẽ được các đường đa giác mô tả doanh số của công ty tại 4 thị trường theo thời gian như sau:

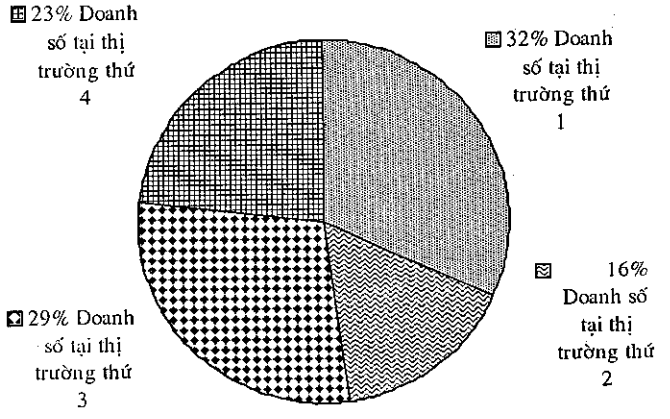


THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Dùng Stata ta thu được biểu đồ hình cột mô tả cơ cấu của 4 thị trường trong tổng doanh số của công ty như sau:



Biểu đồ hình cột doanh số tại các thị trường



Biểu đồ hình bánh xe tỷ trọng doanh số tại các thị trường

Những phương pháp mô tả nói trên phản ánh từng giá trị của mẫu ngẫu nhiên. Chúng có thể được sử dụng nhằm các mục đích khác nhau. Trước hết là để sơ bộ phân tích bản thân các số liệu mẫu, chẳng hạn để lọc đi những giá trị quá "xa" xu hướng chung của dấu hiệu nghiên cứu. Đó là biện pháp tiếp theo để loại trừ các sai sót trong quá trình chọn mẫu và tăng thêm tính đại diện của nó. Sau nữa phân phối thực nghiệm của mẫu cho phép đưa ra những nhận xét sơ bộ về dấu hiệu nghiên cứu trong tổng thể làm cơ sở để tiếp tục khái quát hoá chúng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5. THỐNG KÊ

5.1. Định nghĩa

Để nghiên cứu biến ngẫu nhiên gốc X trong tổng thể, nếu chỉ rút ra một mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$ thì mới chỉ có được một vài kết luận sơ bộ và rời rạc về X, vì các giá trị X_i của mẫu có cùng quy luật phân phối xác suất với X song quy luật này lại thường chưa được xác định hoàn toàn. Song nếu tổng hợp các biến ngẫu nhiên $X_1, X_2, ..., X_n$ này lại thì theo luật số lớn chúng sẽ bộc lộ những tính quy luật mới làm cơ sở để nhận định về biến ngẫu nhiên gốc X trong tổng thể.

Việc tổng hợp mẫu $W = (X_1, X_2, ..., X_n)$ được thực hiện dưới dạng một hàm nào đó của các giá trị $X_1, X_2, ..., X_n$ của mẫu. Nó được gọi là thống kê, ký hiệu là G. Như vậy

$$G = f(X_1, X_2, ..., X_n)$$

Như vậy về thực chất thống kê là một hàm của các biến ngẫu nhiên do đó bản thân nó cũng sẽ là một biến ngẫu nhiên tuân theo một quy luật phân phối xác suất nhất định và có các tham số đặc trưng như E(G), V(G)... Mặt khác, khi mẫu ngẫu nhiên nhận một giá trị cụ thể $w = (x_1, x_2, ..., x_n)$ thì G cũng nhận một giá trị cụ thể là:

$$g = f(x_1, x_2, ..., x_n)$$

Các thống kê cùng với quy luật phân phối xác suất của chúng là cơ sở để suy rộng các thông tin của mẫu cho dấu hiệu nghiên cứu của tổng thể.

Sau đây ta sẽ nghiên cứu một số thống kê thông dụng nhất.

5.2. Một số thống kê đặc trưng của mẫu ngẫu nhiên

Các thống kê đặc trưng của mẫu ngẫu nhiên cũng được chia thành ba loại:

- Các thống kê đặc trưng xu hướng trung tâm của phân phối của mẫu như trung bình mẫu, trung vị, mốt v.v...
- Các thống kê đặc trưng độ phân tán của phân phối của mẫu như khoảng biến thiên, phương sai, độ lệch chuẩn v.v...
- Các thống kê đặc trưng dạng phân phối.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sau đây ta sẽ xem xét một thông số thống kê đặc trưng mẫu quan trọng nhất.

1. Trung bình mẫu: Giả sử từ biến ngẫu nhiên gốc X trong tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

Trung bình mẫu là một thống kê, ký hiệu là $\overline{X}$ và là trung bình số học của các giá trị mẫu:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{6.20}$$

Ta chú ý rằng trung bình mẫu là một thống kê do đó nó là một biến ngẫu nhiên, tuân theo một quy luật phân phối xác suất nào đó với các tham số đặc trưng tương ứng. Khi mẫu ngẫu nhiên nhận một giá trị cụ thể $w = (x_1, x_2, ..., x_n)$ thì trung bình mẫu cũng nhận giá trị cụ thể bằng

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6.21}$$

hoặc

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \tag{6.22}$$

Trung bình mẫu $\overline{X}$ có tính chất sau: Nếu biến ngẫu nhiên gốc có kỳ vọng toán $E(X) = m$ và phương sai $V(X) = \sigma^2$ thì

$$E(\overline{X}) = m \tag{6.23}$$

và

$$V(\overline{X}) = \frac{\sigma^2}{n} \tag{6.24}$$

Thật vậy, theo định nghĩa

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

Áp dụng các tính chất của kỳ vọng toán ta có:

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Song từ (6.15) ta có $E(X_i) = E(X) = m\,(\forall i)$ do đó:

$$E(\overline{X}) = \frac{1}{n} nm = m$$

Mặt khác do X_i là các biến ngẫu nhiên độc lập, do đó theo tính chất của phương sai, ta có:

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$

và cũng từ (6.16) ta có $V(X_i) = V(X) = \sigma^2\,(\forall i)$ nên

$$V(\overline{X}) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Vậy bất kể biến ngẫu nhiên gốc phân phối theo quy luật nào, trung bình mẫu $\overline{X}$ cũng có kỳ vọng toán bằng kỳ vọng toán của biến ngẫu nhiên gốc, tức là $E(\overline{X}) = E(X) = m$, còn phương sai $V(\overline{X})$ của nó nhỏ hơn n lần so với phương sai của biến ngẫu nhiên gốc, $V(\overline{X}) = \frac{\sigma^2}{n}$, nghĩa là các giá trị có thể có của $\overline{X}$ ổn định quanh kỳ vọng toán m hơn các giá trị có thể có của X.

Nếu lấy căn bậc hai của $V(\overline{X})$ ta thu được độ lệch chuẩn

$$\sqrt{V(\overline{X})} = \frac{\sigma}{\sqrt{n}}$$

Độ lệch chuẩn này của $\overline{X}$ thường được dùng để phản ánh sai số ước lượng do đó người ta thường gọi là sai số chuẩn của trung bình mẫu $\overline{X}$. Vậy

$$Se(\overline{X}) = \frac{\sigma}{\sqrt{n}} \qquad (6.25)$$

Ở trên ta đã luôn giả thiết rằng mẫu được rút ra từ tổng thể theo phương thức có hoàn lại. Nếu kích thước tổng thể là vô hạn hoặc kích thước tổng thể là hữu hạn song $n < 0{,}1N$ thì có thể lấy mẫu không hoàn lại mà không ảnh hưởng đến kết quả. Song nếu $n > 0{,}1N$ thì trong các công thức trên phải sử dụng hệ số hiệu chỉnh do mẫu là không lặp (xem mục 5 chương III). Lúc đó:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$V(\overline{X}) = \frac{N-n}{N-1}.\frac{\sigma^2}{n} \quad (6.26)$$

và
$$Se(\overline{X}) = \sqrt{\frac{N-n}{N-1}.\frac{\sigma^2}{n}} \quad (6.27)$$

2. Trung vị: Trung vị, ký hiệu là X_d là giá trị nằm ở chính giữa tức là giá trị chia các số liệu mẫu thành hai phần bằng nhau.

Đối với một mẫu cụ thể, trung vị được xác định như sau:

a. Nếu các số liệu mẫu gồm n giá trị rời rạc được sắp xếp theo trình tự tăng dần và nếu n là một số lẻ thì trung vị là giá trị thứ $\frac{n+1}{2}$ trong dãy số liệu đó.

Thí dụ 1. Giả sử có các số liệu mẫu sau:

240 220 210 225 235 225 270 250 280

Ta có n = 9 số liệu nên trung vị là giá trị thứ $\frac{9+1}{2} = 5$ trong dãy số liệu đã được xếp theo thứ tự tăng dần:

210 220 225 225 235 240 250 270 280

↑

X_d

Còn nếu n là một số chẵn thì trung vị là hai giá trị nằm chính giữa dãy số liệu đó. Nó được gọi là khoảng trung vị.

Thí dụ 2. Giả sử mẫu có thêm số liệu 200 tức là n = 10 do đó hai giá trị nằm chính giữa là giá trị thứ 5 và thứ 6. Tức là

200 210 220 225 225 235 240 250 270 280

$$X_d = [225 - 235]$$

b. Nếu các số liệu mẫu được ghép lớp theo phân phối tần số thì giá trị trung vị có thể tính gần đúng bởi công thức sau:

$$X_d \approx L + \frac{\left(\frac{n}{2} - S\right)}{n_{x_d}} h \quad (6.28)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

trong đó: L - Giới hạn dưới của lớp chứa trung vị

n - Kích thước mẫu

S - Tổng tần số của các lớp đứng trước lớp chứa trung vị

n_{x_d} - Tần số của lớp chứa trung vị

h - Độ dài của lớp chứa trung vị

Thí dụ 3. Tìm trung vị của mẫu được cho bởi phân phối thực nghiệm trong bảng 6.4.

Bảng 6.4

Đoạn giá trị chiều dài h = 5	Tần số n_i	Tần số tích luỹ w_i
5-10	4	4
10-15	6	10
15-20	16	26
20-25	36	62
25-30	24	86
30-35	10	96
35-40	4	100
Tổng số	n = 100	

Ta có n/2 = 100/2 = 50 vậy trung vị nằm ở lớp thứ tư căn cứ vào cột tần số tích luỹ.

Từ đó

$$X_d \approx 20 + \frac{50-26}{36}5 = 23,33$$

Trung vị, cũng như trung bình mẫu, phản ánh xu hướng trung tâm của phân phối mẫu song nó có đặc điểm là không san bằng các chênh lệch giữa các giá trị của mẫu do đó nó thường được dùng để bổ xung hoặc thay thế trung bình mẫu khi không đủ số liệu để tính.

Người ta đã chứng minh được rằng nếu biến ngẫu nhiên X trong tổng thể là liên tục và mẫu rút ra có kích thước khá lớn thì thống kê trung vị X_d sẽ có kỳ vọng toán chính bằng trung bình tổng thể.

$$E(X_d) = m$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

và phương sai được xác định bằng biểu thức

$$V(X_d) = \frac{1}{4n[f(x_d)]^2}$$

trong đó n là kích thước mẫu, còn $f(X_d)$ là hàm mật độ xác suất tại điểm trung vị X_d. Chẳng hạn nếu X trong tổng thể phân phối chuẩn thì trung vị trùng với trung bình và

$$f(X_d) = f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

do đó

$$V(X_d) = \frac{1}{4n[f(x_d)]^2} = \frac{2\pi\sigma^2}{4n} = 1,57\frac{\sigma^2}{n} = 1,57V(\overline{X})$$

*3. **Mốt***. Mốt ký hiệu là X_0 là giá trị có tần số lớn nhất trong dãy số liệu mẫu. Với một mẫu cụ thể, mốt được xác định như sau: Nếu các số liệu mẫu là rời rạc và các giá trị $x_1, x_2,..., x_k$ xuất hiện với tần số tương ứng $n_1, n_2,..., n_k$ thì có thể xác định trực tiếp giá trị của mốt trên bảng phân phối tần số của mẫu.

Thí dụ 4. Bảng sau đây trình bày số nhân khẩu của một gia đình tại một khu vực gồm 200 gia đình (bảng 6.5).

Bảng 6.5

Số nhân khẩu x_i	Số gia đình n_i
1	10
2	30
3	75
4	45
5	20
6	15
7 trở lên	5
Tổng số	N=200

Từ bảng số liệu, ta tìm được X_0=3.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nếu các số liệu mẫu là ghép lớp theo bảng phân phối tần số thì mốt được tính gần đúng theo công thức:

$$X_0 \approx L + \left(\frac{d_1}{d_1 + d_2}\right).h \qquad (6.29)$$

trong đó: L - Giới hạn dưới của lớp chứa mốt;

d_1 - Hiệu số giữa tần số của lớp chứa mốt và tần số của lớp đứng trước;

d_2 - Hiệu số giữa tần số của lớp chứa mốt và tần số của lớp đứng sau;

h - Độ dài của lớp chứa mốt.

Thí dụ 5. Với các số liệu mẫu cho trong bảng 6.4 hãy tìm giá trị mốt.

Từ bảng 6.2 ta tìm thấy ngay lớp chứa mốt là lớp thứ 4.

Từ đó $X_0 \approx 20 + \left(\frac{20}{20+12}\right).5 = 23,125$

Cũng như trung vị, mốt không san bằng, bù trừ chênh lệch giữa các giá trị của mẫu, do đó nó bổ xung hoặc thay thế trung bình mẫu khi việc tính trung bình mẫu gặp khó khăn.

Trung bình, trung vị và mốt là các tham số chủ yếu đặc trưng cho xu hướng trung tâm của mẫu. Sau đây ta sẽ nghiên cứu một vài tham số đặc trưng cho độ phân tán của các giá trị của mẫu.

4. Khoảng biến thiên. Khoảng biến thiên, ký hiệu là R là sai lệch giữa các giá trị lớn nhất và nhỏ nhất của mẫu.

$$R = X_{max} - X_{min} \qquad (6.30)$$

Nếu các số liệu mẫu được ghép lớp thì khoảng biến thiên là hiệu số giữa cận trên của lớp cuối cùng với cận dưới của lớp đầu tiên trong dãy phân phối các giá trị của mẫu.

Việc tính khoảng biến thiên khá đơn giản song không mang lại nhiều thông tin về độ phân tán của các giá trị mẫu.

5. Khoảng tứ phân vị: Trong phân tích kinh tế, xã hội nhiều khi phải tính đến thứ bậc của các đơn vị nghĩa là chia các đơn vị của số liệu mẫu trong bảng phân phối thành các phần bằng nhau. Nếu mẫu được chia thành 4 phần bằng nhau thì có tứ phân vị. Tứ phân vị đầu là giá trị của mẫu đứng ở

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

vị trí cách đơn vị đầu tiên 1/4 số đơn vị của mẫu. Tứ phân vị thứ hai chính là trung vị. Tứ phân vị thứ ba là giá trị của mẫu đứng ở vị trí cách đơn vị đầu tiên 3/4 số đơn vị của mẫu.

Chẳng hạn với các số liệu mẫu

200 210 220 225 225 235 240 250 270 280

thì tứ phân vị đầu là giá trị nằm ở vị trí thứ n/4 = 10/4 = 2,5. Song do thứ tự vị trí phải là nguyên do đó nó là giá trị nằm ở vị trí thứ ba.

$$Q_1 = 220$$

Tứ phân vị thứ ba là giá trị nằm ở vị trí thứ 3n/4 = 3.10/4 = 7,5 do đó nó là giá trị nằm ở vị trí thứ 8.

$$Q_3 = 250$$

Nếu các số liệu mẫu được ghép lớp thì các tứ phân vị được tính gần đúng theo các công thức sau:

$$Q_1 = L_{Q_1} + \frac{\frac{n}{4} - S_{Q_1}}{n_{Q_1}} h_{Q_1} \qquad (6.31)$$

$$Q_3 = L_{Q_3} + \frac{\frac{3n}{4} - S_{Q_3}}{n_{Q_3}} h_{Q_3} \qquad (6.32)$$

trong đó:

L_{Q_1} và L_{Q_3} - Các giới hạn dưới của các lớp chứa Q_1 và Q_3.

S_{Q_1} và S_{Q_3} - Tổng các tần số của các lớp đứng trước lớp chứa Q_1 và Q_3.

n_{Q_1} và n_{Q_3} - Tần số của các lớp chứa Q_1 và Q_3.

h_{Q_1} và h_{Q_3} - Độ dài của các lớp chứa Q_1 và Q_3.

Chẳng hạn với các số liệu của bảng 6.4 ta tìm được lớp chứa Q_1 là lớp thứ ba (100/4 = 25) do đó:

$$Q_1 = 15 + \frac{25 - 10}{16} 5 = 19,6875$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lớp chứa Q_3 là lớp thứ năm (3.100/4 = 75) do đó:

$$Q_3 = 25 + \frac{75-62}{24}5 = 27,7083$$

Lúc đó khoảng tứ phân vị, ký hiệu là IQR được xác định bằng biểu thức:

$$IQR = Q_3 - Q_1 \qquad (6.33)$$

Khoảng tứ phân vị tuy đã nhạy cảm hơn đối với số liệu mẫu so với khoảng biến thiên song nó chỉ có nhiều ý nghĩa khi so sánh hai mẫu, còn với từng mẫu thì cũng không có nhiều ý nghĩa đặc trưng.

6. Tổng bình phương các sai lệch và độ lệch bình phương trung bình

Cho mẫu ngẫu nhiên được xây dựng từ biến ngẫu nhiên gốc X

$$W = (X_1, X_2, ..., X_n)$$

Lúc đó tổng bình phương các sai lệch giữa các giá trị của mẫu và trung bình mẫu được ký hiệu là SS và bằng

$$SS = \sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad (6.34)$$

Giá trị SS thường được sử dụng trong phân tích phương sai.

Nếu đem chia SS cho kích thước mẫu ta thu được trung bình số học của tổng bình phương sai lệch giữa các giá trị của mẫu và trung bình mẫu gọi tắt là *độ lệch bình phương trung bình,* ký hiệu là MS:

$$MS = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad (6.35)$$

Ta chú ý rằng, cũng giống như trung bình mẫu, SS và MS là các biến ngẫu nhiên với quy luật phân phối xác suất xác định và các tham số đặc trưng tương ứng. Còn trên một mẫu cụ thể chúng sẽ nhận những giá trị cụ thể.

Trong thực tế, với mẫu cụ thể, để tiện cho việc tính toán ms thường được tính bằng công thức sau:

$$ms = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$$

hoặc

$$ms = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2 \qquad (6.36)$$

Bạn đọc có thể chứng minh kết quả trên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

MS có tính chất sau: Nếu biến ngẫu nhiên gốc có E(X) = m và $V(X)=\sigma^2$ thì

$$E(MS) = \frac{n-1}{n}\sigma^2 \qquad (6.37)$$

Để chứng minh, trước hết ta biến đổi công thức của MS như sau:

$$MS = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}[(X_i - m) - (\overline{X} - m)]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}[(X_i - m)^2 - 2(\overline{X} - m)(X_i - m) + (\overline{X} - m)^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - 2(\overline{X} - m)\frac{1}{n}\sum_{i=1}^{n}(X_i - m) + (\overline{X} - m)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X} - m)^2$$

Do đó

$$E(MS) = \frac{1}{n}\sum_{i=1}^{n}E(X_i - m)^2 - E(\overline{X} - m)^2$$

Do $E(X_i - m)^2 = V(X_i) = V(X) = \sigma^2$

Và $E(\overline{X} - m)^2 = V(\overline{X}) = \frac{\sigma^2}{n}$

$$\Rightarrow E(MS) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

7. Phương sai mẫu S^2 và phương sai S^{*2}

- Phương sai mẫu, ký hiệu là S^2 được xác định bằng công thức:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right] \qquad (6.38)$$

- Phương sai S^{*2} được xác định bằng biểu thức:

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 \qquad (6.39)$$

trong đó m là trung bình tổng thể.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Dễ dàng thấy rằng giữa phương sai mẫu và MS có mối liên hệ sau:

$$S^2 = \frac{n}{n-1} MS$$

Cũng giống như MS, phương sai mẫu S^2 và phương sai S^{*2} thực chất là các biến ngẫu nhiên với quy luật phân phối xác suất tương ứng, còn giá trị của chúng trên một giá trị cụ thể của mẫu là những số xác định, ký hiệu là s^2 và s^{*2}.

Từ tính chất của MS có thể suy ra tính chất sau đây của phương sai mẫu:

$$E(S^2) = \sigma^2 \tag{6.41}$$

Thật vậy, theo (6.40) ta có:

$$E(S^2) = E\left(\frac{n}{n-1} MS\right) = \frac{n}{n-1} E(MS)$$

Thay giá trị của E(MS) từ (6.37) vào ta có:

$$E(S^2) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$$

Đối với phương sai S^{*2} ta cũng có tính chất tương tự sau:

$$E(S^{*2}) = \sigma^2 \tag{6.42}$$

Thật vậy

$$E(S^{*2}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i - m)^2$$

Do $$E(X_i - m) = V(X_i) = V(X) = \sigma^2$$

nên $$E(S^{*2}) = \sigma^2$$

Nếu lấy căn bậc hai của phương sai mẫu S^2, ta thu được thống kê gọi là *độ lệch chuẩn mẫu*, ký hiệu là S. Như vậy:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \tag{6.43}$$

Còn giá trị của nó trên một giá trị cụ thể của mẫu là một số xác định, ký hiệu là s.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8. Hệ số biến thiên

Hệ số biến thiên, ký hiệu là CV là biểu thức:

$$CV=\left|\frac{S}{\overline{X}}\right|100 \qquad (6.44)$$

Hệ số biến thiên được đo bằng phần trăm và được dùng để nhận xét về độ thuần nhất của phân phối mẫu và qua đó đo mức độ đại diện của trung bình mẫu cho xu hướng trung tâm của phân phối. Nếu CV<15% thì mẫu được xem là khá thuần nhất.

9. Hệ số bất đối xứng

Trên mẫu, hệ số bất đối xứng được xác định bằng biểu thức:

$$a_3=\frac{\sum_{i=1}^{n}\frac{\left(X_i-\overline{X}\right)^3}{n}}{S^3} \qquad (6.45)$$

Trong nhiều trường hợp hệ số bất đối xứng chỉ cần xác định một cách gần đúng bằng công thức:

$$a_3 \approx \frac{3(\overline{X}-X_d)}{S}$$

Giá trị của a_3 càng gần 0 thì phân phối thực nghiệm của các giá trị của mẫu càng đối xứng qua giá trị trung bình mẫu.

10. Hệ số nhọn

Trên mẫu hệ số nhọn được xác định bằng biểu thức:

$$a_4=\frac{\sum_{i=1}^{n}\frac{(X_i-\overline{X})^4}{n}}{S^4} \qquad (6.46)$$

11. Tần suất mẫu

Giả sử tổng thể kích thước N, trong đó M phần tử mang dấu hiệu nghiên cứu, lấy ra một mẫu ngẫu nhiên kích thước n và trong đó thấy có X phần tử mang dấu hiệu nghiên cứu. Lúc đó tần suất mẫu là một thống kê, ký hiệu là f và là tỷ số giữa số phần tử mang dấu hiệu nghiên cứu trong mẫu và kích thước mẫu:

$$f=\frac{X}{n} \qquad (6.47)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Về thực chất, thống kê f cũng là một biến ngẫu nhiên vì nó là hàm của biến ngẫu nhiên X - số lần xuất hiện dấu hiệu trong mẫu, tức là trong n phép thử độc lập còn giá trị của nó trên một mẫu cụ thể là một số xác định.

Ta chú ý rằng, cũng giống như tần suất của tổng thể là trường hợp riêng của trung bình tổng thể, tần suất mẫu là trường hợp riêng của trung bình mẫu $\overline{X}$ khi xem dấu hiệu nghiên cứu trong tổng thể như biến ngẫu nhiên tuân theo quy luật không - một.

Với nhận xét đó có thể chứng minh tính chất sau đây của tần suất mẫu: Nếu biến ngẫu nhiên gốc tuân theo quy luật A(p) với $E(X)=p$ và $V(X)=p(1-p)$ thì:

$$E(f)=p \tag{6.48}$$

$$V(f)=\frac{p(1-p)}{n} \tag{6.49}$$

Thật vậy, theo định nghĩa của tần suất mẫu:

$$E(f)=E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)$$

Song X là số lần xuất hiện dấu hiệu nghiên cứu trong mẫu, tức là số lần xuất hiện biến cố "dấu hiệu nghiên cứu" trong n phép thử độc lập nên X là biến ngẫu nhiên phân phối theo quy luật nhị thức với các tham số là $E(X)=np$ và $V(X)=np(1-p)$ vì vậy:

$$E(f)=\frac{1}{n}np=p$$

Tương tự:

$$V(f)=V\left(\frac{X}{n}\right)=\frac{1}{n^2}V(X)=\frac{1}{n^2}np(1-p)=\frac{p(1-p)}{n}$$

Từ đó sai số chuẩn của tần suất mẫu là:

$$Se(f)=\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Nếu mẫu lấy ra theo phương pháp không hoàn lại thì sai số chuẩn của tần suất mẫu là:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$Se(f)=\sqrt{\frac{N-n}{N-1}\frac{p(1-p)}{n}} \qquad (6.50)$$

Việc tính giá trị của tần suất mẫu f trên một mẫu cụ thể nói chung không có gì khó khăn, còn để tính trung bình mẫu $\overline{X}$, MS và các phương sai S^2, S^{*2} có thể gặp khó khăn khi có nhiều số liệu, do đó người ta thường lập bảng để tính toán cho thuận lợi hơn.

Thí dụ 6. Lấy ngẫu nhiên 100 sản phẩm do một máy sản xuất ra thấy có 5 phế phẩm. Tìm tỷ lệ phế phẩm của mẫu nói trên. Như vậy, dấu hiệu nghiên cứu trong tổng thể là "tính chất phế phẩm" của sản phẩm. Dấu hiệu này xuất hiện 5 lần trong mẫu kích thước bằng 100. Vậy ta có tần suất mẫu, theo (6.47) bằng

$$f=\frac{5}{100}=0,05$$

Thí dụ 7. Trở lại thí dụ về điều tra năng suất lúa. Gặt ngẫu nhiên 365 điểm trồng lúa của huyện thu được bảng số liệu sau:

Năng suất (tạ/ha)	25	30	33	34	35	36	37	39	40
Số điểm gặt tương ứng	6	13	38	74	106	85	30	10	3

Dấu hiệu nghiên cứu trong tổng thể là năng suất lúa trên một ha canh tác. Vậy các số liệu của mẫu nói trên cho phép xác định các thống kê đặc trưng của mẫu như trung bình mẫu (năng suất trung bình), phương sai mẫu (độ phân tán của năng suất) v.v...

Để tiện cho việc tính toán ta có thể lập bảng tính toán sau (bảng 6.6).

Từ đó, theo (6.22), ta có:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}n_i x_i=\frac{12700}{365}=34,795 \text{ tạ/ha}$$

và theo (6.36)

$$ms=\frac{1}{n}\sum_{i=1}^{k}n_i x_i-\bar{x}^2=\frac{443466}{365}-(34,795)^2$$

$$=1214,975-1210,692=4,283$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đó theo (6.40) ta có:

$$s^2 = \frac{n}{n-1}.ms = \frac{365}{364} \cdot 4,283 = 4,295$$

và

$$s = \sqrt{4,295} = 2,072$$

Bảng 6.6

x_i	n_i	$n_i x_i$	$n_i x_i^2$
25	6	150	3750
30	13	390	11700
33	38	1254	41382
34	74	2516	85544
35	106	3710	129850
36	85	3060	110160
38	30	1110	41070
39	10	390	15210
40	3	120	4800
	$\sum n_i = n = 365$	$\sum n_i x_i = 12700$	$\sum n_i x_i^2 = 443466$

Nếu các giá trị thu được của mẫu sai khác nhau rất ít hoặc khi có quá nhiều giá trị, người ta thường xây dựng bảng phân phối tần số ghép lớp.

Lúc đó, để tính các thống kê đặc trưng của mẫu, người ta lấy giá trị giữa của mỗi lớp đại diện cho lớp đó để tính toán.

Thí dụ 8. Lấy ngẫu nhiên 100 thanh niên ở một tỉnh đem đo chiều cao và thu được các số liệu sau:

Chiều cao (cm)	154 -158	158 -162	162 -166	166 -170	170 -174	174 -178	178 -182
Số thanh niên có chiều cao tương ứng	10	14	26	28	12	8	2

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ở đây dấu hiệu nghiên cứu là chiều cao thanh niên. Để xác định các thống kê mẫu như trung bình mẫu, phương sai mẫu... ta lập bảng tính, trong đó các giá trị x_i được thay bằng các giá trị giữa của mỗi lớp (bảng 6.7).

Bảng 6.7

x_i	n_i	$n_i x_i$	$n_i x_i^2$
156	10	1560	243360
160	14	2240	358400
164	26	4264	699296
168	28	4704	790272
172	12	2064	355008
176	8	1408	247808
180	2	360	64800
	$\sum n_i = n = 100$	$\sum n_i x_i = 16600$	$\sum n_i x_i^2 = 2758944$

Vậy $\bar{x} = \frac{16600}{100} = 166\text{cm}$

$$ms = \frac{2758944}{100} - 166^2 = 33,44$$

$$s = \sqrt{\frac{100}{99} 33,44} = 5,812$$

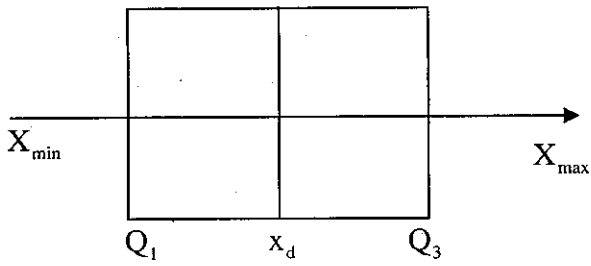
Khi kích thước mẫu khá lớn thì việc tính toán thủ công các thống kê đặc trưng mẫu sẽ gặp khó khăn. Lúc đó nên sử dụng các phần mềm như Excel, SPSS, Stata, MFIT v.v.., để tính các thống kê nói trên một cách nhanh chóng.

5.3. Đồ thị hình hộp (Box-Plot)

Đồ thị hình hộp là phương pháp mô tả và tổng hợp các số liệu mẫu bằng đồ thị, trên đó phản ánh được cùng một lúc cả các đặc trưng về xu hướng trung tâm cũng như độ phân tán của các giá trị của mẫu.

Để xây dựng đồ thị hình hộp, người ta sử dụng các thống kê đặc trưng mẫu là trung vị, các tứ phân vị Q_1, Q_3 và các giá trị x_{max} và x_{min} của phân phối mẫu. Nó có dạng như sau (hình 6.4).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Hình 6.4. Đồ thị hình hộp

Thí dụ 15. Vẽ đồ thị hình hộp với các số liệu cho trong bảng 6.4.

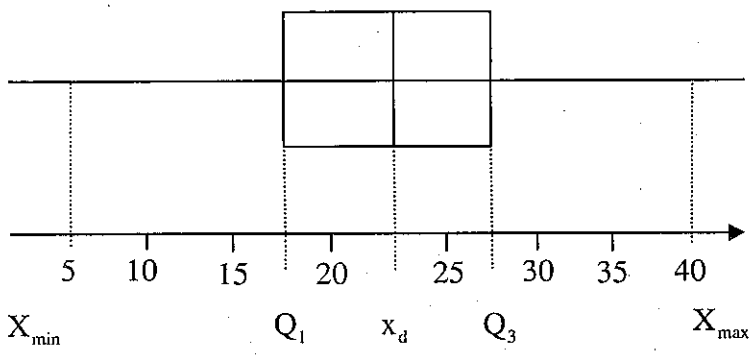
Từ các kết quả tính toán ở trên ta có:

$$m_d = 23{,}33$$

$$Q_1 = 19{,}6875\,,\ Q_3 = 27{,}7083$$

$$X_{min} = 5\,;\ X_{max} = 40$$

Vậy đồ thị hình hộp có dạng sau (hình 6.5).



Hình 6.5

Đặc biệt khi có nhiều mẫu rút ra từ các tổng thể nghiên cứu mà chúng lại có các thống kê đặc trưng khác nhau thì việc vẽ đồng thời các đồ thị hình hộp của các mẫu đó lên cùng một mặt phẳng sẽ cho phép so sánh trực quan các mẫu, từ đó có được những nhận xét sơ bộ về sự khác biệt của các tổng thể nghiên cứu tương ứng.

Trong các phần mềm thống kê luôn có chương trình để vẽ đồ thị hình hộp một cách dễ dàng và nhanh chóng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ. sau đây là các tham số đặc trưng chính của mẫu tìm được bằng phần mềm Excel theo các số liệu của thí dụ A.

Thu nhập cá nhân (USD/năm) vùng 1

Mean	1649.75
Standard Error	4.751254
Median	1643
Mode	1597
Standard Deviation	47.51254
Sample Variance	2257.442
Kurtosis	-0.5482
Skewness	0.200865
Range	208
Minimum	1547
Maximum	1755
Sum	164975
Count	100

Thu nhập cá nhân (USD/năm) vùng 2

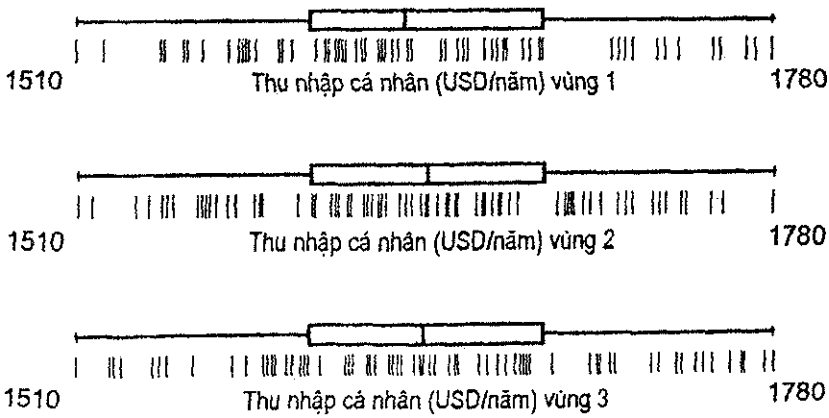
Mean	1683.86
Standard Error	5.932057
Median	1687
Mode	1730
Standard Deviation	59.32057
Sample Variance	3518.93
Kurtosis	-1.18153
Skewness	-0.30734
Range	216
Minimum	1557
Maximum	1773
Sum	168386
Count	100

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thu nhập cá nhân (USD/năm) vùng 3

Mean	1654.03
Standard Error	6.727892
Median	1646
Mode	1650
Standard Deviation	67.27892
Sample Variance	4526.454
Kurtosis	-1.03835
Skewness	0.02813
Range	250
Minimum	1510
Maximum	1760
Sum	165403
Count	100

Các đồ thị hình hộp thu được bằng phần mềm Excel được đặt trong cùng một mặt phẳng để tiện so sánh.



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

6.6. Hãy cho ví dụ về tổng thể nghiên cứu, dấu hiệu nghiên cứu và một mẫu ngẫu nhiên rút ra từ tổng thể đó.

6.7. Hãy phân tích các đặc điểm cơ bản của một mẫu ngẫu nhiên rút ra từ tổng thể nghiên cứu.

6.8. Mẫu ngẫu nhiên khác với mẫu cụ thể như thế nào? Cho ví dụ minh hoạ.

6.9. Hãy phân biệt dấu hiệu nghiên cứu định lượng và dấu hiệu nghiên cứu định tính. Cho ví dụ minh hoạ.

6.10. Để đo lường dấu hiệu nghiên cứu định lượng và định tính người ta dùng các loại thang đo nào? Cho ví dụ.

6.11. Hãy phân tích các phương pháp chọn mẫu chủ yếu được dùng trong phân tích kinh tế xã hội và sự khác biệt của các phương pháp đó. Cho ví dụ minh hoạ.

6.12. Một công ty điện lực phát các phiếu điều tra cho khách hàng với các mẫu câu hỏi sau đây:

a. Tuổi của chủ hộ.

b. Giới tính của chủ hộ.

c. Số người trong hộ.

d. Có dùng điện để đun nấu không (có hoặc không).

e. Nếu có thì đun nấu bình quân mấy giờ trong một ngày.

f. Thu nhập của hộ gia đình.

g. Tiền điện bình quân hàng tháng phải trả.

Hãy dùng các biến ngẫu nhiên để đặc trưng cho các câu hỏi trên và cho biết chúng là các biến định tính hay định lượng.

6.13. Hãy tính trung bình ($\bar{x}$) Trung vị x_m, mốt x_0, phương sai và độ lệch chuẩn của các mẫu cụ thể cho ở các bảng dưới đây:

x_i	-2	1	2	3	4	5
n_i	2	1	2	2	2	1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

x_i	12	13	15	17	18	20
n_i	2	5	8	4	4	2

x_i	4	7	8	12
n_i	5	2	3	10

x_i	21	24	25	26	28	32	34
n_i	10	20	30	15	10	10	5

ĐS:

a. $\overline{x} = 2$; $s^2 = 5,78$; $s = 2,404$

b. $\overline{x} = 15,56$; $s^2 = 5,507$; $s = 2,347$

c. $\overline{x} = 8,9$; $s^2 = 11,884$; $s = 3,447$

d. $\overline{x} = 26$; $s^2 = 10,909$; $s = 3,303$

6.14. Cho 8 kết quả đo đạc về một đại lượng X bởi cùng một máy đo không có sai số hệ thống

369; 378; 315; 420; 385; 401; 372; 383

Hãy tính $\overline{x}$, s^2;

6.15. Theo dõi thời gian hoàn thành một sản phẩm (phút) ở hai nhóm công nhân, ta thu được số liệu sau:

a. Nhóm 1

X(phút)	42	44	45	58	60	64
Số người	4	5	20	10	8	3

b. Nhóm 2

X(phút)	46	48	51
Số người	2	40	8

Hãy mô tả và tìm đặc trưng của hai mẫu cụ thể nói trên. Nêu nhận xét.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6.16. Tại một cửa hàng người ta thu thập số liệu về số lượng sản phẩm mà các nhân viên của cửa hàng bán được. Chọn ngẫu nhiên số sản phẩm bán được trong một ngày của 20 nhân viên bán hàng thu được các số liệu sau:

9 6 12 10 13 15 16 14 14 16 17 16 24 21 23
18 19 18 20 17

a. Tìm trung vị và khoảng tứ phân vị của mẫu nói trên.
b. Tìm trung bình và phương sai của mẫu nói trên.

6.17. Có số liệu sau đây về lãi suất trong một năm của 15 loại cổ phiếu (%)

12,5 13 14,8 11 16,7 9 8,3 -12 -3,9 15,5 16,2 18 11,6
10 9,5

a. Tìm trung vị, mốt, khoảng tứ phân vị của mẫu nói trên.
b. Tìm trung bình và phương sai của mẫu nói trên.

6.18. Theo dõi ngẫu nhiên 36 chuyến bay từ Hà Nội đi thành phố Hồ Chí Minh thu được các số liệu sau về số lượng khách của mỗi chuyến

128 121 134 136 136 118 123 109 120
116 125 128 121 129 130 131 127 119
114 134 110 136 134 125 128 123 128
133 132 136 134 129 132 125 120 129

a. Lập bảng phân phối tần số của số khách cho mỗi chuyến bay.

b. Tìm trung bình, trung vị và mốt và cho nhận xét về mối quan hệ giữa ba tham số đó.

c. Tìm phương sai, độ lệch chuẩn, hệ số biến thiên và cho nhận xét.

d. Tìm hệ số nhọn và hệ số bất đối xứng và cho nhận xét.

6.19. Đo cân nặng của 100 thanh niên tuổi trưởng thành ở một tỉnh, ta thu được bảng số liệu sau:

Cân nặng X(kg)	Dưới 58	58 đến 62	62 đến 66	66 đến 70	70 đến 74	74 đến 78	Trên 78
Số thanh niên	10	12	26	28	14	8	2

a. Vẽ biểu đồ hình cột của phân phối tần số trên.

b. Hãy tính $\bar{x}$, x_0, x_d, s^2, s, CV của mẫu nói trên.

ĐS: $\bar{x} = 66,24$ $s = 5,3069$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6.20. Các kết quả về việc đo độ bền các sợi chỉ ta thu được bảng số liệu dưới đây (đơn vị gam)

Độ bền của sợi	120 đến 140	140 đến 160	160 đến 180	180 đến 200	200 đến 220	220 đến 240	240 đến 260	260 đến 280
Số các sợi	1	4	10	14	12	6	2	1

Hãy xác định độ bền trung bình, phương sai và độ lệch chuẩn của mẫu trên.

ĐS: $\bar{x} = 195,2$ $s^2 = 829,55$ $s = 28,8$

6.21. Để nghiên cứu tuổi thọ (X) của một loại bóng đèn, người ta thắp thử 100 bóng và có số liệu như sau

Tuổi thọ (giờ)	Số bóng tương ứng (ni)
1010 - 1030	2
1030 - 1050	3
1050 - 1070	8
1070 - 1090	13
1090 - 1110	25
1110 - 1130	20
1130 - 1150	12
1150 - 1170	10
1170 - 1190	6
1190 - 1210	1
	n = 100

Hãy vẽ biểu đồ hình cột và tìm các tham số đặc trưng của mẫu nói trên.

6.22. Sau khi cải tiến kỹ thuật người ta thắp thử 100 bóng kết quả là:

Tuổi thọ	1150	1160	1170	1180	1190	1200
Số bóng (ni)	10	15	20	30	15	10

a. Hãy vẽ biểu đồ tần số của bảng phân phối tần số trên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b. So sánh tuổi thọ loại bóng đèn nói trên trước và sau khi cải tiến kỹ thuật qua hai mẫu cụ thể này (bằng cách so sánh 2 đồ thị hình hộp).

6.23. Gặt ngẫu nhiên 155 điểm trồng lúa của một huyện người ta được các số liệu sau đây:

Năng suất (ta/ha)	38	40	43	44	45	46	47	49	50	52
Số điểm gặt (n_i)	6	9	15	24	30	28	25	10	5	3

Hãy tìm các tham số đặc trưng của 365 điểm gặt nói trên.

ĐS: $\bar{x} = 45,148$ $s = 2,144$

6.24. Điều tra doanh số hàng tháng của 100 hộ kinh doanh một ngành hàng, ta thu được bảng số liệu sau:

Doanh số (triệu đ/tháng)	10,1	10,2	10,4	10,5	10,7	10,8	10,9	11	11,3	11,4
Số hộ (ni)	2	3	8	13	25	20	12	10	6	1

a. Hãy vẽ đường đa giác tần suất và biểu đồ tần suất của bảng phân phối thực nghiệm trên.

b. Tìm doanh số trung bình và độ phân tán của các doanh số so với doanh số trung bình của các hộ nói trên.

ĐS: $\bar{x} = 10,74$ $s^2 = 0,0685$

6.25. Tại một trại nuôi lợn người ta áp dụng thử một loại thuốc tăng trọng bổ sung vào khẩu phần thức ăn của lợn. Sau khi nuôi ba tháng thu được kết quả sau đây.

Trọng lượng (kg)	65	67	68	69	70	71	73
Số lợn (n_i)	1	3	4	7	6	2	2

b. Lô đối chứng (không dùng thuốc tăng trọng) cho kết quả sau:

Trọng lượng (kg)	55	56	58	59	60	62	63	65	67
Số lợn (n_i)	1	2	4	5	6	3	2	1	1

Hãy tìm trọng lượng trung bình sau 3 tháng tuổi và mức độ phân tán của trọng lượng so với trọng lượng trung bình của hai lô lợn nói trên.

ĐS: a. $\bar{x} = 69,16$ $s^2 = 3,2233$ b. $\bar{x} = 59,92$ $s^2 = 7,5767$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6.26. Có các số liệu sau đây về sản lượng thép hàng tháng của một tổng công ty thép (đơn vị: ngàn tấn)

7,0 6,9 8,2 7,8 7,7 7,3 6,8 6,7 8,2 8,4 7,0
6,7 7,5 7,2 7,9 7,6 6,7 6,6 6,3 5,6 7,8 5,5
6,2 5,8 5,8 6,1 6,0 7,3 7,3 7,5 7,2 7,2 7,4
7,6 7,3

a. Hãy vẽ đồ thị hình hộp của dãy số liệu trên

b. Phân phối của các số liệu là đối xứng hay bất đối xứng. Nếu là bất đối xứng thì lệch về phía nào?

6.27. Có các số liệu sau đây về trọng lượng bạc thu được từ một tấn quặng (đơn vị: kg) tại hai khu mỏ:

Mỏ A: 34 32 37 35 41 42 43 45 46 45 48
49 51 52 53 60 73 76 85
Mỏ B: 23 24 28 29 32 34 35 37 38 40 43
44 47 48 49 50 51 52 59

Hãy xây dựng đồ thị hình hộp cho mỗi dãy số liệu. Hãy đưa ra các kết luận thông qua việc so sánh hai đồ thị hình hộp đó.

6. MẪU NGẪU NHIÊN HAI CHIỀU

6.1. Khái niệm

Giả sử trên cùng một tổng thể phải nghiên cứu đồng thời hai dấu hiệu định tính hoặc định lượng, trong đó dấu hiệu nghiên cứu thứ nhất có thể xem như biến ngẫu nhiên X, còn dấu hiệu nghiên cứu thứ hai có thể xem như biến ngẫu nhiên Y. Lúc đó, việc nghiên cứu đồng thời hai dấu hiệu trong tổng thể tương tự như việc nghiên cứu một biến ngẫu nhiên hai chiều.

Từ tổng thể lấy ra một mẫu kích thước n, tức là thực hiện n phép thử đối với biến ngẫu nhiên (X,Y). Gọi X_i và Y_i tương ứng là giá trị của biến (X,Y) đo được trên phần tử thứ i của mẫu, ta thu được n biến ngẫu nhiên hai chiều độc lập. Từ đó ta có định nghĩa sau.

Định nghĩa. *Mẫu ngẫu nhiên hai chiều kích thước n là tập hợp của n biến ngẫu nhiên độc lập* $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ *được thành lập từ biến ngẫu nhiên hai chiều* (X,Y) *và có cùng quy luật phân phối xác suất với* (X,Y).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Mẫu ngẫu nhiên hai chiều được ký hiệu là:

$$W = \left[(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\right]$$

Việc thực hiện một phép thử đối với mẫu ngẫu nhiên W là thực hiện một phép thử đối với mỗi thành phần của mẫu.

Giả sử thành phần (X_i, Y_i) nhận giá trị (x_i, y_i) $(i = \overline{1,n})$ ta thu được một mẫu cụ thể là:

$$w = \left[(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\right]$$

Các giá trị X_i $(i = \overline{1,n})$ được gọi là thành phần X_i của mẫu, còn các giá trị Y_i $(i = \overline{1,n})$ được gọi là thành phần Y_i của mẫu. (Cũng giống như mẫu ngẫu nhiên 1 chiều X_i và Y_i có thể đo được bằng những thang khác nhau tùy thuộc vào dấu hiệu nghiên cứu (X,Y) là định tính hay định lượng).

6.2. Phương pháp mô tả mẫu ngẫu nhiên hai chiều

Giả sử từ tổng thể rút ra một mẫu kích thước n, trong đó thành phần X nhận các giá trị $x_1, x_2, ..., x_i, ..., x_h$, còn thành phần Y nhận các giá trị $y_1, y_2, ..., y_j, ..., y_k$ trong đó giá trị (x_i, y_j) xuất hiện với tần số $n_{ij} (i = \overline{1,h}; j = \overline{1,k})$. Lúc đó, sau khi các giá trị x_i và y_j được sắp xếp theo thứ tự tăng dần thì giá trị cụ thể của mẫu w được mô tả bằng bảng phân phối thực nghiệm sau:

X \ Y	y_1	y_2	...	y_j	...	y_k	n_i
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	n_1
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	n_2
⋮							
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	n_i
⋮							
x_h	n_{h1}	n_{h2}	...	n_{hj}	...	n_{hk}	n_h
m_j	m_1	m_2	...	m_j	...	m_k	$\sum = n$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trong đó n_i ký hiệu tổng các tần số của mẫu mang giá trị x_i của thành phần X, m_j ký hiệu tổng các tần số của mẫu mang giá trị y_j của thành phần Y.

Thí dụ 1. Bảng dưới đây chỉ kết quả thu hoạch Y(tạ/ha) và lượng phân bón X (kg/ha) của một loại hoa màu tại 100 thửa ruộng gieo trồng loại hoa màu đó (bảng 6.8).

Bảng 6.8

X \ Y	14	15	16	17	18	n_i
1	10					10
2	8	12				20
3		7	28			35
4			6	8		14
5				9	12	21
m_j	18	19	34	17	12	$\sum = 100$

Các điểm trống trên bảng thay cho tần số bằng 0.

Các số liệu của mẫu hai chiều cũng có thể mô tả thông qua đồ thị. Nếu biểu diễn các cặp giá trị (x_i, y_i) $(i = \overline{1,n})$ lên mặt phẳng thì ta sẽ thu được đồ thị rải điểm của mẫu, còn nếu nối các điểm trên bằng các đoạn thẳng thì thu được đồ thị liền nét của các giá trị của mẫu. Các đồ thị nói trên đều có thể vẽ bằng cách sử dụng các phần mềm thống kê thông thường.

6.3. Một số thống kê đặc trưng của mẫu ngẫu nhiên hai chiều

Từ bảng phân phối tần số thực nghiệm của mẫu ngẫu nhiên hai chiều ta có thể rút ra:

1. Bảng phân phối thực nghiệm của thành phần X

X	x_1	x_2	...	x_i	...	x_h
n_i	n_1	n_2	...	n_i	...	n_h

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

và tính được các thống kê đặc trưng của X:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{h} n_i X_i ; S_X^2 = \frac{1}{n-1}\sum_{i=1}^{h} n_i (X_i - \overline{X})^2 \qquad (6.51)$$

2. Bảng phân phối thực nghiệm của thành phần Y

Y	Y_1	Y_2	...	Y_j	...	Y_k
m_j	m_1	m_2	...	m_j	...	m_k

và tính được các thống kê đặc trưng của Y:

$$\overline{Y} = \frac{1}{n}\sum_{j=1}^{k} m_j Y_j ; S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{k} m_j (Y_j - \overline{Y})^2 \qquad (6.52)$$

3. Bảng phân phối có điều kiện của Y khi $X = X_i$

$Y_{X=X_i}$	Y_1	Y_2	...	Y_j	...	Y_k
n_{ij}	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}

trong đó $\sum_{i=1}^{k} n_{ij} = n_i$

Từ bảng phân phối có điều kiện của Y tính được trung bình có điều kiện của Y:

$$\overline{Y}_{X=X_i} = \frac{1}{n_i}\sum_{j=1}^{k} n_{ij} Y_j \qquad (6.53)$$

4. Tương tự ta có bảng phân phối có điều kiện của X khi $Y = Y_j$

$X_{Y=Y_j}$	X_1	X_2	...	X_i	...	X_h
n_{ij}	n_{1j}	n_{2j}	...	n_{ij}	...	n_{hj}

trong đó $\sum_{i=1}^{h} n_{ij} = m_j$

và tính được trung bình có điều kiện của X

$$\overline{X}_{Y=Y_j} = \frac{1}{m_i}\sum_{i=1}^{h} n_{ij} X_i \qquad (6.54)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 2. Từ bảng (6.8) ta có bảng phân phối có điều kiện của Y khi X = 3.

$Y_{X=3}$	15	16
n_{ij}	7	28

Từ đó $\overline{Y}_{X=3} = \dfrac{15.7 + 16.28}{35} = 15,8$

Tương tự, ta có bảng phân phối có điều kiện của X khi Y = 15.

$X_{Y=15}$	2	3
n_{ij}	12	7

Từ đó $\overline{X}_{Y=15} = \dfrac{2.12 + 3.7}{19} = 2,368$

5. Hệ số tương quan mẫu, ký hiệu là r và được xác định bằng biểu thức:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{SS_x}\sqrt{SS_y}} \qquad (6.55)$$

Đối với mẫu cụ thể mà các số liệu mẫu thường được trình bày dưới dạng bảng thì hệ số tương quan mẫu thường được tính bằng công thức sau:

$$r = \frac{n\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij}x_i y_i - \sum_{i=1}^{h} n_i x_i \sum_{j=1}^{k} m_j y_j}{\sqrt{n\sum_{i=1}^{h} n_i x_i^2 - \left(\sum_{i=1}^{h} n_i x_i\right)^2}\sqrt{n\sum_{j=1}^{k} m_j y_j^2 - \left(\sum_{j=1}^{k} m_j y_j\right)^2}}$$

Hệ số tương quan mẫu phản ánh mức độ kết hợp giữa X và Y. Các thống kê đặc trưng khác của mẫu hai chiều sẽ được đề cập tiếp ở các chương sau (xem chương X).

Trên đây là một số khái niệm liên quan đến mẫu ngẫu nhiên hai chiều. Khi nghiên cứu cùng một lúc n dấu hiệu trong tổng thể bằng phương pháp mẫu thì ta có khái niệm mẫu ngẫu nhiên n chiều và là sự mở rộng tương ứng các khái niệm đã xét ở trên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

6.28. Có các số liệu sau đây về lượng hàng bán ra và chi phí quảng cáo trong 100 tuần của một công ty.

Chi phí quảng cáo (tr.đ)	Lượng hàng bán ra (1000 sản phẩm)		
	10	12	15
1	5	10	5
1,5	8	12	15
2	12	18	15

a. Tính chi phí quảng cáo trung bình và lượng hàng hoá bán ra trung bình hàng tuần.

b. Tìm các phương sai của chi phí quảng cáo và lượng bán.

c. Tìm lượng bán trung bình khi chi phí quảng cáo là 1,5 triệu đồng.

d. Tìm hệ số tương quan mẫu và nhận xét

6.29. Có các số liệu sau đây về thu nhập X và tổng mức chi tiêu Y của một hộ gia đình (đơn vị triệu đồng) trên cơ sở điều tra ngẫu nhiên 100 gia đình:

a. Tìm $\bar{x}$, $\bar{y}$, S_x^2, S_y^2 và cho nhận xét

b. Tìm $\bar{y}/x = 30$: $\bar{y}/x = 40$ và $\bar{y}/x = 50$

c. Tìm r và cho nhận xét.

Y \ X	40	50	60	70
30	7	9	1	11
40	5	15	8	11
50	5	6	4	8

6.30. Điều tra ngẫu nhiên ý kiến của 5000 hành khách thường xuyên đi ô tô về chất lượng phục vụ của 3 hãng vận tải khách thu được kết quả sau đây.

Chất lượng phục vụ	Hãng vận tải khách		
	A	B	C
Rất tốt	325	275	800
Khá	800	1250	950
Bình thường	250	150	300
Kém	35	25	40

Hãy tính các đặc trưng cho từng hãng vận tải và sơ bộ kết luận về chất lượng phục vụ của hãng đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7. QUY LUẬT PHÂN PHỐI XÁC SUẤT CỦA MỘT SỐ THỐNG KÊ ĐẶC TRƯNG MẪU

Vì bản chất của các thống kê đặc trưng mẫu đã xét ở trên là các biến ngẫu nhiên, do đó để nắm được đầy đủ thông tin về các thống kê này cần khảo sát quy luật phân phối xác suất của chúng. Mặt khác, các quy luật phân phối xác suất cũng phản ánh mối liên hệ mật thiết giữa các thống kê của mẫu với các tham số đặc trưng của dấu hiệu nghiên cứu trong tổng thể. Đó là căn cứ cho việc suy rộng các kết quả thu được trên mẫu cho toàn bộ tổng thể nghiên cứu.

Nói chung quy luật phân phối xác suất của các thống kê đặc trưng mẫu phụ thuộc chặt chẽ vào quy luật phân phối xác suất của biến ngẫu nhiên gốc X. Để tiện cho việc sử dụng ở các chương tiếp theo trong một số trường hợp ta sẽ không khảo sát trực tiếp quy luật phân phối xác suất của các thống kê mẫu mà là quy luật phân phối xác suất của các thống kê là hàm của các thống kê nói trên. Mặt khác, ta cũng chỉ dừng lại ở việc thừa nhận các kết luận. Bạn đọc có thể tìm thấy chứng minh các kết luận này ở các tài liệu đầy đủ hơn về thống kê toán.

Sau đây ta xét một số thống kê thông dụng nhất.

7.1. Trường hợp biến ngẫu nhiên gốc phân phối theo quy luật chuẩn

Giả sử dấu hiệu ngiên cứu trong tổng thể có thể xem như một biến ngẫu nhiên tuân theo quy luật chuẩn với $E(X)=\mu$ và $V(X)=\sigma^2$. Các tham số có thể đã biết hoặc chưa biết. Từ tổng thể rút ra một mẫu ngẫu nhiên kích thước n:

$$W=(X_1,X_2,...,X_n)$$

Lúc đó, để xác định quy luật phân phối xác suất của các thống kê đặc trưng mẫu có thể sử dụng định lý sau:

Nếu các biến ngẫu nhiên $X_1,X_2,...,X_n$ độc lập và cùng phân phối theo quy luật chuẩn thì mọi tổ hợp tuyến tính của các biến ngẫu nhiên đó cũng phân phối theo quy luật chuẩn (xem mục 8 chương III).

Như vậy, theo định lý vừa trình bày thì các thành phần $X_1,X_2,...,X_n$ của mẫu ngẫu nhiên thỏa mãn các điều kiện của định lý, do đó có thể sử

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

dụng định lý để xét quy luật phân phối xác suất của các đặc trưng mẫu với tư cách là những tổ hợp tuyến tính của $X_1, X_2, ..., X_n$.

1. Thống kê trung bình mẫu $\overline{X}$ là một tổ hợp tuyến tính của $X_1, X_2, ..., X_n$ do đó nó cũng phân phối theo quy luật chuẩn với các tham số [xem (6.23) và (6.24)] là:

$$E(\overline{X}) = \mu \text{ và } V(\overline{X}) = \frac{\sigma^2}{n}$$

Do đó, nếu xây dựng tiếp thống kê

$$G = U = \frac{\overline{X} - \mu}{Se(\overline{X})} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}$$

thì (xem mục 7 chương III) thống kê U nói trên sẽ phân phối theo quy luật chuẩn hóa: $N(0,1)$.

2. Từ định nghĩa của phương sai S^{*2}

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

suy ra $$nS^{*2} = \sum_{i=1}^{n}(X_i - \mu)^2$$

Đem chia cả hai vế cho σ^2 ta thu được thống kê sau:

$$G = \chi^2 = \frac{nS^{*2}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

Do đó thống kê χ^2 (xem mục 9 chương III) sẽ phân phối theo quy luật "khi bình phương" với n bậc tự do: $\chi^2(n)$.

3. Cũng trên cơ sở thống kê mẫu nói trên, ta xây dựng thống kê

$$G = \chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

Có thể chứng minh được rằng thống kê χ^2 nói trên phân phối theo quy luật "khi bình phương" với $(n-1)$ bậc tự do: $\chi^2(n-1)$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4. Từ các kết quả thu được là:

$$U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \quad \text{phân phối } N(0,1) \text{ và}$$

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \quad \text{phân phối } \chi^2(n-1)$$

nên nếu ta xây dựng tiếp thống kê:

$$G = T = \frac{U}{\sqrt{\frac{\chi^2}{n-1}}} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} : \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = \frac{(\overline{X} - \mu)\sqrt{n}}{S}$$

thì thống kê T (xem mục 10 chương III) sẽ phân phối theo quy luật Student với $(n-1)$ bậc tự do: $T(n-1)$.

Chú ý rằng khi n khá lớn thì quy luật Student hội tụ khá nhanh về quy luật chuẩn hóa, do đó với $n > 30$ thực tế có thể xem thống kê T phân phối xấp xỉ $N(0,1)$.

7.2. Trường hợp có hai biến ngẫu nhiên gốc cùng phân phối theo quy luật chuẩn

Giả sử ta xét cùng một lúc hai tổng thể. Ở tổng thể thứ nhất dấu hiệu nghiên cứu được xem như biến ngẫu nhiên X_1 phân phối chuẩn với $E(X_1) = \mu_1$ và $V(X_1) = \sigma_1^2$; ở tổng thể thứ hai dấu hiệu nghiên cứu được xem như biến ngẫu nhiên X_2 phân phối chuẩn với $E(X_2) = \mu_2$ và $V(X_2) = \sigma_2^2$.

Từ hai tổng thể nói trên rút ra hai mẫu ngẫu nhiên độc lập có kích thước tương ứng là n_1 và n_2.

$$W_1 = (X_{11}, X_{12}, ..., X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, ..., X_{2n_2})$$

1. Xét thống kê $(\overline{X_1} - \overline{X_2})$ là hiệu của hai trung bình mẫu. Đây là một tổ hợp tuyến tính của hai biến ngẫu nhiên độc lập cùng phân phối theo quy luật chuẩn, do đó bản thân nó cũng là biến ngẫu nhiên phân phối chuẩn với kì vọng toán là:

$$E(\overline{X_1} - \overline{X_2}) = E(\overline{X_1}) - E(\overline{X_2}) = \mu_1 - \mu_2$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

và phương sai là:

$$V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

do đó nếu xây dựng thống kê

$$G = U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

thì U sẽ phân phối chuẩn hóa $N(0,1)$.

2. Mặt khác ta thấy các thống kê $\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2}$ và $\chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$ cùng phân phối theo quy luật khi bình phương với các bậc tự do tương ứng là $(n_1 - 1)$ và $(n_2 - 1)$, do đó theo tính chất cộng của χ^2 ta có thống kê

$$\chi^2 = \frac{(n_1 - 1)S_1^2}{\sigma_2^2} + \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

cũng sẽ phân phối theo quy luật khi bình phương với $(n_1 + n_2 - 2)$ bậc tự do.

Ở trên ta đã có thống kê

$$U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

phân phối chuẩn hóa và độc lập với χ^2, vì vậy nếu $\sigma_1^2 = \sigma_2^2 = \sigma^2$ thì thống kê

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} : \frac{\sqrt{\frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}}}{\sqrt{n_1 + n_2 - 2}}$$

$$= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

sẽ phân phối theo quy luật Student với (n_1+n_2-2) bậc tự do và nếu $n_1>30$ và $n_2>30$ thì nó phân phối xấp xỉ chuẩn hóa $N(0,1)$.

Mặt khác nếu xét thống kê

$$T=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

thì người ta đã chứng minh được rằng nó phân phối Student với số bậc tự do là:

$$k=\frac{(n_1-1)(n_2-1)}{(n_2-1)C^2+(1-C)^2(n_1-1)}$$

trong đó $C=\dfrac{S_1^2/n_1}{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$

song nếu $n_1>30$ và $n_2>30$ thì có thể xem như T phân phối xấp xỉ chuẩn hóa $N(0,1)$.

3. Do

$$\chi_1^2=\frac{(n_1-1)S_1^2}{\sigma_1^2} \text{ và } \chi_2^2=\frac{(n_2-1)S_2^2}{\sigma_2^2}$$

phân phối $\chi^2(n_1-1)$ và $\chi^2(n_2-1)$. Do đó nếu ta xây dựng thống kê

$$G=F=\frac{\frac{\chi_1^2}{n_1-1}}{\frac{\chi_2^2}{n_2-1}}=\frac{S_1^2\cdot\sigma_2^2}{S_2^2\cdot\sigma_1^2}$$

thì thống kê F (xem mục 11 chương III) sẽ phân phối theo quy luật Fisher-Snedecor với (n_1-1) và (n_2-1) bậc tự do:

$$F(n_1-1,n_2-1).$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7.3. Trường hợp biến ngẫu nhiên gốc X phân phối theo quy luật không - một

Giả sử trong tổng thể dấu hiệu nghiên cứu có thể xem như biến ngẫu nhiên X phân phối theo một quy luật nào đó khác với quy luật chuẩn. Lúc đó để cho đơn giản người ta thường yêu cầu có thể rút ra một mẫu ngẫu nhiên có kích thước n khá lớn

$$W = (X_1, X_2, ..., X_n)$$

Lúc đó để xác định quy luật phân phối xác suất của các thống kê đặc trưng mẫu, theo định lý giới hạn Lindenberg - Lewi ta có:

$$U = \frac{(\overline{X} - m)\sqrt{n}}{\sigma} \text{ và } U = \frac{(\overline{X} - m)\sqrt{n}}{S}$$

sẽ phân phối xấp xỉ chuẩn hóa $N(0,1)$ *khi n khá lớn.*

Ta sẽ vận dụng định lý trên để xét một vài trường hợp thông dụng nhất trong thực tế.

1. Giả sử trong tổng thể dấu hiệu nghiên cứu có thể xem như biến ngẫu nhiên phân phối theo quy luật không - một. Từ tổng thể lập mẫu ngẫu nhiên kích thước n

$$W = (X_1, X_2, ..., X_n)$$

Ở mục 3 chương III ta đã biết rằng lúc đó tần suất mẫu f sẽ phân phối theo quy luật nhị thức với các tham số đặc trưng là:

$$E(f) = p$$

$$V(f) = \frac{p(1-p)}{n}$$

với p là xác suất (cơ cấu) của tổng thể.

2. Nếu kích thước mẫu n lớn mà p lại rất nhỏ và $np \approx np(1-p)$ thì tần suất mẫu sẽ phân phối xấp xỉ quy luật Poisson với tham số là p.

3. Nếu kích thước mẫu n lớn mà p lại không nhỏ song thỏa mãn điều kiện

$$n > 5 \text{ và } \frac{\left|\sqrt{\frac{p}{p-1}} - \sqrt{\frac{1-p}{p}}\right|}{\sqrt{n}} < 0,3$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

thì tần suất mẫu sẽ phân phối xấp xỉ chuẩn với $E(f) = p$ và $V(f) = \frac{p(1-p)}{n}$.

Do đó biến ngẫu nhiên

$$U = \frac{f-p}{Se(f)} = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}}$$

sẽ phân phối xấp xỉ chuẩn hóa $N(0,1)$.

7.4. Trường hợp có hai biến ngẫu nhiên gốc cùng phân phối theo quy luật không- một

Giả sử có hai tổng thể trong đó dấu hiệu nghiên cứu trong hai tổng thể có thể xem như các biến ngẫu nhiên phân phối không- một với các tham số tương ứng là p_1 và p_2. Từ hai tổng thể nói trên rút ra hai mẫu ngẫu nhiên độc lập kích thước tương ứng là n_1 và n_2.

$$W_1 = (X_{11}, X_{12}, ..., X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, ..., X_{2n_2})$$

Xét thống kê $f_1 - f_2$ là hiệu của hai tần suất mẫu. Lúc đó nếu $n_1 > 30$ và $n_2 > 30$ thì $f_1 - f_2$ sẽ phân phối xấp xỉ chuẩn theo định lí giới hạn trung tâm với các tham số đặc trưng là:

$$E(f_1 - f_2) = p_1 - p_2$$

và

$$V(f_1 - f_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

do đó thống kê

$$U = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

sẽ phân phối xấp xỉ chuẩn hóa $N(0,1)$

Trên đây là quy luật phân phối xác suất của một số thống kê thông dụng nhất. Còn quy luật phân phối xác suất của một số thống kê đặc biệt khác sẽ được đề cập trong từng trường hợp cụ thể.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8. SUY DIỄN THỐNG KÊ

Như phần trên đã thấy, quy luật phân phối xác suất của các thống kê đặc trưng mẫu phản ánh mối liên hệ chặt chẽ giữa các tham số của mẫu với các tham số tương ứng của tổng thể nghiên cứu. Trong thực tế, các kết luận thu được ở trên về phân phối xác suất của các đặc trưng mẫu được sử dụng trong suy đoán thông kê theo hai cách thức trái ngược nhau:

- Suy diễn thống kê: Nếu đã biết quy luật phân phối xác suất cũng như các tham số đặc trưng của tổng thể thì có thể sử dụng các kết luận trên để suy đoán về tính chất của một mẫu ngẫu nhiên rút ra từ tổng thể đó. Nói cách khác đây chính là việc sử dụng thông tin đã biết của tổng thể để suy đoán về một bộ phận của tổng thể đó.

- Quy nạp thống kê: Nếu đã biết các tham số đặc trưng của mẫu thì căn cứ vào các kết luận trên để suy đoán về tính chất của tổng thể nghiên cứu mà từ đó mẫu được rút ra. Nói cách khác đây là việc sử dụng thông tin đã biết của một bộ phận của tổng thể để suy đoán về toàn bộ tổng thể đó.

Ở phần này ta sẽ nghiên cứu một số bài toán suy diễn thống kê quan trọng hơn cả trong thực tiễn. Còn các vấn đề quy nạp thống kê được xét ở các chương tiếp theo.

8.1. Suy diễn về mẫu ngẫu nhiên rút ra từ tổng thể phân phối chuẩn

Giả sử trong tổng thể nghiên cứu biến ngẫu nhiên X phân phối chuẩn với kỳ vọng toán μ và phương sai σ^2 đã biết. Lúc đó nếu từ tổng thể rút ra một mẫu ngẫu nhiên kích thước n thì có thể căn cứ vào thông tin đã biết về tổng thể để suy đoán về mẫu như sau:

1. Suy đoán về giá trị của trung bình mẫu

Ta có thống kê

$$U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim N(0,1)$$

Vì vậy với xác suất $1-\alpha$ có thể tìm được các giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tìm được các giá trị tới hạn $u_{1-\alpha_1}$ và u_{α_2} tương ứng sao cho thỏa mãn điều kiện:

$$P[u_{1-\alpha_1} < U < u_{\alpha_2}] = 1 - \alpha$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đó

$$P\left[u_{1-\alpha_1} < \frac{(\overline{X}-\mu)\sqrt{n}}{\sigma} < u_{\alpha_2}\right] = 1-\alpha$$

Sau khi chuyển vế ta thu được:

$$P\left[\mu - \frac{\sigma}{\sqrt{n}}u_{\alpha_1} < \overline{X} < \mu + \frac{\sigma}{\sqrt{n}}u_{\alpha_2}\right] = 1-\alpha \qquad (6.56)$$

Với những cặp giá trị α_1 và α_2 khác nhau ta thu được khoảng giá trị khác nhau của $\overline{X}$.

Thí dụ 1. Một công ty sản xuất mỳ ăn liền với trọng lượng đóng gói là biến ngẫu nhiên phân phối chuẩn có trọng lượng đóng gói trung bình là 340 gam và độ lệch chuẩn là 10 gam. Lấy ngẫu nhiên 16 gói mỳ để kiểm tra.

a. Tìm xác suất để trọng lượng trung bình của các gói mỳ đó nằm trong khoảng từ 335 gam đến 345 gam.

b. Tìm xác suất để trọng lượng trung bình của các gói mỳ đó lớn hơn 340 gam.

c. Với xác suất 0.99 thì trọng lượng trung bình của các gói mỳ đó nằm trong khoảng nào xung quanh giá trị trung bình.

Giải. Gọi X là trọng lượng mỳ đóng gói. Theo giả thiết X phân phối chuẩn với $\mu = 340$ và $\sigma = 10$. Lấy ngẫu nhiên một mẫu $n = 16$.

a. Ta phải tìm xác suất để trung bình mẫu $\overline{X}$ nằm trong khoảng (335;345). Theo công thức (6.56) ta có:

$$\mu - \frac{\sigma}{\sqrt{n}}u_{\alpha_1} = 340 - \frac{10}{\sqrt{15}}u_{\alpha_1} = 335$$

Từ đó $u_{\alpha_1} = 2$

và

$$\mu + \frac{\sigma}{\sqrt{n}}u_{\alpha_2} = 340 + \frac{10}{\sqrt{16}}u_{\alpha_2} = 345$$

Từ đó $u_{\alpha_2} = 2$

Tra bảng giá trị tới hạn u_α ta có $u_{\alpha_1} = u_{\alpha_2} = 2 = u_{0,0228}$

vậy $\alpha_1 = \alpha_2 = 0{,}0228$ hay $\alpha = \alpha_1 + \alpha_2 = 0{,}0456$.

Từ đó $1-\alpha = 1-0{,}0456 = 0{,}9544$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b. Để tìm xác suất sao cho trung bình mẫu lớn hơn 340 ta lấy $\alpha_2 = 0; \alpha_1 = \alpha$ từ đó $u_{\alpha_2} = +\infty$ và công thức (6.56) trở thành

$$P\left[\overline{X} > \mu - \frac{\sigma}{\sqrt{n}} u_\alpha\right] = 1 - \alpha$$

Từ đó

$$\mu - \frac{\sigma}{\sqrt{n}} u_\alpha = 340 - \frac{10}{\sqrt{16}} u_\alpha = 340$$

Từ đó

$$u_\alpha = 0 = u_{0,5}. \text{ Vậy } \alpha = 0,5 \Rightarrow 1 - \alpha = 0,5$$

c. Ta phải tìm giá trị a và b sao cho

$$p[a < \overline{X} < b] = 0,99 = 1 - \alpha$$

Vậy

$$\alpha = 0,01 \Rightarrow \alpha_1 = \alpha_2 = \frac{\alpha}{2} = 0,005$$

$$\Rightarrow u_{0,005} = 2,58$$

Do đó

$$a = \mu - \frac{\sigma}{\sqrt{n}} U_{\alpha_1} = 340 - \frac{10}{\sqrt{16}} \cdot 2,58 = 333,55 \text{ (gam)}$$

và $b = \mu + \frac{\sigma}{\sqrt{n}} U_{\alpha_2} = 340 + \frac{10}{\sqrt{16}} \cdot 2,58 = 346,45$ (gam)

Vậy với xác suất 0,99 trọng lượng trung bình của các gói mỳ được đem kiểm tra sẽ nằm trong khoảng (333,55 ; 346,45) gam.

2. Suy đoán về giá trị của phương sai mẫu

Ta có thống kê

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Vì vậy với xác suất $1 - \alpha$ có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tìm được các giá trị tới hạn $\chi^{2(n-1)}_{1-\alpha_1}$ và $\chi^{2(n-1)}_{\alpha_2}$ tương ứng sao cho thỏa mãn điều kiện

$$P\left[\chi^{2(n-1)}_{1-\alpha_1} < \chi^2 < \chi^{2(n-1)}_{\alpha_2}\right] = 1 - \alpha$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thay giá trị của χ^2 vào biểu thức và sau khi chuyển vế ta thu được

$$P\left[\frac{\sigma^2}{n-1}\chi^{2(n-1)}_{1-\alpha_1} < S^2 < \frac{\sigma^2}{n-1}\chi^{2(n-1)}_{\alpha_2}\right] = 1-\alpha \qquad (6.57)$$

với mỗi cặp giá trị α_1 và α_2 ta sẽ có một khoảng giá trị tương ứng của S^2.

Thí dụ 2. Xí nghiệp sử dụng một loại nguyên liệu với lượng tạp chất là biến ngẫu nhiên phân phối chuẩn với phương sai là 20 $(gam)^2$ trong 1kg nguyên liệu. Từ một lô nguyên liệu mới nhập về, người ta lấy ngẫu nhiên ra 16kg. Tìm xác suất để độ phân tán (phương sai) của lượng tạp chất trong mẫu hàng nằm trong khoảng $[9,68;33,33]$ $(gam)^2$.

Giải. Gọi X là lượng tạp chất trong nguyên liệu thì X phân phối chuẩn với $\sigma^2 = 20$. Nếu lấy một mẫu ngẫu nhiên $n = 16$ thì theo (6.57) ta có:

$$\frac{\sigma^2}{n-1}\chi^{2(n-1)}_{1-\alpha_1} = \frac{20}{15}\chi^{2(15)}_{1-\alpha_1} = 9,68$$

$$\Rightarrow \chi^{2(15)}_{1-\alpha_1} = 7,26 = \chi^{2(15)}_{0,95}$$

$$\Rightarrow 1-\alpha_1 = 0,95 \Rightarrow \alpha_1 = 0,05$$

$$\text{và } \frac{\sigma^2}{n-1}\chi^{2(n-1)}_{\alpha_2} = \frac{20}{15}\chi^{2(n-1)}_{\alpha_2} = 33,33$$

$$\Rightarrow \chi^{2(n-1)}_{\alpha_2} = 24,9975 = \chi^{2(15)}_{0,05}$$

$$\Rightarrow \alpha_2 = 0,05$$

Từ đó $\alpha = \alpha_1 + \alpha_2 = 0,1 \Rightarrow 1-\alpha = 0,9.$

8.2. Suy diễn về hai mẫu ngẫu nhiên rút ra từ hai tổng thể phân phối chuẩn

Giả sử có hai tổng thể, trong đó các dấu hiệu nghiên cứu là các biến ngẫu nhiên X_1 và X_2 cùng phân phối chuẩn với các kỳ vọng toán tương ứng là μ_1, μ_2 và các phương sai tương ứng là σ_1^2, σ_2^2 đã biết. Nếu từ các tổng thể trên rút ra hai mẫu độc lập kích thước tương ứng là n_1 và n_2 thì có thể tiến hành các suy đoán sau đây về các mẫu ngẫu nhiên đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1. Suy đoán về hiệu hai trung bình mẫu

Ta có thống kê

$$U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Vì vậy với xác suất 1 - α và tiến hành như ở mục trước ta thu được công thức sau về khoảng giá trị của $\overline{X_1} - \overline{X_2}$

$$P\left((\mu_1 - \mu_2) - u_{\alpha_1}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \overline{X_1} - \overline{X_2} < (\mu_1 - \mu_2) + u_{\alpha_2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \quad (6.58)$$
$$= 1 - \alpha$$

Ta sẽ thu được các khoảng giá trị khác nhau của $\overline{X_1} - \overline{X_2}$ với mỗi cặp giá trị α_1 và α_2.

Chú ý rằng nếu không biết quy luật phân phối xác suất của X_1 và X_2 trong các tổng thể song các kích thước mẫu đều đủ lớn ($n_1 > 30$ và $n_2 > 30$) thì thống kê $\overline{X_1} - \overline{X_2}$ vẫn có thể xem như phân phối xấp xỉ chuẩn.

Thí dụ 3. Hai công ty A và B cùng sản xuất một loại bóng đèn có tuổi thọ trung bình tương ứng là 12000 và 11000 giờ với độ lệch chuẩn là 200 và 150 giờ. Nếu kiểm tra ngẫu nhiên 50 bóng đèn của công ty A và 75 bóng đèn của công ty B thì xác suất để tuổi thọ trung bình của các bóng đèn được kiểm tra của A lớn hơn của B ít nhất 950 giờ là bao nhiêu?

Giải. Gọi X_1 và X_2 là tuổi thọ của bóng đèn của công ty A và B. Ta có

$$\mu_1 = 12000; \mu_2 = 11000$$
$$\sigma_1 = 200; \sigma_2 = 150$$
$$n_1 = 50; n_2 = 75$$

Vì các kích thước mẫu là đủ lớn nên không cần giả thiết X_1 và X_2 phân phối chuẩn. Ta phải tìm $P(\overline{X_1} - \overline{X_2} > 950)$.

Nếu trong công thức (6.58) đặt $\alpha_2 = 0$, $\alpha_1 = \alpha$ thì ta có công thức

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P\left(\overline{X_1}-\overline{X_2}>\mu_1-\mu_2-U_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)=1-\alpha$$

Từ đó

$$12000 - 11000 - U_\alpha\sqrt{\frac{200^2}{50}+\frac{150^2}{75}} = 950$$

$$\rightarrow U_\alpha = 1{,}51 \rightarrow \alpha = 0{,}0655 \rightarrow 1 - \alpha = 0{,}9345$$

Như vậy là trong 93,45% trường hợp kiểm tra tương tự thì tuổi thọ trung bình của bóng đèn do công ty A sản xuất sẽ lớn hơn của công ty B ít nhất 950 giờ.

2. Suy đoán về tỷ số giữa hai phương sai mẫu

Ta có thống kê

$$F=\frac{S_1^2}{S_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}\sim F(n_1-1,n_2-1)$$

Do đó với xác suất 1 - α có thể tìm được cặp giá trị α_1 và α_2 và các giá trị tới hạn $f_{1-\alpha_1}^{(n_1-1,n_2-1)}; f_{\alpha_2}^{(n_1-1,n_2-1)}$ thoả mãn điều kiện

$$P\left[f_{1-\alpha_1}^{(n_1-1,n_2-1)}<F<f_{\alpha_2}^{(n_1-1,n_2-1)}\right]=1-\alpha$$

Từ đó thay biểu thức của F vào và chuyển vế ta thu được công thức

$$P\left(\frac{\sigma_1^2}{\sigma_2^2}f_{1-\alpha_1}^{(n_1-1,n_2-1)}<\frac{S_1^2}{S_2^2}<\frac{\sigma_1^2}{\sigma_2^2}f_{\alpha_2}^{(n_1-1,n_2-1)}\right)=1-\alpha \qquad (6.59)$$

Và tiến hành các suy đoán tương ứng về tỷ số S_1^2/S_2^2.

Thí dụ 4. Quay trở lại thí dụ 3. Giả sử ta biết rằng X_1 và X_2 phân phối chuẩn. Lúc đó xác suất để độ phân tán về tuổi thọ của các bóng đèn được kiểm tra của công ty A lớn hơn của công ty B ít nhất 2 lần bằng bao nhiêu?

Giải. Trong công thức (6.59) nếu ta đặt $\alpha_2 = 0$ thì thu được công thức:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P\left(\frac{S_1^2}{S_2^2} > \frac{\sigma_1^2}{\sigma_2^2} f_{1-\alpha}^{(n_1-1,n_2-1)}\right) = 1-\alpha$$

$$\Rightarrow \frac{200^2}{150^2} f_{1-\alpha}^{(49,74)} = 2$$

$$\Rightarrow f_{1-\alpha}^{(49,74)} = 1{,}125$$

Tra bảng giá trị tới hạn F ta được $1 - \alpha = 0{,}1$

8.3. Suy diễn về mẫu ngẫu nhiên rút ra từ tổng thể phân phối không-một

Giả sử biến ngẫu nhiên X trong tổng thể phân phối không- một với tham số là p. Lúc đó nếu từ tổng thể rút ra một mẫu ngẫu nhiên kích thước n thì có thể suy đoán về số lần xuất hiện biến cố trong mẫu đó hoặc tần suất mẫu.

Ta có thống kê

$$U = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim \text{xấp xỉ } N(0,1)$$

nếu thỏa mãn điều kiện

$$n > 5 \text{ và } \left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \Big/ \sqrt{n} < 0{,}3$$

Lúc đó với xác suất $1-\alpha$ có thể tìm được cặp giá trị α_1 và $\alpha_2 (\alpha_1 + \alpha_2 = \alpha)$ và hai giá trị tới hạn $u_{1-\alpha_1}$ và u_{α_2} tương ứng thỏa mãn điều kiện.

$$P[u_{1-\alpha_1} < U < u_{\alpha_2}] = 1-\alpha$$

Thay biểu thức của U và chuyển vế ta thu được công thức:

$$P\left[p - \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_{\alpha_1} < f < p + \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_{\alpha_2}\right] = 1-\alpha \qquad (6.60)$$

Từ đó tiến hành các suy đoán tương ứng đối với f.

Thí dụ 5. Một lô hàng đủ tiêu chuẩn xuất khẩu nếu tỷ lệ sản phẩm loại một của lô hàng đó là 90%. Từ một lô hàng dự định xuất khẩu người ta lấy ngẫu nhiên ra 100 sản phẩm thì trong đó phải có ít nhất bao nhiêu sản phẩm loại một thì đủ tiêu chuẩn xuất khẩu. Hãy kết luận với xác suất 0,9.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Gọi p là tỷ lệ sản phẩm loại một của lô hàng đó, $p = 0,9$. Nếu lấy ra một mẫu sản phẩm $n = 100$ và gọi X là số sản phẩm loại một trong đó thì ta có tần suất mẫu

$$f = \frac{X}{n} = \frac{X}{100}$$

Do $n = 100 > 5$ và $\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \Big/ \sqrt{n}$

$$= \left|\sqrt{\frac{0,9}{0,1}} - \sqrt{\frac{0,1}{0,9}}\right| \Big/ \sqrt{100} = 0,267 < 0,3$$

nên có thể dùng công thức (6.58). Trong biểu thức này nếu lấy $\alpha_2 = 0$ thì ta có:

$$P\left[f > p - \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_\alpha\right] = 1 - \alpha = 0,9$$

từ đó $\alpha = 0,1$ và $u_\alpha = 1,282$

vậy $$f > 0,9 - \frac{\sqrt{0,1 \cdot 0,9}}{\sqrt{100}} 1,282 = 0,86154$$

từ đó $X > 100 \cdot 0,86154 = 86,154$.

Song vì X phải là số nguyên nên $X \geq 87$ sản phẩm. Vậy trong mẫu sản phẩm đó phải có ít nhất 87 sản phẩm loại một thì lô hàng mới xuất khẩu được.

Chú ý rằng việc suy diễn bằng công thức (6.60) chỉ tiến hành được nếu mẫu đủ lớn để quy luật nhị thức có thể xấp xỉ chuẩn. Nếu điều kiện đó không được thỏa mãn thì phải sử dụng trực tiếp quy luật nhị thức hoặc quy luật Poisson.

1. Trường hợp mẫu nhỏ phải áp dụng trực tiếp quy luật nhị thức

Nếu kích thước mẫu khá nhỏ (chẳng hạn chưa đến 1/10 kích thước của tổng thể) thì phải áp dụng trực tiếp quy luật nhị thức.

Thí dụ 6. Tiếp tục thí dụ 5. Giả sử lô hàng xuất khẩu có 1000 sản phẩm mà ta chỉ có thể lấy ra 20 sản phẩm để kiểm tra. Vì mẫu quá nhỏ nên không thể

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

dùng quy luật chuẩn thay cho quy luật nhị thức. Gọi X là số sản phẩm không phải loại một có trong 20 sản phẩm được kiểm tra thì X tuân theo quy luật nhị thức với n = 20 và p = 0,1. Lúc đó X có bảng phân phối xác suất như sau:

X	0	1	2	3	4	5	6	7	8	9
P	0,1216	0,2702	0,2582	0,1901	0,0898	0,0319	0,0089	0,002	0,0004	0,0001

Các xác suất được tìm theo phụ lục 1. Ta không lấy các giá trị X > 9 vì xác suất tương ứng quá nhỏ và có thể coi bằng không.

Từ bảng phân phối xác suất của X có thể xây dựng hàm phân bố xác suất của nó.

X	0	1	2	3	4	5	6	7	8	9
F(x)	0	0,1216	0,3918	0,677	0,8671	0,9569	0,9888	0,9977	0,9997	1

Từ bảng trên ta thấy trong khoảng giá trị từ 0 đến 5 đã chứa đựng 95,69% các giá trị của X. Do đó nếu trong mẫu n = 20 sản phẩm đem kiểm tra mà có từ 0 đến 5 sản phẩm không phải loại một, tức là có từ 15 đến 20 sản phẩm loại một thì có thể cho rằng tỷ lệ sản phẩm loại một của lô hàng đó là 95,69%, do đó có thể xuất khẩu được.

2. Trường hợp áp dụng quy luật Poisson

Nếu ta yêu cầu tỷ lệ sản phẩm loại một của lô hàng xuất khẩu phải đạt 99% tức là chỉ còn 1% sản phẩm không phải loại một và nếu kiểm tra một mẫu n = 100 sản phẩm thì do n lớn song p lại quá nhỏ nên có thể thay thế quy luật nhị thức bằng quy luật Poisson. Gọi X là số sản phẩm không phải loại một trong mẫu sản phẩm được kiểm tra thì có thể xem X phân phối Poisson với $\lambda = np = 1$. Từ đó X có bảng phân phối xác suất như sau:

X	0	1	2	3	4	5	6	7	…
P	0,3679	0,3679	0,1839	0,0613	0,0153	0,0031	0,0005	0,0001	…

Các xác suất được tìm theo phụ lục 2. Ta không lấy các giá trị X > 7 vì xác suất tương ứng quá nhỏ và có thể coi bằng không.

Từ bảng phân phối xác suất của X có thể xây dựng hàm phân bố xác suất của nó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

X	0	1	2	3	4	5	6	7	...
F(x)	0	0,3679	0,7358	0,9197	0,981	0,9963	0,9994	1	...

Từ bảng trên ta thấy trong khoảng giá trị từ 0 đến 5 đã chứa đựng 95,63% các giá trị của X. Do đó nếu trong mẫu n = 100 sản phẩm đem kiểm tra mà có từ 95 đến 100 sản phẩm loại một thì có thể cho rằng tỷ lệ sản phẩm loại một của lô hàng đó là 95,63%, do đó có thể xuất khẩu được.

8.4. Suy diễn về hai mẫu ngẫu nhiên rút ra từ hai tổng thể phân phối không-một

Giả sử có hai tổng thể, trong đó các biến ngẫu nhiên X_1 và X_2 cùng phân phối không một với các tham số là p_1 và p_2. Nếu từ hai tổng thể rút ra hai mẫu ngẫu nhiên độc lập kích thước tương ứng là n_1 và n_2 thì có thể suy đoán về hiệu $f_1 - f_2$ như sau:

Ta có thống kê:

$$U = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{(f_1 - f_2) - (p_1 - p_2)}{S_f} \sim N(0,1)$$

nếu n_1 và n_2 khá lớn. Do đó với xác suất 1 - α có thể tìm được cặp giá trị α_1 và α_2 ($\alpha_1 + \alpha_2 = \alpha$) và hai giá trị tới hạn $u_{1-\alpha_1}$ và u_{α_2} tương ứng thỏa mãn điều kiện.

$$P[u_{1-\alpha_1} < U < u_{\alpha_2}] = 1 - \alpha$$

Thay biểu thức của U và chuyển vế ta thu được công thức

$$P\left[(p_1 - p_2) - U_{\alpha_1} S_f < f_1 - f_2 < (p_1 - p_2) + U_{\alpha_2} S_f\right] = 1 - \alpha \qquad (6.61)$$

Từ đó có thể tiến hành các suy đoán khác nhau về hiệu $f_1 - f_2$.

Thí dụ 7. Tỷ lệ dân chúng ưa thích dùng loại dầu gội mác X ở hai tỉnh A và B tương ứng là 50% và 40%. Phỏng vấn ngẫu nhiên 100 khác hàng ở mỗi tỉnh thì tỉnh A có 52 người và tỉnh B có 38 người ưa thích dùng sản phẩm đó. Tìm xác suất để sự khác biệt về tỷ lệ ưa thích sản phẩm mác X trong một mẫu ngẫu nhiên khác sẽ cao hơn kết quả phỏng vấn nói trên.

Giải. Gọi p_1 và p_2 là tỷ lệ ưa thích sản phẩm mác X ở hai tỉnh A và B. Nếu lấy hai mẫu khách hàng $n_1 = n_2 = 100$ từ mỗi tỉnh và gọi f_1 và f_2 là tỷ lệ

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

khách hàng ưa thích sản phẩm mác X của hai mẫu khách hàng đó thì ta phải đi tìm xác suất để

$$f_1 - f_2 > 0{,}52 - 0{,}38 = 0{,}14$$

Nếu trong công thức (6.61) ta lấy $\alpha_2 = 0$ thì thu được công thức sau

$$P\left[f_1 - f_2 > (p_1 - p_2) - U_\alpha S_f\right] = 1 - \alpha$$

Từ đó

$$(0{,}5 - 0{,}4) - \sqrt{\frac{0{,}5.0{,}5}{100} + \frac{0{,}4.0{,}6}{100}}U_\alpha = 0{,}14$$

$$\Rightarrow U_\alpha = -0{,}571 = -U_{0{,}2843} = U_{0{,}7157}$$

$$\Rightarrow \alpha = 0{,}7157 \Rightarrow 1 - \alpha = 0{,}2843$$

Vậy khả năng để lấy một mẫu ngẫu nhiên khác từ mỗi tỉnh mà chênh lệch về tỷ lệ khách hàng ưa thích sản phẩm X ở tỉnh A so với tỉnh B cao hơn kết quả phỏng vấn thu được là 28,43%.

Bài tập

6.31. Một ngẫu nhiên kích thước n = 64 được rút ra từ tổng thể phân phối chuẩn với trung bình là 50 và độ lệch chuẩn là 4. Tìm xác suất để trung bình mẫu nằm trong khoảng từ 48,5 đến 51,5.

ĐS: 0,9974

6.32. Độ lệch chuẩn của kích thước chi tiết được ước lượng là 4mm. Kích thước các chi tiết được sản xuất là biến ngẫu nhiên phân phối chuẩn song phải tập trung quanh giá trị 40mm. Tìm xác suất để lấy ngẫu nhiên 4 chi tiết để kiểm tra thì kích thước trung bình của chúng nằm trong khoảng từ 35mm đến 45mm.

ĐS: 0,9876

6.33. Một mẫu kích thước n được rút ra từ tổng thể phân phối chuẩn với trung bình là μ và độ lệch chuẩn là 10. Hãy xác định n sao cho:

a. $P(\mu - 10 < \bar{x} < \mu + 10) = 0{,}954$

b. $P(\mu - 5 < \bar{x} < \mu + 5) = 0{,}954$

c. $P(\mu - 2 < \bar{x} < \mu + 2) = 0{,}954$

ĐS: a. n = 4; b. n = 16; c. n = 100

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6.34. Trọng lượng một sản phẩm là biến ngẫu nhiên phân phối chuẩn với trung bình là 20,5 và độ lệch chuẩn là 2. Lấy ngẫu nhiên 4 sản phẩm để kiểm tra thì với xác suất 0,95 trọng lượng trung bình của chúng sai lệch so với trọng lượng qui định tối đa là bao nhiêu.

ĐS: 1,96

6.35. Biến ngẫu nhiên phân phối theo quy luật khi bình phương với 10 bậc tự do. Hãy tìm các giá trị a và b thoả mãn các điều kiện sau:

$P(\chi^2 > a) = 0,975$

Và $P(a < \chi^2 < b) = 0,95$

ĐS: $a = 3,247$; $b = 20,483$

6.36. Lập mẫu ngẫu nhiên kích thước $n = 25$ từ tổng thể phân phối chuẩn với phương sai bằng 24. Tìm xác suất để phương sai mẫu nằm trong khoảng từ 12,401 đến 36,415.

ĐS: 0,925

6.37. Hai mẫu ngẫu nhiên kích thước 100 được rút ra từ tổng thể phân phối chuẩn có cùng trung bình và phương sai tương ứng là 50 và 40. Tìm xác suất để:

a. Trung bình của hai mẫu đó sai lệch nhau ít nhất 2.

b. Phương sai của mẫu thứ nhất lớn hơn của mẫu thứ hai ít nhất 2 lần.

ĐS:a. 0,0348

b. 0,01

6.38. Hai mẫu ngẫu nhiên kích thước $n_1 = 40$ và $n_2 = 50$ được rút ra từ các tổng thể phân phối chuẩn có $\mu_1 = 70$; $\mu_2 = 68$ và các phương sai $\sigma_1^2 = 120$ và $\sigma_2^2 = 150$. Tìm xác suất để:

a. Trung bình của mẫu thứ nhất lớn hơn của mẫu thứ hai ít nhất là 5.

b. Phương sai của mẫu thứ hai lớn hơn của mẫu thứ nhất không quá 1,5 lần.

ĐS:a. 0,1112

b. 0,75

6.39. Tỷ lệ phế phẩm cho phép của một lô hàng là 10%. Vậy với xác suất 0,95 nếu lấy ngẫu nhiên ra 100 sản phẩm để kiểm tra thì tỷ lệ phế phẩm tối đa của mẫu sản phẩm đó là bao nhiêu để có thể chấp nhận lô hàng đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6.40. Tỷ lệ đỗ tốt nghiệp trung học chung của cả nước là 70%. Vậy một trường có 800 học sinh thi tốt nghiệp thì phải có tối thiểu bao nhiêu em đỗ thì sẽ được coi là bình thường. Hãy kết luận với xác suất 0,95.

6.41. Người ta biết rằng 40% các gia đình ở Hà Nội có thu nhập hàng năm nằm trong khoảng từ 600 USD đến 1200 USD. Vậy phải điều tra một mẫu gồm bao nhiêu gia đình để với xác suất 0,95 tỷ lệ các gia đình có thu nhập trong khoảng nói trên sai lệch so với tỷ lệ chung của thành phố không quá 4%.

ĐS: n = 577

6.42. Tỷ lệ đàn ông và phụ nữ ủng hộ việc sử dụng các biện pháp tránh thai là 0,52 và 0,65. Vậy nếu phỏng vấn ngẫu nhiên 400 đàn ông và 400 phụ nữ về vấn đề trên thì xác suất để tỷ lệ đàn ông và phụ nữ trong hai mẫu nói trên khác biệt nhau quá 16% là bao nhiêu.

ĐS: 0,1922

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9. TÓM TẮT CHƯƠNG 6

1. Đối tượng nghiên cứu của thống kê là tổng thể, chủ đề nghiên cứu của nó là những dấu hiệu trên tổng thể mà luôn có thể đại diện bằng các biến ngẫu nhiên, vì vậy mục tiêu cơ bản của việc nghiên cứu một tổng thể thực chất là xác lập quy luật phân phối xác suất và tìm các tham số đặc trưng của các biến ngẫu nhiên đặc trưng cho tổng thể đó.

2. Để nghiên cứu một tổng thể có thể dùng phương pháp điều tra toàn bộ song phương pháp này ít được sử dụng vì có nhiều hạn chế.

3. Phương pháp mẫu là phương pháp nghiên cứu thống kê đối lập với phương pháp điều tra toàn bộ. Sử dụng các mô hình của lý thuyết xác suất, nó sử dụng thông tin của một bộ phận của tổng thể để kết luận về toàn bộ tổng thể đó.

4. Để có thể có được các suy đoán chính xác về tổng thể thì mẫu phải mang tính đại diện, tức là nó phải là một mẫu ngẫu nhiên.

5. Các tham số đặc trưng mẫu thực chất là các biến ngẫu nhiên, vì vậy để có thể tiến hành các suy đoán thông kê thì phải xác lập được quy luật phân phối xác suất và các tham số đặc trưng của chúng.

6. Bài toán suy diễn thống kê là một quá trình của thống kê suy diễn. Nó dùng thông tin đã biết của tổng thể để suy đoán về một bộ phận của tổng thể đó. Đây là một cách tiếp cận khác của bài toán kiểm định về tổng thể.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các ký hiệu và công thức cơ bản

* Tổng thể nghiên cứu kích thước N

* Dấu hiệu nghiên cứu χ được mô hình hóa bằng biến ngẫu nhiên X.

* Tần suất tổng thể

$p_i = \frac{N_i}{N}$ trong đó N_i - số phần tử của tổng thể mang giá trị x_i

* Tần suất tích lũy của tổng thể

$$F(x_i) = \frac{W_i}{N} = \sum_{x_j < x_i} \frac{N_i}{N}$$

* Các tham số cơ bản của tổng thể

+ Trung bình tổng thể (trung bình số học)

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i$$

+ Trung bình điều hòa

$$m_h = \frac{N}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

+ Trung bình nhân

$$m_g = \sqrt[N]{x_1 x_2 ... x_N}$$

+ Phương sai tổng thể

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - m^2$$

* Mẫu ngẫu nhiên kích thước n: $W = (X_1, X_2, ..., X_n)$

* Mẫu cụ thể $w = (x_1, x_2, ..., x_n)$

* Tần suất mẫu

$f_i = \frac{n_i}{n}$ trong đó n_i là số phần tử của mẫu mang giá trị x_i

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

* Tần suất tích lũy của mẫu

$$F^*(x_i) = \frac{w_i}{n} = \sum_{x_j<x_i} \frac{n_i}{n}$$

* Các thống kê đặc trưng mẫu

+ Trung bình mẫu

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

* Kỳ vọng toán của trung bình mẫu

$$E(\overline{X}) = m$$

* Sai số chuẩn của trung bình mẫu

$$Se(\overline{X}) = \sqrt{V(\overline{X})} = \frac{\sigma}{\sqrt{n}}$$

+ Trung vị mẫu (số liệu ghép lớp)

$$X_d \approx L + \frac{(n/2 - S)}{n_{x_d}} h$$

+ Mốt (số liệu ghép lớp)

$$X_0 \approx L + \left(\frac{d_1}{d_1 + d_2}\right) h$$

+ Khoảng biến thiên

$$R = x_{max} - x_{min}$$

+ Khoảng tứ phân vị

$$IQR = Q_3 - Q_1$$

trong đó:

$$Q_1 = L_{Q_1} + \frac{n/4 - S_{Q_1}}{n_{Q_1}} h_{Q_1}$$

$$Q_3 = L_{Q_3} + \frac{3n/4 - S_{Q_3}}{n_{Q_3}} h_{Q_3}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Độ lệch bình phương trung bình

$$MS = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

+ Phương sai mẫu

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

+ Phương sai S^{*2}

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$$

+ Độ lệch chuẩn mẫu S

+ Tần suất mẫu (với số liệu định danh)

$f = \frac{X}{n}$ trong đó X- số phần tử mang dấu hiệu nghiên cứu

- Kì vọng toán của tần suất mẫu $E(f) = p$
- Sai số chuẩn của tần suất mẫu

$$Se(f) = \sqrt{V(f)} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

* Mẫu ngẫu nhiên hai chiều

$$W = \left[(X_1, Y_1)(X_2, Y_2)...(X_n, Y_n)\right]$$

* Hệ số tương quan mẫu

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{SS_x}\sqrt{SS_Y}}$$

* Phân phối xác suất của các thống kê đặc trưng mẫu

+ $U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim N(0,1)$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ $T = \dfrac{(\overline{X} - \mu)\sqrt{n}}{S} \sim T(n-1)$

+ $\chi^2 = \dfrac{nS^{*2}}{\sigma^2} \sim \chi^2(n)$

+ $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

+ $U = \dfrac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim N(0,1)$

+ $U = \dfrac{(f-p)\sqrt{n}}{\sqrt{f(1-f)}} \sim N(0,1)$ nếu $n \geq 100$

+ $U = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$

+ $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim T(n_1+n_2-2)$

nếu $\sigma_1^2 = \sigma_2^2$

+ $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim T(k)$

với $k = \dfrac{(n_1-1)(n_2-1)}{(n_2-1)C^2 + (1-C)^2(n_1-1)}$; $C = \dfrac{S_1^2/n_1}{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}$

+ $F = \dfrac{S_1^2}{S_2^2} \cdot \dfrac{\sigma_2^2}{\sigma_1^2} \sim F(n_1-1, n_2-1)$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ $U = \dfrac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$

* Công thức suy diễn về trung bình mẫu

$$P\left(\mu - \frac{\sigma}{\sqrt{n}} u_{\alpha_1} < \overline{X} < \mu + \frac{\sigma}{\sqrt{n}} u_{\alpha_2}\right) = 1 - \alpha$$

* Công thức suy diễn về phương sai mẫu

$$P\left(\frac{\sigma^2}{n-1}\chi^{2(n-1)}_{1-\alpha_1} < S^2 < \frac{\sigma^2}{n-1}\chi^{2(n-1)}_{\alpha_2}\right) = 1 - \alpha$$

* Công thức suy diễn về hai trung bình mẫu

$$P\left((\mu_1 - \mu_2) - u_{\alpha_1}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \overline{X_1} - \overline{X_2} < (\mu_1 - \mu_2) + u_{\alpha_2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

* Công thức suy diễn về hai phương sai mẫu

$$P\left(\frac{\sigma_1^2}{\sigma_2^2} f^{(n_1-1,n_2-1)}_{1-\alpha_1} < \frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} f^{(n_1-1,n_2-1)}_{\alpha_2}\right) = 1 - \alpha$$

* Công thức suy diễn về tần suất mẫu

$$P\left(p - \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_{\alpha_1} < f < p + \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_{\alpha_2}\right) = 1 - \alpha$$

* Công thức suy diễn về hai tần suất mẫu

$$P\left((p_1 - p_2) - u_{\alpha_1}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} < f_1 - f_2 < (p_1 - p_2) + u_{\alpha_2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right) = 1 - \alpha$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Thống kê toán	Mathematical statistic
Phương pháp nghiên cứu toàn bộ	Total study
Phương pháp mẫu	Sampling
Thống kê mô tả	Descriptive statistic
Thống kê suy diễn	Inferential statistic
Tổng thể	Population
Bảng phân phối tần số	Frequence distribution table
Bảng phân phối tần suất	Relative-frequence distribution table
Tần số tích lũy	Cumulative frequency
Tần suất tích lũy	Cumulative relative-frequency
Trung bình tổng thể	Population mean
Trung bình điều hòa	Harmonical mean
Trung bình nhân	Geometrical mean
Phương sai tổng thể	Population variance
Độ lệch chuẩn tổng thể	Population standard deviation
Tần suất tổng thể	Population Relative-frequency
Mẫu ngẫu nhiên	Random sample
Mẫu đại diện	Representative sample
Mẫu cụ thể	Concret sample
Mẫu hoàn lại	Sampling with replacement
Mẫu không hoàn lại	Sampling without replacement
Mẫu ngẫu nhiên giản đơn	Simple random sample
Mẫu ngẫu nhiên hệ thống	Systematic random sample
Mẫu chùm	Quota sample
Mẫu phân tổ	Cluster sample
Mẫu phân tầng	Stratiffied sample
Thang định danh	Nominal scale
Thang thứ bậc	Ordinal scale
Thang đo khoảng	Interval scale
Thang tỷ lệ	Ratio scale
Đường đa giác	Diagram
Biểu đồ hình cột	Histogram
Biểu đồ hình bánh xe	Pie chart

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Đồ thị hình hộp	Box - plot
Thống kê	Statistic
Trung bình mẫu	Sample mean
Trung bình số học	Arithmetic mean
Trung bình có trọng số	Weight arithmetic mean
Kỳ vọng toán của trung bình mẫu	Expected value of sample mean
Phương sai của trung bình mẫu	Variance of sample mean
Sai số chuẩn của trung bình mẫu	Standard error of sample mean
Trung vị mẫu	Sample median
Mốt của mẫu	Sample mode
Khoảng biến thiên	Range
Khoảng tứ phân vị	Inter - quartile range
Tứ phân vị 1	Lower quartile
Tứ phân vị 3	Upper quartile
Tổng bình phương các sai lệch	Sum squares of variation
Độ lệch bình phương trung bình	Mean squares of variation
Phương sai mẫu	Sample variance
Độ lệch chuẩn mẫu	Sample standard deviation
Hệ số biến thiên của mẫu	Sample coefficient of variation
Hệ số bất đối xứng của mẫu	Sample Skewness
Hệ số nhọn của mẫu	Sample kurtosis
Tần suất mẫu	Sample relative frequency
Kỳ vọng toán của tần suất mẫu	Expected value of relative frequency
Phương sai của tần suất mẫu	Variance of relative frequency
Sai so chuẩn của tần suất mẫu	Standard error of relative frequency
Mẫu ngẫu nhiên hai chiều	Two dimention random sample
Phân phối thực nghiệm của X	Empiric distribution of X
Phân phối thực nghiệm của Y	Empiric distribution of Y
Phân phối thực nghiệm có điều kiện	Conditional empiric distribution
Trung bình có điều kiện	Conditional mean
Hệ số tương quan mẫu	Sample correlation coefficient
Suy diễn thống kê	Deduction statistatic
Quy nạp thống kê	Induction statistatic

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

6.43. Hãy phân biệt sự khác nhau giữa tổng thể và mẫu, giữa các tham số và các thống kê. Cho ví dụ.

6.44. Hãy phân biệt sự khác nhau giữa dấu hiệu nghiên cứu định lượng và định tính. Cho ví dụ. Các dấu hiệu định lượng và định tính có thể mô hình hóa bằng cách nào?

6.45. Thống kê mô tả và thống kê suy diễn khác nhau như thế nào? Trong nghiên cứu thống kê thì loại thống kê nào nằm ở dạng thức cao hơn? Vì sao?

6.46. Một giám đốc xí nghiệp muốn tìm hiểu xem trong một năm bình quân mỗi công nhân của xí nghiệp nghỉ việc bao nhiêu ngày. Song do xí nghiệp có tới hơn 2000 công nhân nên ông ta không thể đủ thời gian để xem sổ chấm công của từng người. Trong trường hợp này giám đốc có thể làm như thế nào để có được thông tin cần thiết?

6.47. Cho thí dụ về một mẫu ngẫu nhiên kích thước $n = 10$ và một mẫu cụ thể của mẫu ngẫu nhiên ấy.

6.48. Cho thí dụ về một mẫu ngẫu nhiên kích thước n được xây dựng từ biến ngẫu nhiên tuân theo quy luật không-một, quy luật nhị thức, quy luật chuẩn.

6.49. Khi nào thì mẫu lặp (có hoàn lại) và mẫu không lặp (không hoàn lại) có thể coi là như nhau?

6.50. Một mẫu ngẫu nhiên phải được chọn theo những nguyên tắc nào? Cho ví dụ.

6.51. Hãy liệt kê các đặc trưng về xu hướng trung tâm, độ phân tán và dạng phân phối của tổng thể và của mẫu.

6.52. Tại sao sau khi đã thu thập được số liệu về tổng thể hoặc mẫu lại phải sắp xếp chúng lại và mô tả theo một hình thức nào đó? Hãy phân tích ưu điểm của các hình thức mô tả sau:

a - Bảng phân phối tần số

b - Bảng phân phối tần số tích lũy

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

c - Bảng phân phối tần suất

d - Đồ thị hình cột

e - Đồ thị hình bánh xe

f - Đường đa giác

6.53. Với những điều kiện nào thì dùng trung bình điều hòa và trung bình nhân tốt hơn trung bình số học.

6.54. Với những điều kiện nào thì dùng trung vị tốt hơn trung bình để đặc trưng cho xu hướng trung tâm.

6.55. Đồ thị hình hộp chứa đựng những thông tin nào về dãy số liệu mà nó mô tả?

6.56. Nếu rút ra một mẫu kích thước n từ tổng thể thì kích thước mẫu sẽ ảnh hưởng như thế nào đến độ phân tán của trung bình mẫu? Độ phân tán của trung bình mẫu sẽ thay đổi như thế nào khi n tăng lên?

6.57. Tiến sĩ A được thuyên chuyển từ vụ thống kê sang vụ kinh tế của một bộ. Người phụ trách vụ thống kê cho rằng như vậy chỉ số IQ của vụ thống kê sẽ giảm đi còn chỉ số IQ của vụ kinh tế sẽ tăng lên. Với những điều kiện nào thì lời nhận xét trên là đúng?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập tổng hợp chương 6

6.58. Chiều cao của thanh niên ở một địa phương là biến ngẫu nhiên phân phối chuẩn với trung bình là 170cm và độ lệch chuẩn là 10cm. Chọn ngẫu nhiên 31 thanh niên ở vùng đó.

a. Tìm xác suất để chiều cao trung bình của số thanh niên nói trên không vượt quá 172cm.

b. Tìm xác suất để độ lệch chuẩn của chiều cao của số thanh niên nói trên lớn hơn 15cm.

6.59. Chiều dài của một loại sản phẩm được sản xuất hàng loạt là biến ngẫu nhiên phân phối chuẩn với $\mu = 100$ mm và $\sigma^2 = 4^2$. Kiểm tra ngẫu nhiên 25 sản phẩm. Khả năng chiều dài trung bình của số sản phẩm kiểm tra nằm trong khoảng từ 98 mm đến 101 mm là bao nhiêu?

ĐS: 0,8882

6.60. Người ta đề nghị 8 nhà phân tích tài chính dự báo chỉ số của thị trường chứng khoán trong tháng tới. Biết rằng kết quả dự báo là biến ngẫu nhiên phân phối chuẩn.

a. Với xác suất 0,05 phương sai của kết quả dự báo của các nhà phân tích trên có thể lớn hơn phương sai thực ít nhất bao nhiêu lần.

b. Tìm giá trị a sao cho với xác suất 0,9 tỷ số giữa phương sai mẫu và phương sai thực nằm giữa a và b, biết rằng b = 2,01.

6.61. Năng suất của hai giống lúa là hai biến ngẫu nhiên tuân theo quy luật chuẩn với $\mu_1 = 70$ tạ/ha, $\mu_2 = 60$ tạ/ha, $\sigma_1^2 = 10^2$, $\sigma_2^2 = 8^2$.

Nếu gặt thử mỗi giống lúa 25 thửa ruộng thì khả năng năng suất trung bình của 25 thửa ruộng trồng giống thứ nhất lớn hơn năng suất trung bình mẫu của giống thứ hai ít nhất là 5 tạ là bao nhiêu?

ĐS: 0,9744

6.62. Giả sử tỷ lệ người dân mua bảo hiểm nhân thọ trong một thành phố là 25%.

a. Tìm xác suất để có trên 28% số người trong một mẫu điều tra gồm 120 người của thành phố này mua bảo hiểm nhân thọ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b. Vẫn sử dụng mẫu 120 người nói trên với xác suất là 0,1 thì tần suất mẫu lớn hơn so với tỷ lệ của cả tổng thể một lượng tối thiểu là bao nhiêu.

6.63. Trọng lượng của một loại gia cầm là biến ngẫu nhiên phân phối chuẩn với trung bình là 2,5 kg/con. Tìm xác suất để lấy ngẫu nhiên từ đàn gia cầm ra 25 con thì trọng lượng trung bình của chúng nằm trong khoảng 2,4 đến 2,6 kg. Biết rằng với xác suất 0,9973 thì trọng lượng của loại gia cầm này nằm trong khoảng sai lệch so với trọng lượng trung bình là 0,3 kg.

6.64. Trọng lượng đóng bao gạo xuất khẩu là biến ngẫu nhiên phân phối chuẩn với trung bình (trọng lượng quy định) là 50 kg và độ lệch chuẩn cho phép là 0,5 kg. Nếu từ một lô hàng xuất khẩu đem cân ngẫu nhiên 16 bao thì với xác suất 0,95 trọng lượng trung bình của chúng chỉ được phép sai lệch so với trọng lượng quy định tối đa là bao nhiêu để có thể cho phép xuất khẩu lô hàng đó.

6.65. Kích thước chi tiết khi gia công là biến ngẫu nhiên phân phối chuẩn với độ lệch chuẩn cho phép là 0,1 cm. Nếu từ một lô chi tiết mới được gia công đem kiểm tra ngẫu nhiên 10 chi tiết thì với xác suất 0,99 độ lệch chuẩn tối đa của chúng có thể bằng bao nhiêu để có thể kết luận rằng lô chi tiết đó là đạt tiêu chuẩn.

6.66. Giả sử tỷ lệ phụ nữ trong độ tuổi sinh đẻ có áp dụng các biện pháp kế hoạch hoá gia đình ở một nước là 80%. Tại một địa phương, nếu điều tra ngẫu nhiên 200 phụ nữ trong độ tuổi sinh đẻ thì tối thiểu tỷ lệ áp dụng các biện pháp kế hoạch hoá gia đình trong số người đó phải là bao nhiêu để có thể kết luận công tác kế hoạch hoá gia đình ở địa phương đó đạt được mức chung của cả nước. Hãy kết luận với xác suất 0,9.

6.67. Nếu cho rằng tỷ lệ cử tri ủng hộ cho ứng cử viên A và B là như nhau thì khi phỏng vấn 2500 người thì khả năng tỷ lệ ủng hộ A và B khác biệt nhau không quá 1% là bao nhiêu?

ĐS: 0,6826

6.68. Tỷ lệ phế phẩm của dây chuyền sản xuất thứ nhất là 5% và của dây chuyền sản xuất thứ 2 là 7%. Nếu sản xuất thử trên mỗi dây chuyền 400 sản phẩm thì tỷ lệ phế phẩm của hai mẫu nói trên sai lệch nhau trên 6% với xác suất bằng bao nhiêu?

ĐS: 0,0087

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6.69. Hai cỗ máy cùng gia công một loại sản phẩm. Kích thước của sản phẩm là biến ngẫu nhiên tuân theo quy luật chuẩn. Nếu coi độ chính xác của 2 máy là như nhau thì trên thực tế tỷ số giữa hai độ lệch tiêu chuẩn của mẫy gồm 25 sản phẩm mà máy thứ nhất là 36 sản phẩm của máy thứ 2 không vượt quá giá trị nào với xác suất 0,95.

ĐS: 1,3838

6.70. Có các số liệu sau đây về các chỉ tiêu cơ bản của nền kinh tế Việt Nam giai đoạn 1980 - 1996

Năm	GDP	EX	IM	GIP	GAP
1990	16,8	0,3	1,3	6,8	9,7
1991	17,2	0,4	1,4	6,9	10,1
1992	18,7	0,5	1,5	7,5	11,2
1993	20,1	0,6	1,5	8,5	11,5
1994	21,8	0,6	1,7	9,6	12,2
1995	23,0	0,7	1,9	10,5	12,5
1996	23,8	0,8	2,2	11,2	13,1
1997	24,7	0,9	2,5	12,3	13,1
1998	25,9	1,0	2,8	14,0	13,7
1999	28,0	1,9	2,6	13,6	14,7
2000	29,5	2,4	2,8	14,0	14,9
2001	31,3	2,1	2,4	15,5	15,4
2002	34,0	2,6	2,6	18,1	16,6
2003	36,7	3,0	3,5	20,4	17,6
2004	40,0	4,1	5,9	23,2	18,6
2005	43,8	5,37	7,67	26,5	19,4
2006	47,9	7,1	11,0	30,23	20,35

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trong đó: GDP là tổng sản phẩm quốc nội

EX là tổng giá trị xuất khẩu.

IM là tổng giá trị nhập khẩu.

GIP là tổng giá trị sản xuất công nghiệp.

GAP là tổng giá trị sản xuất nông nghiệp.

a. Hãy vẽ các đường đa giác tần số mô tả sự biến động của các chỉ tiêu qua các năm.
b. Vẽ biểu đồ hình bánh xe mô tả tỷ trọng của EX và IM trong GDP cho năm 2006.
c. Xây dựng đồ thì hình hộp cho GIP và GAP và cho nhận xét.
d. Tìm trung bình, phương sai cho mỗi chỉ tiêu trong giai đoạn nói trên và cho nhận xét.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như. *Thống kê toán học. NXB Đại học và THCN*. Hà Nội. 1984
2. Nguyễn Tấn Lập - *Giáo trình thống kê toán*. ĐH Kinh tế Kế hoạch. *Hà Nội*. 1972
3. Nguyễn Đình Cử, Nguyễn Cao Văn. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Thống kê. Hà Nội. 1991.
4. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà Nội.2009.
5. Ngô Văn Thứ. *Thống kê thực hành*. NXB Khoa học và kỹ thuật. Hà Nội. 2008.
6. Ngô Văn Thứ. *Thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội. 2011
7. Harald Cramer. *Phương pháp toán học trong thống kê*. NXB Khoa học và Kỹ thuật. Hà Nội. 1970
8. Гмурман В.Е. Теория вероятностей и математическая статистика. Москва. 1972.
9. Уйякс. Математическая статистика. Москва. 1982.
10. Bencrézi J.P. Analyse des données. Dunod. Paris. 1973
11. Baillageon G et Rainville J. Introduction à la Statistique appliquée. Québec. 1976
12. Giard V. Statistique appliquée à la gestion. Economica. Paris. 1992
13. Michel Volle: Analyse des donnees. Economica. 1994.
14. Amir D Aczel. Complete Business Statistics. Irwin Boston 1993.
15. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992
16. OTT L.R. And introduction to statistical methods and data analysis. California. 1993
17. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons. New york. 1990

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 7

ƯỚC LƯỢNG CÁC THAM SỐ CỦA BIẾN NGẪU NHIÊN

1. MỞ ĐẦU

Giả sử cần phải nghiên cứu dấu hiệu χ trong tổng thể. Như đã phân tích ở chương 6, một trong những mục tiêu cơ bản của việc nghiên cứu là xác định các tham số đặc trưng của tổng thể như trung bình, phương sai, cơ cấu của tổng thể theo dấu hiệu nghiên cứu. Đó là những chỉ tiêu tổng hợp để phân tích tổng thể cần nghiên cứu.

Nếu dấu hiệu nghiên cứu trong tổng thể có thể xem như một biến ngẫu nhiên X và giả sử bằng phân tích lý thuyết đã xác định được dạng phân phối xác suất của nó thì vấn đề xác định các tham số đặc trưng của tổng thể sẽ được quy về bài toán xác định các tham số đặc trưng của quy luật phân phối xác suất xác định biến ngẫu nhiên gốc X. Chẳng hạn, nếu đã biết được rằng dấu hiệu nghiên cứu trong tổng thể có thể cỏ thể xem như biến ngẫu nhiên phân phối theo quy luật chuẩn thì bài toán đặt ra là phải ước lượng (tức là xác định một cách gần đúng) các tham số kỳ vọng toán μ và phương sai σ^2 của nó vì các tham số trên hoàn toàn xác định quy luật phân phối chuẩn và thực chất chúng chính là trung bình và phương sai của tổng thể.

Như vậy, bài toán ước lượng tham số có thể phát biểu như sau: Cho biến ngẫu nhiên X với quy luật phân phối xác suất đã biết song chưa biết tham số θ nào đó của nó. Phải ước lượng (xác định một cách gần đúng) giá trị θ.

Phương pháp mẫu cho phép giải quyết bài toán trên bằng quy nạp thống kê như sau: Từ tổng thể nghiên cứu rút ra một mẫu ngẫu nhiên kích thước n và dựa vào đó mà xây dựng một thống kê $\hat{\theta}$ dùng để ước lượng θ bằng cách này hay cách khác. Có hai phương pháp sử dụng $\hat{\theta}$ để ước lượng θ là phương pháp ước lượng điểm và phương pháp ước lượng bằng khoảng tin cậy. Sau đây ta sẽ xem xét những nội dung chính của hai phương pháp đó. Mục 2 sẽ nghiên cứu hai phương pháp cơ bản của ước lượng điểm là phương pháp hàm ước lượng và phương pháp ước lượng hợp lý tối đa. Mục

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3 sẽ nghiên cứu một số nội dung quan trọng nhất của phương pháp ước lượng bằng khoảng tin cậy.

2. PHƯƠNG PHÁP ƯỚC LƯỢNG ĐIỂM

Phương pháp ước lượng điểm chủ trương dùng một giá trị để thay thế cho tham số θ chưa biết của tổng thể, vì bản thân $\hat{\theta}$ là một số xác định. Thông thường giá trị được chọn là một thống kê $\hat{\theta}$ nào đó của mẫu ngẫu nhiên. Có nhiều cách chọn thống kê $\hat{\theta}$ khác nhau tạo nên những phương pháp ước lượng điểm khác nhau.

2.1. Phương pháp hàm ước lượng (phương pháp mômen)

1. Khái niệm: Giả sử cần ước lượng tham số θ của biến ngẫu nhiên gốc X. Từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, \ldots, X_n)$$

Chọn lập thống kê $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$ mà thực chất là một thống kê đặc trưng mẫu tương ứng với tham số θ cần ước lượng. Chẳng hạn, để ước lượng kỳ vọng toán m của biến ngẫu nhiên gốc thì chọn thống kê trung bình mẫu $\overline{X}$, để ước lượng phương sai σ^2 của biến ngẫu nhiên gốc thì chọn thống kê phương sai mẫu S^2... Nếu lập một mẫu cụ thể và tính được giá trị $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$ của thống kê $\hat{\theta}$ trên mẫu cụ thể đó thì ước lượng của θ là giá trị $\hat{\theta}$ vừa tính.

Vì thống kê $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$ thực chất là hàm của các biến ngẫu nhiên nên nó được gọi là hàm ước lượng của θ.

Chất lượng của ước lượng không thể đánh giá qua một giá trị cụ thể của $\hat{\theta}$ vì như vậy chỉ có cách so sánh trực tiếp $\hat{\theta}$ và θ mà θ lại chưa biết. Do đó chỉ có thể đánh giá chất lượng của ước lượng thông qua bản thân thống kê $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$. Rõ ràng là có vô số cách chọn hàm f, tức là có vô số thống kê $\hat{\theta}$ có thể dùng làm ước lượng của θ. Vì vậy, cần đưa ra các tiêu chuẩn để đánh giá chất lượng của các thống kê $\hat{\theta}$, từ đó lựa chọn được thống kê " xấp xỉ một cách tốt nhất" tham số cần ước lượng. Dưới đây ta xét các tiêu chuẩn đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2. Các tiêu chuẩn lựa chọn hàm ước lượng

a.Ước lượng không chệch

Giả sử thống kê $\hat{\theta}$ là ước lượng của tham số θ của biến ngẫu nhiên gốc. Với k mẫu cụ thể rút ra từ tổng thể, thống kê $\hat{\theta}$ sẽ nhận k giá trị cụ thể tương ứng là $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$. Nếu thống kê $\hat{\theta}$ là một ước lượng thừa của θ thì mọi giá trị $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ cũng đều sẽ lớn hơn θ và giá trị trung bình của chúng (tức là kỳ vọng toán của $\hat{\theta}$) cũng sẽ lớn hơn θ: $E(\hat{\theta}) > \theta$. Ngược lại, nếu $\hat{\theta}$ là một ước lượng thiếu của θ thì mọi giá trị của $\hat{\theta}_i (i = \overline{1,k})$ cũng đều sẽ nhỏ hơn θ nên kỳ vọng toán của $\hat{\theta}$ cũng sẽ nhỏ hơn θ: $E(\hat{\theta}) < \theta$.

Như vậy, việc sử dụng một thống kê mà kỳ vọng toán của nó khác với tham số cần ước lượng có thể dẫn đến các sai số hệ thống (tất cả các giá trị của $\hat{\theta}$ đều lớn hơn hoặc nhỏ hơn θ). Vì lý do đó mà đương nhiên ta sẽ yêu cầu là phải tìm được một ước lượng $\hat{\theta}$ sao cho kỳ vọng toán của nó đúng bằng tham số cần ước lượng. Dĩ nhiên yêu cầu trên không loại trừ được hoàn toàn các sai số, song như vậy các sai số khác dấu sẽ xuất hiện tương đối đều nhau, do đó các giá trị của $\hat{\theta}$ sẽ không bị lệch hẳn về một phía so với θ. Từ đó ta có định nghĩa sau:

Định nghĩa. Thống kê $\hat{\theta}$ của mẫu được gọi là ước lượng không chệch của tham số θ của biến ngẫu nhiên gốc X nếu:

$$E(\hat{\theta}) = \theta$$

Ngược lại, nếu $E(\hat{\theta}) \neq \theta$ thì $\hat{\theta}$ được gọi là ước lượng chệch của θ.

Rõ ràng từ phân tích ở trên, ta nên dùng loại ước lượng không chệch.

Chú ý rằng $\hat{\theta}$ là ước lượng không chệch của θ không có nghĩa là mọi giá trị của $\hat{\theta}$ đều trùng khít với θ mà chỉ có nghĩa: Trung bình các giá trị của $\hat{\theta}$ bằng θ. Từng giá trị của $\hat{\theta}$ có thể sai lệch rất lớn so với θ.

Dựa vào các kết quả đã chứng minh được ở mục 5 chương VI có thể rút ra một số kết luận sau:

- Trung bình mẫu $\overline{X}$ là ước lượng không chệch của kỳ vọng toán m của biến ngẫu nhiên gốc $[E(\overline{X}) = m]$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Tần suất mẫu f là ước lượng không chệch của xác suất p của biến ngẫu nhiên gốc [E(f) = p].

- Phương sai mẫu S^2 và phương sai S^{*2} đều là các ước lượng không chệch của phương sai σ^2 của biến ngẫu nhiên gốc:

$$E(S^2) = \sigma^2 \text{ và } E(S^{*2}) = \sigma^2 \quad (7.1)$$

Nếu $\tilde{\theta}$ là một ước lượng chệch của θ song thỏa mãn điều kiện

$$\lim_{n\to\infty} E(\tilde{\theta}) = \theta \quad (7.2)$$

thì $\tilde{\theta}$ được gọi là ước lượng tiệm cận không chệch của θ. Độ chệch của một ước lượng được đo bằng giá trị:

$$BS = \left|E(\hat{\theta}) - \theta\right| \quad (7.3)$$

Hiển nhiên là trong số các ước lượng chệch thì nên chọn ước lượng nào có BS nhỏ nhất.

b. Ước lượng hiệu quả

Như đã phân tích ở trên, dù $\hat{\theta}$ là ước lượng không chệch của θ thì từng giá trị của $\hat{\theta}$ vẫn có thể sai lệch rất lớn so với θ, tức là phương sai $V(\hat{\theta})$ vẫn có thể lớn. Lúc đó, nếu lấy một giá trị của $\hat{\theta}$ tìm được trên một mẫu cụ thể, chẳng hạn $\hat{\theta}_1$ để ước lượng θ thì nó có thể sai lệch rất nhiều so với giá trị trung bình $\hat{\theta}$ tức là bản thân tham số θ cần ước lượng. Như vậy, nếu lấy một giá trị của $\hat{\theta}$, chẳng hạn $\hat{\theta}_1$ để ước lượng θ, thì có thể mắc sai số rất lớn. Còn nếu như đòi hỏi phương sai của $\hat{\theta}$ phải nhỏ thì có thể hạn chế được loại sai số này trong ước lượng. Từ đó ta có định nghĩa sau:

Định nghĩa: Thống kê $\hat{\theta}$ của mẫu được gọi là ước lượng hiệu quả nhất của tham số θ của biến ngẫu nhiên gốc X nếu nó là ước lượng không chệch và có phương sai nhỏ nhất so với mọi ước lượng không chệch khác được xây dựng trên cùng mẫu đó.

Như vậy, để xét xem $\hat{\theta}$ có phải là ước lượng hiệu quả nhất của θ hay không, cần phải tìm được giá trị nhỏ nhất có thể có của phương sai các hàm

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ước lượng. Nếu $\hat{\theta}$ đã là một ước lượng không chệch của θ thì trong nhiều trường hợp giá trị nhỏ nhất của phương sai $V(\hat{\theta})$ có thể tìm được dựa vào *bất đẳng thức Cramer - Rao* được phát biểu như sau:

Cho mẫu ngẫu nhiên $W = (X_1, X_2, \ldots, X_n)$ được xây dựng từ biến ngẫu nhiên gốc X *có hàm mật độ xác suất (hay biểu thức xác suất)* $f(x, \theta)$ *thỏa mãn một số điều kiện nhất định* (chẳng hạn như các phân phối xác suất đã xét ở chương III đều thỏa mãn các điều kiện này) và θ^* là một ước lượng không chệch bất kỳ của θ thì

$$V(\theta^*) \geq \frac{1}{nE\left[\frac{\partial \ln f(x,\theta)}{\partial \theta}\right]^2} \tag{7.4}$$

Chẳng hạn, dựa vào bất đẳng thức trên ta có thể chứng minh được rằng trung bình mẫu $\overline{X}$ là ước lượng hiệu quả nhất của kỳ vọng toán μ của biến ngẫu nhiên gốc X trong tổng thể khi X phân phối chuẩn $N(\mu, \sigma^2)$.

Thật vậy, ở mục 5 chương VI ta đã chứng minh được $V(\overline{X}) = \frac{\sigma^2}{n}$. Mặt khác, do $\overline{X}$ là ước lượng không chệch của μ và do X phân phối $N(\mu, \sigma^2)$ nên ta có:

$$f(x,\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\ln f(x,\theta) = -\ln\sqrt{2\pi}\sigma - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \ln f(x,\theta)}{\partial \theta} = \frac{\partial \ln f(x,\mu)}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

Vậy

$$nE\left[\frac{\partial \ln f(x,\theta)}{\partial \theta}\right]^2 = nE\left(\frac{x-\mu}{\sigma^2}\right)^2 = \frac{n}{\sigma^4}E(x-\mu)^2$$

$$= \frac{n}{\sigma^4}V(X) = \frac{n\sigma^2}{\sigma^4} = \frac{n}{\sigma^2}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Như vậy, $V(\overline{X})$ bằng biểu thức ở vế phải của bất đẳng thức Cramer - Rao. Vậy $\overline{X}$ là ước lượng hiệu quả nhất của μ.

Ta cũng có thể chứng minh được rằng tần suất mẫu f là ước lượng hiệu quả nhất của xác suất p của biến ngẫu nhiên gốc X phân phối theo quy luật A(p). Thật vậy, ở mục 5 chương VI ta đã có $V(f) = \frac{p(1-p)}{n}$. Mặt khác, do f là ước lượng không chệch của p và do X phân phối A(p) nên:

$$f(x,\theta) = P(X = x) = p^x(1-p)^{1-x}$$

$$\ln f(x,\theta) = x\ln p + (1-x)\ln(1-p)$$

Lấy đạo hàm của biểu thức trên theo p ta được

$$\frac{\partial \ln f(x,\theta)}{\partial\theta} = \frac{x}{p} - \frac{1-x}{1-p}$$

do đó

$$\left(\frac{x}{p} - \frac{1-x}{1-p}\right)^2 = \left[\frac{x-p}{p(1-p)}\right]^2$$

và

$$nE\left[\frac{\partial \ln f(x,\theta)}{\partial\theta}\right]^2 = nE\left[\frac{x-p}{p(1-p)}\right]^2 = \frac{n}{p^2(1-p)^2}E(x-p)^2 =$$

$$= \frac{n}{p^2(1-p)^2}V(X) = \frac{n}{p^2(1-p)^2}p(1-p) = \frac{n}{p(1-p)}$$

Vậy V(f) bằng biểu thức ở vế phải của bất đẳng thức Cramer - Rao hay f là ước lượng hiệu quả nhất của p.

Khi hai ước lượng $\hat{\theta}_1$ và $\hat{\theta}_2$ nào đó đều là các ước lượng không chệch của θ song không phải là ước lượng hiệu quả nhất thì có thể so sánh phương sai của hai ước lượng đó để tìm ra ước lượng hiệu quả hơn. Giả sử $V(\hat{\theta}_1) > V(\hat{\theta}_2)$, lúc đó độ hiệu quả của $\hat{\theta}_2$ so với $\hat{\theta}_1$ được xác định bằng biểu thức:

$$EF = \frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)} \qquad (7.5)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ. Từ một ngẫu nhiên kích thước n = 2 ta xét hai ước lượng sau đây của trung bình tổng thể m

$$\overline{X} = \frac{1}{2}(X_1 + X_2)$$

và $$X' = \frac{1}{3}X_1 + \frac{2}{3}X_2$$

a. Xét xem $\overline{X}$ và X' có phải là ước lượng không chệch của m hay không?

b. Ước lượng nào hiệu quả hơn.

Giải:

a. Ta có

$$E(\overline{X}) = E\left(\frac{1}{2}X_1 + \frac{1}{2}X_2\right) = \frac{1}{2}E(X_1) + \frac{1}{2}E(X_2) = \frac{1}{2}m + \frac{1}{2}m = m$$

$$E(X') = E\left(\frac{1}{3}X_1 + \frac{2}{3}X_2\right) = \frac{1}{3}E(X_1) + \frac{2}{3}E(X_2) = \frac{1}{3}m + \frac{2}{3}m = m$$

Vậy cả $\overline{X}$ và X' đều là các ước lượng không chệch của m.

b. Ta có

$$V(\overline{X}) = V\left(\frac{1}{2}X_1 + \frac{1}{2}X_2\right) = \frac{1}{4}V(X_1) + \frac{1}{4}V(X_2) = \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{\sigma^2}{2}$$

$$V(X') = V\left(\frac{1}{3}X_1 + \frac{2}{3}X_2\right) = \frac{1}{9}V(X_1) + \frac{4}{9}V(X_2) = \frac{1}{9}\sigma^2 + \frac{4}{9}\sigma^2 = \frac{5\sigma^2}{9}$$

$V(\overline{X}) < V(X')$. Vậy $\overline{X}$ hiệu quả hơn

$\Rightarrow$ EF của $\overline{X}$ so với X' là:

$$EF = \frac{(5/9)\sigma^2}{(1/2)\sigma^2} = \frac{10}{9} = 1,11 = 111\%$$

Trong thực tế đôi khi người ta còn xây dựng một tổ hợp tuyến tính của hai ước lượng không chệch $\hat{\theta}_1$ và $\hat{\theta}_2$ dạng $\hat{\theta} = \alpha\hat{\theta}_1 + (1-\alpha)\hat{\theta}_2$ nhằm thu được một ước lượng hiệu quả hơn cho θ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

c. Ước lượng vững: Khi xét những mẫu có kích thước lớn thì nảy sinh vấn đề là mẫu càng lớn thì thống kê $\hat{\theta}$ của mẫu phải càng gần tham số θ cần ước lượng. Từ đó ta có định nghĩa:

Định nghĩa. Thống kê $\hat{\theta}$ của mẫu được gọi là ước lượng vững của tham số θ của biến ngẫu nhiên gốc X nếu $\hat{\theta}$ hội tụ theo xác suất đến θ khi $n \to \infty$.

Tức là với mọi ε dương bé tùy ý ta luôn có:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right| < \varepsilon\right) = 1$$

Như vậy theo luật số lớn của Trêbusép (trường hợp riêng) và luật số lớn của Bernoulli (xem chương V) ta suy ra ngay là trung bình mẫu $\overline{X}$ là ước lượng vững của tham số m của biến ngẫu nhiên gốc X và tần suất mẫu f là ước lượng vững của xác suất p của biến ngẫu nhiên gốc X.

Trong trường hợp $\hat{\theta}$ là ước lượng không chệch của θ thì để tìm ước lượng vững có thể sử dụng định lý sau: *Nếu* $\hat{\theta}$ là ước lượng không chệch của θ và $\lim_{n\to\infty} V(\hat{\theta}) = 0$ thì $\hat{\theta}$ là ước lượng vững của θ.

d. Ước lượng đủ: Một ước lượng $\hat{\theta}$ được gọi là ước lượng đủ nếu nó chứa đựng toàn bộ các thông tin trong mẫu về tham số θ của ước lượng.

Chẳng hạn, trung bình mẫu và trung vị mẫu đều là các ước lượng không chệch của trung bình tổng thể song trung bình mẫu là ước lượng đủ của trung bình tổng thể vì nó sử dụng toàn bộ thông tin của mẫu, còn trung vị mẫu không phải là ước lượng đủ vì nó chỉ dùng đến giá trị chính giữa của dãy số liệu mẫu mà thôi.

Hiển nhiên là trong ước lượng ta sẽ chủ trương sử dụng các ước lượng đủ nếu điều đó cho phép.

Mặt khác, trong thực tế $\hat{\theta}$ có thể là ước lượng tuyến tính hoặc phi tuyến tính của các giá trị của mẫu. Nếu $\hat{\theta}$ là ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất thì nó được gọi là ước lượng tuyến tính không chệch tốt nhất (BLUE) của θ. Còn nếu $\hat{\theta}$ là ước lượng phi tuyến song vẫn thỏa mãn điều kiện không chệch và có phương sai nhỏ nhất thì được gọi là ước lượng không chệch tốt nhất (BUE).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3. Một vài kết luận của phương pháp hàm ước lượng

Dùng những tiêu chuẩn nêu trên để đánh giá các thống kê đặc trưng mẫu khác nhau cho phép ta lựa chọn được những thống kê tốt nhất, tức là ước lượng một cách chính xác nhất các tham số đặc trưng của tổng thể. Có thể đưa ra một số kết luận chung sau đây:

- Vì trung bình mẫu $\overline{X}$ là ước lượng không chệch, hiệu quả nhất và vững của trung bình tổng thể m và đồng thời là ước lượng tuyến tính không chệch tốt nhất, do đó nếu chưa biết m có thể dùng $\overline{X}$ để ước lượng nó.

- Vì tần suất mẫu f là ước lượng không chệch, hiệu quả nhất và vững của tần suất tổng thể p và đồng thời là ước lượng tuyến tính không chệch tốt nhất, do đó nếu chưa biết p có thể dùng f để ước lượng nó.

- Vì phương sai mẫu S^2 và phương sai S^{*2} đều là các ước lượng không chệch của phương sai tổng thể σ^2 do đó nếu chưa biết phương sai σ^2 có thể dùng S^2 hoặc S^{*2} để ước lượng nó.

Chú ý rằng MS và phương sai mẫu S^2 chỉ khác nhau chút ít bởi hệ số $\frac{n-1}{n}$. Khi n lớn thì sự sai khác này là không đáng kể. Trong thực tế phương sai mẫu S^2 được sử dụng khi $n < 30$.

Bài tập

7.1. Rút ra một mẫu kích thước n = 100 từ tổng thể có trung bình là 1.065 và độ lệch chuẩn là 500. Tìm kỳ vọng toán và phương sai của trung bình mẫu.

7.2. Giả sử lấy ra một mẫu từ tổng thể có phương sai bằng 1000000. Nếu muốn sai số chuẩn của trung bình mẫu không quá 25 thì kích thước mẫu điều tra phải bằng bao nhiêu.

7.3. Nếu lấy mẫu kích thước n = 120 từ tổng thể có xác suất p = 0,5 thì sai số chuẩn của tần suất mẫu bằng bao nhiêu.

7.4. Từ tổng thể lấy ra một mẫu kích thước n = 5 và thu được các giá trị sau: 10, 12, 16, 18, 19. Hãy tìm ước lượng điểm bằng phương pháp hàm ước lượng của trung bình tổng thể.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7.5. Kiểm tra ngẫu nhiên 50 đĩa mềm từ một lô đĩa mềm mới sản xuất thấy có 3 cái hỏng. Vậy có thể cho rằng tỷ lệ đĩa mềm hỏng của lô sản phẩm đó bằng bao nhiêu nếu dùng phương pháp hàm ước lượng.

7.6. Giả sử có hai ước lượng chệnh của tham số θ cần ước lượng của tổng thể. Tham số thứ nhất có BS bằng 1/n với n là kích thước mẫu, còn tham số thứ hai có BS bằng 0,01. Với điều kiện nào thì ước lượng thứ nhất tốt hơn ước lượng thứ hai.

7.7. Một nhà kinh tế lấy ra một mẫu kích thước n = 500 quan sát từ tổng thể nghiên cứu song lại đánh rơi mất 180 số liệu sau cùng. Lúc đó để ước lượng trung bình của tổng thể thì nên sử dụng trung bình của 320 quan sát còn lại hay trung vị của mẫu.

7.8. Từ một tổng thể phân phối chuẩn rút ra một mẫu ngẫu nhiên kích thước n. Tìm độ hiệu quả của trung vị so với trung bình mẫu, từ đó hãy cho biết xem để ước lượng trung bình tổng thể thì nên dùng trung bình hay trung vị mẫu.

ĐS: 64%. Dùng trung bình mẫu

7.9. Để xác định diện tích của một cánh đồng hình chữ nhật trong đó kết quả đo chiều dài X và chiều rộng Y là các biến ngẫu nhiên độc lập có phân phối xác suất như nhau:

x	8	10	11
P(x)	1/4	1/4	1/2

y	4	6
P(y)	1/2	1/2

Diện tích Z = X.Y cũng là một biến ngẫu nhiên và được dùng để ước lượng diện tích thực của thửa ruộng đó. Vậy nếu chiều dài và chiều rộng thực của thửa ruộng đó là 10 và 5 thì:

a. X và Y có phải là các ước lượng không chệch của chiều dài thực và chiều rộng thực của thửa ruộng đó hay không.

b. Z có phải là một ước lượng không chệch của diện tích thực của thửa ruộng đó hay không.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7.10. Hai mẫu kích thước n = 3 và n = 4 được rút ra từ một tổng thể phân phối A(p).

a. Hãy chứng tỏ rằng các tần suất mẫu f_1 và f_2 đều là các ước lượng không chệch của p.

b. Ước lượng nào hiệu quả hơn. Tìm độ hiệu quả của f_2 so với f_1.

c. Trong lớp các ước lượng tuyến tính $\alpha f_1 + (1 - \alpha)f_2$ hãy tìm một ước lượng hiệu quả hơn cả của p.

7.11. Giả sử có hai nhà kinh tế định ước lượng mức chi tiêu trung bình của các gia đình cho thực phẩm và họ dùng hai ước lượng không chệch (và độc lập nhau) là U và V. Do nhà kinh tế thứ hai kém cẩn thận hơn nên độ lệch chuẩn của V lớn gấp 3 lần độ lệch chuẩn của U. Để kết hợp hai ước lượng lại sao cho thu được một ước lượng chung của m người ta đề nghị ba cách sau.

i. $W_1 = \frac{1}{2}U + \frac{1}{2}V$

ii. $W_2 = \frac{3}{4}U + \frac{1}{4}V$

iii. $W_3 = 1.U + 0.V$

a. Trong các ước lượng trên ước lượng nào là không chệch.

b. Trong các ước lượng trên ước lượng nào tốt nhất (hiệu quả hơn cả).

2.2. Phương pháp ước lượng hợp lý tối đa

Giả sử đã biết quy luật phân phối xác suất tổng quát của biến ngẫu nhiên gốc X dưới dạng hàm mật độ $f(x,\theta)$. Đó cũng có thể là biểu thức xác suất nếu X là biến ngẫu nhiên rời rạc. Cần phải ước lượng tham số θ nào đó của X. Lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

và xây dựng hàm của đối số θ tại một giá trị cụ thể của mẫu:

$$L(x_1, x_2, ..., x_n, \theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdots f(x_n, \theta).$$

Hàm L được gọi là *hàm hợp lý* của tham số θ. Giá trị của hàm hợp lý chính là xác suất hay mật độ xác suất tại điểm $(x_1, x_2, ..., x_n)$, còn giá trị của thống kê θ tại điểm đó

$$\hat{\theta} = f(x_1, ..., x_n)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

được gọi là ước lượng hợp lý tối đa của θ nếu ứng với giá trị này của θ, hàm hợp lý đạt cực đại.

Do hàm L và lnL đạt cực đại tại cùng một giá trị của θ nên có thể tìm giá trị của θ để lnL đạt cực đại với các bước sau:

a. Tìm đạo hàm bậc nhất của lnL theo θ.

b. Giải phương trình $\frac{d\ln L}{d\theta} = 0$

Giả sử nó có nghiệm $\theta = \hat{\theta} = f(x_1, x_2, ..., x_n)$

c. Tìm đạo hàm bậc hai $\frac{d^2 \ln L}{d\theta^2}$

Nếu tại điểm $\theta = \hat{\theta}$ đạo hàm bậc hai âm thì tại điểm này hàm lnL đạt cực đại, do đó $\hat{\theta} = f(x_1, x_2, ..., x_n)$ là ước lượng điểm hợp lý tối đa cần tìm của θ. Chú ý rằng đối số của hàm hợp lý là θ chứ không phải là $(x_1, x_2, ..., x_n)$ do đó nếu thay giá trị của mẫu bằng bản thân mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ thì kết quả thu được vẫn đúng. Từ đó sẽ thu đươc kết quả tổng quát hơn, tức là tìm được hàm ước lượng hợp lý tối đa.

Thí dụ 1. Tìm ước lượng hợp lý tối đa của tham số p trong quy luật phân phối A(p).

Biểu thức xác suất tổng quát của quy luật không-một là $P_x = p^x (1-p)^{1-x}$

Lập mẫu ngẫu nhiên kích thước n: $W = (X_1, X_2, ..., X_n)$ và tại giá trị $w = (x_1, x_2, ..., x_n)$ của nó lập hàm hợp lý với mỗi giá trị x_i ($x_i = 0$ hoặc 1)

$$L(x_1, x_2, ..., x_n, p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

Từ đó

$$\ln L = \sum_{i=1}^{n} [x_i \ln p + (1 - x_i) \ln(1-p)]$$

$$\frac{d\ln L}{dp} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \sum_{i=1}^{n} (1 - x_i)$$

$$\frac{d \ln L}{dp} = 0 \Rightarrow p = \bar{x} = f$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Dễ thấy rằng tại $p = \overline{X}$ thì $\frac{d^2 \ln L}{dp^2} = -\frac{n}{p(1-p)} < 0$ do đó ước lượng hợp lý tối đa của p là f.

Thí dụ 2. Tìm ước lượng hợp lý tối đa của các tham số μ và σ^2 của biến ngẫu nhiên X tuân theo quy luật chuẩn.

Trong trường hợp quy luật chuẩn thì hàm hợp lý có dạng:

$$L(x_1, x_2, ..., x_n, \mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-1/2\sigma^2 \sum_{i=1}^{n} (x_i - \mu)^2}$$

Từ đó

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Tìm các đạo hàm riêng của lnL theo μ và σ^2 và cho bằng không ta có:

$$\begin{cases} \dfrac{\partial \ln L}{\partial \mu} = \dfrac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2} = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^4} = 0 \end{cases}$$

Giải phương trình thứ nhất thu được $\mu = \overline{X}$ và thế kết quả này vào phương trình thứ hai thu được $\sigma^2 = MS$.

Sau nữa tại các điểm $\mu = \overline{X}$ và $\sigma^2 = MS$ ta có:

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0$$

$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{2\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(n\sigma^2) = -\frac{n}{2\sigma^4} < 0$$

$$\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} = \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) = 0$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$\Rightarrow \frac{\partial^2 \ln L}{\partial \mu^2}\frac{\partial^2 \ln L}{\partial \sigma^2} - \left(\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2}\right)^2 = \frac{n^2}{2\sigma^6} > 0$$

Vậy ước lượng hợp lý tối đa của μ là $\overline{X}$, còn của σ^2 là MS. Vậy ước lượng hợp lý tối đa của σ^2 là ước lượng chệch.

Bài tập

7.12. Hãy tìm ước lượng hợp lý tối đa.

a. Cho trung bình tổng thể biết rằng mẫu kích thước n = 4 được rút ra từ tổng thể phân phối chuẩn và thu được các giá trị là 17, 28, 92, 41.

b. Cho tỷ lệ phế phẩm của một lô hàng biết rằng kiểm tra ngẫu nhiên 20 sản phẩm của lô hàng đó thì có 8 phế phẩm.

ĐS: a. $\bar{x} = 44,5$; b. f = 0,4

7.13. Cho biến ngẫu nhiên X trong tổng thể phân phối Poisson với tham số là λ. Để ước lượng λ người ta điều tra một mẫu kích thước n = 3 và thu được các giá trị sau: 15, 8, 13.

a. Viết biểu thức của hàm hợp lý.

b. Tìm giá trị của hàm hợp lý khi λ lần lượt nhận các giá trị bằng 5, 10, 15, 20, 25.

c. Từ kết quả thu được hãy cho biết ước lượng hợp lý tối đa của λ bằng bao nhiêu.

d. Hãy tìm ước lượng hợp lý tối đa của λ của phân phối Poisson trong trường hợp chung.

7.14. Cho biến ngẫu nhiên X trong tổng thể phân phối luỹ thừa với tham số là λ. Để ước lượng λ người ta điều tra một mẫu kích thước n = 5 và thu được các giá trị 1,2; 7,5; 1,8; 3,7; và 0,8.

a. Tìm giá trị của hàm hợp lý khi λ lần lượt bằng 0,1; 0,2; 0,3; 0,4; 0,5 và từ đó xác định xem ước lượng hợp lý tối đa của λ bằng bao nhiêu.

b. Tìm ước lượng hợp lý tối đa cho tham số λ của phân phối luỹ thừa trong trường hợp chung.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7.15. Cho biến ngẫu nhiên X trong tổng thể phân phối nhị thức với các tham số n và p chưa biết. Để ước lượng p người ta lấy ra một mẫu kích thước n = 5 và tìm được tần suất mẫu là 0,6.

a. Viết biểu thức của hàm hợp lý và tính giá trị của nó với p lần lượt bằng 0; 0,1; 0,2; ...; 0,9; 1 và mô tả lên đồ thị.

b. Từ đồ thị hãy chỉ ra xem ước lượng hợp lý tối đa của p bằng bao nhiêu.

c. Hãy chứng tỏ rằng trong trường hợp chung, ước lượng hợp lý tối đa của tham số p của quy luật nhị thức là tần suất mẫu f.

3. PHƯƠNG PHÁP ƯỚC LƯỢNG BẰNG KHOẢNG TIN CẬY

3.1. Khái niệm

Phương pháp ước lượng điểm nói trên có một nhược điểm cơ bản là khi kích thước mẫu nhỏ thì ước lượng điểm tìm được có thể sai lệch rất nhiều so với giá trị của tham số cần ước lượng, tức là sai số của ước lượng có thể rất lớn. Mặt khác dùng các phương pháp trên không thể đánh giá được khả năng mắc sai lầm khi ước lượng bằng bao nhiêu. Do đó khi kích thước mẫu nhỏ người ta thường sử dụng phương pháp *ước lượng bằng khoảng tin cậy.*

Để ước lượng tham số θ của biến ngẫu nhiên gốc X trong tổng thể, phương pháp này chủ trương từ một thống kê G nào đó của mẫu xây dựng một khoảng giá trị (G_1, G_2) sao cho với một xác suất cho trước tham số θ sẽ rơi vào khoảng (G_1, G_2) đó. Chú ý rằng do G là biến ngẫu nhiên nên khoảng (G_1, G_2) cũng là một khoảng ngẫu nhiên, còn θ lại là một số xác định nên phải nói chính xác hơn là khoảng (G_1, G_2) sẽ chứa đựng giá trị θ với một xác suất cho trước. Từ đó ta có định nghĩa sau:

Định nghĩa. Khoảng (G_1, G_2) *của thống kê G được gọi là khoảng tin cậy của tham số* θ *nếu với xác suất bằng* $(1-\alpha)$ *cho trước thỏa mãn điều kiện*

$$P(G_1 < \theta < G_2) = 1 - \alpha \qquad (7.6)$$

xác suất $(1-\alpha)$ được gọi là độ tin cậy của ước lượng, còn $I = (G_2 - G_1)$ được gọi là độ dài khoảng tin cậy.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Như vậy, vấn đề chủ yếu của phương pháp ước lượng bằng khoảng tin cậy là làm thế nào để xác định được khoảng tin cậy (G_1, G_2) thỏa mãn (7.6). Để làm điều đó người ta tiến hành như sau:

Từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

và từ đó xây dựng thống kê $G = f(X_1, X_2, ..., X_n, \theta)$ sao cho quy luật phân phối xác suất của G không phụ thuộc vào các đối số của nó và hoàn toàn xác định. Lúc đó với độ tin cậy bằng $(1-\alpha)$ cho trước có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tương ứng với chúng tìm được cặp giá trị g_{α_1} và g_{α_2} thỏa mãn điều kiện

$$P(G < g_{\alpha_1}) = \alpha_1 \tag{7.7}$$

và

$$P(G > g_{\alpha_2}) = \alpha_2 \tag{7.8}$$

Từ (7.7) và (7.8) suy ra

$$P(g_{\alpha_1} < G < g_{\alpha_2}) = 1 - (\alpha_1 - \alpha_2) = 1 - \alpha \tag{7.9}$$

Như vậy, với độ tin cậy $1-\alpha$ ta xây dựng được khoảng tin cậy $(g_{\alpha_1}, g_{\alpha_2})$ cho G. Bằng các phép biến đổi tương đương bao giờ cũng có thể đưa (7.9) về dạng biểu thức tương đương

$$P(G_1 < \theta < G_2) = 1 - \alpha$$

Đó chính là khoảng tin cậy cần tìm.

Trong thực tế thường yêu cầu độ tin cậy $(1-\alpha)$ khá lớn (chẳng hạn $1-\alpha = 0{,}95$) nên theo nguyên lý xác suất lớn biến cố $(G_1 < \theta < G_2)$ hầu như chắc chắn sẽ xảy ra trong một phép thử. Tiến hành một phép thử với mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$ ta thu được một mẫu cụ thể $w = (x_1, x_2, ..., x_n)$ do đó tính được giá trị của G_1 và G_2 ứng với mẫu cụ thể này, ký hiệu là g_1 và g_2. Lúc đó có thể kết luận là: Qua mẫu cụ thể với độ tin cậy $(1-\alpha)$ tham số θ của biến ngẫu nhiên gốc X sẽ nằm trong khoảng (g_1, g_2) tức là $(g_1 < \theta < g_2)$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phương pháp ước lượng bằng khoảng tin cậy khắc phục được các nhược điểm của phương pháp ước lượng điểm. Chẳng những nó làm tăng độ chính xác của ước lượng mà còn đánh giá được mức độ tin cậy của ước lượng đó nữa. Tuy nhiên nó cũng chứa đựng khả năng mắc sai lầm bằng α.

Dưới đây sẽ áp dụng phương pháp này để ước lượng một vài tham số đặc trưng cơ bản của biến ngẫu nhiên X trong tổng thể.

3.2. Ước lượng kỳ vọng toán của biến ngẫu nhiên phân phối theo quy luật chuẩn

Giả sử trong tổng thể biến ngẫu nhiên gốc X phân phối chuẩn $N(\mu,\sigma^2)$ nhưng chưa biết tham số μ của nó. Để ước lượng μ từ tổng thể ta lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

Để chọn thống kê G thích hợp ta xét hai trường hợp sau:

*1. **Đã biết phương sai σ^2 của biến ngẫu nhiên gốc X trong tổng thể***

Lúc đó ta chọn thống kê

$$G = U = \frac{\overline{X} - \mu}{Se(\overline{X})} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \qquad (7.10)$$

trong đó $\overline{X}$ là trung bình mẫu. Từ mục 7 Chương VI ta đã biết thống kê U phân phối chuẩn hóa $N(0,1)$. Do đó với độ tin cậy $1-\alpha$ cho trước tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ từ đó tìm được hai giá trị tới hạn tương ứng của phân phối chuẩn hóa là $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn các điều kiện

$P(U < u_{1-\alpha_1}) = \alpha_1$ và $P(U > u_{\alpha_2}) = \alpha_2$

Từ đó $P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha$

Vì u_{α_1} có tính chất $u_{\alpha_1} = -u_{1-\alpha_1}$ nên biểu thức trên có thể viết

$$P(-u_{\alpha_1} < U < u_{\alpha_2}) = 1 - \alpha \qquad (7.11)$$

Thay biểu thức của U từ (7.10) vào (7.11) và giải ra μ, ta thu được biểu thức tương đương:

$$P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_1}\right) = 1 - \alpha \qquad (7.12)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ý nghĩa của biểu thức thu được là: Với độ tin cậy bằng $1-\alpha$ tham số μ của biến ngẫu nhiên gốc X sẽ nằm trong khoảng

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\alpha_2};\overline{X}+\frac{\sigma}{\sqrt{n}}u_{\alpha_1}\right) \qquad (7.13)$$

Biểu thức (7.13) mới chỉ cho ta một khoảng tin cậy tổng quát. Với độ tin cậy $1-\alpha$ cho trước hiển nhiên có thể tìm được vô số cặp giá trị α_1 và α_2 thỏa mãn điều kiện $\alpha_1+\alpha_2=\alpha$, từ đó sẽ có vô số khoảng tin cậy tương ứng. Trong thực tế từ biểu thức (7.13) người ta thường chỉ sử dụng một số trường hợp đặc biệt sau:

- Khoảng tin cậy đối xứng:

Nếu lấy $\alpha_1=\alpha_2=\alpha/2$ từ (7.13) suy ra khoảng tin cậy của μ là

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\alpha/2};\overline{X}+\frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right) \qquad (7.14)$$

Nếu ký hiệu $$\varepsilon=\frac{\sigma}{\sqrt{n}}u_{\alpha/2} \qquad (7.15)$$

thì biểu thức của khoảng tin cậy sẽ có dạng $(\overline{X}-\varepsilon;\overline{X}+\varepsilon)$

ε được gọi là *độ chính xác* của ước lượng. Nó phản ánh mức độ sai lệch của trung bình mẫu so với trung bình tổng thể với xác suất $1-\alpha$ cho trước.

- Khoảng tin cậy bên phải:

Nếu lấy $\alpha_1=0$ và $\alpha_2=\alpha$ thì $u_{\alpha_1}=u_0=+\infty$ do đó khoảng tin cậy của μ là

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\alpha};+\infty\right) \qquad (7.16)$$

Biểu thức (7.16) được dùng để ước lượng giá trị tối thiểu của μ.

- Khoảng tin cậy bên trái:

Nếu lấy $\alpha_2=0$ và $\alpha_1=\alpha$ thì $u_{\alpha_2}=u_0=+\infty$ do đó khoảng tin cậy của μ là

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$\left(-\infty;\overline{X}+\frac{\sigma}{\sqrt{n}}u_{\alpha}\right) \tag{7.17}$$

Biểu thức (7.17) được dùng để ước lượng giá trị tối đa của μ.

Với cùng độ tin cậy $1-\alpha$ hiển nhiên khoảng tin cậy nào ngắn hơn sẽ tốt hơn. Trong trường hợp này độ dài khoảng tin cậy I sẽ là ngắn nhất khi khoảng tin cậy là đối xứng. Lúc đó độ dài khoảng tin cậy sẽ bằng hai lần độ chính xác và được xác định bằng biểu thức:

$$I = 2\varepsilon = \frac{2\sigma}{\sqrt{n}}u_{\alpha/2} \tag{7.18}$$

Từ (7.18) ta sẽ thu được công thức xác định kích thước mẫu tối thiểu n sao cho với độ tin cậy bằng $(1-\alpha)$ cho trước, độ dài khoảng tin cậy không vượt quá giá trị I_0 cho trước. Công thức có dạng:

$$n \geq \left[\frac{4\sigma^2}{I_0^2}u_{\alpha/2}^2\right] = \left[\frac{\sigma^2}{\varepsilon_0^2}u_{\alpha/2}^2\right] \tag{7.19}$$

tức là n số nguyên dương nhỏ nhất lớn hơn hoặc bằng:

$$\frac{\sigma^2}{\varepsilon_0^2}u_{\alpha/2}^2$$

Bài toán xác định kích thước mẫu tối thiểu n theo công thức (7.19) thường được đặt ra trước khi chọn mẫu, khi phải xác định kích thước mẫu cần điều tra để đáp ứng những yêu cầu chất lượng cho trước về độ tin cậy và độ chính xác của ước lượng.

Các khoảng tin cậy (7.14) (7.16) và (7.17) vẫn đang còn là khoảng tin cậy ngẫu nhiên. Vì vậy thủ tục ước lượng chưa thể coi là kết thúc. Vì độ tin cậy $(1-\alpha)$ khá lớn nên áp dụng nguyên lý xác suất lớn có thể coi là biến cố

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\alpha_2} < \mu < \overline{X}+\frac{\sigma}{\sqrt{n}}u_{\alpha_1}\right)$$

sẽ xảy ra trong một phép thử đối với mẫu ngẫu nhiên $W=(X_1,X_2,...,X_n)$. Thực hiện một phép thử đối với mẫu này thu được mẫu cụ thể

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$w = (x_1, x_2, ..., x_n)$; từ đó tìm được giá trị cụ thể $\bar{x}$ của trung bình mẫu. Lúc đó với độ tin cậy $(1-\alpha)$ qua một mẫu cụ thể, khoảng tin cậy của μ là:

$$(g_1, g_2) = \left(\bar{x} - \frac{\sigma}{\sqrt{n}} u_{\alpha_2} < \mu < \bar{x} + \frac{\sigma}{\sqrt{n}} u_{\alpha_1} \right)$$

Bằng cách tiến hành tương tự ta cũng thu được khoảng tin cậy cụ thể của các công thức (7.14) (7.16) (7.17) cũng như độ dài cụ thể của khoảng tin cậy theo công thức (7.18).

Đến đây ta thấy rõ vai trò của từng loại mẫu. Mẫu ngẫu nhiên cho phép xác định khoảng tin cậy ngẫu nhiên, còn mẫu cụ thể cho phép tìm ra khoảng tin cậy cụ thể (bằng số) của μ.

Thí dụ 1. Trọng lượng một loại sản phẩm là biến ngẫu nhiên phân phối theo quy luật chuẩn với độ lệch chuẩn là 1 gam. Cân thử 25 sản phẩm loại này ta thu được kết quả sau.

