TRƯỜNG ĐẠI HỌC KINH TẾ THÀNH PHỐ HỒ CHÍ MINH
BỘ MÔN LÝ THUYẾT THỐNG KÊ - THỐNG KÊ KINH TẾ
Chủ biên: HÀ VĂN SƠN

GIÁO TRÌNH LÝ THUYẾT THỐNG KÊ

ỨNG DỤNG TRONG QUẢN TRỊ VÀ KINH TẾ

STATISTICS FOR MANAGEMENT AND ECONOMICS

NHÀ XUẤT BẢN THỐNG KÊ

TRƯỜNG ĐẠI HỌC KINH TẾ THÀNH PHỐ HỒ CHÍ MINH
BỘ MÔN LÝ THUYẾT THỐNG KÊ - THỐNG KÊ KINH TẾ
Chủ biên: HÀ VĂN SƠN

GIÁO TRÌNH

LÝ THUYẾT THỐNG KÊ

ỨNG DỤNG TRONG QUẢN TRỊ VÀ KINH TẾ
(STATISTICS FOR MANAGEMENT AND ECONOMICS)

NHÀ XUẤT BẢN THỐNG KÊ
2004

LỜI NÓI ĐẦU

Là công cụ không thể thiếu được trong hoạt động nghiên cứu và công tác thực tiễn, cho nên thống kê đã trở thành một môn học cần thiết trong hầu hết các ngành đào tạo. Trong các chuyên ngành khối kinh tế-xã hội, Lý thuyết thống kê là một môn học cơ sở bắt buộc có vị trí xứng đáng với lượng thời gian đáng kể.

Cùng với chính sách mở cửa và sự phát triển của kinh tế thị trường chịu sự điều tiết của nhà nước, tình hình kinh tế – xã hội nước ta đã có nhiều chuyển biến. Trước đây công tác thống kê diễn ra chủ yếu trong khu vực kinh tế nhà nước, trong các cơ quan thống kê nhà nước để thu thập thông tin phục vụ cho việc quản lý kinh tế xã hội của các cơ quan chính quyền các cấp.

Hiện nay công tác thống kê đã được chú ý trong các doanh nghiệp ở tất cả các ngành. Việc sử dụng các phương pháp thống kê trở nên cần thiết và phổ biến. Bên cạnh đó, trong xu hướng hội nhập với khu vực và thế giới, giáo dục đại học Việt Nam đang từng bước chuyển mình, và đào tạo thống kê cũng không nằm ngoài quỹ đạo đó. Nhu cầu về một giáo trình thống kê vừa phù hợp với điều kiện giảng dạy và học tập hiện nay, vừa thống nhất với chương trình đào tạo thống kê khá chuẩn mực tại các nước đang tỏ ra cấp bách.

Để đáp ứng yêu cầu nghiên cứu, giảng dạy và học tập của giáo viên và đông đảo sinh viên các chuyên ngành khối kinh tế - xã hội, cũng như yêu cầu tham khảo của đông đảo cựu sinh viên và những người đang làm công tác thực tế, Bộ môn lý thuyết thống kê - thống kê kinh tế tổ chức biên soạn giáo trình Lý thuyết thống kê. Giáo trình này được xây dựng với định hướng ứng dụng trong kinh tế và quản trị theo xu thế hội nhập quốc tế. Với kinh nghiệm giảng dạy được tích lũy qua nhiều năm cộng với nỗ lực nghiên cứu từ các nguồn tài liệu phong phú, giáo trình biên soạn lần này có nhiều thay đổi và bổ sung để đáp ứng yêu cầu nâng cao chất lượng đào tạo đặt ra.

Tham gia biên soạn gồm có:

- ThS. Hà Văn Sơn, chủ biên, biên soạn các chương 6,7,8.
- TS. Trần Văn Thắng biên soạn chương 1.
- TS. Mai Thanh Loan biên soạn chương 5.
- ThS. Nguyễn Văn Trãi biên soạn chương 13.
- ThS. Hoàng Trọng biên soạn chương 2,3,9,10.
- ThS. Võ Thị Lan biên soạn chương 11,12.
- ThS. Đặng Ngọc Lan biên soạn chương 4.

Chúng tôi xin chân thành cảm ơn anh Hoàng Ngọc Nhậm, anh Trần Tuấn Cường, chị Dương Xuân Bình đã đọc và góp ý cho cuốn giáo trình này.

Mặc dù các tác giả đã có nhiều cố gắng, song do khả năng có hạn, cùng với những thay đổi và bổ sung như vậy, chắc chắn việc biên soạn không tránh khỏi những thiếu sót. Chúng tôi rất biết ơn và mong nhận được những ý kiến trao đổi và đóng góp của bạn đọc để lần tái bản sau giáo trình được hoàn thiện hơn. Thư góp ý xin gửi về địa chỉ sau:

Bộ môn Lý Thuyết Thống Kê – Thống Kê Kinh Tế
Khoa Toán - Thống Kê
Đại Học Kinh Tế TP Hồ Chí Minh,
số 91 đường 3/2, quận 10, TP Hồ Chí Minh
Email: hason@uch.edu.vn hoặc hasondhkt@yahoo.com

TP.Hồ Chí Minh, những ngày đầu xuân năm 2004

Các tác giả

MỤC LỤC CHI TIẾT

CHƯƠNG 10: KIỂM ĐỊNH PHI THAM SỐ

CHƯƠNG 11: TƯƠNG QUAN VÀ HỒI QUI

CHƯƠNG 12: DÃY SỐ THỜI GIAN

CHƯƠNG 1

GIỚI THIỆU MÔN HỌC

1.1 THỐNG KÊ LÀ GÌ?

Trong công tác thực tế cũng như trong đời sống hàng ngày chúng ta thường gặp thuật ngữ "Thống kê". Thuật ngữ này có thể hiểu theo hai nghĩa:

Thứ nhất: Thống kê là các số liệu được thu thập để phản ánh các hiện tượng kinh tế - xã hội, tự nhiên, kỹ thuật. Chẳng hạn như sản lượng các loại sản phẩm chủ yếu được sản xuất ra trong nền kinh tế trong một năm nào đó. Mực nước cao nhất và thấp nhất của một dòng sông tại một địa điểm nào đó trong năm...

Thứ hai: Thống kê là hệ thống các phương pháp được sử dụng để nghiên cứu các hiện tượng kinh tế - xã hội, tự nhiên, kỹ thuật.

Thực ra khi hỏi thống kê là gì, có nhiều cách trả lời, ví dụ trả lời như sau khó có thể bắt bẻ. " Thống kê là công việc mà các nhà thống kê làm." [1]. Công việc của nhà thống kê bao gồm các hoạt động trên một phạm vi rộng, có thể tóm tắt thành các mục lớn như sau:

- Thu thập và xử lý số liệu.
- Điều tra chọn mẫu.
- Nghiên cứu mối liên hệ giữa các hiện tượng.
- Dự đoán.
- Nghiên cứu các hiện tượng trong hoàn cảnh không chắc chắn.
- Ra quyết định trong điều kiện không chắc chắn.

Một cách tổng quát, ta đi đến định nghĩa về thống kê như sau:
Thống kê là hệ thống các phương pháp dùng để thu thập, xử lý và phân tích các con số (mặt lượng) của những hiện tượng số lớn để tìm hiểu bản chất và tính quy luật vốn có của chúng (mặt chất) trong điều kiện thời gian và không gian cụ thể.

Mọi sự vật, hiện tượng đều có hai mặt chất và lượng không tách rời nhau, và khi chúng ta nghiên cứu hiện tượng, điều chúng ta muốn biết đó là bản chất

---

[1] "Statistics is what statisticians do"- Paul Newbold, Statistics for Business and Economics, Prentice- Hall International, Inc , 1995-page 1

của hiện tượng. Nhưng mặt chất thường ẩn bên trong, còn mặt lượng biểu hiện ra bên ngoài dưới dạng các đại lượng ngẫu nhiên. Do đó phải thông qua các phương pháp xử lý thích hợp trên mặt lượng của số lớn đơn vị cấu thành hiện tượng, tác động của các yếu tố ngẫu nhiên mới được bù trừ và triệt tiêu, bản chất của hiện tượng mới bộc lộ ra và ta có thể nhận thức đúng đắn bản chất, quy luật vận động của nó.

Thống kê được chia thành hai lãnh vực:

- Thống kê mô tả [1]: Bao gồm các phương pháp thu thập số liệu, mô tả và trình bày số liệu, tính toán các đặc trưng đo lường. Phần thống kê mô tả được trình bày trong các chương 2,3,4.
- Thống kê suy diễn [2]: Bao gồm các phương pháp như ước lượng, kiểm định, phân tích mối liên hệ, dự đoán...trên cơ sở các thông tin thu thập từ mẫu. Phần thống kê suy diễn được trình bày trong các chương còn lại.

Trong lĩnh vực kinh tế – xã hội, thống kê thường quan tâm nghiên cứu các hiện tượng như :

- Các hiện tượng về nguồn tài nguyên, môi trường, của cải tích lũy của đất nước, của một vùng.
- Các hiện tượng về sản xuất, phân phối, lưu thông, tiêu dùng sản phẩm.
- Các hiện tượng về dân số, nguồn lao động.
- Các hiện tượng về đời sống vật chất, văn hoá của dân cư.
- Các hiện tượng về sinh hoạt chính trị xã hội.

1.2 MỘT SỐ KHÁI NIỆM DÙNG TRONG THỐNG KÊ

1.2 .1 Tổng thể thống kê[3] và đơn vị tổng thể:

Tổng thể thống kê (còn gọi là tổng thể chung) là tập hợp các đơn vị (hay phần tử) thuộc hiện tượng nghiên cứu, cần quan sát, thu thập và phân tích mặt lượng của chúng theo một hoặc một số tiêu thức nào đó.

Các đơn vị (hay phần tử) cấu thành tổng thể thống kê gọi là đơn vị tổng thể.

Ví dụ: Muốn tính thu nhập trung bình của một hộ gia đình ở Thành Phố Hồ Chí Minh thì tổng thể sẽ là tổng số hộ của Thành Phố Hồ Chí Minh. Muốn tính chiều cao trung bình của sinh viên nam lớp X thì tổng thể sẽ là toàn bộ

---

[1] Descriptive Statistics

[2] Inferential Statistics

[3] Population

Nam sinh viên của lớp X.
Như vậy thực chất của việc xác định tổng thể thống kê là xác định 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 bộc lộ**. (Ví dụ: Tổng thể sinh viên của một trường ; Tổng thể các doanh nghiệp trên một địa bàn...)
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 xã hội ta thường gặp loại tổng thể này (ví dụ tổng thể những người đồng ý (ủng hộ) một vấn đề nào đó; Tổng thể những người ưa thích nghệ thuật cải lương...)

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 dệt trên một địa bàn thì 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.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 (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).

1.2.2 Tổng thể mẫu [1](mẫu)

Tổng thể mẫu là tổng thể bao gồm một số đơn vị được chọn ra từ tổng thể

---

[1] Sample

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.

1.2.3 Quan sát [1]

Quan sát là cơ sở để thu thập số liệu và thông tin cần nghiên cứu. Chẳng hạn trong điều tra chọn mẫu, mỗi đơn vị mẫu sẽ được tiến hành ghi chép, thu thập thông tin và được gọi là một quan sát.

1.2.4 Tiêu thức thống kê

Tiêu thức thống kê là khái niệm dùng để chỉ các đặc điểm của đơn vị tổng thể.

Ví dụ khi nghiên cứu nhân khẩu, mỗi nhân khẩu có các tiêu thức như: giới tính, độ tuổi, trình độ học vấn, nghề nghiệp, dân tộc, tôn giáo... Khi nghiên các doanh nghiệp, mỗi doanh nghiệp có các tiêu thức như: Số lượng công nhân, vốn cố định, vốn lưu động, giá trị sản xuất ...

Tiêu thức thống kê được chia thành hai loại:

- **Tiêu thức thuộc tính**: là tiêu thức phản ánh tính chất hay loại hình của đơn vị tổng thể, không có biểu hiện trực tiếp bằng các con số. Ví dụ các tiêu thức như giới tính, nghề nghiệp, tình trạng hôn nhân, dân tộc, tôn giáo... là các tiêu thức thuộc tính.
- **Tiêu thức số lượng**: là tiêu thức có biểu hiện trực tiếp bằng con số. Ví dụ: Tuổi, chiều cao, trọng lượng của con người, năng suất làm việc của công nhân...

  Các trị số cụ thể khác nhau của tiêu thức số lượng gọi là **lượng biến**.

  Ví dụ: Tuổi là tiêu thức số lượng, tuổi không phải là lượng biến. Lượng biến là: 18 tuổi, 20 tuổi, 30 tuổi...

  Lượng biến có thể phân biệt thành hai loại:

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

  Ví dụ: Số công nhân trong một doanh nghiệp. Số sản phẩm sản xuất trong ngày của một phân xưởng.

  * **Lượng biến liên tục**: là lượng biến mà 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. Năng suất của một loại cây trồng.

Các tiêu thức thuộc tính hoặc tiêu thức số lượng chỉ có hai biểu hiện không trùng nhau trên một đơn vị tổng thể, được gọi là **tiêu thức thay phiên**. Ví dụ, tiêu thức giới tính là tiêu thức thay phiên vì chỉ có hai biểu hiện là nam

---

[1] Observation

và nữ. Đối với tiêu thức có nhiều biểu hiện ta có thể chuyển thành tiêu thức thay phiên bằng cách rút gọn thành hai biểu hiện. Ví dụ, thành phần kinh tế chia thành nhà nước và ngoài nhà nước. Số công nhân của các doanh nghiệp chia thành: < 500 và ≥ 500.

1.2.5 Chỉ tiêu thống kê

Chỉ tiêu thống kê là các trị số phản ánh các đặc điểm, các tính chất cơ bản của tổng thể thống kê 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.

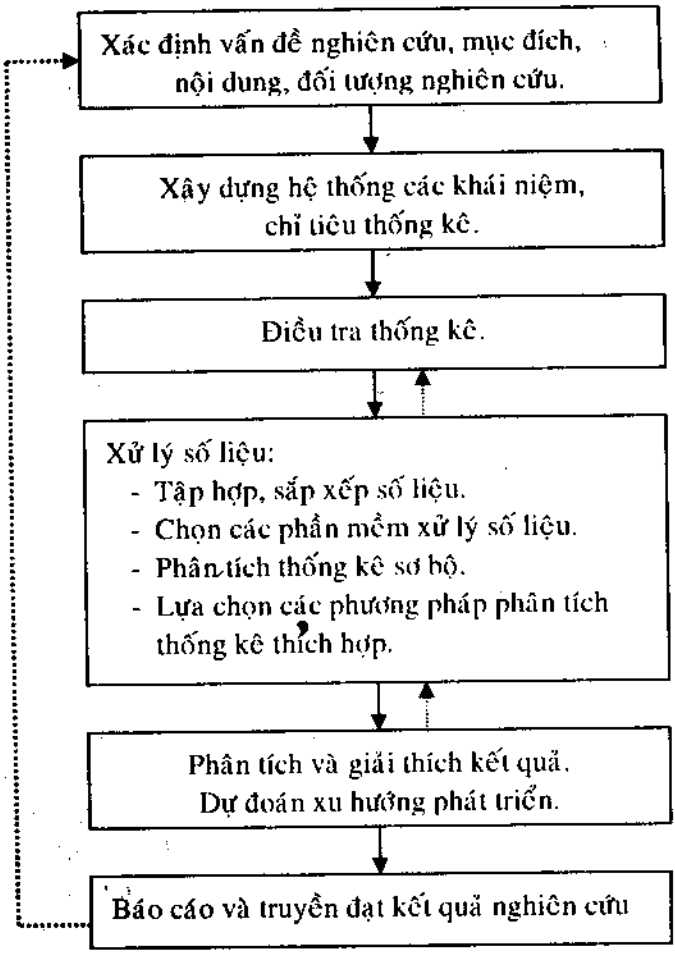
1.2.5.1 Chỉ tiêu khối lượng: là các chỉ tiêu biểu hiện qui mô của tổng thể, ví dụ số nhân khẩu, số doanh nghiệp, vốn cố định, vốn lưu động của một doanh nghiệp, tổng sản phẩm quốc nội (GDP), diện tích gieo trồng, số sinh viên đại học...

1.2.5.2 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 ảnh 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, năng suất cây trồng, tiền lương... 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.

1.3 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 này, 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.

1.4 Các loại thang đo [1]

Để lượng hoá hiện tượng nghiên cứu, thống kê tiến hành đo lường bằng các loại thang đo phù hợp. Tùy theo tính chất của dữ liệu, ta có thể sử dụng các loại thang đo sau:

[1] Scales of Measurement

1.4.1 Thang đo định danh [1]

Thang đo định danh là loại thang đo dùng cho các tiêu thức thuộc tính. Người ta sử dụng các **mã số** [2]để phân loại các đối tượng. Chúng 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. Ta cũng hay gặp thang đo định danh trong các câu hỏi phỏng vấn như sau:

10) 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

Hoặc:

11. Ly dị 33.Có gia đình 55.Trường hợp khác 88. Độc thân

Trong thang đo định danh người ta cũng có thể sử dụng ký tự:

Đ = Độc thân L = Ly dị C = Có gia đình T = Trường hợp khác

1.4.2 Thang đo thứ bậc [3]

Thang đo thứ bậc thường được sử dụng cho các tiêu thức thuộc tính và nó cũng được áp dụng cho các tiêu thức số lượng. Trong thang đo này giữa các biểu hiện của tiêu thức có quan hệ thứ bậc hơn kém. Sự chênh lệch giữa các biểu hiện không nhất thiết phải bằng nhau. Ví dụ, huân chương có ba hạng: Nhất, nhì, ba. Ta cũng hay gặp loại thang đo này trong các câu hỏi phỏng vấn dạng:

12) Anh/chị/ông/bà hãy xếp hạng các chủ đề sau trên báo Phụ Nữ 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)

-Hôn nhân gia đình (.....)

-Thời trang (.....)

- Nuôi dạy con cái (.....)

Hoặc câu hỏi phỏng vấn dạng:

13) Thu nhập của anh/chị/ông/bà hàng tháng:

1. < 3 tr.đồng 2. Từ 3 – 4 tr.đồng 3. > 4 tr.đồng

---

[1] Nominal scale

[2] Assigning codes

[3] Ordinal scale

1.4.3 Thang đo khoảng [1]

Thang đo khoảng thường dùng cho các tiêu thức số lượng và cũng còn áp dụng cho các tiêu thức thuộc tính. Thang đo khoảng là thang đo thứ bậc có các khoảng cách đều nhau. Ví dụ rõ nhất cho loại thang đo này là nhiệt độ. Ví dụ: 32^0c > 30^0c và 80^0c > 78^0c. Sự chênh lệch giữa 32^0c và 30^0c cũng giống như sự chênh lệch giữa 80^0c và 78^0c, đó là cách nhau 2^0c. 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 cũng gặp loại thang đo này trong các câu hỏi phỏng vấn dạng:

13) Đề nghị quý Thầy/Cô cho biết ý kiến của mình về tầm quan trọng của các mục tiêu đào tạo cho sinh viên đại học sau đây bằng cách khoanh tròn các con số tương ứng trên thang đánh giá chỉ mức độ từ 1 đến 5 (1 = không quan trọng; 5 = rất quan trọng).

	Không quan trọng		Bình thường		Rất quan trọng
(1) Đạo đức	1	2	3	4	5
(2) Khả năng biết phê phán	1	2	3	4	5
(3) Năng lực giải quyết vấn đề	1	2	3	4	5
(4) Tư duy logic	1	2	3	4	5
(5) Khả năng làm việc độc lập	1	2	3	4	5
(6) Năng lực nghiên cứu khoa học	1	2	3	4	5
(7) Tinh thần học tập suốt đời	1	2	3	4	5
(8) Kiến thức chuyên môn sâu	1	2	3	4	5
(9) Kỹ năng làm việc theo nhóm	1	2	3	4	5
(10) Sức khoẻ	1	2	3	4	5

1.4.4 Thang đo tỷ lệ [2]

Thang đo tỷ lệ là loại thang đo dùng cho dữ liệu số lượng. Thang đo tỷ lệ có đầy đủ các đặc tính của thang đo khoảng, ngoài ra nó có một trị số 0 " thật". Đây là loại thang đo cao nhất trong các loại thang đo.

Sự khác nhau giữa thang đo khoảng và thang đo tỷ lệ thường bị lẫn 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 không có ý nghĩa.

Ví dụ bạn có 5 ngàn đồng và anh của bạn có 10 ngàn đồng. Như vậy số tiền

---

[1] Interval scale

[2] Ratio scale

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 tiền tệ có trị số 0 thật và là loại thang đo tỷ lệ. Các loại thang đo tỷ lệ khác như m, kg, tấn, tạ...
Trái lại, với nhiệt độ là thang đo khoảng, ví dụ nhiệt độ hôm nay là 12^0c (53,6 ^{0}F) và hôm qua là 6^0c (42,8^{0}F), ta không thể nói rằng hôm nay ấm áp gấp hai lần hôm qua. Nếu ta đổi từ ^{0}c sang ^{0}F thì tỷ lệ không còn là 2/1 (53,6/42,8). Hơn nữa, nếu nhiệt độ là 0^0c, không có nghĩa là không có nhiệt độ. 0^0c dĩ nhiên lạnh hơn 6^0c. 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 tiêu thức số lượng (ví dụ như thu nhập, chi tiêu ...), và ngược lại có thể áp dụng thang đo định lượng đối với tiêu thức 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 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 thành dạng dữ liệu định tính. Ví dụ như từ dữ liệu tuổi (thang đo tỉ lệ và dữ liệu định lượng) ta có thể biến đổi thành dữ liệu về độ tuổi (thang đo thứ bậc và dữ liệu định tính).

CHƯƠNG 2

THU THẬP DỮ LIỆU THỐNG KÊ

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

2.1 XÁC ĐỊNH DỮ LIỆU CẦN THU THẬP

Chúng ta có thể thu thập rất nhiều dữ liệu liên quan đến hiện tượng nghiên cứu. Vấn đề 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 chúng ta sẽ mất rất nhiều thời gian và tiền bạc cho những dữ liệu không quan trọng hay không liên quan đến vấn đề đang nghiên cứu. Xác định dữ liệu cần thu thập xuất phát từ vấn đề nghiên cứu và mục tiêu nghiên cứu. Nếu 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 sinh viên đi làm thêm có ảnh hưởng đến kết quả học tập hay không, hai nhóm dữ liệu chính cần thu thập là (1) đi làm thêm và (2) kết quả học tập. Về nhóm dữ liệu đi làm thêm, có thể thu thập những dữ liệu liên quan như:

- Có đi làm thêm hay không
- Mức độ thường xuyên công việc làm thêm như thế nào
- Thời gian làm thêm hàng ngày, hàng tuần bao nhiêu giờ
- Tính chất công việc, có liên quan với ngành nghề đang được đào tạo không
- Mục đích của việc đi làm thêm
- Nơi làm thêm có xa chỗ ở và chỗ học không
- Có thích thú với công việc không, có giúp ích cho việc học không? Giúp ích ở khía cạnh nào ...

Một số dữ liệu khác về việc đi làm thêm, nhưng không liên quan lắm đến mục tiêu nghiên cứu ảnh hưởng của việc đi làm thêm đến **kết quả học tập** thì không nhất thiết phải thu thập, ví dụ như:

- Đi làm thêm có phải mặc đồng phục không
- Có được huấn luyện trước khi làm không

- Tính chất công việc làm thêm là làm một mình hay làm với nhiều người
- Người phụ trách công việc là nam hay nữ, có phải là cựu sinh viên của trường không
- Người cùng làm là nam hay nữ
- Những người cùng chỗ làm có cùng quê không
- Việc làm thêm này là do tự tìm, hay do quen biết, giới thiệu
- Có phải trả phí môi giới, giới thiệu việc làm không, trả bao nhiêu ...

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 đáp ứng mục tiêu nghiên cứu đã đề ra.

2.2 DỮ LIỆU ĐỊNH TÍNH VÀ DỮ LIỆU ĐỊNH LƯỢNG

Trước khi thu thập dữ liệu, cần phải phân biệt tính chất của dữ liệu. Có hai loại là dữ liệu định tính và dữ liệu định lượng. Dữ liệu định tính phản ảnh 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 ảnh 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 thứ bậc.

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, 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, thang đo phù hợp cần sử dụng trong biểu mẫu hay bảng câu hỏi dùng để thu thập dữ liệu mong muốn.

Ví dụ như 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 TPHCM), 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.

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

2.3 DỮ LIỆU THỨ CẤP VÀ DỮ LIỆU SƠ CẤP

Có hai loại là dữ liệu thứ cấp và sơ cấp phân theo nguồn. Dữ liệu thứ cấp là dữ liệu thu thập từ những nguồn có sẵn, đó chính là những dữ liệu đã qua tổng hợp, xử lý. 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 việc đi làm thêm 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 việc đi làm thêm 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, rẻ, nhưng đôi khi ít chi tiết và

ít 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 rất nhiều.

2.3.1. Nguồn dữ liệu thứ cấp

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

- Nội bộ: các số liệu báo cáo về tình hình 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 về dân số, lao động, việc làm, 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) 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 mắc bệnh tiểu 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 dược sẽ quan tâm đến con số này), số xe tải và xe buýt quá niên hạn cần thay thế ...
- 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. Ví dụ như số lượng học sinh sinh viên các cấp, các hệ bước vào năm học 2003-2004 là bao nhiêu. Số lượng trung tâm ngoại ngữ có phép và cả không phép đang hoạt động.
- Các tổ chức, hiệp hội, viện nghiên cứu ...: ví dụ như số lượng doanh nghiệp có sản xuất ống nước nhựa
- Các công ty nghiên cứu và cung cấp thông tin

2.3.2. Thu thập dữ liệu sơ cấp

Dữ liệu sơ cấp được thu thập qua các cuộc điều tra khảo sát. Các cuộc điều tra khảo sát để thu thập dữ liệu ban đầu có thể được chia thành nhiều loại. Căn cứ vào tính chất liên tục hay không liên tục của việc ghi chép dữ liệu chia ra điều tra thường xuyên hay không thường xuyên. Căn cứ vào phạm vi khảo sát và thu thập thực tế chia ra điều tra toàn bộ và điều tra không toàn bộ.

Điều tra thường xuyên và điều tra không thường xuyên

Điều tra thường xuyên là tiến hành thu thập, ghi chép dữ liệu ban đầu về hiện tượng nghiên cứu một cách có hệ thống theo sát quá trình biến động của hiện tượng. Ví dụ thu thập, ghi chép tình hình biến động nhân khẩu của một địa phương (sinh, tử, đi, đến); Trong phạm vi một doanh nghiệp việc theo dõi, ghi chép hằng ngày về số công nhân đi làm, số lượng sản phẩm sản xuất ra, số lượng sản phẩm tiêu thụ... là điều tra thường xuyên. Dữ liệu của điều tra thường xuyên là cơ sở chủ yếu để lập các báo cáo thống kê theo định kỳ.

Điều tra không thường xuyên là tiến hành thu thập, ghi chép dữ liệu ban đầu một cách không liên tục, mà chỉ tiến hành khi có nhu cầu cần nghiên cứu hiện tượng. Dữ liệu của điều tra không thường xuyên phản ánh trạng thái hiện tượng tại thời điểm nhất định. Ví dụ tổng điều tra dân số, tổng điều tra tra đất đai nông nghiệp, điều tra đàn gia súc, gia cầm, điều tra năng suất cây trồng, những cuộc điều tra nghiên cứu thị trường...là những cuộc điều tra không thường xuyên. Các cuộc điều tra không thường xuyên có thể được tiến hành theo định kỳ nhất định (3 tháng, 6 tháng, 1 năm, 2 năm, 5 năm ...) hay không theo định kỳ.

Điều tra toàn bộ và điều tra không toàn bộ

Điều tra toàn bộ là tiến hành thu thập, ghi chép dữ liệu trên tất cả các đơn vị của tổng thể hiện tượng nghiên cứu. Ví dụ tổng điều tra dân số, tổng điều tra tồn kho vật tư, hàng hoá; tổng điều tra vốn sản xuất, kinh doanh của các doanh nghiệp, điều tra tất cả các chợ trên địa bàn quận, thành phố, điều tra tất cả các cây xăng, tiệm rửa xe,..là điều tra toàn bộ.

Điều tra toàn bộ cung cấp dữ liệu đầy đủ nhất cho nghiên cứu thống kê, nhất là trong nghiên cứu kinh tế và thị trường. Nó giúp ta tính được các chỉ tiêu quy mô, khối lượng một cách khá chính xác. Cho phép nghiên cứu cơ cấu, tình hình biến động, đánh giá thực trạng hiện tượng, dự đoán xu hướng biến động hiện tượng.... Nhưng điều tra toàn bộ đòi hỏi chi phí rất lớn về nhân lực, thời gian, chi phí, vì vậy không thể áp dụng cho tất cả các trường hợp nghiên cứu.

Điều tra không toàn bộ là tiến hành thu thập, ghi chép dữ liệu trên một số đơn vị được chọn ra từ toàn bộ các đơn vị thuộc tổng thể hiện tượng nghiên

cứu. Tùy theo cách chọn số đơn vị để tiến hành điều tra thực tế, điều tra không toàn bộ chia thành 3 loại cụ thể sau: điều tra chuyên đề, điều tra chọn mẫu và điều tra trọng điểm.

Điều tra chuyên đề là tiến hành điều tra trên một số rất ít các đơn vị của tổng thể, nhưng lại đi sâu nghiên cứu nhiều khía cạnh của đơn vị đó. Mục đích là để khám phá, tìm hiểu các yếu tố ảnh hưởng đến hiện tượng nghiên cứu. Dữ liệu của điều tra chuyên đề phục vụ cho nghiên cứu định tính, không dùng để suy rộng, không dùng để tìm hiểu tình hình cơ bản của hiện tượng, mà chỉ rút ra kết luận về bản thân các đơn vị được điều tra. Kết quả điều tra chuyên đề có thể được sử dụng làm cơ sở để thiết kế cho một cuộc điều tra trên quy mô lớn hơn, mang tính chất nghiên cứu định lượng.

Ví dụ điều tra điển hình một số ít sinh viên có đi làm thêm, đạt kết quả học tập tốt và thành tích nghiên cứu khoa học xuất sắc, vài sinh viên có đi làm thêm nhưng kết quả học tập kém, bị tạm dừng học tập. Các kết quả điều tra chuyên đề này giúp ta khám phá những yếu tố quan trọng có ảnh hưởng đến kết quả học tập của sinh viên, trên cơ sở đó xác định các dữ liệu cần thu thập trong nghiên cứu định lượng (điều tra chọn mẫu) tiếp theo để kết luận về ảnh hưởng của việc đi làm thêm đối với kết quả học tập của sinh viên. Hoặc kết quả điều tra chuyên đề giúp người nghiên cứu giải thích được nguyên nhân của các khám phá phát hiện qua cuộc điều tra chọn mẫu hay toàn bộ.

Điều tra chọn mẫu được thực hiện bằng cách chọn ra một số phần tử hay đơn vị thuộc tổng thể đơn vị nghiên cứu để thu thập dữ liệu thực tế. Điều tra chọn mẫu được dùng nhiều nhất trong nghiên cứu vì tiết kiệm thời gian, chi phí và dữ liệu đáng tin cậy. Dữ liệu của điều tra chọn mẫu được dùng để suy rộng thành các đặc trưng chung của toàn bộ tổng thể hiện tượng nghiên cứu.

Điều tra trọng điểm là tiến hành thu thập dữ liệu trên bộ phận chủ yếu nhất, tập trung nhất trong toàn bộ tổng thể hiện tượng nghiên cứu. Kết quả thu được từ điều tra trọng điểm giúp ta nhận biết nhanh tình hình cơ bản của hiện tượng nghiên cứu, chứ không dùng để suy rộng thành các đặc trưng chung tổng thể. Chẳng hạn, khi cần nắm nhanh tình hình cơ bản về sản xuất cao su, cà phê của nước ta, ta có thể chỉ tiến hành điều tra về sản xuất cao su, cà phê ở miền Đông Nam Bộ và Tây nguyên chứ không cần tiến hành điều tra trong cả nước. Tại TP Hồ Chí Minh, cần nhận biết nhanh tình hình tiêu thụ hàng điện lạnh, chỉ cần khảo sát và thu thập dữ liệu tại vài địa điểm

trung tâm mua bán hàng điện lạnh chính yếu.

2.4 CÁC PHƯƠNG PHÁP THU THẬP DỮ LIỆU BAN ĐẦU

2.4.1. Thu thập trực tiếp

Quan sát

Quan sát là thu thập dữ liệu bằng cách quan sát các hành động, thái độ của đối tượng khảo sát trong những tình huống nhất định. Ví dụ quan sát số lượng và thái độ của các khách đến thăm gian hàng của công ty tại một hội chợ hay một cuộc triển lãm; quan sát thứ tự hành động đi đến các kệ hàng của từng khách hàng đi siêu thị.

Phỏng vấn trực tiếp

Người phỏng vấn trực tiếp hỏi đối tượng được điều tra và tự ghi chép dữ liệu vào bản câu hỏi hay phiếu điều tra. Phương pháp phỏng vấn trực tiếp phù hợp với những cuộc điều tra phức tạp cần thu thập nhiều dữ liệu. Ưu điểm là thời gian phỏng vấn có thể ngắn hay dài tuỳ thuộc vào số lượng dữ liệu cần thu thập; và nhân viên trực tiếp phỏng vấn có điều kiện để có thể giải thích một cách đầy đủ, cặn kẽ, đặt những câu hỏi chi tiết để khai thác thông tin và kiểm tra dữ liệu trước khi ghi chép vào phiếu điều tra.

Phương pháp phỏng vấn trực tiếp có ưu điểm lớn là dữ liệu được thu thập đầy đủ theo nội dung điều tra và có độ chính xác khá cao, cho nên được áp dụng phổ biến trong điều tra thống kê. Tuy nhiên, phương pháp này đòi hỏi chi phí lớn, nhất là chi phí về nhân lực và thời gian.

2.4.2. Thu thập gián tiếp

Nhân viên điều tra thu thập tài liệu qua trao đổi bằng điện thoại, hoặc thư gửi qua bưu điện với đơn vị điều tra hoặc qua chứng từ, sổ sách có sẵn ở đơn vị điều tra.

Ví dụ trong điều tra thu chi của hộ gia đình, nhân viên điều tra gặp đại diện hộ gia đình trao phiếu điều tra, giải thích ý nghĩa điều tra, cách trả lời.... Đại diện hộ gia đình xác định các dữ liệu cần thiết và tự ghi vào phiếu điều tra, rồi gởi cho nhân viên điều tra. Trong điều tra về biến động dân số của một địa phương, nhân viên điều tra có thể thu thập tài liệu qua sổ sách theo dõi của cơ quan địa phương về số sinh, tử, chuyển đi, chuyển đến. Trong điều tra về tình hình việc làm của sinh viên ra trường, nhân viên phụ trách điều

tra có thể gửi bản câu hỏi qua đường bưu điện đến địa chỉ của sinh viên đã tốt nghiệp để thu thập dữ liệu về tính chất công việc, khu vực kinh tế đang làm việc, lĩnh vực hoạt động, thu nhập và đãi ngộ, huấn luyện và đào tạo .... Cựu sinh viên, sau khi trả lời xong sẽ gửi lại quan đường bưu điện đến địa chỉ tiếp nhận.

Thu thập gián tiếp ít tốn kém hơn thu thập trực tiếp, nhưng chất lượng dữ liệu không cao, nên thường chỉ được áp dụng trong những trường hợp khó khăn hoặc không có điều kiện thu thập trực tiếp.

2.5 XÂY DỰNG KẾ HOẠCH ĐIỀU TRA THỐNG KÊ

Để thu thập dữ liệu khách quan đáp ứng được yêu cầu nghiên cứu kịp thời và đầy đủ thì điều tra thống kê cần phải được tổ chức một cách khoa học, thống nhất và chu đáo. Vấn đề cơ bản nhất được đặt ra trước khi tiến hành điều tra thực tế là phải xây dựng được kế hoạch điều tra.

Kế hoạch điều tra là một tài liệu dưới dạng văn bản, trong đó đề cập những vấn đề cần giải quyết hoặc cần được hiểu thống nhất, trình tự và phương pháp tiến hành cuộc điều tra, những vấn đề thuộc về chuẩn bị và tổ chức toàn bộ cuộc điều tra.

Đối với mỗi cuộc điều tra thống kê cần phải xây dựng kế hoạch điều tra phù hợp. Nội dung cơ bản của kế hoạch điều tra thường bao gồm một số vấn đề chủ yếu sau đây.

2.5.1. Mô tả mục đích điều tra

Mục đích điều tra là nội dung quan trọng đầu tiên của kế hoạch điều tra, xác định rõ điều tra để tìm hiểu những khía cạnh nào của hiện tượng, phục vụ yêu cầu nghiên cứu hoặc yêu cầu quản lý nào.

Bất kỳ một hiện tượng nào cũng có thể được quan sát, nghiên cứu từ nhiều góc độ khác nhau. Nhưng với mỗi cuộc điều tra ta không thể và cũng không cần thiết phải điều tra tất cả các khía cạnh của hiện tượng, mà chỉ cần khảo sát điều tra những khía cạnh phục vụ yêu cầu nghiên cứu cụ thể.

Việc xác định mục đích điều tra có tác dụng định hướng cho toàn bộ quá trình điều tra. Nó liên quan đến xác định đối tượng, đơn vị và nội dung điều tra. Muốn xác định mục đích điều tra phải căn cứ vào mục đích của toàn bộ

quá trình nghiên cứu.

2.5.2. Xác định đối tượng điều tra và đơn vị điều tra

Đối tượng điều tra

Đối tượng điều tra là tổng thể các đơn vị thuộc hiện tượng nghiên cứu có thể cung cấp những dữ liệu cần thiết khi tiến hành điều tra.

Xác định đối tượng điều tra có nghĩa là quy định rõ phạm vi của hiện tượng nghiên cứu, vạch rõ ranh giới của hiện tượng nghiên cứu với hiện tượng khác, giúp ta xác định đúng đắn số đơn vị cần điều tra thực tế. Xác định chính xác đối tượng điều tra giúp ta tránh được nhầm lẫn khi thu thập dữ liệu, làm cho dữ liệu thu thập và tổng hợp phản ánh đúng hiện tượng cần nghiên cứu.

Khi xác định đối tượng điều tra phải căn cứ mục đích điều tra, đồng thời phải định nghĩa những tiêu chuẩn phân biệt rõ ràng, vì nhiều khi biểu hiện bên ngoài của hiện tượng giống nhau, nhưng thực chất lại khác nhau. Ví dụ trong cuộc tổng điều tra dân số, đối tượng điều tra được xác định là "Nhân khẩu thường trú" trên lãnh thổ Việt Nam. Để phân biệt "nhân khẩu thường trú" với "nhân khẩu tạm trú" và với "nhân khẩu có mặt", tránh đăng ký trùng lắp hay bỏ sót. Kế hoạch điều tra đã nêu ra những tiêu chuẩn cụ thể để xác định thế nào là nhân khẩu thường trú. Một ví dụ khác là khi tiến hành nghiên cứu về các điểm bán dầu nhớt thì ta phải xác định rõ ràng là điều tra loại dầu nhớt nào (dành cho xe gắn máy, xe ô-tô, hay động cơ nổ, máy tàu thủy, máy phát điện ...), tiêu chuẩn phân biệt các điểm bán (cửa hàng xăng dầu, tiệm bán phụ tùng, điểm rửa xe, ...)

Đơn vị điều tra

Đơn vị điều tra là đơn vị thuộc đối tượng điều tra và được xác định sẽ điều tra thực tế. Trong điều tra toàn bộ thì số đơn vị điều tra chính là số đơn vị thuộc đối tượng điều tra. Trong điều tra không toàn bộ thì số đơn vị điều tra là những đơn vị được chọn ra trong số đơn vị của đối tượng điều tra.

Xác định đơn vị điều tra chính là xác định nơi sẽ cung cấp những dữ liệu cần thiết cho quá trình nghiên cứu. Đồng thời đơn vị điều tra là căn cứ để tiến hành tổng hợp, phân tích và dự báo thống kê.

Khi xác định đơn vị điều tra phải căn cứ vào mục đích điều tra và đối tượng điều tra. Đơn vị điều tra có thể là từng doanh nghiệp, từng cửa hàng từng

trường học..., nhưng cũng có thể là từng công nhân, từng học sinh.... Trong một cuộc điều tra cũng có thể dùng nhiều loại đơn vị điều tra để đáp ứng những yêu cầu nghiên cứu khác nhau. Ví dụ trong tổng điều tra dân số thường dùng 2 loại đơn vị điều tra là từng người dân và từng hộ gia đình.

2.5.3. Nội dung điều tra

Nội dung điều tra là mục lục các tiêu thức hay đặc trưng cần thu thập dữ liệu trên các đơn vị điều tra.

Từ đơn vị điều tra ta có thể thu thập được dữ liệu theo nhiều tiêu thức khác nhau. Nhưng trong mỗi cuộc điều tra ta không cần thu thập dữ liệu theo tất cả các tiêu thức, mà chỉ thu thập theo một số tiêu thức. Những tiêu thức này đủ đáp ứng cho mục đích điều tra và mục đích nghiên cứu. Vì vậy trong kế hoạch điều tra phải xác định và thống nhất mục lục các tiêu thức cần thu thập dữ liệu, xác định và thống nhất nội dung điều tra. Khi tiến hành điều tra cần thu thập dữ liệu theo đúng nội dung điều tra từ tất cả các đơn vị điều tra.

Khi xác định nội dung điều tra phải căn cứ vào mục đích nghiên cứu chung, mục đích điều tra cụ thể, đồng thời phải tính đến khả năng về nhân lực, thời gian, chi phí.... Cho nên nội dung điều tra chỉ nên bao gồm những tiêu thức hay đặc trưng quan trọng nhất có liên quan trực tiếp đến mục đích điều tra và có quan hệ chặt chẽ hoặc có thể bổ sung cho nhau, tạo điều kiện thuận lợi cho việc kiểm tra tính chất chính xác của dữ liệu.

Mỗi tiêu thức trong nội dung điều tra phải được diễn đạt thành câu hỏi ngắn gọn, cụ thể, rõ ràng để cả người điều tra và người được điều tra đều hiểu thống nhất.

2.5.4. Xác định thời điểm, thời kỳ điều tra.

Tùy theo tính chất, đặc điểm của hiện tượng nghiên cứu cần phải xác định đúng đắn và chặt chẽ thời gian thu thập dữ liệu về hiện tượng.

Thời điểm điều tra

Thời điểm điều tra là mốc thời gian được xác định để thống nhất đăng ký dữ liệu của toàn bộ các đơn vị điều tra. Xác định thời điểm điều tra là xác định cụ thể ngày, giờ để thống nhất đăng ký dữ liệu, tức là xác định ý muốn nghiên cứu trạng thái hiện tượng ở chính thời điểm đó.

Khi xác định thời điểm điều tra phải căn cứ vào tính chất mỗi loại hiện

tượng, đồng thời phải đảm bảo thuận tiện cho việc đăng ký dữ liệu và tính các chỉ tiêu từ dữ liệu điều tra. Ví dụ tổng điều tra dân số Việt Nam, thời điểm điều tra được xác định là 0 giờ ngày 1 tháng 4 vì ở thời điểm này dân số ít biến động nhất để vừa dễ dàng đăng ký dữ liệu chính xác, vừa tránh đăng ký trùng hoặc bỏ sót đơn vị điều tra khi thu thập dữ liệu. Điều tra thị trường áo mưa không thể tiến hành trong mùa khô vì lúc đó người bán và cả người mua sử dụng không quan tâm để tham gia cung cấp thông tin tốt được.

Thời kỳ điều tra

Là khoảng thời gian được xác định để thống nhất đăng ký dữ liệu của các đơn vị điều tra trong suốt khoảng thời gian đó (ngày, tuần, 10 ngày, 1 tháng, 3 tháng hay 1 năm...). Ví dụ điều tra lượng nguyên liệu tiêu thụ trong sản xuất, số lượng sản phẩm làm ra của 1 kỳ nào đó; số người sinh, chết, chuyển đi, chuyển đến trong 1 năm của 1 địa phương; số lần đi siêu thị trong vòng 1 tháng qua, số lượng tập vở học sinh sử dụng trong năm học qua ... Thời kỳ điều tra có thể dài hay ngắn phụ thuộc vào mục đích nghiên cứu.

Thời hạn điều tra

Là thời gian dành cho việc đăng ký ghi chép tất cả các dữ liệu điều tra, được tính từ khi bắt đầu cho đến lúc kết thúc toàn bộ việc thu thập dữ liệu. Ví dụ tổng điều tra dân số thời hạn điều tra là trong vòng 10 ngày đầu tháng 4.

Thời hạn điều tra dài hay ngắn tùy thuộc quy mô, tính chất phức tạp của hiện tượng, vào nội dung nghiên cứu, lực lượng tham gia điều tra. Nhưng thời hạn điều tra không nên quá dài.

2.5.5. Biểu điều tra và bản giải thích cách ghi biểu.

Biểu điều tra

Biểu điều tra (còn gọi là phiếu điều tra, bản câu hỏi) là loại bản in sẵn theo mẫu quy định trong kế hoạch điều tra, được sử dụng thống nhất để ghi dữ liệu của đơn vị điều tra.

Biểu điều tra phải chứa đựng toàn bộ nội dung cần điều tra, đồng thời phải thuận tiện cho việc ghi chép và kiểm tra dữ liệu, thuận tiện cho tổng hợp.

Trên biểu điều tra, những thang đo định tính sử dụng trong nội dung điều tra cần được mã hoá sẵn tạo điều kiện thuận lợi cho việc nhập liệu vào máy tính. Thường người ta dùng chữ số để mã hoá.

Bản giải thích cách ghi biểu

Kèm theo biểu điều tra là bản giải thích và hướng dẫn cụ thể cách xác định và ghi dữ liệu vào biểu điều tra. Nó giúp cho nhân viên điều tra và đơn vị điều tra nhận thức thống nhất các câu hỏi trong biểu điều tra. Nội dung, ý nghĩa các câu hỏi phải được giải thích một cách khoa học và chính xác. Những câu hỏi phức tạp có nhiều khả năng trả lời cần có ví dụ cụ thể.

Ngoài những nội dung chủ yếu trên, trong kế hoạch điều tra còn cần đề cập và giải thích một số vấn đề thuộc về phương pháp, tổ chức và tiến hành điều tra như:

- Cách thức chọn mẫu
- Phương pháp thu thập dữ liệu và ghi chép ban đầu
- Các bước và tiến độ tiến hành điều tra
- Tổ chức và quy định nhiệm vụ của bộ phận tham gia điều tra
- Bố trí lực lượng điều tra và phân chia khu vực điều tra.
- Tổ chức cuộc họp chuẩn bị và huấn luyện nhân viên điều tra.
- Tiến hành điều tra thử để rút kinh nghiệm.
- Tổ chức tuyên truyền mục đích, ý nghĩa của cuộc điều tra...

2.6 SAI SỐ TRONG ĐIỀU TRA THỐNG KÊ.

Các cuộc điều tra thống kê, dù được tổ chức dưới hình thức nào và thu thập dữ liệu bằng phương pháp nào, đều phải đảm bảo yêu cầu chính xác với mức độ nhất định. Tuy nhiên, trong thực tế điều tra dữ liệu thu thập được thường có sai số.

Sai số trong điều tra thống kê là chênh lệch giữa trị số thu thập được trong điều tra với trị số thực tế của đơn vị điều tra.

Sai số điều tra làm giảm chất lượng của kết quả điều tra và ảnh hưởng đến chất lượng của cả quá trình nghiên cứu thống kê. Nhưng trong thực tế khó có thể xác định được sai số và khó có thể loại bỏ hoàn toàn được sai số trong điều tra thống kê. Vấn đề đặt ra là phải nắm được các nguyên nhân làm phát sinh sai số trong điều tra để chủ động tìm biện pháp khắc phục làm hạn chế sai số. Có hai loại sai số trong điều tra thống kê: sai số do đăng ký và sai số do tính chất đại biểu.

2.6.1 Sai số do đăng ký

Sai số do đăng ký là loại sai số phát sinh do xác định và ghi chép dữ liệu

không chính xác. Có nhiều nguyên nhân dẫn đến loại sai số này như:

- Vạch kế hoạch điều tra sai hoặc không khoa học, không sát với thực tế của hiện tượng
- Do trình độ của nhân viên điều tra, không hiểu chính xác nội dung các câu hỏi, không biết cách khai thác dữ liệu.
- Do đơn vị điều tra không hiểu câu hỏi nên trả lời sai.
- Do ý thức, tinh thần trách nhiệm của nhân viên điều tra hoặc của đơn vị điều tra thấp dẫn đến xác định, cung cấp hoặc ghi chép sai (hồi tưởng, cân, đo, đếm ... sai và ghi sai).
- Do dụng cụ đo lường không chính xác.
- Do công tác tuyên truyền, vận động không tốt dẫn đến đơn vị điều tra không hiểu hết mục đích điều tra nên cung cấp dữ liệu không đúng.
- Do thiếu tính trung thực, khách quan nên cố tình cung cấp hoặc ghi chép sai.
- Do lỗi in ấn biểu mẫu, phiếu và bản giải thích sai...

Nắm được các nguyên nhân cụ thể dẫn đến loại sai số này cũng có nghĩa là tìm được hướng để khắc phục.

2.6.2 Sai số do tính chất đại biểu

Sai số do tính chất đại biểu là loại sai số xảy ra trong điều tra không toàn bộ, nhất là trong điều tra chọn mẫu.

Nguyên nhân của loại sai số này là do việc lựa chọn đơn vị điều tra thực tế không có tính đại diện cao. Trong điều tra chọn mẫu, ta chỉ thu thập dữ liệu từ một số ít đơn vị thuộc đối tượng điều tra rồi căn cứ kết quả điều tra thực tế mà suy rộng thành các đặc trưng của tổng thể. Như vậy tổng thể các đơn vị được chọn nếu khác về kết cấu theo tiêu thức điều tra với tổng thể chung sẽ phát sinh sai số do tính chất đại biểu.

2.6.3 Một số biện pháp chủ yếu nhằm hạn chế sai số trong điều tra thống kê.

- Làm tốt công tác chuẩn bị điều tra: chọn, huấn luyện, kiểm tra nhân viên; in ấn chính xác phiếu và các tài liệu hướng dẫn, phổ biến mục tiêu, ý nghĩa của cuộc khảo sát ...
- Tiến hành kiểm tra một cách có hệ thống toàn bộ cuộc điều tra: chọn ra 20-30% số phiếu để kiểm tra thật sự đối tượng có được khảo sát và phỏng vấn hay không, kiểm tra về mặt logic của dữ liệu bằng cách đọc soát nghiệm thu từng phiếu, kiểm tra việc xác định và tính toán dữ liệu, kiểm

tra tính chất đại biểu và chỉ tiêu mẫu khảo sát (trong điều tra không toàn bộ).

Làm tốt công tác tuyên truyền đối với các đơn vị được điều tra và nâng cao tinh thần trách nhiệm đối với nhân viên điều tra (điều kiện làm việc, thời gian, thù lao, chế độ thưởng phạt ...)

CHƯƠNG 3

TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU

Sau khi tiến hành điều tra thống kê, ta sẽ thu được rất nhiều dữ liệu ban đầu (dữ liệu sơ cấp) trên mỗi đơn vị điều tra. Những dữ liệu này là những dữ liệu thô phản ánh các đặc trưng cá biệt của từng đơn vị, có tính chất rời rạc nên rất khó quan sát để rút ra những nhận xét, kết luận về hiện tượng nghiên cứu, và cũng chưa thể sử dụng ngay vào phân tích và dự đoán thống kê. Vì vậy, phải tiến hành tóm tắt những tài liệu thu được trong điều tra và trình bày chúng dưới những hình thức phù hợp.

Nhiệm vụ cơ bản của tóm tắt dữ liệu thống kê là từ các thông tin riêng biệt trên từng đơn vị, thực hiện sắp xếp, phân loại để giúp cho người nghiên cứu thấy được các đặc trưng chung của mẫu hay toàn bộ tổng thể nghiên cứu. Khi tóm tắt dữ liệu thống kê, nếu số đơn vị điều tra ít, tức là lượng dữ liệu ít, ta có thể tiến hành bằng phương pháp đơn giản là sắp xếp dữ liệu theo một trật tự nào đó: trật tự tăng dần hoặc giảm dần (đối với dữ liệu định lượng); hoặc theo trật tự qui định nào đó (đối với dữ liệu định tính).

Trong trường hợp số lượng đơn vị điều tra lớn, lượng dữ liệu lớn, thì không thể tiến hành theo phương pháp sắp xếp đơn giản như trên vì sẽ gặp nhiều khó khăn mà kết quả sắp xếp cũng không giúp thấy được những đặc trưng cơ bản. Trong trường hợp này cần phải tiến hành phân tổ, tức là sắp xếp các đơn vị vào các tổ nhóm theo một hay một vài tiêu thức hay đặc trưng và tính toán các đại lượng thống kê. Các kết quả sắp xếp này thường được trình bày dưới dạng bảng hay biểu đồ để dễ quan sát, cảm nhận và nhận thức. Chương này sẽ bắt đầu bằng phần lý thuyết căn bản về phương pháp phân tổ. Các phần tiếp theo lần lượt trình bày vận dụng phương pháp phân tổ trong từng trường hợp cụ thể với các ví dụ thực tế. Các công cụ cơ bản được trình bày trong phần này là: bảng tần số, các đại lượng thống kê mô tả, bảng kết hợp, biểu đồ và đồ thị.

3.1 LÝ THUYẾT PHÂN TỔ

3.1.1 Khái niệm: Phân tổ thống kê là căn cứ vào một hay một số tiêu thức (đặc trưng) 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 chia tổng thể hay mẫu nghiên cứu thành các tổ nhóm có tính chất khác nhau.

3.1.2 Các bước tiến hành phân tổ: Để tiến hành phân tổ một tổng thể công việc đầu tiên cần làm là lựa chọn tiêu thức phân tổ từ nhiều tiêu thức có thể sử dụng. Sau khi đã lựa chọn được tiêu thức phân tổ rồi, công việc tiếp theo là nên sắp xếp các đơn vị tổng thể hay mẫu quan sát vào bao nhiêu tổ, nhóm – tức là xác định số tổ cần thiết.

3.1.2.1 Lựa chọn tiêu thức phân tổ: Tiêu thức phân tổ là tiêu thức được chọn làm căn cứ để tiến hành phân tổ. Mỗi đơn vị tổng thể có nhiều tiêu thức khác nhau. Có tiêu thức khi chọn làm căn cứ phân tổ sẽ giúp ta hiểu được tính chất của hiện tượng, nhưng cũng có tiêu thức nếu chọn làm căn cứ phân tổ chẳng những không đáp ứng mục đích nghiên cứu mà còn làm cho ta hiểu sai lệch hiện tượng nghiên cứu qua các kết quả xử lý và tổng hợp. Vì vậy khi tiến hành phân tổ, trước tiên ta phải lựa chọn đúng đắn tiêu thức phân tổ phù hợp.

Để lựa chọn tiêu thức phân tổ, trước hết phải dựa vào phân tích lý thuyết để chọn ra tiêu thức phù hợp đáp ứng được mục đích nghiên cứu. Ngoài ra phải căn cứ vào điều kiện cụ thể của hiện tượng để chọn tiêu thức phân tổ thích hợp.

3.1.2.2 Xác định số tổ: Số tổ được xác định tùy thuộc vào tiêu thức phân tổ là tiêu thức thuộc tính (dữ liệu định tính) hay tiêu thức số lượng (dữ liệu định lượng).

a. Phân tổ theo tiêu thức thuộc tính hay dữ liệu định tính: có hai trường hợp

* Tiêu thức thuộc tính có một vài biểu hiện (loại hình): ví dụ như phân tổ nhân khẩu theo giới tính, phân tổ các doanh ngiệp theo thành phần kinh tế ... Trong trường này việc chia hiện tượng ra làm bao nhiêu tổ khá đơn giản, thông thường cứ mỗi biểu hiện của tiêu thức thuộc tính có thể chia thành một tổ.

* Tiêu thức thuộc tính có nhiều biểu hiện, như phân tổ nhân khẩu theo nghề nghiệp, phân tổ các sản phẩm công nghiệp theo giá trị sử dụng ... Trường hợp này ta ghép nhiều nhóm nhỏ lại với nhau theo nguyên tắc các nhóm ghép lại với nhau phải giống nhau hoặc gần giống nhau.

Ví dụ: Khi phân tổ ngành công nghiệp, các sản phẩm có tính chất giống nhau hoặc gần giống nhau được xếp trong cùng một tổ, như:

- Công nghiệp chế biến, bảo quản thịt và sản phẩm từ thịt.
- Công nghiệp sản xuất bánh, mứt, kẹo, ca cao, sô cô la.
- Công nghiệp gốm sứ và sản phẩm bằng gốm sứ.
- Công nghiệp chế biến gỗ, lâm sản và các sản phẩm từ gỗ, lâm sản.

b. Phân tổ theo tiêu thức số lượng hay dữ liệu định lượng: Ta chia ra hai trường hợp

* Tiêu thức số lượng có ít trị số. Ví dụ phân tổ các hộ gia đình theo số nhân khẩu, phân tổ công nhân trong xí nghiệp theo bậc thợ ... Trong trường hợp này thường cứ mỗi trị số ứng với một tổ.

* Tiêu thức số lượng có nhiều trị số. Ví dụ phân tổ dân số theo độ tuổi, phân tổ công nhân trong một xí nghiệp theo năng suất lao động ... Ta không thể thực hiện giống như trường hợp trên, vì nếu tương ứng với mỗi trị số hình thành một tổ thì số tổ sẽ quá nhiều và người nghiên cứu khó quan sát và thấy rõ sự khác nhau giữa các tổ. Trong trường hợp này ta phân tổ có khoảng cách tổ, và mỗi tổ có hai giới hạn là giới hạn dưới và giới hạn trên. Giới hạn dưới là trị số nhỏ nhất của tổ. Giới hạn trên là trị số lớn nhất của tổ.

Trị số chênh lệch giữa giới hạn trên và giới hạn dưới của mỗi tổ gọi là khoảng cách tổ. Trong thực tế tùy theo đặc điểm của hiện tượng nghiên cứu để quyết định xem phân tổ có khoảng cách tổ đều hay không đều. Đối với các hiện tượng nghiên cứu có lượng biến trên các đơn vị thay đổi một cách đều đặn, có thể phân tổ với khoảng cách tổ đều nhau.

Khi phân tổ có khoảng cách tổ đều, trị số khoảng cách tổ được xác định:

* Đối với trị số quan sát liên tục: $h = \dfrac{x_{max} - x_{min}}{k}$

Trong đó:

h : trị số khoảng cách tổ.

k : số tổ.

x_{max} : trị số quan sát lớn nhất.

x_{min} : trị số quan sát nhỏ nhất.

Trong thực tế số tổ k được xác định chủ yếu dựa vào kinh nghiệm và tùy theo đặc điểm của hiện tượng nghiên cứu. Ngoài ra ta có thể tham khảo cách xác định k bằng công thức: $k = (2\times n)^{1/3}$. Trong đó n là số đơn vị được quan sát.

* Đối với trị số quan sát rời rạc: $h = \dfrac{(x_{max} - x_{min}) - (k-1)}{k}$

Khi tính h người ta thường làm tròn số.

Ví dụ: Có tài liệu về năng suất lúa (tạ/ha) của 50 hộ nông dân cho trong

bảng sau:

35	41	32	44	33	41	38	44	43	42
30	35	35	43	48	46	48	49	39	49
46	42	41	51	36	42	44	34	46	34
36	47	42	41	37	47	49	38	41	39
40	44	48	42	46	52	43	41	52	43

Năng suất lúa của các hộ biến thiên tương đối đều đặn, ta có thể phân tổ có khoảng cách tổ đều nhau. Các bước được thực hiện như sau:

a. Xác định số tổ: $k = (2\times n)^{1/3}$ với n = 50

Ta có : $k = (2\times 50)^{1/3} = 4,6408 \approx 5$ tổ.

b. Xác định khoảng cách tổ: $h = \dfrac{x_{max} - x_{min}}{k}$

Ta có: $h = \dfrac{52 - 30}{5} = 4,4$

h = 4,4 sẽ được lấy tròn thành 5.

Vì nếu chọn h = 4 ta có các tổ như sau:

30 34
34 38
38 42
42 46
46 50

Ta nhận thấy các trị số lớn hơn 50 sẽ không được xếp vào tổ nào. Do vậy để có thể xếp được tất cả các trị số vào trong các tổ, ta chọn khoảng cách tổ bằng 5, khi đó ta có các tổ như sau:

30 - 35
35 - 40
40 - 45
45 - 50
50 - 55

Tần số của mỗi tổ được xác định bằng cách đếm số quan sát rơi vào giới hạn của tổ đó. Cuối cùng ta có bảng phân tổ sau:

Năng suất lúa (tạ / ha)	Số hộ gia đình
30 – 35	5
35 – 40	10
40 – 45	20
45 – 50	12
50 – 55	3
Tổng	50

Trường hợp ví dụ trên đây có số quan sát ít (n = 50), ta có thể đếm trị số quan sát bằng cách sử dụng ký hiệu gạch ⧅ hoặc ~~////~~ , mỗi một gạch tượng trưng cho 1 quan sát.

Nếu như số quan sát lớn, chẳng hạn n = 1000, chúng ta thường không thể đếm bằng tay. Trong trường hợp này người ta sử dụng các chương trình máy tính phổ biến như Excel hay chương trình thống kê chuyên dụng như SPSS for Windows để tiến hành phân tổ và xác định tần số của mỗi tổ sau khi ta đã nhập đầy đủ các số liệu vào máy.

3.1.2.3 Phân tổ mở: là phân tổ mà tổ đầu tiên không có giới hạn dưới, tổ cuối cùng không có giới hạn trên, các tổ còn lại có thể có khoảng cách tổ đều hoặc không đều. Mục đích của phân tổ mở là để tổ đầu tiên và tổ cuối cùng chứa được các đơn vị có trị số lượng biến đột biến, nghĩa là lượng biến nhỏ bất thường hoặc lớn bất thường và tránh việc hình thành quá nhiều tổ.

Trở lại ví dụ trên, giả sử trong 50 hộ gia đình có một hộ có mức năng suất là 10 tạ/ha và một hộ có mức năng suất 60 tạ/ha. Rõ ràng đây là hai trị số bất thường. Để tiến hành phân tổ ta không lấy 2 trị số này làm trị số lớn nhất và nhỏ nhất để tính trị số khoảng cách tổ h. Khi đó ta có thể giải quyết bằng cách phân tổ mở, ta có bảng phân tổ mở như sau:

Năng suất lúa (tạ / ha)	Số hộ gia đình
< 35	5
35 – 40	10
40 – 45	20
45 – 50	12
≥ 50	3
Tổng	50

Khi tính toán đối với tài liệu phân tổ mở người ta qui ước lấy khoảng cách tổ của tổ mở bằng với khoảng cách tổ của tổ nào đứng gần nó nhất.

Trường hợp phân tổ theo tiêu thức số lượng với trị số liên tục thì giới hạn trên và giới hạn dưới của hai tổ kế tiếp phải trùng nhau. Và người ta cũng qui ước là khi có một lượng biến đúng bằng giới hạn trên của một tổ, thì đơn vị đó được xếp vào tổ kế tiếp.

3.2 VẬN DỤNG PHƯƠNG PHÁP PHÂN TỔ TRONG TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU

3.2.1 TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU ĐỊNH TÍNH

3.2.1.1. Bảng tần số

Đối với dữ liệu định tính thu thập từ các tiêu thức thuộc tính như giới tính, ngành học, nghề nghiệp ... hay thu thập từ các tiêu thức số lượng nhưng qua các thang đo định tính như độ tuổi (dưới 18, 18-25, 26-35, 36-45, 46-60), mức thu nhập (dưới 1 triệu đồng, từ 1 đến dưới 2 triệu đồng, 2 đến dưới 4 triệu đồng, trên 4 triệu đồng), người ta thường đếm xem có bao nhiêu đơn vị quan sát có cùng một biểu hiện, và so với tổng số quan sát thì số đơn vị có cùng biểu hiện này chiếm bao nhiêu phần trăm. Kết quả thường được trình bày dưới dạng bảng tần số. Ở dạng cơ bản nhất thì bảng tần số thường bao gồm hai cột tính toán là tần số và tần suất %.

Ví dụ từ kết quả cuộc khảo sát doanh nghiệp và thương hiệu năm 2002, ta có bảng tần số diễn tả mẫu nghiên cứu 498 doanh nghiệp theo tiêu thức vùng địa lý như sau:

Bảng 3.1: Số lượng các doanh nghiệp khảo sát chia theo vùng địa lý

Vùng	Tần số	Tần suất %
Miền Bắc	113	22,7
Miền Trung	34	6,8
Đông Nam Bộ	293	58,8
Đồng bằng sông Cửu Long	58	11,6
Cộng	498	100,0

(Nguồn: Kết quả khảo sát hiện trạng xây dựng thương hiệu tại các doanh nghiệp Việt Nam, 2002, sách Thương Hiệu Việt, Nhà Xuất bản TPHCM)

3.2.1.2. Bảng tần số có ghép nhóm (có phân tổ)

Trong trường hợp có quá nhiều biểu hiện, mỗi biểu hiện là một nhóm thì bảng tần số sẽ rất dài gây khó khăn cho việc quan sát và cảm nhận. Các quan sát cần được phân tổ tức là phân loại và xếp vào một số

tổ, nhóm nhất định. Mỗi tổ bây giờ sẽ bao gồm một hay một số biểu hiện tùy theo đặc điểm của đối tượng nghiên cứu, mục tiêu so sánh và phân tích.

Ví dụ: Trong cuộc khảo sát tái định cư phục vụ dự án nghiên cứu khả thi nâng cấp đô thị và làm sạch kênh Tân Hóa - Lò Gốm do Trường ĐH KHXHNV TPHCM thực hiện năm 2002, có 1037 hộ gia đình được phỏng vấn. Công việc của các chủ hộ được tóm tắt và trình bày trong Bảng 3.2a.

Bảng 3.2a: Công việc của chủ hộ

Công việc của chủ hộ		Tần số	%
Có hoạt động kinh tế	Làm việc trong nhà máy	4	0,4
	Làm việc toàn thời gian trong cty nhà nước hay liên doanh	47	4,5
	Làm việc toàn thời gian trong hộ kinh doanh cá thể	61	5,9
	Làm việc không thường xuyên từ 1-10 giờ 1 tuần	1	0,1
	Làm việc theo thời vụ trong các xưởng chế biến	1	0,1
	Lao động tự do	237	22,9
	Làm việc trong các cơ quan nhà nước	5	0,5
	Làm việc chân tay trong các cơ quan nhà nước	13	1,3
	Làm việc văn phòng trong các cơ quan nhà nước	16	1,5
	Làm việc trong cửa hàng	26	2,5
	Làm việc trong các văn phòng (không phải cơ quan chính phủ, công ty nhà nước hay Liên doanh)	2	0,2
	Buôn bán nhỏ	3	0,3
	Bán đồ kim khí điện máy	3	0,3
	Bán thịt cá	3	0,3
	Bán rau và các loại củ	7	0,7
	Bán thức ăn đã chế biến	6	0,6
	Bán quần áo và các sản phẩm may mặc	4	0,4
	Bán mỹ phẩm, tạp hóa	9	0,9
	Bán các loại khác	14	1,4
	Bán hàng rong các loại khác	7	0,7
	Bán hàng rong thức ăn chưa chế biến và đã chế biến	30	2,9
	Bán hàng rong quần áo, giày dép và tạp hóa	9	0,9
	Bán các loại khác	23	2,2
	Làm việc tại nhà	3	0,3
	Khác	79	7,6
	Tự kinh doanh	14	1,4
	Chế biến thực phẩm (ml)	7	0,7
	May/sửa quần áo	5	0,5
	Làm đồ kim loại	6	0,6
	Làm đồ nhựa	4	0,4
	Sửa chữa đồ	9	0,9
	Tổng	658	63,5
Không hoạt động kinh tế	Hưởng lương hưu	10	1,0
	Nhận trợ cấp của họ hàng	21	2,0
	Thu nhập từ cho thuê nhà	16	1,5

	Không nghề nghiệp không việc làm	332	32,0
	Tổng	379	36,5
Tổng		1037	100,0

Trong Bảng 3.2a chúng ta nhận thấy công việc của chủ hộ rất đa dạng và phong phú, những người nghiên cứu đã ghép nhóm (phân tổ) các công việc có tính chất tương tự lại với nhau. Kết quả được trình bày trong Bảng 3.2b.

Bảng 3.2b: Công việc của chủ hộ (đã ghép nhóm/phân tổ)

Công việc của chủ hộ		Tần số	Tần suất %
Có hoạt động kinh tế	Làm việc trong nhà máy, xí nghiệp, công ty	114	11,0
	Làm nghề tự do	237	22,9
	Làm việc trong các cơ quan nhà nước	34	3,3
	làm việc trong các cửa hàng	26	2,5
	Làm việc trong các văn phòng	2	0,2
	Buôn bán nhỏ	44	4,2
	Bán hàng rong	51	4,9
	Làm việc tại nhà	105	10,1
	Tự kinh doanh	45	4,3
	Tổng	658	63,5
Không hoạt động kinh tế	Thu từ nguồn khác (được chuyển nhượng)	47	4,5
	Không việc làm	332	32,0
	Tổng	379	36,5
Tổng cộng		1037	100,0

Trong Bảng 3.2b này, chúng ta thấy những công việc liên quan đến sản xuất được ghép lại với nhau; làm việc trong cơ quan nhà nước dù lao động chân tay hay văn phòng được gom chung lại; buôn bán có nơi chỗ được sắp xếp vào nhóm buôn bán nhỏ; các công việc sản xuất dịch vụ khác do người chủ hộ tự bỏ vốn, tự chịu rủi ro và tự điều hành thì xếp vào nhóm tự kinh doanh. Những chủ hộ sống nhờ vào lương hưu, trợ cấp của con cái họ hàng và từ thu nhập cho thuê nhà thì không có hoạt động kinh tế, thu từ những người khác chuyển nhượng mà có thì xếp vào loại không hoạt động kinh tế, thu từ nguồn khác.

Một nguyên tắc khi ghép nhóm hay phân tổ là số tổ hay nhóm không nên quá ít hay quá nhiều. Số tổ, nhóm quá ít thì dữ liệu tổng hợp ít chi tiết, số tổ, nhóm quá nhiều thì dữ liệu tổng hợp lại quá nhiều chi tiết khó quan sát và cảm nhận. Các đơn vị sắp xếp vào cùng một tổ, nhóm nên có tính chất càng tương tự nhau càng tốt.

3.2.2 TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU ĐỊNH LƯỢNG

3.2.2.1. Phương pháp nhánh và lá

Nội dung cơ bản của phương pháp nhánh và lá là các dữ liệu thu thập được sẽ được tách thành hai phần: phần nhánh và phần lá. Việc phân chia này chỉ có tính quy ước và khá linh hoạt. Các chữ số bên phải của dữ liệu là lá (có thể là 1 hay 2 chữ số ở hàng đơn vị hay hàng chục), tương ứng các chữ số còn lại bên tay trái (có thể là 1 hay 2 chữ số ở hàng chục hay hàng trăm) là nhánh. Hãy theo dõi ví dụ dưới đây để hình dung phương pháp.

