BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NÔNG NGHIỆP HÀ NỘI

PGS. TS. Ngô Thị Thuận (*Chủ biên*)
TS. Phạm Vân Hùng - TS. Nguyễn Hữu Ngoan

GIÁO TRÌNH
NGUYÊN LÝ THỐNG KÊ KINH TẾ

(Dùng cho sinh viên các ngành kinh tế, kế toán, kinh doanh và quản trị doanh nghiệp)

HÀ NỘI – 2006

LỜI NÓI ĐẦU

Thống kê là một trong các nghiệp vụ không thể thiếu được trong công tác quản lý nhà nước và quản trị kinh doanh của doanh nghiệp. Nó còn được sử dụng như một công cụ bắt buộc trong nghiên cứu khoa học và triển khai các hoạt động thực tiễn. Do vậy, nguyên lý thống kê kinh tế là môn học không thể thiếu được trong hầu hết các ngành đào tạo.

*Trước đây, công tác thống kê ở nước ta chủ yếu được áp dụng trong khu vực kinh tế nhà nước nhằm thu thập các thông tin phục vụ cho việc quản lý kinh tế, xã hội của các ngành, các cấp. Cùng với chính sách mở cửa và cải cách quản lý kinh tế, công tác thống kê ngày càng được chú trọng trong các doanh nghiệp ở tất cả các ngành. Để đáp ứng nhu cầu đào tạo ngày càng cao, phù hợp với xu thế "hội nhập và phát triển", Bộ môn Kinh tế lượng, Khoa Kinh tế & Phát triển nông thôn biên soạn giáo trình "**Nguyên lý thống kê kinh tế**".*

Giáo trình được biên soạn theo chương trình môn học đã được Hội đồng khoa học giáo dục Khoa Kinh tế & Phát triển nông thôn thông qua với phương châm chú trọng thực hành, gắn kết với thực tế, có ứng dụng và khai thác các phần mềm tin học thông dụng.

Giáo trình bao gồm các chương:

Chương I : Giới thiệu môn học
Chương II : Thu thập thông tin thống kê
Chương III : Tổng hợp và trình bày các dữ liệu thống kê
Chương IV : Thống kê mức độ của hiện tượng
Chương V : Điều tra chọn mẫu
Chương VI : Kiểm định thống kê
Chương VII : Thống kê biến động của hiện tượng
Chương VIII : Phân tích tương quan và hồi quy

Từng chương có các bài tập và một số bài có gợi ý cách giải.

*Tham gia biên soạn cuốn giáo trình "**Nguyên lý thống kê kinh tế**" gồm:*

- PGS.TS. Ngô Thị Thuận (chủ biên) và viết các chương I, II, III, IV, VI, VII.

- TS. Phạm Vân Hùng viết chương VIII

- TS. Nguyễn Hữu Ngoan cùng viết chương V.

*Chúng tôi hy vọng cuốn giáo trình "**Nguyên lý thống kê kinh tế**" sẽ phục vụ được đông đảo bạn đọc, các nhà nghiên cứu, các nhà doanh nghiệp và sinh viên các ngành kinh tế, kế toán, quản trị kinh doanh của các trường đại học có các ngành đào tạo này.*

Mặc dù các tác giả đã rất cố gắng, song do khả năng có hạn và cùng với những điểm mới bổ sung, nên nội dung giáo trình biên soạn khó tránh khỏi thiếu sót và những hạn chế nhất định.

Chúng tôi mong muốn nhận được các ý kiến đóng góp của bạn đọc để giáo trình ngày càng hoàn thiện hơn.

CÁC TÁC GIẢ

DANH MỤC CÁC CHỮ VIẾT TẮT

ĐBSCL	Đồng bằng sông Cửu Long
ĐHNNI	Đại học Nông nghiệp I
ĐVT	Đơn vị tính
HTX	Hợp tác xã
NN & PTNT	Nông nghiệp và Phát triển nông thôn
TSCĐ	Tài sản cố định
UBND	Uỷ ban nhân dân

Chương I
GIỚI THIỆU MÔN HỌC

1. SƠ LƯỢC VỀ SỰ RA ĐỜI VÀ PHÁT TRIỂN CỦA THỐNG KÊ HỌC

1.1. Khái niệm về thống kê

Trong thực tế sản xuất kinh doanh, cũng như trong đời sống kinh tế xã hội chúng ta thường sử dụng thuật ngữ "thống kê" như thống kê lại các công việc đã làm trong ngày, các số liệu đã có, các khoản thu, chi... Vậy thống kê học là gì? Trước khi xét đến khái niệm thống kê học, chúng ta quan sát các ví dụ sau:

Ví dụ 1: Kết quả chính thức điều tra mức sống hộ gia đình Việt Nam 2002 và kết quả sơ bộ khảo sát mức sống hộ gia đình Việt Nam năm 2004 của Tổng cục Thống kê về tỷ lệ hộ nghèo cho năm 2002 và 2004 theo chuẩn nghèo được Thủ tướng Chính phủ ban hành áp dụng cho giai đoạn 2006 - 2010 (200 nghìn đồng/người/tháng cho khu vực nông thôn, 260 nghìn đồng/người/tháng cho khu vực thành thị) như sau:

Biểu 1.1. Tỷ lệ hộ nghèo theo chuẩn mới

ĐVT: %

Diễn giải	Năm 2002	Năm 2004
Cả nước	23,0	18,1
Chia theo khu vực		
Thành thị	10,6	8,6
Nông thôn	26,9	21,2
Chia theo vùng		
Đồng bằng sông Hồng	18,2	12,9
Đông Bắc	28,5	23,2
Tây Bắc	54,5	46,1
Bắc Trung Bộ	37,1	29,4
Duyên hải Nam Trung Bộ	23,3	21,3
Tây Nguyên	43,7	29,2
Đông Nam Bộ	8,9	6,1
Đồng bằng sông Cửu Long	17,5	15,3

Số liệu bảng 1.1 cho thấy, tính chung cả nước tỷ lệ hộ nghèo đã giảm từ 23,0% năm 2002 còn 18,1% năm 2004.

Vùng Đồng bằng sông Hồng là một trong những vùng có tỷ lệ số nghèo giảm nhanh nhất, năm 2002 là 18,2%, năm 2004 chỉ còn 12,9%.

Vùng Tây Bắc tỷ lệ hộ nghèo cao nhất, năm 2002 là 54,5%, năm 2004 có giảm nhưng chậm vẫn còn 46,1%.

Vùng Đông Nam Bộ có tỷ lệ hộ nghèo ít nhất.

Ví dụ 2: Có tài liệu về diện tích, dân số của 13 tỉnh ở Đồng bằng sông Cửu Long (ĐBSCL) năm 2003 ở bảng 2.1.

Các số liệu ở bảng 2.1 cho biết: Đồng bằng sông Cửu Long gồm 13 tỉnh, với tổng diện tích là 39.763 km^2; 16,964 triệu dân và 10,164 triệu lao động trong độ tuổi. Bình quân số dân trên 1 đơn vị diện tích là 427 người/km^2. Kiên Giang là tỉnh có diện tích lớn nhất. Tỉnh có số dân đông nhất là An Giang. Thành phố Cần Thơ có diện tích đất ít nhưng số dân tương đối đông, nên mật độ dân số là cao nhất (807 người/km^2).

Bảng 2.1. Diện tích và dân số 13 tỉnh ĐBSCL năm 2003

TT	Tỉnh	Diện tích tự nhiên (km^2)	Dân số (người)	Mật độ (người/km^2)	Lao động trong tuổi (người)
1	TP Cần Thơ	1.390	1.121.141	807	696003
2	Hậu Giang	1.607	772.239	481	470.130
3	Tiền giang	2.367	1.655.000	699	1.178.000
4	Long An	4.493	1.381.305	307	823.119
5	Đồng Tháp	3.238	1.640.309	507	1.016.309
6	Bến Tre	2.322	1.348.137	581	841.726
7	Trà Vinh	2.215	1.009.643	456	606.493
8	Vĩnh Long	1.475	1.038.965	704	665.000
9	An Giang	3.406	2.155.121	633	1.038.520
10	Kiên Giang	6.269	1.623.834	259	832.859
11	Sóc Trăng	3.223	1.243.982	386	771.269
12	Bạc Liêu	2.547	784.462	308	486.366
13	Cà Mau	5.211	1.190.676	228	738.219
Cộng		39.763	16.964.814	427	10.164.696

Nguồn: Niên giám thống kê kinh tế - xã hội tỉnh ĐBSCL năm 2003

Từ các ví dụ nêu trên chúng ta có nhận xét sau:

- Các số liệu thể hiện trong các bảng là các số liệu thống kê. Các số liệu này thu thập được là dựa vào các tài liệu thống kê;

- Tài liệu thống kê có được do kết quả tổng hợp của các cơ quan từ xã - huyện - tỉnh - toàn quốc bằng cách ghi chép quá trình diễn biến trong sản xuất, trong đời sống xã hội, văn hoá... và lập các báo cáo hàng năm;

- Từ các tài liệu thống kê từng năm, ta có thể tính bình quân rồi so sánh giữa các giai đoạn thời gian khác nhau dựa vào số liệu của từng giai đoạn.

- Các số liệu thống kê cho phép đánh giá kết quả (bản chất) của các hiện tượng kinh tế xã hội của một đất nước ở từng năm và xu hướng phát triển của nó qua các năm (theo thời gian).

- Các số liệu này cũng gợi mở cho người sử dụng nó các biện pháp thúc đẩy quá trình sản xuất tốt hơn hoặc dự kiến khả năng đạt được trong giai đoạn tới.

Tóm lại: *Tất cả các công việc từ theo dõi diễn biến của các hiện tượng, ghi chép tài liệu - tổng hợp tài liệu ở phạm vi rộng hơn, phân tích rút ra kết luận về bản chất, tính quy luật và đề ra các biện pháp chỉ đạo... là một quá trình nghiên cứu thống kê.*

Như vậy, thống kê không chỉ là việc cộng dồn đơn thuần các số liệu sẵn có mà là cả một quá trình nghiên cứu theo trình tự nhất định có nội dung, mục đích và phương pháp khoa học để đáp ứng các nhu cầu của xã hội. Một cách tổng quát, chúng ta có thể đi đến khái niệm về thống kê như sau:

Thống kê học là hệ thống các phương pháp dùng để thu thập, xử lý và phân tích các con số (mặt lượng) của hiện tượng kinh tế-xã hội để tìm hiểu bản chất và tính quy luật vốn có của chúng (mặt chất) trong điều kiện thời gian và không gian cụ thể.

Như vậy, từ "Thống kê' có 2 nghĩa: Nghĩa thông thường là thu thập số liệu; nghĩa rộng là một môn khoa học về bố trí, hoạch định các quan sát và thí nghiệm; thu thập và phân tích các số liệu và rút ra kết luận về các số liệu đã phân tích. Do đó, thống kê được coi là một công cụ của nghiên cứu khoa học, quản lý kinh tế và quản lý xã hội. Đây chính là "bộ đồ nghề" của các nhà nghiên cứu và lãnh đạo.