Trọng lượng (gam)	18	19	20	21
Số SP tương ứng	3	5	15	2

Với độ tin cậy 0,95 hãy tìm khoảng tin cậy đối xứng của trọng lượng trung bình của loại sản phẩm nói trên.

Giải. Gọi X là "trọng lượng sản phẩm", theo giả thiết X phân phối chuẩn với $\sigma = 1$. Vậy trọng lượng trung bình của sản phẩm chính là tham số μ. Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng giá trị của tham số μ của phân phối $N(\mu, \sigma^2)$ khi đã biết phương sai của nó. Vậy theo biểu thức (7.14) ta có khoảng tin cậy là:

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha/2}; \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha/2} \right)$$

Lấy từ tổng thể ra một mẫu ngẫu nhiên kích thước n = 25, gọi X_i là trọng lượng của sản phẩm thứ i $(i = \overline{1,25})$ ta có:

$$W = (X_1, X_2, ..., X_{25})$$

Từ đó:

$$\overline{X} = \frac{1}{25} \sum_{i=1}^{25} X_i$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với độ tin cậy $1-\alpha = 0,95$ thì $\alpha/2 = 0,025$

Tra bảng giá trị tới hạn chuẩn có $u_{0,025} = 1,96$

Vậy khoảng tin cậy đối xứng của μ là:

$$\left(\overline{X} - \frac{1}{\sqrt{25}}1,96; \overline{X} + \frac{1}{\sqrt{25}}1,96\right) = (\overline{X} - 0,392; \overline{X} + 0,392)$$

Kết quả thu được cho biết 95% số mẫu khích thước n = 25 sẽ chứa đựng tham số μ trong khoảng $(\overline{X} - 0,392; \overline{X} + 0,392)$

Từ bảng số liệu tìm được trung bình mẫu cụ thể:

$$\overline{x} = \frac{3 \cdot 18 + 5 \cdot 19 + 15 \cdot 20 + 2 \cdot 21}{25} = 19,64$$

Vậy với độ tin cậy 0,95 qua mẫu cụ thể này, khoảng tin cậy đối xứng của μ là:

$$(19,64 - 0,392; 19,64 + 0,392)$$

hay

$$(19,248 < \mu < 20,032)$$

Chú ý rằng không thể viết $P(19,248 < \mu < 20,032) = 0,95$ vì độ tin cậy gắn với khoảng tin cậy ngẫu nhiên chứ không gắn với mẫu cụ thể. Hơn nữa do μ là một hằng số nên nó chỉ có thể thuộc hoặc không thuộc khoảng $(19,248 < \mu < 20,032)$ tức là với một mẫu cụ thể thì biến cố $(19,248 < \mu < 20,032)$ không phải là biến cố ngẫu nhiên. Hoặc nó là biến cố chắc chắn, hoặc nó là biến cố không thể có:

Chú ý: Từ biểu thức (7.18) ta có thể đưa ra các nhận xét sau:

- Khi tăng kích thước mẫu n lên và giữ nguyên độ tin cậy $1-\alpha$ cho trước thì ε giảm đi tức là độ chính xác của ước lượng tăng lên.

- Khi tăng độ tin cậy $1-\alpha$ lên mà giữ nguyên kích thước mẫu n thì giá trị tới hạn chuẩn cũng tăng lên theo đó ε cũng tăng lên làm cho độ chính xác của ước lượng giảm đi.

Thí dụ 2: Trong thí dụ 1, nếu yêu cầu độ chính xác của ước lượng chỉ là 0,1 và giữ nguyên độ tin cậy $1-\alpha = 0,95$ thì phải điều tra một mẫu kích thước bằng bao nhiêu?

Giải: Với $\varepsilon_0 = 0,1$ theo công thức (7.19) ta có:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$n \geq \left[\left(\frac{1^2}{(0,1)^2} \right) (1,96)^2 \right] = 385$$

Vậy để làm tăng độ chính xác của ước lượng từ 0,392 lên đến 0,1 thì kích thước mẫu phải tăng từ 25 lên đến 385.

Nếu mẫu được chọn theo phương thức không hoàn lại từ tổng thể hữu hạn và $n > 0,1N$ thì $Se(\overline{X})$ sẽ nhân thêm hệ số điều chỉnh, do đó biểu thức (7.10) của thống kê G sẽ trở thành:

$$G = U = \frac{(\overline{X} - \mu)}{Se(\overline{X})} = \frac{\overline{X} - \mu}{\sqrt{\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}}}$$

và việc xây dựng các công thức ước lượng cũng tiến hành tương tự.

2. Chưa biết phương sai σ^2 của biến ngẫu nhiên gốc X trong tổng thể và kích thước của mẫu $n < 30$

Lúc đó ta chọn thống kê:

$$G = T = \frac{(\overline{X} - \mu)\sqrt{n}}{S} \qquad (7.20)$$

trong đó S là độ lệch chuẩn mẫu. Từ mục 7 Chương VI ta đã biết thống kê T phân phối theo quy luật Student với $(n-1)$ bậc tự do. Vì vậy, với độ tin cậy bằng $(1-\alpha)$ cho trước có thể tìm được hai giá trị tới hạn Student tương ứng là $t_{1-\alpha_1}^{(n-1)}$ và $t_{\alpha_2}^{(n-1)}$ thỏa mãn điều kiện:

$$P(T < t_{1-\alpha_1}^{(n-1)}) = \alpha_1 \text{ và}$$

$$P(T > t_{\alpha_2}^{(n-1)}) = \alpha_2$$

Từ đó $$P(t_{1-\alpha_1}^{(n-1)} < T < t_{\alpha_2}^{(n-1)}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha$$

Vì giá trị tới hạn Student có tính chất $t_{\alpha_1}^{(n-1)} = -t_{1-\alpha_1}^{(n-1)}$ nên biểu thức có thể viết:

$$P(-t_{\alpha_1}^{(n-1)} < T < t_{\alpha_2}^{(n-1)}) = 1 - \alpha \qquad (7.21)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thay biểu thức của T từ (7.20) vào (7.21) và giải ra μ ta thu được biểu thức tương đương:

$$P\left(\overline{X}-\frac{S}{\sqrt{n}}t_{\alpha_2}^{(n-1)}<\mu<\overline{X}+\frac{S}{\sqrt{n}}t_{\alpha_1}^{(n-1)}\right)=1-\alpha \qquad (7.22)$$

Như vậy, khoảng tin cậy của μ với độ tin cậy $(1-\alpha)$ là:

$$\left(\overline{X}-\frac{S}{\sqrt{n}}t_{\alpha_2}^{(n-1)};\overline{X}+\frac{S}{\sqrt{n}}t_{\alpha_1}^{(n-1)}\right) \qquad (7.23)$$

Từ biểu thức tổng quát (7.23) có thể xây dựng công thức khoảng tin cậy trong những trường hợp đặc biệt sau:

- Khoảng tin cậy đối xứng khi $\alpha_1=\alpha_2=\alpha/2$

$$\left(\overline{X}-\frac{S}{\sqrt{n}}t_{\alpha/2}^{(n-1)};\overline{X}+\frac{S}{\sqrt{n}}t_{\alpha/2}^{(n-1)}\right) \qquad (7.24)$$

- Khoảng tin cậy bên phải khi $\alpha_1=0$ và $\alpha_2=\alpha$

$$\left(\overline{X}-\frac{S}{\sqrt{n}}t_{\alpha}^{(n-1)};+\infty\right) \qquad (7.25)$$

- Khoảng tin cậy bên trái khi $\alpha_2=0$ và $\alpha_1=\alpha$

$$\left(-\infty;\overline{X}+\frac{S}{\sqrt{n}}t_{\alpha}^{(n-1)}\right) \qquad (7.26)$$

Trong trường hợp này độ dài khoảng tin cậy I cũng là ngắn nhất khi khoảng tin cậy là đối xứng, do đó nó cũng bằng hai lần độ chính xác và được xác định bằng biểu thức:

$$I=2\varepsilon=\frac{2S}{\sqrt{n}}t_{\alpha/2}^{(n-1)} \qquad (7.27)$$

Lúc đó việc xác định kích thước mẫu tối thiểu n sao cho với độ tin cậy bằng $(1-\alpha)$ cho trước, độ dài khoảng tin cậy không vượt quá giá trị I_0 cho trước được giải quyết bằng *phương pháp mẫu kép* sau đây:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trước hết điều tra một mẫu sơ bộ kích thước $m \geq 2$: $W_1 = (X_1, ..., X_m)$ và từ đó tìm được phương sai mẫu của mẫu sơ bộ đó:

$$S^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2$$

với $$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$$

Sau đó lập mẫu thứ hai kích thước $n - m$

$$W_2 = (X_{m+1}, ..., X_n)$$

Có thể chứng minh được rằng thống kê:

$$T = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)\sqrt{n}}{S}$$

phân phối theo quy luật Student với $(m-1)$ bậc tự do. Vì vậy, có thể tìm được giá trị $t_{\alpha/2}^{(m-1)}$ sao cho:

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{S}{\sqrt{n}} t_{\alpha/2}^{(m-1)} < \mu < \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{S}{\sqrt{n}} t_{\alpha/2}^{(m-1)}\right) = 1 - \alpha$$

do đó $I_0 = \frac{2S}{\sqrt{n}} t_{\alpha/2}^{(m-1)}$

Từ đây suy ra kích thước mẫu cần tìm:

$$n \geq \left[\frac{S^2}{\varepsilon_0^{\,2}} t_{\alpha/2}^{(m-1)}\right] \qquad (7.28)$$

Như vậy, dựa vào mẫu sơ bộ đã có ta tìm được kích thước mẫu chính thức đáp ứng yêu cầu về chất lượng của ước lượng. Trên thực tế chỉ cần điều tra tiếp mẫu thứ hai kích thước $n - m$ là đủ.

Với độ tin cậy $(1-\alpha)$ khá lớn, để có khoảng tin cậy cụ thể của μ, người ta lập mẫu cụ thể $w = (x_1, x_2, ..., x_n)$, từ đó tính được các giá trị $\overline{x}$ và s, thay chúng vào các công thức khoảng tin cậy vừa tìm được ở trên ta có các khoảng tin cậy cụ thể bằng số phải tìm.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 3. Để xác định trọng lượng trung bình của các bao bột trong kho, người ta đem cân ngẫu nhiên 15 bao của kho đó và tìm được $\bar{x} = 39,8$ kg; $s^2 = 0,144$. Hãy tìm khoảng tin cậy đối xứng của trọng lượng trung bình của các bao bột trong kho với yêu cầu độ tin cậy của việc ước lượng là 99%. Giả thiết trọng lượng đóng bao của các bao bột là biến ngẫu nhiên phân phối chuẩn.

Giải. Gọi X là trọng lượng bột đóng bao, theo giả thiết X phân phối chuẩn. Vậy trọng lượng đóng bao trung bình chính là giá trị μ. Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng giá trị của tham số μ của phân phối $N(\mu, \sigma^2)$ khi chưa biết σ^2 của X. Vậy theo biểu thức (7.27) ta có khoảng tin cậy là:

$$\left(\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)}; \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)}\right)$$

Cân ngẫu nhiên 15 bao bột, gọi $X_i (i = \overline{1,15})$ là trọng lượng của bao thứ i ta có mẫu ngẫu nhiên $W = (X_1, ..., X_{15})$.

Với độ tin cậy $1 - \alpha = 0,99$ thì $\alpha/2 = 0,005$ tra bảng phân phối Student có $t_{0,005}^{(14)} = 2,977$. Vậy với độ tin cậy 0,99 khoảng tin cậy đối xứng của μ là

$$\left(\overline{X} - \frac{S}{\sqrt{15}} 2,977; \overline{X} + \frac{S}{\sqrt{15}} 2,977\right)$$

Với mẫu cụ thể ta tính được $\bar{x} = 39,8$, $s^2 = 0,144 \rightarrow s = 0,397$. Vậy với độ tin cậy 0,99 qua mẫu cụ thể này, khoảng tin cậy đối xứng của μ là:

$$\left(39,8 - \frac{0,379}{\sqrt{15}} 2,977; 39,8 + \frac{0,379}{\sqrt{15}} 2,977\right)$$

hay $(39,5023 < \mu < 40,0977)$.

Thí dụ 4. Phỏng vấn 5 gia đình có 3 người về chi phí hàng tháng cho nhu yếu phẩm thu được các số liệu sau: 150 ngàn đồng, 180 ngàn, 200 ngàn, 250 ngàn, 300 ngàn. Vậy phải phỏng vấn bao nhiêu gia đình cùng loại

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

để với độ tin cậy 95% sai số của việc ước lượng chi phí trung bình hàng tháng cho nhu yếu phẩm không vượt quá 30 ngàn đồng. Giả thiết chi phí hàng tháng cho nhu yếu phẩm là biến ngẫu nhiên phân phối chuẩn.

Giải. Gọi X là chi phí hàng tháng cho nhu yếu phẩm, theo giả thiết X phân phối chuẩn. Vậy chi phí trung bình chính là giá trị μ. Đây là bài toán xác định kích thước mẫu tối thiểu cho việc ước lượng tham số μ của phân phối $N(\mu,\sigma^2)$ khi chưa biết phương sai σ^2.

Theo phương pháp mẫu kép, từ mẫu sơ bộ kích thước m=5 ta tìm được:

$$\bar{x} = \frac{150+180+200+250+300}{5} = 216 \text{ ngàn}$$

$$s^2 = [(150-216)^2 + (180-216)^2 + (200-216)^2 + (250-216)^2 + (300-216)^2] : 4 = 3530$$

$$\varepsilon_0 = 30 \text{ ngàn}$$

$$t^4_{0,025} = 2,776$$

Vậy theo công thức (7.28) có:

$$n \geq \left[\frac{3530}{30^2}(2,776)^2\right] = 31 \text{ gia đình}$$

Như vậy phải phỏng vấn thêm $31-5=26$ gia đình nữa.

Chú ý rằng khi kích thước mẫu $n > 30$ thì phân phối Student đã xấp xỉ phân phối chuẩn hóa do đó khi sử dụng các công thức (7.22 - 7.28) có thể dùng các giá trị phân phối chuẩn hóa để thay thế cho các giá trị của phân phối Student tương ứng.

Thí dụ 5. Để xác định kích thước trung bình của chi tiết do một máy sản xuất người ta lấy ngẫu nhiên 200 chi tiết để đo kích thước và thu được bảng số liệu sau (7.1). Với độ tin cậy 95% hãy ước lượng bằng khoảng tin cậy đối xứng kích thước trung bình của chi tiết do máy đó sản xuất. Giả thiết kích thước chi tiết là biến ngẫu nhiên phân phối chuẩn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bảng 7.1

Kích thước chi tiết (cm)	Số chi tiết tương ứng
54,795 - 54,805	6
54,805 - 54,815	14
54,815 - 54,825	33
54,825 - 54,835	47
54,835 - 54,845	45
54,845 - 54,855	33
54,855 - 54,865	15
54,865 - 54,875	7
	n = 200

Giải. Gọi X là kích thước chi tiết do máy đó sản xuất. Theo giả thiết X phân phối chuẩn. Vậy kích thước trung bình của chi tiết chính là tham số μ. Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng giá trị của tham số μ của phân phối $N(\mu,\sigma^2)$ khi chưa biết phương sai σ^2. Vậy khoảng tin cậy của μ là:

$$(\overline{X}-\frac{S}{n}t_{\alpha/2}^{(n-1)};\overline{X}+\frac{S}{n}t_{\alpha/2}^{(n-1)})$$

Do $n=200>30$ nên với $1-\alpha=0,95$ thì

$t_{\alpha/2}^{(n-1)}=t_{0,025}^{(199)}\approx u_{0,025}=1,96$. Với mẫu cụ thể để tìm $\overline{x}$ và s lập bảng sau:

x_i	n_i	n_ix_i	$n_ix_i^2$
54,80	6	328,80	18018,240
54,81	14	767,34	2057,905
54,82	33	1809,06	99172,669
54,83	47	2577,01	14197,450
54,84	45	1467,80	135334,150
54,85	33	1810,05	99281,242
54,86	15	822,90	45144,294
54,87	7	384,09	21075,018
	n = 200	10967,05	601380,950

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đó $\bar{x} = \frac{10967,05}{200} = 54,83525$

$$ms = \frac{601380,95}{200} - (54,83525)^2 = 0,0002559$$

$$s = \sqrt{\frac{200}{199}0,0002559} = 0,0164$$

Vậy với độ tin cậy 0,95 qua mẫu cụ thể này khoảng tin cậy đối xứng của μ là:

$$\left(54,83525 - \frac{0,0164}{\sqrt{200}}1,96; 54,83525 + \frac{0,0164}{\sqrt{200}}1,96\right)$$

hay $(54,83298 < \mu < 54,83752)$

Thí dụ. Với các số liệu trong thí dụ A và bằng phần mềm Excel chúng ta ước lượng với độ tin cậy 0,95 thu nhập trung bình hàng năm của dân cư ba vùng và thu được kết quả sau:

Khoảng tin cậy 95% cho thu nhập trung bình vùng 1

95%	confidence level
1649.75	mean
4.75	std. dev.
100	n
1.984	t (df = 99)
0.943	half-width
1650.693	upper confidence limit
1648.807	lower confidence limit

Khoảng tin cậy 95% cho thu nhập trung bình vùng 2

95%	confidence level
1623.86	mean
5.93	std. dev.
100	n
1.984	t (df = 99)
1.177	half-width
1625.037	upper confidence limit
1622.683	lower confidence limit

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Khoảng tin cậy 95% cho thu nhập trung bình vùng 3

95%	confidence level
1654.03	mean
6.73	std. dev.
100	n
1.984	t (df = 99)
1.335	half-width
1655.365	upper confidence limit
1652.695	lower confidence limit

Nếu mẫu được chọn từ tổng thể hữu hạn theo phương thức không hoàn lại và $n > 0{,}1N$ thì biểu thức (7.20) của thống kê G trở thành:

$$G = T = \frac{\overline{X} - \mu}{\sqrt{\frac{N-n}{N-1}\frac{S^2}{n}}}$$

và việc xây dựng công thức ước lượng cũng được tiến hành tương tự như mục trước.

Bài tập

7.16. Doanh số của một cửa hàng là biến ngẫu nhiên phân phối chuẩn với độ lệch chuẩn là 2 triệu/tháng. Điều tra ngẫu nhiên doanh số của 600 cửa hàng có quy mô tương tự nhau tìm được doanh số trung bình là 8,5 triệu. Với độ tin cậy 95% hãy ước lượng doanh số trung bình của các cửa hàng thuộc quy mô đó.

7.17. Với độ tin cậy 95%, hãy ước lượng xăng hao phí trung bình cho một ô tô chạy từ A đến B nếu chạy thử 30 lần trên đoạn đường này người ta ghi nhận được lượng xăng hao phí như sau:

Lượng xăng hao phí (lít)	Số lần tương ứng
9,6 - 9,8	3
9,8 - 10,0	5
10,0 - 10,2	10
10,2 - 10,4	8
10,4 - 10,6	4
	n = 30

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Biết rằng lượng xăng hao phí là biến ngẫu nhiên tuân theo quy luật chuẩn.

ĐS: (10,0454; 10,2212)

7.18. Để định mức thời gian gia công một chi tiết máy, người ta theo dõi ngẫu nhiên quá trình gia công 25 chi tiết và thu được số liệu như sau:

Bằng khoảng tin cậy đối xứng hãy ước lượng thời gian gia công trung bình một chi tiết máy với độ tin cậy $1 - \alpha = 0,95$. Giả thiết thời gian gia công chi tiết là biến ngẫu nhiên phân phối chuẩn.

Thời gian gia công (phút)	Số chi tiết máy tương ứng
15 - 17	1
17 - 19	3
19 - 21	4
21 - 23	12
23 - 25	3
25 - 27	2
	n = 25

7.19. Có số liệu về trọng lượng của một loại trứng gà như ở bảng dưới đây. Bằng khoảng tin cậy đối xứng hãy ước lượng trọng lượng trung bình của loại trứng gà này với độ tin cậy 0,95. Giả thiết trọng lượng trứng gà là biến ngẫu nhiên tuân theo quy luật chuẩn.

Trọng lượng trứng (gam)	Số quả trứng tương ứng
25 - 30	2
30 - 35	3
35 - 40	10
40 - 45	8
45 - 50	2
	n = 25

ĐS: $\bar{x} = 38,5$; $s = 5,2042$; (36,352; 40,648).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7.20. Cân thử 25 bao bột mỳ, người ta tính được trọng lượng trung bình một bao là $\bar{x}$ = 40 kg, độ lệch chuẩn mẫu s = 5 kg.

Với độ tin cậy 95%, bằng khoảng tin cậy đối xứng hãy ước lượng trọng lượng trung bình của một bao bột mỳ. Giả thiết trọng lượng của bao bột mỳ là biến ngẫu nhiên tuân theo quy luật chuẩn.

7.21. Để ước lượng chiều dày trung bình của các tấm vật liệu do một xí nghiệp sản xuất, người ta tiến hành đo 5 tấm và thu được kết quả sau:

x_1 = 2,015 mm

x_2 = 2,025 mm

x_3 = 2,015 mm

x_4 = 2,020 mm

x_5 = 2,015 mm

Dựa vào số liệu trên hãy ước lượng chiều dày trung bình các tấm vật liệu do xí nghiệp sản xuất bằng khoảng tin cậy đối xứng với độ tin cậy 1 - α = 0,95.

Biết rằng chiều dày các tấm vật liệu là biến ngẫu nhiên tuân theo quy luật phân phối chuẩn.

ĐS: (2,0141; 2,0219)

7.22. Điều tra doanh số hàng tháng của 100 hộ kinh doanh một loại hàng, ta có bảng số liệu sau:

Doanh số (triệu đồng)	Số hộ tương ứng
11,5	10
11,6	15
11,7	20
11,8	30
11,9	15
12,0	10
	n = 100

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hãy ước lượng doanh số trung bình hàng tháng của các hộ kinh doanh mặt hàng này bằng khoảng tin cậy đối xứng với hệ số tin cậy 0,95 với giả thiết doanh số là biến ngẫu nhiên phân phối chuẩn.

ĐS: $\bar{x}$ = 11,755; s = 0,144; (11, 727; 11,183)

7.23. Để xác định giá trung bình đối với một loại hàng hoá trên thị trường, người ta điều tra ngẫu nhiên tại 100 cửa hàng thu được bảng số liệu sau đây

Giá (đồng)	83	85	87	89	91	93	95	97	99	101
Số cửa hàng	6	7	12	15	30	10	8	6	4	2

Với độ tin cậy 95% hãy ước lượng giá trung bình của loại hàng đó tại thời điểm đang xét. Biết rằng giá hàng hoá là biến ngẫu nhiên tuân theo quy luật phân phối chuẩn.

7.24. Chiều dài (X) loại sản phẩm A do một máy tự động sản xuất là một biến ngẫu nhiên tuân theo quy luật phân phối chuẩn với độ lệch chuẩn σ_x = 3cm.

Để ước lượng chiều dài trung bình của loại sản phẩm nói trên với độ tin cậy 1 - α = 0,95, người ta tiến hành đo 36 sản phẩm.

a. Tìm khoảng tin cậy đối xứng của chiều dài trung bình loại sản phẩm đó.

b. Để ước lượng với độ tin cậy 1 - α = 0,99, độ dài khoảng tin cậy đối xứng không vượt quá 0,6cm thì phải đo bao nhiêu sản phẩm.

ĐS: a. ($\bar{x}$ - 0,98; $\bar{x}$ + 0,98);

b. $n \geq 666$

7.25. Trọng lượng (X) một loại chi tiết là một biến ngẫu nhiên tuân theo quy luật chuẩn với σ_x = 1,2kg. Phải chọn ít nhất bao nhiêu chi tiết để điều tra, nếu muốn độ chính xác của ước lượng là 0,3 và độ tin cậy của ước lượng 1 - α = 0,95.

ĐS: n = 62

7.26. Đo chiều sâu của biển bằng một loại dụng cụ có sai số tuân theo quy luật chuẩn với phương sai bằng 400 (m^2). Phải đo bao nhiêu lần độc lập với nhau để kết quả có sai số không quá 15 (m) với độ tin cậy 0,95.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7.27. Năng suất giống ngô A ở một vùng được báo lên qua 25 điểm thu hoạch và có kết quả như sau:

Năng suất (tạ/ha)	Số điểm thu hoạch
7	2
9	7
11	12
13	3
17	1
	n = 25

Với hệ số tin cậy 95%, hãy tính năng suất trung bình tối thiểu của vùng này, biết rằng năng suất ngô của vùng này là biến ngẫu nhiên tuân theo quy luật chuẩn.

ĐS: $\mu > 9{,}913$

7.28. Để định mức thời gian gia công một chi tiết máy người ta theo dõi ngẫu nhiên quá trình gia công 25 chi tiết và thu được bảng số liệu sau:

Thời gian (phút)	Số chi tiết
14	2
16	6
18	11
20	4
24	2
	n = 25

a. Tìm trung bình và độ lệch chuẩn của mẫu trên.

b. Với độ tin cậy $1 - \alpha = 0{,}95$ hãy ước lượng thời gian gia công trung bình tối đa đối với loại chi tiết đó. Giả thiết thời gian gia công chi tiết đó là biến ngẫu nhiên tuân theo quy luật chuẩn.

3.3. Ước lượng hiệu hai kỳ vọng toán của hai biến ngẫu nhiên phân phối chuẩn

Giả sử có hai tổng thể nghiên cứu, trong đó các biến ngẫu nhiên X_1 và X_2 cùng phân phối chuẩn với các tham số đặc trưng tương ứng là μ_1, μ_2 và

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

σ_1^2, σ_2^2, với μ_1 và μ_2 chưa biết. Để ước lượng sự chênh lệch $\mu_1 - \mu_2$ giữa hai kỳ vọng toán từ hai tổng thể lập hai mẫu ngẫu nhiên độc lập kích thước tương ứng là n_1 và n_2

$$W_1 = (X_{11}, X_{12}, ..., X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, ..., X_{2n_2})$$

từ đó tìm được các thống kê đặc trưng mẫu tương ứng là $\overline{X}_1, \overline{X}_2$ và S_1^2, S_2^2.

Để chọn thống kê G thích hợp ta xét các trường hợp sau:

1. Nếu đã biết các phương sai σ_1^2 và σ_2^2 của các tổng thể

Lúc đó chọn thống kê

$$G = U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Từ mục 7 chương VI ta đã biết U phân phối chuẩn hóa $N(0,1)$. Do đó với độ tin cậy $1-\alpha$ cho trước tìm được cặp giá trị α_1 và $\alpha_2 (\alpha_1 + \alpha_2 = \alpha)$ và hai giá trị tới hạn chuẩn tương ứng là $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn:

$$P(U < u_{1-\alpha_1}) = \alpha_1$$

$$P(U > u_{\alpha_2}) = \alpha_2$$

Từ đó:

$$P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1 - \alpha$$

Bằng các phép biến đổi như đã làm ở các mục trước ta thu được khoảng tin cậy mức $(1-\alpha)$ của $(\mu_1 - \mu_2)$ như sau:

$$P\left[(\overline{X}_1 - \overline{X}_2) - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{\alpha_2} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{\alpha_1}\right] = 1 - \alpha \quad (7.29)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ công thức tổng quát (7.29) có thể xây dựng được các công thức cụ thể như đã làm ở các mục trước:

+ Khoảng tin cậy đối xứng

$$P\left[(\overline{X}_1-\overline{X}_2)-\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}u_{\alpha/2}<\mu_1-\mu_2<(\overline{X}_1-\overline{X}_2)+\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}u_{\alpha/2}\right]=1-\alpha \quad (7.30)$$

+ Khoảng tin cậy bên phải

$$P\left[(\mu_1-\mu_2)>(\overline{X}_1-\overline{X}_2)-\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}u_{\alpha}\right]=1-\alpha \quad (7.31)$$

+ Khoảng tin cậy bên trái

$$P\left[(\mu_1-\mu_2)<(\overline{X}_1-\overline{X}_2)+\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}u_{\alpha}\right]=1-\alpha \quad (7.32)$$

2. Nếu chưa biết các phương sai σ_1^2 và σ_2^2 của các tổng thể, song giả thiết rằng $\sigma_1^2=\sigma_2^2$

Lúc đó chọn thống kê

$$G=T=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \quad (7.33)$$

trong đó $$S_p=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

Ta đã biết thống kê T phân phối $T(n_1+n_2-2)$ do đó tiến hành tương tự như ở các mục trước ta thu được khoảng tin cậy $(1-\alpha)$ của hiệu $\mu_1-\mu_2$ như sau:

$$P\left[(\overline{X}_1-\overline{X}_2)-S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\alpha_2}^{(n_1+n_2-2)}<\mu_1-\mu_2<(\overline{X}_1-\overline{X}_2)+S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\alpha_1}^{(n_1+n_2-2)}\right]=1-\alpha \quad (7.34)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ công thức (7.34) có thể xây dựng các khoảng tin cậy đối xứng, khoảng tin cậy bên phải và bên trái như đã làm ở mục trước.

3. Nếu chưa biết các phương sai σ_1^2 và σ_2^2 của các tổng thể và không có căn cứ để cho rằng chúng bằng nhau

Lúc đó chọn lập thống kê

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad (7.35)$$

Ta đã biết thống kê T phân phối Student với số bậc tự do là: .

$$k = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)C^2 + (1 - C)^2(n_1 - 1)}$$

với
$$C = \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

do đó khoảng tin cậy mức $(1-\alpha)$ của hiệu $\mu_1 - \mu_2$ như sau :

$$P\left[(\overline{X}_1 - \overline{X}_2) - \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\, t_{\alpha_2}^{(k)} < \mu_1 - \mu_2 < \right.$$

$$\left. < (\overline{X}_1 - \overline{X}_2) + \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\, t_{\alpha_1}^{(k)}\right] = 1 - \alpha \qquad (7.36)$$

Từ đó có thể xây dựng các khoảng tin cậy đối xứng, bên phải, bên trái tương ứng.

Việc xác định kích thước mẫu để ước lượng hiệu $\mu_1 - \mu_2$ của hai kỳ vọng toán được tiến hành như sau: Giả sử ta chủ trương lấy ra hai mẫu độc lập có kích thước như nhau và bằng n và giả sử các phương sai tổng thể đã biết và bằng σ_1^2 và σ_2^2. Lúc đó từ (7.30) suy ra kích thước n của hai mẫu đảm bảo với độ tin cậy $1-\alpha$ cho trước độ dài khoảng tin cậy không vượt quá giá trị I_0 cho trước là:

$$n \geq \left[\frac{(\sigma_1^2 + \sigma_2^2)}{\varepsilon_0^2} u_{\alpha/2}^2\right] \qquad (3.37)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

nếu chưa biết σ_1^2 và σ_2^2 thì thay bằng ước lượng của chúng là S_1^2 và S_2^2 với S_1^2 và S_2^2 là phương sai của hai mẫu sơ bộ kích thước m_1 và m_2 được rút ra từ hai tổng thể nghiên cứu.

Thí dụ 6. Từ một chuồng nuôi lợn người ta lấy ra ngẫu nhiên bốn con đem cân và thu được trọng lượng tương ứng của chúng là 64, 66, 89 và 77 kg. Từ một chuồng khác lấy ra ba con đem cân thu được trọng lượng là 56, 71 và 53 kg. Với độ tin cậy 95% hãy ước lượng sự khác biệt về trọng lượng trung bình của hai chuồng lợn đó. Giả thiết trọng lượng của lợn phân phối chuẩn.

Giải. Gọi X_1 và X_2 tương ứng là trọng lượng của lợn ở hai chuồng nói trên, theo giả thiết X_1 và X_2 phân phối chuẩn. Vậy trọng lượng trung bình là μ_1 và μ_2. Đây là bài toán ước lượng hiệu số $\mu_1 - \mu_2$ khi chưa biết các phương sai của tổng thể. Nếu có thể cho rằng phương sai của chúng bằng nhau (chẳng hạn cả hai chuồng cùng nuôi một giống lợn và được chăm sóc như nhau) thì có thể từ công thức ước lượng (7.34) suy ra khoảng tin cậy đối xứng sau :

$$P\left[(\overline{X}_1 - \overline{X}_2) - S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\alpha/2}^{(n_1+n_2-2)} < \mu_1 - \mu_2 < \right.$$

$$\left.(\overline{X}_1 - \overline{X}_2) + S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\alpha/2}^{(n_1+n_2-2)}\right] = 1-\alpha \qquad (7.38)$$

Từ hai mẫu cụ thể ta tìm được:

$$n_1 = 4 \qquad n_2 = 3$$
$$\overline{x}_1 = 74 \qquad \overline{x}_2 = 60$$
$$s_1^2 = 132,67 \qquad s_2^2 = 93$$

Từ đó:

$$S_p^2 = \frac{3.132,67 + 2.93}{4+3-2} = 116,8$$

và $$t_{\alpha/2}^{(n_1+n_2-2)} = t_{0,025}^{(5)} = 2,57$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$14-\sqrt{116,8}.\sqrt{\frac{1}{4}+\frac{1}{3}}2,57<\mu_1-\mu_2<14+\sqrt{116,8}.\sqrt{\frac{1}{4}+\frac{1}{3}}2,57$$

$$[-7,21<\mu_1-\mu_2<35,21]$$

4. Ước lượng hiệu hai kì vọng toán khi mẫu gồm các số liệu theo cặp

Ở các phần trước ta luôn giả thiết rằng có thể lấy ra hai mẫu độc lập từ hai tổng thể nghiên cứu. Trong nhiều trường hợp thực tế phải so sánh hai kỳ vọng toán trong điều kiện hai mẫu được rút ra từ cùng một tổng thể và nói chung là chúng phụ thuộc nhau. Lúc đó nếu hai mẫu bao gồm các cặp giá trị (X_1, Y_1) (X_2, Y_2) ... (X_n, Y_n) thì lúc đó với mỗi cặp giá trị của hai mẫu có thể xác định sự sai lệch trên từng cặp

$$D_i = X_i - Y_i \quad (i=\overline{1,n}) \tag{7.39}$$

Lúc đó có thể coi D là biến ngẫu nhiên phân phối chuẩn với các tham số là μ_D và σ_D^2 và việc xử lý D cũng giống như đã làm với một mẫu ngẫu nhiên. Từ mẫu bao gồm n sai lệch D_i ta tìm được:

$$\overline{D}=\frac{1}{n}\sum_{i=1}^{n}D_i \tag{7.40}$$

$$\text{và}\quad Se(\overline{D})=\frac{S_D}{\sqrt{n}} \qquad \text{với}\quad S_D=\frac{\sqrt{\sum_{i=1}^{n}(D_i-\overline{D})^2}}{\sqrt{n-1}} \tag{7.41}$$

Lúc đó thống kê

$$G=T=\frac{(\overline{D}-\mu_D)\sqrt{n}}{S_D}\sim T(n-1)$$

nên khoảng tin cậy mức $(1-\alpha)$ của μ_D là:

$$P\left[\overline{D}-\frac{S_D}{\sqrt{n}}t_{\alpha_2}^{(n-1)}<\mu_D<\overline{D}+\frac{S_D}{\sqrt{n}}t_{\alpha_1}^{(n-1)}\right]=1-\alpha \tag{7.42}$$

Từ đó có thể xây dựng các khoảng tin cậy đối xứng, bên phải, bên trái tương ứng.

Lúc đó kích thước mẫu tối thiểu n theo độ tin cậy $(1-\alpha)$ cho trước và độ dài của khoảng tin cậy tối đa là I_0 cho trước được xác định bằng công thức:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$n \geq \left[\frac{S_D^2}{\varepsilon_0^2} (t_{\alpha/2}^{(m-1)})^2 \right] \qquad (7.43)$$

trong đó S_D^2 là phương sai của mẫu sơ bộ kích thước m.

Thí dụ 7. Giả sử để theo dõi tốc độ tăng trọng trung bình của một đàn lợn người ta lấy ngẫu nhiên 4 con và cân trọng lượng của chúng vào đầu tháng và cuối tháng. Kết quả thu được như sau:

Thứ tự	TL đầu tháng X_2	TL cuối tháng X_1	d_i	$(d_i - \bar{d})$	$(d_i - \bar{d})^2$
1	57	64	7	- 4	16
2	57	66	9	- 2	4
3	73	89	16	5	25
4	65	77	12	1	1
			$\bar{d} = \frac{44}{4} = 11$		$S_D^2 = \frac{46}{3} = 15,3$

Với độ tin cậy 95% hãy ước lượng mức tăng trọng trung bình của đàn lợn trong tháng đó.

Giải. Gọi X_1 và X_2 là trọng lượng của lợn vào cuối tháng và đầu tháng. Giả sử X_1 và X_2 phân phối chuẩn. Lúc đó μ_1 và μ_2 là trọng lượng trung bình tương ứng và mức tăng trọng trung bình được xác định bằng giá trị $\mu_D = \mu_1 - \mu_2$.

Từ biểu thức (7.42) ta có khoảng tin cậy đối xứng của μ_D như sau:

$$P\left[\bar{D} - \frac{S_D}{\sqrt{n}} t_{\alpha/2}^{(n-1)} < \mu_D < \bar{D} + \frac{S_D}{\sqrt{n}} t_{\alpha/2}^{(n-1)} \right] = 1 - \alpha$$

Từ mẫu cụ thể ta tính được:

$$\bar{d} = 11;\ S_d^2 = 15,3;\ t_{\alpha/2}^{(n-1)} = t_{0,025}^{(3)} = 3,18$$

Vậy $11 - \frac{\sqrt{15,3}}{\sqrt{4}} 3,18 < \mu_D < 11 + \frac{\sqrt{15,3}}{\sqrt{4}} 3,18$

$$\Rightarrow (4,78 < \mu_D < 17,22)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

7.29. Để xác định hiệu quả của một loại thức ăn phụ đối với sự tăng trọng của bò, người ta lấy ngẫu nhiên 8 con bò cùng trọng lượng chia thành hai nhóm, mỗi nhóm 4 con, một nhóm ăn bình thường, một nhóm cho ăn thêm thức ăn phụ. Sau 6 tháng thu được kết quả sau:

Nhóm ăn thêm thức ăn phụ A	Nhóm cho ăn bình thường B
330	290
360	320
400	340
350	370

a. Ước lượng bằng khoảng tin cậy mức 95% cho hiệu số của trọng lượng trung bình của hai nhóm bò.

b. Nếu bán mỗi kg thịt bò là 90 ngàn và chi phí thức ăn phụ trong 6 tháng cho mỗi con bò là 80 ngàn thì có nên dùng loại thức ăn phụ đó không.

Biết rằng trọng lượng của bò là biến ngẫu nhiên phân phối chuẩn.

ĐS: a. $-21{,}71 < \mu_1 - \mu_2 < 81{,}71$

7.30. Điều tra ngẫu nhiên thu nhập của 5 nữ nhân viên và 5 nam nhân viên đang làm việc tại các công ty liên doanh về thu nhập hàng năm của họ thu được các kết quả sau (đơn vị: triệu đồng).

Nữ	Nam
9	16
12	19
8	12
10	11
16	22

a. Với độ tin cậy 95% hãy ước lượng chênh lệch thu nhập trung bình hàng năm giữa nam và nữ nhân viên.

b. Kết quả thu được có phản ánh sự phân biệt đối xử đối với nữ nhân viên hay không.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Biết rằng thu nhập hàng năm phân phối chuẩn.

7.31. Để nghiên cứu ảnh hưởng của rượu đối với trí thông minh của trẻ em, người ta đo chỉ số thông minh của 8 đứa trẻ mà mẹ của chúng uống rượu trong thời gian mang thai, tìm được chỉ số thông minh trung bình của chúng là 78 và $\Sigma(x_i - \bar{x})^2 = 1805$. Một nhóm khác gồm 46 đứa trẻ có điều kiện kinh tế, xã hội giống như ở nhóm trên, chỉ khác biệt duy nhất ở chỗ mẹ của chúng không hề uống rượu khi mang thai. Kết quả đo được chỉ số thông minh trung bình của chúng là 99 và $\sum(x_i - \bar{x})^2 = 11520$.

Với độ tin cậy 95% hãy ước lượng sự chênh lệch giữa chỉ số thông minh trung bình của hai nhóm trẻ trên.

7.32. Để đánh giá hiệu quả của một chiến dịch quảng cáo, người ta so sánh doanh số của công ty tại 6 khu vực thị trường trước và sau chiến dịch quảng cáo và thu được các số liệu sau (đơn vị: triệu đồng/tháng).

Trước khi quảng cáo	Sau khi quảng cáo
620	660
600	620
640	670
630	620
570	580
600	630

Với độ tin cậy 95% hãy ước lượng mức tăng doanh số trung bình sau chiến dịch quảng cáo. Biết rằng doanh số của công ty là biến ngẫu nhiên phân phối chuẩn.

7.33. Để thử nghiệm hiệu quả của một loại thuốc trừ sâu người ta áp dụng thử loại thuốc này đối với 5 thửa ruộng đang bị sâu phá hoại. Số lượng sâu bắt được trước và sau khi dùng loại thuốc trừ sâu nói trên được cho ở bảng sau:

Thửa ruộng	Trước khi phun thuốc	Sau khi phun thuốc
1	109	107
2	68	72
3	82	88
4	104	101
5	93	97

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với độ tin cậy 95% hãy ước lượng số sâu giảm đi trung bình sau khi dùng thuốc. Giả sử số lượng sâu tại mỗi thửa ruộng đều phân phối chuẩn.

7.34. Kết quả khảo sát trong 30 ngày ở một dây chuyền lắp ráp chi tiết cho thấy thời gian dừng máy là biến ngẫu nhiên phân phối chuẩn với trung bình là 65,3 phút và độ lệch chuẩn là 7 phút mỗi ngày giữa các thao tác. Phòng kỹ thuật đã đưa ra một cải tiến quy trình kỹ thuật nhằm giảm thời gian dừng máy giữa các thao tác. Sau khi áp dụng quy trình mới này trong 30 ngày tìm được thời gian dừng máy trung bình là 60,4 phút và độ lệch chuẩn là 6 phút.

Với độ tin cậy 95% hãy ước lượng mức thời gian trung bình mà máy phải dừng đã được giảm đi sau cải tiến kỹ thuật đó.

7.35. Để nghiên cứu lượng tiền gửi tiết kiệm vào ngân hàng của hai thành phố người ta điều tra ngẫu nhiên 230 ngân hàng ở thành phố A và tìm được lượng tiền gửi trung bình của mỗi khách hàng là 1,317 triệu đồng. Ở thành phố B nghiên cứu 302 ngân hàng tìm được lượng tiền gửi trung bình của mỗi khách hàng là 1,512 triệu đồng. Với độ tin cậy 95% hãy ước lượng sự chênh lệch của số tiền gửi tiết kiệm trung bình của dân hai thành phố đó. Biết tiền gửi tiết kiệm của dân hai thành phố đó là biến ngẫu nhiên phân phối chuẩn với các độ lệch chuẩn tương ứng là 0,517 và 0,485 triệu.

ĐS: $0{,}10865 < \mu_1 - \mu_2 < 0{,}28135$

7.36. Điều tra 300 người nghiện thuốc thấy tuổi thọ trung bình của họ là 68,1 năm và độ lệch chuẩn là 8,2. Còn đối với 400 người không hút thuốc được điều tra ngẫu nhiên thì tuổi thọ trung bình của họ là 77,3 năm và độ lệch chuẩn là 5,3 năm. Hãy ước lượng với độ tin cậy 95% sự chênh lệch về tuổi thọ trung bình giữa những người không nghiện và những người nghiện thuốc lá. Biết tuổi thọ là biến ngẫu nhiên phân phối chuẩn.

ĐS: $7{,}8 < \mu_1 - \mu_2 < 10{,}6$

7.37. Nghiên cứu lãi suất ngân hàng của 7 nước công nghiệp phát triển và 11 nước đang phát triển trong 1 năm thì được kết quả như sau: Với các nước công nghiệp phát triển thì lãi suất trung bình là 17,5% và độ lệch chuẩn là 3,2% còn đối với các nước đang phát triển thì lãi suất trung bình là 15,3 và độ lệch chuẩn là 2,9%. Với độ tin cậy 95% hãy ước lượng sự chênh lệch về

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

lãi suất trung bình giữa hai nhóm nước nói trên. Biết rằng lãi suất ngân hàng phân phối chuẩn.

ĐS: $-0,35 < \mu_1 - \mu_2 < 4,75$

7.38. Điều tra ý kiến của 6 nhà kinh tế đang làm việc trong các cơ quan của chính phủ về mức lạm phát trong năm tới thu được các giá trị 4,2%, 5,1%, 3,9%, 4,7%, 4,9% và 5,8%. Trong khi đó hỏi ý kiến 8 nhà kinh tế đang làm việc tại các công ty cùng về vấn đề trên thì thu được các kết quả là 5,7%, 6,1%, 5,2%, 4,9%, 4,6%, 4,5%, 5,2% và 5,5%.

Với độ tin cậy 95% hãy ước lượng độ chênh lệch giữa mức lạm phát do các nhà kinh tế làm việc trong hai khu vực kinh tế nói trên dự báo. Giả thiết mức lạm phát là biến ngẫu nhiên phân phối chuẩn.

3.4. Ước lượng xác suất p của biến ngẫu nhiên phân phối theo quy luật không - một

Giả sử trong tổng thể biến ngẫu nhiên gốc X phân phối theo một quy luật nào đó khác với quy luật chuẩn. Trong trường hợp này, để ước lượng giá trị của kỳ vọng toán m chưa biết ta có thể chọn thống kê

$$G = U = \frac{(\overline{X} - m)\sqrt{n}}{\sigma}$$

nếu đã biết phương sai σ^2 của X, hoặc thống kê

$$G = U = \frac{(\overline{X} - m)\sqrt{n}}{S}$$

nếu chưa biết phương sai σ^2 của X.

Từ mục 7 chương VI ta đã biết nếu kích thước mẫu n đủ lớn, cả hai thống kê trên đều phân phối xấp xỉ chuẩn hóa $N(0,1)$. Do đó vẫn có thể tiến hành thủ tục ước lượng bằng khoảng tin cậy như đã xét ở phần trên. Sau đây ta sẽ xét một trường hợp cụ thể khá thông dụng trong thực tế là bài toán ước lượng kỳ vọng toán của biến ngẫu nhiên phân phối theo quy luật không - một: A(p).

Giả sử trong tổng thể kích thước N có M phần tử mang dấu hiệu nghiên cứu. Nếu lấy ngẫu nhiên ra 1 phần tử và gọi X là số phần tử mang dấu hiệu nghiên cứu được lấy ra thì X là biến ngẫu nhiên phân phối theo quy luật

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

không-một:

X	0	1
P	1-p	p

trong đó p là xác suất để lấy ngẫu nhiên một phần tử thì được phần tử mang dấu hiệu nghiên cứu

$$p = \frac{M}{N}$$

Ta đã biết trong quy luật không - một thì $E(X) = p$ và $V(X) = \frac{p(1-p)}{n}$, như vậy ước lượng kỳ vọng toán của quy luật này cũng chính là ước lượng xác suất p, mà p lại là tần suất của tổng thể, phản ánh cơ cấu của tổng thể theo dấu hiệu nghiên cứu đó. Như vậy đây chính là bài toán ước lượng cơ cấu của tổng thể. Từ mục 7 chương VI ta thấy nếu thỏa mãn điều kiện:

$$n \geq 5 \text{ và } \frac{\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right|}{\sqrt{n}} < 0,3$$

thì thống kê

$$G = U = \frac{(f-p)}{Se(f)} = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} \qquad (7.44)$$

phân phối xấp xỉ N(0,1). Vì vậy, với độ tin cậy $(1-\alpha)$ cho trước có thể tìm được hai giá trị tới hạn chuẩn $u_{1-\alpha/2}$ và $u_{\alpha/2}$ thỏa mãn điều kiện:

$$P(U < u_{1-\alpha/2}) = \frac{\alpha}{2}$$

$$P(U > u_{\alpha/2}) = \frac{\alpha}{2}$$

Từ đó $P(u_{1-\alpha/2} < U < u_{\alpha/2}) = 1-(\alpha/2+\alpha/2) = 1-\alpha$

Do $u_{\alpha/2} = -u_{1-\alpha/2}$ nên

$$P(-u_{\alpha/2} < U < u_{\alpha/2}) = 1-\alpha \qquad (7.45)$$

Phép biến đổi tương đương đối với biểu thức trong ngoặc của (7.45) được tiến hành như sau: Thay U từ (7.44) vào ta có:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$-u_{\alpha/2} < \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} < u_{\alpha/2}$$

tương đương với $\left|\frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}}\right| < u_{\alpha/2}$

Bình phương hai vế thu được $n(f-p)^2 < p(1-p)u_{\alpha/2}^2$. Khai triển và chuyển vế ta thu được bất phương trình:

$$(n+u_{\alpha/2}^2)p^2 - (2nf+u_{\alpha/2}^2)p + nf^2 < 0$$

Giải ra ta được:

$$p_1, p_2 = \frac{nf + \frac{1}{2}u_{\alpha/2}^2 \pm u_{\alpha/2}\sqrt{nf(1-f)+\frac{1}{4}u_{\alpha/2}^2}}{n+u_{\alpha/2}^2} \qquad (7.46)$$

Như vậy với độ tin cậy $(1-\alpha)$ khoảng tin cậy đối xứng của p là:

$$(p_1 < p < p_2) \qquad (7.47)$$

với p_1 và p_2 được xác định từ (7.46)

Việc áp dụng các công thức (7.46) và (7.47) khá phức tạp và chỉ cho phép tìm khoảng tin cậy hai phía của p. Do đó nếu có thể điều tra một mẫu có kích thước n khá lớn $(n \geq 100)$ thì ta có thể chọn thống kê

$$G = U = \frac{(f-p)\sqrt{n}}{\sqrt{f(1-f)}} \qquad (7.48)$$

Nó cũng phân phối xấp xỉ N(0,1), do đó với độ tin cậy $(1-\alpha)$ cho trước có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$, từ đó tìm được hai giá trị tới hạn chuẩn tương ứng là $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn điều kiện:

$$P(U < u_{1-\alpha_1}) = \alpha_1$$

và $$P(U > u_{\alpha_2}) = \alpha_2$$

từ đó $$P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1-(\alpha_1+\alpha_2) = 1-\alpha$$

Thay giá trị của U từ (7.48) vào và sử dụng tính chất $u_{\alpha_1} = -u_{1-\alpha_1}$ sau phép biến đổi tương đương ta có

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P\left(f-\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha_2}<p<f+\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha_1}\right)=1-\alpha \qquad (7.49)$$

Như vậy, với độ tin cậy $(1-\alpha)$ khoảng tin cậy của p có dạng:

$$\left(f-\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha_2};f+\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha_1}\right) \qquad (7.50)$$

Từ biểu thức tổng quát (7.50) có thể xây dựng được các khoảng tin cậy cụ thể:

- Khoảng tin cậy đối xứng khi $\alpha_1=\alpha_2=\frac{\alpha}{2}$

$$\left(f-\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha/2};f+\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha/2}\right) \qquad (7.51)$$

- Khoảng tin cậy bên phải khi $\alpha_1=0;\alpha_2=\alpha$

$$\left(f-\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha};+\infty\right) \qquad (7.52)$$

- Khoảng tin cậy bên trái khi $\alpha_2=0;\alpha_1=\alpha$

$$\left(-\infty;f+\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha}\right) \qquad (7.53)$$

- Độ dài khoảng tin cậy ngắn nhất trong trường hợp khoảng tin cậy đối xứng:

$$I=2\varepsilon=\frac{2\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha/2} \qquad (7.54)$$

- Kích thước mẫu tối thiểu n đảm bảo độ tin cậy $(1-\alpha)$ cho trước và độ dài khoảng tin cậy không vượt quá giá trị I_0 cho trước.

$$n\geq\left[\frac{f(1-f)}{\varepsilon_0^2}u_{\alpha/2}^2\right] \qquad (7.55)$$

trong đó f là tần suất của mẫu sơ bộ kích thước $m > 5$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Việc chuyển từ các khoảng tin cậy ngẫu nhiên sang các khoảng tin cậy bằng số qua một mẫu cụ thể cũng được tiến hành như đã làm ở các phần trên.

Thí dụ 8. Kiểm tra ngẫu nhiên 400 sản phẩm do một máy sản xuất thấy có 20 phế phẩm. với độ tin cậy 0,95 hãy ước lượng tỷ lệ phế phẩm tối đa của máy đó.

Giải. Gọi p là tỷ lệ phế phẩm của máy đó. Như vậy p là cơ cấu của tập hợp sản phẩm do máy đó sản xuất theo dấu hiệu " phế phẩm". Đây là bài toán ước lượng tham số p của quy luật phân phối A(p) bằng khoảng tin cậy bên trái. Vậy khoảng tin cậy của p có dạng (7.53):

$$\left(-\infty; f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_\alpha\right)$$

Qua mẫu cụ thể ta có $f = \frac{20}{400} = 0{,}05$. Với $1-\alpha = 0{,}95 \rightarrow u_{0{,}05} = 1{,}645$.

Vậy với độ tin cậy 0,95 qua mẫu cụ thể này khoảng tin cậy của p là:

$$\left(-\alpha; 0{,}05 + \frac{\sqrt{0{,}05 \cdot 0{,}95}}{\sqrt{400}} 1{,}645\right)$$

hay $p < 0{,}0679$, hay tỷ lệ phế phẩm tối đa của máy đó là 6,79%.

Thí dụ 9. Một vùng có 2000 hộ gia đình. Để điều tra nhu cầu tiêu dùng một loại hàng hóa tại vùng đó người ta nghiên cứu ngẫu nhiên 100 gia đình và thấy có 60 gia đình có nhu cầu về loại hàng hóa trên. Với độ tin cậy 0,95 hãy ước lượng bằng khoảng tin cậy đối xứng số gia đình trong vùng có nhu cầu về loại hàng hóa đó.

Giải. Gọi M là số gia đình trong vùng có nhu cầu về loại hàng hóa đó thì tỷ lệ gia đình có nhu cầu này là :

$$p = \frac{M}{2000}$$

Như vậy bài toán được đưa về việc ước lượng bằng khoảng tin cậy đối xứng cơ cấu p của tổng thể. Theo công thức (7.51) khoảng tin cậy của p là:

$$\left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2}; f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2}\right)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Qua mẫu cụ thể ta có $f = \frac{60}{100} = 0,6; u_{\alpha/2} = u_{0,025} = 1,96$

Vậy với độ tin cậy 0,95 khoảng tin cậy đối xứng của p qua mẫu cụ thể này là:

$$\left(0,6 - \frac{\sqrt{0,6 \cdot 0,4}}{\sqrt{100}} 1,96; 0,6 + \frac{\sqrt{0,6 \cdot 0,4}}{\sqrt{100}} 1,96\right)$$

hay $$(0,504 < p < 0,696)$$

Giải bằng Excel cho kết quả như sau:

95%	confidence level
0.6	proportion
100	n
1.960	z
0.096	half-width
0.696	upper confidence limit
0.504	lower confidence limit

Do $M = pN = p \cdot 2000$ nên ta có khoảng tin cậy đối xứng của M với độ tin cậy 0,95 qua mẫu cụ thể này là $(1008 \le M \le 1392)$.

Nếu sử dụng các công thức (7.46) và (7.47) thì kết quả là:

$$p_1 = \frac{0,6 \cdot 100 + \frac{1}{2}(1,96)^2 - 1,96\sqrt{0,6 \cdot 0,4 \cdot 100 + \frac{1}{4}(1,96)^2}}{100 + (1,96)^2} = 0,502$$

$$p_2 = \frac{0,6 \cdot 100 + \frac{1}{2}(1,96)^2 + 1,96\sqrt{0,6 \cdot 0,4 \cdot 100 + \frac{1}{4}(1,96)^2}}{100 + (1,96)^2} = 0,691$$

vậy $(0,502 < p < 0,691)$

do đó $(1004 \le M \le 1382)$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

7.39. Hãy ước lượng tỷ lệ chính phẩm của một nhà máy bằng khoảng tin cậy đối xứng với độ tin cậy 0,95 biết rằng kiểm tra 100 sản phẩm của nhà máy thì thấy có 10 phế phẩm.

ĐS: $0,8412 < p < 0,9588$

7.40. Bằng khoảng tin cậy đối xứng hãy ước lượng tỷ lệ hạt nảy mầm với độ tin cậy 95% trên cơ sở kết quả thực nghiệm: gieo 1000 hạt, có 860 hạt nảy mầm.

7.41. Mở thử 200 hộp của một kho đồ hộp, người ta thấy có 8 hộp bị biến chất. Với độ tin cậy 0,95, bằng khoảng tin cậy đối xứng, hãy ước lượng tỷ lệ đồ hộp biến chất ở kho đó.

ĐS: $0,0125 < p < 0,0675$

7.42. Thăm dò ý kiến của 200 người ở cơ quan A thấy có 150 người đồng ý về vấn đề B. Bằng khoảng tin cậy đối xứng hãy ước lượng tỷ lệ ý kiến đồng ý vấn đề B ở cơ quan đó với độ tin cậy $1 - \alpha = 0,95$

7.43. Gieo thử 400 hạt giống thì thấy có 20 hạt không nảy mầm. Tỷ lệ hạt giống không nảy mầm tối đa là bao nhiêu. Yêu cầu kết luận với độ tin cậy là 95%.

ĐS: $p \leq 0,068$

7.44. Trong đợt vận động bầu cử tổng thống người ta phỏng vấn ngẫu nhiên 1600 cử tri thì được biết 960 người trong số đó sẽ bỏ phiếu cho ứng cử viên A. Với độ tin cậy 99%, ứng cử viên A sẽ chiếm được tối thiểu bao nhiêu phần trăm số phiếu bầu?

7.45. Kiểm tra ngẫu nhiên 400 sản phẩm do một nhà máy sản xuất thấy có 160 sản phẩm loại 1. Hãy ước lượng tỷ lệ sản phẩm loại 1 tối đa của nhà máy với xác suất tin cậy 0,95.

ĐS: $p \leq 0,482$.

7.46. Tỷ lệ nảy mầm của một loại hạt giống là 90%.

Cần ước lượng tỷ lệ nảy mầm của loại hạt giống đó với độ tin cậy là 0,95 và độ dài khoảng tin cậy không quá 0,02 thì phải gieo bao nhiêu hạt.

ĐS: $n \geq 3456$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.5. Ước lượng hiệu hai tham số p của hai biến ngẫu nhiên phân phối không-một

Giả sử có hai tổng thể nghiên cứu, trong đó các biến ngẫu nhiên X_1 và X_2 phân phối không- một với tham số tương ứng là p_1 và p_2. Nếu từ hai tổng thể rút ra hai mẫu ngẫu nhiên độc lập kích thước n_1 và n_2 và tìm được các tần suất tương ứng là f_1 và f_2 thì có thể chọn lập thống kê

$$G = U = \frac{(f_1 - f_2) - (p_1 - p_2)}{S_f} \qquad (7.56)$$

với $$S_f = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \approx \sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}$$

Từ mục 7 chương VI ta đã biết nếu $n_1 > 30$ và $n_2 > 30$ thì biến ngẫu nhiên U sẽ phân phối xấp xỉ N(0,1). Do đó với cách tiến hành giống như ở mục trước ta thu được khoảng tin cậy mức $(1-\alpha)$ của hiệu $p_1 - p_2$ như sau:

$$P\left[(f_1 - f_2) - S_f u_{\alpha_2} < p_1 - p_2 < (f_1 - f_2) + S_f u_{\alpha_1}\right] = 1 - \alpha \qquad (7.57)$$

Xuất phát từ công thức (7.57) có thể xây dựng các khoảng tin cậy đối xứng, khoảng tin cậy bên phải và bên trái cho hiệu $p_1 - p_2$.

Công thức tìm kích thước mẫu tối thiểu cần điều tra với cả hai mẫu sao cho với độ tin cậy $1-\alpha$ cho trước, độ dài khoảng tin cậy không vượt quá $I_0 = 2\varepsilon_0$ cho trước có dạng:

$$n \geq \frac{\left[f_1(1-f_1) + f_2(1-f_2)\right]}{\varepsilon_0^2} u_{\alpha/2}^2 \qquad (7.58)$$

với f_1 và f_2 là tần suất của các mẫu sơ bộ kích thước m.

Thí dụ 10. Doanh nghiệp dự định đưa sản phẩm của mình vào hai thị trường khác nhau. Bán thử sản phẩm cho 100 khách hàng tiềm năng của thị trường thứ nhất thì có 50 người mua. Còn với thị trường thứ hai, khi bán thử sản phẩm cho 50 khách hàng tiềm năng thì có 20 người mua. Với độ tin cậy 95% hãy ước lượng mức độ chênh lệch về thị phần mà doanh nghiệp có thể đạt được tại hai thị trường đó.

Giải. Gọi p_1 và p_2 tương ứng là thị phần mà doanh nghiệp có thể đạt được ở hai thị trường. Vậy mức độ chênh lệch về thị phần là hiệu $p_1 - p_2$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với n_1 và n_2 đủ lớn, để ước lượng khoảng tin cậy đối xứng của $p_1 - p_2$, trong công thức (7.57) ta lấy $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ và thu được công thức ước lượng

$$P[(f_1 - f_2) - S_f u_{\alpha/2} < p_1 - p_2 < (f_1 - f_2) + S_f u_{\alpha/2}] = 1 - \alpha$$

Với hai mẫu cụ thể

$$n_1 = 100 \qquad n_2 = 50$$
$$f_1 = 0,5 \qquad f_2 = 0,4$$

Ta tìm được.

$$S_f = \sqrt{\frac{0,5.0,5}{100} + \frac{0,4.0,6}{50}} = 0,0854$$

Với $1 - \alpha = 0,95 \Rightarrow u_{\alpha/2} = u_{0,025} = 1,96$

Vậy

$$[(0,5 - 0,4) - 0,0854.1,96 < p_1 - p_2 < (0,5 - 0,4) + 0,0854.1,96]$$

$$[0,0674 < p_1 - p_2 < 0,2674]$$

Vậy với độ tin cậy 95%, thị phần mà doanh nghiệp có thể đạt được ở thị trường thứ nhất cao hơn ở thị trường thứ hai từ 6,74% đến 26,74%.

Bài tập

7.47. Vào năm 1990 phỏng vấn ngẫu nhiên 1500 người thì tỷ lệ hút thuốc là 43%. Sau thời gian dài vận động không hút thuốc lá, phỏng vấn 1000 người vào năm 1997 thì they tỷ lệ hút thuốc lá là 38%.

Với độ tin cậy 95% hãy ước lượng tỷ lệ người hút thuốc lá đã giảm xuống sau 7 năm vận động.

ĐS: $1,06\% < p_1 - p_2 < 8,94\%$

7.48. Kiểm tra ngẫu nhiên 100 sản phẩm của xí nghiệp A thì thấy có 7 phế phẩm. Ở xí nghiệp B kiểm tra ngẫu nhiên 150 sản phẩm thì thấy có 12 phế phẩm. Với độ tin cậy 95% hãy ước lượng sự chênh lệch về tỷ lệ phế phẩm giữa hai xí nghiệp.

7.49. Vào năm 1980 điều tra 1000 sản phụ ở một địa phương thì có 30 người là sinh con thứ ba. Để đánh giá kết quả của cuộc vận động kế

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

hoạch hoá gia đình, vào năm 1995 cũng tại địa phương đó, khi điều tra 800 sản phụ thì có 15 người là sinh con thứ ba. Với độ tin cậy 95% hãy ước lượng tỷ lệ sinh con thứ ba đã giảm xuống ở địa phương đó.

7.50. Để nghiên cứu việc chi tiêu cho một tang lễ người ta đã điều tra 162 gia đình công nhân có người nhà mới chết và thấy 61 gia đình phải chi phí hơn 2 triệu đồng cho một đám tang. Một cuộc điều ta tương tự với 189 gia đình thương nhân thì có 106 gia đình chi phí cho đám tang trên 2 triệu. Với độ tin cậy 95% hãy ước lượng chênh lệch về tỷ lệ các gia đình chi trên 2 triệu đồng cho một đám tang thuộc 2 tầng lớp xã hội là thương nhân và công nhân.

ĐS: $(-0{,}287 < p_1 - p_2 < -0{,}083$

7.51. Phỏng vấn ngẫu nhiên 150 sinh viên xuất thân từ nông thôn thì có 30 người đi làm thêm trong thời gian học ở trường. Tương tự số sinh viên đi làm thêm ở 140 sinh viên xuất thân từ thành phố là 17 người. Với độ tin cậy 95% hãy ước lượng xem tỷ lệ sinh viên nông thôn đi làm thêm cao hơn thành phố là bao nhiêu.

3.6. Ước lượng phương sai của biến ngẫu nhiên phân phối theo quy luật chuẩn

Giả sử trong tổng thể biến ngẫu nhiên gốc X phân phối theo quy luật chuẩn $N(\mu, \sigma^2)$ nhưng chưa biết phương sai σ^2 của nó. Để ước lượng σ^2 từ tổng thể lập mẫu ngẫu nhiên kích thước n

$$W = (X_1, X_2, ...X_n)$$

Để chọn thống kê G thích hợp ta xét hai trường hợp sau:

1. Đã biết kỳ vọng toán μ của biến ngẫu nhiên gốc X trong tổng thể. Lúc đó ta chọn thống kê

$$G = \chi^2 = \frac{nS^{*2}}{\sigma^2} \qquad (7.59)$$

Từ mục 7 chương VI ta đã biết thống kê χ^2 nói trên phân phối theo quy luật "khi bình phương" với n bậc tự do: $\chi^2(n)$. Do đó với độ tin cậy $(1-\alpha)$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

cho trước có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ từ đó tìm được hai giá trị tới hạn "khi bình phương" tương ứng là $\chi^{2(n)}_{1-\alpha_1}$ và $\chi_{\alpha_2}{}^{2(n)}$ thỏa mãn điều kiện:

$$P(\chi^2 < \chi_{1-\alpha_1}{}^{2(n)}) = \alpha_1 \text{ và}$$

$$P(\chi^2 > \chi_{\alpha_2}{}^{2(n)}) = \alpha_2 \qquad \text{do đó}$$

$$P(\chi_{1-\alpha_1}{}^{2(n)} < \chi^2 < \chi_{\alpha_2}{}^{2(n)}) = 1-(\alpha_1+\alpha_2) = 1-\alpha \qquad (7.60)$$

Thay giá trị của χ^2 từ (7.59) và (7.60) và giải ra theo σ^2, ta thu được biểu thức tương đương:

$$P\left(\frac{nS^{*2}}{\chi_{\alpha_2}{}^{2(n)}} < \sigma^2 < \frac{nS^{*2}}{\chi_{1-\alpha_1}{}^{2(n)}}\right) = 1-\alpha \qquad (7.61)$$

Như vậy, với độ tin cậy $(1-\alpha)$ khoảng tin cậy của σ^2 có dạng:

$$\left(\frac{nS^{*2}}{\chi_{\alpha_2}{}^{2(n)}} < \sigma^2 < \frac{nS^{*2}}{\chi_{1-\alpha_1}{}^{2(n)}}\right) \qquad (7.62)$$

Từ khoảng tin cậy tổng quát (7.62) có thể xây dựng được khoảng tin cậy trong một số trường hợp cụ thể sau:

- Nếu $\alpha_1 = \alpha_2 = \alpha/2$, khoảng tin cậy có dạng:

$$\left(\frac{nS^{*2}}{\chi_{\alpha/2}{}^{2(n)}}; \frac{nS^{*2}}{\chi_{1-\alpha/2}{}^{2(n)}}\right) \qquad (7.63)$$

Ta chú ý rằng khoảng tin cậy này không đối xứng.

- Nếu $\alpha_1 = 0$; $\alpha_2 = \alpha$, ta có khoảng tin cậy bên phải của σ^2:

$$\left(\frac{nS^{*2}}{\chi_{\alpha}{}^{2(n)}}; +\infty\right) \qquad (7.64)$$

- Nếu $\alpha_2 = 0$; $\alpha_1 = \alpha$, ta có khoảng tin cậy bên trái của σ^2:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$\left(0;\frac{nS^{*2}}{\chi_{1-\alpha}^{2(n)}}\right) \qquad (7.65)$$

Với một mẫu cụ thể $w=(x_1,x_2,...,x_n)$ có thể xác định được một khoảng tin cậy cụ thể bằng số của σ^2 giống như đã làm ở các phần trên.

Thí dụ 11. Mức hao phí nguyên liệu cho một đơn vị sản phẩm là biến ngẫu nhiên phân phối chuẩn với trung bình là 20 gam. Để ước lượng mức độ phân tán của mức hao phí này người ta cân thử 25 sản phẩm và thu được kết quả sau:

Hao phí nguyên liệu (gam)	19,5	20,0	20,5
Số sản phẩm tương ứng	5	18	2

Với độ tin cậy $1-\alpha=0,90$, hãy ước lượng σ^2 nếu:

$$\alpha_1=\alpha_2=\frac{\alpha}{2}=0,05$$

Giải. Gọi X là mức hao phí nguyên liệu cho 1 đơn vị sản phẩm, X phân phối chuẩn với kỳ vọng toán đã biết $\mu=20$. Đây là bài toán ước lượng phương sai của phân phối $N(\mu,\sigma^2)$ khi đã biết μ. Vậy theo công thức (7.63) khoảng tin cậy của σ^2 là:

$$\left(\frac{nS^{*2}}{\chi_{\alpha/2}^{2(n)}}<\sigma^2<\frac{nS^{*2}}{\chi_{1-\alpha/2}^{2(n)}}\right)$$

Tra bảng giá trị χ^2 ta có:

$$\chi_{\alpha/2}^{2(25)}=\chi_{0,05}^{2(25)}=37,65$$

$$\chi_{1-\alpha/2}^{2(25)}=\chi_{0,95}^{2(25)}=14,61$$

Để tìm s^{*2} ta lập bảng

x_i	n_i	$(x_i-\mu)$	$(x_i-\mu)^2$	$n_i(x_i-\mu)^2$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

19,5	5	- 0,5	0,25	1,25
20,0	18	0,0	0,00	0,00
20,5	2	0,5	0,25	0,50
	n = 25			$\sum = 1,75$

$$s^{*2} = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \mu)^2 = \frac{1,75}{25} = 0,07$$

Vậy với độ tin cậy 0,90, qua mẫu cụ thể này, khoảng tin cậy của σ^2 là:

$$\left(\frac{25.0,07}{37,65}; \frac{25.0,07}{14,61}\right) \text{ hay } (0,0464 < \sigma^2 < 0,1198)$$

Để tìm khoảng tin cậy của σ ta chỉ cần lấy căn bậc hai của cả ba vế trong các công thức ước lượng trên.

2. Chưa biết kỳ vọng toán μ của biến ngẫu nhiên gốc X trong tổng thể. Lúc đó ta chọn thống kê

$$G = \chi^2 = \frac{(n-1)S^2}{\sigma^2} \qquad (7.66)$$

Từ mục 7 chương VI ta đã biết thống kê χ^2 nói trên phân phối theo quy luật "khi bình phương" với $(n-1)$ bậc tự do. Vì vậy, với độ tin cậy $(1-\alpha)$ cho trước, có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$, từ đó tìm được hai giá trị tới hạn "khi bình phương" tương ứng là $\chi_{1-\alpha_1}{}^{2(n-1)}$ và $\chi_{\alpha_2}{}^{2(n-1)}$ thỏa mãn điều kiện:

$$P(\chi^2 < \chi_{1-\alpha_1}{}^{2(n-1)}) = \alpha_1 \text{ và}$$

$$P(\chi^2 > \chi_{\alpha_2}{}^{2(n-1)}) = \alpha_2 \quad \text{do đó}$$

$$P(\chi_{1-\alpha_1}{}^{2(n-1)} < \chi^2 < \chi_{\alpha_2}{}^{2(n-1)}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha \qquad (7.67)$$

Thay giá trị của χ^2 từ (7.66) vào (7.67) và giải ra σ^2 ta có:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$P\left(\frac{(n-1)S^2}{\chi_{\alpha_2}^{2(n-1)}} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\alpha_1}^{2(n-1)}}\right) = 1-\alpha$$

Vậy với độ tin cậy $(1-\alpha)$, khoảng tin cậy của σ^2 là:

$$\left(\frac{(n-1)S^2}{\chi_{\alpha_2}^{2(n-1)}}; \frac{(n-1)S^2}{\chi_{1-\alpha_1}^{2(n-1)}}\right) \qquad (7.68)$$

Từ công thức (7.68) ta cũng xây dựng được các khoảng tin cậy trong những trường hợp cụ thể sau:

- Nếu $\alpha_1 = \alpha_2 = \alpha/2$, khoảng tin cậy của σ^2 là:

$$\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^{2(n-1)}}; \frac{(n-1)S^2}{\chi_{1-\alpha/2}^{2(n-1)}}\right) \qquad (7.69)$$

- Nếu $\alpha_1 = 0$; $\alpha_2 = \alpha$, ta có khoảng tin cậy bên phải của σ^2:

$$\left(\frac{(n-1)S^2}{\chi_{\alpha}^{2(n-1)}}; +\infty\right) \qquad (7.70)$$

- Nếu $\alpha_2 = 0$; $\alpha_1 = \alpha$, ta có khoảng tin cậy bên trái của σ^2:

$$\left(0; \frac{(n-1)S^2}{\chi_{1-\alpha}^{2(n-1)}}\right) \qquad (7.71)$$

Việc xác định các khoảng tin cậy bằng số qua một mẫu cụ thể cũng tiến hành giống như đã làm ở các phần trên.

Thí dụ 12. Với độ tin cậy 0,95 hãy ước lượng phương sai của kích thước các chi tiết trên cơ sở các số liệu mẫu cho trong bảng (7.1), biết $\alpha_1 = \alpha_2 = \alpha/2 = 0{,}025$.

Giải. Đây là bài toán ước lượng phương sai của phân phối $N(\mu, \sigma^2)$ khi chưa biết μ. Vậy khoảng tin cậy của σ^2 có dạng:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Nếu $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, khoảng tin cậy của σ^2 là:

$$\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^{2(n-1)}}; \frac{(n-1)S^2}{\chi_{1-\alpha/2}^{2(n-1)}}\right)$$

Qua mẫu cụ thể ta đã tìm được:

$$s^2 = 0{,}0002689; n = 200$$

Tra bảng χ^2 : $\chi_{0,975}^{2(199)} \approx 198{,}98; \chi_{0,025}^{2(199)} \approx 284{,}8$

Vậy với độ tin cậy 0,95 qua mẫu cụ thể này, khoảng tin cậy của σ^2 là:

$$\left(\frac{199.0{,}0002689}{284{,}8}; \frac{199.0{,}0002689}{198{,}98}\right)$$

hay $(0{,}000188 < \sigma^2 < 0{,}000269)$.

Tương tự như ở mục trước, để tìm khoảng tin cậy của σ ta chỉ cần lấy căn bậc hai của các công thức ước lượng tương ứng.

Bài tập

7.52. Kiểm tra ngẫu nhiên 16 viên thuốc từ một lô thuốc mới nhập về tìm được độ phân tán của thành phần chính trong mỗi viên thuốc là $s^2 = 0{,}0775$ gr^2. Với độ tin cậy 95% hãy ước lượng độ phân tán của thành phần chính trong mỗi viên thuốc của cả lô thuốc đó.

Biết trọng lượng thành phần chính trong mỗi viên thuốc phân phối chuẩn.

ĐS: $0{,}176 < \sigma < 0{,}431$

7.53. Lãi suất cổ phiếu của một công ty trong vòng 5 năm qua là 15% 10% 20% 7% 14%. Với độ tin cậy 90% hãy ước lượng độ phân tán của lãi suất cổ phiếu của công ty đó. Biết lãi suất cổ phiếu là biến ngẫu nhiên phân phối chuẩn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7.54. Để nghiên cứu sự biến động của lượng sữa của mỗi con bò trong chu kỳ vắt sữa người ta lấy ngẫu nhiên 15 con bò và thu được các số liệu sau (đơn vị: lít)

12928	13812	11036
12120	14358	9248
14972	8989	9980
14004	10620	11990
14788	14744	14786

Với độ tin cậy 95% hãy ước lượng mức độ biến động của lượng sữa bò mỗi con trong chu kỳ vắt sữa. Biết lượng sữa của bò là biến ngẫu nhiên phân phối chuẩn.

7.55. Để nghiên cứu độ ổn định của một máy gia công, người ta lấy ngẫu nhiên 25 chi tiết do máy đó gia công, đem đo và thu được các kích thước sau:

24,1	27,2	26,7	23,6	26,4
25,8	27,3	23,2	26,9	27,1
22,7	26,9	24,8	24,0	23,4
24,5	26,1	25,9	25,4	22,9
26,4	25,4	23,3	23,0	24,3

Với độ tin cậy 95% hãy ước lượng độ phân tán của kích thước các chi tiết do máy đó gia công. Biết kích thước chi tiết được gia công là biến ngẫu nhiên phân phối chuẩn.

3.7. Ước lượng tỷ số của hai phương sai của hai biến ngẫu nhiên phân phối chuẩn

Giả sử có hai tổng thể nghiên cứu trong đó các biến X_1 và X_2 cùng phân phối chuẩn với các phương sai σ_1^2 và σ_2^2 chưa biết. Từ các tổng thể trên lập hai mẫu ngẫu nhiên độc lập kích thước n_1 và n_2, từ đó tìm được các phương sai mẫu S_1^2 và S_2^2 và tạo lập thống kê.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$G = F = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \qquad (S_1^2 > S_2^2) \tag{7.72}$$

Từ mục 7 chương VI ta đã biết thống kê F phân phối $F(n_1 - 1, n_2 - 1)$. Do đó với độ tin cậy $1 - \alpha$ cho trước tìm được cặp giá trị α_1 và $\alpha_2 (\alpha_1 + \alpha_2 = \alpha)$ từ đó tìm được hai giá trị tới hạn $f_{1-\alpha_1}^{(n_1-1,n_2-1)}$ và $f_{\alpha_2}^{(n_1-1,n_2-1)}$ thỏa mãn điều kiện

$$P(F < f_{1-\alpha_1}^{(n_1-1,n_2-1)}) = \alpha_1$$

$$P(F > f_{\alpha_2}^{(n_1-1,n_2-1)}) = \alpha_2$$

do đó
$$P(f_{1-\alpha_1}^{(n_1-1,n_2-1)} < F < f_{\alpha_2}^{(n_1-1,n_2-1)}) = 1 - \alpha \tag{7.73}$$

Thay giá trị của F từ (7.72) vào (7.73) và sau khi chuyển vế ta thu được biểu thức

$$P\left[\frac{S_2^2}{S_1^2} f_{1-\alpha_1}^{(n_1-1,n_2-1)} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2} f_{\alpha_2}^{(n_1-1,n_2-1)}\right] = 1 - \alpha$$

Từ đó

$$P\left[\frac{S_1^2}{S_2^2} f_{1-\alpha_2}^{(n_2-1,n_1-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha_1}^{(n_2-1,n_1-1)}\right] = 1 - \alpha \tag{7.74}$$

Từ (7.74) có thể xây dựng được các khoảng tin cậy tương ứng với mỗi cặp giá trị α_1 và α_2.

Việc tìm khoảng tin cậy của $\frac{\sigma_1}{\sigma_2}$ được thực hiện bằng cách lấy căn bậc hai của cả ba vế của biểu thức (7.74).

Thí dụ 13. Giá cổ phiếu của hai công ty A và B là các biến ngẫu nhiên phân phối chuẩn. Theo dõi giá cổ phiếu của hai công ty đó trong 10 ngày tìm được phương sai mẫu tương ứng là 0,51 và 0,2. Với độ tin cậy 0,9 hãy ước lượng tỷ số của hai phương sai của giá cổ phiếu của hai công ty đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Gọi X_1 và X_2 là giá cổ phiếu của hai công ty A và B. Theo giả thiết X_1 và X_2 phân phối chuẩn. Vậy các phương sai là ${\sigma_1}^2$ và ${\sigma_2}^2$. Từ (7.74) ta lấy $\alpha_1 = \alpha_2 = \dfrac{\alpha}{2}$ và thu được khoảng tin cậy hai phía sau đây:

$$P\left[\frac{S_1^2}{S_2^2} f_{1-\alpha/2}^{(n_2-1,n_1-1)} < \frac{{\sigma_1}^2}{{\sigma_2}^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}^{(n_2-1,n_1-1)}\right] = 1-\alpha$$

Từ hai mẫu cụ thể ta có:

$$n_1 = 10 \qquad n_2 = 10$$
$${s_1}^2 = 0{,}51 \qquad {s_2}^2 = 0{,}2$$

và với $1-\alpha = 0{,}9$ thì

$$f_{0,05}^{(9,9)} = 3{,}18 \text{ và } f_{0,95}^{(9,9)} = \frac{1}{f_{0,05}^{(9,9)}} = \frac{1}{3{,}18} = 0{,}31$$

Từ đó $$\frac{0{,}51}{0{,}2} \cdot 0{,}31 < \frac{{\sigma_1}^2}{{\sigma_2}^2} < \frac{0{,}51}{0{,}2} 3{,}18$$

$$\Rightarrow 0{,}79 < \frac{{\sigma_1}^2}{{\sigma_2}^2} < 8{,}11$$

Bài tập

7.56. Với các số liệu của bài tập 7.29 hãy ước lượng với độ tin cậy 95% tỷ số của hai phương sai tổng thể.

ĐS: $0{,}527 < {\sigma_1}^2 / {\sigma_2}^2 < 3{,}244$.

7.57. Với các số liệu của bài tập 7.32 hãy ước lượng với độ tin cậy 95% tỷ số của hai phương sai tổng thể.

7.58. Hai máy giống nhau cùng gia công một loại chi tiết. Lấy ngẫu nhiên 10 chi tiết của mỗi máy đem đo và tìm được các phương sai mẫu tương ứng là $10{,}5\text{mm}^2$ và $12{,}7667\text{mm}^2$. Với độ tin cậy 95% hãy ước lượng tỷ số giữa độ phân tán của 2 máy đó.

ĐS: $0{,}186 < \sigma_2^2 / \sigma_1^2 < 7{,}953$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.8. Ước lượng trung vị của tổng thể nghiên cứu

Cho đến nay, đối với các tham số đặc trưng xu hướng trung tâm của tổng thể nghiên cứu, ta mới chỉ quan tâm đến việc ước lượng trung bình mà chưa xét đến trung vị hay mốt. Đó cũng là điều hợp lý vì tham số của tổng thể mà ta quan tâm chủ yếu là tổng số mà tổng số lại gắn liền với trung bình. Chẳng hạn để đánh giá tổng doanh số bán của các đại lý chỉ cần tìm doanh số trung bình của mỗi đại lý sau đó đem nhân với tổng số đại lý là được.

Không thể ước lượng tổng doanh số dựa trên trung vị vì nó chỉ sử dụng giá trị ở chính giữa và bỏ qua mọi giá trị khác, nhất là những giá trị đột xuất, khác biệt so với các giá trị khác. Trong kinh tế, người ta cũng thường quan tâm đến giá trị tổng số như tổng sản phẩm quốc nội, tổng sản phẩm công nghiệp, nông nghiệp... Trong những trường hợp này, ta thường dùng trung bình vì nó chứa đựng thông tin về mọi phần tử của tổng thể nghiên cứu.

Tuy nhiên trong một số trường hợp trung vị tỏ ra hữu dụng. Chẳng hạn khi số liệu điều tra về tổng thể không đầy đủ hay không hoàn toàn tin cậy thì nên dùng trung vị thay cho trung bình. Hơn nữa trung vị có ưu thế hơn trung bình là nó không đòi hỏi biến ngẫu nhiên trong tổng thể phải phân phối chuẩn, do đó sẽ thuận lợi nhiều trong quy nạp thống kê.

Sau đây ta sẽ xem xét phương pháp ước lượng trung vị của tổng thể.

Để ước lượng trung vị m_d của tổng thể, người ta cũng xuất phát từ x_d của mẫu như là một ước lượng điểm của m_d. Tuy nhiên, để nâng cao độ chính xác của ước lượng ta sẽ đưa ra phương pháp để xây dựng khoảng tin cậy cho m_d.

Ta sẽ chứng minh việc tìm khoảng tin cậy cho m_d qua thí dụ sau: Giả sử lấy ngẫu nhiên 9 gia đình tại một thanh phố và điều tra thu nhập hàng năm X của họ thu được các số liệu được sắp xếp từ thấp đến cao như sau:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0		6 triệu		$x_d = 9$ triệu		13 triệu		20 triệu

từ đó ta có $x_d = x_5 = 9$ triệu. Để xây dựng khoảng tin cậy cho m_d ta có thể dịch chuyển sang hai phía, chẳng hạn hai giá trị, như vậy

$$6 \text{ triệu} \le m_d \le 13 \text{ triệu}. \qquad (7.75)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Song vấn đề là độ tin cậy của khoảng tin cậy (7.75) là bao nhiêu. Để giải đáp vấn đề này ta lập luận như sau: Nếu trung vị của thu nhập là m_d thì một nửa số gia đình của thành phố đó phải có thu nhập hàng năm lớn hơn giá trị đó. Song điều đó không có nghĩa là một nửa thu nhập của mẫu sẽ phải lớn hơn m_d mà chỉ có nghĩa là mỗi giá trị của mẫu ngẫu nhiên rút ra từ tổng thể có xác suất lớn hơn m_d là 0,5 giống như xác suất để được mặt sấp khi tung đồng xu. Vậy là ta có hai biến cố tương đương: "Giá trị của mẫu ngẫu nhiên lớn hơn m_d" tương đương với "được mặt sấp khi tung đồng xu".

Với quan niệm như vậy thì số gia đình có thu nhập cao hơn m_d sẽ là biến ngẫu nhiên phân phối theo quy luật nhị thức với các tham số là n (kích thước mẫu) và p = 0,5. Như vậy theo công thức của quy luật nhị thức, áp dụng vào thí dụ đang xét, có thể tìm được xác suất để có 7 gia đình trong 9 gia đình được điều tra sẽ có thu nhập lớn hơn hoặc bằng trung vị m_d. Nó được tính theo công thức:

$$P(X \geq 7) = P_9(7) + P_9(8) + P_9(9) = 0,09$$

Tương tự như vậy có thể tìm được xác suất để có 7 gia đình có thu nhập nhỏ hơn hoặc bằng m_d. Nó cũng bằng 0,09. Vậy xác suất để khoảng tin cậy (7.75) chứa đựng giá trị trung vị m_d là 1- 2.0,09 = 0,82.

Hiển nhiên là nếu n đủ lớn thì có thể dùng phân phối chuẩn để thay thế cho quy luật nhị thức. Chẳng hạn nếu điều tra thu nhập của 25 gia đình lấy ngẫu nhiên từ một thành phố và thu được các số liệu đã sắp xếp theo trình tự tăng dần như sau:

4, 5, 5, 5, 5, 6, 7, 9, 10, 12, 13, 14, 14, 15, 17, 18, 18, 19, 23, 25, 27, 30, 32, 89, 40, 52

Lúc đó trung vị mẫu là giá trị thứ 13: $x_{13} = 14$. Đó là ước lượng điểm của m_d.

Nếu ta chọn khoảng tin cậy của m_d là $7 \leq m_d \leq 27$

thì độ tin cậy của ước lượng là $1 - 2.P(X \geq 19)$

Ta tìm xác suất $P(X \geq 19)$. Với n đủ lớn X phân phối xấp xỉ chuẩn với

$$\mu = np = 25.0,5 = 12,5$$

và

$$\sigma = \sqrt{np(1-p)} = 2,5$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Do đó

$$P(X \geq 19) = P\left(\frac{X-12,5}{2,5} \geq \frac{19-12,5}{2,5}\right) = P(U \geq 2,6)$$

$$= 0,005 = \alpha \rightarrow \alpha/2 = 0,0025$$

Từ đó, độ tin cậy của ước lượng là $1 - 2.0,0025 = 0,9975$

Ngược lại nếu muốn độ tin cậy của ước lượng là 0,95 thì $\alpha = 0,05$

$\Rightarrow u_\alpha = 1,645 \Rightarrow X_c = 1,645.2,5 + 12,5 = 16.6 \approx 17$

Vậy $12 \leq m_d \leq 19$

Bài tập

7.59. Người ta kiểm tra ngẫu nhiên 20 cái chốt cửa do một máy sản xuất theo khả năng chịu lực của chúng thu được kết quả sau:

10,4	9,6	10,6	11,0	9,9
10,6	10,7	8,9	8,6	9,5
10,1	10,5	8,7	11,1	11,1
8,4	11,5	11,3	10,7	9,1

Với độ tin cậy 95% hãy ước lượng trung vị của khả năng chịu lực của chốt cửa do máy đó sản xuất

7.60. Điều tra ngẫu nhiên thu nhập hàng năm của 10 cặp vợ chồng thu được kết quả sau

(đơn vị: triệu đồng)

Thu nhập của chồng	Thu nhập của vợ
18	28
28	20
32	16
32	28
36	26
38	32
44	24
46	80
50	26
156	48

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Hãy xây dựng khoảng tin cậy phi tham số dưới dạng $X_2 < m_d < X_9$ cho trung vị của:

i. Thu nhập của chồng

ii. Thu nhập của vợ

iii. Chênh lệch thu nhập giữa chồng và vợ.

b. Độ tin cậy của mỗi khoảng tin cậy trên bằng bao nhiêu.

ĐS: a. i. $28 \le m_d \le 50$;

ii. $20 \le m_d \le 48$;

iii. $-10 \le m_d \le 24$

b. 0,9804

7.61. Trở lại bài 7.60. Nếu cho rằng thu nhập của chồng phân phối xấp xỉ chuẩn thì hãy ước lượng trung vị của tổng thể với độ tin cậy 98%.

7.62. Lấy ngẫu nhiên 8 gia đình có hai con (1 trai, 1 gái) đã lớn và đo chiều cao của con trai và con gái thu được kết quả sau (xem bảng)

a. Hãy xây dựng khoảng tin cậy phi tham số dưới dạng $x_2 < m_d < x_7$ cho trung vị của

1. Chiều cao của con trai.
2. Chiều cao của con gái.
3. Chênh lệch chiều cao giữa con trai và con gái.

b. Tìm độ tin cậy cho mỗi khoảng tin cậy nói trên.

Chiều cao con trai	Chiều cao con gái
165	163
167	162
169	134
170	165
171	168
173	166
176	171
177	169

7.63. Tiếp tục bài 7.62. Nếu cho rằng chiều cao của con trai phân phối xấp xỉ chuẩn thì với độ tin cậy 0,93 hãy ước lượng trung vị của tổng thể.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4. TÓM TẮT CHƯƠNG 7

1. Phương pháp ước lượng là một trong hai phương pháp cơ bản để tiến hành quá trình quy nạp thống kê. Nó dùng thông tin đã biết của mẫu rút ra từ tổng thể để suy đoán về các tham số của tổng thể đó.
2. Phương pháp hàm ước lượng là phương pháp ước lượng điểm chủ trương thay thế tham số chưa biết của tổng thể bằng một tham số của mẫu nếu tham số này thoả mãn một cách tốt nhất các tính chất không chệch, hiệu quả nhất, vững và đủ.
3. Phương pháp ước lượng hợp lý tối đa là phương pháp ước lượng điểm chủ trương thay thế tham số chưa biết của tổng thể bằng một tham số của mẫu nếu tham số này làm cho hàm hợp lý đạt cực đại.
4. Phương pháp ước lượng bằng khoảng tin cậy là sự mở rộng của phương pháp ước lượng điểm. Xuất phát từ một ước lượng điểm tốt nhất, nó đi xây dựng một khoảng tin cậy cho tham số chưa biết của tổng thể. Từ đó theo nguyên lý xác suất lớn có thể tìm được một khoảng tin cậy cụ thể cho tham số cần ước lượng.
5. So với phương pháp ước lượng điểm, phương pháp ước lượng bằng khoảng tin cậy không những nâng cao được độ chính xác của ước lượng mà còn chỉ ra được độ tin cậy của các kết luận bằng bao nhiêu.
6. Để ước lượng trung bình tổng thể người ta thường sử dụng trung bình mẫu, song nếu các số liệu bị thiếu hoặc không hoàn toàn đáng tin cậy thì trong nhiều trường hợp dùng trung vị lại tỏ ra chính xác hơn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

* Ước lượng không chệch

$$E(\hat{\theta}) = \theta$$

* Ước lượng vững

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1$$

* Ước lượng hiệu quả nhất: là ước lượng không chệch $\hat{\theta}$ có $V(\hat{\theta})$ đạt min.

* Bất đẳng thức Cramer- Rao:

$$V(\hat{\theta}) \geq \frac{1}{nE\left[\frac{\partial \ln f(x,\theta)}{\partial\theta}\right]^2}$$

* Hàm hợp lý

$$L(x_1, x_2, ..., x_n, \theta) = \prod_{i=1}^{n} f(x_i.\theta)$$

* Khoảng tin cậy của μ:

+ Trong trường hợp đã biết σ^2

$$P\left[\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_1}\right] = 1 - \alpha$$

Độ dài khoảng tin cậy $I = \frac{2\sigma}{\sqrt{n}} u_{\alpha/2}$

Kích thước mẫu n sao cho $I \leq I_0; n \geq \frac{\sigma^2}{\varepsilon_0^2}(u_{\alpha/2})^2$

+ Trường hợp chưa biết σ^2

$$P\left[\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha_2}^{(n-1)} < \mu < \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha_1}^{(n-1)}\right] = 1 - \alpha$$

Độ dài khoảng tin cậy: $I = \frac{2S}{\sqrt{n}} t_{\alpha/2}^{(n-1)}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Kích thước mẫu n sao cho $I \le I_0; n \ge \frac{S^2}{\varepsilon_0^2}(t_{\alpha/2}^{(m-1)})^2$

* Khoảng tin cậy của σ^2

+ Trường hợp đã biết μ:

$$P\left[\frac{nS^{*2}}{\chi^{2(n-1)}_{\alpha_2}} < \sigma^2 < \frac{nS^{*2}}{\chi^{2(n-1)}_{1-\alpha_1}}\right] = 1-\alpha$$

+ Trường hợp chưa biết μ:

$$P\left[\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha_2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha_1}}\right] = 1-\alpha$$

* Khoảng tin cậy của p

+ Trường hợp $n < 100$; $P(p_1 < p < p_2) = 1-\alpha$

Trong đó

$$p_1, p_2 = \frac{2nf + u^2_{\alpha/2} \pm u_{\alpha/2}\sqrt{4nf(1-f) + u^2_{\alpha/2}}}{2(n + u^2_{\alpha/2})}$$

+ Trường hợp $n \ge 100$

$$P\left[f - \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha_2} < p < f + \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha_1}\right] = 1-\alpha$$

Độ dài khoảng tin cậy

$$I = \frac{2\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha/2}$$

Kích thước mẫu n sao cho $I \le I_0; n \ge \frac{f(1-f)}{\varepsilon_0^2}(u_{\alpha/2})^2$

* Khoảng tin cậy cho hiệu $\mu_1 - \mu_2$

+ Trường hợp đã biết σ_1^2 và σ_2^2

$$P\left[(\overline{X}_1 - \overline{X}_2) - u_{\alpha_2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + u_{\alpha_1}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1-\alpha$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Trường hợp chưa biết σ_1^2 và σ_2^2 song giả thiết $\sigma_1^2 = \sigma_2^2$:

$$P\left[(\overline{X}_1 - \overline{X}_2) + t_{\alpha_2}^{(n_1+n_2-2)} S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_{\alpha_1}^{(n_1+n_2-2)} S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = 1 - \alpha$$

trong đó $S_p = \sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$

+ Trường chưa biết σ_1^2 và σ_2^2 song giả thiết $\sigma_1^2 \neq \sigma_2^2$

$$P\left[(\overline{X}_1 - \overline{X}_2) + t_{\alpha_2}^{(k)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_{\alpha_1}^{(k)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right] = 1 - \alpha$$

+ Trường hợp hai mẫu phụ thuộc theo cặp

$$P\left[\overline{D} - \frac{S_D}{\sqrt{n}} t_{\alpha_2}^{(n-1)} < \mu_D < \overline{D} + \frac{S_D}{\sqrt{n}} t_{\alpha_1}^{(n-1)}\right] = 1 - \alpha$$

* Khoảng tin cậy cho hiệu $p_1 - p_2$

$$P[(f_1 - f_2) - S_f . u_{\alpha_2} < p_1 - p_2 < (f_1 - f_2) + S_f . u_{\alpha_1}] = 1 - \alpha$$

trong đó $S_f = \sqrt{\dfrac{f_1(1-f_1)}{n_1} + \dfrac{f_2(1-f_2)}{n_2}}$

* Khoảng tin cậy cho tỷ số $\dfrac{\sigma_1^2}{\sigma_2^2}$

$$P\left[\frac{S_1^2}{S_2^2} f_{1-\alpha_2}^{(n_2-1;n_1-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha_2}^{(n_2-1;n_1-1)}\right] = 1 - \alpha$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Phương pháp ước lượng	Estimation
Phương pháp ước lượng điểm	Point estimation
Phương pháp hàm ước lượng	Method moment
Ước lượng không chệch	Unbiased estimator
Ước lượng chệch	Biased estimator
Độ chệch	Bias
Ước lượng tiệm cận không chệch	Assimptotically unbiased estimator
Ước lượng hiệu quả	Efficient estimator
Ước lượng hiệu quả nhất	The most efficient estimator
Bất đẳng thức Cramer - Rao	Cramer- Rao inegality
Độ hiệu quả	Efficacity
Ước lượng vững	Consistent estimator
Ước lượng đủ	Sufficient estimator
Phương pháp ước lượng hợp lý tối đa	Maximum log likelihood estimation
Hàm hợp lý	Likelihood function
Phương pháp ƯL khoảng tin cậy	Confidence interval estimation
Khoảng tin cậy	Confidence interval
Độ tin cậy	Confidence coefficient (level)
Độ chính xác (sai số) của ước lượng	Precision (Error) of estimation
Khoảng tin cậy hai phía	Two side confidence interval
Khoảng tin cậy đối xứng	Symmetric confidence interval
Khoảng tin cậy bên phải	Right confident interval
Khoảng tin cậy bên trái	Left confidence interval
Kích thước mẫu	Sample size
Phương pháp mẫu kép	Double samples
Mẫu lớn	Large sample
Mẫu nhỏ	Small sample
Mẫu theo cặp	Mached samples

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

7.64. Để ước lượng tham số θ của tổng thể, tại sao người ta dùng thống kê $\hat{\theta}$ của mẫu thỏa mãn các điều kiện

a. Không chệch b. Vững

c. Hiệu quả nhất d. Đủ

7.65. Tại sao trung vị lại là một ước lượng kém hiệu quả hơn trung bình?

7.66. Có sự khác biệt nào trong kết luận của phương pháp hàm ước lượng và phương pháp ước lượng hợp lý tối đa?

7.67. Một nhân viên chọn ngẫu nhiên một mẫu n = 12 hóa đơn trong số các hóa đơn bán hàng của công ty và thu được các giá trị sau (đơn vị: Ngàn đồng): 875, 1231, 453, 522, 2130, 1550, 309, 760, 498, 999, 1320, 1021. Hãy tìm một ước lượng của giá trị trung bình của các hóa đơn bán hàng và ước lượng của phương sai của các giá trị của hóa đơn bán hàng.

7.68. Ở câu hỏi 4, giả sử nhân viên muốn ước lượng tỷ lệ các hóa đơn có giá trị trên 1000. Hãy tìm ước lượng của tham số đó.

7.69. Một nhân viên nghiên cứu thị trường phỏng vấn ngẫu nhiên 18 người xem họ có (C) hay không (K) tiêu dùng một loại sản phẩm và thu được dãy câu trả lời như sau: CKKCCCKCKCCCKCKCCK. Hãy ước lượng tỷ lệ khách hàng tiêu dùng loại sản phẩm đó.

7.70. Cho biến ngẫu nhiên X phân phối chuẩn với $\sigma^2 = 1$. Hãy tìm mối liên hệ giữa độ tin cậy, độ chính xác của ước lượng và kích thước mẫu trong trường hợp ước lượng μ bằng khoảng tin cậy đối xứng.

7.71. Với độ tin cậy 0,95 người ta ước lượng được tham số θ của tổng thể nằm trong khoảng từ 62 đến 69. Từ đó có thể kết luận được rằng xác suất để θ nằm trong khoảng (62;69) bằng 0,95 được không? Tại sao?

7.72. Một chủ cửa hàng muốn ước lượng lợi nhuận trung bình hàng ngày của cửa hàng đó. Người đó lấy mẫu n=60 ngày để ước lượng song không chắc lợi nhuận hàng ngày có phân phối chuẩn hay không do đó cho rằng kết quả ước lượng là không chính xác. Hãy bình luận về điều đó.

7.73. Giám đốc xí nghiệp giao cho nhân viên A ước lượng bằng khoảng tin cậy 95% tỷ lệ phế phẩm của xí nghiệp với một mẫu sản phẩm kích thước

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

n=100. Sau đó đề nghị nhân viên B ước lượng lại với một mẫu n=100 sản phẩm khác và thu được kết quả khác. Giám đốc cho rằng kết quả ước lượng của một trong hai nhân viên là sai. Vậy hai nhân viên A và B phải làm gì để thuyết phục giám đốc?

7.74. Một chủ cửa hàng muốn ước lượng bằng khoảng tin cậy 95% số khách hàng trung bình vào cửa hàng của ông ta mỗi ngày. Song khoảng tin cậy thu được quá rộng nên mất ý nghĩa. Nếu chủ cửa hàng không muốn thay đổi độ tin cậy của ước lượng thì phải làm gì?

Bài tập tổng hợp chương 7

7.75. Để thử nghiệm một công nghệ mới trong sản xuất gạch chịu lửa người ta đem đo độ chịu lửa của 5 viên gạch được sản xuất theo công nghệ mới và 5 viên theo công nghệ cũ và thu được kết quả sau:

Gạch cũ	Gạch mới
475	485
436	438
495	493
483	486
426	433

a. Với độ tin cậy 95% hãy ước lượng nhiệt độ chịu lửa trung bình của loại gạch cũ và gạch mới.

b. Cũng với độ tin cậy trên hãy ước lượng nhiệt độ chịu lửa được tăng thêm của loại gạch mới so với gạch cũ.

Giả thiết độ chịu lửa của gạch là biến ngẫu nhiên phân phối chuẩn.

ĐS: a. (425,415; 500,585); (431,029; 502,971)

b. $-3{,}924 < \mu_2 - \mu_1 < 11{,}924$

7.76. Để ước lượng tỷ lệ phế phẩm p của một lô hàng với độ tin cậy 99% và sai số của ước lượng không quá 0,005 thì phải kiểm tra tối thiểu bao nhiêu sản phẩm.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7.77. Để ước lượng khối lượng gỗ khai thác được từ 380 cây gỗ vừa khai thác người ta đem đo ngẫu nhiên 10% số cây và thu được kết quả sau

Khối lượng gỗ mỗi cây (m^3)	20	40	60	80
Số cây	8	16	12	2

Với độ tin cậy 95% hãy ước lượng:

a. Khối lượng gỗ trung bình mỗi cây

b. Tổng khối lượng gỗ khai thác được.

Giả thiết khối lượng gỗ mỗi cây là biến ngẫu nhiên phân phối chuẩn.

7.78. Để đánh giá trữ lượng cá trong hồ, người ta đánh bắt 2000 con cá, đánh dấu rồi thả xuống hồ. Sau đó bắt lại 400 con thì thấy 80 con có dấu.

a. Với độ tin cậy 95% hãy ước lượng trữ lượng cá hiện có trong hồ.

b. Nếu muốn sai số của ước lượng giảm đi một nửa thì lần sau phải đánh bắt bao nhiêu con cá.

ĐS: a. $8362 \le N \le 12437$

b. 1600 con.

7.79. Rủi ro đầu tư thường được đo bằng phương sai của tỷ lệ thu hồi vốn của dự án. Theo dõi ngẫu nhiên tỷ lệ thu hồi vốn của hai dự án trong 10 năm thu được kết quả sau

	Dự án 1	Dự án 2
Kích thước mẫu	10	10
Tỷ lệ thu hồi trung bình (%)	13,2	14,6
Phương sai mẫu $(\%)^2$	10,9	25,6

a. Với độ tin cậy 95% hãy ước lượng tỷ lệ thu hồi vốn trung bình của hai dự án trên.

b. Hãy ước lượng phương sai của tỷ lệ thu hồi vốn của 2 dự án đó.

Biết rằng tỷ lệ thu hồi vốn là biến ngẫu nhiên phân phối chuẩn.

7.80. Hợp tác xã trường đại học đã điều tra ngẫu nhiên 144 sinh viên về chi phí cho giáo trình ở năm thứ nhất và tìm được chi phí trung bình là 190 ngàn, độ lệch chuẩn là 30 ngàn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Với độ tin cậy 95% hãy ước lượng chi phí trung bình cho giáo trình của sinh viên năm thứ nhất.

b. Nếu với độ tin cậy 95% mà muốn độ chính xác của ước lượng là 3 ngàn thì phải điều tra bao nhiêu sinh viên.

Biết chi phí cho giáo trình là biến ngẫu nhiên phân phối chuẩn.

ĐS: a. 185,1 μ < 194,9

b. 385 sinh viên

7.81. Kết quả thi tốt nghiệp môn toán của 68 học sinh chuyên ban và 64 học sinh không chuyên ban được điều tra ngẫu nhiên cho trong bảng sau:

	Học sinh chuyên ban	Học sinh không chuyên ban
Kích thước mẫu	68	64
Trung bình	7,66	7,32
Độ lệch chuẩn	1,14	1,09

Với độ tin cậy 95% hãy ước lượng sự chênh lệch điểm thi trung bình của học sinh chuyên ban và không chuyên ban.

7.82. Trong các trường hợp sau đây, hãy tìm ước lượng bằng phương pháp hàm ước lượng và ước lượng hợp lý tối đa; nếu giá trị của chúng khác nhau thì cho biết ước lượng nào tốt hơn.

a. Để kiểm tra chất lượng thép ống, người ta lấy ngẫu nhiên 6 thanh thép ống từ lô sản phẩm mới sản xuất, đo độ cứng của chúng và thu được các giá trị 4,30; 4,38; 4,20; 4,24 và 4,20. Giả thiết độ cứng của các thanh thép ống phân phối chuẩn, hãy ước lượng độ cứng trung bình của các thanh thép được sản xuất.

b. Cũng 6 thanh thép trên được đem đo độ dài để xem có đáp ứng độ dài quy định không thì thấy có 2 thanh có độ dài thấp hơn độ dài quy định. Hãy ước lượng tỷ lệ độ dài các thanh thép được sản xuất có độ dài thấp hơn quy định.

c. Để xác định số khách hàng đến một cửa hàng, người ta theo dõi ngẫu nhiên 10 ngày và thu được các số liệu sau:

22 21 14 32 15 19 21 27 16 27

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giả sử số khách hàng vào cửa hàng mỗi ngày là biến ngẫu nhiên phân phối chuẩn, hãy ước lượng tổng số khách hàng đến cửa hàng trong 1 năm, biết trong 1 năm cửa hàng mở cửa 310 ngày.

7.83. Một nhà sản xuất tủ lạnh tuyên bố rằng tỷ lệ tủ lạnh có hỏng hóc của họ không quá 0,03. Kiểm tra ngẫu nhiên 170 tủ lạnh do công ty này sản xuất thấy có 12 chiếc hỏng hóc.

a. Với độ tin cậy 95% hãy ước lượng tỷ lệ tối thiểu các tủ lạnh bị hỏng hóc của công ty đó.

b. Nếu một năm công ty đó sản xuất 1000 cái tủ lạnh thì số tủ lạnh bị hỏng hóc tối đa là bao nhiêu với độ tin cậy 95%.

ĐS: a. $p \geq 0{,}0377$

b. 102

7.84. Một nghiên cứu đối với 15 nhà đầu tư nước ngoài được chọn ngẫu nhiên cho thấy vốn đầu tư trung bình của họ là 11,32 triệu USD và độ lệch chuẩn là 4,4 triệu USD. Với độ tin cậy 95% hãy ước lượng số vốn đầu tư trung bình mà các nhà đầu tư nước ngoài bỏ vào các dự án đầu tư.

Biết số vốn đầu tư vào các dự án là biến ngẫu nhiên phân phối chuẩn.

7.85. Một xí nghiệp sản xuất ắc quy biết rằng độ lệch chuẩn của tuổi thọ của ắc quy do họ sản xuất là 4,5 giờ. Để ước lượng tuổi thọ trung bình của các ắc quy do họ sản xuất với độ tin cậy 95% và sai lệch không quá 1 giờ thì phải kiểm tra tối thiểu bao nhiêu ắc quy.

7.86. Điều tra ngẫu nhiên 128 điểm bán thịt lợn trong thành phố tìm được giá trung bình là 25.500 đồng/kg và độ lệch chuẩn là 6.500 đồng. Với độ tin cậy 99% hãy ước lượng giá thịt lợn trung bình trong thành phố.

ĐS: $24020 < \mu < 26980$

7.87. Nếu trong bài 7.86 ta biết rằng trong thành phố có 500 điểm bán thịt lợn thì kết quả ước lượng ở trên có cần thiết điều chỉnh lại không.

7.88. Vào năm 1963 trong số 200 người Mỹ được phỏng vấn ngẫu nhiên thì có 160 người ủng hộ chính sách của tổng thống Mỹ Giôn Xơn. Đến năm 1968 do sa lầy vào chiến tranh Việt Nam, trong 200 người được hỏi ý kiến chỉ có 70 người ủng hộ chính sách của ông Giôn Xơn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a. Với độ tin cậy 95% hãy ước lượng tỷ lệ dân Mỹ ủng hộ chính sách của ông Giôn Xơn vào năm 1963, 1968.

b. Với độ tin cậy 95% tỷ lệ dân chúng ủng hộ ông Giôn Xơn giữa hai năm đã giảm đi bao nhiêu phần trăm.

7.89. Trước một chiến dịch quảng cáo, tỷ lệ phòng có khách ở trong một khách sạn được theo dõi ngẫu nhiên trong 15 ngày và thu được kết quả là 86 92 83 88 79 81 90 76 80 91 85 89 77 91 và 83 phần trăm. Sau khi kết thúc chiến dịch quảng cáo người ta theo dõi tỷ lệ này trong 16 ngày và thu được kết quả là 88 94 97 99 89 93 92 98 89 90 97 91 87 80 88 và 96 phần trăm.

Giả sử tỷ lệ thuê phòng là phân phối chuẩn. Hãy ước lượng với độ tin cậy 95% hiệu quả của chiến dịch quảng cáo mang lại.

ĐS: $3,171\% < \mu_2 - \mu_1 < 10,869\%$

7.90. Để kiểm tra chất lượng của một lô lớn các màn hình máy tính xuất khẩu, người ta đã lấy ngẫu nhiên 100 màn hình để kiểm tra chất lượng và thấy 4 màn hình có khuyết tật.

a. Với độ tin cậy 99% hãy ước lượng tỷ lệ màn hình có khuyết tật của lô hàng đó.

b. Cũng với độ tin cậy trên hãy ước lượng số màn hình có khuyết tật tối đa nếu lô hàng đó có 10000 màn hình.

7.91. Một doanh nghiệp có dự định đưa một sản phẩm mới vào một thị trường có 1.500.000 người tiêu dùng. Nghiên cứu thị trường đối với 2500 khách hàng tiềm năng thấy 800 người sẵn sàng mua sản phẩm đó.

a. Với độ tin cậy 95% hãy ước lượng thị phần tiềm năng của doanh nghiệp.

b. Số lượng khách hàng tiềm năng mà doanh nghiệp hy vọng sẽ có được ở thị trường mới là bao nhiêu.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Nguyễn Tấn Lập - *Giáo trình thống kê toán*. ĐH Kinh tế Kế hoạch. Hà Nội. 1972
2. Nguyễn Đình Cử, Nguyễn Cao Văn. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Thống kê. Hà Nội. 1991.
3. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Thống kê. Hà Nội.2005.
4. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB ĐH Kinh tế Quốc dân. Hà Nội.2008.
5. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà Nội.2009.
6. Ngô Văn Thứ. *Thống kê thực hành*. NXB Khoa học và Kỹ thuật. Hà Nội. 2008.
7. Ngô Văn Thứ. *Thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội. 2011
8. Harald Cramer. *Phương pháp toán học trong thống kê*. NXB Khoa học và Kỹ thuật. Hà Nội. 1970
9. Гмурман В.Е. Теория вероятностей и математическая статистика. Москва. 1972.
10. Уйякс. Математическая статистика. Москва. 1982.
11. Baillageon G et Rainville J. Introduction à la Statistique appliquée. Québec. 1976
12. Giard V. Statistique appliquée à la gestion. Economica. Paris. 1992
13. Amir D Aczel. Complete Business Statistics. Irwin Boston 1993.
14. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992
15. OTT L.R. And introduction to statistical methods and data analysis. California. 1993
16. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons. New york. 1990
17. Paul Newbold: Statistics for Business & Economics (Fourth Edition). Pretice-Hall International, Inc.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 8
KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

1. MỞ ĐẦU

Ở chương 7, ta đã nghiên cứu các tham số đặc trưng của tổng thể trên cơ sở thông tin của mẫu bằng phương pháp ước lượng. Chương này tiếp tục nghiên cứu dấu hiệu của tổng thể bằng một phương pháp khác là kiểm định giả thuyết thống kê. Với những thông tin bổ sung thay vì ước lượng các tham số của tổng thể người ta đưa ra một giả thuyết về tổng thể và dùng thông tin của mẫu để kiểm định tính đúng đắn của nó, nhờ đó mà phương pháp này cho phép giải quyết nhiều bài toán đa dạng hơn liên quan đến dấu hiệu nghiên cứu trong tổng thể.

Mục 2 sẽ đề cập đến phương pháp luận tổng quát của việc kiểm định một giả thuyết thống kê. Mục 3 sẽ nghiên cứu các bài toán kiểm định tham số có nhiều ứng dụng trong phân tích kinh tế xã hội.

2. KHÁI NIỆM CHUNG

2.1. Giả thuyết thống kê

Giả sử dấu hiệu nghiên cứu trong tổng thể có thể xem như biến ngẫu nhiên X. nếu chưa biết dạng phân phối xác suất của nó, song có cơ sở để giả thiết rằng X phân phối theo một quy luật A nào đó, người ta đưa ra giả thuyết: biến ngẫu nhiên X phân phối theo quy luật A.

Cũng có trường hợp dạng phân phối xác suất của X đã biết song tham số đặc trưng của nó lại chưa biết, nếu có cơ sở để giả thiết rằng giá trị của tham số θ bằng θ_0 (θ_0 là hằng số đã biết), người ta đưa ra giả thuyết: $\theta = \theta_0$.

Khi nghiên cứu hai hay nhiều biến ngẫu nhiên thuộc các tổng thể khác nhau hay thuộc cùng một tổng thể thường phải xét xem chúng độc lập hay phụ thuộc nhau, các tham số đặc trưng của chúng có bằng nhau hay không. Nếu chưa biết một cách chắc chắn song có cơ sở để nhận định về các vấn đề đó cũng có thể đưa ra các giả thuyết tương ứng.

Từ đó có thể đưa ra định nghĩa sau:

Định nghĩa. *Giả thuyết thống kê là giả thuyết về dạng phân phối xác suất của biến ngẫu nhiên, về các tham số đặc trưng của biến ngẫu nhiên hoặc về tính độc lập của các biến ngẫu nhiên.*

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giả thuyết thống kê đưa ra được ký hiệu là H_0 và được gọi là *giả thuyết gốc*.

Khi đưa ra một giả thuyết thống kê, người ta còn nghiên cứu kèm theo nó mệnh đề mâu thuẫn với nó, gọi là *giả thuyết đối* và ký hiệu là H_1 để khi giả thuyết H_0 bị bác bỏ thì thừa nhận giả thuyết H_1. H_0 và H_1 tạo nên cặp giả thuyết thống kê.

Chẳng hạn ta nghiên cứu nhu cầu thị trường về một loại hàng hóa nào đó. Ta có thể đưa ra các cặp giả thuyết thống kê sau:

* H_0: Nhu cầu X của thị trường phân phối theo quy luật chuẩn; H_1: Nhu cầu X của thị trường không phân phối theo quy luật chuẩn.

* H_0: Nhu cầu trung bình về loại hàng hóa này là $\mu = 1000$ đơn vị/tháng, lúc đó các giả thuyết đối tương ứng với nó có thể là

$H_1: \mu > 1000$; $H_1: \mu < 1000$ hoặc $H_1: \mu \neq 1000$.

* H_0: Nhu cầu X của thị trường và thu nhập Y của khách hàng độc lập nhau; H_1: X và Y phụ thuộc nhau...

Trong thực tế người ta còn phân biệt các giả thuyết chứa đựng một mệnh đề hoặc nhiều mệnh đề.

Giả thuyết *đơn* là giả thuyết chỉ chứa đựng một mệnh đề. Chẳng hạn nếu μ là tham số của quy luật chuẩn và ta đưa ra giả thuyết $H_0: \mu = 5$ thì đó là giả thuyết đơn.

Giả thuyết *hợp* là giả thuyết chứa đựng một số hữu hạn hoặc vô hạn các giả thuyết đơn. Chẳng hạn giả thuyết $H_0: \mu > 5$ bao gồm một số vô hạn các giả thuyết đơn dạng $H_0: \mu = b_i$ trong đó b_i là mọi số lớn hơn 5.

Việc kiểm định một giả thuyết hợp thường khá phức tạp, do đó ở đây ta chỉ hạn chế ở việc nghiên cứu trường hợp giả thuyết gốc là giả thuyết đơn.

Vì các giả thuyết thống kê có thể đúng hoặc có thể sai nên cần kiểm định, tức là tìm ra kết luận về tính thừa nhận hay không thừa nhận được của giả thuyết đó. Việc kiểm định này gọi là *kiểm định thống kê* vì nó dựa vào thông tin thực nghiệm của mẫu để kết luận.

Phương pháp chung để kiểm định một giả thuyết thống kê như sau: Trước hết giả sử H_0 đúng và từ đó dựa vào thông tin của mẫu rút ra từ tổng thể tìm được một biến cố A nào đó sao cho xác suất xảy ra biến cố A bằng α bé đến mức có thể sử dụng nguyên lý xác suất nhỏ tức là có thể coi A

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

không xảy ra trong một phép thử về biến cố này. Lúc đó trên một mẫu cụ thể thực hiện một phép thử đối với biến cố A, nếu A xảy ra thì điều đó chứng tỏ H_0 sai và bác bỏ nó, còn nếu A không xảy ra thì ta chưa có cở sở để bác bỏ H_0.

Để cụ thể hóa phương pháp trên ta nghiên cứu một số khái niệm sau:

2.2. Tiêu chuẩn kiểm định giả thuyết thống kê

Từ biến ngẫu nhiên gốc X trong tổng thể lập mẫu ngẫu nhiên kích thước n

$$W = (X_1, X_2, ..., X_n)$$

và chọn lập thống kê

$$G = f(X_1, X_2, ..., X_n, \theta_0)$$

trong đó θ_0 là tham số liên quan đến giả thuyết cần kiểm định. Điều kiện đặt ra đối với thống kê G là nếu H_0 đúng thì quy luật phân phối xác suất của G hoàn toàn xác định. Thống kê G được gọi là tiêu chuẩn kiểm định.

2.3. Miền bác bỏ giả thuyết

Sau khi đã chọn được tiêu chuẩn kiểm định G, do quy luật phân phối xác suất của G đã biết nên với một xác suất khá bé bằng α cho trước (thường α được lấy bằng 0,05 hoặc 0,01) có thể tìm được miền W_α tương ứng sao cho với điều kiện giả thuyết H_0 là đúng xác suất để G nhận giá trị thuộc miền W_α bằng α. Điều kiện này được viết như sau:

$$P(G \in W_\alpha / H_0) = \alpha \qquad (8.1)$$

Biến cố $(G \in W_\alpha)$ đóng vai trò như biến cố A nói trên và vì α khá bé nên theo nguyên lý xác suất nhỏ có thể coi như nó không xảy ra trong một phép thử.

Giá trị α được gọi là *mức ý nghĩa* của kiểm định và miền W_α được gọi là *miền bác bỏ* giả thuyết H_0 với mức ý nghĩa α. Hiển nhiên với một mức ý nghĩa α cho trước có thể tìm được vô số miền bác bỏ tương ứng.

Lúc đó miền giá trị còn lại của G, ký hiệu là $\overline{W}_\alpha$ được gọi là miền không bác bỏ giả thuyết (đôi khi để cho đơn giản người ta gọi là miền thừa nhận giả thuyết). Điểm tới hạn phân chia miền bác bỏ và miền không bác bỏ gọi là giá trị tới hạn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.4. Giá trị quan sát của tiêu chuẩn kiểm định

Thực hiện một phép thử đối với mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$ thu được một mẫu cụ thể $w = (x_1, x_2, ..., x_n)$ và qua đó tính được một giá trị cụ thể của tiêu chuẩn kiểm định G:

$$G_{qs} = f(x_1, x_2, ..., x_n, \theta_\alpha)$$

Giá trị này được gọi là giá trị quan sát của tiêu chuẩn kiểm định.

2.5. Quy tắc kiểm định giả thuyết thống kê

Sau khi đã tính được giá trị quan sát G_{qs} của tiêu chuẩn kiểm định, ta so sánh giá trị này với miền bác bỏ W_α và kết luận theo quy tắc sau:

1. Nếu giá trị quan sát của tiêu chuẩn kiểm định thuộc miền bác bỏ $G_{qs} \in W_\alpha$ thì điều đó có thể giải thích rằng là H_0 sai và do đó ta bác bỏ H_0, thừa nhận H_1.

2. Nếu giá trị quan sát của tiêu chuẩn kiểm định không thuộc miền bác bỏ $G_{qs} \notin W_\alpha$ thì điều đó chưa khẳng định được rằng H_0 đúng mà chỉ có nghĩa là qua mẫu cụ thể này chưa khẳng định được rằng H_0 sai. Do đó ta chỉ có thể nói: Qua mẫu cụ thể này chưa có cơ sở để bác bỏ H_0 (trên thực tế là vẫn thừa nhận H_0).

2.6. Sai lầm loại một và sai lầm loại hai

Với quy tắc kiểm định như trên có thể mắc hai loại sai lầm:

1. Sai lầm loại 1: Bác bỏ giả thuyết H_0 trong khi H_0 đúng. Ta thấy xác suất mắc phải loại sai lầm này đúng bằng mức ý nghĩa α. Thật vậy, mặc dù H_0 đúng thì xác suất để $(G \in W_\alpha)$ vẫn bằng α. Nhưng nếu $G \in W_\alpha$ thì ta lập tức bác bỏ H_0. Như vậy, ta có thể mắc sai lầm loại 1 với xác suất bằng α. Sai lầm này có thể sinh ra do kích thước mẫu quá nhỏ, do phương pháp lấy mẫu v.v...

2. Sai lầm loại 2. Thừa nhận giả thuyết H_0 trong khi nó sai, hay giá trị quan sát G_{qs} không thuộc miền bác bỏ W_α trong khi H_1 đúng.

Giả sử xác suất mắc sai lầm loại 2 là β:

$$P(G \notin W_\alpha / H_1) = \beta \qquad (8.2)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lúc đó biến cố không mắc sai lầm loại 2 là biến cố để G thuộc miền bác bỏ và do đó ta bác bỏ H_0 trong khi H_1 đúng:

$$G \in W_\alpha / H_1$$

Biến cố này đối lập với biến cố $(G \notin W_\alpha / H_1)$ nên xác suất của nó là:

$$P(G \in W_\alpha / H_1) = 1 - \beta \qquad (8.3)$$

xác suất $1-\beta$ được gọi là *lực kiểm định.*

Quan hệ giữa việc kiểm định giả thuyết và các loại sai lầm có thể mô tả trong bảng sau: (Bảng 8.1)

Bảng 8.1

Tình huống / Quyết định	H_0 đúng	H_0 sai
Bác bỏ H_0	Sai lầm loại 1 xác suất $= \alpha$	Quyết định đúng xác suất $= 1-\beta$
Không bác bỏ H_0	Quyết định đúng xác suất $= 1-\alpha$	Sai lầm loại 2 xác suất $= \beta$

Ta thấy rằng sai lầm loại 1 và sai lầm loại 2 mâu thuẫn nhau, tức là với một mẫu kích thước n xác định thì không thể cùng một lúc giảm xác suất mắc cả hai loại sai lầm nói trên được. Khi ta giảm α đi thì đồng thời sẽ làm tăng β và ngược lại. Chẳng hạn nếu lấy $\alpha = 0$ thì sẽ không bác bỏ bất kỳ giả thuyết nào, kể cả giả thuyết sai, như vậy β sẽ đạt cực đại.

Trong thực tế người ta tiến hành như sau: Sau khi đã ấn định một mức ý nghĩa α và với mẫu kích thước n xác định thì trong vô số các miền bác bỏ W_α tương ứng có thể tìm được, ta chọn ra miền bác bỏ W_α sao cho xác suất mắc sai lầm loại 2 là nhỏ nhất hay lực kiểm định là lớn nhất.

Như vậy, cần tìm miền bác bỏ W_α thỏa mãn các điều kiện sau:

$$P(G \in W_\alpha / H_0) = \alpha \text{ cho trước}$$

và

$$P(G \in W_\alpha / H_1) = 1 - \beta \to \max.$$

Dựa vào định lý Neyman - Pearson được trình bày trong các tài liệu đầy đủ hơn có thể tìm được những miền bác bỏ " tốt nhất" như vậy.

Những miền bác bỏ được xây dựng trong những phần dưới đây đều thỏa mãn điều kiện nói trên, tức là đều là những miền bác bỏ "tốt nhất" với mức ý nghĩa và kích thước mẫu xác định trước.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tuy nhiên, vẫn còn lại vấn đề là lựa chọn mức ý nghĩa bằng bao nhiêu?

Điều này phụ thuộc vào từng trường hợp cụ thể và căn cứ vào "hậu quả" mà sai lầm loại 1 và sai lầm loại 2 mang lại.

2.7. Thủ tục kiểm định giả thuyết thống kê

Qua nội dung được trình bày ở phần trên có thể xây dựng một thủ tục để kiểm định giả thuyết thống kê. Trên thực tế việc kiểm định giả thuyết thống kê có thể tiến hành theo hai thủ tục khác nhau.

1. Kiểm định với giá trị cho trước của α

Khi chỉ kiểm soát khả năng mắc sai lầm loại 1 thì thủ tục kiểm định được tiến hành như sau:

a) Xây dựng giả thuyết gốc H_0 cần kiểm định.

b) Từ tổng thể nghiên cứu lập mẫu ngẫu nhiên kích thước n.

c) Chọn tiêu chuẩn kiểm định G và xác định quy luật phân phối xác suất của nó với điều kiện giả thuyết H_0 là đúng.

d) Với mức ý nghĩa α cho trước xác định miền bác bỏ tốt nhất tùy thuộc vào giả thuyết đối H_1.

e) Lập mẫu cụ thể và tìm được giá trị của tiêu chuẩn kiểm định trên mẫu.

g) So sánh giá trị quan sát của tiêu chuẩn kiểm định với miền bác bỏ và kết luận.

h) Đánh giá xác suất mắc sai lầm loại 2 theo các giá trị khác nhau của H_1.

2. Kiểm định với giá trị cho trước của α và β

Khi kiểm soát cả khả năng mắc sai lầm loại 1 và loại 2 thì thủ tục kiểm định được tiến hành như sau:

a) Xây dựng giả thuyết H_0 cần kiểm định.

b) Chọn tiêu chuẩn kiểm định G và xác định quy luật phân phối xác suất của nó với điều kiện giả thuyết H_0 là đúng.

c) Với α và β cho trước xác định kích thước mẫu cần điều tra để việc kiểm định phạm hai sai lầm trên với xác suất không vượt quá mức cho trước.

d) Theo kết quả ở phần c, điều tra một mẫu cụ thể và tiến hành tiếp như ở trường hợp trước.

Sau đây ta sẽ vận dụng các thủ tục kiểm định trên vào một số giả thuyết thống kê thông dụng hơn cả trong nghiên cứu kinh tế - xã hội.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3. KIỂM ĐỊNH THAM SỐ

3.1. Kiểm định giả thuyết về kỳ vọng toán của biến ngẫu nhiên phân phối theo quy luật chuẩn khi đã biết phương sai

Giả sử biến ngẫu nhiên gốc X trong tổng thể phân phối theo quy luật chuẩn $N(\mu,\sigma^2)$ với phương sai đã biết nhưng chưa biết kỳ vọng toán μ. Nếu có cơ sở để giả thiết rằng giá trị của nó bằng μ_0 ta đưa ra giả thuyết thống kê $H_0: \mu = \mu_0$. Để kiểm định giả thuyết trên từ tổng thể lập mẫu kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

Vì đã biết phương sai σ^2 của biến ngẫu nhiên gốc X trong tổng thể nêu tiêu chuẩn kiểm định được chọn là thống kê.

$$G = U = \frac{(\overline{X} - \mu_0)}{Se(\overline{X})} = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma} \qquad (8.4)$$

Nếu giả thuyết H_0 đúng thì ta có:

$$U = \frac{(\overline{X} - \mu_0)}{Se(\overline{X})} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}$$

và từ mục 7 Chương VI ta có U phân phối N(0,1).

1. Kiểm định với giá trị cho trước của α. Nếu cho trước mức ý nghĩa α thì tùy thuộc vào dạng của giả thuyết đối H_1 miền bác bỏ "tốt nhất" được xây dựng theo các trường hợp sau:

a) $H_0: \mu = \mu_0; H_1: \mu > \mu_0$. Lúc đó với α cho trước có thể tìm được giá trị tới hạn chuẩn u_α sao cho

$$P(G \in W_\alpha / H_0) = P(U > u_\alpha) = \alpha$$

Ta thu được miền bác bỏ bên phải W_α được xác định bằng biểu thức:

$$W_\alpha = \left\{ U = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma}; U > u_\alpha \right\} \qquad (8.5)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b) $H_0: \mu = \mu_0; H_1: \mu < \mu_0$

Lúc đó với mức ý nghĩa α cho trước có thể tìm được giá trị tới hạn chuẩn $u_{1-\alpha}$ sao cho

$$P(G \in W_\alpha / H_0) = P(U < u_{1-\alpha}) = P(U < -u_\alpha) = \alpha$$

Ta thu được miền bác bỏ bên trái W_α được xác định bằng biểu thức:

$$W_\alpha = \left\{ U = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma}; U < -u_\alpha \right\} \qquad (8.6)$$

c) $H_0: \mu = \mu_0; H_1: \mu \neq \mu_0$

Lúc đó với mức ý nghĩa α cho trước có thể tìm được hai giá trị tới hạn chuẩn là $u_{1-\alpha/2}$ và $u_{\alpha/2}$ sao cho

$$\begin{aligned} P(G \in W_\alpha / H_0) &= P(U < u_{1-\alpha/2}) + P(U > u_{\alpha/2}) \\ &= P(U < -u_{\alpha/2}) + P(U > u_{\alpha/2}) \\ &= P(|U| > u_{\alpha/2}) = \alpha \end{aligned}$$

Ta thu được miền bác bỏ hai phía được xác định bằng biểu thức:

$$W_\alpha = \left\{ U = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma}; |U| > u_{\alpha/2} \right\} \qquad (8.7)$$

Lập mẫu cụ thể $w_\alpha = (x_1, x_2, ..., x_n)$ và tính giá trị quan sát tiêu chuẩn kiểm định

$$U_{qs} = \frac{(\overline{x} - \mu_0)\sqrt{n}}{\sigma}$$

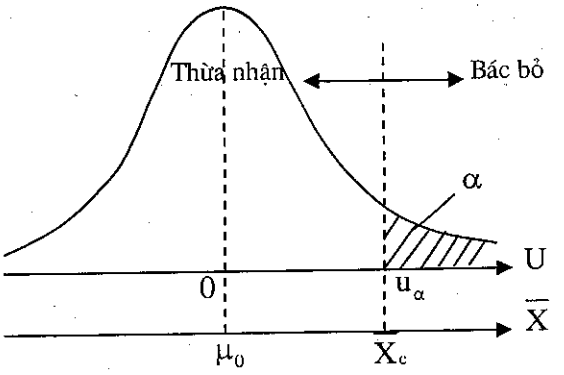
và so sánh với W_α để kết luận:

- Nếu $U_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1:

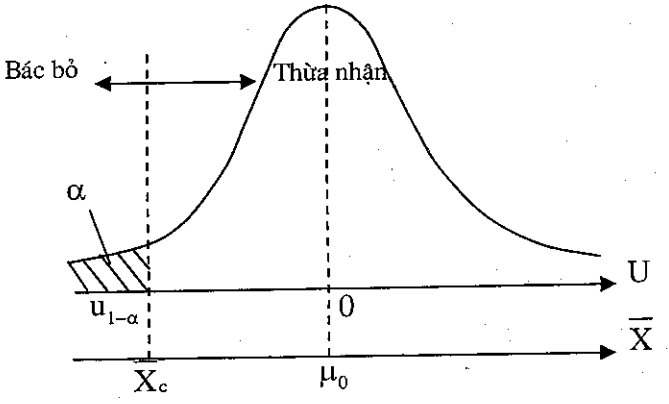
- Nếu $U_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Các miền bác bỏ được xây dựng theo các công thức (8.5) (8.6) và (8.7) có thể mô tả trên đồ thị như sau (H.8.1) trong đó $\overline{x}_c$ là giá trị tới hạn tương ứng của $\overline{x}$ khi $U = u_\alpha$.

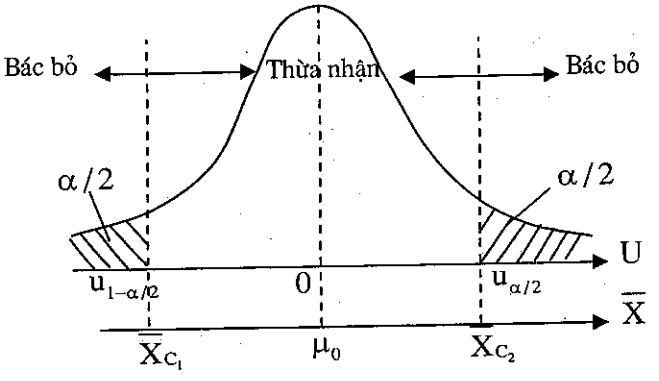
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Hình 8.1a. **Miền bác bỏ bên phải**



Hình 8.1.b. **Miền bác bỏ bên trái**



Hình 8.1.c. **Miền bác bỏ hai phía**

Việc kiểm định bằng cách sử dụng tiêu chuẩn (8.4) thường được gọi là kiểm định U (U-test hoặc Z - test).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 1. Trong năm trước trọng lượng trung bình trước khi xuất chuồng của bò ở một trại chăn nuôi là 380 kg. Năm nay người ta áp dụng thử một chế độ chăn nuôi mới với hy vọng là bò sẽ tăng trọng nhanh hơn. Sau thời gian áp dụng thử người ta lấy ngẫu nhiên 50 con bò trước khi xuất chuồng đem cân và tính được trọng lượng trung bình của chúng là 390 kg. Vậy với mức ý nghĩa $\alpha = 0,01$ có thể cho rằng trọng lượng trung bình của bò trước khi xuất chuồng đã tăng lên hay không? Giả thiết trọng lượng của bò là biến ngẫu nhiên phân phối chuẩn với độ lệch chuẩn là 35,2 kg.

Giải. Gọi X là trọng lượng của bò trước khi xuất chuồng. Theo giả thiết X phân phối chuẩn với $\sigma = 35,2$. Vậy trọng lượng xuất chuồng trung bình là μ. Đây là bài toán kiểm định giá trị của tham số μ của biến ngẫu nhiên phân phối chuẩn khi đã biết phương sai của tổng thể. Vậy từ (8.5) ta có:

Cặp giả thuyết thống kê có dạng:

$$H_0 : \mu = 380; H_1 : \mu > 380$$

Tiêu chuẩn kiểm định

$$U = \frac{(\overline{X} - 380)\sqrt{50}}{35,2}$$

trong đó $\overline{X}$ là trung bình mẫu ngẫu nhiên kích thước n= 50.

Cũng từ (8.5), với $\alpha = 0,01$ ta có:

$$u_\alpha = u_{0,01} = 2,33$$

Vậy miền bác bỏ bên phải là $(2,33; +\infty)$.

Từ mẫu cụ thể ta có $\overline{x} = 390$. Vậy giá trị quan sát của tiêu chuẩn kiểm định.

$$U_{qs} = \frac{(390 - 380)\sqrt{50}}{35,2} = 2,01$$

Như vậy $U_{qs} \notin W_\alpha$.

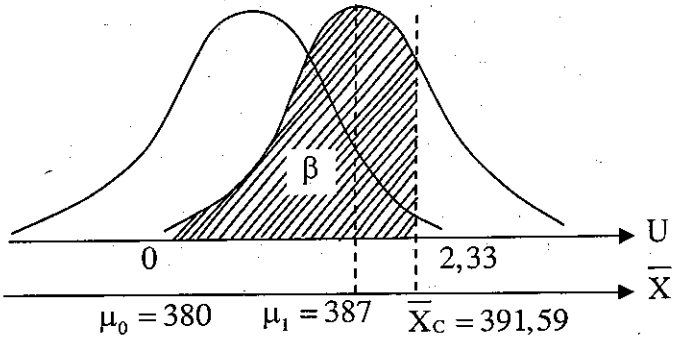
Kết luận: Với mức ý nghĩa $\alpha = 0,01$, qua mẫu cụ thể đã cho ta chưa có cơ sở để bác bỏ H_0 (trên thực tế là vẫn chấp nhận $H_0 : \mu = 380$). Kết quả trên cũng cho thấy trung bình mẫu thu được qua mẫu cụ thể đã cho không khác biệt một cách có ý nghĩa so với trung bình tổng thể.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

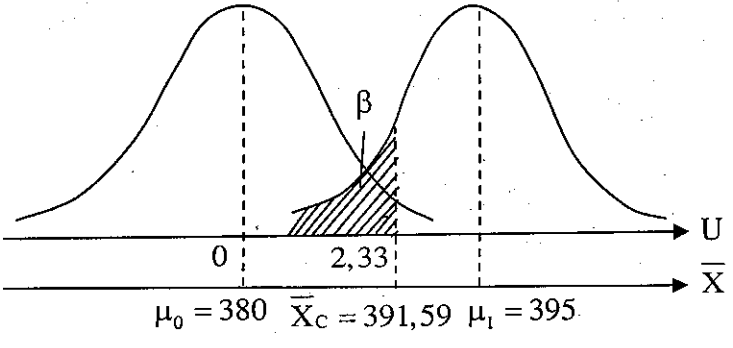
2. Tìm β

Ta cũng có thể tính tiếp giá trị β là xác suất để mắc sai lầm loại 2. Nếu giả thuyết $H_0 : \mu = 380$ thì xác suất để thừa nhận giả thuyết sai H_0 sẽ phụ thuộc vào việc giá trị thực của μ sai lệch nhiều hay ít so với 380. Chẳng hạn giá trị thực của μ là 400 thì β sẽ nhỏ hơn so với trường hợp giá trị thực của μ là 387. Vậy tùy thuộc vào giá trị thực của μ mà ta có các giá trị β khác nhau. Điều đó được minh họa trên hình (8.2) với 3 giá trị thực của μ là 387, 395 và 400.

Giả sử ta vẫn kiểm định cặp giả thuyết $H_0 : \mu = 380$; $H_1 : \mu > 380$ và với mức ý nghĩa $\alpha = 0{,}01$ ta xác định β. Phần gạch chéo trên hình (8.2) biểu diễn giá trị của β vì đó chính là xác suất để $\overline{X}$ rơi vào miền thừa nhận khi giả thuyết H_0 sai và giá trị thực của μ là 387, 395 và 400. Lúc đó lực kiểm định giả thuyết $1 - \beta$ chính là miền còn lại nằm trong miền bác bỏ H_0.

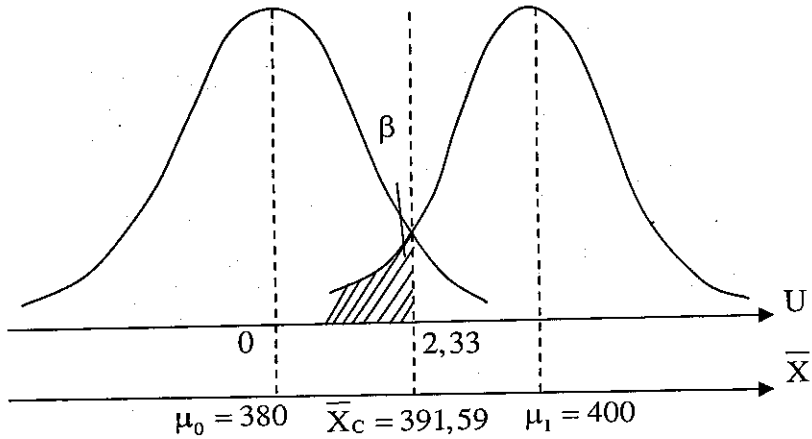


Hình 8.2a. *β* **khi** $H_1 : \mu = 387$



Hình 8.2b. β khi $H_1 : \mu = 395$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Hình 8.2c. β khi $H_1 : \mu = 400$

Nếu ký hiệu như trước đây μ_0 là giá trị giả thuyết của μ và μ_1 là giá trị thực của μ thì với miền bác bỏ bên phải giá trị β được xác định như sau:

Phải kiểm định cặp giả thuyết $H_0 : \mu = \mu_0$; $H_1 : \mu > \mu_0$.

Do H_0 sai nên biến ngẫu nhiên $\dfrac{\overline{X} - \mu_0}{Se(\overline{X})}$ không có phân phối N(0,1).

Thay vào đó, biến ngẫu nhiên $\dfrac{\overline{X} - \mu_1}{Se(\overline{X})}$ sẽ phân phối N(0,1).

Vì thế:

$$\beta = P\left(\frac{\overline{X} - \mu_0}{Se(\overline{X})} < u_\alpha\right) = P\left(\frac{\overline{X} - \mu_0}{Se(\overline{X})} - \frac{\mu_1}{Se(\overline{X})} < u_\alpha - \frac{\mu_1}{Se(\overline{X})}\right)$$

$$= P\left(\frac{\overline{X} - \mu_1}{Se(\overline{X})} < u_\alpha - \frac{\mu_1 - \mu_0}{Se(\overline{X})}\right) = P\left(U < u_\alpha - \frac{\mu_1 - \mu_0}{Se(\overline{X})}\right)$$

$$\Rightarrow \beta = P\left(U < u_\alpha - \frac{\mu_1 - \mu_0}{Se(\overline{X})}\right) \qquad (8.8)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bằng cách chứng minh tương tự, ta thu được biểu thức của β khi miền bác bỏ là bên trái

$$\beta = P\left[U < u_\alpha - \frac{\mu_0 - \mu_1}{Se(\overline{X})}\right] \quad (8.9)$$

Từ đó ta có công thức chung để tìm xác suất mắc sai lầm loại hai β khi miền bác bỏ một phía (bên phải hoặc bên trái) như sau:

$$\beta = P\left[U < u_\alpha - \frac{|\mu_0 - \mu_1|}{Se(\overline{X})}\right] \quad (8.10)$$

Nếu miền bác bỏ là hai phía thì β được xác định bằng công thức:

$$\beta = P\left[U < u_{\alpha/2} - \frac{|\mu_0 - \mu_1|}{Se(\overline{X})}\right] \quad (8.11)$$

Từ đó suy ra giá trị của lực kiểm định $1-\beta$.

Thí dụ 2. Tiếp tục thí dụ 1, tìm xác suất mắc sai lầm loại 2 và lực kiểm định nếu trọng lượng xuất chuồng trung bình của đàn bò thực sự là 395 kg.

Giải. Vì giả thuyết đối là $H_1 : \mu > 380$

nên với $\alpha = 0,01;\ \mu_0 = 380; \mu_1 = 395$, theo công thức (8.10) ta có:

$$\beta = P\left[U < u_{0,01} - \frac{|380 - 395|}{\frac{35,2}{\sqrt{50}}}\right] = P[U < 2,33 - 3,01]$$

$$= P[U < -0,68] = P[U > 0,68] = 0,2483$$

Vậy $1 - \beta = 1 - 0,2483 = 0,7517$.

3. Kiểm định với α và β cho trước

Từ công thức (8.10) và (8.11) ta thấy rằng nếu đã cố định trước xác suất mắc sai lầm loại 1 và với giá trị xác định của μ_1 thì xác suất mắc sai lầm loại 2 sẽ phụ thuộc vào kích thước mẫu n. Kích thước mẫu càng lớn thì càng có nhiều thông tin hơn về μ và giá trị của β sẽ càng nhỏ hơn. Ngược lại, nếu giá trị μ_1 đã xác định trước thì có thể xác định kích thước mẫu n sao cho đảm bảo xác suất mắc sai lầm loại 1 và loại 2 tương ứng là α và β cho trước.

Giả sử ta phải kiểm định cặp giả thuyết $H_0 : \mu = \mu_0$ và $H_1 : \mu > \mu_0$. Hơn

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

nữa giả sử ta muốn xác suất mắc sai lầm loại 1 là α và xác suất mắc sai lầm loại 2 không vượt quá giá trị β với giá trị thực của μ sai lệch so với μ_0 không vượt quá Δ cho trước.

Từ hình (8.2) ta có:

$$u_\alpha + u_\beta = u_\alpha - u_{1-\beta} = \frac{(\overline{X}_c - \mu_0)}{Se(\overline{X})} - \frac{(\overline{X}_c - \mu_1)}{Se(\overline{X})}$$

$$= \frac{(\mu_1 - \mu_0)}{Se(\overline{X})} = \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}$$

suy ra: $$n = \frac{\sigma^2 (u_\alpha + u_\beta)^2}{(\mu_1 - \mu_0)^2}$$

Từ đó nếu đòi hỏi $\mu_1 - \mu_0 \le \Delta$ thì ta có:

$$n \ge \left[\frac{\sigma^2 (u_\alpha + u_\beta)^2}{\Delta^2} \right] \qquad (8.12)$$

Nếu miền bác bỏ là hai phía thì kích thước mẫu tối thiểu được xác định bằng công thức:

$$n \ge \left[\frac{\sigma^2 (u_{\alpha/2} + u_\beta)^2}{\Delta^2} \right] \qquad (8.13)$$

Thí dụ 3. Trở lại thí dụ 1 và với xác suất mắc sai lầm loại 1 cho trước là 0,01, nếu xác suất mắc sai lầm loại 2 không vượt quá 0,05 thì phải đem cân thử tối thiểu là bao nhiêu con bò nếu quả thực trọng lượng trung bình của bò trước khi xuất chuồng là không quá 390 kg.

Giải. theo công thức (8.12) ta có:

$$\alpha = 0,01 \rightarrow u_\alpha = u_{0,01} = 2,33$$

$$\beta = 0,05 \rightarrow u_\beta = u_{0,05} = 1,645$$

$$\mu_1 - \mu_0 = 390 - 380 = 10$$

Từ đó: $$n \ge \left[\frac{(35,2)^2 (2,33 + 1,645)^2}{10^2} \right] = 195,78$$

Vậy $n = 196$ con.

Vậy để tiếp tục kiểm định, trước hết cần đem cân ngẫu nhiên 196 con để thu được trung bình mẫu $\overline{x}$ cho quá trình kiểm định tiếp theo.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4. Giá trị xác suất (P-Value) của kiểm định

Thủ tục kiểm định được trình bày ở trên có tính chất truyền thống và thường được gọi là kiểm định theo cách tiếp cận cổ điển, theo đó ta xác định các bộ phận của một giả thuyết thống kê theo các sai lầm loại 1 và loại 2 tương ứng với xác suất α và β. Trong những năm gần đây nhiều nhà nghiên cứu thường sử dụng một cách tiếp cận khác. Thay vì kiểm định giả thuyết với một giá trị α định trước thì họ cho rằng ta nên định rõ các giả thuyết cơ sở H_0 và giả thuyết đối H_1, sau đó thu thập các số liệu mẫu và xác định mức độ khẳng định việc bác bỏ giả thuyết H_0. Mức độ khẳng định này thường được gọi là giá trị xác suất (P- value) của kiểm định.

Ta sẽ minh họa việc tính giá trị xác suất qua thí dụ sau.

Thí dụ 4: Trở lại thí dụ 1.

a) Thay vì định trước giá trị α, hãy xác định giá trị P của kiểm định.

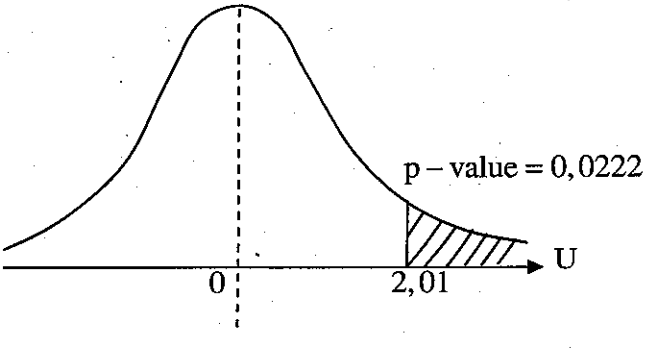
b) Giá trị P sẽ thay đổi như thế nào nếu trung bình mẫu tìm được không phải bằng 390 mà là 397.

Giải: a) Cặp giả thuyết thống kê có dạng $H_0 : \mu = 380; H_1 : \mu > 380$.

Từ mẫu điều tra trên 50 con bò ta tìm được giá trị quan sát của tiêu chuẩn kiểm định là:

$$U_{qs} = \frac{(\overline{X} - 380)\sqrt{50}}{35,2} = \frac{(390 - 380)\sqrt{50}}{35,2} = 2,01$$

Giá trị P của kiểm định (tức là mức độ khẳng định việc bác bỏ H_1) là xác suất để giá trị quan sát của $\overline{X}$ lớn hơn 390 nếu giả thuyết H_0 là đúng. Giá trị này có thể tính bằng cách sử dụng giá trị quan sát 2,01 của tiêu chuẩn kiểm định và tìm xác suất để U lớn hơn 2,01. Nó bằng 0,0222. Giá trị này được minh họa trên đồ thị như sau (Hình 8.3).



***Hình 8.3.* P-value**

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b) Với $\bar{x}$ ta tìm được:

$$U_{qs} = \frac{(397-380)\sqrt{50}}{35,2} = 3,415$$

Vì giá trị gần nhất của bảng giá trị tới hạn U là 3,5 nên giá trị P có thể lấy gần bằng 0,0002. P-value $\approx 0,0002$.

Như đã thấy qua thí dụ 4, giá trị P là xác suất nhỏ nhất để kết quả quan sát được qua mẫu hiện tại mâu thuẫn nhiều hơn với H_0 so với kết quả quan sát mẫu trước đó. Nói cách khác nó khẳng định độ tin cậy được của H_0. Giá trị xác suất này càng nhỏ thì mức độ khẳng định của mẫu về việc bác bỏ H_0 càng rõ rệt hơn hay H_0 càng kém tin cậy hơn. Chẳng hạn kiểm định với giá trị P bằng 0,01 cho thấy mức độ khẳng định để bác bỏ H_0 rõ ràng hơn là kiểm định với giá trị P bằng 0,2.

Trên đây là giá trị P trong kiểm định bên phải. Nếu tiến hành kiểm định bên trái với cặp giả thuyết $H_0 : \mu = 380; H_1 : \mu < 380$ ta cũng thu được kết quả tương tự.

Từ đó ta có công thức tính giá trị P (P-value) cho kiểm định giả thuyết thống kê như sau:

Nếu $H_1 : \mu > \mu_0$ thì

$$\text{P-value} = P(U > U_{qs}) \qquad (8.14)$$

Nếu $H_1 : \mu < \mu_0$ thì

$$\text{P-value} = P(U < U_{qs}) \qquad (8.15)$$

Nếu $H_1 : \mu \neq \mu_0$ thì

$$\text{P-value} = 2P(U > |U_{qs}|) \qquad (8.16)$$

Trong thực tế việc kiểm định theo giá trị P (P-value) thường được tiến hành theo nguyên tắc sau:

- Nếu P-value > 0,1 thì thường người ta thừa nhận H_0.
- Nếu 0,05 < P-value < 0,1 thì cần cân nhắc cẩn thận trước khi bác bỏ H_0.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Nếu 0,01 < P-value < 0,05 thì nghiêng về hướng bác bỏ H_0 nhiều hơn.

- Nếu 0,001 < P-value < 0,01 thì có thể ít băn khoăn trước khi bác bỏ H_0.

- Nếu P-value < 0,001 thì có thể hoàn toàn yên tâm khi bác bỏ H_0.

Mặt khác nếu quy định trước mức ý nghĩa α thì có thể dùng P-value để kết luận theo α. Lúc đó nguyên tắc kiểm định như sau:

- Nếu P-value $< \alpha$ thì bác bỏ H_0, thừa nhận H_1.

- Nếu P-value $> \alpha$ thì chưa có cơ sở bác bỏ H_0.

Theo cách kiểm định này thì việc sử dụng P-value lại chính là kiểm định theo cách tiếp cận truyền thống.

5. Mối liên hệ giữa miền bác bỏ và khoảng tin cậy

Ta có thể chứng tỏ được rằng việc tìm miền bác bỏ với mức ý nghĩa α cũng chính là tìm khoảng tin cậy tương ứng với độ tin cậy $1-\alpha$. Như ở phần trên khi kiểm định cặp giả thuyết $H_0 : \mu = \mu_0$ và $H_1 : \mu \neq \mu_0$ ta đã đòi hỏi là xác suất để tiêu chuẩn kiểm định

$$U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}$$

thuộc vào miền bác bỏ hai phía bằng α, do đó xác suất để U thuộc miền thừa nhận giả thuyết $(-u_{\alpha/2}; u_{\alpha/2})$ bằng $1-\alpha$, tức là thỏa mãn điều kiện

$$P\left[-u_{\alpha/2} < \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} < u_{\alpha/2}\right] = 1 - \alpha$$

hay

$$P\left[\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha/2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha/2}\right] = 1 - \alpha$$

Ta lại thu được công thức ước lượng khoảng tin cậy đối xứng giá trị μ khi đã biết σ.

Chú ý rằng hai thủ tục trên, mặc dù đi đến cùng một kết quả song cách giải thích những kết quả đó lại khác nhau. Miền bác bỏ xác định các giới hạn (giá trị tới hạn) trong đó chứa đựng $(1-\alpha)\%$ số các giả thuyết H_0 được chấp nhận khi lặp lại nhiều lần phép thử. Còn khoảng tin cậy thì xác định

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

các giới hạn trong đó (1-α)% phép thử sẽ chứa giá trị thực của tham số cần ước lượng.

3.2. Kiểm định giả thuyết về kỳ vọng toán của biến ngẫu nhiên phân phối chuẩn khi chưa biết phương sai

Lúc đó tiêu chuẩn kiểm định là thống kê

$$G = T = \frac{(\overline{X} - \mu_0)}{Se(\overline{X})} = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S} \qquad (8.17)$$

Nếu giả thuyết H_0 đúng thì ta có:

$$T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S} = \frac{(\overline{X} - \mu)\sqrt{n}}{S}$$

Và từ mục 7 Chương VI ta đã biết T phân phối T(n - 1). Từ đó, tuỳ thuộc vào dạng của giả thuyết đối H_1, miền bác bỏ "tốt nhất" được xây dựng theo các trường hợp sau:

a) $H_0 : \mu = \mu_0; H_1 : \mu > \mu_0$

Lúc đó với mức ý nghĩa α cho trước có thể tìm được giá trị tới hạn Student $t_\alpha^{(n-1)}$ sao cho

$$P(G \in W_\alpha / H_0) = P(T > t_\alpha^{(n-1)}) = \alpha$$

Ta thu được miền bác bỏ bên phải W_α sau:

$$W_\alpha = \left\{ T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S}; T > t_\alpha^{(n-1)} \right\} \qquad (8.18)$$

b) $H_0 : \mu = \mu_0; H_1 : \mu < \mu_0$

Lúc đó với mức ý nghĩa α cho trước có thể tìm được giá trị tới hạn Student $t_{1-\alpha}^{(n-1)}$ sao cho

$$P(G \in W_\alpha / H_0) = P(T < t_{1-\alpha}^{(n-1)}) = P(T < -t_\alpha^{(n-1)}) = \alpha$$

Ta thu được miền bác bỏ bên trái W_α sau:

$$W_\alpha = \left\{ T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S}; T < -t_\alpha^{(n-1)} \right\} \qquad (8.19)$$

c) $H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lúc đó với mức ý nghĩa α cho trước có thể tìm được hai giá trị tới hạn Student là $t_{\alpha/2}^{(n-1)}$ và $t_{1-\alpha/2}^{(n-1)}$ sao cho

$$P(G \in W_\alpha / H_0) = P(T < t_{1-\alpha/2}^{(n-1)}) + P(T > t_{\alpha/2}^{(n-1)})$$
$$= P(T < -t_{\alpha/2}^{(n-1)}) + P(T > t_{\alpha/2}^{(n-1)})$$
$$= P(|T| > t_{\alpha/2}^{(n-1)}) = \alpha$$

Ta thu được miền bác bỏ hai phía W_α sau:

$$W_\alpha = \left\{ T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S}; |T| > t_{\alpha/2}^{(n-1)} \right\} \qquad (8.20)$$

Với một mẫu cụ thể tính được $\overline{x}$, s; từ đó tính được giá trị quan sát T của tiêu chuẩn kiểm định

$$T_{qs} = \frac{(\overline{x} - \mu_0)\sqrt{n}}{s}$$

So sánh T_{qs} với miền bác bỏ W_α và kết luận:

- Nếu $T_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1;

- Nếu $T_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Thí dụ 5: Trọng lượng đóng bao của các bao gạo trong kho là biến ngẫu nhiên phân phối chuẩn với trọng lượng trung bình theo quy định là 50kg: Nghi ngờ gạo bị đóng thiếu, người ta đem cân ngẫu nhiên 25 bao và thu được các số liệu sau (bảng 8.2).

Trọng lượng bao (kg)	Số bao tương ứng
48,0 - 48,5	2
48,5 - 49,0	5
49,0 - 49,5	10
49,5 - 50,0	6
50,0 - 50,5	2
	n = 25

Với ý nghĩa $\alpha = 0,01$ hãy kết luận về điều nghi ngờ nói trên.

Giải. Gọi X là trọng lượng đóng bao. Theo giả thiết X phân phối chuẩn. Vậy trọng lượng đóng bao trung bình chính là tham số μ. Đây là bài toán

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

kiểm định giả thuyết về tham số μ của phân phối chuẩn $N(\mu,\sigma^2)$ khi chưa biết σ^2.

Cặp giả thuyết thống kê: $H_0 : \mu = 50; H_1 : \mu < 50$

Vậy theo công thức (8.19) ta có tiêu chuẩn kiểm định là:

$$T = \frac{(\overline{X} - 50)\sqrt{25}}{S}$$

trong đó $\overline{X}$ và S là trung bình và độ lệch chuẩn của mẫu ngẫu nhiên kích thước n = 25.

Cũng từ (8.19)

$$-t_{\alpha}^{(n-1)} = -t_{0,01}^{(24)} = -2,402$$

Vậy miền bác bỏ $(-\infty; -2,402)$. Từ mẫu cụ thể ta lập bảng tính $\overline{x}$ và s.

x_i	n_i	$n_i x_i$	$n_i x_i^2$
48,25	2	96,5	4656,125
48,75	5	243,75	11882,8125
49,25	10	492,5	24255,625
49,75	6	298,5	14850,375
50,25	2	100,5	5050,125
	n = 25	$\sum = 1231,75$	$\sum = 60695,062$

Từ đó: $\overline{x} = \frac{1231,75}{25} = 49,27$

$$ms = \frac{60695,062}{25} = (49,27)^2 = 0,27$$

$$s = \sqrt{\frac{25}{24}.0,27} = 0,53$$

Giá trị quan sát của tiêu chuẩn kiểm định

$$T_{qs} = \frac{(49,27 - 50)\sqrt{25}}{0,53} = -6,887$$

Vậy $T_{qs} \in W_{\alpha}$: Bác bỏ H_0, thừa nhận H_1, tức là qua mẫu cụ thể này

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

thừa nhận gạo bị đóng thiếu với mức ý nghĩa 0,01.

Việc kiểm định bằng cách sử dụng tiêu chuẩn (8.17) thường được gọi là kiểm định T (T-test).

Trong trường hợp này xác suất mắc sai lầm loại 2 được tìm bằng các công thức sau:

- Nếu là kiểm định một phía (bên phải hoặc bên trái) thì

$$\beta = P\left[T < t_{\alpha}^{(n-1)} - \frac{|\mu_0 - \mu_1|}{Se(\bar{x})} \right] \tag{8.21}$$

- Nếu là kiểm định hai phía thì

$$\beta = P\left[T < t_{\alpha/2}^{(n-1)} - \frac{|\mu_0 - \mu_1|}{Se(\bar{x})} \right] \tag{8.22}$$

trong đó $Se(\bar{x}) = \frac{s}{\sqrt{n}}$.

Và kích thước mẫu tối thiểu n cần điều tra để xác suất mắc sai lầm loại 1 và loại 2 không quá α và β tương ứng và giá trị thực μ_1 không sai lệch so với μ_0 không quá giá trị Δ được tính bằng các công thức sau:

- Nếu là kiểm định một phía

$$n \geq \left[\frac{s^2}{\Delta^2} (t_{\alpha}^{(m-1)} + t_{\beta}^{(m-1)})^2 \right] \tag{8.23}$$

- Nếu là kiểm định hai phía

$$n \geq \left[\frac{s^2}{\Delta^2} (t_{\alpha/2}^{(m-1)} + t_{\beta}^{(m-1)})^2 \right] \tag{8.24}$$

Trong đó s^2 là phương sai của mẫu sơ bộ kích thước $m \geq 2$ và $t_{\alpha}^{(m-1)}, t_{\alpha/2}^{(m-1)}$ và $t_{\beta}^{(m-1)}$ là các giá trị tới hạn tương ứng mức $\alpha, \frac{\alpha}{2}$ và β và số bậc tự do là m - 1.

Việc tìm các giá trị P để kiểm định theo cách tiếp cận P-value cũng tiến hành tương tự như đã làm ở mục trước.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 6: Tiếp tục thí dụ 5.

a) Tìm xác suất mắc sai lầm loại 2 nếu trọng lượng đóng bao trung bình thực sự là 49,5 kg.

b) Tìm giá trị P.

Giải: a) Vì đây là kiểm định bên trái nên theo công thức (8.21) ta có:

$$\beta = P\left[T < t_{\alpha}^{(n-1)} - \frac{|\mu_0 - \mu_1|}{Se(\bar{x})}\right]$$

$$Se(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{0,53}{\sqrt{25}} = 0,106$$

với $\alpha = 0,01 \Rightarrow t_{\alpha}^{(n-1)} = t_{0,01}^{(24)} = 2,492$.

$$\text{Vậy } \beta = P\left[T < 2,492 - \frac{|50 - 49,5|}{0,106}\right] = P[T < -2,675]$$

$$= P[T > 2,675] \approx 0,005$$

b) Từ thí dụ 5 ta tìm được.

$$T_{qs} = -6,887$$

Vì đây là kiểm định bên trái nên

$$P - \text{value} = P(T < T_{qs}) = P(T < -6,887) = P(T > 6,887)$$

Với số bậc tự do là 24, giá trị tối đa của T là 3,467 do đó P-value < 0,001.