Ví dụ như chúng ta có dữ liệu trong hai mẫu điều tra nhỏ về tuổi của các sinh viên tại chức đang học năm thứ 1 của hai ngành như sau:

Tuổi của 30 sinh viên ngành KTKT

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

Tuổi của 30 sinh viên ngành QTKD

31	23	36	24	20	21	42	33	30	31	37	33	19	40	45
35	26	34	29	38	27	39	25	28	26	33	31	22	32	37

Chúng ta bắt đầu từ việc xác định nhánh và lá. Trong ví dụ trên mỗi dữ liệu chỉ có 2 chữ số, nên việc xác định rất đơn giản, chữ số hàng đơn vị là lá và chữ số hàng chục là nhánh. Ở đây ta có thể thấy ngay là đối với ngành KTKT chỉ có 3 nhánh 1,2,3 vì tuổi thấp nhất là 19 và cao nhất là 39. Đối với ngành QTKD có đến 4 nhánh vì tuổi thấp nhất là 19 và cao nhất là 45. Ta viết ra 3 nhánh của tuổi sinh viên ngành KTKT

```
1 |
2 |
3 |
```

Sau đó lần lượt xếp các dữ liệu quan sát về tuổi sinh viên ngành KTKT vào 3 nhánh này. Chúng ta bắt đầu từ dòng 1 và đi từ trái sang phải. Người thứ nhất là 28 tuổi, như vậy người này thuộc nhánh 2 và có lá là 8. Ta ghi vào biểu đồ nhánh và lá như sau:

```
1 |
2 | 8
3 |
```

Tiếp tục, người thứ hai là 23, thuộc nhánh 2 và có lá là 3, ta ghi tiếp:

```
1 |
2 | 8 3
3 |
```

Tuần tự như vậy đến người thứ 30, ta sẽ có biểu đồ nhánh và lá như sau:

```
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 này dễ nhìn hơn, ta sẽ sắp xếp lại thứ tự của các chữ số trong phần lá tăng dần từ trái qua phải. Biểu đồ nhánh và lá hoàn chỉnh như sau:

```
1  |  9
2  |  0 1 1 1 2 2 2 3 4 5 5 6 6 7 7 7 8 8 9 9 9
3  |  0 0 1 2 3 5 7 9
```

Khi các nhánh có phần lá quá dài (nhiều trị số quan sát), chúng ta có thể tách mỗi nhánh thành 2 nhánh nhỏ, ví dụ như nhánh 2 và 3 có thể tách thành nhánh 2 dưới, nhánh 2 trên. Nhánh 2 dưới nhận các trị số từ 20 đến 24, nhánh 2 trên nhận các trị số từ 25 đến 29. Tương tự như vậy đối với nhánh 3 dưới và 3 trên. Lúc đó biểu đồ thân và lá sẽ có hình dạng như sau:

Hình 3.1: Biểu đồ thân và lá tuổi của sinh viên ngành KTKT

```
KTKT Stem-and-Leaf Plot
 Frequency     Stem &  Leaf
     1.00        1 .  9
     9.00        2 .  011122234
    12.00        2 .  556677788999
     5.00        3 .  00123
     3.00        3 .  579

 Stem width:        10
 Each leaf:       1 case(s)
```

(Ghi chú: kết quả từ xử lý bằng chương trình thống kê SPSS for Windows)

Trong Hình 3.1, chúng ta thấy máy tính đã tự động tách nhánh 2 và 3 thành hai nhánh nhỏ hơn để biểu đồ cân đối hơn, và có thêm cột tần số cho biết có bao nhiêu lá (tần số) trong một nhánh (nhóm).

Tương tự như vậy, chúng ta có kết quả tóm tắt dữ liệu về tuổi của sinh viên ngành QTKD trong Hình 3.2.

Hình 3.2: Biểu đồ thân và lá tuổi của sinh viên ngành QTKD

```
QTKD Stem-and-Leaf Plot
 Frequency     Stem &  Leaf
     1.00        1 .  9
     5.00        2 .  01234
     6.00        2 .  566789
     9.00        3 .  011123334
     6.00        3 .  567789
     2.00        4 .  02
     1.00        4 .  5
```

