

thống kê ứng dụng

trong kinh tế - xã hội



cuuduong thanhcong.com

μ
 σ

hoàng trọng
chunguyễnmộngngọc

HOÀNG TRỌNG – CHU NGUYỄN MỘNG NGỌC

THỐNG KÊ ỨNG DỤNG

trong Kinh tế - Xã hội

cuu duong than cong. com
TÁI BẢN LẦN THỨ NHẤT

NHÀ XUẤT BẢN THỐNG KÊ
NĂM 2008

LỜI NÓI ĐẦU

Thống kê là công cụ không thể thiếu được trong hoạt động nghiên cứu và công tác thực tiễn. Thống kê đã trở thành một môn học cơ bản hay cơ sở trong hầu hết các ngành đào tạo. Trong các chuyên ngành thuộc khối kinh tế - xã hội, đã có môn xác suất thống kê và lý thuyết thống kê. Với định hướng cải tiến chương trình và nội dung gắn liền với thực tiễn, nhiều trường đại học đã bắt đầu giảng dạy môn lý thuyết thống kê theo hướng ứng dụng trong lĩnh vực kinh tế - xã hội và có thực hành trên máy vi tính. Một vài trường đã chuyển sang môn học Thống kê ứng dụng.

Trong bối cảnh đào tạo đại học đang cần, và có những chuyển biến mạnh mẽ về công tác đào tạo, trong đó thời gian lên lớp được giới hạn và sinh viên được khuyến khích tự tham khảo tài liệu và tự học. Điều này đòi hỏi cần có những tài liệu được biên soạn kỹ lưỡng và chi tiết để sinh viên có thể tự nghiên cứu.

Bên cạnh đó, trong xu hướng hội nhập với khu vực và thế giới, giáo dục đại học Việt Nam đang từng bước thay đổi, việc giảng dạy và học tập thống kê cũng không nằm ngoài quy đạo đó. Nhu cầu về một tài liệu giảng dạy và học tập môn thống kê ứng dụng, vừa phù hợp với sinh viên Việt Nam, vừa nhất quán với các môn học thống kê ứng dụng chuẩn mực trên thế giới là rất cần thiết.

Ngoài ra, việc đi sâu vào các môn về phương pháp nghiên cứu, phương pháp phân tích dữ liệu của sinh viên các chuyên ngành khối kinh tế - xã hội, và việc nghiên cứu và tự học của những người đang làm công tác thực tế đang đòi hỏi một quyển sách tham khảo về thống kê ứng dụng được trình bày chặt chẽ và chi tiết.

Hơn nữa, còn nhiều sinh viên coi việc học môn thống kê nói chung và thống kê ứng dụng nói riêng là một việc khó khăn hay gánh nặng. Việc giảng dạy và học tập môn thống kê hiện nay ít đạt hiệu quả hay còn hời hợt xét theo ý nghĩa của việc học thống kê có đem lại niềm vui và sự hiểu biết, có là cơ sở tốt cho người học tiếp cận các môn học khác về sau, cũng như vận dụng hiệu quả trong công việc sau này của người học hay không. Điều này do khá nhiều nguyên nhân. Ở góc độ người biên soạn sách, chúng tôi nghĩ một phần là do tài liệu đáp ứng tốt nhu cầu của người đọc vẫn còn thiếu.

Để đáp ứng các nhu cầu trên, chúng tôi thực hiện biên soạn quyển sách Thống kê ứng dụng trong kinh tế xã hội. Tài liệu này được xây dựng với định hướng ứng dụng trong kinh tế và xã hội với các ví dụ gần gũi và thực tế. Quyển sách được biên soạn theo tinh thần diễn giải chi tiết để người đọc có thể tự mình nắm bắt cặn kẽ phần lớn các vấn đề được trình bày.

Với kinh nghiệm giảng dạy được tích lũy qua nhiều năm, tham gia thực hiện các đề tài nghiên cứu trong lĩnh vực kinh tế - xã hội, cộng với các nguồn tài liệu phong phú, chúng tôi hy vọng quyển sách đáp ứng được nhu cầu học tập của các sinh viên và nhu cầu tham khảo của tất cả những ai có quan tâm đến việc ứng dụng thống kê trong nghiên cứu kinh tế và xã hội.

Chúng tôi hy vọng với quyển sách này bạn đọc không những chỉ biết mà còn hiểu được thống kê. Qua đó có thể cảm thấy lợi ích của thống kê như là một công cụ hữu hiệu cho sinh viên, nhà quản lý, nhà nghiên cứu, người điều hành trong lĩnh vực kinh tế - xã hội. Chúng tôi cũng hy vọng bạn đọc có những giờ phút lý thú cùng với quyển sách này!

Trong lần tái bản này, chúng tôi đã chỉnh sửa và bổ sung một số nội dung. Tuy nhiên, chắc chắn việc biên soạn vẫn không tránh khỏi những thiếu sót. Chúng tôi mong nhận được những ý kiến trao đổi và đóng góp của bạn đọc để lần tái bản tiếp theo quyển sách được hoàn thiện hơn. Thư góp ý xin gửi về hộp thư sau:

hoangtrong@hcm.vnn.vn

chunguyenmongngoc@yahoo.com

TP HCM, tháng 9 năm 2008

Các tác giả

Hoàng Trọng

Chu Nguyễn Mộng Ngọc

MỤC LỤC TỔNG QUÁT

CHƯƠNG	NỘI DUNG	TRANG
1	GIỚI THIỆU MÔN HỌC	1
2	THU THẬP DỮ LIỆU	16
3	TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU BẰNG BẢNG VÀ ĐỒ THỊ	37
4	TÓM TẮT DỮ LIỆU BẰNG CÁC ĐẠI LƯỢNG SỐ	66
5	XÁC SUẤT CĂN BẢN, BIẾN NGẪU NHIÊN VÀ QUY LUẬT PHÂN PHỐI XÁC SUẤT	101
6	PHÂN PHỐI CỦA CÁC THAM SỐ MẪU	170
7	ƯỚC LƯỢNG CÁC THAM SỐ TỔNG THỂ	185
8	KIỂM ĐỊNH GIÁ THUYẾT VỀ THAM SỐ TỔNG THỂ	208
9	PHÂN TÍCH PHƯƠNG SAI	250
10	KIỂM ĐỊNH PHI THAM SỐ	283
11	HỐI QUI TUYẾN TÍNH ĐƠN BIẾN VÀ PHÂN TÍCH TƯƠNG QUAN	304
12	HỐI QUI TUYẾN TÍNH ĐA BIẾN	357
13	CHỈ SỐ	391
14	CHUỖI THỜI GIAN VÀ DỰ BÁO TRÊN CHUỖI THỜI GIAN	419
15	DỰ BÁO BẰNG PHƯƠNG PHÁP BOX-JENKINS	469
	Tài liệu tham khảo	504
	Phụ lục:	506
	Bảng tra 1: Phân phối chuẩn	507
	Bảng tra 2: Phân phối Student	508
	Bảng tra 3: Phân phối Chi bình phương	509
	Bảng tra 4: Phân phối F	511
	Bảng tra 5: Phân phối Hartley	514
	Bảng tra 6: Kiểm định dấu và hạng WILCOXON	515
	Bảng tra 7: Kiểm định tổng và hạng WILCOXON	516
	Bảng tra 8: Durbin Watson	517
	Bảng tra 9: Phân phối Tukey	519

MỤC LỤC CHI TIẾT

CHƯƠNG 1: GIỚI THIỆU MÔN HỌC

1.1 THỐNG KÊ LÀ GÌ?	1
1.1.1 Xuất phát thuật ngữ thống kê	1
1.1.2 Khái niệm Thống kê.....	2
1.1.3 Tổng quan về thống kê.....	3
1.2 CÁC PHƯƠNG PHÁP NGHIÊN CỨU THỐNG KÊ.....	4
1.3 THỐNG KÊ ỨNG DỤNG TRONG KINH TẾ VÀ XÃ HỘI.....	5
1.4 MỘT SỐ KHÁI NIỆM DÙNG TRONG THỐNG KÊ.....	7
1.4.1 Dữ liệu, thông tin và tri thức (Data, information, knowledge)	7
1.4.2 Tổng thể thống kê (Population) và đơn vị tổng thể.....	8
1.4.3 Mẫu (Sample)	9
1.4.4 Đặc điểm thống kê (Characteristic)	9
1.4.5 Chỉ tiêu thống kê	10
1.5 KHÁI QUÁT QUÁ TRÌNH NGHIÊN CỨU THỐNG KÊ	10
1.6 CÁC CẤP BẬC ĐO LƯỜNG VÀ THANG ĐO	11
1.6.1 Thang đo định danh (Nominal scale)	12
1.6.2 Thang đo thứ bậc (Ordinal scale)	12
1.6.3 Thang đo khoảng (Interval scale).....	13
1.6.4 Thang đo tỷ lệ (Ratio scale)	14

CHƯƠNG 2: THU THẬP DỮ LIỆU

2.1 XÁC ĐỊNH DỮ LIỆU CẦN THU THẬP	16
2.2 DỮ LIỆU THỦ CẤP VÀ DỮ LIỆU SƠ CẤP	17
2.2.1 Nguồn dữ liệu thứ cấp	17
2.2.2 Nguồn dữ liệu sơ cấp.....	18
2.3 CÁC PHƯƠNG PHÁP THU THẬP DỮ LIỆU SƠ CẤP	19
2.3.1 Thu thập dữ liệu sơ cấp trong nghiên cứu thực nghiệm.....	20
2.3.2 Thu thập dữ liệu sơ cấp trong nghiên cứu quan sát.....	20
2.3.2.1 Khảo sát qua điện thoại	21
2.3.2.2 Thư hỏi và những khảo sát dạng viết khác.....	21
2.3.2.3 Quan sát trực tiếp và phỏng vấn cá nhân	22
2.3.2.4 Những phương pháp thu thập dữ liệu khác	23
2.4 CÁC KỸ THUẬT LẤY MẪU	23
2.4.1 Kỹ thuật lấy mẫu xác suất (probability sampling)	24
2.4.1.1 Lấy mẫu ngẫu nhiên đơn giản (Simple random sampling)	24
2.4.1.2 Lấy mẫu hệ thống (systematic sampling)	25
2.4.1.3 Lấy mẫu cả khối/cụm (cluster sampling) và lấy mẫu nhiều giai đoạn (multi-stage sampling)	28
2.4.1.4 Lấy mẫu phân tầng (Stratified sampling).....	29
2.4.2 Kỹ thuật lấy mẫu phi xác suất (non-probability sampling).....	33
2.4.2.1 Lấy mẫu thuận tiện (convenient sampling).....	33

2.4.2.2 Lấy mẫu định mức (quota sampling)	33
2.4.2.3 Lấy mẫu phán đoán (Judgement sampling).....	34
2.5 DỮ LIỆU ĐỊNH TÍNH VÀ DỮ LIỆU ĐỊNH LƯỢNG	34
CHƯƠNG 3: TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU BẰNG BẢNG VÀ ĐỒ THỊ	
3.1 TÓM LƯỢC VÀ TRÌNH BÀY DỮ LIỆU BẰNG BẢNG TẦN SỐ	38
3.1.1 Cách lập bảng tần số cho dữ liệu định tính	38
3.1.2 Cách lập bảng tần số cho dữ liệu định lượng.....	39
3.1.2.1 Dữ liệu định lượng mà đặc điểm quan tâm có ít biểu hiện	40
3.1.2.2 Dữ liệu định lượng của đặc điểm quan tâm có nhiều biểu hiện	40
3.1.3 Lập bảng tần số bằng Excel	48
3.2 TÓM LƯỢC VÀ TRÌNH BÀY DỮ LIỆU BẰNG ĐỒ THỊ PHÂN PHỐI TẦN SỐ (HISTOGRAM) VÀ ĐA GIÁC TẦN SỐ	53
3.2.1 Đồ thị phân phối tần số	53
3.2.2 Đa giác tần số	56
3.3 BIỂU ĐỒ THÂN VÀ LÁ	56
3.4 TÓM LƯỢC VÀ TRÌNH BÀY DỮ LIỆU ĐỊNH TÍNH DẠNG PHÂN LOẠI BẰNG ĐỒ THỊ	58
3.4.1 Đồ thị dạng thanh (Bar Chart).....	59
3.4.2 Đồ thị hình tròn (Pie Chart).....	61
3.4.3 Cách vẽ đồ thị bằng Excel	62
3.5 BIỂU ĐỒ PARETO	63

CHƯƠNG 4: TÓM TẮT DỮ LIỆU BẰNG CÁC ĐẠI LƯỢNG SỐ

4.1 CÁC ĐẠI LƯỢNG ĐO LƯỜNG MỨC ĐỘ TẬP TRUNG CỦA TẬP DỮ LIỆU VÀ PHƯƠNG PHÁP MÔ TẢ HÌNH DÁNG CỦA TẬP DỮ LIỆU	66
4.1.1 Các đại lượng đo lường độ tập trung phổ biến	66
4.1.1.1 Trung bình cộng (Arithmetic mean)	66
4.1.1.2 Trung vị (Median) - Me	71
4.1.1.3 Số mode (Mo)	72
4.1.1.4 Trung bình nhân (Geometric mean)	73
4.1.2 Sử dụng Excel để tính toán các đại lượng thống kê mô tả độ tập trung..	73
4.1.3 Nhóm các đại lượng khác mô tả sự phân bố của tập dữ liệu	74
4.1.3.1 Tứ phân vị (Quartiles)	74
4.1.3.2 Phân vị (Percentiles)	75
4.1.4 Hình dáng của phân phối.....	76
4.2 CÁC ĐẠI LƯỢNG ĐO LƯỜNG ĐỘ PHÂN TÁN	79
4.2.1 Khoảng biến thiên (Range) – R	80
4.2.2 Độ tráoi giữa (Interquartile Range) – RQ	80
4.2.3 Phương sai và độ lệch chuẩn.....	81
4.3 CÁC ĐẠI LƯỢNG THỐNG KÊ MÔ TẢ CHO BẢNG TẦN SỐ	83
4.3.1 Trung bình cộng	83
4.3.1.1 Trường hợp bảng tần số cho dữ liệu định lượng không phân tổ.....	84
4.3.1.2 Trường hợp bảng tần số cho dữ liệu định lượng có phân tổ	85
4.3.2 Trung vị	86

4.3.3 Số mode (yếu vị).....	86
4.3.4 Phương sai và Độ lệch chuẩn.....	89
4.4 CÁC ĐẠI LƯỢNG THỐNG KÊ MÔ TẢ CHO TỔNG THỂ	90
4.4.1 Trung bình cộng của tổng thể.....	90
4.4.2 Phương sai và độ lệch chuẩn.....	90
4.5 KHÁM PHÁ DỮ LIỆU QUA BIỂU ĐỒ HỘP VÀ RÂU (BOX PLOT)	91
4.6 SỬ DỤNG KẾT HỢP TRUNG BÌNH VÀ ĐỘ LỆCH TIÊU CHUẨN	95
4.6.1 Hệ số biến thiên (Coefficient of variation) - CV	95
4.6.2 Quy tắc thực nghiệm.....	96
4.6.3 Quy tắc Chebyshev.....	97
4.6.4 Chuẩn hóa dữ liệu.....	98
4.7 PHÂN BIỆT MỘT SỐ CẤP KHÁI NIỆM	99
4.7.1 Phân biệt tham số tổng thể và tham số mẫu.....	99
4.7.2 Phân biệt biến thiên và độ lệch chuẩn.....	100

CHƯƠNG 5: XÁC SUẤT CĂN BẢN, BIẾN NGẪU NHIÊN VÀ LUẬT PHÂN PHỐI XÁC SUẤT

5.1 XÁC SUẤT CĂN BẢN.....	101
5.1.1 Ý nghĩa của xác suất.....	101
5.1.2 Phép thử và biến cố.....	102
5.1.2.1 Các định nghĩa	102
5.1.2.2 Một số loại quan hệ giữa các biến cố	103
5.1.3 Tính xác suất theo các định nghĩa về xác suất	105
5.1.3.1 Định nghĩa cổ điển về xác suất	105
5.1.3.2 Định nghĩa thống kê về xác suất (định nghĩa dựa trên kết quả thực nghiệm)	107
5.1.4 Một vài tính chất của xác suất	108
5.1.5 Tính xác suất theo các quy tắc xác suất	109
5.1.5.1 Quy tắc cộng xác suất	109
5.1.5.2 Quy tắc nhân xác suất	111
5.1.5.3 Quy tắc xác suất đầy đủ	112
5.1.5.4 Định lý Bayes	114
5.2 BIẾN NGẪU NHIÊN VÀ CÁC QUY LUẬT PHÂN PHỐI XÁC SUẤT	116
5.2.1 Biến ngẫu nhiên	116
5.2.1.1 Định nghĩa	117
5.2.1.2 Phân loại biến ngẫu nhiên	117
5.2.2 Phân phối xác suất của biến số ngẫu nhiên	118
5.2.2.1 Phân phối xác suất của biến ngẫu nhiên rời rạc	119
5.2.2.2 Phân phối xác suất của biến ngẫu nhiên liên tục	120
5.2.3 Các đặc trưng cơ bản của biến ngẫu nhiên	123
5.2.3.1 Kỳ vọng	124
5.2.3.2 Phương sai	126
5.2.3.3 Độ lệch chuẩn	127
5.2.4 Ứng dụng kỳ vọng vào việc ra quyết định trong kinh doanh	129
5.2.4.1 Khái niệm ra quyết định	129

5.2.4.2 Lập bảng kết toán và ra quyết định bằng phương pháp EMV	129
5.2.4.3 Lập bảng tổng thết cơ hội & ra quyết định bằng phương pháp EOL	130
5.3 CÁC PHÂN PHỐI LÝ THUYẾT QUAN TRỌNG	132
5.3.1 Phân phối lý thuyết cho biến rời rạc	132
5.3.1.1 Phân phối Nhị thức (Binomial Distribution).....	132
5.3.1.2 Phân phối Poisson (Poisson Distribution).....	140
5.3.2 Phân phối lý thuyết cho biến liên tục.....	146
5.3.2.1 Phân phối Bình thường (Normal Distribution)	146
5.3.2.2 Phân phối bình thường chuẩn hóa (Standard	149
Normal Distribution).....	149
5.3.2.3 Dùng phân phối Bình thường tính xấp xỉ một số phân phối rời rạc	155
5.3.2.4 Phân phối đều (Uniform distribution).....	159
5.3.2.5 Phân phối mũ (Exponential distribution)	161
5.3.2.6 Kiểm tra một tập dữ liệu bất kỳ có phân phối bình thường hay xấp xỉ bình thường không?	163

CHƯƠNG 6: PHÂN PHỐI CỦA CÁC THAM SỐ MẪU

6.1 PHÂN PHỐI CỦA TRUNG BÌNH MẪU	171
6.1.1 Trung bình mẫu là ước lượng không chêch của trung bình tổng thể	171
6.1.2 Sai số chuẩn của trung bình mẫu	172
6.1.3 Chọn mẫu từ một tổng thể có phân phối Bình thường	175
6.1.4 Chọn mẫu từ một tổng thể không có phân phối bình thường	177
6.2 PHÂN PHỐI CỦA TỶ LỆ MẪU	179
6.2.1 Khảo sát phân phối của tỷ lệ mẫu	180
6.2.2 Điều chỉnh sai số chuẩn của tỷ lệ mẫu	183

CHƯƠNG 7: ƯỚC LƯỢNG CÁC THAM SỐ TỔNG THỂ

7.1 ƯỚC LƯỢNG TRUNG BÌNH TỔNG THỂ	185
7.1.1 Ước lượng khoảng trung bình tổng thể (biết phương sai tổng thể).....	187
7.1.2 Ước lượng khoảng trung bình tổng thể (không biết phương sai tổng thể)	190
7.1.2.1 Mô tả phân phối t (Phân phối t Student).....	190
7.1.2.2 Ước lượng khoảng của trung bình tổng thể khi cỡ mẫu nhỏ	192
7.2 ƯỚC LƯỢNG TỈ LỆ TỔNG THỂ	193
7.3 XÁC ĐỊNH CỠ MẪU CHO BÀI TOÁN ƯỚC LƯỢNG.....	195
7.3.1 Quy tắc xác định cỡ mẫu cho ước lượng trung bình tổng thể	195
7.3.2 Quy tắc xác định cỡ mẫu cho ước lượng tỷ lệ tổng thể.....	196
7.3.3 Xác định cỡ mẫu trong tình huống tổng thể hữu hạn	197
7.4 ƯỚC LƯỢNG TRÊN HAI MẪU	198
7.4.1 Ước lượng trung bình hai mẫu	198
7.4.1.1 Ước lượng khác biệt hai trung bình tổng thể, trường hợp mẫu độc lập	198
7.4.1.2 Ước lượng khác biệt hai trung bình tổng thể, trường hợp mẫu cặp	204
7.4.2 Ước lượng tỷ lệ hai mẫu	206

CHƯƠNG 8: KIỂM ĐỊNH GIẢ THUYẾT VỀ THAM SỐ TỔNG THỂ

8.1 CÁC VẤN ĐỀ CHUNG VỀ KIỂM ĐỊNH	208
8.1.1 Đặt giả thuyết về tham số tổng thể	208
8.1.2 Một số nguyên tắc liên quan đến việc đặt giả thuyết	208
8.1.3 Logic của bài toán kiểm định.....	209
8.1.4 Xác suất sai lầm loại I và Xác suất sai lầm loại II	210
8.1.5 Mức ý nghĩa của kiểm định (Significance level).....	211
8.1.6 Giá trị tới hạn (Critical Value)	212
8.1.7 Kiểm định một bên và kiểm định hai bên.	212
8.2 KIỂM ĐỊNH GIẢ THUYẾT MỘT MẪU.....	214
8.2.1 Kiểm định giả thuyết về trung bình tổng thể	214
8.2.1.1 Kiểm định giả thuyết về trung bình tổng thể, biết độ lệch chuẩn tổng thể	214
8.2.1.2 Kiểm định giả thuyết về trung bình tổng thể, không biết độ lệch chuẩn tổng thể.....	215
8.2.2 Kiểm định giả thuyết về tỷ lệ tổng thể	223
8.2.3 Kiểm định giả thuyết về phương sai tổng thể.....	224
8.3 KIỂM ĐỊNH GIẢ THUYẾT HAI MẪU	228
8.3.1 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể	228
8.3.1.1 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể, biết phương sai của hai tổng thể, hai mẫu độc lập.....	230
8.3.1.2 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể, không biết phương sai của tổng thể, hai mẫu độc lập cỡ mẫu lớn... ..	230
8.3.1.3 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể, không biết phương sai của hai tổng thể, hai mẫu độc lập cỡ mẫu nhỏ	231
8.3.1.4 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể, hai mẫu không độc lập (mẫu phối hợp từng cặp hay mẫu cặp).....	234
8.3.1.5 Cách thực hiện bằng Excel.....	238
8.3.2 Kiểm định giả thuyết khác biệt giữa hai tỷ lệ tổng thể	240
8.3.2.1 Phương pháp dùng phân phối z	240
8.3.2.2 Phương pháp dùng phân phối Chi Bình phương	242
8.3.3 Kiểm định giả thuyết cho hai phương sai tổng thể	246

CHƯƠNG 9: PHÂN TÍCH PHƯƠNG SAI

9.1 PHÂN TÍCH PHƯƠNG SAI MỘT YẾU TỐ	250
9.1.1 Trường hợp k tổng thể có phân phối bình thường và phương sai bằng nhau.....	251
9.1.2 Thực hiện ANOVA một yếu tố bằng Excel	260
9.1.3 Kiểm tra các giả định của phân tích phương sai	263
9.1.4 Phân tích sâu ANOVA	265
9.2 PHÂN TÍCH PHƯƠNG SAI HAI YẾU TỐ	267
9.2.1 Trường hợp có một quan sát mẫu trong một ô	267
9.2.2 Trường hợp có nhiều quan sát trong một ô	270
9.2.3 Phân tích sâu trong ANOVA 2 yếu tố	280
9.2.4 Thực hiện ANOVA trên chương trình Excel	281

CHƯƠNG 10: KIỂM ĐỊNH PHI THAM SỐ

10.1 KIỂM ĐỊNH DẤU VÀ HẠNG WILCOXON VỀ TRUNG VỊ TỔNG THỂ	284
10.2 KIỂM ĐỊNH TỔNG HẠNG WILCOXON CHO TRUNG BÌNH HAI MẪU ĐỘC LẬP	288
10.3 KIỂM ĐỊNH DẤU VÀ HẠNG WILCOXON CHO MẪU PHỐI HỢP TÙNG CẶP (2 MẪU PHỤ THUỘC).....	290
10.4 KIỂM ĐỊNH KRUSKAL WALLIS CHO NHIỀU MẪU ĐỘC LẬP	292
10.5 KIỂM ĐỊNH CHI-BÌNH PHƯƠNG VỀ TÍNH ĐỘC LẬP (KIỂM ĐỊNH LIÊN HỆ GIỮA 2 BIẾN ĐỊNH TÍNH).....	296
10.6 KIỂM ĐỊNH CHI-BÌNH PHƯƠNG VỀ SỰ PHÙ HỢP.....	299

CHƯƠNG 11: HỎI QUI TUYẾN TÍNH ĐƠN BIẾN VÀ PHÂN TÍCH TƯƠNG QUAN

11.1 LÀM QUEN VỚI HỎI QUI	304
11.1.1 Khái niệm hỏi qui	304
11.1.2 Phân biệt liên hệ thống kê và liên hệ hàm số khi phân tích hỏi qui ...	305
11.1.3 Một số qui ước về ký hiệu và tên gọi	305
11.1.4 Các dạng liên hệ giữa hai biến X và Y	306
11.2 MÔ HÌNH HỎI QUI TUYẾN TÍNH ĐƠN.....	308
11.2.1 Môđ đầu	308
11.2.2 Các giả định liên quan đến yếu tố nhiễu	309
11.2.3 Ý nghĩa của các hệ số hỏi qui	310
11.2.4 Tính toán các kết quả hỏi qui bằng phần mềm Excel	315
11.2.5 Vấn đề cần chú ý khi dự đoán với mô hình hỏi qui	317
11.2.6 Đo lường biến thiên bằng Hệ số xác định	317
11.2.7 Sai số chuẩn của ước lượng	320
11.2.8 Suy diễn thống kê về hệ số độ dốc.....	321
11.2.8.1 Định lý Gauss – Markov.....	321
11.2.8.2 Khoảng tin cậy cho hệ số độ dốc	323
11.2.8.3 Kiểm định ý nghĩa của hệ số độ dốc	324
11.2.9 Phân tích phần dư	326
11.2.9.1 Kiểm tra tính đúng đắn của mô hình hỏi qui tuyến tính	326
11.2.9.2 Kiểm tra sự vi phạm giả định phương sai bằng nhau	328
11.2.9.3 Kiểm tra giả định phân phối bình thường của phần dư	331
11.2.9.4 Kiểm tra tính độc lập của phần dư	334
11.2.10 Sử dụng phân tích hỏi qui dự đoán giá trị trung bình và giá trị cá biệt của biến phụ thuộc Y	337
11.3 TƯƠNG QUAN TUYẾN TÍNH.....	339
11.3.1 Hệ số tương quan tuyến tính tổng thể	340
11.3.2 Hệ số tương quan tuyến tính mẫu r	341
11.3.3 Tính hệ số tương quan tuyến tính bằng Excel	342
11.3.4 Kiểm định ý nghĩa thống kê của hệ số tương quan tuyến tính	344
11.4 TƯƠNG QUAN GIỮA CÁC BIẾN ĐỊNH TÍNH	347
11.4.1 Tương quan hạng Spearman rs	347
11.4.2 Kendall Tau	349

11.4.3 Tương quan đối với dữ liệu thứ bậc trong dữ liệu đã phân nhóm (tau c , gamma, dxy và dxy)	352
CHƯƠNG 12: HỒI QUI TUYẾN TÍNH ĐA BIẾN	
12.1 PHƯƠNG TRÌNH HỒI QUI TUYẾN TÍNH TỔNG THỂ ĐA BIẾN VỚI K BIẾN ĐỘC LẬP
12.1.1 Phương trình hồi qui tổng thể	359
12.1.2 Các hệ số hồi qui riêng phần.....	360
12.2 PHƯƠNG TRÌNH HỒI QUI TUYẾN TÍNH MẪU ĐA BIẾN VỚI 3 BIẾN ĐỘC LẬP	360
12.2.1 Viết phương trình hồi qui tuyến tính mẫu 3 biến độc lập.....	360
12.2.2 Dùng Microsoft Excel để tính toán các hệ số hồi qui mẫu và các số thống kê khác	361
12.2.3 Đọc các con số thống kê cần thiết trên bảng kết quả	363
12.2.4 Đánh giá sự phù hợp của mô hình	363
12.2.4.1 Tính toán hệ số xác định bội	363
12.2.4.2 Hệ số xác định hiệu chỉnh	364
12.2.4.3 Đánh giá ý nghĩa toàn diện của mô hình	365
12.2.4.4 Tính toán sai số chuẩn của ước lượng	367
12.2.4.5 Đánh giá ý nghĩa của từng biến độc lập riêng biệt	368
12.2.5 Hiện tượng đa cộng tuyến	369
12.2.5.1 Ảnh hưởng của đa cộng tuyến	369
12.2.5.2 Cách phát hiện mô hình có tồn tại hiện tượng đa cộng tuyến.....	370
12.2.5.3 Khắc phục đa cộng tuyến	371
12.2.6 Diễn giải các ý nghĩa các hệ số hồi qui riêng	373
12.2.7 Phân tích phần dư	374
12.2.7.1 Kiểm tra sự phù hợp khi lựa chọn mô hình hồi qui tuyến tính....	374
12.2.7.2 Kiểm tra giả định phương sai không đổi	375
12.2.7.3 Kiểm tra giả định không có tự tương quan giữa các phần dư	376
12.2.8 Dự đoán giá trị cụ thể của biến phụ thuộc	378
12.3 HỒI QUI VỚI BIẾN ĐỘC LẬP ĐỊNH TÍNH.....	378
12.4 LIÊN HỆ PHI TUYẾN.....	384
12.4.1 Dạng hàm bậc 2	385
12.4.1.1 Kết quả chạy hồi qui trên Excel	386
12.4.1.2 Phương trình hồi qui tuyến tính mẫu	387
12.4.1.3 Đánh giá độ phù hợp của mô hình	388
12.4.1.4 Đánh giá tác động bậc 2	389
12.4.2 Dạng log kép	389

CHƯƠNG 13: CHỈ SỐ

13.1 MỘT SỐ VẤN ĐỀ CHUNG VỀ PHƯƠNG PHÁP CHỈ SỐ.....	391
13.1 Khái niệm chỉ số.....	391
13.2 Phân loại chỉ số	391
13.2 CHỈ SỐ CÁ THỂ.....	391
13.2.1 Chỉ số cá thể giá cả	392

13.2.2 Chỉ số cá thể khối lượng	392
13.3 CHỈ SỐ TỔNG HỢP	393
13.3.1 Chỉ số tổng hợp giá cả	393
13.3.1.1 Chỉ số Laspeyres	393
13.3.1.2 Chỉ số Paasche	394
13.3.1.3 Chỉ số Fisher	395
13.3.2 Chỉ số tổng hợp khối lượng	396
13.3.3 Chỉ số của chỉ tiêu chất lượng và chỉ số của chỉ tiêu khối lượng	397
13.4 CHỈ SỐ LIÊN HOÀN VÀ CHỈ SỐ ĐỊNH GỐC	397
13.4.1 Chỉ số liên hoàn	397
13.4.2 Chỉ số định gốc	397
13.5 CHỈ SỐ KHÔNG GIAN (CHỈ SỐ ĐỊA PHƯƠNG)	401
13.5.1 Chỉ số tổng hợp giá cả theo không gian	402
13.5.2 Chỉ số tổng hợp khối lượng theo không gian	402
13.6 HỆ THỐNG CHỈ SỐ	404
13.6.1 Hệ thống chỉ số tổng hợp	404
13.6.2 Hệ thống các chỉ số liên hoàn và định gốc	406
13.6.3 Hệ thống chỉ số nghiên cứu biến động của chỉ tiêu trung bình	408
13.6.4 Hệ thống chỉ số phân tích biến động của chỉ tiêu tổng trị số	411
13.7. MỘT SỐ CHỈ SỐ THƯỜNG GẶP TRONG THỰC TẾ	414
13.7.1 Chỉ số giá tiêu dùng (CPI)	414
13.7.2 Chỉ số chứng khoán VN-Index	415

CHƯƠNG 14: CHUỖI THỜI GIAN VÀ DỰ BÁO TRÊN CHUỖI THỜI GIAN

14.1 CHUỖI THỜI GIAN	419
14.1.1 Khái niệm	419
14.1.1.1 Chuỗi thời kỳ	420
14.1.1.2 Chuỗi thời điểm	420
14.1.2 Các đại lượng mô tả chuỗi thời gian	420
14.1.2.1 Mức độ trung bình theo thời gian	420
14.1.2.2 Lượng tăng (giảm) tuyệt đối	421
14.1.2.3 Tốc độ phát triển	422
14.1.2.4 Tốc độ tăng (giảm)	422
14.1.2.5 Tri tuyệt đối của 1% tăng (giảm) liên hoàn	423
14.2 DỰ BÁO TRÊN CHUỖI THỜI GIAN	423
14.2.1 MỘT SỐ VĂN ĐỀ LIÊN QUAN ĐẾN DỰ BÁO	425
14.2.1.1 Thời đoạn dự báo	425
14.2.1.2 Tầm xa dự báo	425
14.2.1.3 Đánh giá độ phù hợp của mô hình dự báo	425
14.2.2 CÁC PHƯƠNG PHÁP DỰ BÁO ĐƠN GIẢN	429
14.2.2.1 Dự đoán bằng lượng tăng (giảm) tuyệt đối trung bình	429
14.2.2.2 Dự đoán bằng tốc độ phát triển trung bình	429
14.2.2.3 Dự báo bằng phương pháp trung bình trượt (Moving Average)	430
14.2.2.4 Mô hình ngoại suy xu thế	433
14.3 DỰ BÁO BẰNG MÔ HÌNH NHÂN	434

14.4 DỰ BÁO BẰNG HÀM TĂNG TRƯỞNG MŨ	444
14.5 DỰ BÁO BẰNG SAN BẰNG HÀM SỐ MŨ	447
14.5.1 San bằng hàm mũ đơn giản	448
14.5.1.1 Lý thuyết dự báo bằng phương pháp san bằng hàm mũ đơn giản	448
14.5.1.2 Dùng Excel thực hiện phương pháp san bằng hàm mũ đơn giản	452
14.5.2 Phương pháp Holt	457
14.5.3 Phương pháp Holt – Winter	462
CHƯƠNG 15: DỰ BÁO BẰNG PHƯƠNG PHÁP BOX-JENKINS	
15.1 KIỂM TRA TÍNH TƯƠNG QUAN TRONG DỮ LIỆU CHUỖI THỜI GIAN	470
15.1.1 Hệ số tự tương quan	470
15.1.2 Kiểm tra tính tương quan	472
15.2 TÍNH DỨNG CỦA CHUỖI THỜI GIAN	476
15.2.1 Khảo sát tính dừng	476
15.2.2 Loại bỏ tính dừng	479
15.3 HỆ SỐ TỰ TƯƠNG QUAN RIÈNG	481
15.4 MÔ HÌNH BOX – JENKINS (ARIMA) CHO CHUỖI DỨNG VÀ DỰ BÁO	482
15.4.1 Các quá trình tự hồi qui (AR)	483
1.4.1.1 Phương trình	483
1.4.1.2 Khảo sát dấu hiệu nhận dạng mô hình tự hồi qui	483
15.4.2. Các quá trình trung bình trượt (MA)	487
1.4.2.1 Phương trình	487
1.4.2.2 Khảo sát dấu hiệu nhận dạng mô hình trung bình trượt	488
15.4.3 Các quá trình phối hợp tự hồi qui – trung bình trượt (ARMA)	490
15.4.3.1 Phương trình	490
15.4.3.2 Khảo sát dấu hiệu nhận dạng mô hình tự hồi qui – trung bình trượt	490
15.5 MÔ HÌNH BOX – JENKINS ARIMA CHO CHUỖI KHÔNG DỨNG VÀ DỰ BÁO	491
15.6 MÔ HÌNH BOX-JENKINS CHO CHUỖI THỜI GIAN CÓ TÍNH MÙA VỤ	493
15.6.1 Nhận dạng tính mùa trong một chuỗi thời gian	493
15.6.2 Biến đổi chuỗi thời gian có tính mùa thành chuỗi thời gian dừng và dự báo	495

CHƯƠNG 1

GIỚI THIỆU MÔN HỌC

1.1 THỐNG KÊ LÀ GÌ?

Nói đến thống kê nhiều người thường liên tưởng đến các con số, các số liệu được sắp xếp trong các bảng biểu, hay những đồ thị biểu diễn những dữ liệu về kinh tế - xã hội như dân số, việc làm, thất nghiệp, giá vàng, lượng gạo xuất khẩu, GDP... bởi vì hiểu theo nghĩa thông thường, danh từ "thống kê" được đồng nghĩa với số liệu. Ví dụ như chúng ta hay nghe nói tới thống kê tai nạn xe máy, thống kê về giá sinh hoạt, thống kê về vốn đầu tư nước ngoài vào Việt Nam, thống kê về thị trường chứng khoán ... Hiện nay nghĩa thông thường trên không thể diễn tả đầy đủ thống kê hiện đại vì thống kê không còn giới hạn trong việc thu thập dữ liệu hay lập các bảng tổng hợp các dữ liệu.

1.1.1 Xuất phát thuật ngữ thống kê

Thuật ngữ thống kê đầu tiên bắt nguồn từ tiếng Latinh "Statisticum collegium", trong tiếng Anh gọi là "Council of state", đều có nghĩa là "hội đồng chính quyền"; và một từ tiếng Ý là "Statista", tiếng Anh là "Statesman" hay "Politician", có nghĩa là người làm cho chính quyền hay người làm chính trị. Thuật ngữ tiếng Đức "Statistik" lần đầu tiên do Gottfried Achenwall (1749) giới thiệu xuất phát từ việc phân tích dữ liệu về chính quyền, có nghĩa là "khoa học về trạng thái" (sau đó trong tiếng Anh gọi là số học chính trị). Trong thế kỷ 19, thuật ngữ thống kê được hiểu một cách phổ biến là thu thập và phân loại dữ liệu, thuật ngữ này được John Sinclair đưa vào tiếng Anh. Do vậy mục đích chính đầu tiên của thống kê là dữ liệu được sử dụng để đáp ứng nhu cầu của chính phủ và các cơ quan quản lý. Việc thu thập dữ liệu về chính quyền và địa phương diễn ra chủ yếu ở các cơ quan thống kê nhà nước và quốc tế ví dụ như tổng điều tra dân số quốc gia.

Ngày nay việc sử dụng thống kê đã mở rộng hơn rất nhiều so với xuất phát điểm đầu tiên là phục vụ cho chính quyền hay chính phủ. Các tổ chức và cá nhân sử dụng thống kê để tìm hiểu dữ liệu và ra quyết định. Thống kê được sử dụng từ khoa học tự nhiên, cho đến khoa học xã hội, y dược học, kinh doanh và rất nhiều lĩnh vực khác. Thống kê nói chung không được xem là nhánh của toán học, mà là một lĩnh vực riêng, có quan hệ với toán học. Nhiều trường đại học có 2 khoa hay 2 bộ môn riêng về toán và thống kê. Ngoài khoa kinh tế, quản trị kinh doanh, ở các

trường khối kinh tế xã hội, thống kê ~~còn~~ được giảng dạy trong nhiều ngành đào tạo như xã hội học, tâm lý học, giáo dục học, sức khỏe cộng đồng...

1.1.2 Khái niệm Thống kê

Khái niệm thống kê theo nhiều tác giả trong nghiên cứu kinh tế và xã hội:

- Thống kê liên quan đến nhiều vấn đề khác nhau bao gồm phân tích và trình bày dữ liệu, thiết kế nghiên cứu thử nghiệm, và ra quyết định (Wyatt và Bridges, 1967)
- Thống kê liên quan đến việc phát triển và áp dụng các phương pháp, kỹ thuật trong việc thu thập, phân tích, và thảo luận – giải thích những dữ liệu sao cho dựa trên các dữ liệu quan sát được, người ta có thể đưa ra các kết luận đáng tin cậy về một vấn đề nghiên cứu (Ngọc và Tươi, 1974)
- Thống kê có thể được định nghĩa là việc thu thập, trình bày, phân tích và diễn giải các dữ liệu dưới dạng số (Croxton và ctg, 1988)
- Thống kê là ngành cung cấp nhiều phương pháp hỗ trợ cho việc phân tích dữ liệu và ra quyết định (Groebner và ctg, 2005)

Một cách tổng quát, có thể định nghĩa về thống kê như sau:

Thống kê là một nhánh của toán học liên quan đến việc thu thập, phân tích, diễn giải hay giải thích và trình bày các dữ liệu. Thống kê được vận dụng trong nhiều lĩnh vực học thuật khác nhau, từ vật lý, cho đến khoa học xã hội và nhân văn. Thống kê cũng được sử dụng để ra quyết định (và đôi khi cũng bị lạm dụng vì những lý do khác nhau) trong tất cả mọi lĩnh vực kinh doanh và quản lý nhà nước.

Thống kê mô tả, thống kê suy diễn, thống kê ứng dụng

Thống kê mô tả là các phương pháp sử dụng để tóm tắt hoặc mô tả một tập hợp dữ liệu. Còn thống kê suy diễn là các phương pháp mô hình hóa trên các dữ liệu quan sát để giải thích được những biến thiên “dường như” có tính ngẫu nhiên và tính không chắc chắn của các quan sát, và dùng để rút ra các suy diễn về quá trình hay về tập hợp các đơn vị được nghiên cứu. Thống kê mô tả và thống kê suy diễn tạo thành thống kê ứng dụng. Còn thống kê toán là lĩnh vực nghiên cứu cơ sở lý thuyết của khoa học thống kê.

Thuật ngữ “thống kê” (Statistics, số *nhiều* của statistic) cũng được sử dụng như số liệu tổng hợp, là kết quả của một quá trình nghiên cứu thống kê, đối với một tập hợp dữ liệu. Ví dụ như thống kê về việc làm, thống kê tai nạn giao thông ...

1.1.3 Tổng quan về thống kê

Khi ứng dụng thống kê để giải quyết các vấn đề khoa học, công nghiệp, hay xã hội, chúng ta bắt đầu từ đối tượng cần nghiên cứu là một quá trình hay một tổng thể (xem khái niệm ở trang 8). Đối tượng nghiên cứu có thể là một tổng thể như dân số của một quốc gia, hay hàng hóa do một doanh nghiệp sản xuất trong một thời kỳ nhất định. Đối tượng nghiên cứu cũng có thể là một quá trình được quan sát nhiều lần, dữ liệu thu thập từ loại “tổng thể” này gọi là chuỗi thời gian.

Vì những lý do thực tế, giới hạn về thời gian và ngân sách, thay vì thu thập dữ liệu trên toàn bộ tổng thể, chúng ta thường nghiên cứu trên một bộ phận của tổng thể, gọi là mẫu. Dữ liệu thu thập từ mẫu có thể trong bối cảnh quan sát hay thử nghiệm. Sau đó dữ liệu được phân tích thống kê theo hai mục tiêu: mô tả và suy diễn.

Thống kê mô tả được dùng để tóm tắt dữ liệu, để mô tả mẫu nghiên cứu dưới dạng số hay đồ họa. Các công cụ số dùng để mô tả thường dùng nhất là trung bình cộng và độ lệch chuẩn. Các công cụ đồ họa bao gồm các biểu đồ và đồ thị.

Thống kê suy diễn được dùng để mô hình hóa các kiểu biến thiên trong dữ liệu, giải thích những biến thiên có vẻ ngẫu nhiên và rút ra kết luận về tổng thể nghiên cứu mà chúng ta thường không có điều kiện khảo sát hết. Những kết luận này có thể dưới dạng trả lời các câu hỏi yes/no (kiểm định giả thuyết, ví dụ như người mua hàng lớn tuổi chú ý đến yếu tố chất lượng hơn giá cả khi mua hàng), ước lượng các đặc trưng số học (ví dụ ước lượng chi tiêu trung bình hàng tháng của một phụ nữ tuổi từ 30 đến 40 để mua mỹ phẩm), dự đoán các giá trị tương lai, mô tả mối liên hệ (tương quan), hay mô hình hóa mối liên hệ (hồi qui). Ngoài ra còn rất nhiều kỹ thuật mô hình hóa khác nữa như Phân tích phương sai, Chuỗi thời gian, và Khai thác dữ liệu.

Nếu mẫu nghiên cứu là đại diện của tổng thể thì suy diễn và kết luận rút ra từ mẫu có thể mở rộng được cho tổng thể. Vấn đề chính nằm ở chỗ chúng ta xác định mức độ đại diện của mẫu tới đâu. Thống kê cung cấp các phương pháp để ước lượng và hiệu chỉnh những ngẫu nhiên xảy ra trong mẫu và trong cách thu thập dữ liệu, cũng như các phương pháp thiết kế các thử nghiệm đủ mạnh để tăng tính đại diện của mẫu.

Khái niệm toán căn bản được dùng để hiểu tính ngẫu nhiên chính là xác suất. Thống kê toán (còn gọi là thống kê lý thuyết) là một nhánh của toán ứng dụng, sử dụng lý thuyết và phân tích xác suất để nghiên cứu cơ sở lý thuyết của thống kê như các luật phân phối.

Việc sử dụng bất kỳ phương pháp thống kê nào cũng chỉ đúng đắn khi tổng thể nghiên cứu thỏa mãn những giả thiết toán học cần thiết của phương pháp. Việc sử dụng sai kết quả thống kê có thể tạo ra những sai lầm nghiêm trọng, nhưng khó thấy, trong việc mô tả và diễn giải. Khó thấy ở chỗ ngay cả các nhà chuyên môn có kinh nghiệm đôi khi cũng phạm những sai lầm đó, và nghiêm trọng ở chỗ chúng có thể ảnh hưởng đến chính sách xã hội, phác đồ điều trị trong y khoa, độ tin cậy của cầu đường hay nhà máy, quyết định về chính sách, chiến lược kinh doanh ...

Ngay cả khi thống kê được áp dụng đúng thì các kết quả khó mà diễn giải cho những người không có chuyên môn hiểu. Ví dụ như mức ý nghĩa thống kê của xu hướng tìm được trong tập hợp dữ liệu (là thước đo khả năng xu hướng tìm được là do biến thiên ngẫu nhiên trong mẫu, không phải là xu hướng thật sự của tổng thể) có thể không giống với cảm nhận trực tiếp.

1.2 CÁC PHƯƠNG PHÁP NGHIÊN CỨU THỐNG KÊ

Mục đích thông thường của nghiên cứu thống kê là xem xét mối liên hệ nhân quả, và đặc biệt là kết luận về ảnh hưởng của những sự thay đổi của những biến độc lập đến biến phụ thuộc. Có hai loại nghiên cứu thống kê nhân quả: nghiên cứu thử nghiệm (*experimental studies*) và nghiên cứu quan sát (*observational studies*). Trong cả hai loại nghiên cứu này, ảnh hưởng của biến độc lập lên biến thiên của biến phụ thuộc đều được xem xét, nhưng khác nhau ở chỗ cách thức thực hiện nghiên cứu. Cả hai cách đều rất hiệu quả.

Nghiên cứu thử nghiệm thực hiện việc đo lường đối tượng nghiên cứu, thay đổi điều kiện của đối tượng, và đo lường lại đối tượng với cùng một cách đo để xác định xem sự thay đổi được kiểm soát chủ động này có làm thay đổi các giá trị đo đạc hay không. Còn nghiên cứu quan sát lại không thực hiện điều khiển biến nguyên nhân có kiểm soát, mà chỉ thu thập các dữ liệu cần nghiên cứu và khảo sát tương quan giữa biến nguyên nhân và biến kết quả.

Ví dụ về nghiên cứu thử nghiệm là nghiên cứu về ánh sáng tại một nhà máy sản xuất thiết bị điện. Những người nghiên cứu muốn biết tăng cường độ chiếu sáng có làm tăng năng suất lắp ráp sản phẩm của công nhân hay không. Đầu tiên người nghiên cứu đo lường năng suất lắp ráp sản phẩm của các công nhân trong dây chuyền sản xuất, sau đó tăng cường độ sáng trong khu vực nhà máy để xem sự tăng cường độ chiếu sáng có ảnh hưởng đến năng suất lao động hay không. Kết quả là năng

suất lao động được cải thiện trong mọi điều kiện ánh sáng được tăng cường. Tuy nhiên, cuộc nghiên cứu này cũng bị cho là còn thiếu sót trong thủ tục thử nghiệm, nhất là thiếu một nhóm đối chứng và chưa thử nghiệm trong điều kiện ánh sáng tối hơn.

Ví dụ về nghiên cứu quan sát là cuộc nghiên cứu thăm dò tương quan giữa hút thuốc lá và ung thư phổi. Kiểu nghiên cứu tiêu biểu này được thực hiện bằng cách thực hiện một cuộc khảo sát để thu thập dữ liệu quan sát trên cả hai nhóm có hút thuốc lá và không hút thuốc lá, sau đó đếm số trường hợp bị ung thư phổi trong mỗi nhóm và tính ra tỉ lệ xem nhóm nào có tỉ lệ ung thư cao hơn.

Các bước cơ bản để thực hiện một nghiên cứu thử nghiệm bao gồm:

- Lập kế hoạch nghiên cứu: xác định nguồn thông tin, lựa chọn đối tượng nghiên cứu, các vấn đề liên quan đến đạo đức,
- Thiết kế cuộc thử nghiệm tập trung vào mô hình hệ thống và tương tác giữa biến nguyên nhân và biến kết quả,
- Tóm tắt các giá trị quan sát để làm nổi bật những điểm chung của tập dữ liệu thu được,
- Từ thực tế quan sát kết luận về vấn đề đang nghiên cứu
- Viết báo cáo và trình bày các kết quả của cuộc nghiên cứu

1.3 THỐNG KÊ ỨNG DỤNG TRONG KINH TẾ VÀ XÃ HỘI

Thống kê hiện nay đã được ứng dụng vào mọi lĩnh vực. Các phương pháp và công cụ thống kê đã được vận dụng đan xen trong một số nội dung của nhiều môn học. Một số lĩnh vực nghiên cứu sử dụng thống kê ứng dụng nhiều đến mức mỗi ngành đã đưa ra môn học riêng và đặt tên riêng để nói về thống kê ứng dụng trong ngành của mình. Có thể kể ra một số “thống kê ngành” hay môn học:

- Thống kê bảo hiểm
- Thống kê trong kỹ thuật
- Thống kê trong sinh học
- Kinh tế lượng
- Thống kê trong kinh doanh
- Thống kê dân số
- Thống kê trong tâm lý học
- Thống kê trong giáo dục học
- Thống kê xã hội (cho tất cả các ngành khoa học xã hội)
- Phân tích xử lí và chemometric (phân tích dữ liệu từ phân tích hóa học)

- Thống kê độ tin cậy của công nghệ
- Địa lí và hệ thống thông tin địa lí, đặc biệt trong phân tích không gian
- Thống kê trong xử lí hình ảnh
- Thống kê trong thể thao ...

Trong lĩnh vực xã hội nói chung và kinh tế và kinh doanh nói riêng, thống kê đã đóng vai trò là một công cụ cơ bản quan trọng trong việc nhận thức tình hình và hỗ trợ ra quyết định. Thống kê được dùng để nhận ra và hiểu các biến thiên có hệ thống khi đo lường các hiện tượng kinh tế xã hội, để tóm tắt dữ liệu, và để đưa ra quyết định dựa trên dữ liệu.

Một số phán xét về thống kê

“Ông ta dùng thống kê như một người say dùng cột đèn – để dựa thay vì chiếu sáng” Andrew Lang

Một số người thường cho rằng kiến thức thống kê thường bị dùng sai có chủ ý qua việc tìm cách diễn giải dữ liệu sao cho có lợi cho người trình bày. Một câu nói nổi tiếng là: “Có ba kiểu nói dối: nói dối, nói dối trá hình, và thống kê”. Quyển sách “Làm thế nào để chung sống với thống kê” nổi tiếng của Darell Huff bàn về nhiều trường hợp sử dụng thống kê với mục đích lừa dối, phân tích tập trung vào việc sử dụng các đồ thị sai. Bằng việc chọn (hoặc bác bỏ, hoặc thay đổi) một giá trị nào đó, ta có thể điều khiển được kết quả; bỏ đi các giá trị quan sát quá nhỏ hay quá lớn cũng là một cách làm để thay đổi kết quả. Đây có thể là kết quả của sự lừa dối có chủ ý, hoặc của sai lệch không chủ ý và khó nhận thấy của người nghiên cứu. Do đó, Chủ tịch trường Harvard, Lawrence Lowell, vào năm 1909 đã viết như thế này: “thống kê giống như những cái bánh nhân thịt, nó ngon khi bạn biết người đã làm ra chúng, và khi bạn biết rõ thành phần của chúng”

Vì những kết quả thống kê của các nghiên cứu sau mâu thuẫn với những kết quả thống kê của các nghiên cứu đã được công bố trước đó, một số người trở nên cảnh giác với những kết quả nghiên cứu thống kê như vậy. Người ta có thể đọc một nghiên cứu thống kê nói rằng: “làm điều X sẽ giảm huyết áp cao”, sau đó có một nghiên cứu khác nói “làm điều X thật sự sẽ làm nghiêm trọng thêm tình trạng huyết áp cao”. Thông thường các nghiên cứu do nhiều nhóm khác nhau tiến hành với những thủ tục khác nhau, hoặc những những kết quả thú vị khi nghiên cứu với mẫu nhỏ lại không đúng khi thực hiện những nghiên cứu với mẫu lớn. Tuy nhiên, nhiều người có thể không chú ý những khác biệt này, hoặc các phương tiện truyền thông lại đơn giản hóa quá mức thông tin về bối cảnh nghiên

cứu (vốn khá quan trọng này), và sự mất lòng tin của công chúng về thống kê do đó cũng tăng lên.

Trong các ngành tâm lý học và y dược, đặc biệt là đối với việc các cơ quan quản lý thực phẩm và dược phẩm cấp phép sử dụng các loại thuốc mới, những phê phán về phương pháp kiểm định giả thuyết về hiệu quả và tác dụng của dược phẩm, thực phẩm đã tăng trong những năm gần đây. Đáp lại, người ta nhấn mạnh hơn đến giá trị xác suất (*p*-value) so với việc chỉ báo cáo đơn giản là liệu một giả thuyết có bị loại bỏ ở mức ý nghĩa á đã định hay không. Tuy nhiên, một lần nữa, điều này cho thấy nhiều người chỉ chú ý đến bằng chứng kiểm định có ý nghĩa thống kê hay không, không chú ý đến độ chính xác của kiểm định. Thay vào đó một phương pháp ngày càng được dùng rộng rãi gần đây là báo cáo những khoảng tin cậy, vì những khoảng này chỉ ra độ chính xác của kiểm định và khả năng không chắc chắn. Điều này giúp cho việc diễn giải các kết quả, vì khoảng tin cậy ứng với một mức ý nghĩa á đã cho đồng thời chỉ ra cả ý nghĩa thống kê và tính chính xác của kiểm định. Những kết quả công bố cần được phát biểu chi tiết hơn, thay vì chỉ nêu kết luận đúng hoặc sai của kiểm định.

1.4 MỘT SỐ KHÁI NIỆM DÙNG TRONG THỐNG KÊ

1.4.1 Dữ liệu, thông tin và tri thức (Data, information, knowledge)

Nói chung dữ liệu bao gồm các biểu hiện dùng để phản ánh thực tế của đối tượng nghiên cứu. Phần lớn các biểu hiện này là các trị số đo lường hay quan sát về các biến nghiên cứu. Những biểu hiện này bao gồm, con số, từ ngữ hay hình ảnh.

Thông tin là kết quả của việc xử lý, sắp xếp và tổ chức dữ liệu sao cho qua đó cho người đọc có thêm hiểu biết và tri thức. Nói cách khác, đó là nội dung của dữ liệu đã thu thập. Thông tin là một khái niệm mang nhiều ý nghĩa khác nhau, từ trong cuộc sống hàng ngày cho đến trong môi trường kỹ thuật. Nói một cách tổng quát, khái niệm thông tin có liên quan chặt chẽ đến điều kiện ràng buộc của thông tin, truyền thông, kiểm soát, dữ liệu, hình thức, hướng dẫn, hiểu biết, ý nghĩa, kích thích suy nghĩ, các dạng thức, cảm nhận và trình bày.

Tri thức là những điều đã được biết. Giống như các khái niệm khác có liên quan như sự thật, niềm tin, và sự khôn ngoan, không có một định nghĩa riêng lẻ nào được tất cả các học giả thống nhất, mà có nhiều lý thuyết khác nhau và vẫn còn tranh luận nhau về bản chất của tri thức. Tích lũy tri thức là một quá trình nhận thức phức tạp: cảm nhận, học tập,

truyền thông, liên tưởng và sử dụng lý lẽ. Thuật ngữ tri thức cũng còn được dùng để hàm ý về những hiểu biết tin chắc về một sự vật, có thể dùng hiểu biết này để thực hiện một mục tiêu nào đó

1.4.2 Tổng thể¹ thống kê (Population) và đơn vị tổng thể

Tổng thể thống kê là tập hợp các đơn vị (hay phần tử) thuộc hiện tượng nghiên cứu, cần được quan sát, thu thập và phân tích theo một hoặc một số đặc trưng nào đó. Các đơn vị (hay phần tử) tạo thành tổng thể thống kê gọi là đơn vị tổng thể.

Ví dụ: Muốn tính chỉ tiêu trung bình của một hộ gia đình ở Thành Phố Hồ Chí Minh thì tổng thể nghiên cứu sẽ là toàn bộ các hộ gia đình của Thành Phố Hồ Chí Minh. Muốn tính chỉ tiêu trung bình của sinh viên Đại học Kinh Tế TPHCM chỉ cho việc học thì tổng thể sẽ là toàn bộ sinh viên của Đại học Kinh Tế.

Như vậy việc xác định tổng thể thống kê là xác định chính xác phạm vi các đơn vị tổng thể. Đơn vị tổng thể là xuất phát điểm của quá trình nghiên cứu thống kê, vì nó chứa đựng những thông tin ban đầu cần cho quá trình nghiên cứu.

Tổng thể trong đó bao gồm các đơn vị (hay phần tử) mà ta có thể trực tiếp quan sát hoặc nhận biết được gọi là **tổng thể hộc lột**. (Ví dụ: Tổng thể sinh viên của một trường ; Tổng thể các ngân hàng thương mại...)

Khi xác định tổng thể có thể gặp trường hợp các đơn vị tổng thể không trực tiếp quan sát hoặc nhận biết được, ta gọi đó là **tổng thể tiềm ẩn**. Khi nghiên cứu các hiện tượng kinh tế - xã hội ta thường gặp loại tổng thể này (ví dụ tổng thể những người đồng ý / ủng hộ việc bắt buộc đội nón bảo hiểm khi đi xe máy; tổng thể những người ưa thích đi du lịch sinh thái...)

Tổng thể trong đó bao gồm các đơn vị (hay phần tử) giống nhau ở một hay một số đặc điểm chủ yếu có liên quan trực tiếp đến mục đích nghiên cứu được gọi là **tổng thể đồng chất**. Ngược lại, nếu tổng thể trong đó bao gồm các đơn vị (hay phần tử) không giống nhau ở những đặc điểm chủ yếu có liên quan đến mục đích nghiên cứu được gọi là **tổng thể không đồng chất**. Ví dụ mục đích nghiên cứu là tìm hiểu về hiệu quả sử dụng vốn của các doanh nghiệp trên một địa bàn. Tổng thể các doanh nghiệp dệt trên địa bàn là tổng thể đồng chất, nhưng tổng thể tất cả các doanh nghiệp trên địa bàn là tổng thể không đồng chất. Bởi vì các doanh nghiệp

¹ Tổng là tập hợp, thể là cá thể, phần tử, tổng thể là tập hợp nhiều cá thể hay phần tử

ở những ngành kinh tế khác nhau, tùy theo tính chất sản xuất, quy mô vốn đầu tư ban đầu ... sẽ có hiệu quả sử dụng vốn không so sánh được với nhau. Việc xác định một tổng thể là đồng chất hay không đồng chất là tùy thuộc vào mục đích nghiên cứu cụ thể. Các kết luận rút ra từ nghiên cứu thống kê chỉ có ý nghĩa khi nghiên cứu trên tổng thể đồng chất.

Tổng thể thống kê có thể là hữu hạn, cũng có thể được coi là vô hạn (vô hạn là không thể hoặc khó xác định được số đơn vị tổng thể như tổng thể trẻ sơ sinh, tổng thể sản phẩm do một loại máy sản xuất ra ...). Cho nên khi xác định tổng thể thống kê không những phải giới hạn về thực thể (tổng thể là tổng thể gì), mà còn phải giới hạn về thời gian và không gian (tổng thể tồn tại ở thời gian nào, không gian nào).

1.4.3 Mẫu (Sample)

Mẫu là một số đơn vị được chọn ra từ tổng thể chung theo một phương pháp lấy mẫu nào đó. Các đặc trưng mẫu được sử dụng để suy rộng ra các đặc trưng của tổng thể chung.

1.4.4 Đặc điểm thống kê (Characteristic)

Đặc điểm thống kê là các tính chất quan trọng liên quan trực tiếp đến nội dung nghiên cứu và khảo sát, cần thu thập dữ liệu (cần “thống kê”) trên các đơn vị tổng thể.

Ví dụ khi nghiên cứu nhân khẩu, mỗi nhân khẩu có các tính chất riêng về: giới tính, tuổi, tình trạng hôn nhân gia đình, học vấn, nghề nghiệp, dân tộc, tôn giáo, việc làm... Khi nghiên cứu các doanh nghiệp, mỗi doanh nghiệp có các tính chất riêng như: Loại hình công ty, nguồn vốn chủ sở hữu, lĩnh vực hoạt động chính, số lượng nhân viên, vốn cố định, vốn lưu động, có xuất khẩu hay không, ...

Đặc điểm thống kê được chia thành hai loại:

- **Đặc điểm thuộc tính:** là tính chất của đơn vị tổng thể, không có biểu hiện trực tiếp bằng các con số. Ví dụ như giới tính, nghề nghiệp, tình trạng hôn nhân, dân tộc, tôn giáo của con người, hay loại hình doanh nghiệp... là các đặc điểm thuộc tính.
- **Đặc điểm số lượng:** là đặc điểm của đơn vị tổng thể có biểu hiện trực tiếp bằng con số. Ví dụ như tuổi, chiều cao, trọng lượng của con người, năng suất làm việc của công nhân, quy mô vốn của doanh nghiệp... Các trị số quan sát của các đặc điểm số lượng có thể là rời rạc hay liên tục.

* **trị số rời rạc:** các giá trị có thể có của nó là hữu hạn hay vô hạn và có thể đếm được.

Ví dụ: Số nhân viên trong một doanh nghiệp, số nhân khẩu trong một hộ gia đình, số môn thi lại của một sinh viên, số quốc gia mà công ty đã xuất khẩu hàng đến ...

* **trị số liên tục:** các giá trị có thể có của nó có thể lấp kín cả một khoảng trên trục số. Ví dụ: Trọng lượng, chiều cao của sinh viên.

Các đặc điểm chỉ có hai biểu hiện và không trùng nhau trên một đơn vị tổng thể, được gọi là **đặc điểm nhị phân**. Ví dụ, giới tính là đặc điểm nhị phân vì chỉ có hai biểu hiện là nam và nữ. Đối với đặc điểm có nhiều biểu hiện ta có thể chuyển thành đặc điểm nhị phân bằng cách rút gọn thành hai biểu hiện. Ví dụ, thành phần kinh tế chia thành nhà nước và ngoài nhà nước. Căn cứ vào số lượng lao động, các doanh nghiệp chia thành < 100 và ≥ 100 người.

1.4.5 Chỉ tiêu thống kê

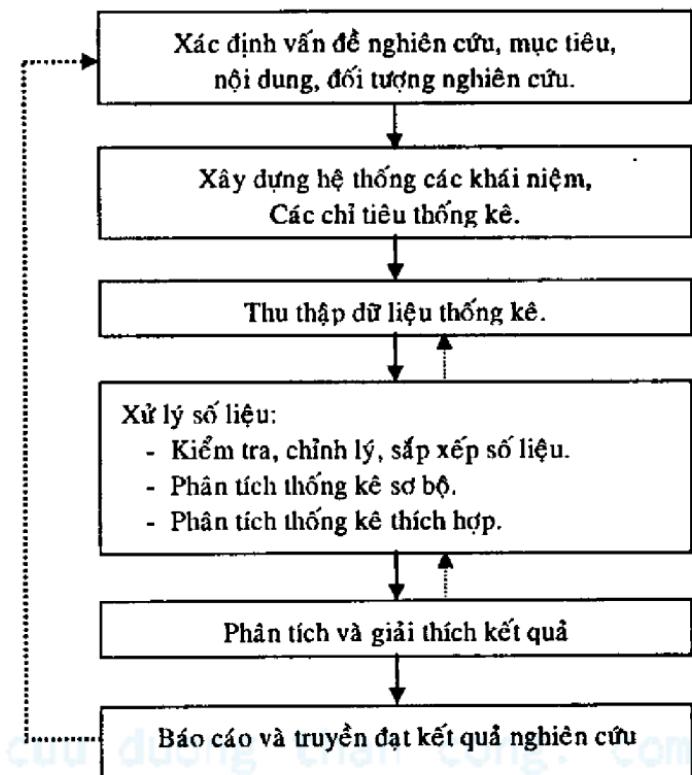
Chỉ tiêu thống kê là con số có ý nghĩa và nội dung trong điều kiện thời gian và không gian xác định. Chỉ tiêu thống kê có thể phân biệt thành hai loại: chỉ tiêu khối lượng và chỉ tiêu chất lượng.

1.4.5.1 Chỉ tiêu khối lượng: là các chỉ tiêu biểu hiện qui mô của tổng thể, ví dụ dân số của một thành phố, số doanh nghiệp trong một ngành, vốn cố định, vốn lưu động của một doanh nghiệp, tổng sản phẩm quốc nội (GDP), số nhân viên có trình độ đại học của doanh nghiệp...

1.4.5.2 Chỉ tiêu chất lượng là các chỉ tiêu biểu hiện tính chất, trình độ phổ biến, quan hệ so sánh trong tổng thể. Ví dụ giá thành đơn vị sản phẩm là một chỉ tiêu chất lượng, nó biểu hiện quan hệ so sánh giữa tổng giá thành và số lượng sản phẩm sản xuất ra, đồng thời nó phản ánh tính chất phổ biến về mức chi phí cho một đơn vị sản phẩm đã được sản xuất ra. Tương tự, các chỉ tiêu năng suất lao động, tiền lương trung bình một nhân viên, tỉ lệ nhân viên có bằng đại học trong doanh nghiệp... là các chỉ tiêu chất lượng. Các chỉ tiêu chất lượng mang ý nghĩa phân tích, trị số của nó được xác định chủ yếu từ việc so sánh giữa các chỉ tiêu khối lượng.

1.5 KHÁI QUÁT QUÁ TRÌNH NGHIÊN CỨU THỐNG KÊ

Quá trình nghiên cứu thống kê hay bất kỳ quá trình nghiên cứu nào, cũng đều trải qua các bước, được khái quát bằng mô hình sau:



Trong mô hình trên, hướng mũi tên từ trên xuống chỉ trình tự tiến hành các công đoạn của quá trình nghiên cứu. Hướng mũi tên từ dưới lên chỉ những công đoạn cần phải kiểm tra lại, bổ sung thông tin hay làm lại nếu chưa đạt yêu cầu.

1.6 CÁC CẤP ĐO LƯỜNG VÀ THANG ĐO

Để thực hiện nghiên cứu, trong thống kê sử dụng 4 cấp bậc đo lường theo mức độ độ thông tin tăng dần đó là thang đo: định danh, thứ bậc, khoảng cách và tỉ lệ. Thang đo định danh (thang đo phân loại) là bậc thấp nhất, không thể hiện sự hơn kém. Thang đo thứ bậc cao hơn ở chỗ thể hiện sự hơn kém, nhưng không thể hiện được chính xác mức độ hơn kém giữa các giá trị. Thang đo khoảng cách thể hiện được khoảng cách (mức độ) hơn kém giữa các giá trị đo lường nhưng không có giá trị 0 có ý nghĩa (ví dụ như chỉ số IQ hay thang đo nhiệt độ là độ C). Thang đo tỉ lệ là cấp bậc đo lường cao nhất, vừa thể hiện khoảng cách hơn kém giữa các giá trị đo lường đồng thời có cả điểm không “tuyệt đối”, do đó có thể tính tỉ lệ so sánh giữa các giá trị quan sát. Thang đo khoảng cách và thang đo tỉ lệ

được nhóm lại thành thang đo định lượng, còn thang đo định danh và thứ bậc được nhóm lại thành thang đo định tính. Stanley Smith Stevens (1946) đề xuất 4 thang đo này và cho rằng các phép tính toán học trên các biến tùy thuộc vào cấp bậc đo lường nào đã sử dụng để thu thập dữ liệu này. Do đó các thống kê mô tả và các kiểm định thống kê phụ thuộc vào biến nghiên cứu được đo lường ở cấp bậc nào.

1.6.1 Thang đo định danh (Nominal scale)

Thang đo định danh là loại thang đo dùng cho các đặc điểm thuộc tính. Người ta thường sử dụng các mã số (code) để phân loại các đối tượng. Ngoài vai trò này các mã số không mang ý nghĩa nào khác. Ví dụ, giới tính, nam ký hiệu số 1, nữ ký hiệu số 2. Giữa các con số ở đây không có quan hệ hơn kém, chỉ dùng để đếm tần số xuất hiện của các biểu hiện. Thước đo độ tập trung duy nhất là mode (đọc là模式). Độ phân tán thống kê có thể đo bằng các tỉ lệ, không tính được độ lệch chuẩn.

Chúng ta hay gặp thang đo định danh trong các câu hỏi về thông tin cá nhân của từng người hay của thông tin về doanh nghiệp. Ví dụ như:

+ Tình trạng hôn nhân của Anh/chị/ông/bà:

1. Có gia đình 2. Độc thân 3. Ly dị 4. Trường hợp khác

Đối với mỗi người, sẽ chọn một trong các mã số 1, 2, 3, 4. Các mã số này là thang đo định danh. Các mã số trên cũng có thể thay đổi như sau:

1. Độc thân 2. Có gia đình 3. Ly dị 4. Trường hợp khác

+ Công ty ông/bà đang hoạt động chính trong lĩnh vực nào?

Sản xuất	<input type="checkbox"/> 1
Xây dựng	<input type="checkbox"/> 2
Dịch vụ	<input type="checkbox"/> 3
Thương mại	<input type="checkbox"/> 4
khác	<input type="checkbox"/> 5

1.6.2 Thang đo thứ bậc (Ordinal scale)

Thang đo thứ bậc thường được sử dụng cho các đặc điểm thuộc tính, và đôi khi cũng được áp dụng cho các đặc điểm số lượng. Trong thang đo này giữa các biểu hiện của đặc điểm có quan hệ thứ bậc hơn kém. Sự chênh lệch giữa các biểu hiện không nhất thiết phải bằng nhau. Thước đo độ tập trung là模式 hay trung vị, trung vị cung cấp nhiều thông tin hơn là模式.

Chúng ta cũng hay gặp loại thang đo này trong các câu hỏi dạng so sánh:

+ Anh/chị/ông/bà hãy xếp hạng các chủ đề sau trên báo Sài Gòn Tiếp Thị tùy theo mức độ quan tâm. (Chủ đề nào quan tâm nhất thì ghi số 1, quan tâm thứ nhì thì ghi số 2, quan tâm thứ ba thì ghi số 3)

- Thông tin thị trường
- Mua sắm
- Gia đình

+ Thu nhập của anh/chị/ông/bà hàng tháng:

1. < 3 triệu đồng 2. Từ 3 – 5 triệu đồng 3. > 5 triệu đồng

+ Xin cho biết mức doanh thu của quý Doanh Nghiệp trung bình hàng tháng?

- | | |
|---------------------|----------------------------|
| Dưới 200 trđ | <input type="checkbox"/> 1 |
| 200 - 500 trđ | <input type="checkbox"/> 2 |
| 500 trđ - 1 tỉ đồng | <input type="checkbox"/> 3 |
| 1 – 3 tỉ đồng | <input type="checkbox"/> 4 |
| Trên 3 tỉ đồng | <input type="checkbox"/> 5 |

1.6.3 Thang đo khoảng (Interval scale)

Thang đo khoảng thường dùng cho các đặc điểm số lượng và đôi khi cũng được áp dụng cho các đặc điểm thuộc tính. Thang đo khoảng là thang đo thứ bậc có các khoảng cách đều nhau. Các phép tính cộng trừ đều có ý nghĩa. Điểm “không” của thang đo này là tùy ý (ví dụ như trong thang đo nhiệt độ thì độ C, độ K và độ F có định nghĩa về điểm “không” hoàn toàn khác nhau), và có thể có giá trị âm. Tỉ số giữa các giá trị thu thập được không có ý nghĩa nên không áp dụng trực tiếp được các phép tính nhân hay chia. Tuy nhiên các khoảng chênh lệch có thể lấy tỉ lệ được, ví dụ như chênh lệch này gấp đôi chênh lệch kia. Khuynh hướng trung tâm của dữ liệu thu thập từ thang đo khoảng có thể là模式, trung vị và trung bình cộng, trong đó trung bình cộng chứa nhiều thông tin nhất.

Ví dụ rõ nhất cho loại thang đo này là nhiệt độ. Ví dụ: $32^{\circ}\text{C} > 30^{\circ}\text{C}$ và $80^{\circ}\text{C} > 78^{\circ}\text{C}$. Sự chênh lệch giữa 32°C và 30°C cũng giống như sự chênh lệch giữa 80°C và 78°C , đó là cách nhau 2°C . Như vậy thang đo khoảng cho phép chúng ta đo lường một cách chính xác sự khác nhau giữa hai giá trị bất kỳ. Còn trong thang đo thứ bậc thì không thể, ta chỉ có thể nói giá trị này lớn hơn giá trị khác. Ta hay gặp loại thang đo này trong các câu hỏi phỏng vấn dạng đánh giá:

Ví dụ: Sau đây là những phát biểu liên quan đến chất lượng dịch vụ đào tạo tại trường ĐH Kinh Tế TPHCM. Xin bạn vui lòng trả lời bằng cách

khoanh tròn một con số ở từng dòng. Những con số này thể hiện mức độ bạn đồng ý hay không đồng ý đối với các phát biểu theo quy ước như sau:

Rất không đồng ý

Rất

	1	2	3	4
1. Chương trình đào tạo của trường phù hợp với yêu cầu của thực tiễn.	1	2	3	4
2. Nội dung các môn học được cập nhật, đổi mới, đáp ứng tối yêu cầu đào tạo.	1	2	3	4
3. Phương pháp giảng dạy của giảng viên phù hợp với yêu cầu của từng môn học.	1	2	3	4
4. GV có kiến thức sâu về môn học đảm trách.	1	2	3	4
5. Cách đánh giá và cho điểm SV công bằng.	1	2	3	4
6. Tổ chức thi cử, giám thị coi thi nghiêm túc.	1	2	3	4
7. Quy mô lớp học (số SV trong một lớp) hợp lý cho việc tiếp thu các môn học.	1	2	3	4
8. Cơ sở vật chất nhà trường (giảng đường, bàn ghế, phương tiện nghe nhìn...) đáp ứng tốt nhu cầu đào tạo và học tập.	1	2	3	4
9. Phòng máy tính đáp ứng tốt nhu cầu thực hành của sinh viên.	1	2	3	4
10. Cơ sở vật chất thư viện tốt (số lượng và chất lượng sách báo, không gian và chỗ ngồi).	1	2	3	4

1.6.4 Thang đo tỷ lệ (Ratio scale)

Thang đo tỷ lệ là loại thang đo dùng cho đặc tính số lượng. Thang đo tỷ lệ có đầy đủ các đặc tính của thang đo khoảng tức là có thể áp dụng các phép tính cộng trừ. Ngoài ra thang đo này có một trị số 0 “thật”, cho phép lấy tỉ lệ so sánh giữa hai giá trị thu thập, cho nên gọi là thang đo tỉ lệ. Đây là loại thang đo ở bậc cao nhất trong các loại thang đo. Tương tự như thang đo khoảng, khuynh hướng trung tâm của dữ liệu thu thập từ thang đo tỉ lệ có thể là một, trung vị và trung bình cộng, trong đó trung bình cộng chứa nhiều thông tin nhất.

Sự khác nhau giữa thang đo khoảng và thang đo tỷ lệ thường bị lẩn lộn vì hai điểm sau:

- Điểm 0 trong thang đo tỷ lệ là một trị số thật.
- Trong thang đo khoảng, sự so sánh về mặt tỷ lệ giữa các giá trị thu thập không có ý nghĩa.

Ví dụ bạn có 5 triệu đồng và anh của bạn có 10 triệu đồng. Như vậy số tiền của anh bạn gấp đôi số tiền của bạn. Nếu ta đổi sang dollars, pounds, lire, yen hoặc marks thì số tiền của anh bạn vẫn gấp đôi số tiền của bạn.

Nếu số tiền của bạn bị mất hay bị đánh cắp thì bạn có 0 đồng. Số 0 ở đây là một trị số thật, Vì thật sự bạn không có đồng nào cả. Như vậy tỷ lệ có trị số 0 thật và là loại thang đo tỷ lệ. Các loại thang đo tỷ lệ khác như mét, kg, tấn, tạ...

Trái lại, với nhiệt độ là thang đo khoảng, ví dụ nhiệt độ hôm nay là 12°C ($53,6^{\circ}\text{F}$) và hôm qua là 6°C ($42,8^{\circ}\text{F}$), ta không thể nói rằng hôm nay ấm áp gấp hai lần hôm qua. Nếu ta đổi từ $^{\circ}\text{C}$ sang $^{\circ}\text{F}$ thì tỷ lệ không còn là 2/1 ($53,6/42,8$). Hơn nữa, nếu nhiệt độ là 0°C , không có nghĩa là không có nhiệt độ. 0°C dĩ nhiên lạnh hơn 6°C . Như vậy nhiệt độ không có trị số 0 thật.

Hai thang đo đầu tiên cung cấp cho chúng ta các dữ liệu định tính, cho nên còn gọi là thang đo định tính. Hai thang đo còn lại cung cấp cho chúng ta dữ liệu định lượng, nên còn gọi là thang đo định lượng. Trong thực tế vấn đề thang đo phức tạp và trở nên quan trọng hơn nhiều, vì chúng ta có thể áp dụng thang đo định tính đối với đặc điểm số lượng (ví dụ như thu nhập, chi tiêu ...), và ngược lại có thể áp dụng thang đo định lượng đối với đặc điểm thuộc tính (đồng ý, không đồng ý). Trong các trường hợp này thì loại dữ liệu ta thu thập được là tùy thuộc vào thang đo, chứ không phải tùy thuộc vào tiêu thức thống kê sử dụng để thu thập dữ liệu.

Ngay cả khi dữ liệu đã thu thập xong, chúng ta còn có thể chuyển đổi dữ liệu định lượng về dạng dữ liệu thứ bậc định tính. Ví dụ như từ dữ liệu thu nhập thật (thang đo tỉ lệ và dữ liệu định lượng) ta có thể biến đổi về dữ liệu về mức thu nhập (thang đo thứ bậc và dữ liệu định tính); quy mô vốn của doanh nghiệp vừa và nhỏ (tỉ đồng) có thể được biến đổi về dạng thứ bậc (dưới 1 tỉ đồng, 1-5 tỉ đồng, 5-10 tỉ đồng, 10-50 tỉ đồng, và trên 50 tỉ đồng). Tuy nhiên việc chuyển đổi ngược lại không thực hiện được. Nghĩa là sau khi đã thu thập thì dữ liệu ở bậc đo lường cao hơn có thể chuyển xuống bậc đo lường thấp hơn, nhưng dữ liệu ở bậc đo lường thấp hơn không thể chuyển lên bậc đo lường cao hơn.

CHƯƠNG 2

THU THẬP DỮ LIỆU

Quá trình nghiên cứu thống kê các hiện tượng kinh tế xã hội cần phải có nhiều dữ liệu. Việc thu thập dữ liệu đòi hỏi nhiều thời gian, công sức và chi phí. Cho nên việc thu thập dữ liệu cần phải được tiến hành một cách có hệ thống để thu thập được dữ liệu cần thiết đáp ứng được mục tiêu nghiên cứu trong khả năng nhân lực, kinh phí và giới hạn thời gian cho phép.

2.1 XÁC ĐỊNH DỮ LIỆU CẦN THU THẬP

Người nghiên cứu có thể thu thập rất nhiều dữ liệu liên quan đến hiện tượng nghiên cứu. Vấn đề quan trọng đầu tiên của công việc thu thập dữ liệu là xác định rõ những dữ liệu nào cần thu thập, thứ tự ưu tiên của các dữ liệu này. Nếu không thì sẽ mất rất nhiều thời gian và chi phí cho những dữ liệu ít quan trọng hay không liên quan đến vấn đề đang nghiên cứu. Xác định dữ liệu cần thu thập xuất phát từ hiểu kỹ vấn đề nghiên cứu và mục tiêu nghiên cứu. Vấn đề nghiên cứu và mục tiêu nghiên cứu càng cụ thể thì xác định dữ liệu cần thu thập càng dễ dàng.

Ví dụ như khi nghiên cứu về vấn đề điều kiện ăn ở sinh hoạt có ảnh hưởng đến kết quả học tập của sinh viên hay không, hai nhóm dữ liệu chính cần thu thập là (1) điều kiện ăn ở sinh hoạt và (2) kết quả học tập. Về nhóm dữ liệu điều kiện ăn ở sinh hoạt, có thể thu thập những dữ liệu liên quan như:

- Ở nhà cha mẹ, ở nhà trọ, ký túc xá, hay ở nhờ nhà bà con, người quen
- Có phòng riêng hay ở chung với người khác, nếu ở với người khác thì bao nhiêu người ở trong một phòng
- Nếu ở chung với cha mẹ hay ở nhờ nhà bà con, người quen thì
 - có phòng riêng hay không, hay ở chung phòng với thành viên khác trong gia đình
 - nếu ở chung phòng với thành viên khác trong gia đình thì có bàn học riêng cho cá nhân hay không
 - có làm việc phụ giúp gia đình không? Có làm việc nhà giúp gia đình không? Thời gian làm mất bao nhiêu?
- Nếu ở nhà trọ hay ký túc xá
 - ở bao nhiêu người trong cùng một phòng
 - có nhà vệ sinh ngay trong phòng hay nhà vệ sinh ở ngoài
- Nơi ở cách chỗ học bao xa

— Chỗ ở có nóng, chật, ồn ào không?

Một số dữ liệu khác về điều kiện ăn ở sinh hoạt, nhưng không liên quan lắm đến mục tiêu nghiên cứu ảnh hưởng của điều kiện ăn ở sinh hoạt đến kết quả học tập thì không nhất thiết phải thu thập, ví dụ như:

— Nhà có nuôi chó mèo không?

— Nhà có trồng cây gì không?

— Nhà có sân không?

— Nhà được xây năm nào? Nền nhà lát bằng vật liệu gì?

— Nhà vệ sinh có hiện đại không, có bồn tắm không?

— Gường ngủ bằng sắt hay bằng gỗ

— Bàn học bằng sắt hay bằng gỗ

— Cả nhà theo tôn giáo gì?

Qua ví dụ trên chúng ta thấy nếu không xác định rõ giới hạn, phạm vi dữ liệu thu thập thì công việc rất nhiều và các dữ liệu thu thập được lại ít ý nghĩa trong việc phân tích để đáp ứng được mục tiêu nghiên cứu đã đề ra.

2.2 DỮ LIỆU THỨ CẤP VÀ DỮ LIỆU SƠ CẤP

Khi thực hiện một nghiên cứu cụ thể, người nghiên cứu có thể sử dụng dữ liệu từ một nguồn có sẵn đã công bố hay chưa công bố, hay tự mình thu thập các dữ liệu cần thiết cho nghiên cứu.

Dữ liệu thu thập từ những nguồn có sẵn, thường là những dữ liệu đã qua tổng hợp, xử lý, gọi là dữ liệu thứ cấp. Dữ liệu sơ cấp là dữ liệu thu thập trực tiếp, ban đầu từ đối tượng nghiên cứu. Ví dụ khi nghiên cứu về ảnh hưởng của điều kiện ăn ở sinh hoạt đến kết quả học tập của sinh viên, những dữ liệu liên quan đến kết quả học tập của sinh viên có thể lấy từ phòng đào tạo hay thư ký khoa như điểm trung bình, số môn thi lại... (dữ liệu thứ cấp) Những dữ liệu có liên quan đến điều kiện ăn ở sinh hoạt của sinh viên thì không có sẵn, chúng ta phải trực tiếp thu thập từ sinh viên (dữ liệu sơ cấp).

Dữ liệu thứ cấp có ưu điểm là thu thập nhanh, ít tốn kém chi phí, nhưng đôi khi ít chi tiết và không đáp ứng đúng nhu cầu nghiên cứu. Ngược lại dữ liệu sơ cấp đáp ứng tốt nhu cầu nghiên cứu nhưng phải tốn kém chi phí và thời gian khá nhiều.

2.2.1 Nguồn dữ liệu thứ cấp

Nguồn dữ liệu thứ cấp khá đa dạng, đối với doanh nghiệp và các tổ chức xã hội có thể sử dụng các nguồn sau:

- Nội bộ: các số liệu báo cáo về tình hình kinh tế như sản xuất, tiêu

thụ, tài chính, nhân sự... của các phòng ban, bộ phận; các số liệu báo cáo từ các cuộc điều tra khảo sát trước đây.

- Cơ quan thống kê nhà nước: các số liệu do các cơ quan thống kê nhà nước (Tổng cục thống kê, Cục thống kê Tỉnh/ Thành phố ...) cung cấp trong Niên giám thống kê. Nội dung chủ yếu là các dữ liệu tổng quát về dân số, lao động, việc làm, giáo dục, mức sống dân cư, tài nguyên, đầu tư, kết quả sản xuất của nền kinh tế, xuất nhập khẩu, ...
- Cơ quan chính phủ: số liệu do các cơ quan trực thuộc chính phủ (Bộ, cơ quan ngang bộ, Ủy ban nhân dân, Ủy Ban Quốc Gia) công bố hay cung cấp. Các số liệu này thường chi tiết hơn và mang tính đặc thù của ngành hay địa phương. Ví dụ như: số lượng người nhập cư, số lượng người mắc bệnh tiêu đường của cả nước hay của TP Hồ Chí Minh (công ty sản xuất, kinh doanh, xuất nhập khẩu sản phẩm y tế hay ngành được săn quan tâm đến con số này), số xe tải và xe buýt đang lưu hành, số người nhiễm HIV, số trẻ em mồ côi, số vụ ly hôn, số tai nạn giao thông đường bộ, số trường hợp kết hôn với người nước ngoài...
- Báo, tạp chí: số liệu mang tính thời sự và cập nhật cao, nhưng mức độ tin cậy phụ thuộc vào nguồn số liệu của chính tờ báo hay tạp chí sử dụng hay cách thức thu thập dữ liệu của các cơ quan này. Ví dụ như số lượng học sinh sinh viên các cấp, các hệ bước vào năm học 2007-2008 là bao nhiêu; số lượng trung tâm ngoại ngữ có phép và cả không phép đang hoạt động; số lượng công ty hay tổ chức làm dịch vụ tư vấn du học, xuất khẩu lao động, tư vấn tâm lý ...
- Các tổ chức, hiệp hội, viện nghiên cứu ...: ví dụ như số lượng doanh nghiệp có sản xuất ống nước nhựa, số lao động trình độ cao trong ngành hay lĩnh vực cụ thể ...
- Các công ty và tổ chức nghiên cứu và cung cấp thông tin theo yêu cầu.

Trong thời đại kỹ thuật số, khá nhiều các dữ liệu thứ cấp đã được nhiều cơ quan chính phủ, các tổ chức, các đơn vị nghiên cứu, các doanh nghiệp, trường học ... đưa lên mạng internet và người nghiên cứu có thể tìm thấy tại các trang web của các đơn vị này. Một cách khác là dùng các máy tìm kiếm (search engine) như google, yahoo ... và các từ khóa (keywords) phù hợp để tìm và chọn lọc các tài liệu trên mạng internet.

2.2.2 Nguồn dữ liệu sơ cấp

Dữ liệu thống kê sơ cấp thường được thu thập theo một quy trình bài bản tùy theo loại nghiên cứu thống kê là nghiên cứu thử nghiệm hay nghiên cứu quan sát.

Trong nghiên cứu thử nghiệm, người nghiên cứu đo đạc và thu thập dữ liệu trên các biến kết quả trong các điều kiện khác nhau của các biến nguyên nhân có ảnh hưởng đang nghiên cứu.

Trong nghiên cứu mang tính quan sát thì các dữ liệu cần thiết có thể thu thập từ nhiều người cung cấp thông tin khác nhau như: người chủ hộ gia đình, người đại diện doanh nghiệp, hay cá nhân ... bằng nhiều hình thức khác nhau. Người thu thập dữ liệu có thể đến gặp người cung cấp thông tin tại địa điểm thuận lợi cho việc thu thập (nhà, văn phòng, trường học...) trực tiếp hỏi và ghi chép các dữ liệu vào phiếu khảo sát hay bản câu hỏi. Hoặc người thu thập có thể gửi bản câu hỏi đến người cung cấp thông tin qua đường bưu điện để người cung cấp thông tin tự trả lời vào lúc thuận tiện.

Dữ liệu sơ cấp trong nghiên cứu quan sát có thể đến từ nội bộ hay từ bên ngoài. Các doanh nghiệp hay tổ chức thường có bộ phận chức năng được giao nhiệm vụ thường xuyên, hàng ngày, hàng tuần, hàng tháng, ghi chép lại các dữ liệu về các hiện tượng, quá trình hay yếu tố cần nghiên cứu. Ví dụ phòng kinh doanh của công ty có người luôn theo dõi, cập nhật và hệ thống số liệu bán hàng của cả công ty. Các tổ chức tài chính, ngân hàng, đầu tư hàng ngày đều theo dõi và ghi chép diễn biến giá vàng, giá ngoại tệ, giá chứng khoán ... trên thị trường. Các bệnh viện, các khoa đều có theo dõi, ghi chép số lượng bệnh nhân đến khám chữa bệnh ...

Khi cần thiết các doanh nghiệp và các tổ chức tiến hành tổ chức thu thập dữ liệu sơ cấp từ bên ngoài, hay thuê các công ty hay tổ chức khác tiến hành thu thập theo yêu cầu của mình. Ví dụ doanh nghiệp có thể tự mình, hay thuê công ty nghiên cứu trường, làm những cuộc điều tra khảo sát để đánh giá mức độ nhận biết sản phẩm thương hiệu, đánh giá về chất lượng sản phẩm của doanh nghiệp. Các tổ chức giáo dục có thể tự mình, hay thuê các tổ chức chuyên nghiệp, thực hiện các đo lường khảo sát về học viên của mình, và cả những người học đang học ở những nơi khác.

2.3 CÁC PHƯƠNG PHÁP THU THẬP DỮ LIỆU SƠ CẤP

Phần này sẽ trình bày những cách thức thu thập dữ liệu sơ cấp khác nhau. Những phương pháp thu thập dữ liệu thường được dùng nhất là:

1. Thực nghiệm
2. Khảo sát qua điện thoại
3. Thư hỏi
4. Quan sát trực tiếp
5. Phỏng vấn cá nhân

2.3.1 Thu thập dữ liệu sơ cấp trong nghiên cứu thực nghiệm

Các công ty và các tổ chức thường thực hiện các thực nghiệm hoặc nhóm các thí nghiệm để thu thập dữ liệu cung cấp cho nhà quản lý để ra những quyết định. Một kế hoạch thực nghiệm dựa trên ý tưởng cơ bản xác định trước yếu tố quan tâm. Một số nhân tố ảnh hưởng được lựa chọn, sẽ được điều khiển hoặc thay đổi sao cho tác động của chúng lên yếu tố quan tâm có thể đo đạc hoặc quan sát được.

Ví dụ một nhà máy chế biến khoai tây chiên cần thực hiện các nghiên cứu về quá trình sản xuất khoai tây. Khách hàng mua khoai chiên của họ đặt ra những tiêu chuẩn nghiêm ngặt về chất lượng khoai mà họ mua vào, một yêu cầu quan trọng là màu sắc của khoai sau khi chiên, chúng phải có màu vàng nâu đồng đều, không quá nhạt màu cũng không quá sậm màu. Khoai tây chiên thành phẩm được làm từ khoai tây đã gọt vỏ, xắt lát, tẩy trắng, nấu chín một phần, và được làm lạnh-khô. Đó không phải là một qui trình đơn giản, vì khoai tây vốn khác nhau ở nhiều mặt (như hàm lượng đường và độ ẩm), thời gian tẩy trắng, nhiệt độ lúc nấu, cùng các nhân tố khác cũng thay đổi từ mẻ này sang mẻ khác.

Nhân viên kỹ thuật bắt đầu thí nghiệm của họ bằng cách nhóm các củ khoai tây sống khác nhau vào những mẻ có tính chất tương tự nhau về môi trường nhiệt độ và thời gian tẩy được thiết lập ở mức độ mà thiết kế thí nghiệm đã xác định. Sau khi kiểm tra thành phẩm của mẻ đó, họ tiếp tục thay môi trường và làm tiếp mẻ khác, rồi lại kiểm tra lại thành phẩm lần nữa. Ghi chép lại kết quả, ví dụ, tỉ lệ phần trăm khoai tây không bị cháy đen thay đổi theo mỗi sự kết hợp loại khoai, thời gian tẩy trắng, và nhiệt độ được ghi lại.

2.3.2 Thu thập dữ liệu sơ cấp trong nghiên cứu quan sát

Nói chung các vấn đề thu thập dữ liệu trong nghiên cứu quan sát như phỏng vấn qua điện thoại, qua thư, quan sát trực tiếp và phỏng vấn cá nhân khá giống nhau. Các nội dung chính để thu thập dữ liệu trong nghiên cứu quan sát bao gồm:

- Từ kế hoạch nghiên cứu, nghĩ ra các câu hỏi và thiết kế thành bản câu hỏi hoàn chỉnh
- Quyết định cách chọn mẫu nếu khảo sát hết toàn bộ tổng thể
- Thực hiện việc thu thập dữ liệu: tiếp cận đối tượng và quan sát, ghi nhận dữ liệu

Phần tiếp theo đây trình bày chi tiết các cách thực hiện thu thập dữ liệu sơ cấp và sau đó là phần nói về các kỹ thuật lấy mẫu.

2.3.2.1 Khảo sát qua điện thoại

Một phương pháp khá đơn giản để thu được dữ liệu về ý kiến của mọi người là khảo sát qua điện thoại. Có thể bạn đã từng gặp một cuộc điện thoại có nội dung như sau “Xin chào. Tôi là Nguyễn Văn A, tôi đại diện cho tổ chức XYZ, chúng tôi đang thực hiện một khảo sát về kết nối internet tại nhà của các hộ gia đình, vui lòng cho tôi nói chuyện với ...”.

Khảo sát qua điện thoại là một công cụ thu thập dữ liệu hữu hiệu và ít tốn kém. Dĩ nhiên, một số người được hỏi sẽ từ chối trả lời, một số người khác không có nhà khi bạn gọi đến, và một số khác nữa không thể liên lạc được qua điện thoại vì lí do này hoặc lí do khác. Như vậy mẫu bạn dự định ban đầu cuối cùng được hoàn thành không như dự kiến.

Khảo sát qua điện thoại thường phải ngắn gọn trong vòng từ 1 tới 3 phút vì phần lớn mọi người sẽ không sẵn lòng nói chuyện lâu qua điện thoại. Các câu hỏi thường là những câu hỏi đóng tức là những câu hỏi yêu cầu người trả lời lựa chọn câu trả lời từ một số lựa chọn xác định.

Ví dụ một câu hỏi đóng có thể như sau “Nhà bạn có kết nối internet không? Trả lời: có hoặc không. Kết nối internet qua hình thức dịch vụ nào? Trả lời: dial-up, ADSL trên đường dây điện thoại, ADSL trên đường cáp truyền hình hay một hình thức nào khác.

Nội dung các câu trao đổi trong cuộc khảo sát nên được thiết kế sẵn thành văn bản, bao gồm phát biểu ngắn ở phần đầu giải thích mục đích của khảo sát và bảo đảm với người trả lời phỏng vấn rằng câu trả lời của họ sẽ được giữ bí mật. Phần đầu nên là những câu hỏi liên quan đến vấn đề trọng tâm của khảo sát. Phần cuối liên quan đến những câu hỏi về thông tin cá nhân người trả lời phỏng vấn (như giới tính, mức thu nhập, trình độ học vấn). Những thông tin này sẽ giúp người nghiên cứu có cái nhìn sâu hơn về kết quả khảo sát.

Thời gian gọi điện khảo sát phù hợp nhất cho một người được phỏng vấn là từ 7 giờ tối đến 9 giờ tối. Tuy nhiên, một số người không ở nhà vào buổi tối và bạn sẽ loại họ khỏi nghiên cứu nếu không có ý định gọi lại.

2.3.2.2 Thư hỏi và những khảo sát dạng viết khác.

Một phương pháp khác được dùng để thu thập ý kiến và dữ liệu thực tế là khảo sát dạng viết. Ví dụ, các khảo sát ở dạng một bức thư hỏi gửi qua đường bưu điện, hoặc các bản khảo sát phát tận tay người được phỏng vấn. Nói chung đây là phương tiện thu thập dữ liệu ít tốn kém nhất. Thư

hỏi và khảo sát dạng viết có thể hỏi chi tiết hơn, do đó cũng cần nhiều thời gian để hoàn thành hơn khảo sát qua điện thoại.

Khảo sát dạng viết gồm cả câu hỏi đóng và mở. Câu hỏi mở là câu hỏi cho phép người được hỏi tự do trả lời theo đánh giá, ngôn ngữ, nhận định của chính họ. Câu hỏi mở tạo cho người được hỏi sự linh hoạt hơn khi trả lời, tuy nhiên, những trả lời như vậy lại khó khăn trong việc xử lý và phân tích. Chú ý rằng, khảo sát qua điện thoại cũng có thể sử dụng cả câu hỏi mở, tuy nhiên, người thực hiện phỏng vấn có thể hiểu sai lời người trả lời khi phải ghi lại một câu trả lời dài.

Nội dung khảo sát dạng viết cũng nên được định hình sao cho dễ hiểu và rõ ràng để người được hỏi có thể cung cấp dữ liệu chính xác và đáng tin cậy. Ngoài ra bản khảo sát dạng viết cần phải dễ nhìn vì bối ngoài của nó như thế nào sẽ ảnh hưởng đến tỉ lệ trả lời, bởi vậy cần phải thiết kế sao cho trông có vẻ chuyên nghiệp.

Nếu một thư hỏi được phát qua bưu điện, bạn thường có một tỉ lệ trả lời thấp, từ 5% đến 20%. Nếu bản khảo sát được phát tận tay người trả lời thì có thể hy vọng một tỉ lệ trả lời cao hơn. Ví dụ, sinh viên thường được đề nghị điền vào một bảng đánh giá khóa học vào cuối học kì, đó là một dạng khảo sát được phát tận tay người được hỏi, và đa số sinh viên sẽ điền đầy đủ vào bảng đó.

Nói chung, các khảo sát dạng viết là phương tiện hiệu quả, ít tốn kém để thu thập dữ liệu nếu bạn có thể khắc phục vấn đề tỷ lệ trả lời thấp.

2.3.2.3 Quan sát trực tiếp và phỏng vấn cá nhân

Quan sát trực tiếp là một công cụ khác, thường được dùng để thu thập dữ liệu. Đúng như tên gọi của nó, quá trình thu thập dữ liệu của kỹ thuật này được thực hiện qua quan sát bằng mắt và dữ liệu được người thu thập ghi lại dựa trên những gì bản thân người thu thập nhận biết được trong quá trình đó.

Có lẽ cách cơ bản nhất để thu được dữ liệu về hành vi con người là quan sát họ. Nếu bạn đang muốn khẳng định phương pháp trình bày sản phẩm ở siêu thị có thể đem lại thoải mái cho người tiêu dùng hoặc không, bạn hãy thay đổi một số trình bày và quan sát phản ứng của khách hàng. Nếu bạn là một nhà sản xuất phim, bạn muốn dự đoán bộ phim mới của bạn có thành công hay không, hãy tổ chức chiếu thử và quan sát phản ứng và nhận xét về bộ phim khi khán giả rời buổi chiếu phim.

Để quan sát có hiệu quả, cần dùng những người quan sát đã được đào tạo bài bản, nhưng điều này làm tăng chi phí. Quan sát cá nhân cũng đòi hỏi

nhiều thời gian. Cuối cùng, khó khăn nữa là nhận thức cá nhân của người thực hiện việc thu thập mang tính chủ quan, những người quan sát khác nhau có thể sẽ nhìn nhận một tình huống không cùng một cách, và báo cáo có thể sẽ không phản ánh cùng một kiểu giống nhau.

Phỏng vấn cá nhân là cách để thu thập dữ liệu từ các đối tượng thông qua hỏi đáp. Phỏng vấn có thể có cấu trúc hoặc không có cấu trúc, và có thể sử dụng hoặc những câu hỏi mở hoặc đóng. Phỏng vấn có cấu trúc là phỏng vấn mà các câu hỏi đã được soạn sẵn thành bản hỏi. Phỏng vấn không có cấu trúc là phỏng vấn bắt đầu với một hoặc nhiều câu hỏi chung, rồi phát triển những câu hỏi sâu hơn dựa trên những câu trả lời trước đã được trả lời như thế nào.

2.3.2.4 Những phương pháp thu thập dữ liệu khác

Thu thập dữ liệu sử dụng những kỹ thuật mới đang dần trở nên thông dụng hơn. Nhiều người cho rằng Wal-Mart là công ty hàng đầu trên thế giới thu thập dữ liệu về hành vi mua sắm của khách hàng một cách tự động bằng cách quét mã vạch trên các sản phẩm khách hàng mua. Không chỉ có bảng kê hàng được cập nhật, mà thông tin về hành vi người mua cũng được ghi lại. Dữ liệu này giúp nhà quản lý tổ chức các cửa hàng của họ sao cho có thể tăng doanh thu bán hàng. Ví dụ, Wal-Mart đã quyết định sắp mặt hàng bia và tã lót dùng 1 lần gần nhau khi họ phát hiện rằng nhiều khách hàng nam cũng mua bia khi đến cửa hàng mua tã lót cho con của họ.

Khi bạn sử dụng thẻ tín dụng hay thẻ ghi nợ, dữ liệu được các nhà bán lẻ cập nhật một cách tự động, hệ thống máy vi tính ngày nay được phát triển để có thể lưu trữ và xử lý dữ liệu để cung cấp thông tin theo ý muốn của người sử dụng.

2.4 CÁC KỸ THUẬT LẤY MẪU

Ở Chương 1 chúng ta đã làm quen với hai thuật ngữ quan trọng là tổng thể và mẫu. Để chọn mẫu từ một tổng thể chúng ta dùng các kỹ thuật lấy mẫu. Vậy tại sao người ta lại chọn mẫu? Việc chọn mẫu không cần thiết trong những trường hợp đối tượng nghiên cứu đồng nhất, khi bạn muốn xét nghiệm máu thì một lọ máu trích ra từ cánh tay cũng không khác gì lọ máu trích ra từ chân bệnh nhân. Tuy nhiên nếu bạn đang nghiên cứu một tổng thể có nhiều yếu tố biến đổi thì khi đó một mẫu được rút ra một cách khoa học từ tổng thể là một điều cần thiết vì nó giúp bạn tiết kiệm được một khoản chi phí và thời gian rất đáng kể so với việc nghiên cứu toàn bộ.

Nếu việc chọn mẫu chỉ nhằm để ít tốn kém hơn và dễ dàng hơn cho nghiên cứu nhưng không thành công trong việc tạo ra những dữ liệu hữu ích thì không có gì đáng để nói. Nhưng các cuộc nghiên cứu căn cứ trên dữ liệu mẫu thực sự có tính đại diện lại thường tốt hơn nghiên cứu toàn bộ tổng thể. Ví dụ khi bạn điều tra một tổng thể là một cộng đồng dân cư khoảng 1.000 hộ gia đình, bạn cần có nhiều phỏng vấn viên để tiếp cận hết mọi người. Các phỏng vấn viên của bạn có thể không sử dụng cùng một cách diễn đạt về các câu hỏi, họ cũng có thể không ứng xử tốt như nhau đối với các đối tượng cần được phỏng vấn thận trọng; họ có thể không cẩn thận như nhau trong việc ghi chép dữ liệu và mã hoá dữ liệu cho việc phân tích.

Ngoài ra một nghiên cứu căn cứ trên một mẫu có tính đại diện thường tốt hơn nghiên cứu toàn bộ tổng thể còn ở chỗ là dữ liệu mẫu có thể có giá trị đo đạc (internal validity) lớn hơn dữ liệu thu thập từ toàn bộ tổng thể..

Từ tổng thể bạn chọn ra mẫu nghiên cứu, tính toán các tham số thống kê đặc trưng trên mẫu (Chương 4 sẽ làm rõ khái niệm tham số thống kê) để từ đó mô tả về tổng thể, hoặc cũng từ các mối quan hệ, các nhận định trên mẫu mà bạn có các suy diễn về tổng thể, như vậy là bạn làm việc với mẫu nhưng mục tiêu cuối cùng của bạn lại là hiểu biết về tổng thể. Mục tiêu đó chỉ đạt được trọn vẹn nếu bảo đảm mẫu được chọn thực sự phản ánh trung thực, đại diện cho toàn bộ tổng thể, cũng như khi chụp hình toàn cảnh một đối tượng, bức hình hoàn hảo là bức hình cho bạn thấy trọn vẹn toàn thể đối tượng được thu nhỏ lại chứ không phải là một góc phía trái hay phía phải, cũng không mất phần chân hay đầu của đối tượng. Các kỹ thuật chọn mẫu đúng đắn sẽ giúp cho bạn chụp được bức hình này.

Có hai nhóm kỹ thuật lấy mẫu là kỹ thuật lấy mẫu xác suất và kỹ thuật lấy mẫu phi xác suất, cả hai nhóm kỹ thuật này đều được sử dụng phổ biến. Phương pháp lấy mẫu xác suất bao gồm các phương pháp chọn mẫu dựa trên nguyên tắc lựa chọn ngẫu nhiên như chọn mẫu ngẫu nhiên đơn giản, chọn mẫu hệ thống, chọn mẫu phân tầng, chọn mẫu cả khối hay nhiều giai đoạn. Nhóm kỹ thuật lấy mẫu phi xác suất bao gồm các phương pháp lấy mẫu thuận tiện, lấy mẫu định mức (quota), lấy mẫu phán đoán.

2.4.1 Kỹ thuật lấy mẫu xác suất (probability sampling)

2.4.1.1 Lấy mẫu ngẫu nhiên đơn giản (Simple random sampling)

Lấy mẫu ngẫu nhiên đơn giản là phương pháp chọn mẫu trong đó mỗi đơn vị của tổng thể được chọn với sự ngẫu nhiên như nhau, hay nói cách khác

là các đơn vị tổng thể được chọn vào mẫu với cơ hội bằng nhau.

Để thực hiện chọn mẫu ngẫu nhiên đơn giản, đầu tiên bạn phải chuẩn bị danh sách các đơn vị của tổng thể cần nghiên cứu, cần thu thập dữ liệu. Danh sách này gọi là khung lấy mẫu hay dàn chọn mẫu (sampling frame). Các đơn vị tổng thể trong danh sách này có thể được sắp xếp theo một trật tự nào đó, ví dụ như theo vần ABC, theo quy mô, theo địa chỉ ... và được gán cho một số thứ tự từ đơn vị thứ 1 đến đơn vị cuối cùng.

Sau khi có khung lấy mẫu, và có số thứ tự từng đơn vị, bạn có thể thực hiện việc lấy đơn vị mẫu ra bằng nhiều cách như bốc thăm, quay số, hay dùng số ngẫu nhiên. Nếu số lượng đơn vị tổng thể ít, khung lấy mẫu ngắn (vài chục hay vài trăm đơn vị), thì bạn có thể viết tên hay số thứ tự của từng đơn vị vào từng lá thăm và bỏ vào một cái hộp, trộn đều lên và bốc ra từng lá thăm, các đơn vị nào có số thứ tự, hay tên, được ghi trong lá thăm được bốc ra là đã được chọn vào mẫu ngẫu nhiên một cách đơn giản rồi.

Khi số lượng đơn vị tổng thể nhiều (vài trăm, vài ngàn hay hơn nữa) thì việc viết hay in danh sách ra và cắt thành các lá thăm trở nên nặng nề và phức tạp, lúc đó bạn có thể dùng cách quay số, hay gần đây là dùng bảng số ngẫu nhiên hay số ngẫu nhiên lấy ra từ hàm ngẫu nhiên trong máy tính bỏ túi, hay chương trình Excel trên máy vi tính.

Trong trường hợp quy mô tổng thể nghiên cứu rất lớn như toàn bộ các sinh viên hệ chính quy của Đại Học Kinh Tế, hay toàn bộ hộ gia đình hay nhân khẩu của Hà Nội hay của TPHCM, cho dù có danh sách tất cả các đơn vị tổng thể thì khối lượng thực hiện khi lấy ra hàng trăm, hàng ngàn mẫu, phải cần hàng trăm hàng ngàn số ngẫu nhiên, thì công việc trở nên rất nặng nề mặc dù quy trình đơn giản. Lúc đó người ta nhờ tới các chương trình máy tính (ví dụ như Excel hay SPSS) để thực hiện việc chọn các đơn vị mẫu một cách ngẫu nhiên. Một giải pháp khác khá đơn giản là dùng cách lấy mẫu hệ thống.

2.4.1.2 Lấy mẫu hệ thống (systematic sampling)

Tuy nhiên trong thực tế nghiên cứu kinh tế - xã hội, nhiều trường hợp bạn không có điều kiện để lấy mẫu ngẫu nhiên đơn giản vì cách làm tuy đơn giản (bốc thăm, quay số, hay dùng bảng số ngẫu nhiên) trở nên nặng nề vì số đơn vị mẫu cần chọn ra khá nhiều hay rất nhiều (vài trăm, vài ngàn đơn vị mẫu). Cho nên các nhà nghiên cứu thường dùng một số phương pháp chọn mẫu ngẫu nhiên khác để thay thế như lấy mẫu hệ thống. Về cơ bản trong lấy mẫu hệ thống, chỉ cần chọn ra một con số ngẫu nhiên là có thể xác định được tất cả các đơn vị mẫu cần lấy ra từ danh sách chọn mẫu (thay vì phải chọn ra n số ngẫu nhiên ứng với n đơn vị mẫu cần lấy ra).

Quy trình thực hiện lấy mẫu hệ thống bao gồm các bước:

- Chuẩn bị danh sách chọn mẫu, xếp thứ tự theo một quy ước nào đó, đánh số thứ tự cho các đơn vị trong danh sách. Tổng số đơn vị trong danh sách là N.
- Xác định cỡ mẫu muốn lấy, ví dụ gồm n quan sát
- Chia N đơn vị tổng thể thành k nhóm theo công thức $k = N/n$, k được gọi là khoảng cách chọn mẫu.
- Trong k đơn vị đầu tiên ta chọn ngẫu nhiên ra 1 đơn vị (bốc thăm hay sử dụng bảng số ngẫu nhiên hay hàm ngẫu nhiên), đây là đơn vị mẫu đầu tiên, các đơn vị mẫu tiếp theo được lấy cách đơn vị này 1 khoảng là k, 2k, 3k ...

Như vậy có thể nói chọn mẫu hệ thống là phương pháp chọn mẫu trong đó các đơn vị chọn mẫu chọn ra cách nhau 1 khoảng là k đơn vị.

Ví dụ: $N = 60$; $n=10$; tính khoảng cách chọn mẫu $k = \frac{N}{n} = \frac{60}{10} = 6$

Chọn 1 số ngẫu nhiên trong khoảng từ 1 đến 6, chẳng hạn chọn được số 4, ta sẽ có các đơn vị mẫu: 4, 10, 16, 22, 28, 34, 40, 46, 52, 58.

Có hai trường hợp chọn mẫu hệ thống:

Trường hợp 1: Lấy mẫu hệ thống đường thẳng (linear systematic sampling) khi k là số nguyên (N chia chẵn cho n)

Ví dụ 1: $N = 64$; $n = 10$;

$$k = \frac{64}{10} = 6,4 \rightarrow \text{lấy } k = 6$$

Chọn 1 số ngẫu nhiên từ 1 đến 6:

Nếu số ngẫu nhiên chọn được là 1, thì các đơn vị lấy ra sẽ là:

1, 7, 13, 19, 25, 31, 37, 43, 49, 55, 64 (chọn được tối 11)

Nếu số ngẫu nhiên chọn được là 2, thì các đơn vị lấy ra sẽ là:

2, 8, 14, 20, 26, 32, 38, 44, 50, 56 (chọn được 10)

...

Nếu số ngẫu nhiên chọn được là 6, thì các đơn vị lấy ra sẽ là:

6, 12, 18, 24, 30, 36, 42, 48, 54, 60 (chọn được 10)

Ví dụ 2: $N = 66$; $n = 10$;

$$k = \frac{66}{10} = 6,6 \rightarrow \text{lấy } k = 7$$

Chọn 1 số ngẫu nhiên từ 1 đến 7:

Nếu số ngẫu nhiên chọn được là 1, thì các đơn vị lấy ra sẽ là:

1, 8, 15, 22, 29, 36, 43, 50, 57, 64 (n = 10)

Nếu số ngẫu nhiên chọn được là 2, thì các đơn vị lấy ra sẽ là:

2, 9, 16, 23, 30, 37, 44, 51, 58, 65 (n = 10)

534

Nếu số ngẫu nhiên chọn được là 4, thì các đơn vị lấy ra sẽ là:

4, 11, 18, 25, 32, 39, 46, 53, 60 (n = 9)

44

Nếu số ngẫu nhiên chọn được là 7, thì các đơn vị lấy ra sẽ là:

7, 14, 21, 28, 35, 42, 49, 56, 63 (n = 9)

Khi N không chia chẵn cho n thì các đơn vị không có cùng 1 xác suất chọn ra như nhau và khi dùng trung bình mẫu để ước lượng trung bình tổng thể thì rất có khả năng bị chêch. Để khắc phục ta sử dụng phương pháp chọn mẫu hệ thống quay vòng.

Trường hợp 2: Lấy mẫu hệ thống quay vòng (Circular systematic sampling) khi k là số thập phân (N không chia chẵn cho n)

Giả sử từ tổng thể N đơn vị cần chọn ra n đơn vị mẫu, thứ tự như sau:

- Tính khoảng cách chọn mẫu $k = \frac{N}{n}$
 - Chọn 1 số ngẫu nhiên trong khoảng từ 1 đến N, đơn vị mẫu đầu tiên có số thứ tự với số đã được chọn ra.
 - Các đơn vị tiếp theo cách đơn vị mẫu đầu tiên 1 khoảng $1k$, $2k$, $3k$, ...
 - Nếu đến hết danh sách N đơn vị chưa đủ n đơn vị mẫu ta quay trở lại đầu danh sách với quy ước: $N + 1$ tương ứng đơn vị thứ nhất; $N + 2$ tương ứng đơn vị thứ 2 trong danh sách.

Ví dụ: $N = 13$; $n = 4$.

Tính khoảng cách chọn mẫu $k = \frac{13}{4} = 3,25 \rightarrow$ chọn $k = 3$

Lấy 1 số ngẫu nhiên trong khoảng từ 1 đến 13, ví dụ chọn được số 6, thì các đơn vị mẫu tương ứng sẽ là: 6, 9, 12, $(12+3-13) = 2$ (Khi số thứ tự lớn hơn N thì số thứ tự = số thứ tự - N)

Cả lấy mẫu ngẫu nhiên đơn giản và lấy mẫu hệ thống cùng đòi hỏi phải cần có danh sách đơn vị. Trong thực tế chọn mẫu ngẫu nhiên hay hệ thống chỉ được áp dụng trong một giai đoạn nào đó hay trong giai đoạn cuối cùng của những thủ tục chọn mẫu khác sẽ được trình bày tiếp theo.

2.4.1.3 Lấy mẫu cả khối/cụm (cluster sampling) và lấy mẫu nhiều giai đoạn (multi-stage sampling)

Đầu tiên tổng thể được chia thành nhiều khối, mỗi khối xem như một tổng thể con, lấy ngẫu nhiên đơn giản m khối, sau đó khảo sát hết các đối tượng trong các khối mẫu đã được lấy ra.

Trong trường hợp danh sách đơn vị (danh sách hộ gia đình hay danh sách nhân khẩu của khu vực khảo sát) không có, người nghiên cứu có thể dùng chọn mẫu cả khối (cluster sampling) hay chọn mẫu nhiều giai đoạn (multistage sampling). Ưu điểm là không cần có danh sách tất cả các đơn vị mà chỉ cần có danh sách của các khối hay của các đơn vị mẫu bậc thấp như danh sách quận, phường, khu phố, tổ dân phố). Khi áp dụng cách chọn cả khối thì do không có danh sách tất cả các đơn vị nên phải dùng danh sách các khối (là một nhóm các đơn vị, ví dụ như đơn vị hành chính: phường, khu phố, tổ dân phố hay khối nhà - block) để chọn ra các khối mẫu. Sau khi chọn ra các khối mẫu thì khảo sát hết tất cả các đơn vị trong khối đó. Ví dụ như quận 3 có 14 phường, sau khi chọn được 2 phường mẫu thì khảo sát hết tất cả các hộ trong phường. Hoặc quận 3, ví dụ có 700 tổ dân phố, sau khi chọn ra 7 tổ dân phố mẫu thì sẽ khảo sát hết tất cả các hộ trong 7 tổ dân phố này.

Trong thực tế thì nếu khảo sát hết tất cả các đơn vị của khối mẫu đã chọn ra thì: một là cỡ mẫu khảo sát thực tế quá lớn và chi phí cao; hai là các đơn vị trong cùng một khối có khuynh hướng khá giống nhau nên không nhất thiết phải khảo sát hết (ví dụ trong cùng một ngõ hẽm, trong cùng một chung cư, trong cùng một khu biệt thự, trong cùng một khu tập thể, trong cùng một lớp ...). Lúc đó trong mỗi khối chọn ra chỉ khảo sát một số đơn vị trong khối này mà thôi. Lúc này mỗi khối chính là đơn vị mẫu bậc 1, mỗi hộ gia đình là đơn vị mẫu bậc 2, và cách chọn mẫu này gọi là chọn mẫu hai giai đoạn.

Ví dụ tổng thể nghiên cứu là quận 3 (để cho đơn giản). Trong quận 3 có 14 phường chia ra 700 tổ dân phố. Đơn vị mẫu bậc 1 có thể là phường hay tổ dân phố. Nếu đơn vị mẫu bậc 1 là phường, ta chọn ngẫu nhiên ra 2 phường mẫu (bằng bốc thăm hay dùng số ngẫu nhiên, hay chọn hệ thống). Nếu đơn vị mẫu bậc 1 là tổ dân phố, giả dụ ta chọn ra 7 tổ dân phố. Trong mỗi phường chọn ra hay mỗi tổ chọn ra, về lý thuyết ta phải đi

lập danh sách các hộ gia đình hay nhân khẩu (đơn vị mẫu bậc 2) từ đó ta chọn ra các đơn vị mẫu bậc 2 là hộ gia đình hay cá nhân. Trong thực tế danh sách này khó lấy được và khó được cập nhật thường xuyên, cho nên người nghiên cứu thường lấy mẫu trên thực địa (field) bằng cách chọn hệ thống theo bước nhảy (ví dụ trong 5 nhà lấy 1 nhà, hay cách 4 nhà khảo sát 1 nhà) hoặc tiến hành theo cách này đến đủ số lượng đơn vị mẫu quy định thì dừng lại.

Phức tạp hơn, thay vì chỉ nghiên cứu ở quận 3, chúng ta sẽ nghiên cứu toàn bộ khu vực nội thành TPHCM. Nếu ta coi quận là đơn vị mẫu bậc 1, hộ gia đình là đơn vị mẫu bậc 2 thì lúc chọn ra được các quận là đơn vị bậc 1 rồi, thì do quy mô của quận quá lớn nên việc chọn hộ gia đình là đơn vị mẫu bậc 2 từ đơn vị mẫu bậc 1 gấp nhiều khăn. Cho nên người ta thường chia nhỏ đơn vị mẫu bậc 1 (quận) thành nhiều đơn vị mẫu bậc 2 ở số lượng vừa phải như phường, khu phố, hay tổ dân phố để có thể chọn ra dễ dàng các đơn vị mẫu bậc 2 này. Như vậy phường, khu phố, tổ dân phố, khối nhà/ô phố là đơn vị mẫu bậc 2. Trong mỗi đơn vị mẫu bậc 2 này như phường, khu phố hay tổ dân phố là một nhóm các đơn vị nghiên cứu cơ bản (hộ, cá nhân) mà chúng ta quan tâm cho gọi là các đơn vị mẫu bậc 3. Lúc này chúng ta có chọn mẫu ba giai đoạn. Cứ tiếp tục như vậy chúng ta có chọn mẫu nhiều giai đoạn.

Chọn mẫu cả khối hay chọn mẫu nhiều giai đoạn giúp chúng ta vượt qua điều kiện đầu tiên của chọn mẫu ngẫu nhiên là phải có danh sách các đơn vị chọn mẫu / khung chọn mẫu (sampling frame) ngay từ đầu.

2.4.1.4 Lấy mẫu phân tầng (Stratified sampling)

Chọn mẫu phân tầng sử dụng khi các đơn vị khác nhau về tính chất liên quan đến vấn đề cần nghiên cứu và khảo sát. Theo phương pháp này tổng thể nghiên cứu được chia thành các tầng lớp, mục tiêu là để các giá trị của các đối tượng tổng thể ta quan tâm thuộc cùng một tầng càng ít khác nhau càng tốt (và như vậy có được sai số lấy mẫu nhỏ hơn chọn mẫu ngẫu nhiên đơn giản hay chọn mẫu hệ thống). Sau đó các đơn vị mẫu được chọn từ các tầng này theo các phương pháp lấy mẫu xác suất thông thường như lấy mẫu ngẫu nhiên đơn giản hay lấy mẫu hệ thống.

Chọn mẫu phân tầng có hai vấn đề quan trọng: phân tầng theo đặc điểm gì và phân bổ số lượng mẫu vào các tầng/lớp khác nhau như thế nào. Đặc điểm dùng để phân tầng phải có liên quan đến nội dung bạn cần nghiên cứu khảo sát. Ví dụ mục đích chính của cuộc nghiên cứu có thể là cách gửi và số tiền gửi ngân hàng trung bình của các cá nhân có gửi tiền. Việc gửi tiền chịu ảnh hưởng của mức sống của hộ gia đình (social economic

class - SEC: loại A, B, C, hay D ...). Chúng ta có thể phân tầng tổng thể nghiên cứu theo SEC rồi tiến hành chọn mẫu trong từng tầng lớp. Phân tầng theo SEC hiện nay chủ yếu là theo tình trạng nhà ở, các vật dụng lâu bền đang sở hữu.

Số đơn vị mẫu trong từng tầng lớp có thể: bằng nhau, theo tỉ lệ của từng class hay phân bổ tối ưu (vừa theo quy mô của tầng lớp và theo mức độ đồng đều của các đơn vị trong cùng một tầng lớp). Phương pháp thường dùng là phân bổ mẫu theo tỉ lệ. Lúc đó chúng ta cần một thông tin đáng tin cậy về cơ cấu của tổng thể theo các tầng lớp này để áp cơ cấu này vào mẫu lấy ra. Quota lấy mẫu cho từng tầng lớp được xác định theo quy tắc tam suât dựa trên quy mô toàn tổng thể, quy mô của từng tầng lớp và quy mô mẫu chung cần lấy ra. Cách làm này sẽ có một mẫu đại diện tốt, tuy nhiên về mặt ước lượng thì không bằng phân bổ tối ưu vì chưa xét đến biến thiên trong từng tầng lớp, và biến thiên trong từng tầng lớp có khả năng rất khác biệt.

Khi quy mô toàn bộ mẫu không lớn lắm, nếu cần tách kết quả ra cho từng tầng lớp thì đôi khi tầng lớp có quy mô nhỏ thì số lượng đơn vị lấy mẫu ra ít, thì kết quả tách riêng ra cho tầng lớp đó ít có ý nghĩa. Lúc đó người nghiên cứu thường dùng phân bổ mẫu cho các tầng lớp đều nhau (mục đích chính là xem kết quả của từng tầng lớp và so sánh giữa các tầng lớp với nhau, mục đích khác là xem xét kết quả của toàn bộ tổng thể), và khi cần có kết quả chung thì sẽ gia trọng (nhân với hệ số) các tầng lớp theo hệ số phản ánh qui mô của từng tầng lớp trong toàn bộ tổng thể.

Giả sử chúng ta cần lấy n đơn vị mẫu từ N đơn vị tổng thể, các đơn vị tổng thể được phân tầng thành k lớp

Nếu dùng phân bổ mẫu đều thì công thức tính số lượng đơn vị mẫu lấy ra trong từng tầng lớp sẽ theo tỉ lệ $\frac{n}{N}$ tức là ($\frac{n_1}{N_1} = \frac{n_2}{N_2} = \dots = \frac{n_k}{N_k} = \frac{n}{N}$)

cụ thể từ tầng lớp thứ i là: $n_i = \frac{n}{N} N_i$

Ví dụ: Tại một trường đại học có 20.000 sinh viên (số liệu đã được làm tròn cho dễ tính và dễ thấy) ở 5 hệ đào tạo và cấp đào tạo khác nhau Bộ phận kiểm định chất lượng tiến hành cuộc khảo sát định kỳ về cảm nhận

về chất lượng và mức độ hài lòng của sinh viên. Mỗi hệ đào tạo cử nhân và cấp đào tạo cao học được coi là một tầng. Số lượng mẫu dự định lấy ra là 1.000 (5% của tổng thể). Nếu phân bổ mẫu vào từng tầng theo tỉ lệ thì chúng ta sẽ lấy 5% đơn vị mẫu ở mỗi tầng như trong bảng sau. Ở dòng thứ nhất lấy 5% của 10.000 sinh viên hệ chính quy thì số lượng mẫu của tầng này là 500. Các tầng khác tính tương tự. Kết quả trong bảng sau:

Hệ đào tạo /cấp đào tạo	Số lượng sinh viên	% sinh viên	Số lượng sinh viên lấy ra từ từng tầng
Cử nhân hệ chính quy	10.000	50%	500
Cử nhân hệ hoàn chỉnh ĐH	2.000	10%	100
Cử nhân hệ văn bằng thứ hai	2.000	10%	100
Cử nhân hệ tại chức	5.000	25%	250
Cao học	1.000	5%	50
	20.000	100%	1.000

Trong bảng này có thể thấy số lượng sinh viên hệ chính quy chiếm tới 50% trong tổng thể ($10.000/20.000$), và mẫu lấy ra là 500, cũng chiếm 50% của toàn bộ mẫu ($500/1.000$). Các tầng khác cũng tương tự. Tức là do tỉ lệ lấy mẫu bằng nhau giữa các tầng nên cơ cấu của mẫu giống hệt như cơ cấu của tổng thể theo các tầng.

Ở đây, ngoài kết quả chung, nếu chúng ta muốn có một kết quả đáng tin cậy cho từng tầng thì có thể nhận thấy, số lượng mẫu lấy ra từ hệ hoàn chỉnh Đại học, văn bằng hai và cao học khá ít. Cho nên có thể áp dụng một cách phân bổ đơn vị mẫu vào các tầng khác, đó là phân bổ đều 1.000 mẫu/5 tầng = 200 đơn vị mẫu ở mỗi tầng (Hoặc là một cách phân bổ khác trong đó tăng số lượng đơn vị mẫu lên ở những tầng có quy mô nhỏ và giảm bớt số lượng mẫu lấy ra ở những tầng có quy mô lớn). Trong trường hợp này dữ liệu khảo sát của từng tầng ở cỡ mẫu 200 sẽ cho kết quả tốt hơn so với lúc phân bổ mẫu theo tỉ lệ mẫu ở các tầng có ít đơn vị mẫu. Tuy nhiên, khi cần tính toán chung cho toàn bộ mẫu thì trong tính toán cần gia trọng các tầng để cơ cấu của mẫu giống với cơ cấu của tổng thể nghiên cứu, để kết quả chung của mẫu đại diện được cho toàn bộ tổng thể.

Trong ví dụ này thì tầng đầu tiên là sinh viên chính quy chiếm đến 50% của tổng thể nhưng mẫu lấy ra chỉ có 200, chiếm tỉ lệ tương đương với các tầng khác là 20% trong mẫu. Do đó cần gia trọng tầng này với trọng số là tỉ trọng của tầng trong tổng thể chia cho tỉ trọng của tầng trong mẫu lấy ra, cụ thể là $50\% / 20\% = 2,5$.

Lúc này sau khi gia trọng với 2,5 thì số mẫu của nhóm sinh viên hệ chính quy khi tham gia kết quả tính toán chung cho toàn bộ mẫu sẽ tương đương với 500 mẫu ($200 \times 2,5 = 500$). Đối với tầng sinh viên cao học, tương tự tính được trọng số của tầng này là $5\% / 20\% = 0,25$. Sau khi gia trọng với 0,25 thì số mẫu của nhóm sinh viên cao học khi tham gia kết quả tính toán chung cho toàn bộ mẫu sẽ tương đương với 50 mẫu ($200 \times 0,25 = 50$). Một cách tổng quát thì trọng số của tầng thứ i được xác định bằng công thức:

$$\frac{N_i}{N} \times \frac{n}{n_i}$$

Với công thức trên thì có thể không cần phân bổ mẫu đều, chúng ta có thể “diều tiết” ở mức độ vừa phải số lượng đơn vị lấy mẫu từ những tầng quá lớn sang những tầng khá nhỏ để số lượng mẫu lấy ra từ tầng nhỏ không quá ít và kết quả của riêng tầng này có ý nghĩa. Lúc đó chúng ta sẽ dùng đến công thức tính trọng số tổng quát như trên khi tính toán kết quả chung.

Ví dụ: Sau khi phân bổ mẫu vào mỗi tầng là 200 sinh viên, tiến hành khảo sát bằng bảng câu hỏi do sinh viên tự trả lời. Từ kết quả tính toán mức độ hài lòng của sinh viên ở từng tầng, ta có thể tính mức độ hài lòng chung của toàn bộ mẫu nghiên cứu được gia trọng như sau:

Hệ đào tạo / cấp đào tạo	Trọng số	Mức độ hài lòng của sinh viên (tính từ quy mô mẫu của mỗi tầng là 200 sinh viên)
Cử nhân hệ chính quy	2,50	4,4
Cử nhân hệ hoàn chỉnh ĐH	0,50	3,8
Cử nhân hệ văn bằng thứ hai	0,50	3,6
Cử nhân hệ tại chức	1,25	4,0
Cao học	0,25	4,1
Cộng	5,00	

Mức độ hài lòng chung của sinh viên sẽ là:

$$\frac{(4,4 \times 2,5) + (3,8 \times 0,5) + (3,6 \times 0,5) + (4 \times 1,25) + (4,1 \times 0,25)}{2,5 + 0,5 + 0,5 + 1,25 + 0,25} = \frac{20,725}{5} = 4,145$$

Nếu chúng ta không gia trọng các kết quả của từng tầng, do số lượng đơn vị mẫu lấy ra ở từng tầng đều bằng nhau nên kết quả chung sẽ bằng với việc lấy 5 mức độ hài lòng ở 5 tầng khác nhau cộng lại chia cho 5, kết quả sẽ không phản ánh đúng đắn tình hình chung (vấn đề trung bình cộng

có trọng số sẽ được đề cập chi tiết trong Chương 4 Tóm tắt dữ liệu bằng các đại lượng thống kê mô tả).

2.4.2 Kỹ thuật lấy mẫu phi xác suất (non-probability sampling)

Trong thực tế nhiều khi chúng ta không có điều kiện về thời gian, thông tin (số lượng đơn vị tổng thể, cơ cấu tổng thể và khung lấy mẫu) và chi phí để thực hiện lấy mẫu ngẫu nhiên. Lúc đó chúng ta có thể sử dụng lấy mẫu phi xác suất (hay gọi là mẫu phi ngẫu nhiên). Mẫu phi xác suất không đại diện để ước lượng cho toàn bộ tổng thể, nhưng được chấp nhận trong nghiên cứu khám phá và trong kiểm định giả thuyết.

2.4.2.1 Lấy mẫu thuận tiện (convenient sampling)

Lấy mẫu thuận tiện được sử dụng trong nghiên cứu khám phá, để có cảm nhận về “điều gì đang diễn ra ở thực tế” và để kiểm tra trước bản câu hỏi nhằm bảo đảm là các đặc điểm cần thu thập dữ liệu trong bảng câu hỏi rõ ràng và không gây lo lắng cho người trả lời. Mẫu thuận tiện còn được dùng khi bạn muốn có một ước lượng sơ bộ về kết quả bạn quan tâm mà không muốn mất nhiều thời gian và chi phí.

Bạn có thể lấy mẫu thuận tiện bằng cách đến những nơi mà bạn có nhiều khả năng gặp được đối tượng mà bạn muốn khai thác thông tin mà bạn cảm thấy tiện lợi. Tuy nhiên, điều này không có nghĩa là bạn có quyền lấy mẫu tùy tiện (hay tùy hứng) hay không theo một nguyên tắc nào cả. Bạn cần suy nghĩ kỹ về thời gian, địa điểm hay hoàn cảnh mà bạn sẽ gặp đối tượng và thu thập dữ liệu ở đó.

Ví dụ nếu bạn hỏi những sinh viên đang trong thư viện là họ cảm thấy như thế nào về một số vấn đề đang được sinh viên quan tâm và tranh luận tại các trường Đại học hiện nay, bạn có thể thu được nhiều câu trả lời phong phú hơn là nếu bạn hỏi các sinh viên đang chơi bài tại quán cafe. Tương tự nếu bạn quan tâm đến việc du lịch hay giải trí của những phụ nữ ở tầng lớp trung lưu ở đô thị lớn, nếu bạn chọn chợ là nơi tiếp xúc thì có thể dễ dàng thấy rằng đó là nơi không phù hợp, vì phụ nữ ở tầng lớp trung lưu này sẽ rất ít khi đi chợ. Trong trường hợp này thì siêu thị, trung tâm mua sắm, nơi chăm sóc tóc, da, các câu lạc bộ sẽ phù hợp hơn vì bạn sẽ dễ gặp các đối tượng mà bạn muốn nghiên cứu ở đó.

2.4.2.2 Lấy mẫu định mức (quota sampling)

Trong lấy mẫu định mức, bạn sẽ quyết định các tổng thể con (tương tự như các tầng lớp trong lấy mẫu phân tầng) cần quan tâm và tỷ lệ của tổng thể con này trong mẫu của bạn lấy ra. Nếu định lấy một mẫu 400 người lớn tại một thành phố, người nghiên cứu có thể quyết định rằng, vì giới

tính là một biến độc lập có ảnh hưởng, và vì phụ nữ tạo thành một nửa của tổng thể, thì một nửa mẫu phải là phụ nữ và một nửa là nam giới. Hơn nữa, bạn lại quyết định rằng một nửa của mỗi phân giới tính phải có tuổi trên 40 và nửa còn lại trẻ hơn; một nửa là lao động tự do và một nửa là làm công ăn lương ...

Lấy mẫu định mức tương tự lấy mẫu xác suất phân tầng ở chỗ đầu tiên người nghiên cứu phải phân chia tổng thể nghiên cứu thành các tầng (tổng thể con). Nhưng điểm khác biệt cơ bản là trong từng tổng thể con những người phỏng vấn được chọn mẫu tại hiện trường theo cách thuận tiện hay phán đoán, trong khi trong mỗi tầng của chọn mẫu phân tầng thì các đơn vị mẫu được chọn ra theo kiểu xác suất (ngẫu nhiên đơn giản hay hệ thống). Một kinh nghiệm quan trọng là cần huấn luyện những nhân viên phỏng vấn để chọn mẫu không bị chêch khi đi lấy cho đủ các đối tượng theo các định mức, đó là, không chọn những người trả lời khá giống chính họ, mà chọn những người trả lời thật sự đại diện cho phạm vi của các biến trong tổng thể.

Nếu quyết định lấy mẫu định mức, hãy cẩn thận rằng không nên chỉ chọn những người mà bạn thích phỏng vấn và tránh những người mà bạn cảm thấy khó chịu hay cảm thấy họ bất hợp tác. Để tránh phỏng vấn những người khó tiếp xúc (những người bận rộn hầu như không có mặt tại nhà). Đặc biệt cẩn thận không nên chỉ chọn những người rất thích được phỏng vấn.

2.4.2.3 Lấy mẫu phán đoán (Judgement sampling)

Trong lấy mẫu phán đoán, bạn chính là người quyết định sự thích hợp các các đối tượng để mời họ tham gia vào mẫu khảo sát. Tuy nhiên vấn đề nằm ở chỗ chính phỏng vấn viên là người trực tiếp phán đoán sự thích hợp của các đối tượng để mời họ. Do đó tính đại diện của mẫu khảo sát thực tế sẽ phụ thuộc nhiều vào kiến thức và kinh nghiệm không những của người nghiên cứu điều tra, mà còn phụ thuộc vào kiến thức và kinh nghiệm của những người đi thu thập dữ liệu trực tiếp.

2.5 DỮ LIỆU ĐỊNH TÍNH VÀ DỮ LIỆU ĐỊNH LƯỢNG

Trước khi thu thập dữ liệu, bạn cũng cần phải phân biệt rõ tính chất của dữ liệu. Có hai loại là dữ liệu định tính và dữ liệu định lượng. Dữ liệu định tính phản ánh tính chất, sự hơn kém của các đối tượng nghiên cứu, ví dụ như giới tính (sinh viên đi làm thêm nam nhiều hay nữ nhiều). Dữ liệu định lượng phản ánh mức độ hay mức độ hơn kém, ví dụ như thời gian làm thêm của sinh viên bao nhiêu giờ một ngày hay tuần. Dữ liệu định

tính thu thập bằng thang đo định danh hay thứ bậc, dữ liệu định lượng thu thập bằng thang đo khoảng cách hay tỷ lệ (các thang đo này đã được trình bày trong phần cuối của Chương 1).

Dữ liệu định tính dễ thu thập hơn dữ liệu định lượng, nhưng dữ liệu định lượng thường cung cấp nhiều thông tin hơn và dễ áp dụng nhiều phương pháp phân tích hơn. Khi thực hiện nghiên cứu, trong giai đoạn lập kế hoạch nghiên cứu và thu thập dữ liệu, người nghiên cứu cần xác định trước các phương pháp phân tích cần sử dụng để phục vụ cho mục tiêu nghiên cứu của mình, và từ đó xác định loại dữ liệu cần thu thập, có nghĩa là, xác định thang đo phù hợp cần sử dụng trong khi thiết kế biểu mẫu hay bảng câu hỏi dùng để thu thập dữ liệu bạn mong muốn.

Ví dụ, chúng ta muốn nghiên cứu ảnh hưởng của việc đi làm thêm đối với kết quả học tập của sinh viên. Các dữ liệu thu thập có thể dưới dạng định tính hay định lượng. Chẳng hạn như dữ liệu sinh viên có đi làm thêm hay không (có và không) là dữ liệu định tính, kết quả học tập của sinh viên có thể là định tính (xếp loại học tập: giỏi, khá, trung bình) hay định lượng (điểm trung bình học tập). Nếu chúng ta không có điều kiện khảo sát và thu thập dữ liệu trên tất cả các sinh viên thuộc tổng thể nghiên cứu (ví dụ như sinh viên của trường ĐH Kinh Tế), mà chỉ có thể khảo sát và thu thập dữ liệu trên một mẫu (ví dụ như 200 sinh viên), thì để rút ra kết luận chung cho toàn bộ sinh viên, chúng ta phải sử dụng những kiểm định thống kê phù hợp.

Nếu nghiên cứu ảnh hưởng của việc có đi làm thêm (dữ liệu định tính) đến kết quả học tập của sinh viên (dữ liệu định tính) thì chúng ta có thể sử dụng 1 kiểm định phi tham số là kiểm định Chi bình phương. Nhưng nếu dữ liệu về kết quả học tập của sinh viên là định lượng (điểm trung bình học tập) thì chúng ta dùng kiểm định t đối với hai trung bình. Các phương pháp vừa nêu sẽ được trình bày chi tiết trong các chương sau. Nếu muốn nghiên cứu thời gian làm thêm nhiều ít có ảnh hưởng đến kết quả học tập không, chúng ta cũng có thể sử dụng kiểm định phi tham số, phân tích phương sai, mô hình hồi quy. Sử dụng công cụ nào tùy thuộc vào tính chất của dữ liệu ta đã thu thập là định tính hay định lượng (Bảng 2.1)

Bảng 2.1: Loại dữ liệu và loại kiểm định thống kê sử dụng khi phân tích

Thời gian làm thêm	Kết quả học tập	Loại kiểm định
Định tính Dưới 6 giờ/tuần 6-12 giờ/tuần trên 12 giờ/tuần	Định tính Trung bình Khá Giỏi	Phi tham số
Định tính Dưới 6 giờ/tuần 6-12 giờ/tuần trên 12 giờ/tuần	Định lượng Điểm trung bình học tập _____	Phân tích phương sai 1 yếu tố
Định lượng Số giờ làm thêm: giờ/tuần	Định lượng Điểm trung bình học tập _____	Hồi quy và kiểm định F

cuu duong than cong. com

cuu duong than cong. com

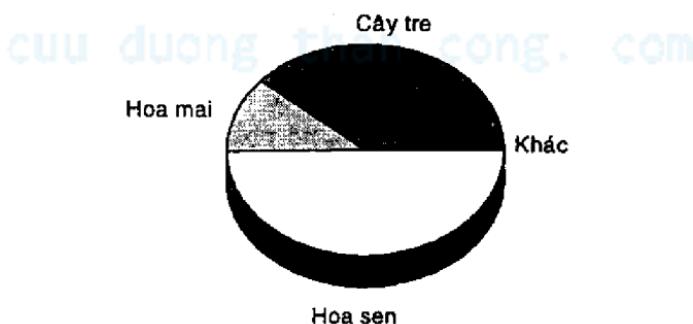
CHƯƠNG 3

TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU BẰNG BẢNG VÀ ĐỒ THỊ

Trong khoảng tháng 7 và 8 năm 2006 Tuổi trẻ Online có làm một khảo sát nhỏ bình chọn của bạn đọc về Quốc hoa của Việt Nam trong 4 lựa chọn là hoa mai, hoa sen, cây tre và đê xuất khác. Sau hơn 2 tháng bình chọn, tổng cộng có 135.097 ý kiến tham gia. Trong đó, 67.008 ý kiến (chiếm 49,6%) chọn hoa sen làm quốc hoa, 47.288 ý kiến, chiếm 35% chọn cây tre, 15.850 ý kiến (11,73%) chọn hoa mai và 4.951 (3,66%) chọn đê xuất khác.

Hình ảnh bạn thấy dưới đây là một cách khác để thể hiện kết quả của cuộc bình chọn mà không phải dùng đến số liệu thống kê. Mặc dù không có con số nhưng đồ thị này vẫn cho bạn cảm nhận rất nhanh rằng Hoa sen đang dẫn đầu trong cuộc đua vào vị trí Quốc hoa Việt Nam qua bình chọn của người đọc Tuổi Trẻ online, kế đến là Cây tre, vị trí thứ 3 là của Hoa mai.

Hình 3.1



Nguồn: TuoiTre online

Đồ thị trong Hình 3.1, như phần sau của chương này sẽ đề cập, được gọi tên là đồ thị Hình tròn, là một trong nhiều phương pháp để tóm lược và trình bày dữ liệu thống kê rất hiệu quả. Bạn đọc gặp các phương tiện tóm lược dữ liệu này rất thường xuyên trên báo chí. Chúng được kết hợp với các bài báo nhằm giúp độc giả nhận thức vấn đề nhanh và rõ hơn. Biểu đồ và đồ thị còn được dùng nhiều trong quảng cáo, tuyên truyền để gia tăng hiệu quả của thông điệp. Chúng ta sẽ lần lượt nghiên cứu các phương pháp tóm lược và trình bày dữ liệu thống kê không chỉ dưới hình thức đồ thị mà còn cả bảng biểu trong nội dung Chương 3 này.

3.1 TÓM LƯỢC VÀ TRÌNH BÀY DỮ LIỆU BẰNG BẢNG TẦN SỐ 1

Bảng tần số là một bảng tổng hợp các biểu hiện có thể có của đặc điểm quan sát, hoặc các khoảng giá trị mà trong phạm vi đó dữ liệu (định lượng) có thể rơi vào và số quan sát tương ứng với mỗi biểu hiện hoặc khoảng giá trị dữ liệu, ngoài ra người ta còn tính xem so với tổng số quan sát thì số đơn vị thuộc cùng biểu hiện hoặc khoảng giá trị này chiếm bao nhiêu phần trăm. Ở dạng cơ bản nhất, bảng tần số thường bao gồm 3 cột, cột đầu tiên mô tả các biểu hiện hoặc các giá trị hay khoảng giá trị được xác định cho dữ liệu, cột thứ 2 mô tả tần số tương ứng với các biểu hiện hay giá trị, và cột thứ 3 là các tần suất (tỷ lệ %).

Bảng tần số có thể được tính thêm một số cột dữ liệu nữa phục vụ cho các mục đích tìm hiểu sâu hơn, khi tiến hành lập bảng tần số cần phải xác định rõ các tình huống lập bảng tần số cho dữ liệu định tính hay định lượng, nếu là dữ liệu định lượng thì là loại dữ liệu rời rạc hay liên tục, ít biểu hiện hay nhiều biểu hiện.. để có những giải pháp phù hợp.

3.1.1 Cách lập bảng tần số cho dữ liệu định tính

Đối với loại dữ liệu định tính như giới tính, vùng địa lý, ngành học.. chúng ta sẽ lập bảng tần số với những thông tin như sau:

- Cột thứ nhất liệt kê tất cả các biểu hiện có thể có của đối tượng theo đặc điểm ta đang muốn lập bảng tần số để tóm tắt dữ liệu. Nếu đặc điểm quan tâm là giới tính rõ ràng nó phải có hai biểu hiện là Nam hoặc Nữ, nếu đặc điểm quan tâm là khu vực địa lý nó có thể được phân chia thành: Miền Bắc, Miền Trung và Tây nguyên, Miền Nam; hoặc chi tiết hơn nữa thành: Đồng Bằng sông Hồng, Đông Bắc Bộ, Tây Bắc Bộ, Bắc Trung Bộ, Duyên hải miền Trung, Tây nguyên, Đông Nam Bộ, Đồng bằng sông Cửu Long.
- Cột thứ hai là cột tần số được điền số liệu bằng cách đếm xem có bao nhiêu quan sát có cùng 1 biểu hiện. Tổng của cột tần số phải bằng đúng số quan sát của tập dữ liệu.
- Cột thứ ba là cột tần suất. Các tần suất được tính bằng cách lấy các tần số chia cho tổng số quan sát của tập dữ liệu, đem kết quả nhân cho 100%, rồi ghi vào cột tần suất tại vị trí tương ứng cùng hàng, nhằm so sánh xem so với tổng số quan sát thì số đơn vị có cùng biểu hiện chiếm bao nhiêu %.

Công thức tính:

¹ Frequency Distribution

Tần suất của biểu hiện thứ i được tính bằng = $\frac{f_i}{n} \times 100\%$

Trong đó: f_i là tần số của biểu hiện thứ i

n là tổng số quan sát của tập dữ liệu, $n = \sum_{i=1}^k f_i$

k là số biểu hiện của đặc điểm ta quan tâm

Tổng của cột tần suất phải bằng đúng 100%, nếu các số liệu tần suất là số lẻ, chúng ta phải làm tròn phù hợp sao cho tổng này bằng đúng 100%.

Chúng ta xem xét ví dụ về bảng tần số cho dữ liệu định tính trong một tình huống đơn giản như sau: người ta khảo sát 1.037 hộ gia đình tại một địa phương về nhiều vấn đề trong đó có đặc điểm “công việc của chủ hộ”. Các chủ hộ có thể có các hoạt động kinh tế như làm lao động tay chân, buôn bán, viên chức Nhà nước, công nhân ... tức là thuộc nhóm có hoạt động kinh tế, họ cũng có thể có thu nhập mà không hoạt động kinh tế như hưởng lương hưu, nhận trợ cấp từ họ hàng hay cho thuê nhà..., và cuối cùng là những người không có việc làm. Từ số liệu thực tế thu được người ta chia ra 3 loại biểu hiện như sau trong tiêu chí công việc của chủ hộ: Có hoạt động kinh tế, không hoạt động kinh tế và cuối cùng là không có việc làm. Sau đó người ta đếm số chủ hộ có cùng loại biểu hiện ở đặc điểm “công việc của chủ hộ” để điền vào cột tần số theo từng biểu hiện tương ứng và sau cùng là tính toán cột tần suất bằng cách lấy số lượng chủ hộ thuộc từng biểu hiện chia cho tổng số 1.037 chủ hộ, rồi lấy kết quả nhân với 100%. Ta được bảng tần số sau đây.

$$= \frac{658}{1037} \times 100\%$$

Bảng 3.1 Công việc của chủ hộ

Công việc của chủ hộ	Tần số (người)	Tần suất (%)
Có hoạt động kinh tế	658	63,45
Không hoạt động kinh tế	47	4,53
Không có việc làm	332	32,02
Tổng	1037	100

3.1.2 Cách lập bảng tần số cho dữ liệu định lượng

Khi lập bảng tần số cho dữ liệu định lượng chúng ta phải phân biệt ra các tình huống cụ thể như sau.

3.1.2.1 Dữ liệu định lượng mà đặc điểm quan tâm có ít biểu hiện

Nếu đặc điểm bạn quan tâm là số con của một cặp vợ chồng trẻ thành thị, bạn có dữ liệu định lượng nhưng các giá trị có thể có của nó rất ít, bạn có thể bao quát được trên 5 đầu ngón tay, bao gồm: 0, 1, 2, hiếm gặp 3 và có lẽ tối đa chỉ đến 4. Hoặc nếu bạn là nhà quản trị của một tờ báo ra hàng ngày, bạn quan tâm đến số lượng tờ báo của bạn mà mỗi độc giả đọc trong tuần, có thể có độc giả không đọc báo của bạn, tức là số tờ báo đọc trong tuần là 0; có thể có độc giả đọc báo của bạn hàng ngày, tức số tờ báo đọc trong tuần là 7. Và giữa 0 với 7 là các tình huống còn lại bao gồm từ 1 đến 6. Rõ ràng bạn “giám sát” được số biểu hiện, như vậy bạn đang có dữ liệu định lượng thuộc loại đặc điểm quan tâm có ít biểu hiện.

Cách lập bảng tần số cho dữ liệu định lượng của đặc điểm quan tâm có ít biểu hiện giống như cách lập bảng tần số cho dữ liệu định tính vừa khảo sát trên, lúc này mỗi giá trị xem như một biểu hiện. Đặc biệt hơn là lúc này tại cột đầu tiên là cột đặc điểm thống kê chúng ta có thể xác định đơn vị tính cho đặc điểm ta quan tâm.

Ví dụ Ban biên tập của một tờ báo ngày A tiến hành khảo sát 200 người về số tờ báo A mà họ đã đọc trong tuần. Như đã phân tích ở trên, đặc điểm “số tờ báo đọc trong tuần” đã thu thập cho chúng ta dữ liệu định lượng với 8 biểu hiện, ta lập bảng tần số mà cột đầu tiên liệt kê 8 biểu hiện này, cột tần số mô tả số người tương ứng với từng biểu hiện và cột tần suất cũng được tính theo cách thức đã biết.

Bảng 3.2 Số tờ báo đọc trong tuần

Số báo đọc (tờ/tuần)	Tần số (người)	Tần suất (%)
0	44	22
1	24	12
2	18	9
3	16	8
4	20	10
5	22	11
6	26	13
7	30	15
Tổng	200	100

3.1.2.2 Dữ liệu định lượng của đặc điểm quan tâm có nhiều biểu hiện

Trong tình huống đặc điểm ta đang thống kê có quá nhiều biểu hiện thì việc liệt kê từng biểu hiện một như ở cách trên không còn phù hợp vì nếu

làm thế bảng tần số sẽ rất dài, mất đi tác dụng tóm lược thông tin, không tạo thuận lợi cho người quan sát thông tin trong việc nhận thức vấn đề. Lúc này chúng ta phải tiến hành thao tác đầu tiên là phân tổ rồi sau đó mới lập bảng tần số trên cơ sở dữ liệu đã phân tổ này.

Phân tổ dữ liệu là căn cứ vào một số đặc điểm nào đó để sắp xếp các đơn vị quan sát vào các tổ, nhóm có tính chất khác nhau, hay nói một cách khác là phân bổ các đơn vị mẫu nghiên cứu vào các tổ có tính chất khác nhau.

Lấy một ví dụ minh họa như sau, bạn khảo sát 1129 dân nhập cư vào tp HCM trong độ tuổi lao động từ 15 đến 60 tuổi, như vậy bạn sẽ gặp 46 biểu hiện tuổi khác nhau trong bộ dữ liệu này. Khi tạo bảng tần số, nếu bạn liệt kê hết 46 biểu hiện tuổi ra bạn sẽ tạo nên một bảng tần số dài tới 48 hàng (bao gồm cả hàng tiêu đề và hàng tổng cộng) như Bảng 3.3 dưới đây, thay vào đó bạn phải nghĩ đến một biện pháp hợp lý hơn, đó là bạn sẽ gom dữ liệu lại theo một quy tắc nào đó, chẳng hạn gom theo đặc trưng là có bao nhiêu người trong độ tuổi 15 đến 20, bao nhiêu người trong phạm vi từ 21 đến 30, từ 31 đến 40... Nếu xử lý theo kiểu này bạn sẽ chỉ còn 5 biểu hiện cho đặc điểm độ tuổi và sau đó có thể tạo được một bảng tần số gọn gàng, khả năng biểu đạt thông tin cũng súc tích hơn.

Bảng 3.3 Tuổi của các đối tượng trong mẫu nghiên cứu

Tuổi	Tần số (người)	Tần suất (%)	Tần suất tích lũy (%)
15	21	1,9	1,9
16	19	1,7	3,5
17	21	1,9	5,4
18	18	1,6	7,0
19	22	1,9	8,9
20	45	4,0	12,9
21	36	3,2	16,1
22	41	3,6	19,8
23	39	3,5	23,2
24	33	2,9	26,1
25	35	3,1	29,2
26	48	4,3	33,5
27	51	4,5	38,0
28	33	2,9	40,9
29	42	3,7	44,6

30	52	4,6	49,2
31	39	3,5	52,7
32	33	2,9	55,6
33	38	3,4	59,0
34	31	2,7	61,7
35	32	2,8	64,6
36	28	2,5	67,1
37	27	2,4	69,4
38	24	2,1	71,6
39	20	1,8	73,3
40	21	1,9	75,2
41	22	1,9	77,1
42	24	2,1	79,3
43	26	2,3	81,6
44	19	1,7	83,3
45	17	1,5	84,8
46	12	1,1	85,8
47	23	2,0	87,9
48	19	1,7	89,5
49	19	1,7	91,2
50	6	,5	91,8
51	13	1,2	92,9
52	10	,9	93,8
53	18	1,6	95,4
54	5	,4	95,8
55	12	1,1	96,9
56	9	,8	97,7
57	4	,4	98,1
58	10	,9	98,9
59	5	,4	99,4
60	7	,6	100,0
Tổng	1129	100	

Nguồn: Nghiên cứu về “Tác động của chính sách cư trú đến việc giảm nghèo đô thị”, Trung tâm nghiên cứu xã hội học, Viện Khoa Học Xã Hội vùng Nam Bộ, năm 2005.

Sau khi gom dữ liệu theo qui tắc đã mô tả chúng ta có bảng tần số thứ hai gọn hơn rất nhiều.

Bảng 3.4 Độ tuổi của các đối tượng trong mẫu nghiên cứu

Độ tuổi (tuổi)	Tần số (người)	Tần suất (%)	Tần suất tích lũy (%)
Từ 15 đến 20	146	12,9	12,9
Từ 21 đến 30	410	36,3	49,2
Từ 31 đến 40	293	26,0	75,2
Từ 41 đến 50	187	16,6	91,8
Từ 51 đến 60	93	8,2	100
Tổng	1129	100	

Kiểu phân tổ vừa mô tả trong ví dụ trên đây là kiểu phân tổ theo kinh nghiệm. Trong thực tế khi gặp dạng dữ liệu định lượng nhiều biểu hiện người ta có quy tắc để tiến hành phân tổ. Sau đây chúng ta sẽ đi vào nghiên cứu phương pháp phân tổ dữ liệu.

Phương pháp phân tổ dữ liệu

Thực ra khi phân tổ dữ liệu, tùy theo mục đích thể hiện dữ liệu cũng như đặc điểm phân bố đều đặn hay không đều đặn của dữ liệu mà có thể tiến hành phân tổ đều hoặc phân tổ không đều. Khái niệm đều hoặc không đều liên quan đến khái niệm khoảng cách tổ. Khi bạn phân tổ thì mỗi tổ sẽ có giới hạn dưới và giới hạn trên, giới hạn dưới là trị số nhỏ nhất của tổ và giới hạn trên là trị số lớn nhất của tổ. Ví dụ với tổ từ 21 đến 30 tuổi, bao gồm những người trong phạm vi 21 đến 30 tuổi thì giới hạn dưới lúc này là 21 tuổi và giới hạn trên là 30 tuổi. Chênh lệch giữa giới hạn dưới và giới hạn trên là trị số khoảng cách tổ. Nếu tất cả các tổ trong bảng tần số đều có khoảng cách tổ bằng nhau thì đó là phân tổ đều và ngược lại, như ví dụ ở Bảng 3.4 là một tình huống phân tổ không đều vì khoảng cách tổ của tổ đầu tiên không bằng khoảng cách tổ của các tổ còn lại. Trong nội dung dưới đây chúng ta sẽ tập trung tìm hiểu phương pháp phân tổ đều.

Một số điều kiện phải tuân thủ khi tiến hành phân tổ:

- Các tổ không được trùng nhau, để cho một quan sát bất kì chỉ thuộc về một tổ
- Tất cả các tổ được phân chia phải bảo đảm bao quát hết tất cả các giá trị hiện có của tập dữ liệu
- Tránh không để tổ rỗng do không có quan sát nào thuộc về tổ đó.

Các bước của thủ tục phân tổ đều:

- Xác định số tổ cần chia k: không có một con số qui định chính xác về số tổ cần chia là bao nhiêu, nhưng theo kinh nghiệm người ta thấy nên

chia trong khoảng từ trên 5 tổ đến dưới 15 tổ. Một công thức có tính tham khảo sau đây giúp chúng ta xác định được số tổ cần chia phù hợp cho từng bộ dữ liệu cụ thể

$$\text{Số tổ cần chia } k = (2 * n)^{1/3} \text{ với } n \text{ là số quan sát của tập dữ liệu. } ^1$$

Chú ý là kết quả tính được nếu là số lẻ thì phải được làm tròn, và vì công thức này có tính tham khảo nên chúng ta có thể làm tròn lên hoặc làm tròn xuống đều được nhưng phải đảm bảo các điều kiện khi tiến hành phân tổ vừa nêu trên. Nói chung là nếu tập dữ liệu nhỏ, ít quan sát thì ta nên tạo ra ít tổ hơn và nếu tập dữ liệu lớn với nhiều quan sát thì ta tạo ra nhiều tổ hơn. Nếu dùng ít tổ quá mức cần thiết có xu hướng làm dữ liệu bị “nén quá chặt” làm mất đi nhiều thông tin.

- Xác định trị số khoảng cách tổ h : căn cứ trên số tổ định chia người ta xác định trị số khoảng cách tổ theo công thức như sau

$$h = \frac{(X_{\max} - X_{\min})}{k}$$

X_{\max} là giá trị lớn nhất và X_{\min} là giá trị nhỏ nhất của tập dữ liệu định phân tổ; còn k là số tổ định chia. Giá trị h tính được nếu là một số lẻ cũng thường được xem xét làm tròn để dễ theo dõi các khoảng cách tổ hơn.

- Xác định giới hạn dưới và giới hạn trên của các tổ: Khi xác định giới hạn dưới của tổ đầu tiên cần đảm bảo giá trị của nó tối đa là bằng, còn không thì phải bé hơn giá trị X_{\min} để có thể bao quát được X_{\min} trong tổ đầu tiên. Đảm bảo giới hạn trên của tổ cuối cùng thì phải lớn hơn X_{\max} để bao gồm được X_{\max} trong tổ cuối cùng. Với các tổ liên tục nhau, giá trị cận trên của tổ trước vừa trùng giá trị cận dưới của tổ sau liền kề. Có thể tham khảo công thức sau

Tổ thứ nhất : $(X_{\min}, X_{\min} + h)$

Tổ thứ hai : $(X_{\min} + h; X_{\min} + h + h) = (X_{\min} + h; X_{\min} + 2h)$

Tổ thứ ba : $(X_{\min} + 2h; X_{\min} + 2h + h) = (X_{\min} + 2h; X_{\min} + 3h)$

...

Tuy nhiên trong thực tế khi xác định các cận trên và dưới của các tổ người ta có thể xử lý linh động hơn để đảm bảo tính khoa học và mỹ thuật.

¹ Có thể tham khảo một công thức khác của H.A Stugres là $k = 1 + 3,3 * \log(n)$ với $\log(n)$ là logarit cơ số 10 của n là cỡ mẫu quan sát hay số dữ liệu thu thập được.

- Phân chia các quan sát vào các tổ: ta điểm qua các quan sát, quan sát có giá trị phù hợp với tổ nào thì ta xếp nó vào tổ đó, quy ước thông thường là khi gặp một quan sát có giá trị bằng đúng cận trên của một tổ thì ta xếp quan sát đó vào tổ kế tiếp. Điều này có nghĩa là chúng ta có thể thể hiện sự tồn tại của một quan sát có giá trị x_i trong tổ của nó theo lối viết toán học như sau:

Cận dưới $< x_i <$ Cận trên

Ví dụ chúng ta có dữ liệu của một mẫu điều tra nhỏ về tuổi của 30 sinh viên tại chức đang học năm 1 ngành Kế toán – Kiểm toán¹ như sau

28 23 30 24 19 21 39 22 22 31 37 33 20 30 35

21 26 27 25 29 27 21 25 28 26 29 29 22 32 27

Để tóm lược lại dữ liệu này chúng ta có thể sử dụng bảng tần số, và để bảng tần số đảm bảo tính khoa học chúng ta cần tiến hành phân tổ dữ liệu rồi mới lập bảng tần số. Trình tự phân tổ đều cho tập dữ liệu này đi qua các bước sau đây:

Xác định số tổ cần chia theo công thức:

$$k = (2 \times n)^{1/3} \text{ với } n = 30 \rightarrow (2 \times 30)^{1/3} = 3,9 \approx 4 \text{ tổ}$$

Xác định trị số khoảng cách tổ h theo công thức

$$h = \frac{(X_{\max} - X_{\min})}{k} \text{ với } X_{\max} = 39 ; X_{\min} = 19 \text{ và } k = 4$$

$$\rightarrow h = \frac{(39 - 19)}{4} = 5 \text{ tuổi}$$

Xác định giới hạn dưới và giới hạn trên của các tổ:

Với $h = 5$ ta xác định cận trên và dưới của các tổ lần lượt như sau :

$$\text{Tổ 1 : } 19 - 19 + 5 = 19 - 24$$

$$\text{Tổ 2 : } 24 - 24 + 5 = 24 - 29$$

$$\text{Tổ 3 : } 29 - 29 + 5 = 29 - 34$$

$$\text{Tổ 4 : } 34 - 34 + 5 = 34 - 39$$

Phân chia các quan sát vào các tổ, chú ý rằng tại tổ cuối cùng giá trị cận dưới bằng 39 đúng bằng giá trị X_{\max} là 39 tuổi, nếu như áp dụng quy tắc phân chia các quan sát theo kiểu nếu quan sát có giá trị bằng đúng cận trên của một tổ thì ta xếp quan sát đó vào tổ kế tiếp thì sinh viên có tuổi 39 đúng ra phải được xếp vào tổ kế tiếp sau tổ 34 – 39, nhưng ta không có tổ này, nên ta có thể giải quyết bằng cách mở tổ cuối cùng như bảng sau:

¹ Ví dụ này sẽ được sử dụng lại nhiều lần trong Chương 3 và Chương 4

Bảng 3.5

Độ tuổi (tuổi)	Số SV
19 - 24	9
24 - 29	10
29 - 34	8
34 trở lên	3
Tổng	30

Khi gấp các tính toán đối với tài liệu phân tách mờ người ta qui ước lấy khoảng cách của tổ mờ bằng với khoảng cách của tổ đứng gần nó nhất. Như vậy tổ thứ 4 trong ví dụ trên được xem như có $h = 5$. Ngoài ra người ta còn có thể mờ giới hạn dưới của tổ đầu tiên, hoặc mờ đồng thời cả giới hạn dưới của tổ đầu tiên với giới hạn trên của tổ sau cùng. Mục đích của phân tách mờ là để tổ đầu tiên và tổ cuối cùng chứa được các đơn vị có giá trị đột biến (lớn hoặc nhỏ bất thường so với các giá trị còn lại trong tập dữ liệu) và tránh việc hình thành quá nhiều tổ.

Từ cơ sở phân tách này bảng tần số được lập như sau

Bảng 3.6 Tuổi của 30 sinh viên tại chức chuyên ngành KTKT

Độ tuổi (tuổi)	Tần số (SV)	Tần suất (%)
19 - 24	9	30,00
24 - 29	10	33,33
29 - 34	8	26,67
34 trở lên	3	10,00
Tổng	30	100,00

Sau đây là một ví dụ khác để bạn đọc hình dung được tính linh hoạt trong phân tách.

Một doanh nghiệp kinh doanh thương mại có 28 cửa hàng bán lẻ trong thành phố. Số liệu về mức tiêu thụ (triệu đồng) của 28 cửa hàng được ghi chép và báo cáo hàng tuần, trong tuần thứ nhất của tháng 7 năm 2007 người ta thấy tình hình tiêu thụ của 28 cửa hàng này như sau:

57,8	57,5	52,4	50,9	50,2	53,3	50,1	43,3	42,5	41,7	41,1	45,8	47,2
56,9	47,5	38,8	50,3	37,6	38,9	52,3	49,2	47,5	47	49,6	46,2	49,8

Theo kinh nghiệm, người lập báo cáo của doanh nghiệp thấy rằng khi phân tách phục vụ cho việc lập bảng tần số nên chia ra 6 tổ là hợp lý nhất. Do đó ta quyết định chọn $k = 6$.

$$\text{Tính trị số khoảng cách tổ } h = \frac{(57,8 - 36,8)}{6} = 3,5 \approx 4$$

Khi xác định các tổ, nếu chọn cận dưới của tổ đầu tiên là giá trị $X_{\min} = 36,8$ thì tổ đầu tiên sẽ là $36,8 - 40,8$. Những số lẻ như vậy có thể khiến người đọc rối mắt và khó cảm nhận, chúng ta có thể lùi cận dưới của tổ đầu tiên xuống giá trị 36. Lúc này các tổ hình thành như sau

Bảng 3.7

Mức tiêu thụ (triệu đồng/tuần)	Số cửa hàng
36 – 40	4
40 – 44	4
44 – 48	7
48 – 52	7
52 - 56	3
56 - 60	3
Tổng	28

Ngoài ra đối với dữ liệu định lượng hoặc định tính dạng thứ bậc, khi thiết lập bảng tần số chúng ta có thể xây dựng thêm cột tần số tích lũy và cột tần suất tích lũy để cung cấp thêm thông tin cho người đọc. Trong bảng tần số với các tổ được sắp theo trật tự tăng dần về giá trị của tiêu thức thống kê đang quan tâm :

- Tần số tích lũy là số liệu tổng cộng thể hiện số quan sát có giá trị bé hơn (hoặc bằng) giới hạn trên của tổ mà nó nằm cùng hàng (khi tiêu thức thống kê quan tâm được sắp xếp theo trật tự tăng dần của giá trị). Dữ liệu của cột tần số tích lũy được tính bằng cách cộng dồn các tần số từ trên xuống cho đến đúng vị trí tương ứng với biểu hiện mà ta đang muốn tính tần số tích lũy.
- Tần suất tích lũy là số liệu tổng cộng thể hiện tỷ lệ % số quan sát có giá trị bé hơn (hoặc bằng) giới hạn trên của tổ mà nó nằm cùng hàng (khi tiêu thức thống kê quan tâm được sắp xếp theo trật tự tăng dần của giá trị). Dữ liệu của cột tần suất tích lũy được tính bằng cách cộng dồn các tần suất từ trên xuống cho đến đúng vị trí tương ứng với biểu hiện mà ta đang muốn tính tần suất tích lũy.

Để cảm nhận được giá trị thông tin mà cột tần số tích lũy và tần suất tích lũy cung cấp chúng ta hãy nghiên cứu ví dụ sau.

Ví dụ Người ta khảo sát 30 gia đình về số người trong hộ gia đình họ, số liệu thu được là dữ liệu định lượng dạng ít biểu hiện, bảng tần số được lập với 2 cột thông tin mới là tần số tích lũy và tần suất tích lũy.

Bảng 3.8 Số người trong hộ gia đình

Quy mô hộ gia đình (người)	Tần số (hộ)	Tần suất (%)	Tần số tích lũy (hộ)	Tần suất tích lũy (%)
4	1	3,33	1	3,33
5	4	13,33	5	16,66
6	3	10,00	8	26,66
7	3	10,00	11	36,66
8	7	23,33	18	59,99
9	6	20,00	24	79,99
10	3	10,00	27	89,99
11	1	3,33	28	93,32
12	2	6,68	30	100
Tổng	30	100		

Nếu muốn biết có bao nhiêu phần trăm số hộ được điều tra có dưới 10 thành viên, chúng ta tìm hàng mang giá trị 10 tại cột thứ nhất, tham chiếu sang cột tần suất tích lũy ta được giá trị 89,99. Như vậy gần 90% các hộ gia đình được hỏi có dưới 10 thành viên, tương ứng với con số tuyệt đối là 27 hộ gia đình.

Cột tần số và tần suất cho biết trong 30 hộ được điều tra, tình trạng phổ biến nhất là có từ 6 đến 7 thành viên trong gia đình, hiếm khi gặp hộ chỉ có 4 người hay hiếm khi gặp hộ có trên 10 người.

3.1.3 Lập bảng tần số bằng Excel

Chúng ta có thể sử dụng phần mềm Excel để lập bảng tần số với kết quả là hai cột số liệu tần số và tần suất tích lũy, tuy nhiên thực sự mà nói thì đây cũng vẫn là một công việc bán thủ công và đòi hỏi không ít công sức. Chúng ta sử dụng ví dụ về tuổi của 30 sinh viên tại chức để minh họa cách thức tiến hành

Nhập dữ liệu gốc vào cửa sổ làm việc của phần mềm Excel

Nhập các giá trị giới hạn trên của các tổ mà chúng ta dự định phân tổ, với ví dụ này các giá trị cận trên lần lượt là 24, 29, 34, 39. Bạn đọc xem cách nhập liệu ở Hình 3.2 a. Các giá trị gốc được nhập trên cột E còn các giá trị giới hạn nhập trên cột M của bảng tính Excel.

Hình 3.2 a

Rõ ràng là dù tiến hành bằng Excel nhưng để xác định được các giá trị này chúng ta phải tiến hành tính toán thủ công, như đã biết ở trên, và Excel chẳng qua chỉ làm giúp chúng ta một việc là đếm các quan sát để xác định cột tần số mà thôi. Chúng ta hoàn toàn phải tự chịu trách nhiệm về tính thẩm mỹ, khoa học và ý nghĩa của bảng tần số mà mình đã dùng Excel lập ra.

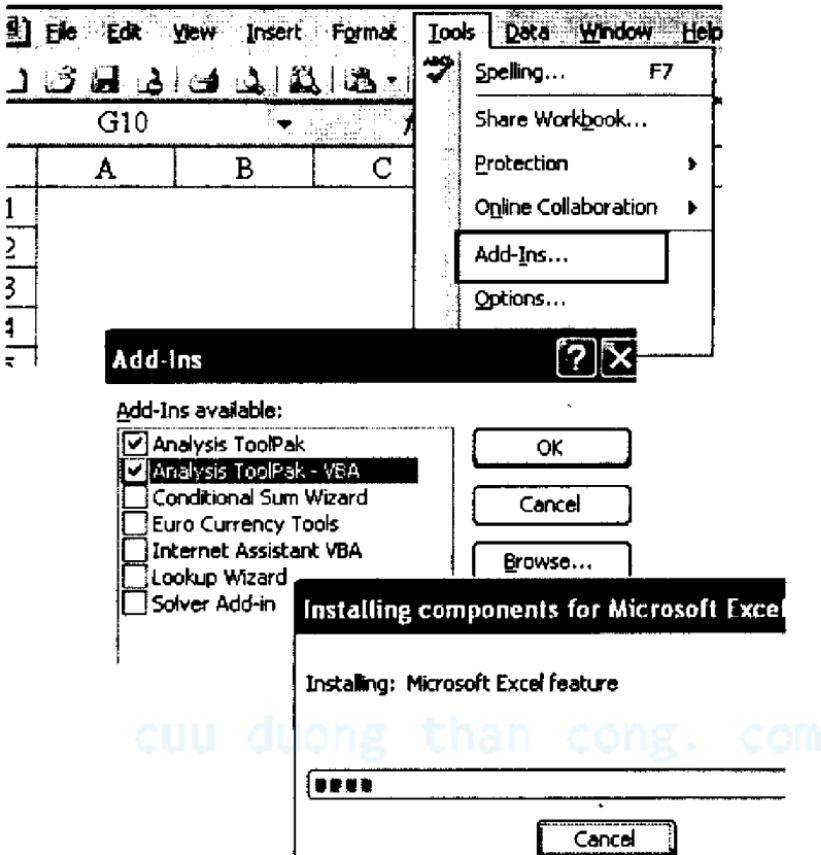
Vào menu Tool chọn lệnh Data Analysis

Nếu trên bạn không tìm thấy Data Analysis nằm sẵn trên menu Tool thì gọi chức năng Data Analysis ra như sau:

- Vào Tools/Add – Ins bạn mở ra cửa sổ Add-Ins
- Trong danh sách các chức năng liệt kê tại mục Add-Ins available bạn nhấp chọn 2 chức năng Analysis Toolpak và Analysis Toolpak VBA nằm ở đầu danh sách sau đó nhấp nút OK, trả lời các yêu cầu (nếu có) và chờ máy cài đặt.

cuuduongthancong.com

Hình 3.2 b



Bạn kiên nhẫn chờ vài phút để máy cài đặt (như hình trên đây) mọi việc sẽ ổn thỏa (với điều kiện Excel trên máy bạn được cài đặt bản đầy đủ - full), quay lại Tools bạn sẽ thấy có Data Analysis.

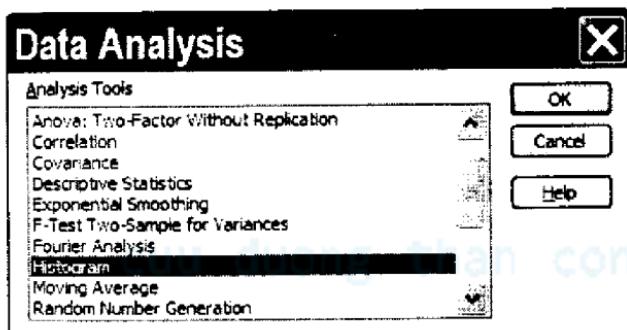
Bấm vào lệnh Data Analysis bạn sẽ mở được hộp thoại như Hình 3.3, trên hộp thoại này chọn sáng lựa chọn Histogram rồi nhấn nút OK để mở cửa sổ Histogram. Xem các khai báo trên cửa sổ Histogram như Hình 3.4

- Tại mục Input Range chúng ta nhập địa chỉ khu vực chứa dữ liệu đã nhập về tuổi của 30 sinh viên, chú ý đưa luôn cả địa chỉ hàng tiêu đề vào.
- Tại mục Bin Range đưa địa chỉ phạm vi chứa các giá trị các cận trên vào, nhớ đưa luôn địa chỉ hàng tiêu đề.
- Nhấn chọn nút Label để loại trừ hàng chứa tiêu đề ra khỏi các tính toán.