Ví dụ A: Bằng Excel việc kiểm định giả thuyết

$H_0 : \mu_0 = 1650$; H_1: $\mu \neq 1650$ cho kết quả sau:

1,650.000	hypothesized value
1,649.750	mean Data
47.513	std. dev.
4.751	std. error
100	n
99	df
-0.05	t
.9581	p-value (two-tailed)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

8.1. Một mẫu kích thước n = 25 được rút ra từ tổng thể phân phối chuẩn với phương sai là 64. Nếu trung bình mẫu tìm được là 55,4 thì với mức ý nghĩa 0,05 hãy tiến hành một thủ tục kiểm định giả thuyết là trung bình tổng thể bằng 52.

ĐS: Uqs = 2,125.

8.2. Thông qua một mẫu gồm 100 gia đình người ta thu được kết quả là chi tiêu trung bình hàng tháng của các gia đình đó là 2,455 triệu đồng với độ lệch chuẩn là 0,3 triệu. Với mức ý nghĩa 0,05 có thể cho rằng chi tiêu trung bình hàng tháng của các gia đình ít hơn 2,5 triệu hay không. Giả thiết mức chi tiêu phân phối chuẩn.

ĐS: Tqs = -1,5. Không bác bỏ giả thuyết.

8.3. Giả sử cần kiểm định cặp giả thuyết

H_0: $\mu = 400$; H_1: $\mu \neq 400$;

trong đó μ là trung bình của tổng thể phân phối chuẩn với độ lệch chuẩn là 45. Lấy một mẫu kích thước 25 tìm được $\bar{x} = 395$.

a. Với mức ý nghĩa 0,05 hãy kiểm định cặp giả thuyết trên.

b. Tìm xác suất mắc sai lầm loại hai β nếu giá trị thực của μ là 390.

8.4. Trọng lượng sản phẩm (X) do nhà máy sản xuất ra là biến ngẫu nhiên phân phối chuẩn với độ lệch chuẩn $\sigma_x = 2$ kg và trọng lượng trung bình là 20kg. Nghi ngờ máy hoạt động không bình thường làm thay đổi trọng lượng trung bình của sản phẩm người ta cân thử 100 sản phẩm và thu được kết quả sau:

Trọng lượng sản phẩm	19	20	21	22	23
Số sản phẩm tương ứng	10	60	20	5	5

Với mức ý nghĩa $\alpha = 0,05$ hãy kết luận về điều nghi ngờ nói trên.

ĐS: Uqs = 1,75. Vậy Uqs $\notin$ Wα. Chưa có cơ sở để cho rằng máy móc hoạt động không bình thường.

8.5. Mỳ chính được đóng gói 453 gam một gói trên máy tự động. Có thể coi trọng lượng các gói mỳ chính tuân theo quy luật phân phối chuẩn với độ

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

lệch chuẩn 36 gam. Kiểm tra ngẫu nhiên 81 gói *thấy* trọng lượng trung bình là 448 gam. Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận trọng lượng các gói mỳ chính có xu hướng bị đóng thiếu không?

ĐS: Uqs = -1,25. Uqs $\notin$ Wα - chưa có cơ sở để kết luận như vậy.

8.6. Trong điều kiện chăn nuôi bình thường, lượng sữa trung bình của một con bò là 14 kg một ngày. Nghi ngờ điều kiện chăn nuôi bò kém đi làm cho lượng sữa giảm xuống. Người ta điều tra ngẫu nhiên 25 con bò và tính được lượng sữa trung bình của một con trong một ngày là 12,5 kg và độ lệch chuẩn mẫu s = 2,5 kg. Với mức ý nghĩa $\alpha = 0,05$ hãy kết luận về điều nghi ngờ nói trên.

Giả thiết lượng sữa bò là biến ngẫu nhiên phân phối chuẩn.

ĐS: Miền bác bỏ: Wα = (-∞, -1,711); Tqs = – 3;

Tqs $\in$ Wα. Bác bỏ giả thiết Ho tức là lượng sữa bò có xu hướng giảm.

8.7. Định mức gia công một loại chi tiết là 44 chi tiết trong một ca. Do chất lượng nguyên liệu cung cấp giảm sút, có ý kiến đề nghị hạ định mức. Người ta theo dõi ngẫu nhiên 25 công nhân gia công loại chi tiết đó và tính được năng suất trung bình của họ là 41,6 chi tiết trong một ca sản xuất, độ phân tán s = 3,2 chi tiết. Với mức ý nghĩa $\alpha = 0,01$ hãy kết luận về đề nghị nói trên. Biết rằng năng suất lao động của công nhân là biến ngẫu nhiên phân phối chuẩn.

8.8. Năng suất trung bình của một giống lúa là 47 tạ/ha. Sau thời gian dài canh tác, người ta nghi ngờ giống lúa đó bị thoái hoá, năng suất giảm. Dựa vào mẫu kích thước bằng 25 tìm được năng suất trung bình từ mẫu là 45,5 tạ/ha và độ lệch chuẩn mẫu là 4 tạ/ha. Hãy kết luận về điều nghi ngờ nói trên với mức ý nghĩa $\alpha = 0,05$. Cho biết năng suất của giống lúa đó là biến ngẫu nhiên tuân theo quy luật chuẩn.

ĐS: Miền bác bỏ: Wα = (- ∞, -1,711); Tqs = -1,875.

Tqs $\in$ Wα - Vậy giống lúa đó bị thoái hoá.

8.9. Mức hao phí xăng (X) cho một loại xe ôtô chạy trên đoạn đường AB là biến ngẫu nhiên phân phối chuẩn có kỳ vọng toán là 50 lít. Do đường được tu sửa lại, người ta cho rằng mức hao phí xăng trung bình đã giảm xuống. Quan sát 30 chuyến xe chạy trên đường AB ta thu được bảng số liệu sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Mức xăng hao phí (lít)	Số chuyến xe
48,5 - 49,0	5
49,0 - 49,5	10
49,5 - 50,0	10
50,0 - 50,5	3
50,5 - 51,0	2

a. Với mức ý nghĩa $\alpha = 0,05$ hãy kết luận về ý kiến nêu trên.

b. Tìm xác suất mắc sai lầm loại hai nếu mức xăng hao phí trung bình thực sự là 48 lít.

8.10. Định mức thời gian hoàn thành một sản phẩm là 14 phút. Có cần thay đổi định mức không, nếu theo dõi thời gian hoàn thành sản phẩm ở 25 công nhân, ta thu được bảng số liệu sau:

Thời gian sản xuất một sản phẩm (phút)	Số công nhân tương ứng
10 - 12	2
12 - 14	6
14 - 16	10
16 - 18	4
18 - 20	3
	n = 25

Yêu cầu kết luận với mức ý nghĩa $\alpha = 0,05$ biết rằng thời gian hoàn thành một sản phẩm là biến ngẫu nhiên phân phối chuẩn.

ĐS: Miền bác bỏ $W\alpha = [(-\infty, -2,064); (2,064, +\infty)]$;

$Tqs = 2,23$. $Tqs \in W\alpha$. Kết luận: cần phải thay đổi định mức.

8.11. Bột mỳ được đóng bao bằng máy tự động có trọng lượng đóng bao theo quy định là 16 kg và độ lệch chuẩn là 1,2 kg. Lấy ngẫu nhiên 25 bao bột để kiểm tra tìm được trọng lượng đóng gói trung bình của chúng là 16,5 kg.

a. Với mức ý nghĩa 0,05 có cần dừng máy để điều chỉnh hay không.

b. Tìm xác suất mắc sai lầm loại 2 nếu giá trị thực của trọng lượng đóng gói trung bình là 15,5 và 16,6.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

c. Nếu muốn kiểm định vấn đề trên với xác suất mắc sai lầm loại 1 là 0,05 và xác suất mắc sai lầm loại 2 là 0,02 trong điều kiện trọng lượng đóng gói thực của bột mỳ sai lệch so với trọng lượng quy định không quá 1 kg thì phải lấy ra bao nhiêu bao bột mỳ để kiểm tra.

Giả thiết trọng lượng đóng bao phân phối chuẩn

ĐS: a. Cần dừng máy.

b. 0,4522; 0,2946. c. 24.

8.12. Một xí nghiệp sản xuất lốp xe hơi dự định dùng một loại sợi tổng hợp mới để tăng độ bền của lốp xe. Sản xuất thử 16 lốp xe với loại sợi mới này và đem thử nghiệm thu được kết quả sau về độ bền $\bar{x}$ = 34625 km ; s = 3200 km.

Giả thiết độ bền của lốp xe phân phối chuẩn.

a. Với mức ý nghĩa 0,05 hãy kiểm định xem độ bền trung bình của lốp xe có khác trước hay không, biết rằng trước đây độ bền trung bình của lốp xe là 32000 km.

b. Hãy kết luật về vấn đề trên bằng cách dùng P - value.

8.13. Người ta đã thực hiện một cải tiến kỹ thuật trong bộ chế hoà khí của xe ôtô với hy vọng sẽ tiết kiệm được xăng hơn. Dùng thử 12 lần thu được kết quả sau về số km chạy được cho 1 lít xăng.

20,6	20,5	20,8
20,8	20,7	20,6
21,0	20,6	20,5
20,4	20,3	20,7

Nếu trước khi cải tiến một lít xăng trung bình chạy được 20,2 km thì có thể kết luận rằng cải tiến trên đã mang lại hiệu quả đáng kể hay không. Hãy kết luận với mức ý nghĩa 0,05.

Giả thiết số km chạy được cho 1 lít xăng phân phối chuẩn.

ĐS: Cải tiến trên thực sự có hiệu quả

8.14. Năm trước tiền lương trung bình của các cử nhân quản trị kinh doanh làm việc trong các công ty liên doanh với nước ngoài là 210 USD một tháng. Năm nay điều tra lương tháng của 25 cử nhân kinh tế đang làm việc cho các công ty đó tìm được $\bar{x}$ = 218 USD; $\Sigma(x_i - \bar{x})^2 = 2400$.

Nếu giả thiết tiền lương là biến ngẫu nhiên phân phối chuẩn thì hãy tìm

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

P-value để kiểm định ý kiến cho rằng các nhân viên đó đã được hưởng mức tiền lương cao hơn năm trước.

ĐS: P-value < 0,005

8.15. Thời gian trước số tiền gửi tiết kiệm bằng ngoại tệ trung bình của mỗi khách hàng là 1000 USD. Để đánh giá xem hiện nay xu hướng này còn giữ nguyên hay không người ta kiểm tra ngẫu nhiên 64 số tiết kiệm và tìm được số tiền gửi trung bình là 990 USD, độ lệch chuẩn là 100 USD.

a. Với mức ý nghĩa 0,05 hãy kiểm định xem số tiền gửi trung bình có thay đổi hay không.

b. Hãy kiểm định vấn đề trên bằng cách sử dụng P-value.

c. Tìm xác suất mắc sai lầm loại 2 nếu số tiền tiết kiệm trung bình của mỗi khách hàng thực sự bằng 1050 USD.

d. Tìm lực kiểm định nếu $\mu = 980$ USD.

e. Nếu đòi hỏi xác suất mắc sai lầm loại 1 và loại hai đều bằng 0,05 và sai lệch giữa giá trị trung bình thực sự với giá trị giả thuyết không quá 30 USD thì phải điều tra một mẫu có kích thước tối thiểu là bao nhiêu.

ĐS: a. Tqs = - 0,8

b. P-value = 0,21

c. $\beta = 0,0207$

d. 0,3594

e. 144

3.3. Kiểm định giả thuyết về hai kỳ vọng toán của hai biến ngẫu nhiên phân phối chuẩn

Giả sử có hai tổng thể nghiên cứu trong đó các biến ngẫu nhiên X_1 và X_2 cùng phân phối chuẩn với các kỳ vọng toán là μ_1, μ_2 và các phương sai là σ_1^2 và σ_2^2. Nếu μ_1 và μ_2 chưa biết song có cơ sở để giả thiết rằng giá trị của chúng bằng nhau người ta đưa ra giả thuyết thống kê.

$$H_0 : \mu_1 = \mu_2$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Để kiểm định giả thuyết trên ta xét một số trường hợp sau:

1. Nếu đã biết các phương sai σ_1^2 *và* σ_2^2 *của các biến ngẫu nhiên gốc trong tổng thể và từ hai tổng thể trên có thể rút ra hai mẫu độc lập kích thước* n_1 *và* n_2 *:*

$$W_1 = (X_{11}, X_{12}, ..., X_{1n1})$$

$$W_2 = (X_{21}, X_{22}, ..., X_{2n2})$$

Lúc đó tiêu chuẩn kiểm định được chọn là thống kê

$$G = U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (8.25)$$

Từ mục 7 Chương VI ta đã biết thống kê U phân phối N(0,1). Nếu giả thuyết H_0 đúng thì thống kê U có dạng

$$U = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (8.26)$$

và cũng phân phối N(0,1). Vì vậy, với mức ý nghĩa bằng α cho trước và tuỳ thuộc vào dạng của giả thuyết đối H_1, với phương pháp xây dựng giống như đã làm ở các phần trên ta thu được các miền bác bỏ W_α tương ứng sau:

a) $H_0 : \mu_1 = \mu_2; H_1 : \mu_1 > \mu_2$

Miền bác bỏ bên phải là

$$W_\alpha = \left\{ U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; U > u_\alpha \right\} \qquad (8.27)$$

b) $H_0 : \mu_1 = \mu_2; H_1 : \mu_1 < \mu_2$

Miền bác bỏ bên trái là

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$W_\alpha = \left\{ U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; U < -u_\alpha \right\} \quad (8.28)$$

c) $H_0 : \mu_1 = \mu_2; H_1 : \mu_1 \neq \mu_2$

Miền bác bỏ 2 phía là

$$W_\alpha = \left\{ U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; |U| > u_{\alpha/2} \right\} \quad (8.29)$$

Lập hai mẫu cụ thể từ X_1 và X_2 và tính được các trung bình mẫu cụ thể

$$\overline{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}; \overline{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$$

và tính được giá trị quan sát của tiêu chuẩn kiểm định

$$U_{qs} = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Xem xét U_{qs} có thuộc W_α không để kết luận.

Thí dụ 7: Tại một xí nghiệp người ta xây dựng hai phương án gia công cùng một loại chi tiết. Để đánh giá xem chi phí trung bình về nguyên liệu theo hai phương án ấy có khác nhau hay không người ta tiến hành sản xuất thử và thu được kết quả sau:

Phương án 1:	2,5	3,2	3,5	3,8	3,5	
Phương án 2:	2,0	2,7	2,5	2,9	2,3	2,6

Với mức ý nghĩa α=0,05, hãy kết luận về vấn đề trên biết rằng chi phí nguyên liệu theo cả hai phương án gia công đều là các biến ngẫu nhiên phân phối chuẩn với $\sigma_1^2 = \sigma_2^2 = 0,16$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Gọi X_1 và X_2 tương ứng là chi phí nguyên liệu theo hai phương án gia công trên. Theo giả thuyết X_1 và X_2 phân phối chuẩn. Vậy chi phí nguyên liệu trung bình theo các phương án đó là μ_1 và μ_2. Vậy đây là bài toán kiểm định $H_0 : \mu_1 = \mu_2$ với $H_1 : \mu_1 \neq \mu_2$ khi đã biết σ_1^2 và σ_2^2.

Theo (8.29) tiêu chuẩn kiểm định là

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{0,16}{5} + \frac{0,16}{6}}}$$

Do $\alpha = 0,05 \Rightarrow u_{\alpha/2} = u_{0,025} = 1,96$.

Vậy miền bác bỏ là $(-\infty; -1,96)$ và $(1,96; +\infty)$.

Từ mẫu cụ thể ta tính được

$$\overline{x}_1 = \frac{2,5 + 3,2 + 3,5 + 3,8 + 3,5}{5} = 3,3$$

$$\overline{x}_2 = \frac{2,0 + 2,7 + 2,5 + 2,9 + 2,3 + 2,6}{6} = 2,5$$

và giá trị quan sát của tiêu chuẩn kiểm định

$$U_{qs} = \frac{3,3 - 2,5}{\sqrt{\frac{0,16}{5} + \frac{0,16}{6}}} = 3,33$$

Vậy $U_{qs} \in W_\alpha$ bác bỏ H_0 thừa nhận H_1 tức là chi phí nguyên liệu theo hai phương án gia công trên thực sự khác nhau.

Việc tìm xác suất mắc sai lầm loại 2, kích thước mẫu n và giá trị P cũng được tiến hành như đã làm ở các mục trước. Chẳng hạn với α và β định trước và nếu muốn $|\mu_1 - \mu_2| \leq \Delta$ thì kích thước mẫu tối thiểu trong trường hợp kiểm định một phía là

$$n \geq \left[\frac{2\sigma^2 (u_\alpha + u_\beta)}{\Delta^2} \right] \qquad (8.30)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Còn trong trường hợp kiểm định hai phía là

$$n \geq \left[\frac{2\sigma^2 (u_{\alpha/2} + u_{\beta})^2}{\Delta^2} \right] \tag{8.31}$$

trong đó $n_1 = n_2 = n$ và $\sigma^2 = \frac{\sigma_1^2 + \sigma_2^2}{2}$.

Chú ý rằng phương pháp kiểm định vừa trình bày ở trên còn có thể áp dụng khi μ_1 và μ_2 khác nhau. Chẳng hạn cần kiểm định giả thuyết $H_0 : \mu_1 - \mu_2 = D_0$ với D_0 là một giá trị định trước thì tiêu chuẩn kiểm định sẽ là:

$$G = U = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{8.32}$$

Và thủ tục kiểm định được tiến hành hoàn toàn giống như đã trình bày ở trên.

Thí dụ: Giả sử trong thí dụ A ta đã biết phương sai của thu nhập của dân cư hai vùng 1 và 2 tương ứng là 2034 và 2240. Với mức ý nghĩa 0,05 hãy kiểm định sự bằng nhau của thu nhập trung bình của hai vùng đó.

Kiểm định bằng Excel cho kết quả sau:

z-Test: Two Sample for Means

	Variable 1	*Variable 2*
Mean	1649.75	1683.86
Known Variance	2034	2240
Observations	100	100
Hypothesized Mean Difference	0	
z	-5.217524061	
P(Z<=z) one-tail	9.06654E-08	
z Critical one-tail	1.644853627	
P(Z<=z) two-tail	1.81331E-07	
z Critical two-tail	1.959963985	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2. Nếu chưa biết các phương sai σ_1^2 *và* σ_2^2 *của các biến ngẫu nhiên gốc trong tổng thể song giả định rằng chúng bằng nhau* $(\sigma_1^2 = \sigma_2^2)$. Việc kiểm định $\sigma_1^2 = \sigma_2^2$ sẽ được xét ở mục 3.7. Mặt khác, ta vẫn giả định rằng có thể điều tra được 2 mẫu độc lập kích thước n_1 và n_2.

Lúc đó tiêu chuẩn kiểm định được chọn là

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{8.33}$$

Với $$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Ta biết rằng T phân phối Student với $n_1 + n_2 - 2$ bậc tự do. Với điều kiện giả thuyết H_0 là đúng thì tiêu chuẩn kiểm định trở thành

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{8.34}$$

và vẫn phân phối $T(n_1 + n_2 - 2)$. Do đó tuỳ thuộc vào giả thuyết đối H_1 ta có miền bác bỏ mức α như sau:

a) $H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 > \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} ; T > t_\alpha^{(n_1+n_2-2)} \right\} \tag{8.35}$$

b) $H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 < \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} ; T < -t_\alpha^{(n_1+n_2-2)} \right\} \tag{8.36}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

c) $H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 \neq \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} ; |T| > t_{\alpha/2}^{(n_1+n_2-2)} \right\} \qquad (8.37)$$

Qua hai mẫu cụ thể tính được $\overline{x}_1, \overline{x}_2, s_1^2, s_2^2$ và giá trị quan sát T_{qs} của tiêu chuẩn kiểm định, từ đó so sánh với W_α và kết luận.

Việc suy ra một cách tương ứng cho việc kiểm định giả thuyết $H_0 : \mu_1 - \mu_2 = D_0$ cũng tiến hành tương tự như ở mục trước.

Việc xác định β, P-value và kích thước mẫu n cũng tiến hành tương tự. Sự thay đổi duy nhất so với công thức (8.30) và (8.31) là tính giá trị của σ^2 theo công thức:

$$\sigma^2 = S_p^2$$

Thí dụ 8. Một nghiên cứu được thực hiện đối với 20 người ở một phường và 19 người ở một phường khác trong thành phố để xem thu nhập trung bình hàng năm (tính bằng triệu đồng) của dân cư hai phường đó có thực sự khác nhau hay không. Các số liệu mẫu thu được như sau:

$n_1 = 20$ $\qquad$ $n_2 = 19$

$\overline{x}_1 = 18,27$ $\qquad$ $\overline{x}_2 = 16,78$

$s_1^2 = 8,74$ $\qquad$ $s_2^2 = 6,58$

Vậy với mức ý nghĩa 0,05 có thể cho rằng thu nhập trung bình của dân cư ở hai phường đó khác nhau hay không? Giả thiết thu nhập hàng năm của dân cư hai phường cùng phân phối chuẩn với phương sai như nhau.

Giải. Gọi X_1 và X_2 tương ứng là thu nhập hàng năm của dân cư hai phường đó. Theo giả thiết X_1 và X_2 phân phối chuẩn với các phương sai $\sigma_1^2 = \sigma_2^2$. Do đó để kiểm định cặp giả thuyết

$$H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 \neq \mu_2$$

ta sử dụng công thức kiểm định (8.37)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}; |T| > t_{\alpha/2}^{(n_1+n_2-2)} \right\}$$

với $\alpha = 0,05 \Rightarrow t_{\alpha/2}^{(n_1+n_2-2)} = t_{0,025}^{(37)} \approx 2,021$

Vậy miền bác bỏ là $(-\infty; -2,021)$ và $(0,021; +\infty)$.

Từ mẫu cụ thể ta tính được

$$S_p = \sqrt{\frac{19.8,74 + 18.6,58}{20+19-2}} = 2,773$$

do đó

$$T_{qs} = \frac{18,27 - 16,78}{2,773\sqrt{\frac{1}{20}+\frac{1}{19}}} = 1,677$$

$T_{qs} \notin W_\alpha$ do đó với mức ý nghĩa 0,05 qua hai mẫu cụ thể đã cho, chưa có cơ sở để bác bỏ H_0, tức là có thể xem thu nhập trung bình hàng năm của dân cư hai phường đó là như nhau.

3. Nếu chưa biết các phương sai σ_1^2 và σ_2^2 của các tổng thể và không thể cho rằng chúng bằng nhau $(\sigma_1^2 \neq \sigma_2^2)$. Lúc đó nếu có thể điều tra được từ tổng thể hai mẫu độc lập kích thước n_1 và n_2 thì chọn lập thống kê

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \qquad (8.38)$$

Từ mục 7 chương VI ta đã biết T phân phối Student với số bậc tự do là

$$k = \frac{(n_1-1).(n_2-1)}{(n_2-1)C^2 + (1-C)^2(n_1-1)}$$

với

$$C = \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nếu giả thuyết H_0 là đúng thì tiêu chuẩn kiểm định trở thành

$$T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{8.39}$$

Và vẫn phân phối T(k). Vì vậy, miền bác bỏ mức α được xác định bằng các công thức sau:

a) $H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 > \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} ; T > t_\alpha^{(k)} \right\} \tag{8.40}$$

b) $H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 < \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} ; T < -t_\alpha^{(k)} \right\} \tag{8.41}$$

c) $H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 \neq \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} ; |T| > t_{\alpha/2}^{(k)} \right\} \tag{8.42}$$

Việc xác định xác suất mắc sai lầm loại hai, P-value và kích thước mẫu điều tra cũng tiến hành giống như ở mục trước. Thay đổi duy nhất là giá trị σ^2 trong công thức xác định n được tính bằng công thức xấp xỉ

$$\sigma^2 = \frac{m_1 S_1^2 + m_2 S_2^2}{m_1 + m_2}$$

Với S_1^2 và S_2^2 là phương sai của các mẫu sơ bộ kích thước m_1 và m_2.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 9. Để kiểm nghiệm hiệu quả của một loại thuốc tẩy giun cho lợn, người ta bắt ngẫu nhiên 14 con lợn từ một trại chăn nuôi và chia thành hai nhóm:

Nhóm I: Cho uống thuốc tẩy giun

Nhóm II: Không cho uống thuốc tẩy giun

Sau thời gian dùng thuốc, khi giết thịt, hai nhóm lợn trên cho kết quả sau về số giun có trong những con lợn thuộc hai nhóm trên.

Nhóm I:	18	43	28	50	16	32	13
Nhóm II:	40	54	26	63	21	37	39

a) Với mức ý nghĩa 0,05 hãy kết luận xem loại thuốc tẩy giun nói trên có thực sự hiệu quả hay không.

b) Tìm β nếu $|\mu_1 - \mu_2| = 10$

c) Tìm P-value

Giả thiết số lượng giun phân phối chuẩn.

Giải: a) Gọi X_1 và X_2 là số giun trong mỗi con lợn thuộc hai nhóm trên. Theo giả thiết X_1 và X_2 phân phối chuẩn với σ_1^2 và σ_2^2 chưa biết và không thể cho rằng chúng bằng nhau. Vậy số giun trung bình là μ_1 và μ_2. Để kiểm định cặp giả thuyết $H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 < \mu_2$ ta dùng công thức (8.41)

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} ; T < -t_\alpha^{(k)} \right\}$$

từ hai mẫu cụ thể tính được

$$n_1 = 7 \qquad n_2 = 7$$
$$\overline{x}_1 = 28,57 \qquad \overline{x}_2 = 40$$
$$s_1^2 = 198,62 \qquad s_2^2 = 215,33$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ta có:

$$C = \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = 0,4798$$

$$\Rightarrow k = \frac{(n_1 - 1).(n_2 - 1)}{(n_1 - 1)C^2 + (1 - C)^2(n_1 - 1)}$$

$$= \frac{6.6}{6.0,4798^2 + 0,5202^2.6} \approx 12$$

với $\alpha = 0,05 \Rightarrow t_{0,05}^{(12)} = 1,782$

Vậy miền bác bỏ là $(-\infty; -1,782)$

$$T_{qs} = \frac{28,57 - 40}{\sqrt{\frac{198,62}{7} + \frac{215,33}{7}}} = -1,49$$

$T_{qs} \notin W_\alpha$ vậy với mức ý nghĩa 0,05 chưa có cơ sở bác bỏ H_0 hay chưa thể nói loại thuốc tẩy giun được thử nghiệm là có hiệu quả.

b) Ta tìm β theo công thức

$$\beta = P\left[T < T_\alpha^{(k)} - \frac{|\mu_1 - \mu_2|}{Se(\overline{X}_1 - \overline{X}_2)}\right]$$

$$V(\overline{X}_1 - \overline{X}_2) \approx \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} = \frac{198,62}{7} + \frac{215,33}{7} = 59,14$$

$$\Rightarrow Se(\overline{X}_1 - \overline{X}_2) \approx 7,69$$

$$\beta = P\left[T < 1,728 - \frac{10}{7,69}\right] = P[T < +0,48)$$

$$= 1 - P[T > 0,48]$$

Với k = 12 giá trị nhỏ nhất của T là 1,356 nên $\beta < 0,9$

c) P-value $= P[T < T_{qs}] = P[T < -1,49] = P[T > 1,49]$

Với k = 12 giá trị 1,49 nằm trong khoảng hai giá trị 1,356 và 1,782. Vậy 0,05 < P-value < 0,1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trong các trường hợp kiểm định trên, ta luôn giả thiết biến ngẫu nhiên trong các tổng thể phân phối chuẩn. Trong thực tế nếu X_1 và X_2 không phân phối chuẩn (việc kiểm định phân phối chuẩn sẽ được đề cập ở chương 9) song nếu điều tra được 2 mẫu độc lập với kích thước $n_1 > 30$ và $n_2 > 30$ thì các thống kê

$$U = \frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

hoặc thống kê

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

vẫn phân phối xấp xỉ N(0,1) theo định lý giới hạn trung tâm. Do đó vẫn có thể áp dụng các thủ tục kiểm định nói trên với các biến ngẫu nhiên X_1 và X_2 phân phối theo các quy luật bất kỳ, chỉ cần thoả mãn hai điều kiện:

- Các mẫu điều tra là độc lập
- Kích thước của cả hai mẫu đều > 30

Thí dụ: Bằng phần mềm Excel, với các số liệu của bảng A, kiểm định $H_0 : \mu_1 = \mu_2$ với các phương sai tổng thể chưa biết song giả thiết bằng nhau cho ta kết quả sau:

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	1649.75	1683.86
Variance	2257.442	3518.929697
Observations	100	100
Pooled Variance	2888.186	
Hypothesized Mean Difference	0	
df	198	
t Stat	-4.48801	
P(T<=t) one-tail	6.09E-06	
t Critical one-tail	1.652586	
P(T<=t) two-tail	1.22E-05	
t Critical two-tail	1.972017	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Và nếu kiểm định $\mu_1 = \mu_2$ khi các phương sai tổng thể khác nhau cho kết quả sau:

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	1649.75	1683.86
Variance	2257.441919	3518.929697
Observations	100	100
Hypothesized Mean Difference	0	
df	189	
t Stat	-4.488013522	
P(T<=t) one-tail	6.2366E-06	
t Critical one-tail	1.652955802	
P(T<=t) two-tail	1.24732E-05	
t Critical two-tail	1.972595036	

Thí dụ 10: Người ta cân trẻ sơ sinh ở hai khu vực thành thị và nông thôn, thu được kết quả sau (Bảng 8.3)

Bảng 8.3

Khu vực	Số trẻ được cân	Trọng lượng trung bình	Phương sai
Nông thôn	$n_1 = 2500$	$\overline{x}_1 = 3,0$	$s_1^2 = 200$
Thành thị	$n_2 = 500$	$\overline{x}_2 = 3,1$	$s_2^2 = 5$

Với mức ý nghĩa 0,01 có thể coi trọng lượng trung bình của trẻ sơ sinh ở hai khu vực bằng nhau được không?

Giải: Gọi trọng lượng trẻ sơ sinh ở nông thôn và thành thị tương ứng là X_1 và X_2. Vậy trọng lượng trẻ sơ sinh trung bình chính là m_1 và m_2. Đây là bài toán kiểm định cặp giả thuyết $H_0 : m_1 = m_2 ; H_1 : m_1 \neq m_2$ khi chưa biết các phương sai σ_1^2 và σ_2^2. Do $n_1 > 30$ và $n_2 > 30$ nên theo công thức (8.42) tiêu chuẩn kiểm định có dạng:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{2500} + \frac{S_2^2}{500}}}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Do $\alpha = 0,01 \rightarrow \frac{\alpha}{2} = 0,005 \Rightarrow u_{0,005} = 2,576$ nên miền bác bỏ có dạng $(-\infty; -2,576)$ và $(2,576; +\infty)$.

Qua mẫu cụ thể tính được

$$U_{qs} = \frac{3,0 - 3,1}{\sqrt{\frac{200}{2500} + \frac{5}{500}}} = -0,33$$

Vậy $U_{qs} \notin W_\alpha$ chưa có cơ sở để bác bỏ H_0, như vậy có thể coi trọng lượng trẻ sơ sinh ở nông thôn và thành thị là như nhau.

Sau đây là kết quả giải bài toán bằng phần mềm Excel:

Hypothesis Test: Independent Groups (t-test, unequal variance)

x1	x2	
3	3.1	mean
14.14	2.24	std. dev.
2500	500	n
	2934	df
	-0.1000	difference (x1 - x2)
	0.3000	standard error of difference
	0	hypothesized difference
	-0.33	t
	.7389	p-value (two-tailed)

4. Trường hợp hai mẫu điều tra phụ thuộc theo từng cặp

Ở các phần trên khi kiểm định về hai kỳ vọng toán ta luôn giả thiết rằng có thể điều tra được hai mẫu độc lập từ các tổng thể nghiên cứu. Trong thực tế có nhiều trường hợp hai mẫu điều tra được rút ra từ cùng một tổng thể thì khả năng chúng phụ thuộc nhau sẽ rất cao. (Vấn đề này sẽ được kiểm định ở chương 9). Lúc đó các phương pháp kiểm định ở trên không thể áp dụng được vì các kết quả kiểm định không còn đáng tin cậy nữa. Ở phần này ta sẽ xét việc kiểm định khi hai mẫu điều tra có cùng kích thước n, trong đó các giá trị của mẫu phụ thuộc tương ứng với nhau theo từng cặp.

Giả sử có hai tổng thể nghiên cứu trong đó có các biến ngẫu nhiên X_1 và X_2 cùng phân phối chuẩn với các phương sai chưa biết. Với mức ý nghĩa α phải kiểm định giả thuyết thống kê

$$H_0 : \mu_1 = \mu_2 \text{ hay } H_0 : \mu_1 - \mu_2 = 0$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ hai tổng thể rút ra hai mẫu ngẫu nhiên kích thước n

$$W_1 = (X_{11}, X_{12}, ..., X_{1n})$$

$$W_2 = (X_{21}, X_{22}, ..., X_{2n})$$

Do các giá trị của 2 mẫu tương ứng với nhau theo từng cặp do đó ta thiết lập biến ngẫu nhiên D với các giá trị có thể có $D_1, D_2, ..., D_n$ là hiệu số của các cặp giá trị tương ứng của 2 mẫu, như vậy

$$D_i = X_{1i} - X_{2i} \quad (i = \overline{1,n}) \tag{8.43}$$

Lúc đó kỳ vọng toán của D là

$$\mu_D = E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2$$

và việc kiểm định giả thuyết $H_0 : \mu_1 - \mu_2 = 0$ có thể viết dưới dạng $H_0 : \mu_D = 0$. Bài toán đưa về dạng kiểm định giá trị của tham số μ như đã xét ở mục 2.1.

Từ các giá trị D_i tìm được ở (8.43) ta xác định trung bình mẫu và phương sai mẫu

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{8.44}$$

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n} (D_i - \overline{D})^2 \tag{8.45}$$

Từ đó chọn lập tiêu chuẩn kiểm định

$$G = T = \frac{\overline{D}\sqrt{n}}{S_D} \tag{8.46}$$

Ta biết rằng T phân phối Student với n - 1 bậc tự do, vì vậy với mức ý nghĩa α ta có các miền bác bỏ sau:

a) $H_0 : \mu_D = 0; H_1 : \mu_D > 0$

$$W_\alpha = \left\{ T = \frac{\overline{D}\sqrt{n}}{S_D}; T > t_\alpha^{(n-1)} \right\} \tag{8.47}$$

b) $H_0 : \mu_D = 0; H_1 : \mu_D < 0$

$$W_\alpha = \left\{ T = \frac{\overline{D}\sqrt{n}}{S_D}; T < -t_\alpha^{(n-1)} \right\} \tag{8.48}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

c) $H_0 : \mu_D = 0; H_1 : \mu_D \neq 0$

$$W_\alpha = \left\{ T = \frac{\overline{D}\sqrt{n}}{S_D}; |T| > t_{\alpha/2}^{(n-1)} \right\} \tag{8.49}$$

Việc tìm β, P-value và kích thước mẫu n cũng tiến hành giống như ở mục 2.1. Chẳng hạn kích thước mẫu n khi kiểm định một phía là:

$$n \geq \left[\frac{S_D^2}{\Delta^2} (t_\alpha^{(m-1)} + t_\beta^{(m-1)})^2 \right] \tag{8.50}$$

Và khi kiểm định 2 phía là:

$$n \geq \left[\frac{S_D^2}{\Delta^2} (t_{\alpha/2}^{(m-1)} + t_\beta^{(m-1)})^2 \right] \tag{8.51}$$

Trong đó S_D^2 là phương sai của mẫu sơ bộ kích thước $m \geq 2$, còn $t_\alpha^{(m-1)}$ và $t_\beta^{(m-1)}$ là các giá trị tới hạn Student tương ứng.

Thí dụ 11. Theo dõi doanh số bán của một công ty (tính bằng triệu đồng) trong 15 ngày đầu tháng 3 và 15 ngày đầu tháng 5 thu được kết quả sau:

Ngày	Tháng 3	Tháng 5	di
1	7,6	7,3	0,3
2	10,2	9,1	1,1
3	9,5	8,4	1,1
4	1,3	1,5	-0,2
5	3,0	2,7	0,3
6	6,3	5,0	1,3
7	5,3	4,9	0,4
8	6,2	5,3	0,9
9	2,2	2,0	0,2
10	4,8	4,2	0,6
11	11,3	11,0	0,3
12	12,1	11,0	1,1
13	6,9	6,1	0,8
14	7,6	6,7	0,9
15	8,4	7,5	0,9

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nếu giả thiết doanh số hằng ngày phân phối chuẩn thì với mức ý nghĩa $\alpha = 0,05$ có thể cho rằng doanh số bán trung bình hàng ngày trong tháng 5 có giảm sút so với tháng 3 hay không?

Giải. Gọi X_1 và X_2 tương ứng là doanh số bán hàng ngày trong tháng 3 và tháng 5. Theo giả thiết X_1 và X_2 phân phối chuẩn. Vậy doanh số trung bình là μ_1 và μ_2. Ta phải kiểm định cặp giả thuyết

$$H_O : \mu_1 - \mu_2 = 0; H_1 : \mu_1 - \mu_2 > 0;$$

hay $$H_O : \mu_D = 0; H_1 : \mu_D > 0;$$

Theo công thức (8.47) công thức kiểm định là

$$W_\alpha = \left\{ T = \frac{\overline{D}\sqrt{n}}{S_D}; T > t_\alpha^{(n-1)} \right\}$$

với $\alpha = 0,05 \Rightarrow t_\alpha^{(n-1)} = t_{0,05}^{(14)} = 1,761$

Vậy miền bác bỏ là $(1,761; +\infty)$

Từ các số liệu mẫu tìm được

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = 0,61$$

$$s_D^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum d_i\right)^2}{n} \right] = 0,156$$

$$s_D = 0,394$$

Từ đó $$T_{qs} = \frac{0,61.\sqrt{5}}{0,394} = 6$$

Vì $T_{qs} \in W_\alpha$ nên bác bỏ H_O, thừa nhận H_1 tức là với mức ý nghĩa 0,05 qua hai mẫu cụ thể trên, doanh số trung bình hàng ngày của tháng 5 thực sự giảm sút so với tháng 3.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Kết quả kiểm định bằng Excel như sau:

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	6.846666667	6.18
Variance	10.26552381	8.743142857
Observations	15	15
Pearson Correlation	0.993446285	
Hypothesized Mean Difference	0	
df	14	
t Stat	5.999143041	
P(T<=t) one-tail	1.62979E-05	
t Critical one-tail	1.761310115	
P(T<=t) two-tail	3.25958E-05	
t Critical two-tail	2.144786681	

Trên đây là một số kiểm định tham số đối với hai kì vọng toán của hai biến ngẫu nhiên phân phối chuẩn. Việc mở rộng kiểm định cho k kì vọng toán (k > 2) của k biến ngẫu nhiên phân phối chuẩn trong k tổng thể nghiên cứu sẽ được giải quyết bằng phương pháp phân tích phương sai (xem chương X). Còn nếu các dấu hiệu không phân phối chuẩn thì việc kiểm định có thể tiến hành bằng các phương pháp kiểm định phi tham số (xem mục 4 của chương IX).

Bài tập

8.16. Hai máy tự động dùng để cắt những thanh kim loại do cùng một kỹ thuật viên phụ trách và căn chỉnh. Từ mỗi máy lấy ra 36 thanh kim loại để kiểm tra và thu được kết quả sau:

Máy 1: $\bar{x}_1 = 12{,}5$ cm, $s_1 = 1{,}2$cm

Máy 2: $\bar{x}_2 = 12{,}2$cm, $s_2 = 1{,}4$cm

Vậy với mức ý nghĩa 0,01 có thể cho rằng chiều dài của các thanh kim loại do máy 1 cắt nói chung lớn hơn chiều dài của các thanh kim loại do máy 2 cắt hay không.

Biết chiều dài của các thanh kim loại do các máy đó cắt phân phối chuẩn.

ĐS: Uqs = 0,976

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.17. Hai lớp sinh viên cùng học môn thống kê và kết quả thi hết môn như sau:

Lớp A	Lớp B
$n_1 = 64$	$n_2 = 68$
$\bar{x}_1 = 73,2$	$\bar{x}_2 = 76,6$
$s_1 = 10,9$	$s_2 = 11,2$

a. Với mức ý nghĩa 0,05 có thể cho rằng kết quả thi trung bình của lớp B cao hơn lớp A được không.

b. Tìm xác suất mắc sai lầm 2 nếu quả thực sai lệch về điểm trung bình của hai lớp là 2 điểm.

ĐS: a. Tqs = -1,75

b. $\beta = 0,0367$

8.18. Để so sánh chiều cao trung bình của nam thanh niên ở thành thị và nông thôn người ta đo thử chiều cao của 10.000 thanh niên và thu được kết quả như sau:

Vùng	Số thanh niên được đo	Chiều cao trung bình	Độ lệch chuẩn
Nông thôn	8.000	165 (cm)	4 (cm)
Thành thị	2.000	168 (cm)	9 (cm)

Với mức ý nghĩa $\alpha = 0,05$, có thể coi chiều cao trung bình của nam thanh niên ở thành phố cao hơn ở nông thôn hay không? Giả thiết chiều cao của thanh niên là biến ngẫu nhiên phân phối chuẩn.

8.19. Người ta thí nghiệm hai phương pháp chăn nuôi gà khác nhau. Sau một tháng kết quả tăng trọng như sau:

Phương pháp	Số gà được theo dõi	Mức tăng trọng trung bình (kg)	Độ lệch chuẩn (kg)
I	100	1,1	0,2
II	150	1,2	0,3

Với mức ý nghĩa $\alpha = 0,05$, có thể kết luận phương pháp II hiệu quả hơn phương pháp I hay không? Giả thiết mức tăng trọng gà phân phối chuẩn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.20. Trồng cùng một giống lúa trên hai thửa ruộng như nhau và bón hai loại phân khác nhau. Đến ngày thu hoạch ta có kết quả như sau:

Thửa thứ nhất lấy mẫu 1000 bông lúa thấy số hạt trung bình là $\bar{x}_1 = 70$ hạt và độ lệch chuẩn mẫu $s_1 = 10$. Thửa thứ hai lấy mẫu 500 bông thấy số hạt trung bình mỗi bông là $\bar{x}_2 = 72$ hạt và $s_2 = 20$.

Hỏi sự khác nhau giữa $\bar{x}_1$ và $\bar{x}_2$ là ngẫu nhiên hay bản chất. Hãy kết luận với mức ý nghĩa $\alpha = 0,05$.

ĐS: $|Uqs| = |-2,108| > U_{0,025} = 1,96$

Vậy sự khác nhau giữa $\bar{x}_1$ và $\bar{x}_2$ là có ý nghĩa.

8.21. Nếp áp dụng biện pháp kỹ thuật thứ nhất thì điều tra ngẫu nhiên $n_1 = 100$ thửa ruộng trồng giống lúa A thu được năng suất trung bình $\bar{x}_1 = 100$ tạ/ha và $s_1 = 10$ tạ/ha.

Còn nếu áp dụng biện pháp kỹ thuật thứ hai thì điều tra ngẫu nhiên $n_2 = 50$ thửa, thu được năng suất trung bình $\bar{x}_2 = 95$ tạ/ha và $s_2 = 9$ tạ/ha.

Hãy kết luận với mức ý nghĩa $\alpha = 0,05$ xem nếu áp dụng biện pháp kỹ thuật thứ nhất thì năng suất giống lúa A cao hơn thực sự so với kết quả áp dụng biện pháp kỹ thuật thứ hai không? Giả thiết năng suất lúa tuân theo quy luật chuẩn.

8.22. Để so sánh sự phát triển trí tuệ của trẻ em ở hai vùng khác nhau người ta chọn ngẫu nhiên từ vùng thứ nhất 38 trẻ em và từ vùng thứ hai 40 trẻ em cùng lứa tuổi. Kiểm tra chỉ số thông minh và thu được kết quả sau:

Vùng thứ nhất	Vùng thứ hai
$\bar{x}_1 = 89,7$	$\bar{x}_2 = 94,5$
$s_1 = 12,2$	$s_2 = 13,05$

Với mức ý nghĩa 0,05 có sự khác biệt đáng kể về sự phát triển trí tuệ của trẻ em thuộc hai vùng nói trên hay không.

ĐS: Uqs = -1,68.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.23. Người ta đề xuất một phương pháp tổ chức lao động mới trong quy trình lắp ráp một chi tiết. Để đánh giá hiệu quả của phương pháp này một kỹ thuật viên đã lắp ráp 10 chi tiết theo phương pháp cũ và 10 chi tiết theo phương pháp mới. Thời gian lắp ráp các chi tiết (tính bằng phút) như sau:

Phương pháp cũ: 5,6 4,9 5,3 5,1 4,8 5,2 5,5 5,3 4,9 5,0

Phương pháp mới: 4,8 4,6 4,7 4,3 4,6 4,5 4,9 4,6 4,7 4,4

a. Với mức ý nghĩa 0,01 có thể cho rằng phương pháp mới đã thực sự đẩy nhanh tốc độ lắp ráp chi tiết hay không.

b. Nếu giả thiết $\sigma_1^2 = \sigma_2^2$ thì việc kiểm định giả thuyết trên sẽ thay đổi như thế nào.

ĐS: a. Tqs = 5,40

b. Tqs = 5,39

8.24. Với các số liệu cho trong bài 7.32 hãy kiểm định xem chiến dịch quảng cáo có thực sự làm tăng doanh số bán của công ty hay không. Mức ý nghĩa được lấy là 0,01.

8.25. Với các số liệu đã cho trong bài 7.35 hãy kiểm định với mức ý nghĩa 0,01 xem lượng tiền gửi tiết kiệm trung bình của dân hai thành phố A và B có thực sự khác nhau hay không. Nếu sự chênh lệch thực sự là 0,3 triệu thì xác suất mắc sai lầm loại hai bằng bao nhiêu.

8.26. Với các số liệu của bài 7.38 hãy kiểm định với mức ý nghĩa 0,05 xem dự báo về mức lạm phát trung bình của hai nhóm nhà kinh tế có thực sự khác nhau hay không. P-value của kiểm định bằng bao nhiêu.

ĐS: Tqs = 0,7299; P-value < 0,1

8.27. Dùng hai dụng cụ đo lường để đo 5 chi tiết thu được các kết quả sau (đơn vị: mm).

Dụng cụ 1:	6	7	8	5	7
Dụng cụ 2:	7	6	8	7	8

Với mức ý nghĩa 0,05 hãy kết luận xem có sự khác biệt đáng kể giữa hai dụng cụ đo lường hay không.

ĐS: Tqs = -1,18

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.28. Người ta dùng hai loại nguyên liệu A và B để sản xuất thử đế của 10 đôi giày trẻ em sau đó cho các em đi thử trong 6 tháng với cường độ sử dụng tương tự nhau. Kết quả thử nghiệm là trọng lượng đế giày còn lại sau 6 tháng sử dụng được cho trong bảng sau:

Thứ tự	Vật liệu	Giày trái	Giày phải
1	A	180	183
2	A	162	154
3	A	203	189
4	A	194	181
5	A	205	200
6	B	189	185
7	B	168	171
8	B	185	179
9	B	176	175
10	B	169	173

Vậy với mức ý nghĩa 0,1 có thể cho rằng dùng loại nguyên liệu A làm đế giày tốt hơn dùng loại nguyên liệu B hay không.

ĐS: Tqs = 1,39.

3.4. Kiểm định giả thuyết về tham số p của biến ngẫu nhiên phân phối không - một

Giả sử trong tổng thể nghiên cứu biến ngẫu nhiên gốc X phân phối không - một với tham số là p. Như đã thấy ở mục 2 chương VII, p chính là cơ cấu của tổng thể theo dấu hiệu nghiên cứu. Nếu chưa biết p song ta có cơ sở giả thiết rằng giá trị của nó bằng p_O, ta đưa ra giả thuyết thống kê

$$H_O : p = p_O$$

Từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Để kiểm định H_O ta xét các trường hợp sau:

1. Nếu n và p thỏa mãn điều kiện

$$n > 5 \text{ và } \frac{\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right|}{\sqrt{n}} < 0{,}3$$

thì lập thống kê

$$G = U = \frac{(f - p_O)\sqrt{n}}{\sqrt{p_O(1-p_O)}} \tag{8.52}$$

Từ mục 7 chương VI ta đã biết nếu giả thuyết H_0 là đúng thì U phân phối xấp xỉ N(0,1). Do đó với mức ý nghĩa α và tùy thuộc và giả thuyết đối H_1, các miền bác bỏ được xác định như sau:

a) $H_O : p = p_O; H_1 : p > p_O;$

$$W_\alpha = \left\{ U = \frac{(f - p_O)\sqrt{n}}{\sqrt{p_O(1-p_O)}}; U > u_\alpha \right\} \tag{8.53}$$

b) $H_O : p = p_O; H_1 : p < p_O;$

$$W_\alpha = \left\{ U = \frac{(f - p_O)\sqrt{n}}{\sqrt{p_O(1-p_O)}}; U < -u_\alpha \right\} \tag{8.54}$$

c) $H_O : p = p_O; H_1 : p \neq p_O;$

$$W_\alpha = \left\{ U = \frac{(f - p_O)\sqrt{n}}{\sqrt{p_O(1-p_O)}}; |U| > u_{\alpha/2} \right\} \tag{8.55}$$

Với mẫu cụ thể $w = (x_1, x_2, ..., x_n)$ tìm được giá trị quan sát U_{qs} của tiêu chuẩn kiểm định, so sánh với W_α và kết luận.

Thí dụ 12. Tỷ lệ khách hàng tiêu dùng 1 loại sản phẩm ở địa phương A là 60%. Sau một chiến dịch quảng cáo người ta muốn đánh giá xem chiến dịch quảng cáo này có thật sự mang lại hiệu quả hay không. Để làm điều đó người ta đã phỏng vấn ngẫu nhiên 400 khách hàng thì thấy có 250 người tiêu dùng loại sản phẩm nói trên. Với mức ý nghĩa 0,05 hãy kết luận về hiệu quả của chiến dịch quảng cáo đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Gọi p là tỷ lệ khách hàng tiêu dùng loại sản phẩm đó ở địa phương A. Đây là bài toán kiểm định tham số p của phân phối A(p). Cặp giả thuyết thống kê có dạng:

$$H_O: p=0,6;\ H_1: p>0,6$$

vì $n>5$ và $\dfrac{\left|\sqrt{\dfrac{0,6}{0,4}}-\sqrt{\dfrac{0,4}{0,6}}\right|}{\sqrt{400}}=0,02<0,3$

nên ta dùng công thức kiểm định (8.53)

Với $\alpha=0,05\Rightarrow u_\alpha=u_{0,05}=1,645$

Vậy miền bác bỏ là $(1,645;+\infty)$

Với $f=\dfrac{250}{400}=0,625$ ta có:

$$U_{qs}=\frac{(0,625-0,6)\sqrt{400}}{\sqrt{0,6.0,4}}=1,02$$

$U_{qs}\notin W_\alpha$ nên chưa có cơ sở bác bỏ H_0, tức là chưa thể nói chiến dịch quảng cáo có hiệu quả. Giải bằng Excel cho kết quả sau:

0.625	0.6	p (as decimal)
250/400	240/400	p (as fraction)
250.	240.	X
400	400	n
	0.0245	std. error
	1.02	z
	.1537	p-value (one-tailed, upper)

Việc tìm xác suất mắc sai lầm loại 2 trong các trường hợp kiểm định một phía hoặc 2 phía, tìm kích thước mẫu n theo $\alpha;\beta$ và $|p_1-p_0|<\Delta$ cho trước cũng như việc xác định giá trị p cũng được tiến hành giống như đã làm ở các mục trước.

Thí dụ 13. Tiếp tục thí dụ 12

a) Tìm xác suất mắc sai lầm loại 2 nếu $p_1=0,65$.

b) Tìm kích thước mẫu cần điều tra để với các điều kiện như câu a, xác suất mắc sai lầm loại 2 chỉ là 0,1.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải: a) Vì là kiểm định bên phải nên ta có công thức tính β như sau:

$$\beta = P\left[U < u_\alpha - \frac{|p_0 - p_1|}{Se(f)}\right]$$

ta có $$Se(f) = \sqrt{\frac{p_1(1-p_1)}{n}} = \sqrt{\frac{0{,}65.0{,}35}{400}} = 0.0245$$

và với $\alpha = 0{,}05 \Rightarrow u_\alpha = u_{0{,}05} = 1{,}645$

do đó

$$\beta = P\left[U < 1{,}645 - \frac{|0{,}6 - 0{,}65|}{0{,}0245}\right]$$

$$= P[U < -0{,}396] = P[U > 0{,}396] = 0{,}3483$$

Theo công thức tìm kích thước mẫu khi kiểm định một phía:

$$n \geq \left[\frac{2f(1-f)}{\Delta^2}(u_\alpha + u_\beta)^2\right]$$

với $\beta = 0{,}1 \rightarrow u_{0{,}1} = 0{,}46$

$$n \geq \left[\frac{2.0{,}625.0{,}375}{(0{,}65 - 0{,}6)^2}(1{,}645 + 0{,}46)^2\right] = 830{,}8 \Rightarrow n = 831$$

Như vậy, để giảm β từ 0,3372 xuống còn 0,1 thì ta phải điều tra thêm 431 khách hàng nữa.

2. Trường hợp mẫu nhỏ

Nếu kích thước mẫu nhỏ và không thể dùng quy luật chuẩn để kiểm định thì từ mục 7 chương VI ta đã biết lúc đó tần suất mẫu f sẽ phân phối theo quy luật nhị thức với biểu thức xác suất là:

$$P_n(x) = C_n^x p^x (1-p)^{n-x} \qquad f = 0, \frac{1}{n}, \frac{2}{n}, ..., 1 \qquad (8.56)$$

Với điều kiện giả thuyết đối H_1 là đúng thì biểu thức xác suất của f có dạng:

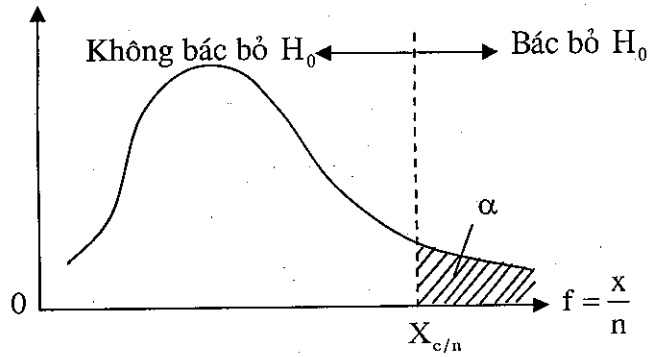
$$P_n(x) = C_n^x p_0^x (1-p_0)^{n-x}$$

Lúc đó tùy thuộc vào giả thuyết đối H_1, miền bác bỏ giả thuyết với mức ý nghĩa α được xác định như sau:

a) $H_0 : p = p_0; H_1 : p > p_0$

Miền bác bỏ bên phải sẽ được mô tả trên hình vẽ như sau (hình 8.3a).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Hình 8.3a

Trong đó $\frac{x_c}{n}$ là giá trị tới hạn. Ở đây, ta dùng dạng liên tục chỉ nhằm mục đích minh họa. Từ đó:

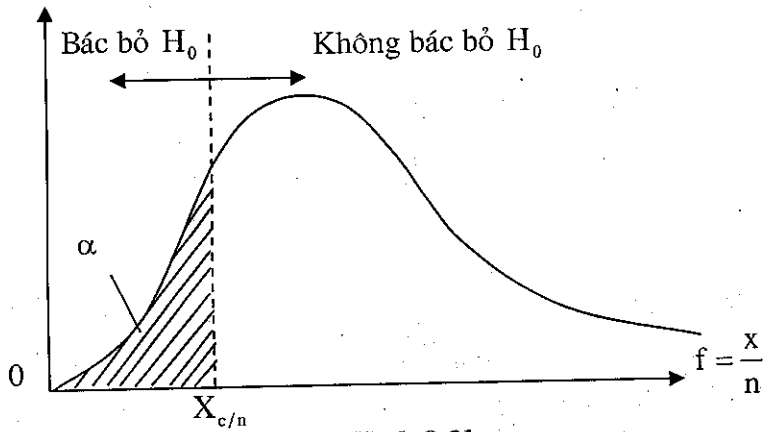
$$P(G \in W_\alpha / H_0) = \sum_{\frac{x}{n}=\frac{x_c}{n}}^{1} C_n^x p_0^x (1-p_0)^{n-x}$$

$$= \sum_{x=x_c}^{n} C_n^x p_0^x (1-p_0)^{n-x} = \alpha \qquad (8.57)$$

Vậy giả thuyết H_0 bị bác bỏ nếu $x \geq x_0$. Với n và p_0 cho trước có thể tìm được giá trị nguyên nhỏ nhất x_0 (bằng cách tra bảng phân phối nhị thức) đảm bảo xác suất mắc sai lầm loại một $\leq \alpha$.

b) $H_0 : p = p_0; H_1 : p < p_0$

Trên đồ thị miền bác bỏ có dạng sau (hình 8.3b).



Hình 8.3b

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

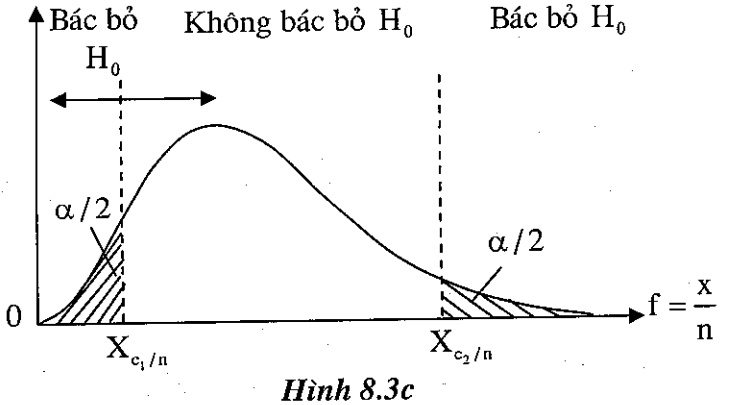
Lúc đó giá trị tới hạn là $x \le x_c$ với x_c là giá trị nguyên lớn nhất thỏa mãn:

$$P(G \in W_\alpha / H_0) = \sum_{x=0}^{x_c} C_n^x p_0^x (1-p_0)^{n-x} = \alpha \qquad (8.58)$$

c) $H_0 : p = p_0; H_1 : p \neq p_0$

Trên đồ thị miền bác bỏ có dạng hình 8.3c.

Lúc đó giá trị tới hạn là giá trị $x \le x_{c_1}$ với x_{c_1} là giá trị nguyên lớn nhất và giá trị $x \ge x_{c_2}$ với x_{c_2} là giá trị nguyên nhỏ nhất thỏa mãn:



Hình 8.3c

$$P(G \in W_\alpha / H_0) = \sum_{x=0}^{x_{c1}} C_n^x p_0^x (1-p_0)^{n-x} + \sum_{x=x_{c2}}^{n} C_n^x p_0^x (1-p_0)^{n-x}$$

$$= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha \qquad (8.59)$$

Thí dụ 14. Một máy gia công một loại chi tiết có tỷ lệ phế phẩm là 10%. Kiểm tra ngẫu nhiên 15 chi tiết thấy có 6 phế phẩm. Vậy, với mức ý nghĩa 0,05 có thể cho rằng tỷ lệ phế phẩm của máy đó đã tăng lên không?

Giải: Gọi p là tỷ lệ phế phẩm của máy đó. Cặp giả thuyết thống kê là:

$$H_0 : p = 0,1; H_1 : p > 0,1$$

Vì mẫu nhỏ và

$$\frac{\left|\sqrt{\frac{0,1}{0,9}} - \sqrt{\frac{0,9}{0,1}}\right|}{\sqrt{15}} = 0,69 > 0,3$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

nên ta dùng công thức (8.57) để kiểm định. Miền bác bỏ được xác định theo bảng sau với n = 15 và p = 0,1 (xem phụ lục 1).

x	$p_n(X)$	$p_n(X \geq x)$	x	$p_n(X)$	$p_n(X \geq x)$
0	0,2059	1	6	0,0019	0,0022
1	0,3432	0,7941	7	0,0003	0,0003
2	0,2669	0,4509	8	0,0000	0
3	0,1285	0,1840	.	.	.
4	0,0428	0,0555	.	.	.
5	0,0105	0,0127	15	0,0000	0

Với $x = 4 \Rightarrow P(x \geq 4) = 0,0555$

Với $x = 5 \Rightarrow P(x \geq 5) = 0,0127$

Để đảm bảo $\alpha \leq 0,05$ thì $x_c = 5$

Do x trong mẫu bằng 6 > 5 nên bác bỏ H_0 thừa nhận H_1 tức là thừa nhận tỷ lệ phế phẩm của máy đó tăng lên.

Bài tập

8.29. Lô hàng đủ tiêu chuẩn xuất khẩu nếu tỷ lệ phế phẩm không vượt quá 3%. Kiểm tra ngẫu nhiên 400 sản phẩm của lô hàng này thấy có 14 phế phẩm. Với mức ý nghĩa 0,05 có cho phép lô hàng xuất khẩu được không.

ĐS: Uqs = 0,586

8.30. Với các số liệu cho trong bài 7.83 hãy kiểm định lại tuyên bố của nhà sản xuất tủ lạnh với mức ý nghĩa 0,01 và tìm khả năng mắc sai lầm loại hai khi kiểm định nếu tỷ lệ tủ lạnh có khuyết tật của nhà sản xuất đó thực sự là 5%.

8.31. Trở lại bài 7.90. Nếu nhà nhập khẩu chỉ chấp nhận lô màn hình đó nếu tỷ lệ các màn hình có khuyết tật tối đa là 5% thì lô hàng đó có thể chấp nhận được không. Hãy kiểm định bằng P-value.

Mặt khác nếu tỷ lệ màn hình có khuyết tật của lô hàng thực sự là 7% và nếu quy định xác suất mắc sai lầm loại 1 và loại 2 tương ứng là 0,05 và 0,02 thì phải kiểm tra bao nhiêu màn hình.

ĐS: P-value = 0,28

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.32. Tỷ lệ phế phẩm do một máy tự động sản xuất là 5%. Kiểm tra ngẫu nhiên 300 sản phẩm thấy có 24 sản phẩm là phế phẩm. Từ đó có ý kiến cho rằng tỷ lệ phế phẩm do máy đó sản xuất có chiều hướng tăng lên. Hãy kết luận ý kiến nêu trên với mức ý nghĩa $\alpha = 0,05$.

ĐS: Miền bác bỏ $W\alpha = (1,645; +\infty)$; $Uqs = 2,39$. Kết luận $Uqs \in W\alpha$ - Có cơ sở nói rằng tỷ lệ phế phẩm tăng lên.

8.33. Nếu áp dụng phương pháp công nghệ thứ nhất thì tỷ lệ phế phẩm là 6%; còn nếu áp dụng phương pháp công nghệ thứ hai thì trong 100 sản phẩm có 5 phế phẩm. Vậy có thể kết luận áp dụng phương pháp công nghệ thứ hai thì tỷ lệ phế phẩm thấp hơn tỷ lệ phế phẩm của phương pháp công nghệ thứ nhất không? Yêu cầu kết luận với mức ý nghĩa $\alpha = 0,05$.

8.34. Tỷ lệ bệnh nhân khỏi bệnh T khi điều trị bằng thuốc A là 85%. Thí nghiệm dùng loại thuốc B để chữa bệnh thì trong số 900 người mắc bệnh T có 810 người được chữa khỏi. Như vậy có thể kết luận thuốc B hiệu quả hơn thuốc A hay không?

ĐS: Miền bác bỏ $W\alpha = (1,645; +\infty)$; $Uqs = 4,17$.

Kết luận: $Uqs \in W\alpha$ - Có cơ sở kết luận dùng thuốc B chữa bệnh T hiệu quả hơn thuốc A.

8.35. Hàm lượng dầu lý thuyết của một loại cây bằng 0,5 (sau một thời gian chăm bón). Lấy ngẫu nhiên một mẫu gồm 100 cây thấy có tỷ lệ dầu $f = 0,7$. Kiểm định giả thuyết Ho: $p = 0,5$ đối với giả thuyết $H_1 : p \neq 0,5$. Với mức ý nghĩa $\alpha = 0,01$.

3.5. Kiểm định giả thuyết về hai tham số p của hai biến ngẫu nhiên phân phối không - một

Giả sử có hai tổng thể nghiên cứu, trong đó các biến ngẫu nhiên X_1 và X_2 cùng phân phối không - một với các tham số tương ứng là p_1 và p_2. Nếu p_1 và p_2 chưa biết song có cơ sở để giả thiết rằng giá trị của chúng bằng nhau, ta đưa ra giả thuyết thống kê:

$H_0 : p_1 = p_2$

Để kiểm định giả thuyết trên, từ các tổng thể rút ra hai mẫu ngẫu nhiên độc lập kích thước n_1 và n_2.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$W_1 = (X_{11}, X_{12}, ..., X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, ..., X_{2n_2})$$

Và chọn lập tiêu chuẩn kiểm định là thống kê:

$$G = U = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Từ mục 7 chương VI ta đã biết thống kê U phân phối xấp xỉ chuẩn hóa nếu $n_1 > 30$ và $n_2 > 30$

Nếu giả thuyết H_0 là đúng $(p_1 = p_1 = p)$ thì tiêu chuẩn kiểm định trở thành:

$$G = U = \frac{f_1 - f_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

thông thường p chưa biết nên được thay bằng ước lượng của nó là:

$$\bar{f} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

Như vậy ta có tiêu chuẩn kiểm định:

$$G = U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (8.60)$$

phân phối xấp xỉ N(0,1) nếu $n_1 > 30$ và $n_2 > 30$. Do đó tùy thuộc vào giả thuyết đối H_1 các miền bác bỏ mức α được xác định như sau:

a) $H_0 : p_1 = p_2; H_1 : p_1 > p_2$

$$W_\alpha = \left\{ U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}; U > u_\alpha \right\} \qquad (8.61)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b) $H_0 : p_1 = p_2; H_1 : p_1 < p_2$

$$W_\alpha = \left\{ U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}; U < -u_\alpha \right\} \qquad (8.62)$$

c) $H_0 : p_1 = p_2; H_1 : p_1 \neq p_2$

$$W_\alpha = \left\{ U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}; |U| > u_{\alpha/2} \right\} \qquad (8.63)$$

Với hai mẫu cụ thể ta tính được các giá trị cụ thể của f_1, f_2 và $\bar{f}$ và giá trị quan sát U_{qs} của tiêu chuẩn kiểm định và so sánh với W_α để kết luận.

Thí dụ 15: Kiểm tra ngẫu nhiên các sản phẩm cùng loại do hai nhà máy sản xuất được các số liệu sau:

Bảng 8.4

Nhà máy	Số sản phẩm được kiểm tra	Số phế phẩm
A	$n_1 = 1000$	$x_1 = 20$
B	$n_2 = 900$	$x_2 = 30$

Với mức ý nghĩa $\alpha = 0{,}05$ có thể coi tỷ lệ phế phẩm của hai nhà máy là như nhau hay không?

Giải: Gọi p_1 và p_2 tương ứng là tỷ lệ phế phẩm của hai nhà máy A và B. Như vậy, đây là bài toán kiểm định cặp giả thuyết $H_0 : p_1 = p_2$ và $H_1 : p_1 \neq p_2$ với n_1 và n_2 khá lớn.

Theo công thức (8.63) tiêu chuẩn kiểm định là:

$$U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{1000} + \frac{1}{900}\right)}}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Do $\alpha = 0,05 \rightarrow u_{\alpha/2} = u_{0,025} = 1,96$ nên miền bác bỏ là $(-\infty; -1,96)$ và $(1,96; +\infty)$

Với hai mẫu cụ thể ta tìm được:

$$f_1 = \frac{20}{1000} = 0,02; f_2 = \frac{30}{900} = 0,033;$$

$$\bar{f} = \frac{20+30}{1000+900} = 0,0263;$$

$$U_{qs} = \frac{0,02 - 0,033}{\sqrt{0,0263.0,9737\left(\frac{1}{1000} + \frac{1}{900}\right)}} = -1,81$$

$U_{qs} \notin W_\alpha$ vậy chưa có cơ sở để bác bỏ H_0, tức là có thể coi tỷ lệ phế phẩm ở hai nhà máy là như nhau.

Giải bằng Excel cho kết quả sau:

p_1	p_2	p_c	
0.02	0.0333	0.0263	p (as decimal)
20/1000	30/900	50/1900	p (as fraction)
20.	29.97	49.97	X
1000	900	1900	n
	-0.0133	difference	
	0.	hypothesized difference	
	0.0074	std. error	
	-1.81	z	
	.0705	p-value (two-tailed)	

Việc mở rộng thủ tục kiểm định cho k (k >2) tỷ lệ p của k tổng thể phân phối không - một sẽ được tiến hành bằng phương pháp kiểm định phi tham số.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

8.36. Tại hai xí nghiệp A và B có số liệu sau về nhân công: Xí nghiệp A có 200 công nhân thì năm 2009 có 30 người xin chuyển đi chỗ khác. Xí nghiệp B có 350 công nhân thì năm 2009 có 65 người thôi việc. Vậy với mức ý nghĩa 0,05 có thể cho rằng tỷ lệ công nhân thôi việc ở xí nghiệp A thấp hơn xí nghiệp B hay không.

ĐS: Uqs = 1,071

8.37. Từ dây chuyền thứ nhất khi kiểm tra 100 chi tiết người ta loại đi 30 chi tiết, từ dây chuyền thứ hai khi kiểm tra 150 chi tiết người ta loại đi 40 chi tiết. Vậy nếu cho rằng chất lượng sản phẩm của hai dây chuyền là như nhau thì P-value bằng bao nhiêu.

ĐS: P-value = 0,28

8.38. Vào lúc 9 giờ lấy ngẫu nhiên 50 chi tiết do một máy sản xuất thì có 5 chi tiết hỏng. Và lúc 12 giờ lấy ngẫu nhiên 40 chi tiết cũng do máy đó sản xuất thì có 7 chi tiết hỏng. Vậy với mức ý nghĩa 0,01 có thể cho rằng tỷ lệ chi tiết hỏng đã thực sự tăng lên theo thời gian sản xuất hay không.

ĐS: Uqs = -1,04.

8.39. Hiện tượng học sinh bỏ học là vấn đề đang được đặc biệt quan tâm, nhất là ở nông thôn. Tại hai trường trung học ở hai vùng nông thôn A và B trong năm học 2009 - 2010 có các số liệu thống kê sau:

Trường	Số học sinh	Số học sinh bỏ học
A	1900	175
B	2600	325

a. Với mức ý nghĩa 0,05 có thể cho rằng tình trạng bỏ học ở trường B là nghiêm trọng hơn ở trường A hay không.

b. Tìm xác suất mắc sai lầm loại 2 nếu tỷ lệ học sinh bỏ học ở hai trường thật sự chênh lệch nhau là 2%.

8.40. Kiểm tra chất lượng sản phẩm ở hai xí nghiệp cùng loại ta có kết quả sau:

Xí nghiệp	Số sản phẩm được kiểm tra	Số phế phẩm
I	1800	54
II	1200	30

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với mức ý nghĩa $\alpha = 0,05$ có thể coi tỷ lệ phế phẩm ở hai xí nghiệp như nhau hay không?

ĐS: Miền bác bỏ $W\alpha = [(-\infty, -1,96); (1,96, +\infty,]$

$Uqs = 0,81$ - Vậy $Uqs \notin W\alpha$. Tỷ lệ phế phẩm ở hai xí nghiệp có thể coi là như nhau.

8.41. Thống kê số tai nạn lao động tại hai xí nghiệp có các số liệu sau:

Xí nghiệp	Số công nhân	Số tai nạn lao động
I	200	20
II	800	120

Với mức ý nghĩa $\alpha = 0,05$ hãy kết luận xem chất lượng của công tác bảo hộ lao động tại hai xí nghiệp trên có khác nhau không?

ĐS: Miền bác bỏ $W\alpha = [(-\infty, -1,96); (1,96, +\infty,];$

$Uqs = -1,82$ - Chưa có cơ sở để kết luận là khác nhau vì $Uqs \notin W\alpha$.

8.42. Hai máy tiện giống nhau nhưng hoạt động trong các điều kiện thời tiết khác nhau. Sau một thời gian sản xuất, người ta nghi ngờ chất lượng hoạt động của chúng khác nhau. Điều đó có đúng không nếu trong 1000 sản phẩm do máy 1 sản xuất có 140 phế phẩm và trong số 2000 sản phẩm do máy 2 sản xuất có 260 phế phẩm. Hãy kết luận với mức ý nghĩa $\alpha = 0,05$.

3.6. Kiểm định giả thuyết về phương sai của biến ngẫu nhiên phân phối chuẩn

Giả sử trong tổng thể biến ngẫu nhiên gốc X phân phối $N(\mu, \sigma^2)$ với σ^2 chưa biết song có cơ sở để giả thiết rằng giá trị của nó bằng σ_0^2. Người ta đưa giả thuyết thống kê $H_0 : \sigma^2 = \sigma_0^2$.

Để kiểm định giả thuyết trên từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

và chọn tiêu chuẩn kiểm định là thống kê:

$$G = \chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \tag{8.64}$$

Nếu giả thuyết H_0 đúng thì theo mục 7 Chương VI thống kê χ^2 phân phối theo quy luật "khi bình phương" với (n-1) bậc tự do. Do đó với mức ý nghĩa α cho trước và tùy thuộc vào dạng của giả thuyết đối H_1 miền bác bỏ W_α được xây dựng theo các trường hợp sau:

a) $H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 > \sigma_0^2$

$$W_\alpha = \left\{ \chi^2 = \frac{(n-1)S^2}{\sigma_0^2}; \chi^2 > \chi_\alpha^{2(n-1)} \right\} \tag{8.65}$$

b) $H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 < \sigma_0^2$

$$W_\alpha = \left\{ \chi^2 = \frac{(n-1)S^2}{\sigma_0^2}; \chi^2 < \chi_{1-\alpha}^{2(n-1)} \right\} \tag{8.66}$$

c) $H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 \neq \sigma_0^2$

$$W_\alpha = \left\{ \chi^2 = \frac{(n-1)S^2}{\sigma_0^2}; \chi^2 < \chi_{1-\alpha/2}^{2(n-1)} \text{ hoặc } \chi^2 > \chi_{\alpha/2}^{2(n-1} \right\} \tag{8.67}$$

Với mẫu cụ thể ta tìm được giá trị cụ thể s^2 và giá trị quan sát χ_{qs}^2 của tiêu chuẩn kiểm định, so sánh với W_α và kết luận.

Thí dụ 16: Để kiểm tra độ chính xác của một máy người ta đo ngẫu nhiên kích thước của 15 chi tiết do máy đó sản xuất và tính được $s^2 = 14,6$. Với mức ý nghĩa $\alpha = 0,01$ hãy kết luận máy móc có hoạt động bình thường không, biết rằng kích thước chi tiết là biến ngẫu nhiên phân phối chuẩn có dung sai theo thiết kế là $\sigma^2 = 12$.

Giải: Gọi X là kích thước chi tiết, theo giả thiết X phân phối chuẩn. Vậy đây là bài toán kiểm định cặp giả thuyết:

$$H_0 : \sigma^2 = 12$$
$$H_1 : \sigma^2 > 12$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ (8.65) tiêu chuẩn kiểm định có dạng:

$$\chi^2 = \frac{14.S^2}{12}$$

Do $\alpha = 0{,}01 \Rightarrow \chi_{\alpha}^{2(14)} = \chi_{0{,}01}^{2(14)} = 29{,}14$

nên miền bác bỏ có dạng $(29{,}14;+\infty)$

Với mẫu cụ thể ta có giá trị quan sát của tiêu chuẩn kiểm định:

$$\chi^2 = \frac{14.14{,}6}{12} = 17{,}033$$

$\chi_{qs}^2 \notin W_{\alpha}$ vậy chưa có cơ sở để bác bỏ H_0 hay có thể nói máy móc vẫn làm việc bình thường.

Dùng Excel cho phép vừa kiểm định, vừa tìm khoảng tin cậy hai phía mức 99%. Kết quả như sau:

Chi-square Variance Test

12.000	hypothesized variance
14.600	observed variance of x
15	n
14	df
17.03	chi-square
.5088	p-value (two-tailed)
6.526	confidence interval 99.% lower
50.164	confidence interval 99.% upper

Bài tập

8.43. Từ một mẫu kích thước n = 15 rút ra từ tổng thể phân phối chuẩn người ta tìm được $s^2 = 144$. Với mức ý nghĩa 0,01 hãy kiểm định cặp giả thuyết.

H_0: $\sigma^2 = 138$; H_1 : $\sigma^2 > 138$

ĐS: $\chi^2 qs = 14{,}61$

8.44. Một nhà sản xuất bóng đèn tuýp cho rằng chất lượng bóng đèn sẽ được coi là đồng đều nếu tuổi thọ của các bóng đèn có độ lệch chuẩn bằng 1000 giờ hoặc ít hơn. Lấy ngẫu nhiên 10 bóng để kiểm tra thì tìm được độ lệch chuẩn của mẫu s = 1150. Vậy với mức ý nghĩa 0,05 có thể coi chất

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

lượng bóng đèn do công ty đó sản xuất là đồng đều hay không biết tuổi thọ của bóng đèn phân phối chuẩn.

ĐS: $\chi^2 qs = 11,9$.

8.45. Với các số liệu của bài 7.52 giả sử độ phân tán của thành phần chính trong mỗi viên thuốc theo quy định không được vượt quá 0,05 thì với mức ý nghĩa 0,01 có thể chấp nhận lô thuốc mới nhập về hay không.

8.46. Trọng lượng của con gà lúc mới nở là biến ngẫu nhiên phân phối chuẩn. Nghi ngờ độ đồng đều về trọng lượng gà con giảm sút người ta cân thử 12 con và tìm được $s^2 = 11,41\ (gam)^2$ - Với mức ý nghĩa $\alpha = 0,05$, hãy kết luận về điều nghi ngờ trên biết rằng bình thường độ phân tán của trọng lượng gà con là $\sigma^2 = 10\ (gam)^2$.

ĐS: Chưa có cơ sở để nghi ngờ rằng độ đồng đều về trọng lượng gà con giảm sút.

8.47. Nếu máy móc hoạt động bình thường thì trọng lượng sản phẩm tuân theo quy luật chuẩn với độ lệch chuẩn 1 kg. Có thể coi máy móc còn hoạt động tốt hay không nếu cân thử 30 sản phẩm ta thấy độ lệch chuẩn mẫu tăng lên tới 1,1 (kg). Yêu cầu kết luận với mức ý nghĩa $\alpha = 0,01$.

ĐS: Miền bác bỏ $W\alpha = (49,6, +\infty)$; $\chi^2 qs = 35,09$. Chưa có cơ sở để nói rằng máy móc hoạt động không bình thường.

3.7. Kiểm định giả thuyết về sự bằng nhau của hai phương sai của hai biến ngẫu nhiên phân phối chuẩn

Giả sử có hai tổng thể nghiên cứu trong đó biến ngẫu nhiên gốc X_1 trong tổng thể thứ nhất phân phối $N(\mu_1, \sigma_1^2)$ và biến ngẫu nhiên gốc X_2 trong tổng thể thứ hai phân phối $N(\mu_2, \sigma_2^2)$. Nếu σ_1^2 và σ_2^2 chưa biết song có cơ sở để giả thiết rằng giá trị của chúng bằng nhau thì ta đưa ra giả thuyết thống kê: $H_0 : \sigma_1^2 = \sigma_2^2$.

Để kiểm định giả thuyết trên từ hai tổng thể rút ra hai mẫu ngẫu nhiên độc lập kích thước tương ứng là n_1 và n_2:

$$W_1 = (X_{11}, X_{12}, ..., X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, ..., X_{2n_2})$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

và chọn tiêu chuẩn kiểm định là thống kê:

$$G = F = \frac{S_1^2.\sigma_2^2}{S_2^2.\sigma_1^2} \qquad (8.68)$$

nếu $S_1^2 > S_2^2$. Từ mục 7 Chương VI ta biết thống kê F phân phối theo quy luật Fisher-Snedecor với $(n_1 - 1)$ và $(n_2 - 1)$ bậc tự do.

Nếu giả thuyết H_0 đúng thì tiêu chuẩn kiểm định có dạng:

$$F = \frac{S_1^2}{S_2^2} \qquad (S_1^2 > S_2^2)$$

và vẫn phân phối $F(n_1 - 1, n_2 - 1)$. Do đó với mức ý nghĩa α cho trước và tùy thuộc vào dạng của giả thuyết đối H_1 có thể xây dựng được các miền bác bỏ W_α tương ứng như sau:

a) $H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma_1^2 > \sigma_2^2$

$$W_\alpha = \left\{ F = \frac{S_1^2}{S_2^2}; F > f_\alpha^{(n_1-1, n_2-1)} \right\} \qquad (8.69)$$

b) $H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma_1^2 \neq \sigma_2^2$

$$W_\alpha = \left\{ F = \frac{S_1^2}{S_2^2}; F < f_{1-\alpha/2}^{(n_1-1, n_2-1)} \text{ hoặc } F > f_{\alpha/2}^{(n_1-1, n_2-1)} \right\} \qquad (8.70)$$

Với hai mẫu cụ thể tìm được s_1^2, s_2^2 và giá trị quan sát F_{qs} của tiêu chuẩn kiểm định, so sánh với W_α và kết luận.

Thí dụ 17: Có hai giống lúa có năng suất trung bình xấp xỉ như nhau song mức độ phân tán về năng suất có thể khác nhau. Để kiểm tra điều đó người ta gặt mẫu tại 2 vùng trồng hai giống lúa đó và thu được kết quả sau:

Bảng 8.4

Giống lúa	Số điểm gặt	Phương sai
A	$n_1 = 41$	$s_1^2 = 11{,}41$
B	$n_2 = 30$	$s_2^2 = 6{,}52$

Với mức ý nghĩa $\alpha = 0{,}05$ hãy kết luận về vấn đề trên, biết năng suất lúa là biến ngẫu nhiên phân phối chuẩn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải: Gọi X_1 và X_2 là năng suất của 2 giống lúa A và B, X_1 và X_2 phân phối chuẩn. Đây là bài toán kiểm định cặp giả thuyết:

$$H_0 : \sigma_1^2 = \sigma_2^2; H_1 : \sigma_1^2 \neq \sigma_2^2$$

Theo (8.70) tiêu chuẩn kiểm định có dạng $F = \frac{S_1^2}{S_2^2}$

Do $\alpha = 0,05$ nên $f_{\alpha/2}^{(40,29)} = f_{0,025}^{(40,29)} = 2,03;$

$$f_{1-\alpha/2}^{(40,29)} = \frac{1}{f_{\alpha/2}^{(29,40)}} = \frac{1}{1,94} = 0,52;$$

Vậy miền bác bỏ là $(0;0,52)$ và $(2,03;+\infty)$

Với mẫu cụ thể ta có:

$$F_{qs} = \frac{11,41}{6,52} = 1,75$$

$F_{qs} \notin W_\alpha$ vậy chưa có cơ sở bác bỏ H_0 hay độ phân tán của năng suất hai giống lúa trên là như nhau. Sau đây là kết quả giải bằng Excel:

F-test for equality of variance

11.42	variance: x1
6.503	variance: x2
1.76	F
.1164	p-value

Thí dụ: Với mức ý nghĩa 0,05 hãy kiểm định xem độ phân tán về thu nhập của dân cư vùng 2 có lớn hơn vùng 1 theo các số liệu của thí dụ A hay không?

Kết quả giải bằng Excel như sau:

F-Test Two-Sample for Variances

	Variable 1	*Variable 2*
Mean	1649.75	1683.86
Variance	2257.442	3518.93
Observations	100	100
df	99	99
F	0.641514	
P(F<=f) one-tail	0.014118	
F Critical one-tail	0.717329	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

8.48. Hai máy cùng gia công một loại chi tiết. Người ta muốn kiểm tra xem hai máy này có độ chính xác như nhau hay không. Để làm điều đó người ta lấy ngẫu nhiên từ mỗi máy 7 chi tiết, đem đo và thu được kết quả sau:

Máy A: 135 138 136 140 138 135 139

Máy B: 140 135 140 138 135 138 140

Với mức ý nghĩa 0,05 có thể cho rằng hai máy có độ chính xác như nhau hay không. Biết kích thước chi tiết phân phối chuẩn.

ĐS: Fqs = 1,28

8.49. Để so sánh độ chính xác của hai thiết bị đo lường người ta tiến hành lấy mẫu và thu được kết quả sau:

Thiết bị A	Thiết bị B
Đo 25 lần	Đo 20 lần
Phương sai sai số = 14,5	Phương sai sai số = 17,2

Giả thiết rằng sai số đo lường là biến ngẫu nhiên phân phối chuẩn. Với mức ý nghĩa 0,05 có thể cho rằng độ chính xác của hai thiết bị là như nhau được không.

ĐS: Fqs = 1,19

8.50. Để so sánh độ phân tán về trọng lượng đóng gói của hai máy đóng bao người ta lấy từ hai máy đó hai mẫu và thu được kết quả sau:

Máy A	Máy B
$n_1 = 12$	$n_2 = 15$
$\bar{x}_1 = 1{,}5$	$\bar{x}_2 = 2{,}0$
$\sum_{i=1}^{12} X_{i1}^2 = 49$	$\sum_{i=1}^{15} X_{i2}^2 = 158$

Hãy tìm P-value để kiểm định xem độ phân tán về trọng lượng đóng bao của hai máy có như nhau hay không. Biết trọng lượng đóng bao là biến ngẫu nhiên phân phối chuẩn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.51. Một số nhà kinh tế cho rằng độ phân tán của thị phần trong các công ty hoạt động có cạnh tranh về giá cả cao hơn trong các công ty độc quyền. Để kết luận về điều đó người ta đã điều tra thị phần của một công ty có cạnh tranh về giá cả trong 4 năm và tìm được phương sai mẫu là 85,576. Đồng thời điều tra thị phần của một công ty độc quyền trong 7 năm tìm được phương sai mẫu là 13,78. Với mức ý nghĩa 0,05 hãy kết luận về ý kiến trên. Giả thiết thị phần của các công ty là biến ngẫu nhiên phân phối chuẩn.

8.52. Độ rủi ro trong đầu tư thường được đo bằng phương sai của tỷ lệ thu hồi vốn. Để quyết định xem nên đầu tư vào ngành nào một người đã thu thập số liệu về hai ngành kinh tế và thu được kết quả sau:

Ngành A	Ngành B
Số dự án điều tra: 10	Số dự án điều tra: 15
Tốc độ hoàn vốn trung bình: 10,48	Tốc độ hoàn vốn trung bình: 11
Phương sai: 1,44	Phương sai: 16

Vậy với mức ý nghĩa 0,05 có thể cho rằng rủi ro đầu tư ở ngành B cao hơn ngành A hay không. Giả thiết tỷ lệ thu hồi vốn là biến ngẫu nhiên phân phối chuẩn.

ĐS: Fqs = 11,11

3.8. Kiểm định k phương sai của k biến ngẫu nhiên phân phối chuẩn

Giả sử có k tổng thể nghiên cứu, trong đó các biến ngẫu nhiên $X_1, X_2, ..., X_k$ cùng phân phối chuẩn. Từ các tổng thể rút ra k mẫu độc lập kích thước n_1, n_2,..., n_k và tìm được các phương sai mẫu tương ứng $S_1^2, S_2^2, ..., S_k^2$.

Với mức ý nghĩa α phải kiểm định cặp giả thuyết:

$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

H_1 : Có ít nhất hai phương sai khác nhau

Như vậy, cần kiểm định xem sự khác biệt của các phương sai mẫu là có ý nghĩa hay không.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Để chọn lập tiêu chuẩn kiểm định ta xét hai trường hợp sau:

1. Nếu các kích thước mẫu $n_1, n_2, ..., n_k$ khác nhau

Lúc đó nếu ký hiệu $\overline{S^2}$ là trung bình số học của các phương sai mẫu:

$$\overline{S^2} = \frac{\sum_{i=1}^{n} h_i S_i^2}{h}$$

trong đó $h = \sum_{i=1}^{k} h_i$ với $h_i = n_i - 1$ tức là số bậc tự do của phương sai mẫu S_i^2 thì tiêu chuẩn kiểm định là thống kê Bartlett sau đây:

$$B = \frac{V}{C} \tag{8.71}$$

trong đó

$$V = 2{,}303\left[h.\lg\overline{S^2} - \sum_{i=1}^{n} h_i \lg S_i^2\right]$$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{n}\frac{1}{h_i} - \frac{1}{h}\right]$$

Với điều kiện giả thuyết H_0 là đúng thì thống kê B phân phối xấp xỉ khi bình phương với k - 1 bậc tự do nếu mọi $h_i > 2$ tức là mọi $n_i > 3$.

Lúc đó miền bác bỏ mức α được xác định bằng biểu thức:

$$W_\alpha = \left\{B = \frac{V}{C}; B > \chi_\alpha^{2(k-1)}\right\} \tag{8.72}$$

Thí dụ 18: Từ 4 mẫu độc lập kích thước $n_1 = 10$, $n_2 = 13$, $n_3 = 15$, $n_4 = 16$ được rút ra từ các tổng thể phân phối chuẩn tìm được các phương sai mẫu tương ứng bằng 0,25; 0,4; 0,36; 0,46. Với mức ý nghĩa 0,05 hãy kiểm định xem phương sai của các tổng thể có bằng nhau hay không.

Giải: Cặp giả thuyết có dạng

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

H_1 : Có ít nhất hai phương sai khác nhau

Theo công thức kiểm định (8.72) ta có với

$$\alpha = 0{,}05 \rightarrow \chi_\alpha^{2(k-1)} = \chi_{0,05}^{2(3)} = 7{,}8$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vậy miền bác bỏ là $(7,8;+\infty)$

Để tìm B_{qs} ta lập bảng tính sau:

Mẫu i	Kích thước n_i	Số bậc tự do h_i	Phương sai S_i^2	$h_i s_i^2$	$\lg s_i^2$	$h_i \lg s_i^2$	$\frac{1}{h_i}$
1	10	9	0,25	2,25	-1,3979	-12,5811	0,11
2	13	12	0,4	4,8	-19,2252	-19,2252	0,08
3	15	14	0,36	5,04	-21,7822	-21,7822	0,07
4	16	15	0,46	6,9	-24,942	-24,942	0,07
$\sum$				18,99		-78,53	0,32

Từ đó

$$\overline{s^2} = \frac{\sum h_i s_i^2}{h} = \frac{18,99}{50} = 0,3798$$

$$\Rightarrow \lg \overline{s^2} = -1,5795$$

$$V = 2,303[h.\lg \overline{s^2} - \sum h_i \lg s_i^2]$$

$$= 2,303[50(-1,5795) + 78,5350] = -1,01$$

$$C = 1 + \frac{1}{3(4-1)}\left[0,32 - \frac{1}{50}\right] = 1,03$$

$$\Rightarrow B = -0,98$$

$B \notin W_\alpha$ nên chưa thể bác bỏ H_0, tức là các phương sai của tổng thể có thể xem như là bằng nhau với mức ý nghĩa 0,05

Sau đây là kết quả giải bằng Stata cho thí dụ A

oneway x v. tabulate

	Summary of Thu nhap ca nhan		
v	Mean	Std.Dev.	Freq.
1	1649.73	47.374897	100
2	1655.99	47.408284	100
3	1628.12	48.399699	100
Total	1644.6133	49.050351	300

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Analysis of Varianace

Source	SS	df	MS	F	Prob>F
Between groups	42763.8867	2	21381.9433	9.39	0.0001
Within groups	676611.26	297	2278.15239		
Total	719375.147	299	2405.93695		

Bartlett's test for eaqual variances: chi2(2) = 0.0585

Prob>chi2=0.971

Nếu chấp nhận H_0 thì có thể lấy $\overline{s^2}$ để ước lượng phương sai đồng đều của các tổng thể. Chẳng hạn, trong bài toán trên có thể ước lượng phương sai của các tổng thể là:

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 \approx \overline{s^2} = 0,3798$$

2. ***Nếu các kích thước mẫu*** $n_1 = n_2 = ... = n_k = n$

Trong trường hợp này ta có thể sử dụng tiêu chuẩn kiểm định Bartlett đã xét ở trên, tuy nhiên do phân phối xác suất của tiêu chuẩn Bartlett chỉ là xấp xỉ nên kết quả kiểm định không thật đáng tin cậy. Với trường hợp kích thước mẫu bằng nhau thì ta dùng tiêu chuẩn kiểm định Cochran sẽ cho kết quả tin cậy hơn. Tiêu chuẩn Cochran là thống kê:

$$G = \frac{s^2_{max}}{s_1^2 + s_2^2 + ... + s_k^2} \qquad (8.73)$$

tức là tỉ số giữa phương sai mẫu lớn nhất và tổng các phương sai mẫu. Phân phối xác suất của thống kê G chỉ phụ thuộc vào số bậc tự do (n-1) và số lượng mẫu k. Lúc đó, miền bác bỏ mức α được xác định bằng biểu thức:

$$W_\alpha = \left\{ G = \frac{s^2_{max}}{s_1^2 + s_2^2 + ... + s_k^2}; G > g_\alpha^{(n-1,k)} \right\} \qquad (8.74)$$

các giá trị tới hạn $g_\alpha^{(n-1,k)}$ được tìm trong bảng tính sẵn (Phụ lục 13). Với k mẫu cụ thể tìm được G_{qs} so sánh với W_α và kết luận.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 19. Từ 4 mẫu độc lập kích thước như nhau và bằng 17 rút ra từ các tổng thể phân phối chuẩn tìm được các phương sai mẫu tương ứng là 0,26; 0,36; 0,40; 0,42.

a) Với mức ý nghĩa 0,05 hãy kiểm định sự bằng nhau của các phương sai tổng thể.

b) Hãy ước lượng các phương sai tổng thể nếu có thể cho rằng chúng bằng nhau.

Giải.

a) Ta tìm giá trị quan sát của tiêu chuẩn Cochran:

$$G_{qs} = \frac{0,42}{0,26+0,36+0,40+0,42} = 0,2917$$

Với $\alpha = 0,05$, số bậc tự do 17-1=16 và số lượng mẫu là 4 ta tìm được giá trị tới hạn là:

$$g_{\alpha}^{(n-1,k)} = g_{0,05}^{(16,4)} = 0,4366$$

vậy miền bác bỏ là $(0,43666;+\infty)$

Do $G_{qs} \notin W_{\alpha}$ nên chưa thể bác bỏ H_0 hay có thể cho rằng các phương sai tổng thể bằng nhau.

b) Vì ta thừa nhận H_0 nên phương sai đồng đều của tổng thể có thể ước lượng bằng trung bình số học của các phương sai mẫu.

$$\sigma^2 \approx \bar{s}^2 = \frac{0,26+0,36+0,40+0,42}{4} = 0,36$$

Bài tập

8.53. Từ 5 mẫu độc lập kích thước tương ứng là $n_1 = 7$; $n_2 = 9$; $n_3 = 10$; $n_4 = 12$; $n_5 = 12$ được rút ra từ các tổng thể phân phối chuẩn tìm được các phương sai mẫu tương ứng là 0,27; 0,32; 0,40; 0,42; 0,48. Với mức ý nghĩa 0,05 hãy kiểm định giả thuyết về sự bằng nhau của các phương sai tổng thể.

ĐS: Vqs = 6.63; $\chi^{2(4)}_{0.05} = 9,5$ => chưa có cơ sở để bác bỏ Ho.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.54. Từ 4 mẫu có cùng kích thước n = 17 rút ra từ các tổng thể phân phối chuẩn tìm được các phương sai mẫu tương ứng là 2,12; 2,32; 2,24; 2,32.

a. Với mức ý nghĩa 0,05 hãy kiểm định giả thuyết về sự bằng nhau của các phương sai tổng thể.

b. Hãy ước lượng phương sai đồng đều của các tổng thể.

ĐS: a. Gqs = 0,36

b. $\delta^2 \approx 3$.

4. TÓM TẮT CHƯƠNG 8

1. Kiểm định giả thuyết thống kê là một phương pháp của thống kê suy diễn, trong đó người ta dùng thông tin của mẫu để kiểm tra tính đúng đắn của một nhận định tiên nghiệm liên quan đến tổng thể.
2. Giả thuyết thống kê được gọi là giả thuyết gốc và mệnh đề đối lập với nó gọi là giả thuyết đối.
3. Giả thuyết thống kê có thể là giả thuyết đơn hoặc giả thuyết kép.
4. Để kiểm định một giả thuyết thống kê, người ta chọn ra một tiêu chuẩn kiểm định và với một mức ý nghĩa cho trước đi tìm miền bác bỏ tốt nhất sau đó so sánh với giá trị quan sát của tiêu chuẩn kiểm định và kết luận.
5. Khi kiểm định một giả thuyết thống kê có thể mắc sai lầm loại một và sai lầm loại hai. Xác suất không mắc sai lầm loại hai gọi là lực kiểm định.
6. Với việc khống chế khả năng mắc sai lầm loại một và loại hai ở một mức cho trước có thể tìm được kích thước mẫu cần điều tra để thỏa mãn các yêu cầu đó.
7. Kiểm định tham số thường dựa trên một giả thiết nào đó về quy luật phân phối xác suất của dấu hiệu nghiên cứu trên tổng thể.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các ký hiệu và công thức cơ bản

* H_0 -Giả thuyết gốc

* H_1 -Giả thuyết đối

* G -Tiêu chuẩn kiểm định

* G_{qs} -Giá trị quan sát của tiêu chuẩn kiểm định

* α -Mức ý nghĩa của kiểm định và là xác suất mắc sai lầm loại một

$$\alpha = P(G \in W_\alpha / H_0)$$

* β -Xác suất mắc sai lầm loại hai

$$\beta = P(G \notin W_\alpha / H_1)$$

* $1-\beta$ -Lực kiểm định

$$1-\beta = P(G \in W_\alpha / H_1)$$

* W_α -Miền bác bỏ giả thuyết. Nếu $G_{qs} \in W_\alpha \rightarrow$ Bác bỏ H_0

* P-value=$P(G>G_{qs})$ nếu là miền bác bỏ bên phải

P-value=$P(G<G_{qs})$ nếu là miền bác bỏ bên trái

P-value=$2P(G>|G_{qs}|)$ nếu là miền bác bỏ hai phía.

* Kiểm định U về μ (σ đã biết)

$H_0 : \mu = \mu_0$;

Tiêu chuẩn kiểm định

$$U = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} \sim N(0,1)$$

$\beta = P[U < u_\alpha - \frac{|\mu_0 - \mu_1|}{Se(\bar{x})}]$ nếu là miền bác bỏ một phía

$\beta = P[U < u_{\alpha/2} - \frac{|\mu_0 - \mu_1|}{Se(\bar{x})}]$ nếu là miền bác bỏ hai phía

$n \geq \left[\frac{\sigma^2(u_\alpha + u_\beta)^2}{\Delta^2}\right]$ nếu là miền bác bỏ một phía

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$n \geq \left[\frac{\sigma^2(u_{\alpha/2}+u_\beta)^2}{\Delta^2}\right]$$ nếu là miền bác bỏ hai phía

* Kiểm định T về μ (σ chưa biết)

$$H_0 : \mu = \mu_0$$

Tiêu chuẩn kiểm định

$$T = \frac{(\bar{x}-\mu_0)\sqrt{n}}{S} \sim T(n-1)$$

$$\beta = P[T < t_\alpha^{(n-1)} - \frac{|\mu_0-\mu_1|}{Se(\bar{x})}]$$ nếu là kiểm định một phía

$$\beta = P[T < t_{\alpha/2}^{(n-1)} - \frac{|\mu_0-\mu_1|}{Se(\bar{x})}]$$ nếu là kiểm định hai phía

$$n \geq \left[\frac{S^2}{\Delta^2}(t_\alpha^{(m-1)} + t_\beta^{(m-1)}\right]$$ nếu là kiểm định một phía

$$n \geq \left[\frac{S^2}{\Delta^2}(t_{\alpha/2}^{(m-1)} + t_\beta^{(m-1)}\right]$$ nếu là kiểm định hai phía

* Kiểm định hai kỳ vọng toán: $H_0 : \mu_1 = \mu_2$

+ Trường hợp đã biết σ_1^2 và σ_2^2

Tiêu chuẩn kiểm định

$$U = \frac{(\bar{x_1}-\bar{x_2})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

+ Trường hợp chưa biết σ_1^2 và σ_2^2 ($\sigma_1^2 = \sigma_2^2$)

Tiêu chuẩn kiểm định

$$T = \frac{(\bar{x_1}-\bar{x_2})}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim T(n_1+n_2-2)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Trường hợp chưa biết σ_1^2 và σ_2^2 $(\sigma_1^2 \neq \sigma_2^2)$

Tiêu chuẩn kiểm định

$$T = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim T(k)$$

+ Trường hợp hai mẫu phụ thuộc theo từng cặp. Tiêu chuẩn kiểm định

$$T = \frac{\overline{D}\sqrt{n}}{S_D} \sim T(n-1)$$

* Kiểm định về P; $H_0 : p = p_0$

Tiêu chuẩn kiểm định

$$U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \sim N(0,1)$$

* Kiểm định về hai tham số p; $H_0 : p_1 = p_2$

Tiêu chuẩn kiểm định

$$U = \frac{f_1 - f_2}{\sqrt{\overline{f}(1-\overline{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

trong đó $\overline{f} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$

* Kiểm định về σ^2, $H_0 : \sigma^2 = \sigma_0^2$

Tiêu chuẩn kiểm định

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

* Kiểm định về hai tham số σ^2; $H_0 : \sigma_1^2 = \sigma_2^2$

Tiêu chuẩn kiểm định

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

* Kiểm định k phương sai

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$$

+ Trường hợp kích thước mẫu khác nhau

Tiêu chuẩn kiểm định Bartlett

$$B = \frac{V}{C} \sim \chi^2(k-1)$$

trong đó

$$V = 2,303\left[h.\lg\overline{S^2} - \sum_{i=1}^{k} h_i \lg S_1^2\right]$$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{h_i} - \frac{1}{h}\right]$$

+ Trường hợp kích thước mẫu bằng nhau

Tiêu chuẩn kiểm định Cochran:

$$G = \frac{S_{max}^2}{S_1^2 + S_2^2 + ... + S_k^2}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Kiểm định giả thuyết thống kê	Hypothesis tests
Giả thuyết thống kê	Statistical hypothesis
Giả thuyết gốc	Null hypothesis
Giả thuyết đối	Alternative hypothesis
Giả thuyết đơn	Simple hypothesis
Giả thuyết hợp	Compose hypothesis
Tiêu chuẩn kiểm định giả thuyết	Test statistic
Giá trị quan sát của TCKĐ	Observed value of test statistic
Miền bác bỏ giả thuyết	Rejection region
Miền không bác bỏ giả thuyết	Non-rejection region
Miền bác bỏ bên phải	Right regection region
Miền bác bỏ bên trái	Left rejection region
Miền bác bỏ hai phía	Two - taied rejection region
Điểm (Giá trị tới hạn)	Critical (Value) point
Mức ý nghĩa của kiểm định	Significance level
Sai lầm loại một	Type I error
Sai lầm loại hai	Type II error
Xác suất mắc sai lầm loại 1	Probability of a type I error
Xác suất mắc sai lầm loại 2	Probability of a type II error
Lực kiểm định	Testing power
Thủ tục kiểm định giả thuyết TK	Decision rule
Giá trị xác suất của kiểm định	P - value
Định lý Neyman – Pearson	Neyman- Pearson theorem
Kiểm định tham số	Parametric hypothesis test

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

8.55. Khi kiểm định một giả thuyết thống kê, nếu kết luận giá trị của thống kê mẫu khác biệt một cách có ý nghĩa so với giá trị giả thuyết thì điều đó có ý nghĩa gì?

8.56. Sai lầm loại một có ý nghĩa như thế nào? Nó ảnh hưởng như thế nào đến giá trị tới hạn của kiểm định?

8.57. Khi tăng kích thước của mẫu lên thì điều đó sẽ ảnh hưởng như thế nào đến:

a) Xác suất mắc sai lầm loại 1?

b) Giá trị tới hạn?

Hãy mô tả điều đó bằng đồ thị.

8.58. Sai lầm loại 2 là gì? Quan hệ của nó đối với sai lầm loại 1 như thế nào?

8.59. Phát biểu sau đây là đúng hay sai

$$P(\text{sai lầm loại 1}) + P(\text{sai lầm loại 2}) = 1$$

8.60. Tại sao khi P-value có giá trị càng nhỏ thì ta lại càng nghiêng về việc bác bỏ giả thuyết H_0.

8.61. Độ lệch chuẩn của mẫu có ảnh hưởng như thế nào đến giá trị tới hạn của kiểm định?

8.62. Hãy phát biểu cặp giả thuyết thống kê cho các tình huống sau đây:

a) Một nhà máy sản xuất kẹo tuyên bố là trọng lượng trung bình của mỗi gói kẹo là 500 gram. Kiểm tra ngẫu nhiên 500 gói kẹo tìm được trọng lượng trung bình mỗi gói là 450 gram và độ lệch chuẩn là 100 gram.

b) Một nhà sản xuất tủ lạnh tuyên bố là tỷ lệ tủ lạnh phải bảo hành khi sử dụng không vượt quá 3%. Theo dõi ngẫu nhiên 170 tủ lạnh đã bán ra thấy có 12 chiếc phải bảo hành.

c) Trước chiến dịch quảng cáo, điều tra ngẫu nhiên 10 tuần tìm được doanh số trung bình đối với một loại mỹ phẩm là 35 triệu/tuần và độ lệch chuẩn là 3 triệu. Sau chiến dịch quảng cáo, theo dõi ngẫu nhiên 12 tuần tìm được doanh số trung bình là 37 triệu/tuần và độ lệch chuẩn là 4 triệu.

d) Ở địa phương A, xét nghiệm ngẫu nhiên 500 người thấy 50 người có ký sinh trùng sốt rét. Ở địa phương B, xét nghiệm ngẫu nhiên 1000 người thấy 120 người có ký sinh trùng sốt rét.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.63. Phân biệt sự khác nhau giữa giả thuyết đơn và giả thuyết kép. Cho ví dụ.

8.64. Tại sao khi giá trị quan sát của tiêu chuẩn kiểm định không thuộc vào miền bác bỏ giả thuyết thì ta lại chỉ có thể nói rằng chưa có cơ sở để bác bỏ giả thuyết H_0.

Bài tập tổng hợp chương 8

8.65. Các công ty bảo hiểm y tế địa phương than phiền rằng số tiền bảo hiểm trung bình phải trả cho người mua bảo hiểm y tế cao hơn so với mức tính toán của tổng công ty bảo hiểm là 2 triệu đồng. Điều tra ngẫu nhiên 100 trường hợp được thanh toán bảo hiểm y tế tìm được tiền thanh toán trung bình cho họ là 2,7 triệu và độ lệch chuẩn là 0,947 triệu. Vậy với mức ý nghĩa 0,01 có thể cho rằng lời than phiền trên của các công ty địa phương là có căn cứ hay không?

ĐS: $T_{qs} = 7,39$

8.66. Một công ty sản xuất hàng điện tử tuyên bố là tỷ lệ sản phẩm phải bảo hành của công ty chỉ là 12%. Người ta điều tra ngẫu nhiên 100 sản phẩm do công ty bán ra và thấy 17 sản phẩm phải bảo hành. Với mức ý nghĩa 0,05 có thể cho rằng tỷ lệ sản phẩm phải bảo hành của công ty thực sự cao hơn so với tuyên bố hay không.

ĐS: $U_{qs} = 1,538$.

8.67. Theo đánh giá của Hội bảo vệ người tiêu dùng thì giá xe máy trong 6 tháng đầu năm đã tăng lên đến 49%. Để kiểm chứng lại điều đó người ta đã đi điều tra ngẫu nhiên ở 18 cửa hàng bán xe máy và tìm được mức tăng giá trung bình so với 6 tháng trước là 38% và độ lệch chuẩn là 14%. Với mức ý nghĩa là 0,05 hãy kết luận về đánh giá trên.

8.68. Nếu như phương sai của thời gian chờ đợi của khách hàng tại một ngân hàng vượt quá 1,5 phút^2 thì cần có biện pháp cải tiến hoạt động tiếp khách. Điều tra ngẫu nhiên 60 khách hàng đến giao dịch tại ngân hàng đó tìm được phương sai của thời gian chờ đợi của họ là 1,78 phút^2. Hãy tìm P-value để kiểm định xem có cần cải tiến hoạt động tiếp khách hay không.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.69. Người ta muốn nghiên cứu xem việc thay đổi bao bì đối với một loại đồ uống có làm tăng lượng sản phẩm tiêu thụ hay không. Số liệu của 15 cửa hàng bán loại đồ uống trên trước và sau khi thay đổi bao bì như sau (đơn vị: sản phẩm/ngày):

Cửa hàng:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Trước:	57	61	12	38	12	69	5	39	88	9	92	26	14	70	22
Sau:	60	54	20	35	21	70	1	65	79	10	90	32	19	77	29

Với mức ý nghĩa 0,05 có thể cho rằng việc thay đổi bao bì có làm tăng thêm lượng đồ uống bán được hay không.

ĐS: $T_{qs} = 1{,}469$

8.70. Một mẫu gồm 2060 khách hàng được phỏng vấn thì 13% trong số họ ưa chuộng loại dầu gội đầu A, một mẫu khác gồm 5000 khách hàng được phỏng vấn thì 19% trong số họ ưa chuộng loại dầu gội đầu B hơn. Với mức ý nghĩa 0,05 có thể cho rằng tỷ lệ khách hàng ưa chuộng loại dầu gội A và B là như nhau được không.

8.71. Một độ đo quan trọng để đánh giá mức độ rủi ro của cổ phiếu là phương sai hay độ lệch chuẩn của giá cổ phiếu. Người ta điều tra ngẫu nhiên giá cổ phiếu của công ty A trong 25 ngày và tìm được phương sai mẫu là 6,52.

Với công ty B điều tra giá cổ phiếu trong 22 ngày tìm được phương sai mẫu là 3,47. Vậy với mức ý nghĩa 0,01 có thể cho rằng cổ phiếu của công ty A có độ rủi ro cao hơn công ty B hay không?

8.72. Số ngày trung bình mà khách du lịch ở các khách sạn của thành phố là 3,4. Điều tra ngẫu nhiên 28 khách du lịch đến thành phố tìm được số ngày ở khách sạn của họ là 5 4 3 2 1 1 5 7 8 4 3 3 2 5 7 1 3 1 1 5 3 4 2 2 2 6 1 7.

Với mức ý nghĩa 0,05 hãy kiểm định xem số ngày trung bình mà khách du lịch lưu lại ở khách sạn có thay đổi hay không.

ĐS: $T_{qs} = 0{,}248$

8.73. Một nghiên cứu về phụ nữ cho rằng 30% chủ các doanh nghiệp nhỏ là phụ nữ. Người ta điều tra ngẫu nhiên 2000 doanh nghiệp nhỏ thì thấy có 852 doanh nghiệp do phụ nữ làm chủ. Tìm P-value để kiểm định xem kết luận trên có đúng hay không.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.74. Theo dõi trong 17 tháng ở cảng A thu được các số liệu sau về số tàu vào cảng xếp dỡ trong tháng là 9, 6, 7, 7, 8, 7, 9, 4,7, 6, 6, 5,7, 5, 9, 8, 6. Với cảng B trong 9 tháng số tàu vào trong tháng là 10, 5, 8, 8, 9, 10, 4, 9, 11. Nếu giả thiết số tàu vào cả hai cảng cùng phân phối chuẩn thì với mức ý nghĩa 0,05 hãy kiểm định xem số tàu trung bình vào hai cảng hàng tháng có như nhau hay không?

8.75. Cũng với mức ý nghĩa 0,05 hãy kiểm định giả thiết là phương sai của số tàu vào hai cảng là như nhau, và nếu chấp nhận giả thuyết đó thì tìm ước lượng của phương sai của số tàu vào hai cảng đó.

ĐS: $F_{qs} = 2,527$.

8.76. Tỷ lệ thu hồi của vốn đầu tư được biết là có phân phối chuẩn với trung bình là 11% một năm và độ lệch chuẩn là 2% một năm. Một công ty muốn kiểm định lại xem điều đó có đúng hay không nên đã thu thập một mẫu ngẫu nhiên sau đây về tỷ lệ thu hồi vốn đầu tư:

8,8%	9,4%	10%	12,3%	12,9%	7%	8%	9%	9,5%	8,6%
13%	14,5%	12%	12,4%	19%	9%	10%	10%	11,7%	15%
10,1%	12,7%	17%	8%	9,9%	11%	12,5%	12,8%	10,6%.	

Với mức ý nghĩa 0,05 hãy kiểm định các giả thuyết sau:

a. Tỷ lệ thu hồi vốn đầu tư trung bình hàng năm là 11%

b. Độ lệch chuẩn của tỷ lệ thu hồi vốn là 2%.

8.77. Một xí nghiệp có hai dây chuyền lắp ráp cùng một loại chi tiết. Quan sát 31 lần lắp ráp tìm được thời gian trung bình mỗi dây chuyền phải dừng giữa hai lần lắp ráp và phương sai như sau:

	Dây chuyền 1	Dây chuyền 2
Trung bình	60,4 phút	65,3 phút
Phương sai	31 phút^2	44 phút^2

Với mức ý nghĩa là 0,05 hãy cho biết.

a. Thời gian dừng máy trung bình của hai dây chuyền có như nhau hay không?

b. Phương sai của thời gian dừng máy của dây chuyền hai có lớn hơn dây chuyền một hay không?

Biết thời gian dừng máy là biến ngẫu nhiên phân phối chuẩn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.78. Một hợp tác xã trồng thử hai giống lúa, mỗi giống trên 31 thửa ruộng và được chăm sóc như nhau. Cuối vụ thu hoạch thu được kết quả như sau:

	$\bar{x}$	s
Giống lúa A	45	2,5
Giống lúa B	46,5	4

Với mức ý nghĩa 0,05, hãy kết luận xem:

a. Năng suất trung bình của hai giống lúa có thể xem là như nhau hay không?

b. Nếu chấp nhận ý kiến ở câu a, thì nên chọn giống lúa nào để đưa vào sản xuất đại trà?

8.79. Điều tra ngẫu nhiên 100 sinh viên tốt nghiệp của mỗi trường đại học A và B sau một năm ra trường cho kết quả sau:

Trường	Đã đi làm	Học tiếp	Chưa có việc làm
A	60	12	28
B	55	10	35

a. Với mức ý nghĩa 0,05 có thể cho rằng tỷ lệ sinh viên đi làm sau khi tốt nghiệp một năm của trường A cao hơn B hay không?

b. Với độ tin cậy 0,95 hãy ước lượng số sinh viên chưa có việc làm của trường A nếu năm đó trường A có 4000 sinh viên tốt nghiệp ra trường.

8.80. Điều tra 100 hộ dân cư để xem xét nước sạch tiêu dùng bình quân theo đầu người trong một tháng, người ta thu được số liệu sau:

Khu vực/ (m^3/người)	1	2	3	4	7	8
Thị xã A	4	10	40	30	10	6
Thị xã B	6	10	30	40	10	4

a. Ước lượng mức độ tiêu dùng nước sạch bình quân theo đầu người của thị xã A bằng khoảng tin cậy đối xứng, với độ tin cậy 0,95.

b. Có ý kiến cho rằng những hộ dùng từ mức 6m^3/ 1người trở lên là dùng lãng phí. Và những hộ dùng không quá 2m^3/ 1người là bị thiếu nước sạch.

Với $\alpha = 0,05$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Hãy ước lượng tỷ lệ số hộ bị thiếu nước sạch tối thiểu ở hai thị xã A và B.

- Có thể nói rằng tỷ lệ số hộ dùng nước lãng phí của thị xã A lớn hơn thị xã B được không?

8.81. Để nghiên cứu chiều cao của nam thanh niên ở nước A và B, người ta lấy ngẫu nhiên ở mỗi nước một mẫu gồm 100 nam thanh niên và tiến hành đo chiều cao (tính bằng cm) thì được kết quả là:

Nước A:

Chiều cao	163-167	167-171	171-175	175-179	179-183	183-187
Số người	6	9	21	35	20	9

Nước B:

Chiều cao	163-167	167-171	171-175	175-179	179-183	183-187
Số người	8	13	17	29	22	11

a. Với độ tin cậy 95%, ước lượng chiều cao trung bình tối đa của nam thanh niên ở mỗi nước.

b. Có ý kiến cho rằng độ đồng đều về chiều cao của nam thanh niên ở hai quốc gia trên là khác nhau. Với mức ý nghĩa 0,05 hãy nhận xét về ý kiến này.

c. Người ta cho rằng tỷ lệ cao trên 175cm của nam thanh niên ở nước A ít nhất là bằng tỷ lệ này ở nước B. Với mức ý nghĩa 5%, hãy nhận xét về ý kiến này.

Giả sử chiều cao nam thanh niên ở mỗi nước là các biến ngẫu nhiên có quy luật phân phối chuẩn.

8.82. a. Người ta cho rằng tỷ lệ thất nghiệp ở thành phố A và thành phố B là khác nhau. Để kiểm tra điều này, ở mỗi thành phố người ta chọn ngẫu nhiên 1000 người đang trong độ tuổi lao động thì thấy rằng có 75 người ở thành phố A và 72 người ở thành phố B bị thất nghiệp. Với mức ý nghĩa 0,05 hãy kết luận về điều nhận định trên.

b. Trong điều tra trên, người ta tính được độ lệch chuẩn mẫu của mức lương của những người đang có việc làm ở thành phố A là 20000 đồng, còn ở

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

thành phố B là 24000. Với mức ý nghĩa là 0,05 hãy kiểm định giả thiết rằng độ đồng đều trong mức lương ở thành phố A cao hơn ở thành phố B.

8.83. Thống kê 1500 đứa trẻ sinh ra trong tháng tư cho thấy có 780 con trai.

Cũng tại địa bàn điều tra cũ, số liệu điều tra vào đầu tháng chín cho thấy trong 1800 đứa trẻ sinh ra có 980 đứa là con trai.

a. Có thể cho rằng xác suất sinh con trai vào tháng chín cao hơn xác suất sinh con trai vào tháng tư không? Hãy kết luận với mức ý nghĩa 5%.

b. Trong tháng chín, xác suất sinh con trai lớn hơn xác suất sinh con gái không? Cho kết luận với mức ý nghĩa 1%.

8.84. Hai cửa hàng A (chủ là người ngoài xí nghiệp) và B (chủ là một cổ đông của xí nghiệp) nhận làm đại lý bán cùng một loại sản phẩm cho một xí nghiệp.

Trước đây, hai cửa hàng được nhận mức hoa hồng như nhau do doanh số bình quân hàng tháng xấp xỉ nhau.

Mới đây cửa hàng A có đề nghị xí nghiệp tăng mức hoa hồng cho mình với lý do doanh số bình quân hàng tháng của mình nhiều hơn cửa hàng B.

Để đi tới quyết định, xí nghiệp đã xem xét bảng thống kê doanh số hàng tháng của hai cửa hàng trong 3 năm gần đây và thu được các số liệu sau:

	Cửa hàng A	Cửa hàng B
Doanh số bình quân	50 triệu đồng	30 triệu đồng
Phương sai mẫu	900 (triệu đồng)^2	700 (triệu đồng)^2

Qua kết quả điều tra này, với mức ý nghĩa 5%, xí nghiệp có nên chấp nhận lời đề nghị của cửa hàng A hay không, biết rằng doanh số hàng tháng của cả hai cửa hàng đều tuân theo quy luật chuẩn.

8.85. Khi biết tin cửa hàng A được tăng mức hoa hồng, cửa hàng B phản đối vì cho rằng xí nghiệp đã có một số quyết định sai lầm, có thể ảnh hưởng tới lãi suất cổ phần của mình.

Một thành viên trong ban quản trị xí nghiệp cho rằng điều phản đối này cũng có thể có lý vì doanh số bình quân hàng tháng của hai cửa hàng có thể vẫn như nhau. Hãy xét xem sự tán đồng của thành viên này có thể đúng ở mức xác suất là bao nhiêu.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội. 1999.
2. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB ĐH Kinh tế Quốc dân. Hà Nội. 2008.
3. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà Nội. 2009.
4. Ngô Văn Thứ. *Thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội. 2011
5. Harald Cramer. *Phương pháp toán học trong thống kê*. NXB Khoa học và Kỹ thuật. Hà Nội. 1970.
6. Гмурман В.Е. Теория вероятностей и математическая статистика. Москва. 1972.
7. Уйякс. Математическая статистика. Москва. 1982.
8. Giard V. Statistique appliquée à la gestion. Economica. Paris. 1992.
9. Amir D Aczel. Complete Business Statistics. Irwin Boston 1993.
10. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992.
11. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons. New York. 1990.
12. Paul Newbold: Statistics for Business & Economics (Fourth Edition). Pretice-Hall International, Inc.

THƯ VIỆN HUBT

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

THƯ VIỆN
HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 9

KIỂM ĐỊNH PHI THAM SỐ

1. MỞ ĐẦU

Ở chương trước, ta đã xét các kiểm định tham số. Đặc điểm chung của các kiểm định này là phải dựa trên một số điều kiện nào đó liên quan đến tổng thể nghiên cứu, chẳng hạn kiểm định T và kiểm định F phải dựa trên giả thiết là tổng thể phân phối chuẩn. Khi các giả thuyết thống kê phụ thuộc vào các giả thiết về tổng thể và các tham số của nó, ta có các kiểm định tham số.

Trong thực tế có nhiều trường hợp ta không đưa ra bất kỳ giả thiết nào liên quan đến tham số của tổng thể hoặc dạng phân phối xác suất của tổng thể. Lúc đó, các kiểm định được xét ở mục trước không thể áp dụng được. Các kiểm định trong điều kiện như vậy gọi là kiểm định phi tham số.

Như vậy *kiểm định phi tham số* là các thủ tục thống kê để kiểm định giả thuyết khi không có được các giả thiết liên quan đến tham số của tổng thể hay dạng phân phối xác suất của tổng thể.

Chương này sẽ xem xét một số kiểm định phi tham số quan trọng nhất.

Mục 2 sẽ nghiên cứu kiểm định khi bình phương và các ứng dụng khác nhau của nó. Mục 3 sẽ đề cập đến một số phương pháp khác để kiểm định quy luật phân phối xác suất của biến ngẫu nhiên trong tổng thể. Mục 4 sẽ nghiên cứu phương pháp kiểm định theo dấu để kiểm định trung vị của tổng thể. Mục 5 và 6 sẽ đề cập đến hai phương pháp kiểm định Wilcoxon về hai trung bình tổng thể. Mục 7 sẽ xem xét phương pháp kiểm định Kruskal - Wallis về k trung bình tổng thể. Mục 8 và 9 sẽ nghiên cứu hai phương pháp kiểm định đoạn mạch về tính ngẫu nhiên của mẫu và sự đồng nhất của hai tổng thể về dạng phân phối xác suất.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2. KIỂM ĐỊNH KHI BÌNH PHƯƠNG

Kiểm định khi bình phương dựa trên việc sử dụng tiêu chuẩn χ^2 của K.Pearson là một trong những kiểm định phi tham số quan trọng nhất. Nó thường được sử dụng để giải quyết các bài toán sau đây trong thực tế.

2.1. Kiểm định giả thuyết về tính độc lập của hai dấu hiệu định tính

Giả sử cần nghiên cứu đồng thời hai dấu hiệu định tính A và B trên cùng một tổng thể. Dấu hiệu A có các phạm trù là $A_1, A_2, ..., A_h$ còn dấu hiệu B có các phạm trù là $B_1, B_2, ..., B_k$. Nếu có cơ sở để giả thiết rằng A và B độc lập ta đưa ra cặp giả thuyết thống kê sau đây:

H_0 : A và B độc lập

H_1 : A và B phụ thuộc

Để kiểm định giả thuyết trên, từ tổng thể lập mẫu kích thước n và trình bày các số liệu mẫu dưới dạng *bảng tiếp liên* sau đây:

A \ B	B_1	B_2	...	B_j	...	B_k	Tổng số n_i
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	n_1
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	n_2
. .	. .	. .	. .	. .	. .	. .	. .
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	n_i
. .	. .	. .	. .	. .	. .	. .	. .
A_h	n_{h1}	n_{h2}	...	n_{hj}	...	n_{hk}	n_h
Tổng số m_j	m_1	m_2	...	m_j	...	m_k	$\sum = n$

Trong đó:

n là kích thước mẫu;

n_i là tổng các tần số ứng với dấu hiệu thành phần A_i;

m_j là tổng các tần số ứng với dấu hiệu thành phần B_j;

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

n_{ij} là tần số ứng với các phần tử đồng thời mang dấu hiệu A_i và B_j.

Với n khá lớn thì theo định nghĩa thống kê về xác suất ta có:

$$P(A_iB_j) \approx \frac{n_{ij}}{n} \qquad i = \overline{1,h}; \qquad j = \overline{1,k}$$

$$P(A_i) \approx \frac{n_i}{n} \qquad i = \overline{1,h};$$

$$P(B_j) \approx \frac{n_j}{n} \qquad j = \overline{1,k}$$

Nếu giả thiết H_0 đúng tức là A và B độc lập thì các dấu hiệu thành phần cũng độc lập nên:

$$P(A_iB_j) = P(A_i).P(B_j)$$

Tức là: $$\frac{n_{ij}}{n} = \frac{n_i}{n}.\frac{m_j}{n} \qquad i = \overline{1,h}; \qquad j = \overline{1,k}$$

Vì thế tiêu chuẩn kiểm định giả thuyết có thể được chọn là thống kê sau:

$$G = \chi^2 = n\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{\left(\frac{n_{ij}}{n} - \frac{n_i}{n}.\frac{m_j}{n}\right)}{\frac{n_i}{n}.\frac{m_j}{n}}$$

Hay $$\chi^2 = n\left[\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i.m_j} - 1\right] \qquad (9.1)$$

Nếu giả thuyết H_0 đúng và n khá lớn thì ta có thể coi thống kê χ^2 phân phối theo qui luật "khi bình phương" với $(h-1)(k-1)$ bậc tự do. Vì vậy, với mức ý nghĩa α miền bác bỏ của H_0 là:

$$W_\alpha = \left\{\chi^2 = n\left[\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i.m_j} - 1\right]; \chi^2 > \chi_\alpha^{2(h-1)(k-1)}\right\} \qquad (9.2)$$

Dựa vào mẫu cụ thể tính được giá trị quan sát χ^2_{qs}, so sánh với W_α và kết luận.

Thí dụ 1. Nghiên cứu sự ảnh hưởng của hoàn cảnh gia đình đối với tình trạng phạm tội của trẻ em ở tuổi vị thành niên qua điều tra ngẫu nhiên 148 em nhỏ thu được kết quả sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bảng 9.1

Hoàn cảnh gia đình / Tình trạng phạm tội	Bố hoặc mẹ đã chết	Bố mẹ li hôn	Còn cả bố mẹ	Tổng cộng n_i
Không phạm tội	20	25	18	63
Phạm tội	29	43	13	85
Tổng cộng m_j	49	68	31	$\sum = 148$

Với mức ý nghĩa = 0,05 có thể kết luận hoàn cảnh gia đình độc lập với tình trạng phạm tội của trẻ em hay không?

Giải. Đây là bài toán kiểm định tính độc lập của hai dấu hiệu định tính với cặp giả thuyết là H_0: Tình trạng phạm tội độc lập với hoàn cảnh gia đình; H_1: Tình trạng phạm tội phụ thuộc hoàn cảnh gia đình;

Tiêu chuẩn kiểm định là:

$$\chi^2 = 148\left[\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i.m_j} - 1\right]$$

Do $\alpha = 0,05 \Rightarrow \chi_{\alpha}^{2(h-1)(k-1)} = \chi_{0,05}^{2(2-1)(3-1)} = 5,991$

Qua mẫu cụ thể tính được $\chi_{qs}^2 = 4,0552$

Kết quả giải bằng Excel như sau:

	Col 1	Col 2	Col 3	Total
Row 1	**20**	**25**	**18**	63
Row 2	**29**	**43**	**13**	85
Total	49	68	31	148

4.04	chi-square
2	df
.1324	p-value

$\chi_{qs}^2 \notin W_{\alpha}$, vậy chưa có cơ sở để bác bỏ H_0, tức là có thể xem hoàn cảnh gia đình và tình trạng phạm tội là độc lập nhau.

Phương pháp kiểm định khi bình phương vừa xét ở trên cho phép ta kết luận xem hai dấu hiệu định tính A và B có độc lập với nhau hay không. Nếu giả thuyết H_0 bị bác bỏ tức là A và B phụ thuộc (tương quan) thì sẽ nảy sinh nhu cầu đánh giá mức độ chặt chẽ của sự phụ thuộc đó. Vấn đề đánh giá mức độ chặt chẽ của sự phụ thuộc giữa hai dấu hiệu định tính, giữa hai hoặc nhiều dấu hiệu định lượng sẽ được giải quyết bằng phương pháp phân tích tương quan ở chương XI.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

9.1. Để nghiên cứu xem quy mô của một công ty có ảnh hưởng đến hiệu quả quảng cáo đối với khách hàng hay không, người ta đã điều tra ý kiến của 356 khách hàng và thu được kết quả sau:

Quy mô công ty	Hiệu quả quảng cáo		
	Mạnh	Vừa phải	Yếu
Nhỏ	20	52	32
Vừa	53	47	28
Lớn	67	32	25

Với mức ý nghĩa $\alpha = 0,1$ có thể cho rằng quy mô của công ty có ảnh hưởng đến hiệu quả của quảng cáo đối với khách hàng hay không?

ĐS: $\chi^2_{qs} = 29,664$; $\chi^2_{qs} \in W\alpha \Rightarrow$ bác bỏ H_0

9.2. Nghiên cứu ảnh hưởng của thành phần thức ăn của bố mẹ (X) đối với giới tính (Y) của con cái, thu được kết quả sau:

Y \ X	Không có Vitamin	Có Vitamin	Tổng cộng
Nam	123	145	268
Nữ	153	150	303
Tổng cộng	276	295	571

Với mức ý nghĩa $\alpha = 0,01$ có thể xem X và Y độc lập với nhau được không?

ĐS: $\chi^2_{qs} = 1,2$; $W\alpha = (6,6 + \infty)$ - Thành phần thức ăn của bố mẹ không ảnh hưởng tới giới tính của con cái.

9.3. Nghiên cứu tình trạng hôn nhân trước ngày cưới của 542 cặp vợ chồng, ta có bảng số liệu sau:

Tình trạng hôn nhân của vợ / Tình trạng hôn nhân của chồng	Chưa kết hôn lần nào	Ly hôn	Góa	Tổng cộng
Chưa kết hôn lần nào	180	34	36	250
Ly hôn	58	76	54	188
Góa	43	34	27	104
Tổng cộng:	281	144	117	542

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với α = 0,05 có thể kết luận tình trạng hôn nhân của vợ và chồng trước ngày cưới độc lập nhau không?

ĐS: Miền bác bỏ Wα = (9,5; ∞); $\chi^2_{qs} = 79,99$

$\chi^2_{qs} \in W\alpha$ - bác bỏ tính độc lập với nhau giữa tình trạng hôn nhân của vợ và của chồng trước ngày cưới.

9.4. Để lập kế hoạch sản xuất một công ty đã cho điều tra sở thích của khách hàng về ba loại mẫu khác nhau là A, B, C của cùng một loại hàng. Kết quả điều tra cho ở bảng sau:

Mẫu hàng / Ý kiến	A	B	C	Σ
Thích	43	30	42	115
Không thích	35	53	39	127
Không có ý kiến	22	17	19	58
Σ	100	100	100	300

Có hay không sự phân biệt về sở thích của khách hàng đối với ba loại mẫu nói trên? Cho kết luận với mức ý nghĩa 5%.

ĐS: $\chi^2_{qs} = 7,606$

9.5. Điều tra ngẫu nhiên thu nhập của 400 công nhân ở Hà Nội và thành phố Hồ Chí Minh, thu được kết quả sau (đơn vị tính: triệu đồng/năm)

Thu nhập / Thành phố	0-5	5-10	10-15	> 15
Hà Nội	28	42	30	24
Tp HCM	44	78	78	76

Với mức ý nghĩa 0,05 hãy kiểm định xem thu nhập của công nhân có phụ thuộc vào thành phố mà họ làm việc hay không?

ĐS: $\chi^2_{qs} = 5,81$

9.6. Điều tra ngẫu nhiên 200 sinh viên thì thấy 60% trong số họ là ở nội trú và 40% là ở ngoại trú. Tỷ lệ sinh viên học khá, trung bình, kém trong số sinh viên nội trú tương ứng là 30%, 40%, 30%. Tỷ lệ này ở các sinh viên ngoại trú là 20%, 50%, 30%. Với mức ý nghĩa là 0,05 có thể cho rằng nơi cư trú của sinh viên có ảnh hưởng đến chất lượng học tập hay không?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.2. Kiểm định giả thuyết về k tham số p của k biến ngẫu nhiên phân phối không- một

Khi mở rộng việc so sánh hai tham số p của phân phối A(p) cho trường hợp có nhiều hơn hai tổng thể thì có thể sử dụng thủ tục kiểm định khi bình phương như sau:

Giả sử có k tổng thể nghiên cứu trong đó các biến ngẫu nhiên $X_1, X_2, ..., X_k$ cùng phân phối không - một với các tham số tương ứng là $p_1, p_2, ..., p_k$. Nếu có cơ sở để giả thiết rằng giá trị của chúng bằng nhau thì ta đưa ra cặp giả thuyết sau:

$$H_0 = p_1 = p_2 = ... = p_k = p$$

H_1: Có ít nhất hai giá trị khác nhau.

Để kiểm định cặp giả thuyết trên từ các tổng thể lập k mẫu ngẫu nhiên độc lập kích thước tương ứng là $n_1, n_2, ..., n_k$ và giả thiết các số liệu mẫu được cho dưới dạng bảng sau (Bảng 9.2)

Bảng 9.2

Mẫu	Số lần xuất hiện biến cố	Số lần không xuất hiện biến cố	Tổng số n_i
1	X_{11}	X_{12}	n_1
.	.	.	.
i	X_{i1}	X_{i2}	n_i
.	.	.	.
k	X_{k1}	X_{k2}	n_k
Tổng số m_j	m_1	m_2	$\sum n_i = n$

Lúc đó nếu $n_1, n_2, ..., n_k$ đều lớn hơn 50 thì các thống kê:

$$U_i = \frac{X_{i1} - n_i p_i}{\sqrt{n_i p_i (1 - p_i)}} \qquad i = \overline{1,k}$$

sẽ phân phối xấp xỉ N(0,1). Vì vậy, thống kê:

$$\chi^2 = \sum_{i=1}^{k} U_i^2 = \sum_{i=1}^{k} \frac{(X_{i1} - n_i p_i)^2}{n_i p_i (1 - p_i)} \qquad (9.3)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

sẽ phân phối khi bình phương với k bậc tự do. Nếu giả thuyết H_0 là đúng $(p_1 = p_2 = ... = p_k)$ thì do p thường chưa biết nên có thể thay bằng ước lượng của nó là:

$$\bar{f} = \frac{X_{11} + X_{21} + ... + X_{k1}}{n_1 + n_k + ... + n_2}$$

và tiêu chuẩn kiểm định (9.3) trở thành:

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_{i1} - n_i \bar{f})^2}{n_i \bar{f}(1 - \bar{f})} \qquad (9.4)$$

và χ^2 sẽ phân phối khi bình phương với (k-1) bậc tự do. Mặt khác, có thể biến đổi (9.4) thành biểu thức tương đương sau đây:

$$\chi^2 = n\left[\sum_{i=1}^{k} \sum_{j=1}^{2} \frac{X_{ij}^2}{n_i m_j} - 1\right] \qquad (9.5)$$

Do đó miền bác bỏ mức α được xác định bằng biểu thức:

$$W_\alpha = \left\{\chi^2 = n\left[\sum_{i=1}^{k} \sum_{j=1}^{2} \frac{X_{ij}^2}{n_i m_j} - 1\right]; \chi^2 > \chi_\alpha^{2(k-1)}\right\} \qquad (9.6)$$

Từ các tổng thể lập k mẫu cụ thể và tìm được χ_{qs}^2, so sánh với W_α và kết luận.

Cần nhấn mạnh rằng để sử dụng tiêu chuẩn kiểm định trên thì phải đáp ứng các yêu cầu sau:

- Các mẫu điều tra là ngẫu nhiên và độc lập nhau.

- Kích thước mỗi mẫu tối thiểu là 50.

- Trong mỗi mẫu số lần xuất hiện biến cố ít nhất là 5.

Thí dụ 2. Trước khi đưa ra thị trường một loại sản phẩm với màu sơn mới, người ta muốn xem xét các lứa tuổi khác nhau phản ứng như thế nào với màu sắc sản phẩm đó. Một cuộc điều tra theo những nhóm lứa tuổi khác nhau về màu sắc sản phẩm mà doanh nghiệp sản xuất thu được kết quả sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

	Thích	Không thích	$\sum$
Dưới 25 tuổi	104	21	125
Từ 25 đến 35	122	28	150
Từ 36 đến 45	68	32	100
Từ 46 trở lên	41	34	75
$\sum$	335	115	450

Với mức ý nghĩa $\alpha = 0,05$ có thể cho rằng tỷ lệ những người thích màu sắc đó của sản phẩm là như nhau đối với mọi lứa tuổi hay không?

Giải. Cặp giả thuyết thống kê:

$H_0: p_1 = p_2 = ... = p_k$

H_1: Có ít nhất hai p_i khác nhau

Với $\alpha = 0,05 \rightarrow \chi_{\alpha}^{2(k-1)} = \chi_{0,05}^{2(3)} = 7,815$

Vậy miền bác bỏ là $(7,815;+\infty)$

Để tìm giá trị quan sát của tiêu chuẩn kiểm định ta lập bảng tính sau:

Mẫu	Thích	Không thích	n_i
1	104 0,258	21 0,031	125
2	122 0,296	28 0,045	150
3	68 0,138	32 0,089	100
4	41 0,067	34 0,134	75
m_j	335	115	n = 450

Tại mỗi ô ta tính giá trị $\frac{x_{ij}^2}{n_i m_j}$ và ghi vào góc ô

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lúc đó $\sum_{i=1}^{1}\sum_{j=1}^{2}\frac{X_{ij}^2}{n_i m_j} = 0,258 + ... + 0,134 = 1,058$

Từ đó $\chi_{qs}^2 = 450[1,058 - 1] = 26,1$

$\chi_{qs}^2 \in W_\alpha$. Vậy với mức ý nghĩa 0,05 từ các mẫu đã cho bác bỏ H_0, thừa nhận H_1 tức là tỷ lệ khách hàng thích màu sản phẩm nói trên là khác nhau theo các lứa tuổi.

Việc tìm β, P-value cũng được tiến hành giống như ở các mục trước.

Chẳng hạn P-value $= P[\chi^2 > \chi_{qs}^2] = P[\chi^2 > 26,1]$

Với số bậc tự do bằng 3. Giá trị lớn nhất là 16,27 do đó P-value < 0,001.

Giải bằng Excel cho kết quả sau:

	Col 1	Col 2	Total
Row 1	**104**	**21**	125
Row 2	**122**	**28**	150
Row 3	**68**	**32**	100
Row 4	**41**	**34**	75
Total	335	115	450
	26.38	chi-square	
	3	df	
	7.93E-06	p-value	

Bài tập

9.7. Để đánh giá xem hiệu quả của một chiến dịch quảng cáo có như nhau đối với các địa phương khác nhau hay không, người ta chọn ngẫu nhiên 5 địa phương, từ mỗi địa phương điều tra 100 khách hàng và thống kê số khách hàng mua sản phẩm được quảng cáo ít nhất 1 lần sau chiến dịch quảng cáo đó. Kết quả như sau: địa phương A: 32 người; địa phương B: 59 người; địa phương C: 78 người; địa phương D: 59 người; địa phương E: 10 người. Với mức ý nghĩa $\alpha = 0,05$ hãy kết luận về vấn đề trên.

9.8. Một nhân viên kế toán muốn đánh giá xem trong 4 khách hàng chính của công ty thì tỷ lệ các hóa đơn làm sai có như nhau không. Người đó đã lấy mẫu và thu được kết quả như sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Số khách hàng	Số hóa đơn	Số hóa đơn sai
A	80	21
B	100	25
C	90	30
D	110	40

Hãy tìm P - value để kiểm định về vấn đề trên.

9.9. Một công ty muốn mở rộng việc bán sản phẩm sang 3 thị trường nước ngoài. Để đánh giá xem thị phần mà công ty có thể chiếm lĩnh được tại ba thị trường đó so với các đối thủ cạnh tranh có khác nhau hay không, người ta đã thử nghiệm thị trường bằng cách bán thử sản phẩm cho 150 khách hàng tiềm năng tại mỗi thị trường và thu được kết quả sau:

	Thị trường		
	A	B	C
Công ty	55	38	24
Đối thủ cạnh tranh 1	28	30	21
Đối thủ cạnh tranh 2	20	18	31
Các đối thủ khác	47	64	74

Hãy tìm P - value để kiểm định xem cơ cấu của ba thị trường trên có khác nhau hay không?

9.10. Người ta kiểm tra ngẫu nhiên các sản phẩm do ba máy sản xuất trong một ngày và thu được kết quả sau:

	Số sản phẩm kiểm tra	Số phế phẩm
Máy A	100	12
Máy B	120	16
Máy C	130	18

Hãy kiểm định giả thuyết về cơ cấu phế phẩm của ba máy trên là như nhau với mức ý nghĩa 0,05.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2.3. Kiểm định giả thuyết về quy luật phân phối xác suất của biến ngẫu nhiên

Giả sử chưa biết quy luật phân phối xác suất của biến ngẫu nhiên gốc X trong tổng thể nghiên cứu song có cơ sở để giả thiết rằng X phân phối theo một quy luật A nào đó. Lúc đó đưa ra cặp giả thuyết thống kê sau:

H_0 : X phân phối theo quy luật A

H_1 : X không phân phối theo quy luật A

Để kiểm định cặp giả thuyết trên tiêu chuẩn khi bình phương được sử dụng như sau:

a) Nếu X là biến ngẫu nhiên rời rạc

Từ tổng thể rút ra một mẫu kích thước n trong đó biến ngẫu nhiên X có bảng phân phối tần số thực nghiệm sau đây:

x_i	x_1	x_2	...	x_k	$\sum_{i=1}^{k} n_i = n$
n_i	n_1	n_2	...	n_k	

Nếu giả thuyết H_0 là đúng thì có thể tính được các xác suất lí thuyết để X nhận các giá trị tương ứng:

$$P_i = P(X = x_i) \qquad i = \overline{1,k}$$

Từ đó tần số lí thuyết của phân phối xác suất sẽ là:

$$n'_i = np_i \qquad (i = \overline{1,k})$$

và bảng phân phối tần số lí thuyết có dạng:

x_i	x_1	x_2	...	x_k	$\sum_{i=1}^{k} n'_i = n$
n'_i	n'_1	n'_2	...	n'_k	

lúc đó tiêu chuẩn kiểm định được chọn là thống kê:

$$G = \chi^2 = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i} \qquad (9.7)$$

Đây là dạng tương đương của (9.1)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Biến ngẫu nhiên χ^2 nói trên phân phối "khi bình phương" với (k - r - 1) bậc tự do trong đó r là số tham số cần ước lượng của quy luật cần kiểm định. Các tham số này được ước lượng bằng phương pháp hợp lí tối đa. Chẳng hạn, nếu quy luật cần kiểm định là quy luật nhị thức hay Poisson thì r =1.

Với mức ý nghĩa bằng α cho trước miền bác bỏ W_α được xác định bằng biểu thức:

$$W_\alpha = \left\{ \chi^2 = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}; \chi^2 > \chi_\alpha^{2(k-r-1)} \right\} \qquad (9.8)$$

Với một mẫu cụ thể tính được giá trị quan sát của χ^2 và so sánh với miền bác bỏ W_α để kết luận.

- Nếu $\chi^2_{qs} \in W_\alpha$ thì bác bỏ giả thuyết về dạng phân phối A của biến ngẫu nhiên X.

- Nếu $\chi^2_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết về dạng phân phối A của X.

Thí dụ 3. Số lời gọi đến một trạm điện thoại (X) trong một phút được cho trong bảng sau:

Số lời gọi x_i	0	1	2	3	4	5	≥ 6
Số phút tương ứng n_i	17	22	26	20	11	2	2

Với mức ý nghĩa $\alpha = 0{,}01$ có thể coi X phân phối theo quy luật Poisson được không?

Giải. Cặp giả thuyết thống kê là:

H_0 : X phân phối theo quy luật Poisson

H_1 : X không phân phối theo quy luật Poisson

Từ mẫu trên tính được $\bar{x} = 2$ là ước lượng hợp lí tối đa của λ trong phân phối Poisson. Để tính giá trị quan sát χ^2 ta lập bảng tính toán sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

x_i	n_i	$p_i = \frac{e^{-2}2^{x_i}}{x_i!}$	$n'_i = np_i$	$\frac{(n_i - n'_i)^2}{n'_i}$
0	17	0,1353	13,53	0,89
1	22	0,2707	27,07	0,95
2	26	0,2707	27,07	0,04
3	20	0,1804	18,04	0,21
4	11	0,0902	9,02	0,43
5	2	0,0361	3,61	0,72
≥ 6	2	0,0166	1,66	0,07
	n=100	$\sum = 1.000$		$\chi^2_{qs} = 3,31$

Do $\alpha = 0,01 \rightarrow \chi^{2(k-r-1)}_{\alpha} = \chi^{2(7-1-1)}_{0,01} = 15,1$. Vậy miền bác bỏ là $(15,1;+\infty)$. $\chi^2_{qs} \notin W_{\alpha}$ nên chưa có cơ sở bác bỏ H_0 hay có thể coi X phân phối theo quy luật Poisson với $\lambda = 2$.

Một bài toán kiểm định về quy luật phân phối xác suất rời rạc khá quan trọng trong thực tế là bài toán kiểm định sự phù hợp của quy luật đa thức. Ta sẽ xét chi tiết hơn bài toán này.

Quy luật đa thức là sự mở rộng của quy luật nhị thức khi thỏa mãn các điều kiện sau:

- Tiến hành n phép thử độc lập.

- Trong mỗi phép thử có thể xảy ra k biến cố $A_1, A_2, ..., A_k$ tạo nên một nhóm đầy đủ các biến cố.

- Xác suất xảy ra biến cố $A_i (i = \overline{1,k})$ trong mỗi phép thử đều bằng $p_i (i = \overline{1,k})$. Lúc đó, xác suất để trong n phép thử nói trên biến cố $A_i (i = \overline{1,k})$ xuất hiện tương ứng n_i lần $\left(\sum_{i=1}^{n} n_i = n\right)$ được tính bằng công thức:

$$P_n(n_1, n_2, ..., n_k) = \frac{n!}{n_1!n_2!...n_k!} P_1^{n_1} P_2^{n_2} ... P_k^{n_k} \qquad (9.9)$$

Lúc đó, biến ngẫu nhiên X chỉ số lần xuất hiện các biến cố $A_1, A_2, ..., A_k$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

trong n phép thử nói trên sẽ phân phối theo quy luật đa thức với các tham số là n và $p_1, p_2, ..., p_k$.

Bài toán kiểm định sự phù hợp của quy luật đa thức được phát biểu như sau: Giả sử trong tổng thể nghiên cứu biến ngẫu nhiên X phân phối đa thức song chưa biết các tham số $p_1, p_2, ..., p_k$. Nếu có cơ sở để giả thiết rằng giá trị của chúng tương ứng bằng $p_1^0, p_2^0, ..., p_k^0$ người ta đưa ra cặp giả thuyết:

$H_0: p_i = p_i^0 (i = \overline{1,k})$

H_1: Có ít nhất một xác suất khác với giá trị giả thuyết.

Để kiểm định giả thuyết trên, từ tổng thể rút ra một mẫu kích thước n và giả sử trong mẫu các biến cố $A_1, A_2, ..., A_k$ xuất hiện tương ứng $n_1, n_2, ..., n_k$ lần.

Tiêu chuẩn kiểm định vẫn là thống kê χ^2 trong biểu thức (9.7).

$$G = \chi^2 = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}$$

trong đó n'_i là tần số lí thuyết tương ứng được tính với điều kiện giả thuyết H_0 là đúng, tức là:

$$n'_i = np_i^0$$

Do đó với mức ý nghĩa α miền bác bỏ W_α vẫn được xác định bằng biểu thức (9.8) và thủ tục kiểm định cũng tiến hành như các mục trước.

Thí dụ 4. Một công ty dược phẩm cho rằng loại thuốc cảm cúm mà họ đang bán có sự biến động khá rõ về nhu cầu theo mùa vụ. Họ đánh giá rằng trong tổng khối lượng thuốc bán hàng năm thì 40% được bán vào mùa đông, 40% vào mùa xuân, 10% vào mùa hè và 10% vào mùa thu. Để đánh giá lại điều đó, người ta thống kê 1000 lô thuốc đã tiêu thụ trong năm và thu được kết quả sau:

Mùa	Tần số
Đông	374
Xuân	292
Hè	169
Thu	165

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với mức ý nghĩa 0,05 hãy kiểm định các tỷ lệ tiêu thụ đã nói ở trên.

Giải. Ta có n =1000 và k = 4

Cặp giả thuyết có dạng:

$H_0: p_1 = 0,4; p_2 = 0,4; p_3 = 0,1; p_4 = 0,1$

H_1: Có ít nhất một xác suất khác với giá trị giả thuyết.

Để tìm giá trị quan sát của tiêu chuẩn kiểm định ta lập bảng tính sau:

Mùa	Tần số thực nghiệm n_i	Tần số lí thuyết n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
Đông	374	400	1,69
Xuân	292	400	29,16
Hè	169	100	47,61
Thu	165	100	42,25
$\sum$	n =1000	n =1000	χ^2_{qs} =120,71

Với $\alpha = 0,05 \rightarrow \chi^{2(k-1)}_{\alpha} = \chi^{2(3)}_{0,05} = 7,815$

Vậy miền bác bỏ là: $(7,815; +\infty)$

Do $\chi^2_{qs} \in W_\alpha$ nên bác bỏ H_0 thừa nhận H_1 tức là tỷ lệ tiêu thụ thuốc không đúng như giả định.

Thí dụ 5. Giả sử với số liệu của bài trước ta muốn kiểm định xem có sự khác biệt hay không giữa các mùa về tỷ lệ thuốc tiêu thụ. Lúc đó cặp giả thuyết có dạng:

$H_0: p_1 = p_2 = p_3 = p_4 = 0,25$

H_1: Có ít nhất một xác suất khác 0,25

Ta có bảng tính sau:

Mùa	tần số thực nghiệm n_i	tần số lí thuyết n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
Đông	374	250	61,504
Xuân	292	250	7,056
Hè	169	250	26,244
Thu	165	250	28,9
			χ^2_{qs} =120,704

$\chi^2_{qs} \in W_\alpha$ nên bác bỏ H_0.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b) Nếu X là biến ngẫu nhiên liên tục

Từ tổng thể rút ra một mẫu kích thước n và giả sử các số liệu mẫu được ghép lớp như sau:

$x_{i-1} - x_i$	$x_0 - x_1$	$x_1 - x_2$	...	$x_{k-1} - x_k$
n_i	n_1	n_2	...	n_k

Lúc đó các xác suất lí thuyết p_i chính là xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng $(x_{i-1}; x_i)$ nếu giả thuyết H_0 là đúng:

$$p_i = P(x_{i-1} < X < x_i) \qquad (i = \overline{1,k}) \qquad (9.10)$$

còn các tần số lý thuyết vẫn được tính bằng công thức:

$$n'_i = np_i \qquad (i = \overline{1,k})$$

Từ đó thủ tục kiểm định được tiến hành giống như ở các mục trước.

Sau đây ta sẽ xét trường hợp quan trọng hơn cả trong thực tế là kiểm định giả thuyết về dạng phân phối chuẩn của biến ngẫu nhiên. Trong trường hợp phân phối chuẩn công thức (9.10) có dạng cụ thể sau:

$$p_i = P(x_{i-1} < X < x_i) = \phi_0\left(\frac{x_i - \mu}{\sigma}\right) - \phi_0\left(\frac{x_{i-1} - \mu}{\sigma}\right) \qquad (i = \overline{1,k})$$

thông thường ta chưa biết μ và σ nên chúng được thay bằng các ước lượng hợp lí tối đa tương ứng là $\bar{x}$ và $\sqrt{MS}$. Lúc đó công thức trở thành:

$$p_i \approx \phi_0\left(\frac{x_i - \bar{x}}{\sqrt{MS}}\right) - \phi_0\left(\frac{x_{i-1} - \bar{x}}{\sqrt{MS}}\right)$$

Thí dụ 6. Gặt 200 thửa ruộng của một vùng thu được các số liệu sau:

Năng suất (Tạ/ha)	Số thửa ruộng tương ứng
4 - 6	15
6 - 8	26
8 - 10	25
10 - 12	30
12 - 14	26
14 - 16	21
16 - 18	24
18 - 20	20
20 - 22	13
	n = 200

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với mức ý nghĩa $\alpha = 0{,}05$ có thể coi năng suất lúa của vùng đó phân phối theo quy luật chuẩn được không?

Giải. Cặp giả thuyết thống kê:

H_0 : Năng suất lúa X phân phối chuẩn

H_1 : Năng suất lúa X không phân phối chuẩn

Qua mẫu cụ thể trên, ước lượng hợp lí tối đa của μ là $\bar{x} = 12{,}63$, của σ^2 là ms=22,04 và $\sqrt{ms} = 4{,}695$.

Để tính các xác suất $p_i = \mathrm{P}(x_{i-1} \le X \le x_i)$ ta áp dụng công thức của quy luật chuẩn:

$$p_i = \mathrm{P}(x_{i-1} < X < x_i) = \phi_0\left(\frac{x_i - \mu}{\sigma}\right) - \phi_0\left(\frac{x_{i-1} - \mu}{\sigma}\right)$$

hay

$$p_i \approx \phi_0\left(\frac{x_i - \bar{x}}{\sqrt{ms}}\right) - \phi_0\left(\frac{x_{i-1} - \bar{x}}{\sqrt{ms}}\right)$$

Với khoảng thứ nhất $(x_0 - x_1)$ ta thay bằng $(-\infty; x_1)$ và khoảng cuối cùng $(x_{k-1} - x_k)$ thay bằng $(x_{k-1}; +\infty)$ để hợp của tất cả các khoảng tạo thành toàn bộ trục số. Lập bảng tính toán sau:

$x_{i-1} - x_i$	$U_{i-1} = \frac{x_{i-1} - \bar{x}}{\sqrt{ms}}$	$U_i = \frac{x_i - \bar{x}}{\sqrt{ms}}$	$\phi_0(U_{i-1})$	$\phi_0(U_i)$	$p_i = \phi_0(U_i) - \phi_0(U_{i-1})$	$n'_i = np_i$
$-\infty - 6$	$-\infty$	-1,41	-0,5	-0,4207	0,0793	15,86
6-8	-1,41	-0,99	-0,4207	-0,3389	0,0818	16,36
8-10	-0,99	-0,156	-0,3389	-0,2123	0,1266	25,32
10-12	-0,156	-0,13	-0,2123	-0,0517	0,1606	32,12
12-14	-0,13	0,29	-0,0517	0,1141	0,1658	33,16
14-16	0,29	0,72	0,1141	0,2642	0,1501	30,02
16-18	0,72	1,14	0,2642	0,3729	0,0197	21,74
18-20	1,14	1,57	0,3729	0,4418	0,0689	13,78
20-$+\infty$	1,57	$+\infty$	0,4418	0,5	0,0582	11,6
					$\sum = 1{,}00$	$\sum = 200$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giá trị quan sát $\chi^2_{qs} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i} = 13,32$

Do $\alpha = 0,05 \rightarrow \chi_\alpha^{2(k-r-1)} = \chi_{0,05}^{2(6)} = 12,6$

Vậy miền bác bỏ là $(12,6;+\infty)$. $\chi^2_{qs} \in W_\alpha \Rightarrow$ Bác bỏ H_0 thừa nhận H_1, tức là thừa nhận X không tuân theo quy luật phân phối chuẩn.

Chú ý rằng tiêu chuẩn χ^2 của Pearson chỉ áp dụng được khi kích thước mẫu đủ lớn (n>50) và các tần số tương ứng với mỗi giá trị hay mỗi khoảng giá trị của mẫu cũng phải đủ lớn $(n_i \geq 5; \forall i)$. Vì vậy, nếu các tần số thực nghiệm n_i quá nhỏ (n_i<5) thì phải ghép các giá trị hay các khoảng giá trị của mẫu lại để tăng giá trị của tần số thực nghiệm lên.

Nếu các điều kiện trên không thỏa mãn thì phải áp dụng các phương pháp kiểm định khác mới thu được kết luận đáng tin cậy.

3. MỘT SỐ KIỂM ĐỊNH KHÁC VỀ QUY LUẬT PHÂN PHỐI XÁC SUẤT

Ở mục trước ta đã xét việc sử dụng tiêu chuẩn khi bình phương để kiểm định giả thuyết về dạng phân phối xác suất của biến ngẫu nhiên. Sau đây sẽ trình bày thêm một số thủ tục kiểm định khác về dạng phân phối xác suất đặc biệt là liên quan đến dạng phân phối chuẩn của tổng thể nghiên cứu.

3.1. Tiêu chuẩn phù hợp của Kolmogorov

Tiêu chuẩn Kolmogorov áp dụng được đối với mọi phân phối liên tục song các tham số của quy luật phân phối lý thuyết trong tổng thể giả định đã biết trước chứ không xác định dựa vào các số liệu mẫu. Giả sử biến ngẫu nhiên X trong tổng thể giả thiết phân phối theo một quy luật lý thuyết liên tục nào đó với hàm phân bố xác suất $F(x)$. Từ tổng thể lập mẫu kích thước n và tìm được hàm phân bố thực nghiệm $F^*(x_i)$ tại mỗi giá trị x_i của mẫu $(i = \overline{1,n})$. Lúc đó tiêu chuẩn kiểm định giả thuyết được chọn là thống kê:

$$D = \max_{x_i} \left| F^*(x_i) - F(x_i) \right| \qquad (9.11)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Khi $n \to \infty$ thì không tùy thuộc vào dạng của F(x) biến ngẫu nhiên $\lambda = \sqrt{n}D$ luôn luôn hội tụ về phân phối Kolmogorov có hàm phân bố xác suất như sau:

$$K(x) = \begin{cases} \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2x^2} & \text{với } x > 0 \\ 0 & x \le 0 \end{cases}$$

do đó với mức ý nghĩa α cho trước có thể tìm được giá trị tới hạn thỏa mãn điều kiện:

$$P(\lambda > \lambda_\alpha) = \alpha$$

Các giá trị của λ_α được tính sẵn thành bảng (Phụ lục 14). Như vậy, miền bác bỏ giả thuyết được xác định bằng biểu thức:

$$W_\alpha = \left\{ \lambda = \sqrt{n} \max_{x_i} \left| F^*(x_i) - F(x_i) \right| ; \lambda > \lambda_\alpha \right\} \qquad (9.12)$$

Thủ tục kiểm định cũng được tiến hành như đã làm ở các mục trước.

Thí dụ 1. Kết quả đo kích thước của 1000 chi tiết được cho dưới dạng bảng phân phối thực nghiệm ghép lớp như sau: (x_i là giá trị giữa lớp).

i	x_i	n_i	i	x_i	n_i
1	98,0	21	6	100,5	201
2	98,5	47	7	101,0	142
3	99,0	87	8	101,5	97
4	99,5	158	9	102,0	41
5	100,0	181	10	102,5	25

Dùng tiêu chuẩn Kolmogorov hãy kiểm định giả thuyết cho rằng kích thước chi tiết là biến ngẫu nhiên phân phối chuẩn với kỳ vọng toán $\mu = 100{,}25$ và độ lệch chuẩn $\sigma = 1$. Mức ý nghĩa được chọn là 0,05.

Giải. Hàm phân bố xác suất lý thuyết của phân phối chuẩn trong trường hợp này có thể viết dưới dạng:

$$F(x) = \frac{1}{2} + \phi_0(x - \mu)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Còn hàm phân bố thực nghiệm $F^*(x_i)$ có thể tính theo công thức:

$$F^*(x_i) = \frac{1}{1000}\left[\sum_{j=1}^{i-1} n_j + 0{,}5n_j\right].$$

Với mỗi giá trị x_i ta tìm hiệu $F^*(x) - F(x)$ và tìm giá trị lớn nhất về giá trị tuyệt đối. Quá trình tính toán được cho trong bảng sau:

i	$x_i - \mu$	$\phi_0(x_i - \mu)$	$F(x_i)$	$F^*(x_i)$	$F^*(x_i) - F(x_i)$
1	-2,25	-0,4877	0,0123	0,0105	0,0018
2	-1,75	-0,4599	0,0401	0,0445	0,0044
3	-1,25	-0,3944	0,1056	0,1115	0,0059
4	-0,75	-0,2734	0,2266	0,2340	0,0074
5	-0,25	-0,0987	0,4013	0,4035	0,0022
6	0,25	0,0987	0,5987	0,5945	0,0042
7	0,75	0,2734	0,7734	0,7660	0,0074
8	1,25	0,3944	0,8944	0,8855	0,0089
9	1,75	0,4599	0,9599	0,9545	0,0054
10	2,25	0,4877	0,9877	0,9875	0,0002

Từ bảng tính toán tìm được:

$$D_{qs} = 0{,}0089$$

do đó $\lambda_{qs} = \sqrt{n}D_{qs} = \sqrt{1000}.0{,}0089 = 0{,}281$

Với $\alpha = 0{,}05$ giá trị tới hạn $\lambda_\alpha = \lambda_{0,05} = 1{,}358$ vậy miền bác bỏ giả thuyết là $(1{,}358;+\infty)$. $\lambda_{qs} \notin W_\alpha$ do đó với mức ý nghĩa 0,05 chưa có cơ sở bác bỏ giả thuyết về dạng phân phối chuẩn của kích thước chi tiết.

Chú ý rằng để có thể áp dụng tiêu chuẩn phù hợp của Kolmogorov thì kích thước mẫu cũng phải đủ lớn $(n \geq 40)$.

Một biến tướng của tiêu chuẩn kiểm định trên là tiêu chuẩn Kolmogorov - Smirnov. Tiêu chuẩn này được sử dụng để kiểm định xem hai mẫu kích thước n_1 và n_2 có cùng được rút ra từ một tổng thể nghiên cứu hay không. Để kiểm định điều đó người ta sử dụng thống kê:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$D_{n_1,n_2} = \max_x \left| F_1^*(x) - F_2^*(x) \right| \qquad (9.13)$$

Trong đó: $F_1^*(x)$ và $F_2^*(x)$ là các hàm phân bố thực nghiệm của hai mẫu tương ứng. Từ đó tiêu chuẩn kiểm định được chọn là:

$$\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1,n_2} \qquad (9.14)$$

và với mức ý nghĩa α miền bác bỏ giả thuyết được xác định bằng biểu thức:

$$W_\alpha = \left\{ \lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \max_x \left| F_1^*(x) - F_2^*(x) \right| ; \lambda > \lambda_\alpha \right\} \qquad (9.15)$$

Thủ tục kiểm định cũng được tiến hành giống như trên.

