BIÊN SOẠN : TS. MAI VĂN NAM

GIÁO TRÌNH

NGUYÊN LÝ THỐNG KÊ KINH TẾ

NHÀ XUẤT BẢN VĂN HÓA THÔNG TIN

MỤC LỤC

Mục lục **Trang**

Mục lục **Trang**

LỜI NÓI ĐẦU

Thống kê là một ngành khoa học có vai trò quan trọng trong hầu hết các lĩnh vực kinh tế xã hội. Nguyên lý thống kê kinh tế, lý thuyết thống kê theo hướng ứng dụng trong lĩnh vực kinh tế và quản trị kinh doanh, là công cụ không thể thiếu được trong hoạt động nghiên cứu và quản lý. Nguyên lý thống kê kinh tế đã trở thành một môn học cơ sở trong hầu hết các ngành đào tạo thuộc khối kinh tế.

Trong bối cảnh đào tạo đại học theo tín chỉ hóa, thời gian lên lớp được giới hạn và sinh viên được khuyến khích tự tham khảo tài liệu và tự học có hướng dẫn của giảng viên. Nhu cầu về một tài liệu giảng dạy và học tập môn nguyên lý thống kê kinh tế, vừa phù hợp với chương trình đào tạo theo tín chỉ, vừa nhất quán với các môn học định lượng trong chương trình đào tạo bậc đại học là cần thiết. Giáo trình này được biên soạn nhằm mục đích giúp cho bạn đọc am hiểu các vấn đề về lý thuyết, chuẩn bị cho những tiết thực hành trên máy tính có hiệu quả, là cơ sở quan trọng cho người học tiếp cận các môn học chuyên ngành kinh tế.

Để đáp ứng nhu cầu trên, Tác giả thực hiện biên soạn quyển sách giáo trình thống kê kinh tế. Tài liệu này được viết trên cơ sở bạn đọc đã có kiến thức về xác suất thống kê toán, cho nên cuốn sách không đi sâu về mặt toán học mà chú trọng đến kết quả và ứng dụng trong lĩnh vực kinh tế và quản trị kinh doanh với các ví dụ gần gũi với thực tế.

Với kinh nghiệm giảng dạy được tích lũy qua nhiều năm, tham gia thực hiện các đề tài nghiên cứu trong lĩnh vực kinh tế xã hội; cùng với sự phối hợp và hỗ trợ của đồng nghiệp, đặc biệt của ThS. Nguyễn Ngọc Lam, Tác giả hy vọng quyển sách này đáp ứng được nhu cầu học tập của các sinh viên và nhu cầu tham khảo của các bạn đọc có quan tâm đến nguyên lý thống kê kinh tế trong nghiên cứu kinh tế xã hội.

Trong quá trình biên soạn chắc chắn không tránh khỏi những thiếu sót, Tác giả rất mong nhận được những ý kiến đóng góp quí báu của bạn đọc để lần tái bản sau quyển sách được hoàn thiện hơn. Xin chân thành cám ơn.

Tác giả
TS.Mai Văn Nam

PHẦN I
GIỚI THIỆU MÔN HỌC

I. NGUỒN GỐC MÔN HỌC

Nếu thống kê được hiểu theo nghĩa thông thường thì ngay từ thời cổ đại con người đã đã chú ý đến việc này thông qua việc ghi chép đơn giản.

Cuối thế kỷ XVII, lực lượng sản xuất phát triển mạnh mẽ làm cho phương thức sản xuất của chủ nghĩa tư bản ra đời. Kinh tế hàng hóa phát triển dẫn đến các ngành sản xuất riêng biệt tăng thêm, phân công lao động xã hội ngày càng phát triển. Tính chất xã hội của sản xuất ngày càng cao, thị trường được mở rộng không chỉ trong một nước mà toàn thế giới. Để phục vụ cho mục đích kinh tế, chính trị và quân sự nhà nước tư bản và các chủ tư bản cần rất nhiều thông tin thường xuyên về thị trường, giá cả, sản xuất, nguyên liệu, dân số,... Do đó, công tác thống kê phát triển nhanh chóng. Chúng ta có thể đưa ra 3 nhóm tác giả được gọi là những người khai sáng cho ngành khoa học thống kê:

- Những người đầu tiên đưa ngành khoa học thống kê đi vào thực tiễn, đại diện cho những tác giả này là nhà kinh tế học người Đức H.Conhring (1606 - 1681), năm 1660 ông đã giảng dạy tại trường đại học Halmsted về phương pháp nghiên cứu hiện tượng xã hội dựa vào số liệu điều tra cụ thể.

- Với những thành quả của người đi trước, bổ sung hoàn chỉnh thành môn học chính thống, đại diện là William Petty, một nhà kinh tế học của người Anh, là tác giả cuốn "Số học chính trị" xuất bản năm 1682, một số tác phẩm có tính chất phân tích thống kê đầu tiên ra đời.

- Thống kê được gọi với nhiều tên khác nhau thời bấy giờ, sau đó năm 1759 một giáo sư người Đức, Achenwall (1719-1772) lần đầu tiên dùng danh từ "Statistics" (một thuật ngữ gốc La tinh "Status", có nghĩa là Nhà nước hoặc trạng thái của hiện tượng) - sau này người ta dịch ra là "Thống kê".

Kể từ đó, thống kê có sự phát triển rất mạnh mẽ và ngày càng hoàn thiện, gắn liền với nhiều nhà toán học - thống kê học nổi tiếng như: M.V.Lomonoxop (nga, 1711-1765), Laplace (Pháp, 1749-1827), I.Fisher, W.M.Pearsons,...

II. THỐNG KÊ LÀ GÌ?

1. Định nghĩa

Thống kê là một hệ thống các phương pháp bao gồm thu thập, tổng hợp, trình bày số liệu, tính toán các đặc trưng của đối tượng nghiên cứu nhằm phục vụ cho quá trình phân tích, dự đoán và ra quyết định.

2. Chức năng của thống kê

Thống kê thường được phân thành 2 lĩnh vực:

- Thống kê mô tả *(Descriptive statistics)*: là các phương pháp có liên quan đến việc thu thập số liệu, tóm tắt, trình bày, tính toán và mô tả các đặc trưng khác nhau để phản ánh một cách tổng quát đối tượng nghiên cứu.

- Thống kê suy luận *(Inferential statistics)*: là bao gồm các phương pháp ước lượng các đặc trưng của tổng thể, phân tích mối liên hệ giữa các hiện tượng nghiên cứu, dự đoán hoặc ra quyết định trên cơ sở thông tin thu thập từ kết quả quan sát mẫu.

3. Phương pháp thống kê

- Thu thập và xử lý số liệu:

Số liệu thu thập thường rất nhiều và hỗn độn, các dữ liệu đó chưa đáp ứng cho quá trình nghiên cứu. Để có hình ảnh tổng quát về tổng thể nghiên cứu, số liệu thu thập phải được xử lý tổng hợp, trình bày, tính toán các số đo; kết quả có được sẽ giúp khái quát được đặc trưng của tổng thể.

- Nghiên cứu các hiện tượng trong hoàn cảnh không chắc chắn:

Trong thực tế, có nhiều hiện tượng mà thông tin liên quan đến đối tượng nghiên cứu không đầy đủ mặc dù người nghiên cứu đã có sự cố gắng. Ví dụ như nghiên cứu về nhu cầu của thị trường về một sản phẩm ở mức độ nào, tình trạng của nền kinh tế ra sao, để nắm được các thông tin này một cách rõ ràng quả là một điều không chắc chắn.

- Điều tra chọn mẫu:

Trong một số trường hợp để nghiên cứu toàn bộ tất cả các quan sát của tổng thể là một điều không hiệu quả, xét cả về tính kinh tế (chi phí, thời gian) và tính kịp thời, hoặc không thực hiện được. Chính điều này đã đặt ra cho thống kê xây dựng các phương pháp chỉ cần nghiên cứu một bộ phận của tổng thể mà có thể suy luận cho hiện tượng tổng quát mà vẫn đảm bảo độ tin cậy cho phép, đó là phương pháp điều tra chọn mẫu.

- Nghiên cứu mối liên hệ giữa các hiện tượng:

Giữa các hiện tượng nghiên cứu thường có mối liên hệ với nhau. Ví dụ như mối liên hệ giữa chi tiêu và thu nhập; mối liên hệ giữa lượng vốn vay và các yếu tố tác động đến lượng vốn vay như chi tiêu, thu nhập, trình độ học vấn; mối liên hệ giữa tốc độ phát triển với tốc độ phát triển của các ngành, lạm phát, tốc độ phát triển dân số,…Sự hiểu biết về mối liên hệ giữa các hiện tượng rất có ý nghĩa, phục vụ cho quá trình dự đoán

- Dự đoán:

Dự đoán là một công việc cần thiết trong tất cả các lĩnh vực hoạt động. Trong hoạt động dự đoán người ta có thể chia ra thành nhiều loại:

(1). Dự đoán dựa vào định lượng và dựa vào định tính. Tuy nhiên, trong thống kê chúng ta chủ yếu xem xét về mặt định lượng với mục đích cung cấp cho những nhà quản lý có cái nhìn mang tính khoa học hơn và cụ thể hơn trước khi ra quyết định phù hợp.

(2). Dự đoán dựa vào nội suy và dựa vào ngoại suy.

- Dự đoán nội suy là chúng ta dựa vào bản chất của hiện tượng để suy luận, ví dụ như chúng ta xem xét một liên hệ giữa lượng sản phẩm sản xuất ra phụ thuộc các yếu tố đầu vào như vốn, lao động và trình độ khoa học kỹ thuật.

- Dự đoán dựa vào ngoại suy là chúng ta chỉ quan sát sự biến động của hiện tượng trong thực tế, tổng hợp lại thành qui luật và sử dụng qui luật này để suy luận, dự đoán sự phát triển của hiện tượng. Ví dụ như để đánh giá kết quả hoạt động của một công ty người ta xem xét kết quả hoạt động kinh doanh của họ qua nhiều năm.

Ngoài ra, người ta còn có thể phân chia dự báo thống kê ra thành nhiều loại khác.

III. CÁC KHÁI NIỆM THƯỜNG DÙNG TRONG THỐNG KÊ

1. Tổng thể thống kê (Populations)

Tổng thể thống kê là tập hợp các đơn vị cá biệt về sự vật, hiện tượng trên cơ sở một đặc điểm chung nào đó cần được quan sát, phân tích mặt lượng của chúng. Các đơn vị, phần tử tạo nên hiện tượng được gọi là các đơn vị tổng thể.

Như vậy muốn xác định được một tổng thể thống kê, ta cần phải xác định được tất cả các đơn vị tổng thể của nó. Thực chất của việc xác định tổng thể thống kê là việc xác định các đơn vị tổng thể.

Trong nhiều trường hợp, các đơn vị của tổng thể được biểu hiện một cách rõ ràng, dễ xác định. Ta gọi nó là tổng thể bộ lộ. Ngược lại, một tổng thể mà các đơn vị của nó không được nhận biết một cách trực tiếp, ranh giới của tổng thể không rõ ràng được gọi là tổng thể tiềm ẩn.

Đối với tổng thể tiềm ẩn, việc tìm được đầy đủ, chính xác gặp nhiều khó khăn. Việc nhầm lẫn, bỏ sót các đơn trong tổng thể dễ xảy ra. Ví dụ như tổng thể là những những mê nhạc cổ điển, tổng thể người mê tín dị đoan,...

2. Mẫu (Samples)

Mẫu là một bộ phận của tổng thể, đảm bảo được tính đại diện và được chọn ra để quan sát và dùng để suy diễn cho toàn bộ tổng thể. Như vậy, tất cả các phần tử của mẫu phải thuộc tổng thể, nhưng ngược lại các phần tử của tổng thể thì chưa chắc thuộc mẫu. Điều này tưởng chừng là đơn giản, tuy nhiên trong một số trường hợp việc xác định mẫu cũng có thể dẫn đến nhầm lẫn, đặc biệt là trong trường hợp tổng thể ta nghiên cứu là tổng thể tiềm ẩn.

Ngoài ra, chọn mẫu như thế nào để làm cơ sở suy diễn cho tổng thể, tức là mẫu phải mang tính đại diện cho tổng thể. Điều này thực sự không dễ dàng, ta chỉ cố gắng hạn chế tối đa sự sai biệt này mà thôi chứ không thể khắc phục được hoàn toàn.

3. Quan sát (Observations)

Là mỗi đơn vị của mẫu ; trong một số tài liệu còn được gọi là quan trắc.

4. Tiêu thức thống kê

Các đơn vị tổng thể thường có nhiều đặc điểm khác nhau, tuy nhiên trong thống kế người ta chỉ chọn một số đặc điểm để nghiên cứu, các đặc điểm này người ra gọi là tiêu thức thống kê. Như vậy, tiêu thức thống kê là khái niệm chỉ các đặc điểm của đơn vị tổng thể. Mỗi tiêu thức thống kê đều có các giá trị biểu hiện của nó, dựa vào sự biểu hiện của nó người ta chia ra làm hai loại:

a) Tiêu thức thuộc tính: là tiêu thức phản ánh loại hoặc tính chất của đơn vị. Ví dụ như ngành kinh doanh, nghề nghiệp,...

b) Tiêu thức số lượng: là đặc trưng của đơn vị tổng thể được thể hiện bằng con số. Ví dụ, năng suất của một loại cây trồng.

Tiêu thức số lượng được chia làm 2 loại:

- Loại rời rạc: là loại các giá trị có thể của nó là hữu hạn hay vô hạn và có thể đếm được.

- Loại liên tục: là loại mà giá trị của nó có thể nhận bất kỳ một trị số nào đó trong một khoảng nào đó.

5. Tham số tổng thể

Là giá trị quan sát được của tổng thể và dùng để mô tả đặc trưng của hiện tượng nghiên cứu. Trong xác suất thống kê toán chúng ta đã biết các tham số tổng thể như trung bình tổng thể (μ), tỷ lệ tổng thể (p), phương sai tổng thể (σ^2). Ngoài ra, trong quá trình nghiên cứu sâu môn thống kê chúng ta còn có thêm nhiều tham số tổng thể nữa như: tương quan tổng thể (ρ), hồi qui tuyến tính tổng thể,…

6. Tham số mẫu

Tham số mẫu là giá trị tính toán được của một mẫu và dùng để suy rộng cho tham số tổng thể. Đó là cách giải thích mang tính chất thông thường, còn đối với xác suất thống kê thì tham số mẫu là ước lượng điểm của tham số tổng thể, trong trường hợp chúng ta chưa biết tham số tổng thể chúng ta có thể sử dụng tham số mẫu để ước lượng tham số tổng thể. Chúng ta có thể liệt kê vài tham số mẫu như sau: trung bình mẫu ($\bar{x}$), tỷ lệ mẫu ($\hat{p}$), phương sai mẫu (S^2), hệ số tương quan mẫu (r),…

IV. CÁC LOẠI THANG ĐO (Scales of Measurement)

Đứng trên quan điểm của nhà nghiên cứu, chúng ta cần xác định các phương pháp phân tích thích hợp dựa vào mục đích nghiên cứu và bản chất của dữ liệu. Do vậy, đầu tiên chúng ta tìm hiểu bản chất của dữ liệu thông qua khảo sát các cấp độ đo lường khác nhau vì mỗi cấp độ sẽ chỉ cho phép một số phương pháp nhất định mà thôi.

1. Khái niệm

- Số đo: là việc gán những dữ kiện lượng hoá hay những ký hiệu cho những hiện tượng quan sát. Chẳng hạn như những đặc điểm của khách hàng về sự chấp nhận, thái độ, thị hiếu hoặc những đặc điểm có liên quan khác đối với một sản phẩm mà họ tiêu dùng.

- Thang đo: là tạo ra một thang điểm để đánh giá đặc điểm của đối tượng nghiên cứu thể hiện qua sự đánh giá, nhận xét.

2. Các loại thang đo

- Thang đo danh nghĩa (Nominal scale)**:**

Là loại thang đo sử dụng cho dữ liệu thuộc tính mà các biểu hiện của dữ liệu không có sự hơn kèm, khác biệt về thứ bậc. Các con số không có mối quan hệ hơn kém, không thực hiện được các phép tính đại số. Các con số chỉ mang tính chất mã hoá. Ví dụ, tiêu thức giới tính ta có thể đánh số 1 là nam, 2 là nữ.

- Thang đo thứ bậc (Ordinal scale):

Là loại thang đo dùng cho các dữ liệu thuộc tính. Tuy nhiên trường hợp này biểu hiện của dữ liệu có sự so sánh. Ví dụ, trình độ thành thạo của công nhân được phân chia ra các bậc thợ từ 1 đến 7. Phân loại giảng viên trong các trường đại học: Giáo sư, P.Giáo sư, Giảng viên chính, Giảng viên. Thang đo này cũng không thực hiện được các phép tính đại số.

- Thang đo khoảng (Interval scale):

Là loại thang đo dùng cho các dữ liệu số lượng. Là loại thang đo cũng có thể dùng để xếp hạng các đối tượng nghiên cứu nhưng khoảng cách bằng nhau trên thang đo đại diện cho khoảng cách bằng nhau trong đặc điểm của đối tượng. Với thang đo này ta có thể thực hiện các phép tính đại số trừ phép chia không có ý nghĩa. Ví dụ như điểm môn học của sinh viên. Sinh viên A có điểm thi là 8 điểm, sinh viên B có điểm là 4 thì không thể nói rằng sinh viên A giỏi gấp hai lần sinh viên B.

- Thang đo tỷ lệ (Ratio scale):

Là loại thang đo cũng có thể dùng dữ liệu số lượng. Trong các loại thang đo đây là loại thang đo cao nhất. Ngoài đặc tính của thang đo khoảng, phép chia có thể thực hiện

được. Ví dụ, thu nhập trung bình 1 tháng của ông A là 2 triệu đồng và thu nhập của bà B là 4 triệu đồng, ta có thể nói rằng thu nhập trung bình trong một tháng của bà B gấp đôi thu nhập của ông A.

Tuỳ theo thang đo chúng ta có thể có một số phương pháp phân tích phù hợp, ta có thể tóm tắt như sau:

Phương pháp phân tích thống kê thích hợp với các thang đo

Loại thang đo	Đo lường độ tập trung	Đo lường độ phân tán	Đo lường tính tương quan	Kiểm định
1. Thang biểu danh	Mốt	Không có	Hệ số ngẫu nhiên	Kiểm định χ^2
2. Thang thứ tự	Trung vị	Sô phần trăm	Dãy tương quan	Kiểm định dấu
3. Thang khoảng	Trung bình	Độ lệch chuẩn	Hệ số tương quan	Kiểm định t, F
4. Thang tỷ lệ	Trung bình tỷ lệ	Hệ số biến thiên	Tất cả các phép trên	Sử dụng tất cả các phép trên

V. THU THẬP THÔNG TIN

Về nguyên tắc, thống kê mô tả chắc hẳn có từ lâu đời cũng gần như chữ viết. Nó liên quan chặt chẽ với nhu cầu của con người muốn sắp xếp lại một cách có trật tự trong vô vàn thông tin sự kiện đã đến với họ để hiểu hơn thực tại hơn nhằm tác động lên nó tốt hơn. Khi nghiên cứu bất kỳ hiện tượng kinh tế xã hội nào công việc đầu tiên là thu thập dữ liệu, sau đó là trình bày dữ liệu và phân tích.

1. Xác định nội dung thông tin

Nói chung, tuỳ thuộc vào mục đích nghiên cứu để xác định những nội dung thông tin cần thu thập. Thông tin sử dụng cho quá trình nghiên cứu phải đảm bảo các yêu cầu cơ bản sau:

- Thích đáng: Số liệu thu thập phải phù hợp, đáp ứng được mục đích nghiên cứu. Số liệu đáp ứng được mục tiêu nghiên cứu có tính chất trực tiếp hoặc gián tiếp. Đối với những thông tin dễ tiếp cận thường thì ta sử dụng số liệu trực tiếp, ví dụ muốn biết được nhu cầu của khách hàng chúng ta có thể hỏi trực tiếp khách hàng. Tuy nhiên, một số nội dung nghiên cứu mang tính chất nhạy cảm hoặc khó thu thập thì chúng ta có thể thu nhập những số liên gián tiếp có liên quan, ví dụ để thu thập thu nhập của cá nhân chúng ta có thể thu thập những nội dung có liên quan như nghề nghiệp, đơn vị công tác, chức vụ, nhà ở, phương tiện đi lại...

- Chính xác: Các thông tin trong quá trình nghiên cứu phải có giá trị, đáng tin cậy để các phân tích kết luận phản ánh được đặc điểm bản chất của hiện tượng.

- Kịp thời: Yêu cầu thông tin không những đáp ứng yêu cầu phù hợp, chính xác mà giá trị thông tin còn thể hiện ở chỗ nó có phục vụ kịp thời cho công tác quản lý và tiến trình ra các quyết định hay không.

- Khách quan: Tức là số liệu thu thập được không bị ảnh hưởng vào tính chủ quan của người thu thập cũng như người cung cấp số liệu và ngay cả trong thiết kế bảng câu hỏi. Yếu tố khách quan tưởng chừng thực hiện rất dễ dàng nhưng thực tế thì chúng ta khó có thể khắc phục vấn đề này một cách trọn vẹn, chúng ta chỉ có thể hạn chế yếu tố chủ quan một cách tối đa. Ví dụ chỉ cần một hành động đơn giản là tiếp cận với đáp viên là ít nhiều cũng ảnh hưởng đến kết quả trả lời của họ.

2. Nguồn số liệu

Khi nghiên cứu một hiện tượng cụ thể, người nghiên cứu có thể sử dụng từ nguồn số liệu đã có sẵn đã được công bố hay chưa công bố hay tự mình thu thập các dữ liệu cần thiết cho nghiên cứu. Dựa vào cách thức này người ta chia dữ liệu thành 2 nguồn: dữ liệu thứ cấp và dữ liệu sơ cấp.

2.1. Dữ liệu thứ cấp (Secondary data):

Dữ liệu thứ cấp là các thông tin đã có sẵn và đã qua tổng hợp, xử lý. Loại dữ kiện này có thể thu thập từ các nguồn sau:

(1) Số liệu nội bộ: là loại số liệu đã được ghi chép cập nhật trong đơn vị hoặc được thu thập từ các cuộc điều tra trước đây.

(2) Số liệu từ các ấn phẩm của nhà nước: Các dữ liệu do các cơ quan thống kê nhà nước phát hành định kỳ như niên giám thống kê, các thông tin cập nhật hàng năm về tình hình dân số lao động, kết quả sản xuất của các ngành trong nền kinh tế, số liệu về văn hoá xã hội.

(3) Báo, tạp chí chuyên ngành: Các báo và tạp chí đề cập đến vấn đề có tính chất chuyên ngành như tạp chí thống kê, giá cả thị trường,...

(4) Thông tin của các tổ chức, hiệp hội nghề nghiệp: Viên nghiên cứu kinh tế, phòng thương mại

(5) Các công ty chuyên tổ chức thu thập thông tin, nghiên cứu và cung cấp thông tin theo yêu cầu.

Số liệu thứ cấp có ưu điểm là có thể chia sẻ chi phí, do đó nó có tính kinh tế hơn, số liệu được cung cấp kịp thời hơn. Tuy nhiên, dữ liệu thứ cấp thường là các thông tin cơ bản, số liệu đã được tổng hợp đã qua xử lý cho nên không đầy đủ hoặc không phù hợp cho quá trình nghiên cứu. Số liệu thứ cấp thường ít được sử dụng để dự báo trong thống kê, số liệu này thường được sử dụng trong trình bày tổng quan nội dung nghiên cứu, là cơ sở để phát hiện ra vấn đề nghiên cứu. Ngoài ra, số liệu thứ cấp còn được sử dụng để đối chiếu lại kết quả nghiên cứu để nhằm kiểm tra lại tính đúng đắn hoặc phát hiện ra những vấn đề mới để có hướng nghiên cứu tiếp.

2.2. Dữ liệu sơ cấp (Primary data):

Là các thông tin thu thập từ các cuộc điều tra. Căn cứ vào phạm vi điều tra có thể chia thành 2 loại: Điều tra toàn bộ và điều tra chọn mẫu.

a) Điều tra toàn bộ: Là tiến hành thu thập thông tin trên tất cả các đơn vị thuộc tổng thể nghiên cứu.

Ưu điểm của điều tra toàn bộ là thu thập được thông tin về tất cả các đơn vị tổng thể. Tuy nhiên, loại điều tra này thường gặp phải một số trở ngại sau:

- Số lượng đơn vị thuộc tổng thể chung thường rất lớn cho nên tiến hành điều tra toàn bộ mất nhiều thời gian và tốn kém.

- Trong một số trường hợp do thời gian kéo dài dẫn đến số liệu kém chính xác do hiện tượng tự biến động qua thời gian.

- Trong một số trường hợp điều tra toàn bộ sẽ không thực hiện được, ví dụ như kiểm tra chất lượng sản phẩm phải phá huỷ các đơn vị thuộc đối tượng nghiên cứu.

b) Điều tra chọn mẫu: Để nghiên cứu tổng thể, ta chỉ cần lấy ra một số phần tử đại diện để nghiên cứu và từ đó suy ra kết quả cho tổng thể bằng các phương pháp thống kê.

Điều tra chọn mẫu thường được sử dụng vì các lý do sau:

- Tiết kiệm chi phí

- Cung cấp thông tin kịp thời cho quá trình nghiên cứu

- Đáng tin cậy. Đây là yếu tố rất quan trọng, nó làm cho điều tra chọn mẫu trở nên có hiệu quả và được chấp nhận. Tuy nhiên, để có sự đáng tin cậy này chúng ta phải có phương pháp khoa học để đảm bảo tính chính xác để chỉ cần chọn ra một số quan sát mà có thể suy luận cho cả tổng thể rộng lớn – đó là nhờ vào các lý thuyết thống kê.

Việc sử dụng điều tra toàn bộ hay điều tra chọn mẫu phụ thuộc vào nhiều yếu tố có liên quan: kích thước tổng thể, thời gian nghiên cứu cứu, khả năng về tài chính và nguồn lực, đặc điểm của nội dung nghiên cứu.

3. Các phương pháp thu thập thông tin

Để thu thập dữ liệu ban đầu, tuỳ theo nguồn kinh phí và đặc điểm của đối tượng cần thu thập thông tin, ta có các phương pháp sau đây:

a) Quan sát: Là phương pháp thu thập dữ liệu bằng cách quan sát hành động, hành vi thái độ của đối tượng được điều tra. Ví dụ, nghiên cứu trẻ con yêu thích màu sắc nào, quan sát thái độ khách hàng khi dùng thử loại sản phẩm. Phương pháp này tỏ ra hiệu quả đối với các trường hợp đối tượng khó tiếp cận và tăng tính khách quan của đối tượng. Tuy nhiên, phương pháp này tỏ ra khá tốn kém nhưng lượng thông tin thu thập được ít.

b) Phương pháp gởi thư: Theo phương pháp này nhân viên điều tra gởi bảng câu hỏi đến đối tượng cung cấp thông tin qua đường bưu điện. Phương pháp gởi thư có thể thu thập thông tin với khối lượng lớn, tiết kiệm chi phí so với các phương pháp khác. Tuy nhiên tỷ lệ trả lời bằng phương pháp này tương đối thấp, đây là một nhược điểm rất lớn của phương pháp này.

c) Phỏng vấn bằng điện thoại: Phương pháp thu thập thông tin bằng cách phỏng vấn qua điện thoại. Phương pháp này thu thập được thông tin một cách nhanh chóng, tuy nhiên phương pháp này có nhược điểm: tốn kém, nội dung thu thập thông tin bị hạn chế.

d) Phỏng vấn trực tiếp:

Phương pháp phỏng vấn trực tiếp thích hợp cho những cuộc điều tra cần thu thập nhiều thông tin, nội dung của thông tin tương đối phức tạp cần thu thập một cách chi tiết. Phương pháp phỏng vấn trực tiếp cho 2 hình thức:

(1) Phỏng vấn cá nhân. Nhân viên điều tra tiếp xúc với đối tượng cung cấp thông tin thường tại nhà riêng hoặc nơi làm việc. Thông thường phỏng vấn trực tiếp được áp dụng khi chúng ta cho tiến hành điều tra chính thức.

(2) Phỏng vấn nhóm. Nhân viên điều tra phỏng vấn từng nhóm để thảo luận về một vấn đề nào đó. Trường hợp này người ta thường sử dụng khi điều tra thử để kiểm tra lại nội dung của bảng câu hỏi được hoàn chỉnh chưa hoặc nhằm tìm hiểu một vấn đề phức tạp mà bản thân người nghiên cứu chưa nắm được một cách đầy đủ mà cần phải có ý kiến cụ thể từ những người am hiểu.

Sau đây ta có bảng tổng hợp một số ưu nhược điểm của các phương pháp thu thập thông tin.

Đặc điểm của các phương pháp thu thập thông tin

Tính chất	Phương pháp gởi thư	Phỏng vấn qua điện thoại	Phỏng vấn trực tiếp
Linh hoạt	Kém	Tốt	Tốt
Khối lượng thông tin	Đầy đủ	Hạn chế	Đầy đủ
Tốc độ thu thập thông tin	Chậm	Nhanh	Nhanh
Tỷ lệ câu hỏi được trả lời	Thấp	Cao	Cao
Chi phí	Tiết kiệm	Tốn kém	Tốn kém

PHẦN II
THỐNG KÊ MÔ TẢ

CHƯƠNG I
TỔNG HỢP VÀ TRÌNH BÀY DỮ LIỆU THỐNG KÊ

Thông tin ban đầu có tính rời rạc, dữ liệu hỗn độn không theo một trật tự nào và có thể quá nhiều nếu nhìn vào đây chúng ta không thể phát hiện được điều gì để phục vụ cho quá trình nghiên cứu. Do đó, chúng ta cần phải trình bày một cách có thể thống với hai mục đích là làm cho bảng dữ liệu gọn lại, hai là thể hiện được tính chất của nội dung nghiên cứu.

I. PHÂN TỔ THỐNG KÊ

1. Khái niệm

Phân tổ còn được gọi là phân lớp thống kê là căn cứ vào một hay một số tiêu thức để chia các đơn vị tổng thể ra thành nhiều tổ (lớp, nhóm) có tính chất khác nhau.

2. Nguyên tắc phân tổ

Một cách tổng quát tổng thể phải được phân chia một cách trọn vẹn, tức là một đơn vị của tổng thể chỉ thuộc một tổ duy nhất và một đơn vị thuộc một tổ nào đó phải thuộc tổng thể.

3. Phân tổ theo tiêu thức thuộc tính

- Trường hợp tiêu thức thuộc tính chỉ có một vài biểu hiện thì mỗi biểu hiện của tiêu thức thuộc tính có thể chia thành một tổ. Ví dụ, tiêu thức giới tính.

- Trường hợp tiêu thức thuộc tính có nhiều biểu hiện, ta ghép nhiều nhóm nhỏ lại với nhau theo nguyên tắc các nhóm ghép lại với nhau có tính chất giống nhau hoặc gần giống nhau. Ví dụ phân tổ trong công nghiệp chế biến: Thực phẩm và đồ uống, thuốc lá, dệt,...

4. Phân tổ theo tiêu thức số lượng

- Trường hợp tiêu thức số lượng có ít biểu hiện, thì cứ mỗi một lượng biến có thể thành lập một tổ.

Ví dụ 1.1: phân tổ công nhân trong một xí nghiệp dệt theo số máy do mỗi công nhân thực hiện.

Số máy/Công nhân	Số công nhân
10	3
11	7
12	20
13	50
14	35
15	15
Tổng	130

- Trường hợp tiêu thức số lượng có nhiều biểu hiện, ta phân tổ khoảng cách mỗi tổ và mỗi tổ có một giới hạn:

- Giới hạn trên: lượng biến nhỏ nhất của tổ.

- Giới hạn dưới: lượng biến lớn nhất của tổ.

Tuỳ theo mục đích nghiên cứu, người ta phân ra 2 loại phân tổ đều và phân tổ không đều.

- Phân tổ đều: Là phân tổ có khoảng cách tổ bằng nhau. Thông thường nếu chỉ vì mục đích nghiên cứu phân phối của tổng thể hoặc làm cho bảng thống kê gọn lại thì ta thường dùng phương pháp này.

Để xác định số tổ hình như không có một tiêu chuẩn tối ưu nó phụ thuộc vào kinh nghiệm. Dưới đây là một cách phân chia tổ mang tính chất tham khảo.

- Xác định số tổ (Number off classes):

Số tổ = $(2 \text{ x } n)^{0,3333}$ n: Số đơn vị tổng thể

- Xác định khoảng cách tổ (Class interval):

$$k = \frac{X_{max} - X_{min}}{\text{So to}}$$

- Xác định tần số (Frequency) của mỗi tổ: bằng cách đếm các quan sát rơi vào giới hạn của tổ đó.

- Một số qui ước khi lập bảng phân tổ:

- Trường hợp phân tổ theo tiêu thức số lượng rời rạc thì giới hạn trên và giới hạn dưới của 2 tổ kế tiếp nhau không được trùng nhau.

Ví dụ 1.2: Các xí nghiệp ở tỉnh X được phân tổ theo tiêu số lượng công nhân:

Số lượng công nhân	**Số xí nghiệp**
≤100	80
101 – 200	60
201 – 500	6
501 – 1.000	4
1.001 – 2.000	1
Tổng	**151**

- Trường hợp phân tổ theo tiêu thức số lượng loại liên tục, thường có qui ước sau:

* Giới hạn trên và giới hạn dưới của 2 tổ kế tiếp trùng nhau.

* Quan sát có lượng biến bằng đúng giới hạn trên của một tổ nào đó thì đơn vị đó được xếp vào tổ kế tiếp.

Ví dụ 1.3: phân tổ các tổ chức thương nghiệp theo doanh thu.

Doanh thu (triệu đồng)	**Số tổ chức thương nghiệp**
≤1.000	2
1.000-2.000	9
2.000-3.000	12
3.000-4.000	7
Tổng	**30**

5. Bảng phân phối tần số (Frequency table)

Sau khi phân tổ chúng ta có thể trình bày số liệu bằng cách sử dụng bảng phân phối tần số để biết được một số tính chất cơ bản của hiện tượng nghiên cứu.

Lượng biến	Tần số	Tần số tương đối	Tần số tích lũy
x_1	f_1	f_1/n	f_1
x_2	f_2	f_2/n	$f_1+ f_2$
...	...	...	...
x_i	f_i	f_i/n	$f_1+ f_2+...+ f_i$
...	...	...	...
x_k	f_k	f_k/n	$f_1+ f_2+...+ f_k$
Cộng	$\sum_{i=1}^{k} f_i = n$	1	

Trong đó lượng biến có thể là giá trị cụ thể hoặc là một khoảng.

6. Các loại phân tổ thống kê

- *Phân tổ kết cấu:*

Trong công tác nghiên cứu thống kê, các bảng phân tổ kết cấu được sử dụng rất phổ biến nhằm mục đích nêu lên bản chất của hiện tượng trong điều kiện nhất định và để nghiên cứu xu hướng phát triển của hiện tượng qua thời gian.

Ví dụ 1.4: Để xem xét cơ cấu giữa các nhóm ngành trong một quốc gia nào đó ta lập bảng như sau:

Bảng 1.1. Cơ cấu tổng sản phẩm của quốc gia X theo nhóm ngành, 2003 -2007

Đơn vị tính: %.

Tổng sản phẩm theo nhóm ngành	2003	2004	2005	2006	2007
Nông, lâm nghiệp và thủy sản	24,53	23,24	23,03	22,54	21,76
Công nghiệp và xây dựng	36,73	38,13	38,49	39,47	40,09
Dịch vụ	38,74	38,63	38,48	37,99	38,15
Tổng	100,00	100,00	100,00	100,00	100,00

Qua bảng kết cấu trên, ta thấy có thấy sự thay đổi về dịch chuyển cơ cấu ngành: Nhóm ngành công nghiệp và xây dựng có xu hướng tăng, nhóm ngành nông, lâm, thuỷ sản có xu hướng giảm,...

- *Phân tổ liên hệ:*

Khi tiến hành phân tổ liên hệ, các tiêu thức có liên hệ với nhau được phân biệt thành 2 loại tiêu thức nguyên nhân và tiêu thức kết quả. Phân tổ liên hệ có thể được vận dụng để nghiên cứu mối liên hệ giữa nhiều tiêu thức: mối liên hệ giữa năng suất với lượng phân bón, nghiên cứu giữa năng suất lao động của công nhân với tuổi nghề, bậc thợ, trình độ trang bị kỹ thuật,...

Ví dụ 1.5: Ta có bảng phân tổ liên hệ sau:

Bảng 1.2. Mối liên hệ giữa năng suất lao động với trình độ kỹ thuật nghề nghiệp của quốc gia X năm 2007

Trình độ kỹ thuật	Tuổi nghề (Năm)	Số công nhân	Sản lượng cả năm (tấn)	Năng suất lao động bình quân (tấn)
Đã được đào tạo kỹ thuật	dưới 5	15	1.125	75
	5-10	40	3.750	94
	10-15	40	4.200	105
	15-20	15	1.725	115
	trên 20	10	1.200	120
Cả tổ	-	120	12.000	100
Chưa được đào tạo kỹ thuật	dưới 5	10	510	51
	5-10	30	2.140	71
	10-15	20	1.540	79
	15-20	10	860	86
	trên 20	10	910	91
Cả tổ	-	80	6.000	75
Chung cho cả doanh nghiệp	-	200	18.000	90

II. BẢNG THỐNG KÊ (Statistical table)

Sau khi tổng hợp các tài liệu điều tra thống kê, muốn phát huy tác dụng của nó đối với phân tích thống kê, cần thiết phải trình bày kết quả tổng hợp theo một hình thức thuận lợi nhất cho việc sử dụng sau này.

1. Khái niệm

Bảng thống kê là một hình thức trình bày các tài liệu thống kê một cách có hệ thống, hợp lý và rõ ràng, nhằm nêu lên các đặc trưng về mặt lượng của hiện tượng nghiên cứu. Đặc điểm chung của tất cả các bảng thống kê là bao giờ cũng có những con số của từng bộ phận và có mối liên hệ mật thiết với nhau.

2. Cấu thành bảng thống kê

a) Về hình thức: Bảng thống kê bao gồm các hàng, cột, các tiêu đề, tiêu mục và các con số.

Các hàng cột thể hiện qui mô của bảng, số hàng và cột càng nhiều thì bảng thống kê càng lớn và càng phức tạp.

Tiêu đề của bảng thống kê phản ánh nội dung, ý nghĩa của bảng và của từng chi tiết trong bảng. Trước hết ta có tiêu đề chung, sau đó là các tiêu đề nhỏ (tiêu mục) là tên riêng của mỗi hàng, cột phản ánh ý nghĩa của cột đó.

b) Phần nội dung: Bảng thống kê gồm 2 phần: Phần chủ đề và phần giải thích.

Phần chủ đề nói lên tổng thể được trình bày trong bảng thống kê, tổng thể này được phân thành những đơn vị, bộ phận. Nó giải đáp: đối tượng nghiên cứu là những đơn vị nào, những loại hình gì. Có khi phần chủ đề phản ánh các địa phương hoặc các thời gian nghiên cứu khác nhau của một hiện tượng.

Phần giải thích gồm các chỉ tiêu giải thích các đặc điểm của đối tượng nghiên cứu, tức là giải thích phần chủ đề của bảng.

Phần chủ đề thường được đặt bên trái của bảng thống kê, còn phần giải thích được đặt ở phía trên của bảng. Cũng có trường hợp ta thay đổi vị trí.

Cấu thành của bảng thống kê có thể biểu hiện bằng sơ đồ sau:

Phần giải thích / Phần chủ đề	Các chỉ tiêu giải thích (tên cột)			
(1)	(2)	(3)	(4)	(5)
Tên chủ đề				

3. Các yêu cầu và qui ước xây dựng bảng thống kê

- ***Qui mô của bảng thống kê:*** không nên quá lớn, tức là quá nhiều hàng, cột và nhiều phân tổ kết hợp. Một bảng thống kê ngắn, gọn một cách hợp lý sẽ tạo điều kiện dễ dàng cho việc phân tích. Nếu thấy cần thiết nên xây dựng hai, ba,... bảng thống kê nhỏ thay cho một bảng thống kê quá lớn

- ***Số hiệu bảng:*** nhằm giúp cho người đọc dễ dàng xác định vị trí của bảng khi tham khảo, đặc biệt là đối với các tài liệu nghiên cứu người ta thường lập mục lục biểu bảng để người đọc dễ tham khảo và người trình bày dễ dàng hơn. Nếu số biểu bảng không nhiều thì chúng ta chỉ cần đánh số theo thứ tự xuất hiện của biểu bảng, nếu tài liệu được chia thành nhiều chương và số liệu biểu bảng nhiều thì ta có thể đánh số theo chương và theo số thứ tự xuất hiện của biểu bảng trong chương. Ví dụ, Bảng II.5 tức là bảng ở chương II và là bảng thứ 5.

- ***Tên bảng:*** yêu cầu ngắn gọn, đầy đủ, rõ ràng, đặt trên đầu bảng và phải chứa đựng nội dung, thời gian, không gian mà số liệu được biểu hiện trong bảng. Tuy nhiên yêu cầu này chỉ mang tính chất tương đối không có tiêu chuẩn rõ ràng nhưng thông thường người ta cố gắng trình bày trong một hàng hoặc tối đa là hai hàng.

- ***Đơn vị tính:***

- Đơn vị tính dùng chung cho toàn bộ số liệu trong bảng thống kê, trường hợp này đơn vị tính được ghi bên góc phải của bảng.

- Đơn vị tính theo từng chỉ tiêu trong cột, trong trường hợp này đơn vị tính sẽ được đặt dưới chỉ tiêu của cột.

- Đơn vị tính theo từng chỉ tiêu trong hàng, trong trường hợp này đơn vị tính sẽ được đặt sau chỉ tiêu theo mỗi hàng hoặc tạo thêm một cột ghi đơn vị tính.

- ***Cách ghi số liệu trong bảng:***

- Số liệu trong từng hàng (cột) có đơn vị tính phải nhận cùng một số lẻ, số liệu ở các hàng (cột) khác nhau không nhất thiết có cùng số lẻ với hàng (cột) tương ứng.

- Một số ký hiệu qui ước:

+ Nếu không có tài liệu thì trong ô ghi dấu gạch ngang "-"

+ Nếu số liệu còn thiếu, sau này sẽ bổ sung sau thì trong ô ghi dấu ba chấm "..."

+ Ký hiệu gạch chéo "x" trong ô nào đó thì nói lên hiện tượng không có liên quan đến chỉ tiêu đó, nếu ghi số liệu vào đó sẽ vô nghĩa hoặc thừa.

- ***Phần ghi chú ở cuối bảng:*** được dùng để giải thích rõ các nội dung chỉ tiêu trong bảng, nói rõ nguồn tài liệu đã sử dụng hoặc các chỉ tiêu cần thiết khác. Đối với các tài liệu khoa học, việc ghi rõ nguồn số liệu được coi như là bắt buộc không thể thiếu được trong biểu bảng.

III. TỔNG HỢP BẰNG ĐỒ THỊ

Phương pháp đồ thị thống kê là phương pháp trình bày và phân tích các thông tin thống kê bằng các biểu đồ, đồ thị và bản đồ thống kê. Phương pháp đồ thị thống kê sử dụng con số kết hợp với các hình vẽ, đường nét và màu sắc để trình bày các đặc điểm số lượng của hiện tượng. Chính vì vậy, ngoài tác dụng phân tích giúp ta nhận thức được những đặc điểm cơ bản của hiện tượng bằng trực quan một cách dễ dàng và nhanh chóng, đồ thị thống kê còn là một phương pháp trình bày các thông tin thống kê một cách khái quát và sinh động, chứa đựng tính mỹ thuật; thu hút sự chú ý của người đọc, giúp người xem dễ hiểu, dễ nhớ nên có tác dụng tuyên truyền cổ động rất tốt. Đồ thị thống kê có thể biểu thị:

- Kết cấu của hiện tượng theo tiêu thức nào đó và sự biến đổi của kết cấu.
- Sự phát triển của hiện tượng theo thời gian.
- So sánh các mức độ của hiện tượng.
- Mối liên hệ giữa các hiện tượng.
- Trình độ phổ biến của hiện tượng.
- Tình hình thực hiện kế hoạch.

Trong công tác thống kê thường dùng các loại đồ thị: Biểu đồ hình cột, biểu đồ tượng hình, biểu đồ diện tích (hình vuông, hình tròn, hình chữ nhật), đồ thị đường gấp khúc và biểu đồ hình màng nhện.

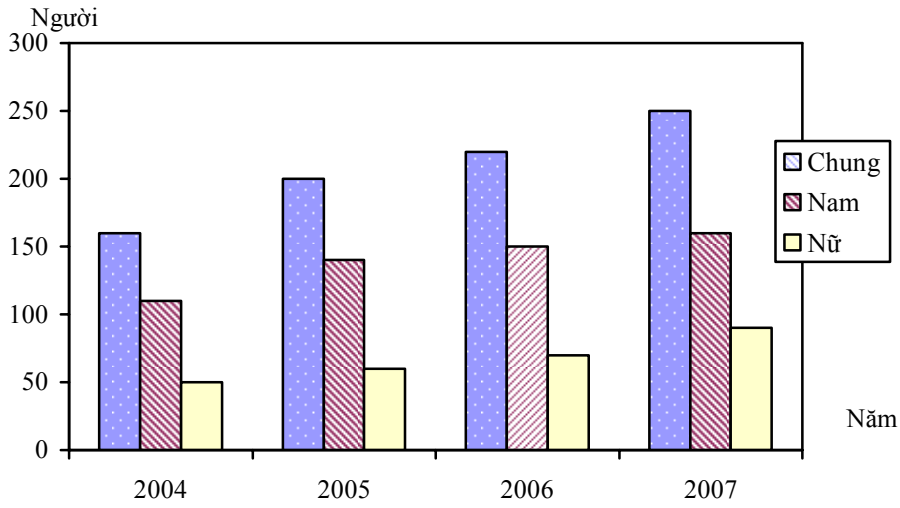
1. Biểu đồ hình cột

Biểu đồ hình cột là loại biểu đồ biểu hiện các tài liệu thống kê bằng các hình chữ nhật hay khối chữ nhật thẳng đứng hoặc nằm ngang có chiều rộng và chiều sâu bằng nhau, còn chiều cao tương ứng với các đại lượng cần biểu hiện.

Biểu đồ hình cột được dùng để biểu hiện quá trình phát triển, phản ánh cơ cấu và thay đổi cơ cấu hoặc so sánh cũng như biểu hiện mối liên hệ giữa các hiện tượng.

Ví dụ 1.6: Biểu diễn số lượng cán bộ khoa học công nghệ của một quốc gia nào đó chia theo nam nữ của 4 năm: 2004, 2005, 2006 và 2007 qua biểu đồ 1.1.

Biểu đồ 1.1: **Hình cột phản ánh số lượng cán bộ khoa học công nghệ của quốc gia X, 2004 - 2007**



Đồ thị trên vừa phản ánh quá trình phát triển của cán bộ khoa học công nghệ vừa so sánh cũng như phản ánh mối liên hệ giữa cán bộ là nam và nữ.

2. Biểu đồ diện tích

Biểu đồ diện tích là loại biểu đồ, trong đó các thông tin thống kê được biểu hiện bằng các loại diện tích hình học như hình vuông, hình chữ nhật, hình tròn, hình ô van,...

Biểu đồ diện tích thường được dùng để biểu hiện kết cấu và biến động cơ cấu của hiện tượng.

Tổng diện tích của cả hình là 100%, thì diện tích từng phần tương ứng với mỗi bộ phận phản ánh cơ cấu của bộ phận đó.

Biểu đồ diện tích hình tròn còn có thể biểu hiện được cả cơ cấu, biến động cơ cấu kết hợp thay đổi mức độ của hiện tượng. Trong trường hợp này số đo của góc các hình quạt phản ánh cơ cấu và biến động cơ cấu, còn diện tích toàn hình tròn phản ánh quy mô của hiện tượng.

Khi vẽ đồ thị ta tiến hành như sau:

- Lấy giá trị của từng bộ phận chia cho giá trị chung của chỉ tiêu nghiên cứu để xác định tỷ trọng (%)của từng bộ phận đó. Tiếp tục lấy 360 (360^0) chia cho 100 rồi nhân với tỷ trọng của từng bộ phận sẽ xác định được góc độ tương ứng với cơ cấu của từng bộ phận.

- Xác định bán kính của mỗi hình tròn có diện tích tương ứng là S: $R = \sqrt{S:\pi}$ vì diện tích hình tròn: $S = \pi.R^2$. Khi có độ dài của bán kính mỗi hình tròn, ta sẽ dễ dàng vẽ được các hình tròn đó.

Ví dụ 1.7: Có số lượng về học sinh phổ thông phân theo cấp học 3 năm 2005, 2006 và 2007 của địa phương X như bảng 1.3:

Bảng 1.3: **Học sinh phổ thông phân theo cấp học của địa phương X, 2005 - 2007**

	2005		2006		2007	
	Số lượng (Người)	Cơ cấu (%)	Số lượng (Người)	Cơ cấu (%)	Số lượng (Người)	Cơ cấu (%)
Tổng số học sinh	**1.000**	**100,0**	**1.140**	**100,0**	**1.310**	**100,0**
Chia ra:						
Tiểu học	500	50,0	600	53,0	700	53,5
Trung học cơ sở	300	30,0	320	28,0	360	27,5
Trung học phổ thông	200	20,0	220	19,0	250	19,0

Từ số liệu bảng 1.3 ta tính các bán kính tương ứng:

Năm 2005: $R = \sqrt{1000/3{,}14} = 17{,}84$

Năm 2006: $R = \sqrt{1140/3{,}14} = 19{,}05$

Năm 2007: $R = \sqrt{1310/3{,}14} = 20{,}42$

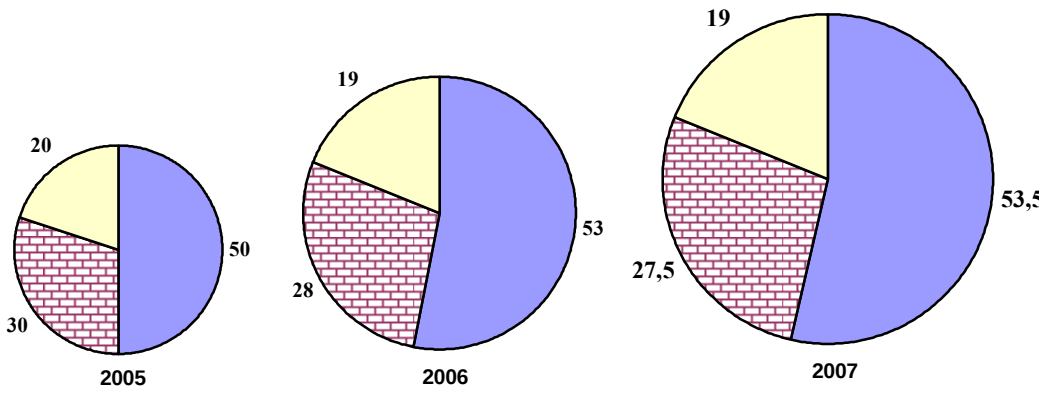
Nếu năm 2005 lấy R = 1,00

Thì năm 2006 có R = 19,05 : 17,84 = 1,067

Năm 2007 có R = 20,42 : 17,84 = 1,144

Kết quả 3 hình tròn được vẽ phản ánh cả quy mô học sinh phổ thông lẫn cơ cấu và biến động cơ cấu theo cấp học của học sinh qua các năm 2005, 2006 và 2007.