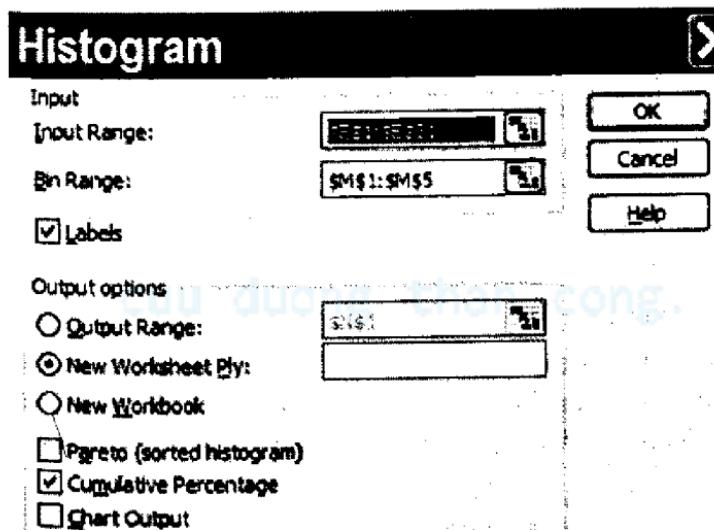
- Chọn Cumulative Percentage để tính tần suất tích lũy, nếu không Excel chỉ tính tần số.
- Nhấp nút OK ta được bảng kết quả sau, nó được đặt trên một Sheet mới nếu ta không xác định trước vị trí đặt kết quả trong nội dung Output Range

<i>Độ tuổi</i>	<i>Frequency</i>	<i>Cumulative %</i>
24	10	33.33%
29	12	73.33%
34	5	90.00%
39	3	100.00%
More	0	100.00%

Hình 3.3



Hình 3.4



Tiến hành các chỉnh sửa cần thiết trên bảng kết quả này, bao gồm hoàn chỉnh thể hiện của các tổ trong cột thứ nhất bằng cách nhập thêm cột trên

của tổ. Sửa chữ More thành chữ tổng cộng, sửa tiêu đề cột Frequency thành chữ tần số, và Cumulative % thành chữ Tần suất tích lũy. Quan trọng là nhớ kiểm tra xem tổng cột Frequency có bằng đúng giá trị n không. Đưa giá trị n vào cuối cột này. Rồi sau cùng xóa giá trị 100% tại cuối cột thứ 3.

Như bạn đọc thấy ở Bảng 3.9, sau khá nhiều thao tác chúng ta có được một bảng tần số chưa hoàn chỉnh cho lắm. Đây là tất cả những gì Excel giúp được bạn. Còn một chú ý nữa là Excel đếm các quan sát của các tổ theo quy tắc ngược với cách đếm thủ công thông thường, đó là khi gấp các quan sát có giá trị đúng bằng cận trên của một tổ, quan sát đó sẽ được Excel xếp vào chính tổ đó (chữ không phải tổ kế tiếp), do đó mà cột tần số do Excel cung cấp cho ví dụ này hơi khác kết quả ta tìm được lúc làm thủ công, do đó dĩ nhiên cột tần suất cũng khác chút ít.

Bảng 3.9

Độ tuổi	Tần số (SV)	Tần suất tích lũy %
19 - 24	10	33.33%
24 - 29	12	73.33%
29 - 34	5	90.00%
34 - 39	3	100.00%
Tổng	30	

Cuối cùng, nói về ưu điểm của bảng tần số đã được xử lý trên cơ sở phân tổ, đó là đặc trưng cơ bản của dữ liệu sẽ được bộc lộ rõ ràng và nhanh chóng cho người xem, Bảng 3.9 trên cho chúng ta biết ngay là phần lớn sinh viên tại chức ngành KT-KT có tuổi dưới 29, chiếm tới 73% số người được hỏi, tập trung nhiều nhất trong tầm 24 – 29 tuổi. Nhưng cũng có nhược điểm là sau khi phân tổ như vậy thì đặc điểm phân bố của dữ liệu trong mỗi tổ sẽ bị che dấu. Ví dụ ta biết có 12 người trong độ tuổi từ 24 – 29 nhưng ta không biết cụ thể là phần lớn họ trên 25 hay dưới 25, hoặc tuổi của họ phân bố đều trong phạm vi này.

3.1.3 Bảng tần số kết hợp hai biến

Bảng tần số có thể được kết cấu phức tạp hơn nữa là mô tả đặc điểm của mẫu nghiên cứu theo một biến dưới sự phân tách của một biến khác, hoặc linh động hơn nữa là phân tổ không đều mà theo chủ ý của người làm báo cáo.

Bảng dưới mô tả kết hợp hai đặc điểm là độ tuổi và khu vực cư trú của 7.584 thanh niên Việt nam trong cuộc Điều tra chọn mẫu quốc gia đầu tiên về Vị thành niên và Thanh niên Việt Nam (SAVY).

Bảng 3.10 a Khu vực cư trú của thanh niên trong mẫu điều tra phân tách theo từng nhóm tuổi

Thanh niên trong mẫu điều tra		Nhóm tuổi					
		(14 - 17) tuổi		(18 – 21) tuổi		(22 - 25) tuổi	
		Tần số (người)	Tần suất (%)	Tần số (người)	Tần suất (%)	Tần số (người)	Tần suất (%)
Khu vực	Thành thị	1020	31,60	919	36,12	723	39,90
	Nông thôn	2208	68,40	1625	63,88	1089	60,10
Tổng		3228	100	2544	100	1812	100

Nguồn: Điều tra chọn mẫu quốc gia đầu tiên về Vị thành niên và Thanh niên Việt Nam (SAVVY)

Ngoài ra bảng kết hợp này còn có thể được xoay theo chiều khác hoặc tách riêng thông tin về tần số và tần suất thành hai bảng riêng biệt, xem Bảng 3.10 b dưới đây.

Bảng 3.10 b Nhóm tuổi của thanh niên trong mẫu điều tra phân tách theo từng khu vực cư trú

Thanh niên trong mẫu điều tra		Khu vực			
		Thành thị		Nông thôn	
		Tần số (người)	Tần suất (%)	Tần số (người)	Tần suất (%)
Nhóm tuổi	(14 - 17) tuổi	1020	38,32	2208	44,86
	(18 – 21) tuổi	919	34,52	1625	33,02
	(22 - 25) tuổi	723	27,16	1089	22,12
Tổng		2662	100	4922	100

Để có thêm hiểu biết về các dạng bảng phức tạp này mời bạn đọc xem thêm Sách Phân tích dữ liệu nghiên cứu với SPSS (cùng tác giả) ở Chương 3.

3.2 TÓM LƯỢC VÀ TRÌNH BÀY DỮ LIỆU BẰNG ĐỒ THỊ PHÂN PHỐI TẦN SỐ (HISTOGRAM) VÀ ĐA GIÁC TẦN SỐ

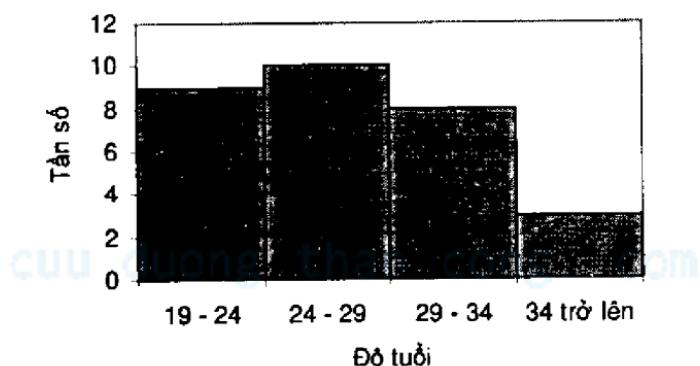
3.2.1 Đồ thị phân phối tần số

Mặc dù bảng tần số rất hữu dụng trong việc tóm lược và trình bày dữ liệu số lớn nhưng để đạt được mục đích thể hiện thông tin sinh động và súc tích hơn thì kĩ thuật Đồ thị phân phối tần số (Histogram) thường được sử dụng để chuyển hóa thông tin trên bảng tần số thành một hình ảnh hấp dẫn trực quan. Nhất là khi kết hợp với màu sắc và các phương pháp thể hiện độc đáo thì thông tin mà đồ thị phân phối tần số cung cấp dễ đi vào trí nhớ của người xem hơn các con số khô khan trên bảng tần số.

Histogram có mối liên kết chặt chẽ với thông tin về bảng tần số, nó được định nghĩa như một loại đồ thị biểu diễn sự phân phối tần số bằng các cột sao cho diện tích của cột tỷ lệ với tần số.

Khi xây dựng đồ thị này, các biểu hiện của tiêu chí hoặc đặc trưng thống kê ta quan tâm (ví dụ độ tuổi, mức tiêu thụ hàng hóa...) được thể hiện trên trục nằm ngang; còn trực đứng thể hiện tần số của các biểu hiện. Các cột của đồ thị phân phối tần số có bề rộng bằng nhau và có thể coi như bằng 1 đơn vị quan sát nên diện tích của cột sẽ đại diện cho số quan sát thuộc về biểu hiện mà cột đó mang tên. Với các quy tắc nêu trên, ta dùng thông tin trên Bảng tần số 3.5 xây dựng được đồ thị sau.

Hình 3.5 Đồ thị phân phối tần số cho tuổi của 30 sinh viên ngành KTKT
Số lượng sinh viên phân theo tuổi



Đồ thị phân phối tần số thể hiện 3 thông tin cơ bản sau:

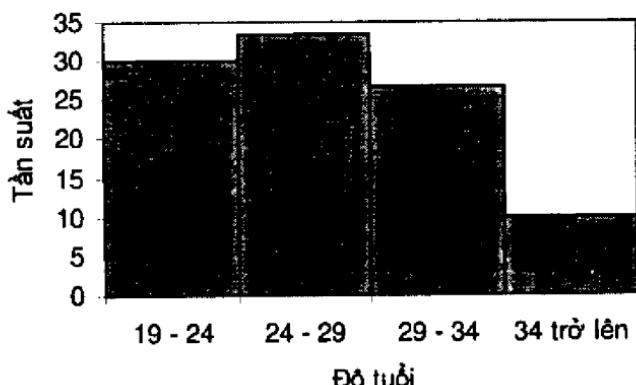
- Nó cho cảm nhận (một cách tương đối) về sự tập trung của tập dữ liệu
- Nó cũng cho thấy mức độ phân tán tương đối của tập dữ liệu
- Nó giúp ta cảm nhận sơ bộ hình dáng của phân phối là bằng phẳng, lệch hay cân đối.

Với đồ thị như Hình 3.5 trên đây cho cảm nhận chung là tuổi của 30 sinh viên được hỏi ít phân tán, tập trung chủ yếu ở độ tuổi 19 – 24 và 24 – 29. Hình dáng của phân phối hơi lệch sang phải (ở Chương 4 chúng ta sẽ làm rõ thế nào là “lệch sang phải”).

Chú ý là nếu chúng ta vẽ lại đồ thị phân phối tần số với thông tin của trực đứng là tần suất chúng ta sẽ nhận được một đồ thị có hình dáng hoàn toàn giống đồ thị đã vẽ, nghĩa là cảm nhận của chúng ta về vấn đề không khác trước (xem Hình 3.6), tuy nhiên đơn vị trên trực đứng là đơn vị %. Nhắc đến điều này vì Histogram vẽ từ thông tin tần suất sẽ có một công dụng hữu ích cho chúng ta ở Chương 5. Các bạn hãy nhớ những kiến thức này.

Hình 3.6

Tần suất sinh viên theo độ tuổi



Cách vẽ Histogram bằng Excel

Phần mềm Excel có thể giúp bạn vẽ đồ thị phân phối tần số nhanh chóng như sau (dùng lại ví dụ về tuổi của 30 sinh viên)

Cách nhập dữ liệu tuổi và dữ liệu về các giới hạn trên của các tổ (cho Bin Range) và các thao tác khác phục vụ cho việc vẽ đồ thị hoàn toàn giống cách đã hướng dẫn để lập bảng tần số bằng Excel cho ví dụ về tuổi của sinh viên.

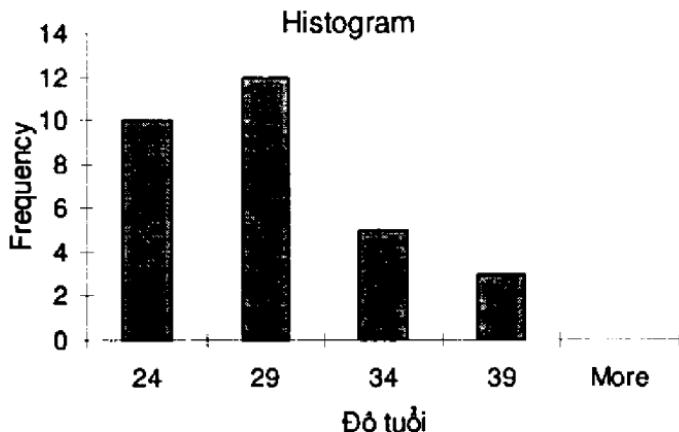
Tuy nhiên ở phần cuối của cửa sổ Histogram bạn chỉ lựa chọn duy nhất mục Chart Output (xem Hình 3.7).

Hình 3.7

- Pareto (sorted histogram)
- Cumulative Percentage
- Chart Output

Sau khi nhấn nút Ok bạn được kết quả ở Hình 3.8. Đồ thị này đã được tác giả chỉnh sửa chứ không phải là kết quả thô (thành thực mà nói là khá rối rắm) do Excel cung cấp, các bạn có thể tiến hành thêm nhiều hiệu chỉnh khác nữa để kết quả hoàn thiện hơn. Đặc biệt bạn phải chú ý chỉnh sửa để giữa các thanh của đồ thị không có khoảng cách vì đây là đặc điểm quan trọng của Histogram. Muốn làm được điều đó bạn trỏ chuột vào các cột đồ thị, bấm chuột trái 1 lần để chọn đối tượng, sau đó bấm chuột phải để lấy menu tắt, chọn mục Format Data Series..., trong cửa sổ này chọn phiếu Option, nhập giá trị 0 vào phần Gap width, rồi nhấn OK.

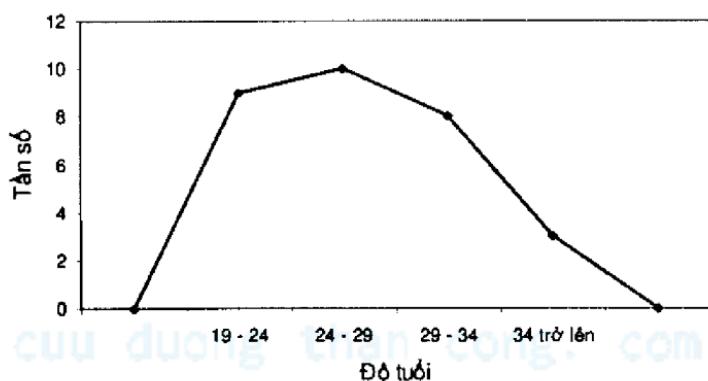
Hình 3.8



3.2.2 Đa giác tần số

Phương pháp thứ hai để biểu diễn phân phối tần số bằng đồ thị là dùng **Đa giác tần số**. Người ta vẽ đa giác tần số bằng cách nối các trung điểm của cạnh đỉnh các cột trong Histogram lại với nhau bằng các đoạn thẳng. Để làm cho đường biểu diễn này không có vẻ lơ lửng trên không ta có thể thêm vào hai bên của Histogram hai tổ có tần số bằng 0. Ta có đa giác tần số tương ứng với Histogram biểu diễn tuổi của 30 sinh viên tại chức như sau đây

Hình 3.9



Đa giác tần số cũng cho cảm nhận là hình dáng phân phối của tập dữ liệu về tuổi của 30 sinh viên tại chức có đuôi hơi dài về bên phải.

3.3 BIỂU ĐỒ THÂN VÀ LÁ

Biểu đồ thân – lá (còn gọi là biểu đồ nhánh – lá) là một công cụ hữu hiệu để tóm lược và trình bày tập dữ liệu mà vẫn giúp người xem thấy được

cách thức phân tán của dữ liệu gốc một cách chi tiết. Quy tắc lập đồ thị này là dữ liệu định lượng dưới dạng những con số sẽ được tách thành 2 phần: thân và lá. Việc phân chia này chỉ có tính qui ước và khá linh hoạt. Các chữ số bên phải của dữ liệu đóng vai trò lá có thể là 1 đến 2 chữ số ở hàng chục hay hàng đơn vị; còn các chữ số tương ứng ở bên tay trái của dữ liệu có thể là 1 hoặc 2 chữ số ở hàng trăm hay hàng chục đóng vai trò thân. Chúng ta hãy sử dụng ví dụ về tuổi của 30 sinh viên tại chức để hình dung cách xây dựng một biểu đồ thân – lá, ban đầu để dễ theo dõi ta chỉ phân tích tuổi của 5 người đầu tiên trong tập dữ liệu

28 23 30 24 19

Duyệt qua tập dữ liệu ta thấy tất cả đều được biểu diễn bằng những con số có 2 chữ số, nên việc phân chia rất đơn giản, chữ số hàng chục làm thân và số hàng đơn vị làm lá. Vì có người 19 tuổi, có người 23 tuổi, người 30 tuổi, nên nếu ta sử dụng con số hàng chục để tạo thân thì có ba loại thân là 1, 2, 3. Viết ra ba thân chính, sau đó ta lần lượt gắn các lá vào các thân tương ứng. Lúc này

19 được biểu diễn thành →

28, 23, 24 đều có thân 2 nên đều biểu diễn thành →

30 được biểu diễn thành →

Thân	Lá
1	9
2	8 3 4
3	0

Với cách tiến hành tương tự ta xây dựng được biểu đồ thân – lá cho toàn bộ tập dữ liệu về tuổi như sau:

Thân	Lá
1	9
2	8 3 4 1 2 2 0 1 6 7 5 9 7 1 5 8 6 9 9 2 7
3	0 9 1 7 3 0 5 2

Để biểu đồ dễ nhìn hơn và hợp lý hơn ta sẽ sắp lại các "lá" theo thứ tự từ số nhỏ đến số lớn theo chiều từ trái sang phải, đồng thời do thân 2 quá "sum suê" ta có thể tách nó thành 2 thân nhỏ, nguyên tắc tách thân là tách giữa số 4 và số 5. Như vậy thân số 2 sau khi tách sẽ thành 2 thân, thân 1 nhận các giá trị từ 20 đến 24 và thân 2 nhận các giá trị từ 25 đến 29. Tách thân 3 theo cách tương tự, kết quả là ta có một biểu đồ cân đối hơn như sau:

Thân	Lá
1	9
2	0 1 1 1 2 2 2 3 4
2	5 5 6 6 7 7 7 8 8 9 9 9
3	0 0 1 2 3
3	5 7 9

Nhìn vào biểu đồ thân – lá trên ta dễ dàng nhận thấy tuổi của sinh viên tại chức ngành KTKT tập trung trong khoảng từ 20 đến 29 vì nhánh 2 có nhiều lá nhất, chỉ có 1 người dưới 20 tuổi và 3 người trên 35 tuổi. Biểu đồ thân – lá đã thể hiện ưu điểm của nó khi cho người xem nhìn thấy cả giá trị thật của dữ liệu bên cạnh các cảm nhận về mức độ tập trung, phân tán. Lợi thế này chỉ phát huy tác dụng trong trường hợp số quan sát không quá lớn. Khi số lượng quan sát trong tập dữ liệu lên đến hàng trăm thì biểu đồ thân – lá lại có thể làm người xem rối mắt, lúc đó bảng tần số hay Histogram tỏ ra phù hợp hơn.

Nếu gặp tập dữ liệu mà các giá trị được thể hiện dưới 3 chữ số hoặc có số lẻ sau dấu phẩy thì để đơn giản người ta thường làm tròn số rồi mới biểu diễn biểu đồ thân – lá.

Ví dụ 1: muốn biểu diễn các số 613, 776, 1224 chúng ta lấy chữ số hàng trăm làm thân và làm tròn con số hàng chục để tạo lá

613 -> 6 | 1

776 -> 7 | 8

1224 -> 12 | 2

Ví dụ 2 : muốn biểu diễn các con số -1,2; 0,7; 4,3; 13,1 chúng ta lấy con số hàng chục làm thân, hàng đơn vị là lá, bỏ qua số lẻ.

-1,2 -> -0 | 1

0,7 và 4,3 -> 0 | 0 4

13,1 -> 1 | 3

3.4 TÓM LƯỢC VÀ TRÌNH BÀY DỮ LIỆU ĐỊNH TÍNH DẠNG PHÂN LOẠI BẰNG ĐỒ THỊ

Trong các nội dung trước chúng ta nghiên cứu phương pháp bảng biểu và đồ thị chủ yếu phục vụ mục đích tóm lược và trình bày dữ liệu định lượng. Tuy nhiên dữ liệu định tính dạng phân loại cũng là một tình huống hay gặp trong thống kê, để tóm lược thông tin của dữ liệu định tính nói chung,

chúng ta cũng có thể dùng bảng tần số (đã nghiên ở phần 3.1.1), tại nội dung 3.4 này chúng ta sẽ tiếp tục tập trung tìm hiểu các phương pháp trình bày dữ liệu định tính dạng phân loại bằng 3 dạng đồ thị cơ bản là đồ thị thanh đứng, đồ thị thanh ngang và đồ thị hình tròn.

Ví dụ khảo sát 500 sinh viên tại khoa Kinh tế của một trường ĐH với 5 chuyên ngành là Kinh tế phát triển, Quản Trị Kinh Doanh, Kế toán, Ngân hàng và Thương mại, người ta tổng hợp được số liệu như Bảng 3.11.

Trước khi tạo đồ thị cho biến phân loại (ở đây là tiêu thức “ngành học”), đầu tiên ta phải kiểm kê và tổng hợp số lượng cho từng loại rồi sau đó mới xây dựng các đồ thị trên cơ sở bảng tổng hợp này. Như vậy từ bảng tổng hợp dưới đây (mà thực ra rất gần với bảng tần số) ta bắt đầu nghiên cứu từng loại đồ thị.

Bảng 3.11

Ngành học	Số sinh viên (người)	Tỷ lệ (%)
Kinh tế phát triển	48	9,6
Kế toán	158	31,6
Ngân hàng	90	18
QTKD	124	24,8
Thương mại	80	16
	500	100

3.4.1 Đồ thị dạng thanh (Bar Chart)

Trên đồ thị này mỗi thanh đại diện một phân loại của biến (tiêu thức thống kê) ta quan tâm, chiều dài của thanh thể hiện tần số hay tỷ lệ phần trăm của các quan sát thuộc về phân loại đó. Đồ thị dạng thanh có hai loại là Đồ thị thanh đứng và đồ thị thanh ngang. Sự lựa chọn giữa hai dạng này hoàn toàn do tính thẩm mỹ và khoa học của người dựng đồ thị quyết định.

Các bước vẽ đồ thị

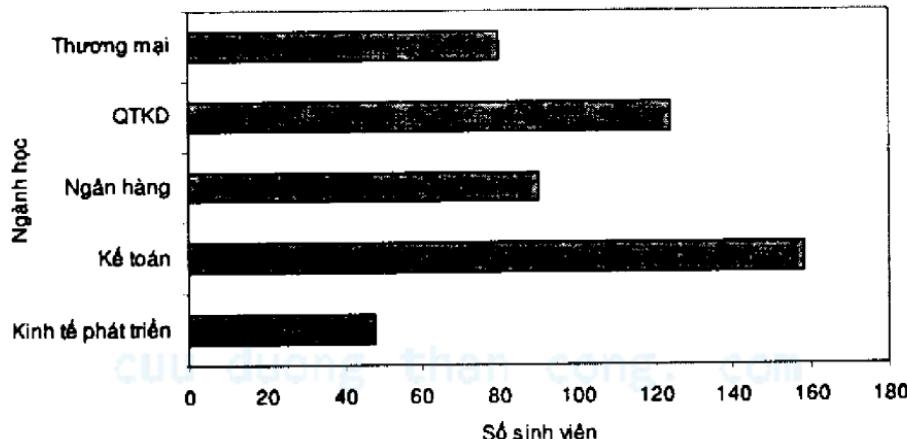
- Xác định các phân loại của biến (tiêu thức thống kê) ta quan tâm
- Xác định số quan sát thuộc về từng phân loại.
- Với đồ thị thanh đứng biến phân loại (tiêu thức thống kê) được đặt trên trục nằm ngang, còn trục đứng thể hiện số quan sát thuộc về các phân loại. Với đồ thị thanh ngang, chức năng của hai trục đảo lại.

- Dựng các thanh đồ thị theo nguyên tắc bề rộng của các thanh bằng nhau còn chiều dài của thanh tương ứng với số quan sát thuộc về phân loại mà thanh đại diện.

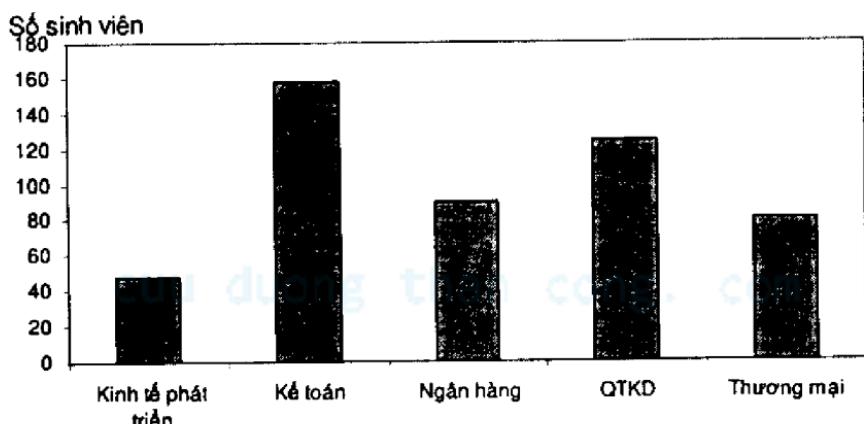
Sau đây là hình ảnh hai đồ thị xây dựng trên dữ liệu của Bảng 3.11

Như bạn thấy, trên đồ thị thanh ngang, thanh đại diện cho ngành học Kinh tế phát triển có chiều dài ngắn nhất, như vậy là ngành Kinh tế phát triển là ngành hiện có ít sinh viên theo học nhất, so sánh chiều dài của các thanh với nhau có thể kết luận là số lượng sinh viên của ngành Kinh tế phát triển chỉ bằng một nửa thậm chí $\frac{1}{3}$ lượng sinh viên các ngành khác.

Hình 3.10 Đồ thị thanh ngang



Hình 3.11 Đồ thị thanh đứng



Như vậy rõ ràng đồ thị thanh cho cảm nhận vấn đề nhanh và hiệu quả hơn con số trong bảng biểu tổng hợp.

Nhiều người vẫn hay nhầm lẫn giữa Histogram và đồ thị dạng thanh đứng. Mặc dù hình dạng rất giống nhau nhưng thực ra đây là hai công cụ

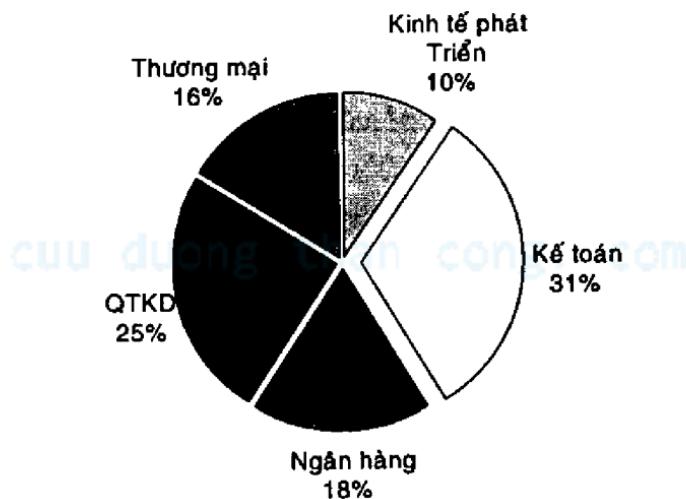
thống kê khá khác biệt, Histogram thể hiện phân phối tần số của dữ liệu định lượng, khi vẽ Histogram giữa các thanh đứng của đồ thị không có khoảng cách. Ngược lại đồ thị hình thanh chủ yếu dùng cho dữ liệu định tính dạng phân loại, giữa các thanh đồ thị phải có khoảng cách vì mỗi thanh là một biểu hiện của biến phân loại.

Chúng ta cũng có thể dùng thông tin về tỷ lệ phần trăm để vẽ đồ thị hình thanh.

3.4.2 Đồ thị hình tròn (Pie Chart)

Đồ thị hình tròn thường được dùng khi muốn tạo ấn tượng về kết cấu của hiện tượng đang quan tâm.

Hình 3.12



Trên đồ thị này, toàn bộ diện tích hình tròn được chia thành nhiều “mảnh” nhỏ hình rẻ quạt, diện tích mỗi “mảnh” tương đương với tỷ lệ của phân loại mà nó đại diện trong toàn thể và mang một màu khác nhau. Thứ tự của các phân loại trên đồ thị (theo chiều kim đồng hồ) là trật tự nó được sắp xếp trong bảng tổng hợp. Dù có hay không có ghi chú về tỷ lệ phần trăm đi kèm trong đồ thị thì chúng ta vẫn có thể đánh giá được là số lượng sinh viên ngành KTPT chỉ bằng nửa số lượng sinh viên ngành Ngân hàng bằng cách ước lượng diện tích mỗi mảnh đại diện. Hơn nữa đồ thị hình tròn cho thấy rõ ràng là tổng tỷ trọng của tất cả các phân loại là 100%.

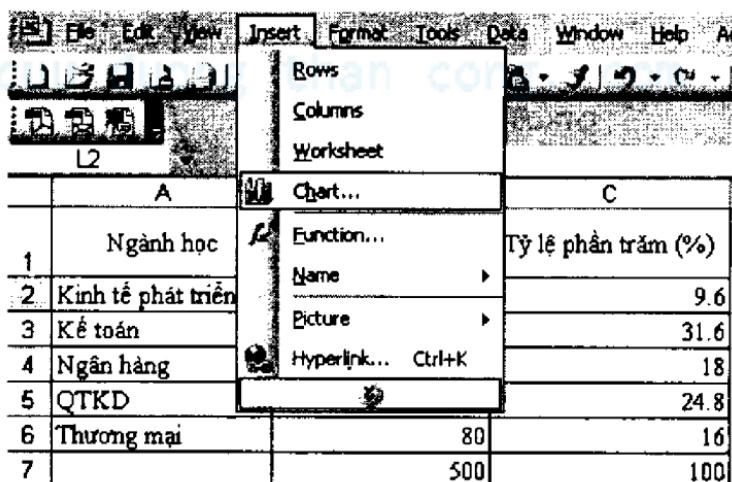
Chú ý rằng có một số nghiên cứu về khả năng nhận thức đồ thị của con người đã nhận thấy rằng đồ thị hình tròn kém hơn đồ thị dạng thanh trong việc thể hiện thông tin trực quan, vì mắt con người đánh giá tốt hơn khi so sánh chiều dài của các thanh trên một thang đo cố định so với so sánh diện tích các mảnh hình rẻ quạt. Chẳng hạn, nếu đồ thị trên đây không thể hiện thông tin về %, các bạn có thể sẽ phân vân không biết tỷ trọng sinh viên ngành Ngân hàng và ngành Thương mại thực ra có khác nhau không.

Tóm lại việc chọn đồ thị dạng thanh hay đồ thị hình tròn là phụ thuộc vào ý định thể hiện thông tin của người dựng đồ thị, nếu mục đích chủ yếu là để so sánh các phân loại thì dùng đồ thị hình thanh là tốt nhất, nếu muốn thể hiện tỷ trọng của từng phân loại trong toàn thể đối tượng thì ta dùng đồ thị hình tròn.

3.4.3 Cách vẽ đồ thị bằng Excel

Trên cửa sổ làm việc của Excel bạn nhập liệu như hình dưới, sau đó vào menu Insert chọn mục Chart..để mở cửa sổ Chart Wizard

Hình 3.13



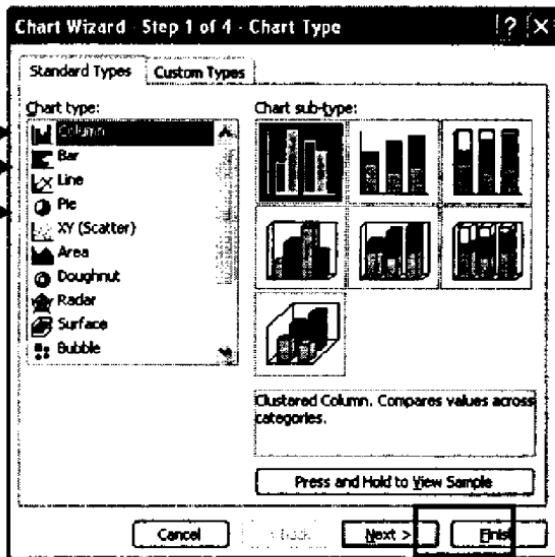
Trong cửa sổ này, tiến hành lựa chọn kiểu đồ thị bạn muốn vẽ trong khu vực Chart type.

Hình 3.14

Đồ thị thanh đứng

Đồ thị thanh ngang

Đồ thị hình tròn



Bấm nút Next, thực hiện tiếp các khai báo mà chương trình đòi hỏi. Cuối cùng nhấn nút Finish bạn sẽ được đồ thị như ý.

Chú ý là với các đồ thị do lệnh Chart Wizard cung cấp, bạn có thể tiến hành nhiều hiệu chỉnh để đạt được một đồ thị cuối cùng ưng ý nhất.

3.5 BIỂU ĐỒ PARETO

Một công cụ đồ thị miêu tả dữ liệu phân loại có thể cung cấp thông tin trực giác hơn cả hai đồ thị thanh đứng và đồ thị hình tròn là biểu đồ Pareto. Biểu đồ Pareto là loại đồ thị hình thanh đứng đặc biệt mà trong đó các thông tin về các quan sát được phân loại và được đưa lên đồ thị theo thứ tự giảm dần của các tần số và kết hợp luôn với đa giác tích lũy trên cùng đồ thị này. Nguyên tắc cơ bản của đồ thị Pareto là nó tách được “một số thông tin quan trọng” từ “rất nhiều thông tin vụn vặt”, giúp người xem tập trung vào những phân loại đáng chú ý của hiện tượng nên nó phát huy tác dụng lớn nhất khi các biến đang xem xét gồm khá nhiều phân loại. Biểu đồ Pareto hay được sử dụng trong quá trình phân tích chất lượng sản phẩm.

Khi hình thành biểu đồ Pareto, trục đứng ở bên trái thể hiện tần số hoặc tần suất, trục đứng bên phải thể hiện tần suất tích lũy (từ 0 ở chân trục đến 100% ở đỉnh trục), trục nằm ngang trình bày các phân loại. Các thanh đứng của đồ thị rộng bằng nhau sẽ thể hiện thông tin của trục đứng trái. Thông tin của trục đứng phải thể hiện dưới dạng đường vạch nối giữa các

chấm tạo nên cái gọi là đa giác tần suất tích lũy (cùng ý tưởng với đa giác tần số). Các chấm trên đa giác tần suất tích lũy tương ứng với mỗi phân loại được đặt ở trung tâm các thanh tương ứng của phân loại đó. Khi nghiên cứu biểu đồ Pareto, cần tập trung vào hai vấn đề: chiều dài của một thanh so sánh với chiều dài của các thanh bên phải nó và tần suất tích lũy của những phân loại gần nhau này.

Hãy xem ví dụ sau mô tả một tình huống trong quản lý. Dữ liệu được lấy từ 1 công ty sản xuất nhựa công nghiệp chuyên sản xuất các thành phần bằng nhựa để lắp ráp bàn phím máy vi tính và lắp ráp tivi. Bảng 3.12 liệt kê các khiếm khuyết có thể gặp ở các bàn phím máy tính được sản xuất trong thời gian 3 tháng gần đây.

Bảng 3.12

Nguyên nhân	Tần số	Tần suất
Chấm đen	413	6,53
Võ	1039	16,43
Không tuyển màu	258	4,08
Dấu kim châm	834	13,19
Xước	442	6,99
Khuôn không đều	275	4,35
Vạch màu bạc	413	6,53
Dấu chìm	371	5,87
Dấu phun	292	4,62
Bị biến dạng	1987	31,42
Tổng cộng	6324	100

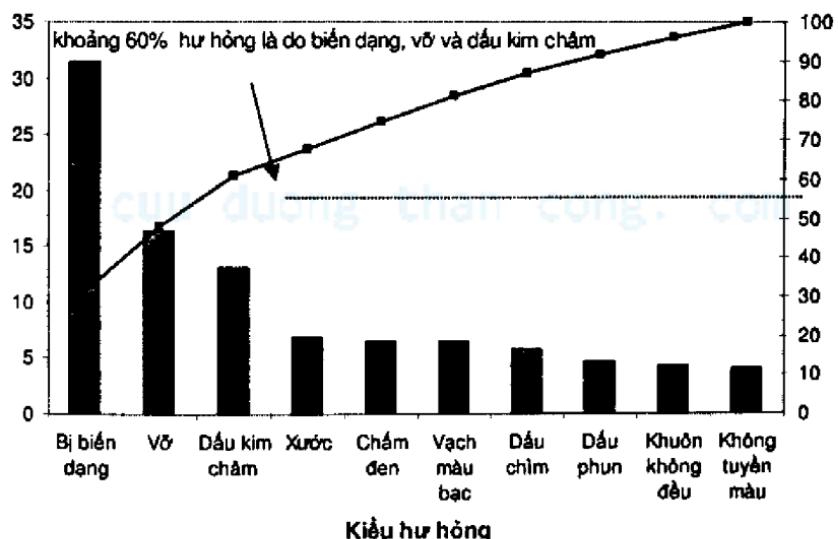
Để vẽ biểu đồ Pareto, đầu tiên phải xây dựng một bảng tóm tắt trong đó các phân loại khiếm khuyết được xếp xếp từ cao xuống thấp dựa trên tần số xuất hiện (thay vì được xếp theo bảng chữ cái). Bảng 3.13 là một bảng được xây dựng theo cách như vậy trên dữ liệu về khiếm khuyết của bàn phím máy tính từ Bảng 3.12. Tần suất tích lũy của những phân loại đã được sắp thứ tự này cũng được gộp vào bảng.

Ta quan sát thấy trên Bảng 3.13, sự biến dạng là hư hỏng đầu tiên được liệt kê (chiếm 31,42% của khiếm khuyết) theo sau đó là võ (16,43%), tiếp đó là dấu kim châm (13,19%). Hai loại thường xảy ra nhất là bị biến dạng và võ chiếm 47,85% các khiếm khuyết; ba loại thường gặp nhất là bị biến dạng, võ và dấu kim châm chiếm 61,04% vấn đề cần cải tiến... Các kết quả thể hiện trong Bảng 3.13 được đưa lên biểu đồ Pareto bằng công cụ đồ thị của Microsoft Excel trong Hình 3.16

Bảng 3.13

Nguyên nhân	Tần số	Tần suất	Tần suất tích lũy
Bị biến dạng	1987	31,42	31,42
Vỡ	1039	16,43	47,85
Dấu kim châm	834	13,19	61,04
Xước	442	6,99	68,03
Chấm đen	413	6,53	74,56
Vạch màu bạc	413	6,53	81,09
Dấu chìm	371	5,87	86,96
Dấu phun	292	4,62	91,58
Khuôn không đều	275	4,35	95,93
Không tuyên màu	258	4,08	100
Tổng công	6324	100	

Hình 3.16



Ở Hình 3.16, trục đứng bên trái được vẽ theo tần số, trục đứng phải là tần suất tích lũy. Nếu bạn đi theo đường đa giác tần suất tích lũy được vẽ qua các điểm giữa các thanh đứng với độ cao tương ứng với tần suất tích lũy, bạn sẽ thấy ba loại hư hỏng đầu tiên chiếm khoảng 61% vấn đề cần điều chỉnh. Vì biểu đồ Pareto xếp xắp các loại hư hỏng theo tần số xuất hiện của chúng, quá trình cải thiện tập trung vào các khiếm khuyết do biến dạng, vỡ và dấu kim châm sẽ giúp giảm tỷ lệ % hư hỏng nhiều nhất, sau đó là đến xước và chấm đen.

CHƯƠNG 4

TÓM TẮT DỮ LIỆU BẰNG CÁC ĐẠI LƯỢNG SỐ

Ở nội dung chương 3 chúng ta đã biết, đồ thị phân phối tần số là một công cụ hữu ích để chuyển hóa dữ liệu định lượng thành thông tin, nó thể hiện cho chúng ta biết dữ liệu của chúng ta tập trung ở đâu, và phân tán đến mức độ nào. Tuy nhiên đó mới chỉ là bước khởi đầu, để có thể mô tả sâu hơn một tập dữ liệu định lượng, chúng ta cần bổ sung vào bộ công cụ thống kê của mình các đại lượng số (đại lượng thống kê mô tả) để đo độ tập trung và độ phân tán. Các đại lượng này kết hợp với đồ thị phân phối tần số sẽ cho chúng ta một bức tranh rõ ràng chi tiết về một tập dữ liệu nghiên cứu.

Cơ bản, nội dung phần 4.1 sẽ tập trung vào các đại lượng số mô tả độ tập trung và phần 4.2 dành cho các đại lượng số mô tả độ phân tán của một tập dữ liệu mẫu được lấy ngẫu nhiên từ một tổng thể. Phần 4.3 khảo sát việc tính các đại lượng thống kê mô tả trong tình huống không có dữ liệu gốc mà phải làm việc với dữ liệu đã được tóm lược và trình bày lại bằng bảng tần số. Phần 4.4 trình bày cách tính các đại lượng thống kê mô tả độ tập trung và phân tán cơ bản cho bộ dữ liệu tổng thể, tuy đây là một tình huống hiếm gặp vì đa phần chúng ta làm việc với các mẫu chứ không phải với tổng thể nhưng việc nghiên cứu để có sự so sánh điểm giống và khác giữa cách tính các đại lượng thống kê mô tả cho dữ liệu mẫu và tổng thể là cần thiết để tạo tiền đề cho các chương phía sau. Các phần còn lại nghiên cứu các nội dung có liên quan đến việc mô tả dữ liệu dưới nhiều góc độ khác nhau.

4.1 CÁC ĐẠI LƯỢNG ĐO LƯỜNG MỨC ĐỘ TẬP TRUNG CỦA TẬP DỮ LIỆU VÀ PHƯƠNG PHÁP MÔ TẢ HÌNH DÁNG CỦA TẬP DỮ LIỆU

4.1.1 Các đại lượng đo lường độ tập trung phổ biến

4.1.1.1 Trung bình cộng (Arithmetic mean)

Gọi là trung bình cộng để phân biệt với trung bình nhân (sẽ trình bày trong phần sau). Trung bình cộng là một đại lượng số mô tả độ tập trung của dữ liệu được sử dụng phổ biến nhất.

Có hai loại số trung bình cộng là trung bình cộng đơn giản (mean) và trung bình cộng có trọng số (weighted mean).

Trung bình cộng đơn giản

Trung bình cộng đơn giản được tính bằng cách cộng tất cả các giá trị quan sát trong tập dữ liệu lại rồi đem kết quả đó chia cho số quan sát. Công thức tính:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

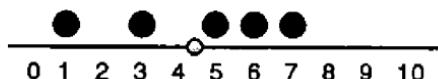
Trong đó:

\bar{x} là trung bình cộng đơn giản

n là số quan sát hay cỡ mẫu

x_i là giá trị trên quan sát thứ i

Ví dụ: Tập dữ liệu mẫu của chúng ta có 5 quan sát với các giá trị như sau



Trung bình cộng đơn giản tính được cho tập dữ liệu này như sau

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{1+3+5+6+7}{5} = \frac{22}{5} = 4,4$$

Với giá trị trung bình là 4,4 ta có thể hình dung rằng 5 quan sát trong tập dữ liệu phân tán xung quanh trung tâm của nó là vị trí có chấm tròn nhỏ trên trực số.

Trung bình cộng có trọng số (Weighted mean)

Khi chúng ta áp dụng công thức tính trung bình cộng đơn giản ở trên chúng ta giả định là mọi quan sát trong tập dữ liệu đều có tầm quan trọng ngang nhau, tuy nhiên có tình huống các giá trị quan sát được có tầm quan trọng khác nhau lúc này chúng ta phải dùng đến một trọng số thể hiện được mức độ quan trọng đó và áp dụng công thức tính trung bình có tính đến trọng số.

Ví dụ ta có một tập dữ liệu mẫu với k loại giá trị quan sát được là $\{x_1, x_2, x_3 \dots x_k\}$ và $\{w_1, w_2, w_3 \dots w_k\}$ lần lượt là những trọng số tương ứng của các loại giá trị quan sát này. Công thức trung bình có trọng số sẽ như sau

$$\bar{X}_w = \frac{\sum w_i x_i}{\sum w_i}$$

Một ví dụ rất dễ hình dung của số trung bình có trọng số là điểm trung bình của một sinh viên sau một học kì với nhiều môn học có số tín chỉ khác nhau. Để tính được điểm trung bình học tập người ta nhân điểm kết thúc môn học với số tín chỉ tương ứng (đóng vai trò trọng số trong công thức tính trung bình có trọng số), cộng các kết quả lại với nhau rồi đem đáp số này chia cho tổng số tín chỉ đã tham gia vào tính toán.

Bảng 4.1

Môn học	Số tín chỉ	Điểm
Dân số học	2	8,0
Nguyên lý kế toán	3	7,1
Marketing căn bản	4	8,4
Thống kê ứng dụng	4	8,0
Tiền tệ ngân hàng	3	5,7
Quản trị học	3	6,0

$$\bar{x}_w = \frac{(2x8) + (3x7,1) + (4x8,4) + (4x8) + (3x5,7) + (3x6)}{2+3+4+4+3+3} = \frac{138}{19} = 7,26$$

Cách tính điểm này rõ ràng sẽ chính xác hơn là tính theo kiểu trung bình đơn giản, tức là đem cộng điểm tổng kết các môn lại rồi chia đều cho số môn học.

Khi tổ hợp nhiều nhóm dữ liệu lại với nhau, ví dụ ta có 3 nhóm dữ liệu với số quan sát lần lượt là n_1, n_2, n_3 ; và các trung bình tương ứng lần lượt là $\bar{x}_1, \bar{x}_2, \bar{x}_3$. Trung bình của toàn bộ tập dữ liệu hình thành do tổ hợp ba nhóm này với nhau là trung bình có trọng số của ba trung bình riêng biệt, với trọng số là cỡ của các nhóm tức là n_1, n_2, n_3 . Do đó trung bình của toàn bộ tập dữ liệu tính theo công thức sau:

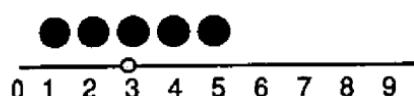
$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$$

Chúng ta sẽ ứng dụng công thức này trong nội dung Chương 9, phân tích phương sai. Ngoài ra một ứng dụng khác của trung bình có trọng số là để tính chỉ số thống kê trong Chương 13 (Chỉ số).

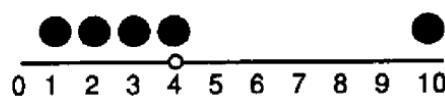
Phương pháp tính số trung bình trọng số thực ra cũng là phương pháp được áp dụng trong việc tính trung bình từ dữ liệu đã tóm tắt bằng bảng tần số mà chúng ta sẽ gặp ở Mục 4.3. Trong mục 4.3, tại nội dung tính giá trị trung bình bạn sẽ thấy một tình huống ứng dụng cụ thể của trung bình có trọng số khi dữ liệu đã được phân tổ và lập thành bảng tần số, trong đó các tần số của các tổ sẽ là trọng số tương ứng của tổ đó.

Tác động của các giá trị ngoại lệ lên số Trung bình cộng

Số trung bình là một sự san bằng bù trừ tất cả các giá trị trong tập dữ liệu vì thế dùng nó làm đại lượng tiêu biểu để mô tả độ tập trung của tập dữ liệu là hoàn toàn hợp lý. Tuy nhiên phải luôn nhớ rằng số trung bình tồn tại một nhược điểm lớn là rất nhạy cảm với các giá trị ngoại lệ, do đó nếu trong tập dữ liệu của chúng xuất hiện giá trị ngoại lệ nó sẽ làm cho giá trị trung bình tính được khác rất nhiều bản chất thật của nó, làm sai lệch cảm nhận về mức độ tập trung của tập dữ liệu. Các bạn quan sát hai ví dụ bằng hình sau đây:



$$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\bar{x} = \frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Giá trị trung bình cộng tính cho hai trường hợp này khác nhau về giá trị, trong khi tình trạng phân bố của các quan sát trong mỗi tập dữ liệu là khá tương tự, ngoại trừ một giá trị ngoại lệ ở tập dữ liệu thứ 2.

Trong tình huống mà có giá trị ngoại lệ làm ảnh hưởng đến cảm nhận về mức độ tập trung như vậy, một đại lượng thứ hai mô tả độ tập trung của tập dữ liệu là số trung vị sẽ được sử dụng đồng thời với số trung bình nhằm giúp “điều chỉnh” sai lệch trong cảm nhận về mức độ tập trung của tập dữ liệu.

Không tính đại lượng trung bình cho dữ liệu định danh

Bạn cần nhận thức rõ cấp bậc dữ liệu bạn đang làm việc trước khi tính toán các đại lượng thống kê mô tả. Một nhầm lẫn thường gặp là tính toán các đại lượng này cho dữ liệu đo lường bằng thang định danh. Ví dụ, một nhà sản xuất các đồ gia dụng điện tử điều tra một nhóm 17 khách hàng để xác định họ thích máy hát vỏ trắng, đen hay có màu. Dữ liệu định danh sau đó được mã hóa thành.

1 = đen

2 = trắng

3 = có màu

Thông tin phản hồi trên 17 người này như sau :

{1,1,3,2,1,2,2,2,3,1,1,1,3,2,2,1,2}

Dùng những con số mã hóa này, người ta tính trị trung bình của mẫu kiểm tra là:

$$\bar{x} = \frac{\sum x}{n} = \frac{30}{17} = 1,77$$

Từ kết quả trung bình tính được, người ta sẽ báo cáo rằng khách hàng thích một màu nằm giữa màu đen và trắng nhưng gần màu trắng hơn, kết luận này rõ ràng nghe vô nghĩa bởi vì không tính được điểm trung bình cho dữ liệu định danh. Kiểu nhầm lẫn này rất hay xảy ra khi người ta quen sử dụng các phần mềm để tính toán. Khi đó người ta có thể dễ dàng yêu cầu Excel, SPSS, hoặc những phần mềm khác tính trung bình, trung vị... cho tất cả các biến trong tập hợp dữ liệu. Sau đó trên các bảng kết quả được máy tính lập ra họ viết hàng loạt các báo cáo và rất dễ để lọt những nội dung vô nghĩa trong các bản báo cáo đó như chúng ta vừa thấy trên đây.

Có nên tính trị trung bình cho dữ liệu định lượng đo lường bằng thang đo khoảng?

Hiện nay vẫn có những bất đồng giữa các nhà thống kê rằng có nên tính trung bình cho dữ liệu định lượng đo lường bằng thang đo khoảng hay không. Ví dụ, ta dùng thang đo 5 điểm để đánh giá quan điểm của khách hàng đối với việc quảng cáo các sản phẩm trên TV, nội dung thang đo như sau:

- 1 = rất đồng ý
- 2 = đồng ý
- 3 = bình thường
- 4 = phản đối
- 5 = hoàn toàn phản đối

Các phản ứng đối với câu hỏi này thu được trên 10 người như sau:

{2,2,1,3,3,1,5,2,1,3}

Điểm đánh giá trung bình tính được cho nhóm này là 2,3. Sau đó chúng ta lại hỏi tương tự, ghi lại thông tin rồi tính trung bình cho nhóm người thứ hai và so sánh hai giá trị trung bình của hai nhóm để đánh giá thái độ của 2 nhóm người này.

Tuy nhiên, nhớ rằng chúng ta tính toán trung bình cho một biến được đo bằng thang khoảng cách, để so sánh được chúng ta sẽ phải có hai giả định cơ bản:

- Xem như khoảng cách giữa sự đánh giá 1 và 2 là bằng với khoảng cách giữa 2 và 3. Chúng ta cũng cho rằng những khoảng cách này là hoàn toàn tương tự với tình huống của nhóm thứ hai mà ta muốn so sánh. Mặc dù xét về mặt số học thì điều này đúng, nhưng xét về vấn

đề mà thang đo đang đánh giá thì 1 điểm cách biệt giữa rất đồng ý với đồng ý có thật bằng với 1 điểm cách biệt giữa đồng ý và bình thường hay không? Nếu không bằng, thì việc so sánh 2 trị trung bình tính được không thật sự là 1 cách hay?

- Chúng ta cũng giả sử là những người tham gia trả lời khảo sát có cùng định nghĩa thế nào là “rất đồng ý” và “đồng ý”. Khi một người chọn số 4 (phản đối) trong khảo sát, cần phải chắc họ có cùng cảm nhận như một người khác cũng chọn số 4 về vấn đề này? Nếu không thì việc so sánh 2 điểm trung bình ở đây sẽ là khập khiễng.

Mặc dù có những vấn đề như vậy với dữ liệu với thang đo khoảng, trong thực tế chúng ta vẫn thấy nhiều tình huống trong đó giá trị trung bình được tính toán phục vụ cho việc ra quyết định. Có thể dùng trung vị làm con số đo lường mức độ tập trung cho dữ liệu loại này bên cạnh số trung bình để điều chỉnh sự hạn chế đó.

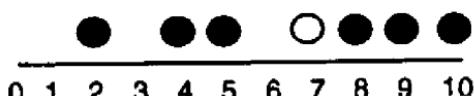
4.1.1.2 Trung vị (Median) - Me

Trong một tập dữ liệu đã được sắp xếp trật tự tăng dần thì trung vị là giá trị đứng giữa của tập dữ liệu, lúc này, không kể trung vị, sẽ có 50% số quan sát của tập dữ liệu có giá trị lớn hơn giá trị của số trung vị và 50% số quan sát của tập dữ liệu có giá trị bé hơn giá trị của số trung vị.

Muốn xác định số trung vị của một tập dữ liệu, đầu tiên là sắp xếp lại các quan sát của tập dữ liệu theo trật tự từ nhỏ đến lớn rồi sau đó xác định số trung vị theo quy tắc sau:

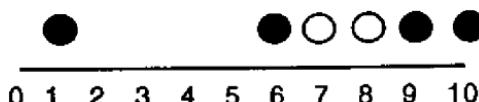
- Nếu số quan sát trong tập dữ liệu (n) là một số lẻ thì quan sát ở vị trí thứ $[(n+1)/2]$ là số trung vị.
- Nếu n là số chẵn, số trung vị là giá trị trung bình cộng của hai quan sát nằm ở vị trí chính giữa tập dữ liệu, tức là một quan sát ở vị trí thứ $n/2$ và một quan sát ở vị trí thứ $[(n+2)/2]$.

Ví dụ: Chúng ta có hai tập dữ liệu đã được sắp trật tự từ bé đến lớn, tập thứ nhất có 7 quan sát nên ta dễ dàng xác định ngay trung vị sẽ ở vị trí quan sát thứ 4 trong tập dữ liệu này đã sắp trật tự này.



Tại vị trí thứ 4 quan sát có giá trị bằng 7 (đứng nhầm lẫn giữa vị trí của quan sát và giá trị của quan sát tại vị trí đó).

Tập dữ liệu thứ hai có 6 quan sát nên trung vị là giá trị trung bình cộng của hai quan sát nằm ở vị trí chính giữa tập dữ liệu, tức là một quan sát ở vị trí thứ 3 và một quan sát ở vị trí thứ 4, ta có trung vị bằng $(7+8)/2 = 7,5$



Như vậy có thể thấy trung vị không bị ảnh hưởng bởi giá trị ngoại lệ cho nên nếu dùng đồng thời 2 đại lượng số trung vị và số trung bình khi mô tả mức độ phân tán sẽ giúp điều chỉnh nhận định của người xem xét số liệu về mức độ phân tán của tập dữ liệu.

Chú ý là khi xác định trung vị nhiều người hay lắn longoose vị trí của trung vị và giá trị của trung vị, nhô là ta xác định vị trí trung vị trước để biết trung vị là quan sát nào, sau đó xem quan sát đó có giá trị bao nhiêu thì đó là giá trị của trung vị.

4.1.1.3 Số mode (Mo)

Số Mode còn được gọi tên là yếu vị, đó là giá trị gặp nhiều lần nhất trong tập dữ liệu. Sau Trung vị, số Mode cũng được dùng để mô tả mức độ tập trung của tập dữ liệu, cũng như trung vị, mode không chịu ảnh hưởng của các giá trị ngoại lệ.

Ví dụ với tập dữ liệu có các giá trị quan sát như sau 1, 2, 3, 3, 4, 4, 4, 5, 6, 7, 7, 8 ta thấy giá trị 4 gặp nhiều lần nhất trong khi các giá trị khác chỉ xảy ra 1 đến 2 lần, do đó theo định nghĩa thì giá trị Mode của tập dữ liệu này là 4.

Có tập dữ liệu có một Mode, thì cũng có tập dữ liệu không có Mode hoặc có nhiều Mode.

Ví dụ:

Tập dữ liệu sau là tập dữ liệu không có Mode: 1, 2, 3, 4, 5, 6

Tập dữ liệu sau là tập dữ liệu có 2 Mode: 1, 2, 3, 3, 4, 5, 6, 6, 6, 7

Vì một tập hợp dữ liệu có thể không có Mode hoặc ngược lại, có nhiều Mode, nên Mode thường như không phải là một số đo về xu hướng tập trung đặc biệt hữu ích, tuy nhiên Mode là đại lượng thống kê mô tả duy nhất có thể vận dụng cho dữ liệu định tính. Hình dung rất đơn giản như sau, bạn thu thập thông tin về giới tính của sinh viên, biến giới tính là một biến định danh với 2 mã hóa: 1 đại diện cho nam và 2 đại diện cho nữ, nếu bạn đếm được nhiều số 2 hơn số 1, tức giá trị của Mode trong tình huống này là 2, và đồng nghĩa sinh viên nữ nhiều hơn sinh viên nam.

4.1.1.4 Trung bình nhân (Geometric mean)

Tuy tên gọi tương tự như trung bình cộng, nhưng đây là một đại lượng rất khác với số trung bình cộng về cách tính toán, và cả bản chất. Do đó, sắp xếp nó trong nhóm các đại lượng thống kê mô tả độ tập trung là không hợp lý lắm, tuy nhiên, vì tên gọi của đại lượng này bắt đầu bằng chữ “trung bình” nên chúng ta sẽ nghiên cứu sơ bộ về nó để phân biệt với số trung bình cộng.

Tổng quát, nếu có n giá trị x_i có quan hệ tích số kiểu $x_1 \times x_2 \times x_3 \dots \times x_n$ thì số trung bình nhân của n giá trị này được tính theo công thức

$$\bar{x} = \sqrt[n]{x_1 x_2 \dots x_n}$$

Số trung bình nhân hay được vận dụng để tính tốc độ phát triển trung bình khi các giá trị x_i là các con số tốc độ phát triển liên hoàn. Những khái niệm nghe chừng phức tạp này sẽ được quay lại bàn luận một cách chi tiết ở Chương 13, sau khi các bạn đã có kiến thức về số tương đối động thái liên hoàn.

4.1.2 Sử dụng Excel để tính toán các đại lượng thống kê mô tả độ tập trung

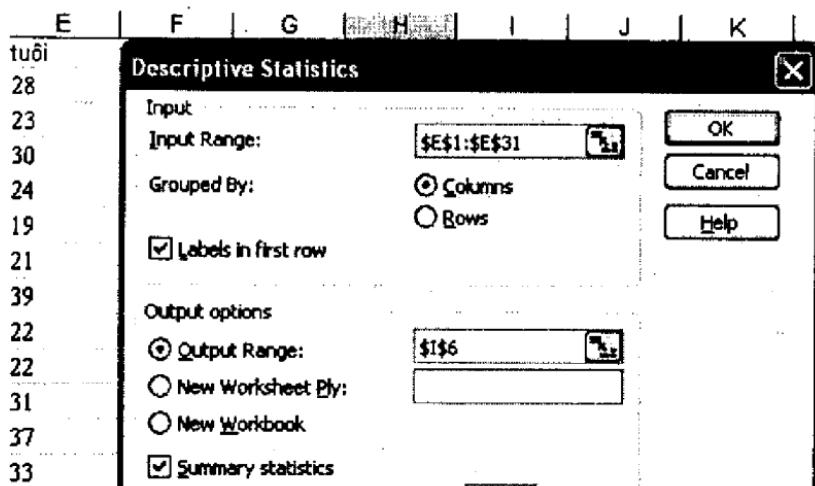
Để nhờ Excel tính toán các đại lượng thống kê mô tả chúng ta nhập dữ liệu vào cửa sổ làm việc theo cách đã quen thuộc. Dùng ví dụ về tuổi của 30 sinh viên tại chức. Sau đó vào menu Tool/ Data Analysis/ Descriptive Statistics mở cửa sổ Descriptive Statistics.

Tiến hành các khai báo như hướng dẫn trong hình dưới, sau đó bấm OK bạn có bảng kết quả 4.1.

Trong bảng kết quả này có một số đại lượng chúng ta chưa nghiên cứu đến vì nó thuộc mục đo lường độ phân tán.

Tuổi trung bình của 30 sinh viên được khảo sát là 26,93 tuổi; một nửa số sinh viên được khảo sát trên 27 tuổi và một nửa còn lại dưới 27 tuổi, như vậy là trung bình < trung vị. Ta thấy rằng tập dữ liệu này có hai mode nhưng Excel chỉ nhận ra được mode thứ nhất. Excel còn tìm được giá trị lớn nhất (Maximum), bé nhất (Minimum) của tập dữ liệu, tính tổng của tất cả các giá trị có trong tập dữ liệu (Sum) và đếm được số quan sát (Count).

Hình 4.1



Bảng 4.1

Tuổi	
Mean	26.933
Standard Error	0.927
Median	27
Mode	21
Standard Deviation	5.078
Sample Variance	25.789
Kurtosis	-0.127
Skewness	0.533
Range	20
Minimum	19
Maximum	39
Sum	808
Count	30

4.1.3 Nhóm các đại lượng khác mô tả sự phân bố của tập dữ liệu

4.1.3.1 Tứ phân vị (Quartiles)

Các tứ phân vị chia một tập dữ liệu đã được sắp xếp trật tự từ bé đến lớn thành 4 phần có số quan sát bằng nhau (trong khi trung vị thì chỉ chia làm 2 phần). Thực ra trình bày tứ phân vị trong nội dung các đại lượng mô tả sự tập trung thì không hợp lý lắm vì nó thuộc nhóm đại lượng phục vụ cho việc khảo sát độ phân tán của tập dữ liệu, tuy nhiên cách xác định tứ phân vị thì lại hơi giống số trung vị, và tứ phân vị thứ 2 cũng chính là trung vị nên ta tạm xếp nó vào nhóm lưỡng tính và gọi chung là đại lượng mô tả sự phân bố của tập dữ liệu.

- Tứ phân vị thứ nhất kí hiệu Q_1 là giá trị của quan sát ở tại vị trí xác định bởi công thức $[25\% \cdot (n+1)]$, với n là số quan sát của tập dữ liệu. Khi xác định được giá trị của Q_1 ta có thể kết luận rằng, không kể Q_1 thì có 25% số quan sát của tập dữ liệu có giá trị bé hơn hoặc bằng Q_1 và 75% số quan sát còn lại có giá trị bằng hoặc lớn hơn Q_1 .
- Tứ phân vị thứ hai kí hiệu Q_2 chính là trung vị.
- Tứ phân vị thứ 3 kí hiệu Q_3 là giá trị của quan sát ở tại vị trí xác định bởi công thức $[75\% \cdot (n+1)]$. Khi xác định được giá trị của Q_3 ta có thể kết luận rằng, không kể Q_3 thì có 75% số quan sát của tập dữ liệu có giá trị bé hơn hoặc bằng Q_3 và 25% số quan sát còn lại có giá trị bằng hoặc lớn hơn Q_3 .

Ví dụ Chúng ta có tập dữ liệu với 8 quan sát như sau

11 12 14 15 16 17 18 21

Xác định giá trị của các tứ phân vị

Q_1 ở vị trí $25\% \cdot (8+1) = 2,25 \rightarrow Q_1$ phải là một giá trị nằm giữa quan sát thứ 2 và quan sát thứ 3 theo tọa độ lệch $\frac{1}{4}$ gần về phía quan sát thứ 2 nên ta xác định giá trị Q_1 như sau

$$Q_1 = 12 + 0,25 \times (14 - 12) = 12,5$$

Q_2 ở vị trí $50\% \cdot (8+1) = 4,5 \rightarrow Q_2$ phải là một giá trị nằm giữa quan sát thứ 4 và quan sát thứ 5, nên ta xác định giá trị Q_2 như sau

$$Q_2 = \frac{15 + 16}{2} = 15,5$$

Q_3 ở vị trí $75\% \cdot (8+1) = 6,75 \rightarrow Q_3$ phải là một giá trị nằm giữa quan sát thứ 6 và quan sát thứ 7 theo tọa độ lệch $\frac{3}{4}$ tính từ quan sát thứ 6 nên ta xác định giá trị Q_3 như sau

$$Q_3 = 17 + 0,75 \times (18 - 17) = 17,75$$

Tứ phân vị được sử dụng để xác định giá trị của Độ trải giữa, một đại lượng thể hiện mức độ phân tán. Ngoài ra nó cũng còn có công dụng như phân vị.

4.1.3.2 Phân vị (Percentiles)

Phân vị thứ p ($0 < p < 100$) trong một dãy số đã sắp trật tự tăng dần là một giá trị chia dãy số làm hai phần, một phần gồm $p\%$ số quan sát có giá trị nhỏ hơn hoặc bằng giá trị phân vị thứ p , phần còn lại có $(100-p)\%$ số quan sát có giá trị bằng hoặc lớn hơn giá trị của phân vị thứ p . Theo cách nói này thì phân vị thứ 50 chính là trung vị. Phân vị thứ 25 chính là Q_1 và phân vị thứ 75 chính là Q_3 .

Công thức xác định vị trí của giá trị phân vị thứ p

$$i = \frac{p}{100} (n+1)$$

Ví dụ phân vị thứ 60 trong một dãy số gồm 19 quan sát đã sắp thứ tự tăng dần là giá trị nằm ở vị trí thứ 12 vì

$$i = \frac{p}{100} (n+1) = \frac{60}{100} (19+1) = 12$$

Nếu vị trí xác định được là một số lẻ ta cũng xử lý giống cách như xác định các tứ phân vị, giả dụ nếu dãy số có 18 quan sát thì phân vị thứ 60 lúc này nằm tại vị trí

$$i = \frac{p}{100} (n+1) = \frac{60}{100} (18+1) = 11,4$$

Vì 11,4 là số thập phân nên ta nội suy rằng giá trị của phân vị thứ 60 là giá trị nằm ở tọa độ lệch 4/10 tính từ quan sát thứ 11 giữa 2 quan sát 11 và 12. Từ đó mà tính được giá trị thật của nó.

Phân vị thường hay được sử dụng khi ta muốn biết mức đạt cao nhất của p% số quan sát, hay mức đạt thấp nhất của (100 - p)% số quan sát khi các quan sát được xếp từ thấp đến cao.

Muốn tính các giá trị tứ phân vị và phân vị bằng Excel chúng ta vào Menu Insert/Function mở cửa sổ các hàm của Excel, chọn loại hàm là Statistical và chọn tên hàm là Quartile hoặc Percentile.

Chú ý là thủ tục tính tứ phân vị và phân vị của Excel không được chuẩn nên các bạn có thể có các đáp số hơi khác với tính kết quả tính thủ công và do các phần mềm khác tính được. Với ví dụ về tuổi của 30 sinh viên tại chức, nếu làm thủ công ta tìm được giá trị $Q_3 = 30$ và phân vị 40 có giá trị 25,4. Hai giá trị này khi tính bằng Excel các bạn được lần lượt là 29,75 và 25,6.

4.1.4 Hình dáng của phân phối

Một đặc điểm khá quan trọng của tập dữ liệu là hình dáng phân phối của nó, chúng ta đã có các hiểu biết về các đại lượng Trung bình, trung vị, số mode, dựa trên các hiểu biết này chúng ta có thể “khảo sát” kiểu cách phân phối của một tập dữ liệu với độ chính xác tương đối. Hình dáng phân phối của một tập dữ liệu thuộc một trong hai kiểu là Cân đối hoặc lệch, trong nhóm phân phối lệch có hai kiểu là lệch trái hoặc lệch phải. Dựa trên giá trị của 2 đại lượng chính là Trung bình và trung vị chúng ta có thể suy luận được phân phối là cân đối hay lệch theo quy tắc sau

Nếu trung bình = trung vị \rightarrow cân đối

Nếu trung bình < trung vị \rightarrow lệch trái

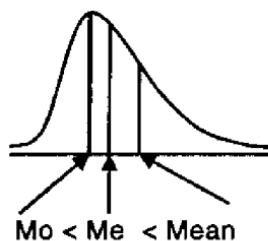
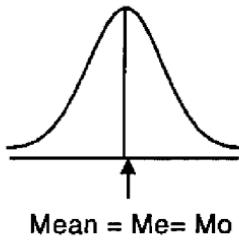
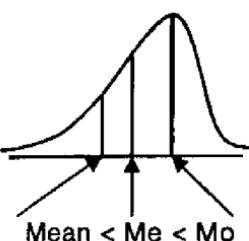
Nếu trung bình > trung vị \rightarrow lệch phải

Khảo sát hình dáng của các đa giác tần số tương ứng với các kiểu phân phối

Lệch trái

Cân đối

Lệch phải



Phân phối bị lệch phải có một cái “đuôi” kéo dài về phía bên phải, do trị trung bình bị một số ít quan sát có giá trị lớn kéo tăng lên khiến cho giá trị của nó trở nên lớn hơn giá trị trung vị; phân phối lệch trái có một cái “đuôi” kéo dài về phía bên trái do trị trung bình bị một số ít quan sát có giá trị nhỏ kéo giảm đi khiến nó bé hơn trung vị. Phân phối sẽ cân đối khi không có các giá trị quá đặc biệt xét theo cả hai hướng lớn hoặc nhỏ.

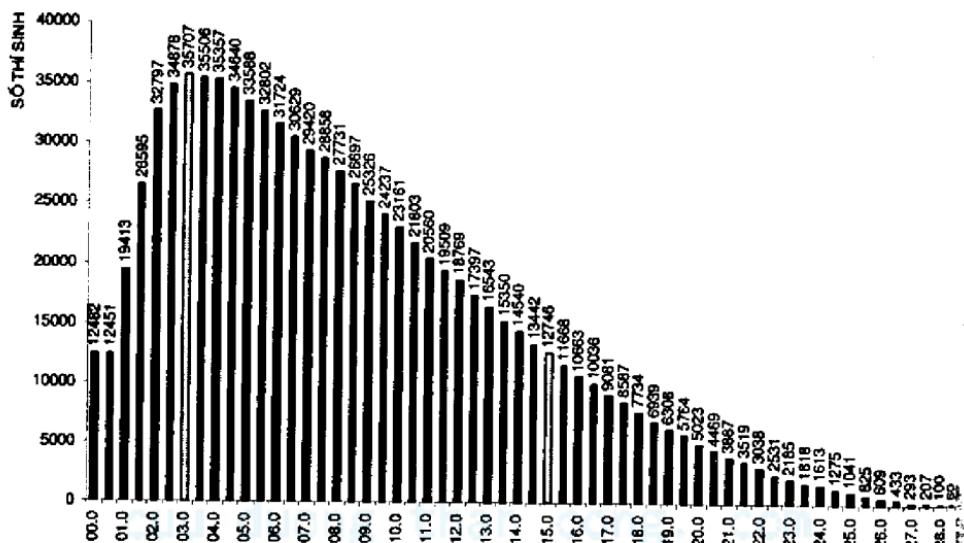
Từ nguyên tắc trên, chúng ta xem xét lại đa giác tần số đã dựng ở Hình 3.9, đuôi của nó kéo dài về phía phải chứng tỏ tập dữ liệu tuổi của 30 sinh viên có phân phối lệch phải (nghĩa là phần lớn các sinh viên có tuổi trẻ hoặc rất trẻ), nếu như vậy thì trị trung bình về tuổi của các sinh viên này phải lớn hơn trị trung vị. Xem lại bảng kết quả 4.1 ở nội dung 4.1.2 (Sử dụng Excel để tính toán các đại lượng thống kê mô tả độ tập trung) để kiểm chứng suy luận này thì ta thấy $TB \approx Me$, đó là do số quan sát quá ít, ta phải có đủ nhiều quan sát mới có thể đạt được tình huống hoàn hảo như lý thuyết.

Một ví dụ thực tế. Trong kì thi tuyển sinh đại học năm 2003, báo Tuổi trẻ đã tổng kết điểm thi của 141 trường đại học bằng một đồ thị phân phối tần số như Hình 4.2 sau đây. Do đặc điểm thống kê quan tâm là “Điểm” có tới 30 biểu hiện nên các thanh của biểu đồ nằm rất sát nhau và do đó không cần phải dùng đa giác tần số ta cũng có thể hình dung rất rõ hình dáng phân phối của dữ liệu về Điểm là một đường cong lệch phải. Vị trí cao nhất của đường cong cho biết giá trị Mode về Điểm thi (tổng cộng 3 môn) là 3 điểm. Đa phần các thí sinh có điểm tổng cộng ba môn dưới 15,

số lượng thí sinh có tổng điểm trên 15 ít. Bạn đọc có thể nhớ lại năm 2003 là một năm có kết quả thi tuyển sinh đại học rất kém, hình dáng phân phối của tập dữ liệu đã thể hiện tình hình này rất rõ.

Hình 4.2

Biểu đồ phân bố điểm của 141 trường đại học năm 2003



Nguồn: Báo Tuổi Trẻ, ngày 4/9/2003.

Trong các kết quả mà Excel tính toán về các đại lượng thống kê mô tả trình bày trong Bảng 4.1 còn có hai đại lượng đáng chú ý là Skewness và Kurtosis. Hai đại lượng này giúp hình dung về hình dáng của phân phối.

Skewness là một đại lượng đo lường mức độ lệch của phân phối, còn gọi tên là hệ số bất đối xứng, quy tắc nhận xét hệ số Skewness là

- Nếu phân phối cân xứng $\text{Skewness} = 0$
- Nếu phân phối lệch phải $\text{Skewness} > 0$
- Nếu phân phối lệch trái $\text{Skewness} < 0$

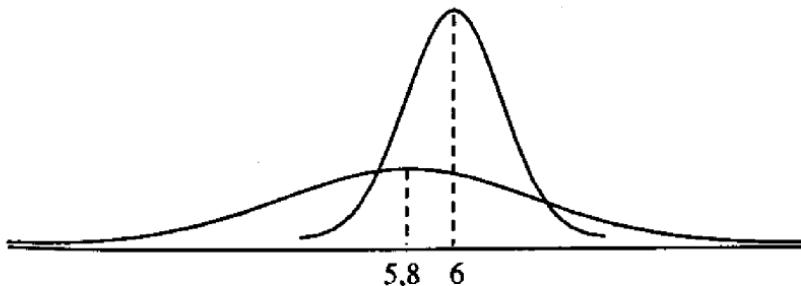
Kurtosis là một đại lượng đo mức độ tập trung tương đối của các quan sát quanh trung tâm của nó trong mối quan hệ so sánh với hai đuôi:

- Khi phân phối tập trung ở mức độ bình thường thì hệ số Kurtosis = 3
- Nếu phân phối tập trung hơn mức bình thường (hình dáng của đa giác tần số trông sẽ khá cao và nhọn với 2 đuôi hẹp) thì Kurtosis > 3
- Nếu Kurtosis < 3 ta sẽ có một đa giác tù hơn với hai đuôi dài.

4.2 CÁC ĐẠI LƯỢNG ĐO LƯỜNG ĐỘ PHÂN TÁN

Bên cạnh các đại lượng mô tả độ tập trung của một tập dữ liệu, còn có các đại lượng mô tả độ phân tán của dữ liệu. Nếu chỉ dùng một trong hai nhóm đại lượng này khi mô tả dữ liệu chúng ta có thể sẽ có những cảm nhận không đầy đủ hay lệch lạc. Chúng ta cùng xem xét ví dụ như sau. Giả sử chúng ta khảo sát hai địa phương A và B có qui mô dân số như nhau, thu nhập hộ gia đình trung bình tại địa phương A là 6 triệu đồng/hộ/tháng và tại địa phương B là 5,8 triệu đồng/hộ/tháng. Nếu một nhà xã hội học muốn biết địa phương nào có nhiều hộ nghèo hơn; hay ngược lại một chuỗi siêu thị muốn tìm hiểu địa phương nào có nhiều hộ gia đình có thu nhập ở mức cao hơn địa phương nào, nếu chỉ dựa vào hai số trung bình này, họ có tìm được câu trả lời đúng không?

Trên hai con số trung bình được cung cấp chúng ta có thể dễ dàng suy luận như sau: thu nhập hộ gia đình trung bình của địa phương A cao hơn địa phương B nên các hộ gia đình ở địa phương A về cơ bản sẽ giàu hơn các hộ gia đình ở địa phương B, như vậy địa phương A nhiều hộ có thu nhập cao hơn và địa phương B nhiều hộ nghèo hơn. Tuy nhiên nếu các bạn được cung cấp thêm hình dáng phân phối của hai tập dữ liệu về thu nhập hộ trung bình của các gia đình tại hai địa phương, các bạn sẽ có kết luận khác hẳn.



Phân phối tần số của tập dữ liệu về thu nhập hộ trung bình của các gia đình tại địa phương A có dáng cao và nhọn, trong khi với địa phương B nó lại rất tù với hai đuôi kéo dài. Vị trí tương đối của đuôi của hai phân phối cho ta thấy rằng số hộ gia đình có thu nhập cao (thậm chí rất cao) tại địa phương B nhiều hơn địa phương A, và địa phương B cũng có số hộ gia đình có thu nhập thấp nhiều hơn hẳn tại địa phương A. Ví dụ này cho thấy ngoài việc mô tả độ tập trung của một tập dữ liệu bạn đọc còn phải tìm hiểu cả độ phân tán của nó, hình dáng phân phối của một tập dữ liệu là một công cụ giúp mô tả độ phân tán của nó, ngoài ra chúng ta còn có thể dùng các đại lượng thống kê mô tả đo lường độ phân tán mà ta sẽ lần lượt tìm hiểu sau đây.

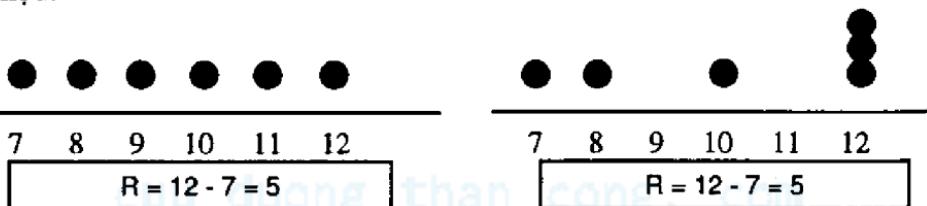
4.2.1 Khoảng biến thiên (Range) – R

Khoảng biến thiên là một đại lượng đo lường mức độ phân tán đơn giản và dễ hiểu nhất, vì nó được tính bằng cách lấy giá trị quan sát lớn nhất trừ đi giá trị quan sát bé nhất của tập dữ liệu

$$R = x_{\max} - x_{\min}$$

Nhược điểm của khoảng biến thiên là chỉ phụ thuộc vào hai giá trị lớn nhất và bé nhất của tập dữ liệu nên nó thay đổi rất nhạy theo các giá trị quan sát ngoại lệ. Bên cạnh đó, dù cho tập dữ liệu của bạn có bao nhiêu quan sát đi nữa thì R cũng chỉ được tính từ duy nhất hai giá trị x_{\max} và x_{\min} nên nó bỏ qua thông tin về cách phân bố nội bộ tập dữ liệu, vì những lý do này mà R cũng được xem là đại lượng đo lường độ phân tán yếu nhất và ít được sử dụng.

Một ví dụ khoảng biến thiên bỏ qua thông tin về cách phân bố của dữ liệu:



Một ví dụ cho thấy khoảng biến thiên rất nhạy với các giá trị ngoại lệ

$$1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,5$$

$$R = 5 - 1 = 4$$

$$1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,120$$

$$R = 120 - 1 = 119$$

4.2.2 Độ trải giữa (Interquartile Range) – R_Q

Đại lượng độ trải giữa (còn gọi là khoảng tứ phân vị) có thể khắc phục nhược điểm của khoảng biến thiên vì nó được tính bằng chênh lệch giữa tứ phân vị thứ ba và tứ phân vị thứ nhất

$$R_Q = Q_3 - Q_1$$

Ở nội dung tứ phân vị chúng ta đã dùng ví dụ về tập dữ liệu với 8 quan sát như sau : 11 12 14 15 16 17 18 21

Trên tập dữ liệu này các tứ phân vị đã được xác định giá trị lần lượt là

$$Q_1 = 12,5$$

$$Q_2 = 15,5$$

$$Q_3 = 17,75$$

Do đó khoảng tứ phân vị được xác định $R_Q = Q_3 - Q_1 = 17,75 - 12,5 = 5,25$

4.2.3 Phương sai và độ lệch chuẩn

Chúng ta đã nghiên cứu cả khoảng biến thiên và khoảng tứ phân vị để đánh giá độ biến thiên của tập dữ liệu, tuy nhiên cả hai đại lượng này đều không xem xét đến cách thức phân bố của tất cả các quan sát trong tập dữ liệu. Vì thế người ta sử dụng rất phổ biến hai đại lượng sau đây để phục vụ cho mục đích đó, chúng đánh giá được mức độ biến thiên của các quan sát quanh trung bình, đó là phương sai và độ lệch chuẩn.

Phương sai mẫu (sample variance) được định nghĩa gần như là trung bình của các biến thiên (đã được lấy) bình phương giữa từng quan sát trong tập dữ liệu so với giá trị trung bình của nó. Còn độ lệch chuẩn (Standard Deviation) là đại lượng được tính bằng cách lấy căn bậc hai của phương sai.

Công thức tính phương sai của một tập dữ liệu mẫu có n quan sát :

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Trong đó

- x_i là các giá trị quan sát thứ i của tập dữ liệu
- \bar{x} là số trung bình số học
- n là số quan sát của tập dữ liệu
- s^2 là phương sai

Lấy căn bậc hai của phương sai thì ta được độ lệch chuẩn, kí hiệu là s

$$s = \sqrt{s^2}$$

Hay viết một cách đầy đủ là:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Độ lệch chuẩn có cùng đơn vị tính với dữ liệu gốc còn với phương sai thì đơn vị tính đã được bình phương, và vì làm việc với đơn vị tính gốc thì dễ hơn đơn vị tính đã bình phương nên độ lệch chuẩn được sử dụng phổ biến hơn.

Chúng ta có thể thắc mắc tại sao lại phải lấy bình phương cho các giá trị độ lệch ($x_i - \bar{x}$) khi tính phương sai để rồi sau đó lại tốn công đi lấy căn bậc hai của phương sai để có độ lệch chuẩn, các bạn hình dung nếu ta chỉ nhầm đến việc tính độ lệch ta có thể dùng công thức sau

$$\frac{\sum_{i=1}^n (x_i - \bar{x})}{n-1}$$

Nhưng tử số của đại lượng này sẽ luôn bằng 0, rất dễ chứng minh điều này

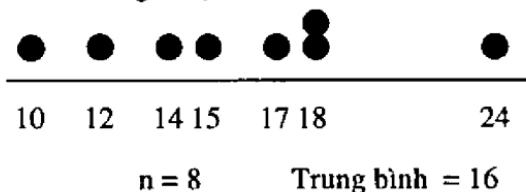
$$\sum(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n \bar{x}$$

$$\text{Mà } \bar{x} = (\sum x_i)/n \text{ hay } n \bar{x} = \sum x_i$$

$$\text{Do đó } \sum(x_i - \bar{x}) = \sum x_i - \sum x_i = 0$$

Để tránh sự triệt tiêu này chúng ta có thể dùng biện pháp lấy trị tuyệt đối cho chênh lệch, trị tuyệt đối tuy có vẻ đơn giản xong lại không có thuận lợi khi sử dụng cho mục đích suy diễn thống kê do sự gò bó trong trị số tuyệt đối đã giới hạn các phép biến đổi toán học. Cách khác để giải quyết vấn đề chính là bình phương các độ lệch.

Sau đây là một ví dụ về cách tính phương sai và độ lệch chuẩn của một tập dữ liệu có 8 quan sát. Quan sát hình vẽ bạn đọc chú ý có hai quả cầu tại vị trí 18 tức là có hai giá trị 18.



Vận dụng công thức tính phương sai ta được:

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{(10-\bar{x})^2 + (12-\bar{x})^2 + \dots + (18-\bar{x})^2 + (18-\bar{x})^2 + (24-\bar{x})^2}{n-1} \\ &= \frac{(10-16)^2 + (12-16)^2 + \dots + (18-16)^2 + (18-16)^2 + (24-16)^2}{8-1} \\ &= \frac{130}{7} = 18,57\end{aligned}$$

Từ đó độ lệch chuẩn được xác định nhờ lấy căn bậc hai của phương sai:

$$s = \sqrt{s^2} = \sqrt{18,57} = 4,31$$

Trong khi tính phương sai chúng ta đã thấy nó là tổng của các khác biệt đã được lấy bình phương giữa các quan sát và trung bình của tập dữ liệu, do đó cả phương sai và độ lệch chuẩn đều không có giá trị âm. Phương sai (và do đó cả độ lệch chuẩn) nhận giá trị bằng zero khi không có biến động trong tập dữ liệu, tức là tất cả các quan sát có giá trị bằng nhau và bằng đúng trung bình, đây là trường hợp hầu như không xảy ra trong thực tế vì dữ liệu mà ta quan tâm đều có biến động (vậy nên ta mới phải nghiên cứu chúng). Khi tìm hiểu một tập dữ liệu ta không chỉ quan tâm đến các đại lượng đo lường độ tập trung mà còn phải nghiên cứu cả các đại lượng đo lường độ phân tán của chúng, trong nhóm đó thì độ lệch chuẩn là đại lượng cơ bản nhất.

Trở lại Bảng 4.1 với ví dụ về tuổi của 30 sinh viên tại chức, phương sai của tập dữ liệu mẫu này là 25,789 (xem Sample Variance), sau đó lấy căn bậc hai của đáp số về phương sai ta sẽ có độ lệch chuẩn (Standard Deviation) $s = 5,078$ tuổi. Như Bảng 4.1 cho thấy, Excel tính toán sẵn hai giá trị này cho chúng ta. Trong bảng này còn có một đại lượng tên Standard Error, nó là kết quả khi lấy Độ lệch chuẩn chia cho căn bậc hai của cỡ mẫu của tập dữ liệu (s/\sqrt{n}), đại lượng này sẽ xuất hiện trong những chương sau.

4.3 CÁC ĐẠI LƯỢNG THỐNG KÊ MÔ TẢ CHO BẢNG TẦN SỐ

Trong thực tế chúng ta rất thường xuyên phải làm việc với dữ liệu đã được lập thành bảng tần số chứ không có dữ liệu gốc. Với những tình huống như vậy chúng ta vẫn có thể tính được (một cách xấp xỉ) các đại lượng thống kê mô tả như trung bình, trung vị, số mode, phương sai và độ lệch chuẩn. Chú ý là các đại lượng này chỉ được tính đúng một cách xấp xỉ vì cùng với việc phân tổ và lập bảng dữ liệu chúng ta đã bị “hao hụt” một lượng thông tin qua việc phân tổ, chia các quan sát vào các tổ...

4.3.1 Trung bình cộng

Để tính trị trung bình cho dữ liệu đã lập bảng tần số chúng ta vận dụng nguyên tắc của trung bình có trọng số, lúc này tần số của tổ nào cũng chính là trọng số của tổ đó; với bảng tần số lập ra trên cơ sở phân tổ thì mỗi tổ có một phạm vi giá trị (dao động từ cận dưới đến cận trên) ta sẽ lấy giá trị giữa của mỗi tổ làm đại diện cho tổ đó.

Công thức cơ bản tính trung bình cộng của bảng tần số:

$$\bar{x} = \frac{\sum_{i=1}^k x_i f_i}{\sum_{i=1}^k f_i}$$

Trong đó

- x_i là giá trị quan sát của tổ thứ i hoặc giá trị giữa của tổ i ($i = 1, 2, \dots, k$)
- Khi x_i là giá trị giữa của tổ thứ i nó được tính bằng cách lấy giá trị cận trên cộng giá trị cận dưới rồi đem kết quả chia 2 sẽ được giá trị đại diện cho tổ i .
- f_i là tần số của tổ thứ i ($i = 1, 2, \dots, k$)

Chú ý là $\sum_{i=1}^k f_i = n$ với n là cỡ mẫu quan sát

Để vận dụng công thức này chúng ta sử dụng hai ví dụ cho hai tình huống là Bảng tần số cho dữ liệu định lượng mà đặc điểm quan tâm ít biểu hiện (không phân tổ) và Bảng tần số cho dữ liệu định lượng mà đặc điểm quan tâm nhiều biểu hiện (có phân tổ).

4.3.1.1 Trường hợp bảng tần số cho dữ liệu định lượng không phân tổ

Sử dụng lại ví dụ về khảo sát tình hình đọc báo ngày của 200 người đã dùng ở Chương 3. Ta giả sử không có dữ liệu gốc, chúng ta dùng thông tin trên bảng tần số được báo cáo này tính trung bình hàng tuần mỗi người đọc bao nhiêu tờ báo A.

Bảng 4.2

Số báo đọc (tờ/tuần) x_i	Tần số (người) f_i	$x_i f_i$
0	44	0
1	24	24
2	18	36
3	16	48
4	20	80
5	22	110
6	26	156
7	30	210
Tổng	200	664

Để việc tính toán hệ thống và ít bị nhầm lẫn người ta hay lập bảng tính gồm các cột như trên, tính giá trị tổng $\sum(x_i f_i)$; $\sum(f_i)$ rồi thế các giá trị cuối cùng vào công thức.

$$\bar{x} = \frac{\sum_{i=1}^8 x_i f_i}{\sum_{i=1}^8 f_i} = \frac{664}{200} = 3,32 \text{ (tờ/tuần)}$$

4.3.1.2 Trường hợp bảng tần số cho dữ liệu định lượng có phân tách

Sử dụng lại ví dụ về tuổi của 30 sinh viên KT-KT đã dùng ở Chương 3

Bảng 4.3

Độ tuổi (tuổi)	Giá trị đại diện (x_i)	Tần số (SV) f_i	$x_i f_i$
19 - 24	21,5	9	193,5
24 - 29	26,5	10	265
29 - 34	31,5	8	252
34 trở lên	36,5	3	109,5
Tổng		30	820

Để tính giá trị đại diện của mỗi tổ ta lấy hai giá trị cận trên và dưới cộng lại rồi chia đôi, tức là nhằm mục đích lấy giá trị trung điểm của mỗi tổ làm đại diện cho tổ đó. Với ba tổ đầu việc tính toán tương tự nhau, ví dụ với tổ thứ 2 giá trị đại diện là $x_2 = (24+29)/2 = 26,5$ tuổi

Tổ cuối cùng là tổ mở, nên ta vận dụng quy tắc khi phân tổ mở, thì khoảng cách của tổ mở bằng với khoảng cách của tổ gần nó nhất, tổ gần nhất ở đây là tổ (29-34) với khoảng cách tổ là 5, do đó ta nội suy giá trị cận trên của tổ cuối cùng sẽ là $34+5 = 39 \rightarrow$ giá trị đại diện của tổ 4 là $x_4 = (34+39)/2 = 36,5$ tuổi

Từ các kết quả này ta lập bảng tính thêm các thông tin cần thiết rồi thay thế các kết quả trung gian vào công thức sau

$$\bar{x} = \frac{\sum_{i=1}^4 x_i f_i}{\sum_{i=1}^4 f_i} = \frac{820}{30} = 27,33 \text{ tuổi}$$

Nhớ lại kết quả trung bình cộng đơn giản do Excel tính được cho ví dụ này là 26,93 tuổi, còn với cách tính trung bình có trọng số kết quả là

27,33 tuổi, sự sai lệch này đã được chú ý ngay từ đầu, đó là kết quả của quá trình “hao hụt” thông tin do ta không làm việc với (hay không có) dữ liệu gốc.

4.3.2 Trung vị

Ta phân biệt các tình huống sau:

- Với bảng tần số không có phân tổ thì trung vị sẽ là giá trị của tổ có tần số tích lũy $= (\sum f_i + 1)/2$
- Với bảng tần số phân tổ có khoảng cách tổ: Cách tính số trung vị (và cả số mode) cho dữ liệu đã lập bảng tần số trên cơ sở phân tổ cần một phép nội suy trong tổ chứa số trung vị (và số mode) và phép nội suy này đòi hỏi một giả định về sự phân phối dữ liệu trong tổ, đó là, trừ khi dữ liệu có trị số bất thường rõ rệt, ta có thể tính được số trung vị (và số mode) khá đúng nếu ta giả định các giá trị quan sát trong mỗi tổ được phân phối đều khắp tổ. Thông thường thì giả định này chỉ được đáp ứng đến một mức độ nhất định nên các đáp số về trung vị (và số mode) cho bảng tần số cũng chỉ là đáp số gần đúng.

Quá trình tính số trung vị đi qua các bước sau:

Bước 1: Xác định tổ chứa trung vị: là tổ có tần số tích lũy vừa $\geq (\sum f_i + 1)/2$

Bước 2: Xác định giá trị gần đúng của trung vị theo công thức sau:

$$Me = x_{Me(\min)} + h_{Me} \frac{\sum f_i / 2 - S_{Me-1}}{f_{Me}}$$

Trong đó

- Me là giá trị trung vị (xấp xỉ) đang cần tính
- $x_{Me(\min)}$ là giá trị cận dưới của tổ chứa trung vị đã xác định ở bước 1
- h_{Me} là khoảng cách tổ của tổ chứa trung vị
- $\sum f_i$ là tổng các tần số của các tổ trong bảng tần số, nguyên tắc là $\sum f_i = n$
- S_{Me-1} là tổng các tần số của tổ đứng trước tổ chứa trung vị
- f_{Me} là tần số của tổ chứa trung vị

4.3.3 Số mode (yếu vị)

Với bảng tần số không có phân tổ thì mode là giá trị có tần số lớn nhất.

Với bảng tần số phân tổ có khoảng cách tổ, quá trình tính số mode đi qua các bước sau

Bước 1: xác định tổ chứa số mode là tổ có tần số lớn nhất

Bước 2: tính giá trị xấp xỉ của mode theo công thức sau

$$Mo = x_{Mo(\min)} + h_{Mo} \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})}$$

Trong đó

- Mo là giá trị mode (xấp xỉ) đang cần tính
- $x_{Mo(\min)}$ là giá trị cận dưới của tổ chứa mode đã xác định được ở bước 1
- h_{Mo} là khoảng cách tổ của tổ chứa mode
- f_{Mo} là tần số của tổ chứa mode
- f_{Mo-1} là tần số của tổ đứng sát trước tổ chứa mode
- f_{Mo+1} là tần số của tổ đứng sát sau tổ chứa mode

Xác định trung vị và mode, trường hợp bảng tần số không phân tổ

Ví dụ: tiếp tục sử dụng ví dụ về tình hình đọc báo vừa nhắc lại ở mục 4.3.1.1 Trung bình cộng

Ta thấy $(\sum f_i + 1)/2 = (200+1)/2 = 100,5$

Vận dụng quy tắc xác định trung vị là giá trị có tần số tích lũy = 100,05. Xét từ biểu hiện thứ nhất trở đi ta thấy tại biểu hiện thứ 4 của đặc điểm quan tâm, với giá trị của "Số báo đọc (tờ/tuần)" là 3, tần số tích lũy tại tổ này là 102 gần nhất với giá trị 100,05 nên ta chọn 3 là giá trị trung vị $\rightarrow Me = 3$ (tờ/tuần)

Trong thực tế nhiều khi việc xác định số trung vị đòi hỏi sự linh hoạt hơn nữa và luôn nhận thức rằng giá trị tìm được chỉ đúng tương đối mà thôi.

Bảng 4.4

Số báo đọc (tờ/tuần) x_i	Tần số (người) f_i	Tần số tích lũy
0	44	44
1	24	68
2	18	86
3	16	102
4	20	122
5	22	144
6	26	170
7	30	200
Tổng	200	

Vận dụng quy tắc xác định mode là giá trị có tần số lớn nhất thì ta thấy giá trị $M_0 = 0$ (tổ/tuần), kết luận này có nghĩa là tình huống phổ biến nhất là người được hỏi không đọc loại báo A.

Xác định trung vị và số mode, trường hợp bảng tần số có phân tổ

Ví dụ Có bảng tần số tóm lược và thể hiện thông tin về thu nhập hàng tuần của công nhân trong một công ty như sau:

Bảng 4.5

Thu nhập hàng tuần (ngàn đồng)	Tần số (người) f_i	Tần số tích lũy (người)
< 520	8	8
520 - 540	12	20
540 - 560	20	40
560 - 580	56	96
580 - 600	18	114
600 - 620	16	130
≥ 620	10	140
Tổng	140	

Trên thông tin mà bảng tần số cung cấp, tính toán số trung vị và số mode.

Trung vị <http://CuuDuongThanCong.com>

Bước 1: Tổ chứa trung vị là tổ có tần số tích lũy vừa $\geq (\sum f_i + 1)/2 = 141/2 = 70,5 \rightarrow$ đó là tổ (560 - 580) vì nó có tần số tích lũy là 96

Từ đó ta xác định được các chi tiết sau

$$x_{Me(min)} = 560; h_{Me} = 20; \sum f_i = 140; S_{Me-1} = 40; f_{Me} = 56$$

Bước 2: Xác định giá trị gần đúng của trung vị theo công thức sau

$$Me = 560 + 20 \frac{\frac{140}{2} - 40}{56} = 570,71 \text{ (ngàn đồng)}$$

Số mode

Bước 1: Tổ có tần số lớn nhất là 56 đó là tổ (560-580) \rightarrow tổ chứa số mode là tổ (560-580)

Từ đó ta xác định được các chi tiết sau

$$x_{Mo(min)} = 560; h_{Mo} = 20; f_{Mo} = 56; f_{Mo-1} = 20; f_{Mo+1} = 18$$

Bước 2: tính giá trị xấp xỉ của số mode theo công thức sau

$$Mo = 560 + 20 \frac{56 - 20}{(56 - 20) + (56 - 18)} = 569,73 \text{ (ngàn đồng)}$$

4.3.4 Phương sai và Độ lệch chuẩn

Nguyên tắc tính phương sai cho bảng tần số không phân tổ cũng giống cách tính phương sai cho bảng tần số lập ra trên cơ sở phân tổ, chỉ có khác biệt là nếu bảng có phân tổ ta sẽ lấy giá trị giữa của mỗi tổ làm đại diện cho tổ đó, còn bảng không phân tổ thì ta không cần tính giá trị đại diện.

Công thức tính phương sai của tập dữ liệu mẫu đã lập bảng tần số như sau:

$$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 f_i}{(\sum_{i=1}^k f_i) - 1}$$

Trong đó:

- x_i là giá trị của các tổ hoặc giá trị đại diện của các tổ được tính bằng cách lấy giá trị cận trên cộng giá trị cận dưới rồi đem kết quả chia đôi ($i=1,2\dots k$)
- \bar{x} là trị trung bình tính được của tập dữ liệu đã lập bảng phân tổ
- f_i là tần số của các tổ tương ứng, $\sum f_i = n$

Ví dụ: Tiếp tục với bảng tần số đã lập để tóm tắt tuổi của 30 sinh viên ngành kế toán kiểm toán, ta đã dùng công thức tính trung bình cộng cho dữ liệu lập bảng tần số này và được kết quả là 27,33 tuổi

Bảng 4.6

Dộ tuổi (tuổi)	Giá trị đại diện (tuổi) x_i	Tần số (SV) f_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \times f_i$
19 - 24	21,5	9	33,99	305,91
24 - 29	26,5	10	0,69	6,90
29 - 34	31,5	8	17,39	139,12
34 trở lên	36,5	3	84,09	252,27
Tổng		30		704,20

Ta lập thêm các cột tính toán các kết quả trung gian, rồi thay số liệu vào công thức tính phương sai ta có

$$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 f_i}{(\sum_{i=1}^k f_i) - 1} = \frac{704,2}{30-1} = 24,28$$

Lấy căn bậc hai của phương sai ta được độ lệch chuẩn:

$$s = \sqrt{24,28} = 4,93 \text{ (tuổi)}$$

Một lần nữa chúng ta thấy kết quả tính được từ bảng tần số có sai lệch với giá trị phương sai tính từ dữ liệu gốc (24,28 so với 25,79).

Khi tính phương sai cho bảng tần số lập ra trên cơ sở không phân tổ ta không cần lấy giá trị giữa của mỗi tổ làm đại diện cho tổ đó mà lấy chính giá trị của tổ làm x_i để thực hiện tính toán tương tự như trên.

4.4 CÁC ĐẠI LƯỢNG THỐNG KÊ MÔ TẢ CHO TỔNG THỂ

Trong toàn bộ những nội dung chúng ta đã nghiên cứu về các đại lượng mô tả độ tập trung và phân tán của một tập dữ liệu vừa qua, chúng ta chỉ xem xét dữ liệu mẫu do đó các đại lượng tính được là các đại lượng thống kê mô tả cho mẫu chứ không phải cho tổng thể. Nếu chúng ta có dữ liệu tổng thể ta cũng có thể tính các đại lượng này với phương pháp và ý tưởng khá tương tự. Chúng ta sẽ tìm hiểu cách tính 2 đại lượng tiêu biểu nhất là trung bình cộng đơn giản và phương sai của tổng thể.

4.4.1 Trung bình cộng của tổng thể

Với tổng thể, trung bình cộng được kí hiệu bằng chữ μ , nó được xác định với công thức sau

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

Trong đó:

- N là số quan sát hay qui mô tổng thể
- X_i là giá trị trên quan sát thứ i

Có thể thấy cách tính toán trung bình tổng thể không khác với cách tính trung bình mẫu, chỉ có các kí hiệu là thay đổi mà thôi, trung bình cộng có trọng số và trung bình trong tình huống dữ liệu đã lập bảng tần số của tổng thể cũng được tính với ý tưởng tương tự.

4.4.2 Phương sai và độ lệch chuẩn

Phương sai tổng thể được kí hiệu bằng chữ σ^2 . Công thức tính như sau:

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

Trong đó

- N là số quan sát hay qui mô tổng thể
- X_i là giá trị trên quan sát thứ i
- μ là trung bình tổng thể

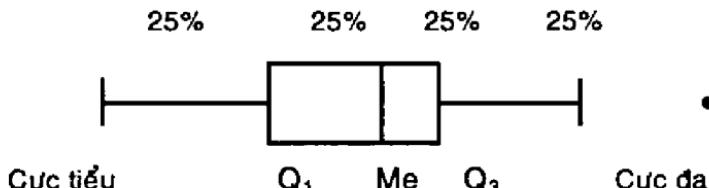
Độ lệch chuẩn tổng thể được kí hiệu là $\sigma = \sqrt{\sigma^2}$

Ngoài khái niệm về các kí hiệu, điểm khác biệt lớn nhất trong công thức tính phương sai tổng thể so với công thức tính phương sai mẫu là ở mẫu số người ta chia cho qui mô tổng thể N chứ không phải (N - 1). Một cách trực giác, chúng thấy rằng việc chia cho n hay N trong công thức tính phương sai xem ra hợp lý hơn là chia cho (n-1) hay (N-1) vì phương sai được hiểu là trung bình của các biến động (đã được) bình phương giữa từng quan sát trong tập dữ liệu so với giá trị trung bình của nó. Đối với tổng thể ta áp dụng công thức tính σ^2 như trên là hoàn toàn phù hợp với định nghĩa của phương sai. Tuy nhiên đối với mẫu, nếu ta không hiệu chỉnh mẫu số trong công thức tính phương sai mẫu thành lượng (n-1) thì trung bình của tất cả các phương sai mẫu của tất cả các mẫu cỡ n rút ra từ một tổng thể sẽ không bằng đúng phương sai tổng thể; cụ thể hơn, nếu ta rút ra được tất cả các tinh huống mẫu có cỡ n từ một tổng thể, với mỗi mẫu ta tính một phương sai mẫu theo công thức $s^2 = \sum(x_i - \bar{x})^2/n$ rồi sau đó tính trung bình của tất cả các phương sai mẫu này ta sẽ được một giá trị bé hơn σ^2 trong khi đúng ra nó phải bằng σ^2 (Chương phân phối của tham số mẫu sẽ trình bày rõ vấn đề này). Hiện tượng đó được gọi là phương sai mẫu đã ước lượng chêch phương sai tổng thể. Mặc dù những lý do toán học của sự hiệu chỉnh cỡ mẫu trong công thức tính phương sai mẫu nằm ngoài phạm vi nghiên cứu của môn học này nhưng bạn đọc có thể chấp nhận lý do sau: vì ta muốn phương sai mẫu là sự đại diện tốt nhất cho phương sai tổng thể nên ta phải tiến hành hiệu chỉnh công thức tính phương sai mẫu thay vì là $s^2 = \sum(x_i - \bar{x})^2/n$ thì phải là $s^2 = \sum(x_i - \bar{x})^2/(n-1)$. Do đó phương sai mẫu mà ta đã nghiên cứu ở nội dung 4.2.3 thực ra được gọi tên đầy đủ là phương sai mẫu hiệu chỉnh, và nó được dùng phổ biến trong thống kê ứng dụng với tên gọi ngắn gọn là phương sai mẫu.

4.5 KHÁM PHÁ DỮ LIỆU QUA BIỂU ĐỒ HỘP VÀ RÂU (BOX PLOT)

Chúng ta đã thảo luận 3 vấn đề chính khi nghiên cứu một tập dữ liệu định lượng là độ tập trung, độ phân tán và hình dáng của phân phối. Bây giờ chúng ta tìm hiểu một công cụ thống kê đặc biệt hay được dùng để khám phá một tập dữ liệu là biểu đồ hộp và râu. Gọi là biểu đồ nhưng thực ra đây là một đối tượng bằng hình ảnh có khả năng thể hiện đồng thời các thông tin là: giá trị cực đại, giá trị cực tiểu, 3 tứ phân vị và đôi khi cả các quan sát ngoại lệ. Các yếu tố này kết hợp lại sẽ tạo nên hình dáng của biểu đồ, và ngược lại, nhìn qua hình dáng của biểu đồ ta có thể hình dung được mối quan hệ tương đối giữa các yếu tố này.

Dưới đây là hình ảnh của một biểu đồ hộp và râu bất kỳ, chúng ta sẽ sử dụng nó để tìm hiểu bản chất của biểu đồ hộp và râu.



Hộp hình chữ nhật thể hiện 50% các quan sát ở giữa tập dữ liệu, bề rộng của hộp bằng độ trai giữa ($Q_3 - Q_1$) vì cạnh bên trái của hộp đi qua giá trị tứ phân vị thứ nhất và cạnh bên phải của hộp là tứ phân vị thứ 3. Đường thẳng đứng trong hộp đi qua giá trị tứ phân vị thứ 2 hay chính là trung vị, phần hộp ở bên trái đường này là 25% các quan sát có giá trị lớn hơn Q_1 và bé hơn trung vị, phần hộp bên phải đường này là 25% số quan sát có giá trị lớn hơn trung vị nhưng bé hơn Q_3 .

Hai râu của đồ thị biểu diễn 25% quan sát phía dưới Q_1 và 25% quan sát phía trên Q_3 . Tức là có 25% giá trị quan sát có trị số nhỏ hơn Q_1 và có 25% giá trị quan sát có trị số lớn hơn Q_3 . Râu trái đi từ Q_1 đến giá trị nhỏ nhất, râu phải đi từ Q_3 đến giá trị lớn nhất, chú ý là giá trị lớn nhất và nhỏ nhất này được xác định bằng giá trị cực đại hoặc cực tiểu thực sự của tập dữ liệu nhưng cũng có khi nó chưa phải là giá trị cực đại hoặc cực tiểu thực sự, nếu tập dữ liệu có các quan sát ngoại lệ thì chiều dài tối đa của 2 râu tính từ mỗi cạnh hộp được xác định bằng 1,5 lần độ trai giữa. Các quan sát ngoại lệ có giá trị vượt ra khỏi giới hạn này sẽ được diễn tả bằng dấu chấm hoặc dấu sao.

Ví dụ, lập biểu đồ hộp và râu mô tả tập dữ liệu về tuổi của 30 sinh viên tại chức kế toán kiểm toán đã sử dụng ở Chương 3.

Bước 1: sắp xếp tập dữ liệu theo thứ tự tăng dần

19 20 21 21 21 22 22 22 23 24 25 25 26 26 27
27 27 28 28 29 29 29 30 30 31 32 33 35 37 39

Bước 2: tính toán giá trị của các tứ phân vị, ta được

$$Q_1 = 22$$

$$Q_2 = 27$$

$$Q_3 = 30$$

$$\rightarrow \text{độ trai giữa} = Q_3 - Q_1 = 30 - 22 = 8$$

Bước 3: Vẽ hộp với bề rộng bằng độ trai giữa

Bước 4: Vẽ đường thẳng nằm trong hộp đi qua giá trị trung vị

Bước 5: Tính toán giá trị cực đại và cực tiểu

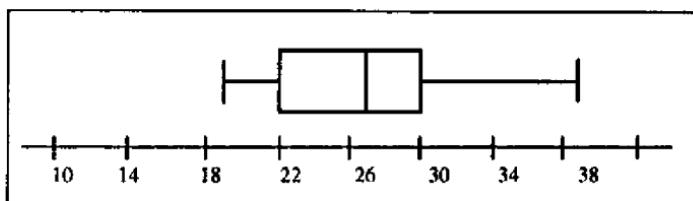
$$\text{Giá trị nhỏ nhất} = Q1 - 1,5(Q3-Q1) = 22 - 1,5*8 = 10$$

$$\text{Giá trị lớn nhất} = Q3 + 1,5(Q3-Q1) = 30 + 1,5*8 = 42$$

Bước 6: Vẽ hai râu

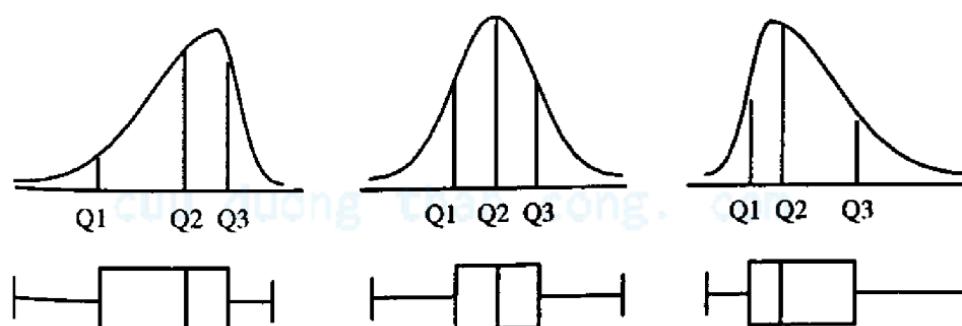
Do giá trị cực đại và cực tiểu thực sự của tập dữ liệu nằm trong giới hạn cực đại và cực tiểu tính theo qui tắc 1,5 lần độ trai giữa (10;42) nên ta chỉ vẽ chiều dài của hai râu đến vị trí thật của nó đó là 19 với cực tiểu và 39 với cực đại.

Đây là hình ảnh biểu đồ hộp và râu về tuổi của 30 sinh viên tại chúc



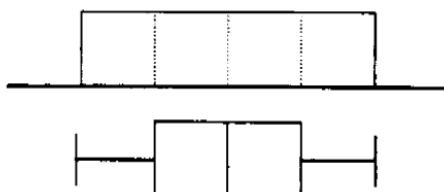
Như chúng ta thấy hộp chữ nhật lệch sang phía trái với râu bên phải rất dài, nhớ lại đa giác tần số đã dựng cho ví dụ này, đa giác này cho thấy phân phối của tuổi 30 sinh viên này lệch phải với một đuôi kéo dài về phía phải. Như vậy giữa đa giác tần số và biểu đồ hộp và râu trên cùng một tập dữ liệu có mối liên hệ, hình sau đây mô tả các tình huống của đa giác tần số và mối liên hệ giữa nó với biểu đồ hộp và râu.

Hình 4.3



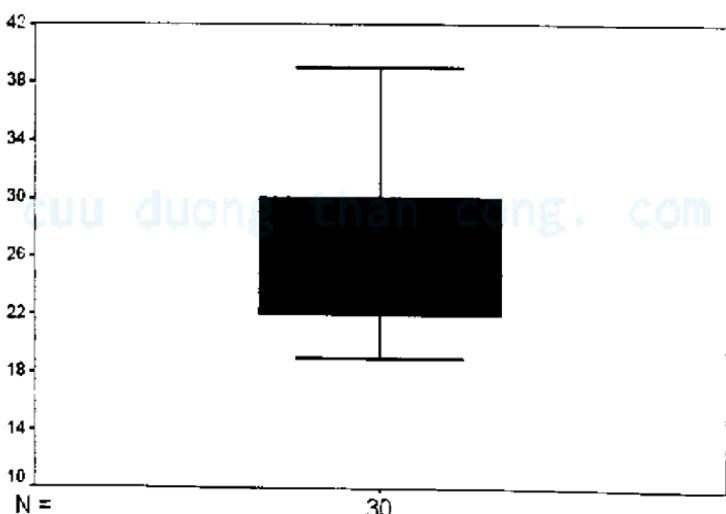
- Khi đa giác tần số cân đối như hình giữa, tức trung bình và trung vị trùng nhau, biểu đồ hộp và râu cũng cân đối với hai râu dài bằng nhau, đường thẳng đi qua trung vị sẽ nằm chính giữa hộp
- Khi đa giác tần số lệch trái như hình bên trái, biểu đồ hộp và râu sẽ có râu bên trái dài hơn râu bên phải, 25% số quan sát sẽ phân bố với mật độ loãng hơn trong khu vực bên trái Q_1 và 25% số quan sát sẽ phân bố với mật độ dày hơn ở khu vực bên phải Q_3 .
- Khi đa giác tần số lệch phải như hình bên phải, biểu đồ hộp và râu sẽ có râu bên phải dài hơn râu bên trái, lúc đó 25% số quan sát sẽ phân bố với mật độ loãng hơn trong khu vực bên phải Q_3 và 25% số quan sát sẽ phân bố với mật độ dày hơn ở khu vực bên trái Q_1 .

Với tình huống đa giác tần số bằng phẳng, biểu đồ hộp và râu sẽ hoàn toàn cân đối, chiều dài của hai râu bằng nhau và bằng $\frac{1}{2}$ độ rộng hộp.



Chú ý là tùy cách vẽ của các phần mềm thống kê khác nhau mà biểu đồ hộp và râu có thể được đặt đứng hoặc nằm ngang, tuy nhiên bản chất và quy tắc thì không thay đổi, dưới đây là biểu đồ hộp và râu biểu diễn tuổi của 30 sinh viên tại chúc được vẽ bằng phần mềm SPSS

Hình 4.4



4.6 SỬ DỤNG KẾT HỢP TRUNG BÌNH VÀ ĐỘ LỆCH TIÊU CHUẨN

4.6.1 Hệ số biến thiên (Coefficient of variation) - CV

Ta đã biết độ lệch chuẩn đo lường sự biến thiên của một tập dữ liệu, khi hai tập dữ liệu có cùng giá trị trung bình tập dữ liệu nào có độ lệch chuẩn lớn hơn sẽ biến thiên nhiều hơn. Tuy nhiên nếu hai tập dữ liệu có giá trị trung bình khác nhau thì không thể kết luận điều này bằng cách so sánh trực tiếp hai độ lệch chuẩn. Lúc đó, hệ số biến thiên được sử dụng để đo lường mức độ biến động tương đối của những tập dữ liệu có giá trị trung bình khác nhau.

Công thức tính hệ số biến thiên cho tập dữ liệu mẫu:

$$CV = \left(\frac{s}{\bar{x}} \right) \cdot 100\%$$

Trong đó \bar{x} là trung bình cộng

Công thức tính hệ số biến thiên cho tập dữ liệu tổng thể:

$$CV = \left(\frac{\sigma}{\mu} \right) \cdot 100\%$$

Khi hệ số biến thiên của hai tập dữ liệu được so sánh với nhau, hệ số biến thiên của tập nào lớn hơn thì tập đó biến động nhiều hơn.

Ví dụ: trong ngành tài chính, hệ số biến thiên hay được sử dụng để đo mức độ rủi ro tương đối của các danh mục vốn đầu tư. Chẳng hạn của một nhà kinh doanh trên thị trường chứng khoán xem xét hai danh mục đầu tư, danh mục A bao gồm các khoản đầu tư có lợi nhuận trung bình 16% với một độ lệch chuẩn là 4% và danh mục B có lợi nhuận trung bình 9% với độ lệch chuẩn 3%. Chúng ta có thể tính giá trị CV cho mỗi danh mục đầu tư như sau:

$$CV_A = \left(\frac{4}{16} \right) \cdot 100\% = 25\%$$

Và

$$CV_B = \left(\frac{3}{9} \right) \cdot 100\% = 33\%$$

Mặc dù danh mục đầu tư B có độ lệch chuẩn bé hơn (khiến ta cảm giác lợi nhuận ít bị biến động hơn) nhưng thực ra xem xét giá trị CV lại cho kết luận danh mục B biến thiên nhiều hơn danh mục A.

Ngoài ra hệ số biến thiên cũng hữu dụng khi so sánh hai tập dữ liệu có đơn vị đo khác nhau vì hệ số biến thiên độc lập với đơn vị đo lường và được tính bằng %.

Ví dụ: một doanh nghiệp kinh doanh dịch vụ vận chuyển hàng hóa cần phải xem xét giữa khối lượng và thể tích các kiện hàng họ vận chuyển, đối tượng nào biến động nhiều hơn. Một mẫu 200 kiện hàng được họ chọn ngẫu nhiên, sau đó đo lường khối lượng (kg) và thể tích (cm^3) của tất cả các kiện hàng trong mẫu rồi tính trị trung bình và độ lệch chuẩn, được kết quả lần lượt là :

- Khối lượng trung bình 11,801 kg với độ lệch chuẩn 1,78 kg
- Thể tích trung bình 4800 cm^3 với độ lệch chuẩn 1100 cm^3

Bằng cách nào họ có thể kết luận được giữa khối lượng và thể tích, yếu tố nào biến thiên nhiều hơn?

Vì hai yếu tố này có đơn vị tính khác nhau nên chúng ta sẽ dùng CV để so sánh mức độ biến động tương đối của chúng, cụ thể

- Với khối lượng $CV_{KL} = (1,78 / 11,801) * 100\% = 15,08\%$
- Với thể tích $CV_{TT} = (1100 / 4800) * 100\% = 22,92\%$

→ thể tích các kiện hàng công ty vận chuyển biến thiên nhiều hơn khối lượng của chúng. Cho nên công ty vận chuyển sẽ tính cước khác nhau theo thể tích của hàng hóa vận chuyển.

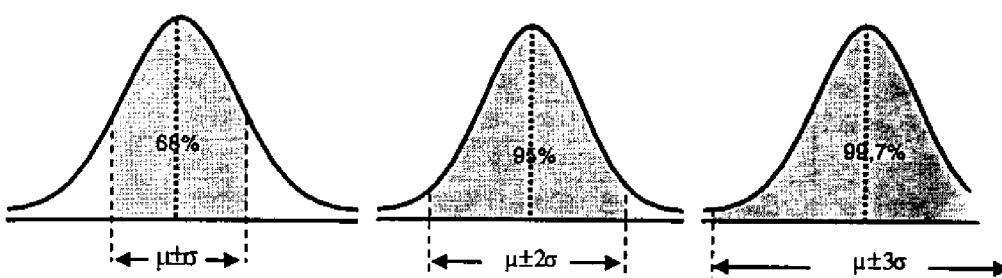
4.6.2 Quy tắc thực nghiệm

Sau khi nghiên cứu mối quan hệ giữa biểu đồ hộp và râu và đa giác tần số ta nhận thấy nếu một tập dữ liệu phân phối tạo thành đa giác tần số cân đối hình chuông (trung bình và trung vị trùng nhau) thì các quan sát có xu hướng tập trung dày hơn quanh khu vực trung tâm nên mới tạo nên hình chuông. Người ta nhận thấy rằng nếu dữ liệu có phân phối hình chuông cân đối thì có một quy tắc thực nghiệm như sau:

- Có khoảng 68% số quan sát của tổng thể hoặc mẫu sẽ tập trung trong phạm vi 1 độ lệch chuẩn so với trị trung bình.
- Có khoảng 95% số quan sát của tổng thể hoặc mẫu sẽ tập trung trong phạm vi 2 độ lệch chuẩn so với trị trung bình.
- Có khoảng 99,7% số quan sát của tổng thể hoặc mẫu sẽ tập trung trong phạm vi 3 độ lệch chuẩn so với trị trung bình

Quy tắc thực nghiệm cho một tổng thể được mô tả hình học như sau:

Hình 4.5



Quy tắc thực nghiệm giúp chúng ta có cơ sở để nhận diện những quan sát bất thường trong một tập dữ liệu. Theo quy tắc này, nếu một tập dữ liệu có phân phối hình chuông cân đối thì chỉ có 5% số quan sát nằm ngoài phạm vi hai lần độ lệch chuẩn tính từ trung bình, vì thế áp dụng quy tắc này ta có thể xem những quan sát có giá trị không nằm trong phạm vi $\mu \pm 2\sigma$ là những quan sát ngoại lệ. Quy tắc thực nghiệm cũng khẳng định nếu một tập dữ liệu có phân phối hình chuông cân đối thì chỉ có 0,3% số quan sát nằm ngoài phạm vi ba lần độ lệch chuẩn tính từ trung bình, vì thế ta có thể khẳng định những quan sát có giá trị không nằm trong phạm vi $\mu \pm 3\sigma$ thực sự là những quan sát ngoại lệ.

Giả dụ một giảng viên có tập dữ liệu về điểm thi kết thúc môn thống kê của một lớp học, dùng đồ thị Histogram mô tả tập dữ liệu này ông ta thấy nó có hình chuông khá cân đối, điểm trung bình tính được cho cả lớp là 5,6 điểm, độ lệch chuẩn là 1,41 điểm. Nếu giảng viên quyết định áp dụng quy tắc thực nghiệm để xét sinh viên xuất sắc trong môn thống kê theo tiêu chuẩn những người có điểm nằm ngoài phạm vi 2 độ lệch chuẩn trên trung bình được coi là học xuất sắc, như vậy học viên nào có điểm từ $(5,6+2 \times 1,41) = 8,42$ điểm trở lên được ông ta đánh giá là xuất sắc trong lớp này.

Chú ý là với những tập dữ liệu mà phân phối không phải là hình chuông cân đối, chúng ta không sử dụng qui tắc thực nghiệm mà nên dùng quy tắc Chebyshev.

4.6.3 Quy tắc Chebyshev

Quy tắc này phát biểu rằng với mọi tập dữ liệu bất kỳ, không cần xét đến hình dáng của phân phối, thì sẽ có ít nhất $(1 - 1/k^2)100\%$ quan sát tập trung trong phạm vi k lần độ lệch chuẩn tính từ trung bình, điều kiện áp dụng

quy tắc này là $k > 1$. Cụ thể hóa quy tắc Chebyshev cho một tổng thể có trung bình là μ và độ lệch chuẩn là σ trong bảng sau:

k	Số quan sát tối thiểu	Phạm vi
1,5	$(1 - 1/1,5^2)100\% = 55,6\%$	$(\mu \pm 1,5\sigma)$
2,0	$(1 - 1/2^2)100\% = 75\%$	$(\mu \pm 2\sigma)$
2,5	$(1 - 1/2,5^2)100\% = 84\%$	$(\mu \pm 2,5\sigma)$
3,0	$(1 - 1/3^2)100\% = 89\%$	$(\mu \pm 3\sigma)$

Bảng trên cho thấy với $k = 2$ có ít nhất $\frac{3}{4}$ hay 75% số quan sát của tổng thể tập trung trong phạm vi 2 lần độ lệch chuẩn xung quanh trung bình.

Tuy nhiên thực tế cho thấy tỷ lệ % của các quan sát rơi vào các khoảng thường cao hơn so với giới hạn mà quy tắc Chebyshev đưa ra, vì thế nó được đánh giá là một quy tắc khá dễ đặt. Ưu điểm của nó là có thể áp dụng cho bất kỳ tập dữ liệu nào mà không cần phải có phân phối cân đối.

4.6.4 Chuẩn hóa dữ liệu

Khi làm việc với dữ liệu số lượng, sẽ có lúc bạn cần biến đổi chúng thành dữ liệu ở một thang đo chuẩn, chẳng hạn nếu bạn muốn so sánh các đối tượng được đo lường bằng những phương pháp đo hay đơn vị đo khác nhau, việc làm này gọi là chuẩn hóa dữ liệu. Giá trị dữ liệu đã chuẩn hóa sẽ cho biết một giá trị quan sát trong tập dữ liệu gốc, lệch khỏi trung bình của nó mấy lần độ lệch chuẩn. Điều này thể hiện qua công thức sau đây

Công thức tính giá trị chuẩn hóa z cho dữ liệu tổng thể

$$z = \frac{x - \mu}{\sigma}$$

Trong đó

- x là giá trị dữ liệu gốc
- μ = trung bình tổng thể
- σ = độ lệch chuẩn của tổng thể
- z = điểm số chuẩn hóa cho biết x cách xa trung bình một khoảng bằng mấy lần độ lệch chuẩn

Công thức tính giá trị chuẩn hóa z cho dữ liệu mẫu

$$z = \frac{x - \bar{x}}{s}$$

Trong đó

- x là giá trị dữ liệu gốc
- \bar{x} là trung bình mẫu
- s là độ lệch chuẩn của mẫu

- z = điểm số chuẩn hóa cho biết x cách xa trung bình một khoảng bằng mấy lần độ lệch chuẩn

Một giá trị z tiến gần đến 0 có nghĩa là quan sát đó ở vị trí rất gần trung bình. Một giá trị z bằng -1 có nghĩa là quan sát thực tế đó ở vị trí lệch một độ lệch chuẩn so với trung bình về phía trái, và bằng +1 tức là nó lệch về phía phải một khoảng cách bằng một độ lệch chuẩn.

Ví dụ một học sinh có điểm thi môn Toán là 8,9 (thang điểm 10) và môn Anh văn là 89 (thang điểm 100), như vậy thì học sinh này học môn nào tốt hơn, hay là học tốt hai môn như nhau?

Với 2 môn thi em học sinh đã tham gia, điểm trung bình và độ lệch chuẩn của điểm lần lượt được tính (cho tập dữ liệu là điểm của tất cả học sinh trong lớp) như sau: Anh văn ($\bar{x} = 65$, $s = 17$) và Toán ($\bar{x} = 5,7$ và $s = 1,6$)

Chúng ta dùng phương pháp chuẩn hóa dữ liệu để xác định học sinh này có kết quả thi môn nào cao hơn, cụ thể:

$$z_T = \frac{8,9 - 5,7}{1,6} = 2 \quad \text{và} \quad z_A = \frac{89 - 65}{17} = 1,4$$

Giá trị chuẩn hóa của điểm thi môn Toán cho thấy điểm toán của học sinh này cao hơn trung bình tới 2 độ lệch chuẩn trong khi điểm chuẩn hóa môn Anh văn cho thấy điểm của em chỉ cao hơn trung bình 1,4 lần. Như vậy em học sinh này học khá môn toán hơn môn Anh nếu so với các học sinh khác trong cùng lớp.

4.7 PHÂN BIỆT MỘT SỐ CẶP KHÁI NIỆM

4.7.1 Phân biệt tham số tổng thể và tham số mẫu

Các đại lượng thống kê tính được có thể thuộc về hai tập hợp dữ liệu là tổng thể hoặc mẫu, với mỗi tập hợp nó có một tên gọi khác nhau.

Tham số mẫu là tên gọi chung cho các đại lượng (như trung bình, tỉ lệ, phương sai, độ lệch chuẩn ...) tính được trên tập dữ liệu của mẫu được chọn từ tổng thể chung. Giá trị của các đại lượng này thay đổi tùy theo mẫu được chọn.

Tham số tổng thể là tên gọi chung cho các đại lượng (như trung bình, tỉ lệ, phương sai, độ lệch chuẩn ...) tính được trên tập dữ liệu của toàn bộ tổng thể. Vì tổng thể là không thay đổi nên giá trị của các thông số tính được cũng không thay đổi. Hay nói cách khác, tham số của tổng thể nếu thu thập được dữ liệu trên toàn bộ các đơn vị sẽ là một con số xác định.

4.7.2 Phân biệt biến thiên và độ lệch chuẩn

Biến thiên là khái niệm diễn tả sự chênh lệch giữa các quan sát riêng lẻ so với trị trung bình của tập dữ liệu, còn độ lệch chuẩn là một thước đo tổng hợp mức độ biến thiên trong đó có sự tham gia tính toán của tất cả các giá trị quan sát. Chỉ với một con số thống kê duy nhất, độ lệch chuẩn đã diễn tả khá đầy đủ được mức độ biến thiên (phân tán) ít hay nhiều của dữ liệu. Hiểu rộng hơn, nó diễn tả chính xác mức độ tương tự hay mức độ khác biệt của các đối tượng trong tập dữ liệu theo một đặc trưng nghiên cứu (so với các thước đo độ phân tán đã trình bày trước nó). Chính vì vậy trong thực tế, độ lệch chuẩn là đại lượng đo lường độ phân tán được sử dụng phổ biến nhất.

cuu duong than cong. com

cuu duong than cong. com

CHƯƠNG 5

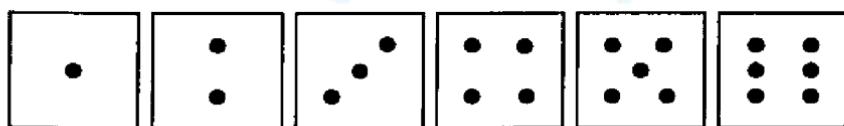
XÁC SUẤT CĂN BẢN, BIẾN NGẪU NHIÊN VÀ LUẬT PHÂN PHỐI XÁC SUẤT

5.1 XÁC SUẤT CĂN BẢN

Vì xác suất giữ vai trò quan trọng trong lý thuyết và ứng dụng của thống kê nên chúng ta phải trang bị những hiểu biết đầy đủ về xác suất trước khi tiếp tục đi vào những chi tiết có tính kỹ thuật của thống kê.

5.1.1 Ý nghĩa của xác suất

Các bạn có biết các nghiên cứu có tính chất toán học về xác suất bắt đầu từ khi nào không? Khoảng hơn 300 năm trước một nhà quý tộc người Pháp tên là Méré đã quan sát và tự đặt ra những câu hỏi về quy luật ẩn sau các trò chơi có tính may rủi, ví dụ như trò tung súc sắc. Nhà toán học nổi tiếng người Pháp Blaise Pascal đã cùng một người bạn là Pierre de Fermat nghiên cứu để tìm ra đáp án cho câu hỏi của Méré, tất nhiên khi bắt tay vào nghiên cứu Pascal và người cộng sự của mình đã tự đặt ra những câu hỏi phức tạp hơn, và cứ thế các nghiên cứu chính thức về xác suất đã được bắt đầu.



Trở lại với vấn đề của chúng ta, các trò chơi có tính may rủi như quay số, rút bài, tung đồng xu... cho chúng ta khái niệm về “phép thử” (còn có thể gọi là thí nghiệm ngẫu nhiên), khi phép thử được thực hiện có thể dẫn đến nhiều kết cục khác nhau, nhưng thông thường ta không thể nào tiên đoán chính xác kết cục nào sẽ xảy ra trước khi thực hiện phép thử mặc dù ta vẫn biết sẽ có những kết cục nào, như với trò tung súc sắc trên đây chúng ta không biết trước là mặt có mấy chấm sẽ ngửa lên nhưng ta có thể chắc chắn số chấm của mặt ngửa sẽ không thể lớn hơn 6 do ta biết cấu trúc của con súc sắc. Tuy nhiên lưu ý rằng nếu lập lại phép thử trên nhiều lần, chẳng hạn n lần và gọi m lần là số lần thành công, ví dụ là số lần mà có mặt 3 chấm ngửa lên, thực nghiệm cho thấy rằng tỉ số $f = m/n$ sẽ tiến tới một giới hạn ổn định nếu như số lần tung súc sắc n ngày càng lớn. Tính ổn định trên chính là nền tảng của lý thuyết xác suất.

5.1.2 Phép thử và biến cố

5.1.2.1 Các định nghĩa

Người ta tổng kết các khái niệm sau:

Việc thực hiện một nhóm các điều kiện cơ bản để quan sát một hiện tượng nào đó có xảy ra hay không được gọi là thực hiện một phép thử, còn kết cục của phép thử đó gọi là biến cố. Trong bài toán xác suất người ta hay đặt tên cho biến cố bằng các kí hiệu viết tắt cho ngắn gọn.

Ví dụ: Tung một đồng xu cân đối và đồng chất là làm phép thử, kết cục mặt số lật lên được gọi là biến cố và ta gọi tên là “biến cố được mặt số”, gọi vậy để phân biệt với kết cục nếu mặt hình lật lên, nó cũng là một biến cố khác của phép thử này và gọi là “biến cố được mặt hình”. Ta có thể quy ước tên gọi của hai biến cố này là biến cố S và biến cố H thay cho chữ “biến cố được mặt số” và “biến cố được mặt hình”.

Biến cố sơ cấp là một kết cục sơ đẳng nhất của phép thử. Còn biến cố cũng là kết cục của phép thử, nhưng nó là một tập hợp các biến cố sơ cấp có chung đặc tính. Như vậy một biến cố có thể là một tập hợp của một số biến cố sơ cấp, biến cố cũng có thể chỉ bao gồm một biến cố sơ cấp duy nhất (tất nhiên lúc này ta chỉ đơn giản gọi là biến cố), và cũng có khi biến cố là một tập hợp rỗng tức nó là một biến cố không thể có (tức là không bao giờ xảy ra được).

Chúng ta vận dụng những định nghĩa trên vào ví dụ tung con súc sắc để làm sáng tỏ:

- Biến cố sơ cấp là một kết cục sơ đẳng nhất của phép thử, vậy biến cố sơ cấp khi tung con súc sắc có thể là biến cố xuất hiện mặt 2 chấm, hoặc là biến cố xuất hiện mặt 5 chấm hay biến cố xuất hiện mặt 6 chấm (có thể kí hiệu gọn là biến cố 2 hay biến cố 5 hoặc biến cố 6).
- Biến cố xuất hiện các mặt có số chấm chẵn được đặt là C, vậy biến cố C là tập hợp các biến cố sơ cấp xuất hiện mặt có số chấm là số chẵn, vậy thì $C = \{2, 4, 6\}$, lúc này C không phải là biến cố sơ cấp.
- Biến cố xuất hiện mặt có số chấm lớn hơn 6 chấm là một biến cố rỗng vì không có biến cố sơ cấp nào tạo nên nó.

Xác suất của một biến cố là một con số đặc trưng cho khả năng xảy ra biến cố đó khi thực hiện phép thử, ví dụ theo trực giác ta luôn tin rằng khả năng xuất hiện mặt số khi tung một đồng xu là 50%, như vậy ta có thể nói “xác suất của biến cố xuất hiện mặt số khi tung đồng xu là 0,5”.

Xác suất được kí hiệu bằng chữ P, nên ta kí hiệu vẫn tắt câu phát biểu trên theo cách viết của xác suất là $P(S) = 0,5$.

Tập hợp của tất cả các kết cục sơ đẳng có thể xảy ra trong một phép thử được gọi là không gian mẫu. Ví dụ tập hợp tất cả các mặt có thể xuất hiện khi tung con súc sắc tạo thành không gian mẫu, nếu kí hiệu không gian mẫu là S thì không gian mẫu của phép thử tung con súc sắc có thể được kí hiệu một cách ngắn tắt là $S = \{1, 2, 3, 4, 5, 6\}$

Người ta phân biệt các biến cố như sau:

- **Biến cố chắc chắn:** là biến cố luôn xảy ra khi thực hiện phép thử, kí hiệu biến cố chắc chắn bằng chữ Ω . Ví dụ phép thử là thả một con súc sắc, “biến cố mặt lật lên có số chấm nhỏ hơn hoặc bằng 6” là một biến cố chắc chắn (vì cấu trúc của con súc sắc là như vậy)
- **Biến cố không thể có :** là biến cố nhất định không xảy ra khi thực hiện phép thử, kí hiệu biến cố không thể có là \emptyset . Ví dụ phép thử là chọn một tờ trong một block lịch mới, biến cố chọn được tờ lịch để ngày 30 tháng 2 là biến cố không thể có.
- **Biến cố ngẫu nhiên :** là biến cố không phải chắc chắn cũng không phải không thể có, cụ thể hơn, khi thực hiện phép thử biến cố ngẫu nhiên có thể xảy ra và cũng có thể không xảy ra. Người ta hay thích kí hiệu biến cố ngẫu nhiên bằng các chữ cái A, B, C... Ví dụ phép thử gieo con xúc xắc, biến cố xuất hiện mặt có 6 chấm là biến cố ngẫu nhiên vì có thể ra mặt 6 chấm mà cũng có thể ra mặt chấm khác.

Tất cả các biến cố chúng ta gặp trong thực tế đều thuộc một trong 3 loại biến cố trên, không có tình huống nào khác, tuy nhiên biến cố ngẫu nhiên là biến cố thường gặp hơn cả

5.1.2.2 Một số loại quan hệ giữa các biến cố

Vì một biến cố cũng là một tập hợp nên chúng ta có thể tổ hợp các biến cố thành các biến cố mới theo những phép tính về tập hợp như sau:

i) **Biến cố tổng $C = A \cup B$ hay $C = A + B$** là biến cố xảy ra khi và chỉ khi có ít nhất một trong hai biến cố thành phần xảy ra

Ví dụ : có hai người thợ săn cùng ngắm bắn một con thú, gọi A là biến cố người thứ nhất bắn trúng con thú, B là biến cố người thứ 2 bắn trúng con thú và C là biến cố con thú trúng đạn. Vậy $C = A \cup B$.

ii) **Biến cố tích $C = A \cap B$ hay $C = A * B$** là biến cố xảy ra khi và chỉ khi A và B cùng xảy ra.

Ví dụ : có hai người thợ săn cùng ngắm bắn một con thú, gọi A là biến cố người thứ nhất bắn trượt con thú, B là biến cố người thứ 2 bắn trượt con thú, C là biến cố con thú không trúng đạn. Vậy $C = A \cap B$

iii) Biến cố xung khắc: hai biến cố A và B được gọi là xung khắc khi và chỉ khi chúng không thể đồng thời xảy ra trong một phép thử (hay $A \cap B = \emptyset$ tức là việc A và B cùng xảy ra là không thể có)

Ví dụ : Trên bia có vẽ các vòng đồng tâm đánh số, một xạ thủ nhảm bắn 1 viên đạn vào bia. Đặt

N là biến cố xạ thủ bắn trúng vòng 5 điểm

B là biến cố xạ thủ bắn trúng vòng 7 điểm

Thì N và B là hai biến cố xung khắc vì một phát đạn không thể vừa gim trúng vòng 5 vừa gim trúng vòng 7 của tấm bia.

iv) Các biến cố xung khắc từng đôi: Khi áp dụng khái niệm xung khắc cho nhóm gồm n biến cố ta có khái niệm xung khắc từng đôi như sau :

n biến cố A_1, A_2, \dots, A_n được gọi là xung khắc từng đôi nếu như 2 biến cố bất kỳ trong n biến cố này xung khắc với nhau, (hay $A_i \times A_j = \emptyset \quad \forall i \neq j$)

Ví dụ : Tung 1 con súc sắc, đặt A_i ($i=1,6$) là biến cố mặt có số chấm i xuất hiện, nhóm A_i là nhóm xung khắc từng đôi vì bất kì hai biến cố nào trong nhóm này cũng xung khắc với nhau, cụ thể (A_1 xung khắc A_2); (A_2 xung khắc A_5); (A_1 xung khắc A_6)...

v) Biến cố đối lập : Biến cố “không xảy ra biến cố A” (kí hiệu \bar{A}) gọi là biến cố đối lập với biến cố A. Như vậy trong phép thử bắt buộc có một và chỉ được một trong A và \bar{A} xảy ra.

Ta kí hiệu $A \cap \bar{A} = \emptyset$ (cả A lẫn \bar{A} là không thể có)

$A \cup \bar{A} = \Omega$ (A hoặc \bar{A} là chắc chắn)

Nhận xét :

- Hai biến cố đối lập thì xung khắc (vì không thể xảy ra đồng thời) nhưng hai biến cố xung khắc chưa chắc đối lập
- Với hai biến cố đối lập : $P(A \cup \bar{A}) = P(\Omega) = 1$ (vì $P(\Omega) = 1$)

Ví dụ : Trong phép thử là một ca sinh bình thường (không tính ca sinh đôi), biến cố sinh trai và biến cố sinh gái (gọi cách khác là biến cố không sinh trai) là đối lập, chúng cũng là biến cố xung khắc (vì một ca sinh chỉ có thể là trai hay gái)

vi) Biến cố độc lập : Hai biến cố A và B được 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 biến cố kia và ngược lại.

Ví dụ : có hai bà mẹ cùng sinh con tại nhà hộ sinh X trong buổi sáng ngày tết Dương lịch, biết xác suất sinh trai tự nhiên mỗi ca sinh là 0,52. Vậy thì

xác suất của biến cố bà mẹ thứ hai sinh trai không chịu ảnh hưởng gì của biến cố bà mẹ thứ nhất sinh gái hay trai (và ngược lại). Do đó nếu ta đặt

T_1 là biến cố con của bà mẹ thứ nhất là trai

T_2 là biến cố con của bà mẹ thứ hai là trai

Thì T_1 và T_2 là hai biến cố độc lập

vii) Nhóm đầy đủ các biến cố (còn gọi là nhóm đầy đủ và xung khắc từng đôi): Các biến cố H_1, H_2, \dots, H_n được gọi là nhóm đầy đủ các biến cố nếu trong kết quả của một phép thử sẽ chắc chắn (tính đầy đủ) xảy ra một và chỉ một (tính xung khắc) trong các biến cố đó.

Ví dụ khi tung con súc sắc, gọi A_i là biến cố xuất hiện mặt i chấm ($i = 1, 6$), thì các A_i lập thành nhóm đầy đủ các biến cố.

viii) Biến cố đồng khả năng: Trong một số trường hợp do tính đối xứng của nhóm điều kiện xác định phép thử mà một số biến cố có khả năng khách quan để xuất hiện như nhau, ta gọi các biến cố đó là biến cố đồng khả năng.

Ví dụ : nếu một con súc sắc cân đối và đồng chất thì khả năng ra mặt nào trong 6 mặt khi nó rơi xuống cũng đều bằng nhau và bằng $1/6$. Vậy 6 biến cố này là 6 biến cố đồng khả năng.

5.1.3 Tính xác suất theo các định nghĩa về xác suất

Giả sử có biến cố A và chúng ta kí hiệu $P(A)$ là xác suất của nó, xác suất $P(A)$ đại diện cho khả năng xuất hiện của biến cố A. Chúng ta sẽ tính xác suất $P(A)$ này theo hai định nghĩa là định nghĩa theo quan điểm cổ điển và định nghĩa theo quan điểm thống kê.

5.1.3.1 Định nghĩa cổ điển về xác suất

Trong một phép thử có n kết cục đồng khả năng và xung khắc, trong đó có m kết cục thuận cho biến cố A xuất hiện thì (theo định nghĩa cổ điển) xác suất của biến cố A là tỷ số $P(A) = m/n$

Ví dụ chúng ta rút một lá bài từ bộ bài 52 lá. Ở đây ta có 52 kết quả đồng khả năng (nếu như bộ bài được đảo cẩn thận) và dĩ nhiên là xung khắc (vì ta chỉ có thể rút được duy nhất một lá trong phép thử của chúng ta). Xác suất của biến cố ta rút được lá "đầm bích" là $1/52$ vì chỉ có duy nhất 1 lá đầm bích trong bộ bài. Muốn biết xác suất của biến cố rút được một lá bích bất kỳ (biến cố A), ta xác định số kết cục thuận lợi cho A là 13 vì có 13 lá bích trong một bộ bài $\rightarrow P(A) = 13/52$.

Trong những trường hợp số kết cục của phép thử rất lớn mà không thể dùng lối suy đoán trực tiếp như trên thì người ta vận dụng các công thức

của giải tích tổ hợp như chỉnh hợp, hoán vị, tổ hợp để giải quyết bài toán xác suất.

Ví dụ: một người khi gọi điện thoại quên mất 3 số cuối và chỉ nhớ được rằng đây là 3 số khác nhau. Tính xác suất chỉ quay một lần mà trúng số định gọi?

Khi quay ngẫu nhiên 3 con số cuối là làm một phép rút ngẫu nhiên 3 con số từ 10 con số (từ 0 – 9) và trật tự sắp xếp 3 số này có tính khác biệt (ví dụ quay số 120 được kết quả khác với quay số 201) nên ta đang làm các chỉnh hợp chapter 3 của 10 phần tử, vậy tính ra ta có $A_{10}^3 = 720$ cách quay 3 con số cuối.

Gọi Đ là biến cố quay một lần trúng số định gọi. Số biến cố sơ cấp đồng khả năng và xung khắc có thể xảy ra là tất cả các phương thức có thể lập lên 3 con số khác nhau từ 10 con số, ta đã tính được ở trên tức $n = 720$

Số biến cố sơ cấp thuận lợi cho Đ thì chỉ có 1, do đó theo định nghĩa cổ điển

$$P(D) = 1/720 = 0,00138$$

Định nghĩa cổ điển trên không tổng quát và không áp dụng được khi số kết quả của phép thử có thể nhiều vô tận, hạn chế khác là ta sẽ không xác định được xác suất nếu các kết cục không đồng khả năng. Tất nhiên trong những trường hợp thoả mãn các điều kiện trên thì định nghĩa cổ điển sẽ giúp chúng ta tính xác suất một cách chính xác mà không cần phải thực hiện phép thử, chỉ cần dựa trên tư duy logic (bạn hãy chú ý điều này để rồi so sánh với định nghĩa thống kê về xác suất). Với các bài toán xác suất chúng ta gặp trong thế giới thực thì những tình huống có điều kiện như vậy không nhiều, giả dụ bạn hiếm khi có các biến cố sơ cấp đồng khả năng khi thực hiện một phép thử, thử hình dung khi bạn lên kế hoạch về một công việc kinh doanh, không gian mẫu của bạn lúc này bao gồm 3 biến cố sơ cấp là thành, huề vốn và bại, $S = \{\text{thành}, \text{huề}, \text{bại}\}$, kết quả có hợp lý không nếu bạn dùng định nghĩa cổ điển về xác suất để xác định khả năng thành công của việc kinh doanh?

Nếu theo định nghĩa cổ điển thì $P(\text{thành}) = 1/3$ có nghĩa là xác suất thành công là 0,333. Và nếu trình tự chúng ta lập luận là đúng thì cơ hội thành công của bất kỳ một kế hoạch làm ăn nào cũng là 33,3%. Dĩ nhiên là không phải thế, do đó các biến cố sơ cấp thành hay bại không thể có cùng khả năng xảy ra. Và do đó chúng ta cần có phương pháp khác để tính toán xác suất cho những tình huống tương tự như thế này.

5.1.3.2 Định nghĩa thống kê về xác suất (định nghĩa dựa trên kết quả thực nghiệm)

Theo cách tiếp cận này mặc dù xác suất vẫn được xác định như tỷ số giữa số kết cục ta quan tâm trên tổng số kết cục, tuy nhiên những kết cục này được căn cứ trên dữ liệu quan sát được qua việc thực hiện phép thử chứ không căn cứ trên suy luận logic nhờ đã hiểu được tiến trình từ trước.

Định nghĩa thống kê về xác suất xác định tần suất xuất hiện biến cố A như tỷ số giữa số lần biến cố A xảy ra chia cho số lần phép thử được thực hiện lặp lại nhiều lần.

$$f_{(A)} = \frac{m}{n}$$

Ký hiệu

- A là biến cố quan tâm
- n là số lần thực hiện phép thử
- m là số lần biến cố A xuất hiện trong n lần thực hiện phép thử
- $f_{(A)}$ là tần suất xuất hiện biến cố A

Người ta tin rằng khi số lần thực hiện lặp lại phép thử đủ nhiều và được tiến hành trong những điều kiện giống nhau thì tần suất của biến cố A sẽ có xu hướng dao động quanh một con số gọi là tần suất lý thuyết của biến cố (và mức độ dao động sẽ càng nhỏ đi nếu số lần thực hiện phép thử càng tăng lên). Người ta gọi tần suất lý thuyết của biến cố chính là xác suất của biến cố.

Định nghĩa thống kê về xác suất không đòi hỏi chặt chẽ như định nghĩa cổ điển là các kết cục phải đồng khả năng và xung khắc, tuy nhiên nó đòi hỏi phải thực hiện phép thử chứ không suy luận được.

Ví dụ : Để nghiên cứu khả năng xuất hiện mặt số khi tung đồng xu cổ đúng là 0,5 hay không (không phải bằng định nghĩa cổ điển nữa mà bằng phương pháp thống kê) 3 nhà khoa học đã tiến hành các thí nghiệm tung một đồng xu rất nhiều lần và thu được kết quả sau :

Người làm thí nghiệm	Số lần tung đồng xu (n)	Số lần xuất hiện mặt số (k)	Tần suất = k/n
Buffon	4.040	2048	0,5069
Pearson	12.000	6019	0,5016
Pearson	24.000	12012	0,5005

Ví dụ trên chứng tỏ khi số phép thử tăng lên thì tần suất xuất hiện mặt số sẽ dao động ngày càng ít quanh giá trị 0,5, điều đó cho ta hy vọng rằng

khi số phép thử tăng lên vô hạn, tần suất sẽ hội tụ về giá trị 0,5. Và ta kết luận được rằng xác suất xuất hiện mặt số khi tung đồng xu là 0,5; điều này cũng phù hợp xác suất tính theo định nghĩa cổ điển về xác suất.

Có thể lấy ví dụ thực tế sau: khi tiến hành điều tra mức độ ủng hộ các ứng cử viên đang tranh cử Tổng thống Mỹ, tổ chức điều tra muốn xác định tỷ lệ cử tri ủng hộ ứng cử viên của đảng Cộng hoà. Lúc này, đặt biến cố A là việc gặp người trả lời ủng hộ ứng cử viên của đảng Cộng hoà, mỗi lần phỏng vấn một cử tri là một lần thực hiện lặp lại phép thử giống nhau nên số lần thực hiện phép thử chính là số người được phỏng vấn (cho rằng phỏng vấn 10.000 người là số phép thử đủ lớn), số lần nhận được câu trả lời ủng hộ ứng cử viên của đảng Cộng hoà chính là số lần biến cố A xuất hiện. Nếu mẫu được chọn cho điều tra cho biết gặp 6070 cử tri ủng hộ cho ứng cử viên của đảng Cộng hoà, thì tần suất xuất hiện biến cố A là 60,7%, ta kết luận $P(A) = 0,607$.

5.1.4 Một vài tính chất của xác suất

Gọi A là biến cố đang định xét xác suất $P(A)$ thì từ định nghĩa về xác suất ta suy ra xác suất $P(A)$ có một số tính chất cơ bản sau đây

i) Với mọi biến cố ngẫu nhiên A : $0 \leq P(A) \leq 1$

Chúng ta đã định nghĩa xác suất như là tần suất gấp biến cố A trên tổng số lần thực hiện phép thử, hoặc nếu theo định nghĩa cổ điển về xác suất thì là tỷ số giữa số lần thuận lợi cho biến cố A xuất hiện trên tổng số kết cục đồng khả năng và xung khắc của phép thử, tức là $P(A) = m/n$; vì số kết cục thuận lợi cho biến cố A tức m luôn thỏa mãn $0 \leq m \leq n$, do đó chia cả 3 vế cho n ta được $0/n \leq m/n \leq n/n \Leftrightarrow 0 \leq P(A) \leq 1$.

ii) Một biến cố chắc chắn xảy ra có xác suất bằng 1 kí hiệu $P(\Omega) = 1$, ví dụ nếu A là biến cố xuất hiện mặt có số chấm nhỏ hơn hoặc bằng 6 khi tung con súc sắc thì $P(A) = 1$.

iii) Một biến cố không thể xảy ra có xác suất bằng 0, kí hiệu $P(\emptyset) = 0$, ví dụ nếu là biến cố xuất hiện mặt có số chấm lớn hơn 6 khi tung con súc sắc thì $P(A) = 0$.

iv) Nếu A_1, A_2, \dots, A_n là một nhóm đầy đủ các biến cố thì

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

Tổng của nhóm đầy đủ các biến cố là một biến cố chắc chắn vì theo bản chất của nhóm đầy đủ các biến cố thì sự hợp của chúng là một biến cố

chắc chắn, hay tổng của chúng là một biến cố chắc chắn, mà biến cố chắc chắn có xác suất bằng 1.

Ví dụ phép thử là Bỏ vốn vào kinh doanh, đặt

A là biến cố kết cục lời

B là biến cố kết cục huề vốn

C là biến cố kết cục lỗ

A, B, C là nhóm đầy đủ các biến cố

Ta thấy $(A + B + C) = \Omega$ (chắc chắn phải có gì đó xảy ra vì không kết cục lời thì kết cục lỗ hoặc kết cục huề vốn)

Mà $P(\Omega) = 1$ nên $P(A + B + C) = 1$

5.1.5 Tính xác suất theo các quy tắc xác suất

5.1.5.1 Quy tắc cộng xác suất

Quy tắc cộng đơn giản

Nếu A và B là hai biến cố xung khắc của một phép thử thì xác suất của tổng hai biến cố bằng tổng xác suất của từng biến cố

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

Hay $P(A \cup B) = P(A) + P(B)$

Ví dụ : chọn ngẫu nhiên 1.000 sản phẩm ngũ cốc đã đóng gói của mẻ sản xuất ngũ cốc trong ngày đầu tiên sau khi bảo trì hệ thống dây chuyền đóng gói, cân 1.000 gói này thì thấy 10 gói thiếu trọng lượng và 30 gói có trọng lượng dư so với trọng lượng quy định trong thiết kế kỹ thuật. Hãy tính xác suất của biến cố dây chuyền đóng gói trọng lượng sai thiết kế?

Đặt

S là biến cố gói cân lên có trọng lượng sai so với thiết kế

D là biến cố gói cân lên trọng lượng dư so với thiết kế

T là biến cố gói cân lên trọng lượng thiếu so với thiết kế

Ta thấy $S = D + T$ (vì dư hay thiếu đều là sai) $\rightarrow P(S) = P(D + T)$

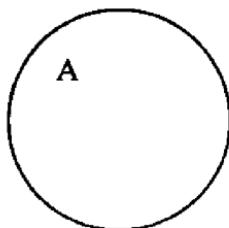
Mà D và T là hai biến cố xung khắc (vì một gói cân được không thể vừa dư vừa thiếu trọng lượng) nên

$$P(S) = P(D) + P(T) = 10/1000 + 30/1000 = 0,04$$

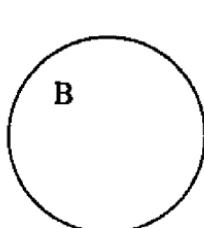
Ta vẽ lên bảng hai vòng tròn A và B như trong Hình 5.1 dưới đây. Gọi A là biến cố phỏng mũi tên lọt vào vòng tròn A và B là biến cố phỏng mũi tên lọt vào vòng tròn B. Theo một nghĩa nào đó thì xác suất của biến cố phỏng mũi tên lọt vào vòng tròn tỷ lệ với diện tích của vòng tròn, như

vậy gọi $P(A)$ là xác suất của biến cố phỏng mũi tên vào vòng A và $P(B)$ là xác suất của biến cố phỏng mũi tên vào vòng B. Chú ý rằng một mũi tên ta phỏng không thể đồng thời lọt vào cả hai vòng tròn A và B, do đó A và B là hai biến cố xung khắc, và xác suất để phỏng mũi tên lọt vào vòng tròn A hoặc B sẽ là tổng số của hai diện tích. Nghĩa là $P(A+B) = P(A) + P(B)$.

Hình 5.1



Hình 5.2



Bây giờ ta xem tiếp Hình 5.2, ở đây A và B có một vùng chung. Mũi tên ta phỏng lúc này có thể lọt vào vùng chung này và dĩ nhiên lúc đó nó được xem là vừa lọt vào vòng A vừa lọt vào vòng B, do đó A và B không phải là hai biến cố xung khắc. Lúc này nếu ta tính xác suất của biến cố $A \cup B$ bằng cách cộng hai diện tích hai vòng tròn như trên thì thành ra ta sẽ cộng hai lần vùng chung của chúng. Do đó chúng ta phải trừ đi một lần diện tích vùng chung này ra khỏi tổng hai diện tích, vùng chung này tương ứng với xác suất của A và B đồng thời xảy ra (biến cố $A \cap B$)

Vậy ta có $P(A+B) = P(A) + P(B) - P(A \cap B)$

Từ hình vẽ 5.2 ta có thể suy rộng công thức cộng đơn giản thành quy tắc cộng tổng quát như sau.

Quy tắc cộng tổng quát

Nếu A và B là hai biến cố bất kì có thể xảy ra trong một phép thử, xác suất của tổng hai biến cố A và B bằng tổng hai xác suất của từng biến cố trừ đi xác suất của tích hai biến cố đó.

Kí hiệu

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\text{Hay } P(A+B) = P(A) + P(B) - P(A \cdot B)$$

Ví dụ : Lớp có 50 sinh viên trong đó có 20 sinh viên giỏi anh văn, 15 sinh viên giỏi pháp văn và 7 sinh viên giỏi cả hai ngoại ngữ. Chọn ngẫu nhiên 1 sinh viên trong lớp, tính xác suất sinh viên này giỏi ít nhất một ngoại ngữ?

Đặt:

A là biến cố sinh viên chọn được giải anh văn

P là biến cố sinh viên chọn được giải pháp văn

C là biến cố sinh viên chọn được giải ít nhất một ngoại ngữ
ta thấy $C = A + P \rightarrow P(C) = P(A + P)$

mà A và P không xung khắc nhau vì có thể đồng thời xảy ra việc ta chọn được 1 sinh viên giỏi cả hai ngoại ngữ nên ta áp dụng công thức cộng tổng quát

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

Với: $P(A) = 20/50$; $P(P) = 15/50$; $P(A.P) = 7/50$

$$\rightarrow P(C) = P(A) + P(P) - P(A.P) = 20/50 + 15/50 - 7/50 = 28/50 = 0,56$$

5.1.5.2 Quy tắc nhân xác suất

Quy tắc nhân đơn giản

Nếu hai biến cố A và B độc lập với nhau, thì xác suất của A và B cùng xảy ra là tích xác suất của A với xác suất của B. Trong trường hợp này về mặt toán học ta viết

$$P(A*B) = P(A) * P(B)$$

Hay $P(A \cap B) = P(A) * P(B)$

Ví dụ phép thử của chúng ta là tung đồng thời hai đồng xu, muốn tính xác suất xuất hiện mặt sấp ở đồng xu thứ nhất và mặt ngửa ở đồng xu thứ hai, ta gọi

- A là biến cố xuất hiện mặt sấp ở đồng xu thứ nhất
- B là biến cố xuất hiện mặt ngửa ở đồng xu thứ hai, khi đó A và B là hai biến cố độc lập hoàn toàn với nhau

Thì $A \cap B$ là biến cố xuất hiện mặt sấp ở đồng xu thứ nhất và mặt ngửa ở đồng xu thứ hai

$$P(A \cap B) = P(A) * P(B) = 1/2 * 1/2 = 1/4$$

Quy tắc nhân tổng quát

* * * Trước khi tìm hiểu quy tắc nhân tổng quát ta phải có khái niệm về xác suất điều kiện:

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

Ví dụ : Trong bình có 3 cầu trắng và 2 cầu đen. Lấy ngẫu nhiên lần lượt 2 quả cầu. Tính xác suất để lần thứ 2 lấy được cầu trắng nếu biết rằng lần một đã lấy được cầu trắng?

Gọi T_1 là biến cố lấy lần 1 được cầu trắng

T_2 là biến cố lần thứ 2 lấy được cầu trắng.

Sau khi lần thứ nhất lấy được cầu trắng (biến cố T_1 đã xảy ra) thì trong bình còn 4 quả cầu trong đó 2 cầu màu trắng. Do đó xác suất có điều kiện của T_2 bằng

$$P(T_2/T_1) = 2/4 = 1/2.$$

* * * Từ đó ta phát biểu Quy tắc nhân tổng quát như sau

Xác suất của tích hai biến cố không độc lập 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.

$$P(A \cdot B) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B)$$

Chúng ta để ý rằng trong trường hợp các biến cố độc lập với nhau thì $P(B/A)$ cũng bằng chính $P(B)$ vì xác suất xảy ra biến cố B không bị ảnh hưởng bởi sự xảy ra hay không xảy ra của A, lúc này công thức của Quy tắc nhân tổng quát lại quay về công thức của Quy tắc nhân đơn giản

Câu hỏi đặt ra là : Khi nào ta chọn biến cố điều kiện là biến cố A hoặc biến cố B. Điều này phụ thuộc vào việc ta dễ tính $P(A/B)$ hơn hay là $P(B/A)$ hơn.

Ví dụ : từ một hộp chứa 4 trái banh trắng và 6 trái banh đen chúng ta lấy không hoàn lại hai trái liên tiếp. Xác suất của hai trái banh lấy ra đều trắng là bao nhiêu?

Biến cố trái thứ nhất lấy ra có màu trắng được đặt tên là A, xác suất của biến cố này là $4/10$

Biến cố trái thứ hai lấy ra có màu trắng được đặt tên là B, xác suất của biến cố này trong tình huống trái banh thứ nhất trắng được tính là $P(B/A) = 3/9$ (vì chỉ còn 3 trái banh trắng trong số 9 trái còn lại)

Vậy xác suất của biến cố hai trái banh lấy ra đều trắng là $P(A \cdot B)$ mà A và B không độc lập nên

$$P(A \cdot B) = P(A) \times P(B/A) = 4/10 * 3/9 = 2/15.$$

5.1.5.3 Quy tắc xác suất đầy đủ

Xét một phép thử có các kết cục $H_1; H_2; \dots; H_n$ tạo thành nhóm đầy đủ các biến cố.

Giả sử biến cố A liên quan đến phép thử này (A có thể xảy ra đồng thời với chỉ một trong các biến cố $H_1; H_2; \dots; H_n$)

Đã biết các $P(H_i)$ và các $P(A/H_i)$

Lúc đó xác suất của biến cố A được tính bằng công thức xác suất đầy đủ sau đây:

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

Chứng minh

Vì các biến cố $H_1; H_2; \dots; H_n$ tạo nên một nhóm đầy đủ các biến cố nên biến cố A chỉ có thể xảy ra đồng thời với duy nhất 1 trong các biến cố này, như vậy không gian mẫu của chúng ta là $[H_1.A \cup H_2.A \cup \dots \cup H_n.A]$

Vậy áp dụng quy tắc cộng đơn giản ta được

$$P(A) = P(H_1.A + H_2.A + \dots + H_n.A) = P(H_1.A) + P(H_2.A) + \dots + P(H_n.A)$$

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

Theo quy tắc nhân xác suất tổng quát thì $P(H_i.A) = P(H_i) \times P(A/H_i)$

Cho nên $P(A) = \sum_{i=1}^n [P(H_i) \times P(A/H_i)]$, đó là điều phải chứng minh.

Ví dụ : Có 3 hộp giống nhau, hộp thứ nhất đựng 10 sản phẩm trong đó có 4 phế phẩm; hộp thứ 2 đựng 15 sản phẩm trong đó có 5 phế phẩm; hộp thứ 3 đựng 20 sản phẩm trong đó có 5 phế phẩm. Chọn ngẫu nhiên lấy một hộp và từ hộp chọn được lấy ngẫu nhiên ra một sản phẩm, hãy tính xác suất chọn được chính phẩm.

Đặt:

C là biến cố lấy được chính phẩm, biến cố C có thể xảy ra đồng thời với chỉ một trong 3 biến cố sau đây:

H_1 : biến cố sản phẩm lấy ra thuộc hộp 1

H_2 : biến cố sản phẩm lấy ra thuộc hộp 2

H_3 : biến cố sản phẩm lấy ra thuộc hộp 3

$\rightarrow H_i$ tạo thành nhóm đầy đủ các biến cố

Mà các H_i đồng khả năng nên ta tính ngay được $P(H_1) = P(H_2) = P(H_3) = 1/3$

Tính xác suất lấy được chính phẩm với điều kiện biết nó được chọn từ hộp nào là tính các xác suất có điều kiện sau:

$$P(C/H_1) = 6/10$$

$$P(C/H_2) = 10/15$$

$$P(C/H_3) = 15/20$$

Cuối cùng vì biến cố C có thể xảy ra đồng thời với một trong các biến cố $H_1; H_2; H_3$. Nhóm H_i là nhóm đầy đủ các biến cố nên xác suất của biến cố C được tính bằng công thức xác suất đầy đủ sau đây:

$$P(C) = \sum_{i=1}^3 [P(H_i) \times P(C/H_i)]$$

$$= 1/3 \times 6/10 + 1/3 \times 10/15 + 1/3 \times 15/20 = 0,672$$

5.1.5.4 Định lý Bayes

Xét một phép thử có các kết cục $H_1; H_2; \dots; H_n$ tạo thành nhóm đầy đủ các biến cố.

Giả sử biến cố A liên quan đến phép thử này (A có thể xảy ra đồng thời với chỉ một trong các biến cố $H_1; H_2; \dots; H_n$)

Để tính xác suất của biến cố H, với điều kiện biến cố A đã xảy ra tức tính $P(H_i/A)$ ta dùng công thức xác suất Bayes như sau

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

Chứng minh

Theo quy tắc nhân xác suất tổng quát ta có

$$P(A \cdot H_i) = P(A) \cdot P(H_i/A) = P(H_i) \cdot P(A/H_i) \quad (i=1, n)$$

Từ $P(A) \cdot P(H_i/A) = P(H_i) \cdot P(A/H_i)$ ta hoán vị và được kết quả

$$P(H_i/A) = P(H_i) \cdot P(A/H_i) / P(A) \quad (*)$$

Mà theo công thức đầy đủ thì

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

Ráp các thành phần vào (*) ta có

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

Chú ý trong công thức Bayes mẫu số chính là $P(A)$ tính theo công thức xác suất đầy đủ và tử số là một thành phần tạo nên mẫu số đó, cũng theo công thức thì tử số là công thức nhân xác suất tổng quát của biến cố ($H_i A$) trong đó H_i là biến cố mà ta đang muốn tìm xác suất với điều kiện A đã xảy ra.

Như vậy bản chất 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.

Ví dụ 1: Dây chuyền đóng gói thành phẩm nhận được các sản phẩm do hai phân xưởng sản xuất. Trung bình phân xưởng thứ nhất cung cấp khoảng 60% sản phẩm, 40% còn lại do phân xưởng 2 đáp ứng. Kiểm tra kỹ thuật cho biết khoảng 90% sản phẩm từ phân xưởng 1 đạt tiêu chuẩn, và 85% sản phẩm từ phân xưởng 2 đạt tiêu chuẩn.

Bốc ngẫu nhiên một sản phẩm từ dây chuyền thấy nó là sản phẩm đạt yêu cầu. tính xác suất để sản phẩm đó từ phân xưởng 2 sản xuất

Đặt D là biến cố sản phẩm đạt yêu cầu, biến cố D có thể xảy ra đồng thời với chỉ một trong hai biến cố thuộc nhóm đầy đủ các biến cố sau đây

P_1 : sản phẩm do phân xưởng 1 sản xuất

P_2 : sản phẩm do phân xưởng 2 sản xuất

Để tính xác suất lấy được một sản phẩm từ phân xưởng 2 sản xuất mà là sản phẩm đạt yêu cầu ta phải xác định xác suất sau $P(P_2/D)$. Ta có điều kiện để áp dụng Công thức Bayes, như sau

$$P(P_2/D) = \frac{P(P_2)P(D/P_2)}{\sum_{i=1}^2 P(P_i)P(D/P_i)}$$

Tính các xác suất thành phần

$$P(P_1) = 0,6$$

$$P(D/P_1) = 0,9$$

$$P(P_2) = 0,4$$

$$P(D/P_2) = 0,85$$

Thay số liệu vào công thức

$$P(P_2/D) = \frac{0,4 \times 0,85}{(0,6 \times 0,9) + (0,4 \times 0,85)} = 0,386$$

Ta thấy trước khi thực hiện phép thử, xác suất chọn được sản phẩm từ phân xưởng 2 là 0,4; sau khi thực hiện phép thử và biết kết quả thì xác suất đó thay đổi, cụ thể, còn 0,386. Đó là do ta có thêm một điều kiện là sản phẩm từ phân xưởng 2 nhưng phải là sản phẩm tốt, mà tỷ lệ sản phẩm tốt của phân xưởng 2 không cao nên nó kéo xác suất này xuống.

Ví dụ 2: Trong 1 trường đại học, có 4% nam sinh viên và 1% nữ sinh viên cao hơn 1m65, ngoài ra ta còn biết 60% sinh viên của trường là nữ. Chọn

ngẫu nhiên 1 sinh viên cao hơn 1m65. Tính xác suất để sinh viên được chọn này là 1 nữ sinh viên.

Đặt: B là biến cố sinh viên chọn được cao hơn 1m65

M là biến cố sinh viên chọn được là nam

W là biến cố sinh viên chọn được là nữ

$\rightarrow M$ và W hợp thành hệ đầy đủ các biến cố

Biến cố B có thể xảy ra đồng thời với chỉ một trong hai biến cố thuộc nhóm đầy đủ các biến cố trên.

Chúng ta muốn tìm $P(W/B)$ là xác suất để cho 1 sinh viên cao hơn 1m65 được chọn là nữ sinh viên. Theo công thức Bayes, ta viết:

$$P(W/B) = \frac{P(W)P(B/W)}{P(W)P(B/W) + P(M)P(B/M)}$$

Chúng ta biết $P(W) = 0,6$ vì ta có 60% nữ sinh viên

$P(B/W) = 0,01$ vì có 1% nữ sinh viên cao hơn 1m65

$P(M) = 0,4$ vì có 40% nam sinh viên

$P(B/M) = 0,04$ vì có 4% nam sinh viên cao hơn 1m65

Vậy: $P(W/B) = \frac{0,6 \times 0,01}{0,6 \times 0,01 + 0,4 \times 0,04} = 0,273$

5.2 BIẾN NGẪU NHIÊN VÀ CÁC QUY LUẬT PHÂN PHỐI XÁC SUẤT

5.2.1 Biến ngẫu nhiên

Chúng ta tìm hiểu một ví dụ khá kinh điển sau đây để có thể hình thành khái niệm về biến ngẫu nhiên, đó là ví dụ về việc thực hiện phép thử tung một lượt hai đồng xu. Ta ký hiệu ngắn gọn S là tình huống đồng xu xuất hiện mặt sấp, N là tình huống đồng xu xuất hiện mặt ngửa, ký hiệu SN ám chỉ biến cố đồng xu thứ nhất rơi xuống sấp trong khi đồng thứ hai rơi xuống ngửa. Như vậy không gian mẫu của phép thử tung một lượt 2 đồng xu sẽ như sau:

$$S = \{SS, SN, NS, NN\}$$

Từ không gian mẫu này ta tổng kết có ba kết cục có thể xảy ra liên quan đến số mặt ngửa xuất hiện trong phép thử tung một lượt hai đồng xu là: không có mặt nào ngửa (SS), có một mặt ngửa (NS hoặc SN) và có hai mặt ngửa (NN).

Thay vì mô tả bằng lời về các kết cục của phép thử, ta tìm cách mô tả đơn giản và hữu ích hơn, đó là mô tả bằng con số, ta gọi X là đại lượng biểu thị số mặt ngửa xuất hiện khi hai đồng xu này rơi xuống, thì tương ứng với không gian mẫu trên đại lượng X có thể nhận ba trị số $X = 0$; $X = 1$; $X = 2$. Trị số mà X nhận được được xác định bởi các kết cục của phép thử nên ta có thể gọi X là biến số ngẫu nhiên, tức là nó có thể có các giá trị khác nhau tùy theo kết cục ngẫu nhiên nào của phép thử mà ta không biết được trước khi hai đồng xu rơi xuống. Ta hệ thống lại các kết cục của phép thử và trị số của biến X tương đương với kết cục đó trong bảng dưới đây

Bảng 5.1

Biến cố	Đồng xu thứ 1	Đồng xu thứ 2	X
A ₁	S	S	0
A ₂	S	N	1
A ₃	N	S	1
A ₄	N	N	2

Trong bảng trên ta để ý rằng mỗi biến cố chỉ cho ta một trị số của X (như vậy sự kiện X nhận các giá trị x_i tương ứng là các biến cố xung khắc), nhưng cùng một trị số của X có thể tương ứng với hai biến cố khác nhau, như X = 1 tương ứng với hai biến cố A₂ và A₃, hai biến cố này đều tạo ra cùng một giá trị X = 1 tức là có một mặt ngửa nhưng rõ ràng biến cố đồng xu thứ nhất rơi xuống sấp trong khi đồng thứ hai rơi xuống ngửa khác hẳn biến cố đồng thứ nhất ngửa trong lúc đồng thứ hai sấp.Ta đi đến định nghĩa về biến ngẫu nhiên.

5.2.1.1 Định nghĩa

Biến ngẫu nhiên là những biến mà giá trị của nó được xác định một cách ngẫu nhiên.

Về mặt toán học nếu mỗi biến cố sơ cấp A thuộc tập hợp biến cố ω nào đấy có thể đặt tương ứng với một đại lượng xác định X = X(A) thì X được gọi là một biến ngẫu nhiên. Biến ngẫu nhiên có thể xem như hàm của biến cố A với miền xác định là ω.

Các biến ngẫu nhiên thường được kí hiệu bằng các chữ cái viết hoa như X, Y, Z..còn các giá trị của chúng được kí hiệu bằng các chữ cái viết thường x, y, z..

5.2.1.2 Phân loại biến ngẫu nhiên

Biến ngẫu nhiên được chia thành 2 loại là biến ngẫu nhiên rời rạc và biến ngẫu nhiên liên tục.

Nếu giá trị của biến ngẫu nhiên X có thể lập thành dãy rời rạc các số x_1, x_2, \dots, x_n (dãy hữu hạn hay vô hạn) thì X được gọi là biến ngẫu nhiên rời rạc. Ví dụ lượng khách hàng đến siêu thị được theo dõi trong 4 ngày cuối tuần của một tháng là biến ngẫu nhiên rời rạc.

Nếu giá trị của biến ngẫu nhiên X có thể lấp đầy toàn bộ khoảng hữu hạn hay vô hạn của trục số Ox thì biến ngẫu nhiên X được gọi là liên tục. Ví dụ nhiệt độ đo tại sa mạc vào 12 giờ trưa trong nhiều ngày liên tục.

5.2.2 Phân phối xác suất của biến số ngẫu nhiên

Mục đích của thống kê là suy diễn từ những thông tin thu thập được trên một mẫu thành những sự hiểu biết về tổng thể, phương pháp này đòi hỏi sự hiểu biết về xác suất kết hợp với các trị số của biến ngẫu nhiên, nói nôm na là chúng ta phải xác định được các xác suất tương ứng với các giá trị có thể có của biến ngẫu nhiên để hoàn toàn xác định nó.

Như vậy chúng ta phải xác định được phân phối xác suất của biến số ngẫu nhiên, quy luật phân phối xác suất của biến ngẫu nhiên X là sự tương ứng giữa các giá trị có thể có của biến ngẫu nhiên và các xác suất tương ứng với các giá trị đó. Để minh họa cho phân phối xác suất của biến ngẫu nhiên chúng ta phát triển tiếp ví dụ tung hai con súc sắc ở trên, tiếp tục từ Bảng 5.1 ta tính toán xác suất tương đương với từng biến cố.