Thí dụ 2. Có hai nhóm chi tiết cùng loại do hai máy sản xuất, mỗi nhóm gồm 60 chi tiết được đem đo và thu được kết quả cho trong bảng sau (Bảng 9.3)

Bảng 9.3

x_i	Số giá trị $x \le x_i$		$F_1^*(x)$	$F_2^*(x)$	$\left\|F_1^*(x) - F_2^*(x)\right\|$
	Nhóm I	Nhóm II			
71,95	1	0	0,0167	0	0,0167
72,11	1	1	0,0167	0,0167	0,0000
72,12	1	2	0,0167	0,0333	0,0167
72,14	3	2	0,0500	0,0333	0,0167
...	...	...	...	...	...
72,53	43	50	0,7167	0,8333	0,1167
72,54	46	52	0,7667	0,8667	0,1000
72,55	46	56	0,7667	0,9333	0,1667
72,56	49	57	0,8167	0,9500	0,1333
72,58	51	57	0,8500	0,9500	0,1000
72,60	52	58	0,8667	0,9667	0,1000
...	...	...	...	...	...
72,69	57	60	0,9500	1,0000	0,0500
72,70	58	60	09,667	1,0000	0,0333
72,72	59	60	0,9833	1,0000	0,0167
72,73	60	60	1,000	1,0000	0,0000

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hãy dùng tiêu chuẩn Kolmogorov - Smirnov để kiểm định với mức ý nghĩa 0,1 xem hai mẫu trên có thuộc cùng một tổng thể hay không, tức là hai máy có sản xuất các chi tiết với kích thước theo cùng một quy luật phân phối xác suất hay không.

Giải. Sắp xếp kích thước của chi tiết thuộc cả hai nhóm theo thứ tự tăng dần và tìm các hàm phân bố thực nghiệm $F_1^*(x)$ và $F_2^*(x)$ từ đó xác định hiệu của chúng về giá trị tuyệt đối (xem bảng).

Từ đó: $D_{n_1,n_2} = 0{,}1667$

và $$\lambda_{qs} = \sqrt{\frac{60.60}{60+60}}0{,}1667 = 0{,}9130$$

với $\alpha = 0{,}1 \to \lambda_\alpha = 1{,}224$ vậy miền bác bỏ giả thuyết là $(1{,}224;+\infty)$. Do $\lambda_{qs} \notin W_\alpha$ nên chưa có có cơ sở để bác bỏ giả thuyết H_0. Nói cách khác phân phối xác suất của kích thước các chi tiết do hai máy đó sản xuất có thể là như nhau.

3.2. Kiểm định Lilliefors về dạng phân phối chuẩn

Khi sử dụng tiêu chuẩn phù hợp của Kolmogorov ta giả thiết rằng kỳ vọng toán và phương sai của phân phối lý thuyết được xác định bên ngoài mẫu và đã biết. Nếu điều kiện này không thỏa mãn thì việc kiểm định sẽ kém chính xác.

H.W. Lilliefors đã áp dụng kiểm định Kolmogorov vào trường hợp phân phối chuẩn khi kỳ vọng toán và phương sai của tổng thể chưa biết. Lúc đó sẽ dùng trung bình mẫu và phương sai mẫu để thay thế cho kỳ vọng toán và phương sai tổng thể khi tính giá trị của hàm phân bố xác suất lý thuyết. Còn tiêu chuẩn kiểm định cũng vẫn là cực đại của sai lệch giữa hàm phân bố thực nghiệm và lý thuyết tại mỗi giá trị của mẫu. Sau đó giá trị này được so sánh trực tiếp với giá trị tới hạn L_α được cho trong phụ lục 15 tương ứng với mức ý nghĩa α và kích thước mẫu n để kết luận.

Thí dụ 3. Theo dõi thời gian thực hiện một hóa đơn cho khách hàng (đơn vị tính là 10 giây) thu được kết quả sau: 51, 50, 45, 53, 46, 49, 47. Với mức ý nghĩa 0,05 có thể cho rằng thời gian thực hiện xong một hóa đơn phân phối chuẩn hay không.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải. Sắp xếp các giá trị của X theo trình tự tăng dần và tiến hành tính toán ta có bảng sau (Bảng 9.4).

Bảng 9.4

i	x_i	x_i^2	$u_i=\frac{x_i-\bar{x}}{s}$	$F^*(x_i)=\frac{i}{7}$	$F(x_i)=P(x<x_i)$	$\lvert F^*(x_i)-F(x_i)\rvert$
1	45	2025	-1,40	0,1429	0,0808	0,0621
2	46	2116	-1,02	0,2857	0,1539	0,1318
3	47	2209	-0,65	0,4286	0,2578	0,1708
4	49	2401	0,11	0,5714	0,5438	0,0276
5	50	2500	0,48	0,7143	0,6844	0,0299
6	51	2601	0,86	0,8571	0,8051	0,0520
7	53	2809	1,61	1,0000	1,0000	0,0000

$\bar{x} = 48,7143$

$s = 2,6573$

Từ đó:

$$D_{qs} = \max_{x_i}\left|F^*_{(x_i)} - F_{(x_i)}\right| = 0,1708$$

với $n=7; \alpha = 0,05$ từ phụ lục 15 tìm được $L_\alpha = L_{0,05} = 0,3$. Vậy miền bác bỏ là $(0,3;+\infty)$. Do $D_{qs} \notin W_\alpha$ nên chưa có cơ sở bác bỏ H_0 hay có thể cho rằng thời gian thực hiện một hóa đơn phân phối chuẩn.

Các kiểm định trên đều có thể thực hiện bằng phần mềm Stata

.ksmirnov x3=normprob ((x3-1628.12)/48.3997)

One-Sample Kolmogorov - Smirnov test against theortical distribution

normprob ((x3-1628.12)/48.3997)

Smaller group	D	P-value	Corrected
x3:	0.0698	0.377	
Cumulative	-0.0603	0.483	
Combined K-S	0.0698	0.714	0.670

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.3. Kiểm định Jarque - Bera về dạng phân phối chuẩn

Một kiểm định khác về dạng phân phối chuẩn, đặc biệt hay được dùng trong phân tích hồi quy để kiểm định dạng phân phối chuẩn của các phần dư là kiểm định Jarque - Bera. Trong kiểm định này người ta sử dụng hệ số bất đối xứng a_3 của mẫu và hệ số nhọn a_4 của mẫu để kiểm định.

Tiêu chuẩn kiểm định được chọn là thống kê:

$$JB = n\left[\frac{a_3^2}{6} + \frac{(a_4 - 3)^2}{24}\right] \qquad (9.16)$$

Nếu giả thuyết H_0 về dạng phân phối chuẩn đúng thì với kích thước mẫu đủ lớn thống kê JB sẽ phân phối xấp xỉ khi bình phương với 2 bậc tự do. Vì vậy với mức ý nghĩa α miền bác bỏ được xác định bằng biểu thức:

$$W_\alpha = \left\{JB = n\left[\frac{a_3^2}{6} + \frac{(a_4 - 3)^2}{24}; JB > \chi_\alpha^{2(2)}\right]\right\} \qquad (9.17)$$

Thủ tục kiểm định cũng tiến hành giống như ở các mục trước.

Thí dụ 4. Với các số liệu đã cho trong thí dụ A, với mức ý nghĩa 0,05 hãy kiểm định xem thu nhập hàng năm của dân cư vùng 3 có phân phối chuẩn hay không.

Giải. Từ các số liệu của vùng 3 ta tìm được:

$$\bar{x} = 1628{,}12; \qquad s^2 = 2342{,}531$$

$$a_3 = 0{,}0724; \qquad a_4 = 2{,}777$$

Từ đó:

$$JB_{qs} = 100\left[\frac{0{,}0724^2}{6} + \frac{(2{,}777 - 3)^2}{24}\right] = 0{,}29$$

Với $\alpha = 0{,}05$ tra bảng được $\chi_{0{,}05}^{2(2)} = 5{,}991$. Vậy miền bác bỏ là $(5{,}991; +\infty)$

Do $JB_{qs} \notin W_\alpha$ nên chưa thể bác bỏ giả thiết là thu nhập của dân cư vùng 3 phân phối chuẩn.

Kết quả giải bằng Stata như sau:

Skewness/Kurtosis tests for Normality

		--------joint--------		
Variable	Pr (Skewness)	Pr (Kurosis)	adj chi-sq(2)	Pr (chi-sq)
x3	0.753	0.825	0.15	0.9289

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

9.11. Phân phối thực nghiệm của 4000 gia đình có 3 con theo số con trai được cho ở bảng sau:

Số con trai	Số gia đình tương ứng
0	450
1	1460
2	1530
3	560

Với mức ý nghĩa $\alpha = 0{,}05$ có thể xem X là số con trai trong gia đình có 3 con phân phối theo quy luật nhị thức được không?

ĐS: H_0 bị bác bỏ, hay X không tuân theo quy luật nhị thức.

9.12. Số con của 2000 phụ nữ thủ đô dưới 25 tuổi cho bảng sau:

X: Số con	0	1	2	3	4
Số phụ nữ	1090	650	220	30	10

Với mức ý nghĩa $\alpha = 0{,}05$ có thể xem X tuân theo quy luật Poisson được không?

ĐS: $\chi^2_{qs} = 0{,}318$; $W\alpha = (6{,}0; +\infty)$. Vậy $\chi^2_{qs} \notin W\alpha$. Số con của phụ nữ dưới 25 tuổi theo quy luật Poisson.

9.13. Một công ty dược phẩm khẳng định loại thuốc A mà họ được đưa ra thị trường đạt kết quả điều trị rất tốt đối với 60% bệnh nhân, kết quả vừa phải đối với 30% bệnh nhân và chỉ có 10% bệnh nhân là không có kết quả. Một ngẫu nhiên gồm 85 bệnh nhân điều trị bằng loại thuốc trên thu được kết quả sau:

Kết quả rất tốt: 30

Kết quả vừa phải: 35

Không có kết quả: 20

Với mức ý nghĩa 0,05 hãy kết luận về lời tuyên bố của công ty dược nói trên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.14. Một báo cáo về tình hình kinh doanh cho biết bình quân 72,9% hàng tồn kho là của các công ty nhà nước, 23% là của các công ty tư nhân và 4,1% là của các công ty liên doanh với nước ngoài. Giả sử vào thời điểm nghiên cứu trong số 1000 sản phẩm tồn kho thì có 610 sản phẩm là của các công ty nhà nước, 290 là của các công ty tư nhân và 100 là của các công ty liên doanh. Với mức ý nghĩa 0,05 hãy kết luận về báo cáo trên.

9.15. Tại một thành phố, người ta thống kê số tai nạn giao thông trong một tháng theo giờ như sau:

Từ 8 giờ đến 10 giờ: 31 vụ;

Từ 10 giờ đến 12 giờ: 30 vụ;

Từ 13 giờ đến 15 giờ: 41 vụ;

Từ 15 giờ đến 17 giờ: 58 vụ;

a. Tìm P-value để kiểm định xem tỷ lệ tai nạn giao thông vào những giờ khác nhau trong ngày có khác nhau không?

b. Với mức ý nghĩa 0,05 thì có bác bỏ giả thuyết H_o hay không?

ĐS: $\chi^2_{qs} = 12,65$ => P-value < 0,01

9.16. Kết quả đo kích thước của 1000 chi tiết cho trong bảng dưới đây.

Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận kích thước chi tiết sản xuất phân phối theo quy luật chuẩn hay không:

Kích thước chi tiết (mm)	Số chi tiết tương ứng
97,75 - 98,25	21
98,25 - 98,75	47
98,75 - 99,25	87
99,25 - 99,75	158
99,75 - 100,25	181
100,25 - 100,75	201
100,75 - 101,25	142
101,25 - 101,75	97
101,75 - 102,25	41
102,25 - 102,75	25

ĐS: Chưa có cơ sở bác bỏ giả thiết về dạng phân chuẩn của kích thước chi tiết.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.17. Các số 0, 1, 2...., 9 trong 8000 số thập phân đầu tiên của số π xuất hiện tương ứng 74, 92, 83, 79, 80, 73, 77, 75, 76 và 91 lần. Hãy dùng tiêu chuẩn Kolmogorov để kiểm định giả thuyết về dạng phân phối đều của các số đó với mức ý nghĩa 0,1 nếu giả thiết rằng xác suất xuất hiện bất kỳ chữ số nào ở bất kỳ chữ số thập phân nào cũng đều bằng 0,1.

ĐS: $D = 0,0138$; $\lambda_{qs} = 0,3903$

9.18. Từ bảng số ngẫu nhiên lấy ra 150 số có 2 chữ số và thu được kết quả sau:

Khoảng giá trị x_{i-1} - x_i	Tần số (n_i)	Tần suất (f_i)
0-9	16	0,107
10-19	15	0,100
20-29	19	0,127
30-39	13	0,087
40-49	14	0,093
50-59	19	0,127
60-69	14	0,093
70-79	11	0,073
80-89	13	0,087
90-99	16	0,107

Với mức ý nghĩa 0,05 hãy dùng tiêu chuẩn Kolmogorov để kiểm định sự phù hợp của các giá trị này với dạng phân phối đều.

ĐS: $D = 0,041$; $\lambda_{qs} = 0,5021$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.19. Từ các sản phẩm của hai máy A và B lấy ngẫu nhiên 2 mẫu kích thước đều bằng 50 chi tiết đem đo và thu được kết quả sau:

Thứ tự	Máy A	Máy B	Thứ tự	Máy A	Máy B
1	51,58	51,50	8	54	55
2	35	36	9	24	33
3	33	69	10	36	56
4	54	60	11	58	68
5	24	54	12	70	39
6	42	42	13	47	42
7	47	54	14	50	15
8	26	48	33	34	55
9	47	46	34	36	51
10	05	42	35	56	55
11	18	39	36	56	10
12	66	46	37	42	44
13	35	42	38	29	19
14	50	39	39	31	24
15	50	16	40	30	31
16	54	51	41	12	51
17	48	50	42	28	46
18	36	50	43	51	39
19	50	48	44	39	39
20	42	53	45	15	30
21	56	25	46	42	30
22	56	48	47	36	42
23	48	36	48	28	55
24	42	53	49	30	44
25	56	23	50	48	24

Với mức ý nghĩa 0,2 hãy dùng tiêu chuẩn Kolmogorov - Smirnov để kiểm định giả thuyết là cả hai mẫu đều thuộc cùng một tổng thể, tức là phân phối xác suất của kích thước chi tiết do hai máy sản xuất là như nhau.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.20. Với các số liệu của bài 9.19, hãy dùng tiêu chuẩn Lilliefors để kiểm định dạng phân phối chuẩn của kích thước do máy A sản xuất với mức ý nghĩa 0,2.

9.21. Với các số liệu của bài 9.19, hãy dùng tiêu chuẩn Jarque - Bera để kiểm định dạng phân phối chuẩn của kích thước chi tiết cho máy B sản xuất với mức ý nghĩa 0,05.

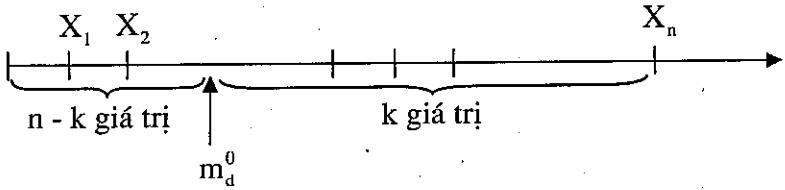
4. KIỂM ĐỊNH THEO DẤU

Kiểm định theo dấu là một loại kiểm định phi tham số được dùng trong các quyết định kinh doanh. Nó thường được sử dụng để kiểm định giả thuyết về giá trị trung vị của tổng thể vì như đã trình bày ở mục 2.8 chương VII, trong nhiều trường hợp thực tế trung vị lại tỏ ra hữu dụng hơn trung bình. Sau đây ta xét hai ứng dụng của kiểm định theo dấu với giá trị trung vị.

4.1. Kiểm định về một trung vị của tổng thể

Giả sử trong tổng thể nghiên cứu, giá trị của trung vị của biến ngẫu nhiên chưa biết song có cơ sở để giả thiết rằng nó bằng m_d^0. Lúc đó, ta đưa ra giả thuyết thống kê: $H_0 : m_d = m_d^0$.

Để kiểm định giả thuyết trên, từ tổng thể lập mẫu ngẫu nhiên kích thước n và sắp xếp các giá trị mẫu theo thứ tự tăng dần $X_1, X_2, ..., X_n$. Giả sử trên trục số các giá trị mẫu có phân phối như sau:



và như trên hình, giả sử có k giá trị mẫu lớn hơn hoặc bằng giá trị giả thuyết m_d^0. Nếu ta gán cho các giá trị lớn hơn hoặc bằng m_d^0 dấu +, còn các giá trị nhỏ hơn m_d^0 dấu - thì k chính là số giá trị mang dấu + trên mẫu.

Như đã phân tích ở mục 2.8 Chương VII, xác suất để mỗi giá trị của mẫu lớn hơn hoặc bằng m_d đều bằng 0,5 do đó nếu giả thuyết H_0 là đúng thì k là biến ngẫu nhiên phân phối nhị thức. Vì vậy, có thể dùng các phương thức sau để kiểm định giả thuyết H_0 :

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

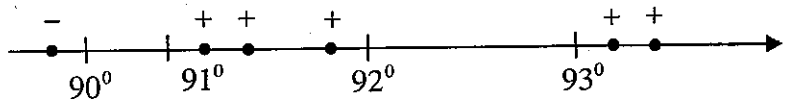
- Do khoảng tin cậy cũng có thể dùng để kiểm định giả thuyết H_0 (xem mục 3.1 Chương VIII) do đó xuất phát từ trung vị x_d của mẫu có thể xây dựng khoảng tin cậy mức $1-\alpha$. Từ đó bác bỏ H_0 với mức ý nghĩa α nếu m_d^0 rơi ra ngoài khoảng tin cậy.

- Dùng công thức Bernoulli để tìm P-value theo công thức

$$P-value = P(x \geq k)$$

và dựa vào P-value để kết luận.

Thí dụ 1. Tia sáng của máy phân cực chiếu qua loại đường lactoza α thì góc quay của kim đồng hồ là 90^0. Một kỹ sư hóa công nghiệp lấy ngẫu nhiên 6 mẫu đường từ một loại đường mới nhập về đem đo và thu được kết quả mô tả trên trục số như sau:



Tìm P-value để kiểm định giả thuyết là loại đường nhập về đúng là đường lactoza α.

Giải. Theo bảng phân phối nhị thức với n = 6; p = 0,5 tìm được:

$$P-value = P(X \geq 5) = p_6(5) + p_6(6) = 0{,}11 = 11\%$$

Vậy nếu lấy mức ý nghĩa là 0,05 thì chưa thể bác bỏ H_0 hay có thể thừa nhận loại đường nhập về đúng là lactoza α.

Kết quả giải bằng Excel như sau:

Sign Test	
90	hypothesized value
91.5	median
1	below
0	equal
5	above
6	n
	binomial
.1094	p-value (one-tailed, upper)
	normal approximation
1.22	z
.1103	p-value (one-tailed, upper)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.2. Kiểm định về hai trung vị của hai tổng thể khi hai mẫu gồm các giá trị theo cặp

Khi kiểm định sự bằng nhau của hai trung vị trong trường hợp hai mẫu rút ra có các giá trị tương ứng theo từng cặp thì có thể đưa bài toán về trường hợp đã xét ở trên bằng cách thiết lập sự sai lệch của từng cặp giá trị:

$$D_i = X_{1i} - X_{2i} \qquad i = \overline{1,n}$$

Lúc đó coi như ta chỉ có một mẫu gồm các giá trị $D_1, D_2, ..., D_n$ và việc kiểm định sự bằng nhau của hai trung vị tương đương với kiểm định trung vị của các sai lệch bằng không: $H_0 : m_d = 0$.

Lúc đó xác suất để với mỗi giá trị của mẫu các sai lệch nhận giá trị dương hoặc âm bằng 0,5 và cũng như trên, xác suất để số quan sát X của mẫu nhận giá trị dương sẽ phân phối nhị thức với các tham số là n (kích thước mẫu) và p=0,5. Từ đó tìm được P-value và kết luận.

Thí dụ 2. Tám người tình nguyện kiểm tra khả năng hít thở của phổi trước và sau khi áp dụng một phương pháp mới để điều trị bệnh hen. Kết quả thu được như sau:

X_1 (trước điều trị)	X_2 (sau điều trị)	$D_i = X_{1i} - X_{2i}$
750	850	+100
860	880	+20
950	930	- 20
830	860	+30
750	800	+50
680	740	+60
720	760	+40
810	800	-10

Hãy tìm P-value để kiểm định giả thuyết cho rằng phương pháp điều trị này không mang lại hiệu quả.

Giải. Phương pháp điều trị không có hiệu quả tương đương với giả thuyết trung vị của các sai lệch bằng không $H_0 : m_d = 0$

Theo bảng phân phối nhị thức với n = 8; p = 0,5 tìm được

$$\text{P-value} = P(d \geq 6) = 0{,}145 = 14{,}5\%$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì P-value khá lớn nên không thể bác bỏ H_0 ở mức ý nghĩa 0,05, thậm chí ở mức 0,1. Sau đây là kết quả giải bằng Excel:

Sign Test	
0	hypothesized value
35	median
2	below
0	equal
6	above
8	n
	binomial
.2891	p-value (two-tailed)
	normal approximation
1.06	z
.2888	p-value (two-tailed)

Bài tập

9.22. Với các số liệu của bài 7.59 hãy tìm P-value để kiểm định giả thuyết là trung vị của khả năng chịu lực của chốt cửa do nhà máy sản xuất bằng 9,75.

ĐS: P-value = 0,1316.

9.23. Với các số liệu của bài 7.60, hãy tìm P-value để kiểm định các giả thuyết sau:

a. Trung vị của thu nhập của chồng là 30 triệu.

b. Trung vị của thu nhập của vợ là 30 triệu.

c. Thu nhập của chồng và của vợ là như nhau.

9.24. Với các số liệu của bài 7.62, hãy tìm P-value để kiểm định các giả thuyết sau:

a. Trung vị của chiều cao của con trai là 166 cm.

b. Trung vị của chiều cao của con gái là 163 cm.

c. Chiều cao của con trai và con gái là như nhau.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.25. Để đánh giá xem sau khi tách công ty ra thành hai công ty nhỏ thì lãi suất cổ phiếu có thay đổi hay không, người ta đã lấy ngẫu nhiên 10 cổ phiếu của các công ty mới tách ra, thống kê lãi suất trước và sau khi tách ra và thu được kết quả sau:

Trước khi tách (%)	0,5	-0,2	0,9	1,1	-0,7
	1,5	2,0	1,3	1,6	2,1
Sau khi tách (%)	1,1	0,3	1,2	1,9	-0,2
	1,4	1,8	1,8	2,4	2,2

Với mức ý nghĩa 0,05 hãy kiểm định xem lãi suất cổ phiếu trước và sau khi tách công ty có khác nhau không?

9.26. Kiểm tra ngẫu nhiên 20 bóng đèn thu được tuổi thọ của chúng như sau:

738	752	710	701	689
779	650	541	902	700
488	555	870	609	745
712	881	599	659	793

Hãy tìm P-value để kiểm định giả thuyết là trung vị của tuổi thọ bóng đèn bằng 750 giờ.

5. KIỂM ĐỊNH TỔNG HẠNG WILCOXON

Như đã trình bày ở mục 3 chương 8, để kiểm định sự bằng nhau của hai kỳ vọng toán của 2 biến ngẫu nhiên trong tổng thể nghiên cứu thì ta luôn giả thiết rằng các dấu hiệu nghiên cứu trong tổng thể phân phối chuẩn. Nếu chúng không phân phối chuẩn thì các mẫu điều tra phải đủ lớn để có thể áp dụng định lý giới hạn trung tâm. Song nếu các kích thước mẫu điều tra không đủ lớn thì không thể áp dụng các phương pháp kiểm định đã xét. Ở phần này ta đưa ra giải pháp cho trường hợp này thông qua kiểm định phi tham số bằng tổng các hạng Wilcoxon (còn gọi là kiểm định Mann-Whitney). Nó dựa trên các giả thiết sau:

- Các biến ngẫu nhiên X_1 và X_2 trong hai tổng thể nghiên cứu có thể phân phối theo một quy luật bất kì không nhất thiết là quy luật chuẩn.

- Các mẫu ngẫu nhiên rút ra từ hai tổng thể phải độc lập nhau song có thể có kích thước tùy ý.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lúc đó, việc kiểm định sự bằng nhau của hai kỳ vọng toán tương đương với việc kiểm định giả thuyết:

H_0 :Hai tổng thể có phân phối giống nhau.

Với các giả thuyết đối tương ứng

H_1: Tổng thể thứ nhất có phân phối lệch sang phải so với tổng thể thứ hai.

H_1: Tổng thể thứ nhất có phân phối lệch sang trái so với tổng thể thứ hai.

H_1: Hai tổng thể 1 và 2 có các tham số đặc trưng vị trí hoàn toàn khác nhau

Nếu giả thuyết H_0 là đúng thì hai mẫu độc lập được rút ra từ hai tổng thể đó sẽ tương đồng với nhau. Để đo mức độ tương đồng này có thể sử dụng các hạng kết hợp (từ nhỏ đến lớn) của các giá trị mẫu kết hợp lại từ hai mẫu điều tra và kiểm tra tổng các hạng của các giá trị của mẫu thứ nhất (hoặc của mẫu thứ hai).

(Nhắc lại rằng xếp hạng các giá trị của một biến là việc gán số thứ tự cho các giá trị đó theo trình tự tăng dần của chúng).

Nếu giả thuyết H_0 là đúng, tức là hai tổng thể là như nhau thì tổng các hạng của mẫu sẽ tỷ lệ thuận với kích thước mẫu. Nếu ký hiệu T là tổng các hạng của mẫu thứ nhất thì về trực quan có thể thấy nếu T quá nhỏ hoặc quá lớn sẽ là chứng cứ để ta bác bỏ H_0.

Nếu H_0 đúng thì thống kê T sẽ có phân phối mẫu với kỳ vọng toán và phương sai như sau:

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \qquad (9.18)$$

$$\sigma_T^2 = \frac{n_1 n_2}{12}(n_1 + n_2 + 1) \qquad (9.19)$$

Hơn nữa nếu cả n_1 và n_2 đều lớn hơn 10 thì T sẽ phân phối xấp xỉ chuẩn.

Kiểm định tổng các hạng Wilcoxon giả thiết rằng biến X trong tổng thể là liên tục, như vậy sẽ không có bất kỳ hai giá trị nào bằng nhau. Trong

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

thực tế, khi điều tra mẫu có thể gặp trường hợp hai hoặc nhiều giá trị mẫu bằng nhau. Lúc đó mỗi giá trị của nhóm giá trị bằng nhau sẽ được gán cho hạng bằng trung bình số học của các hạng trong nhóm. Chẳng hạn nếu hai giá trị bằng nhau nằm ở hạng thứ ba và thứ tư thì mỗi giá trị được gán hạng bằng 3,5 và hạng của giá trị tiếp theo sẽ là 5. Khi có các giá trị mẫu bằng nhau thì công thức của phương sai là:

$$\sigma_T^2 = \frac{n_1 n_2}{12}\left[n_1 + n_2 + 1 - \frac{\sum_j t_j(t_j^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}\right] \qquad (9.20)$$

Trong đó: t_j là tần số của các hạng ghép nhóm trong nhóm thứ j. Chú ý rằng nếu mọi giá trị của mẫu đều khác nhau thì từ (9.20) ta lại thu được (9.19).

Thủ tục kiểm định tổng các hạng của Wilcoxon bao gồm hai trường hợp sau:

α) Nếu $n_1 \le 10$ *và* $n_2 \le 10$

Lúc đó thống kê T là tổng các hạng của mẫu thứ nhất. Với mức ý nghĩa α cho trước dùng phụ lục 11 để tìm các giá trị tới hạn T_L và T_u và tùy thuộc vào giả thuyết đối H_1 các miền bác bỏ được xác định như sau:

a) H_0: Tổng thể 1 và 2 có phân phối giống nhau

H_1: Tổng thể 1 có phân phối lệch sang phải so với tổng thể thứ hai

$$W_\alpha = \{T; T > T_u\} \qquad (9.21)$$

b) H_0: Tổng thể 1 và 2 có phân phối giống nhau

H_1: Tổng thể 1 có phân phối lệch sang trái so với tổng thể thứ hai

$$W_\alpha = \{T; T < T_L\} \qquad (9.22)$$

c) H_0: Tổng thể 1 và 2 có phân phối giống nhau

H_1: Hai tổng thể có tham số đặc trưng vị trí hoàn toàn khác nhau.

$$W_\alpha = \{T; T < T_L \text{ hoặc } T > T_u\} \qquad (9.23)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

β) Nếu $n_1 > 10$ *và* $n_2 > 10$

Lúc đó tiêu chuẩn kiểm định có dạng:

$$G = U = \frac{T - \mu_T}{\sigma_T} \tag{9.24}$$

và W_α tương ứng với 3 cặp giả thuyết trên là:

$$\text{a. } W_\alpha = \left\{ U = \frac{T - \mu_T}{\sigma_T}; U > u_\alpha \right\} \tag{9.25}$$

$$\text{b. } W_\alpha = \left\{ U = \frac{T - \mu_T}{\sigma_T}; U < -u_\alpha \right\} \tag{9.26}$$

$$\text{c. } W_\alpha = \left\{ U = \frac{T - \mu_T}{\sigma_T}; |U| > u_{\alpha/2} \right\} \tag{9.27}$$

Như vậy, kiểm định tổng các hạng của Wilcoxon gián tiếp phản ánh quan hệ của μ_1 và μ_2.

Nếu phân phối của tổng thể 1 lệch sang phải so với tổng thể 2 thì có cơ sở để kết luận rằng $\mu_1 > \mu_2$. Các kết luận tương tự có thể đưa ra với hai trường hợp còn lại.

Thí dụ. Các kỹ sư môi trường quan tâm đến việc là dự án nạo vét một cái hồ trong thành phố có thực sự mang lại hiệu quả hay không. Để làm điều đó, trước và sau khi nạo vét người ta đã lấy 12 mẫu nước mỗi lần và đo lượng oxygen không tan (tính bằng ppm) trong các mẫu nước và thu được kết quả sau:

Trước khi nạo vét		Sau khi nạo vét	
11	11,6	10,2	10,8
11,2	11,7	10,3	10,8
11,2	11,8	10,4	10,9
11,2	11,9	10,6	11,1
11,4	11,9	10,6	11,1
11,5	12,1	10,7	11,3

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với mức ý nghĩa 0,05 hãy kiểm định cặp giả thuyết sau:

H_0 : Phân phối của lượng oxygen không tan trước và sau khi nạo vét là như nhau.

H_1: Phân phối lượng oxygen trước khi nạo vét lệch sang phải so với phân phối sau khi nạo vét.

Giải. Trước hết ta xác định hạng của phân phối kết hợp của 24 giá trị mẫu theo chiều tăng dần và sau đó tính tổng hạng của mẫu thứ nhất. Ta thu được kết quả sau:

Trước khi nạo vét		Sau khi nạo vét	
Giá trị	Hạng	Giá trị	Hạng
11	10	10,2	1
11,2	14	10,3	2
11,2	14	10,4	3
11,2	14	10,6	4,5
11,4	17	10,6	4,5
11,5	18	10,7	6
11,6	19	10,8	7,5
11,7	20	10,8	7,5
11,8	21	10,9	9
11,9	22,5	11,1	11,5
11,9	22,5	11,1	11,5
12,1	24	11,3	16
	T=216		

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì $n_1>10$ và $n_2>10$ ta dùng tiêu chuẩn kiểm định (9.24). Sau khi ghép nhóm các hạng giống nhau ta còn lại 18 nhóm với các tần số như sau:

Hạng	Tần số t_j	Hạng	Tần số t_j
1	1	14	3
2	1	16	1
3	1	17	1
4,5	2	18	1
6	1	19	1
7,5	2	20	1
9	1	21	1
10	2	22,5	2
11,5	2	24	1

Từ đó

$$\mu_T = \frac{n_1(n_1+n_2+1)}{2} = \frac{12(12+12+1)}{2} = 150$$

$$\sigma_T^2 = \frac{n_1 n_2}{12}\left[(n_1+n_2+1) - \frac{\sum_j t_j(t_j^2-1)}{(n_1+n_2)(n_1+n_2-1)}\right]$$

$$= \frac{12.12}{12}\left[(12+12+1) - \frac{6+6+24+6}{(12+12)(12+12+1)}\right] = 298,956$$

$\sigma_T = 17,29$

Từ đó $T_{qs} = \frac{T-\mu_T}{\sigma_T} = \frac{216-150}{17,29} = 3,82$

Với $\alpha = 0,05 \rightarrow u_\alpha = u_{0,05} = 1,645$

Vậy miền bác bỏ là $(1,645;+\infty)$

Do $T_{qs} \in W_\alpha$ nên với mức ý nghĩa 0,05 bác bỏ H_0 thừa nhận H_1 tức là thừa nhận phân phối của lượng oxygen không tan trước khi nạo vét lệch sang phải so với phân phối này sau khi nạo vét. Nó cũng chứng tỏ rằng lượng oxygen không tan trung bình trong 1 ml nước hồ trước khi nạo vét lớn hơn so với sau khi nạo vét hay dự án vét hồ thực sự có hiệu quả.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sau đây là kết quả giải bằng Stata:

.ranksum X_1, by(x2)

Two-sample Wilcoxon rank-sum(Mann-Whitney) test

x2	obs	rank sum	expected
1	12	216	150
2	12	84	150
combined	24	300	300

unadjusted variance	300.00
adjustment for ties	-1.04
adjusted variance	298.96

H_0 :x1 (x2==1)=x1 (x2==2)

z=3.817

Prob > |z| = 0.0001

Phương pháp kiểm định dựa trên việc sử dụng hạng của các giá trị mẫu thay cho bản thân các giá trị đó còn có thể sử dụng đối với các dấu hiệu định tính được đo theo thang thứ bậc hay khoảng cách, khi sự khác biệt về giá trị mẫu không có nhiều ý nghĩa tính toán.

Bài tập

9.27. Điều tra ngẫu nhiên thu nhập hàng năm của 4 người ở tỉnh A và 6 người ở tỉnh B, thu được kết quả sau:

Tỉnh A (triệu đồng)	Tỉnh B (triệu đồng)
6	11
10	13
15	14
29	17
	20
	31

Với mức ý nghĩa 0,05 hãy kiểm định giả thuyết là thu nhập trung bình hàng năm của dân 2 tỉnh A và B như nhau.

ĐS: T = 18. Không bác bỏ giả thuyết H_0

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.28. Hai mác xe A và B được chọn một cách ngẫu nhiên để thử nghiệm độ bền của động cơ (tính bằng số ngàn km sử dụng cho đến khi hỏng). Với các số liệu đã cho, hãy kiểm định với mức ý nghĩa 0,05 giả thuyết là mác xe A không tốt hơn mác xe B.

Mác A	61	30	49	48	41			
Mác B	26	32	22	39	41	30	41	36

9.29. Chất lượng của hai loại rượu vang không đo bằng thang tỷ lệ mà chỉ là thang thứ bậc như tốt hơn, kém hơn. Giả sử, khi thử 10 mẫu rượu M và 5 mẫu rượu C, ta sắp xếp các mẫu rượu được thử theo trình tự từ cao xuống thấp như sau:

$M_7\ M_5\ M_3\ M_{10}\ M_2\ M_6\ C_3\ M_1\ C_5\ M_4\ M_9\ C_4\ M_8\ C_1\ C_2$.

Với mức ý nghĩa 0,05 hãy kiểm định giả thuyết là chất lượng của hai loại rượu là như nhau.

ĐS: T = 97. Bác bỏ giả thuyết H_0.

9.30. Để xác định xem khách du lịch ngoại quốc thích đổi ngoại tệ ở các ngân hàng nhà nước hơn hay ở các quầy đổi tiền của tư nhân hơn, người ta chọn ngẫu nhiên 16 ngân hàng và 17 cửa hàng đổi tiền của tư nhân và thu được các số liệu sau về tỷ lệ đổi tiền:

Cửa hàng tư nhân: 1 1,5 1,3 1,2 2 2,1 1,8 1,4 1,5 2,3 1,7 1,2
1,4 1,1 1,9 2,3 2,6

Ngân hàng nhà nước: 1,8 1,7 1,9 1,5 1,7 1,3 2,4 2,8 1,9
2,1 2,2 2,6 1,9 2,5 2,0 2,1.

Với mức ý nghĩa 0,05 hãy kiểm định giả thuyết là tỷ lệ đổi tiền ở hai hệ thống đổi tiền trên là khác nhau.

6. KIỂM ĐỊNH TỔNG HẠNG THEO DẤU WILCOXON

Phương pháp kiểm định theo từng cặp giá trị phụ thuộc xét ở mục 2 chương 8 chỉ sử dụng được khi các biến ngẫu nhiên trong tổng thể phân phối chuẩn. Khi giả thiết này bị vi phạm thì có thể thay thế bằng phương pháp kiểm định phi tham số thông qua tổng hạng theo dấu của Wilcoxon. Nó là một biến tướng của kiểm định tổng hạng đã xét ở mục 5.

Nếu kí hiệu:

n là tổng số các cặp giá trị của hai mẫu điều tra có hiệu số D_i khác không.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

T_+ là tổng các hạng mang dấu dương

(Nếu không có hạng nào mang dấu dương thì $T_+=0$).

T_- là tổng các hạng mang dấu âm

(Nếu không có hạng nào mang dấu âm thì $T_-=0$).

$T=\min(T_+, |T_-|)$

$$\mu_T = \frac{n(n+1)}{4} \tag{9.28}$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{9.29}$$

Nếu có nhiều giá trị D_i cùng dấu và cùng giá trị thì được ghép lại thành một nhóm và mỗi giá trị được gán cho hạng bằng trung bình các hạng của nhóm đó. Lúc đó phương sai của T được tính bằng công thức.

$$\sigma_T^2 = \frac{1}{24}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_j t_j(t_j-1)(t_j+1)\right] \tag{9.30}$$

Nếu mọi nhóm đều có tần số bằng 1 thì ta lại thu được công thức (9.29).

Việc kiểm định được tiến hành theo các giả thuyết sau:

H_0: Phân phối của các hiệu số D_i là đối xứng qua giá trị 0.

a) H_1: Các hiệu số D_i có phân phối lệch sang phải điểm 0

b) H_1: Các hiệu số D_i có phân phối lệch sang trái điểm 0

b) H_1: Các hiệu số D_i hoặc lệch sang phải hoặc lệch sang trái so với điểm 0

Lúc đó thủ tục kiểm định được tiến hành theo hai trường hợp sau:

α) Nếu $n \le 50$

Tiêu chuẩn kiểm định là thống kê T được xác định tùy thuộc vào giả thuyết đối H_1 như sau:

a) $T=|T_-|$ (9.31)

b) $T=T_+$ (9.32)

c) $T=\min(T_+, |T_-|)$ (9.33)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nguyên tắc kiểm định như sau: Với mức ý nghĩa α (với kiểm định một phía có thể bằng 0,05; 0,025; 0,01; 0,005 và kiểm định hai phía có thể bằng 0,1; 0,05; 0,02; 0,01) và số lượng các giá trị D_i khác không bằng n bác bỏ H_0 nếu giá trị của T nhỏ hơn hoặc bằng giá trị tới hạn được cho trong phụ lục 12.

β) Nếu n>50

Tiêu chuẩn kiểm định là:

$$U = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \qquad (9.34)$$

Với các giả thuyết đối H_1 ở trường hợp a và b thì bác bỏ H_0 nếu $U_{qs} < -u_\alpha$ còn với trường hợp c thì bác bỏ H_0 nếu $U_{qs} < -u_{\alpha/2}$

Thí dụ: Để thử nghiệm năng suất của hai giống lúa, người ta chọn ngẫu nhiên 10 thửa ruộng, mỗi thửa lại chia đôi, một phần trồng giống lúa A, một phần trồng giống lúa B. Sau khi thu hoạch được kết quả sau (đơn vị kg/sào).

Thửa ruộng	Giống lúa A	Giống lúa B	Hiệu số
1	312	346	-34
2	333	372	-39
3	356	392	-36
4	316	351	-35
5	310	330	-20
6	352	364	-12
7	389	375	14
8	313	315	-2
9	316	325	-11
10	346	378	-32

Với mức ý nghĩa 0,05 hãy kiểm định xem phân phối của năng suất hai giống lúa có khác nhau hay không.

Giải. Trước hết ta xếp hạng từ thấp đến cao giá trị tuyệt đối của các hiệu số D_i, sau đó gán cho các hạng dấu tương ứng của chúng. Ta thu được bảng sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thửa ruộng	Hạng của hiệu số $\lvert X_{1i} - X_{2i} \rvert$	Hạng có dấu
1	7	-7
2	10	-10
3	9	-9
4	8	-8
5	5	-5
6	3	-3
7	4	4
8	1	-1
9	2	-2
10	6	-6

Từ đó

$T_{+qs} = 4$

$T_{-qs} = -51$

$T_{qs} = \min\left[T_{+}, |T_{-}|\right] = 4$

Để kiểm định hai phía với n=10 và $\alpha = 0{,}01$, tra bảng tìm được giá trị tới hạn là 8.

Do $T_{qs} \leq 8$ nên với mức ý nghĩa 0,05 bác bỏ H_0 và thừa nhận H_1 tức là phân phối của năng suất hai giống lúa là khác nhau. Điều đó cho thấy rằng năng suất lúa trung bình cũng khác nhau theo hướng giống lúa A thấp hơn giống lúa B vì phân phối năng suất có xu hướng lệch sang trái.

Sau đây là kết quả giải bằng Excel:

Wilcoxon Signed Rank Test

variables:	Group 1 - Group 2
4	sum of positive ranks
51	sum of negative ranks
10	n
27.50	expected value
9.81	standard deviation
-2.40	z
.0166	p-value (two-tailed)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

9.31. Một công ty du lịch quảng cáo hai tua du lịch trọn gói, tua A gồm 5 ngày đêm với giá 799 USD, tua B hai ngày đêm với giá 710 USD. Theo dõi ngẫu nhiên số khách đi trong 12 tua du lịch A và B, thu được các số liệu sau:

Tua A: 56 79 85 77 32 48 88 95 70 52 90 57

Tua B: 60 85 70 82 41 60 89 80 77 86 66 75

Với mức ý nghĩa 0,05 hãy kiểm định xem tua A có được ưa chuộng hơn tua B hay không?

ĐS: T = 51. Không bác bỏ giả thuyết H_0 => Hai tua du lịch được ưa chuộng như nhau.

9.32. Để so sánh hai phương thức quản lý, người ta đã áp dụng thử cả hai phương pháp vào 10 công ty, kết quả quản lý được cho điểm tương ứng như sau:

Phương pháp I: 25 28 37 10 14 51 30 28 33 20

Phương pháp II: 41 18 35 56 15 72 43 66 31 29

Với mức ý nghĩa 0,05 hãy kiểm định xem có sự khác biệt về hiệu quả của hai phương pháp quản lý nói trên hay không.

9.33. Độ dãn nở của nhựa khi gia công phụ thuộc khá mạnh vào công nghệ dùng để gia công. Với hai công nghệ A và B người ta gia công ngẫu nhiên 6 sản phẩm và thu được số liệu sau về độ dãn nở.

Công nghệ A: 6,1 9,2 8,7 8,9 7,6 7,1

Công nghệ B: 9,1 8,2 8,6 6,9 7,5 7,9

Với mức ý nghĩa 0,05 hãy kiểm định xem độ dãn nở trung bình của nhựa theo hai công nghệ trên có khác nhau hay không.

ĐS: T = 9. Không bác bỏ $H_0 \Rightarrow$ Độ dãn nở của nhựa theo hai công nghệ trên là như nhau.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7. KIỂM ĐỊNH KRUSKAL - WALLIS

Khi cần so sánh k (k > 2) kỳ vọng toán của k biến ngẫu nhiên không phân phối chuẩn thì có thể áp dụng kiểm định Kruskal-Wallis sau đây. Nó là sự mở rộng tương ứng của kiểm định tổng hạng Wilcoxon.

Giả sử có k tổng thể nghiên cứu, trong đó các biến ngẫu nhiên $X_1, X_2, ..., X_k$ phân phối theo một quy luật chưa biết, hoặc đã biết song không phải là phân phối chuẩn. Từ các tổng thể trên rút ra k mẫu độc lập kích thước tương ứng là $n_1, n_2, ..., n_k$. Lúc đó việc kiểm định sự bằng nhau của k kỳ vọng toán tương đương với việc kiểm định cặp giả thuyết.

H_0 : k phân phối trong các tổng thể là giống nhau

H_1: Có ít nhất hai phân phối khác nhau

Tiêu chuẩn kiểm định được chọn là thống kê

$$H = \frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n_T+1) \qquad (9.35)$$

Trong đó:

n_i là kích thước mẫu thứ i $(i = \overline{1,k})$

n_T là tổng kích thước của k mẫu

T_i là tổng các hạng của mẫu thứ i sau khi tất cả các mẫu đã được kết hợp lại và xếp hạng chung.

Với điều kiện giả thuyết H_0 là đúng thì thống kê H phân phối khi bình phương với số bậc tự do bằng k-1

Do đó với mức ý nghĩa α miền bác bỏ được xác định bằng biểu thức:

$$W_\alpha = \left\{ H = \frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n_T+1); H > \chi_\alpha^{2(k-1)} \right\} \qquad (9.36)$$

Từ k mẫu cụ thể tìm được H_{qs} và so sánh với W_α và kết luận.

Thí dụ: Ba nhóm học sinh trung học, mỗi nhóm 10 người thuộc các chuyên ban A, B và D được chọn một cách ngẫu nhiên để kiểm tra kiến thức về văn phạm tiếng Việt. Kết quả điểm kiểm tra như sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nhóm chuyên ban A	Nhóm chuyên ban B	Nhóm chuyên ban D
32	26	32
32	21	30
26	15	30
26	15	29
22	14	26
20	14	23
19	14	20
16	11	19
14	9	18
14	8	12

Với mức ý nghĩa 0,05 hãy kiểm định xem học sinh theo học các chuyên ban A, B và D có khác nhau về kiến thức văn phạm tiếng Việt hay không.

Giải. Ta có cặp giả thuyết sau:

H_0: Không có sự khác biệt giữa học sinh các chuyên ban về kiến thức văn phạm tiếng Việt.

H_1: Có ít nhất một chuyên ban khác biệt với các chuyên ban khác.

Để tìm H_{qs} trước hết phải xếp hạng 30 kết quả thu được theo trình tự từ thấp đến cao. Ta thu được kết quả xếp hạng như sau:

A	Hạng	B	Hạng	D	Hạng
32	29	26	24	32	29
32	29	21	18	30	26,5
26	22	15	10,5	30	26,5
26	22	15	10,5	29	25
22	19	14	7	26	22
20	16,5	14	7	23	20
19	14,5	14	7	20	16,5
16	12	11	3	19	14,5
14	7	9	2	18	13
14	7	8	1	12	4
$\sum$	178		90		197

Từ đó

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$H_{qs} = \frac{12}{30(30+1)}\left[\frac{178^2}{10} + \frac{90^2}{10} + \frac{197^2}{10}\right] - 3(30+1) = 8,4$$

Nếu các mẫu có nhiều giá trị trùng nhau thì có thể tính H' thay cho H theo công thức:

$$H' = \frac{H}{1 - \sum_{j=1}^{h}(t_j^3 - t_j)/(n_T^3 - n_T)} \qquad (9.37)$$

trong đó t_j là số giá trị giống nhau của nhóm thứ j.

Chẳng hạn với thí dụ trên ta xếp lại các giá trị của ba mẫu theo hạng và tần số tương ứng:

Hạng	Tần số	Hạng	Tần số
1	1	16,5	2
2	1	18	1
3	1	19	1
4	1	20	1
7	5	22	3
10,5	2	24	1
12	1	25	1
13	1	26,5	2
14,5	2	29	3

lúc đó

$$\frac{\sum_{j=1}^{h}(t_j^3 - t_j)}{n_T^3 - n_T} = \frac{1}{30^3 - 30}[(5^3 - 5) + (2^3 - 2) + (2^3 - 2) + (2^3 - 2) + (3^3 - 3) + (2^3 - 2) + (3^3 - 3)]$$

$$= \frac{192}{26970} = 0,0071$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đó

$$H'_{qs} = \frac{H}{1-0,0071} = \frac{8,4}{0,9929} = 8,46$$

như vậy là hai kết quả xấp xỉ nhau.

Với $\alpha = 0,05 \rightarrow \chi_{\alpha}^{2(k-1)} = \chi_{0,05}^{2(2)} = 5,991$

Vậy miền bác bỏ là $(5,991;+\infty)$

Do $H'_{qs} \in W_{\alpha}$ nên bác bỏ H_0, thừa nhận H_1 tức là thừa nhận có ít nhất một nhóm học sinh có kiến thức về văn phạm tiếng Việt cao hơn các nhóm khác.

Giải bằng Excel cho kết quả sau:

Kruskal-Wallis Test

Median	*n*	*Avg. Rank*	
21.00	10	17.90	Group 1
14.00	10	8.85	Group 2
24.50	10	19.75	Group 3
19.50	30		Total
		8.780	H
		2	d.f.
		.0124	p-value

Bài tập

9.34. Theo dõi ngẫu nhiên sản lượng của 3 máy trong 5 ngày thu được các kết quả sau:

Máy I	Máy II	Máy III
47	55	54
53	54	50
49	58	51
50	61	51
46	52	49

Với mức ý nghĩa 0,05 hãy kiểm định giả thuyết là sản lượng trung bình của ba máy đó là như nhau.

ĐS: $H_{qs} = 56,505$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.35. Để so sánh chi phí quảng cáo trên bốn tờ báo khác nhau (với các điều kiện quảng cáo như nhau) người ta đã lấy mẫu 7 lần quảng cáo trên mỗi tờ báo và thu được các kết quả sau (đơn vị ngàn đồng).

Báo A	57	65	50	45	70	62	48
Báo B	72	81	64	55	90	38	75
Báo C	35	42	58	59	46	60	61
Báo D	73	85	92	68	82	94	66

Hãy tìm P-value để kiểm định xem có sự khác biệt về chi phí quảng cáo giữa các tờ báo nói trên hay không.

9.36. Để so sánh tốc độ tăng trưởng vốn của ba ngành công nghiệp A, B và C người ta lấy ngẫu nhiên 8 công ty của ngành A, 6 công ty của ngành B và 6 công ty của ngành C để điều tra tốc độ tăng trưởng vốn trong năm qua và thu được các số liệu sau:

Ngành A: 1,38 1,55 1,90 2,00 1,22 2,11 1,98 1,61
Ngành B: 2,23 2,50 2,79 3,01 1,99 2,45
Ngành C: 1,06 1,37 1,09 1,65 1,44 1,11

Hãy tìm P-value để kiểm định xem tốc độ tăng trưởng vốn của 3 ngành trên có khác nhau hay không?

9.37. Theo dõi ngẫu nhiên giá thuê nhà tại 5 thành phố (với điều kiện thuê nhà như nhau) thu được các số liệu sau:

Thành phố A: 900 1200 850 1320 1400 1150975
Thành phố B: 625 640 775 1000 690 550 840 750
Thành phố C: 415 400 420 560 780 620 800 390
Thành phố D: 410b 310 320 280 500 385 440
Thành phố E: 340 425 275 210 575 360

Hãy tìm P-value để kiểm định xem có sự khác biệt về giá thuê nhà ở 5 thành phố nói trên hay không.

8. KIỂM ĐỊNH ĐOẠN MẠCH

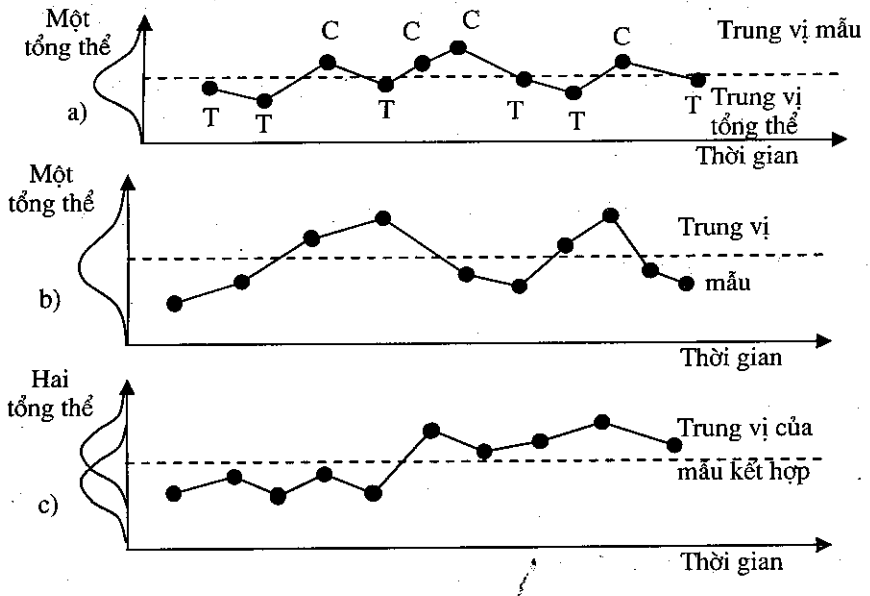
8.1. Kiểm định một mẫu

Một trong các giả thuyết cơ bản mà ta đưa ra ở các mục trước là giả thuyết mẫu được rút ra từ tổng thể một cách ngẫu nhiên. Phần này ta sẽ xét một phương pháp kiểm định tính ngẫu nhiên của mẫu gọi là phương pháp kiểm định đoạn mạch.

Theo định nghĩa mẫu ngẫu nhiên là tập hợp các biến ngẫu nhiên độc lập

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

được xây dựng từ biến X trong tổng thể. Như vậy thì các phần tử của mẫu phải được lấy ra một cách độc lập từ một tổng thể chung. Chẳng hạn nếu giá trị mẫu được mô tả theo trật tự thời gian thì đồ thị của nó có dạng như hình 9.4a. Mặt khác nếu các giá trị mẫu tương quan với nhau thì đồ thị của chúng sẽ chao đảo như ở hình 9.4b. Còn nếu mẫu được rút ra từ hai tổng thể khác nhau thì đồ thị của giá trị mẫu sẽ có dạng như hình 9.4c.



Hình 9.4

Để kiểm định về tính ngẫu nhiên của mẫu ta đưa ra giả thuyết thống kê.

H_0 : Mẫu được lập một cách ngẫu nhiên

H_1 : Mẫu được lập một cách phi ngẫu nhiên

Chú ý rằng nếu H_0 là đúng tức là mẫu được tạo lập một cách ngẫu nhiên thì các giá trị nhỏ hơn trung vị (ký hiệu là T) và các giá trị lớn hơn trung vị (ký hiệu là C) phải phân phối một cách hoàn toàn ngẫu nhiên xung quanh trung vị.

Còn nếu H_0 sai thì sự phân phối đó sẽ mang tính quy luật rõ rệt. Chẳng hạn theo hình 9.4a ta có dãy phân phối trong quan hệ so sánh với trung vị như sau:

TT CT CC TT CT (*)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

còn với hình 9.4b thì dãy phân phối là

TT CC TT CC TT (**)

Từ đó ta có định nghĩa sau: Mỗi đoạn mạch là một dãy liên tiếp các ký hiệu giống nhau mà trước và sau nó hoặc không có có ký hiệu hoặc ký hiệu khác.

Như vậy, trong dãy (*) ta có số đoạn mạch R=7. Còn trong dãy (**) ta có số đoạn mạch R=5.

Giả sử kích thước mẫu là n (nếu n lẻ thì đường trung vị sẽ đi qua điểm trung vị, lúc đó điểm này sẽ không được tính là T hay C. Lúc đó n là số lượng các quan sát còn lại).

Nếu H_0 là đúng thì số lượng của đoạn mạch R sẽ có phân phối xấp xỉ chuẩn với kỳ vọng toán

$$E(R) \approx \frac{n}{2} + 1 \tag{9.38}$$

và sai số chuẩn

$$Se(R) \approx \frac{\sqrt{n-1}}{2} \tag{9.39}$$

Từ đó tiêu chuẩn kiểm định được chọn là

$$U = \frac{R - E(R)}{Se(R)} \tag{9.40}$$

và miền bác bỏ α được xác định bằng biểu thức

$$W_\alpha = \left\{ U = \frac{R - E(R)}{Se(R)}; U < -u_\alpha \right\} \tag{9.41}$$

Từ mẫu cụ thể xác định được R và tìm giá trị quan sát U_{qs} so sánh nó với W_α và kết luận.

Thí dụ 1. Với các số liệu mẫu được cho trong hình 9.4.a hãy kiểm định giả thuyết về tính ngẫu nhiên của mẫu điều tra với mức ý nghĩa 0,05.

Giải. Từ hình 9.4.a ta tìm được

n=10; R=7

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$E(R) \approx \frac{10}{2} + 1 = 6$$

$$Se(R) \approx \frac{\sqrt{10-1}}{2} = 1,5$$

Từ đó

$$U_{qs} = \frac{7-6}{1,5} = 0,67$$

Với $\alpha = 0,05 \rightarrow u_\alpha = u_{0,05} = 1,645$

Vậy miền bác bỏ là $(-\infty; -1,645)$

Do $U_{qs} = 0,67 \notin W_\alpha$ nên chưa thể bác bỏ H_0 tức là có thể cho rằng mẫu được rút ra một cách ngẫu nhiên.

Sau đây là kết quả giải bằng Excel:

Runs Test for Random Sequence

variable: (no label)

n	runs	
6	4	t
4	3	c
10	7	total
5.80	expected value	
1.42	standard deviation	
0.49	z	
.6229	p-value (two-tailed)	

8.2. Kiểm định hai mẫu (Kiểm định Wald-Wolfowitz)

Kiểm định Wald-Wolfowitz là sự mở rộng kiểm định đoạn mạch để xác định xem hai tổng thể có các phân phối xác suất như nhau hay không. Trong trường hợp này cặp giả thuyết thống kê là

H_0 : Hai tổng thể nghiên cứu có cùng một phân phối xác suất

H_1 : Hai tổng thể nghiên cứu có các phân phối xác suất khác nhau

Như vậy đây là phương án kiểm định phi tham số để thay cho kiểm định T về sự bằng nhau của hai trung bình tổng thể. Để tiến hành kiểm định này chỉ yêu cầu hai mẫu rút ra từ hai tổng thể là độc lập và hoàn toàn ngẫu nhiên. Giả sử điều tra hai mẫu như vậy với kích thước tương ứng là n_1 và

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

n_2 và các giá trị của mẫu được sắp xếp đồng thời theo trình tự tăng dần thành một dãy gồm $n=n_1+n_2$ kết quả. Sau đó gọi R là tổng số đoạn mạch của dãy kết quả đó thì ta có thể áp dụng thủ tục kiểm định các đoạn mạch đã trình bày ở mục trước đối với R.

Thí dụ 2. Người quản lý một cửa hàng muốn kiểm định xem hiệu quả làm việc của hai nhân viên bán hàng có như nhau hay không. Để làm điều đó người quản lí đã thống kê số sản phẩm mà mỗi nhân viên đã bán được trong một số ngày và thu được kết quả sau:

Nhân viên A: 35 44 39 50 48 29 60 75 49 66

Nhân viên B: 17 23 13 24 33 21 18 16 32

Hãy tìm P-value để kiểm định xem hoạt động của nhân viên A có thực sự hiệu quả hơn nhân viên B hay không.

Giải. Ta có $n_1=10$ và $n_2=9$. Xếp các giá trị của cả hai mẫu thành một dãy kết quả theo trình tự tăng dần thu được dãy sau:

B B B B B B B A B B A A A A A A A A A

Như vậy tổng số đoạn mạch R=4. Từ đó tìm được

$$E(R)=10{,}47$$

$$\sigma_R = 2{,}1$$

và

$$U_{qs} = \frac{4-10{,}47}{2{,}1} = -3{,}08$$

Tra bảng tìm được P-value ≈ 0.

Như vậy có thể bác bỏ giả thuyết là hiệu quả làm việc của hai nhân viên A và B là như nhau.

Sau đây là kết quả giải bằng Excel:

Runs Test for Random Sequence

variable: (no label)

n	runs	
9	2	b
10	2	a
19	4	total

10.47	expected value
2.11	standard deviation
-2.83	z
.0023	p-value (one-tailed, upper)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

9.38. Để nghiên cứu xem khách hàng vào một nhà hàng có thực sự thích chỗ ngồi này hay chỗ ngồi khác hay không, người quản lý đã điều tra ngẫu nhiên 20 bàn ăn của nhà hàng tại một thời điểm. Nếu ký hiệu B là bàn bận và R là bàn rỗi thì thu được dãy kết quả sau:

BRBRBRRBBBRBRRRRBBRB

Với mức ý nghĩa 0,05 hãy kết luận về vấn đề trên.

ĐS: $U_{qs} = 0{,}918$

9.39. Trong các sản phẩm do một máy sản xuất người ta thấy có phế phẩm. Nếu việc xuất hiện các phế phẩm là không hoàn toàn ngẫu nhiên thì máy phải được đưa đi căn chỉnh lại. Theo dõi ngẫu nhiên quá trình sản xuất 58 sản phẩm thu được dãy kết quả sau:

CCCCCPPPCCCCCCPPPCCCCCCCCCCPP

PPCCCCCCCCCCCPPPCCCCCCCCCPPPP

Với mức ý nghĩa là 0,05 hãy kết luận về vấn đề trên.

9.40. Người ta luôn khẳng định rằng các hồ sơ dự thầu một hợp đồng của chính phủ được mở một cách hoàn toàn ngẫu nhiên. Tại một cuộc đấu thầu có 42 hồ sơ dự thầu, trong đó có 30 hồ sơ là của các công ty trong nước và 12 hồ sơ của các công ty nước ngoài. Nếu ký hiệu D là hồ sơ của công ty trong nước và F là hồ sơ của công ty nước ngoài thì thu được dãy kết quả sau:

DDDDDDDFDDDDDDFFDDDDDDDDFDDFDDDDDDFFFFFFFD

Với mức ý nghĩa 0,05 có thể cho rằng các công ty nước ngoài đã bị phân biết đối xử hay không

9.41. Để so sánh hiệu quả của hai quảng cáo người ta đã đề nghị 8 khách hàng đánh giá bằng cách cho điểm chất lượng quảng cáo 1 theo thanh điểm 10 và 9 khách hàng khác được đề nghị cho điểm đối với quảng cáo 2. Kết quả thu được như sau:

Quảng cáo 1: 7 8 6 7 8 9 9 10

Quảng cáo 2: 3 4 3 5 5 4 2 5 4

Hãy tìm P-value để kiểm định xem theo đánh giá của khách hàng thì hai quảng cáo có chất lượng như nhau hay không?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.42. Có số liệu sau đây về tiền lương của 7 công nhân thuộc một công ty sợi và 8 công nhân thuộc một công ty may được chọn một cách ngẫu nhiên.

Công ty sợi: 175 170 166 168 204 96 147

Công ty may: 89 120 136 160 111 101 98 80

Với mức ý nghĩa 0,05 hãy kiểm định xem tiền lương trung bình của công nhân hai công ty có khác nhau hay không.

ĐS: P-value = 0,298 > 0,05. Không có cơ sở để bác bỏ H_0.

9. TÓM TẮT CHƯƠNG 9

1. Kiểm định phi tham số là phương pháp kiểm định giả thuyết thống kê khi không có được bất kỳ giả thiết nào về tham số của tổng thể cũng như về phân phối xác suất của biến ngẫu nhiên trong tổng thể.
2. Một trong những kiểm định phi tham số quan trọng nhất là kiểm định khi bình phương. Nó cho phép kiểm định tính độc lập của hai dấu hiệu nghiên cứu định tính, kiểm định nhiều cơ cấu tổng thể, kiểm định dạng phân phối xác suất của biến ngẫu nhiên. Ngoài ra còn có thể tìm thấy nhiều ứng dụng khác nữa của kiểm định này.
3. Ngoài kiểm định khi bình phương còn có thể dùng kiểm định Kolmogorov, kiểm định Lillieford hoặc kiểm định Jarque - Bera để kiểm định về dạng phân phối chuẩn của biến ngẫu nhiên trong tổng thể. Đặc biệt kiểm định Jarque - Bera hay được sử dụng trong phân tích kinh tế lượng.
4. Để kiểm định về một hoặc hai trung vị của tổng thể người ta hay sử dụng kiểm định theo dấu.
5. Khi cần kiểm định về sự bằng nhau của hai trung bình tổng thể mà không có được giả thiết là các tổng thể cùng phân phối chuẩn thì có thể dùng kiểm định tổng hạng Wilcoxon hoặc kiểm định tổng hạng theo dấu Wilcoxon.
6. Kiểm định Kruskal - Wallis được dùng để kiểm định sự bằng nhau của nhiều trung bình tổng thể.
7. Các kiểm định đoạn mạch được sử dụng để kiểm định tính ngẫu nhiên của một mẫu hoặc tính đồng nhất về dạng phân phối xác suất của nhiều tổng thể.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

* Kiểm định χ^2 về tính độc lập của 2 dấu hiệu định tính

H_0 : A và B độc lập

Tiêu chuẩn kiểm định

$$\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{h}\frac{h_{ij}^2}{n_i m_j} - 1\right] \sim \chi^2\left[(k-1)(h-1)\right]$$

* Kiểm định χ^2 về k tham số p

$$H_0 : p_1 = p_2 = ... = p_k$$

Tiêu chuẩn kiểm định

$$\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{2}\frac{X_{ij}^2}{n_i m_j} - 1\right] \sim \chi^2(k-1)$$

* Kiểm định χ^2 về quy luật phân phối xác suất

H_0 : X phân phối theo quy luật A

Tiêu chuẩn kiểm định

$$\chi^2 = \sum_{i=1}^{k}\frac{(n_i - n'_i)^2}{n'_i} \sim \chi^2(k-r-1)$$

* Kiểm định Kolmogorov về quy luật phân phối xác suất

Tiêu chuẩn kiểm định

$$\lambda = \sqrt{n}D = \sqrt{n}\max_i\left|F^*(X_i) - F(X_i)\right|$$

* Kiểm định Jarque - Bera về phân phối chuẩn

H_0 : X phân phối chuẩn

Tiêu chuẩn kiểm định

$$JB = n\left[\frac{a_3^2}{6} + \frac{(a_4-3)^2}{24}\right] \sim \chi^2(2)$$

* Kiểm định tổng hạng Wilcoxon (mẫu lớn)

H_0 : Hai tổng thể có phân phối giống nhau

Tiêu chuẩn kiểm định

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$U = \frac{T - \mu_T}{\sigma_T} \sim N(0,1)$$

* Kiểm định tổng hạng theo dấu Wilcoxon (mẫu lớn)

H_0 : Các hiệu số D_i đối xứng qua giá trị 0.

Tiêu chuẩn kiểm định

$$U = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)$$

* Kiểm định Kruskal-Wallis

H_0 : k phân phối trong các tổng thể là giống nhau

Tiêu chuẩn kiểm định

$$H = \frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n_T+1) \sim \chi^2(k-1)$$

* Kiểm định đoạn mạch

+ Trường hợp một mẫu

H_0 : Mẫu được lập một cách ngẫu nhiên

Tiêu chuẩn kiểm định

$$U = \frac{R - E(R)}{Se(R)} \sim N(0,1)$$

+ Trường hợp hai mẫu

H_0 : Hai tổng thể nghiên cứu có cùng phân phối xác suất

Tiêu chuẩn kiểm định Wald-Wolfowitz

$$U = \frac{R - E(R)}{Se(R)} \sim N(0,1)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Kiểm định phi tham số	Nonparametric hypothsis test
Kiểm định Khi bình phương	Chi - squares test
Bảng tiếp liên	Contingent table
Quy luật đa thức	Multinomial distribution
Kiểm định Kolmogorov-Smirnov	Kolmogorov - Smirnov test
Kiểm định Lilliefors	Lilliefors test
Kiểm định Jarque - Bera	Jarque - Bera test
Kiểm định theo dấu	Sign test
Kiểm định tổng hạng Wilcoxon	Wilcoxon rank - sum test
Kiểm định tổng hạng theo dấu Wilcoxon	Wilcoxon signed - rank test
Kiểm định Kruskal - Wallis	Kruskal - Wallis test
Đoạn mạch	Runs
Kiểm định đoạn mạch	Runs test
Kiểm định Wald - Wolfowitz	Wald - Wolfowitz test

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

9.43. Điểm khác biệt cơ bản giữa kiểm định tham số và kiểm định phi tham số là gì?

9.44. Kiểm định khi bình phương có thể dùng để giải quyết những bài toán nào?

9.45. Bài toán kiểm định quy luật đa thức có gì khác biệt so với bài toán kiểm định sự bằng nhau của k ($k > 2$) cơ cấu tổng thể?

9.46. Để kiểm định dạng phân phối chuẩn của biến ngẫu nhiên trong tổng thể có thể sử dụng các phương pháp kiểm định nào?

9.47. Mục đích của phương pháp kiểm định theo dấu là gì?

9.48. Khi nào thì việc kiểm định trung vị lại chính xác hơn kiểm định trung bình?

9.49. Khi kiểm định sự bằng nhau của hai trung bình tổng thể mà các biến ngẫu nhiên trong tổng thể lại không phân phối chuẩn thì có thể dùng những phương pháp nào để kiểm định?

9.50. Khi có hai tổng thể phụ thuộc theo từng cặp song lại không phân phối chuẩn thì làm thế nào để có thể kiểm định hai trung bình của chúng?

9.51. Có thể dùng phương pháp nào để kiểm định sự bằng nhau của k ($k > 2$) trung bình tổng thể khi các biến ngẫu nhiên trong tổng thể không phân phối chuẩn?

9.52. Mục đích của phương pháp kiểm định đoạn mạch là gì?

9.53. Phương pháp kiểm định Wald - Wolfowitz dùng để làm gì?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập tổng hợp chương 9

9.54. Một công ty dự kiến 5 tên gọi cho một sản phẩm mới. Trước khi quyết định người ta muốn đánh giá phản ứng của khách hàng đối với các tên gọi đó có như nhau hay không. Một mẫu ngẫu nhiên gồm 100 khách hàng được đề nghị chọn tên cho sản phẩm và thu được kết quả sau:

Tên:	A	B	C	D	E
Số người chọn:	4	12	34	40	10

Với mức ý nghĩa 0,05 hãy kết luận xem tỷ lệ khách hàng ưa thích các tên gọi đó có như nhau hay không.

ĐS: $\chi^2_{qs} = 50,8$

9.55. Trở lại bài 8.74, với mức ý nghĩa 0,1 hãy dùng kiểm định Lilliefors để kiểm định xem số tàu vào hai cảng có phân phối chuẩn hay không?

9.56. Một nhà kinh tế muốn kiểm định giả thuyết là thị phần của các công ty trong một ngành công nghiệp như sau:

Công ty:	A	B	C	D	E	Các CT khác
Thị phần:	1%	9%	14%	26%	0,5%	49,5%

Việc phân tích được tiến hành trên cơ sở 100 sản phẩm được chọn một cách ngẫu nhiên.

a. Với các số liệu như trên có thể tiến hành kiểm định khi bình phương được không.

b. Nếu không thì phải làm thế nào để có thể áp dụng được kiểm định khi bình phương.

ĐS: a. Không

b. Ghép thị phần của A và B, của E với các công ty khác.

9.57. Tiếp tục bài 9.56. Giả sử 100 sản phẩm được điều tra một cách ngẫu nhiên được phân loại như sau: Công ty A: 0; Công ty B: 0; Công ty C: 26; Công ty D: 15; Công ty E: 2; Các công ty khác: 57. Với mức ý nghĩa 0,05 hãy kiểm định giả thuyết ở bài 9.55.

ĐS: $\chi^2_{qs} = 26,56$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.58. Người ta muốn so sánh hiệu quả của hai phương pháp sản xuất khác nhau. Vì số sản phẩm được sản xuất ra hàng ngày có biến động rất lớn do đó người ta sử dụng trung vị thay vì trung bình để so sánh. Các mẫu ngẫu nhiên về sản lượng hàng ngày của hai phương pháp sản xuất như sau:

Phương pháp A: 5 7 19 8 10 16 14 9 22 4 7 8 15 18 7

Phương pháp B: 8 12 15 28 5 14 19 16 23 19 25 17 20

Với mức ý nghĩa 0,05 hãy kiểm định xem trung vị của hai phương pháp sản xuất trên có như nhau không.

9.59. Giá của một loại cổ phiếu bán trên thị trường chứng khoán trong 100 phiên giao dịch được cho ở bảng sau:

Giá cổ phiếu (1000đ)	13-15	15-17	17-19	19-21	21-23
Số phiên giao dịch	5	18	42	27	8

Dùng kiểm định Jarque - Bera và với mức ý nghĩa 5% có thể cho rằng giá của loại cổ phiếu nói trên là biến ngẫu nhiên tuân theo quy luật phân phối chuẩn hay không?

ĐS: Có thể cho rằng giá của loại cổ phiếu trên tuân theo quy luật chuẩn.

9.60. Gạo đủ tiêu chuẩn xuất khẩu là gạo có phân bố thành phần như sau:

Hạt nguyên: 90% - Hạt vỡ: 6% - Tấm: 4%

Kiểm tra 1000 hạt gạo của một lô gạo người ta thấy trong đó có:

Hạt nguyên: 880 - Hạt vỡ: 60 - Tấm: 60

Hỏi lô gạo đó có đủ tiêu chuẩn xuất khẩu không? Cho kết luận với mức ý nghĩa 5%.

ĐS: Lô gạo đó không đủ tiêu chuẩn xuất khẩu.

9.61. Trong giai đoạn nền kinh tế đang tăng trưởng, các nhà kinh tế cho rằng các công ty nhỏ thường thuê nhiều công nhân hơn và sa thải ít hơn so với các công ty lớn. Để kiểm tra điều này, người ta lấy 1 mẫu ngẫu nhiên gồm 1032 công ty. Bảng sau đây cho biết số công nhân tuyển thêm và số công nhân bị sa thải của các công ty được chia theo quy mô của các công ty:

	Công ty nhỏ	Công ty vừa	Công ty lớn	Tổng số
Số lượng thuê thêm	210	290	325	825
Số lượng sa thải	32	95	80	207
Tổng số	242	385	405	1032

Dựa vào mẫu này có thể khẳng định việc thuê thêm hoặc sa thải nhân công độc lập với quy mô của các công ty hay không? Cho mức ý nghĩa là 5%.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9.62. Sản phẩm sản xuất ra trên một dây chuyền tự động được đóng gói ngẫu nhiên theo quy cách: 3 sản phẩm /1 hộp. Có thể xem số chính phẩm của một hộp là biến ngẫu nhiên tuân theo quy luật nhị thức được không, biết rằng kiểm tra 100 hộp người ta thấy 75 hộp không có phế phẩm, 20 hộp có 1 phế phẩm, 5 hộp có 2 phế phẩm, không có hộp nào có 3 phế phẩm? Cho kết luận với mức ý nghĩa 1%.

ĐS: X tuân theo quy luật nhị thức B (n = 3; p = 0,9).

9.63. Sau một số năm thực hiện chủ trương cho nông dân vay vốn để phát triển sản xuất ở một tỉnh gồm hai vùng trung du và vùng núi, giám đốc Ngân hàng tỉnh muốn kiểm tra xem việc cho vay có sự phân biệt đối xử giữa hai vùng không. Một mẫu ngẫu nhiên gồm 100 chủ hộ của hai vùng đã được điều tra và thu được kết quả như sau:

	Trung du	Vùng núi	Cộng
Số hộ không được vay	10	8	18
Số hộ được vay	60	22	82
Tổng	70	30	100

Với mức ý nghĩa 0,05 hãy kiểm định giả thiết về tính độc lập của đặc tính vùng và việc cho vay đối với các hộ.

9.64. Kết quả 200 phiên giao dịch chứng khoán cho thấy mức độ tăng giá của một loại cổ phiếu (%) có phân phối thực nghiệm như sau:

Stt	Mức tăng giá(%)	Số phiên
1	2 - 4	21
2	4 - 6	16
3	6 - 8	15
4	8 - 10	26
5	10 - 12	22
6	12 - 14	14
7	14 - 16	21
8	16 - 18	22
9	18 - 20	18
10	20 - 22	25

Có ý kiến cho rằng mức độ tăng giá của loại cổ phiếu trên có phân phối đều. Cho kết luận về ý kiến nói trên với mức ý nghĩa $\alpha = 5\%$.

ĐS: Có thể cho rằng mức độ tăng giá của loại cổ phiếu trên tuân theo quy luật phân phối đều R(a=2, b=22).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Nguyễn Đình Cử, Nguyễn Cao Văn. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Thống kê. Hà Nội. 1991.
2. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội. 1999.
3. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà Nội. 2009.
4. Nguyễn Cao Văn, Bùi Dương Hải. *Hướng dẫn trả lời lý thuyết và giải bài tập Kinh tế lượng*. NXB Tài chính. Hà Nội. 2011.
5. Ngô Văn Thứ. *Thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội. 2011
6. Nguyễn Quang Dong: *Kinh tế lượng nâng cao*. NXB KHKT. 2002.
7. PAO C.P. Линейные статистические методы и их применения. Москва. 1968.
8. Уйякс. Математическая статистика. Москва. 1982.
9. Giard V. Statistique appliquée à la gestion. Economica. Paris. 1992.
10. Amir D Aczel. Complete Business Statistics. Irwin Boston 1993.
11. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992.
12. Kendall M and Stuart A, The advanced theory of Statistics. California. 1993.
13. OTT L.R. And introduction to statistical methods and data analysis. California. 1993.
14. Gujarati D. Basic econometrics, McGraw Hill. New York. 1995.
15. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons. New York. 1990.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 10
PHÂN TÍCH PHƯƠNG SAI

1. MỞ ĐẦU

Giả sử chúng ta cần nghiên cứu, phân tích sự biến động của một tổng thể thông qua một biến ngẫu nhiên (chỉ tiêu nghiên cứu). Thông thường, như đã xét ở các chương trước, chúng ta có các bài toán ước lượng, kiểm định giả thuyết về giá trị tham số và quy luật phân phối xác suất của biến ngẫu nhiên đó. Tuy nhiên, phương pháp kiểm định, so sánh tham số đó chỉ sử dụng được khi sự biến động của biến ngẫu nhiên đó chỉ chịu tác động của một nhân tố và cũng ở một hoặc tối đa hai mức cố định. Nếu sự biến động đó được tạo nên do nhiều nhân tố cùng tác động, hoặc do một nhân tố nhưng ở nhiều mức độ khác nhau thì phải tiến hành phân tích phương sai mới thấy được vai trò ảnh hưởng của từng mức nhân tố, hoặc từng nhân tố cũng như sự ảnh hưởng tổng hợp của các nhân tố (nếu có) trong việc tạo nên sự biến đổi, sự sai khác đó.

Nhân tố ở đây được hiểu là các yếu tố, điều kiện khách quan (thời tiết, khí hậu, thiên tai...) hoặc chủ quan (phương pháp thí nghiệm, giống cây trồng, máy móc sử dụng...) có tác động trực tiếp đến sự biến động của biến ngẫu nhiên mà ta nghiên cứu. Ví dụ, chúng ta cần phân tích sự biến động của năng suất một loại cây trồng (X) dưới tác động của hai nhân tố: nơi trồng (ký hiệu là F) và giống (ký hiệu là G). Ta có X là một biến ngẫu nhiên định lượng và thông thường F và G là các biến (ngẫu nhiên hoặc không ngẫu nhiên) định tính. Do đó các giá trị của nhân tố (trong thống kê người ta thường gán cho chúng bằng những số nguyên 1,2...) chỉ có ý nghĩa phân loại và trong phân tích phương sai chúng ta gọi đó là các mức của nhân tố. Ví dụ, F có p giá trị ($F = \overline{1,p}$) tức là F có p *mức* khác nhau. F=1 có nghĩa là cây trồng được trồng ở nơi được đánh số là 1, F=2 là nơi được đánh số là 2...ta cũng giả thiết rằng G có q giá trị (q mức) tức là loại cây trồng trên có q giống khác nhau được đem ra trồng thí nghiệm. Ta có thể mô tả bằng sơ đồ sau đây:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$X = (\text{Năng suất cây trồng}) \xrightarrow[\substack{F = (\text{Nơi trồng}) \\ G = (\text{Giống})}]{\text{Nhân tố tác động}} \{Xij\}\ i = \overline{1,p},\ j = \overline{1,q}$$

Trong đó X_{ij} = (Năng suất cây trồng ở địa điểm i và của giống j).

Nếu chúng ta chỉ xét một nhân tố G có 2 mức (q=2) thì chúng ta có bài toán so sánh năng suất trung bình của hai giống cây trồng (xem Chương VIII). Với q>2 và chúng ta lại giả thiết rằng điều kiện địa lí của các lô ruộng trồng các loại giống cây trồng nói trên là như nhau..., nói chung các điều kiện khác đều giống nhau (điều này trên thực tế rất ít khi xảy ra) thì chúng ta có bài toán so sánh năng suất trung bình của q giống cây trồng khác nhau...

Tùy theo *số lượng nhân tố và điều kiện tác động* của chúng lên chỉ tiêu nghiên cứu (biến ngẫu nhiên) mà người ta có những bài toán phân tích phương sai khác nhau cùng với những mô hình phân tích phương sai khác nhau. Tệp số liệu dùng để minh họa cho các mô hình phân tích phương sai sẽ được trình bày dưới đây có tên gọi là XSTK10_1. Trong tệp số liệu này F có 4 mức (p=4) và G có 2 mức (q=2), sau đây là nội dung tệp số liệu. (Xem bảng 10.1).

Bảng 10.1

Contains data from C:\WINSTATA\XSTK10_1.DTA

obs:	13	
vars:	3	15 Aug 2010 10:54
size:	130 (84.4% of memory free)	
1. x	float %9.0g	năng suất
2. f	byte %8.0g	nơi trồng
3. g	byte %8.0g	giống

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sorted by:

• List

	x	f	g
1.	1.38	1	1
2.	1.38	1	1
3.	1.42	1	2
4.	1.42	1	2
5.	1.41	2	1
6.	1.42	2	2
7.	1.44	2	1
8.	1.45	2	2
9.	1.32	3	1
10.	1.33	3	1
11	1.34	3	2
12.	1.31	4	2
13.	1.33	4	1

Ở mục 2 ta sẽ nghiên cứu bài toán phân tích phương sai một nhân tố, còn mục 3 sẽ đề cập đến bài toán phân tích hai nhân tố trong hai trường hợp: tác động riêng rẽ và tác động tổng hợp.

2. MÔ HÌNH PHÂN TÍCH PHƯƠNG SAI MỘT NHÂN TỐ

Xét biến ngẫu nhiên X tuân theo quy luật phân phối chuẩn $N(\mu,\sigma^2)$ và một nhân tố F tác động lên X có p mức khác nhau. Như vậy, tương ứng mỗi mức nhân tố i ta có biến ngẫu nhiên X_i và chúng cũng tuân theo quy luật phân phối chuẩn $N(\mu_i,\sigma_i^2)$ (ví dụ như X_i là năng suất cây trồng ở nơi thứ i). Nếu tiến hành quan sát X_i bằng cách lấy một mẫu ngẫu nhiên kích thước n_i:

$W_i = (\ X_{1i}, X_{2i}, ..., X_{ki}, ..., X_{n_i})$ $(i = \overline{1,p})$ thì ta có thể viết:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$X_{ki} = \mu + \alpha_i + U_{ki} \qquad k = \overline{1, n_i} \qquad (10.1)$$

Trong đó α_i đặc trưng cho sự khác biệt về giá trị trung bình μ của biến ngẫu nhiên X dưới tác động của nhân tố F ở mức i và U_{ki} là các sai số ngẫu nhiên giả thiết là độc lập với nhau, cùng tuân theo quy luật phân phối chuẩn $N(0, \sigma_{ij}^2)$.

Mô hình (10.1) được gọi là mô hình phân tích phương sai một nhân tố. Để tiến hành phân tích phương sai ta xét cặp giả thuyết sau đây:

$H_0 : (\alpha_i = 0 \forall i, i = \overline{1, p})$

H_1: (Tồn tại ít nhất một cặp $\alpha_i \neq 0$)

Cặp giả thuyết trên tương đương với cặp giả thuyết sau đây:

$H_0 : (\mu_1 = \mu_2 = ... = \mu_p)$

H_1: (Tồn tại ít nhất một cặp ($i \neq i'$) sao cho $\mu_i \neq \mu_{i'}$)

Bảng số liệu dùng để tính toán phân tích phương sai được xây dựng như sau:

Nhân tố F / STT quan sát	1	2	...	i	...	p
1	x_{11}	x_{12}	.	x_{1i}	.	x_{1p}
2	x_{21}	x_{22}	.	x_{2i}	.	x_{2p}
.	.	.	.	.	.	.
.	.	.	.	.		.
n_i	n_1	n_2	.	n_i	.	n_p

Ví dụ 10.1: Với số liệu đã cho ở bảng 10.1 ta có bảng số liệu dùng để phân tích phương sai trường hợp một nhân tố như sau:

F / STT quan sát	1	2	3	4
1	1,38	1,41	1,32	
2	1,38	1,42	1,33	1,31
3	1,42	1,44	1,34	1,33
4	1,42	1,45		
n_i	n_1=4	n_2=4	n_3=3	n_4 = 2

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phương pháp phân tích phương sai về cơ bản dựa trên cơ sở tính toán và phân tích một số đặc trưng mẫu sau đây:

- TSS (Total sum of squares) là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát X_{ki} và giá trị trung bình mẫu chung của chúng. Ký hiệu:

$$\overline{X_i} = \left(\sum_{k=1}^{n_i} X_{ki}\right)/n_i \quad i = \overline{1,p}$$

$$\overline{X} = \left(\sum_i \sum_k X_{ki}\right)/n \quad n = \sum_i n_i \ , \ i = \overline{1,p}, k = \overline{1,n_i}$$

Ta có:

$$TSS = \sum_i \sum_k (X_{ki} - \overline{X})^2 = \sum_i Q_i - \left[\sum_i T_i\right]^2 / n \qquad (10.2)$$

Trong đó: $Q_i = \sum_k X_{ki}^2 ; T_i = \sum_k X_{ki}$

- MSS (Model sum of squares) là tổng bình phương các sai lệch giữa các giá trị trung bình mẫu của các nhóm quan sát (phân theo mức nhân tố i) và trung bình mẫu chung.

$$MSS = \sum_i (\overline{X_i} - \overline{X})^2 n_i = \sum_i (T_i^2 / n_i) - \left[\sum_i T_i\right]^2 / n \qquad (10.3)$$

- RSS (Residual sum of squares) là tổng của tất cả các tổng bình phương các sai lệch của các giá trị X_{ki} so với trung bình mẫu của từng nhóm:

$$RSS = \sum_i \left(\sum_k (X_{ki} - \overline{X_i})^2\right) = \sum_i Q_i - \sum_i (T_i^2 / n_i) \qquad (10.4)$$

Dễ dàng thấy rằng TSS=MSS+RSS và số bậc tự do của TSS là (n-1) do đó số bậc tự do của MSS và RSS tương ứng là (p-1) và (n-p).

Như vậy, ta thấy toàn bộ sự khác biệt của các giá trị X_{ki} so với $\overline{X}$ được chia làm 2 phần: Một phần được tạo nên bởi nhân tố trong mô hình (MSS) và phần còn lại là do bản thân sự khác biệt giữa các giá trị quan sát tạo ra (RSS) (Đó là sự khác biệt hoàn toàn do ngẫu nhiên mà có).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trở lại ví dụ 10.1, ta lập bảng tính toán sau đây:

Bảng 10.2

F / STT quan sát	1		2		3		4		Kết quả
	x_{11}	x_{k1}^2	x_{12}	x_{k2}^2	x_{13}	x_{k3}^2	x_{14}	x_{k4}^2	
1 2 3 4	1,38 1,38 1,42 1,42	1,9044 1,9044 2,0164 2,0164	1,41 1,42 1,44 1,45	1,9881 2,0164 2,0736 2,1025	1,32 1,33 1,34	1,7424 1,7689 1,7956	1,31 1,33	1,7161 1,7689	$n_1 = n_2 = 4$ $n_3 = 3, n_4 = 2$ $n = \sum n_i = 13$
$T_i = \sum_k X_{ki}$	5,6		5,72		3,99		2,64		$\sum_i T_i = 17,95$
$Q_i = \sum_k X_{ki}^2$		7,7416		8,1806		5,3069		3,485	$\sum_i Q_i = 24,8141$
$\frac{T_i^2}{n_i}$	7,84		8,1796		5,3067		3,4848		$\sum_i \frac{T_i^2}{n_i} = 24,8111$

Từ kết quả tính toán trên và các công thức (10.2) (10.3) (10.4) ta có:

$$TSS = \sum Q_i - \left[\sum T_i\right]^2 / n = 0,029292$$

$$MSS = \sum (T_i^2 / n_i) - \left[\sum T_i\right]^2 / n = 0,026292$$

$$RSS = TSS - MSS = 0,003$$

Để kiểm định cặp giả thuyết của mô hình (10.1):

$$H_0 : (\alpha_i = 0 \forall i, i = \overline{1,p})$$

H_1: (Tồn tại ít nhất một $\alpha_i \neq 0$)

ta sử dụng tiêu chuẩn sau đây:

$$F = \frac{MSS/(p-1)}{RSS/(n-p)}$$

Nếu H_0 đúng, F tuân theo quy luật Fisher-Snedecor với bậc tự do là (p-1) và (n-p). Ta có miền bác bỏ để kiểm định cặp giả thuyết trên là:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$W_\alpha = \left\{ F = \frac{MSS/(p-1)}{RSS/(n-p)}; F > f_\alpha(p-1, n-p) \right\}$$

Trong đó:

$f_\alpha(p-1, n-p)$ là giá trị tới hạn mức α của phân phối F(p-1,n-p).

Đối với ví dụ 10.1 ta có:

$$F_{qs} = \frac{0,26292/(4-1)}{0,003/(13-4)} = \frac{0,08764}{0,000333} = 26,31$$

Chọn $\alpha = 0,05$, tra bảng ta được $f_{0,05}(3,9) = 3,86 \Rightarrow F_{qs} \in W_\alpha$

Vậy ta bác bỏ giả thiết về sự bằng nhau của các giá trị năng suất trung bình của giống cây trồng nói trên được trồng ở bốn nơi khác nhau. Nói một cách khác năng suất cây trồng sẽ khác nhau thực sự (có ý nghĩa) nếu được trồng ở bốn nơi có điều kiện môi trường khác nhau nói trên. Tuy nhiên, như chúng ta đã nhận xét từ đầu, năng suất cây trồng không thể chỉ chịu tác động của một yếu tố về địa lí (nơi trồng) mà còn chịu tác động của nhiều nhân tố khác nữa, ví dụ như: giống, phương pháp canh tác... Vậy nếu chỉ xét 1 nhân tố địa lý (F) thì mức độ tác động của nó đối với sự biến động của năng suất cây trồng (X) là bao nhiêu? Để trả lời câu hỏi này chúng ta xét biểu thức sau đây:

$$TSS = MSS + RSS$$

Chia cả 2 vế cho TSS ta được:

$$1 = 100\% = \frac{MSS}{TSS} + \frac{RSS}{TSS}$$

Đặt $R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS}$

Như vậy ta thấy ý nghĩa của R^2 chính là tỷ lệ hay số phần trăm chiếm trong tổng số 100% của toàn bộ sự sai lệch của X_{ki} so với giá trị trung bình của chúng. R^2 được sử dụng để đo mức độ ảnh hưởng của các nhân tố chứa trong mô hình đối với sự biến động của các giá trị của biến ngẫu nhiên X xung quanh giá trị trung bình của nó. R^2 được gọi là *hệ số xác định* của mô hình phân tích phương sai, đó cũng chính là mức độ thích hợp của mô hình. R^2 càng lớn mô hình càng thích hợp, càng giải thích được nhiều hơn sự biến động của các giá trị biến ngẫu nhiên X dưới tác động của các nhân tố có trong mô hình.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trở lại ví dụ đã nêu trên, chúng ta có:

$$R^2 = \frac{MSS}{TSS} = \frac{0,026292}{0,029292} = 0,8976$$

Điều này có nghĩa là nhân tố nơi trồng (F) ảnh hưởng tới 89,76% đến sự biến động của năng suất cây trồng (89,76% sự khác biệt về năng suất trung bình ở các nơi trồng khác nhau được giải thích bởi yếu tố địa lý-nhân tố F).

Tất cả những kết quả phân tích trên có thể thực hiện một cách đơn giản bởi lệnh anova của Stata. Sau đây là kết quả giải bài toán trên (mô hình 10.1) bằng máy tính để các bạn tham khảo và so sánh. Chú ý kết quả tính bằng tay sẽ có một chút sai số không đáng kể do cách làm tròn số.

```
anova x f
Number of obs =13       R-squared =0.8976
Root MSE =0.18257 Adj R-squared=0.8634
```

Source	Partial SS	df	MS	F	Prob>F
Model	.026292296	3	.008764099	26.29	0.0001
f	.026292296	3	.008764099	26.29	0.0001
Residual	.003000004	9	.000333334		
Total	0.029292299	12	.002441025		

Bài tập

10.1. So sánh mức lợi nhuận trung bình trong kinh doanh của một Công ty tại bốn thị trường khác nhau trên cơ sở bảng số liệu thống kê sau đây:

Thị trường / Quý	1	2	3	4
	x_{k1}	x_{k2}	x_{k3}	x_{k4}
1	7.3	5.4	6.4	7.9
2	7.6	7.1	8.1	8.5
3	8.3	7.4		8.6
4	8.3			

Giải: Đặt X = (Mức lợi nhuận) và F là yếu tố thị trường, ta có mô hình sau đây:

$$X_{ki} = \mu + \alpha_i + U_{ki} \qquad k = \overline{1, n_i} \text{ và } i = \overline{1, p} \ (p = 4).$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Lập bảng tính sau đây:

Stt \ F	1		2		3		4		Kết quả
	x_{k1}	x_{k1}^2	x_{k2}	x_{k2}^2	x_{k3}	x_{k3}^2	x_{k4}	x_{k4}^2	
1	7.3	53.29	5.4	29.16	6.4	40.96	7.9	62.41	$n_1 = 4$
2	7.6	57.76	7.1	50.41	8.1	65.61	8.5	72.25	$n_2 = 3$
3	8.3	68.89	7.4	54.76			8.6	73.96	$n_3 = 2$
4	8.3	68.89							$n_4 = 3$
$T_i = \sum_k x_{ki}$	31.5		19.9		14.5		25		$\sum_i T_i$ = 90.9
$Q_i = \sum_k x_{ki}^2$		248.83		134.33		106.57		208.62	$\sum_i Q_i$ = 698.35
T_i^2 / n_i	248.0625		132.0033		105.125		208.3333		$\sum_i \frac{T_i^2}{n_i}$ =693.524

Từ kết quả tính toán trên ta có:

$$TSS = \sum Q_i - [\sum T_i]^2 / n = 9{,}7825$$

$$MSS = \sum (T_i^2 / n_i) - [\sum T_i]^2 / n = 4{,}9567$$

$$RSS = TSS - MSS = 4{,}8258$$

Xét cặp giả thuyết của mô hình:

H_o: (Mức lợi nhuận trung bình tại bốn thị trường là như nhau)

H_1: (Mức lợi nhuận trung bình tại bốn thị trường là khác nhau)

Ta có miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{MSS/(p-1)}{RSS/(n-p)}; F > f_\alpha(p-1, n-p) \right\}$$

Trong đó: p = 4, n = 12 và với α=5% tra bảng ta được $f_{0,05}(3,8) = 4{,}07$

Thay các giá trị tính toán được ở trên ta có $F_{qs} = 2{,}739 \Rightarrow F_{qs} \notin W_\alpha$

Vậy ta chưa có cơ sở bác bỏ giả thiết H_0 tức là có thể coi mức lợi nhuận trung bình trong kinh doanh của Công ty tại bốn thị trường trên là như nhau.

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{4{,}9567}{9{,}7825} = 0{,}50669$$

Điều này có nghĩa là nhân tố thị trường (F) chỉ ảnh hưởng tới 50,67% sự biến động của lợi nhuận và sự ảnh hưởng này chưa đủ để tạo ra sự khác biệt về lợi nhuận kinh doanh.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

10.2. Người ta đo mức độ bụi trong ba phân xưởng của một nhà máy dệt và được số liệu như sau:

Số thứ tự quan sát	Phân xưởng		
	F1	F2	F3
1	37	60	69
2	47	86	100
3	40	67	98
4	60	92	
5		95	
6		98	

Có thể cho rằng mức độ ô nhiễm bụi của ba phân xưởng là như nhau được không?

Cho kết luận với mức ý nghĩa 5%.

Đáp số:

$MSS_F = 4284$
$RSS = 2160$
$TSS = 6444$

$F_{qs}^F = 9.92$
$R^2 = 0.6648$

10.3. Năng suất của một giống ngô đem trồng ở 3 vùng A, B, C có thực sự khác nhau không? Hãy cho kết luận với mức ý nghĩa 1% trên cơ sở bảng số liệu sau đây:

Số thứ tự quan sát	Vùng		
	A	B	C
1	30,56	43,44	31,36
2	32,66	47,51	36,2
3	34,78	53,8	36,38
4	35,5	42,2	
5	36,63		
6	40,2		
7	42,28		

Đáp số:

$MSS_F = 352.083333$
$RSS = 199.130952$
$TSS = 551.214286$

$F_{qs}^F = 9.72$
$R^2 = 0.6387$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

10.4. So sánh năng suất lúa của 4 trang trại trong một huyện ở đồng bằng, biết rằng gặt ngẫu nhiên ở mỗi trang trại 7 điểm khác nhau người ta thu được bảng số liệu sau đây (tạ/ha):

Số thứ tự quan sát	Trang trại			
	I	II	III	IV
1	51	52	56	54
2	59	58	56	58
3	53	66	58	62
4	59	69	58	64
5	63	70	70	66
6	69	72	74	67
7	72	74	78	69

Đáp số:

$MSS_F = 94.964$
$RSS = 1446$
$TSS = 1540.964$

$F_{qs}^F = 0.53$
$R^2 = 0.616$

10.5. Theo dõi mức tăng trọng của 4 nhóm lợn lai kinh tế được nuôi bằng bốn loại thức ăn tổng hợp khác nhau, người ta thu được bảng số liệu sau đây (kg/tháng):

Số thứ tự quan sát	Nhóm lợn thí nghiệm			
	F1	F2	F3	F4
1	27	24	25	28
2	23	25	30	30
3	29	25	35	24
4	29	28	31	31
5	29	32	22	
6	28	22		
7	25	25		

Có thể xem mức tăng trọng trung bình của lợn ở 4 nhóm là như nhau được không? Cho kết luận với mức ý nghĩa 1%.

Đáp số:

$MSS_F = 26.7704969$
$RSS = 229.664286$
$TSS = 256.434783$

$F_{qs}^F = 0.74$
$R^2 = 0.1044$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3. MÔ HÌNH PHÂN TÍCH PHƯƠNG SAI HAI NHÂN TỐ

Giả sử chúng ta xét biến ngẫu nhiên X có phân phối chuẩn và hai nhân tố F,G. Nhân tố F có p mức ($i=\overline{1,p}$) và nhân tố G có q mức ($j=\overline{1,q}$). Chúng ta có hai mô hình phân tích phương sai sau đây:

3.1. Mô hình hai nhân tố tác động riêng rẽ

Trong mô hình này chúng ta đã giả thiết rằng F và G cùng tác động lên X, nhưng một cách riêng rẽ. Tức là giữa F và G không có mối quan hệ cùng tác động đồng thời lên X. Ta có mô hình sau đây:

$$X_{ij} = \mu + \alpha_i + \beta_j + U_{ij} \qquad i=\overline{1,p}; j=\overline{1,q}$$

Trong đó: α_i, β_j là các hằng số đặc trưng cho sự khác biệt về giá trị trung bình μ của X dưới tác động của hai nhân tố F và G tương ứng ở mức i và j.

U_{ij} là các sai số ngẫu nhiên độc lập với nhau và cùng tuân theo quy luật phân phối chuẩn $N(0,\sigma_u^2)$.

Với loại mô hình này ta có bảng số liệu phân tích phương sai như sau:

F \ G	1	2	…	j	…	q
1	X_{11}	X_{12}	…	X_{1j}	…	X_{1q}
.	.	.		.		.
.	.	.		.		.
i	X_{i1}	X_{i2}		X_{ij}		X_{iq}
.	.	.		.		.
.	.	.		.		.
p	X_{p1}	X_{p2}		X_{pj}		X_{pq}

Dễ dàng nhận thấy chúng ta có pq nhóm, trong mỗi một nhóm chỉ có một quan sát. Nếu trong một nhóm có nhiều quan sát ta lấy giá trị trung bình của nhóm đó làm đại diện.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ký hiệu:

$$\overline{X} = (\sum_i \sum_j X_{ij})/pq$$

$$\overline{X_i} = (\sum_j X_{ij})/q = \frac{X_{i\bullet}}{q}$$

$$\overline{X_j} = (\sum_i X_{ij})/p = \frac{X_{j\bullet}}{p}$$

Ta có tổng bình phương các sai lệch chung của toàn bộ quan sát so với giá trị trung bình mẫu là:

$$TSS = \sum_i \sum_j (X_{ij} - \overline{X})^2 = \sum_{ij} X_{ij}^2 - pq(\overline{X})^2 \qquad (10.5)$$

Vế phải có thể biến đổi như sau:

$$\sum_i \sum_j (X_{ij} - \overline{X})^2 = \sum_i q(\overline{X_i} - \overline{X})^2 + \sum_j p(\overline{X_j} - \overline{X})^2 + \sum_i \sum_j (X_{ij} - \overline{X_i} - \overline{X_j} + \overline{X})^2$$

Trong đó tổng bình phương các sai lệch mà trung bình mẫu các nhóm so với trung bình mẫu chung (tức là sai lệch do việc phân chia thành các nhóm bởi hai nhân tố F và G) là:

$$MSS = \sum_i q(\overline{X_i} - \overline{X})^2 + \sum_j p(\overline{X_j} - \overline{X})^2$$

$$= \left(\frac{\sum X_{i\bullet}^2}{q} - \frac{(\sum X_{ij})^2}{pq} \right) + \left(\frac{\sum X_{j\bullet}^2}{p} - \frac{(\sum X_{ij})^2}{pq} \right)$$

$$= MSS_F + MSS_G \qquad (10.6)$$

Trong đó: MSS_F là tổng bình phương các sai lệch về trung bình của các nhóm được tạo nên bởi nhân tố F và MSS_G là tổng bình phương các sai lệch về trung bình của các nhóm được tạo nên bởi nhân tố G.

Phần còn lại chính là tổng bình phương các sai lệch được tạo nên bởi các sai số ngẫu nhiên:

$$RSS = \sum_i \sum_j (X_{ij} - \overline{X_i} - \overline{X_j} + \overline{X})^2$$

$$= TSS - MSS_F - MSS_G \qquad (10.7)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bậc tự do của TSS là (qp-1), của MSS_F là (p-1), của MSS_G là (q-1), do đó bậc tự do của RSS là:

$$pq - 1 - (p - 1) - (q - 1) = (p - 1)(q - 1).$$

Ví dụ 10.2:

Tệp số liệu cho trong ví dụ này có nội dung trong bảng 10.3.

Bảng 10.3

Contains data from C:\WINSTATA\XSTK10_2.DTA

obs:	12	
vars:	3	15 Aug 2010
10:37		
size:	84 (86.2% of memory free)	
1. x	byte %8.0g	chi phi hoc hanh
2. f	byte %8.0g	thanh phan gia dinh
3. g	byte %8.0g	vung

Sorted by:

	X	f	g
1.	22	1	1
2.	17	2	1
3.	28	3	1
4.	34	4	1
5.	21	1	2
6.	15	2	2
7.	27	3	2
8.	33	4	2
9.	31	1	3
10.	19	2	3
11	30	3	3
12.	35	4	3

Trong đó: X là chi phí học hành bình quân cho một đứa trẻ đi học của một gia đình. Hai nhân tố, F là thành phần gia đình (F = 1: Công nhân, F = 2: Nông dân, F = 3: Trí thức, F = 4: Buôn bán) và G là vùng, nơi cư trú (G = 1: Bắc, G = 2: Trung; G = 3: Nam). Từ tệp số liệu này ta có bảng số liệu phân tích phương sai hai nhân tố tác động riêng rẽ như sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

F \ G	1	2	3
1	22	21	31
2	17	15	19
3	28	27	30
4	34	33	35

Để giải bài toán trên, chúng ta lập bảng tính sau đây:

F \ G	1	2	3	$\sum_j x_{ij}$	$\overline{x_i}$	$(\overline{x_i})^2$
1	22	21	31	74	24.66667	608.4444
2	17	15	19	51	17	289
3	28	27	30	85	28.33333	802.7778
4	34	33	35	102	34	1156
$\sum_i x_{ij}$	101	96	115	$\sum_{i,j} x_{ij}$ = 312		$\sum_i (\overline{x_i})^2$ = 2856.222
$\overline{x_j}$	25.25	24	28.75		$\overline{x}$ = 26	
$(\overline{x_j})^2$	637.5625	576	826.5625	$\sum_j (\overline{x_j})^2$ =	2040.125	$(\overline{x})^2$ = 676
$\sum_i (x_{ij})^2$	2713	2484	3447	$\sum_{i,j} (x_{ij})^2$ =	8644	

Từ những kết quả tính được ở bảng trên thay vào công thức (10.5), (10.6), (10.7) ta có:

$$MSS_F = \frac{25706}{3} - \frac{(312)^2}{12} = 456{,}66$$

$$MSS_G = \frac{32642}{4} - \frac{(312)^2}{12} = 48{,}5$$

$$TSS = 8644 - \frac{(312)^2}{12} = 532$$

Vậy ta suy ra: RSS = 532 - 456,66 - 48,5 = 26,84

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Kiểm định giả thuyết của mô hình

a. Cặp giả thuyết

$H_0^F : (\alpha_i = 0 \quad \forall i, i = \overline{1,p})$

H_1^F : (Có ít nhất một $\alpha_i \neq 0$)

Nếu H_0^F đúng thì $F = \dfrac{MSS_F/(p-1)}{RSS/[(p-1)(q-1)]}$ sẽ tuân theo quy luật Fisher-Snedecor với bậc tự do là $(p-1)$ và $(p-1)(q-1)$

Do đó miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{MSS_F/(p-1)}{RSS/[(p-1)(q-1)]}; F > f_\alpha[(p-1),(p-1)(q-1)] \right\}$$

Đối với ví dụ 10.2 thì:

$$F_{qs} = \frac{456,66/(4-1)}{26,84/[(4-1)(3-1)]} = 34,04$$

Chọn $\alpha = 0,05$ tra bảng ta được $f_{0,05}(3,6) = 4,76 \rightarrow F_{qs} \in W_\alpha$.

Ta bác bỏ giả thuyết H_0^F, có nghĩa là thành phần gia đình (nhân tố F) có ảnh hưởng đến mức chi tiêu cho học hành của một gia đình.

b. Cặp giả thuyết

$H_0^G : (\beta_j = 0 \quad \forall j, j = \overline{1,q})$

H_1^G : (Có ít nhất một $\beta_j \neq 0$)

Tương tự ta có miền bác bỏ:

$$W_\alpha = \left\{ F = \frac{MSS_G/(q-1)}{RSS/[(p-1)(q-1)]}; F > f_\alpha[(q-1),(p-1)(q-1)] \right\}$$

Đối với ví dụ 10.2 thì:

$$F_{qs} = \frac{48,5/(3-1)}{26,84/6} = 5,42$$

Tra bảng ta có $f_{0,05}(2,6) = 5,14 \rightarrow F_{qs} \in W_\alpha$ do đó ta bác bỏ giả thuyết H_0^G. Điều này có nghĩa là chúng ta có thể cho rằng nhân tố vùng cũng có ảnh hưởng đến mức chi tiêu cho học hành của một gia đình.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sau đây là kết quả giải bài toán phân tích phương sai hai nhân tố tác động riêng rẽ trên cơ sở tệp số liệu XSTK10_2 đã nêu trên bằng Stata.

.anova x f g

Number of obs =12 R-squared =0.9496

Root MSE =2.11476 Adj R-squared=0.9075

Source	Partial SS	df	MS	F	Prob>F
Model	505.166667	5	101.033333	22.59	0.0008
f	456.666667	3	152.222222	34.04	0.0004
g	48.50	2	24.25	5.42	0.0452
Residual	26.8333333	6	4.47222222		
Total	532.00	11	48.3636364		

Cũng như đối với mô hình phân tích phương sai một nhân tố, ta có hệ số xác định của mô hình phân tích phương sai hai nhân tố tác động riêng rẽ là R^2 được xác định như sau:

$$R^2 = \frac{MSS}{TSS} = \frac{MSS_F + MSS_G}{TSS}$$

Theo ví dụ 10.2 thì:

$$R^2 = \frac{456,66 + 48,5}{532} = \frac{505,15}{532} = 0,9495$$

Có nghĩa là mô hình đưa ra đã giải thích được 94,95% sự biến động của giá trị chi phí cho học hành trung bình của một gia đình dưới tác động của hai nhân tố là thành phần gia đình (F) và vùng cư trú (G). Trong đó F đóng góp 85,83% và G đóng góp 9,12% vào sự biến động của X.

3.2. Mô hình phân tích phương sai hai nhân tố tác động tổng hợp

Trong mô hình này ở mỗi một nhóm tương ứng với một cặp mức nhân tố (i,j) ta cần phải có số quan sát lớn hơn 1. Giả sử ở tất cả các nhóm ta đều có số quan sát là như nhau và bằng m (còn gọi là mẫu cân bằng) thì ta có mô hình hai nhân tố tác động tổng hợp như sau:

$$X_{kij} = \mu + \alpha_i + \beta_j + \delta_{ij} + U_{kij}$$

$$k = \overline{1,m};\ \ i = \overline{1,p};\ \ j = \overline{1,q}$$

Trong đó α_i, β_j là các hằng số đặc trưng cho sự khác biệt về giá trị trung bình μ của X do tác động của hai nhân tố F và G, còn δ_{ij} là hằng số

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

đặc trưng cho sự khác biệt đó nhưng được gây nên bởi tác động tổng hợp của hai nhân tố (F,G) ở mức (i,j). Ta cũng giả thiết $X \sim N(\mu, \sigma_x^2)$ và các sai số ngẫu nhiên U_{kij} độc lập với nhau, cùng tuân theo quy luật phân phối chuẩn $N(\mu, \sigma_u^2)$

Ký hiệu:

$$\overline{X} = \left(\sum_{k,i,j} X_{kij}\right) / mqp$$

$$\overline{X_i} = \left(\sum_{k,j} X_{kij}\right) / mq = X_{i\bullet} / mq$$

$$\overline{X_j} = \left(\sum_{k,i} X_{kij}\right) / mp = X_{j\bullet} / mp$$

$$\overline{X_{ij}} = \left(\sum_{k} X_{kij}\right) / m = X_{ij\bullet} / m$$

Xét tương tự như mục 3.1 ta có:

$$TSS = \sum_{k,i,j} (X_{kij} - \overline{X})^2 = \sum_{k,i,j} X_{kij}^2 - mpq(\overline{X})^2$$

$$= \sum X_{kij}^2 - \frac{\left(\sum X_{kij}\right)^2}{mpq}$$

$$MSS_F = \sum_i mq(\overline{X_i} - \overline{X})^2 = \sum_i mq\left(\overline{X_i}\right)^2 - mpq(\overline{X})^2$$

$$= \frac{\sum X_{i\bullet}^2}{mq} - \frac{\left(\sum X_{kij}\right)^2}{mpq}$$

$$MSS_G = \sum_j mp(\overline{X_j} - \overline{X})^2 = \sum_j mp\left(\overline{X_j}\right)^2 - mpq(\overline{X})^2$$

$$= \frac{\sum X_{j\bullet}^2}{mp} - \frac{\left(\sum X_{kij}\right)^2}{mpq}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$RSS = \sum_{k,i,j}(X_{kij} - \overline{X}_{ij})^2 = \sum_{k,i,j} X_{kij}^2 - m\sum_{i,j}(\overline{X}_{ij})^2$$

$$= \sum X_{kij}^2 - \frac{\sum X_{ij\bullet}^2}{m} \qquad (10.8)$$

$$MSS_{F*G} = TSS - MSS_F - MSS_G - RSS$$

Bậc tự do của TSS là $(mpq-1)$, của MSS_F là $(p-1)$, của MSS_G là $(q-1)$, RSS là $pq(m-1)$ và của MSS_{FxG} là $(p-1)(q-1)$

+ Phân tích sự tác động của hai nhân tố F và G tương đương với việc kiểm định các cặp giả thuyết sau đây:

a. $H_0^F : (\alpha_i = 0 \quad \forall i, i = \overline{1,p})$

H_1^F: (Có ít nhất một $\alpha_i \neq 0$)

Miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{MSS_F/(p-1)}{RSS/[pq(m-1)]}; F > f_\alpha[p-1, pq(m-1)] \right\}$$

b. $H_0^G : (\beta_j = 0 \quad \forall j, j = \overline{1,q})$

H_1^G: (Có ít nhất một $\beta_j \neq 0$)

$$W_\alpha = \left\{ F = \frac{MSS_G/(q-1)}{RSS/[pq(m-1)]}; F > f_\alpha[q-1, pq(m-1)] \right\}$$

c. $H_0^{F*G} : (\sigma_{ij} = 0 \quad \forall i,j : i = \overline{1,p}; j = \overline{1,q})$

H_1^{F*G}: (Tồn tại ít nhất một cặp (i,j): $\delta_{ij} \neq 0$)

$$W_\alpha = \left\{ F = \frac{MSS_{F*G}/[(p-1)(q-1)]}{RSS/[pq(m-1)]}; F > f_\alpha[(p-1)(q-1), pq(m-1)] \right\}$$

+ Tương tự như phần 3.1 chúng ta cũng có hệ số xác định của mô hình trong trường hợp này là:

$$R^2 = \frac{MSS}{TSS} = \frac{MSS_F + MSS_G + MSS_{F*G}}{TSS}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ví dụ 10.3. Chúng ta xét lại các chỉ tiêu và nhân tố như đã nêu ở ví dụ 10.2. Đó là: X = (Chi phí giáo dục bình quân cho một đứa trẻ đi học trong một gia đình). F = (Thành phần gia đình) và G = (Nơi cư trú - vùng). Khác với số liệu ở ví dụ 10.2, ở đây ứng với mỗi mức nhân tố chúng ta có hai quan sát (tức là số liệu của hai gia đình). Như vậy, trong mô hình phân tích phương sai hai nhân tố tác động tổng hợp ta có:

$$i=1,2,3,4;\ j=1,2,3;\ m=1,2.$$

Sau đây là bảng số liệu cụ thể:

G \ F	1	2	3	4
1	22;21	17;19	28;25	34;33
2	21;25	15;18	27;25	33;31
3	31;31	19;20	30;33	35;33

Trước hết chúng ta lập bảng tính toán số liệu trung gian như sau (xem bảng 10.4)

Bảng 10.4

G \ F	1	2	3	4	$X_{j\bullet}$	$X^2_{j\bullet}$
1	22;21	17;19	28;25	34;33	199	39601
2	21;25	15;18	27;25	33;31	195	38025
3	31;31	19;20	30;33	35;33	232	53824
$X_{i\bullet}$	151	108	168	199	$\sum X_{kij}=626$	$\sum X^2_{j\bullet}=131450$
$\sum_{k,j} X^2_{kij}$	3913	1960	4752	6609	$\sum X^2_{kij}=17234$	
$X^2_{i\bullet}$	22801	11664	28224	39601	$\sum X^2_{i\bullet}=102290$	
$\sum_{j} X^2_{ij\bullet}$	7809	3906	9482	13209	$\sum_{ij} X^2_{ij\bullet}=34406$	

Theo các công thức (10.8) ta có:

$$TSS = 17234 - \frac{626^2}{24} = 905,83$$

$$MSS_F = \frac{102290}{6} - \frac{626^2}{24} = 720,17$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$MSS_G = \frac{131450}{8} - \frac{626^2}{24} = 103,08$$

$$RSS = 17234 - \frac{34406}{2} = 31$$

$$\Rightarrow MSS_{F*G} = 905,83 - 720,17 - 103,08 = 51,58$$

Giả sử chúng ta muốn kiểm định xem tồn tại hay không sự tác động tổng hợp của hai nhân tố F và G. Xét cặp giả thiết ở mục C, theo kết quả tính toán ở trên ta có:

$$F_{qs} \frac{MSS_{F*G}/[(p-1)(q-1)]}{RSS/[pq(m-1)]} = \frac{51,58/6}{31/12} = 3,33$$

Với mức ý nghĩa $\alpha = 0,05$, tra bảng chúng ta được $f_{0,05}^{(6,12)} = 3$. Vì $F_{qs} > f_{0,05}^{(6,12)}$ nên ta bác bỏ giả thuyết H_0^{F*G}. Tức là chúng ta có thể cho rằng trên thực tế có tồn tại sự tác động lẫn nhau của hai yếu tố thành phần gia đình và vùng đối với việc chi tiêu cho trẻ em đi học.

Chú ý: Bài toán phân tích phương sai trong trường hợp mô hình tổng quát hơn như số nhân tố lớn hơn hai hoặc số quan sát trong mỗi nhóm không như nhau cũng đã giải quyết một cách đầy đủ (xem tài liệu tham khảo M.G.Kendall, A.Stuart). Tuy nhiên đối với tất cả các bài toán phân tích phương sai việc giải nó trên máy tính là một việc hoàn toàn không có gì khó khăn.

Ví dụ, nếu sử dụng Stata thì để phân tích sự ảnh hưởng của các nhân tố F,G,L đối với biến X chúng ta chỉ cần gõ lệnh:

anova x f g l f*g f*l g*l

Đối với ví dụ 10.3 nói trên, kết quả chạy bằng Stata như sau:

.anova x f g f*g

Number of obs =24 R-squared =0.9658

Root MSE =1.60728 Adj R-squared=0.9344

Source	Partial SS	df	MS	F	Prob>F
Model	874.83333	11	79.530303	30.79	0.0000
f	720.16667	3	0240.055556	92.92	0.0000
g	103.08333	2	51.5416667	19.95	0.0002
f*g	51.583333	6	8.59722222	3.33	0.0362
Residual	31.00	12	2.58333333		
Total	905.83333	23	39:384058		

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

10.6. Hãy phân tích sự biến động của thu nhập ($/tháng/người) trên cơ sở số liệu điều tra về thu nhập trung bình của người lao động thuộc 4 loại hình dịch vụ chăm sóc xe ô tô của các đại lý của ba công ty liên doanh sản xuất ô tô:

Loại hình dịch vụ (F)	Liên doanh(G)		
	1	2	3
1.	212	200	230
2.	222	205	222
3.	241	250	245
4.	240	228	230

Giải: Đặt X = (Thu nhập bình quân) và với giả thiết rằng F và G cùng tác động lên X, nhưng một cách riêng rẽ ta có mô hình:

$$X_{ki} = \mu + \alpha_i + \beta_j + U_{ki} \qquad i = \overline{1,p}\ ;\ j = \overline{1,q}\ \ (p = 4 \text{ và } q = 3)$$

Lập bảng tính sau đây:

F \ G	1	2	3	$\sum_j x_{ij}$	$\overline{x_i}$	$(\overline{x_i})^2$
1	212	200	230	642	214.000	45796.0000
2	222	205	222	649	216.3333	46800.111
3	241	250	245	736	245.333	60188.444
4	240	228	230	698	232.6667	54133.778
$\sum_i x_{ij}$	915	883	927	$\sum_{i,j} x_{ij} = 2725$		$\sum_i (\overline{x_i})^2 =$ 206918.33
$\overline{x_j}$	228.75	220.75	231.75		$\overline{x} = 227.0833$	
$(\overline{x_j})^2$	52326.562	48730.563	53708.062	$\sum_j (\overline{x_j})^2 =$	154765.188	$(\overline{x})^2 =$ 51566.84
$\sum_i (x_{ij})^2$	209909	196509	215109	$\sum_{i,j} (x_{ij})^2 =$	621527	

Từ bảng tính trên ta có:

$$TSS = \sum_{i,j,k} (X_{kij} - \overline{X})^2 = \sum_{k,i,j} X_{kij}^2 - mpq(\overline{X})^2 = 2724{,}918$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$MSS_F = \sum_i mq(\bar{X}_i - \bar{X})^2 = \sum_i mq(\bar{X}_i)^2 - mpq(\bar{X})^2 = 1952{,}917$$

$$MSS_G = \sum_j mp(\bar{X}_j - \bar{X})^2 = \sum_j mq(\bar{X}_j)^2 - mpq(\bar{X})^2 = 258{,}667$$

$RSS = TSS - (MSS_F + MSS_G) = 513{,}333$

Kiểm định giả thuyết của mô hình.

a. Cặp giả thuyết.

H_o^F: ($\alpha_i = 0$ với $\forall i = \overline{1,p}$)

H_1^F: (Có ít nhất một $\alpha_i \neq 0$)

Miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{MSS_F/(p-1)}{RSS/[(p-1)(q-1)]}; F > f_\alpha(p-1),(p-1)(q-1) \right\}$$

$F_{qs} = 7{,}6088$

Chọn $\alpha = 0{,}05$ tra bảng ta được $f_{0,05}(3,6) = 4{,}76 \rightarrow Fqs \in W_\alpha$.

Ta bác bỏ thuyết H_o^F, có nghĩa là loại hình dịch vụ khác nhau (nhân tố F) có ảnh hưởng đến mức thu nhập của người lao động.

b. Cặp giả thuyết:

H_o^G: ($\beta_j = 0$ với $\forall_j = \overline{1,q}$)

H_1^G: (có ít nhất một $\beta_j \neq 0$

Tương tự ta có miền bác bỏ:

$$W_\alpha = \left\{ F = \frac{MSS_G/(p-1)}{RSS/[(p-1)(q-1)]}; F > f_\alpha(q-1),(p-1)f(q-1) \right\}$$

Fqs = 1,5117. Tra bảng ta có $f_{0,05}(2,6) = 5{,}14 \rightarrow Fqs \notin W_\alpha$ do đó ta chưa có cơ sở bác bỏ giả thuyết H_o^G. Điều này có nghĩa là nếu cùng làm một công việc thì ở liên doanh nào thu nhập trung bình cũng như nhau.

Cũng như đối với mô hình phân tích phương sai một nhân tố, ta có hệ số xác định của mô hình phân tích phương sai hai nhân tố tác động riêng rẽ là R^2 là:

$$R^2 = \frac{MSS}{TSS} = \frac{MSS_F + MSS_G}{TSS} = 0{,}8116$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Có nghĩa là mô hình đưa ra đã giải thích được 81,16% sự biến động của thu nhập, trong đó phần quyết định chủ yếu do loại hình dịch vụ và các đơn giá cho các công việc tiến hành ở các đại lý thuộc các liên doanh có thể coi là tương đối đồng đều.

10.7. Sau đây là số liệu về số lượng một loại báo ngày bán được ở 5 quận nội thành:

Ngày khảo sát	Quận nội thành				
	Q1	Q2	Q3	Q4	Q5
Thứ hai	22	18	22	18	18
Thứ ba	21	18	22	18	19
Thứ tư	25	25	25	19	20
Thứ năm	24	24	18	20	22
Thứ sáu	28	19	15	22	25
Thứ bảy	30	22	28	25	25

Lượng báo bán được ở 5 quận có khác nhau thực sự không? Lượng báo bán ra có chịu tác động của yếu tố ngày trong tuần không? Chọn $\alpha = 15\%$.

Đáp số:

$MSS_F = 141,5$

$MSS_G = 78,53$

$RSS = 158,67$

$TSS = 378,7$

$F_{qs}^F = 0,0181$

$F_{qs}^G = 2,47$

$R^2 = 0,581$

10.8. Theo dõi số học sinh đến lớp muộn của năm trường PTTH vào các ngày khác nhau trong tuần, người ta thu được số liệu về số lượng học sinh trung bình đến lớp muộn của các trường đó vào một số ngày tiêu biểu trong tuần như sau:

Ngày trong tuần	Trường PTTH			
	A	B	C	D
Thứ hai	5	4	5	7
Thứ tư	4	5	3	2
Thứ sáu	4	3	4	5
Thứ bảy	4	4	3	2

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bạn có nhận xét gì về số lượng học sinh đến lớp muộn của các trường? Có sự khác biệt gì về số lượng học sinh đến lớp muộn vào các ngày khác nhau trong tuần?

Đáp số:

$MSS_F = 15,6$ $\quad F_{qs}^F = 3,92$

$MSS_G = 3,7$

$RSS = 15,9$ $\quad F_{qs}^G = 0,7$

$TSS = 35,2$ $\quad R^2 = 0,5483$

10.9. So sánh chi phí cho ba loại dịch vụ ở ba thành phố khác nhau bằng phương pháp phân tích phương sai trên cơ sở bảng số liệu sau đây:

Thành phố	Loại dịch vụ		
	I	II	III
I	61	52	69
II	58	51	61
III	68	64	79

Các con số trong ô là chi phí trung bình cho 1 lần dịch vụ (đơn vị: 1000đ)

Đáp số:

$MSS_F = 296.222222$ $\quad F_{qs}^F = 24.92$

$MSS_G = 294.222222$

$RSS = 23.7777778$ $\quad F_{qs}^G = 24.75$

$TSS = 614.22222$ $\quad R^2 = 0.9613$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4. TÓM TẮT CHƯƠNG 10

1. Phân tích phương sai là phân tích sự biến động của một biến ngẫu nhiên do tác động của một hay nhiều nhân tố. Nhân tố ở đây được hiểu là các yếu tố, điều kiện khách quan hoặc chủ quan có tác động trực tiếp đến sự biến động của biến ngẫu nhiên mà ta nghiên cứu.
2. Tùy theo *số lượng nhân tố và cách tác động* của chúng lên chỉ tiêu nghiên cứu (biến ngẫu nhiên) mà người ta có những mô hình phân tích phương sai khác nhau: mô hình phân tích phương sai một nhân tố, hai nhân tố, hai nhân tố tác động tổng hợp...
3. Phương pháp phân tích phương sai về cơ bản dựa trên cơ sở tính toán và phân tích các đặc trưng mẫu: TSS, MSS, RSS. Để kiểm định sự khác biệt về mức độ tác động của một nhân tố hay sự tác động tổng hợp của nhiều nhân tố khác nhau trong mô hình người ta dùng tiêu chuẩn F. Để đánh giá mức độ phù hợp của mô hình chúng ta dùng hệ số xác định R^2.
4. Việc giải bài toán phân tích phương sai trong trường hợp mô hình tổng quát hơn như số nhân tố lớn hơn hai, số quan sát sát trong mỗi nhóm không như nhau, số quan sát lớn... đều có thể làm được nhanh chóng và đơn giản bằng cách sử dụng các phần mềm thống kê như STATA, SPSS...

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các ký hiệu và công thức cơ bản

* Mô hình phân tích phương sai một nhân tố

$X_{ij} = \mu + \alpha_i + U_{ki}$ $\quad i = \overline{1,p};\ k = \overline{1,n_i};\ n = \sum n_i$

Thành phần	Tổng bình phương các sai lệch	Bậc tự do	F
F	$MSS = \sum \frac{X_{i\bullet}^2}{n_i} - \frac{\left(\sum X_{ki}\right)^2}{n}$	$p-1$	$F = \frac{MSS/(p-1)}{RSS(n-p)}$
U	$RSS = \sum X_{ki}^2 - \frac{\sum X_{i\bullet}^2}{n_i}$	$n-p$	
$\sum$	$TSS = \sum X_{ki}^2 - \frac{\left(\sum X_{ki}\right)^2}{n}$	$n-1$	

* Mô hình phân tích phương sai hai nhân tố

a. Hai nhân tố tác động riêng rẽ

$X_{ij} = \mu + \alpha_i + \beta_j + U_{ij}$ $\quad i = \overline{1,p};\ j = \overline{1,q};\ n = p.q$

Thành phần	Tổng bình phương các sai lệch	Bậc tự do	F
F	$MSS_F = \sum \frac{X_{i\bullet}^2}{q} - \frac{\left(\sum X_{ij}\right)^2}{pq}$	$p-1$	$F_1 = \frac{MSS_F/(p-1)}{RSS/[(p-1)(q-1)]}$
G	$MSS_G = \sum \frac{X_{j\bullet}^2}{p} - \frac{\left(\sum X_{ij}\right)^2}{pq}$	$(q-1)$	$F_2 = \frac{MSS_G/(q-1)}{RSS/[(p-1)(q-1)]}$
U	$RSS = TSS - MSS_F - MSS_G$	$(p-1)(q-1)$	
$\sum$	$TSS = \sum X_{ij}^2 - \frac{\left(\sum X_{ij}\right)^2}{pq}$	$pq-1$	

b. Hai nhân tố tác động tổng hợp

$X_{kij} = \mu + \alpha_i + \beta_j + \delta_{ij} + U_{kij}$

$i = \overline{1,p};\ \ j = \overline{1,q}\ \ k = \overline{1,m};\ n = mpq$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thành phần	Tổng bình phương các sai lệch	Bậc tự do	F
F	$MSS_F=\frac{\sum X_{i\bullet}^2}{mq}-\frac{\left(\sum X_{kij}\right)^2}{mpq}$	$p-1$	$F_1=\frac{MSS_F/(p-1)}{RSS/[pq(m-1)]}$
G	$MSS_G=\frac{\sum X_{j\bullet}^2}{mp}-\frac{\left(\sum X_{kij}\right)^2}{mpq}$	$(q-1)$	$F_2=\frac{MSS_G/(q-1)}{RSS/[pq(m-1)]}$
F*G	$MSS_{F*G}=TSS-MSS_F$ $-MSS_G-RSS$	$(p-1)\times$ $(q-1)$	$F_3=\frac{MSS_{F*G}/[(p-1)(q-1)]}{RSS/[pq(m-1)]}$
U	$RSS=\sum X_{kij}^2-\frac{\sum X_{ij\bullet}^2}{m}$	$pq(m-1)$	
$\sum$	$TSS=\sum X_{kij}^2-\frac{\left(\sum X_{kij}\right)^2}{mpq}$	$mpq-1$	

* Lệnh phân tích phương sai dùng trong Stata

+ Phân tích phương sai một nhân tố (nhân tố f):

[**by** (dsb):] ***an*ova** x f [if...] [in...] [, các lựa chọn]

+ Phân tích phương sai hai nhân tố (f và g) tác động riêng rẽ:

[**by** (dsb):] ***an*ova** x f g [if...] [in...] [, các lựa chọn]

+ Phân tích phương sai hai nhân tố - tác động tổng hợp:

[**by** (dsb):] ***an*ova** x f g f*g [if...] [in...] [, các lựa chọn]

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Phân tích phương sai - ANOVA	analysis of variance
Bảng phân tích phương sai	ANOVA table
Nhân tố	Factor
Mô hình hiệu quả cố định	fixed effects models
Mô hình hiệu quả ngẫu nhiên	random effects models
Mô hình một nhân tố	one factor models
Mô hình hai nhân tố	multifactor models
Phân tích phương sai bảng ngẫu nhiên hai chiều	Twoo-Way Analysis of Variance
Bảng ngẫu nhiên hai chiều	Twoo-Way Table

Câu hỏi ôn tập

10.10. Hãy cho biết sự khác nhau giữa hai bài toán kiểm định giả thuyết về phương sai và phân tích phương sai.

10.11. Hãy nêu sự khác nhau của hai tiêu chuẩn χ^2 và F khi nghiên cứu về sự phân tán của tổng thể.

10.12. Mô hình phân tích phương sai của một nhân tố và mô hình phân tích phương sai hai nhân tố khác nhau cơ bản ở điểm nào?

10.13. Bảng số liệu phân tích phương sai và bảng ngẫu nhiên hai chiều có điều gì giống nhau và khác nhau?

10.14. Hãy giải thích về số bậc tự do của TSS, RSS và các MSS trong phân tích phương sai.

10.15. Có thể giải thích như thế nào khi F_G càng lớn thì sự tác động của nhân tố G càng có ý nghĩa và ngược lại.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập tổng hợp chương 10

10.16. Theo dõi doanh thu của 4 cửa hàng thuộc một công ty (triệu đồng/tháng), người ta thu được số liệu như sau:

Tháng kinh doanh	Cửa hàng			
	1	2	3	4
1	12.3	14.2	15.6	17.2
2	12.6	12.4	17.1	15.8
3	11.6	11.5	18.2	12.2
4	15.2	11.6	12.5	
5	18.6		11.8	
6	17.1			

Hãy so sánh doanh thu trung bình/ tháng của các cửa hàng thuộc công ty nói trên.

Đáp số:

MSS_F = 18.9899496

RSS = 90.3795091

TSS = 109.369459

F_{qs}^{F} = 0.98

R^2 = 0.1736

10.17. Người ta tiến hành đo mực nước sông tại một số địa điểm thuộc tỉnh X trong cùng một ngày (số lần đo không giống nhau) và thu được bảng số liệu sau đây:

Số thứ tự quan sát	Địa điểm đo			
	F1	F2	F3	F4
1	5.5	4.9	4.6	4.5
2	5.6	5.1	4.8	6.2
3	5.8	6.5	5.8	4.8
4	5.9	5.4	5.1	4.8
5	6.0	6.1	6.2	6.5
6	6.7		7.1	6.8
7	7.2			

Mực nước sông trung bình/ ngày của các điểm nói trên có thực sự khác nhau không?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Đáp số:

MSS_F = 1.23958311
RSS = 13.7999983
TSS = 15.0395814

F_{qs}^{F} = 0.6
R^2 = 0.0824

10.18. So sánh lãi suất của một loại cổ phiếu được đầu tư vào 5 khu vực khác nhau trên cơ sở bảng số liệu thống kê sau đây:

Tháng quan sát	Khu vực				
	I	II	III	IV	V
1	0.5	0.9	0.5	0.3	0.2
2	0.5	0.8	0.5	0.4	0.6
3	0.6	0.5	0.6	0.7	0.4
4	0.8	0.5	0.8	0.7	0.5
5	0.4	0.7	0.7	0.9	0.7
6	0.7			0.5	0.8

Đáp số:

MSS_F = 0.06364285
RSS = 0.785999967
TSS = 0.849642817

F_{qs}^{F} = 0.47
R^2 = 0.0824

10.19. Kiểm tra sức khỏe của 29 công nhân ở 5 phân xưởng của nhà máy sản xuất pin-ắc quy, người ta đo được mức độ nhiễm chì của họ như sau:

Số thứ tự quan sát	Mức nhân tố				
	F1	F2	F3	F4	F5
1	0.25	0.22	0.25	0.31	0.22
2	0.28	0.25	0.26	0.33	0.28
3	0.32	0.24	0.28	0.30	0.28
4	0.22	0.28	0.25	0.29	0.25
5	0.22	0.31	0.22	0.25	0.30
6		0.21	0.28		
7		0.22	0.31		

So sánh mức độ nhiễm chì đối với công nhân ở các phân xưởng của nhà máy nói trên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Đáp số:

MSS_F = 1.23958311

RSS = 13.7999983

TSS = 15.0395814

F_{qs}^{F} = 0.6

R^2 = 0.0824

10.20. Hãy phân tích vai trò ngành nghề (chính, phụ) trong hoạt động kinh tế của các hộ gia đình ở một vùng nông thôn trên cơ sở bảng số liệu về thu nhập trung bình của một hộ tương ứng với các ngành nghề nói trên như sau:

Nghề chính	Nghề phụ			
	(1)	(2)	(3)	(4)
Trồng lúa (1)	3,5	7,4	8,0	3,5
Trồng cây ăn quả (2)	5,6	4,1	6,1	9,6
Chăn nuôi (3)	4,1	2,5	1,8	2,1
Dịch vụ (4)	7,2	3,2	2,2	1,5

Đáp số:

Number of obs = 16 R- squared = 0.4126

Root MSE = 2.46464 Adj R- squared = 0.0210

Source	Partial SS	df	MS	F	Prob > F
Model	38.400001	6	6.40000016	1.05	0.4529
f	36.385z001	3	12.1283339	2.00	0.1851
g	6	3	671666447	0.11	0.9517
Residual	2.01499934	9	6.07444472		
	54.6700024				
Total	93.0700034	15	6.2046669		

10.21. Hãy phân tích tình hình kinh doanh của một số ngành nghề ở 4 quận nội thành trên cơ sở số liệu về doanh thu của một số cửa hàng như sau:

Đơn vị: 100 triệu đồng

Ngành nghề kinh doanh	Khu vực kinh doanh			
	Q1	Q2	Q3	Q4
Điện lạnh	2,5;2,7;2,0;3,0	3,1;3,5;2,7	2,0;2,4	5,0;5,4
Vật liệu xây dựng	0,6;10,4	15,0	9,5;9,3;9,1	19,5;17,5
Dịch vụ tin học	1,2;1,0;9,8;1,8	2,0;2,2;1,8	1,2;1,3;1,2	5,0;4,8;5,2

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Đáp số:

MSS_F = 297.02167
MSS_G = 52.669164
RSS = 16.3783334
TSS = 366.069167

$F_{qs}^{F} = 54.41$

$F_{qs}^{G} = 6.43$

$R^2 = 0.9553$

10.22. Cho bảng số liệu về chi phí giáo dục bình quân cho một đứa trẻ đi học của một gia đình được thống kê theo hai nhân tố là thành phần gia đình (F = 1: công nhân, F = 2: nông dân, F = 3: trí thức; F = 4: buôn bán) và G là nơi cư trú (G = 1: miền Bắc, G = 2: miền Trung; G = 3: miền Nam) như sau:

F \ G	1	2	3
1	22	21	31
2	17	15	19
3	28	27	30
4	34	33	35

Hãy phân tích sự ảnh hưởng của thành phần gia đình và yếu tố địa lý (môi trường xã hội) đến chi phí dành cho trẻ đi học trong một gia đình.

Giải: Đặt X = (Chi phí bình quân cho một đứa trẻ đi học trong một gia đình) và với giả thiết rằng F và G cùng tác động lên X, nhưng một cách riêng rẽ. Tức là giữa F và G không có mối quan hệ cùng tác động đồng thời lên X. Ta có mô hình:

$$X_{ki} = \mu + \alpha_i + \beta_j + U_{ki} \qquad i = \overline{1,p}\ ;\ j = \overline{1,q}\ \ (p = 4 \text{ và } q = 3)$$

Lập bảng tính như sau:

F \ G	1	2	3	$\sum_j x_{ij}$	$\overline{x_i}$	$(\overline{x_i})^2$
1	22	21	31	74	24.66667	608.4444
2	17	15	19	51	17	289
3	28	27	30	85	28.33333	802.7778
4	34	33	35	102	34	1156

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

F \ G	1	2	3	$\sum_j x_{ij}$	$\overline{x_i}$	$(\overline{x_i})^2$
$\sum_i x_{ij}$	101	96	115	$\sum_{i,j} x_{ij} =$ 312		$\sum_i (\overline{x_i})^2 =$ 2856.222
$\overline{x_j}$	25.25	24	28.75		$\overline{x} = 26$	
$(\overline{x_j})^2$	637.5625	576	826.5625	$\sum_j (\overline{x_j})^2 =$	2040.125	$(\overline{x})^2 = 676$
$\sum_i (x_{ij})^2$	2713	2484	3447	$\sum_{i,j} (x_{ij})^2 =$	8644	

Từ những tính toán trên đây chúng ta có:

$$TSS = \sum_{i,j,k} (X_{kij} - \bar{X})^2 = \sum_{k,i,j} X_{kij}^2 - mpq(\bar{X})^2 = 532$$

$$MSS_F = \sum_i mq(\bar{X}_i - \bar{X})^2 = \sum_i mq(\bar{X}_i)^2 - mpq(\bar{X})^2 = 456{,}6667$$

$$MSS_G = \sum_j mp(\bar{X}_j - \bar{X})^2 = \sum_j mq(\bar{X}_j)^2 - mpq(\bar{X})^2 = 48{,}5$$

$$RSS = TSS - (MSS_F + MSS_G) = 26{,}8333$$

Kiểm định giả thuyết của mô hình.

a. Cặp giả thuyết.

H_o^F: ($\alpha_i = 0$ với $\forall i = \overline{1, p}$)

H_1^F: (Có ít nhất một $\alpha_i \neq 0$)

Miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{MSS_F/(p-1)}{RSS/[(p-1)(q-1)]}; F > f_\alpha(p-1),(p-1)(q-1) \right\}$$

$F_{qs} = 34.0373$

Chọn $\alpha = 0{,}05$ tra bảng ta được $f_{0,05}(3,6) = 4{,}76 \rightarrow Fqs \in W_\alpha$.

Ta bác bỏ thuyết H_o^F, có nghĩa là thành phần gia đình (nhân tố F) có ảnh hưởng đến mức chi tiêu cho học hành của một gia đình.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b. Cặp giả thuyết:

H_o^G: ($\beta_j = 0$ với $\forall_j = \overline{1,q}$)

H_1^G: (có ít nhất một $\beta_j \neq 0$

Tương tự ta có miền bác bỏ:

$$W_\alpha = \left\{ F = \frac{MSS_G/(p-1)}{RSS/[(p-1)(q-1)]}; F > f_\alpha(q-1),(p-1)f(q-1) \right\}$$

Fqs = 5,4224. Tra bảng ta có $f_{0,05}$ (2,6) = 5,14 $\rightarrow$ Fqs $\in W_\alpha$ do đó ta bác bỏ giả thuyết H_o^G. Điều này có nghĩa là chúng ta có thể cho rằng nhân tố vùng cũng có ảnh hưởng đến mức chi tiêu cho học hành của một gia đình.

Cũng như đối với mô hình phân tích phương sai một nhân tố, ta có hệ số xác định của mô hình phân tích phương sai hai nhân tố tác động riêng rẽ là R^2 được xác định như sau:

$$R^2 = \frac{MSS}{TSS} = \frac{MSS_F + MSS_G}{TSS} = 0,9496$$

Có nghĩa là mô hình đưa ra đã giải thích được 94,94% sự biến động của chi phí bình quân cho một đứa trẻ đi học trong một gia đình dưới tác động của hai nhân tố là thành phần gia đình (F) và vùng cư trú (G). Trong đó F đóng góp 85,83% và G đóng góp 14,17% vào sự biến động của X.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội.1999.
2. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB ĐH Kinh tế quốc dân. Hà Nội. 2008.
3. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà Nội.2009.
4. Ngô Văn Thứ. *Thống kê thực hành*. NXB Khoa học và kỹ thuật. Hà Nội. 2008.
5. Гмурман ВЕ Теория вероятностей и математическая статистика. Москва. 1972.
6. Уйякс. математическая статистика. Москва. 1982.
7. Bencrézi J.P. Analyse des données. Dunod. Paris. 1973
8. Michel Volle: Analyse des donnees. Economica. 1994.
9. Kendall M and Stuart A, The advanced theory of Statistics. California. 1993
10. OTT L.R. And introduction to statistical methods and data analysis. California. 1993
11. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons. New york. 1990

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 11
PHÂN TÍCH TƯƠNG QUAN VÀ HỒI QUY

1. MỞ ĐẦU

Khác với các chương trước, từ chương này trở đi chúng ta sẽ nghiên cứu trực tiếp nhiều chỉ tiêu khác nhau đối với một đối tượng nghiên cứu. Các chỉ tiêu này có thể là định lượng (tức là các biến ngẫu nhiên) hay định tính (được gọi là các dấu hiệu hay nhân tố như đã đề cập ở phần phân tích phương sai). Các chỉ tiêu này được xác định cụ thể trên đối tượng quan sát (cần phân biệt đối tượng nghiên cứu và đối tượng quan sát) và được "đo lường" (tức là thu nhận thông tin) trực tiếp trên các đơn vị quan sát. Đó là các phần tử thuộc tập hợp đối tượng quan sát mà ta sẽ trực tiếp thu nhận thông tin từ đó. Ví dụ đối tượng quan sát là lao động trong các ngành công nghiệp chẳng hạn thì đơn vị quan sát là các cán bộ, công nhân đang làm việc trong ngành công nghiệp. Tùy theo vấn đề cần nghiên cứu, phân tích mà các chỉ tiêu có thể là tuổi, thời gian làm việc, năng suất lao động... (chỉ tiêu định lượng)... hay giới tính, tay nghề, loại hình nghề nghiệp, trình độ văn hóa... (chỉ tiêu định tính).

Như vậy để giải quyết vấn đề cụ thể người ta phải nghiên cứu, phân tích nhiều chỉ tiêu (đôi khi còn gọi là chỉ tiêu thống kê) đó là các biến ngẫu nhiên (chỉ tiêu định lượng), các nhân tố, dấu hiệu (chỉ tiêu định tính). Một phương pháp hữu hiệu giúp chúng ta có thể làm rõ hơn bản chất của hiện tượng hay sự việc cần nghiên cứu để tìm ra quy luật, dự đoán được xu thế biến động của hiện tượng, sự việc đó trong tương lai, đó là phương pháp phân tích mối quan hệ phụ thuộc giữa các chỉ tiêu phản ánh nội dung của hiện tượng, sự việc cần nghiên cứu. Cụ thể chúng ta sẽ đề cập đến những vấn đề sau:

1. Mức độ phụ thuộc giữa các biến ngẫu nhiên, giữa các nhân tố, dấu hiệu (có tồn tại sự phụ thuộc giữa chúng không? Mức độ phụ thuộc đó như thế nào nếu chúng tồn tại).

2. Mối phụ thuộc đó thuộc loại nào? Trong thống kê người ta phân biệt hai loại phụ thuộc sau đây:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ *Định nghĩa 1:* Hai biến ngẫu nhiên X và Y được gọi là phụ thuộc hàm nếu tồn tại hàm f sao cho Y=f(X).

+ *Định nghĩa 2:* Hai biến ngẫu nhiên X, Y được gọi là có phụ thuộc thống kê nếu với mỗi giá trị của X ta đều có thể xác định được quy luật phân phối xác suất có điều kiện của Y đối với X là:

$$F(y/x) = P(Y<y/X = x)$$

3. Xác định biểu thức của hàm mô tả mối phụ thuộc nói trên giữa các biến ngẫu nhiên.

Mục 2 sẽ nghiên cứu nội dung của bài toán phân tích tương quan.

Mục 3 sẽ nghiên cứu nội dung của bài toán phân tích hồi quy.

2. PHÂN TÍCH TƯƠNG QUAN

2.1. Phân tích tương quan bảng số liệu định lượng

Chúng ta tiến hành phân tích tương quan chủ yếu dựa trên cơ sở phân tích hai đặc trưng cơ bản là hệ số tương quan (trường hợp hai biến ngẫu nhiên) và hệ số tương quan riêng phần (trường hợp nhiều hơn hai biến ngẫu nhiên).

1. Phân tích hệ số tương quan

Giả sử X và Y là hai biến ngẫu nhiên có V(X)>0 và V(Y)>0. Hệ số tương quan của hai biến ngẫu nhiên là X và Y, ký hiệu là ρ_{xy} và được xác định như sau (xem Chương IV, §6).

$$\rho_{xy} = \frac{E(X - E(X))(Y - E(Y))}{\sqrt{V(X)}\sqrt{V(Y)}}$$

Như chúng ta đã biết có thể dùng ρ_{xy} để đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên. $|\rho|$ càng lớn thì sự phụ thuộc tuyến tính càng rõ. Trường hợp đặc biệt $\rho_{xy} = 0$ thì ta nói giữa X và Y không có sự tương quan với nhau. Nếu X và Y độc lập với nhau thì $\rho_{xy} = 0$, điều ngược lại chưa chắc đã đúng. Tuy nhiên đối với biến ngẫu nhiên có phân phối chuẩn thì sự độc lập và không tương quan là tương đương nhau. Trong trường hợp

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

tổng quát cần phân biệt hai khái niệm độc lập và không tương quan giữa hai biến ngẫu nhiên.

Định nghĩa. Giả sử ta có $\{(X_i, Y_i)\}$ $i = \overline{1,n}$ là một mẫu ngẫu nhiên hai chiều thu được khi quan sát vectơ ngẫu nhiên (X, Y) thì hệ số tương quan mẫu R_{xy} của X và Y được xác định như sau:

$$R_{xy} \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n}} \sqrt{\frac{\sum_i (Y_i - \overline{Y})^2}{n}}} \qquad (11.1)$$

$$\frac{\overline{XY} - \overline{Y}\,\overline{Y}}{\sqrt{\overline{X}^2 - (\overline{X})^2}\sqrt{\overline{Y}^2 - (\overline{Y})^2}} = \frac{\overline{XY} - \overline{Y}\,\overline{Y}}{\sqrt{MS_X}\sqrt{MS_Y}}$$

Trong đó:

$$\overline{X} = \frac{\left(\sum_i X_i\right)}{n}; \quad \overline{X}^2 = \frac{\left(\sum_i X_i^2\right)}{n}$$

$$\overline{Y} = \frac{\left(\sum_i Y_i\right)}{n}; \quad \overline{Y}^2 = \frac{\left(\sum_i Y_i^2\right)}{n}$$

$$\overline{XY} = \frac{\left(\sum_i X_i Y_i\right)}{n}$$

Ví dụ 1. Chúng ta xét tệp số liệu có tên là XSTK11_1 về một số chỉ tiêu phát triển kinh tế Việt Nam trong thời kỳ 1990-2006 (đơn vị tính 1000 tỷ đồng VN năm 1999). Nội dung của tệp số liệu đó như sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bảng 11.1

• d

Contains data from C:\WINSTATA\XSTK10_1.DTA

obs:	17	
vars:	6	11 Aug 2010 19:02
size:	476 (85.9% of memory free)	

1.nam	float	%9.0g
2.gdp	float	%9.2g Tong gia tri san pham trong nuoc
3.ex	float	%9.2g Tong gia tri hang hoa xuat khau
4.im	float	%9.2g Tong gia tri hang hoa nhap khau
5.gip	float	%9.2g Tong gia tri hang hoa cong nghiep
6.gap	float	%9.0g Tong gia tri hang hoa nong nghiep

Sorted by:

• List

	nam	gdp	ex	im	gip	gap
1.	1990	16.80	.30	6.80	6.80	9.7
2.	1991	17.20	.40	6.90	6.90	10.1
3.	1992	18.70	.50	7.50	7.50	11.2
4.	1993	20.10	.60	8.50	8.50	11.5
5.	1994	21.80	.60	9.60	9.60	12.2
6.	1995	23.00	.70	10.50	10.50	12.5
7.	1996	23.80	.80	11.20	11.20	13.1
8.	1997	24.70	.90	12.30	12.30	13.1
9.	1998	25.90	1.00	14.00	14.00	13.7
10.	1999	28.00	1.90	13.60	13.60	14.7
11.	2000	29.50	2.40	14.00	14.00	14.9
12.	2001	31.30	2.10	15.50	15.50	15.4
13.	2002	34.00	2.60	18.10	18.10	16.6
14.	2003	36.70	3.00	20.40	20.40	17.6
15.	2004	40.00	4.10	23.20	23.20	18.6
16.	2005	43.80	5.37	26.50	26.50	19.4
17.	2006	47.90	7.10	30.23	30.23	20.35

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ký hiệu Y = (gdp) và X = (gap), để tính hệ số tương quan (mẫu) r_{xy} ta lập bảng tính toán sau đây:

STT	y_i	x_i	y_i^2	x_i^2	$x_i y_i$
1	16.8	9.70	282.24	94.0900	162.9600
2	17.2	10.10	295.84	102.0100	173.7200
3	18.7	11.20	349.69	125.4400	209.4400
4	20.1	11.50	404.01	132.2500	231.1500
5	21.8	12.20	475.24	148.8400	265.9600
6	23.0	12.50	529.00	156.2500	287.5000
7	23.8	13.10	566.44	171.6100	311.7800
8	24.7	13.10	610.09	176.6100	323.5700
9	25.9	13.70	670.81	187.6900	354.8300
10	28.0	14.70	784.00	216.0900	411.6000
11	29.5	14.90	870.25	222.0100	439.5500
12	31.3	15.40	979.69	237.1600	482.0200
13	34.0	16.60	1150.00	275.5600	564.4000
14	36.7	17.60	1346.89	309.7600	645.9200
15	40.0	18.60	1600.00	345.9600	744.0000
16	43.8	19.40	1918.44	376.3600	849.7200
17	47.9	20.35	2294.41	4144.1225	974.7650
$\sum$	483.2	244.65	15133.04	3686.8125	7432.8850

Từ kết quả trên thu được:

$$\bar{x} = 14{,}3912\,,\quad \overline{x^2} = 216{,}8713$$

$$\bar{y} = 28{,}4235\,,\quad \overline{y^2} = 890{,}1788\,,\quad \overline{xy} = 437{,}2285$$

Áp dụng công thức (11.1) ta được:

$$r_{xy} = \frac{437{,}2285 - 14{,}3912.28{,}4235}{\sqrt{216{,}8713 - 14{,}3912^2}\sqrt{890{,}1788 - 28{,}4235^2}}$$

$$= \frac{28{,}180}{3{,}1248.9{,}07} = 0{,}9942$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trường hợp nếu chúng ta cần xét p biến ngẫu nhiên $X_1, X_2, ..., X_p (p > 2)$ ta ký hiệu $r_{x_i y_i} = r_{ij}$, khi đó ta có tập hợp $\{r_{ij}\}; i, j = \overline{1,p}$ là ma trận hệ số tương quan của p biến ngẫu nhiên nói trên. Dễ dàng thấy rằng $\{r_{ij}\}$ là ma trận đối xứng và các phần tử trên đường chéo là $r_{ii} = 1$. Với số liệu đã cho ở bảng 11.1 Stata cho ta ma trận hệ số tương quan của các biến gdp, ex, im, gip, gap như sau:

corr gdp ex im gip gap

Bảng 11.2

	gdp	ex	im	gip	gap
gdp	1.0000				
ex	0.9653	1.0000			
im	0.8926	0.9605	1.0000		
gip	0.9942	0.9720	0.9244	1.0000	
gap	0.9942	0.9375	0.8508	0.9815	1.0000

Qua bảng hệ số tương quan trên ta thấy các chỉ tiêu (biến ngẫu nhiên) nói trên có sự tương quan với nhau rất mạnh, hầu hết hệ số tương quan giữa các cặp đều lớn hơn 0,9. Điều đó chứng tỏ rất có khả năng giữa chúng tồn tại mối quan hệ phụ thuộc tuyến tính.

a. Khoảng tin cậy cho hệ số tương quan ρ_{xy}

Fisher đã chứng minh rằng ngay với những giá trị n (kích thước mẫu) không lớn lắm thống kê Z sau đây:

$$Z = \frac{1}{2}\ln\frac{1+R_{xy}}{1-R_{xy}}$$

cũng có phân phối xấp xỉ chuẩn với

$$E(Z) = \frac{1}{2}\ln\frac{1+\rho_{xy}}{1-\rho_{xy}} + \frac{\rho_{xy}}{2(n-1)} \text{ và } V(Z) = \frac{1}{n-3}$$

Từ đây ta có khoảng tin cậy $(1-\alpha)$ của $\frac{1}{2}\ln\frac{1+\rho_{xy}}{1-\rho_{xy}}$ là:

$$\left(Z - \frac{R_{xy}}{2(n-1)} - u_{\alpha/2}\frac{1}{\sqrt{n-3}}; Z - \frac{R_{xy}}{2(n-1)} + u_{\alpha/2}\frac{1}{\sqrt{n-3}}\right)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tra bảng các giá trị của hàm $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ sẽ suy ra khoảng tin cậy cho hệ số tương quan ρ_{xy}.

b. Kiểm định giả thuyết về giá trị của ρ

Xét cặp giả thuyết:

$$H_0 : (\rho_{xy} = \rho_0)$$

$$H_1 : (\rho_{xy} \neq \rho_0)$$

Sử dụng thống kê Z nêu trên ta thấy nếu H_0 đúng thì:

$$U = \left(Z - \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1)} \right)\sqrt{n-3}$$

sẽ có phân phối xấp xỉ chuẩn hoá N(0,1). Do đó miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ U = \left(Z - \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1)} \right)\sqrt{n-3}; |U| > u_{\alpha/2} \right\}$$

Trường hợp đặc biệt ta giả thiết X và Y đều tuân theo quy luật phân phối chuẩn. Xét cặp giả thuyết

$$H_0 : (\rho_{XY} = 0)$$

$$H_1 : (\rho_{XY} \neq 0)$$

Với kích thước mẫu bất kỳ ta luôn có:

$$T = R_{xy}\sqrt{\frac{n-2}{1-R_{XY}^2}}$$

tuân theo quy luật phân phối Student với bậc tự do $(n-2)$.

Vậy miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ t = R_{xy}\sqrt{\frac{n-2}{1-R_{XY}^2}}; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2. Phân tích hệ số tương quan riêng phần

Xét tập hợp biến ngẫu nhiên $X_1, X_2, ..., X_p$. Hệ số tương quan riêng phần của X_1 và X_2 khi X_3 cố định ký hiệu là $R_{12,3}$. Khi đó ta có công thức xác định $R_{12,3}$ như sau:

$$R_{12,3} = \frac{R_{12} - R_{13}R_{23}}{\sqrt{(1-R_{13}^2)(1-R_{23}^2)}}$$

Tổng quát ta có hệ số tương quan riêng phần của X_i và X_j khi cố định X_k là $R_{ij,k}$:

$$R_{ij,k} = \frac{R_{ij} - R_{ik}R_{jk}}{\sqrt{(1-R_{ik}^2)(1-R_{jk}^2)}} \qquad (11.2)$$

Hệ số tương quan riêng phần giữa X_1 và X_2 khi cố định X_3, X_4 là:

$$R_{12,34} = \frac{R_{12,3} - R_{14,3}R_{24,3}}{\sqrt{(1-R_{14,3}^2)(1-R_{24,3}^2)}}$$

Tương tự ta sẽ có công thức xác định hệ số tương quan riêng phần giữa X_1 với X_2 khi cố định $X_3, X_4, ...X_p$ là:

$$R_{12,34...p} = \frac{R_{12,34...(p-1)} - R_{1p,34...(p-1)}R_{2p,34...(p-1)}}{\sqrt{(1-R_{1p,34...(p-1)}^2)(1-R_{2p,34...(p-1)}^2)}}$$

Trên thực tế khi cần nghiên cứu một tập biến ngẫu nhiên $X_1, X_2, ..., X_p$ thì chúng ta thường quan tâm đến hệ số tương quan riêng phần của một biến nào đó (thông thường là biến mà chúng ta đang muốn quan tâm phân tích, giải thích sự biến động của nó) với tất cả các biến còn lại. Ví dụ đó là biến X_1 chẳng hạn thì để đơn giản chúng ta chỉ cần viết:

$R_{12,}$ ... (Hệ số tương quan riêng phần của X_1 và X_2 khi $X_3, X_4, ...X_p$ cố định).

...

$R_{1p,}$ ... (Hệ số tương quan riêng phần của X_1 và X_p khi $X_2, X_3, ...X_{p-1}$ cố định).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ví dụ 2. Căn cứ vào ma trận hệ số tương quan (bảng 11.3) chúng ta sẽ tính hệ số tương quan riêng phần của $X_1 = (gdp)$ với $X_2 = (ex)$ và $X_3 = (im)$. Theo công thức (11.2) ta có:

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = \frac{0,9653 - 0,8926.0,9605}{\sqrt{(1-0,8926^2)(1-0,9605^2)}}$$

$$= \frac{0,1079577}{0,1254624} = 0,8605$$

$$r_{13,2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}} = \frac{0,8926 - 0,9653.0,9805}{\sqrt{(1-0,9653^2)(1-0,9605^2)}}$$

$$= \frac{-0,03457}{0,07267} = -0,4757$$

Với Stata chúng ta có thể dễ dàng tính được các hệ số tương quan riêng phần bằng lệnh pcorr sau đây:

pcorr x_1 x_2 x_3 ... x_p

Ở đây có nghĩa là tính hệ số tương quan riêng phần của x_1 với lần lượt các biến x_2, x_3, ... x_p khi cố định các biến còn lại. Theo như ký hiệu ở trên ta sẽ có tất cả các r_{1i},... với $i = \overline{2,p}$.

Sau đây là kết quả chạy bằng Stata cho tệp số liệu XSTK11_1.

- pcor gdp ex im

(obs =17)

Partial correlation of gdp with

Variable	Corr.	sig.
ex	0.8610	0.000
im	-0.4771	0.062

- pcorr gdp ex im gip gap

(obs=17)

Partial correlation of gdp with

Variable	Corr.	sig.
ex	0.7958	0.001
im	-0.5830	0.029
gip	0.8024	0.001
gap	0.8514	0.000

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Theo kết quả nêu trên ta thấy nếu cố định một chỉ tiêu ví dụ (im) chẳng hạn thì hệ số tương quan riêng phần của gdp với ex là 0,8610, còn khi cố định (ex) thì hệ số tương quan riêng phần của gdp với im lại là một số âm: -0,4771 (ở đây có một chút sai số so với tính tay vì làm tròn số).

Ta cũng thấy hệ số tương quan riêng phần của gdp với ex khi cố định các chỉ tiêu còn lại là 0,7958. Dễ dàng nhận thấy rằng, nếu cố định các chỉ tiêu còn lại thì hệ số tương quan riêng phần của gdp với gap là cao nhất và cũng rất lớn: 0,8514. Kết hợp với phương pháp hồi quy, các bạn sẽ thấy chúng ta có thể rút ra được nhiều nhận xét, kết luận rất thú vị.

2.2. Phân tích tương quan bảng số liệu định tính

1. Bảng ngẫu nhiên hai chiều

Bảng này chủ yếu dùng cho trường hợp các chỉ tiêu định tính được "đo" bằng thang đo phân loại. Như các bạn đã biết trong chương 9 chúng ta đã xét bài toán kiểm định sự độc lập của hai dấu hiệu (chỉ tiêu định tính) A và B. Tuy nhiên nếu A và B không độc lập với nhau thì tiêu chuẩn χ^2 không cho chúng ta biết mức độ tương quan giữa A và B.