Biểu đồ 1.2: **Biểu đồ cơ cấu học sinh phổ thông địa phương X từ 2005 – 2007**



Tuy nhiên, nếu chúng ta chỉ vẽ biểu đồ mang tính đơn lẽ thì không cần phải xác định độ lớn của đường kính.

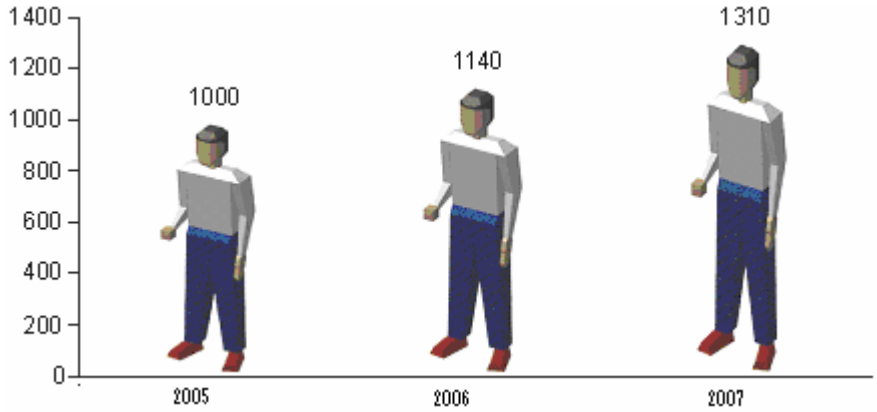
3. Biểu đồ tượng hình

Biểu đồ tượng hình là loại đồ thị thống kê, trong đó các tài liệu thống kê được thể hiện bằng các hình vẽ tượng trưng. Biểu đồ tượng hình được dùng rộng rãi trong việc tuyên truyền, phổ biến thông tin trên các phương tiện sử dụng rộng rãi. Biểu đồ hình tượng có nhiều cách vẽ khác nhau, tuỳ theo sáng kiến của người trình bày mà lựa chọn loại hình vẽ tượng hình cho phù hợp và hấp dẫn.

Tuy nhiên khi sử dụng loại biểu đồ này phải theo nguyên tắc: cùng một chỉ tiêu phải được biểu hiện bằng cùng một loại hình vẽ, còn chỉ tiêu đó ở các trường hợp nào có trị số lớn nhỏ khác nhau thì sẽ biểu hiện bằng hình vẽ có kích thước lớn nhỏ khác nhau theo tỷ lệ tương ứng.

Trở lại ví dụ trên số lượng học sinh phổ thông được biểu diễn như sau:

Biểu đồ 1.3: **Biểu đồ cơ cấu học sinh phổ thông địa phương X từ 2005 – 2007**



4. Đồ thị đường gấp khúc

Đồ thị đường gấp khúc là loại đồ thị thống kê biểu hiện các tài liệu bằng một đường gấp khúc nối liền các điểm trên một hệ toạ độ, thường là hệ toạ độ vuông góc.

Đồ thị đường gấp khúc được dùng để biểu hiện quá trình phát triển của hiện tượng, biểu hiện tình hình phân phối các đơn vị tổng thể theo một tiêu thức nào đó hoặc biểu thị tình hình thực hiện kế hoạch theo từng thời gian của các chỉ tiêu nghiên cứu.

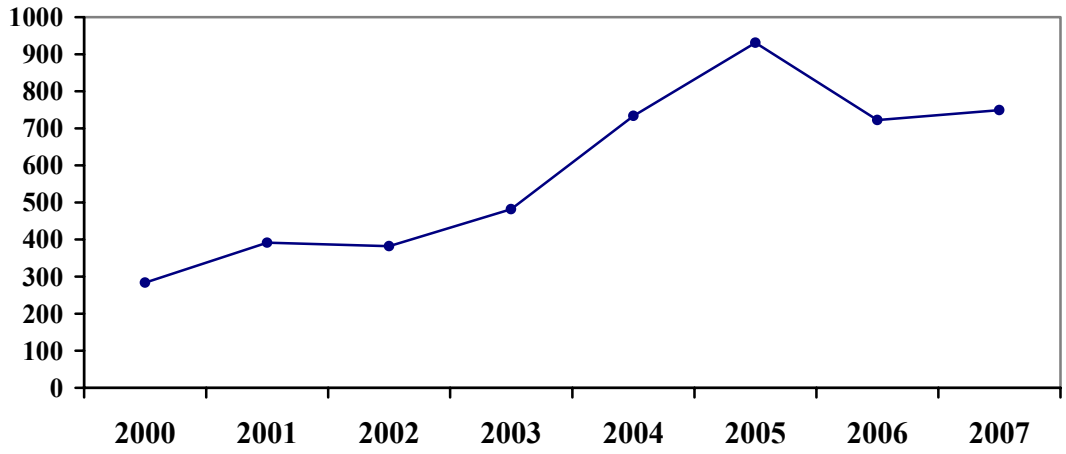
Trong một đồ thị đường gấp khúc, trục hoành thường được biểu thị thời gian, trục tung biểu thị mức độ của chỉ tiêu nghiên cứu. Cũng có khi các trục này biểu thị hai chỉ tiêu có liên hệ với nhau hoặc lượng biến và các tần số (hay tần suất) tương ứng. Độ phân chia trên các trục cần được xác định cho thích hợp vì có ảnh hưởng trực tiếp đến độ dốc của đồ thị. Mặt khác, cần chú ý là trên mỗi trục toạ độ chiều dài của các khoảng phân chia tương ứng với sự thay đổi về lượng của chỉ tiêu nghiên cứu phải bằng nhau.

Ví dụ 1.8: Sản lượng cà phê xuất khẩu của quốc gia X qua các năm từ 2000 - 2007 (ngàn tấn) có kết quả như sau:

Năm	2000	2001	2002	2003	2004	2005	2006	2007
Sản lượng (ngàn tấn)	283,3	391,6	382,0	482,0	733,9	931,0	722,0	749,0

Số liệu trên được biểu diễn qua đồ thị đường gấp khúc sau:

***Đồ thị 1.4:* Sản lượng cà phê xuất khẩu của quốc gia X từ 2000 – 2007**



5. Biểu đồ hình màng nhện

Biểu đồ hình màng nhện là loại đồ thị thống kê dùng để phản ánh kết quả đạt được của hiện tượng lặp đi lặp lại về mặt thời gian, ví dụ phản ánh về biến động thời vụ của một chỉ tiêu nào đó qua 12 tháng trong năm. Để lập đồ thị hình màng nhện ta vẽ một hình tròn bán kính R, sao cho R lớn hơn trị số lớn nhất của chỉ tiêu nghiên cứu (lớn hơn bao nhiêu lần không quan trọng, miễn là đảm bảo tỷ lệ nào đó để hình vẽ được cân đối, kết quả biểu diễn của đồ thị dễ nhận biết). Sau đó chia đường tròn bán kính R thành các phần đều nhau theo số kỳ nghiên cứu (ở đây là 12 tháng) bởi các đường thẳng đi qua tâm đường tròn. Nối các giao điểm của bán kính cắt đường tròn ta được đa giác đều nội tiếp đường tròn. Đó là giới hạn phạm vi của đồ thị. Độ dài đo từ tâm đường tròn đến các điểm xác định theo các đường phân chia đường tròn nói trên chính là các đại lượng cần biểu hiện của hiện tượng tương ứng với mỗi thời kỳ. Nối các điểm xác định sẽ được hình vẽ của đồ thị hình màng nhện.

Ví dụ 1.9: Có số liệu về trị giá xuất, nhập khẩu hải sản của tỉnh X 2 năm 2006 và 2003 như sau:

***Bảng 1.4.* Giá trị xuất khẩu hải sản trong 12 tháng tỉnh X năm 2006 - 2007**

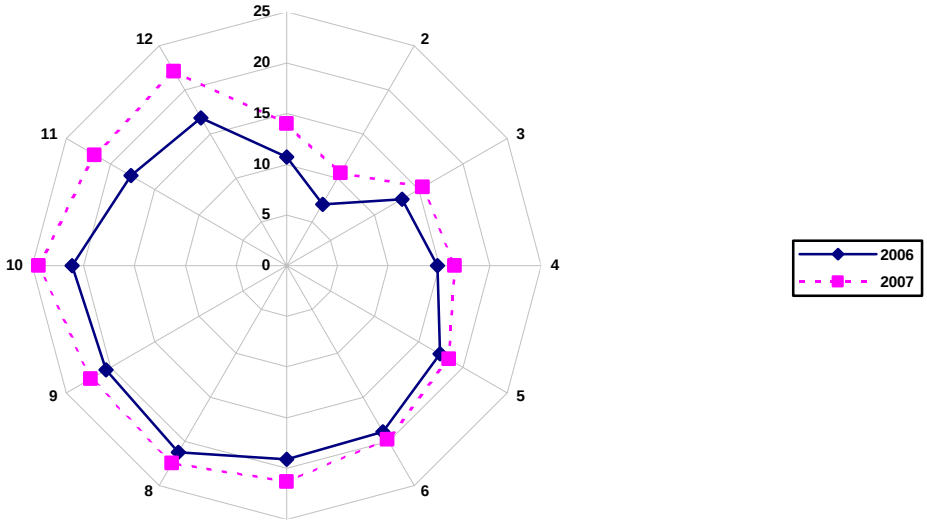
ĐVT: Triệu đồng

Tháng \ Năm	2006	2007	Tháng \ Năm	2006	2007
1	10,7	14,0	7	19,1	21,3
2	7,0	10,5	8	21,2	22,5
3	13,1	15,4	9	20,5	22,2
4	14,8	16,5	10	21,1	24,4
5	17,4	18,4	11	17,7	21,8
6	18,9	19,8	12	16,8	22,1

Chia đường tròn thành 12 phần đều nhau, vẽ các đường thẳng tương ứng cắt đường

tròn tại 12 điểm. Nối các điểm lại có đa giác đều 12 cạnh nội tiếp đường tròn. Căn cứ số liệu của bảng ta xác định các điểm tương ứng với giá trị xuất khẩu đạt được của các tháng trong từng năm rồi nối các điểm đó lại thành đường liền ta được đồ thị hình màng nhện biểu diễn kết quả xuất khẩu qua các tháng trong 2 năm của tỉnh X.

Đồ thị 1.5: **Đồ thị Giá trị xuất khẩu hải sản trong 12 tháng tỉnh X năm 2006 - 2007**



Sự mô tả của đồ thị hình màng nhện cho phép ta quan sát và so sánh không chỉ kết quả xuất khẩu giữa các tháng khác nhau trong cùng một năm, mà cả kết quả sản xuất giữa các tháng cùng tên của các năm khác nhau cũng như xu thế biến động chung về xuất khẩu của các năm.

CHƯƠNG II
CÁC MỨC ĐỘ CỦA HIỆN TƯỢNG KINH TẾ XÃ HỘI

Nghiên cứu các mức độ của hiện tượng kinh tế xã hội là yêu cầu quan trọng của việc tổng hợp, tính toán và phân tích thống kê nhằm biểu hiện mặt lượng trong quan hệ mật thiết với mặt chất của hiện tượng nghiên cứu trong điều kiện thời gian và không gian cụ thể nhờ vào sự trợ giúp của các phương pháp thống kê.

Dưới đây là nội dung, phương pháp tính và điều kiện vận dụng của các đại lượng đó.

I. SỐ TUYỆT ĐỐI

Số tuyệt đối là chỉ tiêu biểu hiện quy mô, khối lượng của hiện tượng hoặc quá trình kinh tế - xã hội trong điều kiện thời gian và không gian cụ thể.

Số tuyệt đối trong thống kê bao gồm các con số phản ánh quy mô của tổng thể hay của từng bộ phận trong tổng thể (số doanh nghiệp, số nhân khẩu, số học sinh đi học, số lượng cán bộ khoa học,...) hoặc tổng các trị số theo một tiêu thức nào đó (tiền lương của công nhân, giá trị sản xuất công nghiệp, tổng sản phẩm trong nước (GDP), v.v...).

Số tuyệt đối dùng để đánh giá và phân tích thống kê, là căn cứ không thể thiếu được trong việc xây dựng chiến lược phát triển kinh tế, tính toán các mặt cân đối, nghiên cứu các mối quan hệ kinh tế - xã hội, là cơ sở để tính toán các chỉ tiêu tương đối và bình quân.

Có hai loại số tuyệt đối: Số tuyệt đối thời kỳ và số tuyệt đối thời điểm.

Số tuyệt đối thời kỳ: Phản ánh quy mô, khối lượng của hiện tượng trong một thời kỳ nhất định. Ví dụ: Giá trị sản xuất công nghiệp trong 1 tháng, quý hoặc năm; Sản lượng lương thực năm 2005, năm 2006, năm 2007,...

Số tuyệt đối thời điểm: Phản ánh quy mô, khối lượng của hiện tượng ở một thời điểm nhất định như: dân số của một địa phương nào đó có đến 0 giờ ngày 01/04/2005; giá trị tài sản cố định có đến 31/12/2007; lao động làm việc của doanh nghiệp vào thời điểm 1/7/2007,...

II. SỐ TƯƠNG ĐỐI

Số tương đối là chỉ tiêu biểu hiện quan hệ so sánh giữa hai chỉ tiêu thống kê cùng loại nhưng khác nhau về thời gian hoặc không gian hoặc giữa hai chỉ tiêu khác loại nhưng có quan hệ với nhau. Trong hai chỉ tiêu để so sánh của số tương đối, sẽ có một số được chọn làm gốc (chuẩn) để so sánh.

Số tương đối có thể được biểu hiện bằng số lần, số phần trăm (%) hoặc phần nghìn (‰), hay bằng các đơn vị kép (người/km^2, người/1000 người; đồng/1000đồng, ...).

Trong công tác thống kê, số tương đối được sử dụng rộng rãi để phản ánh những đặc điểm về kết cấu, quan hệ tỷ lệ, tốc độ phát triển, mức độ hoàn thành kế hoạch, mức độ phổ biến của hiện tượng kinh tế - xã hội được nghiên cứu trong điều kiện thời gian và không gian nhất định.

Số tương đối phải được vận dụng kết hợp với số tuyệt đối. Số tương đối thường là kết quả của việc so sánh giữa hai số tuyệt đối. Số tương đối tính ra có thể rất khác nhau, tuỳ thuộc vào việc lựa chọn gốc so sánh. Có khi số tương đối có giá trị rất lớn nhưng ý nghĩa của nó không đáng kể vì trị số tuyệt đối tương ứng của nó lại rất nhỏ. Ngược lại, có số tương đối tính ra khá nhỏ nhưng lại mang ý nghĩa quan trọng vì trị số tuyệt đối tương ứng của nó có quy mô đáng kể. Ví dụ: 1% dân số Việt Nam tăng lên trong những năm 1960 đồng nghĩa với dân số tăng thêm 300 nghìn người, nhưng 1% dân số tăng lên trong những năm 2000 lại đồng nghĩa với dân số tăng thêm 800 nghìn người.

Căn cứ vào nội dung mà số tương đối phản ánh, có thể phân biệt: số tương đối động

thái, số tương đối kế hoạch, số tương đối kết cấu, số tương đối cường độ, và số tương đối không gian.

1. Số tương đối động thái

Số tương đối động thái là chỉ tiêu phản ánh biến động theo thời gian về mức độ của chỉ tiêu kinh tế - xã hội. Số tương đối này tính được bằng cách so sánh hai mức độ của chỉ tiêu được nghiên cứu ở hai thời gian khác nhau. Mức độ của thời kỳ được tiến hành nghiên cứu thường gọi là mức độ của kỳ báo cáo, còn mức độ của một thời kỳ nào đó được dùng làm cơ sở so sánh thường gọi là mức độ kỳ gốc.

Trong hai mức độ đó, mức độ tử số (y_1) là mức độ cần nghiên cứu (hay còn gọi là mức độ kỳ báo cáo), mức độ ở mẫu số (y_0) là mức độ kỳ gốc (hay mức độ dùng làm cơ sở so sánh).

- Nếu y_0 cố định qua các kỳ nghiên cứu ta có kỳ gốc cố định: dùng để so sánh một chỉ tiêu nào đó ở hai thời kỳ tương đối xa nhau. Thông thường người ta chọn năm gốc là năm đầu tiên của dãy số.

- Nếu y_0 thay đổi theo kỳ nghiên cứu ta có kỳ gốc liên hoàn: dùng để nói lên sự biến động của hiện tượng liên tiếp nhau qua các kỳ nghiên cứu.

Ví dụ 2.1: Sản lượng hàng hóa tiêu thụ (1.000 tấn) của một công ty X qua các năm như sau:

Năm	2001	2002	2003	2004	2005	2006	2007
Sản lượng hàng hóa (1.000 tấn)	240,0	259,2	282,5	299,5	323,4	355,8	387,8
Tốc độ phát triển liên hoàn (lần)		1,08	1,09	1,06	1,08	1,10	1,09

- Mối liên hệ giữa tốc độ phát triển định gốc và tốc độ phát triển liên hoàn. Nếu ta có dãy số sau:

Thời kỳ	1	2	3	...	n-1	n
y_i	y_1	y_2	y_3	…	y_{n-1}	y_n

thì mối liên hệ giữa tốc độ phát triển định gốc và tốc độ phát triển liên hoàn được thể hiện qua công thức sau: $\frac{y_2}{y_1}\frac{y_3}{y_2}....\frac{y_n}{y_{n-1}}=\frac{y_n}{y_1}$.

2. Số tương đối so sánh

Số tương đối so sánh là chỉ tiêu phản ánh quan hệ so sánh giữa hai bộ phận trong một tổng thể hoặc giữa hai hiện tượng cùng loại nhưng khác nhau về điều kiện không gian. Ví dụ: Dân số thành thị so với dân số nông thôn, dân số là nam so với dân số là nữ; giá trị tăng thêm của doanh nghiệp ngoài quốc doanh so với giá trị tăng thêm của doanh nghiệp quốc doanh; năng suất lúa của tỉnh X so với năng suất lúa của tỉnh Y; số học sinh đạt kết quả học tập khá giỏi so với số học sinh đạt kết quả trung bình,...

3. Số tương đối kế hoạch

Số tương đối kế hoạch là chỉ tiêu phản ánh mức cần đạt tới trong kỳ kế hoạch hoặc mức đã đạt được so với kế hoạch được giao về một chỉ tiêu kinh tế - xã hội nào đó. Số tương đối kế hoạch được chia thành hai loại:

+ Số tương đối nhiệm vụ kế hoạch: Phản ánh quan hệ so sánh giữa mức độ đề ra trong kỳ kế hoạch với mức độ thực tế ở kỳ gốc của một chỉ tiêu kinh tế - xã hội.

$$KH=\frac{\text{Mức kế hoạch}}{\text{Mức thực tế kỳ gốc}}x100=\frac{y_{KH}}{y_0}x100$$

+ Số tương đối hoàn thành kế hoạch: Phản ánh quan hệ so sánh giữa mức thực tế đã

đạt được với mức kế hoạch trong kỳ về một chỉ tiêu kinh tế - xã hội.

$$HT = \frac{\text{Mức thực tế đạt được}}{\text{Mức kế hoạch}} x\,100 = \frac{y_I}{y_{KH}} x\,100$$

Mối liên hệ giữa số tương đối động thái và số tương đối kế hoạch:

$$\frac{y_I}{y_0} = \frac{y_{KH}}{y_0} x \frac{y_I}{y_{KH}}$$

4. Số tương đối kết cấu

Số tương đối kết cấu là chỉ tiêu phản ánh tỷ trọng của mỗi bộ phận chiếm trong tổng thể, tính được bằng cách đem so sánh mức độ tuyệt đối của từng bộ phận với mức độ tuyệt đối của toàn bộ tổng thể.

Số tương đối kết cấu thường được biểu hiện bằng số phần trăm. Ví dụ: Tỷ trọng của GDP theo từng ngành trong tổng GDP của nền kinh tế quốc dân; tỷ trọng dân số của từng giới nam hoặc nữ trong tổng số dân,...

$$\text{Số tương đối kết cấu} = \frac{\text{Số tuyệt đối từng bộ phận}}{\text{Số tuyệt đối của tổng thể}} x\,100$$

Ví dụ 2.2: Trong công ty A có 500 công nhân, trong đó có 300 công nhân nam và 200 công nhân nữ.

$$\text{Tỷ trọng nam trong tổng công nhân} = \frac{300}{500}.100\% = 60\%$$

$$\text{Tỷ trọng nữ trong tổng công nhân} = \frac{200}{500}.100\% = 40\%$$

5. Số tương đối cường độ

Số tương đối cường độ là chỉ tiêu biểu hiện trình độ phổ biến của một hiện tượng trong các điều kiện thời gian và không gian cụ thể.

Số tương đối cường độ tính được bằng cách so sánh mức độ của hai chỉ tiêu khác nhau nhưng có quan hệ với nhau. Số tương đối cường độ biểu hiện bằng đơn vị kép, do đơn vị tính ở tử số và ở mẫu số hợp thành. Số tương đối cường độ được tính toán và sử dụng rất phổ biến trong công tác thống kê. Các số tương đối trong số liệu thống kê thường gặp như mật độ dân số bằng tổng số dân (người) chia cho diện tích tự nhiên (km^2) với đơn vị tính là người /km^2; GDP bình quân đầu người bằng tổng GDP (nghìn đồng) chia cho dân số trung bình (người) với đơn vị tính là 1000đ/người; số bác sĩ tính bình quân cho một vạn dân bằng tổng số bác sĩ chia cho tổng số dân tính bằng vạn người với đơn vị tính là người /10.000 người,...

$$\text{Mật độ dân số} = \frac{\text{Số dân}}{\text{Diện tích}} (\text{người/Km}^2)$$

$$\text{Năng suất lao động} = \frac{\text{Tổng sản phẩm}}{\text{Tổng số công nhân}} (\text{Sản phẩm/người})$$

III. SỐ ĐO ĐỘ TẬP TRUNG – SỐ BÌNH QUÂN (Measures of central tendency):

Số bình quân là chỉ tiêu biểu hiện mức độ điển hình của một tổng thể gồm nhiều đơn vị cùng loại được xác định theo một tiêu thức nào đó. Số bình quân được sử dụng phổ biến trong thống kê để nêu lên đặc điểm chung nhất, phổ biến nhất của hiện tượng kinh tế - xã hội trong các điều kiện thời gian và không gian cụ thể. Ví dụ: Tiền lương bình quân một công nhân trong doanh nghiệp là mức lương phổ biến nhất, đại diện cho các mức lương khác nhau của công nhân trong doanh nghiệp; thu nhập bình quân đầu người của

một địa bàn là mức thu nhập phổ biến nhất, đại diện cho các mức thu nhập khác nhau của mọi người trong địa bàn đó.

Số bình quân còn dùng để so sánh đặc điểm của những hiện tượng không có cùng một quy mô hay làm căn cứ để đánh giá trình độ đồng đều của các đơn vị tổng thể.

Xét theo vai trò đóng góp khác nhau của các thành phần tham gia bình quân hoá, số bình quân chung được chia thành số bình quân giản đơn và số bình quân gia quyền.

+ Số trung bình giản đơn: Được tính trên cơ sở các thành phần tham gia bình quân hoá có vai trò về qui mô (tần số) đóng góp như nhau.

+ Số trung bình gia quyền (trung bình có trọng số): Được tính trên cơ sở các thành phần tham gia bình quân hoá có vai trò về qui mô (tần số) đóng góp khác nhau.

Để tính được số trung bình chính xác và có ý nghĩa, điều kiện chủ yếu là nó phải được tính cho những đơn vị cùng chung một tính chất (thường gọi là tổng thể đồng chất). Muốn vậy, phải dựa trên cơ sở phân tổ thống kê một cách khoa học và chính xác. Đồng thời phải vận dụng kết hợp giữa số bình quân tổ với số bình quân chung.

Có nhiều loại số bình quân khác nhau. Trong thống kê kinh tế - xã hội thường dùng các loại sau: Số trung bình số học, số trung bình điều hoà, số trung bình hình học (số trung bình nhân), mốt và trung vị.

Dưới đây là từng loại số bình quân nêu trên.

1. Số trung bình cộng (Mean)

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

x_i: Giá trị lượng biến quan sát

n: Số quan sát

2. Số trung bình gia quyền (Weighted mean)

Với mỗi lượng biến x_i có tần số tương ứng f_i, số trung bình được xác định theo công thức sau:

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

x_i: Giá trị lượng biến quan sát

f_i: Tần số lượng biến quan sát

Ví dụ 2.3: Có tài liệu về mức thu nhập của các hộ theo tháng

Thu nhập hàng tháng (triệu đồng)	**Số hộ**
5.000	3
5.250	8
5.400	9
5.450	10
5.600	12
6.000	30
6.200	15
6.300	7
6.500	6
Tổng	**100**

Ta lập bảng tổng thu nhập hàng tháng của các hộ

Thu nhập hàng tháng (triệu đồng) **(x_i)**	**Số hộ** **(f_i)**	**$x_i f_i$**
5.000	3	15.000
5.250	8	42.000
5.400	9	48.600
5.450	10	54.500
5.600	12	67.200
6.000	30	180.000
6.200	15	93.000
6.300	7	44.100
6.500	6	39.000
Tổng	**100**	**583.400**

Ví dụ 2.4: Có số liệu thu nhập hàng tháng (ngàn đồng) của nhân viên một công ty như sau:

Thu nhập hàng tháng (ngàn đồng)	**Số nhân viên**
500-520	8
520-540	12
540-560	20
560-580	56
580-600	18
600-620	16
Trên 620	10
Tổng	**140**

Chú ý, trường hợp dãy số được phân tổ thì lượng biến x_i là trị số giữa của các tổ. Nếu dãy số có tổ mở thì lấy khoảng cách tổ của tổ gần tổ mở nhất để tính giới hạn trên của tổ mở từ đó xác định được giá trị x_i.

- Đối với tổ không có giới hạn trên: giới hạn dưới của tổ mở cộng với khoảng cách tổ của tổ trước đó mở rồi chia hai.

- Đối với tổ không có giới hạn dưới: giới hạn trên của tổ mở trừ khoảng cách tổ của tổ sau đó mở rồi chia hai. Tùy theo tính chất của nội dung nghiên cứu mà có thể chọn giá trị x_i phù hợp.

Từ bảng trên ta có bảng sau:

Thu nhập hàng tuần (1.000đ)	x_i	f_i	x_if_i
500-520	510	8	4.080
520-540	530	12	6.360
540-560	550	20	11.000
560-580	570	56	31.920
580-600	590	18	10.620
600-620	610	16	9.760
Trên 620	630	10	6.300
Tổng		**140**	**80.040**

Áp dụng công thức ta có:

$$\bar{x} = \frac{80.040}{140} = 571,71$$

Tuy nhiên, việc ước lượng các giá trị x_i có chính xác hay không còn phụ thuộc vào phân phối của từng tổ. Nếu phân phối của từng tổ có tính chất đối xứng thì việc ước lượng x_i có thể chấp nhận được, tuy nhiên đối với các trường hợp phân phối của tổ lệch trái hoặc lệch phải thì kết quả đó khó có thể chấp nhận được. Do đó, trong quá trình tính toán với sự hỗ trợ của các phần mềm máy tính ta nên sử dụng số liệu điều tra và tính với công thức trung bình đơn giản để đảm bảo tính chính xác.

3. Số trung bình điều hòa (Harmonic mean)

Số trung bình điều hòa được sử dụng trong trường hợp biết các lượng biến x_i và tích x_if_i mà chưa biết tần số f_i.

$$\bar{x} = \frac{\sum_{i=1}^{k} M_i}{\sum_{i=1}^{k} \frac{M_i}{x_i}}, \quad M_i = x_if_i$$

Ví dụ 2.5: Có số liệu giá thành sản và chi phí sản xuất của 3 phân xưởng của một doanh nghiệp:

Phân xưởng	Giá thành 1 tấn sản phẩm (1.000đ)	Chi phí sản xuất(1.000đ)
Số 1	18,5	740
Số 2	19,0	855
Số 3	19,4	970

Đặt x_i: Giá thành của phân xưởng i

M_i: Chi phí của phân xưởng i

Giá thành trung bình một tấn sản phẩm của doanh nghiệp được xác định bởi công thức:

$$\bar{x} = \frac{740 + 855 + 970}{\frac{740}{18,5} + \frac{855}{19,0} + \frac{970}{19,4}} = 19\,(1.000đ)$$

4. Số trung bình nhân (Geometric mean)

Số trung nhân hay số trung bình hình học sử dụng để tính tốc độ phát triển trung bình nói riêng và dùng để tính số trung bình trong trường hợp các giá trị x_i có mối liên hệ tích.

$$\bar{x} = \sqrt[n]{x_1 .... x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Ví dụ 2.6: Hãy tính tốc độ phát triển sản lượng hàng hóa tiêu thụ (1.000 tấn) của một công ty qua các năm như sau:

Năm	2001	2002	2003	2004	2005	2006	2007
Sản lượng hàng hóa (1.000 tấn)	240,0	259,2	282,5	299,5	323,4	355,8	387,8
Tốc độ phát triển liên hoàn (lần)		1,08	1,09	1,06	1,08	1,10	1,09

Giữa các tốc độ phát triển liên hoàn có mối quan hệ nhân, do đó ta áp dụng công thức trung bình nhân:

$$\bar{x} = \sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6} = \sqrt[6]{1,6158} = 1,08$$

Như vậy, trung bình mỗi một năm sản lượng hàng hoá năm sau sẽ bằng 1,08 lần năm trước.

5. Số trung vị - Me (Median)

Trong một số trường hợp đặc biệt, nếu dữ liệu có sự biến động lớn hay có sự chênh lệch bất thường thì số trung bình tỏ ra không đại diện cho tổng thể vì các giá trị quá nhỏ hay quá lớn sẽ làm lệch kết quả của số trung bình. Số trung vị là một giá trị bình quân có vẻ đại diện tốt hơn cho số trung bình trong trường hợp này, bởi vì nó sẽ chia tổng thể ra thành hai nhóm có số quan sát bằng nhau: một nhóm có giá trị nhỏ hơn, một nhóm có giá trị lớn hơn.

5.1. Định nghĩa: Số trung vị là lượng biến đứng ở vị trí giữa trong dãy số đã được sắp xếp theo thứ tự tăng dần hay giảm dần.

5.2. Phương pháp xác định số trung vị:

Trước tiên ta sắp xếp lượng biến theo thứ tự tăng dần.

- Tài liệu không phân tổ:

- Trường hợp n lẻ: số trung vị là lượng biến ở vị trí thứ (n+1)/2

$$Me = x_{(n+1)/2}$$

- Trường hợp n chẵn: số trung vị rơi vào giữa hai lượng biến $x_{n/2}$ và $x_{(n+2)/2}$. Trường hợp này qui ước số trung vị là trung bình cộng của hai lượng biến đó.

Ví dụ 2.7: thu nhập hàng hàng tháng của số công nhân sau:

500, 520, 530, 550, 560, 570, 590, 600, 610, 670

Số trung vị là: Me = (560+570)/2 = 565

•Tài liệu phân tổ có khoảng cách tổ:

Trong trường hợp này ta tìm tổ chứa số trung vị. Trước hết ta tính ((f_i/2) và đem so sánh với tần số tích lũy của tổ. Giá trị ((f_i/2) thuộc tổ nào thì tổ đó chứa số trung vị.

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

$x_{Me(min)}$: Giới hạn dưới của tổ chứa số trung vị

k_{Me}: Trị số khoảng cách tổ chứa số trung vị

f_{Me}: Tần số của tổ chứa số trung vị

S_{Me-1}: Tần số tích lũy trước tổ chứa số trung vị

Ví dụ 2.8: Sử dụng số liệu của ví dụ trước ta tìm số trung vị. Ta có bảng:

Thu nhập hàng tháng (ngàn đồng)	**Số nhân viên**	**Tần số tích lũy**
500-520	8	8
520-540	12	20
540-560	20	$40=S_{Me-1}$
560-580	$56=f_{Me}$	96
580-600	18	114
600-620	16	130
Trên 620	10	140
Tổng	**140**	

Như vậy số trung vị rơi vào tổ: 560-580

$X_{Me(min)} = 560$

$f_{Me} = 56$

$S_{Me-1} = 40$

Thay vào công thức, ta có:

$$Me = 560 + 20\frac{\sum 140/2 - 40}{56} = 570{,}714$$

6. Mốt – Mo (mode)

•**Định nghĩa:** Một là lượng biến có tần số xuất hiện lớn nhất trong tổng thể. Số Mo là giá trị thể hiện tính phổ biến của hiện tượng, tức là dữ liệu tập trung nhiều ở một khoảng giá trị nào đó. Trong thực tế người ta có thể sử dụng giá trị này trong sản xuất giày, quần áo may sẵn,…

•**Phương pháp xác định Mo:**

Ta phân biệt 2 trường hợp:

- Trường hợp tài liệu phân tổ không có khoảng cách tổ: (Phân tổ thuộc tính) thì đại lượng là Mo lượng biến có tần số lớn nhất.

- Trường hợp tài liệu phân tổ có khoảng cách tổ: trước hết ta xác định tổ chứa Mo, tổ chứa Mo là tổ có tần số lớn nhất, sau đó trị số gần đúng của Mốt được xác định theo công thức sau:

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

$x_{Mo(min)}$: Giới hạn dưới 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

k_{Mo}: Trị số khoảng cách tổ chứa Mốt

Trở lại ví dụ trước ta tính Mốt về thu nhập:

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

Chúng ta đã nghiên cứu các số đo tập trung biểu thị khuynh hướng tập trung của tổng thể, tức là nghiên cứu đại lượng mang tính chất đại diện cho tổng thể. Không có một số đo duy nhất nào có thể mô tả một cách đầy đủ cho một tổng thể. Tùy theo mục đích nghiên cứu ta cần xem xét để vận dụng các số đo cho phù hợp. Tuy nhiên, trong thực tế số trung bình được sử dụng rộng rãi vì dựa vào số trung bình người ta phát triển nhiều cơ sở suy luận để xây dựng các lý thuyết và tính các số đo khác.

IV. SỐ ĐO ĐỘ PHÂN TÁN (Measure of dispersion)

Độ biến thiên của tiêu thức dùng để đánh giá mức độ đại diện của số bình quân đối với tổng thể được nghiên cứu. Trị số này tính ra càng lớn, độ biến thiên của tiêu thức càng lớn do đó mức độ đại diện của số bình quân đối với tổng thể càng thấp và ngược lại.

Quan sát độ biến thiên tiêu thức trong dãy số lượng biến sẽ thấy nhiều đặc trưng về phân phối, kết cấu, tính đồng đều của tổng thể.

Độ biến thiên của tiêu thức được sử dụng nhiều trong nghiên cứu thống kê như phân tích biến thiên cũng như mối liên hệ của hiện tượng, dự đoán thống kê, điều tra chọn mẫu,...

Khi nghiên cứu độ biến thiên của tiêu thức, thống kê thường dùng các chỉ tiêu như khoảng biến thiên, độ lệch tuyệt đối bình quân, phương sai, độ lệch tiêu chuẩn và hệ số biến thiên. Dưới đây là nội dung và phương pháp tính của các chỉ tiêu đó.

1. Khoảng biến thiên (Range)

Khoảng biến thiên (còn gọi là toàn cự) là chỉ tiêu được tính bằng hiệu số giữa lượng biến lớn nhất và lượng biến nhỏ nhất của một dãy số lượng biến. Khoảng biến thiên càng lớn, mức độ biến động của chỉ tiêu càng lớn. Ngược lại, khoảng biến thiên nhỏ, mức độ biến động của chỉ tiêu thấp, tức là mức độ đồng đều của chỉ tiêu cao.

Công thức:

$$R = X_{max} - X_{min}$$

Trong đó:

R - Toàn cự;

X_{max} - Lượng biến có trị số lớn nhất

X_{min} - Lượng biến có trị số nhỏ nhất

Ví dụ 2.9: Thu nhập của hộ gia đình như sau:

Hộ	1	2	3	4	5	6	7	8
Thu nhập (1000 đồng)	6.000	7.000	85.000	86.000	9.000	9.100	9.500	10.000

Từ số liệu bảng, sử dụng công thức ở trên ta tính được khoảng biến thiên:

$$R = 10.000 – 6.000 = 4000 \text{ (nghìn đồng)}$$

Khoảng biến thiên phản ánh khoảng cách biến động của tiêu thức tuy tính toán đơn giản song phụ thuộc vào lượng biến lớn nhất và nhỏ nhất của tiêu thức, tức là không tính gì đến mức độ khác nhau của các lượng biến còn lại trong dãy số.

2. Độ lệch tuyệt đối trung bình (Mean Absolute Deviation)

Độ lệch tuyệt đối bình quân là số bình quân số học của các độ lệch tuyệt đối giữa các lượng biến với số bình quân số học của các lượng biến đó.

Công thức:

Trường hợp tính giản đơn $\overline{d} = \dfrac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n}$;

Trường hợp có quyền số $\overline{d} = \dfrac{\sum_{i=1}^{k}|x_i - \overline{x}|f_i}{\sum_{i=1}^{k} f_i}$;

Trong đó:

$\overline{d}$ - Độ lệch tuyệt đối bình quân;

x_i - Các trị số của lượng biến;

$\overline{x}$ - Số trung bình số học;

f_i - Quyền số của từng lượng biến x_i;

n - Tổng số lượng biến ($n = \sum_{i=1}^{k} f_i$).

Chỉ tiêu này biểu hiện độ biến thiên của tiêu thức nghiên cứu một cách đầy đủ hơn khoảng biến thiên. Qua đó phản ánh rõ nét hơn tính chất đồng đều của tổng thể: vì nó tính đến độ lệch của tất cả các lượng biến. Về cách tính cũng tương đối đơn giản, nhưng có đặc điểm là phải lấy giá trị tuyệt đối (giá trị dương) của chênh lệch.

Ví dụ 2.10: Có số liệu về năng suất lao động năm của công nhân trong một doanh nghiệp:

STT	Năng suất lao động năm (Triệu đồng /người)	Số công nhân (Ngàn người)	STT	Năng suất lao động năm (Triệu đồng /người)	Số công nhân (Ngàn người)
A	1	2	A	1	2
1	10	10	4	25	10
2	15	20	5	35	10
3	20	50			

a. Số bình quân

$$\bar{x} = \frac{(10.10) + (15.20) + (20.50) + (25.10) + (35.10)}{10 + 20 + 50 + 10 + 10} = 20$$

b. Độ lệch tuyệt đối bình quân

$$\bar{d} = \frac{|10 - 20|10 + |15 - 20|20 + |20 - 20|50 + |25 - 20|10 + |35 - 20|10}{10 + 20 + 50 + 10 + 10}$$

$$= \frac{400}{100} = 4$$

3. Phương sai (Variance)

Phương sai là số bình quân số học của bình phương các độ lệch giữa các lượng biến với số bình quân số học của các lượng biến đó.

Phương sai là sai số trung bình bình phương giữa các lượng biến và số trung bình số học của các lượng biến đó.

3.1. Phương sai tổng thể:

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

x_i: Giá trị lượng biến thứ i

μ: Trung bình tổng thể

N: Số đơn vị tổng thể

3.2. Phương sai mẫu:

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

Nếu dãy số có tần số f_i thì:

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

Trong công thức phương sai mẫu người ta gọi tử số là tổng độ lệch bình phương và mẫu số là bậc tự do.

Chú ý, đối với công thức phương sai mẫu, theo toán học người ta chia ra thành 2 loại là phương sai mẫu và phương sai mẫu điều chỉnh. Tuy nhiên phương sai mẫu (bậc tự do là n) là ước lượng chệch của phương sai của tổng thể, còn phương sai mẫu là ước lượng không chệch. Chính vì vậy, để cho đơn giản chúng ta hiểu phương sai mẫu ở đây là phương sai mẫu đã điều chỉnh theo quan điểm của toán học.

4. Độ lệch chuẩn (Standard deviation)

4.1. Độ lệch chuẩn của tổng thể:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

4.2. Độ lệch chuẩn của mẫu:

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

5. Hệ số biến thiên (Coefficient of Variation)

Hệ số biến thiên là chỉ tiêu tương đối phản ánh mối quan hệ so sánh giữa độ lệch chuẩn với số bình quân số học.

Công thức:

$$V = \frac{\sigma}{\bar{x}}$$

Trong đó:

V - Hệ số biến thiên;

σ - Độ lệch chuẩn;

$\bar{x}$ - Số bình quân số học.

Hệ số biến thiên cũng dùng để đánh giá độ biến thiên của tiêu thức và tính chất đồng đều của tổng thể. Hệ số này biểu hiện bằng số tương đối nên còn có thể được dùng để so sánh cả những chỉ tiêu cùng loại nhưng ở các quy mô khác nhau như so sánh độ đồng đều về thu nhập bình quân của hộ gia đình ở khu vực nông thôn (có thu nhập thấp và số hộ ít hơn) với thu nhập bình quân của hộ gia đình ở thành thị (có mức thu nhập cao hơn và số hộ nhiều hơn), đặc biệt để so sánh được những chỉ tiêu của các hiện tượng khác nhau và có đơn vị đo lường khác nhau như so sánh hệ số biến thiên về bậc thợ với hệ số biến thiên về tiền lương bình quân, hệ số biến thiên về năng suất lao động bình quân, so sánh hệ số biến thiên về chỉ tiêu thu nhập của hộ gia đình với hệ số biến thiên về chi tiêu của hộ gia đình,...

Hệ số biến thiên còn có thể tính theo độ lệch tuyệt đối bình quân, nhưng hệ số biến thiên tính theo độ lệch chuẩn thường được sử dụng rộng rãi hơn, tuy phần tính toán có phức tạp hơn phải sử dụng độ lệch tuyệt đối trung bình.

Hệ số biến thiên tính theo độ lệch tuyệt đối bình quân có công thức tính:

$$V = \frac{\bar{d}}{\bar{x}}$$

Trong đó: $\bar{d}$ - Độ lệch tuyệt đối bình quân.

V. PHƯƠNG PHÁP CHỈ SỐ

Hiện nay các nhà doanh nghiệp có thể nắm bắt thông tin trên nhiều phương tiện thông tin khác nhau, họ quan tâm đến giá cả hay khối lượng sản phẩm từng mặt hàng hay nhiều mặt hàng tăng lên hay giảm xuống qua thời gian trên một thị trường hay nhiều thị trường khác nhau. Những thông tin này được tính toán thông qua phương pháp chỉ số. Một cách tổng quát, chỉ số đo lường sự thay đổi của hiện tượng kinh tế qua hai thời gian và không gian nghiên cứu.

Một số ký hiệu thường dùng trong phương pháp chỉ số:

p: Giá hàng hóa nói chung

z: Giá thành

q: Khối lượng sản phẩm (chỉ tiêu số lượng)

i: Chỉ số cá thể

I: Chỉ số chung, chỉ số tổng hợp

(0): Thể hiện kỳ gốc

(1): Thể hiện kỳ báo cáo hay kỳ nghiên cứu

Các loại chỉ số và cách tính:

Căn cứ vào phạm vi tính toán, có 2 loại chỉ số tương ứng với việc nghiên cứu hai loại chỉ tiêu chất lượng và số lượng.

- Chỉ tiêu chất lượng: giá cả, giá thành sản phẩm, năng suất thu hoạch, năng suất lao động, mức nguyên liệu cần thiết để sản xuất một thành phẩm,...

- Chỉ tiêu khối lượng: lượng hàng hóa tiêu thụ, lượng hàng hóa sản xuất, số lượng lao động,...

1. Chỉ số cá thể

Là loại chỉ số nghiên cứu sự biến động về một chỉ tiêu nào đó của từng loại đơn vị, từng phần tử của hiện tượng.

- Chỉ số cá thể giá cả (chỉ số chất lượng):

$$i_{pi} = \frac{p_{i(1)}}{p_{i(0)}}$$

$p_{i(1)}$: là giá cả mặt hàng thứ i kỳ nghiên cứu

$p_{i(0)}$: là giá cả mặt hàng thứ i kỳ gốc.

- Chỉ số cá thể khối lượng:

$$i_{qi} = \frac{q_{i(1)}}{q_{i(0)}}$$

$q_{i(1)}$: là khối lượng mặt hàng thứ i kỳ nghiên cứu

$q_{i(0)}$: là khối lượng mặt hàng thứ i kỳ gốc.

2. Chỉ số tổng hợp

Là loại chỉ tiêu chỉ nghiên cứu sự biến động về một chỉ tiêu nào đó của nhiều đơn vị, nhiều phần tử của hiện tượng phức tạp.

2.1. Chỉ số tổng hợp giá cả:

- Chỉ số tổng hợp giá cả đơn giản:

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}}$$

$p_{i(1)}$: là giá cả mặt hàng thứ i kỳ nghiên cứu

$p_{i(0)}$: là giá cả mặt hàng thứ i kỳ gốc.

- Chỉ số tổng hợp giá cả có quyền số:

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)} q_i}{\sum_{i=1}^{n} p_{i(0)} q_i}$$

$p_{i(1)}$: là giá cả mặt hàng thứ i kỳ nghiên cứu

$p_{i(0)}$: là giá cả mặt hàng thứ i kỳ gốc.

q_i : là quyền số của mặt hàng thứ i

Quyền số của mặt hàng thứ i có 2 giá trị: giá trị chọn ở kỳ định gốc hoặc là chọn ở kỳ nghiên cứu. Tùy theo cách chọn quyền số ta có 2 phương pháp tính:

a) Phương pháp Laspeyres:

Nếu trọng số là lượng hàng hóa thiêu thụ chọn ở kỳ gốc làm căn bản để so sánh, ta có công thức chỉ số giá cả của Laspeyres:

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)} q_{i(0)}}{\sum_{i=1}^{n} p_{i(0)} q_{i(0)}}$$

Ví dụ 2.11: Ta có số liệu về giá cả và lượng hàng tiêu thụ của 4 mặt hàng sau, tính chỉ số giá theo phương pháp Laspeyres:

TT	Mặt hàng	Giá cả (1.000đ)		Lượng hàng tiêu thụ		$p_{i(0)}q_{i(0)}$	$p_{i(1)}q_{i(0)}$
		2000 ($p_{i(0)}$)	2005 ($p_{i(1)}$)	2000 ($q_{i(0)}$)	2005 ($q_{i(1)}$)		
1	Sữa (hộp)	3,0	5,0	50.000	190.000	150.000	250.000
2	Lúa (kg)	1,6	2,4	100.000	120.000	160.000	240.000
3	Dầu (lít)	2,4	3,6	200.000	360.000	480.000	720.000
4	Máy tính bỏ túi (chiếc)	40,0	25,0	4.000	4.200	160.000	100.000
	Tổng					**950.000**	**1.310.000**

Về số tương đối:

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)} q_{i(0)}}{\sum_{i=1}^{n} p_{i(0)} q_{i(0)}} = \frac{1.310.000}{950.000} = 137,9\%$$

Về số tuyệt đối:

$$\sum_{i=1}^{n} p_{i(1)} q_{i(0)} - \sum_{i=1}^{n} p_{i(0)} q_{i(0)} = 1.310.000 - 950.000 = 360.000$$

Kết luận: Nhìn chung giá cả 4 mặt hàng trên năm 2005 so với năm 2000 bằng 137,9%, tăng 37,9% làm tăng giá trị tiêu thụ (hay doanh số tiêu thụ) một lượng là 360.000 (ngàn đồng).

Cách tính chỉ số giá theo phương pháp Laspeyres có những ưu điểm sau:

Thứ nhất, với trọng số là lượng hàng hóa tiêu thụ ở thời kỳ gốc; ta chỉ cần số liệu lượng hàng hóa tiêu thụ và giá cả của một thời kỳ căn bản nào đó để làm căn cứ so sánh.

Thứ hai, trên cơ sở chỉ số giá cả tính toán so với cùng kỳ gốc, phương pháp Laspeyres cho phép ta xác định sự thay đổi giá cả giữa hai thời gian nghiên cứu bất kỳ.

Tuy nhiên, phương pháp Laspeyres có nhược điểm là không phản ánh, cập nhật được những thay đổi về khuynh hướng, thói quen của người tiêu dùng. Một số mặt hàng nào đó vài năm trước được người tiêu thụ mua với số lượng lớn, nhưng có thể ngày nay không còn quan trọng đối với họ nữa.

b) Phương pháp tính chỉ số Paasche:

Ngược lại với chỉ số Laspeyres, chỉ số Paasche chọn lượng sản phẩm tiêu thụ ở kỳ nghiên cứu làm trọng số. Ta có công thức tính chỉ số giá của Paasche:

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)} q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)} q_{i(1)}}$$

Ví dụ 2.12: Từ số liệu ở ví dụ 2.11, tính chỉ số giá theo phương pháp Paasche:

TT	Mặt hàng	Giá cả (1.000đ)		Lượng hàng tiêu thụ		$p_{i(0)}q_{i(1)}$	$p_{i(1)}q_{i(1)}$
		2000 ($p_{i(0)}$)	2005 ($p_{i(1)}$)	2000 ($q_{i(0)}$)	2005 ($q_{i(1)}$)		
1	Sữa (hộp)	3,0	5,0	50.000	190.000	570.000	950.000
2	Lúa (kg)	1,6	2,4	100.000	120.000	192.000	288.000
3	Dầu (lít)	2,4	3,6	200.000	360.000	860.000	1.296.000
4	Máy tính bỏ túi (chiếc)	40,0	25,0	4.000	4.200	168.00	105.000
	Tổng					**1.794.000**	**2.639.000**

Về số tương đối:

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)} q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)} q_{i(1)}} = \frac{2.639.000}{1.794.000} = 147{,}1\%$$

Về số tuyệt đối:

$$\sum_{i=1}^{n} p_{i(1)} q_{i(1)} - \sum_{i=1}^{n} p_{i(0)} q_{i(1)} = 2.639.000 - 1.794.000 = 845.000$$

Kết luận: Nhìn chung giá cả 4 mặt hàng trên năm 2005 so với năm 2000 bằng 147,1%, tăng 47,1% làm tăng doanh thu một lượng là 845.000 (ngàn đồng).