```
Stem width:      10
Each leaf:     1 case(s)
```

(Ghi chú: kết quả từ xử lý bằng chương trình thống kê SPSS for Windows)

Nhìn vào hai biểu đồ nhánh và lá trên, ta có thể dễ dàng nhận thấy, tuổi của sinh viên tại chức ngành KTKT tập trung trong khoảng 20 đến 29 (nhánh 2 có nhiều lá nhất). Trong khi đó, tuổi của sinh viên tại chức ngành QTKD không tập trung nhiều vào khoảng 20-29 bằng khoảng 30-39, và phân tán nhiều hơn (nhiều nhánh hơn). Như vậy có thể nhận định rằng, tuổi sinh viên tại chức ngành QTKD lớn hơn và biến thiên nhiều hơn ngành KTKT.

Biểu đồ nhánh và lá tương tự như biểu đồ phân phối tần số (nếu chúng ta quay biểu đồ thân và lá 90 độ ngược chiều kim đồng hồ), ta có thể tưởng tượng mỗi nhánh tương ứng với một cột trong biểu đồ cột phân phối tần số, nhưng các lá trong các nhánh còn giúp ta quan sát chi tiết từng trị số thu được. Chi tiết hơn như vậy làm cho biểu đồ nhánh và lá có lợi thế khi tóm tắt và trình bày dữ liệu trong trường hợp có ít quan sát. Khi số lượng quan sát nhiều lên đến hàng trăm hay hàng ngàn, thì chi tiết như vậy làm cho người xem kết quả tóm tắt rối mắt, cần loại bỏ bớt chi tiết, lúc đó bảng tần số thích hợp hơn.

3.2.2.2 Bảng tần số

Đối với dữ liệu định lượng thu thập từ các thang đo định lượng, khi số quan sát khá nhiều lên đến vài chục, hàng trăm hoặc hơn, thì chúng ta cần lập bảng tần số tương tự như trong trường hợp dữ liệu định tính. Trong trường hợp các trị số thu thập được có ít giá trị thì mỗi trị số là một tổ hay nhóm. Trong trường hợp có quá nhiều trị số tức là có quá nhiều nhóm, thì các trị số hay nhóm sẽ được sắp xếp/ghép lại với nhau để số tổ nhóm ít lại, dễ cho việc quan sát và cảm nhận. Bảng 3.3a trình bày số chủ hộ có thu nhập trong 1037 hộ đã khảo sát.

Bảng 3.3a bây giờ có thêm cột cuối là cột tần suất tích lũy. Cột này được tính bằng cách cộng dồn các tần suất % lại theo thứ tự từ tổ nhóm đầu tiên đến tổ nhóm cuối cùng. Cột % tích lũy này có ý nghĩa khi tóm tắt dữ liệu định lượng hay dữ liệu thứ bậc. Khi tóm tắt dữ liệu thu thập từ thang đo định danh ít khi sử dụng % tích lũy này. Nhìn vào Bảng 3.3a, ta có thể dễ dàng thấy số người có thu nhập trong các hộ gia đình khảo sát chủ yếu

là từ 1 đến 4 người. Bảng này có thể gọn và đẹp hơn nếu ta ghép các nhóm cuối lại với nhau. Kết quả ghép nhóm được trình bày trong Bảng 3.3b

Bảng 3.3a: Số người có thu nhập trong hộ gia đình

Số người	Tần số	Tần suất %	Tần suất tích lũy %
0	5	0,5	0,5
1	282	27,2	27,7
2	373	36,0	63,7
3	145	14,0	77,7
4	105	10,1	87,8
5	72	6,9	94,7
6	30	2,9	97,6
7	16	1,5	99,1
8	4	0,4	99,5
9	3	0,3	99,8
10	1	0,1	99,9
11	1	0,1	100,0
Tổng	1037	100,0	

Bảng 3.3b: Số người có thu nhập trong hộ gia đình (đã ghép nhóm)

Số người	Tần số	%	% tích lũy
0	5	0,5	0,5
1	282	27,2	27,7
2	373	36,0	63,6
3	145	14,0	77,6
4	105	10,1	87,8
5-6	102	9,8	97,6
7 trở lên	25	2,4	100,0
Tổng	1037	100,0	

Trong trường hợp các trị số thu thập có rất nhiều thì bảng tần số đối với dữ liệu thu thập được hầu như ít có ý nghĩa tóm tắt dữ liệu, ít tạo điều kiện cho người nghiên cứu quan sát và nhận thức, chúng ta phải ghép các trị số gần nhau lại thành 1 tổ, 1 nhóm. Bảng 3.4a trình bày dữ liệu về thu nhập hàng tháng của các chủ hộ. Bảng này rất dài và có quá nhiều nhóm, người đọc rất khó nắm bắt những điểm cơ bản từ dữ liệu thu thập. Từ bảng này chúng ta có thể ghép các trị số gần lại với nhau và kết quả được trình bày trong Bảng 3.4b.

Bảng 3.4a: Thu nhập của người chủ hộ (ngàn đồng/tháng)

	Tần số	%
Không có thu nhập	332	32,0
80	1	0,1
100	3	0,3
120	1	0,1
125	1	0,1
135	1	0,1
150	1	0,1
200	8	0,8
210	1	0,1
249	1	0,1
250	1	0,1
272	1	0,1
300	33	3,2
350	4	0,4
400	24	2,3
450	15	1,4
500	45	4,3
520	1	0,1
550	1	0,1
600	78	7,5
650	2	0,2
700	37	3,6
750	4	0,4
800	46	4,4
900	57	5,5
920	1	0,1
999	1	0,1
1000	96	9,3
1100	4	0,4
1200	42	4,1
1300	2	0,2
1500	78	7,5
1600	1	0,1
1700	5	0,5
1800	10	1,0
2000	33	3,2
2100	1	0,1
2200	1	0,1
2400	2	0,2
2500	11	1,1
2600	1	0,1
3000	21	2,0
3500	1	0,1
4000	7	0,7
4500	3	0,3
5000	8	0,8
6000	1	0,1
7000	1	0,1
10000	2	0,2
15000	1	0,1
20000	2	0,2
Không trả lời	2	0,2
Tổng	1037	100,0

Bảng 3.4b: Thu nhập của người chủ hộ (ngàn đồng/tháng) (đã ghép lớp)

	Tần số	%
Không có thu nhập	332	32,0
từ 250 trở xuống	19	1,8
251-500	122	11,8
501-750	123	11,9
751-1000	201	19,4
1001-1250	46	4,4
1251-1500	80	7,7
1501-1750	6	0,6
1751-2000	43	4,1
2001-3000	37	3,6
3001-4000	8	0,8
4001-5000	11	1,1
trên 5000	7	0,7
Không trả lời	2	0,2
Tổng	1037	100,0

Bảng 3.4b giúp người đọc nhanh chóng và dễ cảm nhận được mức độ thu nhập của các chủ hộ được khảo sát hơn Bảng 3.4a. Trong khu vực Lò Gốm – Tân Hóa này thì đa số chủ hộ là người không có thu nhập hay thu nhập hàng tháng rất thấp (dưới 1,5 triệu đồng).

Khi phân tổ đối với dữ liệu định lượng thì chúng ta có thể phân tổ có khoảng cách tổ đều hay không đều tùy theo tính chất của hiện tượng nghiên cứu hay tùy theo mục đích so sánh và phân tích của những người nghiên cứu. Nếu mức độ của các đơn vị phân tán đều thì sử dụng phân tổ có khoảng cách đều. Nếu các đơn vị có mức độ phân tán không đều (như trong ví dụ trên), chúng ta có thể phân tổ có khoảng cách không đều chứ không nhất thiết phải phân tổ đều như Bảng 3.4b.

3.2.3 CÁC ĐẠI LƯỢNG THỐNG KÊ MÔ TẢ

Đối với dữ liệu định lượng, chúng ta có thể tóm tắt tốt hơn khi có khối lượng dữ liệu lớn, đó là dùng các đại lượng thống kê mô tả. Các đại lượng thống kê mô tả thường sử dụng nhất được chia làm hai nhóm: nhóm các đại lượng thể hiện mức độ tập trung của dữ liệu, và nhóm các đại lượng thể hiện độ phân tán của dữ liệu. Chúng ta cần phải tính toán cả hai đại lượng đo lường này, vì chúng phản ảnh hai khía cạnh của tập hợp dữ liệu đã thu thập.

Ví dụ như trong trường hợp nghiên cứu ảnh hưởng của việc đi làm thêm đến kết quả học tập của sinh viên, chúng ta khảo sát hai nhóm sinh viên. Nhóm không đi làm thêm gồm 150 sinh viên và nhóm có đi làm thêm có 100 sinh viên. Vấn đề đầu tiên là chúng ta muốn biết nhóm nào có kết quả học tập (điểm trung bình học tập) cao hơn, chúng ta có thể tính điểm trung bình (một chỉ tiêu đo lường mức độ tập trung của dữ liệu) của hai nhóm rồi so sánh với nhau. Có thể nhóm sinh viên có làm thêm kết quả học tập không thấp hơn nhóm sinh viên không đi làm thêm, và nếu dựa vào điểm trung bình học tập để kết luận thì có thể chúng ta đã bỏ qua những thông tin quan trọng.

Một khía cạnh quan trọng khác mà nhiều người ít chú ý đó là mức độ đồng đều, nhóm sinh viên nào có kết quả học tập đồng đều hơn. Có thể nhóm sinh viên có đi làm thêm có kết quả học tập rất không đồng đều, một số sinh viên đi làm thêm điểm trung bình học tập rất cao, trong khi đó, một số sinh viên đi làm thêm lại có điểm trung bình rất thấp. Trong trường hợp này, nếu chỉ dựa vào điểm trung bình thì kết luận về vấn đề nghiên cứu của chúng ta sẽ rất đơn giản so với thực tế đang diễn ra. Cho nên nếu chỉ dựa vào đại lượng đo lường mức độ tập trung thì chưa đủ.

Các đại lượng đo lường mức độ tập trung của dữ liệu thường dùng là trung bình cộng, mốt (mode), trung vị (median). Trong đó trung bình cộng được sử dụng phổ biến nhất. Các đại lượng đo lường độ phân tán (hay độ đồng đều) của dữ liệu thường dùng là khoảng biến thiên, độ lệch tuyệt đối bình quân, phương sai, độ lệch chuẩn và hệ số biến thiên. Trong đó độ lệch chuẩn được sử dụng phổ biến nhất.

Bảng 3.5: Tóm tắt dữ liệu về thu nhập (ngàn đồng/tháng) của chủ hộ bằng các đại lượng thống kê mô tả (dự án Tân Hóa – Lò Gốm, 2002)

Số quan sát	Số người có TN	Thấp nhất	Cao nhất	TB cộng	Trung vị	Mode	Độ lệch chuẩn	Sai số chuẩn
1037	N=703	80	20.000	1.225,4	900	1.000	1.507,8	56,9

(Ghi chú: kết quả tính tóm tắt và trình bày qua xử lý bằng chương trình thống kê SPSS[1])

[1] Sinh viên có thể tham khảo phần hướng dẫn sử dụng chương trình thống kê này trong sách "Xử lý dữ liệu nghiên cứu với SPSS for Windows", Hoàng Trọng, Nhà xuất bản Thống Kê, 2002.

Trong khảo sát 1037 hộ gia đình của dự án Tân Hóa – Lò Gốm, nhìn vào Bảng 3.5, chúng ta thấy thu nhập trung bình tháng của chủ hộ là 1.225,4 ngàn đồng (chỉ tính trên 703 chủ hộ có thu nhập và có trả lời), độ lệch chuẩn (đã hiệu chỉnh) là 1507,8 ngàn đồng. Trong bảng tóm tắt này còn có một số đại lượng khác như trung vị và mode, sai số chuẩn (dùng để ước lượng khoảng). Qua bảng có thể thấy rằng thu nhập trung bình của các chủ hộ không cao, nhưng rất không đồng đều (độ lệch chuẩn rất lớn).

3.2.4 BẢNG KẾT HỢP

Trong các phần trước chúng ta chỉ mới xem xét việc tóm tắt các dữ liệu theo từng tiêu thức, trong phần này chúng ta sẽ nói đến việc tóm tắt dữ liệu theo 2 hay nhiều tiêu thức cùng một lúc, và kết quả sẽ được trình bày trong bảng kết hợp. Bảng kết hợp này giúp chúng ta có cái nhìn chi tiết và sâu hơn về đối tượng đang nghiên cứu.

3.2.4.1. Bảng kết hợp 2 dữ liệu định tính

Ví dụ trong cuộc khảo sát thuộc dự án Tân Hóa – Lò Gốm, chúng ta có bảng tóm tắt về công việc của chủ hộ. Chủ dự án quan tâm đến các kết quả tóm tắt công việc của chủ hộ riêng cho từng khu vực thuộc dự án này.

Bảng 3.6a: Công việc của chủ hộ, chi tiết theo quận

Công việc của chủ hộ		QTân Bình		Q6		Q11	
		Tần số	% cột	Tần số	% cột	Tần số	% cột
Có hoạt động kinh tế	Làm việc trong nhà máy	18	14,3	91	10,5	5	11,6
	Làm nghề tự do	39	31,0	187	21,5	11	25,6
	Làm việc trong các CONN	2	1,6	30	3,5	2	4,7
	làm việc trong các CH	3	2,4	22	2,5	1	2,3
	Làm việc trong VP			2	0,2		
	Buôn bán nhỏ	4	3,2	39	4,5	1	2,3
	Bán hàng rong	5	4,0	45	5,2	1	2,3
	Làm việc tại nhà	10	7,9	88	10,1	7	16,3
	Tự kinh doanh	8	6,3	37	4,3		
	Tổng	89	70,6	541	62,3	28	65,1
Không hoạt động kinh tế	Thu nhập từ nguồn khác	6	4,8	41	4,7		
	không việc làm	31	24,6	286	32,9	15	34,9
	Tổng	37	29,4	327	37,7		
Tổng		126	100,0	868	100,0	43	100,0

Nhìn vào Bảng 3.6a ta có thể thấy ngay quận 11 là quận có tỉ lệ chủ hộ không có việc làm cao nhất, quận Tân Bình là quận có tỉ lệ chủ hộ làm việc liên quan đến sản xuất cao nhất, và làm nghề tự do nhiều nhất. Bảng

3.6a này giống bảng 3.2b, chỉ có khác là tần số và % theo cột được dùng để tóm tắt riêng cho từng quận khảo sát.

Khi lập một bảng kết hợp như Bảng 3.6a, thì vấn đề cơ bản là ngoài các tần số thì thường phải tính tỉ lệ phần trăm. Có hai tỉ lệ % được sử dụng phổ biến nhất là % tính theo cột và % tính theo dòng.

Tỉ lệ % tính theo cột được tính bằng cách lấy từng tần số trong cột chia cho tổng các tần số trong cột đó. Đặc điểm của các % tính theo cột này là thường thì tổng các % trong cùng một cột phải là 100%. Còn tỉ lệ % tính theo dòng được tính bằng cách lấy từng tần số trong dòng chia cho tổng các tần số trong dòng đó. Đặc điểm của các % tính theo dòng này là thường thì tổng các % trong cùng 1 dòng phải là 100%.

Vấn đề đặt ra là khi nào sử dụng % theo cột, khi nào sử dụng % theo dòng. Việc sử dụng % nào là tùy thuộc vào mục đích so sánh của người nghiên cứu và vị trí của các tiêu thức, loại dữ liệu để ở dòng hay cột trong bảng tóm tắt dữ liệu theo 2 tiêu thức này.

Trong Bảng 3.6a, chúng ta có thể dễ dàng nhận thấy công việc của chủ hộ được sắp xếp theo các dòng của bảng, còn địa bàn sinh sống của các chủ hộ này được sắp xếp theo cột. Mục đích lập bảng tổng hợp này là so sánh công việc của chủ hộ giữa các quận khảo sát. Quận chính là yếu tố quan trọng chúng ta cần so sánh về công việc của chủ hộ, thông tin này được sắp xếp ở các cột của bảng, thì trong trường hợp này, chúng ta tính % theo cột. Nếu trong Bảng 3.6a, chúng ta sắp xếp đảo lại, quận được sắp xếp vào các dòng còn công việc của chủ hộ được sắp xếp vào cột thì trong trường hợp này chúng ta sẽ tính % theo dòng. Bảng 3.6b thể hiện nội dung tóm tắt dữ liệu giống như Bảng 3.6a, nhưng hình thức trình bày khác nhau nhau ở chỗ hoán chuyển giữa cột và dòng, cho nên loại % sử dụng cũng khác nhau.

Bảng 3.6b: Công việc của chủ hộ, chi tiết theo quận

Công việc của chủ hộ		Có hoạt động kinh tế										Không hoạt động kinh tế			Tổng
		Làm việc trong NM	Làm nghề tự do	Làm việc trong các CQNN	làm việc trong các CH	Làm việc trong VP	Buôn bán nhỏ	Bán hàng rong	Làm việc tại nhà	Tự kinh doanh	Tổng	Thu từ nguồn khác	không việc làm	Tổng	
Tân Bình	Tần số	18	39	2	3		4	5	10	8	89	6	31	37	126
	% Dòng	14.3	31.0	1.6	2.4		3.2	4.0	7.9	6.3	70.6	4.8	24.6	29.4	100
6	Tần số	91	187	30	22	2	39	45	88	37	541	41	286	327	868

	% Dòng	10,5	21,5	3,5	2,5	0,2	4,5	5,2	10,1	4,3	62,3	4,7	32,9	37,7	100
	Tần số	5	11	2	1		1	1	7		28		15		43
11	% Dòng	11,6	25,6	4,7	2,3		2,3	2,3	16,3		65,1		34,9		100

Nếu chúng ta vẫn giữ nguyên vị trí các đặc trưng hay tiêu thức như trong Bảng 3.6 a, tức là quận được sắp xếp theo cột, mà chúng ta lại sử dụng % tính theo dòng (Bảng 3.6c) thì các kết quả sẽ ít có ý nghĩa hay không có ý nghĩa đối với nội dung nghiên cứu.

Bảng 3.6c: Công việc của chủ hộ, chi tiết theo quận

Công việc của chủ hộ		QTân Bình		Q6		Q11			
		Tần số	% dòng	tần số	% dòng	tần số	% dòng	tần số	% dòng
Có hoạt động Kinh tế	Làm việc trong nhà máy	18	15,8	91	79,8	5	4,4	114	100,0
	Làm nghề tự do	39	16,5	187	78,9	11	4,6	237	100,0
	Làm việc trong các CQNN	2	5,9	30	88,2	2	5,9	34	100,0
	làm việc trong các CH	3	11,5	22	84,6	1	3,8	26	100,0
	Làm việc trong VP			2	100,0			2	100,0
	Buôn bán nhỏ	4	9,1	39	88,6	1	2,3	44	100,0
	Bán hàng rong	5	9,8	45	88,2	1	2	51	100,0
	Làm việc tại nhà	10	9,5	88	83,8	7	6,7	105	100,0
	Tự kinh doanh	8	17,8	37	82,2		0	45	100,0
	Tổng	89	13,5	541	82,2	28	4,3	658	100,0
Không hoạt động kinh tế	Thu nhập từ nguồn khác	6	12,8	41	87,2			47	100,0
	không việc làm	31	9,3	286	86,1	15	4,5	332	100,0
	Tổng	37	10,2	327	89,8			364	100,0
Tổng		126	12,2	868	83,7	43	4,1	1037	100,0

Trong Hình 3.6c thì các con số % theo dòng trong trường trường hợp này ít có ý nghĩa giúp ta so sánh giữa các quận, các % ở quận 6 thường rất cao vì số đơn vị khảo sát ở quận 6 cao nhất (là 868) trong toàn bộ 1037 hộ được khảo sát, nhìn vào các con số % này ta khó có thể cảm nhận được và đưa ra nhận xét đúng đắn được.

3.2.4.2 Bảng kết hợp 3 dữ liệu định tính

Tùy theo yêu cầu so sánh và phân tích, chúng ta có thể tóm tắt và trình bày bằng những bảng kết hợp phức tạp hơn, trong đó tóm tắt dữ liệu kết hợp 3 hay nhiều hơn nữa dữ liệu định tính.

Trong ví dụ về công việc của chủ hộ, chúng ta có thể chi tiết hóa Bảng 3.6a theo hai độ tuổi là: chủ hộ trong độ tuổi lao động và chủ hộ trên tuổi lao động. Hoặc chi tiết hóa bảng 3.6a theo trình độ học vấn hay giới tính. Trong trường hợp này chúng ta có bảng kết hợp 3 dữ liệu định tính. Hay vừa chi tiết hóa Bảng 3.6a theo vừa độ tuổi vừa giới tính, và ta sẽ có bảng

kết hợp đến 4 dữ liệu định tính. Bảng 3.7 trình bày bảng kết hợp 3 dữ liệu định tính, trong đó tóm tắt dữ liệu về công việc của chủ hộ kết hợp với dữ liệu về nơi cư ngụ (quận) và giới tính của chủ hộ.

Bảng 3.7: Công việc của chủ hộ, chi tiết theo quận và giới tính

Công việc của chủ hộ		Tân Bình			6			11		
		Nữ	Nam	Tổng	Nữ	Nam	Tổng	Nữ	Nam	Tổng
Tần số	Làm việc trong nhà máy	3	15	18	30	61	91	1	4	5
	Làm nghề tự do	12	27	39	44	143	187	6	5	11
	Làm việc trong cơ quan NN	1	1	2	5	25	30		2	2
	làm việc trong các cửa hàng	1	2	3	10	12	22		1	1
	Làm việc trong văn phòng					2	2			
	Buôn bán nhỏ	4		4	20	19	39		1	1
	Bán hàng rong	3	2	5	32	13	45		1	1
	Làm việc tại nhà	5	5	10	53	35	88	3	4	7
	Tự kinh doanh	1	7	8	8	29	37			
	Thu nhập từ nguồn khác	4	2	6	25	16	41			
	không việc làm	23	8	31	211	75	286	10	5	15
	Total	57	69	126	438	430	868	20	23	43
%	Làm việc trong nhà máy	5,3	21,7	14,3	6,8	14,2	10,5	5,0	17,4	11,6
	Làm nghề tự do	21,1	39,1	31,0	10,0	33,3	21,5	30,0	21,7	25,6
	Làm việc trong cơ quan NN	1,8	1,4	1,6	1,1	5,8	3,5		8,7	4,7
	làm việc trong các cửa hàng	1,8	2,9	2,4	2,3	2,8	2,5		4,3	2,3
	Làm việc trong văn phòng					0,5	0,2			
	Buôn bán nhỏ	7,0		3,2	4,6	4,4	4,5		4,3	2,3
	Bán hàng rong	5,3	2,9	4,0	7,3	3,0	5,2		4,3	2,3
	Làm việc tại nhà	8,8	7,2	7,9	12,1	8,1	10,1	15,0	17,4	16,3
	Tự kinh doanh	1,8	10,1	6,3	1,8	6,7	4,3			
	Thu nhập từ nguồn khác	7,0	2,9	4,8	5,7	3,7	4,7			
	không việc làm	40,4	11,6	24,6	48,2	17,4	32,9	50,0	21,7	34,9
Total		100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

(Ghi chú: kết quả từ xử lý bằng chương trình thống kê SPSS for Windows)

Trong thực tế, khi cần chúng ta có thể ghép các bảng có liên quan lại với nhau hay tách rời nhau tùy theo mức độ dễ liên hệ và so sánh cho người đọc và tùy theo quy mô của bảng. Nếu ghép lại mà quy bảng không quá lớn và không quá phức tạp thì tạo điều kiện cho người xem một cái nhìn bao quát hơn. Ví dụ như Bảng 3.8, chúng ta có thể tách thành 3 bảng rời, một bảng cho TP Hồ Chí Minh, một bảng cho khu vực thành thị, và một bảng cho khu vực nông thôn của TPHCM. Lúc đó mỗi bảng là một bảng kết hợp 2 dữ liệu định tính (loại nhà và hình thức sở hữu). Nếu chúng ta ghép chung 3 bảng này với nhau thành Bảng 3.8, thì bảng này là một bảng kết hợp 3 dữ liệu định tính (khu vực, loại nhà và hình thức sở hữu).

Trong trường hợp dữ liệu rất nhiều và đa dạng, và để đáp ứng nhu cầu nghiên cứu, việc tóm tắt dữ liệu sẽ cho ra những bảng rất chi tiết. Ví dụ như Bảng 3.8, chúng ta có thể chi tiết theo địa bàn quận huyện và trình bày thành bảng 4 loại dữ liệu định tính. Nguyên tắc cơ bản khi tóm tắt dữ liệu là mức độ kết hợp các loại dữ liệu tùy thuộc vào yêu cầu phân tích, chúng ta không nên tạo ra các bảng thật to mà kém ý nghĩa, hay số đơn vị mẫu quá ít để bảng tóm lược có thể sử dụng được. Bảng tóm lược dữ liệu càng phối hợp nhiều loại dữ liệu thì số lượng mẫu khảo sát phải càng lớn, nếu không thì số lượng quan sát trong mỗi ô của bảng sẽ rất nhỏ và như vậy ít ý nghĩa thống kê.

Bảng 3.8: Số hộ có nhà ở thuộc loại kiên cố, bán kiên cố, khung gỗ chia theo loại nhà và hình thức sở hữu (đơn vị tính: hộ)

Loại nhà	Tổng số	Nhà riêng	Nhà thuê của NN	Nhà thuê mượn của tư nhân	Nhà của tập thể, tôn giáo	Nhà của NN và nhân dân cùng làm	Nhà chưa rõ quyền sở hữu
TP Hồ Chí Minh							
Tổng số	895.186	770.740	83.492	26.341	4.522	2.113	7.978
-Nhà kiên cố	212.866	159.273	46.043	3.894	1.782	516	1.358
-Nhà bán kiên cố	647.009	579.245	36.589	21.118	2.688	1.494	5.875
-Nhà khung gỗ lâu bền, mái lá	35.311	32.222	860	1.329	52	103	745
TPHCM thành thị							
Tổng số	781.323	659.366	83.049	24.997	4.277	1.773	7.861
-Nhà kiên cố	207.405	153.909	46.030	3.845	1.756	510	1.355
-Nhà bán kiên cố	547.189	481.752	36.162	19.857	2.470	1.178	5.770
-Nhà khung gỗ lâu bền, mái lá	26.729	23.705	857	1.295	51	85	736
TPHCM nông thôn							
Tổng số	113.863	111.374	443	1.344	245	340	117
-Nhà kiên cố	5.461	5.364	13	49	26	6	3
-Nhà bán kiên cố	99.820	97.493	427	1.261	218	316	105
-Nhà khung gỗ lâu bền, mái lá	8.582	8.517	3	34	1	18	9

(Nguồn: Kết quả tổng điều tra dân số và nhà ở 1/4/1999, Trung tâm Tính toán Thống kê Trung ương)

3.2.4.3. Bảng kết hợp dữ liệu định lượng với dữ liệu định tính

Đối với dữ liệu định lượng, nếu tóm tắt bằng bảng tần số tương tự như dữ liệu định tính, thì chúng ta có thể tóm tắt dữ liệu định lượng kết hợp với dữ liệu định tính tương tự như bảng kết hợp hai dữ liệu định tính. Nhất là khi chúng ta phân tổ lại, lúc đó dữ liệu định lượng của ta đã thật sự mang tính chất của dữ liệu định tính và các bảng kết hợp dữ liệu định lượng với dữ liệu định tính tương tự như bảng kết hợp 2 dữ liệu định tính đã trình bày ở trên. Các Bảng 3.9 cho thấy rõ điều này.

Bảng 3.9a: Thu nhập trung bình hàng tháng của chủ hộ, chi tiết theo quận

Thu nhập tháng (ngàn đồng)	QTÂN BÌNH		Q6		Q11		tổng	
	Tần số	% cột	Tần số	% cột	Tần số	% cột	Tần số	% cột
Không có thu nhập	31	24.6	286	32.9	15	34.9	332	32.0
1-250			18	2.1	1	2.3	19	1.8
251-500	18	14.3	99	11.4	5	11.6	122	11.8
501-750	24	19.0	92	10.6	7	16.3	123	11.9
751-1000	30	23.8	164	18.9	7	16.3	201	19.4
1001-1250	8	6.3	37	4.3	1	2.3	46	4.4
1251-1500	5	4.0	71	8.2	4	9.3	80	7.7
1501-1750			6	0.7			6	0.6
1751-2000	3	2.4	38	4.4	2	4.7	43	4.1
2001-3000	4	3.2	32	3.7	1	2.3	37	3.6
3001-4000	1	0.8	7	0.8			8	0.8
4001-5000	1	0.8	10	1.2			11	1.1
trên 5000			7	0.8			7	0.7
Không trả lời	1	0.8	1	0.1			2	0.2
Tổng	126	100.0	868	100.0	43	100.0	1037	100.0

Bảng 3.9b: Thu nhập trung bình hàng tháng của chủ hộ, chi tiết theo quận

		Không TN	1-250	251-500	501-750	751-1000	1001-1250	1251-1500	1501-1750	1751-2000	2001-3000	3001-4000	4001-5000	Trên 5000	KTL	Tổng
TB	Tần số	31		18	24	30	8	5		3	4	1	1		1	126
	% Dòng	24.6		14,3	19,0	23.8	6,3	4,0		2,4	3,2	0,8	0,8		0,8	100
6	Tần số	286	18	99	92	164	37	71	6	38	32	7	10	7	1	868
	% Dòng	32,9	2,1	11,4	10,6	18,9	4,3	8,2	0,7	4,4	3,7	0,8	1,2	0,8	0,1	100
11	Tần số	15	1	5	7	7	1	4		2	1					43
	% Dòng	34,9	2.3	11,6	16,3	16,3	2,3	9,3		4,7	2,3					100
Tổng	Tần số	332	19	122	123	201	46	80	6	43	37	8	11	7	2	1037
	% Dòng	32,0	1,8	11,8	11,9	19,4	4,4	7,7	0,6	4,1	3,6	0,8	1,1	0,7	0,2	100

Nhìn vào Bảng 3.9a ta có thể thấy quận 11 là quận có tỉ lệ chủ hộ không có thu nhập cao nhất, và chúng ta cũng có thể cảm nhận ngay là thu nhập của các chủ hộ trong quận này cũng thấp nhất, quận Tân Bình là quận có tỉ lệ chủ hộ không có thu nhập thấp nhất. Bảng 3.9a này giống bảng 3.4b, chỉ có khác là tần số và % theo cột được dùng để tóm tắt riêng cho từng quận khảo sát. Bảng 3.9b thể hiện nội dung tóm tắt dữ liệu giống như Bảng 3.5a, nhưng hình thức trình bày khác nhau nhau ở chỗ hoán chuyển giữa cột và dòng, cho nên loại % sử dụng cũng khác nhau. Còn Bảng 3.9c thì các con số % theo dòng trong trường hợp này là không phù hợp vì kém ý nghĩa.

Bảng 3.9c: Thu nhập trung bình hàng tháng của chủ hộ, chi tiết theo quận

Thu nhập tháng (ngàn đồng)	QTÂN BÌNH		Q6		Q11		Tổng	
	Tần số	% Dòng	Tần số	% Dòng	Tần số	% Dòng	Tần số	% Dòng
Không có thu nhập	31	9,3	286	86,1	15	4,5	332	100,0
1-250			18	94,7	1	5,3	19	100,0
251-500	18	14,8	99	81,1	5	4,1	122	100,0
501-750	24	19,5	92	74,8	7	5,7	123	100,0
751-1000	30	14,9	164	81,6	7	3,5	201	100,0
1001-1250	8	17,4	37	80,4	1	2,2	46	100,0
1251-1500	5	6,3	71	88,8	4	5,0	80	100,0
1501-1750			6	100,0			6	100,0
1751-2000	3	7,0	38	88,4	2	4,7	43	100,0
2001-3000	4	10,8	32	86,5	1	2,7	37	100,0
3001-4000	1	12,5	7	87,5			8	100,0
4001-5000	1	9,1	10	90,9			11	100,0
Trên 5000			7	100,0			7	100,0
Không trả lời	1	50,0	1	50,0			2	100,0
Tổng	126	12,2	868	83,7	43	4,1	1037	100,0

Ngoài bảng tần số, dữ liệu định lượng thường được tóm tắt bằng các đại lượng thống kê mô tả. Trong nhiều trường hợp, người nghiên cứu cần so sánh giữa các nhóm về một hay một số mặt định lượng nào đó. Ví dụ như trung bình và độ lệch chuẩn về thu nhập của chủ hộ cần được tính tóm tắt riêng cho từng quận được khảo sát. Trong trường hợp này chúng ta tóm tắt dữ liệu về thu nhập của chủ hộ theo từng quận, và quận là yếu tố phân nhóm để tóm tắt dữ liệu đưa ra các kết quả chi tiết để phục vụ cho việc so sánh và phân tích.

Bảng 3.10 trình bày tóm tắt dữ liệu định lượng về thu nhập (ngàn đồng/tháng) của chủ hộ theo từng quận và chung cho toàn bộ mẫu khảo sát. Có thể thấy tại quận 6, mức thu nhập trung bình của các chủ hộ có thu nhập cao hơn hai quận Tân Bình và 11, nhưng mức thu nhập của chủ hộ trong khu vực quận 6 ít đồng đều nhất (biến thiên nhiều nhất) thể hiện ở mức chênh lệch lớn giữa giá trị cao nhất và thấp nhất hay độ lệch chuẩn mẫu là lớn nhất trong 3 quận.

Bảng 3.10: Tóm tắt dữ liệu về thu nhập (1000đồng/tháng) của chủ hộ bằng các đại lượng thống kê mô tả (dự án Tân Hóa – Lò Gốm) chia theo quận

Quận	số quan sát	số người có TN	thấp nhất	cao nhất	TB cộng	trung vị	Mốt	độ lệch chuẩn	sai số chuẩn
Tân Bình	126	94	272	5.000	989,6	850	600	748,7	77,2
6	868	581	80	20.000	1.276,6	900	1000	1.621,0	67,3
11	43	28	100	3.000	953,6	850	600	609,5	115,2
Phần tính chung	1037	703	80	20.000	1.225,4	900	1.000	1.507,8	56,9

(Ghi chú: kết quả tính tóm tắt và trình bày qua xử lý bằng chương trình thống kê SPSS)

3.2.5 TRÌNH BÀY KẾT QUẢ TÓM TẮT DỮ LIỆU BẰNG BIỂU ĐỒ

3.2.5.1. Ý nghĩa của biểu đồ

Biểu đồ và đồ thị thống kê là các hình vẽ, đường nét hình học dùng để mô tả có tính qui ước các số liệu thống kê. Bảng thống kê chỉ dùng các con số và cung cấp những thông tin chi tiết, còn biểu đồ và đồ thị thống kê sử dụng con số kết hợp với hình vẽ, đường nét và màu sắc để tóm tắt và trình bày các đặc trưng chủ yếu của hiện tượng nghiên cứu, phản ảnh một cách khái quát các đặc điểm về cơ cấu, mối liên hệ, quan hệ so sánh, xu hướng biến động ...của hiện tượng nghiên cứu.

Do dùng hình vẽ, đường nét và màu sắc để biểu hiện mức độ của hiện tượng, cho nên đồ thị thống kê rất sinh động, có sức hấp dẫn mạnh mẽ, giúp cho người xem nhận thức được những đặc điểm cơ bản của hiện tượng một cách dễ dàng, nhanh chóng, làm cho những người dù ít hiểu biết về thống kê vẫn có thể nhận ra được nội dung chủ yếu của vấn đề được trình bày trên đồ thị.

3.2.5.2. Các loại đồ thị thống kê

Theo nội dung phản ánh của đồ thị thống kê, có thể phân chia đồ thị thống kê thành các loại:

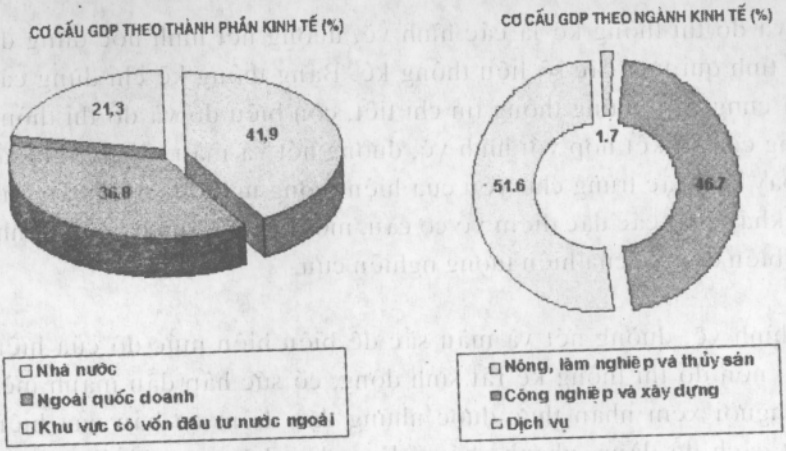
- Đồ thị kết cấu,
- Đồ thị phát triển
- Đồ thị hoàn thành kế hoạch hoặc định mức
- Đồ thị liên hệ
- Đồ thị so sánh,
- Đồ thị phân phối

Theo hình thức biểu hiện, có thể chia đồ thị thành các loại:

- Biểu đồ hình cột
- Biểu đồ tượng hình
- Biểu đồ diện tích (hình tròn, hình vuông, hình chữ nhật)
- Đồ thị đường gấp khúc (đường động thái)
- Bản đồ thống kê

Ví dụ thứ nhất, Hình 3.3 cho thấy hai dạng thức khác nhau của loại biểu đồ hình tròn. Cả hai đều có nội dung phản ảnh là cơ cấu của GDP thành phố Hồ Chí Minh năm 2002.

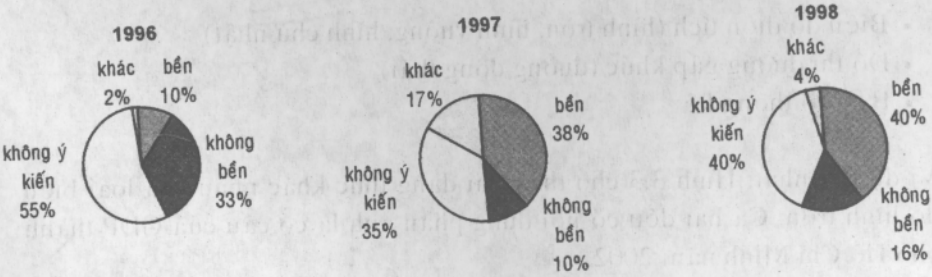
Hình 3.3: Cơ cấu Tổng sản phẩm quốc nội (GDP) theo thành phần kinh tế và ngành kinh tế



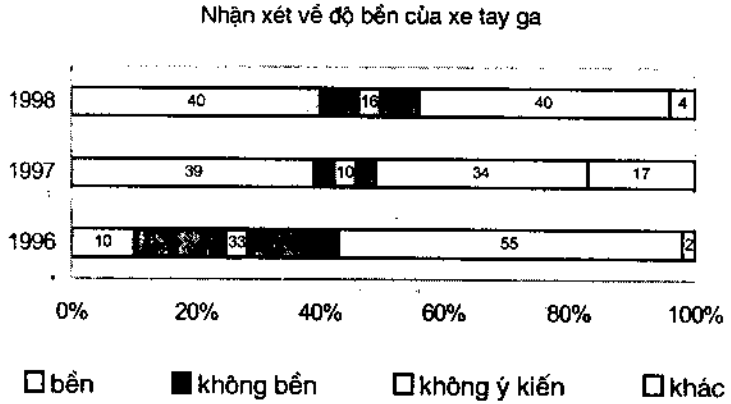
(Nguồn: Cục Thống Kê TP Hồ Chí Minh)

Ví dụ thứ hai, Hình 3.4a về hình thức là biểu đồ hình tròn, về nội dung thì phản ảnh cơ cấu ý kiến nhận xét về độ bền của xe tay ga của 198 người tiêu dùng tại TP Hồ Chí Minh qua các cuộc khảo sát định kỳ hàng năm về thị trường xe gắn máy. Để cho người xem dễ thấy sự thay đổi cơ cấu ý kiến này qua ba năm, chúng ta có thể sử dụng chỉ 1 biểu đồ thanh ngang hay cột dọc, trong đó mỗi thanh hay cột đại diện cho 1 năm, và có độ dài như nhau là 100%. Mỗi thanh lại được cắt ra thành nhiều đoạn nhỏ tượng trưng cho các nhóm ý kiến đánh giá khác nhau về độ bền xe tay ga.

Hình 3.4a: Ý kiến của người tiêu dùng về độ bền của xe tay ga
(Nguồn: Khảo sát thị trường xe gắn máy tại TPHCM 1996-1998, Hoàng Trọng, 1998)



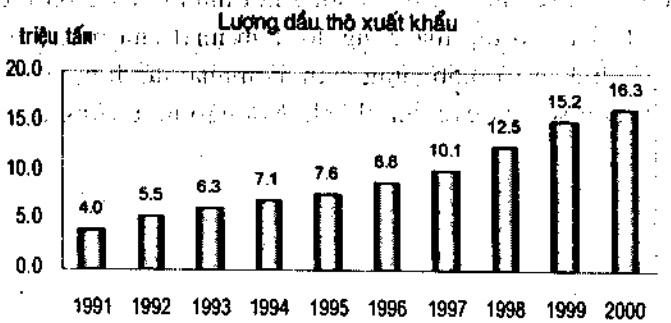
Hình 3.4b: Ý kiến của người tiêu dùng về độ bền của xe tay ga



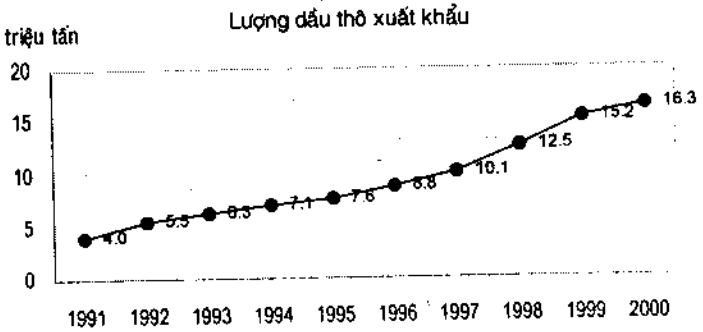
Hình 3.4b là biểu đồ thanh ngang biểu hiện cơ cấu, có cùng nội dung diễn tả như Hình 3.4a. Hình 3.4b giúp người xem dễ dàng so sánh và nhận thấy ý kiến người tiêu dùng ngày càng thay đổi theo hướng thuận lợi hơn cho tiêu thụ xe tay ga. Các biểu đồ trong Hình 3.4a và 3.4b có thể được tạo ra khá dễ dàng nhờ chương trình bảng tính Excel. Ví dụ này cho thấy các dạng biểu đồ khác nhau có thể sử dụng linh hoạt như thế nào trong từng trường hợp cụ thể.

Ví dụ thứ ba, các Hình 3.5a, 3.5b và 3.5c tuy có hình thức khác nhau nhưng có cùng một nội dung thể hiện là lượng dầu thô xuất khẩu của Việt Nam từ 1991 đến 2000. Trong trường hợp có quá nhiều cột thì có thể sử dụng biểu đồ đường gấp khúc thay cho biểu đồ hình cột để biểu đồ gọn dễ nhìn hơn.

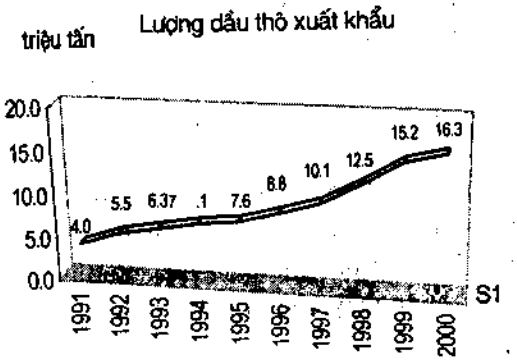
Hình 3.5a: Lượng dầu thô xuất khẩu của Việt Nam giai đoạn 1991-2000 (biểu đồ hình cột hay thanh đứng)



Hình 3.5b: Lượng dầu thô xuất khẩu của Việt Nam giai đoạn 1991-2000 (Biểu đồ đường gấp khúc)

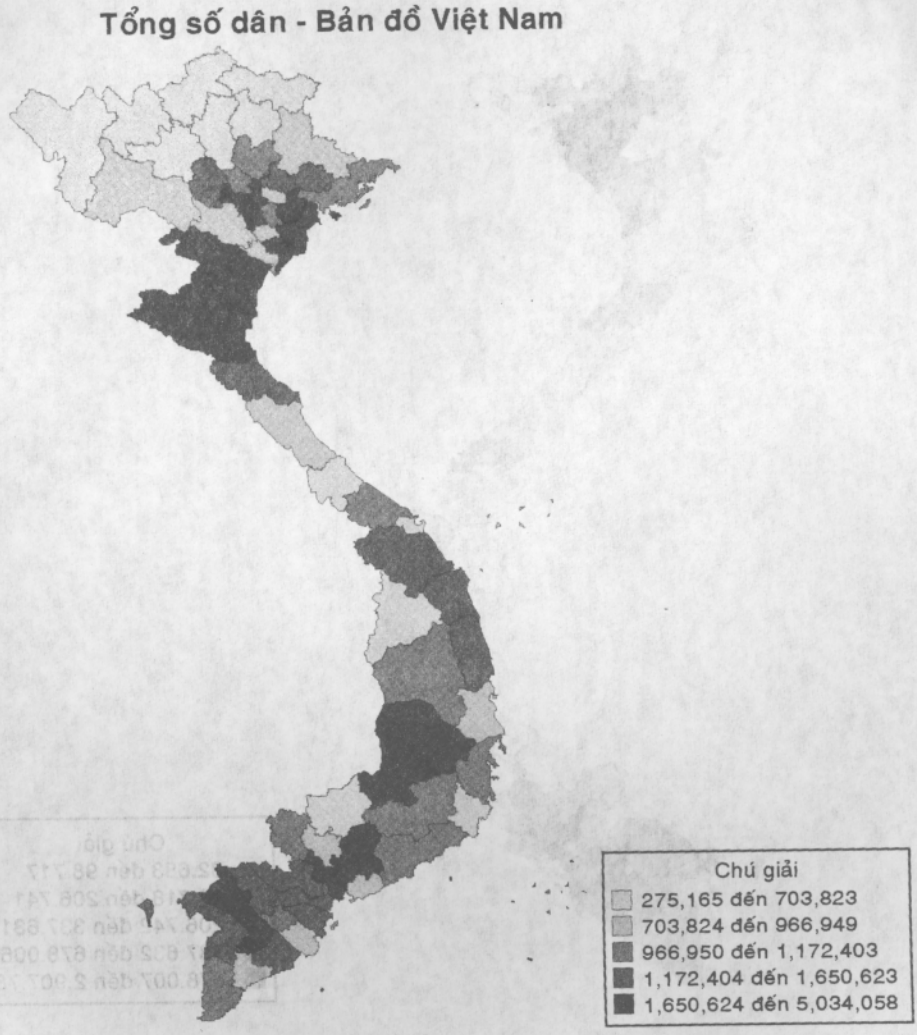


Hình 3.5c: Lượng dầu thô xuất khẩu của Việt Nam giai đoạn 1991-2000 (Biểu đồ đường gấp khúc giả lập 3 chiều thành biểu đồ mặt phẳng)



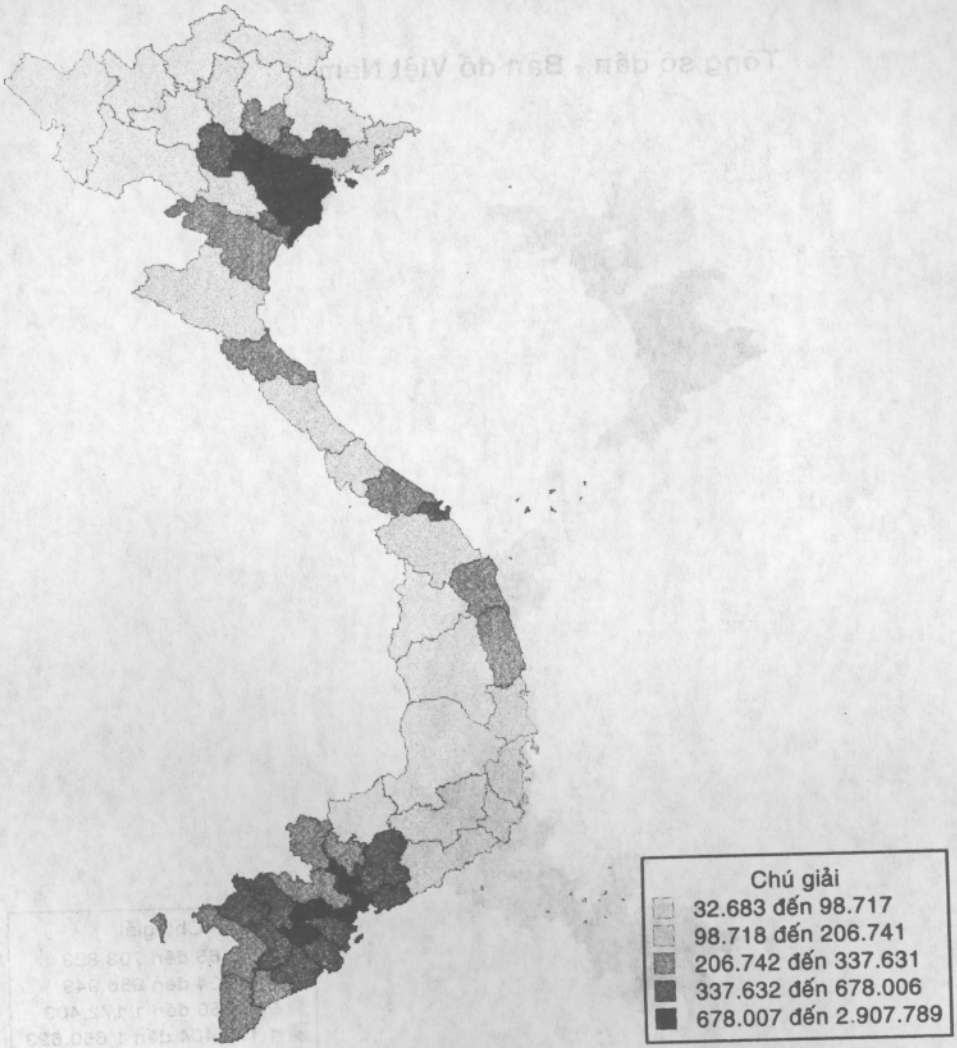
Ví dụ thứ tư là về bản đồ thống kê. Bản đồ thống kê không dùng quy ước về độ lớn của các hình vẽ hay đường nét hình học để biểu diễn mức độ như biểu đồ hay đồ thị, mà dùng độ đậm nhạt của màu sắc để biểu thị mức độ. Hình 3.5a và 3.5b dùng màu đậm nhạt để diễn ta quy mô dân số và mật độ dân số của các tỉnh thành. Nơi nào màu càng đậm thì mức độ càng lớn.

Hình 3.5a: Bản đồ dân số của các tỉnh/thành phố



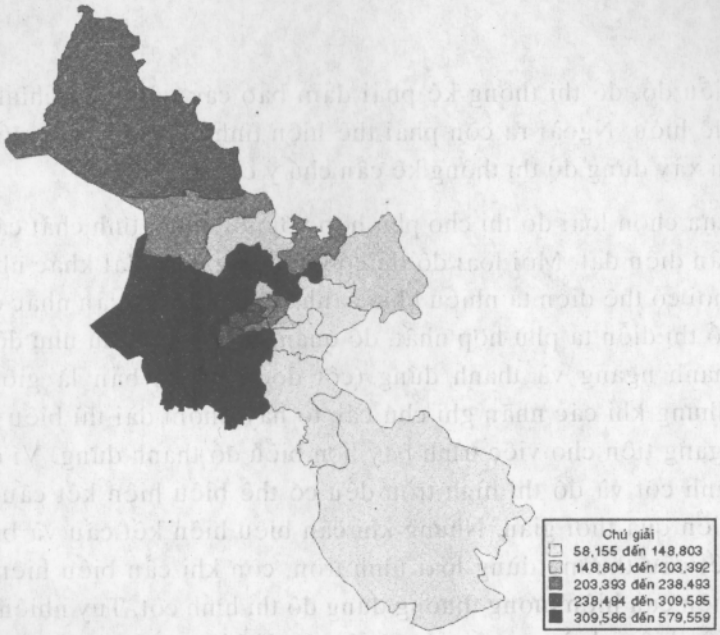
Hình 3.5b: Bản đồ mật độ dân số của các tỉnh/thành phố

Tổng số dân / Diện tích (Ratio): Bản đồ Việt Nam

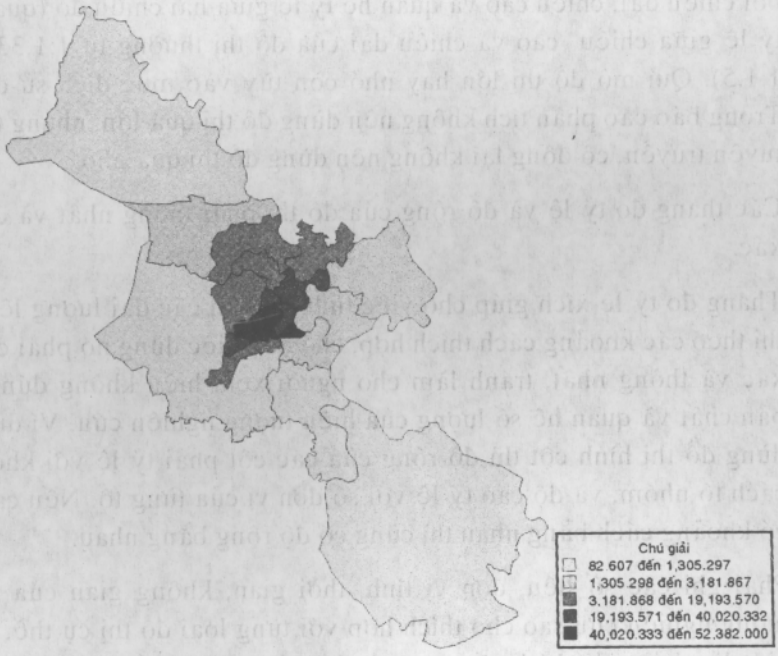


Hình 3.5b cho người xem dễ dàng hình dung là mật độ dân số cao ở hai khu vực châu thổ sông Hồng Và sông Cửu Long.Thưa thớt nhất là các tỉnh ở khu vực Tây Nguyên. Nếu trình bày dưới dạng bảng số liệu mật độ của 61 tỉnh thành phố thì có lẽ rất ít người có thể nắm bắt được những ý chính này nhanh chóng trước một "rừng" số liệu dày đặc. Tương tự, Hình 3.6 a và 3.6b diễn tả quy mô dân số và mật độ dân số của các quận huyện ở TP Hồ Chí Minh.

Hình 3.6a: Bản đồ dân số của các quận huyện ở TP Hồ Chí Minh



Hình 3.6b: Bản đồ mật độ dân số của các quận huyện ở TP Hồ Chí Minh



3.2.5.3. Những vấn đề cần chú ý khi xây dựng biểu đồ và đồ thị thống kê

Một biểu đồ, đồ thị thống kê phải đảm bảo các yêu cầu: chính xác, dễ xem, dễ hiểu. Ngoài ra còn phải thể hiện tính thẩm mỹ của đồ thị. Cho nên khi xây dựng đồ thị thống kê cần chú ý các điểm sau:

1. Lựa chọn loại đồ thị cho phù hợp với nội dung, tính chất các số liệu cần diễn đạt. Mỗi loại đồ thị có khả năng diễn đạt khác nhau, đồng thời có thể diễn tả nhiều khía cạnh. Cho nên cần cân nhắc chọn loại đồ thị diễn tả phù hợp nhất, dễ quan sát nhất. Ví dụ như đồ thị hình thanh ngang và thanh đứng (cột dọc), về cơ bản là giống nhau. Nhưng khi các nhãn ghi chú các tổ hay nhóm dài thì biểu đồ thanh ngang tiện cho việc trình bày hơn biểu đồ thanh đứng. Ví dụ đồ thị hình cột và đồ thị hình tròn đều có thể biểu hiện kết cấu, sự phát triển qua thời gian. Nhưng khi cần biểu hiện kết cấu và biến động kết cấu thường dùng loại hình tròn, còn khi cần biểu hiện sự phát triển của hiện tượng thường dùng đồ thị hình cột. Tuy nhiên, khi cần biểu hiện nhiều cơ cấu cùng một lúc thì biểu đồ hình thanh có ưu thế hơn, gọn, dễ so sánh(Hình 3.3a và Hình 3.3b)
2. Xác định qui mô đồ thị cho thích hợp. Qui mô đồ thị được quyết định bởi chiều dài, chiều cao và quan hệ tỷ lệ giữa hai chiều đó (quan hệ tỷ lệ giữa chiều cao và chiều dài của đồ thị thường từ 1:1,33 đến 1:1,5). Qui mô đồ thị lớn hay nhỏ còn tùy vào mục đích sử dụng. Trong báo cáo phân tích không nên dùng đồ thị quá lớn, nhưng trong tuyên truyền, cổ động lại không nên dùng đồ thị quá nhỏ.
3. Các thang đo tỷ lệ và độ rộng của đồ thị phải thống nhất và chính xác.
4. Thang đo tỷ lệ xích giúp cho việc tính chuyển các đại lượng lên đồ thị theo các khoảng cách thích hợp, cho nên việc dùng nó phải chính xác và thống nhất, tránh làm cho người xem hiểu không đúng về bản chất và quan hệ số lượng của hiện tượng nghiên cứu. Ví dụ khi dùng đồ thị hình cột thì độ rộng của các cột phải tỷ lệ với khoảng cách tổ nhóm, và độ cao tỷ lệ với số đơn vị của từng tổ. Nếu các tổ có khoảng cách bằng nhau thì cũng có độ rộng bằng nhau.
5. Phải ghi các số liệu, đơn vị tính, thời gian, không gian của hiện tượng nghiên cứu sao cho thích hợp với từng loại đồ thị cụ thể. Đặc biệt là phải giải thích rõ ràng các ký hiệu, màu sắc qui ước được

dùng trong đồ thị. Các quy ước về màu sắc thể hiện phải nhất quán trong toàn bộ các biểu đồ và đồ thị. Ví dụ khảo sát tại 3 địa bàn, khi diễn tả trên đồ thị quy ước về màu cho từng địa bàn phải giống nhau giữa các biểu đồ, đồ thị.

Trong thực tế, chương trình máy tính Excel được sử dụng rất phổ biến để thực hiện các biểu đồ hay đồ thị vì chương trình này rất phổ biến, dễ sử dụng, và liên kết tốt với chương trình xử lý văn bản Word hay chương trình trình diễn Power Point [2].

[2] Sinh viên có thể tham khảo cách dùng Excel để vẽ biểu đồ ở trang 48-55 sách "Xử lý dữ liệu nghiên cứu với SPSS for Windows", Hoàng Trọng, Nhà xuất bản Thống Kê, 2002.

CHƯƠNG 4

MÔ TẢ DỮ LIỆU BẰNG CÁC ĐẶC TRƯNG ĐO LƯỜNG

Mọi hiện tượng kinh tế - xã hội đều tồn tại trong những điều kiện không gian, thời gian nhất định và mặt lượng của hiện tượng được biểu hiện ở các mức độ khác nhau với nội dung phản ánh khác nhau. Ta có thể sử dụng các chỉ tiêu để đo lường, tính toán các mức độ này như số tuyệt đối, số tương đối, các đặc trưng đo lường khuynh hướng tập trung, các đặc trưng đo lường độ phân tán nhằm nêu lên đặc trưng phân phối của dãy số.

4.1. SỐ TUYỆT ĐỐI

4.1.1 Khái niệm

Số tuyệt đối là chỉ tiêu biểu hiện qui mô, khối lượng của hiện tượng kinh tế - xã hội trong điều kiện thời gian và địa điểm cụ thể.

Ví dụ: Tổng số dân của nước ta có lúc 0 giờ 1/4/1999 là 76.324.753 người.

Sản lượng lương thực (qui ra thóc) cả nước năm 1998 là 31,85 triệu tấn.

Số tuyệt đối có thể biểu hiện số đơn vị của tổng thể hay bộ phận, như số nhân khẩu, số công nhân, số học sinh, số xí nghiệp... hoặc là trị số của một chỉ tiêu kinh tế nào đó như sản lượng của nhà máy, tổng chi phí sản xuất, tổng mức tiền lương...

4.1.2 Các loại số tuyệt đối

Tùy theo tính chất của hiện tượng có thể phân biệt 2 loại số tuyệt đối:

4.1.2.1. Số tuyệt đối thời điểm

Số tuyệt đối thời điểm phản ánh qui mô, khối lượng của hiện tượng tại một thời điểm nhất định.

Ví dụ: Số lao động của một xí nghiệp có vào ngày 10/01/2003 là 120 người; giá trị TSCĐ của 1 đơn vị có vào ngày 12/02/2003 là 3 tỷ đồng, giá trị hàng hóa tồn kho của cửa hàng X vào ngày 1/1/2003 là 2 tỷ đồng.

Số tuyệt đối thời điểm chỉ phản ánh tình hình của hiện tượng tại một thời điểm nào đó, trước và sau thời điểm trạng thái của hiện tượng có thể khác.

Ta không thể cộng các số tuyệt đối thời điểm của cùng một chỉ tiêu nhưng ở những thời điểm khác nhau, vì không có ý nghĩa.

Ví dụ: Giá trị TSCĐ của Công ty A có vào ngày 1/9/2003 là 100 triệu đồng. Giá trị TSCĐ của Công ty A có vào ngày 2/9/2003 là 100 triệu đồng.

Ta không thể nói giá trị TSCĐ của Công ty A trong 2 ngày là 200 triệu đồng.

4.1.2.2. Số tuyệt đối thời kỳ

Số tuyệt đối thời kỳ phản ánh qui mô, khối lượng của hiện tượng trong một khoảng thời gian nhất định. Nó được hình thành thông qua sự tích lũy (cộng dồn) về lượng của hiện tượng trong suốt thời gian nghiên cứu.

Ví dụ: Số lượng sản phẩm sản xuất ra của 1 doanh nghiệp trong năm 2003 là 10.000 sản phẩm (cộng dồn số sản phẩm sản xuất ra trong 12 tháng của năm 2003 được 10.000 sản phẩm). Doanh số bán của 1 công ty trong năm 2003 là 150 tỷ đồng.

Các số tuyệt đối thời kỳ của cùng một chỉ tiêu có thể cộng được với nhau để có trị số của thời kỳ dài hơn. Thời kỳ tính toán càng dài, trị số của chỉ tiêu càng lớn.

4.1.3 Đơn vị tính của số tuyệt đối

4.1.3.1 Đơn vị hiện vật

Đơn vị hiện vật là đơn vị tính phù hợp với đặc điểm vật lý của hiện tượng. Nó được sử dụng rộng rãi khi xác định qui mô, khối lượng sản phẩm sản xuất và tiêu dùng.

Đơn vị hiện vật bao gồm:

- Đơn vị hiện vật tự nhiên: người, cái, chiếc, con...

- Đơn vị hiện vật qui ước: Kg, tạ, tấn, lít, mét, phút, giờ, ngày, tháng, năm... Đây là đơn vị tính cơ bản của sản phẩm, phản ánh chính xác theo giá trị sử dụng của sản phẩm. Tuy nhiên, nó cũng có nhược điểm là không tổng hợp được các sản phẩm khác loại và những công việc có tính chất dịch vụ.

Để khắc phục một phần nhược điểm của đơn vị hiện vật tự nhiên, người ta sử dụng đơn vị hiện vật qui đổi.

- Đơn vị hiện vật qui đổi:

Sử dụng đơn vị hiện vật qui đổi là chọn một sản phẩm làm gốc rồi qui đổi các sản phẩm khác cùng tên nhưng có qui cách, phẩm chất khác nhau ra sản phẩm đó theo một hệ số qui đổi.

Ví dụ: Sản lượng lương thực (qui thóc) của tỉnh X năm 2000 là 3 triệu tấn.

Ở đây thóc là sản phẩm qui đổi để qui các loại cây lương thực khác (bắp, khoai) ra thóc theo hệ số qui đổi được Nhà nước qui định là:

1 kg thóc = 1 kg bắp hạt = 3 kg khoai tươi.

Cơ sở để xác định hệ số qui đổi là căn cứ vào giá trị sử dụng (công dụng chính) của sản phẩm. Trong thực tế đôi khi người ta lại dùng giá trị sản phẩm (giá cả) để làm cơ sở xác định hệ số qui đổi.

Ý nghĩa của loại đơn vị tính này là dùng để tổng hợp các sản phẩm cùng loại nhưng có qui cách, phẩm chất khác nhau. Tuy nhiên, nó vẫn không thể tổng hợp được tất cả các loại sản phẩm khác tên. Ngoài ra, đơn vị hiện vật qui đổi phản ánh lượng giá trị sử dụng tương đương, không phản ánh lượng giá trị sử dụng thực tế. Do vậy, nó có tính trừu tượng, giảm tính cụ thể của đơn vị hiện vật tự nhiên.

4.1.3.2. Đơn vị tiền tệ

Đơn vị tiền tệ (đồng, rúp, đô la...) được sử dụng để biểu hiện giá trị sản phẩm. Nó giúp cho việc tổng hợp nhiều loại sản phẩm có giá trị sử dụng và đơn vị đo lường khác nhau. Tuy nhiên do giá cả hàng hoá luôn thay đổi, đơn vị tiền trở nên không có tính chất so sánh được qua thời gian.

Ví dụ: có kết quả sản xuất thành phẩm của một nhà máy dệt qua 2 tháng trong năm như sau:

Tháng 5/2002	Số lượng:	100.000m vải
	Đơn giá bán:	50.000 đ/m

Giá trị khối lượng thành phẩm: 5 tỷ đồng

Tháng 6/2002	Số lượng:	100.000m vải
	Đơn giá bán:	60.000 đ/m

Giá trị khối lượng thành phẩm: 6 tỷ đồng.

Rõ ràng cùng số lượng nhưng giá cả khác nhau làm cho giá trị khối lượng khác nhau.

Để khắc phục nhược điểm do ảnh hưởng của thay đổi giá cả như ví dụ trên, người ta dùng giá so sánh hay giá cố định. Ở nước ta, đã có các bảng giá cố định 1970, 1989, 1994. Tuy nhiên hiện nay, người ta có xu hướng bỏ giá cố định và thay bằng việc tính chỉ số lạm phát giá cả để loại trừ ảnh hưởng của giá.

4.1.3.3. Đơn vị thời gian lao động

Đơn vị thời gian lao động như giờ công, ngày công... thường dùng để tính lượng lao động hao phí để sản xuất ra những sản phẩm không thể tổng hợp hoặc so sánh với nhau bằng các đơn vị tính toán khác, hoặc những sản phẩm phức tạp do nhiều người thực hiện qua nhiều giai đoạn khác nhau.

Đơn vị này dùng nhiều trong công tác định mức thời gian cho sản xuất, tính năng suất lao động, quản lý thời gian lao động của đơn vị...

4.2 SỐ TƯƠNG ĐỐI

4.2.1 Khái niệm

Số tương đối trong thống kê là chỉ tiêu biểu hiện quan hệ so sánh giữa hai mức độ của hiện tượng nghiên cứu.

Ta thấy có 2 trường hợp so sánh:

* So sánh hai mức độ cùng loại nhưng khác nhau về thời gian hoặc không gian.

Ví dụ:

$$\frac{\text{Doanh số bán của Công ty A năm 2002}}{\text{Doanh số bán của Công ty A năm 2001}}$$

hoặc

$$\frac{\text{Doanh số bán của Công ty A năm 2002}}{\text{Doanh số bán của Công ty B năm 2002}}$$

* So sánh hai hiện tượng khác loại nhưng có liên quan nhau.

Ví dụ:

$$\text{Mật độ dân số (người/Km}^2\text{)} = \frac{\text{Dân số trung bình}}{\text{Diện tích đất đai}}$$

$$\text{GDP trung bình đầu người (đ/người)} = \frac{\text{GDP}}{\text{Dân số trung bình}}$$

Như vậy, số tương đối là kết quả so sánh hai số tuyệt đối, hai số tương đối hay hai số trung bình với nhau. Nó được sử dụng tùy theo mục đích nghiên cứu chủ quan của người xử lý số liệu. Chẳng hạn như muốn đánh giá sự hơn kém của hai mức độ, đánh giá mức độ hoàn thành kế hoạch của một doanh nghiệp, nghiên cứu cơ cấu của một ngành, cơ cấu doanh thu... Số tương đối còn được sử dụng để công bố khi muốn giữ bí mật của số tuyệt đối.

Hình thức biểu hiện số tương đối là số lần, phần trăm (%), phần nghìn (‰)... hoặc bằng đơn vị đo lường ghép (người/km^2, sản phẩm/người . . .)

4.2.2 Các loại số tương đối:

Căn cứ vào nội dung phản ánh, có thể chia số tương đối thành các loại sau.

4.2.2.1 Số tương đối động thái

Số tương đối động thái (hay tốc độ phát triển) là kết quả so sánh giữa hai mức độ của cùng hiện tượng nhưng khác nhau về thời gian.

Công thức: $$t = \frac{y_1}{y_0} \qquad (4.1)$$

Trong đó: t : số tương đối động thái.

y_0 : mức độ của hiện tượng kỳ gốc

y_1 : mức độ của hiện tượng kỳ nghiên cứu (kỳ báo cáo)

Ví dụ: Sản phẩm sản xuất của xí nghiệp A qua hai năm như sau:

Năm 2001 sản xuất 100 tấn. Năm 2002 sản xuất 150 tấn.

Số tương đối động thái là: $t = \frac{y_1}{y_0} = \frac{150}{100} = 1,5 = 150\%$

Vậy sản phẩm sản xuất của xí nghiệp A năm 2002 so với năm 2001 bằng 1,5 lần hay bằng 150%, tức là tăng lên 50%.

* Nếu ta tính các số tương đối động thái với kỳ gốc y_0 thay đổi và kề ngay trước kỳ báo cáo, ta có các số tương đối động thái liên hoàn (hay tốc độ phát triển liên hoàn)

Ví dụ: Có tài liệu về doanh số bán hàng của một Công ty X qua các năm như sau:

Bảng 4.1

Năm	1999	2000	2001	2002
Doanh số bán (tỷ đồng)	10,00	12,00	14,40	15,84

Các số tương đối động thái liên hoàn về doanh số của Công ty:

$$\frac{y_{2000}}{y_{1999}} = \frac{12}{10} = 1,2 = 120\%$$

$$\frac{y_{2001}}{y_{2000}} = \frac{14,4}{12} = 1,2 = 120\%$$

$$\frac{y_{2002}}{y_{2001}} = \frac{15,84}{14,4} = 1,1 = 110\%$$

* Nếu ta tính các số tương đối động thái với kỳ gốc y_0 cố định, ta có các số tương đối động thái định gốc (hay tốc độ phát triển định gốc).

Giả sử ở đây ta chọn kỳ gốc cố định là doanh số bán năm 1999.

Các số tương đối động thái định gốc về doanh số bán của Công ty:

$$\frac{y_{2000}}{y_{1999}} = \frac{12}{10} = 1,2 = 120\%$$

$$\frac{y_{2001}}{y_{1999}} = \frac{14,4}{10} = 1,44 = 144\%$$

$$\frac{y_{2002}}{y_{1999}} = \frac{15,84}{10} = 1,584 = 158,4\%$$

* Muốn tính số tương đối động thái chính xác cần bảo đảm tính chất so sánh được giữa các mức độ kỳ nghiên cứu và kỳ gốc. Cụ thể phải đảm bảo giống nhau về nội dung kinh tế, về phương pháp tính và đơn vị tính, về phạm vi và độ dài thời gian mà mức độ phản ánh.

4.2.2.2 Số tương đối kế hoạch

Số tương đối kế hoạch bao gồm hai loại: Số tương đối nhiệm vụ kế hoạch và số tương đối hoàn thành kế hoạch.

*** Số tương đối nhiệm vụ kế hoạch:**

Số tương đối nhiệm vụ kế hoạch là tỷ lệ so sánh giữa mức độ kế hoạch với mức độ thực tế của chỉ tiêu ấy ở kỳ gốc.

Công thức: $$t_{NK} = \frac{y_k}{y_0} \qquad (4.2)$$

Với: t_{NK} : số tương đối nhiệm vụ kế hoạch

y_k : mức độ kế hoạch

*** Số tương đối hoàn thành kế hoạch**

Số tương đối hoàn thành kế hoạch là tỷ lệ so sánh giữa mức độ thực tế đạt được trong kỳ nghiên cứu với mức độ kế hoạch đặt ra cùng kỳ của một chỉ tiêu nào đó.

Công thức: $$t_{HK} = \frac{y_1}{y_k} \quad (4.3)$$

Ví dụ : Sản lượng lúa của huyện Y năm 2001 là 250.000 tấn, kế hoạch dự kiến sản lượng lúa năm 2002 là 300.000 tấn , thực tế năm 2002 huyện Y đạt được 330.000 tấn. Như vậy ta có:

a. Số tương đối động thái:

$$t = \frac{y_1}{y_0} = \frac{330.000}{250.000} = 1,32 = 132\%$$

b. Số tương đối nhiệm vụ kế hoạch:

$$t_{NK} = \frac{y_k}{y_0} = \frac{300.000}{250.000} = 1,2 = 120\%$$

c. Số tương đối hoàn thành kế hoạch

$$t_{HK} = \frac{y_1}{y_k} = \frac{330.000}{300.000} = 1,1 = 110\%$$

Ta có mối quan hệ : $\frac{y_1}{y_0} = \frac{y_k}{y_0} \times \frac{y_1}{y_k}$

$$1,32 = 1,2 \times 1,1$$

4.2.2.3 Số tương đối kết cấu

Số tương đối kết cấu xác định tỷ trọng của mỗi bộ phận cấu thành tổng thể.

Công thức: $$d_i = \frac{y_i}{\sum_{i=1}^{n} y_i} \quad (4.4)$$

Với: d_i : tỷ trọng của bộ phận thứ i

y_i : mức độ của bộ phận thứ i

$\sum_{i=1}^{n} y_i$: tổng các mức độ của tổng thể

Qua chỉ tiêu này có thể phân tích được đặc điểm cấu thành của hiện tượng. Nghiên cứu sự thay đổi kết cấu sẽ thấy được xu hướng phát triển của hiện tượng.

4.2.2.4 Số tương đối cường độ

Số tương đối cường độ là kết quả so sánh mức độ của hai hiện tượng khác nhau nhưng có liên quan với nhau.

Ví dụ: Mật độ dân số, GDP tính trên đầu người, số bác sĩ trên 1.000 dân...

Số tương đối cường độ phản ánh trình độ phổ biến của hiện tượng, nó được sử dụng rộng rãi để biểu hiện trình độ phát triển sản xuất, trình độ bảo đảm mức sống vật chất, văn hóa của dân cư một nước. Các chỉ tiêu này thường dùng để so sánh trình độ phát triển sản xuất, đời sống giữa các nước khác nhau.

4.2.2.5 Số tương đối không gian

Số tương đối không gian là kết quả so sánh giữa hai mức độ của một hiện tượng nhưng khác nhau về không gian.

Ví dụ: Giá trị sản xuất công nghiệp của Công ty xi măng Hà Tiên II năm 2002 so với giá trị sản xuất công nghiệp của Công ty luyện cán thép Miền Nam năm 2002.

Số tương đối không gian cũng biểu hiện sự so sánh giữa hai bộ phận trong cùng một tổng thể.

Ví dụ: So sánh số lao động nữ với số lao động nam; số lao động gián tiếp với số lao động trực tiếp trong một đơn vị.

4.3 CÁC ĐẶC TRƯNG ĐO LƯỜNG KHUYNH HƯỚNG TẬP TRUNG[1]

Các đặc trưng đo lường khuynh hướng tập trung như số trung bình, số trung vị, mốt là các chỉ tiêu biểu hiện mức độ đại diện theo một tiêu thức số lượng nào đó của một tổng thể bao gồm nhiều đơn vị cùng loại.

---

[1] Measures of central tendency

4.3.1 Số trung bình cộng (Số trung bình số học [1])

Số trung bình cộng được tính bằng cách đem chia tổng tất cả các trị số của các đơn vị cho số đơn vị tổng thể.

* Số trung bình cộng tính từ tổng thể chung[2]

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (4.5)$$

Trong đó: μ : Trung bình của tổng thể.

x_i : i=1,2,3...,N: lượng biến thứ i.

N : Số đơn vị tổng thể.

* Số trung bình cộng tính từ mẫu[3]

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (4.6)$$

Trong đó: $\bar{x}$: Trung bình của mẫu.

x_i : lượng biến thứ i (i = 1,2,...,n)

n : Cỡ mẫu (tổng số đơn vị của mẫu).

Ví dụ: Tiền lương của 4 công nhân trong tháng 4/2003 của phân xưởng A như sau (1000đ): 1.370, 1.400, 1.420 và 1.500.

Vậy tiền lương trung bình của một công nhân trong tháng 4/2003 là:

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i}{4} = \frac{1370+1400+1420+1500}{4} = 1422{,}5 \text{ ngàn đồng/người}$$

---

[1] The arithmetic mean

[2] Population mean

[3] Sample mean

4.3.2 Số trung bình cộng gia quyền

Số trung bình cộng gia quyền (còn gọi là số trung bình số học có trọng số [1])

áp dụng khi mỗi lượng biến được gặp nhiều lần, nghĩa là có các tần số f_i khác nhau.

* Đối với tổng thể:

$$\mu = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} \qquad (4.7)$$

Trong đó: $\sum_{i=1}^{k} f_i = N$

x_i : lượng biến thứ i (i = 1,2,...,k)

f_i : Tần số của tổ i (i = 1,2,3,...,k)

* Đối với mẫu:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} \qquad (4.8)$$

Trong đó: $\sum_{i=1}^{k} f_i = n$

x_i : lượng biến thứ i (i = 1,2,...,k)

f_i : Tần số của tổ i (i = 1,2,3,...,k)

Ví dụ: Có tình hình tiền lương trong tháng 4/2003 của công nhân thuộc một phân xưởng sản xuất gồm các mức lương sau:

[1] Weighted mean

Bảng 4.2

Tiền lương (1000 đồng)(x_i)	Số công nhân (f_i)	$x_i \times f_i$
1370	10	13700
1400	15	21000
1420	12	17040
1500	3	4500
Tổng	40	56240

Để tính tiền lương trung bình 1 công nhân, ta lập thêm cột $x_i f_i$

Vậy tiền lương trung bình của công nhân trong phân xưởng là:

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i f_i}{\sum_{i=1}^{4} f_i} = \frac{56240}{40} = 1406 \text{ ngàn đồng / người}$$

Trong công thức số trung bình cộng gia quyền, ta lấy $\sum x_i f_i$ chính là tổng tất cả các trị số của các đơn vị chia cho $\sum f_i$ là tổng số đơn vị. Như vậy xét theo nội dung của số trung bình cộng, thì hai công thức (4.6) và (4.8) thực ra chỉ là một.

* Tính số trung bình cộng trong trường hợp tài liệu phân tổ có khoảng cách tổ: Trong trường hợp này mỗi tổ có một phạm vi trị số, ta lấy trị số giữa của mỗi tổ làm lượng biến đại diện cho tổ đó.

Trị số giữa mỗi tổ [1] $x_i = \frac{x_{min} + x_{max}}{2}$

Trong đó: x_{min}, x_{max} là giới hạn dưới và giới hạn trên của từng tổ.

Ví dụ: Tính năng suất lao động trung bình của công nhân theo tài liệu sau:

[1] Midpoint

Bảng 4.3

Năng suất lao động (kg)	Trị số giữa (x_i)	Số công nhân (f_i)	$x_i \times f_i$
200 – 300	250	10	2.500
300 – 400	350	32	11.200
400 – 500	450	43	19.350
500 – 600	550	78	42.900
600 – 700	650	32	20.800
700 - 800	750	5	3.750
Cộng		200	100.500

Năng suất lao động trung bình của công nhân trong xí nghiệp:

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i f_i}{\sum_{i=1}^{6} f_i} = \frac{100.500}{200} = 502,5 \text{ kg}$$

Trị số giữa của tổ 1 bằng 250, có nghĩa là chúng ta coi như cả 10 công nhân trong tổ 1 đều có cùng một mức năng suất là 250 kg, điều này chắc chắn sẽ sai lệch so với năng suất thật của 10 người trong thực tế. Liệu việc giả định như trên có làm cho số trung bình tính ra có sai số lớn? Thực ra theo luật bù trừ của số lớn, ví dụ như trong tổ 1 có người có mức năng suất nhỏ hơn 250kg, nhưng cũng có người có mức năng suất lớn hơn 250kg, các tổ khác cũng tương tự như vậy. Bù trừ qua lại ta có thể xem trị số giữa của mỗi tổ là lượng biến trung bình của tổ đó, do đó năng suất trung bình của công nhân trong xí nghiệp tính ra vẫn có ý nghĩa.

Ta dễ dàng chứng minh các tính chất sau:

* Nếu ta có : $f_1 = f_2 = ... = f_k = f$

Thì $\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{\sum x_i}{k}$ (k là số tổ)

* Nếu đặt $d_i = \frac{f_i}{\sum f_i} \Rightarrow \bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \sum x_i d_i$

* $\sum (x_i - \bar{x}) = 0$

* $\bar{x} = \frac{\sum \bar{x}_i f_i}{\sum f_i}$ Trong đó $\bar{x}_i$ là số trung bình của tổ thứ i

4.3.3 Số trung bình điều hòa[1]

Xét về nội dung kinh tế, số trung bình điều hòa cũng giống như số trung bình cộng. Tuy nhiên trong trường hợp này ta không có tài liệu về số đơn vị tổng thể f_i, mà chỉ có tài liệu về các lượng biến x_i và $M_i = x_i \times f_i$

Công thức: $$\bar{x} = \frac{M_1 + M_2 + M_3 + ... + M_n}{\frac{M_1}{x_1} + \frac{M_2}{x_2} + \frac{M_3}{x_3} + ... + \frac{M_n}{x_n}} = \frac{\sum M_i}{\sum \frac{M_i}{x_i}} \quad (4.9)$$

Nếu ta có: $M_1 = M_2 = M_3 = ... = M_n = M$

Thì : $$\bar{x} = \frac{\sum M_i}{\sum \frac{M_i}{x_i}} = \frac{\sum M}{\sum \frac{M}{x_i}} = \frac{nM}{M\sum \frac{1}{x_i}} = \frac{n}{\sum \frac{1}{x_i}} \quad (4.10)$$

Ví dụ: Có tình hình về doanh số bán của 3 loại gạo tại một cửa hàng gạo như sau:

Bảng 4.4

Loại gạo	Đơn giá (1000đ/kg)	Doanh thu (1000 đồng)
Loại 1	8	24.000
Loại 2	6	24.000
Loại 3	4	24.000

Tính giá trung bình 1kg gạo mà cửa hàng đã bán ra?

Ta áp dụng công thức: $\bar{x} = \frac{\sum M_i}{\sum \frac{M_i}{x_i}}$

[1] Harmonic mean

Trong đó: x_i : Giá gạo loại i

M_i : Doanh thu gạo loại i

Vậy giá trung bình 1kg gạo mà cửa hàng đã bán ra:

$$\bar{x} = \frac{24.000 + 24.000 + 24.000}{\frac{24.000}{8} + \frac{24.000}{6} + \frac{24.000}{4}} = 5,538 \text{ ngàn đồng/kg}$$

Trong ví dụ trên ta thấy: $M_1 = M_2 = M_3 = 24.000$ ngàn đồng

Nên ta có thể tính theo công thức: $\bar{x} = \frac{n}{\sum \frac{1}{x_i}}$

Vậy giá trung bình 1kg gạo mà cửa hàng đã bán ra:

$$\bar{x} = \frac{3}{\frac{1}{8} + \frac{1}{6} + \frac{1}{4}} = 5,538 \text{ ngàn đồng/kg}$$

Thực ra là ta đã đơn giản từ công thức:

$$\bar{x} = \frac{24.000 + 24.000 + 24.000}{\frac{24.000}{8} + \frac{24.000}{6} + \frac{24.000}{4}} = \frac{24.000(1+1+1)}{24.000(\frac{1}{8} + \frac{1}{6} + \frac{1}{4})} = \frac{3}{\frac{1}{8} + \frac{1}{6} + \frac{1}{4}} = 5,538$$

Xét theo nội dung công thức số trung bình điều hòa, tử số $\sum M_i$ chính là tổng tất cả các trị số của các đơn vị. Trong ví dụ trên nó là tổng doanh thu của 3 loại gạo. Còn mẫu số $\sum \frac{M_i}{x_i}$ là tổng số đơn vị. Trong ví dụ trên nó là tổng lượng gạo 3 loại bán ra.

Nếu thay $M_i = x_i \times f_i$ vào công thức (4.9) ta có:

$$\bar{x} = \frac{\sum M_i}{\sum \frac{M_i}{x_i}} = \frac{\sum x_i f_i}{\sum \frac{x_i f_i}{x_i}} = \frac{\sum x_i f_i}{\sum f_i}$$

Nghĩa là số trung bình điều hòa thực chất là số trung bình cộng.

4.3.4 Số trung bình nhân[1] (Số trung bình hình học)

Số trung bình nhân, còn được gọi là tốc độ phát triển trung bình, được tính từ những lượng biến có quan hệ tích số.

$$\bar{t} = \sqrt[m]{t_1 t_2 t_3 \ldots t_m} = \sqrt[m]{\prod_{i=1}^{m} t_i} \qquad (4.11)$$

Với: t_i : tốc độ phát triển liên hoàn thứ i (số tương đối động thái liên hoàn)

m: Số tốc độ phát triển liên hoàn

Trong trường hợp các lượng biến t_i được gặp nhiều lần, nghĩa là có các tần số f_i khác nhau, công thức 4.11 được viết gọn:

$$\bar{t} = \sqrt[\sum_{i=1}^{k} f_i]{t_1^{f_1} \times t_2^{f_2} \times \ldots \times t_k^{f_k}} = \sqrt[\sum_{i=1}^{k} f_i]{\prod_{i=1}^{k} t_i^{f_i}} \qquad (4.12)$$

Với: $$\sum_{i=1}^{k} f_i = m$$

Ví dụ: Có số liệu về doanh thu của một Công ty thương mại từ năm 1997 đến năm 2002

Bảng 4.5

Năm	1997	1998	1999	2000	2001	2002
Doanh thu (tỷ đồng)	200	210	215	222	230	244
Tốc độ phát triển liên hoàn (t_i) (lần)		1,05	1,02	1,03	1,04	1,06

Tốc độ phát triển trung bình về doanh thu của công ty trong giai đoạn 1997-2002:

$$\bar{t} = \sqrt[5]{t_1 t_2 t_3 t_4 t_5} = \sqrt[5]{1{,}05.1{,}02.1{,}03.1{,}04.1{,}06} = 1{,}04 = 104\%$$

Giả sử như trong công thức trên ta có $t_1 = t_2 = 1{,}05$; $t_3 = t_4 = 1{,}04$; $t_5 = 1{,}06$

Tốc độ phát triển trung bình về doanh thu của công ty trong giai đoạn 1997-

[1] Geometric mean

2002 được tính theo công thức 4.12:

$$\bar{t} = \sqrt[5]{t_1 t_2 t_3 t_4 t_5} = \sqrt[5]{1,05.1,05.1,04.1,04.1,06}$$

$$= \sqrt[5]{1,05^2.1,04^2.1,06} = 1,05 = 105\%$$

Trở lại ví dụ trên, ta có tốc độ phát triển định gốc bằng tích các tốc độ phát triển liên hoàn, nghĩa là:

$$1,05.1,02.1,03.1,04.1,06 = \frac{244}{200}$$

Như vậy ta có thể tính nhanh tốc độ phát triển trung bình:

$$\bar{t} = \sqrt[6-1]{\frac{244}{200}} = 1,04 = 104\%$$

Và ta có công thức khác để tính tốc độ phát triển trung bình:

$$\bar{t} = \sqrt[n-1]{\frac{y_n}{y_1}} \qquad (4.13)$$

Trong đó: y_1, y_n là mức độ đầu tiên và mức độ cuối cùng trong dãy số

n : số mức độ

4.3.5 Mốt [1](Mo)

4.3.5.1 Khái niệm

Mốt là biểu hiện của một tiêu thức được gặp nhiều nhất trong tổng thể. Đối với một dãy số lượng biến, mốt là lượng biến có tần số lớn nhất.

4.3.5.2 Cách xác định mốt

Ta chia ra các trường hợp sau:

- Tài liệu phân tổ không có khoảng cách tổ: Mốt là lượng biến có tần số lớn nhất

[1] Mode

Ví dụ: Điểm môn Lý thuyết thống kê của một lớp như sau:

Điểm số	Số sinh viên
4	10
5	15
6	30
7	52
8	15
9	2
Tổng	124

Ta nhanh chóng xác định mốt: $M_o = 7$ điểm

- Tài liệu phân tổ có khoảng cách tổ đều: Trước hết cần xác định tổ chứa mốt, tức là tổ có tần số lớn nhất, sau đó trị số gần đúng của mốt được xác định theo công thức:

$$M_o = x_{Mo(min)} + h_{Mo}\frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} \quad (4.14)$$

Trong đó:

$x_{Mo(min)}$: giới hạn dưới của tổ chứa mốt.

h_{Mo} : Trị số khoảng cách tổ của tổ chứa mốt.

f_{Mo} : Tần số của tổ chứa mốt.

f_{Mo-1} : Tần số của tổ đứng trước tổ chứa mốt.

f_{Mo+1}: Tần số của tổ đứng sau tổ chứa mốt.

Ví dụ: Có tài liệu tổng hợp về doanh số bán của 50 trạm xăng dầu thuộc tỉnh X trong tháng 12/2002 như sau:

Bảng 4.6

Doanh số bán (triệu đồng)	Số trạm
200 – 300	8
300 – 400	10
400 – 500	20
500 – 600	7
600 - 700	5
Tổng	50

Theo tài liệu ở bảng trên, ta xác định mốt ở vào tổ thứ 3 (400 – 500). Áp dụng công thức (4.14) ta tính mốt:

$$M_o = 400 + 100\frac{20-10}{(20-10)+(20-7)} = 443{,}48 \text{ triệu đồng}$$

Như vậy, đa số các trạm xăng dầu của tỉnh trên có mức doanh số trong tháng 12/1998 khoảng 443,48 triệu đồng.

- Tài liệu phân tổ có khoảng cách tổ không đều: Mốt vẫn được tính theo công thức (4.14) nhưng việc xác định tổ có mốt không căn cứ vào tần số, mà căn cứ vào mật độ phân phối (tỷ số giữa các tần số với khoảng cách tổ tương ứng)

Ví dụ: Có tài liệu về doanh thu của 79 cửa hàng trong tháng 12/2002 như sau: Bảng 4.7

Doanh thu (tr.đ)	Cửa hàng (f_i)	Khoảng cách tổ (h_i)	Mật độ phân phối $d_i = \left(\frac{f_i}{h_i}\right)$
200 – 400	8	200	0,04
400 – 500	12	100	0,12
500 – 600	25	100	0,25
600 – 800	25	200	0,125
800 – 1000	9	200	0,045
Tổng	79		

Theo tài liệu bảng trên, ta xác định mốt ở vào tổ 3 (500 – 600) vì có mật độ

phân phối lớn nhất là 0,25. Mốt được xác định như sau:

$$M_o = 500 + 100\frac{0,25 - 0,12}{(0,25 - 0,12) + (0,25 - 0,125)} = 550,9 \text{ triệu đồng}$$

Như vậy, đa số các cửa hàng có mức doanh thu trong tháng 12/2002 khoảng 550,9 trđ.

* Mốt có ưu điểm là không chịu ảnh hưởng của các lượng biến đột xuất, nhưng cũng chính điều này làm cho mốt kém nhạy bén với sự biến thiên của tiêu thức.Trong thực tế mốt được sử dụng ít hơn số trung vị và số trung bình. Có lẽ ứng dụng rõ ràng nhất của mốt là để nghiên cứu nhu cầu của thị trường về một loại kích cỡ sản phẩm nào đó như giày dép, mũ nón, quần áo...Mốt cho biết đa số, khuynh hướng phong trào. Có trường hợp không có mốt vì không có lượng biến nào xuất hiện nhiều nhất, hoặc cũng có trường hợp có 2 mốt, trong đó có mốt chính và mốt phụ.

4.3.6 Số trung vị [1] (Me)

4.3.6.1 Khái niệm

Số trung vị là lượng biến của đơn vị đứng ở vị trí giữa trong dãy số lượng biến đã được sắp xếp theo thứ tự tăng dần. Số trung vị chia dãy số làm hai phần, mỗi phần có số đơn vị tổng thể bằng nhau

4.3.6.2 cách xác định số trung vị

Ta chia ra các trường hợp sau:

- Tài liệu không phân tổ

- Trường hợp n lẻ, thì số trung vị sẽ là lượng biến đứng ở giữa dãy số, tức là đứng ở vị trí thứ $\frac{n+1}{2}$

- Trường hợp n chẵn, số trung vị sẽ là trung bình cộng của hai lượng biến đứng ở giữa, tức là hai lượng biến đứng ở vị trí thứ $\frac{n}{2}$ và $\frac{n+2}{2}$.

Ví dụ: Tiền lương tháng của công nhân ở một tổ sản xuất như sau (1000đ)

* dãy số lẻ: (n=7) 1500; 1600; 1750; 1900; 2150; 2300; 2500.

---

[1] Median

$$Me = x_{(n+1/2)} = x_4 = 1900 \text{ ngàn đồng}$$

* dãy số chẵn: (n=8) 1500; 1600; 1750; 1900; 2150; 2300; 2500; 2800.

$$Me = (x_4 + x_5)/2 = (1900 + 2150)/2 = 2025 \text{ ngàn đồng}$$

- Tài liệu phân tổ có khoảng cách tổ : Trước tiên ta xác định tổ chứa số trung vị, đó là tổ có tần số tích lũy lớn hơn hoặc bằng $\frac{\sum f_i + 1}{2}$. Sau đó trị số gần đúng của số trung vị được tính theo công thức:

$$Me = x_{Me(min)} + h_{Me} \frac{\frac{\sum f_i}{2} - S_{Me-1}}{f_{Me}} \qquad (4.15)$$

Trong đó:

$X_{Me(min)}$: Giới hạn dưới của tổ có số trung vị.

h_{Me} : Trị số khoảng cách tổ có số trung vị.

S_{Me-1} : Tổng các tần số của các tổ đứng trước tổ có số trung vị.

f_{Me} : Tần số của tổ có số trung vị.

$\sum f_i$: Tổng các tần số.

Trở lại ví dụ trong bảng 4.7, ta xác định số trung vị về doanh thu

Bảng 4.8

Doanh thu (tr.đ)	Cửa hàng (f_i)	Tần số tích lũy
200 – 400	8	8
400 – 500	12	20
500 – 600	25	45
600 – 800	25	70
800 – 1000	9	
Tổng	79	

Tổ có chứa số trung vị là tổ 3 (500 – 600) vì có tần số tích lũy bằng 45 >

(79 +1)/2. Thay số vào công thức 4.15 ta có:

$$Me = 500 + 100\frac{\frac{79}{2} - 20}{25} = 578 \text{ triệu đồng}$$

* Cũng như mốt, số trung vị biểu hiện mức độ đại biểu của hiện tượng mà không san bằng bù trừ chênh lệch giữa các lượng biến. Số trung vị có thể dùng để thay thế số trung bình cộng. Số trung vị cũng là một trong các chỉ tiêu dùng để nêu lên đặc trưng phân phối của dãy số.

4.3.7 Tứ phân vị [1]

Sẽ là không thích hợp nếu ta trình bày tứ phân vị trong phần này, vì nó là chỉ tiêu đo lường độ phân tán chứ không phải là chỉ tiêu đo lường khuynh hướng tập trung. Tuy nhiên cách tính toán của nó tương tự như số trung vị, hơn nữa tứ phân vị thứ hai chính là số trung vị, nên ta kết hợp trình bày trong phần này để dễ theo dõi.

4.3.7.1 Khái niệm

Tứ phân vị chia dãy số lượng biến thành 4 phần, mỗi phần có số đơn vị bằng nhau.

4.3.7.2 Cách xác định tứ phân vị

❖ Tài liệu phân tổ không có khoảng cách tổ: dãy số lượng biến có 3 tứ phân vị.

- Q_1: Tứ phân vị thứ nhất là lượng biến đứng ở vị trí thứ (n+1)/4.
- Q_2: Tứ phân vị thứ hai chính là số trung vị, đứng ở vị trí thứ 2(n+1)/4 = (n+1)/2.
- Q_3: Tứ phân vị thứ ba, là lượng biến đứng ở vị trí thứ 3(n+1)/4.

Khi (n+1) không phải là bội số của 4, tứ phân vị được xác định bằng cách cộng thêm vào[2]. Ví dụ, giả sử ta có n = 12 đơn vị, nên (n+1)/4 = (12+1)/4 = $3^1/_4$. Tứ phân vị thứ nhất sẽ bằng lượng biến của đơn vị ở vị trí thứ ba cộng với ¼ giá trị chênh lệch giữa lượng biến ở vị trí thứ tư và lượng biến ở vị trí

---

[1] Quartiles

[2] Interpolation

thứ ba. Tương tự, $3(n+1)/4 = 9^3/_4$, vì vậy tứ phân vị thứ ba sẽ bằng lượng biến của đơn vị ở vị trí thứ chín cộng thêm $^3/_4$ giá trị chênh lệch giữa lượng biến ở vị trí thứ mười và lượng biến ở vị trí thứ chín.

Ví dụ: Tiền lương tháng của 8 công nhân ở một tổ sản xuất như sau (1000đ):

1800; 1900; 2000; 2100; 2200; 2500; 2700; 2800.

Ta có: $n = 8 \Rightarrow (n+1)/4 = (8+1)/4 = 2^1/_4$.

$$Q_1 = 1900 + \frac{1}{4}(2000 - 1900) = 1925 \text{ ngàn đồng.}$$

$Q_2 = (2100 + 2200)/2 = 2150$ ngàn đồng.

$$Q_3 = 2500 + \frac{3}{4}(2700 - 2500) = 2650 \text{ngàn đồng.}$$

❖ Tài liệu phân tổ có khoảng cách tổ:

- Tứ phân vị thứ nhất:

$$Q_1 = X_{Q1min} + h_{Q1}\frac{\frac{1}{4}\sum f - S_{Q1-1}}{f_{Q1}} \qquad (4.16)$$

- Tứ phân vị thứ ba:

$$Q_3 = X_{Q3min} + h_{Q3}\frac{\frac{3}{4}\sum f - S_{Q3-1}}{f_{Q3}} \qquad (4.17)$$

Trong đó:

Q_1, Q_3: Tứ phân vị thứ nhất và tứ phân vị thứ ba.

X_{Q1min} , X_{Q3min} : Giới hạn dưới của tổ chứa tứ phân vị.

h_{Q1} , h_{Q3} : Khoảng cách tổ chứa tứ phân vị.

S_{Q1-1} , S_{Q3-1} : Tần số tích lũy của các tổ đứng trước tổ chứa tứ phân vị.

F_{Q1} , f_{Q3} : Tần số của tổ chứa tứ phân vị.

Ví dụ: Từ số liệu bảng 4.8, tính ra các tứ phân vị:

* Tứ phân vị thứ nhất chứa trong tổ có tần số tích lũy bằng (n+1)/4 = (79+1)/4 = 20

$$Q_1 = 400 + 100\frac{\frac{1}{4}.79 - 8}{12} = 497{,}92 \text{ triệu đồng.}$$

* Tứ phân vị thứ ba chứa trong tổ có tần số tích lũy bằng 3(n+1)/4 = 3(79+1)/4 = 60

$$Q_3 = 600 + 200\frac{\frac{3}{4}.79 - 45}{25} = 714 \text{ triệu đồng.}$$

Trường hợp tài liệu phân tổ có khoảng cách tổ, người ta cũng có thể xác định tứ phân vị bằng cách sau.

Giả sử một tổ nào đó có giới hạn dưới là x_{min}, giới hạn trên là x_{max}, tần số của tổ là f. Nếu các đơn vị được sắp xếp theo thứ tự tăng dần, lượng biến thứ j được ước tính bằng công thức:

$$x_{min} + (j - 1/2)\frac{x_{max} - x_{min}}{f} \quad \text{với } j = 1,2,...,f$$

Trở lại ví dụ trong bảng 4.8, Ta tính tứ phân vị thứ nhất và tứ phân vị thứ ba.

Ta có n = 79 ⇒ (n+1)/4 = (79 + 1)/4 = 20. Từ bảng 4.8 ta thấy quan sát thứ 20 là quan sát thứ 12 trong tổ (400 – 500), nên j = 12.

Tứ phân vị thứ nhất là: $400 + (12 - 1/2)\frac{500 - 400}{12} = 495{,}83$ triệu đồng.

Để xác định tứ phân vị thứ ba, ta tính: 3(n+1)/4 = 3(79+1)/4 = 60.

Ta thấy quan sát thứ 60 là quan sát thứ 15 trong tổ (600 – 800), nên j = 15.

Tứ phân vị thứ ba là: $600 + (15 - 1/2)\frac{800 - 600}{25} = 716$ triệu đồng.

Trường hợp nếu (n+1) không phải là bội của 4, ta giải quyết giống như ví dụ trong mục 4.3.7.2, tuy nhiên việc tính toán phức tạp.

So sánh với cách tính theo công thức (4.16) và (4.17), ta thấy kết qủa không chênh lệch nhau nhiều.

Trong thực tế đôi khi người ta cũng có nhu cầu chia các đơn vị trong dãy số lượng biến thành 10 phần đều nhau và ta có thập phân vị. Cách tính thập phân vị cũng tương tự như cách tính tứ phân vị. Ta có công thức:

$$D_1 = X_{D1min} + h_{D1}\frac{\frac{1}{10}\sum f - S_{D1-1}}{f_{D1}}$$

$$D_2 = X_{D2min} + h_{D2}\frac{\frac{2}{10}\sum f - S_{D2-1}}{f_{D2}} \qquad (4.18)$$

v.v...

Tứ phân vị, thập phân vị được sử dụng trong thực tế khi người ta muốn biết mức đạt cao nhất của 1/10 hay 1/4 số đơn vị xếp từ thấp lên, hoặc mức đạt thấp nhất của 1/10 hay 1/4 số đơn vị tiên tiến xếp từ cao xuống.

4.3.8 Một số vấn đề lưu ý khi sử dụng số tương đối, số tuyệt đối, số trung bình.

- Trong thực tế trừ một số trường hợp mang tính chất bí mật không được phép công bố số tuyệt đối, người ta thường kết hợp số tương đối với số tuyệt đối để nhận thức hiện tượng một cách chính xác. Ví dụ ở một bệnh viện X, giả sử trong một tuần nào đó chỉ có 2 bệnh nhân vào bệnh viện, và một trong hai bệnh nhân đó bị chết. Vậy có bao nhiêu % bệnh nhân vào bệnh viện tuần đó bị chết? Rõ ràng câu trả lời là 50%. Nếu chỉ nhìn vào số tương đối trên ta thấy thật sự "khủng khiếp" và không ai dám vào bệnh viện X để chữa bệnh. Tuy nhiên nếu ta kết hợp với số tuyệt đối, nghĩa là 50% bệnh nhân bị chết tương ứng với 1 bệnh nhân thì sự việc trở nên hết sức bình thường.
- Một nguyên tắc cần nhớ là số trung bình chỉ được tính ra từ tổng thể đồng chất. Nghĩa là "Chỉ có giữa những lượng cùng gọi bằng một tên thì mới có lượng trung bình."[1]

[1] C.Mác : Tư Bản. Quyển 1, tập 2 trang 14.

4.4 CÁC ĐẶC TRƯNG ĐO LƯỜNG ĐỘ PHÂN TÁN[1]

4.4.1 Khái niệm

Các chỉ tiêu số trung bình, số trung vị và mốt mới chỉ cho ta biết được giá trị trung tâm, mức độ đại biểu của hiện tượng, mà không thể nào phản ánh đầy đủ các tính chất đặc thù của dãy số lượng biến. Do vậy ngoài các đặc trưng đo lường khuynh hướng tập trung ta cần đánh giá độ phân tán (độ biến thiên).

Ví dụ: có hai tổ công nhân, mỗi tổ có 5 người với các mức năng suất lao động như sau (kg):

Tổ 1: 200, 250, 300, 350, 400

Tổ 2: 280, 290, 300, 310, 320.

Năng suất lao động trung bình của mỗi tổ đều là 300kg, tuy nhiên các mức năng suất lao động trong tổ 1 chênh lệch nhiều hơn so với tổ hai, nên số trung bình của tổ 1 kém đại diện hơn so với tổ 2.

Ta có thể sử dụng các chỉ tiêu sau để đo độ phân tán của hiện tượng.

4.4.2 Khoảng biến thiên [2](R)

Là chênh lệch giữa lượng biến lớn nhất (x_{max}) và lượng biến nhỏ nhất (x_{min}) của dãy số lượng biến.

Công thức: $R = x_{max} - x_{min}$ (4.19)

Khoảng biến thiên càng nhỏ thì tổng thể càng đồng đều, số trung bình càng có tính đại diện cao và ngược lại.

Nhược điểm của khoảng biến thiên là chỉ phụ thuộc vào hai lượng biến lớn nhất và nhỏ nhất của dãy số lượng biến.

Trở lại ví dụ trong mục 4.4.1, ta có khoảng biến thiên về năng suất lao động:

$R_1 = 400 - 200 = 200$ (kg)

$R_2 = 320 - 280 = 40$ (kg)

$R_1 > R_2$ có nghĩa là các mức năng suất lao động trong tổ 1 biến thiên nhiều hơn trong tổ 2, do đó số trung bình trong tổ 2 đại diện tốt hơn so với tổ 1.

---

[1] Measures of dispersion

[2] Range

4.4.3 Độ trải giữa (Khoảng tứ phân vị)[1] (R_Q)

Độ trải giữa là chênh lệch giữa tứ phân vị thứ ba và tứ phân vị thứ nhất.(xem mục 4.3.7.2)

$$R_Q = Q_3 - Q_1 \qquad (4.20)$$

Ví dụ: có tài liệu về tiền lương của hai tổ công nhân, mỗi tổ có 11 người, được cho trong bảng sau (triệu đồng)

Tổ 1	0,9	1,2	**1,5**	1,8	2,1	**2,4**	2,7	3,0	**3,3**	3,6	3,9
Tổ 2	1,9	2,0	**2,1**	2,2	2,3	**2,4**	2,5	2,6	**2,7**	2,8	2,9
			Q_1			Q_2			Q_3		

Tổ 1 : $R_Q = Q_3 - Q_1 = 3,3 - 1,5 = 1,8$ triệu đồng.

Tổ 2 : $R_Q = Q_3 - Q_1 = 2,7 - 2,1 = 0,6$ triệu đồng.

Độ trải giữa của tổ 1 lớn hơn độ trải giữa của tổ 2, nghĩa là các mức lương trong tổ 1 biến thiên nhiều hơn trong tổ 2.

4.4.4 Độ lệch tuyệt đối trung bình[2] ($\bar{d}$)

Độ lệch tuyệt đối trung bình là số trung bình cộng của các độ lệch tuyệt đối giữa các lượng biến và số trung bình cộng của các lượng biến đó.

Công thức tính độ lệch tuyệt đối trung bình cho tổng thể và cho mẫu về cơ bản là giống nhau, chỉ khác nhau về số đơn vị tổng thể. Hơn nữa giữa khái niệm tổng thể và mẫu, giữa đơn vị tổng thể và tổng thể mang tính chất tương đối, tùy theo giác độ mà chúng ta xét. Do đó để đơn giản ở đây chỉ trình bày công thức tính cho mẫu.

Công thức:

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \qquad (4.21)$$

---

[1] Interquartile range

[2] Mean absolute deviation

Hoặc $\bar{d} = \dfrac{\sum_{i=1}^{n}|x_i - \bar{x}|f_i}{\sum_{i=1}^{n} f_i}$ (khi x_i có các tần số f_i khác nhau) (4.22)

Trở lại ví dụ trong mục 4.4.1, độ lệch tuyệt đối trung bình từng tổ:

$$\bar{d}_1 = \frac{|200-300|+|250-300|+|300-300|+|350-300|+|400-300|}{5}$$

$$= \frac{300}{5} = 60\,\text{kg}$$

$$\bar{d}_2 = \frac{|280-300|+|290-300|+|300-300|+|310-300|+|320-300|}{5}$$

$$= \frac{60}{5} = 12\,\text{kg}$$

Dựa vào độ lệch tuyệt đối trung bình của hai tổ tính được, ta cũng có cùng kết luận như các chỉ tiêu ở trên. Độ lệch tuyệt đối trung bình càng nhỏ, tổng thể càng đồng đều, do đó tính chất đại biểu của số trung bình càng cao. Độ lệch tuyệt đối trung bình có ưu điểm hơn khoảng biến thiên vì nó xét đến tất cả các lượng biến trong dãy số.

4.4.5. Phương sai [1]

Phương sai là số trung bình cộng của bình phương các độ lệch giữa các lượng biến và số trung bình cộng của các lượng biến đó.

- **Phương sai tính từ tổng thể chung** [2]

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \frac{\sum_{i=1}^{N}x_i^2}{N} - \mu^2 = \overline{x^2} - \mu^2 \qquad (4.23)$$

---

[1] Variance

[2] Diễn giải công thức 4.23

$$\sum_{i=1}^{N}(x_i-\mu)^2 = \sum_{i=1}^{N}(x_i^2 - 2\mu x_i + \mu^2)$$

$$= \sum_{i=1}^{N}x_i^2 - 2\mu\sum_{i=1}^{N}x_i + \sum_{i=1}^{N}\mu^2 = \sum_{i=1}^{N}x_i^2 - 2N\mu^2 + N\mu^2 = \sum_{i=1}^{N}x_i^2 - N\mu^2$$

hoặc $\sigma^2 = \dfrac{\sum_{i=1}^{k}(x_i - \mu)^2 f_i}{\sum_{i=1}^{k} f_i}$ (khi x_i có các tần số f_i khác nhau)(4.24)

với $N = \sum_{i=1}^{k} f_i$

- Phương sai mẫu:

$$\hat{S}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad (4.25)$$

Hoặc $\hat{S}^2 = \dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{k} f_i}$ (khi x_i có các tần số f_i khác nhau) (4.26)

Với $n = \sum_{i=1}^{k} f_i$

Phương sai mẫu còn được tính theo công thức:

$$\hat{S}^2 = \overline{x^2} - (\bar{x})^2 \qquad (4.27)$$

- Phương sai mẫu hiệu chỉnh:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (4.28)$$

Hoặc: $$S^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1} \qquad (4.29)$$

Phương sai mẫu hiệu chỉnh được sử dụng nhiều trong thống kê suy diễn, như ước lượng, kiểm định...Do đó trong các chương sau khi nói đến phương sai mẫu là ta đề cập đến phương sai mẫu hiệu chỉnh.

Ví dụ: Số lỗi sai tìm thấy trong một cuốn sách dày 500 trang

Bảng 4.9

Số lỗi	Số trang
0	102
1	138
2	140
3	79
4	33
5	8
Tổng	500

Ta lập bảng tính phương sai về số lỗi sai

Bảng 4.10

x_i	f_i	$x_i f_i$	$x_i^2 f_i$	$(x_i - \mu)^2 f_i$ $\mu = 1,654$
0	102	0	0	279,043
1	138	138	138	59,025
2	140	280	560	16,760
3	79	237	711	143,126
4	33	132	528	181,623
5	8	40	200	89,566
Tổng	500	827	2.137	769,143

Số lỗi trung bình mỗi trang:

$$\mu = \frac{827}{500} = 1,654 \text{ lỗi/trang}$$

Thay số liệu vào công thức 4.24 ta có:

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \mu)^2 f_i}{\sum_{i=1}^{k} f_i} = \frac{769,143}{500} = 1,5383$$

Ta cũng có thể tính phương sai bằng công thức:

$$\sigma^2 = \frac{\sum_{i=1}^{6} x_i^2 f_i}{\sum_{i=1}^{6} f_i} - \mu^2 = \overline{x^2} - \mu^2 = \frac{2.137}{500} - 1,654^2 = 1,538$$

4.4.6 Độ lệch tiêu chuẩn [1]

Độ lệch tiêu chuẩn là căn bậc hai của phương sai.

Công thức tính: $\sigma = \sqrt{\sigma^2}$ (4.30)

Dựa vào ví dụ bảng 4.10, ta tính độ lệch tiêu chuẩn về số lỗi sai:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1,538} = 1,24 \text{ lỗi.}$$

❖ **Ý nghĩa của độ lệch tiêu chuẩn:**

Độ lệch tiêu chuẩn ngoài việc được sử dụng để so sánh độ phân tán của hai tổng thể, nó còn cho biết sự phân phối của các lượng biến trong một tổng thể, thể hiện trên hai quy tắc sau:

- **Quy tắc Tchebychev [2]**

Bất kỳ một tổng thể nào với trung bình là μ và độ lệch tiêu chuẩn là σ, thì có ít nhất $100(1-\frac{1}{m^2})\%$ giá trị rơi vào khoảng $\mu \pm m\sigma$, với $m > 1$.

Bảng 4.11 Các tỉ lệ % ứng với m

m	1,5	2	2,5	3
$100(1-1/m^2)\%$	55,6%	75%	84%	88,9%

Nhìn vào bảng 4.11 ta thấy với m = 1,5, ta có ít nhất 55,6% giá trị rơi vào

[1] Standard deviation

[2] Tchebychev's Rule

khoảng $\mu \pm 1,5\sigma$.

Ví dụ: Giả sử ta có tiền lương hàng năm của 7 nhân viên thuộc bộ phận quản lý:

34,5; 30,7; 32,9; 36,0; 34,1; 33,8; 32,5 (triệu đồng).

Tiền lương trung bình:

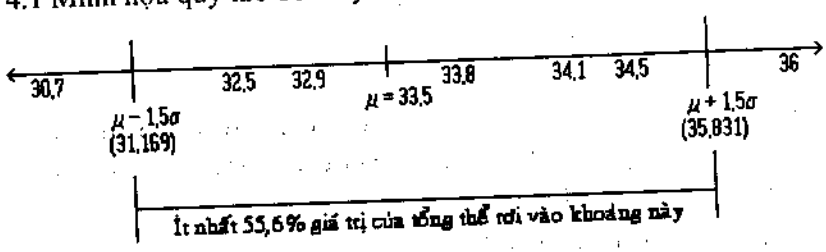
$$\mu = \frac{34,5 + 30,7 + 32,9 + 36,0 + 34,1 + 33,8 + 32,5}{7} = 33,5 \text{ triệu đồng}$$

Độ lệch tiêu chuẩn:

$$\sigma = \sqrt{\frac{(34,5 - 33,5)^2 + ... + (32,5 - 33,5)^2}{7}} = 1,554 \text{ triệu đồng}$$

Theo quy tắc Tchebychev, đối với tổng thể trên, có ít nhất 55,6% mức lương rơi vào khoảng (33,5 ± 1,5.1,554), tức là từ 31,169triệu đồng đến 35,831triệu đồng.

Hình 4.1 Minh họa quy tắc Tchebychev



Thực ra tổng thể trên có $\frac{5}{7} = 0,7143 = 71,43\%$ giá trị rơi vào khoảng $\mu \pm 1,5\sigma$.

Ưu điểm của quy tắc Tchebychev là có thể áp dụng cho bất kỳ tổng thể nào. Tuy nhiên đối với nhiều tổng thể tỉ lệ % rơi vào các khoảng cao hơn nhiều so với giới hạn thấp nhất mà quy tắc Tchebychev đưa ra. Như theo ví dụ trên quy tắc Tchebychev đưa ra 55,6%, nhưng thực tế thì có tới 71,43% giá trị rơi vào.

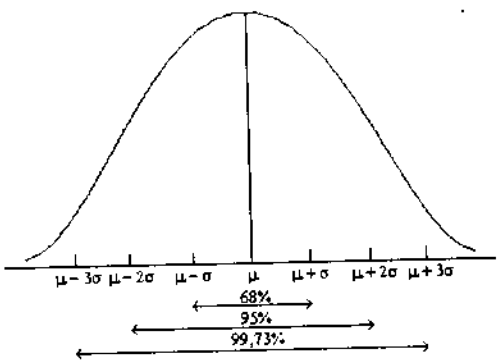
- **Quy tắc thực nghiệm** [1]

Đối với những tổng thể lớn, phân phối chuẩn được sử dụng để mô tả hình dáng của phân phối.

- Khoảng 68% giá trị rơi vào khoảng $\pm \sigma$ so với trung bình.
- Khoảng 95% giá trị rơi vào khoảng $\pm 2\sigma$ so với trung bình.
- Khoảng 99,73% giá trị rơi vào khoảng $\pm 3\sigma$ so với trung bình.

Giả sử chúng ta có một tổng thể lớn về tiền lương, với trung bình là 33,5 triệu đồng, độ lệch tiêu chuẩn là 1,554 triệu đồng. Quy tắc thực nghiệm ước đoán có xấp xỉ 68% mức lương rơi vào khoảng ($\mu \pm \sigma$), tức là từ 31,946 đến 35,054 triệu đồng, và có khoảng 95% mức lương rơi vào khoảng từ 30,392 đến 36,608 triệu đồng.

Hình 4.2 Phân phối các lượng biến trong phân phối chuẩn



4.4.7 Hệ số biến thiên [2] (V)

Hệ số biến thiên là số tương đối tính được bằng cách so sánh giữa độ lệch tiêu chuẩn với số trung bình cộng.

Công thức: $$V = \frac{\sigma}{\mu} \quad \text{hoặc} \quad V = \frac{\hat{s}}{\bar{x}} \qquad (4.31)$$

Trở lại ví dụ trong mục 4.4.1, ta tính hệ số biến thiên:

---

[1] Rule of Thumb

[2] Coefficient of Variation

$$V_1 = \frac{70,7}{300} = 0,2357 = 23,57\%$$

$$V_2 = \frac{14,14}{300} = 0,047 = 4,7\%$$

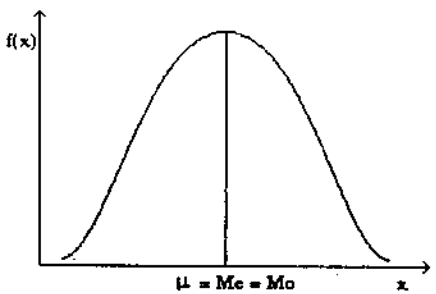
Hệ số biến thiên được dùng để so sánh độ phân tán giữa các hiện tượng có đơn vị tính khác nhau, hoặc giữa các hiện tượng cùng loại nhưng có số trung bình không bằng nhau.

4.4.7 Khảo sát hình dáng phân phối của dãy số

Dựa vào số trung bình, số trung vị và mốt, ta có thể biết được hình dáng phân phối của dãy số.