Bảng 5.2

Biến cố	Đồng xu thứ 1	Đồng xu thứ 2	$P(A_i)$	X
A_1	S	S	$\frac{1}{2} * \frac{1}{2} = \frac{1}{4}$	0
A_2	S	N	$\frac{1}{2} * \frac{1}{2} = \frac{1}{4}$	1
A_3	N	S	$\frac{1}{2} * \frac{1}{2} = \frac{1}{4}$	1
A_4	N	N	$\frac{1}{2} * \frac{1}{2} = \frac{1}{4}$	2

Cũng từ bảng trên chúng ta thấy rằng xác suất của kết cục $X = 1$ là sự kết hợp của hai biến cố xung khắc A_2 và A_3 dưới dạng tổng vì chỉ cần một trong hai biến cố A_2 hoặc A_3 xảy ra là đạt được kết cục $X = 1$, theo Quy tắc Cộng xác suất thì $P(X=1) = P(A_2 \cup A_3) = P(A_2) + P(A_3) = (\frac{1}{4} + \frac{1}{4}) = \frac{1}{2}$. Từ đó ta lập được bảng mô tả phân phối xác suất của biến ngẫu nhiên X biểu thị mối quan hệ giữa các giá trị có thể có của biến ngẫu nhiên với các xác suất tương ứng như sau:

Bảng 5.3

X	Biến cố tương ứng với X	P(X)
0	A ₁	1/4
1	A ₂ , A ₃	1/2
2	A ₄	1/4

Từ hiểu biết ở trên chúng ta tóm lược lại hai vấn đề cơ bản sau cần xác định về một biến ngẫu nhiên:

- Phải xác định được các giá trị có thể có của biến số (trong trường hợp biến rời rạc) hoặc khoảng giá trị có thể có của nó (trong trường hợp biến liên tục)
- Xác định xác suất để nó nhận mỗi một giá trị có thể có (trong trường hợp biến rời rạc) hoặc xác suất để nó nhận giá trị trong một khoảng giá trị (trong trường hợp biến liên tục) nào đó là bao nhiêu.

Tổng quát, bất kỳ một hình thức nào đó (mà thường là đồ thị hoặc bảng số hay công thức) biểu diễn mối quan hệ giữa các giá trị có thể có của một biến ngẫu nhiên và xác suất tương ứng của chúng thì đều được coi là hình thức biểu hiện quy luật phân phối xác suất của biến ngẫu nhiên ấy.

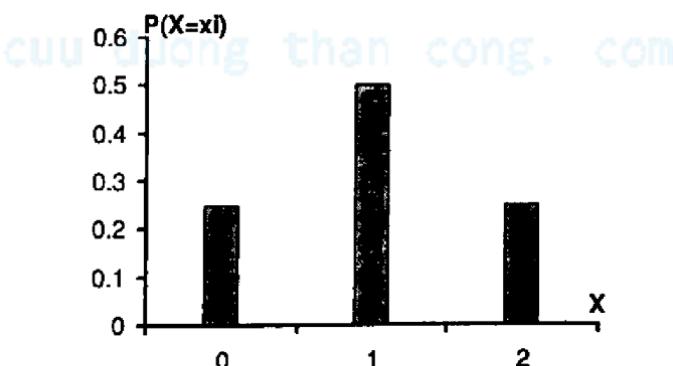
5.2.2.1 Phân phối xác suất của biến ngẫu nhiên rời rạc

Phân phối xác suất của biến ngẫu nhiên X thể hiện sự tương quan giữa các giá trị x_i của của X và các xác suất để X nhận giá trị x_i , sự tương quan đó có thể trình bày bằng bảng, đồ thị hay biểu thức.

Chính Bảng 5.3 là một bảng số biểu diễn quy luật phân phối xác suất của biến ngẫu nhiên rời rạc.

Nếu ta biểu diễn quy luật phân phối của X dưới dạng đồ thị thì đồ thị của chúng ta sẽ trông như Hình 5.3 dưới đây.

Hình 5.3



Trên Hình 5.3, xác suất rời rạc liên quan đến các giá trị mà biến số X nhận có thể được minh họa bằng diện tích các cột của đồ thị có chiều rộng bằng nhau và xem như bằng 1 đơn vị, như vậy chiều cao của cột chính là độ lớn của xác suất, nếu chúng ta cộng các xác suất ở cột $P(X)$ của Bảng 5.3 cũng như cộng các giá trị diện tích của các cột ở đồ thị trên ta đều được tổng số là 1. Điều này cũng có thể lý giải hợp lý bằng tính chất của xác suất là tất cả các giá trị mà biến số X có thể nhận hợp thành một nhóm đầy đủ các biến cố.

Phân phối xác suất của biến ngẫu nhiên rời rạc cũng có thể được mô tả bằng công thức nhưng chúng ta chưa bàn đến công thức trong phần này vì ví dụ chúng ta sử dụng còn rất thô sơ nên hoàn toàn có thể suy luận được mà chưa cần mô thức bằng một phương trình, nhưng cũng cần chú ý rằng :

- Người ta ký hiệu hàm xác suất biểu thị xác suất mà biến ngẫu nhiên rời rạc X nhận giá trị x_i là $P_X(x_i) = P(X = x_i)$, tổng quát ta viết $P_X(x)$ là hàm xác suất của X và xác định với mọi giá trị có thể có của X. Ví dụ nếu bạn thấy con súc sắc và đặt biến ngẫu nhiên X là số chấm xuất hiện thì bạn viết được $P_X(5) = 1/6$
- Phân phối xác suất của biến rời rạc phải thỏa 2 điều kiện :
 - i) $0 \leq P_X(x) \leq 1$
 - ii) $\sum P_X(x) = 1$

5.2.2.2 Phân phối xác suất của biến ngẫu nhiên liên tục

Với biến ngẫu nhiên liên tục, vì X nhận các giá trị liên tục trong một khoảng nào đó nên việc lập một bảng số mô tả quy luật phân phối của nó với một bên là các giá trị cụ thể X sẽ nhận và một bên là các xác suất X nhận giá trị x , tương ứng là điều không thể, trong tình huống này, người ta thường mô tả quy luật phân phối xác suất của biến ngẫu nhiên liên tục bằng một công thức toán được gọi là hàm mật độ xác suất kí hiệu $f_X(x)$, về mặt đồ thị thì hàm này định nên một đường cong gọi là đường cong mật độ xác suất, lần lượt các nội dung sau chúng ta sẽ khảo sát qua nhiều dạng đường cong mật độ xác suất, với mỗi dạng đường cong hàm mật độ xác suất lại có một công thức cụ thể, còn đường cong minh họa ở Hình 5.4 dưới đây chỉ là một ví dụ tiêu biểu.

Nhớ lại những hiểu biết của chúng ta ở Chương 3 về Đồ thị phân phối tần số (Histogram), chúng ta đã thử vẽ lại đồ thị này nhưng trực đứng không lấy thông tin từ cột tần số mà từ cột tần suất trong bảng tần số, và đã thấy hình dáng, cấu trúc và tỷ lệ so sánh tương đối giữa các cột của đồ thị mới vẽ được không khác so với đồ thị phân phối tần số. Bạn cũng dễ nhận

thấy tổng giá trị của tất cả các cột trên đồ thị mới chính bằng tổng tần suất và bằng 100% hay bằng 1. Trong tình huống dữ liệu rời rạc, ta nhận ra các cột rõ ràng, nhưng tương ứng với dữ liệu liên tục thì số cột là rất nhiều và bề ngang các cột rất hẹp đến nỗi nếu nối các đỉnh cột lại (như cách vẽ đa giác tần số) thì khi đó ta có một đường cong nhẵn liên tục biểu diễn các tần suất.

Hình 5.4 a



Tần suất kết hợp với một cột biểu thị khả năng các giá trị trong mẫu lọt vào trong phạm vi giá trị tạo nên 2 thành cột. Khả năng chính là xác suất để ta có một giá trị trong khoảng đó. Vậy nếu diện tích toàn phần dưới đường cong tần suất đã được định chuẩn là 1 (tức 100%) thì một phần diện tích nào đó dưới đường cong sẽ cho ta xác suất tương ứng với khả năng các giá trị trong mẫu lọt vào trong phạm vi giá trị đã xác định nên đoạn đáy của phần diện tích đang xét.

Trên Hình 5.4 b nếu ta muốn tính xác suất để biến số ngẫu nhiên liên tục X nhận giá trị trong khoảng $[a,b]$ là vùng tô sậm màu ta làm như thế nào?

Nói một cách ngắn gọn, xác suất để biến ngẫu nhiên X nhận một giá trị trong khoảng $[a;b]$ nào đó là diện tích bên dưới đường cong mật độ xác suất giới hạn bởi hai đường thẳng vuông góc với trục hoành đi qua hai điểm a và b , trục hoành và phần đường cong mật độ xác suất (xem miền sậm màu trên Hình 5.4b). Nói theo chiều ngược lại, diện tích phần hình được giới hạn bởi trục hoành, đường cong mật độ xác suất và hai đường thẳng song song trục tung đi qua hai giá trị a và b chính bằng xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng $[a;b]$, ký hiệu $P(a \leq X \leq b)$. Về mặt toán học để tính xác suất này tức cũng chính là để tính diện tích phần này ta làm như sau (trong đó $f_X(x)$ là hàm mật độ xác suất)

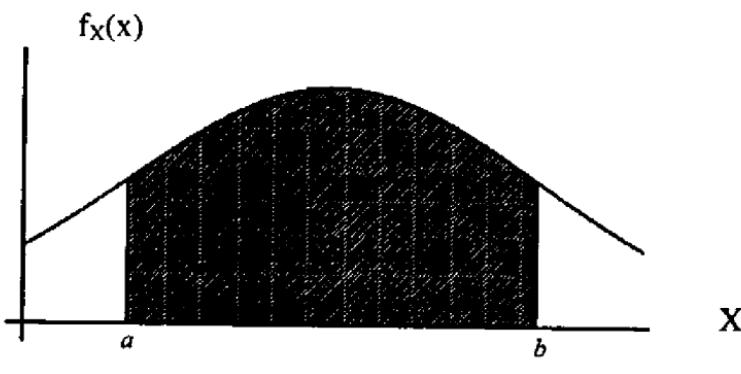
$$\text{Với } a < b, \quad P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

Chú ý Nếu $f_X(x)$ là Hàm mật độ xác suất thì nó phải thỏa mãn hai điều kiện cơ bản

i) $f_X(x) \geq 0$, với mọi x

ii) $\int_{-\infty}^{+\infty} f_X(x) dx = 1$ nghĩa là toàn bộ diện tích dưới đường cong mật độ xác suất bằng 1.

Hình 5.4 b



$$P(a \leq X \leq b) = S$$

Chú ý rằng với biến ngẫu nhiên liên tục, cho bất cứ giá trị x_i xác định nào thì $P(X=x_i) = 0$ có nghĩa là xác suất để biến liên tục nhận một giá trị cụ thể là zero, do vậy khi làm việc với biến liên tục chúng ta phải xác định xác suất cho một khoảng giá trị chứ không phải một giá trị cụ thể. Ví dụ nếu chúng ta hỏi “Xác suất để một nam sinh viên trong lớp nặng đúng 60 kg là bao nhiêu?”, câu trả lời là 0, nhưng nếu chúng ta muốn biết khả năng để có một sinh viên có cân nặng trong khoảng 59,5 kg đến 60,5 kg thì chúng ta có thể tính phần diện tích dưới đường cong mật độ xác suất (đường này biểu diễn mối quan hệ giữa các giá trị có thể có của biến ngẫu nhiên X đại diện chiều cao của sinh viên trong lớp và xác suất tương ứng để X nhận các giá trị chiều cao đó) giữa hai giá trị này. Chúng ta sẽ gặp lại và tìm hiểu chi tiết hơn về điều này ở các nội dung sau, còn ở đây ta ghi nhớ rằng với biến ngẫu nhiên liên tục thì các biểu thức $P(a \leq X \leq b)$ và $P(a < X < b)$ tương đương bởi lẽ $P(X=a)$ hay $P(X=b)$ bao giờ cũng bằng 0.

Ví dụ : Ta có biến ngẫu nhiên liên tục X tuân theo quy luật phân phối xác suất với hàm mật độ xác suất có dạng như sau:

$$f_X(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ 2x & \text{nếu } 0 \leq x \leq 1 \\ 0 & \text{nếu } x > 1 \end{cases}$$

Yêu cầu : Tính $P(0,5 \leq X \leq 0,75)$ = ?

Không cần vẽ đường cong mật độ xác suất mà để tính được xác suất này trước hết ta phải kiểm tra xem $f_X(x)$ có đạt điều kiện là hàm mật độ xác suất không.

Điều kiện 1: rõ ràng ta thấy $f_X(x) \geq 0, \forall x$

Điều kiện 2: $\int_{-\infty}^{+\infty} f_X(x).dx = 1$ có đạt không ?

$$\begin{aligned} \text{Ta thấy } \int_{-\infty}^{+\infty} f_X(x).dx &= \int_{-\infty}^0 f_X(x).dx + \int_0^1 f_X(x).dx + \int_1^{+\infty} f_X(x).dx \\ &= \int_0^1 2x.dx \\ &= x^2 \Big|_0^1 = 1^2 - 0^2 = 1 \end{aligned}$$

Vậy điều kiện 2 cũng đạt, nên $f_X(x)$ là hàm mật độ xác suất

Để tính xác suất cần thiết ta vận dụng công thức

$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b f(x).dx \\ \Rightarrow P(0,5 \leq X \leq 0,75) &= \int_{0,5}^{0,75} 2x.dx = x^2 \Big|_{0,5}^{0,75} = (0,75)^2 - (0,5)^2 = 0,3125 \end{aligned}$$

Như vậy xác suất để giá trị của X rơi vào khoảng $(0,5; 0,75)$ bằng 0,3125

5.2.3 Các đặc trưng cơ bản của biến ngẫu nhiên

Sự phân phối xác suất mô tả ở trên cho ta một mô thức về phân phối thực tế của một biến ngẫu nhiên, như vậy khi ta xác định được quy luật phân phối xác suất của 1 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 qua tâm đến những thông tin cộ đồng phản ánh những đặc trưng quan trọng nhất của biến ngẫu nhiên đang nghiên cứu, những thông tin cộ đồng phản ánh từng phần về biến ngẫu

nhiên được gọi là các tham số đặc trưng hay các đặc trưng cơ bản của biến ngẫu nhiên. Các đặc trưng cơ bản này gồm 2 nhóm chính là đặc trưng cho xu hướng trung tâm và đặc trưng cho độ phân tán. Ý tưởng này bạn đọc đã làm quen qua Chương 4. Với đặc trưng cho xu hướng trung tâm ta có trị số kỳ vọng (chính là trung bình) và đặc trưng cho độ phân tán ta có phương sai (hoặc độ lệch chuẩn) kết hợp với tổng thể mà phân phối này biểu diễn. Chúng ta sẽ tìm hiểu công thức tính các đại lượng này cho cả biến ngẫu nhiên rời rạc và liên tục.

5.2.3.1 Kỳ vọng

Kỳ vọng kí hiệu là $E(X)$.

- Kỳ vọng của một biến ngẫu nhiên rời rạc được xác định bằng cách nhân mỗi giá trị có thể có x_i của X với xác suất tương ứng $P(x_i)$ sau đó cộng tất cả các kết quả đó lại, với biến ngẫu nhiên rời rạc công thức tính kỳ vọng như sau:

$$E(X) = \sum_x x_i P_X(x_i)$$

Trong đó:

$E(X)$ là kỳ vọng của biến ngẫu nhiên

x_i : là giá trị thứ i của biến X

$P_X(x_i)$: là xác suất để X nhận giá trị x_i

Chúng ta nhớ lại rằng trị trung bình chẳng qua là trung bình của một số giá trị, với đại lượng ngẫu nhiên X ta cũng tính được trung bình của nó bởi nó cũng bao gồm các giá trị và khả năng xuất hiện các giá trị này, theo công thức tính thì ta thấy kỳ vọng của biến ngẫu nhiên chính là trị trung bình của biến ngẫu nhiên và được kí hiệu là μ_X , vậy $E(X) = \mu_X$. Quy ước kí hiệu này dùng cho cả tình huống kỳ vọng của biến ngẫu nhiên liên tục

Ta có thể hiểu rõ phương pháp tính kỳ vọng của một biến số ngẫu nhiên qua việc phát triển tiếp ví dụ về phân phối xác suất của biến ngẫu nhiên X là số mặt ngửa xuất hiện khi tung một lượt hai đồng tiền, giả sử lúc này ta tung tất cả 4.000.000 lượt trong thí nghiệm này, theo trực giác ta chờ đợi quan sát được vào khoảng 1.000.000 lượt không có mặt nào ngửa, 2.000.000 lượt có 1 mặt ngửa và 1.000.000 lượt được 2 mặt ngửa. Trị số kỳ vọng của biến số ngẫu nhiên X lúc này được tính

$$= \frac{1000.000 \times (0) + 2000.000 \times (1) + 1000.000 \times (2)}{4000.000} = \frac{1}{4} \times (0) + \frac{1}{2} \times (1) + \frac{1}{4} \times (2) = 1$$

Hãy để ý các cặp số hạng trong biểu thức ở trước dấu $=$ cuối cùng, chúng lần lượt là $P(0)*0$, $P(1)*1$ và $P(2)*2$ rồi xem Bảng 5.3. Trị số trung bình = 1

là trị số hay xảy ra nhất và trong thí nghiệm đây là trị số ta chờ đợi xảy ra nhiều lần nhất.

Kỳ vọng của hàm số của biến ngẫu nhiên rời rạc

Biến ngẫu nhiên rời rạc X có hàm xác suất $P_X(x)$, $g(X)$ là một hàm số của biến ngẫu nhiên X, kỳ vọng của hàm số $g(X)$ được định nghĩa như sau

$$E[g(x)] = \sum_x g(x)P_X(x_i)$$

Kỳ vọng $E(X)$ của một biến ngẫu nhiên liên tục được định nghĩa theo công thức sau

$$E(X) = \int_{-\infty}^{+\infty} xf_X(x)dx$$

Ví dụ : tiếp tục với ví dụ có biến ngẫu nhiên liên tục X tuân theo quy luật phân phối xác suất với hàm mật độ xác suất có dạng như sau:

$$f_X(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ 2x & \text{nếu } 0 \leq x \leq 1 \\ 0 & \text{nếu } x > 1 \end{cases}$$

Yêu cầu : Tính $E(X)$

Từ công thức $E(X) = \int_{-\infty}^{+\infty} xf_X(x)dx$

Theo điều kiện của bài toán ta tách ra như sau:

$$E(X) = \int_{-\infty}^0 x.f_X(x).dx + \int_0^1 x.f_X(x).dx + \int_1^{+\infty} x.f_X(x).dx$$

Vậy $E(X) = \int_0^1 xf_X(x)dx$

$$= \int_0^1 x \cdot 2x dx = 2 \int_0^1 x^2 dx = 2 \left(\frac{x^3}{3} \right) \Big|_0^1 = \frac{2}{3}$$

Như vậy giá trị kỳ vọng của biến ngẫu nhiên X trong ví dụ này bằng $2/3$.

Kỳ vọng của hàm số của biến ngẫu nhiên liên tục

Biến ngẫu nhiên liên tục X có hàm mật độ xác suất $f_X(x)$, $g(X)$ là một hàm số của biến ngẫu nhiên X kỳ vọng của hàm số $g(X)$ được định nghĩa như sau:

$$E[g(x)] = \int_{-\infty}^{+\infty} g(x)f_X(x)dx$$

5.2.3.2 Phương sai

Với biến ngẫu nhiên X có μ_x là kỳ vọng của biến ngẫu nhiên này thì phương sai của biến ngẫu nhiên rời rạc chính là kỳ vọng của $(X - \mu_x)^2$ và được kí hiệu là $V(X)$ với công thức tính.

- Công thức tính phương sai của biến ngẫu nhiên rời rạc

$$V(X) = E[(X - \mu_x)^2] = \sum_x (x_i - \mu_x)^2 P_x(x_i)$$

Trong đó:

- $V(X)$ là phương sai của biến ngẫu nhiên, ta có cách kí hiệu thứ 2 tương tự của phương sai là σ_x^2
- μ_x là kỳ vọng của biến ngẫu nhiên
- x_i : là giá trị thứ i của biến X
- $P_x(x_i)$: là xác suất để X nhận giá trị x_i

Phương sai còn có thể tính theo công thức thứ hai thuận tiện hơn cho tính toán: $V(X) = \sigma_x^2 = [E(X^2) - (\mu_x)^2] = \sum_x x^2 P_x(x_i) - (\mu_x)^2$

- Công thức tính phương sai của biến ngẫu nhiên liên tục

$$V(X) = \sigma_x^2 = E[(X - \mu_x)^2] = \int_{-\infty}^{+\infty} [(x - \mu_x)^2] f_x(x) dx$$

Hay $V(X) = \sigma_x^2 = [E(X^2) - (\mu_x)^2] = \int_{-\infty}^{+\infty} x^2 f_x(x) dx - (\mu_x)^2$

Ví dụ : tiếp tục với ví dụ có biến ngẫu nhiên liên tục X tuân theo quy luật phân phối xác suất với hàm mật độ xác suất có dạng như sau:

$$f_x(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ 2x & \text{nếu } 0 \leq x \leq 1 \\ 0 & \text{nếu } x > 1 \end{cases}$$

Yêu cầu : Tính phương sai σ_x^2

Vận dụng công thức thứ hai

$$\sigma_x^2 = E(X^2) - (\mu_x)^2 = \int_{-\infty}^{+\infty} x^2 f_x(x) dx - (\mu_x)^2$$

Từ phần trên ta đã tính được $E(X) = \mu_x = 2/3 \rightarrow (\mu_x)^2 = (2/3)^2$

Như vậy ta còn phải tính phần $E(X^2) = \int_{-\infty}^{+\infty} x^2 f_x(x) dx$

Ta tách ra như sau :

$$E(X^2) = \int_{-\infty}^0 x^2 f_x(x) dx + \int_0^1 x^2 f_x(x) dx + \int_1^{+\infty} x^2 f_x(x) dx$$

Vậy $E(X^2) = \int_0^1 x^2 f_x(x) dx = \int_0^1 x^2 2x dx = 2 \int_0^1 x^3 dx = 2 \left(\frac{x^4}{4} \right) \Big|_0^1 = \frac{1}{2}$

Như vậy $\sigma_x^2 = [E(X^2) - (\mu_x)^2] = (1/2) - (2/3)^2 = 1/18$

Như vậy giá trị phương sai của biến ngẫu nhiên X trong ví dụ này bằng 1/18.

5.2.3.3 Độ lệch chuẩn

Độ lệch chuẩn được ký hiệu là σ_x với công thức tính

$$\sigma_x = \sqrt{V(X)} = \sqrt{\sigma_x^2}$$

Ví dụ: Tiếp tục với ví dụ trên, nếu phương sai ta đã tính được bằng 1/18 thì độ lệch chuẩn $\sigma_x = \sqrt{1/18}$

Để hiểu rõ hơn cách tính và ý nghĩa của các đặc trưng của biến ngẫu nhiên ta xem xét ví dụ thực tế sau đây.

Giả sử, một cặp vợ chồng dự định sẽ sinh 3 đứa con, và họ quan tâm đến số con gái mà có khả năng họ sẽ có trong ba đứa con (đặt là biến ngẫu nhiên X). Hãy thành lập bảng phân phối xác suất của X, tính các đặc trưng số của X (gồm kỳ vọng và phương sai) ?

Rõ ràng số con gái mà cặp vợ chồng này sẽ có là giá trị của một biến ngẫu nhiên rời rạc với các giá trị có thể nhận là 0, 1, 2, hoặc 3. Tuy nhiên khả năng xảy ra các kết cục này không giống nhau. Để tính được xác suất mà X nhận các giá trị x_i bạn hãy hình dung đến không gian mẫu và các giá trị mà biến X nhận tương ứng với các biến cố sơ cấp trong không gian mẫu như sau

Bảng 5.4 a

Các biến cố sơ cấp			X = số con gái
T	T	T	0
G	T	T	1
T	G	T	1
T	T	G	1
G	G	T	2
G	T	G	2
T	G	G	2
G	G	G	3

Bảng 5.4 b

→

X	0	1	2	3
$P_X(x)$	0,14	0,39	0,36	0,11

Diễn giải:

Nhớ rằng xác suất sinh con trai trong mỗi ca sinh nói chung đã được xác định là 0,52 vì thế xác suất sinh gái phải bằng $(1-0,52) = 0,48$ bởi vì sinh gái hay trai là hai biến cố xung khắc trong một phép thử (mỗi phép thử chính là một ca sinh).

Khi muốn tính xác suất sẽ có 1 gái trong 3 đứa con (hay chính là tính $P(X=1)$) ta phải tính xác suất của một biến cố bao gồm ba biến cố xung khắc GTT, TGT và TTG, tức là $(X=1) = [(GTT) + (TGT) + (TTG)]$, do đó

$$P(X=1) = P(GTT) + P(TGT) + P(TTG) =$$

$$= [(1-0,52)*(0,52)*(0,52)] + [(0,52)*(1-0,52)*(0,52)] + [(0,52)*(0,52)*(1-0,52)] = 0,13 + 0,13 + 0,13 = 0,39$$

Cũng với quy tắc đó ta tính được các xác suất $P(X=x_i)$ còn lại như đã liệt kê ở Bảng 5.4b

Để tính các đặc trưng của biến ngẫu nhiên ta áp dụng công thức tính kỳ vọng và phương sai đã biết, và để thuận tiện và tránh nhầm lẫn thì hệ thống việc tính toán thành 1 bảng số liệu như sau đây là 1 giải pháp hay.

Bảng 5.5

Bảng phân phối xác suất		Tính toán μ_x	Tính toán σ_x^2		
x	$P_x(x)$	$x P_x(x)$	$(x-\mu_x)$	$(x-\mu_x)^2$	$(x-\mu_x)^2 P_x(x)$
0	0,14	0	- 1,44	2,07	0,29
1	0,39	0,39	- 0,44	0,19	0,08
2	0,36	0,72	0,56	0,31	0,11
3	0,11	0,33	1,56	2,43	0,27
		$\mu_x = 1,44$			$\sigma_x^2 = 0,75$

Giá trị kỳ vọng bằng 1,44 nghĩa là số con gái trung bình mà cặp vợ chồng này sẽ có trong 3 đứa con là 1,44 trẻ gái; dĩ nhiên lời giải này không thể hiểu sát nghĩa theo từng chữ mà nên hiểu về mặt trung bình thì là như vậy.

Giá trị phương sai bằng 0,75 thì giá trị độ lệch chuẩn $= \sqrt{0,75} = 0,87$ tức là biến thiên của số con gái họ có là 0,87 trẻ gái so với giá trị trung bình 1,44.

5.2.4 Ứng dụng kỳ vọng vào việc ra quyết định trong kinh doanh

5.2.4.1 Khái niệm ra quyết định

Ra quyết định là một quá trình lựa chọn có ý thức giữa hai hoặc nhiều phương án để chọn ra một phương án và phương án này sẽ tạo được một kết quả mong muốn trong các điều kiện ràng buộc đã biết.

Các nhà quản trị thường phải ra quyết định trong các tình huống không chắc chắn vì nhiều trường hợp không có đủ thông tin, ví dụ :

- Lựa chọn công suất xây dựng nhà máy sẽ là công suất lớn hay nhỏ hoặc trung bình thì sẽ phù hợp với tình hình nền kinh tế trong tương lai.

- Quyết định nhập kho bao nhiêu sản phẩm là vừa với nhu cầu thị trường..

Phương pháp ra quyết định dựa trên hiểu biết về xác suất và kỳ vọng giúp các nhà quản trị có cơ sở lựa chọn trong các tình huống như vậy.

5.2.4.2 Lập bảng kết toán và ra quyết định bằng phương pháp EMV

Bảng kết toán

Bảng kết toán là bảng 2 chiều liệt kê các biến cố có thể xảy ra cho từng Phương án hành động. Trong sự kết hợp một biến cố với một phương án ta cần xác định lợi nhuận của tình huống kết hợp đó.

Ví dụ một doanh nghiệp sản xuất sản phẩm X đang cân nhắc 3 phương án xây dựng nhà máy là phương án xây dựng theo quy mô nhỏ, quy mô lớn hay quy mô vừa. Doanh nghiệp có thể gặp rủi ro khi quyết định sai lầm vì nền kinh tế có thể đổi mới với 1 trong 3 tình huống là kinh tế tăng trưởng mạnh, tăng trưởng ổn định hay bị suy yếu.

Doanh nghiệp đánh giá mức lợi nhuận (nghìn \$) xảy ra cho từng phương án đã chọn khi một biến cố nào đó xảy ra và lập thành bảng kết toán sau:

Bảng 5.6

Các tình huống của nền kinh tế	Lợi nhuận từ các tình huống (nghìn \$)		
	Quy mô lớn	Quy mô vừa	Quy mô nhỏ
Kinh tế mạnh	200	90	40
Kinh tế ổn định	50	110	30
Kinh tế suy yếu	-110	-30	20

Lúc đó các biến cố chính là Các tình huống của nền kinh tế; các phương án là 3 sự lựa chọn của doanh nghiệp về quy mô của nhà máy; các số liệu trong các ô là lợi nhuận của các tình huống kết hợp.

Nếu ký hiệu các phương án hành động là j và các biến cố là i thì x_{ij} là lợi nhuận của việc chọn phương án j khi biến cố i xảy ra

Sử dụng tiêu chuẩn cực đại lợi nhuận kỳ vọng (Expected Monetary Value -EMV) để ra quyết định

Sau khi đã xác định được lợi nhuận của từng phương án sản xuất tương ứng với các biến cố, để đánh giá xem phương án nào là tối ưu ta tính lợi nhuận kỳ vọng của từng phương án theo công thức

$$EMV_j = \sum x_{ij} \cdot P_i$$

Trong đó

EMV_j là lợi nhuận kỳ vọng của phương án j

x_{ij} là lợi nhuận của việc chọn phương án j khi biến cố i xảy ra

P_i : xác suất của biến cố i (có thể có được qua kinh nghiệm các chuyên gia hoặc qua nghiên cứu, khảo sát thị trường...)

Nguyên tắc lựa chọn → phương án tối ưu là phương án có EMV max

Với ví dụ trên, giả sử các chuyên gia dự đoán có 50% khả năng nền kinh tế phát triển ổn định, 30% khả năng nền kinh tế phát triển mạnh, còn lại là suy yếu. Ta dùng bảng kết toán trên để lập bảng tính EMV như sau

Bảng 5.7

Xác suất xảy ra các tình huống của nền kinh tế P_i	Lợi nhuận từ các tình huống (nghìn \$)					
	Quy mô lớn		Quy mô vừa		Quy mô nhỏ	
	x_{i1}	$x_{i1}P_i$	x_{i2}	$x_{i2}P_i$	x_{i3}	$x_{i3}P_i$
Kinh tế mạnh	0,3	200	60	90	27	40
Kinh tế ổn định	0,5	50	25	110	55	30
Kinh tế suy yếu	0,2	-110	-22	-30	-6	20
EMV _j			63		76	31

Theo nguyên tắc lựa chọn ta sẽ chọn phương án tối ưu là phương án nhà máy có Quy mô vừa vì EMV của phương án này cực đại (76 nghìn \$)

5.2.4.3 Lập bảng tổn thất cơ hội và ra quyết định bằng phương pháp EOL

Bảng tổn thất cơ hội

Tổn thất cơ hội là chênh lệch giữa mức lợi nhuận lớn nhất có thể có của một biến cố với mức lợi nhuận đạt được từ một phương án hành động cụ thể nào đó. Tổn thất cơ hội cho thấy mức lợi nhuận bị mất đi khi một phương án hành động tốt nhất đã không được chọn.

Kí hiệu L_{ij} là tổn thất cơ hội của phương án j khi biến cố i xảy ra. Từ bảng kết toán ta xây dựng được bảng tổn thất cơ hội như sau.

Bảng 5.8

Các tình huống của nền kinh tế	Tổn thất từ các tình huống (nghìn \$)		
	Quy mô lớn	Quy mô vừa	Quy mô nhỏ
Kinh tế mạnh	0	110	160
Kinh tế ổn định	60	0	80
Kinh tế suy yếu	130	50	0

Giải thích cách tính các L_{ij} : Biến cố Nền kinh tế mạnh đã xảy ra và nếu doanh nghiệp chọn quy mô nhà máy lớn thì lợi nhuận là đúng 200 nghìn \$ tức là không bị tổn thất cơ hội gì cả, nhưng nếu họ đã chọn quy mô vừa thì khi nền kinh tế mạnh họ chỉ đạt được lợi nhuận 90 nghìn \$ tức là họ tổn thất một khoản cơ hội là $(90 - 200) = -110$ nghìn \$. Còn nếu trước đó đã chọn quy mô nhỏ thì tổn thất của họ lên tới 160 nghìn \$.

Sử dụng tiêu chuẩn cực tiểu tổn thất cơ hội kỳ vọng (*Expected Opportunity Loss - EOL*) để ra quyết định

Sau khi tính toán các mức tổn thất cơ hội ta có thể đánh giá lựa chọn các phương án bằng cách tính giá trị tổn thất cơ hội kỳ vọng theo công thức

$$EOL_j = \sum L_{ij} P_i$$

Trong đó:

EOL_j là tổn thất cơ hội kỳ vọng của phương án j

L_{ij} là tổn thất cơ hội của phương án j khi biến cố i xảy ra

P_i : xác suất của biến cố i

Nguyên tắc lựa chọn \rightarrow phương án tối ưu là phương án có EOL_{min} vì tổn thất ít nhất

Lập bảng tính các EOL_j

Bảng 5.9

Xác suất xảy ra các tình huống của nền kinh tế P_i	Tổn thất từ các tình huống (nghìn \$)					
	Quy mô lớn		Quy mô vừa		Quy mô nhỏ	
	L_{i1}	$L_{i1}P_i$	L_{i2}	$L_{i2}P_i$	L_{i3}	$L_{i3}P_i$
Kinh tế mạnh 0,3	0	0	110	33	160	48
Kinh tế ổn định 0,5	60	30	0	0	80	40
Kinh tế suy yếu 0,2	130	26	50	10	0	0
EOL_j		56		43		88

Theo nguyên tắc lựa chọn ta sẽ chọn phương án tối ưu là phương án nhà máy có Quy mô vừa vì EOL của phương án này cực tiểu (43 nghìn \$)

5.3 CÁC PHÂN PHỐI LÝ THUYẾT QUAN TRỌNG

5.3.1 Phân phối lý thuyết cho biến rời rạc

5.3.1.1 Phân phối Nhị thức (Binomial Distribution)

Qua nội dung quy luật phân phối xác suất nghiên cứu ở phần trước có thể thấy rằng khi chúng ta có sẵn một phương trình toán học, chúng ta có thể tính chính xác xác suất xảy ra bất kỳ giá trị cụ thể nào của biến ngẫu nhiên rời rạc. Như vậy là toàn bộ phân phối xác suất của biến ngẫu nhiên có thể được xác định. Nhiều mô hình toán học đã được hình thành để thể hiện quy luật phân phối của những biến ngẫu nhiên rời rạc mà chúng ta hay gặp trong các tình huống nghiên cứu kinh tế, nghiên cứu xã hội và thậm chí cả khoa học tự nhiên, y học... trong đó có một quy luật phân phối xác suất khá hữu ích là phân phối Nhị thức.

Phân phối Nhị thức là một hàm phân phối xác suất rời rạc có nhiều ứng dụng trong thực tế, nó được sử dụng khi biến ngẫu nhiên rời rạc ta quan tâm là số lần thành công trong n lần thực hiện lặp lại một phép thử giống hệt nhau (ta quy ước gọi theo cách của thống kê là một mẫu n phép thử). Các phép thử này có kết cục là hai biến cố xung khắc Thành công và Thất bại, tổng quát ta gọi tất là thành và bại. Ví dụ ta xem mỗi lần sinh là một lần làm phép thử trong đó biến cố sinh được bé gái là thành, hoặc mỗi lệnh đặt hàng là một phép thử trong đó biến cố lệnh đặt hàng được thực hiện đúng giờ là thành... Như vậy biến cố sinh bé trai là bại hoặc lệnh đặt hàng bị thực hiện trễ giờ là bại.

Chúng ta quy ước gọi p là tỷ lệ thành và $q = (1-p)$ là tỷ lệ bại nói chung trong tổng thể, ví dụ tỷ lệ sinh bé gái là $p=0,48$ hay tỷ lệ mặt ngửa là $p=0,5$. Nếu chúng ta lấy một mẫu cỡ n (ví dụ chọn ngẫu nhiên n lệnh đặt hàng đã được thực hiện trước đây để xem tiến trình thực hiện lệnh có đúng giờ hay không, hay quan sát n ca sinh) từ một phân phối Nhị thức thì số lần thành trong mẫu tạo nên các giá trị rời rạc có thể có của một biến số ngẫu nhiên, và xác suất kết hợp với các giá trị có thể có được của biến ngẫu nhiên sẽ có một sự phân phối gọi là phân phối Nhị thức.

Phân phối Nhị thức có các tính chất cơ bản sau:

- i) Thí nghiệm gồm có n lượt thử y hệt nhau
- ii) Kết quả của mỗi lượt thử rơi vào một trong hai trường hợp mà chúng ta có thể gọi một cách tổng quát là thành và bại

iii) Xác suất thành trong một lượt thử đơn độc bằng p và như nhau trong mọi lượt thử. Tương tự xác suất bại là $q = (1-p)$

iv) Các lần thử đều độc lập, để bảo đảm tính độc lập, mỗi quan sát được chọn ngẫu nhiên không hoàn lại từ một tổng thể vô hạn hoặc có hoàn lại từ một tổng thể hữu hạn (nhằm làm cho xác suất thành hay bại trong mỗi lượt thử đều bằng nhau y hệt).

v) Chúng ta quan tâm đến số lượt thành (thể hiện qua các giá trị x_i mà biến X nhận) quan sát được trong n lượt thử

Trước khi trình bày công thức của phân phối Nhị thức chúng ta sẽ thực hiện một tiến trình từ đơn giản đến phức tạp để tính $P(X)$ với số lần thử lần lượt là $n=1, 2, \dots, 3$. Từ đó mà dẫn suất đến công thức tổng quát. Trong tiến trình này ta quy ước:

- Mỗi phép thử của chúng ta có xác suất thành cố định là p và xác suất bại là $q = (1-p)$
- Ký hiệu gọn T là kết cục thành và B là kết cục bại trong mỗi phép thử.
- Biến số ngẫu nhiên X nhận kết quả = số lượt thành trong n phép thử.

Tiến trình i) Khi $n=1$ (tức chỉ thực hiện phép thử một lần)

Không gian mẫu của thí nghiệm của chúng ta với ($n=1$) có hai biến cố sơ cấp ký hiệu là E_1 và E_2 . Lúc này X nhận hai trị số 0 (là không có lần nào thành tương đương với biến cố E_2) và 1 (là có một lần thành tương đương với biến cố E_1). Quy luật phân phối xác suất được mô tả bằng bảng số như sau

Bảng 5.10

Biến cố sơ cấp	Kết cục	$P(E_i)$	X	X	$P(X)$
E_1	T	p	1	0	q
E_2	B	q	0	1	p

$\Sigma P(X) = (p + q) = 1$

Tiến trình ii) Khi $n = 2$ (tức là thực hiện phép thử hai lần)

Với các ý nghĩa tương tự của E_i , $P(E_i)$, X và $P(X)$, các kết quả của trường hợp thực hiện phép thử hai lần được trình bày trong Bảng 5.11

Xác suất tương ứng với các E_i có thể tính được dễ dàng vì mỗi biến cố là sự giao của hai biến cố độc lập nên ta áp dụng Quy tắc nhân xác suất đã nghiên cứu ở trước để tính các $P(E_i)$

Ví dụ $P(E_1) = P(T) * P(T) = p*p = p^2$

$$P(E_2) = P(T) * P(B) = p*q = pq$$

Khi tính xác suất $P(X)$ cho biến số ngẫu nhiên X , chúng ta để ý rằng $X=1$ tương ứng với hai biến cố sơ cấp E_2 và E_3 , do đó

$$P(X=1) = P(E_2) + P(E_3) = (pq+pq) = 2pq$$

Bảng 5.11

Biến cố sơ cấp	Kết cục	$P(E_i)$	X
E_1	TT	p^2	2
E_2	TB	pq	1
E_3	BT	qp	1
E_4	BB	q^2	0

X	$P(X)$
0	q^2
1	$2pq$
2	p^2

$$\Sigma P(X) = p^2 + 2pq + p^2 = (p+q)^2 = 1$$

Ngoài ra cũng nên lưu ý rằng tổng các xác suất $P(X)$ là các số hạng trong sự khai triển biểu thức $(p+q)^2$.

Tiến trình iii) Khi $n=3$ (tức là thực hiện phép thử ba lần)

Bảng 5.12

Biến cố sơ cấp	Kết cục	$P(E_i)$	X
E_1	TTT	p^3	3
E_2	TTB	p^2q	2
E_3	TBT	p^2q	2
E_4	BTT	p^2q	2
E_5	TBB	pq^2	1
E_6	BTB	pq^2	1
E_7	BBT	pq^2	1
E_8	BBB	q^3	0

X	$P(X)$
0	q^3
1	$3pq^2$
2	$3p^2q$
3	p^3

$$\Sigma P(X) = q^3 + 3pq^2 + 3p^2q + p^3 = (p+q)^3 = 1$$

Trong trường hợp $n=3$, tổng các xác suất $P(X)$ cũng lại là những số hạng trong sự khai triển biểu thức $(p+q)^3$

Điều chúng ta chứng tỏ được qua tiến trình trên thật rõ ràng, phân phôi xác suất cho một thí nghiệm nhị thức gồm n lượt thử có thể tính được bằng cách khai triển nhị thức Newton $(p+q)^n$.

Hãy xem lại phương trình biểu diễn tổng số xác suất của các kết cục trong trường hợp n=3 :

$P(X) = (q^3 + 3pq^2 + 3p^2q + p^3) = 1$, dĩ nhiên phải bằng 1 vì đây là nhóm đầy đủ các biến cố.

Có phải mỗi số hạng trong phương trình trên chính là xác suất $P(X)$ xác định khả năng X nhận một trị số đặc biệt x , ta nhận thấy lần lượt những số hạng này liên quan đến lũy thừa bậc x ($x = 0, 1, 2, 3$) của p trong phương trình nhị thức $(p+q)^3$, tức là thành phần p^x , về mặt tổng quát có thể biểu diễn số hạng chứa thành phần p^x thành $C_x^n p^x q^{n-x}$.

Chúng ta thử lần lượt thế giá trị của x của X vào công thức trên để chứng minh điều này (với trường hợp $(p+q)^3$ tức là $n=3$)

Với $X = 0$ thì $P(X=0) = q^3$. Nếu ta khai triển theo công thức $C_0^3 p^0 q^{3-0}$ thì cũng bằng

$$\frac{3!}{0!(3-0)!} * 1 * q^3 = q^3$$

Với $X = 1$ thì $P(X=1) = 3pq^2$. Nếu ta khai triển theo công thức $C_1^3 p^1 q^{3-1}$ thì cũng bằng

$$\frac{3!}{1!(3-1)!} * p * q^2 = 3pq^2$$

Với $X = 2$ thì $P(X=2) = 3p^2q$. Nếu ta khai triển theo công thức $C_2^3 p^2 q^{3-2}$ thì cũng bằng

$$\frac{3!}{2!(3-2)!} * p^2 * q = 3p^2q$$

Với $X = 3$ thì $P(X=3) = p^3$. Nếu ta khai triển theo công thức $C_3^3 p^3 q^{3-3}$ thì cũng bằng

$$\frac{3!}{3!(3-3)!} * p^3 * 1 = p^3$$

Tóm lại, người ta dùng phương trình toán học sau đây để biểu diễn phân phối xác suất Nhị thức nhằm xác định khả năng thành công x lần trong n lần thực hiện phép thử khi biết xác suất thành trong mỗi phép thử là nhau và bằng p

$$P(X=x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Nhớ lại ví dụ về số con gái mà một cặp vợ chồng muốn xác định khả năng họ sẽ có trong 3 đứa con dự định sinh, lúc này mỗi lần sinh là một lần làm phép thử với xác suất thành công (sinh gái) không đổi là $p = 0,48$; họ dự định sinh 3 đứa con tức là $n=3$, số con gái mà họ sẽ có và xác suất xảy ra kết cục đó rõ ràng có phân phối Nhị thức. Ta vận dụng công thức tính của phân phối Nhị thức để xác định khả năng cặp vợ chồng này sẽ có đúng hai đứa con gái là

$$P(X=2) = \frac{3!}{2!(3-2)!} 0,48^2 (1-0,48)^{3-2} = 3 * 0,48^2 * 0,52 = 0,36$$

Kết quả không khác với cách ta tính toán bằng phương pháp suy luận logic khi chưa nghiên cứu về quy luật của phân phối Nhị thức.

Trong thực tế thì điều mà chúng ta cần biết thường không phải là xác suất của một kết cục đơn lẻ mà là xác suất của một nhóm xuất quả, giả dụ xác suất để X nhận giá trị trong một phạm vi trị số nào đó. Với ví dụ trên, nếu cặp vợ chồng này muốn biết xác suất để họ không có quá 2 đứa con gái là bao nhiêu, lúc này $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$

$$= \frac{3!}{0!(3-0)!} 0,48^0 (1-0,48)^{3-0} + \frac{3!}{1!(3-1)!} 0,48^1 (1-0,48)^{3-1} + \frac{3!}{2!(3-2)!} 0,48^2 (1-0,48)^{3-2} = 0,89$$

Bạn có thể kiểm tra độ chính xác của kết quả trên bằng cách xem lại ví dụ ở trước và cộng bằng máy tính bỏ túi.

Sau khi đã dẫn suất công thức của phân phối Nhị thức, chúng ta sẽ xác định công thức tính các đặc trưng của X , chúng ta cũng xét một số trường hợp đơn giản với $n=1, 2, 3$ để suy ra công thức tổng quát.

- Công thức tính các đặc trưng của biến ngẫu nhiên X theo quy luật phân phối Nhị thức

* Theo công thức tính kỳ vọng của biến số ngẫu nhiên rời rạc

$$\mu = \sum x_i \times P(x)$$

Dùng bảng phân phối $P(X)$ đã xác định ta tính toán

i) Với $n=1$

$$\mu = 0xq + 1xp = p$$

ii) Với $n=2$

$$\mu = 0xq^2 + 1x2pq + 2xp^2 = 2pq + 2p^2 = 2p(q+p) = 2p, \text{ vì } (p+q)=1$$

iii) Với $n=3$

$$\begin{aligned} \mu &= 0xq^3 + 1x3pq^2 + 2x3p^2q + 3xp^3 = 3p(q^2 + 2pq + p^2) = 3p(p+q)^2 \\ &= 3p \end{aligned}$$

Như vậy suy rộng cho trường hợp $n = n$ ta có $\mu = np$

* Theo công thức tính phương sai của biến số ngẫu nhiên rời rạc

$$\sigma^2 = \sum (x_i - \mu)^2 P(x_i)$$

i) Với $n=1$ thì $\mu = p$

$$\begin{aligned}\sigma^2 &= (0-p)^2 x q + (1-p)^2 x p = p^2 q + q^2 p, \text{ vì } (1-p) = q \\ &= pq(p+q) = pq\end{aligned}$$

ii) Với $n=2$ thì $\mu = 2p$

$$\begin{aligned}\sigma^2 &= (0-2p)^2 x q^2 + (1-2p)^2 x 2pq + (2-2p)^2 x p^2 = \\ &= 4p^2 q^2 + (1 - 4p + 4p^2) x 2pq + 2^2(1-p)^2 p^2 \\ &= 4p^2 q^2 + 2pq - 8p^2 q + 8p^3 q + 4q^2 p^2 \\ &= 8p^2 qx(q + p - 1) + 2pq \\ &= 2pq, \text{ vì } (q + p - 1) = 0 \text{ do } (q+p) = 1\end{aligned}$$

iii) Các bạn có thể tiếp tục kiểm chứng với $n=3$ và $\mu = 3p$ thì $\sigma^2 = 3pq$

Suy rộng cho $n = n$ ta được $\sigma^2 = npq$

Dù cho các công thức tính kỳ vọng và phương sai trên không được chứng minh một cách trực tiếp nhưng dẫn suất từ những trường hợp đơn giản trên cũng có thể bảo đảm độ tin cậy của chúng. Chúng ta sẽ vận dụng công thức tính các đặc trưng vừa xác định ở trên để tính lại số con gái trung bình và phương sai của số con gái mà cặp vợ chồng đó có khả năng có, rồi đem so sánh với các đáp số đã tính được bằng phương pháp thủ công xem chúng có bằng nhau không.

$$\mu = np = 3 \times 0,48 = 1,44$$

$$\sigma^2 = npq = 3 \times 0,48 \times 0,52 = 0,75$$

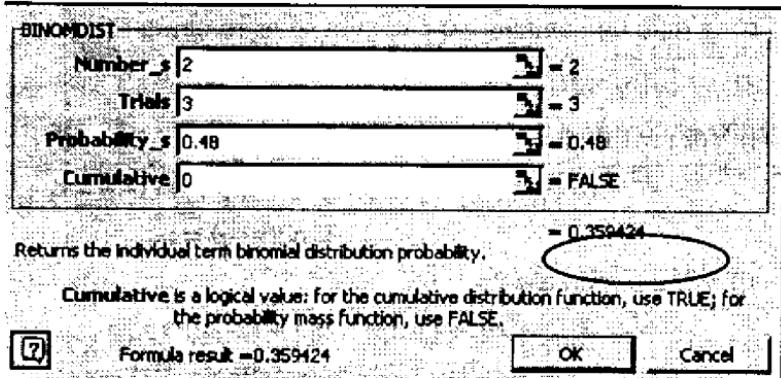
→ Kết quả không khác biệt.

- Đồng thời chúng ta cũng có thể sử dụng lệnh có sẵn trong Excel để tính toán các giá trị xác suất của phân phối Nhị thức mà ta quan tâm.

Các bạn vào lệnh Insert Function của Excel, chọn mục Statistical trong phần Function Category, sau đó chọn lệnh BINOMDIST trong danh sách các hàm thống kê tương ứng xuất hiện bên khung Function name.

Bạn sẽ thấy cửa sổ như sau đây mở ra.

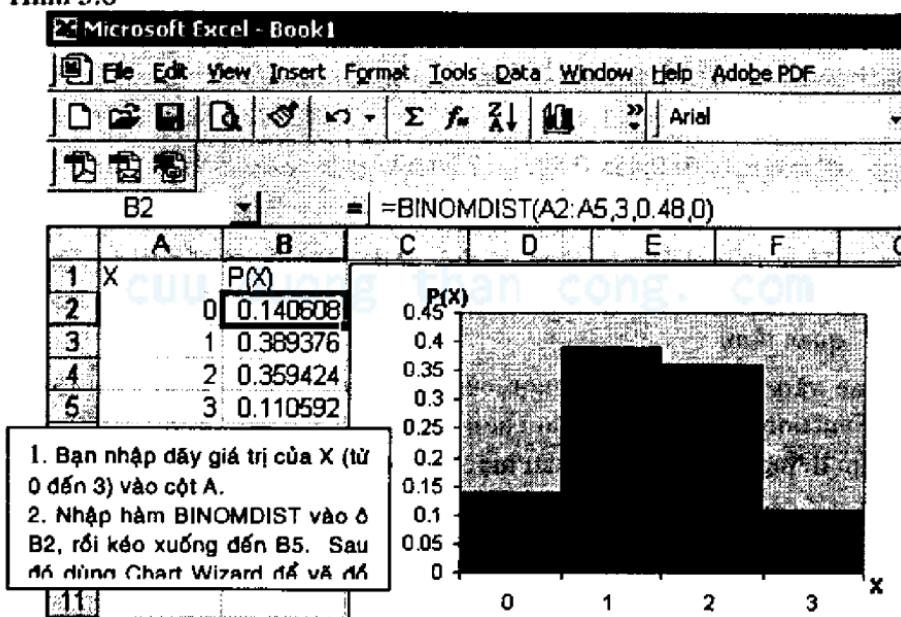
Hình 5.5



Để tính khả năng X nhận giá trị 2 trong ví dụ của chúng ta bạn khai báo như hình trên đây, chú ý rằng trong khung Cumulative nếu bạn nhập số 0 thì Excel tính cho bạn xác suất chính xác để X nhận giá trị 2 tức là khả năng có hai đứa con gái trong ba đứa con, kết quả cũng bằng 0,36. Chú ý rằng bạn có thể thấy ngay kết quả trên cửa sổ lệnh này từ trước khi nhấp nút OK (xem trong hình ovan)

Còn nếu bạn nhập số 1 vào đấy, kết quả của lệnh BINOMDIST lúc này là xác suất tích luỹ $P(X \leq 2)$ chính là khả năng không có quá hai đứa con gái trong ba đứa con, kết quả Excel tìm được (là 0,889408) gần bằng 0,89 cũng chính là đáp số mà chúng ta tính thủ công theo công thức.

Hình 5.6



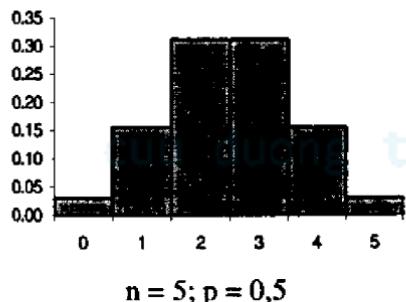
Hình 5.6 trên đây thể hiện hình dáng phân phối của biến X đại diện cho số con gái của cặp vợ chồng, nó khá đối xứng, ở trên hình bạn có thể tìm hiểu cách xây dựng phân phối này một cách tự động bằng Excel.

- Về hình dáng của phân phối Nhị thức

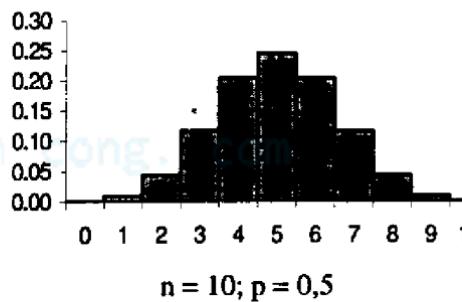
Phân phối này có thể cân đối, nhưng cũng có thể lệch trái hoặc lệch phải. Thực nghiệm chứng minh rằng nếu xác suất thành $p = 0,5$ thì hình dáng của phân phối nhị thức sẽ cân đối dù cho cỡ mẫu là bao nhiêu, điều này thể hiện rõ ở Hình 5.7a với tình huống $n = 5$ và $n = 10$, các phân phối này đối xứng quanh trị trung bình μ của nó.

Khi p khác $0,5$ (lớn hay nhỏ hơn $0,5$) hình dáng của phân phối trở nên lệch, hay p càng gần 0 hoặc 1 thì phân phối càng đặc biệt lệch, tuy nhiên khi cỡ mẫu càng lớn thì phân phối dần trở nên cân đối hơn, thể hiện ở Hình 5.7b.

Hình 5.7a

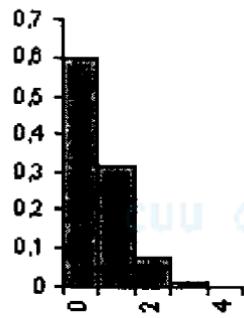


$n = 5; p = 0,5$

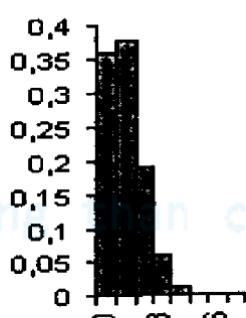


$n = 10; p = 0,5$

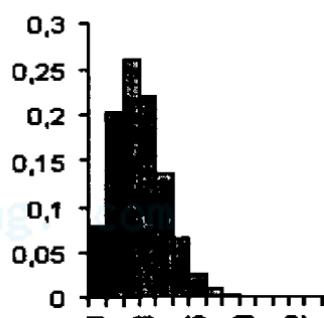
Hình 5.7b



$n = 10; p = 0,05$



$n = 20; p = 0,05$



$n = 50; p = 0,05$

5.3.1.2 Phân phối Poisson (Poisson Distribution)

Phân phối Nhị thức rất hữu dụng trong nhiều tình huống nghiên cứu, tuy nhiên như chúng ta đã xác định trong nội dung nghiên cứu về phân phối Nhị thức, chúng ta phải đảm bảo một số điều kiện mới sử dụng được phân phối này. Nếu điều kiện cho phân phối Nhị thức không thể thoả mãn chúng ta có thể nghĩ đến việc sử dụng một phân phối lý thuyết khác cho biến rời rạc chẳng hạn phân phối Poisson.

Để sử dụng phân phối Nhị thức chúng ta phải đếm được số thành công và số thất bại. Trong các tình huống thực tế chúng ta có thể xác định được số lần thành công, tuy nhiên hầu như khó mà đếm được số thất bại, ví dụ người quản lý đường dây điện thoại cấp cứu khẩn cấp của một thành phố lớn có thể dễ dàng xác định được số cuộc gọi khẩn họ nhận được mỗi giờ đồng hồ nhưng làm sao xác định được những cuộc gọi không thành công trong một giờ là bao nhiêu? Rõ ràng trong những trường hợp này việc xác định số kết cục có thể (thành + bại) là rất khó thậm chí là không thể. Mà nếu không thể xác định được tổng số kết cục thì không thể áp dụng được phân phối Nhị thức, trong những tình huống như vậy chúng ta sẽ dùng Phân phối Poisson.

Phân phối Poisson là một quy luật phân phối rời rạc thích hợp khi bạn quan tâm đến số lần một biến cố cụ thể sẽ xảy ra trong một đơn vị thời gian hay không gian xác định (chẳng hạn như chiều dài, hay diện tích bề mặt...) chúng ta tạm gọi là một phân đoạn (thời gian hay không gian). Ví dụ về các biến số tuân theo định luật phân phối Poisson có thể kể như số lỗi trên một trang đánh máy, số khách hàng đến giao dịch tại Ngân hàng trong mỗi phút vào giờ nghỉ ăn trưa, số cuộc gọi khẩn cấp nhận được mỗi 15 phút... Cũng như khi bạn làm việc với phân phối Nhị thức, bạn có thể gọi việc gấp các kết cục bạn quan tâm là biến cố Thành mặc dù có thể nó là điều mà trong thực tế người ta không mong đợi (ví dụ số lỗi trên một trang đánh máy).

Chúng ta sẽ cần tính xác suất của số lần xảy ra biến cố thành (số lần này dĩ nhiên là những con số nguyên dương như 0, 1, 2..., những con số nguyên dương này là tập các giá trị mà biến ngẫu nhiên rời rạc X của chúng ta nhận), biến số X lúc này là số lần gấp biến cố thành trong phân đoạn). Ví dụ khả năng nhận được đúng 4 cuộc gọi trong mỗi 15 phút là bao nhiêu? Hay xác suất để có đúng 2 lỗi trên một trang đánh máy là bao nhiêu? Chúng ta có thể sử dụng phân phối Poisson để trả lời câu hỏi này nếu chúng ta xác định được những vấn đề sau:

i) Xác suất thành công (khả năng biến cố chúng ta quan tâm xảy ra) trong một phân đoạn là tương tự cho tất cả các phân đoạn có cùng kích thước, ví dụ phân phối xác suất của biến số đại diện cho số các cuộc gọi khẩn cấp nhận được là giống hệt nhau cho bất kỳ phân đoạn 15 phút nào khác.

ii) Số biến cố thành công xảy ra trong một phân đoạn này là độc lập với số biến cố thành công xảy ra trong một phân đoạn khác, ví dụ số cuộc gọi đến trong 15 phút này không chịu ảnh hưởng của số cuộc gọi đến trong 15 phút khác bất kỳ trong ngày.

iii) Thử tưởng tượng bạn chia phân đoạn thành những phân khúc ngày càng nhỏ hơn thì khả năng có hơn một lần thành công (tức là biến cố Thành sẽ xảy ra 2, hoặc 3 lần trở lên trong một phân khúc) sẽ gần như bằng 0, ví dụ nếu bạn chia trang đánh máy A₄ thành bốn phần thì khả năng gấp hai lối trên mỗi $\frac{1}{4}$ trang giấy là không đáng kể, tương tự như vậy, khả năng có 2 cuộc gọi khẩn cấp đến dịch vụ mỗi phút đồng hồ (nếu bạn chia phân đoạn thành 15 phân khúc) về cơ bản là = 0.

iv) Chúng ta phải biết được λ (lambda) là trung bình hay kỳ vọng của số thành công trên một đơn vị được chọn, ví dụ ta biết là trung bình có 8 cuộc gọi khẩn cấp mỗi giờ đồng hồ. Một khi đã xác định được λ chúng ta sẽ xác định được tỷ lệ trung bình của số thành công trong mỗi phân đoạn cỡ t mà ta quan tâm, đó là λt , chú ý rằng λ và t phải cùng đơn vị thời gian. Ở đây ta biết $\lambda=8$ cuộc gọi/giờ thì phân đoạn ta muốn tính toán cũng phải ở dạng giờ, tức là ta đổi 15 phút thành $\frac{1}{4}$ giờ hay $t=1/4$ giờ và $\lambda t=8*1/4=2$ cuộc gọi/15 phút.

Một điều quan trọng cần lưu ý là con số trung bình λt trong mỗi phân đoạn cỡ t không nhất thiết phải là con số chúng ta sẽ thấy nếu chúng ta quan sát tiến trình trong một phân đoạn. Chúng ta sẽ kỳ vọng rằng về trung bình có 2 cuộc gọi đến tổng dài trong mỗi 15 phút đồng hồ bất kỳ nhưng không có nghĩa là sẽ đếm được chính xác con số này trong một khoảng 15 phút ta chọn theo dõi.

▪ Biểu thức toán học của phân phối Poisson

Biểu thức để tính toán xác suất biến ngẫu nhiên X nhận giá trị x với giá trị kỳ vọng (hay trung bình) λt đã biết là :

$$P(X=x) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$$

Trong đó

- X biến ngẫu nhiên rời rạc nhận giá trị là các số nguyên đại diện cho số lần gấp kết cục thành.

- x là giá trị cụ thể của số lần thành công trong một phân đoạn mà ta đang muốn xác định xác suất $P(X=x)$
- λ là số trung bình của thành công trong một phân đoạn
- t là khoảng phân đoạn ta đang quan tâm, t phải cùng đơn vị đo với λ
- e là hằng số toán học xấp xỉ 2,71828

Ví dụ: Biết rằng về trung bình có 3 khách hàng đến giao dịch với Ngân hàng mỗi phút trong khoảng thời gian mở cửa thêm sau giờ làm việc từ 17g chiều đến 19g tối. Vậy thì xác suất để có đúng 2 khách hàng đến giao dịch với Ngân hàng trong mỗi phút vào giờ làm việc thêm là bao nhiêu? Khả năng sẽ có hơn một khách hàng đến trong khoảng thời gian 30 giây vào giờ này là bao nhiêu?

Để trả lời hai câu hỏi này chúng ta sử dụng công thức của phân phối Poisson, nhưng trước khi sử dụng công thức của phân phối Poisson để xác định xác suất chúng ta cần khẳng định lại một số yêu cầu:

Biến cố ta quan tâm là số khách hàng đến Ngân hàng trong một phân đoạn (đây được xác định là mỗi 1 phút). Mỗi phút sẽ có 1, 2 hay 3 ... khách hàng đến?

Hợp lý khi ta giả định rằng xác suất mà khách hàng đến trong 1 phút này cũng như xác suất mà khách hàng đến trong mỗi phút bất kỳ nào khác, và số khách hàng đến giao dịch trong khoảng thời gian 17:32 đến 17:33 không có ảnh hưởng gì đến số khách hàng đến trong khoảng thời gian 17:45 đến 17:46...

Chú ý rằng nếu ta chia nhỏ khoảng thời gian một phút thành 100 phân khúc nhỏ hơn thì hầu như không có khả năng có 2 hoặc hơn 2 khách hàng sẽ đến giao dịch trong khoảng thời gian 0,01 phút (tức xác suất = 0) → như vậy đủ điều kiện để ta áp dụng phân phối Poisson cho việc xác định số khách hàng đến ngân hàng vào sau giờ làm việc và các xác suất tương ứng của chúng.

Vận dụng vào ví dụ của chúng ta, ở trên đã xác định $\lambda = 3$ và t trong tình huống này = 1 do đó λt vẫn bằng 3 thì xác suất có đúng hai khách hàng đến trong 1 phút là

$$P(X=2) = \frac{e^{-3} \times 3^2}{2!} = \frac{9}{2,71828^3 \times 2} = 0,224$$

Để xác định xác suất mà trong mỗi 30 giây bất kỳ có hơn một khách hàng đến, chúng ta xác định xác suất $P(X>1) = P(X=2) + P(X=3) + P(X=4) + \dots + P(X=\infty)$.

Vì tổng tất cả các xác suất trong một phân phối xác suất phải bằng 1, nên phương trình trên được đảo lại để có thể tính toán được như sau :

$$\begin{aligned}P(X > 1) &= 1 - P(X \leq 1) \\&= 1 - [P(X = 0) + P(X = 1)]\end{aligned}$$

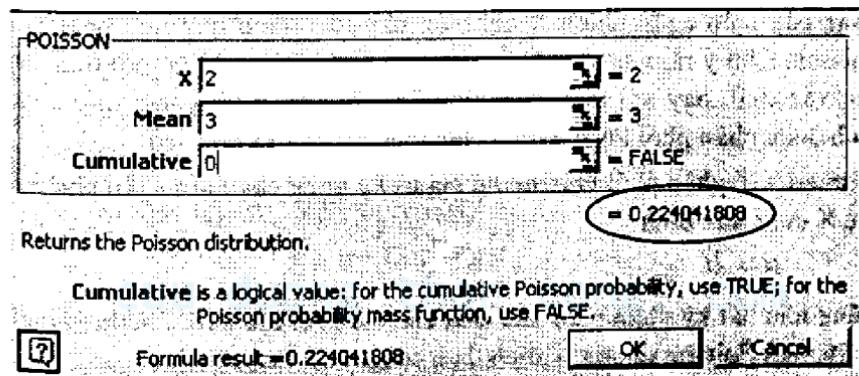
mà trong tình huống này phân đoạn quan tâm là 30 giây tức $\frac{1}{2}$ phút nên ta xác định

$$\begin{aligned}\lambda t &= 3 * 1/2 = 1,5 \\P(X > 1) &= 1 - \left[\frac{e^{-1.5} * 1,5^0}{0!} + \frac{e^{-1.5} * 1,5^1}{1!} \right] \\&= 1 - [0,5578] \\&= 0,4422\end{aligned}$$

Như vậy có một khả năng khoảng 44,2% là hơn một khách hàng trở lên đến giao dịch với ngân hàng mỗi 30 giây trong khoảng thời gian sau giờ làm việc chính thức.

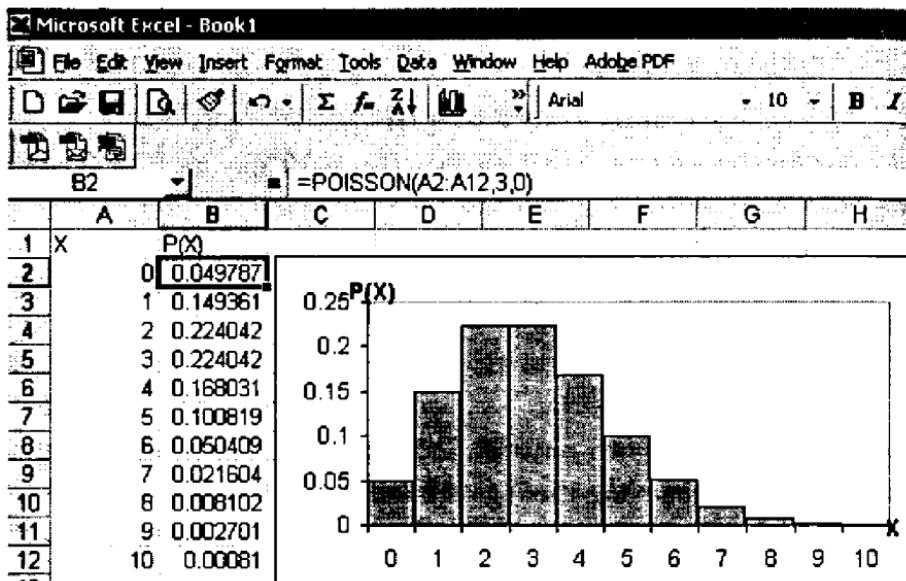
- Nếu bạn sử dụng phần mềm Excel để tính toán nhanh chóng các xác suất bạn sẽ chọn lệnh Poisson trong danh mục lệnh Statistical thuộc mục Function trong menu Insert rồi khai báo như trên hình sau, rồi nhấn nút OK bạn được kết quả (xem trong hình ovan).

Hình 5.8



Đồ thị phân phối xác suất của biến X với các giá trị từ 0 đến 6 lênh phái, trong đó xác suất để có trên 8 khách hàng đến ngân hàng trong vòng một phút vào giờ mở cửa thêm hầu như khó xảy ra.

Hình 5.9



Kỳ vọng và Phương sai của phân phối Poisson

Các bạn hãy nhìn lại Hình 5.9, khung trống sau chữ Mean được nhập giá trị 3, chính là giá trị của đại lượng λt trong công thức xác định xác suất của biến ngẫu nhiên X của ví dụ trên, trở lại ngay từ quá trình dẫn đến đến công thức này bạn cũng có thể hình dung được λt chính là giá trị kỳ vọng của biến ngẫu nhiên X khi biến này tuân theo quy luật phân phối Poisson. Chú ý rằng ta phải biết trước giá trị kỳ vọng thì ta mới tính được các xác suất, hay nói một cách hình ảnh thì giá trị kỳ vọng giúp ta xác định được phân phối Poisson của chúng ta “trông” ra làm sao

Tóm lại, với phân phối Poisson thì giá trị kỳ vọng của biến ngẫu nhiên rời rạc X được xác định

$$\mu = \lambda t$$

Cũng như bất kỳ phân phối xác suất cho biến rời rạc nào, Phương sai của phân phối Poisson được xác định theo công thức như sau

$$\sigma^2 = \sum (x_i - \mu)^2 P(x_i)$$

Với $\mu = \lambda t$, ta có

$$\begin{aligned}\sigma^2 &= \sum (x_i - \lambda t)^2 P(x_i) = \sum [x_i^2 + (\lambda t)^2 - 2(\lambda t)x_i]P(x_i) \\ &= \sum x_i^2 P(x_i) + (\lambda t)^2 \sum P(x_i) - 2(\lambda t) \sum x_i P(x_i)\end{aligned}$$

Chú ý rằng $\sum P(x) = 1$ và $\sum x_i P(x) = \lambda t$

$$\begin{aligned} \text{Ta có } \sigma^2 &= \sum x^2 P(x) + (\lambda t)^2 - 2(\lambda t)^2 \\ &= \sum x^2 P(x) - (\lambda t)^2 \end{aligned}$$

Vậy chúng ta chỉ cần tính được $\sum x^2 P(x)$ là xác định được σ^2

Tiếp tục, từ công thức tính xác suất của phân phối Poisson ta được:

$$\begin{aligned} \sum x^2 P(x) &= e^{-\lambda t} \sum \frac{x^2 (\lambda t)^x}{x!} \\ &= e^{-\lambda t} \left[0 + 1 \frac{\lambda t}{1!} + 4 \frac{(\lambda t)^2}{2!} + 9 \frac{(\lambda t)^3}{3!} + \dots + n^2 \frac{(\lambda t)^n}{n!} \right] \end{aligned}$$

Trong biểu thức trên, các mẫu số được viết dưới dạng $(x!)$, nếu ta tách nó thành $[(x-1)!x]$ thì tại mỗi số hạng tương ứng với x trong biểu thức, ta sẽ lấy x^2 ở tử số chia cho x trong cụm $[(x-1)!x]$ thì còn lại x ở tử số và $(x-1)!$ ở mẫu số. Do đó ta có thể viết

$$\sum x^2 P(x) = e^{-\lambda t} (\lambda t) \left[1 + 2 \frac{\lambda t}{1!} + 3 \frac{(\lambda t)^2}{2!} + 4 \frac{(\lambda t)^3}{3!} + \dots + n \frac{(\lambda t)^{n-1}}{(n-1)!} \right]$$

Tổng số ở trong ngoặc vuông có thể tách ra và xếp lại thành hai tổng riêng như sau:

$$[] = \left[1 + \frac{\lambda t}{1!} + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^3}{3!} + \dots + \frac{(\lambda t)^{n-1}}{(n-1)!} \right] + (\lambda t) \left[1 + \frac{\lambda t}{1!} + \frac{(\lambda t)^2}{2!} + \dots + \frac{(\lambda t)^{n-2}}{(n-2)!} \right]$$

Trong toán học người ta đã chứng minh được rằng, nếu n lớn thì chuỗi số trong dấu ngoặc vuông chính là khai triển của $e^\lambda = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}$

$$\text{Vậy } [] = e^{\lambda t} + (\lambda t)e^{\lambda t}$$

$$\begin{aligned} \text{Và } \sum x^2 P(x) &= e^{-\lambda t} (\lambda t) [e^{\lambda t} + (\lambda t)e^{\lambda t}] \\ &= (\lambda t) + (\lambda t)^2 \end{aligned}$$

$$\begin{aligned} \text{và như vậy } \sigma^2 &= \sum x^2 P(x) - (\lambda t)^2 \\ &= (\lambda t) + (\lambda t)^2 - (\lambda t)^2 \end{aligned}$$

Đưa đến công thức $\sigma^2 = \lambda t$

Ta thấy ngay rằng độ lệch chuẩn của phân phối Poisson ($\sigma = \sqrt{\lambda t}$) chỉ đơn giản là căn bậc hai của giá trị kỳ vọng, vì thế nếu bạn gặp phân phối Poisson, hãy nhớ rằng bạn có thể tìm cách làm giảm độ biến thiên qua việc giảm trung bình λt .

- Về hình dáng đồ thị mô tả quy luật phân phối Poisson, vì phân phối Poisson là một mô thức phân phối thích hợp cho số phần tử tính trong một đơn vị thời gian hay không gian khi số trung bình

của phần tử này trong đơn vị tương đối nhỏ, hình dáng của phân phối này phụ thuộc vào λ và t , λt càng lớn hình dáng của phân phối càng ít lệch.

Ví dụ: Hãng taxi M đã nghiên cứu nhu cầu gọi taxi tại một sân bay địa phương và nhận thấy rằng trung bình mỗi giờ có 6 chiếc taxi được khách gọi. Nếu phòng điều độ của hãng quyết định thường trực đặt 6 xe tại sân bay trong vòng mỗi giờ đồng hồ, hãy xác định khả năng thiếu xe để phục vụ khiến cho có khách phải chờ xe tại sân bay là bao nhiêu?

Ta xác định số taxi được gọi X tuân theo phân phối Poisson.

Vì trung bình mỗi giờ có 6 chiếc taxi được khách gọi:

$$\mu = \lambda t = 6 \times 1 = 6$$

$$\text{Và } \sigma^2 = \lambda t = 6 \times 1 = 6$$

Yêu cầu ở đây là tìm xác suất thiếu xe để phục vụ tức :

$$P(X > 6) = 1 - P(X \leq 6)$$

Dùng Excel ta tính được $P(X \leq 6) = 0,6063$

Vậy $P(X > 6) = 1 - 0,6063 = 0,3937$

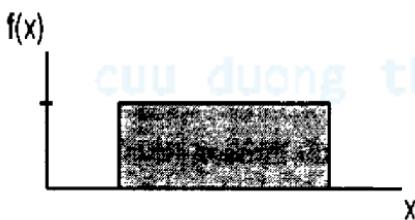
Như vậy có một khả năng tới 39,37% là nhu cầu gọi taxi tại sân bay sẽ vượt quá khả năng phục vụ nếu hãng M chỉ đặt 6 xe tại sân bay.

5.3.2 Phân phối lý thuyết cho biến liên tục

5.3.2.1 Phân phối Bình thường (Normal Distribution)

Ở nội dung trước ta cũng đã biết là biểu thức toán học xác định phân phối của các giá trị của biến số ngẫu nhiên liên tục được gọi tên là hàm mật độ xác suất. Trong nội dung Phân phối lý thuyết cho biến liên tục ta sẽ nghiên cứu 3 phân phối liên tục là phân phối Bình thường, phân phối Đều và phân phối Mũ.

Hình 5.10



Phân phối Đều

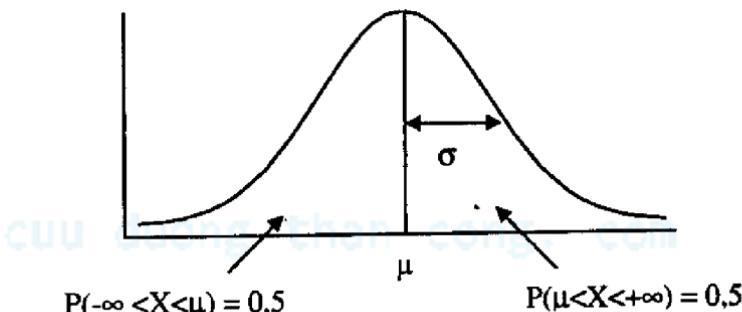


Phân phối Mũ

Phân phối đầu tiên ta xem xét là phân phối Bình thường, đây là kiểu phân phối thông dụng nhất với hình dạng của hàm phân phối là dạng hình

chuông cân đối (Hình 5.11), người ta nhận thấy rằng một số lớn biến số ngẫu nhiên liên tục quan sát được có hàm mật độ xác suất cân đối dạng quả chuông thế này, (chú ý là hình dáng của phân phối Nhị thức khi $p = 0,5$ cũng cân đối nhưng đây là phân phối xác suất của biến ngẫu nhiên rời rạc). Quả chuông cân đối thể hiện phần lớn các giá trị sẽ tập trung quanh trung bình (mà cũng bằng chính Median và bằng Mode), còn độ rộng của chân chuông phụ thuộc vào độ lệch chuẩn, độ lệch chuẩn càng lớn chân chuông càng rộng và ngược lại. Mặc dù trong phân phối Bình thường, các giá trị mà biến số ngẫu nhiên X có thể nhận về nguyên tắc là biến thiên từ trừ vô cùng đến cộng vô cùng nhưng hình dáng của phân phối cho thấy hầu như rất ít có khả năng xuất hiện những giá trị rất lớn hoặc rất nhỏ (giá trị ngoại lệ). Phân phối Bình thường còn có tên gọi khác là phân phối Gauss.

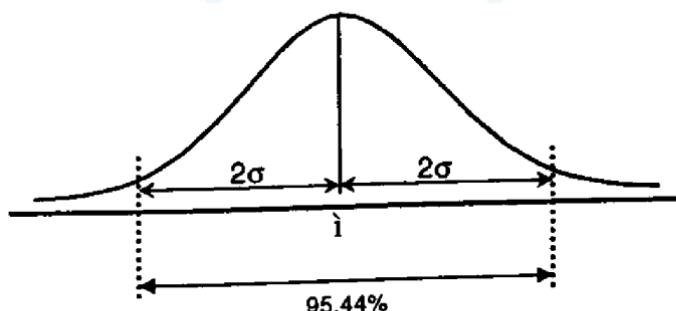
Hình 5.11



Các đặc tính của phân phối Bình thường:

1. Phân phối có dạng đường cong hình chuông cân đối, toàn bộ diện tích dưới đường cong này = 1, vì thế dĩ nhiên diện tích mỗi bên đúng = 0,5
2. Các đại lượng đo lường khuynh hướng tập trung (trung bình, trung vị, Mode) trùng nhau.
3. Khoảng 95% các giá trị mà biến số X nhận tập trung trong vòng hai lần độ lệch chuẩn so với trung bình ($\mu \pm 2\sigma$).
4. Phân phối này biểu diễn các giá trị của biến X trong khoảng $(-\infty ; +\infty)$

Hình 5.12



Nếu biến ngẫu nhiên liên tục X có phân phối xác suất Bình thường với trị trung bình là μ và phương sai là σ^2 (σ là độ lệch chuẩn) thì biểu thức toán học mô tả hàm mật độ xác suất X (ký hiệu $f(x)$) có dạng như sau:

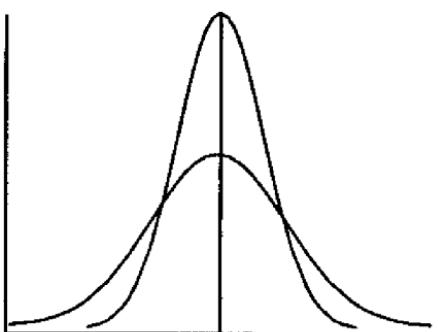
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Trong đó:

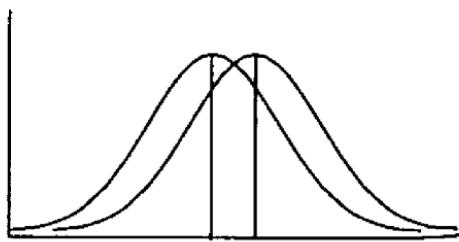
- x là giá trị bất kỳ của biến ngẫu nhiên liên tục X, $(-\infty < x < +\infty)$.
- Đại lượng ngẫu nhiên X phân phối theo quy luật bình thường có μ và σ^2 , được ký hiệu là $X \sim N(\mu ; \sigma^2)$
- $e = 2,71828$
- $\pi = 3,14159$

Trong biểu thức trên chú ý rằng vì e và π là những hằng số toán học nên xác suất tính được phụ thuộc vào hai tham số là μ và σ^2 . Về mặt hình học mỗi sự kết hợp của μ và σ^2 tạo nên những đường cong khác nhau của phân phối Bình thường. Nếu trung bình cố định mà phương sai thay đổi thì hình dạng của phân phối sẽ trở nên nhọn hơn hoặc tù hơn, nếu phương sai không đổi mà trung bình thay đổi thì đường cong của phân phối sẽ dịch sang hai phía.

Hình 5.13 *cuuduongthancong.com*



Phương sai thay đổi



Trung bình thay đổi

Ta nhận thấy đường cong biểu diễn $f(x)$ cực đại khi $x = \mu$ (do lúc này $(e^0 = 1)$ tức là tại giá trị x bằng đúng trị trung bình, khi X nhận những giá trị x lớn hay nhỏ hơn μ thì lượng $(x-\mu)/\sigma$ tăng nên $e^{-1/2((x-\mu)/\sigma)^2}$ giảm, do đó $f(x)$ giảm khi X tăng và giảm nhanh xuống tới 0 theo đà tăng của X tạo thành đường cong, cũng do lượng $(x-\mu)/\sigma$ này mà $f(x)$ có dạng đối xứng và đồ thị có dạng hình chuông).

Như đã biết ở nội dung Phân phối xác suất của biến ngẫu nhiên liên tục, khi muốn tính xác suất để X nhận giá trị trong khoảng $[a;b]$ ta lấy tích phân từ a đến b của biểu thức $f(x)$ để tìm ra giá trị xác suất cần biết $P(a < X < b)$ với các giá trị trung bình và phương sai đã xác định trước, đây thực sự là một công việc nặng nề và nhảm chán. Để khắc phục điều này các nhà thống kê học đã tìm cách xây dựng các bảng số tính toán sẵn để cung cấp các xác suất cần biết, tuy nhiên với các kết hợp bất kỳ (gần như là vô tận) các cặp giá trị μ và σ^2 trong vô vàn tình huống thực tế thì cần lập bao nhiêu bảng mới đủ?

5.3.2.2 Phân phối bình thường chuẩn hóa (Standard Normal Distribution)

Biện pháp chuẩn hóa dữ liệu mà ta đã nghiên cứu ở Chương 4 được các nhà thống kê vận dụng để chỉ cần xây dựng một bảng tra duy nhất cung cấp tất cả những xác suất ta cần tính, bằng cách sử dụng công thức chuẩn hóa đã biết $Z = (X-\mu)/\sigma$ thì bất kỳ một biến ngẫu nhiên bình thường X nào có phân phối Bình thường với trung bình là μ và phương sai là σ^2 cũng có thể được chuyển hóa thành một biến ngẫu nhiên chuẩn hóa Z. Dù cho dữ liệu nguyên thủy của biến X có μ và σ như thế nào đi nữa thì biến Z luôn chỉ có trung bình $\mu=0$ và độ lệch chuẩn $\sigma=1$, tức là phương sai =1, từ đây bạn đọc có thể đoán ra nếu chỉ có một cặp kết hợp duy nhất của trung bình và phương sai thì chỉ cần xây dựng một bảng tra là đủ, ta sẽ trở lại vấn đề này ở phần sau.

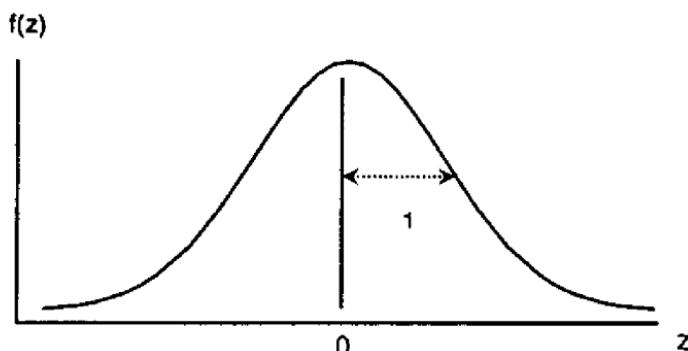
Biến ngẫu nhiên chuẩn hóa Z có phân phối được gọi tên là phân phối Bình thường chuẩn hóa có liên quan mật thiết về bản chất với phân phối Bình thường. Hàm mật độ xác suất của biến số ngẫu nhiên chuẩn hóa Z được gọi là hàm mật độ xác suất của phân phối Bình thường chuẩn hóa với công thức như sau:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

z là giá trị bất kỳ của biến ngẫu nhiên chuẩn hóa Z, $(-\infty < z < +\infty)$. Đại lượng ngẫu nhiên Z phân phối theo quy luật bình thường chuẩn hóa có $\mu=0$ và $\sigma^2 = 1$, được ký hiệu là $Z \sim N(0;1)$.

Hình 5.14 dưới đây biểu diễn đường cong hàm mật độ xác suất của phân phối Bình thường chuẩn hóa với trị trung bình = 0 và độ lệch chuẩn = 1.

Hình 5.14



$$P(-\infty < Z < +\infty) = 1$$

Ta nhận thấy rằng $f(z)$ cực đại khi $Z = 0$ tức tại vị trí của trị trung bình, khi Z dịch sang trái hay phải điêm 0 thì z^2 tăng do đó toàn bộ lượng liên quan đến e giảm, vậy $f(z)$ giảm nhanh đến 0 khi trị số tuyệt đối của z tăng, đường cong $f(z)$ đối xứng vì nó phụ thuộc vào z^2 , nên đường cong này cũng có dạng quả chuông cân đối và một số đặc tính khác như phân phối Bình thường. Như vậy phân phối của Z tương tự như phân phối Bình thường nhưng lại luôn có trung bình bằng 0 và độ lệch chuẩn = 1 nên nó mới được gọi là phân phối Bình thường chuẩn hóa, khái niệm chuẩn hóa này còn nhằm ám chỉ việc chuẩn hóa biến số X bất kì thành một biến số Z mà trung bình luôn = 0 và phương sai luôn = 1. Toàn bộ diện tích dưới đường cong của phân phối Bình thường chuẩn hóa cũng bằng 1.

Sau đây là một ví dụ nhỏ để bạn đọc có thể hình dung được sự chuyển hóa từ biến ngẫu nhiên X thành biến chuẩn hóa Z và hai phân phối tương đương nhau như thế nào: ta có biến số ngẫu nhiên liên tục X có phân phối Bình thường với trung bình bằng 100 và độ lệch chuẩn bằng 50, với một giá trị cụ thể mà biến X nhận là 250, ta tiến hành chuẩn hóa nó thành biến Z như sau:

$$z = \frac{x - \mu}{\sigma} = \frac{250 - 100}{50} = 3$$

Kết quả này cho thấy biến X có giá trị = 250 lệch so với trị trung bình $(250 - 100 = 150)$ về phía bên phải của phân phối đúng 3 lần độ lệch chuẩn $(150/50 = 3)$, ta thấy thực ra chỉ có thang đo là thay đổi còn bản chất vẫn đề như nhau, chúng ta có thể diễn tả vấn đề bằng đơn vị x nguyên thủy hay bằng đơn vị chuẩn hóa z đều được. Chú ý là trong phân phối Bình thường chuẩn hóa độ lệch chuẩn chính là đơn vị đo lường, chúng ta có thể

phát biểu là biến số ta quan sát nhận giá trị ở vị trí trên hoặc dưới trung bình mấy độ lệch chuẩn.

Bây giờ chúng ta quay lại với vấn đề đã khiến ta phải tìm hiểu về phân phối Z, đó là cách tính P($a < X < b$) nhanh nhất thông qua một bảng tính toán được lập sẵn.

Trước hết ta giả sử một tình huống đơn giản là ta có $X \sim N(\mu; \sigma^2)$, tính xác suất để biến X nhận giá trị trong khoảng $[\mu; b]$.

$P(\mu < X < b)$ được xác định theo qui trình sau:

Khi X nhận giá trị = μ thì biến số chuẩn hóa lúc này được xác định

$$Z_\mu = (\mu - \mu)/\sigma = 0$$

Khi X nhận giá trị = b thì biến số chuẩn hóa lúc này được xác định

$$Z_b = (b - \mu)/\sigma$$

Vậy nếu X có trị số trong khoảng $[\mu; b]$ thì biến Z có trị số trong khoảng $[0; Z_b]$. Do đó ta viết lại

$$P(\mu < X < b) = P\left(\frac{\mu - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) = P(0 < Z < Z_b)$$

Lúc này ta chỉ việc sử dụng một bảng tra đã được tính toán sẵn cho các tình huống có thể có của Z_b là có được xác suất muốn tính, vậy nguồn gốc bảng tra này như thế nào?

Theo qui tắc thông thường để tính xác suất của một biến số ngẫu nhiên liên tục ta sẽ lấy tích phân hàm mật độ xác suất của nó, Z được tính toán từ X nên dĩ nhiên nó cũng là một biến ngẫu nhiên liên tục, vì vậy cũng không nằm ngoài qui tắc này, từ đó ta viết được phương trình dưới đây, trong đó vế sau dấu bằng được gọi là tích phân Laplace:

$$P(0 < Z < Z_b) = \frac{1}{\sqrt{2\pi}} \int_0^{Z_b} e^{-z^2/2} dz$$

Với Z_b là một con số cụ thể ta sẽ tính được giá trị xác suất này. Từ ý tưởng đó người ta đã liệt kê tất cả các giá trị có thể xảy ra được với Z_b và tính toán sẵn các xác suất để Z nhận giá trị trong khoảng từ trị trung bình (bằng 0) tới Z_b rồi lập thành bảng tra tên là Bảng phân phối bình thường chuẩn hóa (có thể gọi vẫn tắt là bảng phân phối z) mà chúng ta có thể thấy ở cuối sách trong phần Phụ lục, đó là Bảng tra số 1, bạn đọc có thể tra bảng này để tìm đáp số cho một bài toán $P(0 < Z < Z_b)$ bất kì mà không phải tính toán nhiều.

Xem xét một ví dụ thực tế sau: có $X \sim N(8; 5^2)$, yêu cầu tìm $P(8 < X < 8.6)$.

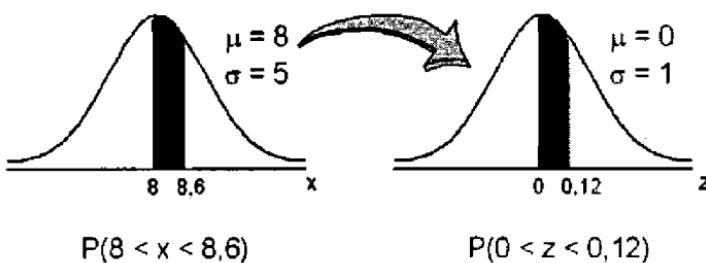
Trước hết ta chuẩn hóa biến X theo qui trình đã xác lập ở trên:

$$Z = \frac{X - \mu}{\sigma} = \frac{8 - 8}{5} = 0$$

$$Z_b = \frac{X - \mu}{\sigma} = \frac{8,6 - 8}{5} = 0,12$$

Lúc này việc tìm xác suất quan tâm trở thành việc tìm diện tích dưới đường cong Bình thường chuẩn hóa từ trị trung bình = 0 cho tới trị số $Z_b = 0,12$; diện tích này cho chúng ta xác suất để Z rơi vào trong khoảng từ 0 tới 0,12; suy ngược trở lại đó là xác suất để X rơi vào trong khoảng từ 8 đến 8,6 tức là: $P(8 < X < 8,6) = P(0 < Z < 0,12)$

Hình 5.15



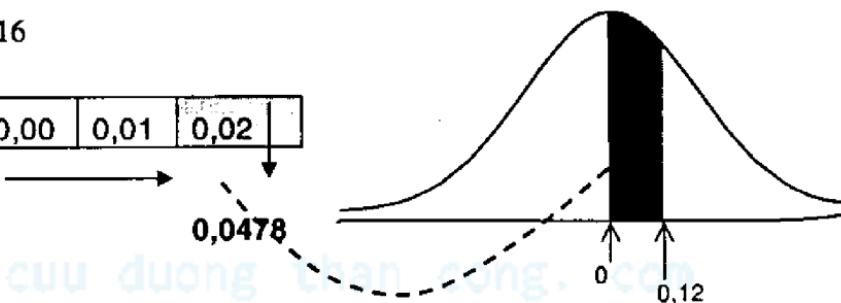
$$P(8 < X < 8,6)$$

$$P(0 < Z < 0,12)$$

Như đã nói, các giá trị xác suất này được tính sẵn và lập thành Bảng tra số 1, bảng tra này cho thấy diện tích dưới đường cong bình thường chuẩn hóa giữa giá trị 0 và trị số $Z = 0,12$ là 0,0478 từ đó suy ngược lại tức là $P(8 < X < 8,6) = 0,0478$. Hình 5.16 trình bày qui tắc tra Bảng tra số 1.

Hình 5.16

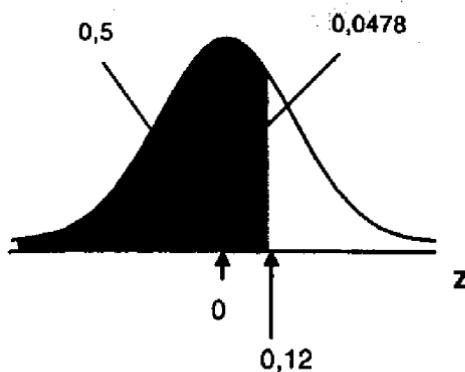
z	0,00	0,01	0,02	
0,0				
0,1				
0,2				



Để hiểu rõ hơn cách sử dụng bảng tra cũng như phương pháp tính toán ta xem thêm một vài ví dụ như sau, cũng vẫn với tình huống của biến số ngẫu nhiên X đã đề cập: Tính $P(X < 8,6)$.

$$P(X < 8,6) = P(Z < 0,12) = P(Z < 0) + P(0 < Z < 0,12) = 0,5 + 0,0478 = 0,5478.$$

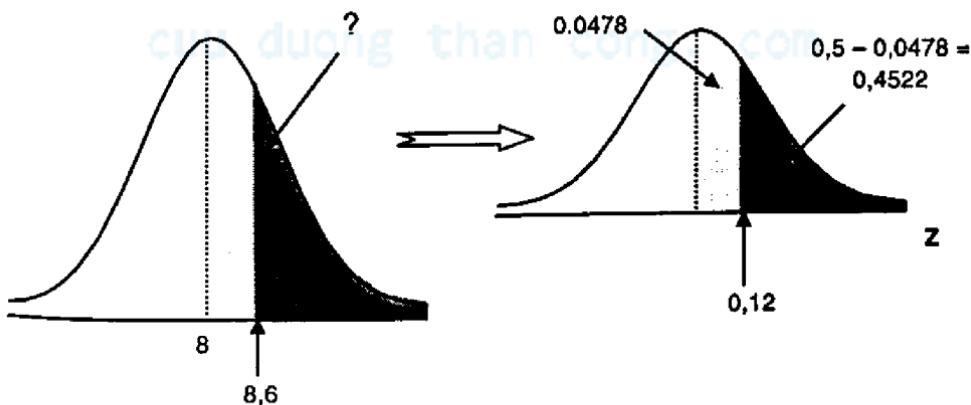
Hình 5.17



* Tính $P(X > 8,6) = ?$

$$\begin{aligned} P(X > 8,6) &= P(Z > 0,12) = P(Z > 0) - P(0 < Z < 0,12) \\ &= 0,5 - 0,0478 = 0,4522 \end{aligned}$$

Hình 5.18

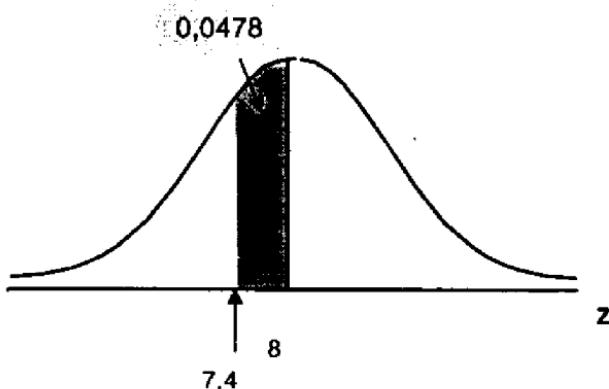


* Tính $P(7,4 < X < 8) = ?$

$$P(7,4 < X < 8) = P(-0,12 < Z < 0) = 0,0478$$

Chú ý là vì phân phối của chúng ta đối xứng nên diện tích tương ứng với khoảng từ $-Z_b$ tới 0 bằng diện tích tương ứng với khoảng từ 0 tới Z_b do đó chúng ta có thể sử dụng cùng Bảng tra số 1 cho những Z_b mang giá trị âm, ta chỉ cần tra theo trị tuyệt đối.

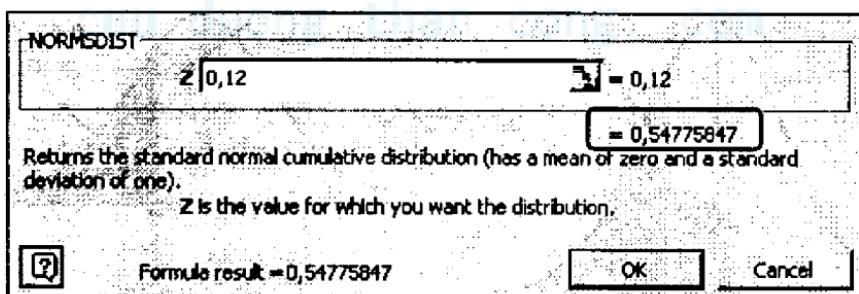
Hình 5.19



Sử dụng Microsoft Excel để tìm các giá trị xác suất của phân phối Bình thường chuẩn hóa:

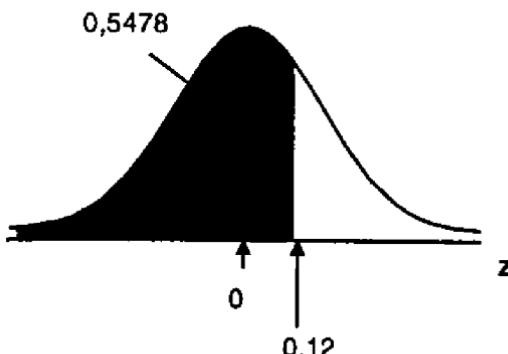
Qui trình cũng tương tự các tìm các phân phối xác suất của các biến rời rạc ta đã nghiên cứu, nhưng ở đây bạn chọn tên lệnh là NORMSDIST để mở hộp thoại Normsdist như hình sau:

Hình 5.20



Ta tiến hành lại việc tìm diện tích dưới đường cong phân phối Bình thường chuẩn hóa cho xác suất $P(0 < Z < 0,12)$ bằng cách nhập giá trị 0,12 vào khung z trên cửa sổ hộp thoại Normsdist, kết quả tìm được là 0,5478. Kết quả này chênh lệch so với kết quả đã tìm được là 0,0478 một lượng = 0,5. Như vậy qui luật của bảng tra Z trên Microsoft Excel là nó trả cho chúng ta toàn bộ diện tích dưới đường cong hàm mật độ xác suất từ vị trí $Z = 0,12$ về $-\infty$ (Hình 5.21), tức là nó cung cấp xác suất $P(Z < 0,12)$, muốn xác định được $P(0 < Z < 0,12)$ ta trừ 0,5 ra khỏi kết quả tìm được là 0,5478. Nấm được qui luật này bạn đọc sẽ tự suy ra các tính các xác suất cần thiết nếu sử dụng bảng tra trên Excel.

Hình 5.21



Chúng ta đã thảo luận xong về phân phối Bình thường và phân phối Bình thường chuẩn hóa, ở nội dung sau chúng ta sẽ nghiên cứu về một ứng dụng rất hữu ích của phân phối Bình thường là dùng chúng để xấp xỉ phân phối Nhị thức và phân phối Poisson nhằm làm cho việc tính toán các giá trị xác suất của hai phân phối rời rạc này trở nên đơn giản hơn.

5.3.2.3 Dùng phân phối Bình thường tính xấp xỉ một số phân phối rời rạc

Xấp xỉ phân phối Nhị thức

Khi cỡ mẫu gia tăng chúng ta sẽ phải mất nhiều công sức hơn để tính những số lũy thừa cao của p và q (tức là của $(1-p)$), và số hạng cần phải tính để cộng lại với nhau cũng nhiều hơn, bạn đọc sẽ lập luận đấy chẳng phải là vấn đề đáng lo vì chúng ta đã có máy vi tính, nhưng có những khi không có sẵn máy vi tính, vì vậy chúng ta cần biết một phương pháp tương đối đơn giản để tính xấp xỉ các giá trị phân phối Nhị thức, phương pháp này dùng tới bảng phân phối Z và gọi là phép tính xấp xỉ bình thường cho phân phối nhị thức.

Chúng ta đã biết hình dáng của phân phối Nhị thức sẽ trở nên cân đối (giống như phân phối Bình thường) khi $p = 0,5$; với $p \neq 0,5$ hình dáng của phân phối sẽ không cân đối nữa, tuy nhiên người ta cũng chứng minh được rằng khi cỡ mẫu càng lớn và giá trị p không quá gần 0 hay 1 thì phân phối Nhị thức sẽ càng trở nên cân đối.

Từ đó quy tắc chung được đưa ra là khi cỡ mẫu đủ lớn (sao cho $np \geq 5$ và $n(1-p) \geq 5$) chúng ta có thể dùng phân phối Bình thường với trung bình bằng np và phương sai bằng $np(1-p)$ để tính xấp xỉ cho phân phối Nhị thức. Lúc này số lượt thành x trong mẫu sẽ phân phối xấp xỉ bình thường với trung bình $\mu = np$ và phương sai $\sigma^2 = np(1-p)$ và chúng ta sẽ dùng sự

phân phối bình thường để tính xác suất mà ta quan tâm tới của biến ngẫu nhiên rời rạc X.

Nếu ta giả định là X phân phối hầu như bình thường thì biến số nhị thức chuẩn hóa

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{npq}} = \frac{X - np}{\sqrt{np(1-p)}}$$

sẽ phân phối xác suất bình thường chuẩn hóa, và xác suất để X nhận giá trị x ở trong khoảng a và b (kể cả a và b) sẽ xác suất bằng xác suất để Z nhận giá trị z trong khoảng Z_a và Z_b với

$$Z_a = \frac{a - np}{\sqrt{np(1-p)}} \quad \text{và} \quad Z_b = \frac{b - np}{\sqrt{np(1-p)}}$$

Phát biểu trên được thể hiện thành biểu thức là $P(a \leq X \leq b) \approx P(Z_a \leq Z \leq Z_b)$

Giả sử chúng ta lấy một mẫu cỡ n=100 từ một tổng thể có phân phối Nhị thức mà tỷ lệ thành là p = 0,36 và chúng ta muốn tính xác suất để số lượng thành trong mẫu sẽ ở trong khoảng từ 24 đến 42 (kể cả 24 và 42). Đối với phân phối trên, các đại lượng trung bình, phương sai và độ lệch chuẩn lần lượt được xác định

$$\mu = np = 100 \times (0,36) = 36$$

$$\sigma^2 = np(1-p) = 100 * 0,36 * (1-0,36) = 23,04$$

$$\sigma = \sqrt{23,04} = 4,8$$

Sử dụng công thức chuẩn hóa $Z = \frac{X - \mu}{\sigma} = \frac{x - 36}{4,8}$

Vậy

$$P(24 \leq X \leq 42) \approx P\left(\frac{24 - 36}{4,8} \leq Z \leq \frac{42 - 36}{4,8}\right) \approx P(-2,5 \leq Z \leq 1,25)$$

Áp dụng qui tắc tính toán và tra bảng đã biết ta tính được $P(-2,5 \leq Z \leq 1,25) = 0,4938 + 0,3944 = 0,8882$

Nếu tính xác suất này đúng theo công thức xác suất Nhị thức hoặc bằng lệnh BINOMDIST ta được kết quả là 0,90738. Như vậy phép tính xác suất sai lệch tới một số thập phân. Khi cỡ mẫu n gia tăng thì độ chính xác của phép tính xác suất trên cũng gia tăng, tuy nhiên chúng ta có thể gia tăng độ chính xác của phép tính xác suất trên khi cỡ mẫu chưa đủ lớn nếu chúng ta chú ý đến một điều là chúng ta đang dùng một phân phối liên tục để tính xác suất cho một biến số rời rạc. Chúng ta đã biết là biến số

ngẫu nhiên rời rạc chỉ có thể nhận một giá trị chính xác (chẳng hạn 24, hay 25 hoặc 40...) trong khi biến số ngẫu nhiên liên tục có thể nhận bất kỳ giá trị nào trong những khoảng giá trị liên tục xung quanh các giá trị cụ thể này. Vì thế khi sử dụng phân phối bình thường xấp xỉ phân phối nhị thức, trong tình huống $n < 50$, kết quả xấp xỉ tìm được sẽ chính xác hơn nếu chúng ta sử dụng phương pháp điều chỉnh tính liên tục cho biến ngẫu nhiên rời rạc X bằng cách trừ bớt đi trị số dưới mà biến X nhận một nửa đơn vị ($x - 0,5$) và cộng thêm vào trị số trên mà biến X nhận một nửa đơn vị ($x + 0,5$), quay lại với ví dụ trên ta có:

$$\begin{aligned} P(24 \leq X \leq 42) &\text{ trở thành } P(24 - 0,5 \leq X \leq 42 + 0,5) \\ &\approx P\left(\frac{23,5 - 36}{4,8} \leq Z \leq \frac{42,5 - 36}{4,8}\right) \\ &= P(-2,6 \leq Z \leq 1,35) = 0,4953 + 0,4115 = 0,9068 \end{aligned}$$

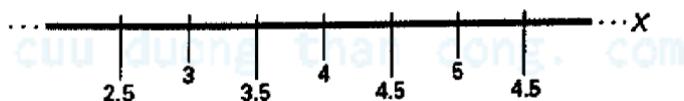
Rõ ràng sai số trong phép tính này nhỏ hơn so với kết quả cũ.

Chú ý một vấn đề nữa là đối với phân phối liên tục, chẳng hạn phân phối Bình thường, xác suất để biến X nhận một giá trị cụ thể là bằng zero, nên khi sử dụng phân phối bình thường để xấp xỉ cho một phân phối rời rạc như phân phối Nhị thức trong tình huống biến rời rạc X nhận một giá trị cụ thể, để đạt được kết quả xấp xỉ tốt nhất ta cũng vận dụng qui tắc trên, cách làm cụ thể như sau

Ví dụ: Tung một đồng xu cân đối đồng chất 20 lần, và quan sát số lần mặt ngửa xuất hiện, gọi biến số X là số mặt ngửa xuất hiện trong 20 lần đồng xu rơi xuống, tính xác xuất để có đúng 4 lần gấp mặt ngửa.

Trong khi biến rời rạc X chỉ có thể nhận một giá trị cụ thể là 4 thì biến liên tục được sử dụng xấp xỉ cho nó có thể nhận giá trị bất kỳ nào đó trong khoảng giá trị quanh 4 (xem hình minh họa sau)

Hình 5.22



Ở đây qui tắc cộng thêm và trừ bớt đi 0,5 khỏi 4 lại được sử dụng, cụ thể chúng ta sẽ phải tìm diện tích dưới đường cong Bình thường từ giá trị $X = 3,5$ đến $X = 4,5$ với phân phối bình thường trong tình huống này có các đặc trưng cơ bản sau:

$$\text{Trung bình } \mu = np = 20 * 0,5 = 10$$

$$\text{Độ lệch chuẩn } \sigma = \sqrt{npq} = \sqrt{20 * 0,5 * 0,5} = 2,24$$

Vậy $P(X = 4)$ trở thành $P(3,5 \leq X \leq 4,5)$

$$\approx P\left(\frac{3,5-10}{2,24} \leq Z \leq \frac{4,5-10}{2,24}\right)$$

$$= P(-2,90 \leq Z \leq -2,46) = 0,4981 - 0,4931 = 0,005$$

Kết quả tìm được bằng lệnh BINOMDIST của $P(X = 4) = 0,0046$

Tóm lại khi dùng phép tính xấp xỉ bình thường chúng ta phải nhớ là bao giờ biến số nhị thức cũng phải nằm trong một phạm vi trị số nào đó, chẳng hạn $n = 95$, $p = 0,91$ và cần tìm xác suất để X nhận giá trị lớn hơn hay bằng 80, chúng ta sẽ viết

$P(X \geq 80) = P(80 \leq X \leq 95)$ vì rõ ràng giá trị x không thể lớn hơn 95 là cỡ mẫu.

$$\text{Với } \mu = 95x(0,91) = 86,45$$

$$\sigma = \sqrt{95x(0,91)(1-0,91)} = 2,789$$

$$\begin{aligned} \text{ta có } P(80 \leq X \leq 95) &\approx P\left(\frac{79,5-86,45}{2,789} \leq Z \leq \frac{95,5-86,45}{2,789}\right) \\ &= P(-2,49 \leq Z \leq 3,24) = 0,4936 + 0,4994 = 0,9930 \end{aligned}$$

Xác suất đúng tính được là 0,9894

Sau cùng, với một p cho sẵn khi cỡ mẫu càng tăng thì phép tính xấp xỉ bình thường càng chính xác. Độ chính xác còn tùy thuộc ở p , nếu p gần bằng 0,5 thì sai số sẽ nhỏ dù cho các mẫu tương đối nhỏ, nhưng khi p tiến tới 0 hoặc 1, để có một độ chính xác định trước người ta thường phải gia tăng cỡ mẫu.

Xấp xỉ phân phối Poisson

Phân phối bình thường cũng có thể được sử dụng để xấp xỉ phân phối Poisson với điều kiện trị trung bình λ (số lần thành công được kỳ vọng) lớn hơn hay bằng 5.

Vì trung bình và phương sai của phân phối Poisson là tương đương $\mu = \sigma^2 = \lambda$, nên độ lệch chuẩn $\sigma = \sqrt{\lambda}$, và biến số X được chuẩn hóa thành biến Z lúc này được tính như sau

$$Z = \frac{X - \mu}{\sigma} = \frac{x - \lambda}{\sqrt{\lambda}}$$

với λ lớn thì biến ngẫu nhiên Z là xấp xỉ phân phối Bình thường.

Để tìm xác suất xấp xỉ tương đương với giá trị mà biến rời rạc X của phân phối Poisson nhận chúng ta cũng sử dụng phương pháp cộng hoặc trừ thêm 0,5. Cụ thể hóa qua một ví dụ như sau:

Ví dụ: Tại một nhà máy, số lần ngừng việc trung bình mỗi ngày vì những vấn đề liên quan đến máy móc trong quá trình sản xuất là 12. Xác định xác suất để có không quá 15 lần ngừng việc vì hỏng máy trong một ngày làm việc bất kỳ.

Lúc này các đặc trưng của phân phối bình thường dùng để xấp xỉ phân phối Poisson được xác định như sau

$$\mu = \sigma^2 = \lambda = 12$$

Biến ngẫu nhiên liên tục Z được chuẩn hóa từ biến ngẫu nhiên rời rạc X đại diện cho số lần gấp kết cục thành công đã được điều chỉnh tính liên tục là 15,5 được tính như sau:

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} = \frac{15,5 - 12}{\sqrt{12}} = 1,01$$

$$P(X \leq 15) \approx P(Z \leq 1,01) = 0,8438$$

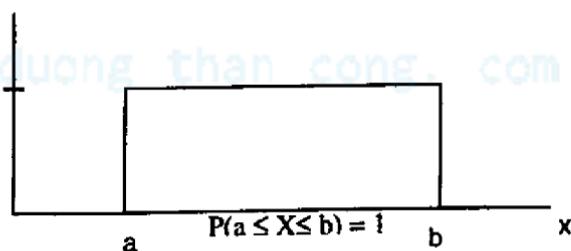
Như vậy xác suất gần đúng để có không quá 15 lần ngừng việc là 0,8438, còn giá trị xác suất chính xác tính theo công thức của phân phối Poisson là 0,8444.

Phân phối Bình thường là phân phối được sử dụng phổ biến nhất trong thống kê, tuy nhiên còn có các phân phối xác suất liên tục khác cũng hay gấp là phân phối đều và phân phối mũ. Nội dung kế tiếp chúng ta sẽ tìm hiểu 2 phân phối này.

5.3.2.4 Phân phối đều (Uniform distribution)

Phân phối đều là một phân phối mà xác suất xảy ra như nhau cho mọi kết cục của biến ngẫu nhiên. Phân phối đều đôi khi còn được gọi là phân phối "hình chữ nhật", bạn sẽ thấy điều này dễ hình dung hơn khi xem hình biểu diễn hình dáng của phân phối đều dưới đây

Hình 5.23 $f(x)$



Vì thế, phân phối đều có biểu thức khá đơn giản như sau:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{Nếu } a \leq x \leq b \\ 0 & \text{khác} \end{cases}$$

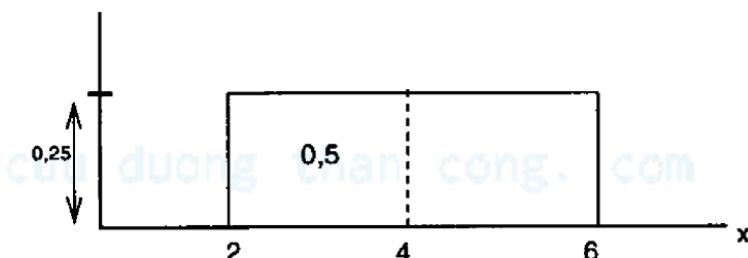
Trong đó

- $f(x)$ là giá trị của hàm mật độ xác suất
- a là giá trị nhỏ nhất mà X nhận
- b là giá trị lớn nhất mà X nhận

Chú ý phân phối đều cũng là 1 phân phối xác suất đơn thuần nên nó cũng thỏa mãn đặc điểm là toàn bộ diện tích dưới đường cong của hàm mật độ xác suất bằng 1.

Ví dụ: Biết ngẫu nhiên X có phân phối đều, với giá trị biến thiên trong phạm vi từ 2 đến 6. Chiều cao của phân phối này bằng $1/(6-2) = 0,25$.

Hình 5.25 $f(x)$



Còn các đặc trưng của phân phối này được tính như sau:

$$\mu = (a+b)/2$$

$$\sigma^2 = (b-a)^2/12$$

Vậy ta có thể xác định được $P(2 \leq X \leq 4) = 0,25 \cdot (4-2) = 0,5$

Ví dụ: Một tổ chức chuyên cung cấp dịch vụ cho hàng hàng không P quan tâm đến khoảng thời gian cần thiết tính từ lúc một chiếc máy bay hạ cánh để bốc dỡ, vệ sinh và chuẩn bị cho nó sẵn sàng cất cánh trở lại. Họ tìm hiểu thông tin từ nhà điều hành sân bay và biết rằng khoảng thời gian này dao động trong khoảng từ 15 đến 45 phút. Không có thêm thông tin gì khác nên nhà cung cấp dịch vụ này quyết định áp dụng phân phối đều cho khoảng thời gian tháo dỡ và chuẩn bị. Từ đó họ xác định được những vấn đề sau:

$$f(x) = 1/(b-a) = 1/(45-15) = 1/30 = 0,0333$$

$$\mu = (a+b)/2 = (15 + 45) / 2 = 30$$

$$\sigma^2 = (b-a)^2/12 = (45 - 15)^2/12 = 75$$

$$\sigma = \sqrt{75} = 8,66$$

Như vậy thời gian tháo dỡ và chuẩn bị trung bình khoảng 30 phút, độ lệch chuẩn là 8,66 phút.

Muốn biết khả năng để khoảng thời gian này không vượt quá 35 phút, họ sẽ tính $P(15 \leq X \leq 35)$.

$$P(15 \leq X \leq 35) = 0,0333 (35-15) = 0,666$$

Như vậy xác suất để khoảng thời gian này không vượt quá 35 phút là gần 67%.

5.3.2.5 Phân phối mũ (Exponential distribution)

Hình 5.10 biểu diễn một dạng của phân phối Mũ, ở nội dung này ta sẽ nghiên cứu sâu về phân phối này, như bạn thấy trên hình thì phân phối mũ là một phân phối liên tục có đặc điểm lệch phải và biến thiên từ zéro đến dương vô cùng. Phân phối mũ hay được sử dụng trong việc đánh giá quá trình sản xuất và cung cấp dịch vụ, nó được vận dụng rộng rãi trong lý thuyết về hàng chờ để mô hình hóa độ dài khoảng thời gian trôi qua giữa các lần xảy ra sự kiện ví dụ như thời gian giữa những lần xe tải đến bến dỡ hàng, thời gian giữa những lần giao dịch của máy ATM, thời gian giữa những lần thực khách bước vào cửa hàng fastfood, thời gian giữa những cuộc gọi đến một tổng đài điện thoại...

Hàm mật độ xác suất của phân phối Mũ có dạng

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

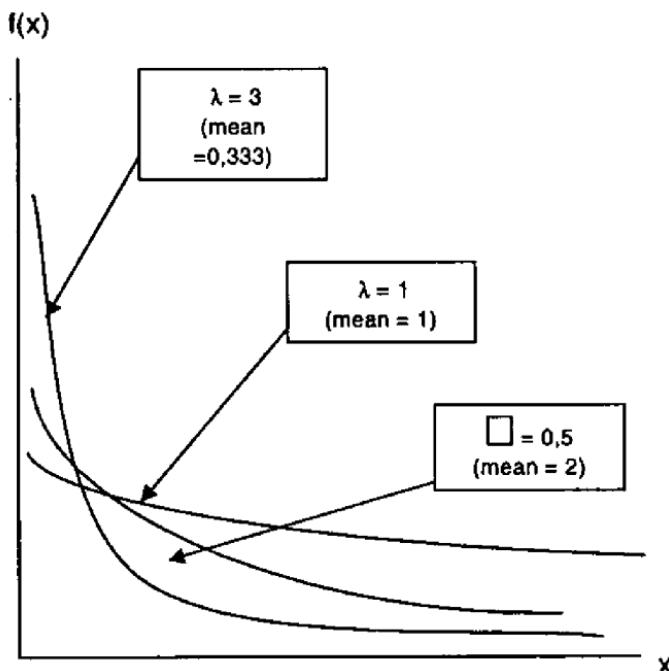
Trong đó: $e = 2,71828$

$1/\lambda$ là thời gian trung bình giữa những sự kiện ($\lambda > 0$)

Chú ý rằng tham số λ trong hàm mật độ xác suất của phân phối Mũ chính là giá trị trung bình ta đã biết trong phân phối Poisson, nó thể hiện số lần xảy ra sự kiện trên mỗi đơn vị thời gian, vậy thì khoảng thời gian trung bình giữa các lần sự kiện xảy ra được tính bằng $1/\lambda$, nó chính là trị trung bình của phân phối Mũ. Chú ý khác là độ lệch chuẩn của bất kỳ một phân phối Mũ nào cũng bằng chính trung bình của nó tức bằng $1/\lambda$.

Nếu ta chọn được một giá trị λ ta có thể vẽ được phân phối Mũ bằng cách thay λ và những giá trị x khác nhau vào hàm mật độ xác suất. Hình dưới đây biểu diễn hàm phân phối Mũ cho các giá trị λ lần lượt bằng 3, 1 và 0,5.

Hình 5.26



Từ hàm mật độ xác suất của phân phối Mũ ta thấy $f(0) = \lambda$ và khi x tăng lên thì $f(x)$ giảm dần đến tiệm cận 0.

Công thức sau hay được sử dụng để tìm xác suất mà X bằng hoặc bé hơn một giá trị a cụ thể tức xác suất $P(0 \leq X \leq a)$:

$$P(0 \leq X \leq a) = 1 - e^{-\lambda a}$$

Xem ví dụ sau: Các khách hàng đến máy ATM với cường độ 20 người mỗi giờ. Nếu một khách hàng vừa đến, cho biết xác suất để khách hàng kế tiếp sẽ đến trong vòng 6 phút?

Ta biết cường độ 20 người/giờ chính là tham số λ trong hàm mật độ xác suất của phân phối Poisson, thời gian giữa những lần đến có phân phối Mũ với trị trung bình (chính là thời gian trung bình giữa các lần đến) bằng $1/\lambda = 1/20 = 0,05$ (giờ).

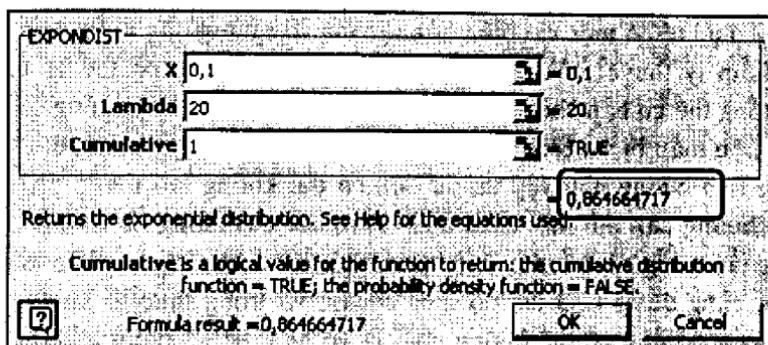
Vấn đề ta quan tâm là tìm xác suất $P(0 \leq X \leq 0,1)$ vì 6 phút được đổi ra đơn vị giờ là 0,1 giờ vậy với $\lambda = 20$ và $a = 0,1$ ta vận dụng công thức trên như sau:

$$P(0 \leq X \leq 0,1) = 1 - e^{-(20)(0,1)} = 1 - 0,1353 = 0,8647$$

Như vậy xác suất mà khách hàng kế tiếp sẽ đến trong vòng 6 phút là 86,47%

Ta thực hiện lại ví dụ trên bằng cách sử dụng chương trình Excel, cách tìm giá trị xác suất ta quan tâm bằng chương trình Excel cũng vẫn tiến hành theo cách ta đã quen thuộc, ta chọn lệnh EXPONDIST để mở cửa sổ hộp thoại dưới này:

Hình 5.27



Nhập các giá trị cần thiết vào vị trí phù hợp như trên hình, trong khung Cumulative ta nhập giá trị 1 vì ta đang cần tìm $P(0 \leq X \leq 0,1)$ kết quả nhận được là 0,864664717.

Còn nếu muốn tìm $P(X = 0,1)$ tức $f(0,1)$ ta nhập giá trị 0 vào khung Cumulative, lúc này ta được đáp số là 2,706705665, bạn hãy thử kiểm tra lại đáp số này bằng cách thay các giá trị phù hợp vào hàm mật độ xác suất và tiến hành tính toán với máy tính tay.

5.3.2.6 Kiểm tra một tập dữ liệu bất kỳ có phân phối bình thường hay xấp xỉ bình thường không?

Chú ý rằng không phải tất cả các biến số ngẫu nhiên liên tục đều có phân phối bình thường, mà thường là chỉ xấp xỉ bình thường vì thế trong phân tích mô tả một tập dữ liệu cụ thể chúng ta luôn có một câu hỏi rất thực tế là : Bằng cách nào chúng ta có thể xác định tập dữ liệu của chúng ta có phân phối gần như bình thường hay không. Có hai phương pháp thăm dò điều này:

- Phương pháp thứ nhất là so sánh các đặc điểm của tập dữ liệu với thuộc tính cơ bản của phân phối bình thường
- Phương pháp thứ hai là xây dựng đồ thị xác suất bình thường (Normal probability plot)

Ta lần lượt đi sâu vào từng phương pháp :

Phương pháp I: Đánh giá các đặc điểm của tập dữ liệu

Phân phối bình thường là một phân phối hình chuông cân đối, tất cả các đại lượng đo lường khuynh hướng tập trung đều bằng nhau, có khoảng từ phân vị (độ tráí giữa) bằng 1,33 lần độ lệch chuẩn, và khoảng biến thiên của giá trị (tức độ tráí rộng của chân chuông) xem như là vô hạn.

Trong thực tế có một vài biến liên tục mà có những đặc điểm xấp xỉ các đặc điểm lý thuyết của phân phối bình thường, lý do có thể là vì phân phối tổng thể cơ bản chỉ xấp xỉ phân phối bình thường hoặc cũng có thể là vì dữ liệu mẫu bị sai lệch so với các đặc điểm lý thuyết mong đợi. Trong những tình huống như vậy thì dữ liệu có thể không tạo thành phân phối hình chuông cân đối một cách hoàn hảo, các đại lượng đo lường khuynh hướng tập trung có thể khác biệt nhau chút ít, và khoảng từ phân vị có thể không bằng đúng 1,33 lần độ lệch chuẩn, và trong thực tế khoảng biến thiên của dữ liệu có thể không phải là vô hạn mà sẽ bằng khoảng 6 lần độ lệch chuẩn.

Nếu những biến số ngẫu nhiên không có phân phối bình thường hoặc cũng không phải xấp xỉ phân phối bình thường thì các thuộc tính mô tả tập dữ liệu sẽ không phù hợp với 4 đặc điểm của phân phối bình thường chúng ta vừa kể trên.

Cách tiếp cận đầu tiên trong việc kiểm tra tính phân phối bình thường của tập dữ liệu là so sánh các đặc điểm thực tế của tập dữ liệu với các đặc điểm tương ứng trong phân phối bình thường, bao gồm các việc như sau:

1. Vẽ đồ thị và quan sát kiểu cách của chúng. Với một tập dữ liệu nhỏ hay có qui mô hạn chế ta có thể vẽ biểu đồ thân - lá hoặc biểu đồ hộp và râu. Với tập dữ liệu có cỡ lớn ta xây dựng Histogram
2. Tính toán các đại lượng thống kê mô tả và so sánh với các đặc điểm lý thuyết của phân phối bình thường. Xác định trung bình và trung vị, so sánh sự đồng nhất hay khác biệt của các đại lượng này. Xác định khoảng từ phân vị và độ lệch chuẩn, xem xét xem khoảng từ phân vị gần bằng 1,33 lần độ lệch chuẩn tới mức nào. Tính khoảng biến thiên và xem nó có đúng là xấp xỉ 6 lần độ lệch chuẩn không.
3. Xem xét phân phối của các giá trị trong tập dữ liệu, tính xem có phải khoảng 2/3 số quan sát tập trung trong phạm vi ± 1 độ lệch chuẩn so với trung bình; 4/5 số quan sát tập trung trong phạm vi $\pm 1,28$ độ lệch chuẩn so với trung bình; 95% số quan sát tập trung trong phạm vi ± 2 độ lệch chuẩn so với trung bình.

Phương pháp 2: Xây dựng đồ thị xác suất bình thường

Cách tiếp cận thứ hai là bằng phương pháp xây dựng đồ thị xác suất bình thường. Quy luật là nếu những điểm được vẽ trên đồ thị xác suất bình thường nằm gần như trên một đường thẳng hoặc rất gần một đường thẳng vô hình hướng từ phía dưới bên tay trái đến phía trên bên tay phải của đồ thị thì có nghĩa là dữ liệu có phân phối bình thường hoặc xấp xỉ phân phối bình thường. Ngược lại nếu những điểm dữ liệu này chệch đi so với đường thẳng vô hình kia và làm thành những kiểu hình dạng cụ thể thì dữ liệu không có phân phối bình thường. Để xây dựng đồ thị xác suất bình thường ta đi theo tiến trình sau :

1. Sắp xếp trật tự của các giá trị trong tập dữ liệu theo thứ tự tăng dần
2. Tính các giá trị chuẩn hóa tương đương với diện tích dưới đường cong bình thường được xác định từ công thức $i/(n+1)$ với i là vị trí của giá trị trong tập dữ liệu gồm n quan sát
3. Vẽ lên đồ thị từng cặp giá trị một của các quan sát thực tế trên trực đứng và các giá trị chuẩn hóa tương ứng của nó trên trực ngang
4. Đánh giá khả năng có phân phối bình thường của tập dữ liệu qua việc xem xét xem các điểm phân tán trên đồ thị có gần như tạo thành một đường thẳng hay không.

Ví dụ: Có tập dữ liệu về điểm thi kết thúc học kì môn Anh văn của 19 học sinh (thang điểm 100), được sắp xếp theo trật tự tăng dần như Bảng 5.9.

Trên bảng số liệu này, lần lượt thực hiện các bước sau

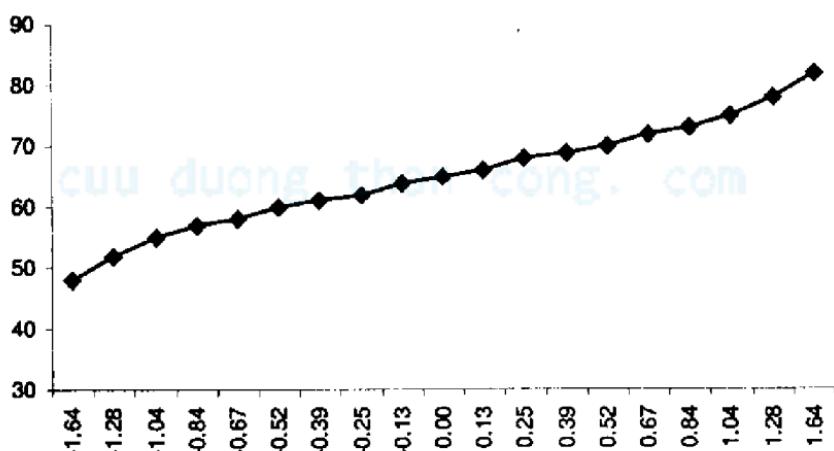
- Đánh số thứ tự theo trật tự tăng dần của dữ liệu
- Tính các giá trị chuẩn hóa tương đương với diện tích dưới đường cong bình thường được xác định từ công thức $i/(n+1)$ với i là vị trí của giá trị trong tập dữ liệu gồm $n=19$ quan sát. Kết quả được ghi trong cột thứ 3, giả dụ
- Tại quan sát thứ nhất, $i=1 \rightarrow 1/(19+1) = 0,05$. Để xác định giá trị chuẩn hóa Z^* sao cho xác suất $P(Z \leq Z^*) = 0,05$ ta dùng hàm =NORMSINV(0,05) trên Excel, được kết quả là -1,64
- Tại quan sát thứ hai, $i=2 \rightarrow 2/(19+1) = 0,1$. Để xác định giá trị chuẩn hóa Z^* sao cho xác suất $P(Z \leq Z^*) = 0,1$ ta dùng hàm =NORMSINV(0,1) trên Excel, được kết quả là -1,28
- Vẽ lên đồ thị từng cặp giá trị của các quan sát thực tế (Điểm) trên trực đứng và các giá trị chuẩn hóa tương ứng của nó (Z^*) trên trực ngang. Xem Hình 5.28

- Xem xét xem các điểm phân tán trên đồ thị ta thấy nó gần như tạo thành một đường thẳng nên có thể kết luận tập dữ liệu gốc có phân phối gần như bình thường.

Bảng 5.12

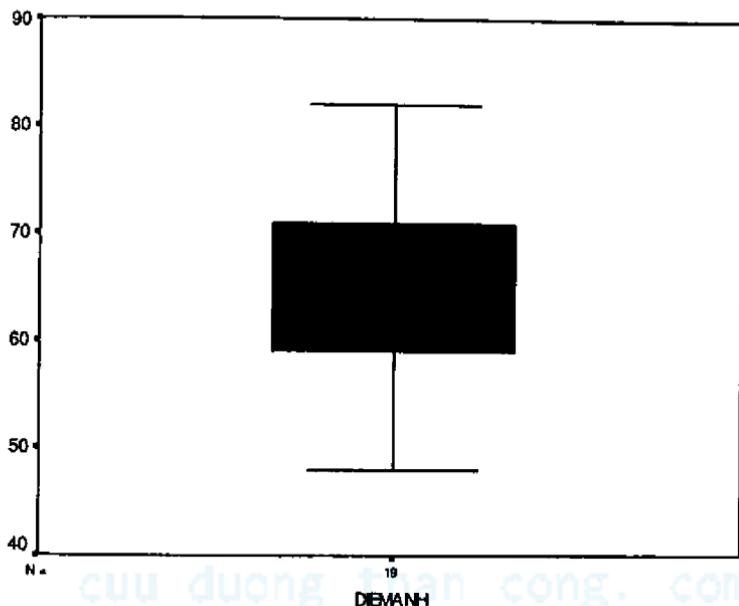
Điểm	Thứ tự	$P(Z \leq Z^*)$	Z^*
48	1	0.05	-1.64
52	2	0.1	-1.28
55	3	0.15	-1.04
57	4	0.2	-0.84
58	5	0.25	-0.67
60	6	0.3	-0.52
61	7	0.35	-0.39
62	8	0.4	-0.25
64	9	0.45	-0.13
65	10	0.5	0.00
66	11	0.55	0.13
68	12	0.6	0.25
69	13	0.65	0.39
70	14	0.7	0.52
72	15	0.75	0.67
73	16	0.8	0.84
75	17	0.85	1.04
78	18	0.9	1.28
82	19	0.95	1.64

Hình 5.28



Nếu chúng ta dùng phương pháp khảo sát các đặc điểm của tập dữ liệu, với cùng tập dữ liệu về điểm môn Anh văn của 19 học sinh, ta dựng biểu đồ Hộp và râu trông như hình sau:

Hình 5.29



Biểu đồ Hộp và râu trông rất cân đối, là dấu hiệu cho thấy dữ liệu gốc có phân phối xem như bình thường, bạn đọc có thể tự khảo sát các dấu hiệu khác căn cứ trên các đại lượng thống kê mô tả được tính trong bảng sau đây để củng cố nhận định của mình.

Bảng 5.13: Các đại lượng thống kê mô tả cho Dữ liệu Điểm môn Anh văn

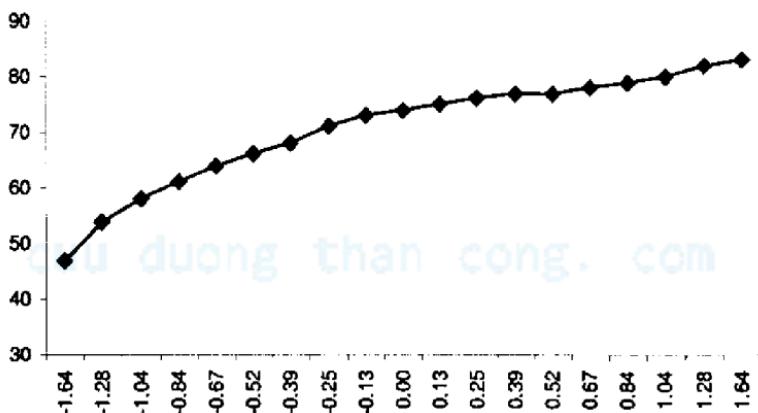
Mean	65
Standard Error	2.0548
Median	65.0000
Mode	N/A
Standard Deviation	8.9567
Sample Variance	80.2222
Kurtosis	-0.4489
Skewness	0
Range	34
Minimum	48
Maximum	82
Sum	1235
Count	19

Bạn đọc xem một tập dữ liệu thứ 2 về điểm môn Toán của 19 học sinh này, tập dữ liệu này cũng được sắp trật tự tăng dần

47 54 58 61 64 66 68 71 73 74 75 76 77 77 78 79 80 82

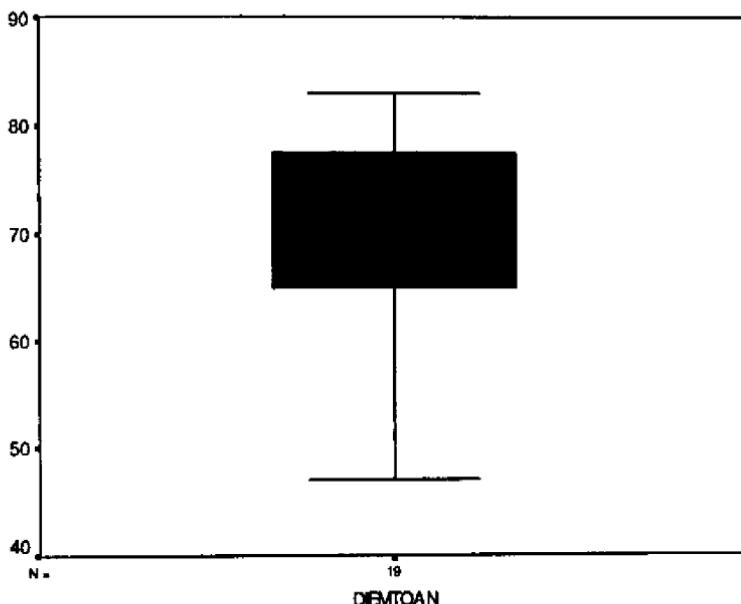
Dưới đây là đồ thị xác suất bình thường xây dựng cho tập dữ liệu điểm môn Toán của 19 học sinh, đồ thị cho thấy một xu hướng không phải là đường thẳng trong trật tự sắp xếp của các điểm kết hợp. Đây là dấu hiệu của sự lệch trái trong phân phối dữ liệu.

Hình 5.30



Khảo sát biểu đồ hộp và râu cung cho kết luận tương tự.

Hình 5.31



cuu duong than cong. com

CHƯƠNG 6

PHÂN PHỐI CỦA CÁC THAM SỐ MẪU

Một nhiệm vụ quan trọng trong thống kê là suy diễn thống kê, có nghĩa là sử dụng những tham số thống kê tính toán được từ mẫu nghiên cứu để ước lượng những tham số của tổng thể. Trong chương này chúng ta sẽ nghiên cứu về trung bình mẫu, một tham số thống kê được sử dụng để ước lượng trung bình tổng thể (là một tham số của tổng thể). Chúng ta cũng nghiên cứu về tỷ lệ mẫu, một tham số thống kê được sử dụng để ước lượng tỷ lệ tổng thể (cũng là một tham số của tổng thể). Chú ý rằng quan tâm chính của chúng ta khi thực hiện thống kê là rút ra kết luận về tổng thể chứ không phải về mẫu. Ví dụ người tiến hành điều tra dư luận xã hội chỉ quan tâm đến kết quả trên mẫu như một phương tiện để họ có thể ước lượng được tỷ lệ phần trăm người dân có hài lòng với các thủ tục hành chính như trong việc xin cấp chủ quyền nhà đất. Tương tự như vậy, giám đốc sản xuất của một công ty chuyên sản xuất mì ăn liền sử dụng thông tin về trọng lượng gói mì trung bình được tính toán từ một mẫu các gói mì ăn liền được chọn ngẫu nhiên để ước lượng trọng lượng trung bình của tất cả các gói mì được sản xuất như một tổng thể.

Với một mẫu ngẫu nhiên lấy ra từ một tổng thể, những tham số thống kê mà chúng ta tính được từ mẫu để suy diễn về tổng thể, là những hàm số của các giá trị của từng quan sát trong mẫu, do đó các tham số này cũng là biến số ngẫu nhiên. Là biến số ngẫu nhiên thì chúng cũng có phân phối, phân phối của các tham số mẫu được gọi chung là phân phối mẫu. Hãy xem trung bình mẫu, trị trung bình trên một mẫu là trung bình cộng của toàn bộ các giá trị của các quan sát ngẫu nhiên được chọn vào trong mẫu này. Nếu bạn tình cờ chọn được một mẫu khác, trị trung bình sẽ được tính từ các giá trị của các quan sát ngẫu nhiên khác, nếu bạn có thể lấy tất cả các mẫu ngẫu nhiên có cỡ n từ một tổng thể gồm N đơn vị, bạn sẽ thấy các trị trung bình tính được thay đổi từ mẫu này sang mẫu khác và do đó tất cả các trị trung bình mẫu tương ứng với tất cả các mẫu cùng một cỡ mà ta có thể lấy từ một tổng thể hợp thành một tập hợp các giá trị mà một biến số ngẫu nhiên có thể nhận. Cũng như các biến số ngẫu nhiên bất kỳ, biến số ngẫu nhiên nhận giá trị là các trị trung bình mẫu cũng có phân phối mà ta gọi đầy đủ là phân phối của trị trung bình mẫu.

6.1 PHÂN PHỐI CỦA TRUNG BÌNH MẪU

6.1.1 Trung bình mẫu là ước lượng không chêch của trung bình tổng thể

Trong Chương 4 chúng ta đã thảo luận về nhiều đại lượng đo lường độ tập trung, ta cũng biết trị trung bình là đại lượng đo lường độ tập trung tiêu biểu nhất. Ở nội dung này ta sẽ nghiên cứu đặc điểm trung bình mẫu là ước lượng không chêch của trung bình tổng thể. Việc thảo luận sâu về vấn đề này nằm ngoài phạm vi của quyển sách này, nhưng ta có thể hiểu một cách đơn giản rằng: Trung bình mẫu được cho rằng là ước lượng không chêch của trung bình tổng thể vì giá trị trung bình tính được từ tất cả các trị trung bình mẫu của các mẫu cỡ n có thể lấy được từ tổng thể cỡ N sẽ bằng đúng trị trung bình của tổng thể. (Ước lượng không chêch là đặc điểm của một tham số thống kê trong đó giá trị trung bình tính được từ tất cả những giá trị có thể của tham số thống kê mẫu bằng đúng tham số tổng thể).

Tính chất này có thể được chứng minh theo cách thực nghiệm với một ví dụ như sau, giả sử có 4 người với 4 độ tuổi làm thành một tập dữ liệu tổng thể về đặc điểm Tuổi có cỡ $N = 4$ như sau {18, 20, 22, 24}

Muốn tính trị trung bình tổng thể về tuổi ta áp dụng công thức $\mu = \Sigma X_i / N$ thay thế số liệu ta có $\mu = (18+20+22+24)/4 = 21$ tuổi

Nếu tất cả các mẫu cỡ 2 người được chọn theo kiểu có hoàn lại từ tổng thể này, thì có 16 mẫu sẽ được chọn ($N^2 = 4^2 = 16$). Những mẫu cụ thể với các đơn vị mẫu cấu thành được thể hiện trong Bảng 6.1 như sau

Bảng 6.1

Các mẫu	18	20	22	24
18	18;18	18;20	18;22	18;24
20	20;18	20;20	20;22	20;24
22	22;18	22;20	22;22	22;24
24	24;18	24;20	24;22	24;24

Chú ý rằng vì chúng ta lấy mẫu có hoàn lại nên trong một mẫu có thể có 2 người 18 tuổi, mẫu này được hiểu là kết quả của tiến trình lấy mẫu ngẫu nhiên như sau, lần đầu lấy được một người 18 tuổi từ tổng thể, sau đó trả người này lại tổng thể rồi lại tiến hành lấy ngẫu nhiên và lại tình cờ lấy đúng người này. Cũng theo cách lấy mẫu này thì một mẫu mà người đầu tiên 18 và người thứ hai 20 tuổi là một kết cục hoàn toàn khác một mẫu ngẫu nhiên mà người đầu tiên 20 và người thứ hai 18 tuổi.

Các trung bình mẫu tính được thể hiện trong bảng dưới, nếu ta tính kỳ vọng của các trung bình mẫu, kí hiệu là $\mu_{\bar{x}}$, theo đúng công thức tính kỳ vọng đã nghiên cứu trong chương trước, thì giá trị này sẽ bằng đúng trung bình tổng thể μ (tức là bằng 21).

Bảng 6.2

Các trị trung bình	18	20	22	24
18	(18+18)/2=18	(18+20)/2=19	(18+22)/2=20	(18+24)/2=21
20	(20+18)/2=19	(20+20)/2=20	(20+22)/2=21	(20+24)/2=22
22	(22+18)/2=20	(22+20)/2=21	(22+22)/2=22	(22+24)/2=23
24	(24+18)/2=21	(24+20)/2=22	(24+22)/2=23	(24+24)/2=24

Vì ta có dữ liệu gốc nên để nhanh chóng ta không cần lập bảng mô tả quy luật phân phối của trung bình mẫu rồi từ bảng đó tính kỳ vọng mà dùng luôn các trị trung bình mẫu để tính kỳ vọng của các trung bình mẫu theo công thức tính trung bình quen thuộc nhất thì thấy nó bằng đúng trung bình tổng thể

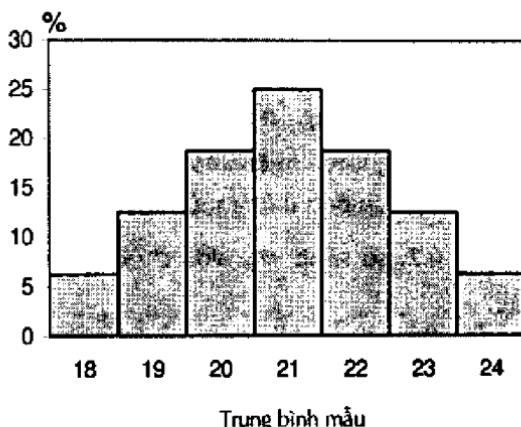
$$\mu_{\bar{x}} = \frac{(18+19+20+\dots+22+23+24)}{16} = \frac{336}{16} = 21 = \mu$$

Vì thế trung bình mẫu được xem như ước lượng không chêch của trung bình tổng thể, thế nên mặc dù chúng ta có thể không biết trị trung bình tính được của một mẫu cụ thể mà ta chọn được gần tới đâu trị trung bình thực của tổng thể nhưng ít nhất chúng ta có thể chắc rằng giá trị trung bình tính được từ tất cả các trị trung bình mẫu của tất cả các mẫu ngẫu nhiên có thể chọn từ tổng thể sẽ bằng đúng trị trung bình tổng thể.

6.1.2 Sai số chuẩn của trung bình mẫu

Có sự biến thiên của 16 trị trung bình mẫu về tuổi do sự thay đổi ngẫu nhiên những người được chọn vào mẫu theo kiểu chọn có hoàn lại, điều này được minh họa trong Hình 6.1, ta mô tả lại quy luật phân phối của trị trung bình mẫu dưới dạng đồ thị thì hình dáng của đồ thị rất giống hình chuông cân đối của phân phối bình thường, tuy đồ thị được vẽ đơn giản chỉ theo tần suất ở trục tung. Độ rộng của chân chuông phản ánh mức độ biến thiên.

Hình 6.1 phân phối của trung bình mẫu



Trong mẫu nhỏ, như ví dụ của chúng ta, mặc dù có nhiều biến thiên trong giá trị của các trung bình mẫu lấy được phụ thuộc vào tuổi của những người được chọn vào mẫu, nhưng thực tế là trung bình mẫu ít biến động hơn dữ liệu của các đơn vị tổng thể. Điều này nảy sinh từ thực tế rằng trị trung bình của các trung bình mẫu được tính từ tất cả các trung bình mẫu nên nó phải ít dao động hơn, vì các bản thân các trung bình mẫu đã ổn định hơn các quan sát riêng biệt. Tổng thể thì bao gồm rất nhiều những giá trị riêng biệt có khoảng biến động rất rộng, thậm chí có những giá trị ngoại lệ, tuy nhiên ngay khi những giá trị ngoại lệ này tình cờ được chọn vào một mẫu bất kỳ thì mặc dù nó có gây ảnh hưởng đến trị trung bình mẫu cụ thể tính được do sự bù trừ qua lại nhưng sự ảnh hưởng này đã bị làm yếu đi chính do sự bù trừ với các giá trị khác khi tính trung bình, cỡ mẫu càng tăng lên sự ảnh hưởng càng giảm đi vì nó phải bù trừ cho càng nhiều giá trị. Từ ý tưởng này ta sẽ nghiên cứu về mức độ biến động của các trung bình mẫu qua đại lượng Sai số chuẩn của trung bình (hay là độ lệch chuẩn của trung bình)

Trong nội dung phân phối mẫu đại lượng đo lường sự biến động của trung bình mẫu được gọi tên là sai số chuẩn của trung bình, ký hiệu $\sigma_{\bar{x}}$. Khi lấy mẫu có hoàn lại hoặc không hoàn lại từ một tổng thể rất lớn hoặc xem như vô hạn, sai số chuẩn của trung bình được xác định bằng công thức sau:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Như vậy khi cỡ mẫu tăng lên sai số chuẩn của trung bình giảm theo một lượng bằng căn bậc hai của cỡ mẫu, điều này cũng phù hợp với lý luận một cách thực nghiệm ở trên là khi cỡ mẫu càng tăng lên sự ảnh hưởng của những giá trị ngoại lệ càng giảm đi vì nó phải bù trừ cho càng nhiều giá trị khiến trung bình mẫu càng ổn định.

Vận dụng lại ví dụ về tổng thể tuổi của 4 người ta tính độ lệch chuẩn tổng thể σ , rồi tính $\sigma_{\bar{x}}$ theo cách tính thông thường dựa trên dữ liệu về các \bar{X} có được trong Bảng 6.2, so sánh $\sigma_{\bar{x}}$ với σ/\sqrt{n} xem thử có chúng có bằng nhau không.

$$\text{Tính } \sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} = \sqrt{\frac{(18-21)^2 + \dots + (24-21)^2}{4}} = \sqrt{5} = 2,236$$

Tính

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum(\bar{X}_i - \mu_{\bar{x}})^2}{N}} = \sqrt{\frac{(18-21)^2 + (19-21)^2 + \dots + (23-21)^2 + (24-21)^2}{16}} = \sqrt{\frac{40}{16}} = 1,58$$

$$\text{So sánh } \sigma/\sqrt{n} = 2,236/\sqrt{2} = 1,58 = \sigma_{\bar{x}}$$

Trên thực tế hầu như mọi cuộc nghiên cứu đều là lấy mẫu cỡ n không hoàn lại từ một tổng thể hữu hạn N, trong trường hợp lấy mẫu không hoàn lại nếu cỡ mẫu n được lấy so với cỡ của tổng thể không quá 5% ($n/N < 0,05$) thì ta vẫn dùng công thức trên, nhưng nếu cỡ mẫu n được lấy lớn hơn 5% so với cỡ của tổng thể thì ta phải dùng yếu tố hiệu chỉnh tổng thể hữu hạn FPC (Finite population correction) nhân thêm vào đại lượng sai số chuẩn của trung bình

$$FPC = \sqrt{\frac{N-n}{N-1}}$$

$$\text{thì Sai số chuẩn của trung bình } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (\text{khi } \frac{n}{N} \times 100 > 5\%)$$

Vì $n > 1$ nên tử số của FPC luôn bé hơn mẫu số nên FPC luôn < 1 , như vậy yếu tố hiệu chỉnh này sẽ làm sai số chuẩn của trung bình bé đi sau khi được hiệu chỉnh.

Những nội dung trên đã giới thiệu cho người đọc ý tưởng về phân phối mẫu và cách tính Sai số chuẩn của trung bình, Hình 6.1 cũng cho thấy các trung bình mẫu có một hình dáng phân phối cụ thể, vậy các trung bình mẫu \bar{X} có phân phối như thế nào.

6.1.3 Chọn mẫu từ một tổng thể có phân phối Bình thường

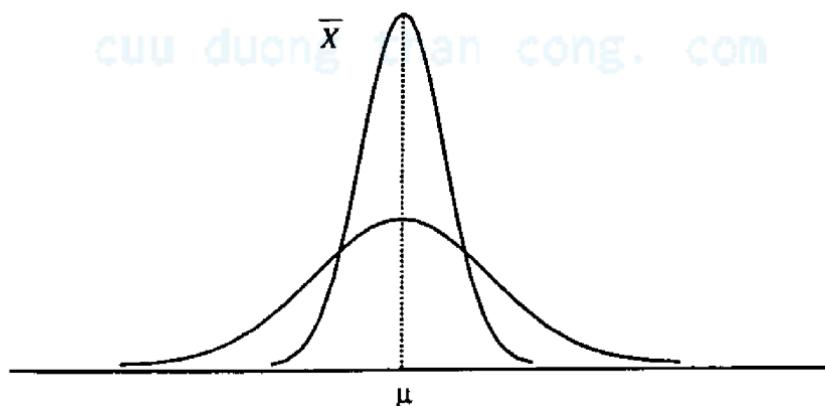
Một hàm số tuyến tính bất kỳ của các biến số bình thường cũng có phân phối bình thường, vậy ta có thể kết luận rằng nếu mẫu được lấy từ một tổng thể có phân phối bình thường với trung bình bằng μ và độ lệch chuẩn bằng σ thì cho dù cỡ mẫu n bằng bao nhiêu, phân phối của trung bình mẫu cũng sẽ có dạng phân phối bình thường với trung bình và độ lệch chuẩn (tức sai số của trung bình) được xác định như sau

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Chú ý rằng khi cỡ mẫu tăng lên, phân phối của trung bình mẫu vẫn là phân phối bình thường với trung bình $\mu_{\bar{X}} = \mu$. Tuy nhiên khi cỡ mẫu gia tăng sai số chuẩn của trung bình giảm đi, điều này có nghĩa là sẽ có một khả năng lớn hơn các trung bình mẫu lấy được có giá trị gần đúng với trung bình thực của tổng thể, lý giải này có thể được xem xét qua hình dưới.

Hình 6.2



Với cùng một trị trung bình, hai phân phối bình thường này có độ rỗng chân khác nhau tùy thuộc vào độ lệch chuẩn, với cỡ mẫu lớn hơn thì $\sigma_{\bar{X}}$ bé hơn nên ta có hình chuông nhọn hơn, ngược lại với cỡ mẫu nhỏ hơn ta có hình chuông tù hơn. Điều này có nghĩa là nếu từ một tổng thể nhất định có trung bình là μ bạn lấy mẫu có cỡ n càng lớn thì khả năng trung bình mẫu bạn lấy được \bar{X} gần bằng trị trung bình thực của tổng thể μ là rất lớn, ngược lại khi bạn lấy mẫu nhỏ, độ lệch chuẩn lớn khiến chân

chuông phình to theo các giá trị ngẫu nhiên mà biến số \bar{X} có thể nhận, có rất nhiều khả năng giá trị trung bình của mẫu lấy được nằm rất xa tâm của nó, tức là xa trung bình thực của tổng thể.

\bar{X} có phân phối bình thường nên nếu ta áp dụng công thức chuẩn hóa cho trung bình mẫu chúng ta cũng được biến số ngẫu nhiên chuẩn hóa Z có phân phối bình thường chuẩn hóa, hiển nhiên sự phân phối của biến số Z có trung bình bằng 0 và phương sai bằng 1, chúng ta cũng có thể dùng Bảng tra số 1 để tính các xác suất cho trung bình mẫu

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \text{ với } \sigma_{\bar{X}} = \frac{\sigma_x}{\sqrt{n}}$$

Ví dụ cho một phân phối bình thường với trung bình $\mu = 50$ và phương sai $\sigma^2 = 100$. Hãy tìm xác suất để trị trung bình \bar{X} của một mẫu có cỡ $n = 25$ được lấy từ tổng thể sẽ khác biệt trung bình của tổng thể ít hơn 4 đơn vị.

Từ yêu cầu trên chúng ta viết ra xác suất cần tính : $P(-4 < \bar{X} - \mu < 4)$.

Chia các vế trong dấu ngoặc cho σ/\sqrt{n} ta được :

$$P(-4/\sigma/\sqrt{n} < (\bar{X} - \mu)/\sigma/\sqrt{n} < 4/\sigma/\sqrt{n})$$

Thể số và các ký tự tương đương ta được :

$$P\left(\frac{-4}{10/\sqrt{25}} < \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma/\sqrt{n}} < \frac{4}{10/\sqrt{25}}\right)$$

Biến đổi tiếp tục ta có xác suất cần tìm là $P(-2 < Z < 2)$

Áp dụng phương pháp tính toán chúng ta đã quen thuộc từ chương trước, kết quả tìm được là $P(-2 < Z < 2) = 0,9545$

Như vậy có một khả năng tới 95,45% là trị trung bình \bar{X} của một mẫu có cỡ $n = 25$ được lấy từ tổng thể sẽ khác biệt trung bình của tổng thể ít hơn 4 đơn vị.

Vẫn tiếp tục với ví dụ trên, nếu giờ đây chúng ta phải tìm 2 giá trị cách đều số trung bình sao cho 90% các giá trị trung bình của các mẫu cỡ $n = 100$ lọt vào trong phạm vi hai trị số trên thì hai giá trị đó là bao nhiêu?

Vận dụng quy trình logic ở ví dụ trên thì lúc này bài toán của ta có dạng

$$P(-Z_a \leq Z \leq Z_a) = 0,90$$

Suy luận từ bảng phân phối bình thường chuẩn hóa (Bảng tra số 1), ta xác định được giá trị của $Z_a = 1,64$ và $-Z_a = -1,64$

Như vậy lúc này:

$$\begin{aligned}
 &= P\left(-1,64 \leq \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma / \sqrt{n}} \leq 1,64\right) \\
 &= P\left(-1,64 \leq \frac{(\bar{X} - 50)}{10 / \sqrt{100}} \leq 1,64\right) \\
 &= P(-1,64 + 50 \leq \bar{X} \leq 1,64 + 50) \\
 &= P(48,36 \leq \bar{X} \leq 51,64)
 \end{aligned}$$

Như vậy hai giá trị cần tìm là 48,36 và 51,64

6.1.4 Chọn mẫu từ một tổng thể không có phân phối bình thường

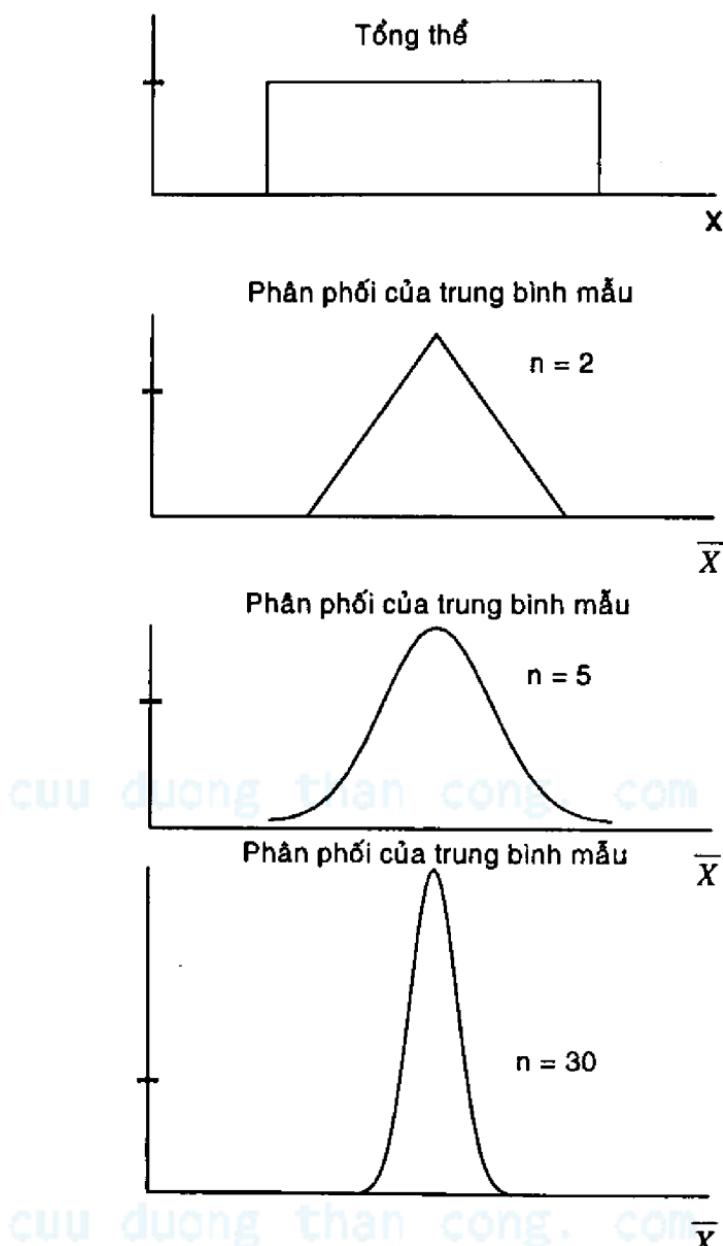
Trong thực tế có khi ta không biết gì về phân phối của tổng thể hoặc tổng thể không có phân phối Bình thường, trong những trường hợp đó định lý giới hạn trung tâm giúp ta giải quyết vấn đề xem xét phân phối của trung bình mẫu

Định lý giới hạn trung tâm (Central limit theorem) phát biểu rằng khi cỡ mẫu n đủ lớn thì phân phối của trung bình mẫu \bar{X} sẽ xấp xỉ phân phối bình thường bất chấp tổng thể có phân phối gì

Cỡ mẫu bao nhiêu là đủ lớn? Có nhiều nghiên cứu thống kê đã đề cập đến vấn đề này. Các nhà thống kê đã rút ra được một qui tắc chung như sau: cho dù tổng thể có phân phối gì thì khi cỡ mẫu không dưới 30, phân phối của trung bình mẫu sẽ xấp xỉ phân phối bình thường. Tuy nhiên chúng ta vẫn áp dụng được Định lý giới hạn trung tâm cho các tình huống có cỡ mẫu nhỏ hơn (thậm chí với n bằng 2 hay 3), nếu ta biết chắc hình dáng của phân phối tổng thể gần như cân đối. Đối với các tình huống đặc biệt hơn, nếu hình dáng của phân phối tổng thể lệch nhiều hoặc có hơn 1 mode thì cỡ mẫu cần phải lớn hơn 30 mới bảo đảm được phân phối bình thường của trung bình mẫu.

Hình 6.3 mô tả phân phối của trung bình mẫu lấy từ một tổng thể có phân phối cân đối. Khi cỡ mẫu $n = 2$, “hiệu ứng giới hạn trung tâm” bắt đầu xuất hiện thể hiện qua hình dạng tam giác của phân phối, khi cỡ mẫu $n = 5$ phân phối của trung bình mẫu có hình chuông và xấp xỉ bình thường. Khi $n = 30$, hình dáng của phân phối rất giống phân phối bình thường. Nói chung, cỡ mẫu càng lớn hình dáng của phân phối của trung bình càng giống phân phối bình thường.

Hình 6.3



Tóm lại, từ định lý giới hạn trung tâm người ta rút được 3 kết luận

1. Nếu tổng thể có phân phối Bình thường thì phân phối của trung bình mẫu \bar{X} cũng là phân phối bình thường cho dù cỡ mẫu bằng bao nhiêu
2. Với kích thước mẫu khá lớn ($n \geq 30$) thì phân phối mẫu sẽ xấp xỉ phân phối bình thường bất chấp phân phối của tổng thể

3. Nếu hình dáng của phân phối tổng thể khá đối xứng, phân phối mẫu sẽ xấp xỉ phân phối bình thường nếu cỡ mẫu $n \geq 15$

Điểm quan trọng của Định lý giới hạn trung tâm là nó giúp chúng ta biết được hình dáng của phân phối trung bình mẫu để từ đó có các suy diễn về trung bình tổng thể mặc dù có thể chúng ta không biết hình dáng phân phối của tổng thể, miễn là đáp ứng được yêu cầu cỡ mẫu phải đủ lớn.

Ví dụ: tại một cửa hàng, các hóa đơn được lưu trữ cho thấy rằng dữ liệu về doanh số của cửa hàng có dạng phân phối lệch phải với trung bình tổng thể bằng 12,5 triệu đồng/khách hàng và độ lệch chuẩn bằng 5,5 triệu đồng/khách hàng. Người quản lý của cửa hàng đã chọn một mẫu ngẫu nhiên 100 hóa đơn bán hàng, người này muốn biết xác suất để trung bình của mẫu lấy được sẽ nằm trong phạm vi từ 12,25 đến 13 triệu đồng là bao nhiêu.

Định lý giới hạn trung tâm áp dụng ở đây cho thấy vì cỡ mẫu đủ lớn ($n=100$) nên có thể khẳng định phân phối của trung bình mẫu sẽ xấp xỉ phân phối bình thường cho dù tổng thể có phân phối lệch phải. Như vậy

$$\mu_{\bar{x}} = 12,5$$

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} = 5,5 / \sqrt{100} = 0,55$$

Xác suất cần tìm được xác định là $P(12,25 \leq \bar{X} \leq 13) = ?$

Dùng phương pháp chuẩn hóa dữ liệu để tìm xác suất cần biết, ta viết lại

$$P\left(\frac{12,25 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \leq \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \leq \frac{13 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$

Thay thế số liệu ta được

$$P\left(\frac{12,25 - 12,5}{0,55} \leq Z \leq \frac{13 - 12,5}{0,55}\right)$$

$$P(-0,45 \leq Z \leq 0,91) = 0,1736 + 0,3186 = 0,4922$$

Như vậy có một khả năng khoảng gần 50% là trung bình của mẫu lấy được sẽ nằm trong phạm vi từ 12,25 đến 13 triệu đồng.

6.2 PHÂN PHỐI CỦA TỶ LỆ MẪU

Trong nhiều tình huống nghiên cứu, mục tiêu của việc lấy mẫu là để ước lượng về tỷ lệ tổng thể, ví dụ một giám đốc sản xuất có thể muốn xác định tỷ lệ của những sản phẩm bị lỗi, một công ty nghiên cứu thị trường muốn ước lượng tỷ lệ khách hàng trả lời rằng hài lòng đối với dịch vụ do một công ty du lịch cung cấp, một nhà quản lý cửa hàng muốn biết tỷ lệ

khách hàng nữ trong số các khách hàng đến mua hàng ... Trong tất cả các tình huống trên người nghiên cứu sẽ phải chọn mẫu, tính toán tỷ lệ mẫu họ quan tâm và ra quyết định căn cứ trên kết quả của mẫu.

Lúc này giống như người nghiên cứu đang làm việc với biến phân loại ở đó mỗi đơn vị hay mỗi quan sát trong tổng thể được phân loại theo kiểu có hoặc không có một đặc tính cụ thể nào đó; ví dụ là nam hoặc không là nam, hài lòng hay không hài lòng với dịch vụ... Hai kết cục có thể xảy ra này được quy ước là 1 và 0 tương ứng đại diện cho tình huống có và không có đặc tính đó. Nếu chỉ có một mẫu ngẫu nhiên n đối tượng được chọn, giá trị trung bình mẫu của biến phân loại này được tính bằng cách cộng dồn tất cả các giá trị của các quan sát lại (tức là cộng tất cả các giá trị 1 và 0 có trong biến trên từng quan sát) rồi đem chia cho n (tức cỡ mẫu). Ví dụ một mẫu có 5 người được hỏi, ba người cho biết hài lòng với dịch vụ và hai người nói không, lúc này biến phân loại của chúng ta có ba giá trị 1 và hai giá trị 0, cộng ba giá trị 1 và hai giá trị 0 lại rồi chia cho 5 ta có giá trị trung bình = 0,6 mà thực ra cũng chính là tỷ lệ người trong mẫu cho biết có hài lòng với dịch vụ. Rõ ràng về mặt phát biểu khi ta bảo tỷ lệ là 0,6 thì sự mô tả có vẻ đời thực hơn bảo 0,6 là trung bình của tất cả các giá trị 0 và 1 của biến phân loại. Bởi vậy khi làm việc với biến phân loại, đại lượng thống kê ta muốn biết là tỷ lệ trong mẫu có thuộc tính quan tâm. Đại lượng thống kê này được gọi là tỷ lệ mẫu và ký hiệu là p_s , tỷ lệ mẫu là ước lượng của tỷ lệ tổng thể. Tham số tổng thể này được kí hiệu là p .

$$\text{Công thức tính } p_s = \frac{X}{n} = \frac{\text{số quan sát có thuộc tính quan tâm}}{\text{cỡ mẫu}}$$

Tỷ lệ mẫu p_s có một đặc điểm là nhận giá trị trong khoảng từ 0 đến 1. Cũng như trung bình mẫu là ước lượng không chênh của trung bình tổng thể thì tỷ lệ mẫu là ước lượng không chênh của tỷ lệ tổng thể.

6.2.1 Khảo sát phân phối của tỷ lệ mẫu

Trong Chương 5 khi đề cập đến phân phối nhị thức ta biết nó có các đặc trưng

$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$

Chúng ta biết quy tắc chung là khi cỡ mẫu đủ lớn (sao cho np và $n(1-p)$ lớn hơn hoặc bằng 5) chúng ta có thể dùng phân phối bình thường với trung bình bằng np và độ lệch chuẩn bằng $\sqrt{np(1-p)}$ để tính xấp xỉ cho

phân phối Nhị thức. Với X là số lượt thành trong n lần thực hiện phép thử, nếu phân phối Nhị thức được coi như xấp xỉ phân phối Bình thường, thì ta có thể vận dụng đại lượng ngẫu nhiên chuẩn hóa Z có công thức

$$Z = \frac{X - \mu}{\sigma}$$

Thay các giá trị trung bình và độ lệch chuẩn của phân phối xấp xỉ Bình thường này vào công thức của Z ta có

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

Đem chia cả tử và mẫu của lượng bên phải dấu bằng cho n , đại lượng Z bên trái không thay đổi

$$Z = \frac{\frac{X - np}{\sqrt{np(1-p)}}}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{X - np}{\sqrt{np(1-p)}}}{\sqrt{\frac{p(1-p)}{n}}}$$

X là số lượt thành trong n lần thực hiện phép thử nên giá trị X cũng chính là số quan sát có thuộc tính quan tâm trong một mẫu cỡ n , ta thay $p_s = X/n$ vào và được

$$Z = \frac{\frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}}}{\sqrt{\frac{p(1-p)}{n}}}$$

Từ các thành phần trong công thức của đại lượng chuẩn hóa ở trên ta suy ra được vai trò của chúng trong phân phối của p_s , tức là phân phối của tỷ lệ mẫu:

$p = \mu_{p_s} =$ Trung bình của phân phối

$$\sqrt{\frac{p(1-p)}{n}} = \sigma_{p_s} =$$
 Độ lệch chuẩn của phân phối

Như vậy phân phối bình thường có thể được sử dụng để khảo sát phân phối của tỷ lệ mẫu với trị trung bình và độ lệch chuẩn của phân phối của tỷ lệ mẫu có công thức như sau

$$\mu_{p_s} = p$$

$$\sigma_{p_s} = \sqrt{\frac{p(1-p)}{n}}$$

Trong hầu hết trường hợp cần nghiên cứu để suy diễn về tỷ lệ tổng thể, cỡ mẫu nghiên cứu được lấy phải đủ lớn để đạt được điều kiện sử dụng phân phối bình thường để khảo sát phân phối của tỷ lệ mẫu ($n \geq 100$). Khi vận dụng phân phối bình thường, ta có thể vận dụng biến đổi chuẩn hóa Z để tìm các xác suất mong muốn liên quan đến tỷ lệ mẫu. Công thức

$$Z = \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Trong đó p là tỷ lệ tổng thể, p_s là tỷ lệ mẫu, n là cỡ mẫu

Ví dụ: H là một công ty chuyên sản xuất các món đồ trang trí giáng sinh, sản phẩm của họ được cung cấp cho các nhà bán lẻ trên toàn quốc. Qua quan sát, ban giám đốc của H đã tổng kết được là khoảng 15% các sản phẩm (ngay cả khi được đóng gói rất kỹ) bị hỏng trong quá trình chuyên chở trước khi đến được tay người bán lẻ; và dường như không có một dạng cụ thể nào của các kiểu hư hỏng, mỗi món đồ hư hỏng theo một kiểu hoàn toàn độc lập với nhau.

Có một nhà bán lẻ phản ánh với công ty rằng trong số 500 món đồ trang trí người đó đặt mua trong chuyến hàng vừa rồi có đến 90 món bị hư hỏng (điều này có nghĩa là tỷ lệ hỏng của mẫu này là $90/500=0,18$). Giả sử rằng tỷ lệ hư hỏng của tổng thể các món đồ được vận chuyển được xác định chung là 15%, vậy ban giám đốc làm sao để xác định được khả năng một mẫu gồm 500 đơn vị có bị hỏng trên 18% số đơn vị là bao nhiêu?

Để trả lời câu hỏi này trước tiên ta phải xác định một vài điều kiện.

Bởi vì

$$np = 500 \times (0,15) = 75 \geq 5$$

$$\text{và } n(1-p) = 500 \times (1-0,15) = 425 \geq 5$$

nên chúng ta có thể dùng phân phối bình thường cho phân phối của tỷ lệ mẫu, ta áp dụng công thức tính trung bình và độ lệch chuẩn của phân phối này như sau:

$$\mu_{p_s} = 0,15$$

$$\sigma_{p_s} = \sqrt{\frac{0,15(1-0,15)}{500}} = 0,016$$

Sau đó áp dụng công thức chuẩn hóa dữ liệu để đưa tỷ lệ mẫu $p_s = 0,18$ về dạng chuẩn hóa

$$Z = \frac{p_s - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{0,18 - 0,15}{0,016} = 1,88$$

Kết quả này có nghĩa là tỷ lệ hứ hỏng 18% mà nhà bán lẻ phản ánh lệch về phía trên khoảng 1,88 độ lệch chuẩn so với tỷ lệ hứ hỏng trung bình, công thức sau cho biết xác suất để điều này xảy ra là bao nhiêu

$$P(p_s \geq 0,18) = P(z \geq 1,88) = 0,5 - 0,4699 = 0,0301$$

Vì đây là xác suất khá thấp nên ban quản trị của H nên kiểm tra phải chăng có điều gì bất thường trong quá trình đóng gói và vận chuyển hàng đó.

6.2.2 Điều chỉnh sai số chuẩn của tỷ lệ mẫu

Cũng như trung bình mẫu, khi lấy mẫu không hoàn lại từ một tổng thể hữu hạn, sai số chuẩn của phân phối của tỷ lệ mẫu cần được điều chỉnh nếu cỡ mẫu n được lấy lớn hơn 5% so với cỡ của tổng thể bằng nhân tố hiệu chỉnh tổng thể hữu hạn có công thức như sau

$$FPC = \sqrt{\frac{N-n}{N-1}}$$

Lúc này sai số chuẩn của tỷ lệ mẫu cho một tổng thể hữu hạn được tính lại

$$\sigma_{p_s} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

Xem xét công thức tính FPC ở cả nội dung phân phối trung bình mẫu và phân phối tỷ lệ mẫu này ta sẽ nhận thấy rằng vì n luôn lớn hơn 1 nên mẫu số luôn lớn hơn tử số, kết quả là FPC luôn bé hơn 1, vì vậy khi được nhân vào với sai số chuẩn của tham số mẫu thì sai số chuẩn sẽ trở nên bé hơn sau khi được hiệu chỉnh, vì thế mà một sự ước lượng sẽ trở nên chính xác hơn khi FPC được sử dụng.

Ví dụ Giám đốc chi nhánh địa phương một ngân hàng xác định là khoảng 40% người gởi tiền tại ngân hàng có nhiều hơn một tài khoản tại đây. Nếu bạn chọn ngẫu nhiên một mẫu 200 người, xác suất để tỷ lệ tổng thể của những người gởi có nhiều tài khoản không quá 30% là bao nhiêu?

Vì $np = 200 \times 0,4 = 80 \geq 5$ và $n(1-p) = 200 \times 0,6 = 120 \geq 5$, nên phân phối của tỷ lệ mẫu xấp xỉ phân phối bình thường, sử dụng công thức tính sai số ta có

$$\sigma_{p_s} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0,4(1-0,4)}{200}} = \sqrt{\frac{0,24}{200}} = 0,0346$$

$$\text{Dùng công thức chuẩn hóa ta tính } Z = \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0,3 - 0,4}{0,0346} = -2,89$$

Theo bảng tích phân Laplace, diện tích dưới đường cong chuẩn hóa cho $P(Z \leq -2,89) = 0,5 - 0,4981 = 0,0019$. như vậy khả năng lấy được một mẫu có tỷ lệ không quá 30% là rất khó xảy ra.

Nếu thông tin khác cho ta biết tổng thể có 2.000 người, lúc này tỷ lệ n/N lớn hơn 0,05 nên ta sẽ phải sử dụng thêm FPC cho tình huống này, lúc đó sai số chuẩn của tỷ lệ mẫu cho một tổng thể hữu hạn được tính lại

$$\sigma_{p_s} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0,4(1-0,4)}{200}} \sqrt{\frac{2000-200}{2000-1}} = \sqrt{\frac{0,24}{200}} \sqrt{\frac{1800}{1999}} = 0,0329$$

như vậy FPC đã điều chỉnh làm cho sai số chuẩn của tỷ lệ mẫu giảm đi.

Khi đó $Z = \frac{0,3 - 0,4}{0,0329} = -3,04$. Lúc đó diện tích dưới đường cong chuẩn hóa cho $P(Z \leq -3,04) = 0,5 - 0,4988 = 0,0012$.

cuu duong thanh cong. com

CHƯƠNG 7

ƯỚC LƯỢNG CÁC THAM SỐ TỔNG THỂ

7.1 ƯỚC LƯỢNG TRUNG BÌNH TỔNG THỂ

Thống kê suy diễn là quá trình sử dụng các thông tin trên mẫu để rút ra kết luận về các đặc điểm của tổng thể, trong nội dung chương này chúng ta nghiên cứu một công cụ của thống kê suy diễn là phương pháp ước lượng các tham số tổng thể, cụ thể là ước lượng trung bình tổng thể và tỷ lệ tổng thể từ các tham số mẫu.