Cách tính chỉ số giá theo phương pháp này khắc phục được nhược điểm của phương pháp Laspeyres. Ngoài ra phương pháp của Paasche còn cho thấy ảnh hưởng một cách cụ thể sự thay đổi giá cả đến người tiêu thụ. Nhưng một khó khăn khi sử dụng phương pháp Paasche là phải thường xuyên thu thập lại lượng hàng hóa tiêu thụ ở kỳ nghiên cứu.

Tuỳ theo mức độ số liệu thể thu thập được mà chúng ta có thể áp dụng một trong 3 trường hợp trên một cách linh hoạt nhưng phương pháp Paasche là phương pháp tỏ ra ưu việt hơn bởi số liệu mang tính cập nhật hơn và đây là phương pháp mà người ta thường dùng. Ngoài ra, người ta còn đưa ra một phương pháp khác mang tính dung hoà hơn đó là tính quyền số là trung bình của lượng hàng hoá tiêu thụ ở hai thời điểm.

2.2. Chỉ số tổng hợp khối lượng:

Chỉ số tổng hợp khối lượng về căn bản giống như chỉ số tổng hợp giá cả nhưng ngược lại nhân tố xem xét là khối lượng sản phẩm còn giá cả đóng vai trò là trọng số.

- Chỉ số tổng hợp khối lượng đơn giản:

$$I_q = \frac{\sum_{i=1}^{n} q_{i(1)}}{\sum_{i=1}^{n} q_{i(0)}}$$

$q_{i(1)}$: là khối lượng mặt hàng thứ i kỳ nghiên cứu.

$q_{i(0)}$: là khối lượng mặt hàng thứ i kỳ gốc.

a) Phương pháp Laspeyres:

$$I_q = \frac{\sum_{i=1}^{n} q_{i(1)} p_{i(0)}}{\sum_{i=1}^{n} q_{i(0)} p_{i(0)}}$$

Ví dụ 2.13: Từ số liệu ở ví dụ 2.11, tính chỉ số khối lượng theo phương pháp Laspeyres.

TT	Mặt hàng	Giá cả (1.000đ)		Lượng hàng tiêu thụ		$q_{i(0)}p_{i(0)}$	$q_{i(1)}p_{i(0)}$
		2000 ($p_{i(0)}$)	2005 ($p_{i(1)}$)	2000 ($q_{i(0)}$)	2005 ($q_{i(1)}$)		
1	Sữa (hộp)	3,0	5,0	50.000	190.000	150.000	570.000
2	Lúa (kg)	1,6	2,4	100.000	120.000	160.000	192.000
3	Dầu (lít)	2,4	3,6	200.000	360.000	480.000	864.000
4	Máy tính bỏ túi (chiếc)	40,0	25,0	4.000	4.200	160.000	168.000
	Tổng					**950.000**	**1.794.000**

Về số tương đối:

$$I_q = \frac{\sum_{i=1}^{n} q_{i(1)} p_{i(0)}}{\sum_{i=1}^{n} q_{i(0)} p_{i(0)}} = \frac{1.794.000}{950.000} = 188{,}8\%$$

Về số tuyệt đối:

$$\sum_{i=1}^{n} q_{i(1)} p_{i(0)} - \sum_{i=1}^{n} q_{i(0)} p_{i(0)} = 1.794.000 - 950.000 = 844.000$$

Kết luận: Nhìn chung lượng hàng tiêu thụ 4 mặt hàng trên năm 2005 so với năm 2000 bằng 188,8%, tăng 88,8% làm tăng giá trị tiêu thụ một lượng là 844.000 (ngàn đồng).

Phương pháp chỉ số tổng hợp khối lượng theo phương pháp Laspeyres thường được chọn vì các lý do sau:

Thứ nhất, với cách chọn giá cả ở kỳ gốc làm quyền số, phương pháp này giúp ta có thể tính toán chỉ số tổng hợp khối lượng một cách nhanh chóng.

Thứ hai, trên cơ sở chỉ số khối lượng tính toán so với kỳ gốc, phương pháp này cho phép ta xác định sự thay đổi khối lượng giữa hai thời kỳ nghiên cứu.

b) Phương pháp tính chỉ số Paasche:

$$I_q = \frac{\sum_{i=1}^{n} q_{i(1)} p_{i(1)}}{\sum_{i=1}^{n} q_{i(0)} p_{i(1)}}$$

Ví dụ 2.14: Từ số liệu ở ví dụ 2.11, tính chỉ số khối lượng theo phương pháp Paasche

TT	Mặt hàng	Giá cả (1.000đ)		Lượng hàng tiêu thụ		$q_{i(0)}p_{i(1)}$	$q_{i(1)}p_{i(1)}$
		2000 ($p_{i(0)}$)	2005 ($p_{i(1)}$)	2000 ($q_{i(0)}$)	2005 ($q_{i(1)}$)		
1	Sữa (hộp)	3,0	5,0	50.000	190.000	250.000	950.000
2	Lúa (kg)	1,6	2,4	100.000	120.000	240.000	288.000
3	Dầu (lít)	2,4	3,6	200.000	360.000	720.000	1.296.000
4	Máy tính bỏ túi (chiếc)	40,0	25,0	4.000	4.200	100.000	105.000
	Tổng					**1.310.000**	**2.639.000**

Về số tương đối:

$$I_q = \frac{\sum_{i=1}^{n} q_{i(1)} p_{i(1)}}{\sum_{i=1}^{n} q_{i(0)} p_{i(1)}} = \frac{2.639.000}{1.310.000} = 201{,}5\%$$

Về số tuyệt đối:

$$\sum_{i=1}^{n} q_{i(1)} p_{i(1)} - \sum_{i=1}^{n} q_{i(0)} p_{i(1)} = 2.639.000 - 1.310.000 = 1.329.000$$

Kết luận: Nhìn chung lượng hàng tiêu thụ 4 mặt hàng trên năm 2005 so với năm 2000 bằng 201,5%, tăng 101,5% làm tăng giá trị tiêu thụ một lượng là 1.329.000 (ngàn đồng).

Theo thương pháp Paasche, xét về nội dung có ý nghĩa kinh tế hơn là chỉ số tổng hợp theo phương pháp Laspeyres. Tuy nhiên, nếu xét trong mối tương quan giữa mặt lượng và mặt chất tác động đến một hiện tượng (Doanh thu phụ thuộc vào giá bán và lượng hàng hoá tiêu thụ) thì người ta sử dụng phương pháp Laspeyres, do đó đây là phương pháp thường được sử dụng.

Nếu chỉ nghiên cứu riêng chỉ số tổng hợp khối lượng thì tuỳ theo số liệu thu thập mà ta có thể chọn phương pháp phù hợp nhưng với công thức đơn giản chúng ta cẩn thận, trong nhiều trường hợp tính toán công thức này không có ý nghĩa.

3. Chỉ số trung bình tính từ chỉ số tổng hợp

3.1. Chỉ số trung bình điều hòa về biến động của chỉ tiêu chất lượng: Trong trường hợp tài liệu chỉ có giá trị ở kỳ báo cáo và chỉ số giá cả cá thể:

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)} q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)} q_{i(1)}} = \frac{\sum_{i=1}^{n} p_{i(1)} q_{i(1)}}{\sum_{i=1}^{n} \frac{p_{i(0)}}{p_{i(1)}} p_{i(1)} q_{i(1)}} = \frac{\sum_{i=1}^{n} p_{i(1)} q_{i(1)}}{\sum_{i=1}^{n} \frac{1}{i_{pi}} p_{i(1)} q_{i(1)}}$$

Ví dụ 2.15: Có số liệu sau đây của một công ty:

Mặt hàng	Doanh thu năm 2007	Thay đổi giá bán năm 2007 so với năm 2006
A	5.408	+4
B	6.175	-5
C	9.996	+2

Tính chỉ số giá của 3 mặt hàng trên.

Từ số liệu trên ta có bảng:

Mặt hàng	Doanh thu năm 2007 ($p_{i(1)}q_{i(1)}$)	Thay đổi giá bán năm 2007 so với năm 2006 (i_{pi})	$(p_{i(1)}q_{i(1)})/ (i_{pi})$
A	5.408	104	5.200
B	6.175	95	6.500
C	9.996	102	9.800
Tổng	**21.579**		**21.500**

Thay vào công thức ta có:

$$I_p = \frac{\sum_{i=1}^{n} p_{i(1)}q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(1)}} = \frac{\sum_{i=1}^{n} p_{i(1)}q_{i(1)}}{\sum_{i=1}^{n} \frac{1}{i_{pi}} p_{i(1)}q_{i(1)}} = \frac{21.579}{21.500} = 100,37\%$$

Kết luận: Giá cả 3 mặt hàng năm 2007 so với năm 2006 bằng 100,37%, tăng 0,37%.

3.2. Chỉ số trung bình số học về biến động của chỉ tiêu khối lượng: Trong trường hợp tài liệu chỉ cho giá trị kỳ gốc và chỉ số khối lượng cá thể.

$$I_q = \frac{\sum_{i=1}^{n} q_{i(1)}p_{i(0)}}{\sum_{i=1}^{n} q_{i(0)}p_{i(0)}} = \frac{\sum_{i=1}^{n} \frac{q_{i(1)}}{q_{i(0)}} q_{i(0)}p_{i(0)}}{\sum_{i=1}^{n} q_{i(0)}p_{i(0)}} = \frac{\sum_{i=1}^{n} i_{qi}q_{i(0)}p_{i(0)}}{\sum_{i=1}^{n} q_{i(0)}p_{i(0)}}$$

Ví dụ 2.16: Có số liệu sau đây của một công ty:

Mặt hàng	Doanh thu năm 2006	Tỷ lệ lượng hàng bán năm 2007 so với năm 2006
A	2.000	104
B	5.000	96
C	3.000	105

Tính chỉ số khối lượng hàng hóa tiêu thụ của 3 mặt hàng trên.

Từ số liệu trên ta có bảng:

Mặt hàng	Doanh thu năm 2006 ($p_{i(0)}q_{i(0)}$)	Thay đổi giá bán năm 2007 so với năm 2006 (i_{qi})	$i_{pi}\ p_{i(0)}q_{i(0)}$
A	2.000	104	2.080
B	5.000	96	4.800
C	3.000	105	3.150
Tổng	**10.000**		**10.030**

Thay vào công thức ta có:

$$I_q = \frac{\sum_{i=1}^{n} q_{i(1)} p_{i(0)}}{\sum_{i=1}^{n} q_{i(0)} p_{i(0)}} = \frac{\sum_{i=1}^{n} i_{qi} q_{i(0)} p_{i(0)}}{\sum_{i=1}^{n} q_{i(0)} p_{i(0)}} = \frac{10.030}{10.000} = 100,3\%$$

Kết luận: Khối lượng hàng hóa tiêu thụ của công ty tính chung 3 mặt hàng năm 2007 so với năm 2006 bằng 100,3%, tăng 0,3%.

4. Chỉ số không gian

Là chỉ số so sánh các hiện tượng cùng loại nhưng qua các điều kiện không gian khác nhau. Ví dụ, nghiên cứu sự biến động về lượng hàng bán ra và giá cả các mặt hàng ở hai thị trường khác nhau.

4.1. Chỉ số tổng hợp nghiên cứu sự biến động của chỉ tiêu chất lượng ở hai thị trường A và B.

$$I_p(A/B) = \frac{\sum_{i=1}^{n} p_{Ai} q_i}{\sum_{i=1}^{n} p_{Bi} q_i}$$

Trong đó, $q_i = q_{Ai} + q_{Bi}$: Khối lượng sản phẩm cùng loại ở hai thị trường A và B.

Ví dụ 2.17: Tình hình tiêu thụ mặt hàng X và Y tại hai chợ A và B. Hãy tính sự biến động về giá cả ở hai thị trường trên.

Mặt hàng	Thị trường A		Thị trường B	
	Lượng bán	Giá đơn vị	Lượng bán	Giá đơn vị
X	480	12.000	520	10.000
Y	300	10.000	200	18.000

Ta có:

$Q_X = q_{AX} + q_{BX} = 480 + 520 = 1.000$

$Q_Y = q_{AY} + q_{BY} = 300 + 200 = 500$

$$I_p(A/B) = \frac{\sum p_{Ai} q_i}{\sum p_{Bi} q_i} = \frac{p_{AX} q_X + p_{AY} q_Y}{p_{BX} q_X + p_{BY} q_Y} =$$

$$= \frac{(12.000 \text{x} 1.000) + (10.000 \text{x} 500)}{(10.000 \text{x} 1.000) + (18.000 \text{x} 500)} = \frac{17.000.000}{19.000.000} = 89,5\%$$

Về số tuyệt đối: (17.000.000 - 19.000.000) = -2.000.000

Kết luận: Nói chung giá cả hai mặt hàng ở thị trường A thấp hơn thị trường B là 10,5%.giá trị tiêu thụ ở thj trường A thấp hơn thị trường B 2.000.000.

4.2. Chỉ số tổng hợp nghiên cứu sự biến động của chỉ tiêu khối lượng ở hai thị trường A và B: Trong trường hợp này có thể có các quyền số khác nhau là các chỉ tiêu chất lượng, chẳng hạn như giá cố định cho từng mặt hàng (p) hoặc tính giá trung bình từng mặt hàng ở hai thị trường).

$$I_q(A/B) = \frac{\sum_{i=1}^{n} q_{Ai} p_i}{\sum_{i=1}^{n} q_{Bi} p_i}$$

Trong đó: p_i là giá cố định cho mặt hàng i

Nếu ta chọn giá cố định là giá trung bình của hai mặt hàng ở từng thị trường, ta có:

$$I_q(A/B) = \frac{\sum q_{Ai} \overline{p_i}}{\sum q_{Bi} \overline{p_i}}$$

$\overline{p_i}$: Giá trung bình của mặt hàng i ở hai thị trường.

Ví dụ 2.18: Sử dụng số liệu của vị dụ 2.17, hãy tính sự biến động về khối lượng hàng tiêu thụ ở hai thị trường trên.

Mặt hàng	Thị trường A		Thị trường B	
	Lượng bán	Giá đơn vị	Lượng bán	Giá đơn vị
X	480	12.000	520	10.000
Y	300	10.000	200	18.000

Sử dụng số bình quân gia quyền ta tính giá cố định của hai mặt hàng.

$$\overline{p}_X = \frac{(12.000x480) + (10.000x520)}{480 + 520} = 10.960$$

$$\overline{p}_Y = \frac{(10.000x300) + (18.000x200)}{300 + 200} = 13.200$$

$$I_q(A/B) = \frac{\sum q_{Ai} \overline{p_i}}{\sum q_{Bi} \overline{p_i}} = \frac{q_{AX} \overline{p_X} + q_{AY} \overline{p_Y}}{q_{BX} \overline{p_X} + q_{BY} \overline{p_Y}}$$

$$\frac{(480x10.960) + (300x13.200)}{(300x10.960) + (200x13.200)} = \frac{9.220.800}{8.339.200} = 110,6\%$$

Về số tuyệt đối: 9.200.800 - 8.339.200 = 881.600.

Kết luận: Nói chung lượng hàng tiêu thụ ở thị trường A cao hơn thị trường B là 10,6%, làm tăng giá trị tiêu thụ của thị trường A cao hơn thị trường B là 881.600.

5. Hệ thống chỉ số liên hoàn 2 nhân tố

Phương pháp chỉ số giúp nghiên cứu sự thay đổi của hiện tượng kinh tế qua thời gian. Ngoài phương pháp chỉ số còn có thể phân tích mối liên hệ giữa các nhân tố và mức độ ảnh hưởng của các nhân tố đến sự thay đổi một chỉ tiêu kinh tế tổng hợp. Ví dụ, doanh số bán của một công ty biến động là do ảnh hưởng của hai nhân tố: giá bán và khối lượng hàng tiêu thụ, sản lượng thu hoạch của một loại cây trồng do ảnh hưởng của hai loại nhân tố: năng suất thu hoạch và diện tích gieo trồng,...

Giả sử cần phân tích tổng mức hàng hóa tiêu thụ biến động qua hai thời kỳ nghiên cứu trong mốt liên hệ giữa hai nhân tố: giá cả và khối lượng hàng hóa tiêu thu. Hệ thống chỉ số thể hiện mối liên hệ giữa các chỉ tiêu trên như sau:

$$I_{pq} = I_p \text{ x } I_q$$

Trong đó:

I_{pq}: Chỉ số tổng mức hàng hóa tiêu thụ

I_p: Chỉ số giá được xác định theo phương pháp Paasche

I_q: Chỉ số khối lượng được xác định theo phương pháp Laspeyres

$$\frac{\sum_{i=1}^{n} p_{i(1)}q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(0)}} = \frac{\sum_{i=1}^{n} p_{i(1)}q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(1)}} \text{ x } \frac{\sum_{i=1}^{n} p_{i(0)}q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(0)}}$$

Về số tuyệt đối:

$$\sum_{i=1}^{n} p_{i(1)}q_{i(1)} - \sum_{i=1}^{n} p_{i(0)}q_{i(0)} = (\sum_{i=1}^{n} p_{i(1)}q_{i(1)} - \sum_{i=1}^{n} p_{i(0)}q_{i(1)}) + (\sum_{i=1}^{n} p_{i(0)}q_{i(1)} - \sum_{i=1}^{n} p_{i(0)}q_{i(0)})$$

Về số tương đối so với giá trị tiêu thụ kỳ gốc:

$$\frac{\sum_{i=1}^{n} p_{i(1)}q_{i(1)} - \sum_{i=1}^{n} p_{i(0)}q_{i(0)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(0)}} = \frac{\sum_{i=1}^{n} p_{i(1)}q_{i(1)} - \sum_{i=1}^{n} p_{i(0)}q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(0)}} + \frac{\sum_{i=1}^{n} p_{i(0)}q_{i(1)} - \sum_{i=1}^{n} p_{i(0)}q_{i(0)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(0)}}$$

Ví dụ 2.19: Từ số liệu ví dụ 2.11, nghiên cứu sự ảnh hưởng của giá bán và lượng bán đến doanh số bán.

Ta có bảng số liệu:

TT	**Mặt hàng**	**Giá cả (1.000đ)**		**Lượng hàng tiêu thụ**		$p_{i(1)}q_{i(1)}$	$p_{i(0)}q_{i(0)}$	$p_{i(0)}q_{i(1)}$
		2000 ($p_{i(0)}$)	**2005 ($p_{i(1)}$)**	**2000 ($q_{i(0)}$)**	**2005 ($q_{i(1)}$)**			
1	Sữa (hộp)	3,0	5,0	50.000	190.000	950.000	150.000	570.000
2	Lúa (kg)	1,6	2,4	100.000	120.000	288.000	160.000	192.00
3	Dầu (lít)	2,4	3,6	200.000	360.000	1.296.000	480.000	864.000
4	Máy tính bỏ túi (chiếc)	40,0	25,0	4.000	4.200	105.000	160.000	168.00
	Tổng					**2.639.000**	**950.000**	**1.794.000**

Nhận xét về số tương đối:

$$\frac{\sum_{i=1}^{n} p_{i(1)}q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(0)}} = \frac{\sum_{i=1}^{n} p_{i(1)}q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(1)}} \text{ x } \frac{\sum_{i=1}^{n} p_{i(0)}q_{i(1)}}{\sum_{i=1}^{n} p_{i(0)}q_{i(0)}}$$

$$\frac{2.639.000}{950.000} = \frac{2.639.000}{1.794.000} \text{ x } \frac{1.794.000}{950.000}$$

277,8% = 147,1% x 188,8%

(tăng 177,8%) (tăng 47%) (tăng 88,8%)

Nhận xét về số tuyệt đối:

$$\sum_{i=1}^{n} p_{i(1)}q_{i(1)} - \sum_{i=1}^{n} p_{i(0)}q_{i(0)} = (\sum_{i=1}^{n} p_{i(1)}q_{i(1)} - \sum_{i=1}^{n} p_{i(0)}q_{i(1)}) + (\sum_{i=1}^{n} p_{i(0)}q_{i(1)} - \sum_{i=1}^{n} p_{i(0)}q_{i(0)})$$

(2.639.000-950.000)=(2.639.000-1.794.000)+(1.794.000-950.000)

1.689.000=845.000+844.000

Kết luận: Tổng mức tiêu thụ hàng hóa tính chung 4 mặt hàng năm 2005 so với năm 2000 bằng 277,8% (tăng 117,8%), mức tăng là 1.689.000 ngàn đồng là do ảnh hưởng của 2 nhân tố liên quan:

- Do giá cả năm 2005 so với năm 2000 tăng 47,1% làm tăng giá trị tiêu thụ là 845.000 ngàn đồng.

- Do khối lượng các mặt hàng bán ra nói chung tăng 88,8% làm tăng giá trị tiêu thụ là 844.000 ngàn đồng.

Hệ thống liên hoàn hai nhân tố người ta có thể mở rộng ra thành hệ thống liên hoàn nhiều nhân tố và chúng ta cũng thực hiện theo nguyên tắc chỉ số chất lượng thì sử dụng phương pháp Paasche còn chỉ số số lượng thì sử dụng phương pháp Laspeyres. Tuy nhiên, việc mở rộng này tỏ ra không hiệu quả cao vì nó có thể làm cho việc phân tích quá phức tạp.

PHẦN III

THỐNG KÊ SUY LUẬN

CHƯƠNG III
PHÂN PHỐI CHUẨN VÀ PHÂN PHỐI MẪU

Đối với các khái niệm và tính chất có liên quan đến phân phối của tổng thể chúng ta đã nghiên cứu ở môn học xác suất thống kê toán. Ở đây chỉ mang tính chất nhắc lại một cách khái quát nhất.

I. PHÂN PHỐI CHUẨN

Phân phối chuẩn chiếm một vị trí rất quan trọng trong lý thuyết thống kê nó liên quan đến các kết luận thống kê suy luận sau này. Trong thực tế, nhiều biến ngẫu nhiên tuân theo qui luật phân phối chuẩn hoặc gần chuẩn, chẳng hạn như trọng lượng và chiều cao của người lớn, mức độ thông minh của trẻ em, điểm thi của các thí sinh, lực chịu đựng của một thanh sắt, các sai số đo đạc, …. Do đó, việc nhắc lại là rất cần thiết.

1. Định nghĩa

Phân phối chuẩn là phân phối của đại lượng ngẫu nhiên liên tục X có miền xác định từ -∞ đến +∞ với hàm mật độ xác suất:

$$f_x = \frac{1}{\sigma.\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$, trong đó: e = 2,71828, π=3,1415

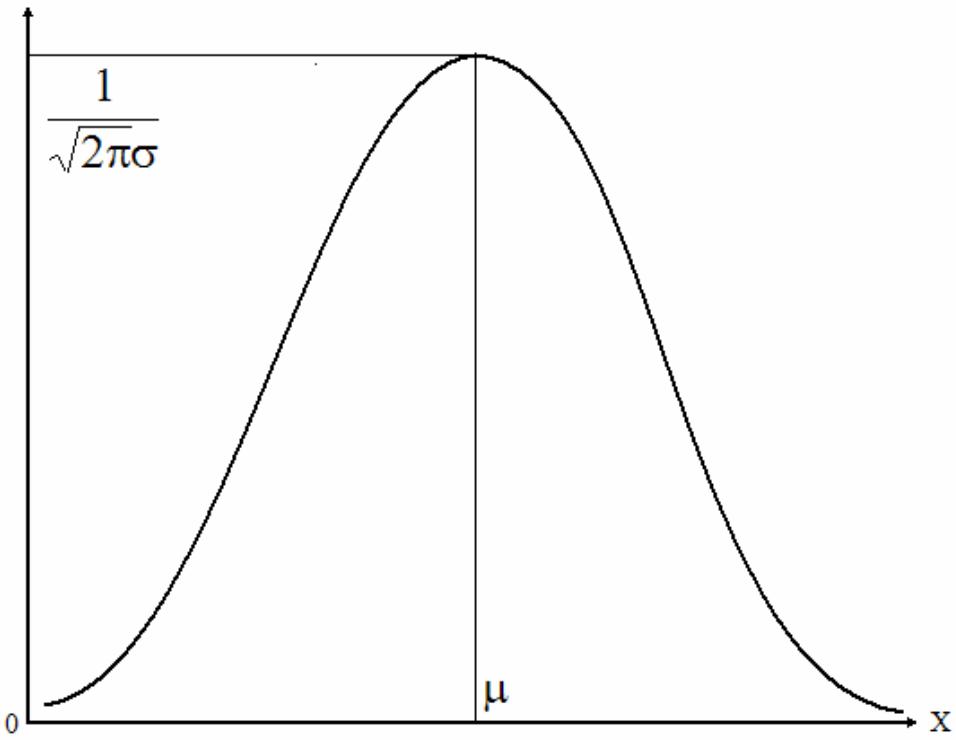
Ký hiệu: $X \sim N(\mu,\sigma^2)$

- *Tính chất của hàm phân phối chuẩn:*

- $\int_{-\infty}^{+\infty} \frac{1}{\sigma.\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} .dx = 1$. Chính là diện tích giới hạn bởi đồ thị f(x) và trục hoành.

- Đồ thị đối xứng với nhau qua đường thẳng x=μ

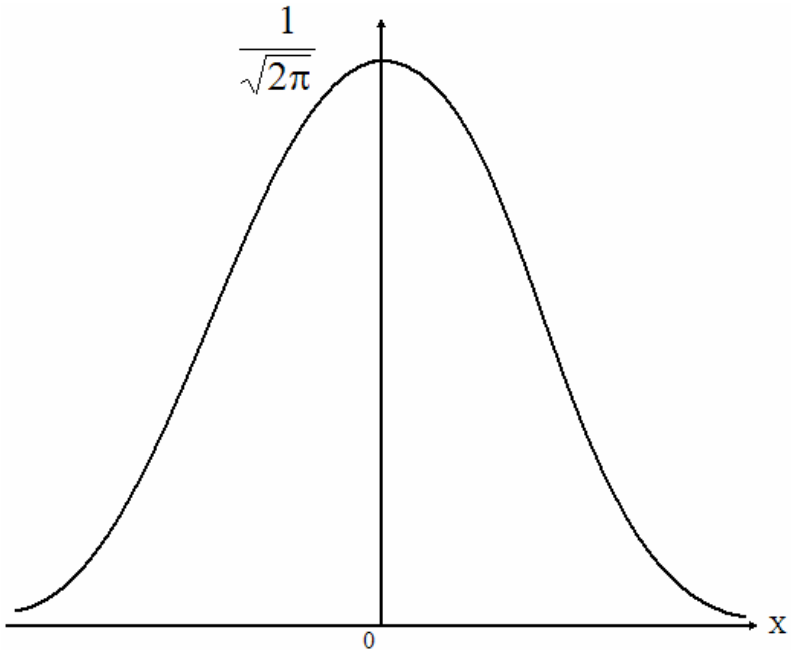
- X có trung bình là μ và phương sai là σ^2


1
√2πσ
0
μ
x

2. Phân phối chuẩn tắc (đơn giản)

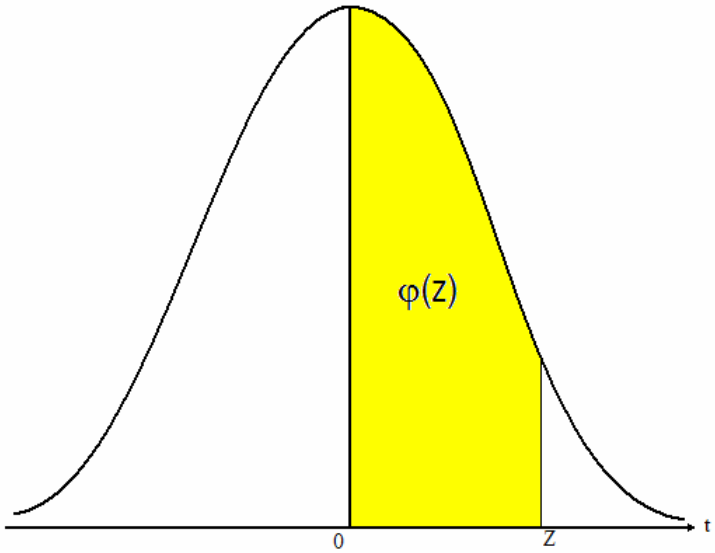
Định nghĩa: là phân phối chuẩn có μ=0 và σ=1. $f_t = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}$

Ta có thể dùng phương pháp đổi biến t=(x-μ)/σ đối với phân phối chuẩn thành phân phối chuẩn tắc.



Bảng phân phối chuẩn tắc (đơn giản):

Hàm số $\varphi(z) = \int_0^z \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}$ được gọi là hàm tích phân Laplace



- Tính chất của hàm tích phân Laplace:

  - φ(Z) = p(0 < t < Z)

  - φ(z) là hàm số lẽ: φ(-z)= -φ(z)

  - $\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} = 0,5$

Để tìm được φ(Z) ta có thể tra bảng ở phục lục 1 bằng phương pháp tọa độ. Ví dụ để tìm φ(1,08) ta thực hiện như sau:

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

$\varphi(1,08) = 0,3599$

Chúng ta có thể sử dụng hàm NORMSDIST() trong Excel, tuy nhiên trong Excel là bảng 1, còn phụ lục của ta là bảng 0,5. Do đó để có kết quả chính xác ta sử dụng (NORMSDIST(Z) – 0,5).

3. Khái niệm Z_α

Z_α là một số sao cho $p(Z>Z_\alpha) = \alpha$. Đây chính là xác suất sai lầm mà ta thường dùng trong thống kê. Ta có thể tìm giá trị Z_α bằng hàm NORMSINV(1- α) trong Excel. Một vài giá trị đặc biệt:

α	0,005	0,010	0,025	0,050	0,100
Z_α	2,575	2,330	1,960	1,645	1,280

4. Một vài công thức xác suất thường dùng

Cách tính xác suất của biến ngẫu nhiên X có phân phối chuẩn tắc:

a) $p(X>a) = 0,5 - \varphi(a)$

b) $p(X<b) = 0,5 + \varphi(b)$

d) p(a<X<b) = ((b) - ((a), với a<b

e) p(X<a, X>b) = p(X<a)+p(X>b), với a<b, X - phân phối chuẩn

II. PHÂN PHỐI CỦA MỘT VÀI ĐẠI LƯỢNG THỐNG KÊ

1. Phân phối Chi bình phương

Nếu $x_1, x_2, ... x_n$ là đại lượng ngẫu nhiên được chọn từ tổng thể phân phối chuẩn thì $(n-1)\frac{S^2}{\sigma^2}$ có phân phối Chi bình phương bậc tự do (n-1).

Ký hiệu: $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$

Không giống như hàm phân phối chuẩn, hàm chi bình phương không đối xứng. Nó có các tính chất sau:

$E(\chi^2_{df}) = df$

$Var(\chi^2_{df}) = 2df$

Hàm phân phối χ^2 có đồ thị gần giống như đồ thị hàm phân phối chuẩn nhưng không đối xứng nhau qua cực trị. Tuy nhiên, nếu n lớn thì phân phối Chi bình phương sẽ sắp sĩ phân phối chuẩn. Phân phối χ^2 do Karl Pearson đưa ra vào năm 1900.

- *Khái niệm* $\chi^2_{n-1;\alpha}$**:** $\chi^2_{n-1;\alpha}$ là một số sao cho $p(\chi^2 > \chi^2_{n-1;\alpha}) = \alpha$. Định nghĩa này tương tự như đối với Z_α. Muốn tìm giá trị này ta có thể tra bảng ở phụ lục 3 cũng bằng phương pháp tọa độ hoặc sử dụng hàm CHIINV(α,df) trong Excel. Trong đó, df là bậc tự do.

PEARSON Karl

London 1857 – London 1936

Nhà toán học thống kê người Anh này đã từng học tại King's College ở Oxford. Năm 1884 ông được bổ nhiệm phụ trách bộ môn Toán ứng dụng và Cơ học tại University College ơ London. Năm 1901 ông thành lập phòng thí nghiệm về sinh trắc và cùng năm ấy với sự tài trợ của Galton, ông cho ra đời tạp chí Biometrika và ông đã điều hành tạp chí này cho đến khi mất. Năm 1911 ông phụ trách luôn bộ môn Eugénisme (Giáo sư Đào Duy Anh dịch là Nhân chủng cải lương học, Ưu sinh học) cho đến khi ông về hưu năm 1933. Kế tục công trình của Galton, ông là một trong những người sáng lập ngành Toán học thống kê hiện đại. Ông nghiên cứu lý thuyết tiến hoá theo mô hình thống kê toán học của ông. Ngoài những giây phúc miệt mài với khoa học, ông còn dành thời gian cho thi ca và hoạt động chính trị.

2. Phân phối Student

Nếu $x_1, x_2, \ldots x_n$ là đại lượng ngẫu nhiên được chọn từ tổng thể phân phối chuẩn thì $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$ có phân phối Student bậc tự do (n-1).

Ký hiệu: $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

- *Khái niệm* $t_{n-1;\alpha}$**:** $t_{n-1;\alpha}$ là một số sao cho $p(t > t_{n-1;\alpha}) = \alpha$. Định nghĩa này tương tự như đối với Z_α. Muốn tìm giá trị này ta có thể tra bảng ở phụ lục 2 cũng bằng phương pháp tọa độ hoặc sử dụng hàm TINV(2α,df) trong Excel.

Hàm phân phối t có đồ thị gần giống như đồ thị hàm phân phối chuẩn nhưng độ nhọn thấp hơn. Tuy nhiên, nếu n lớn thì phân phối phối Student sẽ sắp sĩ phân phối chuẩn.

GOSSET Wiliam Sealy

Canterbury 1876 - Beaconsfied (Ngoại ô luân Đôn) 1937

Thường thường người ta biết ông dưới dạng biệt hiệu Student. Ông là nhà thống kê người Anh. Thời trẻ, Ông học toán ở New college (Oxford). Năm 1899 ông vào làm việc xưởng sản xuất bia cho Guinness Brewery ở Dublin. Nhà máy muốn giảm giá thành sản xuất , nâng cao chất lượng lúa đại mạch và cây Hublon để sản xuất bia nên thành lập một phòng thí nghiệm nghiên cứu. Một bài toán đặt ra là làm sao từ những mẫu nghiên cứu ở phòng thí nghiệm rút ra được những kết luận xác đáng. Gosset tự nguyện tham gia nghiên cứu bài toán này: lý thuyết các mẫu nhỏ để từ đó rút ra kết luận. Gosset lao vào đề tài đặt ra, cùng Karl Pearson ở Luân Đôn miệt mài trong 2 năm 1906, 1907. Năm 1908 ông đưa ra một testn - sau này gọi là test Gosset - dùng để lựa chọn đại lúa mạch. Để bảo mật, nhà máy yêu cầu Gosset dấu tên thật, chỉ dùng biệt hiệu Pupil hay Student. Và như chúng ta biết, Gosset dùng hiệu thứ 2.

2.3. Phân phối Fisher (F)

Giả sử có hai mẫu độc lập có n_x, n_y quan sát lấy từ hai tổng thể có phân phối chuẩn, phương sai tổng thể và phương sai mẫu lần lượt là σ^2_x , σ^2_y , S^2_x , S^2_y thì khi đó $F = \dfrac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$ có phân phối Fisher bậc tự do của tử (n_x-1) và bậc tự do của mẫu (n_y-1).

Ký hiệu: $F \sim F_{v1,v2}$

Trong thực tế khi kiểm định sự bằng nhau của hai phương sai tổng thể và khi đó:

$$F = \frac{S_x^2}{S_y^2}$$

- *Khái niệm $F_{v1;v2;\alpha}$*: $F_{v1,v2;\alpha}$ là một số sao cho $p(F>F_{v1;v2;\alpha}) = \alpha$. Định nghĩa này tương tự như đối với Z_α. Muốn tìm giá trị này ta có thể tra bảng ở phụ lục 4 cũng bằng phương pháp tọa độ hoặc sử dụng hàm $FINV(\alpha,df_1,df_2)$ trong Excel.

FISHER Ra nald Aylmer

Londres 1890 - Adélaide 1962

Ông là một chuyên gia về Lý thuyết tiến hóa, nhà Di truyền học người Anh này đã có công phát triển các phương pháp thống kê so sánh những trung bình của hai mẫu nhờ đó xác định sự khác biệt của chúng có ý nghĩa hay không.

III. PHÂN PHỐI MẪU

1. Khái niệm

Mục đích của phân tích thống kê là sử dụng số liệu thu thập từ mẫu như trung bình và tỷ lệ mẫu để ước lượng giá trị thực của tổng thể. Quá trình khái quát quá kết quả nghiên cứu của mẫu cho tổng thể chung được gọi là suy luận thống kê.

Về lý thuyết, để có thể sử dụng các thông tin mẫu để suy luận các tham số của tổng thể chung, ta nên dựa vào kết quả nghiên cứu của nhiều mẫu nếu có thể. Nếu điều này được thực hiện, phân phối của các kết quả từ mẫu được gọi là phân phối mẫu. Nhưng về mặt thực hành việc ước lượng cho các tham số của tổng thể chỉ căn cứ vào kết quả cụ thể của một mẫu. Cho nên khái niệm về phân phối mẫu cần phải được xem xét để có thể ứng dụng lý thuyết về xác suất cho quá trình suy luận thống kê.

Phân phối mẫu có 3 trường hợp:

- Phân phối của trung bình mẫu để ước lượng trung bình của tổng thể;

- Phân phối của phương sai mẫu để ước lượng phương sai tổng thể;

- Phân phối của tỷ lệ mẫu để ước lượng tỷ lệ của tổng thể.

2. Định lý giới hạn trung tâm

Trong thực tế thường gặp là ta không biết về phân phối của tổng thể hoặc tổng thể không có phân phối chuẩn. Trong những trường hợp đó định lý giới hạn trung tâm giúp ta giải quyết vấn đề khi xem xét phân phối mẫu.

Định lý: Khi cỡ mẫu n đủ lớn thì phân phối của trung bình mẫu $\overline{X}$ sẽ xấp xỉ phân phối chuẩn, bất chấp hình dáng phân phối của tổng thể.

Định lý: Một biến ngẫu nhiên là tổng của của nhiều biến ngẫn nhiên khác sẽ có phân phối xấp xỉ phân phối chuẩn.

Điều này rất hữu ích trong kinh tế lượng bởi vì ta có thể coi sai số của một mô hình là tổng của nhiều tác động ngẫu nhiên.

3. Các tính chất của phân phối mẫu

- Nếu X có phân phối χ^2_m thì $n\overline{X}$ cũng có phân phối χ^2_{nm}

- Nếu X có phân phối chuẩn $N(\mu, \sigma^2)$ thì:

1). $n\overline{X}$ cũng có phân phối chuẩn $N(n\mu, n\sigma^2)$ và $\overline{X} \sim N(\mu, \sigma^2/n)$

2). Với kích thước mẫu khá lớn (n≥30), thì phân phối của trung bình mẫu sẽ xấp xỉ phân phối chuẩn và $z = \frac{\bar{x} - \mu}{\sigma_x / \sqrt{n}}$ có phân phối chuẩn tắc.

3). $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

4). $\overline{X}$ và S^2 độc lập với nhau.

CHƯƠNG IV
ƯỚC LƯỢNG KHOẢNG TIN CẬY
(Confidence Interval Estimation)

Chương này sẽ đề cập đến việc suy luận các đặc trưng của tổng thể dựa trên các đặc trưng của mẫu. Các đặc trưng của tổng thể có thể là giá trị trung bình, phương sai hoặc tỷ lệ các đơn vị tổng thể có một tính chất nào đó. Ví dụ, ta quan tâm đến thu nhập trung bình của lao động ở một ngành nghề nào đó. Vấn đề đặt ra là ước lượng các đặc trưng của tổng thể (chưa biết) từ các đặc trưng của mẫu dữ liệu thu thập được. Trong khái niệm ước lượng có 2 khái niệm là ước lượng điểm và ước lượng khoảng, tuy nhiên trong nội dung này chúng ta chỉ đi sâu vào vấn đề ược lượng khoảng bởi vì đây là một một nội dung rất quan trọng trong bài toán ước lượng mà chúng ta mong muốn giải quyết một vấn đề thực tế.

I. KHÁI NIỆM

Vấn đề ước lượng ở đây nói riêng và thống kê nói chung được chúng ta xem xét toàn diện theo quan điểm của xác suất, có nghĩa là ta xem xét khả năng xảy ra cao nhất chứ không xem xét trong từng trường hợp cụ thể đây là một vấn đề có thể không chính xác lắm nhưng nó cũng gây ra sự hiểu nhầm ít nhiều trong việc đánh giá tính hiệu quả của thống kê. Với quan điểm trên, các nhà toán học đưa ra khái niệm ước lượng khoảng như sau:

Gọi θ là đặc trưng của tổng thể cần ước lượng. Giả sử dựa vào mẫu quan sát, ta tìm được 2 biến ngẫu nhiên A, B sao cho:

$$\mathbf{P(A<\theta<B)=1-\alpha}$$

Trong đó, $(1-\alpha)$ là độ tin cậy

Giả sử a, b là giá trị cụ thể của A, B. Khoảng (a,b) được gọi là khoảng ước lượng với độ tin cậy $(1-\alpha)100\%$ của θ, hay nói ngắn gọn là khoảng tin cậy $(1-\alpha).100\%$ của θ.

II. ƯỚC LƯỢNG TRUNG BÌNH TỔNG THỂ

Để thực hiện ta chọn một mẫu ngẫu nhiên n quan sát $x_1, x_2,... x_n$ từ tổng thể X có trung bình là μ, phương sai σ^2, trung bình mẫu là $\bar{x}$, phương sai mẫu S^2, độ tin cậy $(1-\alpha)$. Ta ước lượng trung bình tổng thể như sau:

1. Khi đã biết phương sai σ^2

Điều kiện tổng thể có phân phối chuẩn hoặc có cỡ mẫu lớn ($n\geq30$)

$$\bar{x}-\frac{z_{\alpha/2}\sigma}{\sqrt{n}}<\mu<\bar{x}+\frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

với z có phân phối chuẩn tắc.

Từ công thức trên ta nhận thấy:

(1). Nếu độ tin cậy và độ lệch chuẩn cố định, kích thước mẫu càng lớn thì khoảng ước lượng càng hẹp, tức là độ chính xác của ước lượng càng cao.

(2). Nếu độ tin cậy và kích thước mẫu cố định, độ lệch chuẩn càng lớn thì khoảng ước lượng càng rộng, tức là độ chính xác của ước lượng càng thấp.

(3). Nếu độ lệch chuẩn và kích thước mẫu cố định, độ tin cậy càng cao thì khoảng khoảng ước lượng càng rộng, tức là độ chính xác của ước lượng càng thấp.

Ví dụ 4.1: Để xác định trọng lượng trung bình của các bao bột mì được đóng bằng máy tự động, người ta chọn ngẫu nhiên 15 bao và tính được trọng lượng trung bình 39,8kg. Tìm khoảng tin cậy 99% của trọng lượng trung bình các bao bột mì. Giả sử trọng lượng bao đường có phân phối chuẩn và phương sai là 0,144.

Gọi μ là trọng lượng trung bình một bao đường.

Ta có: n=15; $\bar{x}$ = 39,8, σ^2 = 0,144, α=1% => $Z_{0,5\%}$=2,575.

$$39,8 - 2,575\frac{\sqrt{0,144}}{\sqrt{15}} < \mu < 39,8 + 2,575\frac{\sqrt{0,144}}{\sqrt{15}}$$

$$39,55 < \mu < 40,05$$

Kết luận, với độ tin cậy 99%, trọng lượng trung bình của mỗi bao bột mì được ước lượng trong khoảng từ 39,55kg đến 40,05kg.

2. Khi chưa biết phương sai σ^2:

a) Trường hợp có cỡ mẫu lớn (n≥30):

Trong trường hợp này chúng ta áp dụng như trường hợp 1 nhưng chúng ta dùng phương sai của mẫu thay cho phương sai của tổng thể để ước lượng, chúng ta không cần điều kiện về phân phối chuẩn của tổng thể.

$$\bar{x} - \frac{z_{\alpha/2}S}{\sqrt{n}} < \mu < \bar{x} + \frac{z_{\alpha/2}S}{\sqrt{n}}$$

b) Trường hợp có cỡ mẫu nhỏ (n<30):

Trong trường hợp này ta sử dụng phân phối Student để ước lượng và điều kiện tổng thể phải có phân phối chuẩn.

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

Với t_{n-1} có phân phối Student với n-1 bậc tự do.

Ví dụ 4.2: Một công ty điện thoại muốn ước lượng thời gian trung bình của một cuộc điện thoại đường dài vào ngày cuối tuần. Mẫu ngẫu nhiên 20 cuộc gọi đường dài vào ngày cuối tuần cho thấy thời gian trung bình là 14,8 phút, độ lệch chuẩn 5,6 phút. Ước lượng thời gian trung bình của một cuộc gọi đường dài vào ngày cuối tuần, với độ tin cậy 95%.

Gọi μ là thời gian trung bình của một cuộc gọi đường dài vào ngày cuối tuần.

Ta có: n=20; $\bar{x}$=14,8; S=5,6; $t_{n-1,\alpha/2}=t_{19,\ 0,025}$= 2,093
Áp dụng công thức:

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

$$14,8 - 2,093\frac{5,6}{\sqrt{20}} < \mu < 14,8 - 2,093\frac{5,6}{\sqrt{20}}$$

$$12,1792 < \mu < 17,4208$$

Vậy, với độ tin cậy 95%, thời gian trung bình của một cuộc điện đàm đường dài vào cuối tuần được ước lượng trong khoảng từ 12,1792 đến 17,4208 phút

III. ƯỚC LƯỢNG TỶ LỆ TỔNG THỂ

Trong nhiều trường hợp ta có thể quan tâm đến tỷ lệ các đơn vị có một tính chất nào đó trong tổng thể. Chẳng hạn, tỷ lệ khách hàng sử dụng một loại sản phẩm nào đó hoặc tỷ lệ phế phẩm trong sản xuất,... Khi đó, ta thực hiện ước lượng cho tỷ lệ p của tổng thể.

Giả sử có mẫu ngẫn nhiên n quan sát và $\hat{p}$ là tỷ lệ các quan sát có tính chất A nào đó. Với mẫu lớn (n≥40), khoảng tin cậy (1-α)100% của tỷ lệ p các quan sát có tính chất A của tổng thể được xác định bởi:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ví dụ 4.3: Một nghiên cứu được thực hiện nhằm ước lượng thị phần của sản phẩm nội địa đối với mặt hàng bánh kẹo. Kết quả điều tra chọn mẫu ngẫu nhiên 100 khách hàng cho thấy có 34 người dùng sản phẩm nội địa. Tìm khoảng tin cậy 95% cho tỷ lệ khách hàng sử dụng bánh kẹo nội địa.

Gọi p là thị phần của sản phẩm nội địa đối với mặt hàng bánh kẹo.

Ta có: n=100, $\hat{p}$=0,34; $z_{\alpha/2}=z_{0,025}=1,96$

Áp dụng công thức ta có:

$$0,34 - 1,96\sqrt{\frac{0,34(1-0,34)}{100}} < p < 0,34 + 1,96\sqrt{\frac{0,34(1-0,34)}{100}}$$

$0,2472 < p < 0,4328$

Vậy, Khoảng tin cậy 95% cho tỷ lệ khách hàng sử dụng bánh kẹo nội địa là khoảng từ 24,72% đến 43,28%.

IV. ƯỚC LƯỢNG PHƯƠNG SAI TỔNG THỂ

Để xem xét độ đồng đều của dữ liệu hoặc chất lượng của sản phẩm, trong một số trường hợp ta có thể sử dụng ước lượng phương sai của tổng thể để xem xét. Để thực hiện bài toán này ta thực hiện như sau:

Chọn một mẫu ngẫu nhiên n quan sát có phân phối chuẩn, với độ tin cậy 1-α ta có ước lượng phương sai:

$$\frac{(n-1).S^2}{\chi^2_{n-1;\alpha/2}} < \sigma^2 < \frac{(n-1).S^2}{\chi^2_{n-1;1-\alpha/2}}$$

Ví dụ 4.4: Một nhà sản xuất quan tâm đến biến thiên của tỷ lệ tạp chất trong một loại hương liệu được cung cấp. Chọn ngẫu nhiên 15 mẫu hương liệu cho thấy độ lệch chuẩn về tỷ lệ tạp chất là 2,36%. Với khoảng tin cậy 95%, ước lượng độ lệch chuẩn về tỷ lệ tạp chất.

Gọi σ là độ lệch chuẩn về tỷ lệ tạp chất.

Ta có: n = 15, S=2,36%, α=5% => $\chi^2_{14;2,5\%} = 26,119$, $\chi^2_{14;97,5\%} = 5,629$

$$\frac{(15-1).2,36^2}{26,119} < \sigma^2 < \frac{(15-1).2,36^2}{5,629}$$

$2,99 < \sigma^2 < 13,85$ => $1,73 < \sigma < 3,72$

Vậy, với độ tin cậy 95%, độ lệch chuẩn về tuỷ lệ tạp chất được ước lượng trong khoảng từ 1,73 – 3,72%.

V. ƯỚC LƯỢNG CHÊNH LỆCH HAI TRUNG BÌNH TỔNG THỂ

Trong nhiều trường hợp ta có thể quan tâm đến sự khác biệt giữa trung bình 2 tổng thể. Chẳng hạn, khác biệt về doanh số trung bình trong tuần từ hai phương pháp trình bày hàng hóa, chào hàng khác nhau hoặc sự khác biệt giữa năng suất cây trồng do sử dụng hai loại phân bón khác nhau,... Phương pháp so sánh trung bình hai tổng thể phụ thuộc vào cách thức lấy mẫu: mẫu phối hợp từng cặp (mẫu phụ thuộc) hoặc mẫu độc lập.

1. Ước lượng khoảng tin cậy dự trên sự phối hợp từng cặp (Matched pair)

Ở phương pháp này, các đơn vị mẫu được chọn từng cặp từ hai tổng thể. Thông thường mẫu phối hợp từng cặp bao gồm các trường hợp sau đây:

- *So sánh giữa "trước" và "sau*", chẳng hạn như mẫu thứ nhất: doanh số bán trước khi thực hiện chiến dịch khuyến mãi; mẫu thứ hai: doanh số bán sau khi thực hiện chiến dịch khuyến mãi. Ở đây, mẫu phối hợp từng cặp theo nghĩa là từng cặp doanh số thu thập ở cùng một cửa hàng.

- *So sánh giữa các đơn vị về một đặc điểm nào đó*, chẳng hạn mẫu thứ nhất: tiền lương của nhân viên nam ở công ty A; mẫu thứ hai: tiền lương của nhân viên nữ ở công ty A.

- *So sánh giữa các đơn vị phối hợp từng cặp theo không gian*, chẳng hạn mẫu thứ nhất: doanh số bán của một loại nước giải khát A ở cửa hàng X; mẫu thứ hai: doanh số bán của một loại nước giải khát B ở cửa hàng X. Ở đây mẫu phối hợp từng cặp theo nghĩa cả hai doanh số của hai nhãn hiệu A, B được thu thập trên cùng một cửa hàng.