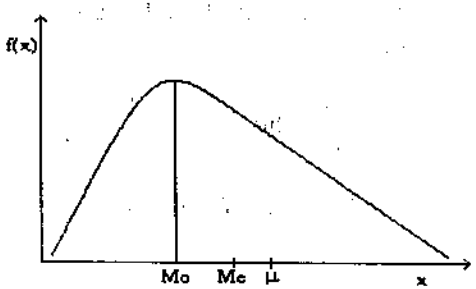
4.4.7.1 Phân phối đối xứng [1]

Phân phối đối xứng khi $\mu = M_e = M_o$



4.4.7.2 Phân phối lệch phải [2]

Phân phối lệch phải khi $\mu > M_e > M_o$

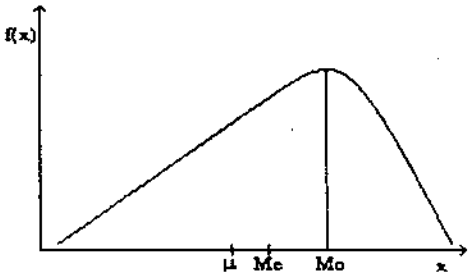


[1] Symmetrical distribution

[2] Skewed–right distribution

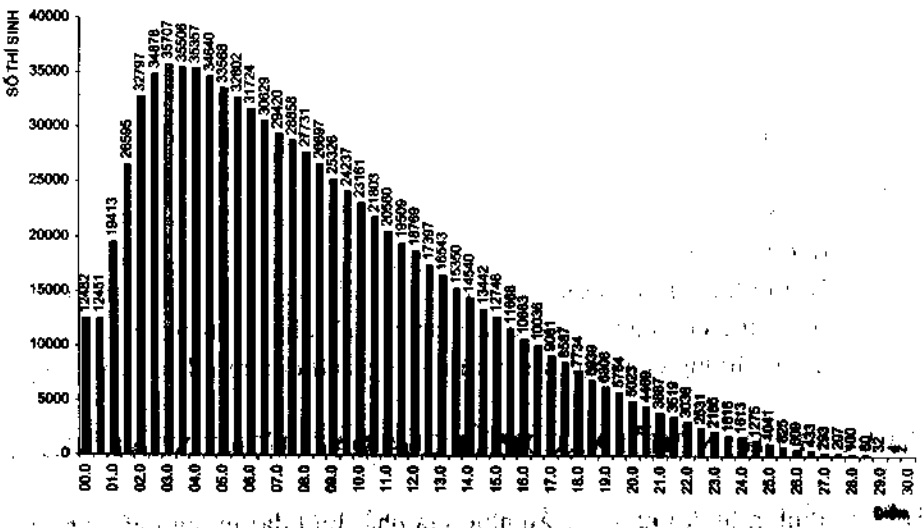
4.4.7.3 Phân phối lệch trái [1]

Phân phối lệch trái khi $\mu < M_e < M_o$



Xem xét phân bố điểm thi tuyển sinh của 141 trường Đại Học, Cao Đẳng năm 2003, ta thấy có dạng phân phối lệch phải. Chỉ có 14% thí sinh có tổng điểm ba môn thi từ 15 điểm trở lên. Có đến 593.999 thí sinh (chiếm 67,9%) đạt tổng điểm ba môn thi 10 điểm trở xuống.

BIỂU ĐỒ PHÂN BỐ ĐIỂM CỦA 141 TRƯỜNG ĐẠI HỌC NĂM 2003



Nguồn: Báo Tuổi Trẻ, phát hành ngày 4/9/2003.

[1] Skewed–left distribution

CHƯƠNG 5

ĐẠI LƯỢNG NGẪU NHIÊN VÀ CÁC QUY LUẬT PHÂN PHỐI XÁC SUẤT THÔNG DỤNG

5.1 KHÁI NIỆM VỀ ĐẠI LƯỢNG NGẪU NHIÊN

Khi gieo một con xúc xắc đều đặn, gọi X là số chấm của mặt xuất hiện trong một lần gieo. Như vậy X sẽ nhận một trong số 6 giá trị có thể có là 1,2,3,4,5,6 và ta gọi X là một đại lượng ngẫu nhiên (biến ngẫu nhiên).

Nói cách khác đại lượng ngẫu nhiên là những đại lượng nhận giá trị này hay giá trị khác trong các giá trị có thể có ở những lần thử khác nhau mà ta không thể khẳng định được trước khi thực hiện phép thử.

Khái niệm tiêu thức số lượng mà ta đề cập trong chương 1 được xem như là các đại lượng ngẫu nhiên và các giá trị có thể có của đại lượng ngẫu nhiên là các lượng biến.

5.2 PHÂN LOẠI ĐẠI LƯỢNG NGẪU NHIÊN

Đại lượng ngẫu nhiên được chia thành hai loại:

- Đại lượng ngẫu nhiên rời rạc[1]: Là đại lượng ngẫu nhiên mà các giá trị có thể có của nó là hữu hạn hay vô hạn và có thể đếm được. Ví dụ: Số chấm ở mặt xuất hiện của con xúc xắc. Số sản phẩm sản xuất ra ở từng phân xưởng...
- Đại lượng ngẫu nhiên liên tục[2]: Là đại lượng ngẫu nhiên mà các giá trị có thể có của nó có thể lấp kín cả một khoảng trên trục số. Ví dụ: Trọng lượng của một sản phẩm, năng suất của một loại cây trồng...

5.3 LUẬT PHÂN PHỐI XÁC XUẤT CỦA ĐẠI LƯỢNG NGẪU NHIÊN

Để xác định một đại lượng ngẫu nhiên ta phải biết đại lượng ngẫu nhiên ấy có thể nhận các giá trị nào và nó nhận các giá trị ấy với xác xuất tương ứng là bao nhiêu.

---

[1] Discrete Random Variables

[2] Continuous Random Variables

5.3.1 Luật phân phối xác suất của đại lượng ngẫu nhiên rời rạc

Nếu X là một đại lượng ngẫu nhiên rời rạc hữu hạn với các giá trị có thể có là $x_1, x_2, ..., x_k$ với các xác suất tương ứng là $p_1, p_2, ..., p_k$. Luật phân phối xác suất của đại lượng ngẫu nhiên rời rạc X được trình bày trong bảng sau:

X	P_X
x_1	p_1
x_2	p_2
.	.
.	.
x_k	p_k
Tổng	1

Trong đó: $p_i = P(X = x_i)$ $i = 1,2,..., k$

$$\sum_{i=1}^{k} p_i = 1$$

Ví dụ: Khi gieo một con xúc xắc đều đặn. Gọi X là số chấm của mặt xuất hiện, ta có bảng phân phối xác suất của đại lượng ngẫu nhiên X như sau:

X	P_X
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Tổng	1

Các đặc trưng của phân phối đại lượng ngẫu nhiên rời rạc như trung bình, phương sai, độ lệch tiêu chuẩn, mốt, số trung vị được trình bày trong chương 4, trong đó p_i tương ứng với $\frac{f_i}{\sum f_i}$.

Ví dụ: Có tài liệu về bậc thợ của 50 công nhân trong một xí nghiệp như sau:

Bậc thợ	1	2	3	4	5	6	7
Số CN	3	8	14	9	6	6	4

Nếu gọi X là bậc thợ của một công nhân chọn ngẫu nhiên từ xí nghiệp này thì X là đại lượng ngẫu nhiên có quy luật phân phối xác suất như sau:

X	1	2	3	4	5	6	7
P_X	0,06	0,16	0,28	0,18	0,12	0,12	0,08

Từ bảng phân phối của X ta tính được:

- Bậc thợ trung bình của xí nghiệp:

$$\mu = \sum_{i=1}^{7} x_i p_i = 1.0,06 + 2.0,16 + 3.0,28 + ... + 7.0,08 = 3,82$$

- Phương sai về bậc thợ:

$$\sigma^2 = \sum_{i=1}^{7} (x_i - \mu)^2 p_i = (1-3,82)^2.0,06 + (2-3,82)^2 + ... + (7-3,82)^2.0,08$$

$$= 2,7476$$

- Độ lệch tiêu chuẩn về bậc thợ:

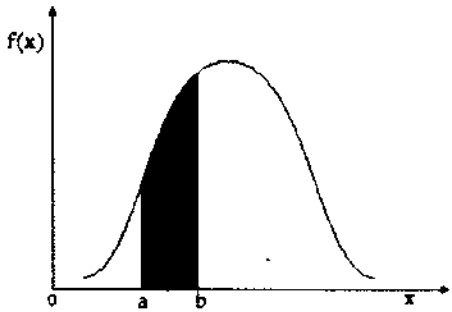
$$\sigma = \sqrt{\sigma^2} = \sqrt{2,7476} = 1,6576$$

- Mốt về bậc thợ:

P(X=3) = 0,28 lớn nhất, vậy Mod(X) = 3

5.3.2 Luật phân phối xác suất của đại lượng ngẫu nhiên liên tục

Gọi X là đại lượng ngẫu nhiên liên tục, xác suất để X nhận giá trị trong đoạn $[a,b]$ nào đó là $P(a \le X \le b) = \int_a^b f(x)dx$. Trong đó f(x) được gọi là hàm mật độ xác suất, với $f(x) \ge 0$ và $\int_{-\infty}^{+\infty} f(x)dx = 1$. Nói cách khác, xác suất để X nhận giá trị trong đoạn $[a;b]$ bằng diện tích hình thang cong được chắn bởi trục hoành, đường cong f(x), và hai đường thẳng x = a và x = b (miền tô đen trên hình vẽ)



Hình 5.1 xác suất để X nhận giá trị trong đoạn $[a,b]$

5.4 MỘT SỐ QUY LUẬT PHÂN PHỐI XÁC SUẤT THÔNG DỤNG

5.4.1 Quy luật phân phối nhị thức [1]

- Bài toán tổng quát:

Giả sử ta tiến hành n phép thử độc lập. Gọi A là biến cố nào đó mà ta quan tâm. Ở mỗi phép thử chỉ có thể xảy ra một trong hai trường hợp: hoặc biến cố A xảy ra (thành công) hoặc A không xảy ra (thất bại). Ở mọi phép thử xác suất để cho A xảy ra luôn bằng hằng số p, tức là P(A) = p và $P(\overline{A}) = 1 - p = q$. Gọi X là số lần biến cố A xảy ra trong n phép thử, thì X là đại lượng ngẫu nhiên rời rạc có thể nhận các giá trị 0,1,2,...,n với các xác suất tương ứng được tính theo công thức Bernoulli:

$$P(X = x) = C_n^x p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \qquad (5.1)$$

Với: $x = 0,1,2,\ldots,n$

$0 < p < 1, q = 1 - p$

- **Định nghĩa**: Phân phối nhị thức là phân phối của đại lượng ngẫu nhiên rời rạc X nhận các giá trị 0,1,2,...,n với các xác suất tương ứng được tính theo công thức 5.1.

đại lượng ngẫu nhiên X phân phối theo quy luật nhị thức với các tham số n và p được ký hiệu là X~ B(n,p).

- Các đặc trưng của phân phối nhị thức:

Giá trị trung bình của X là: $\mu = np$ (5.2)

Phương sai của X là: $\sigma^2 = npq$ (5.3)

[1] The Binomial distribution

Để tính P(X= x) hoặc P(X ≤ x) có thể dùng hàm BINOMDIST trong Excel.

P(X= x) =**BINOMDIST(x,n,p,0)**

P(X≤ x) =**BINOMDIST(x,n,p,1)**

Ví dụ: Một máy sản xuất với tỷ lệ phế phẩm là 2%, lấy 10 sản phẩm để kiểm tra. Tính xác xuất để trong số đó:

a/ Có 2 phế phẩm.

b/ có không quá 2 phế phẩm.

Giải

a/ Gọi X là số phế phẩm trong 10 sản phẩm, thì X~ B(10; 0,02)

Áp dụng công thức 5.1 ta có xác xuất để có 2 phế phẩm:

$$P(X=2)=C_{10}^{2}0{,}02^{2}0{,}98^{8}=\frac{10!}{2!(10-2)!}0{,}02^{2}0{,}98^{8}=0{,}0153$$

b/ Xác suất để có không quá 2 phế phẩm:

$$P(X\le 2)=P(0\le X\le 2)=P(X=0)+P(X=1)+P(X=2)=p_0+p_1+p_2$$

$$P(X=0)=p_0=C_{10}^{0}0{,}02^{0}0{,}98^{10}=0{,}8171$$

$$P(X=1)=p_1=C_{10}^{1}0{,}02^{1}0{,}98^{9}=0{,}1667$$

vậy $P(X\le 2)=0{,}8171+0{,}1667+0{,}0153=0{,}9991$

5.4.2 Quy luật phân phối Poisson

Phân phối Poisson là phân phối của đại lượng ngẫu nhiên rời rạc X nhận một trong các giá trị 0, 1, 2, ..., n ... với các xác suất tương ứng được tính theo công thức:

$$P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!} \qquad (5.4)$$

Trong đó: e: là hằng số Nêpe gần bằng 2,71828

λ = np là hằng số dương.

k = 0,1,2... số trường hợp thành công.

Khi n càng lớn, việc tính toán đối với phân phối nhị thức càng khó khăn, do đó người ta tìm cách tính gần đúng đơn giản. Phân phối Poisson được sử dụng để xấp xỉ cho phân phối nhị thức khi n rất lớn và p rất nhỏ (n ≥ 20 ; p ≤ 0,05).

Nghĩa là: $$P(X=k)=C_n^k p^k q^{n-k}\approx\frac{e^{-\lambda}\lambda^{k}}{k!}$$

X có phân phối Poisson với tham số λ được ký hiệu là X~$\mathcal{P}$ (λ).

Nếu X~ $\mathcal{P}$(λ) để tính P(X=k) hoặc P(X≤ k) ta dùng hàm Poisson trong Excel:

$$P(X = k) = \mathbf{POISSON(k,\lambda,0)}$$
$$P(X \le k) = \mathbf{POISSON(k,\lambda,1)}$$

- Đặc trưng của phân phối Poisson:

Trong phân phối Poisson , trung bình và phương sai bằng nhau và bằng λ .
Trong thực tế phân phối Poisson được sử dụng để mô tả số sự kiện xuất hiện trong một khoảng thời gian, trong một khu vực nào đó. Ví dụ: Số tai nạn giao thông trong một tháng ở một thành phố. Số lần hư hỏng của một máy sản xuất trong một tháng...

Ví dụ: Một máy sản xuất với tỷ lệ phế phẩm là 0,2%. Lấy 1000 sản phẩm để kiểm tra. Tính xác suất để trong số đó có 2 phế phẩm.

Giải:

Nếu gọi X là số phế phẩm có trong 1000 sản phẩm thì X ~ B(1000; 0,002).
Vì n = 1000 khá lớn và p = 0,002 rất nhỏ và $\lambda = np = 1000\times 0{,}002 = 2$ không đổi, nên ta có thể coi X~ $\mathcal{P}$(2).

Xác suất để có 2 phế phẩm trong 1000 sản phẩm:

$$P(X = 2) = \frac{e^{-2}2^2}{2!} = 0{,}2707$$

Ta cũng có thể tính xác xuất mà bài toán yêu cầu bằng phân phối nhị thức:

$$P(X = 2) = C_{1000}^2 0{,}002^2 0{,}998^{998} = 0{,}2734$$

5.4.3 Quy luật phân phối chuẩn [1]

Định nghĩa: Phân phối chuẩn là phân phối của đại lượng ngẫu nhiên liên tục X nhận giá trị từ $-\infty$ đến $+\infty$ với hàm mật độ xác suất của nó có dạng:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (5.5)$$

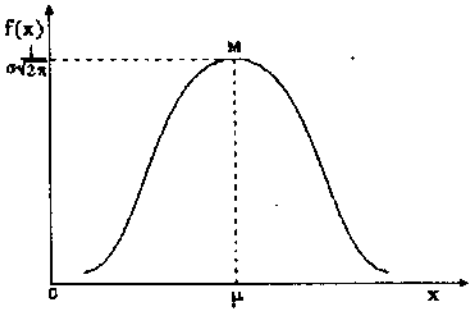
Trong đó: $e = 2{,}71828$
$\pi = 3{,}14159$
μ : trung bình
σ : độ lệch chuẩn
$-\infty < x < +\infty$

Đồ thị của hàm f(x) có dạng hình chuông, đối xứng qua đường thẳng $x = \mu$.

Hàm $f(x) > 0$ ($\forall x$) và đạt cực đại tại điểm M (μ ; $\frac{1}{\sigma\sqrt{2\pi}}$).

[1] Normal distribution

Hình 5.2 Đồ thị của phân phối chuẩn:



Nếu dựa vào công thức 5.5 để tính xác suất khi đại lượng ngẫu nhiên X nhận giá trị trong một khoảng nào đó của trục số ta phải tính μ và σ , công việc tính toán sẽ rất nặng nề, do đó người ta đưa phân phối chuẩn tổng quát về phân phối chuẩn đơn giản (phân phối chuẩn tắc).

- Phân phối chuẩn đơn giản [1]

Giả sử đại lượng ngẫu nhiên X phân phối theo quy luật chuẩn với trung bình là μ và phương sai là σ^2, thì đại lượng ngẫu nhiên:

$Z = \dfrac{X-\mu}{\sigma}$ $(-\infty < Z < +\infty)$ có phân phối chuẩn đơn giản với hàm mật độ xác suất của Z có dạng:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

Đại lượng ngẫu nhiên Z phân phối theo quy luật chuẩn đơn giản có μ = 0 và $\sigma^2 = 1$ được ký hiệu là Z ~ N(0,1). Như vậy ta có thể biến đổi một biến ngẫu nhiên X bất kỳ có phân phối chuẩn với trung bình là μ và phương sai là σ^2 thành biến ngẫu nhiên Z có phân phối chuẩn đơn giản có trung bình μ = 0 và $\sigma^2 = 1$ với phép biến đổi $Z = \dfrac{X-\mu}{\sigma}$ sau đó ta có thể tính xác suất từ bảng đã lập sẵn gọi là bảng tích phân Laplace cho ở cuối sách.

Giả sử ta có $X \sim N(\mu,\sigma^2)$, xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng (a,b), với a < b:

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \quad (5.6)$$

$$= \varphi\left(\frac{b-\mu}{\sigma}\right) - \varphi\left(\frac{a-\mu}{\sigma}\right)$$

[1] Standard normal distribution

Trong đó: $\varphi(z) = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{z^2}{2}}dz$ được gọi là tích phân Laplace.

Hàm $\varphi(z)$ là hàm lẻ. Vì vậy ta chỉ cần quan tâm các giá trị $\varphi(z)$ với $z > 0$ là đủ. Hàm $\varphi(z)$ được tính bằng máy tính và lập bảng sẵn gọi là bảng tích phân Laplace cho ở phụ lục cuối sách.

Ví dụ: Trọng lượng của một loại sản phẩm Là X có phân phối chuẩn với $\mu = 8{,}6$kg, $\sigma^2 = 0{,}36$. Lấy một sản phẩm bất kỳ:

a/ Tính xác suất để sản phẩm ấy có trọng lượng từ 8kg đến 9,8kg.

b/ Tính xác suất để sản phẩm ấy có trọng lượng nhỏ hơn 7,8kg.

Giải:

a/

$$P(8 \le X \le 9{,}8) = P(\frac{8-8{,}6}{0{,}6} \le Z \le \frac{9{,}8-8{,}6}{0{,}6}) = P(-1 \le Z \le 2)$$

$$= \varphi(2) - \varphi(-1) = \varphi(2) + \varphi(1)$$

vì $\varphi(z)$ là hàm lẻ nên $\varphi(-z) = -\varphi(z)$

Tra bảng ta có: $\varphi(1) = 0{,}3413$

$\varphi(2) = 0{,}4772$

Vậy $P(8 \le X \le 9{,}8) = 0{,}4772 + 0{,}3413 = 0{,}8185$

b/

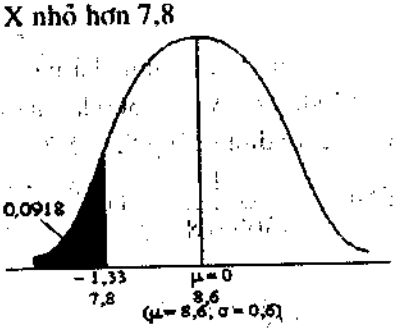
$$P(X < 7{,}8) = P(Z < \frac{7{,}8-8{,}6}{0{,}6}) = P(Z < -1{,}33) = 0{,}5 - \varphi(1{,}33)$$

Không để ý dấu, ta tra bảng, $\varphi(1{,}33) = 0{,}4082$

Vậy $P(X < 7{,}8) = 0{,}5 - 0{,}4082 = 0{,}0918$

Ta có thể minh họa xác suất trên trong hình 5.3.

Hình 5.3: Xác suất để X nhỏ hơn 7,8



Chú ý: Nếu $X \sim N(\mu, \sigma^2)$ để tính $P(X < x)$ ta có thể dùng hàm NORMDIST trong Excel.

$P(X < x)$ = **NORMDIST(x,μ,σ,1)**

$P(a \le X \le b)$ = **NORMDIST(b,μ,σ,1) - NORMDIST(a,μ,σ,1)**

Nếu $Z \sim N(0, 1)$:

$P(Z < z)$ = **NORMSDIST(z)**

$P(\alpha \le Z \le \beta)$ = **NORMSDIST(β) - NORMSDIST(α)**

5.4.4 Dùng phân phối chuẩn để xấp xỉ phân phối nhị thức và phân phối Poisson

5.4.4.1 Xấp xỉ phân phối nhị thức

Khi sử dụng quy luật phân phối nhị thức, nếu n khá lớn thì công việc tính toán sẽ gặp khó khăn, lúc đó nếu p nhỏ đến mức np ≈ npq hoặc theo kinh nghiệm $p \le 0,05$ thì có thể dùng quy luật phân phối Poisson thay thế cho quy luật phân phối nhị thức. Nhưng nếu p không nhỏ ($p > 0,05$), theo kinh nghiệm $np \ge 5$ và $nq \ge 5$ thì ta dùng quy luật phân phối chuẩn để thay thế cho phân phối nhị thức.

Tóm lại, khi n lớn ($n \ge 30$) và p không quá gần 0 và không quá gần 1 thì đại lượng ngẫu nhiên $X \sim B(n,p)$ có thể coi như xấp xỉ phân phối chuẩn với trung bình $\mu = np$ và phương sai $\sigma^2 = npq$.

Ta có: $$P(X = x) = C_n^x p^x q^x \approx \frac{1}{\sqrt{npq}} f(z) \qquad (5.7)$$

Trong đó: $Z = \dfrac{X - np}{\sqrt{npq}}$; $f(z) = \dfrac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$

Ví dụ: Xác suất để một máy sản xuất được sản phẩm loại I là 0,8. Tìm xác suất để trong 300 sản phẩm do máy sản xuất ra có:

a/ 254 sản phẩm loại I.

b/ Có từ 232 đến 250 sản phẩm loại I.

Giải:

a/ Gọi X là sản phẩm loại I có trong 300 sản phẩm do máy sản xuất. X có phân phối nhị thức $X \sim B(300; 0,8)$. Vì n = 300 là khá lớn và p = 0,8 không quá gần 0 và 1 nên ta có thể áp dụng công thức 5.7.

$$P(X = 254) \approx \frac{1}{\sqrt{300.0,8.0,2}} f\left(\frac{254 - 300.0,8}{\sqrt{300.0,8.0,2}}\right)$$

$$= \frac{1}{\sqrt{48}} f(2,02)$$

Tra bảng ta được $f(2,02) = 0,0519$

Vậy $P(X = 254) \approx 0,00749$

b/ Ta tính:

$$P(232 \le X \le 250) \approx \varphi(\frac{250-300.0,8}{\sqrt{300.0,8.0,2}}) - \varphi(\frac{232-300.0,8}{\sqrt{300.0,8.0,2}})$$

$$= \varphi(1,44) - \varphi(-1,15) = 0,1415 + 0,20594 \approx 0,34744$$

5.4.4.2 Xấp xỉ phân phối Poisson

Nếu đại lượng ngẫu nhiên X phân phối theo quy luật Poisson, nhưng có $\lambda \ge 5$ thì có thể xem X có phân phối xấp xỉ chuẩn với $\mu = \lambda$ và $\sigma^2 = \lambda$.

Ta dùng phép biến đổi như sau:

$$Z = \frac{X-\lambda}{\sqrt{\lambda}}$$

Ví dụ: Số tai nạn lao động trung bình trong một năm của một nhà máy là 6,5 vụ. Tính xác suất trong một năm nào đó có tối đa 7 vụ tai nạn lao động.

Giải:

Ta có: $Z = \frac{X-\lambda}{\sqrt{\lambda}}$ với $X = 7; \lambda = 6,5$

Thay số vào: $Z = \frac{7-6,5}{\sqrt{6,5}} = 0,59$

$$P(X \ge 7) = P(\frac{X-\lambda}{\sqrt{\lambda}} \ge \frac{7-6,5}{\sqrt{6,5}}) = P(Z \ge 0,59) = 0,5 + \varphi(0,59)$$

$$= 0,5 + 0,2224 = 0,7224$$

Tra bảng phân phối chuẩn ta có $\varphi(0,59) = 0,2224$

Vậy xác suất để một năm nào đó có tối đa 7 vụ tai nạn lao động là 0,7224

5.4.5 Phân phối Chi bình phương [1] (χ^2)

Giả sử X_i (i = 1,2,..,n) là các biến ngẫu nhiên được chọn từ một phân phối chuẩn. Khi đó:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \text{ trong đó } S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

có phân phối Chi bình phương với n-1 bậc tự do. Khi n càng lớn phân phối Chi bình phương sẽ xấp xỉ phân phối chuẩn.

---

[1] Chi – square distribution

Phân phối Chi bình phương được sử dụng để kiểm định tính độc lập của hai biến hoặc dùng để so sánh sự phù hợp giữa các tần số quan sát và tần số lý thuyết. Ngoài ra nó còn dùng để suy rộng phương sai tổng thể.

5.4.6 Phân phối Student t [1]

Giả sử X_i (i = 1,2,..,n) là các biến ngẫu nhiên được chọn từ một phân phối chuẩn. Trường hợp n < 30 phương sai của tổng thể σ^2 chưa biết, khi đó:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$ có phân phối Student với n-1 bậc tự do.

Phân phối Student là phân phối xác suất có hình dáng gần giống với phân phối chuẩn. Hàm mật độ của phân phối Student là hàm chẵn, đồ thị hàm mật độ đối xứng qua trục tung. Khi n càng lớn thì phân phối Student sẽ tiến rất nhanh về phân phối chuẩn. Vì vậy khi n >30 ta có thể dùng phân phối chuẩn thay cho phân phối Student.

5.4.7 Phân phối Fisher – Snedecor (phân phối F)

Gọi χ_1^2 và χ_2^2 là hai đại lượng ngẫu nhiên có phân phối Chi bình phương với bậc tự do tương ứng là v_1 và v_2, thì:

$$F_{v_1,v_2} = \frac{\chi_1^2 / v_1}{\chi_2^2 / v_2}$$

có phân phối Fisher – Snedecor, với hai bậc tự do là v_1 và v_2.

Giả sử ta có hai mẫu cỡ n_x , n_y được chọn ngẫu nhiên độc lập từ hai tổng thể có phân phối chuẩn X và Y, với phương sai σ_x^2, σ_y^2. Gọi S_x^2, S_y^2 là các phương sai mẫu, thì khi đó:

$$F = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$$ có phân phối Fisher – Snedecor với bậc tự do của tử số là $n_x - 1$ và của mẫu số là $n_y - 1$.

5.5 PHÂN PHỐI MẪU

5.5.1 Mối liên hệ giữa tổng thể chung và tổng thể mẫu

Từ tổng thể chung, nếu theo cách chọn hoàn lại (chọn lặp) hoặc theo cách chọn không hoàn lại (chọn không lặp), có thể xây dựng được nhiều tổng

---

[1] t distribution

thể mẫu khác nhau. Từ tổng thể chung, về mặt lý thuyết, nếu lấy theo cách chọn có trả lại, có thể chọn ra N^n tổng thể mẫu khác nhau và nếu lấy theo cách chọn không trả lại thì có thể chọn ra $C_N^n \left(C_N^n = \frac{N!}{n!(N-n)!} \right)$ tổng thể mẫu khác nhau. Mỗi tổng thể mẫu ấy đều được chọn một cách ngẫu nhiên, nên các tham số của nó (như số trung bình mẫu, tỷ lệ mẫu, phương sai mẫu) là những đại lượng ngẫu nhiên tuân theo những quy luật phân phối nhất định.

Giả định rằng nếu điều tra toàn bộ N đơn vị của tổng thể chung thì cuối cùng sẽ biết được tất cả các trị số cụ thể đó, và từ đó tính ra được các tham số mô tả tổng thể chung.

Các tham số của tổng thể chung và tổng thể mẫu được liệt kê trong bảng sau:

Bảng 5.1

Tham số	Tổng thể chung (X)	Tổng thể mẫu $(x_1, x_2, ..., x_n)$
Trung bình	$\mu = \frac{x_1 + x_2 + ... + x_N}{N}$	$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$
Tỷ lệ	$p = \frac{M}{N}$	$\hat{p} = \frac{m}{n}$
Phương sai	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$

Phương sai mẫu hiệu chỉnh:

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

Đương nhiên điều giả định như trên là không thể có, do ta không điều tra toàn bộ N đơn vị tổng thể chung. Các tham số của tổng thể chung nói trên là chưa biết, nhưng chắc chắn chúng tồn tại khách quan và ta phải xác định chúng bằng phương pháp chọn mẫu, tức là xác định thông qua các tham số của tổng thể mẫu.

5.5.2 Khái niệm phân phối mẫu

Phân phối mẫu hay nói rõ hơn là phân phối các đặc trưng mẫu như số trung bình, tỷ lệ, phương sai mẫu. Để có thể tìm được các tham số của tổng thể chung bằng cách suy đoán từ các tham số của tổng thể mẫu cần phải nắm được quy luật phân phối của các tham số của tổng thể mẫu.

Nhìn chung trong việc ứng dụng phương pháp chọn mẫu trong kinh tế, ta thường quan tâm nhiều nhất đến ba tham số là trung bình, tỷ lệ và phương sai mẫu, vì vậy ở đây cũng chỉ chú ý đến việc nắm quy luật phân phối của các tham số này. Mặt khác, tìm quy luật phân phối của mẫu là một vấn đề rất phức tạp, vì vậy ta đặc biệt chú ý đến giả thuyết là tổng thể chung được phân phối theo quy luật chuẩn, vì đó là trường hợp đơn giản nhất và những quy luật mẫu xuất phát từ giả thuyết đó đều là những quy luật thông dụng.

5.5.2.1 Phân phối của trung bình mẫu

Để nắm được quy luật phân phối của trung bình mẫu ta xét định lý giới hạn trung tâm [1]. Định lý giới hạn trung tâm phát biểu như sau:

Giả sử $X_1, X_2, \ldots, X_n$ là dãy các đại lượng ngẫu nhiên độc lập cùng phân phối, với giá trị trung bình là μ và độ lệch tiêu chuẩn là σ, thì khi đó:

Trung bình $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ sẽ có phân phối xấp xỉ chuẩn với giá trị trung bình là μ và độ lệch tiêu chuẩn là $\dfrac{\sigma}{\sqrt{n}}$. Nghĩa là các giá trị có thể có của $\overline{X}$ ổn định quanh μ hơn các giá trị có thể có của X.

Từ định lý giới hạn trung tâm ta rút ra kết luận:

- Nếu đại lượng ngẫu nhiên X có phân phối chuẩn $X \sim N(\mu, \sigma^2)$ thì $\overline{X}$ cũng có phân phối chuẩn $\overline{X} \sim N(\mu, \dfrac{\sigma^2}{n})$.
- Với kích thước mẫu khá lớn ($n \geq 30$) thì trung bình mẫu $\overline{X}$ sẽ có phân phối chuẩn bất kể quy luật phân phối xác suất của tổng thể như thế nào.

Ví dụ 1: Trọng lượng trung bình của một sản phẩm do một máy sản xuất ra là 78,5g, với độ lệch tiêu chuẩn là 11,2g. Chọn ngẫu nhiên 20 sản phẩm.

---

[1] Central Limit Theorem

Gọi $\overline{X}$ là trọng lượng trung bình của 20 sản phẩm. Tính xác suất để $\overline{X}$ lớn hơn 82g.

Giải:

Theo định lý giới hạn trung tâm, ta có $\overline{X}$ có phân phối xấp xỉ chuẩn với giá trị trung bình là 78,5g và độ lệch tiêu chuẩn là $\frac{11,2}{\sqrt{20}} = 2,504$.

Ta tính: $Z = \frac{82 - 78,5}{2,504} = 1,398$

Tra bảng ta có: $\varphi(1,398) = 0,4192$

Vậy $P(\overline{X} > 82) = 0,5 - 0,4192 = 0,0808$

Ví dụ 2: Để hiểu rõ hơn quy luật phân phối của trung bình mẫu, ta xét tổng thể được tạo ra từ việc tung con xúc xắc đều đặn vô số lần. Gọi X là biến ngẫu nhiên biểu thị số chấm của mặt xuất hiện của một con xúc xắc ở một lần gieo bất kỳ. Luật phân phối xác suất của biến ngẫu nhiên X được trình bày trong bảng sau:

X	1	2	3	4	5	6
P_X	1/6	1/6	1/6	1/6	1/6	1/6

Tổng thể thì lớn vô hạn, vì ta có thể tung con xúc xắc vô số lần. Ta có thể tính trung bình và phương sai của tổng thể.

Trung bình tổng thể:

$$\mu = \sum x_i p_i = 1.1/6 + 2.1/6 + ... + 6.1/6 = 3,5$$

Phương sai tổng thể:

$$\sigma^2 = \sum (x_i - \mu)^2 p_i = (1-3,5)^2.1/6 + (2-3,5)^2.1/6 + ... + (6-3,5)^2.1/6$$
$$= 2,92$$

Bây giờ giả sử trung bình của tổng thể là μ chưa biết và chúng ta muốn suy đoán (ước lượng) giá trị của nó bằng việc sử dụng trung bình mẫu $\overline{x}$ được tính toán từ một mẫu có kích thước n = 2. Trong thực tế chỉ một mẫu được chọn và vì vậy cũng chỉ có một giá trị của $\overline{x}$. Nhưng để hiểu rõ làm thế nào $\overline{x}$ có thể ước lượng giá trị của μ, ta liệt kê tất cả các khả năng có thể có với mẫu có kích thước là 2 và tính toán các trung bình mẫu theo bảng sau. Với cỡ mẫu n = 2, số khả năng có thể thiết lập là 36.

Bảng 5.2 Các mẫu với kích thước n = 2 và trung bình của chúng

Mẫu	$\bar{x}$	Mẫu	$\bar{x}$	Mẫu	$\bar{x}$
1,1	1,0	3,1	2,0	5,1	3,0
1,2	1,5	3,2	2,5	5,2	3,5
1,3	2,0	3,3	3,0	5,3	4,0
1,4	2,5	3,4	3,5	5,4	4,5
1,5	3,0	3,5	4,0	5,5	5,0
1,6	3,5	3,6	4,5	5,6	5,5
2,1	1,5	4,1	2,5	6,1	3,5
2,2	2,0	4,2	3,0	6,2	4,0
2,3	2,5	4,3	3,5	6,3	4,5
2,4	3,0	4,4	4,0	6,4	5,0
2,5	3,5	4,5	4,5	6,5	5,5
2,6	4,0	4,6	5,0	6,6	6,0

Ta nhận thấy $\bar{x}$ có 11 giá trị khác nhau: 1,0 ; 1,5 ; 2,0 ;. . . ; 6,0. Giá trị $\bar{x}$ = 1,0 chỉ gặp có một lần nên xác suất của nó là 1/36. Giá trị $\bar{x}$ = 1,5 xảy ra hai lần vì vậy P(1,5) = 2/36. Các giá trị khác của $\bar{x}$ cũng tính tương tự.

Bảng 5.3 Phân phối mẫu của $\bar{X}$

$\bar{x}$	$P_{\bar{x}}$	$\bar{x}$	$P_{\bar{x}}$	$\bar{x}$	$P_{\bar{x}}$
1,0	1/36	3,0	5/36	5,0	3/36
1,5	2/36	3,5	6/36	5,5	2/36
2,0	3/36	4,0	5/36	6,0	1/36
2,5	4/36	4,5	4/36		

Trung bình của $\bar{x}$:

$$\mu_{\bar{x}} = \sum \bar{x}.p_{\bar{x}} = 1{,}0.1/36 + 1{,}5.2/36 + ... + 6{,}0.1/36 = 3{,}5$$

Phương sai của $\bar{x}$:

$$\sigma^2_{\bar{x}} = \sum (\bar{x} - \mu_{\bar{x}})^2.p_{\bar{x}} = (1{,}0 - 3{,}5)^2.1/36 + (1{,}5 - 3{,}5)^2 + ...$$

$$... + (6{,}0 - 3{,}5)^2.1/36 = 1{,}46$$

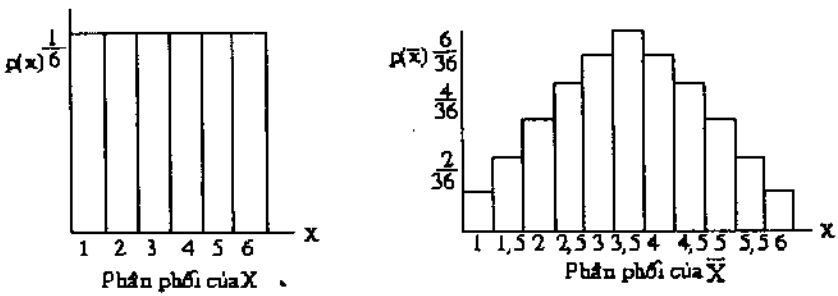
Qua tính toán ta có thể kiểm chứng được các biểu thức:

$$E(\overline{X}) = \mu_{\overline{x}} = \mu = 3,5$$

$$Var(\overline{X}) = \sigma^2_{\overline{x}} = \frac{\sigma^2}{n} = \frac{2,92}{2} = 1,46$$

Ta có thể so sánh phân phối của hai đại lượng ngẫu nhiên X và $\overline{X}$ trong hình vẽ 5.4

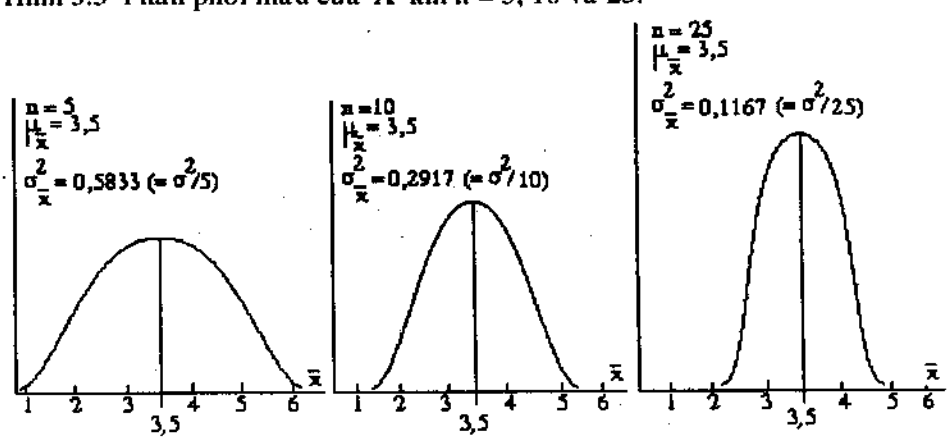
Hình 5.4 Phân phối của X và $\overline{X}$



Nếu bây giờ chúng ta lặp lại quá trình chọn mẫu với cùng một tổng thể chung với các giá trị khác của n, chúng ta sẽ tạo ra một vài phân phối mẫu khác nhau của $\overline{X}$. Hình 5.5 cho ta thấy phân phối mẫu của $\overline{X}$ khi n = 5; 10 và 25. Khi n càng lớn, các giá trị có thể có của $\overline{X}$ cũng càng lớn, vì vậy đồ thị mô tả trong hình 5.5 sẽ được làm trơn (để tránh vẽ quá nhiều hình chữ nhật).

Để ý trong mỗi trường hợp $\mu_{\overline{x}} = \mu$ và $\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$.

Hình 5.5 Phân phối mẫu của $\overline{X}$ khi n = 5, 10 và 25.



5.5.2.2 Phân phối tỷ lệ mẫu

Giả sử một tổng thể chung có N đơn vị trong đó M đơn vị mang dấu hiệu A nào đó mà ta cần quan tâm. Như vậy tỷ lệ các đơn vị mang dấu hiệu A chiếm trong tổng thể gọi là tỷ lệ chung $p = \frac{M}{N}$. Ta lấy ngẫu nhiên n đơn vị từ tổng thể chung, trong đó có m đơn vị mang dấu hiệu A, tỷ lệ mẫu $\hat{p} = \frac{m}{n}$.

$\hat{P}$ là một đại lượng ngẫu nhiên, với n khá lớn có thể xem $\hat{P}$ có phân phối chuẩn với trung bình là p và phương sai là $\frac{pq}{n}$, tức là $\hat{P} \sim N(p, \frac{pq}{n})$.

5.5.2.2 Phân phối phương sai mẫu

Giả sử tổng thể có phân phối chuẩn $X \sim N(\mu, \sigma^2)$, khi đó đại lượng $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ có phân phối Chi bình phương với n-1 bậc tự do.

Ta có kỳ vọng của phương sai mẫu: $E(\hat{S}^2) = \frac{n-1}{n}\sigma^2$

Trong khi đó kỳ vọng của phương sai mẫu có hiệu chỉnh: $E(S^2) = \sigma^2$

Như vậy s^2 là ước lượng không chệch cho σ^2, do đó người ta dùng s^2 để ước lượng cho σ^2.

CHƯƠNG 6

ƯỚC LƯỢNG

Khi nghiên cứu điều tra chọn mẫu, cái chính không phải nhằm nghiên cứu tổng thể mẫu đại diện được chọn ra từ tổng thể chung, mà chính là qua tổng thể mẫu đó để nghiên cứu được tính quy luật và trạng thái của tổng thể chung chứa nó. Nghĩa là dựa vào sự hiểu biết về tham số $\theta'(\overline{X}, \hat{P}, S^2)$ của tổng thể mẫu đã tính ra được để suy ra tham số θ (μ, p, σ^2) của tổng thể chung chưa biết. Việc làm như vậy gọi chung là ước lượng.

6.1. ƯỚC LƯỢNG ĐIỂM [1]:

Thống kê toán đã chứng minh:

$$E(\overline{X}) = \mu$$

$$E(\hat{P}) = p$$

$$E(S^2) = \sigma^2$$

Nghĩa là $\overline{x}$, $\hat{p}$, s^2 là các ước lượng không chệch của μ, p, σ^2.

Để ý rằng phương sai mẫu không dùng ước lượng cho phương sai tổng thể vì đó là ước lượng chệch. Do vậy trong các phần sau khi nói đến phương sai mẫu thay thế cho phương sai tổng thể ta hiểu đó là phương sai mẫu hiệu chỉnh.

Do đó khi đã có mẫu cụ thể ta lấy:

$$\mu \approx \overline{x}$$

$$p \approx \hat{p}$$

$$\sigma^2 \approx s^2$$

6.2. ƯỚC LƯỢNG KHOẢNG [2]:

Trong ước lượng điểm, giá trị ước lượng các đặc trưng của tổng thể phụ thuộc vào một giá trị cụ thể của biến ngẫu nhiên, ví dụ trung bình, tỷ lệ và phương sai mẫu. Ứng với các mẫu khác nhau ta sẽ nhận được các giá trị khác nhau. Do đó chúng không thể biểu thị chính xác của ước lượng. Do vậy ta cần thực hiện ước lượng khoảng, nghĩa là dựa vào số liệu của mẫu,

---

[1] Point Estimation

[2] Interval Esti nation

với một độ tin cậy cho trước, xác định khoảng gía trị mà các đặc trưng của tổng thể có thể rơi vào.

Giả sử tổng thể chung có đặc trưng số θ chưa biết. Căn cứ vào mẫu gồm n đơn vị, ta đưa ra θ_1, θ_2 sao cho θ_1, θ_2 là các đại lượng ngẫu nhiên:
$P(\theta_1 \le \theta \le \theta_2) = 1 - \alpha$

Khoảng (θ_1, θ_2) gọi là khoảng ước lượng của θ.

$1 - \alpha$: độ tin cậy của khoảng ước lượng đó.

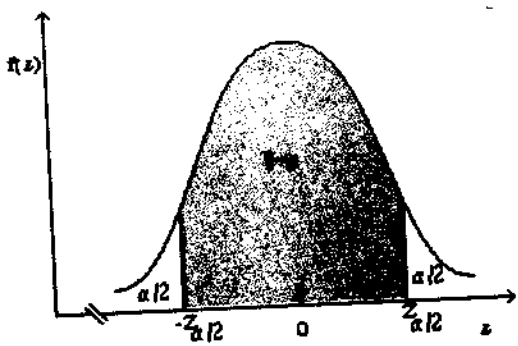
θ_1: giới hạn tin cậy dưới.

θ_2: giới hạn tin cậy trên.

$\frac{1}{2}|\theta_1 - \theta_2|$: độ chính xác của ước lượng khoảng.

Nói chung với cỡ mẫu n cố định thì độ tin cậy và độ chính xác có xu hướng đối lập nhau, tức là khoảng ước lượng càng dài (độ chính xác thấp) càng có cơ hội trúng cao (độ tin cậy cao). Ngược lại khoảng ước lượng càng ngắn (độ chính xác cao) thì càng dễ trật (độ tin cậy thấp).

Ta có thể minh họa ước lượng khoảng bằng hình vẽ sau:



Hình 6.1 $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$ hay $P(Z > z_{\alpha/2}) = P(Z < -z_{\alpha/2}) = \alpha/2$

với $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ là biến ngẫu nhiên có phân phối chuẩn N (0, 1)

6.2.1. Ước lượng trung bình tổng thể [1]:

Ta có các trường hợp sau:

a) $n \ge 30$

[1] Confidence intervals for the mean of a normal population

+ Giả sử chúng ta có một mẫu ngẫu nhiên bao gồm n quan sát được chọn từ một tổng thể có phân phối chuẩn với trung bình là μ chưa biết và phương sai σ^2 đã biết. Trung bình mẫu có phân phối chuẩn $\overline{X} \sim (\mu, \frac{\sigma^2}{n})$.

Với độ tin cậy 1 - α cho trước, trung bình của tổng thể được xác định:

$$\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (6.1)$$

Trong đó Z là biến ngẫu nhiên có phân phối chuẩn.

Ví dụ: Giả sử chúng ta muốn ước lượng khoảng cho trung bình tổng thể với độ tin cậy 90%, từ bảng Z ta có:

$$P(Z < 1{,}645) = 0{,}95$$

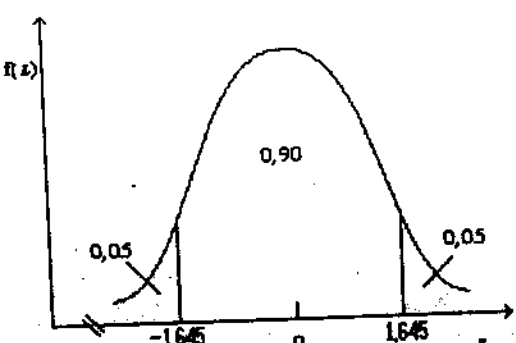
Do vậy $\quad P(Z > 1{,}645) = 0{,}05$

$$P(Z < -1{,}645) = 0{,}05$$

Xác suất để Z nằm trong khoảng (-1,645, 1,645):

$$P(-1{,}645 < Z < 1{,}645) = 1 - P(Z > 1{,}645) - P(Z < -1{,}645)$$
$$= 1 - 0{,}05 - 0{,}05 = 0{,}90$$

Việc tính toán xác suất được minh họa bằng hình vẽ sau:



Hình 6.2 P(-1,645 < Z < 1,645) = 0,9 với Z là biến ngẫu nhiên có phân phối chuẩn.

$$0{,}9 = P(-1{,}645 < Z < 1{,}645) \text{ thay } Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}, \text{ ta có :}$$

$$0{,}9 = P(-1{,}645 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1{,}645)$$

$$= P(-\frac{1,645\sigma}{\sqrt{n}} < \overline{X} - \mu < \frac{1,645\sigma}{\sqrt{n}})$$

$$= P(\overline{X} - \frac{1,645\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{1,645\sigma}{\sqrt{n}})$$

để đơn giản, ta có thể viết, với độ tin cậy 90% trung bình của tổng thể được ước lượng:

$$\overline{x} - \frac{1,645\sigma}{\sqrt{n}} < \mu < \overline{x} + \frac{1,645\sigma}{\sqrt{n}}$$

+ σ^2 chưa biết, khi đó ta thay σ^2 bằng s^2 (phương sai mẫu hiệu chỉnh).

b) $n < 30$

+ tổng thể chung phân phối chuẩn, σ^2 đã biết khi đó:

$$\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

+ Tổng thể chung phân phối chuẩn, σ^2 chưa biết khi đó:

$$\overline{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \overline{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \qquad (6.2)$$

Trong đó $T_{n-1} = \frac{\overline{X} - \mu}{s/\sqrt{n}}$ có phân phối student với n -1 bậc tự do.

cho trước 1 - α và biết n ta tìm được $t_{n-1,\alpha/2}$ theo bảng lập sẵn.

Ví dụ: Để ước lượng tuổi thọ trung bình của một loại sản phẩm, nhân viên kỹ thuật chọn 40 sản phẩm một cách ngẫu nhiên từ kho sản phẩm. Kết quả kiểm tra cho thấy tuổi thọ trung bình là 200 giờ; $s^2 = 5776$. Giả sử rằng tuổi thọ của sản phẩm có phân phối chuẩn, hãy ước lượng tuổi thọ trung bình của sản phẩm trên với độ tin cậy là 95%.

Giải

Ta có: $n = 40$; $s = \sqrt{5776} = 76$; $\alpha = 5\%$; $z_{\alpha/2} = z_{0,025} = 1,96$; $\overline{x} = 200$.

(Với $\alpha = 5\%$ thì $z_{\alpha/2} = z_{0,025}$ =**NORMSINV(1-0.05/2)** = 1,959963 ≈ 1,96)

Với độ tin cậy 95%, tuổi thọ trung bình của sản phẩm được xác định:

$$\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$200 - 1{,}96\frac{76}{\sqrt{40}} \le \mu \le 200 + 1{,}96\frac{76}{\sqrt{40}}$$

$$176{,}4473 \le \mu \le 223{,}5526$$

Nếu gọi $\varepsilon_x = z_{\alpha/2}\frac{s}{\sqrt{n}}$ ta có thể dùng hàm **CONFIDENCE** trong EXCEL để tính.

Ta có : ε_X =**CONFIDENCE(α,s,n)**

Theo ví dụ trên ε_X =**CONFIDENCE(0.05,76,40)** = 23,55219684.

Nghĩa là, với độ tin cậy 95%, tuổi thọ trung bình của sản phẩm nằm trong khoảng từ 176,4473 đến 223,5526 (giờ).

Ví dụ: Một chi nhánh điện lực thực hiện một nghiên cứu để ước lượng sản lượng điện sử dụng trung bình của các hộ gia đình trong một tháng. Một mẫu gồm 15 hộ gia đình được chọn ngẫu nhiên, kết quả cho thấy sản lượng điện tiêu thụ trung bình hàng tháng của mỗi hộ là 395 kwh và $s^2 = 120$. Giả thiết sản lượng điện tiêu thụ của các hộ gia đình là đại lượng ngẫu nhiên có phân phối chuẩn. Hãy ước lượng sản lượng điện tiêu thụ trung bình của một hộ gia đình ở chi nhánh đó với độ tin cậy 95%.

Giải

Với độ tin cậy 95% sản lượng điện tiêu thụ trung bình của một hộ gia đình được xác định:

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

với : $\bar{x} = 395$; $s = \sqrt{120} = 10{,}9544$; $t_{14,2.5\%} = 2{,}145$

(Với $\alpha = 5\%$, n=15 thì $t_{n-1,\alpha/2} = t_{14,2.5\%}$ =**TINV(0.05,14)** = 2,144789≈ 2,145)

$$389 \le \mu \le 401$$

Nghĩa là với độ tin cậy 95%, sản lượng điện tiêu thụ trung bình của một hộ gia đình nằm trong khoảng từ 389 đến 401 (kwh).

6.2.2 Ước lượng tỷ lệ tổng thể [1]:

Trong thực tế nhiều khi ta quan tâm đến tỷ lệ các đơn vị có một tính chất nào đó trong tổng thể chung. Ví dụ: tỷ lệ trẻ em bỏ học, tỷ lệ người mù chữ, tỷ lệ sản phẩm hỏng vv... tức là ta có nhu cầu ước lượng tỷ lệ p của tổng thể chung. Ta biết với n khá lớn, $\hat{P} \sim N(p, \frac{pq}{n})$; cho 1-α ta tìm $z_{\alpha/2}$

Tỷ lệ tổng thể chung p được xác định :

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (6.3)$$

Ví dụ: Một công ty kinh doanh gas thực hiện một nghiên cứu để ước lượng tỷ lệ các hộ gia đình có sử dụng gas làm chất đốt. Kết qủa điều tra mẫu ngẫu nhiên 50 hộ gia đình cho thấy có 35 hộ sử dụng gas làm chất đốt. Với độ tin cậy 95% hãy ước lượng tỷ lệ hộ gia đình sử dụng gas làm chất đốt.

Giải:

Với độ tin cậy 95%, tỷ lệ các hộ gia đình có sử dụng gas làm chất đốt được xác định:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ta có: $\hat{p} = \frac{35}{50} = 0,7; n = 50, z_{\alpha/2} = 1,96$

$$0,7 - 1,96\sqrt{\frac{0,7.0,3}{50}} \le p \le 0,7 + 1,96\sqrt{\frac{0,7.0.3}{50}}$$

$$0,5730 \le p \le 0,827$$

Nghĩa là với độ tin cậy 95%, tỷ lệ hộ có sử dụng gas nằm trong khoảng từ 57,3% đến 82,7%.

6.2.3. Ước lượng phương sai của tổng thể [2]:

Giả sử tổng thể chung có phân phối chuẩn với σ^2 chưa biết, căn cứ vào mẫu gồm n đơn vị ta đưa ra σ_1^2 và σ_2^2 sao cho σ_1^2 và σ_2^2 là các đại lượng ngẫu nhiên:

---

[1] Confidence intervals for the population proportion (large samples)

[2] Confidence intervals for the variance of a normal population

$$P(\sigma_1^2 \le \sigma^2 \le \sigma_2^2) = 1 - \alpha \qquad (6.4)$$

Với : $\sigma_1^2 = \dfrac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}$ $\qquad \sigma_2^2 \dfrac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}$

Trong đó χ^2_{n-1} có phân phối χ^2 với n -1 bậc tự do.

Ví dụ: Một công ty muốn nghiên cứu sự biến thiên về tuổi thọ của một loại sản phẩm. Chọn ngẫu nhiên 15 sản phẩm ta tính được $s^2 = 15,274$. Hãy ước lượng phương sai của tuổi thọ sản phẩm với độ tin cậy 95% (giả thiết tuổi thọ sản phẩm có phân phối chuẩn).

Giải

Ta có $n = 15$, $1 - \alpha = 95\% \Rightarrow \alpha = 5\%$; $\alpha/2 = 2,5\%$

$$\chi^2_{14;0,025} = 26,12 \quad ; \quad \chi^2_{14;0,975} = 5,63$$

$$\sigma_1^2 = \frac{14.15,274}{26,12} = 8,1867$$

$$\sigma_2^2 = \frac{14.15,274}{5,63} = 37,9815$$

$$8,1867 \le \sigma^2 \le 37,9815$$

Với độ tin cậy 95% phương sai về tuổi thọ sản phẩm nằm trong khoảng từ 8,1867 đến 37,9815.

6.2.4. Ước lượng sự khác biệt giữa 2 số trung bình của hai tổng thể[1]:

Trong thực tế nhiều khi ta quan tâm đến sự khác biệt về năng suất lúa trung bình do sử dụng hai loại phân bón khác nhau, hoặc trọng lượng trung bình của 1 con gia súc ở hai giống khác nhau v.v...

Phương pháp so sánh trung bình của hai tổng thể phụ thuộc vào cách thức lấy mẫu: Mẫu phối hợp từng cặp (mẫu phụ thuộc) hoặc mẫu độc lập.

- Mẫu phối hợp từng cặp: Trong phương pháp này các đơn vị mẫu được chọn từng cặp từ hai tổng thể và phụ thuộc nhau. Tính phụ thuộc thể hiện ở các khía cạnh sau:

[1] Confidence intervals for the difference between the means of two normal populations

+ So sánh theo thời gian: Ví dụ: Mẫu thứ nhất: doanh số bán trước khi thực hiện khuyến mãi; mẫu thứ hai: doanh số bán sau khi thực hiện khuyến mãi. Tính phụ thuộc thể hiện ở chỗ từng cặp doanh số được thu thập ở cùng một cửa hàng. Hoặc mẫu thứ nhất là doanh số bán của cửa hàng A ở một tháng nào đó, mẫu thứ hai : doanh số của cửa hàng B cũng ở tháng đó. Tính phụ thuộc thể hiện ở chỗ từng cặp doanh số được thu thập trong cùng một tháng.

+ So sánh theo không gian: Ví dụ: Mẫu thứ nhất: doanh số của mặt hàng A ở 10 cửa hàng, mẫu thứ hai: doanh số của mặt hàng B cũng ở 10 cửa hàng trên. Tính phụ thuộc thể hiện ở chỗ từng cặp doanh số của 2 mặt hàng A và B đều được thu thập từ cùng một cửa hàng.

- Mẫu độc lập: Trong phương pháp này, các giá trị quan sát của mẫu được chọn ngẫu nhiên độc lập từ hai tổng thể không phụ thuộc nhau.

6.2.4.1. Ước lượng sự khác biệt giữa 2 số trung bình của hai tổng thể trong trường hợp mẫu phối hợp từng cặp [1]:

Giả sử ta có mẫu gồm n cặp quan sát lấy ngẫu nhiên từ hai tổng thể X và Y: (x_1, y_1); (x_2, y_2)... (x_n, y_n)

Gọi μ_X, μ_Y là trung bình của hai tổng thể.

$\bar{d}$: là trung bình của n khác biệt $(x_i - y_i)$.

S_d: là độ lệch tiêu chuẩn của n khác biệt $(x_i - y_i)$.

Giả sử rằng sự khác biệt giữa x và y trong tổng thể có phân phối chuẩn.

Với độ tin cậy 1 - α, chênh lệch giữa 2 số trung bình $\mu_X - \mu_Y$ được xác định:

$$\bar{d} - t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}} \le \mu_X - \mu_Y \le \bar{d} + t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}} \qquad (6.5)$$

Biết 1 - α, n, ta tìm $t_{n-1,\alpha/2}$ trong bảng phân phối student.

Ví dụ: Một công ty thực hiện các biện pháp tăng năng suất lao động. Số liệu về năng suất của 10 công nhân được thu thập trước và sau khi thực hiện các biện pháp tăng năng suất lao động được cho trong bảng sau.

---

[1] Confidence intervals based on matched pairs

Công nhân	NSLĐ trước và sau khi thực hiện các biện pháp tăng NSLĐ (kg/ngày)		$d_i = x_i - y_i$
	Trước khi (x_i)	Sau khi(y_i)	
A	50	52	-2
B	48	46	2
C	45	50	-5
D	60	65	-5
E	70	78	-8
F	62	61	1
G	55	58	-3
H	62	70	-8
I	58	67	-9
K	53	65	-12

Giả sử rằng các khác biệt giữa NSLĐ trước và sau khi áp dụng các biện pháp tăng NLSĐ có phân phối chuẩn. Hãy ước lượng sự khác biệt về NSLĐ trung bình trước và sau khi thực hiện các biện pháp tăng NSLĐ với độ tin cậy là 95%.

Dựa vào số liệu trên ta tính được: $\bar{d} = \frac{1}{n}\sum d_i$

$$\bar{d} = \frac{-2+2-5-5-8+1-3-8-9-12}{10} = \frac{-49}{10} = -4,9$$

$$s_d = \sqrt{\frac{1}{n-1}\sum(d_i - \bar{d})^2} = \sqrt{\frac{180,9}{10-1}} = 4,4833$$

Như vậy với độ tin cậy 95%, sự khác biệt về năng suất lao động trung bình trước và sau khi thực hiện các biện pháp tăng năng suất lao động được xác định:

$$\bar{d} - t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}} \le \mu_x - \mu_y \le \bar{d} + t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}}$$

Với: $n = 10$; $\bar{d} = -4,9; S_d = 4,4833$; $t_{n-1,\alpha/2} = t_{9,0,025} = 2,262$

$$-4,9 - 2,262\frac{4,4833}{\sqrt{10}} \le \mu_x - \mu_y \le -4,9 + 2,262\frac{4,4833}{\sqrt{10}}$$

$$-8,1069 \le \mu_x - \mu_y \le -1,6931 \text{ (kg)}$$

$$1,6931 \le \mu_y - \mu_x \le 8,1069 \text{(kg)}$$

Với độ tin cậy 95%, có thể nói các biện pháp tăng NSLĐ đã làm tăng NSLĐ trung bình của 1 công nhân từ 1,693 đến 8,1069 kg/ngày.

6.2.4.2. Ước lượng sự khác biệt giữa 2 số trung bình của hai tổng thể trong trường hợp mẫu độc lập [1]:

Gọi n_x, n_y là các mẫu được chọn ngẫu nhiên độc lập từ hai tổng thể có phân phối chuẩn X và Y có trung bình là μ_x, μ_y, phương sai σ_x^2, σ_y^2, trung bình mẫu $\bar{x}, \bar{y}$.

Với độ tin cậy 1 - α, thì $\mu_x - \mu_y$ được xác định như sau:

$$(\bar{x} - \bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \le \mu_x - \mu_y \le (\bar{x} - \bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \quad (6.6)$$

Trong trường hợp cỡ mẫu lớn $n_x, n_y \ge 30$ ta có thể dùng các phương sai mẫu hiệu chỉnh s_x^2, s_y^2 thay cho phương sai tổng thể σ_x^2, σ_y^2.

Ví dụ: Một trại chăn nuôi tiến hành nghiên cứu hiệu quả của hai loại thức ăn mới A và B. Sau một thời gian nuôi thử nghiệm người ta chọn 50 con gà nuôi bằng thức ăn A và thấy trọng lượng trung bình một con là 1,9 kg; độ lệch chuẩn là 1,25 kg và chọn 40 con gà nuôi bằng thức ăn B, trọng lượng trung bình một con là 1,2 kg; độ lệch chuẩn là 1,02 kg. Hãy ước lượng sự khác biệt về trọng lượng trung bình của 1 con gà đối với 2 loại thức ăn A và B với độ tin cậy là 95%.

[1] Confidence intervals based on independent samples

Giải

Ta có :

$$\bar{x} = 1,9 \ ; s_x = 1,25 \ ; \ n_x = 50 \ ; z_{\alpha/2} = z_{0,025} = 1,96$$

$$\bar{y} = 1,2 \ ; s_y = 1,02 \ ; \ n_y = 40$$

Với độ tin cậy 95%, sự khác biệt về trọng lượng trung bình của 1 con gà đối với 2 loại thức ăn A và B được xác định:

$$(1,9 - 1,2) - 1,96\sqrt{\frac{1,25^2}{50} + \frac{1,02^2}{40}} \le \mu_x - \mu_y$$

$$\le (1,9 - 1,2) + 1,96\sqrt{\frac{1,25^2}{50} + \frac{1,02^2}{40}}$$

$$0,7 - 0,469 \le \mu_x - \mu_y \le 0,7 + 0,469$$

$$0,231 \le \mu_x - \mu_y \le 1,169 \ \text{(kg)}$$

Có nghĩa là với độ tin cậy 95%, trọng lượng trung bình của 1 con gà nuôi bằng loại thức ăn A sẽ tăng hơn so với nuôi bằng thức ăn B từ 0,231 kg đến 1,169 kg.

6.2.5. Ước lượng sự khác biệt giữa hai tỷ lệ tổng thể [1]:

Ta có n_x, n_y là hai mẫu được chọn ngẫu nhiên độc lập từ hai tổng thể X và Y. Gọi p_x, p_y và $\hat{p}_x$, $\hat{p}_y$ lần lượt là tỷ lệ các đơn vị có tính chất nào đó mà ta quan tâm trong tổng thể và trong mẫu. Với mẫu lớn n_x, $n_y \ge 40$, sự khác biệt giữa tỷ lệ hai tổng thể X và Y với độ tin cậy 1 - α được xác định:

$$(\hat{p}_x - \hat{p}_y) - z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \le p_x - p_y \le (\hat{p}_x - \hat{p}_y) + z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \qquad (6.7)$$

Ví dụ: Một công ty đang xem xét việc ứng dụng một phương pháp sản xuất mới nhằm giảm tỷ lệ phế phẩm. Ở phương pháp sản xuất mới người ta

---

[1] Confidence intervals for the difference between two population proportions (large samples)

chọn ngẫu nhiên ra 500 sản phẩm thì thấy có 10 phế phẩm. Ở phương pháp sản xuất cũ người ta chọn ngẫu nhiên 400 sản phẩm thì thấy có 7 phế phẩm. Với độ tin cậy 95% hãy ước lượng sự khác biệt về tỷ lệ phế phẩm ở hai phương pháp sản xuất.

Giải

Ta có:

$$\hat{p}_x = \frac{10}{500} = 0{,}02 \qquad \hat{p}_y = \frac{7}{400} = 0{,}0175$$

$$n_x = 500\ ,\ n_y = 400,\ z_{\alpha/2} = z_{0{,}025} = 1{,}96$$

Ta tính:

$$z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

$$= 1{,}96\sqrt{\frac{0{,}02(1-0{,}02)}{500} + \frac{0{,}0175(1-0{,}0175)}{400}} = 0{,}017768$$

Với độ tin cậy 95% sự khác biệt về tỷ lệ phế phẩm giữa hai phương pháp sản xuất mới và cũ là:

$$(0{,}02 - 0{,}0175) - 0{,}017768 \le p_x - p_y \le (0{,}02 - 0{,}0175) + 0{,}017768$$

$$-0{,}0153 \le p_x - p_y \le 0{,}0203$$

$$-1{,}53\% \le p_x - p_y \le 2{,}03\%$$

Như vậy với độ tin cậy 95% chênh lệch về tỷ lệ phế phẩm của phương pháp sản xuất mới so với phương pháp sản xuất cũ nằm trong khoảng từ -1,53% đến 2,03%, vì khoảng ước lượng có chứa giá trị 0 nên có khả năng xảy ra tỷ lệ phế phẩm của 2 phương pháp sản xuất bằng nhau, tức là $p_x - p_y = 0$, do đó ta không thể kết luận phương pháp sản xuất nào có tỷ lệ phế phẩm thấp hơn.

6.2.6. Ước lượng một bên:

Các ước lượng khoảng ta đã nghiên cứu trong các phần trên là ước lượng đối xứng:

$$P(\theta_1 \le \theta \le \theta_2) = 1 - \alpha$$

Đôi khi người ta chỉ cần tìm giới hạn tin cậy dưới θ_1, hoặc chỉ cần tìm giới hạn tin cậy trên θ_2, tức là ta có khoảng ước lượng một bên:

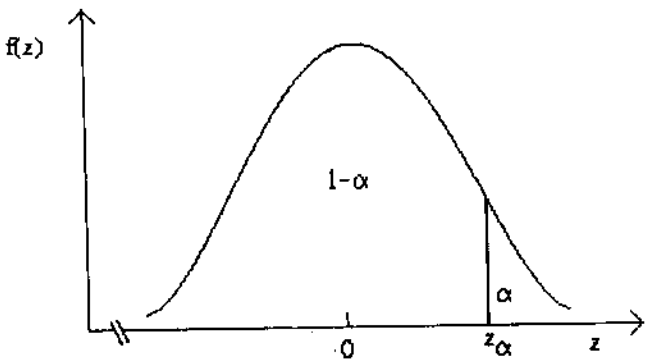
$$P(\theta_1 \le \theta) = 1 - \alpha \quad \text{hoặc} \quad P(\theta \le \theta_2) = 1 - \alpha$$

(ước lượng bên trái) (ước lượng bên phải)

Chẳng hạn ta có ước lượng bên phải:

$$P(\mu \le \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}) = 1 - \alpha \qquad (6.8)$$

(để ý rằng trong ước lượng một bên ta thay $z_{\alpha/2} = z_\alpha$)



Hình 6.3 Ước lượng bên phải

Ví dụ: Một công ty muốn ước lượng giới hạn dưới cho lượng nhiên liệu tiêu thụ trung bình hàng ngày. Một mẫu ngẫu nhiên được chọn gồm 40 ngày cho thấy lượng nhiên liệu tiêu thụ trung bình hàng ngày là 250 lít, độ lệch chuẩn s = 125 lít.

Với độ tin cậy 95% khoảng ước lượng bên trái của lượng nhiên liệu tiêu thụ trung bình hàng ngày được ước lượng:

$$\overline{x} - z_\alpha \frac{s}{\sqrt{n}} \le \mu$$

$$250 - 1{,}645 \frac{125}{\sqrt{40}} \le \mu$$

$$217{,}49 \text{ lít} \le \mu$$

Như vậy với độ tin cậy 95% lượng nhiên liệu tiêu thụ trung bình hàng ngày không thấp hơn 217,49 lít.

CHƯƠNG 7

ĐIỀU TRA CHỌN MẪU [1]

Để thu thập tài liệu ban đầu, hiện nay ngành thống kê thực hiện hai hình thức: báo cáo thống kê định kỳ và điều tra chuyên môn. Chế độ báo cáo thống kê định kỳ áp dụng chủ yếu đối với thành phần kinh tế quốc doanh, các doanh nghiệp. Điều tra chuyên môn được áp dụng để thu thập thông tin đối với những hiện tượng và quá trình kinh tế xã hội không thể hoặc không nhất thiết phải thực hiện chế độ báo cáo thống kê định kỳ. Điều tra chuyên môn có thể tiến hành trên tất cả các đơn vị của hiện tượng nghiên cứu, gọi là điều tra toàn bộ hoặc chỉ tiến hành trên một số đơn vị thuộc đối tượng nghiên cứu gọi là điều tra không toàn bộ. Điều tra không toàn bộ gồm một số loại chủ yếu như điều tra chọn mẫu, điều tra trọng điểm, điều tra chuyên đề.

7.1 KHÁI NIỆM VỀ ĐIỀU TRA CHỌN MẪU:

7.1.1 Khái niệm: Điều tra chọn mẫu là một loại điều tra không toàn bộ, người ta chỉ chọn ra một số đơn vị từ tổng thể chung để điều tra thực tế, rồi sau đó bằng các phương pháp khoa học, tính toán suy rộng cho toàn bộ tổng thể.

Như vậy trong điều tra chọn mẫu người ta đặc biệt lưu ý tới nai vấn đề cơ bản:

- Quy tắc lựa chọn các đơn vị sao cho có thể đại diện cho toàn bộ tổng thể.

- Dùng công thức suy rộng thành các đặc điểm của tổng thể.

Cơ sở khoa học của phương pháp chọn mẫu là lý thuyết xác suất và thống kê toán. Lý thuyết xác suất và thống kê toán đã chứng minh là bằng phương pháp điều tra chọn mẫu ta có thể biết được các tham số của tổng thể theo một đặc trưng nào đó với một mức độ chính xác, mức độ tin cậy tính toán được. Như vậy dựa trên cơ sở khoa học này ta thấy phương pháp điều tra chọn mẫu hoàn toàn có thể thay thế được điều tra toàn bộ trong một số trường hợp.

7.1.2 Ưu điểm và hạn chế của điều tra chọn mẫu.

Trong điều tra chọn mẫu, người ta chỉ thực hiện điều tra trên một bộ phận của tổng thể. Do đó so với điều tra toàn bộ, điều tra chọn mẫu có các ưu điểm chủ yếu trên các mặt sau:

---

[1] Survey sampling methods

- Về chi phí: Do số đơn vị điều tra ít, điều tra chọn mẫu tiết kiệm được khá nhiều chi phí.
- Về thời gian: Tiến độ công việc nhanh hơn có thể đáp ứng được tính khẩn cấp của thông tin cần thu thập.
- Về tính chính xác: Với các phương pháp suy luận thống kê khoa học, thông qua nghiên cứu mẫu vẫn có thể đi đến các kết luận đáng tin cậy mà không cần nghiên cứu toàn bộ tổng thể.

Hơn nữa, nhiều khi điều tra toàn bộ do các điều kiện ràng buộc về chi phí và thời gian, lại đưa đến thông tin không kịp thời và kém chính xác.

Tuy nhiên điều tra chọn mẫu không hoàn toàn có thể thay thế được điều tra toàn bộ vì những lý do sau:

- Trong điều tra toàn bộ, người ta thu thập thông tin trên từng đơn vị tổng thể, do đó có thể nghiên cứu tổng thể và các bộ phận của nó theo tất cả các đặc trưng cần nghiên cứu. Chính vì vậy đối với những nguồn thông tin thống kê quan trọng người ta vẫn phải tiến hành điều tra toàn bộ tức là các cuộc tổng điều tra, ví dụ tổng điều tra dân số, tổng điều tra chăn nuôi...

- Kết quả suy rộng từ mẫu điều tra bao giờ cũng có sai số đại diện nhất định, mà loại sai số này không có trong điều tra toàn bộ.

Điều tra chọn mẫu thường được dùng trong những trường hợp sau đây:

- Dùng để thay thế điều tra toàn bộ. Khi tổng thể nghiên cứu vừa cho phép điều tra chọn mẫu vừa cho phép điều tra toàn bộ thì người ta thường quyết định dùng điều tra chọn mẫu vì những ưu điểm của nó. Khi tổng thể nghiên cứu không cho phép điều tra toàn bộ, đó là khi tổng thể quá lớn hoặc không xác định trước được (như toàn bộ cây lấy gỗ trong một cánh rừng, số trẻ em mới sinh...) hoặc khi điều tra phải phá hủy sản phẩm (bóng đèn, đồ hộp...) thì người ta thường sử dụng điều tra chọn mẫu.

- Dùng để tổng hợp nhanh tài liệu điều tra toàn bộ. Trong các cuộc tổng điều tra, chẳng hạn như tổng điều tra dân số, thường phải tổng hợp rất nhiều chỉ tiêu khác nhau. Muốn hoàn thành khối lượng công việc như vậy phải mất rất nhiều thời gian. Để có thông tin nhanh phục vụ cho công tác quản lý có thể dùng đến điều tra chọn mẫu.

- Điều tra chọn mẫu có thể ứng dụng rộng rãi trong các lĩnh vực nghiên cứu kinh tế xã hội như tình hình thu nhập và chi tiêu của các hộ gia đình, mức sống của các tầng lớp dân cư, nhu cầu tiêu dùng các loại hàng hóa, giá cả thị trường...

7.1.3 Sai số trong điều tra chọn mẫu.

Trong các cuộc điều tra chọn mẫu, sai số toàn bộ bao gồm: sai số chọn mẫu[1] và sai số phi chọn mẫu[2].

- Sai số chọn mẫu còn được gọi là sai số đại diện, tồn tại ngay trong bản thân cuộc điều tra chọn mẫu, bởi vì việc điều tra chỉ thực hiện trên một số ít đơn vị, nhưng kết quả thu được lại được tính toán suy rộng cho toàn bộ tổng thể. Sai số chọn mẫu là điều khó tránh khỏi vì dù cho có tổ chức khoa học chu đáo đến đâu, thì việc lấy ra một tổng thể mẫu có kết cấu giống như kết cấu của tổng thể chung là điều khó thực hiện, mà chỉ cần có sự sai khác nhỏ về kết cấu của hai tổng thể là đã phát sinh sai số rồi.

Sai số chọn mẫu có thể giảm bằng cách tăng quy mô của mẫu. Khi quy mô của mẫu tăng đến mức bằng quy mô tổng thể chung thì sai số chọn mẫu sẽ biến mất.

- Loại sai số thứ hai xuất hiện trong cả điều tra chọn mẫu lẫn trong điều tra toàn bộ, được gọi là sai số phi chọn mẫu. Việc lập danh sách tất cả các nguồn sai số phi chọn mẫu là rất khó. Những sai số này xảy ra do nhiều nguyên nhân:

+ Do đơn vị điều tra trả lời sai vì không hiểu đúng nội dung, hoặc do cố ý khai sai.

+ Do nhân viên điều tra vô tình ghi chép sai.

+ Do tỷ lệ không trả lời qúa cao

+ Do đo lường sai...

Rõ ràng rằng, với một đội ngũ nhân viên được huấn luyện tốt ở cả hai lĩnh vực thu thập và xử lý số liệu, nên các sai số phi chọn mẫu ở các cuộc điều tra chọn mẫu có thể ít nghiêm trọng hơn so với ở các cuộc điều tra toàn bộ.

Nói chung khái niệm sai số chọn mẫu thường được hiểu là sai số ngẫu nhiên. Như vậy đối với mỗi tổng thể mẫu cụ thể được chọn ra một cách ngẫu nhiên từ tổng thể chung sẽ có một trị số cụ thể của sai số và ta có khái niệm sai số trung bình chọn mẫu.

---

[1] Sampling error

[2] Nonsampling error

* Khi nhiệm vụ chọn mẫu là để ước lượng số trung bình về một tiêu thức nào đó, tức là khi mẫu được chọn ngẫu nhiên, giá trị trung bình sẽ khác nhau từ tổng thể mẫu này sang tổng thể mẫu khác. Độ lệch tiêu chuẩn của các giá trị trung bình mẫu dùng để đo lường độ biến thiên giữa các giá trị trung bình mẫu với giá trị trung bình của tổng thể chung gọi là sai số trung bình mẫu[1] (sai số chuẩn) ký hiệu $\sigma_{\bar{x}}$ được xác định theo công thức:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \qquad (7.1)$$

(Nếu σ^2 chưa biết ta thay bằng s^2)

* Khi nhiệm vụ chọn mẫu là để ước lượng tỷ lệ theo một tiêu thức nào đó, sai số trung bình chọn mẫu sẽ là:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad (7.2)$$

(Nếu p chưa biết ta thay bằng $\hat{p}$)

Trong trường hợp chọn không hoàn lại sai số trung bình chọn mẫu sẽ nhân cho hệ số điều chỉnh tổng thể hữu hạn $fpc^2 = \sqrt{1 - \frac{n}{N}}$

Gọi ε là phạm vi sai số chọn mẫu.

- Khi nhiệm vụ chọn mẫu là để ước lượng số trung bình về một tiêu thức nào đó thì:

$$\varepsilon_x = z_{\alpha/2}\sigma_{\bar{x}} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (7.3)$$

- Khi nhiệm vụ chọn mẫu là để ước lượng tỷ lệ theo một tiêu thức nào đó thì

$$\varepsilon_p = z_{\alpha/2}\sigma_{\hat{p}} = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \qquad (7.4)$$

7.2 CÁC BƯỚC CỦA QUÁ TRÌNH NGHIÊN CỨU MẪU:

Khi nghiên cứu mẫu ta thường trải qua các bước sau:

---

[1] Standard error

[2] Finite population correction factor

Bước 1: Xác định mục đích nghiên cứu.

Do nhu cầu thực tế ta cần thông tin về một hiện tượng nào đó. Nếu thông tin đã có sẵn hoặc không có khả năng thu thập được, không có lý do gì để tiến hành quá trình chọn mẫu. Việc xác định rõ mục đích của quá trình chọn mẫu là hết sức quan trọng, nó quyết định đến việc lựa chọn các phương pháp lấy mẫu, cũng như kích thước mẫu cần lấy. Đây là bước khởi đầu ảnh hưởng đến các bước sau.

Bước 2: Xác định tổng thể có liên quan.

Rõ ràng là nếu thông tin của mẫu được dùng để suy rộng cho tổng thể nào thì mẫu phải được lấy ra từ tổng thể đó. Điều này tưởng chừng đơn giản nhưng thực tế có khá nhiều kết luận không có giá trị chỉ vì nguyên tắc cơ bản này không được chú ý. Do vậy trước khi tiến hành lấy mẫu người ta thường lập dàn chọn mẫu trên cơ sở đã xác định rõ phạm vi, tính chất của tổng thể phù hợp với mục đích nghiên cứu.

Bước 3: Xác định kích thước mẫu (cỡ mẫu).

Sau khi đã xác định được tổng thể có liên quan để từ đó chúng ta tiến hành lấy mẫu. Vấn đề quan trọng tiếp theo là nên chọn bao nhiêu đơn vị mẫu – tức là xác định kích thước mẫu. Việc xác định kích thước mẫu phụ thuộc vào nhiều yếu tố, sẽ được trình bày trong phần kế tiếp.

Bước 4: Lựa chọn phương pháp thu thập thông tin.

Trong việc thu thập thông tin từ các đơn vị mẫu, có hai điểm cần quan tâm.

Thứ nhất là tỷ lệ nhận được các câu trả lời. Số người trả lời càng cao ta càng nhận được nhiều thông tin về tổng thể, do đó mẫu chúng ta càng có tính đại diện cao. Để tăng tỷ lệ trả lời chúng ta cần thiết kế câu hỏi thích hợp, giải thích rõ mục đích của cuộc điều tra, làm cho đối tượng điều tra cảm thấy an tâm khi trả lời. Cần có chi phí bồi dưỡng dành cho đối tượng điều tra. Việc thu thập số liệu thông qua đường bưu điện thường có tỷ lệ hồi đáp thấp hơn so với phỏng vấn trực tiếp hoặc thông qua điện thoại. Trong thực tế, tùy theo yêu cầu đòi hỏi về độ chính xác của số liệu mà ta chọn phương pháp thu thập thông tin phù hợp.

Thứ hai là sự chính xác và thành thật của các câu trả lời. Những kết luận được rút ra từ các phương pháp thống kê vô cùng phức tạp nhưng lại dựa trên số liệu không đáng tin cậy sẽ trở thành vô nghĩa. Vấn đề ở đây là các câu hỏi phải được thiết kế rõ ràng, dễ hiểu, nhất là đối với những vấn đề nhạy cảm, tế nhị. Nói chung việc phỏng vấn để thu thập số liệu phải

được nâng lên thành nghệ thuật, nghệ thuật hỏi sao cho nhận được các câu trả lời thành thật, chính xác.

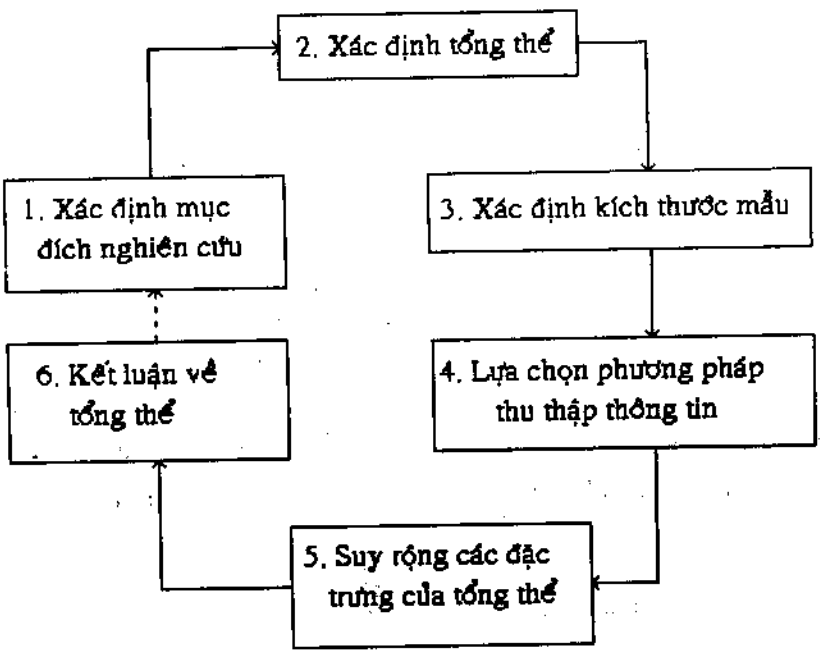
Bước 5: Suy rộng các đặc trưng mẫu thành các đặc trưng tổng thể.

Mẫu sau khi được điều tra, tiếp tục xử lý, tính toán các đặc trưng mẫu, sau đó sử dụng các phương pháp thống kê để suy rộng thành các đặc trưng của tổng thể. Cần hiểu rõ bản chất, nội dung của vấn đề nghiên cứu để chọn phương pháp thống kê phù hợp.

Bước 6: Rút ra các kết luận về tổng thể.

Đây là bước cuối cùng của qúa trình nghiên cứu mẫu. Ta xem xét các kết luận rút ra từ mẫu có thỏa mãn các yêu cầu đặt ra khi bắt đầu nghiên cứu, nghĩa là đối chiếu lại với bước 1 để xem mục đích của việc chọn mẫu có thoả mãn hay không.

Qúa trình nghiên cứu mẫu có thể được minh họa bằng sơ đồ sau:



Hình 7.1 Các bước trong một nghiên cứu mẫu

7.3 XÁC ĐỊNH KÍCH THƯỚC MẪU (CỠ MẪU) [1]:

Việc xác định kích thước mẫu nằm ở bước thứ 3 trong quá trình nghiên cứu mẫu, là một trong các công việc chuẩn bị đầu tiên, tức là ta quyết định xem nên chọn bao nhiêu đơn vị mẫu từ tổng thể.

Kích thước mẫu n phụ thuộc vào các yếu tố sau:

- Phương pháp chọn mẫu sẽ được tiến hành theo phương pháp nào để sử dụng công thức xác định kích thước mẫu phù hợp.
- Xác định phạm vi sai số có thể chấp nhận được (ε).
- Quy định độ tin cậy muốn có trong ước lượng.
- Xác định hệ số tin cậy z từ độ tin cậy mong muốn.
- Ước tính độ lệch tiêu chuẩn của tổng thể.

Ngoài các yếu tố trên còn một yếu tố không kém phần quan trọng ảnh hưởng đến kích thước mẫu đó là kinh phí dành cho cuộc điều tra mẫu. Ở đây chúng ta giả định rằng kinh phí chúng ta dồi dào đủ để điều tra n đơn vị mẫu được xác định thông qua các yếu tố trên.

7.3.1 Các công thức xác định kích thước mẫu (n) :

Yếu tố ảnh hưởng đầu tiên đến kích thước mẫu là ta quyết định chọn mẫu theo phương pháp nào, vì ứng với các phương pháp chọn mẫu khác nhau ta sẽ có kích thước mẫu khác nhau. Có thể nói rằng không có phương pháp chọn mẫu nào là tối ưu cả. Tùy theo mục đích, tính chất, thời gian, kinh phí mà ta sẽ chọn phương pháp chọn mẫu phù hợp.

Giả sử chúng ta có ý định chọn mẫu theo phương pháp chọn ngẫu nhiên đơn giản, công thức xác định kích thước mẫu n được tính như sau:

- Khi nhiệm vụ nghiên cứu là để ước lượng số trung bình theo một tiêu thức nào đó.

+ Trường hợp chọn hoàn lại: $$n = \frac{z^2_{\alpha/2}\sigma^2}{\varepsilon^2_x} \qquad (7.1)$$

+ Trường hợp chọn không hoàn lại: $$n = \frac{z^2_{\alpha/2}\sigma^2 N}{\varepsilon^2_x N + z^2_{\alpha/2}\sigma^2} \qquad (7.2)$$

- Khi nhiệm vụ nghiên cứu là để ước lượng tỷ lệ theo một tiêu thức nào đó:

[1] Estimating the sample size

+ Trường hợp chọn hoàn lại: $n = \frac{z^2_{\alpha/2}pq}{\varepsilon^2_p}$ (7.3)

+ Trường hợp chọn không hoàn lại: $n = \frac{z^2_{\alpha/2}pqN}{\varepsilon^2_p N + z^2_{\alpha/2}pq}$ (7.4)

Các công thức trên được suy ra từ công thức xác định phạm vi sai số chọn mẫu: ε_x hoặc ε_p.

7.3.2 Xác định phạm vi sai số có thể chấp nhận được (ε):

Yếu tố ảnh hưởng đầu tiên đến kích thước mẫu là độ lớn của phạm vi sai số. Độ lớn của phạm vi sai số được xác định căn cứ vào mục đích nghiên cứu cụ thể, kinh nghiệm nghiên cứu và khả năng nghiên cứu. Thông thường nó được xác định dựa vào kinh nghiệm của các chuyên gia thiết kế mẫu. Dựa vào các công thức trên ta thấy phạm vi sai số và kích thước mẫu tỷ lệ nghịch với nhau, nghĩa là phạm vi sai số càng nhỏ thì kích thước mẫu càng phải lớn và ngược lại. Trong thực tế, với mỗi cuộc điều tra chọn mẫu, người ta căn cứ vào nhiều mặt để xác định phạm vi sai số phù hợp.

7.3.3 Xác định độ tin cậy mong muốn từ đó xác định hệ số tin cậy:

Nếu chúng ta muốn có kết quả nghiên cứu với độ tin cậy là 100% thì phải điều tra toàn bộ các đơn vị trong tổng thể chung. Song điều này quá tốn kém và kl ông thực tế. Do vậy, thường phải chấp nhận mức tin cậy thấp hơn 100%. Trong thực tế mức tin cậy thường được sử dụng là 99%; 95% và 90%; trong đó mức tin cậy 95% được sử dụng phổ biến nhất. Mức tin cậy này cho phép kết quả nghiên cứu sai số 5% so với giá trị thực của tổng thể, và mức sai sót này thường được chấp nhận đối với phần lớn các quyết định trong nghiên cứu kinh tế - xã hội.

Từ độ tin cậy mong muốn, ta xác định được hệ số tin cậy z.

Dựa vào công thức xác định kích thước mẫu trên, ta thấy kích thước mẫu tỷ lệ thuận với hệ số tin cậy z, nghĩa là độ tin cậy càng cao thì hệ số tin cậy càng lớn và kích thước mẫu càng tăng.

7.3.4 Ước tính độ lệch tiêu chuẩn:

Trong công thức xác định kích thước mẫu ta thấy có yếu tố độ lệch tiêu chuẩn của tổng thể chung. Vì ta không điều tra toàn bộ nên ta không biết độ lệch tiêu chuẩn, do đó ta có thể ước tính độ lệch tiêu chuẩn theo các cách sau:

- Nếu trước đây đã tiến hành điều tra và được xem là tương tự với lần này thì có thể lấy độ lệch tiêu chuẩn của lần điều tra trước. Nếu trước đây đã tiến hành nhiều lần điều tra, có thể lấy độ lệch tiêu chuẩn lớn nhất. Trường hợp này áp dụng đối với những hiện tượng không có sự thay đổi lớn trong quá trình phát triển.

- Có thể sử dụng độ lệch tiêu chuẩn của cuộc điều tra tương ứng ở nơi khác, nếu hiện tượng nghiên cứu ở nơi đó cũng có những đặc điểm và điều kiện tương tự với hiện tượng ta cần nghiên cứu.

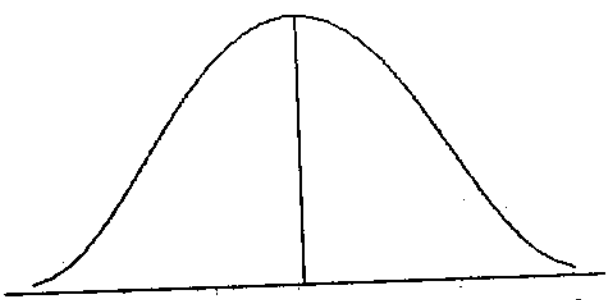
- Tiến hành điều tra thí điểm để tính độ lệch tiêu chuẩn.

- Nếu hiện tượng nghiên cứu có phân phối chuẩn thì có thể ước tính độ lệch tiêu chuẩn theo khoảng biến thiên R.

Nhớ lại theo qui tắc 3σ, nếu $X \sim N(\mu, \sigma^2)$ thì hầu hết các giá trị của X sai lệch với μ không quá 3 lần σ.

Ta có: $R = (x_{max} - x_{min}) = (\mu + 3\sigma) - (\mu - 3\sigma) = 6\sigma$

$$\Rightarrow \quad \sigma = \frac{R}{6} = \frac{x_{max} - x_{min}}{6}$$



μ-3σ μ-2σ μ-1σ μ μ+1σ μ+2σ μ+3σ

Ví dụ 1: Để xác định thu nhập trung bình trong năm của một công nhân ngành may, người ta tiến hành điều tra chọn mẫu với yêu cầu là: phạm vi sai số $\varepsilon \leq 40$ ngàn đồng; độ tin cậy 95%; độ lệch tiêu chuẩn về thu nhập ước tính được là 220 ngàn đồng. Hãy xác định cỡ mẫu cần điều tra?

Giải:

Ta có : $\varepsilon_x = 40$; $\alpha = 5\%$; $z_{\alpha/2} = z_{0,025} = 1,96$; $\sigma = 220$

$$n = \frac{z^2_{\alpha/2}\sigma^2}{\varepsilon^2_x} = \frac{1,96^2 \times 220^2}{40^2} = 116 \text{ công nhân.}$$

Ví dụ 2: Ở một tỉnh miền núi, người ta tổ chức điều tra chọn mẫu để xác định tỷ lệ trẻ em ở cấp tiểu học bỏ học, với yêu cầu phạm vi sai số $\varepsilon \leq 2\%$, độ tin cậy 95%. Ở cuộc điều tra năm trước đã xác định tỷ lệ trẻ em bỏ học của tỉnh là 8%. Hãy xác định cỡ mẫu cần điều tra?

Giải

Ta có : $\varepsilon_p = 2\%$; $p = 8\%$; $q = 92\%$; $\alpha = 5\%$; $z_{\alpha/2} = z_{0,025} = 1,96$

$$n = \frac{z^2_{\alpha/2}pq}{\varepsilon^2_p} = \frac{1,96^2.0,08.0,92}{0,02^2} = 707 \text{ em.}$$

Trường hợp nếu không biết p, song với bất kỳ giá trị nào của p thì p (1-p) không thể vượt quá 0,25 do đó n có thể xác định:

$$n = \frac{0,25z^2_{\alpha/2}}{\varepsilon^2_p} = \frac{0,25.1,96^2}{0,02^2} = 2401 \text{ người}$$

7.4 CÁC PHƯƠNG PHÁP CHỌN MẪU THƯỜNG DÙNG:

7.4.1. Chọn mẫu ngẫu nhiên đơn giản[1]:

Đây là phương pháp chọn mẫu đơn giản nhất trong các phương pháp chọn mẫu ngẫu nhiên. Các đơn vị mẫu được chọn ra từ tổng thể chung bằng cách rút thăm, quay số hoặc theo bảng số ngẫu nhiên và có thể được chọn một lần (không hoàn lại hay không lặp) hoặc chọn nhiều lần (chọn hoàn lại hay chọn lặp).

Khi tính toán sai số trung bình chọn mẫu hoặc ước lượng các tham số của tổng thể chung ta có thể sử dụng các công thức đã trình bày ở phần trên.

Phương pháp chọn ngẫu nhiên đơn giản có thể cho kết quả tốt nếu giữa các đơn vị của tổng thể không có gì khác biệt nhiều. Nếu tổng thể có kết cấu phức tạp thì chọn theo phương pháp này sẽ khó đảm bảo tính đại biểu. Hơn nữa việc đánh số tất cả các đơn vị sẽ hoàn toàn không thực tế trong trường hợp tổng thể chung có quy mô khá lớn.

---

[1] Simple random sampling

7.4.2. Chọn mẫu phân tổ (chọn mẫu phân tầng)[1]

Trong chọn mẫu phân tổ, trước hết tổng thể gồm N đơn vị sẽ được chia thành K tổ, số đơn vị ở mỗi tổ là $N_1, N_2, ... N_K$ với $\sum_{i=1}^{K} N_i = N$. Số đơn vị mẫu n được phân phối cho các tổ lần lượt là $n_1, n_2, ..., n_K$ với $\sum_{i=1}^{K} n_i = n$.

Trong thực tế số đơn vị mẫu ở từng tổ thường được xác định bằng phương pháp tỷ lệ, gọi n_i là số đơn vị mẫu lấy ra từ tổ i ta có:

$$n_i = n\frac{N_i}{N} \Rightarrow \frac{n_i}{n} = \frac{N_i}{N}$$

hoặc $\frac{n_i}{N_i} = \frac{n}{N}$ (là hằng số).

7.4.2.1. Ước lượng trung bình tổng thể

Gọi μ là trung bình của cả tổng thể, $\bar{x}_i$ là trung bình mẫu của tổ thứ i, s_i^2 là phương sai mẫu hiệu chỉnh của tổ thứ i ta có:

- Ước lượng điểm của μ là:. $\bar{x} = \frac{1}{N}\sum_{i=1}^{K} \bar{x}_i N_i$

- Ước lượng khoảng cho μ với độ tin cậy là 1 - α:

$$\bar{x} - z_{\alpha/2}s_{\bar{x}} < \mu < \bar{x} + z_{\alpha/2}s_{\bar{x}} \qquad (7.5)$$

Với :

$$s_{\bar{x}}^2 = \sum_{i=1}^{k} \frac{w_i^2 s_i^2}{n_i}(1 - f_i)$$

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

$$w_i = \frac{N_i}{N} \; ; \; f_i = \frac{n_i}{N_i}$$

Khi các đơn vị mẫu được phân phối theo tỷ lệ thì : $\frac{n_i}{N_i} = \frac{n}{N} = f_i = f$

---

[1] Stratified random sampling

Khi đó $s_{\bar{x}}^2 = (1-f)\sum_{i=1}^{K}\frac{w_i^2 s_i^2}{n_i}$.

Ví dụ: Một cuộc nghiên cứu về thời gian xem tivi trung bình của học sinh trong tuần được thực hiện ở 1 trường tiểu học. Người ta tiến hành chọn mẫu số học sinh từ các khối lớp 5, khối lớp 4, khối lớp 3. Số học sinh mỗi khối và số mẫu được chọn ra từ mỗi khối (theo phương pháp chọn tỷ lệ) lần lượt là: $N_1 = 250$, $n_1 = 50$, $N_2 = 280$, $n_2 = 56$, $N_3 = 300$, $n_3 = 60$.

Ta có giá trị trung bình và độ lệch chuẩn về số thời gian xem tivi trong tuần lần lượt là:

$\bar{x}_1 = 13,54$ giờ; $s_1 = 3,02$ giờ; $\bar{x}_2 = 15,62$ giờ; $s_2 = 4,25$ giờ; $\bar{x}_3 = 14,25$ giờ $s_3 = 2,98$ giờ; $N = 830$; $n = 166$.

Hãy ước lượng điểm và khoảng cho thời gian xem tivi trung bình của học sinh tiểu học trong tuần với độ tin cậy 95%.

Giải:

+ Ước lượng điểm cho μ là:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{K}\bar{x}_i N_i = \frac{13,54x250 + 15,62x280 + 14,25x300}{250 + 280 + 300} = 14,50 \text{ giờ}$$

+ Ước lượng khoảng cho μ, ta tính:

$$s_{\bar{x}}^2 = (1-f)\sum_{i=1}^{K}\frac{w_i^2 s_i^2}{n_i}$$

$$= (1 - \frac{166}{830})\left[\frac{\left(\frac{250}{830}\right)^2 \times 3,02^2}{50} + \frac{\left(\frac{280}{830}\right)^2 \times 4,25^2}{56} + \frac{\left(\frac{300}{830}\right)^2 \times 2,98^2}{60}\right]$$

$$= 0,058$$

$$\Rightarrow s_{\bar{x}} = 0,24$$

Với độ tin cậy 95% trung bình của tổng thể μ:

$$\bar{x} - z_{\alpha/2}s_{\bar{x}} < \mu < \bar{x} + z_{\alpha/2}s_{\bar{x}}$$

$$14,5 - 1,96 \times 0,24 < \mu < 14,5 + 1,96 \times 0,24$$

$$14,03 < \mu < 14,97$$

Nghĩa là thời gian xem tivi trung bình của học sinh tiểu học trong 1 tuần nằm trong khoảng từ 14,03 giờ đến 14,97 giờ với độ tin cậy 95%.

7.4.2.2. Ước lượng tỷ lệ tổng thể:

Gọi p, p_i, lần lượt là tỷ lệ các đơn vị có tính chất nào đó mà ta quan tâm của tổng thể và của tổ thứ i.

- Ước lượng điểm của p được xác định bởi:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{K} N_i p_i = \sum_{i=1}^{k} w_i p_i$$

- Ước lượng khoảng cho p với độ tin cậy 1 - α:

$$\hat{p} - z_{\alpha/2} s_{\hat{p}} < p < \hat{p} + z_{\alpha/2} s_{\hat{p}} \qquad (7.6)$$

với :
$$s_{\hat{p}}^2 = \sum_{i=1}^{K} \frac{w_i^2 p_i(1-p_i)}{n_i - 1}\left(1 - \frac{n_i}{N_i}\right)$$

khi phân phối mẫu theo tỷ lệ thì : $\frac{n_i}{N_i} = \frac{n}{N}$

nên :
$$s_{\hat{p}}^2 = \left(1 - \frac{n}{N}\right)\sum_{i=1}^{K} \frac{w_i^2 p_i(1-p_i)}{n_i - 1}$$

Ví dụ: Một huyện miền núi gồm 3 xã. Người ta tiến hành điều tra tỷ lệ người mù chữ của huyện (≥ 8 tuổi) bằng phương pháp phân tổ số nhân khẩu. Số mẫu và số người mù chữ trong mẫu của các xã tương ứng là:

$N_1 = 1200$; $n_1 = 120$; $m_1 = 6$

$N_2 = 2400$; $n_2 = 240$; $m_2 = 15$

$N_3 = 1800$; $n_3 = 180$; $m_3 = 13$

(phân bổ theo tỷ lệ).

Hãy ước lượng điểm và khoảng cho tỷ lệ người mù chữ của huyện với độ tin cậy là 95%.

Giải

Ta có $N = N_1 + N_2 + N_3 = 1200 + 2400 + 1800 = 5400$

$n = n_1 + n_2 + n_3 = 120 + 240 + 180 = 540$

$$p_1 = \frac{m_1}{n_1} = \frac{6}{120} = 0.05$$

$$p_2 = \frac{m_2}{n_2} = \frac{15}{240} = 0{,}0625$$

$$p_3 = \frac{m_3}{n_3} = \frac{13}{180} = 0{,}0722$$

* Ước lượng điểm cho tỷ lệ p là:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{3} p_i N_i = \frac{0{,}05.1200 + 0{,}0625.2400 + 0{,}0722.1800}{5400} = 0{,}0629$$

Nghĩa là tỷ lệ người mù chữ của toàn huyện là 6,29%.

* Ước lượng khoảng cho tỷ lệ p là:

$$\hat{p} - z_{\alpha/2}s_{\hat{p}} < p < \hat{p} + z_{\alpha/2}s_{\hat{p}}$$

Trong đó:

$$s_{\hat{p}}^2 = (1 - \frac{n}{N})\sum_{i=1}^{3}\frac{w_i^2 p_i(1-p_i)}{n_i - 1}$$

$$= \left(1 - \frac{540}{5400}\right) \times$$

$$\left[\frac{\left(\frac{1200}{5400}\right)^2 0{,}05(1-0{,}05)}{120-1} + \frac{\left(\frac{2400}{5400}\right)^2 0{,}0625(1-0{,}0625)}{240-1} + \frac{\left(\frac{1800}{5400}\right)^2 0{,}0722(1-0{,}0722)}{180-1}\right]$$

$$= 0{,}00009875$$

$$\Rightarrow s_{\hat{p}} = 0{,}009937$$

$$0{,}0629 - 1{,}96 \times 0{,}009937 < p < 0{,}0629 + 1{,}96 \times 0{,}009937$$

$$0{,}0434 < p < 0{,}0824$$

Nghĩa là, với độ tin cậy 95% tỷ lệ người mù chữ của huyện nằm trong khoảng từ 4,34% đến 8,24%.

Phương pháp chọn mẫu phân tổ có ưu điểm là: thực hiện thuận lợi, phân tích số liệu khá toàn diện và hiệu quả hơn lấy mẫu ngẫu nhiên đơn giản. Tuy nhiên nó cũng có nhược điểm là: để thực hiện phương pháp chọn mẫu này đòi hỏi phải có các nguồn thông tin có sẵn và những kiến thức, kinh nghiệm về tổng thể nghiên cứu để có thể phân tổ tổng thể. Phương pháp chọn mẫu

phân tổ phần nào dựa vào những kinh nghiệm phán đoán chủ quan, nên cần phải tuân theo những nguyên tắc chung khi tiến hành phân tổ như sau:

- Phải đảm bảo tính đồng chất của tổ.

- Số tổ không được chia quá nhỏ và quá nhiều, số đơn vị mẫu của từng tổ phải đủ lớn để đảm bảo độ tin cậy cho ước lượng.

7.5 Chọn mẫu cả khối (mẫu cụm) [1]:

Chọn mẫu cả khối là phương pháp tổ chức chọn mẫu trong đó số đơn vị mẫu được rút ra để điều tra không phải là từng đơn vị lẻ tẻ mà là từng khối đơn vị. Như vậy trước hết tổng thể chung được chia thành các khối, sau đó chọn ngẫu nhiên một số khối để điều tra.

Giả sử tổng thể chia thành M khối. Mẫu gồm m khối được chọn ngẫu nhiên từ M khối và điều tra được thực hiện trên tất cả các đơn vị của m khối được chọn. Gọi:

$n_1, n_2, ..., n_m$ lần lượt là số đơn vị tổng thể của khối thứ 1, 2,.., m.

$\bar{x}_1, \bar{x}_2, ..., \bar{x}_m$ lần lượt là trung bình của khối thứ 1, 2,.., m.

$p_1, p_2, ..., p_m$ lần lượt là tỷ lệ các đơn vị có tính chất nào đó trong khối thứ 1, 2,.., m.

+ Ước lượng điểm cho μ và p lần lượt là:

Trung bình:
$$\bar{x} = \frac{\sum_{i=1}^{m} \bar{x}_i n_i}{\sum_{i=1}^{m} n_i}$$

Tỷ lệ:
$$\hat{p} = \frac{\sum_{i=1}^{m} p_i n_i}{\sum_{i=1}^{m} n_i}$$

+ Ước lượng khoảng cho μ và p với độ tin cậy 1 - α lần lượt là:

$$\bar{x} - z_{\alpha/2} s_{\bar{x}} < \mu < \bar{x} + z_{\alpha/2} s_{\bar{x}} \qquad (7.7)$$

$$\hat{p} - z_{\alpha/2} s_{\hat{p}} < p < \hat{p} + z_{\alpha/2} s_{\hat{p}} \qquad (7.8)$$

---

[1] Cluster sampling

với: $$s_x^2 = \frac{M-m}{Mm\bar{n}^2} \times \frac{\sum_{i=1}^{m} n_i^2(\bar{x}_i - \bar{x})^2}{m-1}$$

$$s_{\hat{p}}^2 = \frac{M-m}{Mm\bar{n}^2} \times \frac{\sum_{i=1}^{m} n_i^2(p_i - \hat{p})^2}{m-1}$$

Trong đó: $\bar{n} = \frac{\sum_{i=1}^{m} n_i}{m}$: Số đơn vị tính trung bình cho một khối.

Ví dụ: Một tỉnh có 6 huyện gồm 519 ấp. Người ta chọn ra 10 ấp một cách ngẫu nhiên. Số liệu về thu nhập của hộ gia đình và tỷ lệ hộ có thu nhập hằng năm dưới 7 triệu đồng được cho trong bảng sau. Hãy ước lượng điểm và khoảng về thu nhập trung bình hàng năm và tỷ lệ hộ có thu nhập hằng năm dưới 7 triệu đồng cho toàn tỉnh với độ tin cậy 95%.

Ấp	Số hộ (n_i)	Thu nhập trung bình (trđ) ($\bar{x}_i$)	Tỷ lệ hộ có thu nhập nhỏ hơn 7 triệu đồng (p_i)
1	150	15,234	0,1224
2	125	14,126	0,0982
3	65	12,150	0,2246
4	120	10,280	0,3546
5	140	13,150	0,1825
6	200	16,110	0,2246
7	110	11,210	0,3410
8	85	8,240	0,1240
9	132	9,150	0,1928
10	140	15,320	0,2234

Ta có: M = 519 ; m = 10.

+ Ước lượng điểm cho μ và p lần lượt là:

$$\bar{x} = \frac{\sum_{i=1}^{m} \bar{x}_i n_i}{\sum_{i=1}^{m} n_i} = \frac{15,234x150 + ... + 15,320x140}{150 + 125 + ... + 140} = \frac{16423,300}{1267} = 12,962$$

$$\hat{p} = \frac{\sum_{i=1}^{m} p_i n_i}{\sum_{i=1}^{m} n_i} = \frac{263,0316}{1267} = 0,2076$$

Nghĩa là thu nhập trung bình hàng năm của một hộ gia đình được ước lượng là 12,962 triệu đồng và tỷ lệ hộ có thu nhập hàng năm thấp hơn 7 triệu đồng là 20,76%.

+ Ước lượng khoảng cho μ và p với độ tin cậy 95% .

• Ước lượng khoảng μ:

Ta có: $$\bar{n} = \frac{\sum_{i=1}^{m} n_i}{m} = \frac{1267}{10} = 126,7$$

$$\frac{\sum_{i=1}^{m} n_i^2 (\bar{x}_i - \bar{x})^2}{m-1} = \frac{150^2(15,234-12,962)^2 + ... + 140^2(15,320-12,962)^2}{10-1}$$

$$= 133.464,7482$$

$$s_{\bar{x}}^2 = \frac{M-m}{Mm\bar{n}^2} \frac{\sum_{i=1}^{m} n_i^2 (\bar{x}_i - \bar{x})^2}{m-1}$$

$$= \frac{(519-10).133464,7482}{519.10.126,7^2} = 0,8153869$$

$s_{\bar{x}} = 0,9029878$ triệu đồng ≈ 0,903 triệu đồng.

Với độ tin cậy 95% → $\alpha = 5\%$, $z_{\alpha/2} = z_{0,025} = 1,96$.

Ta có ước lượng khoảng cho μ với độ tin cậy 95% là:

$$12,962 - 1,96 \times 0,903 < \mu < 12,962 + 1,96 \times 0,903$$
$$11,192 < \mu < 14,732$$

Nghĩa là, với độ tin cậy 95% thu nhập trung bình hàng năm của một hộ nằm trong khoảng từ 11,192 triệu đồng đến 14,732 triệu đồng.

• Ước lượng khoảng cho p:

Ta có:
$$\frac{\sum_{i=1}^{m} n_i^2 (p_i - \hat{p})^2}{m-1} = 106,7959839$$

$$s_{\hat{p}}^2 = \frac{M-m}{Mmn^{-2}} \times \frac{\sum_{i=1}^{m} n_i^2 (p_i - \hat{p})^2}{m-1}$$

$$= \frac{(519-10) \times 106,7959839}{519.10.126,7^2} = 6,52457 \times 10^{-4}$$

$$\Rightarrow s_{\hat{p}} = 0,0255$$

Vậy ước lượng khoảng cho với độ tin cậy 95% là:

$$0,2076 - 1,96 \times 0,0255 < p < 0,2076 + 1,96 \times 0,0255$$
$$0,1576 < p < 0,2576$$
$$15,76\% < p < 25,76\%$$

Nghĩa là với độ tin cậy 95%, tỷ lệ hộ có thu nhập hàng năm dưới 7 triệu đồng trong tổng các hộ ở tỉnh nằm trong khoảng từ 15,76% đến 25,76%.

Phương pháp chọn mẫu theo khối có ưu điểm là không cần thiết phải xây dựng một danh sách tất cả các phần tử trong tổng thể nghiên cứu như phải làm trong lấy mẫu ngẫu nhiên đơn giản hay chọn mẫu phân tổ. Các phần tử được chọn điều tra nằm tập trung theo từng khu vực nên hạn chế được chi phí và thời gian đi lại. Tuy nhiên phương pháp chọn mẫu theo khối cũng có nhược điểm là: do các đơn vị mẫu tập trung, không phân bố đều trong tổng thể, tính đại diện của mẫu có thể không cao nên sai số chọn mẫu có khả năng phát sinh lớn hơn.