Có hai loại ước lượng là ước lượng điểm và ước lượng khoảng. Ước lượng điểm là dùng một tham số thống kê mẫu đơn lẻ để ước lượng về giá trị thực của tham số tổng thể. Ví dụ trung bình mẫu \bar{x} là ước lượng của trung bình tổng thể μ và phương sai mẫu hiệu chỉnh s^2 là ước lượng của phương sai tổng thể σ^2 . Ở nội dung chương 6 ta đã biết trung bình mẫu là ước lượng không chêch của trung bình tổng thể, và mặc dù trong thực tế mỗi khi làm việc ta chỉ có duy nhất một mẫu nhưng ta biết rằng trung bình của tất cả các mẫu lấy được từ một tổng thể sẽ bằng đúng trung bình tổng thể. Tương tự như vậy, ở chương 4 ta cũng đã tìm hiểu là trong công thức tính phương sai mẫu (hiệu chỉnh) s^2 , mẫu số của công thức phải là lượng $n-1$ chứ không phải n để phương sai mẫu (hiệu chỉnh) tính được là ước lượng không chêch của phương sai tổng thể, nói cách khác, trung bình của tất cả các s^2 của các mẫu lấy được sẽ bằng đúng σ^2 .

Một tham số mẫu như trung bình mẫu thay đổi từ mẫu này sang mẫu khác tùy theo giá trị của các quan sát được chọn vào mẫu, do đó nếu dùng một giá trị trung bình mẫu của một mẫu cụ thể để ước lượng điểm về trung bình tổng thể sẽ kém tin cậy hơn so với khi chúng ta vận dụng hiểu biết về quy luật phân phối của trung bình mẫu vào quá trình ước lượng trung bình tổng thể qua phương pháp ước lượng khoảng. Khoảng ước lượng được xác định bằng phương pháp ước lượng khoảng để ước lượng giá trị thật của trung bình tổng thể sẽ có một độ tin cậy xác định được tính bằng %.

Trong Chương 6, Định lý giới hạn trung tâm và các hiểu biết về phân phối tổng thể được vận dụng để tính toán tỷ lệ các trung bình mẫu sẽ rơi vào một khoảng giá trị nào đó quanh trung bình tổng thể. Ngược lại, trong ước lượng khoảng, các kết quả từ một mẫu đơn lẻ được vận dụng để rút ra kết luận về thông tin tổng thể với thực tế trung bình tổng thể là một con số ta không biết. Cụ thể là: cho tham số tổng thể μ , ta phải tìm hai giá trị L và

U là hai hàm số của trị \bar{x} sao cho với một xác suất tin cậy là $C < 1$ được ấn định trước khi lấy mẫu, xác suất để μ rơi vào trong khoảng $(L; U)$ bằng C. Khoảng $(L; U)$ được gọi là khoảng ước lượng cho μ với xác suất tin cậy C. Như vậy khoảng ước lượng đã chỉ rõ mức độ chính xác của sự ước lượng, nên nó được ưa dùng hơn ước lượng điểm.

Điểm quan trọng cần nói lại ở đây là xác suất tin cậy C, nếu chúng ta lấy tất cả các mẫu cùng cơ mà ta có thể lấy được từ một tổng thể cho sẵn, với mỗi mẫu ta tính một khoảng ước lượng kiểu như $(L; U)$ tương ứng với xác suất C, ta sẽ gấp 100C% của các khoảng ước lượng tìm được chứa μ , chứ không phải ta hiểu rằng với một khoảng $(L; U)$ cụ thể thì xác suất nó chứa μ là C%.

Một số khái niệm cần nắm:

- Các giá trị L (Lower limit) và U (Upper limit) được gọi là giới hạn dưới và trên của khoảng ước lượng.
- 100C% được gọi là độ tin cậy (Confidence level), từ đây trở đi trong thống kê suy diễn chúng ta kí hiệu độ tin cậy dưới dạng xác suất là $(1-\alpha)$ và dưới dạng phát biểu bằng đơn vị % là $100(1-\alpha)\%$.
- Ngoài ra còn có một khái niệm rất quan trọng cho sự phát triển các nội dung ở sau là khái niệm giá trị tối hạn (Critical value).

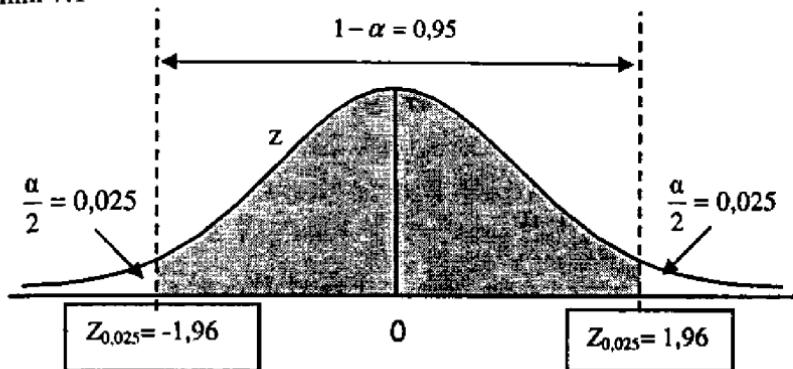
Nếu Z là một biến số bình thường chuẩn hóa thì ta có 2 giá trị tối hạn $z_{\alpha/2}$ và $-z_{\alpha/2}$ tuân theo định nghĩa như sau $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$

Như vậy theo định nghĩa này $z_{\alpha/2}$ là giá trị của biến Z trong phân phối bình thường chuẩn hóa ở phía bên phải của phân phối và $-z_{\alpha/2}$ là giá trị của biến Z ở phía bên trái của phân phối tương ứng với xác suất để mọi trị số Z bé hơn $z_{\alpha/2}$ hoặc lớn hơn $-z_{\alpha/2}$ sẽ bằng đúng $1 - \alpha$.

Minh họa cụ thể bằng hình ảnh trong ví dụ dưới: ta có $-z_{0,025}$ và $z_{0,025}$ là hai trị số trên thang z sao cho phần diện tích dưới đường cong bình thường chuẩn hóa không tô đậm trong hình minh họa bằng đúng $0,025 \times 2 = 0,05$; hay viết theo chiều khác $-z_{0,025}$ và $z_{0,025}$ là hai trị số trên thang z sao cho phần diện tích dưới đường cong bình thường chuẩn hóa được tô đậm trong hình minh họa bằng đúng $(1 - 0,025 \times 2) = 0,95$.

Vậy thì $-z_{0,025}$ và $z_{0,025}$ có giá trị bằng bao nhiêu?

Hình 7.1



Sử dụng Bảng tra số 1 cho nửa bên phải của phân phối ta thấy diện tích phần màu đậm là $(0,5 - 0,025) = 0,475$ thì tương ứng với giá trị $z_h = 1,96$. Như vậy $z_{0,025} = 1,96$ và suy ra $-z_{0,025} = -1,96$ do tính đối xứng của phân phối. Từ quy tắc này ta xác định sẵn một số tình huống về độ tin cậy $(1 - \alpha)$ hay gấp trong thống kê và các giá trị tối hạn tương ứng của nó

Bảng 7.1

$(1 - \alpha)100\%$	$ z_{\alpha/2} $
80%	1,28
85%	1,44
90%	1,645
95%	1,96
98%	2,33
99%	2,58
99.80%	3,08
99.90%	3,27

7.1.1 Ước lượng khoảng của trung bình tổng thể (khi đã biết phương sai tổng thể)

Giả sử chúng ta có một tổng thể bình thường với phương sai đã biết là σ^2 . Đây là một tình huống có tính giả định bởi lẽ đã không biết μ thì ta cũng không biết σ^2 . Tuy nhiên trong thực tế nhiều khi vấn đề khảo sát khá quen thuộc nên bằng kinh nghiệm người ta biết được giá trị gần đúng của phương sai tổng thể nên giả định này có phần nào hợp lý.

Xem xét lại phương trình $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$

Vì phương trình này dùng cho bất cứ biến số bình thường chuẩn hóa nào nên ta cũng có thể áp dụng cho giá trị chuẩn hóa z của biến ngẫu nhiên trung bình mẫu có công thức $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$, thay giá trị này của Z vào phương trình trên, ta được

$$P(-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leq z_{\alpha/2}) = 1 - \alpha$$

Bây giờ ta có thể áp dụng quy tắc đại số cho bất đẳng thức ở vế bên trái của phương trình để có được một biểu thức tương đương chỉ còn μ ở giữa. Hiển nhiên xác suất tương ứng không thay đổi vì thực ra chúng ta chỉ áp dụng phép biến đổi bên trong bất đẳng thức, quá trình như sau:

Nhân 3 vế bất đẳng thức với $\sigma_{\bar{x}}$ ta sẽ được

$$\text{trên } P(-z_{\alpha/2} \sigma_{\bar{x}} \leq \bar{x} - \mu \leq z_{\alpha/2} \sigma_{\bar{x}}) = 1 - \alpha$$

trừ đi \bar{x} trong mỗi vế để được

$$P(-\bar{x} - z_{\alpha/2} \sigma_{\bar{x}} \leq -\mu \leq -\bar{x} + z_{\alpha/2} \sigma_{\bar{x}}) = 1 - \alpha$$

Nhân cả ba vế với -1 , dấu của bất đẳng thức sẽ đổi chiều ta được kết quả mong muốn

$$P(\bar{x} - z_{\alpha/2} \sigma_{\bar{x}} \leq \mu \leq \bar{x} + z_{\alpha/2} \sigma_{\bar{x}}) = 1 - \alpha$$

Chú ý rằng đây là xác suất tin cậy chứ không phải xác suất của một tình huống nhận định nào đó về μ vì μ là một hằng số chứ không phải biến số ngẫu nhiên.

Vậy xác suất để cho μ rơi vào trong khoảng từ $(\bar{x} - z_{\alpha/2} \sigma_{\bar{x}})$ đến $(\bar{x} + z_{\alpha/2} \sigma_{\bar{x}})$ là $1 - \alpha$, đặt $(\bar{x} - z_{\alpha/2} \sigma_{\bar{x}})$ là L và $(\bar{x} + z_{\alpha/2} \sigma_{\bar{x}})$ là U thì ta có thể phát biểu cách khác là chúng ta có thể tin cậy tới $100(1-\alpha)\%$ là khoảng $(L; U)$ trên chứa μ .

Ở nội dung Chương 6 chúng ta cũng biết nếu mẫu khảo sát có cỡ n thì $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ do đó các giới hạn tin cậy thường được viết lại như sau:

$$L = \bar{x} - z_{\alpha/2} \sigma / \sqrt{n}$$

$$U = \bar{x} + z_{\alpha/2} \sigma / \sqrt{n}$$

Các giới hạn này xác định khoảng ước lượng cho trung bình tổng thể với cách trình bày như sau:

$$\bar{x} - z_{\alpha/2} \sigma / \sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2} \sigma / \sqrt{n}$$

Với các độ tin cậy hay gấp trong thống kê là độ tin cậy 95% và 99%, các khoảng ước lượng có thể được xác định như sau

- Nếu $1 - \alpha = 0,95 \rightarrow \alpha = 0,05 \rightarrow \alpha/2 = 0,025$ thì theo liệt kê trên Bảng 7.1 $|z_{\alpha/2}| = 1,96$. Vậy khoảng ước lượng cho μ với độ tin cậy 95% định bởi $\bar{x} - 1,96\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1,96\sigma/\sqrt{n}$
- Nếu $1 - \alpha = 0,99 \rightarrow \alpha = 0,01 \rightarrow \alpha/2 = 0,005$ thì theo liệt kê trên Bảng 7.1 $|z_{\alpha/2}| = 2,58$. Vậy khoảng ước lượng cho μ với độ tin cậy 99% định bởi $\bar{x} - 2,58\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 2,58\sigma/\sqrt{n}$

Ví dụ: một nhà máy sản xuất giấy theo dây chuyền tự động, giấy được sản xuất có chiều dài trung bình 29,7cm và độ lệch tiêu chuẩn của chiều dài là 0,05cm, để kiểm soát tiêu chuẩn giấy thì định kì người ta sẽ chọn mẫu gồm 100 tờ giấy để tiến hành kiểm tra xem chiều dài của các tờ giấy sản xuất còn đạt tiêu chuẩn 29,7cm hay không, nếu không cần phải kiểm tra xem có vấn đề gì xảy ra với dây chuyền sản xuất đã gây ảnh hưởng đến tiêu chuẩn của giấy. Trong lần kiểm tra gần đây nhất chiều dài tờ giấy trung bình tính được từ mẫu là 29,698cm. Hãy xác định khoảng ước lượng với độ tin cậy 95% cho chiều dài giấy trung bình của tổng thể các tờ giấy sản xuất trong giai đoạn giữa lần kiểm tra định kì này với lần kiểm tra kế trước đó.

Ta có $n = 100$; $\bar{x} = 29,698$; $\sigma = 0,05$

Độ tin cậy đặt ra của bài toán là 95% tức $1 - \alpha = 0,95 \rightarrow \alpha = 0,05 \rightarrow \alpha/2 = 0,025 \rightarrow |z_{\alpha/2}| = |z_{0,025}| = 1,96$. Vậy khoảng ước lượng cho μ với độ tin cậy 95% định bởi

$$\bar{x} - 1,96\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1,96\sigma/\sqrt{n}$$

Thay thế số liệu vào công thức trên

$$29,698 - 1,96 \times 0,05/\sqrt{100} \leq \mu \leq 29,698 + 1,96 \times 0,05/\sqrt{100}$$
$$29,6882 \leq \mu \leq 29,7078$$

Như vậy với độ tin cậy 95%, chiều dài trung bình tổng thể các tờ giấy được sản xuất được ước lượng trong khoảng từ 29,6882cm đến 29,7078 cm. Vì giá trị 29,7cm (giá trị cho biết tiến trình sản xuất vẫn ổn định) thuộc khoảng ước lượng này nên ta có thể khẳng định tiến trình sản xuất vẫn bình thường.

Nếu yêu cầu đặt ra của ban quản lý chất lượng là phải xây dựng một khoảng ước lượng với độ tin cậy đến 99% ta tiến hành lại quá trình tính toán như sau với giá trị tối hạn bằng $\pm 2,58$

$$29,698 - 2,58 \times 0,05/\sqrt{100} \leq \mu \leq 29,698 + 2,58 \times 0,05/\sqrt{100}$$

Với khoảng tin cậy này, một lần nữa chúng ta lại khẳng định tiến trình sản xuất vẫn bình thường.

So sánh hai khoảng ước lượng tìm được tương ứng với hai độ tin cậy ta nhận thấy khi xác suất tin cậy càng tăng thì khoảng ước lượng càng rộng, đây là điều chúng ta mong đợi vì khoảng ước lượng mà càng rộng thì khả năng để khoảng này chứa được trung bình tổng thể càng tăng, tuy nhiên nó chưa hẳn là điều đáng mong muốn vì khoảng ước lượng càng rộng thì độ chính xác của ước lượng càng thấp. Với một độ tin cậy đã xác định cách duy nhất để tăng độ chính xác của bài toán ước lượng là tăng cỡ mẫu vì tăng cỡ mẫu làm cho khoảng (L;U) hẹp lại, tuy nhiên do giới hạn của nhân lực, tiền bạc, thời gian mà việc tăng cỡ mẫu không phải bao giờ cũng thực hiện được.

Thông thường nhiều tình huống chúng ta không biết được độ lệch chuẩn σ của tổng thể nên trong các phép tính xấp xỉ các giới hạn của khoảng tin cậy chúng ta có thể dùng độ lệch chuẩn của mẫu s thay cho σ . Phép tính xấp xỉ này khá đúng khi cỡ mẫu của chúng ta lớn hơn hoặc bằng 30. Với tình huống cỡ mẫu nhỏ hơn 30 phép tính xấp xỉ trên không được tốt và chúng ta phải áp dụng lý thuyết về mẫu nhỏ mà chúng ta sẽ xem xét ở nội dung sau

7.1.2 Ước lượng khoảng của trung bình tổng thể (khi không biết phương sai tổng thể)

Ở trên đã nói, khi không biết độ lệch chuẩn σ của tổng thể, nếu cỡ mẫu khá lớn ($n \geq 30$) ta thay bằng s và lúc này khoảng ước lượng xấp xỉ được thiết lập lại với công thức

$$\bar{x} - z_{\alpha/2} s / \sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2} s / \sqrt{n}$$

Nhưng nếu cỡ mẫu nhỏ ($n < 30$) sự thay thế trên dẫn đến một sai số đáng kể trong phép tính xấp xỉ, vậy nên nếu vẫn muốn duy trì mức độ tin cậy 100 (1- α) % thì khoảng ước lượng phải được nới rộng bằng cách dùng phân phối t thay cho phân phối z.

7.1.2.1 Mô tả phân phối t (Phân phối Student)

Chúng ta nhắc lại rằng \bar{x} có một phân phối bình thường và khi biết σ ta chuẩn hóa \bar{x} để được $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

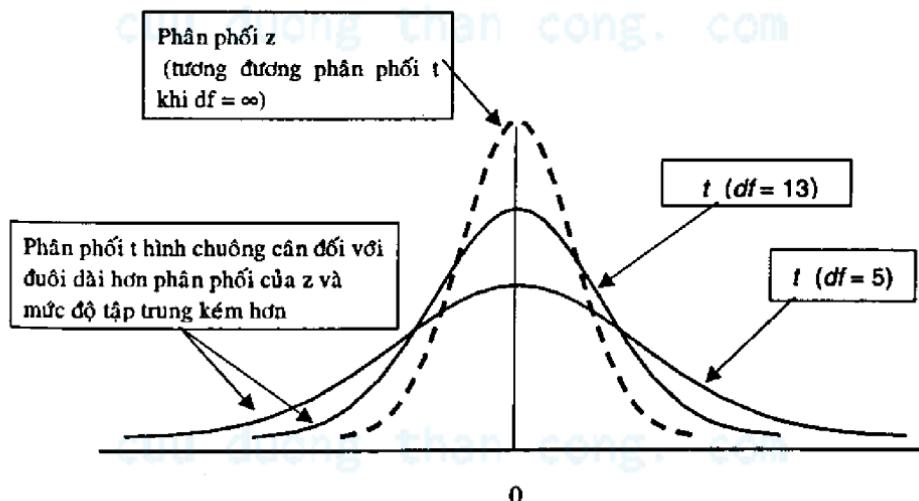
Trong trường hợp không biết σ ta thay bằng s và lúc này thay vì dùng z ta phải dùng một biến số mới định bởi công thức:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Đi nhiên biến số t là biến chuẩn hóa theo một hàm tuyến tính của \bar{x} nên nó sẽ có dạng phân phối tương tự Z tức có hình chuông cân đối, biến số t do ông William S. Gosset tìm ra năm 1908 dưới bút hiệu Student nên ngày nay người ta quen gọi phân phối của t là Phân phối t Student (Student's t distribution).

Đặc điểm của phân phối t: Phân phối t có dạng hình chuông cân đối với hai đuôi dài hơn đuôi của phân phối z và mức độ tập trung ở trung tâm kém hơn, bởi lẽ khi dùng s thay cho σ ta đã đưa vào trong bài toán ước lượng thêm một phần "bất định" nên sự trải rộng hơn này ám chỉ \bar{x} phân tán rộng hơn quanh μ thật. Ngoài ra trong khi chỉ có duy nhất một đường cong ứng với phân phối của z thì chúng ta lại có cả một họ phân phối t theo các bậc tự do, bậc tự do kí hiệu là $df = n-1$. Khi cỡ mẫu nhỏ (tức bậc tự do nhỏ) phân phối t khá khác phân phối z, nhưng khi cỡ mẫu tăng lên phân phối t dần dần tiến đến hội tụ với phân phối z và khi cỡ mẫu từ 30 trở lên hai phân phối xấp xỉ nhau với độ chính xác cao.

Hình 7.2



Phân phối t được tính toán và lập thành bảng tra (nhưng không tra xác suất như phân phối z mà lại tra các giá trị tối hạn của các tình huống hay gấp về xác suất), nhưng có kết hợp với bậc tự do, nên khi tra bảng phân phối t chúng ta phải dùng thêm bậc tự do (df). Phân phối t cũng đổi xứng

như phân phối z nên bảng tra các trị tới hạn của t chỉ cần liệt kê bên phía dương. Trong phần phụ lục nó là Bảng tra số 2

Hướng dẫn cách tra bảng phân phối t tìm giá trị tới hạn t với độ tin cậy 95% và bậc tự do là 8

Ta có $df = 8$; độ tin cậy là 95% tức $1 - \alpha = 0,95 \Rightarrow \alpha = 0,05 \Rightarrow \alpha/2 = 0,025$
 \rightarrow giá trị tới hạn $t_{\alpha/2, df} = t_{0,025, 8} = ?$

Bởi vì phân phối t đối xứng như phân phối z nên ta chỉ cần tra giá trị t phía bên phải của phân phối là ta có thể suy ra giá trị ở phía đối diện, giá trị này bằng giá trị kia về trị tuyệt đối nhưng khác dấu

Dưới đây là phần đầu của Bảng tra số 2 được chúng tôi trích ra để hướng dẫn bạn đọc cách tra, trong đó bậc tự do được tra theo cột đầu tiên và thông tin về độ tin cậy tra theo hàng đầu tiên.

Bảng 7.2

df	$t_{0.125}$	$t_{0.1}$	$t_{0.075}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	2.4142	3.0777	4.1653	6.3138	12.7062	31.8205	63.6567
2	1.6036	1.8856	2.2819	2.9200	4.3027	6.9646	9.9248
3	1.4226	1.6377	1.9243	2.3534	3.1824	4.5407	5.8409
4	1.3444	1.5332	1.7782	2.1318	2.7764	3.7469	4.6041
5	1.3009	1.4759	1.6994	2.0150	2.5706	3.3649	4.0321
6	1.2733	1.4398	1.6502	1.9432	2.4469	3.1427	3.7074
7	1.2543	1.4149	1.6166	1.8946	2.3646	2.9980	3.4995
8	1.2403	1.3968	1.5922	1.8595	2.3060	2.8965	3.3554
9	1.2297	1.3830	1.5737	1.8331	2.2622	2.8214	3.2498

$$|t_{0.025, 8}| = 2,306$$

Vậy với độ tin cậy 95%, bậc tự do là 8 thì hai giá trị t tới hạn sử dụng cho ước lượng khoảng là $-t_{0.025, 8} = -2,3$ và $t_{0.025, 8} = 2,3$

7.1.2.2 Ước lượng khoảng của trung bình tổng thể khi cỡ mẫu nhỏ

Trở lại với vấn đề đặt ra ở mục 7.1.2, khi nếu cỡ mẫu không đủ lớn thì khoảng ước lượng cho trung bình tổng thể phải được nới rộng bằng cách dùng phân phối t.

Để xây dựng khoảng ước lượng cho μ khi chưa biết σ ta cũng tiến hành cùng một cách thức đã khảo sát ở phần ước lượng khoảng mà biết σ , có điều với phân phối t ta viết

$$P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1-\alpha$$

Đối với một mẫu cỡ n lấy từ một tổng thể có phân phối bình thường, biến số $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ có phân phối t với bậc tự do $df = n-1$, vậy

$$P(-t_{\alpha/2;n-1} \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t_{\alpha/2;n-1}) = 1-\alpha$$

Và quãng tin cậy $100(1-\alpha)\%$ cho μ định bởi

$$\bar{x} - t_{\alpha/2;n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$$

Ví dụ ta đo trọng lượng của 25 bé trai 8 tuổi ở 1 ngôi làng, sau khi tính toán tập dữ liệu mẫu này ta có các số liệu sau $\bar{x} = 26,06\text{kg}$ và $s = 1,61\text{kg}$. Hãy ước lượng trọng lượng trung bình của trẻ em trai 8 tuổi ở đây với độ tin cậy 90%.

Ta có $n = 25 \rightarrow df = n-1 = 25-1 = 24$; $\bar{x} = 26,06$; $s = 1,61$

Độ tin cậy đặt ra của bài toán là 90% tức $1-\alpha = 0,9 \rightarrow \alpha = 0,1 \rightarrow \alpha/2 = 0,05$. $\rightarrow |t_{\alpha/2;n-1}| = |t_{0,05;24}| = 1,711$. Vậy khoảng ước lượng cho μ với độ tin cậy 90% định bởi

$$26,06 - 1,711 \frac{1,61}{\sqrt{25}} \leq \mu \leq 26,06 + 1,711 \frac{1,61}{\sqrt{25}}$$

$$25,5091 \leq \mu \leq 26,6109$$

Với độ tin cậy 90%, trọng lượng trung bình của trẻ em trai tại ngôi làng này được ước lượng trong khoảng từ $25,51\text{kg}$ đến $26,61\text{kg}$

7.2 ƯỚC LƯỢNG TỈ LỆ TỔNG THỂ

Trong nội dung này chúng ta tìm hiểu cách ước lượng tỷ lệ tổng thể từ thông tin về tỷ lệ mẫu ký hiệu là p_s , ở Chương 6 trong nội dung Phân phối của tỷ lệ mẫu ta đã thảo luận việc tỷ lệ mẫu p_s là ước lượng không chênh của tỷ lệ tổng thể p . Một lần nữa, bên cạnh việc có thể dùng p_s làm ước lượng điểm cho p ta có thể phát triển ước lượng khoảng cho p với cùng ý tưởng đã làm với μ .

Nhớ lại là phân phối của tỷ lệ mẫu p_s (là phân phối Nhị thức) xấp xỉ phân phối bình thường khi cỡ mẫu đủ lớn, với trung bình là $\mu_{p_s} = p$ và độ lệch chuẩn là $\sigma_{p_s} = \sqrt{\frac{p(1-p)}{n}}$; vậy biến số chuẩn hóa Z định bởi

$$Z = \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Vận dụng lại phương trình

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

Thay công thức chuẩn hóa Z vào phương trình trên

$$P(-z_{\alpha/2} \leq \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}) = 1 - \alpha$$

Sau khi biến đổi bất đẳng thức bên tay trái ta được kết quả sau cùng

$$P(p_s - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq p \leq p_s + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}) = 1 - \alpha$$

Vì p là thông tin tổng thể ta không biết nên ta thay thế bằng thông tin trên mẫu

$$P(p_s - z_{\alpha/2} \sqrt{\frac{p_s(1-p_s)}{n}} \leq p \leq p_s + z_{\alpha/2} \sqrt{\frac{p_s(1-p_s)}{n}}) = 1 - \alpha$$

Lập luận như cách đã xây dựng khoảng ước lượng cho trung bình tổng thể ta có thể xây dựng khoảng ước lượng cho p với độ tin cậy $100(1-\alpha)\%$ như sau

$$p_s - z_{\alpha/2} \sqrt{\frac{p_s(1-p_s)}{n}} \leq p \leq p_s + z_{\alpha/2} \sqrt{\frac{p_s(1-p_s)}{n}}$$

Trong đó

- p_s là tỷ lệ mẫu với công thức

$$p_s = \frac{X}{n} = \frac{\text{số quan sát có thuộc tính quan tâm}}{\text{cỡ mẫu}}$$

- p là tỷ lệ tổng thể
- $z_{\alpha/2}$ là giá trị tối hạn
- n là cỡ mẫu

Ví dụ: Người ta chọn 1 mẫu ngẫu nhiên 100 người và thấy 25% số người này thuận tay trái trong ăn uống, hãy xây dựng một khoảng ước lượng cho tỷ lệ người thuận tay trái trong tổng thể với độ tin cậy 95%

Ta có $n = 100$; $p_s = 0,25$; $1-\alpha = 95\% \rightarrow z_{\alpha/2} = 1,96$

Từ công thức

$$p_s - z_{\alpha/2} \sqrt{\frac{p_s(1-p_s)}{n}} \leq p \leq p_s + z_{\alpha/2} \sqrt{\frac{p_s(1-p_s)}{n}}$$

Vậy khoảng tin cậy này được xác định

$$0,25 - 1,96 \sqrt{\frac{0,25(1-0,25)}{100}} \leq p \leq 0,25 + 1,96 \sqrt{\frac{0,25(1-0,25)}{100}}$$

$$0,25 - 0,0849 \leq p \leq 0,25 + 0,0849$$

$$0,1651 \leq p \leq 0,3349$$

Như vậy chúng ta tin cậy 95% rằng tỷ lệ những người thuận tay trái trong khi ăn uống là từ 16,51% đến 33,49%.

Mặc dù khoảng này có hoặc không thể chứa tỷ lệ thật của tổng thể nhưng 95% của các khoảng ước lượng được xây dựng trên mẫu 100 người theo cách vừa rồi sẽ chứa giá trị thật của tổng thể.

7.3 XÁC ĐỊNH CƠ MẪU CHO BÀI TOÁN ƯỚC LƯỢNG

7.3.1 Quy tắc xác định cơ mẫu cho ước lượng trung bình tổng thể

Thông thường, trước khi lấy mẫu chúng ta thường phải xác định là cơ mẫu n cần bằng bao nhiêu để chúng ta có thể ước tính trung bình với một độ chính xác dự định trước, độ chính xác được đo bằng độ rộng của khoảng ước lượng, đó chính là lượng được cộng vào và trừ ra khỏi trung bình mẫu khi tính giới hạn dưới và trên của khoảng ước lượng, nó được kí hiệu là e và đôi khi còn được gọi là dung sai của ước lượng hoặc sai số của ước lượng. Chúng ta cũng phải xác định trước chúng ta muốn bài toán ước lượng với độ chính xác dự định trước này có độ tin cậy đến đâu, và với thông tin về độ lệch chuẩn tổng thể đã biết, ta xây dựng công thức xác định n là

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{e^2} = \left(\frac{z_{\alpha/2} \sigma}{e} \right)^2$$

Tìm ra n chúng ta sẽ biết được cơ mẫu cần thiết để xây dựng một khoảng ước lượng cho μ có độ tin cậy và độ chính xác phù hợp với dự định.

Nếu bạn không có thông tin trước về σ trong bài toán xác định cỡ mẫu của bạn thì sao, bạn có thể làm các cách sau:

- Tham khảo thông tin về σ trong những nghiên cứu tương tự có trước
- Áp dụng quy tắc 3σ để tìm, nếu phân phối dữ liệu của bạn cân đối hình chuông thì khoảng biến thiên từ giá trị lớn nhất đến giá trị nhỏ nhất R sẽ trải trong vòng 6σ ($\pm 3\sigma$ xung quanh μ) vì thế $\rightarrow \sigma = R/6$.
- Cách khác nữa là làm một cuộc điều tra thí điểm với cỡ mẫu nho nhỏ và tính toán s từ mẫu này để thay thế cho σ .

Chú ý nữa là đại lượng chuẩn hóa z sẽ được sử dụng trong công thức xác định cỡ mẫu n chứ không phải đại lượng t . Vì muốn dùng t ta phải có thông tin về bậc tự do, mà muốn tính bậc tự do thì lại phải có cỡ mẫu là cái ta đang muốn tìm, cho nên không thực hiện được. Trong hầu hết các nghiên cứu, cỡ mẫu cần thiết sẽ đủ lớn để phân phối bình thường chuẩn hóa và phân phối t Student xem như xấp xỉ nhau, và ta có thể dùng chung phân phối z .

Ví dụ trưởng phòng nhân sự một công ty muốn ước lượng số ngày nghỉ bệnh trung bình trong năm của nhân viên công ty, tìm hiểu ở các công ty tương tự thì người này biết tổng thể số ngày nghỉ ốm có phân phối bình thường với độ lệch chuẩn là 3 ngày, nếu muốn khoảng tin cậy 85% của trung bình tổng thể chênh lệch trong khoảng $\pm 0,5$ ngày so trung bình mẫu thì cần chọn mẫu gồm bao nhiêu nhân viên ?

Ta có $\sigma = 3$; $e = 0,5$; $1-\alpha = 85\% \rightarrow \alpha = 0,15 \rightarrow \alpha/2 = 0,075 \rightarrow z_{\alpha/2} = 1,44$

$$n \left(\frac{z_{\alpha/2} \sigma}{e} \right)^2 = \left(\frac{1,44 \times 3}{0,5} \right)^2 = 74,6$$

Như vậy để đạt các yêu cầu đề ra ông này cần nghiên cứu 75 nhân viên. Trong thực tế, lấy mẫu trong nghiên cứu là một quy trình không đơn giản như trên mà nó còn chịu ràng buộc của vô số vấn đề như ngân sách, thời gian, sự sẵn có của đối tượng quan sát.

7.3.2 Quy tắc xác định cỡ mẫu cho ước lượng tỷ lệ tổng thể

Công thức xác định cỡ mẫu cần thiết trong ước lượng tỷ lệ tổng thể cũng cùng ý tưởng với công thức xác định cỡ mẫu cho ước lượng trung bình tổng thể, công thức đó như sau:

$$n = \frac{z_{\alpha/2}^2 p (1-p)}{e^2}$$

Trong đó

- $z_{\alpha/2}$ là giá trị tra bảng phân phối z căn cứ trên độ tin cậy $1-\alpha$
- e là độ rộng của ước lượng
- p là tỷ lệ thành công, nó là tham số tổng thể mà bạn đang phải tìm cách ước lượng, và việc đầu tiên của quá trình ước lượng là xác định cỡ mẫu.

Sự khó khăn này được giải quyết bằng một vài cách sau:

- Sử dụng thông tin từ các nghiên cứu trước hoặc dùng kinh nghiệm để phỏng đoán
- Chọn $p = 0,5$. Khi bạn chọn $p = 0,5$ thì thành phần $p(1-p)$ sẽ lớn nhất so với các tình huống khác của p , như vậy nó sẽ làm cho tử số của công thức tìm n trên đạt cực đại để bảo đảm rằng n được ước lượng có độ lớn "an toàn" nhất. Dĩ nhiên cách làm này kéo theo sự tốn kém chi phí khi lấy số lượng mẫu cao nhất. Có khi nó còn khiến ta lấy mẫu có cỡ lớn quá mức cần thiết khi mà tỷ lệ thực của tổng thể xa giá trị 0,5 nhiều, công dụng của nó chỉ là làm hẹp độ rộng của khoảng ước lượng khiến bài toán của ta có độ chính xác cao.

Ví dụ: Phản ứng của một người trong một loại trắc nghiệm tâm lý có thể phát hiện dưới hai dạng A hoặc B. Nếu người làm trắc nghiệm muốn ước tính xác suất số người có phản ứng loại A trong tổng thể thì anh ta cần làm thí nghiệm với bao nhiêu người. Cho rằng người này bằng lòng với kết quả nếu độ rộng của khoảng ước lượng $e = 0,04$ và độ tin cậy của bài toán là 90%. Anh ta kì vọng p có giá trị khoảng 0,6.

$$1-\alpha = 0,9 \rightarrow z_{\alpha/2} = 1,64; p = 0,6; e = 0,04$$

Áp dụng công thức, ta tính được:

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{e^2} = \frac{1,64^2 \times 0,6 \times (1-0,6)}{0,04^2} = 403,44 \approx 404$$

7.3.3 Xác định cỡ mẫu trong tình huống tổng thể hữu hạn

Trong các nội dung xác định cỡ mẫu ở chương này chúng ta xem như đang làm việc với tình huống lấy mẫu không lặp lại ở một tổng thể vô hạn, nhưng trong thực tế nhiều khi ta phải làm việc với tổng thể hữu hạn. Như ở Chương 6 chúng ta đã thảo luận về việc dùng FPC khi lấy mẫu không lặp lại từ một tổng thể hữu hạn, ở đây cũng vậy, các cỡ mẫu sau khi đã được xác định theo công thức bình thường

$$n_0 = \left(\frac{z_{\alpha/2} \sigma}{e} \right)^2 \quad n_0 = \frac{z_{\alpha/2}^2 p(1-p)}{e^2}$$

Sẽ được điều chỉnh tiếp cho phù hợp bằng cách sử dụng công thức sau

$$n = \frac{n_0 N}{n_0 + (N - 1)}$$

Trong đó n_0 là cỡ mẫu được xác định theo công thức bình thường chưa xét đến việc tổng thể vô hạn hay hữu hạn. N là cỡ của tổng thể hữu hạn.

7.4 ƯỚC LƯỢNG TRÊN HAI MẪU

7.4.1 Ước lượng trung bình hai mẫu

Trong phần này chúng ta xây dựng phương pháp ước lượng sự khác biệt giữa hai trung bình của hai tổng thể dựa trên thông tin của hai mẫu. Chúng ta chia ra hai trường hợp là hai mẫu được rút từ hai tổng thể hoàn toàn độc lập và hai mẫu được rút từ hai tổng thể không độc lập. Vậy trước hết chúng ta làm rõ khái niệm hai mẫu thế nào là độc lập hay không độc lập.

Hai mẫu độc lập là hai mẫu được chọn ra từ hai tổng thể theo cách sao cho một quan sát khi được chọn vào mẫu này không làm ảnh hưởng đến xác suất một quan sát khác được chọn vào mẫu kia.

Mẫu phối hợp từng cặp là mẫu được chọn theo cách một quan sát trên mẫu này có sự tương ứng với một quan sát trên mẫu thứ hai nhằm mục đích kiểm soát những tác nhân ngoại cảnh. Mẫu này còn có tên gọi là mẫu không độc lập hay ngắn gọn là mẫu cặp.

Ví dụ nếu bạn muốn so sánh mức lương giữa nhân viên nam và nữ bạn sẽ thấy rằng mức lương trước hết chịu ảnh hưởng rất lớn của bằng cấp, trình độ ngoại ngữ, chức vụ đảm nhiệm, mức độ năng động, loại hình công việc, thâm niên... chứ không chỉ chịu chi phối của giới tính. Khi đó để hạn chế tối đa sự ảnh hưởng của các yếu tố khác ngoài giới tính ta phải chọn hai mẫu nhân viên nam nữ sao cho có sự tương đồng hoàn toàn theo từng cặp về các yếu tố có khả năng tác động đến vấn đề ta muốn so sánh là mức lương, như vậy điểm khác biệt giữa từng cặp chỉ là giới tính và ta tiến hành ghi chép lại số liệu về mức lương để so sánh nhằm lùm kiểm sự khác biệt nếu có của mức lương giữa hai giới tính. Cách lấy mẫu như vậy gọi là lấy mẫu cặp khi mỗi đối tượng trong mẫu này có một đối tượng tương ứng (đối xứng) trong mẫu kia.

Ta sẽ tiến hành khảo sát phương pháp ước lượng khác biệt hai trung bình tổng thể trong trường hợp mẫu độc lập trước rồi đến mẫu cặp

7.4.1.1 Ước lượng khác biệt hai trung bình tổng thể trong trường hợp mẫu độc lập

Chúng ta gọi

μ_1 là giá trị trung bình của tổng thể thứ nhất

μ_2 là giá trị trung bình của tổng thể thứ hai

σ_1 là độ lệch chuẩn của tổng thể thứ nhất

σ_2 là độ lệch chuẩn của tổng thể thứ hai

Sự khác biệt giữa hai trung bình tổng thể là $(\mu_1 - \mu_2)$

Để ước lượng khoảng cho chênh lệch $(\mu_1 - \mu_2)$ chúng ta chọn hai mẫu ngẫu nhiên:

Mẫu thứ nhất có cỡ là n_1 được chọn từ tổng thể thứ nhất

Mẫu thứ hai có cỡ là n_2 được chọn từ tổng thể thứ hai.

Hai mẫu này được chọn theo cách thức hoàn toàn độc lập với nhau và n_1 với n_2 không nhất thiết phải bằng nhau.

Trên hai mẫu ta tính được \bar{x}_1 với s_1 và \bar{x}_2 với s_2 lần lượt là trung bình và độ lệch chuẩn các mẫu

Lượng $(\bar{x}_1 - \bar{x}_2)$ là khác biệt giữa hai trung bình của hai mẫu. Ta thấy $(\bar{x}_1 - \bar{x}_2)$ là ước lượng điểm cho khác biệt giữa hai trung bình tổng thể. Nếu hai mẫu có cỡ đều lớn (trên 30 quan sát) thì phân phối mẫu của $(\bar{x}_1 - \bar{x}_2)$ có thể xem như tuân theo phân phối bình thường (chuẩn).

Khi xây dựng khoảng ước lượng cho sự khác biệt giữa hai giá trị trung bình hai tổng thể chúng ta xem xét hai trường hợp: đã biết phương sai tổng thể (σ_1^2 với σ_2^2) và không biết phương sai tổng thể (σ_1^2 với σ_2^2); trong trường hợp thứ hai (không biết phương sai) ta lại xét tiếp hai tình huống là cả hai mẫu đều lớn ($n_1 \geq 30$ và $n_2 \geq 30$) và một trong hai mẫu là nhỏ ($n_1 < 30$ và/hoặc $n_2 < 30$)

1. Trường hợp biết phương sai tổng thể

Khi biết phương sai của hai tổng thể thì ta xây dựng khoảng ước lượng với độ tin cậy $(1-\alpha)$ cho chênh lệch $(\mu_1 - \mu_2)$ với giả định tổng thể có phân phối bình thường hoặc nếu cỡ mẫu của hai mẫu đều lớn (≥ 30) thì phân phối của $(\bar{x}_1 - \bar{x}_2)$ xấp xỉ bình thường. Công thức của khoảng ước lượng là

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Ví dụ: Một công ty nghiên cứu thị trường được thuê thực hiện một cuộc khảo sát khách hàng của một chuỗi cửa hàng thực phẩm lớn để ước lượng sự khác biệt trong thời gian trung bình mỗi lần ghé cửa hàng của khách

hàng nam và khách hàng nữ. Các nghiên cứu trước đó cho biết độ lệch chuẩn là 11 phút đối với khách nam và 16 phút đối với khách nữ. Công ty đã chọn mẫu nhiên 100 khách nam và 100 khách nữ vào những thời điểm khác nhau ở các cửa hàng khác nhau trong chuỗi cửa hàng này để khảo sát. Kết quả là thời gian trung bình của khách nam tại cửa hàng là 34,5 phút còn thời gian trung bình của khách nữ là 42,4 phút.

Để xây dựng khoảng ước lượng với độ tin cậy 95% cho khác biệt giữa hai trung bình tổng thể ta thực hiện các bước sau:

Gọi

- Thời gian trung bình tổng thể tại cửa hàng của khách nam là μ_1 và khách nữ là μ_2
- Độ lệch chuẩn tổng thể thời gian tại cửa hàng của khách nam là σ_1 và khách nữ là σ_2

Thời gian trung bình mẫu của khách nam là $\bar{x}_1 = 34,5$ và khách nữ là $\bar{x}_2 = 42,4$

Với độ tin cậy 95% ta có giá trị tối hạn $Z_{\alpha/2} = 1,96$

Khoảng tin cậy cho khác biệt thời gian tại cửa hàng của khách nam và khách nữ ($\mu_1 - \mu_2$) được xác định bằng cách thế số liệu vào công thức

$$\begin{aligned} & (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= (34,5 - 42,4) \pm 1,96 \sqrt{\frac{11^2}{100} + \frac{16^2}{100}} \\ &= -7,9 \pm 3,8056 \end{aligned}$$

Vậy khoảng ước lượng 95% cho khác biệt giữa hai trung bình tổng thể là $-11,7056 \leq \mu_1 - \mu_2 \leq -4,0944$

Như vậy với độ tin cậy 95% ta kết luận là trung bình khách hàng nữ mất nhiều thời gian tại cửa hàng hơn khách hàng nam từ 4,0944 đến 11,7056 phút.

2. Trường hợp không biết phương sai tổng thể, mẫu lớn

Khi mẫu lớn chúng ta áp dụng định lý giới hạn trung tâm nên phân phối của $(\bar{x}_1 - \bar{x}_2)$ xấp xỉ phân phối bình thường, do đó công thức xây dựng khoảng ước lượng với độ tin cậy $(1-\alpha)$ cho chênh lệch $(\mu_1 - \mu_2)$ giống công

thức tinh huống trên nhưng vì không biết phương sai tổng thể nên ta thay thế σ bằng s như sau

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

3. Trường hợp không biết phương sai tổng thể, mẫu nhỏ

Ta phải có giả định là tổng thể có phân phối bình thường, phương sai của các tổng thể bằng nhau. Không biết phương sai tổng thể nên ta sẽ ước lượng nó bằng phương sai mẫu và chúng ta sẽ dùng phân phối t để xây dựng khoảng ước lượng. Do có giả định phương sai hai tổng thể bằng nhau nên ta sẽ gộp lại các phương sai mẫu lại để ước lượng phương sai tổng thể theo công thức tính độ lệch chuẩn mẫu gộp như sau

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

Khoảng ước lượng với độ tin cậy $(1-\alpha)$ cho chênh lệch $(\mu_1 - \mu_2)$ được xây dựng với giá trị tới hạn t gồm $df = (n_1+n_2-2)$ bậc tự do như sau

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2; df} * s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Ví dụ: Một nhà phân tích tài chính của một công ty mới giới chứng khoán muốn phân tích có hay không có sự khác biệt giữa lợi tức của các cổ phiếu được liệt kê trong danh sách cổ phiếu của chỉ số NYSE và NASDAQ. Anh ta tổng hợp được bảng số liệu sau cho các cổ phiếu anh ta chọn vào mẫu nghiên cứu:

Số cổ phiếu quan sát	21	25
Lợi tức trung bình mẫu	3,27	2,53
Độ lệch tiêu chuẩn của lợi tức	1,30	1,16

Giả định là phương sai về lợi tức của hai tổng thể bằng nhau, anh ta có tìm thấy sự khác biệt trong lợi tức cổ phiếu trung bình hay không (chọn $\alpha = 0.05$).

Nhà đầu tư này sẽ giải quyết bài toán thống kê qua các bước:

Đặt μ_1 là lợi tức trung bình của các cổ phiếu trong chỉ số NYSE
còn μ_2 là lợi tức trung bình của các cổ phiếu trong chỉ số
NASDAQ

Đây là tinh huống 2 cỡ mẫu nhỏ, nhỡ có giả định hai phương sai tổng thể bằng nhau nên anh ta tính toán độ lệch chuẩn mẫu gộp như sau

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(21-1)1.30^2 + (25-1)1.16^2}{21+25-2}} = 1,2256$$

Sau đó khoảng ước lượng với độ tin cậy $(1-\alpha) = 95\%$ cho chênh lệch ($\mu_1 - \mu_2$) được xây dựng với giá trị tối hạn t gồm df = $(n_1+n_2-2) = (21+25-2) = 44$ bậc tự do như sau (với giá trị tối hạn $t_{(0,025;44)} = 2,015$ được tra từ bảng tính Excel)

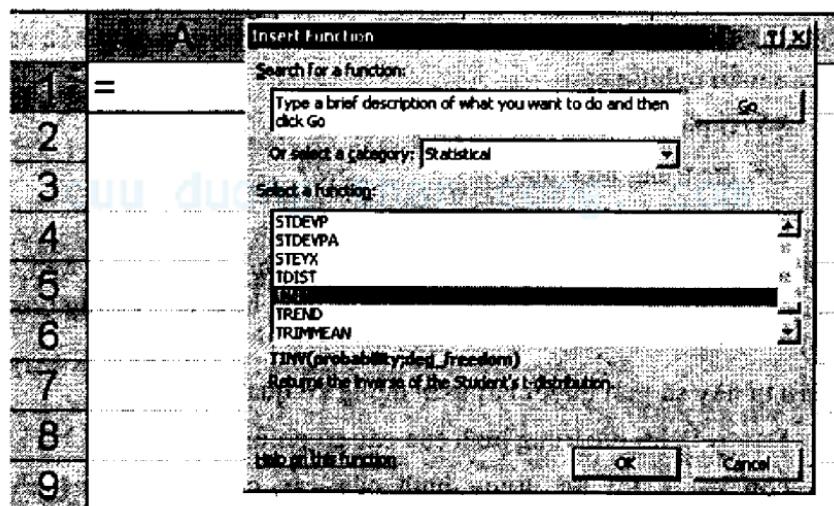
$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{\alpha/2, df} * s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$
$$= (3,27 - 2,53) \pm 2,015 * 1,2256 \sqrt{1/21 + 1/25} = 0,74 \pm 0,731$$
$$= 0,009 \leq \mu_1 - \mu_2 \leq 1,471$$

Như vậy với độ tin cậy 95% có thể kết luận các cổ phiếu trong chỉ số NYSE có mức lợi tức trung bình cao hơn các cổ phiếu trong chỉ số NASDAQ. Mức độ cụ thể của khoảng chênh lệch bạn đọc có thể tự nhận định dựa vào khoảng ước lượng tìm được.

Nhân đây xin hướng dẫn các bạn cách tra các giá trị tối hạn t khi gặp phải tinh huống bậc tự do không có trong bảng tra.

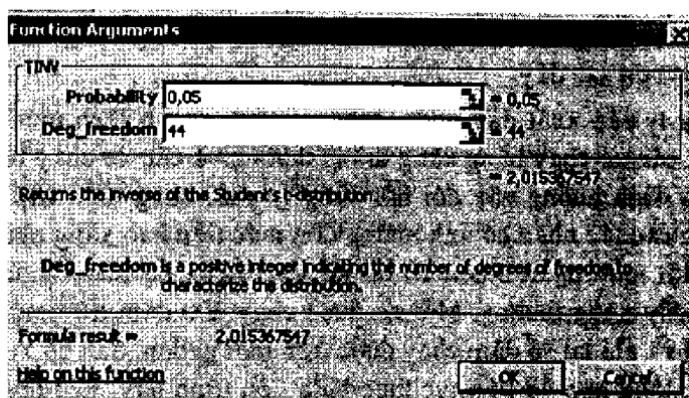
Bạn khởi động màn hình làm việc của Excel lên và nhập dấu = để sẵn sàng cho việc gọi hàm. Sau đó bạn vào menu Insert chọn lệnh Function để mở cửa sổ Insert Function, tiến hành các lựa chọn như trong Hình 7.3.

Hình 7.3



Bạn nhập các thông tin sau vào cửa sổ xuất hiện kế tiếp là cửa sổ trong hình sau.

Hình 7.4



Nhấn nút OK bạn được kết quả. Như vậy là bạn cần phải chú ý khi nhập số liệu về mức ý nghĩa vào khung Probability trong cửa sổ Function Arguments trong hàm tìm giá trị t tới hạn bạn phải nhập giá trị α chứ không phải $\alpha/2$ mặc dù thông tin bạn cần là của $\alpha/2$, Excel sẽ tự động chia đôi mức ý nghĩa và tìm cho bạn giá trị đúng. Điều này có nghĩa là nếu bạn cần một giá trị t tới hạn của mức ý nghĩa 10% và không bị chia đôi (tức $t_{0.1, df}$) bạn cần phải tự động nhập giá trị 20% (tức 0,2) chứ không phải nhập giá trị 0,1 thì mới được giá trị đúng.

*Nếu giả định phương sai tổng thể bằng nhau không đạt được?

Trong tình huống ước lượng cỡ mẫu nhỏ mà không đạt được giả định phương sai hai tổng thể bằng nhau chúng ta không thể ước lượng phương sai tổng thể bằng phương sai gộp được mà phải sử dụng công thức xây dựng khoảng ước lượng như sau

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2; df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Bậc tự do df lúc này được tính theo công thức sau đây

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{s_1^2}{n_1} \right)^2 + \left(\frac{s_2^2}{n_2} \right)^2} \cdot (n_1 - 1) + (n_2 - 1)$$

7.4.1.2 Ước lượng khác biệt hai trung bình tổng thể trong trường hợp mẫu cặp

Ta đã nghiên cứu tình huống kiểm định trung bình hai mẫu độc lập, ta cũng đã biết trong thực tế nghiên cứu có khi ta không dùng mẫu độc lập được như ví dụ sau đây:

Một công ty sản xuất dầu nhớt muốn so sánh chênh lệch mức tiêu hao xăng của các xe chạy bằng xăng thông thường và các xe chạy bằng xăng tổng hợp. Tình huống này đòi hỏi công ty phải dùng mẫu cặp để kiểm soát được các tác nhân có ảnh hưởng đến mức tiêu hao xăng như loại xe, tài xế... Một mẫu ngẫu nhiên 10 tài xế và xe của họ được chọn, các xe được đổ đầy xăng thường. Mỗi xe được chạy 200 dặm trên một lộ trình nhất định và ghi lại số dặm chạy được trên mỗi gallon. Sau đó xe được đổ đầy xăng tổng hợp và tiếp tục thực hiện tiến trình này. Kết thúc nghiên cứu chúng ta có hai mẫu dữ liệu về số dặm đi được trên mỗi gallon cho cả hai loại xăng mà thực ra hai mẫu dữ liệu này là một bộ mẫu cặp vì được ghi trên cùng một tài xế và một loại xe theo từng cặp đối xứng về loại xăng. (Xem Bảng 7.3).

Để có thể xây dựng khoảng ước lượng cho chênh lệch trung bình tổng thể μ_d cho mức tiêu hao xăng của hai loại xăng, đầu tiên chúng ta phải tính được từng cặp chênh lệch trên mẫu theo công thức

$$d_i = x_{1i} - x_{2i}$$

Bài toán của chúng ta có giả định là hai tổng thể có phân phối bình thường, nếu không có giả định này thì ta phải có cỡ mẫu lớn.

Giá trị trung bình của các chênh lệch trên mẫu là \bar{d} được tính theo công thức

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

Trong công thức trên n là số cặp chênh lệch. Giá trị \bar{d} là ước lượng điểm cho trung bình của các chênh lệch tổng thể μ_d .

Độ lệch chuẩn của các chênh lệch được tính theo công thức

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

Sau đó giá trị tối hạn t được sử dụng để xây dựng khoảng ước lượng với độ tin cậy $(1-\alpha)$ cho chênh lệch μ_d theo công thức:

$$\bar{d} \pm t_{(\alpha/2; n-1)} \frac{s_d}{\sqrt{n}}$$

Bảng 7.3

Mức tiêu hao của Xăng tổng hợp (dặm/gallon)	Mức tiêu hao của Xăng thông thường (dặm/gallon)	d_i	$(d_i - \bar{d})^2$
19,8	20,7	-0,9	10,05
28,8	25,8	3	0,53
20,4	27,8	-7,4	93,51
18,7	14,9	3,8	2,34
23,4	21,6	1,8	0,22
27,1	21,1	6	13,91
28,4	28	0,4	3,50
21,4	13	8,4	37,58
26,4	24,4	2	0,07
19,9	14,3	5,6	11,09
Tổng		22,7	172,80

Ta tính được các số liệu trung gian như sau

$$\bar{d} = 22,7/10 = 2,27$$

$$sd = \sqrt{(172,8/9)} = 4,38$$

Chọn độ tin cậy 95% ta có $t(0,025; 9) = 2,2622$

Vậy khoảng ước lượng tính được như sau

$$\bar{d} \pm t_{(\alpha/2; n-1)} \frac{s_d}{\sqrt{n}}$$

$$= 2,27 \pm 2,2622 * (4,38/\sqrt{10})$$

$$= 2,27 \pm 3,13$$

$$-0,86 \leq \mu d \leq 5,4$$

Với độ tin cậy 95% có thể kết luận mức tiêu hao của hai loại xăng này không khác nhau.

Trước khi kết thúc nội dung này chúng ta xác định đâu là chìa khóa để biết nên hay không nên dùng mẫu cắp, đó là khi có hay không có những nhân tố tác động đến kết quả nghiên cứu. Nếu bạn tin rằng không cần kiểm soát các nhân tố tác động bên ngoài thì ta có thể dùng mẫu độc lập.

7.4.2 Ước lượng tỷ lệ hai mẫu

Trong nội dung này chúng ta nghiên cứu phương pháp ước lượng chênh lệch giữa hai tỷ lệ tổng thể vì trong nhiều tình huống nghiên cứu kinh tế đây cũng là một dạng ước lượng thống kê hay gấp.

Một công ty quảng cáo đang muốn kiểm tra mức độ thu hút của một chương trình quảng cáo có như nhau hay không đối với 2 khán giả thị trường khách hàng nam giới và nữ giới trước khi phát nó rộng rãi trên TV.

Hiển nhiên không có cách nào để đo lường thái độ của toàn bộ tổng thể khách hàng nam giới và nữ giới mà họ phải nghiên cứu mẫu bằng cách phát đoạn quảng cáo này cho 2 mẫu ngẫu nhiên 425 khách nam và 370 khách nữ, trong mẫu khách nam có 240 người ưa thích mẫu quảng cáo này còn mẫu khách nữ có 196 người ưa thích. Căn cứ trên thông tin này ta tính được tỷ lệ ưa thích của hai mẫu như sau

p_{s1} là tỷ lệ ưa thích của mẫu nam = $240/425 = 0,565$

p_{s2} là tỷ lệ ưa thích của mẫu nữ = $196/370 = 0,530$

$(p_{s1} - p_{s2}) = 0,565 - 0,530 = 0,035$ là ước lượng điểm cho khác biệt của hai tỷ lệ tổng thể là $(p_1 - p_2)$

Ta thấy rõ ràng 0,035 là ước lượng điểm đơn giản nhất cho khác biệt trong tỷ lệ ưa thích của nam và nữ trên tổng thể đối với chương trình quảng cáo. Tuy nhiên để giảm sai số do chọn mẫu chúng ta nên xây dựng một khoảng ước lượng cho khác biệt tỷ lệ tổng thể theo công thức sau đây (với điều kiện cỡ mẫu đủ lớn). Quy tắc xác định cỡ mẫu thế nào là đủ lớn khá đơn giản, nếu $np_s \geq 5$ và $n(1-p_s) \geq 5$ cho cả hai mẫu

$$(p_{s1} - p_{s2}) \pm z_{\alpha/2} \sqrt{\frac{p_{s1}(1-p_{s1})}{n_1} + \frac{p_{s2}(1-p_{s2})}{n_2}}$$

Trong đó, tổng quát

p_{s1} là tỷ lệ ưa thích của mẫu 1

p_{s2} là tỷ lệ ưa thích của mẫu 2

n_1 là cỡ của mẫu 1

n_2 là cỡ của mẫu 2

Áp dụng trở lại ví dụ của chúng ta, ta tính toán được khoảng ước lượng cho chênh lệch tỷ lệ với độ tin cậy 95% như sau ($z_{\alpha/2} = z_{0,025} = 1,96$)

$$(p_{s1} - p_{s2}) \pm z_{\alpha/2} \sqrt{\frac{p_{s1}(1-p_{s1})}{n_1} + \frac{p_{s2}(1-p_{s2})}{n_2}}$$

$$= (0,035) \pm 1,96 * \sqrt{\frac{0,565(1-0,565)}{425} + \frac{0,53(1-0,53)}{370}}$$

$$= 0,035 \pm 0,069$$

$$- 0,034 \leq p_1 - p_2 \leq 0,104$$

Vì vậy căn cứ trên dữ liệu mẫu và sử dụng phương pháp ước lượng khoảng với độ tin cậy 95% công ty kết luận rằng khác biệt thực sự giữa tỷ lệ tổng thể khách hàng nam ưa thích quảng cáo và khách hàng nữ ưa thích quảng cáo là từ -3,4% đến 10,4%. Một phía, tỷ lệ khách nữ thích quảng cáo này nhiều hơn tỷ lệ khách nam thích quảng cáo là 3,4%. Một phía, tỷ lệ khách nam thích quảng cáo này nhiều hơn tỷ lệ khách nữ là 10,4%. Vì giá trị zero nằm trong khoảng ước lượng này ta có thể kết luận là không có sự khác biệt trong mức độ ưa thích mẫu quảng cáo của cả khách nam và khách nữ.

cuu duong than cong. com

cuu duong than cong. com

CHƯƠNG 8

KIỂM ĐỊNH GIẢ THUYẾT VỀ THAM SỐ TỔNG THỂ

8.1 CÁC VẤN ĐỀ CHUNG VỀ KIỂM ĐỊNH

Chương này tập trung vào nội dung kiểm định giả thuyết, một công cụ khác của thống kê suy diễn, chúng ta sẽ thảo luận từng bước cách thức để có thể suy diễn về một tham số tổng thể căn cứ trên việc dùng các thông tin mẫu phân tích giả thuyết ta đặt ra về các trị số của các tham số tổng thể. Những giả thuyết đặt ra về tham số tổng thể được gọi là giả thuyết thống kê, nó có thể được quá trình kiểm định giả thuyết kết luận là đúng hoặc sai.

8.1.1 Đặt giả thuyết về tham số tổng thể

Thủ tục kiểm định giả thuyết luôn bắt đầu với việc đặt giả thuyết, nó là một phát biểu, một nhận định, một đề xuất về tham số tổng thể, bất kỳ một bài toán kiểm định nào cũng phải có 1 cặp giả thuyết bao gồm giả thuyết không H_0 (Null Hypothesis) và giả thuyết đối H_1 (Alternative Hypothesis). Với kiểm định bạn chỉ có thể đi đến 1 trong 2 quyết định: không bác bỏ H_0 (tức loại H_1) hoặc ngược lại bác bỏ H_0 (tức chấp nhận H_1)

8.1.2 Một số nguyên tắc liên quan đến việc đặt giả thuyết

- Giả thuyết H_0 thường mô tả hiện tượng lúc bình thường, mô tả tình trạng nguyên thủy, hoặc tình trạng không chịu tác động gì của hiện tượng. Khi xây dựng H_0 , trong cấu trúc của nó phải luôn luôn có một dấu bằng, có thể là dấu $=, \leq, \geq$
- Giả thuyết H_1 mô tả tình trạng ngược lại với H_0 , nó thể hiện các nghi ngờ, các nhận định về hiện tượng mà bạn đang muốn chứng minh trong bài toán kiểm định của mình. Khi xây dựng H_1 , trong cấu trúc của nó không được có dấu bằng, nó có thể là dấu $\neq, <, >$ tùy theo tình huống.
- Nếu bạn loại bỏ H_0 tức là bạn có bằng chứng thống kê để cho rằng H_1 đúng.
- Nếu bạn không loại H_0 tức là bạn không chứng minh thống kê được là H_1 đúng.

Tuy nhiên cần nhớ rằng sự thất bại trong việc loại H_0 không đồng nghĩa với việc bạn đã chứng minh được H_0 đúng, chỉ là bạn không đủ bằng chứng thống kê để loại nó mà thôi. Lý do một nhà thống kê không bao giờ có thể chứng minh H_0 hoàn toàn đúng là vì họ chỉ luôn có thông tin

trên một mẫu dữ liệu chứ không phải toàn bộ tổng thể. Tình huống này rất giống các tiền đề trong một phiên tòa, khi mở đầu phiên tòa hội đồng xét xử phải xem như bị cáo đang trong tình trạng vô tội, và tìm bằng chứng để kết tội, nếu hội đồng không có đủ bằng chứng chứng minh bị cáo có tội, họ đi đến kết luận là “không có đủ bằng chứng buộc tội”, các bạn nhớ là “không có đủ bằng chứng buộc tội” chứ không phải “vô tội”, vì các bằng chứng trong phiên tòa không đủ để kết tội bị cáo chưa hẳn có nghĩa là bị cáo vô tội, tất nhiên không loại trừ tình huống vô tội thật sự, nhưng kết luận như thế sẽ chặt chẽ hơn.

Ví dụ: Giám đốc điều hành sản xuất của một nhà máy chế biến các loại thực phẩm ăn liền đang đặc biệt quan tâm đến dây chuyền tự động đóng hộp ngũ cốc dinh dưỡng cho người ăn kiêng. Theo đúng qui định thì trọng lượng tịnh của mỗi hộp ngũ cốc là 368gram. Nhưng ông ta nghi ngờ rằng có thể dây chuyền gấp trực trặc gì đó khiến quy định trên không được bảo đảm. Ông ta chọn một mẫu ngẫu nhiên các hộp ngũ cốc trong kho, dùng các thông tin tìm được trên mẫu này để tiến hành bài toán kiểm định của mình.

Việc đầu tiên của ông ta là đặt giả thuyết cho kiểm định, H_0 được đặt ra trên cơ sở cho rằng dây chuyền sản xuất vẫn đang hoạt động bình thường tức là trọng lượng trung bình của các hộp ngũ cốc vẫn là 368gram, do đó ta viết $H_0: \mu = 368$ gram. Như vậy trong cấu trúc H_0 đã bảo đảm có một dấu bằng, và nó đã mô tả nhận định về một tham số tổng thể, không phải mẫu.

H_1 mô tả tình trạng ngược lại của H_0 , thể hiện nghi ngờ mà ông giám đốc điều hành đang muốn chứng minh. Lúc này nếu cho rằng trực trặc của dây chuyền làm nó đóng hộp một lượng bột nhiều hơn tiêu chuẩn thì ông ta sẽ đặt $H_1: \mu > 368$, nếu ngờ rằng trọng lượng tịnh của bột được đóng dưới quy định thì ông đặt $H_1: \mu < 368$, còn nếu không biết rõ hướng sai lệch mà chỉ nghi ngờ chung là dây chuyền gấp trực trặc trong vấn đề đóng hộp thì ông ta sẽ đặt $H_1: \mu \neq 368$. Như các bạn thấy, các tình huống của H_1 đều không có dấu bằng.

8.1.3 Logic của bài toán kiểm định

Chúng ta sẽ tìm hiểu logic của phương pháp kiểm định giả thuyết qua cách thức sử dụng thông tin mẫu để quyết định về sự đúng đắn của H_0 . Trở lại ví dụ về kiểm định khối lượng ngũ cốc đóng hộp của chúng ta, giả thuyết H_0 là khối lượng tịnh của ngũ cốc đóng hộp trong tổng thể đúng 368gram. Một mẫu các hộp ngũ cốc đã đóng bột được chọn ngẫu nhiên, cân trọng lượng bột của các hộp và tính trọng lượng trung bình. Con số

trung bình này là một thông tin ước lượng cho con số tổng thể mà từ đó mẫu được rút. Ta vẫn biết rằng ngay cả nếu H_0 thực sự đúng như ta đã giả thuyết tức $\mu = 368$ thì con số trung bình mẫu tính được trên mẫu đã rút vẫn có khả năng khác giá trị 368 vì trung bình mẫu là một biến ngẫu nhiên có giá trị khác nhau trong các lần lấy mẫu khác nhau. Tuy nhiên ta hy vọng rằng con số thống kê mẫu sẽ gần với con số thống kê tổng thể đã giả thuyết nếu quả thực H_0 đúng (điều này đã được khám phá và chứng minh trong nội dung Phân phối các tham số mẫu ở Chương 6)

Như trong tình huống trung bình mẫu tính được chỉ là $367,9 < 368$ thì trực giác vẫn có thể khiến ta tin rằng dây chuyền không có vấn đề gì (tức μ vẫn đúng bằng 368) vì trị trung bình mẫu 367,9 rất gần giá trị tổng thể ta đã giả thuyết 368. Chứng tỏ mẫu này phải được chọn từ một tổng thể có trọng lượng trung bình là 368 gram, do đó ta không bác bỏ H_0 .

Ngược lại, nếu có chênh lệch lớn giữa giá trị trung bình mẫu và giá trị tổng thể ta đã giả thuyết thì theo trực giác ta có thể nhận thấy rằng H_0 khó mà đúng, ví dụ khi bạn tính được trọng lượng trung bình mẫu là 340 bạn nhiều phần đoán chắc dây chuyền đóng hộp ngũ cốc đã bị sai lệch với giá trị theo quy định vì 340 quá khác xa 368, theo logic khó mà tin rằng mẫu này được chọn từ một tổng thể có trọng lượng trung bình đúng 368 gram. Vậy là chúng ta bác bỏ H_0 .

Trong các bài toán kiểm định thực tế số liệu giúp ra quyết định không đơn giản và rõ ràng như trong ví dụ trên, vào những lúc ấy, đánh giá thế nào là giá trị mẫu rất gần, gần hay quá xa, xa... giá trị tổng thể đã giả thuyết hoàn toàn phụ thuộc vào cảm tính cá nhân hay sao? Không thể như thế được, cần phải có một tiêu chuẩn chung nào đó để quyết định thế nào là gần hay xa... Lý thuyết kiểm định thống kê sẽ đưa ra quy tắc quyết định giúp cho bạn định lượng kết luận của mình. Điều này đạt được bằng cách tính toán giá trị kiểm định thống kê căn cứ trên các số liệu mẫu đã lấy được, sử dụng phân phối của giá trị kiểm định để quyết định giá trị kiểm định này cho phép bác bỏ hay không bác bỏ H_0 .

8.1.4 Xác suất sai lầm loại I và Xác suất sai lầm loại II

Khi kiểm định một giả thuyết thống kê ta có thể phạm phải những sai lầm như sau: dựa trên những thông tin từ mẫu ta có thể bác bỏ một giả thuyết mà thực ra giả thuyết này đúng hoặc không bác bỏ một giả thuyết trong khi trên thực tế nó sai. Người ta định nghĩa Xác suất sai lầm như sau:

- Xác suất sai lầm loại I kí hiệu là α , là xác suất để chúng ta bác bỏ giả thuyết H_0 trong khi thật sự nó đúng, tức là

$$\alpha = P(\text{sai lầm loại I}) = P(\text{loại } H_0/H_0 \text{ đúng})$$

- Xác suất sai loại II, kí hiệu β là xác suất để chúng ta không bác bỏ H_0 khi nó sai:

$$\beta = P(\text{sai lầm loại II}) = P(\text{không loại } H_0/H_0 \text{ sai})$$

Trong khi kiểm định nếu ta bác bỏ H_0 tức ta đang đứng trước nguy cơ phạm sai lầm loại I còn nếu ta không bác bỏ H_0 thì ta đối mặt với sai lầm loại II.

Giá trị α xác định nên diện tích vùng bác bỏ giả thuyết H_0 , nếu giá trị kiểm định tính toán trên các thông tin mẫu rơi vào vùng này ta sẽ quyết định bác bỏ H_0 , khi ta tăng α thì diện tích vùng loại bỏ tăng, làm khả năng bác bỏ H_0 tăng, dẫn đến khả năng phạm sai lầm loại II giảm tức β giảm. Ngược lại thu nhỏ giá trị α thì khả năng bác bỏ H_0 giảm nên khả năng phạm sai lầm loại I bé đi nhưng tăng nguy cơ mắc sai lầm loại II, đồng nghĩa với việc giá trị β bị tăng lên. Quan hệ này xảy ra trong tình huống cỡ mẫu của nghiên cứu không đổi.

8.1.5 Mức ý nghĩa của kiểm định (Significance level)

Giá trị xác suất phạm sai lầm loại I (α) phải được ấn định trước khi tiến hành kiểm định, trong thực tế các mức $\alpha = 0,1$; $\alpha = 0,05$ hoặc $\alpha = 0,01$ hay được chọn. Việc chọn lựa giá trị của α lớn hay bé tùy thuộc vào mức độ tổn thất người làm kiểm định có thể “chịu đựng” nếu sai lầm loại I xảy ra, nếu sai lầm loại I gây tổn thất không cao thì chọn α lớn. Ví dụ với $\alpha = 0,05$ có nghĩa là có 5% may rủi H_0 đúng và quyết định bác bỏ H_0 là một sai lầm. Quyết định này dựa vào dữ kiện của mẫu chứ không căn cứ trên toàn thể tổng thể nên chúng ta không thể nào hoàn toàn tin chắc vào kết quả kiểm định được và vì thế cần đưa ra thông tin về quyết định với một xác suất sai lầm nào đó mà ta có thể phạm phải. Nếu ta bác bỏ giả thuyết H_0 với mức $\alpha = 0,05$ có nghĩa là nếu lặp lại thử nghiệm 100 lần với 100 mẫu khác nhau và cứ mỗi lần vậy lại bác bỏ giả thuyết H_0 thì sẽ có khoảng 5 lần chúng ta có thể phạm sai lầm loại I (nói cách khác ta có may rủi 5% phạm sai lầm loại I). Chúng ta có 5% khả năng phạm phải lỗi đã bác bỏ H_0 khi thực tế nó đúng, nói theo chiều khác, chúng ta tin cậy 95% là ta đã quyết định đúng. Lúc này ta phát biểu là giả thuyết đã bị bác bỏ ở mức ý nghĩa 5%. Đại lượng α được gọi tên là mức ý nghĩa của kiểm định.

Nếu bạn thấy một nhà thống kê làm kiểm định và kết luận rằng kiểm định của họ “có ý nghĩa thống kê ở mức p%” tức là họ đã đi đến bác bỏ H_0 và có thể sai tối đa chỉ p% với kết luận đó vì “mức có ý nghĩa thống kê” trong một kiểm định là xác suất sai lầm loại I tối đa mà người làm

kiểm định có thể phạm phải khi bác bỏ H_0 . Khái niệm này gắn với nội dung giá trị xác suất p-value sẽ được trình bày ở phần sau.

Giá trị $(1-\alpha)$ được gọi tên là độ tin cậy, ngược với α , nó xác định nên vùng chấp nhận H_0 .

Như nói ở trên, xác suất phạm phải sai lầm loại II được biểu thị bằng kí hiệu β . Không giống như α ta không kiểm soát β được. Thế nhưng có mối liên hệ nghịch giữa α và β nói cách khác α tăng thì β đi nhưng chú ý là lượng thay đổi ở β không tỉ lệ trực tiếp với lượng thay đổi ở α . Nếu muốn giảm đồng thời cả α và β bạn có thể tăng cỡ mẫu.

8.1.6 Giá trị tối hạn (Critical Value)

Khi đã xác định được α thì bạn xác định được vùng bác bỏ và vùng chấp nhận H_0 , tức là bạn cũng suy ra được giá trị tối hạn, nó là biên giới chia đôi hai vùng chấp nhận và bác bỏ H_0 trên phân phối của giá trị kiểm định. Nếu giá trị thống kê kiểm định rơi vào vùng bác bỏ thì chúng ta bác bỏ H_0 và ngược lại, nếu nó rơi vào vùng chấp nhận H_0 thì ta không bác bỏ H_0 .

Vùng bác bỏ là khu vực chứa các giá trị thống kê kiểm định không có khả năng xảy ra nếu thật sự H_0 đúng, do đó nếu một giá trị thống kê kiểm định rơi vào vùng này ta sẽ bác bỏ H_0 vì ta suy luận rằng giá trị này đã không thể xảy ra nếu H_0 đúng.

8.1.7 Kiểm định một bên và kiểm định hai bên.

Ví dụ nếu giám đốc điều hành của công ty sản xuất thực phẩm dinh dưỡng đặt giả thuyết như sau

$$H_0: \mu = 368$$

$$H_1: \mu \neq 368$$

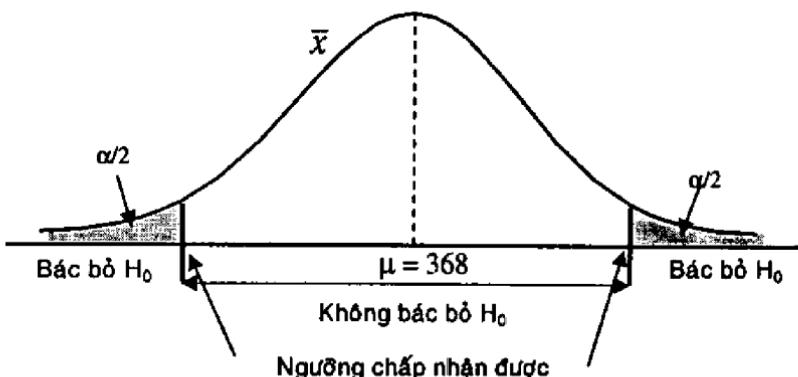
Thì lúc này ta có kiểm định hai bên vì miền bác bỏ nằm ở hai bên của phân phối, các bạn hãy xem lập luận sau:

Ta biết trung bình mẫu \bar{x} có phân phối xấp xỉ bình thường với trung bình của phân phối bằng μ khi cỡ mẫu đủ lớn. Giả thuyết H_0 đặt ra là $\mu = 368$. Hình sau thể hiện phân phối của Trung bình mẫu khi coi như giả thuyết H_0 đúng tức trung bình của phân phối này $= \mu$ và $\sigma = 368$. Nếu H_0 đúng thì \bar{x} không thể quá lệch xa so với 368 mà phải tập trung gần μ trong mức chấp nhận được ở cả 2 bên của phân phối

Khu vực tõ mờ 2 bên hình là vùng bác bỏ H_0 . Diện tích vùng bác bỏ cho ta xác suất có các \bar{x} xa μ quá ngưỡng chấp nhận được (ngưỡng chấp nhận này nếu được chuyển thành đơn vị đo lường độ lệch chuẩn qua phương pháp chuẩn hóa dữ liệu, thì chính là giá trị tối hạn). Nên tổng

diện tích vùng tô mờ chính là xác suất phạm sai lầm loại I, nó là α . Vì chia hai nên mỗi bên còn $\alpha/2$. Giả dụ bạn chọn $\alpha = 0,05$ thì $\alpha/2 = 0,025$ nên nếu chuyển hóa qua phương pháp chuẩn hóa dữ liệu để dùng phân phối Z bạn suy ra được với diện tích 0,025 thì giá trị tối hạn sẽ là $\pm 1,96$

Hình 8.1



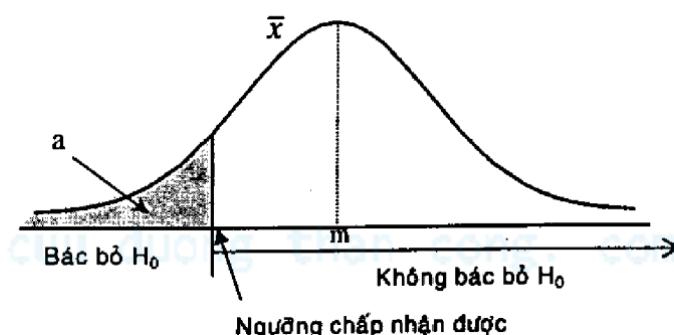
Còn nếu đặt giả thuyết theo kiểu sau:

$$H_0: \mu = 368 \text{ hoặc } H_0: \mu \neq 368$$

$$H_1: \mu < 368$$

Thì ta có kiểm định bên trái vì lúc này miền bác bỏ nằm ở bên trái của phân phối, diện tích miền bác bỏ là α . Xem hình minh họa

Hình 8.2



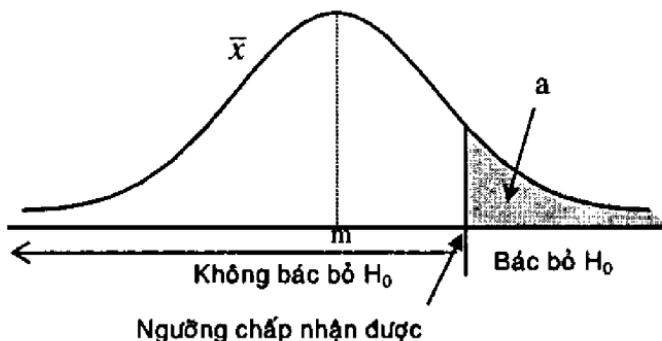
Hoặc đặt giả thuyết

$$H_0: \mu = 368 \text{ hoặc } H_0: \mu \leq 368$$

$$H_1: \mu > 368$$

Thì ta có kiểm định bên phải vì lúc này miền bác bỏ nằm ở bên phải của phân phối, diện tích miền bác bỏ là α .

Hình 8.3



Hai loại kiểm định giả thuyết sau cùng được gọi chung là kiểm định một bên vì miền bác bỏ H_0 chỉ nằm ở một bên của phân phối.

Những vấn đề chúng ta đã khảo sát ở phần 8.1 là những vấn đề chung nhất của bài toán kiểm định, trong thực tế ta có nhiều dạng kết hợp của giả thuyết không và giả thuyết đối, có nhiều mức độ tin cậy lựa chọn, có thể thực hiện kiểm định nhiều loại tham số thống kê tổng thể như kiểm định trị trung bình của một tổng thể, kiểm định tỷ lệ của một tổng thể, hoặc kiểm định kết hợp hiệu số của hai trung bình của hai tổng thể ... đây là những vấn đề hay gặp trong thực tế, phần sau chúng ta sẽ đi vào áp dụng kiểm định giả thuyết trong hai trường hợp là kiểm định giả thuyết một mẫu (dùng thông tin chỉ trên một mẫu cho quá trình kiểm định) và kiểm định giả thuyết hai mẫu (dùng thông tin trên hai mẫu cho quá trình kiểm định).

8.2 KIỂM ĐỊNH GIẢ THUYẾT MỘT MẪU

8.2.1 Kiểm định giả thuyết về trung bình tổng thể

8.2.1.1 Kiểm định giả thuyết về trung bình tổng thể khi biết độ lệch chuẩn tổng thể

Quy trình làm kiểm định đi qua các bước

1. Nhận định tình hình của tham số tổng thể ta muốn làm kiểm định.
2. Đặt giả thuyết không và giả thuyết đối về tham số tổng thể, tùy theo nhận định ở bước 1 mà ta đặt giả thuyết một bên hoặc giả thuyết hai bên.
3. Xác định mức ý nghĩa của bài toán kiểm định là α .
4. Tính toán giá trị kiểm định theo công thức.

$$z_{\alpha} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

5. Xem xét bác bỏ hay không bác bỏ giả thuyết H_0

Sau khi so sánh giá trị kiểm định tính toán được z_{α} với giá trị tối hạn trong tình huống phù hợp, quyết định loại hay không loại H_0 theo quy tắc (phân biệt hai trường hợp) :

- Nếu là kiểm định hai bên thì ta sẽ bác bỏ giả thuyết H_0 nếu $(z_{\alpha} > z_{\alpha/2} \text{ hoặc } z_{\alpha} < -z_{\alpha/2})$
- Nếu là kiểm định một bên thì ta sẽ bác bỏ giả thuyết H_0 nếu :
 - Với kiểm định bên trái $z_{\alpha} < -z_{\alpha}$
 - Với kiểm định bên phải $z_{\alpha} > z_{\alpha}$

6. Kết luận về bài toán kiểm định

8.2.1.2 Kiểm định giả thuyết về trung bình tổng thể khi không biết độ lệch chuẩn tổng thể

Quy trình làm kiểm định đi qua các bước

1. Nhận định tình hình của tham số tổng thể ta muốn làm kiểm định
2. Đặt giả thuyết không và giả thuyết đối về tham số tổng thể, tùy theo nhận định ở bước 1 mà ta đặt giả thuyết một bên hoặc giả thuyết hai bên.
3. Xác định mức ý nghĩa của bài toán kiểm định

4 Tính toán giá trị kiểm định

- Nếu cỡ mẫu lớn ($n \geq 30$) ta vẫn dùng công thức Z_{α} nhưng thay σ bằng s
 - Nếu cỡ mẫu nhỏ ($n < 30$) ta cũng thay σ bằng s và dùng công thức
- $$t_{\alpha} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

5. Xem xét bác bỏ giả thuyết H_0

Sau khi so sánh giá trị kiểm định tính toán được với giá trị tối hạn trong tình huống phù hợp, quyết định loại hay không loại H_0 theo quy tắc (phân biệt hai trường hợp):

- Nếu là kiểm định hai bên thì ta sẽ bác bỏ giả thuyết H_0 nếu $(z_{\alpha} > z_{\alpha/2} \text{ hoặc } z_{\alpha} < -z_{\alpha/2})$ khi cỡ mẫu lớn
- Nếu là kiểm định một bên thì ta sẽ bác bỏ giả thuyết H_0 nếu $(t_{\alpha} > t_{(\alpha/2; n-1)} \text{ hoặc } t_{\alpha} < -t_{(\alpha/2; n-1)})$ khi cỡ mẫu nhỏ

Với kiểm định bên trái

$Z_{tt} < -z_\alpha$ khi cỡ mẫu lớn

$t_{tt} < -t_{(\alpha/2, n-1)}$ khi cỡ mẫu nhỏ

Với kiểm định bên phải

$Z_{tt} > z_\alpha$ khi cỡ mẫu lớn

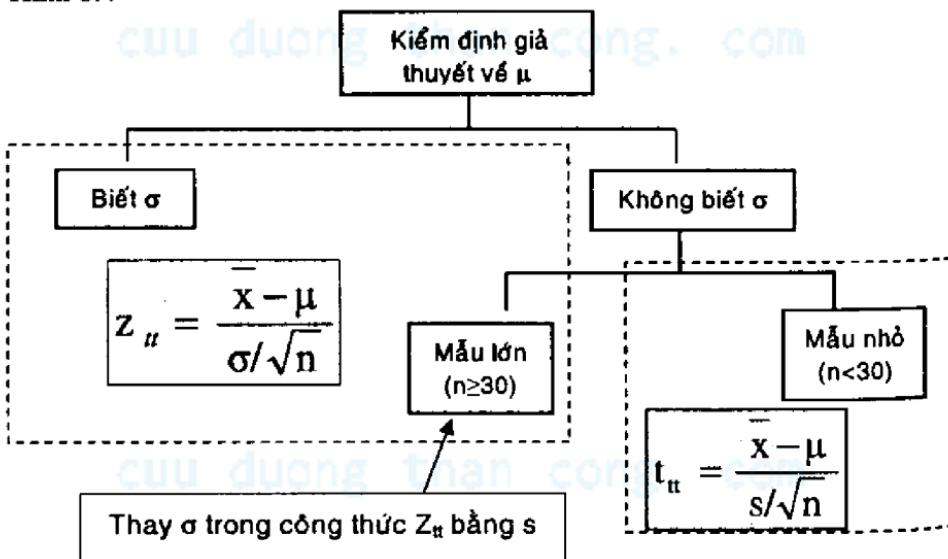
$t_{tt} > t_{(\alpha/2, n-1)}$ khi cỡ mẫu nhỏ

6. Kết luận về bài toán kiểm định

Trong quy trình của bài toán kiểm định về trung bình tổng thể ở trên, chúng ta có vài điểm cần làm rõ trước khi đi vào ví dụ thực tế.

Cũng giống như tình huống của bài toán ước lượng, khi kiểm định giả thuyết về trung bình tổng thể chúng ta phân biệt các tình huống là đã biết phương sai tổng thể và chưa biết phương sai tổng thể, nếu chưa biết phương sai tổng thể ta thay thế phương sai tổng thể bằng phương sai mẫu nhưng nếu kết hợp với tình huống cỡ mẫu của chúng ta không đủ lớn thì giá trị kiểm định của chúng ta sẽ có phân phối t student thay vì phân phối z. Các bạn xem sơ đồ phân biệt sau.

Hình 8.4



Chú ý là tính giá trị kiểm định thực ra là một thủ tục chuyển tham số mẫu (\bar{x}) thành giá trị chuẩn hóa. Quá trình quyết định bác bỏ hay không bác bỏ H_0 là việc đưa giá trị chuẩn hóa này lên phân phối tương ứng của nó (phân phối z hoặc t) để xem nó có rơi vào phạm vi của khu vực bác bỏ giả thuyết không, tức là khu vực chứa những giá trị không thể xảy ra nếu H_0 đúng.

Ví dụ 1

Trở lại ví dụ về trọng lượng trung bình của ngũ cốc đóng hộp, biết rằng ông giám đốc điều hành chọn mẫu ngẫu nhiên 25 hộp ngũ cốc thì thấy trọng lượng trung bình của mẫu này là 372,5gram; Ông ta biết độ lệch chuẩn của tổng thể là 15 gram. Ông ta quyết định chọn độ tin cậy của bài toán kiểm định là 95%. Và nghi ngờ của ông ta là dây chuyền đóng hộp ngũ cốc bị trực trặc nói chung chứ không biết chiều hướng của sai lệch là đóng quá hay đóng thiếu bột so với quy định.

1. Như vậy ông giám đốc muốn làm kiểm định để xem khối lượng trung bình của ngũ cốc đóng hộp trong tổng thể còn bằng đúng 368 gram như quy định hay không, nếu dây chuyền đóng ít bột hơn quy định hay đóng nhiều bột hơn quy định đều không chấp nhận được, khi sai sót thứ nhất xảy ra công ty mất uy tín, nếu sai sót thứ hai xảy ra công ty thiệt hại chi phí sản xuất.

2. Ông ta sẽ đặt giả thuyết không và giả thuyết đối về trung bình tổng thể như sau

$$H_0: \mu = 368$$

$$H_1: \mu \neq 368$$

Như vậy đây là bài toán kiểm định hai bên.

3. Vì ông ta chọn độ tin cậy 95% nên mức ý nghĩa của bài toán kiểm định là 5% tức $\alpha = 0,05$

4. Chúng ta có $n = 25$ là cỡ mẫu nhỏ; $\bar{x} = 372,5$ gram; đã biết $\sigma = 15$ gram. Do đó chúng ta tính toán giá trị kiểm định theo công thức

$$z_u = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{372,5 - 368}{15 / \sqrt{25}} = 1,5$$

5. Xem xét bác bỏ giả thuyết H_0

Với $\alpha = 0,05 \rightarrow \alpha/2 = 0,025 \rightarrow$ Sử dụng Bảng tra số 1 xác định được 2 giá trị tới hạn là $z_{\alpha/2} = 1,96$ và $-z_{\alpha/2} = -1,96$

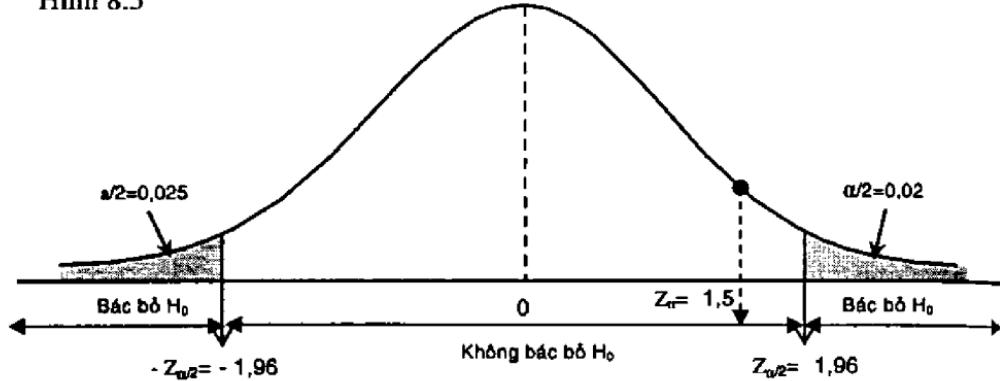
Theo quy tắc kiểm định hai bên ta sẽ bác bỏ giả thuyết H_0 nếu

$$(z_u > z_{\alpha/2} \text{ hoặc } z_u < -z_{\alpha/2})$$

Nhưng $-z_{\alpha/2} = -1,96 < z_u = 1,5 < z_{\alpha/2} = 1,96 \rightarrow$ Như vậy không bác bỏ H_0

6. Với độ tin cậy 95% chúng ta kết luận rằng không có đủ bằng chứng thống kê cho rằng dây chuyền đóng hộp ngũ cốc bị trực trặc.

Hình 8.5



Ví dụ 2

Cán bộ Sở du lịch của thành phố Đ cho rằng giá phòng khách sạn mini trong mùa thấp điểm trung bình không quá 168 ngàn/ngày đêm, một mẫu ngẫu nhiên 25 phòng khách sạn trong thành phố được chọn và người ta tính được $\bar{x} = 172,5$ ngàn đồng/ngày đêm, độ lệch chuẩn tính được từ mẫu này là $s = 15,4$ ngàn đồng/ngày đêm. Hãy kiểm định câu phát biểu của vị cán bộ có đúng hay không, chọn mức ý nghĩa cho bài toán là 5%.

1. Phát biểu của vị cán bộ đề cập đến giá phòng trung bình của các khách sạn mini trong thành phố, cho rằng nó không quá 168 ngàn đồng/ngày đêm, tức là giá phòng trung bình của tổng thể sẽ bằng hoặc dưới 168 ngàn đồng/ngày đêm.

2. Ta đặt giả thuyết như sau

$$H_0: \mu \leq 168$$

$$H_1: \mu > 168$$

Như vậy đây là bài toán kiểm định một bên, kiểm định bên phải

3. Chọn mức ý nghĩa cho bài toán là 5% $\rightarrow \alpha = 0,05$

4. Chúng ta có $n = 25$ là cỡ mẫu nhỏ; $\bar{x} = 172,5$; không biết σ mà biết $s = 15,4$. Do đó chúng ta tính toán giá trị kiểm định theo công thức

$$t_u = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{172,5 - 168}{15,4/\sqrt{25}} = \frac{4,5}{3,08} = 1,46$$

5. Xem xét bác bỏ giả thuyết H_0

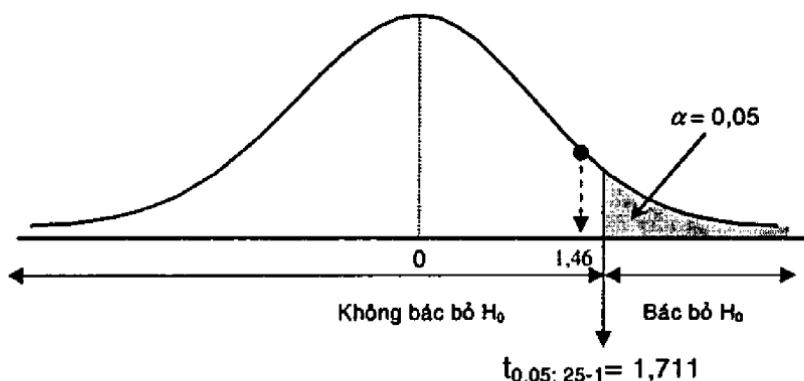
Với $\alpha = 0,05 \rightarrow$ Sử dụng Bảng tra số 2 xác định được giá trị tối hạn là $t_{n-1, \alpha} = t_{0,05; 25-1} = 1,711$

Theo quy tắc kiểm định bên phải với cỡ mẫu nhỏ ta sẽ bác bỏ giả thuyết H_0 nếu $t_u > t_{(n-1, \alpha)}$

Nhưng $t_{\alpha} = 1,46 < t_{0,05; 25-1} = 1,711$. Như vậy không bác bỏ H_0 (xem hình minh họa).

6. Như vậy với độ tin cậy 95% có thể phát biểu rằng không có đủ bằng chứng thống kê để bác bỏ giả thuyết H_0 , nhận xét của cán bộ sở du lịch không sai.

Hình 8.6



Ví dụ 3

Một công ty sản xuất phô mai nghi ngờ các nhà cung cấp sữa cho công ty đã pha thêm nước vào sữa để làm tăng lượng sữa cung cấp. Công ty biết rằng nếu như sữa có nhiều nước quá mức bình thường thì nhiệt độ đông của nó sẽ thấp hơn sữa tự nhiên (điểm đông của sữa tự nhiên trung bình khoảng $-0,545^{\circ}\text{C}$, độ lệch chuẩn của điểm đông sữa tự nhiên khoảng $0,008^{\circ}\text{C}$). Do đó họ sẽ kiểm tra chất lượng sữa bằng cách thực hiện một bài toán kiểm định xem nhiệt độ đông của sữa mà công ty đang chế biến có thấp hơn điểm đông của sữa tự nhiên không. Chọn $\alpha = 0,05$ và thực hiện kiểm tra chọn mẫu trên 25 container sữa thì thấy nhiệt độ đông trung bình của sữa trong mẫu là $-0,55^{\circ}\text{C}$.

1. Nếu sữa tự nhiên nhiệt độ đông trung bình của nó không thể dưới $-0,545^{\circ}\text{C}$, nếu nhiệt độ đông này dưới $-0,545^{\circ}\text{C}$ thì sữa bị pha nước. Như vậy bài toán kiểm định của chúng ta tập trung vào kiểm định giá trị trung bình tổng thể.

2. Từ lập luận đó ta đặt giả thuyết như sau

$$H_0: \mu \geq -0,545$$

$$H_1: \mu < -0,545$$

Như vậy đây là bài toán kiểm định một bên và là kiểm định bên trái

3. Mức ý nghĩa $\alpha = 0,05$

4. Chúng ta có $n = 25$ là cỡ mẫu nhỏ; $\bar{x} = -0,55$; đã biết $\sigma = 0,008$. Do đó chúng ta tính toán giá trị kiểm định theo công thức

$$z_u = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{-0,55 - (-0,545)}{0,008 / \sqrt{25}} = -3,125$$

5. Xem xét bác bỏ giả thuyết H_0

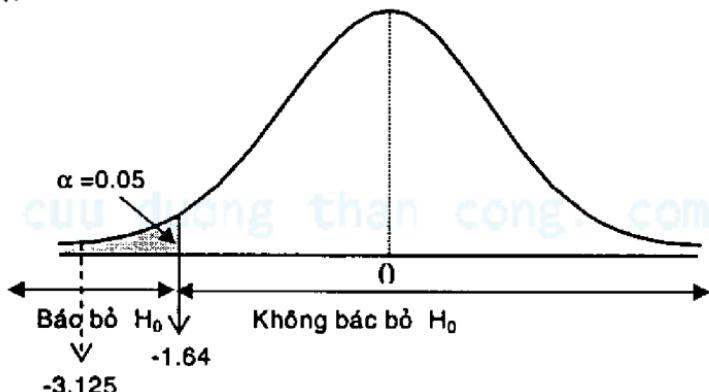
Với $\alpha = 0,05 \rightarrow$ Sử dụng Bảng tra số 1 xác định được giá trị tối hạn là $z_\alpha = z_{0,05} = 1,64$.

Đây là kiểm định bên trái nên ta suy ra giá trị $-z_\alpha$ ở phía trái của phân phối là $-1,64$.

Theo quy tắc kiểm định bên trái ta sẽ bác bỏ giả thuyết H_0 nếu $z_u < -z_\alpha$

Ta có $z_u = -3,125 < -z_\alpha = -1,64 \rightarrow$ Bác bỏ giả thuyết H_0

Hình 8.7



6. Với độ tin cậy 95% có đủ bằng chứng thống kê để bác bỏ giả thuyết H_0 , như vậy nhiệt độ đông của sữa trong mẫu nghiên cứu thấp hơn nhiệt độ đông bình thường của sữa tự nhiên là bằng chứng cho thấy quả thực sữa đã bị pha nước.

Cách tiếp cận p-value trong quy tắc bác bỏ H_0

Những năm gần đây với sự sẵn có của các phần mềm thống kê, cách tiếp cận p-value trong quy tắc quyết định bác bỏ giả thuyết H_0 được sử dụng ngày càng rộng rãi. P-value được gọi là mức ý nghĩa quan sát, là xác suất phạm sai lầm loại I tối đa khi bác bỏ giả thuyết H_0 với một tập dữ liệu mẫu đang quan sát. Quy tắc quyết định bác bỏ H_0 theo cách tiếp cận p – value là:

- Nếu $p\text{-value} \geq \alpha$ đã định \rightarrow không bác bỏ H_0
- Nếu $p\text{-value} < \alpha$ đã định \rightarrow bác bỏ H_0

Để hiểu được bản chất của cách tiếp cận p-value chúng ta xem xét lại các ví dụ đã làm ở trên.

Với ví dụ 1: mục tiêu của chúng ta là kiểm định xem trung bình tổng thể bằng hay khác giá trị 368 gram, giá trị kiểm định $z_u = 1,5$ và chúng ta không bác bỏ H_0 vì z_u nhỏ hơn giá trị tới hạn cận trên và lớn hơn giá trị tới hạn cận dưới, hình vẽ cho ta thấy nó rơi vào khu vực không bác bỏ H_0 .

Bây giờ sử dụng phương pháp p-value, với kiểm định hai bên như tình huống này chúng ta sẽ tìm xác suất có được các giá trị kiểm định z_u lêch bằng hoặc hơn 1,5 lần đơn vị độ lệch chuẩn tính từ trung tâm của một phân phối bình thường chuẩn hóa. Điều này có nghĩa là chúng ta cần tính xác suất để $z_u \leq -1,5$ và $z_u \geq 1,5$ vì ta có kiểm định hai bên. Nếu diễn đạt bằng công thức là $P(|z_u| \geq 1,5) = p\text{-value}$.

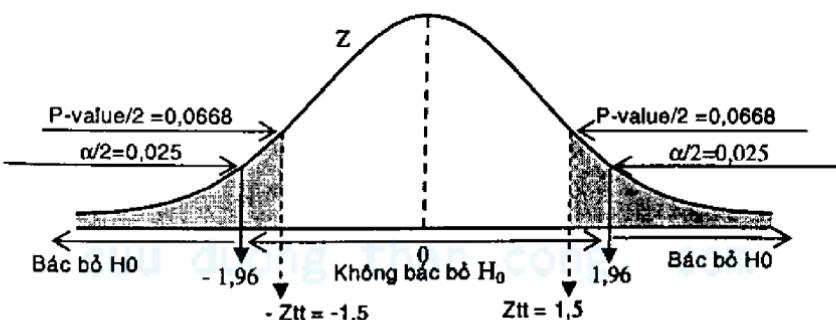
Ta có thể tính riêng $P(z_u \geq 1,5)$ rồi đem kết quả nhân hai lên vì phân phối này đối xứng.

Dùng lệnh Normsdist trên Excel ta có $P(z_u < 1,5) = 0,9332 \rightarrow P(z_u \geq 1,5) = 1 - 0,9332 = 0,0668$

Như vậy $p\text{-value} = P(|z_u| \geq 1,5) = 2 \times 0,0668 = 0,1336$

Kết quả này có nghĩa là xác suất để tính được các giá trị kiểm định lêch bằng hoặc hơn 1,5 lần đơn vị độ lệch chuẩn tính từ trung tâm của một phân phối bình thường chuẩn hóa là 0,13. Vì $p\text{-value} = 0,13 > 0,05$ nên không bác bỏ H_0 . Các bạn xem hình minh họa:

Hình 8.8

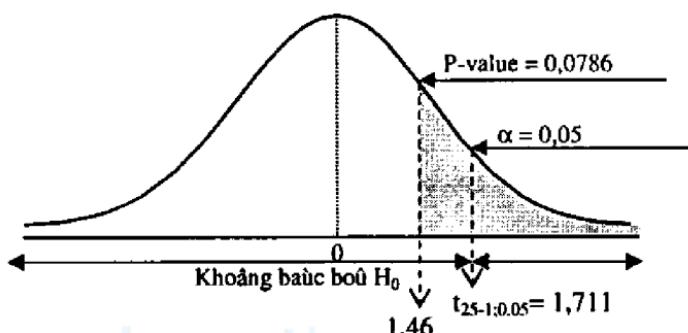


Nhìn trên hình minh họa các bạn dễ cảm nhận hơn rằng nếu giá trị $p\text{-value}/2$ lớn hơn $\alpha/2$ thì giá trị kiểm định tính được từ dữ liệu mẫu tương ứng với giá trị $p\text{-value}$ này phải rơi vào bên trong của giá trị tới hạn tức rơi vào khu vực không bác bỏ H_0 , đó là nguyên nhân khiến ta có quy tắc bác bỏ hay không bác bỏ H_0 theo $p\text{-value}$.

Nếu chúng ta vẫn bác bỏ H_0 với mức ý nghĩa quan sát bằng 0,13 thì điều gì sẽ xảy ra, khi làm bài toán kiểm định chúng ta đã chọn mức ý nghĩa của bài toán là 0,05 tức chúng ta chấp nhận phạm sai lầm loại I tối đa chỉ 5%, giờ đây nếu ta bác bỏ H_0 tại mức ý nghĩa quan sát thì có nghĩa là khả năng phạm sai lầm loại I của chúng ta tới 13%, lớn hơn nhiều so với tối đa mà ta đã chọn nên ta không thể bác bỏ H_0 được vì nguy cơ phạm sai lầm quá lớn.

Với ví dụ 2 ta xác định được giá trị p-value = 0,07 (dùng hàm tdist trên Excel). Vì $p\text{-value} > \alpha \rightarrow$ không bác bỏ H_0 .

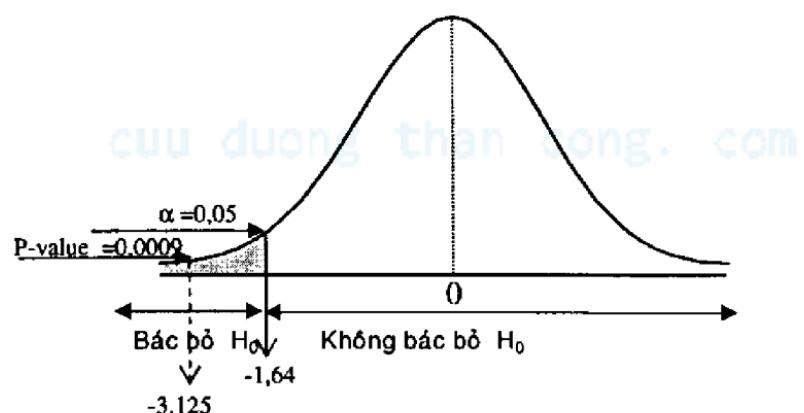
Hình 8.9



Với ví dụ 3 ta xác định được giá trị p-value = 0,0009, vì $p\text{-value} < \alpha$ nên ta bác bỏ H_0 .

Tương tự nếu ta phát biểu kiểm định “có ý nghĩa thống kê ở mức p%” tức là ta đã đi đến bác bỏ H_0 và có thể sai tối đa chỉ p% với kết luận đó, ta tự tin do giá trị p-value < α .

Hình 8.10



Sử dụng phương pháp p-value bạn không cần dùng tới các bảng tra, điều này đặc biệt tiện lợi khi bạn làm việc với phần mềm thống kê, lúc đó các phần mềm sẽ tính toán từ giá trị kiểm định và cung cấp luôn cho bạn giá trị p-value để bạn so sánh với α nhằm quyết định có bác bỏ H_0 hay không.

8.2.2 Kiểm định giả thuyết về tỷ lệ tổng thể

Có những lúc chúng ta cần kiểm định giả thuyết liên quan đến tỷ lệ tổng thể p . Một mẫu ngẫu nhiên có thể được chọn từ tổng thể và giá trị tỷ lệ mẫu $p_s = X/n$ được tính, giá trị mẫu này được đem so sánh với giá trị giả thuyết về tỷ lệ tổng thể p_0 để ra quyết định.

Khi có mẫu đủ lớn (cả np và $n(1-p) \geq 5$), phân phối của tỷ lệ mẫu xấp xỉ phân phối bình thường, để thực hiện kiểm định nhằm đánh giá mức độ khác biệt giữa tỷ lệ mẫu p_s và tỷ lệ tổng thể được giả thuyết p_0 chúng ta sẽ tính toán giá trị kiểm định theo công thức

$$z_n = \frac{p_s - p_0}{\sqrt{p(1-p)/n}}$$

Trong đó

- p_s là tỷ lệ thành công trong mẫu
- p_0 là tỷ lệ thành công giả định cho tổng thể
- vì chúng ta đã giả định $p = p_0$ nên dùng p_0 thay thế vào mẫu số của công thức như một cách tính xấp xỉ

Quyết định bác bỏ hay không bác bỏ H_0 theo quy tắc sau

- Nếu là kiểm định hai bên ta sẽ bác bỏ H_0 khi $z_n < -z_{\alpha/2}$ hoặc $z_n > z_{\alpha/2}$
- Nếu là kiểm định bên trái ta sẽ bác bỏ H_0 khi $z_n < -z_\alpha$
- Nếu là kiểm định bên phải ta sẽ bác bỏ H_0 khi $z_n > z_\alpha$

Ví dụ: Hiện tại cứ trong 10 người uống bia thì chỉ có 1 người thích bia hiệu A. Sau chiến dịch quảng cáo cho loại bia A người ta khảo sát một mẫu ngẫu nhiên 200 người uống bia để kiểm tra hiệu quả của quảng cáo, thì thấy có 26 người thích loại bia A trong 200 người này. Thông tin này đủ để ta kết luận là có sự gia tăng trong tỉ lệ những người thích uống bia hiệu A không, chọn độ tin cậy cho bài toán kiểm định là 90%.

1. Trước khi quảng cáo tỷ lệ tổng thể ưa thích bia A là $1/10=0,1$, sau khi quảng cáo người ta muốn kiểm tra tỷ lệ này có tăng hơn 0,1 hay không
2. Đặt giả thuyết

H_0 : $p = 0,1$

H_1 : $p > 0,1$

Như vậy đây là kiểm định bên phải

3. Độ tin cậy của kiểm định là 90% như vậy mức ý nghĩa $\alpha = 0,1$

4. Cơ mẫu $n = 200$; $p_s = 26/200 = 0,13$; giá trị kiểm định được tính như sau

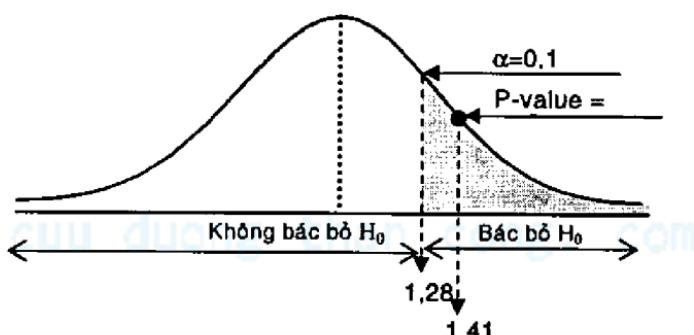
$$z_n = \frac{p_s - p_0}{\sqrt{p(1-p)/n}} = \frac{0,13 - 0,1}{\sqrt{0,1(1-0,1)/200}} = 1,41$$

5. Theo quy tắc kiểm định bên phải ta sẽ bác bỏ H_0 khi $z_u > z_\alpha$

Ta có $\alpha = 0,1 \rightarrow z_\alpha = z_{0,1} = 1,28$

Như vậy $z_u = 1,41 > z_{0,1} = 1,28 \rightarrow$ Bác bỏ giả thuyết H_0

Hình 8.11



6. Với độ tin cậy 90% ta có đủ bằng chứng thống kê để bác bỏ giả thuyết H_0 . Như vậy có thể tin rằng chiến dịch quảng cáo đã làm tăng tỷ lệ người thích uống bia hiệu A.

8.2.3 Kiểm định giả thuyết về phương sai tổng thể

Trong các nội dung đã thảo luận về các công cụ của thống kê suy diễn, chúng ta mới chỉ tập trung vào các tham số tổng thể như trung bình, tỷ lệ mà chưa tìm hiểu công cụ thống kê suy diễn cho một đại lượng mô tả độ phân tán tiêu biểu của tập dữ liệu là độ lệch chuẩn. Tuy nhiên do không có các công thức và phương pháp thống kê phát triển trực tiếp cho tham số thống kê độ lệch chuẩn mà chỉ có của tham số thống kê phương sai, nên ta sẽ chuyển các vấn đề cần kiểm định liên quan đến độ lệch chuẩn thành phương sai bằng phép bình phương đơn giản ta đã biết.

Thủ tục tiến hành kiểm định giả thuyết về phương sai cũng tương tự như thủ tục kiểm định các tham số thống kê khác, việc đầu tiên sau khi xác định được vấn đề cần kiểm định là đặt giả thuyết.

1. Các dạng giả thuyết có thể gấp liên quan đến kiểm định giả thuyết về phương sai tổng thể

Kiểm định giả thuyết hai bên

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 \neq \sigma_0^2$$

Kiểm định giả thuyết bên trái

$$H_0: \sigma^2 \geq \sigma_0^2$$

$$H_1: \sigma^2 < \sigma_0^2$$

Kiểm định giả thuyết bên phải

$$H_0: \sigma^2 \leq \sigma_0^2$$

$$H_1: \sigma^2 > \sigma_0^2$$

2. Chọn độ tin cậy

3 Tính toán giá trị kiểm định. Để kiểm định giả thuyết H_0 cũng như cách làm của các kiểm định khác, chúng ta sẽ so sánh tham số mẫu là s^2 với tham số tổng thể đã giả thuyết là σ_0^2 . Muốn vậy chúng ta phải tính toán giá trị kiểm định theo công thức

$$\chi_{\alpha}^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Trong đó

- χ_{α}^2 là giá trị kiểm định có phân phối Chi Bình phương
- n là cỡ mẫu
- s^2 là phương sai mẫu
- σ_0^2 là phương sai tổng thể đã giả thuyết.

Giá trị kiểm định của chúng ta có phân phối Chi Bình phương, phân phối này cũng bao gồm một họ các phân phối, phụ thuộc vào giá trị bậc tự do liên quan đến bài toán kiểm định, bậc tự do là $df = n-1$.

4. Quyết định bác bỏ hay không bác bỏ giả thuyết H_0 theo quy tắc như sau:

- Nếu là kiểm định bên phải

Giá trị tối hạn tương ứng với mức ý nghĩa α và bậc tự do $df = n-1$ là $\chi_{n-1;\alpha}^2$

Quy tắc là nếu $\chi_n^2 > \chi_{n-1;\alpha}^2 \rightarrow$ bác bỏ giả thuyết H_0

Người ta xây dựng sẵn bảng tra các giá trị tối hạn cho phân phối Chi bình phương tương ứng với một số mức ý nghĩa và bậc tự do cho trước để giúp

việc tìm giá trị tối hạn được nhanh chóng. Trong phụ lục của cuốn sách này nó là Bảng tra số 3

- Nếu là kiểm định bên trái

Giá trị tối hạn tương ứng với mức ý nghĩa α và bậc tự do $df = n-1$ là $\chi^2_{n-1;1-\alpha}$

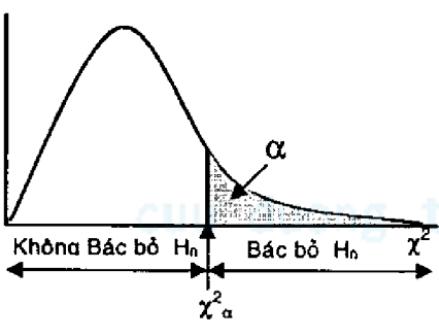
Quy tắc là nếu $\chi^2_n < \chi^2_{n-1;1-\alpha} \rightarrow$ bác bỏ giả thuyết H_0

- Nếu là kiểm định 2 bên

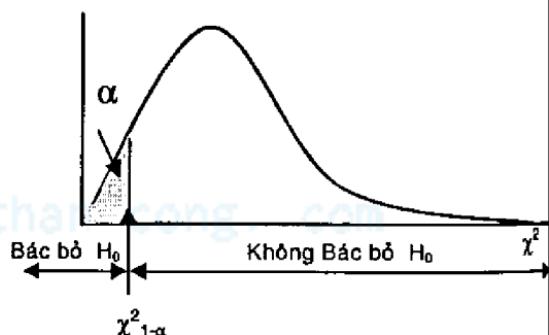
Quy tắc là nếu $\chi^2_n < \chi^2_{n-1;1-\alpha/2}$ hoặc $\chi^2_n > \chi^2_{n-1;\alpha/2} \rightarrow$ bác bỏ giả thuyết H_0

Xem hình minh họa các tình huống ra quyết định.

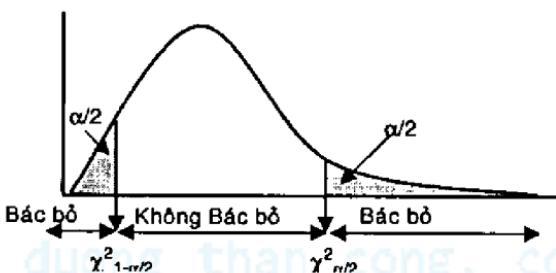
Hình 8.12 Kiểm định bên phải



Hình 8.13 Kiểm định bên trái



Hình 8.14 Kiểm định hai bên



5. Sau khi căn cứ trên các quy tắc để ra quyết định, chúng ta cũng có những kết luận phù hợp về bài toán kiểm định.

Ví dụ: Theo quy định kỹ thuật các tủ lạnh phải giữ được độ lạnh đã cài đặt với mức độ dao động của nhiệt độ là ít nhất thể hiện qua thông số độ lệch chuẩn không quá 4°C . Một mẫu 16 cái tủ lạnh trong kho nhà máy được chọn ngẫu nhiên và qua kiểm tra thấy rằng phương sai của mẫu này là s^2

= 24. Hãy thực hiện kiểm định phù hợp để cho biết độ lệch chuẩn theo yêu cầu kỹ thuật có bị vượt quá không. Chọn $\alpha = 0,05$.

1. Tính huống cần kiểm định là độ lệch chuẩn của các tủ lạnh được sản xuất có đảm bảo theo yêu cầu kỹ thuật không, vì không có các thủ tục kiểm định liên quan đến độ lệch chuẩn nên chúng ta thay thế bằng phương pháp kiểm định phương sai.

2. Đặt giả thuyết

$$H_0: \sigma^2 \leq 16$$

$$H_1: \sigma^2 > 16$$

3. Chọn $\alpha = 0,05$

4. Với $n = 16$, $s^2 = 24$; $\sigma_0^2 = 16$ chúng ta tính toán giá trị kiểm định như sau

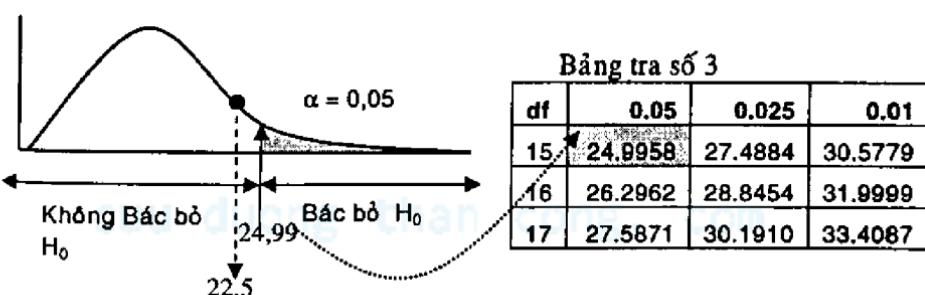
$$\chi_{n-2}^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(16-1)24}{16} = 22,5$$

5. Ta có $\alpha = 0,05$ và $df = 16 - 1 = 15 \rightarrow \chi_{\alpha}^2 = \chi_{15;0,05}^2 = 24,99$

Theo quy tắc là nếu $\chi_{n-2}^2 > \chi_{n-1;\alpha}^2$ thì bác bỏ giả thuyết H_0

Nhưng ta có $\chi_{n-2}^2 = 22,5 < \chi_{15;0,05}^2 = 24,99$ nên không bác bỏ H_0

Hình 8.15



6. Kết luận

Với độ tin cậy 95%, không có đủ bằng chứng thống kê để kết luận rằng độ lệch chuẩn theo yêu cầu kỹ thuật của các tủ lạnh đã bị vượt quá.

8.3 KIỂM ĐỊNH GIẢ THUYẾT HAI MẪU

Trong các tình huống thực tế, phải lựa chọn giữa một trong hai phương án, so sánh giữa hai đối tượng nào đó, là một vấn đề hay gấp, từ người nông dân phải lựa chọn giữa hai loại giống cây trồng cho đến giám đốc của một doanh nghiệp lựa chọn mua chiến dịch quảng cáo của công ty nào trong hai công ty quảng cáo, phụ nữ có học vấn cao và phụ nữ có học vấn trung bình có thái độ giống nhau về vấn đề giáo dục giới tính cho con của họ không. Sự khó khăn của các quyết định lựa chọn này là ở chỗ người ra quyết định chỉ có các thông tin giới hạn, lúc này công cụ kiểm định thống kê giúp người ra quyết định so sánh lựa chọn giữa hai phương án, mà bản chất sự so sánh này là so sánh có tính suy diễn về tham số của 2 tổng thể, ví dụ đánh giá năng suất trung bình của giống cây trồng thứ nhất và năng suất trung bình của giống cây trồng thứ 2, năng suất loại nào lớn hơn, hay khác biệt giữa hai năng suất là bằng 0. Trong nội dung phần này chúng ta sẽ tìm hiểu phương pháp giúp so sánh được 2 tham số của 2 tổng thể.

8.3.1 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể

Giả sử chúng ta có một mẫu cỡ n_1 lấy một cách ngẫu nhiên từ tổng thể thứ nhất, và một mẫu ngẫu nhiên cỡ n_2 rút từ tổng thể thứ 2, dữ liệu thu thập vào hai mẫu đều là dữ liệu định lượng, trong tổng thể thứ nhất trung bình là μ_1 , và độ lệch chuẩn là σ_1 ; trong tổng thể thứ 2 trung bình là μ_2 , và độ lệch chuẩn là σ_2 .

Khi làm kiểm định giả thuyết khác biệt về trung bình của hai tổng thể chúng ta phải phân biệt hai tình huống chính là : 2 mẫu để thực hiện kiểm định là độc lập và 2 mẫu để thực hiện kiểm định là không độc lập. Khái niệm 2 mẫu độc lập với nhau là các mẫu được chọn từ tổng thể theo cách thức sao cho việc một quan sát được chọn vào mẫu này không ảnh hưởng gì đến xác suất một quan sát được chọn vào mẫu kia. Ví dụ hai mẫu sinh viên, một mẫu nam được chọn từ tổng thể sinh viên nam, và một mẫu nữ được chọn từ tổng thể sinh viên nữ, là hai mẫu độc lập; nhưng nếu khi chọn mẫu sinh viên nữ lại theo điều kiện sinh viên đó phải là chị hay em ruột là một sinh viên nam trong mẫu kia thì hai mẫu này như vậy là không độc lập với nhau.

Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể đòi hỏi một số giả định căn bản sau đây

- Mẫu được lấy ngẫu nhiên và độc lập
- Tổng thể có phân phối bình thường
- Hai tổng thể có phương sai như nhau

Thủ tục kiểm định chúng ta sẽ tìm hiểu sau đây sẽ có hiệu lực khi hội đủ các giả định trên.

Thủ tục thực hiện kiểm định về trung bình của hai tổng thể có trình tự tiến hành không khác với kiểm định giả thuyết về tham số đơn lẻ, cũng đi qua các bước

1. Xác định giả thuyết tổng thể quan tâm

2. Xây dựng giả thuyết H_0 và giả thuyết đối H_1 , với loại kiểm định này chúng ta có thể gặp các tình huống kết hợp của giả thuyết sau đây

Kiểm định giả thuyết hai bên

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Tương đương

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Kiểm định giả thuyết bên phải

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Tương đương

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

Kiểm định giả thuyết bên trái

$$H_0: \mu_1 \geq \mu_2$$

$$H_1: \mu_1 < \mu_2$$

Tương đương

$$H_0: \mu_1 - \mu_2 \geq 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

3. Xác định mức ý nghĩa cho kiểm định

4. Tính toán giá trị kiểm định trên các thông tin mẫu của hai mẫu

5. Quyết định bác bỏ hoặc không bác bỏ H_0 theo các quy tắc xác định, chúng ta cũng có thể dùng cách tiếp cận p-value cho bước này

6. Rút ra kết luận.

Đầu tiên, chúng ta nghiên cứu kiểm định trung bình cho hai mẫu độc lập với hai tình huống cụ thể là khi đã biết phương sai tổng thể và khi chưa biết phương sai tổng thể. Sau đó chúng ta nghiên cứu tình huống kiểm định trung bình cho hai mẫu phối hợp từng cặp.

8.3.1.1 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể, biết phương sai của hai tổng thể, hai mẫu độc lập

Tính toán giá trị kiểm định theo công thức

$$z_u = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Trong đó

\bar{x}_1 là trung bình mẫu của mẫu rút từ tổng thể thứ nhất

\bar{x}_2 là trung bình mẫu của mẫu rút từ tổng thể thứ 2

μ_1, μ_2 là trị trung bình tổng thể được giả thuyết của tổng thể thứ nhất và tổng thể thứ 2

σ_1^2, σ_2^2 là phương sai tổng thể thứ nhất và tổng thể thứ 2

n_1 và n_2 là cỡ mẫu của mẫu thứ nhất và mẫu thứ 2

Quy tắc bác bỏ H_0

- Nếu chúng ta có kiểm định hai bên, ta sẽ bác bỏ H_0 khi $z_u < -z_{\alpha/2}$ hoặc $z_u > z_{\alpha/2}$
- Nếu chúng ta có kiểm định bên phải, ta sẽ bác bỏ H_0 khi $z_u > z_\alpha$
- Nếu chúng ta có kiểm định bên trái, ta sẽ bác bỏ H_0 khi $z_u < -z_\alpha$

8.3.1.2 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể, không biết phương sai của hai tổng thể, hai mẫu độc lập cỡ mẫu lớn

Trong phần lớn các kiểm định phương sai tổng thể là cái ta không biết, khi đó nếu cỡ của cả hai mẫu đều ≥ 30 thì tiến trình kiểm định của ta được điều chỉnh theo công thức sau

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Ví dụ: Một hãng viễn thông tiến hành kiểm định có hay không sự khác biệt trong thu nhập trung bình của các khách hàng nam và nữ. Họ chọn mẫu ngẫu nhiên 200 khách nam và 100 khách nữ rồi thực hiện các bước kiểm định sau:

Bước 1: xác định giá trị tổng thể quan tâm

Thu nhập trung bình tổng thể nam kí hiệu μ_1

Thu nhập trung bình tổng thể nữ kí hiệu μ_2

Chúng ta muốn xác định có sự khác biệt hay không trong hai giá trị trung bình tổng thể này

Bước 2 : xây dựng giả thuyết của kiểm định

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Bước 3 : xác định mức ý nghĩa của kiểm định là 0,05

Bước 4 : xác định vùng bác bỏ giả thuyết

Mặc dù không biết phương sai tổng thể nhưng do 2 cỡ mẫu đều trên 30 nên ta sử dụng giá trị kiểm định z. Với kiểm định hai đuôi, $\alpha = 0,05$ t tra bảng phân phối bình thường chuẩn hóa tìm ra hai giá trị tới hạn cho kiểm định hai đuôi là: $-z_{0,025} = -1,96$ và $z_{0,025} = 1,96$. Quy tắc quyết định cho bài toán kiểm định của ta là nếu giá trị kiểm định tính được $< -1,96$ hoặc $> 1,96$ ta sẽ bác bỏ H_0 . Ngược lại ta không bác bỏ H_0

Bước 5 : trên hai mẫu ngẫu nhiên khách hàng chọn được công ty thu thập dữ liệu và tính được các giá trị thống kê mẫu như sau

\bar{x}_1 là thu nhập trung bình của mẫu khách hàng nam, bằng 43390\$

\bar{x}_2 là thu nhập trung bình của mẫu khách hàng nữ, bằng 42400\$

s_1 là độ lệch chuẩn thu nhập của mẫu khách hàng nam, bằng 7300\$

s_2 là độ lệch chuẩn thu nhập của mẫu khách hàng nữ, bằng 8200\$

Bước 6 : Tính giá trị kiểm định z theo công thức

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(43390 - 42400) - 0}{\sqrt{\frac{7300^2}{200} + \frac{8200^2}{100}}} = 1,022$$

Bước 7 : vì $-1,96 < z = 1,022 < 1,96$ nên ta không bác bỏ H_0

Bước 8 : với độ tin cậy 95% ta rút ra kết luận là không có đủ bằng chứng thống kê để nói rằng thu nhập trung bình của khách nam khác thu nhập trung bình của khách nữ.

8.3.1.3 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể, không biết phương sai của hai tổng thể, hai mẫu độc lập cỡ mẫu nhỏ

Trong tình huống không biết phương sai của hai tổng thể và một trong hai cỡ mẫu hoặc cả hai mẫu đều nhỏ hơn 30 quan sát, nếu giả định mẫu được

lấy ngẫu nhiên độc lập từ tổng thể có phân phối bình thường và phương sai bằng nhau vẫn đáp ứng thì chúng ta không sử dụng giá trị kiểm định Z , mà thay bằng giá trị kiểm định t . Lúc này ta không biết 2 phương sai của hai tổng thể nên ta thay thế bằng hai phương sai mẫu, mà giả định của chúng ta là hai tổng thể có phương sai như nhau nên tốt nhất là ta phối hợp 2 phương sai của hai mẫu theo phương pháp trung bình có trọng số để có một phương sai gộp kí hiệu s_p^2 có công thức như sau

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Sau đó giá trị thống kê kiểm định t với $(n_1+n_2 - 2)$ bậc tự do được tính theo công thức

$$t_u = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Trong đó:

\bar{x}_1 là trung bình mẫu của mẫu rút từ tổng thể thứ nhất

\bar{x}_2 là trung bình mẫu của mẫu rút từ tổng thể thứ hai

$\mu_1; \mu_2$ là trị trung bình tổng thể được giả thuyết của tổng thể thứ nhất và tổng thể thứ hai.

$s_1^2; s_2^2$ là phương sai mẫu thứ nhất và mẫu thứ 2

s_p là độ lệch chuẩn gộp

n_1 và n_2 là cỡ mẫu của mẫu thứ nhất và mẫu thứ 2

Quy tắc bác bỏ H_0

- Nếu chúng ta có kiểm định hai bên, ta sẽ bác bỏ H_0 khi

$$t_u < -t_{n_1+n_2-2, \alpha/2} \text{ hoặc } t_u > t_{n_1+n_2-2, \alpha/2}$$

- Nếu chúng ta có kiểm định bên phải, ta sẽ bác bỏ H_0 khi $t_u > t_{n_1+n_2-2, \alpha}$

- Nếu chúng ta có kiểm định bên trái, ta sẽ bác bỏ H_0 khi $t_u < -t_{n_1+n_2-2, \alpha}$

Ví dụ để so sánh độ bền của hai loại sơn phản quang dùng để vẽ các kí hiệu hướng dẫn giao thông trên đường người ta kẻ 12 lần sơn mỗi loại trên một đoạn đường có nhiều xe lưu thông, thứ tự sơn được chọn một cách ngẫu nhiên, sau một thời gian người ta dùng máy đo cường độ phản chiếu của các lần sơn (chỉ số đọc càng cao thì cường độ phản chiếu càng lớn) và ghi lại được các số liệu sau đây.

Sơn A	12,5	11,7	9,9	9,6	10,3	9,6	9,4	11,3	8,7	11,5	10,6	9,7
Sơn B	9,4	11,6	9,7	10,4	6,9	7,3	8,4	7,2	7,0	8,2	12,7	9,2

Người ta cho rằng loại sơn A bền hơn loại sơn B, hãy kiểm định thông tin này với độ tin cậy 90%

Tính toán các tham số mẫu như sau

$$\text{Sơn A } \bar{x}_A = 10,4; s_A^2 = 1,28$$

$$\text{Sơn B } \bar{x}_B = 9,0; s_B^2 = 3,513$$

Giả định phương sai hai tổng thể bằng nhau, tính phương sai gộp

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(12 - 1)1,28 + (12 - 1)3,513}{12 - 1 + 12 - 1} = 2,396$$

$$s_p = \sqrt{2,396} = 1,55$$

1. Người ta đang muốn kiểm tra độ bền của hai loại sơn, sau một thời gian sơn trên đường, nếu loại sơn nào cho chỉ số đọc được đo bằng máy càng lớn thì loại sơn đó phản chiếu càng tốt chứng tỏ nó bền hơn. Chúng ta sẽ tiến hành kiểm định giả thuyết về số đọc trung bình của hai loại sơn với nghi ngờ loại sơn A bền hơn sơn B tức số đọc trung bình của sơn A cao hơn.

2. Giả thuyết đặt ra cho kiểm định này là

$$H_0: \mu_A - \mu_B = 0$$

$$H_1: \mu_A - \mu_B > 0$$

Như vậy đây là kiểm định bên phải

3. Độ tin cậy là 90% tức là mức ý nghĩa $\alpha = 0,1$

4. Tính toán giá trị kiểm định

$$t_{\text{II}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(10,4 - 9) - (0)}{1,55 \sqrt{\frac{1}{12} + \frac{1}{12}}} = 2,2$$

5. Ta có $\alpha = 0,1$; $n_1 + n_2 - 2 = 22$, đây là kiểm định bên phải nên ta sẽ bác bỏ H_0 khi $t_{\text{II}} > t_{n_1+n_2-2, \alpha}$. Theo Bảng tra số 2 thì $t_{22, 0,1} = 1,321$. Như vậy $t_{\text{II}} = 2,2 > t_{22, 0,1} = 1,321$ nên bác bỏ H_0 .

6. Với độ tin cậy 90% ta có đủ bằng chứng thống kê để kết luận rằng loại sơn A bền hơn loại sơn B.

8.3.1.3 Vấn đề với các giả định

Để kiểm tra giả định phân phối bình thường chúng ta có thể vẽ biểu đồ hộp và râu của dữ liệu để đánh giá. Nếu giả định phân phối bình thường

không được đáp ứng chúng ta sẽ dùng kiểm định phi tham số để thay thế cho kiểm định t hoặc z về trung bình của hai tổng thể.

Trong kiểm định giả thuyết về sự bằng nhau của hai trung bình tổng thể tuy là tình huống không biết phương sai tổng thể nhưng chúng ta giả định rằng hai phương sai tổng thể như nhau. Để kiểm tra sự bằng nhau của 2 phương sai tổng thể chúng ta áp dụng phương pháp kiểm định sẽ nghiên cứu ở mục 8.3.3. Trong các tình huống mẫu nhỏ mà phương sai tổng thể không như nhau thì chúng ta phải điều chỉnh công thức t theo cách sau

$$t_n = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Trong công thức này cả hai phương sai mẫu đều tham gia, vì thế nó được gọi tên là kiểm định t với phương sai phân biệt

Bậc tự do v cho đại lượng kiểm định t trường hợp này được tính theo công thức

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$$

Quy tắc bác bỏ H_0 trong tình huống này

- Nếu chúng ta có kiểm định hai bên, ta sẽ bác bỏ H_0 khi $t_u < -t_{v,\alpha/2}$ hoặc $t_u > t_{v,\alpha/2}$
- Nếu chúng ta có kiểm định bên phải, ta sẽ bác bỏ H_0 khi $t_u > t_{v,\alpha}$
- Nếu chúng ta có kiểm định bên trái, ta sẽ bác bỏ H_0 khi $t_u < -t_{v,\alpha}$

8.3.1.4 Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể, hai mẫu không độc lập (mẫu phôi hợp từng cặp hay mẫu cặp)

Trong nghiên cứu không phải lúc nào ta cũng chỉ dùng những mẫu độc lập mà rất nhiều khi ta phải kiểm nghiệm giả thuyết với những mẫu phụ thuộc hay có liên hệ với nhau tức việc lựa các quan sát vào mẫu này có liên hệ hoặc có ảnh hưởng đến việc lựa các quan sát vào mẫu kia. Ví dụ khi muốn so sánh hiệu quả của hai phương pháp học tập người ta có thể chọn 2 mẫu theo kiểu phôi hợp từng cặp, một đối tượng ở mẫu bên này

được chọn sao cho tương ứng với một đối tượng đã chọn ở mẫu bên kia về giới tính, tuổi, điểm kết quả học tập, chỉ số IQ. Sau một thời gian giảng dạy hai nhóm theo hai phương pháp ta sẽ so sánh kết quả thi của hai mẫu. Mục đích của việc lấy mẫu phối hợp từng cặp như vậy là để loại trừ ảnh hưởng của các yếu tố ngoại cảnh ra khỏi so sánh, bằng cách đó ta sẽ tin chắc hơn rằng những khác biệt trong kết quả thi ta thấy giữa hai mẫu là do tác động thực sự của phương pháp giảng dạy chứ không phải do các nguyên nhân khác.

Để hiểu được cách tiến hành của phương pháp kiểm định này ta xuất phát từ cấu trúc dữ liệu dùng cho kiểm định

Giá trị quan sát trên mẫu thứ nhất và mẫu thứ hai từng cặp tương ứng được ghi lại theo cột như sau

Bảng 8.1

Quan sát	Nhóm 1	Nhóm 2	Chênh lệch
1	X_{11}	X_{21}	$d_1 = X_{11} - X_{21}$
2	X_{12}	X_{22}	$d_2 = X_{12} - X_{22}$
3	X_{13}	X_{23}	$d_3 = X_{13} - X_{23}$
...			
n	X_{1n}	X_{2n}	$d_n = X_{1n} - X_{2n}$

Gọi \bar{d} là trung bình của các chênh lệch được giả thuyết, ta có thể gấp các dạng giả thuyết sau đây trong kiểm định trung bình mẫu từng cặp

Giả thuyết hai bên

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

Giả thuyết bên phải

$$H_0: \mu_d \leq 0$$

$$H_1: \mu_d > 0$$

Giả thuyết bên trái

$$H_0: \mu_d \geq 0$$

$$H_1: \mu_d < 0$$

Giá trị kiểm định được tính theo công thức sau đây

$$z_u = \frac{\bar{d} - \mu_d}{\frac{\sigma_d}{\sqrt{n}}}$$

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

Trong đó:

μ_d là trung bình của các chênh lệch được giả thuyết
 σ_d là độ lệch chuẩn tổng thể của các chênh lệch
 n là cỡ mẫu, hai mẫu phải có cỡ bằng nhau

Tuy nhiên tình huống biết σ_d hầu như không bao giờ xảy ra trong thực tế, lúc đó nếu mẫu được chọn từ một tổng thể có phân phối bình thường thì chúng ta có thể thay thế giá trị kiểm định z bằng t , trị kiểm định t có bậc tự do bằng $n-1$

$$t_n = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

Trong đó

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

Quy tắc quyết định bác bỏ H_0

- Nếu là kiểm định hai bên, chúng ta bác bỏ H_0 khi $t_u < -t_{n-1,\alpha/2}$ hoặc $t_u > t_{n-1,\alpha/2}$
- Nếu là kiểm định bên phải, chúng ta bác bỏ H_0 khi $t_u > t_{n-1,\alpha}$
- Nếu là kiểm định bên trái, chúng ta bác bỏ H_0 khi $t_u < -t_{n-1,\alpha}$

Ví dụ: để thiết kế một thử nghiệm nhằm so sánh tốc độ xử lý của hai phần mềm thống kê, một phần mềm mới và một phần mềm hiện đang sử dụng, người ta làm như sau.

Chọn một bộ dữ liệu, đặt ra 10 yêu cầu cần xử lý trên bộ dữ liệu này.

Cài hai phần mềm lên 1 máy tính, chép bộ dữ liệu này lên máy ở hai vị trí khác nhau, lần lượt cho phần mềm thứ nhất xử lý từng lệnh một trên file dữ liệu thứ nhất rồi đến phần mềm thứ 2 xử lý file thứ 2 theo thứ tự lệnh y như nhau, ghi lại thời gian xử lý từng lệnh của từng phần mềm. ta được bảng dữ liệu sau

Bảng 8.2

Lệnh	Thời gian xử lý (giây)		Chênh lệch	$(d_i - \bar{d})^2$
	Phần mềm đang dùng	Phần mềm mới		
1	9,98	9,88	0,10	0,0003
2	9,88	9,86	0,02	0,0041
3	9,84	9,75	0,09	0,0000
4	9,99	9,8	0,19	0,0112
5	9,94	9,87	0,07	0,0002
6	9,84	9,84	0,00	0,0071
7	9,86	9,87	-0,01	0,0088
8	10,12	9,86	0,26	0,0310
9	9,90	9,83	0,07	0,0002
10	9,91	9,86	0,05	0,0012
Tổng			0,84	0,0641

Câu hỏi đặt ra là phần mềm mới có xử lý nhanh hơn phần mềm cũ hay không. Kiểm định thông tin này với độ tin cậy 95%.

Tính toán:

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{0,84}{10} = 0,084$$

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{0,0641}{10-1}} = 0,0844$$

1. Cho rằng tốc độ xử lý của phần mềm hiện đang dùng bằng tốc độ của phần mềm mới. Chúng ta muốn kiểm định để chứng minh tốc độ xử lý của phần mềm mới nhanh hơn.

2. Ta đặt giả thuyết như sau

$$H_0: \mu_d = 0$$

$$H_1: \mu_d > 0$$

Như vậy đây là kiểm định bên phải.

3. Độ tin cậy của kiểm định là 95% như vậy mức ý nghĩa là 5%

4. Tính toán giá trị kiểm định

$$t_n = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{0,084 - 0}{\frac{0,0844}{\sqrt{10}}} = 3,15$$

5. $\alpha = 0,05$, $n = 10 \rightarrow t_{n-1; \alpha} = t_{9; 0,05} = 1,8331$

Đây là kiểm định bên phải, chúng ta bác bỏ H_0 khi $t_u > t_{n-1; \alpha}$

Ta có $t_u = 3,15 > t_{9; 0,05} = 1,8331 \rightarrow$ bác bỏ H_0

6. Với độ tin cậy 95% có đủ bằng chứng thống kê để kết luận là tốc độ xử lý của phần mềm mới nhanh hơn phần mềm hiện đang dùng.

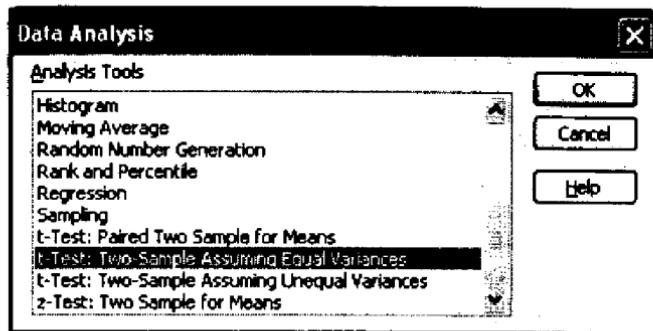
8.3.1.5 Cách thực hiện bằng Excel

Chương trình Excel cung cấp cho chúng ta hàng loạt các lệnh liên quan đến kiểm định trung bình hai mẫu, các lệnh này cũng nằm trong menu Tool/Data Analysis. Khi vào cửa sổ Data Analysis, bạn kéo thanh cuộn đến cuối danh sách các lệnh, bạn sẽ thấy có hàng loạt lựa chọn như sau:

- t-test: Paired two sample for Means Kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể trong tình huống hai mẫu không độc lập (mẫu phối hợp từng cặp)
- t-test: Two sample assuming equal variances dùng cho kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể trong tình huống không biết phương sai của hai tổng thể, hai mẫu độc lập, nếu hai tổng thể có phương sai bằng nhau, trong tình huống phương sai hai tổng thể không bằng nhau dùng lệnh t-test: Two sample assuming unequal variances
- Z test: Two sample for Means dùng cho kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể trong tình huống đã biết phương sai của hai tổng thể, hai mẫu độc lập.

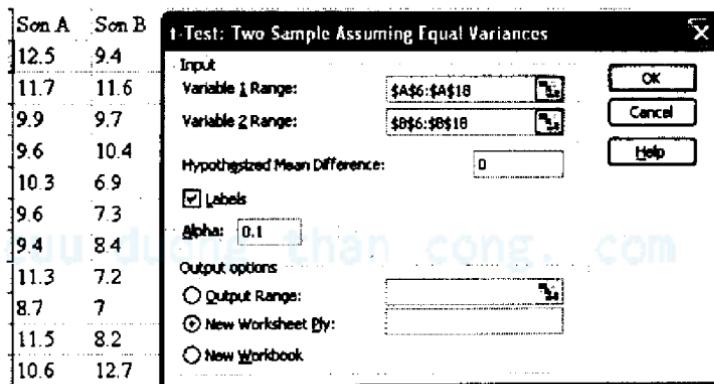
Chúng ta sẽ thực hiện thử một ví dụ, dùng bộ dữ liệu về mức độ phản chiếu của các lăng son ở Mục 8.3.1.2 đây là tình huống kiểm định giả thuyết cho khác biệt của hai trung bình tổng thể trong tình huống không biết phương sai của hai tổng thể, hai mẫu độc lập, cho rằng hai tổng thể có phương sai bằng nhau nên ta dùng t-test: Two sample assuming equal variances. Bấm chọn sáng lệnh này như trong Hình 8.16 để mở cửa sổ t-test: Two sample assuming equal variances.

Hình 8.16



Trong cửa sổ này thực hiện các khai báo như sau, rồi nhấp nút OK chúng ta có kết quả ở Bảng 8.3.

Hình 8.17



Bảng 8.3

	Son A	Son B
Mean	10.400	9.000
Variance	1.280	3.513
Observations	12.000	12.000
Pooled Variance	2.396	
Hypothesized Mean Difference	0.000	
df	22.000	
t Stat	2.215	
P(T<=t) one-tail	0.019	
t Critical one-tail	1.321	
P(T<=t) two-tail	0.037	
t Critical two-tail	1.717	

Trong ví dụ tính thủ công của chúng ta với giả thuyết

$$H_0: \mu_A - \mu_B = 0$$

$$H_1: \mu_A - \mu_B > 0$$

chúng ta đã đi đến kết cục bác bỏ H_0 , ở trong bảng kết quả này ngoài những tham số mẫu ta đã tính toán, có một giá trị mà Excel cung cấp cho chúng ta là giá trị p-value, chúng ta có kiểm định 1 bên nên chúng ta dùng $P(T < t)$ one-tail = 0,019 $< \alpha = 0,1 \rightarrow$ bác bỏ H_0 .

Nếu thực hiện kiểm định hai bên chúng ta sẽ dùng giá trị p-value tại hàng $P(T \leq t)$ two-tail và chú ý là ta vẫn so với α chứ không phải $\alpha/2$.

8.3.2 Kiểm định giả thuyết khác biệt giữa hai tỷ lệ tổng thể

Có hai phương pháp để kiểm định sự khác biệt của hai tỷ lệ tổng thể một phương pháp dùng giá trị kiểm định z với phân phối xấp xỉ bình thường chuẩn hóa và phương pháp thứ hai dùng kiểm định Chi bình phương với 1 bậc tự do, kết quả của hai phương pháp là như nhau

8.3.2.1 Phương pháp dùng phân phối z

Chúng ta có hai mẫu độc lập, một mẫu cỡ n_1 lấy từ một phân phối nhị thức với tỷ lệ thành p_{s1} và một mẫu cỡ n_2 lấy từ một phân phối nhị thức với tỷ lệ thành p_{s2} . Gọi X_1 và X_2 là số lượt thành quan sát được trong hai mẫu nói trên, ta có

$$p_{s1} = X_1/n_1 \text{ và } p_{s2} = X_2/n_2$$

Giả sử chúng ta muốn kiểm định giả thuyết liên quan đến hai tỷ lệ tổng thể p_1 của mẫu thứ nhất và p_2 của mẫu thứ hai, có các dạng giả thuyết sau:

Giả thuyết hai bên

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

Tương đương

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 \neq 0$$

Giả thuyết bên phải

$$H_0: p_1 \leq p_2$$

$$H_1: p_1 > p_2$$

Tương đương

$$H_0: p_1 - p_2 \leq 0$$

$$H_1: p_1 - p_2 > 0$$

Giả thuyết bên trái

$$H_0: p_1 \geq p_2$$

$$H_1: p_1 < p_2$$

Tương đương

$$H_0: p_1 - p_2 \geq 0$$

$$H_1: p_1 - p_2 < 0$$

Chúng ta biết rằng với một mẫu có cỡ n và tỷ lệ thành công p, nếu $np \geq 5$ và $n(1-p) \geq 5$ thì phân phối của tỉ lệ mẫu xấp xỉ phân phối bình thường với trung bình của phân phối bằng p và phương sai của phân phối bằng $p(1-p)/n$.

Như vậy trong trường hợp chúng ta làm việc đồng thời trên hai mẫu thì $(p_{s1} - p_{s2})$ cũng phân phối xấp xỉ bình thường nếu

$$n_1 p_{s1} \geq 5 \text{ và } n_1(1-p_{s1}) \geq 5$$

$$n_2 p_{s2} \geq 5 \text{ và } n_2(1-p_{s2}) \geq 5$$

Chênh lệch trên hai tỷ lệ mẫu $(p_{s1} - p_{s2})$ là ước lượng không chênh của chênh lệch trên hai tỷ lệ tổng thể $(p_1 - p_2)$.

Phương sai của phân phối mẫu của $(p_{s1} - p_{s2})$ được tính theo công thức

$$s_{p_{s1}-p_{s2}}^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = \frac{p_s(1-p_s)}{n_1} + \frac{p_s(1-p_s)}{n_2} = p_s(1-p_s)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Cách tính p_s : Khi kiểm định giả thuyết về hai tỷ lệ tổng thể, ví dụ $H_0: p_1 = p_2$, giả thuyết này cho rằng hai tỷ lệ của tổng thể bằng nhau nhưng không xác định trị số chung của hai tỷ lệ này do đó chúng ta đặt $p_1 = p_2 = p$. Nếu vậy thì theo nguyên tắc p_{s1} và p_{s2} đều là ước lượng không chênh của p, tuy nhiên chúng ta sẽ được ước lượng tốt nhất cho p là p_s , nếu chúng

ta nhập chung hai mẫu lại để có $p_s = \frac{n_1 p_{s1} + n_2 p_{s2}}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$ là ước lượng tốt nhất cho trị số chung p.

Bây giờ chúng ta dùng biến số z cho kiểm định giả thuyết về hai tỷ lệ tổng thể

$$z_n = \frac{(p_{s1} - p_{s2}) - (p_1 - p_2)}{\sqrt{p_s(1-p_s)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Trong đó:

- p_{s1} và p_{s2} là tỷ lệ của mẫu thứ nhất và mẫu thứ 2
- p_s là ước lượng chung cho tỷ lệ của hai mẫu kết hợp
- $p_1 - p_2$ là chênh lệch giữa hai giá trị tỷ lệ tổng thể được giả thuyết
- n_1 và n_2 là cỡ mẫu của mẫu thứ nhất và mẫu thứ 2

Quy tắc bác bỏ hay chấp nhận H_0

- Với kiểm định giả thuyết hai bên, ta bác bỏ H_0 khi $z_n < -z_{\alpha/2}$ hoặc $z_n > z_{\alpha/2}$
- Với kiểm định giả thuyết bên phải, ta bác bỏ H_0 khi $z_n > z_\alpha$
- Với kiểm định giả thuyết bên trái, ta bác bỏ H_0 khi $z_n < -z_\alpha$

Ví dụ: Để kiểm định hiệu quả của một loại thuốc ngừa bệnh người ta tiêm thuốc này lên 150 con vật thí nghiệm, ngoài ra cũng có 150 con vật thuộc nhóm thứ 2 không được tiêm thuốc, sau đó 300 con vật này được gieo bệnh để nghiên cứu, trong nhóm đã tiêm ngừa có 10 chết, trong nhóm không tiêm ngừa có 30 con chết. Như vậy tiêm ngừa có làm giảm tỷ lệ chết vì bệnh không? Chọn $\alpha = 0,01$

1. Gọi p_1 là tỷ lệ chết của các con vật không được tiêm và p_2 là tỷ lệ chết của các con vật được tiêm.

2. Giả thuyết đặt ra cho bài toán kiểm định là

$$H_0: p_1 = p_2$$

$$H_1: p_1 > p_2$$

Tỷ lệ chết trên mẫu được tính :

$$p_{s1} = 30/150 = 0,2$$

$$p_{s2} = 10/150 = 0,067$$

Tỷ lệ chung được tính là $p_s = \frac{X_1 + X_2}{n_1 + n_2} = \frac{10 + 30}{150 + 150} = 0,13$

3. Chọn mức ý nghĩa $\alpha = 1\%$

4. Tính giá trị kiểm định

$$z_{\alpha} = \frac{p_{s1} - p_{s2} - p_1 - p_2}{\sqrt{p_s(1-p_s)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0,2 - 0,067 - 0}{\sqrt{0,13(1-0,13)\left(\frac{1}{150} + \frac{1}{150}\right)}} = 3,42$$

5. $\alpha = 0,01$ đây là kiểm định bên phải, ta bác bỏ H_0 khi $z_{\alpha} > z_{\alpha}$

Theo Bảng tra số 1 thì $z_{0,01} = 2,326 \rightarrow z_{\alpha} = 3,42 > z_{\alpha} = 2,326 \rightarrow$ bác bỏ H_0

6. Với độ tin cậy 99% ta có đủ bằng chứng thống kê để kết luận rằng tiêm ngừa làm giảm tỷ lệ chết vì bệnh của các con vật.

8.3.2.2 Phương pháp dùng phân phối Chi Bình phương

Thay vì dùng kiểm định z cho hai tỷ lệ tổng thể thông qua việc so sánh trực tiếp hai tỷ lệ ta có thể dùng kiểm định Chi Bình phương bằng cách khảo sát tần số thành công và không thành công trong hai nhóm qua một bảng 2 hàng 2 cột thường được gọi tên là bảng chéo có cấu trúc như sau

Bảng 8.4

		Biến trên cột		
Biến trên hàng		Nhóm 1	Nhóm 2	Tổng
Thành công	X_1	X_2	X	
Không thành công	$n_1 - X_1$	$n_2 - X_2$	$n - X$	
Tổng	n_1	n_2	n	

Trong đó:

- X_1 là số thành công trong nhóm 1
- X_2 là số thành công trong nhóm 2
- $n_1 - X_1$ là số không thành công trong nhóm 1
- $n_2 - X_2$ là số không thành công trong nhóm 2
- $X = X_1 + X_2$ là tổng số thành công
- $n - X$ là tổng số không thành công
- n_1 là cỡ mẫu theo nhóm 1
- n_2 là cỡ mẫu theo nhóm 2

Để minh họa phương pháp chúng ta xem ví dụ sau. Một công ty sở hữu hai khu nghỉ dưỡng trên một hòn đảo du lịch đã tiến hành một cuộc khảo sát sự hài lòng của khách hàng sau khi họ nghỉ tại đây, trong bản câu hỏi điều tra có một câu hỏi về việc khách hàng có dự định quay lại đây một lần nữa không, có 163 trên 227 khách được hỏi của khu nghỉ A trả lời có, và 154 trên 262 khách được hỏi của khu B trả lời như vậy. Với mức ý nghĩa 5% có bằng chứng thống kê nào về sự khác biệt trong mức độ hài lòng của khách hàng (đo lường bằng ý định họ sẽ quay trở lại) tại hai khu nghỉ A và B không.

Bảng 8.5

Mô tả các tần số thực tế	Khu nghỉ		
Dự định quay lại	A	B	Tổng
Có	163	154	317
Không	64	108	172
Tổng	227	262	489

Trong bảng này thành công chính là việc khách hàng trả lời có dự định quay lại và không thành công là khi họ nói không quay lại, Biến trên cột là tên của hai khu nghỉ. Tổng theo hàng thể hiện tổng số khách dự định sẽ quay lại và tổng số khách không quay lại, tổng theo cột là số khách được phỏng vấn tại từng khu nghỉ. Tổng hàng hay tổng cột đều bằng nhau và bằng cỡ mẫu.

Để kiểm định giả thuyết

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

Ta dùng trị thống kê kiểm định Chi bình phương có công thức

$$\chi^2 = \sum_{tất cả các ô} \frac{(O - E)^2}{E}$$

Trong đó

- O là tần số thực tế trong một ô cụ thể của bảng
- E là tần số lý thuyết (khi cho rằng H_0 đúng) trong một ô cụ thể của bảng

Giá trị kiểm định có phân phối Chi bình phương với 1 bậc tự do.

Để tính O ta nhớ rằng:

- Số thành công trong nhóm 1 ta lấy X_1 trong nhóm hai ta lấy X_2
- Số không thành công trong nhóm 1 ta lấy $n_1 - X_1$ trong nhóm hai ta lấy $n_2 - X_2$, như vậy bảng tổng hợp trên là bảng trình bày các tần số thực tế O.

Để tính toán E ta lập luận như sau:

Nếu H_0 đúng thì tỷ lệ thành công trong cả hai tổng thể là như nhau, sự khác biệt ta thấy chẳng qua là do tình cờ lấy mẫu và do đó chúng ta dùng một con số thống kê kết hợp cả hai giá trị này là tốt nhất, đó là giá trị p_s thể hiện tỷ lệ thành công chung của hai nhóm kết hợp

$$p_s = (X_1 + X_2) / n_1 + n_2 = X/n$$

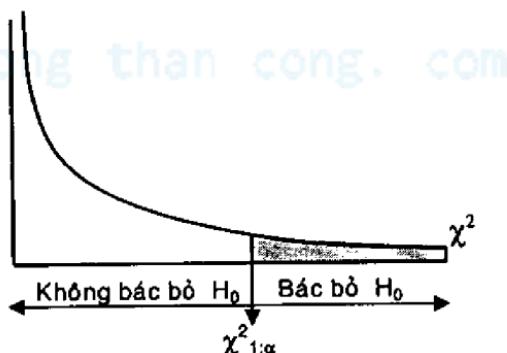
Để tính E mô tả số thành công (dòng đầu của bảng kết hợp) ta lấy tổng cột của từng nhóm nhân với p_s , tức là $E_1 = n_1 \times p_s$ và $E_2 = n_2 \times p_s$

Để tính E mô tả số không thành công (dòng hai của bảng kết hợp) ta lấy tổng cột của từng nhóm nhân với $1-p_s$, tức là

$$E_1 = n_1 \times (1-p_s) \text{ và } E_2 = n_2 \times (1-p_s)$$

Sau khi tính được giá trị kiểm định ta so sánh với giá trị tra bảng $\chi^2_{1,\alpha}$ theo quy tắc sau

Hình 8.18



Tiếp tục với ví dụ về sự hài lòng của khách hàng, giả thuyết kiểm định ta đặt ra là không có khác biệt trong tỷ lệ khách muốn quay trở lại ở hai khu nghỉ dưỡng và giả thuyết đối là

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

Ta tính toán ước lượng cho tỷ lệ khách muốn quay lại xét chung cả hai khu nghỉ là: $p_s = (163+154)/227+262 = 317/489 = 0,648$

Ta tính toán ước lượng cho tỷ lệ khách không muốn quay lại xét chung cả hai khu nghỉ là $(1-p_s) = 1 - 0,648 = 0,352$

Đem các tỷ lệ này nhân với số khách được phỏng vấn tại từng khu nghỉ để có các tần số lý thuyết sau

$$\text{Quay lại} - A = 0,648 \times 227 = 147,1$$

$$\text{Quay lại} - B = 0,648 \times 262 = 169,78$$

$$\text{Không quay lại} - A = (1 - 0,648) \times 227 = 79,9$$

$$\text{Không quay lại} - B = (1 - 0,648) \times 262 = 92,2$$

Bảng 8.6

Mô tả các tần số lý thuyết	Khu nghỉ		
Dự định quay lại	A	B	Tổng
Có	147,1	169,8	316,9
Không	79,9	92,2	172,1
Tổng	227	262	489

Ta lập bảng tính các thành phần của đại lượng kiểm định Chi bình phương

Bảng 8.7

O	E	$(O-E)^2/E$
163	147,1	1,719
154	169,8	1,470
64	79,9	3,164
108	92,2	2,708
Tổng		9,061

$$\text{Như vậy giá trị kiểm định } \chi^2_{\text{II}} = \sum_{\text{tất cả các ô}} \frac{(O-E)^2}{E} = 9,061$$

Tra bảng giá trị tới hạn ta có $\chi^2_{1;0,05} = 3,84$.

Như vậy $\chi^2_{\text{II}} = 9,061 > \chi^2_{1;0,05} = 3,84 \rightarrow$ bác bỏ H_0

Như vậy ta có đủ bằng chứng thống kê để kết luận tỷ lệ khách muôn quay lại hai khu nghỉ này khác nhau, hay mức độ hài lòng của khách tại hai khu khác nhau.

Như đã nói trên, kết quả kiểm định của hai phương pháp là như nhau khi chúng ta dùng để kiểm định giả thuyết hai bên $H_0: p_1 = p_2$ và $H_1: p_1 \neq p_2$. Tuy nhiên với giả thuyết một bên, để xác định cụ thể ví dụ $H_1: p_1 < p_2$ thì ta phải dùng kiểm định z. Tuy nhiên phương pháp kiểm định Chi bình phương và bảng chéo phát triển ở đây sẽ cung cấp tiền đề lý thuyết để các bạn nghiên cứu chương Kiểm định phi tham số tốt hơn.

8.3.3 Kiểm định giả thuyết cho hai phương sai tổng thể

Khi cần có một phương pháp để thực hiện kiểm định hai tổng thể có biến động cùng mức độ như nhau không (ví dụ tính ổn định của hai phương pháp sản xuất, cách cho điểm của hai giảng viên đại học...) chúng ta dùng phương pháp kiểm định phương sai của hai tổng thể độc lập dựa trên một đại lượng F như sau:

$$F_n = \frac{s_1^2}{s_2^2}$$

Trong đó

s_1^2 là phương sai của mẫu thứ nhất, mẫu này có cỡ n_1
 s_2^2 là phương sai của mẫu thứ hai, mẫu này có cỡ n_2

Thông thường để xác định mẫu nào là mẫu thứ nhất và mẫu nào là mẫu thứ 2 ta làm như sau, trong khi tính đại lượng F thì giá trị phương sai lớn hơn sẽ được đặt ở tử số, và như vậy mẫu tương ứng với phương sai đó là mẫu thứ nhất.

Giả thuyết đặt ra là kiểm định hai bên

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Nếu tỉ số F rất lớn hoặc rất nhỏ ta có thể suy diễn rằng hai phương sai tổng thể khó mà bằng nhau, ngược lại nếu tỉ số này gần đến 1 ta sẽ có bằng chứng ủng hộ cho giả thuyết H_0 . Như vậy tỉ lệ F lớn đến đâu thì xem như là đủ bằng chứng bác bỏ H_0 và ngược lại.

Nếu mỗi tổng thể lấy mẫu được giả định là có phân phối bình thường thì tỉ lệ $F = s_1^2/s_2^2$ có phân phối xác suất được gọi tên là phân phối F (theo tên của nhà thống kê Fisher). Cũng như phân phối Chi bình phương, hàm phân phối F có cấu trúc khá phức tạp và không phải là một phân phối cân đối, trong phạm vi môn học này chúng ta không đi vào tìm hiểu bản chất

của phân phối F mà ta sẽ chấp nhận và sử dụng các kết quả do các nhà thống kê lập sẵn về giá trị tối hạn của phân phối F theo các mức ý nghĩa và bậc tự do cho trước. Các giá trị tối hạn của phân phối F phụ thuộc vào hai giá trị bậc tự do, bậc tự do tử số ($df_1 = n_1 - 1$) gắn liền với mẫu thử nhất và bậc tự do mẫu số gắn liền với mẫu thử 2 ($df_2 = n_2 - 1$), các tình huống kết hợp của bậc tự do được lập sẵn bảng, bạn đọc tra tại Bảng tra số 4. Trong bảng tra này, bậc tự do của tử số nằm trên hàng đầu tiên, bậc tự do của mẫu số nằm trên cột đầu tiên

Quy tắc thực sự để quyết định bác bỏ H_0 với kiểm định hai bên khi $df_1 = n_1 - 1$ và $df_2 = n_2 - 1$, mức ý nghĩa α là: giả thuyết H_0 sẽ bị bác bỏ nếu giá trị kiểm định F lớn hơn giá trị tối hạn trên $F_U = F_{df_1; df_2; \alpha/2}$ của phân phối F hoặc bé hơn giá trị tối hạn dưới $F_L = F_{df_1; df_2; 1-\alpha/2}$ của phân phối, tức là $F_U < F_{df_1; df_2; 1-\alpha/2}$ hoặc $F_L > F_{df_1; df_2; \alpha/2}$

Nếu chúng ta có kiểm định bên phải

$$H_0: \sigma_1^2 \leq \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

Quy tắc bác bỏ H_0 là khi $F_U > F_{U(n_1-1; n_2-1; \alpha)}$.

Như vậy khi bài toán kiểm định chọn mức ý nghĩa α , nếu bạn làm kiểm định hai bên bạn sẽ tra giá trị F tại bảng có mức ý nghĩa là $\alpha/2$. Nếu bạn làm kiểm định 1 bên bạn sẽ tra giá trị F tại bảng có mức ý nghĩa đúng α .

Hướng dẫn cách tra các giá trị F_U và F_L cho kiểm định hai bên

Giả sử với mức ý nghĩa $\alpha = 0,1$; và ta có $n_1 = 20$; $n_2 = 10 \rightarrow df_1 = n_1 - 1 = 20 - 1 = 19$; $df_2 = n_2 - 1 = 10 - 1 = 9$, xem Bảng tra số 4 tại hàng mang số 9 và cột mang số 19 ta được giá trị tối hạn trên $F_U = F_{(n_1-1; n_2-1; \alpha/2)} = F_{(19; 9; 0,05)} = 2,95$

Bảng 8.8

df	17	18	19
7	3.48	3.47	3.46
8	3.19	3.17	3.16
9	2.97	2.96	2.95
10	2.81	2.80	2.79
11	2.69	2.67	2.66

Nhưng bảng tra F chỉ liệt kê các giá trị tối hạn trên $F_U = F_{(n_1-1; n_2-1; \alpha/2)}$ nên để tra được giá trị $F_L = F_{(n_1-1; n_2-1; 1-\alpha/2)}$ tương ứng ta làm như sau:

Xác định F_{U*} , theo cách xác định F_U nhưng lúc này ta đảo ngược bậc tự do so với khi tra F_U , tức là lúc này bậc tự do tử số là $n_2 - 1$ và bậc tự do mẫu số là $n_1 - 1$, tức $F_{U*} = F_{(n_2-1, n_1-1; \alpha/2)}$. Rồi tính giá trị F_L theo công thức $F_L = 1/F_{U*}$. Như vậy $F_{U*} = F_{(9; 19; 0.05)} = 2.42 \rightarrow$ Giá trị $F_L = 1/F_{U*} = 1/2.42 = 0.41$

Bảng 8.9

df	8	9	10
17	2.55	2.49	2.45
18	2.51	2.46	2.41
19	2.48	2.42	2.38
20	2.45	2.39	2.35

Tuy nhiên cũng chú ý là khi xây dựng công thức của F chúng ta luôn đặt giá trị phương sai nào lớn hơn trên mẫu nên giá trị F này luôn lớn hơn 1, trong khi theo cách thức xác định giá trị tối hạn dưới có thể nhận thấy F_L luôn có giá trị bé hơn 1, nên giá trị kiểm định này không bao giờ rơi vào miền bác bỏ H_0 dưới ($0; F_L$) nên khi xét quy tắc quyết định bác bỏ H_0 ta chỉ cần xem xét giá trị tối hạn trên F_U , tức nếu $F_u > F_{df1, df2; \alpha/2}$ thì bác bỏ H_0 .

Ví dụ: Một công ty chuyên cung cấp dịch vụ điện thoại di động muốn khảo sát có sự khác biệt trong biến thiên hóa đơn điện thoại trung bình tháng của khách hàng là nhà kinh doanh nam và nữ hay không. Họ tiến hành thu thập một mẫu ngẫu nhiên 20 khách nam và một mẫu ngẫu nhiên 10 khách hàng nữ. Hai mẫu này xem như độc lập nhau vì chi tiêu của một khách hàng nam không có liên quan đến chi tiêu của khách hàng nữ, việc lấy khách hàng nam này vào mẫu nghiên cứu cũng không ảnh hưởng gì đến việc chọn khách hàng nữ kia vào mẫu. Tính toán các tham số độ lệch chuẩn mẫu như sau:

Độ lệch chuẩn mẫu khách nữ $s_1 = 164.000$ VNĐ

Độ lệch chuẩn mẫu khách nam $s_2 = 146.000$ VNĐ

Chọn độ tin cậy của kiểm định là 95%. Kiểm định xem có khác biệt trong biến thiên chi tiêu cho điện thoại di động của khách hàng nam so với khách nữ hay không.

1. Quan tâm của chúng ta là có khác biệt không trong biến thiên chi tiêu cho điện thoại di động hàng tháng của khách hàng thuộc hai giới tính, xem mỗi giới khách hàng như một tổng thể mà từ đó ta lấy được mẫu độc lập, kiểm định của chúng ta quan tâm đến việc kiểm định sự bằng nhau của hai phương sai tổng thể.

2. Đặt giả thuyết

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Như vậy đây là kiểm định hai bên, trong đó mẫu khách hàng nữ xem là mẫu thứ nhất và mẫu khách nam xem là mẫu thứ 2

3. Chọn độ tin cậy là 95% như vậy $\alpha = 0,05$

4. Tính toán giá trị kiểm định

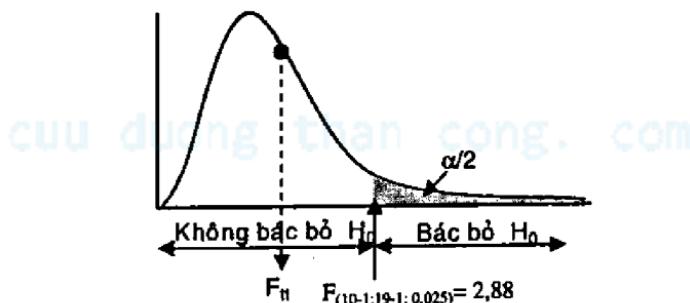
$$F_u = \frac{s_1^2}{s_2^2} = \frac{164^2}{146^2} = 1,26$$

5. Ta có $\alpha = 0,05 \rightarrow F_{(n1-1; n2-1; \alpha/2)} = F_{(10-1; 19-1; 0,025)} = 2,88$

Theo quy tắc nếu $F_u > F_{(n1-1; n2-1; \alpha/2)}$ \rightarrow bác bỏ H_0

Ta có $F_u = 1,26 < F_{(n1-1; n2-1; \alpha/2)} = 2,88 \rightarrow$ không bác bỏ H_0

Hình 8.19



6. Với độ tin cậy 95%, không có đủ bằng chứng thống kê để cho rằng biến thiên chi tiêu cho điện thoại di động của khách nam và của khách nữ khác nhau.

CHƯƠNG 9

PHÂN TÍCH PHƯƠNG SAI

Mục tiêu của phân tích phương sai (Analysis of Variance - ANOVA) là so sánh trung bình của nhiều nhóm (tổng thể) dựa trên các trị trung bình của các mẫu quan sát từ các nhóm này, và thông qua kiểm định giả thuyết để kết luận về sự bằng nhau của các trung bình tổng thể này. Trong nghiên cứu, phân tích phương sai được dùng như một công cụ để xem xét ảnh hưởng của một yếu tố nguyên nhân (định tính) đến một yếu tố kết quả (định lượng). Ví dụ như khi nghiên cứu ảnh hưởng của thời gian tự học đến kết quả học tập của sinh viên. Nếu thời gian tự học của sinh viên được thu thập dạng dữ liệu định tính (dưới 9 giờ/tuần, 9-18 giờ/tuần, trên 18 giờ/tuần); và kết quả học tập của sinh viên là dữ liệu định lượng (điểm trung bình học tập), thì phân tích phương sai là phương pháp phù hợp vì chúng ta có 3 nhóm cần so sánh trị trung bình.

Nếu chứng minh được 3 nhóm sinh viên có mức độ thời gian tự học khác nhau đều có kết quả điểm trung bình học tập bằng nhau, chúng ta kết luận được rằng ảnh hưởng của yếu tố thời gian tự học đến yếu tố kết quả học tập của những nhóm sinh viên có thời gian tự học khác nhau là như nhau. Nếu qua phân tích phương sai chúng ta thấy rằng 3 nhóm sinh viên có kết quả điểm trung bình khác nhau, trong đó nhóm có thời gian tự học nhiều (trên 18 giờ/tuần) có kết quả học tập cao hơn 2 nhóm kia một cách có ý nghĩa thống kê, thì kết luận rút ra là thời gian tự học khác nhau sẽ có ảnh hưởng đến kết quả học tập.

Trong chương này chúng ta đề cập đến hai mô hình phân tích phương sai: phân tích phương sai một yếu tố và hai yếu tố. Cụm từ yếu tố ở đây ám chỉ số lượng yếu tố nguyên nhân ảnh hưởng đến yếu tố kết quả đang nghiên cứu. Vậy thì với ví dụ vừa nêu trên ta có một yếu tố nguyên nhân là thời gian tự học ảnh hưởng đến yếu tố kết quả học tập nên ta có loại phân tích phương sai một yếu tố.

9.1 PHÂN TÍCH PHƯƠNG SAI MỘT YẾU TỐ

Phân tích phương sai một yếu tố (One-way ANOVA) là phân tích ảnh hưởng của một yếu tố nguyên nhân (dạng biến định tính) ảnh hưởng đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu. Ví dụ như xem xét ảnh hưởng của thời gian tự học của sinh viên đến kết quả học tập. Như đã phân tích ở trên, căn cứ vào thời gian tự học ta có 3 nhóm sinh viên cần so sánh về điểm trung bình học tập là nhóm dưới 9 giờ/tuần,

nhóm 9-18 giờ/tuần, và nhóm trên 18 giờ/tuần, cả 3 nhóm này thể hiện các cấp độ của một yếu tố đó là yếu tố thời gian tự học. Xét rộng ra, 3 nhóm sinh viên này như mẫu đại diện của 3 tổng thể sinh viên với thời gian tự học khác nhau, mục đích của chúng ta là tìm hiểu xem điểm trung bình học tập của 3 tổng thể này thực ra giống hay khác nhau để kết luận liệu có hay không sự ảnh hưởng của yếu tố thời gian tự học đến kết quả học tập của sinh viên. Ta đi vào lý thuyết như sau:

9.1.1 Trường hợp k tổng thể có phân phối bình thường và phương sai bằng nhau

Giả sử rằng chúng ta muốn so sánh trung bình của k tổng thể (với ví dụ trên thì $k = 3$) dựa trên những mẫu ngẫu nhiên độc lập gồm n_1, n_2, \dots, n_k quan sát từ k tổng thể này. Cần ghi nhớ ba giả định sau đây về các nhóm tổng thể được tiến hành phân tích ANOVA

- Các tổng thể này có phân phối bình thường
- Các phương sai tổng thể bằng nhau
- Các quan sát được lấy mẫu là độc lập nhau

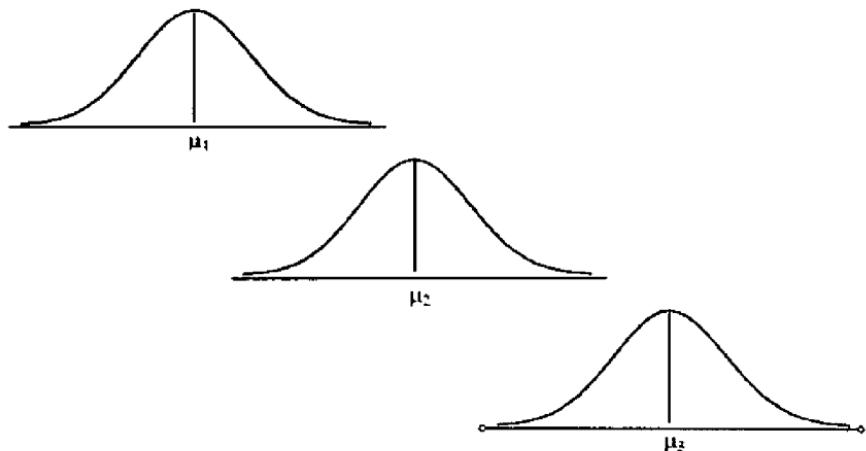
Nếu trung bình của các tổng thể được kí hiệu là $\mu_1, \mu_2, \dots, \mu_k$ thì khi các giả định trên được đáp ứng, mô hình phân tích phương sai một yếu tố ảnh hưởng được mô tả dưới dạng kiểm định giả thuyết như sau:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

Giả thuyết H_0 cho rằng trung bình của k tổng thể đều bằng nhau (về mặt nghiên cứu liên hệ thì giả thuyết này cho rằng yếu tố nguyên nhân không có tác động gì đến vấn đề ta đang nghiên cứu). Và giả thuyết đối là:

$$H_1: \text{Tồn tại ít nhất một cặp trung bình tổng thể khác nhau}$$

Hai giả định đầu tiên để tiến hành phân tích phương sai được mô tả như hình dưới đây, bạn thấy ba tổng thể đều có phân phối bình thường với mức độ phân tán tương đối giống nhau, nhưng ba vị trí chênh lệch của chúng cho thấy ba trị trung bình khác nhau. Rõ ràng là nếu bạn thực sự có các giá trị của 3 tổng thể và biểu diễn được phân phối của chúng như hình dưới thì bạn không cần phải làm gì nữa mà kết luận được ngay là bạn bác bỏ H_0 , hay 3 tổng thể này có trị trung bình khác nhau.



Nhưng bạn chỉ có mẫu đại diện được quan sát, nên để kiểm định giả thuyết này, ta thực hiện các bước sau:

Bước 1: Tính các trung bình mẫu của các nhóm (xem như đại diện của các tổng thể)

Trước hết ta xem cách tính các trung bình mẫu từ những quan sát của k mẫu ngẫu nhiên độc lập (kí hiệu $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$) và trung bình chung của k mẫu quan sát (kí hiệu \bar{x}) từ trường hợp tổng quát như sau:

Bảng 9.1: Bảng số liệu tổng quát thực hiện phân tích phương sai

Tổng thể			
1	2	...	k
x_{11}	x_{21}	...	x_{k1}
x_{12}	x_{22}	...	x_{k2}
...
x_{1n_1}	x_{2n_2}	...	x_{kn_k}

Tính trung bình mẫu của từng nhóm $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ theo công thức

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad (i = 1, 2, \dots, k)$$

Và trung bình chung của k mẫu (trung bình chung của toàn bộ mẫu khảo sát):

$$\bar{x} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{\sum_{i=1}^k n_i}$$

Dĩ nhiên bạn có thể tính trung bình chung của k mẫu theo cách khác là : cộng tất cả các x_{ij} trên Bảng 9.1 lại rồi đem chia cho $\sum n_i$ với ($i = 1, 2, \dots, k$). Kết quả là như nhau.