- *So sánh giữa các đơn vị phối hợp từng cặp theo thời gian*, chẳng hạn mẫu thứ nhất: doanh số bán của nhà hàng X vào tháng 5/2001; mẫu thứ hai: doanh số bán của nhà hàng Y vào tháng 5/2007. Ở đây mẫu phối hợp từng cặp trong cùng một thời gian.

**** Khoảng tin cậy cho sự khác biệt giữa trung bình hai tổng thể:***

Giả sử ta có mẫu n cặp quan sát từ hai tổng thể X và Y. Gọi μ_x , μ_y là trung bình của X, Y; $\bar{d}$ và S_d là trung bình và độ lệch chuẩn của n sự khác biệt (x_i-y_i). Giả sử rằng phân phối của những khác biệt này là phân phối chuẩn, khoảng tin cậy $(1-\alpha)100\%$ của (μ_x , μ_y) được xác định như sau:

$$\bar{d} - t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_x - \mu_y < \bar{d} + t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}$$

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

$$S_d^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} = \frac{\sum_{i=1}^{n} d_i^2 - n.\bar{d}^2}{n-1}$$

với t_{n-1} có phân phối Student với n-1 bậc tự do.

Ví dụ 4.5: Công ty điện lực thực hiện các biện pháp khuyến khích tiết kiệm điện. Lượng điện tiêu thụ ghi nhận ở 12 hộ gia đình trước và sau khi có biện pháp khuyến khích tiết kiệm điện như sau:

Hộ gia đình	Lượng điện tiêu thụ trước và sau khi khuyến khích tiết kiệm (Kwh)	
	Trước	Sau
1	73	69
2	50	54

Hộ gia đình	Lượng điện tiêu thụ trước và sau khi khuyến khích tiết kiệm (Kwh)	
	Trước	Sau
3	83	82
4	78	67
5	56	60
6	74	73
7	74	75
8	87	78
9	69	64
10	72	72
11	77	70
12	75	63

Với độ tin cậy 95%, hãy ước lượng lượng điện tiêu thụ trung bình của hộ gia đình trước và sau khi thực hiện biện pháp khuyến khích tiết kiệm điện.

Gọi μ_x, μ_y là lượng điện tiêu thụ trung bình của hộ gia đình trước và sau khi thực hiện biện pháp khuyên khích tiết kiệm điện.

Ta tính các tham số cần thiết:

Hộ gia đình	Lượng điện tiêu thụ trước và sau khi khuyến khích tiết kiệm (Kwh)		$d_i=(x_i-y_i)$	$(d_i-\bar{d})^2$
	Trước (x_i)	Sau (y_i)		
1	73	69	4	0,3402
2	50	54	-4	55,0074
3	83	82	1	5,8404
4	78	67	11	57,5064
5	56	60	-4	55,0074
6	74	73	1	5,8404
7	74	75	-1	19,5072
8	87	78	9	31,1732
9	69	64	5	2,5068
10	72	72	0	11,6738
11	77	70	7	12,8400
12	75	63	12	73,6730
Tổng			41	330,9167

Ta có:

$$\bar{d} = \frac{41}{12} = 3{,}4167$$

$$s_d = \frac{\sum[d_i - \bar{d}]^2}{n-1} = \frac{330{,}9167}{12-1} = 5{,}4848$$

$t_{n-1,\alpha/2}=t_{11,0,025}=2,201$

Giả sử rằng các khác biệt giữa lượng điện trước và sau khi khuyến khích tiết kiệm có độ lệch chuẩn. Áp dụng công thức ta có:

$$\bar{d} - t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_x - \mu_y < \bar{d} + t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}$$

$$3,4167 - 2,201\frac{5,4848}{\sqrt{12}} < \mu_X - \mu_Y < 3,4167 + 2,201\frac{5,4848}{\sqrt{12}}$$

$-0,0682 < \mu_X-\mu_Y < 6,9016$

Vậy, khoảng tin cậy 95% của sự khác biệt giữa lượng điện tiêu thụ trước và sau khi khuyến khích tiết kiệm được ước lượng từ -0,0682 đến 6,9016 (Kwh).

2. Ước lượng khoảng tin cậy dựa vào mẫu độc lập

Gọi n_x, n_y là số quan sát của các mẫu ngẫu nhiên độc lập $x_1, x_2, ... x_{nx}, y_1, y_2, ... y_{ny}$ từ hai tổng thể X và Y có trung bình μ_x, μ_y và phương sai σ^2_x, σ^2_y. Với trung bình mẫu $\bar{x}$, $\bar{y}$, phương sai mẫu là S^2_x, S^2_y, với độ tin cậy $(1-\alpha)$ ta ước lượng như sau:

2.1. Biết phương sai hai tổng thể:

Điều kiện hai tổng thể phải có phân phối chuẩn hoặc chọn cỡ mẫu lớn (n_x, $n_y \geq 30$)

$$\mu_x - \mu_y \in \{(\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\}$$

2.2. Chưa biết phương sai tổng thể:

a) Có phương sai khác nhau:

* Nếu n_x, $n_y \geq 30$: ta thay phương sai tổng thể bằng phương sai mẫu.

$$\mu_x - \mu_y \in \{(\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}\}$$

* Nếu n_x hoặc $n_y < 30$: Giả sử X, Y có phân phối chuẩn:

$$\mu_x - \mu_y \in \{(\bar{x} - \bar{y}) \pm t_{n;\alpha/2}\sqrt{(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y})}\}$$

với Bậc tự do $n = \dfrac{(s_x^2/n_x + s_y^2/n_y)^2}{\dfrac{(s_x^2/n_x)^2}{n_x - 1} + \dfrac{(s_y^2/n_y)^2}{n_y - 1}}$

b) Có phương sai bằng nhau:

$$\mu_x - \mu_y \in ((\bar{x} - \bar{y}) \pm t_{nx+ny-2;\alpha/2}\sqrt{S^2(\frac{1}{n_x} + \frac{1}{n_y})})$$

Trong đó: $S^2 = \dfrac{(n_x - 1).S_x^2 + (n_y - 1).S_y^2}{(n_x + n_y - 2)}$

Ví dụ 4.6: Một công ty đang muốn xem xét thời gian sản xuất của hai dây chuyên sản xuất mới và cũ.

Ở dây chuyền mới: 40 sản phẩm được được sản xuất với thời gian trung bình 46,5 phút/ sản phẩm, độ lệch chuẩn là 8 phút.

Ở dây chuyền cũ: 38 sản phẩm được sản xuất với thời gian trung bình là 51,2 phút/sản phẩm, độ lệch chuẩn là 9,5 phút.

Hãy ước lượng khỏang tin cậy 95% cho sự khách biệt về thời gian sản xuất giữa hai dây chuyền.

Gọi μ_x, μ_y là thời gian sản xuất trung bình một sản phẩm của dây chuyền cũ và dây chuyên mới.

Ta có: Ở dây chuyền mới: $\bar{x}=46,5$ $s_X=8$ $n_X=40$

Ở dây chuyền cũ: $\bar{y}=51,2$ $s_Y=9,5$ $n_Y=38$

$$z_{\alpha/2}=z_{0,025}=1,96$$

Giả sử ta có 2 phương sai khác nhau, áp dụng công thức ta có:

$$(\bar{x}-\bar{y})-z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}<\mu_x-\mu_y<(\bar{x}-\bar{y})+z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$$

$$(46,5-51,2)-1,96.\sqrt{\frac{8^2}{40}+\frac{9,5^2}{38}}<\mu_x-\mu_y<(46,5-51,2)-1,96.\sqrt{\frac{8^2}{40}+\frac{9,5^2}{38}}$$

$-8,6077 < \mu_X-\mu_Y < -0,7923$

Vậy, Với độ tin cậy 95%, ta ước lượng dây chuyền sản xuất mới rút ngắn thời gian trung bình sản xuất một sản phẩm trong khoảng từ 0,7923 đến 8,6077 phút.

VI. ƯỚC LƯỢNG HAI CHÊNH LỆCH TỶ LỆ TỔNG THỂ (trường hợp n≥40)

Giả sử ta có hai mẫu n_x và n_y được chọn ngẫu nhiên độc lập từ hai tổng thể X và Y có tỷ lệ mẫu là $\hat{p}_X$, $\hat{p}_Y$. Ta có khoảng tin cậy $(1-\alpha)100\%$ của sự khác biệt $(p_x - p_y)$ được xác định như sau:

$$(\hat{p}_x-\hat{p}_y)\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

Ví dụ 4.7: Kết quả điều tra từ mẫu ngẫu nhiên 1000 người ở một thành phố cho thấy năm 2007, tỷ lệ thất nghiệp ở thành phố X là 7,5%, ở thành phố Y là 7,2%. Hãy ước lượng khoảng tin cậy 99% cho sự khác biệt về tỷ lệ thất nghiệp giữa hai thành phố.

Áp dụng công thức ta có:

$$(\hat{p}_x-\hat{p}_y)\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

$$(0,075-0,072)\pm 2,575\sqrt{\frac{0,075(1-0,075)}{1000}+\frac{0,072(1-0,072)}{1000}}$$

$-0,027 < p_X - p_Y < 0,033$

Vậy, với độ tin cậy 99%, có thể nói rằng tỷ lệ thất nghiệp ở thành phối X từ thấp hơn 2,7% đến cao hơn 3,3% so với thành phối Y.

VII. ƯỚC LƯỢNG CỠ MẪU (Estimating the sample size)

Để tăng độ chính xác của ước lượng, theo như chúng ta đã nghiên cứu những phương pháp để ước lượng khoảng tin cậy, thì chỉ có một hướng để đạt được đó là cần xác định cỡ mẫu có kích thước tối thiểu.

1. Cỡ mẫu trong ước lượng khoảng tin cậy của trung bình tổng thể

Xuất phát từ công thức: $\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$

Giả sử một mẫu ngẫu nhiên gồm n quan sát từ một tổng thể có phương sai ta có. Một khoảng tin cậy (1-α)100% của trung bình tổng thể là μ, gọi Δ là một nửa chiều rộng của khoảng tin cậy $\Delta = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Từ đây ta dễ dàng suy ra:

$$n = z_{\alpha/2}^2 \frac{\sigma^2}{\Delta^2}$$

2. Cỡ mẫu trong ước lượng khoảng tin cậy của tỷ lệ tổng thể

Xuất phát từ công thức: $\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Giả sử rằng một mẫu ngẫu nhiên gồm n quan sát, một khoảng tin cậy (1-α)100% cho tỷ lệ tổng thể p cho bởi công thức trên. Gọi Δ là một nửa chiều rộng của khoảng tin cậy $\Delta = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, từ đó ta xác định kích thước mẫu tối thiểu như sau:

$$n = z_{\alpha/2}^2 \frac{\hat{p}(1-\hat{p})}{\Delta^2}$$

Tất nhiên ta chưa thể biết $\hat{p}$, song $\hat{p}(1-\hat{p})$ không vượt quá 0,25. Do đó, ta có công thức:

$$n = z_{\alpha/2}^2 \frac{0{,}25}{\Delta^2}$$

CHƯƠNG V
KIỂM ĐỊNH GIẢ THUYẾT

I. MỘT SỐ KHÁI NIỆM:

Bên cạnh việc ước lượng các đặc trưng của tổng thể - đã nói ở chương trước - các dữ liệu thu thập từ mẫu còn có thể dùng để đánh giá xem một giả thuyết nào đó về tổng thể là đúng hay sai, gọi là kiểm định giả thuyết. Nói cách khác, kiểm định giả thuyết là dựa vào các thông tin mẫu để đưa ra kết luận bác bỏ hay chấp nhận về các giả thuyết đó của tổng thể. Tuy nhiên, chúng ta phải hiểu chấp nhận hay bác bỏ giả thuyết phải hiểu theo nghĩa xác suất.

1. Các loại giả thuyết trong thống kê

Giả thuyết H_0 (null hypothesis):

Gọi θ là một đặc trưng chưa biết của tổng thể (giá trị trung bình, phương sai, tỷ lệ). Ta hình thành giả thuyết H_0 về θ so với giá trị θ_0 cụ thể nào đó.

Giả thuyết H_1 (alternative hypothesis):

Giả thuyết 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. Trong thống kê H_0 được kiểm định dựa trên cơ sở "đối chứng" H_1, H_1 là giả thuyết thể hiện các tình huống không nằm trong H_0.

- Kiểm định dạng hai đuôi (two-tailed):

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

- Kiểm định dạng một đuôi (one-tailed):

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$ Một đuôi phải

$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$ Một đuôi trái

2. Các loại sai lầm trong kiểm định giả thuyết

(1). Sai lầm loại 1: Giả thuyết H_0 đúng nhưng qua kiểm định ta lại kết luận giả thuyết sai, và do vậy bác bỏ giả thuyết H_0 ở mức ý nghĩa α nào đó. Có nghĩa là ta bác bỏ giả thuyết đúng.

(2). Sai lầm loại 2: Giả thuyết H_0 sai nhưng qua kiểm định, ta lại kết luận giả thuyết đúng, và do vậy chấp nhận giả thuyết H_0 ở mức ý nghĩa α nào đó. Có nghĩa là ta chấp nhận một giả thuyết sai.

Tùy theo qua điểm và tính chất mà người ta cho sai lầm loại 1 hoặc loại 2 là nghiêm trọng hơn. Tuy nhiên, thông thường thì người ta sẽ cho rằng sai lầm loại 1 là nghiêm trọng hơn mà thống kê cần tránh.

3. Qui trình tổng quát trong kiểm định giả thuyết

Mục đích chúng ta trong việc giới thiệu về phương thức kiểm định giả thuyết tổng quát là nhằm đi tìm hiểu nội dung cơ bản trong phép suy diễn thống kê. Để đạt đến quyết định cuối cùng, một phương thức có hệ thống sẽ được lập ra với các bước như sau:

Bước 1: Xây dựng giả thuyết. Ta bắt đầu kiểm định giả thuyết với một giả định về một vài tham số tổng thể và sử dụng dữ liệu mẫu để kiểm tra tính logic của giả định đó. Nói cách khác, ta bắt đầu bằng cách giả định có một giá trị tổng thể nào đó và kết luận bằng quyết định bác bỏ hay không bác bỏ giả định đó theo các số liệu mẫu. Việc thiết lập giả thuyết tùy thuộc vào bản chất của tình huống có định hướng hay không định hướng. Nếu tình huống không có định hướng sai biệt, thì giả thuyết là 2 đuôi, nếu có tính định hướng thì ta có giả thuyết 1 đuôi. Để xác định giả thuyết 1 đuôi phải hay 1 đuôi trái chúng ta dựa vào nguyên tắc tránh sai lầm loại 1.

Tuy nhiên cần nhớ rằng sự thất bại trong việc loại H_0 không đồng nghĩa với việc bạn đã chứng minh được H_0 đúng, mà chỉ là bạn không đủ bằng chứng thống kê để loại bỏ mà thôi.

Bước 2: Chọn mức ý nghĩa mong muốn. Khả năng phạm sai lầm loại 1 như ta đã trình bày được gọi là mức ý nghĩa và được ký hiệu là α. Trên thực tế, có 3 mức ý nghĩa thường dùng nhất là 0,1, 0,05 và 0,01 tương ứng với độ tin cậy là 0,90, 0,95, 0,99. Việc lựa chọn α là bao nhiêu phụ thuộc vào tính chủ quan của người nghiên cứu chấp nhận rủi ro ở mức nào. Có một ý có tính chất kinh nghiệm để chúng ta tham khảo:

- Nếu nội dung nghiên cứu đòi hỏi độ chính xác cao thì nên chọn mức α nhỏ, thông thường là 1%.

- Nếu nội dung nghiên cứu số liệu biến động lớn, thu thập thông tin khó chính xác thì ta nên chọn α lớn, tuy nhiên ta không nên tăng α quá lớn sẽ làm tăng khả năng bị sai lầm loại 2 và thông thường theo sự thống nhất chung của các nhà thống kê mức ý nghĩa tối đa là 10%.

- Nếu không quan tâm quá nhiều đến mức ý nghĩa thì ta nên chọn theo mức thông thường là 5%.

Bước 3: Tính trị số thông kê hay giá trị thực tế của kiểm định. Trong bước này, dựa vào các lý thuyết thống kê mà chúng ta lựa chọn công thức phù hợp để qui phân phối mẫu về phân phối nào đó. Một số phân phối thường gặp là phân phối chuẩn, phân phối Student, phân phối Chi bình phương, phân phối Fisher,... Giá trị thực tế của kiểm định là cơ sở để quyết định chấp nhận hay bác bỏ giả thuyết không.

Bước 4: Rút ra kết luận liên quan đến giả thuyết không. Tương ứng với mức ý nghĩa α và phân phối được xác định ở bước 3 ta tìm được giá trị lý thuyết của kiểm định, thông thường là ta tra bảng hoặc sử dụng phần mềm máy tính để tìm được (Z_α, $t_{df,\alpha}$, $F_{v1,v2,\alpha}$,....). Sau đó chúng ta so sánh giữa giá trị thực tế và giá trị lý thuyết của kiểm định để có kết luận phù hợp với giả thuyết không.

Bước 5: Kết luận.Tùy thuộc vào nội dung nghiên cứu chúng ta sẽ đưa ra kết luận phù hợp với mục đích và yêu cầu của vấn đề đặt ra.

II. KIỂM ĐỊNH THAM SỐ

1. Kiểm định trung bình tổng thể

Để thực hiện kiểm định này giả sử ta có một mẫu ngẫu nhiên n quan sát x_1, x_2,... x_n từ tổng thể X có trung bình là μ, phương sai σ^2, trung bình mẫu là $\bar{x}$, phương sai mẫu S^2, mức ý nghĩa kiểm định α và giá trị cho trước là μ_0.

Kiểm định giả thuyết được thực hiện như sau:

a) Trường hợp đã biết phương sai tổng thể:

Điều kiện áp dụng cho trường hợp này là tổng thể X có phân phối chuẩn hoặc có cỡ mẫu n≥30, các bước kiểm định như sau:

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**

1. Đặt giả thuyết	$\begin{cases} H_0 : \mu \le \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu \ge \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \ne \mu_0 \end{cases}$
2. Giá trị kiểm định	$z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$		
3. Quyết định bác bỏ H_0 khi:	$Z > Z_\alpha$	$Z < -Z_\alpha$	$Z > Z_{\alpha/2}$; $Z < -Z_{\alpha/2}$

Đối với trường hợp này ta sử dụng phân phối chuẩn để thực hiện kiểm định giả thuyết thống kê, ta có thể gọi tắt là kiểm định Z và nguyên tắc này đều giống nhau chỉ có giá trị kiểm định là khác nhau. Để khỏi phải nhớ nhiều ta có nguyên tắc bác bỏ H_0 trong kiểm định Z như sau :

Bác bỏ giả thuyết H_0 1 đuôi nếu : $|Z| > Z_\alpha$

Bác bỏ giả thuyết H_0 2 đuôi nếu : $|Z| > Z_{\alpha/2}$

Ví dụ 5.1: Một nghiên cứu được thực hiện để xác định mức độ hài lòng của khách hàng sau khi công ty điện thoại, cải tiến một số dịch vụ khách hàng. Trước khi thay đổi, mức độ hài lòng của khách hàng tính trung bình là 77 điểm, theo thang điểm từ 0 đến 100. 20 khách hàng được chọn ngẫu nhiên để gởi bảng điều tra xin ý kiến sau khi các thay đổi được thực hiện, mức độ hài lòng trung bình tính được là 80. Có thể kết luận khách hàng đã được làm hài lòng ở mức độ cao hơn hay không với mức ý nghĩa 1%. Cho biết tổng thể có phân phối chuẩn với độ lệch chuẩn là 8.

Gọi μ là điểm trung bình về mức độ hài lòng của khách hàng.

Ta có. n=20, $\bar{x}$=80, μ_0=77, σ=8, α=1%.

(1). Đặt giả thuyết: $\begin{cases} H_0 : \mu \le 77 \\ H_1 : \mu > 77 \end{cases}$

(2). Giá trị kiểm định: $Z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{80 - 77}{20 / \sqrt{20}} = 1{,}68$

(3). Quyết định: $|Z| = 1{,}68 < Z_{1\%} = 2{,}33$: Chấp nhận giả thuyết H_0

(4). Kết luận: Với mức ý nghĩa α=1% số liệu của mẫu không đủ bằng chứng để bác bỏ giả thuyết H_0, nghĩa là mức độ hài lòng của khác hàng không tăng lên.

a) Trường hợp chưa biết phương sai tổng thể:

Trường hợp n ≥ 30:

Trường hợp này ta không có phương sai tổng thể thì sử dụng phương sai của mẫu để tính giá trị thực tế của kiểm định. Ngoài ra, dựa vào định lý giới hạn trung tâm trong trường hợp này ta không nhất thiết tổng thể phải có phân phối chuẩn.

Ta có kiểm định Z, với $Z = \dfrac{\bar{x} - \mu_0}{S_x / \sqrt{n}}$

Trường hợp có n<30:

Trong trường hợp này điều kiện tổng thể phải có phân phối chuẩn. Ta có dạng tổng quát kiểm định như sau:

	Một đuôi phải	Một đuôi trái	Hai đuôi
1. Đặt giả thuyết	$\begin{cases} H_0 : \mu \le \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu \ge \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \ne \mu_0 \end{cases}$
2. Giá trị kiểm định	$t = \dfrac{\bar{x} - \mu_0}{S_x / \sqrt{n}}$		
3. Quyết định bác bỏ H_0 khi:	$t > t_{n-1,\alpha}$	$t < - t_{n-1,\alpha}$	$t > t_{n-1,\alpha/2}$; $t < -t_{n-1,\alpha/2}$

Tương tự như đối với kiểm định Z, ta có thể gọi trong trường hợp này là kiểm định Student và ta có thể tóm tắt nguyên tắc bác bỏ H_0 như sau:

Bác bỏ giả thuyết H_0 1 đuôi nếu : $|t| > t_{n-1,\alpha}$

Bác bỏ giả thuyết H_0 2 đuôi nếu : $|t| > t_{n-1,\alpha/2}$

Ví dụ 5.2: Một loại đèn chiếu được nhà sản xuất cho biết có tuổi thọ trung bình thấp nhất là 65 giờ. Kết quả kiểm tra từ mẫu ngẫn nhiên 21 đèn cho thấy tuổi thọ trung bình là 62,5 giờ, với độ lệch chuẩn là 3. Với α= 0,01, có thể kết luận gì về lời tuyên bố của nhà sản xuất? Cho biết tuổi thọ của bóng đèn có phân phối chuẩn.

Gọi μ là tuổi thọ trung bình của bóng đèn.

Ta có: n=21, $\bar{x}$=62,5, μ=65, S=3, α=1%.

(1). Đặt giả thuyết: $\begin{cases} H_0 : \mu \ge 65 \\ H_1 : \mu < 65 \end{cases}$

(2). Giá trị kiểm định: $t = \dfrac{\bar{x} - \mu_0}{S_x / \sqrt{n}} = \dfrac{62,5 - 65}{3/\sqrt{21}} = -3,82$

(3). Quyết định: $|t| = 3,82 > 2,528 = t_{20,\,0,01}$ => Giả thuyết H_0 bị bác bỏ.

(4). Kết luận: Với α=1%, ta có thể kết luận rằng lời tuyên bố của nhà sản xuất là sai.

2. Kiểm định tỷ lệ p tổng thể

Giả sử ta có mẫu ngẫn nhiên n quan sát. Gọi p, $\hat{p}$ lần lượt là tỷ lệ các đơn vị có tính chất nào đó mà ta quan tâm của tổng thể và của mẫu, p_0 là số tỷ lệ cho trước. Điều kiện cỡ mẫu $n \ge 40$.

Trong trường hợp này ta có kiểm định phân phối chuẩn:

	Một đuôi phải	Một đuôi trái	Hai đuôi
1. Đặt giả thuyết	$\begin{cases} H_0 : p \le p_0 \\ H_1 : p > p_0 \end{cases}$	$\begin{cases} H_0 : p \ge p_0 \\ H_1 : p < p_0 \end{cases}$	$\begin{cases} H_0 : p = p_0 \\ H_1 : p \ne p_0 \end{cases}$
2. Giá trị kiểm định	$Z = \dfrac{\hat{p}_x - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$		
3. Quyết định bác bỏ H0 khi:	$Z > Z_\alpha$	$Z < -Z_\alpha$	$Z > Z_{\alpha/2}$; $Z < -Z_{\alpha/2}$

Ví dụ 5.3: Giả sử sản phẩm của một công ty sản xuất vỏ xe ô tô đã chiếm được thị phần 42%. Hiện tại, trước mắt cạnh tranh của đối thủ và những điều kiện thay đổi của môi trường kinh doanh, ban lãnh đạo muốn kiểm tra lại xem thị phần của công ty có còn là

42% hay không. Chọn một mẫu ngẫu nhiên 550 ô tô trên đường, kết quả cho thấy có 219 xe sử dụng vỏ xe của công ty. Có thể kết luận gì, ở mức ý nghĩa α=0,1?

Gọi p là tỷ lệ xe sử dụng vỏ xe của công ty.

Ta có, n=550, $\hat{p}$ =219/550 = 0,398, p_0=0,42, α=10%.

(1). Đặt giả thuyết: $\begin{cases} H_0 : p \geq 0{,}42 \\ H_1 : p < 0{,}42 \end{cases}$

(2). Giá trị kiểm định: $Z = \dfrac{\hat{p}_x - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \dfrac{219/550 - 0{,}42}{\sqrt{0{,}42(1-0{,}42)/550}} = -1{,}037$

(3). Quyết định: $|Z| = 1{,}037 < 1{,}28 = Z_{0,1}$. Chấp nhận giả thuyết H_0.

(4). Kết luận: Ở mức ý nghĩa 10%, ta có thể nói rằng hiện tại công ty chiếm ít nhất 42% thị trường về vỏ xe ô tô.

3. Kiểm định phương sai

Chọn một mẫu ngẫu nhiên n quan sát được chọn ngẫu nhiên từ tổng thể phân phối chuẩn. Gọi S^2 là phương sai của mẫu, kiểm định giả thuyết về phương sai của tổng thể được thực hiện như sau:

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**
1. Đặt giả thuyết	$\begin{cases} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases}$	$\begin{cases} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{cases}$	$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$
2. Giá trị kiểm định	$\chi^2_{n-1} = \dfrac{(n-1).S^2}{\sigma_0^2}$		
3. Quyết định bác bỏ H_0 khi:	$\chi^2_{n-1} > \chi^2_{n-1,\alpha}$	$\chi^2_{n-1} < \chi^2_{n-1,1-\alpha}$	$\chi^2_{n-1} > \chi^2_{n-1,\alpha/2}$; $\chi^2_{n-1} < \chi^2_{n-1,1-\alpha/2}$

Ví dụ 5.4: Bộ phận giám sát chất lượng quan tâm đến đường kính một loại chi tiết sản phẩm. Quá trình sản xuất còn được xem là tốt và chi tiết sản phẩm sản xuất ra được chấp nhận nếu phương sai của đường kính tối đa không quá 1, nếu phương sai vượt quá 1, phải xem xét lại máy móc và sửa chữa. Với mẫu ngẫu nhiên 31 chiết tiết, phương sai đường kính tính được là 1,62. Ở mức ý nghĩa α=0,05, ta có thể kết luận như thế nào về quá trình sản xuất?

Gọi σ^2 là phương sai của đường kính sản phẩm.

Ta có, n=31, S^2=1,62, σ^2_0=1, α=5%

(1). Đặt giả thuyết: $\begin{cases} H_0 : \sigma^2 \leq 1 \\ H_1 : \sigma^2 > 1 \end{cases}$

(2). Giá trị kiểm định: $\chi^2_{30} = \dfrac{(n-1).S_x^2}{\sigma_0^2} = \dfrac{(31-1).1{,}62}{1} = 48{,}6$

(3). Quyết định: $\chi^2_{30} = 48{,}6 > 43{,}77 = \chi^2_{30,0,05}$, bác bỏ giả thuyết H_0.

(4). Kết luận: Ở mức ý nghĩa 5%, ta cần phải xem xét, sửa chửa máy móc.

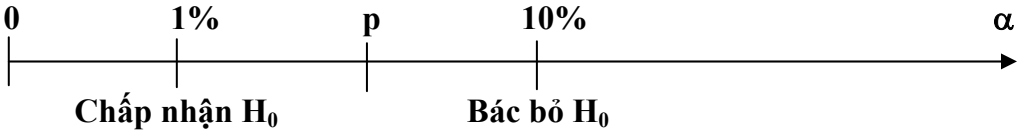
4. Giá trị p của kiểm định (probability value p-value)

Trong ví dụ 5.1, ta có nhận xét như sau:

- Với α=1% : $|Z|$=1,677 < $Z_{1\%}$=2,33 => Chấp nhận H_0

- Với α=10% : $|Z|$=1,677 < $Z_{10\%}$=1,28 => Bác bỏ H_0

Như vậy với mỗi mức ý nghĩa khác nhau chúng ta có thể kết luận khác nhau và theo khuynh hướng nếu mức ý nghĩa càng tăng thì khả năng bác bỏ giả thuyết H_0 càng tăng, từ đó xuất hiện giá trị p là giá trị trung gian giữa 2 miền của α thành miền chấp nhận và miền bác bỏ H_0.



Định nghĩa: Giá trị p của kiểm định là một số sao cho với mọi α>p thì giả thuyết H_0 bị bác bỏ.

Giá trị p đóng vai trò rất quan trọng trong kiểm định bởi vì nó tiện dụng hơn khi quyết định về giả thuyết và cho chúng ta được giá trị tới hạn mà giả thuyết còn có thể chấp nhận được.

Về mặt tính toán thì chúng ta không thể tính toán bằng phương pháp thủ công được mà hiện nay các máy tính xử lý thống kê đều cho chúng ta kết quả này một cách dễ dàng. Chúng ta có thể tìm giá trị p trong trường hợp kiểm định Z bằng cách tra bảng:

- Trường hợp kiểm định 1 đuôi: $p = 0,5 - \varphi(|Z|)$

- Trường hợp kiểm định 2 đuôi: $p = 2(0,5 - \varphi(|Z|))$

Có hai nhận xét quan trọng đối với giá trị p:

- Nếu p quá nhỏ ($p \approx 0$): Giả thuyết H_0 sẽ bị bác bỏ hoàn toàn

- Nếu p quá lớn (p>10%): Giả thuyết H_0 sẽ được chấp nhận hoàn toàn.

Trong trường hợp p quá lớn hoặc quá nhỏ chúng ra kết luận kiểm định thống kê có thể không cần đề cập đến mức ý nghĩa.

5. Kiểm định sự khác nhau của 2 phương sai tổng thể

Chọn 2 mẫu ngẫu nhiên độc lập có n_X, n_Y quan sát từ 2 tổng thể X,Y có phân phối chuẩn. Giả sử $S_x^2 > S_Y^2$, ta có giả thuyết:

Giả thuyết: $\begin{cases} H_0 : \sigma_x^2 = \sigma_y^2 \\ H_1 : \sigma_x^2 > \sigma_y^2 \end{cases}$

Bác bỏ GT H_0: $F = \dfrac{S_x^2}{S_x^2} > F_{n_x-1;n_x-1;\alpha}$

Việc giả sử $S_x^2 > S_Y^2$ điều này không làm mất tính tổng quát của bài toán, khi đó ta sẽ chọn X là tổng thể có phương sai lớn.

Ví dụ 5.5: Một công ty chuyên cung cấp dịch vụ điện thoại di động muốn khảo sát có sự khác biệt trong biến động hóa đơn điện thoại trung bình hàng tháng của khách hàng là nhà kinh doanh nam hay nữ hay không. Họ tiến hành thu thập một mẫu ngẫu nhiên 20 khách hàng nam và một mẫu ngẫu nhiên 10 khách hàng nữ. Tính toán các tham số độ lệch chuẩn mẫu như sau: Khách hàng nam 146.000đ; Khách hàng nữ 164.000đ. Có kết luận gì với mức ý nghĩa 5%.

Gọi σ^2_x , σ^2_y là phương sai về biến động chi phí điện thoại của nữ, nam.

Ta có: n_x=20, n_y=10, S_x=164.000đ, S_y=146.000đ, α=5%.

(1). Giả thuyết: $\begin{cases} H_0 : \sigma_x^2 = \sigma_y^2 \\ H_1 : \sigma_x^2 > \sigma_y^2 \end{cases}$

(2). Giá trị kiểm định: $F = \frac{164.000^2}{146.000^2} = 1{,}26$

(3). Quyết định: $F = 1{,}26 < F_{9,19,5\%} = 2{,}42$ => Chấp nhận H_0

(4). Kết luận: Với α=5%, không đủ bằng chứng để chứng minh rằng chi tiêu điện thoại của nam và nữ là khác nhau.

- Thực hiện trên Excel để xử lý: Các bước thực hiện như sau:

(1). Tools → Data Analysis → F-Test Two-Sample for Variances

(2). Cách nhập liệu cụ thể tham khảo mục 2.6.1.

6. Kiểm định sự khác nhau của hai trung bình tổng thể

6.1. Kiểm định dựa trên phối hợp từng cặp:

Chọn một mẫu ngẫu nhiên có n cặp quan sát (x_i,y_i) từ hai tổng thể X, Y có phân phối chuẩn. D_0 là một giá trị cho trước, ta thực hiện các bước kiểm định như sau:

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**
1. Đặt giả thuyết	$\begin{cases} H_0 : \mu_x - \mu_y \le D_0 \\ H_1 : \mu_x - \mu_y > D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y \ge D_0 \\ H_1 : \mu_x - \mu_y < D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y = D_0 \\ H_1 : \mu_x - \mu_y \ne D_0 \end{cases}$
2. Giá trị kiểm định	$t = \frac{\bar{d} - D_0}{S_d / \sqrt{n}}$		
3. Quyết định bác bỏ H_0 khi:	$t > t_{n-1,\alpha}$	$t < -t_{n-1,\alpha}$	$t > t_{n-1,\alpha/2}$; $t < -t_{n-1,\alpha/2}$

Trong đó:

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

$$S_d^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1} = \frac{\sum_{i=1}^{n} d_i^2 - n.\bar{d}^2}{n-1}$$

Ví dụ 5.6: Một công ty nước giải khát muốn xem xét ảnh hưởng của chiến dịch khuyến mãi đến việc tăng doanh số. 15 cửa hàng trong hệ thống phân phối sản phẩm của công ty được chọn ngẫu nhiên với số liệu về doanh số bán trong tuần lễ trước và sau chiến dịch khuyến mãi được ghi nhận ở bảng sau. Ở mức ý nghĩa 0,05, có thể kết luận chiến dịch khuyến mãi làm tăng doanh số hay không? Cho biết doanh số bán có phân phối chuẩn.

Cửa hàng	**Doanh số trong tuần (triệu đồng)**	
	Trước khuyến mãi	**Sau khuyến mãi**
1	57	60
2	61	54
3	12	20
4	38	35

5	12	21
6	69	70
7	5	1
8	39	65
9	88	79
10	9	10
11	92	90
12	26	32
13	14	19
14	70	77
15	22	29

Gọi μ_x, μ_y là doanh số bán trung bình trước và sau khau khi khuyến mãi. Ta tiến hành tính toán thủ công như sau:

Cửa hàng	**Doanh số trong tuần (triệu đồng)**		**d_i**	**d_i^2**
	Trước khuyến mãi (X)	**Sau khuyến mãi (Y)**		
1	57	60	3	9
2	61	54	-7	49
3	12	20	8	64
4	38	35	-3	9
5	12	21	9	81
6	69	70	1	1
7	5	1	-4	16
8	39	65	26	676
9	88	79	-9	81
10	9	10	1	1
11	92	90	-2	4
12	26	32	6	36
13	14	19	5	25
14	70	77	7	49
15	22	29	7	49
	Cộng		**-48**	**1.150**

Ta có: n=15, $\bar{d}$ = -3,2, D_0=0, S_d=8,43, α=5%

(1). Đặt giả thuyết: $\begin{cases} H_0 : \mu_x - \mu_y \geq 0 \\ H_1 : \mu_x - \mu_y < 0 \end{cases}$

(2). Giá trị kiểm định: $t = \dfrac{\bar{d} - D_0}{S_d / \sqrt{n}} = \dfrac{-3,2 - 0}{8,43/\sqrt{15}} = -1,47$

(3). Quyết định: $|t| = 1{,}47 < t_{14,\,0{,}05} = 1{,}761$ => Chấp nhận H_0

(4). Kết luận: Ở mức ý nghĩa $\alpha = 5\%$, không thể cho rằng sau chiến dịch khuyến mãi doanh số của công ty tăng lên so với trước.

- Thực hiện trên Excel để xử lý: Các bước thực hiện như sau:

(1). Tools → Data Analysis → t-Test: Paired Two Sample for Means

(2). Nhập dữ liệu:

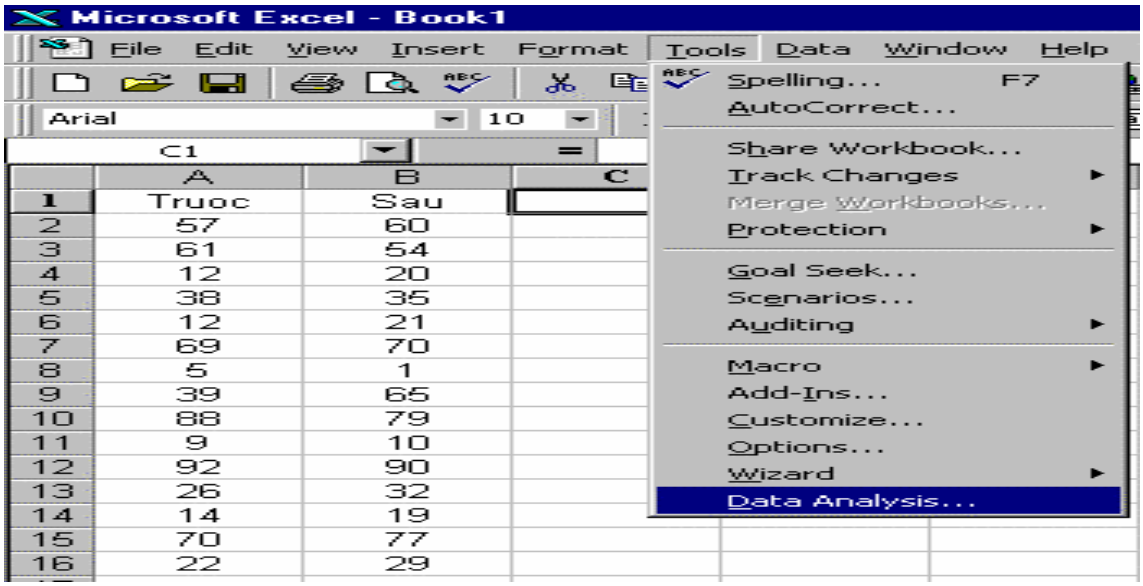
- Nhập số liệu theo cột

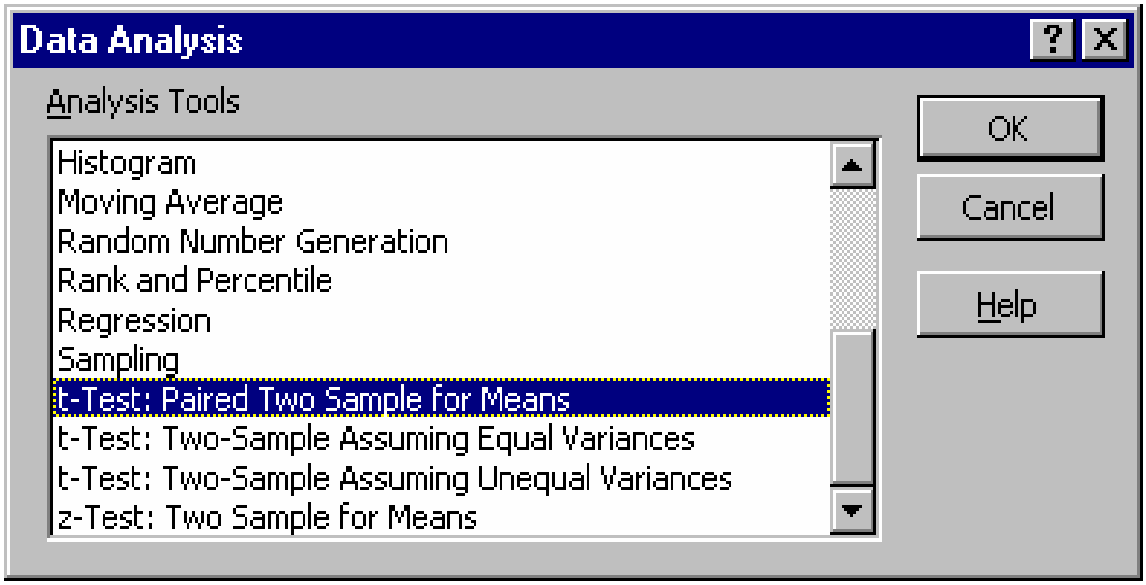
- Variable 1 Range: Chọn vùng xử lý của mẫu 1

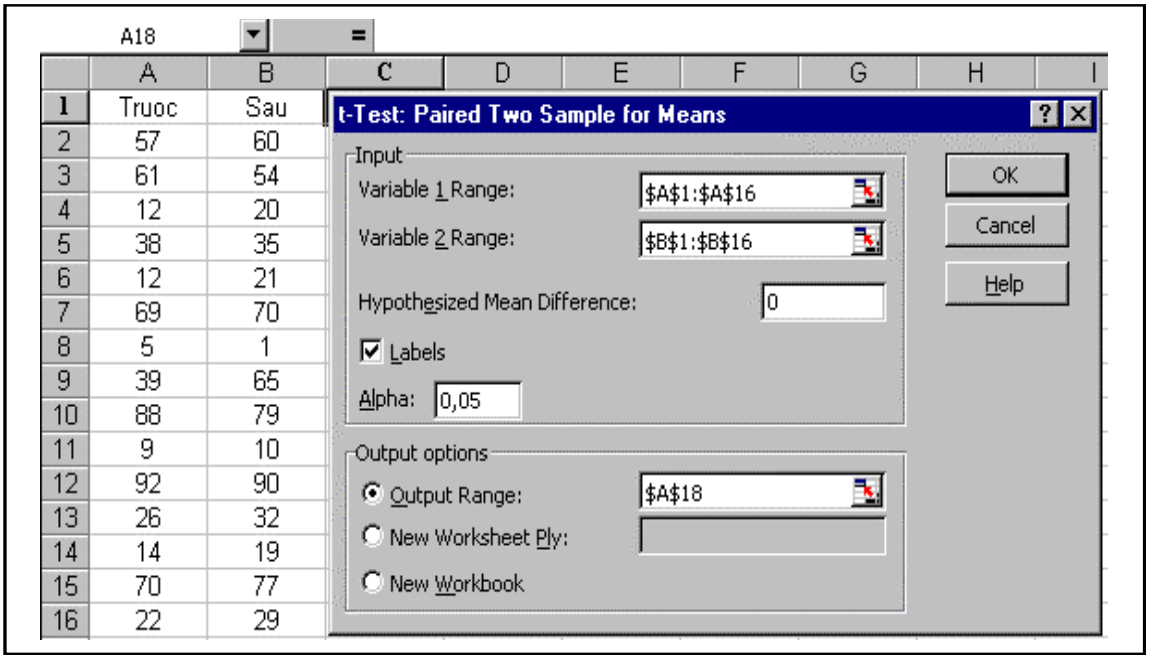
- Variable 2 Range: Chọn vùng xử lý của mẫu 2

- Hypothesized Mean Difference: Giá trị D_0

- Labels: Vùng xử lý có tên biến không.







Kết quả xử lý như sau:

t-Test: Paired Two Sample for Means

	X	*Y*	*Chú thích*
Mean	40,93	44,13	**Trung bình mẫu**
Variance	887,21	791,98	Phương sai mẫu
Observations	15	15	Số quan sát
Pearson Correlation	0,96		Hệ số tương quan
Hypothesized Mean Difference	0		D_0
df	14		Bậc tự do
t Stat	-1,47		Giá trị kiểm định
P(T<=t) one-tail	0,082		Giá trị p – 1 đuôi
t Critical one-tail	1,761		Giá trị t tra bảng – 1 đuôi
P(T<=t) two-tail	0,164		Giá trị p – 2 đuôi
t Critical two-tail	2,145		Giá trị t tra bảng – 2 đuôi

6.2. Kiểm định dựa trên mẫu độc lập:

Gọi n_x, n_y là số quan sát của các mẫu ngẫu nhiên độc lập $x_1, x_2, \ldots x_{nx}, y_1, y_2, \ldots y_{ny}$ từ hai tổng thể X và Y có trung bình μ_x, μ_y và phương sai σ^2_x, σ^2_y. Với trung bình mẫu $\bar{x}$, $\bar{y}$, phương sai mẫu là S^2_x, S^2_y, với mức ý nghĩa α.

6.2.1. Nếu biết phương sai tổng thể:

Điều kiện trong trường hợp này hai tổng thể có phân phối chuẩn hoặc mẫu có cỡ mẫu lớn ($n_x, n_y \geq 30$).

Ta có kiểm định Z như sau:

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**

1. Đặt giả thuyết	$\begin{cases} H_0 : \mu_x - \mu_y \le D_0 \\ H_1 : \mu_x - \mu_y > D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y \ge D_0 \\ H_1 : \mu_x - \mu_y < D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y = D_0 \\ H_1 : \mu_x - \mu_y \ne D_0 \end{cases}$
2. Giá trị kiểm định	$Z = \dfrac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}}$		
3. Quyết định bác bỏ H_0 khi:	$Z > Z_\alpha$	$Z < -Z_\alpha$	$Z > Z_{\alpha/2}$; $Z < -Z_{\alpha/2}$

- Xử lý trên Excel: Tools → Data Analysis → Z-Test: Two Sample for Means

6.2.2. Nếu chưa biết phương sai tổng thể, giả sử 2 phương sai khác nhau:

a) Trường hợp có cỡ mẫu lớn (n_x, $n_y \ge 30$): ta vẫn sử dụng công thức trên với phương sai mẫu thay cho phương sai tổng thể và không cần điều kiện phân phối chuẩn.

Dựa vào định lý giới hạn trung tâm ta không cần điều kiện phân phối chuẩn của tổng thể, khi chưa biết phương sai của tổng thể ta sử dụng phương sai của mẫu để tính giá trị kiểm định.

Ta có kiểm định Z:

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**
1. Đặt giả thuyết	$\begin{cases} H_0 : \mu_x - \mu_y \le D_0 \\ H_1 : \mu_x - \mu_y > D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y \ge D_0 \\ H_1 : \mu_x - \mu_y < D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y = D_0 \\ H_1 : \mu_x - \mu_y \ne D_0 \end{cases}$
2. Giá trị kiểm định	$Z = \dfrac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\dfrac{S_x^2}{n_x} + \dfrac{S_y^2}{n_y}}}$		
3. Quyết định bác bỏ H_0 khi:	$Z > Z_\alpha$	$Z < -Z_\alpha$	$Z > Z_{\alpha/2}$; $Z < -Z_{\alpha/2}$

b) Trường hợp mẫu nhỏ ($n_x < 30$ hoặc $n_y < 30$), với giả định cả hai tổng thể X và Y đều có phân phối chuẩn, ta vẫn sử dụng phương sai mẫu thay cho phương sai tổng thể, nhưng khi đó tiêu chuẩn kiểm định theo phân phối Student với bậc tự do được xác định bởi công thức sau:

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**
1. Đặt giả thuyết	$\begin{cases} H_0 : \mu_x - \mu_y \le D_0 \\ H_1 : \mu_x - \mu_y > D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y \ge D_0 \\ H_1 : \mu_x - \mu_y < D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y = D_0 \\ H_1 : \mu_x - \mu_y \ne D_0 \end{cases}$
2. Giá trị kiểm định	$t = \dfrac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\dfrac{S_x^2}{n_x} + \dfrac{S_y^2}{n_y}}}$, Bậc tự do $n = \dfrac{(S_x^2/n_x + S_y^2/n_y)^2}{\dfrac{(S_x^2/n_x)^2}{n_x - 1} + \dfrac{(S_y^2/n_y)^2}{n_y - 1}}$		
3. Quyết định bác bỏ H_0 khi:	$t > t_{n,\alpha}$	$t < -t_{n,\alpha}$	$t > t_{n,\alpha/2}$; $t < -t_{n,\alpha/2}$

- Xử lý trên Excel: Tools → Data Analysis → t-Test: Two-Sample Assuming Unequal Variances.

Trong trường hợp này máy tính không phân biệt giữa mẫu lớn và mẫu nhỏ và đều sử dụng phân phối Student để kiểm định, bởi vì khi mẫu lớn thì phân phối Student và phân phối chuẩn xấp xỉ nhau.

6.2.3. Giả sử 2 phương sai bằng nhau:

Trường hợp này thuận lợi cho trường hợp mẫu nhỏ và ta cần phải có điều kiện phân phối chuẩn của hai tổng thể.

Kiểm định t, bậc tự do (n_x+n_y-2):

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**
1. Đặt giả thuyết	$\begin{cases} H_0 : \mu_x - \mu_y \le D_0 \\ H_1 : \mu_x - \mu_y > D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y \ge D_0 \\ H_1 : \mu_x - \mu_y < D_0 \end{cases}$	$\begin{cases} H_0 : \mu_x - \mu_y = D_0 \\ H_1 : \mu_x - \mu_y \ne D_0 \end{cases}$
2. Giá trị kiểm định	$t = \dfrac{(\bar{x} - \bar{y}) - D_0}{\sqrt{S^2(\dfrac{1}{n_x} + \dfrac{1}{n_y})}}$, Trong đó $S^2 = \dfrac{(n_x - 1).S_x^2 + (n_y - 1).S_y^2}{n_x + n_y - 2}$		
3. Quyết định bác bỏ H_0 khi:	$t > t_{nx+ny-2,\alpha}$	$t < - t_{nx+ny-2,\alpha}$	$t > t_{nx+ny-2,\alpha/2}$; $t < -t_{nx+ny-2,\alpha/2}$

- Xử lý trên Excel: Tools → Data Analysis → t-Test: Two-Sample Assuming Equal Variances.

Ví dụ 5.7: Một nghiên cứu về hai nhãn hiệu pin X và Y (cùng chủng loại) của hai nhà sản xuất khác nhau được thực hiện. Chọn ngẫu nhiên mỗi nhãn hiệu 100 pin, kết quả được ghi nhận như sau: Pin X có thời gian sử dụng trung bình là 308 phút, độ lệch chuẩn 84 phút; các chỉ số tương ứng của pin Y lần lượt là 254 phút và 67 phút. Có thể kết luận thời gian sử dụng trung bình của pin X lớn hơn pin Y ít nhất là 45 phút được không với mức ý nghĩa α=0,1.