Bên cạnh các phương pháp chọn mẫu ngẫu nhiên trên đây, trong thực tế người ta còn sử dụng các phương pháp chọn mẫu phi ngẫu nhiên. Điều tra chọn mẫu phi ngẫu nhiên là phương pháp lựa chọn các đơn vị của tổng thể vào mẫu điều tra trên cơ sở xem xét chủ quan của nhà thống kê. Chọn mẫu phi ngẫu nhiên không hoàn toàn dựa trên cơ sở toán học như chọn mẫu ngẫu

phi ngẫu nhiên không hoàn toàn dựa trên cơ sở toán học như chọn mẫu ngẫu nhiên mà đòi hỏi phải kết hợp chặt chẽ giữa phân tích lý luận với thực tiễn xã hội. Muốn cho chất lượng tài liệu điều tra tốt phải giải quyết các vấn đề sau:

- Phải đảm bảo phân tổ chính xác đối tượng điều tra. Phân tổ tổng thể giúp cho việc lựa chọn các đơn vị mẫu có khả năng đại diện cao cho cả tổng thể.

- Chọn đơn vị điều tra: Vì là chọn mẫu phi ngẫu nhiên nên các đơn vị mẫu được lựa chọn dựa vào kinh nghiệm của các chuyên gia hoặc qua bàn bạc phân tích tập thể. Thường người ta chọn những đơn vị có mức độ tiêu thức gần với số trung bình của từng bộ phận nhất, đồng thời cũng là mức độ phổ biến nhất trong bộ phận đó, hoặc những đơn vị có kinh nghiệm về một mặt nào đó (điều tra ý kiến chuyên gia). Ví dụ điều tra ý kiến chuyên gia về một số vấn đề cần giải quyết như: vấn đề tiền lương, vấn đề thương binh xã hội, vấn đề bảo hiểm...

- Sai số chọn mẫu: Sai số chọn mẫu trong chọn phi ngẫu nhiên không thể tính được bằng công thức toán học mà phải thông qua nhận xét, so sánh để ước lượng ra. Khi suy rộng kết quả điều tra chọn mẫu phi ngẫu nhiên người ta suy rộng trực tiếp, không suy rộng có phạm vi như chọn ngẫu nhiên.

- Huấn luyện nhân viên tham gia điều tra: Trong chọn mẫu phi ngẫu nhiên ý kiến chủ quan của con người đóng vai trò quan trọng. Chính vì vậy, muốn làm tốt công tác điều tra, người cán bộ không những phải thành thạo về nghiệp vụ, am hiểu về hiện tượng nghiên cứu mà còn cần phải trung thực và làm tốt công tác tổ chức vận động quần chúng. Cán bộ điều tra cần giải thích cho mọi người hiểu rõ mục đích nghiên cứu để họ tích cực và tự giác tham gia, tự giác khai báo.

Điều tra chọn mẫu ngẫu nhiên và phi ngẫu nhiên đều là các phương pháp chọn mẫu có hiệu quả. Mỗi phương pháp có những mặt ưu nhược điểm riêng của nó, thích hợp với từng hiện tượng nghiên cứu. Trong thực tế nếu biết khéo léo kết hợp cả hai phương pháp chọn này, kết quả điều tra sẽ có chất lượng cao và thủ tục làm cũng đơn giản hơn.

CHƯƠNG 8

KIỂM ĐỊNH GIẢ THUYẾT[1]

8.1. KHÁI NIỆM:

Các đặc trưng của mẫu ngoài việc sử dụng để ước lượng các đặc trưng của tổng thể còn được dùng để đánh giá xem một giả thuyết nào đó của tổng thể là đúng hay sai. Việc tìm ra kết luận để bác bỏ hay chấp nhận một giả thuyết gọi là kiểm định giả thuyết.

Ví dụ:

1. Một nhà sản xuất cho rằng trọng lượng trung bình của 1 gói mì là 75g. Để kiểm tra điều này là đúng hay sai, chọn ngẫu nhiên một số gói mì ra để kiểm tra, tính toán.

2. Một nhà sản xuất cho rằng tỷ lệ phế phẩm là 5%. Để kiểm tra điều này là đúng hay sai, chọn ngẫu nhiên một số sản phẩm ra để kiểm tra, tính toán.

3. Một nhà quản trị Marketing muốn kiểm tra giả thuyết: "doanh thu của công ty tăng trung bình ít nhất là 7% sau đợt quảng cáo". Để kiểm tra điều này là đúng hay sai bằng cách liệt kê doanh thu trước và sau chiến dịch quảng cáo để tính toán.

8.2. GIẢ THUYẾT H_0 VÀ GIẢ THUYẾT H_1

8.2.1. Giả thuyết H_0[2]:

Giả sử tổng thể chung có đặc trưng θ chưa biết (như trung bình, tỷ lệ, phương sai). Với giá trị θ_0 cụ thể cho trước nào đó, ta cần kiểm định giả thuyết $H_0 : \theta = \theta_0$ (kiểm định hai bên) hoặc giả thuyết là một dãy giá trị, lúc đó : $H_0 : \theta \geq \theta_0$ hoặc: $H_0 : \theta \leq \theta_0$ (kiểm định một bên).

8.2.2. Giả thuyết H_1[3]:

Giả thiết H_1 là kết quả ngược lại của giả thuyết H_0, nếu giả thuyết H_0 đúng thì giả thuyết H_1 sai và ngược lại. H_1 còn được gọi là giả thuyết đối.

---

[1] Hypothesis testing

[2] Null hypothesis

[3] Alternative hypothesis

Vậy cặp giả thuyết H_0 và H_1 được thể hiện trong các trường hợp kiểm định như sau:

- Trong trường hợp kiểm định hai bên [1]

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

- Trong trường hợp kiểm định một bên[2]

$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

hoặc: $$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

Ví dụ: Một nhà sản xuất cho rằng trọng lượng trung bình của một gói mì là 75gam, để kiểm tra lời tuyên bố này là đúng hay sai ta có thể đặt giả thuyết :

$$\begin{cases} H_0 : \theta = 75 \\ H_1 : \theta \neq 75 \end{cases}$$

8.2.3. Sai lầm loại 1 [3] và sai lầm loại 2 [4]:

Vì chỉ dựa trên một mẫu để kết luận đến các giá trị của tổng thể, nên ta có thể phạm sai lầm khi đưa ra kết luận về giả thuyết H_0. Các sai lầm đó là:

1. Giả thuyết H_0 đúng (tức thực tế $\theta = \theta_0$) nhưng qua kiểm định ta kết luận giả thuyết sai tức là $\theta \neq \theta_0$, và do vậy ta bác bỏ H_0.

2. Giả thuyết H_0 sai nhưng qua kiểm định ta kết luận giả thuyết đúng và do vậy ta chấp nhận giả thuyết H_0.

Người ta quy ước gọi sai lầm ở trường hợp 1 là sai lầm loại I tức là bác bỏ giả thuyết H_0 khi giả thuyết này đúng, còn sai lầm ở trường hợp 2 là sai lầm loại 2 tức là chấp nhận giả thuyết H_0 khi giả thuyết này sai. Như vậy khi ta bác bỏ một giả thuyết là ta có thể mắc phải sai lầm loại I, còn khi ta chấp nhận một giả thuyết là ta có thể phạm phải sai lầm loại II . Thực ra

---

[1] Two- tail test

[2] One- tail test

[3] Type I error

[4] Type II error

sai lầm loại 1 và loại 2 rất tương đối, nó chỉ được xác định khi ta đã đặt giả thuyết H_0 và thông thường sai lầm nào gây ra tổn thất lớn hơn người ta sẽ đặt giả thuyết H_0 sao cho sai lầm đó là loại 1 và định trước khả năng mắc phải sai lầm loại 1 không vượt quá một số α nhỏ nào đó, tức là thực hiện kiểm định giả thuyết H_0 ở mức ý nghĩa α cho trước. Nếu α càng bé thì khả năng phạm phải sai lầm loại I càng ít, tuy nhiên trong trường hợp này xác suất sai lầm loại II sẽ tăng lên. Chẳng hạn nếu lấy $\alpha = 0$ thì sẽ không bác bỏ bất kỳ giả thuyết nào, có nghĩa là không mắc phải sai lầm loại 1 và như vậy xác suất sai lầm loại II sẽ đạt cực đại.

Nếu quyết định xác suất bác bỏ giả thuyết H_0 khi giả thuyết này đúng là α thì xác suất để chấp nhận nó là $(1 - \alpha)$. Người ta gọi α là mức ý nghĩa của kiểm định[1].

Ngược lại với sai lầm loại 1, sai lầm loại II là loại sai lầm của việc chấp nhận giả thuyết H_0 khi giả thuyết này sai. Nếu xác suất của việc quyết định chấp nhận một giả thuyết H_0 sai được ký hiệu là β thì xác suất để bác bỏ giả thuyết này là $(1 - \beta)$.

Những quyết định dựa trên giả thuyết H_0 được tóm tắt như sau:

	Giả thuyết H_0 đúng	**Giả thuyết H_0 sai**
1. Chấp nhận giả thuyết H_0	Xác suất quyết định đúng là $(1 - \alpha)$	Xác suất sai lầm loại 2 là β
2. Bác bỏ giả thuyết H_0	Xác suất sai lầm loại 1 là α	Xác suất quyết định đúng là $(1 - \beta)$

Ví dụ: Một nhà kinh doanh sau khi áp dụng các biện pháp khuyến mãi muốn tìm hiểu xem lợi nhuận có tăng lên hay không. Để thực hiện việc kiểm định giả thuyết ta xét các trường hợp cho trong bảng sau:

[1] Significance level

Giả thuyết H_o	Thực tế	Bác bỏ giả thuyết H_o	Chấp nhận giả thuyết H_o
"Lợi nhuận có tăng"	Lợi nhuận có tăng	Mắc sai lầm loại I Xác suất = α	Kết luận đúng Xác suất = $1-\beta$
	Lợi nhuận không tăng	Kết luận đúng Xác suất = $1-\alpha$	Mắc sai lầm loại II Xác suất = β
"Lợi nhuận không tăng"	Lợi nhuận có tăng	Kết luận đúng Xác suất = $1-\alpha$	Mắc sai lầm loại II Xác suất = β
	Lợi nhuận không tăng	Mắc sai lầm loại I Xác suất = α	Kết luận đúng Xác suất = $1-\beta$

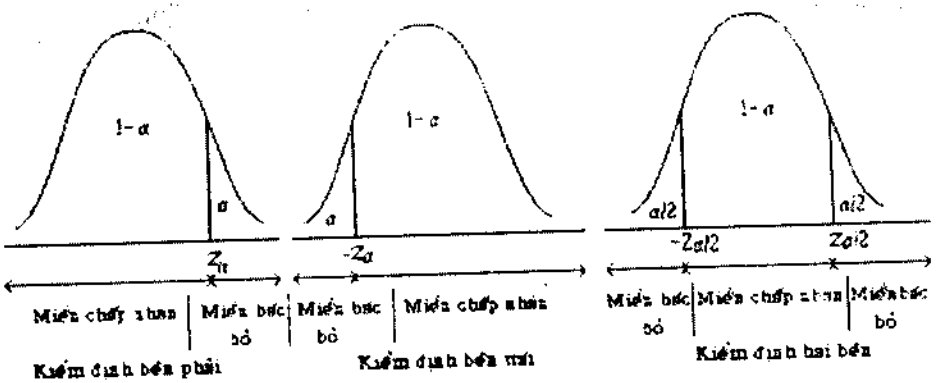
Ở đây ta nên đặt giả thuyết như thế nào? Muốn vậy người ta phải xem xét sai lầm nào quan trọng hơn, tức là khi mắc phải sẽ chịu tổn thất lớn hơn và sẽ đặt bài toán để sai lầm đó là sai lầm loại I. Rõ ràng lợi nhuận không tăng mà bảo có là sai lầm "nghiêm trọng" hơn sai lầm: lợi nhuận có tăng mà bảo không. Do vậy nên đặt giả thuyết H_o: "Lợi nhuận không tăng".

Trong thực tế đôi khi người ta kiểm định giả thuyết với giả thuyết đối dạng $\theta > \theta_o$ hoặc $\theta < \theta_o$. Nếu bằng kinh nghiệm hoặc qua phân tích ta biết được chiều hướng là $\theta > \theta_o$ thì ta có thể đặt giả thuyết đối dạng $\theta > \theta_o$, hoặc ta biết được chiều hướng $\theta < \theta_o$ thì ta có thể đặt giả thuyết đối dạng $\theta < \theta_o$.

Nếu giả thuyết đối có dạng $H_1 : \theta > \theta_o$ thì được gọi là kiểm định bên phải (Vì miền bác bỏ nằm về phía bên phải của miền chấp nhận).

Nếu giả thuyết đối có dạng $H_1 : \theta < \theta_o$ thì được gọi là kiểm định bên trái (Vì miền bác bỏ nằm về phía bên trái của miền chấp nhận).

Nếu giả thuyết đối có dạng $H_1 : \theta \neq \theta_o$ thì được gọi là kiểm định hai bên (Vì miền bác bỏ nằm về hai phía của miền chấp nhận).



Hình 8.1 Miền bác bỏ, miền chấp nhận trong kiểm định

Ví dụ: Một công ty sau khi áp dụng các biện pháp khuyến mãi muốn nghiên cứu xem mức lợi nhuận có tăng lên hay không. Trước khi áp dụng các biện pháp khuyến mãi lợi nhuận trung bình của công ty trong ngày là 10 triệu đồng.

Ở đây ta nên đặt giả thuyết H_0 như thế nào? Rõ ràng bằng kinh nghiệm ta thấy việc áp dụng các biện pháp khuyến mãi sẽ làm tăng lợi nhuận. Do vậy ta đặt giả thuyết đối H_1 : "Lợi nhuận trung bình > 10 triệu đồng" và ta có giả thuyết.

H_0 : " Lợi nhuận trung bình ≤ 10 triệu đồng"

H_1 : " Lợi nhuận trung bình > 10 triệu đồng"

Dĩ nhiên để kiểm định giả thuyết trên ta phải thu thập số liệu trên mẫu để tính toán.

8.3. KIỂM ĐỊNH GIẢ THUYẾT VỀ TỶ LỆ TỔNG THỂ:

Giả sử tổng thể chia hai loại phần tử. Tỷ lệ phần tử có tính chất A là p chưa biết. Gọi $\hat{p}$ là tỷ lệ các đơn vị có tính chất A trong mẫu. Với mẫu lớn, $n \geq 40$, tỷ lệ mẫu $\hat{p}$ có phân phối chuẩn. Kiểm định giả thuyết về p được thực hiện như sau:

- Đặt giả thuyết: Có thể thực hiện theo 1 trong 3 trường hợp sau:

Trường hợp 1: $H_0 : p = p_0$ (p_0 là giá trị cho trước)

$H_1 : p \neq p_0$

Trường hợp 2: $H_0 : p = p_0$ hoặc Ho: $p \geq p_0$

$H_1 : p < p_0$

Trường hợp 3: $H_0 : p = p_0$ hoặc Ho: $p \leq p_0$

$H_1 : p > p_0$

- Tính giá trị của tiêu chuẩn kiểm định (gọi tắt là giá trị kiểm định):

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_o(1-p_o)}{n}}} \qquad (8.1)$$

- Quy tắc kiểm định: Được tóm tắt trong bảng sau:

Giả thuyết	Bác bỏ Ho khi:
$H_o : p = p_o$ $H_1 : p \neq p_o$	$z > z_{\alpha/2}$ hoặc $z < -z_{\alpha/2}$ Hay $\lvert z \rvert > z_{\alpha/2}$
$H_o : p = p_o$ hoặc Ho: $p \geq p_0$ $H_1 : p < p_o$	$z < -z_\alpha$
$H_o : p = p_o$ hoặc Ho: $p \leq p_0$ $H_1 : p > p_o$	$z > z_\alpha$

Giả sử ta có kiểm định hai bên: $H_o : p = p_o$ (p_o là giá trị cho trước)

$H_1 : p \neq p_o$

Ta thực hiện như sau:

Từ α đã biết ta tìm $z_{\alpha/2}$ bằng cách tra bảng hoặc dùng hàm **NORMSINV** trong EXCEL

+ Nếu $\lvert z \rvert \leq z_{\alpha/2}$ ta chấp nhận giả thuyết, coi $p = p_o$.

+ Nếu $\lvert z \rvert > z_{\alpha/2}$ ta bác bỏ giả thuyết, coi $p \neq p_o$.

Và khi đó: - nếu $\hat{p} > p_o$ ta xem $p > p_o$.

- nếu $\hat{p} < p_o$ ta xem $P < P_o$.

Ví dụ: Một nhà máy sản xuất sản phẩm với tỷ lệ sản phẩm loại 1 lúc đầu là 0,20. Sau khi áp dụng phương pháp sản xuất mới, kiểm tra 500 sản phẩm thấy số sản phẩm loại 1 là 150 sản phẩm. Cho kết luận về phương pháp sản xuất mới này với mức ý nghĩa $\alpha = 1\%$.

Giải

Tỷ lệ sản phẩm loại 1 lúc đầu là $p_o = 0,2$.

Tỷ lệ sản phẩm loại 1 khi áp dụng phương pháp mới là p chưa biết:

- Ta đặt giả thuyết: $H_o : p = p_0 = 0,2$

$H_1 : p \neq p_0 = 0,2$

- Kiểm tra giả thuyết:

Giá trị kiểm định : $$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

Ta có : $$\hat{p} = \frac{150}{500} = 0,3; \quad n = 500; \quad z_{0,005} = 2,58$$

$$z = \frac{0.3 - 0.2}{\sqrt{\frac{0,2(1-0,2)}{500}}} \approx 5,59$$

vì $|z| = 5,59 > z_{\alpha/2} = 2,58$ nên ta bác bỏ giả thuyết tức là $p \neq p_0 = 0,2$ nghĩa là phương pháp sản xuất mới đã làm thay đổi tỷ lệ sản phẩm loại 1. Do $\hat{p} = 0,3 > p_o = 0,2$ nên $p > p_o = 0,2$. Vậy phương pháp sản xuất mới có hiệu quả tốt.

Thực ra với giá trị kiểm định tính được khá lớn, giả thuyết Ho bị bác bỏ ở hầu hết các mức ý nghĩa α cho trước.

8.4. KIỂM ĐỊNH GIẢ THUYẾT VỀ TRUNG BÌNH TỔNG THỂ CHUNG:

Giả sử tổng thể có trung bình là μ chưa biết. Ta cần kiểm tra giả thuyết :

H_o: $\mu = \mu_o$ (μ_o cho trước).

H_1: $\mu \neq \mu_o$

Căn cứ vào mẫu gồm n quan sát độc lập ta đưa ra quy tắc chấp nhận hay bác bỏ giả thuyết trên với mức ý nghĩa α.

Ta chia thành hai trường hợp:

a) $n \geq 30$:

a1) σ^2 đã biết; Ta tính giá trị kiểm định:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad (8.2)$$

Dựa vào α ta tìm $z_{\alpha/2}$.

Nếu $|z| > z_{\alpha/2}$ ta bác bỏ giả thuyết.

Nếu $|z| \le z_{\alpha/2}$ ta chấp nhận giả thuyết.

a2) σ^2 chưa biết: Ta thay $\sigma^2 = s^2$ (phương sai hiệu chỉnh mẫu).

b) n < 30:

b1) X có phân phối chuẩn; σ^2 đã biết; ta làm giống như trường hợp a1

b2) X có phân phối chuẩn, σ^2 chưa biết:

Ta tính giá trị kiểm định:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \qquad (8.3)$$

và tìm $t_{n-1,\alpha/2}$ trong bảng phân phối t student hoặc dùng hàm **TINV** trong EXCEL

+ Nếu $|t| > t_{n-1,\alpha/2}$ ta bác bỏ giả thuyết.

+ Nếu $|t| \le t_{n-1,\alpha/2}$ ta chấp nhận giả thuyết.

Chú ý: trong tất cả các trường hợp trên nếu giả thuyết đã bị bác bỏ, tức là $\mu \ne \mu_0$, khi đó:

- Nếu $\bar{x} > \mu_0$ ta kết luận $\mu > \mu_0$.
- Nếu $\bar{x} < \mu_0$ ta kết luận $\mu < \mu_0$.

Trên đây là trường hợp kiểm định hai bên, trong trường hợp kiểm định một bên với $n \ge 30$, ta có bảng tóm tắt sau:

Giả thuyết	Bác bỏ Ho khi:
$H_0 : \mu = \mu_0$ $H_1 : \mu \ne \mu_0$	$z > z_{\alpha/2}$ hoặc $z < -z_{\alpha/2}$ Hay $\lvert z\rvert > z_{\alpha/2}$
$H_0 : \mu = \mu_0$ hoặc Ho: $\mu \ge \mu_0$ $H_1 : \mu < \mu_0$	$z < -z_\alpha$
$H_0 : \mu = \mu_0$ hoặc Ho: $\mu \le \mu_0$ $H_1 : \mu > \mu_0$	$z > z_\alpha$

Trường hợp kiểm định một bên với $n < 30$, σ^2 chưa biết, ta có bảng tóm tắt sau:

Giả thuyết	Bác bỏ Ho khi:
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$t > t_{n-1,\alpha/2}$ hoặc $t < -t_{n-1,\alpha/2}$ Hay $\lvert t \rvert > t_{n-1,\alpha/2}$
$H_0 : \mu = \mu_0$ hoặc Ho: $\mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$t < -t_{n-1,\alpha}$
$H_0 : \mu = \mu_0$ hoặc Ho: $\mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$t > t_{n-1,\alpha}$

Ví dụ 1: Một máy đóng mì gói tự động quy định trọng lượng trung bình là $\mu_0 = 75g$, độ lệch chuẩn là $\sigma = 15g$. Sau một thời gian sản xuất kiểm tra 80 gói ta có trọng lượng trung bình mỗi gói là 72g. Cho kết luận về tình hình sản xuất với mức ý nghĩa $\alpha = 5\%$.

Giải

Trọng lượng trung bình quy định cho mỗi gói mì là $\mu_0 = 75g$.

Trọng lượng trung bình thực tế sản xuất là μ chưa biết.

- Ta đặt giả thuyết: $H_0 : \mu = \mu_0$

  $H_1 : \mu \neq \mu_0$

- Kiểm tra giả thuyết:

$$n = 80 > 30;\ \sigma = 15;\ \alpha = 5\% \Rightarrow z_{\alpha/2} = z_{2,5\%} = 1,96$$

Giá trị kiểm định:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{72 - 75}{\frac{15}{\sqrt{80}}} \approx -1,79$$

Vì $\lvert z \rvert = 1,79 < 1,96$ nên ta chấp nhận giả thuyết H_0, tức là sản xuất diễn ra bình thường.

Giá trị p[1] :

Giả sử trong ví dụ trên ta kiểm định giả thuyết H_0 với mức ý nghĩa $\alpha = 10\%$, ta có cùng kết luận như trên không?

Nếu $\alpha = 10\% \Rightarrow z_{\alpha/2} = z_{5\%} = 1,645 < z = 1,79$ ta bác bỏ giả thuyết H_0. Như vậy có một vấn đề đặt ra ở đây là xác định mức ý nghĩa nhỏ nhất mà ở đó giả thuyết H_0 bị bác bỏ. Mức ý nghĩa nhỏ nhất đó gọi là giá trị p.

Trở lại ví dụ trên với giá trị kiểm định z = 1,79 như vậy giả thuyết H_0 bị bác bỏ ở bất cứ giá trị nào của α mà ở đó $z_\alpha < 1,79$.

Cụ thể ta tìm giá trị p bằng cách tra bảng z, ta có kết quả như sau:

$$\varphi(1,79) = 0,4633 \Rightarrow \alpha/2 = 0,5 - 0,4633 = 0,0367$$

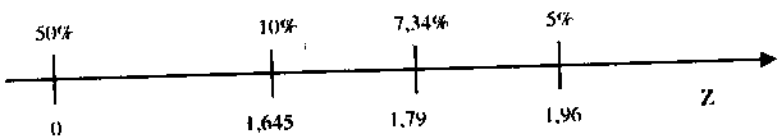
$$\alpha = 2 \times 0,0367 = 7,34\%$$

Do vậy, giả thuyết H_0 sẽ bị bác bỏ ở bất kỳ mức ý nghĩa α nào lớn hơn 7,34%.

Ta cũng có thể tính p-value bằng cách dùng hàm **NORMSDIST** trong EXCEL

Ta có : p-value = P(Z > 1,79) = P(Z < -1,79) = **1- NORMSDIST(1.79)** = 0,0367269

Ta có sơ đồ sau:



Hình 8.1 Miền chấp nhận, miền bác bỏ theo p-value

Trong thực tế việc tính toán được thực hiện bằng các phần mềm thống kê, các kết quả xử lý số liệu bằng máy tính thường luôn thể hiện giá trị p.

Nếu qui định trước mức ý nghĩa α thì có thể dùng p-value để kết luận theo α. Khi đó nguyên tắc kiểm định như sau:

- Nếu p-value < α thì bác bỏ Ho, thừa nhận H_1.
- Nếu p-value ≥ α thì chưa có cơ sở để bác bỏ Ho.

---

[1] Probability value (p-value)

Ngoài ra người ta cũng có thể kiểm định giả thuyết theo p-value và được tiến hành theo nguyên tắc sau:

- Nếu p-value > 0,1 thì thường người ta chấp nhận Ho.
- Nếu 0,05 < p-value ≤ 0,1 thì cần cân nhắc cẩn thận trước khi bác bỏ Ho.
- Nếu 0,01 < p-value ≤ 0,05 thì nghiêng về hướng bác bỏ Ho nhiều hơn.
- Nếu 0,001 < p-value ≤ 0,01 thì ít băn khoăn khi bác bỏ Ho nhiều hơn.
- Nếu p-value ≤ 0,001 thì có thể yên tâm khi bác bỏ Ho.

Ví dụ 2: Một nhà máy sản xuất đèn chụp hình cho biết tuổi thọ trung bình của sản phẩm là 100 giờ. Người ta chọn ngẫu nhiên ra 15 bóng thử nghiệm thấy tuổi thọ trung bình là 99,7 giờ, $s^2 = 0,15$.

Giả thuyết tuổi thọ của đèn có phân phối chuẩn. Cho kết luận về tình hình sản xuất của nhà máy với mức ý nghĩa $\alpha = 5\%$.

Giải:

Trung bình quy định là $\mu_0 = 100$ giờ.

Trung bình thực tế sản xuất là μ chưa biết.

- Đặt giả thuyết: $H_0 : \mu = \mu_0$.

$H_1 : \mu \neq \mu_0$.

- Kiểm tra giả thuyết:

$$n = 15 < 30;\ s^2 = 0,15;\ \overline{x} = 99,7,\ \mu_0 = 100,\ t_{14,0,025} = 2,145$$

Giá trị kiểm định:

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{99,7 - 100}{\frac{\sqrt{0,15}}{\sqrt{15}}} = 3$$

vì $|t| = 3 > 2,145$ nên ta bác bỏ giả thuyết.

Do $\overline{x} = 99,7 < \mu_0 = 100$ nên $\mu < \mu_0 = 100$

Nghĩa là thực tế sản phẩm có tuổi thọ thấp hơn quy định là 100 giờ. Kết luận này đưa ra ở mức ý nghĩa 5%, nghĩa là khả năng ta có thể phạm sai lầm loại I trong kết luận của mình là 5%.

Ta cũng có thể tìm giá trị p bằng cách dùng hàm **TDIST(t,k,1)**

p-value = P(T<-3) = P(T>3) = **TDIST(3,14,1)** = 0,004776

$\alpha/2 = 0{,}004776 \Rightarrow \alpha = 2 \times 0{,}004776 = 0{,}009552 = 0{,}95\%$

Ta cũng có cùng kết luận như trên, nghĩa là giả thuyết Ho sẽ bị bác bỏ ở bất kỳ giá trị nào của $\alpha > 0{,}95\%$.

8.5. KIỂM ĐỊNH GIẢ THUYẾT VỀ PHƯƠNG SAI TỔNG THỂ:

Giả sử ta có mẫu gồm n quan sát được chọn ngẫu nhiên từ tổng thể có phân phối chuẩn với phương sai σ^2 chưa biết.

- Ta cần kiểm định giả thuyết:

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$

Dựa vào mẫu ta đưa ra quy tắc chấp nhận hay bác bỏ giả thuyết trên với mức ý nghĩa α.

- Quy tắc thực hành:

Ta tính giá trị kiểm định: $\chi^2 = \dfrac{(n-1)s^2}{\sigma_o^2}$ (8.4)

Biết α, từ bảng χ^2_{n-1} ta tra được $\chi^2_{n-1,\alpha/2}$ và $\chi^2_{n-1,1-\alpha/2}$

+ Nếu $\chi^2 < \chi^2_{n-1,1-\alpha/2}$ hay $\chi^2 > \chi^2_{n-1,\alpha/2}$ ta bác bỏ giả thuyết.

+ Nếu $\chi^2_{n-1,1-\alpha/2} < \chi^2 < \chi^2_{n-1,\alpha/2}$ ta chấp nhận giả thuyết.

Trong trường hợp giả thuyết bị bác bỏ:

+ Nếu $s^2 > \sigma_o^2$ ta kết luận $\sigma^2 > \sigma_0^2$

+ Nếu $s^2 < \sigma_o^2$ ta kết luận $\sigma^2 < \sigma_0^2$

Trong nhiều trường hợp ta thường quan tâm đến việc kiểm tra xem phương sai của tổng thể σ^2 có vượt quá một giá trị nào đó hay không. Khi đó ta thực hiện kiểm định một bên và là bên phải với giả thuyết.

$$H_0 : \sigma^2 = \sigma_0^2 \text{ hoặc } H_0 : \sigma^2 \leq \sigma_0^2$$

$$H_1 : \sigma^2 > \sigma_0^2$$

Và H_0 sẽ bị bác bỏ ở mức ý nghĩa α nếu :

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{n-1,\alpha}$$

Trường hợp kiểm định bên trái, với giả thuyết:

$$H_0 : \sigma^2 = \sigma_0^2 \text{ hoặc } H_0 : \sigma^2 \geq \sigma_0^2$$

$$H_1 : \sigma^2 < \sigma_0^2$$

H_0 sẽ bị bác bỏ ở mức ý nghĩa α nếu :

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1,1-\alpha}$$

Ví dụ: Một máy tiện tự động quy định phương sai của đường kính trục máy bằng σ_0^2 =36. Người ta tiến hành 25 quan sát về đường kính của trục máy và tính được s^2 = 35,266. Với mức ý nghĩa α = 5% ta có thể kết luận như thế nào về quá trình sản xuất.

Giải

Ta đặt giả thuyết: $H_0 : \sigma^2 = \sigma_0^2$

$H_1 : \sigma^2 \neq \sigma_0^2$

Giá trị kiểm định: $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(25-1)35,266}{36} = 23,5106$

Tra bảng phân phối χ^2 ta có $\chi^2_{24,0,025} = 39,3641$

$$\chi^2 = 23,0156 < 39,3641$$

Nên ta chấp nhận giả thuyết H_0, nghĩa là tình hình sản xuất bình thường.

8.6. KIỂM ĐỊNH GIẢ THUYẾT VỀ SỰ KHÁC NHAU GIỮA 2 SỐ TRUNG BÌNH CỦA HAI TỔNG THỂ:

8.6.1. Kiểm định trong trường hợp mẫu phối hợp từng cặp:

Giả sử ta có mẫu gồm n cặp quan sát lấy ngẫu nhiên từ hai tổng thể X và Y: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Gọi μ_x, μ_y là trung bình của hai tổng thể.

$\bar{d}$: là trung bình của n khác biệt $(x_i - y_i)$.

s_d: là độ lệch tiêu chuẩn của n khác biệt $((x_i - y_i)$.

Giả sử rằng các khác biệt giữa x và y trong tổng thể có phân phối chuẩn.

- Ta cần kiểm định giả thuyết:

$$H_o : \mu_x - \mu_y = D_o \text{ (D_o là giá trị cho trước)}$$

$$H_1 : \mu_x - \mu_y \neq D_o$$

(khi muốn kiểm định giả thuyết $\mu_x = \mu_y$ ta đặt $D_o = o$)

Dựa vào mẫu ta đưa ra quy tắc chấp nhận hay bác bỏ giả thuyết trên với mức ý nghĩa α.

- Quy tắc thực hành:

Tính giá trị kiểm định : $$t = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}} \qquad (8.5)$$

Biết α từ bảng phân phối student ta tìm $t_{n-1,\alpha/2}$

+ $|t| > t_{n-1,\alpha/2}$ ta bác bỏ giả thuyết.

+ $|t| \leq t_{n-1,\alpha/2}$ ta chấp nhận giả thuyết.

Trường hợp kiểm định một bên, ta có bảng tóm tắt sau:

Giả thuyết	Bác bỏ Ho khi:
$H_o : \mu_x - \mu_y = D_0$ $H_1 : \mu_x - \mu_y \neq D_0$	$t > t_{n-1,\alpha/2}$ hoặc $t < -t_{n-1,\alpha/2}$ Hay $\lvert t \rvert > t_{n-1,\alpha/2}$
$H_o : \mu_x - \mu_y = D_0$ hoặc Ho: $\mu_x - \mu_y \geq D_0$ $H_1 : \mu_x - \mu_y < D_0$	$t < -t_{n-1,\alpha}$
$H_o : \mu_x - \mu_y = D_0$ hoặc Ho: $\mu_x - \mu_y \leq D_0$ $H_1 : \mu_x - \mu_y > D_0$	$t > t_{n-1,\alpha}$

Ví dụ: Một công ty thực hiện các biện pháp tăng NSLĐ. Số liệu về NSLĐ của 10 công nhân được thu thập trước và sau khi thực hiện các biện pháp tăng NSLĐ.

Công nhân	NSLĐ trước và sau khi thực hiện các biện pháp tăng NSLĐ (kg/ngày)	
	Trước khi	Sau khi
A	50	52
B	48	46
C	45	50
D	60	65
E	70	78
F	62	61
G	55	58
H	62	70
I	58	67
K	53	65

Quản đốc phân xưởng cho rằng không có sự khác nhau về NSLĐ trung bình trước và sau khi áp dụng các biện pháp tăng NSLĐ. Với mức ý nghĩa 5% có thể kết luận gì về lời tuyên bố của quản đốc?

Giải

Gọi μ_x , μ_y là NSLĐ trung bình sau khi và trước khi thực hiện các biện pháp tăng NSLĐ.

Ta đặt giả thuyết : $H_0 : \mu_x - \mu_y = D_0$

$H_1 : \mu_x - \mu_y \neq D_0$

Từ số liệu trên ta tính được $\bar{d} = 4,9$; $s_d = 4,4833$; $D_0 = 0$.

Giá trị kiểm định : $t = \dfrac{4,9 - 0}{\dfrac{4,4833}{\sqrt{10}}} = \dfrac{4,9}{1,4177} = 3,456$

$t_{n-1,\alpha/2} = t_{9,0,025} = 2,262$ Vì $|t| = 3,456 > 2,262$ nên ta bác bỏ giả thuyết H_0.

Như vậy ta có thể kết luận lời tuyên bố của quản đốc phân xưởng là sai.

Vì $\bar{d} = 4,9 > D_0 = 0$ nên $\mu_x - \mu_y > 0$. Nghĩa là ở mức ý nghĩa 5% các biện pháp tăng NSLĐ đã làm tăng năng suất trung bình.

8.6.2. Kiểm định trong trường hợp mẫu độc lập:

Gọi n_x . n_y là các mẫu được chọn ngẫu nhiên độc lập từ hai tổng thể có phân phối chuẩn X và Y. Có trung bình là μ_x , μ_y ; phương sai là σ_x^2, σ_y^2 ; và trung bình mẫu là $\bar{x}, \bar{y}$.

- Ta cần kiểm định giả thuyết: $H_o : \mu_x - \mu_y = D_o$ (D_o cho trước).

$H_1 : \mu_x - \mu_y \neq D_o$

Dựa vào mẫu ta đưa ra quy tắc chấp nhận hay bác bỏ giả thuyết trên với mức ý nghĩa α.

- Quy tắc thực hành:

Tính giá trị kiểm định :

$$z = \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{ny}}} \qquad (8.6)$$

Biết α ta tìm được $z_{\alpha/2}$.

+ Nếu $|z| > z_{\alpha/2}$ ta bác bỏ giả thuyết.

+ Nếu $|z| \leq z_{\alpha/2}$ ta chấp nhận giả thuyết.

Trường hợp nếu chưa biết phương sai tổng thể, kích thước mẫu lớn (n_x , . $n_y \geq 30$) ta vẫn có thể dùng công thức như trên và thay $\sigma_x^2 = s_x^2, \sigma_y^2 = s_y^2$.

Trường hợp kiểm định một bên, ta có bảng tóm tắt sau:

Giả thuyết	Bác bỏ Ho khi:
$H_o : \mu_x - \mu_y = D_0$ $H_1 : \mu_x - \mu_y \neq D_0$	$z > z_{\alpha/2}$ hoặc $z < -z_{\alpha/2}$ Hay $\|z\| > z_{\alpha/2}$
$H_o : \mu_x - \mu_y = D_0$ hoặc Ho: $\mu_x - \mu_y \geq D_0$ $H_1 : \mu_x - \mu_y < D_0$	$z < -z_\alpha$
$H_o : \mu_x - \mu_y = D_0$ hoặc Ho: $\mu_x - \mu_y \leq D_0$ $H_1 : \mu_x - \mu_y > D_0$	$z > z_\alpha$

Ví dụ: Một trại chăn nuôi chọn 1 giống gà để tiến hành nghiên cứu hiệu quả của hai loại thức ăn mới A và B. Sau một thời gian nuôi thử nghiệm người ta chọn 50 con gà nuôi bằng thức ăn A và thấy trọng lượng trung bình 1 con là 2,2kg độ lệch chuẩn là 1,25 kg và 40 con gà nuôi bằng thức ăn B, trọng lượng trung bình 1 con là 1,2 kg độ lệch chuẩn là 1,02 kg. Giả sử ta muốn kiểm định giả thuyết H_o cho rằng trọng lượng trung bình của 1 con gà sau một thời gian nuôi trong hai trường hợp là như nhau với mức ý nghĩa $\alpha = 0,05$.

Giải:

Gọi μ_x, μ_y là trọng lượng trung bình của 1 con gà nuôi bằng thức ăn A và B.

Ta đặt giả thuyết: $H_o : \mu_x - \mu_y = 0$

$H_1 : \mu_x - \mu_y \neq 0$

Ta có : $\bar{x} = 2.2 ; s_x = 1,25 ; n_x = 50$

$\bar{y} = 1.2 ; s_y = 1,02 ; n_y = 40 \quad z_{\alpha/2} = z_{0,025} = 1,96$

Giá trị kiểm định : $$z = \frac{2,2 - 1,2}{\sqrt{\frac{1,25^2}{50} + \frac{1,02^2}{40}}} = 4,179$$

Vì $|z| = 4,179 > 1,96$ nên ta bác bỏ giả thiết H_o.

Do $\bar{x} > \bar{y}$ nên $\mu_x > \mu_y$.

Như vậy với mức ý nghĩa $\alpha = 0,05$ ta có thể nói rằng trọng lượng trung bình của 1 con gà nuôi bằng thức ăn A lớn hơn nuôi bằng thức ăn B.

Trong trường hợp mẫu nhỏ n_x hoặc n_y hoặc cả hai < 30.

Ta có thể thực hiện kiểm định về sự khác nhau giữa hai số trung bình của hai tổng thể với giả định cả hai tổng thể có phân phối chuẩn và phương sai của hai tổng thể bằng nhau.

- Ta cần kiểm định giả thuyết: $H_o : \mu_x - \mu_y = D_o$

$H_1 : \mu_x - \mu_y \neq D_o$

- Quy tắc thực hành:

Tính giá trị kiểm định : $$t = \frac{\bar{x} - \bar{y} - D_0}{\sqrt{s^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \qquad (8.7)$$

Trong đó: $$s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)}$$

Từ α, tra bảng phân phối student với $n_x + n_y - 2$ bậc tự do để tìm $t_{n_x+n_y-2,\alpha/2}$

+ Nếu $|t| > t_{n_x+n_y-2,\alpha/2}$ ta bác bỏ giả thuyết.

+ Nếu $|t| \le t_{n_x+n_y-2,\alpha/2}$ ta chấp nhận giả thuyết.

Trường hợp kiểm định một bên, ta có bảng tóm tắt sau:

Giả thuyết	Bác bỏ Ho khi:
$H_0 : \mu_x - \mu_y = D_0$ $H_1 : \mu_x - \mu_y \neq D_0$	$t > t_{n_x+n_y-2,\alpha/2}$ hoặc $t < -t_{n_x+n_y-2,\alpha/2}$ Hay $\lvert t\rvert > t_{n_x+n_y-2,\alpha/2}$
$H_0 : \mu_x - \mu_y = D_0$ hoặc Ho: $\mu_x - \mu_y \ge D_0$ $H_1 : \mu_x - \mu_y < D_0$	$t < -t_{n_x+n_y-2,\alpha}$
$H_0 : \mu_x - \mu_y = D_0$ hoặc Ho: $\mu_x - \mu_y \le D_0$ $H_1 : \mu_x - \mu_y > D_0$	$t > t_{n_x+n_y-2,\alpha}$

Ví dụ: Ban lãnh đạo một công ty cho rằng doanh số tăng lên sau khi thực hiện các biện pháp khuyến mãi. Chọn ngẫu nhiên 13 tuần trước khi thực hiện các biện pháp khuyến mãi và 14 tuần sau khi thực hiện các biện pháp khuyến mãi. Doanh số trung bình và độ lệch chuẩn tính được lần lượt là 1234, 1864 và 324,289 (triệu đồng). Hãy kiểm định ý kiến trên với mức ý nghĩa α = 5%.

Giải

Gọi μ_x, μ_y lần lượt là doanh số trung bình sau và trước khi thực hiện các biện pháp khuyến mãi.

- Ta đặt giả thuyết:

$$H_0 : \mu_x - \mu_y = 0$$
$$H_1 : \mu_x - \mu_y \neq 0$$

Trong trường hợp này n_x, $n_y < 30$, ta giả định phương sai của hai tổng thể về doanh số trước và sau khi thực hiện các biện pháp khuyến mãi là bằng nhau.

- Để ước lượng phương sai ta tính:

$$s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)} = \frac{(14-1)289^2 + (13-1)324^2}{(14+13-2)}$$

$$= \frac{2345485}{25} = 93819,4$$

Tính giá trị kiểm định:

$$t = \frac{\bar{x} - \bar{y} - D_0}{\sqrt{s^2(\frac{1}{n_x} + \frac{1}{n_y})}} = \frac{1864 - 1234}{\sqrt{93819,4(\frac{1}{14} + \frac{1}{13})}} = 5,34$$

$$\alpha = 5\%, \quad t_{n_x + n_y - 2, \alpha/2} = t_{25, 2,5\%} = 1,708$$

Vì $|t| = 5,34 > 1,708$ nên ta bác bỏ giả thuyết H_0 nghĩa là $\mu_x \neq \mu_y$, do $\bar{x} > \bar{y}$ nên $\mu_x > \mu_y$. Như vậy với mức ý nghĩa 5% ta có thể nói rằng doanh số trung bình sau khi áp dụng các biện pháp khuyến mãi đã tăng lên.

8.7. KIỂM ĐỊNH GIẢ THUYẾT VỀ SỰ BẰNG NHAU GIỮA HAI PHƯƠNG SAI CỦA TỔNG THỂ:

Giả sử ta có hai tổng thể phân phối chuẩn với phương sai tương ứng là σ_x^2, σ_y^2. Ta cần kiểm định giả thuyết:

$$H_0 : \sigma_x^2 = \sigma_y^2$$

$$H_1 : \sigma_x^2 \neq \sigma_y^2$$

Căn cứ vào hai mẫu n_x, n_y được chọn ngẫu nhiên độc lập từ hai tổng thể với phương sai mẫu tương ứng s_x^2, s_y^2. ta đưa ra quy tắc để kết luận là chấp nhận hay bác bỏ H_0 với mức ý nghĩa α.

- Quy tắc thực hành:

Từ hai mẫu cụ thể ta tính giá trị kiểm định: $\frac{s_x^2}{s_y^2}$ (8.8)

(với giả thiết $s_x^2 > s_y^2$, nếu không ta đặt ngược lại).

Ta tra bảng Fisher – Snedecor với n_x-1 và n_y-1 bậc tự do để tìm $F_{n_x-1,n_y-1,\alpha/2}$.

+ Nếu $\frac{s_x^2}{s_y^2} \le F_{n_x-1,n_y-1,\alpha/2}$ ta chấp nhận giả thuyết.

+ Nếu $\frac{s_x^2}{s_y^2} > F_{n_x-1,n_y-1,\alpha/2}$ ta bác bỏ giả thuyết.

Trong trường hợp bác bỏ giả thuyết nghĩa là $\sigma_x^2 \ne \sigma_y^2$, vì $s_x^2 > s_y^2$ nên ta kết luận $\sigma_x^2 > \sigma_y^2$.

Trường hợp kiểm định một bên, ta có giả thuyết:

$$Ho: \sigma_x^2 = \sigma_y^2 \text{ hoặc } Ho: \sigma_x^2 \le \sigma_y^2$$

$$H_1: \sigma_x^2 > \sigma_y^2$$

Giả thuyết Ho bị bác bỏ nếu $\frac{s_x^2}{s_y^2} > F_{n_x-1,n_y-1,\alpha}$

Ví dụ: để kiểm tra độ chính xác của hai chiếc máy tiện người ta chọn ngẫu nhiên từ máy 1 (x) ra 15 sản phẩm, từ máy 2 (y) ra 13 sản phẩm, phương sai về đường kính sản phẩm tính được lần lượt là 17 và 26. Với mức ý nghĩa α = 5% có thể kết luận hai máy có độ chính xác như nhau không?

Giải:

Ta đặt giả thuyết: $H_o: \sigma_x^2 = \sigma_y^2$

$H_1: \sigma_x^2 \ne \sigma_y^2$

Giá trị kiểm định: $\frac{s_y^2}{s_x^2} = \frac{26}{17} = 1{,}529$

Tra bảng phân phối F ta có $F_{12,14,0.025} = 3{,}05$ vì $1{,}529 < 3{,}05$ nên ta chấp nhận giả thuyết nghĩa là hai máy có độ chính xác như nhau.

8.8. KIỂM ĐỊNH GIẢ THIẾT VỀ SỰ BẰNG NHAU GIỮA HAI TỶ LỆ TỔNG THỂ:

Ta có hai tổng thể X và Y; gọi p_x, p_y và $\hat{p}_x$, $\hat{p}_y$ là tỷ lệ các đơn vị có tính chất nào đó mà ta quan tâm trong tổng thể và trong mẫu.

- Ta cần kiểm định giả thuyết : $H_o : p_x - p_y = 0$

$H_1 : p_x - p_y \neq 0$

Căn cứ vào hai mẫu n_x, n_y với (n_x, $n_y \geq 40$) được chọn ngẫu nhiên độc lập từ hai tổng thể X và Y ta đưa ra quy tắc để chấp nhận hay bác bỏ giả thuyết trên với mức ý nghĩa α.

- Quy tắc thực hành:

Tính giá trị kiểm định:

$$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}_o(1-\hat{p}_o)\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \qquad (8.9)$$

Trong đó $\hat{p}_o$ là tỷ lệ các phần tử có tính chất nào đó chung trong hai mẫu

$$\hat{p}_o = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y}$$

Biết α ta tìm $z_{\alpha/2}$.

+ Nếu $|z| \leq z_{\alpha/2}$ ta chấp nhận giả thuyết.

+ Nếu $|z| > z_{\alpha/2}$ ta bác bỏ giả thuyết.

Trong trường hợp này nếu $\hat{p}_x > \hat{p}_y$ ta kết luận $p_x > p_y$ còn nếu $\hat{p}_x < \hat{p}_y$ ta kết luận $p_x < p_y$.

Trường hợp kiểm định một bên, ta có bảng tóm tắt sau:

Giả thuyết	Bác bỏ Ho khi:
$H_o : p_x - p_y = 0$ $H_1 : p_x - p_y \neq 0$	$z > z_{\alpha/2}$ hoặc $z < -z_{\alpha/2}$ Hay $\lvert z\rvert > z_{\alpha/2}$
$H_o : p_x - p_y = 0$ hoặc Ho: $p_x - p_y \geq 0$ $H_1 : p_x - p_y < 0$	$z < -z_\alpha$

$H_0 : p_x - p_y = 0$ hoặc Ho: $p_x - p_y \le 0$ $H_1 : p_x - p_y > 0$	$z > z_\alpha$

Ví dụ: Để so sánh tỷ lệ trẻ em béo phì ở thành thị và nông thôn người ta tiến hành chọn ngẫu nhiên 200 em ở thành thị thấy có 20 em béo phì và chọn 220 em ở nông thôn thấy có 5 em béo phì.

Hãy kiểm định giả thiết H_0 cho rằng tỷ lệ béo phì của trẻ em thành thị và nông thôn là như nhau với mức ý nghĩa $\alpha = 1\%$.

Giải

- Đặt giả thuyết: $H_0 : p_x - p_y = 0$

$H_1 : p_x - p_y \neq 0$

Trong đó p_x, p_y là tỷ lệ trẻ em béo phì ở thành thị và nông thôn.

- Ta tính:

$$\hat{p}_o = \frac{20+5}{200+220} = 0{,}0595$$

$$\hat{p}_x = \frac{20}{200} = 0{,}1$$

$$\hat{p}_y = \frac{5}{220} = 0{,}0227$$

Giá trị kiểm định:

$$z = \frac{0{,}1 - 0{,}0227}{\sqrt{0{,}0595(1-0{,}0595)(\frac{1}{200}+\frac{1}{220})}} = 3{,}34$$

$$\alpha = 1\% \text{ -> } z_{\alpha/2} = z_{0{,}005} = 2{,}58$$

Vì $|z| = 3{,}34 > 2{,}58$ nên ta bác bỏ giả thiết H_0, tức là $p_x \neq p_y$. Do $\hat{p}_x = 0{,}1 > \hat{p}_y = 0{,}0227$ nên ta kết luận $p_x > p_y$ nghĩa là tỷ lệ béo phì của trẻ em thành thị cao hơn trẻ em nông thôn. Kết luận này có khả năng phạm sai lầm (loại 1) là 1%.

CHƯƠNG 9

PHÂN TÍCH PHƯƠNG SAI [1]

Mục tiêu của phân tích phương sai là so sánh trung bình của nhiều nhóm (tổng thể) dựa trên các trung bình mẫu và thông qua kiểm định giả thuyết để kết luận về sự bằng nhau của các trung bình 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 này đến một yếu tố kết quả kia. Ví dụ như nghiên cưu ảnh hưởng của thời gian đi làm thêm đến kết quả học tập của sinh viên. Nếu thời gian đi làm thêm của sinh viên là dữ liệu định tính (dưới 6 giờ tuần, 6-12 giờ/tuần, trên 12 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 đi làm thêm khác nhau lại có kết quả điểm trung bình học tập bằng nhau, chúng ta kết luận được rằng ảnh hưởng của yếu tố nguyên nhân thời gian đi làm thêm đến yếu tố kết quả học tập của những nhóm sinh viên có thời gian làm thêm 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 đi làm thêm nhiều (trên 12 giờ/tuần) có kết quả học tập thấp hơn 2 nhóm kia một cách có ý nghĩa thống kê, thì kết luận rút ra là thời gian đi làm thêm khác nhau có mức độ ảnh hưởng khác nhau 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. Yếu tố ở đây là số lượng yếu tố nguyên nhân ảnh hưởng đến yếu tố kết quả đang nghiên cứu.

9.1 PHÂN TÍCH PHƯƠNG SAI MỘT YẾU TỐ [2]

Phân tích phương sai một yếu tố là phân tích ảnh hưởng của một yếu tố nguyên nhân (định tính) ảnh hưởng đến một yếu tố kết quả (định lượng) đang nghiên cứu. Ví dụ như xem xét ảnh hưởng của thời gian đi làm thêm của sinh viên đến kết quả học tập. Căn cứ vào thời gian đi làm thêm ta có

---

[1] Analysis of Variance - ANOVA

[2] One-way ANOVA

3 nhóm sinh viên cần so sánh về điểm trung bình học tập.

9.1.1 Trường hợp k tổng thể có phân phối chuẩn và phương sai bằng nhau

Giả sử rằng chúng ta muốn so sánh trung bình của k tổng thể có phương sai bằng nhau dựa trên những mẫu ngẫu nhiên độc lập gồm $n_1, n_2, ..., n_k$ quan sát từ k tổng thể khác nhau có phân phối chuẩn. Nếu trung bình của các tổng thể được kí hiệu là $\mu_1, \mu_2, ..., \mu_k$ thì mô hình phân tích phương sai một yếu tố ảnh hưởng được mô tả dưới dạng kiểm định giả thuyết như sau:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$

Giả thuyết H_0 cho rằng trung bình của k tổng thể đều bằng nhau. Để kiểm định giả thuyết này, ta thực hiện các bước sau:

Bước 1: Tính các trung bình mẫu

Trước hết ta tính các trung bình mẫu từ những quan sát của các mẫu ngẫu nhiên độc lập ($\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$) và trung bình chung của k mẫu quan sát ($\bar{x}$) 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_{k1}
...	...	...	...
x_{1n_1}	x_{2n_2}	...	x_{kn_k}

Tính trung bình mẫu $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ theo công thức

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad (i = 1,2,...,k)$$

Và trung bình chung của k mẫu:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$$

Bước 2: Tính các tổng các độ lệch bình phương

Tính tổng các độ lệch bình phương trong nội bộ nhóm SSW[1] và tổng các độ lệch bình phương giữa các nhóm SSG[2]

- Tổng các độ lệch bình phương trong nội bộ nhóm (SSW) được tính bằng cách tính tổng các độ lệch bình phương giữa các giá trị quan sát với trung bình mẫu trong từng nhóm, rồi lấy tổng cộng tất cả các nhóm lại. SSW phản ảnh phần biến thiên của yếu tố kết quả do ảnh hưởng của các yếu tố khác, không phải do yếu tố nguyên nhân đang nghiên cứu (yếu tố dùng để phân nhóm) [3]

Tổng các độ lệch bình phương của từng nhóm được tính theo công thức:

nhóm 1: $$SS_1 = \sum_{j=1}^{n_1}(x_{1j} - \bar{x}_1)^2$$

nhóm 2: $$SS_2 = \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2$$

Tương tự như vậy ta tính đến nhóm thứ k. Vậy tổng các độ lệch bình phương trong nội bộ các nhóm được tính như sau:

$$SSW = SS_1 + SS_2 + ... + SS_k$$

hay $$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

- Tổng các độ lệch bình phương giữa các nhóm (SSG) được tính bằng cách tính tổng các độ lệch bình phương giữa các trung bình mẫu của từng nhóm với trung bình chung của k 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 độ lệch bình phương toàn bộ SST[4] được tính bằng cách tính tổng các độ lệch bình phương giữa tất cả các giá trị quan sát của toàn bộ mẫu nghiên cứu 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.

---

[1] Sum of squares within group

[2] Sum of squares between group

[3] Diễn giải ý nghĩa của các tổng các độ lệch được trình bày chi tiết trước ví dụ tính toán

[4] Total sum of squares

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$$

Có thể dễ dàng chứng minh là tổng các độ lệch bình phương toàn bộ bằng tổng cộng tổng các độ lệch bình phương trong nội bộ các nhóm và tổng các độ lệch bình phương giữa các nhóm.

$$SST = SSW + SSG$$

Như vậy toàn bộ biến thiên của yếu tố kết quả được phân tích thành 2 phần: phần biến thiên do yếu tố đang nghiên cứu tạo ra và phần biến thiên còn lại do các yếu tố khác tạo ra không nghiên cứu ở đây. Nếu phần biến thiên do yếu tố nguyên nhân tạo ra càng "đáng kể" so với phần biến thiên do các yếu tố khác tạo ra, thì chúng ta càng có cơ sở để bác bỏ H_0 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 (trung bình của các độ lệch bình phương)
Các phương sai được tính bằng cách lấy các tổng các độ lệch bình phương chia cho bậc tự do tương ứng.

Tính phương sai trong nội bộ nhóm (MSW) bằng cách lấy tổng các độ 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 phương sai của yếu tố kết quả do các yếu tố khác tạo ra.

$$MSW = \frac{SSW}{n-k}$$

Tính phương sai giữa các nhóm (MSG) bằng cách lấy tổng các độ lệch bình phương giữa các nhóm chia cho bậc tự do tương ứng là k-1. MSG là ước lượng phần phương sai của yếu tố kết quả do yếu tố nguyên nhân đang nghiên cứu tạo ra.

$$MSG = \frac{SSG}{k-1}$$

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 cuả hai phương sai: phương sai giữa các nhóm (MSG) và phương sai trong nội bộ nhóm (MSW).

$$F = \frac{MSG}{MSW}$$

Bác bỏ giả thuyết H_0 cho rằng trung bình của k tổng thể đều bằng nhau khi:

$$F > F_{k-1,n-k,\alpha}$$

$F_{k-1,n-k,\alpha}$ là giá trị giới hạn tra từ bảng phân phối F với k -1 bậc tự do ở tử số và n -k bậc tự do ở mẫu số ở mức ý nghĩa α.

Sau đây là dạng bảng 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		

Bảng dịch qua tiếng Việt

Nguồn biến thiên	Tổng các độ 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 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 đi làm thêm của các sinh viên đến kết quả học tập của sinh viên đã đề cập ở đầu chương. 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: Ảnh hưởng của thời gian đi làm thêm đến kết quả học tập của sinh viên có thời gian làm thêm dài ngắn khác nhau là như nhau.

Nếu giả thuyết H_0 đúng, ảnh hưởng của thời gian đi làm thêm này là như nhau đối với các nhóm sinh viên có thời gian đi làm thêm khác nhau (tức

là kết quả học tập của các sinh viên này do các yếu tố khác tạo ra như sức khỏe, mức độ liên quan của công việc làm thêm đến ngành đang học, phương pháp học ...) thì trong nội bộ 3 nhóm, điểm trung bình học tập rất phân tán. Cùng nhóm thời gian làm thêm ít (dưới 6 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 rất cao, tính trung bình cả nhóm thì điểm trung bình không cao cũng không thấp, và không khác biệt nhiều với 2 nhóm kia.

Tương tự, trong nhóm thời gian làm thêm nhiều (trên 12 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 rất cao, 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 không nghiên cứu ở đây, các sinh viên có thời gian đi làm thêm như nhau, có kết quả học tập khác nhau do tình trạng sức khỏe, điều kiện ăn ở, sinh hoạt, học tập, công việc làm thêm có liên quan đến ngành học hay không, ... Kết quả là 3 trung bình mẫu của 3 nhóm so sánh rất gần nhau, và rất gần với trung bình chung cả 3 nhóm. Lúc đó tổng các độ lệch bình phương giữa các nhóm (SSG) nhỏ và phương sai giữa các nhóm nhỏ (MSG), còn tổng các độ lệch bình phương trong nội bộ 3 nhóm (SSW) rất lớn (vì điểm trung bình học tập trong cùng 1 nhóm rất khác nhau) và phương sai trong nội bộ nhóm (MSW) lớn. Như vậy khi ảnh hưởng của nguyên nhân (thời gian làm thêm) đến kết quả học tập không khác nhau 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 đi làm thêm không 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 thì ta chấp nhận giả thuyết H_0.

Nếu giả thuyết H_0 sai, tức là quả thật lượng thời gian làm thêm của sinh viên có ảnh hưởng đến kết quả học tập của sinh viên, thì trong nhóm các sinh viên đi làm thêm nhiều (trên 12giờ/tuần), sinh viên nào cũng đều có kết quả điểm trung bình học tập thấp, điểm trung bình trong nhóm này ít phân tán, và khá đồng đều (đều thấp!). Các sinh viên trong nhóm đi làm thêm ít (dưới 6 giờ/tuần), hầu hết đều có kết quả trung bình, khá trở lên. 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 độ lệch bình phương giữa các nhóm (SSG) lớn và phương sai giữa các nhóm (MSG) lớn, còn tổng các độ 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 làm thêm khác nhau có ảnh hưởng khác nhau đến kết quả học tập của sinh viên.

Ví dụ tính toán: Ban giám hiệu trường đại học K muốn nghiên cứu ảnh hưởng của việc đi làm thêm đến kết quả học tập của sinh viên. Một sinh viên năm thứ 3 dưới sự hướng dẫn và khuyến khích của một giáo viên đã nhận đề tài này làm đề tài nghiên cứu khoa học sinh viên. Một trong những nội dung nghiên cứu là xem xét thời gian đi làm thêm khác nhau có ảnh hưởng đến kết quả học tập của sinh viên như nhau không. Một cuộc khảo sát với cỡ mẫu là 120 sinh viên được thực hiện.

Trong 120 sinh viên được thu thập dữ liệu, có 22 sinh viên cho biết có công việc làm thêm đều đặn ít nhất là 16 tuần trong năm học qua. Có 7 sinh viên thời gian làm việc ít, dưới 6 giờ/tuần. Bảy sinh viên khác có thời gian làm việc trung bình, khoảng từ 6 đến 12 giờ/tuần. Còn lại 8 sinh viên làm việc nhiều, trên 12 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 sinh viên thực hiện đề tài được trình bày trong Bảng 9.3.

Bảng 9.3: Điểm trung bình học tập của các sinh viên có đi làm thêm

Nhóm 1 (TG làm thêm ít)	Nhóm 2 (TG làm thêm TB)	Nhóm 3 (TG làm thêm nhiều)
6,3	7,2	6,3
7,0	6,6	5,8
6,5	6,1	6,0
6,6	5,8	5,5
7,2	6,8	5,2
6,9	7,1	6,5
6,4	5,9	5,3
		6,2
T.cộng 46,9	45,5	46,8

Phát biểu giả thuyết H_0: Thời gian làm thêm khác nhau có ảnh hưởng như nhau đến kết quả học tập của sinh viên: điểm học tập trung bình của 3 nhóm sinh viên có thời gian làm thêm khác nhau là bằng nhau.

H_0: $\mu_1 = \mu_2 = \mu_3$

Bước 1: Tính các trung bình của từng nhóm và trung bình chung 3 nhóm Điểm trung bình học tập (ĐTB) của sinh viên:

Nhóm 1: $\bar{x}_1 = \frac{46,9}{7} = 6,7$

Nhóm 2: $\bar{x}_2 = \frac{45,5}{7} = 6,5$

Nhóm 3: $\bar{x}_3 = \frac{46,8}{8} = 5,85$

Cả 3 nhóm: $\bar{x} = \frac{7 \times 6,7 + (7 \times 6,5) + (8 \times 5,85)}{7+7+8} = 6,3273$

Bước 2: Tính các tổng các độ lệch bình phương

- Tổng các độ lệch bình phương nội bộ 3 nhóm: $SSW = SS_1 + SS_2 + SS_3$

$$SS_1 = \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2 = (6,3 - 6,7)^2 + (7,0 - 6,7)^2 + ... + (6,4 - 6,7)^2$$

$$= 0,68$$

Tương tự:

$$SS_2 = (7,2 - 6,5)^2 + (6,6 - 6,5)^2 + ...+ (5,9 - 6,5)^2 = 1,96$$

$$SS_3 = (6,3 - 5,85)^2 + (5,8 - 5,85)^2 + ...+ (6,2 - 5,85)^2 = 1,62$$

$$\Rightarrow SSW = 0,68 + 1,96 + 1,62 = 4,26$$

- Tổng các độ lệch bình phương giữa các nhóm: $SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 = 7(6,7 - 6,3273)^2 + 7(6,5 - 6,3273)^2 + 8(5,85 - 6,3273)^2$$

$$= 3,004$$

Bước 3: Tính các phương sai

Phương sai trong nội bộ nhóm:

$$MSW = \frac{SSW}{n-k} = \frac{4,26}{22-3} = 0,224$$

Phương sai giữa các nhóm:

$$MSG = \frac{SSG}{k-1} = \frac{3,004}{3-1} = 1,502$$

Bước 4: Tính tỉ số F

$$F = \frac{MSG}{MSW} = \frac{1,502}{0,224} = 6,7$$

Tra bảng phân phối F với mức ý nghĩa $\alpha = 0{,}05$, ta có:

$$F_{k-1,n-k,\alpha} = F_{2,\,19,\,0.05} = 3{,}52$$

Vì F = 6,7 > 3,52 cho nên dựa trên dữ liệu đã thu thập, chúng ta có đủ bằng 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 làm thêm khác nhau thì khác nhau. Người nghiên cứu có thể kết luận rằng, thời gian làm thêm khác nhau có ảnh hưởng khác nhau đến kết quả học tập của sinh viên có đi làm thêm. Sau đây là bảng 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

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
it	7	46.9	6.7	0.11333
TB	7	45.5	6.5	0.32667
nhiều	8	46.8	5.85	0.23143

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3.0036	2	1.5018	6.6983	0.0063	3.5219
Within Groups	4.26	19	0.224211			
Total	7.2636	21				

9.1.2 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, 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 nhóm 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[1]. Nội dung của phương pháp này là so sánh từng cặp các trung bình

[1] Honestly Significant Differences

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.

$$C_3^2 = \frac{3!}{2!(3-2)!} = 3$$

Các giả thuyết cần kiểm định 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 (studentized range distribution) ở mức ý nghĩa α, với bậc tự do k và n-k
- 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.

Từ ví dụ tính toán ở phần trước, ta có k = 3, α = 5%, n = 22 và MSW = 0,224. Tra bảng phân phối q (phân phối Tukey) ta có: $q_{0,05,3,19} = 3,59$

Tính giá trị giới hạn Tukey: $T = 3,59\sqrt{\frac{0,224}{7}} = 0,642$

Độ lệch tuyệt đối các cặp trung bình mẫu như sau:

$$|\bar{x}_1 - \bar{x}_2| = 6,7 - 6,5 = 0,2$$

$$|\bar{x}_1 - \bar{x}_3| = 6,7 - 5,85 = 0,85$$

$$|\bar{x}_2 - \bar{x}_3| = 6,500 - 5,85 = 0,65$$

Như vậy, theo điều kiện bác bỏ giả thuyết H_0 thì:

- trung bình tổng thể μ_1 và μ_3 khác nhau vì $|\bar{x}_1 - \bar{x}_3| = 0,85 > T = 0.642$
- trung bình tổng thể μ_2 và μ_3 khác nhau vì $|\bar{x}_2 - \bar{x}_3| = 0,65 > T = 0.642$
- chưa kết luận được μ_1 và μ_2 khác nhau vì $|\bar{x}_1 - \bar{x}_2| = 0,2 < T = 0.642$

vì $\bar{x}_1 > \bar{x}_3$ nên $\mu_1 > \mu_3$, $\bar{x}_2 > \bar{x}_3$ nên $\mu_2 > \mu_3$.

Như vậy chúng ta có thể kết luận rằng điểm trung bình học tập của sinh viên có thời gian đi làm thêm nhiều khác với sinh viên có thời gian đi làm thêm trung bình hay ít. Tuy nhiên chưa thể kết luận có sự khác nhau về điểm trung bình học tập giữa hai nhóm sinh viên có thời gian làm thêm ít và trung bình. Dựa vào trung bình nhóm, chúng ta có thể thấy điểm trung bình học tập của nhóm có thời gian đi làm thêm nhiều thấp hơn hẳn hai nhóm kia (5,85 so với 6,5 và 6,7). Như vậy thời gian đi làm thêm 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 ước lượng khoảng cho chênh lệch giữa các nhóm có khác biệt có ý nghĩa thống kê. Ước lượng khoảng về chêch lệch giữa hai trung bình nhóm có khác biệt là:

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm \left(t_{n-k,\frac{\alpha}{2}} \sqrt{\frac{2MSW}{n_i}} \right)$$

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 phân tích sâu ANOVA. Chúng ta có thể thực hiện phân tích này bằng chương trình SPSS. Sau khi khai báo biến và nhập dữ liệu, ta vào lần lượt chọn các lệnh từ menu và hộp thoại như sau:

Statistics – Compare means – One-way ANOVA... chọn đưa biến yếu tố kết quả vào ô dependent list và biến yếu tố nguyên nhân đưa vào ô factor. Sau đó chọn nút Post Hoc ... và chọn phương pháp Tukey. Nhấn vào nút Continue và OK trên hộp thoại, kết quả sẽ xuất hiện như trong Hình 9.1

Hình 9.1: Kết quả ANOVA 1 yếu tố và kết quả phân tích sâu Tukey

Descriptives

DTB

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	7	6.7000	.3367	.1272	6.3887	7.0113	6.30	7.20
2	7	6.5000	.5715	.2160	5.9714	7.0286	5.80	7.20
3	8	5.8500	.4811	.1701	5.4478	6.2522	5.20	6.50
Total	22	6.3273	.5881	.1254	6.0665	6.5880	5.20	7.20

Test of Homogeneity of Variances

DTB

Levene Statistic	df1	df2	Sig.
1.770	2	19	.197

ANOVA

DTB

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.004	2	1.502	6.698	.006
Within Groups	4.260	19	.224		
Total	7.264	21			

Multiple Comparisons

Dependent Variable: DTB

Tukey HSD

(I) NHOM	(J) NHOM	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	1					
	2	.2000	.2531	.713	-.4430	.8430
	3	.8500*	.2451	.007	.2274	1.4726
2	1	-.2000	.2531	.713	-.8430	.4430
	2					
	3	.6500*	.2451	.040	2.742E-02	1.2726
3	1	-.8500*	.2451	.007	-1.4726	-.2274
	2	-.6500*	.2451	.040	-1.2726	-2.7424E-02
	3					

*. The mean difference is significant at the .05 level.

Homogeneous Subsets

DTB

Tukey HSD[a,b]

NHOM	N	Subset for alpha = .05	
		1	2
3	8	5.8500	
2	7		6.5000
1	7		6.7000
Sig.		1.000	.703

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 7.304.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

9.1.3 Trường hợp các tổng thể được giả định có phân phối bất kỳ (phương pháp phi tham số)

Phân tích phương sai với kiểm định F chỉ có thể áp dụng khi các nhóm (tổng thể) so sánh có phân phối chuẩn và phương sai bằng nhau. Trong trường hợp không thỏa điều kiện này, chúng ta có thể chuyển đổi dữ liệu 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 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 chuẩn, 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...., 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 $n = n_1 + n_2 + ... + n_k$ là tổng các quan sát thuộc các mẫu, và R_1, R_2, ..., 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: $\mu_1 = \mu_2 = ... = \mu_k$: Trung bình của k tổng thể đều bằng nhau. Ở đây ta sử dụng biến W thay cho tỉ số F trong phần tính 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) để so sánh,

giả thuyết H_0 bị bác bỏ khi:

$$W > \chi^2_{k-1,\alpha}$$

Trở lại ví dụ điểm trung bình học tập của ba nhóm sinh viên ta có kết quả như trong bảng sau. Trong cách xếp hạng này, điểm trung bình cao nhất của 1 sinh viên trong cả ba nhóm là 7,3 được xếp hạng 1. Tương tự, hạng được xếp cho đến điểm trung bình thấp nhất là 5,3 ở nhóm 3 là hạng thứ 22.

Bảng 9.5: 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	12,5	7,2	2	6,3	12,5
7,0	4	6,6	7,5	5,8	18,5
6,5	9,5	6,1	15	6,0	16
6,6	7,5	5,8	18,5	5,5	20
7,3	1	6,8	6	5,3	22
6,9	5	7,1	3	6,5	9,5
6,4	11	5,9	17	5,4	21
				6,2	14
	R_1=50,5		R_2=69,0		R_3=133,5

$$W = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{22(22+1)}\left[\frac{(50,5)^2}{7} + \frac{(69,0)^2}{7} + \frac{(133,5)^2}{8}\right] - 3(22+1) = 8,6$$

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 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 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 phần trước. 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:

$$\overline{R}_i = \frac{R_i}{n_i}$$

Bước 2: Tiếp theo chúng ta tính chênh lệch về hạng trung bình giữa 2 nhóm cần so sánh

$$D_{ij} = \left|\overline{R}_i - \overline{R}_j\right|$$

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.

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

$$\overline{R}_{ít} = \frac{R_{ít}}{n_{ít}} = \frac{50,5}{7} = 7,21 \qquad \overline{R}_{TB} = \frac{R_{TB}}{n_{TB}} = \frac{69,0}{7} = 9,86$$

$$\overline{R}_{nhiều} = \frac{R_{nhiều}}{n_{nhiều}} = \frac{133,5}{8} = 16,69$$

* tính các chênh lệch về hạng trung bình giữa từng cặp nhóm

- so sánh nhóm ít với TB: $D_{ít,TB} = |7,21 - 9,86| = 2,65$
- so sánh nhóm ít với nhiều: $D_{ít,nhiều} = |7,21 - 16,69| = 9,48$
- so sánh nhóm TB với nhiều: $D_{TB,nhiều} = |9,86 - 16,69| = 6,83$

* 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:

$$C_K = \sqrt{(\chi^2_{k-1,\alpha})\left(\frac{n(n+1)}{12}\right)\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = \sqrt{5{,}99\left(\frac{22(22+1)}{12}\right)\left(\frac{1}{7}+\frac{1}{8}\right)} = 8{,}23$$

* quyết định:

- $D_{ít,\,TB} = 2{,}65 < 8{,}5$: chấp nhận giả thuyết cho rằng trung bình hai đối tượng bằng nhau. Như vậy điểm trung bình của sinh viên có thời gian làm thêm ít và trung bình không khác biệt có ý nghĩa thống kê.
- $D_{ít,\,nhiều} = 9{,}48 > 8{,}23$: bác bỏ giả thuyết cho rằng trung bình hai đối tượng bằng nhau. Như vậy điểm trung bình của sinh viên có thời gian làm thêm ít và nhiều là có khác biệt có ý nghĩa thống kê.
- $D_{TB,\,nhiều} = 6{,}83 < 8{,}23$: 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 làm thêm trung bình và nhiều không khác biệt có ý nghĩa thống kê.

9.2 PHÂN TÍCH PHƯƠNG SAI HAI YẾU TỐ [1]

Phân tích phương sai hai yếu tố 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 đi làm thêm 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 có đi làm thêm, ví dụ như công việc làm thêm có liên quan hay phù hợp với chuyên ngành đang học không... 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.

[1] Two-way Analysis of Variance

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 x 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)

$$\bar{x}_i = \frac{\sum_{j=1}^{H} x_{ij}}{H} \qquad (i = 1,2,..., K)$$

Trung bình riêng cho từng khối - block (dòng)

$$\bar{x}_j = \frac{\sum_{i=1}^{K} x_{ij}}{K} \qquad (j = 1,2,..., H)$$

Trung bình chung của toàn bộ mẫu quan sát:

$$\bar{x} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{H} x_{ij}}{n} = \frac{\sum_{i=1}^{K} \bar{x}_i}{K} = \frac{\sum_{j=1}^{H} \bar{x}_j}{H}$$

<u>Bước 2</u>: tính tổng các độ lệch bình phương

1. Tổng các độ lệch bình phương chung: SST = SSG + SSB +SSE

$$SST = \sum_{i=1}^{K}\sum_{j=1}^{H}(x_{ij} - \bar{x})^2$$

SST phản ảnh biến thiên của yếu tố định lượng kết quả đang nghiên cứu do ảnh hưởng của tất cả các nguyên nhân.

2. Tổng các độ lệch bình phương giữa các nhóm: between – groups

$$SSG = H\sum_{i=1}^{K}(\bar{x}_i - \bar{x})^2$$

SSG phản ảnh 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 độ lệch bình phương giữa các khối: between – blocks

$$SSB = K\sum_{j=1}^{H}(\bar{x}_j - \bar{x})^2$$

SSB phản ảnh 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 độ 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 ảnh phần biến thiên của yếu tố định lượng kết quả đang nghiên cứu do ảnh hưởng của các yếu tố khác còn lại không nghiên cứu trong phân tích này.

<u>Bước 3</u>: 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:

$$F_1 = \frac{MSG}{MSE} \qquad F_2 = \frac{MSB}{MSE}$$

Bước 5: Có 2 trường hợp trong quyết định bác bỏ giả thuyết H_0 của ANOVA hai yếu tố:

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:

$$F_1 > F_{K-1,(K-1)(H-1),\alpha}$$

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 độ 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	$MSB = \frac{SSB}{H-1}$	$F_2 = \frac{MSB}{MSE}$
Phần 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)

Hàng (blocks)	Nhóm (groups)			
	1	2	. . .	K
1	$x_{111}\ x_{112} \ldots\ x_{11L}$	$x_{211}\ x_{212} \ldots\ x_{21L}$	. . .	$x_{K11}\ x_{K12} \ldots\ x_{K1L}$
2	$x_{121}\ x_{122} \ldots x_{12L}$	$x_{221}\ x_{222} \ldots\ x_{22L}$	. . .	$x_{K21}\ x_{K22} \ldots\ x_{K2L}$
..				
H	$x_{1H1}\ x_{1H2} \ldots x_{1HL}$	$x_{2H1}\ x_{2H2} \ldots\ x_{2HL}$	. . .	$x_{KH1}\ x_{KH2} \ldots\ x_{KHL}$

Bước 1: Tính các trung bình

Trung bình mẫu của từng nhóm – group (cột)

$$\bar{x}_i = \frac{\sum_{j=1}^{H}\sum_{s=1}^{L} x_{ijs}}{H \times L} \quad (j = 1,2, \ldots, H;\ s = 1,2, \ldots, L)$$

Trung bình mẫu của từng khối - block (dòng)

$$\bar{x}_j = \frac{\sum_{i=1}^{K}\sum_{s=1}^{L} x_{ijs}}{K \times L} \qquad (i = 1,2, \ldots, K;\ s = 1,2, \ldots, L)$$

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 độ lệch bình phương

1. Tổng các độ 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 độ lệch bình phương giữa các nhóm: between – groups

$$SSG = HL\sum_{i=1}^{K}(\bar{x}_i - \bar{x})^2$$

3. Tổng các độ lệch bình phương giữa các khối: between – blocks

$$SSB = KL\sum_{j=1}^{H}(\bar{x}_j - \bar{x})^2$$

4. Tổng các độ lệch bình phương giữa các ô (giao nhau giữa các nhóm và khối)

$$SSI = L\sum_{i=1}^{K}\sum_{j=1}^{H}(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$$

5. Tổng các độ lệch bình phương phần dư:

$$SSE = \sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{s=1}^{L}(x_{ijs} - \bar{x}_{ij})^2 = SST - SSG - SSB - SSI$$

<u>Bước 3</u>: Tính các phương sai:

1. Phương sai giữa các nhóm: $MSG = \dfrac{SSG}{K-1}$

2. Phương sai giữa các khối: $MSB = \dfrac{SSB}{H-1}$

3. Phương sai giữa các ô $MSI = \dfrac{SSI}{(K-1)\times(H-1)}$

4. Phương sai dư: $MSE = \dfrac{SSE}{K\times H\times(L-1)}$

<u>Bước 4</u>: 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:

$$F_1 = \frac{MSG}{MSE} \qquad F_2 = \frac{MSB}{MSE} \qquad F_3 = \frac{MSI}{MSE}$$

<u>Bước 5</u>: 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 tính toán: cũng từ ví dụ điểm trung bình học tập và thời gian làm thêm của sinh viên, chúng ta đưa thêm yếu tố mức độ liên quan của việc làm thêm và ngành đang học của sinh viên. Dữ liệu thu thập được trình bày trong bảng dưới đây.

Bảng 9.9: Điểm trung bình học tập của sinh viên phân nhóm theo thời gian làm thêm và mức độ liên quan giữa việc làm thêm và ngành học

Mức độ liên quan giữa việc làm thêm và ngành học	Thời gian làm thêm								
	ít Dưới 6 giờ/tuần			Trung bình 6-12 giờ/tuần			Nhiều Trên 12 giờ/tuần		
ít	6,3	6,5	6,6	6,5	5,9	6,4	5,4	5,3	5,1
Trung bình	7,1	6,8	6,7	6,4	6,8	7,0	6,7	6,5	6,4
nhiều	7,1	7,5	7,2	6,9	7,3	7,5	6,5	6,7	6,6

Các giả thuyết H_0:

1. Điểm trung bình học tập (ĐTB) của sinh viên có thời gian làm thêm khác nhau đều bằng nhau.
2. ĐTB của sinh viên có mức độ liên quan giữa việc làm thêm và ngành học khác nhau đều bằng nhau.
3. Không có ảnh hưởng tương tác giữa thời gian làm thêm và mức độ liên quan giữa việc làm thêm và ngành học đến ĐTB của sinh viên. Nói một cách cụ thể, ảnh hưởng của thời gian làm thêm đến ĐTB là như nhau đối với các nhóm sinh viên có mức độ liên quan giữa việc làm thêm và ngành học khác nhau; và ảnh hưởng của mức độ liên quan giữa việc làm thêm và ngành học đến ĐTB là như nhau đối với

các nhóm sinh viên có thời gian làm thêm 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 làm thêm ít $\bar{x}_1 = \frac{6,3+6,5+...+7,2}{9} = 6,87$

ĐTB của nhóm thời gian làm thêm TB $\bar{x}_2 = \frac{6,5+5,9+...+7,5}{9} = 6,74$

ĐTB của nhóm thời gian làm thêm nhiều $\bar{x}_3 = \frac{5,4+5,3+...+6,6}{9} = 6,13$

- 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 việc làm thêm liên quan ít đến ngành học

$$\bar{x}_1 = \frac{6,3+6,5+...+5,1}{9} = 6,0$$

ĐTB của nhóm việc làm thêm liên quan trung bình đến ngành học

$$\bar{x}_2 = \frac{7,1+6,8+...+6,4}{9} = 6,71$$

ĐTB của nhóm việc làm thêm liên quan nhiều đến ngành học

$$\bar{x}_3 = \frac{7,1+7,5+...+6,6}{9} = 7,03$$

- Trung bình một ô (cell means)

$$\bar{x}_{ij} = \frac{\sum_{s=1}^{L} x_{ijs}}{L}$$

ĐTB của SV có thời gian làm thêm ít và việc làm thêm liên quan ít với ngành học:

$$\bar{x}_{11} = \frac{6,3+6,5+6,6}{3} = 6,47$$

ĐTB của SV có thời gian làm thêm trung bình và việc làm thêm liên quan ít với ngành học:

$$\bar{x}_{12} = \frac{7,1 + 6,8 + 6,7}{3} = 6,87$$

...

ĐTB của SV có thời gian làm thêm nhiều và việc làm thêm liên quan nhiều với ngành học:

$$\bar{x}_{33} = \frac{6,5 + 6,7 + 6,6}{3} = 6,60$$

- Trung bình chung (overall mean):

$$\bar{x} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{H} \sum_{s=1}^{L} x_{ijs}}{K \times H \times L}$$

$$\bar{x} = \frac{6,3 + 6,5 + 6,6 + ... + 6,5 + 6,7 + 6,6}{3 \times 3 \times 3} = 6,58$$

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.

$$\bar{x} = \frac{\sum_{i=1}^{K} \bar{x}_i}{K}$$ (Tổng các trung bình nhóm chia cho số nhóm)

Kết quả tính các trung bình được trình bày tóm tắt trong bảng sau:

Bảng 9.10: Bảng tóm tắt kết quả tính các trung bình (SPSS for Windows)

Điểm TB

		TG làmthêm			Tổng
		ít	TB	nhiều	
liên quan giữa việc làm thêm và ngành học	ít	6.47	6.27	5.27	6.00
	TB	6.87	6.73	6.53	6.71
	nhiều	7.27	7.23	6.60	7.03
Tổng		6.87	6.74	6.13	6.58

Bước 2: Tính các tổng độ lệch bình phương (SS)

1. Tổng các độ lệch bình phương toàn bộ: SST

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{H} \sum_{s=1}^{L} (x_{ijs} - \bar{x})^2 =$$

$(6,3-6,58)^2 + (6,5-6,58)^2 + (6,6-6,58)^2 + ... + (5,4-6,58)^2 + (5,3-6,58)^2 + (5,1-6,58)^2 +$
$(7,1-6,58)^2 + (6,8-6,58)^2 + (6,7-6,58)^2 + ... + (6,7-6,58)^2 + (6,5-6,58)^2 + (6,4-6,58)^2 +$

$(7,1-6,58)^2 + (7,5-6,58)^2 + (7,2-6,58)^2 + \ldots + (6,5-6,58)^2 + (6,7-6,58)^2 + (6,6-6,58)^2$

SST = 9,441

2. Tổng các độ lệch bình phương bình phương giữa các nhóm (between – groups)

$$SSG = HL\sum_{i=1}^{K}(\bar{x}_i - \bar{x})^2 = 3x3x[(6,87-6,58)^2+(6,74-6,58)^2+(6,13-6,58)^2]$$

SSG = 2,779

3. Tổng các độ lệch bình phương giữa các khối (between – blocks)

$$SSB = KL\sum_{j=1}^{H}(\bar{x}_j - \bar{x})^2 = 3x3x[(6,00-6,58)^2+(6,71-6,58)^2+(7,03-6,58)^2]$$

SSB = 5,032

4. Tổng các độ lệch bình phương giữa các ô

$$SSI = L\sum_{i=1}^{K}\sum_{j=1}^{H}(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$$

$3 \times [(6,47-6,87-6,00+6,58)^2 + (6,27-6,74-6,00+6,58)^2 + (5,27-6,13-6,00+6,58)^2 + \ldots + (7,27-6,87-7,03+6,58)^2 + (7,23-6,74-7,03+6,58)^2 + (6,60-6,13-7,03+6,58)^2]$

SSI = 0,717

5. Tổng các độ 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 = 9,441 – 2,779 – 5,032 – 0,717 = 0,913

<u>Bước 3</u>: Tính các phương sai:

1. Phương sai giữa các nhóm: $MSG = \frac{SSG}{K-1} = \frac{2,779}{3-1} = 1,389$

2. Phương sai giữa các khối: $MSB = \frac{SSB}{H-1} = \frac{5,032}{3-1} = 2,516$

3. Phương sai giữa các ô: $MSI = \frac{SSI}{(K-1)\times(H-1)} = \frac{0,717}{(3-1)\times(3-1)} = 0,179$

4. Phương sai dư: $MSE = \frac{SSE}{K \times H \times (L-1)} = \frac{0,913}{3 \times 3 \times (3-1)} = 0,0507$

<u>Bước 4</u>: Tính tỉ số F

1. $F_1 = \frac{MSG}{MSE} = \frac{1,389}{0,0507} = 27,4$

2. $F_2 = \frac{MSB}{MSE} = \frac{2,516}{0,0507} = 49,6$

3. $F_3 = \frac{MSI}{MSE} = \frac{0,179}{0,0507} = 3,53$

Tra bảng F tìm

$F_{K-1,KH(L-1),\alpha} = F_{3-1,3\times3(3-1),0,05} = F_{2,18,0,05} = 3,55$

$F_{H-1,KH(L-1),\alpha} = F_{3-1,3\times3(3-1),0,05} = F_{2,18,0,05} = 3,55$

$F_{(K-1)(H-1),KH(L-1),\alpha} = F_{(3-1)(3-1),3\times3(3-1),0,05} = F_{4,18,0,05} = 2,93$

- Vì $F_1 > F_{K-1,KH(L-1),\alpha} = F_{2,18,0,05} = 3,55$, nên chúng ta có đủ bằng chứng để bác bỏ giả thuyết thứ nhất. Như vậy ĐTB của sinh viên có thời gian làm thêm khác nhau thì không bằng nhau. Nói cách khác, thời gian làm thêm có ảnh hưởng đến kết quả học tập.
- Vì $F_2 > F_{H-1,KH(L-1),\alpha} = F_{2,18,0,05} = 3,55$, nên chúng ta có đủ bằng chứng để bác bỏ giả thuyết thứ hai. Như vậy ĐTB của sinh viên có mức độ liên quan giữa việc làm thêm và ngành học khác nhau thì không bằng nhau. Nói cách khác, mức độ liên quan giữa việc làm thêm và ngành học của sinh viên có ảnh hưởng đến kết quả học tập.
- Vì $F_3 > F_{(K-1)(H-1),KH(L-1),\alpha} = F_{4,18,0,05} = 2,93$, nên chúng ta có đủ bằng chứng để bác bỏ giả thuyết thứ ba. Như vậy có tương tác giữa thời gian làm thêm và mức độ liên quan giữa việc làm thêm và ngành học trong việc ảnh hưởng đến ĐTB của sinh viên. Mức độ ảnh hưởng của thời gian làm thêm đến kết quả học tập còn bị ảnh hưởng bởi mức độ liên quan giữa việc làm thêm và ngành học. Trong Bảng 10.9, chúng ta thấy khi mức độ liên quan giữa việc làm thêm và ngành học ít thì thời gian làm thêm có ảnh hưởng mạnh đến kết quả học tập. Nhưng khi mức độ liên quan giữa việc làm thêm và ngành học nhiều thì ảnh hưởng của thời gian làm thêm đến kết quả học tập không khác nhau nhiều giữa các nhóm sinh viên có thời gian làm thêm 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 tính toán 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 kiểm định F trong ANOVA có nội dung cơ bản như sau:

Bảng 9.11: Bảng ANOVA hai yếu tố tổng quát

Nguồn biến động	Tổng các độ lệch bình phương	Bậc tự do	Phương sai	Tỉ số F
Giữa các nhóm (cột)	SSG	(K-1)	MSG	F_1
Giữa các khối (dòng)	SSB	(H-1)	MSB	F_2
Tương tác giữa 2 yếu tố	SSI	(K-1)(H -1)	MSI	F_3
Phần dư	SSE	KH (L-1)	MSE	
Tổng cộng	SST	KHL -1		

Kết quả ANOVA đầy đủ trong ví dụ tính toán trên thực hiện trên Excel được trình bày trong Bảng 9.12.

Bảng 9.12 : Kết quả phân tích phương sai 2 yếu tố bằng Excel

Anova: Two-Factor With Replication

SUMMARY	ít giờ	giờ TB	nhiều giờ	Total
ít				
Count	3	3	3	9
Sum	19.4	18.8	15.8	54
Average	6.46667	6.26667	5.26667	6
Variance	0.02333	0.10333	0.02333	0.3475
TB				
Count	3	3	3	9
Sum	20.6	20.2	19.6	60.4
Average	6.86667	6.73333	6.53333	6.711111
Variance	0.04333	0.09333	0.02333	0.061111
nhiều				
Count	3	3	3	9
Sum	21.8	21.7	19.8	63.3
Average	7.26667	7.23333	6.60000	7.033333
Variance	0.04333	0.09333	0.01000	0.1425
Total				
Count	9	9	9	
Sum	61.8	60.7	55.2	
Average	6.86667	6.74444	6.13333	
Variance	0.14750	0.24778	0.43750	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	5.03185	2	2.51593	49.5839	0.00000	3.5546
Columns	2.77852	2	1.38926	27.3796	0.00000	3.5546
Interaction	0.71704	4	0.17926	3.5328	0.02694	2.9277
Within	0.91333	18	0.05074			

Khi thực hiện ANOVA trên Excel, trong bảng kết quả ta có thêm một cột mang tên "F Critical". Cột này chính là giá trị giới hạn tra từ bảng thống kê (với mức ý nghĩa 5%) dùng để so sánh với cột "F" để quyết định bác bỏ giả thuyết H_0 hay không. Như vậy nếu sử dụng chương trình máy tính

để thực hiện thì ta không còn cần tra bảng thống kê nữa.

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 đị ıh 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): $T = q_{\alpha,K,KH(L-1)}\sqrt{\dfrac{MSE}{H \times L}}$

So sánh theo yếu tố thứ hai (H khối): $T = q_{\alpha,H,KH(L-1)}\sqrt{\dfrac{MSE}{K \times L}}$

Vận dụng vào ví dụ tính toán trong phần phân tích phương sai 2 yếu tố trên, với α = 0,05, K = 3, H = 3, L =3, MSE = 0,0507, tra bảng phân phối kiểm định Tukey (studentized range distribution) ta có:

$$q_{\alpha,K,KH(L-1)} = q_{0,05,3,18} = 3,61$$

* so sánh giữa các nhóm theo yếu tố thứ nhất (thời gian làm thêm): chúng ta tính giá trị giới hạn Tukey:

$$T = q_{0,05,3,18}\sqrt{\frac{0,0507}{9}} = 3,61(0,0751) = 0,27$$

Ta có các trung bình nhóm lần lượt là: 6,87 ; 6,74 ; 6,13 và các chênh lệch giữa các nhóm là:

$$D_{1,2} = |\bar{x}_1 - \bar{x}_2| = 6,87 - 6,74 = 0,13$$

$$D_{1,3} = |\bar{x}_1 - \bar{x}_3| = 6,87 - 6,13 = 0,74$$

$$D_{2,3} = |\bar{x}_2 - \bar{x}_3| = 6,74 - 6,13 = 0,61$$

Ta thấy $D_{1,3}$ và $D_{2,3}$ lớn hơn giá trị giới hạn Tukey T, cho nên chúng ta có thể nói rằng sinh viên có thời gian đi làm thêm ít và trung bình có điểm trung bình học tập khác với các sinh viên có thời gian đi làm thêm nhiều. Tuy nhiên chúng ta chưa có bằng chứng để cho rằng kết quả học tập khác nhau giữa sinh viên có thời gian đi làm thêm ít và trung bình.

* 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:

$$T = q_{0,05,3,18}\sqrt{\frac{0,0507}{9}} = 3,61(0,0751) = 0,27$$

Ta có các trung bình nhóm lần lượt là: 6,00 ; 6,71 ; 7,03 và các chênh lệch giữa các nhóm là:

$$D_{1,2} = |\bar{x}_1 - \bar{x}_2| = |6,00 - 6,71| = 0,71$$

$$D_{1,3} = |\bar{x}_1 - \bar{x}_3| = |6,00 - 7,03| = 1,03$$

$$D_{2,3} = |\bar{x}_2 - \bar{x}_3| = |6,71 - 7,03| = 0,32$$

Ta thấy cả $D_{1,3}$, $D_{2,3}$ và $D_{2,3}$ lớn hơn giá trị giới hạn Tukey T, cho nên chúng ta có thể nói rằng các nhóm sinh viên có mức độ liên quan giữa việc làm thêm và ngành học khác nhau thì có kết quả học tập khác nhau.

9.2.4 Thực hiện ANOVA trên chương trình Excel

Chúng ta có thể sử dụng các chương trình đang phổ biến hiện nay Excel, SPSS for Windows hoặc các chương trình thống kê khác. 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. Riêng SPSS chúng ta có thể vào menu "Analyze" chọn nội dung thống kê mà bạn muốn xử lý. Tuy nhiên, Excel trong Window có ưu điểm là chỉnh sửa dữ liệu gốc nhanh hơn và đơn giản hơn. 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 1 yếu tố, nhập liệu giống như Bảng 9.3. Đối với ANOVA 2 yếu tố có nhiều quan sát trong một ô, cần chú ý nhập liệu vào máy không giống như Bảng 9.9, mà phải nhập như Bảng 9.13. Nếu không, chương trình sẽ bị trục trặc hoặc cho ra kết quả sai. Kiểm tra kết quả in ra để biết chương trình nhập đúng hay sai bằng cách kiểm tra cột bậc tự do.

Bảng 9.13: Dữ liệu nhập vào Excel để thực hiện ANOVA 2 yếu tố

Mức độ phù hợp của việc làm thêm và ngành học	Thời gian làm thêm của sinh viên		
	ít giờ	giờ TB	nhiều giờ
ít	6.3	6.5	5.4
	6.5	5.9	5.3
	6.6	6.4	5.1
TB	7.1	6.4	6.7
	6.8	6.8	6.5
	6.7	7	6.4
nhiều	7.1	6.9	6.5
	7.5	7.3	6.7
	7.2	7.5	6.6

Bước 2:

Chọn Menu Tool – Add – Ins – đánh dấu 3 mục chọn:

- Analysis Toolpak
- MS Excel 4.0 Analysis Function
- MS Excel 4.0 Analysis Tools.

Bước 3:

- Chọn Tool – Data Analysis – [ ]. Trong khung [ ], chúng ta có thể chọn:
  - *ANOVA: Single Factor.* Phân tích phương sai một yếu tố.
  - Trong hai trường hợp còn lại:
  - *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. Chú ý, khi chọn vùng số liệu (select) thì chọn cả phần chữ (tiêu đề cột và tiêu đề dòng) và phần dữ liệu.
- Chọn vùng số liệu vừa mới nhập.
- Chọn α (α mặc nhiên là 5%)
- Chọn vùng chứa kết quả (nếu chọn New Worksheet thì kết quả được đặt trong một trang mới với đầy đủ các thông tin như được tính trong các công thức phần ví dụ).
- Nhấn phím "Enter"

Trong khung [ ], chúng ta có thể dùng các mục chọn khác cho các mục kiểm định hay phân tích thống kê khác như:

1. t – test paired Two sample for means: Kiểm định (t) trung bình của 2

tổng thể phụ thuộc (mẫu từng cặp).

2. t –test Two samples assuming equal variances: Kiểm định (t) trung bình của 2 tổng thể có phương sai bằng nhau.
3. t –test Two samples assuming unequal variances: Kiểm định (t) trung bình của 2 tổng thể có phương sai không bằng nhau.
4. z – test Two sample for means: Kiểm định (z) cho trung bình tổng thể.
5. Regression: hồi qui

...

CHƯƠNG 10

KIỂM ĐỊNH PHI THAM SỐ

Trong các chương trước, chúng ta thực hiện kiểm định giả thuyết về các tham số tổng thể với điều kiện là tổng thể có phân phối chuẩn. Khi tổng thể nghiên cứu không thỏa điều kiện này thì không thể sử dụng các công cụ đó được. Trong chương này, chúng ta mở rộng vấn đề kiểm định bằng cách xem xét các công cụ giúp kiểm định các giả thuyết khi tổng thể có phân phối bất kỳ, không theo phân phối chuẩn. Các kiểm định này không gắn liền với một tham số nào của tổng thể, cho nên được gọi là kiểm định phi tham số. Các kiểm định phi tham số không mạnh bằng các kiểm định tham số, nhưng chúng phù hợp khi dữ liệu phân tích là loại dữ liệu định tính (thang đo định danh, thứ bậc) hay các dữ liệu định lượng không có phân phối chuẩn một cách rõ ràng.

10.1 Kiểm định dấu

Kiểm định dấu là một thủ tục phi tham số đơn giản nhất được sử dụng cho **hai mẫu liên hệ** (mẫu từng cặp) để kiểm định giả thuyết phân phối của hai tổng thể tương ứng là giống nhau. Kiểm định này không cần giả thuyết nào về hình dạng của hai phân phối này.

Ví dụ như chúng ta cần so sánh mức độ thích chọn ngành học thống kê của sinh viên năm thứ hai trường Đại Học Kinh Tế trước và sau khi học môn Lý thuyết thống kê (LTTK). Mức độ thích được đo bằng thang đo khoảng cách 10 điểm, trong đó 1 là hoàn toàn không thích và 10 là rất thích. Hai tổng thể cần so sánh là mức độ thích chọn ngành học thống kê của các sinh viên trước khi học, và mức độ thích chọn ngành học thống kê của các sinh viên sau khi học. Vì không có điều kiện để thăm dò hết hàng ngàn sinh viên của trường ĐHKT, cho nên chúng ta thu thập ý kiến của 10 sinh viên được chọn ngẫu nhiên. Mười sinh viên này cho chúng ta mức độ thích của mình trước (đầu học kỳ) và sau khi học môn thống kê (cuối học kỳ), và chúng ta có dữ liệu mẫu dưới dạng từng cặp (2 mẫu phụ thuộc) như sau:

Bảng 10.1: Bảng tính toán kiểm định dấu về mức độ thích chọn ngành Thống kê trước và sau khi học LTTK

Sinh viên trả lời	Mức độ ưa thích		chênh lệch	dấu của đánh giá
	Sau khi học LTTK	Trước khi học LTTK		
1	8	6	+ 2	+
2	3	3	0	0
3	6	7	- 1	-
4	7	7	0	0
5	6	4	+ 2	+
6	8	5	+ 3	+
7	7	5	+ 2	+
8	5	6	- 1	-
9	5	6	- 1	-
10	9	7	+ 2	+

Giả thuyết H_0 là xác suất p để một sinh viên có mức độ ưa thích chọn ngành thống kê sau khi học môn LTTK cao hơn mức độ ưa thích chọn ngành thống kê trước khi học môn LTTK là 0,5. Nói một cách khác là tổng thể sinh viên không có khuynh hướng thích chọn ngành thống kê hơn sau khi học xong môn LTTK. Giả thuyết này tương tự nhưng tổng quát hơn giả thuyết trong kiểm định tham số tương ứng là kiểm định t trường hợp mẫu phụ thuộc – kiểm định về 2 trung bình tổng thể mẫu từng cặp. Tuy nhiên kiểm định dấu không đòi hỏi giả thuyết tổng thể (các mức độ thích) có phân phối đối xứng.