Bước 2: Tính các tổng các chênh lệch bình phương (hay gọi tắt là tổng bình phương)

Tính tổng các chênh lệch bình phương trong nội bộ nhóm SSW¹ và tổng các chênh lệch bình phương giữa các nhóm SSG²

- Tổng các chênh lệch bình phương trong nội bộ nhóm (SSW) được tính bằng cách cộng các chênh lệch bình phương giữa các giá trị quan sát với trung bình mẫu của từng nhóm, rồi sau đó lại tính tổng cộng kết quả tất cả các nhóm lại. SSW phản ánh phần biến thiên của yếu tố kết quả do ảnh hưởng của các yếu tố khác, chứ không phải do yếu tố nguyên nhân đang nghiên cứu (là yếu tố dùng để phân biệt các tổng thể/nhóm đang so sánh)

Tổng các chênh lệch bình phương của từng nhóm được tính theo công thức:

$$\text{Nhóm 1: } SS_1 = \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2$$

$$\text{Nhóm 2: } SS_2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2$$

Tương tự như vậy ta tính cho đến nhóm thứ k được SS_k . Vậy tổng các chênh lệch bình phương trong nội bộ các nhóm được tính như sau:

$$SSW = SS_1 + SS_2 + \dots + SS_k$$

Hay viết tổng quát theo công thức ta có

$$SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

- Tổng các chênh lệch bình phương giữa các nhóm (SSG) được tính bằng cách cộng các chênh lệch được lấy bình phương giữa các trung

¹ Sum of squares within group

² Sum of squares between group

bình mẫu của từng nhóm với trung bình chung của k nhóm (các chênh lệch này đều được nhân thêm với số quan sát tương ứng của từng nhóm). SSG phản ảnh phần biến thiên của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân đang nghiên cứu.

$$SSG = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$$

- Tổng các chênh lệch bình phương toàn bộ SST³ được tính bằng cách cộng tổng các chênh lệch đã lấy bình phương giữa từng giá trị quan sát của toàn bộ mẫu nghiên cứu (x_{ij}) với trung bình chung toàn bộ (\bar{x}). SST phản ảnh biến thiên của yếu tố kết quả do ảnh hưởng của tất cả các nguyên nhân.

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

Có thể dễ dàng chứng minh là tổng các chênh lệch bình phương toàn bộ bằng tổng cộng tổng các chênh lệch bình phương trong nội bộ các nhóm và tổng các chênh lệch bình phương giữa các nhóm.

$$SST = SSW + SSG$$

Như vậy công thức trên cho thấy, SST là toàn bộ biến thiên của yếu tố kết quả đã được phân tích thành 2 phần: phần biến thiên do yếu tố đang nghiên cứu tạo ra (SSG) và phần biến thiên còn lại do các yếu tố khác không nghiên cứu ở đây tạo ra (SSW). Nếu phần biến thiên do yếu tố nguyên nhân đang xét tạo ra càng “đáng kể” so với phần biến thiên do các yếu tố khác không xét tạo ra, thì chúng ta càng có cơ sở để bác bỏ H₀ và kết luận là yếu tố nguyên nhân đang nghiên cứu ảnh hưởng có ý nghĩa đến yếu tố kết quả.

Bước 3: Tính các phương sai (là trung bình của các chênh lệch bình phương).

Các phương sai được tính bằng cách lấy các tổng các chênh lệch bình phương chia cho bậc tự do tương ứng.

Tính phương sai trong nội bộ nhóm (MSW) bằng cách lấy tổng các chênh lệch bình phương trong nội bộ các nhóm (SSW) chia cho bậc tự do tương ứng là $n-k$ (n là số quan sát, k là số nhóm so sánh). MSW là ước lượng phần biến thiên của yếu tố kết quả do các yếu tố khác gây ra (hay giải thích).

³ Total sum of squares

$$MSW = \frac{SSW}{n - k}$$

Tính phương sai giữa các nhóm (MSG) bằng cách lấy tổng các chênh lệch bình phương giữa các nhóm chia cho bậc tự do tương ứng là $k - 1$. MSG là ước lượng phần biến thiên của yếu tố kết quả do yếu tố nguyên nhân đang nghiên cứu gây ra (hay giải thích được).

$$MSG = \frac{SSG}{k - 1}$$

Bước 4: Kiểm định giả thuyết

Giả thuyết về sự bằng nhau của k trung bình tổng thể được quyết định dựa trên tỉ số của hai phương sai: phương sai giữa các nhóm (MSG) và phương sai trong nội bộ nhóm (MSW). Tỉ số này được gọi là tỷ số F vì nó tuân theo qui luật Fisher – Snedecor với bậc tự do là $k-1$ ở tử số và $n-k$ ở mẫu số

$$F = \frac{MSG}{MSW}$$

Ta bác bỏ giả thuyết H_0 cho rằng trị trung bình của k tổng thể bằng nhau khi:

$$F > F_{(k-1, n-k), \alpha}$$

$F_{(k-1, n-k), \alpha}$ là giá trị giới hạn tra từ Bảng tra số 4 với bậc tự do $k-1$ tra theo hàng đầu tiên và $n-k$ tra theo hàng đầu tiên, nhớ chọn bảng có mức ý nghĩa phù hợp.

Sau đây là dạng bảng kết quả tổng quát của ANOVA khi phân tích bằng chương trình Excel hay SPSS.

Bảng 9.2 Dạng bảng kết quả ANOVA từ chương trình Excel, SPSS

Bảng gốc bằng tiếng Anh

Source of Variation	Sum of squares (SS)	Degree of Freedom (df)	Mean squares (MS)	F ratio
Between-groups	SSG	$k - 1$	$MSG = \frac{SSG}{k - 1}$	$F = \frac{MSG}{MSW}$
Within-groups	SSW	$n - k$	$MSW = \frac{SSW}{n - k}$	
Total	SST	$n - 1$		

Tạm dịch sang tiếng Việt:

Nguồn biến thiên	Tổng chênh lệch bình phương (SS)	Bậc tự do (df)	Phương sai (MS)	Tỉ số F
Giữa các nhóm	SSG	k - 1	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSW}$
Trong nội bộ các nhóm	SSW	n - k	$MSW = \frac{SSW}{n-k}$	
Toàn bộ	SST	n - 1		

Ý nghĩa của các công thức và logic của các tính toán trong bảng trên cần được hiểu rõ để có thể vận dụng và giải thích các kết quả phân tích một cách súc tích. Giả sử, chúng ta trở lại ví dụ nghiên cứu ảnh hưởng của thời gian tự học của các sinh viên đến kết quả học tập của sinh viên đã đề cập ở đầu chương này. Trong trường hợp này ta có k = 3 (3 nhóm so sánh). Giả thuyết H_0 trong ví dụ này có thể được phát biểu như sau:

H_0 : Thời gian tự học không ảnh hưởng đến kết quả học tập của sinh viên.

H_1 : Thời gian tự học có ảnh hưởng đến kết quả học tập của sinh viên

Các bạn hãy lập luận về logic như sau trước khi dùng số liệu để tính toán cụ thể. Nếu giả thuyết H_0 đúng, ảnh hưởng của thời gian tự học đến kết quả học tập là như nhau đối với các nhóm sinh viên có thời gian tự học khác nhau (tức là kết quả học tập của các sinh viên này khác nhau là do các yếu tố khác như: tình trạng sức khỏe, mức độ yêu thích ngành đang học, phương pháp học ...) thì trong nội bộ 3 nhóm, điểm trung bình học tập sẽ rất phân tán. Cùng nhóm thời gian tự học ít (dưới 9 giờ/tuần), có sinh viên đạt điểm trung bình rất thấp, có sinh viên có điểm bình thường, nhưng cũng có sinh viên đạt điểm cao, tính trung bình cả nhóm thì điểm trung bình không cao cũng không thấp, và không khác biệt nhiều với tình trạng nội bộ của 2 nhóm kia.

Tương tự, trong nhóm thời gian tự học nhiều (trên 18 giờ/tuần), có sinh viên đạt điểm trung bình rất cao, có sinh viên có điểm bình thường, nhưng cũng có sinh viên đạt điểm rất thấp, tính trung bình cả nhóm thì điểm trung bình không cao cũng không thấp, và không khác biệt nhiều với 2 nhóm còn lại. Điều này là do kết quả học tập bị ảnh hưởng bởi những yếu tố khác chưa nghiên cứu ở đây, các sinh viên cùng nhóm có thời gian tự học như nhau, nhưng vẫn có kết quả học tập khác nhau do tình trạng sức khỏe, điều kiện ăn ở, sinh hoạt, học tập, công việc làm thêm, yêu thích ngành học hay không, ... Kết quả là 3 trung bình mẫu của 3 nhóm so sánh khá gần nhau, và rất gần với trung bình chung cả 3 nhóm. Lúc đó tổng các chênh lệch bình phương giữa các nhóm (SSG) nhỏ khiến phương sai giữa các nhóm nhỏ (MSG), còn tổng các chênh lệch bình phương trong

nội bộ 3 nhóm (SSW) rất lớn (vì điểm kết quả học tập trong cùng 1 nhóm rất khác nhau như đã mô tả trên) khiến phương sai trong nội bộ nhóm (MSW) lớn. Như vậy khi ảnh hưởng của nguyên nhân (thời gian tự học) đến kết quả học tập không tạo khác biệt giữa 3 nhóm, thì dấu hiệu để nhận biết là SSG và MSG nhỏ, và SSW và MSW lớn. Kiểm định F được thực hiện bằng cách tính tỉ số F (MSG/MSW), tỉ số F sẽ tiến về 0 khi ảnh hưởng của yếu tố nguyên nhân lượng thời gian tự học không tạo khác nhau đối với kết quả học tập. F càng nhỏ thì càng có khả năng để chấp nhận giả thuyết H_0 . Nếu tỉ số F nhỏ hơn trị số F tra từ bảng thống kê theo các bậc tự do phù hợp và một mức ý nghĩa đã chọn thì ta chấp nhận giả thuyết H_0 .

Nếu giả thuyết H_0 sai, tức là quả thật lượng thời gian tự học của sinh viên có ảnh hưởng đến kết quả học tập của sinh viên, thì trong nhóm các sinh viên tự học nhiều (trên 18 giờ/tuần), sinh viên nào cũng đều có kết quả điểm trung bình học tập cao, điểm kết quả trung bình học tập trong nhóm này ít phân tán, và khá đồng đều (tức đều cao). Các sinh viên trong nhóm tự học ít (dưới 9 giờ/tuần), hầu hết đều có kết quả ở mức trung bình trở xuống. Kết quả là điểm trung bình học tập của các sinh viên trong cùng một nhóm khá đều và điểm trung bình của 3 nhóm khá chênh lệch nhau.

Kết quả là tổng các chênh lệch bình phương giữa các nhóm (SSG) lớn và phương sai giữa các nhóm (MSG) lớn, còn tổng các chênh lệch bình phương trong nội bộ 3 nhóm (SSW) rất nhỏ (điểm trung bình học tập trong cùng 1 nhóm khá giống nhau) và phương sai trong nội bộ nhóm (MSW) nhỏ. Lúc này thì tỉ số F (MSG/MSW) khá lớn. Nếu F lớn quá giá trị giới hạn tra từ bảng thống kê F, thì ta bác bỏ giả thuyết H_0 , kết luận là thời gian tự học khác nhau có ảnh hưởng khác nhau đến kết quả học tập của sinh viên.

Ví dụ tính toán: Một nhóm nghiên cứu muốn xem xét ảnh hưởng của mức độ tự học đến kết quả học tập của sinh viên. Một cuộc khảo sát với cỡ mẫu là 63 sinh viên được thực hiện.

Có 21 sinh viên thời gian tự học ít, dưới 9 giờ/tuần. 21 sinh viên khác có thời gian tự học trung bình, khoảng từ 9 đến 18 giờ/tuần. Còn lại 21 sinh viên tự học nhiều, trên 18 giờ/tuần. Dữ liệu về kết quả trung bình học tập của năm học vừa qua do Phòng đào tạo nhà trường cung cấp theo yêu cầu của nhóm nghiên cứu được trình bày trong Bảng 9.3.

Bảng 9.3 Điểm trung bình học tập của các sinh viên

Nhóm 1 (TG tự học ít)	Nhóm 2 (TG tự học TB)	Nhóm 3 (TG tự học nhiều)
5.8	6.0	6.2
6.2	6.6	5.8
5.4	6.1	6.5
6.0	5.8	6.2
5.2	5.9	6.4
5.3	6.0	5.7
5.4	5.9	6.1
5.6	6.0	6.8
6.2	6.7	7.1
5.7	6.5	6.5
5.5	6.3	7.1
6.1	6.1	7.2
6.0	6.8	6.7
5.2	6.4	7.0
6.4	6.8	7.6
5.5	6.6	7.7
5.0	6.4	7.8
5.6	6.2	6.8
6.2	7.1	7.3
6.1	7.0	7.1
5.3	7.2	7.2
119.7	134.4	142.8

Phát biểu giả thuyết:

H_0 : Thời gian tự học không ảnh hưởng đến kết quả học tập của sinh viên; hay

H_0 : Điểm học tập trung bình của 3 nhóm sinh viên có thời gian tự học khác nhau là bằng nhau;

hay

$H_0: \mu_1 = \mu_2 = \mu_3$

Các giả thuyết trên là tương đương nhau.

Và H_1 được đặt theo tình huống đối nghĩa với H_0

Bước 1: Tính các trung bình của từng nhóm và trung bình chung 3 nhóm

Điểm trung bình học tập (ĐTB) của sinh viên:

$$\text{Nhóm 1: } \bar{x}_1 = \frac{119,7}{21} = 5,7$$

$$\text{Nhóm 2: } \bar{x}_2 = \frac{134,4}{21} = 6,4$$

$$\text{Nhóm 3: } \bar{x}_3 = \frac{142,8}{21} = 6,8$$

$$\text{Cả 3 nhóm: } \bar{x} = \frac{21x5,7 + 21x6,4 + 21x6,8}{21 + 21 + 21} = 6,3$$

Bước 2: Tính các tổng các chênh lệch bình phương

- $SSW = SS_1 + SS_2 + SS_3$

Trong đó

$$SS_1 = \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2 \quad (\text{với } n_1 = 21)$$

$$= (5,8 - 5,7)^2 + (6,2 - 5,7)^2 + \dots + (6,1 - 5,7)^2 + (5,3 - 5,7)^2 = 3,34$$

Tương tự:

$$SS_2 = (6 - 6,4)^2 + (6,6 - 6,4)^2 + \dots + (7,2 - 6,4)^2 = 3,56$$

$$SS_3 = (6,2 - 6,8)^2 + (5,8 - 6,8)^2 + \dots + (7,2 - 6,8)^2 = 7,1$$

$$\Rightarrow SSW = 3,34 + 3,56 + 7,1 = 14$$

- $SSG = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 \quad (\text{với } k = 3)$

$$SSG = 21(5,7 - 6,3)^2 + 21(6,4 - 6,3)^2 + 21(6,8 - 6,3)^2 = 13,02$$

Bước 3: Tính các phương sai

Phương sai trong nội bộ nhóm:

$$MSW = \frac{SSW}{n-k} = \frac{14}{63-3} = 0,233$$

Phương sai giữa các nhóm:

$$MSG = \frac{SSG}{k-1} = \frac{13,02}{3-1} = 6,51$$

Bước 4: Tính tỉ số F

$$F = \frac{MSG}{MSW} = \frac{6,51}{0,233} = 27,94$$

Tra bảng phân phối F với mức ý nghĩa $\alpha = 0,05$ tại các bậc tự do tương ứng:

$$F_{(k-1; n-k); \alpha} = F_{(k-1; n-k); 0.05} = F_{(3-1; 63-3); 0.05} = 3,15$$

Chú ý là gấp những tình huống các bậc tự do không phù hợp với bảng tra chúng ta có thể dùng Excel tìm giá trị cần thiết rất nhanh chóng, bạn sẽ được hướng dẫn cách tra này ở mục kế tiếp.

Vì $F = 27,94 > 3,15$ cho nên dựa trên dữ liệu đã thu thập, chúng ta có đủ bảng chứng để bác bỏ giả thuyết H_0 cho rằng điểm trung bình học tập trung bình của ba nhóm sinh viên bằng nhau ở mức ý nghĩa 5%. Nghĩa là ở độ tin cậy 95% thì điểm trung bình học tập ở ba nhóm có thời gian tự học khác nhau là khác nhau. Người nghiên cứu có thể kết luận rằng, thời gian tự học có ảnh hưởng đến kết quả học tập của sinh viên có tự học.

Sau đây là bảng kết quả phân tích phương sai một yếu tố tính toán từ chương trình Excel.

Bảng 9.4 Bảng kết quả ANOVA một yếu tố từ chương trình Excel

SUMMARY

Groups	Count	Sum	Average	Variance
Ít	21	119.7	5.7	0.167
TB	21	134.4	6.4	0.178
Nhiều	21	142.8	6.8	0.355

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	13.02	2	6.51	27.9	2.7E-09	3.2E+00
Within Groups	14	60	0.233			
Total	27.02	62				

Các bạn có thể đổi chiều các con số tính thủ công ở trên với các số liệu trong hai bảng phụ trong kết quả ANOVA của Excel để từ đó có thể suy ngược lại ý nghĩa của tên gọi các cột trong cấu trúc của bảng kết quả ANOVA trên Excel. Chú ý là giá trị P-value giúp ta quyết định theo nguyên tắc giá trị xác suất. Còn F crit là giá trị tra bảng của $F_{(k-1; n-k); \alpha}$ cả 2 cột dữ liệu cuối cùng này đều có công dụng giúp chúng ta không cần phải lật bảng tra ở phụ lục vẫn quyết định được có bác bỏ H_0 hay không.

9.1.2 Thực hiện ANOVA một yếu tố bằng Excel

Chú ý là để thực hiện được lệnh này trên Excel chúng ta cần tuân theo một quy tắc nhất định khi nhập dữ liệu, các bạn tham khảo cách nhập dữ liệu trên màn hình trong Hình 9.1a, nếu bạn có k nhóm thì bạn có k cột dữ liệu, mỗi cột là tất cả các giá trị quan sát của nhóm tương ứng, mỗi quan sát trên một hàng. Với ANOVA 1 yếu tố, không bắt buộc là số quan sát

của các nhóm phải bằng nhau tuyệt đối, do đó với các tình huống mà n_i khác nhau bạn cứ nhập dữ liệu theo từng hàng trộn vẹn cho mỗi nhóm, dĩ nhiên lúc này về cuối các cột dữ liệu sẽ bị chênh nhau chứ dữ liệu không nằm thẳng trên một hàng ngang như tại hàng 22 của worksheet như ví dụ của chúng ta ở đây. Trên cửa sổ dữ liệu đã nhập bạn vào menu Tool chọn lệnh Data Analysis để mở cửa sổ Data Analysis.

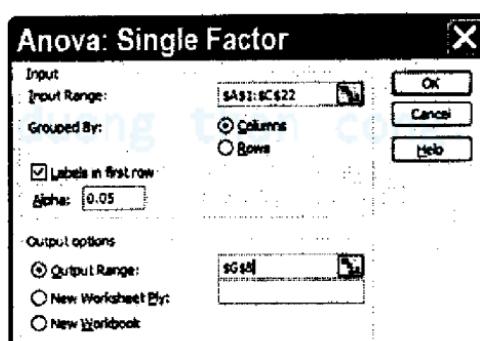
Hình 9.1a

The screenshot shows a Microsoft Excel spreadsheet with 20 rows and 3 columns. The columns are labeled A, B, and C. The first column (A) contains labels: 'ít', 'TB', 'Nhiều'. The second column (B) contains numerical values: 5.8, 6, 6.2, 5.4, 6, 5.2, 5.3, 5.4, 5.6, 6.2, 5.7, 5.5, 6.1, 6, 6.4, 6.3, 6.1, 6.8, 6.4, 6.4, 6.2. The third column (C) contains numerical values: 6.2, 5.8, 6.4, 6.1, 6.2, 5.9, 5.7, 6.1, 6.8, 7.1, 6.5, 7.2, 7.2, 6.7, 7, 7.6, 7.7, 7.8, 6.8, 7.3. To the right of the table, a 'Data Analysis' dialog box is open, showing the 'Analysis Tools' list with options like 'ANOVA: Single Factor', 'ANOVA: Two-Factor With Replication', etc.

	A	B	C	D	E	F
1	ít	TB	Nhiều			
2	5.8	6	6.2			
3	6.2	6.6	5.8			
4	5.4	6.1	6.5			
5	6	5.8	6.2			
6	5.2	5.9	6.4			
7	5.3	6	5.7			
8	5.4	5.9	6.1			
9	5.6	6	6.8			
10	6.2	6.7	7.1			
11	5.7	6.5	6.5			
12	5.5	6.3	7.1			
13	6.1	6.1	7.2			
14	6	6.8	6.7			
15	5.2	6.4	7			
16	6.4	6.8	7.6			
17	5.5	6.6	7.7			
18	5	6.4	7.8			
19	5.6	6.2	6.8			
20	6.2	7.1	7.3			

Trên cửa sổ Data Analysis bạn bấm ngay lựa chọn đầu tiên là ANOVA: Singler Factor để mở cửa sổ ANOVA Singler Factor và tiến hành các khai báo phù hợp như Hình 9.1b

Hình 9.1b



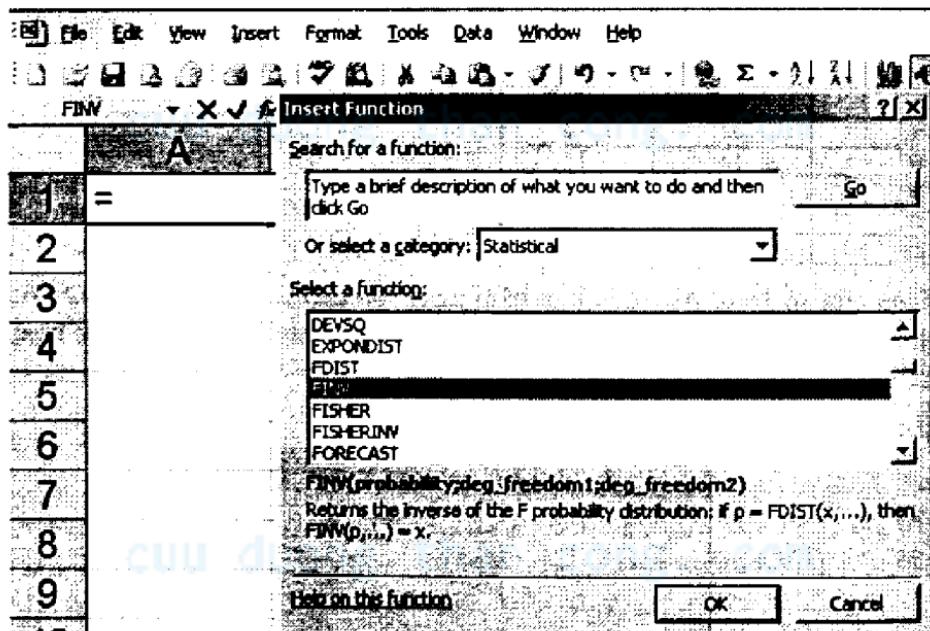
Nhấn nút OK bạn đọc có kết quả mong muốn.

Cách tra giá trị tới hạn F bằng Excel

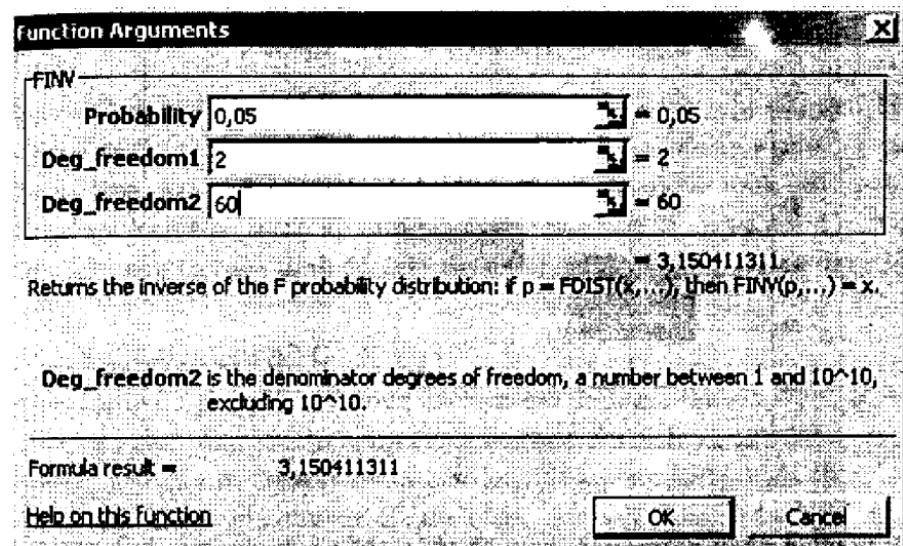
Giả sử nếu giá trị cần tìm $F_{(k-1;n-k);0.05} = F_{(3-1;63-3);0.05}$ không có trong bảng tra bạn có thể tìm thấy nó nhanh chóng trong Excel bằng cách tiến hành lệnh như sau:

- Bước 1: bật cửa sổ làm việc của Excel lên và nhập dấu = để sẵn sàng việc gọi hàm tính toán.
- Bước 2: vào menu Insert chọn lệnh Insert Function để chèn hàm
- Bước 3: khi mở được cửa sổ Insert Function bạn thực hiện các lựa chọn như thể hiện trong Hình 9.2a
- Sau khi nhấn OK bạn mở tiếp cửa sổ thứ 2
- Bước 4: thực hiện các khai báo như hướng dẫn trong Hình 9.2b trên cửa sổ này. Chú ý các số liệu về bậc tự do là các số liệu cuối cùng sau khi đã trừ theo công thức $(k-1)$ và $(n-k)$
- Bước 5: nhấn nút OK bạn được kết quả của giá trị tối hạn F

Hình 9.2a



Hình 9.2b

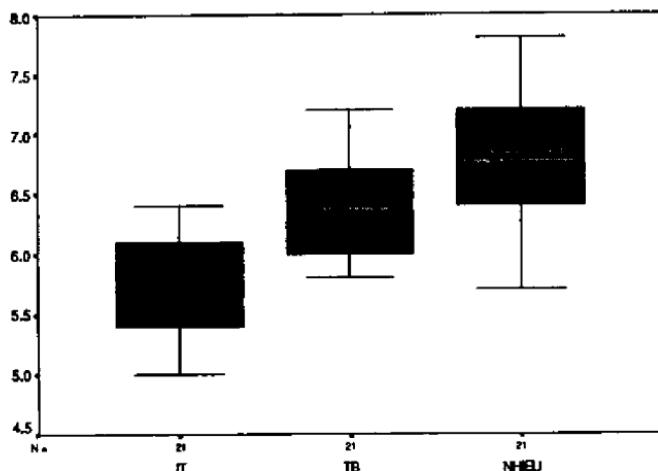


9.1.3 Kiểm tra các giả định của phân tích phương sai

Chúng ta có thể kiểm tra nhanh các giả định này bằng đồ thị. Histogram là phương pháp tốt nhất để kiểm tra giả định về phân phối bình thường của dữ liệu nhưng nó đòi hỏi một số lượng quan sát khá lớn. Biểu đồ thân lá hay biểu đồ hộp và râu là một thay thế tốt trong tình huống số quan sát ít hơn. Nếu công cụ đồ thị cho thấy tập dữ liệu mẫu khá phù hợp với phân phối bình thường thì ta có thể xem giả định phân phối bình thường đã thỏa mãn. Hình dưới mô tả biểu đồ hộp râu cho tập dữ liệu mẫu về ba nhóm sinh viên trong ví dụ của chúng ta. Đồ thị cho thấy ngoại trừ nhóm có thời gian tự học TB có hình dáng phân phối của dữ liệu hơi lệch sang trái, còn hai nhóm còn lại có phân phối khá cân đối. Với số quan sát không nhiều thì biểu hiện như thế này của dữ liệu là khả quan và có thể chấp nhận được.

Để khảo sát giả định bằng nhau của phương sai, biểu đồ hộp và râu cũng cho cảm nhận ban đầu nhanh chóng, với ba biểu đồ này, mức độ phân tán của dữ liệu trong mỗi tập dữ liệu mẫu không khác biệt nhau nhiều.

Hình 9.3



Một phương pháp kiểm định tham số chắc chắn hơn cho giả định phương sai bằng nhau là kiểm định Levene về phương sai của các tổng thể. Kiểm định này xuất phát từ giả thuyết sau.

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

H_1 : Không phải tất cả các phương sai đều bằng nhau

Để quyết định chấp nhận hay bác bỏ H_0 ta tính toán giá trị kiểm định F theo công thức

$$F_{\max} = \frac{s_{\max}^2}{s_{\min}^2}$$

Trong đó s_{\max}^2 là phương sai lớn nhất trong các nhóm nghiên cứu và s_{\min}^2 là phương sai nhỏ nhất trong các nhóm nghiên cứu.

Giá trị F tính được được đem so sánh với giá trị $F_{(k-1, n-k)}; \alpha$ tra được từ bảng phân phối Hartley F_{\max} (là Bảng tra số 5 trong phần Phụ lục). Trong đó k là số nhóm so sánh, bậc tự do df tính theo công thức $df = (n - 1)$. Trong tình

huống các nhóm có n_i khác nhau thì $\bar{n} = \frac{\sum_{i=1}^k n_i}{k}$ (chú ý là nếu kết quả tính \bar{n} là một số thập phân thì ta lấy phần nguyên).

Quy tắc quyết định:

$F_{\max} > F_{(k-1, n-k); \alpha}$ thì bác bỏ giả thuyết H_0 cho rằng phương sai bằng nhau và ngược lại.

Với ví dụ này thì $F_{\max} = \frac{0,36}{0,17} = 2,12$

$F_{(k-1); \alpha} = F_{(3-2; 1-\alpha); 0,05} = F_{(3-2; 1-\alpha); 0,05} = 2,95 > F_{\max} \rightarrow$ chấp nhận H_0

Nếu chúng ta không chắc chắn về các giả định hoặc nếu kết quả kiểm định cho thấy các giả định không được thỏa mãn thì một phương pháp kiểm định thay thế cho ANOVA là phương pháp kiểm định phi tham số Kruskal-Wallis sẽ được áp dụng. Tuy nhiên trong ví dụ này ở đây, ta có thể xem như các giả định để tiến hành phân tích phương sai đã được thỏa mãn.

9.1.4 Phân tích sâu ANOVA

Mục đích của phân tích phương sai là kiểm định giả thuyết H_0 rằng trung bình của các tổng thể bằng nhau. Sau khi phân tích và kết luận, có hai trường hợp xảy ra là chấp nhận giả thuyết H_0 hoặc bác bỏ giả thuyết H_0 . Nếu chấp nhận giả thuyết H_0 thì phân tích kết thúc. Nếu bác bỏ giả thuyết H_0 , bạn kết luận trung bình của các tổng thể không bằng nhau. Vì vậy, vấn đề tiếp theo là phân tích sâu hơn để xác định nhóm (tổng thể) nào khác nhau nào, nhóm nào có trung bình lớn hơn hay nhỏ hơn.

Có nhiều phương pháp để tiếp tục phân tích sâu ANOVA khi bác bỏ giả thuyết H_0 . Trong chương này chỉ đề cập đến 1 phương pháp thông dụng đó là phương pháp Tukey, phương pháp này còn được gọi là kiểm định HSD (Honestly Significant Differences). Nội dung của phương pháp này là so sánh từng cặp các trung bình nhóm ở mức ý nghĩa α nào đó cho tất cả các cặp kiểm định có thể để phát hiện ra những nhóm khác nhau. Nếu có k nhóm nghiên cứu, và chúng ta so sánh tất cả các cặp nhóm thì số lượng cặp cần phải so sánh là tổ hợp chập 2 của k nhóm.

$$C_k^2 = \frac{k!}{2!(k-2)!} \text{ hay } = \frac{k(k-1)}{2}$$

Ví dụ: ta có $k = 3$, thì số cặp so sánh trong kiểm định là 3, vì

$$C_3^2 = \frac{3!}{2!(3-2)!} = 3$$

Các giả thuyết cần kiểm định sẽ là:

- | | | |
|-------------------------|-------------------------|-------------------------|
| 1. $H_0: \mu_1 = \mu_2$ | 2. $H_0: \mu_2 = \mu_3$ | 3. $H_0: \mu_1 = \mu_3$ |
| $H_1: \mu_1 \neq \mu_2$ | $H_1: \mu_2 \neq \mu_3$ | $H_1: \mu_1 \neq \mu_3$ |

Giá trị giới hạn Tukey được tính theo công thức:

$$T = q_{\alpha, k, n-k} \sqrt{\frac{MSW}{n_i}}$$

Trong đó:

- $q_{\alpha, k, n-k}$ là giá trị tra bảng phân phối kiểm định Tukey (Bảng tra số 9) ở mức ý nghĩa α , với bậc tự do k và $n-k$, với n là tổng số quan sát mẫu ($n = \sum n_i$)
- MSW là phương sai trong nội bộ nhóm
- n_i là số quan sát trong 1 nhóm (tổng thể), trong trường hợp mỗi nhóm có số quan sát n_i khác nhau, sử dụng giá trị n_i nhỏ nhất

Tiêu chuẩn quyết định là bác bỏ giả thuyết H_0 khi độ lệch tuyệt đối giữa các cặp trung bình mẫu lớn hơn hay bằng T giới hạn.

Từ ví dụ tính toán ở phần trước, ta có $k = 3$, $\alpha = 5\%$, $n = 63$ và $MSW = 0,233$. Tra bảng phân phối q (phân phối Tukey) ta có: $q_{0,05;3,60} = 3,4$

Tính giá trị giới hạn Tukey: $T = 3,4 \sqrt{\frac{0,233}{21}} = 0,36$

Độ lệch tuyệt đối các cặp trung bình mẫu tính lần lượt như sau:

$$|\bar{x}_1 - \bar{x}_2| = |5,7 - 6,4| = 0,7$$

$$|\bar{x}_1 - \bar{x}_3| = |5,7 - 6,8| = 1,1$$

$$|\bar{x}_2 - \bar{x}_3| = |6,4 - 6,8| = 0,4$$

Như vậy, theo điều kiện bác bỏ giả thuyết H_0 thì, với $T = 0,36$:

- trung bình tổng thể μ_1 và μ_2 khác nhau vì $|\bar{x}_1 - \bar{x}_2| = 0,7 > T$
- trung bình tổng thể μ_2 và μ_3 khác nhau vì $|\bar{x}_2 - \bar{x}_3| = 0,4 > T$
- trung bình tổng thể μ_1 và μ_3 khác nhau vì $|\bar{x}_1 - \bar{x}_3| = 1,1 > T$

Vì $\bar{x}_1 < \bar{x}_2 < \bar{x}_3$ nên ta $\rightarrow \mu_1 < \mu_2 < \mu_3$

Như vậy chúng ta có thể kết luận rằng điểm trung bình học tập của các nhóm sinh viên có thời gian tự học khác nhau là khác nhau. Cụ thể, dựa vào trung bình nhóm, chúng ta có thể thấy điểm trung bình học tập của nhóm có thời gian tự học nhiều cao hơn hẳn hai nhóm kia, nhóm có thời gian tự học ít thấp hơn hẳn hai nhóm kia, nhóm có thời gian tự học trung bình cao hơn nhóm tự học ít nhưng thấp hơn nhóm tự học nhiều. Như vậy thời gian tự học có ảnh hưởng đến kết quả học tập.

Bên cạnh việc kiểm định để phát hiện ra những nhóm khác biệt, chúng ta có thể tìm khoảng ước lượng cho chênh lệch giữa các nhóm có khác biệt có ý nghĩa thống kê. Ước lượng khoảng về chênh lệch giữa hai trung bình nhóm có khác biệt tính theo công thức:

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm \left(t_{n-k, \frac{\alpha}{2}} \sqrt{\frac{2MSW}{n_i}} \right)$$

Trong đó t là giá trị tra từ bảng phân phối Student t với $(n - k)$ bậc tự do.

Trong chương trình Excel không có các lệnh phân tích sâu ANOVA. Chúng ta có thể thực hiện phân tích này bằng chương trình SPSS. Ngoài ra kết quả của SPSS còn cung cấp cho các bạn một kiểm định chính thức về sự bằng nhau của các phương sai tổng thể là kiểm định Levene. (Bạn đọc có thể xem cách thức tiến hành kiểm tra giả định của phân tích ANOVA một yếu tố và phân tích sâu ANOVA trong sách Phân tích dữ liệu nghiên cứu với SPSS của cùng tác giả).

Phân tích phương sai với kiểm định F chỉ có thể áp dụng khi các nhóm so sánh có phân phối bình thường và phương sai bằng nhau. Trong trường hợp không thỏa điều kiện này, chúng ta có thể chuyển đổi dữ liệu của yếu tố kết quả từ dạng định lượng về dạng định tính (dữ liệu thứ bậc) và áp dụng một kiểm định phi tham số phù hợp tên là Kruskal – Wallis. Bạn đọc có thể tìm hiểu về kiểm định này ở Chương 10, Kiểm định phi tham số.

9.2 PHÂN TÍCH PHƯƠNG SAI HAI YẾU TỐ

Phân tích phương sai hai yếu tố (Two-way Analysis of Variance) xem xét cùng một lúc hai yếu tố nguyên nhân (dưới dạng dữ liệu định tính) ảnh hưởng đến yếu tố kết quả đang nghiên cứu (dưới dạng dữ liệu định lượng). Ví dụ như trong phân tích phương sai một yếu tố cho ta biết kết quả thời gian tự học ảnh hưởng đến kết quả học tập của sinh viên. Trường hợp này ta chưa nghiên cứu đến những điều kiện khác của sinh viên, ví dụ như mức độ yêu thích ngành học... Phân tích phương sai hai yếu tố sẽ giúp chúng ta đưa thêm yếu tố này vào trong phân tích, làm cho kết quả nghiên cứu càng có giá trị.

9.2.1 Trường hợp có một quan sát mẫu trong một ô

Giả sử chúng ta nghiên cứu ảnh hưởng của 2 yếu tố nguyên nhân định tính đến một yếu tố kết quả định lượng nào đó. Theo yếu tố nguyên nhân thứ nhất chúng ta có thể sắp xếp các đơn vị mẫu nghiên cứu thành K nhóm. Theo yếu tố nguyên nhân thứ hai ta có thể sắp xếp các đơn vị mẫu nghiên cứu thành H khối. Nếu đồng thời sắp xếp các đơn vị mẫu theo 2

yếu tố nguyên nhân này, ta sẽ có bảng kết hợp gồm K cột và H dòng, và bảng sẽ có K x H ô dữ liệu. Nếu chúng ta chỉ có 1 mẫu quan sát trong 1 ô thì tổng số đơn vị mẫu quan sát là $n = K \times H$. Dạng tổng quát của bảng này như sau:

Bảng 9.6 Quan sát mẫu của phân tích phương sai hai yếu tố.

Dòng (khối - blocks)	Cột (nhóm - groups)				
	1	2	3	K
1	x_{11}	x_{21}	x_{K1}
2	x_{12}	x_{22}		x_{K2}
.					...
.					...
H	x_{1H}	x_{2H}			x_{KH}

Để thực hiện (1) kiểm định giả thuyết cho rằng trung bình của K tổng thể tương ứng với K nhóm mẫu là bằng nhau, và (2) kiểm định giả thuyết cho rằng trung bình của H tổng thể tương ứng với H khối mẫu là bằng nhau, ta thực hiện theo các bước sau:

Bước 1: Tính các trung bình

Trung bình của riêng từng nhóm – group (cột)

$$\bar{x}_i = \frac{\sum_{j=1}^H x_{ij}}{H} \quad (i=1,2,\dots, K)$$

Trung bình riêng cho từng khối - block (dòng)

$$\bar{x}_j = \frac{\sum_{i=1}^K x_{ij}}{K} \quad (j=1,2,\dots, H)$$

Trung bình chung của toàn bộ mẫu quan sát:

$$\bar{x} = \frac{\sum_{i=1}^K \sum_{j=1}^H x_{ij}}{n} = \frac{\sum_{i=1}^K \bar{x}_i}{K} = \frac{\sum_{j=1}^H \bar{x}_j}{H}$$

Bước 2: tính tổng các chênh lệch bình phương

1. Tổng các chênh lệch bình phương chung: $SST = SSG + SSB + SSE$

$$SST = \sum_{i=1}^K \sum_{j=1}^H (x_{ij} - \bar{x})^2$$

SST phản ánh biến thiên của yếu tố định lượng kết quả đang nghiên cứu

do ảnh hưởng của tất cả các nguyên nhân.

2. Tổng các chênh lệch bình phương giữa các nhóm (between – groups)

$$SSG = H \sum_{i=1}^K (\bar{x}_i - \bar{x})^2$$

SSG phản ánh phần biến thiên của yếu tố định lượng kết quả đang nghiên cứu do ảnh hưởng của yếu tố nguyên nhân thứ nhất, yếu tố dùng để phân nhóm ở cột.

3. Tổng các chênh lệch bình phương giữa các khối (between – blocks)

$$SSB = K \sum_{j=1}^H (\bar{x}_j - \bar{x})^2$$

SSB phản ánh phần biến thiên của yếu tố định lượng kết quả đang nghiên cứu do ảnh hưởng của yếu tố nguyên nhân thứ hai, yếu tố dùng để phân nhóm ở dòng.

4. Tổng các chênh lệch bình phương phần dư (error)

$$SSE = \sum_{i=1}^K \sum_{j=1}^H (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2 = SST - SSG - SSB$$

SSE phản ánh phần biến thiên của yếu tố định lượng kết quả đang nghiên cứu do ảnh hưởng của các yếu tố khác còn lại không đưa vào nghiên cứu trong phân tích này.

Bước 3: Tính các phương sai:

1. Phương sai giữa các nhóm: $MSG = \frac{SSG}{K-1}$

2. Phương sai giữa các khối: $MSB = \frac{SSB}{H-1}$

3. Phương sai dư: $MSE = \frac{SSE}{(K-1)(H-1)}$

Bước 4: Kiểm định giả thuyết về ảnh hưởng của yếu tố nguyên nhân thứ nhất (cột) và yếu tố nguyên nhân thứ hai (dòng) đến yếu tố kết quả bằng các tỉ số F:

$$F_1 = \frac{MSG}{MSE}$$

$$F_2 = \frac{MSB}{MSE}$$

Bước 5: Có 2 trường hợp trong quyết định bác bỏ giả thuyết H_0 của ANOVA hai yếu tố:

- Đối với F_1 ở mức ý nghĩa α , giả thuyết H_0 cho rằng trung bình của K tổng thể theo yếu tố nguyên nhân thứ nhất (cột) bằng nhau bị bác bỏ khi:

$$F_1 > F_{K-1,(K-1)(H-1),\alpha}$$

2. Đối với F_2 ở mức ý nghĩa α , giả thuyết H_0 cho rằng trung bình của H tổng thể theo yếu tố nguyên nhân thứ hai (dòng) bằng nhau bị bác bỏ khi:

$$F_2 > F_{H-1,(K-1)(H-1),\alpha}$$

Trong đó:

- $F_{K-1,(K-1)(H-1),\alpha}$ là giá trị tra trong bảng phân phối F với K-1 bậc tự do ở tử số và (K-1)(H-1) bậc tự do ở mẫu số.
- $F_{H-1,(K-1)(H-1),\alpha}$ là giá trị tra trong bảng phân phối F với H-1 bậc tự do ở tử số và (K-1)(H-1) bậc tự do ở mẫu số.

Thường phân tích phương sai hai yếu tố được thực hiện trên chương trình máy tính (Excel hoặc SPSS). Kết quả có dạng tổng quát như sau:

Bảng 9.7 Bảng kết quả tổng quát ANOVA hai yếu tố

Nguồn biến thiên	Tổng các chênh lệch bình phương	Bậc tự do	Phương sai	Tỉ số F
Giữa các nhóm	SSG	K - 1	$MSG = \frac{SSG}{K - 1}$	$F_1 = \frac{MSG}{MSE}$
Giữa các khối	SSB	H - 1	$MSB = \frac{SSB}{H - 1}$	$F_2 = \frac{MSB}{MSE}$
Phản dư	SSE	(K-1)x(H-1)	$MSE = \frac{SSE}{(K - 1)(H - 1)}$	
Tổng cộng	SST	n-1		

9.2.2 Trường hợp có nhiều quan sát trong một ô

Để tăng tính chính xác khi kết luận về ảnh hưởng của hai yếu tố nguyên nhân đến yếu tố kết quả của mẫu cho một tổng thể, ta tăng cỡ mẫu quan sát trong điều kiện cho phép. Gọi L là số quan sát trong một ô, ta có dạng tổng quát của L quan sát trong một ô như sau:

Bảng 9.8 Bảng dữ liệu quan sát mẫu ANOVA 2 yếu tố (nhiều quan sát)

Dòng (blocks)	Nhóm (groups)			
	1	2	...	K
1	$x_{111} x_{112} \dots x_{11L}$	$x_{211} x_{212} \dots x_{21L}$...	$x_{K11} x_{K12} \dots x_{K1L}$
2	$x_{121} x_{122} \dots x_{12L}$	$x_{221} x_{222} \dots x_{22L}$...	$x_{K21} x_{K22} \dots x_{K2L}$
..				
H	$x_{1H1} x_{1H2} \dots x_{1HL}$	$x_{2H1} x_{2H2} \dots x_{2HL}$...	$x_{KH1} x_{KH2} \dots x_{KHL}$

Một ô

Bước 1: Tính các trung bình

Trung bình mẫu của từng nhóm – group (cột)

$$\bar{x}_i = \frac{\sum_{j=1}^H \sum_{s=1}^L x_{ijs}}{H \times L} \quad (i = 1, 2, \dots, K)$$

Trung bình mẫu của từng khối - block (dòng)

$$\bar{x}_j = \frac{\sum_{i=1}^K \sum_{s=1}^L x_{ijs}}{K \times L} \quad (j = 1, 2, \dots, H)$$

Trung bình mẫu của từng ô

$$\bar{x}_{ij} = \frac{\sum_{s=1}^L x_{ijs}}{L}$$

Trung bình chung của toàn bộ mẫu quan sát:

$$\bar{x} = \frac{\sum_{i=1}^K \sum_{j=1}^H \sum_{s=1}^L x_{ijs}}{K \times H \times L}$$

Bước 2: tính tổng các chênh lệch bình phương

1. Tổng các chênh lệch bình phương toàn bộ:

$$SST = SSG + SSB + SSI + SSE$$

$$SST = \sum_{i=1}^K \sum_{j=1}^H \sum_{s=1}^L (x_{ijs} - \bar{x})^2$$

2. Tổng các chênh lệch bình phương giữa các nhóm: between – groups

$$SSG = HL \sum_{i=1}^K (\bar{x}_i - \bar{x})^2$$

SSG phản ánh phần biến thiên của yếu tố định lượng kết quả đang nghiên cứu do ảnh hưởng của yếu tố nguyên nhân thứ nhất, yếu tố dùng để phân nhóm ở cột.

3. Tổng các chênh lệch bình phương giữa các khối: between – blocks

$$SSB = KL \sum_{j=1}^H (\bar{x}_j - \bar{x})^2$$

SSB phản ánh phần biến thiên của yếu tố định lượng kết quả đang nghiên

cứu do ảnh hưởng của yếu tố nguyên nhân thứ hai, yếu tố dùng để phân nhóm ở dòng.

4. Tổng các chênh lệch bình phương giữa các ô (giao nhau giữa các nhóm và khối)

$$SSI = L \sum_{i=1}^K \sum_{j=1}^H (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$$

SSI phản ánh phần biến thiên do tác động qua lại giữa hai yếu tố đang nghiên cứu.

5. Tổng các chênh lệch bình phương phần dư:

$$SSE = \sum_{i=1}^K \sum_{j=1}^H \sum_{s=1}^L (x_{ijs} - \bar{x}_{ij})^2 = SST - SSG - SSB - SSI$$

Bước 3: Tính các phương sai:

1. Phương sai giữa các nhóm: $MSG = \frac{SSG}{K-1}$

2. Phương sai giữa các khối: $MSB = \frac{SSB}{H-1}$

3. Phương sai giữa các ô $MSI = \frac{SSI}{(K-1) \times (H-1)}$

4. Phương sai dư: $MSE = \frac{SSE}{K \times H \times (L-1)}$

Bước 4: Kiểm định giả thuyết về ảnh hưởng của yếu tố nguyên nhân thứ nhất (cột), yếu tố nguyên nhân thứ hai (dòng), tương tác giữa hai yếu tố đến yếu tố kết quả bằng các tỉ số F:

$$F_1 = \frac{MSG}{MSE} \quad F_2 = \frac{MSB}{MSE} \quad F_3 = \frac{MSI}{MSE}$$

Bước 5: Nguyên tắc quyết định trong ANOVA hai yếu tố:

1. Đối với F_1 ở mức ý nghĩa α , giả thuyết H_0 cho rằng trung bình của k tổng thể theo yếu tố nguyên nhân thứ nhất (cột) bằng nhau bị bác bỏ khi:

$$F_1 > F_{K-1, KH(L-1), \alpha}$$

2. Đối với F_2 ở mức ý nghĩa α , giả thuyết H_0 cho rằng trung bình của H tổng thể theo yếu tố nguyên nhân thứ hai (dòng) bằng nhau bị bác bỏ khi:

$$F_2 > F_{H-1, KH(L-1), \alpha}$$

3. Đối với F_3 ở mức ý nghĩa α , giả thuyết H_0 cho rằng không có tác động qua lại giữa yếu tố thứ nhất (cột) và yếu tố thứ hai (dòng) bị bác bỏ khi:

$$F_3 > F_{(K-1)(H-1), KH(L-1), \alpha}$$

Trong đó:

- $F_{K-1, KH(L-1), \alpha}$ là giá trị tra trong bảng phân phối F với K-1 bậc tự do ở tử số và KH(L-1) bậc tự do ở mẫu số.
- $F_{H-1, KH(L-1), \alpha}$ là giá trị tra trong bảng phân phối F với H-1 bậc tự do ở tử số và KH(L-1) bậc tự do ở mẫu số.
- $F_{(K-1)(H-1), KH(L-1), \alpha}$ là giá trị tra trong bảng phân phối F với (K-1)(H-1) bậc tự do ở tử số và KH(L-1) bậc tự do ở mẫu số.

Ví dụ: cũng từ ví dụ điểm trung bình học tập và thời gian tự học của sinh viên, chúng ta đưa thêm vào yếu tố mức độ yêu thích ngành đang học của sinh viên. Dữ liệu thu thập được trình bày trong Bảng 9.9 sau đây.

Bảng 9.9: Điểm trung bình học tập của sinh viên phân nhóm theo thời gian tự học và mức độ yêu thích ngành học

Mức độ yêu thích ngành học	Thời gian tự học		
	Ít giờ	TB	Nhiều giờ
Không thích lăm	5,8	6,0	6,2
	6,2	6,6	5,8
	5,4	6,1	6,5
	6,0	5,8	6,2
	5,2	5,9	6,4
	5,3	6,0	5,7
	5,4	5,9	6,1
	5,6	6,0	6,8
Thích	6,2	6,7	7,1
	5,7	6,5	6,5
	5,5	6,3	7,1
	6,1	6,1	7,2
	6,0	6,8	6,7
	5,2	6,4	7,0
	6,4	6,8	7,6
	5,5	6,6	7,7
Rất thích	5,0	6,4	7,8
	5,6	6,2	6,8
	6,2	7,1	7,3
	6,1	7,0	7,1
	5,3	7,2	7,2

Các giả thuyết H₀ đặt ra:

- Điểm trung bình học tập (ĐTB) của sinh viên có thời gian tự học khác nhau đều bằng nhau.
- ĐTB của sinh viên có mức độ yêu thích ngành đang học khác nhau đều bằng nhau.
- Không có ảnh hưởng tương tác giữa thời gian tự học và mức độ yêu thích ngành đang học của sinh viên. Nói một cách cụ thể, ảnh hưởng của thời gian tự học đến ĐTB là như nhau đối với các nhóm sinh viên có mức độ yêu thích ngành đang học khác nhau; và ảnh hưởng của mức độ yêu thích ngành đang học đến ĐTB là như nhau đối với các nhóm sinh viên có thời gian tự học khác nhau.

Bước 1: Tính các trung bình

- Trung bình mẫu của từng nhóm (group means):

$$\bar{x}_i = \frac{\sum_{j=1}^H \sum_{s=1}^L x_{ijs}}{H \times L}$$

ĐTB của nhóm thời gian tự học ít

$$\bar{x}_1 = \frac{5,8 + 6,2 + 5,4 + \dots + 6,2 + 6,1 + 5,3}{3 \times 7} = 5,7$$

ĐTB của nhóm thời gian tự học trung bình

$$\bar{x}_2 = \frac{6 + 6,6 + 6,1 + \dots + 7,1 + 7 + 7,2}{3 \times 7} = 6,4$$

ĐTB của nhóm thời gian tự học nhiều

$$\bar{x}_3 = \frac{6,2 + 5,8 + 6,5 + \dots + 7,3 + 7,1 + 7,2}{3 \times 7} = 6,8$$

- Trung bình mẫu của từng khối (block means)

$$\bar{x}_j = \frac{\sum_{i=1}^K \sum_{s=1}^L x_{ijs}}{K \times L}$$

ĐTB của nhóm không yêu thích ngành học lăm

$$\bar{x}_1 = \frac{5,8 + 6 + 6,2 + \dots + 5,4 + 5,9 + 6,1}{3 \times 7} = 5,93$$

ĐTB của nhóm yêu thích ngành học

$$\bar{x}_2 = \frac{5,6 + 6 + 6,8 + \dots + 5,2 + 6,4 + 7}{3 \times 7} = 6,36$$

ĐTB của nhóm rất yêu thích ngành học

$$\bar{x}_3 = \frac{6,4 + 6,8 + 7,6 + \dots + 5,3 + 7,2 + 7,2}{3 \times 7} = 6,6$$

- Trung bình một ô (cell means)

$$\bar{x}_{ij} = \frac{\sum_{s=1}^L x_{ijs}}{L}$$

ĐTB của SV có thời gian tự học ít và không yêu thích ngành học là

$$\bar{x}_{11} = \frac{5,8 + 6,2 + 5,4 + 6 + 5,2 + 5,3 + 5,4}{7} = 5,61$$

Tính tương tự cho các khối còn lại gồm: Trung bình - không thích là: 6,04 và nhiều-không thích là: 6,13

ĐTB của SV có thời gian tự học trung bình và yêu thích ngành học:

$$\bar{x}_{22} = \frac{6 + 6,7 + 6,5 + 6,3 + 6,1 + 6,8 + 6,4}{7} = 6,4$$

Tính tương tự cho các khối còn lại gồm: ít – thích là: 5,76 và nhiều - thích là: 6,91

ĐTB của SV có thời gian tự học nhiều và rất yêu thích ngành học

$$\bar{x}_{33} = \frac{7,6 + 7,7 + 7,8 + 6,8 + 7,3 + 7,1 + 7,2}{7} = 7,36$$

Tính tương tự cho các khối còn lại gồm: ít – rất thích là :5,73 và trung bình – rất thích là : 6,76

- Trung bình chung (overall mean):

$$\bar{x} = \frac{\sum_{i=1}^K \sum_{j=1}^H \sum_{s=1}^L x_{ijs}}{K \times H \times L}$$

$$\bar{x} = \frac{5,8 + 6 + 6,2 + 6,2 + 6,6 \dots 7 + 7,1 + 5,3 + 7,2 + 7,2}{3 \times 3 \times 7} = 6,3$$

Để đơn giản ta có thể tính trung bình chung theo công thức như dưới đây với điều kiện số quan sát trong mỗi nhóm đều bằng nhau.

$$\bar{x} = \frac{\sum_{i=1}^K \bar{x}_i}{K} \quad (\text{Tổng các trung bình nhóm chia cho số nhóm})$$

Kết quả tính các trung bình được trình bày tóm tắt trong Bảng 9.10.

Bước 2: Tính các tổng chênh lệch bình phương (SS)

1. Tổng các chênh lệch bình phương toàn bộ: SST

$$\begin{aligned} SST &= \sum_{i=1}^K \sum_{j=1}^H \sum_{s=1}^L (x_{ijs} - \bar{x})^2 \\ &= (5,8-6,3)^2 + (6,2-6,3)^2 + (5,4-6,3)^2 + (6-6,3)^2 + (5,2-6,3)^2 + (5,3-6,3)^2 + \\ &\quad (5,4-6,3)^2 + (6-6,3)^2 + (6,6-6,3)^2 + (6,1-6,3)^2 + (5,8-6,3)^2 + (5,9-6,3)^2 + \\ &\quad (6-6,3)^2 + (5,9-6,3)^2 + \dots + (6,8-6,3)^2 + (6,6-6,3)^2 + (6,4-6,3)^2 + (6,2-6,3)^2 + \\ &\quad (7,1-6,3)^2 + (7-6,3)^2 + (7,2-6,3)^2 + (7,6-6,3)^2 + (7,7-6,3)^2 + (7,8-6,3)^2 + \\ &\quad (6,8-6,3)^2 + (7,3-6,3)^2 + (7,1-6,3)^2 + (7,2-6,3)^2 \\ &= 27,02 \end{aligned}$$

2. Tổng các chênh lệch bình phương bình phương giữa các nhóm (between – groups)

$$\begin{aligned} SSG &= HL \sum_{i=1}^K (\bar{x}_i - \bar{x})^2 = 3 \times 7 \times [(5,7-6,3)^2 + (6,4-6,3)^2 + (6,8-6,3)^2] \\ &= 21 \times 0,62 = 13,02 \end{aligned}$$

3. Tổng các chênh lệch bình phương giữa các khối (between – blocks)

$$\begin{aligned} SSB &= KL \sum_{j=1}^H (\bar{x}_j - \bar{x})^2 = 3 \times 7 \times [(5,93-6,3)^2 + (6,36-6,3)^2 + (6,6-6,3)^2] \\ &= 21 \times 0,2305 = 4,84 \end{aligned}$$

4. Tổng các chênh lệch bình phương giữa các ô

$$SSI = L \sum_{i=1}^K \sum_{j=1}^H (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$$

Để dễ theo dõi, trước khi tính toán SSI ta gom các trung bình đã tính được vào một bảng như bảng sau:

Bảng 9.10: Bảng tóm tắt kết quả tính các trung bình

		Thời gian tự học			\bar{x}_j
		Ít	Trung bình	Nhiều	
Yếu thích	Không	5,61	6,04	6,13	5,93
	Thích	5,76	6,4	6,91	6,36
	Rất	5,73	6,76	7,36	6,60
	\bar{x}_i	5,70	6,40	6,80	6,30

$$\begin{aligned} SSI &= 7 \times [(5,61-5,7-5,93+6,3)^2 + (5,76-5,7-6,36+6,3)^2 + (5,73-5,7-6,6+6,3)^2 + \dots + (6,13-6,8-5,93+6,3)^2 + (6,91-6,8-6,36-6,3)^2 + (7,36-6,8-6,6-6,3)^2] \\ &= 7 \times 0,3187 = 2,23 \end{aligned}$$

5. Tổng các chênh lệch bình phương phần dư:

$$SSE = \sum_{i=1}^K \sum_{j=1}^H \sum_{s=1}^L (x_{ijs} - \bar{x}_{ij})^2 = SST - SSG - SSB - SSI$$

$$SSE = 27,02 - 13,02 - 4,84 - 2,23 = 6,93$$

Bước 3: Tính các phương sai:

$$1. \text{ Phương sai giữa các nhóm: } MSG = \frac{SSG}{K-1} = \frac{13,02}{3-1} = 6,51$$

$$2. \text{ Phương sai giữa các khối: } MSB = \frac{SSB}{H-1} = \frac{4,84}{3-1} = 2,42$$

3. Phương sai giữa các ô:

$$MSI = \frac{SSI}{(K-1)(H-1)} = \frac{2,23}{(3-1)(3-1)} = 0,558$$

4. Phương sai dư:

$$MSE = \frac{SSE}{KH(L-1)} = \frac{6,93}{3 \times 3 \times (7-1)} = 0,128$$

Bước 4: Tính tỉ số F

$$1. F_1 = \frac{MSG}{MSE} = \frac{6,51}{0,128} = 50,86$$

$$2. F_2 = \frac{MSB}{MSE} = \frac{2,42}{0,128} = 18,91$$

$$3. F_3 = \frac{MSI}{MSE} = \frac{0,558}{0,128} = 4,36$$

Tra bảng F tìm

$$F_{K-1; KH(L-1); \alpha} = F_{3-1; 3 \times 3(7-1); 0,05} = F_{2; 54; 0,05} = 3,17$$

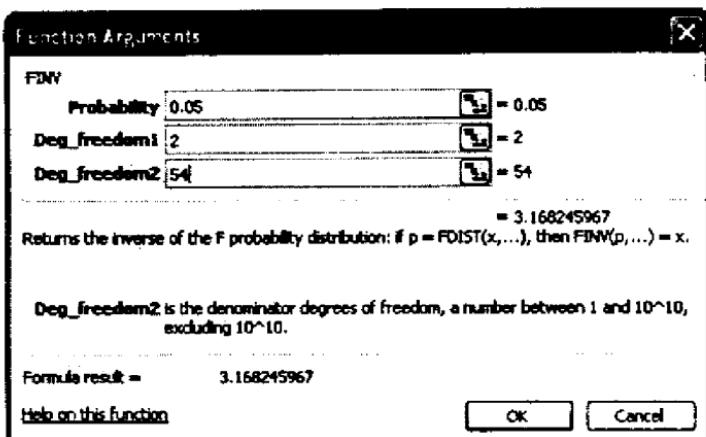
$$F_{H-1; KH(L-1); \alpha} = F_{3-1; 3 \times 3(7-1); 0,05} = F_{2; 54; 0,05} = 3,17$$

$$F_{(K-1)(H-1); KH(L-1); \alpha} = F_{(3-1)(3-1); 3 \times 3(7-1); 0,05} = F_{4; 54; 0,05} = 2,54$$

Cách tra giá trị tối hạn F bằng Excel:

Nếu các bậc tự do của bài toán vượt quá giá trị cho trong các bảng tra thì bạn đọc có thể vào menu Insert/Function chọn lệnh Finv trong nhóm hàm Statistical và khai báo như sau trong cửa sổ hàm. Nhấp nút OK ta có giá trị tối hạn cần tìm.

Hình 9.4



- Vì $F_1=50,86 > F_{2,54;0,05}$ nên chúng ta có đủ bằng chứng để bác bỏ giả thuyết thứ nhất. Như vậy ĐTB của sinh viên có thời gian tự học khác nhau thì không bằng nhau. Nói cách khác, thời gian tự học có ảnh hưởng đến kết quả học tập.
- Vì $F_2=18,91 > F_{2,54;0,05}$ nên chúng ta có đủ bằng chứng để bác bỏ giả thuyết thứ hai. Như vậy ĐTB của sinh viên có mức độ yêu thích ngành học khác nhau thì không bằng nhau. Nói cách khác, mức độ yêu thích ngành học của sinh viên có ảnh hưởng đến kết quả học tập.
- Vì $F_3=4,36 > F_{4,54;0,05}$, nên chúng ta có đủ bằng chứng để bác bỏ giả thuyết thứ ba. Như vậy có tương tác giữa thời gian tự học và mức độ yêu thích ngành học trong việc ảnh hưởng đến ĐTB của sinh viên. Mức độ ảnh hưởng của thời gian tự học đến kết quả học tập còn bị ảnh hưởng bởi mức độ yêu thích ngành học. Trong Bảng 9.10, chúng ta thấy khi mức độ yêu thích ngành học ít thì thời gian tự học ít ảnh hưởng đến kết quả học tập. Nhưng khi mức độ yêu thích ngành học cao thì ảnh hưởng của thời gian tự học đến kết quả học tập tăng giữa các nhóm sinh viên có thời gian tự học khác nhau.

Trong thực tế, khối lượng tính toán khi sử dụng ANOVA, nhất là ANOVA 2 yếu tố, khá lớn, người ta thường sử dụng các chương trình máy tính như Excel và SPSS để ra kết quả nhanh chóng. Khi thực hiện bằng Excel, bên cạnh các kết quả tính toán trung bình, chúng ta được bảng cuối cùng là bảng kiểm định F trong ANOVA có nội dung cơ bản như sau:

Bảng 9.11: Bảng ANOVA hai yếu tố tổng quát

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
Sample				
Columns				
Interaction				
Within				
Total				

Để đọc được các kết quả này mà không nhầm lẫn bạn đọc cần nắm được ý nghĩa các thuật ngữ thống kê Excel dùng trong trình bày kết quả, tạm dịch như sau:

Nguồn biến thiên	Tổng chênh lệch bình phương	Bậc tự do	Phương sai	Tỉ số <i>F</i>
Giữa các khối	SSB	(H-1)	MSB	F_2
Giữa các nhóm	SSG	(K-1)	MSG	F_1
Tương tác giữa 2 yếu tố	SSI	(K-1)(H-1)	MSI	F_3
Phản dư	SSE	KH (L-1)	MSE	
Tổng cộng	SST	KHL -1		

Kết quả ANOVA đầy đủ cho ví dụ trên thực hiện trên Excel được trình bày trong Bảng 9.12. Các bạn có thể so sánh các kết quả trung bình tính thủ công với kết quả do Excel tính toán (nền xám).

Bảng 9.12 : Kết quả phân tích phương sai 2 yếu tố bằng Excel**Anova: Two-Factor With Replication**

SUMMARY	ít giờ	trung bình	Nhiều giờ	Total
		không thích lắm		
Count	7	7	7	21
Sum	39.3	42.3	42.9	124.5
Average	5.6143	6.0429	6.1286	5.8286
Variance	0.1481	0.0695	0.0857	0.1441

thích				
Count	7	7	7	21
Sum	40.3	44.8	48.4	133.5
Average	5.7571	6.4000	6.9143	6.3571
Variance	0.1295	0.0867	0.0648	0.3196

rất thích				
Count	7	7	7	21
Sum	40.1	47.3	51.5	138.9
Average	5.7286	6.7571	7.3571	6.6353
Variance	0.2657	0.1395	0.1295	0.6353

xem tiếp →

Total

Count	21	21	21
Sum	119.7	134.4	142.8
Average	5.7	6.4	6.8
Variance	0.167	0.178	0.355

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Sample	5.04	2	2.52	20.2672	0.0000	3.1682
Columns	13.02	2	6.51	52.3570	0.0000	3.1682
Interaction	2.2457	4	0.5614	4.5153	0.0032	2.5429
Within	6.7143	54	0.1243			
Total	27.02	62				

Đối chiếu các kết quả trong bảng ANOVA với kết quả tính thủ công chúng ta thấy có sai số do khi tính thủ công chúng ta làm tròn số nhiều hơn.

Khi thực hiện ANOVA trên Excel, trong bảng kết quả ta có thêm cột p-value và F crit. Cột F crit chính là giá trị tối hạn tra từ bảng thống kê (với mức ý nghĩa của bài toán kiểm định do ta khai báo trong lúc tiến hành lệnh ANOVA) dùng để so sánh với cột "F" để quyết định bác bỏ giả thuyết H_0 hay không. Bên cạnh đó ta có thể dùng luôn kết quả của cột p-value để quyết định bác bỏ H_0 hay không theo quy tắc $p\text{-value} < \alpha \rightarrow$ bác bỏ H_0 .

9.2.3 Phân tích sâu trong ANOVA 2 yếu tố

Trong phân tích phương sai 2 yếu tố sau khi đã xác định có sự khác biệt giữa các nhóm so sánh, chúng ta có thể dùng kiểm định Tukey để xác định các cặp trung bình tổng thể khác nhau xét theo yếu tố thứ nhất (so sánh giữa K nhóm) hay xét theo yếu tố thứ hai (so sánh giữa H khối). Kiểm định Tukey vẫn được thực hiện theo nguyên tắc giống như phần trước, với giá trị giới hạn Tukey được tính như sau:

So sánh theo yếu tố thứ nhất (K nhóm): $T = q_{\alpha, K, KH(L-1)} \sqrt{\frac{MSE}{H \times L}}$

So sánh theo yếu tố thứ hai (H khối): $T = q_{\alpha, H, KH(L-1)} \sqrt{\frac{MSE}{K \times L}}$

Vận dụng vào ví dụ tính toán trong phần phân tích phương sai 2 yếu tố trên, với $\alpha = 0,05$, $K = 3$, $H = 3$, $L = 7$, $MSE = 0,128$ tra bảng phân phối kiểm định Tukey (Bảng tra số 9) ta có:

$$q_{\alpha;K;KH(L-1)} = q_{0,05;3;54} = 3,4$$

* so sánh giữa các nhóm theo yếu tố thứ nhất (thời gian tự học): chúng ta tính giá trị giới hạn Tukey:

$$T = q_{0,05;3;54} \sqrt{0,128/21} = 3,4 \times 0,078 = 0,265$$

Ta có các trung bình nhóm lần lượt là: 5,7; 6,4; 6,8 và các chênh lệch giữa các nhóm là:

$$D_{lt,trung\ binh} = |5,7-6,4|=0,7$$

$$D_{lt,nhiều} = |5,7-6,8|=1,1$$

$$D_{trung\ binh,nhiều} = |6,4-6,8|=0,4$$

Ta thấy các chênh lệch đều lớn hơn giá trị giới hạn Tukey T, cho nên chúng ta có thể nói rằng sinh viên có thời gian tự học khác nhau có điểm trung bình học tập khác nhau. Theo giá trị trung bình mẫu thì ta kết luận được thời gian tự học càng tăng, điểm trung bình học tập càng cao.

* So sánh giữa các nhóm theo yếu tố thứ hai (mức độ liên quan giữa việc làm thêm và ngành học): chúng ta tính giá trị giới hạn Tukey:

$$q_{\alpha;H;KH(L-1)} = q_{0,05;3;54} = 3,4$$

Ta có các trung bình nhóm lần lượt là: 5,93 ; 6,36 ; 6,6 và các chênh lệch giữa các nhóm là:

$$D_{không,thích} = |5,93-6,36|=0,43$$

$$D_{không,rất\ thích} = |5,93-6,6|=0,67$$

$$D_{thích,rất\ thích} = |6,36-6,6|=0,24$$

Ta thấy chỉ có chênh lệch giữa nhóm thích và rất thích $D_{thích,rất\ thích}$ bé hơn giá trị giới hạn Tukey T, cho nên chúng ta có thể nói rằng các nhóm sinh viên có mức độ yêu thích ngành học nhiều hay rất nhiều thì có kết quả học tập không khác biệt nhau đáng kể. Riêng nhóm không thích ngành mình đang học có kết quả học tập kém hơn hẳn hai nhóm thích và rất thích ngành đang học.

9.2.4 Thực hiện ANOVA trên chương trình Excel

Chúng ta có thể sử dụng Excel để giải quyết phân tích ANOVA. Chương trình bảng tính Excel khá đa năng nên những xử lý thống kê rất hạn chế và đơn giản. Vì vậy, nếu nguồn dữ liệu lớn và xử lý thống kê phức tạp hơn, chúng ta nên dùng chương trình SPSS. Chúng ta cũng cần làm quen trước từ chuyên môn bằng tiếng Anh trong thống kê để có thể dễ dàng hiểu bảng kết quả xử lý. Phần này chỉ giới thiệu thao tác thực hiện ANOVA trên phần mềm Excel cho cả hai trường hợp ANOVA một yếu tố và hai yếu tố.

Bước 1: mở chương trình Excel, và nhập dữ liệu. Đối với ANOVA 2 yếu tố có nhiều quan sát trong một ô, cần chú ý nhập liệu không giống như Bảng 9.9, mà phải nhập như Dữ liệu trong Hình 9.5. Nếu không, chương trình sẽ không thực hiện được hoặc cho ra kết quả sai. Kiểm tra cột bậc tự do để biết kết quả chương trình xuất ra đúng hay sai.

Bước 2: Chọn Tool – Data Analysis, chúng ta có các lựa chọn sau:

ANOVA: Single Factor. Phân tích phương sai một yếu tố.

ANOVA: Two-factor without replication. Phân tích phương sai hai yếu tố với một quan sát trong một ô

ANOVA: Two-factor with replication. Phân tích phương sai hai yếu tố với nhiều quan sát trong một ô.

Bước 3: Chọn vùng số liệu vừa mới nhập. Chú ý, khi chọn vùng số liệu thì chọn cả phần chữ (tiêu đề cột và tiêu đề dòng) và phần dữ liệu

Bước 4: Tiến hành các khai báo như trong hình sau

- α mặc nhiên là 5%
- Chọn vùng chứa kết quả, nếu bạn chọn New Worksheet thì kết quả được đặt trong một trang mới với đầy đủ các thông tin như được trình bày trong Bảng 9.12
- Nhấn phím “OK”

Hình 9.5

	A	B	C	D	E	F	G	H
1	ít giờ	trung bình	nhiều giờ					
2	không thích	5.8	6.0	6.2				
3		6.2	6.6	5.8				
4		5.4	6.1					
5		6.0	5.8					
6		5.2	5.9					
7		5.3	6.0					
8		5.4	5.9					
9	thích	5.6	6.0					
10		6.2	6.7					
11		5.7	6.5					
12		5.5	6.3					
13		6.1	6.1					
14		6.0	6.8					
15		5.2	6.4					
16	rất thích	6.4	6.8	7.6				
17		5.5	6.6	7.7				

Anova: Two-Factor With Replication

Input Range: \$A\$1:\$F\$17

Rows per sample: 7

Alpha: 0.05

Output options:

Output Range: \$G\$1:\$H\$17

New Worksheet By:

New Workbook

OK Cancel Help

CHƯƠNG 10

KIỂM ĐỊNH PHI THAM SỐ

Khi phân tích không phải lúc nào bạn cũng gặp được các tình huống thỏa mãn hoàn toàn các giả định cần thiết cho các kiểm định đã nghiên cứu, đặc biệt khi bạn chỉ có các mẫu nhỏ. Lúc này bạn phải dùng những kiểm định đòi hỏi những giả định ít nghiêm ngặt hơn về phân phối của dữ liệu, những thủ tục này được gọi là kiểm định với phân phối bất kỳ hay còn gọi là kiểm định phi tham số (Nonparametric test).

Nhược điểm của kiểm định phi tham số là khả năng tìm ra được những sai biệt thật sự của chúng kém hơn trong những trường hợp mà các giả định của thủ tục kiểm định có tham số (Parametric test) được thoả mãn. Nói cách khác kiểm định phi tham số không mạnh như những kiểm định có tham số vì nó bỏ qua một số thông tin có giá trị. Như vậy kiểm định phi tham số chỉ hữu dụng cho những trường hợp chúng ta không thể sử dụng các kiểm định tham số như với tình huống tổng thể không đảm bảo giả định là có phân phối bình thường. Các kiểm định phi tham số cũng hữu dụng khi mẫu có những giá trị quan sát bất thường (outliers) vì những giá trị nằm xa trung tâm này sẽ không gây ảnh hưởng lớn đến kết quả như khi chúng được sử dụng trong các thủ tục kiểm định căn cứ trên những tham số thống kê dễ bị ảnh hưởng như trung bình (vì gắn liền với những tham số nên chúng mới có tên là kiểm định tham số).

Kiểm định phi tham số cũng phù hợp trong các trường hợp dữ liệu hiện có của chúng ta là loại dữ liệu định danh (nominal) hay dữ liệu thứ bậc (ordinal).

Bảng sau liệt kê một số kiểm định phi tham số và kiểm định tham số tương ứng

Phi tham số	Tham số
Kiểm định dấu và hạng Wilcoxon (Wilcoxon signed rank test)	Kiểm định giả thuyết về trị trung bình tổng thể Kiểm định sự bằng nhau của 2 trị trung bình trong trường hợp mẫu phối hợp từng cặp (Paired-Samples t test)
Kiểm định tổng hạng Wilcoxon (Wilcoxon rank sum test)	Kiểm định sự bằng nhau của 2 trị trung bình trong trường hợp 2 mẫu độc lập Independent-Samples t test
Kiểm định Kruskal-Wallis	Phân tích ANOVA

Ngoài ra chúng ta còn thảo luận thêm về kiểm định kiểm định Chi-bình phương để kiểm định giả thuyết về phân phối của tổng thể và kiểm định giả thuyết về mối liên hệ (hay tính độc lập).

10.1 KIỂM ĐỊNH DẤU VÀ HẠNG WILCOXON VỀ TRUNG VỊ CỦA MỘT TỔNG THỂ

Chúng ta đã tìm hiểu thủ tục kiểm định về một giá trị tổng thể đơn như trung bình tổng thể, nếu dữ liệu của chúng ta là dữ liệu dạng khoảng cách hay tỷ lệ, hoặc tổng thể có phân phối bình thường thì chúng ta mới dùng thống kê z hoặc t để kiểm định giả thuyết về giá trị của trung bình tổng thể, nếu không đáp ứng được các giả định này thì thống kê t hoặc z không phù hợp, chúng ta sẽ dùng một kiểm định phi tham số tên là kiểm định dấu và hạng Wilcoxon, kiểm định này không đòi hỏi về hình dáng phân phối của tổng thể. Trong nội dung này ta nghiên cứu kiểm định dấu và hạng Wilcoxon cho một tham số tổng thể đơn, nội dung sau chúng ta sẽ áp dụng nó để kiểm định trên hai tổng thể có quan hệ.

Khác với kiểm định t hoặc z về trị trung bình tổng thể, kiểm định Wilcoxon kiểm định về trung vị của tổng thể. Logic của phương pháp kiểm định này là: do trung vị là giá trị chính giữa trong một tổng thể nên chúng ta kì vọng có một nửa các quan sát của mẫu sẽ nằm dưới giá trị trung vị tổng thể này và một nửa sẽ nằm ở trên. Giá trị trung vị giả thuyết sẽ bị loại nếu dữ liệu thực sự trong tập dữ liệu phân bố quá khác định hướng này.

Kiểm định dấu và hạng Wilcoxon được chia ra 2 tình huống là kiểm định với cỡ mẫu nhỏ ($n \leq 20$) và kiểm định với cỡ mẫu lớn ($n > 20$)

Chúng ta tìm hiểu kiểm định này khi cỡ mẫu nhỏ qua một ví dụ cụ thể như sau:

Giám đốc trung tâm hỗ trợ việc làm của một trường đại học muốn làm kiểm tra để xác định giá trị trung vị của phân phối thu nhập của sinh viên tốt nghiệp sau 2 năm làm việc ở khu vực có vốn đầu tư nước ngoài có vượt quá con số 350 đô la hay không. Người ta vẫn thường tin rằng phân phối thu nhập là một phân phối lệch phải vì thế ông giám đốc không muốn sử dụng những kiểm định tham số thông thường, thay vào đó ông ta dùng kiểm định Wilcoxon, ông chọn 10 sinh viên cũ của trường để tiến hành nghiên cứu này. Và chọn mức ý nghĩa là 5%

Theo định hướng của cuộc kiểm tra chúng ta đặt giả thuyết như sau:

$$H_0: \text{trung vị} \leq 350$$

$$H_1: \text{trung vị} > 350$$

Chú ý là các giả thuyết cho dạng kiểm định này cũng rất đa dạng, có thể là 2 bên hoặc 1 bên, giả sử gọi giá trị trung vị đang xét là M_{e_0} thì ta có thể có các dạng tổng quát như:

$$H_0: \text{trung vị} = M_{e_0}$$

$$H_1: \text{trung vị} \neq M_{e_0}$$

$$H_0: \text{trung vị} \leq M_{e_0}$$

$$H_1: \text{trung vị} > M_{e_0}$$

$$H_0: \text{trung vị} \geq M_{e_0}$$

$$H_1: \text{trung vị} < M_{e_0}$$

Hoặc ta có thể đặt giả thuyết so sánh hai giá trị trung vị của hai tổng thể với nhau, quy trình tiến hành cũng tương tự.

Thủ tục kiểm định sẽ đi qua các bước sau

- **Bước 1:** thu thập thông tin mẫu
- **Bước 2:** tính toán chênh lệch d_i giữa từng giá trị quan sát được và giá trị trung vị giả thuyết hoặc chênh lệch giữa giá trị quan sát được trên hai mẫu (chọn thứ tự và đặt phép trừ nhất quán theo giả thuyết đã đặt).
- **Bước 3:** lấy trị tuyệt đối của chênh lệch
- **Bước 4:** xếp hạng từng d_i , quy ước giá trị d_i nhỏ nhất có hạng là 1, các $d_i = 0$ không tham gia vào quá trình xếp hạng. Nếu các d_i có giá trị ngang nhau thì tính hạng trung bình cho tất cả các quan sát có giá trị d_i bằng nhau này.
- **Bước 5:** với các giá trị d_i dương thì ta đặt hạng của nó vào cột R+, với các giá trị d_i âm thì đặt hạng của nó vào cột kí hiệu R-
- **Bước 6:** tính giá trị thống kê W theo quy tắc

Nếu kiểm định 2 bên thì W được xác định là tổng hạng nhỏ hơn tức là: $W = \min[\sum(\text{cột R+}); \sum(\text{cột R-})]$

Nếu kiểm định 1 bên thì nếu kiểm định bên phải $W = \sum(\text{cột R+})$

Nếu kiểm định 1 bên thì nếu kiểm định bên trái $W = \sum(\text{cột R-})$

- **Bước 7:** Quy tắc quyết định là bác bỏ H_0 nếu $W \leq W_\alpha$ với W_α là giá trị tra bảng Kiểm định dấu và hạng Wilcoxon thuộc phần phụ lục (là bảng tra số 6). Bảng tra này liệt kê cả giá trị cận dưới và cận trên ứng với các mức ý nghĩa (tùy trường hợp kiểm định một bên hay hai bên mà xác định cột phù hợp) và n, chú ý rằng n bằng với số lượng d_i khác không chứ không nhất thiết bằng đúng cỡ mẫu. Với kết quả tra bảng

có cả cận dưới và cận trên ta chỉ xét cận dưới vì kiểm định này luôn thực hiện ở bên trái.

Chúng ta xem xét các bước từ 1 – 6 của thủ tục kiểm định được tập hợp trong bảng sau cho ví dụ của chúng ta

Bảng 10.1

Lương X_i	$d_i = X_i - 350$	$ d_i $	Hạng	R+	R-
364	14	14	2	2	
385	35	35	3	3	
270	-80	80	8		8
350	0	0			
290	-60	60	6,5		6,5
400	50	50	5	5	
520	170	170	9	9	
340	-10	10	1		1
389	39	39	4	4	
410	60	60	6,5	6,5	
Tổng					15,5

Với kiểm định bên phải ta xác định $W = 29,5$

Sau đó tiến hành Bước 7 là so sánh giá trị W tính toán được với giá trị tối hạn tra từ bảng tra số 6 ứng với mức ý nghĩa $\alpha = 0,05$ của kiểm định 1 bên và $n = 9$, (chúng ta có 1 giá trị $d_i = 0$ nên $n = 10 - 1 = 9$)

Theo bảng tra 6 tại cột một bên $\alpha = 0,05$, hàng số $n = 9$ ta có cặp 8;37 tức giá trị tối hạn trên là 37 và giá trị tối hạn dưới là 8. Ta dùng giá trị cận dưới tức $W_\alpha = 8$

Tiến hành so sánh để quyết định bác bỏ H_0 theo quy tắc $W = 29,5 > W_\alpha = 8 \rightarrow$ không bác bỏ H_0 .

Như vậy với độ tin cậy 95% chúng ta không đủ bằng chứng để kết luận rằng lương trung vị của sinh viên đã tốt nghiệp 5 năm vượt quá 350\$.

Khi cỡ mẫu lớn, chú ý là giá trị kiểm định W sẽ xấp xỉ phân phối bình thường nếu cỡ mẫu tăng lên, đó là khi cỡ mẫu trên 20 quan sát, kiểm định Wilcoxon có thể sử dụng xấp xỉ phân phối bình thường với giá trị kiểm định z tính theo công thức sau:

$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Trong đó W là tổng các hạng cột R+; n là số giá trị d_i khác 0

Quy tắc quyết định là bắc bỏ H₀ ở mức ý nghĩa α khi

Z < - Z_α nếu là kiểm định 1 bên

Z < - Z_{α/2} nếu là kiểm định 2 bên

Ví dụ: Một tờ báo đánh giá là trung vị về giá căn hộ Penthouse tại thành phố hiện nay không quá 176.200 USD. Họ tiến hành một khảo sát để kiểm chứng thông tin này. Một mẫu ngẫu nhiên 25 căn hộ Penthouse được chọn và người ta tiến hành kiểm định bằng phương pháp dấu và hạng Wilcoxon với mẫu lớn. chọn mức ý nghĩa là 1%.

1. Đặt giả thuyết như sau

H₀: Trung vị = 176.200

H₁: Trung vị < 176.200

2. Mức ý nghĩa của kiểm định là 1%

3. Tính toán giá trị W như sau:

Bảng 10.2

Giá X _i	d _i	d _i	Hạng	R+	R-	Giá X _i	d _i	d _i	Hạng	R+	R-
173000	-3200	3200	1		1	203000	26800	26800	14	14	
169900	-6300	6300	2		2	204900	28700	28700	15	15	
163500	-12700	12700	3		3	145900	-30300	30300	16		16
160600	-15600	15600	4		4	143500	-32700	32700	17		17
159200	-17000	17000	5		5	137650	-38550	38550	18		18
157200	-19000	19000	6		6	216250	40050	40050	19	19	
156500	-19700	19700	7		7	134500	-41700	41700	20		20
155400	-20800	20800	8		8	128900	-47300	47300	21		21
155200	-21000	21000	9		9	117000	-59200	59200	22		22
197750	21550	21550	10	10		112400	-63800	63800	23		23
154200	-22000	22000	11		11	104500	-71700	71700	24		24
200750	24550	24550	12	12		102600	-73600	73600	25		25
149500	-26700	26700	13		13				Tổng		70

4. Tính toán giá trị thống kê kiểm định z:

$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{70 - \frac{25(25+1)}{4}}{\sqrt{\frac{25(25+1)(2\times 25+1)}{24}}} = -2,49$$

5. Với mức ý nghĩa 0,01, bảng tra số 1 cho ta giá trị tối hạn z_α của kiểm định bên trái là $-z_\alpha = -2,33$

Vì $Z_u = -2,49 < -z_\alpha = -2,33 \rightarrow$ bác bỏ H_0

6. Với độ tin cậy 99% ta có đủ bằng chứng thống kê để kết luận rằng trung vị về giá nhà không quá 176200\$

10.2 KIỂM ĐỊNH TỔNG HẠNG WILCOXON CHO TRUNG BÌNH HAI MẪU ĐỘC LẬP

Chúng ta đã khảo sát phương pháp so sánh trung bình của hai tổng thể độc lập bằng kiểm định t, khi cỡ mẫu nhỏ hoặc các tổng thể lấy mẫu không bảo đảm phân phối bình thường hoặc khi chúng ta có dữ liệu thứ tự chúng ta sẽ thay bằng kiểm định tổng hạng Wilcoxon, một phương pháp kiểm định phi tham số nhằm kiểm tra sự giống nhau của hai trung vị tổng thể.

Để thực hiện kiểm định này, các quan sát từ cả hai mẫu được kết hợp với nhau và xếp hạng từ giá trị nhỏ nhất đến giá trị lớn nhất (tính trên mẫu kết hợp). Giá trị nhỏ nhất trong mẫu kết hợp được xếp hạng 1, những trường hợp đồng hạng thì được thay thế bằng hạng trung bình.

Để thuận tiện, trong trường hợp n_1 và n_2 không bằng nhau chúng ta quy ước gọi n_1 là cỡ của mẫu nhỏ và n_2 là cỡ của mẫu lớn và giá trị kiểm định tổng hạng Wilcoxon T_1 được tính là tổng của tất cả các hạng trong mẫu 1, nếu hai mẫu bằng cỡ nhau tính giá trị kiểm định T_1 từ mẫu nào cũng được.

Kiểm định tổng hạng Wilcoxon có thể là kiểm định một đuôi cũng có thể là kiểm định hai đuôi với các giả thuyết như sau

$H_0: M_1 = M_2$ (kiểm định 2 đuôi)

$H_1: M_1 \neq M_2$

$H_0: M_1 \leq M_2$ (kiểm định bên phải)

$H_1: M_1 > M_2$

$H_0: M_1 \geq M_2$ (kiểm định bên trái)

$H_1: M_1 < M_2$

Trong đó M_1 là trung vị được giả thuyết của tổng thể thứ nhất và M_2 là trung vị của tổng thể thứ 2.

Khi cỡ mẫu n_1 và n_2 đều bé hơn 10 chúng ta sử dụng Bảng tra số 7 để tìm giá trị tối hạn so sánh với giá trị kiểm định T_1 .

- Với kiểm định hai bên tại mức ý nghĩa α , quy tắc quyết định là bác bỏ H_0 nếu $T_1 \geq$ giới hạn trên hoặc $T_1 \leq$ giới hạn dưới.
- Với kiểm định một đuôi bên phải quy tắc quyết định là bác bỏ H_0 nếu $T_1 \geq$ giới hạn trên
- Với kiểm định một đuôi bên trái quy tắc quyết định là bác bỏ H_0 nếu $T_1 \leq$ giới hạn dưới

Nếu cỡ mẫu lớn, giá trị kiểm định T_1 xấp xỉ phân phối bình thường với trung bình và độ lệch tiêu chuẩn như sau:

$$\mu_{T_1} = \frac{n_1(n+1)}{2}$$

$$\sigma_{T_1} = \sqrt{\frac{n_1 n_2 (n+1)}{12}}$$

Từ đó giá trị chuẩn hóa z có thể được sử dụng theo công thức:

$$z = \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}}$$

Công thức chuẩn hóa z được sử dụng khi cỡ mẫu vượt ra khỏi phạm vi của Bảng tra số 7, căn cứ trên mức ý nghĩa đã chọn giả thuyết H_0 sẽ bị bác bỏ nếu giá trị z tính toán rơi vào khu vực bác bỏ H_0 tùy theo đó là kiểm định một bên hay hai bên.

Ví dụ : Để kiểm định tác động của việc trưng bày hàng hóa đến doanh số, người ta chọn 2 mẫu ngẫu nhiên, mẫu thứ nhất gồm 10 gian hàng trưng bày bình thường, mẫu thứ hai gồm 10 gian hàng trưng bày đặc biệt, ghi chép doanh số của các gian hàng trong mẫu ta được số liệu như trong bảng sau.

Vì cỡ mẫu của quan sát nhỏ, có thể các giả định không đảm bảo nên chúng ta không dùng kiểm định t mà dùng kiểm định tổng hạng Wilcoxon để đánh giá có sự khác biệt không trong trung vị về doanh số của hai mẫu, giả thuyết đặt ra là

$$H_0: M_1 = M_2$$

$$H_1: M_1 \neq M_2$$

Bảng 10.3

Doanh số tuần Trung bày bình thường (triệu đồng)	Hạng kết hợp	Doanh số tuần Trung bày đặc biệt (triệu đồng)	Hạng kết hợp
22	1	52	5,5
34	3	71	14
52	5,5	76	15
62	10	54	7
30	2	67	13
40	4	83	17
64	11	66	12
84	18,5	90	20
56	8	77	16
59	9	84	18,5
	72		138

Hai mẫu cùng có nên ta tính T_1 từ mẫu nào cũng được

Cách sử dụng Bảng tra số 7, chúng ta bắt đầu với việc xác định giá trị mức ý nghĩa, sau đó lên hàng đầu tiên của bảng tìm giá trị tương ứng với cỡ mẫu n_1 và đến cột đầu tiên của bảng tìm giá trị ứng với n_2 , nơi gặp nhau của hàng và cột cho ta giá trị ứng với mức ý nghĩa α , ta có giá trị cận trên và dưới cho kiểm định hai bên mức ý nghĩa 0,05 trong tình huống này là 78;132, như vậy nếu $T_1 \geq 132$ hoặc $T_1 \leq 78$ ta sẽ bác bỏ H_0 .

Vì $T_1 = 72 < 78$ nên ta bác bỏ H_0 , như vậy có bằng chứng thống kê về sự khác biệt có ý nghĩa giữa trung vị về doanh số của hai cách trưng bày. Vì cách trưng bày đặc biệt cho một tổng hạng cao hơn chúng ta có thể kết luận là cách trưng bày đặc biệt tạo ra trung vị doanh số có hạng cao hơn.

Có một số phần mềm thống kê cung cấp kiểm định Mann – Whitney là một kiểm định phi tham số tương đương kiểm định tổng hạng Wilcoxon. Bạn đọc có thể tìm hiểu thêm về kiểm định này trong các sách thống kê khác hoặc sách Phân tích dữ liệu nghiên cứu với SPSS của cùng tác giả.

10.3 KIỂM ĐỊNH DẤU VÀ HẠNG WILCOXON CHO MẪU PHỐI HỢP TỪNG CẶP (2 MẪU PHỤ THUỘC)

Khi muốn kiểm định mẫu phối hợp từng cặp mà các giả định cũng không bảo đảm, chúng ta có thể sử dụng kiểm định tham số dấu và hạng Wilcoxon mà bản chất là so sánh hai trung vị. Các bước thực hiện như sau:

- Tính khác biệt D_i cho từng cặp quan sát

- Xác định các giá trị tuyệt đối $|D_i|$
- Xác định cỡ mẫu thực tế là $n' = [n - (\text{số chênh lệch bằng } 0)]$
- Sắp hạng từ 1 đến n' cho các $|D_i|$, $|D_i|$ nhỏ nhất mang hạng 1. Nếu hai D_i cùng giá trị chúng ta sẽ tính hạng trung bình
- Tách riêng các hạng + và - theo dấu của D_i gốc
- Tính tổng cộng hạng riêng cho các chênh lệch dương, đó là trị thống kê kiểm định.

$$W = \sum_{i=1}^{n'} R_i^{(+)}$$

Các giả thuyết đặt ra căn cứ trên giả thuyết không là khác biệt trung vị tổng thể $M_D = 0$

$$H_0: M_D = 0 \text{ (kiểm định 2 đuôi)}$$

$$H_1: M_D \neq 0$$

$$H_0: M_D \leq 0 \text{ (kiểm định bên phải)}$$

$$H_1: M_D > 0$$

$$H_0: M_D \geq 0 \text{ (kiểm định bên trái)}$$

$$H_1: M_D < 0$$

Quy tắc chấp nhận hay bác bỏ H_0 cho tình huống cỡ mẫu $n' \leq 20$ với mức ý nghĩa α xác định:

- Với kiểm định hai bên, nếu giá trị $W \geq$ giới hạn trên hoặc $W \leq$ giới hạn dưới thì bác bỏ H_0
- Với kiểm định bên phải nếu $W \geq$ giới hạn trên thì bác bỏ H_0
- Với kiểm định bên trái nếu $W \leq$ giới hạn dưới thì bác bỏ H_0

Khi $n' > 20$ giá trị W xấp xỉ phân phối bình thường với trung bình và độ lệch tiêu chuẩn như sau:

$$\mu_W = \frac{n'(n'+1)}{4}$$

$$\sigma_W = \sqrt{\frac{n'(n'+1)(2n'+1)}{24}}$$

Từ đó giá trị thống kê z được tính như sau

$$z = \frac{W - \mu_W}{\sigma_W}$$

Công thức xấp xỉ này sẽ được sử dụng làm kiểm định trong tình huống cỡ mẫu lớn vượt quá các giá trị cho trong bảng tra số 6 (Phần phụ lục). Trong

trường hợp mẫu lớn cũng có thể dùng phân phối bình thường thay cho phân phối của kiểm định Wilcoxon. Lúc đó căn cứ trên mức ý nghĩa đã chọn ta sẽ bác bỏ H_0 nếu giá trị z rơi vào khu vực bác bỏ tùy thuộc tình huống ta có kiểm định một bên hay hai bên.

Ví dụ: Để minh họa ta sử dụng lại ví dụ về kiểm định tốc độ xử lý của hai phần mềm mới và hiện dùng, nếu chúng ta không đảm bảo tổng thể của các chênh lệch có phân phối bình thường, ta không dùng kiểm định t mà sử dụng kiểm định phi tham số Wilcoxon, đầu tiên ta đặt giả thuyết:

$$H_0: \mu_D = 0$$

$$H_1: \mu_D > 0$$

Các bước tiến hành kiểm định được thể hiện trong bảng sau:

Bảng 10.4

Lệnh	Thời gian xử lý (giây)		D_i	$ D_i $	R_i	Đầu
	Phần mềm đang dùng	Phần mềm mới				
1	9,98	9,88	+0,1	0,1	7	+
2	9,88	9,86	+0,02	0,02	2	+
3	9,84	9,75	+0,09	0,09	6	+
4	9,99	9,8	+0,19	0,19	8	+
5	9,94	9,87	+0,07	0,07	4,5	+
6	9,84	9,84	+0,00	0,00		
7	9,86	9,87	-0,01	0,01	1	-
8	10,12	9,86	+0,26	0,26	9	+
9	9,9	9,83	+0,07	0,07	4,5	+
10	9,91	9,86	+0,05	0,05	3	+

Tính giá trị kiểm định $W = 7+2+6+8+4,5+9+4,5+3=44$

Vì có một giá trị $D_i = 0$ nên $n' = 9$, theo bảng tra số 6 (Phần phụ lục) chúng ta xác định giá trị tối hạn trên tại mức ý nghĩa 0,05 cho kiểm định 1 bên là 37. Vì $W = 44 > 37$ nên ta bác bỏ H_0 .

Như vậy chúng ta có đủ bằng chứng thống kê để ủng hộ giả thuyết là tốc độ xử lý của phần mềm mới nhanh hơn phần mềm đang dùng.

10.4 KIỂM ĐỊNH KRUSKAL WALLIS CHO NHIỀU MẪU ĐỘC LẬP

Trong Chương 9 Phân tích phương sai, chúng ta thấy kiểm định F chỉ có thể áp dụng khi các nhóm so sánh có phân phối bình thường và phương sai bằng nhau. Trong trường hợp không thỏa điều kiện này, chúng ta có thể chuyển đổi dữ liệu của yếu tố kết quả từ dạng định lượng về dạng

định tính (dữ liệu thứ bậc) và áp dụng một kiểm định phi tham số phù hợp tên là Kruskal-Wallis. Kiểm định này không yêu cầu dữ liệu phải thỏa điều kiện các tổng thể (nhóm) đem ra so sánh phải có phân phối bình thường cho nên kiểm định này có thể áp dụng cho các tổng thể có phân phối bất kỳ.

Giả sử rằng chúng ta có các mẫu ngẫu nhiên độc lập gồm n_1, n_2, \dots, n_k quan sát từ k tổng thể có phân phối bất kỳ. Ta sử dụng kiểm định Kruskal – Wallis bằng cách xếp hạng các quan sát mẫu. Mặc dù số quan sát của n_k mẫu là khác nhau nhưng khi xếp hạng thì được sắp xếp một cách liên tục từ nhỏ đến lớn, nếu giá trị quan sát trùng nhau thì hạng giống nhau bằng cách dùng số trung bình cộng các hạng của chúng để chia đều.

Đặt $n = n_1 + n_2 + \dots + n_k$ là tổng các quan sát thuộc các mẫu, và R_1, R_2, \dots, R_k là tổng của các hạng ở từng mẫu được xếp theo thứ tự của k mẫu. Kiểm định giả thuyết ở mức ý nghĩa α cho trường hợp này là:

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$: Trung bình của k tổng thể đều bằng nhau. Ở đây ta sử dụng đại lượng W thay cho tỉ số F trong phân tích toán giá trị kiểm định.

$$W = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1)$$

Sau đó chúng ta sử dụng bảng phân phối χ^2 (Chi – Square) với $k-1$ bậc tự do để so sánh với giá trị kiểm định, giả thuyết H_0 bị bác bỏ khi:

$$W > \chi^2_{k-1, \alpha}$$

Ví dụ: chúng ta so sánh điểm trung bình học tập của ba nhóm sinh viên có thời gian đi làm thêm khác nhau. Lúc này yếu tố kết quả là điểm trung bình kết quả học tập nhưng yếu tố nguyên nhân là thời gian đi làm thêm với 3 phân nhóm là ít, TB và nhiều. Ta có dữ liệu và kết quả xếp hạng như trong bảng sau. Trong cách xếp hạng này, điểm trung bình thấp nhất của 1 sinh viên (xét trong cả ba nhóm) là 5,3 được xếp hạng 1. Tương tự, hạng được xếp cho đến điểm trung bình cao nhất là 7,3 của sinh viên ở nhóm 1 là hạng thứ 22. Chú ý nữa là nhóm làm thêm ít và trung bình có 7 sinh viên ở mỗi nhóm còn nhóm làm thêm nhiều có 8 sinh viên được quan sát.

Bảng 10.5 Xếp hạng các dữ liệu về điểm trung bình học tập của sinh viên

TG làm thêm ít	Hạng	TG làm thêm TB	Hạng	TG làm thêm nhiều	Hạng
6,3	10,5	7,2	21	6,3	10,5
7,0	19	6,6	15,5	5,8	4,5
6,5	13,5	6,1	8	6,0	7
6,6	15,5	5,8	4,5	5,5	3
7,3	22	6,8	17	5,3	1
6,9	18	7,1	20	6,5	13,5
6,4	12	5,9	6	5,4	2
	R _i =110,5		R ₂ =92		R ₃ =50,5

$$W = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{22.(22+1)} \left[\frac{110,5^2}{7} + \frac{92^2}{7} + \frac{50,5^2}{8} \right] - 3(22+1) = 8,6$$

Ở đây chúng ta có bậc tự do (k-1) = 2 và nếu kiểm định ở mức ý nghĩa 0,05 (5%), khi tra bảng phân phối χ^2 ta tìm được: $\chi^2_{2;0,05} = 5,99$

Vì $W = 8,6 > \chi^2_{2;0,05} = 5,99$ nên giả thuyết H_0 bị bác bỏ ở mức ý nghĩa 0,05 nghĩa là điểm trung bình học tập ở ba nhóm sinh viên có thời gian đi làm thêm khác nhau là không bằng nhau. Chúng ta kết luận rằng với dữ liệu mẫu này, ở độ tin cậy 95% thì thời gian đi làm thêm nhiều ít khác nhau có ảnh hưởng khác nhau đến kết quả học tập của sinh viên có đi làm thêm.

Khi giả thuyết về trung bình của k tổng thể giống nhau bị bác bỏ thì vấn đề tiếp theo là trung bình của tổng thể nào khác tổng thể nào? Chúng ta sẽ dùng một phương pháp so sánh tương tự như phương pháp Tukey trong Chương 9 Phân tích phương sai. Sau đây là tóm tắt các bước thực hiện:

Bước 1: Trước hết chúng ta tính hạng trung bình cho từng nhóm muốn so sánh theo công thức tổng quát sau:

$$\bar{R}_i = \frac{R_i}{n_i}$$

Bước 2: Tiếp theo chúng ta tính chênh lệch về hạng trung bình giữa 2 nhóm cần so sánh

$$D_{ij} = |\bar{R}_i - \bar{R}_j|$$

D được coi như giá trị để kiểm định giả thuyết về sự bằng nhau của trung bình hai tổng thể i và j đang so sánh.

Bước 3: Tính giá trị giới hạn C_K theo công thức:

$$C_K = \sqrt{(\chi^2_{k-1,\alpha}) \left(\frac{n(n+1)}{12} \right) \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

trong đó $\chi^2_{k-1,\alpha}$ là giá trị đã sử dụng khi thực hiện kiểm định Kruskal – Wallis trong phần trước. Còn n_i và n_j là số quan sát của hai nhóm được đếm so sánh.

Bước 4: Nguyên tắc quyết định: Bác bỏ giả thuyết Ho về sự bằng nhau của hai trung bình tổng thể khi D > C_K

Ví dụ tính toán: trở lại ví dụ trên chúng ta lần lượt so sánh giữa ba nhóm sinh viên có thời gian đi làm thêm khác nhau: ít, TB và nhiều.

* tính hạng trung bình cho từng nhóm

$$\overline{R_{it}} = \frac{R_1}{7} = \frac{110,5}{7} = 15,786$$

$$\overline{R_{tb}} = \frac{R_2}{7} = \frac{92}{7} = 13,143$$

$$\overline{R_{nhiều}} = \frac{R_3}{8} = \frac{50,5}{8} = 6,3125$$

* tính các chênh lệch về hạng trung bình giữa từng cặp nhóm

- so sánh nhóm ít với TB: $D_{it, TB} = |15,786 - 13,143| = 2,643$

- so sánh nhóm ít với nhiều: $D_{it, nhiều} = |15,786 - 6,3125| = 9,4735$

- so sánh nhóm TB với nhiều: $D_{TB, nhiều} = |13,143 - 6,3125| = 6,8305$

* tính các giá trị giới hạn C_K

- so sánh nhóm ít với TB:

$$C_K = \sqrt{(\chi^2_{k-1,\alpha}) \left(\frac{n(n+1)}{12} \right) \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} = \sqrt{5,99 \left(\frac{22(22+1)}{12} \right) \left(\frac{1}{7} + \frac{1}{7} \right)} = 8,5$$

- so sánh nhóm ít với nhiều:

$$C_K = \sqrt{(\chi^2_{k-1,\alpha}) \left(\frac{n(n+1)}{12} \right) \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} = \sqrt{5,99 \left(\frac{22(22+1)}{12} \right) \left(\frac{1}{7} + \frac{1}{8} \right)} = 8,23$$

- so sánh nhóm TB với nhiều:

$$C_K = \sqrt{\left(\chi^2_{k-1,\alpha}\right)\left(\frac{n(n+1)}{12}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = \sqrt{5,99\left(\frac{22(22+1)}{12}\right)\left(\frac{1}{7} + \frac{1}{8}\right)} = 8,23$$

* quyết định:

- $D_{u,TB} = 2,643 < 8,23$: chấp nhận giả thuyết cho rằng trung bình hai đối tượng bằng nhau. Như vậy điểm trung bình của sinh viên có thời gian đi làm thêm ít và trung bình không khác biệt có ý nghĩa thống kê.
- $D_{u,nhiều} = 9,4735 > 8,23$: bác bỏ giả thuyết cho rằng trung bình hai đối tượng bằng nhau. Như vậy điểm trung bình của sinh viên có thời gian đi làm thêm ít và nhiều là có khác biệt có ý nghĩa thống kê.
- $D_{TB,nhiều} = 6,8305 < 8,23$: chấp nhận giả thuyết cho rằng trung bình hai đối tượng bằng nhau. Như vậy điểm trung bình học tập của sinh viên có thời gian đi làm thêm trung bình và nhiều không khác biệt có ý nghĩa thống kê.

10.5 KIỂM ĐỊNH CHI-BÌNH PHƯƠNG VỀ TÍNH ĐỘC LẬP (KIỂM ĐỊNH LIÊN HỆ GIỮA 2 BIẾN ĐỊNH TÍNH)

Trong thực tế nhiều khi bạn có thể gặp một số tình huống đòi hỏi phải tìm hiểu mối liên hệ giữa các biến định tính, chẳng hạn mối liên hệ giữa kết quả học tập và việc có hay không có người yêu, mối liên hệ giữa độ bền của cuộc hôn nhân với thời gian yêu nhau trước khi kết hôn, mối liên hệ giữa việc thuận tay trái hay tay phải với giới tính ... Kiểm định Chi bình phương sẽ giúp chúng ta tiến hành việc tìm hiểu các mối liên hệ này có thật sự tồn tại hay không.

Kiểm định sự độc lập hay quan hệ giữa hai biến định tính là một tình huống hay gặp, giả thuyết chung cho loại kiểm định này là:

H_0 : Hai biến định tính độc lập (nghĩa là không có mối liên hệ giữa hai biến này)

H_1 : Hai biến định tính không độc lập (nghĩa là có mối liên hệ giữa hai biến này)

Đại lượng Chi bình phương dùng cho kiểm định này được tính theo công thức sau

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Trong đó:

- O_{ij} là tần số quan sát thực tế của ô ở địa chỉ ij
- E_{ij} là tần số lý thuyết của ô ở địa chỉ ij , được tính theo công thức:

$$E_{ij} = \frac{(\text{Tổng hàng } i) \times (\text{Tổng cột } j)}{\text{Cỡ mẫu}}$$

- i là số hàng của bảng
- j là số cột của bảng
- i là kí hiệu của hàng ($i = 1, 2, \dots, r$)
- j là kí hiệu của cột ($j = 1, 2, \dots, c$)
- số bậc tự do của đại lượng Chi bình phương là $df = (r-1) \times (c-1)$

Ý tưởng của kiểm định này là so sánh số quan sát thực tế và số quan sát theo lý thuyết tại các ô, số quan sát theo lý thuyết là số quan sát sẽ xảy ra nếu giả thuyết H_0 đúng, tức là không có mối liên hệ giữa hai đại lượng ta quan tâm, đại lượng χ^2_{tt} sẽ lớn nếu các tần số quan sát và tần số lý thuyết khác biệt nhiều tức là khả năng H_0 bị bác bỏ sẽ lớn và ngược lại nếu χ^2_{tt} rất nhỏ thì có nghĩa là nhiều khả năng H_0 đúng vì O_{ij} xảy ra như giả thuyết.

Quy tắc quyết định là bác bỏ giả thuyết H_0 tại mức ý nghĩa α nếu giá trị kiểm định χ^2_{tt} lớn hơn giá trị tối hạn trên tra từ bảng phân phối Chi bình phương với bậc tự do $df = (r-1)(c-1)$, tức là:

Nếu $\chi^2_{\text{tt}} > \chi^2_{(r-1)(c-1), \alpha} \rightarrow$ bác bỏ H_0 .

Ví dụ cho kiểm định Chi bình phương về sự độc lập.

Người ta nghiên cứu 200 cặp vợ chồng có thời gian kết hôn trên 5 năm để tìm hiểu có mối liên hệ giữa thời gian tìm hiểu trước hôn nhân và tình trạng hiện tại của cuộc hôn nhân hay không, có 3 dạng mức độ của thời gian tìm hiểu là ngắn, trung bình và dài. Cũng có 3 tình trạng hôn nhân là hạnh phúc, không hạnh phúc, và ly dị/lý thân. Từ số liệu thu được người ta lập một bảng 3 dòng và 3 cột để mô tả các tình huống kết hợp của tình trạng hôn nhân và thời gian tìm hiểu như sau. Chú ý là số liệu trong các ô nền xám là tần suất quan sát thực tế, ví dụ O_{21} có giá trị là 12, O_{33} có giá trị là 2. Chọn mức ý nghĩa cho kiểm định là 5%.

Giả thuyết đặt ra là:

H_0 : Không có mối liên hệ giữa thời gian tìm hiểu trước hôn nhân và tình trạng hiện tại của cuộc hôn nhân

H_1 : Có mối liên hệ giữa thời gian tìm hiểu trước hôn nhân và tình trạng hiện tại của cuộc hôn nhân

Bảng 10.6

	Ngắn	Trung bình	Dài	Tổng hàng
Hạnh phúc	38	58	54	150
Không hạnh phúc	12	14	4	30
Ly dị/lý thân	10	8	2	20
Tổng cột	60	80	60	200

Trước khi tính toán giá trị kiểm định chúng ta phải tính tần số lý thuyết như sau. Ví dụ tần số lý thuyết E_{21}

$$E_{21} = \frac{(Tổng hàng 2) \times (Tổng cột 1)}{Cỡ mẫu} = \frac{30 \times 60}{200} = 9$$

Bảng sau liệt kê các tần số lý thuyết tại từng ô.

Bảng 10.7

	Ngắn	Trung bình	Dài	Tổng hàng
Hạnh phúc	45	60	45	150
Không hạnh phúc	9	12	9	30
Ly dị/lý thân	6	8	6	20
Tổng cột	60	80	60	200

Tính toán giá trị kiểm định

$$\begin{aligned} \chi^2 &= \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(38-45)^2}{45} + \frac{(58-60)^2}{60} + \frac{(54-45)^2}{45} + \frac{(12-9)^2}{9} + \dots + \frac{(8-8)^2}{2} + \frac{(2-6)^2}{2} \\ &= 1,089 + 0,067 + 1,800 + 1 + \dots + 0 + 2,667 \\ &= 12,4 \end{aligned}$$

Tra bảng Chi bình phương tìm giá trị tối hạn $\chi^2_{(r-1)(c-1); \alpha} = \chi^2_{(3-1)(3-1); 0,05} = \chi^2_{4; 0,05} = 9,48$.

Vì $\chi^2_n = 12,4 > \chi^2_{4; 0,05} = 9,48 \rightarrow$ bác bỏ H_0 .

Như vậy với độ tin cậy 95%, có đủ bằng chứng thống kê để kết luận là có mối liên hệ giữa thời gian tìm hiểu trước hôn nhân và tình trạng hiện tại của cuộc hôn nhân.

Có một giới hạn với kiểm định Chi bình phương là tần số lý thuyết phải không được bé hơn 5, nếu không thì giá trị Chi bình phương tính toán được sẽ có khả năng bị phóng đại làm tăng khả năng bác bỏ H_0 . Do đó nếu có tần số lý thuyết bé hơn 5 mà kết quả kiểm định đi đến bác bỏ H_0 thì chúng ta phải cẩn thận. Gặp tình huống này chúng ta có thể khắc phục bằng cách tăng cỡ mẫu hoặc nhóm các hàng và các cột lại một cách phù hợp.

10.6 KIỂM ĐỊNH CHI-BÌNH PHƯƠNG VỀ SỰ PHÙ HỢP

Kiểm định Chi-bình phương được sử dụng khá phổ biến đối với các biến định tính (phân loại). Trong mục 10.5 chúng ta đã xem xét việc sử dụng bảng chéo và kiểm định Chi-bình phương để xem xét sự liên hệ của một biến định tính này với một biến định tính khác, ví dụ trình độ học vấn có ảnh hưởng (tức là có liên quan) đến sự đánh giá của một người nào đó về tầm quan trọng của tiền bạc trong cuộc sống không... Kiểm định Chi-bình phương còn được vận dụng để giải quyết nhiều yêu cầu nghiên cứu khác nữa. Trong phần này chúng ta sẽ sử dụng kiểm định Chi-bình phương để xem xét dữ liệu của chúng ta phù hợp (thích hợp) đến mức độ nào với giả thuyết về phân phối của tổng thể.

Trong các kiểm định tham số đã nghiên cứu, rất thường xuyên chúng ta gặp giả định dữ liệu lấy từ tổng thể có phân phối bình thường. Vậy làm thế nào để kiểm định dữ liệu của chúng ta có phân phối bình thường, hay có một phân phối nào đó như dự kiến. Ta dùng kiểm định Chi bình phương về sự phù hợp để xác định dữ liệu mẫu có đúng được chọn từ một tổng thể có phân phối giả thuyết không, trình tự tiến hành kiểm định như sau:

Ví dụ 1: Một công ty muốn nghiên cứu các vụ tai nạn lao động có xảy ra như nhau vào các ngày làm việc trong tuần không hay là nó có xu hướng tăng cao vào các ngày thứ Hai và các ngày cuối tuần. Ta lập luận rằng nếu giả thuyết cho rằng "các vụ tai nạn xảy ra với xác suất như nhau trong 6 ngày làm việc của tuần" là đúng thì số tai nạn phải phân phối đều với xác suất xảy ra tai nạn mỗi ngày phải bằng nhau và bằng $1/6$. Với tổng số 32 vụ tai nạn lao động công ty đó thu thập được trong vòng 5 năm qua tại các nhà máy của công ty, số lượng các vụ tai nạn trong từng ngày phải bằng nhau và $=1/6 \times 32=5,33$ vụ.

Bảng 10.8 cho thấy trên thực tế 32 vụ tai nạn phân bố như thế nào vào các ngày trong tuần, dường như các vụ tai nạn xảy ra không đồng đều giữa 6 ngày làm việc trong tuần (xem cột %)

Giả thuyết:

H_0 : tai nạn lao động vào các ngày trong tuần có phân phối đều

H_1 : tai nạn lao động vào các ngày trong tuần không có phân phối đều

Bảng 10.8

THỨ	Thực tế		Giả thuyết		$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
	Số vụ O_i	%	Số vụ E_i	%		
hai	7	21,9	5,33	16,66	2,79	0,523
ba	3	9,4	5,33	16,66	5,29	0,998
tư	3	9,4	5,33	16,66	5,29	0,998
năm	2	6,3	5,33	16,66	10,89	2,055
sáu	5	15,6	5,33	16,66	0,09	0,017
bảy	12	37,5	5,33	16,66	44,89	8,470
Total	32	100,0	32	100,0		13,061

Ta sử dụng kiểm định Chi-bình phương một mẫu theo thủ tục như sau

Trước tiên các dữ liệu được phân thành các nhóm, trong ví dụ này là phân theo các ngày xảy ra tai nạn, sau đó ta tính tần số lý thuyết hay còn gọi là tần số kỳ vọng (Expected frequency) xảy ra tai nạn tại các ngày trong tuần, Tần số lý thuyết là tần số xảy ra nếu giả thuyết H_0 là đúng, Tần số lý thuyết đó chính là 5,33 vụ tai nạn/ngày đã nói ở trên. Đại lượng thống kê Chi-bình phương được tính như công thức sau:

$$\chi^2_n = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Trong đó:

- O_i là tần số quan sát thực tế của loại thứ i (ở đây là ngày)
- E_i là tần số lý thuyết của loại thứ i
- k là số phân loại (đây là số ngày làm việc trong tuần k = 6)

Đại lượng χ^2_n sẽ lớn nếu các tần số quan sát và tần số lý thuyết khác nhau nhiều, nên giả thuyết H_0 sẽ có khả năng bị bác bỏ nếu χ^2 tính được lớn, nếu đại lượng này rất gần 0 thì khả năng H_0 đúng lớn.

Do đó ta tra bảng Chi-bình phương với $k-1$ bậc tự do và mức ý nghĩa α để tìm giá trị giới hạn trên và so sánh với χ^2_n theo quy tắc nếu $\chi^2_n > \chi^2_{k-1;\alpha}$ thì bác bỏ H_0 .

Ta có $\chi^2_n = 13,061 > \chi^2_{6-1;0,05} = 11,07 \rightarrow$ bác bỏ H_0

Như vậy ta có bảng chứng để bác bỏ giả thiết tai nạn xảy ra các ngày trong tuần theo phân phối đều. Theo bảng tổng hợp, căn cứ trên cột O; ta thấy tai nạn có nhiều khả năng xảy ra vào ngày đầu tuần và nhất là 2 ngày cuối tuần, do đó công ty nên áp dụng các biện pháp đặc biệt để phòng tai nạn lao động vào những ngày này.

Ví dụ thứ 2: Công ty TT chuyên sản xuất các thiết bị gỗ. Bước đầu tiên trong tiến trình sản xuất là họ xẻ gỗ thành tấm với độ rộng khác nhau để phục vụ cho những mục đích sản xuất khác nhau. Ví dụ như sản xuất khung cửa chính và cửa sổ, các đồ gỗ trong nhà... Trong tiến trình tự động hóa sản xuất, công ty quyết định chọn một số độ rộng chuẩn, lưu trữ vào máy tính để các máy xẻ gỗ tự động cắt xẻ gỗ nguyên liệu theo các kích cỡ chuẩn này. Theo quy định kỹ thuật, chênh lệch trung bình của các tấm gỗ được xẻ là $= 0$ và các chênh lệch này có phân phối bình thường với độ lệch tiêu chuẩn $= 0,01$ inch.

Để bảo đảm các máy xẻ gỗ cắt gỗ đúng qui định kỹ thuật, bộ phận kiểm soát chất lượng quyết định tiến hành điều tra, họ chọn ngẫu nhiên 300 tấm vừa ra khỏi máy xẻ. Để có một độ rộng làm giá trị đối chứng, người ta chọn độ rộng của tấm gỗ đầu tiên được lấy vào mẫu, nó rộng 2,875 inches, các tấm còn lại được đo chiều rộng, tính chênh lệch với chiều rộng đối chứng và ghi lại giá trị chênh lệch, sau đó người ta dùng kiểm định Chi bình phương về sự phù hợp để kiểm định giả thuyết sau:

H_0 : các chênh lệch có phân phối bình thường với $\mu = 0$ và $\sigma = 0,01$

H_1 : các chênh lệch không có phân phối bình thường với $\mu = 0$ và $\sigma = 0,01$

Chúng ta đang kiểm định về một phân phối liên tục, do đó trước tiên chúng ta tổ chức dữ liệu thành bảng tần số, sau khi cân nhắc về việc phân tổ dựa trên dữ liệu về chênh lệch chúng ta có kết quả sau

Hình 10.1

A	B	C	D	E	F
Thứ tự	Độ rộng thực	Độ rộng đối chứng	Chênh lệch	Bin	Frequency
2	1	2.870	2.875	dưới -0,02	0
3	2	2.863	2.875	(-0,02 ; -0,01)	42
4	3	2.885	2.875	(-0,01; 0)	133
5	4	2.872	2.875	(0; 0,01)	75
6	5	2.891	2.875	(0,01; 0,02)	47
7	6	2.893	2.875	trên 0,02	3
8	7	2.868	2.875		300

Hình trên là một phần của bộ dữ liệu cho nghiên cứu được cắt ra từ màn hình Excel, bên là bảng phân tổ được lập bằng Excel

Từ bảng phân tách này chúng ta lập bảng tính toán các số liệu cho kiểm định:

Bảng 10.9

Chênh lệch (X _i)	Tần số quan sát	Xác suất tích lũy của phân phối bình thường	Xác suất của phân phối bình thường	Tần số giả thuyết	(O _i -E _i) ² /E _i
(1)	(2)	(3)	(4)	(5)	(6)
dưới -0,02	0	0,0228	0,0228	6,8250	6,8250
(-0,02; -0,01)	42	0,1587	0,1359	40,7715	0,0370
(-0,01; 0)	133	0,5000	0,3413	102,4034	9,1418
(0; 0,01)	75	0,8413	0,3413	102,4034	7,3332
(0,01; 0,02)	47	0,9772	0,1359	40,7715	0,9515
trên 0,02	3	1,0000	0,0227	6,8250	2,1436
	300			1	300
					26,4322

Cách tính một số giá trị trong bảng như sau:

Dữ liệu trên cột thứ 3 được tính bằng hàm Normdist của Excel, hàm này cho ta giá trị xác suất tích lũy để X_i nhận giá trị từ $-\infty$ đến giá trị cận trên của tổ, nếu như phân phối bình thường đã giả thuyết là đúng, ví dụ giá trị 0,1587 được tính bằng hàm =NORMDIST(-0,01,0,0,01,1). Trong đó -0,01 là giá trị cận trên của tổ; 0 là giá trị trung bình của phân phối; 0,01 là giá trị độ lệch tiêu chuẩn của phân phối; 1 để yêu cầu Excel cho ta xác suất tích lũy.

Dữ liệu trên cột thứ 4 được tính bằng cách lấy giá trị Xác suất tích lũy của tổ sau trừ Xác suất tích lũy của tổ kế trước, như vậy nó chính là diện tích dưới đường cong bình thường trong phạm vi giá trị giới hạn bởi hai giá trị cận trên và dưới của mỗi tổ. Ví dụ giá trị 0,3413 = 0,5 - 0,1587.

Dữ liệu trên cột thứ 5 được tính bằng cách lấy tổng số quan sát (300) nhân với xác suất để X_i nhận giá trị trong một tổ nào đó nếu phân phối đã giả thuyết là đúng.

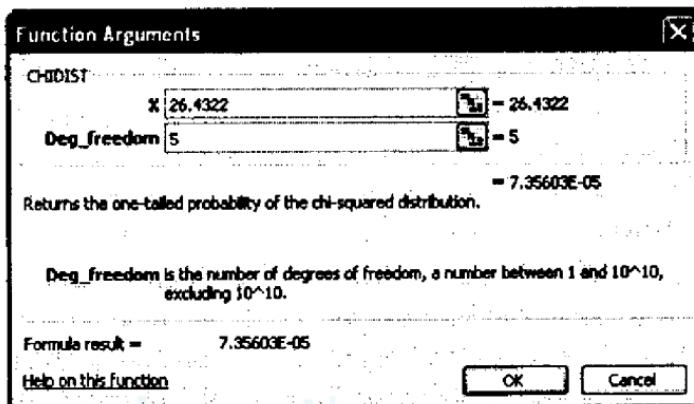
Bảng số liệu cho thấy giá trị thống kê kiểm định $\chi^2_{\text{t}} = 26,43$

Có một chú ý là kiểm định Chi bình phương chỉ chính xác nếu các ô có tần số giả thuyết đều lớn hơn 5, nếu có ô nào có tần số giả thuyết bé hơn 5 chúng ta cần làm lại việc phân chia các tổ theo hướng kết hợp sao cho không còn ô nào có dưới 5 quan sát nữa. Cả hai ví dụ của chúng ta đều đảm bảo yêu cầu này.

Trong ví dụ này chúng ta có $k = 6$ vì ta chia 6 tổ, nếu chúng ta dùng lệnh Chidist trong menu Function của Excel (xem Hình 10.2) ta có thể tìm ra giá trị p-value ứng với giá trị kiểm định 26,43 là $p\text{-value} = 0,0001$. Dù cho chọn mức ý nghĩa là 0,01 thì chúng ta vẫn có thể bác bỏ H_0 vì $p\text{-value} = 0,0001 < 0,01$.

Nếu tra bảng giá trị tối hạn Chi bình phương với mức ý nghĩa 0,01 ta được $\chi^2_{6-1;0,01} = 15,0863$. Vì $\chi^2_{n} = 26,43 > \chi^2_{6-1;0,01} = 15,0863 \rightarrow$ bác bỏ H_0

Hình 10.2



Như vậy với độ tin cậy 99% chúng ta có đủ bằng chứng thống kê để kết luận rằng các máy xé gỗ đã không đạt được yêu cầu kỹ thuật đề ra vì các chênh lệch không có phân phối bình thường với trung bình = 0 và độ lệch tiêu chuẩn = 0,01.

CHƯƠNG 11

HỒI QUI TUYẾN TÍNH ĐƠN BIẾN VÀ PHÂN TÍCH TƯƠNG QUAN

11.1 LÀM QUEN VỚI HỒI QUI

11.1.1 Khái niệm hồi qui

Bạn có biết nguồn gốc của thuật ngữ “hồi qui”? Thuật ngữ này được nhà nghiên cứu Francis Galton sử dụng lần đầu tiên vào cuối thế kỷ 19 trong một nghiên cứu nhằm tìm hiểu tại sao có sự ổn định trong chiều cao trung bình của dân số, nguyên văn là cụm từ “regression to mediocrity” – “hồi qui về trung bình”, kể từ đó trở đi vấn đề hồi qui được nhiều người quan tâm và hoàn thiện qua những ứng dụng có nội dung rộng hơn nhiều so với nghiên cứu ban đầu mà nó được sử dụng.

Trong cụm từ “regression to mediocrity”, từ “regression” chính là “hồi qui”, theo nghĩa Hán Việt thì “hồi qui” có thể hiểu nôm na là cách thức qui các điểm dữ liệu quan sát về một đường lý thuyết đã biết phương trình biểu diễn để có thể dễ dàng tính toán (nội suy hay ngoại suy), hay nói cách khác là dùng một đường lý thuyết để mô tả luật biến thiên của các điểm dữ liệu quan sát, giúp nhìn thấy mối liên hệ giữa các biến nghiên cứu diễn ra theo qui luật nào.

Hiểu theo nghĩa Hán Việt dĩ nhiên không thể đi quá xa định nghĩa về “hồi qui” của thống kê. Vậy quan điểm thống kê hiện đại định nghĩa “hồi qui” như thế nào?

Phân tích hồi qui là nghiên cứu mối liên hệ phụ thuộc của một biến (gọi là biến phụ thuộc) vào một hay nhiều biến khác (gọi là các biến độc lập), với ý tưởng ước lượng và/hoặc dự đoán giá trị trung bình (tổng thể) của biến phụ thuộc trên cơ sở các giá trị biết trước (trong mẫu) của các biến độc lập.

Toàn bộ quan điểm về phân tích hồi qui vừa nêu trên sẽ được làm sáng tỏ dần trong nội dung chương này, nhưng ví dụ sau đây sẽ giúp chúng ta có hình dung ban đầu:

Một người nghiên cứu muốn dùng phương pháp phân tích hồi qui nghiên cứu mối liên hệ phụ thuộc giữa trình độ học vấn của một người (anh ta đã điệu này bằng số năm đi học) vào thu nhập của người đó, thông tin về mối liên hệ phụ thuộc sau đó có thể giúp dự đoán thu nhập của một người căn cứ trên trình độ học vấn, như vậy ở đây thu nhập phụ thuộc vào học vấn nên nó đóng vai trò là biến phụ thuộc (tạm đặt là Y) và trình độ học

vấn là biến độc lập (tạm đặt là X), với sự khảo sát liên hệ giữa một biến phụ thuộc vào chỉ một biến độc lập ta có tình huống của hồi qui đơn biến. Muốn dùng phương pháp hồi qui chúng ta phải biết hoặc cho rằng đã biết dạng của mối liên hệ giữa biến số Y vào X được biểu diễn dưới dạng hàm số trong đó Y được thể hiện bằng một biểu thức toán học chỉ tùy thuộc ở biến số X và một vài thông số, chẳng hạn nếu dạng liên hệ giữa học vấn và thu nhập là đường thẳng thì ta có phương trình $Y = a + bX$.

Ai cũng có thể nhận thấy học vấn không chỉ là yếu tố duy nhất ảnh hưởng đến thu nhập của một người, mà còn có những yếu tố khác như loại hình công việc người đó làm, mức độ năng động của cá nhân đó ... tức là còn có một vài biến độc lập khác có thể tham gia vào phân tích hồi qui này, lập luận đó đã dẫn dắt bạn đi từ ví dụ của một phân tích hồi qui đơn biến sang một ví dụ của phân tích hồi qui đa biến. Phân tích hồi qui đa biến cũng có thể là tình huống khi một nhà nông học nghiên cứu sự phụ thuộc của sản lượng vụ mùa lúa (chính là biến phụ thuộc) vào các yếu tố: nhiệt độ, lượng mưa, nắng, và chế độ bón phân (có thể kể như các biến độc lập). Một phân tích hồi qui về sự phụ thuộc của một biến phụ thuộc Y vào nhiều biến độc lập X như vậy có thể cho phép nhà nông học dự đoán được sản lượng vụ mùa trung bình khi biết được thông tin về các biến độc lập.

11.1.2 Phân biệt liên hệ thống kê và liên hệ hàm số khi phân tích hồi qui

Chú ý rằng trong phân tích hồi qui, chúng ta quan tâm tới các mối liên hệ phụ thuộc thống kê, chứ không phải sự phụ thuộc hàm số như trong toán học.

Trong liên hệ hàm số, ví dụ như phương trình của hàm bậc nhất $Y = (aX + b)$, với một giá trị X thế vào hàm số chúng ta tìm được duy nhất một giá trị Y.

Phân tích hồi qui không xét các liên hệ hàm số như thế, mô hình hồi qui đơn biến ($Y = a + bX$) mô tả sự phụ thuộc của thu nhập vào học vấn về bản chất mang tính chất thống kê ở chỗ biến giải thích X, mặc dù quan trọng, cũng không cho phép người nghiên cứu dự đoán thu nhập Y chính xác sẽ là bao nhiêu bởi vì còn có một số các yếu tố khác (tức là các biến giải thích khác), tuy chưa được kể ra trong phương trình nhưng cùng đồng thời tác động tới thu nhập khiến cho có một mức độ biến thiên ngẫu nhiên trong thu nhập của những người có trình độ học vấn như nhau mà thu nhập không hoàn toàn giống nhau.

11.1.3 Một số qui ước về ký hiệu và tên gọi

Cặp khái niệm biến phụ thuộc và biến độc lập vừa nói trên còn có một cách gọi khác là biến được giải thích và biến giải thích. Như vậy nếu ta nghiên

cứu sự phụ thuộc của một biến được giải thích vào một biến giải thích duy nhất, ví dụ như thu nhập phụ thuộc vào học vấn, hay chiều cao của cây có thể được dự đoán qua mối liên hệ với đường kính thân cây... nghiên cứu đó được gọi là phân tích hồi qui đơn biến (nguyên nhân), hay còn gọi là hồi qui hai biến (1 nguyên nhân, 1 kết quả). Còn nếu ta nghiên cứu sự phụ thuộc của một biến được giải thích bởi hai hay nhiều biến giải thích, như trong ví dụ sản lượng vụ mùa, lượng mưa, nhiệt độ, nắng, và phân hóa học, thì nó được gọi là phân tích hồi qui đa biến, hay hồi qui bội. Quy ước là trong phân tích hồi qui đơn biến, chỉ có một biến giải thích, còn trong hồi qui bội, có nhiều hơn một biến giải thích, còn số biến được giải thích trong mọi trường hợp dĩ nhiên đều là một biến.

Trong phân tích hồi qui người ta thường ký hiệu Y là biến phụ thuộc (hay còn gọi tên là biến được giải thích) và X_k là biến độc lập (hay biến giải thích) thứ k . Trong đó biến phụ thuộc Y là một đại lượng ngẫu nhiên có quy luật phân phối xác suất, còn các X_i không ngẫu nhiên mà giá trị của chúng đã được xác định trước, ứng với mỗi giá trị cụ thể của X có một phân phối các giá trị của Y .

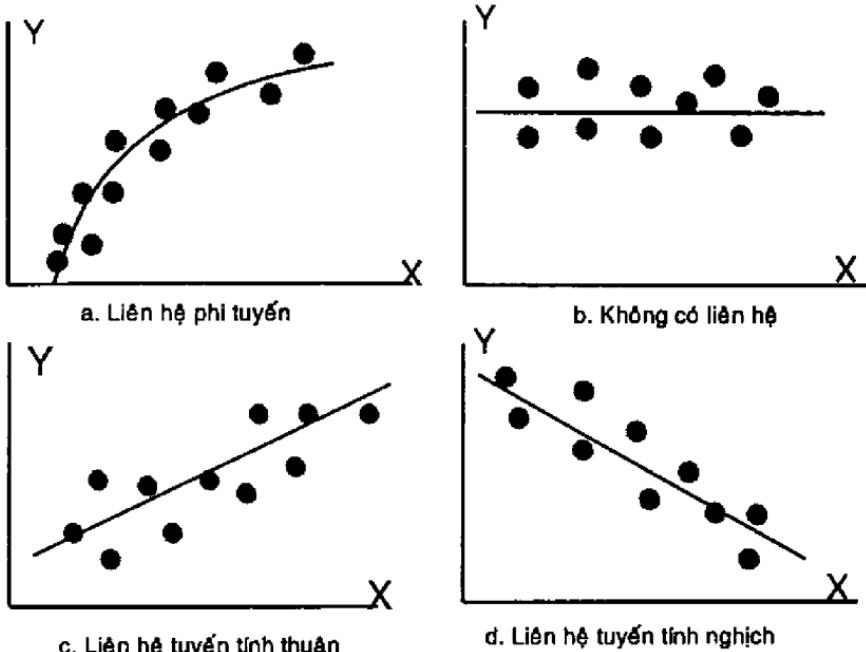
Trừ khi có định nghĩa khác đi, còn trong mô hình hồi qui tuyến tính ký tự i hoặc t sẽ biểu thị lần quan sát thứ t hay giá trị thứ i , vậy X_{ki} hoặc X_{ti} sẽ biểu thị quan sát thứ i hay t của biến X_k . Chữ n biểu thị cho tổng số các quan sát trong một mẫu. Theo quy ước, ký tự con i sẽ được dùng cho số liệu chéo (số liệu thu thập tại một thời điểm ở nhiều không gian hoặc trên nhiều đối tượng) và ký tự con t sẽ được dùng cho số liệu chuỗi thời gian (số liệu thu thập về một đối tượng qua một khoảng thời gian).

Với mô hình hồi qui đơn biến, $Y = a + bX$ thì a và b được gọi là các hệ số hồi qui, cách ký hiệu ban đầu này rất đơn giản và kém chặt chẽ về mặt thống kê, chỉ sử dụng để bạn đọc làm quen với khái niệm hồi qui, còn trong các nội dung sau chúng ta sẽ phân biệt phương trình hồi qui mẫu và phương trình hồi qui tổng thể một cách rõ ràng qua các ký hiệu đại diện và cách xây dựng phương trình.

11.1.4 Các dạng liên hệ giữa hai biến X và Y

Ở trên chúng ta có nói rằng hiểu nôm na thì hồi qui như là cách thức qui các điểm dữ liệu quan sát về một đường lý thuyết đã biết phương trình biểu diễn để có thể dễ dàng tính toán (nội suy hay ngoại suy). Vậy có các dạng đường lý thuyết mô tả liên hệ giữa Y và X như thế nào, hãy xem hình minh họa dưới đây

Hình 11.1



Các chấm trên hình là các điểm dữ liệu phân tán, mỗi chấm là một sự kết hợp giữa Y và X cho ta một cặp giá trị cụ thể. Các đường liền nét trong hình là đường lý thuyết cho ta thấy dạng liên hệ giữa hai biến số. Ở Hình 11.1c ta thấy khi Y tăng X cũng tăng, nếu xét liên hệ giữa chiều cao của cây và đường kính thân cây ta sẽ cảm nhận được mối liên hệ này, đường kính thân cây càng to thì cây đó phải càng cao; ngược lại ở Hình 11.1d thì Y giảm theo chiều tăng của X , đây có thể là dạng của mối liên hệ giữa giá cả một loại hàng hóa nào đó và tổng sản lượng hàng hóa được tiêu thụ, mối liên hệ của chúng ở dạng nghịch. Với Hình 11.1b ta thấy Y nhận một giá trị duy nhất với mọi tình huống của X , còn Hình 11.1a mô tả một dạng liên hệ phi tuyến đặc biệt mà khi X tăng Y cũng tăng nhưng tốc độ tăng chậm dần và có thể di đến tiệm cận một ngưỡng giá trị xác định, đây là đường biểu diễn liên hệ giữa Y là chỉ tiêu cho một loại hàng hóa và X là tổng thu nhập của người tiêu dùng, giao điểm của đường cong với trục hoành là ngưỡng thu nhập tối hạn mà dưới đó thì người tiêu dùng không mua loại hàng hóa này nữa, có một mức tiêu dùng bão hòa mà tối đa đó thì người tiêu dùng sẽ không mua hàng hóa này nữa cho dù thu nhập có cao bao nhiêu, mức này tạo nên đường tiệm cận phía trên đồ thị.

Nội dung nghiên cứu của chúng ta ở chương này sẽ chủ yếu tập trung vào các đường lý thuyết như ở Hình 11.1c và Hình 11.1d, tức là các mối liên

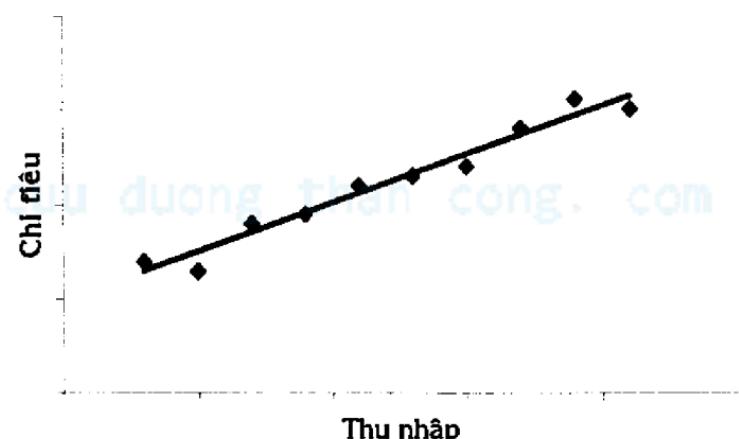
hệ theo dạng đường thẳng giữa X và Y, Với dạng liên hệ phi tuyến giữa X và Y ta sẽ gặp lại ở Chương 12.

11.2 MÔ HÌNH HỒI QUI TUYẾN TÍNH ĐƠN

11.2.1 Mở đầu

Những đồ thị rải điểm ở Hình 11.1 đã mô hình hóa cho chúng ta thấy các dạng liên hệ giữa X và Y, bản chất của mối liên hệ giữa hai biến số này có thể có nhiều hình dạng, từ những dạng đơn giản nhất là dạng đường thẳng có thể mô hình hóa bằng phương trình tuyến tính bậc nhất cho đến những dạng đường cong có hàm số phức tạp. Mỗi liên hệ đơn giản nhất giữa hai đại lượng X và Y là mối liên hệ tuyến tính, ta cụ thể hóa chúng bằng một ví dụ như sau.

Hình 11.2



Hình trên mô tả mối liên hệ giữa chi tiêu theo thu nhập, thông tin được thu thập trên 10 hộ gia đình, lẽ thông thường, chưa xét đến những điều kiện đặc biệt, khi thu nhập tăng chi tiêu sẽ tăng. Qui luật này được thể hiện bằng mối liên hệ tuyến tính thuận chiều, những chấm hình thoi là các cặp giá trị cụ thể của chi tiêu và thu nhập quan sát được trên từng hộ gia đình. Các điểm dữ liệu này gần như nằm trên một đường thẳng và ta có thể làm "thích hợp" chúng bằng một đường thẳng tương đối như trên hình. Đây là đường lý thuyết ta đã nhắc đến ở trên, đường lý thuyết này có dạng tuyến tính thuận chiều, có công thức là

$$E(Y|X_i) = b_0 + b_1 X_i$$

Mối liên hệ giữa hai điểm giá trị bất kỳ X_i và Y_i thực tế được mô tả như sau: $Y_i = b_0 + b_1 X_i + e_i$ (11.1)

Tên gọi đầy đủ của công thức (11.1) là phương trình hồi qui tuyến tính đơn biến tổng thể. Các ký hiệu trong công thức được giải thích như sau:

X_i và Y_i là các giá trị của biến độc lập và biến phụ thuộc tại cặp quan sát thứ i .

Các b_k như đã qui ước là các hệ số hồi qui tổng thể, cụ thể:

b_0 : là hệ số tung độ gốc (hay hệ số chặn)

b_1 : hệ số độ dốc (hay hệ số góc)

e_i là thành phần ngẫu nhiên hay yếu tố nhiễu, nó là chênh lệch giữa giá trị Y_i thực tế và giá trị $E(Y|X_i)$ được xác định từ đường lý thuyết bằng cách thay thế giá trị X_i vào phương trình có dạng $E(Y|X_i) = b_0 + b_1 X_i$; về mặt hình học có thể nhận thấy giá trị của các e_i này được xác định bởi khoảng cách giữa đường lý thuyết và điểm dữ liệu thực tế.

Ở đây ta cần phân biệt rõ giá trị Y_i thực tế với giá trị tính toán được từ đường lý thuyết, đường lý thuyết là đường thẳng thích hợp hóa mối liên hệ giữa X và Y , nó thể hiện mối liên hệ cơ bản giữa X và Y là liên hệ tuyến tính. Giá trị của biến phụ thuộc tính ra từ đường lý thuyết khi thay thế giá trị X_i tương ứng được gọi tên là giá trị trung bình của Y với điều kiện X_i , kí hiệu $E(Y|X_i)$, gọi là giá trị trung bình của Y là vì giá trị $E(Y|X_i)$ này mang tính đại diện, cơ bản cho nhiều giá trị Y_i thực sự có khả năng xảy ra với cùng một mức độ của X_i , do còn các yếu tố ảnh hưởng khác cũng tác động đến Y . Điểm dữ liệu thực tế Y_i (chính là các chấm phân tán trên đồ thị) có sự khác biệt với giá trị do đường lý thuyết tạo ra, sự khác biệt này được thể hiện bằng yếu tố nhiễu e_i , điều này được cụ thể hóa về mặt công thức $e_i = Y_i - E(Y|X_i)$.

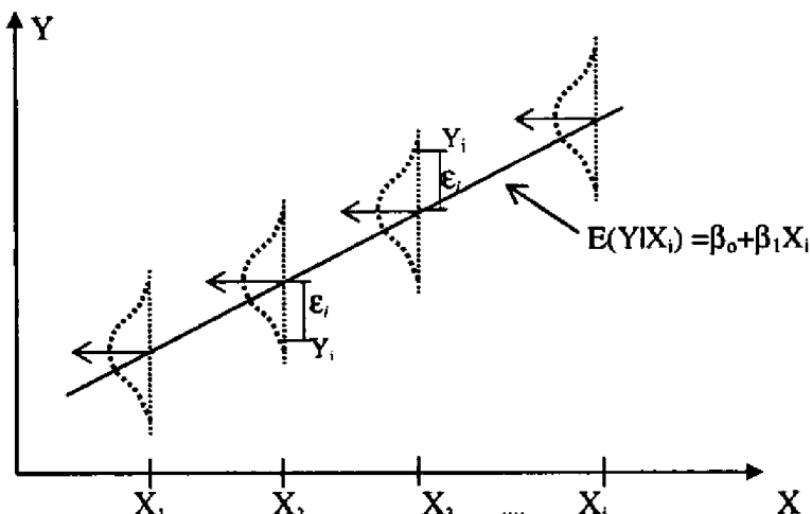
11.2.2 Các giả định liên quan đến yếu tố nhiễu

Với một giá trị cho trước của X có thể xác định nhiều giá trị khác nhau của Y , điều này dẫn đến một suy diễn là với một giá trị X_i cụ thể không phải ta chỉ xác định được một e_i duy nhất mà có thể có nhiều tình huống về e_i , điều này sẽ dễ hiểu hơn nếu bạn nhớ rằng còn có nhiều yếu tố đồng thời tác động đến Y ngoài X , vì vậy với cùng một giá trị X_i có thể có những giá trị Y_i khác nhau do ảnh hưởng của các yếu tố ngoài X này, nhiều Y_i khác nhau thì dẫn đến nhiều e_i khác nhau tuy cùng một X_i , các e_i tại mỗi X_i tạo thành một phân phối bình thường. Các phân phối bình thường của các e_i tại mỗi X_i có phương sai như nhau.

Không có sự tương quan giữa các nhiễu hay nói cách khác các e_i độc lập với nhau.

Các vấn đề vừa đề cập đến ở trên được minh họa rõ hơn qua Hình 11.3 sau.

Hình 11.3



Đường thẳng hồi qui nối liền các giá trị trung bình của Y tại các giá trị khác nhau của biến độc lập X_i , ký hiệu $E(Y|X_i)$, chỉ khi phân phối của nhiễu xung quanh đường hồi qui tại mỗi X_i có phân phối không khống phân phối bình thường thì các suy diễn về đường hồi qui và hệ số hồi qui sau đó mới hợp lý. Đường hồi qui tổng thể xác định bởi 2 giá trị là b_0 và b_1 gọi là các hệ số hồi qui tổng thể. Nếu thỏa những giả định này, các hệ số hồi qui xác định phương trình tổng thể đúng. Với mỗi quan sát, giá trị thật của biến phụ thuộc Y là tổng của hai thành phần: một là thành phần tuyến tính, hai là thành phần ngẫu nhiên.

$$Y_i = b_0 + b_1 X_i + e_i$$

Thành phần tuyến tính Thành phần ngẫu nhiên

Thành phần ngẫu nhiên e_i có thể dương, âm hoặc bằng zero phụ thuộc vào việc Y cụ thể nằm trên, nằm dưới hay nằm ngay trên đường hồi qui tổng thể.

11.2.3 Ý nghĩa của các hệ số hồi qui

- b_1 là hệ số độ dốc của đường hồi qui tổng thể, do lưỡng lượng thay đổi trung bình trong biến phụ thuộc Y , cho mỗi đơn vị thay đổi của X . Hệ số độ dốc tổng thể có thể là dương, bằng zéro hoặc âm phụ thuộc vào mối liên hệ giữa X và Y , giả dụ một hệ số độ dốc bằng 12 có nghĩa là khi X gia tăng 1 đơn vị chúng ta có thể kỳ vọng trung bình Y

tăng 12 đơn vị. ngược lại nếu biết $b_1 = -12$ thì chúng ta kỳ vọng trung bình Y giảm 12 đơn vị cho mỗi đơn vị gia tăng của X.

b_0 là hệ số tung độ gốc (có khi nó còn được gọi là hệ số chặn hay hệ số tự do), cho ta biết giá trị trung bình của Y khi X bằng 0, tuy nhiên sự suy diễn này chỉ hợp lý nếu trong tổng thể X có nhận giá trị 0, khi điều này không xảy ra thì sự diễn giải ý nghĩa của b_0 trong mô hình hồi qui không hợp lý lắm, người ta chỉ có thể coi nó như ảnh hưởng trung bình của tất cả các biến số khác không có mặt trong mô hình mặc dù nó cũng có ảnh hưởng lên Y và thông thường trong diễn dịch ý nghĩa của các hệ số hồi qui người ta không đề cập nhiều đến hệ số tung độ gốc.

Ví dụ: Giám đốc tiếp thị của một công ty có dữ liệu mẫu về 12 đại diện bán hàng được liệt kê trong Bảng 11.1, ông ta cho rằng có một mối liên hệ tuyến tính có ý nghĩa giữa doanh số bán được và số năm kinh nghiệm làm việc với công ty của đại diện bán hàng do đó ông ta quyết định xây dựng một mô hình hồi qui tuyến tính mô tả mối liên hệ này. Mục tiêu kế tiếp của ông ta là dựa trên mô hình hồi qui xây dựng được mà có thể phỏng đoán về doanh số của đại diện bán hàng nếu biết trước số năm kinh nghiệm.

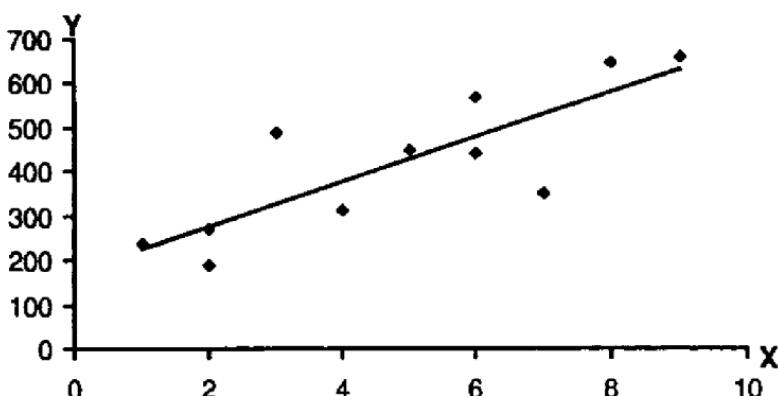
Bảng 11.1

Số tính	Doanh số (triệu đồng)		Số năm	Số tính	Doanh số (triệu đồng)		Số năm
	Y	X			Y	X	
1	487	3	7	346		7	
2	445	5	8	238		1	
3	272	2	9	312		4	
4	641	8	10	269		2	
5	187	2	11	655		9	
6	440	6	12	563		6	

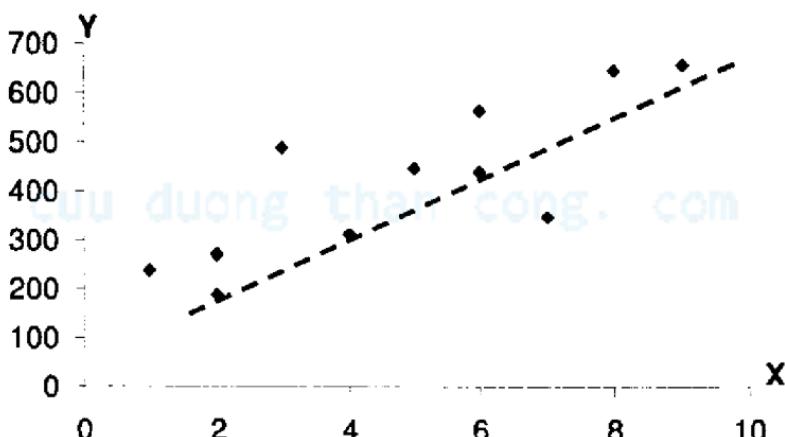
Đồ thị phân tán được vẽ cho hai biến: số năm kinh nghiệm làm việc với công ty và doanh số bán được.

Chúng ta cần dùng dữ liệu mẫu để tính toán các tham số mẫu ước lượng cho b_0 và b_1 là hệ số tung độ gốc và hệ số độ dốc tổng thể của mô hình hồi qui giữa hai biến. Đường thẳng hồi qui xuyên qua các điểm phân tán trên đồ thị là sự ước lượng tốt nhất của đường hồi qui tổng thể. Tuy nhiên ta có thể kẻ nhiều đường thẳng xuyên qua tập dữ liệu này. Xem hình sau mô tả một đường thẳng khác nữa xuyên qua tập dữ liệu của chúng ta, vậy trong hai đường này, thì đường nào có thể sử dụng để ước lượng mô hình hồi qui tuyến tính thật sự của tổng thể?

Hình 11.4



Hình 11.5



Chúng ta sẽ nghiên cứu điều kiện để lựa chọn đường thẳng tốt nhất, đó là điều kiện bình phương bé nhất, điều kiện bình phương bé nhất giúp xác định đường hồi qui dựa trên nguyên tắc *cực tiểu hóa tổng các phần dư bình phương* (còn gọi là nguyên tắc bình phương bé nhất, hay gọi tắt là nguyên tắc OLS – ordinary least square). Nội dung kế tiếp sẽ giúp chúng ta làm rõ về khái niệm phần dư và nguyên tắc OLS.

Dữ liệu chúng ta có trên Bảng 11.1 chỉ là dữ liệu của một mẫu được chọn ngẫu nhiên từ một tổng thể, nếu các giả định liên quan đến mô hình hồi qui ở công thức (11.1) được thỏa mãn, thì hệ số tung độ gốc (kí hiệu b_0) và hệ số độ dốc (kí hiệu b_1) của đường hồi qui mẫu có thể được sử dụng để ước lượng các hệ số hồi qui tổng thể b_0 và b_1 .

Vì thế phương trình hồi qui tuyến tính mẫu được sử dụng để ước lượng mô hình hồi qui tổng thể $E(Y|X_i) = b_0 + b_1 X_i$ sẽ có công thức như sau:

$$\hat{Y}_i = b_0 + b_1 X_i \quad (11.2)$$

Trong đó:

\hat{Y}_i là giá trị ước lượng cho giá trị của biến Y ở quan sát thứ i.

X_i là giá trị của biến X ở quan sát thứ i.

Công thức trên đòi hỏi phải xác định được hai hệ số b_0 và b_1 mới tìm được giá trị ước lượng \hat{Y}_i . Một khi tìm được b_0 và b_1 thì đường thẳng hồi qui mẫu được xác định và có thể chỉ ra nó là đường nào trên đồ thị. Có thể tìm thấy vài cặp giá trị b_0 và b_1 , và vì vậy có thể vẽ được vài đường thẳng hồi qui mẫu, sau đó chúng ta có thể so sánh bằng mắt để xem đường nào phù hợp nhất với các điểm phân tán (đại diện cho dữ liệu mẫu) trên đồ thị bằng cách xem thử đường nào nằm “gần” một cách tương đối với các điểm dữ liệu nhất, đường nào nằm xa nhất. Cách đơn giản nhất là tìm đường thẳng mà sự khác biệt giữa giá trị thực Y_i và giá trị được tìm thấy từ đường hồi qui \hat{Y}_i là nhỏ nhất có thể, tức là xác định $S(Y_i - \hat{Y}_i)^2$ à min, tuy nhiên vì sự khác biệt này có thể mang dấu âm hay dương tùy vị trí của điểm phân tán thực nằm phía nào so với đường thẳng nên người ta phải xác định $S(Y_i - \hat{Y}_i)^2$ à min, công thức này có nghĩa là cực tiểu hóa tổng của các giá trị khác biệt đã được bình phương, mà các khác biệt này chính là phần dư, được ký hiệu e_i , chính là ước lượng trên mẫu của thành phần nhiễu e_i .

$$S(Y_i - \hat{Y}_i)^2 = S[Y_i - (b_0 + b_1 X_i)]^2$$

Trong phương trình trên đây 2 yếu tố b_0 và b_1 chưa biết nên phương trình ở trên trở thành hàm số của b_0 và b_1

Phương pháp bình phương bé nhất là một kĩ thuật để xác định được cặp giá trị b_0 và b_1 sao cho $S(Y_i - \hat{Y}_i)^2$ đạt cực tiểu bằng các phát triển tiếp hàm $S[Y_i - (b_0 + b_1 X_i)]^2$ à min, kết quả tìm được hình thành công thức tính giá trị của các hệ số hồi qui mẫu, được trình bày như sau:

$$b_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} \quad (11.3)$$

Cũng còn một công thức khác để xác định b_1 . Công thức này chỉ là kết quả biến đổi từ công thức trên mà thôi nên kết quả là như nhau, nhưng trong tính toán thủ công nếu dùng công thức thứ hai này sẽ giảm bớt khối lượng phải tính toán.

$$b_1 = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} \quad (11.4)$$

$$\text{Và } b_0 = \bar{Y} - b_1 \bar{X} \quad (11.5)$$

Với ví dụ của chúng ta, giờ đây cần tính hai hệ số hồi qui mẫu b_0 và b_1 , bảng dữ liệu sau sẽ trình bày cách tính toán các thông tin cần thiết để ráp vào công thức tính (thứ hai) b_1 , đây cũng là phương pháp chung để tính được các hệ số hồi qui bằng cách tính thủ công.

Bảng 11.2

STT	Y	X	XY	X^2	Y^2
1	487	3	1461	9	237169
2	445	5	2225	25	198025
3	272	2	544	4	73984
4	641	8	5128	64	410881
5	187	2	374	4	34969
6	440	6	2640	36	193600
7	346	7	2422	49	119716
8	238	1	238	1	56644
9	312	4	1248	16	97344
10	269	2	538	4	72361
11	655	9	5895	81	429025
12	563	6	3378	36	316969
Tổng	4855	55	26091	329	2240687

$$\bar{Y} = \frac{\sum Y}{n} = \frac{4855}{12} = 404,58$$

$$\bar{X} = \frac{\sum X}{n} = \frac{55}{12} = 4,58$$

$$b_1 = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{26091 - \frac{55(4855)}{12}}{329 - \frac{(55)^2}{12}} = 49,91$$

$$b_0 = \bar{Y} - b_1 * \bar{X} = 404,58 - 49,91(4,58) = 175,99$$

Đường hồi qui tìm được nhờ phương pháp bình phương bé nhất vì vậy được viết như sau: $\hat{Y}_i = 175,99 + 49,91X_i$

11.2.4 Tính toán các kết quả hồi qui bằng phần mềm Excel

Trên thực tế người ta có thể sử dụng một chương trình phổ thông là Excel để tính toán các hệ số hồi qui và thiết lập phương trình hồi qui mẫu. Cách thức tiến hành như sau:

- Mở cửa sổ làm việc của Excel, nhập dữ liệu về doanh số và số năm kinh nghiệm vào hai cột của sheet hiện hành, rồi từ menu Tools, chọn lệnh Data analysis, từ cửa sổ Data Analysis lựa chọn tiếp lệnh Regression để mở cửa sổ hộp thoại Regression, sau đó nhập vào các khung địa chỉ của các ô chứa dữ liệu.
- Nhập địa chỉ của biến Y vào Input Y Range (ở ví dụ này là A2:A14)
- Nhập địa chỉ của biến X vào Input X Range (ở ví dụ này là B2:B14)
- Nhớ nhấp chọn mục Labels vì ở trên bạn đã đưa cả ô chứa tiêu đề X và Y vào khung Input
- Bạn có thể để các lựa chọn khác ở chế độ mặc định, ví dụ độ tin cậy là 95%, hoặc chọn các tùy chọn theo ý mình. Trong mục New Worksheet Ply bạn có thể đặt tên cho worksheet chứa kết quả nếu muốn
- Nhấp nút OK để hoàn tất.

Hình 11.6

The screenshot shows the 'Regression' dialog box in Excel. On the left, there is a table with data in columns A and B. Column A is labeled 'Doanh số' and column B is labeled 'Số năm'. The data points are: (487, 3), (445, 5), (272, 2), (641, 8), (187, 2), (440, 6), (346, 7), (238, 1), (312, 4), (269, 2), (655, 9), and (563, 6). The 'Regression' dialog box has several tabs: 'Input' (selected), 'Output options', 'Residuals', and 'Normal Probability'. Under 'Input', 'Labels' is checked. Under 'Output options', 'New Worksheet Ply' is selected. Other options like 'Output Range' and 'New Workbook' are also available. The 'OK' button is at the top right of the dialog box.

Trên màn hình Excel chúng ta sẽ có kết quả như sau:

Bảng 11.3

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,8325
R Square	0,6931
Adjusted R Square	0,6624
Standard Error	92,1055
Observations	12

ANOVA

	df	SS	MS	F	Significance F
Regression	1	191601	191601	22,585	0,00078
Residual	10	84834,3	8483,4		
Total	11	276435			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	→ 175,8288	54,9899	3,1975	0,0095	53,3037	298,3539
X	→ 49,9101	10,5021	4,7524	0,0008	26,5100	73,3102

Trên bảng cuối cùng, mục Coefficients (nghĩa là hệ số hồi qui) ta xem theo hàng ngang, ở vị trí của hàng thứ nhất (tương đương với chữ "Intercept") là giá trị của hệ số tung độ gốc, xấp xỉ với giá trị chúng ta vừa tính được bằng phương pháp thủ công ở trên. Ở vị trí của hàng thứ hai (tương đương với chữ X) là giá trị của hệ số độ dốc. Như vậy khi bạn đọc kết quả, hàng đầu tiên kí hiệu tên biến độc lập bằng chữ gì thì giá trị trên cột thứ hai chính là hệ số độ dốc đứng trước biến đó, không kể đến hàng đầu tiên luôn là vị trí của hệ số chặn Intercept. Từ phương trình hồi qui ước lượng trên dữ liệu mẫu $\hat{Y}_i = 175,8288 + 49,9101X_i$, ta phát biểu về ý nghĩa của các hệ số hồi qui của phương trình như sau:

- Hệ số độ dốc cho biết khi số năm kinh nghiệm làm việc với công ty tăng thêm 1 năm thì doanh số sẽ tăng trung bình khoảng 49,9101 triệu đồng.
- Hệ số tung độ gốc cho biết khi số năm kinh nghiệm bằng zero tức là một đại diện bán hàng vừa mới làm việc với công ty thì cũng vẫn có thể đạt được một doanh số trung bình khoảng 175,8288 triệu đồng. Nhưng như đã nói, cách diễn đạt này về hệ số tung độ gốc không hẳn đã hợp lý.

11.2.5 Vấn đề cần chú ý khi dự đoán với mô hình hồi qui

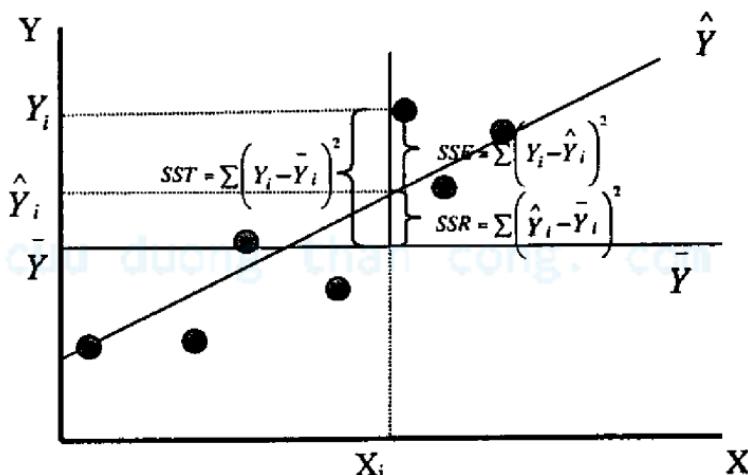
Khi sử dụng mô hình hồi qui để dự đoán, cần chú ý rằng bạn chỉ nên dự đoán trong phạm vi những giá trị của biến độc lập X (từ giá trị nhỏ nhất đến giá trị lớn nhất) trên dữ liệu mẫu đã được sử dụng xây dựng đường hồi qui. Không nên ngoại suy Y với những giá trị X nằm ngoài quá xa khoảng giá trị của X, như với ví dụ của chúng ta thì bạn không nên dự đoán doanh số đạt được với những giá trị X quá lớn hơn 9 năm kinh nghiệm hay nhỏ hơn 1 năm kinh nghiệm.

Ở ví dụ trên, với một đại diện bán hàng có số năm kinh nghiệm bất kỳ ví dụ 5 năm thì ước đoán hay dự đoán tốt nhất cho doanh số tương ứng là $\hat{Y} = 175,8228 + 49,9101 \times 5 = 425,3733$ triệu đồng

11.2.6 Đo lường biến thiên bằng Hệ số xác định

Để khảo sát khả năng sử dụng biến độc lập để dự đoán về biến phụ thuộc cần phải đo lường một số sự biến thiên trong mô hình. Sự biến thiên đầu tiên được gọi tên là tổng biến thiên của biến phụ thuộc (kí hiệu SST) được tính bằng cách lấy tổng chênh lệch bình phương của các giá trị Y_i xung quanh trung bình của chúng. Trong phân tích hồi qui, tổng biến thiên được chia làm hai phần là biến thiên giải thích được và biến thiên không giải thích được, hay còn được gọi tên khác là biến thiên của hồi qui (SSR) và biến thiên của phần dư (SSE). Các thành phần này được biểu diễn hình học như sau:

Hình 11.7



SSR đại diện cho khác biệt giữa giá trị do đường hồi qui tính toán được \hat{Y}_i

và \bar{Y} (giá trị trung bình của Y). SSE đại diện cho thành phần biến thiên trong Y mà không được giải thích bởi hồi qui, nó được hình thành dựa trên chênh lệch giữa Y_i và \hat{Y}_i . Còn SST là chênh lệch giữa mỗi giá trị quan sát Y_i và \bar{Y} .

Các biến thiên được thể hiện về mặt công thức như sau:

$$SST = SSR + SSE$$

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (11.6)$$

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (11.7)$$

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (11.8)$$

Tự bản thân SST, SSR hay SSE cũng cung cấp cho chúng ta ít nhiều thông tin để kết luận về sự biến thiên trong mô hình hồi qui của chúng ta để từ đó mà suy diễn, tuy nhiên tỷ lệ giữa SSR và SST cho chúng ta một đại lượng đánh giá mô hình hồi qui tốt hơn, đại lượng này được gọi tên là hệ số biến thiên (kí hiệu R^2) được xác định bằng công thức như sau:

$$R^2 = \frac{SSR}{SST} \quad (11.9)$$

Hệ số xác định đo lường tỷ lệ biến thiên của Y được giải thích bởi biến độc lập X trong mô hình hồi qui. Ý nghĩa nôm na của nó có thể diễn đạt như sau: có phải thông thường để mô tả, ước lượng về một đại lượng nào đó thì số trung bình là đại lượng hay được nghĩ đến nhất (khi bạn không có thông tin gì khác), tất nhiên dùng số trung bình để ước lượng cho vấn đề thì không tránh khỏi sai lệch, và SST chính là chênh lệch giữa Y_i và \bar{Y} khi ta dùng giá trị trung bình để ước lượng cho Y. Khi bạn xây dựng được đường hồi qui mô tả Y theo X tức là bạn đã có thông tin khác để ước lượng về Y chứ không còn phải dùng đến \bar{Y} nữa. Vậy thông tin này tốt đến đâu so với lúc dùng trị trung bình để ước lượng, sự tốt hơn đến đâu này được thể hiện qua tỉ số R^2 .

Chúng ta trở lại với ví dụ về doanh số và số năm kinh nghiệm, muốn tính hệ số biến thiên ta cần tính được SSR và SST. Để áp dụng công thức tính SSR, SST và SSE ở trên ta cần tính các cột số liệu sau, chú ý là với dữ liệu tính toán bằng máy tính tay thì đầu tiên cần phải tính thêm số liệu cho cột $\hat{Y}_i = 175,99 + 49,91 \times X_i$

Bảng 11.4

STT	Y	X	\hat{Y}_i	$(\hat{Y}_i - \bar{Y})^2$	$(Y_i - \bar{Y})^2$	$(Y_i - \hat{Y}_i)^2$
1	487	3	325,72	6218,8996	6793,0564	26011,2384
2	445	5	425,54	439,3216	1633,7764	378,6916
3	272	2	275,81	16581,7129	17577,4564	14,5161
4	641	8	575,27	29135,0761	55894,4164	4320,4329
5	187	2	275,81	16581,7129	47341,0564	7887,2161
6	440	6	475,45	5022,5569	1254,5764	1256,7025
7	346	7	525,36	14587,8084	3431,6164	32170,0096
8	238	1	225,9	31926,5424	27748,8964	146,4100
9	312	4	375,63	838,1025	8571,0564	4048,7769
10	269	2	275,81	16581,7129	18381,9364	46,3761
11	655	9	625,18	48664,3600	62710,1764	889,2324
12	563	6	475,45	5022,5569	25096,8964	7665,0025
Tổng	4855	55		191600,3631	276434,9168	84834,6051

Hàng dưới cùng của cột thứ 5 của Bảng 11.4 cho ta kết quả về SSR, ở cột kế tiếp là SST và cuối cùng là SSE. Nhớ rằng $SST = SSR + SSE$ nên chúng ta có thể tính SSE theo con đường thứ hai này, ta thử lấy $276434,9168 - 191600,3631 = 84834,55$ tức là gần bằng đúng đáp số tính theo công thức chính thức.

Thực ra Excel cũng đã tính toán sẵn cho chúng ta các đáp số này ngay trong quá trình thực hiện lệnh ANOVA, trở lại Bảng 11.3, trích bảng con thứ 2 có tên ANOVA ra xem xét, chúng ta thấy

ANOVA

	df	SS	MS	F	Significance F
Regression	1	191600,6220	191600,6220	22,585	0,000777
Residual	10	84834,2947	8483,4295		
Total	11	276434,9167			

Cột đầu tiên của bảng liệt kê tên của các đối tượng mà bảng tính toán, chữ Regression chính là ký hiệu cho thành phần SSR, chữ Residual đại diện cho thành phần SSE và chữ Total đại diện cho SST, dò theo hàng ngang này đến cột thứ 3 bạn sẽ thấy các số liệu tương ứng mà Excel đã tính toán, chúng gần bằng kết quả tính thủ công của chúng ta, sự chênh lệch xảy ra là do việc làm tròn số trong tính thủ công mà thôi.

Tính hệ số xác định:

$$R^2 = \frac{SSR}{SST} = \frac{191600,6220}{276434,9167} = 0,6931$$

Hệ số xác định bằng 0,6931 cho ta biết 69,31% biến thiên doanh số của các đại diện bán hàng có thể giải thích được bởi biến thiên trong số năm kinh nghiệm làm việc với công ty. Đây là một ví dụ về mối liên hệ tuyến tính thuận chiều tương đối mạnh giữa hai biến vì việc sử dụng mô hình hồi qui đã làm giúp dự đoán được 69,31% doanh thu của các đại diện bán hàng căn cứ trên số năm kinh nghiệm, còn $(100 - 69,31)\% = 30,69\%$ biến thiên trong doanh thu mà việc sử dụng số năm kinh nghiệm không giải thích được.

Lệnh Regression cũng đã tính toán luôn cho chúng ta величина Hệ số xác định, xem trong bảng đầu tiên có tên Regression Statistics, tại cột thứ nhất có величина R Square, dò sang cột thứ hai bạn sẽ thấy giá trị 0,6931.

Regression Statistics	
Multiple R	0,8325
R Square	0,6931
Adjusted R Square	0,6624
Standard Error	92,1055
Observations	12

11.2.7 Sai số chuẩn của ước lượng

Việc kế tiếp là chúng ta phải xác định sai số trong sự ước lượng, điều này đòi hỏi chúng ta phải biết được s^2 là phương sai của các nhiễu e_i , để đạt mục đích này có vẻ hợp lý nhất là ta dùng ước lượng trên mẫu của s^2 là s^2_{yx} , ta lập luận như sau:

Mặc dù phương pháp bình phương bé nhất đưa đến một đường thẳng phù hợp nhất với tập dữ liệu trên cơ sở cực tiểu hóa tổng độ lệch bình phương, nhưng trừ khi tất cả những điểm dữ liệu quan sát thực tế đều nằm chính xác trên đường thẳng này, còn không thì hàm hồi qui vẫn không phải là một ước lượng hoàn hảo. Và điều này khó mà xảy ra, từ đó cần phải tính toán một đại lượng thống kê đo lường sự chênh lệch của giá trị Y thực và giá trị Y do đường hồi qui tính toán ra. Cũng cùng một ý tưởng như khi tính toán độ lệch chuẩn như một đại lượng đo lường sự biến thiên của mỗi quan sát xung quanh trị trung bình của nó, độ lệch chuẩn xung quanh đường hồi qui được gọi là sai số chuẩn của hồi qui (ký hiệu s_{yx}) được tính bằng cách lấy tổng của các chênh lệch bình phương chia cho bậc tự do rồi lấy căn bậc hai kết quả tìm được (dĩ nhiên lúc này là chênh lệch giữa giá trị Y thực tế và giá trị Y do đường hồi qui ước lượng được, cũng như ý tưởng tính chênh lệch giữa các giá trị quan sát thực tế và trị trung bình của nó khi tính phương sai rồi lấy căn bậc hai để được độ lệch chuẩn).

$$s_{Y/X} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} \quad (11.10)$$

Bình phương $s_{Y/X}$ ta được $s^2_{Y/X}$ là ước lượng tốt cho s^2 căn cứ trên $(n-2)$ bậc tự do. Chúng ta để ý rằng $\hat{Y}_i = b_0 + b_1 X_i$, do đó trong phép tính $s_{Y/X}$ ta cần hai tham số ước lượng b_0 và b_1 , kết quả là bậc tự do của s sẽ bớt đi hai, nghĩa là còn $(n-2)$.

Chỉ số Y/X dùng để chỉ rõ $s^2_{Y/X}$ là ước lượng cho phương sai của Y khi ta có sự hồi qui Y theo X.

Vận dụng lại ví dụ doanh số - số năm kinh nghiệm

$$s_{Y/X} = \sqrt{84834,6051/(12-2)} = \sqrt{8483,4605} = 92,1057$$

Sai số chuẩn của hồi qui này bằng 92,1057 triệu đồng cũng được tính toán sẵn với lệnh Regression của Excel trong bảng Regression Statistics, bạn hãy xem bảng dưới và so sánh kết quả mà phần mềm tính toán với kết quả chúng ta tính thủ công.

<i>Regression Statistics</i>	
Multiple R	0,8325
R Square	0,6931
Adjusted R Square	0,6624
Standard Error	92,1055
Observations	12

Sai số chuẩn của ước lượng thể hiện sự đo lường biến thiên xung quanh đường hồi qui. Nó cùng đơn vị tính với biến Y. Sự diễn giải ý nghĩa của sai số chuẩn của ước lượng tương tự như độ lệch chuẩn, cũng như độ lệch chuẩn đo lường sự biến thiên của các quan sát thực tế xung quanh trung bình, sai số chuẩn của ước lượng đo lường sự biến thiên của các giá trị Y thực tế xung quanh đường hồi qui. Sai số càng lớn thì biến thiên càng nhiều, mà biến thiên càng nhiều thì đường hồi qui càng ít sát với các điểm dữ liệu. Trong các nội dung sau bạn sẽ thấy rõ hơn công dụng của sai số chuẩn của ước lượng được dùng để đánh giá có tồn tại không mối liên hệ thống kê có ý nghĩa giữa hai biến cũng như tiến hành những suy diễn về giá trị tương lai của Y.

11.2.8 Suy diễn thống kê về hệ số độ dốc

11.2.8.1 Định lý Gauss – Markov

Vì hệ số độ dốc của mô hình hồi qui tổng thể tức là b_1 cho ta biết sự liên

hệ giữa X và Y, giả dụ nếu hệ số này bằng 0 tức là X và Y chẳng có liên hệ gì cả, nên ta sẽ thực hiện một số suy diễn thống kê về hệ số độ dốc căn cứ trên ước lượng trên mẫu của nó là b_1 , lúc này chúng ta cần một vài giả định về sự phân phối của các dữ kiện, nói rõ hơn ta muốn biết các ước lượng trên mẫu b_1 phân phối quanh b_1 như thế nào.

Nếu ta giả định là sự phân phối của Y bình thường thì kết quả là các tham số ước lượng như b_0 và b_1 cũng sẽ phân phối bình thường. Người ta có thể chứng minh rằng trung bình và phương sai b_1 lần lượt bằng:

$$\text{Trung bình: } E(b_1) = b_1 \quad (11.11)$$

$$\text{Phương sai: } s_{b_1}^2 = \frac{s_{Y/X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (11.12)$$

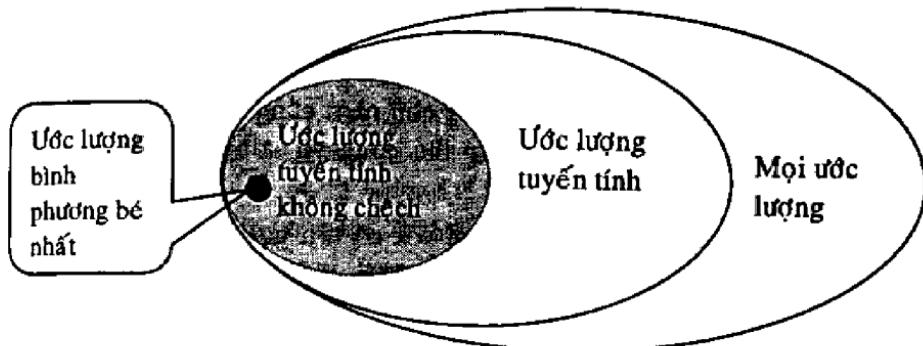
Mẫu số của công thức trên có thể đổi thành một dạng thuận tiện hơn cho tính toán là $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$

Định lý Gauss – Markov phát biểu như sau:

Trong các ước lượng tuyến tính không chêch cho hệ số hồi qui tổng thể, ước lượng tìm được bằng phương pháp bình phương bé nhất có phương sai cực tiểu.