Trong bài toán này chưa đề cập đến việc phương sai của hai tổng thể này có phương sai giống nhau hay khác nhau. Để giải quyết trường hợp này nếu chúng ta đã có thông tin trước về phương sai của tổng thể thì ta căn cứ vào đó để lựa chọn công thức phù hợp. Cụ thể trong trường hợp này nếu chưa biết, chúng ta có thể thực hiện kiểm định về phương sai trước:

- *Kiểm định phương sai:*

(1). Giả thuyết: $\begin{cases} H_0 : \sigma_x^2 = \sigma_y^2 \\ H_1 : \sigma_x^2 > \sigma_y^2 \end{cases}$

(2). Giá trị kiểm định: $F = \dfrac{84^2}{67^2} = 1{,}57$

(3). Quyết định: F = 1,26 < $F_{99,99,10\%}$=1,295 => Chấp nhận H_0

(4). Kết luận: với α=10%, phương sai hai tổng thể là bằng nhau.

- *Kiểm định trung bình:*

Ta có: n_x=n_y=100, $\bar{x}$=308, $\bar{y}$=254, D_0=45, S_x=84, S_y=67, α=10%.

(1). Đặt giả thuyết: $\begin{cases} H_0 : \mu_x - \mu_y \le 45 \\ H_1 : \mu_x - \mu_y > 45 \end{cases}$

(2). Giá trị kiểm định: $z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}} = \frac{308 - 254 - 45}{\sqrt{\frac{84^2}{100} + \frac{67^2}{100}}} = 0{,}838$

(3). Quyết định: $|Z| = 0{,}838 < Z_{0,1} = 1{,}28$, chấp nhận giả thuyết H_0.

(4). Kết luận: Ở mức ý nghĩa 10%, không đủ chứng cớ để kết luận thời gian sử dụng trung bình của pin X lớn hơn pin Y ít nhất là 45 phút.

7. Kiểm định sự khác biệt của hai tỷ lệ tổng thể (với cỡ mẫu lớn ≥ 40)

Chọn 2 mẫu ngẫu nhiên độc lập từ 2 tổng thể X, Y có cỡ mẫu lớn có tỷ lệ tổng thể và tỷ lệ mẫu lần lượt là p_x, p_y, $\hat{p}_x$, $\hat{p}_y$. Giá trị kiểm định cho trước là p_0 và mức ý nghĩa α.

Ta có các trường hợp sau:

7.1. Trường hợp 1: Chênh lệch hai tỷ lệ bằng $p_0=0$.

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**
1. Đặt giả thuyết	$\begin{cases} H_0 : p_x - p_y \le 0 \\ H_1 : p_x - p_y > 0 \end{cases}$	$\begin{cases} H_0 : p_x - p_y \ge 0 \\ H_1 : p_x - p_y < 0 \end{cases}$	$\begin{cases} H_0 : p_x - p_y = 0 \\ H_1 : p_x - p_y \ne 0 \end{cases}$
2. Giá trị kiểm định	$Z = \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_x} + \frac{1}{n_y})}}$		
3. Quyết định bác bỏ H_0 khi:	$Z > Z_\alpha$	$Z < -Z_\alpha$	$Z > Z_{\alpha/2}$; $Z < -Z_{\alpha/2}$

Trong đó: $\hat{p} = \frac{n_x \hat{p}_y + n_y \hat{p}_y}{n_x + n_y}$

Nếu $\hat{p}_x = \frac{m_x}{m_x}$, $\hat{p}_y = \frac{m_y}{m_y}$ => $\hat{p} = \frac{m_x + m_y}{n_x + n_y}$

7.2. Trường hợp 2: Chênh lệch hai tỷ lệ bằng $p_0 \ne 0$

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**
1. Đặt giả thuyết	$\begin{cases} H_0 : p_x - p_y \le p_0 \\ H_1 : p_x - p_y > p_0 \end{cases}$	$\begin{cases} H_0 : p_x - p_y \ge p_0 \\ H_1 : p_x - p_y < p_0 \end{cases}$	$\begin{cases} H_0 : p_x - p_y = p_0 \\ H_1 : p_x - p_y \ne p_0 \end{cases}$
2. Giá trị kiểm định	$Z = \frac{(\hat{p}_x - \hat{p}_y) - p_0}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}}$		
3. Quyết định bác bỏ H_0 khi:	$Z > Z_\alpha$	$Z < -Z_\alpha$	$Z > Z_{\alpha/2}$; $Z < -Z_{\alpha/2}$

Ví dụ 5.8: Một công ty nước giải khát đang nghiên cứu việc đưa vào một công thức mới để cải tiến sản phẩm của mình. Với công thức cũ, khi cho 500 người dùng thử thì có 120 người tỏ ra ưa thích nó. Với công thức mới, khi cho 1000 người khác dùng thử thì có

300 người tỏ ra ưa thích nó. Hãy kiểm định xem công thức mới đưa vào có làm tăng tỷ lệ những người ưa thích nước giải khát hay không với α=5%?

Gọi p_x, p_y là tỷ lệ người ưa thích sản phẩm cũ, mới.

Ta có: n_x=500, n_y=1.000, $\hat{p}_x$=120/500=0,24, $\hat{p}_y$=300/1000=0,3, p_0=0, α=5%.

(1). Đặt giả thuyết: $\begin{cases} H_0 : p_x - p_y \geq 0 \\ H_1 : p_x - p_y < 0 \end{cases}$

(2). Giá trị kiểm định:

Ta có: $\hat{p} = \dfrac{120 + 300}{500 + 1500} = 0{,}28$

$$z = \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_x} + \frac{1}{n_y})}} = \frac{0{,}24 - 0{,}3}{\sqrt{0{,}28(1-0{,}28)(\frac{1}{500} + \frac{1}{1.000})}} = -2{,}44$$

(3). Quyết định: $|z| = 2{,}44 > 1{,}645 = Z_{0,05}$, bác bỏ H_0.

(4). Kết luận: Với α=5%, ta có thể kết luận xu hướng khách hàng ưa chuộng sản phẩm với công thức mới hơn.

III. KIỂM ĐỊNH PHI THAM SỐ:

Ở phần II, chúng ta đã nói đến kiểm định giả thuyết về các đặc trưng - trung bình, tỷ lệ, phương sai - của tổng thể và thường giả định tổng thể phân phối chuẩn. Trong phần này, cũng với giả thuyết H_0 về tham số tổng thể, chúng ta sẽ đề cập đến các kiểm định mà phần lớn không gắn liền tới tham số nào đó của mẫu, và vì vậy, chúng được gọi là kiểm định phi tham số. Kiểm định phi tham số thường không yêu cầu điều kiện giả định phân phối chuẩn của tổng thể, do đó có nhiều ứng dụng hơn. Tuy nhiên, phương pháp kiểm định phi tham số khó mở rộng và kém chính xác hơn so với kiểm định tham số.

1. Kiểm định Willcoxon (Kiểm định T)

Kiểm định Wilcoxon được áp dụng trong trường hợp chúng ta kiểm định về sự bằng nhau của hai trung bình tổng thể đối với mẫu phối hợp từng cặp.

Trước khi đi vào phương pháp giải quyết ta định nghĩa hạng (rank) của phần tử.

Giả sử ta có một dãy các số thực được xếp theo thứ tự tăng dần, trong dãy này không có giá trị nào bằng nhau:

$$x_1 < x_2 < ... < x_n$$

Khi đó rank(x_1) = 1, rank(x_2) = 2, ..., rank(x_n) = n

Trong trường hợp các phần tử có giá trị bằng nhau thì hạng của nó là hạng trung bình của các hạng liên tiếp.

a) Trường hợp mẫu nhỏ (n ≤20):

Chọn ngẫu nhiên n cặp quan sát (x_i,y_i) từ hai tổng thể X, Y. Với mức ý nghĩa α ta có các bước kiểm định sau:

(1). Đặt giả thuyết: $\begin{cases} H_0 : \mu_x - \mu_y = 0 \\ H_1 : \mu_x - \mu_y \neq 0 \end{cases}$

(2). Giá trị kiểm định:

- Tính các chênh lệch giữa các cặp: $d_i = x_i - y_i$

- Xếp hạng giá trị tuyệt đối các chênh lệch $|d_i|$ theo thứ tự tăng dần, các giá trị bằng nhau sẽ nhận hạng trung bình, bỏ qua trường hợp chênh lệch bằng 0.

- Gọi n^+ là số các $d_i \neq 0$

- Tìm tổng các hạng được xếp của d_i mang dấu dương $T^+ = \sum_{di>0} rank(|d_i|)$

- Tìm tổng các hạng được xếp của d_i mang dấu âm $T^- = \sum_{di<0} rank(|d_i|)$

- Kiểm định $T=\min(T^+, T^-)$

(3). Qui tắc quyết định: Ở mức ý nghĩa α, bác bỏ H_0 nếu $T<T_{n+,\alpha}$, với $T_{n+,\alpha}$ là giá trị của kiểm định Wilcoxon, n^+ là số cặp quan sát có $d_i \neq 0$.

Ví dụ 5.9: Mẫu 9 khách hàng được chọn ngẫu nhiên và yêu cầu họ cho biết sở thích về 2 sản phẩm cùng loại A, B thông qua thang điểm từ 1 (thấp nhất) đến 5 (cao nhất). Giả thuyết cho rằng không có xu hướng nghiên về loại nào trong sở thích của 2 loại kem đánh răng A, B với α=5%. Kết quả thu thập số liệu như sau:

Khách hàng	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Sản phẩm A	4	5	2	3	3	1	3	2	2
Sản phẩm B	3	5	5	2	5	5	3	5	5

(1). Đặt giả thuyết: $\begin{cases} H_0 : \mu_x - \mu_y = 0 \\ H_1 : \mu_x - \mu_y \neq 0 \end{cases}$

(2). Giá trị kiểm định:

Khách hàng	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**T**
Kem A	4	5	2	3	3	1	3	2	2	
Kem B	3	5	5	2	5	5	3	5	5	
Chênh lệch	1	0	-3	1	-2	-4	0	-3	-3	
Hạng +	1,5			1,5						3
Hạng -			5		3	7		5	5	25

$T = \min(T^+, T^-) = \min(3, 25) = 3$

$n^+ = 7$

(3). Quyết định: $T = 3 < T_{7,5\%} = 4$ => Bác bỏ giả thuyết H_0.

(4). Kết luận: Với α=5%, có thể kết luận rằng có sự khác biệt trong việc ưa chuộng hai loại kem đánh răng.

b) Trường hợp mẫu lớn (n>20):

Nếu n lớn thì phân phối Wilcoxon gần như phân phối chuẩn, lúc này trung bình và phương sai được tính như sau:

Giá trị kiểm định: $Z = \dfrac{T - \mu_T}{\sigma_T}$

Trung bình: $\mu_T = \dfrac{n(n+1)}{4}$

Phương sai: $\sigma_T^2 = \dfrac{n(n+1)(2n+1)}{24}$

2. Kiểm định Mann - Whitney (Kiểm định U)

Cũng như kiểm định T, nhưng kiểm định U xem xét trường hợp các mẫu độc lập.

Chọn 2 mẫu ngẫu nhiên độc lập có n_1, n_2 quan sát từ hai tổng thể có trung bình là μ_1, μ_2. Với mức ý nghĩa α, các bước kiểm định:

(1). Đặt giả thuyết $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$

(2). Giá trị kiểm định:

- Xếp hạng tất cả các giá trị của hai mẫu theo thứ tự tăng dần. Những giá trị bằng nhau sẽ nhận hạng trung bình hai hạng liên tiếp.

- Cộng các hạng của tất cả các giá trị ở mẫu thứ nhất. Ký hiệu: R_1

- Giá trị kiểm định:

$$U_1 = n_1.n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad U_2 = n_1n_2 - U_1$$

$U = \min(U_1, U_2)$

Khi cỡ mẫu lớn ($n_1, n_2 \geq 10$) phân phối U được xem là phân phối chuẩn

Giá trị kiểm định: $Z = \dfrac{U - \mu_U}{\sigma_U}$

Với: $\mu_U = \dfrac{n_1n_2}{2}$, $\sigma_U^2 = \dfrac{n_1.n_2(n_1+n_2+1)}{12}$

Ví dụ 5.10: Tại một trang trại nuôi lợn người ta thử áp dụng một loại thuốc tăng trọng bổ sung vào khẩu phần thức ăn của 10 con lợn, sau 3 tháng người ta thu thập số liệu về trọng lượng của heo (X). Trong khi đó 15 con lợn khác không dùng thuốc tăng trọng có trọng lượng, sau 3 tháng người ta thu thập số liệu (Y). Hãy kiểm tra xem trọng lượng có như nhau hay không khi thử nghiệm với α=5%.

																Tổng
X	60	61	62	62	63	63	68	64	64	65						
Y	56	56	57	57	58	58	58	59	59	60	60	60	61	61	62	
rank(x)	11,5	15	18	18	20,5	20,5	25	22,5	22,5	24						**197,5**
rank(y)	1,5	1,5	3,5	3,5	6	6	6	8,5	8,5	11,5	11,5	11,5	15	15	18	**127,5**

(1). Đặt giả thuyết $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$

(2). Giá trị kiểm định:

$$U_1 = 10x15 + \frac{10(10+1)}{2} - 197,5 = 7,5 \qquad U_2 = 10x15 - 7,5 = 142,5$$

$U = \min(U_1, U_2) = \min(7,5; 142,5) = 7,5$

$$\mu_U = \frac{10x15}{2} = 75, \quad \sigma_U^2 = \frac{10x15(10+15+1)}{12} = 325$$

$$Z = \frac{7,5 - 75}{\sqrt{325}} = -3,744$$

(3). Quyết định: $|Z| = 3,744 > Z_{2,5\%} = 1,96$ => Bác bỏ H_0.

(4). Kết luận: Với α=5%, trọng lượng có thay đổi khi sử dụng thuốc tăng trọng.

3. Kiểm định Kruskal – Wallis

Đây là trường hợp mở rộng của kiểm định Mann – Whitney, chúng ta sẽ thực hiện bài toán kiểm định về sự bằng nhau của k trung bình tổng thể.

Chọn k mẫu ngẫu nhiên độc lập có $n_1,...,n_k$ quan sát, gọi $n = \Sigma n_i$. Xếp hạng tất cả các quan sát theo thứ tự tăng dần, những giá trị bằng nhau sẽ nhận hàng trung bình. Gọi $R_1,...,R_k$ là tổng hạng của từng mẫu.

(1). Giả thuyết: $\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_k \\ H_1 : \exists \mu_i \neq \mu_j (i \neq j) \end{cases}$

(2). Giá trị kiểm định: $W = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$

(3). Quyết định bác bỏ H_0: $W > \chi^2_{k-1,\alpha}$

Ví dụ 5.11: Một nhà nghiên cứu muốn xem xét tổng giá trị sản phẩm sản xuất của 3 ngành A, B, C có giống nhau không, người ta chọn một số xí nghiệp hoạt động trong ngành các ngành này và có bảng số liệu sau:

Ngành A	1,38	1,55	1,90	2,00	1,22	2,11	1,98	1,61
Ngành B	2,33	2,50	2,79	3,01	1,99	2,45		
Ngành C	1,06	1,37	1,09	1,65	1,44	1,11		

Có thể kết luận gì ở 0,5%.

(1). Giả thuyết: H_0: $\mu_A=\mu_B=\mu_C$

(2). Giá trị kiểm định:

Ngành A	1,38	1,55	1,90	2,00	1,22	2,11	1,98	1,61	Tổng
Ngành B	2,33	2,50	2,79	3,01	1,99	2,45			
Ngành C	1,06	1,37	1,09	1,65	1,44	1,11			
rank(A)	6	8	11	14	4	15	12	9	79
rank(B)	16	18	19	20	13	17			103
rank(C)	1	5	2	10	7	3			28

$$W = \frac{12}{20(20+1)}\left(\frac{79^2}{8} + \frac{103^2}{6} + \frac{28^2}{6}\right) - 3(20+1) = 13{,}54$$

(3). Quyết định: $W=13,54 > \chi^2_{2;0,5\%}=10,597$ => Bác bỏ H_0

(4). Kết luận, với $\alpha=0,5\%$, tổng giá trị sản phẩm trung bình của các ngành là khác nhau.

4. Kiểm định sự phù hợp

Kiểm định sự phù hợp là kiểm định xem giả thuyết về phân phối của tổng thể và số liệu thực tế phù hợp (thích hợp) đến mức độ nào với giả định về phân phối tổng thể.

a) Kiểm định sự phù hợp trong trường hợp giả định đã biết các tham số của tổng thể:

Giả thuyết H_0 có phân phối xác suất p_i ($\Sigma p_i=1$) để một quan sát rơi vào nhóm i. Chọn một mẫu ngẫu nhiên n quan sát, được chia thành k nhóm khác nhau: mỗi quan sát chỉ thuộc vào một nhóm thứ i nào đó (i=1,..,k). O_i là số lượng quan sát ở nhóm thứ i. Vấn đề đặt ra là kiểm định giả thuyết H_0 về phân phối của tổng thể.

(1). Giả thuyết: H_0: Tổng thể có phân phối xác suất p_i

(2). Giá trị kiểm định:

Giá trị kiểm định: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ với $E_i=n.p_i$

Kiểm định sự phù hợp chỉ có ý nghĩa khi $E_i \geq 5$

(3). Qui tắc bác bỏ giả thuyết H_0: $\chi^2 > \chi^2_{k-1,\alpha}$

Ví dụ 5.12: Ở một bar, có 4 nhãn hiệu bia khác nhau. 160 khách hàng được chọn ngẫu nhiên cho thấy sự lựa chọn về các nhãn hiệu như sau:

Nhãn hiệu	A	B	C	D
Số khách hàng	34	46	29	51

Có thể kết luận sự ưa chuộng của khách hàng về 4 loại bia là như nhau được không ở mức ý nghĩa 2,5%.

Ta có bảng sau:

Nhãn hiệu (x)	A	B	C	D	
Số khách hàng (O_i)	34	46	29	51	160
Giả thuyết Ho (p_i)	0,25	0,25	0,25	0,25	1
$E_i=n.p_i$	40	40	40	40	
$(0_i-E_i)^2/E_i$	0,90	0,90	3,03	3,03	7,85

(1). Giả thuyết H_0: $p_A= p_B= p_C= p_D=0,25$

(2). Giá trị kiểm định: $\chi^2 = 7,85$

$E_i = 40>5$ (i=1,..,4), kiểm định có ý nghĩa

(3). Quyết định: $7,85 = \chi^2 < \chi^2_{k-1,\alpha} = \chi^2_{3,0,025} = 9,348$, chấp nhận giả thuyết H_0.

(4). Kết luận: Ở mức ý nghĩa 2,5% sự ưa chuộng của khác hàng về 4 nhãn hiệu bia là như nhau.

b) Kiểm định sự phù hợp trong trường hợp chưa biết các tham số của tổng thể:

Ở phần a), là phương pháp kiểm định sự phù hợp với xác suất để một quan sát rơi vào nhóm thứ i (p_i) đã được định rõ trong giả thuyết H_0.

Phần này ta nghiên cứu việc kiểm định giả thuyết các quan sát tuân theo một qui luật phân phối nào đó. Trong trường hợp này ta phải xác định p_i xác suất để một quan sát rơi vào nhóm thứ i. Sau đó áp dụng phương pháp tương tự như phần a).

5. Kiểm định về sự độc lập, kiểm định về mối liên hệ

Trong phần này, ta sẽ đề cập đến kiểm định trong việc xét xem giữa hai tiêu thức của tổng thể có mối liên hệ hay không. Ví dụ, mối liên hệ giữa giới tính với hành vi tiêu dùng,...

Giả sử có mẫu ngẫu nhiên gồm n quan sát, được phân nhóm kết hợp thành 2 tiêu thức với nhau, hình thành nên bảng phân nhóm kết hợp gồm r hàng và c cột. Gọi n_{ij} là số lượng quan sát tương ứng với hàng i và cột j. n là tổng quan sát của r hàng đồng thời cũng là tổng quan sát của c cột.

Phân nhóm theo	Phân nhóm theo tiêu thức thứ nhất

tiêu thức thứ hai	1	2	...	c	Σ
1	n_{11}	n_{12}	...	n_{1c}	R_1
2	n_{21}	n_{22}	...	n_{2c}	R_2
...	...	...	...	...	...
r	n_{r1}	n_{r2}	...	n_{rc}	R_r
Σ	C_1	C_2	...	C_c	n

Giả thuyết: H_0: không có mối liên hệ giữa hai tiêu thức

H_1: Tồn tại có mối liên hệ giữa hai tiêu thức

Giá trị kiểm định: $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$

Trong đó, $E_{ij} = C_j \frac{R_i}{n}$

Qui tắc kiểm định: Ở mức ý nghĩa α, bác bỏ giả thuyết H0 khi: $\chi^2 > \chi^2_{(r-1)(c-1),\alpha}$

Ví dụ 5.13: Một nghiên cứu được thực hiện nhằm xem xét mối liên hệ giữa giới tính và sự ưa thích các nhãn hiệu nước giải khát, một mẫu ngẫu nhiên 2.425 người tiêu dùng với các nhãn hiệu nước giải khát được ưa thích như sau:

	Nhãn hiệu ưa thích		
Giới tính	Coca	Pepsi	7Up
Nam	308	177	114
Nữ	502	627	697

Kiểm định giả thuyết không có mối liên hệ nào giữa giới tính và sự ưa thích nhãn hiệu nước giải khát ở mức ý nghĩa 0,5%.

(1). Giả thuyết H_0: Không có mối liên hệ giữa giới tính và sự ưa thích các nhãn hiệu nước giải khát.

(2). Giá trị kiểm định:

Giới tính	**Nhãn hiệu ưa thích**				
	Coke	Pepsi	7Up	R_i	χ^2
Nam	308	177	114	599	
$E_{1,j}$	200,08	198,60	200,33	599	
$(n_{1,j}-E_{1,j})^2/E_{1,j}$	58,21	2,35	37,20		**97,76**
Nữ	502	627	697	1826	
$E_{2,j}$	609,92	605,40	610,67	1826	
$(n_{2,j}-E_{2,j})^2/E_{2,j}$	19,10	0,77	12,20		**32,07**
Cj	810	804	811	2425	

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_i)^2}{E_i} = 97,76 + 32,07 = 129,83$$

(3). Quyết định: $129,83 = \chi^2 > \chi^2_{(2-1)(3-1),0,005} = \chi^2_{2,0,005} = 10,597$, bác bỏ giả thuyết H_0.

(4). Kết luận: Ở mức ý nghĩa 0,5%, giả thuyết H_0 bị bác bỏ, có nghĩa là có mối liên hệ giữa giới tính và sự ưa thích các nhãn hiệu nước giải khát.

IV. PHÂN TÍCH PHƯƠNG SAI (ANOVA)

Phân tích phương sai thực chất là bài toán kiểm định về sự bằng nhau của nhiều trung bình tổng thể, bài toán này đã được giải quyết trong phần III nhưng trong điều kiện tổng thể không có phân phối chuẩn và phương sai bằng nhau. Đây là điều kiện làm cho bài toán này khó được ứng dụng. Bài toán này được gọi là phân tích tích phương sai bởi vì khi giải quyết bài toán này người ta chủ yếu dựa vào tính chất của phương sai.

1. Phân tích phương sai một chiều:

Phân tích phương sai một chiều là phân tích dựa trên ảnh hưởng của một nhân tố.

Giả sử ta có k nhóm $n_1, ..., n_k$ quan sát được chọn ngẫu nhiên độc lập từ k tổng thể phân phối chuẩn và phương sai bằng nhau. Bảng giá trị quan sát sau:

1	2	...	k
$x_{1,1}$	$x_{2,1}$	...	$xk_{,1}$
...	...	...	...
$x_{1,n1}$	$x_{2,n2}$	...	$xk_{,nk}$

Kiểm định giả thuyết: H_0: $\mu_1 = ... = \mu_k$ (Trung bình theo cột bằng nhau)

Bước 1: Tính số trung bình.

- Trung bình từng cột: $\bar{x}_i = \dfrac{\sum_{j=1}^{ni} x_{ij}}{n_i}$ (i=1,...,k)

- Trung bình chung: $\bar{x} = \dfrac{\sum_{i=1}^{k} n_i \bar{x}_i}{n}$ với $n = \sum_{i=1}^{k} n_i$

Bước 2: Tính tổng độ lệch bình phương.

- Tổng độ lệch bình phương được sinh ra bởi yếu tố cột:

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

- Tổng độ lệch bình phương sai số:

+ Tổng độ lệch bình phương từng cột: $SS_i = \sum_{j=1}^{ni} (x_{ij} - \bar{x}_i)^2$

+ Tổng độ lệch bình phương của k cột: $SSW = \sum_{i=1}^{k} SS_i = \sum_{i=1}^{k} \sum_{j=1}^{ni} (x_{ij} - \bar{x}_i)^2$

- Tổng độ lệch bình phương chung: $SST = \sum_{i=1}^{k} \sum_{j=1}^{ni} (x_{ij} - \bar{x})^2$

SST = SSW + SSG

Đẳng thức trên thể hiện tinh thần của bài toán phân tích phương sai:

SST: Thể hiện sự biến thiên của hiện tượng nghiên cứu

SSG: Thể hiện sự biến thiên do yếu tố cột tạo ra

SSW: Thể hiện sự biến thiên do các yếu tố khác.

Như vậy hiện tượng nghiên cứu phụ thuộc vào hai phần: do yếu tố đang xem xét và những yếu tố khác tác động. Nếu như sự tác động của yếu đang nghiên cứu cũng như các

yếu tố khác tác động thì ta có thể kết luận hiện tượng nghiên cứu không phụ thuộc vào yếu tố đang xem xét. Điều này dẫn đến trung bình theo cột bằng nhau.

Bước 3: Tính phương sai:

- Phương sai được sinh ra bởi yếu tố cột: $MSG = \frac{SSG}{k-1}$

- Phương sai được sinh ra bởi yếu tố ngẫu nhiên khác: $MSW = \frac{SSW}{n-k}$

Bước 4: Giá trị kiểm định: $F = \frac{MSG}{MSW}$

Bước 5: Bác bỏ giả thuyết H_0 khi $F>F_{k-1,n-k,\alpha}$

Bảng tổng quát phân tích ANOVA

Biến thiên	Tổng độ lệch bình phương	Bậc tự do	Phương sai	Giá trị kiểm định
Giữa các nhóm	SSG	k-1	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSW}$
Trong nội bộ nhóm	SSW	n-k	$MSW = \frac{SSW}{n-k}$	
Tổng cộng	SST	n-1		

Ví dụ 5.14: Một nghiên cứu được thực hiện nhằm xem xét năng suất lúa trung bình của 3 giống lúa có bằng nhau hay không. Kết quả thu thập qua 4 năm như sau:

Năm	**A**	**B**	**C**
1	65	69	75
2	74	72	70
3	64	68	78
4	83	78	76

Hãy cho nhận xét với mức ý nghĩa α=5%.

Để thực hiện kiểm định này ta có thể áp dụng các công thức trên để giải quyết. Tuy nhiên điều này sẽ mất khá nhiều thời gian, đặc biệt là khi số lượng quan sát lớn. Ta có thể sử dụng phần mềm Excel giải quyết rất đơn giản:

- Các bước thực hiện trên Excel:

(1). Tools → Data Analysis – Anova: Single Factor

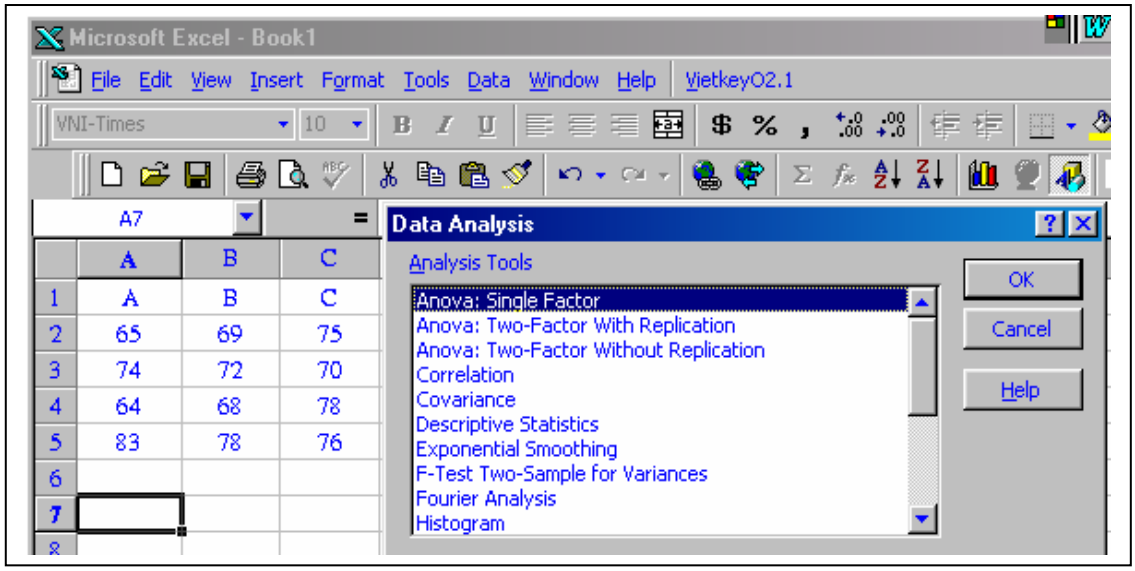
(2). Nhập số liệu:

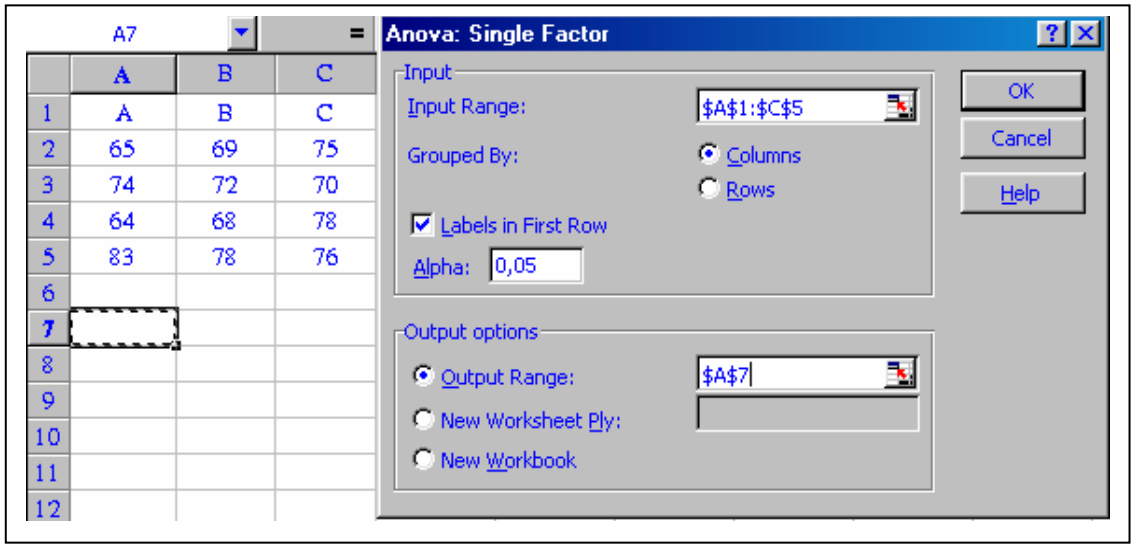
- Nhập số liệu theo cột hoặc hàng.

- Input Range: Chọn tất cả hàng, cột đưa vào vùng xử lý.

- Grouped By: Dữ liệu được nhập theo cột/hàng

- Labels in First Row: Vùng dữ liệu có tên biến không.





Kết quả xử lý của Excel với α=5%.

Anova: Single Factor

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	26,167	2	13,083	0,354	0,711	4,256
Within Groups	332,500	9	36,944			
Total	358,667	11				

(1). Giả thuyết : Giả thuyết ta có thể trình bày một trong 3 cách sau:

H_0: Năng suất trung bình của các giống lúa bằng nhau

H_0: $\mu_A=\mu_B=\mu_C$

H_0: Năng suất không phụ thuộc vào giống lúa

(2). Giá trị kiểm định: Trong trường hợp này máy tính đã cho chúng ta giá trị p do đó chúng ta có thể không cần quan tâm đến các giá trị F. Chính vì vậy ta có thể bỏ qua bước 2.

(3). Quyết định: p=71,1%, quá lớn => Chấp nhận H_0 hoàn toàn.

(4). Kết luận: Năng suất trung bình của 3 giống lúa là như nhau.

Với p=71,1% là quá lớn dẫn đến kết quả chấp nhận giả thuyết H_0.

2. Phân tích phương sai hai chiều

Phân tích phương sai hai chiều là xét đến hai yếu tố ảnh hưởng đến hiện tượng nghiên cứu.

2.1. Trường hợp có một quan sát trong cùng một ô:

Trường hợp này tương ứng với sự tác động của yếu tố cột và yếu tố hàng chúng ta chỉ chọn một quan sát. Đây là trường hợp mở rộng của phân tích phương sai một yếu tố, có nghĩa là ta vừa kiểm định giả thuyết trung bình theo cột bằng nhau vừa kiểm định trung bình theo hàng bằng nhau.

Kết quả chọn mẫu được lập thành bảng kết hợp 2 yếu tố như sau:

Yếu tố thứ hai (hàng)	**Yếu tố thứ nhất (cột)**			
	1	**2**	**...**	**k**
1	$x_{1,1}$	$x_{2,1}$	...	$x_{k,1}$
...	...	...	...	...
h	$x_{1,h}$	$x_{2,h}$	...	$x_{k,h}$

Giả thuyết H_0: - Trung bình của tổng thể theo chỉ tiêu cột bằng nhau,

- Trung bình của tổng thể theo chỉ tiêu hàng bằng nhau.

Ta có các bước sau:

Bước 1: Tính số trung bình.

- Tính trung bình từng cột: $\bar{x}_i = \dfrac{\sum_{j=1}^{h} x_{ij}}{h}$ (i=1,...,k)

- Tình trung bình từng hàng: $\bar{x}_j = \dfrac{\sum_{j=1}^{k} x_{ij}}{k}$ (j=1,...,h)

- Tính trung bình chung: $\bar{x} = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{h} x_{ij}}{n} = \dfrac{\sum_{i=1}^{k} \bar{x}_i}{k} = \dfrac{\sum_{j=1}^{h} \bar{x}_j}{h}$ (n = k.h)

Bước 2: Tính tổng độ lệch bình phương.

- Tổng độ lệch bình phương chung: $SST = \sum_{i=1}^{k}\sum_{j=1}^{h}(x_{ij} - \bar{x})^2$

- Tổng độ lệch bình phương sinh ra bởi yếu tố cột: $SSG = h\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2$

- Tổng độ lệch bình phương sinh ra bởi yếu tố hàng: $SSB = k\sum_{j=1}^{h}(\bar{x}_j - \bar{x})^2$

- Tổng độ lệch bình phương sai số: $SSE = \sum_{i=1}^{k}\sum_{j=1}^{h}(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$

SST = SSG + SSB + SSE

Tương tự như đối với phân tích phương sai một chiều nhưng bây giờ hiện tượng nghiên cứu phụ thuộc vào yếu tố cột, yếu tố hàng và các yếu tố ngẫu nhiên khác. Nếu sự biến động của yếu tố cột cũng như đối với những biến động ngẫu nhiên khác thì ta kết luật hiện tượng nghiên cứu không phụ thuộc vào yếu tố cột và dẫn đến trung bình theo cột bằng nhau. Tương tự như vậy, ta sẽ có kết luận đối với yếu tố hàng.

Bước 3: Tính phương sai.

- Phương sai sinh ra bởi yếu tố cột : $MSG = \frac{SSG}{k-1}$

- Phương sai sinh ra bởi yếu tố hàng : $MSB = \frac{SSB}{h-1}$

- Phương sai sinh ra bởi yếu ngẫu nhiên : $MSE = \frac{SSE}{(k-1)(h-1)}$

Bước 4: Giá trị kiểm định từ hai tỷ số F

- Kiểm định theo cột: $F_1 = \frac{MSG}{MSE}$

- Kiểm định theo hàng: $F_2 = \frac{MSB}{MSE}$

Bước 5: Quyết định bác bỏ giả thuyết H_0:

- Bác bỏ giả thuyết theo chỉ tiêu cột: $F_1 > F_{k-1,(k-1)(h-1),\alpha}$

- Bác bỏ giả thuyết theo chỉ tiêu hàng: $F_2 > F_{h-1,(k-1)(h-1),\alpha}$

Trong đó, $F_{v1,v2,\alpha}$ có phân phối FISHER.

Bảng tổng quát phân tích ANOVA

Biến thiên	Tổng độ lệch bình phương	Bậc tự do	Phương sai	Giá trị kiểm định F
Giữa các cột	SSG	k-1	$MSG = \frac{SSG}{k-1}$	$F_1 = \frac{MSG}{MSE}$
Giữa các hàng	SSB	h-1	$MSB = \frac{SSB}{h-1}$	$F_2 = \frac{MSB}{MSE}$
Sai số	SSE	(k-1)(h-1)	$MSE = \frac{SSE}{(k-1)(h-1)}$	
Tổng cộng	SST	n-1		

Ví dụ 5.11: Một nghiên cứu được thực hiện nhằm xem xét sự liên hệ giữa loại phân bón, giống lúa và năng suất. Năng suất lúa được ghi nhận từ các thực nghiệm sau:

Loại phân bón	**Giống lúa**		
	A	**B**	**C**
1	65	69	75
2	74	72	70
3	64	68	78
4	83	78	76

Hãy cho nhận xét với mức ý nghĩa 5%.

- Thực hiện bằng Excel:

(1). Tools → Data Analysis → Anova: Two-Factor Without Replication.

(2). Nhập số liệu:

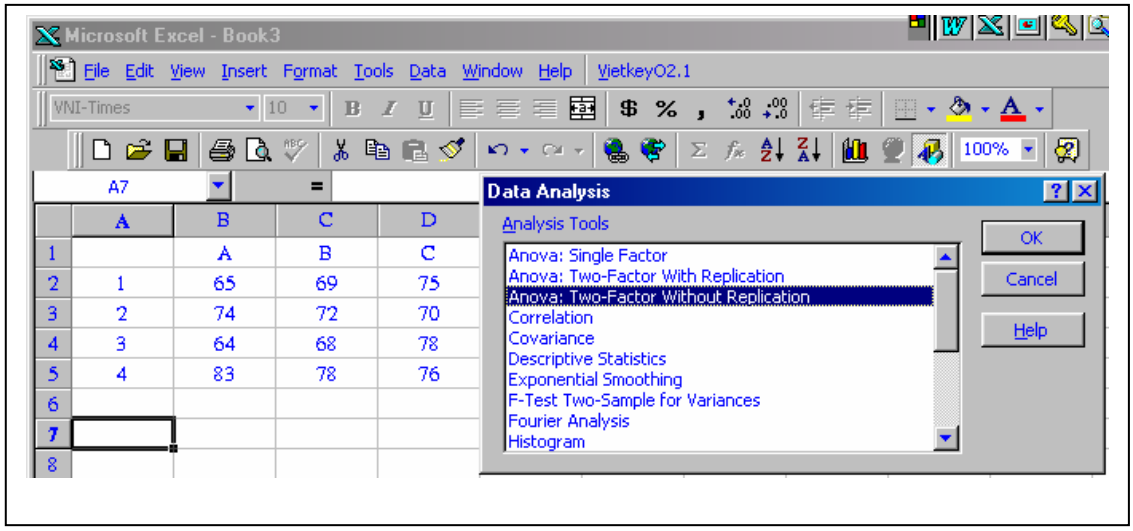
- Nhập số liệu nếu có tiêu đề thì phải có đủ cả hàng và cột.

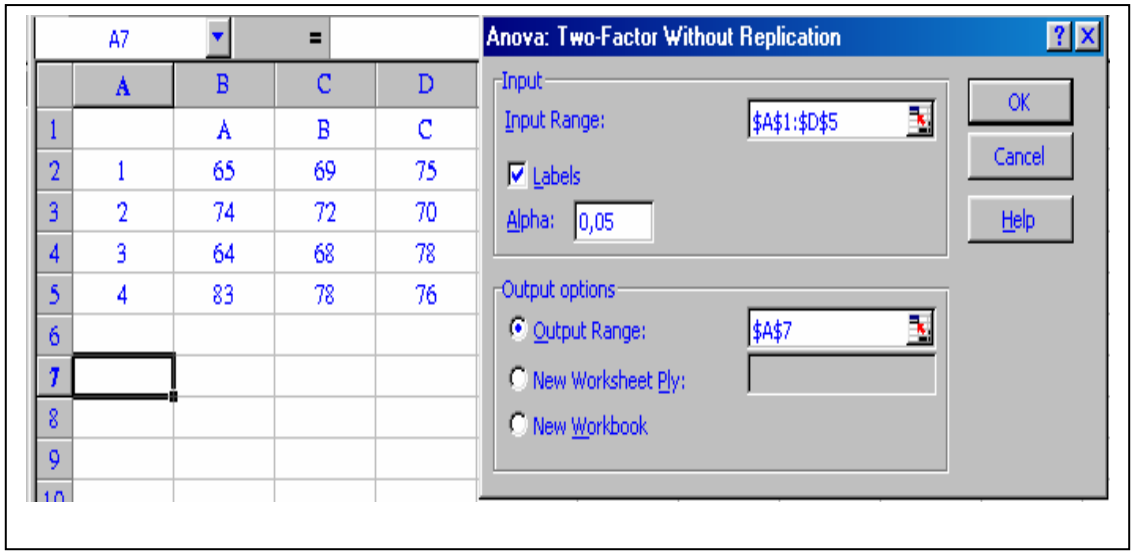
- Input Range: Chọn tất cả hàng, cột đưa vào vùng xử lý.

- Grouped By: Dữ liệu được nhập theo cột/hàng

- Labels in First Row: Nếu chọn thì máy tính hiểu rằng hàng đầu, cột đầu tiên là tên biến, không nằm trong vùng dữ liệu để tính toán.

Thực hiện trên Excel:





Kết quả phân tích ANOVA từ Excel α=5% như sau:

Anova: Two-Factor Without Replication

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	170,000	3	56,667	2,092	0,203	4,757
Columns	26,167	2	13,083	0,483	0,639	5,143
Error	162,500	6	27,083			

Total	358,667	11

Giải thích kết quả: Ta sẽ trình bày hai bài toán kiểm định:

- **Kiểm định theo cột:**
  (1). Giả thuyết: Năng suất không phụ thuộc vào giống.
  (2). Quyết định: p=63,9%, quá lớn => Chấp nhận H_0 hoàn toàn.
  (3). Kết luận: Năng suất không phụ thuộc vào giống.

- **Kiểm định theo hàng:**
  (1). Giả thuyết: Năng suất không phụ thuộc vào phân bón.
  (2). Quyết định: p=20,3%, quá lớn => Chấp nhận H_0 hoàn toàn.
  (3). Kết luận: Năng suất không phụ thuộc vào phân bón.

3. Trường hợp có hơn một tham số trong một ô

Trường hợp tương ứng với mỗi yếu tố cột và yếu hàng ta có thể chọn nhiều qua sát. Trong bài toán này, ngoài việc kiểm định về trung bình theo cột bằng nhau, trung bình theo hàng bằng nhau mà chúng ta còn xem xét sự tương tác giữa yếu tố hàng và yếu tố cột có ảnh hưởng đến hiện tượng nghiên cứu hay không.

Ta có bảng kết hợp 2 tiêu thức như sau:

Yếu tố thứ hai (hàng)	**Yếu tố thứ nhất (cột)**			
	1	**2**	**...**	**k**
1	x_{111} x_{112} ... x_{11l}	x_{211} x_{212} ... x_{21l}	...	x_{k11} x_{k12} ... x_{k1l}
...	...	...	...	...
h	x_{1h1} x_{1h2} ... x_{1hl}	x_{2h1} x_{2h2} ... x_{2hl}	...	x_{kh1} x_{kh2} ... x_{khl}

Giả thuyết H_0: - Trung bình của tổng thể theo chỉ tiêu cột bằng nhau.

- Trung bình của tổng thể theo chỉ tiêu hàng bằng nhau.

- Không có sự tương tác giữa yếu tố cột và hàng.

Ta có các bước sau:

Bước 1: Tính số trung bình.

- Trung bình từng cột: $\bar{x}_i = \dfrac{\sum_{j=1}^{h}\sum_{s=1}^{l} x_{ijs}}{h.l}$ (i=1,...,k)

- Trung bình từng hàng: $\bar{x}_j = \dfrac{\sum_{i=1}^{k}\sum_{s=1}^{l} x_{ijs}}{k.l}$ (j=1,...,h)

- Trung bình từng ô: $\bar{x}_{ij} = \dfrac{\sum_{s=1}^{l} x_{ijs}}{l}$

- Trung bình: $\bar{x} = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{h}\sum_{s=1}^{l} x_{ijs}}{k.h.l}$

Bước 2: Tính tổng độ lệch bình phương.

- Tổng độ lệch bình phương chung: $SST = \sum_{i=1}^{k}\sum_{j=1}^{h}\sum_{s=1}^{l}(x_{ijs} - \bar{x})^2$

- Tổng độ lệch sinh ra bởi yếu tố cột: $SSG = h.l\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2$

- Tổng độ lệch sinh ra bởi yếu tố hàng: $SSB = k.l\sum_{j=1}^{h}(\bar{x}_j - \bar{x})^2$

- Tổng độ lệch sinh ra bởi sự tương tác giữa hàng và cột:

$$SSI = l\sum_{i=1}^{k}\sum_{j=1}^{h}(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$$

- Tổng độ lệch bình phương sinh ra bởi sai số: $SSE = \sum_{i=1}^{k}\sum_{j=1}^{h}\sum_{s=1}^{l}(x_{ijs} - \bar{x})^2$

$SST = SSG + SSB + SSI + SSE$

Tương tự như hai bài toán trước, ta chỉ cần lý giải thêm đối với sự tương tác. Nếu sự tương tác biến thiên cũng như các yếu tố ngẫu nhiên khác thì ta kết luận hiện tượng nghiên cứu không phụ thuộc vào sự tương tác giữa yếu tố hàng và yếu tố cột. Điều này có thể kết luận thêm yếu tố hàng và yếu tố cột là độc lập với nhau trong sự tác động đến hiện tượng nghiên cứu.

Bước 3: Tính phương sai.

- Phương sai sinh ra bởi yếu tố cột: $MSG = \frac{SSG}{k-1}$

- Phương sai sinh ra bởi yếu tố hàng: $MSB = \frac{SSB}{h-1}$

- Phương sai sinh ra bởi sự tương tác: $MSI = \frac{SSI}{(k-1)(h-1)}$

- Phương sai sinh ra bởi yếu tố ngẫu nhiên: $MSE = \frac{SSE}{k.h.(l-1)}$

Bước 4: Giá trị kiểm định từ hai tỷ số F

- Kiểm định theo hàng: $F_1 = \frac{MSG}{MSE}$

- Kiểm định theo cột: $F_2 = \frac{MSB}{MSE}$

- Kiểm định sự tương tác hàng và cột: $F_3 = \frac{MSI}{MSE}$

Bước 5: Quyết định bác bỏ giả thuyết H_0:

- Bác bỏ giả thuyết theo chỉ tiêu cột: $F_1 > F_{k-1,kh(l-1),\alpha}$

- Bác bỏ giả thuyết theo chỉ tiêu hàng: $F_2 > F_{h-1,kh(l-1),\alpha}$

- Bác bỏ giả thuyết không có sự tương: $F_3 > F_{(k-1)(h-1),kh(l-1),\alpha}$

Ví dụ 5.12: Một nghiên cứu được thực hiện nhằm xem xét sự liên hệ giữa loại phân bón, giống lúa và năng suất. Năng suất lúa được ghi nhận từ các thực nghiệm sau:

Loại phân bón	**Giống lúa**								
	A			**B**			**C**		
1	65	68	62	69	71	67	75	75	78

2	74	79	76	72	69	69	70	69	65
3	64	72	65	68	73	75	78	82	80
4	83	82	84	78	78	75	76	77	75

Hãy cho nhận xét với α=5%.

- Thực hiện bằng Excel:

(1). Tools → Data Analysis → Anova: Two-Factor With Replication

(2). Nhập số liệu:

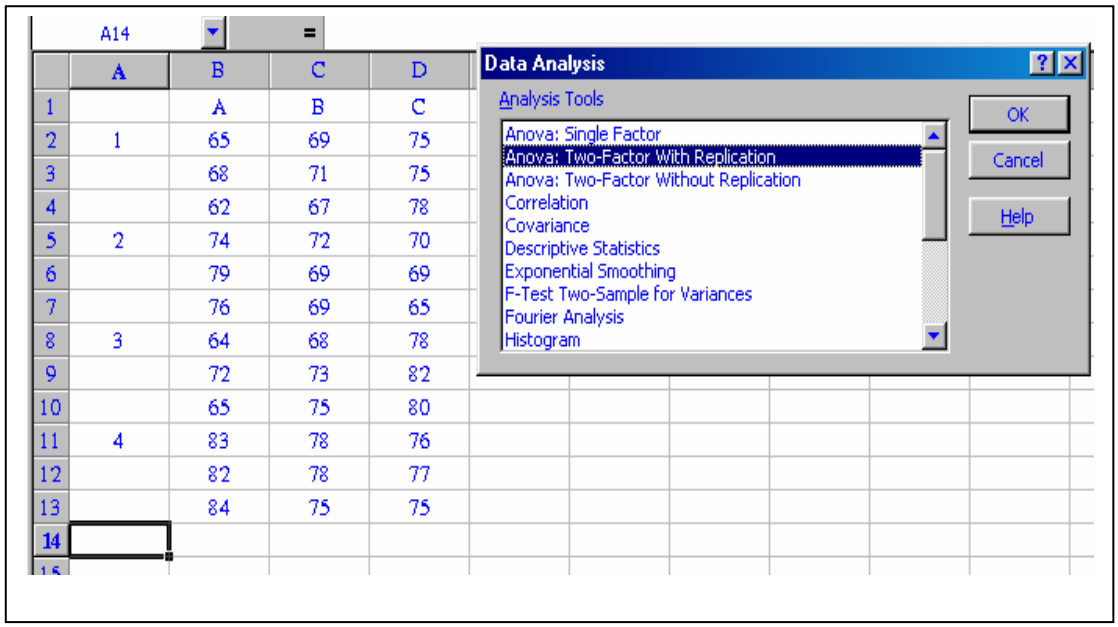
- Nhập số liệu cũng như phân tích phương sai hai yếu tố có một quan sát nhưng phải nhập số liệu theo cột và phải có tiêu đề hàng, cột.