Chúng ta định nghĩa p là xác suất mà X (mức độ thích chọn ngành thống kê sau khi học LTTK) lớn hơn Y (mức độ thích chọn ngành thống kê trước khi học LTTK).

$$P = P(X > Y)$$

Theo giả thuyết Ho, xác suất X > Y hay xác suất Y > X đều là 0,5. Chúng ta bỏ qua những trường hợp X = Y. Khi X > Y ta có chênh lệch dương và ký hiệu dấu (+), khi X < Y ta có chênh lệch âm và ký hiệu dấu (-). Dưới dạng dấu các chênh lệch thì giả thuyết H_0 trong kiểm định này là khả năng có một dấu + (X>Y) bằng với khả năng có một dấu – (X<Y) và bằng 0,5. Các giả thuyết được trình bày như sau:

$$H_0: \quad p = 0,5$$

$$H_1: \quad p \neq 0,5$$

Kiểm định này giả thuyết rằng (1) các cặp dữ liệu (X,Y) là độc lập với nhau (sự thay đổi mức độ thích của một sinh viên này không ảnh hưởng đến sự thay đổi mức độ thích của một sinh viên khác, nói cách khác là các sinh

viên đều có ý kiến độc lập riêng của mình) và (2) thang đo sử dụng ít ra là thang đo thứ bậc trở lên. Sau khi loại bỏ các mẫu có mức độ bằng nhau, chúng ta chỉ còn lại số lượng các dấu (+) và (-). Đại lượng kiểm định được xác định:

T = số lượng các dấu cộng

Chúng ta xác định n là số cặp dữ liệu (đã loại bỏ các cặp có mức độ bằng nhau). Kiểm định này dựa trên phân phối nhị thức với p=0,5 tức là khi giả thuyết p = 0,5 đúng thì phân phối của số lượng các dấu cộng là phân phối nhị thức với p = 0,5 và n là số phép thử. Chúng ta tra bảng phân phối nhị thức và xác định các giá trị giới hạn tương ứng càng gần với mức ý nghĩa α càng tốt. Sau đó chúng ta so sánh đại lượng T với các giá trị giới hạn này. Trường hợp trên là loại kiểm định 2 phía, chúng ta tra bảng phân phối nhị thức để tìm 1 giá trị giới hạn đầu tiên sao cho càng gần với mức ý nghĩa $\alpha/2$ càng tốt. Giá trị giới hạn này gọi là C_1, và chúng ta xác định giá trị giới hạn thứ hai $C_2 = n - C_1$.

Tiêu chuẩn quyết định là:

$C_1 \le T \le C_2$ -> chấp nhận giả thuyết Ho

$T \le C_1$ hay $T \ge C_2$ -> bác bỏ giả thuyết Ho

Với các dữ liệu trong Bảng 10.1, ta có: T = 5 ; n = 8; p = 0,5. Chúng ta muốn kiểm định giả thuyết với độ tin cậy 95% (α = 0,05), tra bảng phân phối nhị thức với n = 8 và p = 0,5 thì thấy chỉ có 0,035 là gần với mức ý nghĩa $\alpha/2$ = 0,025 nhất, tương ứng với giá trị giới hạn C_1 = 1. Từ đây tính ra giá trị giới hạn thứ hai là $C_2 = n - C_1 = 8 - 1 = 7$.

Do $C_1 \le T \le C_2$ (1 < 5 < 7), nên ta chấp nhận giả thuyết H_o. Với mức ý nghĩa α = 2 x 0,035 = 0,07 (độ tin cậy 93%) chúng ta chưa có đủ bằng chứng để kết luận rằng sinh viên thích chọn ngành thống kê sau khi học môn LTTK.

Khi mẫu quan sát lớn, chúng ta có thể thực hiện kiểm định này dựa trên phân phối chuẩn thay vì phân phối nhị thức. Kiểm định này chính là kiểm định tỉ lệ tổng thể trong trường hợp mẫu lớn. Đại lượng kiểm định là:

$$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

Với $\hat{p} = \frac{T}{n}$ và po = 0,5 thì:

$$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} = \frac{2T - n}{\sqrt{n}}$$

Ví dụ: 50 người đi làm được chọn ngẫu nhiên để thu thập ý kiến về mức thu nhập trung bình của họ trong năm 2004 sắp đến sẽ như thế nào so với năm 2003. Kết quả thu thập được tổng hợp trong Bảng 10.2.

Bảng 10.2: Ý kiến về mức thu nhập của người đi làm

Ý kiến về mức thu nhập	Số người trả lời
Sẽ tăng lên	27
Sẽ không đổi	10
Sẽ giảm xuống	13

Hãy kiểm định giả thuyết H_o cho rằng tỷ lệ cho rằng thu nhập sẽ tăng lên bằng với tỷ lệ cho rằng thu nhập sẽ giảm xuống.

Giả thuyết: H_o: $p = 0,5$

H_1: $p \neq 0,5$

Áp dụng công thức $$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} = \frac{2T - n}{\sqrt{n}}$$

$$z = \frac{2(27) - 40}{\sqrt{40}} = 2,21$$

Với độ tin cậy 95% (α=0,05), giá trị giới hạn $z_{\alpha/2} = 1,96$; vì z =2,21 > 1,96 chúng ta bác bỏ giả thuyết H_o.

10.2 Kiểm định dấu và hạng Wilcoxon (kiểm định T)

Kiểm định dấu chỉ sử dụng chiều hướng (- hay +) của chênh lệch giữa các

cặp quan sát và bỏ qua độ lớn của các chênh lệch này. Kiểm định dấu và hạng Wilcoxon sử dụng luôn các thông tin về độ lớn của các chênh lệch và vì vậy, mạnh hơn kiểm định dấu. Để thực hiện kiểm định dấu và hạng Wilcoxon, các chênh lệch được xếp hạng theo độ lớn không tính đến dấu của chúng. Trong trường hợp chênh lệch bằng nhau thì hạng của chúng được tính bình quân (ví dụ hai chênh lệch đồng hạng 7 và 8 thì hạng của chúng là 7,5 và 7,5). Sau đó ta tính tổng các hạng đối với các chênh lệch dương và đối với các chênh lệch âm.

10.2.1.Trường hợp mẫu nhỏ (n $\leq$20)

Các bước tiến hành kiểm định:

1. Phát biểu giả thuyết H_0: Phân phối của hai tổng thể giống nhau
2. Tính chênh lệch giữa các cặp : $d_i = x_i - y_i$
3. Xếp hạng các d_i theo độ lớn của giá trị tuyệt đối $|d_i|$
4. Tìm tổng các hạng được xếp của d_i mang dấu dương: $\sum(+d_i)$
5. Tìm tổng các hạng được xếp của d_i mang dấu âm: $\sum(-d_i)$
6. T là giá trị nhỏ hơn trong 2 tổng các hạng

T = min [$\sum(+d_i), \sum(-d_i)$] và so sánh với T_α là giá trị tra từ bảng Wilcoxon

7. Bác bỏ giả thiết H_0 khi $T \leq T_\alpha$

Ví dụ: sử dụng các dữ liệu về mức độ thích chọn ngành thống kê trong phần trước, lần này chúng ta sử dụng không những dấu của chênh lệch mà còn sử dụng thêm độ lớn của các chênh lệch dưới dạng hạng của các chênh lệch. Chúng ta lập bảng tính toán như sau:

Bảng 10.3: Bảng tính toán kiểm định dấu và hạng về mức độ thích chọn ngành Thống kê trước và sau khi học LTTK

SV trả lời	Mức độ thích vào ngành thống kê		chênh lệch (d_i)	Xếp hạng $\|d_i\|$	Tổng $\sum(+d_i)$	Tổng $\sum(-d_i)$
	Sau khi học LTTK	Trước khi học LTTK				
1	8	6	+ 2	5,5	5,5	
2	3	3	0			
3	6	7	- 1	2		2
4	7	7	0			

5	6	4	+ 2	5,5	5,5	
6	8	5	+ 3	8	8	
7	7	5	+ 2	5,5	5,5	
8	5	6	- 1	2		2
9	5	6	- 1	2		2
10	9	7	+ 2	5,5	5,5	
Tổng					30	6

Giả thuyết H_0: Mức độ thích ngành học thống kê sau và trước khi học môn LTTK của sinh viên là như nhau.

Ta có:

$$\sum(+d_i) = 30 \text{ và } \sum(-d_i) = 6$$

$$\Rightarrow \quad T = \min\left|\sum(+d_i)\ ,\ \sum(-d_i)\right| = 6$$

Tra bảng phân phối của kiểm định Wilcoxon với n=8 ta có $T_{0,05} = 6$. Vậy giả thuyết H_0 mức độ thích ngành học thống kê sau và trước khi học môn LTTK của sinh viên là như nhau bị bác bỏ ở mức ý nghĩa 5%. Có nghĩa là nếu ở độ 95% thì có thể kết luận rằng sinh viên sau khi học môn Lý thuyết thống kê thì thích chọn ngành học thống kê hơn.

10.2.2. Trường hợp mẫu lớn (n >20):

Nếu n lớn thì chúng ta có thể dùng phân phối chuẩn thay cho phân phối Wilcoxon (T), lúc này trung bình và phương sai của phân phối Wilcoxon được tính như sau:

$$\mu_T = \frac{n(n+1)}{4}$$

và

$$\sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$

Đại lượng kiểm định: $Z = \dfrac{T - \mu_T}{\sigma_T}$

Trong trường hợp này ta có thể quyết định bác bỏ giả thuyết H_0 ở mức ý nghĩa α khi:

$Z < -z_\alpha$ Nếu kiểm định dạng 1 bên.

$Z < -z_{\alpha/2}$ Nếu kiểm định dạng 2 bên.

Ví dụ : Trở lại ví dụ ở trường hợp 1, thay vì khảo sát thăm dò 10 sinh viên, chúng ta thăm dò 80 sinh viên. Trong 80 sinh viên trả lời, có 20 sinh viên mức độ thích chọn ngành học thống kê không thay đổi, còn lại 60 sinh viên có chênh lệnh. Trong các mức độ chênh lệch tìm được thì giá trị nhỏ nhất

của T (minimum) là 625. Hãy kiểm định giả thuyết H_0 cho rằng mức độ thích chọn ngành thống kê sau khi học bằng hay thấp hơn hơn trước khi học môn học lý thuyết thống kê (kiểm định 1 bên).

Ta có n = 60, T = 625 và giả thuyết H_0 đúng thì phân phối Wilcoxon có trung bình và phương sai như sau:

$$\mu_T = \frac{n(n+1)}{4} = \frac{60(61)}{4} = 915$$

và $$\sigma_T^2 = \frac{n(n+1)(2n+1)}{24} = \frac{60(61)(121)}{24} = 18.452,5$$

=> $$\sigma_T = \sqrt{18.452,5} = 135,84$$

Vậy $$z = \frac{625 - 915}{135,84} = -2,13$$

Nếu sử dụng độ tin cậy 95% ($\alpha = 0,05$), tra bảng z, ta có $z_{0,05} = 1,645$. Do z < -1,645 nên giả thuyết H_0 bị bác bỏ.

Trong trường hợp chưa có mức ý nghĩa α để so sánh, ta phải tìm giá trị p để kết luận thay cho mức ý nghĩa α.

Theo giá trị kiểm định ta có:

$z_\alpha = -2,13$

Suy ra: $\alpha = 0,5 - \varphi(2,13) = 0,5 - 0,4834 = 0,0166$

Hay $\alpha = 1,66\%$ (đây chính là giá trị p). Vậy, giả thuyết H_0 có thể bị bác bỏ ở bất kỳ mức ý nghĩa nào lớn hơn 1,66%. Với những dữ liệu đã thu thập, chúng ta bằng chứng khá rõ ràng để kết luận rằng giả thuyết H_0 gần như luôn luôn sai và hoàn toàn có căn cứ để kết luận mức độ ưa thích đã tăng lên.

10.3. Kiểm định Mann-Whitney (kiểm định U)

Kiểm định Mann-Whitney, còn được gọi là kiểm định Wilcoxon, không yêu cầu các giả định về hình dạng của phân phối đang xem xét. Nó được dùng để kiểm định giả thuyết trên *hai mẫu độc lập* có xuất phát từ hai tổng thể có phân phối giống nhau không. Hình dạng của phân phối không cần phải xác định. Kiểm định này không đòi hỏi dữ liệu nghiên cứu phải là dữ liệu khoảng cách, mà chỉ cần dữ liệu xếp hạng là đủ.

10.3.1 Trường hợp mẫu nhỏ ($n \le 10$ và $n_1 \le n_2$)

Giả sử ta có hai mẫu ngẫu nhiên độc lập gồm n_1 và n_2 quan sát từ tổng thể

thứ nhất và tổng thể thứ hai (n_1 và n_2 có thể khác nhau). Ta có:

- Giả thuyết H_0 : Phân phối của hai tổng thể là giống nhau, hay
  H_0 : $\mu_1 = \mu_2$

- Đại lượng kiểm định :

$$U = n_1.n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

Trong đó:

n_1: là số quan sát mẫu từ tổng thể thứ 1.

n_2: là số quan sát mẫu từ tổng thể thứ 2.

R_1: là tổng các hạng của quan sát mẫu từ tổng thể thứ 1.

(chú ý rằng chúng ta không cần cộng hạng ở cả hai mẫu vì từ tổng cộng hạng của một mẫu chúng ta có thể suy ra tổng cộng hạng của mẫu kia. Ví dụ như nếu tổng số quan sát của cả hai mẫu là 20, thì tổng cộng hạng của cả hai mẫu là tổng các con số từ 1 đến 20, tức là $\frac{20(21)}{2} = 210$. Do đó kiểm định chỉ cần dựa trên tổng cộng hạng của một mẫu)

Tiếp theo, tra bảng phân phối U để tìm: $F(U) = F_{n_1,n_2}(U)$

Và quyết định bác bỏ giả thuyết H_0 khi : $\alpha > 2\,F(U)$.

Chú ý: 2 F(U) là giá trị p của kiểm định, vì vậy giả thuyết H_0 có thể bị bác bỏ ở mức ý nghĩa α lớn hơn giá trị p.

Ví dụ: Giả sử chúng ta nghiên cứu xem yêu cầu về lương khởi điểm của sinh viên sắp tốt nghiệp nhà ở TP Hồ Chí Minh có khác biệt gì với của sinh viên sắp tốt nghiệp nhà ở các tỉnh/TP khác không. Chọn ngẫu nhiên 12 sinh viên ngành quản trị kinh doanh trong đó có 5 sinh viên nhà ở TPHCM để hỏi ý kiến, chúng ta có kết quả được trình bày trong Bảng 10.4.

Bảng 10.4: Xếp hạng từ nhỏ đến lớn yêu cầu về lương khởi điểm (ngàn đồng) của sinh viên năm cuối ngành quản trị kinh doanh.

Sinh viên	Nhà ở TPHCM	Xếp hạng	Sinh viên	Nhà ở các tỉnh/TP khác	Xếp hạng
1	2.000	12	6	1.200	3.5
2	1.700	9	7	1.900	11
3	1.600	8	8	1.150	2
4	1.800	10	9	1.300	5
5	1.400	6	10	1.200	3.5
			11	1.500	7
			12	1.100	1

Tổng		45			33
Trung bình		9			4,71

Chú ý: Hạng của các mức độ trùng nhau của hai nhóm sinh viên cũng được xếp bằng nhau và bằng trung bình cộng của giá trị 2 hạng liên tiếp đó.

Giả thuyết H_0: trung bình lương khởi điểm yêu cầu của sinh viên nhà ở TPHCM và nhà ở các tỉnh/TP khác là bằng nhau.

Ta có: $n_1 = 5$ $n_2 = 7$ $R_1 = 45$

Suy ra: $U = n_1.n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (5\times 7) + \frac{5(6)}{2} - 45 = 5$

Tra bảng phân phối U ta có

$F_{n_1,n_2}(U) = F_{5,7}(5) = 0{,}024$

Tìm giá trị p: $p = 2F(U) = 2(0{,}024) = 0{,}048$ hay 4,8%

Vì vậy, giả thuyết H_0 có thể bị bác bỏ ở bất kỳ giá trị nào của α lớn hơn 4,8%. Với độ tin cậy 95% (α = 0,05 = 5%), chúng ta có đủ bằng chứng thống kê để bác bỏ giả thuyết H_0 cho rằng yêu cầu về lương khởi điểm của sinh viên quản trị kinh doanh nhà ở TPHCM bằng với sinh viên nhà ở các tỉnh/TP khác.

10.3.2 Trường hợp mẫu lớn ($n_1, n_2 > 10$)

Khi tăng số quan sát lên, phân phối U sẽ tiệm cận phân phối chuẩn với trung bình và phương sai được tính như sau:

$$\mu_U = \frac{n_1 n_2}{2}$$

$$\sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Đại lượng kiểm định được tính như sau:

$$Z = \frac{U - \mu_U}{\sigma_U}$$

Tiêu chuẩn quyết định:

Trường hợp kiểm định 2 bên: $H_0: \mu_1 = \mu_2$

- Chấp nhận Ho khi $-z_{\frac{\alpha}{2}} \le Z \le +z_{\frac{\alpha}{2}}$
- Bác bỏ Ho khi $Z < -z_{\frac{\alpha}{2}}$ hay $Z > +z_{\frac{\alpha}{2}}$

Trường hợp kiểm định 1 bên: $H_0: \mu_1 \ge \mu_2$, bác bỏ Ho khi $Z < -z_\alpha$

Trường hợp kiểm định 1 bên: $H_0: \mu_1 \le \mu_2$, bác bỏ Ho khi $Z > z_\alpha$

Nếu chưa quyết định sử dụng mức ý nghĩa α nào, chúng ta có thể tính giá trị p để kết luận.

Ví dụ: Trở lại ví dụ yêu cầu tiền lương khởi điểm của sinh viên năm cuối. chọn ngẫu nhiên 100 sinh viên năm cuối ở nhiều ngành, trong đó có 40 sinh viên nhà ở TPHCM. Yêu cầu về tiền lương được xếp hạng từ nhỏ đến lớn, và tổng cộng hạng được xếp cho sinh viên nhà ở TPHCM là R_1=2.328

Đặt giả thuyết:

H_0: Yêu cầu lương khởi điểm trung bình của hai nhóm sinh viên bằng nhau, hay

H_0: $\mu_1 = \mu_2$

Đại lượng thống kê Mann-Whitney được tính như sau:

$$u = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 40 \times 60 + \frac{40 \times 41}{2} - 2.328 = 892$$

Với trung bình $\mu_U = \frac{n_1 . n_2}{2} = \frac{40 \times 60}{2} = 1.200$

Và phương sai $\sigma^2 = \frac{n_1 . n_2 (n_1 + n_2 + 1)}{12} = \frac{40 \times 60 \times 101}{12} = 20.200$

Giá trị kiểm định $z = \frac{u - \mu_U}{\sigma_U} = \frac{892 - 1.200}{\sqrt{20.200}} = -2,167$

Với độ tin cậy 95%, ta có $z_{0.05/2} = 1,96$. Do z = -2,167 < -1,96 nên ta bác bỏ giả thuyết Ho.

Nếu kiểm định 1 bên: $H_0: \mu_1 \ge \mu_2$

Với độ tin cậy 95%, ta có: $z_{0.05} = 1,645$. Do z = -2,167 < -1,645 nên Ho được chấp nhận. Chúng ta kết luận rằng sinh viên có nhà ở TP HCM yêu cầu lương khởi điểm cao hơn sinh viên nhà ở các tỉnh/TP khác.

Chúng ta cũng có thể tính ra giá trị p như sau:

$z_{\alpha/2} = -2,169$, tra bảng phân phối chuẩn suy ra:

$\alpha/2 = 0,5 - 0,485 = 0,015$,$\alpha = 0,015 \times 2 = 0,03$ hay 3 %

Dựa vào giá trị p này, chúng ta có thể nói rằng giả thuyết H_0 yêu cầu về tiền lương khởi điểm của sinh viên nhà ở TPHCM và sinh viên nhà ở các

Tỉnh/TP khác bằng nhau có thể bị bác bỏ ở bất kỳ mức ý nghĩa α nào lớn hơn 3%.

10.4 Kiểm định Kruskal-Wallis

Kiểm định Mann-Whitney được sử dụng để xem xét sự khác biệt về phân phối giữa hai tổng thể từ các dữ liệu của hai mẫu độc lập. Để kiểm định sự khác biệt về phân phối giữa ba (hay nhiều hơn ba) tổng thể từ các dữ liệu mẫu của chúng, chúng ta sử dụng kiểm định Mann-Whitney mở rộng, được gọi là phân tích phương sai một yếu tố Kruskal-Wallis.

Thủ tục tính toán kiểm định Kruskal-Wallis tương tự như thủ tục kiểm định Mann-Whitney. Tất cả các quan sát của ba nhóm được gộp lại với nhau để xếp hạng. Sau đó hạng của các quan sát trong từng nhóm được cộng lại, và đại lượng thống kê Kruskal-Wallis W được tính từ các tổng này. Đại lượng W này xấp xỉ một phân phối Khi-bình phương (Chi-Square) với giả thiết là cả ba nhóm có phân phối giống nhau. Để hiểu rõ về tính toán trong kiểm định phi tham số Kruskal-Wallis, xin xem chương phân tích phương sai, phần các tổng thể có phân phối bất kỳ.

10.5 Kiểm định Chi bình phương - χ^2

Trong phần này chúng ta sẽ xem xét hai kiểm định dùng phân phối Chi bình phương. Kiểm định này phù hợp khi chúng ta có dữ liệu định tính từ thang đo danh nghĩa dưới dạng tần số.

10.5.1 Kiểm định sự phù hợp [1]

Kiểm định sự phù hợp là kiểm định xem dữ liệu thu thập được phù hợp đến mức độ nào với giả định về phân phối của tổng thể. Ở đây ta dùng phân phối "Chi" bình phương $\left(\chi^2\right)$ để so sánh trong quá trình kiểm định. Một kiểm định χ^2 thường bao gồm các bước sau đây:

1. Thiết lập giả thuyết H_0 và H_1 về tổng thể.
2. Tính toán các tần số lý thuyết theo giả thuyết H_0.
3. Tính toán các chênh lệch giữa tần số lý thuyết – hay còn gọi là tần số kỳ vọng (E) và tần số quan sát thực tế (O). Từ đó, xác định giá trị kiểm định Chi bình phương theo công thức: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

[1] Goodness-of-fit test

O_i: Tần số quan sát của nhóm thứ i.

E_i: Tần số lý thuyết của nhóm thứ i (tính theo giả thuyết H_0)

4. So sánh giá trị kiểm định tính được với giá trị trong bảng phân phối χ^2 và kết luận.

Kiểm định sự phù hợp trong trường hợp đã biết các tham số của tổng thể.

Giả sử có một mẫu ngẫu nhiên với n quan sát, mỗi quan sát có thể được phân vào một trong k nhóm.

- Gọi $O_1, O_2, \ldots, O_k$ là số lượng quan sát ở nhóm thứ 1,2, . . ., k.
- Gọi $p_1, p_2, \ldots, p_k$ là xác suất giả thuyết để một quan sát rơi vào nhóm thứ 1, 2, . . ., k (giả thuyết H_0). Các xác suất này có thể bằng nhau hay không bằng nhau. Số lượng quan sát lý thuyết ở nhóm thứ i, theo giả thuyết H_0, là:

$$E_i = n.p_i \qquad (i=1, 2, ..., k)$$

- Kiểm định chỉ có ý nghĩa khi $E_i \geq 5$

Với mức ý nghĩa α, kiểm định giả thuyết H_0 được thực hiện như sau:

Bác bỏ H_0 nếu: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{k-1,\alpha}$

Với $\chi^2_{k-1,\alpha}$ là giá trị tra ở bảng phân phối χ^2 với (k-1) bậc tự do.

Ví dụ: một công ty sản xuất xe gắn máy dự định đưa ra thị trường một kiểu dáng xe mới với 4 màu sắc khác nhau. Phòng tiếp thị muốn tìm hiểu thị hiếu khách hàng về màu sắc xe để xem thử các khách hàng mục tiêu có thích một màu nào nhiều hơn hay thích cả 4 màu là như nhau. Thông tin này khá quan trọng đối với kế hoạch sản xuất sắp tới của công ty. Bộ phận nghiên cứu của phòng tiếp thị đã chọn ngẫu nhiên 120 khách hàng tiềm năng để thu thập ý kiến. Mỗi khách hàng được xem hình ảnh xe mẫu với các màu sắc khác nhau và cho biết ý kiến họ thích màu xe nào nhất. Kết quả như sau:

Màu xe	Trắng	Đen	Vàng	Xanh	Tổng cộng
Số khách hàng chọn	18	60	12	30	120

Giả thuyết H_0: Sự ưa thích của khách hàng về 4 màu xe là bằng nhau, nghĩa là xác suất (tỷ lệ) khách hàng chọn một trong 4 màu là bằng nhau:

$$p_1 = p_2 = p_3 = p_4 = 0,25$$

Giả thuyết H_1: Sự ưa thích đối với 4 màu xe là khác nhau, nghĩa là xác suất khách hàng chọn lựa đối với 4 màu là không bằng nhau.

Theo giả thuyết H_0, số lượng khách hàng (tần số lý thuyết) chọn màu thứ i là $E_i = n.p_i$. Do đó, ta có $E_1 = E_2 = E_3 = E_4 = (120)(0,25) = 30$

Giá trị kiểm định χ^2 được tính toán như sau:

$$\chi^2 = \sum_{i=1}^{k} \frac{(0_i - E_i)^2}{E_i} = \frac{(18-30)^2}{30} + \frac{(60-30)^2}{30} + \frac{(12-30)^2}{30} + \frac{(30-30)^2}{30} = 45,6$$

Tra bảng phân phối χ^2, ta có : $\chi^2_{k-1,\alpha} = \chi^2_{4-1,1\%} = 11,34$.

Vì giá trị kiểm định $\chi^2 > \chi^2_{k-1,\alpha}$, ta kết luận rằng ở mức ý nghĩa 1% (độ tin cậy 99%) giả thuyết H_0 bị bác bỏ, nghĩa là sự ưa thích đối với 4 màu xe là khác nhau. Quan sát các tần số thực tế, chúng ta có thể thấy màu đen được ưa thích hơn hẳn màu trắng và màu vàng.

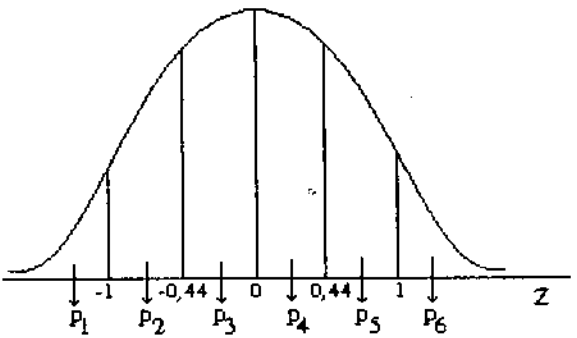
Kiểm định sự phù hợp trong trường hợp chưa biết các tham số tổng thể. Trong phần trên, chúng ta đã thực hiện kiểm định giả thuyết về việc các quan sát thực tế có phân phối với các xác suất đã xác định nào đó. Trong trường hợp chưa biết các tham số tổng thể, để kiểm định giả thuyết cho rằng các đơn vị tổng thể tuân theo một phân phối nào đó, ví dụ như phân phối nhị thức, phân phối Poisson, hay phân phối chuẩn, chúng ta có thể dùng các dữ liệu thu thập được để ước lượng tham số tổng thể.

Trước hết, phải xác định xác suất để một quan sát "rơi" vào nhóm thứ i theo như luật phân phối muốn kiểm định, nghĩa là xác định các p_i. Sau đó, tính các tần số quan sát lý thuyết theo luật phân phối này E_i, tính giá trị kiểm định χ^2 và áp dụng qui tắc kiểm định giống như đã nói ở phần trên. Cần chú ý rằng trong trường hợp này, số bậc tự do giảm đi 1 cho mỗi tham số tổng thể được ước lượng.

Ví dụ: một nhóm nghiên cứu muốn kiểm định xem phân phối của số tiền chi mua sắm của khách du lịch quốc tế trong một ngày lưu trú tại Việt Nam có tuân theo phân phối chuẩn hay không. Một mẫu ngẫu nhiên 150 khách du lịch thuộc nhiều quốc tịch khác nhau được chọn để thu thập dữ liệu. Kết quả cho thấy số tiền chi trung bình 1 ngày của 1 du khách là $\bar{x} = 50$ USD và độ lệch chuẩn s = 20 USD.

- Giả thuyết H_0 : Số tiền chi hàng ngày của du khách có phân phối chuẩn.
- Giả thuyết H_1 : Số tiền chi hàng ngày không có phân phối chuẩn.

Trước tiên, chúng ta xác định các xác suất để một đại lượng phân phối chuẩn có trị số rơi vào các khoảng nhất định. Từ bảng phân phối chuẩn, ta xác định được các xác suất của đại lượng phân phối chuẩn Z. Chẳng hạn, tra bảng phân phối chuẩn ta có xác suất để đại lượng phân phối chuẩn Z rơi vào khoảng từ 0 đến 1 là 0,3413 và gần phân nửa của xác suất này là 0,1700 ứng với trị số giới hạn z = 0,44. Vậy xác suất Z có trị số rơi vào khoảng từ 0,44 đến 1 bằng 0,1713; và xác xuất Z rơi vào khoảng từ 1→ ∞ sẽ bằng 0,1587 (hay 0,5 - 0,3413).



Tương tự chúng ta xác định được các trị số giới hạn của biến Z và các xác xuất để Z nhận các trị số nằm giữa các trị số giới hạn này đối xứng qua 0:

z = -1 z = -0,44 z = 0 z = 0,44 z = 1

$p_1 = 0,1587$, $p_2 = 0,1713$, $p_3 = 0,17$, $p_4 = 0,17$, $p_5 = 0,1713$, $p_6 = 0,1587$

Từ công thức $E_i = n.p_i$, các trị số lý thuyết E_i có kết quả tính toán như sau:

$E_1 = 23,8$, $E_2 = 25,7$, $E_3 = 25,5$, $E_4 = 25,5$, $E_5 = 25,7$, $E_6 = 23,8$

Dựa vào công thức $X = \mu + \sigma Z$ (suy ra từ công thức chuẩn hóa các trị số quan sát), chuyển các trị số giới hạn của đại lượng ngẫu nhiên có phân phối chuẩn Z thành trị số của yếu tố đang nghiên cứu là số tiền chi mua sắm 1 ngày của khách du lịch. Chúng ta có thể dùng $\bar{x}$ và s *(tham số mẫu)* thay cho μ và σ *(tham số tổng thể)* . Do đó, trị số giới hạn của các nhóm được xác định như sau:

$x_1 = 50 + (-1)(20) = 30,0$
$x_2 = 50 + (-0,44)(20) = 41,2$
$x_3 = 50 + (0)(20) = 50,0$
$x_4 = 50 + (0,44)(20) = 58,8$
$x_5 = 50 + (1)(20) = 70,0$

Từ số liệu thu thập được, chúng ta xác định được số lượng các quan sát rơi vào từng nhóm, nghĩa là xác định các tần số thực tế 0_i. Như vậy, ta đã xác định được các nhóm, xác xuất để một quan sát rơi vào nhóm thứ i (p_i), số lượng quan sát thực tế (0_i) và số lượng quan sát theo lý thuyết (E_i).

Tiếp theo tính giá trị kiểm định χ^2 theo công thức $\chi^2 = \sum_{i=1}^{k} \frac{(0_i - E_i)^2}{E_i}$. Các số liệu được trình bày trong Bảng 10.5 như sau :

Bảng 10.5: Bảng tính toán đại lượng kiểm định χ^2

X_i (USD)	P_i	Tần số lý thuyết $E_i = (n.p_i)$	Tần số thực tế 0_i	$(0_i - E_i)^2/E_i$
Dưới 30,0	0,1587	23,8	21	0,329
30,0 – 41,2	0,1713	25,7	27	0,066
41,2 – 50,0	0,1700	25,5	26	0,009
50,0 - 58,8	0,1700	25,5	30	0,794
58,8 – 70,0	0,1713	25,7	26	0,004
≥ 70,0	0,1587	23,8	20	0,607
Tổng cộng	1,0000	150,0	150	1,809

Từ bảng trên ta có $\chi^2 = 1,809$ và trong 6 nhóm có hai tham số được ước lượng ($\bar{x}$ cho μ và s cho σ) nên số bậc tự do là (k -1 - số tham số) = 6 -1 -2 = 3.

Tra bảng phân phối χ^2, ta có: $\chi^2_{3,0,05} = 7,815 > 1,809$. Do vậy ta chấp nhận giả thuyết H_0 ở mức ý nghĩa 5%, tức là chưa có bằng chứng để nói rằng tổng thể không có phân phối chuẩn. Như vậy số tiền chi mua sắm 1 ngày của một khách du lịch nước ngoài được coi như tuân theo phân phối chuẩn.

10.5.2 Kiểm định tính độc lập

Kiểm định Chi bình phương còn được vận dụng để nghiên cứu xem có tồn tại mối liên hệ giữa hai yếu tố (hay hai biến) hay không. Ví dụ, xem xét mối liên hệ giữa thời gian đi làm thêm và kết quả học tập của sinh viên, thời gian cha mẹ dành cho con và kết quả học tập của học sinh, độ tuổi và hành vi tiêu dùng, quy mô doanh nghiệp và thu nhập của nhân viên, giới tính và phương thức tiết kiệm ...

Giả sử chúng ta có một mẫu ngẫu nhiên gồm n quan sát, được phân tổ kết

hợp theo hai tiêu thức, tạo nên nên bảng phân tổ kết hợp gồm r dòng và c cột. Gọi O_{ij} là quan sát ứng với dòng thứ i và cột thứ j, R_i là tổng quan sát ở dòng thứ i, C_j là tổng số quan sát ở cột thứ j, n là tổng quan sát của r dòng đồng thời cũng là tổng số quan sát của c cột.

Bảng 10.6: Dạng tổng quát của một bảng phân tổ kết hợp hai tiêu thức

Phân tổ theo tiêu thức thứ 2	Phân tổ theo tiêu thức thứ 1					
	1	2	3	...	c	$\sum$
1	O_{11}	O_{12}	O_{13}	....	O_{1c}	R_1
2	O_{21}	O_{22}	O_{23}			R_2
3	O_{31}	O_{32}	O_{33}			R_3
...	...	...	...			...
r	O_{r1}	O_{r2}	O_{r3}	...	O_{rc}	R_r
Tổng cộng	C_1	C_2	C_3	...	C_c	n

Để kiểm định xem có mối liên hệ giữa hai yếu tố hay không, trước hết ta lập giả thuyết H_0 và H_1:

- Giả thuyết H_0: Không có mối liên hệ giữa hai yếu tố.
- Giả thuyết H_1: Có tồn tại mối liên hệ giữa hai yếu tố.

Để kiểm định giả thuyết H_0 trước hết chúng ta cần tính các tần số lý thuyết E_{ij} tại từng ô trong bảng phân tổ kết hợp trên cơ sở giả thiết rằng giả thuyết H_0 đúng. Các tần số lý thuyết này được tính theo công thức tổng quát như sau:

$$E_{ij} = \frac{R_i \times C_j}{n}$$

Sau đó chúng ta sẽ tính chênh lệch giữa tần số thực tế với tần số lý thuyết trong từng ô, bình phương lên rồi chia cho tần số lý thuyết, và cộng tất cả các kết quả tính toán ở từng ô này lại với nhau. Thể hiện bằng công thức chúng ta có:

Đại lượng kiểm định : $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Nguyên tắc quyết định là:

Bác bỏ H_0 nếu $\chi^2 > \chi^2_{(r-1)(c-1),\alpha}$

với $\chi^2_{(r-1)(c-1),\alpha}$ là trị số tra từ bảng phân phối Chi bình phương với số bậc tự do là số dòng trừ 1 nhân với số cột trừ 1 và mức ý nghĩa α.

Ví dụ: Để nghiên cứu mối liên hệ giữa thời gian tự học và kết quả học tập của sinh viên, người ta chọn ngẫu nhiên 200 sinh viên để thu thập dữ liệu. Kết quả phân nhóm theo hai tiêu thức kết quả học tập và thời gian tự học của sinh viên được trình bày trong bảng sau:

Bảng 10.7: Bảng phân tổ kết hợp kết quả học tập và thời gian tự học

Kết quả học tập	Thời gian tự học (giờ/ ngày)			
	Ít (dưới 2g/ngày)	Trung bình (2 – 4 giờ/ngày)	Nhiều (trên 4 giờ/ngày)	Tổng cộng (R_i)
Trung bình	70	25	5	100
TB khá	15	35	10	60
Khá, giỏi	5	10	25	40
Tổng cộng (C_i)	90	70	40	200

Giả thuyết H_0 : Không có mối liên hệ giữa thời gian tự học và kết quả học tập của sinh viên.
Giả thuyết H_1 : Có mối liên hệ giữa thời gian tự học và kết quả học tập của sinh viên

Từ cột tổng cộng, chúng ta có thể dễ dàng tính được tỷ lệ sinh viên đạt kết quả học tập trung bình (TB) là 50% (100/200), TB khá là 30% và khá giỏi là 20%. Nếu giả thuyết H_0 đúng, thời gian tự học không có liên hệ với kết quả học tập, vậy thì tỷ lệ đạt kết quả TB đối với các nhóm sinh viên có thời gian tự học khác nhau cũng phải bằng nhau, tức là bằng 50%. Tương tự, tỷ lệ đạt loại TB khá và loại khá giỏi lần lượt là 30% và 20% trong từng nhóm sinh viên có thời gian tự học khác nhau.

Trên cơ sở giả thiết này, chúng ta có thể tính được tần số lý thuyết về số sinh viên đạt loại TB trong nhóm những sinh viên có thời gian tự học ít (E_{11}) là bằng số lượng sinh viên có thời gian tự học ít (90), nhân với tỷ lệ sinh viên đoạt loại kết quả học tập TB là 50%, tức là 90 x 50%. Mà 50% chính bằng số lượng sinh viên đạt loại TB chia cho tổng số sinh viên được khảo sát (50% = 100/200). Do đó chúng ta có thể viết lại là:

$$E_{11} = 90 \times 50\% = 90 \times \frac{100}{200}$$

Như vậy chúng ta có thể thấy tần số lý thuyết (với điều kiện giả thuyết H_0 đúng) của ô đầu tiên chính bằng tích của tổng của dòng 1 nhân với tổng của cột 1 rồi chia cho tổng số quan sát như đã trình bày ở công thức tổng quát trong trang trước. Tương tự như vậy chúng ta có thể tính nhanh các tần số lý

thuyết ở các ô khác theo công thức tổng quát này. Bảng 10.8 trình bày các kết quả tính toán tần số lý thuyết.

Bảng 10.8: Bảng tính toán các tần số lý thuyết trên cơ sở giả thuyết H_0 đúng

Kết quả học tập	Thời gian tự học (giờ/ ngày)			
	Ít (dưới 2g/ngày)	Trung bình (2 – 4 giờ/ngày)	Nhiều (trên 4 giờ/ngày)	Tổng cộng (R_i)
Trung bình	45	35	20	100
TB khá	27	21	12	60
Khá, giỏi	18	14	8	40
Tổng cộng (C_i)	90	70	40	200

Từ các tần số thực tế ở Bảng 10.7 và các tần số lý thuyết ở Bảng 10.8, chúng ta tính toán đại lượng kiểm định χ^2 như sau:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$= (70-45)^2/45 + (25-35)^2/35 + ... + (10-14)^2/14 + (25-8)^2/8 = 89,65.$

Với r = 3, c = 3, số bậc tự do là: (r-1)(c-1) = (3-1)(3-1) = 4

Tra bảng phân phối χ^2, ta có $\chi^2_{4,0,05} = 9;4877 < 89,65$

Do vậy, ở mức ý nghĩa 5%, giả thuyết H_0 bị bác bỏ, nghĩa là có tồn tại mối liên hệ giữa thời gian tự học và kết quả học tập của sinh viên.

Để thấy rõ hơn mối liên hệ này, chúng ta có thể tính ra các số phần trăm như trong Bảng 10.9

Bảng 10.9: Bảng tính toán các tỷ lệ % kết quả học tập của từng nhóm sinh viên có thời gian tự học khác nhau

Kết quả học tập	Thời gian tự học (giờ/ ngày)			
	Ít (dưới 2g/ngày)	Trung bình (2 – 4 giờ/ngày)	Nhiều (trên 4 giờ/ngày)	Tổng cộng (R_i)
Trung bình	78%	36%	13%	50%
TB khá	17%	50%	25%	30%
Khá, giỏi	6%	14%	63%	20%
Tổng cộng (C_i)	100%	100%	100%	100%

Từ Bảng 10.9, chúng ta có thể quan sát thấy rằng, đa số (gần 80%) các sinh viên có thời gian tự học ít chỉ đạt kết quả học tập loại trung bình, rất ít sinh

viên có thời gian tự học ít mà đạt được kết quả khá giỏi. Trong khi đó, gần 2/3 số sinh viên có thời gian tự học nhiều lại đạt kết quả khá giỏi. Có thể nói thời gian tự học càng nhiều thì kết quả học tập càng cao.

CHƯƠNG 11

TƯƠNG QUAN VÀ HỒI QUI

Mặc dù có nhiều tình huống trong thực tế chỉ liên quan đến một biến, song lại có rất nhiều trường hợp khác những nhà quyết định lại cần quan tâm, xem xét đến mối quan hệ giữa hai hay nhiều biến. Chẳng hạn như, người kỹ sư nông nghiệp nghiên cứu mối quan hệ giữa giống, nước, phân bón, kỹ thuật chăm sóc, ... với năng suất thu hoạch của từng loại nông sản. Nhà môi giới đầu tư thì chú ý rất nhiều đến mối quan hệ giữa giá cả thị trường chứng khoán và cổ tức của nó. Một giám đốc tiếp thị thì quan tâm đến mối quan hệ giữa doanh thu bán sản phẩm với chi phí quảng cáo của công ty. Người làm công tác định giá bất động sản như nhà cửa thì quan tâm thu thập giá bán gần đây của các nhà cửa tương tự, kích thước và tình trạng của ngôi nhà. Tương quan và hồi qui được giới thiệu ở chương này là những kỹ thuật thống kê mà người kỹ sư, nhà môi giới đầu tư, vị giám đốc tiếp thị, những người làm công tác định giá, ... sẽ cần trong phân tích, đánh giá của họ. Những kỹ thuật này lại cũng rất quan trọng đối với những người ra quyết định trong việc xác định mối liên hệ giữa các biến.

11.1 TƯƠNG QUAN

Phương pháp tương quan được dùng để nghiên cứu mối quan hệ giữa hai hay nhiều biến ngẫu nhiên. Mục tiêu của phương pháp tương quan tuyến tính là đo lường cường độ của mối quan hệ giữa hai biến X và Y, hai biến này được xem là 2 biến ngẫu nhiên " ngang nhau" – không phân biệt biến độc lập hay biến phụ thuộc.

11.1.1 Hệ số tương quan [1]

Giả sử có 2 biến ngẫu nhiên X và Y có phân phối chuẩn, với trung bình μ_x, μ_y và phương sai σ_x^2, σ_y^2. Hệ số tương quan ρ là khái niệm được dùng để thể hiện cường độ và chiều hướng của mối liên hệ tuyến tính giữa X và Y. ρ được gọi là hệ số tương quan của tổng thể.

Giá trị của ρ nằm trong đoạn [-1, +1]

- $\rho < 0$: giữa X và Y có mối liên hệ nghịch, tức là khi biến X tăng lên (giảm đi) thì biến Y giảm đi (tăng lên), hoặc ngược lại biến Y tăng lên (giảm đi) thì biến X giảm đi (tăng lên).

[1] Correlation coefficient

- $\rho > 0$: giữa X và Y có mối liên hệ thuận, tức là khi biến X tăng lên (giảm đi) thì biến Y cũng tăng lên (giảm đi), hoặc ngược lại biến Y giảm đi (tăng lên) thì biến X cũng giảm đi (tăng lên).
- $\rho = 0$: giữa X và Y không có mối liên hệ tuyến tính

Trị tuyệt đối của ρ càng lớn, thì mối liên hệ tuyến tính giữa X và Y càng chặt chẽ.

Trong thực tế, ta không biết được giá trị của ρ, vì vậy phải ước lượng nó từ dữ liệu mẫu thu thập được.

Gọi $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ là mẫu n cặp giá trị quan sát thu thập ngẫu nhiên từ X và Y. hệ số tương quan tổng thể ρ được ước lượng từ hệ số tương quan mẫu r. Công thức tính hệ số tương quan mẫu r (còn gọi là hệ số tương quan Pearson)[1]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (11.1)$$

hoặc

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y})}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}} \qquad (11.2)$$

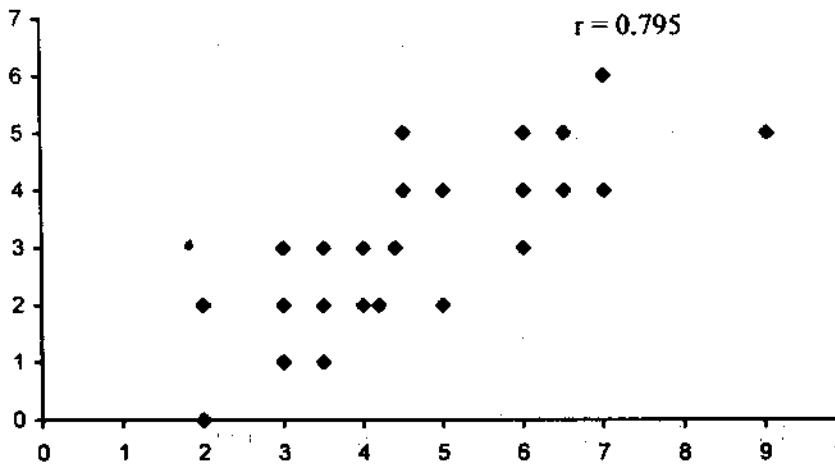
Ví dụ 1: Khảo sát ngẫu nhiên 30 khách hàng nữ đi siêu thị, ta thu được các dữ liệu sau:

- **q1 : số lần đi siêu thị trong tháng qua**
- **q2 : thu nhập hộ trung bình/tháng (triệu đồng)**
- **q3 : tuổi của người trả lời**
- **q4 : số món hàng mua ngoài dự định trong tháng qua**

STT	q1	q2	q3	q4	STT	q1	q2	q3	q4
1	1	3,0	54	2	*16*	2	3,0	32	3
2	1	3,5	48	2	*17*	2	3,0	44	2
3	4	5,0	35	4	*18*	4	6,0	33	4
4	3	4,0	29	3	*19*	2	3,5	29	3

[1] Pearson correlation

5	4	6,5	37	5	*20*	2	4,0	29	2
6	1	3,0	38	1	*21*	4	4,5	38	4
7	2	6,0	44	3	*22*	2	3,5	28	2
8	2	5,0	45	2	*23*	2	4,0	26	3
9	1	2,0	25	0	*24*	3	4,5	31	5
10	2	3,5	37	1	*25*	1	6,5	41	4
11	6	9,0	28	5	*26*	4	7,0	27	4
12	4	7,0	32	6	*27*	5	6,0	40	5
13	5	6,0	36	4	*28*	2	4,2	27	2
14	4	6,0	35	5	*29*	2	4,0	50	2
15	2	2,0	45	2	*30*	3	4,4	33	3



Hình 11.1: Đồ thị phân tán[1] của số món hàng mua ngoài dự định theo thu nhập hộ trung bình/tháng (triệu đồng)

11.1.2 Kiểm định giả thuyết về mối liên hệ tương quan

Hệ số tương quan mẫu r còn được dùng để xét xem giữa hai biến X và Y có liên hệ tương quan với nhau không, bằng cách kiểm định giả thuyết H_0 cho rằng hệ số tương quan của tổng thể ρ bằng không.

Giả sử có mẫu gồm n cặp quan sát chọn ngẫu nhiên từ X, Y có phân phối chuẩn.

[1] Scatter diagram

Giả thuyết: $H_0 : \rho = 0$ (X và Y không có liên hệ)

$H_1 : \rho \neq 0$ (có liên hệ giữa X và Y)

Giá trị kiểm định: $$t = \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}} \qquad (11.3)$$

Qui tắc quyết định: đối với kiểm định cả 2 bên ở mức ý nghĩa α , bác bỏ H_0 , nếu:

$$\frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}} < -t_{n-2,\alpha/2} \quad \text{hay} \quad \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}} > t_{n-2,\alpha/2}$$

Ta tính r theo công thức (11.1) bằng cách lập bảng tính toán 11.2 (hoặc có kết quả nhanh (bảng 11.1) bằng cách sử dụng phần mềm Excel, trên Excel vào menu **Tools – Data Analysis – Correlation**)

Bảng 11. 1: Kết quả hệ số tương quan mẫu r từ Excel

	q3	*q4*
q3	1	
q4	0.7950	1

Bảng 11.2: Bảng tính toán hệ số tương quan mẫu r

STT	q2 (x)	q4 (y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	3,0	2	-1,7	-1,1	2,7335	1,21	1,8187
2	3,5	2	-1,2	-1,1	1,3302	1,21	1,2687
3	5,0	4	0,3	0,9	0,1202	0,81	0,3120
4	4,0	3	-0,7	-0,1	0,4268	0,01	0,0653
5	6,5	5	1,8	1,9	3,4102	3,61	3,5087
6	3,0	1	-1,7	-2,1	2,7335	4,41	3,4720
7	6,0	3	1,3	-0,1	1,8135	0,01	-0,1347
8	5,0	2	0,3	-1,1	0,1202	1,21	-0,3813
9	2,0	0	-2,7	-3,1	7,0402	9,61	8,2253
10	3,5	1	-1,2	-2,1	1,3302	4,41	2,4220
11	9,0	5	4,3	1,9	18,8935	3,61	8,2587

12	7,0	6	2,3	2,9	5,5068	8,41	6,8053
13	6,0	4	1,3	0,9	1,8135	0,81	1,2120
14	6,0	5	1,3	1,9	1,8135	3,61	2,5587
15	2,0	2	-2,7	-1,1	7,0402	1,21	2,9187
16	3,0	3	-1,7	-0,1	2,7335	0,01	0,1653
17	3,0	2	-1,7	-1,1	2,7335	1,21	1,8187
18	6,0	4	1,3	0,9	1,8135	0,81	1,2120
19	3,5	3	-1,2	-0,1	1,3302	0,01	0,1153
20	4,0	2	-0,7	-1,1	0,4268	1,21	0,7187
21	4, 5	4	-0,2	0,9	0,0235	0,81	-0,1380
22	3,5	2	-1,2	-1,1	1,3302	1,21	1,2687
23	4,0	3	-0,7	-0,1	0,4268	0,01	0,0653
24	4,5	5	-0,2	1,9	0,0235	3,61	-0,2913
25	6,5	4	1,8	0,9	3,4102	0,81	1,6620
26	7,0	4	2,3	0,9	5,5068	0,81	2,1120
27	6,0	5	1,3	1,9	1,8135	3,61	2,5587
28	4,2	2	-0,5	-1,1	0,2055	1,21	0,4987
29	4,0	2	-0,7	-1,1	0,4268	1,21	0,7187
30	4,4	3	-0,3	-0,1	0,0642	0,01	0,0253
cộng	**139,6**	**93**			**78,3947**	**60,7**	**54,84**

Trong đó: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ và $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$ (chính là các trung bình mẫu)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{139,6}{30} = 4,653 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{93}{30} = 31$$

với kết quả tính trên bảng 11.2

ta có: $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{54,84}{\sqrt{78,3947.60,7}} = 0,795$$

Ở ví dụ này ta có thể thực hiện kiểm định như sau:

Đặt giả thuyết: $H_0 : \quad \rho \le 0$

$H_1 : \quad \rho > 0$

Do đó giá trị kiểm định: $t = \dfrac{r}{\sqrt{\dfrac{(1-r^2)}{(n-2)}}} = \dfrac{0,795}{\sqrt{\dfrac{1-(0,795)^2}{(30-2)}}} = 9,2911$

Tra bảng phân phối t, hoặc trên Excel vào **Insert – Function**, dùng hàm **TINV** ta có $t_{28,0,05} = 1,7011$. Như vậy, ở mức ý nghĩa 5%, giả thuyết bị bác bỏ, vì vậy có thể nói rằng có mối quan hệ tương quan thuận giữa số món hàng mua ngoài dự định và thu nhập hộ trung bình/tháng (triệu đồng).

Sử dụng phần mềm SPSS (vào **Analyze –Correlation**) ta dễ dàng xác định hệ số r và giá trị p của kiểm định (bảng 11.3), và cũng rút ra kết luận tương tự: bác bỏ H_0 ở các mức ý nghĩa lớn hơn 0,000.

Correlations

		so mon hang mua ngoai du dinh	thu nhap ho TB thang (trd)
Pearson Correlation	so mon hang mua ngoai du dinh	1.000	.795
	thu nhap ho TB thang (trd)	.795	1.000
Sig. (1-tailed)	so mon hang mua ngoai du dinh	.	.000
	thu nhap ho TB thang (trd)	.000	.
N	so mon hang mua ngoai du dinh	30	30
	thu nhap ho TB thang (trd)	30	30

Bảng 11.3: Hệ số tương quan và giá trị r từ SPSS

11.1.3 Hệ số tương quan hạng

Trong trường hợp hai biến X, Y không có phân phối chuẩn, hoặc dữ liệu được thể hiện dưới hình thức xếp hạng, ta có thể đo lường mối liên hệ giữa X và Y bằng hệ số tương quan hạng Spearman, kí hiệu r_s.

Trước tiên, ta xếp hạng R_X, R_Y các giá trị quan sát x_i, y_i theo thứ tự tăng dần (từ 1 trở đi) (nếu có các giá trị quan sát bằng nhau, thì được xếp đồng hạng và hạng sẽ là hạng trung bình).

Hệ số tương quan hạng Spearman r_s, chính là hệ số tương r giữa các hạng của x_i và y_i, tức là vẫn dùng công thức (11.1) hay (11.2) để tính r_s, trong đó, thay x_i, y_i bằng các hạng của chúng.

Cần lưu ý rằng, nếu không xãy ra trường hợp các giá trị x_i hay y_i bằng không, tức là không xãy ra trường hợp đồng hạng, r_s có thể được tính bằng công thức sau đơn giản hơn:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad (11.4)$$

trong đó: n là số lượng các cặp (x_i, y_i)

$d_{R_i} = R_{X_i} - R_{Y_i}$: chênh lệch tuyệt đối giữa từng cặp thứ hạng của x_i và y_i

Hệ số tương quan hạng r_s có các đặc điểm sau:

- Hệ số tương quan hạng r_s biến thiên trong đoạn [-1, +1]
- Nếu $r_s \to 1$ và càng lớn thì mối quan hệ tương quan hạng khá chặt chẽ theo chiều hướng thuận.
- Nếu $r_s \to -1$ và càng lớn thì mối quan hệ tương quan hạng khá chặt chẽ theo chiều hướng nghịch.
- Nếu $r_s \to 0$ và càng nhỏ thì mối quan hệ tương quan hạng không chặt chẽ.
- Nếu $r_s = 0$, giữa các biến nghiên cứu không có quan hệ tương quan hạng.

Ví dụ 2: Có số liệu về số công nhân y_i (người) và giá trị sản xuất x_i (triệu đồng) qua một tháng cao điểm xuất khẩu ở 17 doanh nghiệp chuyên về thủ công mỹ nghệ cao cấp.

x_i	R_{x_i}	y_i	R_{y_i}	d_i	d_i^2
115.300	4	11.597	5	1	1
136.100	2	15.355	2	0	0
16.600	13	1.643	13	0	0
4.877	17	897	15	-2	4
105.300	6	9.344	7	1	1
33.700	11	3.325	11	0	0
6.473	16	707	16	0	0
99.300	7	11,370	6	-1	1
70.000	8	5.165	10	2	4

112.400	5	13.146	3	-2	4
48.200	10	6.053	8	-2	4
27.605	12	1.877	12	0	0
116.148	3	12.123	4	1	1
10.300	15	663	17	2	4
14.200	14	1.365	14	0	0
61.000	9	5.661	9	0	0
190.000	1	17.170	1	0	0
cộng					24

$$r_s = 1 - \frac{6.(24)}{17[(17)^2 - 1]} = 0{,}971$$

Như vậy, giá trị sản xuất và số công nhân có tương quan hạng khá chặt chẽ và theo chiều hướng thuận.

Ngoài ra, r_s cũng có thể dùng để kiểm định về mối liên hệ giữa X và Y. Giả định rằng có mối liên hệ nào đó giữa các hạng của các giá trị của hai biến X và Y, xét ở góc độ tổng thể. Mối liên hệ này được thể hiện bằng hệ số tương quan hạng của tổng thể, ρ_s. Kiểm định giả thuyết H_0 tức là xét xem có tồn tại hay không mối liên hệ giữa hai biến X và Y.

Giả thuyết: $H_0 : \rho_s = 0$ (X và Y không có liên hệ)

$H_1 : \rho_s \neq 0$ (có liên hệ giữa X và Y)

Qui tắc quyết định: đối với kiểm định cả hai bên,ở mức ý nghĩa α , bác bỏ H_0 , nếu:

$$r_s < -r_{n,\alpha/2} \text{ hoặc } r_s > r_{n,\alpha/2}$$

(tra bảng phân phối Spearman ứng với n và mức ý nghĩa $\alpha/2$)

Sử dụng ví dụ 2 để kiểm định giả thuyết sau:

Giả thuyết: $H_0 : \rho_s \leq 0$ (không có mối tương quan giữa giá trị sản xuất và số công nhân)

$H_1 : \rho_s > 0$ (có mối tương quan thuận giữa doanh thu và số công nhân)

Tra bảng phân phối của hệ số Spearman, ta có $r_{n,\alpha} = r_{17,0,05} = 0{,}412$. Vì vậy, giả thuyết H_0 bị bác bỏ ở mức ý nghĩa 5%, và ta có thể kết luận rằng có mối tương quan thuận giữa giá trị sản xuất và số công nhân).

Với SPSS (vào **Analyze -Correlate – Bivarite**), ta cũng có kết quả tương tự.

Correlations

			so cong nhan	gia tri san xuat (trieu dong)
Spearman's rho	so cong nhan	Correlation Coefficient	1.000	.971**
		Sig. (1-tailed)	.	.000
		N	17	17
	gia tri san xuat (trieu dong)	Correlation Coefficient	.971**	1.000
		Sig. (1-tailed)	.000	.
		N	17	17

**. Correlation is significant at the .01 level (1-tailed).

Trong trường hợp n lớn (n > 30), ta có thể dùng phân phối chuẩn thay cho phân phối của Spearman, với giá trị kiểm định được tính theo công thức:

$$Z = r_s\sqrt{n-1} \qquad (11.5)$$

11.2 HỒI QUI

Hồi qui được dùng để xem xét mối liên hệ tuyến tính giữa hai biến X và Y, trong đó X được xem là biến độc lập (ảnh hưởng đến biến Y) còn Y được xem là biến phụ thuộc (chịu ảnh hưởng bởi biến X). Mục tiêu của phân tích hồi qui là mô hình hóa mối liên hệ, nghĩa là từ các dữ liệu mẫu thu thập được ta cố gắng xây dựng một mô hình toán học nhằm thể hiện một cách tốt nhất mối liên hệ giữa hai biến X và Y. Phân tích hồi qui xác định sự liên quan định lượng giữa hai biến ngẫu nhiên Y và X, kết quả của phân tích hồi qui được dùng cho dự đoán, ngược lại phân tích tương quan khảo sát khuynh hướng và mức độ của sự liên quan, được dùng để đo lường tính bền vững của mối liên hệ giữa các biến đặc biệt là các biến định lượng.

11.2.1 Mô hình hồi qui tuyến tính đơn giản của tổng thể

Giả sử có hai biến X và Y, trong đó Y được xem như phụ thuộc tuyến tính vào X. Với một giá trị x_i nào đó của biến X, giá trị tương ứng y_i của biến Y được thể hiện bằng công thức:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad (11.6)$$

α, β là các hằng số, α thể hiện giá trị ước lượng của Y khi giá trị của biến X bằng không, nghĩa là giá trị của Y không phụ thuộc vào X, β là độ dốc của đường hồi qui, thể hiện mức tăng lên của Y khi X tăng một đơn vị.

ε_i là sai số ngẫu nhiên thể hiện ảnh hưởng của các yếu tố khác (không được nghiên cứu) đến Y. ε_i được xem là biến ngẫu nhiên có phân phối chuẩn với trung bình bằng không, phương sai bằng nhau và độc lập không có liên hệ với nhau.

Một cách tổng quát: mô hình hồi qui tuyến tính đơn giản của tổng thể thể hiện mối liên hệ tuyến tính giữa X và Y là: $y = \alpha + \beta x + \varepsilon$ (11.7)

Đầu tiên, ta xem xét mối liên hệ giữa q4 (số món hàng mua ngoài dự định) với q2 (thu nhập hộ trung bình/ tháng (triệu đồng)) bằng cách xem xét mối liên hệ giữa một biến phụ thuộc (q4) với một biến độc lập (q2). Đồ thị phân tán có thể gợi ý cho chúng ta loại hàm số toán học thích hợp để mô tả mối liên hệ. Từ đồ thị phân tán (hình 11.1) cho thấy số món hàng mua ngoài dự định có xu hướng tăng tuyến tính cùng với sự gia tăng của thu nhập. Như vậy ta có thể sử dụng phương trình đường thẳng để mô tả mối liên hệ này. Nếu đồ thị không gợi ý được cho chúng ta một đường thẳng, thì chúng ta phải sử dụng một phương trình khác hay phương pháp phân tích khác như chuyển các số liệu này về dạng tuyến tính (tuyến tính hóa).

Ứng với một mức thu nhập hộ trung bình/tháng x_i nào đó, số món hàng mua ngoài dự định y theo công thức (11.6) là: $y = \alpha + \beta x_i + \varepsilon_i$

Như vậy số món hàng mua ngoài dự định bao gồm 2 phần: $\alpha + \beta x_i$ thể hiện mối liên hệ tuyến tính giữa chúng.

Độ dốc β chính là lượng tăng giảm của số món hàng mua ngoài dự định điều chỉnh (hay số dự đoán, số lý thuyết theo mô hình) do lượng tăng giảm của mức thu nhập.

Hằng số α (chính là tung độ của điểm tại đó đường thẳng cắt trục tung) là số món hàng mua ngoài dự định lý thuyết khi thu nhập bằng 0.

Và giá trị ε_i thể hiện ảnh hưởng của các yếu tố khác (không được nghiên cứu) đến số món hàng mua ngoài dự định.

Trong thực tế ta không thể xác định một cách chính xác các tham số α , β của phương trình hồi qui tuyến tính của tổng thể mà chỉ có thể ước lượng chúng từ các giá trị quan sát của mẫu.

11.2.2 Phương trình hồi qui tuyến tính của mẫu

Giả sử ta có $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ là mẫu gồm n cặp quan sát thu thập ngẫu nhiên từ X và Y. Ta sẽ phải tìm các giá trị a, b để ước lượng cho các tham số α , β . Bởi vì như trên đồ thị phân tán, tất cả các điểm dữ liệu quan sát không nằm trên cùng một đường thẳng. Chúng ta có thể kẻ nhiều

đường thẳng xuyên qua các điểm dữ liệu này, vấn đề là ta chỉ chọn ra một đường thẳng mô tả sát nhất xu hướng này. Phương pháp dùng để xác định đường thẳng này là phương pháp bình phương bé nhất [1]. Phương pháp này sẽ tìm ra được một đường thẳng cực tiểu hóa được tổng các độ lệch bình phương giữa tung độ của các điểm dữ liệu quan sát và đường thẳng. Đường thẳng được xem là "thích hợp" nhất khi tổng bình phương các chênh lệch giữa giá trị thực tế y_i với giá trị $\hat{y}_i$ là nhỏ nhất, tức là:

Với $y_i = a + bx_i + e_i$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \min \qquad (11.8)$$

Với điều kiện này ta tính được:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y})}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2} \qquad (11.9)$$

$$a = \bar{y} - b\bar{x} \qquad (11.10)$$

Trong đó: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ và $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ (chính là các trung bình mẫu)

Đường hồi qui tuyến tính của mẫu có dạng $\hat{y} = a + bx$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{139,6}{30} = 4,653 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{93}{30} = 31$$

Bảng 11.4: Bảng tính tham số a, b

STT	q2 (x)	q4 (y)	x^2	(x.y)
1	3,0	2	9,00	6,0
2	3,5	2	12,25	7,0
3	5,0	4	25,00	20,0
4	4,0	3	16,00	12,0
5	6,5	5	42,25	32,5
6	3,0	1	9,00	3,0
7	6,0	3	36,00	18,0

[1] Least squares

8	5,0	2	25,00	10,0
9	2,0	0	4,00	0,0
10	3,5	1	12,25	3,5
11	9,0	5	81,00	45,0
12	7,0	6	49,00	42,0
13	6,0	4	36,00	24,0
14	6,0	5	36,00	30,0
15	2,0	2	4,00	4,0
16	3,0	3	9,00	9,0
17	3,0	2	9,00	6,0
18	6,0	4	36,00	24,0
19	3,5	3	12,25	10,5
20	4,0	2	16,00	8,0
21	4, 5	4	20,25	18,0
22	3,5	2	12,25	7,0
23	4,0	3	16,00	12,0
24	4,5	5	20,25	22,5
25	6,5	4	42,25	26,0
26	7,0	4	49,00	28,0
27	6,0	5	36,00	30,0
28	4,2	2	17,64	8,4
29	4,0	2	16,00	8,0
30	4,4	3	19,36	13,2
cộng	**139,6**	**93**	**728**	**487,6**

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n(\overline{x})(\overline{y})}{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2} = \frac{487,6 - 30(4,653)(3,1)}{728 - 30(4,653)^2} = 0,6995$$

$$a = \overline{y} - b\overline{x} = 3,1 - (0,6995)(4,653) = -0,1552$$

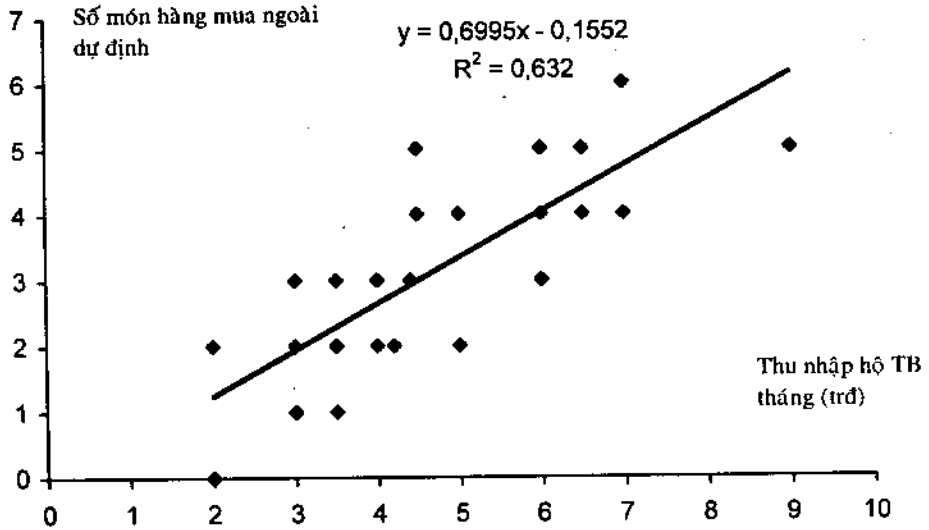
Do đó, ta có phương trình hồi qui tuyến tính thể hiện mối liên hệ giữa số món hàng mua ngoài dự định (q4) và mức thu nhập trung bình tháng (q2) là:

$\hat{y}$ = - 0,1552 + 0,6995(mức thu nhập trung bình/tháng)

Với hệ số b = 0,6995, ta có thể nói rằng khi mức thu nhập trung bình tháng tăng lên 1 triệu đồng, thì số món hàng mua ngoài dự định trung bình tăng 0,6995 món.

Nếu dùng Excel ta có thể nhanh chóng tìm ra đường hồi qui tuyến tính mẫu bằng cách: sau khi vẽ đồ thị phân tán, để con trỏ trên đồ thị phân tán nhấp chuột hai lần rồi chọn **Trendline** trên menu **Insert**, chọn loại đường tuyến

tính, tiếp tục chọn **Option** để tính và cho phép thể hiện hệ số tương quan r trên đồ thị.



Hình 11.2: Đồ thị hồi qui tuyến tính

Hoặc ta có thể dùng công cụ **Data Analysis** (chọn **Regression**) trong menu ***Tools*** trên Excel để tính toán như bảng 11.2. Nhìn vào cột *Coefficients* ta có thể viết được phương trình hồi qui tuyến tính mẫu, với hệ số a chính là *Intercept*, còn b thể hiện là *q2* trên bảng kết quả

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**-0.1552**	0.4969	-0.3123	0.7571
q2	**0.6995**	0.1009	6.9345	0.0000

Bảng 11.5: các hệ số của phương trình hồi qui tuyến tính mẫu

11.2.3. Hệ số xác định và kiểm định F trong phân tích hồi qui đơn giản

11.2.3.1 Hệ số xác định [1]

Mô hình hồi qui tuyến tính được xây dựng nhằm để giải thích sự biến thiên của biến phụ thuộc Y vào biến độc lập X nhưng liệu mô hình này đã thể hiện một cách tốt nhất mối liên hệ giữa Y và X chưa? Hoặc bao nhiêu phần trăm biến thiên của Y có thể giải thích bởi sự phụ thuộc tuyến tính của Y vào X? Hệ số xác định R^2 sẽ giúp trả lời điều này.

[1] Coefficient of determination

Ta có giá trị thực tế $y_i = a + bx_i + e_i$

Giá trị dự đoán theo phương trình hồi qui tuyến tính: $\hat{y}_i = a + bx_i$

Do đó: $y_i = \hat{y}_i + e_i$ (11.11)

Điều này có nghĩa là giá trị thực tế và giá trị dự đoán theo phương trình hồi qui tuyến tính có sự khác biệt e_i. e_i thể hiện phần biến thiên của Y không thể giải thích bởi mối liên hệ tuyến tính giữa Y và X

dùng các biến đổi toán học, ta có:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}e_i^2$$

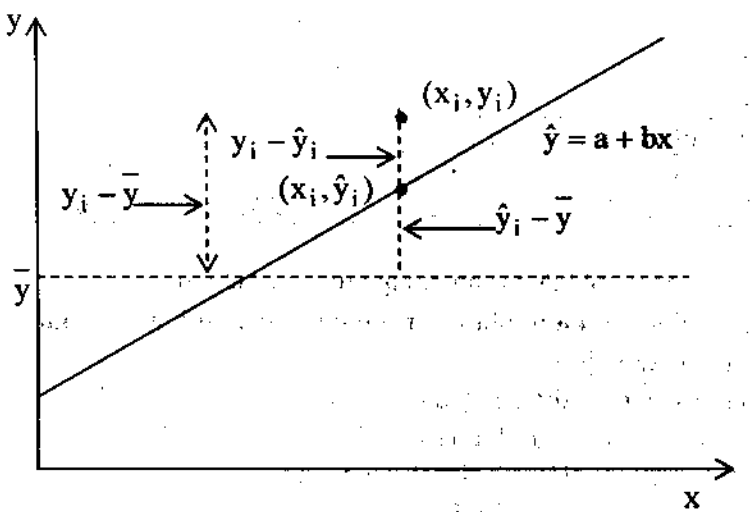
hay SST = SSR + SSE (11.12)

Ý nghĩa của các đại lượng này:

SST[1] : $\sum_{i=1}^{n}(y_i - \bar{y})^2$: thể hiện toàn bộ biến thiên của Y

SSR[2] : $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$: thể hiện phần biến thiên của Y được giải thích bởi biến X.

SSE[3] : $\sum_{i=1}^{n}e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: thể hiện phần biến thiên của Y do các nhân tố không được nghiên cứu đến.



[1] Total sum of squares

[2] Sum of squares for regression

[3] Sum of squares for error

Hình 11.3 : Minh họa SST, SSR và SSE

Do đó: hệ số xác định R^2 thể hiện phần tỉ lệ biến thiên của Y được giải thích bởi mối liên hệ tuyến tính của Y theo X, và được xác định theo công thức:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (11.13)$$

Ta có: $0 \leq R^2 \leq 1$ và R^2 chính là bình phương của hệ số tương quan r

R^2 càng lớn thì mô hình hồi qui tuyến tính đã xây dựng được xem là càng thích hợp và dĩ nhiên là càng có ý nghĩa trong việc giải thích sự biến thiên của Y.

11.2.3.2 Kiểm định F

Kiểm định F được sử dụng nhằm kiểm định giả thuyết về sự tồn tại của mối liên hệ tuyến tính giữa X và Y.

Biến thiên	Tổng các chênh lệch bình phương	Bậc tự do	Trung bình các chênh lệch bình phương (phương sai)	Giá trị kiểm định F
Hồi qui	SSR	1	MSR = $\frac{SSR}{1}$	$F_{(1,n-2)} = \frac{MSR}{MSE}$
Sai số	SSE	n - 2	MSE = $\frac{SSE}{n-2}$	
Tổng cộng	SST	n -1		

Bảng 11.5: Bảng ANOVA trong phân tích hồi qui tuyến tính đơn giản

Với ví dụ trên ta có thể dùng công cụ **Data Analysis** (chọn **Regression**) trong menu **Tools** trên Excel để tính toán như bảng kết quả dưới đây

Regression Statistics	
Multiple R	0.7950
R Square	**0.6320**
Adjusted R Square	0.6189
Standard Error	0.8932
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	38.3626	38.3626	48.0877	**0.0000**
Residual	28	22.3374	0.7978		
Total	29	60.7			

Bảng 11.6: Kết quả tính từ Excel

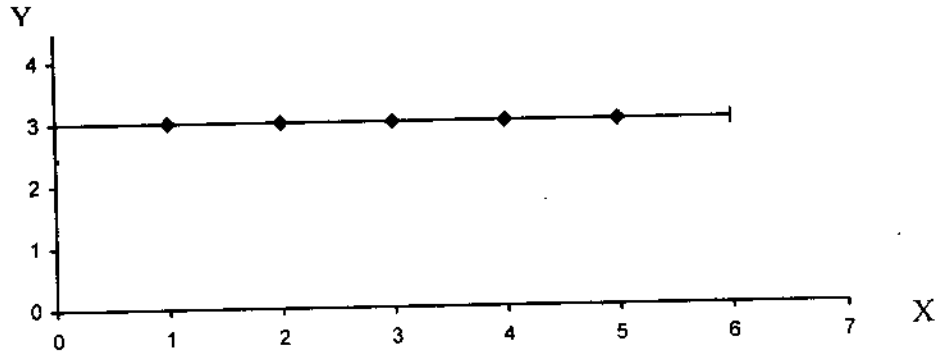
Trên bảng này R Square chính là hệ số xác định R^2, ở đây $R^2 = 0{,}6320$ hay 63,20%, ta có thể nói 63,20% biến thiên về số món hàng mua ngoài dự định có thể được giải thích bởi mối liên hệ tuyến tính giữa sự thay đổi của số món hàng mua ngoài dự định và mức thu nhập trung bình/tháng (triệu đồng). Giá trị P (*Significance F)* tính được là rất nhỏ (0,0000). Do vậy, có thể kết luận rằng có mối liên hệ tuyến tính giữa số món hàng mua ngoài dự định và mức thu nhập trung bình/tháng.

11.2.4. Kiểm định giả thuyết về mối liên hệ tuyến tính (kiểm định t)

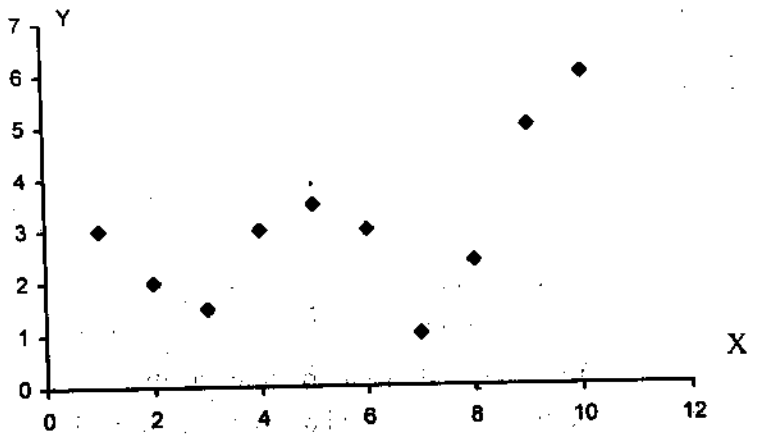
Trong hồi qui tuyến tính đơn giản, kiểm định F và kiểm định t là tương đương nhau.Với phương trình hồi qui tuyến tính của tổng thể thể hiện mối liên hệ tuyến tính giữa X và Y là: $\hat{y} = \alpha + \beta x + \varepsilon$

Ta nhận thấy $\beta = 0$ khi:

- Y là hằng số (không đổi) với bất kỳ giá trị của X
- Khi X và Y không có tương quan với nhau nghĩa là khi X tăng, Y có thể tăng, có thể giảm hoặc Y không đổi
- X và Y có liên hệ phi tuyến (không phải là tuyến tính) và lúc này ta có thể kết luận là X và Y không có liên hệ tuyến tính



Hình 11.4: Y không đổi với mọi giá trị của X



Hình 11.5: X và Y không có liên hệ tuyến tính

Do đó ta có thể dùng β để kiểm định

Giả thuyết: H_0 : $\beta = 0$ (Y và X không có liên hệ tuyến tính)

H_1 : $\beta \neq 0$ (Y và X có liên hệ tuyến tính)

Giá trị kiểm định: $$\frac{b}{S_b} \qquad (11.14)$$

với $$S_b = \sqrt{\frac{S_e^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}} \quad (11.15)$$

và $$S_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2} = MSE \quad (11.16)$$

trong đó: S_e^2 : phương sai của sai số

Qui tắc quyết định: đối với kiểm định cả 2 phía ở mức ý nghĩa α, bác bỏ H_0, nếu:

$$\frac{b}{S_b} < -t_{n-2,\alpha/2} \qquad \text{hay} \qquad \frac{b}{S_b} > t_{n-2,\alpha/2}$$

Ở ví dụ trên, ta có b = 0,6995, $\sum_{i=1}^{n} x_i^2 = 728$, $\overline{x} = 4{,}653$

S_e^2 = MSE = 0,7978 (kết quả trên bảng 11.4),

$$S_b = \sqrt{\frac{S_e^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}} = \sqrt{\frac{0{,}7978}{728-(30\times(4{,}653)^2)}} = 0{,}1009$$

$$t_{n-2} = \frac{b}{S_b} = \frac{0{,}6995}{0{,}1009} = 6{,}9345$$

và tra bảng phân phối t, ta thấy giá trị kiểm định tính được nằm rất xa trong vùng bác bỏ, giả thuyết H_0 bị bác bỏ và do đó, có thể kết luận tồn tại mối liên hệ tuyến tính giữa số món hàng mua ngoài dự định và mức thu nhập trung bình/tháng.

Trên Excel có thể nhanh chóng có kết luận này bằng cách trên bảng 11.2 nhìn vào dòng q2 cột *Standard Error* để có giá trị S_b và cột *t Stat* để có giá trị t_{n-2}

11.2.5. Khoảng tin cậy của các hệ số hồi qui

Với phương pháp bình phương bé nhất ta đã có các ước lượng điểm của α, β. Trong thực tế, ta thường quan tâm đến ước lượng khoảng của hệ số β, chính là độ dốc của đường hồi qui.

- Ước lượng khoảng của β với độ tin cậy (1- α) 100% là:

$$b \pm t_{n-2,\alpha/2} S_b \qquad (11.17)$$

trong đó $t_{n-2,\alpha/2}$ là giá trị của đại lượng ngẫu nhiên T có phân phối Student với n-2 bậc tự do,

S_b sai số chuẩn ước lượng của b theo công thức (11.15)

Tra bảng phân phối t, hoặc trên Excel vào **Insert – Function**, dùng hàm **TINV** ta có $t_{28,0,25} = 2,0484$.

Với ví dụ 1, khoảng ước lượng của β với độ tin cậy 95% là:

$$0,6995 - (2,0484)(0,1009) < \beta < 0,6995 + (2.0484)(0,1009)$$

$$0,4929 < \beta < 0,79062$$

Kết quả này cũng có thể dễ dàng nhận ra khi nhìn vào cột *Lower 95%* và *Upper 95%* dòng q2 (dùng công cụ **Data Analysis** (chọn **Regression**) trong menu **Tools** trên Excel)

	Coefficients	*Lower 95%*	*Upper 95%*
Intercept	-0.1552	-1.1731	0.8627
q2	0.6995	**0.4929**	**0.9062**

Bảng 11.7: khoảng tin cậy của hệ số hồi qui

Do đó với độ tin cậy 95%, có thể nói rằng khi mức thu nhập trung bình/tháng tăng 1 triệu đồng thì số món hàng mua ngoài dự định sẽ tăng trong khoảng từ 0,49 đến 0,97 món hàng.

Ngoài ra, khoảng ước lượng này khác không ($\beta \neq 0$): ta lại có thể khẳng định một lần nữa là tồn tại mối liên hệ tuyến tính giữa mức thu nhập trung bình/tháng và số món hàng mua ngoài dự định.

- Ước lượng khoảng của α với độ tin cậy $(1 - \alpha)100\%$ là:

$$a \pm t_{n-2,\alpha/2} S_a \qquad (11.18)$$

trong đó: t_{n-2} có phân phối Student với n-2 bậc tự do

S_a sai số chuẩn ước lượng của a

$$S_a = \sqrt{S_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)} \qquad (11.19)$$

11.3 HỒI QUI BỘI

Trong thực tế, không chỉ có một biến X ảnh hưởng đến Y, sự thay đổi của biến phụ thuộc Y có thể sẽ được giải thích toàn diện, đầy đủ hơn nếu được đặt trong mối liên hệ với nhiều biến độc lập X. Đây chính là phương pháp phân tích hồi qui bội. Ví dụ như đối với một nước nông nghiệp thì xem xét tốc độ tăng trưởng của nền kinh tế sẽ thật sự tốt hơn khi ta không chỉ nghiên cứu tốc độ tăng trưởng của nông nghiệp mà còn xem xét đến tốc độ tăng trưởng của kim ngạch xuất khẩu, tỉ lệ lạm phát, ... Ở ví dụ 11.1 với mô hình hồi qui đơn giản, ta thấy số món hàng mua ngoài dự định có thể được dự đoán bằng thu nhập hộ gia đình. 63,20% (trên kết quả bảng 11.4 dòng R square : 0,6320) biến thiên của số món hàng mua ngoài dự định của khách hàng có thể được giải thích bởi sự khác biệt về mức thu nhập hộ trung bình/tháng. Nếu chúng ta đưa thêm các biến tuổi (q3), số lần đi siêu thị (q1) vào mô hình thì chúng sẽ có ảnh hưởng như thế nào đối với số món hàng mua ngoài dự định (q4)?

11.3.1 Mô hình hồi qui bội của tổng thể

Giả sử ta có biến Y phụ thuộc vào k biến độc lập $X_1, X_2, ..., X_k$. Nếu giá trị của k biến độc lập $X_1, X_2, ..., X_k$ lần lượt là $x_{1i}, x_{2i}, ..., x_{ki}$ thì giá trị của biến phụ thuộc y_i, thể hiện qua mô hình hồi qui bội dạng tuyến tính như sau: $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$ (11.20)

Trong đó: α, β_j (j = 1,2, . k): là các hằng số

α là giá trị ước lượng của biến Y khi k biến X có giá trị = 0

x_{pi} : biểu hiện giátrị của biến độc lập thứ p tại quan sát thứ i.

β_j (j = 1,2, ... k): là các tham số chưa biết, được gọi là các hệ số hồi qui riêng[1], thể hiện mức thay đổi của trung bình biến Y khi các biến X_j thay đổi một đơn vị, các biến còn lại không thay đổi. Như vậy, β_j cho thấy ảnh hưởng của riêng biến X_i đến trung bình biến Y.

ε_i là sai số, chính là biến độc lập ngẫu nhiên có phân phối chuẩn, với trung bình = 0, phương sai không đổi (bằng nhau) và là độc lập, không có liên hệ với nhau.

---

[1] Partial regression coefficients

[2] Xem thêm " Xử lý dữ liệu nghiên cứu với SPSS for Windows "– Hoàng Trọng, NXB Thống Kê

Mô hình này cho rằng biến phụ thuộc có phân phối chuẩn đối với bất kỳ kết hợp nào của các biến độc lập trong mô hình[2].

11.3.2. Phương trình hồi qui bội của mẫu

Trong thực tế ta không thể xác định một cách chính xác các hệ số α , β_j (j = 1,2, . k) của phương trình hồi qui bội của tổng thể mà ta chỉ có thể ước lượng chúng từ các giá trị quan sát của mẫu thu thập được

Phương pháp bình phương bé nhất vẫn được dùng để ước lượng cho α , β_j (j = 1,2,... k) bằng cách xác định các hệ số a, $b_1, b_2, ..., b_k$

$$\sum_{i=1}^{n}(y_i - a - b_1x_{1i} - b_2x_{2i} - ... - b_kx_{ki})^2 = \min \qquad (11.19)$$

Phương trình hồi qui bội của mẫu có dạng:

$$\hat{y} = a + b_1x_1 + b_2x_2 + ... + b_kx_k$$

Trước khi xác định mô hình hồi qui bội thì cần xây dựng một ma trận tương quan cho tất cả các biến nghiên cứu. Sử dụng số liệu ví dụ 1, nghiên cứu xem số món hàng mua ngoài dự định phụ thuộc như thế nào vào mức thu nhập, số lần đi siêu thị và tuổi của khách hàng bằng cách xây dựng ma trận tương quan.

11.3.3 Ma trận tương quan [1]

Ma trân này cho biết tương quan giữa biến phụ thuộc với từng biến độc lập cũng như giữa các biến độc lập với nhau.

Bảng 11.8: Ma trận tương quan giữa các biến số món hàng mua ngoài dự định, số lần đi siêu thị và thu nhập hộ trung bình/tháng.(Vào Analyze – Regression- Linear trên SPSS)

[1] Correlation matrix

Correlations

		so mon hang mua ngoai du dinh	so lan di ST trong thang qua	thu nhap ho TB thang (trd)
Pearson Correlation	so mon hang mua ngoai du dinh	1.000	.783	.795
	so lan di ST trong thang qua	.783	1.000	.747
	thu nhap ho TB thang (trd)	.795	.747	1.000
Sig. (1-tailed)	so mon hang mua ngoai du dinh	.	.000	.000
	so lan di ST trong thang qua	.000	.	.000
	thu nhap ho TB thang (trd)	.000	.000	.
N	so mon hang mua ngoai du dinh	30	30	30
	so lan di ST trong thang qua	30	30	30
	thu nhap ho TB thang (trd)	30	30	30

Trong bảng 11.8 ta thấy tương quan giữa số món hàng mua ngoài dự định và số lần đi siêu thị trong tháng qua với r = 0,783 là thấp hơn so với tương quan giữa số món hàng mua ngoài dự định và thu nhập hộ trung bình/tháng với r= 0,795.

Với bảng 11.9, xét thêm 1 biến phụ thuộc là biến tuổi, ta cũng thấy tương quan giữa số món hàng mua ngoài dự định và thu nhập hộ trung bình/tháng có r = 0,795 hơi cao hơn so với tương quan giữa số món hàng mua ngoài dự định và số lần đi siêu thị trong tháng qua với r = 0,783. Ngoài ra, giữa tuổi và số món hàng mua ngoài dự định cũng có liên hệ tương quan nghịch và yếu với R = - 0,168.

11.3.4 Kiểm định F

Được dùng để kiểm định giả thiết về sự tồn tại của mối liên hệ tuyến tính giữa biến phụ thuộc Y với bất kỳ một biến độc lập X_j nào đó.

Correlations

		so mon hang mua ngoai du dinh	so lan di ST trong thang qua	tuoi	thu nhap ho TB thang (trd)
Pearson Correlation	so mon hang mua ngoai du dinh	1.000	.783	-.168	.795
	so lan di ST trong thang qua	.783	1.000	-.288	.747
	tuoi	-.168	-.288	1.000	-.165
	thu nhap ho TB thang (trd)	.795	.747	-.165	1.000
Sig. (1-tailed)	so mon hang mua ngoai du dinh	.	.000	.187	.000
	so lan di ST trong thang qua	.000	.	.062	.000
	tuoi	.187	.062	.	.192
	thu nhap ho TB thang (trd)	.000	.000	.192	.
N	so mon hang mua ngoai du dinh	30	30	30	30
	so lan di ST trong thang qua	30	30	30	30
	tuoi	30	30	30	30
	thu nhap ho TB thang (trd)	30	30	30	30

Bảng 11.9: Ma trận tương quan giữa các biến số món hàng mua ngoài dự định , số lần đi siêu thị, thu nhập hộ trung bình/tháng và tuổi.(Vào **Analyze – Regression- Linear** trên SPSS)

Giả thiết $H_0: \beta_1 = \beta_2 = ... = \beta_k = 0$

H_1: có ít nhất một $\beta_j \neq 0$ (không phải tất cả $\beta_j = 0$)

Nếu chấp nhận H_0, tức là không tồn tại mối liên hệ tuyến tính giữa Y với bất kỳ một biến X_j nào đó. Và ngược lại, bác bỏ H_0 ta có thể kết luận có mối liên hệ tuyến tính giữa Y với ít nhất một trong các biến X_j.

Biến thiên	Tổng các chênh lệch bình phương	Bậc tự do	Trung bình các chênh lệch bình phương (phương sai)	Giá trị kiểm định F
Hồi qui	SSR	k	$MSR = \frac{SSR}{k}$	$F_{k,n-(k+1)} = \frac{MSR}{MSE}$
Sai số	SSE	n – (k + 1)	$MSE = \frac{SSE}{n-(k+1)}$	
Tổng cộng	SST	n -1		

Bảng 11.10: kiểm định F trong phân tích hồi qui bội

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	38.363	1	38.363	48.088	.000[a]
	Residual	22.337	28	.798		
	Total	60.700	29			
2	Regression	43.306	2	21.653	33.612	.000[b]
	Residual	17.394	27	.644		
	Total	60.700	29			

a. Predictors: (Constant), thu nhap ho TB thang (trd)

b. Predictors: (Constant), thu nhap ho TB thang (trd), so lan di ST trong thang qua

c. Dependent Variable: so mon hang mua ngoai du dinh

Bảng11.11: ANOVA [1] và kiểm định F (Vào **Analyze – Regression – Linear** trên SPSS)

Trong cả 2 trường hợp mô hình 1 (chỉ có biến thu nhập hộ trung bình tháng trong mô hình) và mô hình 2 (thêm biến số lần đi siêu thị trong tháng qua vào mô hình) thì giá trị p của kiểm định F (cột Sig.) là rất nhỏ (0,000), do đó ta bác bỏ giả thiết cho rằng cả 2 hệ số β_1, β_2 đều bằng không, nghĩa là ta có thể kết luận rằng tồn tại mối liên hệ tuyến tính giữa số món hàng mua ngoài dự định với ít nhất một trong các biến, số lần đi siêu thị, thu nhập hộ trung bình/tháng hoặc cả hai yếu tố này.

Nếu tiếp tục xem xét mối quan hệ giữa cả 4 biến: số món hàng mua ngoài dự định, tuổi, số lần đi siêu thị, thu nhập hộ trung bình/tháng, ta có kết quả kiểm định ANOVA và kiểm định F như bảng 11.12

[1] Phân tích phương sai (Analysis of variance)

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43.382	3	14.461	21.710	.000[a]
	Residual	17.318	26	.666		
	Total	60.700	29			
2	Regression	43.306	2	21.653	33.612	.000[b]
	Residual	17.394	27	.644		
	Total	60.700	29			

a. Predictors: (Constant), tuoi, thu nhap ho TB thang (trd), so lan di ST trong thang qua

b. Predictors: (Constant), thu nhap ho TB thang (trd), so lan di ST trong thang qua

c. Dependent Variable: so mon hang mua ngoai du dinh

Bảng 11.12 : ANOVA và kiểm định F (Vào **Analyze – Regression – Linear** trên SPSS)

Nhìn vào mô hình 1 trên bảng 11.12 thì giá trị p của kiểm định F *(cột Sig.)* là rất nhỏ (0,000), do đó ta bác bỏ giả thiết cho rằng cả 3 hệ số $\beta_1, \beta_2, \beta_3$ đều bằng không, nghĩa là ta có thể kết luận rằng tồn tại mối liên hệ tuyến tính giữa số món hàng mua ngoài dự định với ít nhất một trong các biến, số lần đi siêu thị, thu nhập hộ trung bình/tháng, tuổi hoặc cả ba yếu tố này.

11.3.5. Hệ số hồi qui từng phần [1]

Các ký hiệu B trên Hình 11.13 được gọi là hệ số hồi qui từng phần, bởi vì hệ số của một biến nào đó được điều chỉnh theo các biến độc lập khác trong mô hình).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-.155	.497		-.312	.757			
	thu nhap ho TB thang (trd)	.700	.101	.795	6.935	.000	.795	.795	.795
2	(Constant)	-.088	.447		-.197	.845			
	thu nhap ho TB thang (trd)	.418	.136	.475	3.064	.005	.795	.508	.318
	so lan di ST trong thang qua	.455	.164	.429	2.770	.010	.783	.470	.285

a. Dependent Variable: so mon hang mua ngoai du dinh

Bảng 11.13: **các thông số thống kê của từng biến trong phương trình (vào Analyze – Regression- Linear, trên SPSS (phương pháp Stepwise))**

[1] Partial regression coefficients

Theo bảng 11.13 ta có thể có các phương trình thể hiện mối liên hệ giữa số món hàng mua ngoài dự định phụ thuộc vào các biến:

- Theo mô hình 1:

Y = -0,155 + 0,700 (thu nhập hộ trung bình/tháng)

- Theo mô hình 2:

Y = -0,088 + 0,418 (thu nhập hộ trung bình/tháng) + 0,455 (số lần đi siêu thị)

Nếu tiếp tục xem xét mối quan hệ giữa cả 4 biến: số món hàng mua ngoài dự định, tuổi, số lần đi siêu thị, thu nhập hộ trung bình/tháng, ta có bảng 11.14

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-.358	.923		-.388	.701			
	so lan di ST trong thang qua	.470	.173	.443	2.72	.011	.783	.471	.285
	thu nhap ho TB thang (trd)	.414	.139	.470	2.98	.006	.795	.504	.312
	tuoi	.007	.020	.037	.337	.739	-.168	.066	.035
2	(Constant)	-.088	.447		-.197	.845			
	so lan di ST trong thang qua	.455	.164	.429	2.77	.010	.783	.470	.285
	thu nhap ho TB thang (trd)	.418	.136	.475	3.06	.005	.795	.508	.316

a. Dependent Variable: so mon hang mua ngoai du dinh

Bảng 11.14: các thông số thống kê của từng biến trong phương trình (vào *Analyze – Regression- Linear*, trên SPSS (phương pháp Stepwise))

Ta có phương trình thể hiện mối liên hệ giữa số món hàng mua ngoài dự định phụ thuộc vào các biến:

- Theo mô hình 1:

Y = -0,358 + 0,470 (số lần đi siêu thị) + 0,414 (thu nhập hộ trung bình/tháng) +0,007 (tuổi)

Kết quả mô hình 2 giống như trên bảng 11.13.

Dùng Excel, ta cũng có kết quả tương tự

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.84466
R Square	0.713451
Adjusted R Square	0.692226
Standard Error	0.802623
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	43.3065	21.6532	33.6124	0.0000
Residual	27	17.3935	0.6442		
Total	29	60.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.0882	0.4472	-0.1972	0.8452	-1.0058	0.8294
q1	0.4555	0.1644	2.7703	0.0100	0.1181	0.7929
q3	0.4176	0.1363	3.0638	0.0049	0.1379	0.6972

11.3.6. Kiểm định giả thiết về các hệ số hồi qui (kiểm định t)

Do kiểm định F đóng vai trò xem xét một cách tổng quát, vì vậy cần thực hiện các kiểm định t riêng biệt để đánh giá ý nghĩa của từng biến khác nhau.

Giả thiết: $H_0 : \beta_j = 0$ (j = 1,2, . . . , k)

$H_1 : \beta_j \neq 0$

Giá trị kiểm định: $\dfrac{b_j}{S_{b_j}}$

Qui tắc quyết định: ở mức ý nghĩa α, bác bỏ giả thuyết H_0 nếu:

$$\frac{b_j}{S_{b_j}} < t_{n-(k+1),\alpha/2} \quad \text{hay} \quad \frac{b_j}{S_{b_j}} > t_{n-(k+1),\alpha/2}$$

Các kiểm định t này sẽ cho ta biết biến X_j nào không có ảnh hưởng đến Y ($\beta_j = 0$), X_j nào có ý nghĩa trong việc giải thích biến thiên của Y ($\beta_j \neq 0$), và do đó nên được thể hiện trong phương trình hồi qui.

Ở bảng 11.12, với mô hình 1 giá trị p khá nhỏ (= 0,005) (cột *Sig*, dòng thu nhập hộ trung bình/tháng), vì vậy ta có thể nói biến thu nhập hộ trung bình/tháng có ý nghĩa trong việc giải thích biến thiên của số món hàng mua ngoài dự định.
Cùng bảng 11.12, với mô hình 2 giá trị p của dòng kiểm định *số lần đi siêu thị trong tháng* cũng khá nhỏ (p = 0,010) (cột *Sig*, dòng *số lần đi siêu thị trong tháng*), do đó có thể nói biến số lần đi siêu thị cũng có ý nghĩa trong việc giải thích biến thiên của số món hàng mua ngoài dự định.
Như vậy, với mô hình 2 ta có thể kết luận cả 2 biến số lần đi siêu thị và thu nhập hộ trung bình/tháng đều có ý nghĩa trong việc giải thích biến thiên của số món hàng mua ngoài dự định.

Ở bảng 11.13, ở mô hình 1 với 3 biến độc lập thì chỉ có giá trị p được thể hiện ở cột *Sig* tương ứng với dòng biến tuổi có giá trị p khá lớn (0,739), vì vậy với mức ý nghĩa 5%, ta có thể nói biến số lần đi siêu thị, thu nhập hộ trung bình/tháng có ý nghĩa trong việc giải thích biến thiên của số món hàng mua ngoài dự định, còn biến tuổi không có ý nghĩa trong việc giải thích biến thiên của số món hàng mua ngoài dự định.

11.3.7. Hệ số xác định và hệ số xác định đã điều chỉnh [1]

Hệ số xác định R^2 đo lường phần biến thiên của Y có thể được giải thích bởi các biến độc lập X, đây chính là đại lượng thể hiện sự thích hợp của mô hình hồi qui bội đối với dữ liệu. R^2 càng lớn thì mô hình hồi qui bội xây dựng được xem là càng thích hợp và càng có ý nghĩa trong việc giải thích sự biến thiên của Y. Tuy nhiên cần lưu ý khi R^2 đặc biệt cao, như $R^2 = 0,99$ hay 0,999, vì khi đó số lượng các biến độc lập X trong mô hình hồi qui bội tăng lên do đó giá trị của R^2 cũng tăng lên. Lúc này thì dường như mô hình hồi qui là rất tốt nhưng thực tế lại không thích hợp vì không thể sử dụng mô hình để dự đoán.
Do đó muốn đo lường mức độ thích hợp của mô hình hồi qui bội, ta phải dùng đến hệ số R^2 có tính đến bậc tự do của SSE và SST, được gọi là hệ số xác định đã điều chỉnh $\overline{R}^2$, tính theo công thức sau:

$$\overline{R}^2 = 1 - \frac{SSE/[n-(k+1)]}{SST/(n-1)} \qquad (11.20)$$

Trừ khi số lượng biến X là tương đối lớn so với n, R^2 và $\overline{R}^2$ sẽ không chênh lệch nhau nhiều lắm. Vì vậy ta hầu như dùng $\overline{R}^2$ khi muốn xem xét việc có

[1] Adjusted R Square

nên đưa thêm một biến giải thích X_j nào đó vào mô hình hồi qui bội hay không. Nếu $\overline{R}^2$ tăng lên chứng tỏ là việc đưa thêm biến X_j vào mô hình đã làm tăng ý nghĩa của mô hình và vì vậy cần thiết để X_j trong mô hình.

Do đó để đánh giá tầm quan trọng tương đối của các biến độc lập ta cần xem xét mức độ tăng R^2 của khi một biến được đưa vào phương trình khi phương trình đã chứa sẵn các biến độc lập khác. Mức tăng [1] này là: $R^2_{change} = R^2 - R^2_{(j)}$

Trong đó $R^2_{(j)}$: bình phương hệ số tương quan bội khi tất cả các biến độc lập có trong mô hình ngoại trừ biến j. Mức độ thay đổi do một biến của R^2 lớn cho thấy biến này cung cấp những thông tin độc nhất về biến phụ thuộc mà các biến độc lập khác trong phương trình không có.

11.3.8. Hệ số tương quan từng phần [2], hệ số tương quan riêng[3] và hệ số tương quan bội [4]

Khi lấy căn bậc hai của mức gia tăng này ta sẽ có hệ số tương quan từng phần ($\sqrt{R^2_{change}}$), đây chính là hệ số nói lên tương quan giữa Y và X_j khi ảnh hưởng tuyến tính của các biến độc lập khác đối với biến độc lập X_j bị loại bỏ khỏi mô hình. Nếu tất cả các biến độc lập không có tương quan với nhau thì mức độ thay đổi của R^2 khi một biến được đưa vào phương trình đơn giản chỉ là bình phương của hệ số tương quan giữa biến này và biến phụ thuộc.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.795[a]	.632	.619	.89	.632	46.088	1	28	.000
2	.845[b]	.713	.692	.80	.081	7.674	1	27	.010

a. Predictors: (Constant), thu nhap ho TB thang (trd)

b. Predictors: (Constant), thu nhap ho TB thang (trd), so lan di ST trong thang qua

Bảng 11.15: kết quả thay đổi của R^2 (R Square) (theo phương pháp *Stepwise* trên SPSS)

---

[1] R Square change

[2] Part correlation coefficient

[3] Partial correlation cosfficient

[4] Coefficient of multiple correlation

Mức độ thay đổi của R^2 cho thấy nếu thêm biến thu nhập vào mô hình 1 (đã chứa biến số lần đi siêu thị) tức là sử dụng mô hình 2 (có cả 2 biến số lần đi siêu thị và thu nhập hộ trung bình/tháng, thì sẽ làm cho R^2 thay đổi và cụ thể là tăng lên 0,081 (0,713- 0,632 = 0,081).
(Con số 0,632 trên dòng Model 1, cột R Square Change, số 0,713 trên dòng Model 2 cột R Square). Giá trị này chỉ cho biết R^2 tăng lên bao nhiêu khi thêm một biến vào mô hình hồi qui, chứ nó không chỉ ra được tỉ lệ của phần biến thiên mà riêng một mình biến thu nhập có thể giải thích được. Trong trường hợp các biến thiên đã được giải thích bởi các biến khác rồi thì những biến còn lại chỉ có thể làm R^2 thay đổi rất ít, không đáng kể (như trường hợp của biến tuổi nếu ta thêm vào mô hình)[1]
Hệ số đo lường được phần biến thiên giảm xuống (hay phần biến thiên giải thích được) của một biến là $P_{r_j^2} = \dfrac{R^2 - R^2_{(j)}}{1 - R^2_{(j)}} = \dfrac{R^2_{change}}{1 - R^2_{change}} < \left|\sqrt{R^2_{change}}\right|$

1- R^2_{change} : tỉ lệ biến thiên không giải thích được nguyên nhân khi tất cả các biến có mặt trong mô hình hồi qui ngoại trừ biến j

$\sqrt{P_{r_j^2}}$: hệ số tương quan riêng. Hệ số này được giải thích như là tương quan giữa biến độc lập thứ j và biến phụ thuộc khi ảnh hưởng tuyến tính của các biến độc lập khác đối với cả Y và X_j bị loại bỏ (giữ không đổi).