Với những bảng ngẫu nhiên hai chiều cấp (2 x 2) dạng:

A \ B	b_1	b_2	$\sum$
a_1	n_1	n_2	(n_1+n_2)
a_2	n_3	n_4	(n_3+n_4)
$\sum$	(n_1+n_3)	(n_2+n_4)	$n_1+n_2+n_3+n_4=n$

Chúng ta có thể đo mối tương quan giữa A và B bằng các hệ số Q và F được xác định như sau:

Hệ số Q:
$$Q = \frac{n_1n_4 - n_2n_3}{n_1n_4 + n_2n_3} \qquad (11.3)$$

Hệ số F:
$$F = \frac{n_1n_4 - n_2n_3}{\sqrt{(n_1+n_2)(n_1+n_3)(n_2+n_4)(n_3+n_4)}}$$

Ví dụ 3. Trên cơ sở bảng số liệu điều tra về thái độ của sinh viên đối với việc chấp hành nội quy thì (tệp số liệu XSTK11_2) chúng ta có bảng ngẫu nhiên hai chiều sau đây:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bảng 11.3a

- tab thái độ giới tính

Chấp hành nội quy thi	Giới tính		Total
	Nam	Nữ	
0	40	40	80
1	20	0	20
Total	60	40	100

Bảng 11.3b

- tab thái độ vùng

Chấp hành nội quy thi	Nông thôn hay thành thị		Total
	nt	tt	
0	43	37	80
1	8	12	20
Total	51	49	100

Trong đó A: (thái độ) có hai giá trị $a_1 = 0$ là không vi phạm, $a_2 = 1$ là có vi phạm, B = (giới tính) có hai giá trị $b_1 = 0$ là nữ, $b_2 = 1$ là nam và C = (vùng) có hai loại: nt là nông thôn, tt là thành thị. Như vậy riêng đối với chỉ tiêu C chúng ta không "đo" bằng thang đo phân loại mà vẫn giữa nguyên các ký hiệu phân loại.

Phân tích bảng 11.3a ta có:

$$Q = \frac{40.0 - 40.20}{40.0 + 40.20} = -1$$

$$F = \frac{40.0 - 40.20}{80.60.20.40} = -0.4$$

Nhận xét: Ta nhận thấy giá trị của Q và F được tính trên cùng một bảng số liệu lại rất khác nhau: Giá trị của Q lớn hơn hai lần giá trị của F (về giá trị tuyệt đối). Do đâu mà như vậy? Trước hết ta thấy rằng các hệ số Q và F đo các mặt khác nhau của mối liên hệ trong bảng bốn ô (bảng ngẫu nhiên 2 chiều cấp (2x2). Trong ví dụ này hệ số $|Q| = +1$ vì tất cả các trường hợp vi

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

phạm nội quy trong suốt quá trình khảo sát (điều tra) đều là các sinh viên nam. Song sẽ hoàn toàn không đúng nếu giải thích Q có nghĩa là hầu như tất sinh viên nam đều vi phạm nội quy thi.

Hệ số F cho phép phản ánh mức độ liên hệ qua lại hai chiều giữa các dấu hiệu nghiên cứu, trong lúc đó hệ số Q chỉ phản ánh mối liên hệ một chiều. Giá trị của các hệ số chỉ trùng nhau khi có mối liên hệ qua lại hai chiều hoàn toàn.

Nếu $n_1 = n_4 = 0$ thì F=-1. Nếu $n_2 = n_3 = 0$ thì F=+1. Chúng ta có thể dễ dàng kiểm tra ý nghĩa của F thông qua tiêu chuẩn χ^2 vì $\chi^2 = nF^2$.

Đối với bảng ngẫu nhiên hai chiều cấp (p x q). Giả sử chúng ta đã tính được giá trị của tiêu chuẩn χ^2 (tức là giá trị χ^2_{qs}) và đã kết luận được hai dấu hiệu A và B không độc lập với nhau. Khi đó, chúng ta có thể đo mối liên hệ (mức độ tương quan) giữa hai dấu hiệu đó bằng hệ số liên hợp Pearson (P)

$$P = \sqrt{\frac{\chi^2}{n + \chi^2}} \qquad 0 \le P \le 1 \qquad (11.4)$$

Ví dụ 4. Nghiên cứu tình trạng hôn nhân trước ngày cưới của 542 cặp vợ chồng ta có bảng số liệu sau:

Tình trạng hôn nhân của vợ / Tình trạng hôn nhân của chồng	Chưa kết hôn lần nào	Ly hôn	Góa	$\sum$
Chưa kết hôn lần nào	180	34	36	250
Ly hôn	58	76	54	188
Góa	43	34	27	104
$\sum$	281	144	117	542

Ta tính được $\chi^2_{qs} = 79,96; \chi^2_{0,05}(4) = 9,5 \rightarrow$ Bác bỏ tính độc lập giữa tình trạng hôn nhân của vợ (A) và tình trạng hôn nhân của chồng (B) trước ngày cưới.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vậy: $P = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{79,96}{542+79,76}} = 0,358$

P = 0 khi χ^2=0, tức là khi có sự độc lập hoàn toàn giữa các dấu hiệu. Nhược điểm của hệ số P là giá trị cực đại của nó phụ thuộc vào quy mô của bảng ngẫu nhiên hai chiều. Do đó nẩy sinh những khó khăn nhất định cho việc giải thích và về thực chất nên coi P như một chỉ số bằng số chứ không phải như một độ đo. Để khắc phục nhược điểm trên ta có hệ số Kramer (K)

$$K = \left\{\frac{\chi^2}{n.\min(p-1,q-1)}\right\}^{1/2}$$

Hệ số K bao giờ cũng có thể đạt +1 không phụ thuộc vào dạng của bảng ngẫu nhiên hai chiều.

Theo ví dụ 11.4 ta có:

$$K = \left\{\frac{79,96}{542.\min(2,2)}\right\}^{1/2} = 0,272$$

Như vậy, ta có thể nói rằng tình trạng hôn nhân của vợ và chồng trước khi cưới có phụ thuộc vào nhau nhưng mức độ phụ thuộc không lớn lắm. Có lẽ đó chỉ là yếu tố mang tính chất tham khảo chứ không mang tính chất quyết định.

2. Các độ đo sự tương quan ở cấp đo theo thang thứ bậc

Trước hết, ta xét một ví dụ. Cho tệp số liệu tên gọi là XSTK 11_3 với nội dung như sau:

Bảng 11.4

Contains data from: C\WINSTATA\XSTK11_3.DTA
Obs: 10
Vars: 2 17 Aug 2010 15:26
Size: 60(86.0% of memory free)

1. diem 1 byte %8.0g Diem cua giam khao nguoi Canada
2. diem 2 byte %8.0g Diem cua giam khao nguoi Nhat

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sorted by:

diem 1	95	90	86	84	75	70	62	60	57	50
diem 2	92	93	83	80	55	60	45	72	62	70

Trong đó A= (diem 1), B= (diem 2) là các điểm do hai giám khảo người Canada và Nhật chấm cho một vận động viên trong một cuộc thi quốc tế về trượt băng nghệ thuật.

Các giá trị A và B được đo theo thang đo thứ bậc. Bây giờ ta sẽ xếp hạng cho các giá trị của A theo thứ tự (tăng hoặc giảm dần, sau đó ta gán các hạng cho các giá trị đó đúng bằng số thứ tự đã sắp xếp.

A = (diem 1)	95	90	86	84	75	70	62	60	57	50
Hạng (i)	1	2	3	4	5	6	7	8	9	10

Làm tương tự như đối với A ta có các hạng của B như sau:

B = (diem 2)	93	92	83	80	72	70	62	60	55	45
Hạng (i)	1	2	3	4	5	6	7	8	9	10

Tiếp theo ta sắp xếp các hạng của A và B theo đúng thứ tự các giá trị của A và B đã cho ở bảng 11.4.

Hạng A:i	1	2	3	4	5	6	7	8	9	10
Hạng B: k_i	2	1	3	4	9	8	10	5	7	6

Như vậy là tương ứng với mỗi cặp giá trị (a_i, b_i) của (A,B) ta có một cặp hạng tương ứng là (i, k_i)

+ Hệ số tương quan hạng r_s của Spearman cho hai dấu hiệu A và B được đo cùng thang đo thứ bậc xác định như sau:

$$R_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \qquad (11.5)$$

Trong đó $d_i = i - k_i$ là hiệu giữa hai giá trị của cặp hạng (i, k_i) $(i = \overline{1,n})$, n là số cặp hạng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ký hiệu ρ_s là hệ số tương quan hạng Spearman của tổng thể thì để kiểm định cặp giả thuyết:

$$H_0:(\rho_s=0),\ H_1:(\rho_s\neq 0)$$

Ta sử dụng tiêu chuẩn

$$T=\sqrt{\frac{R_s^2(n-2)}{1-R_s^2}}$$

Nếu H_0 đúng thì T có phân phối Student với (n-2) bậc tự do. Do đó, ta có miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha=\left\{t=\sqrt{\frac{R_s^2(n-2)}{1-R_s^2}};|t|>t_{\alpha/2}^{(n-2)}\right\}$$

Ví dụ 5. Sử dụng số liệu đã cho ở bảng 11.4 ta tính hệ số tương quan hạng giữa hai dãy điểm số đã cho của hai giảm khảo người Canada và Nhật.

Theo cách xếp hạng trên ta được:

$$\sum_i d_i^2=1+1+16+4+9+9+4+16=60$$

$$r_s=1-\frac{6\sum d_i^2}{n(n^2-1)}=1-\frac{6.60}{10(10^2-1)}=0,6364$$

Để kiểm định xem cách cho điểm của hai giám khảo nói trên có độc lập với nhau hay không ta tính:

$$t_{qs}=\sqrt{\frac{r_s^2(n-2)}{1-r_s^2}}=\sqrt{\frac{0,6364^2.8}{1-0,6364^2}}=2,356$$

Với $\alpha=0,05$ tra bảng ta được $t_{0,025}^{(8)}=2,306, t_{qs}>t_{0,025}^{(8)}$. Ta có thể bác bỏ giả thuyết $H_0:(\rho_s=0)$. Tuy nhiên, ta thấy giá trị quan sát t_{qs} chênh lệch so với giá trị tới hạn quá ít và nếu chọn mức ý nghĩa 0,01 thì giá trị tới hạn $t_{\alpha/2}^{(8)}=t_{0,005}^{(8)}=3,335$ thì $t_{qs}\notin W_\alpha$, tức là ta không có cơ sở để bác bỏ giả thuyết H_0. Trong những trường hợp như vậy, ta có thể phân tích thêm bằng hệ số tương quan hạng Kendall.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Hệ số tương quan hạng Kendall ký hiệu là τ và được xác định bằng công thức sau:

$$\tau = \frac{S}{\frac{n(n-1)}{2}} \qquad (11.6)$$

Việc tính giá trị của S được minh họa qua ví dụ cụ thể sau đây.

Trên cơ sở số liệu đã xếp hạng về điểm số của hai giám khảo người Canada và Nhật trong ví dụ 11.5 ta lập bảng tính S như sau:

Hạng của A	Hạng của B	S_i^+	S_i^-	$S_i^+ - S_i^-$
1	2	8	1	7
2	1	8	0	8
3	3	7	0	7
4	4	6	0	6
5	9	1	4	-3
6	8	1	3	-2
7	10	0	3	-3
8	5	2	0	2
9	7	0	1	-1
10	6	-	-	-
				$S = \sum(S_i^+ - S_i^-) = 21$

Cách tính S_i^+ và S_i^- như sau. Trước hết, lấy giá trị đầu tiên của cột hạng của B (tức là k_1) và đếm số hạng có giá trị lớn hơn k_1 đó chính là S_i^+, số hạng còn lại có giá trị nhỏ hơn k_1 là S_i^-. Tiếp tục như vậy cho đến giá trị k_{n-1} thì kết thúc. Tổng quát ta có thể mô tả cách tính S_i^+ và S_i^- như sau:

S_i^+ = số hạng $k_j (j > 1)$ thỏa mãn: $k_j > k_i$

S_i^- = số hạng $k_j (j > 1)$ thỏa mãn: $k_j < k_i$

Cuối cùng là $S = \sum_i (S_i^+ - S_i^-)$. (S còn được gọi là số điểm Kendall-Kendall's score)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ký hiệu ρ_k là hệ số tương quan hạng Kendall của tổng thể thì để kiểm định cặp giả thuyết:

$$H_0 : (\rho_k = 0); H_1 : (\rho_k \neq 0)$$

Ta sử dụng miền bác bỏ sau đây:

$$W_\alpha = \left\{ U = \frac{\tau}{\sqrt{\frac{4n+10}{9n(n-1)}}}; |U| > u_{\alpha/2} \right\}$$

Với số liệu đã tính ở trên ta có:

$$\tau = \frac{S}{n(n-1)/2} = \frac{21}{10(10-1)/2} = 0{,}467$$

và $U_{qs} = \frac{0{,}467}{\sqrt{\frac{40+10}{90.9}}} = 1{,}88$

Với $\alpha = 0{,}05$ ta có $u_{0{,}025} = 1{,}96 \rightarrow U_{qs} \notin W_\alpha$. Vậy không có cơ sở bác bỏ giả thuyết $H_0 : (\rho_k = 0)$. Có nghĩa là có thể coi điểm của hai giám khảo người Canada và Nhật là độc lập với nhau với mức ý nghĩa 5%.

Sau đây là kết quả chạy bằng Stata đối với ví dụ 5.

- spearman diem1 diem2
  Number of obs = 10
  Spearman's rho = 0.6364
  Test of Ho: diem1 and diem2 independent
  Pr > |t| = 0.0479

- ktau diem1 diem2
  Number of obs= 10
  Kendall's tau-a= 0.4667
  Kendall's tau-b= 0.4667
  Kendall's score= 21

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

SE of score = 11.180

Test of Ho: diem1 and diem2 independent

Pr > |z| = 0.0736 (continuity corrected)

Bài tập

10.1. Người ta xếp 12 dải màu sắc theo thứ tự (hạng) tăng dần tương ứng với mức độ tăng dần của độ đậm và lấy đó làm chuẩn. Sau đó kiểm tra một sinh viên bằng cách để người đó tự xếp lại 12 dải màu sắc đó và được bảng sau:

Hạng đúng	1	2	3	4	5	6	7	8	9	10	11	12
Hạng của sinh viên	6	3	4	2	1	10	7	8	9	5	11	12

Tìm hệ số tương quan hạng Spearman giữa cách xếp hạng của các chuyên gia (hạng đúng) và cách xếp hạng của sinh viên được kiểm tra.

ĐS: $R_s = 0.7527$

10.2. Sau đây là điểm chấm của hai giám khảo trong hội đồng thi nâng tay nghề 10 sản phẩm của một thí sinh:

G. khảo 1	18,7	3,3	7,4	8,0	3,5	5,8	5,5	0,2	6,8	5,0
G. khảo 2	6,9	4,6	2,8	7,6	5,6	6,1	6,1	0,1	5,9	9,5

Tìm hệ số tương quan hạng Spearman và Kendall cho điểm chấm của hai giám khảo nói trên.

10.3. Chuyên gia của hai nhà máy được trưng cầu ý kiến về 9 giải pháp 1 dự án kinh tế. Các chuyên gia này đã xếp hạng cho 9 giải pháp đó như sau:

C. gia nhà máy A	1	2	3	3	3	4	4	4	4
C.gia nhà máy B	2	1	4	3	5	6	6	7	8

Ý kiến đánh giá có sự khác biệt nhau không? Cho kết luận với mức ý nghĩa 1%.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

10.4. Cho X là thu nhập (100$/năm) và Y là số tiền chi cho hàng xa xỉ phẩm ($/tháng), trên cơ sở mẫu sau đây hãy đánh giá mức độ liên quan giữa X và Y bằng các hệ số tương quan hạng. Giải thích vì sao có thể làm như vậy?

X	23	17	34	56	49	31	26	80	65	40	26
Y	10	50	120	225	90	60	55	340	170	25	80

ĐS: $R_s = 0{,}9746$

10.5. Hai chuyên gia A, B cùng xem xét và đánh giá tay nghề của 11 công nhân. Sau đây bảng sắp xếp thứ tự tay nghề của 2 chuyên gia nói trên (những người để trong ngoặc được coi là cùng trình độ).

	1	2	(3	4	5)	6	(7	8	9)	10	11
	2	1	3	5	4	7	(6	9)	8	10	11

Kết quả đánh giá của hai chuyên gia có phù hợp với nhau không?

3. PHÂN TÍCH HỒI QUY

Trên đây chúng ta đã trình bày cách phân tích mối tương quan giữa p biến ngẫu nhiên: $X_1, X_2, ..., X_p$. Tuy nhiên trên thực tế, nhiều khi chúng ta muốn đi sâu nghiên cứu rõ hơn sự tác động của một, hay nhiều biến trong số đó đến sự biến động của một biến nào đó, ví dụ như biến ngẫu nhiên X_1 chẳng hạn. Có nghĩa là chúng ta muốn giải thích được nhiều hơn và cụ thể hơn sự biến động của X_1 thông qua sự tác động của những biến đó. Đó chính là mục đích của phương pháp phân tích hồi quy. Để tiện cho việc trình bày, chúng ta vẫn xét một tập hợp p biến ngẫu nhiên, nhưng trong đó X_1 được thay thế bằng Y. Tức là ta có: $Y, X_2, X_3, ..., X_p$.

3.1. Hàm hồi quy

Giả sử ta muốn tìm hàm f nào đó của $X_2, X_3, ..., X_p$ sao cho nó xấp xỉ Y tốt nhất theo nghĩa cực tiểu sai số bình phương trung bình. Có nghĩa là:

$$E(Y - f(X_2, X_3, ..., X_p))^2 = E(Y - f(X))^2$$
$$= \min_{f(x)} E(Y - f(X))^2$$

Trong đó ký hiệu $X = (X_2, X_3, ..., X_p)$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Người ta đã chứng minh được rằng $E(Y-f(X))^2$ sẽ đạt cực tiểu khi hàm f(X) là kỳ vọng có điều kiện của Y đối với $X_2, X_3, ..., X_p$. Tức là khi:

$$f(X) = E(Y / X_2, X_3, ..., X_p)$$

Định nghĩa: Hàm hồi quy (hay hàm hồi quy kỳ vọng của Y đối với một vectơ $X = (X_2, X_3, ..., X_p)$ là kỳ vọng có điều kiện của Y đối với X

$$f_Y(X) = E(Y / X_2, X_3, ..., X_p) = E(Y / X)$$

Hàm hồi quy có thể có nhiều dạng khác nhau. Ví dụ với p = 2 và ta có hai biến ngẫu nhiên là Y, X thì hàm hồi quy có dạng:

$$f_Y(X) = aX + b$$

được gọi là hàm *hồi quy tuyến tính đơn*. Trường hợp tổng quát (p>2) ta có:

$$f_Y(X_2, X_3, ..., X_p) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... \beta_p X_p$$

và nó được gọi là *hàm hồi quy tuyến tính bội*. Ngoài ra, hàm hồi quy còn có các dạng khác như hàm đa thức bậc 2, bậc 3,... hay dạng hàm logarit...Các hằng số $\beta_1, \beta_2, ..., \beta_p$ được gọi là *tham số* của hàm hồi quy.

Nội dung của bài toán phân tích hồi quy gồm một số điểm cơ bản như sau:

Dựa trên số liệu điều tra thống kê (mẫu ngẫu nhiên) thu được khi quan sát $(Y, X_2, X_3, ..., X_p)$ hãy:

1. Ước lượng các tham số của hàm hồi quy
2. Kiểm định giả thuyết về giá trị của các tham số đó.
3. Đánh giá các sai số ước lượng và kiểm tra tính phù hợp (hay đúng đắn) của hàm hồi quy
4. Dự báo (hay dự đoán) các trị của Y theo $X_2, X_3, ..., X_p$.

3.2. Mô hình hồi quy tuyến tính đơn

Trong phần này chúng ta xét trường hợp p = 2, tức là chúng ta xét hàm hồi quy có dạng sau đây:

$$f_Y(X) = aX + b$$

$f_Y(X)$ còn được gọi là hàm hồi quy lý thuyết. Giả sử ta có một mẫu ngẫu nhiên kích thước n thu được khi quan sát (Y,X) là:

$$\{(Y_i, X_i)\} = \{(Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n)\}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Khi đó chúng ta có thể viết: $f_{Yi}(X_i) = E(Y_i / X_i) = aX_i + b$

Hay:

$$Y_i = aX_i + b + U_i \quad i = \overline{1,n} \qquad (11.7)$$

Trong đó U_i là các sai số ngẫu nhiên và giả thiết rằng chúng độc lập với nhau, cùng tuân theo quy luật phân phối chuẩn $N(0,\sigma^2)$.

(11.7) được gọi là mô hình hồi quy tuyến tính đơn.

1. Ước lượng các hệ số a, b

Ký hiệu $\hat{a},\hat{b}$ là các ước lượng của a và b thì $\hat{Y} = \hat{a}X + \hat{b}$ được gọi là hàm hồi quy ước lượng (của hàm hồi quy lý thuyết $f_Y(X) = aX + b$)

Xét mô hình (11.7) và ta đi tìm $\hat{a},\hat{b}$ bằng phương pháp bình phương bé nhất tức là tìm $\hat{a},\hat{b}$ sao cho cực tiểu hóa được tổng các bình phương sai số (e_i^2):

$$S(\hat{a},\hat{b}) = \sum_i (Y_i - \hat{Y})^2 = \sum_i (Y_i - \hat{a}X_i - \hat{b})^2 = \sum_i e_i^2$$
$$= \min_{a,b} \sum_i (Y_i - aX_i - b)^2$$

Dễ dàng nhận thấy rằng $\hat{a},\hat{b}$ là nghiệm của hệ phương trình sau:

$$\begin{cases} \dfrac{\partial S(a,b)}{\partial a} = \dfrac{\partial \sum (Y_i - aX_i - b)^2}{\partial a} = 0 \\ \dfrac{\partial S(a,b)}{\partial b} = \dfrac{\partial \sum (Y_i - aX_i - b)^2}{\partial b} = 0 \end{cases}$$

Hay

$$\begin{cases} nb + a\sum X_i = \sum Y_i \\ b\sum X_i + a\sum X_i^2 = \sum X_i Y_i \end{cases} \qquad (11.8)$$

Giải hệ phương trình (11.8) ta được:

$$\hat{a} = \frac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{X^2} - (\overline{X})^2} = r_{xy}\frac{\sqrt{MS_Y}}{\sqrt{MS_X}} \qquad (11.9)$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trong đó:

$$MS_X = \frac{\sum (X_i - \overline{X})^2}{n} = \overline{X^2} - (\overline{X})^2$$

$$MS_Y = \frac{\sum (Y_i - \overline{Y})^2}{n} = \overline{Y^2} - (\overline{Y})^2$$

$$\overline{XY} = \sum_i (X_i, Y_i)/n$$

$$\overline{X}^2 = (\sum_i X_i^2)/n; \overline{Y}^2 = (\sum_i Y_i^2)/n$$

Người ta cũng chứng minh được rằng $\hat{a}, \hat{b}$ là các ước lượng không chệch tốt nhất của a, b và:

$$V(\hat{a}) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma^2/n}{\overline{X^2} - (\overline{X})^2} = \frac{\sigma^2/n}{MS_X}$$

$$V(\hat{b}) = \frac{\sigma^2 \overline{X}^2}{\sum (X_i - \overline{X})^2} = \frac{(\sigma^2/n)\overline{X}^2}{\overline{X^2} - (\overline{X})^2} = \frac{(\sigma^2/n)\overline{X}^2}{MS_X}$$

Nếu σ^2 chưa biết ta dùng ước lượng không chệch của nó là:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

Với các giả thuyết đã nêu ra đối với mô hình (11.7) ta cũng có:

+ $\hat{a} \sim N(a, V(\hat{a}))$ hay:

$$U = \frac{\hat{a} - a}{\sqrt{V(\hat{a})}} \sim N(0,1)$$

+ $\hat{b} \sim N(b, V(\hat{b}))$ hay:

$$U = \frac{\hat{b} - b}{\sqrt{V(\hat{b})}} \sim N(0,1)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trường hợp σ^2 chưa biết, thay thế nó bằng $\hat{\sigma}^2$, ta suy ra rằng:

$$T = \frac{\hat{a} - a}{Se(\hat{a})} \sim T(n-2)$$

(Phân phối Student với (n-2) bậc tự do)

$$T = \frac{\hat{b} - b}{Se(\hat{b})} \sim T(n-2)$$

Trong đó

$$Se(\hat{a}) = \sqrt{\frac{\hat{\sigma}^2}{\sum (X_i - \overline{X})^2}} = \sqrt{\frac{\hat{\sigma}^2 / n}{\overline{X^2} - (\overline{X})^2}} \qquad (11.10)$$

$$Se(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2 \overline{X}^2}{\sum (X_i - \overline{X})^2}} = \sqrt{\frac{(\hat{\sigma}^2 / n)\overline{X}^2}{\overline{X^2} - (\overline{X})^2}}$$

Từ đây ta có khoảng tin cậy $(1-\alpha)$ cho các tham số a, b là:

$$\left(\hat{a} - Se(\hat{a}) t_{\alpha/2}^{(n-2)} < a < \hat{a} + Se(\hat{a}) t_{\alpha/2}^{(n-2)}\right) \qquad (11.11)$$

Và $$\left(\hat{b} - Se(\hat{b}) t_{\alpha/2}^{(n-2)} < b < \hat{b} + Se(\hat{b}) t_{\alpha/2}^{(n-2)}\right)$$

Ví dụ 6. Trở lại với tệp số liệu XSTK11_1 (Xem bảng 11.1).

Ta gọi chỉ tiêu gdp là Y và gap là X, dựa vào kết quả đã tính toán được ở ví dụ 1 ta có:

$$\hat{a} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - (\overline{X})^2} = \frac{437,23 - 14,391.28,424}{216,87 - 14,391^2} = 2,885$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X} = 28,427 - 2,885.14,391 = -13,095$$

Vậy: $\hat{Y} = \hat{a}X + \hat{b} = 2,885X - 13,095$

Để tính được $\hat{\sigma}^2$ mà không cần phải tính các giá trị $e_i = Y_i - \hat{Y}_i$ ta xét biểu thức sau đây:

$$TSS = \sum_i (Y_i - \overline{Y}_i)^2 = \sum_i \left[(Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \overline{Y})^2\right] = \qquad (11.12)$$

$$= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \overline{Y})^2 = ESS + RSS$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trong đó: $ESS = \sum_i (\hat{Y}_i - \overline{Y})^2$ (Explained sum of squares)

$$RSS = \sum_i (Y_i - \hat{Y}_i)^2 \text{ (Residual sum of squares)}$$

Chú ý rằng

$$ESS = \sum_i (\hat{Y}_i - \overline{Y})^2 = \sum_i (\hat{a}X_i + b - \overline{Y})^2 =$$

$$= \sum_i (\hat{a}X_i + \overline{Y} - \hat{a}\overline{X} - \overline{Y})^2 = (\hat{a})^2 \sum_i (X_i - \overline{X})^2 =$$

$$= (\hat{a})^2 \left[\sum_i X_i^2 - n(\overline{X})^2 \right]$$

Tiếp tục ví dụ 6, theo bảng số liệu đã tính toán ở ví dụ 1 ta có:

$$TSS = \sum_i (Y_i - \overline{Y}_i)^2 = \sum_i Y_i^2 - n(\overline{Y})^2$$

$$= 15133,04 - 17.(28,4235)^2 = 1398,8$$

$$ESS = (\hat{a})^2 \left[\sum_i X_i^2 - n(\overline{X})^2 \right]$$

$$= 2,885^2(3686,8125 - 17.144,391^2) = 1380,4$$

Vì RSS = TSS - ESS nên ta có:

RSS = TSS - ESS = 1398,8 - 1380,4 = 16,4

Vậy: $\hat{\sigma}^2 = \dfrac{RSS}{n-2} = \dfrac{16,4}{15} = 1,09$

Theo công thức (11.10) ta cũng có:

$$Se(\hat{a}) = \sqrt{\frac{\frac{\hat{\sigma}^2}{n}}{\overline{X}^2 - (\overline{X})^2}} = \sqrt{\frac{\frac{1,09}{17}}{216,8713 - 14,391}} = 0,081$$

$$Se(\hat{b}) = \sqrt{\frac{\left(\frac{\hat{\sigma}^2}{n}\right)\overline{X}^2}{\overline{X}^2 - (\overline{X})^2}} = \sqrt{\frac{\left(\frac{1,09}{17}\right).216,8713}{9,77}} = 1,91$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đây ta có thể tính được các khoảng tin cậy cho a và b theo công thức (11.11) như sau:

$(2,885 - 0,081.2,131 < a < 2,885 + 0,081.2,131)$

$\rightarrow \quad (2,713 < a < 3,058)$

$(-13,905 - 1,191.2,131 < b < -13,095 + 1,191.2,131)$

$\rightarrow \quad (-15,633 < b < -10,556)$

2. Kiểm định giả thuyết đối với a, b

+ Với cặp giả thuyết

$$H_0 : (a = a_0); H_1 : (a \neq a_0)$$

Ta thấy nếu H_0 đúng thì:

$$T = \frac{\hat{a} - a_0}{Se(\hat{a})} \sim T(n-2)$$

Do đó miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ t = \frac{\hat{a} - a_0}{Se(\hat{a})}; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

+ Với cặp giả thuyết: $H_0 : (b = b_0); H_1 : (b \neq b_0)$

Tương tự ta có miền bác bỏ là:

$$W_\alpha = \left\{ t = \frac{\hat{b} - b_0}{Se(b)}; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

Trở lại ví dụ 6, giả sử ta muốn kiểm định cặp giả thuyết

$$H_0 : (a = a_0 = 3); H_1 : (a \neq 3)$$

Trước hết ta tính:

$$t_{qs} = \frac{\hat{a} - a_0}{Se(\hat{a})} = \frac{2,885 - 3}{0,081} = -1,419$$

Ta thấy $|t_{qs}| < t_{\alpha/2}^{(n-2)} = t_{0,025}^{(15)} = 2,131$ do đó $t_{qs} \notin W_\alpha$. Chưa có cơ sở bác bỏ giả thuyết H_0, do đó có thể xem như hệ số a = 3;

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3. Kiểm định sự phù hợp của hàm hồi quy

Phần này chúng ta sẽ làm tương tự như đối với bài toán phân tích phương sai (xem chương X).

Theo công thức (11.12) ta có:

$$TSS = ESS + RSS$$

Như vậy nếu chúng ta lấy TSS để đo mức độ biến động của các giá trị của biến phụ thuộc Y xung quanh giá trị trung bình $(\overline{Y})$ của nó thì ta thấy TSS được chia thành 2 phần là ESS và RSS. Trong đó:

$$ESS = \sum_i (\hat{Y}_i - \overline{Y})^2$$

có thể coi như là ước lượng của TSS và RSS là sai số do ước lượng mô hình hồi quy mà có. Có thể thấy ngay mô hình hồi quy sẽ càng tốt nếu kết quả ước lượng cho ta RSS càng nhỏ, tức là ESS càng gần với TSS.

Chia hai vế cho TSS ta được

$$1 = 100\% = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

Giá trị $R^2 = \frac{ESS}{TSS}$ được gọi là *hệ số xác định* (hay xác định bội trong trường hợp p>2) của mô hình hồi quy.

Ý nghĩa của R^2: R^2 cho ta biết trong 100% của toàn bộ sự sai lệch (biến động) của Y so với giá trị trung bình của nó thì có bao nhiêu phần trăm là do biến X (còn gọi là biến giải thích) gây nên. Nói một cách khác mô hình đã giải thích được (R^2.100) phần trăm của sự biến động của Y xung quanh giá trị trung bình của nó, số phần trăm còn lại do sai số ngẫu nhiên và do các yếu tố khác (nếu có) mà ta chưa đưa vào mô hình để xem xét.

Cũng với ví dụ 6 ta có:

$$R^2 = \frac{ESS}{TSS} = \frac{1382,4}{1398,8} = 0,988$$

Như vậy nếu R^2 càng lớn thì mô hình càng tốt. Mô hình càng giải thích được nhiều về sự biến động của Y (còn gọi là biến được giải thích). Việc kiểm định sự phù hợp hay đúng đắn của hàm hồi quy chính là kiểm định cặp giả thuyết sau đây:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$H_0 : (R^2 = 0); H_1 : (R^2 \neq 0) \qquad (11.13)$$

Cặp giả thuyết trên tương đương với cặp giả thuyết :

$$H_0 : (a = 0); H_1 : (a \neq 0)$$

Để kiểm định cặp giả thuyết (11.13) người ta sử dụng tiêu chuẩn sau đây:

$$F = \frac{\left(\frac{R^2}{1}\right)}{\frac{(1-R^2)}{(n-2)}}$$

Nếu H_0 đúng, F có phân phối F(1, n - 2)

Vậy miền bác bỏ để kiểm định giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{\frac{R^2}{(1-R^2)}}{(n-2)}; F > f_\alpha(1, n-2) \right\}$$

Nhận xét: Đối với mô hình hồi quy tuyến tính đơn (p = 2) ta luôn có $R^2 = r_{xy}^2$

Trở lại ví dụ 6, ta đã tính được r_{XY}=0,9942 do đó

$$r_{XY}^2 = 0{,}9942^2 = 0{,}9884 = R^2.$$

4. Dự báo

Chúng ta có $\hat{Y} = \hat{a}X + \hat{b}$ là hàm hồi quy ước lượng của hàm hồi quy lý thuyết

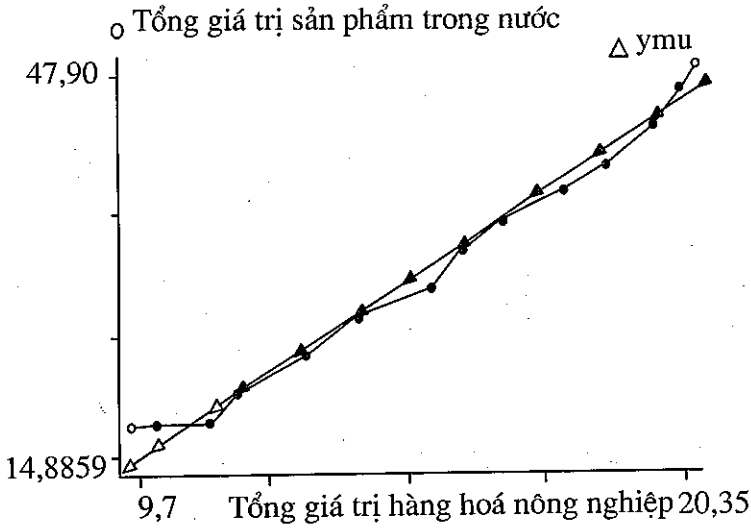
$$f_Y(X) = aX + b$$

Tất nhiên ta cũng thấy Y cũng có các quan hệ chặt chẽ với $\hat{Y}$. Với cùng các giá trị $\{X_i\}$ ta tính các giá trị hồi quy ước lượng:

$$\hat{Y}_i = \hat{a}X_i + \hat{b}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sau đó ta vẽ đồ thị Y và các điểm $\hat{Y}_i$ trên cùng một hệ tọa độ, lúc đó ta sẽ thấy rõ hơn mối quan hệ giữa các giá trị thực tế $\{Y_i\}$ và các giá trị hồi quy ước lượng $\{\hat{Y}_i\}$. Với ví dụ 6 ta có hình ảnh sau đây:



Có thể xem $\hat{Y}$ như là đường xu thế của Y. $\hat{Y}$ là một công cụ đắc lực trong công tác dự báo ngắn hạn. Ta có hai loại dự báo sau đây (trên cơ sở sử dụng $\hat{Y}$):

- Dự báo trung bình có điều kiện của Y khi X nhận giá trị X_0

- Dự báo giá trị cá biệt của Y khi $X = X_0$

+ Dự báo giá trị trung bình:

Giả sử $X=X_0$ ta có $\hat{Y} = \hat{a}X + \hat{b}$ là ước lượng điểm của các $f_Y(X_0) = E(Y/X = x_0) = aX_0 + b$

$\hat{Y}_0$ là ước lượng hiệu quả của $f_Y(X_0)$. Phương sai của $\hat{Y}_0$ là:

$$V(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2} \right]$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thay σ^2 bằng $\hat{\sigma}^2$ ta có:

$$T = \frac{\hat{Y}_0 - (aX_0 + b)}{Se(\hat{Y}_0)} \sim T(n-2)$$

Trong đó $Se(\hat{Y}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]}$

Do đó khoảng tin cậy (1-α) của $f_Y(X_0)$ là:

$$\left(\hat{Y}_0 - t_{\alpha/2}^{(n-2)} se(\hat{Y}_0) < f_Y(X_0) < \hat{Y}_0 + t_{\alpha/2}^{(n-2)} se(\hat{Y}_0)\right)$$

Ví dụ: Dự báo giá trị trung bình của tổng sản phẩm trong nước khi tổng giá trị hàng hóa nông nghiệp đạt mức 25 triệu tấn (X_0 =25)

$$Se(\hat{Y}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]}$$

$$= \sqrt{1{,}09\left[\frac{1}{17} + \frac{(25 - 14{,}391)^2}{166{,}097}\right]} = 0{,}896$$

$(1-\alpha) = 0{,}95, t_{\alpha/2}^{(n-2)} = t_{0{,}025}^{(15)} = 2{,}131$

$\hat{Y}_0 = \hat{a}X_0 + \hat{b} = 2{,}885.25 - 13{,}095 = 59{,}03$. Ta cũng có khoảng tin cậy cho $f_Y(25)$ là:

$$59{,}03 - 2{,}131.0{,}896 < f_Y(25) < 59{,}03 + 2{,}131.0{,}896$$

$$57{,}12 < f_Y(25) < 60{,}94$$

+ *Dự báo giá trị cá biệt*

Ký hiệu giá trị của Y ứng với $X=X_0$ là Y_0, ta cũng lấy $\hat{Y}_0 = \hat{a}X_0 + \hat{b}$ làm ước lượng điểm cho giá trị Y_0.

Vì $V(Y_0 - \hat{Y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nên ta có: $T = \dfrac{Y_0 - \hat{Y}_0}{se(Y_0 - \hat{Y}_0)} \sim T(n-2)$

Trong đó $Se(Y_0 - \hat{Y}_0) = \sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]}$

Do đó khoảng tin cậy của Y_0 là:

$$\left(\hat{Y}_0 - t_{\alpha/2}^{(n-2)} Se(Y_0 - \hat{Y}_0) < Y_0 < \hat{Y}_0 + t_{\alpha/2}^{(n-2)} Se(Y_0 - \hat{Y}_0)\right)$$

Ví dụ: Ta đi tìm khoảng tin cậy cho giá trị tổng sản phẩm trong nước khi tổng giá trị hàng hóa nông nghiệp đạt mức 25 triệu tấn ($X = X_0 = 25$)

$$Se(Y_0 - \hat{Y}_0) = \sqrt{1,09\left[1 + \frac{1}{17} + \frac{(25 - 14,391)^2}{166,097}\right]} = 1,376$$

$$59,03 - 2,131.1,376 < Y_0 < 59,03 + 2,131.1,376$$

$$\rightarrow \qquad 56,098 < Y_0 < 61,962$$

Sau đây là kết quả phân tích hồi quy đối với bài toán đã nêu ở ví dụ 6 được chạy bởi Stata. Các bạn hãy thử đọc kết quả được in ra và so sánh với những kết quả mà chúng ta đã tính bằng tay ở trên.

- Regress gdp gap

Source	SS	df	MS	Number of obs =	17
Model	1382.47717	1	1382.47717	F(1,15)=	1271.17
Residual	16.313456	15	1.08756373	Prob>F =	0.0000
				R-squared =	0.9883
Total	1398.79063	16	87.4244141	Adj R-squared =	0.9876
				Root MSE =	1.0429

gdp	Coef.	Std.Er.	t	P>|t|	[95% Conf. Interval]	
gap	2.885761	.0809392	35.653	0.000	2.713243	3.058278
cons	-13.105	1.19195	-10.995	0.000	-15.64655	-10.56537

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.3. Mô hình hồi quy tuyến tính bội

Ta xét trường hợp tổng quát: p > 2, tức là ta có p biến ngẫu nhiên : $Y, X_2, X_3, ..., X_p$.

Giả sử hàm hồi quy (lý thuyết) của Y đối với các biến $X_2, X_3, ..., X_p$ có dạng sau đây:

$$f_Y(X_2, X_3, ..., X_p) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p$$

$$= \beta_1 + \sum_{j=2}^{p} \beta_j X_j$$

Với ma trận quan sát là:

$$\begin{pmatrix} Y_1 & X_{12} & X_{13} & ... & X_{1p} \\ Y_2 & X_{22} & X_{23} & ... & X_{2p} \\ & & ..... & & \\ Y_n & X_{n2} & X_{n3} & ... & X_{np} \end{pmatrix}$$

Ta có mô hình hồi quy như sau:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_p X_{ip} + U_i \quad i = \overline{1,n} \quad (11.14)$$

Trong đó ta cũng giả thiết các sai số ngẫu nhiên U_i độc lập với nhau và cùng tuân theo quy luật $N(0, \sigma^2)$

Ký hiệu:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{pmatrix}; X = \begin{pmatrix} 1 & X_{12} & ... & X_{1p} \\ 1 & X_{22} & ... & X_{2p} \\ & & ..... & \\ 1 & X_{n2} & ... & X_{np} \end{pmatrix}; X_j = \begin{pmatrix} X_{1j} \\ X_{2j} \\ ... \\ X_{nj} \end{pmatrix} \quad j = \overline{2,p}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_p \end{pmatrix} \text{ và } U = \begin{pmatrix} U_1 \\ U_2 \\ ... \\ U_n \end{pmatrix}; I = \begin{pmatrix} 1 \\ 1 \\ ... \\ 1 \end{pmatrix}$$

Ta có thể viết mô hình (11.14) dưới dạng ma trận như sau:

$$Y = X\beta + U$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1. Ước lượng vectơ tham số β

Ta đi tìm ước lượng $\hat{\beta}$ theo phương pháp bình phương bé nhất, tức là tìm $\hat{\beta}$ sao cho:

$$\sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_k X_{pi})^2 =$$
$$= \sum_i e_i^2 = \min S(\beta_1, \beta_2, ..., \beta_p)$$

Hoặc viết dưới dạng vectơ:

$$\sum_i (Y_i - \hat{Y}_i)^2 = \left\| Y - \hat{Y} \right\|^2 = \left\| e \right\|^2 = \min$$

Dễ dàng thấy rằng $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ là nghiệm của hệ phương trình tuyến tính sau đây:

$$\begin{cases} \dfrac{\partial S(\beta_1, \beta_2, ..., \beta_p)}{\partial \beta_j} = 0 \\ j = \overline{1, p} \end{cases}$$

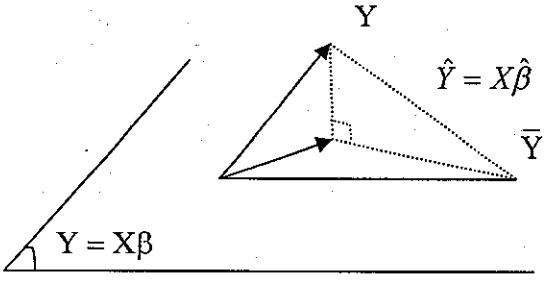
Tuy nhiên việc giải hệ phương trình tuyến tính trên một cách tổng quát sẽ rất cồng kềnh và phức tạp. Có thể sử dụng hình học giải tích và một số phép tính cơ bản của ma trận để tìm ra biểu thức tính véctơ $\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$ một cách nhanh chóng và đơn giản hơn xét về mặt mô tả ký hiệu.

Nếu coi β như một vectơ biến thì $F = X\beta$ chính là phương trình của một siêu phẳng trong không gian R^n với các cơ sở là các vectơ $I, X_2, X_3, ..., X_p$. Tức là ta cũng phải giả thiết rằng $I, X_2, ..., X_p$ là độc lập tuyến tính hay ma trận X cơ sở hạng là p.

Vì $\left\| Y - X\hat{\beta} \right\|^2 = \left\| Y - \hat{Y} \right\|^2 = \min$ nên ta suy ra rằng $\hat{Y}$ là hình chiếu của Y lên siêu phẳng F. Tức là ta có: $(Y - \hat{Y})$ trực giao với F. Vì $I, X_2, ..., X_p$ là cơ sở của F do đó $(Y - \hat{Y})$ cũng trực giao với tất cả những vectơ đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ta có thể mô tả bằng hình vẽ sau đây:



Hình 11.1

Vì $(Y-\hat{Y})$ vuông góc với $I, X_2, ..., X_p$ nên tích vô hướng của $(Y-\hat{Y})$ với những vectơ đó phải bằng không. Tức là ta có:

$$(X_1, (Y-\hat{Y})) = \sum X_{1i}(Y_i - \hat{Y}_i) = 0$$

$$\text{Và } (I, (Y-\hat{Y})) = \sum (Y_i - \hat{Y}_i) = 0$$

Sử dụng ký hiệu ma trận ta có thể viết lại như sau:

$$(X_1, (Y-\hat{Y})) = X_1'(Y_1 - \hat{Y}_1) = 0$$

$$(I, (Y-\hat{Y})) = I'(Y-\hat{Y}) = 0$$

Hay : $(X', (Y-\hat{Y})) = 0$. Thay $\hat{Y} = X\hat{\beta}$ vào ta được

$$(X'(Y-\hat{Y})) = X'(Y - X\hat{\beta}) = 0 \rightarrow X'Y = X'X\hat{\beta}$$

Vì hạng của X bằng p nên $(X'X)$ cũng có hạng bằng p, do đó sẽ tồn tại ma trận nghịch đảo $(X'X)^{-1}$. Ta có:

$$(X'X)^{-1}X'Y = (X'X)^{-1}X'X\hat{\beta}$$

Vậy: $\hat{\beta} = (X'X)^{-1}X'Y$ (11.15)

$\hat{\beta}$ có ma trận hiệp phương sai là:

$$\text{cov}(\hat{\beta}) = \{\text{cov}(\hat{\beta}_i, \hat{\beta}_j)\} = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

Thay $\hat{\beta} = (X'X)^{-1}X'(X\beta + u)$ vào biểu thức trên ta được:

$$\text{cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \{\sigma\}..k, i = \overline{1, p}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì σ^2 chưa biết, chúng ta thay nó bằng ước lượng không chệch:

$$\hat{\sigma}^2 = \frac{\left\|Y - \hat{Y}\right\|^2}{n-p} = \frac{\|e\|^2}{n-p} = \frac{\sum e_i^2}{n-p} \qquad (11.16)$$

Các phần tử nằm trên đường chéo của ma trận hiệp phương sai $\{\sigma_{kj}\} = \sigma^2(X'X)^{-1}$ chính là các phương sai của các ước lượng $\hat{\beta}$. Tức là:

$$V(\hat{\beta}_j) = \text{cov}(\hat{\beta}_j, \hat{\beta}_j) = \sigma_{jj}^2$$

Do đó ta có: $se(\hat{\beta}_j) = \sqrt{V(\hat{\beta}_j)} = \sigma_{jj}$

2. Khoảng tin cậy và kiểm định giả thuyết về các tham số hồi quy

a. Khoảng tin cậy cho tham số hồi quy

Thông thường thì σ^2 chưa biết, do đó chúng ta thay bằng $\hat{\sigma}^2$. Khi đó các thống kê:

$$T = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \ldots\ldots j = \overline{1,p}$$

sẽ có phân phối Student với (n - p) bậc tự do: T(n - p). Vậy khoảng tin cậy (1 - α) của các β_j sẽ là:

$$\hat{\beta}_j - Se(\hat{\beta}_j)t_{\alpha 2}(n-k) < \beta_j < \hat{\beta}_j + Se(\hat{\beta}_j)t_{\alpha 1}(n-k) \qquad (j = \overline{1,k})$$

b. Kiểm định giả thuyết về tham số hồi quy

+ Nếu muốn biết X_j có độc lập với Y không chúng ta kiểm định cặp giả thuyết sau:

$$H_0 : (\beta_j = 0); H_1 : (\beta_j \neq 0)$$

với miền bác bỏ được xác định như sau:

$$W_\alpha = \left\{ T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}; |T| > t_{\alpha/2}^{(n-p)} \right\}$$

+ Muốn kiểm định về mức độ ảnh hưởng của X_j đối với sự biến động của Y chúng ta xét cặp giả thuyết:

$$H_0 : (\beta_j = \beta^*); H_1 : (\beta_j \neq \beta^*)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Miền bác bỏ trong trường hợp này là:

$$W_\alpha = \left\{ T = \frac{\hat{\beta}_j - \beta^*}{se(\hat{\beta}_j)}; |T| > t_{\alpha/2}^{(n-p)} \right\}$$

+ Để so sánh mức độ ảnh hưởng của hai biến X_j và X_k đối với sự biến động của Y chúng ta có thể tiến hành kiểm định cặp giả thuyết sau:

$$H_0 : (\beta_j \pm \beta_k = \beta^*); H_1 : (\beta_j \pm \beta_k \neq \beta^*)$$

với miền bác bỏ là:

$$W_\alpha = \left\{ T = \frac{\hat{\beta}_j \pm \hat{\beta}_k - \beta^*}{se(\hat{\beta}_j \pm \hat{\beta}_k)}; |T| > t_{\alpha/2}^{(n-p)} \right\}$$

Trong đó $Se(\hat{\beta}_i \pm \hat{\beta}_j)$ xác định như sau:

$$Se(\hat{\beta}_i \pm \hat{\beta}_j) = \sqrt{Var(\hat{\beta}_i \pm \hat{\beta}_j)} = \sqrt{Var(\hat{\beta}_i) \pm 2Cov(\hat{\beta}_i, \hat{\beta}_j) + Var(\hat{\beta}_j)}$$

3. Hệ số xác định bội

Từ hình vẽ 11.1 dễ dàng thấy rằng:

$$\left\| Y - \overline{Y} \right\|^2 = \left\| Y - \hat{Y} \right\|^2 + \left\| \hat{Y} - \overline{Y} \right\|^2$$

Hay $\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \hat{Y})^2 + \sum (\hat{Y} - \overline{Y})^2$

Giống như trường hợp chúng ta xét mô hình hồi quy đơn, ở đây chúng ta cũng có: TSS = RSS + ESS

Chia 2 vế cho TSS ta thu được:

$$1 = 100\% = \frac{RSS}{TSS} + \frac{ESS}{TSS}$$

Đặt: $R^2 = \frac{EES}{TSS} = \frac{\left\| \hat{Y} - \overline{Y} \right\|^2}{\left\| Y - \overline{Y} \right\|^2} = 1 - \frac{RSS}{TSS}$

R^2 chính là tỉ lệ (số phần trăm) sự biến động của Y xung quanh giá trị trung bình $\overline{Y}$ được giải thích bởi mô hình hồi quy. Mặt khác, từ ý nghĩa hình học ta thấy R^2 cũng chính là cosin của góc được tạo bởi hai vectơ

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$(\hat{Y} - \overline{Y})$ và $(Y - \overline{Y})$. Góc đó càng nhỏ thì độ chính xác của ước lượng càng cao, tức là R^2 càng lớn thì mô hình càng tốt.

+ Xét cặp giả thuyết:

$$H_0 : (R^2 = 0); H_1 : (R^2 \neq 0)$$

Cặp giả thuyết này tương đương với cặp giả thuyết sau:

$H_0 : (\beta_2 = \beta_3 = ... = \beta_p = 0)$

$H_1 : (\text{Tồn tại ít nhất một } \beta_j \neq 0)$

Chúng ta sử dụng tiêu chuẩn: $$F = \frac{\frac{R^2}{(p-1)}}{\frac{(1-R^2)}{(n-p)}}$$

F sẽ tuân theo quy luật F(p — 1, n — p) nếu H_0 đúng.

Vậy miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = F = \left\{ \frac{\frac{R^2}{(p-1)}}{\frac{(1-R^2)}{(n-p)}}; F > f_\alpha(p-1, n-p) \right\}$$

Chú ý:

1. Về mặt thực hành, việc áp dụng mô hình hồi quy tuyến tính cùng với phương pháp bình phương bé nhất chỉ có ý nghĩa (đáng tin cậy) khi các giả thiết về mô hình được thoả mãn. Để kiểm tra xem các giả thuyết này có được thoả mãn không người ta có thể sử dụng các tiêu chuẩn khác nhau. Ví dụ, để kiểm định giả thuyết về sự phân phối chuẩn của các U_i, chúng ta có thể sử dụng các tiêu chuẩn χ^2, Kolmogorov, Sktest... như đã trình bày ở chương 9. Về vấn đề này có thể tham khảo các sách về kinh tế lượng.

2. Trên thực tế số liệu thống kê cần phân tích thường rất lớn do đó việc tính toán bằng tay là hầu như không thể làm được. Để có thể sử dụng thống kê như một công cụ phân tích có hiệu quả thì việc sử dụng các phần mềm thống kê là gần như bắt buộc. Các phần mềm thống kê hay được sử dụng và được phổ biến rộng rãi hiện nay là: SPSS, STATA,...

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Cuối cùng, để minh hoạ cho phần này chúng ta quay trở lại với tệp số liệu XSTK11_1 (xem §2), ta có kết quả phân tích hồi quy mô hình sau đây:

$$gdp_i = \beta_i ex_i + \beta_2 em_i + \beta_3 gip_i + \beta_4 gap_i + \beta_5 + U_i \quad (i = \overline{1,17})$$

bằng Stata như sau:

regress gdp gap gip ex im

Source	SS	df	MS	Number of obs = 17
Model	1397.57741	4	349.394352	F(4,12) = 3455.88
Residual	1.21321847	12	.101101539	Prob > F = 0.0000
Total	1398.79063	16	87.4244141	R-squared = 0.9991 Adj R-squared = 0.9988 Root MSE =.31796

gdp	Coef.	Std.Err	t	P > |t|	[95% Conf.	Interval
gap	1.232316	.2191744	5.623	0.000	.7547759	1.709856
gip	.6063006	.1301608	4.658	0.001	.3227045	.8898966
ex	1.181809	.2596261	4.552	0.001	.6161321	1.747485
im	-.43687	.175753	-2.486	0.029	-.3198064	-.0539405
cons	.84561	1.687897	0.501	0.625	-2.832	4.623224

Muốn kiểm định giả thuyết về các hệ số của hàm hồi quy, ví dụ muốn kiểm định giả thuyết hệ số của biến ex bằng 1 ta gõ lệnh test ex = 1, hoặc kiểm định giả thuyết về hệ số tự do (hệ số chặn) bằng 0 ta gõ lệnh test cons = 0. Sau đây là kết quả của các kiểm định trên:

```
. test ex = 1
(1) ex = 1.0
          F(1, 12) = 0.49
          Prob > F = 0.4971
. test_cons = 0
(1)_cons = 0.0
          F(1, 12) = 0,25
          Prob > F = 0.6254
```

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Để kết thúc phần này, ta xét thêm một ví dụ nữa để tham khảo. Nếu có điều kiện, độc giả hãy thử chạy chương trình Stata và so sánh với kết quả in ra dưới đây:

Contains data from C:\WINSTATA\XSTK11_4.dta

obs : 16

vars : 5 25 Aug 2010 10:08

size : 384 (85.9% of memory free)

1. nam	float %9.0g	
2. gdp	float %9.2g	Tong s.p trong nuoc (1000 ty)
3. dien	float %9.2	San luong dien (ty Kwh)
4. ximang	float %9.2	San luong xi mang (trieu tan)
5. dien thoai	float %9.2	So may dien thoai (1000 chiec)

Sorted by:

list	nam	gdp	dien	ximang	dthoai
1	1980	16.30	3.60	.60	90.60
2	1981	17.20	3.80	.60	97.40
3	1982	18.70	4.10	.70	98.50
4	1983	20.10	4.30	1.00	103.80
5	1984	21.80	5.00	1.30	108.70
6	1985	23.00	5.20	1.50	103.10
7	1986	23.80	5.70	1.50	113.50
8	1987	24.70	6.20	1.70	116.10
9	1988	25.90	7.00	2.00	109.10
10	1989	28.00	7.90	2.10	110.70
11	1990	29.50	8.80	2.50	114.40
12	1991	31.30	9.30	3.10	121.10
13	1992	34.00	9.80	3.90	132.10
14	1993	36.70	10.90	4.80	268.30
15	1994	40.00	12.50	5.40	420.00
16	1995	43.80	14.70	5.90	600.00

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

. correlate gdp dien ximang dthoai

(obs = 16)

	gdp	dien	ximang	dthoai
gdp	1.0000			
dien	0.9946	1.0000		
ximang	0.9888	0.9821	1.0000	
dthoai	0.8176	08338	0.8488	1.0000

. pcorr gdp dien ximang dthoai

(obs = 16)

Partial correlation of gdp with

Varible	Corr.	Sig.
dien	0.8719	0.000
ximang	0.7207	0.004
dthoai	-0.5075	0.064

. regress gdp dien ximang dthoai

Source	SS	df	MS	Number of obs = 16
Model	990.84684	3	330.28228	F(3,12) = 808.44
Residual	4.90251119	12	.408542599	Prob < F = 0.0000
Total	995.749351	15	66.38329	R-squared = 0.9951
				Adj R-squared = 0.9938
				Root MSE =.63917

gdp	Coef.	Std.Err	t	P > |t|	[95% Conf.	Interval
dien	1.623106	.2631666	6.168	0.000	1.049715	2.196497
ximang	1.908267	-5298553	3.601	0.004	.753812	3.062723
dthoai	-.004454	.0021831	-2.041	0.064	-.009211	.0003019
_cons	11.30476	.8021872	14.092	0.000	9.556943	13.05257

. test dthoai = 0

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

(1) dthoai = 0.0

F(1,12) = 4.16

Prob > F = 0.0639

. test dien = 2

(1) dien = 2.0

F(1,12) = 2.05

Prob > F = 0.1776

3.4. Một số dạng hàm hồi quy phi tuyến có thể đưa về dạng tuyến tính

1. Hồi quy luỹ thừa

$$f_y(X_2, X_3, ..., X_p) = \alpha X_2^{\beta_2} X_3^{\beta_3} ... X_p^{\beta_p}$$

Lấy logarit hai vế ta được:

$$\ln f_y(X_2, ..., X_p) = \ln \alpha + \beta_2 \ln X_2 + ... + \beta_p \ln X_p$$

Đặt: $\ln \alpha = \beta_1, \ln X_j = V_j$ và $\ln Y = Z$ ta có mô hình hồi quy tuyến tính sau đây:

$$Zi = \beta_1 + \beta_2 V_{i2} + \beta_3 V_{i3} + ... + \beta_p V_{ip} + U_i \qquad i = \overline{1,n}$$

Ví dụ: Xét hàm sản xuất Cobb — Douglas;

$$Q = aK^{\alpha} L^{\beta} e^{yt} X^{\delta} e^{u}$$

Ta có thể ước lượng các hệ số của hàm này bằng cách sử dụng mô hình sau đây:

$$\ln Q_i = \ln a + \alpha \ln K_i + \beta \ln L_i + yt_i + \delta \ln X_i + U_i \qquad i = \overline{1,n}$$

2. Hồi quy mũ

$$f_y(X_2, ..., X_p) = e^{\beta_1 + \beta_2 X_2 + ... + \beta_p X_p}$$

Lấy Logarit hai vế ta có hàm hồi quy tuyến tính:

$$\ln f_y(X_2, ..., X_p) = \beta_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Đặt $\ln Y = Z$ ta có mô hình sau đây:

$$Zi = \beta_1 + \beta_2 X_{i2} + ... + \beta_p X_{ip} + U_i$$

Khi đó:

$$\overline{Z} = \overline{\beta}_1 + \overline{\beta}_2 X_2 + ... + \overline{\beta}_p X_p$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Là ước lượng của $\ln f_y(X_2,...,X_p) = \ln E(Y/X_2,...,X_p)$

3. Hồi quy parabol

$$f_y(X) = aX^2 + bX + c$$

Đặt $X = V_1, X^2 = V_2$ ta có mô hình tuyến tính:

$$Y_i = aV_{i2} + bV_{i1} + C + U_i \qquad i = \overline{1,n}$$

4. Hàm quy hyperbol bội

$$f_y(X) = \beta_1 + \frac{\beta_2}{X_2} + \frac{\beta_3}{X_3} + ... + \frac{\beta_p}{X_p}$$

Đặt $\frac{1}{X_j} = V_j (j = \overline{2,p})$ ta có mô hình tuyến tính:

$$Y_i = \beta_1 + \beta_2 V_{i2} + \beta_3 V_{i3} ... + \beta_p V_{ip} + U_i \qquad i = \overline{1,n}$$

5. Hồi quy logarit

Trên thực tế chúng ta gặp nhiều trường hợp hàm hồi quy không phải là tuyến tính nhưng nếu thực hiện phép logarit hoá các biến xem xét thì người ta lại thấy chúng thỏa mãn mô hình hồi quy tuyến tính. Tức là chúng ta có mô hình sau đây:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_{i2} + ... + \beta_p \ln X_{ip} + U_i \qquad i = \overline{1,n}$$

Chúng ta có thể áp dụng phương pháp bình phương bé nhất như đã làm ở phần 3.3 để tìm các ước lượng $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_p$ bằng cách đặt $\ln Y = Z$ và $\ln X_j = V_j$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

10.6. Theo dõi mức đầu tư (Y) và lợi nhuận của 10 xí nghiệp khác nhau trong cùng một ngành, người ta thu được bảng số liệu sau đây:

X	1	1	2	3	3	5	6	7	8
Y	2.3	2.5	2.6	3	3.1	3.5	3.7	4.5	5

a. Tính hệ số tương quan giữa Y và X.

b. Tìm hàm hồi quy ước lượng của Y đối với X.

c. Tính R^2 và giải thích ý nghĩa của nó.

ĐS:

Source	SS	df	MS
Model	64.8977583	1	64.8977583
Residual	3.50224172	8	.43778015
Total	68.40		

Number of obs = 10
F (1,8) = 148.24
Prob > F = 0.0000
R – squared = 0,9488
AdjR- Squared = 0,9454
Root MSE =.66165

Y	Coef.	Std. Err	t	P > | t |	[95% Conf.	Interval
X	3.015695	.2476855	12.175	0.0000	2.444531	3.586859
Cons	-5.913667	.5725424	-6.778	0.0000	-7.925763	-3.901591

10.7. Thống kê năng suất một loại cây trồng (Y) và lượng đầu tư cho cải tạo đất X (triệu đ/ha) tại tỉnh A trong vòng 11 năm ta được bảng sau đây:

X	20	21	21	23	25	25	26	28	30	30	31
Y	2	2	3	3	4	5	6	6	7	8	8

a. Tính hệ số tương quan giữa Y và X.

b. Tìm hàm hồi quy ước lượng của Y đối với X.

c. Ước lương năng suất trung bình khi ta tăng mức đầu tư cho cải tạo đất lên 10 triệu đ/ha.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ĐS:

Source	SS	df	MS	Number of obs = 11
Model	146.818344	1	146.818344	F (1,9) = 167.07
Residual	7.90892857	9	.878769841	Prob > F = 0.0000
Total	154.77273	10	15.4727273	R – squared = 0,9489
				Adj R – squared = 0.9432
				Root MSE =.93743

Y	Coef.	Std. Err	t	P > | t |	[95% Conf.	Interval
X	1.698214	.1313832	12.926	0.000	1.401005	1.995424
Cons	17.11786	.7041855	24.309	0.000	15.52488	18.71084

10.8. Tìm hiểu mức tiêu dùng bia (Y) và thu nhập (X) của 14 người, ta thu được bảng số liệu sau đây:

X	1	1	2	3	3	5	6	7	8	8	9	10	10	10
Y	2.3	2.5	2.6	3	3.2	3.5	4	4	4.5	5	5.2	5.5	5.5	6

a. Tìm hàm hồi quy ước lượng của Y đối với X.
b. Nếu thu nhập là 6.5 triệu đồng/tháng thì ước tính lượng tiêu thụ sẽ là bao nhiêu?

ĐS:

Source	SS	df	MS	Number of obs = 14
Model	145.671724	1	145.671724	F (1, 12) = 332.53
Residual	5.25684699	12	.438070528	Prob > F = 0.0000
Total	150.928571	13	11.6098901	R – squared = 0,9652
				Adj R – squared = 0.9623
				Root MSE =.66187

Y	Coef.	Std. Err	t	P > | t |	[95% Conf.	Interval
X	2.689797	.147504	18.235	0.000	2.368413	3.011181
Cons	-4.98432	.624041	-7.987	0.000	-6.343988	-3.624651

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

10.9. Theo dõi sản lượng (Y) và lượng điện tiêu thụ cho 1 đơn vị sản phẩm (X) của một nhà máy trong 10 tháng, người ta thu được bảng số liệu sau đây:

X	100	110	115	120	130	130	135	140	145
Y	20	22	22	20	20	18	15	141	421

Sử dụng phương pháp phân tích hồi quy nghiên cứu mối quan hệ giữa sản lượng và chi phí điện năng cho 1 đơn vị sản phẩm của nhà máy nói trên. Điện năng tiêu thụ cho sản xuất ảnh hưởng như thế nào đến sự biến động về sản lượng của nhà máy? Liệu chúng ta có thể giảm mãi chi phí điện năng cho 1 đơn vị sản phẩm không? Tại sao?

10.10. Cho Y = (Doanh thu - triệu đ), X1 = (Giá của 1 đơn vị sản phẩm - 1000đ), X2 = (Số lượng công nhân). Sau đây là số liệu thống kê trong 10 tháng gần đây nhất:

Y	11	11	13	15	15	16	17	18	18	19
X_1	6.5	6.3	6.1	5.8	5.5	5.1	4.8	4.6	4.1	3.5
X_2	50	50	52	52	52	55	60	60	65	65

a. Tính ma trận hệ số tương quan

b. Tính hệ số tương quan riêng phần của Y đối với X1, X2.

c. Tìm hàm hồi quy ước lượng của Y đối với X1, X2.

d. Trong hai yếu tố: giá sản phẩm và lượng lao động, yếu tố nào ảnh hưởng đến sự biến động của doanh thu nhiều hơn? Tại sao?

ĐS:

	Y	X1	X2
Y	1.0000		
X1	-9.522	1.0000	
X2	0.9014	-0.9647	1.0000

Pratial correlation of Y with

Variable	Corr.	Sig.
X1	-0.7248	0.027
X2	-0.2141	0.580

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Source	SS	df	MS
Model	67.5078204	2	33.7539102
Residual	6.59217962	7	.941739946
Total	74.10	9	8.23333333

Number of obs = 10
F (2, 7) = 35.84
Prob > F = 0.0002
R – squared = 0,9110
Adj R – squared = 0.8856
Root MSE =.97043

Y	Coef.	Std. Err	t	P > | t |	[95% Conf.	Interval
X1	-3.462349	1.244083	-2.783	0.027	-6.40439	-.5205593
X2	-.1204985	.2077472	-0.580	0.580	-.6117426	.3707456
Cons	40.16806	17.01593	2.230	0.061	-2.432839	82.76895

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4. TÓM TẮT CHƯƠNG 11

1. Mối quan hệ phụ thuộc giữa các biến ngẫu nhiên có thể phân chia thành quan hệ phụ thuộc hàm số và quan hệ phụ thuộc thống kê, trong lớp các quan hệ phụ thuộc thống kê thì phụ thuộc tương quan và hồi quy có nhiều ứng dụng hơn cả trong thực tế.
2. Bảng số liệu thống kê dùng trong phân tích hồi quy và tương quan gồm bảng số liệu định lượng và bảng số liệu định tính.
3. Mục đích của phân tích tương quan là đo lường mức độ phụ thuộc giữa các biến ngẫu nhiên thông qua thông tin của mẫu, tức là trả lời câu hỏi xem các biến ngẫu nhiên có phụ thuộc nhau hay không và nếu có thì phụ thuộc chặt chẽ đến mức độ nào.
4. Để đo lường mức độ tương quan giữa các chỉ tiêu định lượng người ta dùng các loại hệ số tương quan khác nhau.
5. Để đo lường mức độ tương quan giữa các chỉ tiêu định tính người ta dùng các hệ số Q, hệ số F, hệ số liên hợp Pearson, Hệ số Kramer, hệ số tương quan hạng Spearman hoặc hệ số tương quan hạng Kendall.
6. Mục đích của phân tích hồi quy là xác định dạng thức của mối liên hệ phụ thuộc giữa các biến ngẫu nhiên trên cơ sở thông tin của mẫu, tức là trả lời câu hỏi xem các biến ngẫu nhiên phụ thuộc nhau như thế nào.
7. Các hàm hồi quy có thể là hồi quy đơn hoặc hồi quy bội, hồi quy tuyến tính hoặc phi tuyến.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

** Phân tích tương quan bảng số liệu định lượng*

+ Hệ số tương quan mẫu:

$$R_{XY} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\sqrt{\sum (Y_i - \overline{Y})^2}} = \frac{\overline{XY} - \overline{XY}}{\sqrt{MS_X MS_Y}}$$

+. Khoảng tin cậy cho hệ số tương quan ρ_{XY}.

$$(Z - \frac{R_{XY}}{2(n-1)} - U_{\alpha/2}\frac{1}{\sqrt{n-3}}; Z - \frac{R_{XY}}{2(n-1)} + U_{\alpha/2}\frac{1}{\sqrt{n-3}})$$

Trong đó $Z = \frac{1}{2}\ln\frac{1+R_{XY}}{R_{XY}}$ và người ta tính các giá trị của hàm $\frac{1}{2}\ln\frac{1+\rho}{\rho}$ bằng cách tra bảng.

+ Kiểm tra giả thuyết về giá trị của ρ.

Xét cặp giả thuyết:

H_0: ($\rho_{XY} = \rho_0$) H_1: ($\rho_{XY} \neq \rho_0$)

Ta có miền bác bỏ như sau:

$$W_\alpha = \left\{u = \left(\frac{1}{2}\ln\frac{1+R_{XY}}{1-R_{XY}} - \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1)}\right)\sqrt{n-3}; |u| > u_{\alpha/2}\right\}$$

Trường hợp đặc biệt, nếu X và Y đều tuân theo quy luật phân phối chuẩn thì với kích thước mẫu bất kỳ ta luôn có miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$T = R_{XY}\sqrt{\frac{n-2}{1-R_{XY}^2}}$$

Vậy miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{t = R_{XY}\sqrt{\frac{n-2}{1-R_{XY}^2}}; |t| > t_{\alpha/2}^{(n-2)}\right\}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Hệ số tương quan riêng phần.

Hệ số tương quan riêng phần của X_1 và X_2 khi X_3 cố định ký hiệu là $R_{12,3}$. Khi đó ta có công thức xác định $R_{12,3}$ như sau:

$$R_{12,3} = \frac{R_{12} - R_{13}R_{23}}{\sqrt{(1-R_{13}^2)(1-R_{23}^2)}}$$

Tổng quát ta có hệ số tương quan riêng phần của X_i và X_j khi cố định X_k là $R_{ij,k}$:

$$R_{ij,k} = \frac{R_{ij} - R_{ik}R_{jk}}{\sqrt{(1-R_{ik}^2)(1-R_{jk}^2)}}$$

Hệ số tương quan riêng phần giữa X_1 và X_2 khi cố định X_3, X_4 là:

$$R_{12,34} = \frac{R_{12,3} - R_{14,3}R_{24,3}}{\sqrt{(1-R_{14,3}^2)(1-R_{24,3}^2)}}$$

**Phân tích tương quan bảng số liệu định tính.*

+ Bảng ngẫu nhiên hai chiều.

A \ B	b_1	b_2	$\sum$
a_1	n_1	n_2	(n_1+n_2)
a_2	n_3	n_4	(n_3+n_4)
$\sum$	(n_1+n_3)	(n_2+n_4)	$n_1+n_2+n_3+n_4=n$

Chúng ta có thể đo mối tương quan giữa A và B bằng các hệ số Q và F xác định như sau:

+ Hệ số Q: $$Q = \frac{n_1n_4 - n_2n_3}{n_1n_4 + n_2n_3}$$

+ Hệ số F: $$F = \frac{n_1n_4 - n_2n_3}{\sqrt{(n_1+n_2)(n_1+n_3)(n_2+n_4)(n_3+n_4)}}$$

+ Hệ số liên hợp Pearson (P):

$$P = \sqrt{\frac{\chi^2}{n+\chi^2}} \qquad 0 \le P \le 1$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ *Hệ số Kramer (K)* :

$$K = \left\{ \frac{\chi^2}{n.\min(p-1, q-1)} \right\}^{1/2}$$

* *Các độ đo sự tương quan ở cấp đo theo thang thứ bậc.*

+ Hệ số tương quan hạng R_s của Spearman cho hai dấu hiệu A và B được đo cùng thang do thứ bậc xác định như sau:

$$R_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$$

Ký hiệu ρ_s là hệ số tương quan hạng Spearman của tổng thể thì để kiểm định cặp giả thuyết:

$$H_0 : (\rho_s = 0),\ H_1 : (\rho_s \neq 0)$$

Ta có miền bác bỏ như sau:

$$W_\alpha = \left\{ t = \sqrt{\frac{R_s^2(n-2)}{1-R_s^2}};\ |t| > t_{\alpha/2}^{(n-2)} \right\}$$

+ Hệ số tương quan hạng Kendall ký hiệu là τ và được xác định bằng công thức sau:

$$\tau = \frac{S}{n(n-1)/2}$$

$S = \sum (S_i^* - S_i^-) =$ Số điểm Kendall

$S_i^* =$ số hạng $k_j (j > i)$ thoả mãn $k_j > k_i$

$S_i^- =$ số hạng $k_j (j > i)$ thoả mãn $k_j < k_i$

Ký hiệu ρ_k là hệ số tương quan hạng Kendall của tổng thể thì để kiểm định cặp giả thuyết:

$$H_0 : (\rho_K = 0);\ H_1 : (\rho_K \neq)$$

Ta sử dụng miền bác bỏ sau đây:

$$W_\alpha = \left\{ u = \frac{\tau}{\sqrt{\frac{4n+10}{9n(n-1)}}};\ |u| > u_{\alpha/2} \right\}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$S = \sum (S_i^* - S_i^-)$ = Số điểm Kendall

S_i^* = số hạng $k_j (j > i)$ thoả mãn $k_j > k_i$

S_i^- = số hạng $k_j (j > i)$ thoả mãn $k_j < k_i$

* Phân tích hồi quy

+ Hàm hồi quy

$$f_Y(X_2, X_3, ..., X_p) = E(Y / X_2, X_3, ..., X_p)$$

+ Hàm hồi quy tuyến tính:

$$f_Y(X_2, X_3, ..., X_p) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p$$

+ Mô hình hồi quy tuyến tính đơn:

$$Y_i = aX_i + b + U_i \qquad i = \overline{1,n}$$

- Ước lượng tham số hồi quy:

$$\hat{a} = \frac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{X^2} - (\overline{X})^2} = r_{xy} \frac{\sqrt{MS_Y}}{\sqrt{MS_X}}$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X}$$

- Khoảng tin cậy của tham số hồi quy:

$$\left(\hat{a} - Se(\hat{a}) t_{\alpha/2}^{(n-2)} < a < \hat{a} + Se(\hat{a}) t_{\alpha/2}^{(n-2)}\right)$$

- Kiểm định giả thuyết về tham số hồi quy

$$H_0 : (a = a_0); H_1 : (a \neq a_0)$$

$$W_\alpha = \left\{ t = \frac{\hat{a} - a_0}{Se(\hat{a})}; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

Chú ý: Nếu thay a bằng b ta sẽ có công thức khoảng tin cậy và kiểm định giả thuyết về tham số b.

- Kiểm định sự phù hợp của hàm hồi quy:

$$H_0 : (R^2 = 0); H_1 : (R^2 \neq 0)$$

$$W_\alpha = \left\{ F = \frac{R^2}{\frac{(1-R^2)}{(n-2)}}; F > f_\alpha(1, n-2) \right\}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Dự báo giá trị trung bình $E(Y/X=X_0)$

$$f_Y(X_0) = E(Y/X = x_0) = aX_0 + b$$

$$\hat{Y}_0 = X_0'\hat{\beta}$$

$$Se(\hat{Y}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]}$$

$$\left(\hat{Y}_0 - t_{\alpha/2}^{(n-2)} se(\hat{Y}_0) < f_Y(X_0) < \hat{Y}_0 + t_{\alpha/2}^{(n-2)} se(\hat{Y}_0)\right)$$

+ Mô hình hồi quy tuyến tính bội:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + U_i$$

+ *Khoảng tin cậy :*

$$\left(\hat{\beta}_j - t_{\alpha/2}^{(n-k)} Se(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{\alpha/2}^{(n-k)} Se(\hat{\beta}_j)\right) \quad j = \overline{1,k}$$

+ *Kiểm định giả thuyết đối với* β_j *:*

Giả thiết	*Miền bác bỏ*
$H_0 : (\beta_j = \beta^*)$ $H_1 : (\beta_j < \beta^*)$	$W_\alpha = \left\{ t = \dfrac{\hat{\beta}_j - \beta^*}{Se(\hat{\beta}_j)}; t < -t_\alpha^{(n-k)} \right\}$
$H_0 : (\beta_j = \beta^*)$ $H_1 : (\beta_j > \beta^*)$	$W_\alpha = \left\{ t = \dfrac{\hat{\beta}_j - \beta^*}{Se(\hat{\beta}_j)}; t > t_\alpha^{(n-k)} \right\}$
$H_0 : (\beta_j = \beta^*)$ $H_1 : (\beta_j \neq \beta^*)$	$W_\alpha = \left\{ t = \dfrac{\hat{\beta}_j - \beta^*}{Se(\hat{\beta}_j)}; \lvert t \rvert > t_{\alpha/2}^{(n-k)} \right\}$

+ Sai số hồi quy:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} = \frac{RSS}{n-k}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Hệ số xác định của mô hình:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

+ *Kiểm định sự phù hợp của hàm hồi quy:*

$H_0 : (R^2 = 0)\ H_1 : (R^2 > 0)$

$$W_\alpha = \left\{ F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}; f > f_\alpha(k-1, n-k) \right\}$$

Các thuật ngữ cơ bản

Hàm hồi quy	Regression function
Hàm hồi quy tuyến tính	Linear Regression function
Hàm hồi quy đơn	Simple Regression function
Mô hình hồi quy đơn	Simple Regression Model
Hàm hồi quy bội	Multiple Regression function
Mô hình hồi quy bội	Multiple Regression Model
Hàm hồi quy tổng thể	Population Regression Function
Hàm hồi quy mẫu	Sample Regression function
Biến độc lập	independent variable
Biến phụ thuộc	dependent variable
Phương pháp bình phương bé nhất	method of least squares
Ước lượng (tham số) hồi quy	regression estimators
Sai số chuẩn	standard errors
Tham số hồi quy	regression parameters
Hệ số tương quan	correlation coeficient
Dự báo	prediction
Biến dự báo	predictor
Hệ số tương quan riêng phần	partial correlation coeficient
Hệ số xác định	coeficient of determination
Hiệu chỉnh (Điều chỉnh)	adjusted

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Câu hỏi ôn tập

11.11. Cho biết sự giống nhau và khác nhau giữa hàm hồi quy và mô hình hồi quy.

11.12. Khi nào các ước lượng bình phương bé nhất của các tham số hồi quy là các ước lượng không chệch tốt nhất? Hãy giải thích rõ các tính chất đó có nghĩa là gì?

11.13. Phân tích hồi quy và phân tích tương quan giống nhau và khác nhau ở những điểm nào?

11.14. Nội dung của phương pháp bình phương bé nhất trong phân tích hồi quy là gì? Phân biệt e_i và U_i như thế nào?

11.15. Khi nào thì 2 cặp giả thuyết sau đây là tương đương nhau:

a) $H_0 : (\beta_j = 0)$ $\quad H_1 : (\beta_j \neq 0)$

b) $H_0 : (R^2 = 0)$ $\quad H_1 : (R^2 > 0)$

11.16. Cho mô hình hồi quy:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad i = \overline{1,n}$$

Trong đó Y là nhu cầu hàng hoá A, X_2 là thu nhập của người tiêu dùng, X_3 là giá của hàng hoá A.

a) Viết hàm hồi quy ước lượng. Dấu các ước lượng của các tham số hồi quy như thế nào là phù hợp với thực tế?

b) Nếu thu nhập tăng một đơn vị thì nhu cầu trung bình tăng tối đa là bao nhiêu? (Viết công thức xác định).

c) Nếu giá của A tăng một đơn vị thì nhu cầu trung bình giảm tối đa bao nhiêu? (Viết công thức xác định).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập tổng hợp chương 11

11.17. Thống kê về thu nhập (X) và nhu cầu về một loại sản phẩm của 10 cá nhân được chọn một cách ngẫu nhiên người ta được bảng số liệu sau đây:

Y(SP/năm)	10	10	12	13	15	18	18	19	20	20
X(triệu đồng/năm)	30	32	35	35	38	40	42	45	48	50

Trên cơ sở số liệu đã cho, hãy tìm hàm hồi quy ước lượng của Y đối với X và trả lời một số câu hỏi sau đây (chọn α=5%):

1/ Cho biết ý nghĩa của các hệ số hồi quy. Thu nhập có thực sự ảnh hưởng đến nhu cầu về loại sản phẩm nói trên không?

2/ Thu nhập tăng 1 triệu đồng thì nhu cầu tăng ít nhất bao nhiêu?

3/ Có thể cho rằng thu nhập tăng 1 triệu đồng thì nhu cầu tăng 0,6 đơn vị được không?

4/ Thu nhập tăng thì nhu cầu có tăng thực sự không? Thu nhập tăng 1 triệu đồng thì nhu cầu tăng tối đa 0,6 đơn vị?

5/ Có ý kiến cho rằng mô hình hồi quy trên không phù hợp, theo bạn điều đó có đúng không? Ngoài yếu tố thu nhập còn yếu tố nào ảnh hưởng đến nhu cầu nữa không?

6/ Háy dự báo (ước lượng) nhu cầu (trung bình) khi thu nhập đạt mức 100 triệu đồng bằng khoảng tin cậy 95%?

Giải:

Lập bảng tính toán sau đây:

Stt	Y_i	X_i	Y_i^2	X_i^2	X_iY_i	Y^	(Y-Y^)^2
1	10	30	100	900	300	10.070	0.005
2	10	32	100	1024	320	11.213	1.471
3	12	35	144	1225	420	12.928	0.861
4	13	35	169	1225	455	12.928	0.005
5	15	38	225	1444	570	14.643	0.128
6	18	40	324	1600	720	15.786	4.903
7	18	42	324	1764	756	16.929	1.147
8	19	45	361	2025	855	18.644	0.127
9	20	48	400	2304	960	20.359	0.129
10	20	50	400	2500	1000	21.502	2.256
Tổng	155	395	2547	16011	6356		11.031
Tr. bình	15.5	39.5	254.7	1601.1	635.6		

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ *Từ bảng tính trên ta có:*

$MS_X = 40,85 \quad MS_Y = 14,45 \qquad S_X = 6,74 \quad S_Y = 4,01$

+ Hệ số tương quan mẫu: $R_{XY} = \dfrac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{MS_X}\sqrt{MS_Y}}$

$= \dfrac{635,6 - 15,5.39,5}{\sqrt{40,85}\sqrt{14,45}} = 0,8650$

$+ \hat{\beta}_1 = \dfrac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - (\overline{X})^2} = \dfrac{\overline{XY} - \overline{X}\,\overline{Y}}{MS_X} = \dfrac{635,6 - 39,5.15,5}{1601,1 - 39,5^2} = 0,5716$

$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 15,5 - 0,5716.39,5 = -7,0783$

+ Hàm hồi quy ước lượng: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = -7,0783 + 0,5716X$

1. $\hat{\beta}_1 = 0,5716$ có nghĩa là khi thu nhập tăng 1 đơn vị (1 triệu đồng) thi nhu cầu (trung bình) tăng (xấp xỉ) 0,5716 đơn vị.

$\hat{\beta}_0 = -7,0783$ có nghĩa là sản phẩm này là hàng hoá không thiết yếu, nhu cầu về hàng hóa này chỉ phát sinh (Y> 0) khi thu nhập đạt mức tối thiểu là $\dfrac{7,0783}{0,5716} = 12,3833$ (triệu đồng).

Để chứng tỏ thu nhập có thực sự ảnh hưởng đến nhu cầu về mặt hàng này không ta đi kiểm định cặp giả thiết sau đây:

$H_0 : (\beta_1 = 0) \quad H_1 : (\beta_1 \neq 0)$

Miền bác bỏ để kiểm định cặp giả thiết trên là:

$$W_\alpha = \left\{ t = \frac{\hat{\beta}_1 - 0}{Se(\hat{\beta}_1)}; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

Ta có $$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{RSS}{n-2} = \frac{11,031}{8} = 1,3789$$

$$Se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum (X_i - \overline{X})^2}} = \sqrt{\frac{\hat{\sigma}^2 / n}{MS_X}} = 0,0581$$

Do đó $t_{qs} = \dfrac{\hat{\beta}_1 - 0}{Se(\hat{\beta}_1)} = 9,84$, *tra bảng ta có* $t_{0,025}^{(8)} = 2,306$

$\rightarrow t_{qs} \in W_\alpha \rightarrow$ *Bác bỏ giả thiết* H_0.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Kết luận: Thu nhập thực sự ảnh hưởng đến nhu cầu về loại sản phẩm nói trên.

2. Ta phải tìm khoảng tin cậy tối thiểu 95% cho tham số β_i. Khoảng tin cậy đó là:

$$\left(\beta_1 \geq \hat{\beta}_1 - t_\alpha^{(n-2)} Se(\hat{\beta}_1)\right)$$

Thay các giá trị đã tính được ở phần trên ta được:

$\beta_1 \geq 0{,}5716 - 1{,}86.0{,}0581 \rightarrow \beta_1 \geq 0{,}46$

Có thể cho rằng thu nhập tăng 1 triệu đồng thì nhu cầu về loại sản phẩm nói trên tăng ít nhất 460.000 đ.

3. Ta phải kiểm định cặp giả thiết:

$$H_0: (\beta_1 = 0{,}6)\ H_1: (\beta_1 \neq 0{,}6)$$

Miền bác bỏ để kiểm định cặp giả thiết trên là:

$$W_\alpha = \left\{ t = \frac{\hat{\beta}_1 - 0{,}6}{Se(\hat{\beta}_1)}; |\ t\ | > t_{\alpha/2}^{(n-2)} \right\}$$

Tính giá trị quan sát và kết luận

$$t_{qs} = \frac{\hat{\beta}_1 - 0{,}6}{Se(\hat{\beta}_1)} = \frac{0{,}5716 - 0{,}6}{0{,}0581} = -0{,}49$$

$$|\ t_{qs}\ | < t_{\alpha/2}^{(n-2)} = t_{0{,}025}^{(8)} = 2{,}306$$

$\rightarrow t_{qs} \notin W_\alpha \rightarrow$ Chưa có cơ sở bác bỏ H_0.

Kết luận: Có thể cho rằng thu nhập tăng 1 triệu đồng thì nhu cầu tăng 0,6 đơn vị.

4. Ta phải kiểm định cặp giả thiết:

a/ $H_0: (\beta_1 = 0)\ H_1: (\beta_1 > 0)$

Miền bác bỏ để kiểm định cặp giả thiết trên là:

$$W_\alpha = \left\{ t = \frac{\hat{\beta}_1 - 0}{Se(\hat{\beta}_1)}; t > t_\alpha^{(n-2)} \right\}$$

Tính giá trị quan sát và kết luận

$t_{qs} = 9{,}84 \rightarrow t_{qs} > t_\alpha^{(n-2)} = t_{0{,}05}^{(8)} = 1{,}86 \rightarrow t_{qs} \in W_\alpha$

$\rightarrow$Bác bỏ H_0.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Kết luận: Thu nhập tăng thì nhu cầu có tăng thực sự.

b/ $H_0: (\beta_1 \le 0{,}6)$ $H_1: (\beta_1 > 0{,}6)$

Miền bác bỏ để kiểm định cặp giả thiết trên là:

$$W_\alpha = \left\{ t = \frac{\hat{\beta}_1 - 0{,}6}{Se(\hat{\beta}_1)}; t > t_\alpha^{(n-2)} \right\}$$

Tính giá trị quan sát và kết luận

$t_{qs} = 0{,}49$ và $t_{qs} < t_\alpha^{(n-2)} = t_{0,05}^{(8)} = 1{,}86 \rightarrow t_{qs} \notin W_\alpha$

→ Chưa có cơ sở bác bỏ H_0.

Kết luận: Thu nhập tăng 1 triệu đồng thì nhu cầu tăng tối đa 0,6 đơn vị là đúng.

5. Trước hết ta tính hệ số xác định của mô hình:

$$TSS = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n(\bar{Y})^2 = (n-1)S_Y^2 = 144{,}72$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{11{,}031}{144{,}72} = 0{,}9237$$

Để kết luận mô hình đang xem xét có ý nghĩa (phù hợp) không ta kiểm định cặp giả thiết sau đây:

$H_0: (R^2 = 0)$ $H_1: (R^2 > 0)$

Miền bác bỏ để kiểm định cặp giả thiết trên là:

$$W_\alpha = \left\{ F = \frac{R^2}{(1-R^2)/(n-2)}; f > f_\alpha(1, n-2) \right\}$$

Tính giá trị quan sát và kết luận.

$f_{qs} = 96{,}8 > f_\alpha^{(1,n-2)} = f_{0,05}^{(1,8)} = 5{,}32 \rightarrow f_{qs} \in W_\alpha$

→ Bác bỏ giả thiết H_0.

Điều này chứng tỏ mô hình là phù hợp và $R^2 = 0{,}9237$ có nghĩa là mô hình (tức là yếu tố thu nhập) giải thích được khoảng 92% sự biến động của nhu cầu về loại sản phẩm nói trên, còn lại khoảng 8% sự biến động của nhu cầu là do sai số ngẫu nhiên và các yếu tố khác chưa đưa vào mô hình. Các yếu tố đó có thể là giá của sản phẩm đó, giá của sản phẩm thay thế...

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6. Ta phải tìm khoảng tin cậy 95% cho $E(Y/X=X_0)$, với $X_0 = 100$:

$$\left(\hat{Y}_0 - t_{\alpha/2}^{(n-2)} Se(\hat{Y}_0) < E(Y/X=X_0) < \hat{Y}_0 + t_{\alpha/2}^{(n-2)} Se(\hat{Y}_0)\right)$$

Ta có: $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 =$ -7,0783 + 0,5716.100 = 50,082 và :

$$Se(\hat{Y}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right)} = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{nMS_X}\right)}$$

$$= \sqrt{1,3789\left(\frac{1}{10} + \frac{(100-39,5)^2}{10.40,85}\right)} = 3,523$$

Thay vào công thức khoảng tin cậy nói trên, ta được:

$$\left(41,9578 < E(Y/X=100) < 58,2062\right)$$

11.18. Cho số liệu về một số chỉ tiêu cơ bản của nền kinh tế quốc dân Việt Nam trong thời gian từ năm-1980 đến 1996 như sau:

Năm	gdp	ex	im	gip	gap
1980	16.80	0.30	1.30	6.80	9.70
1981	17.20	0.40	1.40	6.90	10.10
1982	18.70	0.50	1.50	7.50	11.20
1983	20.10	0.60	1.50	8.50	11.50
1984	21.80	0.60	1.70	9.60	12.20
1985	23.00	0.70	1.90	10.50	12.50
1986	23.80	0.80	2.20	11.20	13.10
1987	24.70	0.90	2.50	12.30	13.10
1988	25.90	1.00	2.80	14.00	13.70
1989	28.00	1.90	2.60	13.60	14.70
1990	29.50	2.40	2.80	14.00	14.90
1991	31.30	2.10	2.40	15.50	15.40
1992	34.00	2.60	2.60	18.10	16.60
1993	36.70	3.00	3.50	20.40	17.60
1994	40.00	4.10	5.90	23.20	18.60
1995	43.80	5.37	7.67	26.50	19.40
1996	47.90	7.10	11.00	30.23	20.35

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hãy phân tích mối quan hệ giữa thu nhập quốc dân(GDP) với các chỉ tiêu về tổng giá trị của các ngành kinh tế cơ bản: xuất khẩu (EX), Nhập khẩu (IM), sản xuất nông nghiệp (GAP) và sản xuất công nghiệp (GIP) bằng phương pháp phân tích tương quan và hồi quy.

11.19. Điều tra số lượng hành khách đi đến Singapo và Pháp cùng với hãng hàng không mà họ lựa chọn người ta được bảng số liệu sau:

Hãng hàng không	*Nước đến*	
	Xingapo	*Pháp*
I	*25*	*32*
II	*30*	*13*

Các lựa chọn của khách hàng có liên quan đến nhau không. Nếu có thì mức độ của nó như thế nào?

Đáp số:

$\chi^2 = 6.6470$ V = -0.2578 $\tau = -0.2578$

11.20. Hai nhà quản lý được mời tham gia đánh giá về trình độ quản lý của 12 cơ sở sản xuất. Sau đây là bảng xếp hạng về trình độ quản lý của 12 nhà máy đó của hai chuyên gia nói trên:

Chuyên gia 1	1	2	2	2	3	4	4	4	5	6	7	8
Chuyên gia 2	3	1	2	6	4	5	7	8	11	10	9	12

Cách đánh giá của hai nhà quản lý có sự khác biệt nhau không?

Đáp số:

Spearman's rho = 0.8866

Test of Ho: X, Y and y are independent

Prob > |t| = 0.0001

S = 46 $\rightarrow \tau$ = 0.6970

Test of Ho: x and y are independent

Prob > |z| = 0.0017 (continuity corrected)

11.21. Chuyên gia của hai bộ được trưng cầu ý kiến về 9 giải pháp khắc phục khủng hoảng tài chính. Các chuyên gia này đã xếp hạng cho 9 giải pháp đó như sau:

Chuyên gia bộ A	1	2	3	3	3	4	4	4	4
Chuyên gia bộ B	2	1	4	3	5	6	6	7	8

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Đáp số:

spearman x y = 0.9260

Test of Ho: x and y are independent

Prob > |t| = 0.0003

S = 25 $\rightarrow \tau$ = 0.6944

Test of Ho: x and y are independent

Prob > |z| = 0.0069 (continuity corrected)

11.22. Một hãng kinh doanh sơn có hệ thống đại lý ở 5 thành phố. Thống kê số lượng sơn các loại bán được ở các thành phố trên trong một năm (ngàn thùng) người ta được bảng số liệu như sau:

Thành phố (Địa bàn kinh doanh)	*Loại sơn*			
	Trắng	*Xanh*	*Vàng*	*Hồng*
A	*21*	*27*	*40*	*15*
B	*14*	*18*	*28*	*8*
C	*11*	*13*	*21*	*7*
D	*3*	*33*	*30*	*9*
E	*30*	*11*	*34*	*10*

Hãy phân tích mối quan hệ giữa địa bàn kinh doanh và số lượng các loại sơn bán được.

Đáp số: $\chi^2 = 33.9579$; V = 0.1719; τ = -0.0319

11.23. Ba người trong ban giám khảo cùng cho điểm về bài thi của 10 thí sinh trong một cuộc thi như sau:

Giám khảo 1	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Giám khảo 2	*3*	*10*	*7*	*2*	*8*	*5*	*6*	*9*	*1*	*4*
Giám khảo 3	*6*	*2*	*1*	*3*	*9*	*4*	*5*	*7*	*10*	*8*

Hãy chọn ra 2 giám khảo trong số đó mà bạn cho là cách đánh giá cho điểm phù hợp với nhau nhất.

Đáp số: $r_s(1,2) = -0.2121$; $r_s(1,3) = 0.6364$; $r_s(2,3) = -0,2970$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

11.24. Hãy đánh giá mức độ tương quan giữa tổng giá trị hàng hoá xuất khẩu (EX) và tiền trợ cấp hưu trí (HT) trên cơ sở số liệu của 10 năm như sau:

Cách đánh giá của hai nhà quản lý có khác biệt nhau không? Hãy cho kết luận với mức ý nghĩa 1%.

11.25. Cho tệp số liệu sau đây:
Contains data from C:\WINSTATA\XSTK11-5.DTA
obs: 28
vars: 2 8 Nov 2010 14:59
size: 280 (86.1% of memory free)

1. chiphi - Tổng chi phí các công ty ngành dệt
2. gia - Giá của một đơn vị sản phẩm

Stt	*chiphi*	*gia*	*Stt*	*chiphi*	*Gia*
1.	*493*	*8.20*	*15.*	*196*	*3.17*
2.	*410*	*7.39*	*16.*	*238*	*2.36*
3.	*451*	*7.68*	*17.*	*269*	*2.33*
4.	*723*	*9.88*	*18.*	*256*	*2.76*
5.	*329*	*5.65*	*19.*	*6.5*	*8.97*
6.	*432*	*7.10*	*20.*	*246*	*2.77*
7.	*294*	*5.17*	*21.*	*222*	*3.14*
8.	*270*	*3.34*	*22.*	*204*	*2.47*
9.	*311*	*5.63*	*23.*	*356*	*6.77*
10.	*194*	*1.39*	*24.*	*378*	*7.00*
11.	*640*	*9.30*	*25.*	*177*	*1.69*
12.	*217*	*2.21*	*26.*	*263*	*4.41*
13.	*272*	*2.88*	*27.*	*549*	*8.60*
14.	*401*	*6.94*	*28.*	*267*	*4.71*

Hãy phân tích sự biến động của tổng chi phí đối với giá sản phẩm bằng phương pháp hồi quy tuyến tính?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Đáp số:

Source	SS	df	MS	Number of obs =	28
				F(1, 26)	24.23
Model	294258.894	1	294258.894	Prob > F	0
Residual	315805.347	26	12146.3595	R-squared	0.4823
				Adj R-squared	0.4624
Total	610064.241	27	22594.9719	Root MSE	110.21

TC	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
P	39.71919	8.069723	4.92	0.000	23.13163	56.30674
_cons	119.5897	46.41137	2.58	0.016	24.18977	214.9896

11.26. Dựa vào hàm sản xuất Cobb- Duglass xây dựng mô hình hồi quy tuyến tính của sản lượng phụ thuộc vào vốn và lao động và kết quả ước lượng mô hình đó trên cơ sở số liệu sau đây để trả lời các câu hỏi:

Contains data from C:\WINSTATA\XSTK11-6.DTA
obs: 20
vars: 3 8 Nov 2010 15:15
size: 260 (86.1% of memory free)

1. san luong
2. von
3. lao dong

Stt	*sanluong*	*von*	*laodong*	*Stt*	*sanluong*	*von*	*laodong*
1.	*46226*	*2.00*	*2*	*11.*	*70778*	*2.00*	*4*
2.	*58865*	*5.66*	*2*	*12.*	*70658*	*5.66*	*2*
3.	*37382*	*2.00*	*4*	*13.*	*118409*	*5.66*	*4*
4.	*107915*	*5.66*	*4*	*14.*	*90536*	*2.00*	*6*
5.	*133026*	*2.00*	*6*	*15.*	*73843*	*10.38*	*4*
6.	*127848*	*10.38*	*4*	*16.*	*159804*	*5.66*	*6*
7.	*154107*	*5.66*	*6*	*17.*	*225100*	*10.38*	*4*
8.	*226500*	*10.38*	*4*	*18.*	*167678*	*10.38*	*6*
9.	*146649*	*10.38*	*6*	*19.*	*88699*	*5.66*	*4*
10.	*31448*	*2.00*	*2*	*20.*	*53852*	*2.00*	*2*

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

a) Khi số lượng lao động không đổi, nếu tăng vốn lên 1% thì sản lượng tăng ít nhất bao nhiêu%?

b) Khi vốn không đổi, nếu giảm lao động 1% thì sản lượng giảm tối đa bao nhiêu%?

c) Có thể cho rằng hiệu quả của vốn và lao động là như nhau không?

Đáp số:

Source	SS	df	MS	Number of obs =	20
				F(2, 17)	22.7
Model	4.502255	2	2.25112741	Prob > F	0
Residual	1.685961	17	.099174192	R-squared	0.7276
				Adj R-squared	0.6955
Total	6.188216	19	.325695583	Root MSE	0.31492

lnQ	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
lnK	0.435006	.109124	3.99	0.001	.204774	0.665237
lnL	0.724967	.18043	4.02	0.001	.3442932	1.105641
_cons	9.817008	.2600127	37.76	0.000	9.26843	10.36559

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội.1999.
2. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB ĐH Kinh tế quốc dân. Hà Nội.2008.
3. Nguyễn Cao Văn, Trần Thái Ninh. *Hướng dẫn trả lời lý thuyết và giải bài tập xác suất và thống kê toán*. NXB Tài chính. Hà Nội.2009.
4. Nguyễn Cao Văn, Bùi Dương Hải. *Hướng dẫn trả lời lý thuyết và giải bài tập Kinh tế lượng*. NXB Tài chính. Hà Nội. 2011.
5. Ngô Văn Thứ. *Thống kê thực hành*. NXB Khoa học và kỹ thuật. Hà Nội. 2008.
6. Nguyễn Quang Dong: *Kinh tế lượng nâng cao*. NXB KHKT. 2002.
7. Гмурман ВЕ Теория вероятностей и математическая статистика. Москва. 1972.
8. РАО С.Р. Линейные статистические методы и их применения. Москва. 1968.
9. Giard V. Statistique appliquée à la gestion. Economica. Paris. 1992
10. Amir D Aczel. Complete Business Statistics. Irwin Boston 1993.
11. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992
12. OTT L.R. And introduction to statistical methods and data analysis. California. 1993
13. Gujarati D. Basic econometrics, McGraw Hill. New york. 1995
14. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons. New york. 1990

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 12

PHÂN TÍCH CHUỖI THỜI GIAN

1. MỞ ĐẦU

Giả sử chúng ta cần nghiên cứu, phân tích, dự báo về một chỉ tiêu thống kê (biến ngẫu nhiên) nào đó, ví dụ đó là doanh số bán hàng của một công ty chẳng hạn. Trong các chương 10, 11 chúng ta đã nghiên cứu sự biến động của một biến ngẫu nhiên dưới tác động của một hay nhiều yếu tố khác nhau. Các yếu tố đó có thể là các chỉ tiêu định lượng hay định tính, tùy theo từng trường hợp mà ta có các bài toán phân tích thống kê tương ứng (phân tích phương sai, phân tích tương quan hay phân tích hồi quy). Trên thực tế có rất nhiều chỉ tiêu nghiên cứu mà sự biến động của nó có liên quan đến yếu tố thời gian. Nếu ta loại sự ảnh hưởng trực tiếp của các yếu tố khác (mà công việc này đôi khi là bắt buộc vì chúng ta không có hoặc không có đủ các thông tin cần thiết về các yếu tố đó) và chỉ giữ lại yếu tố thời gian thì chúng ta có một bài toán đặc thù gọi là bài toán phân tích chuỗi thời gian.

Định nghĩa: Chuỗi thời gian là tập hợp các giá trị quan sát được của một biến ngẫu nhiên được sắp xếp theo thời gian.

Chuỗi thời gian đôi khi còn được gọi là dãy số thời gian. Đơn vị thời gian có thể là ngày, tháng, quý, năm... Người ta ký hiệu chuỗi thời gian bằng các chữ cái X_t, Y_t, Z_t... Trong đó t được gọi là chỉ số thời gian. Ví dụ: X_t là "Năng suất cây trồng tại năm t", Y_t = (Doanh thu bán hàng của Công ty trong tháng t)... và t=1, 2, 3,... n.

Phân tích chuỗi thời gian là sử dụng các phương pháp thống kê khác nhau để làm rõ cấu trúc của chuỗi thời gian (tức là các thành phần của chuỗi) trong sự biến động của nó. Trên cơ sở đó người ta có thể thấy rõ hơn bản chất cũng như quy luật của hiện tượng, sự việc đang được nghiên cứu thông qua một chỉ tiêu thống kê cụ thể. Một ứng dụng phổ biến nữa của phân tích chuỗi thời gian đó là dự báo ngắn hạn về các giá trị của chuỗi thời gian.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giả sử giá trị của chuỗi thời gian tại thời điểm t và (t+1) là Y_t và Y_{t+1} thì khi đó hiển nhiên chúng ta có thể cho rằng:

$$Y_{t+1} = Y_t + \Delta_t$$

Trong đó Δ_t là lượng biến động ngẫu nhiên của Y khi chuyển từ thời điểm t sang (t+1). Ta cũng có thể viết:

$$\Delta_t = Y_{t+1} - Y_t$$

Δ_t còn được gọi là bước nhảy ngẫu nhiên của chuỗi thời gian tại thời điểm t. Tất nhiên Δ_t chưa biết và không thể xác định được một cách chính xác, do đó cách tốt nhất là chúng ta phải tập trung nghiên cứu bản thân chuỗi thời gian. Có thể thấy rõ rằng bất kỳ chuỗi thời gian nào cũng chứa đựng trong sự biến động của mình ít nhất một trong bốn thành phần (yếu tố) sau đây:

- Xu thế biến động
- Biến động theo mùa (hoặc thời vụ)
- Biến động theo chu kỳ
- Biến động ngẫu nhiên (không theo quy luật)

Tức là ta có thể nói rằng cấu trúc của chuỗi thời gian sẽ bao gồm bốn thành phần nói trên. Ta ký hiệu 4 thành phần nói trên như sau:

T_t - Thành phần xu thế cho ta biết xu hướng biến động của chuỗi (tăng hay giảm như thế nào) trong một khoảng thời gian tương đối dài.

S_t - Thành phần mùa vụ cho ta biết sự biến động của chuỗi trong hai hay nhiều khoảng thời gian (độ dài có thể khác nhau) liền nhau được lặp đi lặp lại trong suốt thời kỳ mà ta xem xét.

C_t - Thành phần chu kỳ cho ta biết mức độ biến động của chuỗi trong một khoảng thời gian nào đó cố định (được gọi là chu kỳ) và sẽ lặp đi lặp lại trong suốt thời kỳ nghiên cứu.

ε_t - Thành phần ngẫu nhiên cho ta biết mức độ biến động của chuỗi sau khi đã loại bỏ sự ảnh hưởng của tất cả các yếu tố nói trên.

Về mặt cấu trúc, chúng ta chỉ xem xét và nghiên cứu hai mô hình chuỗi thời gian là mô hình cộng và mô hình nhân sau đây.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1. Mô hình cộng: là mô hình mà các giá trị (thực) của chuỗi thời gian được viết dưới dạng biểu thức:

$$Y_t = T_t + S_t + C_t + \varepsilon_t$$

Mô hình này trên thực tế ít khi được sử dụng để phân tích vì nó không cho phép chúng ta phân tích được sự ảnh hưởng lẫn nhau giữa các yếu tố (thành phần) cấu thành nên chuỗi. Người ta chỉ sử dụng mô hình cộng khi có thể cho rằng chuỗi chỉ bao gồm hai trong ba thành phần (T_t,S_t,C_t) và ε_t đồng thời giả thiết rằng các thành phần đó tác động độc lập với nhau.

2. Mô hình nhân: là mô hình mà các giá trị (thực) của chuỗi thời gian được viết dưới dạng biểu thức:

$$Y_t = T_t.S_t.C_t.\varepsilon_t$$

Đây là mô hình hay được sử dụng nhất, trong đó T_t được biểu hiện bằng giá trị cùng đơn vị đo với Y_t còn các thành phần còn lại được đo bằng %. Ví dụ: Y_t = (Doanh thu của Công ty A tháng t) và:

$Y_9 = T_9.S_9.C_9.\varepsilon_9$ = 35 triệu đồng. 1,55. 0,92. 0,8 = 39,928 triệu đồng

Các con số trên được tính toán thế nào và ý nghĩa của chúng sẽ được trình bầy cụ thể trong các phần tiếp theo. Các ví dụ cụ thể để minh họa cho các phần sau chỉ sử dụng hai tệp số liệu sau đây:

Bảng 12.1

```
Contains data from C:\LUU TRU\SolieuSTATA\XSTK24.DTA
  obs:            12
 vars:             2                          7 Dec 2007 11:15
 size:            72 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
-------------------------------------------------------------------------------
thang           byte   %8.0g                 Tháng trong năm
y               byte   %8.0g                 Doanh thu trong tháng (triệu đ)
-------------------------------------------------------------------------------
     | thang    y |
  1. |     1   12 |
  2. |     2   18 |
  3. |     3   27 |
```

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.	4	22
5.	5	38
6.	6	42
7.	7	30
8.	8	45
9.	9	40
10.	10	47
11.	11	44
12.	12	48

Bảng 12.2

Contains data from C:\LUU TRU\SolieuSTATA\XSTK25.DTA

obs: 12
vars: 3 7 Dec 2007 11:15
size: 132 (99.9% of memory free)

variable name	storage type	display format	value label	variable label
nam	int	%8.0g		Năm kinh doanh
q	str3	%9s		Qúy trong năm
y	int	%8.0g		Doanh thu của Cty (triệu đ/qúy)

	nam	q	y
1.	1995	I	25
2.	.	II	29
3.	.	III	20
4.	.	IV	36
5.	1996	I	28
6.	.	II	32
7.	.	III	24
8.	.	IV	42
9.	1997	I	22
10.	.	II	35
11.	.	III	29
12.	.	IV	38

Mục 2 sẽ nghiên cứu các phương pháp phân tích chuỗi thời gian bằng hàm xu thế, mục 3 sẽ nghiên cứu phương pháp phân tích sự biến động mùa vụ, mục 4 sẽ đề cập đến phương pháp phân tích và dự báo Holt - Wilters, mục 5 sẽ trình bày phương pháp tách thành phần chu kỳ ra khỏi chuỗi thời gian.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2. PHÂN TÍCH BẰNG HÀM XU THẾ

2.1. Phân tích hồi quy chuỗi thời gian

Người ta cố gắng cụ thể T_t bằng một hàm số cụ thể nào đó, sao cho nó có thể chỉ ngay cho chúng ta thấy rõ xu thế biến động nói chung (trong một khoảng thời gian tương đối dài) của Y_t. Phổ biến người ta dùng phương pháp phân tích hồi quy, có nghĩa là người ta thay T_t bằng hàm hồi quy của Y đối với biến thời gian (t=1,2...,n). Ta có mô hình hồi quy tuyến tính như sau:

$$Y_t = \alpha t + \beta + \varepsilon_t \qquad (12.1)$$

Bằng phương pháp bình phương bé nhất chúng ta tìm được hàm hồi quy ước lượng $\hat{Y}_t = \hat{\alpha} t + \hat{\beta}$

$\hat{Y}_t$ chính là một dạng hàm xu thế cụ thể của chuỗi Y_t, đó cũng chính là một ước lượng của thành phần xu thế T_t. Tất nhiên chúng ta có thể xét mô hình hồi quy không phải là tuyến tính, miễn sao nó phải phản ánh được rõ nhất xu thế biến động của các giá trị chuỗi thời gian. Thông thường người ta dựa vào hình ảnh "đám mây" các quan sát của Y_t để lựa chọn mô hình hồi quy cho thích hợp.

Ví dụ 12.1. Sử dụng số liệu đã cho ở bảng 12.1, ta có kết quả phân tích hồi quy bằng phương pháp bình phương bé nhất của Y là doanh thu bán hàng đối với t là tháng được thực hiện bằng STATA như sau:

```
reg y t
```

Source	SS	df	MS	Number of obs = 12
Model	1320.21154	1	1320.21154	F(1, 10) = 40.16
Residual	328.705128	10	32.8705128	Prob > F = 0.0001
				R-squared = 0.8007
Total	1648.91667	11	149.901515	Adj R-squared = 0.7807
				Root MSE = 5.7333

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
thang	3.038462	.4794411	6.34	0.000	1.9702	4.106723
_cons	14.66667	3.528591	4.16	0.002	6.804477	22.52886

Từ kết quả trên ta có $\hat{Y}_t = \hat{\alpha} t + \hat{\beta} = 3{,}038462t + 14{,}66667$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Để vẽ đồ thị và so sánh các hàm xu thế khác nhau của Y_t ta tính các giá trị $\hat{Y}_t$ (ymu) bằng cách thay các giá trị của t lần lượt là 1,2,3...,n vào hàm hồi quy ước lượng nói trên. Sau đây là kết quả tính bằng STATA:

```
. predict ymu
. list ymu
```

	ymu ($\hat{Y}$)
1.	17.70513
2.	20.74359
3.	23.78205
4.	26.82051
5.	29.85897
6.	32.89743
7.	35.9359
8.	38.97436
9.	42.01282
10.	45.05128
11.	48.08974
12.	51.1282

2.2. Phương pháp bình quân di động (trung bình trượt) (MA)

Giả sử ta có chuỗi thời gian Y_t trong đó t = 1,2,3,...,n. Khi đó bình quân di động bậc hai (MA(2)) của Y_t (ký hiệu là $Y2_t$) được tính như sau:

$$Y2_1 = \frac{Y_1 + Y_2}{2}$$

$$Y2_2 = \frac{Y_2 + Y_3}{2}$$

.....

$$Y2_{n-1} = \frac{Y_{n-1} + Y_n}{2}$$

Tương tự ta tính MA(k) của Y_t theo công thức sau đây:

$$YK_1 = \frac{Y_1 + Y_2 + ... + Y_k}{k}$$

$$YK_2 = \frac{Y_2 + Y_3 + ... + Y_{k+1}}{k}$$

.....

$$YK_{n-k} = \frac{Y_{n-k+1} + Y_{n-k+2} + ... + Y_n}{k}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ví dụ 12.2. Cũng với số liệu đã cho ở bảng 12.1 ta tính các giá trị bình quân di động bậc 3 của Y_t như sau:

$$Y3_1 = \frac{12+18+27}{3} = 19$$

$$Y3_2 = \frac{18+27+22}{3} = 22{,}33$$

.....

$$Y3_{10} = \frac{47+44+48}{3} = 46{,}33$$

Làm tương tự ta sẽ có các giá trị của $Y5_t$. Sau đây là bảng các giá trị của $\hat{Y}_t$, $Y3_t$, $Y5_t$ được thực hiện bằng STATA:

```
. egen y3 = ma(y), t(3)
. egen y5 = ma(y), t(5)
. list ymu y3 y5
```

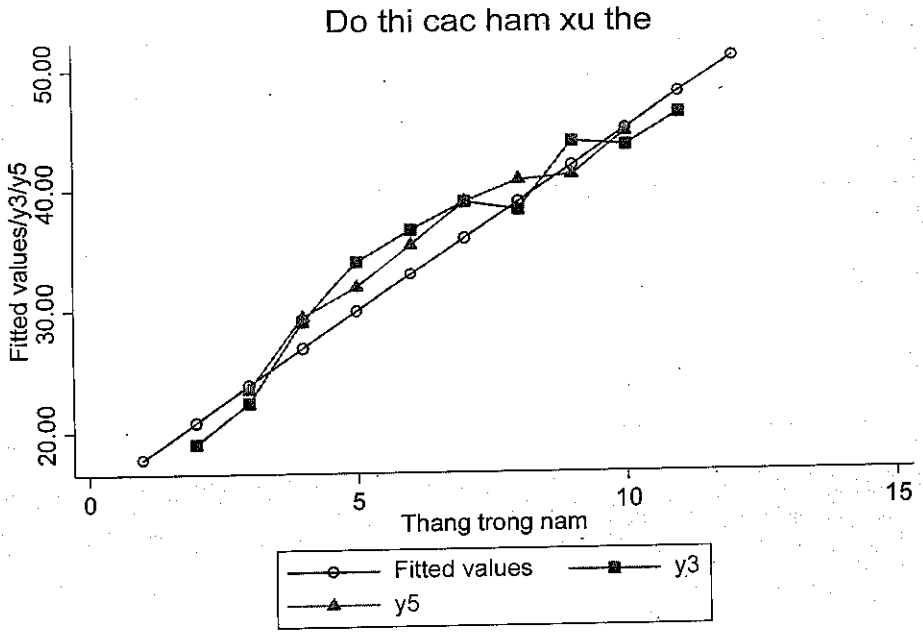
	ymu	y3	y5
1.	17.71	.	.
2.	20.74	19.00	.
3.	23.78	22.33	23.40
4.	26.82	29.00	29.40
5.	29.86	34.00	31.80
6.	32.90	36.67	35.40
7.	35.94	39.00	39.00
8.	38.97	38.33	40.80
9.	42.01	44.00	41.20
10.	45.05	43.67	44.80
11.	48.09	46.33	.
12.	51.13	.	.

Phương pháp bình quân di động thực chất là một phương pháp làm trơn số liệu (còn gọi là san bằng số liệu). Bằng cách làm này người ta loại bớt sự ảnh hưởng của các yếu tố khác đối với chuỗi thời gian. Với bậc càng cao thì dãy số liệu của chúng ta càng trơn hơn, tuy nhiên lúc đó số lượng các giá trị bình quân di động cũng ít đi. Việc lựa chọn số bậc trong phương pháp bình quân di động phải dựa vào đặc điểm biến động của các giá trị của chuỗi thời gian cũng như độ dài của chúng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sau đây là đồ thị của $\hat{Y}_t$ (Fitted values), $Y3_t$, $Y5_t$ được vẽ trên cùng một hệ trục tọa độ để có thể thấy rõ hơn mức độ biểu hiện xu thế của các phương pháp khác nhau đối với cùng một chuỗi thời gian.

```
. twoway (connected ymu thang, msymbol(circle_hollow)) (connected y3
thang, msymbol(square)) (connected y5 thang, msymbol(triangle)), title(Do
thi cac ham xu the)
```



2.3. Phương pháp san mũ

Đây cũng là một phương pháp làm trơn số liệu của chuỗi thời gian. Các giá trị san mũ (được dùng để thay thế cho T_t) được ký hiệu là F_t và xác định bằng công thức sau đây:

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t \quad t = \overline{1,n} \text{ và } F_1 = Y_1 \qquad (12.2)$$

Trong đó α là tham số nhận giá trị trong khoảng (0,1) và được gọi là hệ số san mũ. Ý nghĩa và cách chọn giá trị của α sẽ được nói rõ sau ví dụ về cách tính F_t sau đây.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Ví dụ 12.3. Sử dụng số liệu ở bảng 12.1 và chọn $\alpha = 0,2$ ta tính được các giá trị của F_t như sau:

$$F_2 = 0,2Y_1 + (1-0,2)F_1 = 0,2.12 + 0,8.12 = 12$$

$$F_3 = 0,2Y_2 + (1-0,2)F_2 = 0,2.18 + 0,8.12 = 13,2$$

.....

$$F_{12} = 0,2Y_{11} + (1-0,2)F_{11} = 0,2.44 + 0,8.35,06 = 36,85$$

Ta có bảng các giá trị của Y_t, $\hat{Y}_t$, F_t như sau:

Tháng	Y_t	$\hat{Y}_t$	F_t
1	12	17.71	12.00
2	18	20.74	12.00
3	27	23.78	13.20
4	22	26.82	15.96
5	38	29.86	17.17
6	42	32.9	21.33
7	30	35.94	25.47
8	45	38.97	26.37
9	40	42.01	30.10
10	47	45.05	32.08
11	44	48.09	35.06
12	48	51.13	36.85

Từ cách tính F_t chúng ta thấy F_t cũng như $\hat{Y}_t$ ngoài tác dụng làm rõ xu thế biến động của chuỗi thời gian người ta có thể dùng chúng để ước lượng giá trị của chuỗi trong tương lai. Tức là có thể dùng F_t và $\hat{Y}_t$ để dự báo ngắn hạn cho chuỗi thời gian (tất nhiên chưa tính đến tác động của các yếu tố khác như mùa vụ, chu kỳ...)

Một số nhận xét về phương pháp san mũ:

+ Thực chất giá trị san mũ F_t là trung bình cộng gia quyền của các giá trị thực tế Y_{t-1} và giá trị ước lượng (san mũ) F_{t-1}:

$$F_t = \alpha Y_{t-1} + (1-\alpha)F_{t-1}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thay biểu thức trên vào công thức (12.2) ta có:

$$F_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + (1-\alpha)^2 F_{t-1}$$

Bằng phương pháp quy nạp dễ dàng thấy rằng:

$$F_{t+1} = \alpha \sum_{k=0}^{n-1} (1-\alpha)^k Y_{t-k} + (1-\alpha)^n F_{t-(n-1)}$$

Vì $(o < \alpha < 1)$ nên khi $n \to \infty$ ta sẽ được:

$$F_{t+1} = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k Y_{t-k} \qquad (12.3)$$

+ Từ công thức trên ta thấy hệ số san mũ α có ý nghĩa như sau: Nếu α được chọn càng lớn thì (1 - α) càng nhỏ, điều này có nghĩa là các giá trị của chuỗi thời gian càng về quá khứ càng ít được chú trọng (ảnh hưởng của nó đến giá trị hiện tại của chuỗi càng ít). Ngược lại nếu α càng nhỏ thì các giá trị quá khứ của chuỗi càng được chú ý hơn.

+ Công thức (12.2) có thể viết lại như sau:

$$F_{t+1} = F_t + \alpha Y_t - \alpha F_t = F_t + \alpha(Y_t - F_t)$$

Đặt $e_t = Y_t - F_t$ = Sai số của ước lượng bằng san mũ tại thời điểm t và khi đó:

$$F_{t+1} = F_t + \alpha e_t$$

Từ đây ta cũng suy ra rằng nên chọn α sao cho $\sum_1^n e_t^2$ là nhỏ nhất. Tuy nhiên trong thực hành để xác định được α sao cho thỏa mãn điều kiện trên không phải là dễ. Hiện nay bằng các phần mềm thống kê người ta dễ dàng xác định được α và còn tính được tổng bình phương các sai số $\sum_1^n e_t^2$. Nếu làm bằng tay thì theo kinh nghiệm của các chuyên gia chúng ta nên chọn α nằm trong khoảng (0,1 - 0,4).

+ Cuối cùng chúng ta thấy phương pháp san mũ được thực hiện theo phép đệ quy, tức là để tính được F_{t+1} thì ta phải có F_t. Do đó giá trị ban đầu F_1 thường được chọn là Y_1 hoặc là giá trị trung bình của một số giá trị đầu

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

tiên của chuỗi thời gian. Trong các chương trình thống kê như STATA, EVIEWS... thì F_1 là giá trị trung bình *của một nửa số quan sát đầu tiên.*

Sau đây là kết quả san mũ với số liệu ở các bảng 12.1 và 12.2 bằng STATA với $\alpha = 0,2$ và $\alpha = 0,1$ tương ứng.

```
. use "C:\Program Files\Stata8\SolieuSTATA\XSTK24.DTA"
. gen t=m(2010m1)+_n-1
. format t %tm
. tsset t
        time variable:  t, 2010m1 to 2010m12
. tssmooth exp ysan=y,par(0.2)
        exponential coefficient  =       0.2000
        sum of squared residuals =       1359.1
        root mean squared error  =       10.642
. list y ysan
     y     ysan(F)
------------------
1.   12        26.5
2.   18        23.6
3.   27       22.48
4.   22      23.384
5.   38     23.1072
6.   42    26.08576
7.   30    29.26861
8.   45    29.41488
9.   40    32.53191
10.  47    34.02552
11.  44    36.62042
12.  48    38.09634
```

Nếu không có lựa chọn giá trị của α thì chương trình sẽ tự động chọn α sao cho $\sum_1^n e_t^2$ là nhỏ nhất. Kết quả sau đây cho thấy $\alpha = 0,6072$ thì $\sum_1^n e_t^2$ là nhỏ nhất và bằng 902,61609:

```
. tssmooth exp F=y
  computing optimal exponential  coefficient [0,1]
  optimal exponential coefficient =       0.6072
  sum of squared residuals        =    902.61609
  root mean squared error         =    8.6728316
. list y F
     +---------------+
     |  y          F |
     |---------------|
  1. | 12       26.5 |
  2. | 18   17.69535 |
```

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

```
  3. | 27    17.88034 |
  4. | 22    23.41796 |
  5. | 38    22.55695 |
     |----------------|
  6. | 42    31.93424 |
  7. | 30    38.04634 |
  8. | 45    33.16047 |
  9. | 40    40.34964 |
 10. | 47    40.13733 |
     |----------------|
 11. | 44    44.30446 |
 12. | 48    44.11959 |
     +----------------+

. use "C:\Program Files\Stata8\SolieuSTATA\XSTK25.DTA"
. gen t=q(1995q1)+_n-1
. format t %tq
. tsset t
        time variable:  t, 1995q1 to 1997q4
. tssmooth exp F=y,par(0.1)
        exponential coefficient  =       0.1000
        sum of squared residuals =       678.96
        root mean squared error  =        7.522
. list y F
     +----------------+
     |  y           F |
     |----------------|
  1. | 25   28.33333 |
  2. | 29         28 |
  3. | 20       28.1 |
  4. | 36      27.29 |
  5. | 28     28.161 |
     |----------------|
  6. | 32    28.1449 |
  7. | 24   28.53041 |
  8. | 42   28.07737 |
  9. | 22   29.46963 |
 10. | 35   28.72267 |
     |----------------|
 11. | 29    29.3504 |
 12. | 38   28.31536 |
     +----------------+
```


THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập

12.1. Cho số liệu về lượng tiền đã cho vay (tỷ đồng/năm) của một ngân hàng trong các năm từ 2001 - 2008 như sau :

276 294 234 352 368 441 386 440

Tìm hàm xu thế cho dãy số liệu nói trên.

12.2. Cho chuỗi số liệu sau :

Y = 53 65 74 85 92 105 120 128 144 158 179 195

Trong đó Y là số lượng của một tờ báo được phát hành (1000 tờ) từ năm 1996 - 2007. Hãy ước lượng thành phần xu thế của chuỗi thời gian nói trên bằng hàm hồi quy.

12.3. Theo dõi doanh thu (Triệu đồng/năm) từ cây mía của một trang trại từ Quý I/2007 đến quý IV/2009 thu được các số liệu sau :

126 122 125 119 121 120 119 117 118 116 122 115

Hãy ước lượng thành phần xu thế của chuỗi thời gian nói trên bằng hàm hồi quy và bình quân di động bậc 3.

12.4. Ước lượng thành phần xu thế của chuỗi thời gian sau đây bằng phương pháp san mũ với $\alpha = 0,4$ và $\alpha = 0,8$

925 940 924 925 912 908 910 912 915 924 943 962 960 958 955

12.5. Sau đây là số liệu về tiền lương trung bình (triệu đồng/người) của một doanh nghiệp từ Quý I/2007 đến quý IV/2010

0,54 1,15 1,13 1,69 1,53 1,79 1,34 1,38

2,02 2,38 2,27 1,82 2,07 3,15 2,92 3,73

Hãy ước lượng thành phần xu thế của chuỗi thời gian trên bằng phương pháp san mũ với $\alpha = 0,2$ và $\alpha = 0,6$.

3. PHÂN TÍCH BIẾN ĐỘNG MÙA VỤ

Như đã nêu ở §1, yếu tố mùa vụ (đôi khi gọi tắt là mùa) phải được hiểu là theo một nghĩa rộng. Mùa ở đây có thể là mùa thời tiết trong năm: xuân, hạ, thu, đông; có thể là các khoảng thời gian trong kinh doanh như: vụ tết, sau tết, cuối năm... hay đơn giản là các quý trong một năm...

Những khái niệm chúng ta sẽ xét trong phần này là: chỉ số mùa vụ (ký hiệu là I_S), tỷ lệ bình quân di động RMA (Ratio to Moving Average). Để

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

đơn giản cho việc trình bày cách xác định chỉ số mùa vụ I_S và tỷ lệ bình quân di động RMA chúng ta xét ví dụ cụ thể sau đây.

Ví dụ 12.4. Sử dụng số liệu đã cho ở bảng 12.2. Trước hết ta tìm hàm hồi quy ước lượng của Y đối với t (t = 1,2,3,...,12) tức là ta tìm hàm xu thế của Y. Các giá trị tính toán cho ở bảng sau:

Stt	Y	t	t^2	Y.t	$\hat{Y}$
1	25	1	1	25	25.92
2	29	2	4	58	26.66
3	20	3	9	60	27.41
4	36	4	16	144	28.15
5	28	5	25	140	28.89
6	32	6	36	192	29.63
7	24	7	49	168	30.37
8	42	8	64	336	31.11
9	22	9	81	198	31.85
10	35	10	100	350	32.59
11	29	11	121	319	33.34
12	38	12	144	456	34.08
Tổng	360	78	650	2446	
Tr.bình	30	6.5	54.17	203.83	

$$\hat{\beta}_1 = \frac{\overline{yt} - \bar{y}.\bar{t}}{\overline{t^2} - (\bar{t})^2} = \frac{203{,}83 - 30.6{,}5}{54{,}17 - 6{,}5^2} = 0{,}74$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{t} = 30 - 0{,}74 1.6{,}5 = 25{,}18$$

$$\rightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 t = 25{,}18 + 0{,}74t$$

Tiếp theo chúng ta tính tiếp bình quân di động bậc bốn Y4 của Y_t (vì một năm có 4 quý) và sau đó tính giá trị trung tâm bình quân di động CMA -Centered MA-(thực chất là bình quân di động bậc hai của Y4), tức là ta đã làm trơn chuỗi thời gian Y_t hai lần nhằm mục đích loại hết các yếu tố xu thế, chu kỳ có ảnh hưởng đến giá trị của chuỗi mà chỉ giữ lại ảnh hưởng của yếu tố mùa vụ (S_t) và yếu tố ngẫu nhiên (ε_t).

Giả sử chuỗi của chúng ta được cấu trúc theo mô hình nhân:

$$Y_t = T_t.S_t.C_t.\varepsilon_t$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Và nếu cho rằng với cách làm như trên thì CMA đã loại được hai yếu tố là T_t,C_t (tất nhiên ta cũng có thể thay thế CMA bằng Y4) thì khi đó ta có:

$$Y_t = T_t.S_t.C_t.\varepsilon_t = CMA.S_t.\varepsilon_t$$

hay $RMA_t = \frac{Y_t}{CMA_t} = S_t.\varepsilon_t$. RMA_t được gọi là tỷ số bình quân di động. Để đơn giản về mặt ký hiệu, khi không thật cần thiết, chúng ta sẽ bỏ chỉ số t. Có nghĩa là RMA bao gồm hai thành phần của chuỗi là S_t và ε_t. Sau đây là bảng tính các giá trị $\hat{Y}$ (Ymũ), Y4, CMA, RMA:

Bảng 12.3

Stt	Năm	Quý	Y_t	$\hat{Y}_t$	Y4	CMA	RMA
1	1995	I	25	25.92			
2	-	II	29	26.66	27.500		
3	-	III	20	27.40	28.250	27.875	0.717
4	-	IV	36	28.14	29.000	28.625	1.258
5	1996	I	28	28.88	30.000	29.500	0.949
6	-	II	32	29.62	31.500	30.750	1.041
7	-	III	24	30.36	30.000	30.750	0.780
8	-	IV	42	31.10	30.750	30.375	1.383
9	1997	I	22	31.84	32.000	31.375	0.701
10	-	II	35	32.58	31.000	31.500	1.111
11	-	III	29	33.32			
12	-	IV	38	34.06			

Theo ví dụ trên ta có 8 giá trị của CMA, ta xếp các gí trị này vào trung tâm của chuỗi, tức là từ quan sát thứ 3 đến quan sát thứ 10, sau đó lấy giá trị Y_t tương ứng chia cho giá trị của CMA_t ta được giá trị RMA_t. Ví dụ với Y_{10}=35 và CMA_{10}= 31,5 ta có RMA_{10}= 35/31,5 = 1,111, có nghĩa là với t thuộc quý II năm 1997 thì RMA = 1,111...

Tiếp theo chỉ số mùa vụ I_S (trong đó s = I,II.III,IV; tức là I_S có 4 giá trị tương ứng với 4 "mùa" kinh doanh) được tính như sau:

$$I_s = \frac{TrungbìnhRMA_t (t \in QuýS)}{\sum TrungbìnhRMA_t / 4}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sau đây là bảng tính các giá trị I_S:

Quý	1995	1996	1997	TbRMA	Is
I		0.949	0.701	0.825	0.831
II		1.041	1.111	1.076	1.084
III	0.717	0.780		0.749	0.755
IV	1.258	1.383		1.320	1.330
				3.970	4.000

Ví dụ ta có:

$$I_{IV} = \frac{TrungbìnhRMA_t (t \in QuýIV)}{\sum TrungbìnhRMA_t / 4} = \frac{(1{,}258 + 1{,}383)/2}{3{,}970/4} = 1{,}33$$

Ta có thể phân tích ý nghĩa của chỉ số mùa vụ như sau:

- Trong 4 quý của cả năm thì quý một (chỉ số mùa vụ là 0,831) yếu tố mùa vụ đóng góp (ảnh hưởng) 83,1% tới doanh thu của quý này sau khi đã tính tác động của yếu tố xu thế nhưng chưa tính đến tác động của yếu tố chu kỳ (vì đã bị loại bỏ trong quá trình tính I_S). Tương tự ta thấy quý bốn, yếu tố mùa vụ tác động rõ nét nhất tới kết quả kinh doanh, nó làm cho doanh thu của quý IV tăng thêm trung bình đến 33% so với mức doanh thu của quý này nếu chỉ tính tác động của yếu tố xu thế. Điều này có nghĩa là: theo số liệu trên vào quý I năm 1997 theo xu thế chung thì doanh thu sẽ đạt khoảng 31,84 triệu đồng, nhưng do ảnh hưởng của yếu tố mùa vụ doanh thu chỉ còn lại là 31,84x0,831 = 26,459 triệu đồng. Như vậy so với doanh thu thực tế quý I năm 1997 là 22 triệu đồng thì vẫn còn chênh lệch khoảng 4 triệu đồng. Sai số này là do ta chưa tính đến tác động của yếu tố chu kỳ và yếu tố ngẫu nhiên (bất quy tắc). Tương tự, ta cũng có thể phân tích doanh thu quý IV năm 1997 như sau: Theo xu thế thì doanh thu quý này là 34,06 triệu đồng, nhưng do tác động của yếu tố mùa vụ (ví dụ quý bốn là thời gian mua sắm của các doanh nghiệp hay dân cư để chuẩn bị cho Tết...) nên doanh thu đã tăng lên thành 34,06x33%= 11,24 triệu đồng, tức là sẽ đạt 45,3 triệu đồng. So với doanh thu thực tế quý IV là 38 triệu thì chênh lệch khoảng -7 triệu đồng, điều này có nghĩa là hai yếu tố còn lại: yếu tố chu kỳ và yếu tố ngẫu nhiên có tác động ngược chiều đến kết quả kinh doanh.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Từ những phân tích dễ dàng thấy rằng chúng ta sẽ tính được giá trị của chuỗi thời gian sau khi loại bỏ tác động của yếu tố mùa vụ bằng cách lấy giá trị thực của chuỗi chia cho chỉ số mùa vụ tương ứng. Ví dụ doanh thu quý I năm 1997 sau khi loại bỏ yếu tố (thành phần) mùa vụ là: $\frac{22}{0,831} = 26,47$ triệu đồng. Giá trị này chúng ta sẽ gọi là giá trị không mùa vụ. Tương tự ta có giá trị không mùa vụ của quý IV năm 1997 là 28,57 triệu đồng. So sánh hai giá trị này (hoặc chỉ số mùa vụ ta thấy yếu tố mùa vụ của quý I và quý IV tác động ngược chiều đến kết quả kinh doanh của công ty.

- Dự báo giá trị của chuỗi thời gian. Như đã trình bày ở phần §2, nếu dùng $\hat{Y}$ để ước lượng các giá trị thực của chuỗi thời gian thì những giá trị đó chỉ mới bao gồm sự ảnh hưởng của yếu tố xu thế. Để dự báo chính xác hơn, chúng ta cần kết hợp với chỉ số mùa vụ để tác động của yếu tố này cũng được đưa vào xem xét trong quá trình dự báo. Ví dụ với t= 13 ta có giá trị ước lượng doanh thu của công ty vào quý I năm 1998 khi chưa xem xét sự ảnh hưởng của yếu tố mùa vụ là:

$$\hat{Y}_t = \hat{\alpha}t + \hat{\beta} = 3,038462t + 14,66667$$

$$= 3,038462 \times 13 + 14,66667 = 54,167$$

Với giả thiết chuỗi thời gian thỏa mãn điều kiện mô hình nhân thì chúng ta có giá trị ước lượng $\tilde{Y}_{13}$ có tính đến tác động của yếu tố mùa vụ là:

$$\tilde{Y}_{13} = \hat{Y}.I_S = 54,167.\ 0,831 = 45,01$$

Như trên đã phân tích giá trị dự báo trên chưa tính đến tác động của yếu tố chu kỳ và yếu tố ngẫu nhiên. Sau đây là bảng các giá trị $Y_t, \hat{Y}_t, \tilde{Y}_t$ để các bạn hãy so sánh.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bảng 12.4

Stt	Năm	Quý	Y_t	$\hat{Y}_t$	I_S	$\widetilde{Y}_t = \hat{Y}.I_S$
1	1995	I	25.0	25.92	0.831	21.549
2	-	II	29.0	26.66	1.084	28.898
3	-	III	20.0	27.40	0.755	20.676
4	-	IV	36.0	28.14	1.330	37.428
5	1996	I	28.0	28.88	0.831	24.010
6	-	II	32.0	29.62	1.084	32.107
7	-	III	24.0	30.36	0.755	22.910
8	-	IV	42.0	31.10	1.330	41.366
9	1997	I	22.0	31.84	0.831	26.471
10	-	II	35.0	32.58	1.084	35.315
11	-	III	29.0	33.32	0.755	25.143
12	-	IV	38.0	34.06	1.330	45.303

Chú ý: Trong trường hợp tổng quát yếu tố mùa vụ bao gồm những khoảng thời gian không đều nhau thì chúng ta có thể áp dụng kỹ thuật hồi quy theo biến giả để tính $\widetilde{Y}$. Tức là ước lượng giá trị của chuỗi thời gian bao gồm cả hai yếu tố xu thế và mùa vụ bằng cách đưa vào mô hình hồi quy các biến độc lập là t và các biến giả (Về kỹ thuật hồi quy theo biến giả các bạn có thể tham khảo cuốn giáo trình Kinh tế lượng của trường ĐHKTQD). Trở lại với ví dụ nói trên, giả sử ta biến mùa vụ là v (v nhận các giá trị là Q1,Q2,Q3,Q4 - 4 quý trong năm) thì với sự trợ giúp của STATA chúng ta dễ dàng tạo ra các biến giả tương ứng với v là v1,v2,v3,v4 và sau đó hồi quy theo t và các biến giả đó. Giá trị ước lượng được ký hiệu là ytv. Sau đây là các kết quả chạy bằng STATA để các bạn tham khảo và so sánh.

```
. tab v,gen(v)

      v |      Freq.     Percent        Cum.
--------+-----------------------------------
      I |          3       25.00       25.00
     II |          3       25.00       50.00
    III |          3       25.00       75.00
     IV |          3       25.00      100.00
--------+-----------------------------------
  Total |         12      100.00

. regress y t v1 v2 v3
      Source |       SS       df       MS              Number of obs =      12
```


THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

```
-------------+------------------------------          F( 4,    7) =  10.70
       Model |  433.166667     4   108.291667          Prob > F      =  0.0042
    Residual |  70.8333333     7   10.1190476          R-squared     =  0.8595
-------------+------------------------------          Adj R-squared =  0.7791
       Total |       504    11   45.8181818          Root MSE      =  3.181

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |       .4375   .2811673     1.56   0.164    -.2273551    1.102355
          v1 |  -12.35417   2.730847    -4.52   0.003    -18.81159   -5.896739
          v2 |  -5.791667    2.65749    -2.18   0.066    -12.07563    .4922981
          v3 |  -13.89583   2.612487    -5.32   0.001    -20.07338   -7.718284
       _cons |   35.16667   2.903884    12.11   0.000     28.30007    42.03326
------------------------------------------------------------------------------

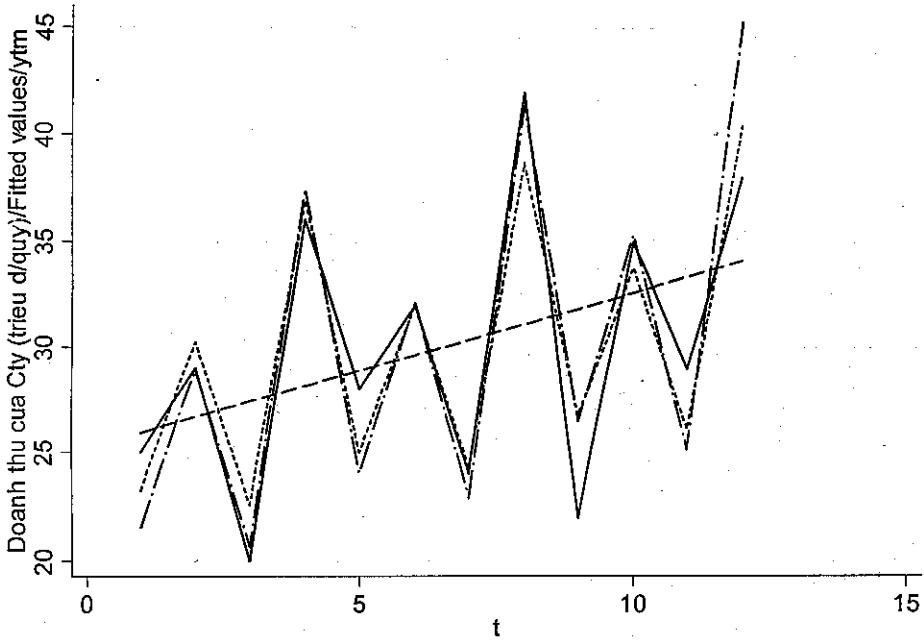
. predict Ytv
```

Trên cơ sở bảng các giá trị của $Y_t, \hat{Y}_t, \widetilde{Y}_t = \hat{Y}.I_S, Ytv$ chúng ta vẽ đồ thị của các biến đó trên cùng một hệ trục tọa độ để so sánh (Xem bảng 12.5 và hình vẽ sau đây).

Bảng 12.5

Stt	Năm	Quý	Y_t	$\hat{Y}_t$	$\widetilde{Y}_t = \hat{Y}.I_S$	Ytv
1	1995	I	25.0	25.92	21.549	23.25
2	-	II	29.0	26.66	28.898	30.25
3	-	III	20.0	27.40	20.676	22.58
4	-	IV	36.0	28.14	37.428	36.92
5	1996	I	28.0	28.88	24.010	25.00
6	-	II	32.0	29.62	32.107	32.00
7	-	III	24.0	30.36	22.910	24.33
8	-	IV	42.0	31.10	41.366	38.67
9	1997	I	22.0	31.84	26.471	26.75
10	-	II	35.0	32.58	35.315	33.75
11	-	III	29.0	33.32	25.143	26.08
12	-	IV	38.0	34.06	45.303	40.42

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



4. PHƯƠNG PHÁP SAN MŨ HOLT - WINTERS

4.1. Phương pháp san mũ Holt-Winters không mùa vụ

Đây cũng là một phương pháp làm trơn số liệu của chuỗi thời gian. Các giá trị san mũ theo Hol-Winters (được dùng để ước lượng cho Y_t) được cấu tạo bởi hai thành phần là Y_t^* và T_t. Trong đó T_t là lượng biến động do xu thế (có thể hiểu là Y_t-Y_{t-1}) và ước lượng của phần còn lại ngoài xu thế là Y_t^*. Y_t^* xác định bằng công thức san mũ sau đây:

$$Y_t^* = \alpha Y_t + (1-\alpha)(Y_{t-1}^* + T_{t-1}) \qquad (12.1)$$

Và vì T_t cũng được coi như là một chuỗi thời gian nên chúng ta cũng có thể xác định T_t theo công thức san mũ giản đơn:

$$T_t = \beta(Y_t^* - Y_{t-1}^*) + (1-\beta)T_{t-1} \qquad (12.2)$$

Để tính Y_t^* và T_t ta phải có hai giá trị xuất phát là Y_1 và T_1. Có nhiều cách để chọn, ví dụ trong Stata chúng được chọn là ước lượng hệ số chặn và

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

hệ số góc của hàm hồi quy Y theo t được xác định trên tập hợp một nửa số quan sát đầu tiên. Cách đơn giản hơn người ta chọn $Y_1^* = Y_1$ và $T_1 = Y_2 - Y_1$.

Ví dụ 12.5: Trên cơ sở bảng số liệu 12.2 với α=0,2 và β=0,8 chúng ta tính được các Y_t^* và T như sau:

Stt	Năm	Quý	Y	Y*	T
1	1995	I	25	25	4
2	-	II	29	29	4
3	-	III	20	30.4	1.92
4	-	IV	36	33.056	2.5088
5	1996	I	28	34.05184	1.298432
6	-	II	32	34.68022	0.762388
7	-	III	24	33.15408	-1.06843
8	-	IV	42	34.06853	0.517866
9	1997	I	22	32.06911	-1.49596
10	-	II	35	31.45853	-0.78766
11	-	III	29	30.33669	-1.055
12	-	IV	38	31.02535	0.33993

Dự báo giá trị của chuỗi tại thời điểm t=n+k là Y_{n+k}^* và Y_{n+k}^* xác định bằng công thức sau đây:

$$Y_{n+k}^* = Y_n^* + kT_n$$

Ví dụ với t = 13,14 ta có giá trị ước lượng doanh thu của công ty vào quý I và II năm 1998 khi chưa xem xét sụ ảnh hưởng của yếu tố mùa vụ là:

$$Y_{n+k}^* = Y_{13}^* = Y_{12}^* + 1.T_{12} = 31,02535 + 0,33993 = 31,059343$$

$$Y_{n+k}^* = Y_{14}^* = Y_{12}^* + 2.T_{12} = 31,02535 + 2.0,33993 = 31,399273$$

4.2. Phương pháp san mũ Holt-Winters có mùa vụ

Trong phần này chúng ta giả thiết rằng chuỗi thời gian Y được cấu tạo ngoài hai thành phần là Y_t^* và T_t còn thêm thành phần I_t là yếu tố mùa vụ. Nếu trong một năm có s thời kỳ thì I sẽ có s giá trị khác nhau, có nghĩa là tại thời điểm t năm nay cũng yếu tố của thời vụ đó nhưng tính cho năm trước sẽ là I_{t-s}. Cũng với giả thiết cấu trúc của chuỗi là mô hình nhân ta công thức xác định các giá trị Y_t^*, T_t và I_t như sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$Y_t^* = \alpha \frac{Y_t}{I_{t-s}} + (1-\alpha)(Y_{t-1}^* + T_{t-1})$$

$$T_t = \beta(Y_t^* - Y_{t-1}^*) + (1-\beta)T_{t-1} \qquad (12.3)$$

$$I_t = \gamma \frac{Y_t}{Y_t^*} Y_t + (1-\gamma) I_{t-s}$$

Ví dụ 12.6: Vẫn sử dụng số liệu đã cho ở bảng 12.2, vì có 4 quý nên ta có s = 4. để tính được I_t theo công thức trên ta phải có 4 giá trị xuất phát của yếu tố mùa vụ. Để tính các giá trị này ta tính bình quân di động bậc bốn MA(4) của Y_t và sau đó tính giá trị trung tâm bình quân di động CMA (tức là ta đã làm trơn chuỗi thời gian Y_t hai lần nhằm mục đích loại hết các yếu tố xu thế, chu kỳ có ảnh hưởng đến giá trị của chuỗi - xem thêm §3). Tiếp theo ta tính Y_t^*, T_t và I_t như sau:

Vì s = 4 nên giá trị CMA đầu tiên sẽ ứng với t=3 là 27,875 do đó giá trị đầu tiên của Y_t^* là $Y_3 = 27,875$. Giá trị T_t tương ứng sẽ là $T_3 = (CMA_4 - CMA_3) = 28,625 - 27,875 = 0,75$. Giá trị mùa vụ tương ứng đầu tiên I_3(tức là quý III) sẽ là $I_t = \frac{Y_t}{CMA_t} = \frac{Y_3}{CMA_3} = \frac{20}{27,875} = 0,717489$, tương tự ta tính tiếp các giá trị I_4, $I_{5-4} = I_1$, $I_{6-4} = I_2$. Tiếp theo chúng ta tính các giá trị của Y_t^*, T và I_t theo công thức (12.3) trong đó $\alpha = 0,2$; $\beta = 0,9$ và $\gamma = 0,1$. Sau đây là kết quả tính toán được thực hiện trên EXCEL:

Stt	Năm	Quý	Y	MA(4)	CMA	Y*	T	Is
1	1995	I	25					
2	-	II	29	27.500				
3	-	III	20	28.250	27.875	27.875	0.750	0.717489
4	-	IV	36	29.000	28.625	28.625	0.75	1.257642
5	1996	I	28	30.000	29.500	29.400	0.7725	0.949153
6	-	II	32	31.500	30.750	30.288	0.87645	1.04065
7	-	III	24	30.000	30.750	31.622	1.287849	0.721637
8	-	IV	42	30.750	30.375	33.007	1.375405	1.259125
9	1997	I	22	32.000	31.375	32.141	-0.64123	0.922685
10	-	II	35	31.000	31.500	31.927	-0.25735	1.046211
11	-	III	29			33.373	1.275715	0.736371
12	-	IV	38			33.755	0.471336	1.245789

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Công thức dự báo giá trị của chuỗi tại thời điểm n+k (trong đó k=1,2,...s) tức là dự báo cho khoảng thời gian tròn một chu kỳ gồm s thời kỳ có tính đến các yếu tố xu thế và thời vụ như sau:

$$Y^*_{n+k} = (Y^*_n + kT_n)I_{n+k-s} \quad k= 1,2,...s$$

Tổng quát ta có $Y^*_{n+k} = (Y^*_n + kT_n)I_{n+k-ms}$ nếu $t=n+k$ là thời điểm rơi vào chu kỳ thứ m tính từ thời điểm n.

Ví dụ ta cần dự báo doanh thu của công ty vào quý I và II năm 1998 có xem xét sụ ảnh hưởng của yếu tố mùa vụ là:

$$Y^*_{12+1} = (Y^*_{12} + 1.T_{12})I_{12+1-4} = (33,755 + 1.0,471336).0,922685 = 31,58$$

$$Y^*_{12+2} = (Y^*_{12} + 2.T_{12})I_{12+2-4} = (33,755 + 2.0,471336).1,046211 = 36,301$$

Bạn đọc hãy so sánh kết quả dự báo bằng phương pháp san mũ Holt-Winters với kết quả dự báo đã làm ở §2, §3. Sau đây là kết quả thực hiện bằng STATA

```
.tssmooth hw yxuthe=y,par(0.2 0.8)
Specified weights:
                alpha = 0.2000
                 beta = 0.8000
sum of squared residuals = 601.0504
 root mean squared error = 7.077254
. tssmooth sh ymuavu=y
 Optimal weights:
                alpha = 0.0844
                 beta = 1.0000
                gamma = 0.0000
   sum of squared residuals = 147.4381
   root mean squared error  = 3.505212
. list y yxuthe ymuavu
     |  y        yxuthe      ymuavu |
  1. |  25     24.90476     25.17125 |
  2. |  29     26.31048     29.00401 |
  3. |  20     28.66537     20.88547 |
```

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.	36	27.36283	36.84353
5.	28	30.90274	28.50001
6.	32	31.67023	32.45913
7.	24	33.13699	23.08506
8.	42	31.24848	40.84006
9.	22	35.05791	31.73615
10.	35	32.01619	34.43169
11.	29	32.66022	24.04944
12.	38	31.38982	42.86272

Bài tập

12.6. Với các số liệu của bài tập 12.3:

a. Hãy tính trung bình di động 4 thời kỳ

b. Hãy tính chỉ số mùa vụ

c. Hãy tìm chỉ số mùa vụ chung

d. Tìm chuỗi doanh thu đã loại trừ yếu tố mùa vụ

e. Hãy dùng phương pháp Holt - Winters để mô hình hóa chuỗi có yếu tố xu thế với $\alpha = 0,2$ và $\beta = 0,6$

12.7. Với các số liệu của bài tập 12.5:

a. Hãy vẽ đồ thị của chuỗi thời gian. Có dấu hiệu của yếu tố mùa vụ trong chuỗi này hay không?

b. Hãy dùng phương pháp Holt - Winters để mô hình hóa chuỗi có yếu tố xu thế, có yếu tố thời vụ với $\alpha = 0,2$; $\beta = 0,6$ và $\gamma = 0,75$. Tính các sai số dự báo.

c. Hãy tìm các giá trị tối ưu của α, β và γ. So sánh sai số dự báo với các sai số này ở câu b.

d. Hãy dự báo cho các quý của năm 2011.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5. PHÂN TÍCH BIẾN ĐỘNG CHU KỲ

Với giả thiết chuỗi thời gian cấu trúc theo mô hình nhân: $Y_t = T_t.S_t.C_t.\varepsilon_t$

chúng ta có: $$C_t.\varepsilon_t = \frac{Y_t}{T_t.S_t} = \frac{Y}{TS}$$

Như trên đã trình bày, chúng ta đã ước lượng được các thành phần T_t và S_t bằng $\hat{Y}$ và I_S do đó ta có thể ước lượng được mức độ ảnh hưởng của hai thành phần còn lại của chuỗi là yếu tố chu kỳ và ngẫu nhiên $(C_t.\varepsilon_t)$ bằng cách chia Y_t cho $\tilde{Y}_t = \hat{Y}.I_S$. Quá trình tính toán được thể hiện trong bảng 12.6 sau đây:

Bảng 12.6

Stt	Năm	Quý	Y_t	$\hat{Y}_t$	I_S	$\tilde{Y}_t = \hat{Y}.I_S$	$(Y_t / \tilde{Y}_t)$	$\hat{C}_t$ =MA(4)
1	1995	I	25.0	25.92	0.831	21.549	1.160	
2	-	II	29.0	26.66	1.084	28.898	1.004	
3	-	III	20.0	27.40	0.755	20.676	0.967	1.023
4	-	IV	36.0	28.14	1.330	37.428	0.962	1.025
5	1996	I	28.0	28.88	0.831	24.010	1.166	1.023
6	-	II	32.0	29.62	1.084	32.107	0.997	1.043
7	-	III	24.0	30.36	0.755	22.910	1.048	1.056
8	-	IV	42.0	31.10	1.330	41.366	1.015	0.973
9	1997	I	22.0	31.84	0.831	26.471	0.831	0.971
10	-	II	35.0	32.58	1.084	35.315	0.991	0.998
11	-	III	29.0	33.32	0.755	25.143	1.153	0.954
12	-	IV	38.0	34.06	1.330	45.303	0.839	

Sau khi tính được $(Y_t / \tilde{Y}_t)$, đây là ước lượng của $(C_t.\varepsilon_t)$ và để loại sự ảnh hưởng của yếu tố ngẫu nhiên ta tính trung bình trượt bậc 4 của dãy số $(Y_t / \tilde{Y}_t)$. Giá trị nhận được ở cột cuối cùng trong bảng trên sẽ là ước lượng của thành phần chu kỳ và được ký hiệu là $\hat{C}_t$.

Ví dụ: Với giả thiết chuỗi thời gian trên thỏa mãn điều kiện mô hình nhân thì chúng ta có giá trị ước lượng doanh thu của công ty vào quý I năm 1998 (t=13) có tính đến tác động của ba yếu tố xu thế, mùa vụ và chu kỳ là:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$Y^*_{13} = \hat{Y}_{13}.I_{13}.\hat{C}_{13} = 54{,}167.\ 0{,}831.0{,}971 = 43{,}707$$

Hoặc chúng ta có thể thay $\hat{C}_t$ bằng $(Y_t / \tilde{Y}_t)$ nếu không muốn loại tác động của yếu tố ngẫu nhiên - điều này tùy thuộc vào điều kiện thực tế. Có nghĩa là ta cũng có thể dự báo doanh thu của công ty vào quý I năm 1998 là:

$$Y^*_{13} = \hat{Y}_{13}.I_{13}.(Y_{13} / \tilde{Y}_{13}) = 54{,}167.\ 0{,}831.0{,}831 = 37{,}406$$

Để kết thúc, chúng tôi xin đưa ra kết quả phân tích về lợi nhuận của một hãng bia (Y- trăm triệu đồng/tháng) trong ba năm 2008, 2009, 2010. Trước hết các bạn hãy thử làm trên Excel, sau đó hãy so sánh với kết quả của chúng tôi và tự mình đưa ra những kết luận, dự báo cho tình hình kinh doanh của hãng trong năm tới. Sau đây là các kết quả đó.

$$\hat{Y}_t = \hat{\alpha}t + \hat{\beta} = 24.37 + 0.0955*t$$

Stt	Năm	Tháng	Y_t	$\hat{Y}_t$	Y3	CMA	RMA
1	2008	1	19.6	24.47			
2	.	2	18.6	24.56	20.46667		
3	.	3	23.2	24.66	22.1	21.283	1.090
4	.	4	24.5	24.75	25.13333	23.617	1.037
5	.	5	27.7	24.85	27.4	26.267	1.055
6	.	6	30	24.94	28.8	28.100	1.068
7	.	7	28.7	25.04	30.83333	29.817	0.963
8	.	8	33.8	25.13	29.2	30.017	1.126
9	.	9	25.1	25.23	27	28.100	0.893
10	.	10	22.1	25.33	23	25.000	0.884
11	.	11	21.8	25.42	21.6	22.300	0.978
12	.	12	20.9	25.52	22	21.800	0.959
13	2009	1	23.3	25.61	21.43333	21.717	1.073
14	.	2	20.1	25.71	23.83333	22.633	0.888
15	.	3	28.1	25.80	24.93333	24.383	1.152
16	.	4	26.6	25.90	27.76667	26.350	1.009
17	.	5	28.6	25.99	29.5	28.633	0.999
18	.	6	33.3	26.09	32.06667	30.783	1.082
19	.	7	34.3	26.18	32.2	32.133	1.067

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Stt	Năm	Tháng	Y_t	$\hat{Y}_t$	Y3	CMA	RMA
20	.	8	29	26.28	29.9	31.050	0.934
21	.	9	26.4	26.38	26.83333	28.367	0.931
22	.	10	25.1	26.47	24.6	25.717	0.976
23	.	11	22.3	26.57	22.56667	23.583	0.946
24	.	12	20.3	26.66	22.4	22.483	0.903
25	2010	1	24.6	26.76	22.56667	22.483	1.094
26	.	2	22.8	26.85	25.26667	23.917	0.953
27	.	3	28.4	26.95	26.13333	25.700	1.105
28	.	4	27.2	27.04	28.06667	27.100	1.004
29	.	5	28.6	27.14	28.36667	28.217	1.014
30	.	6	29.3	27.24	32.06667	30.217	0.970
31	.	7	38.3	27.33	33.2	32.633	1.174
32	.	8	32	27.43	31.73333	32.467	0.986
33	.	9	24.9	27.52	28.2	29.967	0.831
34	.	10	27.7	27.62	24.93333	26.567	1.043
35	.	11	22.2	27.71	23.8	24.367	0.911
36	.	12	21.5	27.81			

Tháng	2008	2009	2010	TbRMA	Is
1		1.073	1.094	1.084	1.083
2		0.888	0.953	0.921	0.920
3	1.090	1.152	1.105	1.116	1.115
4	1.037	1.009	1.004	1.017	1.016
5	1.055	0.999	1.014	1.022	1.021
6	1.068	1.082	0.970	1.040	1.039
7	0.963	1.067	1.174	1.068	1.067
8	1.126	0.934	0.986	1.015	1.014
9	0.893	0.931	0.831	0.885	0.884
10	0.884	0.976	1.043	0.968	0.967
11	0.978	0.946	0.911	0.945	0.944
12	0.959	0.903		0.931	0.930
				12.010	12.000

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Năm	Tháng	Y_t	$\hat{Y}_t$	I_S	$\widetilde{Y}_t = \hat{Y}.I_S$	$(Y_t / \widetilde{Y}_t)$	$\hat{C}_t$ = MA(3)
2008	1	19.6	24.466	1.083	26.48677	0.740	
.	2	18.6	24.561	0.920	22.59411	0.823	0.802
.	3	23.2	24.657	1.115	27.48979	0.844	0.880
.	4	24.5	24.752	1.016	25.14821	0.974	0.976
.	5	27.7	24.848	1.021	25.38103	1.091	1.074
.	6	30	24.943	1.039	25.91093	1.158	1.108
.	7	28.7	25.039	1.067	26.71554	1.074	1.186
.	8	33.8	25.134	1.014	25.49501	1.326	1.175
.	9	25.1	25.230	0.884	22.30796	1.125	1.118
.	10	22.1	25.325	0.967	24.48291	0.903	0.979
.	11	21.8	25.421	0.944	23.9958	0.908	0.897
.	12	20.9	25.516	0.930	23.73045	0.881	0.877
2009	1	23.3	25.612	1.083	27.72745	0.840	0.857
.	2	20.1	25.707	0.920	23.64833	0.850	0.889
.	3	28.1	25.803	1.115	28.76748	0.977	0.946
.	4	26.6	25.898	1.016	26.31255	1.011	1.022
.	5	28.6	25.994	1.021	26.55164	1.077	1.106
.	6	33.3	26.089	1.039	27.1014	1.229	1.178
.	7	34.3	26.185	1.067	27.9383	1.228	1.181
.	8	29	26.280	1.014	26.65747	1.088	1.149
.	9	26.4	26.376	0.884	23.32126	1.132	1.067
.	10	25.1	26.471	0.967	25.5908	0.981	1.001
.	11	22.3	26.567	0.944	25.07758	0.889	0.896
.	12	20.3	26.662	0.930	24.79626	0.819	0.852
2010	1	24.6	26.758	1.083	28.96813	0.849	0.864
.	2	22.8	26.853	0.920	24.70256	0.923	0.906
.	3	28.4	26.949	1.115	30.04517	0.945	0.953
.	4	27.2	27.044	1.016	27.4769	0.990	0.989
.	5	28.6	27.140	1.021	27.72225	1.032	1.019
.	6	29.3	27.235	1.039	28.29187	1.036	1.127
.	7	38.3	27.331	1.067	29.16105	1.313	1.166
.	8	32	27.426	1.014	27.81993	1.150	1.162
.	9	24.9	27.522	0.884	24.33455	1.023	1.070
.	10	27.7	27.617	0.967	26.69869	1.038	0.970
.	11	22.2	27.713	0.944	26.15935	0.849	0.906
.	12	21.5	27.808	0.930	25.86206	0.831	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6. TÓM TẮT CHƯƠNG 12

1. Chuỗi thời gian là tập hợp các giá trị quan sát được của một biến ngẫu nhiên được sắp xếp theo thời gian. Chuỗi thời gian đôi khi còn được gọi là dãy số thời gian thường được kí hiệu là Y_t và $t=1, 2, 3,... n$.
2. Có thể thấy rõ rằng sự biến động của bất kỳ chuỗi thời gian nào cũng chứa đựng trong mình sự tác động của ít nhất một trong bốn thành phần (yếu tố) sau đây: biến động theo xu thế, biến động theo mùa (hoặc thời vụ), biến động theo chu kỳ, biến động ngẫu nhiên (không theo quy luật), tức là ta có thể nói rằng cấu trúc của chuỗi thời gian sẽ bao gồm bốn thành phần nói trên và ký hiệu tương ứng là T_t, S_t, C_t, ε_t.
3. Về mặt cấu trúc, chúng ta chỉ xem xét và nghiên cứu hai mô hình chuỗi thời gian là mô hình cộng và mô hình nhân.
4. Các phương pháp phân tích thành phần xu thế bao gồm: phân tích hồi quy chuỗi thời gian, phương pháp bình quân di động (MA), phương pháp san mũ.
5. Để phân tích sự biến động mùa vụ người ta dùng chỉ số mùa vụ, giá trị trung tâm bình quân di động, tỷ lệ số bình quân di động và phương pháp san mũ Holt - Winters.
6. Sau khi loại bỏ các thành phần trên ta thu được thành phần phản ánh sự biến động theo chu kỳ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các ký hiệu và công thức cơ bản

* Mô hình cộng: $Y_t = T_t + S_t + C_t + \varepsilon_t$

* Mô hình nhân: $Y_t = T_t.S_t.C_t.\varepsilon_t$

* Phân tích hồi quy chuỗi thời gian

$$Y_t = \alpha t + \beta + \varepsilon_t$$

* Phương pháp bình quân di động (MA)

* Bình quân di động bậc k MA(k) của chuỗi Y_t:

$$YK_1 = \frac{Y_1 + Y_2 + ... + Y_k}{k}$$

$$YK_2 = \frac{Y_2 + Y_3 + ... + Y_{k+1}}{k}$$

.....

$$YK_{n-k} = \frac{Y_{n-k+1} + Y_{n-k+2} + ... + Y_n}{k}$$

* Phương pháp san mũ

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t \quad t = \overline{1,n} \text{ và } F_1 = Y_1$$

* Phân tích biến động mùa vụ

Giả sử chuỗi của chúng ta được cấu trúc theo mô hình nhân:

$$Y_t = T_t.S_t.C_t.\varepsilon_t$$

Thì khi đó ta có:

$$Y_t = T_t.S_t.C_t.\varepsilon_t = CMA.S_t.\varepsilon_t$$

Hay $$RMA_t = \frac{Y_t}{CMA_t} = S_t.\varepsilon_t$$

$$I_S = \frac{\text{Trung bình RMA}_t\ (t \in \text{Mùa S})}{\sum \text{Trung bình RMA}_t / \text{Số mùa}}$$

* Phương pháp san mũ Holt-Winters

* San mũ Holt-Winters không mùa vụ

$$Y_t^* = \alpha Y_t + (1-\alpha)(Y_{t-1}^* + T_{t-1})$$

$$T_t = \beta(Y_t^* - Y_{t-1}^*) + (1-\beta)T_{t-1}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Dự báo giá trị của chuỗi tại thời điểm t = n+k:

$$Y^*_{n+k} = Y^*_n + kT_n$$

* San mũ Holt-Winters có mùa vụ

Với giả thiết cấu trúc của chuỗi là mô hình nhân ta công thức xác định các giá trị Y^*_t, T_t và I_t(yếu tố mùa vụ) như sau:

$$\begin{aligned} Y^*_t &= \alpha \frac{Y_t}{I_{t-s}} + (1-\alpha)(Y^*_{t-1} + T_{t-1}) \\ T_t &= \beta(Y^*_t - Y^*_{t-1}) + (1-\beta)T_{t-1} \\ I_t &= \gamma \frac{Y_t}{Y^*_t} Y_t + (1-\gamma) I_{t-s} \end{aligned} \qquad (12.3)$$

Công thức dự báo giá trị của chuỗi tại thời điểm n+k (trong đó k=1,2,...s) tức là dự báo cho khoảng thời gian tròn một chu kỳ gồm s thời kỳ có tính đến các yếu tố xu thế và thời vụ như sau:

$Y^*_{n+k} = (Y^*_n + kT_n) I_{n+k-s}$ k= 1,2,...s

* Phân tích biến động chu kỳ

Với giả thiết chuỗi thời gian cấu trúc theo mô hình nhân:

$Y_t = T_t.S_t.C_t.\varepsilon_t$ và ký hiệu ước lượng của Y_t là Y^*_t thì chúng ta có:

$Y^*_t = \hat{Y}_t.I_t.\hat{C}_t$ trong đó $\hat{C}_t$ được tính như sau:

Hoặc $\hat{C}_t = (Y_t / \tilde{Y}_t) = Y_t / (\hat{Y}_t.I_S)$ nếu không muốn loại tác động của yếu tố ngẫu nhiên hoặc $\hat{C}_t$ là trung bình trượt bậc k (k là số mùa vụ) của dãy số $(Y_t / \tilde{Y}_t)$ nếu muốn loại tác động của yếu tố ngẫu nhiên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Chuỗi thời gian	time series
Phân tích xu thế	trend analysis
Phân tích mùa vụ	seasonality analysis
Phân tích chu kỳ	cyclycal analysis
Phương pháp bình quân di động	moving average method
Phương pháp san mũ	exponential smoothing methods
Mô hình nhân	multiplicative model
Mô hình cộng	additive model
Phương pháp san mũ Holt - Winters	Holt - Winters exponential smoothing methods

Câu hỏi ôn tập

12.8. Chuỗi thời gian có là biến ngẫu nhiên không? Phân biệt chuỗi thời gian với biến ngẫu nhiên nói chung như thế nào?

12.9. Phân tích chuỗi thời gian và phân tích hồi quy khác nhau và giống nhau ở những điểm nào?

12.10. Căn cứ vào đặc điểm gì của chuỗi thời gian để người ta phân tích sự biến động của chúng. Người ta chia chuỗi thời gian thành mấy loại để phân tích. Đặc điểm của mỗi loại đó là gì?