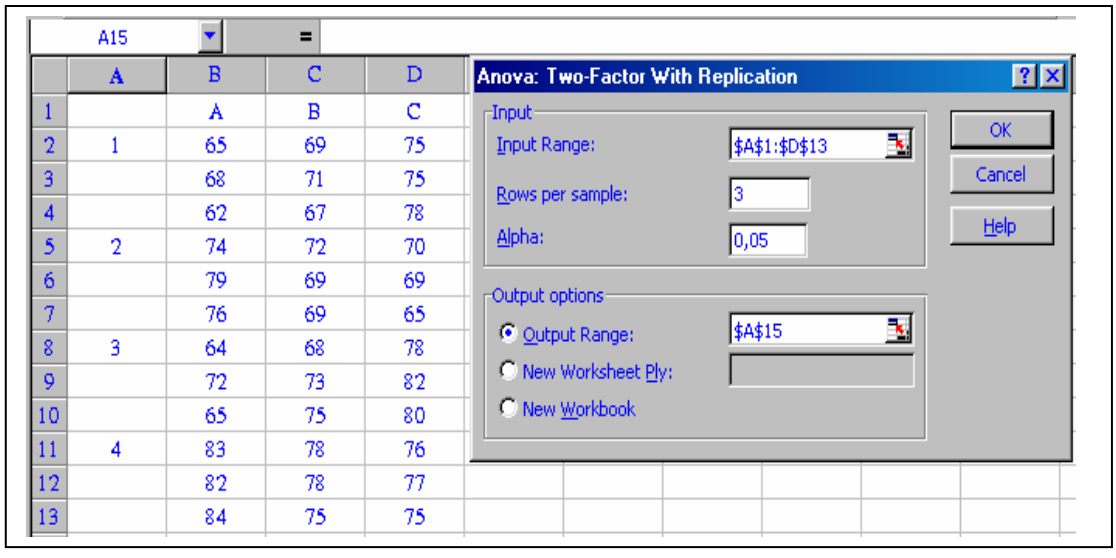
- Input Range: Chọn tất cả hàng, cột đưa vào vùng xử lý bao gồm tên tiêu đề.

- Grouped By: Dữ liệu được nhập theo cột/hàng

- Rows per sample: Số quan sát trong một ô.

Thực hành trên Exel:





Anova: Two-Factor With Replication

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	389,000	3	129,667	21,218	0,000	3,009
Columns	57,556	2	28,778	4,709	0,019	3,403
Interaction	586,000	6	97,667	15,982	0,000	2,508
Within	146,667	24	6,111			
Total	1179,222	35				

Giải thích kết quả: Ta sẽ trình bày ba bài toán kiểm định:

- **Kiểm định theo cột:**
  (1). Giả thuyết: Năng suất không phụ thuộc vào giống
  (2). Quyết định: α=5% > 1,9% = p => Bác bỏ H_0.
  (3). Kết luận: Với α=5%, năng suất phụ thuộc vào giống.

- **Kiểm định theo hàng:**
  (1). Giả thuyết: Năng suất không phụ thuộc vào phân bón.
  (2). Quyết định: p=0%, quá nhỏ => Bác bỏ H_0 hoàn toàn.
  (3). Kết luận: Năng suất không phụ thuộc vào phân bón.

- **Kiểm định về sự tương tác:**
  (1). Giả thuyết: Không có sự tương tác giữa yếu tố giống và phân bón đến năng suất.
  (2). Quyết định: p=0%, quá nhỏ => Bác bỏ H_0 hoàn toàn.
  (3). Kết luận: Có sự tương tác giữa yếu tố giống và phân bón đến năng suất.

  Có sự tương tác có thể giải thích rằng không phải khi sử dụng giống tốt nhất, phân tốt nhất thì có thể cho năng suất bởi vì có thể phân bón này sẽ không phù hợp cho giống nào đó.

CHƯƠNG VI
TƯƠNG QUAN VÀ HỒI QUI TUYẾN TÍNH

Trong chương này ta sẽ nói đến việc nghiên cứu mối liên hệ giữa hai hay nhiều biến ngẫu nhiên với hai phương pháp tương quan và hồi qui.

I. HỆ SỐ TƯƠNG QUAN

Hệ số đo lường mức độ tuyến tính giữa hai biến không phân biệt biến nào là phụ thuộc biến nào là độc lập.

1. Hệ số tương quan

Giả sử X, Y là 2 biến ngẫu nhiên hệ số tương quan tổng thể ρ_{xy} là khái niệm dùng để thể hiện cường độ và chiều hướng của mối liên hệ tuyến tính giữa X và Y nếu nó thoải mãn 5 điều kiện sau:

* $-1 \le \rho \le 1$

* $\rho < 0$: Giữa X và Y có mối liên hệ nghịch, nghĩa là nếu X tăng thì Y giảm và ngược lại.

* $\rho > 0$: Giữa X và Y có mối liên hệ thuận, nghĩa là nếu X tăng thì Y tăng và ngược lại.

* $\rho = 0$: Giữa X và Y không có mối liên hệ tuyến tính.

* $|\rho|$: càng lớn thì mối liên hệ giữa X và Y càng chặt chẽ.

Trong thực tế ta không biết chính xác được hệ số tương quan tổng thể mà phải ước lượng từ dữ liệu mẫu thu thập được.

Gọi (x_i,y_i) là mẫu n cặp quan sát thu thập ngẫu nhiên từ X và Y. Hệ số tương quan tổng thể ρ_{xy} được ước lượng từ hệ số tương quan mẫu r_{xy}. R_{xy} còn được gọi là hệ số tương quan Pearson, được xác định bởi công thức sau:

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

2. Kiểm định giả thuyết về mối liên hệ tương quan

Giả sử có mẫu n cặp quan sát chọn ngẫu nhiên từ X và Y có phân phối chuẩn. Kiểm định giả thuyết về hệ số tương quan của tổng thể $\alpha = 0$, tức là không có mối liên hệ giữa các biến X và Y. Các dạng kiểm định như sau:

	Một đuôi phải	**Một đuôi trái**	**Hai đuôi**
1. Đặt giả thuyết	$\begin{cases} H_0 : \rho \le 0 \\ H_1 : \rho > 0 \end{cases}$	$\begin{cases} H_0 : \rho \ge 0 \\ H_1 : \rho < 0 \end{cases}$	$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \ne 0 \end{cases}$
2. Giá trị kiểm định	$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$		
3. Quyết định bác bỏ H_0 khi:	$t > t_{n-2,\alpha}$	$t < -t_{n-2,\alpha}$	$t > t_{n-2,\alpha/2}$; $t < -t_{n-2,\alpha/2}$

Hệ số tương quan có một vài ứng dụng quan trọng trong việc kiểm định mô hình hồi qui tuyến tính, do đó chúng ta cũng cần quan tâm đúng mức khi đi sâu vào lĩnh vực kinh tế lượng.

II. MÔ HÌNH HỒI QUI TUYẾN TÍNH ĐƠN GIẢN

Như phần tương quan tuyến tính dùng để đo lường mức độ liên hệ tuyến tính giữa hai biến ngẫu nhiên X và Y nhưng trong đó X và Y có tính đối xứng (tức là X phụ thuộc vào Y thì Y cũng phụ thuộc vào X). Trong phần này ta cũng nghiên cứu mối liên hệ tuyến tính giữa X và Y, trong đó X ảnh hưởng đến Y và Y được xem là phụ thuộc vào X. Mối liên hệ giữa X và Y đã được xác định bằng một qui luật khách quan đã có. Mục tiêu của phân tích hồi qui là mô hình hoá mối liên hệ bằng một mô hình toán học nhằm thể hiện một cách tốt nhất mối liên hệ giữa X và Y.

Để bắt đầu, chúng ta hãy tìm hiểu các khái niệm cơ bản.

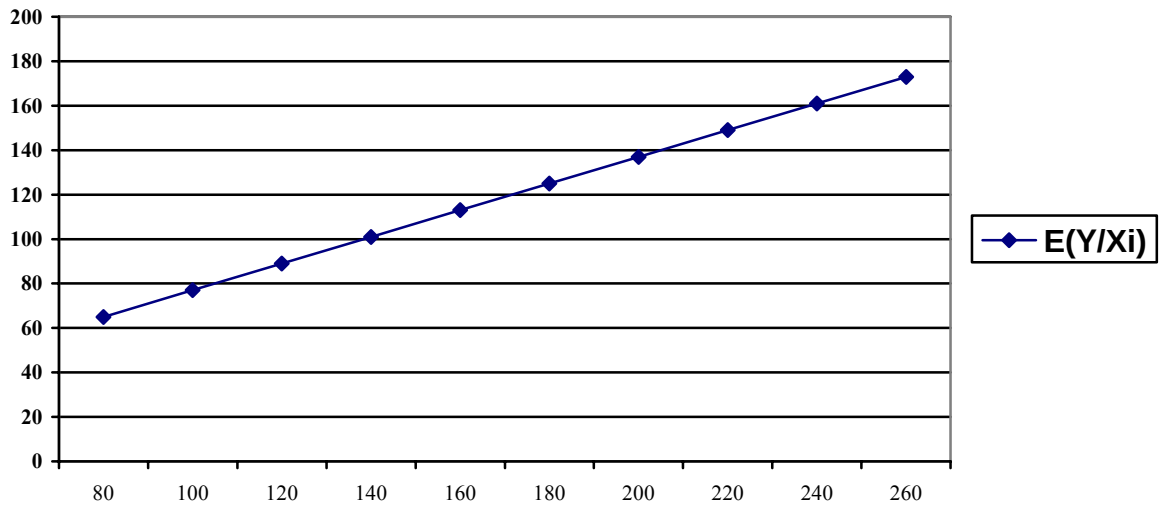
1. Mô hình hồi qui tuyến tính một chiều (tuyến tính đơn giản)

Ví dụ 6.1: Tìm hiểu mối liên hệ giữa chi tiêu (Y) và thu nhập sau khi trừ thuế của hộ gia đình (X):

X_i / Y	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**
	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
		88		113	125	140		160	189	185
				115				162		191
$E(Y/X_i)$	**65**	**77**	**89**	**101**	**113**	**125**	**137**	**149**	**161**	**173**

Với số liệu trên ta có nhận xét, tương ứng với mỗi mức thu nhập các hộ gia đình sẽ có phản ứng khác nhau với mức thu nhập đó. Việc tìm ra mối liên hệ đó quả là khó khăn, người ta đã đưa ra khái niệm kỳ vọng có điều kiện để xây dựng mối liên hệ này. Tức là người ta xây dựng nên mối liên hệ giữa thu nhập ảnh hưởng như thế nào đến mức chi tiêu trung bình.

Nếu ta biểu diễn các điểm (X, $E(Y/X_i)$ lên hệ toạ độ, ta có độ thị sau:



Với kết quả trên chúng ta nhận thấy rằng trung bình có điều kiện của Y sẽ phụ thuộc vào các giá trị của X, do đó ta có thể biểu diễn như sau:

$E(Y/X_i) = f(X_i)$

Hàm số này được gọi là hàm hồi qui tổng thể. Như vậy hồi qui là thể hiện mối quan hệ trung bình của Y phụ thuộc vào X.

- Một vài khái niệm cần chú ý:

- Nếu $f(X_i)$ là một hàm tuyến tính thì ta gọi là Hàm hồi qui tuyến tính:

$E(Y/X_i) = f(X_i) = \alpha + \beta X_i$ hoặc

$Y = f(X_i) + U_i = \alpha + \beta X_i + U_i$

- Nếu f là hàm 1 biến thì ta gọi là hàm Hồi qui tuyến tính đơn giản, nếu f là hàm nhiều biến thì ta gọi là hàm Hồi qui tuyến tính bội.

- X, Y: được gọi là biến. Trong đó:

X được gọi là biến giải thích (độc lập).

Y: biến được giải thích (phụ thuộc).

- α, β: được gọi là tham số của hồi qui. Trong đó:

α: được gọi là tham số tự do hay tham số chặn.

β: được gọi là tham số của biến.

- U_i: là biến ngẫu nhiên và còn gọi là yếu tố ngẫu nhiên (nhiễu). Thành phần yếu tố ngẫu nhiên có thể bao gồm:

. Các biến giải thích bị bỏ sót hay là các yếu tố khác mà ta chưa xem xét.

. Sai số khi đo lường biến phụ thuộc.

. Tính ngẫu nhiên vốn có của biến phụ thuộc.

- X,Y không có mối quan hệ hàm số mà là mối quan hệ nhân quả và thống kê, trong đó X là nguyên nhân và Y là kết quả.

- X,Y không có mối quan hệ hàm số mà là mối quan hệ thống kê, có nghĩa là tương ứng với mỗi giá trị của X ta có ngẫu nhiên giá trị của Y.

- Hồi qui tuyến tính được hiểu là hồi qui tuyến tính theo tham số, ta đang xem xét trường hợp đặc biệt vừa tuyến tính với biến, vừa tuyến tính với tham số.

2. Phương trình hồi qui tuyến tính mẫu

Để mô toán học hoá mối liên hệ giữa X và Y tức là ta phải tìm được giá trị của tham số hồi qui, và ta cũng chỉ có thể thực hiện được điều này thông qua các quan sát mẫu.

Giả sử (x_i,y_i) là mẫu n cặp quan sát thu thập ngẫu nhiên từ X và Y. Ta mong muốn tìm giá trị a, b để ước lượng cho các tham số α, β. Nói cách khác, ta mong muốn tìm một đường thẳng $y = a + bx$ "thích hợp" nhất đối với các giá trị (x_i,y_i). Đường thẳng $\hat{y} = a + bx$ được xem là "thích hợp" nhất khi tổng bình phương các chênh lệch giữa giá trị thực tế y_i với $\hat{y}_i$ là nhỏ nhất. Sau đây là phương pháp bình phương bé nhất:

$$f(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min$$

Từ điều kiện trên, ta xác định các hệ số a, b như sau:

$$\begin{cases} \dfrac{\partial f(a,b)}{\partial a} = 0 \\ \dfrac{\partial f(a,b)}{\partial b} = 0 \end{cases}$$

Giải hệ phương trình ta tìm được a, b

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n.\bar{x}.\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n.\bar{x}^2}}$$

$$a = \bar{y} + b\bar{x}$$

Đường thẳng $\hat{y}_i = a + bx_i$ được gọi là đường hồi qui tuyến tính mẫu.

- Một số tính chất của phương pháp bình phương bé nhất:

(1) a, b được xác định là duy nhất tương ứng với mẫu.

(2) a, b là ước lượng điểm của α, β.

- $\hat{y}_i = a + bx_i$ có các tính chất sau:

(1) Đi qua trung bình mẫu ($\bar{x}$, $\bar{y}$)

(2) Trung bình của $\hat{y}_i$ bằng trung bình của quan sát.

(3) Trung bình của các phần dư bằng 0: $\sum e_i = 0$

(4) Các phần dư không tương quan với $\hat{y}_i$: $\sum \hat{y}_i e_i = 0$

(5) Các phần dư không tương quan với X_i: $\sum x_i e_i = 0$

- Với kết quả bình phương bé nhất chúng ta chỉ mới ước lượng điểm của tham số hồi qui, chúng ta không biết được chất lượng của các ương lượng này như thế nào, chất lượng ước lượng phụ thuộc vào nhiều nội dung như:

- Dạng hàm của mô hình hồi qui được lựa chọn.

- Phụ thuộc vào X_i và U_i.

- Phụ thuộc vào kích thước mẫu.

Ở đây ta sẽ nói đến các giả thuyết mang tính chất toán học đối với X và U để đảm a, b bằng phương pháp bình phương nhỏ nhất là các ước lượng tuyến tính, không chệch, có phương sai nhỏ nhất.

Giả thuyết 1: Kỳ vọng của yếu tố ngẫu nhiên bằng 0:

$E(u_i)=E(u/X_i)0$.

Giả thuyết 2: Phương sai của các yếu tố ngẫu nhiên không đổi:

$Var(u/X_i)=Var(u/X_j)=\sigma^2$, với mọi $i \neq j$

Giả thuyết 3: Không có sự tương quan giữa các u.

Giả thuyết 4: u và X không có sự tương quan.

Giả thuyết 5: Các biến giải thích là các biến phi ngẫu nhiên, tức là các giá trị của nó đã được xác định sẵn.

Các giả thuyết ở đây chỉ mang tính chất giới thiệu, trong phần kinh tế lượng sẽ giải quyết vấn đề này một cách cụ thể hơn. Ở đây ta giả sử rằng các điều kiện trên đã được thoả mãn.

3. Khoảng tin cậy của các hệ số hồi qui

- Ước lượng khoảng của β với độ tin cậy (1-α)100% là:

$$b - t_{n-2,\alpha/2}S_b < \beta < b + t_{n-2,\alpha/2}S_b$$

Trong đó:

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

$$S_e^2 = \frac{\sum_{i=1}^{n}e_i^2}{n-2} = \frac{SSE}{n-2}$$

- Ước lượng khoảng của α với độ tin cậy (1-α)100% là:

$$a - t_{n-2,\alpha/2}S_a < \alpha < a + t_{n-2,\alpha/2}S_a$$

Trong đó: $S_a^2 = S_e^2.(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2})$

4. Kiểm định tham số hồi qui tổng thể (β)

Kiểm định về mối quan hệ giữa X và Y. Trường hợp β=0 thì X và Y không có mối quan hệ nào, trường hợp β>0 (β<0) giữa X và Y có mối quan hệ thuận (nghịch).

Ở mức ý nghĩa α, giả thuyết H_0 kiểm định ở các trường hợp sau:

Giả thuyết	$\begin{cases}H_0 : \beta \le 0 \\ H_0 : \beta > 0\end{cases}$	$\begin{cases}H_0 : \beta \ge 0 \\ H_0 : \beta < 0\end{cases}$	$\begin{cases}H_0 : \beta = 0 \\ H_0 : \beta \ne 0\end{cases}$
Giá trị kiểm định	$t = \frac{b}{S_b}$		
Bác bỏ H_0	$t>t_{n-2,\alpha}$	$T<-t_{n-2,\alpha}$	$t>t_{n-2,\alpha/2}; t<-t_{n-2,\alpha/2}$

2.5. Phân tích phương sai hồi qui:

a) Hệ số xác định: R^2 là hệ số nhằm xác định mức độ quan hệ giữa X và Y có quan hệ hay không, hoặc bao nhiêu phần trăm sự biến thiên của Y có thể giải thích bởi sự phụ thuộc tuyến tính của Y vào X.

Giá trị thực tế $y_i = a + bx_i + e_i$

Giá trị dự đoán theo phương trình hồi qui: $\hat{y} = a + bx_i$

$$\Rightarrow y_i = \hat{y}_i + e_i$$

Vậy e_i là sự khác biệt giữa giá trị thực tế với giá trị dự đoán của phương trình hồi qui tuyến tính. Như vậy, e_i thể hiện phần biến thiên của Y không thể giải thích bởi mối liên hệ tuyến tính giữa Y và X.

Ta có:

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum e_i^2$$

hay SST = SSR + SSE

* SSR càng lớn thì mô hình hồi qui tuyến tính càng có độ tin cậy cao trong việc giải thích sự biến động của Y.

* Hệ số xác định $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ là phần trăm biến động của Y được giải thích bởi mối quan hệ tuyến tính của Y vào X.

b) Phân tích phương sai:

Kiểm định giả thuyết về mối quan hệ tuyến tính giữa X và Y, tức là giả thuyết H0: có sự tồn tại mối liên hệ tuyến tính giữa X và Y.

Bảng ANOVA trong phân tích hồi qui tuyến tính đơn giản:

Biến thiên	Tổng độ lệch bình phương	Bậc tự do	Phương sai	Giá trị kiểm định F
Hồi qui	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Sai số	SSE	n-2	$MSE = \frac{SSE}{n-2}$	
Tổng cộng	SST	n-1		

Qui tắc bác bỏ giả thuyết H_0: $F > F_{1,n-2}$,(trong đó $F_{1,n-2}$ có phân phối F).

Ví dụ 6.2: Nghiên cứu mối quan hệ giữa số tiền chi tiêu trong hộ gia đình trong một năm với thu nhập của họ như thế nào. Ta có số liệu sau:

Thu nhập X	**Chi tiêu Y**	**Thu nhập X**	**Chi tiêu Y**
9.098	5.492	11.307	5.907
9.138	5.540	11.432	6.124
9.094	5.305	11.449	6.186
9.282	5.507	11.697	6.224
9.229	5.418	11.871	6.496
9.347	5.320	12.018	6.718
9.525	5.538	12.523	6.921
9.756	5.692	12.053	6.471
10.282	5.871	12.088	6.394
10.662	6.157	12.215	6.555
11.019	6.342	12.494	6.755

- Thực hiện trên Excel để xử lý: Các bước thực hiện như sau:

(1). Tools → Data Analysis → Regression.

(2). Nhập dữ liệu:

- Nhập số liệu theo cột, mỗi cột một biến.

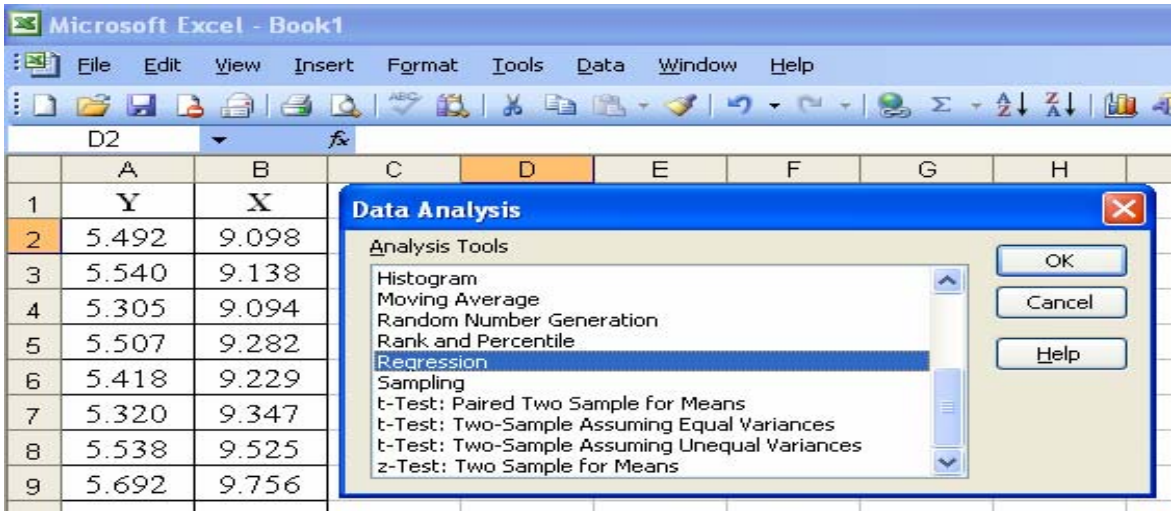
- Input Y Range: Chọn vùng xử lý của biến phụ thuộc.

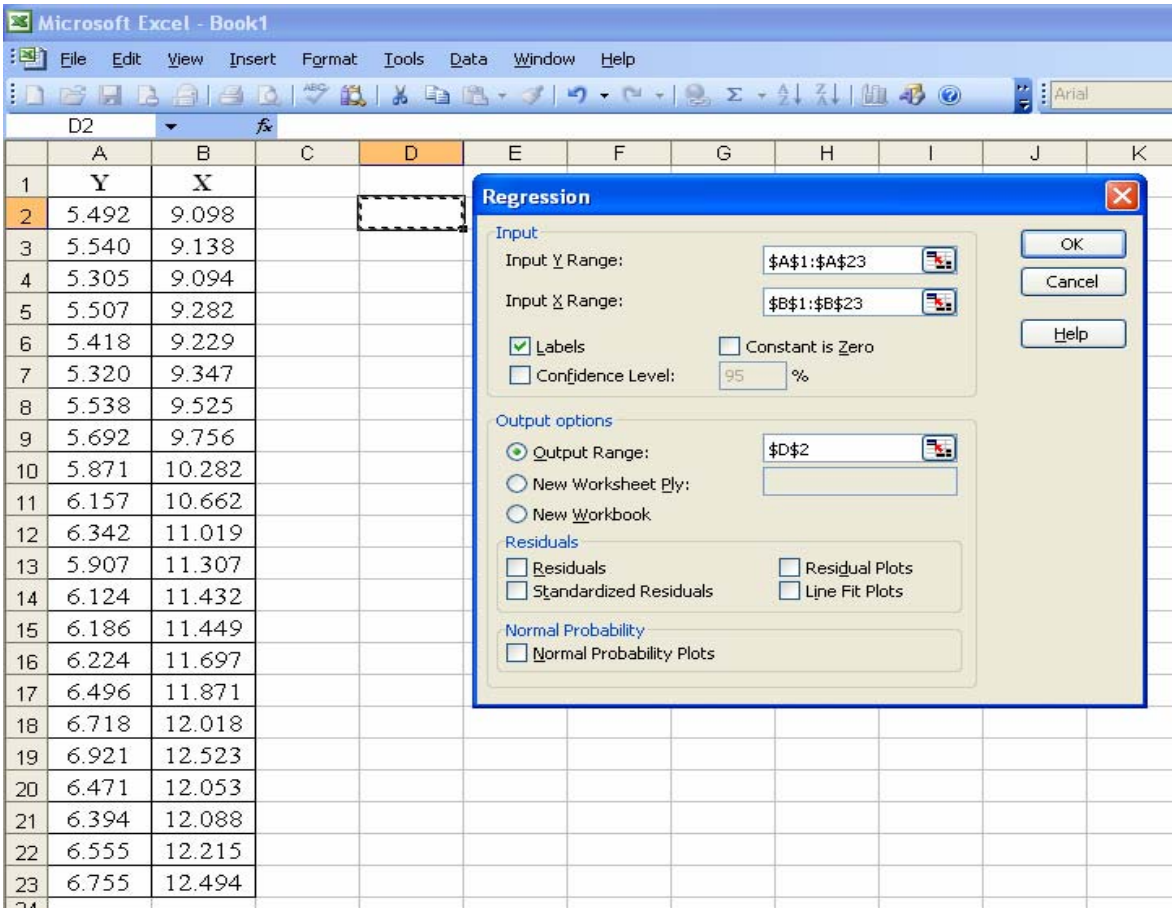
- Input X Range: Chọn vùng xử lý của biến độc lập, nếu nhiều biến thì chọn nhiều cột.

- Labels: Vùng xử lý có tên biến không.

- Constant is Zero: Đây là trường hợp hồi với với $\alpha=0$.

- Confidence Level: Độ tin cậy.





	A	B
1	Y	X
2	5.492	9.098
3	5.540	9.138
4	5.305	9.094
5	5.507	9.282
6	5.418	9.229
7	5.320	9.347
8	5.538	9.525
9	5.692	9.756
10	5.871	10.282
11	6.157	10.662
12	6.342	11.019
13	5.907	11.307
14	6.124	11.432
15	6.186	11.449
16	6.224	11.697
17	6.496	11.871
18	6.718	12.018
19	6.921	12.523
20	6.471	12.053
21	6.394	12.088
22	6.555	12.215
23	6.755	12.494

Kết quả xử lý của Excel như sau:

Regression Statistics	
Multiple R	0,9587488
R Square	0,9191993
Adjusted R Square	0,9151592
Standard Error	147,66972
Observations	22

r = 0,9587 thể hiện mối tương quan thuận giữa thu nhập và chi tiêu là chặt chẽ.

R^2 = 0,9191 có nghĩa là 91,91% sự biến thiên của chi tiêu có thể được giải thích từ mối liên hệ tuyến tính giữa chi tiêu với thu nhập.

ANOVA

	df	***SS***	***MS***	***F***	***Significance F***
Regression	1	4961434,406	4.961.434,41	227,52	2,1713E-12
Residual	20	436126,9127	21.806,35		
Total	21	5397561,318			

Vì giá trị p quá nhỏ (Sig.F=2,1713E-12), do đó ta có thể kết luận giữa X và Y có mối quan hệ tuyến tính.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 90,0%*	*Upper 90,0%*
Intercept	1922,39	274,9493	6,991	9E-07	1448,18	2396,60
X	0,3815	0,025293	15,08	2E-12	0,33789	0,42514

Phương trình hồi qui tuyến tính giữa X và Y:

$$\hat{y} = 1922,39 + 0,3815x$$

Dựa vào phương trình này ta biết được mức độ về mối liên hệ giữa chi tiêu và thu nhập dựa vào hệ số a và b.

- Hệ số b: dựa vào dấu của hệ số này ta có thể biết được giữa X và Y có mối liên hệ thuận hay nghịch và dựa và độ lớn của b ta biết được cường độ hay là mức độ ảnh hưởng của X đến Y lớn hay nhỏ.

- Hệ số a: Đây là hệ số khá phức tạp trong việc giải thích, nó phụ thuộc vào bản chất của nội dung nghiên cứu, phụ thuộc vào mô hình và trong một số trường hợp ta không thể giải thích được hoặc nếu có giải thích chỉ là đơn giản khi X=0 thì Y=a (chính điều này mà người ta còn gọi a là hệ số chặn). Do đó, tuỳ từng trường hợp mà ta các cách giải thích cụ thể.

6. Dự báo trong phương pháp hồi qui tuyến tính đơn giản

- Ước lượng khoảng giá trị thực của y_{n+1} với độ tin cậy (1-α):

$$\hat{y} \pm t_{n-2,\alpha/2} S_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}$$

- Ước lượng khoảng giá trị trung bình của yn+1 với độ tin cậy (1-α):

$$\hat{y} \pm t_{n-2,\alpha/2} S_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}$$

7. Mở rộng mô hình hồi qui 2 biến

7.1. Mô hình tuyến tính logarit:

Trong sản xuất, người ta thường xem xét lượng đầu ra phụ thuộc vào các yếu tố đầu vào và người ta thường dùng mô hình hồi qui mũ như sau:

$$Y = \alpha X^{\beta} e^{U}$$

Ta có thể viết lại như sau:

$$\ln Y = \ln\alpha + \beta \ln X + U$$

Nếu đặt $Y^* = \ln Y_i$, $\alpha^* = \ln\alpha$, $X^* = \ln X_i$ khi đó ta có thể viết lại dưới dạng tuyến tính đơn giản:

$$Y^* = \alpha^* + \beta X^* + U$$

Mô hình này được gọi tên là log – log hoặc log kép. Với hệ số co giản chính là β.

7.2. Mô hình bán logarit (similog) :

Trong lý thuyết tài chính tiền tệ và ngân hàng người ta sử dụng công thức tính lãi suất gộp:

$$Y = Y_0(1+r)^t$$

Ta viết lại dưới dạng logarit:

$$\ln Y = \ln Y_0 + t\ln(1+r)$$

Nếu đặt $\alpha = \ln Y_0$, $\beta = \ln(1+r)$, ta có thể viết lại dưới dạng:

$$\ln Y = \alpha + \beta t + U$$

7.3. Một số mô hình khác:

- Mô hình lin – log: $Y = \alpha + \beta \ln X + U$
- Mô hình nghịch đảo: $Y = \alpha + \beta(1/X) + U$

III. HỒI QUI TUYẾN TÍNH BỘI

1. Mô hình hồi bội

Giả sử Y phụ thuộc vào vào k biến độc lập $X_1....X_k$. Nếu giá trị của của k biến độc lập $X_1....X_k$ mô hình hồi qui bội dưới dạng tuyến tính sau:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + U$$

β_j : được gọi là các hệ số hồi qui riêng, thể hiện mức độ biến thiên Y khi biến X_j thay đổi một đơn vị, các biến còn lại không đổi.

U: là sai số. Tương tự như đối với hồi qui đơn giản.

2. Phương trình hồi qui bội của mẫu

Gọi các hệ số a, $b_1,....b_k$ ước lượng cho α, β_1, ... β_k được xác định bởi phương pháp bình phương bé nhất:

$$f = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - ... - b_k x_{ki})^2 \rightarrow \min$$

Từ điều kiện trên ta có hệ:

$$\begin{cases} \partial f / \partial a = 0 \\ \partial f / \partial b_1 = 0 \\ ... \\ \partial f / \partial b_K = 0 \end{cases}$$

Giải hệ phương trình ta sẽ tìm được nghiệm (a, b_1,..., b_k)

Phương trình $\hat{y} = a + b_1x_1 + ... + b_kx_k$ được gọi là phương trình hồi qui bội của mẫu.

Chúng ta cũng có thể tìm được nghiệm (a, b_1,..., b_k) bằng phương pháp ma trận, tuy nhiên dù phương pháp nào đi nữa thì việc tìm nghiệm bằng phương pháp thủ công là rất phức tạp. Với công nghệ máy tính phát triển, các phần mền thống kê được phát triển thì việc tìm nghiệm trở nên dễ dàng hơn. Chính vì vậy, chúng ta không nên quá quan tâm đến việc tìm nghiệp bằng phương pháp thủ công như thế nào.

Tương tự như đối với hồi qui tuyến tính đơn giản, phương pháp bình phương bé nhất phải thoả mãn 5 điều kiện, ngoài ra còn phải thoả mãn thêm điều kiện:

- U có phân phối chuẩn $N(0,\sigma^2)$

- Các biến X_j độc lập với nhau

3. Khoảng tin cậy của các hệ số hồi qui

Mô hình hồi qui bội có dạng:

$$Y = \alpha + \beta_1X_1+\beta_2X_2+....+\beta_kX_k+U$$

Tương tự như đối với hồi qui đơn giản, ước lượng khoảng của các hệ số như sau:

- Ước lượng khoảng của (i với độ tin cậy $(1-\alpha)100\%$ là:

$$b_i - t_{n-k-1,\alpha/2}S_{bi} < \beta_i < b_i + t_{n-k-1,\alpha/2}S_{bi}$$

- Ước lượng khoảng của (với độ tin cậy $(1-\alpha)100\%$ là:

$$a - t_{n-k-1,\alpha/2}S_a < \alpha < a + t_{n-k-1,\alpha/2}S_a$$

4. Kiểm định từng tham số hồi qui tổng thể (βi)

Tương tự như đối với kiểm định của hồi qui đơn giản.

Trường hợp $\beta_i=0$ thì X_i và Y không có mối quan hệ nào, trường hợp $\beta_i>0$ ($\beta_i<0$) giữa X_i và Y có mối quan hệ thuận (nghịch).

Ở mức ý nghĩa α, giả thuyết H_0 kiểm định ở các trường hợp sau:

Giả thuyết	$\begin{cases} H_0 : \beta_i \le 0 \\ H_0 : \beta_i > 0 \end{cases}$	$\begin{cases} H_0 : \beta_i \ge 0 \\ H_0 : \beta_i < 0 \end{cases}$	$\begin{cases} H_0 : \beta_i = 0 \\ H_0 : \beta_i \ne 0 \end{cases}$
Giá trị kiểm định	$t = \frac{b_i}{S_{bi}}$		
Bác bỏ H_0	$t>t_{n-k-1,\alpha}$	$t<-t_{n-k-1,\alpha}$	$t>t_{n-k-1,\alpha/2}$; $t<-t_{n-k-1,\alpha/2}$

Đây là một phương pháp xây dựng mô hình hồi qui, được gọi là phương pháp loại biến dần. Chúng ta sẽ loại từng biến một dựa vào giá trị p kiểm định lớn ra trước.

5. Phân tích phương sai hồi qui

a) Hệ số xác định:

Tương tự như đối với Hồi qui đơn giản, ta có

* Hệ số xác định $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Nhưng ở đây, hệ số R^2 là nói lên tính chặt chẽ giữa biến phụ thuộc Y và các biến độc lập X_i, tức là nó thể hiện phần trăm biến thiên của Y có thể được giải thích bởi sự biến thiên của tất cả các biến X_i.

Đối với người nghiên cứu thì họ mong muốn hệ số R^2 càng lớn càng tốt, tuy nhiên R^2 là một hàm không giảm theo số lượng biến đưa vào. Điều này có thể dẫn đến một trò chơi về số R^2 bằng cách đưa vào mô hình càng nhiều biến để có hệ số R^2 lớn. Để khắc phục nhược điểm này, người ta đưa ra hệ số xác định điều chỉnh đánh giá mức độ phụ thuộc của Y vào các biến X chính xác hơn.

* Hệ số xác định đã điều chỉnh:

$$\overline{R}^2 = \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - (1-R^2)(\frac{n-1}{n-k-1})$$

Xét về mặt ý nghĩa thì giữa R^2 và $\overline{R}^2$ là như nhau, thông thường thì hai hệ số này chênh lệch nhau không nhiều. Trong một số trường hợp số lượng biến X tương đối lớn so với n, khi đó ta nên dùng hệ số xác định có điều chỉnh để đo lường mức độ thích hợp của mô hình hồi qui bội.

Đây cũng là một phương pháp xây dựng mô hình hồi qui, được gọi là phương pháp đưa biến vào dần. Chúng ta sẽ đưa lần lượt các biến có trị tuyệt đối hệ số tương quan $r_{y,xi}$ lớn vào trước, nếu $\overline{R}^2$ tăng lên thì ta chấp nhận biến, còn ngược lại thì ta loại ra và kết thúc quá trình.

b) Phân tích ANOVA hồi qui bội:

Đặt giả thuyết: H_0: $\beta_1 = \beta_2 = .. = \beta_k$

H_1: Không phải tất cả $\beta_i = 0$

Bảng ANOVA trong phân tích hồi qui tuyến tính bội:

Biến thiên	Tổng độ lệch bình phương	Bậc tự do	Phương sai	Giá trị kiểm định F
Hồi qui	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Sai số	SSE	n-(k+1)	$MSE = \frac{SSE}{n-(k+1)}$	
Tổng cộng	SST	n-1		

Qui tắc bác bỏ giả thuyết H_0: $F > F_{k,n-k-1,\alpha}$, trong đó $F_{k,n-k-1}$, có phân phối F.

* Trong trường hợp ta đã có hệ số xác định R^2 giá trị kiểm định được tính bởi công thức sau:

$$F = \frac{n-k-1}{k} . \frac{R^2}{1-R^2}$$

Ví dụ 6.2: Tốc độ phát triển nền kinh tế (Y) phụ thuộc vào tốc độ phát triển của nông nghiệp (X_1), tốc độ tăng trưởng của kim ngạch xuất khẩu (X_2) và tỷ lệ lạm phát (X_3) được thu thập ở 48 nước dưới đây:

Y	NN	XK	LP	Y	NN	XK	LP
1,3	3,4	-2,7	13,0	4,1	2,3	8,7	9,5
1,0	1,4	-6,0	10,5	-5,0	1,2	-2,0	1,1
0,4	0,1	-3,6	15,9	2,1	2,7	5,6	11,2
4,9	1,8	13,6	3,2	7,7	3,0	2,0	8,9
9,8	5,6	27,3	5,4	9,3	3,3	6,2	7,5
-2,1	2,2	2,6	5,2	-1,7	2,0	-1,7	18,2
2,0	2,3	-9,5	8,7	5,8	4,7	-0,2	2,1
5,8	3,0	4,4	1,4	3,9	-3,9	-2,5	3,4
5,2	2,9	9,2	3,0	5,6	3,9	6,4	13,9
-1,1	-2,3	-6,3	14,9	6,9	1,3	11,6	6,4
0,2	0,3	12,0	20,3	-4,6	0,8	-9,8	21,5
1,1	1,4	-7,2	19,8	-2,6	1,7	-6,6	6,7
-12,0	4,8	-5,5	8,6	1,1	3,9	3,8	7,7
-1,6	-0,4	-2,5	11,3	4,6	3,0	-3,5	8,6
0,5	1,9	1,6	19,0	-0,6	2,5	2,0	11,5
2,2	-3,5	4,7	1,9	8,2	1,9	3,8	7,8
8,0	3,1	10,9	37,3	4,1	0,9	1,3	5,6
6,5	3,3	-0,6	8,9	12,6	7,9	11,7	3,8
0,2	0,1	8,4	29,5	4,1	2,8	-0,9	9,9
7,8	5,3	10,4	8,1	0,6	2,8	-2,1	23,3
2,5	2,3	4,9	22,6	2,0	0,5	-3,1	33,5
-0,2	3,1	7,9	20,2	0,0	0,4	6,9	32,6
6,1	10,3	-19,0	-1,3	-2,6	-1,3	3,4	7,7
2,9	-0,6	5,4	7,5	-3,4	7,9	-7,9	45,4

Kết quả xử lý của Excel như sau:

Regression Statistics	
Multiple R	0,6088296
R Square	0,3706735
Adjusted R Square	0,3277648
Standard Error	3,6899289
Observations	48

R^2= 0,37 có nghĩa là 37% sự biến thiên của tốc độ phát triển kinh tế có thể được giải thích từ mối liên hệ tuyến tính giữa tốc độ phát triển kinh tế với tốc độ biến thiên của nông nghiệp, xuất khẩu và lạm phát.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	352,8613631	117,6205	8,63867	0,000127635
Residual	44	599,0853036	13,61558		
Total	47	951,9466667			

Với giá trị p=0,0001 là rất nhỏ, ta có thể bác bỏ giả thuyết H_0, có nghĩa là có tồn tại mối liên hệ tuyến tính giữa tốc độ phát triển kinh tế với ít nhất một trong các yếu tố: nông nghiệp, xuất khẩu và lạm phát.

	Coeffi-cients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2,033019	0,993121	2,0471	0,046649	0,031514	4,034525
NN	0,500738	0,205653	2,434859	0,019021	0,086269	0,915206
XK	0,268085	0,068954	3,887867	0,000337	0,129117	0,407054
LP	-0,10474	0,053080	-1,97337	0,054755	-0,21172	0,002229

Phương trình hồi qui bội:

$$\hat{y} = 2,033 + 0,5007x_i + 0,268x_2 - 0,1047x_3$$

Từ phương trình hồi qui bội ta có nhận xét:

- Nếu tốc độ tăng trưởng của xuất khẩu và lạm phát không đổi, 1% tăng trưởng nông nghiệp sẽ làm tăng 0,5007% tăng trưởng của nền kinh tế.

- Nếu tốc độ tăng trưởng của nông nghiệp và lạm phát không đổi, 1% tăng trưởng xuất khẩu sẽ làm tăng 0,268% tăng trưởng của nền kinh tế.

- Nếu tốc độ tăng trưởng của nông nghiệp và xuất khẩu không đổi, tỷ lệ lạm phát tăng 1% sẽ làm cho nền kinh tế giảm 0,1047%.

- Nếu tốc độ tăng trưởng của nông nghiệp, xuất khẩu, lạm phát bằng 0 thì nền nền kinh tế tăng trưởng 2,033%.

Tóm lại, Hồi qui tuyến tính là một nội dung rất rộng và sâu, được ứng dụng trong nhiều lĩnh vực. Tuy nhiên để hiểu một cách đầy đủ thì trong phạm vi của môn học không thể trình bày đầy đủ được mà sự mở rộng này thuộc phạm vi của một môn học khác đó là Kinh tế lượng. Chính vì vậy, ở đây chỉ giới thiệu mang tính chất nhập môn không đi sâu về mặt lý thuyết và kỹ thuật.

CHƯƠNG VII
DÃY SỐ THỜI GIAN

Trong chương này sẽ nói đến phương pháp phân tích biến động của hiện tượng qua thời gian. Như trong chương trước, chúng ta sử dụng mô hình hồi qui tuyến định để dự báo, phương pháp này còn được gọi là phương pháp dự báo dự vào nội suy, có nghĩa là ta dự báo dựa vào bản chất của hiện tượng. Tuy nhiên, đối với phương pháp này chúng ta không thể nào đưa tất cả các yếu tố ảnh hưởng đến hiện tượng vào mô hình được bởi nó quá nhiều và cũng không thể biết hết. Phương pháp dự báo dự vào dãy số thời gian là chúng ta quan sát hiện tượng biến đổi qua thời gian rồi tìm ra qui luật và dùng qui luật đó để suy luận, phương này gọi là phương pháp dự báo dựa vào ngoại suy. Trong thực tế có rất nhiều hiện tượng phụ thuộc vào thời gian như: Lượng tiêu thụ lương thực thực phẩm phụ thuộc vào độ tuổi, chu kỳ sống của sản phẩm,... với lý luận như vậy ta có thể xem thời gian như là một biến độc lập tác động đến hiện tượng nghiên cứu.

Như vậy, ta có thể xem nghiên cứu dãy số thời gian như là chúng ta có thêm một phương án lựa chọn để dự báo.

I. DÃY SỐ THỜI GIAN

1. Định nghĩa

Dãy số thời gian là một dãy các giá trị của hiện tượng nghiên cứu được sắp.

t_i	t_1	t_2		t_n
y_i	y_1	y_2		y_n

2. Phân loại

Căn cứ vào đặc điểm thời gian người ta thường chia dãy số thời gian thành hai loại:

- Dãy số thời kỳ: là dãy số biểu hiện sự thay đổi của hiện tượng qua từng thời kỳ nhất định. Ví dụ, giá trị hàng xuất khẩu của một quốc gia vào các năm từ 1990 đến 1995.

- Dãy số thời điểm: là dãy số biểu hiện mặt lượng của hiện tượng vào một thời điểm nhất định. Ví dụ, tổng giá trị tài sản của doanh nghiệp vào các thời điểm cuối năm 31/12/19xx.

3 Phương pháp luận dự báo thống kê

Để xây dựng một mô hình dự báo thì người nghiên cứu cần thu thập số liệu về vấn đề cần dự báo. Phương pháp thu thập dữ liệu và tiến hành dự báo phụ thuộc vào nhiều nhân tố được mô tả ở sơ đồ sau:

	----------Dự liệu lịch sử----------		
	Y_1------Mẫu------- Y_n	Y_{n+1}----------Y_N	
Dự báo lùi	Dự báo trong mẫu	Dự báo hậu nghiệm	Dự báo tiền nghiệm
$\hat{Y}_{1-m}$ -------- $\hat{Y}_{1-1}$	$\hat{Y}_1$ ---------- $\hat{Y}_n$	$\hat{Y}_{n+1}$ ---------- $\hat{Y}_N$	$\hat{Y}_{N+1}$ ---------- $\hat{Y}_{N+k}$

- Dự liệu lịch sử: là dữ liệu mới nhất của chuỗi thời gian thu thập được
- Mẫu: Dữ liệu dùng để xây dựng mô hình
- Giai đoạn dự báo được chia thành dự báo hậu nghiệm và dự báo tiền nghiệm.

- Dự báo hậu nghiệm, đặc trưng quan trọng của nó là đã có các giá trị quan sát thực tế của đối tượng dự báo, nó cho phép các nhà nghiên cứu đánh giá được độ chính xác của mô hình.

- Dự báo tiền nghiệm: các giá trị thực tế không có do đó không xác định được độ chính xác của những dự báo tiền nghiệm.

- Dự báo lùi: chúng ta cũng có thể dự báo lùi cho những thời kỳ trước. Dự báo lùi nhằm tạo ra các giá trị bổ sung cho dãy số lịch sử trong quá trình phân tích.

4. Đo lượng độ chính xác của dự báo

Sai số dự báo là thước đo phản ánh giá trị dự báo gần với giá trị thực tế bao nhiêu. Sai số dự báo là chênh lệch giữa giá trị dự báo và giá trị thực tế tương ứng:

$$e_i = y_i - \hat{y}_i$$

- Sai số tuyệt đối trung bình: $MAE = \frac{\sum_{t=1}^{n}|e_t|}{n}$
- Phần trăm tuyệt đối: $MAPE = \frac{\sum_{t=1}^{n}\frac{|e_t|}{Y_t}}{n}$
- Phương sai: $MSE = \frac{\sum_{t=1}^{n}e_t^2}{n}$
- Độ lệch chuẩn: $RMSE = \sqrt{MSE}$

5. Sự lựa chọn công thức tính sai số dự báo

1). Nếu dữ liệu có một vài sai số dự báo lớn thì không nên sử dụng MSE.

2). Các sai số xấp xỉ bằng nhau thì nên dùng MSE.

3). Khi có đồng thời MAE, MSE, RMSE thì chọn chỉ tiêu nào có giá trị nhỏ nhất.

II. MỘT SỐ CHỈ TIÊU CƠ BẢN VỀ DÃY SỐ THỜI GIAN

1. Mức độ trung bình theo thời gian

1.1. Mức độ trung bình của dãy số thời kỳ:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

1.2. Mức độ trung bình của dãy số thời điểm:

- Nếu khoảng cách giữa các thời điểm bằng nhau:

$$\bar{x} = \frac{\frac{1}{2}x_1 + x_2 + ... + \frac{1}{2}x_n}{n}$$

- Nếu khoảng cách giữa các thời điểm không bằng nhau:

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

t_i: độ dài thời gian có mức độ x_i

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 hai thời điểm nghiên cứu.

2.1. Số tuyệt đối từng kỳ (liên hoàn): Biểu hiện lượng tăng giảm tuyệt đối giữa hai thời kỳ kế tiếp nhau.

$$\Delta_i = x_i - x_{i-1}$$

2.2. Số tuyệt đối định gốc: Biểu hiện lượng tăng giảm tuyệt đối giữa kỳ nghiên cứu và kỳ được chọn làm gốc.

$$\Delta'_i = x_i - x_{(1)} \qquad x_{(1)}: \text{kỳ được chọn làm gốc.}$$

* Mối liên hệ giữa số tăng giảm tuyệt đối liên hoàn và định gốc:

$$\sum_{i=2}^{n} \Delta_i = \Delta'_i$$

2.3. Số tuyệt đối trung bình:

$$\overline{\Delta} = \frac{\sum_{i=2}^{n} \Delta_i}{n-1} = \frac{\Delta'_i}{n-1} = \frac{x_i - x_{(1)}}{n-1}$$

3. Tốc độ phát triển (lần, %)

Là chỉ tiêu biểu hiện sự biến động của hiện tượng xét về mặt tỷ lệ.

3.1. Tốc độ phát triển từng kỳ (liên hoàn): Biểu hiện sự biến động về mặt tỷ lệ của hiện tượng nghiên cứu qua hai thời kỳ liên tiếp nhau:

$$t_i = \frac{x_i}{x_{i-1}}$$

3.2. Tốc độ phát triển định gốc: Biểu hiện sự biến động về mặt tỷ lệ của hiện tượng giữa kỳ nghiên cứu với kỳ được chọn làm gốc:

$$t'_i = \frac{x_i}{x_{(1)}}$$

* Mối liên hệ giữa tốc độ phát triển từng kỳ và định gốc:

$$\prod_{i=2}^{n} t_i = t'_n$$

3.3. Tốc độ phát triển trung bình:

$$\overline{t} = \sqrt[n-1]{\prod_{i=2}^{n} t_i} = \sqrt[n-1]{t'_n} = \sqrt[n-1]{\frac{x_n}{x_{(1)}}}$$

4. Tốc độ tăng giảm

Là chỉ tiêu biểu hiện số tăng lên hay giảm xuống về mặt tỷ lệ của hiện tượng nghiên cứu.