1.2. Sơ lược về sự ra đời và phát triển của thống kê

Thống kê ra đời từ bao giờ và quá trình phát triển của nó ra sao? Để trả lời câu hỏi này các nhà khoa học chuyên nghiên cứu sự hình thành và phát triển của thống kê học đã đưa ra nhận định sau: *Thống kê học ra đời và phát triển theo yêu cầu của xã hội* . Để chứng minh cho nhận định này người ta thường điểm lại lịch sử phát triển của xã hội loài người qua các thời kỳ:

- Thời kỳ cộng sản nguyên thuỷ: Thời kỳ này chưa có sản xuất, chưa có sở hữu tư nhân về tư liệu sản xuất, của cải do thiên nhiên cung cấp và là của chung, loài người chưa có tính toán, nên chưa có nhu cầu về thống kê.

- Thời kỳ chiếm hữu nô lệ: Thời kỳ này, có sở hữu tư nhân về tư liệu sản xuất , đất, nông nô, có sản xuất, có dư thừa, của cải thuộc về người chiếm hữu tư liệu sản xuất (chủ nô) nên chủ nô hoặc trực tiếp hoặc gián tiếp ghi chép, tính toán những tài sản thuộc quyền chiếm hữu của mình như: Có bao nhiêu ruộng đất, trâu bò, nhà cửa... Thực tế có

di tích cổ mà người ta đã tìm thấy ở Trung Quốc, Hy Lạp, Ai Cập, La Mã... thì những ghi chép và tính toán này còn đơn giản, mang tính chất cộng dồn, trong phạm vi hẹp, có thể nói rằng mới là công việc sơ khai của thống kê.

- Thời kỳ phong kiến: Thời kỳ này, sản xuất phát triển hơn, sản phẩm nhiều hơn, phạm vi chiếm hữu tư liệu sản xuất mở rộng hơn nên yêu cầu tính toán nhiều hơn và phức tạp hơn.

Các tài liệu cũ cho biết, hầu hết các nước ở châu Âu, châu Á đã tổ chức việc đăng ký kê khai về ruộng đất, nhân khẩu, tài sản... Những công việc này đã thể hiện tính chất thống kê. Sản xuất nông nghiệp ngày càng phát triển, sản phẩm dồi dào dẫn đến nhu cầu trao đổi hàng hoá, các ngành nghề thủ công ra đời... từ đó công việc ghi chép mở rộng ra ngoài lĩnh vực mỗi ngành, nhưng thống kê học chưa được hình thành.

- Thời kỳ tư bản chủ nghĩa cũ: Thời kỳ này, lực lượng sản xuất phát triển hơn, các ngành sản xuất mới ra đời, công nghiệp, giao thông vận tải, thương nghiệp... Các hoạt động kinh tế xã hội ngày càng phức tạp hơn, sự phân công lao động xã hội cũng phát triển, phân chia giai cấp và đấu tranh giai cấp cùng gay gắt. Để phục vụ cho giai cấp thống trị, đòi hỏi phải theo dõi mọi mặt của xã hội (kinh tế, chính trị). Người ta đã đi sâu nghiên cứu về lý luận và phương pháp thu thập, tính toán các tài liệu sao cho phản ánh đúng hiện tượng và giúp cho người làm công tác quản lý kinh tế, quản lý xã hội điều hành tốt các công việc của mình.

Cuối thế kỷ 17, một số tài liệu sách báo của thống kê được xuất bản hoặc một số trường đã bắt đầu giảng môn lý luận thống kê. Năm 1660, H.Cohring - nhà kinh tế học người Đức giảng bài tại Trường đại học Holmsted về phương pháp nghiên cứu hiện tượng xã hội dựa vào số liệu điều tra cụ thể. Năm 1682, cuốn sách "Số học chính trị" của William Petty – nhà kinh tế học người Anh; năm 1759, G.Achen Wall (1719-1772) -giáo sư người Đức dùng từ "statistik", "status" (Thống kê). Ở thời kỳ này, sự phát triển của toán học, nhất là lý thuyết xác suất cũng rất mạnh mẽ đã góp phần trang bị thêm phương pháp tính toán và quản lý công việc của các nhà thống trị.

Trong hoàn cảnh đó, thống kê đã được hình thành. Như vậy, thống kê học hình thành vào cuối thế kỷ 17, đầu thế kỷ 18 và chủ nghĩa tư bản cũ đã tạo điều kiện cho thống kê ra đời và phát triển.

Nhưng trong xã hội có giai cấp, sự phân hoá giàu nghèo rất rõ rệt, đặc biệt là trong chiến tranh giữa các nước, các cường quốc, giai cấp thống trị thường sử dụng các tài liệu thống kê như một công cụ để phục vụ cho giai cấp mình, để xoa dịu đấu tranh giai cấp hoặc che dấu bí mật kinh doanh, nên họ thường đưa ra những tài liệu thống kê không trung thực và khách quan lắm. Vì lý do đó mà giai đoạn cuối của chủ nghĩa tư bản cũ (chủ nghĩa đế quốc) thống kê không phát huy được vai trò tiến bộ của mình.

- Thời kỳ hình thành và phát triển của hệ thống XHCN: Theo quan điểm của CNXH muốn cho toàn dân hiểu được thực tế khách quan về sản xuất, kinh tế và xã hội để mỗi người đều có trách nhiệm góp phần của mình vào việc thúc đẩy xã hội tiến lên, CNXH đã tạo điều kiện cho khoa học thống kê phát huy tác dụng tích cực và ngày càng hoàn thiện về lý luận và phương pháp để có thể phản ánh đúng thực tế khách quan xã hội.

- Ngày nay, do sự phát triển của xã hội loài người, do sự tiến triển của khoa học - kỹ thuật đòi hỏi khoa học thống kê cũng ngày càng hoàn thiện về lý luận, về phương pháp, có nhiều thông tin nhanh, phong phú, phương tiện tổng hợp tốt hơn, phương pháp phân tích, đánh giá và dự báo ngày càng hiện đại hơn...

Thống kê chính là một công cụ mạnh mẽ nhất để nhận thức xã hội. Tuy nhiên, tuỳ theo mục đích khác nhau mà thứ công cụ này phục vụ có khác nhau.

- Ở nước ta: Trong kháng chiến chống Pháp (1945-1954), chúng ta đã sử dụng công tác thống kê với các thành tựu của khoa học thống kê thế giới để lên án chế độ thực dân, phong kiến, động viên toàn dân làm kháng chiến thắng lợi. Cùng với sự phát triển của đất nước, thống kê học ngày càng hoàn thiện dần về mạng lưới thống kê, về phương pháp tổ chức, về kỹ thuật tổng hợp, phân tích. Song do nền kinh tế nước ta chưa ổn định, chuyển hướng liên tục... nên thống kê học ở nước ta còn có những hạn chế nhất định.

2. ĐỐI TƯỢNG NGHIÊN CỨU CỦA THỐNG KÊ

Các nhà thống kê học nổi tiếng trên thế giới đều thống nhất đưa ra nhận định sau đây về đối tượng nghiên cứu của thống kê.

Thống kê học là môn khoa học xã hội, nghiên cứu mặt lượng trong mối liên hệ chặt chẽ với mặt chất của các hiện tượng kinh tế- xã hội số lớn, trong điều kiện thời gian và địa điểm cụ thể.

Từ nhận định này, chúng ta cần hiểu đúng đối tượng nghiên cứu của thống kê ở các điểm chính sau.

2.1. Thống kê học là một môn khoa học xã hội

Thống kê học là một môn khoa học xã hội, bởi vì thống kê nghiên cứu các hiện tượng kinh tế - xã hội hay quá trình kinh tế xã hội. Các hiện tượng và quá trình đó thường là:

* Các hiện tượng về quá trình tái sản xuất mở rộng như cung cấp nguyên liệu, quy trình công nghệ, chế biến sản phẩm...

* Các hiện tượng về phân phối, trao đổi, tiêu dùng sản phẩm (marketing) như giá cả, lượng hàng xuất, nhập hàng hoá, nguyên liệu...

* Các hiện tượng dân số, lao động như tỷ lệ sinh, tử, nguồn lao động, sự phân bố dân cư, lao động...

* Các hiện tượng về văn hoá, sức khoẻ như trình độ văn hoá, số người mắc bệnh, các loại bệnh, phòng chống bệnh...

* Các hiện tượng về đời sống chính trị, xã hội, bầu cử, biểu tình...

* Ngoài ra thống kê còn nghiên cứu ảnh hưởng của các hiện tượng tự nhiên đến sự phát triển của các hiện tượng kinh tế xã hội, như ảnh hưởng của khí hậu, thời tiết, của các biện pháp kỹ thuật tới quá trình sản xuất nông nghiệp, kết quả sản xuất nông nghiệp và đời sống nhân dân.

2.2. Thống kê nghiên cứu mặt lượng trong mối liên hệ chặt chẽ với mặt chất của số lớn hiện tượng và quá trình kinh tế xã hội

a) Mặt lượng (những biểu hiện cụ thể, đo lường được):

* Quy mô của hiện tượng: Các mức độ to nhỏ, lớn bé, rộng hẹp.

Ví dụ: Diện tích canh tác của 1 doanh nghiệp nông nghiệp A năm 2005 là 500 ha, dân số trung bình của Việt Nam 2003 là 80,90 triệu người (Niên giám thống kê 2003), tổng số sinh viên của 1 lớp năm học 2005 - 2006 là 80 người.

* Kết cấu của hiện tượng: Hiện tượng tạo nên từ các bộ phận nào, mỗi bộ phận chiếm bao nhiêu %;

Ví dụ: Lớp có 50 học sinh, nam là 40 học sinh, chiếm 80%, nữ là 10, chiếm 20%.

* Tốc độ phát triển của hiện tượng: So sánh mức độ của hiện tượng theo thời gian để thấy mức độ tăng hay giảm của hiện tượng;

* Trình độ phổ biến của hiện tượng: Tính cụ thể phạm vi xảy ra hiện tượng, cá biệt hay phổ biến từ đó thấy được ảnh hưởng của nó tới hiện tượng lớn hơn.

Ví dụ: Tỷ lệ tai nạn giao thông xe máy năm 2004 là 2%, có nghĩa là cứ 100 người đi xe máy thì có 2 người tai nạn...

* Mối quan hệ tỷ lệ giữa các hiện tượng hoặc giữa các tiêu thức của cùng một hiện tượng.

b) Liên hệ chặt chẽ với mặt chất của số lớn hiện tượng:

* Thông qua các mặt lượng của hiện tượng để đánh giá bản chất của hiện tượng như quy mô to nhỏ, bộ phận nào nhiều hay ít, xu hướng tiến lên hay giảm đi, mức độ phổ biến của hiện tượng thế nào... nhưng để đánh giá một cách khách quan bản chất của hiện tượng thì mặt lượng của hiện tượng phải được thể hiện ở số lớn đơn vị chứ không phải ở từng đơn vị cá biệt.