12.11. Mô tả sự biến động của chuỗi thời gian theo cách nào là đơn giản nhất? Có những phương pháp nào để phân tích chuỗi thời gian?

12.12. Phân tích thành phần xu thế là gì? Có bao nhiêu phương pháp phân tích thành phần xu thế? Nêu nội dung chính của các phương pháp đó.

12.13. Làm thế nào để người ta có thể tách yếu tố mùa vụ ra khỏi chuỗi thời gian để nghiên cứu? Ước lượng thành phần mùa vụ trong mô hình nhân của chuỗi thời gian được ký hiệu là gì?

12.14. Chỉ số mùa vụ I_S của một chuỗi thời gian cụ thể có bao nhiêu giá trị? Cho ví dụ minh họa.

12.15. Ước lượng thành phần chu kỳ trong mô hình nhân của chuỗi thời gian được ký hiệu là gì? Công thức dự báo giá trị của chuỗi (Mô hình nhân) với đầy đủ các thành phần được viết như thế nào? Nêu ý nghĩa của các thành phần đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập tổng hợp chương 12

12.16. Cho số liệu về doanh thu (Y-triệu đồng/quý) của một doanh nghiệp theo quý như sau:

Stt	1	2	3	4	5	6	7	8	9	10	11	12
Năm	1995	-	-	-	1996	-	-	-	1997	-	-	-
Quý	I	II	III	IV	I	II	III	IV	I	II	III	IV
Y	221.3	206.2	195.7	203	199.1	193.7	196.3	199	189.7	182.7	189.6	190.3

Hãy phân tích chuỗi thời gian nói trên với giả thiết cho rằng cấu trúc của chuỗi là mô hình nhân.

ĐS:

$\hat{Y}_t = \hat{\alpha}t + \hat{\beta}$ = 211.8894 - 2.257*t

Stt	Năm	Quý	Y_t	$\hat{Y}_t$	MA(4)	CMA	RMA
1	1995	I	221.30	209.63			
2	-	II	206.20	207.38	206.550		
3	-	III	195.70	205.12	201.000	203.775	0.960
4	-	IV	203.00	202.86	197.875	199.438	1.018
5	1996	I	199.10	200.60	198.025	197.950	1.006
6	-	II	193.70	198.35	197.025	197.525	0.981
7	-	III	196.30	196.09	194.675	195.850	1.002
8	-	IV	199.00	193.83	191.925	193.300	1.029
9	1997	I	189.70	191.58	190.250	191.088	0.993
10	-	II	182.70	189.32	188.075	189.163	0.966
11	-	III	189.60	187.06			
12	-	IV	190.30	184.81			

Quý	1995	1996	1997	TbRMA	Is
I		1.006	0.993	0.999	1.005
II		0.981	0.966	0.973	0.979
III	0.960	1.002		0.981	0.987
IV	1.018	1.029		1.024	1.029
				3.978	4.000

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Năm	Quý	Y_t	$\hat{Y}_t$	I_S	$\tilde{Y}_t = \hat{Y}.I_S$	$(Y_t / \tilde{Y}_t)$	$\hat{C}_t$ =MA(4)
1995	I	221.3	209.6	1.005	210.664	1.050	
-	II	206.2	207.4	0.979	202.966	1.016	
-	III	195.7	205.1	0.987	202.428	0.967	1.001
-	IV	203.0	202.9	1.029	208.838	0.972	0.986
1996	I	199.1	200.6	1.005	201.592	0.988	0.981
-	II	193.7	198.3	0.979	194.130	0.998	0.993
-	III	196.3	196.1	0.987	193.518	1.014	0.999
-	IV	199.0	193.8	1.029	199.544	0.997	0.999
1997	I	189.7	191.6	1.005	192.519	0.985	0.996
-	II	182.7	189.3	0.979	185.294	0.986	0.999
-	III	189.6	187.1	0.987	184.608	1.027	1.000
-	IV	190.3	184.8	1.029	190.250	1.000	

12.17. Cho số liệu về sản lượng than (ngàn tấn/quý) khai thác được của một liên doanh như sau:

Năm	1995	-	1996	-	-	-	1997	-	-	-	1998	-
Quý	III	IV	I	II	III	IV	I	II	III	IV	I	II
	11387	12264	9817	11493	12211	14367	10440	12527	13832	15399	11276	13531

Hãy phân tích chuỗi thời gian nói trên với giả thiết cho rằng cấu trúc của chuỗi là mô hình nhân.

ĐS: $\hat{Y}_t = \hat{\alpha}t + \hat{\beta} = 10958.62 + 218.4685*t$

Stt	Năm	Quý	Y_t	$\hat{Y}_t$	MA(4)	CMA	RMA
1	1995	III	11387.00	11177.09			
2	-	IV	12264.00	11395.56	11240.250		
3	1996	I	9817.00	11614.03	11446.250	11343.250	0.865
4	-	II	11493.00	11832.49	11972.000	11709.125	0.982
5	-	III	12211.00	12050.96	12127.750	12049.875	1.013
6	-	IV	14367.00	12269.43	12386.250	12257.000	1.172
7	1997	I	10440.00	12487.90	12791.500	12588.875	0.829
8	-	II	12527.00	12706.37	13049.500	12920.500	0.970
9	-	III	13832.00	12924.84	13258.500	13154.000	1.052
10	-	IV	15399.00	13143.31	13509.500	13384.000	1.151
11	1998	I	11276.00	13361.77			
12	-	II	13531.00	13580.24			

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Quý	1996	1997	1998	TbRMA	Is
I	0.865	0.829		0.847	0.844
II	0.982	0.970		0.976	0.971
III	1.013	1.052		1.032	1.028
IV	1.172	1.151		1.161	1.157
				4.017	4.000

Năm	Quý	Y_t	$\hat{Y}_t$	I_S	$\tilde{Y}_t = \hat{Y}.I_S$	$(Y_t / \tilde{Y}_t)$	$\hat{C}_t$ =MA(4)
1995	III	11387.0	11177.1	1.028	11491.813	0.991	
-	IV	12264.0	11395.6	1.157	13179.116	0.931	
1996	I	9817.0	11614.0	0.844	9800.466	1.002	0.981
-	II	11493.0	11832.5	0.971	11495.043	1.000	0.979
-	III	12211.0	12051.0	1.028	12390.293	0.986	1.000
-	IV	14367.0	12269.4	1.157	14189.764	1.012	0.997
1997	I	10440.0	12487.9	0.844	10537.882	0.991	1.001
-	II	12527.0	12706.4	0.971	12343.995	1.015	1.015
-	III	13832.0	12924.8	1.028	13288.774	1.041	1.015
-	IV	15399.0	13143.3	1.157	15200.411	1.013	1.017
1998	I	11276.0	13361.8	0.844	11275.299	1.000	1.020
-	II	13531.0	13580.2	0.971	13192.947	1.026	

12.18. Sau đây là bảng theo dõi giá trung bình (Y) của một loại cổ phiếu từ tháng 1 năm 2008 đến tháng 12 năm 2010:

Stt	Năm	Tháng	Y	Stt	Năm	Tháng	Y	Stt	Năm	Tháng	Y
1	2008	1	55.00	13	2009	1	30.00	25	2010	1	27.00
2	.	2	57.00	14	.	2	30.00	26	.	2	29.00
3	.	3	54.00	15	.	3	29.00	27	.	3	32.00
4	.	4	53.00	16	.	4	33.00	28	.	4	31.00
5	.	5	45.00	17	.	5	32.00	29	.	5	34.00
6	.	6	44.00	18	.	6	29.00	30	.	6	36.00
7	.	7	36.00	19	.	7	28.00	31	.	7	35.00
8	.	8	35.00	20	.	8	25.00	32	.	8	33.00
9	.	9	34.00	21	.	9	22.00	33	.	9	38.00
10	.	10	31.00	22	.	10	24.00	34	.	10	39.00
11	.	11	32.00	23	.	11	27.00	35	.	11	36.00
12	.	12	33.00	24	.	12	28.00	36	.	12	35.00

Sử dụng bảng Excel để tính toán phân tích chuỗi thời gian trên và dự báo giá cổ phiếu trung bình cho 3 tháng quý 1 năm 2011.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ĐS: $\hat{Y}_t = \hat{\alpha}t + \hat{\beta} = -0{,}4101673.t + 42{,}3381$

Stt	Năm	Tháng	Y_t	$\hat{Y}_t$	MA(4)	CMA	RMA
1	2008	1	55.00	41.93			
2	.	2	57.00	41.52			
3	.	3	54.00	41.11			
4	.	4	53.00	40.70			
5	.	5	45.00	40.29			
6	.	6	44.00	39.88	42.417		
7	.	7	36.00	39.47	40.333	41.375	0.870
8	.	8	35.00	39.06	38.083	39.208	0.893
9	.	9	34.00	38.65	36.000	37.042	0.918
10	.	10	31.00	38.24	34.333	35.167	0.882
11	.	11	32.00	37.83	33.250	33.792	0.947
12	.	12	33.00	37.42	32.000	32.625	1.011
13	2009	1	30.00	37.01	31.333	31.667	0.947
14	.	2	30.00	36.60	30.500	30.917	0.970
15	.	3	29.00	36.19	29.500	30.000	0.967
16	.	4	33.00	35.78	28.917	29.208	1.130
17	.	5	32.00	35.37	28.500	28.708	1.115
18	.	6	29.00	34.96	28.083	28.292	1.025
19	.	7	28.00	34.54	27.833	27.958	1.001
20	.	8	25.00	34.13	27.750	27.792	0.900
21	.	9	22.00	33.72	28.000	27.875	0.789
22	.	10	24.00	33.31	27.833	27.917	0.860
23	.	11	27.00	32.90	28.000	27.917	0.967
24	.	12	28.00	32.49	28.583	28.292	0.990
25	2010	1	27.00	32.08	29.167	28.875	0.935
26	.	2	29.00	31.67	29.833	29.500	0.983
27	.	3	32.00	31.26	31.167	30.500	1.049
28	.	4	31.00	30.85	32.417	31.792	0.975
29	.	5	34.00	30.44	33.167	32.792	1.037
30	.	6	36.00	30.03	33.750	33.458	1.076
31	.	7	35.00	29.62			
32	.	8	33.00	29.21			
33	.	9	38.00	28.80			
34	.	10	39.00	28.39			
35	.	11	36.00	27.98			
36	.	12	35.00	27.57			

Tháng	2008	2009	2010	TbRMA	Is
1		0.947	0.935	0.941	0.972
2		0.970	0.983	0.977	1.009
3		0.967	1.049	1.008	1.041
4		1.130	0.975	1.052	1.087
5		1.115	1.037	1.076	1.111
6		1.025	1.076	1.051	1.085

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tháng	2008	2009	2010	TbRMA	Is
7	0.870	1.001		0.936	0.967
8	0.893	0.900		0.896	0.926
9	0.918	0.789		0.854	0.882
10	0.882	0.860		0.871	0.899
11	0.947	0.967		0.957	0.989
12	1.011	0.990		1.001	1.033
				11.618	12.000

Năm	Tháng	Y_t	$\hat{Y}_t$	I_S	$\widetilde{Y}_t = \hat{Y}.I_S$	$(Y_t / \widetilde{Y}_t)$	$\hat{C}_t$ =MA(12)
2008	1	55	41.928	0.972	40.760	1.349	
.	2	57	41.518	1.009	41.883	1.361	
.	3	54	41.108	1.041	42.795	1.262	
.	4	53	40.697	1.087	44.240	1.198	
.	5	45	40.287	1.111	44.763	1.005	
.	6	44	39.877	1.085	43.267	1.017	
.	7	36	39.467	0.967	38.146	0.944	1.059
.	8	35	39.057	0.926	36.149	0.968	1.016
.	9	34	38.647	0.882	34.071	0.998	0.971
.	10	31	38.236	0.899	34.383	0.902	0.930
.	11	32	37.826	0.989	37.392	0.856	0.901
.	12	33	37.416	1.033	38.668	0.853	0.885
2009	1	30	37.006	0.972	35.975	0.834	0.864
.	2	30	36.596	1.009	36.917	0.813	0.855
.	3	29	36.186	1.041	37.671	0.770	0.840
.	4	33	35.775	1.087	38.889	0.849	0.819
.	5	32	35.365	1.111	39.294	0.814	0.810
.	6	29	34.955	1.085	37.927	0.765	0.808
.	7	28	34.545	0.967	33.389	0.839	0.807
.	8	25	34.135	0.926	31.593	0.791	0.809
.	9	22	33.725	0.882	29.732	0.740	0.817
.	10	24	33.314	0.899	29.957	0.801	0.835
.	11	27	32.904	0.989	32.526	0.830	0.841
.	12	28	32.494	1.033	33.582	0.834	0.857
2010	1	27	32.084	0.972	31.190	0.866	0.885
.	2	29	31.674	1.009	31.952	0.908	0.917
.	3	32	31.264	1.041	32.547	0.983	0.953
.	4	31	30.853	1.087	33.539	0.924	1.016
.	5	34	30.443	1.111	33.825	1.005	1.077
.	6	36	30.033	1.085	32.586	1.105	1.116
.	7	35	29.623	0.967	28.632	1.222	1.149
.	8	33	29.213	0.926	27.038	1.221	
.	9	38	28.803	0.882	25.392	1.497	
.	10	39	28.392	0.899	25.531	1.528	
.	11	36	27.982	0.989	27.661	1.301	
.	12	35	27.572	1.033	28.495	1.228	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Nguyễn Cao Văn, Trần Thái Ninh. *Giáo trình lý thuyết xác suất và thống kê toán*. NXB Khoa học và Kỹ thuật. Hà Nội.1999.
2. Nguyễn Cao Văn, Bùi Dương Hải. *Hướng dẫn trả lời lý thuyết và giải bài tập Kinh tế lượng*. NXB Tài chính. Hà Nội. 2011.
3. Nguyễn Quang Dong: *Kinh tế lượng nâng cao*. NXB KHKT. 2002.
4. Bencrézi J.P. Analyse des données. Dunod. Paris. 1973
5. Michel Volle: Analyse des donnees. Economica. 1994.
6. Amir D Aczel. Complete Business Statistics. Irwin Boston 1993.
7. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992
8. OTT L.R. And introduction to statistical methods and data analysis. California. 1993
9. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons. New york. 1990
10. Paul Newbold: Statistics for Business & Economics (Fourth Edition). Pretice-Hall International, Inc.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phần thứ ba
PHÂN TÍCH NHÂN TỐ

Phần này sẽ giới thiệu một nhóm phương pháp phân tích thống kê nhiều chiều qua các mô hình phân tích thống kê hiện đại. Hàm lượng toán học trong các phương pháp phân tích nhân tố tương đối lớn, cũng như vậy mặc dù việc ứng dụng các phương pháp này trong thực tế rất hiệu quả nhưng chúng cũng đòi hỏi những kỹ năng xử lý dữ liệu cao với sự trợ giúp của các phần mềm thống kê chuyên dụng.

Phân tích nhân tố là một nhóm phương pháp phân tích thống kê với số liệu lớn và không đồng nhất về nhiều thuộc tính. Phân tích nhân tố có những đặc điểm tốt hơn các phân tích thống kê nhiều chiều được trình bày ở các chương trước. Thứ nhất là: Phân tích nhân tố không đòi hỏi sự đồng nhất về độ đo của các biến; Thứ hai là: Phân tích nhân tố kết hợp được cả phân tích các cá thể và phân tích các biến trong cùng một mô hình.

Phân tích nhân tố bao gồm nhiều phương pháp, hiện nay các phương pháp chính được sử dụng là: Phương pháp thành phần chính – nhân tố chính; Phương pháp phân tích tương ứng; Phương pháp phân lớp tự động; Phương pháp phân tích khác biệt. Lựa chọn một số nội dung giới thiệu trong phần này, các tác giả hy vọng người sử dụng cuốn sách này có được một số tiếp cận mới về phân tích thống kê và có thể là phần mở đầu cho nhóm các mô hình phân tích nhân tố hiện nay đang được phát triển ngày càng phong phú hơn. Mặc dù hạn hẹp trong 3 chương nhưng các nội dung lựa chọn cũng có thể mang lại cho người sử dụng nhận biết cách sử dụng các phương pháp này.

Kỹ năng sử dụng phần mềm chuyên dụng cho các nội dung tương ứng được giới thiệu trên SPSS, một phần mềm chuyên dụng đang ngày càng được nhiều người quan tâm.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Một số mô hình thực nghiệm cũng được trình bày không chỉ có tính mô tả phương pháp, mà còn tập trung gắn với các bài toán thực tế và sử dụng dữ liệu thực tế của Việt Nam.

Để có thể sử dụng được các kiến thức trong phần này đòi hỏi người học phải có các kiến thức đại số tuyến tính, giải tích và xác suất thống kê tương đối đầy đủ. Ngoài ra người học cần làm quen với việc khai thác các phần mềm chuyên dụng của thống kê, toán học và các bảng tính thông thường như Excel, Lotus hay một số Macro thống kê viết trên Excel như: Megastat, NCSS, ...

Vì thiết kế các bài tập trong phần này rất phức tạp, chúng tôi chỉ đưa ra các bài tập đơn giản có tính chất mô tả phương pháp và hướng dẫn người học kỹ năng sử dụng phần mềm. Các mô hình phân tích thực nghiệm được trình bày thông qua các thí dụ ở các chương. Những nội dung lý thuyết đầy đủ hơn xin xem ở phần tài liệu tham khảo.

Chương 13 sẽ nghiên cứu một số nội dung cơ bản của phương pháp phân tích thành phần chính.

Chương 14 sẽ giới thiệu một số nội dung của phương pháp phân tích khác biệt.

Chương 15 sẽ giới thiệu các phương pháp phân lớp.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 13
PHƯƠNG PHÁP THÀNH PHẦN CHÍNH

1. MỞ ĐẦU

Đây là chương mở đầu cho 3 chương phân tích dữ liệu với cùng một cơ sở là phân tích nhân tố và các ứng dụng. Trong chương này mô hình phân tích nhân tố cũng được thực hiện qua một phương pháp cụ thể là phương pháp phân tích thành phần chính.

Bài toán phân tích nhân tố được giới thiệu với cấu trúc của các phép chiếu trong không gian tuyến tính. Các khái niệm cơ bản của bài toán phân tích nhân tố và các kiến thức cơ sở giải lớp bài toán này được mô tả dưới ngôn ngữ ma trận và các phép biến đổi trong không gian tuyến tính với các đọ đo phù hợp.

Mô hình tổng quát của bài toán phân tích nhân tố được xây dựng cùng với thuật giải và lời giải, tạo cơ sở cho các ứng dụng trong chương này và chương tiếp theo.

Phương pháp thành phần chính sẽ được giới thiệu đầy đủ cả về lý thuyết và thực hành. Phân tích kết quả của bài toán phân tích nhân tố bằng phương pháp thành phần chính là nội dung lớn thứ hai của chương này. Các kiểm định cần thiết cũng được giới thiệu như cách thức nhận biết tính phù hợp của bộ số liệu với mô hình thống kê tương ứng.

Tiếp cận phân tích thành phần chính trong chương này kết hợp các phép chiếu trong không gian tuyến tính với các kỹ thuật hồi qui để dễ dàng nhận được cả các thành phần chính và các nhân tố chính tương ứng. Với các thí dụ thực tế các ứng dụng phân tích thành phần chính như một phép rút gọn thông tin được nhấn mạnh trong các kết quả.

Với nội dung chương này, bạn đọc có thể nắm được cách tiếp cận chung của nhóm mô hình phân tích nhân tố. Đối với phân tích thành phần chính, bạn đọc sẽ làm quen với các nội dung sau:

- Cách sử dụng ma trận độ đo;

- Các giá trị riêng và tỷ lệ giải thích trên siêu phẳng chiếu

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Định danh các thành phần chính
- Các hệ số đo sự đóng góp của các biến
- Các kỹ thuật kiểm định và hiệu chỉnh mô hình.
- Thực hiện được kỹ thuật hồi qui thành phần chính và cách tính tác động của các yếu tố ngoại sinh.
- Sử dụng thành thạo các thủ tục có liên quan trên SPSS

2. BÀI TOÁN PHÂN TÍCH NHÂN TỐ

Trong khi nghiên cứu tổng thể bao gồm các cá thể mang nhiều đặc trưng khác nhau, người ta có thể quan tâm đến hai cách thường gọi là phân tích dọc và phân tích ngang. Một mặt phân tích dọc có mục đích chỉ ra các đặc tính hay dấu hiệu khác nhau tồn tại trong tổng thể có quan hệ với nhau như thế nào, trong khi phân tích ngang lại đặt chính các cá thể với những dấu hiệu nói trên trong mối quan hệ chung cũng như quan hệ của từng cặp hay từng nhóm cá thể. Hãy xét một thí dụ cụ thể sau đây: Các doanh nghiệp vừa và nhỏ ở thành phố A thường được xem xét như các cá thể của một tổng thể, để có thể mô tả một doanh nghiệp người ta dùng 4 chỉ tiêu:

- X(1): Giá trị tài sản cố định
- X(2): Lượng lao động thường xuyên được sử dụng
- X(3): Tỷ lệ lợi nhuận thô
- X(4): Lương bình quân của người lao động/tháng.

Rõ ràng là hệ thống chỉ tiêu này có thể dùng so sánh các doanh nghiệp và có thể nó cho biết các quan hệ của chính các chỉ tiêu này trong phạm vi đang xét chẳng hạn lượng lao động và giá trị tài sản cố định quan hệ chặt chẽ với nhau và cùng chiều nhưng có thể giá trị tài sản cố định và lương bình quân quan hệ cũng rất chặt chẽ nhưng lại ngược chiều... Điều đó có thể đúng với các doanh nghiệp vừa và nhỏ ở thành phố A mà không đúng với các doanh nghiệp vừa và nhỏ ở thành phố B chẳng hạn.

Về phía các doanh nghiệp, với những chỉ báo cụ thể nói trên người ta có thể và rất muốn định vị các doanh nghiệp vừa và nhỏ của thành phố A từ đó thấy sự khác biệt của mỗi doanh nghiệp với doanh nghiệp khác.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trong chừng mực nào đó toàn bộ phân tích dọc - ngang như vậy có thể coi như một cặp phân tích đối ngẫu (trong các phần sau ta sẽ chỉ ra điều đó). Để làm được các phân tích trên, người ta có thể dùng phân tích tương quan, tuy nhiên phân tích tương quan nhiều chiều hầu như không giải quyết được bằng các công cụ thông thường. Giải quyết bài toán trên ở cả phân tích dọc và ngang dựa trên tưởng tượng cơ bản sau đây:

Mỗi cá thể coi như một phần tử của tổng thể (P), số phần tử (M) hữu hạn hay vô hạn, mỗi chỉ tiêu coi như một biến, đo một mặt hay một đặc trưng của cá thể.

Mô tả mỗi cá thể bởi một véc tơ p chiều mà mỗi thành phần là giá trị của một biến.

Phân tích dọc - ngang: Tìm cách tổ hợp các biến, hay tổ hợp p véc tơ M chiều thành các véc tơ khác nhau đôi một trực giao. Mỗi véc tơ này được coi là một thành phần hay một biến mới đo lường các cá thể trên một mặt nào đó. Với số tổ hợp xác định trước, chúng ta thu được một hệ thống chỉ báo, nói chung không có thực nhưng với hệ thống này mỗi chỉ tiêu ban đầu, cũng là mỗi biến ban đầu có phần đóng góp của mình tạo nên các chỉ báo đó, theo một tiêu chuẩn nào đó.

Chẳng hạn với p tiêu thức (biến) ban đầu ta lập:

$$Y_k = u_{k1}X_1 + u_{k2}X_2 + \quad ... + u_{kp}X_p$$

Mỗi bộ số (véc tơ) u_k xác định một biến Y_k và tương ứng với Y_k ta có các giá trị tương ứng các cá thể.

Mỗi cá thể có thể thể hiện bởi một véc tơ mới mà mỗi thành phần của véc tơ này là hệ số phân tích qua một biến mới (gọi là một thành phần chính). Nhờ đó có được mô tả về quan hệ của chúng.

3. CÁC Ý TƯỞNG CƠ BẢN

3.1. Một thí dụ không chuẩn mực

Thí dụ không chuẩn mực ở đây hiểu theo nghĩa là các thành phần chính sẽ không hình thành theo một tiêu chuẩn khách quan mà hình thành theo ý tưởng chủ quan của người thiết kế (không phải người phân tích).

Sau đây là số liệu của 13 nước với các chỉ tiêu P tổng số dân (1000 người), F là mật độ dân số (người/ km^2) và U là tỷ lệ dân thành thị.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

	P	F	U
Afghanistan	20500	25.0	18
Argentina	33900	12.0	86
Armenia	3700	126.0	68
Australia	17800	2.3	85
Austria	8000	94.0	58
Azerbaijan	7400	86.0	54
Bahrain	600	828.0	83
Bangladesh	125000	800.0	16
Barbados	256	605.0	45
Belarus	10300	50.0	65
Belgium	10100	329.0	96
Bolivia	7900	6.9	51

Có hai quan điểm nhân khẩu học muốn áp đặt để phân tích các tiêu thức cũng như các quốc gia này. Các trọng số được ấn định như sau:

	P	F	U
Quan điểm 1:	-0,15	0,05	0,25
Quan điểm 2:	-0,2	-0,1	0,40

Các trọng số này không xuất phát từ chính số liệu mà xuất phát từ quan niệm (sau này ta sẽ mô tả cách xác định chúng một cách khách quan theo những tiêu chuẩn nhất định).

Mỗi quốc gia được mô tả bởi 1 véc tơ 2 chiều S = [S(1), S(2)] với

S(1) = -0,15P + 0,05F + 0,25U

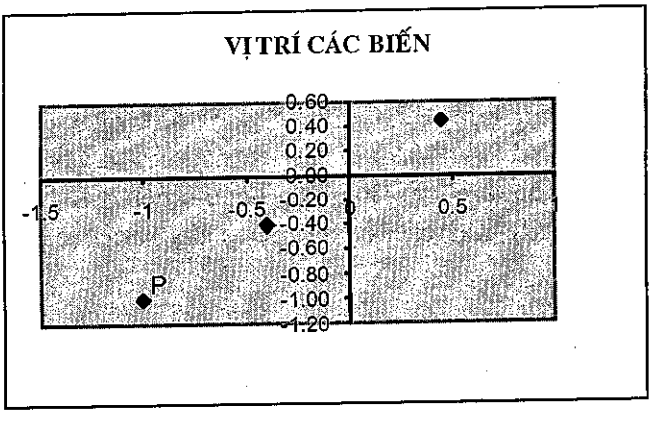
S(2) = -0,20P -0,1F + 0,40U

	P	F	U	S1	S2
Afghanistan	20500	25	18	-3069.25	-4095.3
Argentina	33900	12	86	-5062.9	-6746.8
Armenia	3700	126	68	-531.7	-725.4
Australia	17800	2.3	85	-2648.64	-3526.23
Austria	8000	94	58	-1180.8	-1586.2
Azerbaijan	7400	86	54	-1092.2	-1467
Bahrain	600	828	83	-27.85	-169.6

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

	P	F	U	S1	S2
Bangladesh	125000	800	16	-18706	-25073.6
Barbados	256	605	45	3.1	-93.7
Belarus	10300	50	65	-1526.25	-2039
Belgium	10100	329	96	-1474.55	-2014.5
Bolivia	7900	6.9	51	-1171.91	-1560.29

Trong không gian hai chiều (S1,S2) ta có thể mô tả các biến. Mỗi biến tương ứng là 1 véc tơ n chiều: P,F,U, người ta tính các hệ số tương quan (r_{PS1}, r_{PS2}); (r_{FS1}, r_{FS2}); (r_{US1}, r_{PU2}), mỗi cặp hệ số này xác định 1 điểm trong siêu phẳng tạo bởi S1 và S2. Với thí dụ này, ta có vị trí của các biến P, F, U như sau:



Với kết quả này, ta giải thích ý nghĩa các véc tơ S1, S2 dễ dàng hơn. Có thể xem S1 là tỷ lệ dân cư thành thị và S2 là mật độ dân số. Sau này chúng ta sẽ làm rõ cách phân tích khi đã có đủ cơ sở lý thuyết, ở thí dụ không chuẩn mực này chúng ta không có được nhận xét gì. Mục đích của thí dụ chỉ là cung cấp một cách tiếp cận mà thôi.

Như vậy rất rõ ràng là tỷ lệ dân số thành thị, dù có quan niệm khác nhau cũng khác biệt lớn với số dân và mật độ dân cư. Hay các nước đông dân tỷ lệ đô thị hoá thấp và ngược lại (tuy nhiên đây không phải là một phân tích thực sự).

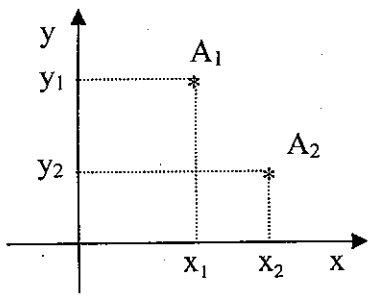
Kỹ thuật phân tích thành phần chính là một bộ phận của phân tích nhân tố (Factoriel Analyse). Phân tích nhân tố nói chung và phân tích thành phần chính nói riêng, được sử dụng cho phân tích những bộ số liệu lớn. Về mặt

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

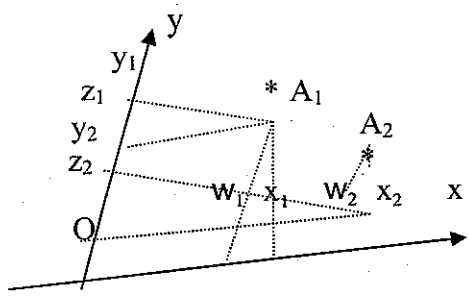
công cụ, người ta cần sử dụng các phép biến đổi trong không gian tuyến tính. Đặc biệt với mục đích phân tích dọc - ngang, véc tơ riêng và giá trị đặc trưng trở thành yếu tố không thể thiếu trong việc xác định các thành phần chính, các nhân tố chính cũng như các thành phần khác trong nhóm phương pháp này. Cũng chính vì lý do này, người ta chỉ có thể thực hiện các mô hình phân tích nhân tố thực sự có ý nghĩa khi có sự hỗ trợ của các phần mềm thống kê chuyên dụng. Các phân tích kinh tế xã hội mà bắt đầu từ các phân tích nhân khẩu học kinh tế và nhân khẩu học xã hội, đã trở thành các mô hình mẫu mực có tính chất giáo khoa trong lĩnh vực này. Các thí dụ sẽ được sử dụng cũng chủ yếu nhận được từ các tệp số liệu như vậy. Về mặt kỹ thuật tính toán trong đại đa số các trường hợp, phân tích nhân tố như một Modul riêng có và hoàn chỉnh của hai phần mềm thống kê chính hiện nay là SAS (Systems programe for data analysis) và SPSS (Systems package of analyis for data socio - economis).

3.2. Bài toán phân tích nhân tố trong không gian 2 và 3 chiều

Hãy xem xét trường hợp có hai biến X và Y được quan sát trên n cá thể. Phương pháp thống kê đã cung cấp lý thuyết mẫu, lý thuyết phân tích tương quan để xem xét mối quan hệ của hai biến ngẫu nhiên. Mỗi cá thể A_i có thể mô tả bởi 1 điểm trong R^2 nhờ các toạ độ (x_i, y_i) hãy xem các biểu đồ:



Hình 1a



Hình 1b

Hình 1a cho thấy sự khác nhau của A_1, A_2 có thể đặc trưng bằng khoảng cách A_1A_2 tuy nhiên khi chiếu vuông góc lên x (không quan tâm đến y) thì khoảng cách này gần hơn khi chiếu vuông góc lên y và cả hai khoảng cách sau phép chiếu đều nhỏ hơn A_1A_2, nói chính xác hơn là $d^2(A_1A_2) = d^2(x_1, x_2) + d^2(y_1, y_2)$. Nếu không muốn mất thông tin thì tốt nhất là chọn một đường thẳng song song với đường thẳng đi qua A_1, A_2. Khi chiếu vuông góc hai điểm này lên đường thẳng như vậy, khoảng cách

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ảnh bằng khoảng cách ban đầu.

Hình 1b cũng cho kết luận tương tự nhưng có thể chứng tỏ rằng theo các toạ độ ban đầu khi Ox không vuông góc với Oy thì:

$$d^2(A_1A_2) = d^2(x_1, x_2) + d^2(y_1, y_2)$$

$$= d^2(u_1, u_2) + d^2(w_1, w_2) - 2(u_1 - u_2)(w_2 - w_1)\text{Cosin}(xOy).$$

Như vậy nếu muốn phản ánh gần nhất liên hệ của hai cá thể A_1 và A_2 cần phải chọn một trục sao cho khoảng cách của các hình chiếu xấp xỉ tốt nhất khoảng cách ban đầu. Với n >2 cá thể ý tưởng trên xuất phát từ việc bảo toàn tối đa sự khác biệt của các cá thể khi xem xét trên cả p tiêu thức.

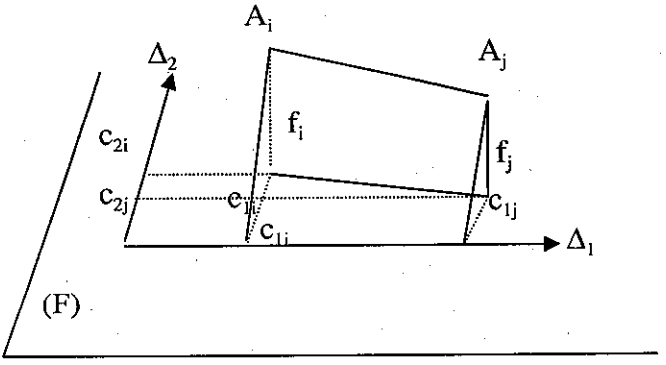
Để có thể tưởng tượng rõ ràng hơn cách tiếp cận này, chúng ta xét trường hợp các cá thể trong R^3. Giả sử mỗi cá thể được đặc trưng bởi 3 tiêu thức X_1, X_2, X_3; Xét hai cá thể A_i, A_j được mô tả bởi hai điểm trong R^3. Gọi Δ_1 và Δ_2 là hai đường thẳng trực giao tạo nên mặt phẳng (F).

Khoảng cách của ảnh (A_i,A_j) sau phép chiếu là khoảng cách (f_i, f_j) khoảng cách này có thể xác định như sau:

$$d^2(f_i,f_j) = d^2(c_{1i},c_{1j}) + d^2(c_{2i},c_{2j}) <= d^2(A_i,A_j)$$

Như vậy nếu chọn một trục để chiếu ta sẽ chọn trục nào có khoảng cách hai toạ độ chiếu lớn hơn.

Có thể mô tả phép chiếu như sau:



Trường hợp có n điểm cần chọn các trục Δ_1, Δ_2 sao cho trung bình của tổng bình phương các khoảng cách giữa các điểm f_i (ảnh của A_i) lớn nhất.

3.3. Bài toán trong không gian p chiều

Xét bài toán nói trên trong trường hợp mỗi cá thể có thể xem xét qua p

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

chỉ tiêu. Một không gian p chiều gồm các trục: Δ_1, Δ_2,... Δ_p đôi một trực giao có thể được hình thành nhờ việc chọn lần lượt các trục Δ_1, Δ_2,... theo thứ tự giảm dần của tổng bình phương các khoảng cách của các hình chiếu. Việc lựa chọn số chiều không gian chiếu phụ thuộc vào nhiều yếu tố khác nhau, như mức phân tán trong các mối quan hệ của p tiêu thức ban đầu, yêu cầu về bảo toàn thông tin,... . Tuy vậy do ý muốn quan sát trực quan, người ta mong muốn có thể bảo toàn thông tin ở mức chấp nhận được với số chiều không gian chiếu nhỏ nhất. Người ta luôn hướng đến không gian chiếu 3 hay 3 chiều, vì trong các không gian này việc mô tả hình học thuận tiện đối với người sử dụng.

Trong phân tích nhân tố người ta sử dụng các khái niệm sau:

Mỗi véc tơ Δ_i gọi là một trục chính của đám mây số liệu ban đầu.

Mỗi điểm A_i chiếu lên p trục chính sẽ có p giá trị tương ứng, các giá trị này lập nên véc tơ C_i, p chiều.

Với n có thể (n điểm) A_1, A_2, ..., A_n ta có ma trận $C=(c_{ij})$. Mỗi cột của ma trận này là một véc tơ trong R^n, gọi là một thành phần chính.

Mỗi thành phần chính C_k có thể biểu diễn qua các cột tương ứng với các tiêu thức ban đầu (X_j) dưới dạng:

$$C_k = u_{k1}X_1 + ... + u_{kp}X_p.$$

Véc tơ các hệ số tổ hợp nói trên (U_k) gọi là nhân tố chính thứ k.

Trường hợp bản thân các tiêu thức ban đầu hoàn toàn độc lập, phân tích này cũng cho phép lựa chọn một không gian có số chiều ít hơn mà liên hệ của các cá thể sai lệch ít nhất. Trong trường hợp các tiêu thức ban đầu không độc lập kết quả tốt hơn rất nhiều, đó là tìm được cách diễn đạt mỗi cá thể qua các tiêu thức độc lập trong toàn bộ.

Khó khăn ở đây chính là ban đầu mỗi cá thể được thể hiện qua p tiêu thức quan sát được nhưng kết quả mỗi cá thể có thể biểu hiện qua q tiêu thức không qua sát được. Việc xác định ý nghĩa kinh tế, xã hội của các tiêu thức này luôn là một khó khăn đối với người thực hành.

Các phân tích tương ứng, phân tích tương quan chính tắc, phân nhóm... cũng được xây dựng trên ý tưởng cơ bản này. Tuy nhiên cách thức và điều kiện áp dụng tuỳ thuộc vào tình trạng dữ liệu. Những vấn đề cơ bản để thực hiện nhóm các phân tích này sẽ được trình bày ở chương này.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4. MÔ TẢ, TÓM TẮT SỐ LIỆU

Trước tiên cần mô tả số liệu cho các phân tích thống kê nhiều chiều nói chung và mô tả số liệu cho phân tích nhân tố nói riêng. Trong các mô tả này ta sẽ tìm cách tóm tắt số liệu cho mục đích riêng có của phân tích nhân tố. Các mô tả này được trình bày dưới ngôn ngữ véc tơ và ma trận trong các không gian tuyến tính thực.

4.1. Số liệu và các đặc trưng

1. Bảng số liệu

Giả sử có n quan sát (n cá thể lập nên 1 mẫu), mỗi quan sát có p tiêu thức (biến). Số liệu có thể trình bày trên bảng sau:

$$X=\begin{bmatrix} x_{11} & x_{12} & \ldots\ldots & x_{1p} \\ x_{21} & x_{22} & \ldots\ldots & x_{2p} \\ \ldots & \ldots & \ldots\ldots & \ldots \\ x_{i1} & x_{i2} & \ldots\ldots & x_{ip} \\ \ldots & \ldots & \ldots\ldots & \ldots \\ x_{n1} & x_{m2} & \ldots\ldots & x_{np} \end{bmatrix}$$

Gọi: - X^*_i là dòng thứ i của X với các phần tử: x_{i1}, x_{i2} . ... , x_{ip} đó là véc tơ các giá trị của các biến quan sát được ở cá thể thứ i (trong trường hợp dùng chỉ số qui ước i ta có thể bỏ dấu * ở trên). Một dòng của X còn có thể gọi một cách ngắn gọn là một cá thể.

- X_j là cột thứ j của X nó bao gồm các giá trị của 1 biến tại tất cả các cá thể. Một cột của X có thể gọi là một biến hay một tiêu thức.

Nói cách khác mỗi cá thể biểu hiện bằng 1 véc tơ dòng p chiều, mỗi tiêu thức thể hiện qua n cá thể bởi một véc tơ cột n chiều.

2. Ma trận trọng số

Nếu như mỗi dòng của X tương ứng duy nhất 1 cá thể thì trọng số của các cá thể như nhau và có thể lấy bằng 1/n. Tuy nhiên, thực tế có thể 1 dòng của X ứng với một số các thể nào đó, như vậy để thể hiện đúng vai trò của đám đông trong phân tích thống kê ta có thể đặt cho mỗi dòng (mỗi cá thể đại diện) một trọng số p_i. Các trọng số như vậy có thể mô tả bởi một ma trận gọi là ma trận trọng số như sau:

$$\begin{bmatrix} p_1 & 0 & \ldots\ldots & 0 \end{bmatrix}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$D = \begin{bmatrix} 0 & p_2 & ..... & 0 \\ 0 & ... & ..... & ... \\ ... & ... & ..... & ... \\ 0 & 0 & ..... & 0 \\ 0 & ... & ..... & ... \\ 0 & 0 & ..... & p_n \end{bmatrix}$$

Các trọng số này thông thường là các tần suất dòng trong bảng số liệu thô. Mỗi dòng chỉ mang thông tin của 1 cá thể thì có thể viết $D = E.\frac{1}{n}$, trong đó E là ma trận đơn vị cấp n.

3. Điểm trung bình hay trung tâm của đám mây số liệu và ma trận trung tâm hoá

Trước tiên ta qui định cách viết mặc định cho mỗi véc tơ là véc tơ cột và chuyển vị của một véc tơ a sẽ ký hiệu là a^T là một véc tơ dòng. Với các phân tích sau này sự phân biệt cách viết này là cần thiết.

Mỗi dòng của X có thể xem là 1 điểm trong không gian R^p, n điểm tạo nên một tập hợp gọi là một ***đám mây số liệu***, ta gọi véc tơ trung tâm của X hay ***điểm trung tâm*** của đám mây này là g thì g có thể tính như sau:

$$g = X^T DI \qquad (13.1)$$

trong đó I là véc tơ có tất cả các thành phần bằng 1 trong không gian R^n. Dễ dàng kiểm tra thấy g là một véc tơ p chiều.

Chẳng hạn, cho ma trận X ở bảng sau:

X_1 = Tuổi thọ trung bình

X_2 = Tỷ lệ tăng dân số

X_3 = Tỷ lệ chết trẻ em

X_4 = GDP/đầu dân cư

i	X_1	X_2	X_3	X_4
1	44	2.8	168.0	205
2	75	1.3	25.6	3408
3	75	1.4	27.0	5000
4	80	1.4	7.3	16848

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5	79	.2	6.7	18396
6	75	1.4	35.0	3000
7	74	2.4	25.0	7875
8	53	2.4	106.0	202
9	78	.2	20.3	6950
10	76	.3	19.0	6500
Tổng	**709**	**13.8**	**439.9**	**68384**

Giả sử các số liệu này thống kê ở 10 nước khác nhau. Ta chọn các trọng số đều là 0,1; véc tơ trung tâm của đám mây số liệu có các thành phần là trung bình của các thành phần quan sát. Véc tơ chuyển vị của véc tơ trung tâm là:

$$g^T = (70.9 \quad 1.38 \quad 43.99 \quad 6838.4)$$

Thực tế là các thành phần này có thể tính nhờ các thông tin ở dòng tổng của bảng trên.

Việc trung tâm hoá đám mây số liệu thực hiện bằng cách lập các véc tơ chênh lệch của các cá thể với véc tơ g.

Đặt ma trận: $Y = X - Ig^T$ g^T là chuyển vị của g. (13.2)

Tích Ig^T là một ma trận có các dòng là đều là véc tơ g^T.

Trong thí dụ trên tích Ig^T là một ma trận có n dòng (n=10) như nhau:

$$\begin{bmatrix} 70.9 & 1.38 & 43.99 & 6838.4 \\ 70.9 & 1.38 & 43.99 & 6838.4 \\ 70.9 & 1.38 & 43.99 & 6838.4 \\ 70.9 & 1.38 & 43.99 & 6838.4 \\ 70.9 & 1.38 & 43.99 & 6838.4 \\ 70.9 & 1.38 & 43.99 & 6838.4 \\ 70.9 & 1.38 & 43.99 & 6838.4 \\ 70.9 & 1.38 & 43.99 & 6838.4 \\ 70.9 & 1.38 & 43.99 & 6838.4 \\ 70.9 & 1.38 & 43.99 & 6838.4 \end{bmatrix}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Từ đó, thực hiện phép trừ ta có ma trận Y

$$Y=\begin{bmatrix} -26.9 & 1.42 & 124.01 & -6633.4 \\ 4.1 & -0.08 & -18.39 & -3430.4 \\ 4.1 & 0.02 & -16.99 & -1838.4 \\ 9.1 & 0.02 & -36.69 & 10009.6 \\ 8.1 & -1.18 & -37.29 & 11557.6 \\ 4.1 & 0.02 & -8.99 & -3838.4 \\ 3.1 & 1.02 & -18.99 & 1036.6 \\ -17.9 & 1.02 & 62.01 & -6636.4 \\ 7.1 & -1.18 & -23.69 & 111.6 \\ 5.1 & -1.08 & -24.99 & -338.4 \end{bmatrix}$$

Ma trận này mô tả thô sự khác biệt hay sự phân tán của các cá thể so với mức trung bình.

4. Ma trận hiệp phương sai và ma trận hệ số tương quan

Đặc trưng quan trọng nói lên mức phân tán của mỗi biến và độ liên hệ giữa chúng là ma trận hệ số tương quan và ma trận hiệp phương sai. Các ma trận này có thể tính như sau.

+ Ma trận hiệp phương sai

Với các trọng số 1/n ta có thể tính ma trận hiệp phương sai (V) theo công thức:

$$V = X^T D X - gg^T. \qquad (13.3)$$

Một cách tổng quát ta có các phần tử của X^TDX được tính như sau:

$$q_{ij} = \sum_{k=1}^{n} p_k x_{ki} x_{kj}$$

véc tơ trung tâm g cũng được tính có trọng số như sau: $g_j^T = \sum_{i=1}^{n} p_i x_{ij}$

Với thí dụ trên ta có ma trận hiệp phương sai V như sau:

$$V=\begin{bmatrix} 132.89 & -7.672 & -561.821 & 44687.44 \\ -7.672 & 0.8056 & 31.9138 & -2817.43 \\ -561.821 & 31.9138 & 2421.443 & -191741 \\ 44687.44 & -2817.43 & -191741 & 35289619 \end{bmatrix}$$

Như đã biết $Cov(X_i,X_i) = Var(X_i)$ nên từ ma trận này dễ dàng tìm được các phương sai của các biến (các cột của X).

Có thể chứng tỏ rằng: $\mathbf{V = X^TDX - gg^T = Y^TDY}$, đẳng thức này cho thấy một biến đổi tuyến tính của Y với ma trận D.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Ma trận hệ số tương quan

Gọi ma trận $D_{1/S}$ là ma trận đường chéo với các thành phần là $1/Se(X_j)=1/s_j$.

$$D_{1/S}=\begin{bmatrix} 1/s_1 & 0 & ..... & 0 \\ 0 & 1/s_2 & ..... & 0 \\ o & ... & ..... & .. \\ ..... & .... & ..... & ... \\ 0 & 0 & ..... & 0 \\ 0 & .... & ..... & ... \\ 0 & 0 & ..... & 1/s_p \end{bmatrix}$$

Chuẩn hoá ma trận X ta có ma trận $Z = (z_{ij})$

trong đó: $z_{ij} = \frac{y_{ij}}{s_j}$

Có thể nhận được ma trận này nhờ công thức:

$$Z= YD_{1/S} \qquad (13.4)$$

Nhờ các công thức trên ta có ma trận chuẩn hoá của ma trận X ở thí dụ trên là:

$$Z = \begin{bmatrix} -2.33349 & 1.582081 & 2.520111 & -1.11664 \\ 0.355662 & -0.08913 & -0.37372 & -0.57746 \\ 0.355662 & 0.022283 & -0.34527 & -0.30947 \\ 0.789397 & 0.022283 & -0.74561 & 1.684974 \\ 0.70265 & -1.31469 & -0.7578 & 1.945558 \\ 0.355662 & 0.022283 & -0.18269 & -0.64614 \\ 0.268915 & 1.136424 & -0.38591 & 0.174497 \\ -1.55277 & 1.136424 & 1.260157 & -1.11714 \\ 0.615903 & -1.31469 & -0.48142 & 0.018786 \\ 0.442409 & -1.20327 & -0.50784 & -0.05696 \end{bmatrix}$$

Ma trận hệ số tương quan tuyến tính của các biến (R) được tính như sau:

$$\mathbf{R} = \mathbf{D_{1/S}VD_{1/S}=Z^TDZ}. \qquad (13.5)$$

Và ma trận hệ số tương quan nhận được là:

$$R=\begin{bmatrix} 1.0000 & -0.7415 & -0.9904 & 0.6526 \\ -0.7415 & 1.0000 & 0.7226 & -0.5284 \\ -0.9904 & 0.7226 & 1.0000 & -0.6559 \\ 0.6526 & -0.5284 & -0.6559 & 1.0000 \end{bmatrix}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chú ý là Z có véc tơ trung tâm là véc tơ 0, nên R cũng chính là ma trận hiệp phương sai của Z.

4.2. Không gian các cá thể

Các cá thể của tổng thể được mô tả bằng các véc tơ với các thành phần là các giá trị của p biến như đã nêu ở trên. Ta sẽ coi mỗi cá thể như vậy là một điểm trong không gian F^p (với số chiều bằng p). Tập hợp một số các cá thể tạo nên một đám mây trong F với g là điểm trung tâm của đám mây này.

Để có thể phân tích thống kê đối với các đám mây trong F ta thống nhất một số khái niệm làm cơ sở cho phân tích sau này.

1. Khoảng cách

Khoảng cách trong không gian tuyến tính thông thường được sử dụng là khoảng cách Ơcơlit. Thông thường người ta định nghĩa khoảng cách giữa hai điểm X_1 và X_2 theo công thức Pithagore như sau:

$$d^2 = (x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + ... + (x_{1j} - x_{2j})^2 + ... + (x_{1p} - x_{2p})^2$$

Tuy nhiên, nếu như mỗi cột của X có một trọng số $a_j > 0$ thì khoảng cách trên có thể được tính bằng cách cho mỗi hạng tử ở trên một nhân tử tương ứng, tức là:

$$d^2 = a_1(x_{11} - x_{21})^2 + a_2(x_{12} - x_{22})^2 + ... + a_i(x_{1j} - x_{2j})^2 + ... + a_p(x_{1p} - x_{2p})^2.$$

$$\text{Hay } d^2 = (X_1-X_2)^T \begin{bmatrix} a_1 & 0 & .....0 \\ 0 & a_2 & .....0 \\ & ..... & \\ 0 & 0 & .....a_p \end{bmatrix} (X_1- X_2) \qquad (13.6)$$

2. Khoảng cách M (Metric M)

Một cách tổng quát có thể chọn một ma trận đường chéo dương M và khoảng cách hai điểm trong không gian được xác định theo công thức sau:

$$d^2 = (X_i - X_j)^T M (X_i - X_j)$$

trong đó M là một ma trận đường chéo dương cấp p.

Trong trường hợp này tích vô hướng của hai véc tơ cũng xác định tương tự:

$$< X_i , X_j> = X_i^T M X_j$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chuẩn của một véc tơ được tính theo công thức: $\| X\|_M = (X^T M X)^{1/2}$

Trong phân tích thống kê nếu chỉ quan tâm đến độ phân tán của các biến, thì M thường xác định qua ma trận hiệp phương sai. Để có một độ đo khoảng cách theo nghĩa mức tương ứng, trong phân tích thành phần chính ta có thể sử dụng ma trận M có đường chéo là nghịch đảo của các phương sai của các biến. Điều đó tương đương với việc chia tất cả các biến cho độ lệch tiêu chuẩn của nó. Cách làm này làm cho sự khác biệt về độ đo không còn nữa. Ma trận hiệp phương sai của các biến đã biến đổi chính là ma trận hệ số tương quan của các biến ban đầu.

Cách hiệu chỉnh này hàm ý rằng những biến ổn định (có phương sai nhỏ) sẽ đóng vai trò lớn hơn trong công thức xác định khoảng cách.

Với cách này ta có:

$$M = D_{1/S}^2 = \begin{bmatrix} 1/s^2_1 & 0 & ..... & 0 \\ 0 & 1/s^2_2 & ..... & 0 \\ o & ... & ..... & .. \\ .... & .... & ..... & ... \\ 0 & 0 & ..... & 0 \\ 0 & .... & ..... & ... \\ 0 & 0 & ..... & 1/s^2_p \end{bmatrix}$$

ma trận M được xác định như trên là một ma trận xác định dương nên nó luôn tồn tại biểu diễn $M = T^T T$ với T là ma trận vuông cấp p.

Thay biểu thức này vào công thức trên ta sẽ có:

Tích vô hướng: $< X_i , X_j> = X_i^T M X_j = X_i^T T^T T X_j = (TX_j)^T (TX_i)$.

Biểu thức này cho thấy mối liên hệ giữa các độ đo. Với độ đo M bất kỳ nhờ phép biến đổi trên ta có thể qui về độ đo M = E, đây chính là độ đo Ơcơlit thông thường.

3. Quán tính

Khái niệm tâm đám mây số liệu đã gợi ý một ý niệm hướng về tâm của mỗi điểm trong đám mây này.

Người ta gọi **tổng quán tính** của đám mây với tâm của nó là tổng các khoảng cách từ các điểm đến tâm của đám mây đó (đo bằng Metric- M) đại lượng này được tính như sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$I_g = \sum_{i=1}^{n} p_i (X_i - g)^T M (X_i - g) \qquad (13.7)$$

Nếu ta chọn một điểm a nào đó thay cho g ta sẽ tính được I_a theo công thức trên. Theo công thức Huyghens ta có: $I_a = I_g + (g-a)^T M (g-a) = I_g + \|g - a\|^2$, trong độ $\|g - a\|$ coi là chuẩn của véc tơ (g-a) theo Metric M.

Ngoài ra khi g=0 ta có $I_g = \sum_{i=1}^{n} p_i X_i^T M X_i$

Người ta còn chứng minh được rằng:

$$2I_g = \sum_{J=1}^{n} \sum_{i=1}^{n} p_i p_j \left\| X_i - X_j \right\|^2 \qquad (13.8)$$

Tức là 2 lần tổng quán tính bằng trung bình bình phương của các khoảng cách giữa các cá thể.

Theo công thức xác định ma trận hiệp phương sai, ta có thể tính tổng quán tính I_g qua ma trận V và M như sau:

I_g =TraceMV =TraceVM, trong đó TraceA (vết của A) là tổng các phần tử trên đường chéo của A.

Như vậy:

- Nếu M=E (E là ma trận đơn vị) thì tổng quán tính chính là tổng phương sai của các biến.

- Nếu $M = D_{1/S}^2$ thì tổng này sẽ là tổng các phần tử trên đường chéo ma trận R hay tổng quán tính bằng p (số biến số), mà không phụ thuộc vào các giá trị của các biến số đó.

Trở lại thí dụ trên nếu ta chọn khoảng cách Ơcơlit cho không gian các cá thể thì tổng quán tính của đám mây số liệu nói trên là tổng các phần tử trên đường chéo của V, tính được như sau:

$I_g = 132.89 + 0.8056 + 2421.4429 + 35289619.24 = 35292174.3785$

Còn nếu ta chọn $M = D_{1/S}^2$ thì $I_g = 4$. Như vậy quán tính này chỉ đo thứ bậc của tổng thể mà không đo sự biến thiên của nó.

4.3. Không gian các biến

Tương tự không gian các cá thể, mỗi véc tơ cột trong X tương ứng 1 điểm trong không gian E với số chiều là n. Để phân tích, chúng ta cũng phải xác định một Metric, tức là tìm 1 ma trận đường chéo dương cấp n, với

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

ma trận này ta xác định khoảng cách, chuẩn của các véc tơ tương tự như trong không gian các cá thể.

Tích vô hướng của hai véc tơ X_j và X_k được tính theo công thức:

$$X_j^T D X_k = \sum_{i=1}^{n} p_i x_{ji} x_{ki} \qquad (13.9)$$

Với cách chọn Metric này, tích vô hướng trên chính là hiệp phương sai của các cột trong X đã được trung tâm hoá (S_{kj}).

Ngoài ra có thể thấy chuẩn của các véc tơ cột chính là phương sai của các cột này (S^2_j), từ đó ta cũng có thể tính độ dài của véc tơ cột trong X.

Đối với các biến ta quan tâm đến quan hệ tương quan của chúng (điều này không được quan tâm đối với các cá thể vì ta luôn xem mẫu nghiên cứu là mẫu ngẫu nhiên). Tương quan của các biến có thể biểu diễn hình học bởi góc giữa các véc tơ trong không gian E n chiều.

Nếu gọi θ_{kj} là góc tạo bởi X_k và X_j thì ta có:

$$\text{Cos}\theta_{kj} = \frac{< X_k, X_j >}{\|X_j\| \|X_k\|} = \frac{S_{kj}}{S_k S_j} \qquad (13.10)$$

Thực chất đây là hệ số tương quan tuyến tính của X_k và X_j, hệ số này có tính đối xứng.

Với 4 biến trên và các số liệu quan sát trong X ta có Cosin các góc tạo bởi các véc tơ như sau:

	X_1	X_2	X_3
X_2	-0.74149		
X_3	-0.99041	0.722573	
X_4	0.652553	-0.52841	-0.65593

5. TẠO BIẾN VÀ PHÉP CHIẾU TRONG KHÔNG GIAN TUYẾN TÍNH

5.1. Tạo biến mới

Trong không gian tuyến tính mỗi véc tơ n chiều là một tập hợp có thứ tự các hệ số phân tích của chính véc tơ đó theo một hệ cơ sở đơn vị. Tuy

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

nhiên không gian n chiều có vô số các hệ cơ sở (mỗi hệ n véc tơ độc lập tuyến tính là một cơ sở). Như vậy một véc tơ X_j khác không bất kỳ có thể trở thành một véc tơ cơ sở của một cơ sở nào đó. Mặt khác cho trước một hệ p véc tơ của R^n ta luôn có thể tạo ra các véc tơ khác nhờ phép tổ hợp tuyến tính từ hệ này. Trong phân tích thống kê thì mỗi cách tổ hợp như vậy tạo nên một thống kê từ một mẫu có kích thước p. Thực chất của một thống kê là một cách tổng hợp thông tin, theo đó người ta có được những hiểu biết về tổng thể mà mỗi cá thể không thể hiện được.

Trở lại với không gian cá thể F đã nói ở trên, ta có thể tạo nên những véc tơ mới từ n véc tơ đã có nhờ các phép tổ hợp tuyến tính, mà theo đó các cá thể được phản ảnh tập trung hơn ở một khía cạch nào đó. Chẳng hạn khi xác định véc tơ g (véc tơ trung bình mẫu hay véc tơ trung tâm) ta đã có một véc tơ mới mà từ đó có thể mô tả các cá thể nhờ khoảng cách từ nó đến g, hay nhờ véc tơ chỉ sự khác biệt của mỗi cá thể với trung bình chung. Tuy nhiên mục đích của chúng ta, như đã nêu từ đầu là tóm tắt thông tin p chiều thành thông tin có số chiều ít hơn. Phép chiếu từ R^p lên một siêu phẳng cho phép chúng ta tóm tắt thông tin như vậy.

Biến mới được tạo ra ở đây là một tổ hợp tuyến tính của các biến ban đầu (các cột của X), như vậy sự liên hệ của các biến ban đầu với biến được tạo mới chính là các hệ số tổ hợp (hay ngược lại là các hệ số phân tích).

5.2. Phép chiếu

Với không gian các biến F, phép tổ hợp tuyến tính các véc tơ X_j (j=1..p) tạo nên các véc tơ mới trong F. Với các véc tơ này các cá thể trong $E \subset R^p$ được biểu diễn bằng các véc tơ mới trong $E' \subset R^k$ (E' thường có số chiều nhỏ hơn E rất nhiều). Các véc tơ mới này chính là các hình chiếu của các véc tơ ban đầu trên hệ toạ độ mới của.F.

Chẳng hạn, nếu ta lấy tổ hợp tuyến tính của các cột trong X là:

$$\Delta = 0{,}1X_1 - 0{,}2\,X_2 + X_3 + 0{,}5\,X_4$$

Thì ta có Δ là một véc tơ n chiều trong F, mỗi cá thể có một giá trị theo Δ như sau:

Cá thể	X1	X2	X3	X4	Hình chiếu
1	44	2.8	168	205	274.34
2	75	1.3	25.6	3408	1736.84
3	75	1.4	27	5000	2534.22

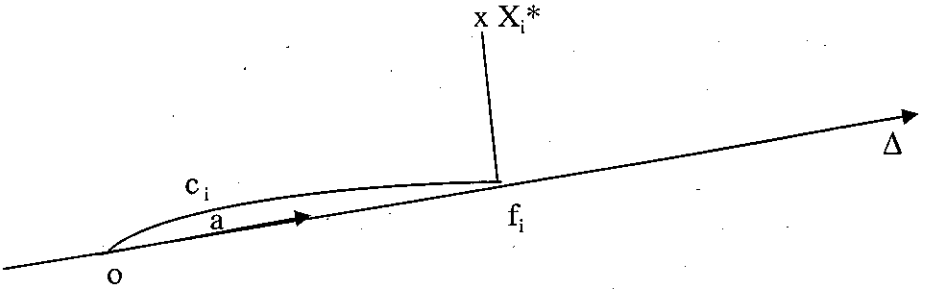
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4	80	1.4	7.3	16848	8439.02
5	79	0.2	6.7	18396	9212.56
6	75	1.4	35	3000	1542.22
7	74	2.4	25	7875	3969.42
8	53	2.4	106	202	211.82
9	78	0.2	20.3	6950	3503.06
10	76	0.3	19	6500	3276.54
HS tổ hợp	0.1	-0.2	1	0.5	

Thực tế là khi xác định một véc tơ mới như Δ, chúng ta đã đưa ra một cách qui đổi chung cho các biến theo một cách nào đó, với một mục đích nào đó, một ánh xạ từ R^4 về R^1 hay trên một trục.

Hình chiếu nhận được của các cá thể như một trong các chỉ tiêu tổng hợp để phân biệt các cá thể theo quan điểm tương ứng, chúng ta cũng có thể tạo một ánh xạ khác tương tự như trên và nhận được ảnh của các các thể trên một trục khác.

Có thể mô tả hình học như sau:



Một điểm (một dòng của X) ứng với cá thể i: X_i^* có ảnh qua phép chiếu (mà sau này được xác định là phép chiếu vuông góc) trên trục Δ là f_i. Trên trục Δ xác định một véc tơ chỉ phương a có độ dài bằng 1 (theo Metric M), lúc đó X_i^* tương ứng với một giá trị c_i là khoảng cách từ gốc O đến f_i. Với n cá thể ta có n giá trị c_1 , c_2, ... , c_n lập nên một véc tơ c (Δa) trong F.

Như vậy: $c_i = a^T M X_i^* = X_i^{*T} M a = <a, X_i^*>_M$ (13.11)

từ đó ta có: $c = X.M.a$ đặt $Ma = u$, u là một véc tơ trong F, ta có:

$c = Xu$, véc tơ c là hình chiếu của X trên u (13.12)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nếu ký kiệu $u = \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ u_p \end{bmatrix}$ ta có: $c = \sum_{j=1}^{p} u_j X_j$ như vậy c là tổ hợp tuyến tính của các cột X_j trong ma trận X (ma trận các biến).

Với phép biến đổi này, ta có tương ứng mỗi véc tơ đơn vị a một véc tơ hình chiếu c có các toạ độ là các hình chiếu của các véc tơ dòng trọng X (ứng với các cá thể) qua phép biến đổi XMa. Có thể xác định tối đa p véc tơ đơn vị trong F (không gian p chiều) và tương ứng ta có một tập các véc tơ c là các tổ hợp tuyến tính khác nhau của các cột trong X. Nhờ phép biến đổi này, ta có thể chiếu các cá thể trong F (n cá thể) lên một không gian có số chiều ít hơn. Điều đó cho phép nhận diện rõ hơn các quan hệ giữa các cá thể này cũng như các chỉ tiêu (các biến) ban đầu.

Quan hệ của véc tơ chỉ phương a và véc tơ u có thể tóm tắt bằng biểu thức sau:

$$a = M^{-1}u.$$

như vậy Metric của a là: a^TMa thì Metric của u là: $u^T M^{-1}u$. Có thể tìm được phương sai của véc tơ c như sau:

$$V(c) = c^TDc = (Xu)^TD(Xu) = u^TX^TDXu = u^TVu \qquad (13.13)$$

trong đó V là ma trận hiệp phương sai của X. Đây là biểu thức hết sức quan trọng cho thấy khả năng bảo tồn sự biến động của các cá thể cũng như các biến đã được mô tả trong X, bởi véc tơ c.

6. PHÂN TÍCH THÀNH PHẦN CHÍNH

6.1. Phép chiếu lên không gian con

1. Phép chiếu lên không gian con:

Nội dung cơ bản nhất của phân tích thành phần chính là tìm cách chiếu các điểm của một đám mây n điểm trong không gian p chiều thành một đám mây n điểm trong không gian con r chiều (không gian có số chiều nhỏ hơn, $r < p$). Phép chiếu này phải thực hiện theo những tiêu chuẩn nhất định, ở đây người ta tìm cách chiếu đám mây số liệu sao cho quán tính của

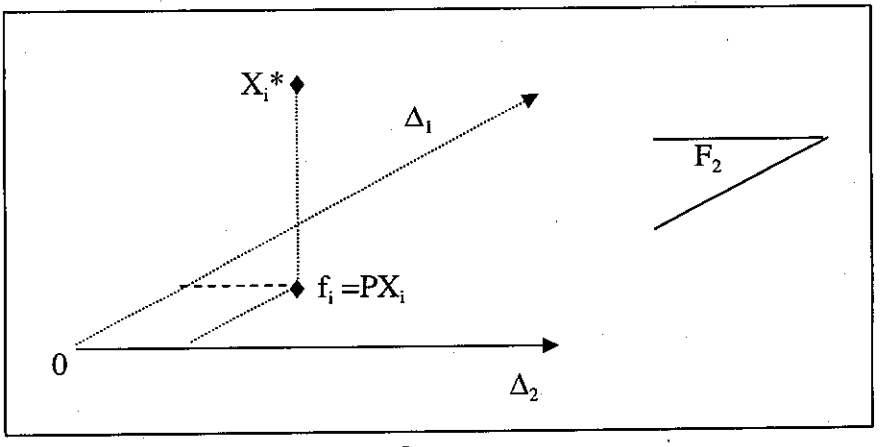
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

đám mây hình chiếu (đám mây ảnh) lớn nhất. Nói cách khác là phép chiếu đảm bảo trung bình của tổng bình phương các khoảng cách của các điểm trong không gian con lớn nhất có thể được.

Với phép chiếu P mỗi véc tơ dòng của X sẽ được biến đổi thành một véc tơ trong F_r qua công thức:

$$f_i = PX_i^* \text{ (một véc tơ cột) hay } f_i^T = X_i P^T \text{ (một véc tơ dòng).}$$

Có thể mô tả như sau:



2. Phép chiếu vuông góc các cá thể của R^p lên không gian F_r

Xét không gian con E của các cá thể y, P là một phép chiếu M vuông góc trên E nếu $Py \in G$ và $(Py, y-Py) = 0$, nói cách khác P là phép chiếu M-vuông góc xuống không gian con F_r, thì P có các tính chất sau đây:

+ $P^2 = P$ (P luỹ đẳng)
+ $MP = P^T M$.

Giả sử $P_1, P_2, ..., P_r$ là các phép chiếu M –vuông góc thì $(P_1 + P_2 + ... + P_r)$ là một phép chiếu M- vuông góc khi và chỉ khi $P_i P_j = 0$ với mọi $i \neq j$. Nói cách khác nếu G là F_r thì không gian con này thuộc không gian trực giao với E.

Với mỗi phép chiếu như vậy, ma trận hiệp phương sai của đám mây ảnh sẽ là:

$$(X P^T)^T D (X P^T) = P V P^T \qquad (13.14)$$

Quán tính tổng của đám mây ảnh là:

$$\text{Trace}(PVP^T M)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Nhờ một vài biến đổi ta có: $\text{Trace}(PVP^TM) = \text{Trace}(VMP)$ (13.15)

Vấn đề còn lại là tìm P sao cho với r cho trước, tức là tìm **không gian chiếu F_r,** Trace(VMP) lớn nhất.

Một mệnh đề đã được chứng minh*:*

Cho không gian con F_r có quán tính lớn nhất thì một không gian con r+1 chiều có quán tính lớn nhất bằng tổng vuông góc của không gian 1 chiều và không gian F_r có quán tính lớn nhất.

Với mệnh đề này thuật toán tìm không gian chiếu F_r sẽ bắt đầu từ không gian F_1 có quán tính lớn nhất.

Có thể xác định P như sau: Giả sử W là không gian ảnh các dòng của X, như đã nói ở trên P phải thỏa mãn điều kiện (y-Py) vuông góc với mọi véc tơ của W. Mặt khác các véc tơ trong W có dạng Xu và P là một phép biến đổi tuyến tính thì Py = Xb.

Điều kiện vuông góc trở thành $(Xu_i, y\text{-}Py) = 0$ cho mọi i = 1..n từ đó ta có:

$$u_i^T X^T M(y\text{-}Py) = 0 \text{ với mọi } i$$

hay $X^TMy = X^TMPy$, thay Py = Xb ta có: $X^TMy = X^TMXb$.

Suy ra $b = (X^TMX)^{-1} X^TMy$.

$$\Rightarrow \quad Xb = X(X^TMX)^{-1} X^TMy.$$

$$\text{Vậy} \quad P = X(X^TMX)^{-1} X^TM \qquad (13.16)$$

Với mỗi cá thể $x = X_i^*$ ta có: $P = x(x^TMx)^{-1}x^TM = \dfrac{x^TxM}{x^TMx}$

6.2. Trục chính, thành phần chính và nhân tố chính

1. Trục chính

Trong không gian $F \subset R^p$ các cá thể, ta phải tìm một đường thẳng đi qua g (véc tơ trung tâm của đám mây số liệu) sao cho quán tính của đám mây ảnh (ảnh đám mây số liệu ban đầu) trên trục này lớn nhất.

Giả sử a là véc tơ chỉ phương của đường thẳng nói trên, theo (7.16) phép chiếu M vuông góc lên đường thẳng này xác định như sau:

$$P = a\,(a^T M\, a\,)^{-1}a^TM \qquad (13.17)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Quán tính của đám mây ảnh sẽ là:

$$\text{TraceVMP} = \text{Trace VM a } (a^T M a)^{-1} a^T M$$

$$= \frac{1}{a^T Ma}\text{Trace}(VMaa^T M) = \frac{\text{Trace}(a^T MVMa)}{a^T Ma} \qquad (13.18)$$

Ma trận MVM được gọi là ma trận quán tính của đám mây, nó xác định một dạng toàn phương, với tất cả các véc tơ có chuẩn (độ dài) theo độ đo M bằng 1, biểu diễn quán tính của đám mây ảnh trên trục tạo bởi véc tơ chỉ phương a. Ma trận này chính là ma trận hiệp phương sai nếu M=E (ma trận đơn vị).

Phương pháp thành phần chính qui về bài toán tìm a cực đại TraceVMP. Giải bài toán này người ta nhận được:

$$VMa = \lambda a \qquad (13.19)$$

Như vậy a là véc tơ riêng (Eigen-Vector) của ma trận VM còn $\lambda = \frac{a^T MVMa}{a^T Ma}$ là 1 giá trị đặc trưng (hay còn gọi là giá trị riêng: Eigen-value) của VM. Giá trị riêng này lại chính là quán tính của đám mây ảnh trên trục chứa véc tơ chỉ phương a.

Điều kiện $\|a\|_M=1$ phương trình (13.18) trở thành:

$$VMa = (a^T MVMa)a = \lambda a \qquad (13.20)$$

Vậy cần tìm a (trục thứ nhất) tương ứng với giá trị riêng lớn nhất. Người ta chứng minh được tiêu chuẩn bậc 2 (điều kiện đủ) được thoả mãn, tại giá trị riêng này.

Kết luận: *Không gian r chiều (F_r) có cơ sở là r véc tơ chỉ phương tương ứng với r giá trị riêng lớn nhất (theo thứ tự giảm dần) của ma trận VM.*

Ta gọi ***các véc tơ riêng a của VM là các trục chính.***

Khó khăn chính trong ứng dụng phương pháp này cũng như các phương pháp phân tích nhân tố sau này chính là (13.19) dẫn đến bài toán tìm giá trị riêng và véc tơ riêng của một ma trận vuông cấp k (thường thì k đủ lớn). Giải bài toán này cần có các thuật toán riêng với khối lượng tính toán lớn, hầu như không thể tính toán thủ công. Chính điều này dẫn đến nhóm phương pháp phân tích nhân tố cần đến sự trợ giúp của các phần

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

mềm chuyên dụng. Có nhiều phần mềm như vậy, trong cuốn sách này các bài toán sẽ được thực hiện trên SPSS và cùng chỉ được trình bày tóm tắt.

2. Nhân tố chính

Như đã biết, một véc tơ riêng a tìm được tương ứng một dạng tuyến tính u (theo kết quả ở phần trên). Véc tơ u là một phần tử của không gian R^{*P}, không gian đối ngẫu của không gian cá thể F. Nó được xác định bởi một tổ hợp tuyến tính của các cột trong X.

Với trục chính a ta xác định một **véc tơ u = Ma gọi là nhân tố chính**. Có thể thấy rằng u là véc tơ riêng của MV. Thật vậy:

$$VMa = \lambda a \text{ từ đó ta có: } MVMa = \lambda Ma = \lambda u$$

Tức là: $$MVu = \lambda u \qquad (13.21)$$

biểu thức này chứng tỏ u là véc tơ riêng của MV và MV có cùng giá trị riêng λ.

Thực tế là R^P xác định một độ đo Metric M và không gian đối ngẫu của nó R^{*P} cũng được xác định một Metric M^{-1}. Trong đó $u^TM^{-1}u=1$ (u là véc tơ có chuẩn bằng 1).

Các véc tơ u lập nên các nhân tố chính M vuông góc.

3. Các thành phần chính

Véc tơ C là một tổ hợp tuyến tính các cột của X gọi là một thành phần chính của X (chuẩn hoá hay không chuẩn hoá) nếu C thoả mãn một số tính chất nào đó về phương sai. Một cách cụ thể tiêu chuẩn của C là phương sai của phép biến đổi tuyến tính đã nói ở trên lớn nhất.

Trở lại với phép chiếu đã nói ở trên, sau khi có các trục chính ta cũng có được các nhân tố chính u_i tương ứng.

Véc tơ các thành phần chính (C) là các véc tơ nhận được từ phép biến đổi (phép chiếu) X lên các trục u_i.

$C_i = Xu_i$ có thể thấy các thành phần chính này có các đặc trưng sau:

$$V(C_i) = \lambda_i \qquad (i=1..r) \qquad (13.22)$$

Nếu gọi C là ma trận lập bởi các thành phần chính thì phương sai của chúng chính là véc tơ λ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Cũng như các trục chính và các nhân tố chính, các thành phần chính là các véc tơ riêng của ma trận XMX^TD.

Có thể tóm tắt toàn bộ nội dung trên như sau:

Thành phần phân tích	Phương trình xác định	Độ đo
Nhân tố chính u	MVu = ?u	Chuẩn M-1
Trục chính a	VM a = ?a	Chuẩn M
Thành phần chính c	XMXTc = ?c	D- trực giao
Các liên hệ	c =Xu, ?; u=Ma	

6.3. Xác định λ và chọn số thành phần chính

Thực tế khi chúng ta phân tích một tổng thể qua một mẫu, việc chọn số thành phần chính tuỳ thuộc vào yêu cầu sử dụng và phân tích kết quả, tuy nhiên thông thường để nhìn thấy trực quan kết quả phân tích thành phần chính người ta thường chọn r=2 hoặc 3. Tất nhiên càng chọn nhiều thành phần chính thì mức giải thích càng cao. Số giá trị khác nhau của λ theo các phương trình nói ở cuối phần b là số nghiệm của đa thức cấp p.

Chẳng hạn, khi tìm trục chính ta cần giải phương trình: VM a = λa

Phương trình này tương đương với phương trình: (VM-Eλ) a = 0

Phương trình trên có nghiệm khác không khi và chỉ khi | VM-Eλ | =0. Định thức |VM-Eλ| là một đa thức bậc p của λ.

Theo các kết quả nói trên, để đám mây ảnh có quán tính lớn nhất thì trục chính thứ nhất là véc tơ a ứng với giá trị riêng lớn nhất của ma trận VM. Trục chính thứ hai là véc tơ a ứng với giá trị riêng lớn thứ hai của ma trận VM...

Để minh hoạ một cách đơn giản cho tất cả các nội dung nói trên ta xét một thí dụ nhỏ sau đây. Hãy xét chỉ hai biến ở thí dụ đang xét ở phần trên

X_1	44	75	75	80	79	75	74	53	78	76
X_2	2.8	1.3	1.4	1.4	0.2	1.4	2.4	2.4	0.2	0.3

Ta thực hiện tìm các giá trị riêng của VM với các trọng số của các cá thể đều bằng 1/10. Để đơn giản ta chọn M=E (độ đo thông thường).

Có thể tính được:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$V=\begin{bmatrix} 132.89 & -7.672 \\ -7.672 & 0.8056 \end{bmatrix}$$

Phương trình xác định λ là:

$$48.1966 \quad -133.6956\lambda + \lambda^2 = 0$$

Nghiệm của phương trình này là:

$$\lambda_1 = 133,33 \qquad \lambda_2 = 0,361.$$

Nếu ta chọn λ_1 để chiếu các cá thể lên một trục, thì trục này có véc tơ chỉ phương a xác định qua phương trình:

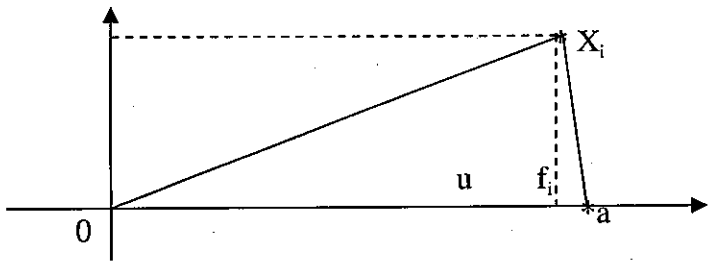
$$(132.88-133.33)a_1 - 7.672a_2 = 0$$
$$-7.672a_1 \quad (0.8056-133.33)a_2 = 0$$

$a_1^2 + a_2^2 = 1$ (điều kiện chuẩn a bằng 1)

Ta có: $a_2 = (-0,45/-7,672)a_1 = 0,058655a_1$.

$a_1 = 0,99657$, $a_2 = 0,058454$.

Có thể mô tả trục u trên R^2 như sau:



Trục chính u được chọn và 1 điểm X_i (cá thể i) chiếu xuống u ta có f_i.

Các toạ độ của 10 cá thể trên trên u có thể tính theo công thức: C=Xa

(44.0128 74.8189 74.8247 79.8076 78.7408 74.8247 73.8866 52.9586 77.7443 75.757)

6.4. Tái hiện dữ liệu

Khi tiến hành thủ tục phân tích thành phần chính, ta đã biến X thành một đám mây ảnh của nó trong các không gian có số chiều nhỏ hơn, với điều kiện tối đa hoá sự khác biệt của các cá thể (các điểm của đám mây ban đầu). Để có thể đánh giá trở lại vai trò của các biến ban đầu, ta cần tìm lại một số công thức liên hệ ngược, các công thức này thực tế là các phép chiếu ngược của phép chiếu đã xét ở trên, nói như vậy hoàn toàn không có nghĩa

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

831 là từ ảnh của một đám mây, ta hoàn toàn có thể tìm lại chính đám mây đó một cách dễ dàng.

Khi chọn M=E (ma trận đơn vị) thì

$$X = \sum_{j=1}^{p} c_j u_j^T = \sum_{j=1}^{p} \sqrt{\lambda_j} z_j v_j^T \qquad (13.23)$$

trong đó z_j là các véc tơ riêng của XX^T và v^T_j là các véc tơ riêng của X^TX.

Một kết quả nhận được từ Định lý EKar-Yuong là: *Tổng r số hạng đầu trong các biểu thức trên chính là xấp xỉ tốt nhất r chiều của X.*

6.5. Phân tích thành phần chính với ma trận hệ số tương quan

Lựa chọn Metric luôn là vấn đề cốt lõi trong mọi phân tích nhân tố. Thực tế chỉ có người thực hành mới biết chính xác cơ sở lựa chọn độ đo này. Thật vậy, trong kinh tế, xã hội sự khác biệt của các cá thể đôi khi không đo được bằng các phép tính thông thường và ngay cả khi nghiên cứu cùng một vấn đề ở các tổng thể khác nhau độ đo này cũng khác nhau. Một khu vực mới đầu tư với lợi nhuận siêu ngạch quá lớn làm cho người ta xác định sự khác nhau của các chủ đầu tư bởi khối lượng đầu tư, nhưng một khu vực đã no vốn đầu tư người ta có thể phân biệt sự khác nhau của các chủ đầu tư bởi lĩnh vực đầu tư gắn với tỷ suất lợi nhuận ròng. Với một quốc gia còn quá ít người tốt nghiệp đại học, người ta có thể chỉ xét đến điểm trung bình của các cử nhân để đánh giá họ, tuy nhiên khi số lượng này quá lớn thì hàng loạt chỉ số phụ mà đôi khi là chỉ số điều kiện được sử dụng (mặc dù có thể các chỉ số này quá ít ý nghĩa).

Trong các Metric hay được sử dụng, để có thể bỏ qua thứ nguyên (đơn vị đo) của các biến, người ta thường chọn một kiểu đo trên cơ sở chuẩn hoá. Lý do này làm cho việc chọn Metric $M = D_{1/S^2}$ trở thành thông dụng. Có thể nói việc sử dụng Metric này tương đương với việc rút gọn đám mây số liệu (***hay còn gọi là phân tích thành phần chính với ma trận hệ số tương quan***). Cũng vì lý do này người ta có thể xem phương pháp thành phần chính là một mở rộng của phương pháp phân tích tương quan.

Điều này tương đương với dùng Metric M=E đối với bảng số liệu đã chuẩn hoá (Z) tương ứng với bảng số liệu ban đầu (X).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì ma trận hiệp phương sai của bảng số liệu đã trung tâm hoá và chuẩn hoá chính là ma trận hệ số tương quan R, nên các nhân tố chính sẽ là các véc tơ riêng của R. Véc tơ u được xác định nhờ phương trình Ru = λu, với chuẩn của u bằng 1.

Thành phần chính đầu tiên c là tổ hợp tuyến tính của các biến đã được chuẩn hoá có độ phân tán cực đại: c =Xu. *Người ta chứng minh được tổng bình phương các hệ số tương quan tuyến tính của c với các cột cuả ma trận X là cực đại*. Nó được đo bởi λ tương ứng.

Tức là: $\sum_{j=1}^{p} r^2(c, X_j)$ đạt giá trị cực đại.

Điều này có nghĩa là véc tơ thành phần chính c có liên hệ chặt chẽ nhất với các biến ban đầu.

Đến đây có thể phát biểu tóm tắt phương pháp thành phần chính như sau:

Phân tích nhân tố bằng phương pháp thành phần chính là việc thay thế p biến ban đầu, có quan hệ tương quan với nhau, bằng một số các biến mới là tổ hợp tuyến tính của chúng. Các biến này không tương quan với nhau sao cho giữ được sự khác biệt tối đa giữa các cá thể, nhưng lại đảm bảo được sự liên hệ tối đa của chính các biến ban đầu. Đây thực chất là một cách phân tích nhân tố tuyến tính, nhờ các phép biến đổi tuyến tính.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7. PHÂN TÍCH VÀ ĐÁNH GIÁ KẾT QUẢ CỦA VIỆC ÁP DỤNG PHƯƠNG PHÁP THÀNH PHẦN CHÍNH TRONG PHÂN TÍCH NHÂN TỐ

Từ kết luận của mục V ta thấy nếu ta có p biến hoàn toàn độc lập thì việc phân tích nhân tố bằng phương pháp đã trình bày ở trên hoàn toàn không đem lại một chút lợi ích nào.

Có thể phân tích và đánh giá kết quả của một phân tích nhân tố bằng phương pháp thành phần chính qua các nội dung sau đây.

7.1. Các kiểm định chung

Trong phân tích nhân tố người ta thực hiện các kiểm định về sự hợp lý của các biến tham gia phân tích. Giả thuyết quan trọng của phân tích nhân tổ chính là các biến có quan hệ tương quan tuyến tính với nhau. Trong trường hợp ngược lại phân tích nhân tố không có ý nghĩa rút gọn số liệu mà chỉ là "bỏ bớt số liệu". Hai kiểm định thường dùng là kiểm định khi bình phương qua thống kê Bartlett và KMO.

+ Kiểm định Bartlett:

H_0: Ma trận hệ số tương quan R=E

H_1: Ma trận hệ số tương quan R≠E

Thống kê Bartlett:

$$\chi^2 = -(W - 1 - \frac{2p+5}{6}) \ln|R| \qquad (13.24)$$

Trong đó W là tổng trọng số của X, p là số biến dùng phân tích, |R| là định thức của ma trận hệ số tương quan.

Thống kê này phân phối khi bình phương với p(p-1)/2 bậc tự do. H_0 bị bác bỏ nếu giá trị quan sát của thống kê này lớn hơn $\chi^2_{(\alpha,p(1-p)/2)}$ (giá trị tới hạn mức α của phân phối khi bình phương).

+ Kiểm định KOM (Kaiser- Mayer - Olkin)

Kiểm định KMO là tỷ lệ của hệ số tương quan trong tổng hệ số tương quan và hệ số tương quan riêng, KMO sử dụng để kiểm tra tính phù hợp của mẫu đối với phân tích nhân tố.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thống kê KMO được tính như sau:

Với mỗi biến j

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^{2*}}$$

Với p biến

$$KMO = \frac{\sum_{j=1}^{p} \sum_{i \neq j} r_{ij}^2}{\sum_{j=1}^{p} \sum_{i \neq j} r_{ij}^2 + \sum_{j=1}^{p} \sum_{i \neq j} a_{ij}^{2*}} \qquad (13.25)$$

Trong đó: r_{ij} là các hệ số tương quan của các biến; a_{ij}^* là các hệ số của ma trận hệ số tương quan riêng phần của các cặp biến trong điều kiện của tất cả các biến khác (anti-image correlation). Tất cả các hệ số này được cung cấp từ các phần mềm chuyên dụng.

Thông thường người ta yêu cầu KMO >0,5.

Trong thực hành tính toán KMO khá phức tạp vì vậy người ta phải nhờ đến sự trợ giúp của các phần mềm chuyên dụng.

Trường hợp R suy biến các hệ số trên không tính được, trong trường hợp này phân tích nhân tố không có ý nghĩa rút gọn số liệu.

7.2. Tương quan của các thành phần và các biến ban đầu

Để phản ánh mức tương quan này, ta có thể tính hệ số tương quan của các thành phần chính với các biến ban đầu. Hệ số tương quan tuyến tính của c và X_j: $r(c,X_j)$ được tính như sau:

Hãy chọn Metric $D_{1/s}{}^2$ hay nói cách khác là ta chọn Metric M=E cho số liệu trung tâm hoá Z. Ta có: $r(c,X_j) = r(c,Z_j)$ với $Z_j = X_j/S_j$

$$r^2(c,Z_j) = \frac{\left[\text{cov}(c, Z_j)\right]^2}{V(c)V(Z_j)} = \frac{\left[(c^T D Z_j)\right]^2}{V(c)}$$

trong đó D là ma trận trọng số, và $V(Z_j)=1$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì V(c) = λ ta có: $r(c,X_j) = (c^T DZ_j)/\sqrt{\lambda}$ (13.26)

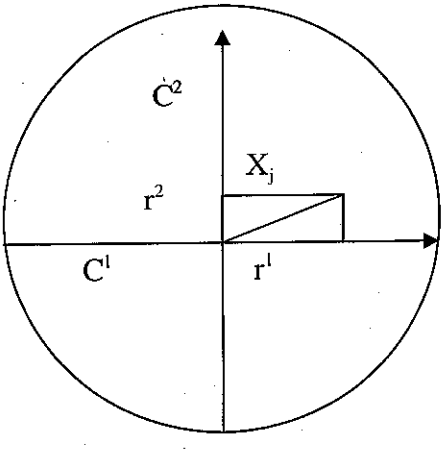
Nếu u là nhân tố chính tương ứng với giá trị λ thì: c =Zu (λ là giá trị riêng của R). Từ đó ta có:

$$r(c,X_j) = \frac{u^T Z^T DZ_j}{\sqrt{\lambda}} = \frac{Z_j^T DZu}{\sqrt{\lambda}}$$

Vì Ru= λu dễ dàng thấy tử số chính là thành phần chính thứ j của R.

Vậy: $r(c,X_j) = \frac{\lambda u_i}{\sqrt{\lambda}} = \sqrt{\lambda}\, u_j.$ (13.27)

Rõ ràng phân tích trên chỉ thực sự có ý nghĩa khi ta so sánh hai thành phần chính. Giả sử c^1 và c^2 là hai thành phần chính tương ứng với hai giá trị riêng lớn nhất, một biến X_j sẽ có tương ứng hai hệ số tương quan r^1 và r^2 theo hai thành phần chính này. Hai hệ số này được biểu diễn bởi 1 điểm trên hệ trục toạ độ hai chiều như sau:



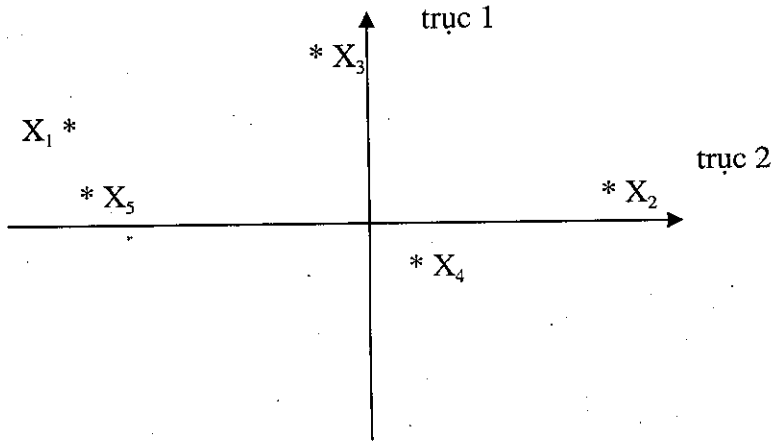
$r^2=(c^2,X_j)$

$r^1=(c^1,X_j)$

Một hình ảnh như vậy cho thấy c^1 quan hệ với X_j chặt chẽ hơn c^2, nó mang nhiều thông tin về X_j hơn c^2.

Một hình ảnh tương tự cũng có thể thiết lập đối với các trục chính, ta biết rằng c = Xu, vì vậy mỗi biến X_j cũng có thể mô tả tương tự trên mặt phẳng với hai trục chính (chẳng hạn u_1 và u_2).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Một hình ảnh như vậy cho thấy X_1, X_5 và X_2 được giải thích bởi trục 2 là chính, mặc dù chúng tương quan ngược chiều nhau, còn X_3 và X_4 thì được giải thích bởi trục 1 là chính. Ngoài ra cũng có thể thấy X_1, X_5 gần như độc lập với X_3.

Để đo tỷ lệ đóng góp của biến X_j theo một trục chính u, người ta dùng tỷ lệ:

$$(r(u_j))^2 = \frac{r^2(c, X_j)}{\lambda} \qquad (13.28)$$

Thực chất đây là tỷ lệ tương quan của X_j với c trong tổng bình phương quán tính của đám mây ảnh theo c.

Với thành phần chính thứ k ta có:

$$\lambda_k = \sum_{j=1}^{p} r^2(c_k, X_j)$$

7.3. Các hệ số phản ánh liên hệ của các cá thể và các thành phần chính

Ta nói rằng c^l quan hệ rất chặt với X_j có nghĩa là các cá thể có toạ độ lớn trên trục thứ nhất được đặc trưng bởi một giá trị lớn hơn giá trị trung bình (người ta nói rằng gốc của trục này chính là tâm của đám mây số liệu).

Ngược lại, nếu các cá thể đồng nhất, nó cho phép ta giải thích rõ hơn về ý nghĩa của thành phần chính và trục chính. Để thấy rõ hơn cách giải thích này ta xét trường hợp một số cá thể nằm ngay trên một trục nào đó. Có thể tính cho mỗi trục như vậy mức đóng góp của các cá thể trên trục đó. Gọi thành phần chính thứ k là c^k, tức là c^k_i là giá trị của thành phần chính này

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

tương ứng với cá thể thứ i (i=1..n). Ta có:

$$\sum_{i=1}^{n} p_i (c_i^k)^2 = \lambda_k$$

Đóng góp của cá thể i giải thích cho thành phần chính thứ k, ký hiệu $CTR_k(i)$ sẽ là:

$$CTR_k(i) = \frac{p_i (c_i^k)^2}{\lambda_k} \qquad (13.29)$$

Ngược lại để đo mức độ đánh giá cá thể i nhờ ảnh của nó bởi thành phần chính thứ k, người ta sử dụng chỉ số CO2, được tính như sau:

$$CO2_k(i) = \frac{\left(c_i^k\right)^2}{\left\|X_i^* - g\right\|^2} \qquad (13.30)$$

Khi sử dụng ma trận hệ số tương quan phân tích thành phần chính thì:

$$\left\|X_i^* - g\right\|^2 = \sum_{j=1}^{p} z_{ij}^2$$

Tổng lượng thông tin phản ánh được qua r thành phần chính đối với cá thể i là:

$$QLT_r(i) = \sum_{k=1}^{r} CO2_k(i) \qquad (13.31)$$

7.4. Thí dụ bằng số

Trong thí dụ này ngoài việc mô tả một số thành phần cơ bản trong phân tích nhân tố bằng phương pháp thành phần chính, chúng ta sẽ xem xét cụ thể việc tính toán các thành phần chính và các nhân tố chính.

+ Số liệu ban đầu

Xét bảng số liệu tiêu dùng lương thực, thực phẩm của tám tầng lớp xã hội ở Pháp năm 1972 được mô tả như sau:

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
AGRI	167	1	163	23	41	8	6	6

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

SAAG	162	2	141	12	40	12	4	15
PRIN	119	6	69	56	39	5	13	41
CSUP	87	11	63	111	27	3	18	39
CMOY	103	5	68	77	32	4	11	30
EMPL	111	4	72	66	34	6	10	28
OUVR	130	3	76	52	43	7	7	16
INAC	138	7	117	74	53	8	12	20

Trong đó:

AGRI - Nông dân	PAO — Bánh mỳ truyền thống
SAAG- CN nông nghiệp	PAA — Bánh mỳ khác
PRIN - Viên chức độc lập	VIO - Vang truyền thống
CSUP - Viên chức có CMKT -	VIA -Vang khác
CMOY- Viên chức NN	POT - Khoai tây
EMPL - Chủ DN	LEC - Các loại rau
OUVR - Công nhân	RAI - Vang-Nho dễ bán
INAC - Không HĐ kinh tế -Xã hội	PLP - Món khai vị

+ *Ma trận Z*

1.429	-1.218	1.724	-1.146	0.3017	0.4867	-0.9261	-1.5004
1.2498	-0.904	1.1569	-1.497	0.1747	1.9025	-1.3751	-0.76551
-0.291	0.3536	-0.699	-0.092	0.0476	-0.575	0.64547	1.35751
-1.438	1.9249	-0.854	1.6645	-1.477	-1.283	1.76802	1.1942
-0.865	0.0393	-0.725	0.5788	-0.842	-0.929	0.19645	0.45931
-0.578	-0.275	-0.622	0.2275	-0.587	-0.221	-0.0281	0.296
0.103	-0.589	-0.519	-0.22	0.5557	0.1327	-0.7016	-0.68386
0.3897	0.6678	0.5382	0.483	1.8259	0.4867	0.42096	-0.35724

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Ma trận hệ số tương quan của các biến (R)

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
PAO	1.000	-0.774	0.926	-0.906	0.656	0.889	-0.833	-0.856
PAA	-0.774	1.000	-0.604	0.904	-0.333	-0.673	0.959	0.771
VIO	0.926	-0.604	1.000	-0.750	0.517	0.792	-0.669	-0.828
VIA	-0.906	0.904	-0.750	1.000	-0.419	-0.839	0.924	0.720
POT	0.656	-0.333	0.517	-0.419	1.000	0.603	-0.410	-0.554
LEC	0.889	-0.673	0.792	-0.839	0.603	1.000	-0.824	-0.751
RAI	-0.833	0.959	-0.669	0.924	-0.410	-0.824	1.000	0.834
PLP	-0.856	0.771	-0.828	0.720	-0.554	-0.751	0.834	1.000

Hai giá trị riêng lớn nhất của R là: 6.20795; 0.87968

Hai véc tơ riêng tương ứng là:

U	1	2
PAO	-0.1571	0.14695
PAA	0.13994	0.46975
VIO	-0.1401	0.21503
VIA	0.14996	0.27754
POT	-0.0989	0.79307
LEC	-0.1464	0.1365
RAI	0.14973	0.34756
PLP	0.14516	-0.0536

- Tìm các thành phần chính: Ta biết các thành phần chính có thể nhận được từ công thức c=Xu. Nhân lần lượt ma trận số liệu ban đầu (X) với các nhân tố chính vừa tính được ta có các thành phần chính như sau:

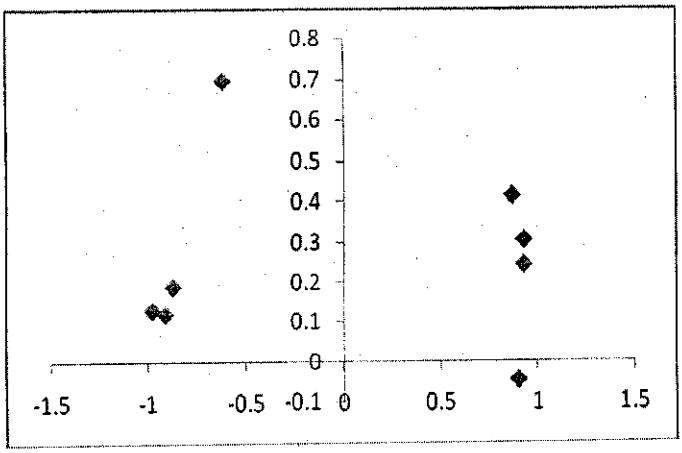
C(1)	C(2)
-1.266	-0.245
-1.322	-0.446
0.553	0.058
1.636	0.176
0.645	-0.854
0.303	-0.806
-0.338	-0.183
-0.211	2.301

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Hệ số tương quan các biến ban đầu và các thành phần chính - Mô tả đồ thị

Biến	C(1)	C(2)
PAO	-.975	.129
PAA	.869	.413
VIO	-.870	.189
VIA	.931	.244
POT	-.614	.698
LEC	-.909	.120
RAI	.929	.306
PLP	.901	-.047

Đồ thị các biến trên hệ toạ độ các thành phần chính



- Các hệ số đo quan hệ của các cá thể và các thành phần chính

Bảng giá trị $\left\|X_i^* - g\right\|^2 = \sum_{j=1}^{p} z_{ij}^2$ của các cá thể

AGRI	SAAG	PRIN	CSUP	CMOY	EMPL	OUVR	AGRI
12.084	11.2462	3.29974	17.6511	3.4306	1.33054	1.96154	4.99629

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các hệ số CO2, CTR và QLP

	QLT	Axe1	CO2	CTR	Axe2	CO2	CTR
PAO	0.1478	-1.266	0.1425	0.03226	-0.245	0.0053	0.0008
PAA	0.1611	-1.322	0.1447	0.0352	-0.446	0.0165	0.0023
VIO	0.0936	0.553	0.0926	0.00615	0.058	0.0010	0.0001
VIA	0.1535	1.636	0.1517	0.05392	0.176	0.0017	0.0002
POT	0.3340	0.645	0.1213	0.00838	-0.854	0.2128	0.0302
LEC	0.5576	0.303	0.0689	0.00185	-0.806	0.4887	0.0694
RAI	0.0751	-0.338	0.0581	0.00229	-0.183	0.0170	0.0024
PLP	1.0683	-0.211	0.0089	0.0009	2.301	1.0594	0.1505

Như vậy biến VIA, PAA và PAO có lượng tin được phản ánh cao nhất qua hai thành phần chính thứ nhất; PLP và POT PAO có lượng tin được phản ánh cao nhất qua hai thành phần chính thứ hai. Bảng trên cũng cho biết đóng góp của các cá thể đối với sự khác biệt của chúng qua hai trục chính đầu tiên.

Khi kiểm định Kaiser và KMO đối với phân tích trên, người ta thấy định thức ma trận R bằng 0 vì vậy tồn tại ít nhất 1 biến là tổ hợp tuyến tính của các biến còn lại ít nhất là với quan sát mẫu đang xét. Vì vậy cần tìm một biến nào đó lại khỏi tập biến trên. Có thể thấy PAO là biến có hệ số tương quan cao với hầu hết các biến khác, biến này có thể bị loại khỏi phân tích trên. Dưới đây là các kết quả chính phân tích thành phần chính khi loại PAO. Bạn đọc có thể dựa trên các kết quả này tính toán các chỉ số, 2 thành phần chính, kiểm định lại các giả thuyết nói trên.

Correlation Matrix(a)

		paa	vio	via	pot	lec	rai	plp
Correlation	paa	1.000	-.604	.904	-.333	-.673	.959	.771
	vio	-.604	1.000	-.750	.517	.792	-.669	-.828
	via	.904	-.750	1.000	-.419	-.839	.924	.720
	pot	-.333	.517	-.419	1.000	.603	-.410	-.554
	lec	-.673	.792	-.839	.603	1.000	-.824	-.751
	rai	.959	-.669	.924	-.410	-.824	1.000	.834
	plp	.771	-.828	.720	-.554	-.751	.834	1.000

a Determinant = 3.88E-006

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.563
Bartlett's Test of Sphericity	Approx. Chi-Square	47.762
	df	21
	Sig.	.001

Anti-image Matrices

		paa	vio	via	pot	lec	rai	plp
Anti-image Correlation	paa	.631(a)	.374	-.191	.479	-.868	-.805	.308
	vio	.374	.506(a)	.717	.588	-.621	-.794	.930
	via	-.191	.717	.728(a)	.303	-.038	-.395	.749
	pot	.479	.588	.303	.411(a)	-.681	-.678	.638
	lec	-.868	-.621	-.038	-.681	.560(a)	.880	-.533
	rai	-.805	-.794	-.395	-.678	.880	.535(a)	-.777
	plp	.308	.930	.749	.638	-.533	-.777	.545(a)

a Measures of Sampling Adequacy(MSA)

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.267	75.244	75.244	5.267	75.244	75.244
2	.861	12.295	87.539	.861	12.295	87.539
3	.404	5.772	93.312			
4	.301	4.303	97.615			
5	.150	2.137	99.753			
6	.014	.199	99.952			
7	.003	.048	100.000			

Extraction Method: Principal Component Analysis.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Component Matrix(a)

	Component	
	1	2
paa	.884	.382
vio	-.853	.195
via	.931	.225
pot	-.600	.724
lec	-.907	.146
rai	.944	.271
plp	.904	-.077

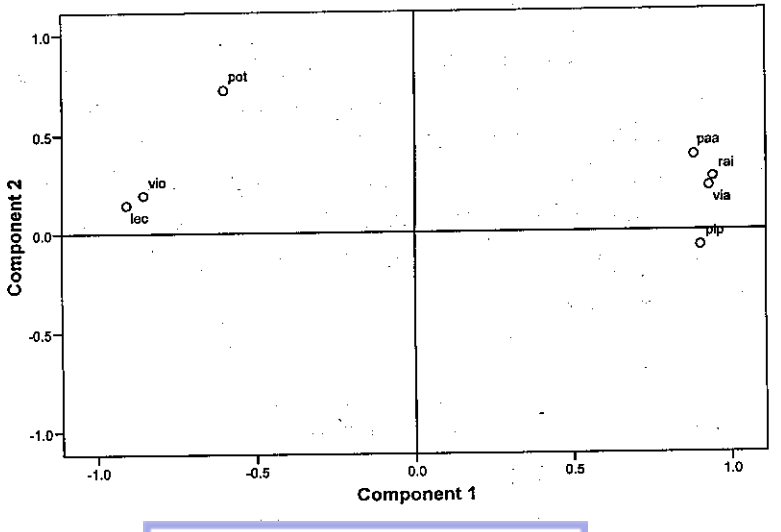
Extraction Method: Principal Component Analysis.

a 2 components extracted.

Component Score Coefficient Matrix

	Component	
	1	2
paa	.168	.444
vio	-.162	.227
via	.177	.262
pot	-.114	.842
lec	-.172	.169
rai	.179	.315
plp	.172	-.090

Component Plot



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8. TỶ LỆ GIẢI THÍCH TRÊN CÁC SIÊU PHẲNG CHIẾU

Như đã nói ở đầu chương, phân tích thành phần chính cho phép biểu diễn đám mây số liệu trên một không gian có số chiều nhỏ hơn. Các không gian như vậy là các siêu phẳng trong không gian các biến ban đầu p chiều. Chúng ta sẽ chỉ ra tỷ lệ được giải thích trên các siêu phẳng như vậy cũng như các bộ phận của chúng.

8.1. Mức giải thích chung (tổng cộng)

Chúng ta có thể chọn r nhân tố để mô tả, giải thích cho p biến ban đầu. Các nhân tố này tương ứng với r giá trị riêng λ_1, λ_2 ,..., λ_r. Tỷ lệ giải thích được là:

$$T_r = \frac{\lambda_1 + \lambda_2 + ... + \lambda_r}{I_g}$$

Chẳng hạn với hai thành phần chính ta tính được: $T_2 = \frac{\lambda_1 + \lambda_2}{I_g} = 0{,}67$ thì có nghĩa là có 67% đặc trưng của số liệu được giải thích bởi hai thành phần đầu tiên. Đám mây ảnh mang 67% đặc tính của đám mây số liệu ban đầu.

8.2. Tỷ lệ bộ phận

Khi tìm các thành phần chính, ta luôn hướng đến việc tăng dần tổng mức giải thích như trên. Khi đã có k thành phần, tức là ta xác định một không gian k chiều, một thành phần tiếp theo được chọn thuộc một không gian 1 chiều vuông góc với không gian đã có. Vậy chỉ tiêu chọn số chiều có thể đo bằng khoảng tỷ lệ của mỗi biến X_j với hình chiếu của nó trên siêu phẳng mới với phần còn lại của quán tính chưa được giải thích bởi k thành phần chính đã chọn, tỷ lệ này là:

$$\frac{d(X_j, f_j)}{\sqrt{I_g - \lambda_1 - \lambda_2 - ..... - \lambda_k}}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

9. TIÊU CHUẨN CHỌN SỐ THÀNH PHẦN CHÍNH CHO MỘT PHÂN TÍCH

Một trong những vấn đề luôn được quan tâm nữa là số trục chính hay số thành phần chính được chọn khi phân tích. Nhiều tiêu chuẩn khác nhau đã được sử dụng, trong đó có tiêu chuẩn tỷ lệ giải thích đã nói ở trên. Tuy nhiên, tiêu chuẩn này không có cơ sở khách quan, vì nó chỉ đáp ứng mong muốn của người phân tích mà không cho biết chất lượng phân tích thực sự. Sau đây là một trong cách cách chọn có cơ sở lý thuyết và ứng dụng của nó.

9.1. Tiêu chuẩn lý thuyết

Một cách thông thường người ta sẽ dùng r thành phần chính tương ứng với r giá trị riêng đầu tiên xếp theo thứ tự giảm dần λ_1 , λ_2 ,..., λ_r, với điều kiện các giá trị riêng λ_{r+1} , λ_{r+2} ,..., λ_p sai khác nhau không đáng kể. Điều này hàm ý rằng các nhân tố tiếp theo không phản ảnh được các đặc trưng khác nhau của các cá thể.

Giả thuyết được kiểm định là:

H_0: $\lambda_{r+1} = \lambda_{r+2} = ... = \lambda_p$

H_1: λ_{r+t} không đồng nhất.

Thống kê:

$$\chi^2 = (n - \frac{2p+11}{6})(p-r)\ln(\frac{a}{g}) \qquad (13.32)$$

trong đó: a là trung bình số học còn g là trung bình hình học của p-r giá trị riêng này, phân phối χ^2 với (p-r+2)(p-r-1)/2 bậc tự do.

Giả thuyết H_0 bị bác bỏ nếu giá trị quan sát của thống kê này lớn hơn giá trị tới hạn mức ý nghĩa α (α cho trước).

9.2. Tiêu chuẩn Kaiser

Kaiser đưa ra tiêu chuẩn về việc chọn số thành phần chính được nhiều người sử dụng, tiêu chuẩn này dựa trên chênh lệch của các giá trị riêng liên tiếp.

Thực tế là ông dựa trên cơ sở xem xét quan hệ tổng hợp thông tin của các thành phần chính c^k. Vì $V(c)^{\cdot} = \lambda$ nên các thành phần chính kế tiếp có

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

phương sai khác nhau đúng bằng chênh lệch của các giá trị riêng tương ứng.

Người ta lập các hiệu $e_1 = \lambda_1 - \lambda_2$, $e_2 = \lambda_2 - \lambda_3$,...

Căn cứ vào các chênh lệch $e_1 - e_2$, $e_2 - e_3$, ... người ta bỏ đi các thành phần có chính bắt đầu từ khi hiệu $e_r - e_{r-1} < \varepsilon$ cho trước.

10. MÔ HÌNH HỒI QUI THÀNH PHẦN CHÍNH

Sử dụng kỹ thuật kinh tế lượng ước lượng hồi qui tuyến tính nhiều biến, chúng ta thường gặp hiện tượng đa cộng tuyến. Hiện tượng này là cho các ước lượng của các tham số hồi qui có thể không chính xác và đôi khi sai dấu. Nguyên nhân của hiện tượng này là các biến độc lập quan hệ phụ thuộc tương quan tuyến tính lẫn nhau, thông tin giải thích cho biến độc lập chồng chéo.

Kỹ thuật hồi qui thành phần chính là một trong nhiều giải pháp khắc phục hiện tượng này. Trong phần này chúng ta sẽ tìm hiểu mô hình hồi qui thành phần chính như mô hình thống kê với giai đoạn cụ thể và cách phân tích kết quả của mô hình này.

10.1. Đa cộng tuyến và hồi qui thành phần chính

Giả sử chúng ta cần sử dụng các biến X_1, X_2, ...,X_k để giải thích cho Y nhờ mô hình hồi qui tuyến tính:

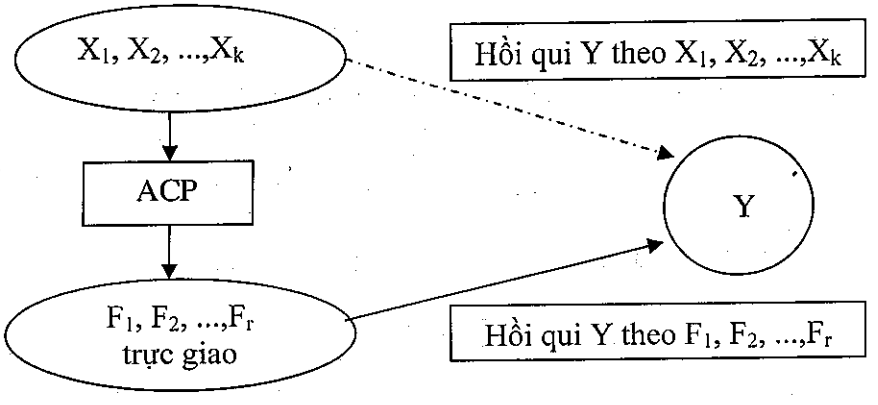
$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + U_i \qquad (13.33)$$

Nếu X_1, X_2, ...,X_k phụ thuộc tương quan tuyến tính lẫn nhau, kết quả ước lượng hồi qui (7.36) sẽ nhận được các ước lượng chệch của các tham số. Việc phân tích hồi qui không hiệu quả và thậm chí gặp các sai lầm đáng tiếc. Thực tế người ta có thể tìm cách bỏ đi một số biến mà vai trò giải thích cho Y không đủ lớn. Tuy vậy, có hai cản trở chính: Thứ nhất là việc lựa chọn biến loại khỏi mô hình không dễ dàng; Thứ hai là không thể loại một số biến mà về mặt lý thuyết nhất thiết các biến này phải có mặt trong mô hình. Chẳng hạn khi ước lượng hàm cầu một mặt hàng người ta không thể bỏ biến giá cho dù biến này tác động không đủ lớn đến lượng cầu.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giải pháp ở đây là tìm r biến độc lập là tổ hợp của k biến ban đầu và hồi qui Y theo r biến này. Các biến đó chính là các thành phần chính trong nhận được từ kết quả phân tích thành phần chính.

Sơ đồ tổng quát của quá trình này là:



10.2. Tác động của các biến độc lập đến biến phụ thuộc

Mục tiêu quan trọng nhất của phân tích hồi qui chính là xác nhận và đo sự tác động của các biến độc lập đến biến phụ thuộc.

Với mô hình ACP (phân tích thành phần chính) chúng ta lựa chọn r thành phần chính sao cho tỷ lệ bảo tồn phương sai đủ lớn (không ngoại trừ r = k). Tương ứng với r thành phần chính chúng ta nhận được r nhân tố chính $U_1, U_2, ...,U_r$. Mối liên hệ của các biến ban đầu với các thành phần chính được mô tả qua hệ sau:

$$F_{vi} = \sum_{j=1}^{k} u_{vj} \frac{X_{ij} - \bar{X}_j}{S(X_j)} \qquad v = 1,2,..,r; \quad i = 1,2,...,n \tag{13.34}$$

trong đó $u_{vj} = \beta_{vj} \frac{S(X_j)}{S(F_v)} = \beta_{vj} S(X_j)$ trong hồi qui $F_{vi} = \beta_0 + \sum_{j=1}^{k} \beta_{vj} X_{ij}$

Như vậy (13.34) trở thành:

$$F_{vi} = \sum_{j=1}^{k} \beta_{vj} (X_{ij} - \bar{X}_j) \qquad v = 1,2,..,r; \quad i = 1,2,...,n \tag{13.35}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$\frac{\partial F_v}{\partial X_j} = (1-\frac{1}{n})\beta_{vj} \quad (13.36)$$

Vậy khi X_j tăng 1 đơn vị thì F_v thay đổi 1 lượng là: $(1-1/n)\beta_{vj}$.

Gọi hồi qui Y theo $F_1, F_2, ..., F_r$ là: $Y = \alpha_0 + \sum_{v=1}^{r} \alpha_v F_v + \varepsilon$ (13.37)

Khi X_j tăng 1 đơn vị sẽ tác động đến Y thay đổi một lượng:

$$d_j Y = \sum_{v=1}^{r} \alpha_v (1-\frac{1}{n})\beta_{vj} \quad j=1,2,...,k \quad (13.38)$$

Khi n đủ lớn có thể bỏ qua hạng tử 1/n và (13.38) xấp xỉ:

$$d_j Y = \sum_{v=1}^{r} \alpha_v \beta_{vj} \quad j=1,2,...,k \quad (13.39)$$

10.3. Một số thí dụ thực nghiệm

Thí dụ 1: Mô hình hiệu quả sử dụng tài sản của doanh nghiệp

Trong thí dụ này, chúng ta xem xét mô hình hồi qui với:

Biến phụ thuộc là:

Lợi nhuận sau thuế của doanh nghiệp/Tổng tài sản (**LN_TTS**)

Các biến độc lập là: Tuổi doanh nghiệp (**tuoi_dn**); - Số năm kinh nghiệm của tổng giám đốc (**knghiem_gd**); - Trình độ của tổng giám đốc (**trinhdo_gd**): Đại học = 2; Trên đại học = 1; Vòng quay hàng tồn kho (**VQ_HTK**); - Kỳ thu tiền bình quân (**KTT_BQ**); Hiệu quả sử dụng tài sản (**HQSD_TS**); - Nợ phải trả trên Vốn chủ sở hữu (**NPT_VCSH**); - Lợi nhuận sau thuế trên doanh thu (**LN_DT**); Lợi nhuận sau thuế trên vốn chủ sở hữu (**LN_VCSH**); - Tỷ số nợ (**tyso_N**); Hệ số đòn bẩy (**HS_donbay**); - Khả năng thanh toán lãi vay (**KN_TTLV**); Quy mô hoạt động (**quymo**); Lớn =1; Trung bình = 2; Nhỏ = 3;Vòng quay các khoản phải thu (VQ_KPT).

Kết quả hồi qui tuyến tính sau đây cho thấy mô hình có đa cộng tuyến (Hệ số Tolerance và VIF).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	6.47711	5.890		1.100	.275		
	nganh_kt	.27767	.960	.028	.289	.773	.779	1.284
	tphan_kt	.32984	1.167	.030	.283	.778	.628	1.592
	Tuoi_dn	.01792	.121	.017	.149	.882	.525	1.904
	knghiem_gd	.36833	.142	.258	2.586	.012	.711	1.406
	trinhdo_gd	-.09642	1.547	-.006	-.062	.950	.664	1.505
	KNTT_NH	1.21646	1.268	.455	.959	.341	.031	31.866
	KNTT_N	-1.28971	1.339	-.453	-.963	.339	.032	31.312
	VQ_HTK	-.01995	.019	-.098	-1.073	.287	.841	1.190
	KTT_BQ	-.00038	.001	-.041	-.382	.704	.603	1.657
	HQSD_TS	.54337	.503	.106	1.081	.283	.728	1.373
	LN_DT	.00002	.003	.001	.007	.995	.816	1.225
	Tyso_N	-7.30002	3.083	-.254	-2.368	.021	.613	1.633
	HS_donbay	.00912	.015	.058	.619	.538	.808	1.238
	KN_TTLV	.19379	.044	.408	4.427	.000	.834	1.199
	Quymo	-1.67035	.845	-.202	-1.976	.052	.678	1.475
	VQ_KPT	-.00004	.000	-.101	-1.141	.258	.898	1.114

a. Dependent Variable: LN_TTS

Tiến hành phân tích thành phần chính với các biến độc lập với 8 thành phần chính, chúng ta có 8 biến mới F_1, F_2,..., F_8 với ma trận hệ số tổ hợp như sau:

Component Score Coefficient Matrix

	Component							
	1	2	3	4	5	6	7	8
Tuoi_dn	.020	-.190	.412	-.164	-.072	.135	.028	.003
knghiem_gd	.059	-.216	.280	-.072	-.062	.182	.385	.270
trinhdo_gd	-.200	-.072	-.144	.080	-.362	-.097	.246	-.033
KNTT_NH	.310	.115	.027	.186	-.024	.080	-.056	.016
KNTT_N	.313	.120	.051	.131	-.044	.030	.003	.035
VQ_HTK	.054	-.059	.109	.106	.249	-.641	.651	.011
KTT_BQ	.201	.148	.096	-.304	-.093	-.013	.105	-.083
HQSD_TS	-.024	-.124	.096	.249	.456	-.177	-.340	-.372
NPT_VCSH	-.138	.322	.277	.183	-.020	.058	.084	-.037
LN_DT	-.061	-.135	.096	.365	.096	.232	-.093	.652
Tyso_N	-.167	.114	.142	-.331	.083	-.078	-.047	.031
HS_donbay	-.138	.322	.277	.183	-.020	.058	.084	-.037
KN_TTLV	.014	-.078	-.027	.326	-.456	-.018	.166	-.383
Quymo	-.005	.190	-.331	-.006	.143	.009	.289	.456
VQ_KPT	-.026	-.020	-.122	.011	.326	.629	.511	-.430

Extraction Method: Principal Component Analysis.
Component Scores.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tiến hành hồi qui LN_TTS theo F_1, F_2,..., F_8 sau khi loại bỏ các biến không tác động đến LN_TTS ta có:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	3.688	.568		6.495	.000		
	F1	.880	.571	.131	1.541	.127	1.000	1.000
	F2	-2.302	.571	-.341	-4.030	.000	1.000	1.000
	F3	1.142	.571	.169	1.999	.049	1.000	1.000
	F4	2.627	.571	.389	4.600	.000	1.000	1.000
	F5	-2.150	.571	-.319	-3.765	.000	1.000	1.000
	F8	-1.078	.571	-.160	-1.888	.063	1.000	1.000

a. Dependent Variable: LN_TTS

Có thể nhận thấy hồi qui này hoàn toàn không có đa cộng tuyến.

Sau đây là bảng kết quả tính tác động của các biến độc lập đến LN_TTS:

Biến độc lập	Hệ số β_{ij}						Hệ số α_v	Tác động đến LN_TSS
	F1	F2	F3	F4	F5	F8		
Tuoi_dn	0.020	-0.190	0.412	-0.164	-0.072	0.003	0.880	0.603
knghiem_gd	0.059	-0.216	0.280	-0.072	-0.062	0.270	-2.302	0.487
trinhdo_gd	-0.200	-0.072	-0.144	0.080	-0.362	-0.033	1.142	0.792
KNTT_NH	0.310	0.115	0.027	0.186	-0.024	0.016	2.627	0.525
KNTT_N	0.313	0.120	0.051	0.131	-0.044	0.035	-2.150	0.428
VQ_HTK	0.054	-0.059	0.109	0.106	0.249	0.011	-1.078	0.036
KTT_BQ	0.201	0.148	0.096	-0.304	-0.093	-0.083		-0.527
HQSD_TS	-0.024	-0.124	0.096	0.249	0.456	-0.372		0.419
NPT_VCSH	-0.138	0.322	0.277	0.183	-0.020	-0.037		0.017
LN_DT	-0.061	-0.135	0.096	0.365	0.096	0.652		0.386
Tyso_N	-0.167	0.114	0.142	-0.331	0.083	0.031		-1.241
HS_donbay	-0.138	0.322	0.277	0.183	-0.020	-0.037		0.017
KN_TTLV	0.014	-0.078	-0.027	0.326	-0.456	-0.383		2.250
Quymo	-0.005	0.190	-0.331	-0.006	0.143	0.456		-1.525
VQ_KPT	-0.026	-0.020	-0.122	0.011	0.326	-0.430		-0.303

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 2: Bài toán phân tích thành phần chính trong mô hình tính toán hệ số cạnh tranh cấp tỉnh (PCI).

Trong thí dụ này, chúng tôi giới thiệu tóm tắt quá trình tính toán PCI, trong đó có việc sử dụng phân tích thành phần chính để xác định các trọng số (khách quan theo dữ liệu, không bị can thiệp bởi các yếu tố chủ quan).

1. Quy trình xử lí dữ liệu để xây dựng chỉ số tổng hợp PCI

Dữ liệu dùng trong xây dựng chỉ số PCI gồm hai loại: dữ liệu mềm và dữ liệu cứng. Trong đó

- *Dữ liệu mềm là các dữ liệu thu được từ việc việc tổng hợp các phiếu điều tra từ các doanh nghiệp dân doanh;*
- *Dữ liệu cứng bao gồm các số liệu liên quan được thu thập từ các niên giám thống kê và một số nguồn khác như ngân hàng nhà nước, các công ty kinh doanh bất động sản và các hiệp hội doanh nghiệp.*

Quá trình phân tích xử lý dữ liệu để xác định một chỉ số tổng hợp PCI gồm hai bước:

Bước 1: Chọn và tính toán ra được các tiêu chí thành phần dựa trên tập dữ liệu

Bước 2: Từ các tiêu chí thành phần này xây dựng chỉ số tổng hợp phản ánh năng lực cạnh tranh cấp tỉnh.

Năm 2005, chỉ số tổng hợp PCI ở Việt nam được tính toán dựa trên 9 chỉ số thành phần. Từ năm 2006 – 2007 việc tính toán được dựa trên 10 tiêu chí thành phần, bao gồm:

1. *Chi phí gia nhập thị trường*
2. *Tiếp cận đất đai*
3. *Tính minh bạch*
4. *Chi phí thời gian*
5. *Chi phí không chính thức*
6. *Ưu đãi doanh nghiệp nhà nước*
7. *Tính năng động*

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8. *Chính sách phát triển kinh tế khu vực kinh tế tư nhân*

9. *Đào tạo lao động*

10. *Thiết chế Pháp lý*

Bước 1: Sử dụng dữ liệu thu thập được để xây dựng và tính toán các tiêu chí thành phần, trong đó tiêu chí thành phần được tính như sau

Mỗi tiêu chí thành phần cấu thành bởi các chỉ tiêu thể hiện các khía cạnh khác nhau, và quy trình tính các tiêu chí thành phần này là như sau:

Các chỉ tiêu của mỗi tiêu chí thành phần được tính toán trực tiếp từ dữ liệu mềm, dữ liệu cứng hoặc theo phương pháp chuyên gia của nhóm nghiên cứu PCI

Sau đó các chỉ tiêu này được chuẩn hóa sao cho tỉnh được đánh giá tốt nhất theo chỉ tiêu này sẽ nhận giá trị 10, tỉnh kém nhất nhận giá trị 1. Việc chuẩn hóa được thực hiện bởi một trong hai công thức sau đây:

Công thức 1: $x_{ij}^{*} = 9(x_{ij} - \min_j)/(\max_j - \min_j) + 1$ (CT1)

Công thức 2: $x_{ij}^{*} = 11 - [9(x_{ij} - \min_j)/(\max_j - \min_j) + 1]$ (CT2)

Trong đó x_{ij} thể hiện giá trị của chỉ tiêu thứ j cho tỉnh i; min_j là giá trị nhỏ nhất trong các giá trị x_{ij} và max là giá trị lớn nhất trong các giá trị x_{ij} của cả nước. Giá trị x_{ij}* là giá trị chuẩn hóa của x_{ij}.

Bước 2: Tổng hợp các tiêu chí thành phần để tính ra tiêu chí tổng hợp PCI

Việc tổng hợp các tiêu chí thành phần để thu được một tiêu chí tổng hợp PCI được thực hiện như sau: trước hết năng lực cạnh tranh được các nhà nghiên cứu cho rằng được cấu thành từ 3 yếu tố chính sau đây: Số doanh nghiệp trên 1000 dân, đầu tư bình quân đầu người, và lợi nhuận bình quân của doanh nghiệp.

Ý tưởng để tính trọng số là: trọng số của mỗi tiêu chí thành phần phụ thuộc vào mức đóng góp của tiêu chí này lên mỗi trong 3 yếu tố nói trên. Trong đó mức đóng góp này được tính thông qua việc sử dụng hồi quy đa biến trong đó các biến độc lập chính là các tiêu chí thành phần nói trên, và

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

biến phụ thuộc lần lượt là một trong 3 yếu tố cấu thành nên năng lực cạnh tranh của tỉnh. Tuy nhiên giữa các biến độc lập này thường có tương quan cao. Do đó cần phải thực hiện thông qua các bước như sau:

Bước 1: *Sử dụng phương pháp phân tích nhân tố để phân chia 10 tiêu chí thành phần thành các nhân tố không tương quan với nhau.*

Bước 2: Tính trọng số cho các nhân tố:

Chạy 3 mô hình hồi quy đa biến với biến phụ thuộc tương ứng là logarit của các yếu tố chính nói trên: Số doanh nghiệp trên 1000 dân, đầu tư bình quân đầu người, và lợi nhuận bình quân của doanh nghiệp (số liệu của năm trước). Các biến phụ thuộc bao gồm: nhân tố 1, nhân tố 2, nhân tố 3 và các biến điều khiển thể hiện sự khác biệt ban đầu giữa các tỉnh (bao gồm:…).

Sử dụng các hệ số hồi quy để tính ra được phần trăm đóng góp của mỗi nhân tố lên mỗi biến phụ thuộc

Lấy trung bình của lượng phần trăm đóng góp này để thu được trọng số của mỗi nhân tố

Bước 3: Tính trọng số ban đầu cho mỗi tiêu chí thành phần, như sau:

Đóng góp của tiêu chí (j) = loading (j1)* trọng số F1+ loading (j2)* trọng số F2+ loading (j3)* trọng số F3

Trọng số của tiêu chí j = đóng góp của tiêu chí j/ tổng đóng góp của tất cả tiêu chí thành phần

Bước 4: Tính trọng số cho mỗi tiêu chí thành phần: làm tròn số cho các trọng số vừa tính ở bước 3 thành 3 nhóm với các trọng số tương ứng là 15%, 10% và 5%, trong đó phải đảm bảo tổng các trọng số là 100%

Bước 5: Tổng hợp PCI: trung bình có trọng số của các chỉ tiêu thành phần, với trọng số được tính ở bước 4.

Trong năm 2006- 2007, các tiêu chí được phân thành 3 nhân tố.

Tính trọng số cho các tiêu chí thành phần: Do vai trò của các tiêu chí thành phần trong việc xác định năng lực cạnh tranh là khác nhau nên mỗi tiêu chí sẽ có trọng số khác nhau trong cơ cấu PCI. Năng lực cạnh tranh, về phần mình lại được cơ cấu từ 3 yếu tố:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1. Số doanh nghiệp trên 1000 dân

2. Đầu tư bình quân đầu người

3. Lợi nhuận doanh nghiệp

Mỗi tiêu chí thành phần thường được tính toán dựa trên các chỉ tiêu thể hiện các khía cạnh khác nhau của tiêu chí thành phần. Các chỉ tiêu này được tính toán trực tiếp dựa trên các dữ liệu cứng, dữ liệu mềm hoặc theo phương pháp chuyên gia của nhóm nghiên cứu PCI.

11. SPSS VỚI PHÂN TÍCH THÀNH PHẦN CHÍNH

Để sử dụng SPSS cho phân tích thành phần chính trước tiên tất cả các biến trong tệp số liệu tham gia phân tích đều phải là các biến số hoặc nếu là các biến định tính thì nhãn của chúng phải có giá trị.

Chúng ta có thể sử dụng một tài liệu riêng bao gồm nhiều nội dung của phầm mềm này. Trong giới hạn của mục này ta chỉ xem xét cách thức và các kết quả phân tích nhân tố nhờ SPSS. Với các kết quả này chúng ta sẽ đề cập đến cách mô tả kết quả và phân tích một số đặc trưng của chúng. Để đơn giản ta sẽ xét 1 bộ số liệu nhỏ thường đã cài đặt sẵn các bộ Setup của SPSS for Windows.

Số liệu dân số kinh tế thế giới

Tệp số liệu Word95.sav bao gồm 26 chỉ tiêu dân số - kinh tế thống kê từ 109 nước trên thế giới (xem phụ lục 1). Mỗi nước tương ứng 1 quan sát hay 1 cá thể, sau đây là một vài phân tích trên cơ sở từng bộ phận của tệp số liệu này.

a. Vùng châu Á - Thái Bình Dương

Các biến được chọn để phân tích là:

1- URBAN People living in cities (%)

2- LIFEEXPF Average female life expectancy

3- LITERACY People who read (%)

4- POP_INCR Population increase (% per year))

5- BABYMORT Infant mortality (deaths per 1000 live births)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6- BIRTH_RT Birth rate per 1000 people

7- DEATH_RT Death rate per 1000 people

8- LOG_GDP Log (base 10) of GDP_CAP

9- B_TO_D Birth to death ratio

10-FERTILTY Fertility: average number of kids

11-LOG_POP Log (base 10) of Population.

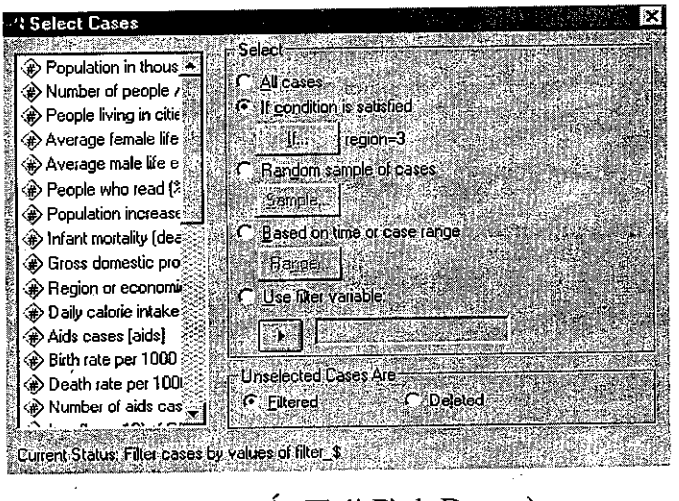
12- GDP_CAP Gross domestic product/capital

Các nước thuộc vùng này là:

1	Afghanistan	9	Malaysia
2	Bangladesh	10	N. Korea
3	Cambodia	11	Pakistan
4	China	12	Philippines
5	Hong Kong	13	S. Korea
6	India	14	Singapore
7	Indonesia	15	Thailand
8	Japan	16	Vietnam

Từ thực đơn Data chọn:

- Selecte Cases



Thủ tục này chọn vùng 3 (Châu Á- Thái Bình Dương)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

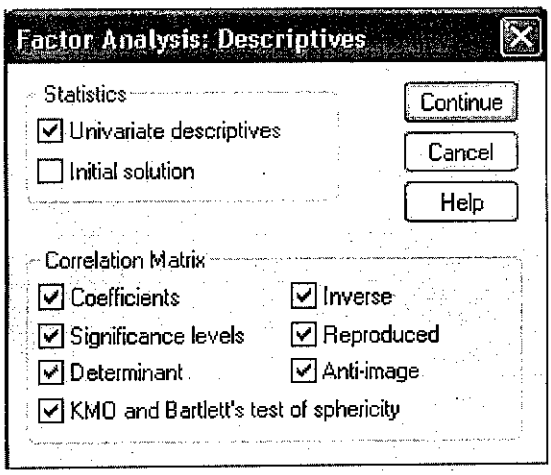
Từ thực đơn hãy chọn:

- Analyze\Data reduction\Factor

- Factor analyze

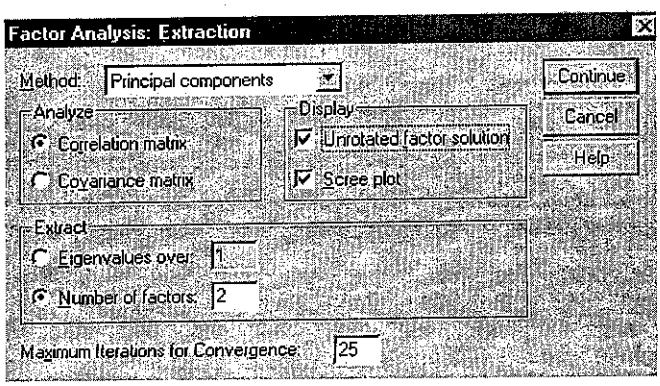
Variables: `urban lifeexpf literacy pop_incr babymort birth_rt death_rt log_gdp b_to_d fertilty log_pop.`

- Discriptives:



Chúng ta có thể lựa chọn các mô tả thống kê của các biến. Ngoài ra có thể hiển thị các đặc trưng của ma trận hệ số tương quan của các biến, nhất là trong trường hợp lựa chọn phương pháp ma trận hệ số tương quan trong việc tìm các thành phần chính (extraction). Với lựa chọn trên, chúng ta có tất cả các thông tin về ma trận hệ số tương quan; kiểm định KMO và Kaier.

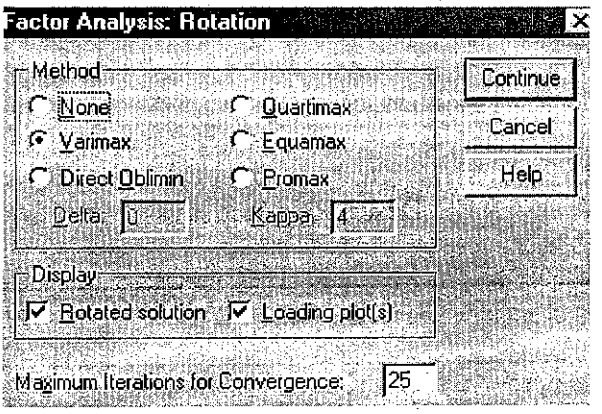
- Extraction



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

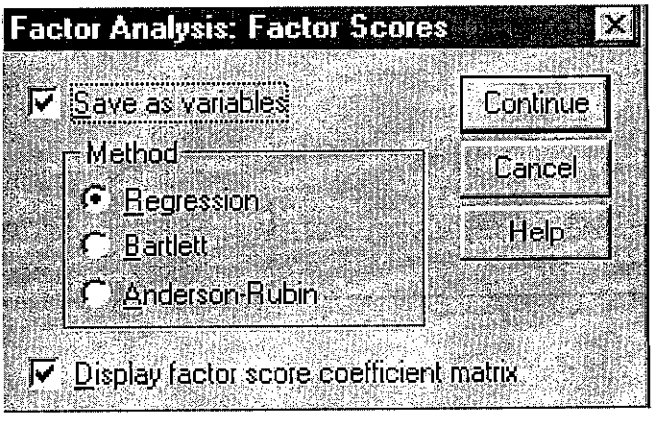
Mặc định SPSS chọn phương pháp phân tích từ ma trận hệ số tương quan. Phương pháp này phù hợp với tình trạng các biến có thang đo khác nhau. Phương pháp ma trận hiệp phương sai thường sử dụng khi chúng ta cho rằng các biến có phương sai khác nhau và chính nguyên nhân này có thể làm giảm hiệu quả phân tích. Thực tế có thể lựa chọn cả hai phương pháp và căn cứ vào tỷ lệ giải thích trên siêu phẳng chiếu để chọn kết quả sử dụng cho các phân tích tiếp theo.

- Rotation



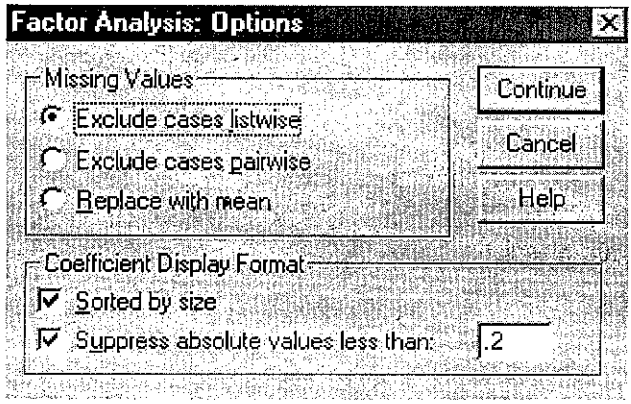
Các phép quay hệ trục trong F cho phép tăng tối đa hệ số tương quan lớn và giảm tối đa hệ số tương quan nhỏ. Thủ tục này nhằm tìm những giải thích có thể cho các trục chính trong phân tích nhờ ý nghĩa của các biến ban đầu.

- Factor Score



- Option

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Kết quả phân tích cơ bản như sau:

- Bảng mô tả thống kê

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
People living in cities (%)	43.13	28.654	16
Average female life expectancy	66.75	10.884	16
People who read (%)	71.81	25.545	16
Population increase (% per year))	1.696	.8635	16
Infant mortality (deaths per 1000 live births)	56.931	46.1738	16
Gross domestic product / capita	4088.50	6454.734	16
Birth rate per 1000 people	26.938	11.8573	16
Death rate per 1000 people	8.84	4.479	16
Log (base 10) of GDP_CAP	3.0613	.71726	16
Log (base 10) of Population	4.7481	.71584	16
Fertility: average number of kids	3.383	1.7961	16
Birth to death ratio	3.1752	1.00157	16

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Kiểm định sự phù hợp của số liệu mẫu

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.713
Bartlett's Test of Sphericity	Approx. Chi-Square	284.630
	df	66
	Sig.	.000

- Bảng tỷ lệ bảo tồn phương sai của đám mây ban đầu trên siêu phẳng chiếu

Total Variance Explained

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.530	68.457	68.457	7.121	64.735	64.735
2	1.509	13.716	82.173	1.918	17.439	82.173

Extraction Method: Principal Component Analysis.

- Ma trận các thành phần (Các véc tơ riêng) - Toạ độ các biến

Component Matrix(a)

	Component	
	1	2
People living in cities (%)	-.845	.362
Average female life expectancy	-.981	-.138
People who read (%)	-.856	-.342
Population increase (% per year))	.900	-.127
Infant mortality (deaths per 1000 live births)	.961	.213
Gross domestic product / capita	-.783	.551
Birth rate per 1000 people	.959	.157
Death rate per 1000 people	.811	.528
Log (base 10) of GDP_CAP	-.896	.312
Log (base 10) of Population	.228	-.554
Fertility: average number of kids	.935	.196
Birth to death ratio	.208	-.665

Extraction Method: Principal Component Analysis.

a 2 components extracted.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Bảng tọa độ các biến trên hai trục chính sau phép quay

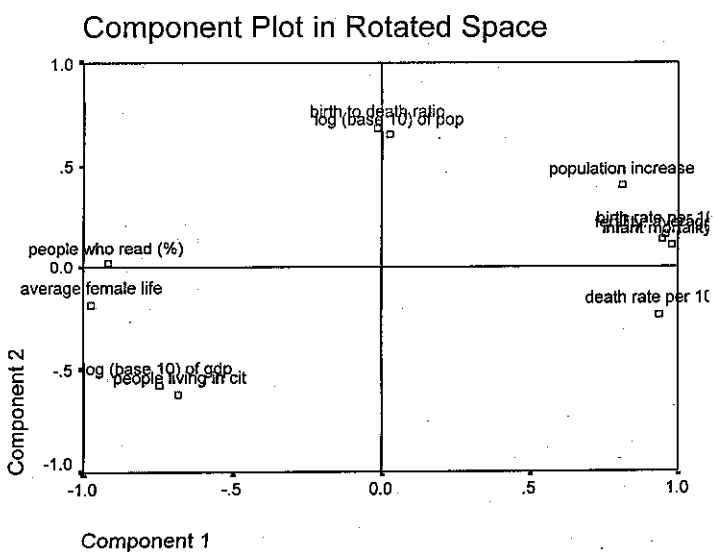
	Component	
	1	2
People living in cities (%)	-.639	.661
Average female life expectancy	-.957	.253
People who read (%)	-.921	.017
Population increase (% per year))	.781	-.466
Infant mortality (deaths per 1000 live births)	.969	-.176
Gross domestic product / capita	-.508	.811
Birth rate per 1000 people	.945	-.227
Death rate per 1000 people	.952	.173
Log (base 10) of GDP_CAP	-.705	.635
Log (base 10) of Population	-.005	-.599
Fertility: average number of kids	.937	-.182
Birth to death ratio	-.066	-.694

- Ma trận của phép quay

Component Transformation Matrix

Component	1	2
1	.922	-.387
2	.387	.922

- Biểu đồ các biến trên siêu phẳng chiếu



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Bảng tỷ lệ giải thích được của các biến nhờ hai factor

	Extraction
People living in cities (%)	.845
Average female life expectancy	.980
People who read (%)	.849
Population increase (% per year))	.827
Infant mortality (deaths per 1000 live births)	.970
Gross domestic product / capita	.916
Birth rate per 1000 people	.945
Death rate per 1000 people	.936
Log (base 10) of GDP_CAP	.900
Log (base 10) of Population	.359
Fertility: average number of kids	.912
Birth to death ratio	.485

Extraction Method: Principal Component Analysis.

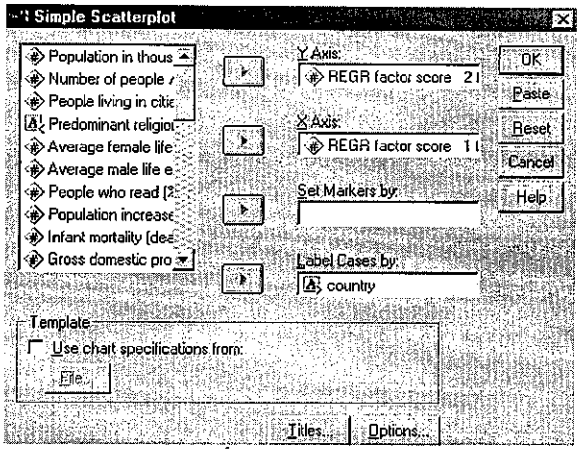
- Bảng hệ số tổ hợp của các biến (các véc tơ nhân tố chính)

	Component	
	1	2
People living in cities (%)	-.019	.224
Average female life expectancy	-.141	-.023
People who read (%)	-.170	-.132
Population increase (% per year))	.075	-.107
Infant mortality (deaths per 1000 live births)	.155	.062
Gross domestic product / capita	.028	.316
Birth rate per 1000 people	.143	.034
Death rate per 1000 people	.205	.229
Log (base 10) of GDP_CAP	-.036	.201
Log (base 10) of Population	-.092	-.291
Fertility: average number of kids	.148	.054
Birth to death ratio	-.118	-.347

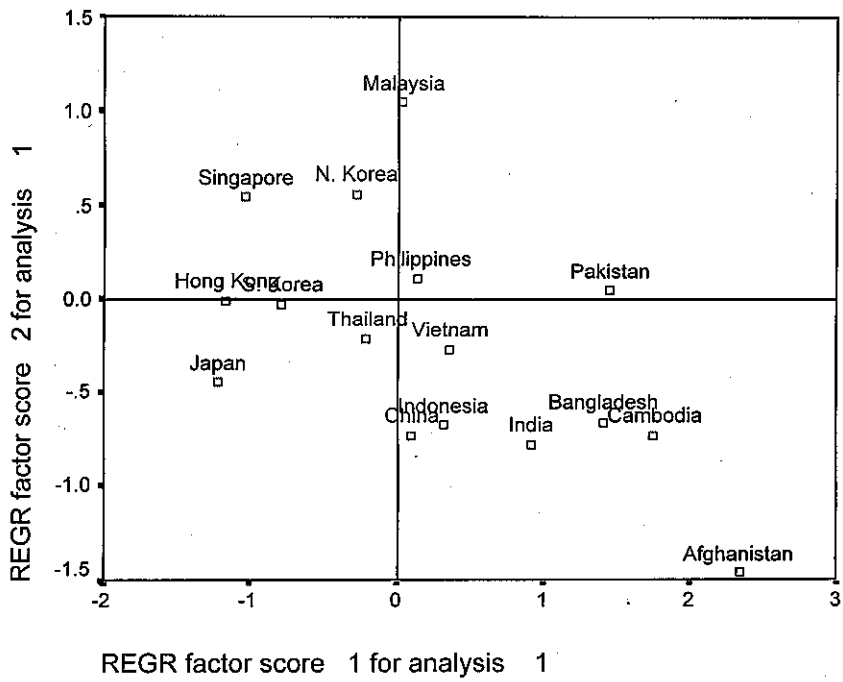
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Để có thể biểu diễn các quốc gia trên các thành phần chính, ta có thể tiến hành như sau:

Từ thực đơn Graph chọn Scatter: Scattterplot chọn Simple - Define



Kết quả ta có biểu đồ sau:



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

1- Biểu đồ **Scree Plot**

Biểu đồ này cho ta các giá trị riêng tương ứng với các thành phần chính. Có thể thấy rằng từ giá trị riêng thứ 5 trở đi sự sai khác của chúng không đáng kể. Điều này cho thấy nếu chúng ta muốn tăng số thành phần chính để tăng mức giải thích của đám mây ảnh thì cũng không nên dùng quá 4 thành phần.

2- Bảng tổng các phương sai được giải thích (**Total Variance Explained**)

Bảng này cho biết 64% quán tính của đám mây số liệu được phản ảnh bởi thành phần chính thứ nhất và 17 % được phản ánh bởi thành phần chính thứ 2. Tổng tỷ lệ của hai thành phần là 82%.

3- Bảng toạ độ các biến trên hai trục chính sau phép quay: bảng này cho cosin của góc tạo bởi các biến với các trục chính, đó chính là cơ sở lập biểu đồ các biến (trong lựa chọn Option ta chọn loại bỏ các giá trị nhỏ hơn 0.2 vì vậy một số giá trị không được in ra).

4- Biểu đồ các biến trên siêu phẳng chiếu (**Component Plot in Rotated Space**)

Biểu đồ này chỉ ra các vị trí của các biến trên mặt phẳng toạ độ với các trục là các thành phần chính. Trên biểu đồ này ta có thể thấy một số quan hệ chủ yếu sau:

- Theo trục 1: Tỷ lệ chết trên 1000 dân biến đổi ngược chiều với tỷ lệ biết đọc, biết viết trong dân cư cũng như tuổi thọ của phụ nữ.

- Theo một phương khác (tổ hợp lại hai thành phần chính): Tỷ lệ tăng dân số biến động ngược chiều với logarit của GDP trên đầu dân cư.

- Trên trục 2: Dân số và tỷ lệ chết của trẻ ngay sau khi sinh biến đổi cùng chiều.

- Các biến có thể chia thành 4 nhóm chính mà quan hệ trong mỗi nhóm tương đối chặt chẽ như sau:

Nhóm 1: Tỷ lệ chết trẻ em sau khi sinh và số dân

Nhóm 2: Học vấn và kỳ vọng sống của nữ

Nhóm 3: Thu nhập và tỷ lệ dân thành thị

Nhóm 4: Số dân - Tỷ lệ chết trẻ em - Tỷ lệ sinh - Tỷ lệ chết.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chúng ta có thể sử dụng các phân tích cận biên để đánh giá sự biến động của các biến ảnh hưởng như thế nào đến vị trí của các cá thể. Phần tiếp theo cho chúng ta mối liên hệ này.

5- Bảng hệ số tổ hợp của các biến (các véc tơ nhân tố chính): Bảng này cung cấp hai véc tơ nhân tố chính với các toạ độ là các hệ số tổ hợp của các biến ban đầu. Mỗi hệ số cho biết khi biến tương ứng thay đổi một đơn vị thì mỗi thành phần chính thay đổi bao nhiêu đơn vị.

6- Biểu đồ các quốc gia: Biểu đồ này thể hiện vị trí trên mặt phẳng chiếu với các trục tương tự như trên. Các phân tích như với biểu đồ các biến. Tuy nhiên điều nổi bật ở đây là có thể chia một cách sơ bộ các nước của khu vực này thành 3 nhóm:

Nhóm 1: Gồm một số nước mà đặc trưng kinh tế dân số gần giống Việt Nam và Indonesia.

Nhóm 2: Gồm các nước giàu nhất của khối như Nam Triều Tiên, Nhật, Hồng Kông và Singapore.

Nhóm 3: Gồm hai nước Campuchia và Afghanistan.

b. Các nước Châu Phi

Với các biến trên phân tích nhân tố bằng phương pháp thành phần chính đối với khu vực Châu Phi người ta nhận được các kết quả trang sau. Trong các kết quả này ngoài các thông tin như với khu vực châu Á - Thái Bình Dương chúng ta quan tâm thêm một số thông tin khác. Chẳng hạn như các hình chiếu của các biến trên các trục chính; các hệ số góc của các phép chiếu đối với các biến (CO2).

Với 19 nước Châu Phi, nếu chỉ dùng hai thành phần chính chúng ta thu được đám mây ảnh giải thích được 72% đặc tính của số liệu ban đầu. Tuy nhiên nếu căn cứ vào giá trị của các Score trên biểu đồ thì người ta có thể dùng tới 6 thành phần chính để mô tả các số liệu ban đầu.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
URBAN People living in cities (%)	29.26	15.06	19
LIFEEXPF Average female life expectancy	54.26	7.98	19
LITERACY People who read (%)	47.26	17.86	19
POP_INCR Population increase (% per year))	2.723	.456	19
BABYMORT Infant mortality (deaths per 1000 live birt	94.179	28.653	19
GDP_CAP Gross domestic product / capita	998.68	1178.26	19
BIRTH_RT Birth rate per 1000 people	42.000	6.412	19
DEATH_RT Death rate per 1000 people	14.74	5.05	19
LOG_GDP Log (base 10) of GDP_CAP	2.7719	.4345	19
B_TO_D Birth to death ratio	3.1020	.8407	19
FERTILTY Fertility: average number of kids	6.081	1.134	19
LOG_POP Log (base 10) of Population	3.9898	.5709	19

- Giá trị riêng và tỷ lệ quán tính được phản ánh trên siêu phẳng chiếu:

Total Variance Explained

	Initial Eigenvalues			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.387	53.224	53.224	5.672	47.264	47.264
2	2.344	19.535	72.759	3.059	25.495	72.759
3	1.212	10.097	82.856			
4	.759	6.324	89.180			
5	.610	5.085	94.265			
6	.431	3.594	97.858			
7	9.588E-02	.799	98.657			
8	7.132E-02	.594	99.252			
9	5.582E-02	.465	99.717			
10	2.466E-02	.206	99.922			
11	6.871E-03	5.726E-02	99.980			
12	2.432E-03	2.026E-02	100.000			

Extraction Method: Principal Component Analysis.

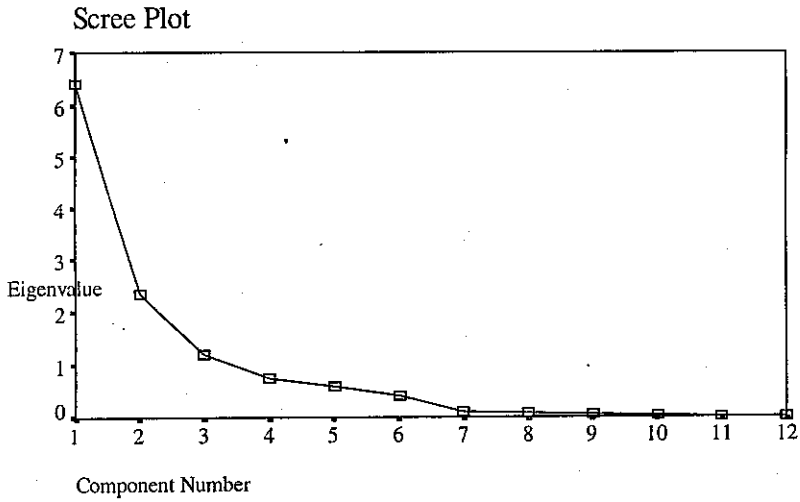
Rotated Component Matrix [a]

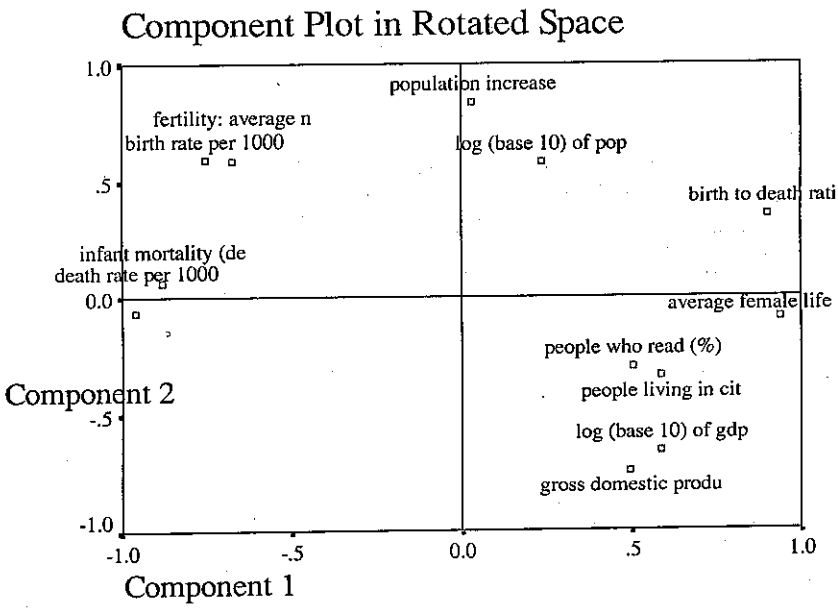
	Component	
	1	2
DEATH_RT Death rate per 1000 people	-.961	-6.30E-02
LIFEEXPF Average female life expectancy	.940	-9.32E-02
B_TO_D Birth to death ratio	.904	.352
BABYMORT Infant mortality (deaths per 1000 live births)	-.883	6.245E-02
BIRTH_RT Birth rate per 1000 people	-.755	.595
FERTILTY Fertility: average number of kids	-.676	.586
URBAN People living in cities (%)	.586	-.334
LITERACY People who read (%)	.504	-.296
POP_INCR Population increase (% per year))	3.071E-02	.835
GDP_CAP Gross domestic product / capita	.493	-.742
LOG_GDP Log (base 10) of GDP_CAP	.587	-.657
LOG_POP Log (base 10) of Population	.235	.585

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Một kết quả rất đáng chú ý ở đây là tỷ lệ tăng dân số và số dân hầu như không tương quan với tuổi thọ trung bình của nữ. Trong khi tỷ lệ chết sơ sinh và tuổi thọ trung bình của nữ biến đổi hoàn toàn ngược chiều nhau.





THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Component Score Coefficient Matrix

	Component	
	1	2
URBAN People living in cities (%)	.085	-.066
LIFEEXPF Average female life expectancy	.183	.062
LITERACY People who read (%)	.072	-.060
POP_INCR Population increase (% per year))	.092	.320
BABYMORT Infant mortality (deaths per 1000 live births)	-.174	-.067
GDP_CAP Gross domestic product / capita	.024	-.230
BIRTH_RT Birth rate per 1000 people	-.093	.148
DEATH_RT Death rate per 1000 people	-.203	-.123
LOG_GDP Log (base 10) of GDP_CAP	.052	-.188
B_TO_D Birth to death ratio	.221	.227
FERTILTY Fertility: average number of kids	-.078	.152
LOG_POP Log (base 10) of Population	.108	.246

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Component Scores.

MỘT SỐ PHÂN TÍCH

+ Bảng **Total Variance Explained**

Kết quả trên liệt kê toàn bộ các kết quả về tỷ lệ giải thích đối với bộ số liệu ban đầu, trên 12 trục chính ứng với 12 giá trị riêng của VM.

- Khi chọn đủ 12 trục (Hay lấy chính các biến ban đầu làm trục chính): Trục thứ nhất giải thích được 53,224%; trục thứ hai: 19,535%...

- Khi chọn 2 trục chính tổng tỷ lệ giải thích vẫn là 72,759% nhưng đóng góp của các trục có thay đổi đôi chút. Lý do thay đổi là do khi tiến hành các tổ hợp tuyến tính để xác định các trục chính các phép chiếu tương ứng đã làm thay đổi phương sai trên từng trục của đám mây số liệu ban đầu. Ngoài ra một phép quay trục toạ độ đối với các trục chính đã được thực hiện (phép quay được trình bày ở phần sau).

+ Biểu đồ **Scree Plot**

Biểu đồ này mô tả các giá trị trong cột Total của bảng **Total Variance Explained.** Tiêu chuẩn Kaiser được sử dụng dựa trên biểu đồ này. Diện tích phía dưới đường gấp khúc cho biết mức giải thích về số thành phần chính được chọn.

+ Bảng **Rotation Component Matrix**

Trong phần này, ta nhận được toạ độ của các biến theo các thành phần

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

chính. Các toạ độ này xác định vị trí của các biến trên biểu đồ tiếp theo. Đây cũng là kết quả cuối cùng của phép chiếu mà chúng ta đã xét ở trên.