4.1. Tốc độ tăng giảm từng kỳ:

$$a_i = \frac{x_i - x_{i-1}}{x_{i-1}} = t_i - 1$$

4.2. Tốc độ tăng giảm định gốc:

$$a'_i = \frac{x_i - x_{(1)}}{x_{(1)}} = t'_i - 1$$

4.3. Tốc độ tăng giảm trung bình:

$$\bar{a} = \bar{t} - 1$$

5. Giá trị tuyệt đối của 1% tăng giảm

Chỉ tiêu này biểu hiện mối quan hệ giữa chỉ tiêu lượng tăng giảm tuyệt đối và chỉ tiêu tiêu tốc độ tăng giảm. Nghĩa là xem xét 1% tăng giảm của hiện tượng sẽ tương ứng với một lượng giá trị tuyệt đối tăng giảm là bao nhiêu.

$$g_i = \frac{\Delta_i}{a_i} = \frac{\Delta_i}{\frac{\Delta_i}{x_{i-1}} x100} = \frac{x_{i-1}}{100}$$

III. MỘT SỐ MÔ HÌNH DỰ BÁO

1. Dự đoán dựa vào lượng tăng giảm tuyệt đối trung bình

Phương pháp này thường được sử dụng khi hiện tượng biến động với một lượng tuyệt đối tương đối đều, nghĩa là các lượng tăng giảm tuyệt đối từng kỳ xấp xỉ bằng nhau.

Công thức dự đoán:

$$\hat{y}_{n+L} = y_n + L.\bar{\Delta}$$

$\hat{y}_{n+L}$: Giá trị dự đoán ở thời điểm n+L

y_n : Giá trị thực tế thời điểm n

$\bar{\Delta}$: Lượng tăng giảm tuyệt đối trung bình

L : Tầm xa dự đoán

Ví dụ 7.1: Giá trị xuất khẩu mặt hàng X của quốc gia trong các năm như sau:

Năm	2002	2003	2004	2005	2006	2007
Giá trị xuất khẩu (tỷ đồng)	2,0	2,2	1,7	1,5	2,8	2,9

Dự đoán giá trị xuất khẩu năm 2008 và 2009 dựa vào lượng tăng giảm tuyệt đối trung bình.

$$\bar{\Delta} = \frac{\sum_{i=2}^{n} \Delta_i}{n-1} = \frac{0,9}{6-1} = 0,18$$

* Dự đoán giá trị xuất khẩu năm 2008:

$$\hat{y}_{2008} = y_{2007} + 1.\bar{\Delta} = 2,9 + 1.0,18 = 3,08$$

* Dự đoán giá trị xuất khẩu năm 2009:

$$\hat{y}_{2009} = y_{2007} + 2.\bar{\Delta} = 2,9 + 2.0,18 = 3,26$$

2. Dự đoán dựa vào tốc độ phát triển trung bình

Phương pháp này thường được sử dụng khi hiện tượng biến động với một nhịp độ tương đối ổn định, nghĩa là tốc độ phát triển từng kỳ xấp xỉ nhau.

Công thức dự đoán:

$$\hat{y}_{n+L} = y_n.(\bar{t})^L$$

$\hat{y}_{n+L}$: Giá trị dự đoán ở thời điểm n+L

y_n : Giá trị thực tế thời điểm n

$\bar{t}$: Tốc độ phát triển trung bình

L : Tầm xa dự đoán

Ví dụ 7.2: Sử dụng số liệu ở ví dụ 7.1 dự đoán giá trị xuất khẩu năm 2008 và 2009 dựa vào tốc độ phát triển trung bình.

$$\bar{t} = \sqrt[n-1]{\frac{x_n}{x_{(1)}}} = \sqrt[6-1]{\frac{2,9}{2,0}} = 1,0771$$

- Dự đoán tốc độ phát triển năm 2008:

$$\hat{y}_{2008} = y_{2007}.(\bar{t})^1 = 2,9.1,0071 = 3,1235$$

- Dự đoán tốc độ phát triển năm 2009:

$$\hat{y}_{2009} = y_{2007}.(\bar{t})^2 = 2,9.(1,0071)^2 = 3,3644$$

3. Phương pháp làm phẳng số mũ đơn giản

Phương pháp này dùng để dự đoán dãy số thời gian không có xu hướng hoặc tính thời vụ rõ rệt.

Bước 1: Làm phẳng dãy số bằng công thức:

$$\bar{y}_t = (1-\alpha)y_t + \alpha(1-\alpha)y_{t-1} + \alpha^2(1-\alpha)y_{t-2} + ...$$

$\bar{y}_t$: Giá trị của dãy số đã được làm phẳng ở thời điểm t.

$0 < \alpha < 1$: hằng số làm phẳng.

Công thức có thể viết:

$$\bar{y}_t = \alpha\bar{y}_{t-1} + (1-\alpha)y_t \qquad (*)$$

Bước 2: Công thức dự đoán:

$$\hat{y}_{t+n} = \bar{y}_t \qquad (**)$$

$\hat{y}_{t+n}$: Giá trị dự đoán ở thời điểm t+n

$\bar{y}_t$: Giá trị được làm phẳng ở thời điểm t.

Cách lựa chọn α:

+ Nếu dãy số có nhiều biến ngẫu nhiên thể hiện các mức độ của dãy số lên xuống bất thường: ta chọn α lớn.

+ Nếu dãy số có ít biến ngẫu nhiên: ta chọn α nhỏ.

+ Nếu cùng một tài liệu phải chọn giữa nhiều α, thì ta chọn α có sai số nhỏ nhất.

$$MSE = \frac{\sum_{i=2}^{n}(y_i - \hat{y}_i)^2}{n-1}$$

Ví dụ 7.3: Dự báo số lượng bán trong một ngày, với hệ số làm phẳng α=0,4. Nếu chọn α=0,6, hãy xem xét hệ số làm phẳng nào tốt hơn.

Ngày	Doanh số	Dự đoán $\bar{y}_t = \alpha\bar{y}_{t-1} + (1-\alpha)y_t$		$(y_i - \hat{y}_i)^2$	
		α=0,4	α=0,6	α=0,4	α=0,6
1	925				
2	940	925,00	925,00	225,00	225,00
3	924	934,00	931,00	100,00	49,00
4	925	928,00	928,20	9,00	10,24
5	912	926,20	926,92	201,64	222,61
6	908	917,68	920,95	93,70	167,75
7	910	911,87	915,77	3,50	33,31
8	912	910,75	913,46	1,57	2,14
9	915	911,50	912,88	12,25	4,50
10	924	913,60	913,73	108,16	105,54
11	943	919,84	917,84	536,39	633,23
12	962	933,74	927,90	798,86	1.162,70
13	960	950,69	941,54	86,59	340,74
14	958	956,28	948,92	2,97	82,36
15	955	957,31	952,55	5,34	5,98
16		955,92	953,53		
		TỔNG		**2.184,98**	**3.045,11**

* Dự đoán doanh số ngày 16 với α=0,4

$$\hat{y}_{11} = \bar{y}_{15} = 955{,}92$$

* Nhìn vào kết quả ta có MSE với α=0,4 nhỏ hơn MSE với α=0,6, do đó hệ số làm phẳng α=0,4 tốt hơn.

4. Dự báo bằng hàm xu hướng

Tuỳ theo tính chất của hiện tượng nghiên cứu hoặc kết hợp với kinh nghiệm ta có thể xây dựng hoặc chọn một hàm số phù hợp biểu hiện sự biến động của hiện tượng qua thời gian. Ta có thể tham khảo một số mô hình sau:

Hàm tuyến tính: $\hat{y}_t = b_1 + b_2 t$

Hàm Parabol: $\hat{y}_t = b_1 + b_2 t + b_3 t^2$

Hàm bậc 3: $\hat{y}_t = b_1 + b_2 t + b_3 t^2 + b_4 t^3$

Hàm Hyperbol: $\hat{y}_t = b_1 + \frac{b_2}{t}$

Hàm mũ: $\hat{y}_t = b_1 b_2^t$

....

Công thức dự đoán: $\hat{y}_{n+L} = f(n+L)$

Để tìm được các hệ số b_i chúng ta biến đổi các phương trình trên thành phương trình Hồi qui tuyến tính và sử dụng Excel. Một số tác giả đã trình bày từng trường hợp cụ thể để tìm bằng phương pháp thủ công, tuy nhiên nó có thể không còn phù hợp trong thời buổi công nghệ máy tính vượt trội.

Ví dụ 7.4: Với số liệu ví dụ 7.3. Sử dụng hàm hồi qui tuyến tính đơn giả để dự báo ngày 16.

Ta tìm được hàm hồi qui mẫu như sau:

$\hat{y}$ =909,4 + 2,8t

Dự báo số lượng bán ngày 16:

$\hat{y}$ =909,4 + 2,8x16 = 953,7

IV. PHÂN TÍCH TÍNH THỜI VỤ CỦA DÃY SỐ THỜI GIAN

1. Các yếu tố ảnh hưởng đến biến động của dãy Số thời gian

- Tính xu hướng (T): thể hiện chiều hướng biến động, tăng hoặc giảm, của hiện tượng trong một thời gian dài.

- Tính chu kỳ (C): Biến động của hiện tượng được lặp lại với một chu kỳ nhất định. Thường kéo dài từ 2 đến 10 năm và trải qua 4 giai đoạn: phục hồi và phát triển, thịnh vượng, suy thoái, đình trệ.

- Tính thời vụ (S): Là sự biến động của hiện tượng ở một số thời điểm nào đó trong năm được lặp đi lặp lại qua nhiều năm.

- Tính ngẫu nhiên hay bất thường (I): Là biến động không có qui luật và hầu như không dự đoán được.

Gọi X_i là giá trị của hiện tượng ở thời gian i, ta có:

$$X_i = T_i \,.\, C_i \,.\, S_i \,.\, I_i$$

2. Phân tích chỉ số thời vụ

Trong mô hình dự báo, đại lượng ngẫu nhiên I ta không dự đoán được, và yếu tố chu kỳ cũng khó xác định. Do đó, ta chủ yếu dự đoán sự ảnh hưởng của chỉ số thời vụ dựa vào 2 yếu tố T và S.

Ví dụ 7.4: Trưởng phòng kinh doanh của một công ty muốn phân tích tính chất thời vụ trong hoạt động của công ty để dự báo cho những năm tới. Doanh số bán hàng quí từ năm 2004- 2007 được thu thập.

Năm	Quí	Doanh số	Năm	Quí	Doanh số
2004	I	170	2006	I	157
	II	148		II	145
	III	141		III	128
	IV	150		IV	134
2005	I	161	2007	I	160
	II	137		II	139
	III	132		III	130
	IV	158		IV	144

2.1 Tính xu hướng của dãy số thời gian (T): Dựa vào các mô hình toán học, có thể làm tuyến tính hoặc hàm phi tuyến tính. Nhưng thông thường ta sử dụng hàm tuyến tính theo thời gian.

$$\hat{y} = b_0 + b_1 t$$

Ta chọn biến t=1 - 16, ta tìm được hàm hồi qui tuyến tính như sau:

$$\hat{y} = 155{,}275 - 1{,}1059t$$

2.2. Tính thời vụ của dãy số thời gian (S): Là sự biến động của dãy số thời gian trong một năm, nên ta có thể xem xét theo tháng (12 tháng) hoặc quí (4 quí)

a) Số trung bình di động:

Số trung bình di động được tính từ nhóm m mức độ được tính như sau:

$$x^*_{i+(m-1)/2} = \frac{1}{m}.\sum_{j=1}^{m-1} x_{i+j}$$

Số trung bình loại bỏ tất cả các yếu tố bất thường, và số trung bình của các số hạng liên tiếp thể hiện xu hướng của dãy số thời gian làm cho dãy số trở nên phẳng hơn, nó sẽ bỏ các yếu tố bất thường và yếu tố ngắn hạn: X* = TC

* Trường hợp nhóm mức độ là số chẵn: Sau khi tính trung bình m mức độ ta tính thêm trung bình 2 mức độ.

Năm	Quí	Doanh số	Trung bình 4 mức độ	Trung bình 2 mức độ (x*)
2004	I	170		
	II	148		
	III	141	152,25	151,13
	IV	150	150,00	148,63
2005	I	161	147,25	146,13
	II	137	145,00	146,00
	III	132	147,00	146,50
	IV	158	146,00	147,00
2006	I	157	148,00	147,50
	II	145	147,00	144,00
	III	128	141,00	141,38
	IV	134	141,75	141,00
2007	I	160	140,25	140,50
	II	139	140,75	142,00
	III	130	143,25	
	IV	144		

b) Tính chỉ số thời vụ (SI) theo thời gian trong năm:

$$SI = \frac{TCSI}{TC} = \frac{x}{x^*}$$

Năm	Quí	Doanh số (x=TCSI)	TC (x*)	SI=x/x*(%)
2004	I	170		
	II	148		
	III	141	151,13	93,3
	IV	150	148,63	100,9
2005	I	161	146,13	110,2
	II	137	146,00	93,8
	III	132	146,50	90,1
	IV	158	147,00	107,5
2006	I	157	147,50	106,4

	II	145	144,00	100,7
	III	128	141,38	90,5
	IV	134	141,00	95,0
2007	I	160	140,50	113,9
	II	139	142,00	97,9
	III	130		
	IV	144		

c) Tính chỉ số thời vụ điều chỉnh (S):

Năm	Quí I	Quí II	Quí III	Quí IV	
2004			93,3	100,9	
2005	110,2	93,8	90,1	107,5	
2006	106,4	100,7	90,5	95,0	
2007	113,9	97,9			
Tổng	330,5	292,4	273,9	303,4	
Chỉ số thời vụ trung bình (SI)	110,17	97,47	91,30	101,13	400,07
Chỉ số thời vụ điều chỉnh (S)	110,15	97,45	91,28	101,12	

* Theo quí : $\text{Chỉ số thời vụ điều chỉnh} = (SI) \times \frac{400}{\text{Tổng chỉ số thời vụ TB}}$

* Theo tháng : $\text{Chỉ số thời vụ điều chỉnh} = (SI) \times \frac{1.200}{\text{Tổng chỉ số thời vụ TB}}$

5.23. Tính chu kỳ:

$$C = \frac{TC}{T} = \frac{x^*}{T}$$

Năm	**Quí**	**t**	**Doanh số**	**x*=TC**	**T** y = 155,275 - 1,1059t	**C(%)**
2004	I	1	170			
	II	2	148			
	III	3	141	151,13	151,96	99,46
	IV	4	150	148,63	150,85	98,53
2005	I	5	161	146,13	149,75	97,59
	II	6	137	146,00	148,64	98,22
	III	7	132	146,50	147,53	99,30
	IV	8	158	147,00	146,43	100,39
2006	I	9	157	147,50	145,32	101,50
	II	10	145	144,00	144,22	99,85

Năm	Quí	t	Doanh số	x*=TC	T y = 155,275 - 1,1059t	C(%)
	III	11	128	141,38	143,11	98,79
	IV	12	134	141,00	142,00	99,29
2007	I	13	160	140,50	140,90	99,72
	II	14	139	142,00	139,79	101,58
	III	15	130			
	IV	16	144			

Từ kết quả vừa tính ta có thể sử dụng để dự báo cho các quí trong năm 2008.

CHƯƠNG VIII
PHƯƠNG PHÁP CHỌN MẪU[1]

I. ĐIỀU TRA CHỌN MẪU

Quá trình nghiên cứu thống kê gồm các giai đoạn: Thu thập số liệu, xử lý tổng hợp và phân tích, dự báo.

Trong thu thập số liệu thường áp dụng hai hình thức chủ yếu: Báo cáo thống kê định kỳ và điều tra thống kê.

Báo cáo thống kê định kỳ là hình thức thu thập số liệu thống kê được tiến hành thường xuyên, định kỳ theo nội dung, phương pháp cũng như hệ thống biểu mẫu thống nhất, được quy định thành chế độ báo cáo do cơ quan có thẩm quyền quyết định và áp dụng cho nhiều năm.

Điều tra thống kê là hình thức thu thập số liệu được tiến hành theo phương án quy định cụ thể cho từng cuộc điều tra. Trong phương án điều tra quy định rõ mục đích, nội dung, đối tượng, phạm vi, phương pháp và kế hoạch tiến hành điều tra. Điều tra thống kê được áp dụng ngày càng rộng rãi trong điều kiện nền kinh tế thị trường có nhiều thành phần kinh tế.

Điều tra thống kê được phân thành điều tra toàn bộ và điều tra không toàn bộ. Điều tra toàn bộ nhằm tiến hành thu thập số liệu ở tất cả các đơn vị của tổng thể. Trong khi đó điều tra không toàn bộ chỉ tiến hành thu thập số liệu của một bộ phận các đơn vị trong tổng thể. Trong điều tra không toàn bộ còn chia ra điều tra trọng điểm, điều tra chuyên đề và điều tra chọn mẫu.

Điều tra trọng điểm và điều tra chuyên đề khác với điều tra chọn mẫu ở chỗ kết quả của nó không dùng để suy rộng cho tổng thể. Kết quả của điều tra chọn mẫu được dùng để mô tả đặc điểm của tổng thể.

1. Điều tra chọn mẫu, ưu điểm, hạn chế và điều kiện vận dụng:

1.1. Khái niệm điều tra chọn mẫu:

Điều tra chọn mẫu là loại điều tra không toàn bộ, trong đó người ta chọn một cách ngẫu nhiên một số đủ lớn đơn vị đại diện trong toàn bộ các đơn vị của tổng thể để điều tra rồi dùng kết quả thu thập được tính toán, suy rộng thành các đặc điểm của toàn bộ tổng thể. Ví dụ: để có năng suất và sản lượng lúa của một địa bàn điều tra nào đó người ta chỉ tiến hành thu thập số liệu về năng suất và sản lượng lúa thu trên diện tích của một số hộ gia đình được chọn vào mẫu của huyện để điều tra thực tế, sau đó dùng kết quả thu được tính toán và suy rộng cho năng suất và sản lượng lúa của toàn huyện.

Điều tra chọn mẫu được ứng dụng rất rộng rãi trong thống kê kinh tế - xã hội như: Điều tra năng suất, sản lượng lúa; Điều tra lao động - việc làm; Điều tra thu nhập, chi tiêu của hộ gia đình; Điều tra biến động thường xuyên dân số; Điều tra chất lượng sản phẩm công nghiệp.

Ngoài ra, trong tự nhiên, trong đời sống sinh hoạt của con người, trong y học, v.v... chúng ta cũng đã gặp rất nhiều ví dụ thực tế đã áp dụng điều tra chọn mẫu; chẳng hạn: Khi đo lượng nước mưa của một khu vực nào đó người ta chỉ chọn ra một số điểm trong khu vực và đặt các ống nghiệm để đo lượng nước mưa qua các trận mưa trong từng tháng và cả năm, sau đó dựa vào kết quả nước mưa đo được từ mẫu là các ống nghiệm để tính toán suy rộng về lượng nước trung bình các tháng và cả năm cho cả khu vực; khi nghiên cứu ảnh

[1] Chương được tham khảo từ tài liệu *Một số vấn đề phương pháp luận thống kê,* năm 2005 của Viện Khoa học thống kê.

hưởng của hút thuốc lá đối với sức khoẻ con người, người ta chọn ra một số lượng cần thiết người hút thuốc lá để kiểm tra sức khoẻ và dùng kết quả kiểm tra từ một số người đó để kết luận về ảnh hưởng của hút thuốc lá tới sức khoẻ cộng đồng, v.v...

1.2. Ưu điểm của điều tra chọn mẫu:

Do chỉ tiến hành điều tra trên một bộ phận đơn vị mẫu trong tổng thể nên điều tra chọn mẫu có những ưu điểm sau:

- Tiến hành điều tra nhanh gọn, bảo đảm tính kịp thời của số liệu thống kê.

- Tiết kiệm nhân lực và kinh phí trong quá trình điều tra.

- Cho phép thu thập được nhiều chỉ tiêu thống kê, đặc biệt đối với các chỉ tiêu có nội dung phức tạp, không có điều kiện điều tra ở diện rộng. Nhờ đó kết quả điều tra thu được sẽ phản ánh được nhiều mặt, cho phép nghiên cứu các mối quan hệ cần thiết của hiện tượng nghiên cứu.

- Làm giảm sai số phi chọn mẫu như: sai số do công cụ đo lường, tính trung thực và trình độ của phỏng vấn viên, dữ liệu biến động,…. Trong thực tế công tác thống kê sai số phi chọn mẫu luôn luôn tồn tại và ảnh hưởng không nhỏ đến chất lượng số liệu thống kê, nhất là các chỉ tiêu có nội dung phức tạp, việc tiếp cận để thu thập số liệu khó khăn, tốn nhiều thời gian trong quá trình phỏng vấn, ghi chép và đặc biệt hơn là đối với các chỉ tiêu điều tra không có sẵn thông tin mà đòi hỏi phải hồi tưởng để nhớ lại. Đối với những loại thông tin như trên, chỉ có tiến hành điều tra mẫu mới có điều kiện tuyển chọn điều tra viên tốt hơn; hướng dẫn nghiệp vụ kỹ hơn, thời gian dành cho một đơn vị điều tra nhiều hơn, tạo điều kiện cho các đối tượng cung cấp thông tin trả lời chính xác hơn, tức là làm cho sai số phi chọn mẫu ít hơn.

- Cho phép nghiên cứu các hiện tượng kinh tế - xã hội, môi trường... không thể tiến hành theo phương pháp điều tra toàn bộ: Ví dụ như nghiên cứu trữ lượng khoáng sản, thuỷ sản, kiểm tra phá huỷ,…

1.3. Hạn chế của điều tra chọn mẫu:

- Do điều tra chọn mẫu chỉ tiến hành thu thập số liệu trên một số đơn vị, sau đó dùng kết quả để suy rộng cho toàn bộ tổng thể nên kết quả điều tra chọn mẫu luôn tồn tại sai số ta gọi đó là sai số chọn mẫu - Sai số do tính đại diện. Sai số chọn mẫu phụ thuộc vào độ đồng đều của chỉ tiêu nghiên cứu, vào cỡ mẫu và phương pháp tổ chức chọn mẫu. Có thể làm giảm sai số chọn mẫu bằng cách tăng cỡ mẫu ở phạm vi cho phép và lựa chọn phương pháp tổ chức chọn mẫu thích hợp nhất.

- Kết quả điều tra chọn mẫu không thể tiến hành phân nhỏ theo mọi phạm vi và tiêu thức nghiên cứu như điều tra toàn bộ, mà chỉ thực hiện được ở mức độ nhất định tuỳ thuộc vào cỡ mẫu, phương pháp tổ chức chọn mẫu và độ đồng đều giữa các đơn vị theo các chỉ tiêu được điều tra.

1.4. Điều kiện vận dụng của điều tra chọn mẫu:

Điều tra chọn mẫu thường được vận dụng trong các trường hợp sau:

- Thay thế cho điều tra toàn bộ trong những trường hợp quy mô điều tra lớn, nội dung điều tra cần thu thập nhiều chỉ tiêu, thực tế ta không đủ kinh phí và nhân lực để tiến hành điều tra toàn bộ, hơn nữa nếu điều tra toàn bộ sẽ mất quá nhiều thời gian, không đảm bảo tính kịp thời của số liệu thống kê như điều tra thu nhập, chi tiêu hộ gia đình, điều tra năng suất, sản lượng lúa, điều tra vốn đầu tư của các đơn vị ngoài quốc doanh...; hoặc không tiến hành được điều tra toàn bộ vì không thể xác định được tổng thể như điều tra đánh giá mức độ ô nhiễm môi trường nước của một số sông, hồ nào đó,…

- Quá trình điều tra gắn liền với việc phá huỷ sản phẩm như điều tra đánh giá chất lượng thịt hộp, cá hộp, y tá lấy máu của bệnh nhân để xét nghiệm, v.v... Các trường hợp trên đây

nếu điều tra toàn bộ thì sau khi điều tra toàn bộ sản phẩm sản xuất ra có thể làm tăng chỉ phí và giảm chất lượng hàng hoá hoặc lượng máu có trong cơ thể của bệnh nhân sẽ bị phá huỷ hoàn toàn. Đây là điều khó cho phép thực hiện trong thực tế.

- Để thu thập những thông tin tiên nghiệm trong những trường hợp cần thiết nhằm phục vụ cho yêu cầu của điều tra toàn bộ. Ví dụ: để thăm dò mức độ tín nhiệm của các ứng cử viên vào một chức vị nào đó thì chỉ có thể điều tra chọn mẫu ở một lượng cử tri nhất định và phải được tiến hành trước khi bầu cử chính thức thì mới có ý nghĩa.

- Thu thập số liệu để kiểm tra, đánh giá và chỉnh lý số liệu của điều tra toàn bộ. Trong thực tế có những cuộc điều tra toàn bộ có quy mô lớn hoặc điều tra rất phức tạp như Tổng Điều tra Dân số và Nhà ở, Tổng Điều tra Nông thôn, Nông nghiệp và Thuỷ sản... thì sai số do khai báo, thu thập thông tin thường xuyên tồn tại và ảnh hưởng đáng kể đến chất lượng số liệu. Vì vậy cần có điều tra chọn mẫu với quy mô nhỏ hơn để xác định mức độ sai số này, trên cơ sở đó tiến hành đánh giá độ tin cậy của số liệu và nếu ở mức độ cần thiết có thể phải chỉnh lý lại số liệu thu được từ điều tra toàn bộ.

2. Sai số chọn mẫu và phạm vi sai số chọn mẫu:

2.1. Sai số chọn mẫu:

Sai số chọn mẫu là sự khác nhau giữa giá trị ước lượng của mẫu và giá trị của tổng thể. Sai số chọn mẫu còn gọi là sai số do tính đại diện. Sai số này chỉ xảy ra trong điều tra chọn mẫu do chỉ điều tra một số ít đơn vị mà kết quả lại suy cho cả tổng thể. Sai số chọn mẫu có hai loại:

- Sai số có hệ thống: Sai số xảy ra khi áp dụng phương pháp chọn có hệ thống, làm cho kết quả điều tra luôn bị lệch so với số thực tế về một hướng.

- Sai số ngẫu nhiên: Sai số chỉ xuất hiện trong trường hợp các đơn vị của tổng thể được chọn theo nguyên tắc ngẫu nhiên, không phụ thuộc vào ý định của người điều tra.

***2.2. Phạm vi sai số chọn mẫu**:*

Phạm vi sai số chọn mẫu (ký hiệu là Δ_x, L_x, ε_x) bằng tích của hệ số tin cậy (t_α) và sai số chọn mẫu (σ_x)

$$\Delta_x = t_\alpha.\sigma_x$$

Trong đó: Hệ số tin cậy là xác suất để giá trị thực tế của chỉ tiêu nghiên cứu (θ: tham số tổng thể) còn nằm trong khoảng tin cậy ($\hat{\theta} \pm t_\alpha.\sigma_x$). Trong đó, $\hat{\theta}$ là tham số mẫu.

Ý nghĩa của hàm xác suất này được biểu hiện như sau:

$$P\left[\left|\hat{\theta} - \theta\right| \le \Delta_x\right] = 1 - \alpha$$

***2.3. Ý nghĩa của việc tính toán sai số chọn mẫu**:*

- Sai số chọn mẫu dùng để ước lượng chỉ tiêu nghiên cứu theo khoảng tin cậy.

- Sai số chọn mẫu dùng để đánh giá tính đại diện của chỉ tiêu nghiên cứu qua tính toán tỷ lệ sai số chọn mẫu (CV) như sau:

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

CV càng nhỏ thì chỉ tiêu có tính đại diện càng cao và ngược lại.

- Là cơ sở để xác định cỡ mẫu cho các cuộc điều tra được tiến hành về sau.

3. Đơn vị chọn mẫu, dàn chọn mẫu:

3.1. Đơn vị chọn mẫu:

Đơn vị chọn mẫu là các đơn vị cơ bản hoặc nhóm đơn vị cơ bản được xác định rõ ràng, tương đối đồng đều và có thể quan sát được, thích hợp cho mục đích chọn mẫu. Ví dụ: Doanh nghiệp, hộ gia đình, đơn vị diện tích gieo trồng, xã, phường, xóm, bản...

Nếu chọn mẫu một cấp thì có một loại đơn vị chọn mẫu, còn nếu chọn mẫu nhiều cấp thì sẽ có nhiều loại đơn vị chọn mẫu. Tức là lược đồ chọn mẫu theo bao nhiêu cấp thì có bấy nhiêu loại đơn vị chọn mẫu.

3.2. Dàn chọn mẫu:

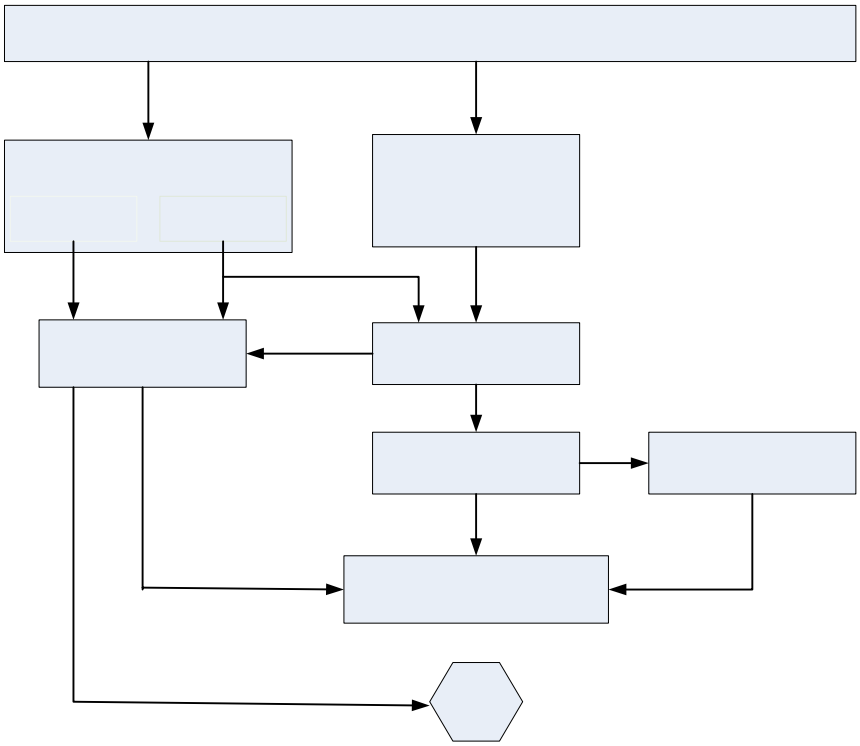
Dàn chọn mẫu có thể là danh sách các đơn vị chọn mẫu với những đặc điểm nhận dạng của chúng hoặc là bản đồ chỉ ra ranh giới của các đơn vị được dùng làm căn cứ để tiến hành chọn mẫu. Khi tổ chức điều tra thống kê.

Trong tổng thể nghiên cứu, tuỳ thuộc vào lược đồ chọn mẫu mà sẽ có các loại dàn chọn mẫu khác nhau. Nếu điều tra mẫu một cấp (giả định điều tra các hộ trên địa bàn huyện) thì dàn chọn mẫu là danh sách các hộ gia đình của tất cả các xã trong huyện. Còn nếu điều tra mẫu hai cấp, cấp I là xã và cấp II là hộ gia đình thì có hai loại dàn chọn mẫu: Dàn chọn mẫu cấp I là danh sách tất cả các xã trong huyện, còn dàn chọn mẫu cấp II là danh sách các hộ gia đình của những xã được chọn ở mẫu cấp I.

4. Các hình thức và phương pháp chọn mẫu:

4.1. Các hình thức thu thập số liệu thống kê:

Sơ đồ 8.1. Các hình thức thu thập số liệu thống kê



a) Chọn mẫu phi xác suất:

- Chọn mẫu thuận tiện: Các đơn vị mẫu được chọn ở tại một địa điểm và vào một thời gian nhất định.

Ví dụ: Chọn mẫu những người đi mua sắm ở siêu thị METRO Cần Thơ và tiếp cận họ khi họ bước vào siêu thị hoặc khi họ vừa mua sắm món hàng mà ta muốn khảo sát.

Cách chọn mẫu này là không ngẫu nhiên vì không phải tất cả mọi người dân Cần Thơ đều có xác suất như nhau để được chọn vào mẫu.

(1) Ưu điểm: dễ dàng để tập hợp các đơn vị mẫu.

(2) Nhược điểm: không đạt được độ xác thực cao.

- Chọn mẫu phán đoán: Các đơn vị mẫu được chọn dựa vào sự phán đoán của người nghiên cứu mà họ nghĩ rằng những mẫu này có thể đại diện cho tổng thể.

Ví dụ: Chọn mẫu một số ít liên doanh lớn có thể chiếm phần lớn tổng sản lượng ngành công nghiệp cả nước.

Cách chọn mẫu này được dùng phổ biến khi nghiên cứu định tính *(nghiên cứu khách hàng công nghiệp hay kênh phân phối)*.

(1) Ưu điểm: Chọn được đúng một số phần tử rất quan trọng của tổng thể vào trong mẫu.

(2) Nhược điểm: Có khả năng phát sinh những sai lệch lớn.

- Chọn mẫu chỉ định: Là chọn mẫu theo tỷ lệ gần đúng của các nhóm đại diện trong tổng thể hoặc theo số mẫu được chỉ định cho mỗi nhóm.

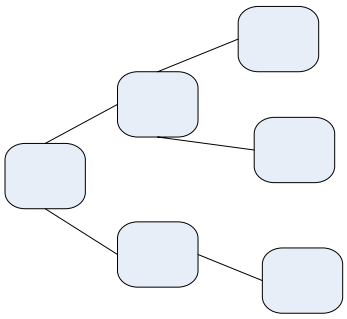
Ví dụ: Chọn 100 phần tử cho mỗi nhãn hiệu nước giải khát để so sánh kết quả thống kê về thái độ khách hàng. Hoặc tổng thể nghiên cứu bao gồm 1.000 công ty, trong đó 600 công ty vừa và nhỏ, 300 trung bình và 100 qui mô lớn. Số mẫu chỉ định là 10% trên tổng thể, ta sẽ chọn 60 công ty vừa và nhỏ, 30 trung bình và 10 công ty lớn.

(1) Ưu điểm: Đảm bảo được số mẫu cần thiết cho từng nhóm trong tổng thể phục vụ cho phân tích.

(2) Nhược điểm: Có thể cho kết quả sai lệch.

- Chọn mẫu theo mạng quan hệ: Người nghiên cứu sẽ thông qua người trả lời đầu tiên để tiếp cận những người trả lời kế tiếp.

Ví dụ: Phỏng vấn An, khoảng gần kết thúc phỏng vấn có thể đề nghị An giới thiệu một hoặc hai người nữa có thể am hiểu về chủ đề khảo sát.



+ Ưu điểm: giúp cho người nghiên cứu chọn được các mẫu mà họ cần nghiên cứu.

b) Chọn mẫu xác suất: Dựa vào lý thuyết xác suất để lấy mẫu ngẫu nhiên.

- Chọn mẫu ngẫu nhiên đơn giản: Là cách chọn mẫu mà mỗi phần tử trong tổng thể có cùng cơ hội được chọn với xác suất như nhau. Để chọn được mẫu, người nghiên cứu phải có danh sách tổng thể nghiên cứu.

Ví dụ: Chọn 100 mẫu sinh viên trong tổng số 4.000 sinh viên khoa kinh tế & QTKD. Sử dụng máy tính gieo một số gồm 4 chữ số. Mỗi lần gieo, số nào xuất hiện sẽ được chọn vào mẫu.

- Chọn mẫu có hệ thống: Là cách chọn mẫu mà mẫu đầu tiên là ngẫu nhiên, sau đó cứ cách k đơn vị ta lại chọn một phần tử. Với k = *N/n (N: độ lớn tổng thể và n kích cở mẫu)*.

Ví dụ: Như ví dụ chọn mẫu ngẫu nhiên đơn giản, danh sách sinh viên không sẵn có trên máy tính, ta sẽ sử dụng cách chọn mẫu hệ thống. Tính k = 4.000/100 = 40. Ta sẽ gieo số ngẫu nhiên trong khoảng 1 đến 40. Ví dụ được số 18. Mẫu được chọn là 18. Các mẫu tiếp theo sẽ là 18 + 40 = 58, 58 + 40 = 98 …

- Chọn mẫu ngẫu nhiên phân tầng: Là phân chia các đối tượng nghiên cứu thành các nhóm, tầng theo các đặc tính, sau đó lấy mẫu theo tầng, nhóm.

Ví dụ: Chúng ta muốn biết chi tiêu cho nước giải khát giữa sinh viên và giảng viên trong khoa KT & QTKD. Chúng ta cần ấn định trước số mẫu cho mỗi nhóm. Chẳng hạn, số mẫu gồm 70 sinh viên và 30 giảng viên. Sau đó ta sẽ chọn mẫu theo phương pháp ngẫu nhiên đơn giản từ 2 nhóm độc lập

4.2. Các phương pháp chọn mẫu thường được sử dụng:

a) Phương pháp chọn mẫu ngẫu nhiên (Probability Sampling Methods):

- Chọn mẫu ngẫu nhiên đơn giản: Chọn mẫu ngẫu nhiên đơn giản là chọn các đơn vị từ tổng thể vào mẫu hoàn toàn hú hoạ. Cách đơn giản nhất của chọn mẫu ngẫu nhiên là rút thăm hoặc sử dụng bảng số ngẫu nhiên.

- Chọn mẫu hệ thống (Systematic Sampling): Chọn mẫu hệ thống là chọn các đơn vị từ tổng thể vào mẫu theo một khoảng cách cố định sau khi đã chọn ngẫu nhiên một nhóm nào đó trên cơ sở các đơn vị điều tra được sắp xếp thứ tự theo một tiêu thức nhất định.

Ví dụ: Trường đại học X có 2000 sinh viên (N = 2000). Cần chọn 100 sinh viên (n = 100) để điều tra mức sống của họ. Nếu chọn hệ thống sẽ tiến hành như sau:

- Lập danh sách 2000 sinh viên của trường theo thứ tự nào đó, chẳng hạn theo vần A, B, C... của tên gọi.

- Chia tổng số sinh viên của trường thành 100 nhóm đều nhau và sẽ có số sinh viên mỗi nhóm là 20 sinh viên: (K = N: n = 2000: 100).

- Chọn ngẫu nhiên một sinh viên ở nhóm thứ nhất, chẳng hạn rơi vào sinh viên có số thứ tự 15.

- Mỗi nhóm khác còn lại sẽ chọn 1 sinh viên có số thứ tự: nhóm 2: (15+K), nhóm 3: (15+2K)...; nhóm 100: (15+99K).

Một nhược điểm của chọn mẫu hệ thống là có thể dẫn đến mẫu được chọn có thể bị lệch một cách có hệ thống. Lý do, nếu danh sách các phần tử được bố trí theo một kiểu tuần hoàn trùng hợp với khoảng cách lấy mẫu, đó có lẽ là một mẫu bị lệch có thể được lựa chọn. Vì vậy, trong việc xem xét một mẫu có hệ thống cần phải khảo sát thận trọng bản chất của danh sách.

Tuy nhiên, phương pháp lấy mẫu có hệ thống có thể tạo ra một mẫu ngẫu nhiên đơn giản. Nó sẽ đảm bảo có sự đại diện từ các khoảng cao nhất đến thấp nhất (nếu dàn chọn mẫu đã được sắp xếp thứ tự theo một tiêu chí nào đó). Một mẫu ngẫu nhiên đơn giản khó thực hiện được như vậy.

- Chọn mẫu theo tổ: Để thực hiện theo phương pháp này, trước hết phải phân tổ các đơn vị tổng thể chung thành nhiều tổ theo tiêu thức có liên quan trực tiếp đến nội dung nghiên cứu. Sau đó, sau đó xác định cỡ mẫu cho từng tổ. Việc phân chia này được thực hiện một trong ba phương pháp:

- Số quan sát điều tra trong từng tổ được chia đều, tức là cỡ mẫu bằng nhau ở các tổ.

- Số quan sát được chia theo tỷ lệ số lượng đơn vị tổng thể của từng tổ trong tổng thể chung.

- Số quan sát được chia theo tỷ lệ số lượng đơn vị tổng thể của từng tổ trong tổng thể chung và độ lệch chuẩn của từng tổ.

- Chọn mẫu theo khối (chùm) (Cluster Sampling): Theo phương pháp tổ chức chọn mẫu này thì trước tiến các đơn vị của tổng thể được chia thành R khối với số lượng đơn vị bằng nhau hoặc không bằng nhau. Từ R khối chọn ra r khối ngẫu nhiên theo phương pháp ngẫu nhiên đơn giản hoặc hệ thống và điều tra tất cả các đơn vị của r khối.

Chọn cả khối có ưu điểm là tổ chức gọn nhẹ, giảm được kinh phí. Song vì số đơn vị được chọn để điều tra chỉ tập trung vào một số khối nên có thể có sai số lớn nếu giữa các khối có sự khác biệt nhau.

- Chọn mẫu ngẫu nhiên phân tầng (Stratified Random Sampling): Phương pháp chọn mẫu phân tầng (còn được gọi là chọn mẫu nhiều cấp) là phương pháp chọn mẫu phải thông qua ít nhất hai cấp chọn trung gian. Đầu tiên cần phải xác định các đơn vị mẫu cấp 1, sau đó là các đơn vị mẫu cấp 1 lại được chia thành các đơn vị chọn mẫu cấp 2 và cứ như thế cho đến cấp cuối cùng. Về bản chất, phương pháp này là sự biến thể của phương pháp chọn mẫu khối. Thật vậy, nếu điều tra chọn mẫu hai cấp thì ở cấp 1, tổng thể được chia thành các khối, sau đó chọn mẫu ngẫu nhiên một số khối nhất định. Ở cấp 2, thay vì điều tra toàn bộ các đơn vị của các chùm được chọn ra, người ta chỉ chọn và điều tra một số đơn vị của các chùm đó mà thôi.

Kết quả chọn được 100 sinh viên như vậy được gọi là chọn hệ thống.

- Chọn mẫu theo phương pháp phân tích chuyên gia là chọn mẫu trên cơ sở phân tích xem xét chủ quan của người điều tra. Cách chọn này thường áp dụng cho tổng thể có ít đơn vị mẫu hoặc trị số của chỉ tiêu nghiên cứu giữa các đơn vị mẫu chênh lệch nhau nhiều.

b) Phương pháp chọn mẫu phi ngẫu nhiên (Nonprobability Sampling Methods):

Điều tra chọn mẫu phi ngẫu nhiên là điều tra chọn mẫu mà trong đó các đơn vị của tổng thể được chọn ra trên cơ sở phân tích đặc điểm của hiện tượng và kinh nghiệm thực tế. Do đó, để đảm bảo chất lượng của tài liệu điều tra, cần phải giải quyết tốt các vấn đề sau đây:

- Phân tích chính xác hiện tượng nghiên cứu: Hiện tượng nghiên cứu thường có kết cấu phức tạp, gồm nhiều tổ, nhiều bộ phận và tính chất khác nhau. Trên cơ sở phân tổ chính xác hiện tượng nghiên cứu, các đơn vị có đặc điểm và tính chất giống nhau (hoặc gần giống nhau) sẽ được đưa vào một tổ. Từ mỗi tổ sẽ chọn ra các đơn vị đại diện (còn gọi là điển hình) cho tổ đó. Tập hợp các đơn vị đại diện của các tổ tạo thành tổng thể mẫu.

- Lựa chọn các đơn vị điều tra: Các đơn vị được lựa chọn để điều tra thực tế thường là những đơn vị có mức độ của tiêu thức xấp xỉ với mức độ trung bình của tổ. Khi lựa chọn các đơn vị để điều tra thực tế cần phải thông qua việc phân tích, bàn bạc tập thể của những người có kinh nghiệm, am hiểu tình hình thức tế.

- Suy rộng kết quả điều tra: Sau khi thu thập được dữ liệu ở các đơn vị điều tra thì tiến hành tính toán suy rộng trực tiếp cho toàn bộ hiệu tượng. Vì các đơn vị điều tra được lựa chọn đại diện cho từng tổ nên khi suy rộng phải chú ý đến tỷ trọng của mỗi tổ chiếm trong toàn bộ hiện tượng.

4.3. Các phương pháp tổ chức chọn mẫu:

Có nhiều phương pháp, tổ chức chọn mẫu khác nhau. Mỗi phương pháp có những ưu, nhược điểm riêng và được áp dụng trong những điều kiện nhất định. Tuy nhiên gọi là phương pháp này hay phương pháp kia là đứng trên những giác độ khác nhau và cũng chỉ có ý nghĩa tương đối.

a) Xét theo cấp chọn mẫu:

Có phương pháp tổ chức chọn mẫu một cấp và tổ chức chọn mẫu hai cấp hay nhiều cấp:

- Chọn mẫu một cấp: Là từ một loại danh sách của tất cả các đơn vị thuộc tổng thể, tiến hành chọn mẫu một lần trực tiếp đến các đơn vị điều tra không qua một phân đoạn nào khác. Chọn mẫu một cấp chỉ có một loại đơn vị chọn mẫu và một dàn chọn mẫu. Đối với mẫu một cấp có thể dùng cách chọn ngẫu nhiên, nhưng cũng có thể dùng cách chọn hệ thống hoặc chọn theo phương pháp chuyên gia. Tuy nhiên, trong thực tế nếu là điều tra mẫu một cấp thì phổ biến là dùng cách chọn ngẫu nhiên và thường được gọi tắt là chọn mẫu ngẫu nhiên đơn giản. Chọn mẫu ngẫu nhiên đơn giản đảm bảo số mẫu được rải trên toàn địa bàn điều tra nên sai số chọn mẫu sẽ nhỏ. Song khó khăn là việc lập danh sách các đơn vị (dàn chọn mẫu) để tiến hành chọn mẫu khá lớn, tốn nhiều thời gian và công sức. Hơn nữa khi tổ chức điều tra phải thực hiện ở địa bàn rất rộng.

- Chọn mẫu nhiều cấp: Là tiến hành điều tra theo nhiều công đoạn, trong đó mỗi công đoạn là một cấp chọn mẫu. Có bao nhiêu cấp điều tra thì có bấy nhiêu loại đơn vị chọn mẫu cũng như có bấy nhiêu loại dàn chọn mẫu. Phương pháp tổ chức chọn mẫu nhiều cấp thuận tiện cho việc lập dàn chọn mẫu và tổ chức điều tra: Ở cấp sau chỉ phải lập dàn chọn mẫu cho cấp đó trong phạm vi mẫu cấp trước được chọn, phạm vi điều tra được thu hẹp sau mỗi cấp điều tra. Tuy nhiên, với phương pháp tổ chức chọn mẫu nhiều cấp số liệu thu thập được thường có độ tin cậy thấp hơn so với chọn mẫu ngẫu nhiên đơn giản.

b) Xét theo chọn mẫu phân tổ:

Nếu trước khi chọn mẫu, tiến hành phân chia tổng thể thành những tổ khác nhau theo một hay một số tiêu thức nào đó liên quan đến tiêu thức điều tra, sau đó phân bổ cỡ mẫu cho từng tổ và trong mỗi tổ lập một danh sách riêng và chọn đủ số mẫu phân bổ cho tổ đó. Cách chọn như vậy gọi là chọn mẫu phân tổ:

- Phương pháp chọn mẫu phân tổ: Nếu việc phân tổ được tiến hành khoa học thì tổng thể mẫu sẽ có kết cấu gần tổng thể, do đó sai số chọn mẫu sẽ giảm đi, tính chất đại diện của tổng thể mẫu được nâng cao.

Tuy nhiên, chọn mẫu phân tổ cũng khó khăn trong việc lập dàn chọn mẫu như chọn mẫu ngẫu nhiên đơn giản. Hơn nữa tổ chức điều tra phải tiến hành trên địa bàn rộng, thậm chí còn phức tạp hơn cả chọn mẫu ngẫu nhiên đơn giản.

Nếu điều tra chia thành nhiều cấp, các cấp tiến hành trước thì chọn từng đơn vị mẫu, nhưng ở cấp cuối cùng không chọn ra từng đơn vị, mà chọn cả nhóm các đơn vị để điều tra. Cách chọn như vậy gọi là chọn mẫu chùm (hay chọn mẫu cả khối).

Nếu cùng cỡ mẫu như nhau, chọn mẫu chùm so với các phương pháp tổ chức chọn mẫu nêu trên sẽ thuận tiện nhất cho việc lập dàn chọn mẫu và tổ chức điều tra. Tuy nhiên, độ tin cậy của số liệu thu thập được sẽ thấp hơn; tức là có sai số chọn mẫu lớn nhất.

5. Xác định cỡ mẫu, phân bổ mẫu và tính sai số chọn mẫu:

5.1. Xác định cỡ mẫu (số đơn vị mẫu):

Xác định cỡ mẫu (số đơn vị mẫu) chính là xác định số lượng đơn vị điều tra trong tổng thể mẫu để tiến hành thu thập số liệu. Yêu cầu của cỡ mẫu là vừa đủ để vừa đảm bảo độ tin cậy cần thiết của số liệu điều tra vừa đảm bảo phù hợp với điều kiện về nhân lực và kinh phí và có thể thực hiện được, tức là có tính khả thi.

Dưới đây sẽ trình bày cách xác định cỡ mẫu đơn thuần theo lý thuyết và việc xác định cỡ mẫu trong thực tế các cuộc điều tra thống kê.

a) Xác định cỡ mẫu theo các công thức lý thuyết:

Một tổng thể khi tiến hành điều tra không chia thành các tổng thể nhỏ (các tổ) thì chỉ có một cách xác định cỡ mẫu trên cơ sở thông tin về quy mô và phương sai của tổng thể. Đối

với một tổng thể khi điều tra có chia thành các tổng thể nhỏ có hai cách xác định cỡ mẫu: Cách thứ nhất xác định cỡ mẫu như trường hợp không phân tổ, sau đó phân bổ số mẫu chung cho các tổ theo nguyên tắc phân bổ mẫu. Cách thứ hai xác định cỡ mẫu trên cơ sở quy mô và phương sai của từng tổ.

Sau đây sẽ giới thiệu công thức xác định cỡ mẫu theo hai cách nói trên nhưng chỉ cho trường hợp tổ chức chọn mẫu ngẫu nhiên đơn giản hoặc có phân tổ và được áp dụng cho nghiên cứu chỉ tiêu bình quân với cách chọn không lặp làm ví dụ.

- Cách thứ nhất xác định cỡ mẫu trên cơ sở các thông tin về quy mô và phương sai của tổng thể:

$$n = \frac{N.t_\alpha^2.S^2}{N.\Delta_x^2 + t_\alpha^2.S^2}$$

Trong đó:

N - Số đơn vị tổng thể;

n - Số đơn vị mẫu;

t_α - Hệ số tin cậy;

Δ_x - Phạm vi sai số chọn mẫu;

S^2 - Phương sai của tổng thể.

- Cách thứ hai xác định cỡ mẫu trên cơ sở các thông tin về quy mô và phương sai của các tổ t:

$$n = \frac{\sum_{t=1}^{K} w_j S_j^2}{\frac{\Delta_x^2}{t_\alpha^2} + \frac{1}{N}\sum_{j=1}^{k} w_j S_j^2}$$

Trong đó:

N - Số đơn vị tổng thể;

n - Số đơn vị mẫu;

t_α - Hệ số tin cậy;

Δ_x - Phạm vi sai số chọn mẫu;

W_j - Tỷ trọng số đơn vị của tổ t trong tổng thể;

k - Số lượng tổ (j = 1, 2...k);

S_j^2 - Phương sai tổng thể của tổ j.