Ví dụ, đánh giá kết quả học tập 2 sinh viên A, B cần dựa vào kết quả học tập nhiều học kỳ, nhiều môn; dựa vào ý thức phấn đấu, sự tham gia các phong trào đoàn, quan hệ bạn bè... Việc làm như vậy người ta gọi là nghiên cứu mặt lượng ở số lớn .

Nhưng để hiểu sâu sắc hơn bản chất của hiện tượng, người ta cũng nghiên cứu những đơn vị tiên tiến, hoặc lạc hậu là những biểu hiện cá biệt.

* Thống kê không nghiên cứu bản chất và quy luật của hiện tượng, mà thông qua mặt lượng có thể đánh giá được bản chất và tính quy luật của hiện tượng.

2.3. Thống kê nghiên cứu các hiện tượng và quá trình kinh tế xã hội trong điều kiện địa điểm và thời gian cụ thể

Mỗi hiện tượng, hay quá trình kinh tế xã hội ở thời gian, địa điểm khác nhau thì mặt lượng cũng khác nhau. Do đó, đối tượng nghiên cứu của thống kê học cũng cần cụ thể hoá ở thời gian nào, địa điểm nào hay trả lời câu hỏi bao giờ ? và ở đâu ?

3. PHƯƠNG PHÁP NGHIÊN CỨU CỦA THỐNG KÊ

3.1. Phương pháp luận của thống kê

- Khái niệm: Tổng hợp về mặt lý luận các phương pháp chuyên môn của thống kê gọi là phương pháp luận của thống kê học

- Cơ sở phương pháp luận: Dựa vào định luật số lớn trong lý thuyết xác suất đã xác định.

Định luật này được vận dụng và thể hiện là *quan sát số lớn các đơn vị cá biệt đến mức đủ lớn để có thể tổng hợp, phân tích, đánh giá bản chất khách quan và tính quy luật của hiện tượng*. Vì từ sự kiện cá biệt, ngẫu nhiên quan sát số lớn giúp chúng ta suy ra sự kiện chung. Qua tổng hợp số lớn, sự kiện cá biệt sẽ bù trừ cho nhau.

- Mức độ lớn phụ thuộc vào hiện tượng và mục đích nghiên cứu.

Phương pháp luận này của thống kê được thể hiện rất rõ trong các phương pháp chuyên môn của thống kê.

3.2. Các phương pháp chuyên môn của thống kê

- Điều tra thống kê: Điều tra toàn bộ, điều tra chọn mẫu, điều tra trực tiếp, điều tra gián tiếp;

- Tổng hợp thống kê: Hệ thống hoá các tài liệu, phân tổ thống kê.

- Phân tích thống kê: Phân tích mức độ, động thái, mối liên hệ...

3.3. Tính quy luật của thống kê

Tính quy luật của thống kê là tính quy luật số lớn các đơn vị trong đó có sự chênh lệch về lượng của từng đơn vị cá biệt. Tính quy luật này cũng phụ thuộc vào địa điểm và thời gian nhất định.

4. MỘT SỐ KHÁI NIỆM THƯỜNG DÙNG TRONG THỐNG KÊ

4.1. Tổng thể thống kê và đơn vị tổng thể

a) Tổng thể thống kê:

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

Xác định tổng thể là xác định phạm vi của đối tượng nghiên cứu. Tuỳ theo mục đích nghiên cứu mà tổng thể xác định có khác nhau.

Ví dụ, dân số trung bình của Việt Nam năm 2003 là 80,9 triệu người thì tổng số dân trung bình năm 2003 là tổng thể thống kê; hoặc số mẫu đất phân tích tính chất lý hoá để lập bản đồ nông hoá thổ nhưỡng của 1 xã năm 2004 là 300 mẫu thì tổng số mẫu đất cần phân tích năm 2004 là một tổng thể.

b) Đơn vị tổng thể:

Các đơn vị cá biệt (hay phần tử) cấu thành nên tổng thể thống kê gọi là đơn vị tổng thể. Tuỳ mục đích nghiên cứu mà xác định tổng thể và từ tổng thể xác định được đơn vị tổng thể.

Ví dụ (quay lại ví dụ trên): Đơn vị tổng thể là người dân, là từng mẫu đất. Đơn vị tổng thể bao giờ cũng có đơn vị tính phù hợp.

Đơn vị tổng thể là xuất phát điểm của quá trình nghiên cứu thống kê, bởi vì nó chứa đựng những thông tin ban đầu cần cho quá trình nghiên cứu. Trên thực tế có xác định được đơn vị tổng thể thì mới xác định được tổng thể. Thực chất xác định tổng thể là xác định các đơn vị tổng thể.

c) Các loại tổng thể thống kê:

* Tổng thể bộc lộ: Tổng thể trong đó bao gồm các đơn vị (hay phân tử) mà ta có thể quan sát hoặc nhận biết trực tiếp được.

Thí dụ: Tổng số sinh viên của Trường đại học Nông nghiệp I năm học 2005-2006.

* Tổng thể tiềm ẩn: Tổng thể trong đó bao gồm các đơn vị (hay phân tử) mà ta không thể quan sát hoặc nhận biết trực tiếp được.

Thí dụ: Tổng số sinh viên yêu ngành nông nghiệp.

* Tổng thể đồng chất: Tổng thể trong đó bao gồm các đơn vị (hay phân tử) giống nhau ở một hay một số đặc điểm chủ yếu có liên quan đến mục đích nghiên cứu.

Thí dụ: Sản lượng lúa của Việt Nam năm 2004.

* Tổng thể không đồng chất: Tổng thể trong đó bao gồm các đơn vị (hay phân tử) không giống nhau ở một hay một số đặc điểm chủ yếu có liên quan đến mục đích nghiên cứu.

Thí dụ: Sản lượng các loại cây hàng năm.

* Tổng thể mẫu: Tổng thể bao gồm một số đơn vị được chọn ra từ tổng thể chung theo một phương pháp lấy mẫu nào đó.

Thí dụ: Số sinh viên được chọn tham dự Đại hội Đảng bộ Trường ĐHNNI Hà Nội năm 2005 là 150 người.

4.2. Tiêu thức

Tiêu thức thống kê là chỉ đặc tính của đơn vị tổng thể.

Ví dụ, mỗi người dân có tiêu thức giới tính, độ tuổi, trình độ văn hoá, nghề nghiệp. Mỗi doanh nghiệp có các tiêu thức như số lao động, diện tích đất, vốn cố định, vốn lưu động...

Mỗi đơn vị tổng thể có nhiều tiêu thức. Mỗi tiêu thức có thể biểu hiện giống nhau hoặc khác nhau ở các đơn vị tổng thể.

Tiêu thức được phân chia thành các loại sau:

** Tiêu thức bất biến và tiêu thức biến động*

- Tiêu thức bất biến biểu hiện giống nhau ở mọi đơn vị tổng thể, căn cứ vào tiêu thức này người ta tập hợp các đơn vị tổng thể để xây dựng nên tổng thể.

Ví dụ: Tiêu thức quốc tịch "Việt Nam" xây dựng tổng số dân Việt Nam. Giới tính "nam", "nữ" xây dựng tổng thể dân số nữ, dân số nam.

- Tiêu thức biến động là tiêu thức biểu hiện của nó không giống nhau ở các đơn vị tổng thể. Ví dụ độ tuổi, trình độ văn hoá...

** Tiêu thức số lượng và tiêu thức chất lượng*

- Tiêu thức số lượng là tiêu thức thể hiện trực tiếp bằng con số. Ví dụ độ tuổi, mức lương...

- Tiêu thức chất lượng là tiêu thức thể hiện không bằng con số. Ví dụ giới tính, quốc tịch, trình độ ngoại ngữ.

* Tiêu thức thay phiên chỉ có 2 biểu hiện không trùng nhau. Thí dụ: giới tính, sinh tử...

* Chú ý: Có những tiêu thức thể hiện tương đối tổng hợp nhiều đặc tính của đơn vị tổng thể thì có thể trùng với chỉ tiêu thống kê như năng suất lúa, năng suất lao động, giá thành...

4.3. Lượng biến

Lượng biến là biểu hiện cụ thể về lượng của các đơn vị tổng thể theo tiêu thức số lượng.

Ví dụ: Độ tuổi 3, 4, 5, 10, 20 tuổi là lượng biến của tiêu thức độ tuổi, biểu hiện mức độ của tiêu thức số lượng.

Có hai loại lượng biến. Lượng biến rời rạc và lượng biến liên tục.

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

Thí dụ: Số công nhân trong một doanh nghiệp; số sản phẩm sản xuất ra trong một ngày của 1 phân xưởng may.

- Lượng biến liên tục: Là lượng biến mà các giá trị có thể có của nó được lấp kín cả một khoảng trên trục số.

Thí dụ: năng suất cây trồng; giá bán hàng hoá.

4.4. Chỉ tiêu thống kê

** Khái niệm*:

Chỉ tiêu thống kê là một khái niệm thể hiện tổng hợp mối quan hệ giữa lượng và chất của hiện tượng hay quá trình kinh tế xã hội trong điều kiện địa điểm và thời gian cụ thể.

** Đặc điểm của chỉ tiêu thống kê:*

- Phản ánh kết quả nghiên cứu thống kê.

- Mỗi chỉ tiêu thống kê phản ánh nội dung mặt lượng trong mối liên hệ với mặt chất về một khía cạnh, một đặc điểm nào đó của hiện tượng.

- Đặc trưng về lượng biểu hiện bằng những con số cụ thể, khác nhau trong điều kiện thời gian và địa điểm cụ thể, có đơn vị đo lường và phương pháp tính đã quy định.

Ví dụ: Tổng diện tích trồng trọt toàn quốc tính bình quân 3 năm 1989 - 1990 là 8.933.000 ha. Tổng diện tích gieo trồng toàn quốc là chỉ tiêu thống kê, nó có nội dung kinh tế, có ý nghĩa, có lượng là 8933000 ha, là một con số cụ thể gọi là số liệu thống kê, thời gian bình quân 3 năm 1989-1990, địa điểm toàn quốc, phương pháp tính bình quân, đơn vị tính ha.

** Các loại chỉ tiêu thống kê:*

- Chỉ tiêu thống kê khối lượng: Phản ánh quy mô về lượng của hiện tượng nghiên cứu. Ví dụ tổng số dân, diện tích gieo trồng, số học sinh.

- Chỉ tiêu chất lượng: Phản ánh các đặc điểm về mặt chất của hiện tượng như trình độ phổ biến, mức độ tốt xấu và quan hệ của các tiêu thức. Ví dụ giá thành, giá cả, hiệu quả sử dụng vốn.

* Hình thức đơn vị đo lường: Có 2 hình thức hiện vật và giá trị

- Chỉ tiêu hiện vật là chỉ tiêu thể hiện bằng các số liệu có đơn vị đo lường tự nhiên như cái, con, đơn vị đo chiều dài, trọng lượng.