Để giải thích định lý này chúng ta hãy xem xét b_1 là ước lượng bình phương bé nhất của b_1 . Như trên ta đã biết b_1 là ước lượng tuyến tính, và chúng ta cũng sẽ tự giới hạn trong các ước lượng tuyến tính. Hơn nữa trong các ước lượng tuyến tính chúng ta chỉ xem xét ước lượng không chêch. Ước lượng bình phương bé nhất không những ở trong loại ước lượng trên mà trong mọi ước lượng cùng loại nó còn có phương sai cực tiểu. Do đó người ta thường gọi b_1 là ước lượng không chêch tuyến tính tốt nhất của b_1 . Ta có thể biểu thị điều vừa nói trên trong giản đồ như ở trang sau.

Sau khi đã thiết lập trung bình, phương sai của b_1 , bây giờ ta có thể suy diễn về tham số tổng thể b_1 vì một lần nữa nhắc lại là b_1 cho ta biết về sự liên hệ giữa X và Y.



11.2.8.2 Khoảng tin cậy cho hệ số độ dốc

Khoảng tin cậy $100(1-\alpha)\%$ cho hệ số độ dốc b_1 cũng có cách tính như bất cứ một khoảng tin cậy cho tham số thống kê tổng thể nào tức là có dạng $[b_1 \pm t_{(n-2; \alpha/2)} \times s_{b_1}]$

Như vậy nếu muốn tìm khoảng tin cậy 95% cho hệ số độ dốc tổng thể cho doanh số trong ví dụ thì ta làm như sau:

- Tính sai số ước lượng của hệ số độ dốc, dùng những thông tin ở Bảng 11.2 để tính toán giá trị ở mẫu số

$$s_{b_1}^2 = \frac{s_{r/x}^2}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{(92,1057)^2}{329 - \frac{55^2}{12}} = \frac{8483,4599}{76,9167} = 110,2941$$

vậy $s_{b_1} = \sqrt{110,2941} = 10,5021$

- Giá trị $t_{(n-2; \alpha/2)} = t_{(12-2; 0,05/2)} = 2,228$

Như vậy khoảng tin cậy 95% cho giá trị thực của hệ số độ dốc được xác định là

$$(49,9101 \pm 2,228 \times 10,5021) = (49,9101 \pm 23,3987) = (26,5114; 73,3088)$$

Một lần nữa bạn yên tâm là tất cả những thông tin này đều được phần mềm Excel tính toán cũng chỉ với một lệnh Regression thôi, quay lại với bảng kết quả 11.3 nhưng bây giờ cắt riêng phần cuối cùng của bảng nơi chứa những thông tin liên quan đến việc tính toán vừa rồi của chúng ta.

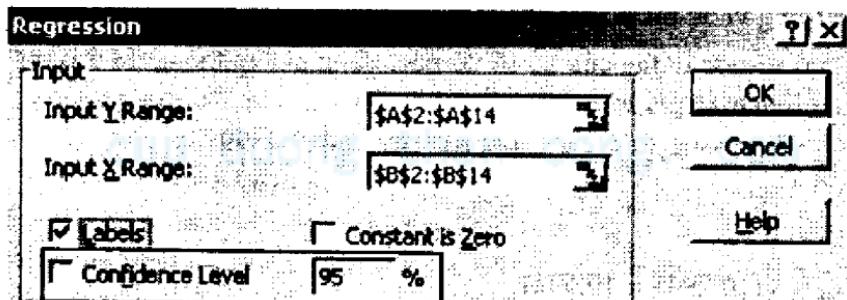
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	175,8288	54,9899	3,1975	0,0095	53,3037	298,3539
X	49,9101	10,5021	4,7524	0,0008	26,5100	73,3102

Cột thứ 3 có tên là Standard Error chính là sai số chuẩn của các hệ số hồi qui, hàng tương ứng với vị trí của biến X chứa thông tin về sai số chuẩn của hệ số độ dốc, số liệu trong khung chữ nhật cũng chính bằng số liệu chúng ta đã tính bằng phương pháp thủ công về ước lượng của sai số chuẩn của hệ số độ dốc.

Nhìn về cuối bảng chúng ta thấy hai cột có tên gọi là Lower 95% và Upper 95% trong đó Lower chính là cận dưới và Upper chính là cận trên của khoảng tin cậy 95% của hệ số hồi qui tổng thể, bạn có thể thấy chúng giống kết quả tính thủ công của chúng ta tới 2 số lẻ sau dấu phẩy, phần chênh lệch còn lại chỉ là do việc làm tròn số trong tính toán mà thôi.

Vậy nếu bạn cần tìm khoảng tin cậy 90% cho hệ số hồi qui tổng thể thì sao, chỉ cần có một lựa chọn phù hợp khi thực hiện lệnh Regression mà thôi, đó là bạn chọn sang mục Confidence Level rồi sửa giá trị mặc định 95% thành 90% hoặc giá trị nào khác về độ tin cậy mà bạn muốn.

Hình 11.8



11.2.8.3 Kiểm định ý nghĩa của hệ số độ dốc

Thực ra kiểm định thông tin về hệ số hồi qui tổng thể có thể tiến hành với giả thuyết bất kỳ về giá trị của b_1 , giả dụ $H_0: b_1 = b_*$, như mọi bài toán kiểm định giả thuyết thống kê thông thường.

- Trước hết ta chuẩn hóa b_1 theo công thức chuẩn hóa thông thường

$$Z = (b_1 - b_*) / s_{b_1} \quad (11.13)$$

Chú ý rằng (ở công thức tính phuơng sai của hệ số b_1) vì ta đã dùng ước lượng trên mẫu $s_{b_1}^2$ thay cho phuơng sai thực trên tổng thể mà ta chưa biết nên b_1 thay vì có phân phối bình thường sẽ có phân phối t với $n-2$ bậc tự do vì ước lượng s_{b_1} được tính từ tổng số bình phuơng độ lệch từ đường hồi qui với $n-2$ bậc tự do, vậy

$$t = \frac{b_1 - \beta_*}{s_{b_1}} \quad (11.14)$$

Trong đó:

b_1 là hệ số hồi qui mẫu

β_* là giá trị của hệ số hồi qui tổng thể được giả định

s_{b_1} là ước lượng của sai số chuẩn của hệ số độ dốc

- Tiến hành so sánh giá trị t này với giá trị t tra bảng theo qui tắc nếu $|t| < t_{(n-2, \alpha/2)}$ thì chúng ta không thể bác bỏ giả thuyết H_0

Đối với một mô hình hồi qui đơn biến (chỉ có một biến độc lập) việc kiểm định thông tin về hệ số độ dốc trong mô hình hồi qui thường được tiến hành với giả thuyết rằng $b_1 = 0$. Bản chất của giả thuyết này là nhằm đặt ra một giả định rằng phải chăng trên thực tế X không có ảnh hưởng gì đến Y , X và Y chẳng có liên hệ gì cả và giá trị b_1 ta nhận được trên đường hồi qui mẫu chỉ là kết quả một sự tình cờ do lấy mẫu mà thôi, vì thế người ta gọi kiểm định này là kiểm định ý nghĩa của hệ số hồi qui. Giả thuyết đặt ra là:

$$H_0: b_1 = 0$$

$$H_1: b_1 \neq 0$$

Ví dụ: Trở lại ví dụ doanh số - số năm kinh nghiệm của chúng ta, để kiểm định giả thuyết về ý nghĩa của hệ số hồi qui tổng thể chúng ta tính toán giá trị:

$$t = \frac{b_1}{s_{b_1}} = \frac{49,9101}{10,5021} = 4,7524$$

Đem so sánh giá trị t vừa tính toán được với giá trị $t_{(12-2, 0,05/2)} = 2,228$ ta thấy $|t| > t_{(n-2, \alpha/2)}$ nên ta bác bỏ giả thuyết H_0 và do đó kết luận được với độ tin cậy 95% là hệ số hồi quy tổng thể khác không. Hay kết luận theo một cách khác rằng thực sự số năm kinh nghiệm làm việc với công ty có ảnh hưởng đến doanh số của các đại diện bán hàng.

Trong bảng kết quả mà Excel cung cấp cũng có cả giá trị t tính toán cho kiểm định ý nghĩa của hệ số hồi qui, đó là thông tin trên cột thứ tư có tên $t Stat$, các bạn sẽ thấy giá trị 4,7524 ở dòng thứ 2, đó cũng chính bằng giá trị chúng ta tính toán bằng phép tính thủ công.

Đồng thời Excel cũng tiến hành tính toán luôn giá trị p-value tương ứng của giá trị t tính toán = 4,7524 được kết quả là 0,0008. Mức ý nghĩa này nhỏ hơn rất nhiều so với mức ý nghĩa 5% ta đã chọn cho phép kiểm định nên ta cũng đi tới cùng một kết luận là bác bỏ giả thuyết H_0 .

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower</i>	<i>Upper</i>
Intercept	175,8288	54,9899	3,1975	0,0095	53,3037	298,3539
X	49,9101	10,5021	4,7522	0,0008	26,5100	73,3102

11.2.9 Phân tích phần dư

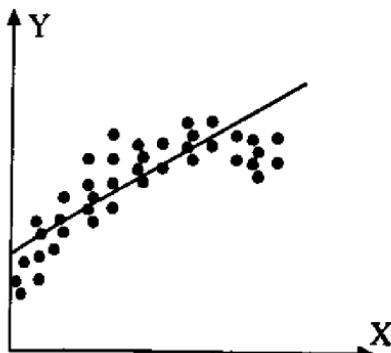
Trong nội dung này, cách phân tích phần dư qua phương pháp tiếp cận chủ yếu bằng đồ thị được sử dụng để đánh giá xem mô hình hồi qui tuyến tính chúng ta đã xây dựng trên dữ liệu thực tế có phải là một mô hình hợp lý không, ngoài ra sự phân tích phần dư này cũng có thể giúp chúng ta suy đoán xem có khả năng xảy ra sự vi phạm giả định nào đó của mô hình hồi qui không.

11.2.9.1 Kiểm tra tính đúng đắn của mô hình hồi qui tuyến tính

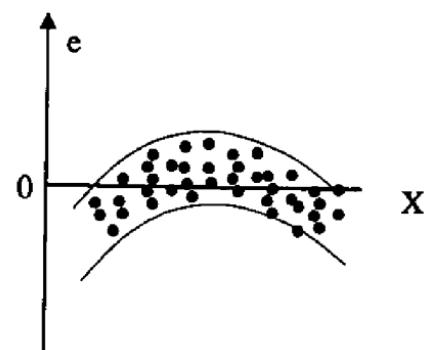
Phần dư hay giá trị sai số e_i được định nghĩa như là sự khác biệt giữa giá trị Y_i quan sát được và giá trị \hat{Y}_i , ước lượng từ đường hồi qui với một giá trị X_i xác định. Nói một cách hình ảnh, trên đồ thị phân tán phần dư được mô tả là khoảng chênh lệch theo phương thẳng đứng giữa giá trị Y thực tế và đường thẳng được định nghĩa bởi phương trình hồi qui tuyến tính đơn biến chúng ta đã xây dựng được bằng phương pháp bình phương bé nhất trên tập dữ liệu. Về mặt định lượng, phần dư được định nghĩa bằng công thức $e_i = (Y_i - \hat{Y}_i)$. Tính đúng đắn của mô hình hồi qui được xem xét bằng cách vẽ đồ thị mà phần dư được đặt trên trục đứng và biến độc lập X được đặt ở trục ngang. Nếu mô hình đã xây dựng là đúng đắn thì đồ thị e_i theo X sẽ không thể hiện một hình dạng rõ ràng nào của các chấm phân tán. Ngược lại, nếu mô hình đã xây dựng không đúng đắn thì sẽ có một dạng liên hệ nào đó giữa X_i và e_i .

Xem xét Hình 11.9, Hình 11.9a mô tả tình huống mà Y gia tăng theo chiều tăng của X , tuy nhiên mối liên hệ này dường như có dạng đường cong vì đi lên phía trên xu thế chung có vẻ chững lại rồi giảm đi mặc dù X vẫn tiếp tục tăng. Điều này cho nhận định đường cong bậc hai có lẽ phù hợp với mối liên hệ giữa X và Y hơn là đường thẳng hồi qui tuyến tính bậc một. Liên hệ bậc hai này càng nổi bật hơn ở Hình 11.9 b, trong đồ thị hình b chúng ta thấy một liên hệ rõ ràng giữa X_i và e_i . Trên đồ thị phần dư, xu hướng phi tuyến tính của X và Y đã lộ ra ra, bằng cách này chúng ta đã chỉ rõ được sự kém phù hợp của mô hình hồi qui tuyến tính, vì vậy mô hình bậc hai trong trường hợp này tốt hơn và nên được sử dụng thay vì mô hình tuyến tính. Nếu mô hình tuyến tính thực sự phù hợp như tình huống đồ thị c thì đồ thị phần dư theo X sẽ có dạng như đồ thị hình d.

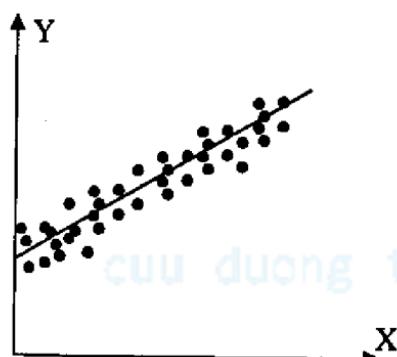
Hình 11.9



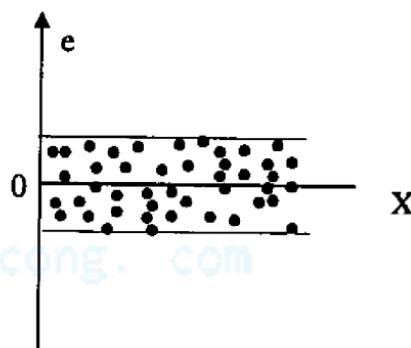
Hình a



Hình b



Hình c

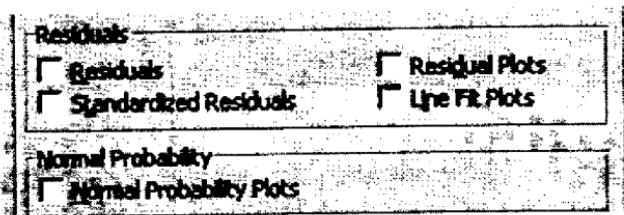


Hình d

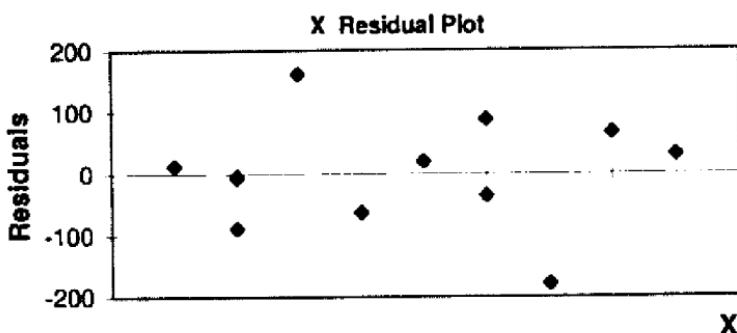
Vậy để xác định xem mô hình hồi qui tuyến tính có phù hợp hay không, người ta sẽ vẽ đồ thị phần dư theo biến độc lập X. Hoặc có thể vẽ đồ thị của phần dư theo các giá trị dự đoán của biến phụ thuộc (\hat{Y}_i), cả hai đồ thị phần dư này đều có tác dụng như nhau trong việc xem xét mô hình hồi qui tuyến tính có phù hợp với tập dữ liệu quan sát. Với Excel chúng ta có thể yêu cầu thực hiện việc này một cách tự động ngay trong quá trình thực hiện lệnh Regression để xây dựng mô hình hồi qui tuyến tính giữa doanh số của đại diện bán hàng và số năm kinh nghiệm làm việc với công ty bằng cách lựa chọn như sau:

Trong khu vực kế dưới cùng của cửa sổ lệnh Regression (khu vực Residuals) chúng ta chọn mục Residual Plots bằng cách nhấp chuột vào khung vuông nằm trước chữ Residual Plots. Sau đó trong các kết quả mà lệnh cho ra, chúng ta sẽ có đồ thị như Hình 11.11

Hình 11.10



Hình 11.11



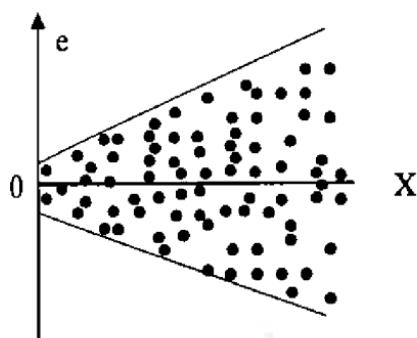
Quan sát đồ thị chúng ta thấy các chấm dữ liệu phân tán khá rộng trong đồ thị phần dư xung quanh trục 0, không có một dạng hình cụ thể nào cho mối liên hệ giữa X_i và e_i . Các phần dư dường như bố trí ở cả hai bên trên và dưới của trục 0 cho các giá trị khác nhau của X . Nhận xét đó cho chúng ta kết luận rằng mô hình dường thẳng chúng ta đã xây dựng là phù hợp đối với tập dữ liệu về doanh số của chúng ta.

11.2.9.2 Kiểm tra sự vi phạm giả định phương sai bằng nhau

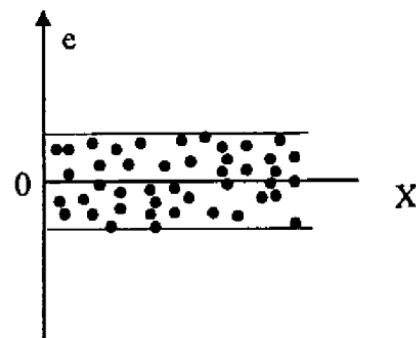
Trong các giả định liên quan đến phần dư, giả định thứ ba bảo rằng phương sai của các phần dư bằng nhau tại các giá trị khác nhau của biến độc lập. Giả định này còn được gọi tên theo cách khác là giả định phương sai không đổi. Việc chuẩn đoán xem giả định phương sai không đổi có bị vi phạm không cũng có thể được thực hiện ngay trên đồ thị phần dư theo biến độc lập ở trên. Với ví dụ của chúng ta, đồ thị ở Hình 11.11 không thể hiện một sự khác biệt lớn nào trong sự biến thiên của phần dư tại các giá trị khác nhau của biến X , vì thế chúng ta có thể kết luận rằng mô hình hồi qui tuyến tính chúng ta đã xây dựng tỏ ra không vi phạm một cách rõ ràng giả định về sự bằng nhau của phương sai của phần dư tại mọi mức độ của biến độc lập X .

Để quan sát một tình huống trong đó giả định phương sai không đổi bị vi phạm các bạn xem xét đồ thị có tính giả thuyết giữa phần dư và biến độc lập dưới đây. Trong đồ thị Hình 11.12a có một dải hình quạt thể hiện biến thiên của phần dư càng lúc càng gia tăng theo chiều tăng của X. Điều đó chứng tỏ giả định phương sai không đổi đã bị vi phạm, cũng đồng thời thể hiện sự dao động của giá trị thực tế Y_i quanh đường hồi qui đã lớn dần theo chiều tăng của X. Đồ thị 11.12b lại thể hiện tình huống phương sai không đổi. Nó chính là đồ thị đã nghiên cứu để đánh giá sự đúng đắn của mô hình hồi qui tuyến tính đã xây dựng.

Hình 11.12



Hình a



Hình b

Ở nội dung chương 12 bạn sẽ gặp tình huống mô hình hồi qui với không chỉ một biến giải thích, nó có tên là mô hình hồi qui bội, nếu với nhiều biến giải thích ta có thể vẽ phần dư theo từng biến giải thích mà ta ngờ gây ra hiện tượng phương sai thay đổi hoặc tốt hơn là vẽ phần dư theo \hat{Y} là giá trị ước lượng được từ mô hình.

Trên thực tế không có một phương pháp chắc chắn nào để phát hiện phương sai thay đổi mà chỉ có thể dùng một vài phương pháp chuẩn đoán thôi, ngoài phương pháp dùng đồ thị chúng ta còn có kiểm định sau để tìm phương sai thay đổi:

Kiểm định Park: bao gồm các bước sau

1. Ước lượng hồi qui gốc
2. Tính sai số e_i của hồi qui gốc sau đó bình phương chúng lấy e_i^2 rồi sau đó lấy $\ln(e_i^2)$
3. Ước lượng mô hình $\ln(e_i^2) = a_1 + a_2 * \ln(X_i) + n_i$ với X_i là biến độc lập của hồi qui gốc bằng dữ liệu trên mẫu. Nếu ở tình huống

hồi qui đa biến ta có nhiều biến giải thích thì chúng ta hồi qui mô hình trên với từng biến giải thích để tìm phương sai thay đổi do biến nào gây ra. Hoặc đơn giản hơn là hồi qui $\ln(e_i^2)$ theo \hat{Y}

- Kiểm định giả thuyết $H_0: a_2 = 0$ nếu cho thấy a_2 có ý nghĩa về mặt thống kê sẽ cho kết luận là có phương sai thay đổi trong số liệu, bằng ngược lại ta có thể kết luận phương sai không thay đổi.

Ví dụ: Sử dụng ví dụ doanh số - số năm kinh nghiệm, ta biến đổi số liệu về phần dư của hồi qui gốc được kết quả sau:

Bảng 11.5

e	X	$\ln(e_i^2)$	$\ln(X_i)$
161,4410	3	10,1683	1,0986
19,6208	5	5,9532	1,6094
-3,6490	2	2,5889	0,6931
65,8906	8	8,3760	2,0794
-88,6490	2	8,9694	0,6931
-35,2893	6	7,1272	1,7918
-179,1993	7	10,3770	1,9459
12,2611	1	5,0129	0,0000
-63,4691	4	8,3011	1,3863
-6,6490	2	3,7889	0,6931
29,9805	9	6,8011	2,1972
87,7107	6	8,9481	1,7918

Kết quả ước lượng mô hình $\ln(e_i^2) = a_1 + a_2 \ln(X_i)$

Coefficients	Standard		t Stat	P-value
	Error			
Intercept	4.8671	1.4665	3.3189	0.0078
$\ln(X_i)$	1.7526	0.9865	1.7766	0.1060

Kiểm định ý nghĩa của hệ số độ dốc của mô hình $\ln(e_i^2) = 4,8671 + 1,7526 * \ln(X_i)$ bằng phương pháp sử dụng p-value cho thấy về mặt tổng thể hệ số này không khác 0 ($p\text{-value} = 0,1 > 0,05$) nên với độ tin cậy 5% ta có thể kết luận là $\ln(X_i)$ không có liên hệ với $\ln(e_i^2)$, như vậy hiện tượng phương sai thay đổi không xảy ra trong ví dụ của chúng ta.

Nếu xảy ra hiện tượng phương sai thay đổi, một trong những biện pháp để khắc phục hiện tượng phương sai thay đổi là:

- Biến đổi biến phụ thuộc thành Y/X_i và biến độc lập thành $1/X_i$, rồi áp dụng phương pháp bình phương nhỏ nhất hồi qui Y/X_i theo $1/X_i$.
- Lấy log cơ số e cả biến phụ thuộc và biến độc lập rồi hồi qui lại mô hình theo các biến đã được lấy log này, việc ước lượng hồi qui theo các biến đã logarit hóa có thể làm giảm phương sai thay đổi do tác động của phép biến đổi Logarit. Phép biến đổi này không thể thực hiện được nếu một số giá trị của X hoặc Y là âm. Chú ý rằng với mô hình hồi qui mà cả hai đều được lấy log thì hệ số hồi qui lúc này do độ co giãn của Y đối với biến độc lập X chứ không còn là hệ số góc như ở mô hình hồi qui thông thường.
- Một phương pháp nữa để khắc phục phương sai thay đổi là lấy căn bậc hai cho biến độc lập.

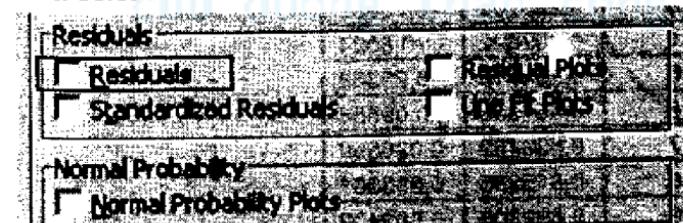
Tuy nhiên chú ý rằng các cách thức biến đổi trên là tùy từng tình huống cụ thể, mà hiệu quả của chúng phụ thuộc vào bản chất của vấn đề. Với mô hình hồi qui bội thì việc chọn biến nào để biến đổi cần phải có những xem xét cẩn thận.

11.2.9.3 Kiểm tra giả định phân phối bình thường của phần dư

Mức độ đáp ứng giả định phân phối Bình thường của phần dư quanh đường hồi qui có thể được đánh giá cũng bằng cách phân tích phần dư. Ở nội dung này ta sử dụng đồ thị xác suất bình thường (Normal probability plot) để xem phần dư có phân phối bình thường hay xấp xỉ bình thường không. Tiến trình vẽ đồ thị này đã được hướng dẫn tại Chương 5, khi Ứng dụng cho đánh giá phần dư của hồi qui chúng ta xem các giá trị phần dư như một tập dữ liệu bất kỳ cần được xác định xem có phân phối bình thường hoặc xấp xỉ bình thường hay không.

Để lưu lại được các giá trị phần dư mà không cần phải tính toán thủ công thì trên cửa sổ Regression chúng ta chọn mục Residuals trong phần Residuals. Xem hình 11.13.

Hình 11.13



Sau đó trong các kết quả chúng ta nhận được sẽ có dãy giá trị phần dư được trình bày trong bảng Residual output

Bảng 11.6 RESIDUAL OUTPUT

Thứ tự của quan sát trong bộ dữ liệu gốc	Giá trị Y được ước lượng từ đường hồi qui	Phần dư
Observation	Predicted Y	Residuals
1	325,559	161,441
2	425,3792	19,6208
3	275,649	-3,64897
4	575,1094	65,89057
5	275,649	-88,649
6	475,2893	-35,2893
7	525,1993	-179,199
8	225,7389	12,26111
9	375,4691	-63,4691
10	275,649	-6,64897
11	625,0195	29,9805
12	475,2893	87,71073

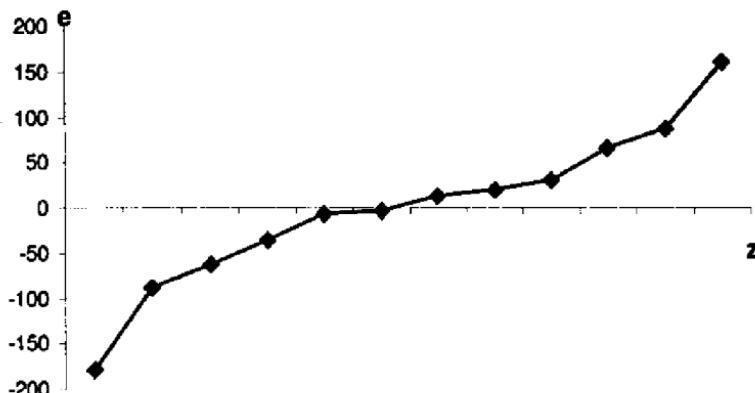
Từ số liệu về phần dư ở bảng này ta thực hiện các bước công việc để xây dựng đồ thị xác suất bình thường nhằm kiểm tra giả định về phân phối bình thường của phần dư, bao gồm:

1. Sắp lại trật tự cho phần dư theo sự tăng dần của giá trị.
2. Đánh số thứ tự của phần dư trong tập dữ liệu đã sắp thứ tự.
3. Xác định diện tích dưới đường cong của phân phối z mà phần dư ở vị trí i tạo ra theo công thức $i/(n+1)$.
4. Tra bảng tích phân Laplace để tìm giá trị chuẩn hóa z tương đương với diện tích tính được.
5. Bước cuối cùng là vẽ đồ thị phân tán mà các giá trị phần dư e được biểu diễn ở trục đứng và giá trị chuẩn hóa z ở trục ngang, ta có đồ thị sau.

Bảng 11.7

e	Thứ tự	$i/(n+1)$	Giá trị z
-179,199	1	0,076923	-1,42608
-88,649	2	0,153846	-1,02008
-63,4691	3	0,230769	-0,73632
-35,2893	4	0,307692	-0,5024
-6,64897	5	0,384615	-0,29338
-3,64897	6	0,461538	-0,09656
12,26111	7	0,538462	0,096559
19,6208	8	0,615385	0,293381
29,9805	9	0,692308	0,502402
65,89057	10	0,769231	0,736316
87,71073	11	0,846154	1,020076
161,441	12	0,923077	1,426077

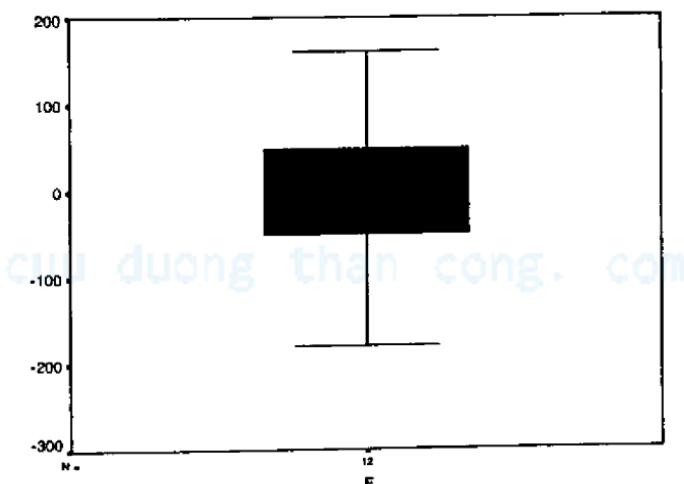
Hình 11.14



Các chấm phân tán trên đồ thị chưa cho ta cảm nhận rõ ràng là phần dư có phân phối bình thường vì đường nối các chấm phân tán không tuân theo xu hướng tuyến tính một cách rõ ràng theo hướng từ gốc dưới bên trái đến gốc trên bên phải. Tuy nhiên với một tập dữ liệu chỉ có 12 quan sát thì chúng ta không thể kỳ vọng một thể hiện hoàn hảo hơn.

Để kết luận lại về phân phối xấp xỉ bình thường của phần dư chúng ta có thể thực hiện cách đơn giản hơn là vẽ đồ thị hộp và râu cho tập dữ liệu về phần dư. Đồ thị hộp và râu lại cho thấy một phân phối khá đối xứng của dữ liệu phần dư trong ví dụ hồi qui doanh số theo số năm kinh nghiệm vì hình dáng và tỷ lệ của biểu đồ khá cân đối.

Hình 11.15

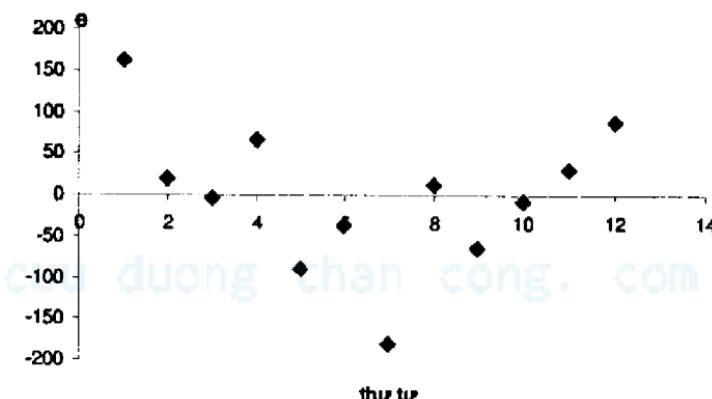


11.2.9.4 Kiểm tra tính độc lập của phần dư

Giả định về tính độc lập của phần dư có thể được kiểm tra bằng cách vẽ đồ thị phần dư theo trật tự của các giá trị mà chúng ta thu thập được theo thời gian. Với dữ liệu thu thập được theo thời gian thường hay xuất hiện một hiện tượng mà người ta gọi tên là tương quan chuỗi giữa các quan sát, đó là mối liên hệ tồn tại giữa các phần dư liên tiếp nhau. Nếu có mối liên hệ như vậy thì ta sẽ thấy một hình dạng cụ thể trên đồ thị vẽ phần dư theo thời gian. Ngoài ra hiện tượng này còn có thể kiểm tra được qua một đại lượng thống kê tên là Durbin-Watson.

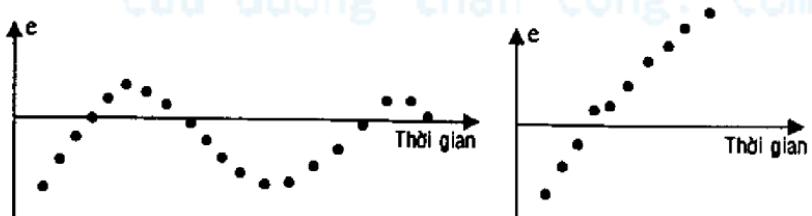
Kiểm tra tính độc lập bằng đồ thị: vẽ đồ thị phần dư theo trật tự dữ liệu như sau (số liệu của Bảng 11.6)

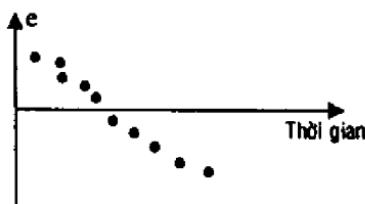
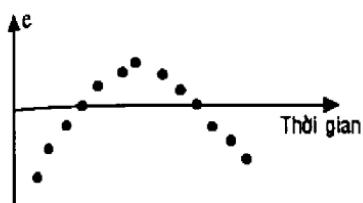
Hình 11.16



Đồ thị không cho thấy một mẫu hình nào của mối liên hệ của các phần dư nên giả định về sự độc lập của các phần dư đã không bị vi phạm. Điều này cũng dễ hiểu vì dữ liệu doanh số và số năm kinh nghiệm của các đại diện bán hàng được thu thập vào cùng một thời điểm tức nó không là dữ liệu chuỗi thời gian nên khó vi phạm giả định này.

Các bạn tham khảo các dạng liên hệ tự tương quan có thể xảy ra trong các hình sau đây:





Cũng giống như tình huống của phương sai thay đổi, ngoài phương pháp đồ thị bạn có thể dùng một kiểm định chính thức để tìm kiếm tương quan chuỗi có tên là Kiểm định Durbin-Watson. Số thống kê sử dụng cho kiểm định này có ký hiệu là D với công thức thiết lập như sau:

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} \quad (11.15)$$

Trong đó

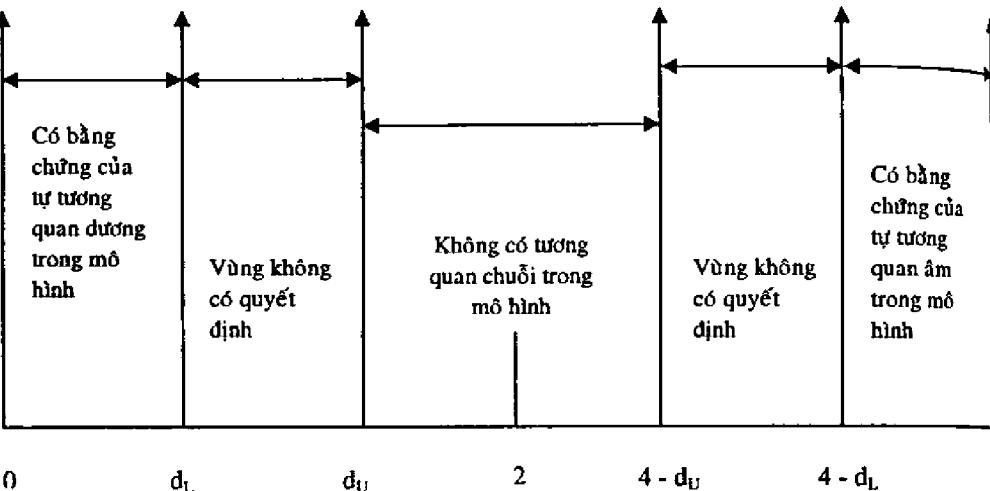
e_i : phần dư tại quan sát i

n số quan sát

Giá trị $0 \leq D \leq 4$

Nhìn vào công thức trên chúng ta thấy rằng nếu những giá trị liên tục của phần dư gần xấp xỉ nhau thì giá trị của D sẽ nhỏ. Tình huống ấy cho thấy các phần dư có khả năng tương quan dương. Trường hợp ngược lại nếu có sự khác biệt lớn trong giá trị của các phần dư thì giá trị D sẽ rất lớn ngụ ý tồn tại một tương quan âm.

Trong Bảng tra 8 Durbin Watson trong phần phụ lục, chúng ta thấy các giá trị d_L và d_U tương ứng với các tình huống cỡ mẫu và số biến độc lập cụ thể. Sau khi tính toán được giá trị D chúng ta so sánh với giá trị ($d_L; d_U$) tra được từ bảng và quyết định về hiện tượng tương quan chuỗi trong mô hình theo qui tắc ở dưới đây với độ tin cậy tương ứng:



Trong thực tế khi tiến hành kiểm định Durbin Watson người ta có thể áp dụng một qui tắc kiểm định đơn giản như sau:

Nếu $1 < D < 3$ thì kết luận mô hình không có tự tương quan

Nếu $0 < D < 1$ thì kết luận mô hình có tự tương quan dương

Nếu $3 < D < 4$ thì kết luận mô hình có tự tương quan âm

Các tình huống đặc biệt:

Giá trị gần đúng của D	Tình huống
$D = 4$	Tương quan hoàn hảo, âm
$D = 0$	Tương quan hoàn hảo, dương
$D = 2$	Không có tự tương quan

Chú ý rằng kiểm định Durbin Watson không đáng tin cậy khi cỡ mẫu nhỏ hơn 15, với ví dụ của chúng ta chỉ có 12 quan sát nên chúng ta không tiến hành kiểm định này, mà sẽ thực hiện khi gặp các ví dụ phù hợp ở những nội dung sau. Tuy nhiên bạn đọc cũng biết rằng tương quan chuỗi là một hiện tượng hay gặp trong dữ liệu thu thập theo chuỗi thời gian liên tục, trong đó quán tính của các hiện tượng hay tồn tại, thực tế này cộng với nhận định trên đồ thị cho kết luận ví dụ của chúng ta không bị hiện tượng phuơng sai thay đổi.

11.2.10 Sử dụng phân tích hồi qui dự đoán giá trị trung bình và giá trị cá biệt của biến phụ thuộc Y

Ta đã biết phương trình hồi qui tổng thể có dạng $E(Y|X_i) = b_0 + b_1 X_i$, và nó được ước lượng bằng mô hình hồi qui mẫu $\hat{Y}_i = b_0 + b_1 X_i$

Từng cặp đối ứng của hai phương trình trên cho thấy đường hồi qui mẫu $\hat{Y}_i = b_0 + b_1 X_i$ cho ta một ước lượng điểm của giá trị trung bình có điều kiện của Y, tức $E(Y/X_i)$. Như vậy muốn dự báo giá trị trung bình có điều kiện của Y với điều kiện $X = X_0$, tức là $E(Y/X_0)$, ta chỉ việc thế giá trị X_0 vào phương trình đường hồi qui mẫu để tìm ra giá trị dự báo \hat{Y}_0 đó là giá trị ước lượng điểm cần tìm. Đồng thời đường hồi qui mẫu này cũng cho ta một giá trị dự báo cá biệt là Y_0 tại điểm $X = X_0$, theo cùng cách là thế X_0 vào phương trình hồi qui mẫu. Dĩ nhiên ta được kết quả giống nhau về số liệu cho cả dự báo điểm giá trị cá biệt và dự báo điểm giá trị trung bình của biến phụ thuộc với cùng một điều kiện về giá trị của biến độc lập, vậy có khác biệt gì giữa hai loại dự báo này không?

Với ví dụ của chúng ta, giả sử giờ đây muốn ước đoán xem nếu một đại diện có 5 năm kinh nghiệm làm việc sẽ có doanh số trung bình là bao nhiêu. Với $X_0 = 5$, ta dễ dàng tính được doanh số trung bình của đại diện sẽ vào khoảng $\hat{Y}_0 = 175,99 + 49,91 \times 5 = 425,54$ (triệu đ) nếu số năm kinh nghiệm là 5 năm. Nếu muốn ước đoán một đại diện có 5 năm kinh nghiệm làm việc sẽ có doanh số cụ thể là bao nhiêu ta cũng làm tương tự, kết quả cũng là 425,54 triệu đồng.

Bạn đừng quên những kết quả tính toán được trên đây của chúng ta chỉ là một ước lượng điểm của giá trị trung bình tổng thể, hoặc ước lượng điểm cho giá trị cá biệt tổng thể của biến phụ thuộc. Ở phương pháp ước lượng thống kê, chúng ta đã biết khoảng tin cậy được xem như một ước lượng đáng tin hơn ước lượng điểm về giá trị của tham số tổng thể. Do đó chúng ta cần phát triển khoảng tin cậy cho trung bình có điều kiện $E(Y/X_0)$ và khoảng tin cậy cho giá trị cá biệt Y_0 .

Công thức chung của hai khoảng tin cậy nói trên lần lượt được viết như sau:

Khoảng tin cậy cho trung bình có điều kiện $E(Y/X_0)$ (11.16)

$$\hat{Y}_0 - t_{(n-2;\alpha/2)} [s_{Y/X}] \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \leq E(Y/X_0) \leq \hat{Y}_0 + t_{(n-2;\alpha/2)} [s_{Y/X}] \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Khoảng tin cậy cho giá trị cá biệt Y_0

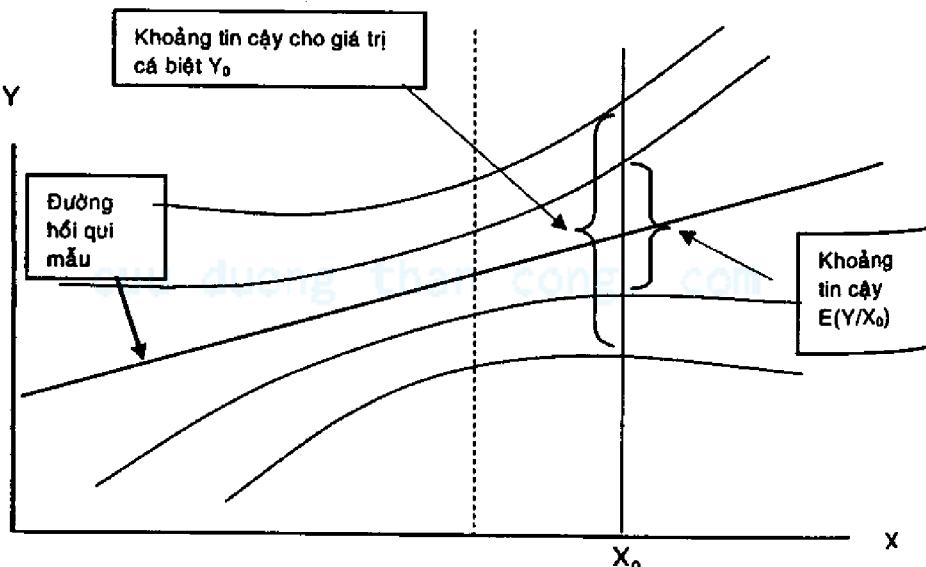
(11.17)

$$\hat{Y}_0 - t_{(n-2,\alpha/2)} [s_{Y/X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}] \leq Y_0 \leq \hat{Y}_0 + t_{(n-2,\alpha/2)} [s_{Y/X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}]$$

Nhìn vào công thức bạn sẽ thấy các thành phần nằm trong ngoặc vuông chính là sai số chuẩn. Giá trị sai số chuẩn trong trường hợp thứ hai lớn hơn trường hợp thứ nhất một lượng $+1$, như vậy nó sẽ tạo khác biệt là khoảng tin cậy cho ước lượng giá trị trung bình sẽ hẹp hơn khoảng tin cậy cho giá trị cá biệt với cùng một độ tin cậy khi tiến hành ước lượng, đây chính là câu trả lời cho câu hỏi có khác biệt gì giữa hai loại dự báo hay không mà chúng ta đặt ra ở trên. Bản chất của việc khoảng tin cậy rộng hơn là thông tin ước lượng kém chính xác hơn, bắt nguồn từ việc sai số chuẩn lớn hơn, tức phương sai lớn hơn. Điều này là do có nhiều biến thiên hơn trong việc dự đoán một giá trị cá biệt so với việc ước lượng một giá trị trung bình tổng thể, điều này rõ ràng rất hợp logic.

Một vấn đề nữa cần chú ý là khoảng tin cậy cho cả hai trường hợp đều sẽ hẹp nhất khi $X_0 = \bar{X}$ và càng lúc càng rộng nếu X_0 đi xa giá trị trung bình của biến độc lập, có nghĩa là việc dự đoán về biến phụ thuộc của chúng ta sẽ kém chính xác đi rất nhiều nếu chúng ta sử dụng những giá trị của biến độc lập quá xa trung bình của nó.

Hình 11.17



Chúng ta sẽ xây dựng khoảng tin cậy 95% cho cả trung bình có điều kiện $E(Y/X_0)$ và giá trị cá biệt Y_0 với $X_0 = 5$ cho ví dụ của chúng ta về doanh số và số năm kinh nghiệm.

Khoảng tin cậy 95% cho trung bình có điều kiện $E(Y/X_0=5)$

Để việc tính toán nhanh chóng, chúng ta vận dụng hàng loạt các kết quả đã tính được ở các nội dung trước

$$\hat{Y}_0 = 425,54; t_{(n-2,\alpha/2)} = t_{(10,0,025)} = 2,228; s_{Y/X} = 92,1057; n = 12; X_0 = 5$$

$$(X_0 - \bar{X})^2 = (5 - \frac{55}{12})^2 = (5 - 4,5833)^2 = 0,1736$$

Nhớ lại rằng ở phần tính toán về s^2 , chúng ta đã tính được lượng ở mẫu số $\sum_{i=1}^n (X_i - \bar{X})^2 = 76,9167$

Từ các số liệu trên ta hình thành công thức như sau cho khoảng tin cậy 95% cho trung bình có điều kiện $E(Y/X_0 = 5)$

$$425,54 - 2,228[92,1057 \sqrt{\frac{1}{12} + \frac{0,1736}{76,9167}}] \leq E(Y/X_0) \leq 425,54 + 2,228[92,1057 \sqrt{\frac{1}{12} + \frac{0,1736}{76,9167}}]$$

$$\rightarrow 425,54 - 60,0363 \leq E(Y/X_0) \leq 425,54 + 60,0363$$

Vậy khoảng tin cậy cần tìm là (365,5037; 485,5763) triệu đồng.

Khoảng tin cậy 95% cho giá trị cá biệt Y_0 khi $X_0 = 5$

$$425,54 - 2,228[92,1057 \sqrt{1 + \frac{1}{12} + \frac{0,1736}{76,9167}}] \leq Y_0 \leq 425,54 + 2,228[92,1057 \sqrt{1 + \frac{1}{12} + \frac{0,1736}{76,9167}}]$$

$$\rightarrow 425,54 - 213,8133 \leq Y_0 \leq 425,54 + 213,8133$$

Vậy khoảng tin cậy cần tìm là (211,7267; 639,3533) triệu đồng.

Kết quả tính toán bằng số liệu thực cũng chứng tỏ tuy với cùng độ tin cậy nhưng khoảng tin cậy cho giá trị cá biệt rộng hơn nhiều (chênh lệch tới hàng trăm triệu đồng) so với khoảng tin cậy cho giá trị trung bình của biến phụ thuộc tại cùng một giá trị của biến độc lập.

11.3 TƯƠNG QUAN TUYẾN TÍNH

Lý thuyết hồi qui chúng ta vừa khảo sát ở trên đã cho chúng ta thấy các biến số X và Y liên hệ tuyến tính với nhau như thế nào. Liên quan mật thiết với phân tích hồi qui nhưng rất khác biệt về khái niệm là lý thuyết phân tích tương quan. Mục tiêu chủ yếu của phân tích tương quan là tính độ mạnh hay mức độ liên hệ tuyến tính giữa hai biến số kể trên.

Ví dụ, ta có thể quan tâm tới việc tìm (hệ số) tương quan giữa giữa số năm đi học và thu nhập; giữa doanh thu và chi phí cho quảng cáo, giữa điểm thi tuyển sinh môn toán vào Đại học và kết quả điểm thi môn Thống kê... Trong phân tích hồi qui, ta không quan tâm chủ yếu tới đại lượng hệ số tương quan, thay vào đó, ta cố gắng ước tính một hàm số toán học (tức là phương trình hồi qui) rồi trên phương trình đó ước lượng giá trị của biến phụ thuộc dựa vào các giá trị xác định của biến độc lập trên cơ sở có mối liên hệ nhân quả giữa các biến. Còn trong phân tích tương quan ta tính toán một chỉ số mà chỉ số này cho chúng ta một hình ảnh về sự biến chuyển cùng với nhau của hai biến số. Trong thuyết tương quan chúng ta không cần để ý đến sự liên hệ nhân quả, chỉ số cho biết sự tương quan giữa X và Y được ước tính bất kể là:

- X có ảnh hưởng lên Y hay ngược lại
- Cả Y và X đều có ảnh hưởng qua lại đến nhau
- Không có biến số nào ảnh hưởng trực tiếp lên biến số kia, nhưng cả hai đều biến chuyển cùng nhau vì một biến số thứ ba có tác động lên cả hai

Tuy tương quan là một kỹ thuật kém sức mạnh hơn hồi qui, nhưng hai vấn đề này có một mối liên hệ toán học rất chặt chẽ và thường thì tương quan được sử dụng như công cụ bổ trợ hữu ích cho hồi qui.

11.3.1 Hệ số tương quan tuyến tính tổng thể ρ

Hệ số tương quan ρ (rho) là một số đo về sự hiệp biến tuyến tính của các biến số, nghĩa là số đo về mức độ kết hợp tuyến tính giữa các biến số. Theo định nghĩa hệ số tương quan tuyến tính ρ được xác định bởi

$$\rho_{xy} = \frac{Cov(X, Y)}{\sqrt{Var(X)} * \sqrt{Var(Y)}} \quad (11.18)$$

Trong đó:

$Cov(X, Y)$ là giá trị đồng phương sai giữa X và Y (ký hiệu σ_{XY}) được xác định như sau $Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E(XY) - \mu_x\mu_y$

Thay vì phương sai chỉ là một đại lượng đo lường mức độ phân tán của một biến ngẫu nhiên xung quanh giá trị trung bình, đại lượng đồng phương sai giữa hai biến ngẫu nhiên sẽ là đại lượng đo lường mức độ liên kết chung giữa chúng. Mặc dù đại lượng đồng phương sai rất có ích trong việc xác định tính chất của mối liên kết giữa X và Y nhưng nó vẫn tồn tại nhược điểm là nhạy cảm với đơn vị tính của X và Y nên người ta tránh vấn đề này bằng cách sử dụng đại lượng đồng phương sai được "chuẩn hóa" chính là hệ số tương quan tuyến tính.

$\text{Var}(X)$ và $\text{Var}(Y)$ không xa lạ gì, chính là phương sai của X và Y được ký hiệu lần lượt là σ_x^2 , σ_y^2 . Công thức hệ số tương quan được viết lại theo ký hiệu như sau

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (11.19)$$

Thường chúng ta không biết được các đại lượng thống kê này vì chúng là tham số tổng thể, do đó chúng ta phải dùng tham số mẫu, vậy σ_{xy} được ước lượng bằng s_{xy} với công thức tính như sau

$$s_{xy} = \frac{1}{n} \sum (X - \bar{X})(Y - \bar{Y}) \quad (11.20)$$

s_x và s_y được ước lượng lần lượt bởi s_x và s_y với công thức tính phương sai của X và Y là

$$s_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \quad (11.21)$$

$$\text{và } s_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n} \quad (11.22)$$

Điều đó cùng có nghĩa là hệ số tương quan tuyến tính tổng thể ρ đã được ước lượng bằng hệ số tương quan tuyến tính mẫu (ký hiệu r).

11.3.2 Hệ số tương quan tuyến tính mẫu r

Theo tiến trình trên, hệ số tương quan mẫu cho bởi công thức sau:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad (11.23)$$

Lần lượt thế công thức của các tham số mẫu ta được

$$r_{xy} = \frac{\frac{1}{(n-1)} \sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\frac{\sum (X - \bar{X})^2}{(n-1)}} \sqrt{\frac{\sum (Y - \bar{Y})^2}{(n-1)}}} \quad (11.24)$$

Giản ước công thức tính r ở trên ta được kết quả cuối cùng sau đây

$$r_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (11.25)$$

Bảng phương pháp biến đổi toán học thông thường chúng ta chuyển được công thức tính r ở trên thành một công thức đơn giản hơn khi tính toán vì có thể vận dụng bảng số liệu đã lập cho quá trình tính các hệ số hồi qui mẫu, đó là

$$r_{xy} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum (X^2) - (\sum X)^2] \times [n \sum (Y^2) - (\sum Y)^2]}} \quad (11.26)$$

Giả sử chúng ta cần tính hệ số tương quan mẫu giữa doanh số và số năm kinh nghiệm của mẫu 12 đại diện, sử dụng lại số liệu đã lập tại Bảng 11.2 ta ráp vào công thức tính r

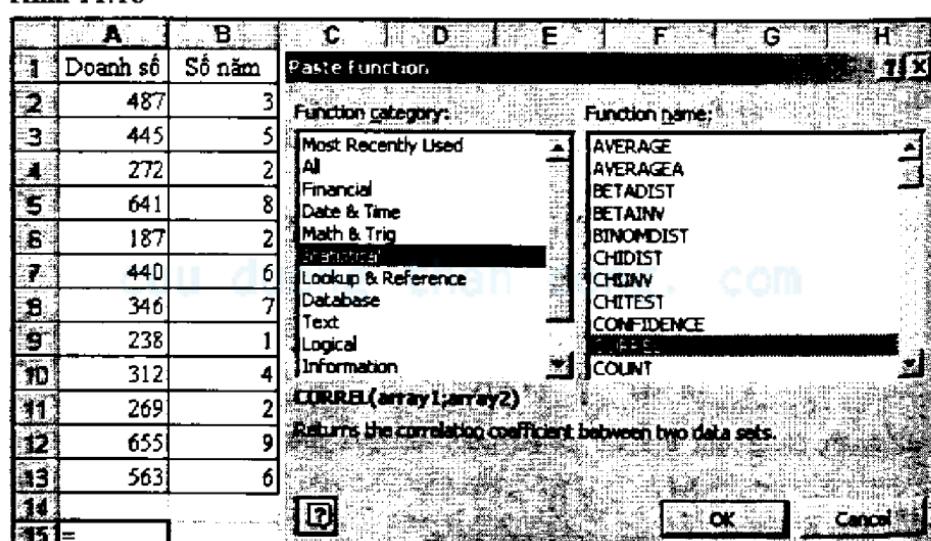
$$r_{xy} = \frac{12 \times 26091 - 55 \times 4855}{\sqrt{[12 \times 329 - 55^2] \times [12 \times 2240687 - 4855^2]}} = \frac{46067}{55333,4721} = 0,83$$

11.3.3 Tính hệ số tương quan tuyến tính bằng Excel

Việc tính hệ số tương quan r có thể tiến hành nhanh chóng nhờ sử dụng lệnh Correl thuộc nhóm hàm Statistical trong lệnh Function của menu Insert trên phần mềm Excel như sau:

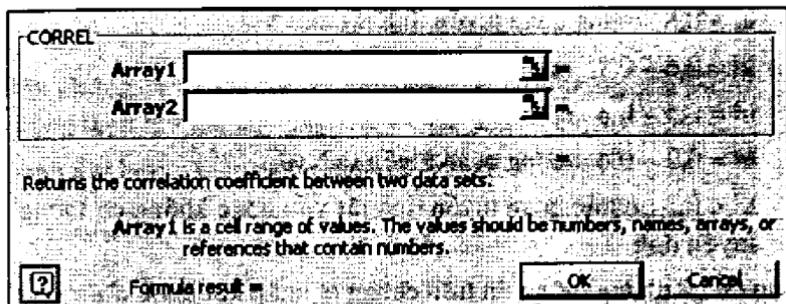
- Vào menu Insert chọn lệnh Function để mở cửa sổ Paste Function
- Trong cửa sổ này bạn chọn sang nhóm hàm Statistical trong khu vực Function Category bên trái, sau đó chọn tiếp lệnh Correl trong khung Function name bên tay phải (xem hình dưới)

Hình 11.18



- Sau các lựa chọn trên, nếu bạn nhấp nút OK thì Excel mở tiếp cửa sổ sau để bạn nhập tiếp dãy chứa dữ liệu về Doanh số vào Array1 và Số năm vào Array2, bạn hoàn toàn có thể làm ngược lại là nhập Doanh số vào Array2 và Số năm vào Array1, kết quả là như nhau

Hình 11.19



- Bạn tiến hành quét dữ liệu theo cách quen thuộc mà chúng ta vẫn làm, và bạn sẽ thấy ngay đáp số thậm chí chưa cần nhấp nút OK nếu bạn nhập đủ dữ liệu, như trong Hình 11.20 Đáp số tính được là 0,832534056; không khác kết quả chúng ta đã tính thủ công là 0,83.

Hình 11.20

A	B	C	D	E	F	G	H	I
Doanh số	Số năm							
487	3							
445	5							
272	2							
641	8							
187	2							
440	6							
346	7							
238	1							
312	4							
269	2							
655	9							
563	6							

=CORREL(A2:A13;B2:B13)

Hệ số tương quan tuyến tính r có thể biến thiên từ -1 (thể hiện một mối liên hệ tuyến tính nghịch hoàn hảo) đến +1 (thể hiện một mối liên hệ tuyến tính thuận hoàn hảo). Dễ hình dung về mối liên hệ tuyến tính hoàn

hảo giữa X và Y nếu bạn tưởng tượng tất cả các chấm phân tán trên đồ thị đều nằm trên một đường thẳng. Dấu của hệ số tương quan chỉ ra hướng của mối liên hệ là thuận hay nghịch, nếu r âm là mối liên hệ nghịch và ngược lại. Nếu giữa X và Y không có liên hệ tuyến tính thì r bằng 0; nếu r càng xa giá trị 0 thì mối liên hệ tuyến tính giữa X và Y càng mạnh dần, theo một quy tắc thực nghiệm như sau:

$|r| > 0,8$: tương quan tuyến tính rất mạnh

$|r| = 0,6 - 0,8$: tương quan tuyến tính mạnh

$|r| = 0,4 - 0,6$: có tương quan tuyến tính

$|r| = 0,2 - 0,4$: tương quan tuyến tính yếu

$|r| < 0,2$: tương quan tuyến tính rất yếu hoặc không có tương quan tuyến tính

Kết hợp tất cả những thông tin trên chúng ta nhận xét được rằng giá trị $r = 0,83$ cho thấy có một mối liên hệ tương quan tuyến tính thuận chiều rất mạnh giữa doanh số bán được và số năm kinh nghiệm làm việc của người đại diện bán hàng. Như vậy là xuyên suốt từ đầu chương đến đây, thông qua đồ thị phân tán và qua phân tích hồi qui chúng ta đã khẳng định được giữa hai đại lượng này có mối liên hệ tuyến tính thuận chiều, tại nội dung này, hệ số tương quan lại khẳng định lại điều này một lần nữa và đồng thời cho chúng ta biết độ mạnh của mối liên hệ đó. Như vậy hẳn bạn đọc nhận thấy ngay rằng dấu của hệ số độ dốc b_1 của mô hình hồi qui tuyến tính mẫu giữa X và Y và dấu của r_{XY} phải giống nhau, vì hệ số độ dốc âm phản ánh một mối liên hệ nghịch chiều giữa hai đại lượng X và Y, và hệ số độ dốc dương phản ánh điều ngược lại. Sự tương tự nhau này được giải thích về mặt toán học như sau: xem công thức chính tắc tính r_{XY} và công thức tính b_1 , mẫu số của hai công thức này luôn là số dương, còn hai tử số thì có cấu trúc giống như nhau nên hai đại lượng này phải cùng dấu, một suy luận khác là khi $b_1 = 0$ thì r cũng bằng 0, mà $b_1 = 0$ có nghĩa là giữa hai đại lượng X và Y chẳng có liên hệ tuyến tính.

11.3.4 Kiểm định ý nghĩa thống kê của hệ số tương quan tuyến tính

Mặc dù ta tính được hệ số tương quan tuyến tính bằng 0,83 nhưng bạn đừng quên đó là hệ số tương quan tuyến tính mẫu căn cứ trên mẫu chỉ gồm 12 quan sát, rất có thể giá trị thực của hệ số tương quan tuyến tính tổng thể (tức là ρ) là một giá trị lớn hoặc nhỏ hơn 0,83 nhiều và thậm chí có thể bằng 0 có nghĩa là giữa doanh số bán được và số năm kinh nghiệm làm việc với công ty của đại diện bán hàng không hề có mối liên hệ. Như vậy kết quả tính được trên mẫu của chúng ta là một kết luận sai lầm, là

hậu quả của sai số lấy mẫu ngẫu nhiên. Như vậy cần phải có một thủ tục kiểm định giả thuyết để kiểm tra ý nghĩa của hệ số tương quan tuyến tính tổng thể giữa doanh số bán được và số năm kinh nghiệm làm việc với công ty của đại diện bán hàng.

Giả thuyết đặt ra cho kiểm định này là

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Giờ đây chúng ta cần kiểm tra xem dữ liệu có ủng hộ giả thuyết không hay không theo thủ tục sau đây

Tính toán величин thống kê t theo công thức

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad (11.27)$$

Trong đó

- t là величина статистики t, kiểm định tuân theo phân phối t với bậc tự do bằng (n-2)
- r là hệ số tương quan mẫu
- n là cỡ mẫu

Với ví dụ của chúng ta kết quả tính được về величина статистики t là

$$t = \frac{0,83}{\sqrt{\frac{1-0,83^2}{12-2}}} = \frac{0,83}{\sqrt{0,03111}} = 4,7057$$

Quy tắc quyết định cũng tương tự như quy tắc kiểm định hai bên bất kỳ, đó là nếu $|t| > t_{(n-2,\alpha/2)}$ thì chúng ta bác bỏ giả thuyết H_0 và ngược lại, với α là mức ý nghĩa đã chọn cho phép kiểm định.

Trong ví dụ này chúng ta chọn mức ý nghĩa bằng 5% và do đó $t_{(12-2,0,05/2)} = 2,228$

Ta thấy $|t| > 2,228$ nên ta quyết định bác bỏ H_0 .

Căn cứ trên thông tin mẫu chúng ta kết luận rằng có một mối liên hệ tương quan tuyến tính thuận chiều có ý nghĩa thống kê giữa doanh số bán và số năm kinh nghiệm của các đại diện bán hàng.

Có một số vấn đề đáng lưu ý về r như sau

- Giá trị của r = 0 cho biết không có mối liên hệ tuyến tính giữa 2 biến chứ chưa có nghĩa là 2 biến đó không có mối liên hệ, vì chúng có thể có liên

hệ phi tuyến. Do đó hệ số tương quan tuyến tính chỉ nên được sử dụng để biểu thị mức độ chất chẽ của liên hệ tương quan tuyến tính.

- Ngoài ra cần phải cẩn thận xem xét đồng thời hệ số tương quan và cả đồ thị phân tán giữa X và Y bởi vì hệ số tương quan có thể có cùng một giá trị trong khi hình dạng của mối liên hệ lại rất khác nhau.
- Một lỗi thông thường khi giải thích hệ số tương quan tuyến tính là cứ cho rằng có liên hệ tương quan có nghĩa là lúc nào cũng có mối liên hệ nhân quả. Kỹ thuật tương quan tuyến tính là một kỹ thuật đối xứng, mối liên hệ giữa X và Y cũng tương tự như liên hệ giữa Y và X chứ nó không phải là liên hệ nhân quả theo một chiều như trong kỹ thuật hồi qui.
- Hệ số tương quan tuyến tính không có đơn vị đo lường.
- Trong mô hình hồi qui tuyến tính đơn biến $\hat{Y}_i = b_o + b_1 X_i$, nếu lấy căn bậc hai của hệ số xác định R^2 bạn sẽ được hệ số tương quan r_{XY} , cụ thể:

$$r = +\sqrt{R^2} \text{ nếu } b_1 > 0$$

$$r = -\sqrt{R^2} \text{ nếu } b_1 < 0$$

Bạn có thể kiểm tra ngay thông tin này qua ví dụ của chúng ta bằng cách lấy căn bậc hai giá trị R^2 tức là $+\sqrt{0,6931} = 0,83 = r$. Điều này hình thành một quy tắc là nếu giữa hai đại lượng X và Y có mối liên hệ tương quan tuyến tính với độ mạnh 0,83 và nếu bạn có thể xác định được mối liên hệ nhân quả giữa chúng (chẳng hạn theo chiều Y phụ thuộc vào X) thì nếu bạn xây mô hình hồi qui tuyến tính đơn $\hat{Y}_i = b_o + b_1 X_i$ chắc chắn mô hình này sẽ giúp bạn giải thích được $0,83^2 = 0,69$ biến thiên của Y dựa trên thông tin về X.

Chúng ta cần phân biệt công dụng của hồi qui và tương quan để hiểu chúng rõ hơn. Cụ thể, hồi qui được sử dụng để:

1. Ước lượng thay đổi của biến phụ thuộc với giá trị xác định của biến độc lập
2. Dự đoán giá trị trung bình và cá biệt của biến phụ thuộc khi biết giá trị của biến độc lập.
3. Kiểm định giả thuyết về tính chất của sự phụ thuộc
4. Kết hợp các vấn đề trên

Còn hệ số tương quan có công dụng đo lường chiều hướng và độ mạnh của mối liên hệ tuyến tính giữa hai biến.

11.4 TƯƠNG QUAN GIỮA CÁC BIẾN ĐỊNH TÍNH

Phân tích tương quan giữa 2 biến đã xem xét trong phần trên áp dụng đối với 2 biến định lượng. Trong phần này chúng ta xem xét các đại lượng dùng để đo lường độ mạnh mối liên hệ giữa 2 yếu tố được đo bằng 2 thang đo danh nghĩa, 1 bảng thang đo danh nghĩa và 1 thang đo khoảng, hoặc cả 2 đều là thang đo khoảng.

Có ba đại lượng đo lường mức độ liên hệ là Spearman r, Kendall tau và gamma, được dùng để tính tương quan giữa hai biến thứ bậc. Khi cả hai biến đang xem xét liên hệ đều có thể xếp hạng được thì chúng ta có thể áp dụng các cách đo lường này dựa trên cơ sở tương quan đơn tuyến tính.

Cách đo lường thứ bậc thích hợp khi mối liên hệ giữa biến X và biến Y, được cho là liên hệ hoặc tăng đơn điệu hoặc giảm đơn điệu. Dĩ nhiên khái niệm về liên hệ tuyến tính thì không thích hợp trong trường hợp thang đo thứ bậc. Tuy nhiên chúng ta có thể áp dụng những mối liên hệ tăng đơn điệu (hoặc giảm đơn điệu). Hàm tăng đơn điệu là hàm hoặc luôn tăng hoặc không đổi khi X tăng. Hay nói cách khác khi X tăng, Y không giảm. Hàm tuyến tính là trường hợp đặc biệt của hàm tăng (hoặc giảm) đơn điệu, ngoại trừ hàm logarit $Y=a+b(\log X)$ thì khác. Chúng ta biết hai loại quan hệ phi tuyến tính, một loại là đơn điệu và loại không đơn điệu. Dĩ nhiên loại không đơn điệu của mối quan hệ phi tuyến tính sẽ có một hay nhiều chỗ cong hoặc đảo chiều, ví dụ như parabol hay phương trình bậc 3.

Chúng ta thường gặp những giả thuyết dưới dạng “Khi X tăng thì Y tăng” Cách nói này ngụ ý rằng mối quan hệ giữa X và Y là đơn điệu, nhưng không nêu rõ hình thức cụ thể. Biến thứ bậc thích hợp với kiểu liên hệ này.

11.4.1 Tương quan hạng Spearman r,

Nguyên lý của đo lường Spearman khá đơn giản. Chúng ta so sánh việc sắp xếp hạng 2 cặp dữ liệu bằng cách tính chênh lệch của các hạng, tính tổng bình phương các chênh lệch này, các hạng là hoàn toàn thuận khi kết quả tính toán hệ số tương quan hạng nhận giá trị là +1, các hạng là hoàn toàn nghịch khi kết quả tính toán hệ số tương quan hạng nhận giá trị -1 và các hạng là không có liên hệ nếu kết quả tính toán là bằng không.

Nếu đặt D_i là chênh lệch giữa từng cặp hạng, sau đó ta tính giá trị $\sum_{i=1}^N D_i^2$ và tính hệ số tương quan hạng bằng công thức:

$$r_s = 1 - \frac{6 \sum_{i=1}^N D_i^2}{N(N^2 - 1)} \quad (11.28)$$

Công thức tính hệ số tương quan hạng này xuất phát công thức tương quan đơn tuyến tính, nhưng được áp dụng trên dữ liệu về hạng thay vì trên dữ liệu gốc, và vì thế, hệ số tương quan hạng Spearman cũng được được giải thích như hệ số tương quan đơn tuyến tính.

Chúng ta xem một ví dụ minh họa sau đây. Các thành viên của một nhóm tham gia cắm trại để huấn luyện tinh thần đồng đội, được xếp hạng từ cao xuống thấp theo 2 biến là mức độ được yêu mến, và mức độ tham gia vào hoạt động nhóm. Hạng của cả hai biến được xếp theo quy ước 1 là có điểm số cao. Trường hợp các mức độ bằng nhau sẽ có hạng bằng nhau được tính bằng cách lấy trung bình cộng của các hạng kế tiếp cho các mức độ này. Giá trị của D_i được tính toán như mô tả trong Bảng 11.8. Nếu số lượng các hạng bằng nhau ít, chúng ta giữ nguyên công thức tương quan hạng. Nếu số lượng các hạng bằng nhau nhiều, cần tính thêm 1 hệ số hiệu chỉnh.

Bảng 11.8 Tính hệ số tương quan hạng của Spearman's

Tên	Hạng của mức độ được yêu mến	Hạng của mức độ tham gia	D_i	D_i^2
Lan	1	5,5	4,5	20,25
Dũng	2,5	5,5	3,0	9
Tuấn	2,5	1	-1,5	2,25
Thùy	4	2	-2,0	4
Hùng	5	3	-2,0	4
Hân	6	9,5	3,5	12,25
Lân	7	5,5	-1,5	2,25
Bích	8	13,5	5,5	30,25
Hoa	9	9,5	0,5	0,25
Yến	10	16	6,0	36
Anh	11,5	5,5	-6,0	36
Bảo	11,5	11,5	0,0	0
Thúy	13,5	8	-5,5	30,25
Sanh	13,5	15	1,5	2,25
Tuệ	15	11,5	-3,5	12,25
Thường	16	13,5	-2,5	6,25
Total			0,0	207,50

Chúng ta tính:

$$r_s = 1 - \frac{6(207,50)}{16(255)} = 1 - 0,305 = 0,695$$

Cần lưu ý rằng nếu việc xếp hạng hoàn toàn thuận, $\sum_{i=1}^N D_i^2$ sẽ nhận giá trị

0 do đó giá trị của hệ số tương quan hạng là 1, trong trường hợp 2 biến liên hệ nghịch hoàn toàn, giá trị của thành phần thứ hai của công thức sẽ là -2.0 và do đó hệ số tương quan hạng r_s sẽ là -1.0.

Trong trường hợp không có liên hệ, thành phần thứ hai sẽ bằng 1 và hệ số tương quan hạng sẽ bằng 0. Nếu $N > 10$ phân phối mẫu của tương quan hạng xấp xỉ phân phối bình thường với độ lệch tiêu chuẩn là $1/\sqrt{N-1}$.

Trong ví dụ trên sai số chuẩn sẽ là $1/\sqrt{15}$. Việc kiểm tra giả thuyết không có mối liên hệ trong mẫu, chúng ta có thể tính toán Z như sau:

$$Z = \frac{r_s - 0}{1/\sqrt{N-1}} = 0,695 \sqrt{15} = 2,69$$

Sử dụng Bảng tra 1 trong phần phụ lục, chúng ta sẽ thấy mối liên hệ là có ý nghĩa tại mức 0.01.

11.4.2 Kendall Tau

Khi tính toán hệ số tương quan hạng Spearman chúng ta dùng bình phương chênh lệch giữa các hạng. Kendall Tau dựa vào cách tính toán khác, Tau cũng nằm trong khoảng giữa -1.0 và 1.0. Đầu tiên chúng ta tính giá trị thống kê S bằng cách quan sát tất cả các cặp trường hợp có thể có và chú ý xem các hạng có cùng thứ tự hay không. Ví dụ, giả sử chúng ta có các cặp hạng như sau:

	a	b	c	d
A	1	2	3	4
B	2	3	1	4

Trong ví dụ này chúng ta có 2 yếu tố (biến) nghiên cứu và 4 đơn vị quan sát. Dữ liệu theo yếu tố A gọi là tập A và dữ liệu theo yếu tố B được gọi là tập B. Dữ liệu ở tập A được sắp xếp theo thứ tự tăng dần (hạng tăng dần), chúng ta có thể tính toán S bằng cách xem xét từng hạng của tập B. Chú ý vào giá trị đầu tiên ở dòng B (đơn vị a), Chúng ta thấy rằng hạng ở yếu tố B có trật tự phù hợp với hạng ở yếu tố A cho từng cặp (a,b) và (a,d). Hay nói cách khác, đơn vị a có hạng thấp hơn b và d tính cho cả hai biến A và B. Mặt khác, dữ liệu ở tập B không tuân theo thứ tự này (xét theo thứ bậc của yếu tố A) cho cặp (a,c) khi a có hạng thấp hơn c trong

tập A, và ngược lại c có hạng thấp hơn a trong tập B.

Hãy đánh dấu +1 mỗi lần khi mỗi cặp này được sắp thứ tự cùng một cách (ý nói cặp phù hợp) cho cả A và B và đánh dấu -1 chúng có trật tự ngược lại (ý nói cặp không phù hợp). Giá trị của S có được bằng cách cộng +1 và -1 cho tất cả các cặp. Do đó S bằng số cặp C phù hợp trừ đi số cặp D không phù hợp. Tính toán các cặp có thể có của đơn vị a với các đơn vị khác bao gồm các cặp (a,b), (a,c) và (a,d) là $+1-1+1 = (2-1) = 1$. Để tính cho các cặp còn lại chúng ta xem toàn bộ bảng. Chúng ta thấy rằng phân phối của các cặp (b,c) và (b,d) là (-1+1) hoặc 0. Cuối cùng phân phối của cặp (c,d) là +1. Lưu ý rằng chúng ta có được tổng giá trị của S bằng cách đầu tiên sắp xếp A vào thứ tự thích hợp kế đó tiếp tục xem xét việc xếp hạng ở dòng B, mỗi lần tính toán số hạng đặt vào bên phải là bên thứ tự đúng và trừ đi các thứ tự ngược lại. Vì thế trong ví dụ này chúng ta có:

$$S = C - D = (2 - 1) + (1 - 1) + (1 - 0) = 2$$

Nếu bây giờ chúng ta chia S cho giá trị lớn nhất có thể, đó là $(N - 1) + (N - 2) + \dots + 2 + 1 = N(N-1)/2$, chúng ta được một hệ số, hệ số này có thể biến thiên trong phạm vi từ -1 tới +. Vì thế chúng ta định nghĩa hệ số tau a (theo Kendall), thích hợp khi không có mối liên hệ nào, công thức như sau:

$$\tau_a = \frac{S}{1/2N(N-1)} = \frac{C - D}{1/2N(N-1)} \quad (11.29)$$

Rõ ràng, nếu có sự ngược nhau hoàn toàn giữa hai hệ thống xếp hạng (tức là nếu B được xếp hạng là hoàn toàn ngược lại A, cụ thể là 4,3,2,1), giá trị của S sẽ là $-1/N(N-1)$ và hệ số tương quan hạng r sẽ là -1.0. Nếu hai biến hoàn toàn không có mối liên hệ, các cặp phù hợp và không phù hợp sẽ bù trừ nhau, và τ sẽ bằng không.

Để minh họa trường hợp xếp hạng mối liên hệ chúng ta hãy làm lại ví dụ nhóm cắm trại trên. Chúng ta sắp xếp các cá thể theo hàng ngang và đặt tên cho các ký tự. Do đó việc xếp hạng thành:

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
A	1	2.5	2.5	4	5	6	7	8	9	10	11.5	11.5	13.5	13.5	15	15
B	5.5	5.5	1	2	3	9.5	5.5	13.5	9.5	16	5.5	11.5	8	15	11.5	13.5

Bây giờ chúng ta theo nguyên tắc là bất cứ cặp nào có hạng bằng nhau ở cả yếu tố A hoặc B, thì giá trị đóng góp vào tổng S sẽ là 0. Đầu tiên ta quan sát tất cả các cặp đơn vị được lập thành với đơn vị (người) a, ta thấy

rằng các cặp (a,b), (a,g), và (a,k) sẽ không có đóng góp giá trị gì cho S vì hạng theo yếu tố B của tất cả các đơn vị này đều bằng 5.5, Do đó giá trị đóng góp của các cặp còn lại sẽ là:

a,c	a,d	a,e	a,f	a,h	a,i	a,j	a,l	a,m	a,n
-1	-1	-1	+1	+1	+1	+1	+1	+1	+1

$$= 9 - 3 = 6$$

Chúng ta tiếp tục so sánh hạng của đơn vị b lần lượt với hạng của các đơn vị bên phải của b. Để ý rằng, b và c có hạng bằng nhau theo yếu tố A. Vì b và c có hạng bằng nhau theo yếu tố A nên chúng ta phải loại trừ cặp (b,c). Tương tự như vậy, các cặp (b,g) và (b,k) có hạng bằng nhau theo yếu tố B và không có đóng góp cho tổng S. Vì vậy để tính các cặp bắt đầu bằng b, chúng ta tính tổng 9 - 2 hay 7. Chúng ta tiếp tục tính:

$$S = C - D = (9 - 3) + (9 - 2) + (13 - 0) + (12 - 0) + (11 - 0) + (6 - 3) + (8 - 0) + (2 - 5) + (5 - 2) + (0 - 6) + (4 - 0) + (2 - 1) + (2 - 0) + (0 - 2) + (1 - 0) = 60$$

Bây giờ chúng ta điều chỉnh mẫu số của τ_{α} để hiệu chỉnh các cặp ngang hàng. Sự điều chỉnh này có tác động làm tăng giá trị số liệu của τ_{α} , mặc dù sự tăng này không đáng kể trừ khi số các cặp ngang hàng nhiều.

Công thức tau b được khái quát như sau:

$$\tau_b = \frac{S}{\sqrt{1/2N(N-1) - T\sqrt{1/2N(N-1) - U}}} \quad (11.30)$$

Trong đó: $T = \frac{1}{2} \sum t_i(t_i - 1)$, t_i là số các cặp ngang hạng trong A, và $U = \frac{1}{2} \sum u_i(u_i - 1)$, u_i là số các cặp ngang hạng trong B

Trong ví dụ trên, chúng ta có ba cặp ngang hàng của mỗi cặp của biến A. Vì thế

$$T = \frac{1}{2}[2(1)+2(1)+2(1)] = 3$$

Tương tự như vậy, có ba cặp ngang hạng và một đơn vị ngang hạng với bốn đơn vị khác của biến B. Do đó

$$U = \frac{1}{2}[2(1)+2(1)+2(1)+4(3)] = 9$$

$$\text{Và } \tau_b = \frac{60}{\sqrt{[8(15)-3][8(15)-9]}} = \frac{60}{\sqrt{(117)(111)}} = \frac{60}{114,0} = 0,526$$

Kiểm tra mức ý nghĩa của tau

Kendall đã chỉ ra rằng các cỡ mẫu từ 10 trở lên, phân phối S của mẫu sẽ xấp xỉ bình thường với trung bình bằng không và phương sai được tính

bảng công thức:

$$s_s^2 = 1/18N(N-1)(2N+5) \quad (11.31)$$

Nói một cách chính xác, công thức trên chỉ đúng khi không có các cặp ngang hạng, nhưng cũng có thể sử dụng an toàn khi số lượng các cặp ngang hạng ít. Nếu số lượng các cặp ngang hạng rất nhiều thì cần thêm vào nhân tố hiệu chỉnh.

Để kiểm tra mức ý nghĩa của τ_{α} cho các dữ liệu nhóm cắm trại trên, đầu tiên chúng ta tính phương sai s_s^2 như sau:

$$s_s^2 = 1/18(16)(15)(37) = 493,3$$

Lấy căn bậc hai chúng ta được: $s_s = 22,21$

Độ lệch chuẩn trên được dùng làm mẫu số của Z trong kiểm tra giả thuyết H_0 rằng A và B không có mối liên hệ. Do đó

$$Z = \frac{S - 0}{\sigma_s} = \frac{60,0}{22,21} = 2,70$$

Và chúng ta có thể thấy rằng giá trị của τ_{α} là 0,526 là có ý nghĩa tại mức 0,01.

11.4.3 Tương quan đối với dữ liệu thứ bậc trong dữ liệu đã phân nhóm (τ_{α} c, gamma, dyx và dxy)

Một cải tiến của τ_{α} so với hệ số tương quan hạng r, là ở chỗ τ_{α} được sử dụng khi số lượng cặp ngang hạng nhiều. Mặc dù việc tính toán theo thủ tục được mô tả ở trên rất nhảm chán như ngay trong ví dụ trên, chúng ta có thể đơn giản hóa thủ tục tính toán này khi cả hai biến nghiên cứu có thể nhóm lại vào một số nhóm thô (lớn hơn). Ví dụ, những người ở năm tầng lớp xã hội, những người ở cùng tầng lớp có địa vị xã hội ngang nhau. Nếu biến thứ hai được phân chia theo cách nhóm lại tương tự này, chúng ta có thể chỉnh sửa công thức τ_{α} và do đó tận dụng thông tin của dữ liệu xếp hạng tốt hơn thay vì chỉ phân loại giản đơn.

Chúng ta có thể tính $S = C - D$ (tổng các cặp thuận hạng trừ tổng các cặp nghịch hạng) bằng thủ tục được mô tả dưới đây. Sử dụng công thức cho trên, chúng ta sẽ nhận thấy rằng giới hạn trên của τ_{α} sẽ là 1 chỉ khi số dòng bằng với số cột. Để hiệu chỉnh khả năng là $r \neq c$ chúng ta tính tỉ lệ:

$$\tau_c = \frac{S}{1/2N^2[(m-1)/m]} \quad (11.32)$$

Trong đó $m = \text{Min}(r, c)$

Chúng ta lại theo Kendall sử dụng ký hiệu τ_c để phân biệt công thức (11.32) từ hai công thức trên. Vậy giờ chúng ta hãy xem xét hệ số tương quan τ_c được tính như thế nào.

Bảng 18.4 Bảng phân tách để tính Kendall tau từ dữ liệu đã được phân nhóm.

Mức độ mong muốn làm việc trong các công ty nước ngoài (A)	Mối quan tâm đến khả năng lực bản thân (B)				Tổng cộng
	Cao	Cao vừa phải	Thấp vừa phải	Thấp	
Cao	18	19	12	8	57
Cao vừa phải	16	16	12	10	54
Thấp vừa phải	11	14	18	16	59
Thấp	5	5	15	22	47
Tổng cộng	50	54	57	56	217

Dữ liệu trong bảng 18.4 mô tả việc xếp hạng của 217 sinh viên tại đại học Kinh Tế. Biến B thể hiện mối quan tâm muốn thể hiện năng lực bản thân. Biến A thể hiện mong muốn làm việc trong các công ty nước ngoài. Vì việc đo lường hai biến này còn khá thô sơ, mỗi biến được phân làm bốn nhóm thứ bậc: cao, cao vừa phải, thấp vừa phải, thấp. Kết quả được tóm tắt trong bảng phân tách.

Khi tính toán S, chúng ta sẽ tính toán trước cho C và D. Đầu tiên chúng ta lưu ý rằng dữ liệu về A được xếp hạng lại từ cao đến thấp, có 57 đơn vị “ngang hàng” ở mức cao, 54 cao vừa phải, 59 thấp vừa phải và 47 thấp. Đầu tiên nhìn vào những tần số ở mức cao ở biến A, chúng ta thấy có 18 cũng là cao ở biến B, 19 cao vừa phải và tiếp tục 12 thấp vừa phải, 8 thấp. Để tính toán C hoặc D, chúng ta để ý rằng vì tất cả các đơn vị trong loại cao của biến A đều ngang hàng (xét theo biến A), không có cặp nào trong những cặp này có đóng góp cho C hoặc D. Tương tự như vậy không có cặp nào trong cùng một cột sẽ đóng góp C hoặc D bởi vì tất cả các cặp này đều ngang hàng theo B.

Nếu chúng ta nhìn vào 1 ô bất kỳ nào trong bảng, tất cả tần số nằm trong các ô phía dưới bên tay phải sẽ đóng góp vào các cặp phù hợp C. Ngược lại những tần số nằm trong các ô phía dưới nó về phía bên trái sẽ đóng góp vào số cặp không phù hợp D. Ví dụ, mỗi đơn vị của 18 đơn vị ở trong ô trên cùng phía bên trái sẽ tạo ra các cặp phù hợp với bất kỳ

$$16 + 14 + 5 + 12 + 18 + 15 + 10 + 16 + 22$$

đơn vị nằm phía dưới về phía tay phải của ô này. Cộng lại tất cả, đóng

góp của ô này vào số cặp phù hợp C sẽ là:

$$18 \times (16 + 14 + 5 + 12 + 18 + 15 + 10 + 16 + 22) = 18 (128)$$

Kế tiếp chúng ta tập trung vào 16 trường hợp ngay bên dưới ô ở góc trên cùng về phía tay trái. Mỗi đơn vị này cũng có hạng ở B cao. Để đếm số lượng các cặp đóng góp vào số cặp phù hợp C chúng ta cộng cộng các tần số trong các ô ở phía dưới bên tay phải của ô này và được

$$16 (14 + 5 + 18 + 15 + 16 + 22) = 16 (90)$$

Khi di chuyển sang cột thứ nhì kế tiếp, chúng ta bắt đầu, chúng ta tính số trường hợp có hạng phù hợp và không phù hợp để tính C và D. Bởi vì các cột bên tay trái có hạng B cao hơn, vì thế ở ô đầu tiên ở cột thứ hai chúng ta tính số trường hợp C:

$$19(12 + 18 + 15 + 10 + 16 + 22) = 19 (93)$$

Và vì thế số trường hợp D là:

$$19(16 + 11 + 5) = 19 (32)$$

Tính tương tự cho hết bảng, chúng ta tính được S khá đơn giản như sau:

$$C = 18(128) + 16(90) + 11(42) + 19(93) + 16(71) + 14(37) + 12(48) + 12(38) + 18(22) = 9055$$

$$D = 19(32) + 16(16) + 14(5) + 12(67) + 12(35) + 18(10) + 8(112) + 10(68) + 16(25) = 4314$$

$$\text{Do đó } S = 9055 - 4314 = 4741$$

$$\text{Vì thế } \tau_c = \frac{4741}{1/2(217)^2[(4-1)/4]} = 0,268$$

Lưu ý rằng mẫu số của τ_c chỉ dựa vào số dòng và số cột, không dựa vào các tổng tần số dòng hay cột, dĩ nhiên việc này quyết định số cặp ngang hàng, điều này làm τ_c khó giải thích hơn và về phương diện này ít thỏa đáng hơn τ_b ³. Cũng có một vài thước đo khác về phương diện xử lý các cặp ngang hàng ở mẫu số. Cách đo lường thông dụng nhất là gamma (γ) nó loại trừ các cặp ngang hàng khỏi mẫu số và nó cũng được áp dụng cho dữ liệu không phân nhóm. Công thức gamma là:

$$\gamma = \frac{C - D}{C + D} \quad (11.33)$$

Trong ví dụ này ta có:

$$= \frac{9055 - 4314}{9055 + 4314} = 0,354$$

Vì gamma, τ_a , và τ_b có cùng tử số, và vì mẫu số của gamma loại trừ tất cả

các cặp ngang hạng, ta dễ dàng nhận ra rằng $\gamma > \tau_b > \tau_c$. Nói chung, chỉ khi nào các tổng dòng và cột của A và B rất khác nhau, trị số của gamma mới lớn hơn τ_b đáng kể. Ví dụ như trong trường hợp của bảng giả thuyết sau:

A	B			Tổng cộng
	Cao	Vừa	Thấp	
Cao	100	80	0	180
Vừa	0	20	80	100
Thấp	0	0	20	20
Tổng cộng	100	100	100	300

Chú ý rằng trong trường hợp này không có các cặp nghịch hạng nhau, do đó $\gamma = 1.0$. Tuy nhiên $\tau_b = 0.77$ và $T_c = 0.68$. Mối liên hệ này là hoàn hảo hay không sẽ dựa vào giả định vì sao phân phối các tổng dòng và cột không giống nhau.

Bên cạnh các thước đo tau và gamma, ta có hai cách đo lường đối xứng nhau d_{yx} và d_{xy} được giới thiệu bởi Somers và được định nghĩa như sau:

$$d_{yx} = \frac{C - D}{C + D + T_y} \quad (11.34)$$

và

$$d_{xy} = \frac{C - D}{C + D + T_x} \quad (11.35)$$

Trong đó:

- T_x là số cặp ngang hạng theo biến X nhưng không ngang hạng theo biến Y
- T_y là số cặp ngang hạng theo biến Y nhưng không ngang hạng theo biến X

Nếu chúng ta đặt T_{xy} là số cặp ngang hạng với cả hai biến X và Y, sau đó áp dụng công thức tính T_b ta có $T = T_x + T_{xy}$ và $U = T_y + T_{xy}$ và do đó vì tổng số các cặp $1/2N(N-1) = C + D + T_x + T_y + T_{xy}$, ta có $C + D + T_y = 1/2N(N-1) - (T_x + T_{xy}) = 1/2N(N-1) - T$.

Tương tự như vậy, mẫu số của d_{xy} là $C + D + T_x = 1/2N(N-1) - U$. Vì thế tích $d_{yx} d_{xy} = T_b^2$. Giả sử ta muốn tiên đoán thứ bậc của 1 cặp trường hợp theo biến B. Nếu loại bỏ các trường hợp ngang hạng, xác suất sai lầm khi không biết thêm thông tin gì là sẽ là 0.5. Nếu ta biết thứ bậc của cặp này

theo biến A, thì trị tuyệt đối của gamma bằng với số khả năng sai kỳ vọng khi biết về A, trừ đi số khả năng sai kỳ vọng khi không biết về A, chia cho số khả năng sai kỳ vọng khi không biết về A.

Chúng ta có một số thước đo liên hệ thứ bậc chỉ khác nhau về cách xử lý các cặp ngang hạng ở mẫu số. Tuy nhiên, chúng ta không có quy tắc quyết định rõ ràng cho việc chọn lựa cái nào trong số các thước đo này bởi vì lý do của các cặp ngang hạng vẫn còn chưa rõ. Wilson đã chỉ ra rằng tính chất giảm sai số của gamma sẽ thất bại nếu ta thừa nhận rằng các sai số có thể phạm phải khi tiên đoán một trật tự thứ bậc theo B nếu thực sự cặp này ngang hạng theo B. Có vẻ như vấn đề xử lý các cặp ngang hạng này không có giải pháp đơn giản. Có lẽ kinh nghiệm thông thường là sử dụng biến có càng nhiều phân hạng càng tốt, do đó sẽ giảm số lượng các cặp ngang hạng giữa các biến do lưỡng khac nhau.

cuu duong than cong. com

cuu duong than cong. com

CHƯƠNG 12

HỒI QUI TUYẾN TÍNH ĐA BIẾN

Trong chương trước chúng ta đã tập trung nghiên cứu mô hình trong đó một biến độc lập hay một biến giải thích X được sử dụng để dự đoán giá trị biến phụ thuộc Y. Nhớ lại phương trình hồi qui tuyến tính tổng thể:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Trong đó:

X_i và Y_i là các giá trị của biến độc lập và biến phụ thuộc của cặp quan sát thứ i

β_0 : là hệ số tung độ gốc (hay hệ số chẵn)

β_1 : hệ số độ dốc (hay hệ số góc)

Trên tập dữ liệu mẫu ta xây dựng phương trình hồi qui tuyến tính mẫu đơn biến có công thức như sau để ước lượng mô hình hồi qui tổng thể :

$$\hat{Y}_i = b_0 + b_1 X_i$$

Trong đó:

\hat{Y}_i là giá trị ước lượng cho giá trị của biến Y ở quan sát thứ i.

X_i là giá trị của biến X ở quan sát thứ i.

Các b_k là các hệ số hồi qui mẫu.

Trong chương này, chúng ta mở rộng thành mô hình hồi qui đa biến (hay còn gọi là hồi qui bội) trong đó không phải chỉ một mà nhiều biến giải thích có thể được sử dụng để dự đoán giá trị của biến phụ thuộc. Hãy nghiên cứu ví dụ sau đây:

Giám đốc nhân sự của một tập đoàn kinh doanh cố gắng xác định những năng lực cá nhân nào là cần thiết đối với một nhà quản lý để chuyển từ vị trí quản lý cấp trung lên cấp quản lý cao hơn. Trong một thời gian dài, bà hay nghe được rằng khuyết điểm thường gặp nhất ở các nhân viên quản lý làm tại tập đoàn là kỹ năng giao tiếp, vì thế bà ta quyết định do lưỡng xem các kỹ năng này có phải là các nhân tố quyết định hay không.

Bà giám đốc nhân sự dự định xây dựng một mô hình hồi qui bội giữa Điểm đánh giá kết quả làm việc của nhân viên quản lý và khả năng giao tiếp của họ. Bà ta chọn ra một mẫu ngẫu nhiên các nhân viên quản lý cấp trung đã giữ chức vụ đó trong khoảng trên 1 năm cho tới 5 năm. Ghi nhận điểm đánh giá kết quả làm việc gần nhất của họ. Sau đó các nhà quản lý

này được cung cấp một số tình huống để phân tích và được yêu cầu trình bày các đề xuất cho các tình huống đó bằng cả hai cách là thuyết trình miệng và báo cáo viết. Một ban giám khảo gồm các nhà quản lý cấp cao sẽ đánh giá họ dựa vào các kết quả phân tích, khả năng trình bày miệng và bài viết. Những kết quả đánh giá này được đem so sánh với điểm đánh giá kết quả làm việc gần nhất của các nhân viên quản lý cấp trung để tìm xem chúng có mối liên hệ với nhau không (theo suy luận của bà giám đốc thì nó phải là một mối liên hệ thuận chiều). Xem dữ liệu được tổng hợp trong bảng sau:

Bảng 12.1

Nhân viên quản lý cấp trung	Điểm đánh giá kết quả làm việc Y	Điểm phân tích tình huống X_1	Điểm khả năng trình bày viết X_2	Điểm khả năng trình bày miệng X_3
1	87	8,4	8,7	9,2
2	93	8,2	9,4	9,4
3	91	9,3	9,7	9,5
4	85	7,9	8,1	8,7
5	86	8,1	8,3	8,8
6	97	9,4	9,3	9,6
7	90	9,1	9,0	9,2
8	93	8,9	9,2	9,5
9	88	8,6	8,4	8,5
10	96	9,7	9,5	9,6
11	86	8,3	7,9	8,4
12	89	8,7	8,5	8,7
13	94	9,2	9,1	9,6
14	91	8,1	9,5	9,2
15	95	9,3	9,1	9,7

Với tình huống trên, bà giám đốc nhân sự đang muốn sử dụng ba biến độc lập là: Điểm phân tích tình huống (ký hiệu X_1), Điểm khả năng trình bày viết (ký hiệu X_2), Điểm khả năng trình bày miệng (ký hiệu X_3) để giải thích cho biến phụ thuộc là Điểm đánh giá kết quả làm việc (như thường lệ vẫn ký hiệu là Y), do vậy chúng ta sẽ xây dựng một mô hình hồi qui tuyến tính bội, trong đó 3 biến độc lập X_k được sử dụng để giải thích cho Y. Các thủ tục chúng ta đã nghiên cứu ở chương trước sẽ lần lượt được mở rộng cho mô hình hồi qui bội trong chương này, về cơ bản chúng ta thực hiện các bước sau:

- Sử dụng phần mềm máy tính để tính toán các hệ số hồi qui và các con số thống kê cần thiết sử dụng để đánh giá mô hình
- Đánh giá sự phù hợp của mô hình, có mấy phương pháp thống kê để tiến hành điều này là : tính sai số chuẩn của ước lượng, hệ số xác định, và dùng số thống kê F để kiểm định ý nghĩa toàn diện của mô hình, đánh giá ý nghĩa của từng biến độc lập riêng biệt.
- Nếu chúng ta hài lòng với độ phù hợp của mô hình và những điều kiện cần thiết được thỏa mãn chúng ta có thể diễn dịch ý nghĩa của các hệ số hồi qui
- Sau đó chúng ta cũng có thể sử dụng mô hình để dự đoán hoặc ước lượng giá trị trung bình của biến phụ thuộc
- Chuẩn đoán sự vi phạm các giả định trong mô hình, nếu điều này xảy ra thì tiến hành khắc phục.

12.1 PHƯƠNG TRÌNH HỒI QUI TUYẾN TÍNH TỔNG THỂ ĐA BIẾN VỚI K BIẾN ĐỘC LẬP

12.1.1 Phương trình hồi qui tổng thể

Phương trình hồi qui tổng thể với k biến độc lập có dạng như sau:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

Trong đó:

β_0 : là hệ số tung độ gốc

β_1 : hệ số độ dốc của Y theo biến X_1 giữ các biến $X_2, X_3...X_k$ không đổi

β_2 : hệ số độ dốc của Y theo biến X_2 giữ các biến $X_1, X_3...X_k$ không đổi

β_3 : hệ số độ dốc của Y theo biến X_3 giữ các biến $X_1, X_2...X_k$ không đổi

...
 β_k : hệ số độ dốc của Y theo biến X_k giữ các biến $X_1, X_2, X_3...X_{k-1}$ không đổi
 ε_i là thành phần ngẫu nhiên (yếu tố nhiễu) với tính chất đã biết ở chương trước.

Với ví dụ của chúng ta, chúng ta cần xây dựng mô hình hồi qui bội với 3 biến độc lập để giải thích cho Y với phương trình như sau

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad (12.1)$$

Trong đó:

X_{1i} là Điểm phân tích tình huống của nhân viên thứ i

X_{2i} là Điểm khả năng trình bày viết của nhân viên thứ i

X_{3i} là Điểm khả năng trình bày miệng của nhân viên thứ i

Y_i là Điểm đánh giá kết quả làm việc của nhân viên thứ i

Với $i = (1; 2; 3; \dots; 15)$

12.1.2 Các hệ số hồi qui riêng phần

Trong phương trình (12.1) các hệ số hồi qui tổng thể được diễn đạt cụ thể như sau:

β_0 : là hệ số tung độ gốc

β_1 : hệ số độ dốc của Y theo biến X_1 , giữ các biến X_2, X_3 không đổi

β_2 : hệ số độ dốc của Y theo biến X_2 , giữ các biến X_1, X_3 không đổi

β_3 : hệ số độ dốc của Y theo biến X_3 , giữ các biến X_1, X_2 không đổi

ϵ_i là thành phần ngẫu nhiên hay yếu tố nhiễu

Như vậy điểm cần chú ý là trong mô hình hồi qui đơn biến, hệ số độ dốc β_1 mô tả thay đổi trong giá trị trung bình của Y trên mỗi đơn vị thay đổi của X mà không cần quan tâm đến tác động của biến độc lập nào khác vì đó là mô hình hồi qui tuyến tính đơn. Còn với mô hình hồi qui tuyến tính bội với 3 biến độc lập như trên thì hệ số độ dốc β_1 thể hiện thay đổi trong trị trung bình của Y trên mỗi đơn vị thay đổi của X_1 không kể đến ảnh hưởng của X_2, X_3 vì thế β_1 được gọi tên là hệ số hồi qui riêng phần. Ta cũng lập luận tương tự như thế cho hai hệ số còn lại. Chúng ta sẽ trở lại thảo luận về các hệ số hồi qui riêng phần ở các nội dung Diễn dịch ý nghĩa của hệ số hồi qui.

12.2 PHƯƠNG TRÌNH HỒI QUI TUYẾN TÍNH MẪU ĐA BIẾN VỚI 3 BIẾN ĐỘC LẬP

12.2.1 Viết phương trình hồi qui tuyến tính mẫu 3 biến độc lập

Cũng như qui trình đã nghiên cứu ở Chương 11 khi xây dựng mô hình hồi qui mẫu đơn biến, ở đây chúng ta cũng xây dựng được mô hình hồi qui tuyến tính mẫu 3 biến cho ví dụ của chúng ta, nó có phương trình như sau:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i}$$

Ý nghĩa các ký hiệu của các thành phần trong phương trình trên như sau:

- \hat{Y}_i là giá trị được ước lượng về Điểm đánh giá kết quả làm việc của nhân viên thứ i căn cứ trên mô hình hồi qui
- b_0 là hệ số tung độ gốc ước lượng của β_0 căn cứ trên dữ liệu mẫu
- b_1 là hệ số độ dốc ước lượng của β_1 căn cứ trên dữ liệu mẫu
- b_2 là hệ số độ dốc ước lượng của β_2 căn cứ trên dữ liệu mẫu
- b_3 là hệ số độ dốc ước lượng của β_3 căn cứ trên dữ liệu mẫu

12.2.2 Dùng Microsoft Excel để tính toán các hệ số hồi qui mẫu và các số thống kê khác

Chúng ta hoàn toàn có thể sử dụng lệnh Regression quen thuộc để tính toán các hệ số hồi qui ước lượng từ dữ liệu thực tế ta ghi chép được trên 15 nhân viên quản lý, tiến trình thực hiện cũng tương tự như với quá trình tính toán các hệ số hồi qui ước lượng của mô hình hồi qui đơn, có một điểm cần chú ý đó là với mô hình hồi qui đơn bạn chỉ cần phải quét một cột dữ liệu của một biến độc lập duy nhất, còn với mô hình hồi qui bội (chẳng hạn với ví dụ của chúng ta là 3 biến) bạn có 3 cột chứa dữ liệu, bạn cần sắp xếp ba cột chứa dữ liệu này liên tục nhau, không được có một cột trống hay cột chứa dữ liệu của một biến nào khác nằm xen vào giữa các cột này thì Excel mới xử lý cho bạn kết quả đúng đắn, xem hình 12.1.

Hình 12.1

A	B	C	D	E	F
Nhân viên quản lý cấp trung	Điểm đánh giá kết quả lâm việc	Điểm phản hồi tích thíc h hướng	Điểm khái ng niệm nâng cao	Điểm khái ng niệm tay nghề	
1					
2	TT	X ₁	X ₂	X ₃	
3	1	87	8.4	8.7	9.2
4	2				
5	3				
6	4				
7	5				
8	6				
9	7				
10	8				
11	9				
12	10				
13	11				
14	12				
15	13				
16	14				
17	15				

Regression

Input:

Input Y Range: \$C\$2:\$C\$17

Input X Range: \$C\$2:\$E\$17

Labels

Constant is Zero

Confidence Level: 95 %

Output options:

Output Range: \$A\$18:\$E\$23

New Worksheet By: New Worksheet

New Workbook

Residuals:

Residuals

Residual Plots

OK Cancel Help

Kết quả lệnh Regression cho bạn 3 bảng kết quả như sau đây:

Bảng 12.2

Regression Statistics	
Multiple R	0.905486545
R Square	0.819905883
Adjusted R Square	0.770789305
Standard Error	1.849512571
Observations	15

 R^2 $S_{Y/X}$

SSR

Bảng 12.3

ANOVA

	df	SS	MS	F	Significance F
Regression	3	171.3056691	57.10189	16.69306	0.000208868
Residual	11	37.62766427	3.420697		
Total	14	208.9333333			

SSE

S_{bj}

SST

t_{tt}**Bảng 12.4** (đã được làm tròn 4 số lẻ sau dấu phẩy)

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	21.4805	10.5312	2.0397	0.0661	-1.6984	44.6594
X1	2.3640	1.1839	1.9967	0.0712	-0.2418	4.9698
X2	1.5313	1.7735	0.8634	0.4063	-2.3722	5.4349
X3	3.8074	2.4930	1.5272	0.1549	-1.6797	9.2945

Từ Bảng 12.4 chúng ta thấy các hệ số hồi qui mẫu b_i có giá trị lần lượt là:

$$b_0 = 21,4805$$

$$b_1 = 2,3640$$

$$b_2 = 1,5313$$

$$b_3 = 3,8074$$

Mô hình hồi qui mẫu được viết lại như sau:

$$\hat{Y}_i = 21,4805 + 2,3640X_{1i} + 1,5313X_{2i} + 3,8074X_{3i}$$

Trong đó xin nhắc lại

- \hat{Y}_i là giá trị được ước lượng về Điểm đánh giá kết quả làm việc của nhân viên thứ i căn cứ trên mô hình hồi qui
- X_{1i} là Điểm phân tích tình huống của nhân viên thứ i
- X_{2i} là Điểm khả năng trình bày viết của nhân viên thứ i
- X_{3i} là Điểm khả năng trình bày miệng của nhân viên thứ i

12.2.3 Đọc các con số thống kê cần thiết trên bảng kết quả

Cũng với ý nghĩa như ta đã tìm hiểu ở nội dung hồi qui đơn, chúng ta có các con số thống kê được tính toán sẵn trên các bảng số liệu như sau:

Trên Bảng 12.2

- Hệ số xác định bội $R^2 = 0,8199$
- Hệ số xác định hiệu chỉnh $R^2_{adj} = 0,7708$
- Sai số chuẩn của ước lượng $s_{yx} = 1,8495$

Trên Bảng 12.3, tại cột thứ 3, từ trên xuống dưới chúng ta lần lượt có số liệu của SSR, SSE và SST bao gồm:

- $SSR = 171,3057$
- $SSE = 37,6277$
- $SST = 208,9333$

Còn tại cột 5 và 6 ta lần lượt có giá trị $F = 16,6930$ và Significance F = 0,0002

Trên Bảng 12.4 cung cấp cho chúng ta thông tin sau:

- Cột thứ 3 có tên Standard Error là s_b tức sai số chuẩn của hệ số độ dốc của biến độc lập thứ j
- Cột thứ 4 có tên t Stat chính là giá trị t tính toán phục vụ cho kiểm định về ý nghĩa của từng biến độc lập riêng biệt
- Cột thứ 5 có tên p-value là sự hoán đổi từ giá trị t tính toán được thành mức ý nghĩa phục vụ cho kiểm định về ý nghĩa của từng biến độc lập riêng biệt
- Hai cột cuối cùng cung cấp thông tin về cận trên và dưới của khoảng tin cậy ước lượng của hệ số hồi qui.

Lần lượt ở các nội dung kế tiếp chúng ta sẽ đi sâu vào việc tìm hiểu bản chất và cách tính toán các đại lượng này.

12.2.4 Đánh giá sự phù hợp của mô hình

Có một số phương pháp thống kê để tiến hành đánh giá sự phù hợp của mô hình là: tính toán hệ số xác định, dùng số thống kê F để đánh giá mức ý nghĩa toàn diện của mô hình, tính toán sai số chuẩn của ước lượng và đánh giá ý nghĩa của từng biến độc lập riêng biệt.

12.2.4.1 Tính toán hệ số xác định bội

Trong chương trước chúng ta đã biết là hệ số xác định R^2 đo lường phần biến thiên trong biến phụ thuộc được giải thích bởi mối liên hệ giữa biến

phụ thuộc và một biến độc lập đơn lẻ. Khi chúng ta có nhiều biến độc lập trong mô hình thì R^2 vẫn được sử dụng để xác định phần biến thiên trong biến phụ thuộc được giải thích bởi mối liên hệ giữa biến phụ thuộc và tất cả các biến độc lập trong mô hình, tuy nhiên lúc này R^2 được gọi là hệ số xác định bội, công thức tính toán hệ số xác định bội thì vẫn là

$$R^2 = \frac{SSR}{SST}$$

Sử dụng các số liệu đã xác định trên các bảng kết quả do Excel cung cấp ta tính lại số liệu về hệ số xác định, rồi so sánh với kết quả tính từ Excel

$$R^2 = \frac{SSR}{SST} = \frac{171,3057}{208,9333} = 0,8199$$

Kết quả này cho biết 81,99% biến thiên trong điểm đánh giá kết quả làm việc của nhân viên quản lý cấp trung có thể được giải thích bởi mối liên hệ tuyến tính giữa biến phụ thuộc với 3 biến độc lập trong mô hình hồi qui, tuy nhiên chú ý rằng không phải cả 3 biến độc lập này đều có tầm quan trọng ngang nhau đối với khả năng giải thích cho biến thiên trong biến phụ thuộc của mô hình.

12.2.4.2 Hệ số xác định hiệu chỉnh

Hệ số xác định hiệu chỉnh ký hiệu R^2_{adj} là một cách khác để đo lường tỷ lệ phần trăm của biến thiên được giải thích trong biến phụ thuộc mà có tính đến mối liên hệ giữa cỡ mẫu và số biến độc lập trong mô hình hồi qui bội, công thức của nó như sau:

$$R^2_{adj} = 1 - (1 - R^2) \left[\frac{n-1}{n-k-1} \right]$$

Trong đó n là cỡ mẫu và k là số biến độc lập trong mô hình

Vì sao ta lại xem xét hệ số xác định hiệu chỉnh: người ta thấy rằng với mô hình hồi qui, việc đưa thêm biến độc lập vào mô hình sẽ luôn làm tăng R^2 , thậm chí ngay cả khi các biến độc lập được đưa vào không có mối liên hệ hoặc có mối liên hệ không đáng kể với biến phụ thuộc. Vì thế khi số biến độc lập tăng lên (chưa kể đến chất lượng của biến), chắc chắn R^2 sẽ luôn luôn tăng, tuy nhiên mỗi biến độc lập thêm vào sẽ làm mất đi một bậc tự do, điều này được xem như một phần chi phí phải trả khi thêm biến độc lập. Sự gia tăng trong R^2 có thể không bù đắp được thiệt hại do mất thêm bậc tự do khi thêm biến, thế nhưng R^2_{adj} có tính đến chi phí này và điều chỉnh giá trị R^2_{adj} theo nó một cách phù hợp. R^2_{adj} sẽ luôn bé hơn R^2 . Khi một biến độc lập được thêm vào không có đóng góp xứng đáng vào khả năng giải thích cho biến phụ thuộc thì R^2_{adj} sẽ luôn

luôn giảm đi mặc dù R^2 thì tăng. Hệ số xác định hiệu chỉnh là một đại lượng đo lường quan trọng khi số biến độc lập lớn một cách tương đối so với cỡ mẫu, nó tính đến mối liên hệ giữa cỡ mẫu và số biến, nếu số biến độc lập là khá lớn so với cỡ mẫu thì R^2 sẽ thổi phồng khả năng giải thích cho biến phụ thuộc của mô hình một cách giả tạo.

Điều đó cho thấy với mô hình hồi qui đa biến, nhất là khi số biến độc lập khá lớn trong tương quan với cỡ mẫu thì ta nên dùng R^2_{adj} để đánh giá khả năng giải thích của mô hình. Ở Bảng 12.2 bên dưới đại lượng R square bạn thấy có một đại lượng là Adjusted R Square chính là R^2_{adj} chúng ta vừa nghiên cứu, giá trị của nó là 0,7708. bây giờ chúng ta thử tính theo phương pháp thủ công xem đáp số có đúng như vậy không.

$$R^2_{adj} = 1 - (1 - 0,8199) \left[\frac{15-1}{15-3-1} \right] = 1 - (0,1801) \frac{14}{11} = 0,7708$$

Hệ số xác định hiệu chỉnh luôn nhỏ hơn Hệ số xác định bội, như ở đây ta thấy hệ số xác định hiệu chỉnh bằng 0,7708 cho biết 77,08% biến thiên trong biến phụ thuộc có thể được giải thích bởi mô hình hồi qui bội mà ta đã xây dựng, rõ ràng nhận định về độ phù hợp của mô hình qua hệ số xác định hiệu chỉnh không bị thổi phồng như qua hệ số xác định bội. Vì vậy thông thường khi đánh giá độ phù hợp của mô hình hồi qui bội, bên cạnh thông tin về R^2 người ta cũng dùng thêm thông tin về R^2_{adj} để tham khảo.

12.2.4.3 Đánh giá ý nghĩa toàn diện của mô hình

Bạn luôn nhớ rằng mô hình hồi qui mà chúng ta xây dựng là dựa trên dữ liệu của một mẫu lấy từ tổng thể vì vậy nó có thể bị ảnh hưởng của sai số lấy mẫu, vì thế chúng ta cần kiểm định ý nghĩa thống kê của toàn bộ mô hình, phần trên chúng ta đã thảo luận là hệ số xác định bội R^2 , đại lượng cho biết có bao nhiêu phần biến thiên trong biến phụ thuộc có thể được giải thích bởi mô hình hồi qui, là một số thống kê trên mẫu có thể được sử dụng để suy diễn về việc mô hình toàn diện có ý nghĩa về mặt thống kê hay không trong việc giải thích cho biến thiên của biến phụ thuộc. Với ý tưởng này chúng ta thiết lập giả thuyết H_0 và giả thuyết đối H_1 như sau

$$H_0: R^2 = 0$$

$$H_1: R^2 \neq 0$$

Bản chất của giả thuyết H_0 này có nghĩa là mô hình hồi qui đa biến tổng thể mà chúng ta xây dựng với tất cả các biến độc lập được đưa vào để giải thích cho biến phụ thuộc thực ra không giải thích được chút nào cho những biến thiên trong biến phụ thuộc. Tương tự, chúng ta có thể xây dựng lại một giả thuyết có dạng biểu hiện khác như sau:

$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$

$H_1: \text{Có ít nhất một hệ số } \beta_i \text{ khác } 0$

Nếu giả thuyết H_0 trên đúng nghĩa là tất cả các hệ số độ dốc đều đồng thời bằng 0 thì mô hình hồi qui hồi qui bội đã xây dựng không hề có tác dụng trong việc dự đoán hay mô tả về biến phụ thuộc.

Trong bảng kết quả ANOVA có đại lượng thống kê F chính là con số thống kê được sử dụng để kiểm định giả thuyết về ý nghĩa toàn diện của mô hình hồi qui, công thức của đại lượng F được hình thành như sau

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

Trong công thức này các đại lượng SSR và SSE ta đã biết ở mục trên; n và k lần lượt là cỡ mẫu và số biến độc lập; cần chú ý là để quyết định ta phải tra bảng thống kê F tìm giá trị giới hạn tương ứng với mức ý nghĩa ta chọn trước, mà muốn tra bảng F ta phải có thêm thông tin về bậc tự do ở tử số và mẫu số, ta qui ước bậc tự do của tử số là $D_1 = k$ và bậc tự do của mẫu số $D_2 = (n - k - 1)$.

Ta vận dụng lại qui trình đánh giá ý nghĩa toàn diện của mô hình cho ví dụ của chúng ta:

Đặt giả thuyết

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

$H_1: \text{Có ít nhất một hệ số } \beta_i \text{ khác } 0$

Chọn độ tin cậy cho kiểm định là 95% ta có mức ý nghĩa $\alpha = 5\%$.

Với $n = 15$ và $k = 3$ ta có $D_1 = 3$ và $D_2 = (15 - 3 - 1) = 11$

Vậy tra Bảng tra số 4 của phân phối F ta tìm được giá trị F giới hạn

$F_{(D_1=3 \text{ và } D_2=11; \alpha=5\%)} = 3,587$

Tính toán giá trị kiểm định

$$F_u = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{\frac{171,3057}{3}}{\frac{37,6277}{11}} = \frac{57,1019}{3,4207} = 16,6930$$

Cũng theo qui tắc thông thường trong quyết định vì $F_u = 16,69 > F_{\text{giới hạn}} = 3,59 \rightarrow$ bác bỏ giả thuyết H_0

Như vậy chúng ta có thể kết luận rằng mô hình hồi qui bội với các biến độc lập ta đưa vào có thể giải thích một cách có ý nghĩa cho biến thiên

trong điểm đánh giá khả năng làm việc của nhân viên cấp trung. Như vậy toàn bộ mô hình có ý nghĩa về mặt thống kê. Điều này cũng có nghĩa là để ước lượng điểm đánh giá khả năng làm việc của nhân viên cấp trung, nếu sử dụng mô hình hồi qui bội đã xây dựng ta sẽ đạt được kết quả khả quan hơn khi dùng số trung bình về điểm đánh giá khả năng làm việc của nhân viên cấp trung như một con số ước lượng đơn giản nhất.

Trong Bảng 12.3 tại cột số 5 chúng ta thấy Excel cũng cung cấp luôn giá trị F tính toán được là 16,69306. Để giảm bớt phiền toái của việc phải tra bảng F chúng ta sử dụng luôn giá trị Significance F = 0,0002 tại cột số 6 của bảng ANOVA, ta thấy với mức ý nghĩa 5% đã chọn thì phép so sánh Significance F = 0,0002 < $\alpha = 0,05 \rightarrow$ bác bỏ giả thuyết H_0 . Như vậy ta có cùng kết luận với kết quả kiểm định thủ công vừa rồi. Chú ý giá trị Significance F chính là giá trị p-value.

12.2.4.4 Tính toán sai số chuẩn của ước lượng

Mục tiêu của việc xây dựng mô hình hồi qui là để có thể xác định được giá trị của biến phụ thuộc khi biết trước các giá trị cụ thể của biến độc lập. Một số thống kê cho thấy mô hình hồi qui thực hiện mục tiêu này tốt đến đâu là độ lệch chuẩn của mô hình hồi qui (còn gọi tên là Sai số chuẩn của ước lượng) đo lường sự phân tán của các giá trị thực tế đo lường được của biến phụ thuộc quanh những giá trị của biến phụ thuộc được dự đoán bằng đường hồi qui. Ví dụ sự phân tán của điểm đánh giá khả năng làm việc của 15 nhân viên, tức là Y_i , quanh những giá trị về điểm đánh giá khả năng làm việc được dự đoán bởi mô hình hồi qui. Giá trị ước lượng từ thông tin mẫu của độ lệch chuẩn của mô hình hồi qui (sai số chuẩn của ước lượng) được tính toán như sau đây:

$$s_{Y/X} = \sqrt{\frac{SSE}{n - k - 1}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - k - 1}}$$

Bản chất của đại lượng này đã được chúng ta thảo luận ở Chương 11, các ký hiệu trong công thức cũng không hề mới lạ, với n là cỡ mẫu và k là số biến độc lập trong mô hình.

Quay lại ví dụ của chúng ta, tính toán thủ công đại lượng này ta được kết quả:

$$s_{Y/X} = \sqrt{\frac{SSE}{n - k - 1}} = \sqrt{\frac{37,6277}{15 - 3 - 1}} = \sqrt{3,4207} = 1,8495$$

Trong Bảng 12.2 bạn đọc cũng thấy đại lượng này được tính toán sẵn với kết quả không khác biệt.

12.2.4.5 Đánh giá ý nghĩa của từng biến độc lập riêng biệt

Ở kiểm định F chúng ta đã kết luận được mô hình toàn diện có ý nghĩa. Điều này có nghĩa là có ít nhất một biến độc lập trong mô hình có thể giải thích được một cách có ý nghĩa cho biến thiên trong biến phụ thuộc. Tuy nhiên điều này không có nghĩa là tất cả các biến độc lập đưa vào mô hình đều có ý nghĩa, để xác định biến độc lập nào có ý nghĩa chúng ta kiểm định giả thuyết sau:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Với ví dụ của chúng ta có ba biến độc lập trong mô hình, vậy j lần lượt nhận các giá trị 1,2,3.

Chúng ta có thể dùng kiểm định t để kiểm định ý nghĩa của mỗi hệ số hồi qui với độ tin cậy 95%, giá trị t tính toán được sẽ được so sánh với giá trị t giới hạn tra từ bảng phân phối student với $(n-k-1)$ bậc tự do và mức ý nghĩa $\alpha/2 = 0,05/2 = 0,025$.

Nhắc lại rằng giá trị t tính toán được xác định bằng cách chia hệ số hồi qui mẫu cho ước lượng độ lệch tiêu chuẩn của hệ số hồi qui.

$$t = \frac{b_j - 0}{s_{b_j}}$$

Với b_j là hệ số độ dốc trong mô hình hồi qui mẫu cho biến độc lập thứ j

s_{b_j} là sai số chuẩn ước lượng của hệ số độ dốc của biến độc lập thứ j

Các đại lượng này đều được Excel tính toán sẵn và cung cấp trong bảng Coefficient (Bảng 12.4)

Bây giờ chúng ta thực hiện kiểm định ý nghĩa của biến độc lập X_3 , biến có tên “Điểm khả năng trình bày miệng” xem thử biến này có thể giải thích được một cách có ý nghĩa cho biến thiên trong biến phụ thuộc Y hay không.

Đặt giả thuyết

$$H_0: \beta_3 = 0$$

$$H_1: \beta_3 \neq 0$$

Chọn độ tin cậy 95% $\rightarrow \alpha/2 = 0,05/2 = 0,025$

Bậc tự do $df = n - k - 1 = 15 - 3 - 1 = 11$

Tra bảng phân phối Student ta tìm được giá trị giới hạn $t_{(11; 0,025)} = 2,201$

Tính toán giá trị t cho kiểm định:

$$t_n = \frac{3,8074 - 0}{2,4930} = 1,5272$$

Giá trị $t_u = 1,5272 < t_{(11; 0,025)} = 2,201$ nên ta chấp nhận giả thuyết Ho và kết luận rằng Điểm khả năng trình bày miệng không có khả năng giải thích cho Điểm đánh giá khả năng làm việc của nhân viên.

Nếu vận dụng phương pháp kiểm định qua giá trị p-value ta cũng có cùng kết luận.

Chú ý rằng tất cả b_j , s_{bj} , t_{bj} , p-value đều có sẵn trên Bảng 12.4

Tại sao hệ số xác định và kiểm định F cho thấy mô hình tổng thể khá phù hợp mà kiểm định ý nghĩa của từng hệ số riêng biệt lại cho thấy không chỉ X_3 , mà cả X_2 và X_1 đều không có khả năng giải thích cho Y nếu chọn mức ý nghĩa cho kiểm định t là 5%. Chúng ta sẽ có giải đáp ở phần tiếp theo đây.

12.2.5 Hiệu tượng đa cộng tuyến

12.2.5.1 Ảnh hưởng của đa cộng tuyến

Một trong những yêu cầu của mô hình hồi qui bội tuyến tính là các biến độc lập không có tương quan chặt với nhau, nếu yêu cầu này không được thỏa mãn, người ta bảo rằng đã xảy ra hiện tượng đa cộng tuyến trong mô hình hồi qui bội. Vậy yêu cầu này có được thoả mãn không với những dữ liệu ở ví dụ trên, thông thường chúng ta dùng phương pháp gì để kiểm tra sự tồn tại của hiện tượng đa cộng tuyến. Xảy ra đa cộng tuyến thì mô hình chịu tác động gì, ta phải xử lý như thế nào nếu trong mô hình có hiện tượng đa cộng tuyến?

Đa cộng tuyến giữa các biến giải thích trong một mô hình hồi qui bội là tình huống trong đó hai hoặc hơn hai biến độc lập có tương quan tuyến tính chặt chẽ với nhau. Trong tình huống này các biến có tương quan tuyến tính chặt chẽ với nhau không cung cấp được thông tin gì mới và cũng không thể xác định được ảnh hưởng riêng biệt của từng biến độc lập lên biến phụ thuộc.

Khi xảy ra đa cộng tuyến trong mô hình sẽ làm phương sai của các ước lượng hệ số hồi qui (S^2_{bj}) có giá trị rất lớn, người ta chứng minh được trong mô hình hồi qui bội với hai biến giải thích X_1 , X_2 dùng để giải thích cho Y, phương sai của hệ số hồi qui b_2 được xác định bằng một công thức mà ở tử số có lượng $(1 - r^2_{12})$ với r_{12} là hệ số tương quan tuyến tính giữa hai biến X_1 và X_2 . Chính vì vậy nếu r_{12} lớn sẽ làm cho $(1 - r^2_{12})$ nhỏ và vì lượng này nằm ở mẫu nên sẽ làm cho phương sai của b_2 lớn tức sai số chuẩn ước lượng của hệ số độ dốc b_2 cũng lớn. Hãy liên tưởng đến công thức t trong

nội dung Đánh giá ý nghĩa của từng biến độc lập riêng biệt, nếu bạn muốn đánh giá ý nghĩa của biến X_2 , sử dụng công thức tính t với s_{b2} nằm ở mẫu số, nếu s_{b2} lớn thì t sẽ nhỏ khiến cho bạn dễ dàng kết luận chấp nhận giả thuyết H_0 là biến X_2 không có ý nghĩa giải thích cho Y mặc dù trên thực tế thì tỉ số R^2 lại cao, như ví dụ về Điểm đánh giá khả năng làm việc của nhân viên đã cho thấy. Thậm chí đa cộng tuyến có thể gây ra hiện tượng làm sai dấu của hệ số hồi qui so với lý thuyết giả dụ thay vì một hệ số dương cho thấy thu nhập tăng thì chi tiêu tăng nhưng do sự có mặt của biến "sự giàu có" vốn tương quan chặt với thu nhập nên làm cho hệ số hồi qui đứng trước biến thu nhập lại mang dấu âm hàm ý thu nhập tăng thì chi tiêu giảm khiến cho bạn khó khăn trong việc diễn đạt ý nghĩa của nó.

12.2.5.2 Cách phát hiện mô hình có tồn tại hiện tượng đa cộng tuyến

- Dấu hiệu dễ gây nghi ngờ nhất là R^2 của mô hình cao mà kiểm định t lại bảo một vài biến độc lập nào đó không có ý nghĩa trong việc giải thích cho Y , từ nghi ngờ này người ta dùng phương pháp đơn giản nhất để phát hiện mô hình có tồn tại hiện tượng đa cộng tuyến không là xem xét hệ số tương quan tuyến tính giữa các biến độc lập, quay lại với ví dụ của chúng ta, sử dụng Excel tính toán hệ số tương quan tuyến tính giữa 3 biến độc lập X_1 , X_2 và X_3 ta có kết quả sau:

r_{12}	r_{13}	r_{23}
0.5781	0.6761	0.8675

Kết quả này cho thấy giữa biến X_3 và biến X_1 có mối liên hệ tuyến tính đáng kể, giữa X_1 và X_2 có mối liên hệ tương quan khá chặt.

- Một phương pháp thứ hai để xác định đa cộng tuyến là dùng nhân tố phỏng đại phương sai VIF có công thức như sau đây với mô hình hồi qui có k biến giải thích

$$VIF_j = \frac{1}{1 - R_j^2}$$

Trong đó R_j^2 là giá trị hệ số xác định trong hàm hồi qui của biến giải thích thứ j theo $(k-1)$ biến giải thích còn lại, nếu có cộng tuyến của X_j với các biến giải thích khác thì R_j^2 sẽ gần bằng 1 và do đó VIF_j sẽ lớn, giá trị VIF_j càng lớn thì biến X_j càng cộng tuyến cao, nhưng VIF_j bằng bao nhiêu thì có thể xem như xảy ra

hiện tượng đa cộng tuyến, như một qui tắc kinh nghiệm nếu VIF_j bằng hoặc vượt quá 5 (khi đó $R_j^2 > 0,8$) thì xem như có đa cộng tuyến giữa X_j và các biến độc lập kia. Cũng chú ý rằng có thể trong một số sách thống kê bạn thấy người ta bảo rằng nếu VIF_j bằng hoặc vượt quá 10 (khi đó $R_j^2 > 0,9$) thì xem như có đa cộng tuyến giữa các biến độc lập, khác biệt này đơn giản là xuất phát từ quan điểm khác nhau của các nhà thống kê về mức độ liên kết tuyến tính thế nào là “chặt”.

Với một số phần mềm thống kê chẳng hạn SPSS thì bạn có thể yêu cầu tính toán và cho luôn các VIF_j , nhưng với Excel bạn phải tính thủ công.

Với ví dụ của chúng ta, có 3 biến độc lập, bây giờ thử hồi qui X_3 theo hai biến còn lại là X_1 và X_2 , lấy hệ số xác định R^2 để tính toán VIF_3 .

Thực hiện hồi qui bội trong đó X_3 ở vị trí biến độc lập còn X_1 và X_2 là hai biến giải thích ta thấy hệ số xác định bội $R^2 = 0,7983$ (bạn đọc tự kiểm tra bằng Excel). Từ đây ta thay thế $R^2_3 = 0,7983$ vào công thức tính VIF_3 được kết quả là:

$$VIF_3 = \frac{1}{1 - R^2_3} = \frac{1}{1 - 0,7983} = 4,9579$$

Có thể làm tròn $VIF_3 = 5$ và như vậy ta thấy VIF cho thấy X_3 có tương quan tuyến tính chặt với hai biến độc lập còn lại. Như vậy X_3 có khả năng gây ra đa cộng tuyến.

Nếu bạn tiếp tục thực hiện hồi qui bội X_1 theo X_2 với X_3 và lặp lại tiến trình trên bạn được $VIF_1 = 1/(1 - 0,4574) = 1,8429 < 5$. Như vậy X_1 không có khả năng gây ra đa cộng tuyến.

Nếu bạn tiếp tục thực hiện hồi qui bội X_2 theo X_1 với X_3 và lặp lại tiến trình trên bạn được $VIF_2 = 1/(1 - 0,7527) = 4,0437 < 5$. Như vậy X_2 không có khả năng gây ra đa cộng tuyến.

Kết hợp đồng thời kết quả của hai phương pháp dò tìm đa cộng tuyến vừa nghiên cứu xong, ta nhận định sơ bộ là X_3 gây đa cộng tuyến do nó tương quan chặt chẽ với X_2 và X_1 .

12.2.5.3 Khắc phục đa cộng tuyến

Để khắc phục đa cộng tuyến, có một số biện pháp mà biện pháp đơn giản nhất là hồi qui lại mô hình hồi qui bội này mà bỏ đi biến độc lập gây ra đa cộng tuyến tức là ở đây ta bỏ đi X_3 . Ngoài ra người ta còn có thể khắc phục đa cộng tuyến bằng cách lấy thêm số liệu hoặc chọn lại một mẫu mới (tuy

nhiên phương pháp này không đảm bảo lắm), hoặc hồi qui sai phân cấp 1 của Y theo sai phân cấp 1 của các X_j , tuy nhiên phương pháp có thể gây ra các vấn đề nghiêm trọng khác. Nếu bỏ bớt biến giải thích X, thì mô hình hồi qui của chúng ta lúc này có kết quả như sau:

Bảng 12.5

Regression Statistics	
Multiple R	0.884149
R Square	0.78172
Adjusted R Square	0.74534
Standard Error	1.949487
Observations	15

Bảng 12.6

ANOVA

	df	SS	MS	F	Significance F
Regression	2	163.327	81.6637	21.4876	0.0001082
Residual	12	45.606	3.8005		
Total	14	208.933			

Bảng 12.7

Coefficients	Standard		P-value	Lower		Upper	
	Error	t Stat		95%	95%		
Intercept	30.4628	9.2079	3.3083	0.0062	10.4005	50.5252	
X1	3.1418	1.1266	2.7888	0.0164	0.6872	5.5964	
X2	3.8788	1.1393	3.2290	0.0072	1.1964	6.1611	

Các bạn chú ý một điều là đa cộng tuyến khiến cho đánh giá của chúng ta về tác động của từng biến độc lập lên biến phụ thuộc có thể bị sai lệch nhưng nó không làm giảm hệ số xác định, tức là tác động gộp của tất cả các biến độc lập lên biến phụ thuộc trong việc giải thích biến thiên của biến phụ thuộc không bị ảnh hưởng xấu bởi đa cộng tuyến. Do đó trừ phi bạn dự định xây dựng mô hình hồi qui để thực hiện chính sách, tức là định lượng xem tác động vào các X_j như thế nào để cho Y thay đổi theo ý muốn, còn nếu bạn dùng mô hình hồi qui để dự đoán Y khi biết trước các giá trị của X_j thì đa cộng tuyến không phải là vấn đề nghiêm trọng, nhất là khi lý thuyết cho bạn thấy sự tồn tại của biến độc lập mà bạn định loại trừ đi khi khắc phục đa cộng tuyến là cần thiết đối với biến phụ thuộc. Như tình huống trong ví dụ của chúng ta đây, sau khi bạn loại trừ X_3 thì kiểm định t cho thấy cả hai biến độc lập đều có ý nghĩa (cột p-value của Bảng 12.7) nhưng hệ số xác định giảm xuống, sai số chuẩn của hồi qui tăng lên (Bảng 12.5). Như vậy việc loại bỏ biến X_3 tuy khắc phục được đa cộng tuyến nhưng nhìn chung đã làm giảm độ phù hợp của mô hình. Mục

tiêu của bà giám đốc nhân sự là muốn sử dụng ba biến độc lập là Điểm phân tích tình huống (ký hiệu X_1), Điểm khả năng trình bày viết (ký hiệu X_2), Điểm khả năng trình bày miệng (ký hiệu X_3) để giải thích cho biến phụ thuộc là Điểm đánh giá kết quả làm việc (ký hiệu là Y), như vậy ta nên giữ lại X_3 để tăng khả năng giải thích cho Y , nhất là sự có mặt của X_3 trong mô hình là hợp lý về mặt lý thuyết.

12.2.6 Diễn giải các ý nghĩa các hệ số hồi qui riêng

Trong chương trước đầu tiên chúng ta đã thảo luận về cách diễn dịch ý nghĩa của các hệ số hồi qui, sau đó chúng ta tìm hiểu về cách đánh giá độ phù hợp của mô hình. Trong thực tế thường chúng ta đảo ngược tiến trình. Đó là đầu tiên chúng ta xác định mô hình phù hợp đến đâu, nếu độ phù hợp của mô hình thấp thì không cần thiết phân tích xa hơn ý nghĩa các hệ số của mô hình. Một bước kế tiếp là khảo sát xem mô hình có tồn tại hiện tượng đa cộng tuyến hay không, nếu có thì phải cải tiến mô hình. Rồi sau cùng chúng ta ta mới diễn dịch ý nghĩa của các hệ số hồi qui riêng phần.

Xem lại mô hình hồi qui mẫu

$$\hat{Y}_i = 21,4805 + 2,3640 X_{1i} + 1,5313 X_{2i} + 3,8074 X_{3i}$$

Hệ số chặn $b_0 = 21,4805$ chính là giá trị ước lượng về điểm đánh giá kết quả làm việc của một nhân viên nếu người đó có Điểm phân tích tình huống = 0; Điểm trình bày miệng = 0 và Điểm trình bày viết = 0. Trong nhiều trường hợp khoảng giá trị mà các biến độc lập nhận nằm ngoài phạm vi giá trị 0 nên sự suy diễn về hệ số chặn thường tỏ ra không phù hợp.

Ý nghĩa của các hệ số độ dốc: Trong nội dung trước chúng ta đã biết các hệ số hồi qui trong tình huống này là hệ số hồi qui riêng phần, bạn sẽ thấy chúng được diễn tả như thế nào:

- Hệ số $b_1 = 2,3640$ cho biết khi Điểm phân tích tình huống tăng thêm 1 điểm, trong điều kiện Điểm khả năng trình bày viết và Điểm khả năng trình bày miệng của nhân viên không đổi, thì điểm đánh giá kết quả làm việc của nhân viên được ước lượng sẽ tăng thêm trung bình 2,3640 điểm.
- Hệ số $b_2 = 1,5313$ cho biết khi Điểm khả năng trình bày viết tăng thêm 1 điểm, trong điều kiện Điểm phân tích tình huống và Điểm khả năng trình bày miệng của nhân viên không đổi, thì điểm đánh giá kết quả làm việc của nhân viên được ước lượng sẽ tăng thêm trung bình 1,5313 điểm.
- Hệ số $b_3 = 3,8074$ cho biết khi Điểm khả năng trình bày miệng tăng thêm 1 điểm, trong điều kiện Điểm phân tích tình huống và Điểm

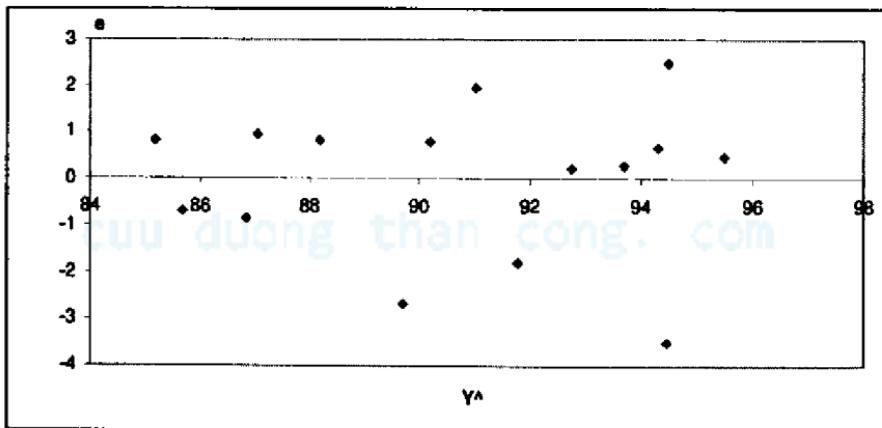
khả năng trình bày viết của nhân viên không đổi, thì điểm đánh giá kết quả làm việc của nhân viên được ước lượng sẽ tăng thêm trung bình 3,807 điểm.

12.2.7 Phân tích phần dư

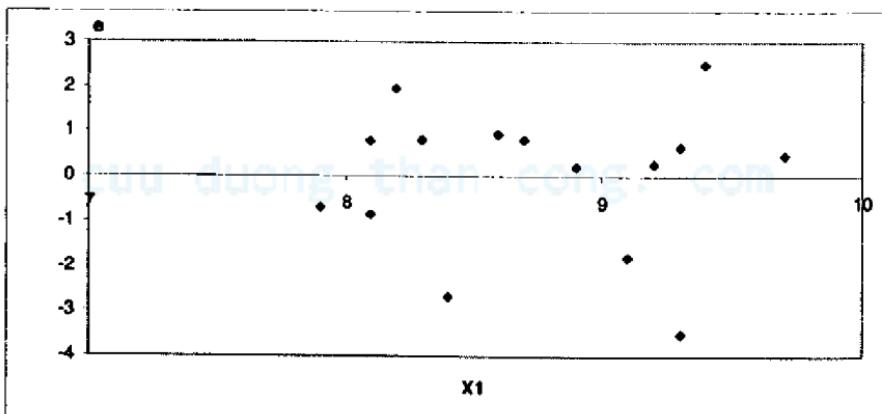
12.2.7.1 Kiểm tra sự phù hợp khi lựa chọn mô hình hồi qui tuyến tính

Vẽ đồ thị phân dư lần lượt theo giá trị \hat{Y} ước lượng được từ mô hình và từng biến độc lập, nếu các điểm phân tán trên đồ thị này không thể hiện một hình dáng cụ thể nào cho mối liên hệ giữa phần dư và các biến độc lập cũng như mối liên hệ giữa phần dư và giá trị dự đoán từ mô hình của biến phụ thuộc, thì như vậy sơ bộ ta kết luận là mô hình hồi qui bội mô tả liên hệ tuyến tính là khá phù hợp với tình huống nghiên cứu của chúng ta.

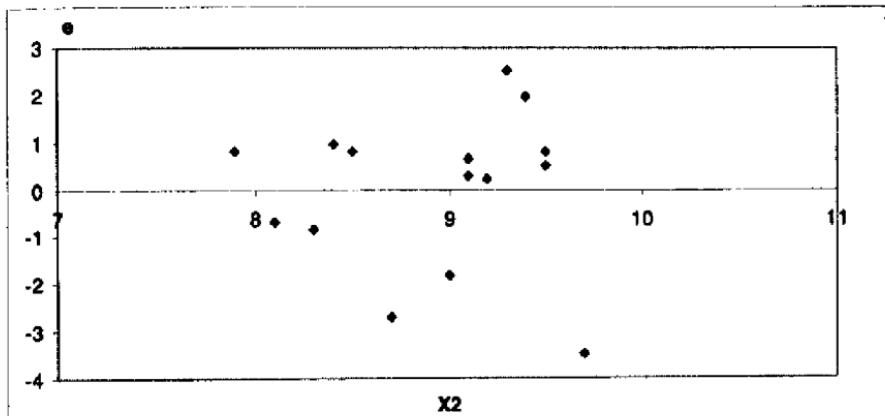
Hình 12.2



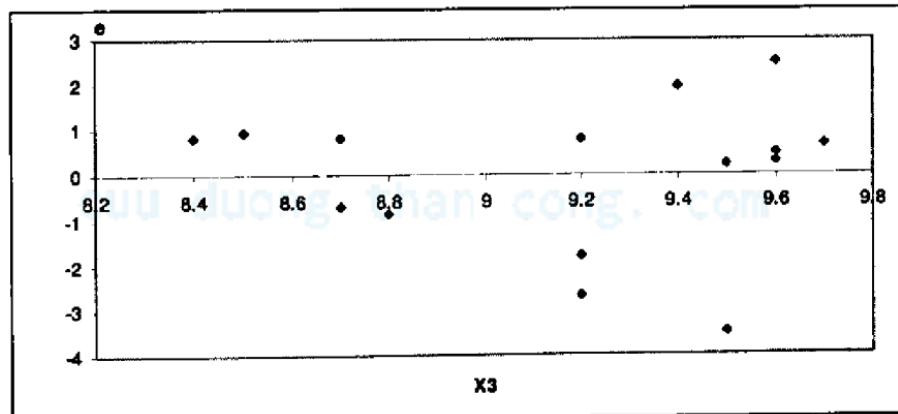
Hình 12.3



Hình 12.4



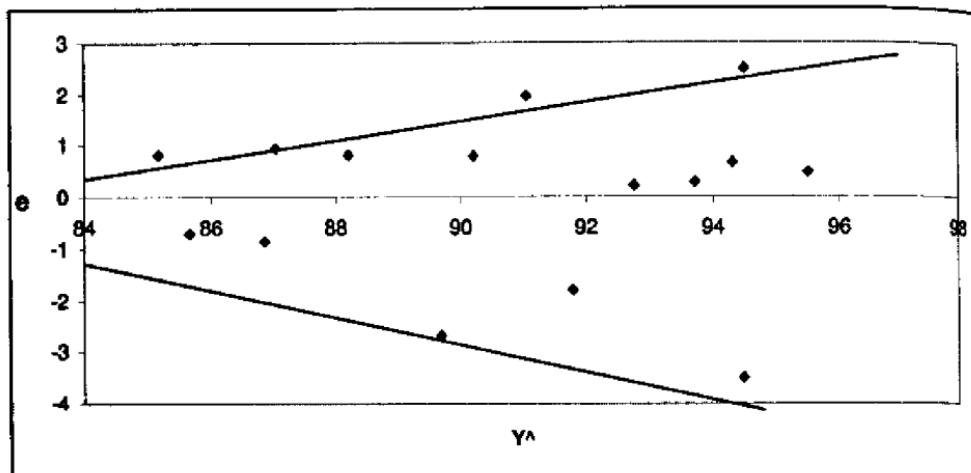
Hình 12.5



12.2.7.2 Kiểm tra giả định phương sai không đổi

Vẽ đồ thị phần dư theo giá trị \hat{Y} ước lượng từ mô hình hồi qui, cảm nhận ban đầu cho ta thấy hình như có tồn tại hiện tượng phương sai thay đổi trong mô hình nhưng kết luận này không rõ ràng lắm. Do đó ta sẽ tiến hành kiểm định Park trong đó ta chạy mô hình hồi qui $\ln(e^2)$ theo \hat{Y} .

Hình 12.6



Kết quả chạy hồi qui như sau

Coefficients	Standard					
	Error	t Stat	P-value	Lower 95%	Upper 95%	
Intercept	-0.8795	11.5288	0.0763	0.9404	-25.7859	24.0270
\hat{Y}^A	0.0087	0.1270	0.0684	0.9465	-0.2656	0.2830

Rõ ràng hệ số độ dốc đứng trước biến \hat{Y} không có ý nghĩa thống kê nên ta kết luận là hiện tượng phương sai thay đổi không xảy ra. Nếu bạn đọc bở thời gian thực hiện mô hình hồi qui $\ln(e^2)$ theo các biến độc lập đã lấy Logarit cơ số e thì bạn cũng có kết luận tương tự.

12.2.7.3 Kiểm tra giả định không có tự tương quan giữa các phần dư

Với một số phần mềm thống kê chẳng hạn SPSS thì bạn có thể yêu cầu tính toán và cho luôn kết quả về số thống kê Durbin Watson, còn với Excel bạn phải tự tính trên các số liệu về phần dư e_i . Chúng ta vận dụng lại công thức tính toán giá trị D của kiểm định Durbin Watson ở Chương 11.

cuuduongthancong.com

Bảng 12.8

e_i	$e_i - e_{i-1}$	$(e_i - e_{i-1})^2$	e_i^2
-2,68866	/	/	7,228883
1,950721	4,639379	21,52384	3,805312
-3,48981	-5,44054	29,59942	12,1788
-0,68416	2,805652	7,871682	0,468078
-0,84397	-0,15981	0,025538	0,712283
2,505587	3,349556	11,21953	6,277969
-1,80286	-4,30845	18,5627	3,250297
0,221457	2,024315	4,097852	0,049043
0,963114	0,741657	0,550054	0,927588
0,49012	-0,47299	0,223723	0,240217
0,818724	0,328604	0,107981	0,670308
0,812103	-0,00662	4,38E-05	0,659512
0,284656	-0,52745	0,278201	0,081029
0,795462	0,510806	0,260923	0,632759
0,667518	-0,12794	0,016369	0,445581
Tổng		94,3379	37,62766

Thể số vào công thức

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} = \frac{94,3379}{37,6277} = 2,5071$$

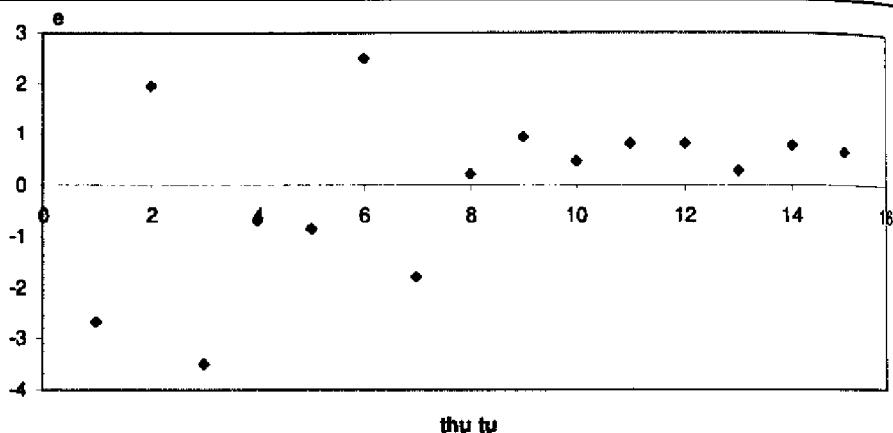
Nếu tra theo bảng Durbin Watson với $n = 15$ và $k = 3$, $\alpha = 5\%$ ta được $d_L = 0,82$ $d_U = 1,75$. Theo như hình vẽ ta lập được các khoảng giá trị như sau:



Như vậy giá trị D rơi vào vùng không có quyết định, nhưng nếu vận dụng qui tắc kiểm định đơn giản ta thấy $1 < D = 2,5071 < 3$ nên kết luận mô hình không có tự tương quan giữa các phần dư.

Dùng phương pháp vẽ đồ thị phần dư theo thứ tự quan sát ta cũng không thấy một hình dạng đặc biệt nào trên đồ thị, nên có thể tin không xảy ra hiện tượng tự tương quan.

Hình 12.7



12.2.8 Dự đoán giá trị cụ thể của biến phụ thuộc

Một nhà quản lý cấp trung đã đạt được các điểm số sau: phân tích tình huống: 9,1; khả năng trình bày viết: 9,4; khả năng trình bày nói: 9,3. Căn cứ trên dữ liệu này hãy ước lượng điểm đánh giá kết quả làm việc của nhân viên đó là bao nhiêu?

$$\hat{Y}_i = 21,4805 + 2,3640 \times 9,1 + 1,5313 \times 9,4 + 3,8074 \times 9,3 = 92,7959$$

Như vậy chúng ta kết luận điểm trung bình đánh giá kết quả làm việc của nhân viên cấp trung sẽ là 92,7959 điểm nếu Điểm phân tích tình huống là 9,1 điểm; Điểm trình bày viết là 9,4 và điểm khả năng trình bày miệng là 9,3.

12.3 HỒI QUI VỚI BIẾN ĐỘC LẬP ĐỊNH TÍNH

Trong những mô hình hồi qui ta đã nghiên cứu, chúng ta mới chỉ gặp tình huống hồi qui với biến độc lập là biến định lượng chứ chúng ta chưa làm việc với biến độc lập dạng định tính, tuy nhiên trong thực tế điều này rất hay xảy ra, ví dụ như những biến mô tả tình trạng hôn nhân, giới tính, khu vực địa lý, mùa vụ trong năm.. Vậy làm thế nào để ta vẫn có thể sử dụng các biến này trong mô hình hồi qui tuyến tính. Câu trả lời nằm trong phương pháp sử dụng biến giả. Tên gốc tiếng Anh của biến giả là Dummy vì vậy người ta vẫn thường ký hiệu nó là D.

Do các biến định tính được sử dụng trong hồi qui (như giới tính nam hay nữ, loại hình doanh nghiệp trong hay ngoài quốc doanh, chủng tộc da trắng hay da màu..) thường mô tả sự xuất hiện hoặc vắng mặt của một

“tính chất” hay đặc điểm (như nam hay không phải nam, trong quốc doanh hay không phải trong quốc doanh, da trắng hay không phải da trắng...), phương pháp “lượng hoá” các thuộc tính được thực hiện bằng cách thiết lập các biến nhân tạo với giá trị 1 biểu thị việc có thuộc tính đó và 0 là tình huống ngược lại. Ví dụ nếu bạn quan tâm đến tiêu thức giới tính, bạn sẽ thấy rằng chỉ xảy ra hai tình huống là có thuộc tính nữ và không có thuộc tính nữ (tức là nam); nếu quan tâm đến tiêu thức chủng tộc bạn có thể qui ước có thuộc tính da trắng và không có thuộc tính da trắng (tức là da màu). Các biến nhận giá trị 0 và 1 được gọi là các biến giả. Biến giả còn có tên gọi là biến nhị phân, biến chỉ định, và dĩ nhiên còn gọi là biến định tính, biến giả được định nghĩa như sau:

Biến giả là một biến nhận hai giá trị 0 hoặc 1, phụ thuộc vào việc quan sát có hay không có một thuộc tính cụ thể nào đó.

Từ đây, giới tính có thể được chuyển hóa thành biến D với qui ước như sau:

$D = 1$ nếu là nữ

$D = 0$ nếu là nam (không là nữ)

Như vậy trong tập dữ liệu của bạn, giới tính của người được quan sát lúc này được thể hiện qua biến D trong đó nếu người được quan sát là nam giới thì biến D mang giá trị 0 còn nếu là nữ thì biến D mang giá trị 1.

Nhưng nếu bạn có biến định tính mà có hơn hai biểu hiện ví dụ biến định tính mô tả 4 mùa trong năm hay biến tình trạng hôn nhân với các biểu hiện Chưa lập gia đình – Có gia đình – Ly thân/Ly dị - Góa. Lúc này bạn phải dùng đến hơn một biến D theo qui tắc sau, nếu tiêu thức quan tâm có m biểu hiện thì phải tạo $(m-1)$ biến D. Các biến định tính lúc này được mã hóa như sau:

$D_1 = 1$ nếu chưa lập gia đình; $= 0$ nếu là tình huống khác

$D_2 = 1$ nếu có gia đình; $= 0$ nếu là tình huống khác

$D_3 = 1$ nếu Ly thân/Ly dị; $= 0$ nếu là tình huống khác

Chú ý rằng bạn không cần phải tạo đến 4 biến D vì nếu muốn mô tả một cá nhân đang trong tình trạng ở góa thì bạn chỉ cần sắp xếp cho các biến $D_1 = 0$, $D_2 = 0$, $D_3 = 0$. Ở đây tình trạng ở góa được gọi tên là phân loại cơ sở.

Để minh họa cách sử dụng biến giả bạn hãy xem xét các tình huống sau:

Ý dụ biến giả trong trường hợp có 2 phân loại

Người ta lấy một mẫu gồm 15 cán bộ quản lý cấp trung làm việc tại các công ty đa quốc gia lớn và thu thập thông tin về mức lương (đô la/năm) ký hiệu Y, số năm kinh nghiệm (năm) ký hiệu X. Mục tiêu của nhà nghiên

cứu là tìm cách giải thích cho mức lương bằng cách sử dụng thông tin về số năm kinh nghiệm, để mở rộng mô hình nhằm giải thích tốt hơn những biến thiên trong mức lương của cấp quản lý, người ta thu thập thêm thông tin về việc người đó có bằng MBA (Thạc sĩ Quản trị kinh doanh) hay không, biến giả D sẽ được dùng đại diện cho thông tin này, trong đó 1 là có bằng Thạc sĩ và 0 là không có bằng Thạc sĩ, ta có bảng số liệu như sau:

Bảng 12.9

STT	Y_i	X_i	D_i
1	23100	14	0
2	23000	11	1
3	25000	15	0
4	28000	15	1
5	19500	9	0
6	24000	10	1
7	29500	16	1
8	21000	12	0
9	25000	13	1
10	22000	12	0
11	26500	14	1
12	26000	16	0
13	27500	17	0
14	31500	18	1
15	29000	18	0

Chúng ta sẽ xây dựng một mô hình hồi qui tuyến tính bội trên dữ liệu mẫu giúp giải thích cho mức lương nhân viên, có phương trình như sau

$$\hat{Y} = b_0 + b_1 X + b_2 D$$

Sử dụng Excel thực hiện hồi qui theo cách thức quen thuộc, nhớ rằng trong tiến trình hồi qui đó chúng ta đổi xử với D như với một biến độc lập bình thường, ta thu được các kết quả như sau:

Bảng 12.10

Regression Statistics	
Multiple R	0.9795
R Square	0.9594
Adjusted R Square	0.9526
Standard Error	733.2659
Observations	15

Bảng 12.11

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	1.5E+08	7.6E+07	141.64729	4.50E-09
Residual	12	6.5E+06	5.4E+05		
Total	14	1.6E+08			

Bảng 12.12

	<i>Coefficients</i>	<i>Standard</i>				
		<i>Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	8993.5232	1022.1508	8.7986	0.0000	6766.4479	11220.5985
XI	1072.1400	69.9985	15.3166	0.0000	919.6264	1224.6535
DI	2935.3946	379.9589	7.7256	0.0000	2107.5353	3763.2539

Từ số liệu thu được, mô hình hồi qui tuyến tính trên mẫu được viết lại

$$\hat{Y} = 8993,5232 + 1072,14X + 2935,3946D$$

Chú ý là độ phù hợp của mô hình là rất tốt và kiểm định t cho chúng ta thấy biến giả có khả năng giải thích tốt cho mức lương của nhân viên vì nó có ý nghĩa thống kê (xem giá trị p-value trong Bảng 12.12).

Vì biến giả được mã hóa với hai giá trị duy nhất là 0 và 1 tùy theo tình trạng thực tế là người nhân viên đó có bằng Thạc sĩ hay không nên khi kết hợp với mô hình hồi qui trên ta xác định được 2 mô hình tuyến tính đơn biến có cùng hệ số dốc nhưng khác nhau về hệ số chặn như sau:

Khi người đó không có bằng Thạc sĩ (tức D = 0)

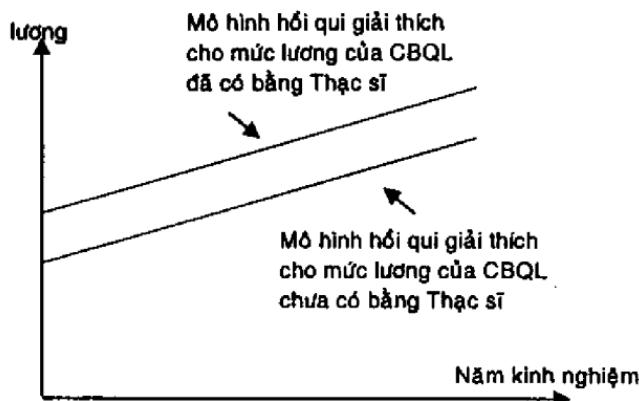
$$\hat{Y} = 8993,5232 + 1072,14 \times X$$

Khi người đó có bằng Thạc sĩ (tức D = 1)

$$\hat{Y} = 11928,917 + 1072,14 \times X$$

Hàm số mức lương của các nhân viên có cùng độ dốc nhưng tung độ gốc thì khác nhau, nói cách khác nhân viên chưa có bằng thạc sĩ hay đã có bằng thạc sĩ đều có tốc độ thay đổi mức lương trung bình theo số năm kinh nghiệm như nhau, chỉ khác nhau về mức lương khởi điểm (do bằng cấp dẫn đến). Cụ thể, nhân viên đã có bằng thạc sĩ sẽ có mức lương khởi điểm cao hơn những nhân viên chưa có bằng thạc sĩ là 2935,3946 đô/năm.

Hình 12.8



Một số vấn đề liên quan đến việc dùng biến giả

- Để phân biệt m phân loại người ta dùng m - 1 biến giả
- Việc gán giá trị 1 và 0 cho phân loại nào không quan trọng. Nếu dữ liệu là đúng đắn thì kết quả sẽ hợp lý, điều then chốt là phải biết các giá trị được gán như thế nào trong khi giải thích kết quả hồi qui
- Phân loại nhận giá trị 0 được gọi tên là phân loại cơ sở, gọi là cơ sở xét trên khía cạnh ta thực hiện các so sánh với phân loại đó.
- Hệ số gắn với biến giả D được gọi là hệ số tung độ gốc chênh lệch

Ví dụ biến giả trong trường hợp có hơn 2 phân loại

Chúng ta xét đến một ví dụ thứ hai trong đó vấn đề ta quan tâm có hơn 2 phân loại nên phải sử dụng hơn một biến định tính. Một người kinh doanh xe hơi cũ nhận thấy rằng trong nhiều yếu tố ảnh hưởng đến giá bán xe thì màu sắc của xe và số km xe đã đi (thể hiện trên đồng hồ cây số) là các yếu tố có ảnh hưởng trực tiếp, vì thế anh ta đã muốn xây dựng một mô hình hồi qui bội trong đó sử dụng các biến độc lập là màu xe và số km xe đã được sử dụng để ước lượng giá bán của xe. Anh ta tiến hành lấy thông tin trên 44 chiếc xe cũ đã tiêu thụ được (Biến Y đại diện cho giá bán xe, Biến X đại diện cho số km xe đã được đi) các xe này cùng một hãng sản xuất, tương đối ngang bằng về các điều kiện như điều hòa không khí, máy hát, hộp số tự động... Riêng về màu xe anh ta nhận thấy rằng xe màu trắng và xe màu bạc có giá bán cao hơn các loại xe màu khác, vì thế anh ta sử dụng hai biến giả để mô tả màu xe như sau:

$D_1 = 1$ nếu màu trắng

= 0 nếu màu khác

$D_2 = 1$ nếu màu bạc

= 0 nếu màu khác

Như vậy phân loại cơ sở ở đây là xe có màu khác ngoài màu trắng và màu bạc, tức $D_1 = 0$ và $D_2 = 0$.

Kết quả hồi qui có được như sau

Bảng 12.13

Regression Statistics	
Multiple R	0.8227
R Square	0.6769
Adjusted R Square	0.6526
Standard Error	151.0825
Observations	44

Bảng 12.14

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	1912583	637527.6	27.92996	6.585E-10
Residual	40	913037.5	22825.94		
Total	43	2825620			

Bảng 12.15

	Coefficients	Standard			Upper	
		Error	t Stat	P-value	Lower 95%	95%
Intercept	6447.418	139.545	46.203	0.000	6165.386	6729.450
X	-0.029	0.004	-7.576	0.000	-0.037	-0.021
D1	34.865	55.681	0.626	0.535	-77.669	147.400
D2	147.991	64.096	2.309	0.028	18.449	277.533

Mô hình hồi qui tuyến tính mẫu được viết lại như sau:

$$\hat{Y} = 6447,418 - 0,029X + 34,865D_1 + 147,991D_2$$

- Hệ số hồi qui đứng trước biến X mang dấu âm có nghĩa là nếu đồng hồ cây số cho thấy xe đã được đi nhiều hơn 1 km thì giá bán xe sẽ giảm đi trung bình khoảng 0,029 đô la (hay 2,9 cent) chưa xét đến màu của xe, điều này hoàn toàn hợp lý.
- Hệ số hồi qui đứng trước biến D_1 và biến D_2 cho thấy với số km đã được sử dụng của các xe là như nhau thì nếu xe có màu trắng nó sẽ được trả nhiều hơn xe màu khác 34,865 đô la, còn nếu xe màu bạc thì nó được trả thêm 147,991 đô la so với xe màu khác.

Tóm lại:

- Nếu xe có màu không phải màu trắng cũng không phải màu bạc thì phương trình hồi qui dự đoán giá bán xe sẽ như sau:
$$\hat{Y} = 6447,418 - 0,029X$$
- Nếu xe có màu trắng tức là $D_1 = 1$ và $D_2 = 0$ thì phương trình hồi qui mô tả giá bán xe sẽ có dạng:
$$\hat{Y} = 6482,283 - 0,029X$$
- Nếu xe có màu bạc tức là $D_1 = 0$ và $D_2 = 1$ thì phương trình hồi qui mô tả giá bán xe sẽ có dạng
$$\hat{Y} = 6595,409 - 0,029X$$

Ta thấy khả năng giải thích của mô hình chấp nhận được, tuy nhiên căn cứ trên giá trị p-value ở Bảng 12.15 thì biến D_1 tỏ ra không có ý nghĩa trong việc giải thích cho giá bán xe (đã kiểm tra và thấy rằng không có khả năng xảy ra da cộng tuyển giữa các biến giải thích) do đó về mặt tổng thể, ta không thể suy diễn được rằng xe có màu trắng sẽ có giá bán cao hơn xe có màu khác; ngược lại, ta có thể kết luận rằng nếu xe có màu bạc, nó sẽ được bán với giá cao hơn xe có các màu khác trung bình khoảng 147 đô la nếu các điều kiện khác là như nhau.

12.4 LIÊN HỆ PHI TUYẾN

Ở nội dung hồi qui, cho đến lúc này chúng ta đã làm việc trên giả định rằng mối liên hệ giữa biến phụ thuộc Y và biến giải thích X là tuyến tính, tuy nhiên trong thực tế có nhiều tình huống liên hệ giữa Y và X không phải là tuyến tính mà là một dạng đường cong nào đó, ví dụ nhu cầu sử dụng điện sẽ gia tăng theo dạng hàm mũ trong mối liên hệ với sự tăng trưởng dân số chứ không phải một sự gia tăng tuyến tính thông thường, hay sự gia tăng nỗ lực trong đánh bắt cá sẽ có mối liên hệ dạng hàm Parabol lồi với sản lượng cá đánh bắt được nhằm mô tả một thực tế là trong quá trình gia tăng nỗ lực đánh bắt liên tục, nếu sự gia tăng này vượt qua ngưỡng cho phép thì sản lượng thu được bắt đầu giảm dần.

Trong những tình huống mà việc sử dụng mô hình hồi qui cho liên hệ phi tuyến là cần thiết thì chúng ta phải tuân thủ, tuy nhiên rõ ràng chúng ta cảm nhận là hồi qui cho liên hệ phi tuyến sẽ phức tạp hơn nhiều so với liên hệ tuyến tính. Nếu bạn là người thực hiện nghiên cứu để phục vụ cho mục tiêu ra quyết định của cấp trên, rõ ràng họ sẽ dễ quyết định hơn nếu vấn đề đơn giản dễ hiểu và sự phức tạp sẽ khiến họ lưỡng lự. Nguyên tắc vừa đủ trong thống kê cũng khuyên bạn nên giải quyết vấn đề bằng cách đơn giản nhất có thể cho đến khi sự đơn giản đó tỏ ra không còn phù hợp.

12.4.1 Dạng hàm bậc 2

Một trong những kiểu liên hệ phi tuyến phổ biến là mối liên hệ bậc 2 giữa 2 biến, mối liên hệ này giữa X và Y có thể được phân tích bằng mô hình hồi qui bậc 2 có công thức định nghĩa như sau:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

β_0 là hệ số chặn

β_1 là hệ số của ảnh hưởng tuyến tính của X lên Y

β_2 là hệ số của ảnh hưởng bậc 2 của X lên Y

ε_i là sai số tương ứng với mỗi quan sát

Ta thấy rằng mô hình hồi qui bậc 2 này tương tự như một mô hình hồi qui bội trong đó có hai biến giải thích, biến thứ nhất đứng sau hệ số β_1 xem như là biến X_1 , và biến thứ hai đứng sau hệ số β_2 xem như là biến X_2 (mà được tạo ra bằng cách lấy bình phương biến X)

Ví dụ: Một nhà nghiên cứu cho một công ty xăng dầu muốn xây dựng một mô hình để dự đoán lượng xăng xe hơi tiêu thụ (tính bằng số dặm đi được trên mỗi gallon) căn cứ vào tốc độ xe chạy trong điều kiện đường cao tốc.

Một cuộc thí nghiệm được tiến hành với 28 chiếc xe, các xe tham gia được cho chạy trên đường cao tốc với các mức tốc độ khác nhau, biến thiên trong khoảng từ 10 dặm/giờ đến 75 dặm/giờ. Dữ liệu về tốc độ và lượng xăng xe tiêu thụ được ghi nhận lại như sau:

Bảng 12.16

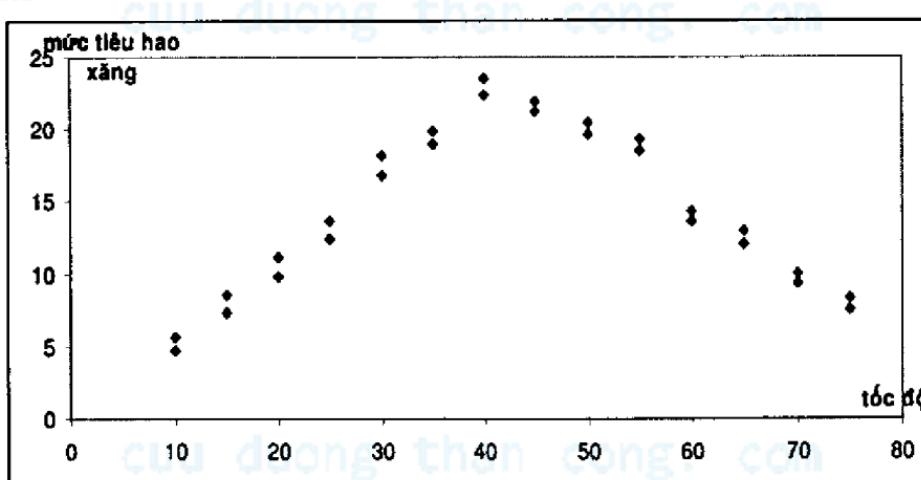
STT	Mức tiêu hao xăng (Dặm/gallon)	Tốc độ (dặm/giờ)	STT	Mức tiêu hao xăng (Dặm/gallon)	Tốc độ (dặm/giờ)
1	21,3	45	15	4,8	10
2	22,0	45	16	5,7	10
3	20,5	50	17	8,6	15
4	19,7	50	18	7,3	15
5	18,6	55	19	9,8	20
6	19,3	55	20	11,2	20
7	14,4	60	21	13,7	25
8	13,7	60	22	12,4	25
9	12,1	65	23	18,2	30
10	13,0	65	24	16,8	30
11	10,1	70	25	19,9	35
12	9,4	70	26	19,0	35
13	8,4	75	27	22,4	40
14	7,6	75	28	23,5	40

Nhằm giúp cho việc lựa chọn mô hình phù hợp mô tả mối liên hệ giữa biến X đại diện cho tốc độ của xe và biến Y đại diện cho mức tiêu hao xăng ta vẽ đồ thị phân tán như Hình 12.9. Nhận định ban đầu cho ta thấy khi tốc độ của xe càng lớn dần, lượng xăng tiêu hao sẽ giảm dần (thể hiện ở số dặm đi được trên mỗi gallon xăng nhiều hơn), nhưng tới một giới hạn nào đó về tốc độ thì lượng xăng tiêu hao lại bắt đầu tăng lên khi xe chạy càng nhanh. Điều đó cho thấy mô hình bậc hai có vẻ phù hợp với mối liên hệ giữa tốc độ và lượng xăng tiêu hao hơn là mô hình tuyến tính. Chúng ta sẽ sử dụng dữ liệu trên mẫu để ước lượng mô hình hồi qui mẫu với dạng như sau:

$$\hat{Y} = b_0 + b_1 X + b_2 X^2$$

Để ước lượng các hệ số hồi qui mẫu chúng ta cũng sử dụng phương pháp bình phương bé nhất thông thường với qui ước chúng ta đổi xử với biến độc lập X^2 như một biến độc lập thứ 2 trong mô hình hồi qui tuyến tính có hai biến giải thích. Tất nhiên muốn chạy được mô hình bằng Excel thì đầu tiên chúng ta phải tạo ra dữ liệu về biến X^2 bằng cách rất đơn giản là bình phương biến “Tốc độ”.

Hình 12.9



12.4.1.1 Kết quả chạy hồi qui trên Excel

Từ bộ dữ liệu trong Bảng 12.16, sử dụng phần mềm Excel ta thu được kết quả như sau:

Bảng 12.17

Regression Statistics	
Multiple R	0.9585
R Square	0.9188
Adjusted R Square	0.9123
Standard Error	1.6634
Observations	28

Bảng 12.18

ANOVA

	df	SS	MS	F	Significance F
Regression	2	782.8247	391.4123	141.4596	2.33771E-14
Residual	25	69.17387	2.766955		
Total	27	851.9986			

Bảng 12.19

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-7.5555	1.4241	-5.3055	0.0000	-10.4886	-4.6225
X	1.2717	0.0757	16.7920	0.0000	1.1157	1.4277
X ²	-0.0145	0.0009	-16.6325	0.0000	-0.0163	-0.0127

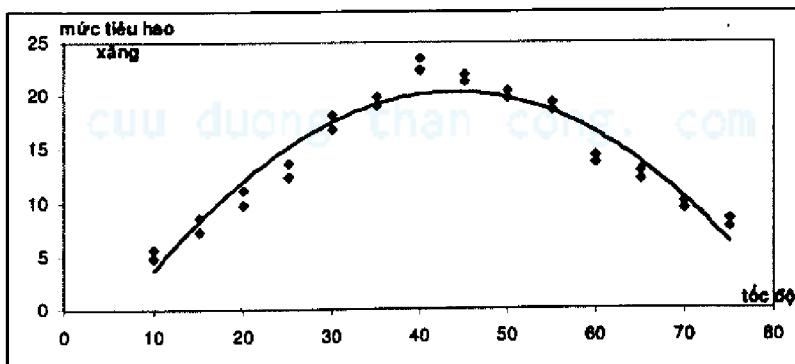
12.4.1.2 Phương trình hồi qui tuyến tính mẫu

Từ các hệ số hồi qui mẫu tính được $b_0 = -7,5555$; $b_1 = 1,2717$; $b_2 = -0,0145$, ta xây dựng mô hình hồi qui mẫu là:

$$\hat{Y} = -7,5555 + 1,2717 * X - 0,0145 * X^2$$

Phương trình bậc hai này được thể hiện về mặt hình ảnh như sau:

Hình 12.10



Từ phương trình này hệ số chặn b_0 được diễn tả như là mức tiêu hao xăng trung bình khi tốc độ của xe bằng 0, như chúng ta đã thảo luận ở các nội dung trước, cách diễn đạt như thế về hệ số chặn gần như không có ý nghĩa vì 0 nằm ngoài khoảng biến thiên của các giá trị mà biến độc lập có thể nhận.

Để diễn dịch ý nghĩa của hệ số b_1 và b_2 ta xem xu thế biến thiên của đường hồi qui, sau xu thế giảm lúc đầu thì mức tiêu hao xăng dần tăng lên khi tốc độ xe tăng. Mỗi liên hệ phi tuyến này có thể được minh họa rõ hơn bằng cách dự đoán lượng tiêu thụ xăng trung bình với các mức tốc độ 28, 48, 68 dặm/giờ.

Với $X = 28$: $\hat{Y} = -7,5555 + 1,2717 \times 28 - 0,0145 \times 28^2 = 16,6841$

Với $X = 48$: $\hat{Y} = -7,5555 + 1,2717 \times 48 - 0,0145 \times 48^2 = 20,0781$

Với $X = 68$: $\hat{Y} = -7,5555 + 1,2717 \times 68 - 0,0145 \times 68^2 = 11,8721$

Như vậy mức tiêu hao xăng được dự đoán là 16,6841 dặm/gallon khi tốc độ là 28 dặm/giờ, khi tốc độ tăng lên 48 dặm/giờ thì mức tiêu hao xăng giảm còn 20,0781 dặm/galllon nhưng khi tốc độ xe là 68 dặm/giờ thì xe còn hao xăng hơn cả mức 28 dặm/giờ.

12.4.1.3 Đánh giá độ phù hợp của mô hình

Giá trị R^2 hiệu chỉnh = 0,9123 cho thấy 91,23% biến thiên của mức tiêu thụ xăng có thể được giải thích bởi mối liên hệ bậc hai giữa mức tiêu hao xăng và tốc độ xe, mô hình ta xây dựng có độ phù hợp cao. Bạn cũng nên sử dụng kết hợp với thông tin về hệ số xác định bội hiệu chỉnh. Kiểm định về ý nghĩa toàn diện của mô hình cũng được tiến hành trên cơ sở kiểm định F với giả thuyết là:

$$H_0 : R^2 = 0$$

$$H_1 : R^2 \neq 0$$

Công thức tính giá trị F vẫn là công thức bạn đã nghiên cứu ở phần Đánh giá ý nghĩa toàn diện của mô hình, bạn thử dùng số liệu về SSR, SSE trên bảng ANOVA và thay thế các giá trị $k = 2$, $n = 28$ vào công thức tính F xem thử có đúng $F = 141,4596$ hay không.

Nếu bạn chọn dùng phương pháp so sánh mức ý nghĩa, vì giá trị Significance F trên bảng ANOVA là rất nhỏ nên ngay cả với một độ tin cậy cao tới 99% ta cũng vẫn có thể an toàn để bác bỏ giả thuyết Ho và kết luận là về mặt tổng thể có một mối liên hệ dạng hàm bậc 2 có ý nghĩa giữa mức tiêu hao xăng và tốc độ.

12.4.1.4 Đánh giá tác động bậc 2

Khi bạn sử dụng mô hình hồi qui để khảo sát mối liên hệ giữa 2 biến thì nhiệm vụ không chỉ là tìm ra một mô hình phù hợp nhất về mặt thống kê mà bạn phải chọn được mô hình càng đơn giản càng tốt, ở đây bạn sẽ cần kiểm tra xem hiệu ứng bậc 2 thực ra có cần thiết hay không, đơn giản chỉ bằng cách kiểm định ý nghĩa thống kê của hệ số hồi qui đứng trước biến X^2 .

Giả thuyết đặt ra cho kiểm định này như sau:

$H_0: \beta_2 = 0$ (tức là việc bao hàm tác động bậc hai không có tác dụng cải thiện mô hình một cách có ý nghĩa).

$H_1: \beta_2 \neq 0$ (tức là việc bao hàm tác động bậc hai có tác dụng cải thiện mô hình một cách có ý nghĩa).

Kiểm định t cũng được thực hiện như cách thông thường, để tìm hiểu ý nghĩa của hệ số hồi qui đứng trước biến X^2 , ta sử dụng giá trị p-value = 0,000 rất nhỏ nên ta an toàn để bác bỏ giả thuyết H_0 với độ tin cậy cao đến 99 %. Ta có thể kết luận rằng mô hình bậc hai có ý nghĩa hơn mô hình bậc 1 trong việc mô tả mối liên hệ giữa mức tiêu hao xăng và tốc độ xe.