```
- - P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S - -

Controlling for..    Q2

                    Q4          Q1

Q4               1.0000       .4705
                (     0)    (    27)
                P= .        P= .010

Q1                .4705     1.0000
                (    27)    (     0)
                P= .010     P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed

- - -  P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S  -
- -

Controlling for..    Q1
```

[1] Xem thêm " Xử lý dữ liệu nghiên cứu với SPSS for Windows "– Hoàng Trọng, NXB Thống Kê

```
                    Q4          Q2

Q4             1.0000       .5079
               (    0)     (   27)
               P= .        P= .005

Q2               .5079     1.0000
               (   27)     (    0)
               P= .005     P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

Bảng 11.16: Hệ số tương quan riêng (Vào **Analyze – Correlate – Partial** trên SPSS)

Hệ số tương quan riêng giữa số món hàng mua ngoài dự định và số lần đi siêu thị trong tháng qua khi thu nhập hộ trung bình/tháng được giữ không đổi là $(0,4705)^2 = 0,2214$, điều này có nghĩa là nếu như thu nhập hộ trung bình/ tháng không đổi thì chỉ có 22,14% biến thiên trong số món hàng mua ngoài dự định có thể giải thích được bởi sự phụ thuộc tuyến tính vào số lần đi siêu thị trong tháng qua.

Ngoài ra, để đo lường cường độ của mối liên hệ giữa Y với các biến độc lập X, tức là liên hệ giữa các giá trị thực tế y_i với các giá trị dự đoán $\hat{y}_i$ theo mô hình hồi qui bội, ta dùng hệ số tương quan bội R. hệ số tương quan bội đo lường một cách tổng quát cường độ của mối liên hệ tương tự như hệ số tương quan giữa hai biến X, Y.

$$R = \sqrt{R^2} = \sqrt{0,713} = 0,845$$

Kết quả này ta có thể nhìn trên bảng 11.14, cột R, dòng Model 2

11.3.9 Khoảng tin cậy của các hệ số hồi qui bội

Ước lượng khoảng của các hệ số β_j với độ tin cậy (1 - α)100% là:

$$b_j \pm t_{n-k+1,\alpha/2} S_b$$

trong đó: $t_{n-k+1,\alpha/2}$ là giá trị của biến ngẫu nhiên T có phân phối Student với n-k+1 bậc tự do

S_{b_j} sai số chuẩn ước lượng của b_j

Coefficients[a]

Model		95% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	-1.173	.863
	thu nhap ho TB thang (trd)	.493	.906
2	(Constant)	-1.006	.829
	thu nhap ho TB thang (trd)	.138	.697
	so lan di ST trong thang qua	.118	.793

a. Dependent Variable: so mon hang mua ngoai du dinh

Bảng 11.17 : Kết quả khoảng tin cậy của các hệ số hồi qui

Như vậy với độ tin cậy 95%, ta có thể nói:

Mô hình 1:

- với thu nhập hộ trung bình/tháng cố định, thì số lần đi siêu thị trong tháng qua khi tăng 1 lần, số món hàng mua ngoài dự định trung bình sẽ tăng từ 0,493 đến 0,906 món. (cột *Lower Bound* và *Upper Bound dòng thu nhập hộ TB tháng*).

Mô hình 2:

- Với thu nhập hộ trung bình/tháng (triệu đồng) cố định, khi số lần đi siêu thị tăng lên 1 lần thì số món hàng mua ngoài dự định trung bình sẽ tăng lên từ 0,138 đến 0,697 món. (cột *Lower Bound* và *Upper Bound dòng thu nhập hộ TB tháng*).
- với số lần đi siêu thị trong tháng qua cố định, thì khi thu nhập hộ trung bình/tháng tăng 1 triệu đồng, số món hàng mua ngoài dự định trung bình sẽ tăng từ 0,118 đến 0,793 món. (cột *Lower Bound* và *Upper Bound dòng số lần đi siêu thị trong tháng*).

11.3.10 Dự đoán trong phân tích hồi qui bội

Tương tự như trong hồi qui đơn giản, trong hồi qui bội, giá trị dự đoán của Y có được, ứng với các giá trị cho trước của k biến X bằng cách thay các giá trị của k biến X vào phương trình hồi qui bội.

Các giá trị cho trước của k biến X lần lượt là $x_{1,n+1}, x_{2,n+2}, \ldots, x_{k,n+1}$ là thì giá trị dự đoán $\hat{y}_{n+1}$ sẽ là:

$$\hat{y}_{n+1} = a + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \ldots + b_k x_{k,n+1} \qquad (11.21)$$

Với ví dụ 1, nếu ta muốn dự đoán số món hàng mua ngoài dự định với khách hàng có số lần đi siêu thị trong tháng là 4, thu nhập hộ trung bình/tháng là 3 triệu đồng:

$$\hat{y}_{n+1} = a + b_1 x_{1,n+1} + b_2 x_{2,n+1}$$

$$= -0{,}088 + 0{,}455\,(4) + 0{,}418\,(3)$$

$$= 2{,}986$$

Như vậy số món hàng mua ngoài dự định được dự đoán là khoảng 3 món.

Chương 12 DÃY SỐ THỜI GIAN

Để phân tích biến động của hiện tượng qua thời gian, ta thường dùng phương pháp phân tích dãy số thời gian. Trong phương pháp này các giá trị quan sát không độc lập với nhau, ngược lại sự phụ thuộc của các giá trị quan sát trong dãy số là đặc điểm, cơ sở cho việc xây dựng các phương pháp nghiên cứu và dự đoán về dãy số thời gian. Các phương pháp dự đoán định lượng có thể được phân chia thành hai loại: phân tích các mức độ qua thời gian và phân tích liên hệ nguyên nhân- kết qủa[1]. 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 và cả hiện tại. Phân tích mối liên hệ nhân qủa 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 các mức độ qua thời gian được dựa trên giả định cơ bản là các yếu tố ảnh hưởng đến biến động của hiện tượng trong quá khứ và hiện tại sẽ còn tiếp tục tồn tại với cùng tính chất, đặc điểm, cường độ như vậy đối với biến động của hiện tượng trong tương lai. Do đó, mục tiêu chính của phân tích dãy số thời gian là nhận ra và tách riêng các yếu tố ảnh hưởng này phục vụ cho mục đích dự đoán cũng như cho việc kiểm soát và hoạch định trong quản lý.

12.1 ĐỊNH NGHĨA

Dãy số thời gian là một dãy các giá trị của hiện tượng nghiên cứu được sắp xếp theo thứ tự thời gian. Ví dụ như giá cả hàng ngày một cổ phiếu nào đó ở thị trường chứng khoán X ở thời điểm đóng cửa. Các ấn bản hàng tháng về chỉ số giá cả ở một thành phố, ...

Một dãy số 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

[1] Causal

Căn cứ vào đặc điểm về mặt thời gian của dãy số, ta có thể chia ra 2 loại dãy số: dãy số thời kỳ và dãy số thời điểm.

12.1.1 Dãy số thời kỳ: là dãy số biểu hiện sự biến động của hiện tượng nghiên cứu qua từng thời kỳ.

Các mức độ trong dãy số thời kỳ có thể cộng lại với nhau qua thời gian, để phản ánh mặt lượng của hiện tượng nghiên cứu trong một thời kỳ dài hơn.

Ví dụ 1:

Năm	97	98	99	2000	2001	2002
sản lượng (tấn)	1200	1340	1512	1570	1653	1615

12.1.2 Dãy số thời điểm: là dãy số biểu hiện sự biến động của hiện tượng nghiên cứu qua các thời điểm nhất định.

Các mức độ trong dãy số thời điểm không thể cộng lại theo thời gian vì con số cộng này không có ý nghĩa kinh tế.

Ví dụ 2:

Ngày	1/1/03	1/2/03	1/3/03	1/4/03
sản lượng tồn kho (tấn)	380	395	350	420

12.2 CÁC THÀNH PHẦN CỦA DÃY SỐ THỜI GIAN:

Biến động của một dãy số thời gian có thể được xem như là kết quả hợp thành của 4 yếu tố thành phần sau:

- *Xu hướng* (T)[1] thể hiện chiều hướng biến động, tăng hoặc giảm của hiện tượng nghiên cứu trong một thời gian dài. Nguyên nhân của những biến động có tính xu hướng có thể là 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, ...
- *Thời vụ* (S)[2] biểu hiện qua sự giảm hay tăng mức độ của hiện tượng ở một số thời điểm (tháng hay quí) nào đó được lặp đi lặp lại qua nhiều năm. Biến động thời vụ thường do các nguyên nhân như điều kiện thời tiết, khí hậu, tập quán xã hội, tín ngưỡng, ... Biến động thời vụ được xem xét khi dữ liệu được thu thập theo tháng, quí, tức là khi chu kỳ biến động là một năm nếu chu kỳ lớn hơn 1 năm ta sẽ có biến động chu kỳ.

[1] Trend component

[2] Seasonal component

- *Chu kỳ* (C)[1] biến động của hiện tượng được lặp lại với một chu kỳ nhất định, thường kéo dài từ 2 – 10 năm. Biến động theo chu kỳ là do tác động tổng hợp của nhiều yếu tố khác nhau.
- *Ngẫu nhiên* (I)[2] biến động không có qui luật và hầu như không thể dự đoán. Loại biến động này thường xảy ra trong thời gian ngắn và gần như không lặp lại, do ảnh hưởng của thiên tai, động đất, nội chiến, chiến tranh, ...

Bốn thành phần trên có thể kết hợp với nhau theo mô hình nhân[3]

$$y_i = T_i.S_i.C_i.I_i \qquad (12.1)$$

T_i : thành phần xu hướng ở thời gian i

S_i : thành phần thời vụ ở thời gian i

C_i : thành phần chu kỳ ở thời gian i

I_i : thành phần ngẫu nhiên ở thời gian I

12.3 CÁC CHỈ TIÊU MÔ TẢ DÃY SỐ THỜI GIAN

12.3.1 Mức độ trung bình theo thời gian

Là số trung bình của các giá trị của hiện tượng nghiên cứu trong dãy số thời gian. Đây là chỉ tiêu biểu hiện mức độ điển hình, chung nhất của hiện tượng trong thời gian nghiên cứu.

Giả sử ta có dãy số thời gian $y_1, y_2, ..., y_n$

Gọi $\bar{y}$: mức độ trung bình của dãy số

12.3.1.1 Dãy số thời kỳ:

$$\bar{y} = \frac{y_1 + y_2 + ... + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad (12.2)$$

12.3.1.2 Dãy số thời điểm: có 2 trường hợp

- *Khoảng cách thời gian giữa các thời điểm bằng nhau*

$$\bar{y} = \frac{\frac{1}{2}y_1 + y_2 + ... + y_{n-1} + \frac{1}{2}y_n}{n-1} \qquad (12.3)$$

(n-1: số các khoảng cách thời gian)

- *Khoảng cách thời gian giữa các thời điểm không bằng nhau& thời gian nghiên cứu là liên tục*

---

[1] Cyclical component
[2] Irregular component, random component
[3] Multiplicative structure

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i} \qquad (12.4)$$

trong đó: y_i: mức độ thứ i trong dãy số

t_i : độ dài thời gian tương ứng với mức độ thứ I

12.3.2 Lượng tăng (giảm) tuyệt đối: là chỉ tiêu biểu hiện sự thay đổi về giá trị tuyệt đối của hiện tượng 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 dãy số

$$\delta_i = y_i - y_{i-1} \qquad (i = 2, \ldots, n) \qquad (12.5)$$

- *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 dãy số)

$$\Delta_i = y_i - y_1 \qquad (i = 2, \ldots, n) \qquad (12.6)$$

Giữa lượng tăng (giảm) tuyệt đối liên hoàn và lượng tăng (giảm) tuyệt đối định gốc có mối liên hệ sau:

$$\sum_{i=2}^{n} \delta_i = \Delta_n \qquad (12.7)$$

- *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.

$$\bar{\delta} = \frac{\sum_{i=2}^{n} \delta_i}{n-1} = \frac{\Delta_n}{n-1} = \frac{y_n - y_1}{n-1} \qquad (12.8)$$

Chỉ tiêu 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.

12.3.3 Tốc độ phát triển: là chỉ tiêu tương đối động thái (phát triển) dùng để đánh giá hiện tượng 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 hiện tượng giữa 2 kỳ liền nhau

$$t_i = \frac{y_i}{y_{i-1}} \qquad (i = 2, 3, \ldots, n) \qquad (12.9)$$

- *Tốc độ phát triển định gốc*: thể hiện tốc độ phát triển của hiện tượng giữa kỳ nghiên cứu với kỳ được chọn làm gốc so sánh

$$T_i = \frac{y_i}{y_1} \qquad (i = 2, 3, \ldots, n) \qquad (12.10)$$

Giữa tốc độ phát triển liên hoàn và tốc độ phát triển định gốc có các mối liên hệ sau:

Tích các tốc độ phát triển liên hoàn bằng tốc độ phát triển định gốc

$$\prod_{i=2}^{n} t_i = T_n \qquad (12.11)$$

Tỉ số giữa hai tốc độ phát triển định gốc liền nhau trong dãy số bằng tốc độ phát triển liên hoàn

$$\frac{T_i}{T_{i-1}} = t_i \qquad (12.12)$$

- *Tốc độ phát triển trung bình*

Là chỉ tiêu thể hiện nhịp độ phát triển đại diện của hiện tượng trong suốt thời kỳ nghiên cứu

$$\bar{t} = \sqrt[n-1]{\prod_{i=2}^{n} t_i} = \sqrt[n-1]{\frac{y_n}{y_1}} \qquad (12.13)$$

Chỉ tiêu này chỉ có ý nghĩa khi các tốc độ phát triển liên hoàn xấp xỉ nhau nghĩa là trong suốt thời kỳ nghiên cứu hiện tượng phát triển với một tốc độ tương đối đều.

12.3.4 Tốc độ tăng (giảm):

Là chỉ tiêu phản ánh mức độ của hiện tượng giữa 2 thời gian nghiên cứu đã tăng (giảm) bao nhiêu lần (%)

- Tốc độ tăng (giảm) liên hoàn

$$a_i = \frac{y_i - y_{i-1}}{y_{i-1}} = \frac{\delta_i}{y_{i-1}} = t_i - 1 \quad (i = 2,3, \ldots, n) \qquad (12.14)$$

- Tốc độ tăng (giảm) định gốc

$$A_i = \frac{y_i - y_1}{y_1} = \frac{\Delta_i}{y_1} = T_i - 1 \quad (i = 2,3, \ldots, n) \qquad (12.15)$$

- Tốc độ tăng (giảm) trung bình

$$\bar{a} = \bar{t} - 1 \qquad (12.16)$$

12.3.5 Giá trị tuyệt đối của 1% tăng (giảm) liên hoàn: phản ánh 1% tăng (giảm) của 2 thời kỳ đứng liền nhau của hiện tượng nghiên cứu tương ứng với một lượng giá trị tuyệt đối là bao nhiêu.

$$g_i = \frac{\delta_i}{a_i(\%)} = \frac{y_i - y_{i-1}}{\frac{y_i - y_{i-1}}{y_{i-1}} 100} = \frac{y_{i-1}}{100} \qquad (12.17)$$

12.4 CÁC PHƯƠNG PHÁP BIỂU HIỆN XU HƯỚNG BIẾN ĐỘNG CỦA DÃY SỐ THỜI GIAN

12.4.1 Phương pháp số trung bình di động (Số bình quân trượt) [1]

Trong thực tế, có một số biến động qua thời gian của hiện tượng tỏ ra bất thường, không thể hiện xu hướng một cách rõ rệt do ảnh hưởng của các yếu tố ngẫu nhiên quá lớn. Để giảm bớt hoặc triệt tiêu các ảnh hưởng ngẫu nhiên này, vạch rõ xu thế phát triển cơ bản của hiện tượng, ta có thể sử dụng phương pháp số trung bình di động. Đây là phương pháp thể hiện xu hướng đơn giản, được xây dựng trên cơ sở cho rằng ảnh hưởng của yếu tố ngẫu nhiên ở một thời điểm nào đó sẽ bị hạn chế, loại trừ nếu giá trị quan sát ở thời điểm đó được tính trung bình với các giá trị quan sát lân cận. Phương pháp số trung bình di động làm cho dãy số thực tế trở nên "bằng phẳng"[2] hơn.

Số trung bình di động là số trung bình cộng của một nhóm nhất định các mức độ của dãy số thời gian, được tính bằng cách lần lượt loại trừ dần các mức độ đầu, đồng thời thêm vào các mức độ tiếp theo, sao cho tổng số các mức độ tham giá tính số trung bình cộng không thay đổi.

Giả sử có dãy số thời gian: $y_1, y_2, \ldots, y_n$

Nếu tính số trung bình di động $\bar{y}_t$ ứng với thời điểm t, và tính với nhóm (2m+1) mức độ

[1] Moving averages method

[2] Smooth

$$\bar{y}_t = \frac{y_{t-m} + y_{t-m+1} + ... + y_t + ... + y_{t+m-1} + y_{t+m}}{2m+1}$$

$$= \frac{1}{2m+1}\sum_{i=-m}^{m} y_{t+i} \qquad (t = m+1, m+2, \ldots, n-m) \qquad (12.18)$$

Kết quả của dãy số trung bình di động luôn có ít hơn dãy số ban đầu 2m số hạng (m số hạng đầu và m số hạng cuối).

Việc lựa chọn nhóm bao nhiêu mức độ để tính số trung bình di động đòi hỏi phải dựa vào đặc điểm biến động của hiện tượng và số lượng các mức độ của dãy số thời gian. Nếu biến động của hiện tượng không lớn và số mức độ của dãy số không nhiều thì số trung bình di động có thể tính từ một nhóm 3 mức độ. Nếu biến động của hiện tượng khá lớn và dãy số có nhiều mức độ thì số trung bình di động nên được tính với nhiều mức độ hơn (5, 7, . . . mức độ)

Nếu biến động của hiện tượng mang tính chu kỳ thì nên chọn thời kỳ tính số trung bình di động bằng với độ dài thời gian (hay bội số) của chu kỳ. Ví dụ nếu chu kỳ biến động là 3 năm thì có thể tính số trung bình di động với 3 hay 6 mức độ. Đối với dãy số biến động thời vụ tháng hay quí thì nên tính số trung bình di động từ nhóm 12 (tháng) hay 4 (quí) mức độ.

Cần lưu ý là nếu tính số trung bình di động từ một nhóm có ít mức độ thì ảnh hưởng ngẫu nhiên sẽ ít bị loại trừ, lúc đó ta sẽ có nhiều số trung bình di động, tạo điều kiện dễ dàng đánh giá xu hướng phát triển của hiện tượng. Và ngược lại nếu số trung bình di động được tính ra từ một nhóm có khá nhiều mức độ thì khả năng hạn chế, loại bỏ ảnh hưởng ngẫu nhiên sẽ lớn hơn, song số lượng số trung bình di động tính được sẽ ít đi và như vậy sẽ khó khăn trong việc đánh giá xu hướng phát triển của hiện tượng.

Ví dụ 3: có số liệu doanh số thực tế của một công ty trong khoảng thời gian 15 năm như dưới đây, tính doanh số trung bình di động (với 5 mức độ)

Thời gian	doanh số (triệu đồng)	doanh số trung bình di động (triệu đồng)
1	1806	-
2	1644	-
3	1814	1710,4
4	1770	1569,8
5	1518	1494,2
6	1103	1426,0
7	1266	1356,6

8	1473	1406,4
9	1423	1618,0
10	1767	183,0
11	2161	2057,8
12	2336	2276,8
13	2602	2450,8
14	2518	-
15	2637	-

Số trung bình di động với 5 mức độ có thể tính lần lượt như sau:

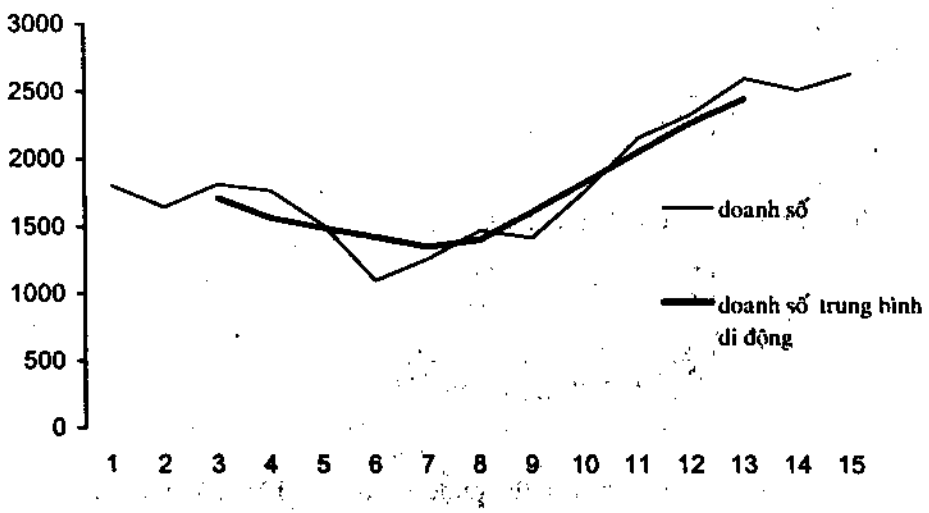
$$\frac{1806+1644+1814+1770+1518}{5}=1710,4 \text{ triệu đồng}$$

$$\frac{1644+1814+1770+1518+1103}{5}=1569,8 \text{ triệu đồng}$$

và cứ tiếp tục cho đến hết dãy số.

Trên Excel, ta có thể tính nhanh hơn bằng cách vào **Tools – Data Analysis – Moving Average.**

Trong thực tế, người ta thường tính các số trung bình di động với một số lẻ các mức độ để con số trung bình di động tính ra được đặt vào vị trí giữa của khoảng cách san bằng.Trong trường hợp số mức độ là số lẻ, người ta thường tính số trung bình di động 2 lần.



Hình 12.1: đồ thị biến động của doanh số thực tế và doanh số trung bình di động.

12.4.2 Phương pháp thể hiện xu hướng bằng hàm số (phương pháp hàm xu thế)

Tiến trình thể hiện xu hướng bằng phương pháp hàm số được thực hiện qua 3 bước: nhận ra mô hình, lựa chọn mô hình và điều chỉnh mô hình.

Nội dung cơ bản của phương pháp hàm xu thế là khái quát hóa chiều hướng biến động của hiện tượng nghiên cứu bằng một hàm số toán học, nhằm mô tả một cách sát nhất, gần đúng nhất biến động thực tế của hiện tượng. Cụ thể là, thông qua việc xem xét đồ thị biến động thực tế của hiện tượng kết hợp với kinh nghiệm, sự hiểu biết thực tế về hiện tượng, ta chọn một hàm số có tính chất lý thuyết để thể hiện một cách tốt nhất xu hướng phát triển của hiện tượng.

Dưới đây là một số hàm thường được sử dụng:

12.4.2.1 Hàm số tuyến tính (phương trình đường thẳng) [1]

Phương trình đường thẳng thường được sử dụng khi hiện tượng biến động với một lượng tăng (giảm) tuyệt đối liên hoàn tương đối đều đặn.

Hàm số có dạng: $$\hat{y}_t = a_0 + a_1 t \qquad (12.19)$$

$\hat{y}_t$: giá trị của hiện tượng tại thời gian t xác định bằng hàm số tuyến tính

t : thứ tự thời gian (t = 1,2, . . . , n)

a_0, a_1 các tham số qui định vị trí của đường thẳng

Theo phương pháp bình phương bé nhất, $\hat{y}_t$ là " thích hợp nhất" đối với dãy số thực tế khi:

$$\sum_{i=1}^{n}(y_i - \hat{y}_{t_i})^2 = \sum_{i=1}^{n}(y_i - a_0 - a_1 t_i)^2 = \min \qquad (12.20)$$

Từ điều kiện này ta có hệ hai phương trình:

$$\begin{cases} \sum_{i=1}^{n} y_i = na_0 + a_1 \sum_{i=1}^{n} t_i \\ \sum_{i=1}^{n} y_i t_i = a_0 \sum_{i=1}^{n} t_i + a_1 \sum_{i=1}^{n} t_i^2 \end{cases} \qquad (12.21)$$

Giải hệ hai phương trình này ta tìm được a_0, a_1 cho hàm số tuyến tính.

[1] Linear trend model

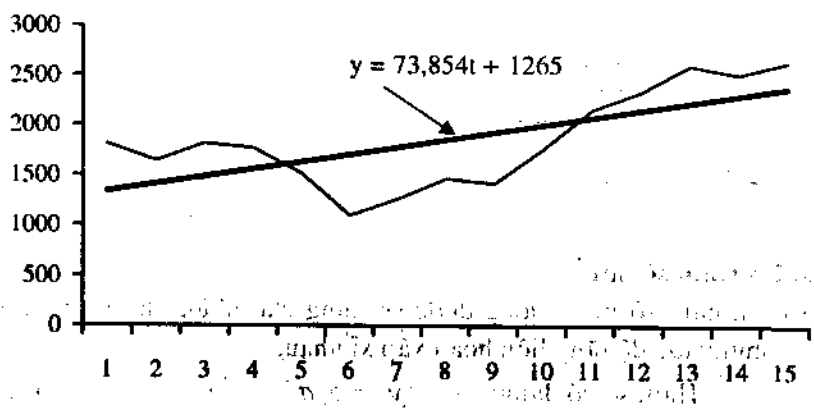
Trong thực tế, do t là thứ tự thời gian trong dãy số , nên ta có thể tìm a_0, a_1 đơn giản hơn bằng cách đánh số thứ tự sao cho $\sum t_i = 0$

- Nếu thứ tự thời gian là một số lẻ, thì lấy thời gian ở vị trí giữa bằng 0, các thời gian đứng trước lần lượt là –1, -2, -3, . . . và thời gian đứng sau lần lượt là 1, 2, 3, ...
- Nếu thứ tự thời gian là một số chẵn, thì lấy hai thời gian ở vị trí giữa bằng –1 và 1, các thời gian đứng trước lần lượt là –3, -5, -7, . . . và thời gian đứng sau lần lượt là 3, 5, 7, ...

Khi đó hệ phương trình (12.21) trở nên đơn giản hơn:

$$\begin{cases} \sum_{i=1}^{n} y_i = na_0 \\ \sum_{i=1}^{n} y_i t_i = a_1 \sum_{i=1}^{n} t_i^2 \end{cases} \qquad (12.22)$$

Ví dụ 4: sử dụng lại số liệu ở ví dụ 3 nhưng dùng phương pháp đường thẳng



Hình 12.2: đồ thị biến động thực tế và hàm xu thế tuyến tính

12.4.2.2 Hàm số bậc 2 (phương trình parabôn bậc 2)[1]

Phương trìnhparabôn bậc 2 thường được sử dụng khi hiện tượng tăng hoặc giảm với tốc độ phát triển liên hoàn xấp xỉ bằng nhau.

Hàm số có dạng: $\hat{y}_t = a_0 + a_1 t + a_2 t^2$ (12.23)

các tham số a_0, a_1, a_2 có thể được xác định thông qua hệ phương trình:

$$\begin{cases} \sum_{i=1}^{n} y_i = na_0 + a_1 \sum_{i=1}^{n} t_i + a_2 \sum_{i=1}^{n} t_i^2 \\ \sum_{i=1}^{n} y_i t_i = a_0 \sum_{i=1}^{n} t_i + a_1 \sum_{i=1}^{n} t_i^2 + a_2 \sum_{i=1}^{n} t_i^3 \\ \sum_{i=1}^{n} y_i t_i^2 = a_0 \sum_{i=1}^{n} t_i^2 + a_1 \sum_{i=1}^{n} t_i^3 + a_2 \sum_{i=1}^{n} t_i^4 \end{cases} \quad (12.24)$$

Giải hệ phương trình trên ta tìm được a_0, a_1, a_2. vì t là thứ tự thời gian nên ta cũng có thể tìm a_0, a_1, a_2 nhanh chóng bằng cách đánh số thứ tự sao cho $\sum t_i = 0$ và vì vậy $\sum t_i^3 = 0$ và hệ phương trình trên trở nên đơn giản hơn để xác định a_0, a_1, a_2

$$\begin{cases} \sum_{i=1}^{n} y_i = na_0 + a_2 \sum_{i=1}^{n} t_i^2 \\ \sum_{i=1}^{n} y_i t_i = a_1 \sum_{i=1}^{n} t_i^2 \\ \sum_{i=1}^{n} y_i t_i^2 = a_0 \sum_{i=1}^{n} t_i^2 + a_2 \sum_{i=1}^{n} t_i^4 \end{cases} \quad (12.25)$$

12.4.2.3 Hàm số mũ[2]

Phương trình hàm số mũ thường được sử dụng cho những hiện tượng biến động với những tốc độ tăng liên hoàn xấp xỉ nhau.

Hàm số có dạng: $\hat{y}_t = a_0 a_1^t$ (12.26)

a_0: điểm gốc của phương trình hồi qui

a_1: tốc độ phát triển trung bình theo đơn vị thời gian (số lần)

Lấy logarit cho cả 2 vế của phương trình, ta có hệ phương trình:

---

[1] Quadratic model, second degree polynimial

[2] Exponential trend

$$\begin{cases} n\lg a_0 + \lg a_1 \sum_{i=1}^{n} t_i = \sum_{i=1}^{n} \lg y_i \\ \lg a_0 \sum_{i=1}^{n} t_i + \lg a_1 \sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} t_i \lg y_i \end{cases} \quad (12.27)$$

Để đơn giản ta cũng có thể đánh số thứ tự sao cho $\sum_{i=1}^{n} t_i = 0$

Trong thực tế, ta còn có thể xây dựng nhiều dạng hàm số khác như hàm bậc 3, hàm hy pec bôn, . . . tùy theo đặc điểm và tính chất biến động của hiện tượng.

12.5 PHÂN TÍCH BIẾN ĐỘNG CÁC THÀNH PHẦN CỦA DÃY SỐ THỜI GIAN

Với dãy số thời gian có quan hệ theo mô hình nhân: y = T.S.C.I, ta có thể xem xét biến động của từng yếu tố thành phần.

12.5.1 Biến động thời vụ :

Do số trung bình di động có tác dụng hạn chế, loại bỏ các biến động mang tính ngẫu nhiên, vì vậy nó thường được áp dụng trong việc tính toán các chỉ số thời vụ (CSTV) nhằm thể hiện biến động thời vụ của các dãy số thời gian.

Đầu tiên, tính số trung bình di động từ một nhóm 4 mức độ (nếu số liệu là các quí), hoặc 12 mức độ (nếu số liệu là hàng tháng). Dãy số thời gian tính được sẽ chỉ bao hàm 2 thành phần là xu hướng và chu kỳ vì 2 thành phần kia (thời vụ và ngẫu nhiên) đã bị loại bỏ nhờ cách tính số trung bình di động. Từ đây, ta sẽ tính ảnh hưởng của 2 yếu tố thành phần: thời vụ và ngẫu nhiên.

$$SI = \frac{TCSI}{TC} = \frac{y_t}{\bar{y}_t} \quad (12.28)$$

y_t : giá trị quan sát ở thời gian t

$\bar{y}_t$: số trung bình di động ứng với giá trị quan sát ở thời gian t

Ví dụ 5: có số liệu về doanh số (triệu đồng) giai đoạn 1992 – 2002, để xem xét tính thời vụ, ta tính số trung bình di động hai lần (4 mức độ và 2 mức độ, tính hai lần TBDĐ nhằm có sự tương ứng giữa doanh số thực tế và doanh số trung bình di động, thuận tiện trong so sánh)

năm	quí	t	y_t (TSCI)	TBDĐ 4 mức độ	TBDĐ 2 mức độ (y_t^*) (TC)	(SI=TSCI/TC)
(1)	(2)	(3)	(4)	(5)	(6)	(7) =(4)/(6) *100%
1992	Q1	1	2513			
	Q2	2	2142			
	Q3	3	2131	2245,25	2228,00	95,65
	Q4	4	2195	2210,75	2219,13	98,91
1993	Q1	5	2375	2227,50	2235,50	106,24
	Q2	6	2209	2243,50	2188,75	100,93
	Q3	7	2195	2134,00	2080,13	105,52
	Q4	8	1757	2026,25	2011,88	87,33
1994	Q1	9	1944	1997,50	1961,88	99,09
	Q2	10	2094	1926,25	1960,50	106,81
	Q3	11	1910	1994,75	2007,38	95,15
	Q4	12	2031	2020,00	2071,13	98,06
1995	Q1	13	2045	2122,25	2163,25	94,53
	Q2	14	2503	2204,25	2249,75	111,26
	Q3	15	2238	2295,25	2362,50	94,73
	Q4	16	2395	2429,75	2479,50	96,59
1996	Q1	17	2583	2529,25	2555,50	101,08
	Q2	18	2901	2581,75	2590,00	112,01
	Q3	19	2448	2598,25	2605,88	93,94
	Q4	20	2461	2613,50	2624,38	93,77
1997	Q1	21	2644	2635,25	2700,25	97,92
	Q2	22	2988	2765,25	2762,38	108,17
	Q3	23	2968	2759,50	2753,88	107,78
	Q4	24	2438	2748,25	2755,38	88,48
1998	Q1	25	2599	2762,50	2781,88	93,43
	Q2	26	3045	2801,25	2831,63	107,54
	Q3	27	3123	2862,00	2830,00	110,35
	Q4	28	2681	2798,00	2763,38	97,02
1999	Q1	29	2343	2728,75	2686,63	87,21
	Q2	30	2768	2644,50	2607,13	106,17
	Q3	31	2786	2569,75	2609,25	106,77
	Q4	32	2382	2648,75	2663,38	89,44
2000	Q1	33	2659	2678,00	2652,13	100,26
	Q2	34	2885	2626,25	2643,00	109,16
	Q3	35	2579	2659,75	2622,75	98,33
	Q4	36	2516	2585,75	2573,13	97,78
2001	Q1	37	2363	2560,50	2571,25	91,90
	Q2	38	2784	2582,00	2528,88	110,09

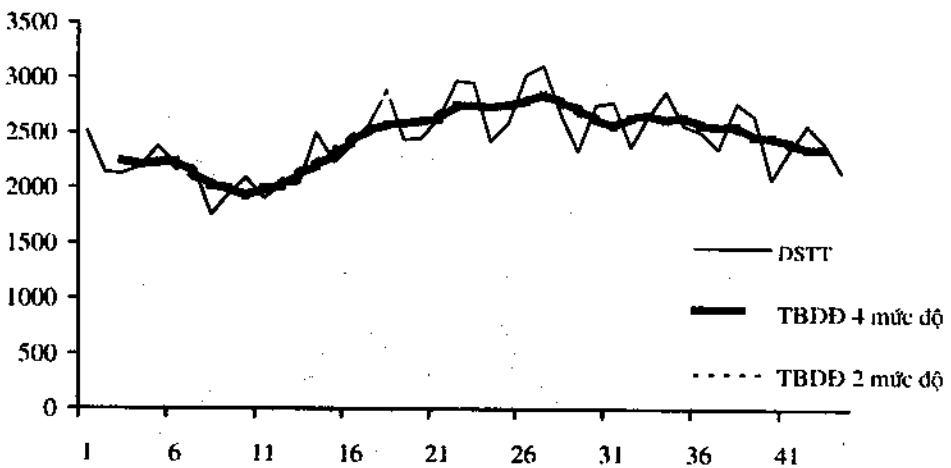
	Q3	39	2665	2475,75	2474,00	107,72
	Q4	40	2091	2472,25	2447,00	85,45
2002	Q1	41	2349	2421,75	2390,25	98,27
	Q2	42	2582	2358,75	2366,50	109,11
	Q3	43	2413	2374,25		
	Q4	44	2153			

Doanh số TBDĐ 4 mức độ được tính như sau:

$\frac{2513 + 2142 + 2131 + 2195}{4} = 2245,25$ và tiếp tục cho đến hết các mức độ của dãy số thực tế.

Doanh số TBDĐ 2 mức độ được tính từ kết quả của dãy số TBDĐ 4 mức độ $\frac{2245,25 + 2210,75}{2} = 2228,0$ và tiếp tục cho đến hết các mức độ của dãy số TBDĐ 4 mức độ.

Kế tiếp, ta loại bỏ yếu tố ngẫu nhiên bằng cách tính CSTV trung bình quí (nếu số liệu hàng quí) hoặc CSTV trung bình tháng (nếu số liệu hàng tháng).



Hình 12.3: Doanh số thực tế và doanh số TBDĐ 4 mức độ và 2 mức độ

Sau cùng, cần điều chỉnh các CSTV quí (hoặc tháng) sao cho trung bình của chúng bằng 100.

Bảng 12.1: chỉ số thời vụ trung bình và chỉ số thời vụ điều chỉnh

năm	quí			
	I	II	III	IV
1992			95,65	98,91
1993	106,24	100,93	105,52	87,33
1994	99,09	106,81	95,15	98,06
1995	94,53	111,26	94,73	96,59
1996	101,08	112,01	93,94	93,77
1997	97,92	108,17	107,78	88,48
1998	93,43	107,54	110,35	97,02
1999	87,21	106,17	106,77	89,44
2000	100,26	109,16	98,33	97,78
2001	91,9	110,09	107,72	85,45
2002	98,27	109,11		
cộng	969,93	1081,25	1015,94	932,83
Chỉ số thời vụ trung bình (%)	96,99	108,12	101,59	93,28
Chỉ số thời vụ điều chỉnh (%) (I_s)	97	108	102	93

Tổng cộng chỉ số thời vụ trung bình:

96,99 + 108,12 + 101,59 +93,28 = 399,98

Chỉ số thời vụ điều chỉnh = (chỉ số thời vụ trung bình × 400) / 399,98

12.5.2 Biến động xu hướng

Muốn thể hiện xu hướng của dãy số thời gian có tính thời vụ, ta cần phải loại bỏ yếu tố thời vụ ra khỏi dãy số.

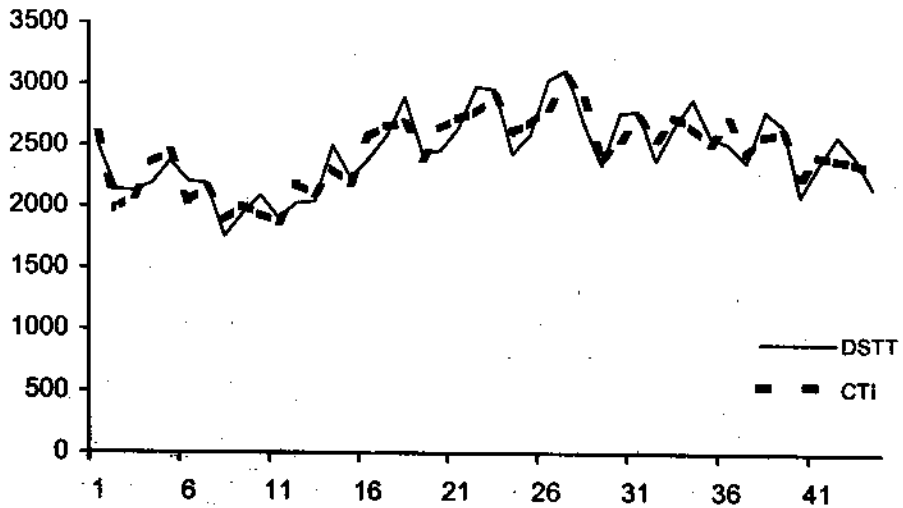
Để loại bỏ biến động thời vụ, ta chia các giá trị thực tế của dãy số cho các CSTV tương ứng.

$$CTI = \frac{CTSI}{S} = \frac{y_t}{I_S} \qquad (12.29)$$

Ví dụ 6: nghiên cứu biến động thời vụ cho số liệu dãy số thực tế (DSTT) ở ví dụ 5

năm	quí	t	y_t	CSTV điều chỉnh I_S	DSTT đã loại bỏ yếu tố thời vụ CTI
(1)	(2)	(3)	(4)	(5)	(6) =[(4)/(5)] *100
1992	Q1	1	2513	97	2590,72
	Q2	2	2142	108	1983,33
	Q3	3	2131	102	2089,22
	Q4	4	2195	93	2360,22
1993	Q1	5	2375	97	2448,45
	Q2	6	2209	108	2045,37
	Q3	7	2195	102	2151,96
	Q4	8	1757	93	1889,25
1994	Q1	9	1944	97	2004,12
	Q2	10	2094	108	1938,89
	Q3	11	1910	102	1872,55
	Q4	12	2031	93	2183,87
1995	Q1	13	2045	97	2108,25
	Q2	14	2503	108	2317,59
	Q3	15	2238	102	2194,12
	Q4	16	2395	93	2575,27
1996	Q1	17	2583	97	2662,89
	Q2	18	2901	108	2686,11
	Q3	19	2448	102	2400,00
	Q4	20	2461	93	2646,24
1997	Q1	21	2644	97	2725,77
	Q2	22	2988	108	2766,67
	Q3	23	2968	102	2909,80
	Q4	24	2438	93	2621,51
1998	Q1	25	2599	97	2679,38
	Q2	26	3045	108	2819,44

	Q3	27	3123	102	3061,76
	Q4	28	2681	93	2882,80
1999	Q1	29	2343	97	2415,46
	Q2	30	2768	108	2562,96
	Q3	31	2786	102	2731,37
	Q4	32	2382	93	2561,29
2000	Q1	33	2659	97	2741,24
	Q2	34	2885	108	2671,30
	Q3	35	2579	102	2528,43
	Q4	36	2516	93	2705,38
2001	Q1	37	2363	97	2436,08
	Q2	38	2784	10	2577,78
	Q3	39	2665	102	2612,75
	Q4	40	2091	93	2248,39
2002	Q1	41	2349	97	2421,65
	Q2	42	2582	108	2390,74
	Q3	43	2413	102	2365,69
	Q4	44	2153	93	2590,72

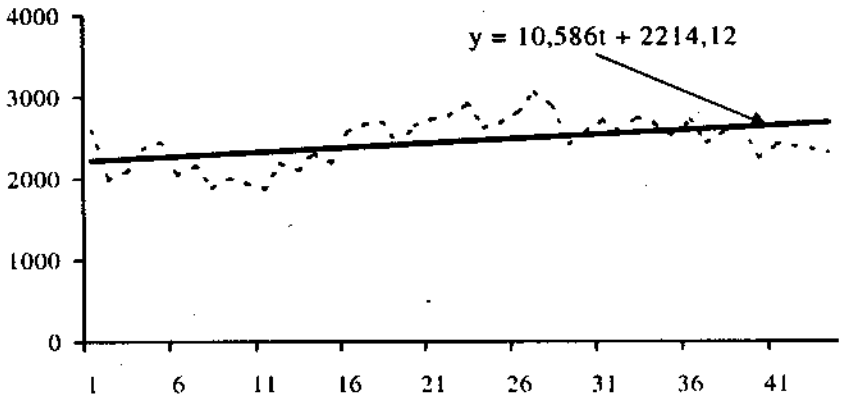


Hình 12.4: đồ thị doanh số thực tế và doanh số đã loại bỏ yếu tố thời vụ.

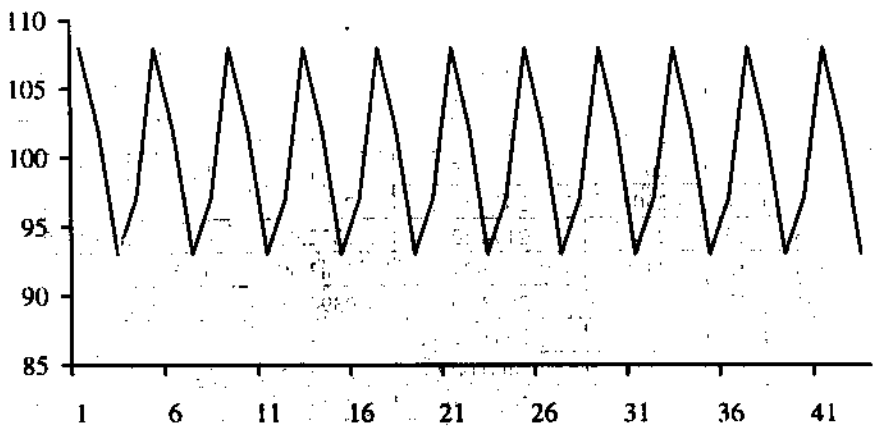
Do vẫn còn bị ảnh hưởng bởi yếu tố biến động ngẫu nhiên nên đường biểu diễn ở Hình 12.4 "gồ ghề" hơn so với đường biểu diễn ở Hình 12.3.

Sau khi đã loại bỏ yếu tố thời vụ, dùng phương pháp hàm xu thế tuyến tính để thể hiện một cách tốt nhất xu hướng biến động của hiện tượng. Trên Excel, dùng công cụ **Trendline** dễ dàng tìm ra phương trình đường thẳng:

y = 10,586t +2214,1. Như vậy, bên cạnh những dao động có tính thời vụ, trung bình mỗi quí doanh số tăng 10,586 triệu đồng



Hình 12.5: doanh số đã loại bỏ yếu tố thời vụ và hàm tuyến tính thể hiện xu hướng



Hình 12.6: chỉ số thời vụ

12.5.3 Biến động chu kỳ

Được thể hiện qua chỉ số chu kỳ, để tính chỉ số chu kỳ trước hết ta chia các giá trị của dãy số đã loại bỏ biến động thời vụ cho các giá trị của thành phần xu hướng tương ứng nhằm loại bỏ yếu tố xu hướng, nghĩa là:

$$CI = \frac{CTI}{T} \qquad (12.30)$$

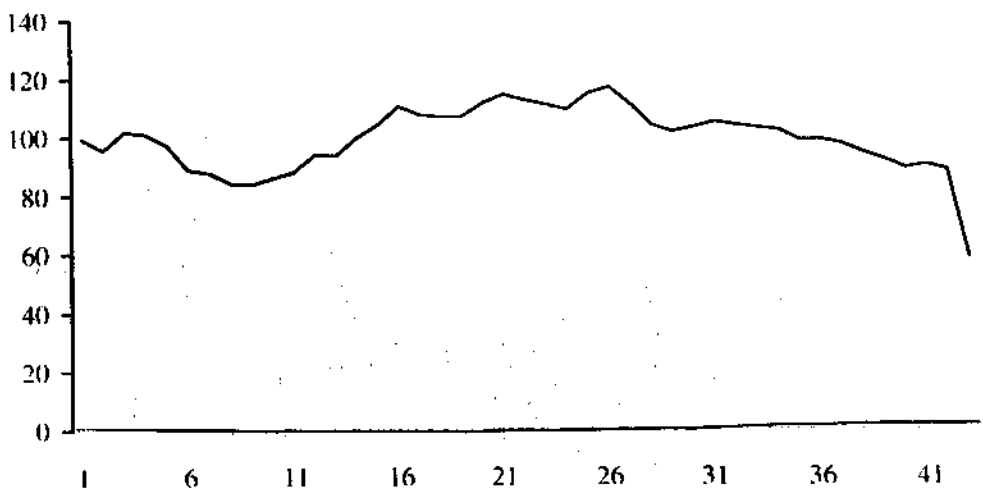
Sau đó loại bỏ yếu tố thành phần ngẫu nhiên (I) bằng cách dùng các chỉ số chu kỳ theo cách tính số trung bình di động. Ví dụ như tính :

$$I_{C_3} = \frac{1,16 + 0,89 + 0,93}{3} 100 = 99,40\%, ...$$

Bảng 12.2: tính chỉ số chu kỳ

năm	qui	t	y_t (TSCI)	DSTT đã loại bỏ yếu tố thời vụ (CTI)	PT đường thẳng (xu hướng) (T)	Biến động chu kỳ và ngẫu nhiên (CI)	Chỉ số chu kỳ I_c (%) (TBDĐ 3 mức độ)
(1)	(2)	(3)	(4)	(5)	(6)	(7) =(5)/(6)	(8)
1992	Q1	1	2513	2590,72	2224,71	1,16	
	Q2	2	2142	1983,33	2235,29	0,89	99,40
	Q3	3	2131	2089,22	2245,88	0,93	95,45
	Q4	4	2195	2360,22	2256,46	1,05	101,88
1993	Q1	5	2375	2448,45	2267,05	1,08	100,80
	Q2	6	2209	2045,37	2277,64	0,90	97,28
	Q3	7	2195	2151,96	2288,22	0,94	88,68
	Q4	8	1757	1889,25	2298,81	0,82	87,67
1994	Q1	9	1944	2004,12	2309,39	0,87	84,18
	Q2	10	2094	1938,89	2319,98	0,84	83,57
	Q3	11	1910	1872,55	2330,57	0,80	85,74
	Q4	12	2031	2183,87	2341,15	0,93	87,76
1995	Q1	13	2045	2108,25	2351,74	0,90	93,68
	Q2	14	2503	2317,59	2362,32	0,98	93,41
	Q3	15	2238	2194,12	2372,91	0,92	99,54
	Q4	16	2395	2575,27	2383,50	1,08	103,91
1996	Q1	17	2583	2662,89	2394,08	1,11	110,33
	Q2	18	2901	2686,11	2404,67	1,12	107,43
	Q3	19	2448	2400,00	2415,25	0,99	106,72
	Q4	20	2461	2646,24	2425,84	1,09	106,78
1997	Q1	21	2644	2725,77	2436,43	1,12	111,34
	Q2	22	2988	2766,67	2447,01	1,13	114,45
	Q3	23	2968	2909,80	2457,60	1,18	112,56
	Q4	24	2438	2621,51	2468,18	1,06	110,90

1998	Q1	25	2599	2679,38	2478,77	1,08	109,19
	Q2	26	3045	2819,44	2489,36	1,13	114,61
	Q3	27	3123	3061,76	2499,94	1,22	116,85
	Q4	28	2681	2882,80	2510,53	1,15	111,04
1999	Q1	29	2343	2415,46	2521,11	0,96	103,96
	Q2	30	2768	2562,96	2531,70	1,01	101,49
	Q3	31	2786	2731,37	2542,29	1,07	103,00
	Q4	32	2382	2561,29	2552,87	1,00	104,90
2000	Q1	33	2659	2741,24	2563,46	1,07	103,68
	Q2	34	2885	2671,30	2574,04	1,04	102,85
	Q3	35	2579	2528,43	2584,63	0,98	101,95
	Q4	36	2516	2705,38	2595,22	1,04	98,52
2001	Q1	37	2363	2436,08	2605,80	0,93	98,75
	Q2	38	2784	2577,78	2616,39	0,99	97,16
	Q3	39	2665	2612,75	2626,97	0,99	94,41
	Q4	40	2091	2248,39	2637,56	0,85	92,05
2002	Q1	41	2349	2421,65	2648,15	0,91	88,87
	Q2	42	2582	2390,74	2658,73	0,90	90,00
	Q3	43	2413	2365,69	2669,32	0,89	88,31
	Q4	44	2153	2590,72	2679,90	0,86	



Hình 12.7: Chỉ số chu kỳ.

Cần lưu ý là, không giống như biến động thời vụ xảy ra tương đối đều đặn với chu kỳ 1 năm, biến động chu kỳ khá phức tạp, đôi khi thất thường cả về độ lớn lẫn chu kỳ của biến động, vì vậy việc dự đoán rất khó khăn. Và đây là trở ngại cho việc dự đoán mức độ tương lai của hiện tượng.

12.5.4 Biến động ngẫu nhiên

Sau khi đã có được các yếu tố thành phần T, S, C, ta có thể xác định biến động của yếu tố ngẫu nhiên bằng cách tính:

$$I_i = \frac{y_t}{T.I_S.I_C} \qquad (12.31)$$

I_i : chỉ số thể hiện yếu tố ngẫu nhiên

y_t : giá trị thực tế của hiện tượng

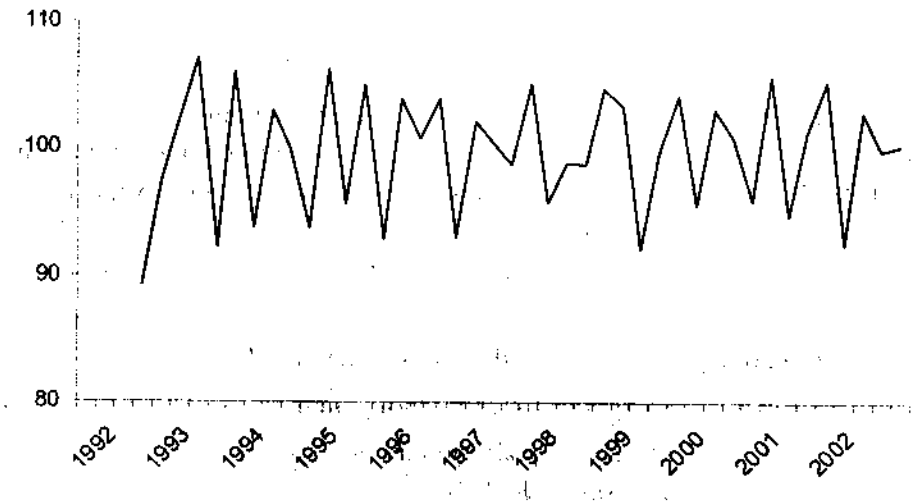
I_S : chỉ số thời vụ

I_C : chỉ số chu kỳ

Bảng 12.3: Tính chỉ số ngẫu nhiên

năm	quí	t	y_t (TSCI)	CSTV điều chỉnh (%) (I_S)	PT đường thẳng (xu hướng) (T)	Chỉ số chu kỳ (%) (I_C)	Chỉ số biến động ngẫu nhiên (I_i) (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8) = (4)/(5)(6)(7)
1992	Q1	1	2513	97	2224,71		
	Q2	2	2142	108	2235,29	99,40	89,26
	Q3	3	2131	102	2245,88	95,45	97,46
	Q4	4	2195	93	2256,46	101,88	102,67
1993	Q1	5	2375	97	2267,05	100,80	107,14
	Q2	6	2209	108	2277,64	97,28	92,31
	Q3	7	2195	102	2288,22	88,68	106,05
	Q4	8	1757	93	2298,81	87,67	93,74
1994	Q1	9	1944	97	2309,39	84,18	103,09
	Q2	10	2094	108	2319,98	83,57	100,01
	Q3	11	1910	102	2330,57	85,74	93,72
	Q4	12	2031	93	2341,15	87,76	106,29
1995	Q1	13	2045	97	2351,74	93,68	95,70
	Q2	14	2503	108	2362,32	93,41	105,03
	Q3	15	2238	102	2372,91	99,54	92,89
	Q4	16	2395	93	2383,50	103,91	103,98
1996	Q1	17	2583	97	2394,08	110,33	100,82
	Q2	18	2901	108	2404,67	107,43	103,98
	Q3	19	2448	102	2415,25	106,72	93,11
	Q4	20	2461	93	2425,84	106,78	102,16
1997	Q1	21	2644	97	2436,43	111,34	100,48
	Q2	22	2988	108	2447,01	114,45	98,79

	Q3	23	2968	102	2457,60	112,56	105,19
	Q4	24	2438	93	2468,18	110,90	95,77
1998	Q1	25	2599	97	2478,77	109,19	99,00
	Q2	26	3045	108	2489,36	114,61	98,82
	Q3	27	3123	102	2499,94	116,85	104,81
	Q4	28	2681	93	2510,53	111,04	103,41
1999	Q1	29	2343	97	2521,11	103,96	92,16
	Q2	30	2768	108	2531,70	101,49	99,74
	Q3	31	2786	102	2542,29	103,00	104,31
	Q4	32	2382	93	2552,87	104,90	95,64
2000	Q1	33	2659	97	2563,46	103,68	103,14
	Q2	34	2885	108	2574,04	102,85	100,91
	Q3	35	2579	102	2584,63	101,95	95,95
	Q4	36	2516	93	2595,22	98,52	105,81
2001	Q1	37	2363	97	2605,80	98,75	94,67
	Q2	38	2784	108	2616,39	97,16	101,41
	Q3	39	2665	102	2626,97	94,41	105,35
	Q4	40	2091	93	2637,56	92,05	92,61
2002	Q1	41	2349	97	2648,15	88,87	102,90
	Q2	42	2582	108	2658,73	90,00	99,91
	Q3	43	2413	102	2669,32	88,31	100,36
	Q4	44	2153	93	2679,90		



Hình 12.8: Chỉ số biến động ngẫu nhiên

Với kết quả tính toán từ Bảng 12.3 và Hình 12.8, ta có thể kết luận:

- Doanh số có xu hướng tăng rõ rệt trong thời kỳ 1992-2002, trung bình mỗi quí tăng 10,586 triệu đồng.

- Doanh số tăng cao vào quí 2 và quí 3, giảm thấp vào quí 1 và quí 4.
- Trong thời kỳ 1992-2002, doanh số tăng cao nhất vào quí 1 năm 1993 và giảm thấp nhất vào quí 2 năm 1992.

12.6 DỰ ĐOÁN BIẾN ĐỘNG CỦA DÃY SỐ THỜI GIAN

Giả sử có dãy số thời gian : $y_1, y_2, ..., y_n$. Với dãy số này ta có thể dùng các phương pháp dự đoán ngắn hạn sau:

12.6.1 Dự đoán dựa vào 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.

$$\hat{y}_{n+L} = y_n + \bar{\delta}L \qquad (12.32)$$

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

12.6.2 Dự đoán dựa vào 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.

$$\hat{y}_{n+L} = y_n(\bar{t})^L \qquad (12.33)$$

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

12.6.3 Ngoại suy hàm xu thế

Dựa trên cơ sở chiều hướng biến động của hiện tượng nghiên cứu được khái quát hoá bằng một hàm số tuyến tính có dạng:

$$\hat{y}_t = a_0 + a_1 t \qquad (12.34)$$

$\hat{y}_t$: giá trị của hiện tượng tại thời gian t xác định bằng hàm số tuyến tính

t : thứ tự thời gian (t = 1,2, . . . , n)

a_0, a_1 : các tham số qui định vị trí của đường thẳng

Tổng quát $\hat{y}_t = f(t)$

Do đó: $\hat{y}_{n+L} = f(n + L)$ (12.35)

Trong đó: $\hat{y}_{n+L}$: giá trị dự đoán ở thời gian n + L

L : tầm xa dự đoán

Trở lại ví dụ 6, với hàm xu thế tìm được : $\hat{y}_t = 10{,}586t + 2214{,}1$, ta dự đoán doanh số quí 4 năm 2003 (t = 48)

$\hat{y}_{48} = 10{,}586(48) + 2214{,}1 = 2722{,}23$ triệu đồng.

12.6.4 Dự đoán dựa trên mô hình nhân

Mô hình dự đoán này được xây dựng dựa trên cơ sở phân tích các yếu tố thành phần tác động đến hiện tượng nghiên cứu: xu hướng (T), thời vụ (S) và chu kỳ (C), còn yếu tố thứ tư: yếu tố ngẫu nhiên (I) thì không thể dự đoán được).

Đầu tiên, ta dự đoán từng yếu tố thành phần rồi nhân kết quả của chúng lại với nhau

$$\hat{y} = T.S.C \qquad (12.36)$$

Trở lại ví dụ 6, dự đoán doanh số quí 4 năm 2003 (t = 48):

$\hat{y}_{48} = 10{,}586(48) + 2214{,}1 = 2722{,}23$ triệu đồng.

Chỉ số thời vụ quí 4 = 93%.

Quan sát Bảng 12.2 và Hình 12.7, ta thấy biến động chu kỳ tương đối nhỏ và đây là yếu tố khó dự đoán, vì vậy để đơn giản, ở đây ta bỏ qua biến động chu kỳ, tức là cho chỉ số chu kỳ bằng 1.

Doanh số quí 4 năm 2003 có thể đạt được là:

$\hat{y} = T.S.C = (2722{,}23)(0{,}93)(1) = 2531{,}67$ triệu đồng.

12.6.5 Dự đoán bằng phương pháp san bằng mũ [1]

Phương pháp san bằng mũ có nhiều hình thức khác nhau, được sử dụng tùy theo biến động của dãy số có hay không biến động xu hướng hoặc thời vụ.

12.6.5.1 Phương pháp san bằng mũ đơn giản[2]

---

[1] Exponential smoothing method

[2] Simple exponential smoothing method

Phương pháp này thường được sử dụng trong dự đoán ngắn hạn đối với dãy số thời gian không có xu hướng hoặc biến động thời vụ rõ rệt. Theo phương pháp này, ở thời gian t nào đó, dựa vào các giá trị thực tế đã biết để ước lượng giá trị hiện tại (thời gian t) của hiện tượng và dùng giá trị hiện tại này để dự đoán giá trị tương lai (thời gian t+1). Thực chất của phương pháp san bằng mũ đơn giản là ứng dụng tính chất của số trung bình di động – san bằng biến động bất thường, ngẫu nhiên của dãy số, làm "bằng phẳng" dãy số thực tế và dùng dãy số mới này (dãy số đã được làm "bằng phẳng") để dự đoán giá trị tương lai. Nhưng chú ý rằng theo phương pháp này không phải tất cả các giá trị quá khứ đều có ảnh hưởng ngang nhau đến việc dự đoán giá trị tương lai, mà các giá trị càng "mới", càng gần với thời gian dự đoán thì giá trị thông tin mới càng cao và do vậy càng có ảnh hưởng đến giá trị dự đoán, tức là các giá trị càng gần với thời gian dự đoán thì được gán cho trọng số càng lớn.

$$\hat{y}_{t+1} = S_t \qquad (12.37)$$

$\hat{y}_{t+1}$: giá trị dự đoán của hiện tượng ở thời gian t + 1

S_t: trung bình có trọng số của các giá trị thực tế $y_t, y_{t-1}, y_{t-2}, \ldots, y_1$

Theo phương pháp san bằng mũ đơn giản, ta có:

$$\hat{y}_{t+1} = w(y_t) + w(1-w)(y_{t-1}) + w(1-w)^2(y_{t-2}) + \ldots$$

hay

$$\hat{y}_{t+1} = w(y_t) + (1-w)(\hat{y}_t)$$

$$\hat{y}_{t+1} = (y_t) + (1-w)(\hat{y}_t - y_t) \qquad (12.38)$$

trong đó: w trọng số[1] (hằng số san bằng mũ)[2] và $0 < w < 1$

- *Xác định* $\hat{y}_1$ *và* w

Ta nhận xét rằng: ảnh hưởng của giá trị dự đoán đầu tiên ngày càng giảm dần (tức là thời điểm dự đoán càng xa thì ảnh hưởng của giá trị dự đoán đầu tiên càng giảm), do đó để đơn giản, người ta thường chọn:

$$\hat{y}_1 = y_t \qquad (12.39)$$

Giá trị của w càng lớn thì dãy số dự đoán càng theo sát với biến động của dãy số thực tế, ngược lại, w càng nhỏ thì dãy số dự đoán càng ít sát với những biến động của dãy số thực tế. Do đó, nếu dãy số có nhiều biến động ngẫu nhiên, bất thường, ta có thể chọn w nhỏ, và ngược lại đối với dãy số tương đối "bằng phẳng", ổn định, ta chọn w lớn. Trong thực tế, w có thể được chọn lựa một cách chủ quan, tức là dựa vào kinh nghiệm để sao cho w đạt được giá trị dự đoán chính xác nhất.