- Chỉ tiêu giá trị là chỉ tiêu biểu hiện số liệu có đơn vị đo lường là tiền.

4.5. Hệ thống chỉ tiêu thống kê

Hệ thống chỉ tiêu thống kê là tập hợp nhiều chỉ tiêu có quan hệ mật thiết với nhau, có thể phản ánh nhiều mặt của hiện tượng hay quá trình kinh tế xã hội trong điều kiện thời gian và địa điểm cụ thể.

- Ai xác định? Tổng cục thống kê.

- Cho từng ngành và toàn nền kinh tế quốc dân.

- Nó được thay đổi và bổ sung, hoàn chỉnh trong các điều kiện lịch sử cụ thể.

5. CÁC LOẠI THANG ĐO

Để lượng hoá hiện tượng nghiên cứu, tuỳ theo tính chất của dữ liệu, thống kê đo lường bằng các loại thang đo sau.

5.1. Thang đo định danh

Thang đo định danh là thang đo dùng các mã số để phân loại các đối tượng. Thang đo dịnh danh không mang ý nghĩa nào cả mà chỉ để lượng hoá các dữ liệu cần cho nghiên cứu. Nó thường được sử dụng cho các tiêu thức thuộc tính. Người ta thường dùng các chữ số tự nhiên như 1, 2, 3, 4... để làm mã số.

Thí dụ: Giới tính: người ta thường mã số nam là 1; nữ là 2.

Tình trạng gia đình: 1: Độc thân ; 2: Kết hôn; 3: Ly dị; 4: Khác.

5.2. Thang đo thứ bậc

Thang đo thứ bậc là thang đo sự chênh lệch giữa các biểu hiện của tiêu thức có quan hệ thứ bậc hơn kém. Sự chênh lệch này không nhất thiết phải bằng nhau. Nó được dùng cho cả tiêu thức thuộc tính và tiêu thức số lượng.

Thí dụ:

- Tiền lương của công nhân trong doanh nghiệp hàng tháng là: < 800 ngàn đồng; từ 800-1000 ngàn đồng; từ 1000-1500 ngàn đồng và > 1500 ngàn đồng.

- Mức độ khó khăn của nông dân Việt Nam:

Thứ nhất : Thiếu vốn

Thứ hai: Thiếu kiến thức

Thứ ba: Thiếu lao động

....

5.3. Thang đo khoảng

Thang đo khoảng là thang đo thứ bậc có khoảng cách đều nhau. Nó được dùng cho cả tiêu thức thuộc tính và tiêu thức số lượng. Thang đo khoảng cho phép chúng ta đo lường một cách chính xác sự khác nhau giữa hai giá trị.

Thí dụ: Đề nghị sinh viên hãy cho biết ý kiến của mình về tầm quan trọng của các vấn đề sau đây trong dạy học ở đại học bằng cách khoanh tròn các con số tương ứng trên thang đánh giá chỉ *mức độ từ 1 đến 5 như sau:*

Các vấn đề	Không quan trọng	Bình thường	Rất quan trọng

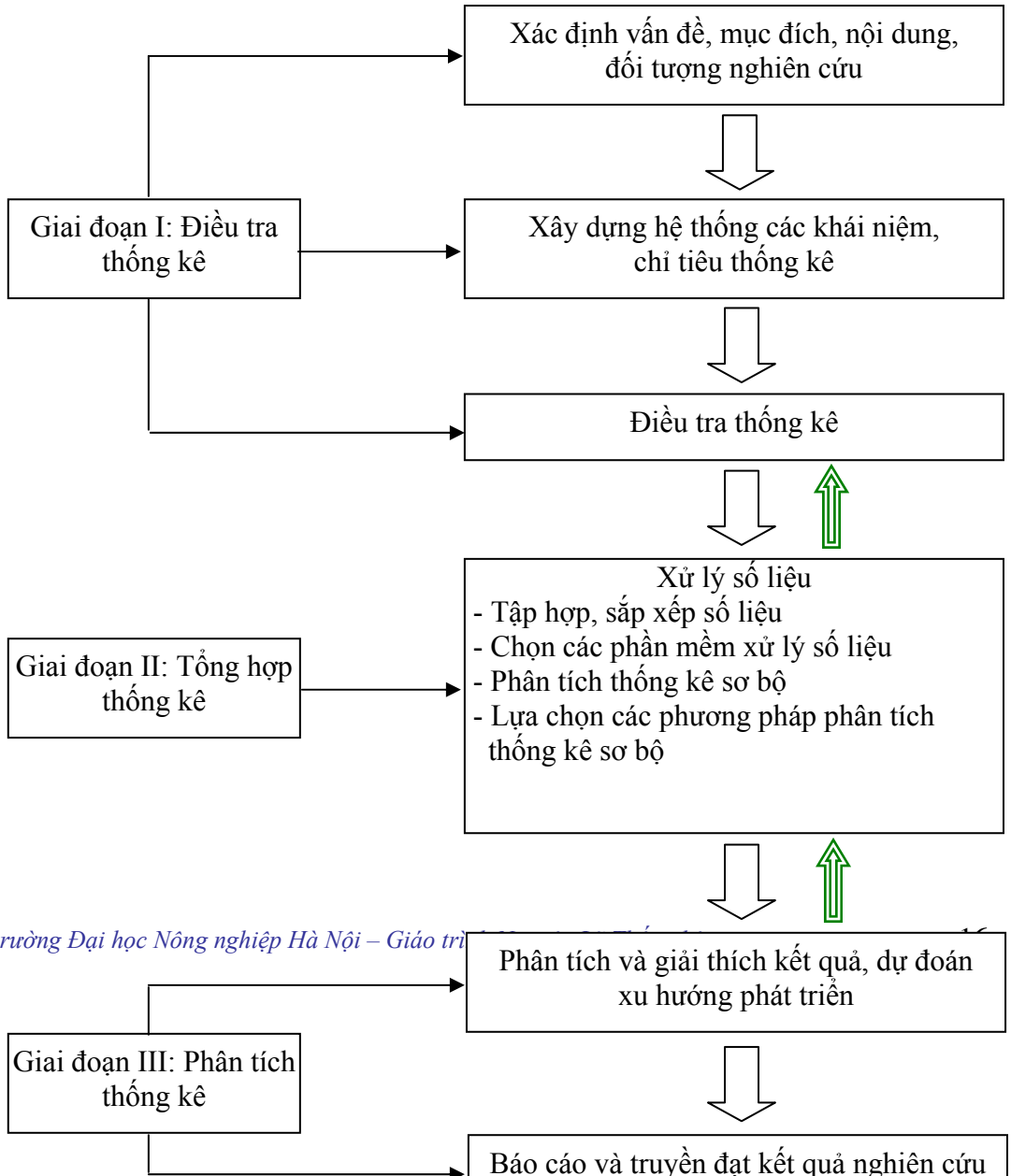
1. Giảng viên giỏi	1	2	3	4	5
2. Có giáo trình, tài liệu	1	2	3	4	5
3. Sự say sưa và ham học của sinh viên	1	2	3	4	5
4. Giảng đường hiện đại	1	2	3	4	5
5. Môi trường không khí trong lành	1	2	3	4	5

5.4. Thang đo tỷ lệ

Thang đo tỷ lệ là loại thang đo cao nhất trong thống kê. Nó sử dụng các số tự nhiên như từ 1 đến 9 và 0 để lượng hoá các dữ liệu. Nó được sử dụng chủ yếu cho các tiêu thức số lượng. Thí dụ: Doanh thu của một cửa hàng bán văn phòng phẩm Trâu Quỳ tháng 1/2005 là 200 triệu đồng; Nhiệt độ ngày 2/12/2005 là 23 ^{o}C.

Trong thực tế thang đo rất phức tạp và quan trọng, vì đôi khi chúng ta có thể áp dụng thang đo định tính cho tiêu thức số lượng và ngược lại.

6. KHÁI QUÁT QUÁ TRÌNH NGHIÊN CỨU THỐNG KÊ



Sơ đồ 1.1. Quá trình nghiên cứu thống kê

Quá trình nghiên cứu thống kê theo trình tự được khái quát hoá bằng sơ đồ 1.1. Theo sơ đồ này, quá trình nghiên cứu thống kê được chia thành 6 bước theo 3 giai đoạn với trình tự từ trên xuống. Hai mũi tên có hướng đi từ dưới lên nhằm chỉ rõ các cộng đoạn cần phải kiểm tra lại, bổ sung thông tin hoặc làm lại nếu dữ liệu chưa đạt yêu cầu

Giai đoạn I: Điều tra thống kê bao gồm thu thập các thông tin ban đầu về các tiêu thức ở từng đơn vị tổng thể;

Giai đoạn II: Tổng hợp thống kê bao gồm tổng hợp và hệ thống hoá các tài liệu đã thu thập được từ giai đoạn I;

Giai đoạn III: Phân tích thống kê nhằm sử dụng những phương pháp chuyên môn của thống kê để phát hiện các vấn đề làm cơ sở đề xuất các giải pháp.

Các bước và các giai đoạn này đều có mối liên hệ rất chặt chẽ. Kết quả và chất lượng kết quả của bước trước làm cơ sở và có ảnh hưởng đến chất lượng bước sau.

CÂU HỎI THẢO LUẬN CHƯƠNG I

1. Đối tượng nghiên cứu của thống kê là gì? Giải thích và chứng minh?
2. Các khái niệm thường dùng trong thống kê là gì? Giải thích, cho ví dụ cụ thể?

Chương II

THU THẬP THÔNG TIN THỐNG KÊ

Theo quá trình nghiên cứu thống kê, sau khi xác định được hướng, mục đích, nội dung và đối tượng nghiên cứu, thì việc thu thập các thông tin phục vụ cho quá trình nghiên cứu là bước rất cần thiết và quan trọng. Công việc thu thập thông tin đòi hỏi nhiều thời gian, công sức và chi phí cho nên việc thu thập thông tin cần được tiến hành một cách có hệ thống, theo một kế hoạch thống nhất để thu thập các thông tin sao cho vừa đáp ứng mục tiêu, nội dung và vừa phù hợp với khả năng nhân lực và kinh phí trong giới hạn cho phép.

1. THÔNG TIN THỐNG KÊ

1.1. Khái niệm và ý nghĩa

a) Khái niệm:

Thông tin là gì ? *Thông tin là một phạm trù được dùng để mô tả các tin tức của một hiện tượng, một sự vật, một sự kiện, một quá trình… đã xuất hiện ở mọi lúc, mọi nơi trong các hoạt động kinh tế- xã hội của con người.*

Thông tin thống kê là gì? *Thông tin thống kê là tin tức của hiện tượng hay quá trình kinh tế- xã hội do cơ quan thống kê thu thập trong điều kiện thời gian và không gian cụ thể.*

Như vậy, thông tin thống kê là một trong các loại thông tin, nên nó cũng mang những đặc trưng và giá trị của thông tin nói chung như: nội dung mới (không có cái mới thì không có thông tin); hình thức biểu hiện đa dạng (ngôn ngữ, con số, chữ viết); vật dẫn thông tin (sóng âm, trang giấy, băng đĩa từ) và có nội dung tin tức (thể hiện ý định, biểu đạt).

b) Ý nghĩa:

Thông tin thống kê là một nguồn lực của sản xuất kinh doanh, là nguồn lực vô giá. Nó có thể sử dụng cho nhiều mục tiêu và sử dụng nhiều lần. Với các giá trị này, khi sử dụng thông tin cần xử lí thông tin và xây dựng ngân hàng cơ sở dữ liệu cho nề nếp.