12.4.2 Dạng log kép

Sự biến đổi logarit các biến trong mô hình hồi qui cũng là một tình huống hay gặp, có khi người ta sử dụng phương pháp này để khắc phục hiện tượng phương sai thay đổi, có khi nó được sử dụng vì lý do trong mô hình dạng log kép các hệ số hồi qui có một ý nghĩa đặc biệt là nó đo độ co giãn của Y theo X tức là tỷ lệ phần trăm thay đổi của Y với tỷ lệ một phần trăm thay đổi của X. Ta xem xét mô hình tổng thể như sau

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \varepsilon_i$$

Giả sử ta muốn ước lượng nhu cầu tiêu thụ thịt gà căn cứ trên thông tin về thu nhập bình quân đầu người và giá bán lẻ thịt gà trên thị trường. Với dữ liệu mẫu thu thập được về nhu cầu tiêu thụ thịt gà ký hiệu Y (đơn vị tính là kg), thu nhập khả dụng bình quân đầu người ký hiệu X_1 (ngàn đồng), giá bán lẻ thịt gà ký hiệu X_2 (ngàn/kg) người ta tiến hành biến đổi các biến từ dữ liệu gốc thành dạng Logarit cơ số e rồi tiến hành chạy mô hình hồi qui trên những dữ liệu này, từ kết quả hồi qui nhận được mô hình hồi qui mẫu được viết như sau:

$$\hat{\ln(Y)} = 2,0328 + 0,4515 \ln(X_1) - 0,37221 \ln(X_2)$$

Từ kết quả trên bạn có thể phát biểu rằng:

- Hệ số co giãn của cầu thịt gà theo thu nhập là 0,45 nghĩa là với 1% gia tăng trong thu nhập, nhu cầu về thịt gà bình quân tăng 0,45% nếu giá thịt gà không đổi.
- Hệ số co giãn của cầu thịt gà theo giá thịt gà cho thấy với 1% gia tăng trong giá cả thịt gà, nhu cầu thịt gà bình quân giảm đi 0,37% nếu thu nhập giữ nguyên; dấu của hệ số hồi qui là cơ sở cho bạn có thể phát biểu như vậy.

Bạn cũng sử dụng các thông tin về hệ số xác định bội, kiểm định F và kiểm định t như cách quen thuộc trong mô hình hồi qui bội bất kỳ, cùng với các kiểm tra để đánh giá về sự đáp ứng các giả định đã biết. Cách đơn giản để không thấy vấn đề này rắc rối là bạn hãy xem các biến đã được lấy Logarit cơ số e như một biến số bất kỳ, ví dụ bạn gọi chúng dưới một cái tên khác.

$$\ln(Y) = Y'$$

$$\ln(X_1) = X_1'$$

$$\ln(X_2) = X_2'$$

Từ đó bạn xem như mình làm việc với mô hình hồi qui bội có dạng $Y' = \beta_0 + \beta_1 X_1' + \beta_2 X_2' + \varepsilon_i$, Dĩ nhiên bạn chỉ cần chú ý một chút khi diễn dịch ý nghĩa của hệ số hồi qui, vì nhớ rằng nó là hệ số co giãn của Y theo X.

cuu duong than cong. com

cuu duong than cong. com

CHƯƠNG 13

CHỈ SỐ

13.1 MỘT SỐ VẤN ĐỀ CHUNG VỀ PHƯƠNG PHÁP CHỈ SỐ

13.1 Khái niệm chỉ số

Chỉ số trong thống kê là số tương đối biểu hiện quan hệ so sánh giữa các mức độ của một chỉ tiêu hay hiện tượng kinh tế - xã hội. Cụ thể chỉ số được tính bằng cách so sánh hai mức độ của hiện tượng ở hai thời gian hoặc hai không gian khác nhau nhằm biểu hiện mức độ biến động của chỉ tiêu hay hiện tượng qua thời gian hoặc không gian. Phương pháp chỉ số ngày càng được sử dụng rộng rãi trong đời sống, ví dụ chỉ số giá chứng khoán VN-Index là một loại chỉ số được quan tâm từng ngày ở nước ta hiện nay. Một điều cần lưu ý là chỉ số thể hiện quan hệ so sánh nên cần xác định rõ gốc so sánh khi sử dụng chỉ số.

13.2 Phân loại chỉ số

Có một số hình thức phân loại chỉ số chủ yếu như sau:

- Nếu căn cứ theo phạm vi tính toán của chỉ số người ta chia thành chỉ số cá thể và chỉ số tổng hợp (bạn đọc xem mục Chỉ số cá thể và chỉ số tổng hợp)
- Nếu căn cứ vào tính chất của chỉ tiêu được nghiên cứu người ta chia thành hai loại là chỉ số chỉ tiêu chất lượng và chỉ số chỉ tiêu khối lượng (bạn đọc xem mục Chỉ số chỉ tiêu chất lượng và chỉ số chỉ tiêu khối lượng)
- Nếu căn cứ gốc so sánh người ta chia thành chỉ số liên hoàn và chỉ số định gốc
- Nếu căn cứ vào hình thức biểu biện người ta chia thành chỉ số ở dạng cơ bản và chỉ số ở dạng biến đổi (bạn đọc xem chỉ số tổng hợp và chỉ số bình quân thuộc mục Hệ thống chỉ số)
- Ngoài ra còn có dạng chỉ số không gian để so sánh sự khác biệt của hiện tượng qua không gian.

13.2 CHỈ SỐ CÁ THỂ

Là loại chỉ số đơn giản nhất, nó thể hiện sự biến động của từng phần tử, từng đơn vị cá biệt trong một tổng thể phức tạp. Về cơ bản thì chỉ số cá thể chính là số tương đối.

13.2.1 Chỉ số cá thể giá cả

Bạn muốn đánh giá sự thay đổi giá cả của một loại hàng hóa cụ thể, và biết giá bán mặt hàng này tại kỳ nghiên cứu là p_1 và tại kỳ gốc là p_0 thì công thức của chỉ số cá thể giá cả của mặt hàng này được tính theo công thức

$$i_p = \frac{p_1}{p_0} \times 100\%$$

Ví dụ, giá bán một chai dầu ăn hiệu T.A dung tích 1 lít vào thời điểm tháng 1 năm 2000 là 11.500 đồng. Vào thời điểm tháng 1 năm 2007 giá của loại sản phẩm này là 19.500 đồng. Vậy chỉ số cá thể giá cả của dầu ăn được tính vào thời điểm tháng 1 năm 2007 (chính là kỳ nghiên cứu) như sau

$$i_p = \frac{19500}{11500} \times 100\% = 169,57\%$$

Ở đây kỳ gốc nghiên cứu là tháng 1 năm 2000, chú ý là chỉ số giá này không phản ánh mức giá mà đo lường mức độ biến động của giá giữa hai khoảng thời gian, nghĩa là so với thời điểm gốc tháng 1/2000 thì giá 1 chai dầu ăn T.A loại 1 lít tại tháng 1/2007 đã tăng thêm 69,57%, còn về số tuyệt đối nó đã tăng $19.500 - 11.500 = 8.000$ đồng.

13.2.2 Chỉ số cá thể khối lượng

Gọi q_1 và q_0 lần lượt là khối lượng của một loại sản phẩm hay hàng hóa cụ thể được sản xuất hoặc tiêu thụ tại kỳ muôn nghiên cứu và kỳ gốc. Chỉ số cá thể khối lượng được tính theo công thức sau

$$i_q = \frac{q_1}{q_0} \times 100\%$$

Ví dụ cùng một mặt hàng A, trong tháng 1 công ty tiêu thụ được 20.000 tấn, sang tháng 2 lượng tiêu thụ tăng hơn tháng 1 là 2.500 tấn, như vậy chỉ số cá thể khối lượng tại kỳ nghiên cứu là tháng 2 được tính căn cứ trên thông tin của kỳ gốc là tháng 1 như sau

$$i_q = \frac{20.000 + 2.500}{20.000} \times 100\% = 112,5\%$$

Vậy khối lượng sản phẩm tiêu thụ tháng 2 tăng 2.500 tấn, tương đương tăng 12,5% so với tháng 1.

Nói chung chỉ số cá thể tính nhanh và đơn giản nhưng rất hạn chế ở chỗ chỉ phản ánh biến động riêng của từng phần tử mà không cho phép ta nghiên cứu biến động chung của các phần tử trong một tổng thể gồm

nhiều phần tử không thể trực tiếp cộng với nhau để so sánh, như giá cả chẵng hạn. Ví dụ nếu một cửa hàng tiêu thụ đồng thời 3 loại mặt hàng khác nhau về đơn vị tính và giá trị sử dụng như: bột ngọt (tính bằng kg), dầu ăn (tính bằng lít) và trứng (tính bằng chục) thì chỉ số cá thể chỉ cho phép tính toán tốc độ phát triển của riêng từng mặt hàng chứ không cho phép cộng trực tiếp 3 mặt hàng đó lại với nhau nhằm xác định tốc độ phát triển chung của 3 loại mặt hàng này do khác đơn vị tính. Chỉ số tổng hợp là loại số tương đối được phát triển để khắc phục hạn chế trên.

13.3 CHỈ SỐ TỔNG HỢP

Trong phương pháp chỉ số tổng hợp, biểu hiện của các phần tử trong hiện tượng phức tạp được chuyển về dạng đồng nhất để có thể cộng trực tiếp với nhau, dựa trên cơ sở mối quan hệ giữa yếu tố nghiên cứu với yếu tố khác. Ví dụ như khối lượng các sản phẩm khác loại vốn không thể cộng trực tiếp với nhau do khác đơn vị tính khi được chuyển sang dạng giá trị, bằng cách nhân với yếu tố giá cả, thì có thể cộng được với nhau.

Chỉ số tổng hợp là loại số tương đối được sử dụng để đánh giá sự thay đổi của một số hoặc tất cả các phần tử thuộc tổng thể nghiên cứu. Khi nghiên cứu chỉ số tổng hợp có một khái niệm quan trọng là quyền số, nó là yếu tố được chọn để giúp chuyển các phần tử không thể cộng trực tiếp với nhau thành một dạng chung có thể cộng được, bên cạnh đó nó còn có công dụng thể hiện vai trò mạnh yếu của từng phần tử trong toàn bộ tổng thể. Quyền số có thể được chọn ở các kỳ khác nhau (có khi là kì gốc, có khi là kì báo cáo, có khi là một kỳ nào đó phù hợp) là tùy theo mục đích nghiên cứu.

13.3.1 Chỉ số tổng hợp giá cả

Khi tính chỉ số tổng hợp giá cả người ta hay chọn quyền số là lượng hàng hóa tiêu thụ, với một số tài liệu bạn đọc sẽ thấy đôi khi quyền số được gọi là trọng số. Liên quan đến kì chọn quyền số người ta có hai loại chỉ số giá cả là chỉ số Laspeyres và chỉ số Paasche.

13.3.1.1 Chỉ số Laspeyres

Nếu quyền số được chọn là khối lượng sản phẩm hàng hóa tiêu thụ ở kì gốc người ta có chỉ số tổng hợp giá cả theo phương pháp Laspeyres có công thức:

$$I_p = \frac{\sum_{i=1}^n p_i q_{i(0)}}{\sum_{i=1}^n p_{i(0)} q_{i(0)}} \times 100\%$$

Trong đó p_i và q_i là giá cả và khối lượng tiêu thụ của loại hàng hóa thứ i trong tổng thể gồm n phần tử, số 1 và số 0 trong ngoặc ám chỉ thông tin của kì nghiên cứu và kì gốc. Như vậy $q_{i(0)}$ là khối lượng tiêu thụ của mặt hàng i tại kì gốc. Để đơn giản thì người ta hay viết công thức này ngắn gọn như sau, trong đó dấu Σ dù ám chỉ rằng thành phần $p_i q_0$ hay $p_0 q_0$ không phải chỉ mô tả một phần tử mà tạo nên từ nhiều phần tử.

$$I_p = \frac{\sum p_i q_0}{\sum p_0 q_0} \times 100\%$$

13.3.1.2 Chỉ số Paasche

Nếu quyền số được chọn là lượng hàng hóa tiêu thụ tại kì báo cáo chỉ số tổng hợp giá cả theo phương pháp Paasche được thiết lập công thức như sau

$$I_p = \frac{\sum p_i q_i}{\sum p_0 q_i} \times 100\%$$

Để minh họa cách tính chỉ số giá Laspeyres và Paasche ta xem ví dụ sau.

Chúng ta có bảng sau liệt kê giá cả và lượng hàng tiêu thụ tương ứng của một số mặt hàng tại cửa hàng A ở kì gốc 2000 và kì nghiên cứu 2005

Bảng 13.1

Tên hàng hóa	ĐVT	Giá (ngàn đồng)		Số lượng tiêu thụ (ngàn ĐVT)		Giá trị (triệu đồng)			
		Kì gốc (p_0)	Kì n/cứu (p_1)	Kì gốc (q_0)	Kì n/cứu (q_1)	$p_1 q_0$	$p_0 q_0$	$p_1 q_1$	$p_0 q_1$
X	Kg	5	6	10	13	60	50	78	65
Y	Lít	10	12,2	5	5,5	61	50	67,1	55
Z	Chục	8	10	0,25	0,32	2,5	2	3,2	2,56
Tổng						123,5	102	148,3	122,56

Vận dụng công thức tính toán các chỉ số, chúng ta lập cột tính toán các con số tổng cộng thành phần rồi thay thế vào công thức tính:

- Chỉ số Laspeyres:

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100\% = \frac{123,5}{102} \times 100\% = 121,08\%$$

Kết quả này cho thấy giá của nhóm 3 mặt hàng ở năm 2005 so với 2000 tăng 21,08%.

- Chỉ số Paasche:

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100\% = \frac{148,3}{122,56} \times 100\% = 121\%$$

Kết quả này cho thấy giá của nhóm 3 mặt hàng ở năm 2005 so với năm 2000 đã tăng 21%.

Nhận xét chung: nếu bạn đọc muốn so sánh trong hai chỉ số giá cả Laspeyres và Paasche chỉ số nào có ý nghĩa hơn thì không có câu trả lời tuyệt đối. Với Laspeyres vì quyền số được cố định ở kì gốc nên thông tin đã có sẵn, sự sẵn có này đảm bảo kết quả tính toán về chỉ số giá cả được cung cấp nhanh chóng, ngoài ra một cách trực quan người tiếp nhận thông tin chủ động hơn trong nhận biết chỉ số giá vì đã nắm được vai trò của từng yếu tố trong tổng thể qua các quyền số quen dùng. Tuy nhiên Paasche lại theo sát kì nghiên cứu nên bảo đảm phản ánh được những thay đổi trong xu hướng tiêu dùng, bởi vì rất có thể một mặt hàng vào kì chọn làm gốc có thể được ưa chuộng và tiêu dùng nhiều nhưng ngày nay đã trở nên không còn quan trọng. Với phạm vi các doanh nghiệp khi mà số lượng các mặt hàng cần xác định để làm quyền số không quá nhiều và phức tạp cộng với sự phát triển của công nghệ thông tin thì ở một mức độ nào đó việc thu thập nhanh chóng thông tin về q_1 cũng trở nên dễ dàng hơn.

13.3.1.3 Chỉ số Fisher

Để trung hòa các vấn đề trên người ta có thể chọn quyền số kết hợp cả thông tin của kỳ báo cáo và kỳ gốc, lúc này ta có chỉ số tổng hợp giá cả theo phương pháp Fisher là trung bình nhân của cả hai chỉ số trên.

Công thức tính

$$I_p = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

Theo số liệu đã có, áp dụng công thức trên ta tính được chỉ số Fisher

$$I_p = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} = \sqrt{\frac{123,5}{102} \times \frac{148,3}{122,56}} = 1,2104 \text{ hoặc } 121,04\%$$

Trong nhiều trường hợp khi tính toán với các quyền số cố định ở các thời kì khác nhau theo phương pháp Laspeyres hoặc Paasche dẫn đến các kết quả quá chênh lệch thì việc sử dụng chỉ số Fisher là cần thiết, tuy nhiên bạn mất nhiều công sức hơn.

13.3.2 Chỉ số tổng hợp khối lượng

Khi cần nghiên cứu sự thay đổi khối lượng sản phẩm của một nhóm nhiều loại sản phẩm người ta dùng loại chỉ số tổng hợp khối lượng với ý tưởng xây dựng công thức của chỉ số này y hệt phương pháp chỉ số tổng hợp giá cả, tuy nhiên lúc này yếu tố giá đóng vai trò trọng số.

Ta có công thức tính chỉ số tổng hợp khối lượng theo phương pháp Laspeyres

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100\%$$

Ta có công thức tính chỉ số tổng hợp khối lượng theo phương pháp Paasche

$$I_q = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100\%$$

Ta có công thức tính chỉ số tổng hợp khối lượng theo phương pháp Fisher

$$I_q = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

Vận dụng tiếp ví dụ ở trên để tính chỉ số khối lượng sản phẩm tiêu thụ

Bảng 13.2

Tên hàng hóa	ĐVT	Giá (ngàn đồng)		Số lượng tiêu thụ (ngàn ĐVT)		Giá trị (triệu đồng)			
		Kì gốc (p_0)	Kì n/cứu (p_1)	Kì gốc (q_0)	Kì n/cứu (q_1)	$q_0 p_1$	$q_0 p_0$	$q_1 p_1$	$q_1 p_0$
X	Kg	5	6	10	13	60	50	78	65
Y	Lít	10	12,2	5	5,5	61	50	67,1	55
Z	Chục	8	10	0,25	0,32	2,5	2	3,2	2,56
Tổng						123,5	102	148,3	122,56

Lần lượt tính các chỉ số tổng hợp khối lượng:

- Theo phương pháp Laspeyres

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100\% = \frac{122,56}{102} \times 100\% = 120,16\%$$

- Theo phương pháp Paasche

$$I_q = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100\% = \frac{148,3}{123,5} \times 100\% = 120,08\%$$

- Theo phương pháp Fisher

$$I_q = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \sqrt{120,16 * 120,08} = 120,12\%$$

Kết quả tính chỉ số theo phương pháp Fisher cho thấy khối lượng tiêu thụ của nhóm 3 mặt hàng ở năm 2005 tăng 12% so với năm 2000.

13.3.3 Chỉ số của chỉ tiêu chất lượng và chỉ số của chỉ tiêu khối lượng

Chỉ số chỉ tiêu chất lượng là loại chỉ số dùng để nghiên cứu sự thay đổi của các chỉ tiêu chất lượng ví dụ như chỉ số giá thành sản phẩm, chỉ số giá cả tiêu dùng... Theo phương pháp phân chia này thì chỉ số tổng hợp giá cả theo phương pháp Laspeyres hay Paasche cũng đều là chỉ số chỉ tiêu chất lượng...

Chỉ số chỉ tiêu khối lượng là loại chỉ số dùng để nghiên cứu sự thay đổi của các chỉ tiêu khối lượng ví dụ như chỉ số khối lượng sản phẩm sản xuất, chỉ số khối lượng hàng hóa tiêu thụ...

Tuy nhiên nhiều khi sự phân chia thành chỉ tiêu chất lượng và chỉ tiêu khối lượng chỉ có ý nghĩa tương đối.

13.4 CHỈ SỐ LIÊN HOÀN VÀ CHỈ SỐ ĐỊNH GỐC

13.4.1 Chỉ số liên hoàn

Khi tính chỉ số cho các thời kì liên tiếp nhau người ta có chỉ số liên hoàn trong đó mỗi chỉ số đều so sánh đối tượng ở thời kì nghiên cứu với thời kì liền kề trước đó.

13.4.2 Chỉ số định gốc

Chỉ số định gốc là chỉ số tính cho nhiều thời kì khác nhau so với một thời kì được chọn làm gốc cố định để so sánh đối tượng ở các thời kì khác nhau với một cái gốc giống nhau giúp cho quá trình nhận định tình hình tăng giảm của đối tượng loại trừ được ảnh hưởng của các yếu tố ngoại cảnh.

Ta có số liệu về giá bán lẻ trung bình của một cân măng cụt tại thành phố X trong thời gian từ năm 1995 đến 2002 được liệt kê thành bảng như sau:

Bảng 13.3

Năm	Giá (VNĐ)	Chỉ số cá thể giá cả liên hoàn (%)	Chỉ số cá thể giá cả định gốc (%)
1995	13230	-	91,68
1996	11030	83,37	76,44
1997	12130	109,97	84,06
1998	14520	119,70	100,62
1999	19040	131,13	131,95
2000	14430	75,79	100,00
2001	14140	97,99	97,99
2002	14510	102,62	100,55

Ta muốn tính chỉ số liên hoàn cho giá măng cụt trung bình thì ta áp dụng ngay công thức của chỉ số cá thể giá cả đã nghiên cứu ở nội dung trên là

$$i_p = \frac{P_1}{P_0} \times 100\%$$

Trong đó kì gốc được chọn thay đổi liên tục theo kì nghiên cứu với qui tắc kì gốc nằm liền kề với kì nghiên cứu. Giả dụ muốn tính chỉ số giá cả theo kiểu liên hoàn tại năm 1997 ta lấy giá măng cụt năm 1997 chia cho giá măng cụt năm 1996 (kì gốc lúc này là 1996) rồi đem kết quả nhân 100: $12130/11030 * 100 = 109,97\%$. Tiếp tục, muốn tính chỉ số này cho năm 1998 ta lấy giá măng cụt năm 1998 chia cho giá măng cụt năm 1997 (kì gốc lúc này là 1997) rồi nhân 100: $14520/12130 * 100 = 119,70\%$.

Nếu theo qui tắc này thì chỉ số cá thể giá cả liên hoàn tại năm 1995 không tính được vì không có thông tin năm 1994

Khi muốn tính chỉ số định gốc cho giá măng cụt chúng ta cũng vẫn áp dụng công thức $i_p = \frac{P_1}{P_0} * 100$ nhưng trước tiên chúng ta phải chọn một cái gốc cố định, giả dụ ở đây ta chọn gốc là năm 2000. Lúc đó ta có chỉ số giá cả thời kì gốc = 100% (chính là chỉ số giá cả định gốc tính tại năm 2000 bằng cách chia giá măng cụt năm 2000 cho gốc chính là nó). Chỉ số giá cả định gốc cho năm 1995 được tính bằng cách lấy giá măng cụt trung bình tại năm 1995 chia cho giá măng cụt trung bình năm 2000 rồi đem kết quả nhân lên 100 được $13230/14430 * 100 = 91,68\%$. Kết quả này nói lên rằng giá măng cụt tại năm 1995 bằng 91,68% giá măng cụt tại năm cơ sở 2000, hay giá măng cụt năm 1995 thấp hơn giá năm 2000 là (100 -

$91,68\% = 8,32\%$). Tuy nhiên ta không nói ngược lại được là giá măng cụt năm 2000 cao $8,32\%$ so với giá năm 1995 mà thực sự giá măng cụt năm 2000 tăng so với năm 1995 là: $(14430 / 13230 \times 100 - 100) = 9,07\%$

Như vậy có thể thấy được một chỉ số cụ thể vừa có thể là chỉ số chỉ tiêu chất lượng (vì xem xét sự biến động của giá cả), vừa có thể là chỉ số cá thể (vì không có quyền số), vừa có thể là chỉ số liên hoàn nếu ta xét trên các cơ sở phân loại khác nhau.

Các loại chỉ số cá thể liên hoàn và định gốc vừa tính còn được gọi bằng cặp tên gọi khác trong thống kê là số tương đối động thái liên hoàn và số tương đối động thái định gốc; hoặc phổ biến hơn là tốc độ phát triển liên hoàn và tốc độ phát triển định gốc.

Vận dụng số trung bình nhân để tính tốc độ phát triển trung bình

Ở nội dung số trung bình nhân ở Chương 4 chúng ta biết rằng Số trung bình nhân được vận dụng để tính tốc độ phát triển trung bình của hiện tượng trong thời kì nghiên cứu khi các giá trị x_i là các con số tốc độ phát triển liên hoàn trong thời kì đó. Ý nghĩa của tốc độ phát triển trung bình thể hiện nhịp độ phát triển đại diện của hiện tượng trong suốt thời kì nghiên cứu. Nếu x_i là các tốc độ phát triển liên hoàn của kì thứ i ($i=1,2,\dots,n$) chú ý là khi tham gia tính số trung bình x , phải được xết theo đơn vị lần chữ không theo %. Thì công thức tính tốc độ phát triển trung bình $\bar{x} = \sqrt[n]{x_1 x_2 \dots x_n}$

Ví dụ có chỉ số giá bán lẻ gạo liên hoàn các tháng trong năm 2005 tại một thành phố như sau:

Bảng 13.4

Tháng	Chỉ số giá bán lẻ gạo
Tháng 1	101,1
Tháng 2	102,5
Tháng 3	100,1
Tháng 4	100,6
Tháng 5	100,5
Tháng 6	100,4
Tháng 7	100,4
Tháng 8	100,4
Tháng 9	100,8
Tháng 10	100,4
Tháng 11	100,4
Tháng 12	100,8

Muốn tính tốc độ tăng bình quân tháng của giá gạo qua các tháng trong năm 2005, trước tiên ta áp dụng công thức trung bình nhân trong đó x_i là các chỉ số liên hoàn của tháng thứ i ($i=1,2,\dots,12$) tính theo đơn vị lần

$$\bar{X} = \sqrt[12]{X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10} X_{11} X_{12}}$$

$$\bar{X} = \sqrt[12]{1,011 \times 1,025 \times 1,001 \times \dots \times 1,004 \times 1,004 \times 1,008} = \sqrt[12]{1,087} = 1,007$$

Trong năm 2005, tốc độ tăng trung bình của giá gạo hàng tháng là 0,007 lần hay 0,7%/tháng.

Cách tính số trung bình nhân bằng Excel

Bạn có thể dùng hàm Geomean trong menu Insert/Function của Excel để tính số trung bình nhân, lúc đó bạn không cần chuyển các số liên hoàn thành phần về đơn vị lần. Tham khảo cách thực hiện trong Hình 13.1

Hình 13.1

A	B	C	D	E
1 Tháng	Chỉ số giá bán lẻ gạo			
2 Tháng 1	101.1			
3 Tháng 2	102.5			=GEOMEAN(B2:B13)
4 Tháng 3	100.1			
11 Tháng 10	100.4			
12 Tháng 11	100.4			
13 Tháng 12	100.8			

Ví dụ thứ 2: xem xét một khoản đầu tư 100 triệu đồng mà một người bỏ vào kinh doanh chứng khoán. Vào cuối tháng đầu tiên, khoản tiền này giảm xuống còn có 50 triệu, sau đó nó lại tăng trở lại đúng số vốn gốc 100 triệu vào cuối tháng thứ 2. Hệ thống các thông tin này lại như sau

Thời gian	Đầu tháng 1	Cuối tháng 1	Cuối tháng 2
Tổng số tiền có (triệu đồng)	100	50	100
Tốc độ phát triển liên hoàn (lần)	NA	50/100 = 0,5	100/50 = 2

Vậy muốn biết mức độ sinh lời trung bình của khoản tiền đầu tư này trong khoảng thời gian 2 tháng (bỏ qua yếu tố mất giá của đồng tiền) ta sẽ tính mức bình quân của Tốc độ phát triển trong suốt khoảng thời gian trên bằng cách áp dụng công thức trung bình nhân như sau

$$\bar{x} = \sqrt[2]{x_1 x_2} = \sqrt[2]{0,5 \times 2} = 1 \text{ hay } 100\%$$

→ Tốc độ phát triển bình quân là 100% tức là mức độ sinh lợi bình quân là 0%.

Chú ý là nếu trong tình huống này ta áp dụng trung bình cộng để tính mức độ bình quân của Tốc độ phát triển thì kết quả lại khác. Hãy xem

$$\bar{x} = \frac{0,5 + 2}{2} = 1,25 \text{ hay } 125\% \text{ tức là khoản đầu tư trên đã sinh lợi}$$

25% chứ không phải là “đậm chân tại chỗ” như kết quả trung bình nhân chỉ ra. Rõ ràng trung bình nhân đã phản ánh chính xác hơn so với trung bình cộng về sự thay đổi trong giá trị của khoản đầu tư qua 2 tháng.

Lý do là vì Tốc độ phát triển bình quân được tính từ các con số tốc độ phát triển liên hoàn thành phần, mà các con số liên hoàn thì không có cùng gốc so sánh như con số định gốc, do đó việc đem cộng trực tiếp với nhau trong quá trình tính toán số trung bình cộng là không hợp lý, do đó số trung bình cộng đã cho kết quả không đúng, ngược lại, phải áp dụng số trung bình nhân (đem m thành phần nhân với nhau rồi lấy căn bậc m của đáp số) ta mới tìm được đúng con số Tốc độ phát triển bình quân cho toàn thời kì nghiên cứu.

Chú ý

Phương pháp tính chỉ số liên hoàn hay định gốc này không chỉ áp dụng cho tình huống chỉ số cá thể mà còn cho cả tình huống chỉ số tổng hợp.

Ví dụ muốn tính chỉ số khối lượng sản phẩm sản xuất liên hoàn cho ba tháng 2, 3, 4 với quyền số là giá cố định của sản phẩm được chọn theo giá một kí nào đó ổn định kí hiệu là p_s , ta thiết lập công thức như sau:

$$I_{q(3/2)} = \frac{\sum q_3 p_s}{\sum q_2 p_s} \text{ và } I_{q(4/3)} = \frac{\sum q_4 p_s}{\sum q_3 p_s}$$

Muốn tính chỉ số định gốc cho khối lượng sản phẩm sản xuất tháng 2, 3 so với tháng 1 với quyền số là giá cố định của sản phẩm được chọn theo giá một kí nào đó ổn định kí hiệu là p_s , ta thiết lập công thức như sau

$$I_{q(2/1)} = \frac{\sum q_2 p_s}{\sum q_1 p_s} \text{ và } I_{q(3/1)} = \frac{\sum q_3 p_s}{\sum q_1 p_s}$$

13.5 CHỈ SỐ KHÔNG GIAN (CHỈ SỐ ĐỊA PHƯƠNG)

Là loại chỉ số so sánh các hiện tượng cùng loại nhưng qua điều kiện không gian khác nhau ví dụ so sánh giá cả của một nhóm mặt hàng ở các chợ khác nhau trong địa phương hay so sánh khối lượng sản phẩm công nghiệp do hai tỉnh làm ra trong năm. Ta cũng có hai nhóm là chỉ số tổng hợp giá cả theo không gian và chỉ số tổng hợp khối lượng theo không gian.

13.5.1 Chỉ số tổng hợp giá cả theo không gian

Giả sử bạn có nhu cầu so sánh giá cả của cùng một nhóm mặt hàng tại hai chợ A và B chênh lệch ra sao. Lúc này chúng ta xây dựng một chỉ số tổng hợp mà quyền số được chọn là tổng khối lượng sản phẩm (hàng hóa) cùng loại trên cả hai chợ, nếu kí hiệu q_A tổng khối lượng tiêu thụ của một mặt hàng tại chợ A và q_B là tổng khối lượng tiêu thụ của một mặt hàng tại chợ B thì quyền số cho mặt hàng đó kí hiệu là $Q = q_A + q_B$

Công thức được xác định chỉ số:

$$I_p = \frac{\sum p_A Q}{\sum p_B Q} \times 100\%$$

13.5.2 Chỉ số tổng hợp khối lượng theo không gian

Khi tính chỉ số tổng hợp khối lượng tiêu thụ theo không gian tại hai chợ A và B ta cũng thiết lập công thức theo các quy tắc quen thuộc nhưng quyền số trong tình huống này là yếu tố giá cả có thể được chọn là giá cố định do Nhà nước ban hành (p_s) hoặc giá tính trung bình cho cả hai chợ (\bar{p}).

$$I_q = \frac{\sum q_A P_s}{\sum q_B P_s} \times 100\%$$

Với q_A tổng khối lượng tiêu thụ các mặt hàng tại chợ A và q_B là tổng khối lượng tiêu thụ các mặt hàng tại chợ B, p_s là giá cố định do Nhà nước ban hành.

Nếu dùng \bar{p} , gọi p_A là giá bán của mặt hàng tại chợ A và p_B là giá bán của mặt hàng tại chợ B, thì \bar{p} được tính cho mặt hàng đó như sau:

$$\bar{p} = \frac{p_A q_A + p_B q_B}{q_A + q_B}$$

Chỉ số tổng hợp lúc này trở thành

$$I_q = \frac{\sum q_A \bar{p}}{\sum q_B \bar{p}} \times 100\%$$

Tất cả các công thức này sẽ được làm rõ qua ví dụ sau.

Có số liệu về giá cả và khối lượng hàng tiêu thụ của 3 mặt hàng tại hai chợ A và B của địa phương trong cùng một kì như sau:

Bảng 13.5

Tên Hàng hóa	Chợ A		Chợ B		Q	\bar{P}	Giá trị (triệu đồng)			
	Giá đơn vị (ngàn đ/kg)	Lượng tiêu thụ (tấn)	Giá đơn vị (ngàn đ/kg)	Lượng tiêu thụ (tấn)			$p_A Q$	$p_B Q$	$q_A \bar{P}$	$q_B \bar{P}$
X	5	250	4,8	262	512	4,9	2560	2457,6	1225	1283,8
Y	4,6	430	4,9	392	822	4,7	3781,2	4027,8	2021	1842,4
Z	6,9	187	6,8	213	400	6,8	2760	2720	1271,6	1448,4
Tổng							9101,2	9205,4	4517,6	4574,6

Trước khi tính Chỉ số tổng hợp theo không gian ta tiến hành tính toán các số liệu cần thiết như sau:

Tính tổng khối lượng tiêu thụ trên từng mặt hàng $Q = q_A + q_B$

$$Q_X = q_A + q_B = 250 + 262 = 512$$

$$Q_Y = q_A + q_B = 430 + 392 = 822$$

$$Q_Z = q_A + q_B = 187 + 213 = 400$$

Các số liệu trên được điền vào cột có tiêu đề cột là Q ở Bảng 13.5 ở trên.

Tính giá trung bình từng mặt hàng tại cả hai chợ $\bar{p} = \frac{p_A q_A + p_B q_B}{q_A + q_B}$

$$\bar{p}_X = \frac{5 \times 250 + 4,8 \times 262}{250 + 262} = 4,9$$

$$\bar{p}_Y = \frac{4,6 \times 430 + 4,9 \times 392}{430 + 392} = 4,7$$

$$\bar{p}_Z = \frac{6,9 \times 187 + 6,8 \times 213}{187 + 213} = 6,8$$

Các số liệu trên được điền vào cột có tiêu đề cột là \bar{p} ở Bảng 13.5 trên.

Áp dụng công thức tính chỉ số giá cả không gian cho hai chợ A và B

$$I_p = \frac{\sum p_A Q}{\sum p_B Q} \times 100\% = \frac{9101,2}{9205,4} \times 100\% = 98,87\%$$

Như vậy nhìn chung giá của, ba mặt hàng trên tại chợ A chỉ bằng 98,87% chợ B, rẻ hơn khoảng 1,13%

Áp dụng công thức tính chỉ số khối lượng không gian cho hai chợ A và B

$$I_q = \frac{\sum q_A \bar{p}}{\sum q_B \bar{p}} \times 100\% = \frac{4517,6}{4574,6} \times 100\% = 98,75\%$$

Như vậy nhìn chung lượng tiêu thụ ba mặt hàng trên tại chợ A so với chợ B bằng 98,75%, hay nói cách khác là ít hơn 1,25%.

13.6 HỆ THỐNG CHỈ SỐ

Hệ thống chỉ số là dãy các chỉ số có liên hệ với nhau, hợp thành một đẳng thức nhất định. Có nhiều loại hệ thống chỉ số, trong thực tế công tác thống kê có 2 loại là hệ thống chỉ số tổng hợp và hệ thống chỉ số nghiên cứu biến động của chỉ tiêu bình quân.

13.6.1 Hệ thống chỉ số tổng hợp

Bên cạnh việc nghiên cứu sự thay đổi của hiện tượng qua thời gian và không gian phương pháp chỉ số còn có thể dùng để phân tích mức độ ảnh hưởng của các yếu tố đến sự thay đổi của một chỉ tiêu tổng hợp bằng cách kết hợp các chỉ số riêng lẻ lại thành hệ thống chỉ số, ví dụ như ta có mối quan hệ giữa giá cả và khối lượng tiêu thụ như sau:

$$\text{Giá đơn vị} \times \text{Khối lượng tiêu thụ} = \text{Giá trị mức tiêu thụ}$$

Thì ta phát triển được hệ thống chỉ số tổng hợp như sau:

$$\text{Chỉ số giá cả} \times \text{Chỉ số khối lượng} = \text{Chỉ số giá trị mức tiêu thụ}$$

Giả sử ta cần phân tích biến động của tổng mức tiêu thụ hàng hóa qua hai kỳ nghiên cứu do ảnh hưởng của cả yếu tố giá cả và lượng hàng hóa tiêu thụ, căn cứ trên quan hệ đã xác lập ta xây dựng hệ thống chỉ số

$$I_p \times I_q = I_{pq}$$

Trong đó:

I_{pq} là chỉ số giá trị mức tiêu thụ

I_p là chỉ số giá tính theo phương pháp Laspeyres hoặc Paasche

I_q là chỉ số khối lượng tính theo phương pháp Laspeyres hoặc Paasche

- Nếu chỉ số giá theo Paasche và chỉ số khối lượng theo Laspeyres thì công thức hệ thống chỉ số được viết lại như sau

$$\frac{\sum p_1 q_1}{\sum p_0 q_1} \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

- Nếu chỉ số giá theo Laspeyres và chỉ số khối lượng theo Paasche thì công thức hệ thống chỉ số được viết lại như sau

$$\frac{\sum p_1 q_0}{\sum p_0 q_0} \frac{\sum q_1 p_1}{\sum q_0 p_1} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Dù viết theo cách nào thì quá trình rút gọn cũng vẫn đảm bảo quan hệ

$$\text{Chỉ số giá cả} \times \text{Chỉ số khối lượng} = \text{Chỉ số giá trị}$$

Trở lại ví dụ trình bày ở Bảng 13.1 liệt kê giá cả và lượng hàng tiêu thụ tương ứng của một số mặt hàng tại cửa hàng A ở kì gốc 2000 và kì nghiên cứu 2005 tại nội dung Chỉ số tổng hợp. Dưới đây trình bày lại bảng số liệu 13.1 để bạn đọc tiện theo dõi

Tên hàng hóa	ĐVT	Giá (ngàn đồng)		Số lượng tiêu thụ (ngàn ĐVT)		Giá trị (triệu đồng)			
		Kì gốc (p ₀)	Kì n/cứu (p ₁)	Kì gốc (q ₀)	Kì n/cứu (q ₁)	p ₁ q ₀	p ₀ q ₀	p ₁ q ₁	p ₀ q ₁
X	Kg	5	6	10	13	60	50	78	65
Y	Lít	10	12,2	5	5,5	61	50	67,1	55
Z	Chục	8	10	0,25	0,32	2,5	2	3,2	2,56
Tổng						123,5	102	148,3	122,56

Với ví dụ này ta chỉ sử dụng:

Chỉ số giá Laspeyres

$$\frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{123,5}{102} = 1,2108 \text{ và}$$

Chỉ số khối lượng Paasche

$$\frac{\sum q_1 p_1}{\sum q_0 p_1} = \frac{148,3}{123,5} = 1,2008$$

Xây dựng hệ thống chỉ số

$$\frac{123,5}{102} \times \frac{148,3}{123,5} = \frac{148,3}{102}$$

$$1,2108 \times 1,2008 = 1,454$$

- Xác định số tuyệt đối về giá tăng giá trị

$$(\sum p_1 q_1 - \sum p_0 q_0) = (\sum p_1 q_0 - \sum p_0 q_0) + (\sum p_1 q_1 - \sum p_1 q_0)$$

$$(148,3 - 102) = (123,5 - 102) + (148,3 - 123,5)$$

$$46,3 = 21,5 + 24,8$$

- Xác định số tương đối về giá tăng giá trị bằng cách chia các chênh lệch tuyệt đối cho lượng $\sum p_0 q_0$

$$\frac{(\sum p_1 q_1 - \sum p_0 q_0)}{\sum p_0 q_0} = \frac{(\sum p_1 q_1 - \sum p_1 q_0)}{\sum p_0 q_0} + \frac{(\sum p_1 q_0 - \sum p_0 q_0)}{\sum p_0 q_0}$$

$$\frac{(148,3 - 102)}{102} = \frac{(123,5 - 102)}{102} + \frac{(148,3 - 123,5)}{102}$$

$$0,454 = 0,211 + 0,243$$

Nhân 100% vào hai vế của phương trình trên ta được kết quả cuối cùng

$$45,4\% = 21,1\% + 24,3\%$$

Từ hàng loạt kết quả tính toán này ta có nhận xét tổng hợp như sau :

Tổng giá trị mức tiêu thụ hàng hóa của cửa hàng A tại năm 2005 so với năm 2000 là 1.454 tức là tăng 45,4% tương ứng tăng 46,3 (triệu đồng) là do hai yếu tố:

- Do giá cả các mặt hàng nói chung ở năm 2005 so với 2000 tăng 21,08% đã làm cho tổng giá trị mức tiêu thụ hàng hóa tăng thêm 21,5 (triệu đồng), tương ứng tăng 21,1% về tương đối.
- Do khối lượng tiêu thụ các mặt hàng nói chung ở năm 2005 tăng 20,08% so với năm 2000 làm cho tổng giá trị mức tiêu thụ hàng hóa tăng thêm 24,8 (triệu đồng), tương ứng tăng 24,3% về tương đối

Tóm lại hệ thống chỉ số tổng hợp được dùng để phân tích ảnh hưởng của các yếu tố thành phần đối với chỉ tiêu phức tạp (ví dụ yếu tố giá và yếu tố khối lượng tiêu thụ ảnh hưởng đến giá trị mức tiêu thụ) cho ta thông tin mới về sự biến động của hiện tượng theo sự tác động của các yếu tố cấu thành đó. Vì vậy phương pháp này còn dùng để phân tích nhiều loại quan hệ khác như:

- Giá thành bình quân của một SP x Số SP sản xuất = Giá thành toàn bộ SP
 - Năng suất lao động của một công nhân x Số công nhân = Số SP sản xuất
- Ngoài ra đôi lúc người ta còn tận dụng mối quan hệ cấu thành nên hệ thống chỉ số để tính ra một chỉ số chưa biết trong khi đã biết các chỉ số còn lại trong hệ thống đó.

13.6.2 Hệ thống các chỉ số liên hoàn và định gốc

Khi có nhiều số liệu qua thời gian người ta có thể tính toán các chỉ số liên hoàn và chỉ số định gốc. Ví dụ như chỉ số giá cả.

Nếu chúng ta so sánh giá cả qua 5 thời kỳ, giả sử như 5 năm, thì có thể xây dựng các dãy chỉ số sau:

Năm	0	1	2	3	4	5
Dãy các chỉ số liên hoàn, quyền số thay đổi	-	$\frac{\sum p_1 q_1}{\sum p_0 q_1}$	$\frac{\sum p_2 q_2}{\sum p_1 q_2}$	$\frac{\sum p_3 q_3}{\sum p_2 q_3}$	$\frac{\sum p_4 q_4}{\sum p_3 q_4}$	$\frac{\sum p_5 q_5}{\sum p_4 q_5}$
Dãy các chỉ số liên hoàn, quyền số cố định	-	$\frac{\sum p_1 q_0}{\sum p_0 q_0}$	$\frac{\sum p_2 q_0}{\sum p_1 q_0}$	$\frac{\sum p_3 q_0}{\sum p_2 q_0}$	$\frac{\sum p_4 q_0}{\sum p_3 q_0}$	$\frac{\sum p_5 q_0}{\sum p_4 q_0}$
Dãy các chỉ số định gốc, quyền số cố định	-	$\frac{\sum p_1 q_0}{\sum p_0 q_0}$	$\frac{\sum p_2 q_0}{\sum p_0 q_0}$	$\frac{\sum p_3 q_0}{\sum p_0 q_0}$	$\frac{\sum p_4 q_0}{\sum p_0 q_0}$	$\frac{\sum p_5 q_1}{\sum p_0 q_0}$

Nếu chúng ta tính chỉ số liên hoàn với quyền số thay đổi thì giữa các chỉ số này không có liên hệ tích số với nhau thành một hệ thống. Nếu chúng ta tính chỉ số liên hoàn với quyền số cố định (trong ví dụ này là quyền số của

kỳ gốc số 0, là năm đứng trước năm số 1), thì các chỉ số này có liên hệ tích số với nhau thành một hệ thống chỉ số.

Lấy ví dụ, nếu chúng ta đem nhân các chỉ số liên hoàn (quyền số cố định) của năm thứ 1 và năm thứ 2, thì kết quả sẽ là chỉ số giá của năm thứ 2 so với năm lấy làm gốc so sánh, năm 0. Cụ thể công thức sẽ là:

$$\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_2 q_0}{\sum p_1 q_0} = \frac{\sum p_2 q_0}{\sum p_0 q_0}$$

Hoặc nếu chúng ta muốn so sánh giá cả giữa 2 năm bất kỳ, ví dụ như năm thứ 5 so với năm thứ 2, có thể dễ dàng thấy rằng chỉ cần đem chỉ số liên hoàn (quyền số cố định) của các năm thứ 3, thứ 4 thứ 5 nhân với nhau là chúng ta có được chỉ số giá định gốc của năm thứ 5 so với năm thứ 2. Cụ thể công thức sẽ là:

$$\frac{\sum p_5 q_0}{\sum p_2 q_0} = \frac{\sum p_3 q_0}{\sum p_2 q_0} \times \frac{\sum p_4 q_0}{\sum p_3 q_0} \times \frac{\sum p_5 q_0}{\sum p_4 q_0}$$

Như vậy các chỉ số dùng quyền số cố định có ưu điểm là có quan hệ tích số, và chính nhờ điều này mà chúng ta có thể dễ dàng so sánh giữa 2 thời gian bất kỳ, và tạo ra các hệ thống chỉ số liên hoàn và định gốc. Hơn nữa, vì các chỉ số liên hoàn có quan hệ tích số nên chúng ta có thể áp dụng công thức tính trung bình nhân để tính ra chỉ số giá trung bình (hàng tháng, hàng năm trong một giai đoạn nào đó).

Hiện tại chỉ số giá tiêu dùng ở nước ta đang dùng kiểu quyền số cố định, cho nên chúng ta có thể dễ dàng so sánh giá cả giữa 2 thời gian bất kỳ.

Chúng ta có chỉ số giá tiêu dùng hàng tháng cho trong bảng sau:

	2001	2002	2003	2004	2005	2006
Tháng trước = 100%						
Tháng 1	100.3	101.1	100.9	101.1	101.1	101.2
Tháng 2	100.4	102.2	102.2	103	102.5	102.1
Tháng 3	99.3	99.2	99.4	100.8	100.1	99.5
Tháng 4	99.5	100	100	100.5	100.6	100.2
Tháng 5	99.8	100.3	99.9	100.9	100.5	100.6
Tháng 6	100	100.1	99.7	100.8	100.4	100.4
Tháng 7	99.8	99.9	99.7	100.5	100.4	100.4
Tháng 8	100	100.1	99.9	100.6	100.4	100.4
Tháng 9	100.5	100.2	100.1	100.3	100.8	100.3
Tháng 10	100	100.3	99.8	100	100.4	100.2
Tháng 11	100.2	100.3	100.6	100.2	100.4	100.6
Tháng 12	101	100.3	100.8	100.6	100.8	100.5
Bình quân tháng	100.1	100.3	100.2	100.8	100.7	

(Nguồn: Cục thống kê TP Hồ Chí Minh)

Từ các số liệu chi tiết hàng tháng trên, chúng ta có thể tính:

- Chỉ số giá tháng 12 năm 2005 so với tháng 12 năm 2004 bằng cách lấy tích số của 12 chỉ số giá từng tháng trong năm 2005:

$$1,011 \times 1,025 \times 1,001 \times 1,006 \times 1,005 \times 1,004 \times 1,004 \times 1,008 \times 1,004 \times 1,004 \times 1,004 \times 1,008 \\ = 1,0871 = 108,71\%$$

- Chỉ số giá trung bình hàng tháng trong năm 2005 bằng cách lấy trung bình nhân của 12 chỉ số giá liên hoàn hàng tháng trong năm 2005:

$$\sqrt[12]{1,011 \times 1,025 \times 1,001 \times 1,006 \times 1,005 \times 1,004 \times 1,004 \times 1,008 \times 1,004 \times 1,004 \times 1,004 \times 1,008} \\ = 1,007 = 100,7\%.$$

Như vậy giá cả tiêu dùng tăng trung bình hàng tháng là 0,7% trong năm 2005.

- Chỉ số giá của tháng 6 so với tháng 1 năm 2005 bằng cách lấy tích số của các chỉ số giá hàng tháng của tháng 2, 3, 4, 5, 6 nhân với nhau, cụ thể là:

$$1,025 \times 1,001 \times 1,006 \times 1,005 \times 1,004 = 1,0415 = 104,15\%.$$

Như vậy giá cả tiêu dùng tháng 6 so với tháng 1 trong năm 2005 đã tăng 4,15%.

- Chỉ số giá của tháng 6/2005 so với tháng 6/2004 (thường gọi là so với cùng kỳ năm trước) bằng cách lấy tích số của các chỉ số giá hàng tháng từ tháng 7/2004 đến tháng 6/2005 nhân với nhau, cụ thể là:

$$1,005 \times 1,006 \times 1,003 \times 1,000 \times 1,002 \times 1,006 \times 1,011 \times 1,025 \times 1,001 \times 1,006 \times 1,005 \times 1,004 \\ = 1,0763 = 107,63\%$$

Như vậy giá cả tiêu dùng tháng 6/2005 so với tháng 6/2004 đã tăng 7,63%.

13.6.3 Hệ thống chỉ số nghiên cứu biến động của chỉ tiêu trung bình

Chúng ta nhớ lại công thức tính giá trị trung bình có trọng số

$$\bar{X}_w = \frac{\sum w_i x_i}{\sum w_i}$$

Thay kí hiệu của trọng số w_i bằng f_i ta viết lại công thức này

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

Nếu gọi \bar{x}_1 và \bar{x}_0 là số trung bình của kì nghiên cứu và kì gốc thì chỉ số $I_{\bar{x}} = \frac{\bar{x}_1}{\bar{x}_0}$ cho biết biến động của chỉ tiêu trung bình giữa kì nghiên cứu và kì gốc.

Từ công thức tính trung bình có trọng số tổng quát ở trên, chúng ta phán đoán được rằng nếu dùng phương pháp hệ thống chỉ số để phân tích ảnh hưởng của các yếu tố cấu thành đối với biến động chỉ tiêu trung bình \bar{x} thì ta phải chỉ ra được trị trung bình chịu ảnh hưởng của các yếu tố nào. Cũng từ công thức ta thấy giá trị \bar{x} tính được lớn hay bé phụ thuộc hai yếu tố:

- Giá trị của các x_i ra sao
- Tỷ lệ $\frac{f_i}{\sum f_i}$ thay đổi như thế nào, cụ thể hơn, qua kiến thức ở nội dung số trung bình ta đã biết kết quả cuối cùng \bar{x} sẽ chịu ảnh hưởng của các giá trị x_i có lượng $\frac{f_i}{\sum f_i}$ chiếm tỷ trọng lớn.

Muốn sử dụng phương pháp hệ thống chỉ số để phân tích ảnh hưởng biến động của các yếu tố đến chỉ tiêu trung bình, trước hết ta qui ước lại:

- f_1 và f_0 là quyền số của kì nghiên cứu và kì gốc
- x_1 và x_0 là các giá trị quan sát thuộc kì nghiên cứu và kì gốc

Ta phân tích chỉ số $\frac{\bar{x}_1}{\bar{x}_0}$ thành hệ thống chỉ số như sau:

$$\frac{\bar{x}_1}{\bar{x}_0} = \frac{\frac{\sum x_1 f_1}{\sum f_1}}{\frac{\sum x_0 f_0}{\sum f_0}} = \frac{\sum x_1 f_1}{\sum f_1} \times \frac{\sum x_0 f_0}{\sum f_0}$$

Viết thu gọn hệ thống chỉ số trên dưới dạng kí hiệu phù hợp ta có

$$\frac{\bar{x}_1}{\bar{x}_0} = \frac{\bar{x}_1}{\bar{x}_{0_1}} \times \frac{\bar{x}_{0_1}}{\bar{x}_0}$$

$$I_{\bar{x}} = I_x \times I_f / \Sigma f$$

Trong đó:

- $I_{\bar{x}}$ là chỉ số của chỉ tiêu trung bình, phản ánh biến động của chỉ tiêu trung bình giữa hai kì nghiên cứu.
- I_x là chỉ số cấu thành cố định phản ánh biến động của chỉ tiêu trung bình do ảnh hưởng của riêng đặc điểm thống kê ta nghiên cứu.
- $I_{f_1 f_0}$ là chỉ số ảnh hưởng kết cấu phản ánh biến động của chỉ tiêu trung bình do ảnh hưởng của riêng yếu tố kết cấu có liên quan.

Để làm rõ phần vừa trình bày ở trên, ta xem một ví dụ. Một nhà máy có 3 phân xưởng A, B và C cùng sản xuất một loại sản phẩm với số liệu trong bảng sau:

Bảng 13.6

Phân xưởng	Kì gốc		Kì nghiên cứu		Giá trị sản xuất (1000 đ)		
	Giá thành đơn vị (1000đ)	Sản lượng (cái)	Giá thành đơn vị (1000đ)	Sản lượng (cái)	$x_1 f_1$	$x_0 f_0$	$x_0 f_1$
	x_0	f_0	x_1	f_1			
A	10	1000	9	8000	72000	10000	8000
B	12	2500	11,5	3000	34500	30000	36000
C	13	4500	12,5	1000	12500	58500	13000
Tổng		8000		12000	119000	98500	129000

Gọi giá thành trung bình của nhà máy là \bar{x} thì giá thành này được tính từ giá thành đơn vị của từng phân xưởng theo công thức sau

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

Với quyền số f_i là sản lượng của từng phân xưởng tương ứng.

Ta lần lượt tính các giá thành trung bình cho kì nghiên cứu, kì gốc, và giá thành trung bình của kì gốc mà quyền số là kì nghiên cứu.

$$\bar{x}_1 = \frac{\sum x_1 f_1}{\sum f_1} = \frac{119000}{12000} = 9,92$$

$$\bar{x}_0 = \frac{\sum x_0 f_0}{\sum f_0} = \frac{98500}{8000} = 12,31$$

$$\bar{x}_{0,1} = \frac{\sum x_0 f_1}{\sum f_1} = \frac{129000}{12000} = 10,75$$

Thể các số liệu này vào hệ thống chỉ số thì ta có:

$$\frac{9,92}{12,31} = \frac{9,92}{10,75} \times \frac{10,75}{12,31}$$

$$0,806 = 0,923 \times 0,873$$

Tính số chênh lệch tuyệt đối về giá thành trung bình

$$(\bar{x}_1 - \bar{x}_0) = (\bar{x}_1 - \bar{x}_{0_1}) + (\bar{x}_{0_1} - \bar{x}_0)$$

$$(9,92 - 12,31) = (9,92 - 10,75) + (10,75 - 12,31)$$

$$(-2,39) = (-0,83) + (-1,56)$$

Tính số chênh lệch tương đối về giá thành trung bình:

$$\frac{\bar{x}_1 - \bar{x}_0}{\bar{x}_0} = \frac{\bar{x}_1 - \bar{x}_{0_1}}{\bar{x}_{0_1}} + \frac{\bar{x}_{0_1} - \bar{x}_0}{\bar{x}_0}$$

$$\frac{(9,92 - 12,31)}{12,31} = \frac{(9,92 - 10,75)}{12,31} + \frac{(10,75 - 12,31)}{12,31}$$

$$(-0,194) = (-0,067) + (-0,127)$$

Nhận xét:

Giá thành trung bình kì nghiên cứu so với kì gốc bằng 80,6% tức là giảm 19,4% hay giảm 2,39 ngàn đồng, là do ảnh hưởng của hai yếu tố:

- Do giá thành đơn vị của các phân xưởng giảm $(1-0,923)100 = 7,7\%$ làm giá thành trung bình giảm đi 0,83 ngàn đồng, về mặt tương đối giá thành trung bình đã giảm 6,7%
- Do kết cấu sản lượng sản xuất thay đổi làm cho giá thành trung bình giảm 1,56 ngàn đồng hay về mặt tương đối giảm 12,7%. hay giảm 1,56 ngàn đồng.

16.6.4 Hệ thống chỉ số phân tích biến động của chỉ tiêu tổng trị số

Trong nhiều trường hợp phân tích, chỉ tiêu trung bình là một thành phần của một chỉ tiêu tổng quát hơn, đó là chỉ tiêu tổng trị số (hay còn gọi là tổng giá trị). Để dễ hình dung, chúng ta xem diễn giải sau:

Từ công thức số trung bình có trọng số (mà chúng ta vừa xem cách thức phân tích biến động của nó qua thời gian) $\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$, đem nhân mẫu số

của vế phải công thức lên vế trái, ta có:

$$\bar{x} \times \sum f_i = \sum x_i f_i$$

Trong công thức này, ở vế phải là chỉ tiêu tổng trị số (của một đặc điểm đang nghiên cứu), ví dụ như tổng giá thành sản xuất (hay là tổng chi phí sản xuất, nếu như không có sản phẩm đỡ dang). Vế trái là 2 chỉ tiêu thành phần: giá thành trung bình đơn vị sản phẩm và tổng lượng sản phẩm đã sản xuất, và hai thành phần này có quan hệ tích số. Theo quy tắc lập thành hệ thống chỉ số tổng hợp (mục 13.6.1), chúng ta có thể áp dụng để lập thành hệ thống chỉ số phân tích biến động của chỉ tiêu tổng trị số ở vế phải theo 2 yếu tố ảnh hưởng là chỉ tiêu trung bình và chỉ tiêu tổng lượng. Theo nguyên tắc này có thể dễ dàng lập thành hệ thống chỉ số sau:

$$\frac{\sum x_1 f_1}{\sum x_0 f_0} = \frac{\bar{x}_1 \times \sum f_1}{\bar{x}_0 \times \sum f_0} = \frac{\bar{x}_1 \times \sum f_1}{\bar{x}_0 \times \sum f_1} \times \frac{\bar{x}_0 \times \sum f_1}{\bar{x}_0 \times \sum f_0}$$

Đơn giản bớt các thành phần giống nhau ở tử và mẫu số

$$\frac{\sum x_1 f_1}{\sum x_0 f_0} = \frac{\bar{x}_1 \times \sum f_1}{\bar{x}_0 \times \sum f_0} = \frac{\bar{x}_1 \times \sum f_1}{\cancel{\bar{x}_0 \times \sum f_1}} \times \frac{\cancel{\bar{x}_0 \times \sum f_1}}{\cancel{\bar{x}_0 \times \sum f_0}}$$

Công thức được viết gọn lại thành:

$$\frac{\sum x_1 f_1}{\sum x_0 f_0} = \frac{\bar{x}_1}{\bar{x}_0} \times \frac{\sum f_1}{\sum f_0}$$

Và được ký hiệu như sau:

$$I_{\Sigma f} = I_{\bar{x}} \times I_{\Sigma f}$$

Chỉ số của
chỉ tiêu tổng trị số

Chỉ số của
chỉ tiêu trung bình

Chỉ số của
chỉ tiêu tổng lượng

Các số chênh lệch tuyệt đối của hệ thống chỉ số này cũng được tính bằng cách lấy tử số trừ cho mẫu số của hệ thống chỉ số đầy đủ, cụ thể như sau:

$$\sum x_1 f_1 - \sum x_0 f_0 = (\bar{x}_1 \sum f_1 - \bar{x}_0 \sum f_1) + (\bar{x}_0 \sum f_1 - \bar{x}_0 \sum f_0)$$

Hay đơn giản hơn là rút các thừa số chung ra ngoài, chúng ta được:

$$\underbrace{\sum x_1 f_1 - \sum x_0 f_0}_{\begin{array}{l} \text{Mức tăng (giảm) của} \\ \text{chỉ tiêu tổng trị số} \end{array}} = \underbrace{(\bar{x}_1 - \bar{x}_0) \sum f_1}_{\begin{array}{l} \text{Mức tăng (giảm)} \\ \text{của chỉ tiêu tổng trị số} \\ \text{do ảnh hưởng tăng (giảm) của} \\ \text{chỉ tiêu trung bình} \end{array}} + \underbrace{(\sum f_1 - \sum f_0) \bar{x}_0}_{\begin{array}{l} \text{Mức tăng (giảm)} \\ \text{của chỉ tiêu tổng trị số} \\ \text{do ảnh hưởng tăng (giảm) của} \\ \text{chỉ tiêu tổng lượng} \end{array}}$$

Mức tăng (giảm) của
chỉ tiêu tổng trị số

Mức tăng (giảm)
của chỉ tiêu tổng trị số
do ảnh hưởng tăng (giảm) của
chỉ tiêu trung bình

Mức tăng (giảm)
của chỉ tiêu tổng trị số
do ảnh hưởng tăng (giảm) của
chỉ tiêu tổng lượng

Các số chênh lệch tương đối của hệ thống chỉ số này cũng được tính bằng cách lấy các chênh lệch tuyệt đối chia cho mức độ kỳ gốc, cụ thể như sau:

$$\frac{\sum x_1 f_1 - \sum x_0 f_0}{\sum x_0 f_0} = \frac{(\bar{x}_1 - \bar{x}_0) \sum f_1}{\sum x_0 f_0} + \frac{(\sum f_1 - \sum f_0) \bar{x}_0}{\sum x_0 f_0}$$

Sử dụng lại các số liệu ở Bảng 13.6 trong phần trước, áp dụng hệ thống chỉ số này chúng ta phân tích biến động của chỉ tiêu tổng giá thành của toàn nhà máy.

Thay số liệu vào công thức $\frac{\sum x_1 f_1 - \sum x_0 f_0}{\sum x_0 f_0} = \frac{\bar{x}_1 - \bar{x}_0}{\bar{x}_0} \times \frac{\sum f_1}{\sum f_0}$, ta có

$$\frac{119000}{98500} = \frac{9,92}{12,31} \times \frac{12000}{8000}$$

$$1,208 = 0,805 \times 1,5$$

Tính số chênh lệch tuyệt đối về giá thành trung bình

$$\sum x_1 f_1 - \sum x_0 f_0 = (\bar{x}_1 - \bar{x}_0) \sum f_1 + (\sum f_1 - \sum f_0) \bar{x}_0$$

$$(119000 - 98500) = (9,92 - 12,31) \times 12000 + (12000 - 8000) \times 12,31$$

$$20500 = - 28680 + 49240$$

Lý do của sự chênh lệch giữa hai vé là do số liệu khi tính các giá trị trung bình đã bị làm tròn dẫn đến sai số trong các phép tính sau, nếu các bạn tính lại các trị trung bình \bar{x}_1 và \bar{x}_0 và lấy tới ba số lẻ sau dấu phẩy thì $\bar{x}_1 = 9,917$ và $\bar{x}_0 = 12,313$

Chúng ta lại thay thế các số liệu này vào công thức trên thì được

$$(119000 - 98500) = (9,917 - 12,313) \times 12000 + (12000 - 8000) \times 12,313$$

$$20500 = - 28752 + 49252$$

Nhận xét:

- Chỉ tiêu tổng giá thành sản xuất tăng 20,8% do ảnh hưởng kết hợp của 2 yếu tố là chỉ tiêu giá thành trung bình giảm 19,5% và chỉ tiêu tổng lượng sản phẩm sản xuất tăng 50%.
- Mức tăng tuyệt đối của tổng giá thành là 20500 ngàn đồng, do kết quả của giá thành trung bình giảm làm tổng giá thành giảm 28752 ngàn đồng và tổng sản lượng sản phẩm tăng làm tổng giá thành tăng 49252 ngàn đồng.

13.7. MỘT SỐ CHỈ SỐ THƯỜNG GẶP TRONG THỰC TẾ

13.7.1 Chỉ số giá tiêu dùng (CPI)

Chỉ số giá tiêu dùng đo lường sự biến động của giá tiêu dùng, nó là một chỉ tiêu kinh tế quan trọng, thường được sử dụng trong phân tích kinh tế, đánh giá tình hình lạm phát, quan hệ cung cầu, sức mua của dân cư, là cơ sở tham khảo cho việc điều chỉnh lãi suất ngân hàng, tiền lương, tính toán điều chỉnh tiền công trong các hợp đồng sản xuất kinh doanh...

Cần xác định rõ giá tiêu dùng là giá mà người tiêu dùng mua hàng hoá hoặc chi trả cho các dịch vụ phục vụ trực tiếp cho đời sống hàng ngày. Giá tiêu dùng được biểu hiện bằng giá bán lẻ hàng hoá trên thị trường và giá dịch vụ phục vụ sinh hoạt đời sống; không bao gồm giá đất đai, giá hàng hoá bán cho sản xuất và các công việc có tính chất sản xuất kinh doanh.

Phương pháp tính chỉ số giá tiêu dùng

Để tính chỉ số giá tiêu dùng, cần thu thập giá của các mặt hàng và các dịch vụ đại diện, phổ biến tiêu dùng của dân cư, các mặt hàng và dịch vụ này được xác định theo một danh mục – mà người ta thường gọi một cách形象 là "rổ" hàng hoá, dịch vụ (bao gồm một số nhóm cơ bản là thực phẩm, nhà ở, quần áo, giao thông, dịch vụ y tế, giải trí... mà được cụ thể hóa thành hàng trăm mặt hàng). Các hàng hóa trong rổ này có thể được thay đổi qua thời gian để theo kịp thói quen tiêu dùng của xã hội và sự xuất hiện của những sản phẩm hoặc dịch vụ mới. Chỉ số giá tiêu dùng được tính từ giá bán lẻ hàng hoá và giá dịch vụ tiêu dùng của rổ hàng hoá và dịch vụ này với quyền số là cơ cấu chi tiêu của các hộ gia đình.

Chỉ số giá tiêu dùng tính được sẽ phản ánh xu hướng và mức độ biến động giá của "rổ" hàng hoá và dịch vụ tiêu dùng đại diện nói trên, khi giá của các mặt hàng, nhóm hàng trong "rổ" có thay đổi.

Trong điều kiện về vật chất, kỹ thuật, nguồn kinh phí và cũng phù hợp với phương pháp của nhiều nước, Chỉ số giá tiêu dùng ở nước ta hiện nay được tính theo công thức Laspeyres. Công thức tổng quát của chỉ số giá tiêu dùng:

$$I_p = \frac{\sum q_{2000} p_i}{\sum q_{2000} P_{2000}} \times 100\%$$

Trong đó:

- I_p : chỉ số giá tiêu dùng
- p_i : giá kỳ báo cáo

- p_{2000} là giá năm 2000
- q_{2000} là cơ cấu chi tiêu của các hộ gia đình năm 2000

Hiện nay cơ cấu chi tiêu của các hộ gia đình năm 2000 được chọn làm quyền số để tính Chỉ số giá tiêu dùng, nó được xác định từ cơ cấu chi tiêu của hộ gia đình với các nguồn số liệu sau đây:

- Kết quả "Điều tra mức sống dân cư Việt Nam 1997-1998" do Tổng cục Thống kê thực hiện trong năm 1998.
- Kết quả "Điều tra bổ sung về chi tiêu hộ gia đình tại 10 tỉnh năm 1999" do Tổng cục Thống kê thực hiện trong năm 1999.

Cần chú ý khi hiểu ý nghĩa chỉ số giá tiêu dùng là nó không phản ánh mức giá mà nó đo lường mức độ biến động giá giữa hai khoảng thời gian. Ví dụ: Chỉ số giá tháng 4/2003 so với tháng 3/2003 của nhóm hàng "Thiết bị đồ dùng gia đình" là 100,5% và Chỉ số giá nhóm hàng "Dược phẩm, Y tế" là 101,3% không có nghĩa là trong tháng 4 "hàng y tế" đắt hơn "thiết bị đồ dùng gia đình" mà chỉ là: so với tháng 3, trong tháng 4 giá các mặt hàng y tế tăng mạnh hơn giá các mặt hàng thiết bị đồ dùng gia đình.

13.7.2 Chỉ số chứng khoán VN-Index

"....Chỉ số chứng khoán Việt Nam (VN-Index) trong phiên giao dịch ngày 25-4 đã phục hồi trở lại, đạt 923,89 điểm, tăng 18,36 điểm (tương đương 2%)...". Bạn đọc có thể gặp các thông báo về tình hình thị trường chứng khoán như thế này hàng ngày trên các phương tiện truyền thông đại chúng, trong đó chỉ số VN-Index được nhắc tới thường xuyên. Vậy VN-Index được tính toán như thế nào?

Về công thức, chỉ số giá chứng khoán VN-Index cơ bản được tính như sau

$$VN - Index = \frac{\text{Tổng giá trị thị trường của các cổ phiếu niêm yết hiện tại}}{\text{Tổng giá trị của các cổ phiếu niêm yết cơ sở}} \times 100$$

Trong công thức này có khái niệm cần làm rõ là khái niệm "cơ sở". Ngày được chọn làm cơ sở là ngày 28/7/2000 hay được gọi là kỳ gốc, và tại ngày này bạn đọc dễ dàng hình dung được ngay giá trị VN-Index cơ sở là 100% mà người ta hay gọi ngắn gọn là 100 điểm.

Giả sử kết quả phiên giao dịch trên thị trường chứng khoán Việt Nam vào ngày 28 tháng 7 năm 2000 được liệt kê như sau:

Bảng 13.7

Tên cổ phiếu	Giá thực hiện	Số lượng CP niêm yết	Giá trị thị trường (Đồng)
REE	16000	15.000.000	240.000.000.000
SAM	17000	12.000.000	204.000.000.000
Tổng			444.000.000.000

Vậy chỉ số VN – Index được tính cho ngày này là

$$VN - Index = \frac{444.000.000.000}{444.000.000.000} \times 100\% = 100$$

Sau đó, vào ngày 2 tháng 8 năm 2000 kết quả giao dịch như sau:

Bảng 13.8

Tên cổ phiếu	Giá thực hiện	Số lượng CP niêm yết	Giá trị thị trường (Đồng)
REE	16600	15.000.000	249.000.000.000
SAM	17500	12.000.000	210.000.000.000
Tổng:			459.000.000.000

Ta tính chỉ số giá chứng khoán cho phiên giao dịch hiện tại là ngày 2 tháng 8, bằng chênh lệch giữa giá trị thị trường ngày 2 tháng 8 với giá trị niêm yết của phiên giao dịch cơ sở ngày 28 tháng 7.

$$VN - Index = \frac{459.000.000.000}{444.000.000.000} \times 100\% = 103,38$$

Kết quả này nói lên một điều rằng giá cả 2 loại cổ phiếu nói trên ở ngày 2/8 đã tăng 3,38% so với ngày giao dịch đầu tiên, do giá tăng mà tổng giá trị thị trường tăng $459 - 444 = 15$ tỉ đồng

Chú ý: Từ hai bảng số liệu về tình hình giao dịch trên thị trường ở trên chúng ta tập hợp lại thành một bảng tổng hợp giá và số lượng, trong đó giá và lượng kí gốc (28/7) kí hiệu là p_0 và q_0 , còn kí nghiên cứu (2/8) kí hiệu p_1 và q_1

Bảng 13.9

Tên CP	Giá (đồng)		Số lượng (ngàn CP)		Giá trị thị trường (ngàn đồng)	
	Kì gốc (p ₀)	Kì n/cứu (p ₁)	Kì gốc (q ₀)	Kì n/cứu (q ₁)	p ₁ q ₀	p ₀ q ₀
REE	16000	16600	15.000	15.000	249.000.000	240.000.000
SAM	17000	17500	12.000	12.000	210.000.000	204.000.000
Tổng					459.000.000	444.000.000

Áp dụng phương pháp tính chỉ số giá theo Laspeyres đã biết, ta được

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100\% = \frac{459000000}{444000000} \times 100\% = 103,38\%$$

Như vậy có thể thấy phương pháp tính chỉ số giá chứng khoán VN – Index giống phương pháp Chỉ số giá Laspeyres trong tình huống đơn giản nhất là cơ cấu số cổ phiếu niêm yết không thay đổi. Tuy nhiên trên thực tế thường xảy ra tình huống làm thay đổi cơ cấu số cổ phiếu niêm yết như: thêm, bớt cổ phiếu giao dịch vào cơ cấu tính toán. Điều này sẽ làm phát sinh tính không liên tục của chỉ số do số lượng cổ phiếu (mà bản chất của phương pháp chỉ số Laspeyres cho biết nó là quyền số) cấu thành tử số đã khác so với số lượng cổ phiếu cấu thành mẫu số của công thức tính VN-Index. Do đó, số chia (mẫu số) trong công thức tính chỉ số VN-Index phải được điều chỉnh nhằm duy trì tính liên tục cần có của chỉ số. Nguyên tắc điều chỉnh được tính như sau:

$$\text{số chia được điều chỉnh} = \frac{\text{Tổng giá trị thị trường các cổ phiếu sau thay đổi}}{\text{Tổng giá trị thị trường các cổ phiếu trước thay đổi}} \cdot \text{số chia cũ}$$

Ví dụ ngày 4 tháng 8 năm 2000 có thêm hai loại cổ phiếu Hapaco (HAP) và Transimex (TMS) được đưa vào giao dịch, do đó ta phải tìm số chia mới trước khi tính VN-Index cho ngày 4 tháng 8. Xem bảng tổng hợp sau:

Bảng 13.10

Mã cổ phiếu	Giá thực hiện	Số lượng CP niêm yết	Giá trị thị trường (đồng)
REE	16900	15.000.000	253.500.000.000
SAM	17800	12.000.000	213.600.000.000
HAP	16000	1.008.000	16.128.000.000
TMS	14000	2.200.000	30.800.000.000
Tổng			514.028.000.000

$$\text{số chia } \alpha = \frac{514.028.000.000}{253.500.000.000 + 213.600.000.000} = 444.000.000.000 = 488.607.219.010$$

Cuối cùng chỉ số VN – Index ngày 4 tháng 8 năm 2000 được tính theo công thức chính với mẫu số bây giờ là số chia đã điều chỉnh như sau

$$VN - Index = \frac{514.028.000.000}{488.607.219.010} \times 100\% = 105,2$$

Trong thời gian đầu chỉ số chứng khoán nước ta được tính toán sẽ đại diện cho tất cả các cổ phiếu được niêm yết và giao dịch trên thị trường chứng khoán vì số lượng chứng khoán được giao dịch trên sàn chưa nhiều. Một chỉ số chứng khoán của nước ngoài chỉ liệt kê một số loại chứng khoán tiêu biểu, ví dụ chỉ số Dow Jones là chỉ số giá chung của 65 chứng khoán đại diện, thuộc nhóm hàng đầu trong các chứng khoán được niêm yết tại Sở giao dịch chứng khoán New York.

Chỉ số chứng khoán VN-Index được tính toán và công bố dưới dạng đơn vị tính là điểm chứ không phải đơn vị tính %.

Khi phân tích thông tin về chỉ số chứng khoán bạn đọc luôn nhớ chỉ số giá cổ phiếu là thông tin thể hiện giá chứng khoán bình quân hiện tại so với giá bình quân thời kỳ gốc đã chọn được lấy là 100. Ví dụ, ta phân tích dòng thông báo “....Chỉ số chứng khoán Việt Nam (VN-Index) trong phiên giao dịch ngày 25/4/2007 đã phục hồi trở lại, đạt 923,89 điểm, tăng 18,36 điểm (tương đương 2%)...” như sau:

Thông báo này đề cập đến chỉ số VN-Index ngày 25 tháng 4 năm 2007 là 923,89 điểm tức là ngũ ý nói rằng về cơ bản giá cổ phiếu trên sàn giao dịch của ngày 25/4/2007 so với ngày gốc đã chọn đã tăng 9,2389 lần (để có cảm giác so sánh bạn đọc tham khảo thông tin này: từ mức 100, muốn lớn gấp 5 lần, chỉ số Dow Jones phải đợi nửa thế kỷ, đạt 500 điểm vào năm 1956). “Đã phục hồi trở lại” chứng tỏ chỉ số chứng khoán ngày kể trước (ngày 24/4) phải thấp hơn, cụ thể nó sẽ là $(923,89 - 18,36) = 905,53$ điểm, mức tăng tuyệt đối này được qui đổi sang mức tăng tương đối là:

$$\frac{923,89 - 905,53}{905,53} \times 100\% = 2,02\%$$

CHƯƠNG 14

CHUỖI THỜI GIAN VÀ DỰ BÁO TRÊN CHUỖI THỜI GIAN

14.1 CHUỖI THỜI GIAN

Để phân tích biến động của chỉ tiêu nghiên cứu qua thời gian, người ta thường tập hợp các số liệu được thu thập qua thời gian gọi là chuỗi số thời gian. Trong phương pháp này các giá trị quan sát không độc lập với nhau, mà ngược lại có sự phụ thuộc của các giá trị quan sát trong dãy số là đặc điểm, cơ sở cho việc xây dựng các phương pháp dự báo trên chuỗi thời gian. Các phương pháp dự báo định lượng có thể được phân chia thành hai loại: phân tích các giá trị qua thời gian và phân tích liên hệ nguyên nhân - kết quả. Phương pháp dự đoán bằng phân tích các mức độ qua thời gian liên quan đến việc tính toán các giá trị tương lai của yếu tố nghiên cứu dựa trên toàn bộ các quan sát có được ở quá khứ. Phân tích mối liên hệ nhân quả liên quan đến việc xác định các yếu tố ảnh hưởng đến yếu tố ta muốn dự đoán, như phân tích hồi qui bội để xem GDP phụ thuộc vào lượng đầu tư trong nước, lượng đầu tư ở nước ngoài, dân số... Phân tích mối liên hệ nhân quả đã được khảo sát kỹ trong Chương 11 và 12. Chương này tập trung vào việc mô tả, phân tích chuỗi thời gian và dự báo dựa vào chính các giá trị của chuỗi thời gian. Phân tích này được dựa trên giả định cơ bản là các yếu tố ảnh hưởng đến biến động của chỉ tiêu nghiên cứu trong quá khứ sẽ còn tiếp diễn tương tự như vậy trong tương lai.

14.1.1 Khái niệm

Chuỗi thời gian là một chuỗi các giá trị của một chỉ tiêu nghiên cứu được sắp xếp theo thứ tự thời gian. Ví dụ như giá cả hàng ngày một mã cổ phiếu nào đó ở thị trường chứng khoán ở thời điểm đóng cửa, chỉ số giá tiêu dùng hàng tháng của cả nước, lượng tiêu thụ điện hàng tháng ở một thành phố, số vụ tai nạn giao thông đường bộ, số vụ tự tử hàng năm...

Một chuỗi thời gian có dạng tổng quát như sau:

t_i	t_1	t_2	...	t_n
Y_i	Y_1	Y_2	...	Y_n

t_i ($i = \overline{1, n}$) : thời gian thứ i

Y_i ($i = \overline{1, n}$) : giá trị của chỉ tiêu tương ứng với thời gian thứ i

Căn cứ vào đặc điểm về mặt thời gian của dãy số, có thể chia ra 2 loại dãy số: dãy số thời kỳ và dãy số thời điểm.

14.1.1.1 Chuỗi thời kỳ

Là chuỗi số liệu biểu hiện biến động của chỉ tiêu nghiên cứu qua từng thời kỳ. Các mức độ trong chuỗi thời kỳ có thể cộng lại với nhau qua thời gian, phản ánh mức độ của chỉ tiêu nghiên cứu trong một thời kỳ dài hơn.

Ví dụ 1: Sản lượng cà phê xuất khẩu của Việt Nam từ 2001 đến 2005

Năm	2001	2002	2003	2004	2005
Sản lượng (ngàn tấn)	931,1	722,2	749,4	976,2	892,4

Nguồn: Tổng Cục Thống Kê

14.1.1.2 Chuỗi thời điểm

Là chuỗi số liệu biểu hiện biến động của chỉ tiêu nghiên cứu qua các thời điểm nhất định.

Các mức độ trong chuỗi thời điểm không thể cộng lại theo thời gian vì con số cộng này không có ý nghĩa.

Ví dụ 2: Giá vàng SJC tại TPHCM trong tuần cuối tháng 7 năm 2007

Ngày	23/7	24/7	25/7	26/7	27/7	28/7
Ngàn đồng/chỉ	1.317,0	1.316,5	1.310,0	1.307,5	1.294,0	1.294,0

14.1.2 CÁC ĐẠI LƯỢNG MÔ TẢ CHUỖI THỜI GIAN

14.1.2.1 Mức độ trung bình theo thời gian

Là số trung bình của các trị số của chỉ tiêu nghiên cứu trong chuỗi thời gian, là biểu hiện mức độ điển hình của chỉ tiêu nghiên cứu trong thời gian nghiên cứu.

Giả sử ta có chuỗi số thời gian Y_1, Y_2, \dots, Y_n

Gọi \bar{Y} : mức độ trung bình của chuỗi

Đối với chuỗi thời kỳ, áp dụng công thức trung bình cộng đơn giản

$$\bar{Y} = \frac{Y_1 + Y_2 + \dots + Y_n}{n} = \frac{\sum_{i=1}^n Y_i}{n} \quad (14.1)$$

Đối với chuỗi thời điểm, có 2 trường hợp tính toán giá trị trung bình :

- Nếu khoảng cách thời gian giữa các thời điểm bằng nhau, áp dụng công thức tính gần đúng là:

$$\bar{Y} = \frac{\frac{1}{2} Y_1 + Y_2 + \dots + Y_{n-1} + \frac{1}{2} Y_n}{n-1} \quad (14.2)$$

(n-1: số các khoảng cách thời gian)

- Khoảng cách thời gian giữa các thời điểm không đều nhau & thời gian nghiên cứu là liên tục, thì áp dụng công thức trung bình cộng có trọng số:

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i t_i}{\sum_{i=1}^n t_i} \quad (14.3)$$

trong đó: Y_i : mức độ thứ i trong dãy số

t_i : độ dài thời gian tương ứng với mức độ thứ i

14.1.2.2 Lượng tăng (giảm) tuyệt đối

Là đại lượng biểu hiện sự thay đổi về giá trị tuyệt đối của chỉ tiêu nghiên cứu giữa hai thời kỳ hoặc thời điểm nghiên cứu.

Tùy theo mục đích nghiên cứu, ta có:

- Lượng tăng (giảm) tuyệt đối liên hoàn*: thể hiện lượng tăng (giảm) tuyệt đối giữa hai thời gian đứng liền nhau trong chuỗi số.

$$\delta_i = Y_i - Y_{i-1} \quad (i = 2, \dots, n) \quad (14.4)$$

- Lượng tăng (giảm) tuyệt đối định gốc*: thể hiện lượng tăng giảm giữa kỳ so sánh với kỳ chọn làm gốc cố định cho mọi lần so sánh (thường là mức độ đầu tiên trong chuỗi số)

$$\Delta_i = Y_i - Y_1 \quad (i = 2, \dots, n) \quad (14.5)$$

Giữa lượng tăng (giảm) tuyệt đối liên hoàn và lượng tăng (giảm) tuyệt đối định gốc có mối liên hệ là tổng của các lượng tăng giảm tuyệt đối liên hoàn thì bằng lượng tăng giảm tuyệt đối định gốc tương ứng, thể hiện bằng công thức như sau:

$$\sum_{i=2}^n \delta_i = \Delta_n \quad (14.6)$$

- Lượng tăng giảm tuyệt đối trung bình* là số trung bình cộng của các lượng tăng giảm tuyệt đối liên hoàn, biểu hiện một cách chung nhất lượng tăng (giảm) tính trung bình cho cả một thời kỳ nghiên cứu.

$$\bar{\delta} = \frac{\sum_{i=2}^n \delta_i}{n-1} = \frac{\Delta_n}{n-1} = \frac{Y_n - Y_1}{n-1} \quad (14.7)$$

Đại lượng này chỉ có ý nghĩa khi các lượng tăng (giảm) tuyệt đối liên hoàn xấp xỉ nhau, nghĩa là trong suốt thời kỳ nghiên cứu, hiện tượng tăng (giảm) với một lượng tương đối đều.

14.1.2.3 Tốc độ phát triển

Là tỉ số dùng để đánh giá chỉ tiêu nghiên cứu qua một thời gian nhất định đã phát triển được với tốc độ cụ thể bao nhiêu (lần hay %).

- *Tốc độ phát triển liên hoàn:* thể hiện tốc độ phát triển của chỉ tiêu nghiên cứu giữa 2 kỳ liền nhau

$$t_i = \frac{Y_i}{Y_{i-1}} \quad (i = 2, 3, \dots, n) \quad (14.8)$$

- *Tốc độ phát triển định gốc:* thể hiện tốc độ phát triển của chỉ tiêu nghiên cứu giữa kỳ nghiên cứu với kỳ được chọn làm gốc so sánh

$$T_i = \frac{Y_i}{Y_1} \quad (i = 2, 3, \dots, n) \quad (14.9)$$

Giữa tốc độ phát triển liên hoàn và định gốc có các mối liên hệ là:

- tích các tốc độ phát triển liên hoàn bằng tốc độ phát triển định gốc tương ứng, thể hiện bằng công thức sau:

$$\prod_{i=2}^n t_i = T_n \quad (14.10)$$

- Tỉ số giữa hai tốc độ phát triển định gốc liên nhau trong chuỗi số bằng tốc độ phát triển liên hoàn.

$$\frac{T_i}{T_{i-1}} = t_i \quad (14.11)$$

- *Tốc độ phát triển trung bình:* thể hiện nhịp độ phát triển đại diện của chỉ tiêu trong suốt thời kỳ nghiên cứu.

$$\bar{t} = \sqrt[n]{\prod_{i=2}^n t_i} = \sqrt[n-1]{\frac{Y_n}{Y_1}} \quad (14.12)$$

Đại lượng này chỉ có ý nghĩa khi các tốc độ phát triển liên hoàn xấp xỉ nhau nghĩa là trong suốt thời kỳ nghiên cứu chỉ tiêu đã phát triển với một tốc độ tương đối đều.

14.2.3.4 Tốc độ tăng (giảm)

Là tỉ số dùng để đánh giá của chỉ tiêu giữa 2 thời gian nghiên cứu đã tăng (giảm) bao nhiêu lần (%).

- *Tốc độ tăng (giảm) liên hoàn*

$$a_i = \frac{Y_i - Y_{i-1}}{Y_{i-1}} = \frac{\delta_i}{Y_{i-1}} = t_i - 1 \quad (i = 2, 3, \dots, n) \quad (14.13)$$

- Tốc độ tăng (giảm) định gốc

$$A_i = \frac{Y_i - Y_1}{Y_1} = \frac{\Delta_i}{Y_1} = T_i - 1 \quad (i = 2, 3, \dots, n) \quad (14.14)$$

- Tốc độ tăng (giảm) trung bình

$$\bar{a} = \bar{t} - 1 \quad (14.15)$$

14.2.3.5 Trị tuyệt đối của 1% tăng (giảm) liên hoàn

Là đại lượng phản ánh 1% tăng (giảm) của 2 thời kỳ đứng liền nhau của hiện tượng nghiên cứu tương ứng với một mức độ tuyệt đối là bao nhiêu. Đại lượng này có tác dụng giúp tính nhanh hay tính nhầm ra một đại lượng khác nếu biết các đại lượng khác khi mô tả chuỗi số.

$$g_i = \frac{\delta_i}{a_i(\%)} = \frac{Y_i - Y_{i-1}}{\frac{Y_i - Y_{i-1}}{100} 100} = \frac{Y_{i-1}}{100} \quad (14.16)$$

14.2 DỰ BÁO TRÊN CHUỖI THỜI GIAN

Người ta vẫn hay nhầm lẫn giữa dự báo và làm kế hoạch. Lên kế hoạch là một tiến trình xác định cách thức làm việc với tương lai. Một khía cạnh dự báo lại là tiến trình dự đoán xem tương lai sẽ như thế nào. Dự đoán thường được dùng như dữ liệu đầu vào cho tiến trình làm kế hoạch.

Nhu cầu về dự báo hiện nay hầu như có ở khắp các lĩnh vực, ví dụ chính phủ phải dự đoán được những vấn đề tương lai như thất nghiệp, lạm phát, sản lượng công nghiệp, các khoản thu từ thuế thu nhập của các doanh nghiệp và cá nhân để có thể xây dựng chính sách. Nhà điều hành Marketing của một tập đoàn bán lẻ phải dự đoán được nhu cầu đối với các mặt hàng, doanh số bán, thái độ của khách hàng, hàng tồn kho... để ra những quyết định đúng lúc liên quan đến những hoạt động hiện tại và tương lai và để hỗ trợ cho việc lập chiến lược kinh doanh. Ngay cả người quản lý của một hãng hàng không cũng cần dự đoán được số lượng hành khách để đảm bảo cung cấp đủ các phương tiện phục vụ cho các nhu cầu cá nhân. Và Ban giám hiệu của một trường ĐH cũng muốn dự đoán được số lượng sinh viên sẽ thi vào trường mình kết hợp với việc xem xét xu hướng của sự phát triển của công nghệ (vốn có ảnh hưởng đến chương trình học được xây dựng căn cứ trên điều kiện khoa học kỹ thuật thực tế) để có thể lên kế hoạch xây dựng ký túc xá và các phương tiện phục vụ học tập khác cũng như đảm bảo các điều kiện có liên quan đến cả nhu cầu của người dạy và người học.

Các chuyên gia đều đồng ý rằng kế hoạch tốt là điều cốt yếu để 1 tổ chức đạt được thành công. Vì dự báo là một phần quan trọng của việc lên kế hoạch nên chúng ta cần làm quen với các phương pháp dự báo. Có hai loại cơ bản của kỹ thuật dự báo: định tính và định lượng. Kỹ thuật dự báo định tính dựa trên quan điểm và phán đoán của các chuyên gia. Kỹ thuật dự báo định lượng dựa trên các phương pháp thống kê để phân tích số liệu lịch sử. Nội dung nghiên cứu về dự báo trong cuốn sách của chúng ta tập trung vào các phương pháp dự báo định lượng.

Nói chung thì phương pháp dự báo định lượng được sử dụng khi các điều kiện sau được bảo đảm: số liệu lịch sử liên quan đến vấn đề có sẵn, số liệu lịch sử đó có thể lượng hóa được, và một giả định rất quan trọng nữa là qui luật vận động của hiện tượng trong quá khứ sẽ tiếp tục trong hiện tại, điều kiện cuối cùng này được xem như là sự thừa nhận về tính liên tục, nó là một giả thuyết cơ bản của tất cả các phương pháp dự báo định lượng và nhiều phương pháp dự báo định tính. Những người chưa quen với dự báo có thể nghĩ rằng mọi thứ luôn luôn thay đổi do đó không thể dự báo một cách chính xác được tương lai dựa trên qui luật vận động của hiện tượng trong quá khứ, tuy nhiên sau khi nghiên cứu về kĩ thuật dự báo bạn sẽ thấy rằng, tuy không giữ nguyên như cũ nhưng một vài khía cạnh của quá khứ có thể lập lại, áp dụng phương pháp dự báo đúng đắn có thể giúp xác định được dạng của mối quan hệ giữa biến được dự báo và thứ tự thời gian nó vận động, hay giữa biến được dự báo với các biến khác, trên cơ sở đó có thể dự báo vấn đề thành công.

Có nhiều phương pháp dự báo định lượng khác nhau được xây dựng nhưng tất cả đều nhắm vào mục tiêu cuối cùng là dự đoán những sự kiện tương lai để những người lên kế hoạch có thể kết hợp thông tin đó vào quá trình lập kế hoạch và chiến lược. Trong các chương liên quan đến nội dung về hồi qui chúng ta biết một trong những công dụng của mô hình hồi qui được xây dựng là sử dụng để dự đoán giá trị của biến phụ thuộc khi biết được thông tin về một số biến độc lập. Phương pháp dự báo dựa trên mô hình hồi qui như vậy được gọi chung là phương pháp dự báo nhân quả. Định hướng của dự báo nhân quả là tìm ra dạng quan hệ giữa các hiện tượng và sử dụng nó vào việc ước lượng giá trị tương lai của biến cần dự báo. Trong nội dung chương này một lần nữa ý tưởng phân tích hồi qui được sử dụng phục vụ cho mục đích dự báo (trong phương pháp ngoại suy xu thế); ngoài ra ta sẽ nghiên cứu thêm một số kỹ thuật thống kê khác được áp dụng cho dữ liệu về một đối tượng nhưng được thu thập qua một thời kì liên tục, dữ liệu như vậy được gọi là dữ liệu chuỗi thời gian. (ví dụ doanh thu hàng tháng trong một năm, nhiệt độ cao nhất trong ngày được

ghi lại trong nhiều ngày liên tục, sản lượng lúa từng vụ mùa trong một giai đoạn của kế hoạch 5 năm). Bạn đọc xem Bảng 14.1 minh họa một chuỗi thời gian. Vì dữ liệu chuỗi thời gian về các hiện tượng trong đời sống kinh tế, xã hội có nhiều biến hiện khác nhau nên trong Chương 14 này chúng ta sẽ nghiên cứu nhiều phương pháp dự báo khác nhau áp dụng cho dữ liệu chuỗi thời gian. Không giống như dự báo nhân quả, dự báo chuỗi thời gian không cố gắng tìm ra các yếu tố tác động đến hành vi của hệ thống mà việc dự đoán tương lai sẽ dựa vào các giá trị quá khứ của chính biến đang cần dự báo và các sai số trong quá khứ để tìm ra kiểu thức vận động của biến trong giai đoạn đã qua và ngoại suy tiếp kiểu đó cho tương lai.

14.2.1 MỘT SỐ VẤN ĐỀ LIÊN QUAN ĐẾN DỰ BÁO

14.2.1.1 Thời đoạn dự báo

Là tần suất thời gian mà số liệu phục vụ cho dự báo được thu thập, thời đoạn dự báo có thể là tháng, quý, năm. Thời đoạn dự báo này phụ thuộc vào bản chất của đối tượng ta cần thu thập thông tin, nếu đối tượng là tỷ giá ngoại tệ thì ta phải thu thập theo ngày vì tỷ giá thay đổi từng ngày; ngược lại, nếu đối tượng là GDP của một quốc gia thì thời đoạn phải là năm vì GDP được định nghĩa theo năm. Nếu chúng ta sử dụng kỹ thuật dự báo định lượng thì các số liệu lịch sử phải ở cùng thời đoạn, giả sử bạn muốn dự báo cho tuần, bạn phải có tất cả dữ liệu lịch sử thu thập theo tuần, nếu vì những lý do khách quan khiến dữ liệu được thu thập không phải theo tuần bạn phải có những phương pháp xử lý thích hợp để chuyển dữ liệu về thời đoạn phù hợp.

14.2.1.2 Tầm xa dự báo

Là khoảng thời gian tương lai mà giá trị dự báo được thực hiện, tầm xa của dự báo có thể gồm một hoặc nhiều thời đoạn dự báo, chẳng hạn thời đoạn là tuần thì tầm xa dự báo có thể là 1 tuần mới kế tiếp, hoặc thời đoạn dự báo là $\frac{1}{2}$ tháng kế tiếp (tức chính là 2 tuần kế tiếp); nhìn chung người ta phân loại tầm xa dự báo một cách tương đối như sau:

- Dự báo ngay tức thì : tầm xa dự báo dưới 1 tháng
- Dự báo ngắn hạn : từ 1 tháng đến 3 tháng
- Dự báo trung hạn : 3 tháng đến 2 năm
- Dự báo dài hạn : 2 năm trở lên

14.2.1.3 Đánh giá độ phù hợp của mô hình dự báo

Cần nhận thức được rằng đối với một bộ dữ liệu lịch sử thu thập được liên quan đến đối tượng cần dự báo, người ta có thể vận dụng không chỉ một

mà là một vài phương pháp dự báo khác nhau để thực hiện mục tiêu dự báo tương lai. Không có phương pháp dự báo nào là hoàn hảo nhất mà tùy vào bản chất của hiện tượng, độ dài dự báo, độ dài của chuỗi thời gian, cùng với kinh nghiệm thực tế là những yếu tố cần thiết để cân nhắc xem trong từng bài toán dự báo thì mô hình dự báo nào là phù hợp hơn cả. Mức độ phù hợp này được xem xét trên khía cạnh mô hình dự báo nào cho ra các kết quả dự báo chính xác hơn, trong phần lớn tình huống sự chính xác được xem như tiêu chuẩn cơ bản để chọn lựa một phương pháp dự báo phù hợp và vì thế chúng ta có thể dùng lẫn nhau giữa hai thuật ngữ “chính xác” và “phù hợp” để chỉ việc mô hình dự báo đã xây dựng được trên dữ liệu lịch sử có thể tái tạo lại dữ liệu gần giống đến mức nào với những dữ liệu thật đã có.

Có nhiều chỉ tiêu đo lường mức độ chính xác của mô hình dự báo mà trong nội dung này ta sẽ nghiên cứu một số chỉ tiêu tiêu biểu. Các chỉ tiêu này đều xây dựng trên thông tin về sai số dự báo, ký hiệu e_t , đó là chênh lệch giữa giá trị thực tế và giá trị dự báo ở cùng thời đoạn t . Về mặt công thức nếu Y_t là ký hiệu cho quan sát thực sự và F_t là ký hiệu cho giá trị dự báo ở cùng thời kỳ thì sai số dự báo được hình thành như sau $e_t = (Y_t - F_t)$. Nếu chuỗi thời gian của bạn có n thời đoạn tức là bạn có n giá trị quan sát Y_t , thì khi xây dựng xong một mô hình dự báo nào đó, dựa trên mô hình đó bạn sẽ tính toán được n giá trị dự báo F_t và do đó suy ra được n giá trị sai số $e_t = (Y_t - F_t)$. Trên các giá trị sai số này bạn có thể tính toán các đại lượng đo lường sai số dự báo sau.

1. MAE (sai số tuyệt đối trung bình Mean Absolute Error)

Công thức tính của chỉ tiêu này như sau:

$$MAE = \frac{\sum_{t=1}^n |e_t|}{n} \quad (14.17)$$

Chú ý là khi tính các chỉ tiêu đo độ chính xác của dự báo thì các xử lý đối với e_t đều phải lấy trị tuyệt đối hoặc bình phương để tránh triệt tiêu do trái dấu vì có thể $\sum e_t = 0$

2. MAPE (sai số phần trăm tuyệt đối trung bình Mean Absolute Percent Error)

$$MPAE = \frac{\sum_{t=1}^n |e_t| / Y_t}{n} \times 100\% \quad (14.18)$$

Công thức này giúp ta khử đơn vị tính trong tử số của công thức MAE để có một đại lượng có đơn vị tính là %, giúp dễ so sánh giữa MAPE của các mô hình dự báo trên các chuỗi dữ liệu khác nhau về đơn vị tính.

3. MSE (sai số bình phương trung bình Mean Square error)

$$MSE = \frac{\sum_{t=1}^n e_t^2}{n} \quad (14.19)$$

So sánh 2 công thức MSE và MAE thì công thức MSE có nhược điểm là nó làm sai số bị bình phương lên nên giá trị cuối cùng của MSE rất lớn, tuy nhiên ưu điểm của MSE là nó giúp cho các phép toán liên quan trở nên dễ xử lý hơn so với khi dùng MAE, nên MSE thông dụng hơn.

Để khắc phục nhược điểm phỏng đại của MSE, có thể dùng chỉ tiêu RMSE = \sqrt{MSE} để cho giá trị của đại lượng tính ra không quá lớn.

Đôi khi các phần mềm thống kê, như SPSS chẳng hạn, lại không dùng MSE mà dùng SSE là tên gọi chỉ lượng ở mẫu số của công thức tính MSE ở trên như một sự thay thế cho MSE, điều này cũng dễ chấp nhận vì, với n đã biết, nếu SSE lớn thì MSE cũng lớn và ngược lại.

4. Chỉ số U

Một thước đo độ chính xác dự báo khác là hệ số không ngang bằng U, là tỷ số giữa RMSE của mô hình dự báo muốn xét và mô hình dự báo Naive (tức là mô hình dự báo thô)

$$\text{Công thức } U = \frac{\text{RMSE của mô hình dự báo đang xem xét}}{\text{RMSE của mô hình thô "Naive"}} \quad (14.20)$$

Quy tắc quyết định là $U > 1$ nghĩa là mô hình đang xét có độ phù hợp rất kém. Nếu $U < 1$ thì mô hình dự báo ta đang xét có thể sử dụng được vì nó tốt hơn mô hình dự báo thô, U càng tiến về 0 thì mô hình dự báo đang xét càng chính xác, trong thực tế giá trị của $U \leq 0,55$ thì mô hình dự báo được đánh giá là tốt.

Các chỉ tiêu làm thước đo cho sai số dự báo (từ 1 đến 3) ở phía trên chỉ có ý nghĩa khi tính toán để so sánh từ hai mô hình trở lên với qui tắc càng nhỏ càng tốt vì điều đó chứng tỏ giá trị dự báo rất sát với giá trị thực tế. Vì vậy người ta đưa ra Chỉ số U để không cần so sánh mô hình dự báo của ta với một mô hình nào phức tạp cả mà so sánh chính nó với mô hình dự báo Naive có phương pháp thực hiện rất đơn giản là sử dụng chính giá trị thực tế t làm giá trị dự báo cho thời kỳ kế tiếp ($t+1$). Diễn tả về mặt công thức là : $F_{t+1} = Y_t$.

Ví dụ: Có chuỗi thời gian Y_t và một mô hình dự báo cần xem xét độ chính xác, giả sử áp dụng mô hình dự báo đó chúng ta tính được các giá trị F_t trình bày trong cột thứ 3 của bảng dưới. Vậy giờ ta sẽ dùng phương pháp chỉ số U để quyết định có chọn dùng mô hình dự báo này hay không.

Bảng 14.1

t	Y_t	F_t	F_t (Naive)	e_t	e_t (Naive)	e_t^2	e_t^2 (Naive)
1	16	16	-	0	-	0	-
2	17	18	16	-1	1	1	1
3	20	18	17	2	3	4	9
4	22	21	20	1	2	1	4
5		24	22				
Tổng						6	14

Đầu tiên ta tính giá trị dự báo của phương pháp dự báo thô căn cứ trên chuỗi thời gian đang có, các giá trị F_t (Naive) này được tính theo công thức $F_{t+1} = Y_t$.

- F_1 (Naive) = $F_{0+1} = Y_0$ mà ta không có giá trị Y_0 nên ta không tính được F_1 (Naive)
- F_2 (Naive) = $F_{1+1} = Y_1$
- ...
- F_5 (Naive) = $F_{4+1} = Y_4$
- F_6 (Naive) = $F_{5+1} = Y_5$ mà ta không có giá trị Y_5 nên ta không tính được F_6

Như vậy là với phương pháp dự báo thô ta chỉ có thể dự báo về trước một thời đoạn, sau đó cập nhật giá trị mới nhất và dự báo tiếp

Sau đó tính RMSE của từng mô hình qua các bước sau: đầu tiên là tính e_t , sau đó tính e_t^2 , rồi tính đến MSE và sau cùng là RMSE. Vì mô hình Naive không có giá trị dự báo đầu tiên nên ta sẽ bắt đầu tính e_t từ $t = 2$. Và vì giai đoạn thứ 5 không có giá trị thực tế mà chỉ có các giá trị dự báo nên ta không có cơ sở tính e_t .

$$\text{vậy } \text{RMSE}_{DB} = \sqrt{\text{MSE}} = \sqrt{6/4} = 1,22$$

$$\text{RMSE}_{\text{Naive}} = \sqrt{\text{MSE}_{\text{Naive}}} = \sqrt{14/3} = 2,16$$

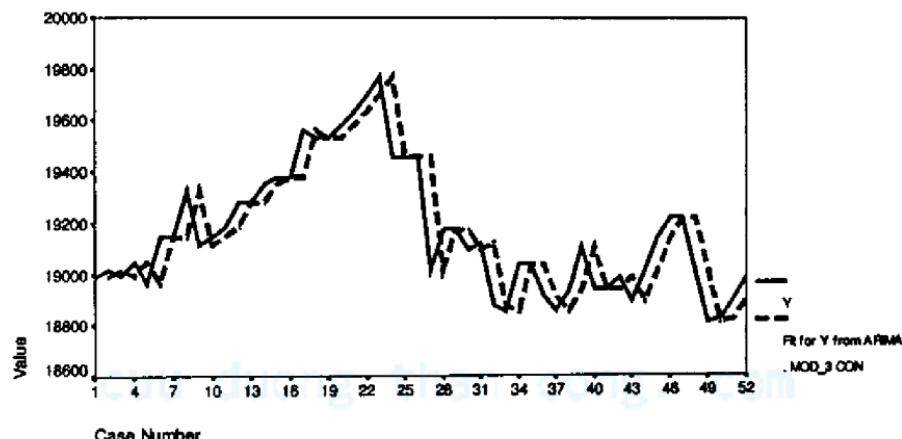
$$\rightarrow \text{Chỉ số U} \text{ được tính như sau } U = \frac{1,22}{2,16} = 0,56$$

Vì $0,55 < U = 0,56 < 1$ nên ta kết luận mô hình đang xét có thể sử dụng được do nó tốt hơn mô hình Naive (vì $U < 1$), tuy nhiên độ chính xác của nó chưa cao lắm (vì $U > 0,55$)

5 Đánh giá trực quan bằng đồ thị

Dùng đồ thị trình diễn đồng thời cả giá trị thực tế và giá trị dự báo là phương pháp nhanh nhất trong việc đánh giá mô hình dự báo phù hợp tốt đến mức độ nào dữ liệu quá khứ, tuy nhiên độ chính xác của phương pháp này chỉ tương đối do cảm nhận hoàn toàn bằng mắt, trong hình dưới đường liền nét mô tả giá trị dữ liệu thật và đường đứt nét mô tả giá trị dự báo, có thể nhận thấy đường đứt nét đi rất sát theo vận động của đường liền nét nhưng trễ đi một chút.

Hình 14.1



14.2.2 CÁC PHƯƠNG PHÁP DỰ BÁO ĐƠN GIẢN

Giả sử có dãy số thời gian: Y_1, Y_2, \dots, Y_n . Với dãy số này ta có thể dùng các phương pháp dự đoán ngắn hạn sau:

14.2.2.1 Dự đoán bằng lượng tăng (giảm) tuyệt đối trung bình

Phương pháp này thường được sử dụng khi biến động của hiện tượng có lượng tăng (giảm) tuyệt đối liên hoàn xấp xỉ nhau.

$$\hat{Y}_{n+L} = Y_n + \bar{\delta}L \quad (14.21)$$

Trong đó:

- \hat{Y}_{n+L} : giá trị dự đoán ở thời gian $n + L$
- Y_n : giá trị thực tế ở thời gian n
- $\bar{\delta}$: lượng tăng (giảm) tuyệt đối trung bình
- L : tầm xa dự đoán

14.2.2.2 Dự đoán bằng tốc độ phát triển trung bình

Phương pháp này sử dụng khi hiện tượng nghiên cứu biến động với một nhịp độ tương đối ổn định, tức là các tốc độ phát triển liên hoàn xấp xỉ

bằng nhau.

$$\hat{Y}_{n+L} = Y_n(\bar{t})^L \quad (14.22)$$

Trong đó:

- \hat{Y}_{n+L} : giá trị dự đoán ở thời gian $n + L$
- Y_n : giá trị thực tế ở thời gian n
- \bar{t} : tốc độ phát triển trung bình
- L : tầm xa dự đoán

14.2.2.3 Dự báo bằng phương pháp trung bình trượt (Moving Average)

Trung bình trượt, còn gọi là trung bình trượt, công thức của phương pháp dự báo này tính như sau:

$$F_{t+1} = \frac{Y_t + Y_{t-1} + \dots + Y_{(t-k)+1}}{k} \quad (14.23)$$

Trong đó

- Y_t là giá trị quan sát thực tế vào thời điểm t
- F_{t+1} là giá trị dự báo vào thời điểm $t+1$
- Số lượng quan sát thực tế Y_t đưa vào vị trí tử số của công thức trên đúng bằng k quan sát. Với k là kí hiệu của khoảng trượt, k được lựa chọn = 3, 4, 5, 6, 7... mà tối thiểu là 3; k bằng bao nhiêu có nghĩa là ta mất đi bấy nhiêu số quan sát đầu tiên khi tính toán dự báo.

Phương pháp này chỉ dự báo được tiếp 1 thời đoạn so với chuỗi dữ liệu đã có. Giả dụ ta có dữ liệu đến 2003 thì ta chỉ dự báo được đến 2004.

Ví dụ: ta có một chuỗi dữ liệu như sau:

t	Y_t
1	Y_1
2	Y_2
3	Y_3
4	Y_4
5	Y_5

Áp dụng phương pháp MA để dự báo các giá trị tương lai

Nếu chọn $k = 3$, sử dụng phương pháp trung bình trượt chúng ta sẽ không tính toán được từ F_1 đến F_3 do ta không có Y_0 (xem công thức của F_3)

$$F_3 = F_{2+1} = \frac{Y_2 + Y_{2-1} + Y_{2-(3-1)}}{3} = \frac{Y_2 + Y_1 + Y_0}{3}$$

Tiếp tục áp dụng công thức tính F_{t+1} để tính các giá trị dự báo còn lại, ví dụ:

$$F_6 = F_{5+1} = \frac{Y_5 + Y_{5-1} + Y_{5-(3-1)}}{3} = \frac{Y_5 + Y_4 + Y_3}{3}$$

Như đã nói phương pháp này chỉ dự báo được về trước một thời đoạn nên ta chỉ dự báo được đến F_6 chứ không có đủ dữ liệu dự báo cho F_7 trở đi (xem công thức của F_7).

$$F_7 = F_{6+1} = \frac{Y_6 + Y_{6-1} + Y_{6-(3-1)}}{3} = \frac{Y_6 + Y_5 + Y_3}{3}$$

Muốn dự báo được F_7 thì ta phải cập nhật thông tin mới nhất Y_6 và dự báo tiếp.

Ưu điểm của phương pháp dự báo này thể hiện rõ khi áp dụng cho những chuỗi dữ liệu có dao động nhiều nhưng không có tính xu thế rõ ràng vì bản chất của trung bình trượt là tính bình quân nên đã gạt bỏ những dao động bất thường để bộc lộ kiểu vận động cơ bản của hiện tượng.

Chú ý: Sự lựa chọn khoảng trượt k là bao nhiêu phụ thuộc vào đặc điểm biến động của hiện tượng và số mức độ của dãy số. Nếu biến động của hiện tượng tương đối đều đặn và mức độ của dãy số không nhiều thì có thể tính số trung bình trượt từ 3 mức độ, nếu sự biến động của hiện tượng lớn và dãy số có nhiều mức độ thì có thể tính số trung bình trượt từ 5 hoặc 7 mức độ. Độ chính xác của dự báo phụ thuộc rất lớn vào cách lựa chọn k . Thông thường nhà nghiên cứu lựa chọn k theo phương pháp thử và sai với mục đích tìm ra kết quả dự báo có sai số thấp nhất, với mỗi lựa chọn k ta sẽ có một mô hình dự báo, với mỗi mô hình ta tính các chỉ tiêu đo lường độ phù hợp của mô hình và so sánh các chỉ tiêu này của từng mô hình để chọn mô hình phù hợp nhất (là mô hình có chỉ tiêu đo lường độ phù hợp bé nhất).

Ví dụ: Bảng 14.2 thể hiện số liệu chuỗi thời gian được thu thập từ năm 1987 đến năm 2006 về sản lượng bột ngọt tiêu thụ của một công ty (đơn vị tính là tấn). Trên Bảng 14.2 đồng thời chúng ta cũng thể hiện kết quả dự báo của phương pháp dự báo trung bình trượt đơn giản với hai tình huống là khoảng trượt $k = 3$ và $k = 7$ cho cả năm 2007

Bảng 14.2

Năm	Sản lượng	$F_t(k=3)$	$F_t(k=5)$	$e_t^2(k=3)$	$e_t^2(k=5)$
1987	1587,7	-	-	-	-
1988	1558,0	-	-	-	-
1989	1752,5	-	-	-	-
1990	1407,5	1632,73	-	50730,05	-
1991	1309,9	1572,67	-	69046,32	-
1992	1424,0	1489,97	1523,12	4351,601	9824,77
1993	1676,6	1380,47	1490,38	87694,95	34677,89
1994	1936,9	1470,17	1514,1	217840	178759,84
1995	1684,7	1679,17	1550,98	30,61778	17881,04
1996	1488,0	1766,07	1606,42	77321,07	14023,30
1997	1562,2	1703,20	1642,04	19881	6374,43
1998	1618,5	1578,30	1669,68	1616,04	2619,39
1999	1686,6	1556,23	1658,06	16995,47	814,53
2000	1840,9	1622,43	1608	47727,68	54242,41
2001	1865,2	1715,33	1639,24	22460,02	51057,92
2002	1636,7	1797,57	1714,68	25878,08	6080,88
2003	1652,8	1780,93	1729,58	16418,15	5895,17
2004	1699,0	1718,23	1736,44	369,9211	1401,75
2005	1698,0	1662,83	1738,92	1236,694	1674,45
2006	1523,0	1683,27	1710,34	25685,4	35096,28
Tổng		1640,00	1641,9		
				685283,1	420424,04

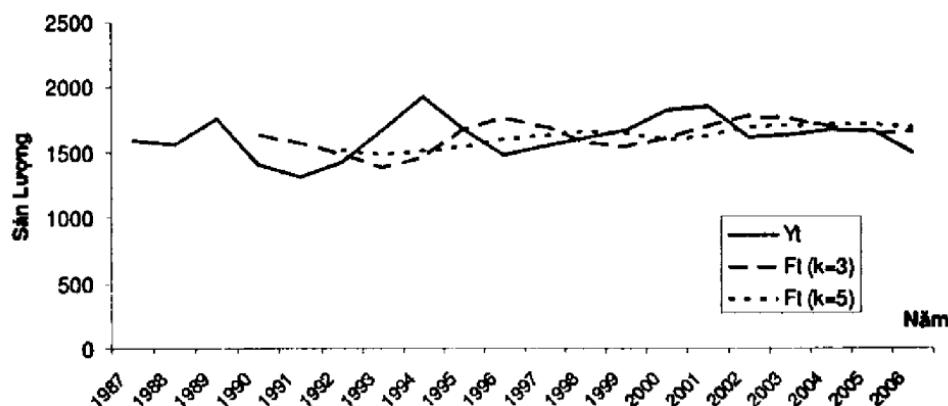
Từ kết quả tổng sai số bình phương tính được (dòng tổng của Bảng 14.2) ta xác định được MSE của hai mô hình như sau:

$$MSE_{(k=3)} = \frac{685283,1}{20} = 34264,16$$

$$MSE_{(k=5)} = \frac{420424,04}{20} = 21021,20$$

Mô hình dự báo bằng phương pháp trung bình trượt với khoảng trượt $k = 5$ tỏ ra phù hợp hơn vì nó đem lại một giá trị MSE bé hơn nên ta quyết định chọn giá trị $k = 5$. Các bạn có thể làm lại ví dụ này nhưng thử với các tinh huống k khác để xem có tinh huống nào khả quan hơn $k = 5$ hay không.

Hình 14.2



Nhìn vào 3 đường biểu diễn trên đồ thị có thể cảm nhận rõ ràng là các đường dự báo trơn tru hơn hẳn đường dữ liệu thật, và giá trị k càng lớn hơn thì đường dự báo càng trơn hơn do các dao động được san bằng nhiều hơn.

14.2.2.4 Mô hình ngoại suy xu thế

Đôi khi mô hình này còn được gọi tên là Mô hình xu hướng tuyến tính. Thay cho việc ước lượng mô hình hồi qui 2 biến quen thuộc với Y là biến độc lập và X là biến phụ thuộc $\hat{Y}_t = b_0 + b_1 X_t$, các nhà nghiên cứu thường dùng mô hình tương tự với các ký hiệu sau:

$$\hat{Y}_t = b_0 + b_1 t \quad (14.24)$$

Trong đó t được gọi tên là biến xu thế hoặc biến thời gian. Giá trị của t được thiết lập theo qui tắc t nhận giá trị 1 cho thời đoạn quan sát đầu tiên (bất kể tần suất ghi chép dữ liệu là ngày hay tháng hay năm), nhận giá trị 2 cho thời đoạn kế tiếp và cứ thế cho đến thời đoạn sau cùng.

Lúc này, thay cho việc hồi quy dữ liệu chuỗi thời gian Y_t theo một biến giải thích X nào đó, họ tính hồi quy chính Y_t theo thời gian. Thuật ngữ *xu hướng* trong tên gọi Mô hình xu hướng tuyến tính có nghĩa là một dịch chuyển đi lên hay đi xuống bền vững trong hành vi của biến số Y . Nếu hệ số độ dốc b_1 dương, Y có xu hướng đi lên, trái lại nếu hệ số độ dốc âm, Y có xu hướng đi xuống, ý nghĩa của hệ số độ dốc b_1 được giải thích là mức độ thay đổi tuyệt đối trung bình của biến số Y trong khoảng thời gian nghiên cứu. Ý tưởng của việc dùng t làm biến giải thích cho Y là ở chỗ biến xu hướng là sự thay thế cho các biến số có khả năng ảnh hưởng đến

Y. Những biến số này có thể không quan sát được, hoặc nếu có quan sát được thì dữ liệu về nó hoặc là không có sẵn, hoặc là khó mà thu thập được. Trong nhiều trường hợp, có thể tin rằng những biến ảnh hưởng đến Y này tương quan với thời gian gần đến mức ta giới thiệu biến thời gian còn dễ hơn là giới thiệu biến cơ bản.

Sự so sánh mô hình xu thế với mô hình hồi qui trình bày ở trên gợi cho chúng ta ý tưởng rằng ta có thể dùng phương pháp hồi qui với lệnh có sẵn trên phần mềm Excel để tính toán các thông số của mô hình xu thế. Đồng thời chúng ta hoàn toàn có thể vận dụng các phương pháp đánh giá chất lượng mô hình hồi qui đã nghiên cứu ở Chương 11 và 12 để kiểm tra chất lượng của mô hình xu thế này. Trong nội dung kế tiếp là Phương pháp dự báo bằng mô hình nhân chúng ta sẽ có một ví dụ cụ thể để vận dụng phương pháp mô hình ngoại suy xu thế.

14.3 DỰ BÁO BẰNG MÔ HÌNH NHÂN

Dữ liệu chuỗi thời gian về một hiện tượng hay chỉ tiêu nghiên cứu được thu thập qua thời gian có thể được xem xét là sự kết hợp của một số thành phần. Phương pháp cơ bản nhất giúp chúng ta nhận diện các bộ phận này có lẽ chính là mô hình nhân được sử dụng cho dữ liệu thu thập theo thời đoạn hàng năm, hàng quý, hàng tháng. Về cơ bản một chuỗi thời gian Y, có thể được mô tả qua các thành phần như sau

- Thành phần xu thế - Trend (Kí hiệu T_t)
- Thành phần mùa - Seasonal (Kí hiệu S_{tj})
- Thành phần chu kỳ - (Kí hiệu C_{tj})
- Thành phần bất thường (Kí hiệu E_t)

Cụ thể:

- **Thành phần xu thế:** thể hiện chiều hướng biến động tăng hoặc giảm của chỉ tiêu nghiên cứu theo một quy luật nào đó trong thời gian dài. Xu thế phản ánh sự tăng trưởng hay giảm sút dài hạn trong chuỗi thời gian do các nguyên nhân như do lạm phát, sự tăng dân số, tăng thu nhập cá nhân, sự tăng trưởng hay giảm sút của thị trường hoặc có sự thay đổi về công nghệ, lối sống, môi trường,
- **Thành phần mùa:** thể hiện biến động của chỉ tiêu nghiên cứu theo một quy luật nào đó giữa các thời điểm trong năm và lặp lại tương tự trong các năm kế tiếp. Biến động mùa được xem xét khi dữ liệu được thu thập theo tháng, quý, nếu chỉ có dữ liệu theo năm chúng ta sẽ không có biến động mùa. Ví dụ lượng tiêu thụ vở học trò sẽ

tăng mạnh vào các tháng 8, 9 tức là trong quý 3 của năm. Biến động thời vụ thường do các nguyên nhân như điều kiện thời tiết, khí hậu, tập quán xã hội, tín ngưỡng,

- Thành phần chu kỳ: thể hiện biến động của chuỗi thời gian theo một quy luật nào đó nếu xét trong một khoảng thời gian tương đối dài tính bằng năm, từ 2 đến 10 năm. Mùa và chu kỳ đều là quy luật, mùa là quy luật diễn ra giữa các thời điểm trong năm với tần suất quan sát là quý hay tháng. Còn chu kỳ là quy luật diễn ra trong khoảng thời gian dài vài năm đến chục năm, với tần suất quan sát là năm và chuỗi thời gian phải đủ dài thì mới phát hiện được quy luật chu kỳ này.
- Thành phần bất thường: Là những dao động bất thường hay những sai biệt không dự đoán được chiều hướng trong chuỗi thời gian. Nó không có sự liên kết với các thành phần mùa, chu kỳ hay xu thế có thể do ảnh hưởng của tin đồn, thiên tai, động đất, nội chiến, khủng bố, các sự kiện bất thường khác.

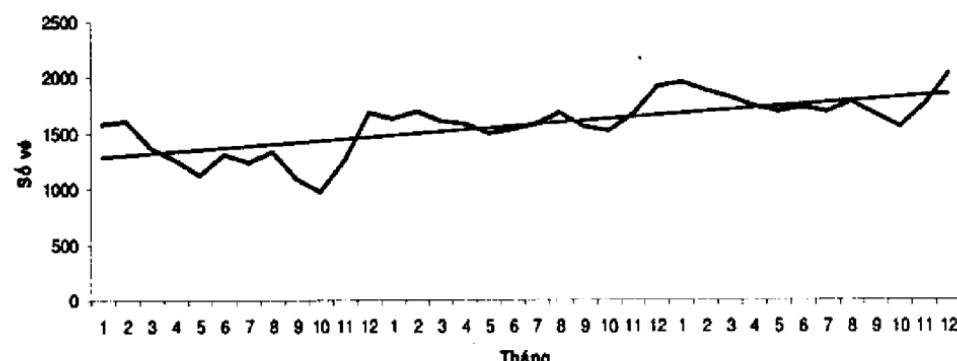
Trong nội dung sau với mô tả cụ thể và ví dụ đi kèm bạn đọc sẽ thấy phương pháp dự báo bằng mô hình nhân xử lý các thành phần của chuỗi thời gian như thế nào. Nhưng luôn nhớ rằng bước đầu tiên quan trọng của tiến trình nhận diện các thành phần của chuỗi thời gian là vẽ đồ thị mô tả chuỗi nhằm nhận diện các thành phần một cách nhanh chóng nhất.

Để minh họa phương pháp này ta khai thác ví dụ về số lượng vé bán ra trong tháng của một nhà hát. Một công ty sở hữu một chuỗi các rạp chiếu phim và sân khấu vừa mua thêm một nhà hát nữa, người quản lý của công ty đã xây dựng một tập dữ liệu về số lượng vé tiêu thụ được trong từng tháng của 3 năm vừa qua cho nhà hát vừa mua này. Ông ta dự định phân tích các số liệu này và sử dụng kết quả phân tích đó để xây dựng kế hoạch trong 12 tháng tới theo số lượng vé được dự đoán. Số liệu gốc thể hiện trong 2 cột đầu tiên của Bảng 14.8. Cột đầu tiên là thứ tự của các tháng trong từng năm và cột thứ hai là số vé bán của từng tháng. Chúng ta sẽ vẽ đồ thị minh họa dữ liệu về số vé tiêu thụ để phục vụ cho việc phân tích và nhận diện các yếu tố của chuỗi thời gian.

Xem Hình 14.3 ta thấy dữ liệu dao động có tính qui luật mùa ở chỗ cứ đến tháng 10 của từng năm lượng vé tiêu thụ giảm hẳn và số vé bán ra tăng cao trong tháng 12 sang tháng 1 của các năm. Bên cạnh đó trong từng năm số vé tiêu thụ có những dao động lên xuống bất thường nhưng xu hướng vận động cơ bản của số lượng vé tiêu thụ là tăng (nhẹ) qua 3 năm

thể hiện trực quan qua hình ảnh đường chéo kẻ xuyên qua đồ thị. Chuỗi dữ liệu chỉ có trong 3 năm nên chưa cho thấy yếu tố chu kỳ.

Hình 14.3



Chúng ta sẽ sử dụng phương pháp mô hình nhân để dự đoán số lượng vé tiêu thụ theo nguyên tắc phương pháp mô hình nhân cho rằng bất kỳ dữ liệu chuỗi thời gian nào ít nhiều cũng được cấu thành từ các thành phần đã mô tả ở trên. Do đó khi gặp một chuỗi thời gian mà đồ thị gợi ý cho bạn nhận diện các thành phần mùa, chu kỳ, xu thế, bất thường (đặc biệt là mùa) thì bạn nên nghĩ đến phương pháp dùng mô hình nhân. Khi đó một quan sát bất kỳ có thể được biểu diễn lại theo công thức sau:

$$Y_t = Tr_t \times Cl_t \times Sn_t \times E_t \quad (14.25)$$

Để dự đoán được bằng mô hình nhân cho chuỗi thời gian này ta phải tìm cách để nhận diện được các thành phần của chuỗi sau đó ráp chúng lại với nhau để có giá trị dự báo mong muốn. Từ ý tưởng này phương pháp phân tích mô hình nhân đi qua những bước sau:

Bước 1 : xác định (Tr_t, Cl_t) bằng cách tách yếu tố mùa và dao động bất thường (Sn_t, E_t) khỏi chuỗi thời gian bằng phương pháp trung bình trượt trung tâm (kí hiệu CMA) để loại (Sn_t và E_t).

Quá trình tính trung bình trượt trung tâm CMA phải thực hiện qua hai bước để đạt được CMA phù hợp.

- Đầu tiên ta tính trung bình trượt đơn giản (MA) với khoảng trượt L là số chẵn, nếu thời đoạn là tháng ta chọn $L = 12$ và nếu thời đoạn là quý thì chọn $L = 4$. Công thức tính MA như sau :

$$MA_t = \frac{Y_{t-(L/2)+1} + \dots + Y_t + \dots + Y_{t+(L/2)}}{L} \quad (14.26)$$

Phương pháp MA với L chẵn sẽ làm ta mất đi L quan sát

- Kết tiếp ta mới dùng các MA_t vừa tính được để xác định tiếp

$$CMA_t = \frac{MA_{t-1} + MA_t}{2} \quad (14.27)$$

Về mặt ý nghĩa, lúc này chuỗi CMA_t chỉ còn gồm hai yếu tố là chu kỳ và xu thế → nếu viết dưới dạng công thức thì $CMA_t = (Tr_t, Cl_t)$

Ví dụ với $L = 4$ ta tiến hành tính CMA như sau

- Đầu tiên tính MA_t (ta bắt đầu tính được từ $t = 2$, bạn đọc thử lập công thức tính MA_t với $t = 1$ xem có đủ dữ liệu để thực hiện không)

$$MA_2 = \frac{Y_{2-(4/2)+1} + Y_2 + Y_3 + Y_{2+(4/2)}}{4}$$

$$MA_3 = \frac{Y_{3-(4/2)+1} + Y_3 + Y_4 + Y_{3+(4/2)}}{4}$$

....

- Sau đó tính tiếp CMA_t (ta bắt đầu tính được từ hàng thứ 3)

$$CMA_3 = (MA_2 + MA_3)/2$$

Hãy xem trong Bảng 14.3 qui trình tính toán CMA_t cho một bộ 10 dữ liệu với $L = 4$, những ô được tô xám là những ô không tính được vì không có thông tin.

Bảng 14.3

TT	Y_t	MA_t	CMA_t
1	Y1		
2	Y2	$(Y_1+Y_2+Y_3+Y_4)/4$	
3	Y3	$(Y_2+Y_3+Y_4+Y_5)/4$	$(MA_2 + MA_3)/2$
4	Y4	$(Y_3+Y_4+Y_5+Y_6)/4$	$(MA_3 + MA_4)/2$
5	Y5	$(Y_4+Y_5+Y_6+Y_7)/4$	$(MA_4 + MA_5)/2$
6	Y6	$(Y_5+Y_6+Y_7+Y_8)/4$	$(MA_5 + MA_6)/2$
7	Y7	$(Y_6+Y_7+Y_8+Y_9)/4$	$(MA_6 + MA_7)/2$
8	Y8	$(Y_7+Y_8+Y_9+Y_{10})/4$	$(MA_7 + MA_8)/2$
9	Y9		
10	Y10		

Bước 2 : Quay lại tách (Sn_t, E_t) khỏi Y_t theo công thức sau

$$(Sn_t, E_t) = Y_t / (Tr_t, Cl_t) \quad (14.28)$$

Bản chất của công thức này là Y_t / CMA_t với CMA_t vừa tính ở Bước 1.

Bước 3 : Từ kết quả về (Sn_t, E_t) mới tách được ở bước 2 ta tiếp tục lọc yếu tố sai biệt E_t bằng cách tính trung bình cho mỗi mùa (vì trung bình bao giờ cũng san bằng những sai biệt để làm lộ ra yếu tố chính).

Cách thực hiện là lập bảng có dạng sau với 4 cột nếu mùa là quý và 12 cột nếu mùa là tháng. Với $L = 4$ ta sẽ có bảng sau.

Bảng 14.4

Quý 1	Quý 2	Quý 3	Quý 4
(S_{n_t}, E_t)
...
$(\sum (S_{n_t}, E_t))_1$	$(\sum (S_{n_t}, E_t))_2$	$(\sum (S_{n_t}, E_t))_3$	$(\sum (S_{n_t}, E_t))_4$

Ta nhận những giá trị (S_{n_t}, E_t) tương ứng thuộc quý nào đưa vào cột của quý đó, nếu với mỗi quý ta gấp m giá trị (S_{n_t}, E_t) thì ngoại trừ hàng tiêu đề và hàng tính tổng, lúc này bảng của ta sẽ có m hàng. Ta tính giá trị tổng $(\sum (S_{n_t}, E_t))_{Qi}$ của từng quý, sau đó tính chỉ số mùa theo công thức sau

$$\bar{S}_{n_{Qi}} = \frac{(\sum (S_{n_t} * E_t))_{Qi}}{m} \quad (14.29)$$

Bước 4 : Với các giá trị $\bar{S}_{n_{Qi}}$ vừa tính được từ bước 3 ta kiểm tra xem tổng chỉ số mùa $\sum (\bar{S}_{n_{Qi}})$ có bằng đúng L hay không.

Nguyên tắc là tổng các chỉ số mùa phải bằng đúng L là độ dài khoảng trượt, nếu $\sum (\bar{S}_{n_{Qi}}) \neq L$ thì ta phải thực hiện tiếp việc điều chỉnh cho tổng này = L

Bước 5 : Cách điều chỉnh là nhân từng $\bar{S}_{n_{Qi}}$ với một hệ số điều chỉnh có giá trị = $\frac{L}{\sum (\bar{S}_{n_{Qi}})}$ để bảo đảm rằng sau khi điều chỉnh thì tổng các chỉ số

mùa đã điều chỉnh này bằng đúng L, về mặt công thức ta biểu diễn như sau $\sum ((\bar{S}_{n_{Qi}})_a) = L$ với $(\bar{S}_{n_{Qi}})_a$ là chỉ số mùa đã điều chỉnh

Bước 6 : Vì hiện tượng mùa lặp lại như nhau trong các năm nên ta thế các giá trị Chỉ số mùa (kết quả bước 3) hoặc Chỉ số mùa đã điều chỉnh (kết quả bước 5) vào lại bảng dữ liệu về Y_t theo đúng thứ tự của mùa trong năm thứ nhất và lặp lại y hệt cho các năm tiếp theo.

Bước 7 : Loại bỏ mùa khỏi dữ liệu bằng công thức

$$d_t = \frac{Y_t}{Chỉ số mùa} \quad (14.30)$$

Chú ý d_t có bản chất là chuỗi Y_t lúc này đã loại bỏ tính mùa
Chỉ số mùa là các giá trị được điền vào bảng dữ liệu tại bước 6.

Bước 8 : Thực hiện dự báo d_t bằng mô hình xu thế tuyến tính và làm các kiểm định thích hợp.

Xem lại tiến trình từ bước 1 đến bước 8 thì kết quả ta có lúc này là bộ dữ liệu dạng chuỗi thời gian d_t có nguồn gốc từ chuỗi Y_t do $d_t = (\text{Tr}_t, \text{Cl}_t, E_t)$

Bước 9 : Ta sẽ dùng phương pháp ngoại suy xu thế tuyến tính cho chuỗi d_t để xác định các giá trị dự báo cho thành phần không còn tính mang mùa này. Từ mô hình dự báo xây dựng được tính giá trị dự báo \hat{d}_t .

Bước 10 Nhưng giá trị dự báo ta đang muốn tìm là giá trị dự báo F_t của đối tượng Y_t chứ không phải là \hat{d}_t , để tính được F_t ta lấy kết quả dự báo mới tính được \hat{d}_t nhân thêm với chỉ số mùa.

Bây giờ chúng ta sẽ thực hành các bước lý thuyết vừa tìm hiểu của phương pháp mô hình nhân và nghiên cứu sâu hơn về phương pháp ngoại suy xu thế với chuỗi thời gian bằng ví dụ về số vé bán ra của nhà hát.

Trên Bảng 13.8 hai cột đầu tiên mô tả thông tin về tháng và số vé bán của từng tháng như ta đã biết.

- Bước 1: Tính CMA_t

Qua phân tích đồ thị ta thấy số liệu mùa ở đây là theo tháng nên khi tính MA ta chọn $L = 12$, với $L = 12$ ta tính được từ giai đoạn $t = 6$, lúc đó MA_6 là trung bình của 12 quan sát từ Y_1 đến Y_{12} , kế tiếp MA_7 là trung bình của 12 quan sát từ Y_2 đến Y_{13} ... Cột thứ 3 của Bảng 13.8 thể hiện các giá trị MA này.

Tính các CMA_t là trung bình của các MA_t . Cột thứ 4 của Bảng 13.8 thể hiện các CMA_t giả dụ CMA_7 là trung bình của hai giá trị MA_6 và MA_7 ; CMA_8 là trung bình của hai giá trị MA_7 và MA_8 ...

- Bước 2: Tách (S_n, E_t)

Cách thực hiện là ta đem Y_t/CMA_t , vì CMA_t chỉ bắt đầu có từ tháng thứ 7 nên thành phần (S_n, E_t) tách được cũng bắt đầu tính từ tháng 7.

Cột thứ 5 của Bảng 13.8 trình bày dữ liệu về thành phần (S_n, E_t) .

- Bước 3: Tính chỉ số mùa để lọc bỏ tiếp E_t .

Muốn vậy đầu tiên ta tạo một bảng có 12 cột tương ứng với 12 tháng của năm rồi nhặt các giá trị (S_n, E_t) của từng tháng điền vào vị trí phù hợp theo thứ tự. Xem Bảng 14.5

- Giả dụ giá trị đầu tiên của cột $(S_n * E_t)$ trong Bảng 13.8 thuộc tháng thứ 7 nên ta đặt nó (0,938) vào hàng thứ nhất của cột tháng 7 trong Bảng 14.5.
- Dò tiếp ở năm thứ 2 ta có giá trị $(S_n * E_t)$ của tháng 7 là 0,964. Ta đặt nó vào hàng thứ 2 của cột tháng 7.

- Xem tiếp ở năm thứ 3 ta thấy số liệu về ($S_n * E$) chỉ có đến tháng thứ 6 chứ không còn tháng thứ 7 nên việc dò tìm chấm dứt, như vậy ở đây $m = 2$

Thực hiện tương tự cho các tháng khác ta điền đầy bảng dưới và từ đó tính được hàng tổng cộng (hàng in đậm cuối cùng)

Bảng 14.5

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
1.129	1.155	1.069	1.027	0.943	0.954	0.933	1.009	0.812	0.718	0.903	1.183
1.130	1.078	1.039	0.995	0.958	0.976	0.964	1.012	0.931	0.899	0.979	1.115
2.259	2.233	2.108	2.022	1.901	1.930	1.902	2.021	1.743	1.617	1.882	2.298

Đem các giá trị tổng cộng này chia cho m tức chia cho 2 ta được các chỉ số mùa như Bảng 14.6 dưới đây:

Bảng 14.6

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
1.130	1.117	1.054	1.011	0.951	0.965	0.951	1.011	0.872	0.809	0.941	1.149

Ý nghĩa của các chỉ số mùa này có thể hiểu như sau: chỉ số mùa của tháng 1 bằng 1,13 có nghĩa là số lượng vé tiêu thụ được trong tháng này cao hơn mức trung bình của năm khoảng 13%.

- Bước 4:* Kiểm tra tổng chỉ số mùa có bằng đúng độ dài khoảng trượt không.

Cộng các chỉ số mùa đã tính được trên Bảng 14.6 ta được kết quả tổng cộng là $11,96 < 12$ (là độ dài khoảng trượt)

- Bước 5:* Điều chỉnh chỉ số mùa

Chúng ta tiến hành điều chỉnh chỉ số mùa bằng cách nhân từng chỉ số mùa đã tính được với hệ số điều chỉnh, hệ số này được tính = $12/11,96$

Các hệ số mùa đã điều chỉnh được tính lại và trình bày ở Bảng 14.7 dưới đây. Ví dụ, chỉ số mùa đã điều chỉnh của tháng 2 được tính = $1,117 \times 12/11,96 = 1,120$

Lúc này tổng các chỉ số mùa đã điều chỉnh sẽ bằng 11,998 (xem như bằng 12 nếu làm tròn).

Bảng 14.7

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
1.133	1.120	1.058	1.014	0.954	0.968	0.954	1.014	0.874	0.811	0.944	1.153

- Bước 6:* Trên Bảng 13.8 chúng ta điền các giá trị của chỉ số mùa (đã điều chỉnh) vào cột 6 cho từng tháng tương ứng của năm thứ 1 và lặp lại y hệt cho năm thứ 2, thứ 3.

- Bước 7: Loại mùa*

Loại mùa ra khỏi dữ liệu gốc bằng cách lấy các Y_t của từng tháng chia cho chỉ số mùa của tháng đó. Chúng ta được cột d_t trên Bảng 13.8

- Bước 8: Xây dựng mô hình dự báo bằng phương pháp ngoại suy xu thế*

Trên bộ dữ liệu d_t này chúng ta sẽ áp dụng phương pháp ngoại suy xu thế tuyến tính để xây dựng mô hình dự báo của thành phần không có tính mùa.

Ta xem chuỗi d_t trong Bảng 14.8 như một chuỗi thời gian bất kỳ, ta trích riêng chuỗi này ra để nhập vào bảng tính Excel và tạo thêm một cột dữ liệu về biến t. Biến t này nhận giá trị 1 ở tháng 1 của năm đầu tiên, giá trị 2 ở tháng 2 của năm đó và cứ liên tục như thế cho đến tháng 12 của năm thứ 3 thì nhận giá trị 36. Xem bộ dữ liệu này trên Bảng 14.9, bên cạnh đó là các kết quả xử lý hồi qui xu thế cho chuỗi d_t này.

Bảng 14.8

Tháng	Số vé	MA (L=12)	CMA (Tr ^o Cl)	(S _n *E)	Chỉ Số mùa	Chuỗi loại bỏ Mùa (d _t)	Giá trị dự báo của chuỗi d _t	Giá trị dự báo của chuỗi Y _t
1	1580				1.133	1394.53	1244.00	1409.45
2	1608				1.120	1435.71	1263.05	1414.62
3	1370				1.058	1294.90	1282.11	1356.47
4	1260				1.014	1242.60	1301.16	1319.38
5	1125				0.954	1179.25	1320.21	1259.48
6	1306	1319.92			0.968	1349.17	1339.26	1296.40
7	1324.03	1322.00	0.954		0.954	1299.79	1358.31	1295.83
8	1340	1331.75	1327.92	1.009	1.014	1321.50	1377.37	1396.65
9	1090	1351.75	1341.75	0.812	0.874	1247.14	1396.42	1220.47
10	980	1379.25	1365.50	0.718	0.811	1208.38	1415.47	1147.95
11	1260	1410.33	1394.79	0.903	0.944	1334.75	1434.52	1354.19
12	1680	1429.83	1420.08	1.183	1.153	1457.07	1453.57	1675.97
1	1630	1458.17	1444.00	1.129	1.133	1438.66	1472.63	1668.49
2	1700	1486.50	1472.33	1.155	1.120	1517.86	1491.66	1670.68
3	1610	1525.67	1506.08	1.069	1.058	1521.74	1510.73	1598.35
4	1590	1570.67	1548.17	1.027	1.014	1568.05	1529.78	1551.20
5	1498	1604.83	1587.75	0.943	0.954	1570.23	1548.83	1477.58
6	1540	1624.83	1614.83	0.954	0.968	1590.91	1567.89	1517.72
7	1580	1652.33	1638.58	0.964	0.954	1656.18	1586.94	1513.94
8	1680	1667.33	1659.83	1.012	1.014	1656.80	1605.99	1628.47
9	1560	1684.83	1676.08	0.931	0.874	1784.90	1625.04	1420.28
10	1520	1698.17	1691.50	0.899	0.811	1874.23	1644.09	1333.36
11	1670	1714.17	1706.17	0.979	0.944	1769.07	1663.15	1570.01
12	1920	1730.00	1722.08	1.115	1.153	1665.22	1682.20	1939.58

1	1960	1739.17	1734.58	1.130	1.133	1729.92	1701.25	1927.52
2	1880	1747.50	1743.33	1.078	1.120	1678.57	1720.30	1926.74
3	1820	1756.67	1752.08	1.039	1.058	1720.23	1739.35	1840.23
4	1750	1760.00	1758.33	0.995	1.014	1725.84	1758.41	1783.03
5	1690	1767.50	1763.75	0.958	0.954	1771.49	1777.46	1695.70
6	1730	1777.50	1772.50	0.976	0.968	1787.19	1796.51	1739.02
7	1690				0.954	1771.49	1815.56	1732.04
8	1780				1.014	1755.42	1834.61	1860.29
9	1670				0.874	1910.76	1853.67	1620.11
10	1560				0.811	1923.55	1872.72	1518.78
11	1760				0.944	1864.41	1891.77	1785.83
12	2040				1.153	1769.30	1910.82	2203.18

Bảng 14.9

t	d _t
1	1394.53
2	1435.71
3	1294.9
4	1242.6
5	1179.25
6	1349.17
7	1299.79
8	1321.5
9	1247.14
10	1208.38
11	1334.75
12	1457.07
13	1438.66
14	1517.86
15	1521.74
16	1568.05
17	1570.23
18	1590.91
19	1656.18
20	1656.8
-	-
35	1864.41
36	1769.3

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.908
R Square	0.825
Adjusted R Square	0.820
Standard Error	93.694
Observations	36

ANOVA

	df	SS	MS	F
Regression	1	1410175.885	1410175.885	160.637
Residual	34	298473.450	8778.631	
Total	35	1708649.335		

	Coefficients	Standard Error	t Stat	P-value
Intercept	1224.949	31.894	38.407	1.36E-29
t	19.052	1.503	12.674	1.95E-14

- Viết lại mô hình hồi qui xu thế xây dựng được cho mục đích dự báo: $\hat{d}_t = 1224.949 + 19.052t$
- Khả năng giả thích của mô hình này là 82,5%
- Giá trị p – value = 1,95E-14 cho thấy biến t có mối liên hệ với Y_t một cách có ý nghĩa.
- Giá trị của hệ số độ dốc bằng 19,052 cho biết trong khoảng thời

gian 3 năm đang nghiên cứu, số vé tiêu thụ được mỗi tháng tăng trung bình, 19 vé.

- Ta cũng có thể thực hiện các kiểm định về tự tương quan, phương sai thay đổi trên mô hình này...
- *Bước 9* : Tính giá trị dự báo cho chuỗi d_t .

Trên mô hình hồi qui xây dựng được $\hat{d}_t = 1224,949 + 19,052 * t$ chúng ta lần lượt thay thế các thứ tự của t vào để xác định giá trị dự báo.

Ví dụ tại tháng 2 của năm thứ 2 thì t nhận giá trị 14, thay thế giá trị 14 này vào phương trình $\hat{d}_t = 1224,949 + 19,052 \times 14 = 1491,68$

- *Bước 10*: Dự báo F_t .

Để tính được F_t ta lấy kết quả dự báo mới tính được \hat{d}_t , nhân thêm với chỉ số mùa.

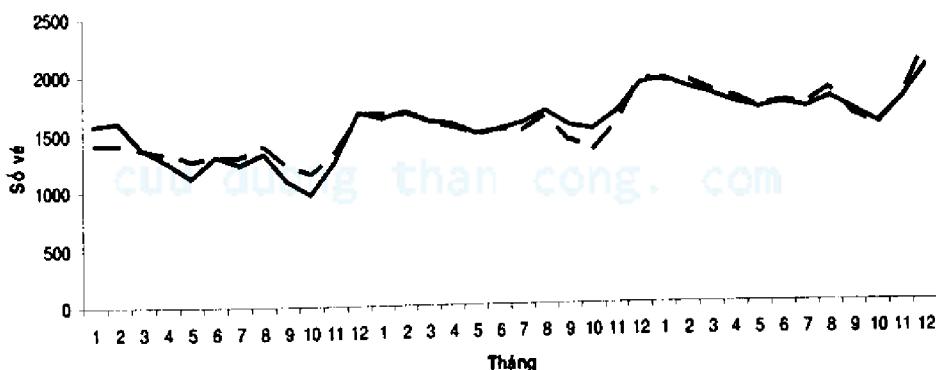
Ví dụ: Muốn dự báo về số vé bán được ở tháng 1 năm thứ tư ta phải làm 2 bước như sau:

- Tính d_{37} ở thời đoạn 37 $\rightarrow \hat{d}_{37} = 1224,949 + 19,052 * 37 = 1929,87$
- Tính $F_{37} = 1929,873 \times 1,133 = 2186,54$ vé

Tiến hành tương tự cho các tháng còn lại để xác định các giá trị dự báo cần tìm.

Đánh giá sơ bộ về mức độ phù hợp của mô hình dự báo bằng cách vẽ đồ thị giá trị dự báo (đường đứt nét trong Hình 14.4) xem nó trùng với đường dữ liệu thực tế (đường liền nét) đến mức nào.

Hình 14.4



Nếu bạn tiến hành lại ví dụ trên nhưng lấy nhiều số thập phân sau dấu phẩy hơn thì mức độ chênh lệch giữa hai đường thực tế và dự báo sẽ giảm đi.

14.4 DỰ BÁO BẰNG HÀM TĂNG TRƯỞNG MŨ

Hàm tăng trưởng mũ phản ánh chiều hướng của chuỗi thời gian sẽ coi như luôn tăng (hay giảm) ở một tỷ lệ không đổi, mô hình này phản ánh chính xác một số tình huống kinh tế và kinh doanh trong thế giới thực.

Để minh họa ta xem xét tình huống của Western Steakhouses, một chuỗi cửa hàng thức ăn nhanh khai trương năm 1978. Mỗi năm từ 1978 đến 1992, số cửa hàng hoạt động trong chuỗi được ghi lại. Một nhà phân tích của công ty muốn sử dụng số liệu này để dự báo số cửa hàng thuộc chuỗi sẽ hoạt động vào năm 1993. Xem dữ liệu ở Bảng 14.10

Mô hình tăng trưởng mũ có dạng phương trình:

$$Y_t = a \times e^{bt} \quad (14.31)$$

với $e = 2,71828$.

Với giai đoạn $(t-1)$ ta có $Y_{t-1} = a \times e^{b \times (t-1)}$

$$\text{Tỷ lệ } \frac{Y_t}{Y_{t-1}} = e^{bt - b(t-1)} = e^b = \text{const}$$

Ở mọi giai đoạn của chuỗi thời gian, tỷ lệ của quan sát sau trên quan sát trước luôn là e^b . Nên mô hình này hay áp dụng cho những chuỗi thời gian có tốc độ tăng trưởng gần như không đổi theo thời gian.

Bảng 14.10

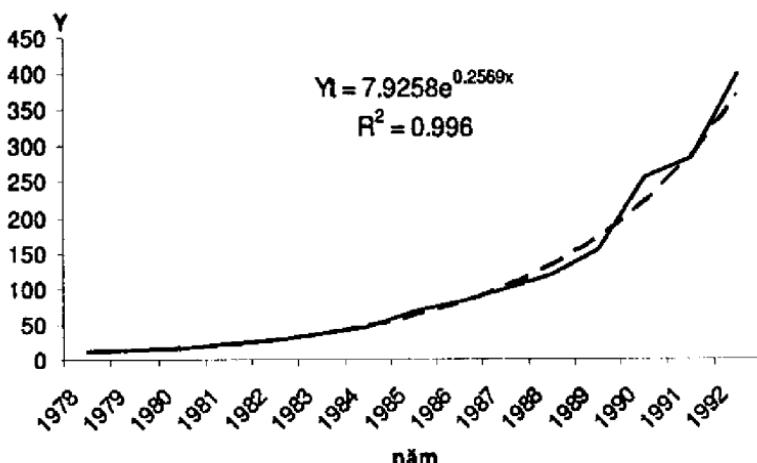
Năm	t	Y_t	Y_t/Y_{t-1}	Năm	t	Y_t	Y_t/Y_{t-1}
1978	1	11	-	1986	9	82	1.22
1979	2	14	1.27	1987	10	99	1.21
1980	3	16	1.14	1988	11	119	1.20
1981	4	22	1.38	1989	12	156	1.31
1982	5	28	1.27	1990	13	257	1.65
1983	6	36	1.29	1991	14	284	1.11
1984	7	46	1.28	1992	15	403	1.42
1985	8	67	1.46				

Khi xác định xu thế của một chuỗi thời gian có tuân theo quy luật tăng trưởng mũ hay không, ngoài phương pháp nghiên cứu đồ thị (xem dạng đồ thị ở Hình 14.5) có một cách bổ sung để quyết định lựa chọn dạng hàm, đó là phương pháp phân tích tỷ số Y_t/Y_{t-1} của chuỗi dữ liệu ban đầu, nếu các tỷ số đều có giá trị xấp xỉ bằng nhau thì ta dùng dạng hàm tăng trưởng mũ.

Với ví dụ của ta, tỷ lệ Y_t/Y_{t-1} tính được trên dữ liệu không chênh lệch nhau là mấy, bên cạnh đó khi đưa dữ liệu lên đồ thị Hình 14.5, đường dữ

liệu thật (đường đứt nét) cho thấy sự gia tăng của dữ liệu theo hình dạng của hàm tăng trưởng mũ. Nên ta quyết định dùng hàm tăng trưởng mũ để mô hình hóa sự vận động của chuỗi thời gian này.

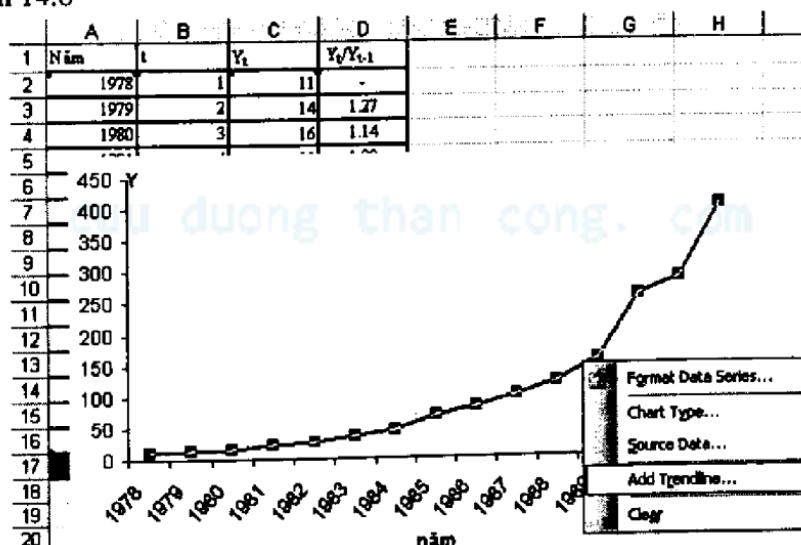
Hình 14.5



Để thể hiện được đường xu thế cơ bản dạng hàm tăng trưởng mũ (đường liền nét) trên đồ thị cũng như tính toán được phương trình hàm tăng trưởng mũ ta có thể sử dụng phần mềm Excel để thực hiện 1 cách nhanh chóng như sau

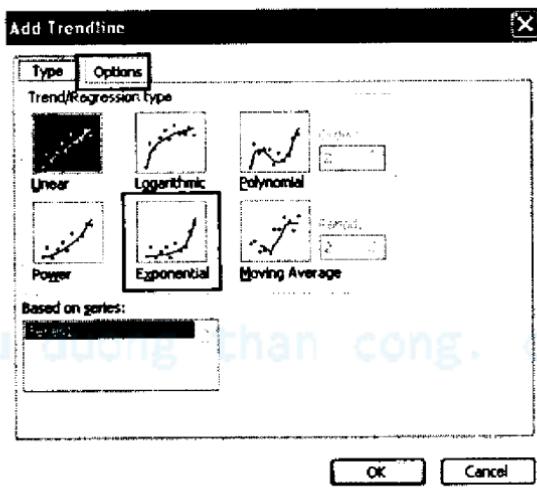
- Bước 1 vẽ đồ thị chuỗi thời gian Y_t theo năm hoặc theo biến thời gian Trên đồ thị này ta bấm chọn sáng đường thể hiện dữ liệu rồi nhấp chuột phải lấy lệnh Add Trendline như hình dưới

Hình 14.6

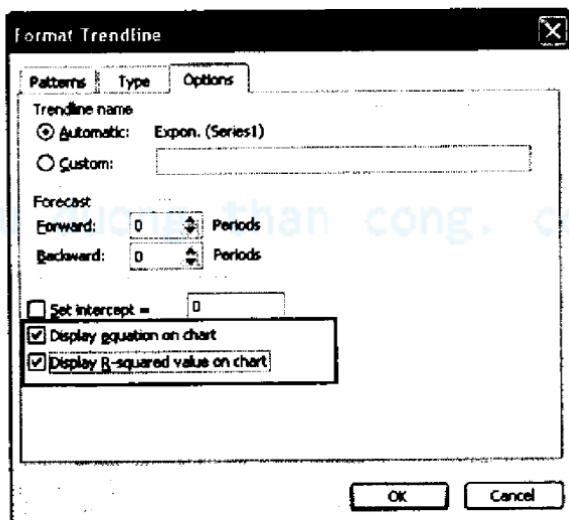


- **Bước 2** Lệnh này mở ra cửa sổ Add trenline, trên cửa sổ này nhấp chọn ô Exponential, rồi sau đó nhấp phiếu Option (xem Hình 14.7)
- **Bước 3** Vào trong thẻ Option này chọn tiếp 2 mục Display Equation on Chart và Display R-squared value on Chart (xem Hình 14.8)
- **Bước 4** Nhấp nút OK trên các cửa sổ để có được đồ thị và phương trình thể hiện như Hình 14.5

Hình 14.7



Hình 14.8



Ta viết lại phương trình dự báo bằng hàm tăng trưởng mũ trong Hình 14.5 nhưng thay kí hiệu X bằng t vì X là kí hiệu chung của Excel dành cho biến độc lập trong mô hình. Với ví dụ của chúng ta thực ra biến độc lập là biến thời gian t. Kí hiệu \hat{Y}_t là của giá trị dự báo.

$$\hat{Y}_t = 7,9258 \cdot e^{0,2569 \cdot t}$$

$$R^2 = 0,996$$

Độ phù hợp của mô hình tăng trưởng mũ này khá cao, tới 99,6%. Ở đây hệ số b = 0,2569. Cần chú ý nghĩa thực sự của b đó là *tốc độ tăng trưởng trung bình hàng năm* (hoặc tháng, quý là tùy theo tần suất quan sát) của đối tượng kinh tế trong giai đoạn có dữ liệu, chú ý là b không đổi trong suốt khoảng thời gian ta nghiên cứu, và b càng lớn thì độ dốc của hàm tăng trưởng mũ càng lớn. Vậy trung bình mỗi năm số cửa hàng trong chuỗi gia tăng với tốc độ khoảng 25,7%.

Muốn tính toán giá trị dự báo cho năm 1993 ta thay thế giá trị t = 16 vào hàm:

$$\hat{Y}_{16} = 7,9258 \cdot e^{0,2569 \cdot 16} = 483,245 \text{ cửa hàng}$$

Nếu cần đánh giá chất lượng của mô hình dự báo này, bạn có thể dùng phương pháp trung gian như sau

- Từ phương trình $Y_t = a \times e^{b \times t}$ bạn lấy log của hai vế:
 $\ln(Y_t) = \ln(a) + \ln(e^b \cdot t)$
- Biến đổi tiếp bạn sẽ được kết quả $\ln(Y_t) = \ln(a) + b \cdot t$
- Nếu bạn đặt $\ln(Y_t)$ như một biến Y'_t và $\ln(a)$ như một hệ số tung độ gốc b_0 thì bạn viết lại được phương trình $Y'_t = b_0 + b \cdot t$ là một mô hình hồi qui xu thế thông thường, lúc này bạn hoàn toàn có thể thực hiện các đánh giá chất lượng mô hình trên hàm hồi qui trung gian này.

14.5 DỰ BÁO BẰNG SAN BẰNG HÀM SỐ MŨ

Là một ứng dụng mở rộng của phương pháp trung bình trượt, với phương pháp trung bình trượt với khoảng trượt là k thì giá trị trung bình của k quan sát quá khứ được sử dụng để dự báo giá trị hiện tại, trong phương pháp đó ngũ ý rằng k điểm dữ liệu quá khứ đều tham gia vào việc tính giá trị tương lai với trọng số như nhau là 1/k. Tuy nhiên với mục đích dự báo thì những quan sát mới nhất thường cung cấp những gợi ý tốt hơn về tương lai vì thế chúng ta cần gán trọng số phù hợp vào các quan sát trong chuỗi nhằm giảm tầm ảnh hưởng đối với những quan sát cũ hơn và tận

dụng tốt hơn thông tin trên những quan sát mới nhất, thay vì các quan sát đều có cùng một trọng số là $1/k$.

Trong nội dung này chúng ta nghiên cứu một loạt các phương pháp dự báo căn cứ trên nguyên tắc gán trọng số giảm theo kiểu hàm số mũ cho những quan sát cũ hơn vì thế chúng ta gọi là thủ tục san bằng hàm số mũ. Có nhiều phương pháp san bằng hàm số mũ, tất cả chúng đều có một thuộc tính chung là những giá trị gần hiện tại nhất được gán trọng số đáng kể hơn những quan sát ở xa.

14.5.1 San bằng hàm mũ đơn giản

14.5.1.1 Lý thuyết về dự báo bằng phương pháp san bằng hàm mũ đơn giản

Giả sử bạn muốn dự báo giá trị kế tiếp trong chuỗi thời gian Y_t đã quan sát được. Giá trị dự báo của chúng ta được ký hiệu là F_t . Khi giá trị quan sát thực tế Y_t có sẵn thì sai số dự báo được xác định bởi công thức $e_t = Y_t - F_t$. Phương pháp san bằng hàm mũ đơn giản tạo ra giá trị dự báo bằng cách lấy giá trị dự báo ở thời kì trước điều chỉnh đi một lượng sai số dự báo, công thức cho giá trị dự báo tương lai là:

$$F_{t+1} = F_t + \alpha(Y_t - F_t) \quad (14.32)$$

Trong đó

F_{t+1} là giá trị dự báo ở giai đoạn $t+1$

Y_t là giá trị thực tế ở giai đoạn t

α là trọng số với tính chất $0 < \alpha < 1$, nhưng không bằng đúng 0 hoặc 1

Nhìn công thức (14.32) có thể nhận ra rằng giá trị dự báo mới đơn giản là bằng giá trị dự báo cũ cộng với một lượng điều chỉnh liên quan đến sai số dự báo đã xảy ra ở lần dự báo kế trước đó. Lượng điều chỉnh dự báo theo hướng nếu giá trị thực tế $>$ giá trị dự báo (tức độ lớn của sai số ở giai đoạn t dương) thì giá trị dự báo ở giai đoạn $t+1$ sẽ được điều chỉnh tăng, và ngược lại; nếu giai đoạn t xảy ra sai số dự báo âm cho dự báo cao hơn thực tế thì dự báo ở giai đoạn t phải được điều chỉnh giảm đi.

Phương trình (14.32) được biến đổi thành một dạng khác như sau

$$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t \quad (14.33)$$

Trong (14.33) giá trị dự báo được tính căn cứ trên giá trị thực tế gần nhất được lấy trọng số là α và giá trị dự báo gần nhất được lấy trọng số là $(1 - \alpha)$.

Khi giá trị của α gần 1 thì giá trị dự báo mới bao gồm một phần đáng kể của lượng điều chỉnh, ngược lại nếu α gần 0 thì giá trị dự báo mới sẽ cần

rất ít lượng điều chỉnh và có thể xem như gần bằng giá trị dự báo ở giai đoạn sát trước. Như vậy tương tự như k trong phương pháp trung bình trượt, ta nên chọn α gần 0 nếu dãy số có nhiều biến đổi bất thường, ngược lại nên chọn gần 1 nếu muốn kết quả dự báo kết hợp với những thay đổi gần nhất trong số liệu.

Công thức (14.32) hay (14.33) đều cho thấy rằng khi sử dụng phương pháp San bằng hàm mũ đơn giản để dự báo cho giai đoạn $t+1$ chúng ta chỉ cần các giá trị thực tế và giá trị dự báo gần nhất cùng với giá trị α chứ không cần làm việc với tất cả chuỗi dữ liệu quá khứ.

Hàm ý của phương pháp san bằng hàm số mũ đơn giản có thể được nhận thấy rõ ràng hơn nếu bạn thay thế F_t trong (14.33) bằng các thành phần tương ứng như sau

$$\begin{aligned} F_{t+1} &= \alpha * Y_t + (1 - \alpha) * [\alpha * Y_{t-1} + (1 - \alpha) * F_{t-1}] \\ &= \alpha * Y_t + (1 - \alpha) * \alpha * Y_{t-1} + (1 - \alpha)^2 * F_{t-1} \end{aligned}$$

Tiến trình này tiếp tục bằng cách thay thế F_{t-1} bằng lượng

$$F_{t-1} = \alpha * Y_{t-2} + (1 - \alpha) * F_{t-2}$$

$$\begin{aligned} Vào F_{t+1} &= \alpha * Y_t + (1 - \alpha) * \alpha * Y_{t-1} + (1 - \alpha)^2 * [\alpha * Y_{t-2} + (1 - \alpha) * F_{t-2}] \\ &= \alpha * Y_t + (1 - \alpha) * \alpha * Y_{t-1} + (1 - \alpha)^2 * \alpha * Y_{t-2} + (1 - \alpha)^3 * F_{t-2} \end{aligned}$$

Tiếp tục thay thế F_{t-2} bằng lượng $F_{t-2} = \alpha * Y_{t-3} + (1 - \alpha) * F_{t-3}$ vào kết quả trên

...
Cuối cùng kết quả khai triển trên toàn bộ chuỗi thời gian là

$$\begin{aligned} F_{t+1} &= \alpha Y_t + \alpha(1 - \alpha) Y_{t-1} + \alpha(1 - \alpha)^2 Y_{t-2} + \alpha(1 - \alpha)^3 Y_{t-3} + \alpha(1 - \alpha)^4 Y_{t-4} + \\ &\quad \alpha(1 - \alpha)^5 Y_{t-5} + \dots + \alpha(1 - \alpha)^{t-1} Y_1 + (1 - \alpha)^t F_1 \end{aligned} \quad (14.34)$$

Từ kết quả khai triển (14.34) ta thấy rõ ràng F_{t+1} là trung bình trượt có trọng số của tất cả các quan sát quá khứ mặc dù trong thể hiện của công thức (14.32) hay (14.33) ta chỉ thấy giá trị thực và giá trị dự báo trước đó một giai đoạn, nhưng thực ra tất cả các thông tin quá khứ đã bao gồm trong đó rồi.

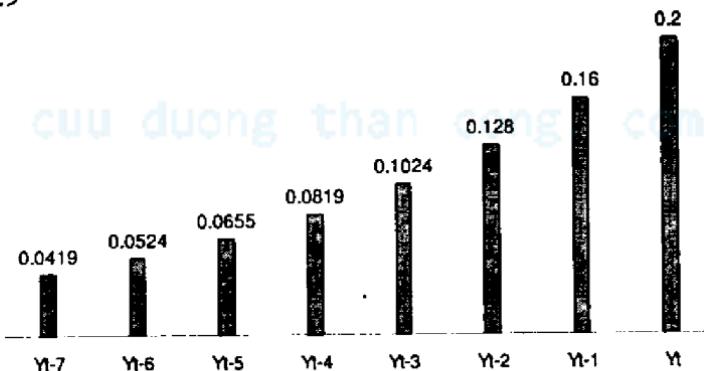
Giả sử bạn chọn $\alpha = 0,2$ và bạn có 8 quan sát trên chuỗi thời gian, thì các giá trị trọng số tính cho các quan sát từ Y_1 đến Y_{t-7} sẽ được xác định như sau (bạn đang ở thời điểm hiện tại là t và quá khứ gồm 7 quan sát từ Y_{t-1} đến Y_{t-7})

Bảng 14.11

STT	Quan sát	Trọng số
1	Y_t	$\alpha = 0,2$
2	Y_{t-1}	$\alpha(1-\alpha) = 0,2*0,8 = 0,16$
3	Y_{t-2}	$\alpha(1-\alpha)^2 = 0,2*0,8^2 = 0,128$
4	Y_{t-3}	$\alpha(1-\alpha)^3 = 0,2*0,8^3 = 0,1024$
5	Y_{t-4}	$\alpha(1-\alpha)^4 = 0,2*0,8^4 = 0,08192$
6	Y_{t-5}	$\alpha(1-\alpha)^5 = 0,2*0,8^5 = 0,065536$
7	Y_{t-6}	$\alpha(1-\alpha)^6 = 0,2*0,8^6 = 0,0524288$
8	Y_{t-7}	$\alpha(1-\alpha)^7 = 0,2*0,8^7 = 0,041943$

Nếu biểu diễn các giá trị trọng số này dưới dạng hình ảnh có thể nhận thấy rằng chúng giảm dần từ thời đoạn t đến $t-7$ theo dạng hàm mũ vì thế chúng được gọi là san bằng hàm mũ. Xem Hình 14.9.

Hình 14.9



Để vận dụng phương pháp san bằng hàm mũ đơn giản chúng ta sử dụng lại chuỗi thời gian được thu thập từ năm 1987 đến năm 2006 về sản lượng bột ngọt tiêu thụ của một công ty (đơn vị tính là tấn) như Bảng 14.12 sau:

Bạn có thể tìm các giá trị dự báo của phương pháp san bằng hàm mũ theo cả hai công thức (14.32) hoặc (14.33). Ở đây ta chọn sử dụng công thức (14.33). Muốn bắt đầu việc tính toán bạn cần có các giá trị khởi đầu F_1 và Y_1 cũng như α , có các giá trị này bạn sẽ bắt đầu tính được F_2 .

Giá trị F_1 người ta thường qui ước chọn bằng chính Y_1 hoặc bằng trung bình cộng của tất cả các quan sát trong chuỗi thời gian hoặc trung bình của 4 hay 5 giá trị dự báo ban đầu. Ở đây ta dùng phương án chọn $F_1 = Y_1$ với Y_1 đã có sẵn, nó chính là giá trị thực tế.

Giá trị α được chọn ngẫu nhiên cho ví dụ này là bằng 0,1 (nội dung kế tiếp chúng ta sẽ bàn về cách chọn α phù hợp). Với các giá trị khởi đầu vừa lựa chọn trên ta tính được kết quả sau

Bảng 14.12

Năm	Thời đoạn	Sản lượng	Giá trị F_t	e^2
1987	1	1587,7	1587,7	0
1988	2	1558	1587,7	882,09
1989	3	1752,5	1584,73	28146,77
1990	4	1407,5	1601,507	37638,72
1991	5	1309,9	1582,106	74096,27
1992	6	1424	1554,886	17131,06
1993	7	1676,6	1541,797	18171,82
1994	8	1936,9	1555,277	145635,8
1995	9	1684,7	1593,44	8328,451
1996	10	1488	1602,566	13125,3
1997	11	1562,2	1591,109	835,7372
1998	12	1618,5	1588,218	916,987
1999	13	1686,6	1591,246	9092,312
2000	14	1840,9	1600,782	57656,77
2001	15	1865,2	1624,794	57795,25
2002	16	1636,7	1648,834	147,2392
2003	17	1652,8	1647,621	26,82417
2004	18	1699	1648,139	2586,87
2005	19	1698	1653,225	2004,815
2006	20	1523	1657,702	18144,73
		Tổng		492363,825

Tính toán ta có:

$$F_1 = Y_1 = 1587,7$$

$$F_2 = F_{1+1} = \alpha x Y_1 + (1 - \alpha) x F_1 = 0,1 \times 1587,7 + (1 - 0,1) \times 1587,7 = 1587,7$$

$$F_3 = F_{2+1} = \alpha x Y_2 + (1 - \alpha) x F_2 = 0,1 \times 1558 + (1 - 0,1) \times 1587,7 = 1584,73$$

$$F_4 = F_{3+1} = \alpha x Y_3 + (1 - \alpha) x F_3 = 0,1 \times 1752,5 + (1 - 0,1) \times 1584,73 = 1601,507$$

...

$$F_{20} = F_{19+1} = \alpha x Y_{19} + (1 - \alpha) x F_{19} = 0,1 \times 1698 + (1 - 0,1) \times 1653,225 = 1657,7025$$

Tương tự các bạn có thể tính giá trị dự đoán cho đến tận thời đoạn 20 là khi còn giá trị thực tế để tham chiếu. Vậy nếu muốn dự đoán được sản lượng tiêu thụ cho năm 2007 tức là thời đoạn 21 bạn tiếp tục thay thế các giá trị phù hợp vào công thức:

$$F_{21} = F_{20+1} = \alpha x Y_{20} + (1 - \alpha) x F_{20} = 0,1 \times 1523 + (1 - 0,1) \times 1657,702 = 1644,232$$

Sau đó tính toán giá trị sai số dự báo để làm căn cứ tính tiếp các chỉ tiêu đo lường độ chính xác của mô hình, ở đây ta dùng chỉ tiêu MSE để đánh giá độ phù hợp của mô hình (xem Bảng 14.12 cột cuối cùng)

$$MSE = \sum(e^2)/20 = 492363,825/20 = 24618,19$$

Trong tiến trình làm việc thực tế, người ta thường chọn một số trọng số α theo kinh nghiệm hay những thông tin tiên nghiệm rồi sau đó thử xem giá trị α nào cho ra chỉ tiêu MSE bé nhất thì giá trị đó được chọn để xây dựng mô hình dự báo cuối cùng. Thủ tướng tương với chỉ 3 giá trị α tùy chọn, bạn phải lặp đi lặp lại một khối lượng tính toán không nhỏ để chọn ra giá trị phù hợp cuối cùng, mà vẫn chưa chắc mình đã chọn được trọng số có giá trị khả thi nhất trên cơ sở nó xây dựng cho chúng ta một mô hình có MSE bé nhất vì có thể xảy ra vô vàn tình huống về α trong khoảng giá trị từ 0 đến 1. Tuy nhiên nếu có sẵn chương trình máy tính người ta có thể ra lệnh cho máy tính chạy tự động để chọn giá trị α khả thi nhất với thời gian rất ngắn. Sử dụng Excel bạn có thể dùng Tool có tên Solver để giải quyết vấn đề.

14.5.1.2 Dùng Excel để thực hiện phương pháp san bằng hàm mũ đơn giản

Muốn sử dụng Excel để giải quyết ví dụ san bằng hàm mũ đơn giản cho chuỗi thời gian về doanh số công ty ở trên trước tiên bạn phải thiết lập một bảng tính phù hợp cho ví dụ của chúng ta rồi sau đó mới sử dụng chức năng Solver của Excek để dò tìm tự động.

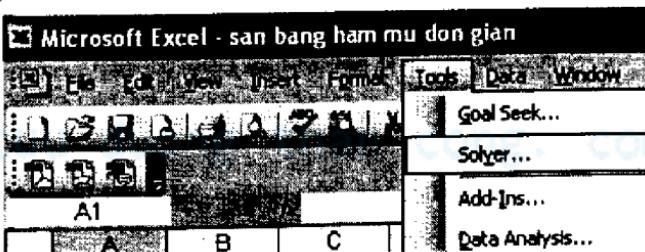
- Thiết lập bảng tính
 - Trên worksheet của Excel tiến hành nhập dữ liệu về số thời đoạn, năm, doanh số trong phạm vi từ ô N1 đến P21 (tham khảo Hình 14.12)
 - Tại cột Q và R tạo nhãn f và e^2 để xác định đây là cột chứa giá trị dự báo bằng phương pháp san bằng hàm mũ và sai số bình phương của phương pháp đó.
 - Tại ô T1 bạn nhập giá trị α đã chọn ngẫu nhiên là 0,1 và tại ô S1 bạn nhập dòng chú giải là “alpha”
 - Chúng ta qui ước chọn giá trị dự báo đầu tiên F1 bằng Y1, nên tại ô Q2 bạn nhập dòng =P2 rồi nhấn nút Enter.
 - Tại ô R2 bạn nhập công thức tính toán giá trị sai số bình phương như sau =(P2-Q2)^2, sau đó nhấn enter

- Để bắt đầu tính toán các giá trị dự báo, bạn trở lại ô Q3 nhập dòng công thức $=\$T\$1*P2+(1-\$T\$1)*Q2$. Nhập xong công thức bạn nhấn Enter rồi rê chuột kéo hết cột Q cho tới dòng 21 để điền đầy đủ các giá trị dự báo tính được bằng phương pháp san bằng hàm mũ đơn giản với trọng số là 0,1
- Sau khi có các F_t, bạn sang ô R2 rê chuột kéo hết cột R cho tới dòng 21 để điền đầy các giá trị sai số bình phương.
- Tại ô T2 bạn nhập công thức tính MSE như sau $=SUM(R2:R21)/20$ rồi nhấn Enter
- Tại ô S2 nhập dòng chú giải “MSE” cho kết quả này. Trên Hình 14.12 bạn có thể thấy hàng loạt các kết quả về Ft, e² và MSE không khác biệt với kết quả chúng ta đã tính thủ công.
- Trên bảng tính đã thiết lập, gọi chức năng Solver như thế nào?

Bạn vào menu Tools tìm chức năng Solver (xem Hình 14.10). Nếu bạn chưa thấy chức năng Solver hiện trong thực đơn Tools trên máy tính của bạn thì hãy quay lại bổ sung chức năng này vào Excel, các bước để bổ sung chức năng Solver như sau

- Vào Tools/Add – Ins bạn mở ra cửa sổ Add-Ins
- Trong danh sách các chức năng tại mục Add-Ins available bạn nhấp chọn chức năng Solver Add-in nằm ở cuối danh sách, sau đó nhấp nút OK
- Lúc này trở lại menu Tools bạn sẽ thấy có chức năng Solver hiện lên như Hình 14.10, hãy bấm vào đó để mở cửa sổ Solver Parameters

Hình 14.10

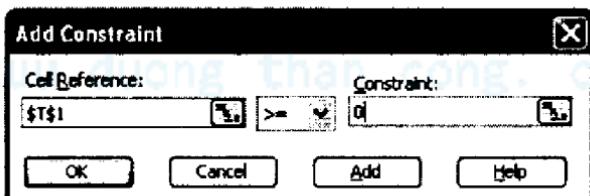


- Trên cửa sổ Solver bạn lần lượt khai báo các thông tin sau
 - Set Target Cell: là nơi để nhập địa chỉ của hàm mục tiêu, bạn đang muốn dò tìm giá trị α tốt nhất trên cơ sở nó xây dựng một mô hình dự báo có MSE bé nhất, vậy MSE là hàm mục

tiêu của bạn, công thức tính MSE nằm tại ô T2 nên bạn nhấp vào mũi tên ở cuối ô Set Target Cell để quét địa chỉ T2 vào

- Equal to: cho biết trạng thái mà hàm mục tiêu của bạn muốn đạt tới là Max, Min hay bằng một giá trị cụ thể nào đó. Ở đây tình trạng bạn muốn đạt được là Min.
- By Changing Cell: là nơi nhập vào địa chỉ chứa các biến của bài toán cần giải, ở đây biến bạn muốn tìm là α , giá trị tạm chọn của nó thuộc ô T1 nên bạn đưa địa chỉ T1 vào By Changing Cell.
- Subject to the constraints là nơi nhập vào các ràng buộc của bài toán, bạn không tìm giá trị α bất kỳ mà bạn chỉ tìm các $0 < \alpha < 1$, muốn nhập các ràng buộc ấy bạn bắt đầu bằng cách (chú ý lúc này bạn chưa nhập một ràng buộc nào nên khung Subject to the constraints hoàn toàn trống trơn chứ chưa có nội dung gì bên trong đâu) nhấp nút Add bên cạnh Subject to the constraints để mở ra cửa sổ Add Constraint

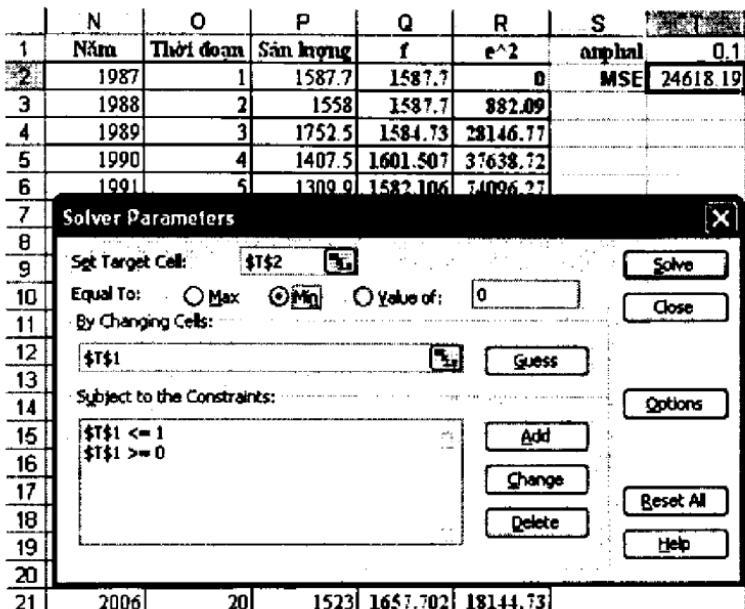
Hình 14.11



Trên cửa sổ Add Constraint tiến hành:

- ✓ Nhập địa chỉ ô T1 vào khung Cell Reference
- ✓ Nhập dấu $>=$ bằng cách nhấp vào mũi tên Drop-down để chọn lựa
- ✓ Nhập số 0 vào khung Constraint
- ✓ Nhấp nút Add để cập nhật ràng buộc vừa thiết lập vào danh sách
- ✓ Lặp lại lần nữa việc nhập T1 vào khung Cell Reference, chọn dấu $<=1$, nhập số 1 vào khung Constraint, rồi nhấp nút Add để cập nhật ràng buộc vào danh sách

Hình 14.12

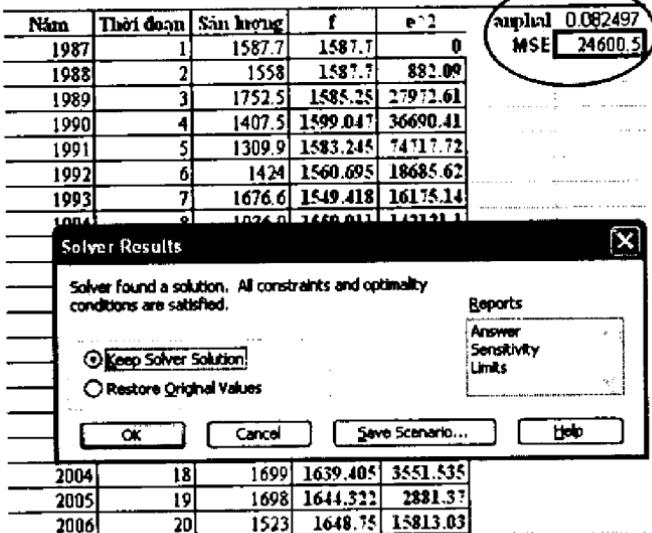


Sau khi hoàn tất những khai báo này bạn nhấp nút Cancel trên cửa sổ Add Constraint để trở lại cửa sổ Solver Parameters, lúc này nó có thể hiện hoàn chỉnh như Hình 14.12 trên, bạn nhấp nút Solver để hoàn tất công việc, lúc đó bạn sẽ được nhắc nhớ là lưu lại giải pháp tối ưu mà Solver tìm ra hay khôi phục giá trị ban đầu (xem Hình 14.13). Bạn có thể xem trước giải pháp tối ưu về α và MSE tương ứng của giải pháp trên nền worksheet (khu vực nằm trong hình ovan), bạn chọn lựa mục Keep Solver Solution như Hình 14.13 dưới đây rồi nhấp OK.