+ Bảng **Component Score Coefficient Matrix và Toạ độ các cá thể**

Bảng này cho ta các hệ số cấu thành hệ số nhân tố của các cá thể, trong thủ tục này các hệ số này ký hiệu là fac1_1 và fac1_2.

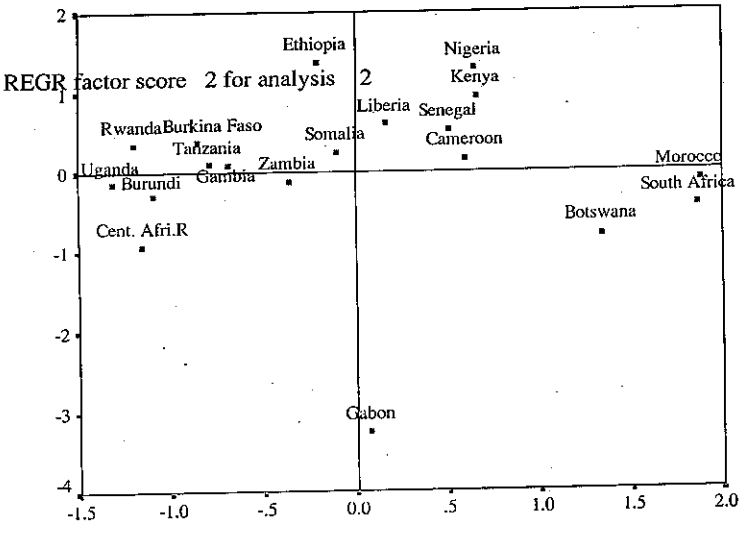
Các biến mới này có thể dùng phân tích như các số liệu vì các giá trị tính được như một tổi hợp tuyến tính của các biến ban đầu, cụ thể là: Nếu gọi f1(j,k) là giá trị fac1_1 của cá thể k tại thành phần chính thứ j thì $f1(j,k) = \sum_{i=1}^{r} w_{ji} z_{ik}$ trong đó w_{ji} nhận được từ bảng trên còn z_{ik} là giá trị X_{jk} đã chuẩn hoá. Các hệ số fac1_1 và fac1_2 được SPSS tính sẵn và ghi lại như các biến mới. Nó chỉ ra các toạ độ của các cá thể theo các thành phần chính đã tìm được. Bảng sau cho kết quả SPSS đã tính.

Toạ độ các quốc gia theo các thành phần chính

Quốc gia	Component	
	1	2
Botswana	1.3428	-0.79851
Burkina Faso	-0.85303	0.37748
Burundi	-1.08908	-0.30687
Cameroon	0.59813	0.15898
Cent. Afri.R	-1.1543	-0.92969
Ethiopia	-0.21015	1.38721
Gabon	0.07382	-3.25342
Gambia	-0.68856	0.09208
Kenya	0.65957	0.9511
Liberia	0.16611	0.61068
Morocco	1.88096	-0.12514
Nigeria	0.64699	1.31116
Rwanda	-1.19631	0.35745
Senegal	0.51336	0.52697
Somalia	-0.09863	0.24426
South Africa	1.86378	-0.44297
Tanzania	-0.78712	0.101
Uganda	-1.31412	-0.1446
Zambia	-0.3542	-0.11718

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với các số liệu này phân bố các quốc gia trên mặt phẳng chiếu như sau:



Với biểu đồ trên, ta thấy các nước Châu Phi tạm chia thành 3 nhóm: Nhóm 1 với đa số nước, nhóm 2 với Maroc, Nam Phi và Bosoana; nhóm 3 chỉ có một nước là Gabon.

c. Tất cả các quốc gia

Trong phần này, chúng ta sẽ trình bày chi tiết hơn một số vấn đề cần chú ý trong phân tích thành phần chính. Với số liệu của 109 quốc gia thuộc 6 khu vực trên thế giới. Những phân tích cần thiết trước khi phân tích nhân tố cũng sẽ được đề cập kỹ hơn đôi chút. Đương nhiên với sự hỗ trợ của phầm mềm SPSS hoặc các phần mềm khác tương đương.

Trước hết chúng ta hãy xem lại kết quả phân tích thành phần chính với 12 biến như các phần a và b.

FactorAnalysis

Chọn Rotation (Một phép quay) nên trong bảng tổng bình phương các tỷ lệ đóng góp vào việc giải thích cho các biến, các cá thể được mô tả bởi hai phần. Phần tỷ lệ đóng góp giả thích khi chưa thực hiện phép quay và tỷ lệ này khi đã thực hiện phép quay.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.316	60.965	60.965	7.316	60.965	60.965	7.291	60.759	60.759
2	2.121	17.673	78.638	2.121	17.673	78.638	2.146	17.879	78.638
3	1.119	9.329	87.967						
4	.681	5.671	93.638						
5	.286	2.385	96.022						
6	.172	1.437	97.459						
7	.096	.798	98.257						
8	.072	.603	98.860						
9	.062	.520	99.380						
10	.036	.298	99.678						
11	.025	.210	99.888						
12	.013	.112	100.00						

Extraction Method: Principal Component Analysis.

Component Matrix [a]

	Component	
	1	2
BIRTH_RT Birth rate per 1000 people	.956	.185
BABYMORT Infant mortality (deaths per 1000 live births)	.946	-.208
LIFEEXPF Average female life expectancy	-.945	.255
FERTILTY Fertility: average number of kids	.929	.159
LITERACY People who read (%)	-.911	
LOG_GDP Log (base 10) of GDP_CAP	-.893	
URBAN People living in cities (%)	-.779	.320
GDP_CAP Gross domestic product / capita	-.759	
POP_INCR Population increase (% per year))	.759	.600
B_TO_D Birth to death ratio	.329	.916
DEATH_RT Death rate per 1000 people	.502	-.754
LOG_POP Log (base 10) of Population		-.250

Extraction Method: Principal Component Analysis.

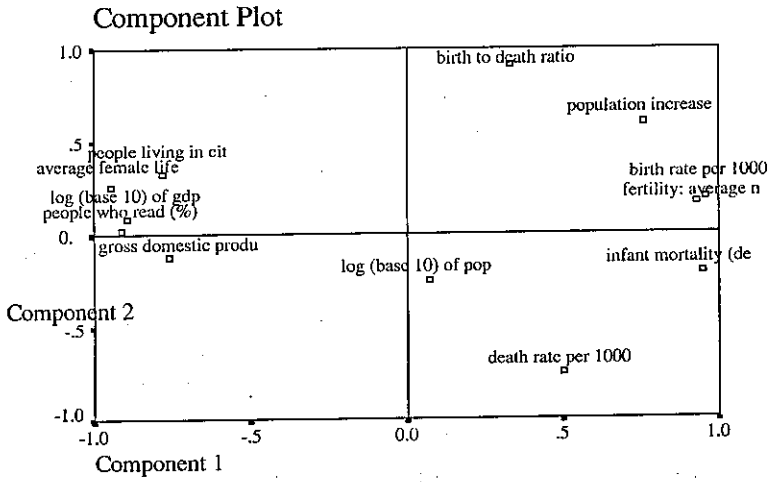
a. 2 components extracted.

Trong thí dụ này hai tỷ lệ không khác nhau đáng kể (60,985 và 60,759 cho thành phần chính thứ nhất, 17,673 và 17,879 cho thành phần chính thứ hai. Khi sử dụng phần mềm SPSS để phân tích thành phần chính, hoàn toàn như lý thuyết đã trình bày ở phần trước, trước tiên ta nhận được kết quả là các véc tơ thành phần chính, các nhân tố chính. Tuy nhiên do chúng ta không chọn tất cả p thành phần chính (có thể nói là không bao giờ), mà chỉ

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

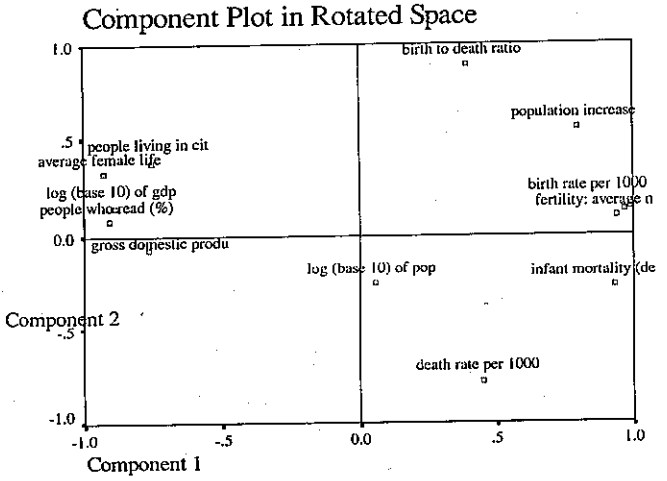
chọn 2 hoặc 3 thành phần có khả năng giải thích cao nhất nên các biến cũng như các cá thể có thể bị nhóm thành các nhóm không thật gần các trục toạ độ tương ứng. Để có thể dễ dàng hơn trong phân tích, SPSS thực hiện một thuật toán quay hệ trục sao cho các biến có toạ độ lớn càng lớn hơn, đồng thời các biến có toạ độ nhỏ càng nhỏ hơn. Sự khác nhau có thể thấy qua các bảng **Component Matrix** và **Rotation Component Matrix**:

Biểu đồ các biến khi chưa quay



Với thí dụ này, ta có thể thấy phép quay gần như không làm thay đổi đáng kể vị trí (theo các trục) của các biến - (xem các biểu đồ sau).

Biểu đồ các biến khi đã quay



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Biểu đồ thứ nhất thiết lập trên cơ sở số liệu của bảng toạ độ các biến ban đầu theo từng thành phần chính. Do mục đích của mô tả các đặc trưng của đám mây số liệu trên không gian r chiều là các trục không trực tiếp mang ý nghĩa kinh tế xã hội, người ta tìm cách tăng các toạ độ lớn nếu có thể và giảm các toạ độ nhỏ tương ứng. Kết quả mà người ta mong đợi trong phép quay này là tìm được một cách nào đó giải thích ý nghĩa kinh tế- xã hội cho các thành phần chính.

Trong nhiều trường hợp người ta có thể không chọn phép quay như trên, nếu như các đám mây ảnh đã khá tập trung ở từng nhóm khác nhau. Tuy nhiên, điều đó chỉ biết sau phân tích nên chúng ta có thể chọn ngay các phép quay từ lần phân tích đầu tiên. Những phân tích chi tiết cũng có thể sử dụng cả hai kết quả, vì với kết quả ban đầu (chưa thực hiện phép quay) các biến có vị trí thực trong hệ trục mới, các vị trí này là các vị trí tuyệt đối chính xác theo các trục chính và các thành phần chính; mặt khác nếu muốn giải thích ý nghĩa các thành phần chính hay các trục chính thì người ta cần tìm một số biến có vị trí ngay sát hoặc nằm ngay trên các trục. Với phép quay thì vị trí tương đối của các biến không đổi, điều đó cho thấy kết quả sau khi quay hoàn toàn có thể dùng để phân tích các mối quan hệ của các biến (mục tiêu chủ yếu của phân tích nhân tố).

Như đã trình bày ở phần b, mỗi cá thể có thêm hai giá trị chính là toạ độ của chúng theo mỗi nhân tố. Điều đó cho phép lập một biểu đồ đối với các cá thể này theo hai nhân tố tìm được.

12. VAI TRÒ CỦA CÁC BIẾN VÀ CÁC CÁCH THỨC LỰA CHỌN BIẾN

Chúng ta thấy trong các thí dụ trên, các biến được đưa vào phân tích thành phần chính hình như có vai trò rất khác nhau. Như đã nói ở các bài trước, nến tất cảc các biến ban đầu hoàn toàn độc lập với nhau, thì mỗi cá thể đã được mô tả bởi một điểm trong không gian p chiều nhờ p toạ độ vuông góc. Không có một biến nào mang thông tin của biến khác, phân tích nhân tố coi như không có tác dụng gì. Vậy trước khi dùng một số biến để phân tích ta cần có những nội dung tiền phân tích như thế nào? đó là nội dung chúng ta sẽ đề cập trong phần này. Sau đó là những vấn đề có tính chất chi tiết khi lựa chọn các độ đo trong phân tích, chọn lọc các kết quả... . Sau đây chúng ta sẽ nêu tóm tắt các bước thông thường được sử dụng cho một

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

quá trình phân tích nhân tố nói chung và phân tích thành phần chính nói riêng. Các dẫn chứng cần thiết sẽ sử dụng đối với tệp số liệu Word95 và thí dụ cuối cùng sẽ thực hiện với tệp số liệu DHS97 (Điều tra nhân khẩu học và sức khoẻ ở Việt Nam 1997).

12.1. Xác định mục đích phân tích và chọn biến ban đầu

Xác định mục đích phân tích là nội dung đầu tiên của mọi quá trình phân tích thống kê. Nó quyết định toàn bộ nội dung, các bước phân tích cũng như độ phức tạp của quá trình phân tích. Trong các thí dụ trên, chúng ta đã áp đặt nội dung phân tích là: Phân tích mối quan hệ của các biến dân số kinh tế đối với 109 nước ở 6 khu vực khác nhau trên thế giới. Vậy mục đích là gì?- Với các kết quả trên, chúng ta thấy các phân tích cho chúng ta hình ảnh của các biến ảnh hưởng hay không ảnh hưởng đến nhau trong quá trình vận động. Trong các biến như vậy sẽ có những biến là biến cần giải thích và có những biến là biến giải thích, trong phân tích này không có sự phân biệt theo kiểu của kinh tế lượng hay mô hình toán. Tuy vậy, nhờ những phân tích này chúng ta thấy quan hệ tác động qua lại của các biến trên các cá thể quan sát được. Như vậy ngay từ đầu chúng ta phải làm rõ các vấn đề sau:

- Bản chất các chỉ tiêu (các biến) kinh tế-xã hội và môi trường tồn tại chúng (đặc điểm các cá thể).

- Một số vấn đề - biểu hiện bởi các biến cần xem xét, thông thường đây là các biến có tính phụ thuộc trong các biến đã có.

- Nhóm các biến có liên quan với nhau và liên quan tới biến đang được quan tâm. Việc lựa chọn như vậy trước hết dựa trên cảm tính, kinh nghiệm hay sử dụng các phương pháp chuyên gia... .

- Những ý niệm tiền định về các liên hệ ngược.

12.2. Phân tích tương quan và lọc biến ban đầu

Có thể áp dụng phân tích tương quan để xác định sơ bộ các biến cần loại khỏi phân tích nhân tố. Các biến cần loại khỏi phân tích nhân tố là các biến hầu như không tương quan với các biến khác, thủ tục phân tích tương quan của SPSS cho phép chúng ta làm công việc này dễ dàng ngay trong phân tích nhân tố hoặc trong thủ tục phân tích tương quan riêng biệt. Chẳng hạn, trong các thí dụ trên với mức ý nghĩa 5%, biến LOG_POP hầu như không tương quan với các biến khác (xem phụ lục 1), vậy có thể loại biến này khỏi các biến sẽ phân tích.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

12.3. Lựa chọn mô hình phân tích

Trong phương pháp thành phần chính thường có hai mô hình. Thực chất là hai Metric khác nhau xác định khoảng cách, góc trong không gian tuyến tính.

- Phương pháp ma trận hệ số tương quan (Corelation Matrix): Phương pháp này phân tích thành phần chính dựa trên Metrix khi bình phương. Đây là phương pháp thông dụng nhất, lý do đơn giản là khi chiếu các cá thể từ R^P xuống R^r, chúng ta dựa trên tiêu chuẩn tối đa hoá sự khác biệt của các biến cũng như các cá thể. Độ đo Khi bình phương vừa trung tâm hoá các cá thể vừa chuẩn hoá chúng vì vậy sự khác biệt ở đây bao hàm cả khác biệt tuyệt đối và tương đối. Trung tâm hoá cho sự khác biệt tuyệt đối còn chuẩn hoán cho sự khác biệt tương đối (không thứ nguyên).

- Phương pháp ma trận hiệp phương sai (Covariances Matrix): Phương pháp này ít được sử dụng hơn, đối tượng của phương pháp này thường là các phân tích trong đó các biến có cùng một nội dung kinh tế xã hội hay kinh tế kỹ thuật, nhưng được quan sát hay được đo bởi các tiêu chuẩn, thước đo khác nhau. Trong trường hợp này quán tính của đám mây số liệu bằng số biến P. Chẳng hạn, khi đánh giá kết cấu của một loại vật liệu mới, người ta có thể thống nhất các chỉ tiêu nhưng không thống nhất các chỉ số qui đổi, các hệ số (thông thường đã được tính sẵn). Trường hợp này cần xác định một hay một vài cách đo chính-cách đo giải thích được nhiều nhất sự mất ổn định về các chỉ số đặc trưng cho loại vật liệu này, trong điều kiện cụ thể nào đó. Phương pháp ma trận hiệp phương sai có thể đáp ứng tốt yêu cầu này.

12.4. Phân tích nhân tố bằng phương pháp thành phần chính

1. Phân tích tương quan và chọn biến

Một cách thông thường chúng ta không có ngay những biến thực sự có ý nghĩa trong mỗi phân tích, vì vậy sau khi phân tích sơ bộ ma trận X bằng phân tích tương quan chúng ta có thể tìm ra một số biến ban đầu dùng cho phân tích.

Thí dụ: Với tệp DHS97 một nhà nhân khẩu học-xã hội (NKHXH) qua tâm đến mức sinh của các bà mẹ (TFR), nhưng cuộc điều tra này chỉ có được các số liệu đối với phụ nữ đã có chồng, tuy nhiên một cách tự nhiên nhà nhân khẩu học - xã hội này có thể có được những suy đoán có lý về TFR nếu biết được các nhân tố tác động đến số con của phụ nữ có chồng. Nhà NKHXH có

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

thể nghĩ đến rất nhiều yếu tố tác động đến số con của một phụ nữ có chồng như: tuổi, tôn giáo, học vấn, học vấn của chồng, thu nhập, tác động của các hình thức truyền thông về kế hoạch hoá gia đình,... Tiến hành thủ tục phân tích tương quan của các biến ban đầu, ta nhận được bảng hệ số tương quan cùng mức ý nghĩa của kiểm định (trang sau).

Như vậy nếu chỉ quan tâm đến các nhân tố tác động đến **V201** (Total children ever born) thì nhà NKHXH có thể bỏ đi các biến có mức ý nghĩa lớn hơn mức cho phép, chẳng hạn là 5%. Các biến HH40_60 (40-60% of Households); HH60_80 (60-80% of Households); WIFE (Wife Decides HH Financial Decisions) có thể loại khỏi phân tích này. Đương nhiên nếu chúng ta quan tâm chủ yếu đến biến khác chẳng hạn biến học vấn (V133 Education in single years) thì việc chọn biến ban đầu có thể sẽ khác đi đôi chút, xem phụ lục 2.

Nhưng cần chú ý rằng một phân tích nhân tố thường không chỉ chú ý đến một hay hai chỉ tiêu nào đó, mà mục tiêu còn là vẽ lại bức tranh tổng thể. Hơn thế nữa phân tích tương quan như trên chỉ kiểm định tương quan cặp mà không xét đến tương quan có điều kiện hay quan hệ tổng hợp của các biến. Phân tích nhân tố khắc phục yếu điểm này trong phân tích tương quan (bảng sau).

Bảng hệ số tương quan của các biến với V201

	V201
V012 Current age - respondent	0.6042
V102 Type of place of residence	0.1733
V130 Religion	0.0744
V131 Ethnicity	0.0972
V133 Education in single years	-0.3000
V384A Heard FP on radio last months	-0.0485
V384B Heard FP on TV last months	-0.1172
V384C Heard FP newspaper last months	-0.1346
V384D Heard FP on poster last months	-0.1007
V384E Heard FP brochures last months	-0.0558
V394 Visited health facil. last 12m	-0.0980

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

URBAIN Khu vuc (TT- NT)	-0.1733
REGION Region Recoded	-0.0539
HHLOW20 Poorest 20% of Households	0.0950
HH20_40 20-40% of Households	0.0453
HH40_60 40-60% of Households	0.0031
HH60_80 60-80% of Households	-0.0208
HHHIGH20 Richest 20% of Households	-0.1153
WIFE Wife Decides HH Financial Decisions	-0.0067

2. Phân tích sơ bộ

Thủ tục phân tích không có gì đặc biệt so với các thí dụ trên, phần đáng chú ý ở đây là cách phân tích kết quả với một tệp số liệu quá lớn (5664 quan sát). Trước tiên ta hãy lần lượt xem các kết quả chính mà SPSS đã cung cấp.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.246	21.642	21.642	3.246	21.642	21.642	3.135	20.901	20.901
2	1.694	11.296	32.938	1.694	11.296	32.938	1.806	12.037	32.938
3	1.418	9.450	42.388						
4	1.329	8.860	51.248						
5	1.088	7.252	58.500						
6	1.048	6.986	65.486						
7	.942	6.281	71.767						
8	.895	5.966	77.733						
9	.665	4.433	82.166						
10	.613	4.084	86.250						
11	.502	3.349	89.599						
12	.458	3.054	92.653						
13	.420	2.801	95.454						
14	.396	2.640	98.094						
15	.286	1.906	100.000						

Extraction Method: Principal Component Analysis.

Như vậy với 2 thành phần chính tổng tỷ lệ giải thích chỉ ở mức 32,938%, một tỷ lệ nhỏ mà người phân tích hầu như không hài lòng. Làm thế nào để có được một mức giải thích cao hơn, rõ ràng là chỉ có 2 cách:

Thứ nhất là tăng số thành phần chính hãy để ý đến việc tăng số thành

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

phần chính và kết quả ở bảng sau:

Số TP chính tăng	1	2	3	4
Tỷ lệ giải thích	42.388	51.248	58.500	65.486

Như vậy để có 65,486 % khác biệt ban đầu được giải thích bởi các thành phần chính, ta cần đến 6 thành phần chính?!.

Thứ hai là có thể trở lại xem xét việc bỏ bớt biến khi phân tích. Có hai tình trạng của một số biến có thể và không nên để trong một phân tích nhân tố hay phân tích thành phần chính. Đó là, các biến lập mà mình nó tạo nên một nhóm gần như độc lập với các biến còn lại; các biến quan hệ quá chặt chẽ với một số biến trong nhóm nào đó, với các biến loại thứ hai này cần có những kiểm định khác đầy đủ hơn chẳng hạn phân tích hồi quy, tương quan...

Component Matrix[a]

	Component	
	1	2
V133 Education in single years	.734	-.126
V384C Heard FP newspaper last months	.652	6.437E-02
HHHIGH20 Richest 20% of Households	.611	.206
V384B Heard FP on TV last months	.606	6.481E-02
URBAIN Khu vuc (TT- NT)	.573	.108
HHLOW20 Poorest 20% of Households	-.564	-7.72E-02
V384D Heard FP on poster last months	.535	2.272E-02
V384A Heard FP on radio last months	.475	9.639E-02
V384E Heard FP brochures last months	.406	1.391E-02
HH20_40 20-40% of Households	-.198	-9.99E-02
V131 Ethnicity	-.169	9.328E-02
V130 Religion	-.148	.102
V012 Current age - respondent	4.542E-02	.899
V201 Total children ever born	-.331	.820
V394 Visited health facil. last 12m	.139	-.302

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

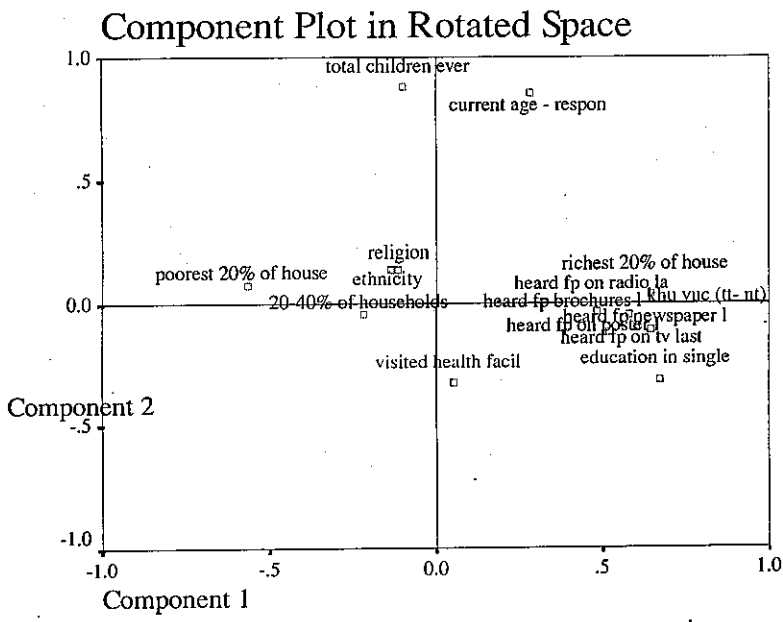
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Rotated Component Matrix

	Component	
	1	2
V133 Education in single years	.673	-.318
V384C Heard FP newspaper last months	.645	-.112
HHHIGH20 Richest 20% of Households	.644	3.516E-02
V384B Heard FP on TV last months	.602	-9.985E-02
URBAIN Khu vuc (TT- NT)	.581	-4.925E-02
HHLOW20 Poorest 20% of Households	-.564	7.658E-02
V384D Heard FP on poster last months	.521	-.121
V384A Heard FP on radio last months	.483	-3.413E-02
V384E Heard FP brochures last months	.395	-9.527E-02
HH20_40 20-40% of Households	-.217	-4.333E-02
V131 Ethnicity	-.138	.135
V201 Total children ever born	-9.939E-02	.878
V012 Current age - respondent	.284	.854
V394 Visited health facil. last 12m	5.358E-02	-.328
V130 Religion	-.115	.138

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Component Score Coefficient Matrix

	Component	
	1	2
V012 Current age - respondent	.155	.507
V130 Religion	-.028	.070
V131 Ethnicity	-.035	.067
V133 Education in single years	.198	-.132
V201 Total children ever born	.031	.493
V384A Heard FP on radio last months	.156	.016
V384B Heard FP on TV last months	.190	-.013
V384C Heard FP newspaper last months	.204	-.017
V384D Heard FP on poster last months	.162	-.031
V384E Heard FP brochures last months	.123	-.026
V394 Visited health facil. last 12m	-.006	-.183
HHLOW20 Poorest 20% of Households	-.180	.003
HH20_40 20-40% of Households	-.075	-.041
HHHIGH20 Richest 20% of Households	.214	.067
URBAIN Khu vuc (TT- NT)	.187	.014

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Component Scores.

3. Các phương án điều chỉnh

Điều chỉnh là một quá trình công phu mà kết quả của việc phân tích không phải lúc nào cũng phản ánh đúng ý muốn của người phân tích. Chất lượng điều chỉnh phụ thuộc rất lớn nếu không muốn nói là duy nhất vào hiểu biết của người phân tích. Một chuyên gia phân tích thống kê có thể sai lầm ngày càng nhiều hơn nếu chuyên gia này không đủ kiến thức xã hội, không am hiểu hoàn cảnh, điều kiện của các cá thể đã được quan sát. Chẳng hạn với tệp số liệu trên ngay việc chia mức sống thành 5 cấp theo khoảng cách 20% có thể là không cần thiết đối với một phân tích cụ thể mà chúng ta đang quan tâm, một cơ sở dữ liệu cố định có thể không dùng được cho nhiều mục đích khác nhau. Sau đây chúng ta sẽ quan sát và phân tích sơ bộ một vài kết quả hiệu chỉnh.

- Tăng số thành phần chính

Với các biến nêu trên ta tăng số thành phần chính (r = 3). Với hai mục đích, một là hy vọng có được tỷ lệ giải thích cao hơn cho tình trạng của đám mây số liệu này; hai là mô tả một số kết quả phân tích với nhiều thành phần chính hơn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Vì một số kết quả đã được giải thích ở các thí dụ trước, trong thí dụ này ta chỉ liệt kê những kết quả cần thiết tối thiểu. Trên những kết quả như vậy ta sẽ làm một số phân tích sâu hơn, làm cơ sở cho việc điều chỉnh ở mục b (kết quả đầy đủ xem phụ lục số 4).

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.25	21.642	21.642	3.25	21.642	21.642	2.41	16.072	16.072
2	1.69	11.296	32.938	1.69	11.296	32.938	2.20	14.642	30.714
3	1.42	9.450	42.388	1.42	9.450	42.388	1.75	11.674	42.388
4	1.33	8.860	51.248						
5	1.09	7.252	58.500						
6	1.05	6.986	65.486						
7	.942	6.281	71.767						
8	.895	5.966	77.733						
9	.665	4.433	82.166						
10	.613	4.084	86.250						
11	.502	3.349	89.599						
12	.458	3.054	92.653						
13	.420	2.801	95.454						
14	.396	2.640	98.094						
15	.286	1.906	100.00						

Extraction Method: Principal Component Analysis.

Nhìn vào bảng Total Variances Explained ta thấy với 3 thành phần chính tỷ lệ giải thích được là 42,388%. Ngoài ra có những nhóm biến có toạ độ trên cùng một trục rất gần nhau, nói cách khác là có những nhóm biến quan hệ tương quan rất chặt. Để thấy rõ hơn bản chất các quan hệ như vậy ta có thể xem các biểu đồ hình chiếu trên 3 mặt phẳng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Rotated Component Matrix[a]

	Component		
	1	2	3
HHHIGH20 Richest 20% of Households	.743	.122	7.618E-02
URBAIN Khu vuc (TT- NT)	.726	5.650E-02	-1.38E-02
HH20_40 20-40% of Households	-.581	.338	-4.55E-02
V384C Heard FP newspaper last months	.560	.350	-6.41E-02
V384D Heard FP on poster last months	.473	.263	-8.27E-02
V384E Heard FP brochures last months	.362	.196	-6.62E-02
HHLOW20 Poorest 20% of Households	-.117	-.728	2.142E-02
V384B Heard FP on TV last months	.191	.703	-4.33E-02
V384A Heard FP on radio last months	7.760E-02	.645	1.373E-02
V133 Education in single years	.448	.544	-.262
V130 Religion	-2.26E-02	-.164	.126
V012 Current age - respondent	.132	.178	.875
V201 Total children ever born	-.197	-2.99E-02	.871
V394 Visited health facil. last 12m	-3.26E-02	.157	-.320
V131 Ethnicity	-9.94E-02	-.111	.124

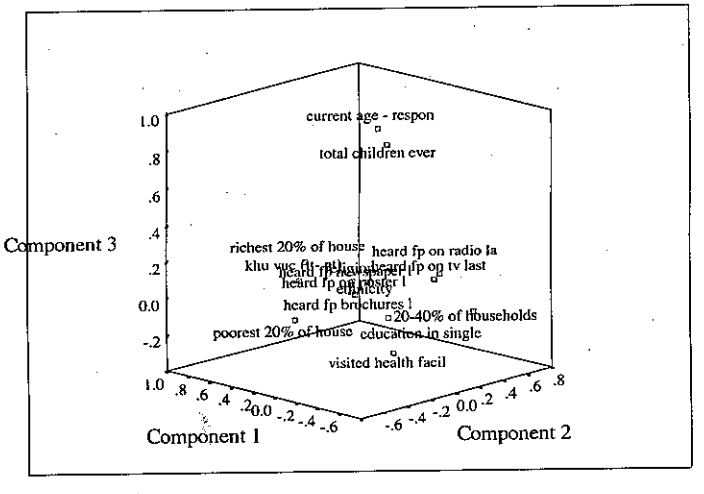
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

Chúng ta thấy rằng khi chiếu lên các mặt phẳng (R^2), các biến thuộc nhóm truyền thông luôn cụm vào một khu vực. Xem lại nội dung thông tin có trong các biến này, ta thấy có thể chọn một vài biến làm đại diện nếu muốn phân tích theo mục đích ban đầu chúng ta đã nêu, đó là các nhân tố kinh tế xã hội ảnh hưởng đến mức sinh.

Biểu đồ trên R^3 sau đây gợi ý chúng ta nên bỏ bớt một số biến trong nhóm nói trên.

Component Plot in Rotated Space



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Giảm số biến tham gia phân tích nhân tố

Chúng ta có thể tăng tỷ lệ giải thích của các thành phần chính bằng cách giảm đến mức tối đa các biến trong phân tích trên. Sau đây là các kết quả tóm tắt.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.516	25.2	25.16	2.516	25.2	25.16	2.447	24.5	24.47
2	1.685	16.9	42.01	1.685	16.9	42.01	1.754	17.5	42.01
3	1.298	13.0	54.99						
4	1.040	10.4	65.39						
5	.978	9.783	75.18						
6	.713	7.130	82.31						
7	.633	6.328	88.63						
8	.445	4.453	93.09						
9	.404	4.037	97.12						
10	.288	2.876	100.0						

Extraction Method: Principal Component Analysis.

Với việc bỏ đi một số biến truyền thông, chỉ giữ lại một biến Heard FP newspaper... và trong hai biến vùng và tôn giáo ta cũng bỏ đi một biến, tỷ lệ giải thích của hai thành phần chính đã đạt được 42%. Ngoài ra biểu đồ phân phối các biến trên các trục sau đây cũng đủ thông tin phân tích theo mục đích ban đầu. Các biến rải trên hai chiều của hai trục, đặc biệt là trên trục thứ nhất (Component 1), mặt khác trên trục thứ hai ta thấy số con không chỉ phụ thuộc vào tuổi của phụ nữ - như một điều đương nhiên mà nó còn phụ thuộc vào chất lượng chăm sóc sức khoẻ và dịch vụ y tế.

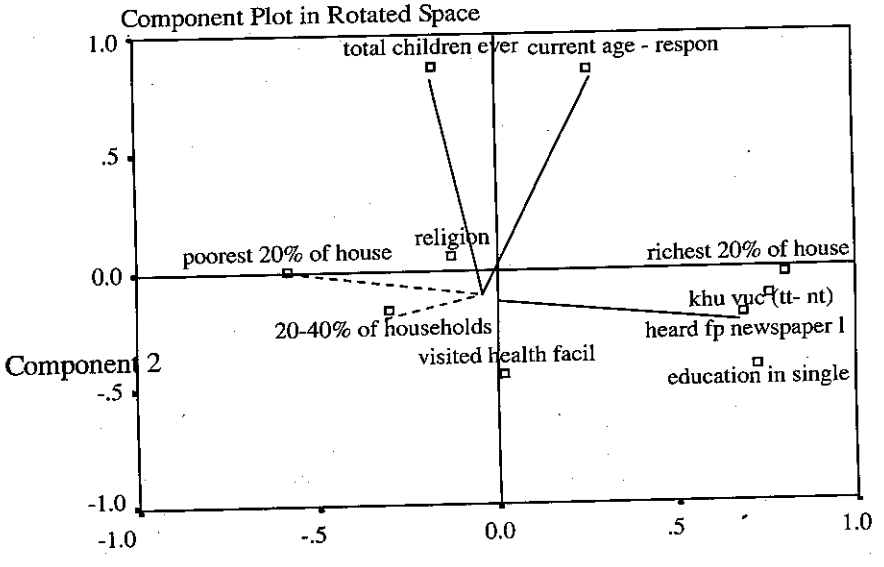
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Rotated Component Matrix[a]

	Component	
	1	2
HHHIGH20 Richest 20% of Households	.759	5.419E-02
URBAIN Khu vuc (TT- NT)	.716	-2.89E-02
V133 Education in single years	.682	-.308
V384C Heard FP newspaper last months	.650	-.100
HHLOW20 Poorest 20% of Households	-.540	8.297E-02
HH20_40 20-40% of Households	-.280	-6.84E-02
V201 Total children ever born	-.160	.873
V012 Current age - respondent	.247	.861
V394 Visited health facil. last 12m	1.951E-02	-.337
V130 Religion	-.114	.135

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.



Các nhân tố kinh tế thể hiện khá rõ trên trục thứ nhất, đó là sự đối lập giàu nghèo, nông thôn thành thị và trình độ học vấn. Tuy nhiên truyền thông hầu như chỉ đến với người thành thị, người giàu và có học vấn cao. Phải chăng là truyền thông hầu như không ảnh hưởng nhiều đến số con của phụ nữ có chồng, ít nhất là trong giới hạn của cuộc điều tra này. Trong khi thế hệ phụ nữ ảnh hưởng rất lớn đến số con mong muốn của họ (không chỉ

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

với số con đã có). Tuy nhiên, nếu ta chú ý đến góc tạo bởi các đường nối hai biến này về gốc toạ độ thì rõ ràng là Cosin của góc này (CO2) âm nhưng không nhỏ về giá trị tuyệt đối, điều đó cho thấy truyền thông trong mức độ nào đó làm cho mức sinh giảm. Cũng như vậy khi ta xét tác động của học vấn đến số con. Mặt khác khi quan tâm đến tầng lớp dân cư thu nhập thấp thì ta thấy rất rõ là 20% dân cư thu nhập thấp nhất có số con cao hơn 20% dân cư thu nhập ở ngay mức kế tiếp. Hãy quan sát cả góc tạo bởi hai biến này với số con của phụ nữ và độ dài của các cạnh tương ứng, ta thấy tác động của nghèo đói đến mức sinh không chỉ chặt chẽ hơn mà cường độ tác động cũng cao hơn. Có thể so sánh bởi độ dài hai hình chiếu vuông góc của Total chidrens ever born xuống các đường nối góc toạ độ và các biến Poorest 20% of Households, 20-40% of Househodls.

Kết quả phân tích với 3 thành phần chính có thể mô tả như sau:

Total Variance Explained

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.516	25.160	25.160	2.344	23.443	23.443
2	1.685	16.853	42.013	1.714	17.137	40.581
3	1.298	12.977	54.990	1.441	14.409	54.990

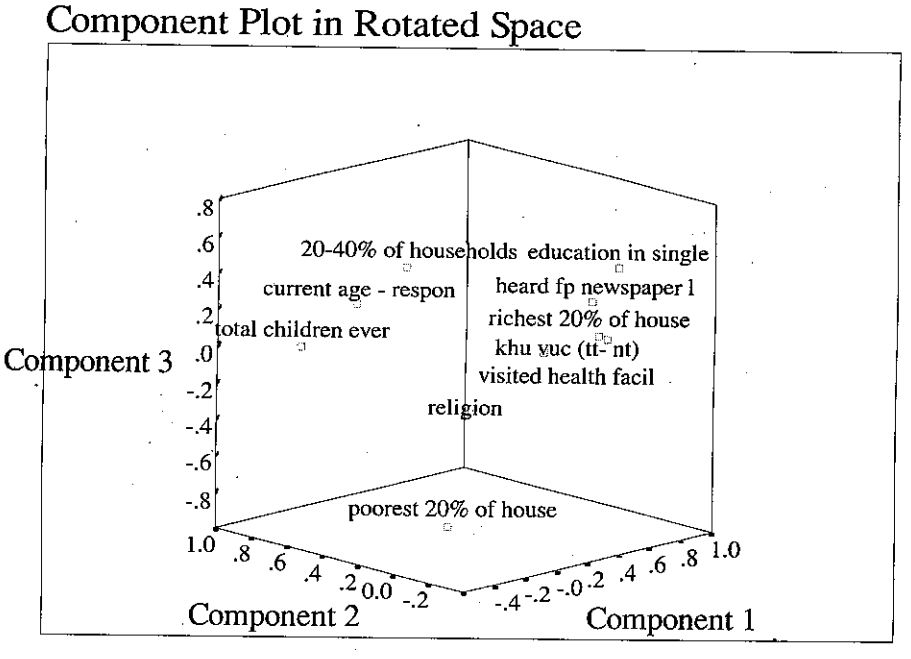
Extraction Method: Principal Component Analysis.

Rotated Component Matrix

	Component		
	1	2	3
HHHIGH20 Richest 20% of Households	.796	7.676E-02	-7.38E-03
URBAIN Khu vuc (TT- NT)	.755	-7.18E-03	1.997E-03
V384C Heard FP newspaper last months	.609	-4.36E-02	.247
V133 Education in single years	.581	-.217	.482
V012 Current age - respondent	.160	.891	8.454E-02
V201 Total children ever born	-.220	.869	-5.97E-02
V394 Visited health facil. last 12m	-7.50E-03	-.314	.166
HHLOW20 Poorest 20% of Households	-.311	-5.34E-02	-.784
HH20_40 20-40% of Households	-.509	2.300E-02	.653
V130 Religion	-4.22E-02	9.234E-02	-.264

a. Rotation converged in 5 iterations.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

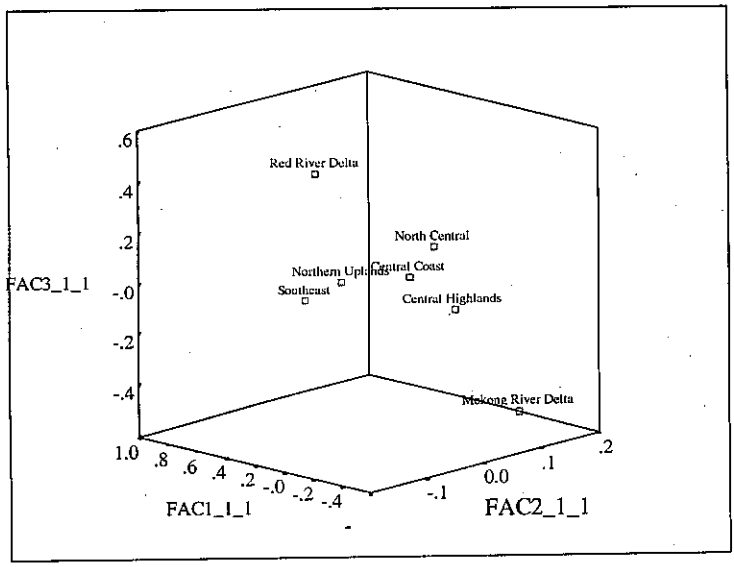


Vì số quan sát quá lớn, đồng thời chúng ta cũng không thể phân tích được những vị trí khác biệt của từng cá thể (nếu muốn có thể dùng các phân tích khác). Để mô tả sự khác biệt chung nhất khi dùng các biến nói trên làm các đặc trưng, chúng ta nhóm các cá thể theo vùng kinh tế, việc chia vùng đã được qui định bởi các văn bản pháp quy (1997). Năm 1997 trở về trước người ta chia Việt Nam thành 7 vùng kinh tế, kết quả tính ở mức trung bình cho các cá thể theo vùng trong phân tích cuối cùng này như sau:

REGION	FAC1_1	FAC2_1	FAC3_1
Northern Uplands	-.115	-.125	.116
North Central	-.226	.006	.195
Central Highlands	-.003	.100	-.143
Central Coast	.178	.065	-.017
Southeast	.757	.030	-.172
Mekong River Delta	-.353	.121	-.511
Red River Delta	.192	-.093	.482

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Mô tả trong R^3 ta có hình ảnh dưới đây:



Dễ dàng nhận thấy sự khác biệt rất lớn về các đặc trưng dân số-kinh tế của hai trung tâm dân cư lớn nhất là Đồng bằng sông Hồng và Đồng Bằng sông Cửu Long. Ngoài ra các vùng còn lại sự khác biệt không lớn lắm.

4. Những tóm tắt cơ bản

Mặc dù các thí dụ nêu trên không đủ các tình huống cần xử lý trong phân tích nhân tố bằng phương pháp thành phần chính, nhưng qua đó chúng ta có một số nhận xét tóm tắt có tính chất khái quát về nội một quá trình phân tích như sau:

+ Thực hiện các phân tích lựa chọn số biến thích hợp theo mục đích chính của người phân tích là nội dung quan trọng đầu tiên và cũng là công việc xuyên suốt quá trình phân tích.

+ Tiến hành các phân tích bổ trợ để có thể làm rõ đặc trưng thể hiện trên kết quả phân tích là hết sức cần thiết nhưng không nên quá cầu toàn vì bản chất khách quan của đám mây số liệu là giới hạn cao nhất của mọi điều chỉnh.

+ Chọn số thành phần chính ít nhất có thể, vì các phân tích hình học cuối cùng là rất cần thiết và không bỏ qua được.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Tìm cách mô tả kết quả rõ ràng cũng là một trong những nội dung quan trọng để có thể trình bày kết quả một cách dễ hiểu và phong phú.

+ Các biểu đồ của các cá thể hầu như không mô tả được, vì thực tế khi đã phân tích nhân tố là chúng ta đã làm việc với những bộ số liệu lớn, số cá thể rất lớn. Muốn mô tả chúng cần có những tiêu thức phân loại khác nhằm giảm số chiều của không gian biến. Thí dụ trên đã sử dụng biến vùng kinh tế để tóm tắt các cá thể.

13. TÓM TẮT CHƯƠNG 13

1. Mô hình phân tích nhân tố trước tiên là mô hình rút gọn thông tin về các cá thể của tổng thể, trong đó mỗi cá thể được đặc trưng bởi một số lượng lớn các biến. Phép rút gọn này được thực hiện trên cơ sở tạo một số biến mới là tổ hợp tuyến tính của các biến ban đầu (sơ cấp), các biến mới này đôi một độc lập với nhau.
2. Công cụ tạo các biến mới (còn gọi là các thành phần chính) nói trên là các phép chiếu vuông góc trong không gian tuyến tính. Qua các phép chiếu mỗi cá thể sẽ có một ảnh trên không gian con.
3. Tiêu chuẩn lựa chọn các biến mới là cực đại phương sai của đám mây ảnh trong điều kiện phương sai của các cá thể biểu hiện bằng tổng quán tính đã biết.
4. Mối quan hệ của các thành phần chính với các biến ban đầu được thể hiện qua các hệ số tương quan và đó cũng là cơ sở nhận biết ý nghĩa của các thành phần chính.
5. Phân tích thành phần chính có nhiều ứng dụng, trước hết là có thể nhận biết, phân nhóm các cá thể thông qua ảnh của chúng trong không gian có số chiều ít nhất có thể.
6. Phân tích thành phần chính có thể ứng dụng trong việc tránh đa cộng tuyến trong các hồi qui bội.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

CÁC KÝ HIỆU VÀ CÔNG THỨC CƠ BẢN

- Các công thức mô tả số liệu:

Tâm đám mây số liệu (véc tơ trung tâm): $g = X^T DI$

Số liệu trung tâm hóa: $Y = X - Ig^T$

Số liệu chuẩn hóa: $Z = \{z_{ij}\} = \left\{\frac{y_{ij}}{s_j}\right\}$

Ma trận hiệp phương sai: $V = X^T D X - gg^T$.

Ma trận hệ số tương quan: $R = D_{1/S} V D_{1/S} = Z^T DZ$.

Khoảng cách giữa các cá thể $d(X_1, X_2)$:

$$d^2 = (X_1 - X_2)^T \begin{bmatrix} a_1 & 0 & \ldots\ldots & 0 \\ 0 & a_2 & \ldots\ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots\ldots & a_p \end{bmatrix} (X_1 - X_2)$$

$d^2 = (X_i - X_j)^T M (X_i - X_j)$. trong đó M là một ma trận đường chéo dương.

- Công thức đo quán tính đám mây số liệu (các cá thể):

Tổng quán tính qui tâm (tâm g):

$$I_g = \sum_{i=1}^{n} p_i (X_i - g)^T M (X_i - g)$$

- Phép chiếu M vuông góc lên đường thẳng này xác định như sau:

$$P = a (a^T M a)^{-1} a^T M$$

- Quán tính đám mây ảnh qua phép chiếu P:

$$\text{TraceVMP} = \frac{Trace(a^T MVMa)}{a^T Ma}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Nghiệm bài toán cực đại TraceVMP: Nghiệm của $VMa = \lambda a$

- Mối quan hệ của thành phần chính, nhân tố chính

Thành phần phân tích	Phương trình xác định	Độ đo
Nhân tố chính u	MVu = λu	Chuẩn M-1
Trục chính a	VM a = λa	Chuẩn M
Thành phần chính c	XMXTc =λc	D- trực giao
Các liên hệ	c =Xu, λ; u=Ma	

- Thống kê Bartlett:

$$\chi^2 = -(W - 1 - \frac{2p+5}{6}) \ln|R|$$

- Hệ số tương quan của thành phần chính với 1 biến:

$$r(c,X_j) = \frac{\lambda u_i}{\sqrt{\lambda}} = \sqrt{\lambda}\, u_j.$$

- Tỷ lệ đóng góp của biến X_j cho một trục chính u:

$$(r(u_j))^2 = \frac{r^2(c,X_j)}{\lambda}$$

- Đóng góp của cá thể i giải thích cho thành phần chính thứ k:

$$CTR_k(i) = \frac{p_i (c_i^k)^2}{\lambda_k}$$

- Mức độ đánh giá cá thể i nhờ ảnh của nó bởi thành phần chính thứ k:

$$CO2_k(i) = \frac{\left(c_i^k\right)^2}{\left\|X_i^* - g\right\|^2}$$

Khi sử dụng ma trận hệ số tương quan phân tích thành phần chính:

$$\left\|X_i^* - g\right\|^2 = \sum_{j=1}^{p} z_{ij}^2$$

- Tổng lượng thông tin phản ánh được qua r thành phần chính đối với cá thể i:

$$QLT_r(i) = \sum_{k=1}^{r} CO2_k(i)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Ảnh của các thể	Image of the case
Biểu đồ thành phần chính	Component Plot
Biểu đồ thành phần chính khi đã quay	Component Plot in Rotated Space
Giá trị riêng	Eigenvalue
Ma trận chiếu	Matrix of projection
Ma trận hệ số tương quan của thành phần chính và các biến	Component matrix (of principal components to the variables)
Ma trận trung tâm hóa (qui tâm)	Matrix centre
Nhân tố chính	Principal factor
Phân tích nhân tố	Factor analysis
Phân tích thành phần chính	Principal component analysis
Phương pháp quay	Rotation method
Phép quay trực giao	orthogonal rotation
Phương pháp triết xuất (dữ liệu)	Extration mehod
Quán tính	Inertia
Thành phần chính	Principal component
Véc tơ riêng	Eigenvector
Véc tơ trung tâm	Centre vector
Vết ma trận	Trace of matrix

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

CÂU HỎI ÔN TẬP

13.1. Mục đích chung của bài toán phân tích nhân tố bằng phương pháp thành phần chính là gì?

13.2. Hãy cho biết các mệnh đề sau đúng hay sai?

a. Dữ liệu của mô hình phân tích nhân tố phải là các dữ liệu bằng số

b. Không phải đặc trưng nào của các cá thể trong tổng thể cũng có thể biểu hiện bằng số

c. Bài toán phân tích nhân tố chấp nhận phân tích các cá thể qua các đặc trưng có độ đo khác nhau.

13.3. Có ý kiến cho rằng mô tả dữ liệu của bài toán phân tích nhân tố việc gọi mỗi dòng của X tương ứng một cá thể và mỗi cột tương ứng 1 biến chỉ có tính qui ước. Bạn có nhận xét gì về ý kiến này? Nêu thí dụ cụ thể trong nghiên cứu kinh tế - xã hội minh họa cho ý kiến của bạn..

13.4. Trọng số của quan sát mẫu có thể mang nội dung gì?

13.5. Hãy lấy một thí dụ bằng số về ma trận dữ liệu X và tính toán: véc tơ trung tâm (tâm đám mây số liệu); bảng số liệu trung tâm hóa; ma trận hiệp phương sai và ma trận hệ số tương quan.

13.6. Ma trận M có vai trò gì trong việc tính khoảng cách giữa hai cá thể. Nêu một thí dụ trong đó các phần tử của M mang một ý nghĩa thực.

13.7. Ma trận D_{1/s^2} dùng trong bài toán ở chương này cho đặc trưng khi tính tổng quán tính?

13.8. Phép chiếu P trong bài toán phân tích nhân tố bằng phương pháp thành phần chính có những tính chất nào?

13.9. Trục chính thực chất là gì? Nhân tố chính thực chất là gì?

13.10. Bài toán xác định trục chính và nhân tố chính dựa trên tiêu chuẩn nào?

13.11. Nghiệm của bài toán phân tích thành phần chính là gì?

13.12. Cơ sở lý thuyết nào cho phép tăng dần số chiều của không gian chiếu mà không cần giải lại bài toán?

13.13. Những tiêu chuẩn nào có thể sử dụng để chọn số thành phần chính trong một phân tích?

13.14. Kiểm định nhờ thống kê Bartlett và KMO dùng để làm gì trong bài toán phân tích thành phần chính?

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

13.15. Để định danh các thành phần chính, người ta dựa vào thông tin nào?
13.16. Hãy nêu qui trình phân tích thành phần chính nhờ SPSS
13.17. Hồi qui thành phần chính là gì? Hãy chọn 1 bộ số liệu và thực hiện 1 tính toán cụ thể.

BÀI TẬP TỔNG HỢP CHƯƠNG 13

Với tệp Hhexpn98.sav (dta):*www.mfe.edu.vn\Thư viện\Dữ liệu-phần mềm*

13.18- a- Hãy thực hiện chương trình sau và cho những nhận xét có thể về các biến.

```
FACTOR
 /VARIABLES ricexpd nonrice langcomp educnexp insrx12m durbuser waterexp
 /MISSING LISTWISE /ANALYSIS ricexpd nonrice langcomp educnexp insrx12m
  durbuser waterexp
 /PRINT EXTRACTION ROTATION FSCORE  /PLOT EIGEN ROTATION
 /CRITERIA FACTORS(2) ITERATE(25)  /EXTRACTION PC
 /CRITERIA ITERATE(25)  /ROTATION VARIMAX  /SAVE REG(ALL)
 /METHOD=CORRELATION .
```

b- Thực hiện chương trình trên với vùng 2 và cho biết các tỉnh thành có thể phân nhóm như thế nào theo các thành phần chính.

13.19- Thực hiện lại chương trình trên cho vùng 6 và sắp xếp các biến, các tỉnh thành theo thứ tự giảm dần của mức đóng góp của các biến này trong việc hình thành các thành phần chính và các trục chính.

13.20- Hãy sử dụng SPSS phân tích nhân tố đối với nhóm biến sau đây: quint98b tobaco elecexp educnexp hhsize reg8 urban98 educyr98, từ tệp số liệu HHEXP98.sav với hai thành phần chính.

Nêu ý nghĩa của các kết quả phân tích

Nếu có thể tiến hành các hồi qui thì nên chọn các biến như thế nào để giải thích cho biến chi cho giáo dục của các hộ gia đình?.

Nếu chọn 3 thành phần chính thì có thể giải thích được bao nhiêu % quan hệ của các biến?.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sử dụng tiêu chuẩn lý thuyết kiểm tra việc lựa chọn số thành phần chính.

Gợi ý: Văn bản chương trình

```
FACTOR  /VARIABLES quint98b tobaco elecexp educnexp hhsize reg8 urban98 educyr98
 /MISSING LISTWISE /ANALYSIS quint98b tobaco elecexp educnexp hhsize reg8
  urban98 educyr98
 /PRINT INITIAL CORRELATION SIG EXTRACTION ROTATION
 /PLOT ROTATION  /CRITERIA FACTORS(2) ITERATE(25)  /EXTRACTION PC
/CRITERIA ITERATE(25)  /ROTATION VARIMAX
 /METHOD=CORRELATION .
```

13.21- Tiến hành phân tích thành phần chính đối với các biến: số năm đến trường của chủ hộ, tuổi của chủ hộ, vùng (7 vùng), kiểu hộ (farm) và phân loại theo ngũ phân vị (quint98b). Sử dụng ma trân hệ số tương quan và ma trận hiệp phương sai. Lựa chọn số nhân tố tối thiểu theo tiêu chuẩn lý thuyết (mức ý nghĩa 5%).

Trên cơ sở kết quả nhận được, hãy cho biết nếu 2 chủ hộ có số năm đến trường khác nhau 1 và các tiêu thức khác như nhau thì các toạ độ của hộ trên siêu phẳng chiếu thay đổi như thế nào. Nêu ý nghĩa của các thay đổi đó.

13.22- Phân tích thành phần chính sử dụng ma trận hệ số tương quan với các biến: agegroup, hhsize, ricexpd, durbuser, elecexp. Từ bảng hệ số tương quan của các biến với nhân tố chính thứ nhất, hãy xác định tỷ lệ đóng góp của mỗi biến trong phân tích thành phần chính nếu chọn một nhân tố.

Để có trên 80% quán tính của đám mây số liệu này được bảo toàn thì phải chọn bao nhiêu nhân tố chính (trong hai trường hợp: sử dụng ma trận tương quan và sử dụng ma trận hiệp phương sai).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Ngô Văn Thứ. Thống kê thực hành. NXB Khoa học và kỹ thuật. Hà Nội. 2008.
2. Tô Cẩm Tú: Phân tích số liệu nhiều chiều. NXB Nông nghiệp. Hà Nội 1996.
3. Benzecri. L'analyse des donnes en sociologie: Imprimerie des Presses Universitaires de France. 1990.
4. Gilbert Saporta: Analyse des donnees. Ecole national de la statistique et de l'administration economique. 1992-1993
5. Jean-marie bouroche et Gilbert Saporta: L'analyse des donnees . Que sais - Je. 1998.
6. Helene Erkel-Rousse: Analyse des donnees Applications et Programmation. L'ensae.1992.
7. Michel Volle: Analyse des donnees. Economica. 1994.
8. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992
9. Anderson T.W. An introduction to multivarate statistical analysis. New york. 1958
10. Dillon W and Goldstein M. Multivariate analysis Methods and Applications. New york . 1984
11. Dale J. Piorier: Intermadiate statistics and econometrics. 1995
12. David W. Hosmer and Stanley Lemeshow: Applied Logistic Regression. Wiley.1989.
13. L.Simar: An introduction to multivariate data analysis. Université catholique de Louvain
14. SPSS Base 8.0 Application guide. 1998.
15. T.W. Anderson: An introduction to multivariate statistical analysis. Wiley. 1984.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 14

PHÂN TÍCH TƯƠNG ỨNG

1. MỞ ĐẦU

Phân tích tương ứng được J.P. Benzécri đề xuất vào những năm 80 của thế kỷ XX khi nghiên cứu sự liên hệ của hai biến định tính. Ông đã sử dụng phân tích nhân tố về sự tương ứng cho việc phân tích nhu cầu quạt điện cho các hộ dân cư 80 quận của Paris, với nhân tố phân tích là lĩnh vực hoạt động xã hội của họ.

Phân tích tương ứng có thể xem như một ứng dụng của phương pháp phân tích nhân tố, trong đó khoảng cách giữa các cá thể (cũng như các biến) thể hiện bởi Metric Khi bình phương. Trong một giới hạn nào đó, đây chỉ là một vận dụng của phân tích nhân tố bằng phương pháp thành phần chính đã nêu ở chương 13. Điều khác biệt ở đây chính là đối tượng phân tích, phân tích tương ứng thực hiện hiệu quả hơn đối với các biến định tính. Một trong những lý do người ta sử dụng phân tích tương ứng như một phương pháp riêng là các đặc trưng kinh tế xã hội đôi khi rất mờ. Chẳng hạn, khi xem xét quan hệ giữa đầu tư cho giáo dục, đào tạo tại hai khu vực A và B, người ta có thể có các chỉ tiêu phân loại như giàu, đủ ăn, nghèo và rất nghèo; có việc ổn định, không ổn định, ít việc,.. Còn đầu tư cho giáo dục có thể chỉ biết là ít, nhiều và không có ý định trước chẳng hạn. Như vậy rõ ràng các quan hệ định tính có thể giống nhau trong khi các quan hệ định lượng khác nhau. Trường hợp này các phân tích nhân tố bằng phương pháp thành phần chính sẽ phức tạp hơn, đặc biệt là khả năng rút ra những kết luận xác đáng từ kết quả phân tích. Phân tích nhân tố bằng phân tích tương ứng có thể hiệu quả hơn vì phân tích này không đòi hỏi thang đo thống nhất. Phân tích tương ứng phù hợp hơn với các bảng tiếp liên với số liệu tần số.

Như tên gọi của nó, phân tích tương ứng trả lời câu hỏi: Có hay không sự tương ứng về khả năng tồn tại ở các trạng thái của các biến với các cá thể trong một tổng thể?. Nếu có thì đo sự tương ứng này như thế nào?.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với nội dung chương này bạn đọc sẽ nắm được:

- Cách nhận biết những vấn đề kinh tế xã hội có thể mô hình hoá và giải quyết được bằng phân tích tương ứng.
- Kỹ năng sử dụng công cụ, phần mềm để có lời giải
- Phân tích kết quả

Ở một mức cao hơn bạn đọc có thể tìm hiểu cơ sở lý thuyết, các biến đổi toán học và việc sử dụng thang đo khoảng cách trong phân tích. Phân tích và tìm cách vận dụng hiệu quả phương pháp này với việc chọn các độ đo phù hợp cho các lớp bài toán cụ thể.

2. BẢNG TIẾP LIÊN VÀ CÁC ĐỘ ĐO ĐỐI VỚI ĐÁM MÂY SỐ LIỆU

2.1. Mô tả hình học một bảng tiếp liên

Số liệu tần số về các cá thể với các chỉ tiêu định tính có thể mô tả trên một bảng cấp $m_1 x\ m_2$ trong đó m_1 và m_2 là số dấu hiệu (đặc trưng khác nhau của các cá thể). Chẳng hạn trên hai phương diện khác nhau có thể mô tả một tổng thể (sinh viên tốt nghiệp ở các trường năm 1999) nhờ bảng sau:

NHOMTR * LOAITN

Count

	LOAITN				
	Trung bình	Trung bình khá	Khá	Giỏi	Xuất xắc
Kỹ thuật và công nghệ	519	492	416	50	1
Khoa học cơ bản - Đa ngành	904	570	838	136	6
Nông - Lâm - Ngư nghiệp	265	215	201	22	
Kinh tế - Pháp lý	408	431	732	51	
Y - Dược - TD Thể thao	61	209	185	15	
Văn hoá - Nghệ thuật	53	1	60	26	2
Sư phạm	858	329	614	29	

Mỗi sinh viên TN được quan sát trên hai chỉ báo là nhóm trường và loại tốt nghiệp. Liệu có hay không một quan hệ tương ứng giữa nhóm trường và xếp loại tốt nghiệp của các sinh viên đã được quan sát.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Một cách tổng quát, ta có bảng số (pxq) dạng liệu sau:

Bảng N

A \ B	B_1	B_2	.	B_q
A_1	n_{11}	n_{12}		n_{1q}
A_2	n_{21}	n_{22}		n_{2q}
.				
A_p	n_{p1}	n_{p2}		n_{pq}

Trong đó A có p dấu hiệu và B có q dấu hiệu.

Đặt: $n_i=\sum_{j=1}^{q} n_{ij}$; $m_j=\sum_{i=1}^{p} n_{ij}$; $N=[n_{ij}]_{pxq}$

Phân bố xác suất của A là:

A	A_1	A_2	.	.	A_p
P	n_1/n	n_2/n	.	.	n_p/n

Phân bố xác suất của B là:

B	B_1	B_2	.	.	B_q
P	m_1/n	m_2/n	.	.	m_q/n

2.2. Mô tả ma trận và cấu trúc bài toán

Để tiện cho các mô tả sau này, chúng ta gọi đám mây điểm dòng là tập các điểm tương ứng với p véc tơ dòng của bảng N, đám mây điểm dòng thuộc R^q. Tương tự như vậy đám mây điểm cột tương ứng tập các véc tơ cột của bảng N, chúng thuộc R^p.

Gọi ma trận đường chéo các tần số biên theo từng biến A, B tương ứng là D_1 (ma trận biên dòng) và D_2 (ma trận biên cột) ta có:

$$D_1=\begin{pmatrix} n_1 & 0 & \ldots\ldots & 0 \\ 0 & n_2 & \ldots\ldots\ldots & 0 \\ \ldots\ldots\ldots & \ldots\ldots & \ldots\ldots & \\ 0 & 0 & \ldots\ldots\ldots & n_p \end{pmatrix} \qquad D_2=\begin{pmatrix} m_1 & 0 & \ldots\ldots & 0 \\ 0 & m_2 & \ldots\ldots\ldots & 0 \\ \ldots\ldots\ldots\ldots & \ldots\ldots & & \\ 0 & 0 & \ldots\ldots\ldots & m_q \end{pmatrix} \qquad (14.1)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

và N là ma trận lập bởi các phần tử n_{ij} từ bảng trên, dễ dàng nhận thấy n_i chính là tần số của dấu hiệu A_i và m_j là tần số của dấu hiệu B_j.

Chú ý rằng các ma trận tần số biên D_1 và D_2 là các ma trận đường chéo, mà mỗi ma trận đường chéo không suy biến A bất kỳ thì:

- Nghịch đảo A là các ma trận đường chéo, có các phần tử đường chéo là nghịch đảo của ma trận gốc.

- Chuyển vị của A là A và chuyển vị của A^{-1} là A^{-1}.

$D_1^{-1}N$ là ma trận tỷ lệ dòng với các véc tơ dòng có các phần tử là $\frac{n_{ij}}{n_i}$;

$$D_1^{-1}N = \begin{bmatrix} \frac{n_{11}}{n_1} & \frac{n_{12}}{n_1} & \cdots & \frac{n_{1q}}{n_1} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{n_{p1}}{n_p} & \frac{n_{p2}}{n_p} & \cdots & \frac{n_{pq}}{n_p} \end{bmatrix} \qquad (14.2)$$

Mỗi dòng i trong ma trận trên có các thành phần là tần suất có điều kiện các dấu hiệu của B theo A xác định tại A_i: $P(B=B_j/(A=A_i))= n_{ij}/n_i$.

Ngoài ra chú ý rằng: $(D_1^{-1}N)^T = N^T D_1^{-1}$; $D_1^{-1}N^T = (ND_1^{-1})^T$

$D_2^{-1}N$ là ma trận tỷ lệ theo cột với các véc tơ cột có các phần tử là $\frac{n_{ij}}{m_j}$;

$$D_2^{-1}N = \begin{bmatrix} \frac{n_{11}}{m_1} & \frac{n_{12}}{m_2} & \cdots & \frac{n_{1q}}{m_q} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{n_{p1}}{m_1} & \frac{n_{p2}}{m_2} & \cdots & \frac{n_{pq}}{m_q} \end{bmatrix} \qquad (14.3)$$

Mỗi cột j trong ma trận trên có các thành phần là tần suất có điều kiện các dấu hiệu của A với B xác định tại B_j: $P(A=A_i/(B=B_j))= n_{ij}/m_j$.

Ngoài ra cũng chú ý rằng: $(D_2^{-1}N)^T = N^T D_2^{-1}$; $D_2^{-1}N^T = (ND_2^{-1})^T$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Những đặc điểm nói trên cho phép thực hiện các tính toán sau này thuận tiện hơn.

Với số liệu ở trên ta có ma trận N:

$$\begin{bmatrix} 519 & 492 & 416 & 50 \\ 904 & 570 & 838 & 136 \\ 265 & 215 & 201 & 22 \\ 408 & 431 & 732 & 51 \\ 61 & 209 & 185 & 15 \\ 53 & 1 & 60 & 26 \\ 858 & 329 & 614 & 29 \end{bmatrix}$$

$$D_1 = \begin{bmatrix} 1477 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2448 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 703 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1622 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 470 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 140 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1830 \end{bmatrix}$$

$$D_2 = \begin{bmatrix} 3068 & 0 & 0 & 0 \\ 0 & 2247 & 0 & 0 \\ 0 & 0 & 3046 & 0 \\ 0 & 0 & 0 & 329 \end{bmatrix}$$

Trung tâm của đám mây điểm cột là véc tơ các tần suất biên của các dòng, có thể xác định véc tơ này như sau:

$$g_r = \frac{1}{n}(D_1^{-1}N)'D_1 I = \begin{pmatrix} \frac{n_1}{n} \\ \frac{n_2}{n} \\ . \\ . \end{pmatrix} = \begin{pmatrix} p_{*1} \\ p_{*2} \\ . \\ . \\ p_{*q} \end{pmatrix} \qquad (14.4)$$

Với thí dụ trên, ma trận (tần suất dòng) tỷ lệ theo dòng $D^{-1}{}_1N$ và véc tơ trung

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

tâm của các véc tơ dòng g_d tính được như sau:

Ma trận tần suất dòng $D^{-1}{}_1N$ có các phần tử: $\frac{n_{ij}}{n_i}$ (j=1,p; i=1,p)

$D^{-1}{}_1N$			
0.351388	0.333108	0.281652	0.033852
0.369281	0.232843	0.34232	0.055556
0.376956	0.305832	0.285917	0.031294
0.251541	0.265721	0.451295	0.031443
0.129787	0.444681	0.393617	0.031915
0.378571	0.007143	0.428571	0.185714
0.468852	0.179781	0.335519	0.015847

$(D^{-1}{}_1N)^T/n$						
4.04359E-05	4.24949E-05	4.33781E-05	2.89461E-05	1.49E-05	4.36E-05	5.4E-05
3.83323E-05	2.67944E-05	3.51936E-05	3.05778E-05	5.12E-05	8.22E-07	2.07E-05
3.2411E-05	3.93924E-05	3.29019E-05	5.19326E-05	4.53E-05	4.93E-05	3.86E-05
3.89556E-06	6.39304E-06	3.6012E-06	3.61826E-06	3.67E-06	2.14E-05	1.82E-06

$((D^{-1}{}_1N)^T/n)D_1$							g_r
0.059724	0.104028	0.030495	0.046951	0.00702	0.006099	0.098734	**0.353049**
0.056617	0.065593	0.024741	0.049597	0.024051	0.000115	0.03786	**0.258573**
0.047871	0.096433	0.02313	0.084235	0.021289	0.006904	0.070656	**0.350518**
0.005754	0.01565	0.002532	0.005869	0.001726	0.002992	0.003337	**0.03786**

Tương tự ma trận (tần suất cột) tỷ lệ theo cột:

$$N^TD^{-1}{}_2 = \begin{bmatrix} 0.1692 & 0.2190 & 0.1366 & 0.1520 \\ 0.2947 & 0.2537 & 0.2751 & 0.4134 \\ 0.0864 & 0.0957 & 0.0660 & 0.0669 \\ 0.1330 & 0.1918 & 0.2403 & 0.1550 \\ 0.0199 & 0.0930 & 0.0607 & 0.0456 \\ 0.0173 & 0.0004 & 0.0197 & 0.0790 \\ 0.2797 & 0.1464 & 0.2016 & 0.0881 \end{bmatrix}$$

véc tơ trung tâm của tập véc tơ cột là g_c (có các toạ độ ký hiệu là p_{i*}).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$(g_c)^T$ = (0.17 0.2817 0.0809 0.1867 0.0541 0.0161 0.2106)

Các phép đặt trên thực chất đã mô tả các cá thể thành hai tập trên hai không gian R^p và R^q, nói cách khác là các cá thể đã mang hai sắc thái là "cá thể dòng" và "cá thể cột". Tuy nhiên, một điểm quan trọng khác cần chú ý là trong các không gian này, chẳng hạn, trong R^q thì các điểm độc lập tối đa là q-1, vì tổng các tần suất bằng 1, hơn nữa chúng ta đã chỉ ra điểm trung tâm của đám mây, nên chỉ cần thêm q -1 điểm là đủ nhận biết toàn bộ đám mây này. Tính đối xứng của hai hệ véc tơ dòng và cột cho phép tiến hành các phân tích đồng thời tương đối thuận tiện.

Nếu A và B độc lập: $\frac{n_{ij}}{n_i} = \frac{m_j}{n}$ và $\frac{n_{ij}}{m_j} = \frac{n_i}{n}$

3. ĐỘ ĐO KHOẢNG CÁCH KHI BÌNH PHƯƠNG

Đối với bài toán phân tích nhân tố trong trường hợp này, người ta có thể sử dụng phân tích thành phần chính nhưng việc chọn độ đo cho mỗi không gian là khác nhau. Phân tích tương ứng chọn một độ đo tương đối đặc biệt đó là độ đo Khi bình phương.

Khoảng cách giữa hai điểm i và k trong không gian các véc tơ dòng q chiều của bảng N, $d_i = (n_{i1}, n_{i2}, ..., n_{iq})$ và $d_k = (n_{k1}, n_{k2}, ..., n_{kq})$ được xác định theo công thức tổng quát sau:

$$d^2(i,k) = ||x_i - x_k||^2 = \sum_{j=1}^{q} \frac{1}{q_j}(x_{ij} - x_{kj})^2$$

Nếu chọn $1/q_j = n/m_j$, với hai dòng i, k của bảng N ta có:

$$d^2_{\chi^2}(i,k) = \sum_{j=1}^{q} \frac{n}{m_j}\left(\frac{n_{ij}}{n_i} - \frac{n_{kj}}{n_k}\right)^2 \qquad (14.5)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Đây là khoảng cách với metric M là ma trận $Q^{-1}= nD_2^{-1}$, xác định như sau:

$$Q^{-1}=nD_2^{-1}=\begin{bmatrix} \frac{n}{m_1} & 0..... & 0 \\ 0 & \frac{n}{m_2}..... & 0 \\ 0 & 0..... & 0 \\ ..... & ..... & .. \\ 0 & 0..... & ...\frac{n}{m_q} \end{bmatrix} \quad (14.6)$$

Chẳng hạn với dòng 1 và 3 của bảng N trong mục trên ta có:

d_1	519	492	416	50	n_1=1477
d_3	408	431	732	51	n_3=1622
m_j	3068	2247	3046	329	n=8690
n/m_j	2.832464	3.867379	2.852922	26.41337	

n_{1j}/n_1	0.351388	0.333108	0.281652	0.033852
n_{3j}/n_3	0.251541	0.265721	0.451295	0.031443
hiệu BF	0.009969	0.004541	0.028779	5.807E-06

Tính các số hạng của tổng (8.5): $\frac{n}{m_j}(\frac{n_{1j}}{n_1}-\frac{n_{3j}}{n_3})^2$

j	1	2	3	4
	0.028238	0.017561	0.082103	0.0001534

Khoảng cách Khi bình phương: $d^2_{\chi^2}(1,3)=0.128056$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Khoảng cách Khi bình phương nói trên có một vài tính chất đáng chú ý sau:

Tính chất 1: Khoảng cách giữa hai véc tơ tỷ lệ luôn bằng không

Thật vậy: nếu hai dòng i và k tỷ lệ thì $n_{ij} = \alpha n_{kj}$ và $n_i = \alpha n_k$ nên tất cả các chêch lệch trong bình phương trên đều bằng không hay $\frac{n_{ij}}{n_i} - \frac{n_{kj}}{n_k} \equiv 0$.

Vậy $d^2_{\chi^2}(i,k) = 0$.

Tính chất 2: Trọng số $\frac{n}{m_j}$ *làm tăng sai khác* $\frac{n_{ij}}{n_i} - \frac{n_{kj}}{n_k}$ *khi* m_j *nhỏ, như vậy khi mj/n nhỏ thì một sai khác nhỏ của dòng i và dòng k được điều chỉnh tăng nhờ* $\frac{n}{m_j}$ *lớn.*

Tính chất 3: Nếu có hai cột tỷ lệ thì khi thay hai cột đó bằng 1 cột là tổng của chúng, khoảng cách của các dòng không thay đổi.

Với cách định nghĩa khoảng cách như trên, chúng ta có thể giảm số chiều của hai đám mây chiếu trên hai không gian các dòng và các cột, nếu có những dòng hay những cột tỷ lệ.

Định nghĩa khoảng cách của hai véc tơ cột trong R^q hoàn toàn tương tự và cũng có những tính chất tương tự như trên.

Tức là, khoảng cách giữa hai cột j, h được tính theo công thức sau:

$$d^2_{\chi^2}(j,h) = \sum_{i=1}^{p} \frac{n}{n_i} \left(\frac{n_{ij}}{m_j} - \frac{n_{ih}}{m_h}\right)^2 \qquad (14.7)$$

Với độ đo này, tổng quán tính của các dòng theo véc tơ trung tâm của các dòng (g_r) có thể tính theo công thức sau:

$$I_{g_r} = \sum_{i=1}^{p} \frac{n_i}{n} d^2_{\chi^2}(i,g_r) = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_i}{n} \frac{n}{m_j} \left(\frac{n_{ij}}{n_i} - \frac{m_j}{n}\right)^2 \qquad (14.8)$$

Biến đổi (14.8) ta có:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$I_{g_r} = \sum_{i=1}^{p} \frac{n_i}{n} d^2_{\chi^2}(i, g_r) = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_i}{n}\frac{n}{m_j}\left(\frac{n_{ij}}{n_i} - \frac{m_j}{n}\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n}{n_i}\frac{n}{m_j}\left(\frac{n_i}{n}\frac{n_{ij}}{n_i} - \frac{n_i}{n}\frac{m_j}{n}\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n}{n_i}\frac{n}{m_j}\left(\frac{n_{ij}}{n} - \frac{n_i}{n}\frac{m_j}{n}\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n}{n_i}\frac{n}{m_j}\frac{1}{n^2}\left(n_{ij} - \frac{n_i m_j}{n}\right)^2$$

$$I_{g_r} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_i m_j}{n}\right)^2}{\frac{n_i m_j}{n}} \qquad (14.9)$$

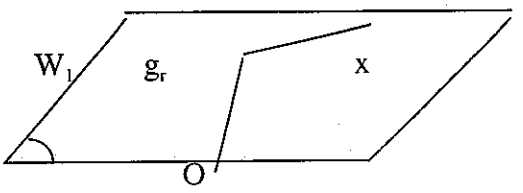
nI_{g_r} chính là thống kê Khi bình phương dùng trong kiểm định tính độc lập của hai dấu hiệu định tính. Thống kê này phân phối Khi bình phương với bậc tự do k= (q-1)(p-1). Để nhận thấy ý nghĩa của thống kê này, chúng ta có thể xem xét trường hợp hạng của hệ véc tơ dòng trong N bằng 1. Chẳng hạn dòng 1 là tổ hợp tuyến tính của các dòng còn lại, ta bỏ dòng 1 nhận được bảng (N-1). Tiếp tục như vậy cho đến khi chỉ còn 1 dòng, lúc này véc tơ trung tâm của hệ véc tơ dòng chính là véc tơ đó, vậy $I_{gr} = 0$. Thực tế là có thể biểu hiện theo B trên các chỉ báo của A như nhau về cơ cấu, ta có thể kết luận ngay chỉ báo A không tương quan với chỉ báo B.

Kiểm định Khi bình phương ở đây chính là kiểm định tính độc lập nếu có của các cá thể, khi mô tả chúng bởi hai dấu hiệu A và B, mặt khác đây cũng là kiểm định sự độc lập của hai đám mây số liệu trong hai không gian chiếu đã nói ở phần I.

Chú ý rằng, đám mây điểm tỷ lệ theo dòng nằm trong không gian con W_1, véc tơ Og_r vuông góc với không gian này tại tâm theo Metric Khi bình phương. Thật vậy, cho x là một véc tơ của W_1 thì:

$(x-g_r)nD_2^{-1}g_rI = 0$ vì $(x-g_r)$ là véc tơ sai lệch cá biệt hay véc tơ qui tâm của véc tơ tỷ lệ theo dòng, mặt khác $nD_2^{-1}g_r$ là véc tơ có mọi thành phần bằng *1*.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Ngoài ra mọi véc tơ của W_1 có độ dài bằng 1, hay xI=1, và $\|g_r\|^2=1$.

4. PHÂN TÍCH THÀNH PHẦN CHÍNH TRONG CÁC KHÔNG GIAN CHIẾU

Hai phân tích thành phần chính cần phải thực hiện là:

- Phân tích thành phần chính theo dòng với bảng số liệu $X_r= D_1^{-1}N$;

  Metric $M_r=nD_2^{-1}$ và ma trận trọng số $D_d=D_1/n$.

- Phân tích thành phần chính theo cột với bảng số liệu Xc= D2-1N;

  Metric $M_c=nD_1^{-1}$ và ma trận trọng số $D_c=D_2/n$.

Các ma trận M (metric) theo dòng và cột được tính từ thí dụ trên là:

$$M_r=nD_2^{-1}=\begin{bmatrix} 2.832464 & 0 & 0 & 0 \\ 0 & 3.867379 & 0 & 0 \\ 0 & 0 & 2.852922 & 0 \\ 0 & 0 & 0 & 26.41337 \end{bmatrix}$$

$$M_c= nD_1^{-1}=\begin{bmatrix} 5.88354 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.54983 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 12.3613 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5.35758 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 18.4893 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 62.0714 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4.74863 \end{bmatrix}$$

4.1. Phân tích thành phần chính không trung tâm hoá và nhân tố tầm thường theo số liệu sơ cấp đối với các dòng và các cột

Vì phân tích theo dòng và cột có tính chất đối xứng, ta xét đám mây số

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

liệu tương ứng với hệ véc tơ dòng.

Đặc trưng của chúng chính là ma trận tỷ lệ theo dòng $X_r=D_1^{-1}N$.

$$D_1^{-1}N=\begin{bmatrix} \frac{n_{11}}{n_1} & \frac{n_{12}}{n_1} & \cdots\cdots & \frac{n_{1q}}{n_1} \\ \cdots\cdots & & & \\ \frac{n_{p1}}{n_p} & \frac{n_{p2}}{n_p} & \cdots\cdots & \frac{n_{pq}}{n_p} \end{bmatrix}, \text{ tức là: } X_i=(x_{ij}=\frac{n_{ij}}{n_i});\ i=1,p$$

Chúng ta có ma trận hiệp phương sai của đám mây số liệu là:

$$V=X^TDX-gg^T=X_r^TDX_r-g_r g_r^T, \text{ trong đó } D_r=D_1/n.$$

Theo kết quả ở trên, ta thấy Og vuông góc với mọi bộ phận của đám mây số liệu, có thể xem đó là một thành phần chính, nói cách khác véc tơ riêng của ma trận VM tương ứng với $\lambda=0$.

Các véc tơ riêng của VM chính là các véc tơ riêng của X^TDXM có cùng các giá trị riêng, trừ g tương ứng với giá trị riêng là 1 (g không phải là véc tơ 0).

Dễ dàng nhận thấy gg^TM luôn có hạng bằng 1,

Với: $X^TDXM=VM+gg^TM$,

ta cõ: $X^TDXMg=VMg+g\,g^TMg=0+g\|g\|^2\chi^2=g$ (g đã chuẩn hoá).

Hay: $X^TDXMg=1g$

ta thấy g là véc tơ riêng của X^TDXM với giá trị riêng 1, nhưng không phải là véc tơ riêng của VM với giá trị riêng1 mà giá trị riêng 0.

Vì: $VM=X_r^TDX_rM-g_r g_r^TM$

suy ra: $VMg=X_r^TDX_rMg-g_r g_r^TMg=0$

hay: $VMg=0g$

- Thật vậy, có thể xét các véc tơ dòng của ma trận N trong thí dụ trên. Từ bảng tần số ta đã tính được các thành phần $X=D_1^{-1}N$;

$D^{-1}{}_1N$			
0.351388	0.333108	0.281652	0.033852

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

0.369281	0.232843	0.34232	0.055556
0.376956	0.305832	0.285917	0.031294
0.251541	0.265721	0.451295	0.031443
0.129787	0.444681	0.393617	0.031915
0.378571	0.007143	0.428571	0.185714
0.468852	0.179781	0.335519	0.015847

Ma trận trọng số $D=D_1/n$:

$$\begin{bmatrix} 0.169965 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.281703 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.080898 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.186651 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.054085 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.01611 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.210587 \end{bmatrix}$$

Ma trận độ đo: Metric $M=nD_2^{-1}$

$$\begin{bmatrix} 2.832464 & 0 & 0 & 0 \\ 0 & 3.867379 & 0 & 0 \\ 0 & 0 & 2.852922 & 0 \\ 0 & 0 & 0 & 26.41337 \end{bmatrix}$$

Ta có:

$$nX^TD = \begin{bmatrix} 0.05972 & 0.10403 & 0.03049 & 0.04695 & 0.00702 & 0.00610 & 0.09873 \\ 0.05662 & 0.06559 & 0.02474 & 0.04960 & 0.02405 & 0.00012 & 0.03786 \\ 0.04787 & 0.09643 & 0.02313 & 0.08423 & 0.02129 & 0.00690 & 0.07066 \\ 0.00575 & 0.01565 & 0.00253 & 0.00587 & 0.00173 & 0.00299 & 0.00334 \end{bmatrix}$$

(Do X^TD nhỏ, ta nhân thêm n)

$$nX^TDX = \begin{bmatrix} 0.132218 & 0.086834 & 0.120844 & 0.013153 \\ 0.086834 & 0.072380 & 0.090075 & 0.009283 \end{bmatrix}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$\begin{bmatrix} 0.120844 & 0.090075 & 0.126167 & 0.013432 \\ 0.013153 & 0.009283 & 0.013432 & 0.001992 \end{bmatrix}$$

$$nX^TDXM = \begin{bmatrix} 3254.4396 & 2918.2802 & 2995.941 & 3019.043 \\ 2137.3454 & 2432.5136 & 2233.14 & 2130.815 \\ 2974.4574 & 3027.2112 & 3127.916 & 3083.018 \\ 323.7501 & 311.98852 & 332.9983 & 457.1408 \end{bmatrix}$$

$$nX^TDXMg_r$$

$$\begin{bmatrix} 3067.997 \\ 2246.998 \\ 3045.998 \\ 329.0007 \end{bmatrix}$$

suy ra:

$$X^TDXMg_r$$

$$\begin{bmatrix} 0.353049 \\ 0.258573 \\ 0.350518 \\ 0.03786 \end{bmatrix}$$

Kết quả là $X^TDXMg_r = 1g_r$.

Véc tơ riêng này không được sử dụng khi trung tâm hoá các bảng số liệu chiếu cũng như khi chúng ta phân tích thành phần chính không trung tâm hoá: giá trị riêng 1 nhận được sau các phép khử. Giá trị riêng tầm thường này tương ứng với trục chính g và nhân tố chính không đổi: Mg=I. Trong trường hợp này phân tích không có tác dụng rút gọn thông tin, phân tích như vậy không mang lại lợi ích gì. Chính nguyên nhân này đòi hỏi chúng ta phải tiến hành phân tích trên không gian các tổ hợp tuyến tính của các dòng hay các cột trong N. Các tổ hợp này hình thành nhờ các phân tích

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

nhân tố, vì vậy chúng ta có thể dùng phương pháp thành phần chính để phân tích tương ứng giữa các dấu hiệu của biến dòng và các dấu hiệu của biến cột.

4.2. Phân tích thành phần chính hình chiếu của các đám mây

Sử dụng các kết quả ở chương 13 cho phân tích này:

Đối với các dòng ta có: $X = D_1^{-1}N$; Metric $M = nD_2^{-1}$ và ma trận trọng số $D = D_1/n$.

Suy ra $X^TDX = [N^TD_1^{-1}N]/n$ và $MX^TDX = D_2^{-1}N^TD_1^{-1}N$; $XMX^TD = D_1^{-1}N\,D_2^{-1}N^T$

Phân tích thành phần chính đối với ma trận $X = D_1^{-1}N$ thực chất là phân tích các dấu hiệu dòng được đo bởi các dấu hiệu cột. Nói cách khác nếu xem mỗi dòng của N là 1 cá th. mỗi cột là một đặc trưng thì phân tích này (như trong chương 13) sẽ là co sở nhận biết sự khác nhau của các dấu hiệu dòng. Các dấu dòng có thể phân nhóm nhờ phân tích này trên cơ sở các thành phần chính (là tổ hợp tuyến tính của các dấu hiệu cột).

Tương tự đối với các cột, ta có: $X = N\,D_2^{-1}$, Metric $M = nD_1^{-1}$ và ma trận trọng số $D = D_2/n$.

suy ra $X^TDX = [ND_2^{-1}N^T]/n$ và $MX^TDX = D_1^{-1}ND_2^{-1}N^T$; $XMX^TD = D_2^{-1}N^T\,D_1^{-1}N$

Phân tích thành phần chính đối với ma trận $X = N\,D_2^{-1}$ thực chất là phân tích các dấu hiệu cột được đo bởi các dấu hiệu dòng. Các dấu hiệu cột có thể phân nhóm nhờ phân tích này trên cơ sở các thành phần chính (là tổ hợp tuyến tính của các dấu hiệu dòng).

Mặt khác phân tích theo dòng hay theo cột trên cơ sở bảng N luôn bắt đầu từ một ma trận (MX^TDX) và chuyển vị của nó. Điều này cho thấy trong thực tế chỉ cần thực hiện một phân tích, kết quả phân tích thứ hai nhận được từ kết quả phân tích thứ nhất.

Có thể tóm tắt như sau

	ACP các véc tơ dòng	ACP các véc tơ cột
Các nhân tố chính	Véc tơ riêng của $MX^TDX = D_2^{-1}N^T\,D_1^{-1}N$	Véc tơ riêng của $XMX^TD = D_1^{-1}N\,D_2^{-1}N^T$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thành phần chính	Véc tơ riêng của $XMX^TD=D_1^{-1}N\,D_2^{-1}N^T$ Chuẩn hoá bởi a sao cho: $a^T[D_1/n]a=\lambda$	Véc tơ riêng của $MX^TDX=D_2^{-1}N^T\,D_1^{-1}N$ Chuẩn hoá bởi b sao cho: $b^T[D_2/n]b=\lambda$

Ta nhận thấy hai phân tích (theo dòng và theo cột) cùng có các giá trị riêng như nhau và các thành phần chính của phân tích theo dòng chính là các nhân tố chính của phân tích theo cột. Ngoài ra các ma trận có các véc tơ riêng là các thành phần chính là tích của hai ma trận $D_1^{-1}N$ và $D_2^{-1}N^T$ với thứ tự khác nhau.

Với các số liệu trên nhờ phần mềm Mathematica 3.0 ta tính các giá trị riêng và các véc tơ riêng như sau: MX^TDX

0.37450396	0.245955	0.342285	0.037255
0.33582053	0.279921	0.348356	0.035902
0.34475729	0.256978	0.359944	0.03832
0.34741577	0.245203	0.354778	0.052605

Các giá trị riêng tìm được là:(1, .0371015, .019459, .010415)

Các véc tơ riêng tương ứng (***các nhân tố chính trong phân tích theo dòng***):

(0.5, 0.5, 0.5, 0.500001),

(-0.643585, 0.712913, 0.147817, -0.236016),

(-0.155203, -0.202854, 0.203888, 0.945088),

(0.0502937, 0.189277, -0.291418, 0.936333)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Sau đây là đoạn chương trình trên Mathematica:

```
m = {{0.374504, 0.245955, 0.342285, 0.037255}, {0.33582053, 0.279921,
0.348356, 0.035902}, {0.34475729, 0.256978, 0.359944, 0.03832}, {0.34741577,
0.245203, 0.354778, 0.052605}};
Eigenvalues[ m ];Eigenvectors [m];
```

Các toạ độ của các điểm dòng và các điểm cột nhận được từ các véc tơ riêng của các tích của hai ma trận tỷ lệ theo dòng và theo cột.

Các thành phần chính tương ứng với 4 nhân tố chính là:

c_1	c_2	c_3	c_4
0.377964	0.113914	0.154016	-0.32552
0.377964	-0.08656	-0.08364	-0.15989
0.377964	0.026088	0.153977	-0.24487
0.377964	0.219962	-0.13565	0.41978
0.377964	0.719719	-0.00038	-0.06302
0.377964	-0.55479	-0.95516	-0.74463
0.377964	-0.32354	0.121832	0.271765

Tính đối xứng trong phân tích thành phần chính theo dòng và theo cột dẫn đến sự trùng lặp của hai phân tích và cuối cùng là có thể tái hiện lại các biến dòng và cột của bảng tiếp liên ban đầu trên cùng một mặt phẳng các thành phần chính.

Giới hạn của quan hệ tương quan không bó hẹp trong điều kiện phân tích các biến định tính, việc giải thích các cấu thành tương quan sẽ rất tiện lợi khi sử dụng mức đóng góp của các đặc trưng vào quán tính của các trục hay các véc tơ riêng.

Vì: $\lambda = \frac{1}{n}\sum_{i=1}^{p} n_i a_i^2 = \frac{1}{n}\sum_{j=1}^{q} m_j b_j^2$ Trong đó λ là tổng quán tính của các đám mây dòng cũng là tổng quán tính của đám mây cột.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chúng ta gọi *tỷ lệ đóng góp của dòng i* trong tổng quán tính là:

$$CTR(i) = \frac{\frac{n_i}{n} a_i^2}{\lambda} \qquad (14.10)$$

Tương tự *tỷ lệ đóng góp của cột j* là:

$$CTR(j) = \frac{\frac{m_j}{n} b_j^2}{\lambda} \qquad (14.11)$$

Kết quả tương ứng nhận được trên SPSS như sau

Summary

Dimension	Statistics						
				Proportion of Inertia		Confidence Singular Value	
	Inertia	Chi Square	Sig.	Accounted for	Cumulative	Standard Deviation	Correlation 2
1	.037			.554	.554	.010	-.022
2	.019			.291	.844	.013	
3	.010			.156	1.000		
Total	.067	582.014	.000[a]	1.000	1.000		

a. 18 degrees of freedom

Chú ý rằng SPSS đã bỏ đi giá trị riêng lớn nhất $\lambda_1=1$.

Vì trong phân tích thành phần chính đối với các dòng cũng như các cột thường có rất nhiều đặc trưng và cũng có thể có nhiều đặc trưng có tỷ lệ đóng góp lớn thể hiện như một trục chính, nên người ta chỉ chú ý đặc biệt đến các dòng hay cột có $CTR(i) > \frac{n_i}{n}$ (tương tự với các cột). Tuy nhiên, tỷ lệ đóng góp cần được bổ sung dấu của các toạ độ vì các đặc trưng có thể có những đóng góp có ý nghĩa trái ngược nhau. Điều này giải thích được qua tính chuẩn hoá của a và b.

4.3. Công thức chuyển đổi

Có thể nhận thấy các thành phần chính đã chuẩn hoá (a và b) có liên hệ rất chặt chẽ với nhau. Sau đây ta xác định công thức chuyển của chúng.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Toạ độ của điểm-dòng và điểm-cột tồn tại trong quan hệ rất chặt chẽ vì vậy có thể lợi dụng chính điểm này để tìm thành phần chính theo chiều nhỏ hơn của bảng tần số. Chẳng hạn với p << q, chúng ta xác định a qua phương trình:

$$D_1^{-1}ND_2^{-1}N^Ta=\lambda a \qquad (*)$$

Nhân hai vế với $D_2^{-1}N^T$ ta có một véc tơ tỷ lệ với b:

$$D_2^{-1}N^T D_1^{-1}ND_2^{-1}N^Ta=\lambda D_2^{-1}N^T a$$

Đặt $b = kD_2^{-1}N^T a$ cần xác định k sao cho $b^TD_2b/n=\lambda$.

Tức là: $k^2a^TN D_2^{-1} D_2 D_2^{-1}N^Ta/n= k^2a^TN D_2^{-1}N^Ta/n=\lambda$.

Từ (*) ta có: $N D_2^{-1}N^Ta=\lambda D_1a$. Vậy $k^2a^T\lambda D_1a /n=\lambda$,

nhưng $a^TD_1a /n=\lambda$ nên cuối cùng ta có: $k^2 =1/\lambda$.

Ta có công thức chuyển đổi như sau:

$$b = \frac{1}{\sqrt{\lambda}} D_2^{-1}N^T a \text{ tức là } b_j = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{p} \frac{n_{ij}}{n_i} a_i \qquad (14.12)$$

$$a = \frac{1}{\sqrt{\lambda}} D_1^{-1}Nb \text{ tức là } a_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{q} \frac{n_{ij}}{m_j} b_j \qquad (14.13)$$

trong đó:

$$a = \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_{m1} \end{bmatrix} \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_{m2} \end{bmatrix}$$

Với mỗi λ_k ta có:

a^k là nhân tố chính thứ k trong R^p (không gian các cá thể)

b^k là nhân tố chính thứ k trong R^q (không gian các biến).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

4.4. Vết ma trận và việc tái hiện dữ liệu

1. Phân tích tổng quán tính

Tổng quán tính của đám mây số liệu theo dòng cũng như theo cột, đã nói ở trên, là như nhau.

Bỏ qua giá trị riêng ứng với nhân tố chính tầm thường ta có thể tính theo công thức sau:

$$Ig = \sum_{k=1}^{p-1} \lambda_k \qquad \text{với } p << q. \text{ (Chú ý là } \lambda_p=1 \text{ bị bỏ qua)}$$

Với thí dụ trên ta có I_g=0,066975.

Mỗi một trục chính tương ứng sẽ giải thích một phần có tính chất độc lập trong tổng quán tính này. Các tỷ lệ tương ứng sẽ là: λ_k/Ig. (k=1..p-1)

P.Cibois đã chứng minh một mệnh đề quan trọng, mệnh đề này chỉ ra rằng phân tích tương ứng nghiên cứu sự khác biệt trong cấu trúc độc lập với chính sự khác biệt này. Đó là, tồn tại một bảng N* với các phần tử:

$$n_{ij}^{*} = \frac{n_i m_j}{n} + \alpha\left(n_{ij} - \frac{n_i m_j}{n}\right)$$

mà với $0<\alpha<1$ thì cấu trúc của sự khác biệt hoàn toàn độc lập với α. Mặt khác N* hoàn toàn tương ứng với N theo nghĩa là: Chúng có cùng đồ thị, cùng tỷ lệ quán tính khi phân tích tương ứng. Sự khác biệt chỉ là tổng quán tính $I^* = \alpha^2 I$, tức là $\lambda^* = \alpha^2 \lambda$.

Kết quả này cho phép tái hiện dữ liệu một cách đồng thời trên cùng một hệ toạ độ nếu các phân tích theo dòng và theo cột đều chọn số thành phần chính bằng nhau.

2. Công thức biểu diễn

Trong chương 13, chúng ta đã có công thức (7.24): $X = \sum_k c_k u_k^T M^{-1}$.

Sử dụng công thức này với X là ma trận tỷ lệ theo dòng, c_k thay bởi a^k với các toạ độ dòng trên trục chính thứ k, nhân tố chính u_k nhận được từ véc tơ b^k chuẩn hoán nhờ phép chia cho căn bậc 2 của λ_k, còn ma trận M được thay bởi nD_2^{-1}. Từ đó ta có:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$\frac{n_{ij}}{n_i} = \sum_{k=1}^{p-1} \frac{a_i^k b_j^k}{\sqrt{\lambda_k}} \frac{m_j}{n}$$

$$n_{ij} = \frac{n_i m_j}{n} \sum_{k=1}^{p-1} \frac{a_i^k b_j^k}{\sqrt{\lambda_k}}$$

Chú ý rằng nếu chúng ta sử dụng đủ m_1 nhân tố trong phân tích thì phải đưa thêm nhân tố ứng với giá trị riêng $\lambda=1$ (tương ứng với $\lambda=0$ khi qui tâm ma trận X). Lúc đó các tần số được tính theo công thức sau:

$$n_{ij} = \frac{n_i m_j}{n}(1 + \sum_{k=1}^{p-1} \frac{a_i^k b_j^k}{\sqrt{\lambda_k}})$$

Công thức này tái hiện toàn bộ dữ liệu ban đầu của bảng tần số N.

Kết quả quan trọng ở đây chính là nhờ mô tả trên, chúng ta có thể biểu diễn đồng thời cả hai đặc trưng ban đầu (dòng và cột) trên cùng một siêu phẳng chiếu. Tính chất về các quan hệ giữa các dòng như các cá thể và các cột như các biến không thay đổi vì thực tế chúng ta chỉ điều chỉnh lại độ đo theo cùng tỷ lệ trên tất cả các trục của một siêu phẳng chiếu, mà cụ thể là siêu phẳng chiếu của các biến (ta luôn giả thiết p>>q).

5. BIỂU DIỄN ĐỒNG THỜI

Do tính đối xứng trong phân tích tương ứng nhờ độ đo khi bình phương chúng ta có thể biểu diễn đồng thời các cá thể (các dấu hiệu A_1) và các biến (các dấu hiệu của A_2) trên cùng một siêu phẳng. Để có biểu diễn này, ta chồng khít một cách hình thức k trục chính trong phân tích theo dòng và phân tích theo cột trên cùng một siêu phẳng.

5.1. Hình chiếu của các dòng và các cột trên cùng một siêu phẳng

Biểu diễn hình chiếu của các dòng và các cột trên siêu phẳng chiếu với k thành phần chính đầu tiên sao cho tỷ lệ giải thích đủ lớn là một cách thông thường được sử dụng. Trong thực tế, người ta chỉ chọn k=2 hoặc 3, vì với số chiều như trên các biểu diễn hình học tương đối thuận lợi. Nhờ đó có được hình ảnh trực quan về các dấu hiệu dòng và cột của số liệu ban đầu. Các phân tích cũng dễ dàng hơn khi chúng ta không gian chiếu có số chiều đủ nhỏ.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

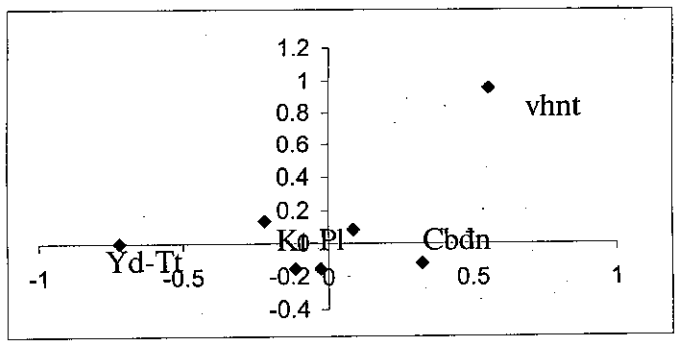
Với những kết quả trên, ta thấy có thể mô tả hình học các điểm dòng và các điểm cột trên cùng một không gian k chiều nào đó. Vì các nhân tố chính của phân tích theo dòng là các thành phần chính trong phân tích theo cột nên ta chỉ cần tiến hành một phân tích theo dòng là đủ. Có thể biểu diễn các điểm dòng và các điểm cột với các toạ độ là các thành phần chính của chúng, các điểm tương ứng với các dòng, cột trên siêu phẳng chiếu cung cấp cho chúng ta hai thông tin chủ yếu đó là:

- Các nhân tố chính có thể mang ý nghĩa chủ yếu của các biến nào.

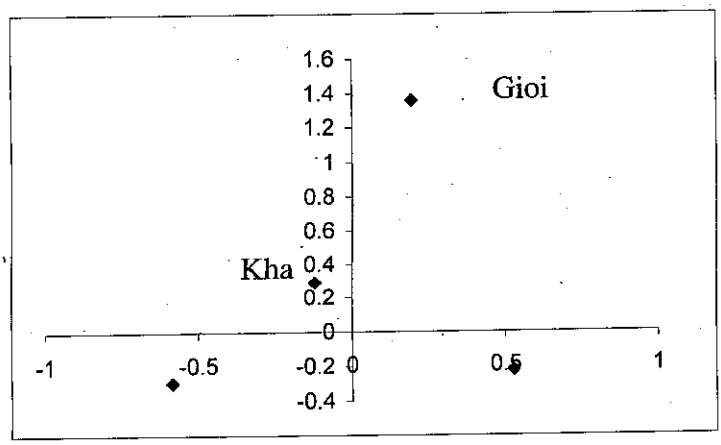
- Quan hệ của các biến dòng và các biến cột nào có thể xem là chặt chẽ và những biến nào có thể xem là gần như độc lập với nhau.

Chúng ta có thể mô tả các phân tích này qua thí dụ trên như sau:

- Các điểm dòng (nhóm trường theo phân loại TN)



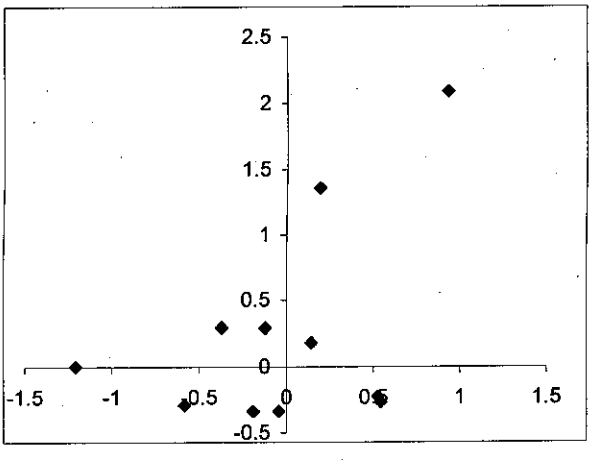
- Các điểm cột (Phân loại TN theo nhóm trường)



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Biểu diễn đồng thời

Thông thường khi phân tích thành phần chính, chúng ta chuẩn hoá các véc tơ riêng để tạo các trục chính vì vậy các kết quả tính toán có thể mô tả không thật chính xác hình ảnh của các điểm dòng và cột trên siêu phẳng chiếu. Đặc biệt là trong trường hợp số điểm dòng quá nhiều hơn số điểm cột và ngược lại. SPSS cung cấp thủ tục cho phép tái hiện hình ảnh không chuẩn hoá hoặc chuẩn hoá theo các độ đo khác nhau (Chúng ta sẽ nghiên cứu ở cuối chương). Cách làm như vậy làm cho việc mô tả đồng thời tốt hơn, sau đây là kết quả mô tả đồng thời không chuẩn hoán từ SPSS:



Một sự tương ứng (và phía sau có thể là sự không tương ứng giữa các nhóm trường và phân loại SVTN) thể hiện rõ ở biểu đồ trên.

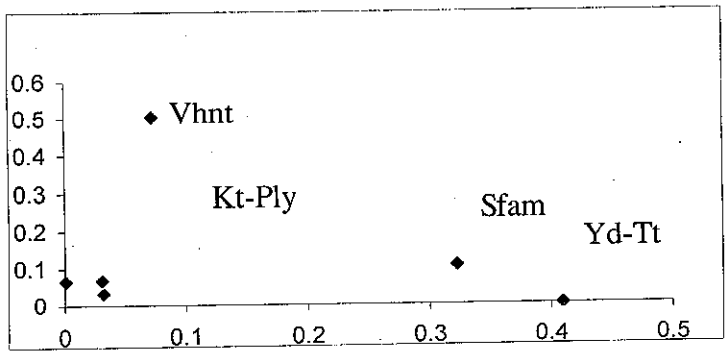
5.2. Mô tả đồng thời mức đóng góp của các điểm dòng, cột đối với các nhân tố chính

Trong phân tích nhân tố, chúng ta đã biết các nhân tố chính hoàn toàn chỉ là các tổ hợp tuyến tính của các véc tơ cần phân tích. Các nhân tố chính vì vậy có thể không có nội dung kinh tế-xã hội. Tuy vậy có thể đánh giá sự dịch chuyển của siêu phẳng chiếu nhờ phân tích mức đóng góp của các tiêu thức cá biệt vào quán tính của từng trục chính như đã nêu ở trên.

Thực tế mô tả đồng thời mức đóng góp vào quán tính của các trục chỉ thực sự cần thiết khi hai tiêu chí dòng và cột tác động tương hỗ. Nhờ các mô tả này có thể thấy cá thể nào hay biến nào thay đổi sẽ làm dịch chuyển siêu phẳng chiếu và cá thể hay biến nào không có tác dụng lớn trong vấn đề này.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hãy xem các biểu đồ vị trí các nhóm trường khi phân tích loại tốt nghiệp:



Theo kết quả này, quan hệ nhóm trường và đánh giá tốt nghiệp chung sẽ biến động nhiều nếu nhóm trường văn hoá nghệ thuật, sư phạm, y-dược và thể dục thể thao thay đổi mức đánh giá của mình. Sự thay đổi của các trường ở các nhóm nông lâm, khoa học cơ bản đa ngành, khoa học công nghệ không tác động nhiều đến kết quả phân tích nói trên. Tuy nhiên nếu căn cứ vào thứ tự các giá trị riêng trong phân tích thành phần chính theo dòng thì các nhóm trường y-dược và thể dục thể thao, sư phạm, kinh tế-pháp lý có vai trò lớn hơn, cơ cấu chính trong phân loại SVTN sẽ thay đổi nhiều nếu các nhóm trường này thay đổi cách đánh giá của mình.

6. ÁP DỤNG PHÂN TÍCH TƯƠNG ỨNG CHO PHÂN TÍCH TƯƠNG QUAN CHÍNH TẮC ĐỐI VỚI HAI BIẾN ĐỊNH TÍNH

Phân tích tương quan chính tắc hay còn gọi là phân tích chính tắc là một trong các kỹ thuật phân tích mở rộng của phân tích tương quan nhiều chiều, nó kết hợp một cách hình thức phân tích hồi qui nhiều chiều với phân tích phân biệt và phân tích tương ứng. Cụ thể là thủ tục này sẽ phân tích tương ứng trên các thành phần chính của các biến nhờ việc quan sát cá thể. Chúng ta sẽ nghiên cứu bài toán loại này trong chương sau, ở đây chúng ta chỉ xét một bài toán đơn giản mà việc phân tích tương quan của hai biến định tính có thể thực hiện tốt nhờ phân tích tương ứng.

6.1. Mô tả định lượng các biến định tính

Cho biến định tính X(1) với m dấu hiệu lập thành nhóm đầy đủ đối với mỗi cá thể của tổng thể. Mỗi cá thể i có thể nhận một và chỉ một dấu hiệu X_{1i}, bảng tiếp liên sau mô tả trạng thái X(1) của n cá thể:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$X(1) = \begin{bmatrix} 1 & 0 & & 0 & 0 \\ 0 & 0 & & 1 & 0 \\ & & & & \\ 0 & & & 0 & 1 \end{bmatrix}$$

Trên bảng n dòng, m cột mỗi cá thể i nhận dấu hiệu j thì ta cho phần tử này bằng 1 các phần tử khác bằng 0. Như vậy tổng một dòng luôn bằng 1, tuy nhiên tổng mỗi cột luôn là số nguyên không âm, thực tế có thể giả thiết tổng mỗi cột là số nguyên dương.

Giả sử có hai biến X(1) có m_1 dấu hiệu và X(2) có m_2 dấu hiệu. Quan sát n cá thể về hai biến này ta có hai ma trận X_1 (n x m_1) và X_2 (n x m_2). Lập ma trận N như sau:

$$N = X_1^T X_2. \qquad (14.14)$$

Bảng N cấp m_1 x m_2 là bảng tần số trong đó n_{ij} là số cá thể có đồng thời hai dấu hiệu X_{1i} và X_{2j}.

Các ma trận D_1, D_2 như trong 13.1 như sau:

$$D_1 = X_1^T X_1; \qquad D_2 = X_2^T X_2. \qquad (14.15)$$

Thí dụ: Cho

X_1			
1	0	0	0
1	0	0	0
0	1	0	0
1	0	0	0
0	0	1	0
0	1	0	0
0	0	1	0
0	0	0	1
0	1	0	0
0	0	0	1
x_1^1	x_1^2	x_1^3	x_1^4

X_2		
0	1	0
1	0	0
0	0	1
1	0	0
0	1	0
0	1	0
1	0	0
1	0	0
0	1	0
0	0	1
x_2^1	x_2^2	x_2^3

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

N			D_1				D_2		
2	1	0	3	0	0	0	4	0	0
0	2	1	0	3	0	0	0	4	0
1	1	0	0	0	2	0	0	0	2
1	0	1	0	0	0	2			

6.2. Mã hoá các biến định tính và phân tích tương ứng

Trước khi tiến hành các phân tích chính tắc cho X(1) và X(2) chúng ta nghiên cứu một cách mã hoá biến định tính, hay tạo một biến định lượng từ một biến định tính m dấu hiệu.

Xét biến X với m dấu hiệu, gọi a_j là giá trị mã hoá tương ứng dấu hiệu j (j=1...m), biến X được mã hoá bởi x theo công thức sau:

$$x = \sum_{j=1}^{m} a_j 1^j$$ trong đó 1^j là véc tơ cột thứ j của X.

Véc tơ x (giá trị số) có n toạ độ xác định như sau:

$$x_i = \sum_{j=1}^{m} a_j 1^j(i) \qquad i = 1..n$$

Như vậy ta có: x = Xa với $a = (a_j)_{m \times 1}$. Véc tơ x là một tổ hợp tuyến tính của các cột của ma trận X, nó thuộc R^n như các cột 1^j của X.

Chúng ta có thể hiểu đơn giản cách mã hoá này thuần tuý là hình thức hoá toán học, tuy nhiên đôi khi có thể tạo véc tơ x như một phép chiếu trong R^n.

Xét mô hình sau:

Giả sử mỗi thành viên của một tổ n người chỉ có thể chọn 1 trong 2 hình thức nghỉ cuối tuần là: tham quan thành cổ (A_1), nghỉ biển (A_2). Các thành viên có thể đi cùng là một trong các đối tượng sau: vợ (chồng) (B_1), con (B_2), không ai cả (B_3).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tổng hợp đăng ký nghỉ cuối tuần là bảng sau:

STT	A_1	A_2	B_1	B_2	B_3
1	1	0	1	0	0
2	1	0	0	1	0
3	0	1	0	0	1
n	0	1	0	1	0

Nếu ta gán cho A_1, A_2 các chi phí hỗ trợ du lịch; B_1, B_2, B_3 các mức phức tạp của bản thân thành viên trong tổ khi có hoặc không có người cùng đi ta có hai biến x và y mã hoá A và B như sau:

x	y
a_1	b_1
a_1	b_2
a_2	b_3
a_n	b_n

hệ số tương quan của x và y cho chúng ta quan hệ của hai biến định tính A và B, bạn có thể chọn a và b sao cho hệ số này tối thiểu hoặc tối đa với những ràng buộc có thể.

Thí dụ: ta có 10 cá thể với lựa chọn và chi phí tương ứng với A và B như sau:

STT	A_1	A_2	B_1	B_2	B_3	x	y
1	1	0	1	0	0	4	3.00
2	1	0	0	1	0	4	2.00
3	0	1	0	0	1	3	1.00
4	1	0	1	0	0	4	3.00
5	0	1	0	1	0	3	2.00
6	0	1	0	1	0	3	2.00
7	1	0	1	0	0	4	3.00
8	0	1	0	0	1	3	1.00
9	0	1	0	1	0	3	2.00
10	1	0	1	0	0	4	3.00
	4	***3***	***3***	***2***	***1***		

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hệ số tương quan của x (chi phí hỗ trợ) và y (độ phức tạp của chuyến đi): R(x,y)= 0,801784. Điều này cho thấy những người muốn đi thăm thành cổ thì có xu thế đi cùng vợ (chồng) hoặc con, còn những người muốn đi biển thì có vẻ thích đi một mình hơn.

Về mặt toán học có thể kiểm tra thấy rằng với biến hai dấu hiệu thì các giá trị a_i chỉ có tính thứ tự, còn với cấc biến hơn hai dấu hiệu thì các giá trị b_i chỉ có ý nghĩa ở tỷ lệ sai khác của chúng đối với các dấu hiệu khác nhau mà thôi. Chúng ta có thể minh họa bởi bộ (a;b) = (10, 4; 10,6,2).

6.3. Phân tích chính tắc với hai nhóm chỉ số

Hai biến X(1), X(2) nói trên thực chất là hai nhóm chỉ số. Phân tích chính tắc tìm siêu phẳng chiếu sao cho hệ số tương quan của hai biến x_1 và x_2 trên siêu phẳng này lớn nhất. Theo cách hình thức hoá ở phần 1, chúng ta có thể lập các ma trận sau:

$$V_{11} = D_1/n; \quad V_{22} = D_2/n; \quad V_{12} = N/n; \quad V_{21} = N^T/n$$

Các nhân tố chính tắc của X_1 là các véc tơ riêng của:

$$D_1^{-1}ND_2^{-1}N^T = V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}$$

Đó là:

0.417	0.167	0.25	0.167
0.167	0.5	0.167	0.167
0.375	0.25	0.25	0.125
0.25	0.25	0.125	0.375

Các giá trị riêng: {1.0006, 0.317163, 0.224164, 0.0000731401}

Các véc tơ riêng là:

(0.500241, 0.5003, 0.499769, 0.49969),

(-0.58678, 0.7250, -0.336465, 0.12941),

(-0.08657, -0.2811, -0.34168, 0.89259),

(0.49956, 0.16723, -0.8334, -0.16675)

Các véc tơ này cho biết toạ độ trên các trục chính trong phân tích chính tắc và người ta gọi là các véc tơ hay các **nhân tố chính tắc**.

Như vậy nhân tố chính tắc của X(1) là một thành phần chính trong phân tích nhân tố theo dòng của N với một hệ số nhân nào đó.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hoàn toàn tương tự đối với X(2), các nhân tố chính tắc xác định được nhờ các véc tơ riêng của $D_2^{-1}N^TD_1^{-1}N^T=V_{22}^{-1}V_{21}\ V_{11}^{-1}V_{12}$. Các giá trị riêng tương ứng chính là bình phương hệ số tương quan chính tắc, giá trị riêng tầm thường có giá trị 1 tương ứng với hai véc tơ lập từ tổng các cột của X(1) và X(2).

Các nhân tố chính tắc là tổ hợp tuyến tính của các biến chỉ số cấu thành X(1) và X(2), vì vậy với hai biến định lượng X^1, X^2 ta có thể giải thích phân tích tương ứng như việc tìm một phép biến đổi đồng thời cho cả hai biến này thành các biến bằng số có hệ số tương quan lớn nhất.

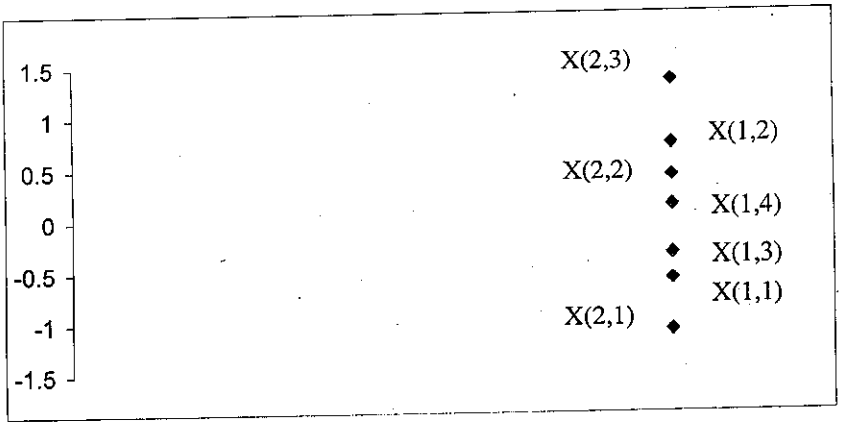
Thực tế như đã xét ở phần 2, các công thức chuyển đổi cho phép tìm các véc tơ nhân tố chính tắc của nhóm này từ nhân tố chính tắc của nhóm kia. Chẳng hạn nếu ta chọn nhân tố chính tắc thứ hai của phân tích cho X(1):

a^2= (-0.586781, 0.725074, -0.336465, 0.12941), nhờ (8.12) ta sẽ tính được nhân tố chính tắc thứ hai của phân tích cho X(2) là:

b^2=(-1.089, 0.415, 1.346).

tương tự cho các nhân tố chính khác.

Các dấu hiệu của X(1) và X(2) được mô tả trên một trục chính như sau:



Bằng trực quan chúng thấy các dấu hiệu thứ hai và thứ 4 của X(1) quan hệ chặt chẽ với dấu hiệu thứ hai của X(2) và lỏng hơn với hai dấu hiệu còn lại của X(2); dấu hiệu thứ nhất của X(2) ảnh hưởng nhiều hơn đến dấu hiệu thứ nhất và thứ ba của X(1). Như vậy qua X(1) chúng ta có thể phỏng đoán được một phần X(2) và ngược lại (chú ý rằng các phỏng đoán này chỉ dựa trên quan hệ về sự xuất hiện dấu hiệu mà không liên quan gì đến quan hệ nhân quả (nếu có) của hai dấu hiệu định tính nói trên).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6.4. Biểu diễn đồng thời tối ưu các dấu hiệu của các cá thể

Trong phần này chúng ta xem xét cách thức ghép X(1) và X(2) thành một biến có m_1+m_2 dấu hiệu và việc lợi dụng phân tích chính tắc để mô tả mỗi cá thể theo cả hai biến định tính này. Những kết quả thu được trong mục này làm rõ hơn các kết quả đã xét ở các mục trước, đồng thời tạo cơ sở cho các phân tích chính tắc đồng thời được trình bày ở chương sau.

Các dấu hiệu của X(1) và X(2) xác định các nhóm cá thể số lượng n_i ($i=1...m_1$) và m_j ($j=1...m_2$). Nếu chúng ta sử dụng một biến z có trung bình bằng không để biểu diễn các toạ độ của n cá thể trên một trục nào đó ta cũng sẽ biểu diễn dấu hiệu thứ i của X(1) chẳng hạn bởi một điểm mà toạ độ a_i của nó là trung bình của các toạ độ của n_i cá thể tương ứng, a_i có thể xác định theo công thức:

$$a_i = \frac{1}{n_i}\sum_{k=1}^{n} z_k 1^i(k) = \frac{1}{n_i} z^T x_1^i \qquad (14.16)$$

Hãy xem mô tả sau cho thí dụ đang xét: (bảng trang sau)

$a_1 = (z_1+z_2+z_4)/3$; $a_2 = (z_3+z_6+z_9)/3$; $a_3 = (z_5+z_7)/2$; $a_4 = (z_8+z_{10})/2$;...

ta có: $a = D_1^{-1} X_1^T z$

$$a = \begin{bmatrix} 0.333 & 0.333 & 0 & 0.333 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.333 & 0 & 0 & 0.333 & 0 & 0 & 0.333 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 \end{bmatrix} z$$

X_1					X_2		
1	2	3	4		1	2	3
1	0	0	0	z_1	0	1	0
1	0	0	0	z_2	1	0	0
0	1	0	0	z_3	0	0	1
1	0	0	0	.	1	0	0
0	0	1	0	.	0	1	0
0	1	0	0	.	0	1	0
0	0	1	0	.	1	0	0
0	0	0	1	.	1	0	0
0	1	0	0	.	0	1	0
0	0	0	1	z_{10}	0	0	1
x_1^1	x_1^2	x_1^3	x_1^4		x_2^1	x_2^2	x_2^3

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Hoàn toàn tương tự với các dấu hiệu của X(2) ta có thể lập véc tơ:

$$b = D_2^{-1}X_2^T z.$$

Việc sử dụng biến z cho phép chúng ta tách các a_i (các toạ độ của các dấu hiệu) một cách tuỳ ý. Tức là phương sai của a (a^TD_1a/n) sẽ có thể rất lớn. Giá trị lớn nhất của phương sai này đạt được nếu tất cả các cá thể có cùng dấu hiệu trên X(1) nhận cùng một giá trị của z.

Chúng ta sẽ tìm véc tơ z và các véc tơ a, b sao cho: trung bình của a^TD_1a và b^TD_2b đạt giá trị lớn nhất. Như vậy cùng lúc ta sẽ biểu diễn tối ưu tất cả các dấu hiệu của cả hai biến X(1) và X(2) trên cùng một trục.

Vì: $a^TD_1a = z^TX_1(X_1^TX_1)^{-1}X_1z$ và $b^TD_2b = z^TX_2(X_2^TX_2)^{-1}X_2z$

đặt $A_1=X_1(X_1^TX_1)^{-1}X_1$ và $A_2= X_2(X_2^TX_2)^{-1}X_2$

thì: $a^TD_1a=z^TA_1z$ và $b^TD_2b= z^TA_2z$.

Trung bình của a^TD_1a và b^TD_2b trở thành $[z^T(A_1+A_2)z]/2$. Vậy ta phải tìm z cực đại giá trị này.

Với mỗi z có phương sai V(z) xác định thì $[z^T(A_1+A_2)z]/2$ lớn nhất khi z là véc tơ riêng của $(A_1+ A_2)/2$. Nếu ký hiệu μ là giá trị riêng thì ta có phương trình sau:

$$(A_1+A_2)z=2\mu z$$

Thay $A_1z= X_1D_1^{-1}X_1^Tz=X_1a$ và $A_2z= X_2D_2^{-1}X_2^Tz=X_2b$

ta có: $X_1a +X_2b=2\mu z$. Từ đó suy ra:

$$D_1^{-1}X_1^TX_1a + D_1^{-1}X_1^TX_2b = D_1^{-1}X_1^T2\mu z$$

hay: $a + D_1^{-1}X_1^TX_2b =2\mu a$ (vì theo 8.15: $D_1 =X_1^TX_1$) (14.17)

$$a + D_1^{-1}Nb=2\mu a \quad (\text{vì theo 8.14: } N =X_1^TX_2)$$

$$D_1^{-1}Nb =(2\mu-1)a \qquad (*) \qquad (14.18)$$

Thực hiện các biến đổi tương tự ta có:

$$D_2^{-1}N^Ta =(2\mu-1)b \qquad (**) \qquad (14.19)$$

Từ (*) và (**) ta có: $D_1^{-1}N\, D_2^{-1}N^Ta =(2\mu-1)^2a$ (14.20)

$$D_2^{-1}N^T\, D_1^{-1}Nb =(2\mu-1)^2b \qquad (14.21)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trở lại với mục 3 ta có các giá trị riêng của $D_I^{-1}N\ D2^{-1}N^T$ là λ thì:

$$\lambda = (2\mu - 1)^2.$$

Như vậy ngoài việc phân tích chính tắc theo dòng và theo cột như trong mục 3 ta có thể sử dụng véc tơ z như một sự kết hợp của hai véc tơ chính tắc riêng biệt. Các toạ độ của các điểm (các dấu hiệu) được xác định trên một trục nhờ phân tích tương ứng tối ưu. Các trục tiếp theo có thể chọn nhờ các biến z ứng với các giá trị riêng nhỏ hơn, vuông góc với trục thứ nhất... .Việc mô tả đồng thời hai dấu hiệu định tính nhờ phân tích tương ứng theo cách này, đã làm cho các dấu hiệu của các biến được đặt ở trung tâm của nhóm các cá thể mang các dấu hiệu đó.

Việc mô tả các cá nhân có dấu hiệu i của X(1) và dấu hiệu j của X(2) trên trục chính thứ s (tương ứng giá trị riêng thứ s), ở toạ độ sau:

$$r^s(i,j) = (a^s{}_i + b^s{}_j)/2\mu = (a^s{}_i + b^s{}_j)/(1 + \lambda_s{}^{0.5}).$$

Sau đây là các tính toán trên cơ sở phân tích bảng N cho ở mục trước với hai biến định tính. Biến thứ nhất có 4 dấu hiệu và biến thứ hai có 3 dấu hiệu. Một số thông tin cần thiết đã được tính ở mục trên, ở đây chúng ta chỉ tìm các toạ độ của các cá thể và mô tả chúng trên biểu đồ.

Chẳng hạn với giá trị riêng thứ 2 (λ_2 =0.317163) ta có a^2 và b^2 được tính ở trên là:

a^2	b^2
-0.58678	-1.08858
0.725074	0.415324
-0.33647	1.346428
0.129	

Ta có r^2:

X^2 \ X^1	1	2	3	4
1	-1.07177	-0.23254	-0.91164	-0.61387
2	-0.10969	0.729541	0.050448	0.348218
3	0.485965	1.325191	0.646098	0.943868

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với giá trị riêng thứ 3 (0.221464)

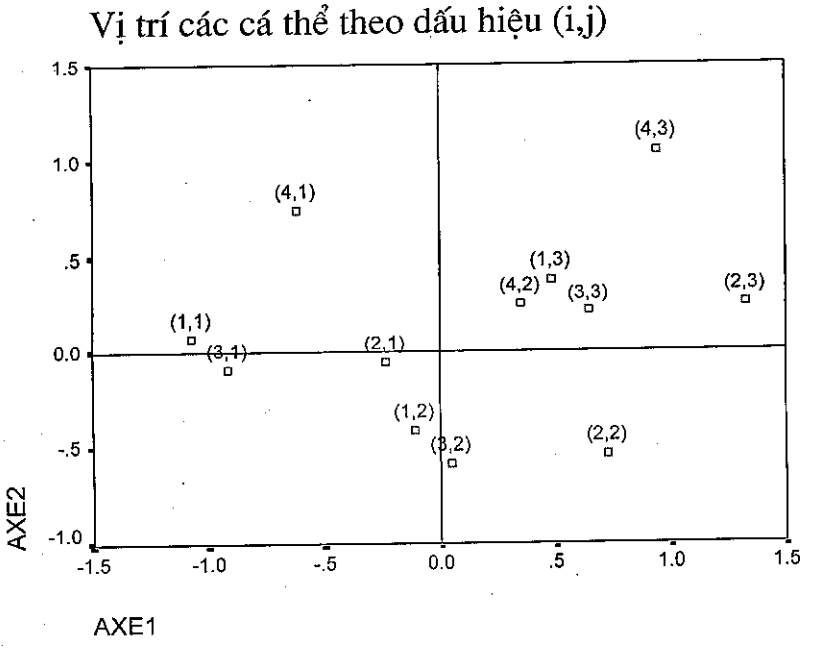
ta có a^3= (-0.0865747, -0.281125, -0.341683, 0.892595)

và b^3 = (0.200682 , -0.52619, 0.649671)

Ta có r^3:

X^2 \ X^1	1	2	3	4
1	0.077592	-0.0547	-0.09588	0.743422
2	-0.41668	-0.54897	-0.59015	0.24915
3	0.382903	0.250609	0.20943	1.048733

Biểu đồ các cá thể trên hai trục chính như sau:



6.5. Mô tả đồng thời các dấu hiệu

Trên hai trục các dấu hiệu của X(1) và X(2) có thể mô tả với các toạ độ như sau:

Dấu hiệu thứ i của X(1), ký hiệu A_i, có toạ độ: (a^2_i, a^3_i)

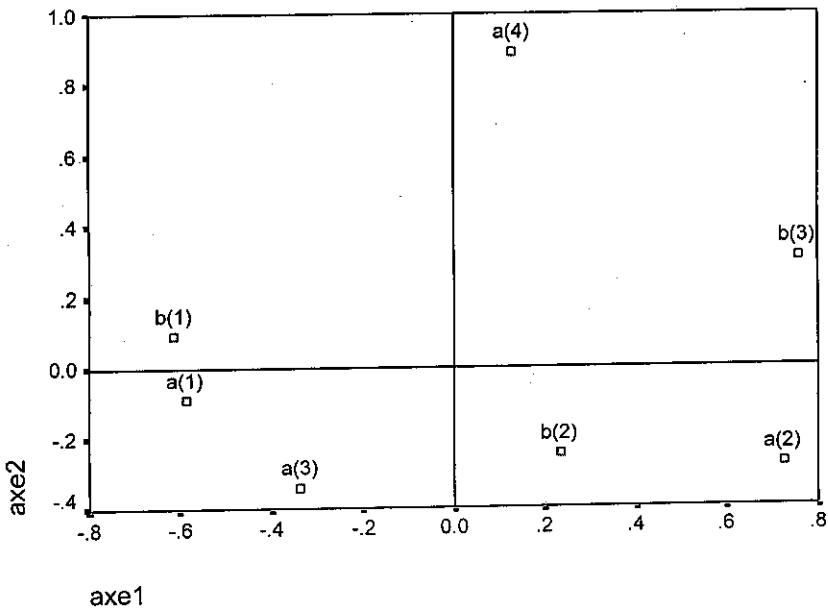
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Dấu hiệu thứ j của X(2), ký hiệu B_j, có toạ độ: ($\sqrt{\lambda_2} b_j^2$, $\sqrt{\lambda_3} b_j^3$) .

Với thí dụ trên ta có các toạ độ này là:

	axe1	axe2
A(1)	-0.58678	-0.08657
A(2)	0.725074	-0.28113
A(3)	-0.33647	-0.34168
A(4)	0.129	0.8926
B(1)	-0.61306	0.094441
B(2)	0.233899	-0.24762
B(3)	0.758271	0.305735

Biểu đồ các dấu hiệu



Ngoài ra có thể thay thế biểu đồ trên bằng biểu đồ hệ số tương quan của các hình chiếu của các dấu hiệu với các trục.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trên hai trục các dấu hiệu của X(1) và X(2) theo hệ số tương quan với các trục có thể mô tả với các toạ độ như sau:

Dấu hiệu thứ i của X(1), ký hiệu A_i, có toạ độ: $(\sqrt{\lambda_2}\,a_i^2, \sqrt{\lambda_3}\,a_i^3)$.

Dấu hiệu thứ j của X(2), ký hiệu B_j, có toạ độ: (b^2_j, b^3_j).

6.6. Mô tả đồng thời theo hai độ đo

Khi mô tả các dấu hiệu theo hai trục nhờ các toạ độ, chúng ta thấy rằng nếu dấu hiệu của X(1) được mô tả nhờ độ đo khi bình phương thì dấu hiệu của X(2) phải mô tả nhờ độ đo hệ số tương quan. Tính đối xứng trong phân tích tương ứng gây nên sự đấn đo khi dùng cả hai mô tả này để phân tích. Gittins (R.) đề nghị một giải pháp sử dụng một hệ toạ độ trung gian, toạ độ trên trục ứng với giá trị riêng thứ k của mỗi dấu hiệu được tính theo công thức sau:

Với A_i: $\frac{(1+\sqrt{\lambda_k})}{2} a_i^k$; với B_j: $\frac{(1+\sqrt{\lambda_k})}{2} b_j^k$.

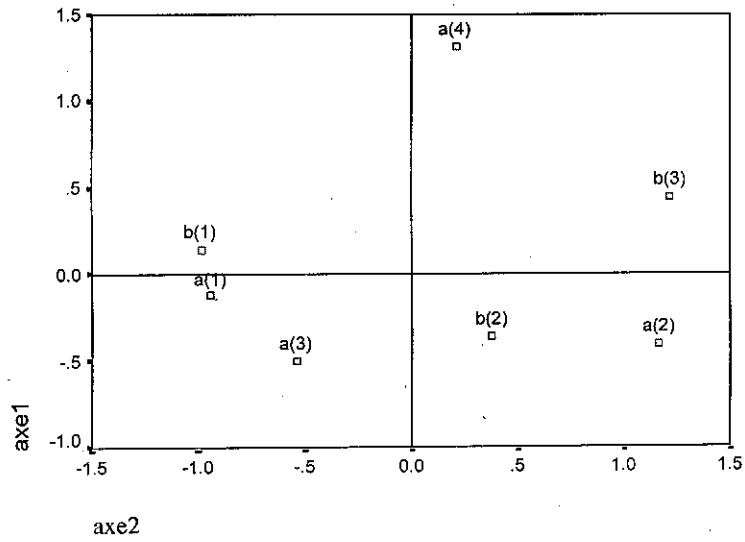
Ta gọi biểu đồ với các toạ độ như trên là biểu đồ dạng MAMC

Với thí dụ trên các toạ độ tính theo hai trục đang xét là:

	axe1	axe2
A(1)	-0.94449	-0.12732
A(2)	1.16709	-0.41342
A(3)	-0.54159	-0.50248
A(4)	0.20764	1.31265
B(1)	-0.98679	0.138885
B(2)	0.376487	-0.36416
B(3)	1.220524	0.449614

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

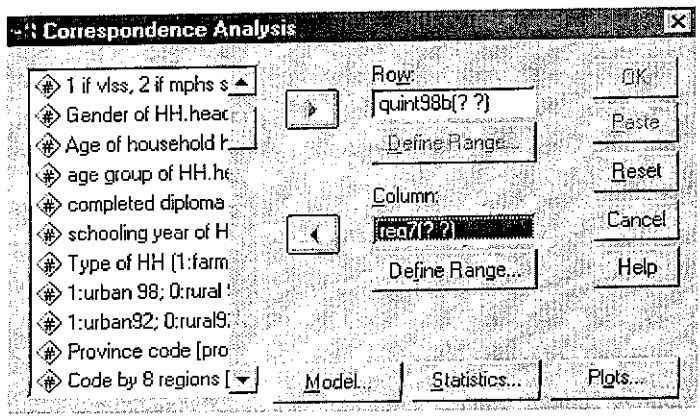
Biểu đồ các dấu hiệu -MAMC



7. PHÂN TÍCH TƯƠNG ỨNG TRÊN SPSS

SPSS cung cấp thủ tục phân tích tương ứng ở thực đơn Analyze (hoặc Statistic). Cũng như các thủ tục khác phân tích tương ứng trên SPSS quá phong phú, ở đây chúng ta chỉ tìm hiểu những nội dung chính của thủ tục này. Tệp số liệu minh hoạ là tệp chi tiêu hộ của VLSS98.

Để bắt đầu phân tích tương ứng với một tệp số liệu chúng ta chọn mục Data Reduction trong Analyze sau đó chọn Correspondence analysis. Thực đơn của thủ tục này như sau:

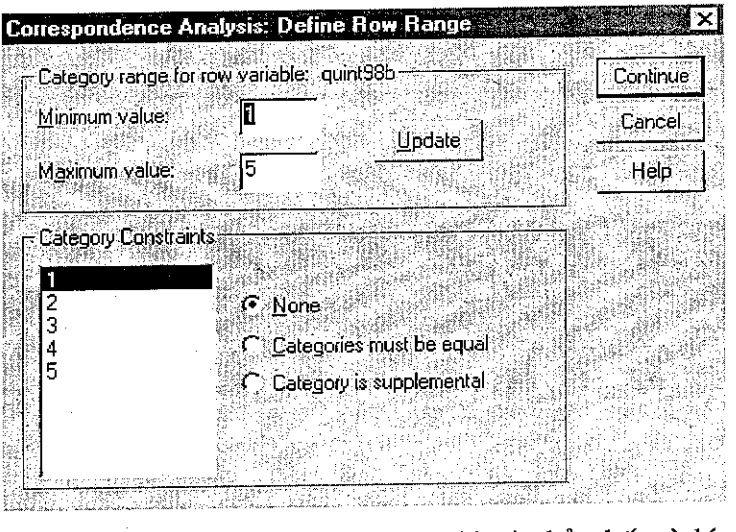


THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Trước hết ta lựa chọn hai biến nào đó để tiến hành phân tích tương ứng, chẳng hạn nếu chúng ta muốn phân tích quan hệ tương ứng của mức chi tiêu (coi như một đặc trưng đo mức giàu nghèo) của các hộ dân cư và chỉ báo vùng kinh tế (7 vùng) ta có thể chọn biến quin98b làm biến dòng và biến reg7 làm biến cột cho bảng chọn trên.

Mỗi biến dòng (cột) cần chọn vùng phân tích của chúng nhờ các nút Define Range..

Chẳng hạn, khi kích nút này ta khai báo trong bảng như sau:



Trong Define Row Range ta khai báo giá trị nhỏ nhất và lớn nhất của biến Quin98b là 1 và 5 vì trong tệp số liệu này, người ta chia mức tiêu dùng hộ theo ngũ phân vị. Cũng trong bảng khai báo này, chúng ta có thể khai báo các ràng buộc (như các giả thiết), tuy nhiên việc khai báo này thường được làm ở các phân tích sau phân tích toàn bộ (không ràng buộc) này. Kết thúc khai báo này nhờ chọn nút Continue. Tương tự ta chọn nút Define range cho biến reg7 và khai báo các giá trị nhỏ nhất là lớn nhất là 1 và 7.

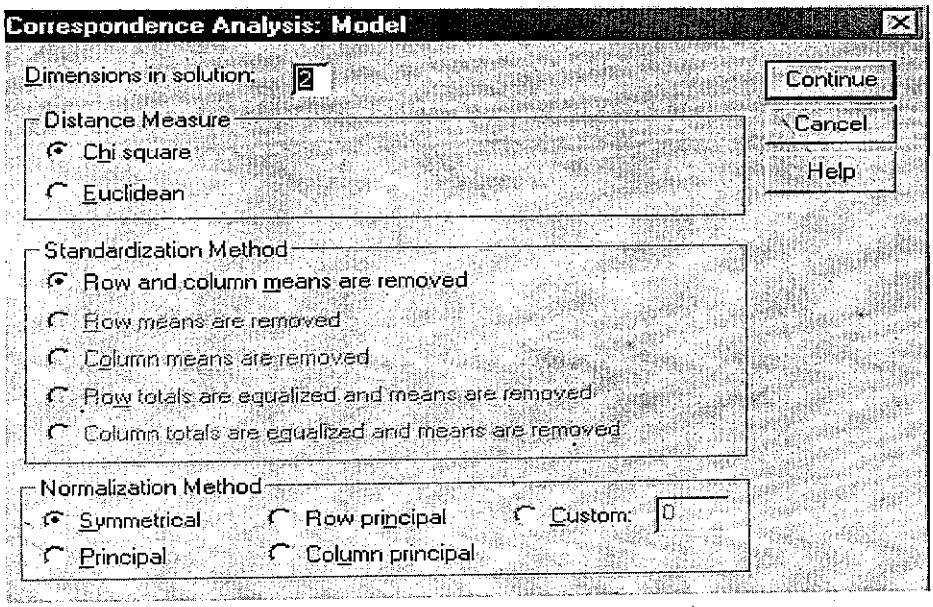
Kết thúc khai báo này, chúng ta có thể chọn tiếp các mục Model, Statistic và Plot. Sau đây là các bảng chọn ở lựa chọn mặc định (default) và các giả thích tương ứng cho các lựa chọn (option).

Bảng chọn Model

- Cửa sổ Dimensions in solution: cho phép chọn số nhân tố chính trong phân tích (mặc định là 2).

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Distance Measure (độ đo khoảng cách): Cho phép chọn một trong hai độ đo khi bình phương hoặc Ơcơlit. Lựa chọn mặc định là khi bình phương như đã trình bày trong bài 1 và 2, tuy nhiên nếu chúng ta phân tích tương ứng chỉ với hai biến cùng thang bậc và mỗi cá thể cùng dấu hiệu trên mỗi cặp biến có trọng số như nhau thì có thể lựa chọn độ đo Ơcơlit.

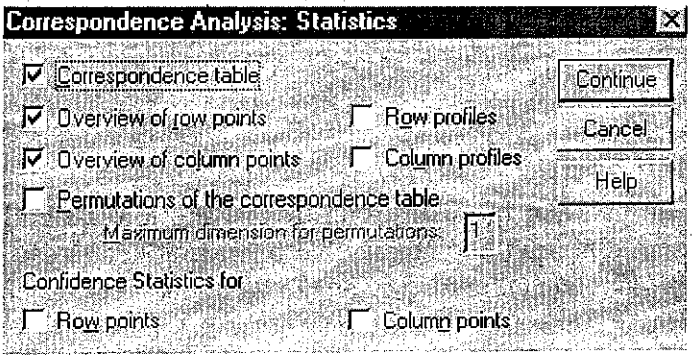


- Standardizaton Method: Với lựa chọn độ đo khi bình phương, chúng ta chỉ có một cách chọn đó là trung tâm hoá cả dòng và cột, như vậy nhân tố tầm thường tương ứng với hai trục chính là các véc tơ trong tâm g_d và g_c bị bỏ qua. Nếu chọn độ đo Ơcơlit, chúng ta có thể chọn các cách thức chuẩn hoá khác nhau theo cách phân tích theo dòng thì trung tâm hoá còn phân tích theo cột thì không hoặc ngược lại, ngoài ra có thể chọn cách phân tích bỏ qua cả hai véc tơ trung tâm của dòng và cột.

- Normalization Method: Chúng ta có thể chọn các phương pháp chuẩn hoá khác nhau (mặc định là Symetrical). Thực chất đây là các phương pháp lựa chọn trọng số khi chuẩn hoá các toạ độ chiếu của các cá thể cũng như các biến sau phân tích. Các giải thích tương ứng cho các lựa chọn có thể xem trực tiếp nhờ kích chuột phải vào lựa chọn này.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bảng chọn Statistic...

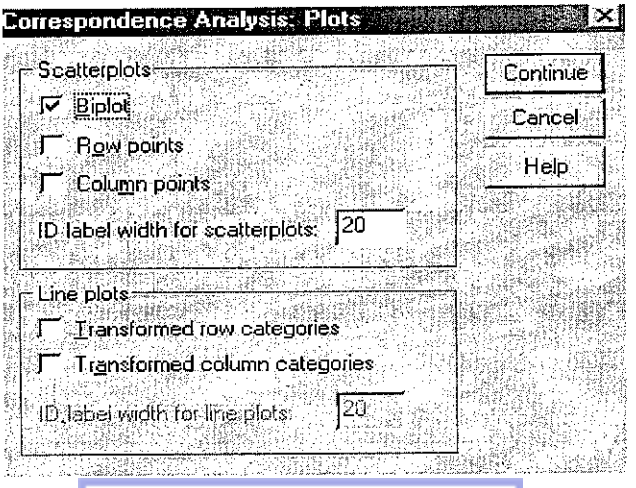


Bảng chọn này cho phép hiển thị các thống kê khác nhau trong kết quả, các lựa chọn không loại trừ nhau. Bảng trên là cách lựa chọn mặc định nó yêu cầu hiển thị bảng tiếp liên (N), các điểm dòng và cột. Nếu hài lòng với lựa chọn này, ta có thể chọn Continue để về bảng chọn chính của thủ tục này.

Bảng chọn Plots...

Bảng chọn này cho phép hiển thị các biểu đồ đối với các điểm dòng và cột riêng biệt hay đồng thời dạng Scatter plots.

Ngoài ra chúng ta có thể lựa chọn đồ thị dạng đường, loại đồ thị này cho chúng ta biểu diễn thành phần chính của biến dòng theo các dấu hiệu của biến này hoặc thành phần chính của biến cột theo các dấu hiệu của chúng.



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với các lựa chọn như trên ta có chương trình tương ứng bằng ngôn ngữ SPSS và kết quả phân tích trên SPSS như sau:

```
CORRESPONDENCE
  TABLE = quint98b(1 5)  BY reg7(1 7)
  /DIMENSIONS = 2  /MEASURE = CHISQ  /STANDARDIZE = RCMEAN
  /NORMALIZATION = SYMMETRICAL  /PRINT = TABLE RPOINTS CPOINTS
  /PLOT = NDIM(1,2) BIPLOT(20) .
```

Thông tin ban đầu rất quan trọng trong bảng tóm tắt chính là mức ý nghĩa của kiểm định tính độc lập của hai biến đang xét và các giá trị riêng được sắp xếp giảm dần cùng tổng tỷ lệ giải thích được khi lựa chọn số thành phần chính tương ứng. Cột hệ số tương quan cho thấy hai trục chính gần như độc lập với nhau nếu ta chọn hai véc tơ riêng cho phân tích này.

Overview Row Points

best quintile for 1998	Mass	Score in Dimension 1	Score in Dimension 2	Inertia	Contribution: Of Point to Inertia of Dimension 1	Contribution: Of Point to Inertia of Dimension 2	Contribution: Of Dimension to Inertia of Point 1	Contribution: Of Dimension to Inertia of Point 2	Contribution: Tota l
nghèo nhất	.153	-.839	.636	.051	.266	.523	.852	.143	.996
nghèo	.166	-.543	-.166	.023	.121	.039	.864	.024	.888
trung bình	.194	-.222	-.378	.008	.024	.235	.503	.428	.931
giàu	.220	.047	-.237	.003	.001	.104	.073	.552	.626
giàu nhất	.266	.944	.210	.098	.588	.099	.985	.014	.999
Active Total	1.000			.182	1.000	1.000			

a. Symmetrical normalization

Overview Column Points[a]

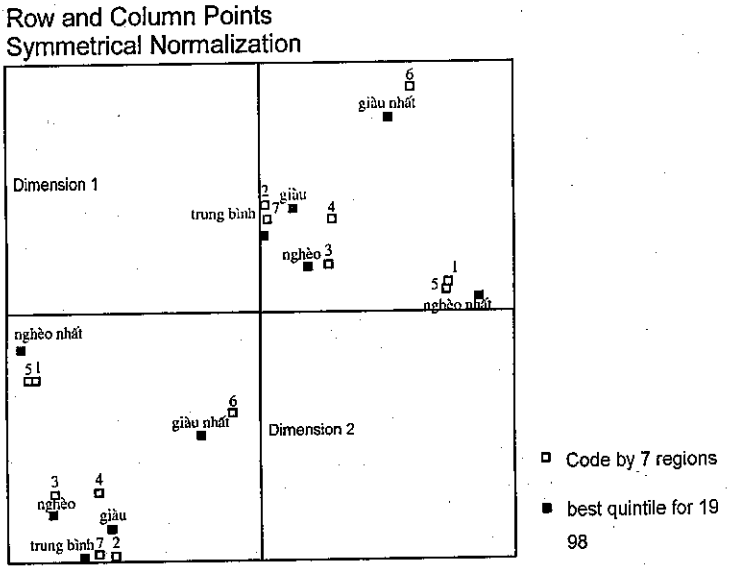
Code by 7 regions	Mass	Score in Dimension 1	Score in Dimension 2	Inertia	Contribution: Of Point to Inertia of Dimension 1	Contribution: Of Point to Inertia of Dimension 2	Contribution: Of Dimension to Inertia of Point 1	Contribution: Of Dimension to Inertia of Point 2	Contribution: Total
1	.143	-.695	.491	.033	.171	.292	.856	.125	.981
2	.196	.086	-.365	.004	.004	.221	.149	.781	.930
3	.118	-.514	-.070	.013	.077	.005	.958	.005	.963
4	.126	-.074	-.059	.003	.002	.004	.099	.019	.118
5	.061	-.766	.486	.017	.089	.122	.869	.102	.970
6	.171	1.247	.321	.109	.655	.148	.981	.019	1.000
7	.185	-.071	-.364	.003	.002	.208	.112	.867	.979
Active Total	1.000			.182	1.00	1.000			

a. Symmetrical normalization

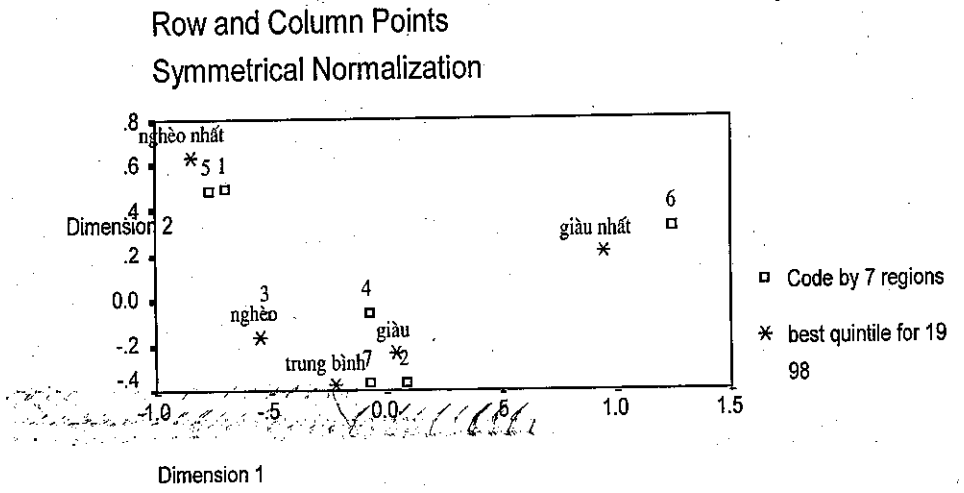
THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các bảng trên cho các phân tích thành phần chính riêng biệt theo dòng và theo cột.

Biểu đồ sau cho vị trí đồng thời của các dấu hiệu của hai biến trên cùng mặt phẳng chiếu.



Biểu đồ này có thể được cải thiện nhờ việc lựa chọn lại cách thể hiện (xử lý trên màn hình output). Biểu đồ sau nhận được từ biểu đồ trên với việc lựa chọn lại dạng biểu đồ, phông chữ .



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Để thấy sự khác nhau khi sử dụng các độ đo khác nhau, chúng ta có thể thực hiện đoạn chương trình sau (dành cho người học):

```
CORRESPONDENCE
  TABLE = quint98b(1 5)  BY reg7(1 7)  /DIMENSIONS = 2
/MEASURE = EUCLID  /STANDARDIZE = RCMEAN  /NORMALIZATION =
SYMMETRICAL  /PRINT = TABLE RPOINTS CPOINTS  /PLOT =
NDIM(1,2) BIPLOT(20) RPOINTS(20) CPOINTS(20) TRROWS(20)
TRCOLUMNS (20).
```

Những nội dung cơ bảng trong chương này đòi hỏi người sử dụng phải kiên trì kết hợp khai thác lý thuyết và làm việc với một bảng tính kèm theo. Chúng ta có thể lợi dụng SPSS tìm các giá trị riêng và các véc tơ riêng, các phân tích khác nếu được thực hiện trên bảng tính (chẳng hạn là excel) sẽ làm sáng tỏ hơn nội dung của chương này. Chúng tôi cho rằng, một khi các thủ tục chưa được tiến hành từng bước trên excel hay một bảng tính nào đó có tính chất hỗ trợ tư duy thì việc dùng SPSS có thể sẽ trở nên máy móc, dù chúng ta có thu được kết quả đủ để phân tích.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

8.TÓM TẮT CHƯƠNG 14

Chương 14 đề cập đến các nội dung sau:

1. Các vấn đề kinh tế xã hội dẫn đến bài toán phân tích tương ứng
2. Cách thức mô hình hóa các biến dấu hiệu trong bài toán tìm sự tương ứng và công cụ mô tả dữ liệu trong các trường hợp khác nhau, với mục đích có thể sử dụng phân tích nhân tố giải quyết bài toán phân tích tương ứng.
3. Mối liên hệ hình thức của cá thể và các đặc trưng trong bài toán phân tích tương ứng
4. Khoảng cách và quán tính trong phân tích tương ứng, cách sử dụng các độ đo trong các trường hợp cụ thể.
5. Chuyển bài toán phân tích tương ứng theo dòng và theo cột về các bài toán phân tích nhân tố bằng phương pháp thành phần chính đã trình bày ở chương 13.
6. Kỹ thuật phân tích tương ứng trên SPSS
7. Các công thức chuyển đổi đối với các thành phần chính theo dòng và theo cột
8. Mô tả đồng thời các cá thể và các đặc trưng trên cùng một mặt phẳng chiếu
9. Phân tích kết quả qua hệ số tương quan và biểu đồ mô tả đồng thời
10. Áp dụng phân tích tương ứng cho bài toán phân tích tương quan.
11. Các kỹ thuật mô tả đồng thời theo hai độ đo đối với các cá thể và các đặc trưng.
12. Cách thức phân tích kết quả

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

CÁC KÝ HIỆU VÀ CÔNG THỨC CƠ BẢN

- Ma trận tần số theo các đặc trưng: $N = (n_{ij})_{pxq}$
- Ma trận đường chéo các tần số biên:

$$D_1 = \begin{pmatrix} n_1 & 0 & \ldots\ldots & 0 \\ 0 & n_2 & \ldots\ldots\ldots & 0 \\ \ldots\ldots\ldots & \ldots\ldots & \ldots\ldots & \\ 0 & 0\ldots\ldots\ldots & \ldots . & n_p \end{pmatrix} \qquad D_2 = \begin{pmatrix} m_1 & 0 & \ldots\ldots & 0 \\ 0 & m_2 & \ldots\ldots\ldots & 0 \\ \ldots\ldots\ldots\ldots & \ldots\ldots & & \\ 0 & 0\ldots\ldots\ldots\ldots & & m_q \end{pmatrix}$$

- Ma trận tỷ lệ dòng:

$$D_1^{-1}N = \begin{bmatrix} \frac{n_{11}}{n_1} & \frac{n_{12}}{n_1} & \ldots\ldots & \frac{n_{1q}}{n_1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & \\ \frac{n_{p1}}{n_p} & \frac{n_{p2}}{n_p} & \ldots\ldots\ldots & \frac{n_{pq}}{n_p} \end{bmatrix}$$

Mỗi dòng i trong ma trận trên có các thành phần là tần suất có điều kiện các dấu hiệu của B theo A xác định tại A_i: $P(B=B_j/(A=A_i)) = n_{ij}/n_i$.

- Khoảng cách giữa hai điểm i và k trong không gian các véc tơ dòng q chiều của bảng N, $d_i = (n_{i1}, n_{i2}, \ldots, n_{iq})$ và $d_k = (n_{k1}, n_{k2}, \ldots, n_{kq})$:

$$d^2(i,k) = \|x_i - x_k\|^2 = \sum_{j=1}^{q} \frac{1}{q_j}(x_{ij} - x_{kj})^2$$

- Cấu trúc hai bài toán phân tích thành phần chính:

Phân tích thành phần chính theo dòng với bảng số liệu $X_r = D_1^{-1}N$;

Metric $M_r = nD_2^{-1}$ và ma trận trọng số $D_d = D_1/n$.

Phân tích thành phần chính theo cột với bảng số liệu Xc= D2-1N;

Metric $M_c = nD_1^{-1}$ và ma trận trọng số $D_c = D_2/n$.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Cấu trúc lời giải của các bài toán:

	ACP các véc tơ dòng	ACP các véc tơ cột
Các nhân tố chính	Véc tơ riêng của $MX^TDX=D_2^{-1}N^T D_1^{-1}N$	Véc tơ riêng của $XMX^TD=D_1^{-1}N D_2^{-1}N^T$
Các thành phần chính	Véc tơ riêng của $XMX^TD=D_1^{-1}N D_2^{-1}N^T$ Chuẩn hoá bởi a sao cho: $a^T[D_1/n]a=\lambda$	Véc tơ riêng của $MX^TDX=D_2^{-1}N^T D_1^{-1}N$ Chuẩn hoá bởi b sao cho: $b^T[D_2/n]b=\lambda$

- Các tỷ lệ đóng góp trong tổng quán tính

Tỷ lệ đóng góp của dòng i:

$$CTR(i)=\frac{\frac{n_i}{n}a_i^2}{\lambda}$$

Tỷ lệ đóng góp của cột j:

$$CTR(j)=\frac{\frac{m_j}{n}b_j^2}{\lambda}$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Biến chính tắc	Canonical variable
Biến định tính	Qualitative variable
Độ đo khi bình phương	Chi-Square metrice
Độ đo khoảng cách	Distance Measure
Điểm cột	Point of column
Phân tích chính tắc (tương quan chính tắc)	Canonical correlation analysis
Điểm dòng	Point of row
Mô tả đồng thời	Multi-presentation
Nhân tố chính tắc	Canonical factor
Phân tích phân biệt	Discriminant analysis
Phân tích tương ứng	Correspondence analysis
Phương pháp chuẩn hóa	Normalization Method

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

CÂU HỎI ÔN TẬP

14.1. Nội dung chính của phân tích tương ứng là gì?

14.2. Làm thế nào để hình thành các ma trận tần số biên từ một bảng tiếp liên?

14.3. Mô tả ma trận $\{\frac{n_{ij}}{n_{i}}\}$ và chọn một bộ số liệu tính toán ma trận này theo công thức ma trận, nhờ bảng tính Excel.

14.4. Trung tâm của đám mây điểm cột được xác định như thế nào, minh họa bằng một thí dụ cụ thể.

14.5. Độ đo khoảng cách khi bình phương được dùng trong phân tích tương ứng vì lý do nào?

14.6. Nêu các tính chất của khoảng cách khi bình phương đối với hệ véc tơ, ý nghĩa của các tính chất đó là gì?

14.7. Bảng số liệu phân tích thành phần chính khi áp dụng cho bài toán phân tích tương ứng được hình thành như thế nào? Nêu một thí dụ bằng số.

14.8. Nêu cách thức mô tả định lượng các biến định tính trong bài toán phân tích tương ứng bằng phân tích tương quan chính tắc.

14.9. Nêu các công thức xác định tọa độ của các dấu hiệu dòng và dấu hiệu cột đối với hai biến định tính nhờ kết quả phân tích tương ứng bằng phương pháp thành phần chính khi lựa chọn hai thành phần chính.

14.10. Nêu cách tính các tọa độ trên biểu đồ dạng MAMC.

14.11. Nêu qui trình sử dụng SPSS để phân tích tương ứng, thực hiện qua 1 bộ số liệu.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bài tập tổng hợp chương 14

Với tệp GSS93 Subset.sav

14.12- Có hay không một quan hệ tương ứng giữa tình trạng hôn nhân và mức ưa thích nhạc đồng quê Châu Âu?. Nếu có hãy nêu các kết luận có thể với hai thành phần chính trong phân tích tương ứng.

14.13- Sử dụng chương trình sau và cho biết sự tương ứng nếu có của tình trạng việc làm và bằng cấp.

```
CORRESPONDENCE
 TABLE = wrkstat(1 7)  BY degree(0 4)
 /DIMENSIONS = 2
 /MEASURE = CHISQ
 /STANDARDIZE = RCMEAN
 /NORMALIZATION = SYMMETRICAL
 /PRINT = TABLE RPOINTS CPOINTS
 /PLOT = NDIM(1,2) BIPLOT(20) RPOINTS(20) CPOINTS(20) .
```

Nếu chỉ dùng một thành phần chính thì có thể nói gì về sự tương ứng này. Mô tả trên biểu đồ.

Với tệp Hhexpn98.sav (dta)

14.14- Dựa vào kết quả thực hiện chương trình sau

```
CORRESPONDENCE
 TABLE = reg7(1 7)  BY quint98b(1 5)
 /DIMENSIONS = 2
 /MEASURE = EUCLID
 /STANDARDIZE = RCMEAN
 /NORMALIZATION = SYMMETRICAL
 /PRINT = TABLE RPOINTS CPOINTS
 /PLOT = NDIM(1,2) BIPLOT(20) .
```

a- Hãy phân nhóm vùng theo tiêu thức giàu-nghèo. Việc phân nhóm như vậy có thể chính xác đến bao nhiêu %.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

b- Sử dụng độ đo khi bình phương, tiến hành lại phân tích trên. Tính tỷ lệ đóng góp của các dấu hiệu trong hai chỉ tiêu trên trong tổng quán tính. Chỉ ra các dấu hiệu của các chỉ tiêu có tỷ lệ đóng góp lớn hơn tần suất, nêu ý nghĩa của tình trạng này.

14.15- Phân tích tương ứng nhóm tuổi chủ hộ và giàu-nghèo với 2 thành phần chính. Nêu các nhận xét chính và giải thích.

14.16- Tiến hành các bài tập 2 và 3 trên SPSS và nêu các kết quả chính.