Thông tin thống kê cũng có các tính chất sau: khách quan, phụ thuộc, lan truyền, cùng hưởng, có hiệu lực, biến động, khuyếch tán và thu gọn.

Thông tin cần thu thập là gì?

Thông tin cần thu thập là những thông tin phục vụ cho vấn đề và mục đích cần nghiên cứu.

Xác định thông tin cần thu thập là xác định rõ những dữ liệu nào, thứ tự ưu tiên của các dữ liệu này và phạm vi dữ liệu cần thu thập.

Tại sao phải xác định thông tin cần thu thu thập?

Trong thực tế có rất nhiều thông tin liên quan đến hiện tượng hay quá trình kinh tế xã hội. Tuỳ theo vấn đề và mục tiêu nghiên cứu mà xác định những thông tin hay dữ liệu nào cần thiết. Do đó, vấn đề đầu tiên của công việc thu thập thông tin là xác định rõ và cụ thể những dữ liệu nào cần thu thập, thứ tự ưu tiên của các dữ liệu này. Nếu không thực hiện được điều này sẽ dẫn đến tình trạng dữ liệu thu thập được rất nhiều nhưng dữ liệu đáp ứng cho mục đích nghiên cứu thì ít hoặc thiếu, gây lãng phí thời gian, tiền bạc.

Thí dụ: Nghiên cứu mối liên hệ giữa tình hình tự học và kết quả học tập của sinh viên Đại học Nông nghiệp I Hà Nội, hai nhóm dữ liệu cần thu thập là: tình hình tự học và kết quả học tập. Về nhóm dữ liệu tình hình tự học, có thể thu thập các dữ liệu sau:

1. Có tự học ở nhà không?
2. Thời gian dành cho tự học ở nhà thế nào? (hàng ngày, hàng tuần)
3. Phương pháp sử dụng thời gian tự học ở nhà thế nào?
4. Mục đích tự học?
5. Hình thức tự học: học một mình, học nhóm ?
6. Khó khăn và thuận lợi khi tự học?
7. Kết quả và hiệu quả tự học?
8. Các yếu tố ảnh hưởng đến tự học.

Có nhiều dữ liệu khác có liên quan đến tự học, nhưng không liên quan lắm đến mục đích nghiên cứu "mối liên hệ giữa tự học với kết quả học tập" thì không nhất thiết phải thu thập. Thí dụ:

- Bạn thường mặc quần áo gì khi tự học?
- Người cùng học với bạn quê ở đâu?
- Bạn có uống nước hay ăn gì trong giờ tự học không?
- Ai nhắc nhở bạn tự học?

1.2. Các loại thông tin cần thu thập

Có nhiêu tiêu chí để phân loại thông tin. Tuỳ thuộc vào mục đích, ý nghĩa và phạm vi ứng dụng mà người ta có thể lựa chọn những tiêu thức phù hợp. ở đây trình bày một số phân loại thông tin được sử dụng chủ yếu trong nghiên cứu thống kê.

a) Căn cứ tính chất của thông tin:

Có hai loại dữ liệu chủ yếu là dữ liệu định tính và dữ liệu định lượng.

* Dữ liệu định tính là dữ liệu phản ánh tính chất và sự hơn kém về tính chất của đối tượng nghiên cứu. Thí dụ như giới tính của sinh viên (nam, hay nữ); thời gian tự học ở nhà dài hay ngắn (dưới 2 giờ; từ 2 đến 4 giờ; trên 4 giờ).

Dữ liệu định tính được thu thập dễ hơn và người ta thường dùng các thang đo định danh hay thứ bậc để xác định.

* Dữ liệu định lượng là dữ liệu phản ánh mức độ hay mức độ hơn, kém theo một tiêu thức số lượng nào đó của đối tượng nghiên cứu. Thí dụ như độ tuổi của sinh viên, thời gian tự học 1 ngày, 1 tuần.

Dữ liệu định lượng trong nghiên cứu thống kê thường gặp nhiều hơn, dễ áp dụng những phương pháp tính toán, phân tích hơn. Khi xác định các dữ liệu định tính, người ta thường dùng thang đo khoảng cách hay thứ bậc.

Mục đích của cách phân loại này nhằm giúp cho người nghiên cứu xác định trước các phương pháp xử lý, tổng hợp và phân tích cần sử dụng cho từng loại dữ liệu sao cho phù hợp và đáp ứng mục tiêu nghiên cứu đặt ra.

Thí dụ: Các dữ liệu và phương pháp phân tích có thể áp dụng trong nghiên cứu mối liên hệ giữa tự học và kết quả học tập của sinh viên cho ở bảng 1.2.

Bảng 1.2.

Tự học ở nhà/ngày	Kết quả học tập	Thang đo	Phương pháp phân tích
Định tính: - Dưới 2 giờ - Từ 2 đến 4 giờ - Trên 4 giờ	Định tính - Khá giỏi - Trung bình - yếu kém	Thứ bậc Định danh	Phân tổ
Định tính - Dưới 2 giờ - Từ 2 đến 4 giờ - Trên 4 giờ	Định lượng - Điểm trung bình chung học tập/1 sinh viên	Thứ bậc Khoảng cách	Phân tích phương sai 1 yếu tố
Định lượng - Số giờ tự học 1 tuần	Định lượng - Điểm trung bình chung học tập/1 sinh viên	Khoảng cách	Phân tích hồi quy và tương quan

b) Căn cứ nguồn cung cấp:

Theo nguồn cung cấp thông tin có hai loại dữ liệu: dữ liệu thứ cấp và dữ liệu sơ cấp.

* *Dữ liệu thứ cấp* là dữ liệu thu thập từ những nguồn có sẵn. Những dữ liệu này đã qua tổng hợp, xử lý công bố hay xuất bản.

Thí dụ: Những dữ liệu về kết quả học tập của sinh viên có thể lấy ở phòng đào tạo hay trợ lý đào tạo của từng khoa là dữ liệu thứ cấp.

Dữ liệu thứ cấp có ưu điểm là thu thập nhanh, rẻ nhưng thiếu chi tiết và đôi khi không đáp ứng đúng yêu cầu nghiên cứu.

Nguồn dữ liệu thứ cấp khá phong phú thường gặp ở các nguồn chủ yếu sau:

- Nội bộ: Các số liệu báo cáo về tình hình sản xuất, tiêu thụ, tài chính, vật tư, nhân sự... của các phòng ban, bộ phận; các số liệu báo cáo từ các cuộc điều tra khảo sát trước đây ở từng đơn vị (doanh nghiệp, cơ quan, ban, ngành...).

- Cơ quan thống kê nhà nước: Các số liệu do các cơ quan thống kê nhà nước (Tổng cục Thống kê, Cục Thống kê, Phòng Thống kê...) cung cấp trong các niên giám thống kê.

- Cơ quan chính phủ: Số liệu do các cơ quan trực thuộc Chính phủ (Bộ, cơ quan ngang bộ, Uỷ ban nhân dân các cấp) công bố hay cung cấp. Các số liệu này thường chi tiết hơn, mang tính chất đặc thù của ngành hay địa phương.

- Sách, báo, tạp chí đã xuất bản. Các số liệu này thường mang tính thời sự và cập nhật cao, mức độ tin cậy tuỳ thuộc vào nguồn số liệu của từng tờ báo hay tạp chí;

- Các tổ chức, hiệp hội, viện nghiên cứu, trường đại học;

- Các công ty nghiên cứu và cung cấp thông tin.

** Dữ liệu sơ cấp (thông tin gốc)* là dữ liệu không có sẵn, dữ liệu ban đầu thu thập trực tiếp từ đối tượng nghiên cứu.

Thí dụ: Các dữ liệu có liên quan đến việc tự học của sinh viên là các dữ liệu sơ cấp, không có sẵn mà chúng ta muốn có phải điều tra từ sinh viên.

- Dữ liệu sơ cấp có ưu điểm là chi tiết, độ tin cậy cao đối với các tình huống cụ thể. Song hạn chế của nó là thu thập tốn kém, phụ thuộc vào trình độ chủ quan của người nghiên cứu (nhất là những tình huống dự báo).

- Dữ liệu sơ cấp được thu thập bằng các cuộc điều tra khảo sát khác nhau.

Dựa vào tính chất liên tục hay không liên tục của thu thập dữ liệu sơ cấp, người ta chia thành 2 loại là điều tra thường xuyên và điều tra không thường xuyên.

+ Điều tra thường xuyên là loại điều tra nhằm thu thập các thông tin ban đầu về hiện tượng cần nghiên cứu một cách có hệ thống theo sát với sự biến động của hiện tượng.

Thí dụ: Ghi chép tình hình sinh, tử, chuyển đến, chuyển đi trong theo dõi và quản lý nhân khẩu của một địa phương. Việc theo dõi, ghi chép hàng ngày về số lượng công nhân đi làm, số lượng sản phẩm bán ra, mua vào... trong công ty thương mại (Bách hoá Trâu Quỳ).

Dữ liệu của điều tra thường xuyên làm cơ sở để lập báo cáo thống kê định kỳ.

+ Điều tra không thường xuyên là loại điều tra thống kê nhằm thu thập các dữ liệu ban đầu về hiện tượng nghiên cứu một cách không thường xuyên, không liên tục mà chỉ tiến hành khi có nhu cầu cần nghiên cứu.

Thí dụ: Điều tra dân số, điều tra thị trường, điều tra đất đai nông nghiệp, điều tra lao động và việc làm... .

Dữ liệu của điều tra không thường xuyên phản ánh trạng thái của hiện tượng tại một thời điểm nhất định. Nó có thể được tiến hành định kỳ (3 tháng, 6 tháng, 2 năm, 5 năm, 10 năm) hoặc không theo định kỳ.

Dựa theo phạm vi điều tra thống kê người ta chia thành 2 loại: Điều tra toàn bộ và điều tra không toàn bộ.

+ *Điều tra toàn bộ* là điều tra thống kê nhằm thu thập dữ liệu ban đầu ở tất cả các đơn vị tổng thể hiện tượng nghiên cứu (còn gọi là tổng điều tra, tổng kiểm kê). Ví dụ tổng điều tra dân số, tổng kiểm kê tài chính cuối năm, báo cáo kết quả học từng môn tất cả sinh viên học kỳ I, II.