Từ các công thức trên, để xác định cỡ mẫu trong quá trình chuẩn bị phương án điều tra phải có được những thông tin sau:

N: Số đơn vị tổng thể. Chỉ tiêu này có đầy đủ ở phần lớn các cuộc điều tra thống kê;

w_j: Tỷ trọng số đơn vị của tổ t trong tổng thể. Đại lượng này xác định được trên cơ sở so sánh số đơn vị từng tổ (N_j) với số đơn vị toàn bộ tổng thể (N);

t_α, Δ_x: Hệ số tin cậy và phạm vi sai số chọn mẫu là những thông tin của chỉ tiêu điều tra và được ấn định từ trước do yêu cầu thuộc chủ quan của những người quản lý và tổ chức điều tra;

S_j^2: Phương sai của từng tổ j. Số liệu để tính các phương sai trên, cần có trước khi điều tra, song thực tế lại không có, do vậy thường phải dùng số liệu điều tra toàn bộ của các cuộc

điều tra trước (nếu có). Trường hợp không có số liệu của các cuộc điều tra trước thì phải tiến hành điều tra mẫu nhỏ. Tuy nhiên, việc điều tra mẫu nhỏ cũng khá phức tạp, mất nhiều thời gian, nhiều khi còn ảnh hưởng đến tiến độ thực hiện của cuộc điều tra chính.

Một khó khăn nữa là trong một cuộc điều tra chọn mẫu thường tiến hành thu thập thông tin về nhiều chỉ tiêu. Các chỉ tiêu khác nhau sẽ có quy luật phân phối và độ biến thiên khác nhau, tức là có phương sai khác nhau. Và do vậy, mỗi chỉ tiêu tính ra sẽ có một cỡ mẫu riêng (mặc dù yêu cầu về độ tin cậy của các chỉ tiêu điều tra như nhau). Nói cách khác, có bao nhiêu chỉ tiêu điều tra thì phải tính bấy nhiêu cỡ mẫu, sau đó sẽ chọn ra cỡ mẫu lớn nhất dùng chung cho điều tra tất cả các chỉ tiêu. Với nhiều cỡ mẫu đòi hỏi phải tính nhiều phương sai nên công việc tính toán càng trở nên phức tạp, tốn nhiều công sức, khó thực hiện.

Vì những đặc điểm trên đây, trong thực tế điều tra chọn mẫu còn ít khi áp dụng một cách trực tiếp các công thức trên để xác định cỡ mẫu.

Trong thống kê đã có một số cuộc điều tra chọn mẫu mà các chuyên gia chọn mẫu đã dựa vào thông tin của các cuộc điều tra có liên quan trước đó để xác định cỡ mẫu theo công thức lý thuyết. Song kết quả thu được còn khiêm tốn.

b) Xác định cỡ mẫu theo kinh nghiệm điều tra thực tế:

Trong thực tế nhiều khi các chuyên gia thống kê thường căn cứ vào cỡ mẫu của các cuộc điều tra có điều kiện và quy mô tương tự đã thực hiện thành công trước đó ở trong nước hoặc trên thế giới để xác định cỡ mẫu cho cuộc điều tra sau. Có nhiều cách xác định cỡ mẫu nhưng phổ biến nhất vẫn dựa vào tỷ lệ mẫu chung đã được điều tra và bổ sung thêm một tỷ lệ mẫu dự phòng nào đó.

Cách làm này đơn giản, nhanh chóng và dễ thực hiện, tức là có tính khả thi cao. Tuy nhiên làm như vậy chủ yếu vẫn là theo chủ nghĩa kinh nghiệm và gần như chưa tính đến mức độ biến động của các chỉ tiêu nghiên cứu.

c) Xác định cỡ mẫu dựa theo cỡ mẫu của cuộc điều tra đã thực hiện:

Có điều kiện, quy mô tương tự và đã được tiến hành thành công, nhưng có điều chỉnh (tăng lên hoặc giảm đi) trên cơ sở phân tích tỷ lệ sai số chọn mẫu của một số chỉ tiêu chủ yếu. Quá trình này được tiến hành theo hai hướng:

(1) Trước hết liệt kê những chỉ tiêu chủ yếu cùng được tổ chức thu thập số liệu trong cả 2 cuộc điều tra (cuộc điều tra trước đó đã hoàn chỉnh và cuộc điều tra lần này đang chuẩn bị); trong đó chọn ra một chỉ tiêu trong cuộc điều tra lần trước có tỷ lệ sai số chọn mẫu lớn nhất (từ đây chỉ tiêu được chọn gọi là chỉ tiêu nghiên cứu).

(2) Tiếp theo, tiến hành xem xét tỷ lệ sai số chọn mẫu của chỉ tiêu nghiên cứu tính được của cuộc điều tra lần trước và xử lý như sau:

- Nếu tỷ lệ sai số chọn mẫu đó lớn hơn mức độ cho phép thì phải điều chỉnh cỡ mẫu của cuộc điều tra lần này tăng lên so với cuộc điều tra trước;

- Nếu tỷ lệ sai số chọn mẫu đó nhỏ hơn mức độ cho phép thì có thể điều chỉnh cỡ mẫu giảm đi.

Chú ý:

- So sánh tỷ lệ sai số chọn mẫu là căn cứ quan trọng để điều chỉnh cỡ mẫu. Song đó không phải là căn cứ duy nhất, mà thực tế còn phải dựa vào một số yếu tố khác như sự thay đổi về quy mô tổng thể, thay đổi về số lượng chỉ tiêu điều tra...

- Điều kiện để áp dụng cách điều chỉnh cỡ mẫu trên đây là trong cuộc điều tra kỳ trước phải tính được tỷ lệ sai số chọn mẫu cho các chỉ tiêu chủ yếu.

Cách ước lượng này đơn giản và thuận tiện hơn nhiều so với cách tính cỡ mẫu theo lý thuyết, nhưng lại có cơ sở chắc chắn hơn so với cách xác định cỡ mẫu có tính chất ước đoán thuần tuý theo kinh nghiệm.

d. Cách xác định cỡ mẫu chủ yếu dựa vào khả năng về kinh phí:

Công thức xác định cỡ mẫu (n) trong trường hợp này như sau:

$$n = \frac{C - C_0}{Z}$$

Trong đó:

C - Tổng kinh phí được cấp;

C_0 - Kinh phí chi cho các khâu chuẩn bị, tập huấn nghiệp vụ thu thập, xử lý và các chi phí chung khác;

Z - Chi phí cần thiết cho tất cả các khâu điều tra tính cho một đơn vị điều tra.

5.2. Phân bổ mẫu:

Nếu địa bàn điều tra được chia thành các khu vực hoặc các tổ khác nhau và tiến hành điều tra trên tất cả các khu vực hoặc các tổ thì phải thực hiện phân bổ mẫu cho từng khu vực hoặc từng tổ đó.

Có nhiều cách phân bổ mẫu khác nhau, dưới đây chỉ giới thiệu một số cách phân bổ chủ yếu.

a) Phân bổ mẫu tỷ lệ thuận với quy mô tổng thể:

Công thức xác định cỡ mẫu của từng tổ t (n_t) như sau:

$$n_j = \frac{N_j}{N} n = N_j f$$

Trong đó:

j - Chỉ số thứ tự tổ (t = 1, 2...k)

n - Số đơn vị mẫu chung

n_j - Số đơn vị mẫu của tổ j

N - Số đơn vị của tổng thể

N_j - Số đơn vị của tổ t

f - Tỷ lệ mẫu ($f = \frac{n}{N}$)

Cách phân bổ mẫu tỷ lệ thuận với quy mô thường được áp dụng khi quy mô của các tổ tương đối đồng đều, phương sai và chi phí cho các tổ không khác nhau nhiều. Cách phân bổ này có ưu điểm: Dễ làm, không phải tính lại theo quyền số thực tế khi suy rộng kết quả là chỉ tiêu bình quân hoặc tỷ lệ cho tổng thể. Tuy nhiên, khi quy mô của các tổ khác nhau nhiều thì phân bổ tỷ lệ thuận với quy mô dễ làm cho các tổ có quy mô nhỏ thường không đủ số lượng mẫu để đại diện cho tổ đó, ngược lại các tổ có quy mô lớn lại "thừa" cỡ mẫu. Mặt khác, việc tổ chức điều tra cũng như kinh phí cần thiết cho điều tra ở các tổ có quy mô lớn sẽ rất nặng nề, còn việc tổ chức điều tra cũng như kinh phí cần thiết cho điều tra ở các tổ có quy mô nhỏ lại quá nhẹ nhàng.

b) Phân bổ mẫu tỷ lệ với căn bậc hai của quy mô tổng thể:

Công thức tính số đơn vị mẫu (n_t) của tổ t như sau:

$n_j = n. w_j$

Trong đó:

N: Số đơn vị của tổng thể

w_j: Tỷ lệ giữa căn bậc hai số đơn vị của tổ t ($\sqrt{N_j}$) và tổng căn bậc hai số đơn vị

của tất cả các tổ ($\sum_{j=1}^{k}\sqrt{N_j}$).

Như vậy công thức sẽ biến đổi như sau:

$$n_j = n.w_j = \frac{\sqrt{N_j}}{\sum_{j=1}^{k}\sqrt{N_j}}$$

Cách phân bổ này sẽ khắc phục nhược điểm của phân bổ tỷ lệ với quy mô tổng thể nhưng khi suy rộng phải tính lại theo quyền số thực tế.

c) Phân bổ Neyman:

Phân bổ Neyman được coi là phân bổ tối ưu theo nghĩa thống kê thuần tuý. Cỡ mẫu vừa tính theo tỷ lệ của quy mô, vừa tính đến sự khác nhau về độ biến động của chỉ tiêu nghiên cứu các tổ.

Công thức xác định cỡ mẫu (n_t) cho tổ t như sau:

$$n_j = n.\frac{N_j S_j}{\sum_{j=1}^{k} N_j S_j} \text{ với (j = 1, 2... k)}$$

Trong đó:

N_j - Tổng số đơn vị của tổ j;

S_j - Độ lệch chuẩn của tổ thứ j.

Công thức trên cho thấy quy mô mẫu của các tổ tỷ lệ thuận với quy mô và phương sai của chúng. Tổ có phương sai lớn sẽ được phân nhiều đơn vị mẫu hơn tổ có phương sai nhỏ, tổ có quy mô lớn sẽ được phân nhiều đơn vị hơn các tổ có quy mô nhỏ.

d) Phân bổ mẫu tối ưu:

Đây là cách phân bổ mẫu tối ưu đầy đủ hơn vì nó không những đề cập tới sự khác biệt về quy mô, sự biến động của chỉ tiêu được nghiên cứu giữa các tổ mà còn đề cập tới khả năng kinh phí của từng tổ. Công thức phân bổ mẫu tối ưu có dạng:

$$n_j = n.\left(\frac{N_j S_j / \sqrt{c_j}}{\sum_{j=1}^{k} N_j S_j / \sqrt{c_j}}\right) \text{ với j = 1, 2... k}$$

Trong đó: c_j - Chi phí điều tra cho tổ j.

Công thức trên cho thấy quy mô mẫu của các tổ tỷ lệ thuận với quy mô và phương sai của chúng. Mặt khác tỷ lệ nghịch với căn bậc hai của chi phí có thể có để thực hiện điều tra trên phạm vi của tổ. Vì vậy, phương pháp phân bổ mẫu này thường được áp dụng khi quy mô, phương sai và khả năng kinh phí của các tổ tương đối khác nhau.

e) Phân bổ mẫu có ưu tiên cho các tổ được đánh giá là quan trọng:

Cách phân bổ mẫu này thường được áp dụng khi có sự khác nhau đáng kể giữa các tổ về hàm lượng thông tin cần thiết. Theo nguyên tắc này, các tổ có hàm lượng thông tin thấp được phân bổ cỡ mẫu nhỏ. Tư tưởng này thường ứng dụng trong điều tra các doanh nghiệp. Các doanh nghiệp thuộc tổ có quy mô lớn (có sản lượng hoặc số lượng công nhân chiếm tỷ trọng lớn trong tổng sản lượng hoặc tổng số công nhân của các doanh nghiệp) thì phân bổ theo tỷ lệ mẫu lớn hơn. Ngược lại các doanh nghiệp có quy mô nhỏ hơn thì phân bổ tỷ lệ mẫu nhỏ hơn.

Tóm lại, phân bổ mẫu trong thực tế cần dựa vào việc phân tích đặc điểm cụ thể của các chỉ tiêu thống kê cần thu thập ở từng tổ. Mặt khác, cũng cần xét tới điều kiện thực tế diễn ra ở từng tổ. Điều này đặc biệt cần lưu ý trong khi phân bổ cỡ mẫu cho điều tra nhiều cấp.

5.3. Cách tính sai số chọn mẫu:

Dưới đây sẽ trình bày công thức tính sai số chọn mẫu tương ứng với các phương pháp tổ chức chọn mẫu ngẫu nhiên đơn giản, mẫu phân tổ, mẫu 2 cấp và mẫu chùm.

Cách trình bày công thức tính sai số chọn mẫu được bắt đầu từ một ví dụ giả định về danh sách các xã, phường (XP) ở địa phương Y với số hộ gia đình có vốn đầu tư cho sản xuất, kinh doanh như sau:

Bảng 8.1. **Danh sách các xã, phường ở địa phương Y với số hộ có đầu tư sản xuất, kinh doanh**

TT	Tên XP	Số hộ	Khu vực	TT	Tên XP	Số hộ	Khu vực
1	A	18	1	11	N	20	2
2	I	20	2	12	E	26	1
3	D	22	3	13	P	22	3
4	B	22	1	14	F	22	2
5	K	24	1	15	G	24	1
6	Y	24	2	16	Q	18	3
7	C	18	3	17	Z	20	2
8	L	20	2	18	J	16	1
9	V	22	1	19	H	26	1
10	M	20	1	20	S	28	2
				Tổng số		**432**	

a) Phương pháp tổ chức chọn mẫu ngẫu nhiên đơn giản:

* Tổ chức chọn mẫu:

Khi tiến hành chọn mẫu ngẫu nhiên đơn giản chỉ việc lập danh sách các hộ gia đình có tên chủ hộ, địa chỉ và kèm theo số thứ tự từ 1 đến 432 của chung 20 xã, phường kể trên. Sau đó dùng bảng số ngẫu nhiên hoặc rút thăm chọn ngẫu nhiên không lặp lại từ danh sách được lập trong bảng để được số hộ cần điều tra (ở đây là chọn 40 hộ).

* Cách tính sai số chọn mẫu:

Gọi i là số thứ tự của hộ gia đình trên địa bàn điều tra.

i = 1, 2..... N (N = 432 - Tổng số hộ của địa bàn điều tra)

i = 1, 2..... n (n = 40 - Số hộ chọn mẫu trên địa bàn)

x_i: Vốn đầu tư sản xuất, kinh doanh của hộ thứ i

Từ đó có công thức:

Vốn đầu tư bình quân một hộ:

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

Phương sai mẫu:

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

Sai số chọn mẫu:

$$\mu = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$$

b) Phương pháp tổ chức chọn mẫu phân tổ:

* Tổ chức chọn mẫu:

Trở lại ví dụ bảng 8.1 phân các xã, phường thành 3 khu vực 1,2,3. Các khu vực này có điều kiện kinh tế xã hội khác nhau và do đó có mức độ đầu tư cho sản xuất, kinh doanh của dân cư cũng khác nhau. Như vậy, việc phân chia các xã, phường theo khu vực sẽ liên quan nhiều đến vốn đầu tư cho sản xuất kinh doanh của dân cư.

Gọi j là số thứ tự của các tổ (j = 1, 2... k = 3 - Số tổ của địa bàn điều tra);

Tổ 1: j = 1 (khu vực 1); Tổ 2: j = 2 (khu vực 2); Tổ 3: j = 3 (khu vực 3)

N_j: Số hộ gia đình tổ (vùng) j

N: Tổng số hộ gia đình của địa bàn điều tra ($N = \sum_{j=1}^{k} N_j$)

n_j: Số hộ chọn mẫu của tổ (khu vực) j

n: Tổng số hộ chọn mẫu của địa bàn ($n = \sum_{j=1}^{k} n_j$)

Cỡ mẫu mỗi tổ (n_j) có thể được chọn theo tỷ lệ đều nhau hoặc chọn không theo tỷ lệ đều nhau. Nếu chọn theo tỷ lệ đều nhau thì tỷ lệ chọn mẫu ở các tổ đều bằng f ($f = \frac{n}{N}$).

* Cách tính sai số chọn mẫu:

Gọi i là số thứ tự của hộ gia đình trong mỗi tổ

i = 1,2..... N_j đối với tổng thể

i = 1,2..... n_j đối với tổng thể mẫu

x_{ij} - Vốn đầu tư của hộ thứ i thuộc tổ j

Từ đó ta có công thức tính:

Vốn đầu tư bình quân của các đơn vị thuộc tổ j:

$$\bar{x}_j = \frac{1}{n_j}\sum_{j=1}^{n_j} x_{ij}$$

Vốn đầu tư bình quân của tất cả các đơn vị điều tra:

- Chọn theo tỷ lệ:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{j} \bar{x}_j n_j$$

- Chọn không theo tỷ lệ:

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{j} \bar{x}_j N_j$$

Phương sai mẫu của các đơn vị trong tổ j:

$$s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2$$

Sai số chọn mẫu:

- Chọn theo tỷ lệ:

$$\mu = \sqrt{\frac{\bar{s}_j^2}{n}\left(1 - \frac{n}{N}\right)}$$

Trong đó: $\bar{s}_j^2 = \dfrac{\sum_{j=1}^{k} s_j^2 n_j}{\sum_{j=1}^{k} n_j}$

- Chọn không theo tỷ lệ:

$$\mu = \frac{1}{N}\sqrt{\sum_{j=1}^{k}\frac{s_j^2}{n_j}\left(1 - \frac{n_j}{N_j}\right)N_j^2}$$

c. Phương pháp tổ chức chọn mẫu 2 cấp:

* Tổ chức chọn mẫu:

Cũng số liệu đã cho ở bảng 8.1 tiến hành chọn mẫu 2 cấp như sau: từ danh sách 20 xã, phường chọn ngẫu nhiên không lặp lấy 4, tức là 20% số xã, phường (chẳng hạn chọn được các xã, phường số 1, 5, 12 và 19). Các xã phường được chọn là mẫu cấp I. Tiếp theo lập danh sách các hộ gia đình của 4 xã, phường này, rồi từ các danh sách đó chọn ngẫu nhiên không lặp ra số hộ đều nhau cho mỗi xã, phường (10 hộ) để tiến hành điều tra. Như vậy tổng số hộ được chọn là 40 (hộ là mẫu cấp II).

* Cách tính sai số chọn mẫu:

Gọi j là số thứ tự của đơn vị mẫu cấp I (xã, phường)

j = 1, 2, 3..., M (M = 20 - Tổng số xã, phường của địa bàn điều tra)

j = 1, 2, 3..., m (m = 4 - Số xã, phường được chọn vào mẫu cấp I)

i: Số thứ tự của đơn vị cấp II (Hộ gia đình)

n: Tổng số đơn vị mẫu cấp II (Hộ gia đình)

n*: Số đơn vị mẫu cấp II trong mỗi đơn vị mẫu cấp I (các đơn vị mẫu cấp I có số đơn vị mẫu cấp II bằng nhau: n* = n: m).

x_{ij}: Vốn đầu tư của hộ gia đình (đơn vị mẫu cấp II) thứ i thuộc xã, phường (đơn vị mẫu cấp I) thứ j.

Ta có công thức tính:

Vốn đầu tư bình quân của các đơn vị mẫu cấp II thuộc mẫu cấp I thứ j:

$$\bar{x}_j = \frac{1}{n^*}\sum_{i=1}^{n^*} x_{ij}$$

Vốn đầu tư bình quân của tất cả các đơn vị điều tra:

$$\bar{x} = \frac{1}{m}\sum_{j=1}^{m}\bar{x}_j = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n^*} x_{ij}$$

Phương sai mẫu cấp II (hộ) thuộc từng đơn vị mẫu cấp I (xã, phường) thứ j:

$$s_j^2 = \frac{1}{(n^* - 1)}\sum_{i=1}^{n^*}\left(x_{ij} - \bar{x}_j\right)^2$$

Bình quân các phương sai mẫu cấp II:

$$\bar{s}_j^2 = \frac{1}{m}\sum_{j=1}^{m} s_j^2$$

Phương sai mẫu cấp I:

$$s_b^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{x}_j - \bar{x}\right)^2$$

Sai số chọn mẫu:

$$\mu = \sqrt{\frac{s_b^2}{m}\left(1-\frac{m}{M}\right) + \frac{\bar{s}_j^2}{m.n^*}\left(1-\frac{n^*}{N^*}\right)}$$

Trong đó: Số đơn vị cấp II thực tế có bình quân trong mỗi đơn vị cấp I (N): N* = N: M.

d. Phương pháp tổ chức chọn mẫu chùm:

Trong mẫu chùm có hai loại: Mẫu chùm có kích thước bằng nhau và mẫu chùm có kích thước khác nhau. Sự khác nhau về kích thước của mẫu chùm liên quan đến sự khác nhau về cách tổ chức chọn mẫu và công thức tính các tham số chọn mẫu.

* Tổ chức chọn mẫu:

Tiếp tục nghiên cứu ví dụ 8.1. Nếu xác định chùm là một xã, phường và cũng tiến hành điều tra cỡ mẫu n = 40 hộ gia đình thì cách tiến hành như sau:

- Với cỡ mẫu có kích thước các chùm bằng nhau (do người tổ chức điều tra ấn định) thì số chùm (m) cần chọn được xác định bằng cách chia tổng số mẫu cần điều tra (n) cho số mẫu qui định trong một chùm (n*), tức là n: n* = m. Cũng với ví dụ trên, cần điều tra 40 hộ (n = 40) và giả sử qui định mỗi chùm chọn 20 hộ (n* = 20) thì số chùm (xã, phường) phải điều tra: m = 40: 20 = 2 chùm.

Sau khi xác định được số chùm cần chọn, ta lập danh sách tất cả các chùm rồi chọn ngẫu nhiên không lặp lại từ danh sách đã lập cho 2 chùm (xã, phường) để tiến hành điều tra thực tế các đơn vị thuộc các chùm đó.

- Với cỡ mẫu có kích thước các chùm khác nhau thì quá trình chọn mẫu được tiến hành qua các bước sau đây:

Chia tổng số hộ gia đình của địa bàn điều tra cho số xã, phường để xác định số hộ bình quân có trong một chùm:

N* = 432: 20 ≈ 22

Chia số mẫu (Hộ gia đình) cần chọn cho số hộ có trong một chùm để xác định số chùm cần điều tra (m):

m = 40: 22 ≈ 2 chùm

Trên cơ sở danh sách các xã, phường ở bảng 8.1, tiến hành chọn 2 chùm, rồi tổ chức điều tra thực tế toàn bộ số hộ gia đình của 2 chùm đó.

Khi chọn mẫu chùm có kích thước khác nhau để điều tra sẽ có những trường hợp sau đây:

Nếu ở 2 chùm có vừa đủ 20 hộ gia đình thì điều tra hết 20 hộ.

Nếu ở 2 chùm có số hộ gia đình lớn hơn 20 thì điều tra hết 20 hộ, số dư ra bỏ lại không điều tra tiếp.

Nếu ở 2 chùm có số hộ gia đình nhỏ hơn 20 thì điều tra hết số hộ gia đình ở 2 xã, phường đã chọn. Sau chọn thêm một xã, phường thứ ba trong số 18 xã, phường còn lại và điều tra thêm số hộ cho đủ 20.

* Cách tính sai số chọn mẫu:

Gọi j là thứ tự các chùm (xã, phường), ở đây: j = 1, 2, 3..., M (M = 20 - toàn bộ số xã, phường có trong địa bàn điều tra) và j = 1, 2, 3..., m (m = 2 - số chùm chọn mẫu).

Gọi i là số thứ tự của hộ gia đình, ở đây i = 1, 2, 3..., n_j (n_j là số hộ có của một chùm – xã, phường).

Trong đó: $\sum_{j=1}^{m} n_j = n$ (n là số mẫu điều tra).

Nếu chọn mẫu chùm có kích thước bằng nhau thì các n_j bằng nhau và bằng n * (n* là số đơn vị trong một chùm).

Gọi x_{ij}: Vốn đầu tư của hộ thứ i thuộc chùm j.

Ta có công thức tính cho hai trường hợp:

- Chùm có kích thước bằng nhau:

Vốn đầu tư bình quân của các đơn vị trong mỗi chùm thứ j.

$$\bar{x}_j = \frac{1}{n^*}\sum_{i=1}^{n^*} x_{ij}$$

Vốn đầu tư bình quân của tất cả các đơn vị điều tra

$$\bar{x} = \frac{1}{m}\sum_{j=1}^{m} \bar{x}_j$$

Phương sai giữa các chùm

$$s_b^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{x}_j - \bar{x}\right)^2$$

Sai số chọn mẫu

$$\mu = \sqrt{\frac{s_b^2}{m}\left(1 - \frac{m}{M}\right)}$$

- Chùm có kích thước khác nhau:

Vốn đầu tư bình quân của các đơn vị trong mỗi chùm thứ j:

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$$

Vốn đầu tư bình quân của tất cả các đơn vị điều tra:

$$\bar{x} = \frac{\sum_{j=1}^{m} \bar{x}_j n_j}{\sum_{j=1}^{m} n_j} = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{ij}$$

Phương sai giữa các chùm:

$$s_b^2 = \frac{1}{\left(n - \frac{n}{m}\right)}\sum_{j=1}^{m}\left(\bar{x}_j - \bar{x}\right)^2 n_j$$

II. SAI SỐ TRONG ĐIỀU TRA THỐNG KÊ

Trong điều tra thống kê có hai loại sai số: Sai số chọn mẫu (sai số do tính đại diện của số liệu vì chỉ chọn một bộ phận các đơn vị để điều tra) và sai số phi chọn mẫu (sai số thuộc về lỗi của các quy định, hướng dẫn, giải thích tài liệu điều tra, do sai sót của việc cân đong, đo đếm, cung cấp thông tin, ghi chép, đánh mã, nhập tin...) từ đây gọi là "sai số điều tra".

Sai số chọn mẫu chỉ phát sinh trong điều tra chọn mẫu khi tiến hành thu thập ở một bộ phận các đơn vị tổng thể (gọi là mẫu) rồi dùng kết quả suy rộng cho toàn bộ tổng thể. Sai số chọn mẫu phụ thuộc vào cỡ mẫu (mẫu càng lớn thì sai số càng nhỏ), vào độ đồng đều của chỉ tiêu nghiên cứu (độ đồng đều cao thì sai số chọn mẫu càng nhỏ) và phương pháp tổ chức điều tra chọn mẫu. Còn sai số điều tra xảy ra cả trong điều tra chọn mẫu và điều tra toàn bộ.

Trong thực tế công tác điều tra thống kê hiện nay, phương pháp chọn mẫu được áp dụng ngày càng nhiều và có hiệu quả. Số liệu thu được từ điều tra chọn mẫu ngày càng phong phú, đa dạng và phục vụ kịp thời các yêu cầu sử dụng. Bên cạnh đó chất lượng số liệu của điều tra chọn mẫu cũng còn những hạn chế nhất định. Có một số ý kiến hiện nay đánh giá không công bằng và thiếu khách quan về kết quả điều tra chọn mẫu, cho rằng số liệu chưa sát với thực tế vì chỉ điều tra một bộ phận rồi suy rộng cho tổng thể.

Tất nhiên cũng phải thấy rằng đã là điều tra chọn mẫu thì không thể tránh khỏi sai số chọn mẫu nhưng mức độ sai số chọn mẫu của phần lớn những chỉ tiêu trong các cuộc điều tra thống kê hiện nay thường là ở phạm vi cho phép nên chấp nhận được. Hơn nữa khi cần thiết ta có thể chủ động giảm được sai số chọn mẫu bằng cách điều chỉnh cỡ mẫu và tổ chức chọn mẫu một cách khoa học, tuân thủ đúng nguyên tắc chọn mẫu.

Điều đáng nói và cần quan tâm hơn trong điều tra thống kê chính là sai số phi chọn mẫu. Loại sai số này xảy ra ở cả ba giai đoạn điều tra, liên quan đến tất cả các đối tượng tham gia điều tra thống kê và ảnh hưởng đáng kể đến chất lượng số liệu thống kê.

Dưới đây sẽ đi sâu nghiên cứu về sai số phi chọn mẫu - sai số điều tra, xảy ra trong cả ba giai đoạn nhưng chỉ đề cập đến sai số liên quan tới những công việc, những đối tượng thường gặp nhiều hơn.

1. Sai số trong quá trình chuẩn bị điều tra thống kê:

Trong công tác điều tra thống kê, chuẩn bị điều tra giữ một vai trò cực kỳ quan trọng. Chất lượng của khâu chuẩn bị điều tra sẽ ảnh hưởng cả đến quá trình thu thập số liệu và cuối cùng là đến chất lượng của số liệu điều tra. Một cuộc điều tra được chuẩn bị kỹ lưỡng, chu đáo và đầy đủ sẽ là cơ sở đầu tiên để giảm sai số điều tra nhằm nâng cao chất lượng của số liệu thống kê.

***1.1. Sai số điều tra liên quan tới việc xác định mục đích, nội dung và đối tượng điều tra**:*

Xác định mục đích điều tra là làm rõ yêu cầu của cuộc điều tra phải trả lời những câu hỏi gì, đạt được những mục tiêu nào của công tác quản lý. Yêu cầu của mục đích điều tra phải rõ ràng, dứt khoát và đó chính là căn cứ để xác định nội dung cũng như đối tượng điều tra một cách đúng đắn, đầy đủ, phù hợp, không bị chệch hướng.

Cùng một đơn vị điều tra, nếu có mục đích điều tra khác nhau với cách tiếp cận thu thập thông tin khác nhau thì sẽ có nội dung cũng như đối tượng điều tra khác nhau.

Xác định đúng nội dung và đối tượng điều tra, một mặt làm cho số liệu thu thập được sẽ đáp ứng những yêu cầu sử dụng, số liệu đảm bảo "vừa đủ". Mặt khác, xác định đúng nội dung và đối tượng điều tra là cơ sở để thiết kế bảng hỏi một cách khoa học và có điều kiện thuận lợi để tiếp cận với đối tượng cung cấp thông tin, đảm bảo thông tin thu được phù hợp và phản ánh đúng thực tế khách quan.

Tóm lại việc xác định đúng mục đích, nội dung và đối tượng điều tra làm cho cuộc điều tra thực hiện đúng hướng, đúng yêu cầu là một trong những điều kiện tiên quyết để đảm bảo chất lượng số liệu, giảm sai số trong điều tra thống kê.

1.2. Sai số liên quan tới việc xây dựng các khái niệm, định nghĩa dùng trong điều tra:

Khái niệm, định nghĩa dùng trong điều tra giúp cho hiểu rõ nội dung, bản chất cũng như phạm vi xác định thông tin của số liệu thống kê cần thu thập.

Như ta đã biết thống kê nghiên cứu mặt lượng trong quan hệ mật thiết với mặt chất của hiện tượng kinh tế - xã hội số lớn. Chính các khái niệm, định nghĩa là phản ánh về mặt chất của hiện tượng, là cơ sở để nhận biết, phân biệt hiện tượng này với hiện tượng khác cũng như xác định phạm vi của hiện tượng nghiên cứu. Nếu khái niệm, định nghĩa chuẩn xác, rõ ràng, được giải thích đầy đủ, cặn kẽ là cơ sở để xác định và thu thập số liệu thống kê phản ánh đúng thực tế khách quan. Ngược lại nếu khái niệm, định nghĩa không đúng, mập mờ, thiếu rõ ràng thì việc xác định, đo tính (lượng hoá) hiện tượng sẽ bị sai lệch.

Ví dụ: Khi điều tra cán bộ khoa học công nghệ có trình độ sau đại học, xét về chất, sau đại học phải là những người đã tốt nghiệp và có bằng thạc sĩ, tiến sĩ và tiến sĩ khoa học. Trong thực tế có cuộc điều tra thống kê ở nước ta chỉ đưa ra khái niệm sau đại học còn chung chung, thiếu cụ thể. Điều này làm cho những người tham gia điều tra (kể cả điều tra viên lẫn đối tượng trả lời) hiểu khái niệm cán bộ khoa học công nghệ có trình độ sau đại học rất khác nhau. Một số ít người đã hiểu đúng với nghĩa trình độ sau đại học phải gồm những người có bằng thạc sĩ, tiến sĩ và tiến sĩ khoa học; phần đông còn lại đã hiểu không đúng và cho là sau đại học gồm những người đã tốt nghiệp đại học sau đó được đi thực tập sinh sau đại học và thậm chí còn cả những người đã tốt nghiệp đại học nhưng chỉ được đi tập trung để đào tạo bồi dưỡng thêm về nghiệp vụ một vài tháng.

Thực tế này đã làm cho số liệu điều tra được về cán bộ khoa học công nghệ có trình độ sau đại học tăng lên hơn hai lần so với số thực tế có tại thời điểm điều tra.

Như vậy, những lỗi trong việc xây dựng các khái niệm, định nghĩa và nội dung thông tin về tiêu thức, chỉ tiêu thống kê sẽ ảnh hưởng trực tiếp đến chất lượng số liệu thống kê. Đây là hiện tượng khá phổ biến trong điều tra thống kê ở nước ta hiện nay.

Để có số liệu tốt, giảm bớt sai số điều tra, một vấn đề có tính chất nguyên tắc đó là phải chuẩn hoá các khái niệm, định nghĩa về các tiêu thức, chỉ tiêu của điều tra thống kê. Đồng thời phải giải thích rõ ràng, đầy đủ và cụ thể hoá các khái niệm, định nghĩa cho phù hợp với từng cuộc điều tra riêng biệt.

1.3. Sai số điều tra liên quan tới thiết kế bảng hỏi, xây dựng các bảng danh mục và mã số dùng trong điều tra:

Trong điều tra thống kê, bảng hỏi là vật mang tin, là công cụ giúp điều tra viên điền thông tin hoặc đánh dấu, đánh mã vào các ô, dòng, cột phù hợp theo nội dung trả lời của các câu hỏi tương ứng với các tiêu thức ghi ở bảng hỏi dùng trong điều tra.

Nếu các câu hỏi phức tạp, khó hiểu, khó trả lời, khó xác định hoặc khó điền thông tin thì khi đó thông tin thu được sẽ kém chính xác, không đáp ứng yêu cầu của số liệu điều tra.

Cùng với bảng hỏi, các bảng danh mục và các mã số có vai trò quan trọng trong quá trình tổng hợp số liệu thống kê. Thông tin thu được dù đảm bảo độ tin cậy cần thiết, nhưng nếu bảng danh mục dùng cho điều tra không chuẩn xác, các mã số không rõ ràng, khó áp dụng dẫn tới việc đánh sai, đánh nhầm và tất nhiên như vậy số liệu tổng hợp sẽ bị sai lệch.

Để giảm sai số điều tra, bảng hỏi phải được thiết kế một cách khoa học, đáp ứng đầy đủ nhu cầu thông tin theo nội dung điều tra đã được xác định, bảo đảm mối liên hệ logic và tính thống nhất giữa các câu hỏi. Mặt khác, các câu hỏi phải đơn giản, dễ hiểu, dễ trả lời, dễ ghi chép, phù hợp với trình độ của điều tra viên và đặc điểm về nguồn thông tin của từng loại câu hỏi. Thiết kế bảng hỏi còn phải đảm bảo thuận lợi cho việc áp dụng công nghệ thông tin. Các bảng danh mục phải có nội dung phù hợp với những thông tin cần thu thập và được mã hoá một cách khoa học theo yêu cầu tổng hợp của điều tra. Danh mục vừa phải phù hợp với yêu cầu của từng cuộc điều tra, vừa phải đáp ứng và thống nhất với danh mục phục vụ cho tổng hợp chung của công tác thống kê. Nội dung bảng danh mục và cách mã hoá phải được giải thích đầy đủ và hướng dẫn cụ thể.

1.4. Sai số điều tra liên quan tới việc lựa chọn điều tra viên và hướng dẫn nghiệp vụ:

Điều tra viên là người trực tiếp truyền đạt mục đích, nội dung, yêu cầu điều tra đến các đối tượng cung cấp thông tin, đồng thời trực tiếp phỏng vấn, lựa chọn thông tin để ghi vào bảng hỏi (nếu là điều tra trực tiếp). Vì vậy, điều tra viên có vai trò rất quan trọng trong việc đảm bảo chất lượng số liệu trong điều tra.

Nếu điều tra viên không nắm vững mục đích của cuộc điều tra, không hiểu hết nội dung thông tin cần thu thập thì sẽ truyền đạt không đúng các yêu cầu cần thiết cho đối tượng trả lời. Ngay cả khi điều tra viên nắm được nghiệp vụ, nhưng nếu thiếu ý thức trách nhiệm, chỉ phỏng vấn và ghi chép cho xong việc, hoặc cách tiếp cận với đối tượng điều tra không tốt thì cũng sẽ dẫn đến kết quả số liệu điều tra thu được không theo ý muốn.

Như vậy, việc lựa chọn điều tra viên không tốt cũng là nguyên nhân không kém phần quan trọng làm cho sai số điều tra tăng lên, ảnh hưởng đến chất lượng số liệu. Vì vậy, muốn giảm bớt loại sai số điều tra này, cần tuyển chọn điều tra viên có trình độ nhất định, nắm được nghiệp vụ, có kinh nghiệm thực tế về điều tra thống kê, đồng thời phải có ý thức và tinh thần trách nhiệm cao.

Sau khi lựa chọn được điều tra viên cần tổ chức tập huấn nghiệp vụ đầy đủ và thống nhất. Trong lớp tập huấn bên cạnh giải thích biểu mẫu điều tra cần cung cấp thêm những kiến thức về xã hội, phổ biến những kinh nghiệm thực tế và cách tiếp cận đối tượng điều tra, cách ứng xử trong thực tế. Đối với các cuộc điều tra thống kê có nội dung phức tạp và quy mô lớn, cần tiến hành điều tra thử để kịp thời rút kinh nghiệm, đảm bảo hướng dẫn nghiệp vụ gắn với điều tra thực địa.

Trong điều tra chọn mẫu, khi hướng dẫn nghiệp vụ cần chỉ rõ lộ trình điều tra theo từng cấp chọn mẫu, xác định địa bàn điều tra, lập danh sách địa bàn và đối tượng điều tra chọn mẫu (có địa chỉ cụ thể), quy định rõ những trường hợp mất mẫu phải thay đổi như thế nào, thay đổi đến đâu để tránh tình trạng điều tra viên thay đổi mẫu tuỳ tiện theo ý chủ quan của họ, v.v...

2. Sai số trong quá trình tổ chức điều tra:

2.1. Sai số điều tra liên quan đến quan hệ giữa yêu cầu về nội dung thông tin và quỹ thời gian, các điều kiện vật chất cần cho thu thập số liệu:

Nếu trong các cuộc điều tra thống kê phải thu thập quá nhiều chỉ tiêu có nội dung thông tin phức tạp, tốn nhiều thời gian để giải thích, phỏng vấn và ghi chép; trong khi đó quỹ thời gian và kinh phí dành cho công việc này lại không tương xứng, làm cho điều tra viên không đủ điều kiện để tiếp cận tìm hiểu tình hình thực tế, giải thích một cách đầy đủ, cặn kẽ về mục đích, yêu cầu và nội dung điều tra, ..., cho người cung cấp thông tin thì có thể họ sẽ không khai báo, hoặc khai báo qua loa, sai với thực tế. Đặc biệt có những loại thông tin phải hồi tưởng thì càng không đủ thời gian để nhớ lại. Tất cả những điều đó làm cho số liệu thu thập được sai số nhiều, không phản ánh đúng thực tế khách quan.

Để nâng cao chất lượng số liệu thống kê, giảm sai số khi tổ chức điều tra, phải cân đối giữa nhu cầu thu thập thông tin với khả năng về điều kiện kinh phí và quỹ thời gian dành cho điều tra. Không nên tổ chức một cuộc điều tra đòi hỏi thu thập quá nhiều chỉ tiêu; đặc biệt phải giới hạn những chỉ tiêu thu thập quá khó và tính toán phức tạp. Hơn nữa tuỳ thuộc vào đặc điểm và nội dung thông tin của các chỉ tiêu khác nhau, thuộc các đối tượng khác nhau để có cách tiếp cận thu thập thông tin cho hợp lý. Có thể chỉ tiêu này cần thu thập từ những nội dung chi tiết rồi tổng hợp chung lại, nhưng chỉ tiêu kia chỉ cần lấy số liệu khái quát. Không nên cho rằng bất kỳ chỉ tiêu nào, nội dung thông tin nào cũng phải lấy từ số liệu chi tiết mới là chính xác.

***2.2. Sai số điều tra liên quan đến điều tra viên*:**

Như trên đã nói để nâng cao chất lượng số liệu, giảm sai số điều tra, một trong những yêu cầu là phải chọn những người điều tra đủ tiêu chuẩn về chuyên môn và tinh thần trách nhiệm.

Ngoài những yêu cầu trên, điều tra viên khi được phân công về địa bàn điều tra, còn đòi hỏi phải làm quen với địa bàn, tìm hiểu thực tế về phong tục, tập quán, về điều kiện đi lại, sinh hoạt của địa phương.

Khi điều tra, điều tra viên phải kết hợp được kiến thức chuyên môn về điều tra đã được hướng dẫn với tình hình thực tế ở địa bàn điều tra, vừa phải giữ đúng nguyên tắc quy định cho điều tra, vừa phải có được những xử lý linh hoạt và hài hoà. Phần lớn những thắc mắc của đối tượng điều tra, điều tra viên phải tự mình tìm ra hướng giải đáp. Chỉ những trường hợp cần thiết mới ghi lại để xin ý kiến về cách xử lý của cấp chỉ đạo cao hơn.

2.3. Sai số điều tra liên quan đến ý thức, tâm lý và khả năng hiểu biết của người trả lời:

Ở đây việc trả lời câu hỏi có thể không tốt do ba nguyên nhân thuộc người cung cấp thông tin như sau:

- Về ý thức của người trả lời: Nếu họ không có tinh thần trách nhiệm cao, cho là cung cấp thông tin thế nào cũng được, nói cho xong việc thì có thể khi điều tra, người cung cung cấp thông tin sẽ lấy lý do này, lý do khác để không trả lời hoặc trả lời không hết, không đúng sự thật. Không ít trường hợp người trả lời còn cố tình khai không đúng vì lợi ích kinh tế và mục đích khác.

- Về tâm lý, nhiều người cung cấp thông tin không muốn trả lời những câu hỏi liên quan đến đời tư, đến mức sống, đến sự bí mật kín đáo của họ, của đơn vị họ. Ví dụ: khi điều tra thu thập thông tin mức thu nhập của hộ gia đình, phần lớn các chủ hộ nhất là những người có thu nhập cao thường không muốn nói thật, nói hết mức thu nhập của mình. Một ví dụ khác một người phụ nữ đi phá thai trong trường hợp giấu gia đình họ sẽ không muốn khai vì không muốn cho những người thân trong gia đình biết đến.

- Về nhận thức của người trả lời, nhiều người do nhận thức có hạn, không thấy rõ được mục đích, yêu cầu điều tra, không hiểu được nội dung câu trả lời... do vậy họ không thể trả lời hoặc trả lời không đúng với yêu cầu câu hỏi.

Qua đây cho thấy, để giảm bớt sai số điều tra, điều tra viên phải có cách tiếp cận hợp lý với từng loại đối tượng điều tra, ngoài kiến thức chuyên môn còn phải hiểu biết về xã hội, giải thích cho người được phỏng vấn về mục đích, ý nghĩa, về nguyên tắc cung cấp và bảo mật thông tin riêng, về trách nhiệm và quyền hạn của người cung cấp thông tin, giải thích cho họ hiểu nội dung câu hỏi một cách thuận tiện nhất, gợi ý cho họ những cách trả lời để đi đến có được số liệu thật.

2.4. Sai số điều tra liên quan đến các phương tiện cân, đong, đo lường:

Tất cả các khâu khác chuẩn bị tốt, nhưng nếu các loại phương tiện như cân, thước đo, dụng cụ đo huyết áp... dùng cho các chỉ tiêu phải thực hiện kiểm tra, đo, đếm trực tiếp mà không được chuẩn bị tốt thì cũng sẽ sai sót dẫn đến sai số trong điều tra. Ví dụ: điều tra để xác định mức độ suy dinh dưỡng của trẻ em. Nếu ta dùng loại cân không chuẩn thì sẽ cân không chính xác, dẫn đến số liệu tổng hợp về tỷ lệ trẻ em suy dinh dưỡng sẽ không đúng, hoặc là cao hơn, hoặc là thấp hơn thực tế.

Như vậy, việc chuẩn bị tốt các phương tiện đo lường, sử dụng đơn vị đo lường tiêu chuẩn, tránh sử dụng đợn vị đo lường địa phương khi điều tra cũng là biện pháp cần thiết để giảm sai số điều tra.

3. Sai số liên quan đến quá trình xử lý thông tin:

Sai số điều tra còn có thể xảy ra vì sai sót trong khâu đánh mã, nhập tin trong quá trình tổng hợp, xử lý số liệu.

Số liệu thu về phải được kiểm tra sơ bộ trước khi đánh mã, nhập thông tin. Việc kiểm tra này có thể phát hiện ra những trường hợp hiểu đúng nhưng ghi chép sai như nhầm đơn vị tính: 1 cái ghi sai thành 1 ngàn cái, 1 đồng thành 1 ngàn đồng; điền sai vị trí của thông tin,

v.v. Bằng kinh nghiệm nghề nghiệp cũng như quan hệ logic tính toán giữa các câu hỏi, người kiểm tra có thể phát hiện được những loại sai sót kiểu này. Kiểm tra sơ bộ còn có thể phát hiện những trường hợp có "số liệu lạ" (quá cao hoặc quá thấp so với mức bình quân chung). Những loại sai sót trên đây nhân viên kinh tế có thể tự điều chỉnh hoặc nếu trong những trường hợp cần thiết phải kiểm tra xác minh lại. Làm tốt khâu kiểm tra sơ bộ cũng là công việc góp phần quan trọng để giảm sai số điều tra.

Cần kiểm tra sơ bộ công đoạn đánh mã và nhập thông tin. Số liệu ghi đúng, ghi đầy đủ được kiểm tra kỹ lưỡng, nhưng nếu đánh mã sai, hoặc nhập thông tin sai thì cũng dẫn đến kết quả tổng hợp sai.

Sai sót trong đánh mã có thể là lựa chọn mã không phù hợp với nội dung của thông tin, hoặc là do bảng mã không cụ thể, khó xác định, hoặc là khả năng liên hệ vận dụng mã của người đánh mã không tốt; , đánh mã sai (mã này lẫn với mã kia) hoặc có mã đúng nhưng lộn số (ví dụ 51 thành 15), v.v...

Để khắc phục sai sót trong khâu đánh mã, trước hết phải có bảng mã tốt, cụ thể, phù hợp với nội dung thông tin cần thu thập. Bên cạnh những mã cụ thể cần có những mã chung để cho người đánh mã có cơ sở vận dụng cho những trường hợp thực tế xảy ra nhưng chưa có mã trong danh mục mã cụ thể (gọi là các trường hợp khác). Mặt khác, người đánh mã phải được hướng dẫn đầy đủ về yêu cầu, nguyên tắc và kỹ thuật đánh mã, khi thực hiện phải biết vận dụng và xử lý linh hoạt nhưng tuyệt đối không được tuỳ tiện, người đánh mã còn kết hợp chặt chẽ với các bộ phận khác trong cùng khâu tổng hợp, xử lý số liệu.

Sau đánh mã là khâu nhập thông tin và khâu này cũng thường xuyên xảy ra sai số. Loại sai sót này thường xảy ra trong các trường hợp sau: Nhập tin đúp hoặc bỏ qua không nhập thông tin, nhập mã sai, ấn lộn số, v.v...

Để khắc phục những sai sót khi nhập tin, thông trước hết phải lựa chọn những nhân viên nhập tin có khả năng nhập tốt, ít nhầm lẫn, có tinh thần trách nhiệm cao, tuân thủ nghiêm túc những quy trình và nguyên tắc nhập thông tin đã được hướng dẫn thống nhất.

Trên góc độ công nghệ thông tin, phải có chương trình nhập hợp lý, khoa học, có được những lệnh cho phép tự kiểm tra để phát hiện những lỗi nhập thông tin.

Trong nhiều trường hợp phải phân công chéo để nhập thông tin hai lần rồi so sánh đối chiếu số liệu nhập để tìm ra những trường hợp không thống nhất thuộc về lỗi nhập thông tin.

Đối với các cuộc điều tra thống kê thực tế hiện nay, những lỗi nhập thông tin ảnh hưởng đến sai số điều tra không phải là nhỏ. Tuy nhiên, sai số do lỗi nhập thông tin, nếu có chuẩn bị tốt hoàn toàn có khả năng khắc phục.

TÀI LIỆU THAM KHẢO

Đào Hữu Hồ (2001), *Xác suất thống kê,* NXB Đại học quốc gia Hà Nội.

Hoàng Trọng, Chu Nguyễn Mộng Ngọc (2007), *Thống kê ứng dụng trong kinh tế xã hội,* NXB Thống kê.

Khoa dự báo và phát triển Trường Đại học Kinh tế Quốc dân (2003), *Giáo trình Dự báo phát triển kinh tế - xã hội,* NXB Thống kê.

Nguyễn Đình Hương (1999), *Thống kê ứng dụng trong quản lý*, NXB Thanh niên.

Nguyễn Khắc Minh (2002), *Các phương pháp phân tích và dự báo trong kinh tế,* NXB Khoa học và kỹ thuật.

Nguyễn Thành Long (2005), *Giáo trình dự doán kinh tế,* Đại học Đà Nẵng.

Trần Bá Nhẫn, Đinh Thái Hoàng (1998*), Lý thuyết thống kê ứng dụng trong quản trị, kinh doanh và nghiên cứu kinh tế*, NXB Thống kê.

Trần Ngọc Phát, Trần thị Kim Thu (2006), *Giáo trình Lý thuyết thống kê,* NXB Thống kê.

Viện Khoa học thống kê (2005), *Một số vấn đề phương pháp luận thống kê.*

Võ Thị Thanh Lộc (2000), *Thống kê ứng dụng và dự báo trong kinh doanh và kinh tế*, NXB Thống kê.

Joseph F. Healy (2002), *Statistics: A tool for Social Research*, Wadsworth Publishing Company (An International Thomson Publishing Company).