Như vậy là lệnh Solver đã giúp chúng ta tìm được giá trị α tối ưu nhất có thể sử dụng cho ví dụ về doanh số của chúng ta là 0,082, với giá trị đó bạn sẽ đạt được MSE min là 24600,5.

Cách thức hoạt động của Solver là nó thay đổi giá trị của các biến tại By Changing Cell cho đến lúc giá trị hàm mục tiêu tại Set target Cell đạt một giá trị qui định đã chọn tại Equal to, đồng thời phải thỏa mãn tập các ràng buộc tại Subject to the constraints

Hình 14.13



Bạn thử tính lại những giá trị F_t , e^2 , MSE bằng phương pháp thủ công rồi so sánh với các kết quả có thể nhìn thấy ở Hình 14.13.

Xác định giá trị dự báo cho thời đoạn 21 như sau:

$$F_{21} = F_{20+1} = \alpha Y_{20} + (1 - \alpha)F_{20} = 0,082 \times 1523 + (1 - 0,082) \times 1648,75 = 1638,439$$

Chú ý là:

- Phương pháp San bằng hàm số mũ đơn giản chỉ có thể giúp dự báo về sau một thời kỳ, với khoảng dự báo xa hơn xem như hàm dự báo là “phẳng” tức là $F_{t+h} = F_{t+1}$ ($h = 2,3\dots$). Hàm dự báo “xem như là “phẳng” vì phương pháp san bằng hàm mũ áp dụng tốt nhất cho dữ liệu không có xu hướng, chu kỳ, và không có một kiểu mẫu cơ bản nào trong dữ liệu.
- Một vấn đề nữa là nếu có tính xu hướng tăng trong chuỗi dữ liệu những giá trị dự báo bằng phương pháp san bằng hàm mũ đơn giản sẽ thiếu tính xu hướng nên “tụt hậu” so với giá trị thực tế, giá trị α càng bé sự “tụt hậu” lại càng xa (do sự góp mặt của giá trị hiện tại Y_t vào giá trị dự báo F_{t+1} càng yếu). Như vậy trong một chuỗi thời gian có xu hướng tăng các giá trị dự báo sẽ luôn thấp hơn thực tế dẫn đến các sai số dự báo luôn luôn dương, vì vậy khi chuỗi có tính xu hướng thì phương pháp san bằng hàm mũ phải được điều chỉnh đi, lúc đó ta có phương pháp Holt.

14.5.2 Phương pháp Holt

Fương pháp Holt là sự mở rộng phương pháp san bằng hàm mũ đơn giản để cho phép dự báo với dữ liệu có tính xu hướng. Trong thực tế dữ liệu không chuyển vận một cách đơn giản để có thể chỉ áp dụng phương pháp san bằng hàm mũ đơn được, khi tính xu hướng xuất hiện trong dãy số mà vẫn áp dụng phương pháp san bằng hàm mũ đơn giản sẽ cho ra sai số dự báo rất lớn. Nếu dùng phương pháp Holt trong đó tính xu hướng đã được nhận diện và điều chỉnh thì kết quả sẽ tốt hơn. Muốn tính giá trị dự báo cho phương pháp Holt ta dùng tới 2 hằng số san bằng là α và β (có giá trị biến động trong khoảng 0 đến 1) và 3 phương trình để có thể phân tích riêng được tính xu hướng trong dữ liệu như sau:

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (14.35a)$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \quad (14.35b)$$

$$F_{t+m} = L_t + b_t m \quad (14.35c)$$

Trong đó

- L_t đại diện cho ước lượng của mức độ được san bằng của chuỗi thời gian tại thời điểm t
- b_t đại diện cho ước lượng của độ dốc của chuỗi thời gian tại thời điểm t .
- Còn α là hằng số mũ để san bằng và β là hằng số mũ cho ước lượng xu hướng.
- Kí hiệu m là số giai đoạn dự báo trong tương lai.

Trong 3 công thức trên

- Công thức thứ nhất điều chỉnh trực tiếp giá trị L_t theo xu hướng của thời kỳ trước là b_{t-1} bằng cách cộng thêm vào b_{t-1} một giá trị được san bằng ở thời kỳ trước là L_{t-1} . Điều này giúp loại trừ được sự tụt hậu và đưa L_t đến mức xấp xỉ giá trị hiện tại.
- Công thức thứ 2 cập nhật cho xu hướng, xu hướng được thể hiện như sự khác biệt giữa hai giá trị được san bằng gần nhất. Điều này là phù hợp vì nếu có xu hướng trong dữ liệu thì giá trị mới cần thể hiện được là cao hơn hoặc thấp hơn giá trị trước. Cách nói khác là vì có thể tồn tại một vài sự ngẫu nhiên nên xu hướng cần được điều chỉnh bằng cách xét đến lượng chênh lệch trong mức độ trước đó $\beta (L_t - L_{t-1})$, và cộng thêm giá trị ước lượng của xu hướng trước đó nhân với $(1-\beta)$.
- Về cơ bản phương pháp dự báo bằng hàm mũ Holt này tương đương với dạng của phương pháp đơn giản ở phần trước nhưng được cập nhật thêm lượng xu hướng.

- Cuối cùng công thức thứ 3 ở trên được sử dụng để dự báo tiếp, xu hướng b_1 được nhân với số thời kì về sau định dự báo là m và cộng vào với giá trị cơ bản là L_1

Ví dụ: Để vận dụng phương pháp này ta dùng dữ liệu về nhu cầu một loại hàng hóa (đvt: tấn) được theo dõi qua 24 thời đoạn (tháng) trong bảng sau:

Bảng 14.13

Thời kì	Giá trị quan sát Y_t	L_t	b_t	Giá trị dự báo F_t ($m=1$)	e^2
1	143	143	9	-	-
2	152	152	9	152,000	0,000
3	161	161	9	161,000	0,000
4	139	154,4690	7,8818	170,000	961,000
5	137	149,6500	6,9673	162,351	642,661
6	174	165,3261	7,5943	156,617	302,157
7	142	157,4293	6,4790	172,920	956,071
8	141	152,4312	5,6526	163,908	524,788
9	162	160,0458	5,7939	158,084	15,336
10	180	172,9340	6,3047	165,840	200,513
11	164	171,6041	5,7550	179,239	232,219
12	171	174,1732	5,5256	177,359	40,438
13	206	192,8757	6,4743	179,699	691,753
14	193	196,1687	6,2453	199,350	40,323
15	207	204,7116	6,4107	202,414	21,032
16	218	214,5680	6,6588	211,122	47,303
17	229	225,1212	6,9392	221,227	60,422
18	225	228,5231	6,6845	232,060	49,849
19	204	219,5726	5,5588	235,208	973,918
20	227	226,0676	5,6262	225,131	3,492
21	223	227,3382	5,3126	231,694	75,582
22	242	237,3347	5,6498	232,651	87,408
23	239	240,9883	5,5061	242,985	15,877
24	266	256,2667	6,2097	246,494	380,468
Tổng					6322,609

- Muốn tính toán các giá trị dự báo của Holt cần phải có các giá trị xuất phát, có một vài cách xác định những giá trị này

— Người ta thường qui ước chọn

$$L_1 = Y_1 = 143$$

$$b_1 = Y_2 - Y_1 = 152 - 143 = 9$$

— Cũng còn cách xác định khác là $b_1 = (Y_4 - Y_1)/3$

— Một phương án khác nữa là hồi qui tuyến tính theo biến thời gian một số giá trị quan sát thực tế đầu tiên của chuỗi để xác định L_1 và b_1 ,

- Dùng thủ tục Solver để tìm các giá trị trọng số α và β tối ưu. Lúc này bạn có tối hai giá trị trọng số cần tìm nhưng thủ tục Solver cũng không vì thế mà phức tạp thêm bao nhiêu:
- Đầu tiên bạn cũng thiết lập một worksheet với các trọng số được chọn ngẫu nhiên là $\alpha = 0.1$ và $\beta = 0.1$ (tham khảo Hình 14.14 dưới)
- Nhập các giá trị khởi động cho cột L_t là $=B2$ và b_t là $=B3-B2$
- Nhập giá trị dự báo cho cột F_t (bắt đầu từ dòng E3) là $=C2+D2*1$
- Nhập giá trị sai số bình phương cho cột F bắt đầu từ dòng thứ 3 là $=(B3-E3)^2$
- Lần lượt nhập tại ô H1 và ô H2 các giá trị trọng số alpha, beta đều là 0,1
- Bắt đầu tại ô C3 bạn nhập công thức sau $=$H$1*B3+(1-$H$1)*(C2+D2)$ để xác định các L_t
- Tại D3 bạn nhập công thức sau $=$H$2*(C3-C2)+(1-$H$2)*D2$ để xác định các b_t
- Kéo rẽ các công thức đã tạo xuống dọc theo các cột để điền đầy các cột dữ liệu về L_t , b_t , F_t và e_t .
- Tại H3 nhập công thức tính MSE là $=SUM(F3:F25)/23$
- Trên worksheet này bạn gọi hàm Solver rồi khai báo như hình trên. Sau khi kết thúc lệnh Solver chúng ta được kết quả tối ưu là $\alpha = 0.501$ và $\beta = 0.072$

Hình 14.14

A	B	C	D	E	F	G	H
Thời gian	Giá trị quan sát y_t	L_t	b_t	Giá trị dự báo F_t ($m=1$)	e^2	alpha	
1							0.1
2	1	143	143	9			0.1
3	2	152	152	9	152.000		
4	3	161	161	9	161.000		
5	4	139	166.9000	8.6900	170.000	0.000	590.8526
6	5	137	171.7310	8.3041	175.590	0.000	961.000
7	6						1489.168
8	7						
9	8						
10	9						
11	10						
12	11						
13	12						
14	13						
15	14						
16	15						
17	16						
18	17						
19	18	225	232.3884	5.6609	233.209	67.394	

Solver Parameters

Set Target Cell:

Equal To: Max Min Value of:

By Changing Cells:

Subject to the Constraints:

Hình dưới đây thể hiện kết quả mà Solver tìm được, chú ý khu vực nằm trong vòng tròn.

Hình 14.15

	A	B	C	D	E	F	G	H
1	Thời kì quan sát Y_t	Giá trị L _t	b _t	Giá trị dự báo F _t (m=1)	e^2	alpha		
2	1	143	143	9				
3	2	152	152	9	152.000			
4	3	161	161	9	161.000			
5	4	139	154.4666	7.8770	170.000			
6	5	137	149.6446	6.9589	162.344			

Với các trọng số san bằng α và β được Solver chọn trên cơ sở tối thiểu hóa MSE cụ thể là $\alpha = 0,501$ và $\beta = 0,072$ chúng ta lần lượt tính toán lại các giá trị dự báo từ F₂ trở đi bằng phương pháp thủ công. Các kết quả này được tổng hợp trên Bảng 14.13

F₂ = ?

$$F_2 = F_{1+1} = L_1 + b_1 * I = 143 + 9 * 1 = 152$$

F₃ = ?

$$L_2 = \alpha Y_2 + (1-\alpha)(L_1 + b_1) = 0,501 * 152 + (1-0,501) * (143+9) = 152$$

$$b_2 = \beta(L_2 - L_1) + (1-\beta)b_1 = 0,072 * (152-143) + (1-0,072) * 9 = 9$$

$$F_3 = F_{2+1} = L_2 + b_2 * I = 152 + 9 * 1 = 161$$

F₄ = ?

$$L_3 = \alpha Y_3 + (1-\alpha)(L_2 + b_2) = 0,501 * 161 + (1-0,501) * (152+9) = 161$$

$$b_3 = \beta(L_3 - L_2) + (1-\beta)b_2 = 0,072 * (161-152) + (1-0,072) * 9 = 9$$

$$F_4 = F_{3+1} = L_3 + b_3 * I = 161 + 9 * 1 = 170$$

F₅ = ?

$$L_4 = \alpha Y_4 + (1-\alpha)(L_3 + b_3) = 0,501 * 139 + (1-0,501) * (161+9) = 154,47$$

$$b_4 = \beta(L_4 - L_3) + (1-\beta)b_3 = 0,072 * (154,47-161) + (1-0,072) * 9 = 7,88$$

$$F_5 = F_{4+1} = L_4 + b_4 * I = 154,47 + 7,88 * 1 = 162,35$$

...

Tính toán tương tự ta có được F₂₃ = 242,97

F₂₄ = ?

$$L_{23} = \alpha Y_{23} + (1-\alpha)(L_{22} + b_{22}) =$$

$$0,501 * 239 + (1-0,501) * (237,33 + 5,64) = 240,98$$

$$b_{23} = \beta(L_{23} - L_{22}) + (1-\beta)b_{22} = 0,072 * (240,98 - 237,33)$$

$$+ (1-0,072) * 5,64 = 5,5$$

$$F_{24} = F_{23+1} = L_{23} + b_{23} * I = 240,98 + 5,5 * 1 = 246,48$$

F₂₅ = ?

$$L_{24} = \alpha Y_{24} + (1-\alpha)(L_{23} + b_{23}) = \\ 0,501 * 266 + (1-0,501) * (240,98 + 5,5) = 256,26$$

$$b_{24} = \beta(L_{24} - L_{23}) + (1-\beta)b_{23} = 6,20$$

$$F_{25} = F_{24+1} = L_{24} + b_{24} * 1 = 256,26 + 6,2 * 1 = 262,46$$

Giá trị MSE = $\sum (e^2)/23 = 6322,609/23 = 274,896$

Chú ý là bắt đầu tính từ giá trị dự báo ở thời đoạn 26 trở đi chúng ta phải căn cứ trên giá trị dự đoán ở thời đoạn 24 vì ta không còn những giá trị thực tế nữa, với F₂₆ ta có m=2, với F₂₇ ta có m=3...

F₂₆ = ?

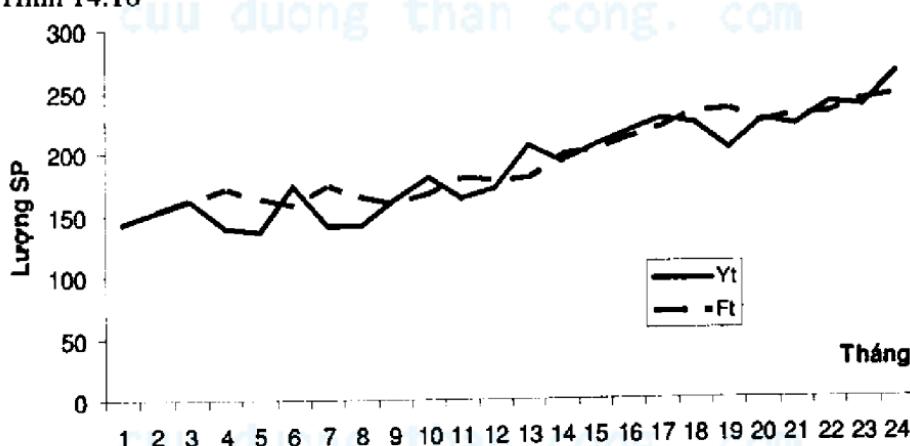
$$F_{26} = F_{24+2} = L_{24} + b_{24} * 2 = 256,26 + 6,2 * 2 = 268,66$$

F₂₇ = ?

$$F_{27} = F_{24+3} = L_{24} + b_{24} * 3 = 256,26 + 6,2 * 3 = 274,86$$

Xem Hình 14.16 mô tả đồng thời giá trị thực tế và giá trị dự báo, chúng ta nhận thấy đường mô tả giá trị dự báo đã bắt kịp xu hướng trong giá trị thực tế của chuỗi thời gian.

Hình 14.16



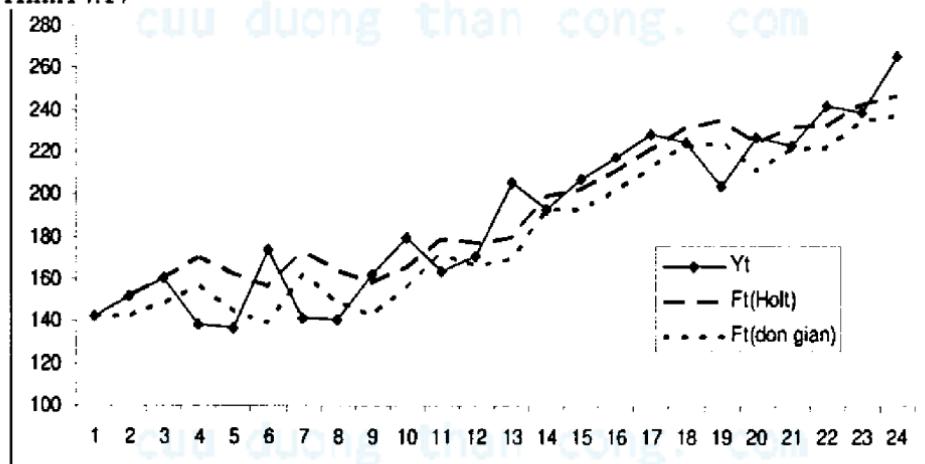
Chú ý là:

- Nếu bạn dùng phương pháp chọn lựa giá trị đầu tiên L₁ và b₁ theo kiểu hồi qui xu thế trên các giá trị thực tế ban đầu cho ví dụ của chúng ta, nếu dữ liệu vận động ổn định thì không có vấn đề gì, nhưng hãy khảo sát các quan sát đầu tiên trong bảng số liệu của chúng ta bạn sẽ thấy rằng xu thế tăng từ quan sát thứ 1 đến quan sát thứ 3 bị “hụt” tại quan sát thứ 4, nếu bước hụt này được bao gồm trong tiến trình xác định giá trị độ dốc ban đầu của chúng ta, nó có thể khiến cho hệ thống dự báo phải mất một thời

gian dài để khắc phục ảnh hưởng của chỉ một thay đổi giảm trong khi toàn bộ xu hướng là tăng này.

- Cũng như phương pháp san bằng hàm mũ đơn giản, các trọng số α và β được chọn qua quá trình cực tiểu hóa MSE hoặc một số điều kiện khác. Các cặp kết hợp bất kỳ của α và β được hình thành để tính toán các giá trị dự báo rồi sau đó cặp kết hợp cho MSE nhỏ nhất sẽ được chọn.
- Nếu so sánh kết quả của phương pháp Holt với kết quả tìm được bằng phương pháp san bằng hàm mũ đơn giản với cùng một chuỗi thời gian về nhu cầu một loại hàng hóa ở ví dụ này. Kết quả của phương pháp Holt tốt hơn, điều này không có gì đáng ngạc nhiên, Holt được thiết kế để có thể xử lý được xu thế trong khi san bằng hàm mũ đơn giản lại giả định rằng chuỗi bằng phẳng (không có xu thế). Bạn đọc thử dùng Solver tìm kết quả dự báo theo phương pháp san bằng hàm mũ đơn giản với $F_1 = Y_1$ sẽ nhận được $\alpha = 0,65$ và các giá trị dự báo tương ứng. Về đồng thời giá trị dự báo này lên đồ thị để có sự so sánh về mức độ phù hợp của 2 phương pháp. Xem Hình 14.17
- Phương pháp Holt có khi còn được gọi là phương pháp san bằng hàm mũ kép (double exponential smoothing).

Hình 14.17



14.5.3 Phương pháp Holt – Winter

Phương pháp san bằng hàm số mũ Holt chúng ta vừa nghiên cứu ở trên sẽ không còn phù hợp nếu dữ liệu của chúng ta ngoài xu thế còn có thêm tính mùa. Vì thế một nhà nghiên cứu tên là Winter đã tiếp tục mở rộng phương pháp Holt thành một mô hình có 3 tham số và 4 phương trình, được gọi tên là Mô hình Holt-Winter.

$$L_t = \alpha(Y/S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (14.36a)$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \quad (14.36b)$$

$$S_t = \gamma(Y/L_t) + (1-\gamma)S_{t-s} \quad (14.36c)$$

$$F_{t+m} = (L_t + b_t * m)S_{t-s+m} \quad (14.36d)$$

Trong công thức

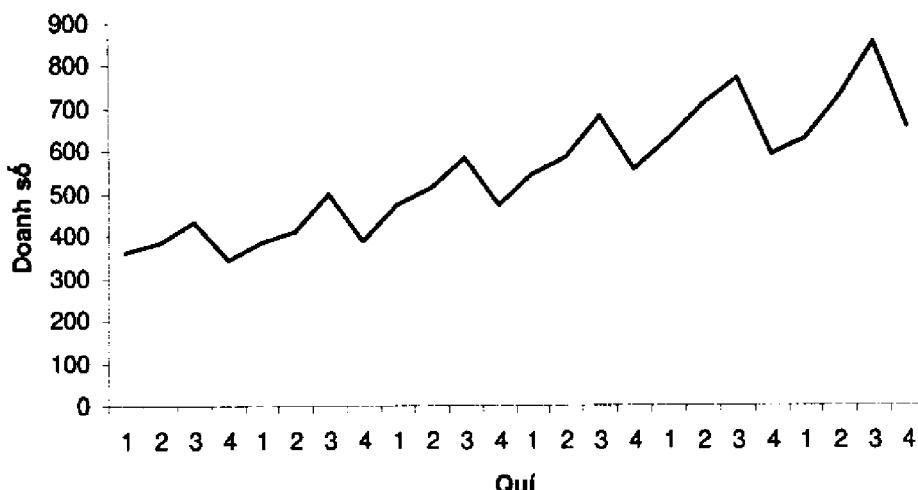
- s là số lượng giai đoạn trong một vòng thời vụ (giả dụ 4 quý trong một năm hay 12 tháng trong năm)
- L_t đại diện cho mức độ của chuỗi thời gian
- b_t đại diện cho xu hướng
- S_t là bộ phận mùa vụ
- F_{t+m} là giá trị dự báo cho m thời đoạn về sau.

Trong công thức của phương pháp Holt-Winter

- Công thức thứ 3 giống như chỉ số mùa mà được tính như tỷ lệ của giá trị hiện tại của chuỗi, Y_t , chia cho giá trị được san bằng hiện tại L_t , nếu $Y_t > L_t$ thì tỉ lệ này lớn hơn 1, và ngược lại. Điểm mấu chốt để hiểu được phương pháp này là nhận thức được L_t là giá trị được san bằng (trung bình) của chuỗi không còn bao gồm tính mùa vụ trong đó (điều này đồng nghĩa với việc nói rằng lúc này dữ liệu đã được điều chỉnh tính mùa). Một khía cạnh khác giá trị gốc của dữ liệu Y_t còn bao gồm tính mùa. Cũng cần nhớ rằng Y_t bao gồm cả những yếu tố ngẫu nhiên. Để san bằng được yếu tố ngẫu nhiên này thì công thức thứ 3 gán trọng số γ cho nhân tố mùa mới nhất vừa tính được và chỉ số mùa gần nhất tương ứng với một vòng lặp mùa được gán với trọng số $(1-\gamma)$. (chỉ số mùa gần nhất này được tính trong thời đoạn $t-s$, với s là chiều dài của mùa)
- Công thức thứ 2 giống hệt công thức xử lý tính xu thế trong phương pháp Holt.
- Còn công thức thứ nhất thì hơi khác ở chỗ thành phần thứ nhất được đem chia cho chỉ số mùa S_{t-s} , việc làm này gọi là loại trừ sự dao động mùa khỏi Y_t . Sự điều chỉnh này có thể được minh họa bằng cách xem xét tình huống $S_{t-s} > 1$ (tình huống này xảy ra khi giá trị thực tế tại giai đoạn $t-s$ lớn hơn giá trị trung bình mùa). Đem Y_t chia cho một con số lớn hơn 1 ta được một giá trị nhỏ hơn giá trị gốc. Sự điều chỉnh theo hướng ngược lại xảy ra khi chỉ số mùa bé hơn 1. Ở đây giá trị S_{t-s} được sử dụng chứ không phải S_t vì lúc này S_t chưa thể tính được cho đến khi biết được L_t (phụ thuộc vào công thức đầu tiên).

Để minh họa phương pháp Holt – Winter chúng ta xem xét một bộ dữ liệu có tính mùa là tình hình xuất khẩu qua các quý của của một công ty (xem xét bằng doanh số), dữ liệu được lưu trữ qua 6 năm trong Bảng 14.14 và được biểu diễn hình ảnh như sau.

Hình 14.18



Ta nhận thấy ngoài xu thế thì chuỗi thời gian trong ví dụ này còn có tính mùa vụ mà đỉnh mùa rơi vào quý 3 của mỗi năm.

Tại Bảng 14.14, tại cột 5 và 6 trình bày kết quả dự báo và sai số của phương pháp hàm số mũ đơn giản (bạn có thể đạt được kết quả này nếu nhờ Excel chạy hàm Solver với giá trị $F_1 = Y_1$, sẽ nhận được $\alpha = 0,464$ và các giá trị dự báo tương ứng). Xem trong dãy sai số này ta thấy có một tính hệ thống là cứ 4 sai số dương thì lại xảy ra một sai số âm (ngoại trừ tại thời điểm 21 có một sai số âm xảy ra không theo quy luật là do yếu tố ngoại lệ). Nếu thử dùng phương pháp Holt bạn cũng sẽ vẫn thấy một tính hệ thống trong dãy sai số (không trình bày ở đây), điều đó chứng tỏ cần phải có một phương pháp phù hợp để loại tính hệ thống trong phần dư, tức là phải xử lý được tính mùa trong dữ liệu. Đó là phương pháp Holt – Winter.

Bảng 14.14

Năm	Quí	Thời đoạn	Doanh số (ngàn\$)	Phương pháp đơn giản		Phương pháp Holt - Winter				
				F _t	e _t	L _t	b _t	S _t	F _t	b _t
1	1	1	362	362,00	0	-	-	0,953	-	-
	2	2	385	362,00	23,00	-	-	1,013	-	-
	3	3	432	372,67	59,33	-	-	1,137	-	-
	4	4	341	400,20	-58,20	380	9,75	0,897	-	-
2	1	5	382	372,73	9,27	399,004	10,247	0,953	371,432	10,568
	2	6	409	377,03	31,97	404,662	10,001	1,013	414,571	-5,571
	3	7	498	391,87	106,13	434,132	11,047	1,137	471,471	26,529
	4	8	387	441,11	-54,11	433,713	10,431	0,897	399,325	-12,325
3	1	9	473	416,00	57,00	487,688	12,770	0,953	423,269	49,731
	2	10	513	442,45	70,55	505,430	13,037	1,013	506,965	6,035
	3	11	582	475,19	106,81	512,965	12,742	1,137	589,498	-7,498
	4	12	474	524,75	-50,75	527,978	12,864	0,897	471,559	2,441
4	1	13	544	501,20	42,80	565,864	14,208	0,953	515,422	28,578
	2	14	582	521,06	60,94	575,449	13,960	1,013	587,614	-5,614
	3	15	681	549,34	131,66	597,366	14,387	1,137	670,157	10,843
	4	16	557	610,43	-53,43	619,435	14,800	0,897	548,743	8,257
5	1	17	628	585,64	42,36	654,876	15,909	0,953	604,426	23,574
	2	18	707	605,29	101,71	693,434	17,126	1,013	679,506	27,494
	3	19	773	652,49	120,51	684,942	15,749	1,137	807,906	-34,906
	4	20	592	708,40	-116,40	666,718	13,924	0,897	628,520	-36,520
6	1	21	627	654,39	-27,39	661,684	12,906	0,953	648,652	-21,652
	2	22	725	641,68	83,32	708,891	14,749	1,013	683,359	41,641
	3	23	854	680,34	173,66	746,553	15,980	1,137	822,778	31,222
	4	24	661	760,92	-99,92	741,144	14,830	0,897	683,992	-22,992

Bảng dữ liệu trên cho thấy với phương pháp Holt-Winter trong phần dư không còn tính hệ thống nữa.

Để thực hiện phương pháp Holt-Winter chúng ta cần các giá trị khởi đầu của L_t, S_t và b_t. Mà muốn xác định giá trị ước lượng ban đầu của chỉ số mùa chúng ta cần sử dụng ít nhất là một bộ dữ liệu hoàn chỉnh về mùa, ví dụ nếu mùa là quý thì cần 4 quan sát, như vậy ta xuất phát tại thời đoạn thứ s = 4. Bởi vậy cho nên chúng ta cũng phải lấy giá trị ban đầu của xu hướng và mức độ bắt đầu tại thời đoạn s = 4. Cách thức cụ thể như sau:

— Xác định giá trị ban đầu:

- Mức độ L_t được xác định giá trị ban đầu bằng cách tính trung bình của các giá trị thuộc mùa đầu tiên, tức $L_t = \frac{1}{s}(Y_1 + Y_2 + \dots + Y_s)$
- Chú ý rằng đây chính là việc tính trung bình của s quan sát nên nó sẽ loại bỏ tính mùa trong dữ liệu về L_t.

- Để bắt đầu xu hướng, thuận tiện nhất là sử dụng dữ liệu của 2 mùa hoàn chỉnh. $b_s = \frac{1}{s} \left[\frac{Y_{s+1} - Y_1}{s} + \frac{Y_{s+2} - Y_2}{s} + \dots + \frac{Y_{s+s} - Y_s}{s} \right]$

Mỗi số hạng trong công thức trên là một ước lượng của xu hướng qua một vòng mùa hoàn chỉnh, và giá trị ước lượng b_s là trung bình của s số hạng này

- Cuối cùng, những chỉ số mùa đầu tiên được xác định là tỷ số giữa một số giá trị dữ liệu đầu tiên với trung bình của năm đầu tiên, đó là:

$$S_1 = \frac{Y_1}{L_s}; S_2 = \frac{Y_2}{L_s}; \dots; S_s = \frac{Y_s}{L_s}$$

Vận dụng lý thuyết trên đây chúng ta lần lượt xác định các giá trị khởi đầu cho ví dụ của chúng ta như sau:

$$\begin{aligned} L_4 &= \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4) = \frac{1}{4}(362 + 385 + 432 + 341) = 380 \\ b_4 &= \frac{1}{4} \left[\frac{Y_{4+1} - Y_1}{4} + \frac{Y_{4+2} - Y_2}{4} + \frac{Y_{4+3} - Y_3}{4} + \frac{Y_{4+4} - Y_4}{4} \right] \\ &= \frac{1}{4} \left[\frac{Y_5 - Y_1}{4} + \frac{Y_6 - Y_2}{4} + \frac{Y_7 - Y_3}{4} + \frac{Y_8 - Y_4}{4} \right] \\ &= \frac{1}{4} \left[\frac{382 - 362}{4} + \frac{409 - 385}{4} + \frac{498 - 432}{4} + \frac{387 - 341}{4} \right] = 9,75 \end{aligned}$$

$$S_1 = \frac{Y_1}{L_4} = \frac{362}{380} = 0,953; S_2 = \frac{Y_2}{L_4} = \frac{385}{380} = 1,013$$

$$S_3 = \frac{Y_3}{L_4} = \frac{432}{380} = 1,137; S_4 = \frac{Y_4}{L_4} = \frac{341}{380} = 0,897$$

— Thiết lập các công thức trên worksheet làm việc của Excel:

- Điền các giá trị khởi đầu về L_4 , b_4 và S_1 , S_2 , S_3 , S_4 vào các ô trong bảng tính theo đúng trật tự về vị trí.
- Điền ba giá trị trọng số san bằng alpha, beta, gama trên cột T, ta ngẫu nhiên chọn cả ba bằng 0,1.
- Thấy ngay rằng chúng ta không có đủ dữ liệu để tính các giá trị dự báo từ giai đoạn $t = 4$ về trước, giả sử muốn tính $F_4 = F_{3+1} = (L_3 + b_3 * 1) * S_{3+1}$ bạn không có dữ liệu về L_3 , b_3 , và S_0 .
- Công đoạn kế tiếp, bạn nhập công thức tính F_5 vào ô Q6 với nội dung (Bạn đọc tự điều chỉnh thành các địa chỉ ô tương ứng khi nhập công

thức cho phù hợp với bản chất của công thức trình diễn có tính lý thuyết ở dưới đây):

$$F_5 = (L_4 + b_4 * 1) * S_1$$

- Nhập công thức tính L_5 vào ô N6 với nội dung: $L_5 = \alpha Y_5 / S_1 + (1 - \alpha)(L_4 + b_4)$
 - Nhập công thức tính b_5 vào ô O6 với nội dung: $b_5 = \beta * (L_5 - L_4) + (1 - \beta)b_4$
 - Nhập công thức tính S_5 vào ô P6 với nội dung: $S_5 = \gamma(Y_5 / L_5) + (1 - \gamma)S_1$
 - Rê chuột kéo tất cả các công thức được nhập trên các cột L_i , b_i , S_i , F_i để diễn đầy đủ dữ liệu cho cột
 - Nhập công thức tính e^2 cho cột R, chú ý là ta cũng chỉ tính được từ thời đoạn thứ 5
 - Nhập công thức tính MSE vào ô T4
- Gọi lệnh Solver với những khai báo như Hình 14.19 dưới đây, hoàn thành lệnh Solver bạn nhận được các giá trị về trọng số san bằng tối ưu như trong khung hình chữ nhật.

Các tham số α , β , γ được chọn theo cách nhóm kết hợp nào cho MSE nhỏ nhất sẽ được chọn. Với thủ tục Solver trên Excel bạn sẽ mất công nhiều hơn, còn với các phần mềm máy tính chuyên xử lý dữ liệu phát triển hơn như SPSS chẳng hạn việc dò tìm một cách tự động các giá trị này để tìm ra cặp kết hợp tốt nhất không phải là một vấn đề nan giải. Ở ví dụ này, sau khi chạy Solver người ta chọn được các giá trị như sau $\alpha = 0,834$; $\beta = 0,054$; $\gamma = 0$.

Chú ý là khi $\gamma = 0$ có nghĩa là các chỉ số mùa sẽ lặp lại y hệt cho các quý của năm sau, bạn thử kiểm tra lại dữ liệu trên cột S_i mà xem.

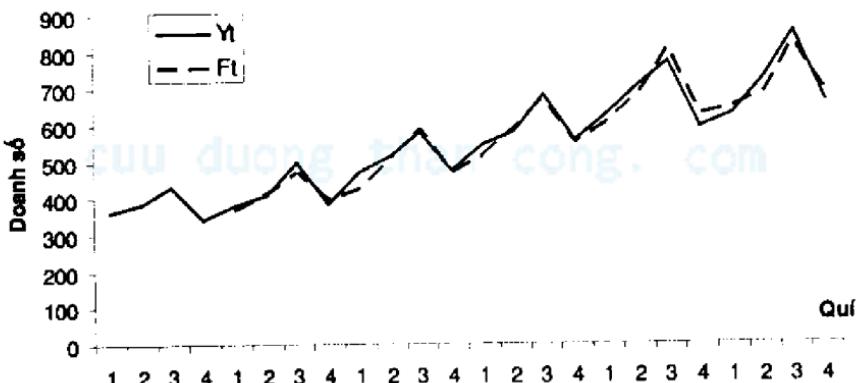
cuu duong than cong. com

Hình 14.19

L	M	N	O	P	Q	R	S
Thời ki	Giá trị quan sát Y_t	L_t	b_t	S_t	F_t	$e^{t/2}$	
1	362	-	-	0.953	-		alpha 0.834
2	385	-	-	1.013	-		beta 0.054
3	432	-	-	1.137	-		gama 0.000
4	341	380	9.75	0.897	-		MSE 608.714;
5	5382	399.004	10.247	0.953	371.432	111.688	
6	6409	404.662	10.001	1.013	414.571	31.036	
7	Solver Parameters						X
8	Set Target Cell:	\$T\$1	\$S\$1				Solve
9	Equal To:	<input checked="" type="radio"/> Max	<input checked="" type="radio"/> Min	<input type="radio"/> Value of:	0		Close
10	By Changing Cells:	\$T\$1:\$T\$3					
11						Guess	
12	Subject to the Constraints:					Add	
13		\$T\$1 <= 1				Change	
14		\$T\$1 >= 0				Reset All	
15		\$T\$2 <= 1					
16		\$T\$2 >= 0					
17		\$T\$3 <= 1					
18		\$T\$3 >= 0					
19	773	684.942	15.749	1.137	807.906	1218.443	
20							

Đồ thị dưới mô tả đồng thời giá trị dự báo và giá trị thực tế, ta có thể nhận thấy giá trị dự báo đã thể hiện được cả tính mùa và tính xu thế.

Hinh 14.20



CHƯƠNG 15

DỰ BÁO BẰNG PHƯƠNG PHÁP BOX-JENKINS

Tên gọi của phương pháp này là sự kết hợp của tên của hai tác giả (George Box và Gwilym Jenkins) đã nghiên cứu bao quát những mô hình dự báo áp dụng cho việc phân tích, dự báo và kiểm soát các chuỗi thời gian trên cơ sở bao quát các tình huống tự hồi qui, sai phân, trung bình trượt. Trên lý thuyết phương pháp này có thể áp dụng để dự báo cho bất kỳ chuỗi thời gian nào nhưng có lẽ thích hợp nhất khi các thành phần mô tả chuỗi thời gian thay đổi tương đối nhanh theo thời gian. Phương pháp Box-Jenkins là một thủ tục với các bước như sau:

- **Bước Nhận dạng:** số liệu quá khứ của chuỗi thời gian được dùng để nhận dạng thử nghiệm một mô hình Box – Jenkins thích hợp
- **Bước Ước lượng:** sử dụng số liệu chuỗi thời gian quá khứ để ước lượng các thông số của mô hình được nhận dạng thử nghiệm ở bước trên
- **Bước Kiểm tra:** dùng nhiều cách chuẩn đoán khác nhau để kiểm tra mô hình thử nghiệm đã ước lượng được, nếu cần thiết có thể đề xuất một mô hình khả thi hơn (trên cơ sở cải thiện được các tiêu chí chuẩn đoán). Mô hình được đề xuất này xem như một mô hình mới nhận dạng thử nghiệm và như thế chúng ta quay trở lại bước đầu tiên.
- **Bước Dự báo:** sử dụng mô hình đạt được cuối cùng qua bước Kiểm tra để dự báo các giá trị tương lai của chuỗi thời gian.

Nội dung lý thuyết của phương pháp Box – Jenkins khá phức tạp, nên chúng ta sẽ xuất phát từ mô hình dự báo Box – Jenkins cổ điển mô tả một chuỗi thời gian dừng, để đơn giản hơn nữa chúng ta làm việc với các mô hình Box-Jenkins không có tính mùa. Để nắm được các bước của phương pháp luận Box-Jenkins chúng ta cần có các hiểu biết về hàm tự tương quan và hàm tự tương quan riêng phần, nhiễu tráng, sai phân, cách phân biệt một chuỗi dừng và không dừng... vì vậy những nội dung sau đây đều tiên sê tập trung tìm hiểu các kiến thức đó.

15.1 KIỂM TRA TÍNH TƯƠNG QUAN TRONG DỮ LIỆU CHUỖI THỜI GIAN

15.1.1 Hệ số tự tương quan

Khi phân tích tính tương quan trong một chuỗi thời gian, con số thống kê cơ bản được sử dụng chính là hệ số tự tương quan, đó là hệ số tương quan giữa chuỗi thời gian Y_t với chuỗi thời gian thứ hai là chính nó được lùi lại 1, 2 hay hơn 2 thời đoạn. Để hiểu điều này hãy xem ví dụ sau đây về một chuỗi gồm 10 quan sát.

Bảng 15.1

Thời đoạn	Chuỗi Y_t	Chuỗi trễ 1 thời đoạn Y_{t-1}	Chuỗi trễ 2 thời đoạn Y_{t-2}
1	Y_1	-	-
2	Y_2	Y_1	-
3	Y_3	Y_2	Y_1
4	Y_4	Y_3	Y_2
5	Y_5	Y_4	Y_3
6	Y_6	Y_5	Y_4
7	Y_7	Y_6	Y_5
8	Y_8	Y_7	Y_6
9	Y_9	Y_8	Y_7
10	Y_{10}	Y_9	Y_8

Các chuỗi Y_{t-1} và Y_{t-2} được tạo ra bằng cách trượt các giá trị của chuỗi Y_t xuống 1 hoặc 2 thời đoạn, theo mức độ lùi k thì chuỗi mới sẽ mất đi đúng k quan sát đầu tiên, chẳng hạn chuỗi Y_{t-2} sẽ mất 2 quan sát đầu tiên.

Hệ số tương quan giữa chuỗi Y_t và chuỗi Y_{t-1} là hệ số tự tương quan bậc 1, hệ số tương quan giữa chuỗi Y_t và chuỗi Y_{t-2} là hệ số tự tương quan bậc 2, theo trình tự này thì hệ số tự tương quan bậc k là hệ số tương quan giữa chuỗi Y_t và chuỗi Y_{t-k} . Hệ số tự tương quan bậc 1 sẽ cho ta thấy các giá trị liên tiếp của Y_t tương quan với nhau ra sao, và hệ số tự tương quan bậc 2 sẽ cho thấy các giá trị của chuỗi cách nhau 2 thời đoạn có tương quan với nhau như thế nào

Liên tưởng đến công thức tính hệ số tương quan giữa X và Y đã nghiên cứu ở Chương 11, ở đây thay vì X ta có Y_t và thay vì Y ta có Y_{t-1} nên công thức tính hệ số tự tương quan bậc 1 được thiết lập như sau:

$$r_{Y_t Y_{t-1}} = \frac{\sum_{i=2}^n (Y_i - \bar{Y})(Y_{i-1} - \bar{Y})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=2}^n (Y_{i-1} - \bar{Y})^2}}$$

Chú ý chỉ số bên dưới dấu \sum của tử số và mẫu số được thiết lập phù hợp với số liệu thực tế của chuỗi. Để dễ dàng tính toán, một giả thuyết được đặt ra là chuỗi Y_t dùng cả về trung bình và phương sai do đó hai giá trị trung bình \bar{Y} và \bar{Y}_{t-1} có thể xem như bằng nhau và lượng phương sai của chúng có thể được ước lượng chung một lần bằng cách sử dụng tất cả những dữ liệu đã có của chuỗi Y_t (khi nghiên cứu về tính dừng của một chuỗi thời gian bạn đọc sẽ hình dung rõ hơn về điều này)

Với giả định đó công thức tính hệ số tự tương quan bậc nhất được viết lại

$$r_{Y_t Y_{t-1}} = \frac{\sum_{i=2}^n (Y_i - \bar{Y})(Y_{i-1} - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Để thuận tiện trong trình bày, người ta qui ước hệ số tự tương quan bậc 1 kí hiệu là r_1 , tương tự hệ số tự tương quan đối với các độ trễ 2,3...k thời đoạn được kí hiệu $r_2, r_3, \dots r_k$. Tổng quát thì công thức tính hệ số r_k là :

$$r_k = \frac{\sum_{i=k+1}^n (Y_i - \bar{Y})(Y_{i-k} - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Vì là hệ số tương quan nên các r_k có tính chất của hệ số tương quan tuyến tính

Ví dụ: Từ dữ liệu của Bảng 15.4 ở trang sau, bạn đọc hãy vận dụng công thức trên để tính r_1 của chuỗi thời gian?

Biết rằng ta tính được giá trị trung bình của chuỗi Y_t bằng 51,03 và từ đó hệ số r_1 được tính như sau:

$$\begin{aligned} r_1 &= \frac{(2-51,03)*(66-51,03)+(55-51,03)*(2-51,03)+\dots+(49-51,03)*(97-51,03)}{(66-51,03)^2+(2-51,03)^2+\dots+(49-51,03)^2} \\ &= \frac{-5415,6}{36096,97} = -0,150 \end{aligned}$$

Vận dụng công thức tính r_k bạn đọc sẽ tiếp tục tính được các r_k còn lại, chọn độ trễ tối đa là 10. Trình bày các kết quả tính toán lên Bảng 15.3.

Bảng 15.3

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
-0,15	0,152	-0,278	-0,051	-0,189	-0,004	0,072	-0,113	-0,01	0,000

15.1.2 Kiểm tra tính tương quan

Sau khi nắm được cách tính hệ số tự tương quan, giờ đây chúng ta quay lại với mục đích ban đầu là Kiểm tra tính tương quan trong dữ liệu chuỗi thời gian, có hai cách kiểm tra tự tương quan trong một chuỗi thời gian căn cứ trên các r_k , cách thứ nhất là nghiên cứu xem từng r_k cụ thể có khác 0 một cách đáng kể hay không; cách thứ hai là xem xét toàn bộ tập hợp các giá trị của r_k , kiểm tra xem tập này có khác 0 một cách có ý nghĩa hay không bằng phương pháp kiểm định.

Giả sử chúng ta có một chuỗi thời gian gồm 36 quan sát như trong Bảng 15.4. Chúng ta sẽ sử dụng từng phương pháp để kiểm tra mức độ tự tương quan trong chuỗi. Trước hết bạn đọc cần biết rằng đây không phải là một chuỗi dữ liệu có thật về doanh thu hay lợi nhuận của một tổ chức nào cả, chuỗi này được tạo ra bằng lệnh phát số ngẫu nhiên Random Number Generation trên Tool Data Analysis của Excel, điều đó có nghĩa là chuỗi hoàn toàn mang tính ngẫu nhiên. Nhưng bây giờ xem như bạn chưa biết điều đó, bạn chỉ biết là có một chuỗi thời gian với $n = 36$ và bạn cần kiểm tra mức độ tự tương quan của nó, dĩ nhiên nếu các phương pháp kiểm tra mức độ tự tương quan được tiến hành đúng bạn phải có kết luận là chuỗi này không có tính tự tương quan.

Bảng 15.4

Thứ tự	Y_t						
1	66	10	4	19	91	28	69
2	2	11	24	20	92	29	33
3	55	12	32	21	88	30	49
4	21	13	70	22	15	31	20
5	74	14	37	23	6	32	79
6	38	15	32	24	20	33	41
7	100	16	79	25	62	34	20
8	43	17	30	26	100	35	97
9	100	18	97	27	2	36	49

Trên lý thuyết, nếu một chuỗi thời gian là độc lập hay nói cách khác không có tự tương quan trong chuỗi dữ liệu thì tất cả các hệ số tự tương quan r_k phải bằng 0, nếu có một r_k bất kỳ khác 0 đồng nghĩa với việc chuỗi Y_t và chuỗi trễ k thời đoạn có tương quan nhau. Chúng ta cũng ch

ý là do số liệu ta làm việc là số liệu của mẫu vì 36 số ngẫu nhiên được chọn chỉ là một tình huống cụ thể xảy ra của vô hạn tập 36 số ngẫu nhiên có thể được Excel cung cấp nên các r_k tính được có thể khác 0 nhiều ít chứ không chính xác bằng 0, điều đó dẫn đến một yêu cầu là phải kiểm định xem một r_k nào đó khác 0 có phải là “khác 0 một cách có ý nghĩa thống kê” hay không.

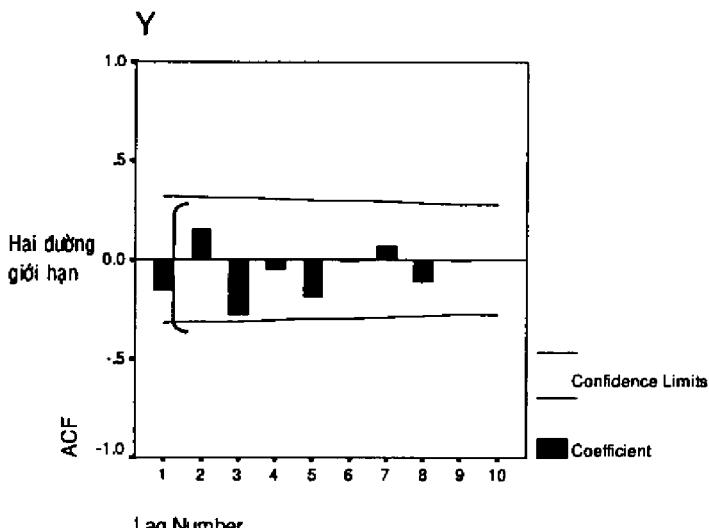
Giả sử chúng ta lấy được tất cả các tập số gồm 36 số ngẫu nhiên, tính được tất cả các hệ số tự tương quan r_k với độ trễ 1, 2, 3 đến tối đa là 10, tính trung bình tất cả các r_k của các mẫu lấy được tại từng độ trễ cụ thể thì giá trị hệ số tự tương quan tổng thể ở độ trễ k sẽ rất gần bằng 0. Kí hiệu ρ_k là hệ số tự tương quan tổng thể ở độ trễ k, nó diễn tả hệ số tự tương quan của toàn tổng thể thì các hệ số tự tương quan của các mẫu khác nhau sẽ tạo thành một phân phối xung quanh ρ_k gọi là phân phối mẫu của hệ số tự tương quan. Chúng ta thừa nhận kết luận rằng nếu chuỗi thời gian đảm bảo tính độc lập thì các hệ số tự tương quan mẫu có phân phối bình thường với trung bình bằng 0 và độ lệch tiêu chuẩn bằng $1/\sqrt{n}$. Từ đây chúng ta kì vọng rằng khoảng 95% các hệ số tự tương quan của các mẫu sẽ nằm trong một phạm vi được xác định bởi $\pm 1,96$ lần độ lệch tiêu chuẩn của phân phối mẫu (tức $\pm 1,96/\sqrt{n}$).

Như vậy một chuỗi thời gian có thể được kết luận không có hiện tượng tự tương quan nếu tất cả các hệ số tương quan mẫu tính toán được nằm trong giới hạn $(-1,96\sqrt{1/n} \leq r_k \leq +1,96\sqrt{1/n})$, lúc đó ta kết luận các hệ số tự tương quan mẫu khác 0 tính được chỉ là do tình cờ. Với chuỗi gồm 36 quan sát trong ví dụ trên chúng ta xác định được các giới hạn là $\pm 1,96\sqrt{1/36} = \pm 0,327$; vậy nếu tất cả các r_k (với $k = 1,2,3\dots 10$) thỏa mãn điều kiện $(-0,327 \leq r_k \leq +0,327)$ thì ta kết luận được chuỗi là ngẫu nhiên. Với 10 giá trị r_k đã được tính toán và liệt kê trong Bảng 15.3 (chính là 10 giá trị r_k của tập dữ liệu 36 quan sát này) bạn có thể so sánh với giới hạn trên và thấy ngay không có giá trị nào vượt khỏi khoảng giới hạn.

Một công cụ phục vụ đắc lực cho sự đánh giá này là đồ thị hệ số tự tương quan, đồ thị này vẽ ra giá trị của tất cả các r_k ta đã tính toán được với độ trễ tối đa do ta chọn (giả sử là 10), và thể hiện sẵn hai đường giới hạn tương ứng với $\pm 1,96\sqrt{1/n}$ (với n là cỡ mẫu tức là số thời đoạn ta đã biết). Nếu có một r_k bất kỳ vượt qua hai đường giới hạn này thì ta kết luận rằng trên thực tế giá trị r_k đó khác 0 không phải do tình cờ mà “có ý nghĩa thống kê”. Ngược lại nếu tất cả 10 hệ số tự tương quan đều nằm trong đường giới hạn thì chuỗi Y_t là ngẫu nhiên. Với phần mềm SPSS, dùng lệnh Time Series/Autocorrelations của menu Graphs bạn có thể phác họa

được đồ thị của 10 hệ số tự tương quan nhanh chóng, đồ thị này được gọi tên là ACF.

Hình 15.1



Trên ACF chúng ta thấy các cột màu sẫm có độ cao ứng với giá trị tuyệt đối của r_k và hướng của cột phụ thuộc vào dấu thực tế của r_k . Hai đường kẻ song song chính là hai đường giới hạn. Trên ACF ta không thấy có một cột nào cao vượt khỏi hai đường giới hạn, điều đó có nghĩa là chuỗi hoàn toàn độc lập.

Sau khi nghiên cứu nội dung chương Phương pháp Box-Jenkins bạn sẽ nhận ra rằng ACF là một công cụ đắc dụng khi dự báo với chuỗi thời gian nhiều hơn những gì chúng ta vừa tìm hiểu ở trên đây, giả dụ để đánh giá chất lượng của một mô hình dự báo được áp dụng thì ngoài các chỉ tiêu đánh giá độ phù hợp của mô hình chúng ta còn phải xem xét tính chất của sai số dự báo e_t , xem thử có tồn tại một mẫu hình nào trong sai số dự báo sau khi áp dụng mô hình dự báo hay không, nếu có thì nhà dự báo phải chọn một mô hình tiến bộ hơn (bạn đọc có thể liên tưởng đến tính hệ thống trong sai số dự báo của chuỗi thời gian được dùng làm ví dụ minh họa cho phương pháp Holt – Winter mà được đem xử lý bằng phương pháp hàm số mũ đơn giản). Nói ngắn gọn là nếu dùng ACF phân tích phần dư (xem phần dư e_t như một chuỗi thời gian Y_t bất kỳ) mà kết quả cho thấy tất cả các hệ số tự tương quan đều nằm trong giới hạn thì có nghĩa là mô hình dự báo đó đã thỏa điều kiện cho ra một sai số có tính “nhiều trắng” vì không có một thông tin nào còn sót lại trong dữ liệu mà

mô hình không chỉ ra được. Những khái niệm phức tạp này sẽ trở nên dễ hiểu hơn khi bạn gặp lại kiến thức về nhiễu trắng và phân tích tính mùa trong chuỗi thời gian ở phần sau.

Một phương pháp thứ hai để kiểm tra mức độ tương quan trong một chuỗi thời gian là xem xét toàn bộ tập hợp các giá trị của r_k , kiểm tra xem tập này có khác 0 một cách có ý nghĩa hay không bằng phương pháp kiểm định χ^2 . Kiểm định này dựa trên con số thống kê được phát triển bởi Ljung và Box, công thức của kiểm định như sau:

$$Q^* = n(n+2) \sum_{k=1}^h \left[\frac{1}{(n-k)} r_k^2 \right]$$

Trong đó

- n là số quan sát của chuỗi
- k là độ trễ xét r_k
- r_k là hệ số tự tương quan bậc k
- h là độ trễ lớn nhất được xem xét

Nếu chuỗi là hoàn toàn ngẫu nhiên thì Q có phân phối χ^2 với $(h-m)$ bậc tự do trong đó m là số tham số trong mô hình phù hợp đã được xây dựng trên chuỗi dữ liệu, khi áp dụng công thức này cho dữ liệu thật tức là không có mô hình phù hợp nào được xây dựng thì ta chọn $m = 0$

Với ví dụ của chúng ta $h = 10$. Lúc đó căn cứ trên số liệu đã tính toán được về r_k ta tính được lượng $\sum_{k=1}^h \left[\frac{1}{(n-k)} r_k^2 \right] = 0,005537$ nên giá trị

$$Q^* = 36 \times 38 \times (0,005537) = 7,57$$

So sánh giá trị Q^* với giá trị $\chi^2_{(0,05;10)} = 18,3$ ta thấy Q^* bé hơn nên ta chấp nhận giả thuyết H_0 cho rằng tập giá trị r_k là không khác 0 một cách có ý nghĩa.

Chú ý là SPSS sẽ cung cấp luôn cho chúng ta các giá trị Q^* của kiểm định Ljung – Box kèm với ACF khi ta thực hiện lệnh Time Series/Autocorrelations của menu Graphs.

Chúng ta đã nghiên cứu xong 2 phương pháp kiểm tra tính độc lập của một chuỗi thời gian, việc kiểm tra mức độ tương quan trong một chuỗi thời gian được trình bày cẩn kẽ ở đây để bạn đọc có thể vận dụng cho việc kiểm tra sự vi phạm giả định tương quan chuỗi trong phần dư của mô hình hồi qui tuyến tính. Chú ý rằng Durbin-Watson chỉ là một kiểm định chính thức về tự tương quan bậc nhất trong phần dư, phương pháp chuẩn để xác định phần dư có ngẫu nhiên hay không là xem xét các hệ số tự tương quan của các phần dư ở nhiều độ trễ liên tiếp nhau xem có tồn tại

một kiểu mẫu nào trong hệ thống sai số không, hay các sai số độc lập nhau, bằng hai phương pháp chúng ta vừa tìm hiểu ở trên.

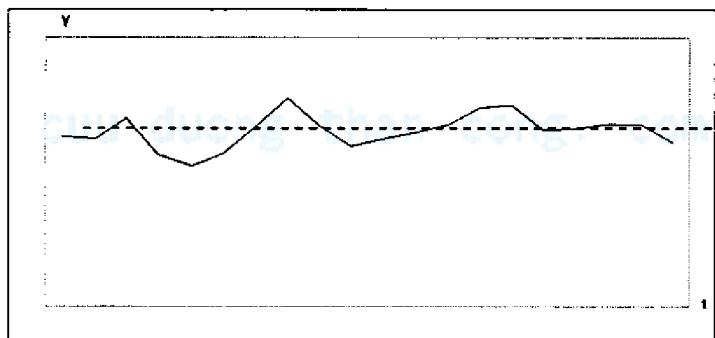
15.2 TÍNH DỪNG CỦA CHUỖI THỜI GIAN

15.2.1 Khảo sát tính dừng

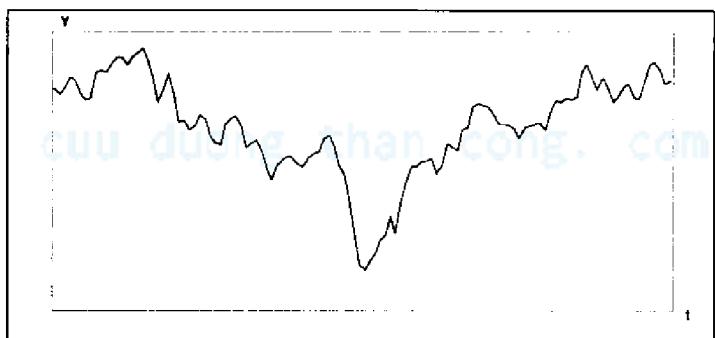
Tính dừng trong một chuỗi dữ liệu được hiểu nôm na là không có sự tăng trưởng hay suy thoái trong dữ liệu mà dữ liệu dao động gần như tập trung xung quanh một trục nằm ngang theo chiều tăng của thời gian, nói cách khác nữa là dữ liệu biến động xung quanh giá trị trung bình không đổi và độ lớn của phương sai thể hiện biến động về cơ bản cũng giữ nguyên theo thời gian. Như vậy khái niệm dừng của một chuỗi thời gian gồm hai nội dung là dừng theo trung bình và dừng theo phương sai. Các bạn xem các hình minh họa dưới đây cho cả hai tình huống chuỗi dừng và không dừng.

Hình 15.2

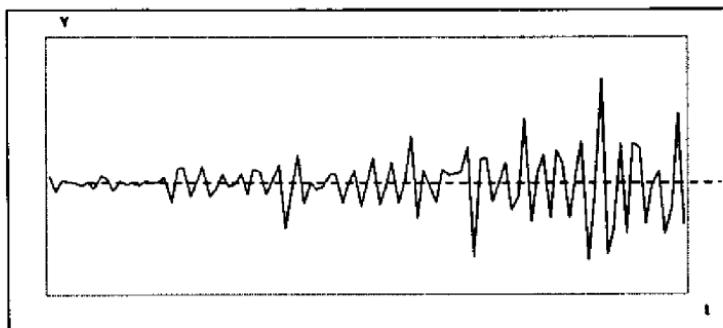
(a)



(b)



(c)



Xem Hình 15.2 a bạn có thể cảm nhận thấy là dữ liệu tuy thay đổi liên tục theo chiều của thời gian nhưng đường như biên độ dao động không có sự thay đổi lớn, đồ thị cũng không cho thấy một xu hướng tăng hay giảm trong dữ liệu mà chúng biến động theo phương ngang gần như tập trung quanh 1 đường thẳng cố định. Đây là một chuỗi dừng về cả trung bình và phương sai.

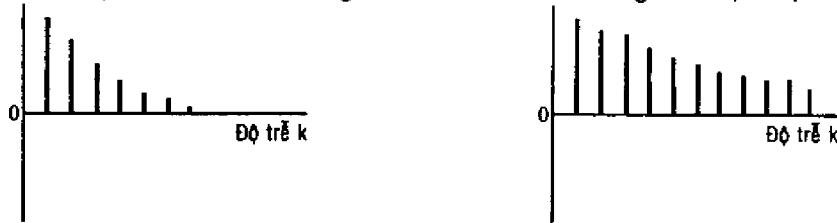
Còn với Hình 15.2 b, ta thấy rằng theo chiều thời gian, giá trị gốc của chuỗi thời gian không dao động xung quanh một trị trung bình cố định và biên độ dao động của đối tượng cũng thay đổi, chuỗi ở hình b không dừng cả về trung bình và phương sai.

Hình 15.2 c cho bạn đọc hình dung về hình dạng của một chuỗi dừng về trung bình mà không dừng phương sai khi dữ liệu dao động dường như theo phương ngang với biên độ dao động lớn dần theo thời gian.

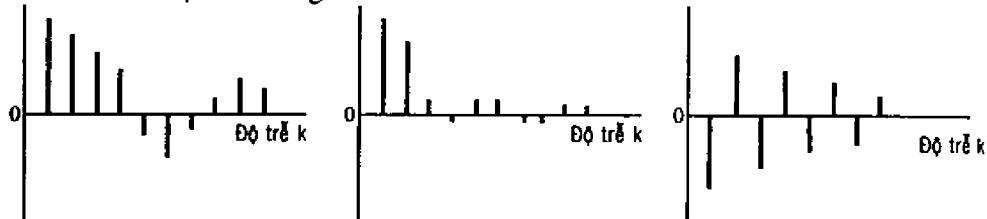
Như vậy đồ thị của một chuỗi tuần tự theo thời gian có thể giúp chúng ta nhận ra chuỗi dừng hay không bao gồm tình huống dừng cả trung bình và phương sai. Ngoài ra đồ thị ACF cũng là một phương tiện tốt để khảo sát tính chất dừng của một chuỗi. Tổng quát, về mặt hình ảnh bạn hình dung là với một chuỗi dừng thì đặc trưng trên đồ thị ACF là bạn có thể thấy một xu hướng giảm nhanh hoặc rất nhanh từ trái sang phải theo chiều tăng của độ trễ. Ngược lại, ACF thể hiện xu hướng giảm thật chậm từ trái sang phải theo chiều tăng của độ trễ nghĩa là chuỗi gốc không dừng (ở dưới đây có minh họa một số dạng ACF của một số chuỗi thời gian dừng hoặc không dừng).

Ý niệm chính xác của sự “rất nhanh” hay “thật chậm” được xem xét một cách khá linh hoạt, ngoài ra kinh nghiệm cho thấy nếu chuỗi dừng, việc ACF giảm rất nhanh thường xảy ra sau các hệ số tự tương quan bậc 1 hoặc 2, còn với chuỗi không dừng thì với nhiều độ trễ r_k vẫn khác 0 một cách có ý nghĩa.

Hình 15.3 Phân biệt ACF giảm rất nhanh và ACF giảm thật chậm

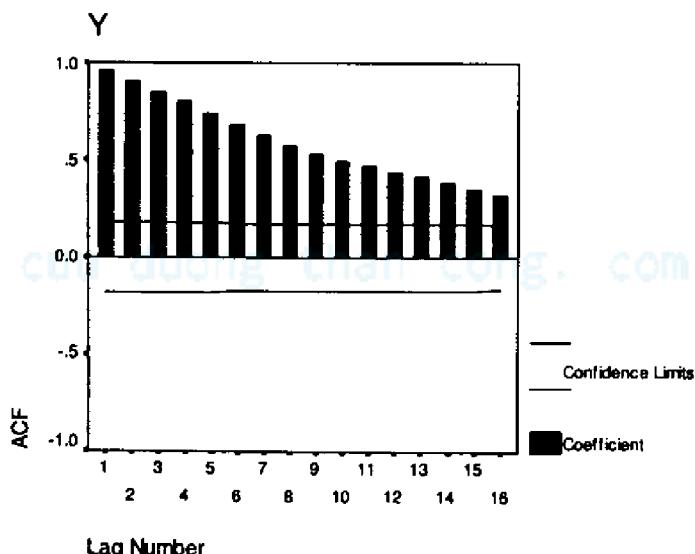


Hình 15.4 Một số kiểu giảm nhanh khác của ACF



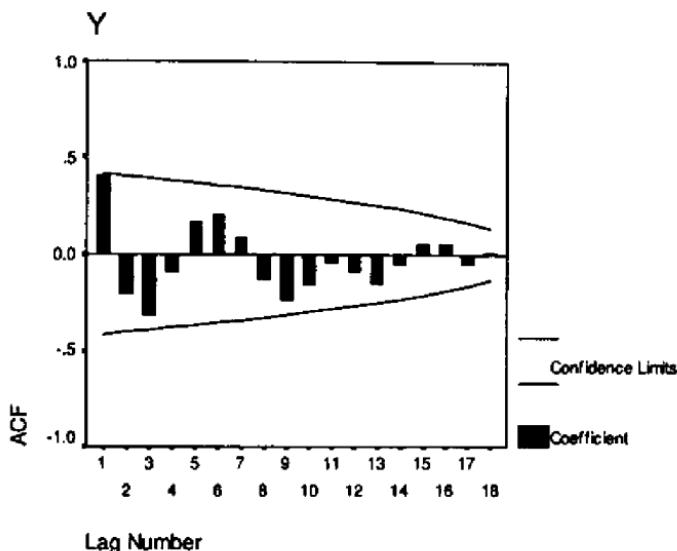
Ví dụ: Nếu vẽ ACF với độ trễ tối đa là 16 cho dữ liệu đã tạo nên chuỗi trong Hình 15.2(b) chúng ta có Hình 15.5. Trên hình này, xem xét từ trái sang phải của ACF chúng ta thấy một xu hướng giảm rất chậm của các r_k thể hiện qua chiều cao của các cột, và tất cả các r_k đều vượt ra ngoài đường giới hạn, điều đó chứng minh tính không dừng trong chuỗi gốc. Cính tính xu hướng sẽ dẫn đến việc các hệ số tự tương quan dương nổi trội trong ACF cho thấy dữ liệu gốc có sự tương quan.

Hình 15.5



Tiếp tục vẽ ACF của dữ liệu của đồ thị Hình 15.2(a) với 26 độ trễ, ta có được Hình 15.6, trong hình này ta thấy các r_k giảm khá nhanh theo dạng hình sin, nên nếu căn cứ trên ACF mà không xem đồ thị dữ liệu gốc ta cũng kết luận được là chuỗi gốc dừng.

Hình 15.6



15.2.2 Loại bỏ tính dừng

Chúng ta đã xác định là ban đầu sẽ làm việc với mô hình Box-Jenkins cổ điển trên chuỗi dữ liệu dừng, vậy khi gặp một chuỗi dữ liệu không dừng ta làm sao? Khi gặp một chuỗi không dừng cần loại bỏ tính không dừng trước khi tiến hành các phân tích kế tiếp. Có thể dễ dàng làm dừng một chuỗi bằng phương pháp sai phân.

Chúng ta định nghĩa sai phân bậc 1 qua công thức sau

$$Y'_t = Y_t - Y_{t-1}$$

Trong đó:

- Y_t và Y_{t-1} là giá trị của chuỗi tại thời đoạn t hoặc $t-1$
- Y'_t là kí hiệu của sai phân bậc 1 tại thời đoạn t

Chúng ta định nghĩa sai phân bậc 2 qua công thức sau

$$Y''_t = Y'_t - Y'_{t-1} = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$$

Trong đó:

- Y_t , Y_{t-1} và Y_{t-2} là giá trị của chuỗi tại thời đoạn t , $t-1$, hoặc $t-2$
- Y''_t là kí hiệu của sai phân bậc 2 tại thời đoạn t
- Y'_t và Y'_{t-1} là kí hiệu của sai phân bậc 1 tại thời đoạn t và $t-1$

Chuỗi sai phân Y'_t sẽ dừng nếu xu hướng của chuỗi gốc là tuyến tính và nó chỉ còn $n-1$ quan sát do Y'_1 không thể tính được mà phải bắt đầu từ Y'_2 . Nếu sau khi lấy sai phân bậc một mà các kiểm tra vẫn cho thấy dữ liệu chưa dừng thì phải tiếp tục lấy sai phân bậc 2. Chuỗi sai phân bậc 2 có $n-2$ quan sát. Trong thực tế người ta thấy rằng hiếm khi phải tính các sai phân bậc cao hơn vì thông thường sau khi lấy đến sai phân bậc 2 là chuỗi đã dừng.

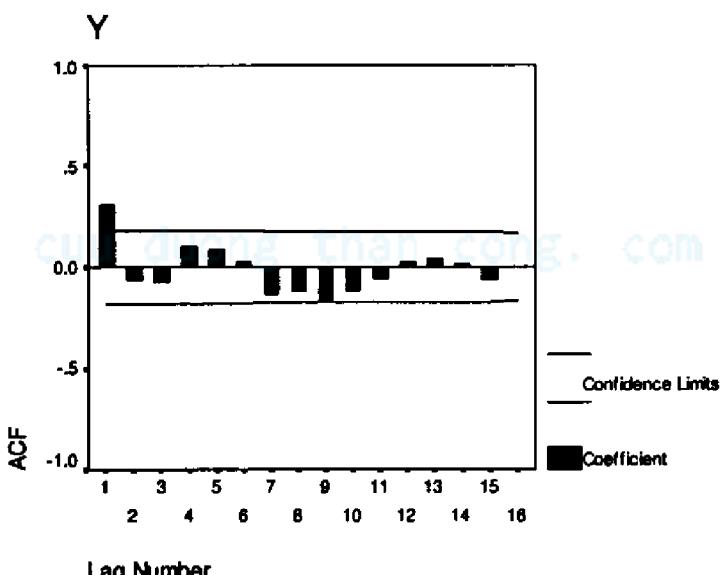
Bảng số liệu sau đây giúp bạn hình dung rõ hơn về cách lấy sai phân.

Bảng 15.5

Thời đoạn	Chuỗi Y_t	Chuỗi $Y'_t = Y_t - Y_{t-1}$	Chuỗi $Y''_t = Y'_t - Y'_{t-1}$
1	Y_1	-	-
2	Y_2	$Y_2 - Y_1$	-
3	Y_3	$Y_3 - Y_2$	$(Y_3 - Y_2) - (Y_2 - Y_1)$
4	Y_4	$Y_4 - Y_3$	$(Y_4 - Y_3) - (Y_3 - Y_2)$
...
$n-1$	Y_{n-1}	$Y_{n-1} - Y_{n-2}$	$(Y_{n-1} - Y_{n-2}) - (Y_{n-2} - Y_{n-3})$
n	Y_n	$Y_n - Y_{n-1}$	$(Y_n - Y_{n-1}) - (Y_{n-1} - Y_{n-2})$

Với chuỗi thời gian không dừng ở Hình 15.2(b), sau khi lấy sai phân bậc 1 của chuỗi, ta vẽ lại ACF cho chuỗi sai phân bậc 1 này thì được Hình 15.7 dưới đây. Tính dừng trong chuỗi thể hiện rõ ràng, đó là r_k khác 0 đáng kể nhưng các r_k còn lại nhanh chóng tắt về 0. Như vậy ta kết luận được quá trình lấy sai phân bậc 1 đã làm dừng chuỗi không dừng này.

Hình 15.7



Ngoài ra, nếu chuỗi của bạn có sự biến thiên thay đổi theo thời gian tức là không dừng theo phương sai thì chúng ta sử dụng phương pháp biến đổi là lấy Logarit tự nhiên hoặc lấy căn bậc 2 hay căn bậc 4 của Y, áp dụng phép biến đổi nào là tốt nhất tùy thuộc đánh giá của bạn sau khi đã biến đổi và xem xét lại đồ thị chuỗi thời gian xem còn tồn tại tính không dừng theo phương sai không.

15.3 HỆ SỐ TỰ TƯƠNG QUAN RIÊNG

Hệ số tự tương quan riêng phần được dùng để đo mức độ của sự tương quan giữa Y_t và Y_{t-k} khi ảnh hưởng của các độ trễ khác như 1,2,3...(k-1) được tách riêng ra.

Hệ số tự tương quan riêng bậc k ký hiệu là α_k có thể được tính bằng cách hồi qui Y_t theo $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k}$ trong phương trình như sau

$$\hat{Y}_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + \dots + b_k Y_{t-k}$$

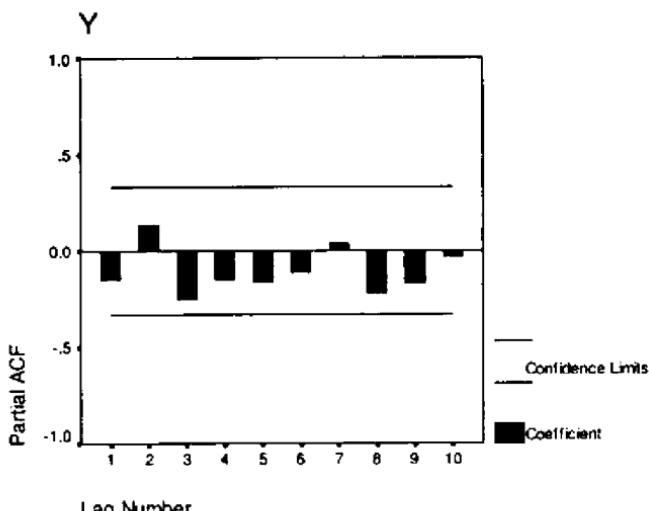
Hệ số tự tương quan riêng α_k chính là hệ số b_k được ước lượng từ phương trình hồi qui bội trên.

Đây là một phương trình hồi qui tuyến tính đặc biệt ở chỗ các biến giải thích ở bên phải chính là giá trị trễ của Y_t , vì thế tên gọi mô hình tự hồi qui được dùng để chỉ tình huống này. Khái niệm tự hồi qui là Autoregressive vì thế viết tắt là AR và theo bậc của độ trễ cuối cùng trong các biến giải thích mà ta viết AR(p). ví dụ $\hat{Y}_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2}$ là phương trình AR(2). Muốn ước lượng hệ số tự tương quan riêng đến bậc p ta chạy mô hình AR(p).

Còn có công thức tính hệ số này theo cách khác, tuy nhiên chúng ta không cần vất vả nghiên cứu công thức tính hệ số tự tương quan riêng phần mà nên tập trung vào công dụng của nó vì các phần mềm ví dụ SPSS tự động tính cho chúng ta tất cả các hệ số tự tương quan riêng phần cho đến độ trễ mà ta mong muốn.

Với chuỗi 36 số ngẫu nhiên ví dụ của chúng ta (Bảng 15.4), ta dùng SPSS (cũng lệnh Time Series/Autocorrelations của menu Graphs) phác họa đồ thị của 10 hệ số tự tương quan riêng phần như hình sau (Đồ thị này được gọi tên là PACF). SPSS đồng thời tính luôn cho chúng ta các giá trị của hệ số tự tương quan riêng phần.

Hình 15.8



Quy tắc đánh giá PACF cũng giống như ACF là giá trị giới hạn $1.96\sqrt{n}$ được sử dụng để đánh giá mức độ khác 0 có ý nghĩa của các hệ số tự tương quan riêng phần để kết luận chuỗi thời gian có ngẫu nhiên không. Nên bạn cũng có thể dùng thêm PACF để đánh giá về tính độc lập của chuỗi thời gian.

Với Hình 15.9, PACF cho thấy các α_k không khác 0 một cách có ý nghĩa vì tất cả các cột đều nằm trong đường giới hạn. Như vậy chuỗi 36 số ngẫu nhiên này đảm bảo không có tự tương quan.

Tuy nhiên công dụng khác PACF là để giúp quyết định chọn dạng mô hình Box-Jenkins thử nghiệm phù hợp, chứ không chỉ dùng để kiểm tra tự tương quan trong chuỗi dữ liệu. Trong nội dung kế tiếp sau đây chúng ta sẽ thấy sự vận dụng PACF cho mục đích thứ hai này.

15.4 MÔ HÌNH BOX – JENKINS (ARIMA) CHO CHUỖI DỪNG VÀ DỰ BÁO

Sau khi nắm được các kiến thức bổ trợ, chúng ta có thể bắt đầu đi vào nghiên cứu các dạng mô hình Box – Jenkins cụ thể, mô hình Box – Jenkins là tên gọi chung của một họ rất nhiều mô hình khác nhau do sự tồn tại riêng lẻ hoặc kết hợp đồng thời của quá trình tự hồi qui (Autoregressive viết tắt là AR) và trung bình trượt (Moving Average viết tắt là MA - các bạn sẽ biết rằng MA này tuy giống về cách viết nhưng khác bản chất với phương pháp trung bình trượt Moving Average mà đã nghiên cứu ở nội dung Chương 14) và có thể kết hợp cả quá trình lấy sai

phân nếu chuỗi là không dừng (tức nếu phải lấy sai phân để tịnh hóa dữ liệu thì ta phải xét thêm đến việc lấy sai phân bậc mấy).

Chúng ta đã xác định sẽ làm việc với chuỗi dữ liệu có tính dừng trước tức là lúc này chúng ta chưa bận tâm đến việc phải lấy sai phân do đó đầu tiên chúng ta sẽ chỉ lần lượt nghiên cứu nhóm mô hình AR, MA và nhóm kết hợp cả AR và MA vào một mô hình với tên gọi tự hồi qui-trung bình trượt Autoregressive-Moving Average viết tắt ARMA.

Sau đó ở nội dung với chuỗi không dừng ta xét thêm việc phải lấy sai phân nên lúc này tên gọi của mô hình là ARIMA là viết tắt của Autoregressive Integrated Moving Average, từ Integrated ám chỉ thành phần sai phân. ARIMA là tên gọi đầy đủ cho họ mô hình Box – Jenkins trong tình huống có đủ các thành phần và trong thực tế mô hình ARIMA còn được dùng phổ biến như một tên gọi khác của mô hình Box – Jenkins.

15.4.1 Các quá trình tự hồi qui (AR)

1.4.1.1 Phương trình

Ý tưởng của tự hồi qui chúng ta đã tìm hiểu ở nội dung hệ số tự hồi qui riêng phần. Tổng quát, đối với mô hình tự hồi qui bậc thứ p kí hiệu AR (p) chúng ta sẽ thiết lập phương trình như sau cho mục đích dự báo:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + e_t$$

Như vậy biến Y_t được ước lượng qua một mối quan hệ với hàng loạt các biến trễ 1 đến p thời đoạn và một thành phần sai số e_t bao hàm trong đó tác động của các yếu tố khác đến Y_t ngoài các trễ (các e_t này độc lập lẫn nhau theo thời gian) và còn có thêm một số hạng hằng số c . Các $\phi_1, \phi_2, \dots, \phi_p$ gọi là các tham số của mô hình tự hồi qui.

Trên thực tế chúng ta hay gặp hai trường hợp $p=1$, $p=2$ tương ứng AR(1), AR(2) có phương trình như sau:

$$\text{AR}(1) \quad Y_t = c + \phi_1 Y_{t-1} \quad \text{với ràng buộc } -1 < \phi_1 < 1$$

$$\text{AR}(2) \quad Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} \quad \text{với } -1 < \phi_2 < 1 \text{ và } \phi_2 + \phi_1 < 1 \text{ và } \phi_2 - \phi_1 < 1$$

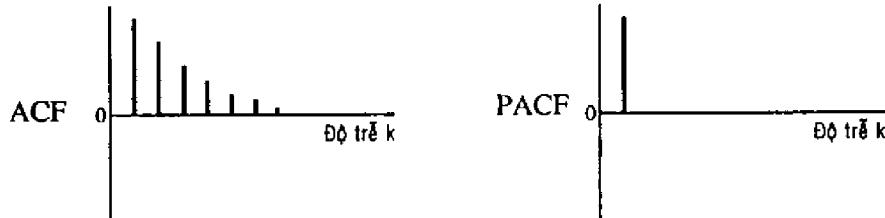
1.4.1.2 Khảo sát dấu hiệu nhận dạng mô hình tự hồi qui

Bạn có một chuỗi thời gian, và bạn muốn biết sử dụng mô hình AR để dự báo có phù hợp không, ngay cả khi bạn đã biết mô hình AR phù hợp thì việc xét một mô hình tự hồi qui đến trễ thứ mấy là vừa cũng không phải là việc đơn giản, chúng ta sẽ khảo sát những dấu hiệu nhận dạng mô hình tự hồi qui phù hợp dựa vào thể hiện của ACF và PACF như sau.

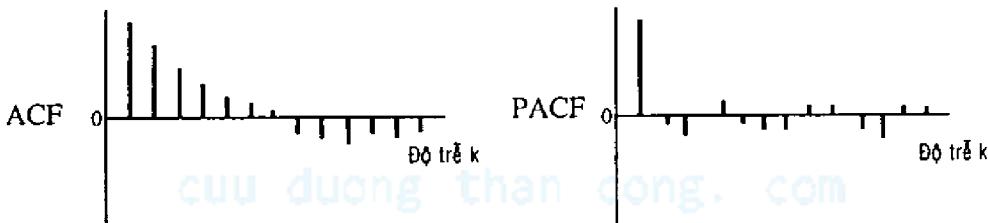
- Về mặt lý thuyết khi ACF có dạng giảm nhanh dần (theo dạng hàm số mũ) và PACF chỉ có duy nhất một hệ số ở trễ 1 có ý nghĩa thì

AR(1) là mô hình tốt. Tuy nhiên thực tế do sai số trong dữ liệu nên có khi ACF của các chuỗi thời gian sẽ không tắt theo dạng mũ hoàn hảo, tức là nó giảm nhanh đến 0 nhưng sau đó chưa tắt hoàn toàn. Và PACF cũng có thể có vài hệ số khác 0 ngẫu nhiên sau trễ đầu tiên (xem Hình 15.10).

Hình 15.9 Minh họa cho ACF và PACF lý thuyết khi nhận dạng AR(1)



Hình 15.10 Minh họa ACF và PACF cho một tình huống thực tế nhưng vẫn cho nhận dạng AR(1)



- Về mặt lý thuyết khi ACF giảm theo sóng hình sin tắt dần và PACF có chính xác 2 đỉnh nhọn ở độ trễ 1 và 2 và tắt hết về 0 sau độ trễ 2 thì đó là dấu hiệu nhận dạng của mô hình AR(2). Tuy nhiên sự nhận dạng các quá trình AR(2) qua các ACF và PACF không phải luôn luôn đơn giản như thế, trong những tình huống ấy nếu PACF có hai hệ số tự tương quan riêng phần đầu tiên khác 0 rõ rệt và các hệ số còn lại không khác 0 nhiều sẽ luôn là một gợi ý tốt cho AR(2)

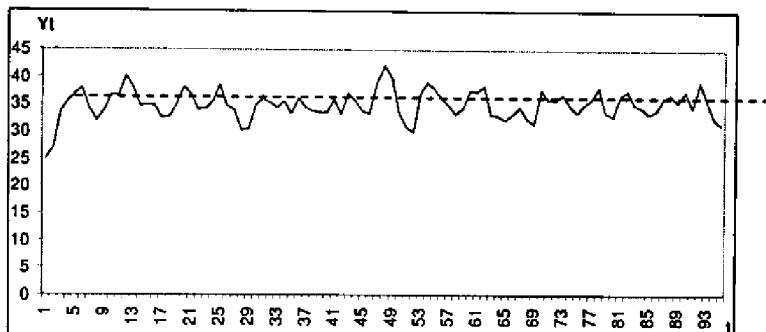
Hình 15.11 Minh họa cho ACF và PACF lý thuyết khi nhận dạng AR(2)



Nhận định chung là ACF của một mô hình phù hợp với dạng AR(p) với $p \geq 2$ sẽ thể hiện một dạng suy giảm theo dạng hàm mũ hay hình sin, còn và PACF có đỉnh nhọn ở độ trễ $1, 2, \dots, p$ và tắt về 0 sau độ trễ p .

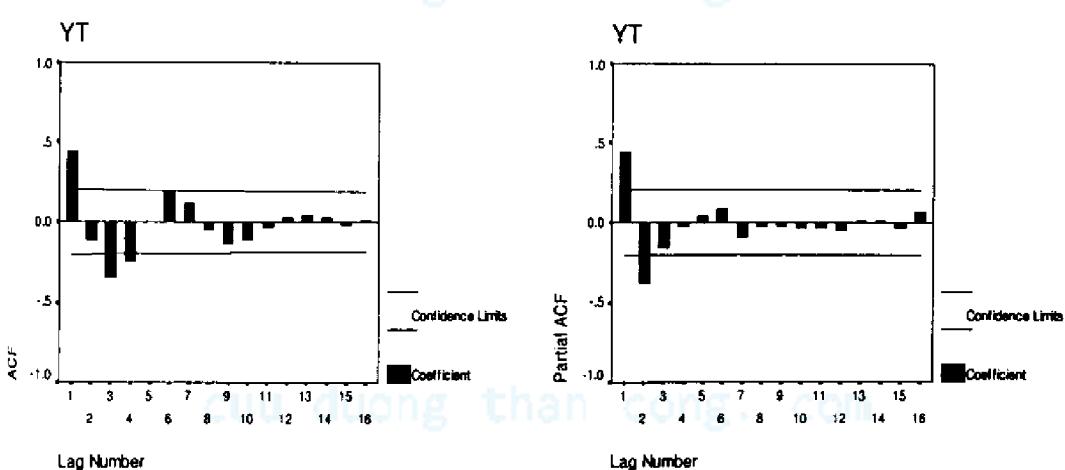
Ví dụ: Ghi nhận số liệu hàng ngày về sản lượng tiêu thụ của một sản phẩm trong vòng 95 ngày và trình bày dữ liệu này lên đồ thị chuỗi thời gian như sau:

Hình 15.12



Đồ thị cho thấy giá trị gốc của chuỗi thời gian dường như dao động quanh một giá trị hằng số và như vậy tạm kết luận là chuỗi này có tính dừng. Để có thêm nhận định về mô hình phù hợp có thể áp dụng cho dữ liệu ta sẽ nghiên cứu biểu hiện trên ACF và PACF xây dựng từ dữ liệu chuỗi thời gian gốc.

Hình 15.13



ACF giảm khá nhanh sau trễ đầu tiên theo dạng hình sin, ta kết luận chuỗi thời gian dừng. Đồng thời PACF có chính xác hai đỉnh nhọn khác 0 ở độ trễ 1,2 và tắt nhanh về 0 sau trễ thứ 2 nên ta đi đến kết luận có thể dùng mô hình AR(2) để thiết lập mô hình mô tả chuỗi thời gian gốc và phục vụ cho mục đích dự báo.

Đến lúc này bạn sẽ thắc mắc làm cách nào có thể tính toán được các giá trị tham số ϕ_1 và ϕ_2 , đây là một khối lượng tính toán không đơn giản chút nào và chúng ta phải nhờ máy tính giúp đỡ. Các phần mềm máy tính sẽ bắt đầu với ước lượng điểm ban đầu của các tham số ước lượng sau đó áp dụng kỹ thuật tìm kiếm lặp dựa vào hàm tổng các bình phương để thu được các tham số cuối cùng trên cơ sở cực tiểu hóa tổng bình phương phần dư. Chẳng hạn với phần mềm SPSS chúng ta có thể thiết lập mô hình với các tham số dự định và nhanh chóng có được kết quả mong muốn trên cửa sổ Output, đồng thời trên cửa sổ Data của phần mềm này chúng ta có kết quả dự báo điểm (đến thời đoạn mong muốn mà bạn xác nhận với SPSS), tương ứng với các kết quả dự báo đó là sai số dự báo, và cả giới hạn trên và giới hạn dưới của dự báo khoảng.

Sau đây là một phần của kết quả xử lý của SPSS trong ví dụ này:

	B	SEB	T-RATIO	APPROX. PROB.
AR1	.682099	.09826285	6.94158	.00000000
AR2	-.433062	.09426190	-4.59424	.00001377
CONSTANT	34.946334	.29669604	117.78497	.00000000

Từ kết quả trên ta viết lại phương trình AR(2)

$$Y_t = 34,9463 + 0,6821xY_{t-1} - 0,4331xY_{t-2}$$

Trên màn hình Output do SPSS truy xuất cho kết quả xử lý mô hình còn nhiều nội dung khác mà chúng ta sẽ lần lượt làm sáng tỏ ở các phần sau.

Dưới đây là một phần của màn hình Data của SPSS được cắt ra để các bạn tham khảo.

Hình 15.14

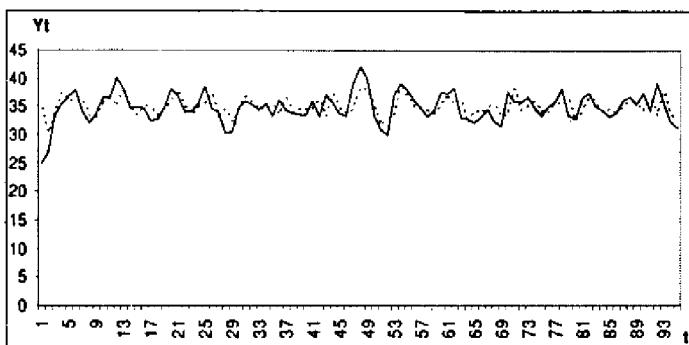
t	y _t	fit _{_1}	err _{_1}	lcl _{_1}	ucl _{_1}	sep _{_1}
1	25	34.94633	-9.94033	9.48755	40.40312	2.74851
2	27	30.21214	-3.21214	25.42944	35.99485	2.40810
3	31	33.83352	-31.932	20.50880	38.15819	2.17748
4	35	37.41073	-1.91453	32.08607	41.73539	2.17748
5	37	35.94160	.96130	31.61694	40.26626	2.17748

Giá trị dự báo
Sai số dự báo
Giới hạn dưới của khoảng DB
Giới hạn trên của khoảng

SPSS cũng cho chúng ta giá trị dự báo tiếp các thời đoạn sau nếu chúng ta yêu cầu, ví dụ giá trị dự báo của số sản phẩm tiêu thụ vào thời đoạn 96 được SPSS tính toán và cho biết là 33,55 sản phẩm. Kết quả dự báo khoảng với độ tin cậy 95% là (29,23;37,88) sản phẩm.

Dùng dữ liệu yt và fit_1 trên màn hình Data của SPSS ta vẽ được đồ thị mô tả đồng thời giá trị thực tế (đường liền nét) và giá trị dự báo (đường đứt nét) để so sánh độ phù hợp của mô hình.

Hình 15.15



15.4.2. Các quá trình trung bình trượt (MA)

1.4.2.1 Phương trình

Không như quá trình AR hồi qui ngược lại các giá trị trước đó của chuỗi thời gian để phục vụ cho dự báo, có một mô hình hồi qui cho chuỗi thời gian mà sử dụng các giá trị sai số trong quá khứ như là biến giải thích, mô hình như vậy được gọi là trung bình trượt kí hiệu MA, quá trình MA tổng quát bậc q có thể viết dưới dạng phương trình như sau

$$Y_t = c + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q}$$

Trong đó:

- e_{t-k} là thành phần sai số ở thời đoạn $t-k$ có tính chất không tương quan qua các thời kì
- Các θ_k là các tham số trung bình trượt. Lưu ý ta qui ước đặt dấu trừ trước các θ_j trong công thức nên cần chú ý khi diễn dịch ý nghĩa của chúng với số liệu thực tế.
- c cũng đại diện cho hằng số.

Thực tế cũng cho thấy ta hay gặp nhất là tình huống $q \leq 2$ tương ứng MA(1) hoặc MA(2).

Với MA(1) ta có phương trình $Y_t = c + e_t - \theta_1 e_{t-1}$

Với MA(2) ta có phương trình $Y_t = c + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$

Những giới hạn đối với tham số của quá trình tự hồi qui cũng giống như với quá trình trung bình trượt đó là nếu $q=1$ thì $-1 < \theta_1 < 1$ còn với $q=2$ thì $-1 < \theta_2 < 1$ và $\theta_2 + \theta_1 < 1$ và $\theta_2 - \theta_1 < 1$.

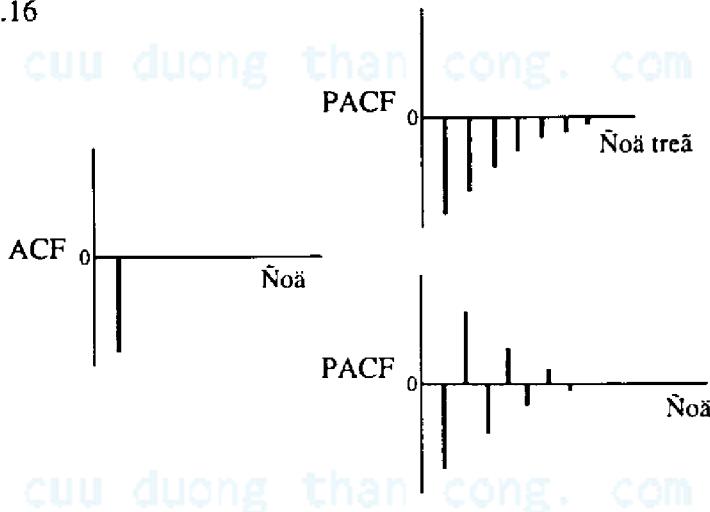
1.4.2.2 Khảo sát dấu hiệu nhận dạng mô hình trung bình trượt

Nếu chuỗi dữ liệu thời gian phù hợp với dạng mô hình trung bình trượt thì ACF và PACF về mặt lý thuyết trông như thế nào. Chú ý rằng ACF cho ta thông tin về xét đoán MA nhiều hơn còn PACF sẽ cho nhiều thông tin hơn về sự phán xét AR phù hợp, điều này là hợp lý vì các hệ số tự tương quan riêng phần đã được biết là hệ số của mô hình tự hồi qui.

- Theo lý thuyết khi ACF có chính xác một đỉnh khác 0 ở độ trễ 1 và tắt về 0 hoàn toàn sau 1, đồng thời PACF giảm nhanh dần theo hàm mũ (hoàn toàn âm hoặc liên tục đổi dấu) thì MA(1) là lựa chọn tốt. Tuy nhiên nhiều khi do bị nhiễu bởi sai số mà dữ liệu không hình thành rõ rệt các thể hiện trên ACF và PACF như thế, ví dụ ACF có thể có đỉnh đầu tiên khác 0 ý nghĩa và một vài hệ số ở độ trễ lớn hơn 1 vẫn khác 0 đáng kể còn PACF không cho thấy giảm rõ ràng theo kiểu mũ.

Hình 15.16 Minh họa cho ACF và PACF lý thuyết khi nhận dạng MA(1), chú ý là với hình dưới ta có cả 2 minh họa cho tình huống PACF giảm theo dạng hàm mũ âm hoàn toàn hoặc tình huống PACF giảm theo dạng hàm mũ đổi dấu liên tục.

Hình 15.16



- Theo lý thuyết ACF có chính xác 2 đỉnh khác 0 ở độ 1, 2 và tắt về 0 sau đó đồng thời PACF giảm theo dạng sóng hình sin tần tần thì mô hình phù hợp là MA(2), nếu so sánh bạn sẽ thấy đây là tình huống ngược lại của AR(2) thuần túy.

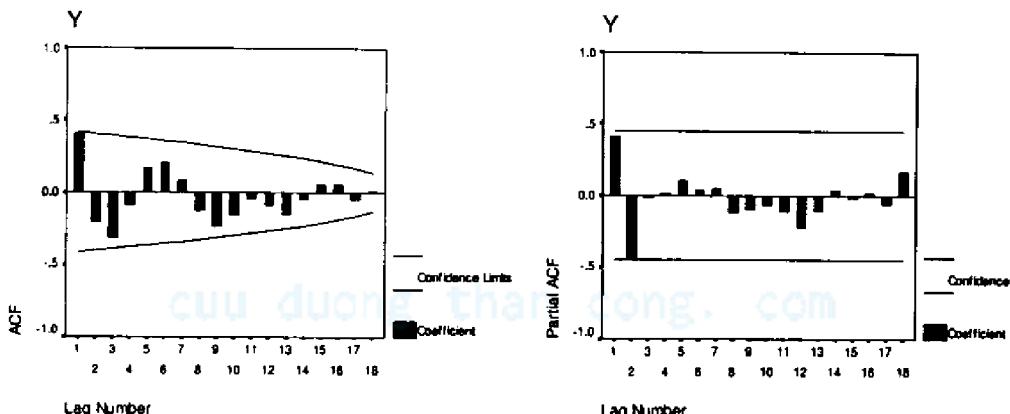
Tổng quát hơn một bộ dữ liệu thời gian phù hợp mô hình MA(q) sẽ có một ACF với các hệ số bằng 0 ngay sau trễ q và PACF tắt theo sóng hình

sin hoặc hàm mũ. Cũng chú ý luôn là với sự hiện diện của nhiễu thì việc nhận dạng một MA(q) phù hợp đối với các chuỗi dữ liệu thực không dễ dàng và rõ ràng.

Để vận dụng phương pháp này chúng ta sử dụng lại chuỗi thời gian được thu thập từ năm 1987 đến năm 2006 về tổng sản lượng bột ngọt tiêu thụ (đơn vị tính là tấn) đã được dùng trong Chương 14.

Đồ thị mô tả chuỗi dữ liệu gốc cho tình huống này đã được vẽ trong Hình 15.2(a) mô tả một chuỗi thời gian dừng. Chúng ta tiếp tục khảo sát ACF và PACF để có nhận định về mô hình phù hợp có thể áp dụng cho mục đích dự báo.

Hình 15.17



ACF có chính xác một định gần như khác 0 tại trễ 1 và sau đó nhanh chóng tắt về 0, do các nhiễu mà PACF không tạo thành hàm mũ giảm rõ ràng. Chúng ta thấy có thể chọn hàm MA(1) cho chuỗi thời gian này.

Kết quả xử lý của SPSS như sau:

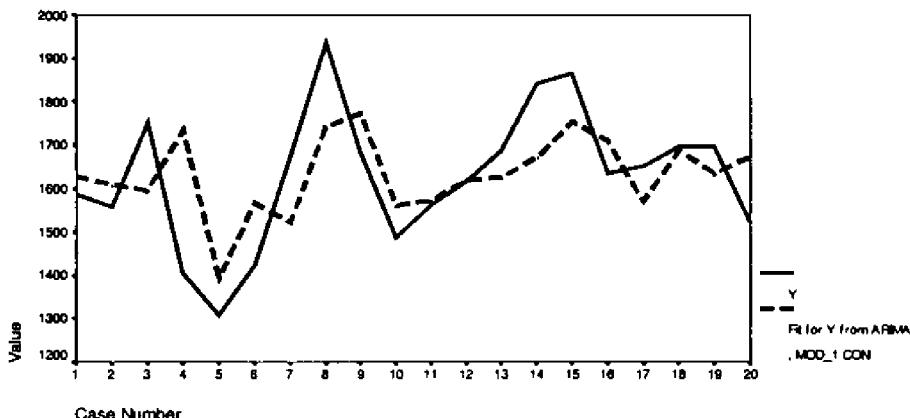
Variables in the Model:

	B	SEB	T-RATIO	APPROX. PROB.
MA1	-.75005	.221563	-3.385279	.00329725
CONSTANT	1628.63640	50.477726	32.264457	.00000000

Kết quả dự đoán cho năm 2007 được SPSS xác định là 1514,608 tấn.

Dưới đây là đồ thị minh họa đồng thời giá trị gốc và giá trị dự báo, bạn đọc có thể tự mình vẽ đồ thị so sánh đồng thời giá trị gốc với giá trị dự báo tạo ra từ phương pháp san bằng hàm mũ đơn giản và so sánh với đồ thi dưới bạn sẽ nhận thấy mô hình MA(1) cho kết quả tốt hơn hẳn, kiểm tra bằng tiêu chí MSE cũng cho cùng một kết luận vì MSE tương ứng với phương pháp MA(1) bé hơn nhiều MSE của phương pháp san bằng hàm mũ.

Hình 15.18



15.4.3 Các quá trình phối hợp tự hồi qui – trung bình trượt (ARMA)

15.4.3.1 Phương trình

Trên thực tế có những mô hình dự báo cho chuỗi thời gian là sự kết hợp đồng thời của quá trình trung bình trượt và tự hồi qui với bậc bất kỳ. Mô hình mô hình phối hợp trung bình trượt - tự hồi qui có dạng phương trình sau còn được gọi một cách tổng quát là mô hình ARMA(p,q)

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q}$$

Trong đó:

- $\phi_1; \phi_2; \dots; \phi_p$ gọi là các tham số của mô hình tự hồi qui.
- $Y_{t-1}; Y_{t-2}; \dots; Y_{t-p}$ là các biến trễ 1 đến p thời đoạn
- e_{t-k} là thành phần sai số ở thời đoạn $t-k$ có tính chất không tương quan qua các thời kì
- $\theta_1; \theta_2; \dots; \theta_q$ là các tham số trung bình trượt.
- c cũng đại diện cho hằng số.

Giả sử nếu chúng ta có quá trình ARMA(1,1) thì phương trình được viết lại như sau: $Y_t = c + \phi_1 Y_{t-1} + e_t - \theta_1 e_{t-1}$

15.4.3.2 Khảo sát dấu hiệu nhận dạng mô hình tự hồi qui - trung bình trượt

Nếu dữ liệu phù hợp với mô hình ARMA, ACF lý thuyết sẽ giảm nhanh theo dạng hàm mũ và PACF giảm theo một dạng vượt trội bởi dạng phân hủy mũ, khi đó ARMA (1,1) tạm thời là một lựa chọn tốt. Nếu ACF giảm thật nhanh đột ngột sau độ trễ q thì ta nên chọn MA(q) và nếu PACF giảm thật nhanh đột ngột sau trễ p thì ta nên xác định AR(p). Nếu cả hai giảm đột ngột như nhau thì theo kinh nghiệm có nhà thống kê thấy rằng MA(q) thường hay cho kết quả tốt hơn AR(p), nhưng tốt nhất ta xem xét cả hai mô hình và lựa chọn mô hình tốt nhất bằng các kỹ thuật giúp lựa chọn mô hình.

15.5 MÔ HÌNH BOX – JENKINS ARIMA CHO CHUỖI KHÔNG DỪNG VÀ DỰ BÁO

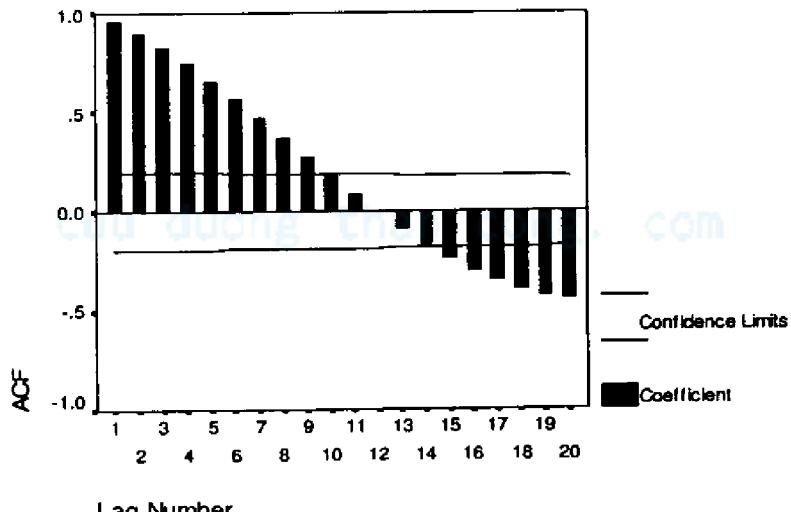
Chúng ta đã nghiên cứu xong mô hình Box – Jenkins cho chuỗi không dừng, ta cũng biết rằng khi chuỗi có tính không dừng (mà đây là tình huống hay gặp) thì ta dùng biện pháp lấy sai phân để xử lý tính không dừng và khi đó yếu tố thứ 3 phải được kết hợp vào mô hình là yếu tố sai phân, lúc đó ta có mô hình tổng quát ARIMA (p, d, q).

Xét trường hợp đơn giản nhất ARIMA (1,1,1) là một chuỗi dữ liệu có tính không dừng về trung bình, sau khi lấy sai phân bậc 1 thì chuỗi trở thành dừng và kết quả phân tích ACF, PACF của chuỗi dừng cho thấy AR(1) kết hợp MA(1) là phù hợp với dạng vận động của chuỗi dữ liệu dừng, do vậy mô hình ARIMA tổng quát lúc này là $p = 1; d = 1; q = 1$. Trong đó $d = 1$ cho ta biết ta lấy sai phân bậc 1 trên chuỗi dữ liệu gốc.

Cần biết rằng mô hình ARIMA (p, d, q) sẽ sinh ra vô số dạng mẫu hình trong ACF và PACF cho nên không thể đưa ra qui luật lý thuyết cho ACF và PACF để nhận dạng mô hình ARIMA cụ thể. Tuy nhiên cũng biết rằng trên cùng một chuỗi thời gian có thể dùng nhiều mô hình dự báo với kết cấu khác nhau để mang lại những kết quả dự báo như nhau cho nên đừng nản lòng khi nghĩ rằng chúng ta đang mò kim đáy bể.

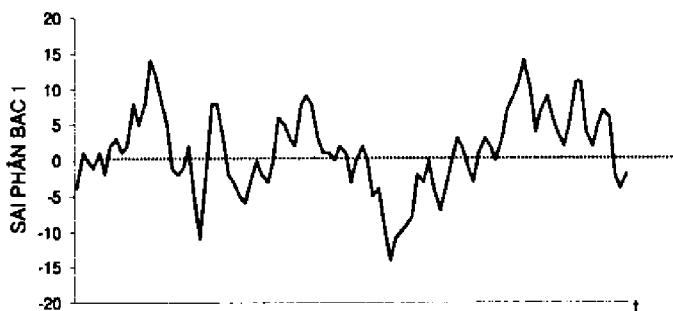
Ví dụ: Người ta theo dõi số người truy cập vào máy chủ Internet mỗi phút trong vòng 100 phút liên tục và lưu trữ thông tin dạng chuỗi thời gian. Phân tích ban đầu trên ACF của dữ liệu gốc cho thấy cho bằng chứng rõ ràng về chuỗi không dừng khi ACF có dạng giảm dần rất chậm.

Hình 15.19



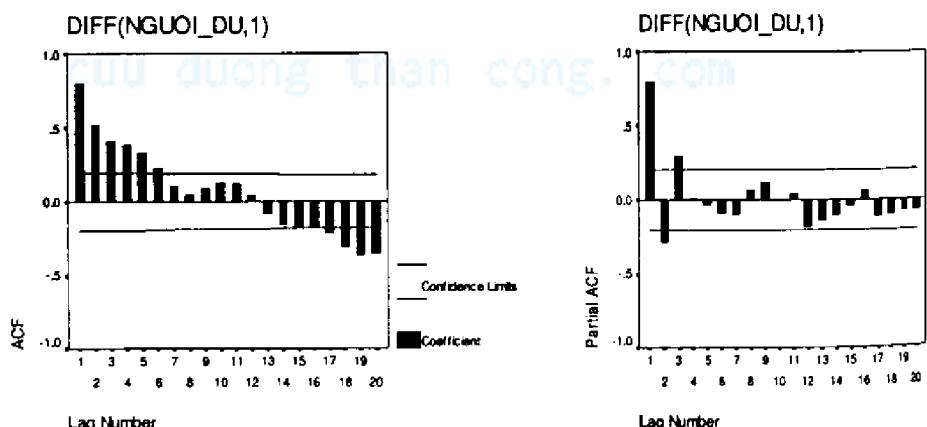
Do vậy chúng ta lấy sai phân bậc 1 của dữ liệu gốc và xem xét lại đồ thị thời gian của chuỗi sai phân này. Đồ thị cho thấy chuỗi sau khi lấy sai phân bậc 1 đã trở nên dừng.

Hình 15.20 Đồ thị của chuỗi sai phân bậc 1



Đồng thời nghiên cứu ACF và PACF của chuỗi đã lấy sai phân để xác định một mô hình Box – Jenkins phù hợp trên chuỗi đã dừng.

Hình 15.21



Chú ý là ACF thể hiện một dạng giảm phoi hợp của hàm mũ và sóng hình sin, trong PACF lại có đúng 3 đỉnh khác 0 và gần như tất hết về sau trễ 3. Vậy thì chuỗi dữ liệu dừng phù hợp với mô hình AR(3) → mô hình tổng quát có thể phù hợp cho dữ liệu gốc là ARIMA (3,1,0)

Khi dùng SPSS chạy mô hình này bạn phải chú ý là đưa chuỗi dữ liệu gốc vào xử lý chứ không phải đưa chuỗi đã lấy sai phân vào xử lý, sau khi đưa chuỗi gốc vào bạn khai báo cho phần mềm biết thông tin về d là 1, p là 3 và q là 0.

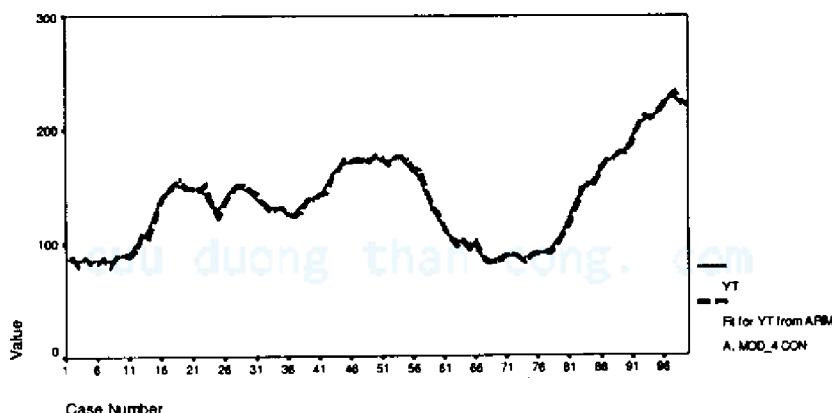
Sau đây là kết quả xử lý mô hình ARIMA (3,1,0) cho chuỗi dữ liệu thời gian về người dùng Internet:

Variables in the Model:

	B	SEB	T-RATIO	APPROX. PROB.
AR1	1.1459977	.0964728	11.878978	.00000000
AR2	-.6592933	.1363958	-4.833676	.00000515
AR3	.3346051	.0961145	3.481316	.00075560
CONSTANT	.9799166	1.6773273	.584213	.56046167

Chúng ta cũng vẫn có thể yêu cầu SPSS tính toán các giá trị dự báo cho các thời đoạn tương lai như mong muốn. Sau đây là đồ thị vẽ đồng thời hai giá trị thực tế và dự báo. Như các bạn thấy, về mặt trực quan mô hình dự báo tỏ ra rất phù hợp với dữ liệu thực tế.

Hình 15.22



Chú ý là kinh nghiệm cho thấy một chuỗi thời gian không dừng nhưng sau khi lấy sai phân bậc 1 mà ACF và PACF đều cho thấy các hệ số nằm trong đường giới hạn thì ARIMA (0,1,0) là lựa chọn tốt nhất.

15.6 MÔ HÌNH BOX-JENKINS CHO CHUỖI THỜI GIAN CÓ TÍNH MÙA VỤ

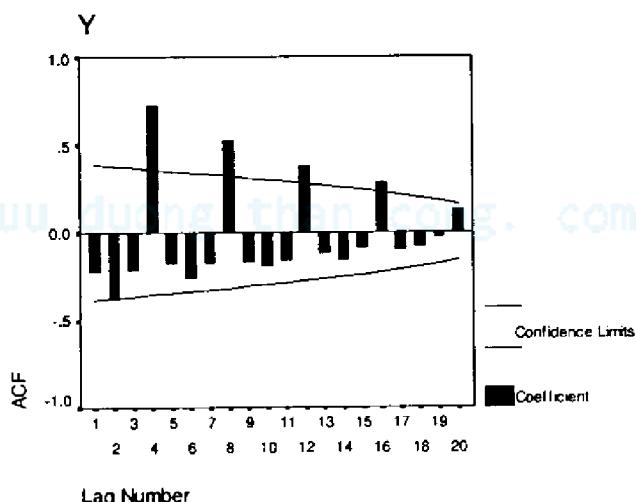
15.6.1 Nhận dạng tính mùa trong một chuỗi thời gian

Trong nội dung của Chương 14 chúng ta đã nắm được khái niệm về tính mùa, đối với chuỗi dữ liệu dừng tính mùa có thể được nhận dạng bằng cách xem ACF, ta biết rằng các hệ số r_k (với k lớn hơn 2 hoặc 3) mà khác 0 một cách có ý nghĩa đều ám chỉ sự hiện diện của một dạng thức nào đó trong dữ liệu nên để nhận dạng tính mùa người ta phải xem xét các hệ số r_k cao, nếu dạng thức mùa theo quí là nhất quán thì các hệ số tự tương quan của các độ trễ cách nhau 4 bậc sẽ có giá trị dương lớn biểu thị sự

hiện diện của tính mùa, còn nếu nó không khác 0 một cách có ý nghĩa thì có thể kết luận là các quý cách nhau 1 năm là không tương quan hay không có một dạng thức nhất quán từ năm này sang năm khác, dữ liệu như vậy thì không có tính mùa.

Trong phương pháp Holt – Winter ở Chương 14 chúng ta có xem xét một bộ dữ liệu có tính mùa là tình hình xuất khẩu qua các quý của một công ty, dữ liệu được lưu trữ qua 6 năm. Nếu sử dụng phương pháp hàm số mũ đơn giản để dự báo ta nhận định được là có tính mùa trong phần dư do phương pháp hàm số mũ đơn giản không xử lý được tính mùa. Mùa ở đây là quý, bây giờ ta xem chuỗi sai số do phương pháp dự báo hàm số mũ đơn giản tạo ra như một chuỗi thời gian bất kỳ và vẽ ACF cho nó với 20 độ trễ, ta thấy các r_4 , r_8 , r_{12} , r_{16} có giá trị dương khác 0 vượt trội, điều đó thể hiện tính mùa theo quý rõ rệt trong phần dư.

Hình 15.23

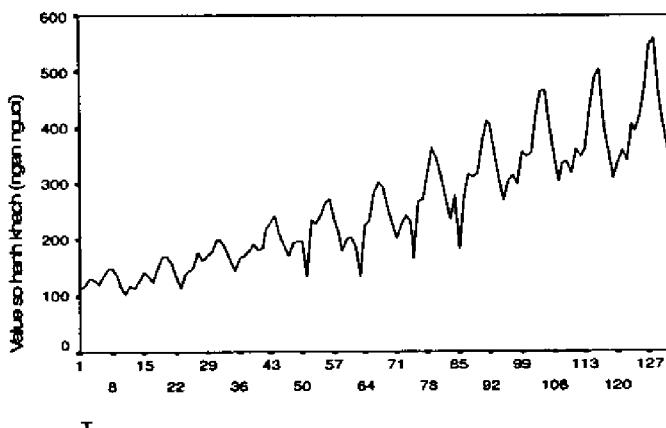


Tuy nhiên quan sát ACF không phải lúc nào cũng dễ dàng cho bạn thấy ngay tính mùa bởi vì tính mùa nhiều khi không phải là dạng thức hiện hữu duy nhất trong chuỗi, giả dụ chuỗi còn có thêm tính xu thế cùng tính mùa thì ACF sẽ cho thấy một loạt các r_k dương, tính xu thế càng mạnh thì hệ số r_k càng lớn nhiều khi lần át làm cho không nhận thấy rõ các r_k dương thể hiện tính mùa. Vì thế nguyên tắc là nếu dữ liệu không dùng cần phải chuyển sang chuỗi dừng trước khi xác định tính mùa.

Để minh họa nhận định này, chúng ta xem xét một ví dụ kinh điển được trình bày bởi chính Box và Jenkins. Ghi nhận số liệu về tổng số hành khách tháng (ngàn người) trong các chuyến bay quốc tế từ năm 1949 đến 1959.

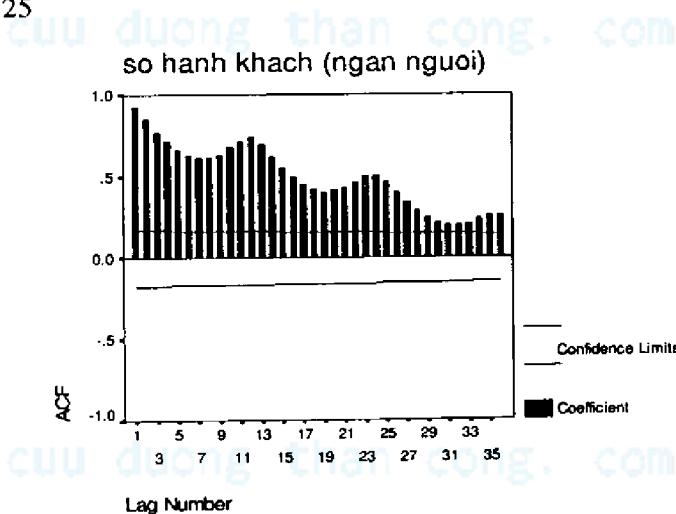
Đồ thị chuỗi gốc theo thời gian cho chúng ta thấy dữ liệu không chỉ không dừng về trung bình và phương sai mà còn có tính mùa.

Hình 15.24



ACF trên dữ liệu gốc cho thấy chuỗi không dừng, hơn nữa còn có đỉnh cao tại các trễ tương ứng với 12, 24, 36 cho thấy sự kết hợp giữa tính không dừng và tính mùa.

Hình 15.25



15.6.2 Biến đổi chuỗi thời gian có tính mùa thành chuỗi thời gian dừng và dự báo

Trong nội dung chương này chúng ta tiếp tục thảo luận cách sử dụng sai phân để biến một chuỗi thời gian không dừng có tính mùa thành chuỗi dừng. Nếu gặp tình huống này thì phương pháp lấy “sai phân bình thường

bậc 1 và sai phân có tính mùa bậc 1" được mô tả trong bảng sau sẽ tạo ra một chuỗi dừng. Trong đó L là số thời đoạn trong một vòng chu kỳ ví dụ 12 nếu là mùa tháng và 4 nếu mùa là quý.

Bảng 15.6

Y_t	$Y_t - Y_{t-1}$	$(Y_t - Y_{t-1}) - (Y_{t-L} - Y_{t-L-1})$
Y_1	-	-
Y_2	$Y_2 - Y_1$	-
...		-
Y_L	$Y_L - Y_{L-1}$	-
Y_{L+1}	$Y_{L+1} - Y_L$	-
Y_{L+2}	$Y_{L+2} - Y_{L+1}$	$(Y_{L+2} - Y_{L+1}) - (Y_2 - Y_1)$
Y_{L+3}	$Y_{L+3} - Y_{L+2}$	$(Y_{L+3} - Y_{L+2}) - (Y_3 - Y_2)$
Y_n	$Y_n - Y_{n-1}$	$(Y_n - Y_{n-1}) - (Y_{n-L} - Y_{n-L-1})$

Tức là sau khi lấy sai phân bậc 1 để làm dừng tính xu hướng, ta tiếp tục lấy sai phân bậc 1 có tính mùa trên chuỗi kết quả vừa tính được để loại tính mùa vụ, rồi sau đó tiến hành khảo sát chuỗi cuối cùng này.

Đôi khi, chúng ta tính sai phân bậc 1 có tính mùa $Y_t - Y_{t-L}$ là đã có thể đạt được cả hai mục đích làm dừng chuỗi và loại tính xu thế.

Bảng 15.7

Y_t	$Y_t - Y_{t-L}$
Y_1	-
Y_2	-
...	-
Y_L	-
Y_{L+1}	$Y_{L+1} - Y_1$
Y_{L+2}	$Y_{L+2} - Y_2$
...	
Y_n	$Y_n - Y_{n-L}$

Quá trình xử lý để nhận dạng một mô hình Box-Jenkins phù hợp với chuỗi thời gian vừa có tính xu thế vừa có tính mùa đòi hỏi người dự báo phải có kinh nghiệm và tính phán đoán tốt, nhưng chúng ta cũng hãy tham khảo một vài nguyên tắc định hướng sau:

Bước 1 Làm cho chuỗi dừng nếu nó không dừng (người ta còn gọi tên khác là tịnh hóa dữ liệu) Đồ thị dạng đường của dữ liệu gốc vẽ theo trình tự thời gian sẽ là một chỉ dẫn hữu ích về việc chuỗi thời gian của chúng ta

có bản chất dừng hay không dừng về trung bình và phương sai để chọn phương pháp tịnh hóa phù hợp.

Sai phân (mùa hay không mùa) sẽ tịnh hóa một chuỗi không dừng trung bình và log tự nhiên hay chuyển hóa căn bậc 2 hoặc bậc 4 sẽ tịnh hóa một chuỗi không dừng phương sai (chú ý là mô hình Box-Jenkins chỉ dùng được với chuỗi dừng về phương sai).

Bước 2 Xem xét thành phần không có tính mùa vụ Sự khảo sát trên ACF và PACF (của chuỗi thu được sau cùng qua các biến đổi ở bước 1) tại các trễ không là bội số của độ dài mùa L sẽ giúp kết luận AR hay MA là khả thi cho thành phần không có tính mùa

Bước 3 Xem xét thành phần có tính mùa vụ Sự khảo sát trên ACF và PACF (của chuỗi thu được sau cùng qua các biến đổi ở bước 1) tại các trễ là bội số của độ dài mùa L sẽ giúp kết luận AR hay MA là khả thi cho thành phần có tính mùa, tuy lúc này việc nhận diện không dễ dàng và rõ ràng như ở thành phần không tính mùa nhưng chúng ta vẫn có thể áp dụng qui tắc cơ bản để khảo sát dạng AR hay MA đã biết trên các trễ là bội số của mùa.

Mô hình tổng quát lúc này là ARIMA $(p,d,q)(P,D,Q)^L$

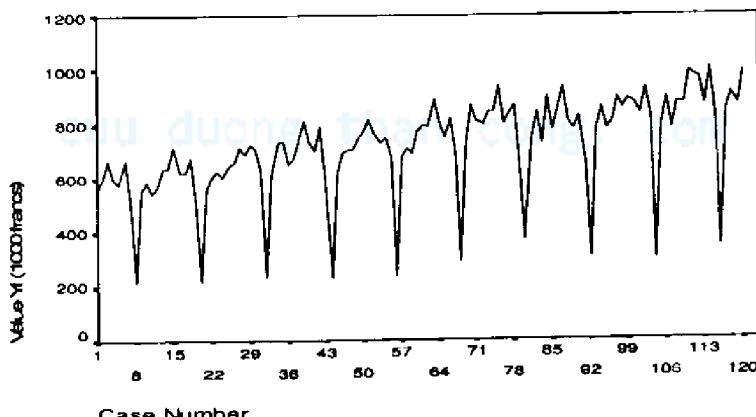
cuuduongthancong.com

Thành phần không mùa của mô hình

Thành phần mùa của mô hình

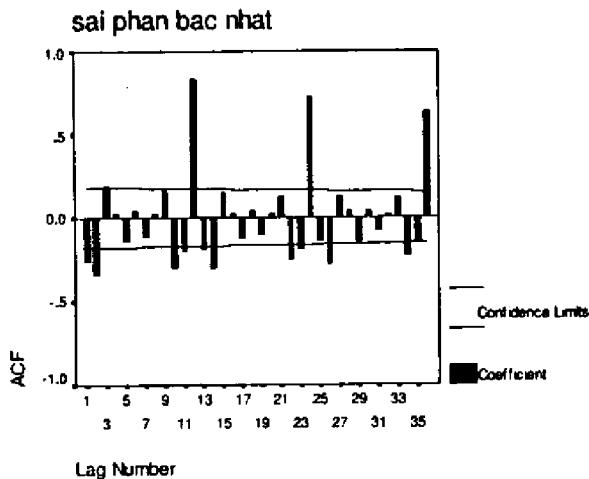
Ví dụ: Doanh số hàng tháng của ngành công nghiệp giấy (đơn vị tính ngàn Frans Pháp) được ghi nhận từ đầu năm 1963 đến cuối năm 1972. Đồ thị chuỗi dữ liệu theo thời gian có dạng như sau:

Hình 15.26



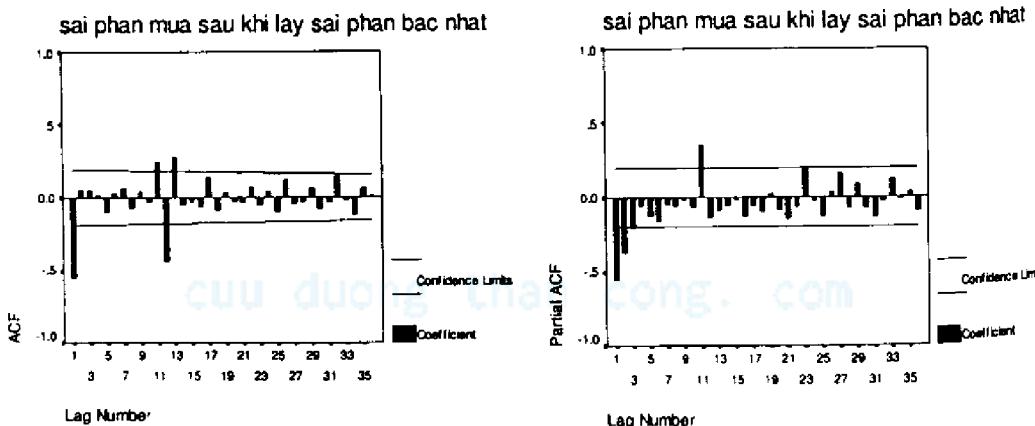
Trên đồ thị tính mùa thể hiện rất rõ bên cạnh tính xu hướng, chúng ta lấy sai phân bậc 1 trên chuỗi gốc, vẽ ACF của chuỗi đã lấy sai phân bình thường bậc 1 này thì thấy kết quả như sau

Hình 15.27



Các hệ số tự tương quan ở các trễ là bội số của 12 (mùa ở đây là tháng) khác 0 và cho thấy một sự giảm rất chậm, như vậy tính mùa vẫn tồn tại, ta lấy tiếp sai phân bậc 1 tính mùa trên chuỗi sai phân bình thường bậc 1 và dựng ACF, PACF cho chuỗi này.

Hình 15.28



Khảo sát ACF và PACF cho thấy các đỉnh mùa trên ACF đã biến mất, chuỗi lúc này đã đạt yêu cầu và ta có thể căn cứ trên ACF và PACF mà khảo sát dạng hàm phù hợp.

Trước tiên ta nhận định mô hình tổng quát phải có dạng như sau ARIMA $(p,1,q)(P,1,Q)$ ¹²

Ta xác định các tham số p,d và P, Q như thế nào?

- PACF cho thấy một sự giảm sút theo hàm mũ tại 3 hệ số trễ đầu tiên kết hợp với 1 hệ số đầu tiên khác 0 tại ACF chỉ ra nên chọn MA(1) cho thành phần không có tính mùa, vậy $p=0$ và $q=1$.
- Trên ACF có một r_{12} khác 0 và một sự suy giảm khá nhanh trong các hệ số ở các trễ mùa còn lại, ngoài ra trễ 12, 24, 36 của PACF giảm dần → cho chúng ta nhận định mô hình MA(1) là phù hợp cho thành phần mùa. Tức là $P=0$ và $Q=1$

Vậy cuối cùng mô hình tổng quát được chọn là ARIMA $(0,1,1)(0,1,1)$ ¹²

Trên SPSS muốn xử lý được mô hình đã chọn chúng ta đưa chuỗi gốc vào xử lý, khai báo các hệ số ương ứng $p=0, d=1, q=1$ tại các khu vực của thành phần không mùa và $P=0, D=1, Q=1$ tại thành phần mùa, ta có được kết quả như sau:

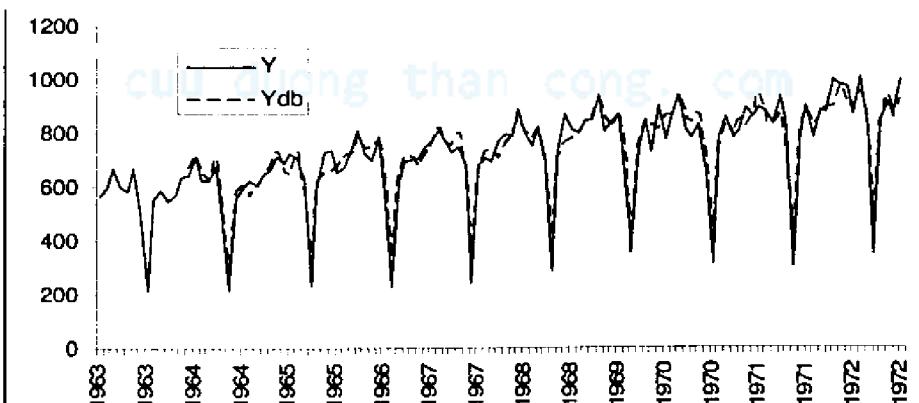
Variables in the Model:

	B	SEB	T-RATIO	APPROX. PROB.
MA1	.84007193	.05576948	15.063292	.00000000
SMA1	.63489836	.10323373	6.150106	.00000001
CONSTANT	.02424350	.33501270	.072366	.94244969

Trong dãy kết quả trên SMA1 là tham số của thành phần MA của thành phần mùa.

Giá trị dự báo cũng được SPSS tính toán theo chỉ định của người sử dụng và cho thấy vào tháng 1 năm 1973 doanh số là 944,64 ngàn Frans Pháp, giá trị dự báo của tháng 2 năm 1973 là 993,539 ngàn Frans Pháp. Đồ thị dưới trình bày giá trị gốc và giá trị dự báo, chú ý là giá trị dự báo chỉ có từ giai đoạn thứ 13 trở đi.

Hình 15.28



Trên màn hình kết quả Output các bạn sẽ thấy còn rất nhiều thông tin, một phần trong các thông tin đó là để phục vụ cho việc lựa chọn một mô hình ARIMA phù hợp, rõ ràng mô hình ARIMA gồm một họ rất lớn các mô hình và căn cứ trên sự thể hiện của dữ liệu chúng ta phải lựa chọn một mô hình hợp lý trong một số mô hình thử nghiệm bằng các kỹ thuật lựa chọn mô hình phù hợp sau:

- Sử dụng ma trận hệ số tương quan giữa các tham số.

Mặc dù các tham số ước lượng cuối cùng của mô hình Box-Jenkins luôn tương quan nhau nhưng những tương quan mạnh (trên 0,9) có thể làm cho các ước lượng điểm có chất lượng kém và lúc đó chúng ta phải xem xét đến việc loại biến đứng bên phải mô hình mà liên quan đến tham số đó.

- Kiểm định t xem một tham số nào đó có là vô nghĩa không.

Đi kèm với ước lượng các tham số trong mô hình Box-Jenkins, SPSS cũng cung cấp sai số chuẩn và giá trị t cho kiểm định.

H_0 : Tham số tổng thể của mô hình bằng zero

H_1 : Tham số tổng thể của mô hình khác zero,

Chúng ta cũng căn cứ vào giá trị xác suất để kết luận có nên đưa tham số đó vào mô hình không theo cách thức quen thuộc như kiểm định ý nghĩa của các hệ số hồi qui riêng của mô hình hồi qui. Nếu một tham số nào đó tỏ ra không cần có mặt trong mô hình thì chúng ta nên xem xét nghiêm túc việc loại bỏ tham số ra khỏi mô hình vì sự có mặt của nó có thể tạo ra một mô hình không thỏa đáng, các mô hình thỏa đáng rất quan trọng vì nó sẽ tạo ra kết quả dự báo chính xác hơn.

- Chuẩn đoán phần dư

Một phương pháp tốt để kiểm tra sự thích hợp của mô hình Box-Jenkins thử nghiệm là phân tích phần dư thu được từ mô hình bằng ACF, PACF hay dùng kiểm định Box-Ljung. Nếu kiểm định có ý nghĩa thì nghĩa là mô hình thử nghiệm đã tạo ra phần dư đó không phù hợp và cần cân nhắc một mô hình khác. Hình dạng của các đỉnh khác 0 trong ACF và PACF có thể là một gợi ý về cách cải tiến mô hình ví dụ các đỉnh khác 0 tại các trễ mùa vụ chỉ ra mô hình cũ thiếu thành phần mùa vụ. Hay các đỉnh khác 0 tại những độ trễ nhỏ gợi ý nên tăng thêm thành phần AR hoặc MA không mùa vụ của mô hình

Khi có một vài mô hình đều tỏ ra hợp lý với các chỉ tiêu trên, ta cần lựa chọn tiếp bằng cách xét đến một số thước đo sự thích hợp toàn bộ

- Sai số chuẩn (Standard error)

Với công thức $S = \sqrt{[SSE/(n-n_p)]}$ trong đó n_p là số tham số trong mô hình. Giá trị sai số chuẩn càng nhỏ thì tính thích hợp của toàn bộ mô hình càng tốt vì nó tạo ra các giá trị dự báo chính xác hơn.

- Chỉ tiêu AIC (Akaike Information Criteria)

Chú ý rằng bản thân AIC chỉ có ý nghĩa khi đem so sánh với một mô hình thích hợp hóa khác trên cùng dãy thời gian, mô hình nào có AIC cực tiểu nhất là tốt nhất. Một sự khác biệt ≤ 2 đơn vị giữa các AIC không được xem là đáng kể và lúc đó bạn có thể chấp nhận để chọn một mô hình đơn giản hoặc dễ hiểu hơn. Hoặc đôi khi có thể chấp nhận một mô hình không có AIC bé nhất nhưng có biểu hiện tốt trong kết quả chuẩn đoán phần dư.

Từ lý thuyết trên chúng ta xem lại các kết quả của mô hình ARIMA (3,1,0) về lượng người truy cập Internet, chú ý là trong một họ các mô hình thì người ta đã tiến hành so sánh và thấy mô hình ARIMA (3,1,0) cho AIC bé nhất.

FINAL PARAMETERS:

Number of residuals	99	Sai số chuẩn
Standard error	3.1191805	
Log likelihood	-251.84844	
AIC	511.69688	Tiêu chuẩn AIC
SBC	522.07736	

Analysis of Variance:

	DF	Adj. Sum of Squares	Residual Variance
Residuals	95	938.95538	9.7292871

Variables in the Model:

	B	SEB	T-RATIO	APPROX. PROB.
AR1	1.1459977	.0964728	11.878978	.00000000
AR2	-.6592933	.1363958	-4.833676	.00000515
AR3	.3346051	.0961145	3.481316	.00025560
CONSTANT	.9799166	1.6773273	.584213	.56167

Giá trị xác suất để phục vụ cho kiểm định xem một tham số nào đó có vô nghĩa không

Covariance Matrix:

	AR1	AR2	AR3
AR1	.00930699	-.00977352	.00286045
AR2	-.00977352	.01860383	-.00968528
AR3	.00286045	-.00968528	.00923800

Correlation Matrix:

	AR1	AR2	AR3
AR1	1.0000000	-.7427543	.3084896
AR2	-.7427543	1.0000000	-.7387915
AR3	.3084896	-.7387915	1.0000000

Matrice hệ số
tương quan giữa
các tham số

```
Regressors Covariance Matrix:  
    CONSTANT  
CONSTANT      2.8134268  
Regressors Correlation Matrix:  
    CONSTANT  
CONSTANT      1.0000000  
The following new variables are being created:  
Name          Label  
FIT_1         Fit for NGUOITC from ARIMA, MOD_7 CON  
ERR_1         Error for NGUOITC from ARIMA, MOD_7 CON  
LCL_1         95% LCL for NGUOITC from ARIMA, MOD_7 CON  
UCL_1         95% UCL for NGUOITC from ARIMA, MOD_7 CON  
SEP_1         SE of fit for NGUOITC from ARIMA, MOD_7 CON
```

cuu duong than cong. com

cuu duong than cong. com

TÀI LIỆU THAM KHẢO

TÀI LIỆU NUỐC NGOÀI

1. Aczel, Amir D., *Complete Business Statistics*, Irwin, 1993
2. Berenson M. L., Levine D. M., Krehbiel T. C., *Basic Business Statistics*, 9th Edition, Pearson Prentice Hall, 2004.
3. Berenson, Mark L., Levine, David M., *Basic Business Statistics*, 4th Edition, Prentice Hall 1986
4. Bernard, Russel H., *Research Methods in Anthropology, Qualitative and Quantitative Approaches*, 3rd Edition, Altamira Press, 2002
5. Blalock, Hubert M., *Social Statistics*, 2nd Edition, McGraw-Hill Book Company, 1972
6. Bowerman, Bruce L., O'Connell, Richard T., *Forecasting and Time Series, An Applied Approach*, 3rd Edition, Duxbury Press, 1993
7. Cochran, William G., *Sampling Techniques*, John Wiley & Sons Inc., 1953
8. Cooper D. R., Schindler P.S., *Business Research Methods*, 8th Edition, McGraw-Hill, 2003.
9. Croxton, Frederick E., Cowden, Dudley J., Klein, Sidney, *General Applied Statistics*, 3rd Edition, Prentice Hall of India, New Dehli, 1988
10. Groebner, D.F., Shannon P.W., Fry P.C., and Smith K.D. (2005), *Business Statistics, A Decision Making Approach*, Updated 6th Edition, Pearson Prentice Hall.
11. Gujarati, Damodar N., *Basic Econometrics*, 3rd Edition, Mc-GrawHill Higher Education, 1995.
12. Levin, Richard I., *Statistics for Management*, 4th Edition, Prentice Hall, 1989
13. Makridakis, Spyros, Wheelwright, Steven C., Hyndman, Rob J., *Forecasting: Method and Applications*, 3rd Edition, John Wiley & Sons, 1998
14. Neuman William Lawrence, *Social Research Methods, Qualitative and Quantitative Approaches*, Allyn & Bacon, 2000
15. Newbold, Paul, *Statistics for Business and Economics*, Prentice Hall International, Inc., 1991
16. Pindyck, Robert S., Rubinfeld, Daniel L., *Econometric Models and Economic Forecasts*, 3rd. Edition, McGraw-Hill College, 1990
17. Ramanathan, Ramu, *Introductory Econometric with Applications*, 5th Edition, Harcourt, 2002 (Bản dịch tiếng Việt của Chương trình giảng dạy kinh tế Fulbright tại Việt Nam).
18. Sirkin Mark R., *Statistics for the Social Sciences*, 2nd Edition, Sage Publications, 1999

19. Siegel, Sidney, Castellan N. John, Jr., *Nonparametric Statistics for the Behavioral Sciences*, 2nd Edition, McGraw Hill, 1988
20. Wonnacott, Thomas H., Wonnacott, Ronald J., *Introductory Statistics for Business and Economics*, 4th Edition, John Wiley & Son, 1990
21. Wyatt, Woodrow W., Bridges, Charles M., *Statistics for the Behavioral Sciences*, DC Health and Company, 1967

TÀI LIỆU TRONG NUỐC

22. Đào Hữu Hò, *Thống kê xã hội học (Xác suất thống kê B)*, NXB Đại Học Quốc Gia Hà Nội, 2000
23. Dương Thiệu Tông, *Thống kê ứng dụng trong nghiên cứu khoa học giáo dục*, NXB Khoa Học Xã Hội, 2005
24. Hà Văn Sơn và các tác giả, *Giáo trình Lý thuyết thống kê ứng dụng trong quản trị và kinh tế*, NXB Thông Kê, 2004
25. Hoàng Trọng, Chu Nguyễn Mộng Ngọc, *Phân tích dữ liệu nghiên cứu với SPSS*, NXB Thông Kê, 2005
26. Nguyễn Cao Văn và Trần Thái Ninh, *Giáo trình Lý thuyết xác suất và thống kê toán*, NXB Thông kê, 2005
27. Nguyễn Ngọc Kiềng, *Thống kê học trong nghiên cứu khoa học*, NXB Giáo Dục, 1996
28. Tập thể tác giả, *Từ Điển Thống Kê*, Tổng cục thống kê, 1977
29. Trần Bá Nhẫn, Đinh Thái Hoàng, *Lý thuyết thống kê ứng dụng trong quản trị kinh doanh và nghiên cứu kinh tế*, NXB Thông Kê, 1998
30. Trần Chung Ngọc, Trần Văn Tươi, *Thống kê căn bản*, Phân khoa Khoa học xã hội, ĐH Vạn Hạnh, 1974.
31. Trần Văn Thắng và các tác giả, *Giáo trình Lý thuyết thống kê*, NXB Thông Kê, 1998
32. Võ Thị Thanh Lộc, *Thống kê ứng dụng và dự báo trong kinh doanh và kinh tế*, NXB Thông Kê, 1998

INTERNET

Quyền sách có sử dụng một số tư liệu trích từ các trang web

33. www.gso.gov.vn
34. www.pso.hochiminhcity.gov.vn
35. www.ssc.gov.vn
36. www.eximbank.com.vn

PHỤ LỤC

Bảng tra 1: Phân phối chuẩn

Bảng tra 2: Phân phối Student

Bảng tra 3: Phân phối Chi bình phương

Bảng tra 4: Phân phối F

Bảng tra 5: Phân phối Hartley

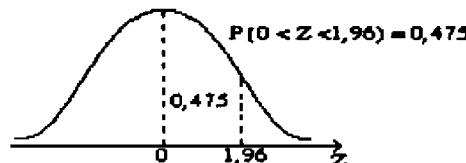
Bảng tra 6: Kiểm định dấu và hạng WILCOXON

Bảng tra 7: Kiểm định tổng và hạng WILCOXON

Bảng tra 8: Durbin Watson

Bảng tra 9: Phân phối Tukey

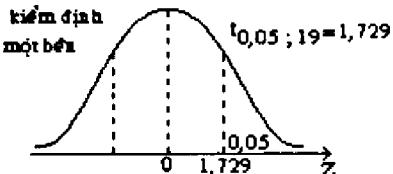
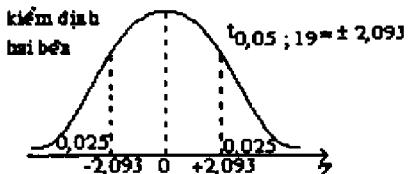
BẢNG TRA 1 : PHÂN PHỐI BÌNH THƯỜNG CHUẨN HÓA



Bảng phân phối bình thường chuẩn hóa $P(0 < Z < z_b)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

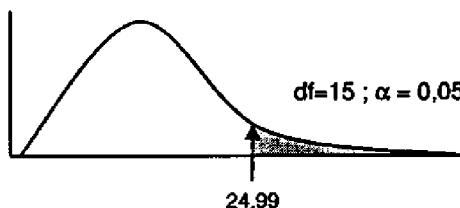
BẢNG TRA 2 : PHÂN PHỐI STUDENT



Bảng tra phân phối t Student cho giá trị t tới hạn $t_{(\alpha, df)}$ sao cho $P(t > t_{(\alpha, df)}) = \alpha$

df	Giá trị α						
	0.125	0.1	0.075	0.05	0.025	0.01	0.005
1	2.4142	3.0777	4.1653	6.3138	12.7062	31.8205	63.6567
2	1.6036	1.8856	2.2819	2.9200	4.3027	6.9646	9.9248
3	1.4226	1.6377	1.9243	2.3534	3.1824	4.5407	5.8409
4	1.3444	1.5332	1.7782	2.1318	2.7764	3.7469	4.6041
5	1.3009	1.4759	1.6994	2.0150	2.5706	3.3649	4.0321
6	1.2733	1.4398	1.6502	1.9432	2.4469	3.1427	3.7074
7	1.2543	1.4149	1.6166	1.8946	2.3646	2.9980	3.4995
8	1.2403	1.3968	1.5922	1.8595	2.3060	2.8965	3.3554
9	1.2297	1.3830	1.5737	1.8331	2.2622	2.8214	3.2498
10	1.2213	1.3722	1.5592	1.8125	2.2281	2.7638	3.1693
11	1.2145	1.3634	1.5476	1.7959	2.2010	2.7181	3.1058
12	1.2089	1.3562	1.5380	1.7823	2.1788	2.6810	3.0545
13	1.2041	1.3502	1.5299	1.7709	2.1604	2.6503	3.0123
14	1.2001	1.3450	1.5231	1.7613	2.1448	2.6245	2.9768
15	1.1967	1.3406	1.5172	1.7531	2.1314	2.6025	2.9467
16	1.1937	1.3368	1.5121	1.7459	2.1199	2.5835	2.9208
17	1.1910	1.3334	1.5077	1.7396	2.1098	2.5669	2.8982
18	1.1887	1.3304	1.5037	1.7341	2.1009	2.5524	2.8784
19	1.1866	1.3277	1.5002	1.7291	2.0930	2.5395	2.8609
20	1.1848	1.3253	1.4970	1.7247	2.0860	2.5280	2.8453
21	1.1831	1.3232	1.4942	1.7207	2.0796	2.5176	2.8314
22	1.1815	1.3212	1.4916	1.7171	2.0739	2.5083	2.8188
23	1.1802	1.3195	1.4893	1.7139	2.0687	2.4999	2.8073
24	1.1789	1.3178	1.4871	1.7109	2.0639	2.4922	2.7969
25	1.1777	1.3163	1.4852	1.7081	2.0595	2.4851	2.7874
26	1.1766	1.3150	1.4834	1.7056	2.0555	2.4786	2.7787
27	1.1756	1.3137	1.4817	1.7033	2.0518	2.4727	2.7707
28	1.1747	1.3125	1.4801	1.7011	2.0484	2.4671	2.7633
29	1.1739	1.3114	1.4787	1.6991	2.0452	2.4620	2.7564
30	1.1731	1.3104	1.4774	1.6973	2.0423	2.4573	2.7500
31	1.1723	1.3095	1.4761	1.6955	2.0395	2.4528	2.7440
32	1.1716	1.3086	1.4749	1.6939	2.0369	2.4487	2.7385
33	1.1710	1.3077	1.4738	1.6924	2.0345	2.4448	2.7333
34	1.1703	1.3070	1.4728	1.6909	2.0322	2.4411	2.7284
35	1.1698	1.3062	1.4718	1.6896	2.0301	2.4377	2.7238
36	1.1692	1.3055	1.4709	1.6883	2.0281	2.4345	2.7195
37	1.1687	1.3049	1.4701	1.6871	2.0262	2.4314	2.7154
38	1.1682	1.3042	1.4692	1.6860	2.0244	2.4286	2.7116
39	1.1677	1.3036	1.4685	1.6849	2.0227	2.4258	2.7079
40	1.1673	1.3031	1.4677	1.6839	2.0211	2.4233	2.7045
50	1.1639	1.2987	1.4620	1.6759	2.0086	2.4033	2.6778
60	1.1616	1.2958	1.4582	1.6706	2.0003	2.3901	2.6603
70	1.1600	1.2938	1.4555	1.6669	1.9944	2.3808	2.6479
80	1.1588	1.2922	1.4535	1.6641	1.9901	2.3739	2.6387
90	1.1578	1.2910	1.4519	1.6620	1.9867	2.3685	2.6316
100	1.1571	1.2901	1.4507	1.6602	1.9840	2.3642	2.6259

BÀNG TRA 3: PHÂN PHỐI CHI BÌNH PHƯƠNG (χ^2)



Bậc tự do (df)	$\chi^2_{0,995}$	$\chi^2_{0,990}$	$\chi^2_{0,975}$	$\chi^2_{0,950}$	$\chi^2_{0,900}$
1	0.000039	0.000157	0.000982	0.003932	0.015791
2	0.010025	0.020100	0.050636	0.102586	0.210721
3	0.071723	0.114832	0.215795	0.351846	0.584375
4	0.206984	0.297107	0.484419	0.710724	1.063624
5	0.411751	0.554297	0.831209	1.145477	1.610309
6	0.675733	0.872083	1.237342	1.635380	2.204130
7	0.989251	1.239032	1.689864	2.167349	2.833105
8	1.344403	1.646506	2.179725	2.732633	3.489537
9	1.734911	2.087889	2.700389	3.325115	4.168156
10	2.155845	2.558199	3.246963	3.940295	4.865178
11	2.603202	3.053496	3.815742	4.574809	5.577788
12	3.073785	3.570551	4.403778	5.226028	6.303796
13	3.565042	4.106900	5.008738	5.891861	7.041500
14	4.074659	4.660415	5.628724	6.570632	7.789538
15	4.600874	5.229356	6.262123	7.260935	8.546753
16	5.142164	5.812197	6.907664	7.961639	9.312235
17	5.697274	6.407742	7.564179	8.671754	10.085183
18	6.264766	7.014903	8.230737	9.390448	10.864937
19	6.843923	7.632698	8.906514	10.117006	11.650912
20	7.433811	8.260368	9.590772	10.850799	12.442601
21	8.033602	8.897172	10.282907	11.591316	13.239596
22	8.642681	9.542494	10.982330	12.338009	14.041490
23	9.260383	10.195689	11.688534	13.090505	14.847954
24	9.886199	10.856349	12.401146	13.848422	15.658679
25	10.519647	11.523951	13.119707	14.611396	16.473405
26	11.160218	12.198177	13.843881	15.379163	17.291880
27	11.807655	12.878468	14.573373	16.151395	18.113889
28	12.461281	13.564666	15.307854	16.927876	18.939235
29	13.121067	14.256406	16.047051	17.708381	19.767740
30	13.786682	14.953484	16.790756	18.492667	20.599245
40	20.706577	22.164201	24.433058	26.509296	29.050516
50	27.990825	29.706725	32.357385	34.764236	37.688637
60	35.534397	37.484796	40.481707	43.187966	46.458885
70	43.275305	45.441700	48.757536	51.739263	55.328945
80	51.171933	53.539983	57.153152	60.391459	64.277842
100	67.327533	70.064995	74.221882	77.929442	82.358127

BẢNG TRA 3 (Tiếp theo)

Bậc tự do (df)	$\chi^2_{0.100}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$	$\chi^2_{0.005}$
1	2.7055	3.8415	5.0239	6.6349	7.8794
2	4.6052	5.9915	7.3778	9.2104	10.5965
3	6.2514	7.8147	9.3484	11.3449	12.8381
4	7.7794	9.4877	11.1433	13.2767	14.8602
5	9.2363	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5475
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5345	20.0902	21.9549
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9872	18.3070	20.4832	23.2093	25.1881
11	17.2750	19.6752	21.9200	24.7250	26.7569
12	18.5493	21.0261	23.3367	26.2170	28.2997
13	19.8119	22.3620	24.7356	27.6882	29.8193
14	21.0641	23.6848	26.1189	29.1412	31.3194
15	22.3071	24.9958	27.4884	30.5780	32.8015
16	23.5418	26.2962	28.8453	31.9999	34.2671
17	24.7690	27.5871	30.1910	33.4087	35.7184
18	25.9894	28.8693	31.5264	34.8052	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5821
20	28.4120	31.4104	34.1696	37.5663	39.9969
21	29.6151	32.6706	35.4789	38.9322	41.4009
22	30.8133	33.9245	36.7807	40.2894	42.7957
23	32.0069	35.1725	38.0756	41.6383	44.1814
24	33.1962	36.4150	39.3641	42.9798	45.5584
25	34.3816	37.6525	40.6465	44.3140	46.9280
26	35.5632	38.8851	41.9231	45.6416	48.2898
27	36.7412	40.1133	43.1945	46.9628	49.6450
28	37.9159	41.3372	44.4608	48.2782	50.9936
29	39.0875	42.5569	45.7223	49.5878	52.3355
30	40.2560	43.7730	46.9792	50.8922	53.6719
40	51.8050	55.7585	59.3417	63.6908	66.7660
50	63.1671	67.5048	71.4202	76.1538	79.4898
60	74.3970	79.0820	83.2977	88.3794	91.9518
70	85.5270	90.5313	95.0231	100.4251	104.2148
80	96.5782	101.8795	106.6285	112.3288	116.3209
100	118.4980	124.3421	129.5613	135.8069	140.1697

BẢNG TRA 4 : PHÂN PHỐI FISHER ($\alpha = 0.05$)

df	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03

df	11	12	13	14	15	16	17	18	19	20
1	242.98	243.91	244.69	245.36	245.95	246.46	246.92	247.32	247.69	248.01
2	19.40	19.41	19.42	19.42	19.43	19.43	19.44	19.44	19.44	19.45
3	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66
4	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80
5	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56
6	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87
7	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44
8	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15
9	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94
10	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77
11	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65
12	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54
13	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46
14	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39
15	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33
16	2.46	2.42	2.40	2.37	2.35	2.33	2.32	2.30	2.29	2.28
17	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23
18	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22	2.20	2.19
19	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.16
20	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15	2.14	2.12
25	2.20	2.16	2.14	2.11	2.09	2.07	2.05	2.04	2.02	2.01
30	2.13	2.09	2.06	2.04	2.01	1.99	1.98	1.96	1.95	1.93
35	2.07	2.04	2.01	1.99	1.96	1.94	1.92	1.91	1.89	1.88
40	2.04	2.00	1.97	1.95	1.92	1.90	1.89	1.87	1.85	1.84
45	2.01	1.97	1.94	1.92	1.89	1.87	1.86	1.84	1.82	1.81
50	1.99	1.95	1.92	1.89	1.87	1.85	1.83	1.81	1.80	1.78

BẢNG TRA 4 (Tiếp theo)

($\alpha = 0.025$)

df	1	2	3	4	5	6	7	8	9	10
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
35	5.48	4.11	3.52	3.18	2.96	2.80	2.68	2.58	2.50	2.44
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
45	5.38	4.01	3.42	3.09	2.86	2.70	2.58	2.49	2.41	2.35
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32

df	11	12	13	14	15	16	17	18	19	20
1	973.03	976.71	979.84	982.53	984.87	986.92	988.73	990.35	991.80	993.10
2	39.41	39.41	39.42	39.43	39.43	39.44	39.44	39.44	39.45	39.45
3	14.37	14.34	14.30	14.28	14.25	14.23	14.21	14.20	14.18	14.17
4	8.79	8.75	8.71	8.68	8.66	8.63	8.61	8.59	8.58	8.56
5	6.57	6.52	6.49	6.46	6.43	6.40	6.38	6.36	6.34	6.33
6	5.41	5.37	5.33	5.30	5.27	5.24	5.22	5.20	5.18	5.17
7	4.71	4.67	4.63	4.60	4.57	4.54	4.52	4.50	4.48	4.47
8	4.24	4.20	4.16	4.13	4.10	4.08	4.05	4.03	4.02	4.00
9	3.91	3.87	3.83	3.80	3.77	3.74	3.72	3.70	3.68	3.67
10	3.66	3.62	3.58	3.55	3.52	3.50	3.47	3.45	3.44	3.42
11	3.47	3.43	3.39	3.36	3.33	3.30	3.28	3.26	3.24	3.23
12	3.32	3.28	3.24	3.21	3.18	3.15	3.13	3.11	3.09	3.07
13	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.95
14	3.09	3.05	3.01	2.98	2.95	2.92	2.90	2.88	2.86	2.84
15	3.01	2.96	2.92	2.89	2.86	2.84	2.81	2.79	2.77	2.76
16	2.93	2.89	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.68
17	2.87	2.82	2.79	2.75	2.72	2.70	2.67	2.65	2.63	2.62
18	2.81	2.77	2.73	2.70	2.67	2.64	2.62	2.60	2.58	2.56
19	2.76	2.72	2.68	2.65	2.62	2.59	2.57	2.55	2.53	2.51
20	2.72	2.68	2.64	2.60	2.57	2.55	2.52	2.50	2.48	2.46
25	2.56	2.51	2.48	2.44	2.41	2.38	2.36	2.34	2.32	2.30
30	2.46	2.41	2.37	2.34	2.31	2.28	2.26	2.23	2.21	2.20
35	2.39	2.34	2.30	2.27	2.23	2.21	2.18	2.16	2.14	2.12
40	2.33	2.29	2.25	2.21	2.18	2.15	2.13	2.11	2.09	2.07
45	2.29	2.25	2.21	2.17	2.14	2.11	2.09	2.07	2.04	2.03
50	2.26	2.22	2.18	2.14	2.11	2.08	2.06	2.03	2.01	1.99

BẢNG TRA 4 (Tiếp theo)

($\alpha = 0.01$)

df	1	2	3	4	5	6	7	8	9	10
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70

df	11	12	13	14	15	16	17	18	19	20
1	6083.32	6106.32	6125.86	6142.67	6157.28	6170.10	6181.43	6191.53	6200.58	6208.73
2	99.41	99.42	99.42	99.43	99.43	99.44	99.44	99.44	99.45	99.45
3	27.13	27.05	26.98	26.92	26.87	26.83	26.79	26.75	26.72	26.69
4	14.45	14.37	14.31	14.25	14.20	14.15	14.11	14.08	14.05	14.02
5	9.96	9.89	9.82	9.77	9.72	9.68	9.64	9.61	9.58	9.55
6	7.79	7.72	7.66	7.60	7.56	7.52	7.48	7.45	7.42	7.40
7	6.54	6.47	6.41	6.36	6.31	6.28	6.24	6.21	6.18	6.16
8	5.73	5.67	5.61	5.56	5.52	5.48	5.44	5.41	5.38	5.36
9	5.18	5.11	5.05	5.01	4.96	4.92	4.89	4.86	4.83	4.81
10	4.77	4.71	4.65	4.60	4.56	4.52	4.49	4.46	4.43	4.41
11	4.46	4.40	4.34	4.29	4.25	4.21	4.18	4.15	4.12	4.10
12	4.22	4.16	4.10	4.05	4.01	3.97	3.94	3.91	3.88	3.86
13	4.02	3.96	3.91	3.86	3.82	3.78	3.75	3.72	3.69	3.66
14	3.86	3.80	3.75	3.70	3.66	3.62	3.59	3.56	3.53	3.51
15	3.73	3.67	3.61	3.56	3.52	3.49	3.45	3.42	3.40	3.37
16	3.62	3.55	3.50	3.45	3.41	3.37	3.34	3.31	3.28	3.26
17	3.52	3.46	3.40	3.35	3.31	3.27	3.24	3.21	3.19	3.16
18	3.43	3.37	3.32	3.27	3.23	3.19	3.16	3.13	3.10	3.08
19	3.36	3.30	3.24	3.19	3.15	3.12	3.08	3.05	3.03	3.00
20	3.29	3.23	3.18	3.13	3.09	3.05	3.02	2.99	2.96	2.94
25	3.06	2.99	2.94	2.89	2.85	2.81	2.78	2.75	2.72	2.70
30	2.91	2.84	2.79	2.74	2.70	2.66	2.63	2.60	2.57	2.55
35	2.80	2.74	2.69	2.64	2.60	2.56	2.53	2.50	2.47	2.44
40	2.73	2.66	2.61	2.56	2.52	2.48	2.45	2.42	2.39	2.37
45	2.67	2.61	2.55	2.51	2.46	2.43	2.39	2.36	2.34	2.31
50	2.63	2.56	2.51	2.46	2.42	2.38	2.35	2.32	2.29	2.27

BẢNG TRẠM 5 : PHÂN PHỐI HARTLEY

($\alpha = 0.05$)

df	k	2	3	4	5	6	7	8	9	10	11	12
2	39,0	87,5	142	202	266	333	403	475	550	626	704	
3	15,4	27,8	39,2	50,7	62,0	72,9	83,5	93,9	104	114	124	
4	9,60	15,5	20,6	25,2	29,5	33,6	37,5	41,1	44,6	48,0	51,4	
5	7,15	10,8	13,7	16,3	18,7	20,8	22,9	24,7	26,5	28,2	29,9	
6	5,82	8,38	10,4	12,1	13,7	15,0	16,3	17,5	18,6	19,7	20,7	
7	4,99	6,94	8,44	9,70	10,8	11,8	12,7	13,5	14,3	15,1	15,8	
8	4,43	6,00	7,18	8,12	9,03	9,78	10,5	11,1	11,7	12,2	12,7	
9	4,03	5,34	6,31	7,11	7,80	8,41	8,95	9,45	9,91	10,3	10,7	
10	3,72	4,85	5,67	6,34	6,92	7,42	7,87	8,28	8,66	9,01	9,34	
12	3,28	4,16	4,79	5,30	5,72	6,09	6,42	6,72	7,00	7,25	7,48	
15	2,86	3,54	4,01	4,37	4,68	4,95	5,19	5,40	5,59	5,77	5,93	
20	2,46	2,95	3,29	3,54	3,76	3,94	4,10	4,24	4,37	4,49	4,59	
30	2,07	2,40	2,61	2,78	2,91	3,02	3,12	3,21	3,29	3,36	3,39	
60	1,67	1,85	1,96	2,04	2,11	2,17	2,22	2,26	2,30	2,33	2,36	
∞	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

cuuduongthancong.com

($\alpha = 0.01$)

df	k	2	3	4	5	6	7	8	9	10	11	12
2	199	448	729	1036	1362	1705	2063	2432	2813	3204	3605	
3	47,5	85	120	151	184	21(6)	24(9)	28(1)	31(0)	33(7)	36(1)	
4	23,2	37	49	59	69	79	89	97	106	113	120	
5	14,9	22	28	33	38	42	46	50	54	57	60	
6	11,1	15,5	19,1	22	25	27	30	32	34	36	37	
7	8,89	12,1	14,5	16,5	18,4	20	22	23	24	26	27	
8	7,50	9,9	11,7	13,2	14,5	15,8	16,9	17,9	18,9	19,8	21	
9	6,54	8,5	9,9	11,1	12,1	13,1	13,9	14,7	15,3	16,0	16,6	
10	5,85	7,4	8,6	9,6	10,4	11,1	11,8	12,4	12,9	13,4	13,9	
12	4,91	6,1	6,9	7,6	8,2	8,7	9,1	9,5	9,9	10,2	10,6	
15	4,07	4,9	5,5	6,0	6,4	6,7	7,1	7,3	7,5	7,8	8,0	
20	3,32	3,8	4,3	4,6	4,9	5,1	5,3	5,5	5,6	5,8	5,9	
30	2,63	3,0	3,3	3,4	3,6	3,7	3,8	3,9	4,0	4,1	4,2	
60	1,96	2,2	2,3	2,4	2,4	2,5	2,5	2,6	2,6	2,7	2,7	
∞	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

BẢNG TRA 6 : Cận dưới và cận trên của W trong kiểm định dấu và hạng WILCOXON

	Một bên $\alpha = .05$ Hai bên $\alpha = .10$	$\alpha = 0.25$ $\alpha = .05$	$\alpha = 0.1$ $\alpha = .02$	$\alpha = .005$ $\alpha = .01$
(cận dưới; cận trên)				
5	0;15	—	—	—
6	2;19	0;21	—	—
7	3;25	2;26	0;28	—
8	5;31	3;33	1;35	0;36
9	8;37	5;40	3;42	1;44
10	10;45	8;47	5;50	3;52
11	13;53	10;56	7;59	5;61
12	17;61	13;65	10;68	7;71
13	21;70	17;74	12;79	10;81
14	25;80	21;84	16;89	13;92
15	30;90	25;95	19;101	16;104
16	35;101	29;107	23;113	19;117
17	41;112	34;119	27;126	23;130
18	47;124	40;131	32;139	27;144
19	53;137	46;144	37;153	32;158
20	60;150	52;158	43;167	37;173

cuuduongthancong.com

BẢNG TRA 7: Cận dưới và cận trên của T_1 trong kiểm định tổng và hạng WILCOXON

n_2	Mức ý nghĩa α		n_1						
	Một bên	Hai bên	4	5	6	7	8	9	10
4	0,05	0,10	11;25						
	0,025	0,05	10;26						
	0,01	0,02							
	0,005	0,01							
5	0,05	0,10	12;28	19;36					
	0,025	0,05	11;29	17;38					
	0,01	0,02	10;30	16;39					
	0,005	0,01		15;40					
6	0,05	0,10	13;31	20;40	28;50				
	0,025	0,05	12;32	18;42	26;52				
	0,01	0,02	11;33	17;43	24;54				
	0,005	0,01	10;34	16;44	23;55				
7	0,05	0,10	14;34	21;44	29;55	39;66			
	0,025	0,05	13;35	20;45	27;57	36;69			
	0,01	0,02	11;37	18;47	25;59	34;71			
	0,005	0,01	10;38	16;49	24;60	32;73			
8	0,05	0,10	15;37	23;47	31;59	41;71	51;85		
	0,025	0,05	14;38	21;49	29;61	38;74	49;87		
	0,01	0,02	12;40	19;51	27;63	35;77	45;91		
	0,005	0,01	11;41	17;53	25;65	34;78	43;93		
9	0,05	0,10	16;40	24;51	33;63	43;76	54;90	66;105	
	0,025	0,05	14;38	22;53	31;65	40;79	51;93	62;109	
	0,01	0,02	13;43	20;55	28;68	37;82	47;97	59;112	
	0,005	0,01	11;45	18;57	26;70	35;84	45;99	56;115	
10	0,05	0,10	17;43	26;54	35;67	45;81	56;96	69;111	82;128
	0,025	0,05	15;45	23;57	32;70	42;84	53;99	65;115	78;132
	0,01	0,02	13;47	21;59	29;73	39;87	49;103	61;119	74;136
	0,005	0,01	12;48	19;61	27;75	37;89	47;105	58;105	71;139

BÀNG TRẠM 8 : Durbin Watson

Durbin-Watson Statistic: 5 %

n	K = 1		K = 2		K = 3		K = 4		K = 5		K = 10		K = 15	
	dL	dU	dL	dU	dL	dU								
15	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21				
16	1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62	2.15	.16	3.30		
17	1.13	1.38	1.02	1.54	.90	1.71	.78	1.90	.67	2.10	.20	3.18		
18	1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71	2.06	.24	3.07		
19	1.18	1.40	1.08	1.53	.97	1.68	.86	1.85	.75	2.02	.29	2.97		
20	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99	.34	2.89	.06	3.68
21	1.22	1.42	1.13	1.54	1.03	1.67	.93	1.81	.83	1.96	.38	1.81	.09	3.58
22	1.24	1.43	1.15	1.54	1.05	1.66	.96	1.80	.86	1.94	.42	2.73	.12	3.55
23	1.26	1.44	1.17	1.54	1.08	1.66	.99	1.79	.90	1.92	.47	2.67	.15	3.41
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	.93	1.90	.51	2.61	.19	3.33
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89	.54	2.57	.22	3.25
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	.98	1.88	.58	2.51	.26	3.18
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86	.62	2.47	.29	3.11
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85	.65	2.43	.33	3.05
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84	.68	2.40	.36	2.99
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83	.71	2.36	.39	2.94
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83	.74	2.33	.43	2.99
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82	.77	2.31	.46	2.84
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81	.80	2.28	.49	2.80
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81	.82	2.26	.52	2.75
35	1.40	1.52	1.34	1.53	1.28	1.65	1.22	1.73	1.16	1.80	.85	2.24	.55	2.72
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80	.87	2.22	.58	2.68
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80	.89	2.20	.60	2.65
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79	.91	2.18	.63	2.61
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79	.93	2.16	.65	2.59
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79	.95	2.15	.68	2.56
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77	1.11	2.04	.88	2.35
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77	1.17	2.01	.96	2.28
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77	1.22	1.98	1.03	2.23
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77	1.27	1.96	1.09	2.18
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77	1.30	1.95	1.14	2.15
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77	1.34	1.94	1.18	2.12
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77	1.37	1.93	1.22	2.09
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77	1.40	1.92	1.26	2.07
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78	1.42	1.91	1.29	2.06
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78	1.44	1.90	1.32	2.04
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78	1.46	1.90	1.35	2.03

cuuduongthancong.com

BÀNG TRA 8 : Durbin Watson (tiếp theo)

Durbin-Watson Statistic: 10%											
	K = 1		K = 2		K = 3		K = 4		K = 5		
n	dL	dU									
15	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.39	1.96	
16	.84	1.09	.74	1.25	.63	1.44	.53	1.66	.44	1.90	
17	.87	1.10	.77	1.25	.67	1.43	.57	1.63	.48	1.85	
18	.90	1.12	.80	1.26	.71	1.42	.61	1.60	.52	1.80	
19	.93	1.13	.83	1.26	.74	1.41	.65	1.58	.56	1.77	
20	.95	1.15	.86	1.27	.77	1.41	.68	1.57	.60	1.74	
21	.97	1.16	.89	1.27	.80	1.41	.72	1.55	.63	1.71	
22	1.00	1.17	.91	1.28	.83	1.40	.75	1.54	.66	1.69	
23	1.02	1.19	.94	1.29	.86	1.40	.77	1.53	.70	1.67	
24	1.04	1.20	.96	1.30	.88	1.41	.80	1.53	.72	1.66	
25	1.05	1.21	.98	1.30	.90	1.41	.83	1.52	.75	1.65	
26	1.07	1.22	1.00	1.31	.93	1.41	.85	1.52	.78	1.64	
27	1.09	1.23	1.02	1.32	.95	1.41	.88	1.51	.81	1.63	
28	1.10	1.24	1.04	1.32	.97	1.41	.90	1.51	.83	1.62	
29	1.12	1.25	1.05	1.33	.99	1.42	.92	1.51	.85	1.61	
30	1.13	1.26	1.07	1.34	1.01	1.42	.94	1.51	.88	1.61	
31	1.15	1.27	1.08	1.34	1.02	1.42	.96	1.51	.90	1.60	
32	1.16	1.28	1.10	1.35	1.04	1.43	.98	1.51	.92	1.60	
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	.94	1.59	
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	.95	1.59	
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	.97	1.59	
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	.99	1.59	
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59	
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58	
39	1.24	1.34	1.19	1.39	1.14	1.4	1.09	1.52	1.03	1.58	
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58	
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58	
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59	
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59	
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60	
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61	
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61	
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62	
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62	
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63	
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64	
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64	
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65	

BÀNG TRA 9 : PHÂN PHỐI TUKEY (Studentized Range Distribution)

($\alpha = 0,05$)

n-k	k									
	2	3	4	5	6	7	8	9	10	
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07	
2	6.08	8.33	9.80	10.88	11.74	12.44	13.03	13.54	13.99	
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	
20	2.95	3.58	3.96	4.23	4.46	4.62	4.77	4.90	5.01	
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.56	4.65	
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	

n-k	k									
	11	12	13	14	15	16	17	18	19	20
1	50.59	51.96	53.20	54.33	55.36	56.32	57.22	58.04	58.83	59.56
2	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.37	16.57	16.77
3	9.72	9.95	10.2	10.3	10.5	10.7	10.8	11.0	11.1	11.2
4	8.03	8.21	8.37	8.52	8.66	8.79	8.91	9.03	9.13	9.23
5	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21
6	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59
7	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.10	7.17
8	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87
9	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64
10	5.72	5.83	5.93	6.03	6.11	6.19	6.27	6.34	6.40	6.47
11	5.61	5.71	5.81	5.90	5.98	6.06	6.13	6.20	6.27	6.33
12	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21
13	5.43	5.53	5.63	5.71	5.79	5.86	5.93	5.99	6.05	6.11
14	5.36	5.46	5.55	5.64	5.71	5.79	5.85	5.91	5.97	6.03
15	5.31	5.40	5.49	5.57	5.65	5.72	5.78	5.85	5.90	5.96
16	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90
17	5.21	5.31	5.39	5.47	5.54	5.61	5.67	5.73	5.79	5.84
18	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79
19	5.14	5.23	5.31	5.39	5.46	5.53	5.59	5.65	5.70	5.75
20	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71
24	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.49	5.55	5.59
30	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.47
40	4.82	4.90	4.98	5.04	5.11	5.16	5.22	5.27	5.31	5.36
60	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24
120	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13
∞	4.55	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01

BẢNG TRA 9 : PHÂN PHỐI TUKEY (Studentized Range Distribution)

($\alpha = 0,01$)

n-k	k									
	2	3	4	5	6	7	8	9	10	
1	90.03	135	164.3	185.6	202.2	215.8	227.2	237.0	245.6	
2	14.04	19.02	22.29	24.72	26.63	28.20	29.53	30.68	31.69	
3	8.26	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69	
4	6.51	8.12	9.17	9.96	10.58	11.10	11.55	11.93	12.27	
5	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.24	
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	
8	4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	
11	4.39	5.14	5.62	5.97	6.25	6.48	6.67	6.84	6.99	
12	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	
15	4.17	4.83	5.25	5.56	5.80	5.99	6.16	6.31	6.44	
16	4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.35	
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	
24	3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	
40	3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	
60	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	
120	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	
∞	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	

n-k	k									
	11	12	13	14	15	16	17	18	19	20
1	253	260	266	272	277	282	286	290	294	298
2	32.6	33.4	34.1	34.8	35.4	36.0	36.5	37.0	37.5	37.9
3	17.1	17.5	17.9	18.2	18.5	18.8	19.1	19.3	19.5	19.8
4	12.6	12.8	13.1	13.3	13.5	13.7	13.9	14.1	14.2	14.4
5	10.5	10.7	10.9	11.1	11.2	11.4	11.6	11.7	11.8	11.9
6	9.30	9.49	9.65	9.81	9.95	10.1	10.2	10.3	10.4	10.5
7	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65
8	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03
9	7.65	7.78	7.91	8.03	8.13	8.23	8.32	8.41	8.49	8.57
10	7.36	7.48	7.60	7.71	7.81	7.91	7.99	8.07	8.15	8.22
11	7.13	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95
12	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73
13	6.79	6.90	7.01	7.10	7.19	7.27	7.34	7.42	7.48	7.55
14	6.66	6.77	6.87	6.96	7.05	7.12	7.20	7.27	7.33	7.39
15	6.55	6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26
16	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15
17	6.38	6.48	6.57	6.66	6.73	6.80	6.87	6.94	7.00	7.05
18	6.31	6.41	6.50	6.58	6.65	6.72	6.79	6.85	6.91	6.96
19	6.25	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89
20	6.19	6.29	6.37	6.45	6.52	6.59	6.65	6.71	6.76	6.82
24	6.02	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61
30	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41
40	5.69	5.77	5.84	5.90	5.96	6.02	6.07	6.12	6.17	6.21
60	5.53	5.60	5.67	5.73	5.79	5.84	5.89	5.93	5.98	6.02
120	5.38	5.44	5.51	5.56	5.61	5.66	5.71	5.75	5.79	5.83
∞	5.23	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65

THỐNG KÊ ỨNG DỤNG
trong kinh tế - xã hội
HOÀNG TRỌNG - CHU NGUYỄN MỘNG NGỌC

NHÀ XUẤT BẢN THỐNG KÊ

98 Thụy Khê, Quận Tây Hồ, TP. Hà Nội

Điện Thoại – Fax: 04-8457290

Email: nxbthongke_cbi@fpt.vn

Chịu trách nhiệm xuất bản

TRẦN HỮU THỰC

Biên tập

HOÀNG TRỌNG

Trình bày

HOÀNG TRỌNG

Sửa bản in

MỘNG NGỌC

Bìa

VŨ XUÂN KHANH

In 2.000 cuốn khổ 16x24 cm tại Công Ty Cổ Phần In Khánh Hội

(360 Bến Vân Đồn, P1, Q4). GPXB số:85-2008/CXB/254.1 – 134/Tk

Ngày 28/08/2008. In xong và nộp lưu chiểu quý IV năm 2008