Ưu điểm của điều tra toàn bộ là cung cấp dữ liệu khá đầy đủ, phong phú và đảm bảo tin cậy. Các dữ liệu này giúp ta tính toán các chỉ tiêu thể hiện quy mô, cơ cấu, biến động và dự đoán xu hướng biến động của hiện tượng.

Nhược điểm của điều tra toàn bộ là chi phí tốn kém, thời gian kéo dài, không áp dụng cho mọi trường hợp được và mức độ chính xác không đồng đều.

Điều tra không toàn bộ là điều tra thống kê nhằm thu thập dữ liệu ban đầu ở một số đơn vị của tổng thể hiện tượng nghiên cứu. Yêu cầu của điều tra không toàn bộ cần xác định rõ 3 vấn đề:

- Số đơn vị điều tra: Tuỳ theo yêu cầu và điều kiện nghiên cứu, người ta có thể chọn từ tổng thể hiện tượng nghiên cứu một số đơn vị để điều tra là nhiều hay ít.

- Phương pháp chọn số đơn vị mẫu điều tra: Chọn ngẫu nhiên hay phi ngẫu nhiên (lí thuyết xác suất).

- Các đơn vị được chọn ra phải đáp ứng được mục đích và yêu cầu nghiên cứu để kết quả điều tra có thể suy rộng cho tổng thể chung.

Ưu điểm của điều tra không toàn bộ là chi phí ít tốn kém, thời gian nhanh, khả năng thu thập tài liệu cũng tỉ mỉ, đảm bảo chính xác, kịp thời và áp dụng cho những trường hợp nghiên cứu mà hiện tượng đó không thể áp dụng điều tra toàn bộ.

Nhược điểm chủ yếu là tài liệu nếu thu thập từ các đơn vị điều tra được chọn không đáp ứng yêu cầu, mục đích nghiên cứu thì phản ánh không đúng thực tế khách quan. Vì vậy khâu chọn đơn vị điều tra rất quan trọng.

Ví dụ: Điều tra năng suất, sản lượng cây trồng, gia súc, điều tra chi phí, giá thành sản phẩm, điều tra mức sống, điều tra chất lượng sản phẩm.

Tuỳ theo cách chọn đơn vị điều tra mà điều tra không toàn bộ được chia thành 3 loại sau:

- Điều tra chọn mẫu: Loại điều tra chỉ tiến hành thu thập dữ liệu ở một số đơn vị được chọn ra từ tổng thể hiện tượng nghiên cứu. Các đơn vị này phải mang tính chất đại biểu cho tổng thể. Kết quả điều tra chọn mẫu có thể suy ra kết quả chung cho cả tổng thể.

Hiện nay đây là loại điều tra không toàn bộ khoa học nhất được áp dụng nhiều nhất trong nghiên cứu kinh tế - xã hội.

Ví dụ: Điều tra mức sống dân cư, điều tra kinh tế hộ, điều tra năng suất cây trồng...

- Điều tra trọng điểm: Loại điều tra chỉ tiến hành điều tra ở bộ phận tập trung lớn nhất của tổng thể hiện tượng nghiên cứu. Kết quả điều tra của bộ phận này không có ý nghĩa suy rộng mà chỉ dùng làm căn cứ để nhận định, đánh giá chung về các đặc điểm, nội dung chủ yếu của tổng thể.

Ví dụ: Điều tra tình hình sản xuất cây ăn quả đặc sản như nhãn lồng, vải thiều thì thực hiện chủ yếu ở vùng Hưng Yên, Lục Ngạn; cà phê, hạt tiêu chủ yếu ở Đắc Lắc.

- Điều tra chuyên đề: Loại điều tra chỉ tiến hành điều tra ở một hoặc một số đơn vị tổng thể điển hình (thường là một đơn vị tiên tiến hay lạc hậu) về một đặc tính nào đó, nghiên cứu tỉ mỉ và nhiều khía cạnh. Kết quả điều tra nhằm rút ra kinh nghiệm và phổ biến kinh nghiệm để có thể vận dụng chung cho các điều kiện tương tự.

Ví dụ: Điều tra báo cáo kết quả học tập, kinh nghiệm học tập, người tốt, việc tốt.

1.3. Chất lượng thông tin

Thông tin có thể được phát sinh, lưu trữ, truyền đi, được tìm kiếm, sao chép, xử lý và nhân bản. Mặt khác, thông tin cũng có thể biến dạng, sai lệch, hoặc bị phá huỷ. Vì vậy chất lượng thông tin có thể bị ảnh hưởng mà nguyên nhân là do:

- Các sự cố vật lí: Các sự cố về kỹ thuật gây ra hoặc sự cố về môi trường. Muốn khắc phục sự cố này cần kiểm tra kỹ thuật thường xuyên.

- Do ngữ nghĩa: Do ngôn ngữ mà xuất hiện những từ đồng âm dị nghĩa, đồng nghĩa khác âm hoặc ngôn ngữ bất đồng mà dẫn đến hiểu không đồng nhất về các khái niệm, văn phạm không rõ làm cho con người hiểu biết và nhận thức khác nhau về hiện tượng hay đối tượng nghiên cứu.

- Do tính thực dụng của con người: Xuất phát từ lợi ích nào đó trong quan hệ xã hội mà các thông tin đưa ra không chính xác, sai lệch sự thật. Nguyên nhân này xảy ra rất nhiều và thường xuyên trong nền kinh tế thị trường.

Trong nghiên cứu thống kê, thông tin là nguyên liệu đầu vào của mô hình phân tích nên rất cần những thông tin có ích.

Thông tin có ích là những thông tin có độ chính xác cao, độ bất định thấp. Thông tin có ích là thông tin có chất lượng phải đảm bảo 3 yêu cầu: đầy đủ, chính xác và kịp thời.

* Đầy đủ: Đủ, đúng các nội dung, các đơn vị hoặc các hiện tượng thuộc phạm vi nghiên cứu. Yêu cầu này có thể bị ảnh hưởng của cả 3 nguyên nhân nói trên.

* Chính xác: Phản ánh đúng thực tế tình hình các đơn vị, các nội dung mà con người cần biết. Yêu cầu này bị ảnh hưởng bởi tất cả các nguyên nhân.

* Kịp thời: Thông tin phản ảnh đúng lúc mà con người cần sử dụng.

2. PHƯƠNG PHÁP THU THẬP DỮ LIỆU BAN ĐẦU

2.1. Hình thức tổ chức thu thập dữ liệu ban đầu

Có hai hình thức tổ chức thu thập các dữ liệu ban đầu là báo cáo thống kê định kỳ và điều tra chuyên môn.

a) Báo cáo thống kê định kỳ:

* Khái niệm: Là hình thức tổ chức thu thập dữ liệu ban đầu một cách thường xuyên, định kỳ theo hình thức, nội dung, phương pháp và chế độ báo cáo đã quy định.

Ví dụ: Báo cáo kết quả thi và kiểm tra môn học của sinh viên; báo cáo tài chính cuối tháng, cuối năm; báo cáo số người đi làm từng ngày...

* Yêu cầu của báo cáo thống kê định kỳ: Đúng biểu mẫu, đúng kỳ hạn, nội dung có thể mở rộng hoặc thu hẹp...

* Phạm vi áp dụng: Hình thức này áp dụng chủ yếu cho các doanh nghiệp nhà nước, hoặc đối với các hiện tượng và quá trình kinh tế xã hội do địa phương hay nhà nước quản lý. Trong nền kinh tế thị trường, hình thức này áp dụng chủ yếu trong nội bộ doanh nghiệp.

* Cách lập các báo cáo thống kê định kỳ: Báo cáo thống kê định kỳ được lập theo trình tự sau:

- Mỗi cơ sở sản xuất tổ chức theo dõi quá trình sản xuất, ghi chép các diễn biến của nó vào các sổ sách. Công việc này được gọi là ghi chép ban đầu.

Ví dụ: Ghi các khoản thu, chi hàng ngày, phiếu xuất kho, phiếu thu, chi, bảng chấm công...

- Đến thời hạn báo cáo, người ta tập hợp các tài liệu ban đầu theo nội dung và phương pháp tính được chỉ dẫn trong báo cáo. Bản giải thích các biểu mẫu báo cáo thống kê định kỳ do Tổng cục Thống kê ban hành.

- Ghi các số liệu vào biểu mẫu và báo cáo.

- Các báo cáo này được lưu trữ nhiều năm, khi cần nghiên cứu người ta có thể lấy tài liệu từ các báo cáo đó phục vụ cho mục đích nghiên cứu.

b) Điều tra chuyên môn:

* Khái niệm: Là hình thức tổ chức thu thập các dữ liệu ban đầu không thường xuyên, không định kỳ mà tiến hành theo một kế hoạch và phương pháp quy định riêng cho mỗi lần điều tra.

- Điều tra chuyên môn chỉ thu thập tài liệu vào thời kỳ hoặc thời điểm có yêu cầu nghiên cứu. Ví dụ: Điều tra dân số, điều tra gia súc, điều tra tội phạm...

Các cuộc điều tra chuyên môn trên phạm vi toàn quốc như điều tra dân số, điều tra tình hình kinh tế và đời sống nông thôn, điều tra năng lực sản xuất công nghiệp của các thành phần kinh tế ngoài quốc doanh, thường gọi là tổng điều tra.

* Phạm vi áp dụng: Dùng để thu thập tài liệu về các vấn đề mà báo cáo thống kê định kỳ không thu thập hoặc không thể thu thập được. Cụ thể là các hiện tượng nằm ngoài kế hoạch, hoặc ít liên quan đến kế hoạch, các hiện tượng xảy ra bất thường và chủ yếu đối với các xí nghiệp ngoài quốc doanh như các tập đoàn tư nhân, các gia đình và cá nhân có doanh nghiệp riêng.

Đối với nông nghiệp nước ta, từ khi thực hiện Chỉ thị khoán 10 của Bộ Chính trị, hình thức này áp dụng phổ biến nhằm thu thập các thông tin ban đầu phục vụ cho lãnh đạo và chỉ đạo sản xuất nông nghiệp của Đảng và chính quyền các cấp.

* Ý nghĩa:

- Tài liệu thu thập rộng khắp và phong phú hơn.
- Kiểm tra chất lượng các báo cáo thống kê định kỳ.

* Tổ chức điều tra chuyên môn: Tiến hành một điều tra chuyên môn, người ta thường xây dựng phương án điều tra gồm các nội dung sau:

- Mục đích yêu cầu
- Đối tượng điều tra
- Nội dung điều tra và giải thích cách ghi chép
- Kế hoạch tiến hành.

2.2. Phương pháp thu thập dữ liệu ban đầu

a) Phương pháp trực tiếp:

Theo phương pháp này, người làm công tác điều tra phải tự mình trực tiếp quan sát, phỏng vấn thực tế, cân, đong, đo đếm và tự ghi chép tài liệu.

Ví dụ: Trong điều tra dân số, theo dõi thí nghiệm, điều tra năng suất cây trồng, khối lượng gia súc người điều tra đều phải trực tiếp phỏng vấn, đo, đếm để thu thập dữ liệu.