---

[1] Weighting factor

[2] Smoothing constant

- *Chọn lựa phương pháp dự đoán thích hợp*

Không có phương pháp nào là tuyệt đối trong tất cả các trường hợp, tùy vào bản chất của hiện tượng, thời gian cần dự đoán cùng với kinh nghiệm thực tế là những yếu tố cần thiết để cân nhắc, sử dụng phương pháp nào là phù hợp hơn. Thông thường để xác định một phương pháp dự đoán có thích hợp hay không, người ta xem xét các sai số dự đoán bằng cách so sánh giá trị dự đoán và giá trị thực tế. Đó chính là chênh lệch giữa giá trị thực tế tại thời điểm t và giá trị dự đoán tại thời điểm t được gọi là sai số dự đoán[1] hay phần dư[2]. Chênh lệch càng nhỏ tức là dự đoán càng chính xác. Chênh lệch này là nhỏ nhất khi đường biểu diễn dự đoán gần sát nhất với đường biểu diễn dãy số thực tế. Ta có thể dùng các đại lượng: trung bình bình phương sai số dự đoán (*MSE*)[3] hay trung bình độ lệch tuyệt đối của sai số dự đoán (MAD)[4], căn bậc hai của trung bình bình phương sai số dự đoán (RMSE)[5] hoặc trung bình của các trị tuyệt đối của phần trăm sai số (MAPE)[6].

$$MSE = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n} \quad (12.40) \quad ; \quad MAD = \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|}{n} \quad (12.41)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}} \quad (12.42) \quad ; \quad MAPE = \frac{\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|}{n} \times 100\% \quad (12.43)$$

n: số các mức độ trong dãy số thời gian

Ví dụ 12.7: tính RMSE cho ví dụ 5

năm	quí	t	y_t	TBDĐ 4 mức độ	TBDĐ 2 mức độ	CS TV	CSTV điều chỉnh	DSTT đã loại bỏ yếu tố thời vụ	PT đường thẳng	$\hat{y}_t$	$(y_t - \hat{y}_t)^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)=(8)(10)/(10)	(12)
1992	Q1	1	2513				97	2590,72	2224,71	2157,96	126049,98
	Q2	2	2142				108	1983,33	2235,29	2414,12	74046,77
	Q3	3	2131	2245,3	2228,0	95,6	102	2089,22	2245,88	2290,80	25534,62

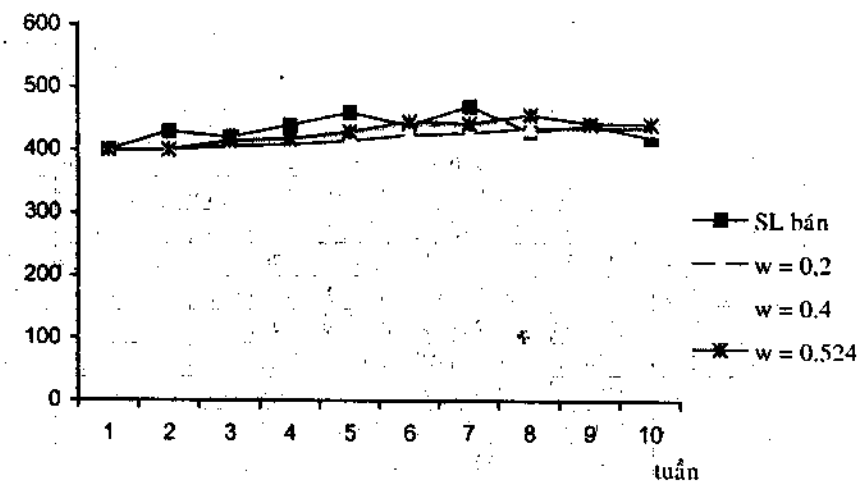
[1] Forecast error
[2] Residual
[3] Mean square error
[4] Mean absolute deviation
[5] Root mean square error
[6] Mean absolute percent error

	Q4	4	2195	2210,8	2219,1	98,9	93	2360,22	2256,46	2098,51	9310,03
1993	Q1	5	2375	2227,5	2235,5	106,2	97	2448,45	2267,05	2199,04	30962,45
	Q2	6	2209	2243,5	2188,8	100,9	108	2045,37	2277,64	2459,85	62924,16
	Q3	7	2195	2134,0	2080,1	105,5	102	2151,96	2288,22	2333,99	19317,23
	Q4	8	1757	2026,3	2011,9	87,3	93	1889,25	2298,81	2137,89	145078,29
1994	Q1	9	1944	1997,5	1961,9	99,1	97	2004,12	2309,39	2240,11	87682,42
	Q2	10	2094	1926,3	1960,5	106,8	108	1938,89	2319,98	2505,58	169396,78
	Q3	11	1910	1994,8	2007,4	95,1	102	1872,55	2330,57	2377,18	218254,65
	Q4	12	2031	2020,0	2071,1	98,1	93	2183,87	2341,15	2177,27	21395,31
1995	Q1	13	2045	2122,3	2163,3	94,5	97	2108,25	2351,74	2281,19	55783,76
	Q2	14	2503	2204,3	2249,8	111,3	108	2317,59	2362,32	2551,31	2333,85
	Q3	15	2238	2295,3	2362,5	94,7	102	2194,12	2372,91	2420,37	33258,16
	Q4	16	2395	2429,8	2479,5	96,6	93	2575,27	2383,50	2216,65	31808,27
1996	Q1	17	2583	2529,3	2555,5	101,1	97	2662,89	2394,08	2322,26	67985,59
	Q2	18	2901	2581,8	2590,0	112,0	108	2686,11	2404,67	2597,04	92390,81
	Q3	19	2448	2598,3	2605,9	93,9	102	2400,00	2415,25	2463,56	242,08
	Q4	20	2461	2613,5	2624,4	93,8	93	2646,24	2425,84	2256,03	42012,21
1997	Q1	21	2644	2635,3	2700,3	97,9	97	2725,77	2436,43	2363,33	78773,84
	Q2	22	2988	2765,3	2762,4	108,2	108	2766,67	2447,01	2642,77	119181,71
	Q3	23	2968	2759,5	2753,9	107,8	102	2909,80	2457,60	2506,75	212751,60
	Q4	24	2438	2748,3	2755,4	88,5	93	2621,51	2468,18	2295,41	20331,59
1998	Q1	25	2599	2762,5	2781,9	93,4	97	2679,38	2478,77	2404,41	37866,47
	Q2	26	3045	2801,3	2831,6	107,5	108	2819,44	2489,36	2688,50	127089,06
	Q3	27	3123	2862,0	2830,0	110,4	102	3061,76	2499,94	2549,94	328396,80
	Q4	28	2681	2798,0	2763,4	97,0	93	2882,80	2510,53	2334,79	119860,64
1999	Q1	29	2343	2728,8	2686,6	87,2	97	2415,46	2521,11	2445,48	10502,27
	Q2	30	2768	2644,5	2607,1	106,2	108	2562,96	2531,70	2734,24	1140,01
	Q3	31	2786	2569,8	2609,3	106,8	102	2731,37	2542,29	2593,13	37198,17
	Q4	32	2382	2648,8	2663,4	89,4	93	2561,29	2552,87	2374,17	61,29
2000	Q1	33	2659	2678,0	2652,1	100,3	97	2741,24	2563,46	2486,55	29737,53
	Q2	34	2885	2626,3	2643,0	109,2	108	2671,30	2574,04	2779,97	11031,82
	Q3	35	2579	2659,8	2622,8	98,3	102	2528,43	2584,63	2636,32	3285,88
	Q4	36	2516	2585,8	2573,1	97,8	93	2705,38	2595,22	2413,55	10495,82
2001	Q1	37	2363	2560,5	2571,3	91,9	97	2436,08	2605,80	2527,63	27102,36
	Q2	38	2784	2582,0	2528,9	110,1	108	2577,78	2616,39	2825,70	1738,81
	Q3	39	2665	2475,8	2474,0	107,7	102	2612,75	2626,97	2679,51	210,64
	Q4	40	2091	2472,3	2447,0	85,5	93	2248,39	2637,56	2452,93	130993,90
2002	Q1	41	2349	2421,8	2390,3	98,3	97	2421,65	2648,15	2568,70	48268,80
	Q2	42	2582	2358,8	2366,5	109,1	108	2390,74	2658,73	2871,43	83770,05
	Q3	43	2413	2374,3			102	2365,69	2669,32	2722,70	95916,79
	Q4	44	2153				93	2590,72	2679,90	2492,31	115131,76
2003	Q1	45					97		2690,49	2609,78	
	Q2	46					108		2701,08	2917,16	
	Q3	47					102		2711,66	2765,90	
	Q4	48					93		2722,25	2531,69	

RMSE	259,66

Ví dụ 8: theo dõi số liệu sản lượng bán qua 10 tuần tại một cửa hàng
Ta tính giá trị dự đoán theo công thức với w = 0,2, và 0,4. (Hướng dẫn: vào menu **Tools – Add in – Solver** trên Excel ta có thể tìm giá trị w sao cho MSE là nhỏ nhất, w đạt điều kiện này bằng 0,524)

Tuần	SL bán	Dự đoán			$y_t - \hat{y}_t$			$(y_t - \hat{y}_t)^2$		
		w = 0,2	w = 0.4	w = 0.524	w = 0,2	w = 0.4	w = 0.524	w = 0,2	w = 0.4	w = 0.524
1	400	400	400	400	0	0	0	0	0	0
2	430	400	400	400	30	30	30	900	900	900
3	420	406	412	415,72	14	8	4,28	196	64	18,3184
4	440	408,8	415,2	417,96	31,2	24,8	22,04	973,44	615,04	485,64
5	460	415,04	425,12	429,51	44,96	34,88	30,49	2021,40	1216,61	929,62
6	440	424,03	439,07	445,49	15,97	0,928	-5,49	254,98	0,86	30,11
7	470	427,23	439,44	442,61	42,77	30,56	27,39	1829,65	933,72	750,12
8	430	435,78	451,67	456,96	-5,78	-21,67	-26,96	33,41	469,41	727,01
9	440	434,62	443,00	442,83	5,38	-3,00	-2,83	28,90	9,00	8,03
10	420	435,70	441,80	441,35	-15,70	-21,80	-21,35	246,47	475,23	455,79
Cộng					162,80	82,70	57,56	6484,25	4683,87	4304,64
							MSE	648,43	468,39	430,46



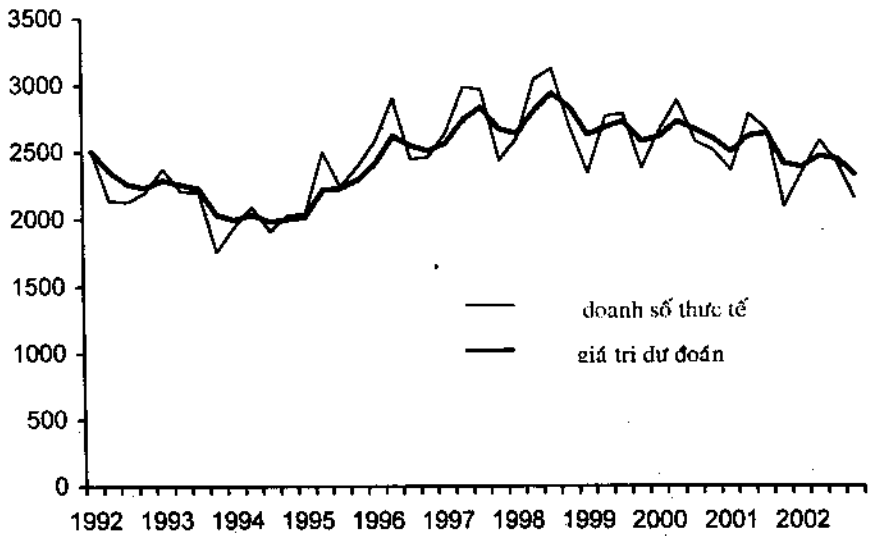
Hình 12.9: số lượng sản phẩm bán thực tế và giá trị dự đoán với w = 0,2, w = 0,4 và w = 0,524

Ví dụ 9: sử dụng số liệu ví dụ 5, dùng phương pháp dự báo san bằng mũ đơn giản để dự báo cho các quí năm 2003.

(Hướng dẫn: vào menu **Tools – Add in – Solver** trên Excel ta có thể tìm giá trị w tối ưu sao cho giá trị MSE là nhỏ nhất, w đạt điều kiện này bằng 0,4165065)

năm	quí	y_t	$\hat{y}_t$ với **w=0.1**	$(y_t - \hat{y}_t)^2$	$\hat{y}_t$ với **w=0.5**	$(y_t - \hat{y}_t)^2$	$\hat{y}_t$ với **W tối ưu**	$(y_t - \hat{y}_t)^2$
			0,1		0,5		**0,3417**	
1992	Q1	2513	2513,00		2513,00		2513,00	
	Q2	2142	2513,00	137641,00	2513,00	137641,00	2386,21	137641,00
	Q3	2131	2475,90	145924,00	2327,50	145924,00	2302,75	65132,89
	Q4	2195	2441,41	78904,81	2229,25	17556,25	2244,06	11610,64
1993	Q1	2375	2416,77	4410,29	2212,13	21243,06	2227,29	17146,19
	Q2	2209	2412,59	43167,96	2293,56	9,77	2277,77	334,58
	Q3	2195	2392,23	47346,32	2251,28	9714,57	2254,27	6850,98
	Q4	1757	2372,51	403520,82	2223,14	244313,95	2234,01	247275,86
1994	Q1	1944	2310,96	183620,48	1990,07	77919,49	2071,00	84107,85
	Q2	2094	2274,26	47071,05	1967,04	10801,38	2027,59	529,23
	Q3	1910	2256,24	132687,37	2030,52	3253,01	2050,29	13828,53
	Q4	2031	2221,61	50731,48	1970,26	0,23	2002,35	372,05
1995	Q1	2045	2202,55	31192,10	2000,63	5586,25	2012,14	1819,43
	Q2	2503	2186,80	90269,26	2022,81	252376,23	2023,37	24094 5,54
	Q3	2238	2218,42	2621,81	2262,91	46304,71	2187,28	46066,68
	Q4	2395	2220,38	31181,64	2250,45	17448,47	2204,61	43146,92
1996	Q1	2583	2237,84	131496,83	2322,73	110587,06	2269,68	143175,55
	Q2	2901	2272,35	439784,39	2452,86	334399,85	2376,76	39856 6,95
	Q3	2448	2335,22	30851,58	2676,93	23, 65	2555,92	5075,78
	Q4	2461	2346,50	15821,00	2562,47	46626,50	2519,04	9008,86
1997	Q1	2644	2357,95	88508,27	2511,73	6647,82	2499,20	15616,18
	Q2	2988	2386,55	396966,86	2577,87	226830,32	2548,69	238923,65
	Q3	2968	2446,70	338081,49	2782,93	152204,18	2698,82	175823,96
	Q4	2438	2498,83	75,64	2875,47	118978,93	2790,81	68027,56
1998	Q1	2599	2492,74	10034,57	2656,73	76433,79	2670,24	36791,98
	Q2	3045	2503,37	304986,04	2627,87	150751,02	2645,89	140445,50
	Q3	3123	2557,53	383941,18	2836,43	245157,03	2782,29	227630,63
	Q4	2681	2614,08	15244,07	2979,72	24159,52	2898,73	10259,13
1999	Q1	2343	2620,77	73484,26	2830,36	405408,11	2824,32	308830,56
	Q2	2768	2592,99	21676,14	2586,68	3888,56	2659,83	3171,72
	Q3	2786	2610,50	37251,07	2677,34	39728,79	2696,80	15919,24
	Q4	2382	2628,05	52210,05	2731,67	87225,47	2727,28	99096,48
2000	Q1	2659	2603,44	958,17	2556,83	5280,90	2609,28	4662,33
	Q2	2885	2609,00	79275,42	2607,92	107692,34	2626,27	76020,39

	Q3	2579	2636,60	899,82	2746,46	836,22	2714,69	2234,74
	Q4	2516	2630,84	14543,71	2662,73	53111,22	2668,32	39478,71
2001	Q1	2363	2619,35	71736,96	2589,36	89837,69	2616,26	93220,21
	Q2	2784	2593,72	27108,37	2476,18	37882,91	2529,71	28135,09
	Q3	2665	2612,75	5081,06	2630,09	35652,11	2616,61	18302,83
	Q4	2091	2617,97	272219,49	2647,55	290619,29	2633,15	276270,58
2002	Q1	2349	2565,27	72345,89	2369,27	89129,47	2447,87	80741,30
	Q2	2582	2543,65	279,73	2359,14	45252,86	2414,08	17990,49
	Q3	2413	2547,48	17068,71	2470,57	2901,29	2471,47	1,17
	Q4	2153	2534,03	155616,47	2441,78	100849,56	2451,49	101421,76
2003	**Q1**		2534,03		**2441,78**			
	Q2		2534,03		**2441,78**			
	Q3		2534,03		**2441,78**			
	Q4		2534,03		**2441,78**			
		MSE		104368,32		90190,44		82596,55
		RMSE		**323,06**		**300,32**		**287,40**



Hình 12.10: Sản lượng bán thực tế và giá trị dự đoán với w tối ưu = 0,3417

12.6.5.2 Phương pháp san bằng mũ Holt-Winters

- *Biến động của dãy số thời gian có tính xu hướng*

Đầu tiên ta ước lượng giá trị S_t của hiện tượng ở thời điểm t, sau đó ước lượng thành phần thể hiện xu hướng T_t .

Giả sử ta có dãy số thời gian $y_1, y_2, ..., y_n$ với biến động có tính xu hướng.

Đặt $S_2 = y_2$ và $T_2 = y_2 - y_1$

Ta có: $S_t = \alpha(y_t) + (1-\alpha)(S_{t-1} + y_{t-1})$ với $0 < \alpha, \beta < 1$ (12.44)

$T_t = \beta(S_t - S_{t-1}) + (1-\beta)(T_{t-1})$ t = 3,4, . . . ,n (12.45)

α, β : các hằng số san bằng mũ

Muốn dự đoán giá trị của hiện tượng ở thời điểm n + h, dùng công thức:

$\hat{y}_{n+h} = S_n + hT_n$ với h = 1,2,3, . . . (12.46)

Ví dụ 10: có số liệu sản phẩm tiêu thụ thời kỳ 1988 – 2002, dùng Solver trên Excel ta tìm ra các giá trị $\alpha = 0,88$, $\beta = 0,01$

Đầu tiên đặt: $S_2 = y_2 = 61,5$, $T_2 = y_2 - y_1 = 6,1$

Áp dụng các công thức (12.44) và (12.45) để tính các giá trị tiếp theo của S và T, ta có được bảng số liệu như dưới đây.

Để dự đoán cho tương lai, dùng công thức (12.46)

Năm	sản phẩm tiêu thụ (ngàn cái)	t	S_t	T_t	$\hat{y}_t$ với α, β	$(y_t - \hat{y}_t)^2$
1988	55,4	1				
1989	61,5	2	61,50	6,10		
1990	68,7	3	68,57	6,11	67,60	0,94
1991	87,2	4	85,70	6,22	80,79	24,18
1992	90,4	5	90,58	6,21	104,36	189,96
1993	86,2	6	87,46	6,11	115,41	780,86
1994	94,7	7	94,57	6,12	118,03	550,59
1995	103,2	8	102,90	6,15	131,30	806,83
1996	119	9	117,81	6,23	145,92	789,93
1997	122,4	10	122,60	6,22	167,67	2032,08
1998	131,6	11	131,27	6,24	178,56	2236,81
1999	157,6	12	155,20	6,42	193,70	1481,93
2000	181	13	178,69	6,59	225,82	2221,61
2001	217,8	14	213,92	6,88	257,77	1923,34
2002	244,1	15	241,32	7,08	303,32	3843,97
2003					340,47	
					MSE	**129,93**

- *Biến động của dãy số thời gian có tính xu hướng và tính thời vụ*

Đây chỉ là một trường hợp mở rộng của phương pháp Holt-Winters ở trên nhưng có xét thêm tính xu hướng và thời vụ của dãy số.

Giả sử ta có dãy số thời gian $y_1, y_2, ..., y_n$ biến động có tính xu hướng và thời vụ, với s thời kỳ trong năm (nếu dãy số theo quí thì s = 4, nếu là tháng thì s = 12). Phương pháp Holt-Winters trong trường hợp này được thực hiện theo 2 bước sau:

a.Thông qua số trung bình di động y_1^* để xác định các giá trị S_t, T_t, F_t đầu tiên

Với $$y_1^* = \frac{y_{t-s/2} + 2(y_{t-(s/2)+1} + ... + y_{t+(s/2)-1}) + y_{t+s/2}}{2s} \qquad (12.47)$$

T = (s/2) + 1, (s/2) + 2, . . ., (5s/2)

Đặt các giá trị S_t, T_t, F_t đầu tiên

$$y^*_{5s/2} = S_{5s/2}$$

$$T_{5s/2} = S_{5s/2} - S_{(5s/2)-1}$$

$$F_{(5s/2)-j} = \frac{1}{2}\left(\frac{y_{(5s/2)-j}}{S_{(5s/2)-j}} + \frac{y_{(3s/2)-j}}{S_{(3s/2)-j}}\right)$$

j = 0, 1, 2, 3, . . . , s-1

b. Bắt đầu ở thời kỳ thứ [5(s/2) +1], các giá trị S_t, T_t, F_t được xác định như sau:

$$S_t = (1-\alpha)(S_{t-1} + T_{t-1}) + \alpha\frac{y_t}{F_{t-s}} \qquad (0 < \alpha < 1) \qquad (12.48)$$

$$T_t = (1-\beta)T_{t-1} + \beta(S_t - S_{t-1}) \qquad (0 < \beta < 1) \qquad (12.49)$$

$$F_t = (1-\gamma)F_{t-s} + \gamma\frac{y_t}{S_t} \qquad (0 < \gamma < 1) \qquad (12.50)$$

Ở thời điểm n, muốn dự đoán giá trị hiện tượng ở thời điểm n + h:

$$\hat{y}_{n+h} = (S_n + hT_n)F_{n+h-s} \qquad (h = 1,2, . . . , s) \qquad (12.51)$$

hay $$\hat{y}_{n+h} = (S_n + hT_n)F_{n+h-2s} \qquad (h = s+1, s+2, . . . , 2s)$$

Ví dụ 11: với bảng số liệu ở ví dụ 5 nhưng dùng phương pháp dự báo Holt – Winters để dự báo cho các năm 2003, 2004, 2005 và 2006. (Hướng dẫn: vào menu **Tools – Add in – Solver** trên Excel ta có thể tìm đồng thời các giá trị α, β và γ sao cho MSE hay RMSE là nhỏ nhất). Ta thấy phương pháp này cho MSE và RMSE nhỏ hơn so với phương pháp san bằng mũ đơn giản (ví dụ 9).

Cách tính toán như sau:

a. đầu tiên tính các số trung bình di động 4 mức độ, ta tiếp tục tính các số trung bình di động 2 mức độ để xác định các giá trị đầu tiên S_t, T_t, F_t.

$y_{10}^{\bullet} = S_{10} = 1960,5$

$T_{10} = S_{10} - S_9 = 1960,5 - 1961,9 = -1,4$

$$F_7 = \frac{1}{2}\left(\frac{y_3}{y_3^{\bullet}}\right) = \frac{1}{2}\left(\frac{2131}{2228} + \frac{2195}{2080}\right) = 1,005$$

b. bắt đầu từ thời gian t = 11, với các giá trị của $\alpha = 0,2563$, $\beta = 0,4212$ và $\lambda = 0,2813$, ta tính các giá trị mới của S_t, T_t, F_t theo các công thức (12.48), (12.49), (12.50) (Hướng dẫn: vào menu **Tools – Add in – Solver** trên Excel ta có thể tìm giá trị α, β, λ tối ưu sao cho giá trị MSE hay RMSE là nhỏ nhất)

$$S_{11} = (1-\alpha)(S_{10} + T_{10}) + \alpha\left(\frac{y_{11}}{F_7}\right)$$

$$= (1 - 0,2563)(1960,5 + (-1,4)) + 0,2563\left(\frac{1910}{1,0058}\right)$$

$= 1943,7$

$$T_{11} = (1-\beta)T_{10} + \beta(S_{11} - S_{10})$$

$= (1 - 0,4212)(-1,4) + 0,4212(1943,7 - 1960,5)$

$= -7,9$

$$F_{11} = (1-\gamma)F_7 + \gamma\left(\frac{y_{11}}{S_{11}}\right)$$

$$= (1 - 0,2813)1,0058 + 0,2813\left(\frac{1910}{1943,7}\right) = 0,9993$$

Để dự đoán sản lượng có thể đạt được trong tương lai, ta dùng công thức: (12.51)

$\hat{y}_{11} = (S_{10} + T_{10})F_7 = (1960,5 + (-1,4))1,0058 = 1970,5744$

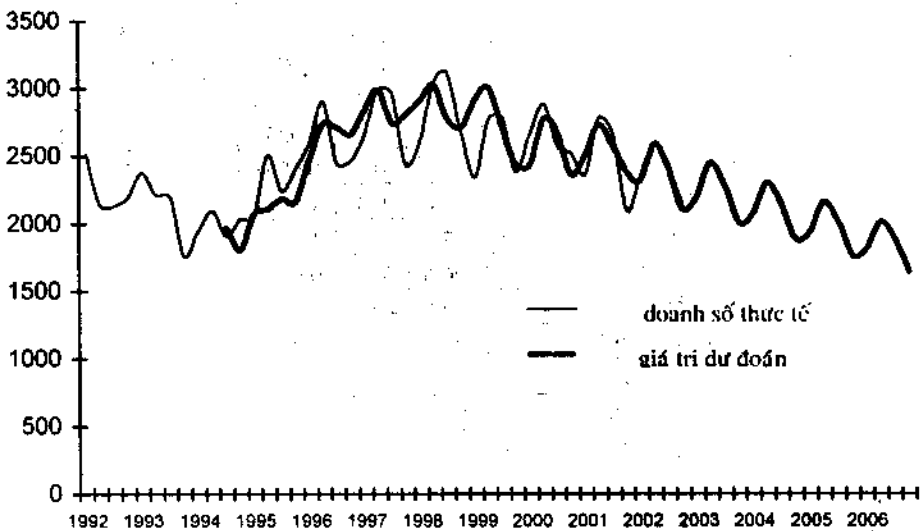
$\hat{y}_{45} = (S_{44} + T_{44})F_{41} = [2317,6 + (-33,1)]0,9619 = 2197,6$

$\hat{y}_{46} = [S_{44} + (46-44)T_{41}]F_{42} = [2317,6 + (46-44)(-33,1)]1,0858 = 2444,5$

Năm	quí	t	y	TBDĐ 4 mức độ	TBDĐ 2 mức độ y^*	S_t	T_t	F_t	$\hat{y}_t$ với α, β, λ	$(y_t - \hat{y}_t)^2$
1992	Q1	1	2513							
	Q2	2	2142	2245,25						
	Q3	3	2131	2210,75	2228,00					
	Q4	4	2195	2227,50	2219,13					
1993	Q1	5	2375	2243,50	2235,50					
	Q2	6	2209	2134,00	2188,75					
	Q3	7	2195	2026,25	2080,13			1,0058		
	Q4	8	1757	1997,50	2011,88			0,9312		
1994	Q1	9	1944	1926,25	1961,88			1,0266		
	Q2	10	2094	1994,75	1960,50	1960,5	-1,4	1,0387		
	Q3	11	1910	2020,00		1943,7	-7,9	0,9993	1970,5744	3669,25
	Q4	12	2031	2122,25		1998,7	18,6	0,9551	1802,6684	52135,30
1995	Q1	13	2045	2204,25		2010,8	15,9	1,0239	2071,0110	676,57
	Q2	14	2503	2295,25		2124,8	57,2	1,0779	2105,0047	158400,28
	Q3	15	2238	2429,75		2196,8	63,4	1,0048	2180,6107	3293,53
	Q4	16	2395	2529,25		2323,6	90,1	0,9764	2158,8134	55784,10
1996	Q1	17	2583	2581,75		2441,7	101,9	1,0335	2471,5551	12419,97
	Q2	18	2901	2598,25		2581,5	117,9	1,0908	2741,6088	25405,56
	Q3	19	2448	2613,50		2631,9	89,5	0,9838	2712,2566	69831,55
	Q4	20	2461	2635,25		2669,9	67,8	0,9610	2657,1231	38464,26
1997	Q1	21	2644	2765,25		2691,7	48,4	1,0191	2829,3388	34350,46
	Q2	22	2988	2759,50		2739,9	48,3	1,0907	2988,8436	0,71
	Q3	23	2968	2748,25		2846,8	73,0	1,0003	2743,0068	50621,95
	Q4	24	2438	2762,50		2821,7	31,7	0,9337	2806,0545	135464,12
1998	Q1	25	2599	2801,25		2775,7	-1,1	0,9958	2907,8112	95364,38
	Q2	26	3045	2862,00		2779,0	0,8	1,0921	3026,3331	348,45
	Q3	27	3123	2798,00		2867,5	37,7	1,0253	2780,6950	117172,68
	Q4	28	2681	2728,75		2896,6	34,1	0,9314	2712,7403	1007,45
1999	Q1	29	2343	2644,50		2782,5	-28,3	0,9525	2918,3358	331011,34
	Q2	30	2768	2569,75		2697,9	-52,0	1,0735	3007,9378	57570,14
	Q3	31	2786	2648,75		2664,2	-44,3	1,0310	2712,7877	5360,05
	Q4	32	2382	2678,00		2603,8	-51,1	0,9268	2440,2219	3389,78
2000	Q1	33	2659	2626,25		2613,9	-25,3	0,9707	2431,6051	51708,45
	Q2	34	2885	2659,75		2614,0	-14,6	1,0820	2778,9403	11248,66
	Q3	35	2579	2585,75		2574,2	-25,2	1,0228	2680,0250	10206,06
	Q4	36	2516	2560,50		2591,5	-7,3	0,9392	2362,3183	23618,05
2001	Q1	37	2363	2582,00		2545,8	-23,5	0,9588	2508,6116	21202,75
	Q2	38	2784	2475,75		2535,3	-18,0	1,0865	2729,0910	3014,99
	Q3	39	2665	2472,25		2539,9	-8,5	1,0303	2574,7191	8150,65

	Q4	40	2091	2421,75		2453,2	-41,4	0,9147	2377,3891	82018,70
2002	Q1	41	2349	2358,75		2421,6	-37,3	0,9619	2312,3797	1341,05
	Q2	42	2582	2374,25		2382,3	-38,1	1,0858	2590,6159	74,23
	Q3	43	2413			2343,6	-38,4	1,0301	2415,0643	4,26
	Q4	44	2153			2317,6	-33,1	0,9187	2108,6931	1963,10
2003	**Q1**	**45**							**2197,6**	
	Q2	**46**							**2444,5**	
	Q3	**47**							**2285,0**	
	Q4	**48**							**2007,5**	
2004	**Q1**	**49**							**2070,1**	
	Q2	**50**							**2300,6**	
	Q3	**51**							**2148,4**	
	Q4	**52**							**1885,8**	
2005	**Q1**	**53**							**1942,6**	
	Q2	**54**							**2156,7**	
	Q3	**55**							**2011,9**	
	Q4	**56**							**1764,0**	
2006	**Q1**	**57**							**1815,1**	
	Q2	**58**							**2012,8**	
	Q3	**59**							**1875,4**	
	Q4	**60**							**1642,2**	

MSE	43126,26
RMSE	**207,67**



Hình 12.11: doanh số thực tế và dự đoán

CHƯƠNG 13

CHỈ SỐ [1]

13.1. GIỚI THIỆU

13.1.1. Giới thiệu: Phương pháp chỉ số ngày càng được sử dụng rộng rãi trong quản lý và nghiên cứu kinh tế. Chẳng hạn chỉ số có thể cho biết giá cả, khối lượng sản phẩm từng loại hay nhiều loại, tăng lên hay giảm xuống qua thời gian hoặc qua những không gian khác nhau. Một cách tổng quát, trong thống kê chỉ số là phương pháp biểu hiện quan hệ so sánh giữa hai mức độ nào đó của một hiện tượng kinh tế – xã hội.

13.1.2. Phân loại chỉ số

*** Căn cứ vào phạm vi tính toán chỉ số được chia thành hai loại:**

- **Chỉ số cá thể** : Là loại chỉ số đo lường sự thay đổi của từng phần tử, từng đơn vị trong tổng thể nghiên cứu. Ví dụ chỉ số giá của từng mặt hàng đang bán trên thị trường, chỉ số khối lượng của từng loại sản phẩm sản xuất trong xí nghiệp v.v...
- **Chỉ số tổng hợp**: Là loại chỉ số đo lường sự thay đổi của một số hoặc tất cả các phần tử thuộc tổng thể nghiên cứu. Ví dụ chỉ số giá của toàn bộ các mặt hàng bán lẻ tại một thị trường. Chỉ số khối lượng của các sản phẩm sản xuất trong một xí nghiệp v.v...

*** Căn cứ theo tính chất của chỉ tiêu nghiên cứu chỉ số được chia thành hai loại :**

- 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 chỉ số khối lượng sản phẩm sản xuất, chỉ số khối lượng hàng hóa tiêu thụ v.v..
- 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 chỉ số giá thành sản phẩm, chỉ số giá cả v.v...

13.2. CHỈ SỐ CÁ THỂ

13.2.1. Chỉ số cá thể giá cả

Gọi giá một mặt hàng nào đó ở kỳ nghiên cứu, kỳ gốc là p_1 và p_0, công thức tính chỉ số cá thể giá cả :

$$i_p = \frac{p_1}{p_0} \qquad (13.1)$$

[1] Index Number

Ví dụ: Có số liệu về giá bán trung bình (ngàn đồng/kg) của loại gạo thường trên một thị trường từ năm 1995 đến năm 2000 , cho trong bảng 13.1

Bảng 13.1

Năm	Giá bán (ngàn đồng / kg)	Chỉ số giá (i_p) (%)
1995	**2,80**	**100,00**
1996	2,90	103,57
1997	2,95	105,36
1998	3,20	114,29
1999	3,40	121,43
2000	3,50	125,00

Từ số liệu của bảng 13.1, nếu chọn kỳ gốc là năm 1995, thì chỉ số giá của loại gạo thường được xác định bởi công thức 13.1 , kết quả tính toán ở cột 3 của bảng 13.1.

Ví dụ: Chỉ số giá năm 1999: $i_p = \frac{p_1}{p_0} = \frac{3,4}{2,8} = 1,2143 = 121,43\%$

Ý nghĩa: Giá bán 1kg gạo năm 1999 so với năm 1995 bằng 121,43% , tăng 21,43%. Về số tuyệt đối, giá 1kg tăng lên 3,4 – 2,8 = 0,6 ngàn đồng.

Khi chọn kỳ gốc làm cơ sở để so sánh, cần lưu ý hai nguyên tắc sau đây:

- Thứ nhất, kỳ gốc so sánh nên chọn thời kỳ nền kinh tế tương đối ổn định.
- Thứ hai, kỳ gốc nên chọn gần kề với kỳ nghiên cứu để kết quả so sánh không chịu ảnh hưởng của sự thay đổi bởi tiến bộ khoa học kỹ thuật, và những điều kiện sản xuất, tiêu dùng khác.

13.2.2. Chỉ số cá thể khối lượng

Gọi q_1, q_0 là khối lượng một loại sản phẩm hay mặt hàng được sản xuất hay tiêu thụ trong kỳ nghiên cứu, kỳ gốc. Chỉ số cá thể khối lượng được tính theo công thức :

$$i_q = \frac{q_1}{q_0} \qquad (13.2)$$

Ví dụ: Có số liệu về khối lượng bia được sản xuất ở Việt Nam từ năm 1995 đến 2000 và nếu chọn kỳ gốc là năm 1995 , thì chỉ số khối lượng được tính theo công thức 13.2 cho kết quả trong Bảng 13.2 dưới đây .

Bảng 13.2

Năm	Khối lượng bia sản xuất (triệu lít)	Chỉ số khối lượng i_q (%)
1995	**465,0**	**100,0**
1996	533,4	114,7
1997	581,0	124,9
1998	670,0	144,1
1999	689,8	148,3
2000	779,0	167,5

Ví dụ: Chỉ số cá thể khối lượng năm 2000:

$$i_q = \frac{q_1}{q_0} = \frac{779,0}{465,0} = 1,675 = 167,5\%$$

Ý nghĩa: Khối lượng bia tiêu thụ năm 2000 so với năm 1995 bằng 167,5%, tăng 67,5%. Về số tuyệt đối tăng 779 – 465 = 314 triệu lít.

13.3 CHỈ SỐ TỔNG HỢP

Chỉ số tổng hợp là loại chỉ số được sử dụng để đánh giá sự thay đổi của một số hoặc tất cả các phần tử thuộc tổng thể nghiên cứu .

13.3.1 Chỉ số tổng hợp giá cả

Chỉ số tổng hợp giá cả nêu lên sự biến động về giá cả của một nhóm hoặc tất cả các mặt hàng trên một thị trường hay ở các thị trường khác nhau .

Ví dụ: Có số liệu về giá cả và lượng hàng tiêu thụ của 3 mặt hàng : Đường, vải, dầu ăn trên một thị trường cho trong bảng 13.3:

Bảng 13.3

Mặt hàng	Đơn vị tính	Giá đơn vị (ngàn đồng)		Lượng tiêu thụ (ngàn đơn vị)	
		Năm 1995	Năm 2000	Năm 1995	Năm2000
Đường	Kg	5,0	6,0	10,0	13,0
Vải	Mét	40,0	50,0	20,0	25,0
Dầu ăn	Lít	10,0	12,2	5,0	5,5

13.3.1.1 Chỉ số tổng hợp giá đơn giản [1]

Chỉ số tổng hợp giá đơn giản (còn gọi là chỉ số tổng hợp giá không có trọng số) được xác định theo công thức:

[1] Simple Aggregative Price Index

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}} \qquad (13.3)$$

Trong đó:

$P_{i(0)}$: Giá của mặt hàng thứ i ở kỳ gốc.

$P_{i(1)}$: Giá của mặt hàng thứ i ở kỳ nghiên cứu.

Công thức 13.3, có thể viết đơn giản thành:

$$I_p = \frac{\sum p_1}{\sum p_0}$$

Từ số liệu bảng 13.3, để tính chỉ số tổng hợp giá đơn giản, ta áp dụng công thức 13.3, kết quả :

$$I_p = \frac{\sum p_1}{\sum p_0} = \frac{(6{,}0 + 50{,}0 + 12{,}2)}{(5{,}0 + 40{,}0 + 10{,}0)} = 1{,}24 = 124\%$$

Kết quả trên cho thấy giá cả của ba mặt hàng năm 2000 so với năm 1995 bằng 124 %, hay tăng 24 % .

Chỉ số tổng hợp giá đơn giản không mang đầy đủ ý nghĩa và tính đại diện cho sự thay đổi giá. Nếu thay đổi cách chọn giá cả đơn vị của các mặt hàng thì có thể làm thay đổi kết quả tính chỉ số giá. Ví dụ: 10.000 đồng/kg = 5000 đồng/500g, tuy nhiên chỉ số giá tính trong hai trường hợp sẽ khác nhau.

Chỉ số tổng hợp giá đơn giản không phản ánh được tầm quan trọng của các mặt hàng khác nhau, nghĩa là không để ý đến lượng hàng tiêu thụ của từng mặt hàng.

13.3.1.2. Chỉ số tổng hợp giá cả có trọng số [1]

Chỉ số tổng hợp giá có trọng số (quyền số), trong đó trọng số có thể là điểm quan trọng, hoặc là thứ bậc ưu tiên của mặt hàng đó. Chẳng hạn trong chỉ số tổng hợp giá người ta có thể lấy trọng số là lượng hàng hoá tiêu thụ.

13.3.1.2.1 Chỉ số giá theo phương pháp Laspeyres

Theo Laspeyres, khi tính chỉ số tổng hợp giá cả, trọng số được chọn là lượng hàng hóa tiêu thụ ở kỳ gốc. Ta có công thức chỉ số tổng hợp giá cả của Laspeyres:

---

[1] Weighted Aggregative Price Index

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)} q_{i(0)}}{\sum_{i=1}^{n} p_{i(0)} q_{i(0)}} \qquad (13.4)$$

Để đơn giản, từ đây về sau, công thức dạng 13.4 được viết thành:

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0}$$

Trở lại ví dụ bảng 13.3, để tính chỉ số tổng hợp giá cho ba mặt hàng theo phương pháp Laspeyres ta lập bảng 13.4

Bảng 13.4

Mặt hàng	Đơn vị tính	Giá đơn vị (1000đ)		Lượng tiêu thụ (1000đv)		p_0q_0 tr.đ	p_1q_0 tr.đ	p_0q_1 tr.đ	p_1q_1 tr.đ
		1995 p_0	2000 p_1	1995 q_0	2000 q_1				
Đường	kg	5,0	6,0	10,0	13,0	50	60	65	78,0
Vải	mét	40,0	50,0	20,0	25,0	800	1000	1000	1250,0
Dầu ăn	lít	10,0	12,2	5,0	5,5	50	61	55	67,1
Cộng		-	-	-	-	900	1121	1120	1395,1

Thay số liệu vào công thức 13.4, ta có chỉ số Laspeyres :

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{1121}{900} = 1,245 = 124,5\%$$

Ý nghĩa: Giá cả của nhóm ba mặt hàng nói trên năm 2000 so với năm 1995 bằng 124,5 %, tức là tăng 24,5 %. Do giá tăng nên làm cho tổng mức tiêu thụ hàng hoá tăng 1121 - 900 = 221triệu đồng.

Trước đây hay dùng công thức Laspeyres để tính chỉ số tổng hợp giá vì nó không đòi hỏi phải tính ngay $\sum p_1q_1$ và thường q_0 lúc nào cũng có sẵn, tuy nhiên phương pháp Laspeyres có nhược điểm là không phản ánh, cập nhật được những sự thay đổi về khuynh hướng, thói quen của người tiêu dùng. Một số mặt hàng nào đó nhiều năm trước đây được người tiêu dùng ưa chuộng và tiêu dùng với số lượng lớn, nhưng ngày nay có thể không còn quan trọng đối với họ. Điều nầy làm cho chỉ số giá tính theo Laspeyres không còn thích hợp .

13.3.1.2..2 Chỉ số giá theo phương pháp Paasche

Theo Paasche, khi tính chỉ số tổng hợp giá cả, trọng số được chọn là lượng hàng hóa tiêu thụ ở kỳ nghiên cứu. Ta có công thức chỉ số tổng hợp giá cả của Paasche:

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} \qquad (13.5)$$

Từ kết quả tính được trong bảng 13.4, ta tính chỉ số giá Paasche theo công thức 13.5:

$$I_p = \frac{1395,1}{1120,0} = 1,246 = 124,6\%$$

Kết quả trên cho thấy, giá cả của nhóm ba mặt hàng năm 2000 so với năm 1995 bằng 124,6 % hay tăng 24,6 %.

Cách tính chỉ số giá Paasche khắc phục được nhược điểm của phương pháp Laspeyres. Hiện nay với sự phát triển mạnh mẽ của công nghệ thông tin, việc thu thập và tổng hợp quyền số q_1 dễ dàng hơn, nên người ta ưa dùng chỉ số Paasche.

❖ Trong thực tế chỉ số giá của Paasche có thể biến đổi như sau:

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{\sum p_1 q_1}{\sum \frac{p_0}{p_1} p_1 q_1} = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}} \qquad (13.6)$$

Chỉ số 13.6 được gọi là chỉ số trung bình điều hòa, vì có dạng giống số trung bình điều hòa.

Nếu đặt: $d_1 = \frac{p_1 q_1}{\sum p_1 q_1}$

Thì công thức 13.6 được viết lại:

$$I_p = \frac{1}{\sum \frac{d_1}{i_p}} \qquad (13.7)$$

Trong đó i_p là các chỉ số cá thể về giá của các mặt hàng.

Dựa vào số liệu của Bảng 13.4, ta lập Bảng 13.5 để tính chỉ số giá theo công thức chỉ số trung bình điều hòa.

Bảng 13.5

Mặt hàng	Đơn vị tính	Giá đơn vị (1000đ)		Lượng tiêu thụ (1000dv)		p_1q_1 tr.đ	d_1	i_p
		1995 p_0	2000 p_1	1995 q_0	2000 q_1			
Đường	kg	5,0	6,0	10,0	13,0	78,0	0,0559	1,2
Vải	mét	40,0	50,0	20,0	25,0	1250,0	0,8960	1,25
Dầu ăn	lít	10,0	12,2	5,0	5,5	67,1	0,0481	1,22
Cộng						1395,1	1	

Thay số liệu trong bảng 13.5 vào công thức 13.6 và 13.7 ta có

$$I_p = \frac{\sum p_1q_1}{\sum \frac{p_1q_1}{i_p}} = \frac{78+1250+67,1}{\frac{78}{1,2}+\frac{1250}{1,25}+\frac{67,1}{1,22}} = \frac{1395,1}{1120} = 1,246 = 124,6\%$$

Hoặc:

$$I_p = \frac{1}{\sum \frac{d_1}{i_p}} = \frac{1}{\frac{0,0559}{1,2}+\frac{0,8960}{1,25}+\frac{0,0481}{1,22}} = \frac{1}{0,8028} = 1,246 = 124,6\%$$

❖ Chỉ số giá của Laspeyres cũng có thể biến đổi như sau:

$$I_p = \frac{\sum p_1q_0}{\sum p_0q_0} = \frac{\sum \frac{p_1}{p_0} p_0q_0}{\sum p_0q_0} = \frac{\sum i_p p_0q_0}{\sum p_0q_0} \qquad (13.8)$$

Công thức 13.8 còn gọi là chỉ số trung bình số học, vì có dạng giống số trung bình số học.

13.3.1.2.3 Chỉ số giá cả của Fisher

Chỉ số Fisher là trung bình nhân của hai chỉ số Laspeyres và Paasche.

Trong nhiều trường hợp kết quả tính toán của hai chỉ số Laspeyres và Paasche quá chênh lệch, việc sử dụng chỉ số Fisher là cần thiết.

Công thức: $$I_p = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \qquad (13.9)$$

Dựa vào kết quả tính trong bảng 13.4, chỉ số giá Fisher:

$$I_p = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} = \sqrt{\frac{1121}{900} \times \frac{1395,1}{1120}} = 1,2455 = 124,55\%$$

13.3.2 Chỉ số tổng hợp khối lượng

Chỉ số tổng hợp khối lượng được sử dụng để nghiên cứu sự thay đổi khối lượng sản phẩm một nhóm hay toàn bộ số lượng sản phẩm sản xuất hoặc tiêu thụ. Phương pháp xây dựng chỉ số tổng hợp khối lượng về căn bản giống như phương pháp xây dựng chỉ số tổng hợp giá cả. Tuy nhiên trong trường hợp này nhân tố giá đóng vai trò trọng số .

- Chỉ số tổng hợp khối lượng theo phương pháp Laspeyres:

$$I_q = \frac{\sum q_1p_0}{\sum q_0p_0} \qquad (13.10)$$

- Chỉ số tổng hợp khối lượng theo phương pháp Paasche:

$$I_q = \frac{\sum q_1p_1}{\sum q_0p_1} \qquad (13.11)$$

- Chỉ số tổng hợp khối lượng Fisher:

$$I_q = \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}} \qquad (13.12)$$

* Công thức 13.10 có thể biến đổi thành:

$$I_q = \frac{\sum q_1p_0}{\sum q_0p_0} = \frac{\sum \frac{q_1}{q_0} q_0p_0}{\sum q_0p_0} = \frac{\sum i_q q_0p_0}{\sum q_0p_0} \qquad (13.13)$$

Nếu đặt: $d_0 = \frac{q_0p_0}{\sum q_0p_0}$

Thì công thức 13.13 được viết lại: $I_q = \sum i_q d_0$ (13.14)

Công thức 13.14 còn gọi là chỉ số trung bình số học.

* Cũng như vậy công thức 13.11 có thể biến đổi thành:

$$I_q = \frac{\sum q_1p_1}{\sum q_0p_1} = \frac{\sum q_1p_1}{\sum \frac{q_0}{q_1} q_1p_1} = \frac{\sum q_1p_1}{\sum \frac{q_1p_1}{i_q}} \qquad (13.15)$$

Công thức 13.15 còn gọi là chỉ số trung bình điều hòa.

Trở lại ví dụ 13.4, ta tính chỉ số tổng hợp khối lượng hàng hóa tiêu thụ của cả 3 mặt hàng theo phương pháp Laspeyres:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{1120}{900} = 1,244 = 124,4\%$$

Kết quả cho thấy lượng tiêu thụ của ba mặt hàng trên năm 2000 so với năm 1995 bằng 124,4 % hay tăng 24,4 %.

Dựa vào số liệu của bảng 13.4, ta lập bảng 13.6 để tính chỉ số tổng hợp khối lượng theo công thức chỉ số trung bình số học 13.13 và 13.14

Bảng 13.6

Mặt hàng	Đơn vị tính	Giá đơn vị (1000đ)		Lượng tiêu thụ (1000đv)		p_0q_0 (tr.đ)	d_0	I_q
		1995 p_0	2000 p_1	1995 q_0	2000 q_1			
Đường	kg	5,0	6,0	10,0	13,0	50	0,0555	1,3
Vải	mét	40,0	50,0	20,0	25,0	800	0,8889	1,25
Dầu ăn	lít	10,0	12,2	5,0	5,5	50	0,0555	1,1
Cộng						900	1	

Thay số liệu trong bảng 13.6 vào công thức 13.13 và 13.14 ta có:

$$I_q = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0} = \frac{1,3 \times 50 + 1,25 \times 800 + 1,1 \times 50}{50 + 800 + 50} = \frac{1120}{900} = 1,244 = 124,4\%$$

hoặc:

$$I_q = \sum i_q d_0 = 1,3 \times 0,0555 + 1,25 \times 0,8889 + 1,1 \times 0,0555 = 1,244 = 124,4\%$$

Chỉ số tổng hợp khối lượng Laspeyres thường được sử dụng vì với cách chọn giá ở kỳ gốc làm trọng số, sẽ giúp chúng ta tính chỉ số tổng hợp khối lượng một cách nhanh chóng, vì p_0 lúc nào cũng có sẵn.

13.3 VẤN ĐỀ CHỌN QUYỀN SỐ (TRỌNG SỐ) CHO CHỈ SỐ TỔNG HỢP

Quyền số của chỉ số tổng hợp là đại lượng được cố định giống nhau ở tử số và mẫu số. Quyền số của chỉ số có tác dụng biểu hiện vai trò quan trọng của mỗi phần tử hay bộ phận trong toàn bộ tổng thể.

Trên đây chúng ta đã nghiên cứu cách xây dựng chỉ số tổng hợp giá cả và chỉ số tổng hợp khối lượng, còn các chỉ tiêu khác, chỉ số được xây dựng như thế nào? Như ta biết, các chỉ tiêu nghiên cứu được chia thành hai nhóm. Nhóm các chỉ tiêu chất lượng và nhóm các chỉ tiêu khối lượng. Chỉ tiêu giá cả nằm trong nhóm các chỉ tiêu chất lượng, nên việc xây dựng chỉ số của các chỉ tiêu còn lại cũng được thực hiện tương tự như chỉ tiêu giá cả. Cũng vậy, chỉ tiêu khối lượng sản phẩm nằm trong nhóm các chỉ tiêu khối lượng, nên việc xây dựng chỉ số của các chỉ tiêu còn lại cũng được thực hiện tương tự như chỉ tiêu khối lượng sản phẩm.

Cụ thể, khi dùng chỉ số tổng hợp để nghiên cứu biến động của chỉ tiêu chất lượng, thì quyền số thường là chỉ tiêu khối lượng có liên quan và cố định ở kỳ nghiên cứu. Còn để nghiên cứu biến động của chỉ tiêu khối lượng, thì quyền số thường là chỉ tiêu chất lượng có liên quan và cố định ở kỳ gốc.

- Các chỉ số tổng hợp chỉ tiêu chất lượng:

Chỉ số tổng hợp giá cả: $$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

Chỉ số tổng hợp giá thành: $$I_z = \frac{\sum z_1 q_1}{\sum z_0 q_1}$$

Chỉ số tổng hợp NSLĐ: $$I_w = \frac{\sum W_1 T_1}{\sum W_0 T_1}$$

Chỉ số tổng hợp NS thu hoạch: $$I_N = \frac{\sum N_1 D_1}{\sum N_0 D_1}$$

v.v...

- Các chỉ số tổng hợp chỉ tiêu khối lượng:

Chỉ số tổng hợp KLSP SX: $$I_q = \frac{\sum q_1 z_0}{\sum q_0 z_0}$$

Chỉ số tổng hợp KLSP tiêu thụ: $$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}$$

Chỉ số tổng hợp số công nhân: $$I_T = \frac{\sum T_1 W_0}{\sum T_0 W_0}$$

Chỉ số tổng hợp diện tích: $$I_D = \frac{\sum D_1 N_0}{\sum D_0 N_0}$$

v.v...

13.4 CHỈ SỐ KHÔNG GIAN

Chỉ số không gian là chỉ số so sánh các hiện tượng cùng loại nhưng qua điều kiện không gian khác nhau. Ví dụ, So sánh lượng hàng bán ra và giá cả các mặt hàng ở hai thị trường TP.HCM và Hà Nội.

13.5.1. Chỉ số tổng hợp khối lượng không gian

Khi tính chỉ số tổng hợp khối lượng theo không gian có thể dùng giá cố định (Giá so sánh) do Nhà nước ban hành hoặc giá trung bình từng mặt hàng ở hai thị trường làm trọng số.

Công thức chỉ số tổng hợp khối lượng ở hai thị trường A và B:

$$I_{q(A/B)} = \frac{\sum q_A p_c}{\sum q_B p_c} \qquad (13.16)$$

Trong đó: p_c là giá so sánh các mặt hàng.

Hoặc: $$I_{q(A/B)} = \frac{\sum q_A \bar{p}}{\sum q_B \bar{p}} \text{ trong đó: } \bar{p} = \frac{p_A q_A + p_B q_B}{q_A + q_B} \qquad (13.17)$$

13.5.2. Chỉ số tổng hợp giá cả không gian

Khi tính chỉ số tổng hợp giá cả không gian, thì trọng số là khối lượng sản phẩm (hàng hóa) cùng loại của hai đơn vị hoặc hai thị trường cần so sánh.

Công thức chỉ số tổng hợp giá cả ở hai thị trường A và B:

$$I_{p(A/B)} = \frac{\sum p_A Q}{\sum p_B Q} \qquad (13.18)$$

Trong đó: $Q = q_A + q_B$: Khối lượng sản phẩm (hàng hóa) cùng loại của hai thị trường A và B.

Ví dụ: Tài liệu về giá cả và khối lượng hàng tiêu thụ của ba mặt hàng tại hai thành phố X và Y trong cùng một kỳ như sau :

Bảng 13.7

Loại hàng hóa	Thành phố X		Thành phố Y	
	Giá đơn vị (1000đ)	Lượng hàngtiêu thụ (tấn)	Giá đơn vị (1000đ)	Lượng hàng tiêu thụ (tấn)
A	5,0	250	4,8	262
B	4,6	430	4,9	392
C	6,9	187	6,8	213

Giá trung bình đơn vị của từng mặt hàng:

Mặt hàng A: $\bar{p}_A = \dfrac{5 \times 250 + 4{,}8 \times 262}{250 + 262} = 4{,}9$ ngàn đồng

Mặt hàng B: $\bar{p}_B = \dfrac{4{,}6 \times 430 + 4{,}9 \times 392}{430 + 392} = 4{,}7$ ngàn đồng

Mặt hàng C: $\bar{p}_C = \dfrac{6{,}9 \times 187 + 6{,}8 \times 213}{187 + 213} = 6{,}8$ ngàn đồng

Chỉ số tổng hợp khối lượng hàng hoá tiêu thụ thành phố X so với thành phố Y:

$$I_{q(X/Y)} = \frac{\sum q_X \bar{p}}{\sum q_Y \bar{p}} = \frac{(250 \times 4{,}9) + (430 \times 4{,}7) + (187 \times 6{,}8)}{(262 \times 4{,}9) + (392 \times 4{,}7) + (213 \times 6{,}8)} = 0{,}9875$$

$$= 98{,}75\%$$

Lượng hàng tiêu thụ ba mặt hàng trên tại thành phố X so với thành phố Y bằng 98,75%, ít hơn 1,25%.

Chỉ số tổng hợp giá cả hàng hoá tiêu thụ thành phố X so với thành phố Y:

$$I_{p(X/Y)} = \frac{\sum p_X Q}{\sum p_Y Q} = \frac{5(250 + 262) + 4{,}6(430 + 392) + 6{,}9(187 + 213)}{4{,}8(250 + 262) + 4{,}9(430 + 392) + 6{,}8(187 + 213)}$$

$$= 0{,}9887 = 98{,}87\%$$

Giá của cả ba mặt hàng nói trên, tại thành phố X thấp hơn thành phố Y là 1,13%.

13.6 HỆ THỐNG CHỈ SỐ

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 nhân tố đến sự thay đổi của một chỉ tiêu kinh tế tổng hợp bằng cách kết hợp các chỉ số nhân tố lại thành hệ thống chỉ số.

Cơ sở để hình thành hệ thống chỉ số là mối liên hệ thực tế giữa các chỉ tiêu, thường được biểu hiện dưới dạng các công thức như:

Giá thành đơn vị sản phẩm	×	Khối lượng sản phẩm	=	Chi phí sản xuất

Ta có hệ thống chỉ số tương ứng:

Chỉ số giá thành	×	Chỉ số khối lượng SP	=	Chỉ số chi phí sản xuất

Hoặc:

Giá bán lẻ đơn vị	×	Lượng hàng tiêu thụ	=	Mức tiêu thụ hàng hoá

Ta có hệ thống chỉ số tương ứng:

Chỉ số giá cả	×	Chỉ số lượng hàng tiêu thụ	=	Chỉ số mức tiêu thụ hàng hoá

Trở lại ví dụ của bảng 13.4, giả sử ta cần phân tích biến động tổng mức tiêu thụ hàng hoá qua hai kỳ nghiên cứu do ảnh hưởng bởi nhân tố giá cả và lượng hàng hóa tiêu thụ, Ta có hệ thống chỉ số:

$$I_{pq} = I_p \times I_q$$

Trong đó:

I_{pq} : Chỉ số tổng mức tiêu thụ hàng hóa.

I_p : Chỉ số giá tính theo phương pháp của Paasche.

I_q : Chỉ số khối lượng tính theo phương pháp của Laspeyres.

$$\frac{\sum p_1q_1}{\sum p_0q_0} = \frac{\sum p_1q_1}{\sum p_0q_1} \times \frac{\sum p_0q_1}{\sum p_0q_0} \qquad (13.19)$$

Thay số liệu trong bảng 13.4 vào hệ thống chỉ số 13.19:

$$\frac{1395,1}{900} = \frac{1395,1}{1120} \times \frac{1120}{900}$$

$$1,55 = 1,246 \times 1,244$$

Số tuyệt đối:

$$\sum p_1q_1 - \sum p_0q_0 = (\sum p_1q_1 - \sum p_0q_1) + (\sum p_0q_1 - \sum p_0q_0)$$

$$1395,1 - 900 = (1395,1 - 1120) + (1120 - 900)$$

$$495,1 = 275,1 + 220 \text{ (triệu đồng)}$$

Số tương đối:

$$\frac{\sum p_1q_1 - \sum p_0q_0}{\sum p_0q_0} = \frac{(\sum p_1q_1 - \sum p_0q_1)}{\sum p_0q_0} + \frac{(\sum p_0q_1 - \sum p_0q_0)}{\sum p_0q_0}$$

$$\frac{495,1}{900} = \frac{275,1}{900} + \frac{220}{900}$$

$$55\% = 30,6\% + 24,4\%$$

Nhận xét:

Tổng mức tiêu thụ hàng hoá năm 2000 so với năm 1995 bằng 155%, tức là tăng 55%, tương ứng tăng 495,1 triệu đồng là do ảnh hưởng của hai nhân tố:

- Do giá cả các mặt hàng nói chung năm 2000 so với năm 1995 tăng 24,6%, làm cho tổng mức tiêu thụ hàng hoá tăng 30,6%, tương ứng tăng 275,1 triệu đồng.
- Do khối lượng các mặt hàng tiêu thụ năm 2000 so với năm 1995 tăng 24,4%, làm cho tổng mức tiêu thụ hàng hoá tăng 24,4%, tương ứng tăng 220 triệu đồng.

* Hệ thống chỉ số ngoài việc được sử dụng để phân tích biến động một hiện tượng do ảnh hưởng bởi các nhân tố cấu thành, trong nhiều trường hợp người ta lợi dụng 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.

❖ Hệ thống chỉ số phân tích biến động của chỉ tiêu trung bình

Chỉ tiêu trung bình chịu ảnh hưởng biến động của hai nhân tố: tiêu thức nghiên cứu và kết cấu tổng thể. Ví dụ, biến động của tiền lương trung bình của công nhân trong xí nghiệp là do biến động của bản thân tiền lương (tiêu thức nghiên cứu) và biến động kết cấu công nhân (kết cấu tổng thể) có các mức lương khác nhau.

Một cách tổng quát, dựa vào công thức số trung bình số học: $\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$,

ta thấy giá trị $\bar{x}$ lớn hay bé, phụ thuộc vào hai nhân tố:

- Giá trị của các x_i lớn hay bé, làm cho $\bar{x}$ lớn hay bé.
- $\frac{f_i}{\sum f_i}$ thay đổi, làm cho $\bar{x}$ thay đổi theo. Cụ thể là giá trị $\bar{x}$ có xu hướng nghiêng về lượng biến x_i nào có $\frac{f_i}{\sum f_i}$ chiếm tỷ trọng lớn.

Hệ thống chỉ số được sử dụng để phân tích ảnh hưởng biến động của các nhân tố đến biến động của chỉ tiêu trung bình.

Ta sử dụng các ký hiệu sau:

x_1 và x_0 : Lượng biến của tiêu thức kỳ nghiên cứu và kỳ gốc.

$\bar{x}_1$ và $\bar{x}_0$: Số trung bình kỳ nghiên cứu và kỳ gốc.

f_1 và f_0 : Số đơn vị tổng thể kỳ nghiên cứu và kỳ gốc.

Ta có hệ thống chỉ số:

$$\frac{\frac{\sum x_1 f_1}{\sum f_1}}{\frac{\sum x_0 f_0}{\sum f_0}} = \frac{\frac{\sum x_1 f_1}{\sum f_1}}{\frac{\sum x_0 f_1}{\sum f_1}} \times \frac{\frac{\sum x_0 f_1}{\sum f_1}}{\frac{\sum x_0 f_0}{\sum f_0}} \qquad (13.20)$$

Công thức 13.20 viết gọn thành:

$$\underset{(1)}{\frac{\bar{x}_1}{\bar{x}_0}} = \underset{(2)}{\frac{\bar{x}_1}{\bar{x}_{01}}} \times \underset{(3)}{\frac{\bar{x}_{01}}{\bar{x}_0}} \qquad (13.21)$$

(1) Chỉ số cấu thành khả biến, nêu lên biến động của chỉ tiêu trung bình giữa hai kỳ nghiên cứu.

(2) Chỉ số cấu thành cố định, nêu lên biến động của chỉ tiêu trung bình do ảnh hưởng của riêng tiêu thức nghiên cứu.

(3) Chỉ số ảnh hưởng kết cấu, nêu lên biến động của chỉ tiêu trung bình do ảnh hưởng của riêng kết cấu tổng thể.

Ví dụ: Một xí nghiệp có 3 phân xưởng cùng sản xuất một loại sản phẩm, số liệu cho trong Bảng 13.8

Bảng 13.8

Phân xưởng	Kỳ gốc		Kỳ nghiên cứu	
	Sản lượng(cái) (q_0)	Giá thành đv (1000đ)(Z_0)	Sản lượng(cái) (q_1)	Giá thành đv (1000đ)(Z_1)
A	1000	10	8000	9
B	2500	12	3000	11,5
C	4500	13	1000	12,5
Tổng	8000		12000	

Để phân tích biến động của giá thành trung bình do ảnh hưởng bởi các nhân tố có liên quan, hệ thống chỉ số 13.21 được viết lại:

$$\frac{\overline{Z}_1}{\overline{Z}_0} = \frac{\overline{Z}_1}{\overline{Z}_{01}} \times \frac{\overline{Z}_{01}}{\overline{Z}_0}$$

Với:

$$\bar{z}_1 = \frac{\sum z_1 q_1}{\sum q_1} = \frac{119000}{12000} = 9{,}92 \text{ ng.đ}$$

$$\bar{z}_0 = \frac{\sum z_0 q_0}{\sum q_0} = \frac{98500}{8000} = 12{,}31 \text{ ng.đ}$$

$$\bar{z}_{01} = \frac{\sum z_0 q_1}{\sum q_1} = \frac{129000}{12000} = 10{,}75 \text{ ng.đ}$$

Thay số vào hệ thống chỉ số:

$$\frac{9{,}92}{12{,}31} = \frac{9{,}92}{10{,}75} \times \frac{10{,}75}{12{,}31}$$

$$0{,}806 = 0{,}9228 \times 0{,}873$$

Số tuyệt đối:

$$(\bar{z}_1 - \bar{z}_0) = (\bar{z}_1 - \bar{z}_{01}) + (\bar{z}_{01} - \bar{z}_0)$$

$$9{,}92 - 12{,}31 = (9{,}92 - 10{,}75) + (10{,}75 - 12{,}31)$$

$$-2{,}39 = (-0{,}83) + (-1{,}56)(\text{ngàn đồng})$$

Số tương đối:

$$\frac{(\bar{z}_1 - \bar{z}_0)}{\bar{z}_0} = \frac{(\bar{z}_1 - \bar{z}_{01})}{\bar{z}_0} + \frac{(\bar{z}_{01} - \bar{z}_0)}{\bar{z}_0}$$

$$\frac{-2,39}{12,31} = \frac{-0,83}{12,31} + \frac{-1,56}{12.31}$$

$$-0,194 = (-0,067) + (-0,127)$$

$$-19,4\% = (-6,7\%) + (-12,7\%)$$

Nhận xét:

Giá thành trung bình kỳ nghiên cứu so với kỳ gốc bằng 80,6%, tức là giảm 19,4%, hay giảm 2,39 ngàn đồng, là do ảnh hưởng bởi hai nhân tố:

- Do giá thành nói chung của các phân xưởng giảm 7,72% làm cho giá thành trung bình giảm 6,7%, hay giảm 0,83 ngàn đồng.
- Do kết cấu sản lượng thay đổi, làm cho giá thành trung bình giảm 12,7%, hay giảm 1,56 ngàn đồng.

❖ **Hệ thống chỉ số phân tích biến động tổng lượng biến tiêu thức có sử dụng chỉ tiêu trung bình**

Trong nhiều trường hợp chỉ tiêu trung bình có quan hệ với tổng lượng biến tiêu thức. Nó là một nhân tố cấu thành tổng lượng biến tiêu thức. Ví dụ:

Tổng sản lượng	=	NSLĐ trung bình 1 công nhân	×	Tổng số công nhân
Tổng chi phí sản xuất	=	Giá thành trung bình 1 ĐVSP	×	Tổng sản phẩm

Ta xây dựng được hệ thống chỉ số tương ứng:

Chỉ số tổng sản lượng	=	Chỉsố NSLĐ trung bình	×	Chỉ số tổng số công nhân
Chỉ số tổng chi phí sản xuất	=	Chỉ số Giá thành trung bình 1 ĐVSP	×	Chỉ số tổng sản phẩm

Tổng quát: $M = \bar{x} \times \sum f$

Hệ thống chỉ số: $$\frac{\bar{x}_1\sum f_1}{\bar{x}_0\sum f_0} = \frac{\bar{x}_1\sum f_1}{\bar{x}_0\sum f_1} \times \frac{\bar{x}_0\sum f_1}{\bar{x}_0\sum f_0} \qquad (13.22)$$

Dựa vào ví dụ cho trong bảng 13.8, sử dụng hệ thống chỉ số để phân tích biến động tổng chi phí sản xuất có liên quan đến biến động của giá thành trung bình.

Công thức 13.22 được viết lại:

$$\frac{\bar{z}_1\sum q_1}{\bar{z}_0\sum q_0} = \frac{\bar{z}_1\sum q_1}{\bar{z}_0\sum q_1} \times \frac{\bar{z}_0\sum q_1}{\bar{z}_0\sum q_0}$$

Thay số liệu trong bảng 13.8 vào công thức 13.22 ta có:

$$\frac{9{,}92\times 12000}{12{,}31\times 8000} = \frac{9{,}92\times 12000}{12{,}31\times 12000} \times \frac{12{,}31\times 12000}{12{,}31\times 8000}$$

$$1{,}2088 = 0{,}806\times 1{,}5$$

Số tuyệt đối:

$$\bar{z}_1\sum q_1 - \bar{z}_0\sum q_0 = (\bar{z}_1 - \bar{z}_0)\sum q_1 + (\sum q_1 - \sum q_0)\bar{z}_0$$

$$9{,}92.12000 - 12{,}31.8000 = (9{,}92 - 12{,}31)12000 + (12000 - 8000)12{,}31$$

$$20560 = -28680 + 49240 \text{ (ngàn đồng)}$$

Số tương đối:

$$\frac{\bar{z}_1\sum q_1 - \bar{z}_0\sum q_0}{\bar{z}_0\sum q_0} = \frac{(\bar{z}_1 - \bar{z}_0)\sum q_1}{\bar{z}_0\sum q_0} + \frac{(\sum q_1 - \sum q_0)\bar{z}_0}{\bar{z}_0\sum q_0}$$

$$\frac{20560}{98480} = \frac{-28680}{98480} + \frac{49240}{98480}$$

$$0{,}2088 = -0{,}2912 + 0{,}5$$

$$20{,}88\% = -29{,}12\% + 50\%$$

Nhận xét:

Tổng chi phí kỳ nghiên cứu so với kỳ gốc bằng 120,88%, tức là tăng 20,88%, tương ứng tăng 20560 ngàn đồng, là do ảnh hưởng bởi hai nhân tố:

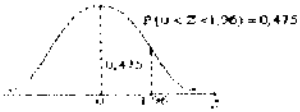
- Do giá thành trung bình giảm 19,4% làm cho tổng chi phí giảm 29,12% hay giảm 28680 ngàn đồng.
- Do sản lượng tăng 50%, làm cho tổng chi phí tăng 50% hay tăng 49240 ngàn đồng.

BẢNG 1 **BẢNG GIÁ TRỊ HÀM MẬT ĐỘ** $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

Phần thập phân của z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3925	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3725	0.3712	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.3467	0.3448	0.3429	0.3410	0.3391	0.3372	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2897	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1.0	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1456	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.1040	0.1023	0.1006	0.0989	0.0973	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551
2.0	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0194	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0088	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.0061
2.9	0.0060	0.0058	0.0056	0.0055	0.0053	0.0051	0.0050	0.0048	0.0047	0.0046
3.0	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006
3.6	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004
3.7	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003
3.8	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002
3.9	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001

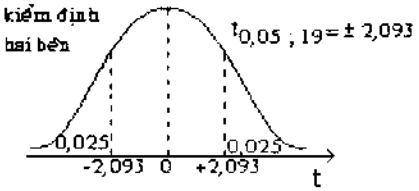
BẢNG 2 **PHÂN PHỐI CHUẨN**

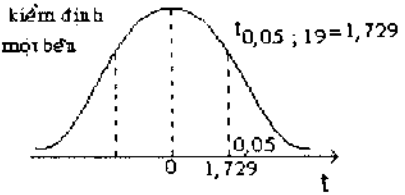


Phần thập phân của z

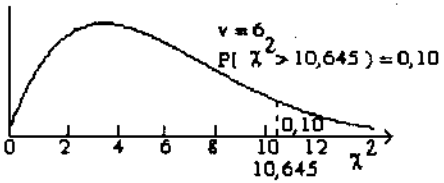
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	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 3 PHÂN PHỐI STUDENT





Bậc tự do (v)	$t_{0,10}$	$t_{0,05}$	$t_{0,025}$	$t_{0,01}$	$t_{0,005}$
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

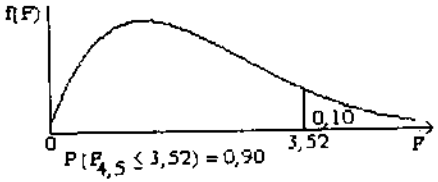


Bậc tự do (V)	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.900}$
1	0.000039	0.000157	0.000982	0.003932	0.015791
2	0.010025	0.020100	0.050636	0.102586	0.210721
3	0.071723	0.114832	0.215795	0.351846	0.584375
4	0.206984	0.297107	0.484419	0.710724	1.063624
5	0.411751	0.554297	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	14.953464	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 4 (Tiếp theo)

Bậc tự do (V)	$\chi^2_{0.100}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$	$\chi^2_{0.005}$
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	79.4893
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 5 **PHÂN PHỐI F ($\alpha = 0.1$)**



Bậc tự do của tử số (V_1)

Bậc tự do của mẫu số (V_2)	1	2	3	4	5	6	7	8	9
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63

ẢNG 5 (tiếp theo) **($\alpha = 0.1$)**

Bậc tự do của tử số (V_1)

Bậc tự do của mẫu số (V_2)	10	12	15	20	24	30	40	60	120	∞
1	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

BẢNG 5 (Tiếp theo) **($\alpha = 0.05$)**

Bậc tự do của tử số (V_1)

Bậc tự do của mẫu số (V_2)	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

BẢNG 5 (Tiếp theo) **($\alpha = 0.05$)**

Bậc tự do của tử số (V_1)

Bậc tự do của mẫu số (V_2)	10	12	15	20	24	30	40	60	120	∞
1	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.5
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

BẢNG 5 (Tiếp theo) **($\alpha = 0.025$)**

Bậc tự do của mẫu số (V_2)	Bậc tự do của tử số (V_1) 1	2	3	4	5	6	7	8	9
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.6	963.3
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11

BẢNG 5 (Tiếp theo) **($\alpha = 0.025$)**

Bậc tự do của tử số (V_1)

Bậc tự do của mẫu số (V_2)	10	12	15	20	24	30	40	60	120	∞
1	968.6	976.7	984.9	993.1	997.3	1001	1006	1010	1014	1018
2	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.14
8	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

BẢNG 5 (Tiếp theo) **($\alpha = 0.01$)**

Bậc tự do của tử số (V_1)

Bậc tự do của mẫu số (V_2)	1	2	3	4	5	6	7	8	9
1	4052	4999	5404	5624	5764	5859	5928	5981	6022
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

BẢNG 5 (Tiếp theo) **($\alpha = 0.01$)**

Bậc tự do của tử số (V_1)

Bậc tự do của mẫu số (V_2)	10	12	15	20	24	30	40	60	120	∞
1	6056	6107	6157	6209	6235	6260	6286	6313	6339	6336
2	99.40	99.42	99.43	99.45	99.46	99.47	99.48	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

BẢNG 6 **PHÂN PHỐI WILCOXON**

n	α				
	0.005	0.010	0.025	0.050	0.100
4	0	0	0	0	1
5	0	0	0	1	3
6	0	0	1	3	4
7	0	1	3	4	6
8	1	2	4	6	9
9	2	4	6	9	11
10	4	6	9	11	15
11	6	8	11	14	18
12	8	10	14	18	22
13	10	13	18	22	27
14	13	16	22	26	32
15	16	20	26	31	37
16	20	24	30	36	43
17	24	28	35	42	49
18	28	33	41	48	56
19	33	38	47	54	63
20	38	44	53	61	70

BẢNG 7 **PHÂN PHỐI SPEARMAN**

n	α			
	0.050	0.025	0.010	0.005
5	0.900	-	-	-
6	0.829	0.886	0.943	-
7	0.714	0.786	0.893	-
8	0.643	0.738	0.833	0.881
9	0.600	0.683	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.523	0.623	0.736	0.818
12	0.497	0.591	0.703	0.780
13	0.475	0.566	0.673	0.745
14	0.457	0.545	0.646	0.716
15	0.441	0.525	0.623	0.689
16	0.425	0.507	0.601	0.666
17	0.412	0.490	0.582	0.645
18	0.399	0.476	0.564	0.625
19	0.388	0.462	0.549	0.608
20	0.377	0.450	0.534	0.591
21	0.368	0.438	0.521	0.576
22	0.359	0.428	0.508	0.562
23	0.351	0.418	0.496	0.549
24	0.343	0.409	0.485	0.537
25	0.336	0.400	0.475	0.526
26	0.329	0.392	0.465	0.515
27	0.323	0.385	0.456	0.505
28	0.317	0.377	0.448	0.496
29	0.311	0.370	0.440	0.487
30	0.305	0.364	0.432	0.478

BẢNG 8 PHÂN PHỐI TUKEY (Studentized Range Distribution) ($\alpha = 0,05$)

n-r	r								
	2	3	4	5	6	7	8	9	10
1	18.0	27.0	32.8	37.1	40.4	43.1	45.4	47.4	49.1
2	6.08	8.33	9.80	10.9	11.7	12.4	13.0	13.5	14.0
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	4.45	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	4.55	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

BẢNG 8 (tiếp theo) **($\alpha = 0,05$)**

n-r	r									
	11	12	13	14	15	16	17	18	19	20
1	50.6	52.0	53.2	54.3	55.4	56.3	57.2	58.0	58.8	59.6
2	14.4	14.7	15.1	15.4	15.7	15.9	16.1	16.4	16.6	16.8
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 8 (tiếp theo) **($\alpha = 0,01$)**

n-r	r								
	2	3	4	5	6	7	8	9	10
1	90.0	135	164	186	202	216	227	237	246
2	14.0	19.0	22.3	24.7	26.6	28.2	29.5	30.7	31.7
3	8.26	10.6	12.2	13.3	14.2	15.0	15.6	16.2	16.7
4	6.51	8.12	9.17	9.96	10.6	11.1	11.5	11.9	12.3
5	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.2
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

BẢNG 8 (tiếp theo) **($\alpha = 0,01$)**

n-r	r									
	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	6.32
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

TÀI LIỆU THAM KHẢO

1. Paul Newbold, *Statistics for Business and Economics*, Prentice Hall International, Inc., 1995
2. Keller, Warrack, *Statistics for Management and Economics*, Brooks, Cole publishing Company, 1999
3. Patrick W.Shannon, David F.Groebner, Phillip C.Fry, Kent D.Smith, *A Course in Business Statistics*, Prentice Hall, Inc., 2002
4. Allen L.Webster, *Applied Statistics for Business and Economics*, The McGraw-Hill Companies, Inc.,1998
5. Amir D. Aczel, *Complete Business Statistics*, Irwin, 1993
6. Mark L. Berenson, David M. Levine, *Basic Business Statistics*, Prentice Hall 1986
7. Richard I. Levin, *Statistics for Management*, Prentice Hall, 1989
8. Fredrerick E. Croxton, Dudley J. Cowden, Sidney Klein, General Applied Statistics, Prentice Hall of India, New Dehli, 1988
9. Neuman William Lawrence, *Social Research Methods, Qualitative and Quantitative Approaches*, Allyn & Bacon, 2000
10. Trần Văn Thắng và các tác giả khác, *Giáo trình Lý thuyết thống kê*, NXB Thống Kê, 1998
11. Nguyễn Ngọc Kiểng, *Thống kê học trong nghiên cứu khoa học*, NXB Giáo Dục, 1996
12. Võ Thị Thanh Lộc, *Thống kê ứng dụng và dự báo trong kinh doanh và kinh tế*, NXB Thống Kê, 1998
13. Đặng Hấn, *Xác suất thống kê*, NXB Thống kê, 1996
14. Hoàng Ngọc Nhậm, Giáo trình Lý thuyết xác suất và thống kê-toán, NXB Thống kê, 2003
15. Trần Bá Nhẫn, Đinh Thái Hoàng, *Lý thuyết thống kê ứng dụng trong quản trị kinh doanh và nghiên cứu kinh tế*, NXB Thống Kê, 1998
16. Tập thể tác giả, *Từ điển thống kê*, Tổng cục thống kê, 1977
17. Võ Văn Huy, Võ Thị Lan, Hoàng Trọng, *Ứng dụng SPSS For Windows để xử lý và phân tích dữ kiện nghiên cứu*, NXB Khoa Học và Kỹ Thuật, 1997
18. Hoàng Trọng, *Xử lý dữ liệu nghiên cứu với SPSS for Windows*, NXB Thống Kê, 2002

GIÁO TRÌNH

LÝ THUYẾT THỐNG KÊ

ỨNG DỤNG TRONG QUẢN TRỊ VÀ KINH TẾ

(STATISTICS FOR MANAGEMENT AND ECONOMICS)

Chịu trách nhiệm xuất bản: Cát Văn Thành
Trình bày: Hà Văn Sơn
Biên tập: Hà Văn Sơn

NHÀ XUẤT BẢN THỐNG KÊ

In 2.000 cuốn, khổ 16 x 24cm tại Xí nghiệp In Machinco - Số 21 Bùi Thị Xuân, Q.1, TP. HCM. Giấy phép xuất bản số: 06-205/XB-QLXB do Cục Xuất Bản cấp ngày 03.03.2003.
In xong và nộp lưu chiểu tháng 01/2004.