TÀI LIỆU THAM KHẢO

1. Ngô Văn Thứ. Thống kê thực hành. NXB Khoa học và kỹ thuật. Hà Nội. 2008.
2. Ngô Văn Thứ. Thống kê toán. NXB Khoa học và Kỹ thuật. Hà Nội. 2011
3. Tô Cẩm Tú: Phân tích số liệu nhiều chiều. NXB Nông nghiệp. Hà Nội 1996.
4. Bencrézi J.P. Analyse des données. Dunod. Paris. 1973
5. Benzecri. L'analyse des donnes en sociologie: Imprimerie des Presses Universitaires de France. 1990.
6. Gilbert Saporta: Analyse des donnees. Ecole national de la statistique et de l'administration economique. 1992-1993
7. Helene Erkel-Rousse: Analyse des donnees Applications et Programmation. L'ensae.1992.
8. Michel Volle: Analyse des donnees. Economica. 1994.
9. Pascal Ardilly: Theorie des sondages. L'ensae. 1994.
10. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992
11. Anderson T.W. An introduction to multivarate statistical analysis. New york. 1958
12. Dillon W and Goldstein M. Multivariate analysis Methods and Applications. New york . 1984

THƯ VIỆN HUBT

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

13. Kendall M and Stuart A, The advanced theory of Statistics. California. 1993

14. Gujarati D. Basic econometrics, McGraw Hill. New york. 1995

15. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons. New york. 1990

16. Base Syntax guid SPSS Ver. 7.5 for wwindows. SPSS Inc. 1997.

17. Dale J. Piorier: Intermadiate statistics and econometrics. 1995

18. David W. Hosmer and Stanley Lemeshow: Applied Logistic Regression. Wiley.1989.

19.L.Simar: An introduction to multivariate data analysis. Université catholique de Louvain

20. SPSS Base 8.0 Application guide. 1998.

21. T.W. Anderson: An introduction to multivariate statistical analysis. Wiley. 1984.

THƯ VIỆN HUBT

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Chương 15

CÁC PHƯƠNG PHÁP PHÂN LỚP

1. MỞ ĐẦU

Phân lớp là một trong những bài toán được quan tâm rất sớm trong lịch sử. Từ xa xưa người ta đã tiến hành phân lớp trong khắp mọi lĩnh vực. Trong trường hợp tổng thể chỉ có một đặc trưng việc phân lớp hoàn toàn hình thành tự động hay theo một quan điểm chủ quan nào đó. Ngay trong trường hợp chỉ dùng một đặc trưng, ý tưởng phân lớp đã rõ ràng và nó cũng mặc nhiên vượt khỏi giới hạn tổng thể 1 đặc trưng. Chẳng hạn, khi quan sát, nghiên cứu thu nhập của cư dân. Cho dù các điều kiện xã hội, kinh tế, chính trị cũng như các điều kiện khác là thuần nhất thì người nghiên cứu thu nhập với mục đích tìm thị trường cho một loại hàng (kể cả hàng thiết yếu như lương thực, thực phẩm, may mặc,...) cũng thường trực một ý niệm là mức hay tỷ lệ chi cho tiêu dùng mặt hàng mà họ quan tâm có thể khác nhau theo giới. Tổng thể mặc nhiên được phân lớp: nam và nữ. Có rất nhiều trường hợp bài toán phân lớp tự bộc lộ lời giải ngay trong quá trình vận động của tổng thể. Sẽ không ai thắc mắc tại sao giới tính của cư dân lại chia 2 lớp, hàng hoá lại chia thành hàng hoá thiết yếu, thông thường và xa xỉ cũng như cách phân chia chúng trong mỗi thời đại. Tuy nhiên, khi mỗi cá thể của tổng thể có quá nhiều đặc trưng nhất là có nhiều đặc trưng mới mà chúng ta không thể hiểu cặn kẽ; khi chính các đặc trưng này lại vận động trong mối quan hệ tác động qua lại đồng thời thì việc phân lớp trở thành vấn đề phức tạp. Có thể thấy vấn đề không chỉ giới hạn ở việc có quá nhiều đặc trưng cho mỗi cá thể mà chính trong điều kiện này người ta không thể áp đặt một quan niệm hay mục đích chủ quan cho một bài toán phân lớp. Trong chương này, chúng ta sẽ nghiên cứu những cơ sở của các phương pháp khác nhau giải quyết bài toán phân lớp khách quan, không phụ thuộc vào các quan điểm chủ quan mà chỉ phụ thuộc vào chính sự biểu hiện của các cá thể trong quá trình vận động của chúng.

Tiêu chuẩn cao nhất trong phân lớp một tổng thể là tạo ra các lớp (tập con) với sự thuần nhất tối đa có thể trong từng lớp cũng như sự khác biệt tối đa của các cá thể khác lớp.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với nội dung chương này, bạn đọc sẽ nắm được cách tiếp cận chung của nhóm mô hình phân tích nhân tố cho việc phân lớp các cá thể và các biến. Đối với phần phân lớp tự động, bạn đọc sẽ được tìm hiểu:

- Quan điểm tiếp cận phân lớp theo phương pháp phân tích nhân tố;
- Sự khác biệt của phân lớp cá thể và phân lớp biến.
- Các chỉ số phân lớp
- Nội dung phân lớp theo thứ bậc.
- Các kỹ thuật phân lớp và phương pháp Wald.
- Sử dụng thành thạo các thủ tục có liên quan trên SPSS

2. MỘT SỐ KHÁI NIỆM CƠ SỞ

2.1. Khoảng cách và độ khác biệt

Dữ liệu cho phân lớp một tập hợp n cá thể có thể được cho dưới dạng một bảng số (số liệu thô) hay một bảng quan hệ n x n. Các quan hệ thông thường được thể hiện dưới hai dạng: khoảng cách hoặc độ khác biệt (độ phát tán) của các cặp cá thể.

1. Khoảng cách

Xét tập n cá thể, mỗi cá thể có thể đặc trưng bởi p đặc trưng (biến). Gọi E là tập các cá thể cần phân lớp.

Khoảng cách giữa cá thể i và cá thể j là một số thực d(i,j) thoả mãn các điều kiện:

+ d(i,j) = d(j,i)

+ d(i,j) >= 0

+ d(i,j) = 0 nếu và chỉ nếu i =j

+ d(i,j) <= d(i,k) + d(k,j) cho mọi k thuộc E.

Có thể mở rộng định nghĩa trên với việc bỏ đi điều kiện thứ tư, ta được một khoảng cách phi Ơcơlit.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

2. Độ khác biệt (hay độ phát tán)

Độ khác biệt của i và j là số thực s(i,j) xác định thoả mãn các điều kiện sau:

+ s(i,j) =s(j,i)

+ s(i,j) >=0

+ s(i,i) >= s(i,j)

Độ đo này tương tự độ đo nhận được từ ma trận hiệp phương sai.

3. Các khoảng cách khác cho tập các biến nhị phân

Nếu ta có n cá thể được thể hiện bởi p (đặc trưng) biến định tính, ta có thể xem xét sự khác biệt của hai cá thể i, j nhờ thông tin các đặc trưng có xuất hiện ở các cá thể này hay không.

Gọi: a là số đặc trưng sơ cấp của toàn bộ tập hợp cá thể

b là số đặc trưng có ở i mà không có ở j

c là số đặc trưng có ở j mà không có ở i

d là số đặc trưng có ở i và cả ở j

Một số chỉ số độ đo phát tán lấy phần bù đơn vị được sử dụng như sau:

Jaccard	$\frac{a}{a+b+c}$	(J)	(15.1)
Czekanowski (Dice)	$\frac{2a}{2a+b+c}$	(C-D)	(15.2)
Ochiai	$\frac{a}{\sqrt{(a+b)(a+c)}}$	(O)	(15.3)
Russel - Rao	$\frac{a}{a+b+c+d}$	(R-R)	(15.4)
Rogers - Tanismoto	$\frac{a+d}{a+d+2(b+c)}$	(R-T)	(15.5)

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ: Thống kê biến động giá cổ phiếu 10 phiên liên tiếp của 6 công ty cho ở bảng dưới đây (1 = tăng hơn 6,2% , 0 = tăng không quá 6,2%)

Phien	1	2	3	4	5	6	7	8	9	10
PREE	0	1	0	0	0	0	0	0	0	1
PSAM	1	0	0	0	0	0	0	0	0	1
PHAP	0	1	0	1	0	0	0	0	0	1
PTMS	1	1	0	0	0	0	0	0	0	1
PLAF	1	1	0	0	0	0	0	0	0	0
PSGH	1	1	1	1	1	1	1	1	1	1

Có thể dùng các độ đo trên xem xét quan hệ về hiện tượng "tăng giá hơn 6,2%/ phiên" của các cổ phiếu. Kết quả tính toán cho ở bảng dưới đây:

		PREE	PSAM	PHAP	PTMS	PLAF
PSAM	J	0.833				
	C-D	0.909				
	O	0.909				
	R-R	0.769				
	R-T	0.733				
PHAP	J	0.909	0.769			
	C-D	0.952	0.870			
	O	0.953	0.870			
	R-R	0.769	0.714			
	R-T	0.857	0.647			
PTMS	J	0.909	0.909	0.833		
	C-D	0.952	0.952	0.909		
	O	0.953	0.953	0.909		
	R-R	0.769	0.769	0.714		
	R-T	0.857	0.857	0.750		
PLAF	J	0.833	0.833	0.833	0.909	
	C-D	0.909	0.909	0.909	0.952	
	O	0.909	0.909	0.909	0.953	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

		PREE	PSAM	PHAP	PTMS	PLAF
	R-R	0.769	0.769	0.769	0.769	
	R-T	0.733	0.733	0.733	0.857	
PSGH	J	0.556	0.556	0.833	0.588	0.556
	C-D	0.714	0.714	0.909	0.741	0.714
	O	0.745	0.745	0.913	0.767	0.745
	R-R	0.500	0.500	0.667	0.500	0.500
	R-T	0.429	0.429	0.765	0.481	0.429

Các độ đo trên đều biến đổi ngược chiều với sự khác biệt của i và j. Như vậy một giá trị lớn của mỗi độ đo cho thấy hai cá thể i, j có khả năng xếp vào một lớp.

Nếu lấy độ đo J làm căn cứ, người ta có thể xếp lớp tạm thời như sau:

Lớp 1: REE - HAP - TMS

Lớp 2: LAF - SAM

Lớp 3: SGH.

Với số liệu trên có thể sử dụng ma trận Covariance để đo độ phát tán. Kết quả tính toán là:

	PREE	*PSAM*	*PHAP*	*PTMS*	*PLAF*	*PSGH*
PREE	0.400					
PSAM	0.245	0.400				
PHAP	0.374	0.200	0.458			
PTMS	0.374	0.374	0.332	0.458		
PLAF	0.245	0.245	0.200	0.374	0.400	
PSGH	0.000	0.000	0.000	0.000	0.000	0.000

2.2. Vấn đề tổ hợp trong phân nhóm

Về mặt lý thuyết khi tập E là hữu hạn, chúng ta có thể cung cấp mọi cách phân nhóm, sau đó dùng một tiêu chuẩn nào đó để chọn cách phân nhóm tối ưu. Tuy nhiên ngay điều này cũng không dễ dàng gì, hơn thế nếu có sẵn một tiêu thức có vai trò hàm mục tiêu thì tiêu thức này cũng chỉ cung cấp một cách phân nhóm chủ quan. Số cách tổ hợp khác nhau từ n phần tử

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

quá lớn, chưa nói đến mỗi cách tổ hợp sinh ra một số lớn các tổ hợp. Hãy xét tình trạng về số tổ hợp và mối quan hệ của chúng.

1. Số cách phân chia tập n cá thể thành các lớp k cá thể

Ký hiêu $P_{n,k}$ là số cách chia n cá thể thành các lớp có k cá thể ta thấy:

Có thể chứng minh rằng: $P_{n,k} = P_{n-1,k-1} + kP_{n-1,k}$

và:
$$P_{n,k} = \frac{1}{n!}\sum_{i=1}^{k} C_k^i (-1)^{k-i} i^n \qquad (15.6)$$

2. Tổng số cách chia n cá thể

Gọi P_n là tổng số cách chia n cá thể ta có thể xác định P_n theo công thức sau:

$$P_n = \sum_{i=1}^{n} C_{n-1}^{i-1} P_{n-i} \qquad \text{với } P_0 = 0. \qquad (15.7)$$

3. CÁC PHƯƠNG PHÁP CHIA LỚP

3.1. Các phương pháp kiểu đám mây động

Các phương pháp loại này cho phép giải quyết nhanh bài toán phân lớp, đối với các tập hợp, theo một tiêu chuẩn tối ưu địa phương được sử dụng như một độ đo quán tính. Chúng ta giả sử rằng các cá thể của tập hợp được mô tả bởi các điểm trong R^p xác định một khoảng cách Ơcơlít.

1- Quán tính giữa các lớp và trong từng lớp

Cho đám mây n điểm với tâm g, giả sử rằng đám mây này được chia thành k lớp. Gọi $g_1, g_2, \ldots, g_k$ là các tâm của các lớp và $I_1, I_2, \ldots, I_k$ là quán tính của các lớp, chúng ta sẽ gọi là quán tính trong của các lớp để phân biệt với quán tính chung của cả đám mây điểm. Các quán tính này được tính theo các tâm lớp.

Tổng các quán tính của các lớp gọi là quán tính trong của các lớp, ký hiệu I_W.

$$I_W = \sum_{j=1}^{k} P_j I_j = \sum_{j=1}^{k} P_j \frac{1}{n_j}\sum_{i=1}^{n_j} d^2(x_{ji}, g_j) \qquad (15.8)$$

Trong đó P_j là trọng số của lớp j; x_{ji} điểm mô tả cá thể i của lớp j.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Tổng quán tính giữa các lớp, ký hiệu I_B, xác định như sau:

$$I_B = \sum_{i=1}^{n_j} P_j d^2(g_j, g) \qquad (15.9)$$

Nhờ công thức Huygens ta có thể nhận được công thức tính tổng quán tính chung như sau: $I = I_B + I_W$. (15.10)

Tiêu chuẩn sử dụng phân lớp là cực tiểu hoá I_W hoặc tương đương là cực đại hoá I_B (vì I hoàn toàn xác định).

Chú ý rằng tiêu chuẩn này dựa trên giả thiết k xác định, nếu k không xác định có thể dẫn đến việc chia tập n điểm thành n lớp (I_W =0).

2. Phương pháp tâm di động

Bước 1: Chọn ngẫu nhiên k điểm $c_1, c_2, \ldots, c_k$ của E, k điểm này xác định một cách chia E thành k nhóm $E_{C1}, E_{C2}, \ldots, E_{Ck}$. Một điểm i nào đó sẽ có k khoảng cách tương ứng với k điểm $c_1, c_2, \ldots, c_k$, điểm i thuộc nhóm nào có khoảng cách này nhỏ hơn.

Bước 2: Với mỗi nhóm ở bước 1 ta có một tâm nhóm $g_j(1)$, (j =1..k), cách chia nhóm ở bước 2 lặp lại như bước 1 với việc thay c_j bằng $g_j(1)$. Kết quả ta có k nhóm $\{E_{gj(1)}\}$.

Tiếp tục tìm các tâm nhóm mới của các nhóm nhận được từ bước 2 và phân nhóm lại như ở bước 2. Thuật toán sẽ hội tụ và dấu hiệu hội tụ là các tâm nhóm ở bước sau chính là các tâm nhóm ở bước trước. Dễ dàng chứng minh rằng thuật toán này hội tụ vì mỗi lần chọn tâm nhóm mới, chúng ta làm giảm tổng quán tính của các nhóm vì nếu một tâm nhóm được xác định thì quán tính trong nhóm không tăng. Thật vậy, với c_i ta lập nhóm E_{Ci}, gọi tâm nhóm này là g_i thì theo công thức Huyghens ta có:

$$I_i(g_i) = I_i(c_i) + d^2(c_i, g_i) \qquad (15.11)$$

$d^2(c_i, g_i) >= 0$ nên $I_i(g_i) <= I_i(c_i)$ dấu bằng chỉ xảy ra khi $c_i = g_i$.

Thí dụ: Với số liệu giá XMHT và FE6 theo tháng, ta muốn phân theo tháng thành 4 nhóm, với việc chọn ngẫu nhiên ban đầu các tâm là các tháng 1, 4, 7, 10. Kết quả phân lớp bước 1 cho ở cột nhóm (1) bảng sau:

Tháng	XMHT	FE6	nhóm (1)	nhóm (2)	nhóm (3)	nhóm
1	42000	3800	1	1	1	1
2	41955.08	3745.989	4	1	1	1

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3	41919.21	3742.787	4	1	1	1
4	41916.23	3792.421	4	4	4	4
5	42868.43	3888.067	7	7	7	7
6	42831.85	3884.749	7	7	7	7
7	42816.63	3833.582	7	7	7	7
8	42786.22	3830.859	7	7	7	7
9	42665.01	3820.007	7	7	7	7
10	42186.97	3777.206	10	10	10	10
11	41748.44	3737.942	4	4	4	4
12	41561.23	3721.18	4	4	4	4

Tính các tâm nhóm từ các nhóm (1) ban đầu (bước 1)

g1	42000	3800
g2	41820.04	3748.063661
g3	42793.63	3851.452906
g4	42186.97	3777.205882

Với các tâm nhóm bước1 các quan sát chia nhóm theo nhóm(2)

Tính tâm nhóm bước 2

g1	41958.1	3762.925408
g2	41741.97	3750.514026
g3	42793.63	3851.452906
g4	42186.97	3777.205882

Tính tâm nhóm bước 3

g1	41958.1	3762.925408
g2	41741.97	3750.514026
g3	42793.63	3851.452906
g4	42186.97	3777.205882

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Kết quả phân nhóm lần 3 (nhom (3)) không có gì khác lần 2. Vì vậy đây là kết quả phân nhóm cuối cùng.

Tổng quán tính của đám mây E là: I =219237.7

Quán tính của đám mây các tâm nhóm là: I_B = 166243.6 (75.828%I)

Tổng quán tính của các nhóm là: I_W = 52994.1

3. Phương pháp đám mây di động

Phương pháp này do E.Diday đề nghị vào năm 1989. Đám mây di động xuất phát là một mở rộng của phương pháp tâm di động, tuy vậy người ta nhận được kết quả tổng quát hơn khi xem xét độ đo khoảng cách của các điểm.

Có thể tóm tắt phương pháp này như sau: Thay vì chọn 1 tâm nhóm, người ta có thể chọn một "hạt nhân" nhóm. Hạt nhân này có thể là một tập q điểm của E, có thể là một trục chính hay một thành phần chính,... Trên cơ sở này tiến hành chia E thành k nhóm bao gồm các điểm tập trung quanh các hạt nhân. Tính toán lại các khoảng cách theo một độ đo nào đó và thiết lập các hạt nhân mới sao cho các nhóm thuần nhất hơn. Tiếp tục như vậy cho đến khi không thể cải thiện hơn nữa cách chia nhóm.

Để làm được như vậy, mỗi bước lặp cần tiến hành các thủ tục sau:

- Tính toán khoảng cách từ các cá thể của E đến các hạt nhân.
- Xác định k nhóm tương ứng với k hạt nhân.
- Đánh giá (đo) chất lượng của cách phân nhóm hiện có.

Cách phân nhóm như vậy cho phép sử dụng một "khoảng cách" mở rộng, không nhất thiết là khoảng cách Ơcơlit như đối với phương pháp tâm di động.

Phương pháp này thường sử dụng khi E quá lớn hoặc các biến đặc trưng cho các cá thể không thật phù hợp với độ đo khoảng cách Ơcơlit. Với việc chọn các hạt nhân, người ta làm giảm đáng kể khối lượng tính toán, tuy nhiên khó khăn là ở việc chọn độ đo khoảng cách phù hợp để có thể giải thích được kết quả phân nhóm. Ngày nay với các phần mềm chuyên dụng, người ta ít sử dụng phương pháp này với các bộ số liệu có thể sử dụng được khoảng cách Ơcơlit.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

3.2. Phân lớp với các biến nhị phân

Các phương pháp phân lớp một tập hợp n cá thể được mô tả qua các chỉ tiêu định tính đã được P. Michaud và F. Marcotorchino xây dựng. Người ta gọi các phương pháp này là các phương pháp mô tả số liệu dưới dạng quan hệ nhị phân.

1. Số liệu cặp thiết lập nhờ một quan hệ nhị phân

Cho tập E có n cá thể, quan hệ nhị phân B xác định một bảng C, n dòng, n cột với c_{ij} =1 nếu i và j có quan hệ B (i B j) và c_{ij} = 0 nếu i và j không có quan hệ B (i không B j).

Một cách chia lớp là một quan hệ nhị phân tương đương kiểu B. Như vậy nếu chỉ có 1 biến định tính có m trạng thái mô tả các cá thể, thì các cá thể của tập n cá thể nói trên được chia thành m lớp.

Một quan hệ tương đương với cách phân lớp nói trên được đặc trưng bởi 3 tính chất: phản xứng, đối xứng và bắc cầu, có thể được thiết lập nhờ một bảng C thoả mãn các tính chất sau:

$c_{ii} = 1$

$c_{ij} = c_{ji}$

$c_{ij} + c_{jk} - c_{ik} <= 1.$

Như vậy, nếu X là ma trận chỉ số lớp n dòng m cột thì ma trận C chính là XX^T. Vậy thực chất của việc tìm cách phân lớp là xác định một ma trận C thoả mãn các điều kiện nói trên.

Nếu có một ma trận C như vậy, chúng ta có thể chứng tỏ rằng số lớp (số đặc trưng của X) là:

$$m = \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} c_{ij}} \qquad (15.12)$$

Thật vậy, giả sử có n cá thể và một quan hệ nhị phân mô tả bởi C thoả mãn các điều kiện trên.

Nếu c_{ij} = 0 cho mọi i khác j thì chắc chắn là tập hợp n cá thể này phải chia thành n lớp vì mỗi cá thể có một trạng thái không giống bất kể cá thể nào khác. Ta suy ra biến định tính có đúng n dấu hiệu. Công thức trên đúng.

Nếu có đúng 2 cá thể có $c_{ij} = c_{ji}$ =1, không mất tính tổng quát ta chọn

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

i=1, j=2. Lúc này cá thể 1 và cá thể 2 hoàn toàn như nhau (có thể coi chúng là một cá thể) và khác tất cả n-2 cá thể còn lại. Biến định tính có n-1 dấu hiệu, lúc này tổng mỗi cột đầu bằng 2 còn tổng các cột khác bằng 1. Tổng n giá trị nghịch đảo đúng bằng n-1. Công thức trên nghiệm đúng.

Tiếp tục như vậy (sử dụng tính chất bắc cầu khi có từ 3 cá thể trở lên như nhau) ta có kết quả trên.

2. Phân lớp theo trung tâm theo qui tắc số đông với ràng buộc kiểu Condorcet

Cho n cá thể được mô tả bởi p biến định tính có m_1, m_2, ..., m_p dấu hiệu sơ cấp. Nếu theo cách phân lớp ở phần a ta có p cách khác nhau cùng muốn đặt lên một tập hợp n cá thể. Cần tìm một cách phân lớp có tính thoả hiệp xuất phát từ p cách chia ban đầu.

Gọi C^1, C^2, ... , C^p là các bảng quan hệ nhị phân tương ứng với các biến X_1, X_2,..., X_p.

Đặt $C = C^1 + C^2 + ... + C^p$. phần tử c_{ij} là số lần i và j thuộc cùng một nhóm nếu tiến hành p cách phân nhóm độc lập.

Gọi C' =(2C-p), đối với một biến X_k nào đó ta có:

Ta có: $c'_{ij} = 1$ nếu i và j cùng nhóm trong số lớn cách phân nhóm,

$c'_{ij} = -1$ nếu i, j cùng nhóm trong một số ít cách phân nhóm;

và $c'_{ij} = 0$ nếu i, j cùng nhóm theo một số biến và khác nhóm theo một số biến khác.

Một phép chia xác định theo C' nói chung không thoả mãn tính chất bắc cầu trong khi các phép chia ban đầu thoả mãn tính chất này.

Người ta tìm cách áp đặt các ràng buộc sao cho phép chia nhóm cuối cùng thoả mãn tính chất bắc cầu với qui tắc số đông được tôn trọng.

Bài toán được thiết lập như sau:

Nếu Y là ma trận chỉ số phân cần tìm lớp thoả mãn các điều kiện:

$$\begin{cases} y_{ij} = y_{ji} \\ y_{ij} + y_{jk} - y_{ik} \le 1 \\ y_{ij} = 0 \text{ hoặc } 1 \end{cases} \qquad (15.13)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

thì cho mỗi biến X_k có thể tính khoảng cách của Y và C^k như sau:

$$d(Y, C^k) = \sum_i \sum_j (y_{ij} - c_{ij}^k)^2 = \sum_i \sum_j (c_{ij}^k)^2 - \sum_i \sum_j y_{ij}(2c_{ij}^k - y_{ij})$$

Vì các phần tử của C^k chỉ là 0 hoặc 1 và với một biến X_k thì

$$2c_{ij}^k - y_{ij} = c_{ij}^{'k}.$$

$$d(Y, C^k) = \sum_i \sum_j c_{ij}^k - \sum_i \sum_j c_{ij}^{'k} y_{ij}$$

Tiêu chuẩn cho Y là cực tiểu tổng sai khác của Y với các C^k.

Tức là tìm Y: $\min_Y \sum_{k=1}^{p} d(Y, C^k) = \min \left(\sum_i \sum_j c_{ij}^k - \sum_i \sum_j c_{ij}^{'k} y_{ij}\right)$

Cuối cùng ta có bài toán:

Tìm Y - ma trận chỉ số phân nhóm sao cho Max $\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^{'} y_{ij}$ (15.14)

Với ràng buộc (15.13)

Đây là bài toán qui hoạch tuyến tính với các biến nhị phân. Ma trận C' xác định theo công thức C' =2C -p.

3. Các tiêu chuẩn kết hợp tối đa

Về mặt phương pháp, có thể tổng quát hoá tiêu chuẩn nêu ở mục trước như sau: Tìm véc tơ định tính Y làm cực đại hàm $A = \sum_k \Phi(Y, X^k)$, trong đó Φ là tiêu chuẩn kết hợp của các biến định tính.

Nói chung không phải lúc nào cũng có thể qui bài toán trên về dạng bài toán thông thường nếu không biết số lớp của Y. Tuy vậy nếu sử dụng Φ là độ đo kết hợp của Y và các C^k như trên với các tính chất phản xứng, đối xứng và bắc cầu thì F.Marcotorchino đã chứng minh được rằng bài toán trên có thể qui về bài toán qui hoạch tuyến tính.

Chỉ số đo độ kết hợp của Y và X^k có thể được chọn khác nhau. Hiện nay chỉ số được chấp nhận rộng rãi là chỉ số Belson có dạng sau:

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

$$\Phi(Y, X^k) = \sum_{i=1}^{n}\sum_{j=1}^{n}(n_{ij} - \frac{n_i m_j}{n})^2 \qquad (15.15)$$

trong đó: n_{ij} là tần số cá biệt của Y và X^k trên bảng tiếp liên.

Thực hiện các biến đổi tương tự như ở mục b ta có:

$$\Phi(Y, X^k) = \sum_{i=1}^{n}\sum_{j=1}^{n}(c_{ij}^k - \frac{c_i^k + c_j^k}{n} + \frac{c_{ij}^k}{n^2})y_{ij} = \sum_i \sum_j a_{ij}^k y_{ij} \qquad (15.16)$$

Bài toán qui về dạng: Tìm Y làm cực đại hàm

$$A = \sum_k \Phi(Y, X^k) = \sum_i \sum_j a_{ij}\ y_{ij} \quad \text{trong đó: } a_{ij} = \sum_k a_{ij}^k \qquad (15.17)$$

Với ràng buộc (15.13)

4. PHÂN LỚP THEO THỨ BẬC

Các phương pháp phân lớp được trình bày ở các mục trước luôn dựa trên cơ sở số lớp đã ấn định, cho dù trong II- c ta không biết trước số lớp khi có p biến định tính. Phương pháp phân lớp theo thứ bậc chỉ ra tất cả các cách phân lớp mà người sử dụng có thể chọn số lớp họ cho là phù hợp.

Với tập n cá thể, người ta có thể phân thành n, n-1, n-2, ... lớp mà lớp này có thể lồng vào lớp khác khi số lớp cần thiết theo nghĩa sau: Một phép chia k lớp là việc ghép hai lớp nào đó của phép chia k+1 lớp. Như vậy có thể có n-2 cách phân lớp có ý nghĩa (trừ hai trường hợp: chia tổng thể thành 1 lớp và n lớp). Ưu điểm quan trọng nhất của cách thức phân lớp này là chúng ta có thể lựa chọn số lớp sau khi phân lớp vì trong thực tế có nhiều bài toán phân lớp người ta không đủ thông tin để chọn số lớp ngay từ đầu.

4.1. Thứ bậc trong phép phân lớp

1. Thứ bậc của các lớp trong tập hợp cá thể cho trước

Một họ H các lớp của tập E là một họ có thứ bậc nếu:

- E và các lớp một phần tử đều thuộc H
- Cho mọi A, B thuộc H ta có $A \cap B \in \{A, B, \emptyset\}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Người ta có thể thiết lập các chỉ số (i) đối với các lớp, cho mỗi phép chia lớp có thứ bậc theo cách sau: Nếu A là một lớp con của B thì i(A) <= i(B), chỉ số i như vậy sẽ tăng theo tính không thuần nhất trong lớp.

Một phép phân lớp phù hợp với H là phép phân lớp mà các lớp là các thành phần của H. Nói cách khác phép phân lớp này có thể nhận được nhờ việc cắt cây phân lớp theo chiều ngang và tập hợp các phần của nó.

Một thứ tự H sẽ được gọi là *thứ tự phân tầng* nếu a được ghép với b trước c được ghép với d theo một thứ tự bao hàm, hay nói cách khác dù chúng ta xem xét việc ghép lớp bắt đầu từ cặp hai phần tử bất kỳ.

Một cách phân lớp có chỉ số nếu nó là một họ H có thứ bậc và tồn tại một cách lập các chỉ số (i) tăng. Có nghĩa là nếu A chứa B thì i(A) <= i(B). Như vậy có thể nói rằng một cách phân lớp có chỉ số tương ứng một phép phân lớp có thứ tự phân tầng.

Vấn đề đặt ra với mỗi thứ bậc H là nếu hai cá thể a, b được ghép vào một lớp trước c và d trong thuật toán phân lớp thì tính chất i(a,b) < i(c,d) có thoả mãn hay không. Nếu điều kiện này thoả mãn thì có thể mô tả H một cách tường minh, xác định. Ngược lại sẽ có nhiều H mô tả tương ứng với một phép chia lớp.

2. Khoảng cách siêu Metric

Mỗi thứ bậc H tương ứng một chỉ số với khoảng cách giữa các phần tử của H: d(A,B) đo mức độ kết hợp giữa A và B tức là chỉ số của bộ phận nhỏ nhất của H đồng thời chứa cả A và B.

Khoảng cách này có tính chất: d(a,b) <= Sup{d(a,c); d(b,c)} với mọi a,b,c. Tính chất này được gọi là tính siêu Metric.

Thực tế tính chất siêu Metric thể hiện như sau: trong lần đầu tiên a và b được nhóm với nhau có thể:

- Một là: Một điểm c không còn được nhóm với a và cũng như vậy nó không được nhóm với b. Nó có thể được nhóm với (a,b) về sau, theo cách tính trên d(a,c) = d(b,c) và nói chung lớn hơn d(a,b).

- Hai là: Điểm c được nhóm với a hoặc b trước (chẳng hạn nhóm với a) thì d(a,c) lớn hơn d(a,b). Nhưng d(b,c) = d(a,b) vì c và a cùng nhóm với b trong một lần.

Có thể thấy rằng khoảng cách dạng này không thông thường và có

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

những điều kiện rất lỏng. Tuy vậy, trong thực tế có nhiều trường hợp khoảng cách dạng này cũng được sử dụng trong phân lớp.

4.2. Các độ đo kết hợp dựa trên sự phát tán (khác biệt)

Cũng như các độ đo thông thường, độ đo kết hợp dựa trên độ khác biệt hay độ phát tán, vấn đề luôn đặt ra là xác định khoảng cách của một nhóm (a,b) với một phần tử thứ 3 (c).

1. Khoảng cách tối thiểu

Khoảng cách này xác định theo cách sau: d((a,b),c) = inf(d(a,c), d(b,c)). Theo cách này khoảng cách của hai lớp là khoảng cách nhỏ nhất của hai phần tử của hai lớp đó.

Một khoảng cách như vậy là một siêu Metric như đã nêu ở trên. Thực chất đây là một cách xác định độ đo theo kiểu "trội dưới".

Thật vậy d(a,b) <= d(a,c) và d(a,b) <= d(b,c) cách chọn d((a,b),c) như trên là cách chọn không thể nhỏ hơn (maximine) và vẫn thoả mãn cách định nghĩa siêu Metric.

2. Khoảng cách tối đa

Ngược lại với khoảng cách tối thiểu, người ta có thể định nghĩa khoảng cách tối đa như sau: d((a,b),c) = Sup(d(a,c), d(b,c)). Tương tự như trên cách định nghĩa này cũng cho một siêu Metric.

3. Khoảng cách hỗn hợp

Khoảng cách này được Lance và William đề nghị và sau này Jambu tổng quát hoá. Công thức hỗn hợp như sau:

$$d((a,b),c) = a_1 d(a,c) + a_2\, d(b,c) + a_3\, d(a,b) + a_4\, i(a) + a_5 i(b) + a_6 i(c) + a_7 |\, d(a,b)-d(b,c)\,| \qquad (15.18)$$

Với các hệ số thoả mãn điều kiện sau:

$$\begin{cases} a_1 + a_2 + a_3 \geq 1 \\ a_7 \geq -\min(a_1, a_2) \\ a_i \geq 0 \quad (i = 1..6) \end{cases}$$

4.3. Tiêu chuẩn quán tính và phương pháp Ward

Cho tập các cá thể mô tả bởi tập điểm E trong không gian Ơcolit với g

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

là tâm của E. Ward xem xét cách thức ghép k+1 lớp thành k lớp nhờ đánh giá độ mất mát quán tính. Đây là phương pháp dễ thực hiện và hiện được sử dụng tương đối phổ biến trong việc phân lớp theo thứ bậc.

1. Mức giảm quán tính khi ghép lớp và khoảng cách giữa hai lớp

Giả sử có hai lớp A tỷ trọng P_A và B tỷ trọng P_B. Trước khi ghép hai lớp này tổng quán tính hai lớp có thể tính bằng trung bình bình phương khoảng cách từ các tâm lớp đến tâm đám mây điểm, cụ thể là:

$P_A d^2(g_A,g) + P_B d^2(g_B, g)$. Nếu ghép A và B quán tính của lớp ghép là bình phương khoảng cách từ tâm lớp ghép đến tâm của đám mây điểm, cụ thể là: $(P_A + P_B)\, d^2(g_{AB}, g)$.

Phần quán tính mất đi sẽ là:

$$I_{AB} = P_A d^2(g_A, g) + P_B d^2(g_B, g) - (P_A + P_B)\, d^2(g_{AB}, g) \qquad (15.19)$$

trong đó: $g_{AB} = \dfrac{P_A g_A + P_B g_B}{P_A + P_B}$

Vì:

$$d^2(g_{AB}, g) = \frac{P_A}{P_A + P_B} d^2(g_A, g) + \frac{P_B}{P_A + P_B} d^2(g_B, g) - \frac{P_A P_B}{(P_A + P_B)^2} d^2(g_A, g_B)$$

nên: $I_{AB} = \dfrac{P_A P_B}{P_A + P_B} d^2(g_A, g_B)$

Chúng ta có thể coi khoảng cách của A và B là:

$$\delta(A,B) = \frac{P_A P_B}{P_A + P_B} d^2(g_A, g_B) \qquad (15.20)$$

Nếu C là một lớp thứ 3 có thể xác định khoảng cách của C và AUB theo công thức sau:

$$\delta(C,(A \cup B)) = \frac{(P_A + P_C)\delta(A,C) + (P_B + P_C)\delta(B,C) - P_C \delta(A,B)}{P_A + P_B + P_C} \qquad (15.21)$$

2. Thuật toán Ward

Thuật toán Ward thực hiện việc phân lớp đám mây (cá thể) điểm như sau:

Với đám mây n điểm trong R^p ở các chương trước ta có tương ứng ma trận khoảng cách D. Thay các khoảng cách này bởi khoảng cách giữa hai

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

điểm i, j là:

$$\delta(i,j) = \frac{p_i p_j}{p_i + p_j} d^2(i,j) \qquad (15.22)$$

Tìm hai cá thể có $\delta(i,j)$ nhỏ nhất ghép vào một lớp có tỷ trọng (p_i+p_j) với thứ bậc tương ứng $\delta(i,j)$.

Tính khoảng cách δ của các cá thể khác với lớp đã có theo công thức nói trên giống như đám mây n-1 điểm trong đó có một điểm là tâm của lớp đã có. Tìm cá thể k gần nhất với lớp đã có và ghép thành một lớp bao hàm với thứ bậc của lớp này tính theo công thức trên. Tiếp tục như vậy cho đến khi tất cả các cá thể đều được phân lớp.

Thí dụ: Cho số liệu 8 nước Đông Âu với 3 tiêu thức mật độ dân số (Density), tỷ lệ dân sống ở thành thị (urban), tỷ lệ dân biết đọc biết viết (Literacy)

COUNTRY	DENSITY	URBAN	LITERACY
Belarus	50.0	65	99
Bosnia	87.0	36	86
Bulgaria	79.0	68	93
Croatia	85.0	51	97
Estonia	36.0	72	99
Georgia	81.0	56	99
Hungary	111.0	64	99
Latvia	40.0	71	99

Thực hiện phân lớp theo thuật toán Ward

+ Trước hết ta tính ma trận D^2 (Bình phương khoảng cách Ơcơlit)

	Belarus	Bosnia	Bulgaria	Croatia	Estonia	Georgia	Hungary
Belarus							
Bosnia	2379						
Bulgaria	886	1137					

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Croatia	1425	350	341				
Estonia	245	4066	1901	2846			
Georgia	1042	605	184	45	2281		
Hungary	3722	1529	1076	849	5689	964	
Latvia	136	3603	1566	2429	17	1906	5090

+ Ma trận khoảng cách: $\delta(i, j)$

Chú ý rằng các tỷ trọng của các quan sát bằng nhau nên $\delta(i, j)$=d^2(i,j)/2n.

	Belarus	Bosnia	Bulgaria	Croatia	Estonia	Georgia	Hungary
Belarus							
Bosnia	148.69						
Bulgaria	55.38	71.06					
Croatia	89.06	21.88	21.31				
Estonia	15.31	254.13	118.81	177.88			
Georgia	65.13	37.81	11.50	2.81	142.56		
Hungary	232.63	95.56	67.25	53.06	355.56	60.25	
Latvia	8.50	225.19	97.88	151.81	**1.06**	119.13	318.13

+ Trước hết nhóm Latvia với Estonia, 8 nước chia thành 7 nhóm.

Để thực hiện chia 8 nước thành 6 nhóm ta tiến hành các bước sau:

+ Tính lại ma trận D^2 với Esto_latvi là tâm của hai điểm Estonia và Latvia

	Belarus	Bosnia	Bulgaria	Croatia	Georgia	Hungary
Belarus						
Bosnia	2379					
Bulgaria	886	1137				

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Croatia	1425	350	341			
Georgia	1042	605	184	45		
Hungary	3722	1529	1076	849	964	
Esto-Latvi	186.25	3830.25	1729.25	2633.25	2089.25	5385.25

Từ đó tính ma trận $\delta(i, j)$ với tỷ trọng của Esto_Latvi là 2/n còn các điểm khác là 1/n vì nhóm Esto_latvi có hai điểm.

.	Belarus	Bosnia	Bulgaria	Croatia	Georgia	Hungary
Belarus						
Bosnia	148.69					
Bulgaria	55.38	71.06				
Croatia	89.06	21.88	21.31			
Georgia	65.13	37.81	11.50	**2.81**		
Hungary	232.63	95.56	67.25	53.06	60.25	
Esto-Latvi	15.52	319.19	144.10	219.44	174.10	448.77

+ Georgia và Croatia được nhóm lại với nhau, 8 nước đã nhóm thành 6 nhóm.

Tiếp tục như trên với tỷ trọng lớp Cro-Geor và Esto-Latvi là 2/n .

+ Ma trận D^2

	Belarus	Bosnia	Bulgaria	Hungary	Esto-Latvi
Belarus					
Bosnia	2379				
Bulgaria	886	1137			
Hungary	3722	1529	1076		
Esto-Latvi	186.25	3830.25	1729.25	5385.25	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Cro-Geor	1222.25	466.25	251.25	895.25	2350

Từ đó tính ma trận $\delta(i, j)$

	Belarus	Bosnia	Bulgaria	Hungary	Esto-Latvi
Belarus					
Bosnia	148.69				
Bulgaria	55.38	71.06			
Hungary	232.63	95.56	67.25		
Esto-Latvi	**15.52**	319.19	144.10	448.77	
Cro-Geor	101.85	38.85	20.94	74.60	293.75

Belarus được nhóm vào lớp Esto-Latvi gọi là nhóm Esto-Lat-Bel.

Tiếp tục như trên với tỷ trọng lớp Cro-Geor là 2/n và Esto-Lat -Bel là 3/n .

+ Ma trận D^2

	Bosnia	Bulgaria	Hungary	Cro-Geor
Bosnia				
Bulgaria	1137.00			
Hungary	1529.00	1076.00		
Cro-Geor	466.25	251.25	895.25	
Esto-lat-bel	3305.11	1406.78	4789.44	1932.69

Từ đó tính ma trận $\delta(i, j)$

	Bosnia	Bulgaria	Hungary	Cro-Geor
Bosnia				
Bulgaria	71.06			
Hungary	95.56	67.25		
Cro-Geor	38.85	**20.94**	74.60	
Esto-lat-bel	309.85	131.89	449.01	134.29

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Bulgaria được nhóm vào lớp Cro-Geor (Cro-Geor-Bul).

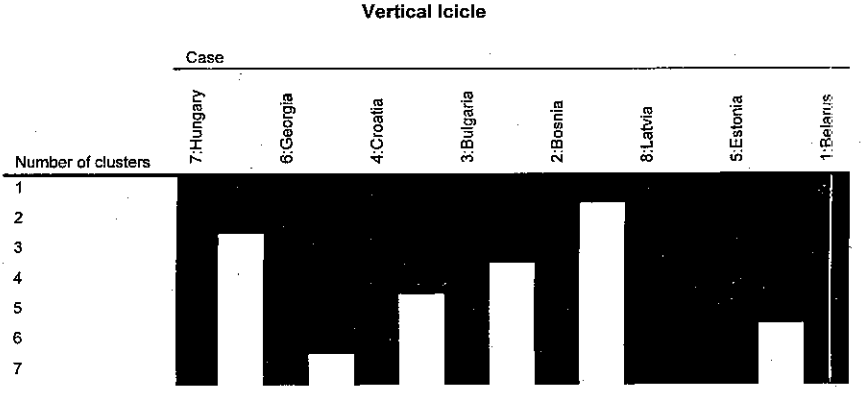
Kết quả tóm tắt trên SPSS như sau: (Các hệ số khoảng cách được tính với bội số n so với các kết quả trên)

+ Quá trình nhóm các cá thể vào các lớp:

Agglomeration Schedule

	Statistics					
	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	5	8	8.500	0	0	3
2	4	6	31.000	0	0	4
3	1	5	155.167	0	1	7
4	3	4	322.667	0	2	5
5	2	3	798.167	0	4	6
6	2	7	1548.667	5	0	7
7	1	2	6038.625	3	6	0

Biểu đồ phân lớp:



4.4. Trường hợp biến định tính

Với các biến định tính chúng ta có bảng số liệu như sau:

$$X=\begin{pmatrix} 0100 & 010 & 1000 \\ 0010 & 001 & 0100 \\ 0001 & 100 & 0001 \end{pmatrix}$$

Khoảng cách được đề nghị là khi bình phương, cụ thể là hai cá thể i, j có

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

khoảng cách tính theo công thức sau:

$$d^2(i,j) = \chi^2 = \sum_{k=1}^{n} \frac{n}{n_k} \left(\frac{n_{ik} - n_{jk}}{p}\right)^2 \qquad (15.23)$$

trong đó: n_{ik}, n_{jk} là các giá trị trên các dòng i và j; n_k là tần số cột k; p là số biến cần phân lớp.

5. PHÂN LỚP ĐỐI VỚI CÁC BIẾN

Trong các mục trước, chúng ta đã xem xét các kỹ thuật phân lớp với các cá thể. Về mặt hình thức, nếu quan niệm mỗi biến là một "cá thể" thì mỗi cá thể có thể coi là một biến. Sự khác biệt về mặt nội dung hàm chứa trong lý thuyết mẫu và suy diễn thống kê. Chúng ta không thể áp đặt một cách hình thức "một tổng thể biến", dù cho về mặt toán học điều đó là hoàn toàn có thể được, theo cách sử dụng cấu trúc hệ thống. Mặc dù vậy, bài toán phân lớp các biến vẫn có ý nghĩa nhất định và trong thực tế người ta không thể bỏ qua.

Phân lớp các biến trên một tổng thể, nhằm xếp các nhóm đặc trưng có quan hệ chặt chẽ hơn trong cùng một lớp. Các thuật toán phân lớp đặc biệt có ý nghĩa khi chúng ta bổ sung các tiêu thức (biến) cho một tổng thể. Hãy xét bài toán lý thuyết sau để thấy rõ nội dung này.

Giả sử đã có p biến đặc trưng trên một tổng thể. Các biến này đã được phân thành k lớp hoặc đã xác định cây phân lớp có thứ bậc. Một biến khác (X_{p+1} chẳng hạn) được quan tâm bổ sung, vấn đề sẽ là:

- Hai biến nào được nhóm thành một lớp?.
- Có thể ghép X_{p+1} vào lớp nào trong k lớp ?.
- Hai nhóm biến có thể ghép với nhau với độ đo khác biệt hay chỉ số phân lớp như thế nào?.

Về mặt kỹ thuật, trả lời các câu hỏi trên không quá khó khăn, nhưng thực tế áp dụng các kỹ thuật này cho các biến không thật đơn giản nhất là khả năng phân tích kết quả. Một số sai lầm đáng kể có thể xảy ra khi áp dụng những phân tích máy móc của các cá thể cho các biến. Trong phần *tiếp theo chúng ta tập trung sự quan tâm* vào những nội dung chủ yếu để tránh những sai lầm như vậy.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

5.1. Phân lớp các biến - tách tổng thể theo giác độ nghiên cứu

Ngay khi xác định tổng thể, một cách chung nhất chúng ta đã ngầm định rằng mỗi cá thể mang p đặc tính thể hiện bởi tình trạng, mức độ của p biến ngẫu nhiên X_1, X_2,..,X_p. Như vậy tổng thể đã xác định hay ít nhất là đã xác định một cách nhìn tổng thể, số cá thể dù được quan sát ít hay nhiều chỉ ảnh hưởng tới độ chính xác trong phân lớp. Tuy nhiên, khi thêm 1 biến X_{p+1} hay ngược lại là phân chia p biến thành các lớp thực chất là việc thay đổi tổng thể hay phân tách việc nghiên cứu tổng thể theo những phương diện khác nhau. Tư tưởng chính của bài toán phân lớp khi áp dụng cho các biến có ý nghĩa hoàn toàn khác với việc phân lớp đối với các cá thể.

5.2. Độ đo khoảng cách đối với các biến định lượng

Có thể nhận thấy một hình ảnh phân lớp trong phân tích thành phần chính với biểu đồ của các biến. Nói cách khác phân tích nhân tố theo các thành phần chính có thể coi là một công cụ phân lớp, mặc dù kết quả phân lớp không được trình bày rõ ràng. Trong phân tích thành phần chính, chúng ta đã sử dụng hệ số tương quan tuyến tính của các biến để thể hiện mức độ "gần gũi" của các biến với các biến mới (các trục chính) và chính bằng cách này sự "gần gũi" của các biến ban đầu cũng phần nào được thể hiện. Trong kỹ thuật phân lớp Gilbert Saporta đề nghị sử dụng hệ số tương quan tuyến tính r làm độ đo khoảng cách cho các biến và (1-r) là độ đo sự khác biệt của các biến loại này.

Có thể thấy rằng nếu chọn $d(i,j) = (1- r_{ij})$ thì các tính chất phản xứng và đối xứng được nêu ở phần I luôn thoả mãn. Tính chất bắc cầu nói chung là không được khẳng định đối với độ đo này. Như vậy, với độ đo r chúng ta chỉ có thể phân lớp các biến nhờ phương pháp phân lớp có thứ bậc.

5.3. Độ đo khoảng cách đối với các biến định tính

Với các biến định tính khái niệm "gần gũi" được thể hiện qua tính chất không độc lập. Chính đặc tính này gợi ý sử dụng độ đo Khi bình phương. Tuy nhiên, với các biến định tính có cùng số dấu hiệu thì hoàn toàn có thể sử dụng độ đo r hoặc r^2 hay các độ đo tương tự khác. Với các biến định tính không cùng số dấu hiệu việc phân lớp hầu như cho những kết quả khác nhau khi dùng các độ đo khác nhau.

5.4- Tiếp cận Lerman và thuật toán dựa trên tính đúng đắn của các liên hệ

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

I.C Lerman đã đề nghị một tiếp cận khác khi phân lớp đối với các biến. Ông thay các độ đo nói trên bằng khả năng các biến độc lập (không có mối liên hệ nào). Chẳng hạn, thay cho việc dùng hệ số tương quan tuyến tính r, ông sử dụng P(R<r).

Cách tiếp cận này khá hiệu quả cho trường hợp các biến định tính. Với các biến định tính ta có thể xác định độ đo khi bình phương, từ đó xác định bảng các hệ số đo độ khác biệt $d(i,j) = P(\chi^2_k > \chi^2(i,j))$. Hai biến i và j được ghép vào một nhóm nếu d(i,j) nhỏ nhất.

Khi đã có hai nhóm biến A và B với p và q biến thì tiêu chuẩn ghép hai lớp này thành một lớp lớn phụ thuộc cách chọn độ đo khoảng cách của hai lớp này. Thông thường người ta chọn $\inf\{P(\chi^2_k > \chi^2(i,j))\}$ với $i \in A$ và $j \in B$ làm độ đo khác biệt của A và B.

6. BÀI TOÁN PHÂN LỚP TRÊN SPSS

SPSS thực hiện phân lớp nhờ thủ tục Classify trong thực đơn Analysis (hoặc Statistic) và Cluster trên Winstata. Sau đây chúng ta tập trung mô tả cách dùng thủ tục này trên SPSS là chính.

6.1. Thủ tục K - Means Cluster...

Thủ tục này thực hiện phân lớp các cá thể với số lớp xác định trước (k). Thực đơn chính như sau:

- Cửa sổ biến cho phép khai báo các biến đặc trưng chia tổng thể trong phân lớp. Việc xác định các biến này phụ thuộc vào phạm vi quan tâm của người nghiên cứu hay vấn đề cần phân tích.

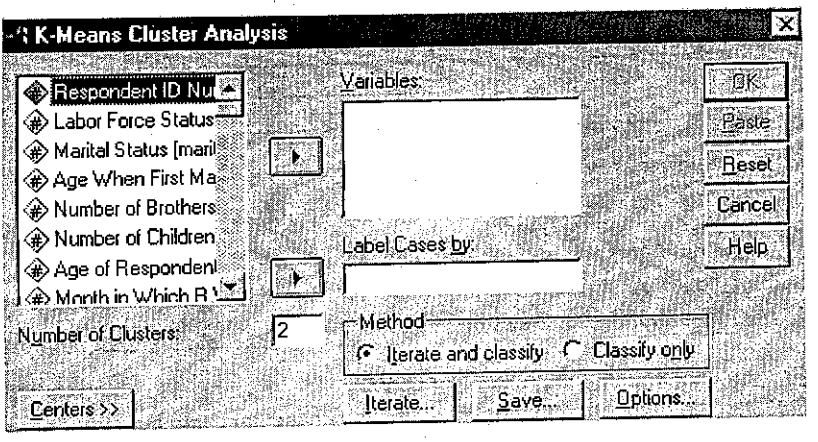
- Số lớp (k) được xác định trong cửa số Number of Clusters.

- Yêu cầu kết quả phân nhóm có thể là toàn bộ quá trình hoặc chỉ in kết quả cuối cùng được xác định ở mục: Method

- Tên cá thể kèm theo số thứ tự Cases được xác định ở ô Label Cases by.

- Nút Centers cho phép ghi lại các tâm lớp thành một tệp nếu có nhu cầu.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



- Nút Iterate cho phép xác định số bước lặp và sử dụng trung tâm nhóm nếu có thể.

- Nút Save cho phép ghi lại từng cá thể là thành phần của nhóm nào trong k nhóm.

- Nút Option... cho phép lựa chọn các mô tả thống kê trong quá trình phân tích.

Sau đây là chương trình tương ứng với bảng chọn đơn giản và kết quả phân lớp các cá thể từ tệp GSS93.save.

```
QUICK CLUSTER
  childs sibs educ income91 chldidel
  /MISSING=LISTWISE  /CRITERIA= CLUSTER(3) MXITER(10)
CONVERGE(0)
  /METHOD=KMEANS(UPDATE )  /PRINT ID(id ) ANOVA.
```

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Final Cluster Centers

	Cluster		
	1	2	3
Number of Children	3	2	2
Number of Brothers and Sisters	9	4	3
Highest Year of School Completed	11	12	14
Total Family Income	13	8	18
Ideal Number of Children	3	3	3

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Number of Children	79.359	2	2.729	921	29.078	.000
Number of Brothers and Sisters	1992.580	2	5.388	921	369.800	.000
Highest Year of School Completed	963.094	2	7.575	921	127.147	.000
Total Family Income	9637.910	2	8.689	921	1109.221	.000
Ideal Number of Children	22.453	2	2.351	921	9.550	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	119.000
	2	279.000
	3	526.000
Valid		924.000
Missing		576.000

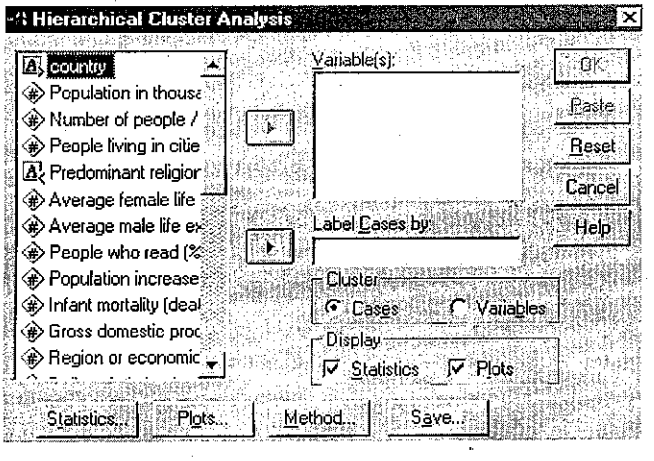
Bảng ANOVA cho thấy tổng sai số bình phương tính theo từng biến trong kết quả phân lớp. Với mỗi biến Cluster\Mean Square cho trung bình bình phương khoảng cách giữa các tâm nhóm với tâm của toàn bộ đám mây điểm; Error\ Mean Square cho trung bình khoảng cách của các cá thể đến tâm các nhóm. Đây là hai đại lượng mang thông tin của quán tính I_B và I_W đã được thiết lập trong mục I. Ngoài ra SPSS thực hiện phân tích phương sai với giả thiết H_0 là "Các tâm lớp không khác nhau" tương đương với giả thiết "Không cần chia thành 3 lớp". Chúng ta có thể nhận được nhiều thông tin hơn nếu xác nhận các lựa chọn khác trên thực đơn.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

6.2. Phân lớp có thứ bậc các cá thể

Phân lớp có thứ bậc được thực hiện với nhiều độ đo khoảng cách khác nhau, như trong các mục trên đã trình bày. Vì vậy khi sử dụng phân lớp có thứ bậc cần chú ý đến đặc điểm của các biến dùng để phân lớp. Có hai lựa chọn quan trọng hơn cả là lựa chọn phương pháp và lựa chọn độ đo mà sau đây chúng ta sẽ xem xét một số cách lựa chọn như vậy.

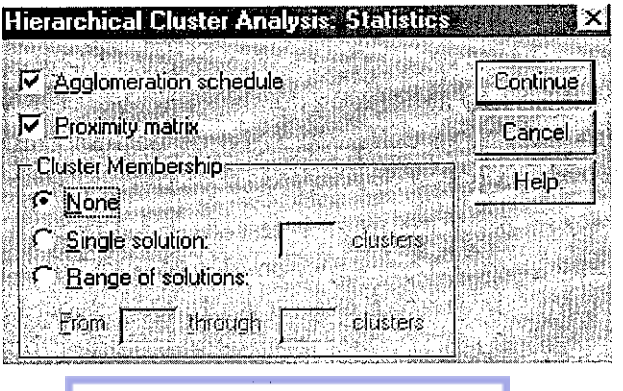
Thủ tục phân lớp có thứ bậc nhận được từ thực đơn Analysis\Classify\ Hierarchical với bảng chọn sau:



Bảng chọn này có cửa sổ biến và nhãn như bảng chọn trên, vì vậy các giải thích chi tiết chỉ thực hiện đối với các lựa chọn mới.

- Lựa chọn Cluster cho phép phân lớp các cá thể (Cases) hoặc các biến (Variables).

- Nút Statistics cho các lựa chọn mô tả thống kê



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

+ Lựa chọn Agglomeration schedule: Liệt kê toàn bộ quá trình ghép lớp

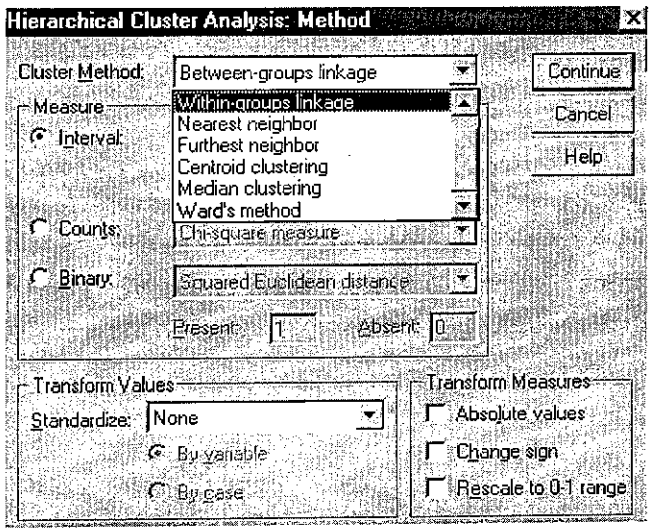
+ Lựa chọn Proximity matrix cho phép hiện bảng độ đo theo cách lựa chọn ở nút Method.

+ Bảng chọn nhỏ Cluster Membership cho phép lựa chọn các cách thức liệt kê các cá thể vào các lớp.

- Nút Plots cho phép định dạng cách hiện các lớp ghép theo thứ bậc.

- Nút Method cho bảng chọn sau: (bảng chọn Hierarchical Cluster...).

+ Bảng chọn này cho phép chọn các phương pháp khác nhau để phân lớp, phương pháp Ward là một trong các phương pháp được liệt kê trong cửa sổ Cluster Method (như bảng trên). Tuỳ thuộc người sử dụng lựa chọn phương pháp, sự khác nhau của các phương pháp chỉ là tốc độ phân lớp.

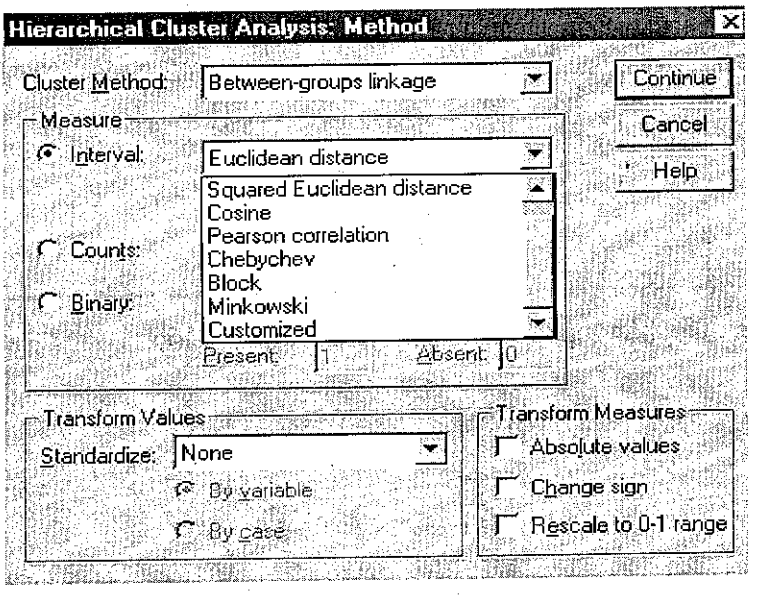


+ Việc sử dụng độ đo được liệt kê trong cửa sổ Measure. Lựa chọn độ đo trong nhiều trường hợp ảnh hưởng đến kết quả phân lớp. Vì vậy cần phải chọn độ đo thích hợp. SPSS cung cấp 3 lớp độ đo tương ứng với 3 trường hợp sử dụng biến khác nhau:

- Đối với các biến nói chung khi sử dụng khoảng cách thông thường, với lựa chọn Interval ta có bảng chọn độ đo khoảng cách. Trong đó khoảng cách Ơcơlit với các bậc khác nhau thường được sử dụng khi các biến làm cơ sở phân lớp là các biến định lượng. Trường hợp tập hợp biến

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

dùng để phân lớp các cá thể bao gồm cả biến định lượng và định tính có thể sử dụng khoảng cách khác như Chebychev, Minkowski,... như ở bảng liệt kê dưới đây.



- Đối với các biến định tính nhiều dấu hiệu chúng ta có thể chọn độ đo Khi bình phương từ Counts.

- Với các biến nhị phân độ đo được chọn từ bảng tiếp liên 2x2, có nhiều độ đo khác nhau như Jaccard, Hamann,... tuy vậy độ đo Khi bình phương được coi là thông dụng hơn cả.

Sau đây là một số kết quả cụ thể tiến hành trên tệp Wordl95.sav (các nước trung cận đông).

Thí dụ 1: Các biến chọn làm cơ sở phân lớp là: urban lifeexpf babymort gdp_cap calories lit_male

```
CLUSTER  urban lifeexpf babymort gdp_cap calories lit_male
 /METHOD BAVERAGE  /MEASURE= SEUCLID  /ID=country  /PRINT
SCHEDULE CLUSTER(3)  /PLOT DENDROGRAM VICICLE.
```

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Kết quả quá trình phân lớp

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	60	87	132931.3	0	0	4
2	52	53	293772.0	0	0	3
3	52	58	560322.0	2	0	5
4	60	64	797890.1	1	0	7
5	33	52	970878.7	0	3	6
6	33	99	5991379	5	0	7
7	33	60	2.3E+07	6	4	0

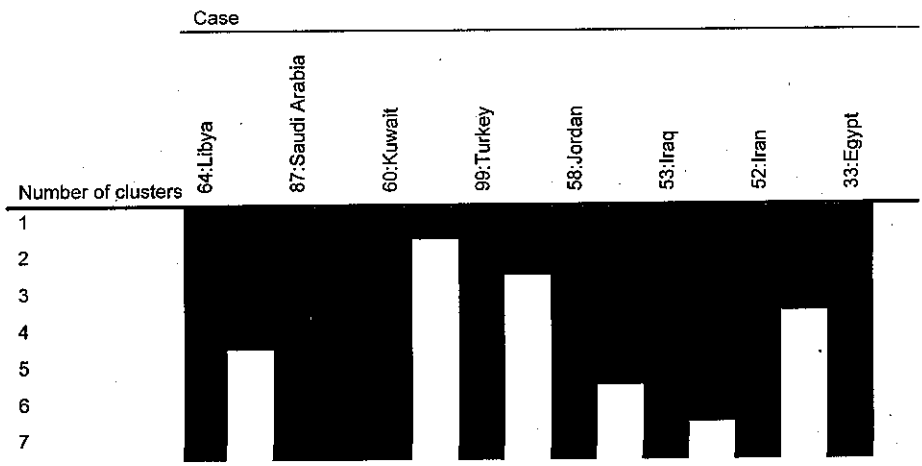
- Kết quả phân lớp khi chia 8 quốc gia thành 3 lớp

Cluster Membership

Case	3 Clusters
33:Egypt	1
52:Iran	1
53:Iraq	1
58:Jordan	1
60:Kuwait	2
64:Libya	2
87:Saudi Arabia	2
99:Turkey	3

- Biểu đồ cây phân lớp

Vertical Icicle



THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Thí dụ 2:

Các biến chọn làm cơ sở phân lớp là: urban lifeexpf babymort lit_male climate density và độ đo khoảng cách Minkowski

CLUSTER urban lifeexpf babymort lit_male climate density

/METHOD BAVERAGE

/MEASURE= **MINKOWSKI(2)**

/ID=country

/PRINT SCHEDULE CLUSTER(3)

/PRINT DISTANCE

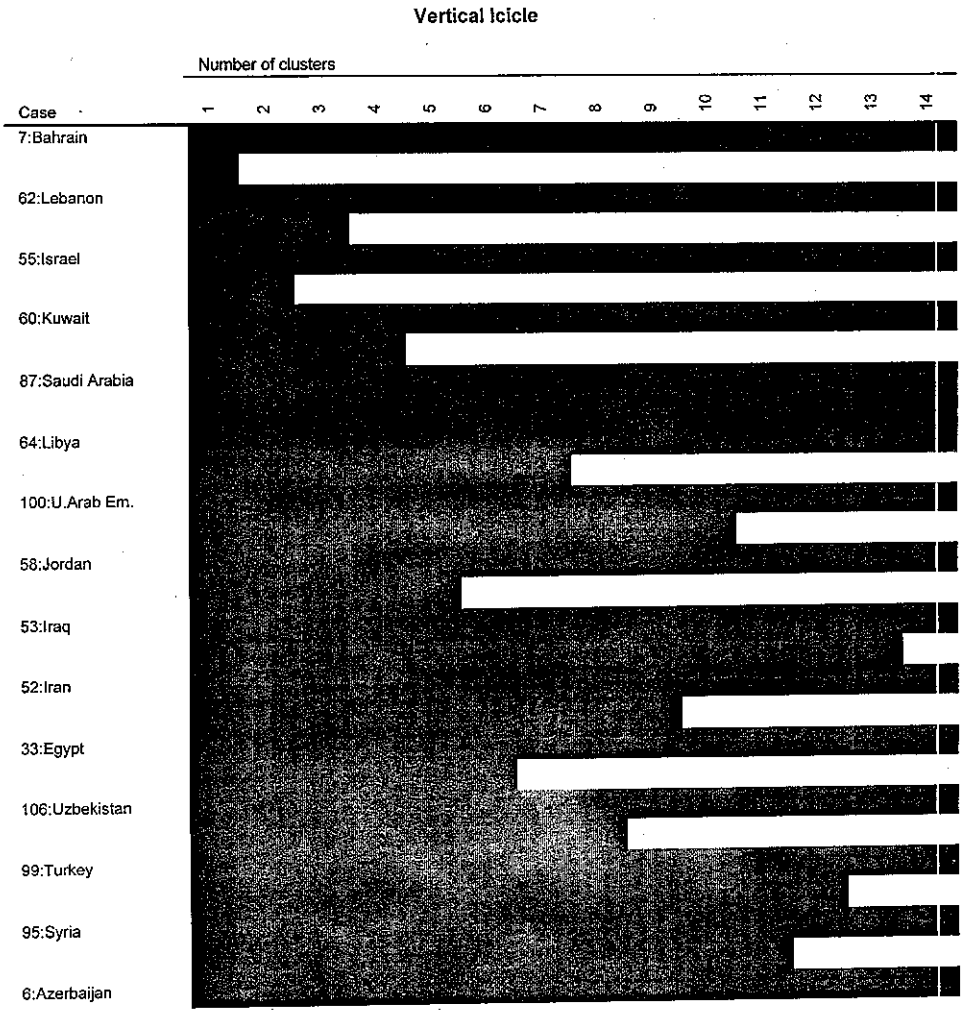
/PLOT DENDROGRAM VICICLE.

Kết quả phân lớp tóm tắt như sau:

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	64	87	14.967	0	0	8
2	52	53	18.358	0	0	6
3	95	99	20.000	0	0	4
4	6	95	24.077	0	3	7
5	58	100	27.839	0	0	8
6	33	52	30.662	0	2	9
7	6	106	38.406	4	0	9
8	58	64	46.842	5	1	10
9	6	33	49.958	7	6	10
10	6	58	58.619	9	8	11
11	6	60	78.228	10	0	13
12	55	62	110.353	0	0	13
13	6	55	245.043	11	12	14
14	6	7	744.308	13	0	0

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ



Thí dụ 3: Thí dụ này sử dụng số liệu 1% của tệp số liệu di dân đến Hà Nội (didanHN1.sav) với các thông tin từ nơi xuất cư, nhằm mô tả cách chọn độ đo cho phân lớp. Các kết quả nhận được không cho phép rút ra kết luận đúng. Lý do duy nhất là kích cỡ các bảng kết quả quá lớn, không cho phép thể hiện trên văn bản này. Tuy vậy chương trình kèm theo cho phép chạy với tệp 100% quan sát và các kết luận từ kết quả này chắc chắn cho phép rút ra những kết luận đáng tin cậy.

Các biến được lựa chọn để phân lớp các cá thể là: quan hệ với chủ hộ (qh); tình trạng hôn nhân (marrital); trình độ chuyên môn (profess); học vấn

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

(edulevel); nghề nghiệp (occu) và nơi chuyển đi (ogiplace). Các biến trên là các biến định tính có số dấu hiệu khác nhau vì vậy chúng ta chọn độ đo Khi bình phương.

CLUSTER qh marrital profess edulevel occu ogiplace

/METHOD BAVERAGE /MEASURE= **CHISQ** /ID=thutuho

/PRINT SCHEDULE /PRINT DISTANCE

/PLOT DENDROGRAM HICICLE /SAVE CLUSTER(4) .

Kết quả tóm tắt

Case Processing Summary

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
12	70.6	5	29.4	17	100.0

a. Average Linkage (Between Groups)

- Độ đo Khi bình phương (cơ sở tính d(i,j))

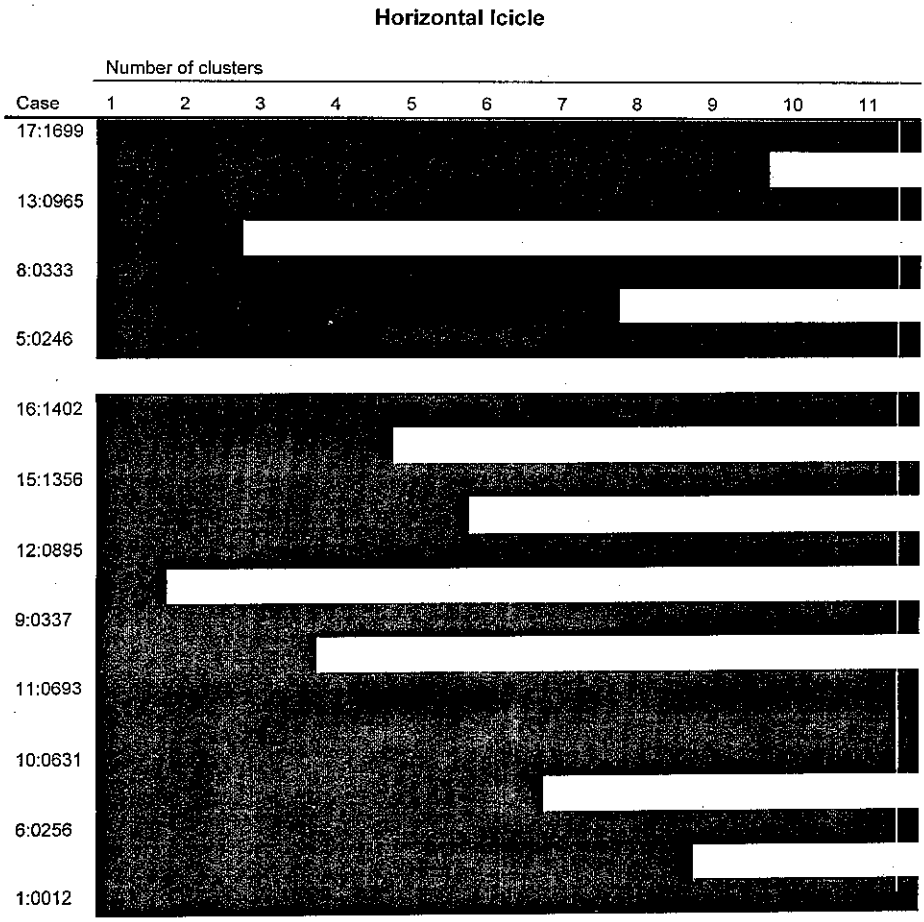
Proximity Matrix

	Chi-square between Sets of Frequencies											
Case	1	5	6	8	9	10	11	12	13	15	16	17
1:0012		1.466	.559	1.666	1.360	.984	.984	1.272	1.609	1.585	1.805	1.485
5:0246	1.466		1.748	.580	1.405	1.517	1.52	1.824	1.365	1.817	1.572	1.140
6:0256	.559	1.748		2.033	1.669	.841	.841	1.566	1.950	1.969	2.195	1.752
8:0333	1.666	.580	2.033		1.525	1.856	1.86	1.894	1.623	1.778	1.280	1.374
9:0337	1.360	1.405	1.669	1.525		1.151	1.15	1.213	2.224	1.904	1.623	2.196
10:0631	.984	1.517	.841	1.856	1.151		.000	1.183	2.119	1.810	1.935	1.957
11:0693	.984	1.517	.841	1.856	1.151	.000		1.183	2.119	1.810	1.935	1.957
12:0895	1.272	1.824	1.566	1.894	1.213	1.183	1.18		2.625	1.038	1.367	2.500
13:0965	1.609	1.365	1.950	1.623	2.224	2.119	2.12	2.625		2.658	2.748	.556
15:1356	1.585	1.817	1.969	1.778	1.904	1.810	1.81	1.038	2.658		1.170	2.490
16:1402	1.805	1.572	2.195	1.280	1.623	1.935	1.94	1.367	2.748	1.170		2.502
17:1699	1.485	1.140	1.752	1.374	2.196	1.957	1.96	2.500	.556	2.490	2.502	

This is a dissimilarity matrix

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Cây phân lớp



Chú ý: Cây phân lớp trên chia thành hai khối cho thấy rằng việc ghép toàn bộ các cá thể vào một lớp là không chấp nhận được.

- Kết quả chi tiết phân lớp các cá thể khi chỉ chọn 4 lớp được ghi lại trong tệp số liệu nhờ lựa chọn (SAVE CLUSTER(4)).

3. Phân lớp có thứ bậc các biến

Trong bảng chọn phân lớp có thứ bậc nói trên, chọn phân lớp theo biến CLUSTER

/MATRIX IN ('C:\WINDOWS\TEMP\spssclus.tmp')

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

/METHOD BAVERAGE

/PRINT SCHEDULE

/PRINT DISTANCE

/PLOT DENDROGRAM HICICLE.

- Ma trận hệ số tương quan

Proximity Matrix

Case	Matrix File Input										
	code relation	Sex	Marrital status	Professional skill	Income	Registration in Hanoi	Age	Education level	Occupation	University and higher	Administrator and relating
code relation		.508	.000	-.282	-.109	.737	.561	-.186	.592	-.049	-.290
Sex	.508		-.447	.239	.006	.363	.326	-.079	.051	.135	.200
Marrital status	.000	-.447		-.267	.378	.000	.397	.177	.684	-.302	.000
Professional skill	-.282	.239	-.267		.504	-.303	.308	.378	-.152	.806	.837
Income	-.109	.006	.378	.504		-.066	.491	.285	.431	.573	.605
Registration in Hanoi	.737	.363	.000	-.303	-.066		.449	.229	.462	-.098	-.363
Age	.561	.326	.397	.308	.491	.449		.489	.661	.313	.361
Education level	-.186	-.079	.177	.378	.285	.229	.489		-.040	.213	.316
Occupation	.592	.051	.684	-.152	.431	.462	.661	-.040		-.034	-.051
University and higher	-.049	.135	-.302	.806	.573	-.098	.313	.213	-.034		.674
Administrator and rela	-.290	.200	.000	.837	.605	-.363	.361	.316	-.051	.674	

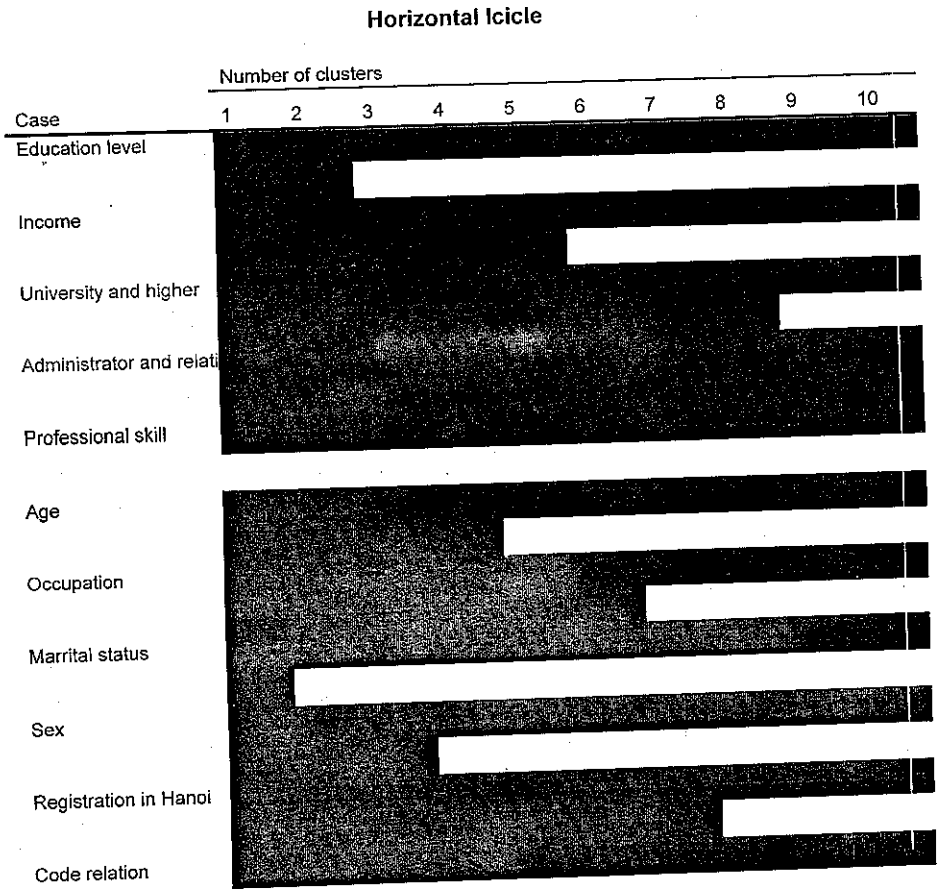
- Quá trình phân lớp

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	11	.837	0	0	2
2	4	10	.740	1	0	5
3	1	6	.737	0	0	7
4	3	9	.684	0	0	6
5	4	5	.560	2	0	8
6	3	7	.529	4	0	9
7	1	2	.435	3	0	9
8	4	8	.298	5	0	10
9	1	3	.221	7	6	10
10	1	4	.036	9	8	0

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Cây phân lớp



Điều dễ nhận ra nhất và cũng chấp nhận được là các biến về học vấn, nghề nghiệp, khu vực làm việc và thu nhập tạo thành một nhóm hầu như độc lập với nhóm các biến về nhân khẩu học. Như vậy di dân về Hà Nội, phải chăng, xuất phát từ hai căn nguyên chính như kết quả phân lớp nói trên?.

Phân lớp là một trong những bài toán đơn giản về mặt ý tưởng nhưng khá phức tạp về kỹ thuật tính toán. Ngày nay các phần mềm chuyên dụng đã được thiết kế cho các bài toán loại này. Tuy vậy, điều đáng chú ý khi sử dụng các phần mềm này là việc lựa chọn đúng các độ đo phù hợp nhất có thể, đối với từng bài toán cụ thể. SPSS có thể cung cấp nhiều hơn các kỹ thuật phân lớp, đặc biệt là ở các phiên bản sau 10.0, khi sử dụng các kỹ thuật mới cần tìm hiểu đầy đủ hơn cơ sở lý thuyết và đặc biệt là điều kiện sử dụng các phương pháp cũng như các độ đo khi phân lớp.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

7. TÓM TẮT CHƯƠNG 15

Nội dung chính của chương này là các phương pháp và các kỹ năng phân lớp các cá thể và các biến, bao gồm:

1. Các khái niệm khoảng cách khác nhau có thể sử dụng trong phân lớp
2. Tính chất của khoảng cách Ơ - cơ- lit và phi Ơ - cơ - lit.
3. Các chỉ số phân lớp đơn giản dựa trên các biến nhị phân
4. Các phương pháp kiểu đám mây động: phương pháp tâm di động; phương pháp đám mây di động
5. Phân lớp với các biến nhị phân
6. Phân lớp theo thứ bậc (các khái niệm, các khoảng cách, tính chất,...)
7. Tiêu chuẩn quán tính và phương pháp Ward: mức giảm quán tính khi ghép lớp; thuật toán Ward.
8. Trường hợp biến định tính
9. Phân lớp các biến
10. Kỹ năng sử dụng SPSS để phân lớp các cá thể và các biến: phân lớp với số lớp đã chỉ định; phân lớp thứ bậc.
11. Cách phân tích kết quả phân lớp.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

- Một số chỉ số độ đo phát tán:

Jaccard	$\frac{a}{a+b+c}$	(J)
Czekanowski (Dice)	$\frac{2a}{2a+b+c}$	(C-D)
Ochiai	$\frac{a}{\sqrt{(a+b)(a+c)}}$	(O)
Russel - Rao	$\frac{a}{a+b+c+d}$	(R-R)
Rogers - Tanismoto	$\frac{a+d}{a+d+2(b+c)}$	(R-T)

- Số cách chia nhóm n cá thể: $P_n = \sum_{i=1}^{n} C_{n-1}^{i-1} P_{n-i}$ với $P_0 = 0$.

- Tổng các quán tính của các lớp:

$$I_W = \sum_{j=1}^{k} P_j I_j = \sum_{j=1}^{k} P_j \frac{1}{n_j} \sum_{i=1}^{n_j} d^2(x_{ji}, g_j)$$

Trong đó P_j là trọng số của lớp j; x_{ji} điểm mô tả cá thể i của lớp j.

- Tổng quán tính giữa các lớp, ký hiệu I_B:

$$I_B = \sum_{i=1}^{n_j} P_j d^2(g_j, g)$$

- Công thức Huygens: $I = I_B + I_W$.

- Mức giảm quán tính khi ghép lớp

$$I_{AB} = P_A d^2(g_A, g) + P_B d^2(g_B, g) - (P_A + P_B)\, d^2(g_{AB}, g)$$

trong đó: $g_{AB} = \frac{P_A g_A + P_B g_B}{P_A + P_B}$

- Khoảng cách của A và B: $\delta(A,B) = \frac{P_A P_B}{P_A + P_B} d^2(g_A, g_B)$

- Khoảng cách này bởi khoảng cách giữa hai điểm i, j:

$$\delta(i,j) = \frac{p_i p_j}{p_i + p_j} d^2(i,j)$$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Các thuật ngữ cơ bản

Biểu đồ thang bậc	Agglomeration schedule
Các phương pháp phân lớp	Classifications methods
Phân lớp theo thứ bậc	Classification hierarchical
Phương pháp tâm di động	Method to centre dinamic
Phương pháp đám mây di động	Method the nuages dinamic

CÂU HỎI ÔN TẬP

15.1. Mục đích chung của bài toán phân lớp là gì?

15.2. Thế nào là khoảng cách phi Ơ- cơ - lit?

15.3. Giải thích ý nghĩa của các công thức 15.1- 15.5

15.4. Phương pháp kiểu tâm di động dựa trên tiêu chuẩn nào?

15.5. Phương pháp tâm di động và đám mây di động khác nhau cơ bản là gì?

15.6. Phân lớp theo thứ bậc khác với phương pháp kiểu tâm di động là gì?

15.7. Khi nào một họ các lớp của một tập hợp là họ có thứ bậc?

15.8. Thế nào là một siêu Metric?

15.9. Tóm tắt tiêu chuẩn quán tính

15.10. Thuật toán Ward dựa trên tiêu chuẩn nào?

15.11. Nêu quá trình thực hiện thuật toán Ward qua một thí dụ bằng số.

15.12. Phân lớp các biến có ý nghĩa gì?

15.13. Trong phân lớp các biến người ta dùng độ đo khoảng cách nào? Tại sao?

15.14. Nêu qui trình sử dụng SPSS đối với bài toán phân lớp. Tiến hành với 1 bộ số liệu cụ thể.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Với tệp DHS97.sav

15.15. Đối với số liệu trung bình theo vùng, phân nhóm phụ nữ hiện đã có con theo các chỉ tiêu sau: số con mong muốn, tuổi kết hôn lần đầu, số năm đến trường, số năm đến trường của chồng.

a- Thành 2 nhóm

b- Thành 7 nhóm

c- Theo thứ bậc

15.16. Với kết quả bài tập 1 (câu a và b) hãy so sánh kết quả phân nhóm này với tiêu thức nông thôn-thành thị và tiêu thức vùng.

Với tệp Hhexpn98.sav

15.17. Tạo tệp trung bình theo 10 vùng đối với các chỉ tiêu tiêu dùng hộ và số khẩu trong hộ. Phân lớp theo thứ bậc hộ đại diện của các vùng theo các chỉ tiêu chi tiêu của hộ theo giá hiện hành và số khẩu của hộ.

15.18. Tạo tệp trung bình của các chỉ tiêu như ở bài 3 với biến vùng và biến quint98, tiến hành phân nhóm các hộ thành 7 nhóm, xem xét quan hệ của kết quả phân nhóm này với chỉ tiêu vùng.

15.19. Tệp trung bình với hai tiêu thức theo vùng từ Hhexpn98.sav như sau:

Vùng	Số khẩu	Chi tiêu/tháng
1	4.61	10961.50
2	4.12	11174.12
3	4.61	11587.45
4	4.99	13137.30
5	5.48	14033.93
6	5.28	15664.63
7	5.11	13152.71

Sử dụng toán Ward phân lớp các vùng theo hai chỉ tiêu trên.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

15.20. Với chương trình và kết quả cơ bản sau, hãy cho biết các kết quả này mô tả nội dung gì? Có thể nhận được thông tin gì trên tệp số liệu mới?

```
GET
  FILE='F:\GTRINH\tketh\didanHN1.sav'.
EXECUTE .
QUICK CLUSTER
  qh marrital income edulevel
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(3) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER DISTANCE
  /PRINT INITIAL ANOVA.
Save outfile 'F:\GTRINH\tketh\didanHN1.sav'.
```

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	116.823	41.139	143.785
2	17.891	.740	55.481
3	19.027	.000	78.207
4	58.512	10.640	13.734
5	16.842	1.478	55.039
6	21.281	4.759	.000
7	1.482	.120	5.511
8	1.886	.000	10.670
9	.000	.000	.000

a. Convergence achieved due to no or small distance change. The maximum distance by which any center has changed is .000. The current iteration is 9. The minimum distance between initial centers is 499.069.

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Final Cluster Centers

	Cluster		
	1	2	3
M· quan hÖ	5	6	3
Marrital status	2	1	2
Income	246.3	23.3	636.7
Education level	3	3	3

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
M· quan hÖ	165.026	2	8.144	1220	20.264	.000
Marrital status	8.621	2	.350	1220	24.631	.000
Income	1.0E+07	2	2773.659	1220	3698.174	.000
Education level	19.668	2	.905	1220	21.741	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	214.000
	2	972.000
	3	37.000
Valid		1223.000
Missing		577.000

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TÀI LIỆU THAM KHẢO

1. Ngô Văn Thứ. *Thống kê thực hành*. NXB Khoa học và kỹ thuật. Hà Nội. 2008.
2- Tô Cẩm Tú: *Phân tích số liệu nhiều chiều*. NXB Nông nghiệp. Hà Nội 1996.
3. Benzecri. L'analyse des donnes en sociologie: Imprimerie des Presses Universitaires de France. 1990.
4. Gilbert Saporta: Analyse des donnees. Ecole national de la statistique et de l'administration economique. 1992-1993
5. Jean-marie bouroche et Gilbert Saporta: L'analyse des donnees. Que sais - Je. 1998.
6. Helene Erkel-Rousse: Analyse des donnees Applications et Programmation. L'ensae.1992.
7. Michel Volle: Analyse des donnees. Economica. 1994.
8. Alen Webster. Applied Statistics for business and ecconomics. Boston. 1992
9. Anderson T.W. An introduction to multivarate statistical analysis. New york. 1958
10. Dillon W and Goldstein M. Multivariate analysis Methods and Applications. New york . 1984
11. Kendall M and Stuart A, The advanced theory of Statistics. California. 1993
12. Base Syntax guid SPSS Ver. 7.5 for wwindows. SPSS Inc. 1997.
13. Dale J. Piorier: Intermadiate statistics and econometrics. 1995
14. David W. Hosmer and Stanley Lemeshow: Applied Logistic Regression. Wiley.1989.
15. L.Simar: An introduction to multivariate data analysis. Université catholique de Louvain
16. SPSS Base 8.0 Application guide. 1998.
17. T.W. Anderson: An introduction to multivariate statistical analysis. Wiley. 1984.

THƯ VIỆN HUBT

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

PHẦN PHỤ LỤC

CÁC BẢNG SỐ

Phụ lục 1: Phân phối nhị thức

Phụ lục 2: Phân phối Poisson

Phụ lục 3: Giá trị hàm e^{-x}

Phụ lục 4: Giá trị hàm

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

Phụ lục 5: Giá trị hàm

$$\Phi_0(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{z^2}{2}} dz$$

Phụ lục 6: Giá trị tới hạn U_α

Phụ lục 7: Giá trị tới hạn $\chi_\alpha^{2(n)}$

Phụ lục 8: Giá trị tới hạn $T_\alpha^{(n)}$

Phụ lục 9: Giá trị tới hạn $F_\alpha^{(n_1,n_2)}$

Phụ lục 10: Bảng số ngẫu nhiên

Phụ lục 11: Giá trị kiểm định tổng hạng Welcoxon

Phụ lục 12: Giá trị kiểm định tổng hạng theo dấu Welcoxon

Phụ lục 13: Giá trị phân phối Cochran

Phụ lục 14: Giá trị kiểm định Kolmogorov

Phụ lục 15: Giá trị kiểm định Lilliefors

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 1: Phân phối nhị thức
P(X = x) = b(x;n,p)

		p									
n	x	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1	0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0025	0,0100	0 0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,8574	0,7290	0,6141	0,5120	0,4239	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0312
	1	0,2036	0,3280	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1562
	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0000	0,0004	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1562
6	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0312
	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	00305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3310	0,2780	0,2344
	3	0,0021	0,0146	0,0415	0,0819	0,1338	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,0406	0,1240	0,2097	0,2153	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,2793	0,3826	0,3847	0,3355	0,2760	0,1977	0,1373	0,0896	0,0548	0,0312
	2	0,0515	0,1488	0,2376	0,2936	0,3135	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0231	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0000	0,0001	0,004	0,0012	0,0033	0,0079	0,0164	0,0312
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0007	0,0039

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 1*(tiếp theo)*

n	x	p = 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0277	0,0101	0,0046	0,0020
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	,0,1612	0,1110	0,0703
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2308	0,2119	0,16,41
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2315	0,1757	0,1209	0,0763	0,0439
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
11	0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
	1	0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054
	2	0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0533	0,0269
	3	0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806
	4	0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611
	5	0,0003	0,0025	0,0132	0,0388	0,0803	0,1231	0,1830	0,2207	0,2360	0,2256
	6	0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256
	7	0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611
	8	0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806
	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005
12	0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
	1	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
	2	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
	3	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
	4	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
	5	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
	6	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
	7	0,0000	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934
	8	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
	9	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0066	0,0161
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,00O0	0,0001	0,0003	0,0010	0,0029
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 1*(tiếp theo)*

n	x	p = 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
13	0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
	1	0,3312	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016
	2	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095
	3	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349
	4	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873
	5	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571
	6	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095
	7	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095
	8	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571
	9	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873
	10	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
14	0	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
	1	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009
	2	0,1229	0,2570	0,2912	0,2501	0,1502	0,1134	0,0634	0,0317	0,0141	0,0056
	3	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222
	4	0,0037	0,0349	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611
	5	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222
	6	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833
	7	0,0000	0,0002	0,0019	0,0092	0,0200	0,0618	0,1082	0,1574	0,1952	0,2095
	8	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833
	9	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222
	10	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
	2	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
	3	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
	4	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1266	0,0780	0,0417
	5	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916
	6	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
	7	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,201,3	0,1956
	8	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
	9	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527
	10	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
	13	0,0000	0,0000	0,0000	0,0060	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
16	0	0,4401	0,1853	0,0743	0,0283	0,0100	0,0033	0,0010	0,0003	0,0001	0,0000
	1	0,3706	0,3294	0,2097	0,1126	0,0535	0,0228	0,0087	0,0030	0,0009	0,0002
	2	0,1463	0,2745	0,2775	0,2111	0,1336	0,0732	0,0353	0,0150	0,0056	0,0018

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 1*(tiếp theo)*

		p									
n	x	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
16	3	0,0359	0,1423	0,2285	0,2463	0,2079	0,1465	0,0008	0,0468	0,0215	0,0083
	4	0,0061	0,0514	0,1311	0,2001	0,2252	0,2040	0,1553	0,1014	0,0572	0,0278
	5	0,0008	0,0137	0,0555	0,1201	0,1802	0,2099	0,2008	0,1623	0,1123	0,0667
	6	0,0001	0,0028	0,0108	0,0550	0,1101	0,1649	0,1982	0,1983	0,1684	0,1222
	7	0,0000	0,0004	0,0045	0,0197	0,0524	0,1010	0,1524	0,1889	0,1969	0,1746
	8	0,0000	0,0001	0,0009	0,0053	0,0197	0,0487	0,0923	0,1417	0,1812	0,1964
	9	0,0000	0,0000	0,0001	0,0012	0,0058	0,0185	0,0442	0,0840	0,1318	0,1746
	10	0,0000	0,0000	0,0000	0,0002	0,0014	0,0056	0,0167	0,0392	0,0755	0,1222
	11	0,0000	0,0000	0,0000	0,0000	0,0002	0,0013	0,0049	0,0142	0,0337	0,0667
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0040	0,0115	0,0278
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0029	0,0085
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
17	0	0,4181	0,1668	0,0631	0,0225	0,0075	0,0023	0,0007	0,0002	0,0000	0,0000
	1	0,3741	0,3150	0,1893	0,0957	0,0426	0,0169	0,0060	0,0019	0,0005	0,0001
	2	0,1575	0,2800	0,2673	0,1914	0,1136	0,0581	0,0260	0,0102	0,0035	0,0010
	3	0,0415	0,1556	0,2359	0,2393	0,1893	0,1245	0,0701	0,0341	0,0144	0,0052
	4	0,0076	0,0605	0,1457	0,2093	0,2209	0,1868	0,1320	0,0796	0,0411	0,0182
	5	0,0010	0,0175	0,0668	0,1361	0,1914	0,2081	0,1849	0,1379	0,0875	0,0472
	6	0,0001	0,0039	0,0236	0,0680	0,1276	0,1784	0,1991	0,1839	0,1432	0,0944
	7	0,0000	0,0007	0,0065	0,0267	0,0668	0,1201	0,1685	0,1927	0,1841	0,1484
	8	0,0000	0,0001	0,0014	0,0084	0,0279	0,0644	0,1143	0,1606	0,1883	0,1855
	9	0,0000	0,0000	0,0003	0,0021	0,0093	0,0276	0,0611	0,1070	0,1540	0,1855
	10	0,0000	0,0000	0,0000	0,0004	0,0025	0,0095	0,0263	0,0571	0,1008	0,1404
	11	0,0000	0,0000	0,0000	0,0001	0,0005	0,0026	0,0090	0,0242	0,0525	0,0944
	12	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0081	0,0215	0,0472
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0021	0,0068	0,0182
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000,	0,0001	0,0004	0,0016	0,0052
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
18	0	0,3972	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
	1	0,3763	0,3002	0,1704	0,0811	0,0338	0,0126	0,0042	0,0012	0,0003	0,0001
	2	0,1683	0,2835	0,2556	0,1723	0,0958	0,0458	0,0190	0,0069	0,0022	0,0006
	3	0,0473	0,1680	0,2406	0,2297	0,1704	0,1046	0,0547	0,0246	0,0095	0,0033
	4	0,0093	0,0700	0,1592	0,2153	0,2130	0,1681	0,1104	0,0614	0,0291	0,0117
	5	0,0014	0,0218	0,0787	0,1507	0,1988	0,2017	0,1664	0,1146	0,0666	0,0327
	6	0,0002	0,0052	0,0316	0,0816	0,1436	0,1873	0,1941	0,1655	0,1181	0,0708
	7	0,0000	0,0010	0,0091	0,0350	0,0820	0,1376	0,1792	0,1892	0,1657	0,1214
	8	0,0000	0,0002	0,0022	0,0120	0,0376	0,0811	0,1327	0,1734	0,1864	0,1669
	9	0,0000	0,0000	0,0004	0,0033	0,0139	0,0386	0,0794	0,1284	0,1694	0,1855
	10	0,0000	0,0000	0,0001	0,0008	0,0042	0,0149	0,0385	0,0771	0,1248	0,1669
	11	0,0000	0,0000	0,0000	0,0001	0,0010	0,0046	0,0151	0,0374	0,0742	0,1214
	12	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0047	0,0145	0,0354	0,0708
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0045	0,0134	0,0327
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0039	0,0117

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 1*(tiếp theo)*

n	x	p									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0031
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
19	0	0,3774	0,1351	0,0456	0,0044	0,0042	0,0011	0,0003	0,0001	0,0000	0,0000
	1	0,3774	0,2852	0,1529	0,0685	0,0268	0,0093	0,0029	0,0008	0,0002	0,0000
	2	0,1787	0,2852	0,2428	0,1540	0,0803	0,0358	0,0138	0,0046	0,0013	0,0003
	3	0,0533	0,1796	0,2428	0,2182	0,1517	0,0869	0,0422	0,0175	0,0062	0,0018
	4	0,0112	0,0798	0,1714	0,2182	0,2023	0,1491	0,0909	0,0467	0,0203	0,0074
	5	0,0018	0,0266	0,0907	0,1636	0,2023	0,1916	0,1468	0,0933	0,0497	0,0222
	6	0,0002	0,0069	0,0374	0,0955	0,1374	0,1916	0,1844	0,1451	0,0949	0,0518
	7	0,0000	0,0014	0,0122	0,0443	0,0974	0,1525	0,1844	0,1797	0,1443	0,0961
	8	0,0000,	0,0002	0,0032	0,0166	0,0487	0,0981	0,1489	0,1797	0,1771	0,1442
	9	0,0000	0,0000	0,0007	0,0051	0,0198	0,0514	0,0980	0,1464	0,1771	0,1762
	10	0,0000	0,0000	0,0001	0,0013	0,0066	0,0220	0,0528	0,0976	0,1449	0,1762
	11	0,0000	0,0000	0,0000	0,0003	0,0018	0,0077	0,0233	0,0532	0,0970	0,1442
	12	0,0000	0,0000	0,0000	0,0000	0,0004	0,0022	0,0083	0,0237	0,0529	0,0961
	13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0024	0,0085	0,0233	0,0518
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0082	0,0222
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0022	0,0074
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	0	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
	1	0,3774	0,2702	0,1368	0,0576	0,0211	0,0068	0,0020	0,0005	0,0001	0,0000
	2	0,1887	0,2852	0,2293	0,1369	0,0669	0,0278	0,0100	0,0031	0,0008	0,0002
	3	0,0596	0,1901	0,2428	0,2054	0,1339	0,0716	0,0323	0,0123	0,0040	0,0011
	4	0,0133	0,0898	0,1821	0,2182	0,1897	0,1304	0,0738	0,0350	0,0139	0,0046
	5	0,0022	0,0319	0,1028	0,1746	0,2023	0,17*9	0,1272	0,0746	0,0365	0,0148
	6	0,0003	0,0089	0,0454	0,1091	0,1686	0,1916	0,1712	0,1244	0,0746	0,0370
	7	0,0000	0,0020	0,0160	0,0545	0,1124	0,1643	0,1844	0,1659	0,1221	0,0739
	8	0,0000	0,0004	0,0046	0,0222	0,0609	0,1144	0,1614	0,1797	0,1623	0,1201
	9	0,0000	0,0001	0,0011	0,0074	0,0271	0,0654	0,1158	0,1597	0,1771	0,1602
	10	0,0000	0,0000	0,0002	0,0020	0,0099	0,0308	0,0686	0,1171	0,1593	0,1762
	11	0,0000	0,0000	0,0000	0,0005	0,0030	0,0120	0,0336	0,0710	0,1185	0,1602
	12	0,0000	0,0000	0,0000	0,0001	0,0008	0,0039	0,0136	0,0355	0,0727	0,1201
	13	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0045	0,0146	0,0366	0,0739
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0049	0,0130	0,0370
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0049	0,0148
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000.	0,0003	0,0013	0,0046
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 2: Phân phối poisson p(X;λ)

x	λ									
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1433	0,1647	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0000	0,0001	0,0002	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020	0,0031
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

x	λ									
	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496	0,1351
1	0,3662	0,3614	0,3543	0,3452	0,3347	0,3230	0,3106	0,2975	0,2842	0,2707
2	0,2014	0,2169	0,2303	0,2417	0,2510	0,2584	0,2640	0,2678	0,2700	0,2702
3	0,0738	0,0867	0,0998	0,1128	0,1255	0,1378	0,1496	0,1607	0,1710	0,1804
4	0,0203	0,0260	0,0324	0,0395	0,0471	0,0551	0,0636	0,0723	0,0812	0,0902
5	0,0045	0,0062	0,0034	0,0111	0,0141	0,0176	0,0216	0,0260	0,0309	0,0361
6	0,0008	0,0002	0,0018	0,0026	0,0035	0,0047	0,0061	0,0078	0,0098	0,0120
7	0,0001	0,0002	0,0003	0,0005	0,0008	0,0011	0,0015	0,0020	0,0027	0,0034
8	0,0000	0,0000	0,0000	0,0001	0,0001,	0,0002	0,0003	0,0005	0,0006	0,0009
9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002

x	λ									
	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608	0,0550	0,0498
1	0,2572	0,2438	0,2306	0,2177	0,2052	0,1931	0,1815	0,1703	0,1596	0,1494
2	0,2700	0,2681	0,2652	0,2613	0,2565	0,2510	0,2450	0,2384	0,2314	0,2240
3	0,1890	0,1966	0,2033	0,2090	0,2138	0,2176	0,2205	0,2225	0,2237	0,2240
4	0,0992	0,1002	0,1169	0,1254	0,1336	0,1414	0,1408	0,1557	0,1622	0,1680
5	0,0417	0,0476	0,0538	0,0602	0,0668	0,0735	0,0804	0,0872	0,0940	0,1008
6	0,0146	0,0174	0,0206	0,0241	0,0278	0,0319	0,0362	0,0407	0,0455	0,0540
7	0,0044	0,0056	0,0068	0,0083	0,0099	0,0118	0,0139	0,0163	0,0188	0,0216
8	0,0011	0,0015	0,0019	0,0025	0,0031	0,0038	0,0047	0,0057	0,0068	0,0081
9	0,0003	0,0004	0,0005	0,0007	0,0009	0,0011	0,0014	0,0018	0,0022	0,0027
10	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0004	0,0005	0,0006	0,0003
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002	0,0002
12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

x	λ									
	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
0	0,0450	0,0408	0,0369	0,0344	0,0302	0,0273	0,0247	0,0224	0,0202	0,0183
1	0,1397	0,1304	0,1217	0,1135	0,1057	0,0984	0,0915	0,0850	0,0789	0,0733
2	0,2165	0,2087	0,2008	0,1929	0,1050	0,1771	0,1692	0,1615	0,1539	0,1465
3	0,2237	0,2226	0,2209	0,2186	0,2158	0,2125	0,2087	0,2046	0,2001	0,1954
4	0,1734	0,1781	0,1823	0,1850	0,1888	0,1912	0,1931	0,1944	0,1951	0,1954

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 2 *(tiếp theo)*

x	λ									
	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
5	0,1075	0,1140	0,1203	0,1264	0,1322	0,1377	0,1429	0,1477	0,1522	0,1563
6	0,0555	0,0608	0,0662	0,0716	0,0771	0,0826	0,0881	0,0936	0,0989	0,1042
7	0,0246	0,0278	0,0312	0,0348	0,0385	0,0425	0,0466	0,0508	0,0551	0,0595
8	0,0095	0,0111	0,0129	0,0148	0,0169	0,0191	0,0215	0,0241	0,0269	0,0298
9	0,0033	0,0040	0,0047	0,0056	0,0066	0,0076	0,0089	0,0102	0,0116	0,0132
10	0,0010	0,0013	0,0016	0,0019	0,0023	0,0028	0,0033	0,0039	0,0045	0,0053
11	0,0003	0,0004	0,0005	0,0006	0,0007	0,0009	0,0011	0,0013	0,0016	0,0019
12	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006
13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

x	λ									
	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
0	0,0166	0,0150	0,0136	0,0123	0,0111	0,0101	0,0091	0,0082	0,0074	0,0067
1	0,0679	0,0630	0,0583	0,0540	0,0500	0,0462	0,0427	0,0395	0,0365	0,0337
2	0,1393	0,1323	0,1254	0,1188	0,1125	0,1063	0,1005	0,0948	0,0894	0,0842
3	0,1904	0,1852	0,1790	0,1743	0,1687	0,1631	0,1574	0,1517	0,1460	0,1404
4	0,1951	0,1944	0,1933	0,1917	0,1898	0,1875	0,1849	0,1820	0,1789	0,1755
5	0,1600	0,1633	0,1662	0,1687	0,1708	0,1725	0,1738	0,1747	0,1753	0,1755
6	0,1093	0,1143	0,1191	0,1237	0,1281	0,1323	0,1362	0,1398	0,1432	0,1462
7	0,0640	0,0686	0,0732	0,0778	0,0824	0,0869	0,0914	0,0959	0,1002	0,1044
8	0,0328	0,0360	0,0393	0,0428	0,0463	0,0500	0,0537	0,0575	0,0614	0,0653
9	0,0150	0,0168	0,0188	0,0209	0,0232	0,0255	0,0280	0,0307	0,0334	0,0363
10	0,0061	0,0071	0,0081	0,0092	0,0104	0,0118	0,0132	0,0147	0,0164	0,0181
11	0,0023	0,0027	0,0032	0,0037	0,0043	0,0049	0,0056	0,0064	0,0073	0,0082
12	0,0008	0,0009	0,0011	0,0014	0,0016	0,0019	0,0022	0,0026	0,0030	0,0034
13	0,0002	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013
14	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005
15	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002

x	λ									
	5,1	5,2	5,3	3,4	5,5	5,6	5,7	5,8	5,9	6,0
0	0,0061	0,0055	0,0050	0,0045	0,0041	0,0037	0,0033	0,0030	0,0027	0,0025
1	0,0311	0,0287	0,0265	0,0244	0,0225	0,0207	0,0191	0,0176	0,0162	0,0149
2	0,0793	0,0746	0,0701	0,0659	0,0618	0,0580	0,0544	0,0509	0,0477	0,0446
3	0,1348	0,1293	0,1239	0,1105	0,1133	0,1082	0,1033	0,0985	0,0938	0,0892
4	0,1719	0,1681	0,1641	0,1600	0,1558	0,1515	0,1472	0,1428	0,1333	0,1339
5	0,1753	0,1748	0,1740	0,1728	0,1714	0,1697	0,1678	0,1656	0,1632	0,1606
6	0,1490	0,1515	0,1537	0,1555	0,1571	0,1584	0,1594	0,1601	0,1605	0,1606
7	0,1086	0,1125	0,1163	0,1200	0,1234	0,1267	0,1298	0,1326	0,1353	0,1377
8	0,0692	0,0731	0,0771	0,0810	0,0849	0,0087	0,0925	0,0962	0,0998	0,1033
9	0,0392	0,0423	0,0454	0,0486	0,0519	0,0552	0,0586	0,0620	0,0654	0,0688

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 2 *(tiếp theo)*

x	λ 5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	6,0
10	0,0200	0,0220	0,0241	0,0262	0,0285	0,0309	0,0334	0,0359	0,0386	0,0413
11	0,0093	0,0104	0,0116	0,0129	0,0143	0,0157	0,0173	0,0190	0,0207	0,0225
12	0,0039	0,0045	0,0051	0,0058	0,0065	0,0073	0,0082	0,0092	0,0102	0,0113
13	0,0015	0,0018	0,0021	0,0024	0,0025	0,0032	0,0036	0,0041	0,0046	0,0052
14	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013	0,0015	0,0017	0,0019	0,0022
15	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009
16	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001

x	λ 6,1	6,2	6,3	6,4	6,5	6,6	6,7	6,8	6,9	7,0
0	0,0022	0,0020	0,0018	0,0017	0,0015	0,0014	0,0012	0,0011	0,0010	0,0009
1	0,0137	0,0126	0,0116	0,0106	0,0098	0,0090	0,0082	0,0076	0,0070	0,0064
2	0,0417	0,0390	0,0364	0,0340	0,0318	0,0296	0,0276	0,0258	0,0240	0,0223
3	0,0848	0,0806	0,0765	0,0726	0,0688	0,0652	0,0617	0,0584	0,0552	0,0521
4	0,1294	0,1249	0,1205	0,1162	0,1118	0,1076	0,1034	0,0992	0,0952	0,0912
5	0,1579	0,1549	0,1519	0,1487	0,1454	0,1420	0,1385	0,1349	0,1314	0,1277
6	0,1605	0,1601	0,1595	0,1586	0,1575	0,1562	0,1546	0,1529	0,1511	0,1490
7	0,1399	0,1418	0,1435	0,1450	0,1462	0,1472	0,1480	0,1486	0,1489	0,1490
8	0,1066	0,1099	0,1130	0,1160	0,1188	0,1215	0,1240	0,1263	0,1284	0,1304
9	0,0723	0,0757	0,0791	0,0825	0,0858	0,0891	0,0923	0,0954	0,0985	0,1014
10	0,0441	0,0469	0,0498	0,0528	0,0558	0,0388	0,0618	0,0649	0,0679	0,0710
11	0,0245	0,0265	0,0285	0,0307	0,0330	0,0353	0,0377	0,0401	0,0426	0,0452
12	0,0124	0,0137	0,0150	0,0164	0,0179	0,0194	0,0210	0,0227	0,0245	0,0264
13	0,0058	0,0065	0,0073	0,0081	0,0089	0,0098	0,0108	0,0119	0,0130	0,0142
14	0,0025	0,0029	0,0033	0,0037	0,0041	0,0046	0,0052	0,0058	0,0064	0,0071
15	0,0010	0,0012	0,0014	0,0016	0,0018	0,0020	0,0023	0,0026	0,0029	0,0033
16	0,0004	0,0005	0,0003	0,0006	0,0007	0,0008	0,0010	0,0011	0,0013	0,0014
17	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006
18	0,0000	0,0001	%0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001

x	λ 7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
0	0,0008	0,0007	0,0007	0,0006	0,0006	0,0005	0,0005	0,0004	0,0004	0,0003
1	0,0059	0,0054	0,0049	0,0045	0,0041	0,0038	0,0035	0,0032	0,0029	0,0027
2	0,0208	0,0194	0,0180	0,0167	0,0156	0,0145	0,0134	0,0125	0,0116	0,0107
3	0,0492	0,0464	0,0438	0,0413	0,0339	0,0366	0,0343	0,0324	0,0305	0,0286
4	0,0874	0,0836	0,0799	0,0764	0,0729	0,0696	0,0663	0,0632	0,0602	0,0573
5	0,1241	0,1204	0,1167	0,1130	0,1094	0,1057	0,1021	0,0986	0,0951	0,0916
6	0,1468	0,1445	0,1420	0,1394	0,1367	0,1339	0,1311	0,1282	0,1252	0,1221
7	0,1489	0,1486	0,1481	0,1474	0,1465	0,1454	0,1442	0,1428	0,1413	0,1396
8	0,1321	0,1337	0,1351	0,1363	0,1373	0,1382	0,1388	0,1392	0,1395	0,1396
9	0,1042	0,1070	0,1096	0,1121	0,1144	0,1167	0,1187	0,1207	0,1224	0,1241
10	0,0740	0,0770	0,0800	0,0829	0,0058	0,0887	0,0914	0,0941	0,0967	0,0993
11	0,0478	0,0504	0,0531	0,0558	0,0585	0,0613	0,0640	0,0667	0,0695	0,0722

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 2 *(tiếp theo)*

x	λ									
	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
12	0,0283	0,0303	0,0323	0,0344	0,0366	0,0388	0,0411	0,0434	0,0457	0,0481
13	0,0154	0,0168	0,0181	0,0196	0,0211	0,0227	0,0243	0,0260	0,0278	0,0296
14	0,0078	0,0086	0,0095	0,0104	0,0113	0,0723	0,0134	0,0145	0,0157	0,0169
15	0,0037	0,0041	0,0046	0,0051	0,0057	0,0062	0,0069	0,0075	0,0083	0,0090
16	0,0016	0,0019	0,0021	0,0024	0,0026	0,0030	0,0033	0,0037	0,0041	0,0045
17	0,0007	0,0008	0,0009	0,0010	0,0012	0,0013	0,0015	0,0017	0,0039	0,0021
18	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0006	0,0009
19	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0003	0,0004
20	0,0000	0,0000	0,0001	0,0001	0,0001	0,0000	0,0001	0,0001	0,0001	0,0002
21	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0001

x	λ									
	8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,9	9,0
0	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001
1	0,0025	0,0023	0,0021	0,0019	0,0017	0,0016	0,0014	0,0013	0,0012	0,0011
2	0,0100	0,0092	0,0086	0,0079	0,0074	0,0068	0,0063	0,0058	0,0054	0,0050
3	0,0269	0,0252	0,0237	0,0222	0,0208	0,0195	0,0183	0,0171	0,0160	0,0150
4	0,0544	0,0517	0,0493	0,0466	0,0443	0,0420	0,0398	0,0377	0,0357	0,0337
5	0,0882	0,0849	0,0816	0,0784	0,0752	0,0722	0,0692	0,0663	0,0635	0,0607
6	0,1191	0,1160	0,1128	0,1097	0,1066	0,1034	0,1003	0,0972	0,0941	0,0911
7	0,1378	0,1358	0,1338	0,1317	0,1294	0,1271	0,1247	0,1222	0,1197	0,1171
8	0,1395	0,1392	0,1338	0,1382	0,3375	0,1366	0,1356	0,1344	0,1332	0,1318
9	0,1256	0,1269	0,1280	0,1290	0,1299	0,1306	0,1311	0,1315	0,1311	0,1318
10	0,1017	0,1040	0,1063	0,1084	0,1104	0,1123	0,1140	0,1157	0,1172	0,1186
11	0,0749	0,0776	0,0802	0,0828	0,0853	0,0870	0,0902	0,0925	0,0948	0,0970
12	0,0505	0,0530	0,0555	0,0579	0,0604	0,0629	0,0654	0,0679	0,0703	0,0728
13	0,0335	0,0334	0,0354	0,0374	0,0395	0,0416	0,0438	0,0459	0,0481	0,0504
14	0,0182	0,0196	0,0210	0,0225	0,0240	0,0256	0,0272	0,0289	0,0306	0,0324
15	0,0098	0,0107	0,0116	0,0126	0,0136	0,0147	0,0158	0,0169	0,0182	0,0194
16	0,0050	0,0055	0,0060	0,0066	0,0072	0,0079	0,0086	0,0093	0,0101	0,0109
17	0,0024	0,0026	0,0029	0,0033	0,0036	0,0040	0,0044	0,0048	0,0053	0,0058
18	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019	0,0021	0,0024	0,0026	0,0029
19	0,0005	0,0005	0,0006	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014
20	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0005	0,0006
21	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003
22	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

x	λ									
	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10
0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0000
1	0,0010	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0005	0,0005	0,0005
2	0,0046	0,0043	0,0040	0,0037	0,0034	0,0031	0,0029	0,0027	0,0025	0,0023
3	0,0140	0,0131	0,0123	0,0115	0,0107	0,0100	0,0093	0,0087	0,0081	0,0076
4	0,0319	0,0302	0,0285	0,0269	0,0254	0,0240	0,0226	0,0213	0,0201	0,0189

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 2 *(tiếp theo)*

x	λ									
	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10,0
5	0,0581	0,0555	0,0530	0,0504	0,0483	0,0460	0,0439	0,0410	0,0393	0,0378
6	0,0881	0,0851	0,0322	0,0793	0,0764	0,0736	0,0709	0,0682	0,0656	0,0631
7	0,1145	0,1118	0,1091	0,1064	0,1037	0,1010	0,0932	0,0955	0,0928	0,0901
8	0,1302	0,1236	0,1269	0,1251	0,1232	0,1212	0,1191	0,1170	0,1148	0,1326
9	0,1317	0,1315	0,1311	0,1306	0,1300	0,1293	0,1284	0,1274	0,1263	0,1251
10	0,1198	0,1210	0,1219	0,1228	0,1235	0,1241	0,1245	0,1249	0,1250	0,1251
11	0,0991	0,1012	0,1031	0,1049	0,1067	0,1083	0,1093	0,1112	0,1125	0,1137
12	0,0752	0,0776	0,0799	0,0822	0,0844	0,0866	0,0888	0,0908	0,0928	0,0948
13	0,0526	0,0549	0,0572	0,0594	0,0417	0,0440	0,0662	0,0685	0,0707	0,0729
14	0,0342	0,0361	0,0380	0,0399	0,0419	0,0439	0,0459	0,0479	0,0500	0,0521
15	0,0208	0,0221	0,0235	0,0250	0,0265	0,0281	0,0297	0,0313	0,0330	0,0347
16	0,0118	0,0127	0,0137	0,0147	0,0157	0,0160	0,0180	0,0192	0,0204	0,0217
17	0,0063	0,0069	0,0075	0,0081	0,0088	0,0095	0,0103	0,0111	0,0119	0,0128
18	0,0032	0,0035	0,0039	0,0042	0,0046	0,0051	0,0055	0,0060	0,0065	0,0071
19	0,0015	0,0017	0,0019	0,0021	0,0023	0,0026	0,0028	0,0031	0,0034	0,0037
20	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019
21	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
22	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004
23	0,0000	0,0001	0,0001	0,0001	0,0003	0,0001	0,0001	0,0001	0,0002	0,0002
24	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001

x	λ									
	11	12	13	14	15	16	17	18	19	20
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0010	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3	0,0037	0,0018	0,0008	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000
4	0,0102	0,0053	0,0027	0,0013	0,0006	0,0003	0,0001	0,0001	0,0000	0,0000
5	0,0224	0,0127	0,0070	0,0037	0,0019	0,0010	0,0005	0,0002	0,0001	0,0001
6	0,0411	0,0255	0,0152	0,0087	0,0048	0,0026	0,0014	0,0007	0,0004	0,0002
7	0,0646	0,0437	0,0281	0,0174	0,0104	0,0060	0,0034	0,0018	0,0010	0,0005
8	0,0888	0,0655	0,0457	0,0304	0,0194	0,0120	0,0072	0,0042	0,0024	0,0013
9	0,1085	0,0874	0,0661	0,0473	0,0324	0,0213	0,0135	0,0003	0,0050	0,0029
10	0,1194	0,1048	0,0859	0,0663	0,0486	0,0341	0,0230	0,0150	0,0095	0,0058
11	0,1194	0,1144	0,1015	0,0844	0,0663	0,0496	0,0355	0,0245	0,0164	0,0106
12	0,1094	0,1144	0,1099	0,0884	0,0029	0,0661	0,0504	0,0368	0,0259	0,0176
13	0,0926	0,1056	0,1099	0,1060	0,0956	0,0814	0,0658	0,0509	0,0378	0,0271
14	0,0728	0,0905	0,1021	0,1060	0,1024	0,0930	0,0800	0,0655	0,0514	0,0307
15	0,0534	0,0724	0,0885	0,0992	0,1024	0,0985	0,0906	0,0786	0,0650	0,0516
16	0,0367	0,0543	0,0719	0,0866	0,0960	0,0992	0,0963	0,0884	0,0772	0,0646
17	0,0237	0,0383	0,0550	0,0713	0,0847	0,0934	0,0963	0,0936	0,0863	0,0760
18	0,0145	0,0256	0,0397	0,0554	0,0706	0,0030	0,0909	0,0936	0,0911	0,0844
19	0,0084	0,0161	0,0272	0,0409	0,0557	0,0699	0,0814	0,0887	0,0911	0,0888
20	0,0046	0,0097	0,0177	0,0286	0,0418	0,0559	0,0692	0,0798	0,0366	0,0888
21	0,0024	0,0055	0,0109	0,0191	0,0299	0,0426	0,0560	0,0684	0,0783	0,0846
22	0,0012	0,0030	0,0065	0,0121	0,0204	0,0310	0,0433	0,0560	0,0676	0,0769
23	0,0006	0,0016	0,0037	0,0074	0,0133	0,0216	0,0320	0,0438	0,0559	0,0669

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 2 *(tiếp theo)*

x	λ									
	11	12	13	14	15	16	17	18	19	20
24	0,0003	0,0008	0,0020	0,0083	0,0033	0,0144	0,0226	0,0328	0,0442	0,0557
25	0,0001	0,0004	0,0010	0,0024	0,0050	0,0092	0,0154	0,0237	0,0336	0,0446
26	0,0000	0,0002	0,0005	0,0013	0,0029	0,0057	0,0101	0,0164	0,0246	0,0343
27	0,0000	0,0001	0,0002	0,0007	0,0016	0,0034	0,0063	0,0109	0,0173	0,0254
28	0,0000	0,0000	0,0001	0,0003	0,0009	0,0019	0,0038	0,0070	0,0117	0,0181
29	0,0000	0,0000	0,0001	0,0002	0,0004	0,0011	0,0023	0,0044	0,0077	0,0125
30	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0013	0,0026	0,0049	0,0083
31	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0007	0,0015	0,0030	0,0054
32	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0004	0,0009	0,0018	0,0034
33	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0010	0,0020
34	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0012
35	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0007
36	0,0000	4,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004
37	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
31	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
39	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

Phụ lục 3: Phân phối luỹ thừa

x	e^{-x}	x	e^{-x}	x	e^{-x}
0,0	1,000	1,5	0,223	3,0	0,050
0,1	0,905	1,6	0,202	3,1	0,045
0,2	0.819	1,7	0,183	3,2	0,041
0,3	0,741	1,8	0,165	3,3	0,037
0,4	0,670	1,9	0,150	3,4	0,033
0,5	0,607	2,0	0,135	3,5	0,030
0,6	0,549	2,1	0,122	3,6	0,027
0,7	0,497	2,2	0,111	3,7	0,025
0,8	0,449	2,3	0,100	3,8	0,022
0,9	0,407	2,4	0,091	3,9	0,020
1,0	0,368	2,5	0,082	4,0	0,018
1,1	0,333	2,6	0,074	4,5	0,011
1,2	0,301	2,7	0,067	5,0	0,007
1.3	0,273	2,8	0,061	6,0	0,002
1,4	0,247	2,9	0,055	7,0	0,001

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 4: Giá trị hàm $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$

	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0303	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 5: Giá trị hàm $\Phi_0(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-z^2/2} dZ$

u	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0,00	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.10	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.20	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.30	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.40	.1554	.1591	1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.50	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.60	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.70	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.80	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.90	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.00	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.10	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.20	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.30	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.40	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.50	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.60	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.70	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.80	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.90	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.00	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.10	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.20	.4861	.4864	.4868	.4871	.4875	.4878	.4881	4884	.4887	.4890
2.30	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.40	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.50	.4938	.4940	.4941	.4943	4945	.4946	.4948	.4949	.4951	.4952
2.60	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.70	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.80	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.90	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.00	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

u	Area
3.50	.49976737
4.00	.49996833
4.50	.49999660
5.00	.49999971

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 6: Giá trị tới hạn chuẩn

u	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.10	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.20	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.30	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.40	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.50	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.60	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.70	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.80	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.90	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.00	.1587	.1562	.1539	.1515	.1492	.1469	1446	.1423	.1401	.1379
1.10	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.20	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.30	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.40	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.50	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.60	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.70	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.80	.0359	.0331	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.90	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.00	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.10	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.20	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.30	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.40	.0082	.0080	.0078	.0073	.0073	.0071	.0069	.0068	.0066	.0064
2.50	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.60	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.70	.0035	.0031	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.80	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.90	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.00	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

u	Area
3.500	.00023263
4.000	.00003167
4.500	.00000340
5.000	.00000029

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 7: Giá trị tới hạn χ^2

df	α = .999	α = .995	α = .99	α = .975	α = .95	α = .9
1	.000002	.000039	.000157	.000982	.003932	.01579
2	.002001	.01003	.02010	.05064	.1026	.2107
3	.02430	.07172	.1148	.2158	.3518	.5844
4	.09080	.2070	.2971	.4844	.7107	1.064
5	.2102	.4117	.5543	.8312	1.145	1.610
6	.3811	.6757	.8721	1.237	1.635	2.204
7	.5985	.9893	1.239	1.690	2.167	2.833
8	.8571	1.344	1.646	2.180	2.733	3.490
9	1.152	1.735	2.088	2.700	3.325	4.168
10	1.479	2.156	2.558	3.247	3.940	4.865
11	1.834	2.603	3.053	3.816	4.575	5.578
12	2.214	3.074	3.571	4.404	5.226	6.304
13	2.617	3.565	4.107	5.009	5.892	7.042
14	3.041	4.075	4.660	5.629	6.571	7.790
15	3.483	4.601	5 229	6.262	7.261	8.547
16	3.942	5.142	5.812	6.908	7.962	9.312
17	4.416	5.697	6.408	7.564	8.672	10.09
18	4.905	6.265	7.015	8.231	9 390	10.86
19	5.407	6.844	7,633	8.907	10.12	11.65
20	5.921	7 .434	8.260	9.591	10.85	12.44
21	6.447	8.034	8.897	10.28	11.59	13.24
22	6.983	8.643	9.542	10.98	12.34	14.04
23	7.529	9.260	10.20	11.69	13.09	14.85
24	8.085	9.886	10.86	12.40	1385	15.66
25	8.649	10.52	11.32	13.12	14.61	16.47
26	9.222	11.16	12.20	13.84	15.38	17.29
27	9.803	11.81	12.88	14.57	16.15	18.11
28	10.39	12.46	13.56	15 31	16.93	18.94
29	10.99	13.12	14.26	16.06	17.71	19.77
30	11.59	13.79	14.95	16.79	18.49	20.60
40	17.92	20.71	22.16	24.43	26.51	29.05
50	24.67	27.99	29.71	32.36	34.76	37.69
60	31.74	35.53	37.48	40.48	43.19	46.46
70	39.04	43.28	45.44	48.76	51.74	55.33
80	46.52	51.17	53.54	57.13	60.39	64.28
90	54.16	59.20	61.75	65.65	69.13	73.29
100	61.92	67.33	70.06	74.22	77.93	82.36
120	77.76	83.85	86.92	91.57	95.70	100.62
240	177.95	187.32	191.99	198.98	205.14	212.39

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 7 *(tiếp theo)*

α = .1	α = .05	α = .025	α = .01	α = 0.005	α = .001	df
2.706	3.841	5.024	6.635	7.879	10.83	1
4.605	5.991	7.378	9.210	10.60	13.82	2
6.251	7.815	9.348	11.34	12.84	16.27	3
7.779	9.488	11.14	13.28	14.86	18.47	4
9.236	11.07	12.83	15.09	16.75	20.52	5
10.64	12.59	14.45	16.81	18.53	22.46	6
12.02	14.07	16.01	18.48	20.28	24.32	7
13.36	15.51	17.51	20.09	21 95	26.12	8
14.68	16.92	19.02	21.67	23.59	27.88	9
16.99	18.31	20.48	23.21	25.19	29.59	10
17.28	19.68	21.92	24.72	26.76	31.27	11
18.55	21.03	23.34	26.22	38.10	32.91	12
19.81	22.36	24 74	27.69	29.82	34.53	13
21.06	23.68	26.12	29.14	11.32	36.12	14
22.31	25.00	27.49	30.58	32.80	37.70	15
23.54	26.30	28.8.5	32.00	34.27	39.25	16
24.77	27 59	30.19	3341	35.72	40.79	17
25.99	28.87	31.53	34.81	37.16	42.31	18
27.20	30.14	32.85	36.19	38.58	43.82	19
28.41	31.41	34.17	37.57	40.00	45.31	20
29.62	32.67	35.48	38.93	41.40	46.80	21
30.81	33.92	36.78	40.29	42.80	48.27	22
32.01	35.17	38.08	41.64	44.18	49.73	23
33.20	36,42	39.36	42.98	45.56	51.18	24
34.38	37.65	40.65	44.31	46.93	52.62	25
35.56	38.89	41.92	45.64	48.29	54.05	26
36.74	40.11	43.19	46.96	49.65	55.48	27
37.92	41.34	44.46	48.28	50.99	56.89	28
39.09	42.56	45.72	49.59	52.34	58.30	29
40.26	43.77	46.98	50.89	53.67	59.70	30
51.81	55.76	59.34	63.69	66.77	73.40	40
63.17	67.50	71.42	76.15	79.49	86.66	60
74.40	79.08	83.30	88.38	91.95	99.61	60
85.53	90.53	95.02	100.43	104.21	112.32	70
96.58	101.88	106.63	112.33	116.32	124.84	80
107.57	113.15	118.14	124.12	128.30	137.21	90
118.50	124.34	129.56	135.81	140.17	149.45	100
140.23	146.57	152.21	158.95	163.65	173.62	120
268.47	277.14	284.80	293.89	300.18	313.44	240

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 8: Giá trị tới hạn student

df	α = .1	α = .05	α = .025	α = .01	α = .005	α = 001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.041	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
60	1.296	1.671	2.000	2.390	2.660	3.232
120	1.289	1.658	1.980	2.358	2.617	3.160
240	1.285	1.651	1.970	2.342	2.596	3.125
inf.	1.282	1.645	1.960	2.326	2.576	3.090

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: Giá trị tới hạn fisher

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
1	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32
	.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
	.05	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
	.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963 3	968.6
	.01	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056
2	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	.025	38.51	39.00	39.17	3925	39.30	39.33	39.36	39.37	39.39	39.40
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4
	.001	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4	999.4
3	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
	.01	34.12	30.82	29.46	28.71	28.24	27.91	2767	27.49	27.35	27.23
	.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43 69
	.001	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129 2
4	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05
5	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92
6	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77
	10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

df_1											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.83	9.85	.25	1
60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.19	63.33	.10	
243.9	245.9	248.0	249 1	250.1	251.1	252.2	253.3	253.8	254.3	.05	
976.7	984.9	993.1	997.2	1001	1006	1010	1014	1016	1018	.025	
6106	6157	6209	6235	6261	6287	6313	6339	6353	6366	.01	
3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.47	3.48	.25	2
9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49	9.49	.10	
19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.49	19.50	.05	
39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.49	39.50	.025	
99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50	99.50	.01	
199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5	199.5	.005	
999.4	999.4	9994	999.5	999.5	999.5	999.5	999.5	999.5	999.5	.001	
2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47	2.47	.25	3
5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	.5.14	3.13	.10	
8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.54	8.53	.05	
14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.92	13.90	.025	
27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.17	26.13	.01	
43.39	4308	42.78	42.62	42.47	42.31	42.15	41.99	41.91	41.83	.005	
128.3	1274	126.4	125.9	123.4	125.0	124.5	124.0	123.7	123.5	.001	
2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	.25	4
3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.77	3.76	.10	
5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.64	5.63	.05	
8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.28	8.26	.025	
14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.51	13.46	.01	
20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.40	19.32	.005	
47.41	46.76	46.10	45.77	45.43	45.09	44.75	44.40	44.23	44.05	.001	
1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87	1.87	.25	5
3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11	3.10	.10	
4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.38	4.36	.05	
6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.04	6.02	.025	
9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.07	9.02	.01	
13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.21	12.14	.005	
26.42	25.91	25.39	25.13	24.87	24.60	24.33	24.06	23.92	23.79	.001	
1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74	1.74	.25	6
2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.73	2.72	.10	
4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.69	3.67	.05	
5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.88	4.85	.025	
7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.92	6.88	.01	
10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.94	8.88	.005	
17.99	17.56	17.12	16.90	16.67	16.44	16.21	15.98	15.86	15.75	.001	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

df_2	α	df_1 1	2	3	4	5	6	7	8	9	10
7	.25	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69
	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08
8	.25	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63
	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54
9	.25	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59
	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89
10	.25	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55
	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75
11	.25	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52
	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	.005	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92
12	.25	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50
	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

df_1											df_2
12	15	20	24	30	40	60	120	240	inf.	α	
1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65	1.65	.25	7
2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.48	2.47	.10	
3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.25	3.23	.05	
4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.17	4.14	.025	
6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.69	5.65	.01	
8.18	7.97	7.75	7.64	7.53	7.42	7.31	7.19	7.13	7.08	.005	
13.71	13.32	12.93	12.73	12.53	12.33	12.12	11.91	11.80	11.70	.001	
1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58	1.58	.25	8
2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.30	2.29	.10	
3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.95	2.93	.05	
4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.70	3.67	.025	
5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.90	4.86	.01	
7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	6.01	5.95	.005	
11.19	10.84	10.48	10.30	10.11	9.92	9.73	9.53	9.43	9.33	.001	
1.58	1.57	1.56	1.56	1.55	1.54	1.64	1.53	1 53	1.53	.25	9
2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.17	2.16	.10	
3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.73	2.71	.05	
3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.36	3.33	.025	
5.11	4.96	4.81	4.73	4.65	4.57	4.48	4:40	4.35	4.31	.01	
6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.24	5.19	.005	
9.57	9.24	8.90	8.72	8.55	8.37	8.19	8.00	7.91	7.81	.001	
1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.49	1 48	.25	10
2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.07	2.06	.10	
2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.56	2.54	.05	
3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.11	3.08	.025	
4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.95	3.91	.01	
5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.69	4.64	.005	
8.45	8.13	7.80	7.64	7.47	7.30	7.12	6.94	6.85	6.76	.001	
1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45	1.45	.25	11
2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.99	1.97	.10	
2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.43	2.40	.05	
3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.91	2.88	.025	
4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.65	3.60	.01	
5.24	5.05	4.86	4.76	4.65	4.55	4.45	4.34	4.28	4.23	.005	
7.63	7.32	7.01	6.85	6.68	6.52	6.35	6.18	6.09	6.00	.001	
1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.43	1.42	.25	12
2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.92	1.90	.10	
2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.32	2.30	.05	
3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.76	2.72	.025	
4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.41	3.36	.01	
4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.96	3.90	.005	
7.00	6.71	6.40	6.25	6.09	5.93	5.76	5.59	5.51	5.42	.001	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: (*tiếp theo*)

df_2	α	df_1 1	2	3	4	5	6	7	8	9	10
13	.25	1.45	1.55	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1 48
	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
	.005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80
14	.25	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46
	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	.025	6 30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
	.005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40
15	.25	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45
	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08
16	.25	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44
	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	.005	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81
17	.25	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43
	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	.005	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58
18	.25	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42
	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	.025	5.98	4.56	395	3.61	3.38	3.22	3.10	3.01	2.93	2.87
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	.005	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

df_1											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40	1.40	.25	13
2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.86	1.85	.10	
2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.23	2.21	.05	
3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.63	2.60	.025	
3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.21	3.17	.01	
4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.70	3.65	.005	
6.52	6.23	5.93	5.78	5.63	5.47	5.30	5.14	5.05	4.97	.001	
1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38	1.38	.25	14
2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.81	1.80	.10	
2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.15	2.13	.05	
3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.52	2.49	.025	
3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.05	3.00	.01	
4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.49	3.44	.005	
6.13	5.85	5.56	5.41	5-25	5.10	4.94	4.77	4.69	4.60	.001	
1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36	1.36	.25	15
2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.77	1.76	.10	
2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.09	2.07	.05	
2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.43	2.40	.025	
3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.91	2.87	.01	
4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.32	3.26	.005	
5.81	5.54	5.25	5.10	4.95	4.80	4.64	4.47	4.39	4.31	.001	
1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.31	1.35	1.34	.25	16
1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.73	1.72	.10	
2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.03	2.01	.05	
2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.35	2.32	.025	
3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.80	2.75	.01	
4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.17	3.11	.005	
5.55	5.27	4.99	4.85	4.70	4.54	4.39	4.23	4.14	4.06	.001	
1.41	1.40	1.39	1 38	1.37	1.36	1.35	1.34	1.33	1.33	.25	17
1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.70	1.69	.10	
2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.99	1.96	.05	
2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.28	2.25	.025	
3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.70	2.65	.01	
3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	3.04	2.98	.005	
5.32	5.05	4.78	4.13	4.48	4.33	4.18	4.02	3.93	3.85	.001	
1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.32	.25	18
1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.67	1.66	.10	
2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.94	1.92	.05	
2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.22	2.19	.025	
3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.61	2.57	.01	
3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.93	2.87	.005	
5.13	4.87	4.59	4.45	4.30	4.15	4.00	3.84	3.75	3.67	.001	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

df_2	α	df_1 1	2	3	4	5	6	7	8	9	10
19	.25	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41
	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22
20	.25	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40
	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08
21	.25	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39
	.10	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
	.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
	.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95
22	.25	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39
	.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	.005	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83
23	.25	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38
	.10	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
	.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
	.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73
24	.25	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38
	.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

df_1											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.30	.25	19
1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.65	1.63	.10	
2 31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.90	1.88	.05	
2.72	2.62	2.51	2.45	2. 39	2.33	2.27	2.20	2.17	2.13	.025	
3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.54	2.49	.01	
3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.83	2.78	.005	
4.97	4.70	4.43	4.29	4.14	3.99	3.84	3.68	3.60	3.31	.001	
1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.30	1.29	.25	20
1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.63	1.61	.10	
2.28	2.20	2.12	2.08	2.04	1 99	1.95	1.90	1.87	1.04	.03	
2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.12	2.09	.025	
3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.47	2.42	.01	
3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.75	2.69	.005	
4.82	4.56	4.29	4.15	4.00	3.86	3.70	3.54	3.46	3.38	.001	
1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.29	1.28	.25	21
1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.60	1.59	.10	
2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.84	1.81	.05	
2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.08	2.04	.025	
3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.41	2.36	.01	
3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.67	2.61	.005	
4.70	4.44	4.17	4.03	3.88	3.74	3.58	3.42	3.34	3.26	.001	
1.37	1.36	1.34	1.33	1 32	1.31	1.30	1.29	1.28	1.28	.25	22
1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.59	1.57	.10	
2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.81	1.78	.05	
2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.04	2.00	.025	
3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.35	2.31	.01	
3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.60	2.55	.005	
4.58	4.33	4.06	3.92	3.78	3.63	3.48	3.32	3.23	3.15	.001	
1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.28	1.27	.25	23
1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.57	1.35	.10	
2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.79	1.76	.05	
2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	2.01	1.97	.025	
3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.31	2.26	.01	
3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.54	2.48	.005	
4.48	4.23	3.96	3.82	3.68	3.53	3.38	3.22	3 14	3.05	.001	
1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.27	1.26	.25	24
1.83	1.78	1.73	1.70	1.07	1.64	1.61	1.57	1.35	1.53	.10	
2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.76	1.76	1.73	.05	
2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.97	1.94	.025	
3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.26	2.21	.01	
3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.49	2.43	.005	
4.39	4.14	3.87	3.74	3 59	3.45	3.29	3.14	3.05	2.97	.001	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

df_2	α	df_1 1	2	3	4	5	6	7	8	9	10
25	.25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37
	.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
	.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
	.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
	.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56
26	.25	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37
	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
	.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48
27	.25	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36
	.10	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
	.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
	.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
	.005	9 34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41
28	.25	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36
	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
	.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41
	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35
29	.25	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35
	.10	2.89	2.50	2.28	2.15	206	1.99	1.93	1.89	1.86	1.83
	.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
	.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
	.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38
	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29
30	.25	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35
	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

df_1											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.26	1.25	.25	25
1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.54	1.52	.10	
2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.74	1.71	.05	
2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.94	1.91	.025	
2.99	2.85	2.70	2 62	2.54	2.45	2.36	2.27	2.22	2.17	.01	
3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.44	2.38	.005	
4.31	4.06	3.79	3.66	3.52	3.37	3.22	3.06	2.98	2.89	.001	
1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26	1.26	1.25	.25	26
1.81	1.76	1.71	1 68	1.65	1.61	1.58	1.54	1.52	1.50	10	
2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.72	1.69	.05	
2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.92	1.88	.025	
2.96	2.81	2.66	2.58	2 50	2.42	2.33	2.23	2.18	2.13	.01	
3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.39	2.33	.005	
4.24	3.99	3.72	3.59	3.44	3.30	3.15	2.99	2.90	2.82	.001	
1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.25	1.24	.25	27
1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.51	1.49	.10	
2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.70	1.67	.05	
2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.89	1.85	.025	
2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.15	2.10	.01	
3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.35	2.29	.005	
4.17	3.92	3.66	3.52	3.38	3.23	3.08	2.92	2.84	2.75	001	
1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24	1.24	.25	28
1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.50	1.48	.10	
2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.68	1.65	.05	
2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.87	1.83	.025	
2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.12	2.06	.01	
3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.31	2.25	.005	
4.11	3.86	3.60	3.41	3.32	3.18	3.02	2.86	2.78	2.69	.001	
1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.24	1.23	.25	29
1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.49	1.47	.10	
2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.67	1.64	.05	
2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.85	1.81	.025	
2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.09	2.03	.01	
3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.27	2.21	.005	
4.05	3.80	3.54	3.46	3.27	3.12	2.97	2.81	2.73	2.64	.001	
1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23	1.23	.25	30
1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.48	1.46	.10	
2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.65	1.62	.05	
2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.83	1.79	.025	
2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.06	2.01	.01	
3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.24	2.18	.005	
4.00	3.75	3.49	3.36	3.22	3.07	2.92	2.76	2.68	2.59	.001	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
40	.25	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2,12	2.08
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2-89	2.80
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87
60	.25	1.35	1.42	1.41	1.38	1.37	1.35	1.33.	1.32	1.31	1.30
	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
	01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54
90	.25	1.34	1.41	1.39	1.37	1.35	1.33	1.32	1.31	1.30	1.29
	.10	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
	.05	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	.025	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19
	.01	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
	.005	8.28	5.62	4,57	3.99	3.62	3.35	3.15	3.00	2.87	2.77
	.001	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
	.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
	.025	5.15	3.80	3.23.	2.89	2.67	2.52	2.39	2.30	2.22	2.16
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71
	.001	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24
240	.25	1.33	1.39	1.38	1.36	1.34	1.32	1.30	1.29	1.27	1.27
	.10	2.73	2.32	2.10	1.97	1.87	1.80	1.74	1.70	1.65	1.63
	.05	3.88	3.03	2.64	2.41	2.25	2.14	2.04	1.98	1.92	1.87
	.025	5.09	3.75	3.17	2.84	2.62	2.46	2.34	2.25	2.17	2.10
	.01	6.74	4.69	3.86	3.40	3.09	2.88	2.71	2.59	2.48	2.40
	.005	8.03	5.42	4.38	3.82	3.45	3.19	2.99	2.84	2.71	2.61
	.001	11.10	7.11	5.60	4.78	4.25	3.89	3.62	3.41	3.24	3.09
inf	.25	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52
	.001	10.83	6,91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 9: *(tiếp theo)*

df_1											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.20	1.19	.25	40
1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.40	1.38	.10	
2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.54	1.51	.05	
2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.68	1.64	.025	
2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.86	1.80	.01	
2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	2.00	1.93	.005	
3.64	3.40	3.14	3.01	2.87	2.73	2.57	2.41	2.32	2.23	.001	
1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.16	1.15	.25	60
1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.32	1.29	.10	
1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.43	1.39	.05	
2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.53	1.48	.025	
2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.67	1.60	.01	
2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.76	1.69	.005	
3.32	3.08	2.83	2.69	2.55	2.41	2.25	2.08	1.99	1.89	.001	
1.27	1.25	1.23	1.22	1.20	1.19	1.17	1.15	1.13	1.12	.25	90
1.62	1.56	1.50	1.47	1.43	1.39	1.35	1.29	1.26	1.23	.10	
1.86	1.78	1.69	1.64	1.59	1.53	1.46	1.39	1.35	1.30	.05	
2.09	1.98	1.86	1.80	1.73	1.66	1.58	1 48	1.43	1.37	.025	
2.39	2.24	2.09	2.00	1.92	1.82	1.72	1 60	1.53	1.46	.01	
2.61	2.44	2.25	2.15	2.05	1.94	1.82	1.68	1.61	1.52	.005	
3.11	2.88	2.63	2.50	2.36	2.21	2.05	1.87	1.77	1.66	.001	
1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.12	1.10	.25	120
1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.23	1.19	.10	
1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.31	1.25	.05	
2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.38	1.31	.025	
2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.46	1.38	.01	
2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.52	1.43	.005	
3.02	2.78	2.53	2.40	2.26	2.11	1.95	1.77	1.66	1.54	.001	
1.25	1.23	1.21	1.19	1.18	1.16	1.14	1.11	1.09	1.07	.25	240
1.57	1.52	1.45	1.42	1.38	1.33	1.28	1.22	1.18	1.13	.10	
1.79	1.71	1.61	1.56	1.51	1.44	1.37	1.29	1.24	1.17	.05	
2.00	1.89	1.77	1.70	1.63	1.55	1.46	1.35	1.29	1.21	.025	
2.26	2.11	1.96	1.87	1.78	1.68	1.57	1.43	1.35	1.25	.01	
2.45	2.28	2.09	1.99	1.89	1.77	1.64	1.49	1.40	1.28	.005	
2.88	2.65	2.40	2.26	2.12	1.97	1.80	1.61	1.49	1.35	.001	
1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.06	1.00	.25	inf
1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1 12	1.00	.10	
1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.15	1.00	.05	
1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.19	1.00	.025	
2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.22	1.00	.01	
2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.25	1.00	.005	
2.74	2.51	2.27	2.13	1.99	1.84	1.66	1.45	1.31	1.00	.001	

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 10: Bảng số ngẫu nhiên 4 chữ số

1559	9068	9290	8303	8508	8954	1051	6677	6415	0342
5550	6245	7313	0117	7652	5069	6354	7668	1096	5780
4735	6214	8037	1385	1882	O828	2957	0530	9210	0177
5333	1313	3063	1134	8676	6241	9960	5304	1582	6198
8495	2956	1121	8484	2920	7934	0670	5263	0968	0069
1947	3353	1197	7363	9003	9313	3434	4261	0066	2714
4785	6325	1868	5020	9100	0823	7379	7391	1250	5501
9972	9163	5833	0100	5758	3696	6496	6297	5653	7782
0472	4629	2007	4464	3312	8728	1193	2497	4219	5339
4727	6994	1175	5622	2341	8562	5192	1471	7206	2027
3658	3226	5981	9025	1080	1437	6721	7331	0792	5383
6906	9758	0244	0259	4609	1269	5957	7556	1975	7898
3793	6916	0132	8873	8987	4975	4814	2098	6683	0901
3376	5966	1614	4025	0721	1537	6695	6090	8083	5450
6126	0224	7169	3596	1593	5097	7286	2686	1796	1150
0466	7566	1320	8777	8470	5448	9575	4669	1402	3905
9908	9832	8185	8835	0384	3699	1272	1181	8627	1968
7594	3636	1224	6808	1184	3404	6752	4391	2016	6167
5715	9301	5847	3524	0077	6674	8061	5438	6508	9673
7932	4739	4567	6797	4540	6488	3639	9777	1621	7244
6311	2025	5250	6099	6718	7539	9681	3204	9637	1091
0476	1624	3470	1600	0675	3261	7749	4195	2660	2150
5317	3903	6098	9438	3482	5505	5167	9993	8191	8488
7474	8876	1918	9828	2061	6664	0391	9170	2776	4025
7460	6800	1967	2758	0737	6880	1500	5763	2061	9373
1002	4494	9972	3877	6104	4006	0477	0669	8557	0513
5449	6891	9047	6297	1075	7762	8091	7153	8881	3367
9453	0809	7151	9982	0411	1120	6129	5090	2053	7570
0471	2725	7588	6573	0546	0110	6132	1224	3124	6563
5469	2668	1996	2249	3857	6637	8010	1701	3141	6147
2782	9603	1877	4159	9809	2570	4544	0544	2660	6737
3129	7217	5020	3788	0853	9465	2186	3945	1696	2286
7092	9885	3714	8557	7804	9524	6226	7774	6674	2775
9566	0501	8352	1062	0634	2401	0379	1697	7153	6208
5863	7000	1714	9276	7218	6922	1032	4838	1954	1680
5881	9151	2321	3147	6755	2510	5759	6947	7102	0097
6416	9939	9569	0439	1705	4680	9881	7071	9596	8758
9568	3012	6316	9065	0710	2158	1639	9149	4848	8634
0452	9538	5730	1893	1186	9245	6558	9562	8534	9321
8762	5920	8989	4777	2169	7073	7082	9495	1594	8600
0194	0270	7601	0342	3897	4133	7650	9228	5558	3597
3306	5478	2797	1605	4996	0023	9780	9429	3937	7573
7198	3079	2171	6972	0926	6599	9328	0597	5948	5753

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 10: *(tiếp theo)* Bảng số ngẫu nhiên 4 chữ số

8350	4846	1309	0612	4584	4968	4642	4430	9481	9048
7449	4279	4224	1018	2496	2091	9750	6086	1955	9860
6126	5399	0852	5491	6557	4946	9918	1541	7894	1843
1851	7940	9908	3660	1536	8011	4314	7269	7047	0382
7698	4218	2726	5130	3132	1722	8592	9662	4795	7718
0810	0118	4979	0458	1059	5739	7919	4557	0245	4861
6647	7149	1409	6809	3313	0082	9024	7477	7320	5822
3867	7111	5549	9439	3427	9793	3071	6651	4267	8099
1172	7278	7527	2492	6211	9457	5120	4903	1023	5745
6701	1668	5067	0413	7961	7825	9261	8572	0634	1140
8244	0620	8736	2649	1429	6253	4181	8120	6500	8127
8009	4031	7884	2215	2382	1931	1252	8088	2490	9122
1947	8315	9755	7187	4074	4743	6669	6060	2319	0635
9562	4821	8050	0106	2782	4665	9436	4973	4879	8900
0729	9026	9631	8096	8906	5713	3212	8854	3435	4205
6904	2569	3251	0079	8838	8738	8503	6333	0952	1641

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 11: Bảng giá trị kiểm định tổng hạng wilcoxon

a. $\alpha = .025$ một phía $\alpha = .05$ hai phía

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

b. $\alpha = .05$ một phía $\alpha = .01$ hai phía

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 12: Bảng giá trị kiểm định tổng hạng theo dấu wilcoxon

Một phía	Hai phía	n = 5	n = 6	n = 7	n = 8	n = 9
p = .1	p = .2	2	3	5	8	10
p = .05	p = .1	0	2	3	5	8
p = .025	p = .05		0	2	3	5
p = .01	p = .02			0	1	3
p = .005	p = .01				0	1
p = .0025	p = .005					0
p = .001	p = .002					

Một phía	Hai phía	n = 15	n = 16	n = 17	n = 18	n = 19
p = .1	p = .2	36	42	48	55	62
p = .05	p = .1	30	35	41	47	53
p = .025	p = .05	25	29	34	40	46
p = .01	p = .02	19	23	27	32	37
p = .005	p = .01	15	19	23	27	32
p = .0025	p = .005	12	15	19	23	27
p = .001	p = .002	8	11	14	18	21

Một phía	Hai phía	n = 25	n = 26	n = 27	n = 28	n = 29
p = .1	p = .2	113	124	134	145	157
p = .05	p = .1	100	110	119	130	140
p = .025	p = .05	89	98	107	116	126
p = .01	p = .02	76	84	92	101	110
p = .005	p = .01	68	75	83	91	100
p = .0025	p = .005	60	67	74	82	90
p = .001	p = .002	51	58	64	71	79

Một phía	Hai phía	n = 35	n = 36	n = 37	n = 38	n = 39
p = .1	p = .2	235	250	265	281	297
p = .05	p = .1	213	227	241	256	271
p = .025	p = .05	195	208	221	235	249
p = .01	p = .02	173	185	198	211	224
p = .005	p = .01	159	171	182	194	207
p = .0025	p = .005	146	157	168	180	192
p = .001	p = .002	131	141	151	162	173

Một phía	Hai phía	n = 45	n = 46	n = 47	n = 48	n = 49
p = .1	p = .2	402	422	441	462	482
p = .05	p = .1	371	389	407	426	446
p = .025	p = .05	343	361	378	396	415
p = .01	p = .02	312	328	345	362	379
p = .005	p = .01	291	307	322	339	355
p = .0025	p = .005	272	287	302	318	334
p = .001	p = .002	249	263	277	292	307

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 12: *(Tiếp theo)*

Một phía	Hai phía	n = 10	n = 11	n = 12	n = 13	n = 14
p = .1	p = .2	14	17	21	26	31
p = .05	p = .1	10	13	17	21	25
p = .025	p = .05	8	10	13	17	21
p = .01	p = .02	5	7	9	12	15
p = .005	p = .01	3	5	7	9	12
p = .0025	p = .005	1	3	5	7	9
p = .001	p = .002	0	1	2	4	6

Một phía	Hai phía	n = 20	n = 21	n = 22	n = 23	n = 24
p = .1	p = .2	69	77	86	94	104
p = .05	p = .1	60	67	75	83	91
p = .025	p = .05	52	58	65	73	81
p = .01	p = .02	43	49	55	62	69
p = .005	p = .01	37	42	48	54	61
p = .0025	p = .005	32	37	42	48	54
p = .001	p = .002	26	30	35	40	45

Một phía	Hai phía	n = 30	n = 31	n = 32	n = 33	n = 34
p = .1	p = .2	169	181	194	207	221
p = .05	p = .1	151	163	175	187	200
p = .025	p = .05	137	147	159	170	182
p = .01	p = .02	120	130	140	151	162
p = .005	p = .01	109	118	128	138	148
p = .0025	p = .005	98	107	116	126	136
p = .001	p = .002	86	94	103	112	121

Một phía	Hai phía	n = 40	n = 41	n = 42	n = 43	n = 44
p = .1	p = .2	313	330	348	365	384
p = .05	p = .1	286	302	319	336	353
p = .025	p = .05	264	279	294	310	327
p = .01	p = .02	238	252	266	281	296
p = .005	p = .01	220	233	247	261	276
p = .0025	p = .005	204	217	230	244	258
p = .001	p = .002	185	197	209	222	235

Một phía	Hai phía	n = 50	n = 51	n = 52	n = 53	n = 54
p = .1	p = .2	503	525	547	569	592
p = .05	p = .1	466	486	507	529	514
p = .025	p = .05	434	453	473	494	514
p = .01	p = .02	397	416	434	454	473
p = .005	p = .01	373	390	408	427	445
p = .0025	p = .005	350	367	384	402	420
p = .001	p = .002	323	339	355	372	389

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Phụ lục 13: Bảng giá trị tới hạn của phân phối cochran

k - số bậc tự do, l - số lượng mẫu; mức ý nghĩa $\alpha = 0,05$

l \ k	1	2	3	4	5	6	7	8	9	10	16	36	144	+∞
2	0,985	0,9750	0,9392	0,9057	0,8772	0,8534	0,8332	0,8159	0,8010	0,7880	0,7341	0,6602	0,5813	0,5000
3	9669	8709	7977	7457	7071	6771	6530	6333	6167	6025	5466	4748	4031	3333
4	9065	7679	6841	6287	5895	5598	5365	5175	5017	4884	4366	3720	3093	2500
5	0,8412	0,6338	0,5981	0,5440	0,5063	0,4783	0,4564	0,4387	0,4241	0,4118	0,3645	0,3066	0,2013	0,2000
6	7808	6161	5321	4803	4447	4184	3980	3817	3682	3568	3135	2612	2119	1667
7	7271	5612	4800	4307	3974	3726	3535	3384	3259	3154	2756	2278	1833	1429
8	0,6798	0,5157	0,4377	0,3910	0,3595	0,3362	0,3185	0,3043	0,2926	0,2829	0,2462	0,2022	0,1616	0,1250
9	6385	4775	4027	3584	3286	3067	2901	2768	2659	2568	2226	1820	1446	1111
10	6020	4450	3733	3311	3029	2823	2666	2541	2439	2353	2032	1655	1308	1000
12	0,5410	0,3924	0,3624	0,2880	0,2624	0,2439	0,2299	0,2187	0,2098	0,2020	0,1737	0,1403	0,1100	0,0833
15	4709	3346	2758	2419	2195	2034	1911	1815	1736	1671	1429	1144	0889	0667
20	3894	2705	2205	1921	1735	1602	1501	1422	1357	1303	1108	0879	0675	0500
24	0,3434	0,2354	0,1907	0,1656	0,1493	0,1374	0,1286	0,1216	0,1160	0,1113	0,0942	0,0743	0,0567	0,0417
30	2929	1980	1593	1377	1237	1137	1061	1002	0958	0921	0771	0604	0457	0333
40	2370	1576	1259	1082	0968	0887	0827	0780	0745	0713	0595	0462	0347	0250
60	0,1737	0,1131	0,0895	0,0765	0,0682	0,0623	0,0583	0,0552	0,0520	0,0497	0,0411	0,0316	0,0234	0167
120	0998	0632	0495	0419	0371	0337	0312	0292	0279	0266	0218	0165	0120	0083
+∞	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ
THƯ VIỆN HUBT

Phụ lục 13: Bảng giá trị tới hạn của phân phối cochran

k - số bậc tự do; l - số lượng mẫu; mức ý nghĩa $\alpha = 0,01$

l \ k	1	2	3	4	5	6	7	8	9	10	16	36	144	+∞
2	0,9999	0,9950	0,9794	0,9586	0,9373	0,9172	0,8988	0,8823	0,8674	0,8539	0,7949	0,7067	0,6062	0,5000
3	9933	9423	8831	8335	7938	7606	7335	7107	6912	6743	6059	5153	4230	3333
4	9676	8643	7814	7212	6761	6410	6129	5897	5702	5536	4884	4057	3251	2500
5	9279	0,7885	6957	0,6329	0,5875	0,5531	0,5259	0,5037	0,4854	0,4697	0,4094	0,3351	0,2644	0,2000
6	8828	7218	6258	5635	5195	4866	4608	4401	4229	4084	3529	2858	2229	1667
7	8376	6644	5685	5080	4659	4347	4105	3911	3751	3616	3105	2494	1929	1429
8	0,7945	0,6152	0,5209	0,4627	0,4226	0,3932	0,3704	0,3522	0,3373	0,3248	0,2779	0,2214	0,1700	0,1250
9	7544	5727	4810	4251	3870	3592	3378	3207	3007	2950	2514	1992	1521	1111
10	7175	5358	4469	3934	3572	3308	3106	2945	2810	2704	2297	1811	1376	1000
12	0,6528	0,4751	0,3919	0,3428	0,3099	0,2861	0,2680	0,2535	0,2413	0,2320	0,1961	0,1535	0,1157	0,0833
15	5747	4069	3317	2882	2593	2386	2228	2104	2002	1918	1612	1251	0934	0667
20	4799	3297	2654	2288	2048	1877	1748	1646	1567	1501	1248	0960	0709	0500
24	0,4247	0,2871	0,2295	0,1970	0,1759	0,1608	0,1495	0,1406	0,1338	0,1263	0,1060	0,0810	0,0595	0,0417
30	3632	2412	1913	1635	1454	1327	1232	1157	1100	1054	0867	0658	0480	0333
40	2940	1915	1508	1281	1135	1033	0957	0898	0853	0816	0668	0503	0363	0250
60	2151	0,1371	0,1069	0,0902	0,0796	0,0722	0,0668	0,0625	0,0594	0,0567	0,0461	0,0344	0,0245	0,0167
120	1225	0759	0585	0489	0429	0387	0357	0334	0316	0302	0242	0178	0125	0083
+∞	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ
THƯ VIỆN HUBT

Phụ lục 14: Bảng giá trị tới hạn của phân phối kolmogorov

α	0.5	0.4	0.3	0.2	0,1	0,05	0,02	0,01	0,001
λ_α	0.828	0,895	0,974	1.073	1.224	1,358	1.520	1.627	1.950

Phụ lục 15: Bảng kiểm định Lilliefors

n	*a*				
	0,2	0,15	0,1	0,05	0,01
4	0,300	0,319	0,352	0,381	0,417
5	0,285	0,299	0,315	0,337	0,405
6	0,265	0,277	0,294	0,319	0,364
7	0,247	0,258	0,276	0,300	0,348
8	0,233	0,244	0,261	0,285	0,331
9	0,223	0,233	0,249	0,271	0,311
10	0,215	0,224	0,239	0,258	0,294
11	0,206	0,217	0,230	0,249	0,284
12	1,199	0,212	0,223	0,242	0,275
13	0,190	0,202	0,214	0,234	0,268
14	0,183	0,194	0,207	0,227	0,261
15	0,177	0,187	0,201	0,220	0,257
16	0,173	0,182	0,195	0,213	0,250
17	0,169	0,177	0,189	0,206	0,245
18	0,166	0,173	0,184	0,200	0,239
19	0,163	0,169	0,179	0,195	0,235
20	0,160	0,166	0,174	0,190	0,231
25	0,142	0,147	0,158	0,173	0,200
30	0,131	0,136	0,144	0,161	0,187
>30	0,736	0,768	0,805	0,886	1,031
	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$

THƯ VIỆN HUBT
TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

Giáo trình
LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ TOÁN

NHÀ XUẤT BẢN ĐẠI HỌC KINH TẾ QUỐC DÂN

Địa chỉ: 207 Đường Giải Phóng, Quận Hai Bà Trưng, Hà Nội

Website: http//nxb.neu.edu.vn E-mail: nxb@neu.edu.vn

Điện thoại/Fax: (024) 36282486

Chịu trách nhiệm xuất bản: **TS. NGUYỄN ANH TÚ**, *Giám đốc Nhà xuất bản*

Chịu trách nhiệm nội dung: **GS.TS. NGUYỄN THÀNH ĐỘ**, *Tổng biên tập*

Biên tập: TRỊNH THỊ QUYÊN – ĐỖ THỊ NGỌC LAN

Chế bản: NGUYỄN LAN

Thiết kế bìa: TRẦN MAI HOA

Sửa bản in và đọc sách mẫu: TRỊNH THỊ QUYÊN – ĐỖ THỊ NGỌC LAN

In 3.500 bản, khổ 16 x 24cm tại Xưởng in Trung tâm Dịch vụ Hỗ trợ Đào tạo Trường Đại học Kinh tế Quốc dân.

Địa chỉ: Số 207 đường Giải Phóng, phường Đồng Tâm, quận Hai Bà Trưng, Hà Nội.

Mã số ĐKXB: 1279-2018/CXBIPH/2-208/ĐHKTQD cấp ngày 17/04/2018 và ISBN: 978-604-946-421-8.

Số quyết định xuất bản: 264/QĐ-NXBĐHKTQD - Ngày 18 tháng 5 năm 2018

In xong và nộp lưu chiểu quý II năm 2018

THƯ VIỆN HUBT

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ

TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN *Kết nối tri thức - Vững bước tương lai*

www.neu.edu.vn ❖ www.neu.edu.vn ❖ www.neu.edu.vn ❖ www.neu.edu.vn ❖ www.neu.edu.vn

Giáo trình

LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ TOÁN

(Tái bản lần thứ 6)

MỜI ĐỘC GIẢ TÌM ĐỌC NHỮNG CUỐN SÁCH HAY
TRONG BỘ SÁCH GIÁO TRÌNH
CỦA TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN



Tổng phát hành tại:

NHÀ SÁCH KINH TẾ QUỐC DÂN
TRUNG TÂM DỊCH VỤ HỖ TRỢ ĐÀO TẠO

207 Đường Giải Phóng, Quận Hai Bà Trưng, Hà Nội
Tel: (84-4) 36280 280/ 5212 - Fax: (84-4) 3628 3799
Email: nhasachkinhtequocdan@gmail.com
Website: www.neu.edu.vn/tdichvu.neu.edu.vn

9 786049 464218

Giá: 145.000VNĐ

THƯ VIỆN HUBT

TÀI LIỆU PHỤC VỤ THAM KHẢO NỘI BỘ