Ưu điểm của phương pháp này là tài liệu đảm bảo chính xác nên thường được áp dụng phổ biến. Tuy nhiên phương pháp này có nhược điểm chủ yếu là tốn nhiều kinh phí (cả về nhân lực và thời gian).

b) Phương pháp gián tiếp:

Theo phương pháp này, người điều tra thu thập tài liệu theo các nội dung cần nghiên cứu phải thông qua một phương tiện trung gian như điện thoại, thư tín, hoặc các chứng từ sổ sách đã ghi chép ở thời gian trước. Ví dụ điều tra thu chi trong doanh nghiệp, điều tra tình hình sinh tử, điều tra tài sản...

Ưu điểm của phương pháp này là đỡ tốn kém, nhưng có nhược điểm là mức độ đầy đủ và chính xác không cao, nên chỉ áp dụng trong những trường hợp khó khăn hoặc không có điều kiện thu thập trực tiếp.

3. KẾ HOẠCH THU THẬP DỮ LIỆU BAN ĐẦU

Để thu thập các dữ liệu ban đầu đảm bảo đầy đủ, khách quan và kịp thời thì điều tra thống kê cần được tổ chức một cách khoa học, thống nhất và chu đáo. Muốn vậy, trước khi tiến hành thu thập dữ liệu cần xây dựng kế hoạch.

Kế hoạch thu thập dữ liệu ban đầu (gọi tắt là kế hoạch điều tra) là một tài liệu dưới dạng văn bản, trong đó trình bày những nội dung, trình tự, phương pháp tiến hành, các công việc cụ thể cần chuẩn bị và tiến hành điều tra thống kê.

Đối với mỗi loại dữ liệu, cũng như mỗi hình thức tổ chức điều tra thống kê cần xây dựng kế hoạch điều tra phù hợp.

3.1. Dữ liệu thứ cấp

Nội dung cơ bản của kế hoạch thu thập dữ liệu thứ cấp cần trả lời các câu hỏi: Những tài liệu nào cần thu thập? Tài liệu đó ở đâu? cấp nào? được thể hiện qua ví dụ ở bảng 2.2.

Bảng 2.2. Nguồn gốc và phương pháp thu thập tài liệu thứ cấp

Cấp nào?	Ở đâu?	Tài liệu nào?
Bộ, tỉnh	Sở Kế hoạch và đầu tư	Các số liệu về các dự án, các chương trình phát triển kinh tế của cả nước, tỉnh...
	Cục Thống kê	Các số liệu thống kê về kinh tế, xã hội
	Sở Nông nghiệp & PTNT Cục Định canh, định cư	Các tài liệu về tình hình sản xuất nông nghiệp, nông thôn
	Sở Địa chính	Các tài liệu về đất đai
	Hiệp hội	Nông dân làm kinh tế giỏi
	...	...
Huyện	Phòng thống kê	Các số liệu thống kê về tình hình kinh tế xã hội của huyện
	Phòng nông lâm	Các số liệu về tình hình sản xuất nông nghiệp của huyện
	...	...
Trường, viện nghiên cứu	Trường ĐHNN I Hà Nội	Các đề tài nghiên cứu khoa học đã nghiệm thu Các luận văn, luận án đã bảo vệ Các kết quả ứng dụng tiến bộ khoa học...
Xã	UBND xã	Các tài liệu về tình hình kinh tế xã hội của xã
	Thôn, hộ nông dân	Các tài liệu của thôn, hộ nông dân

3.2. Dữ liệu sơ cấp

Để thu thập các dữ liệu sơ cấp, người ta thường tổ chức hình thức điều tra chuyên môn. Vì vậy, kế hoạch điều tra bao gồm các nội dung sau:

a) Xác định mực đích điều tra:

Xác định mục đích điều tra là nhằm thu thập những dữ liệu ở khía cạnh nào của hiện tượng, phục vụ cho yêu cầu nghiên cứu nào? và yêu cầu quản lý nào?

Mục đích điều tra là nội dung quan trọng đầu tiên của kế hoạch điều tra. Nó có tác dụng định hướng cho toàn bộ quá trình điều tra. Nó giúp chúng ta xác định đối tượng, đơn vị và nội dung điều tra.

Bất kỳ một hiện tượng nào khi nghiên cứu cũng được quan sát, tìm hiểu ở nhiều góc độ khác nhau. Song, trong điều tra thống kê thì không thể và không nhất thiết phải điều tra tất cả các khía cạnh của hiện tượng mà chỉ nên tập trung khảo sát những khía cạnh có liên quan trực tiếp, phục vụ yêu cầu nghiên cứu .

Thí dụ: Nghiên cứu các yếu tố ảnh hưởng đến kết quả học tập của sinh viên Đại học Nông nghiệp I. Mục đích điều tra là nhằm thu thập các dữ liệu phản ánh kết quả học tập của sinh viên từ 1-3 học kỳ gần đây và các yếu tố ảnh hưởng đến kết quả học tập. Các dữ liệu khác có liên quan đến sinh viên nhưng không cần thu thập như sinh viên quê quán ở đâu? Là con thứ mấy trong gia đình?

b) Xác định đối tượng điều tra và đơn vị điều tra:

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

Xác định đối tượng điều tra là quy định rõ phạm vi, ranh giới của hiện tượng nghiên cứu so với hiện tượng khác.

Trong thí dụ trên, đối tượng điều tra là các sinh viên đang học ít nhất 3 học kỳ gần đây của Trường Đại học Nông nghiệp I.

Xác định đối tượng điều tra đúng giúp chúng ta xác định đúng số đơn vị cần điều tra, tránh được những nhầm lẫn khi thu thập dữ liệu.

Để xác định đúng đắn đối tượng điều tra, cần dựa vào các căn cứ sau:

- Dựa vào mục đích điều tra.

- Các tiêu chuẩn phân biệt. Những tiêu chuẩn này chúng ta khi xác định đối tượng điều tra cần định nghĩa và đưa ra.

Thí dụ: Tiêu chuẩn đưa ra là sinh viên của Trường Đại học Nông nghiệp I đang học khác với đã học, học tập trung tại trường chứ không phải hệ vừa học vừa làm.

* Đơn vị điều tra: Là từng đơn vị cá biệt thuộc đối tượng điều tra và được xác định sẽ điều tra thực tế.

Trong điều tra toàn bộ, số đơn vị điều tra cũng chính là số đơn vị thuộc đối tượng điều tra. Trong điều tra không toàn bộ thì số đơn vị điều tra là những đơn vị được chọn ra từ tổng số các đơn vị thuộc đối tượng điều tra.

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

Tuỳ thuộc vào mục đích và đối tượng điều tra mà đơn vị điều tra được xác định khác nhau.

Thí dụ: Trong điều tra dân số, đơn vị điều tra là hộ gia đình và từng người dân; trong điều tra sản xuất và kinh doanh rau an toàn, đơn vị điều tra có thể là doanh nghiệp, hợp tác xã, hộ nông dân hoặc từng người dân có sản xuất và kinh doanh rau an toàn.

c) Nội dung điều tra:

Nội dung cần điều tra là những danh mục về các tiêu thức hay đặc trưng của các đơn vị điều tra cần thu thập.

Mỗi đơn vị điều tra có rất nhiều tiêu thức khác nhau. Nhưng trong mỗi cuộc điều tra dữ liệu sơ cấp không nhất thiết thu thập tất cả các tiêu thức, mà chỉ thu thập theo một số tiêu thức chủ yếu, những tiêu thức quan trọng nhất đáp ứng cho mục đích điều tra và mục đích nghiên cứu. Do đó, trong kế hoạch điều tra cần xác định và thống nhất danh mục các tiêu thức cần thu thập. Những danh mục này không thể thiếu khi tiến hành điều tra.

Thí dụ: Điều tra mức sống dân cư năm 2002 của Tổng cục Thống kê gồm các nội dung điều tra như sau:

- Tình hình cơ bản của các hộ gia đình
- Tình hình thu và cơ cấu các nguồn thu
- Tình hình chi và cơ cấu các khoản chi
- Tình hình thu nhập
- Ý kiến của hộ gia đình về khó khăn, thuận lợi, nguyện vọng.

Để xác định được đúng, đủ nội dung cần điều tra nên dựa trên các căn cứ sau:

- Mục đích nghiên cứu
- Mục đích điều tra
- Khả năng về nhân lực, chi phí và thời gian cho phép.

Mỗi tiêu thức trong danh mục các tiêu thức cần điều tra phải được diễn đạt thành câu hỏi ngắn gọn, dễ hiểu, cụ thể, rõ ràng để cả người điều tra và đơn vị điều tra đều hiểu một cách thống nhất.

d) Xác định thời điểm và thời kỳ điều tra:

* Thời điểm điều tra: Mốc thời gian được xác định để thống nhất đăng ký dữ liệu cho toàn bộ các đơn vị điều tra.

Thí dụ: Thời điểm điều tra dân số năm 1999 là 0 giờ ngày 1 tháng 04 năm 1999.

Xác định thời điểm điều tra là xác định cụ thể giờ, ngày để thống nhất đăng ký dữ liệu nhằm nghiên cứu trạng thái của hiện tượng tại thời điểm đó.

Tuỳ theo tính chất, đặc điểm của hiện tượng nghiên cứu mà xác định thời điểm điều tra. Tuy nhiên, khi xác định thời điểm điều tra người ta thường chọn thời điểm mà tại đó hiện tượng ít biến động nhất và gắn kết với những kế hoạch của địa phương.

Thí dụ: Điều tra thị trường áo bơi tại Việt Nam thì không thể chọn vào mùa đông.

* Thời kỳ điều tra: Khoảng thời gian được xác định để thống nhất đăng ký dữ liệu của các đơn vị điều tra trong suốt khoảng thời gian đó (cả ngày, cả tuần, 5 ngày, 10 ngày, 1 tháng, 3 tháng, 1 năm...).

Thí dụ: Điều tra số người vi phạm luật giao thông đường bộ 1 ngày, 1 tuần, 1 tháng của một địa phương.

Thời kỳ điều tra dài hay ngắn phụ thuộc vào mục đích nghiên cứu.

* Thời hạn điều tra: Là thời gian dành cho việc đăng ký thu thập tất cả các dữ liệu điều tra, được tính từ bắt đầu cho đến khi kết thúc toàn bộ công việc thu thập dữ liệu.

Thí dụ: Điều tra dân số, thời hạn điều tra trong vòng 10 ngày.

Điều tra số lượng áo bơi bán trên thị trường Hà Nội trong 1 tháng của Công ty may Thăng Long, thời hạn điều tra 5 ngày.

Như vậy, thời hạn điều tra dài hay ngắn phụ thuộc vào quy mô, tính chất phức tạp của hiện tượng, nội dung nghiên cứu và lực lượng tham gia, nhưng không nên quá dài.

e) Biểu mẫu điều tra và bản giải thích cách ghi biểu mẫu:

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

Yêu cầu của biểu mẫu điều tra là:

- Có đầy đủ các nội dung cần điều tra

- Các thang đo định tính sử dụng trong nội dung điều tra cần được mã hoá sẵn

- Các câu hỏi được thiết kế cụ thể, khoa học thuận lợi cho việc kiểm tra và tổng hợp dữ liệu.

* Bản giải thích cách ghi biểu mẫu là bản giải thích và hướng dẫn cụ thể cách xác định và ghi dữ liệu vào biểu mẫu điều tra. Nội dung, ý nghĩa của các câu hỏi phải được giải thích khoa học và chính xác. Những câu hỏi phức tạp có nhiều khả năng trả lời cần có ví dụ cụ thể.

Ngoài những nội dung chủ yếu nêu trên, bản giải thích còn đề cập tới một số vấn đề về phương pháp, cách tổ chức và tiến hành điều tra như sau:

- Cách chọn mẫu

- Phương pháp thu thập và ghi chép dữ liệu ban đầu
- Các bước và tiến độ điều tra
- Tổ chức và quy định nhiệm vụ của cán bộ tham gia điều tra
- Phân công khu vực điều tra
- Tổ chức tập huấn cán bộ điều tra
- Điều tra thử để rút kinh nghiệm
- Tổ chức tuyên truyền mục đích, ý nghĩa và yêu cầu của cuộc điều tra.

4. SAI SỐ TRONG THU THẬP DỮ LIỆU THỐNG KÊ

4.1. Khái niệm, ý nghĩa

Trong thu thập dữ liệu thống kê (gọi tắt là điều tra thống kê) dù tổ chức bằng hình thức nào, trong phạm vi nào và theo phương pháp nào bao giờ cũng chỉ đảm bảo yêu cầu chính xác với mức độ nhất định, hay nói cách khác dữ liệu thống kê thu thập được thường có sai số.

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

Sai số trong điều tra thống kê là sai số vốn có, được phép trong phạm vi sai số là 5%. Tuy nhiên, sai số càng lớn càng làm giảm chất lượng của kết quả điều tra và chất lượng của cả quá trình nghiên cứu thống kê. Vấn đề đặt ra trong điều tra thống kê là phải tìm ra các nguyên nhân làm phát sinh sai số để chủ động tìm biện pháp khắc phục.

4.2. Các loại sai số

* Sai số do đăng ký là loại sai số phát sinh do xác định và ghi chép dữ liệu không chính xác. Các nguyên nhân dẫn đến sai số này thường là:

- Lập kế hoạch điều tra sai hoặc không khoa học, không sát với thực tế của hiện tượng.
- Do trình độ của nhân viên điều tra không hiểu chính xác nội dung các câu hỏi, không biết cách khai thác số liệu.
- Do đơn vị điều tra không hiểu câu hỏi nên trả lời sai.
- Do ý thức, tinh thần trách nhiệm của cán bộ điều tra hoặc của đơn vị điều tra thấp dẫn đến việc xác định, cung cấp và ghi chép sai.
- Do dụng cụ đo lường không chính xác.
- Do công tác tuyên truyền, vận động không tốt dẫn đến đơn vị điều tra không hiểu hết hoặc hiểu sai mục đích điều tra nên cung cấp dữ liệu sai.
- Do thiếu tinh thần trung thực, khách quan nên cố tính cung cấp hoặc ghi chép sai dữ liệu.

- Do lỗi in ấn biểu mẫu, phiếu điều tra và bản giải thích sai.

- Những nguyên nhân khác...

* Sai số do tính chất đại biểu là sai số xảy ra trong điều tra không toàn bộ do chọn mẫu không đảm bảo tính chất đại diện.

Như vậy, nguyên nhân chính của sai số này là do việc lựa chọn đơn vị điều tra thực tế không có tính đại diện cao.

Thí dụ: Trong điều tra chọn mẫu về kinh tế hộ, 2 vấn đề đặt ra khi chọn các hộ là đơn vị điều tra là số lượng hộ là bao nhiêu? Kết cấu các loại hộ (khá, trung bình, nghèo)? Nếu chọn số hộ điều tra thực tế quá ít, kết cấu các hộ điều tra không phù hợp thì từ kết quả điều tra các hộ này suy rộng thành kết quả của tổng thể sẽ xuất hiện sai số do tính chất đại biểu.

4.3. Biện pháp chủ yếu khắc phục sai số trong điều tra thống kê

Sai số trong điều tra thống kê là sai số vốn có. Vì thế chúng ta chỉ tìm các biện pháp khắc phục tới mức thấp nhất các sai số nói trên trong điều tra thống kê. Các biện pháp chủ yếu là:

* Quán triệt mục đích ý nghĩa và yêu cầu từng cuộc điều tra. Cần tổ chức tốt công tác tuyên truyền cho đơn vị điều tra và nâng cao tinh thần trách nhiệm đối với cán bộ điều tra thông qua trang bị điều kiện làm việc, thời gian, thù lao và chế độ thưởng phạt.

* Làm tốt công tác chuẩn bị: Chọn, huấn luyện nhân viên, in ấn chính xác phiếu điều tra và các tài liệu hướng dẫn.

* Kiểm tra một cách có hệ thống các tài liệu thu thập được:

+ Kiểm tra tính logic của tài liệu.

+ Kiểm tra về mặt tính toán.

+ Kiểm tra tính đại biểu của đơn vị mẫu (cụ thể trong điều tra chọn mẫu).

CÂU HỎI THẢO LUẬN CHƯƠNG II

1. Thế nào là thông tin thống kê? Các loại thông tin thường dùng trong nghiên cứu kinh tế - xã hội?
2. Hãy nêu các phương pháp thu thập thông tin kinh tế - xã hội? Cho ví dụ ?
3. Chất lượng thông tin là gì? Các nguyên nhân ảnh hưởng đến chất lượng thông tin? Biện pháp khắc phục?

Chương III

TỔNG HỢP VÀ TRÌNH BÀY CÁC SỐ LIỆU THỐNG KÊ

1. TỔNG HỢP THỐNG KÊ

1.1. Khái niệm, ý nghĩa, nhiệm vụ

a) Khái niệm:

Kết quả của giai đoạn điều tra thông tin ban đầu cho chúng ta các dữ liệu thô về các đặc trưng riêng biệt của từng đơn vị tổng thể. Các dữ liệu này mang tính chất rời rạc, rất khó quan sát để đưa ra các nhận xét chung cho cả hiện tượng nghiên cứu và cũng không thể sử dụng ngay vào phân tích và dự báo thống kê được.

Ví dụ: Nghiên cứu tình hình trang bị máy tính của trường ta, ở giai đoạn điều tra thống kê cho ta những tài liệu ban đầu về từng đơn vị, số lượng máy, năm sản xuất, năm trang bị, nơi sản xuất, công suất, mã hiệu, hãng, tình trạng máy... Bây giờ chúng ta cần trả lời các câu hỏi:

- Trường có bao nhiêu máy tính?
- Mỗi khoa, phòng bao nhiêu?
- Loại máy, công suất?
- Nơi sản xuất?
- Khó khăn và thuận lợi?
-...

Muốn có được các tài liệu phản ánh chung cho cả tổng thể nghiên cứu như trên thì từ các thông tin riêng biệt của từng đơn vị chúng ta phải sắp xếp lại, hệ thống hoá, phân loại theo những tiêu thức cần nghiên cứu để thấy được các đặc trưng chung của tổng thể mẫu hay toàn bộ tổng thể nghiên cứu. Toàn bộ những công việc đó, người ta gọi là tổng hợp thống kê.

Tổng hợp thống kê là sự tập trung, chỉnh lý và hệ thống hoá các tài liệu ban đầu thu thập được trong điều tra thống kê của từng đơn vị tổng thể thành tài liệu phản ánh đặc trưng chung của cả tổng thể.

b) Ý nghĩa:

Tổng hợp thống kê là giai đoạn thứ 2 của quá trình nghiên cứu thống kê, không thể thiếu được, cũng không thể không khoa học và không thể không đúng phương pháp, nó là cơ sở rất quan trọng cho giai đoạn phân tích thống kê.

c) Nhiệm vụ:

Nhiệm vụ của giai đoạn này là:

- Tập trung và sắp xếp các tài liệu theo một trình tự nhất định.

Nếu tài liệu điều tra thu thập được ở số ít các đơn vị người ta thường sắp xếp dữ liệu này theo một trình tự nào đó (thứ tự tăng dần về lượng biến của 1 tiêu thức số lượng nào đó, hoặc theo trật tự quy định nào đó đối với dữ liệu định tính).

- Sắp xếp các đơn vị vào các tổ nhóm theo một hay một vài tiêu thức đặc trưng và tính toán các đại lượng thống kê đặc trưng cho tổ nhóm và toàn bộ tổng thể.

Nhiệm vụ này thường gặp khi tài liệu điều tra thu thập được ở số lớn các đơn vị, khối lượng dữ liệu nhiều.

Ví dụ: Trong điều tra dân số, tài liệu thu thập được ở từng người dân rất lớn, người ta thường tổng hợp theo cách sắp xếp người dân theo độ tuổi, trình độ văn hoá hay nghề nghiệp... sau đó tính các chỉ tiêu thống kê mô tả từng tổ như số lượng trung bình, nhiều nhất, ít nhất, tần số hay tần suất.

- Trình bày dữ liệu tổng hợp dưới hình thức bảng hay đồ thị thống kê.

1.2. Nội dung của tổng hợp thống kê

Theo trình tự nội dung của tổng hợp thống kê bao gồm xác định mục đích phân tích; nội dung tổng hợp; kiểm tra tài liệu; phân chia các đơn vị thành các tổ hay tiểu tổ và trình bày kết quả tổng hợp. Chương này trình bày chủ yếu 2 bước là phân tổ thống kê và trình bày kết quả tổng hợp thống kê dưới hình thức bảng hay đồ thị thống kê.

Xác định mục đích của tổng hợp thống kê là cụ thể hoá tiêu thức cần sắp xếp và phân loại. Đây là bước quan trọng vì tổng thể nghiên cứu có biểu hiện khác nhau. Mặt khác mục đích tổng hợp thống kê làm cơ sở cho phân tích thống kê nên rất cần cụ thể hoá mục đích tổng hợp.

Ví dụ: Tổng thể dân số có biểu hiện về nghề nghiệp, lứa tuổi, trình độ văn hoá, ngoại ngữ, quê quán, tôn giáo... Do vậy, khi nghiên cứu tổng thể ở đặc tính nào thì tổng hợp thống kê mới khái quát hoá, sắp xếp, hệ thống hoá theo các đặc trưng và khía cạnh đó.

* Xác định mục đích tổng hợp thường dựa vào mục đích nghiên cứu của thống kê. Có thể nói rằng mục đích nghiên cứu của thống kê xuyên suốt cả 3 giai đoạn, hay nói cách khác cả 3 giai đoạn này đều nhằm đáp ứng yêu cầu của nghiên cứu thống kê.

* Xác định nội dung của tổng hợp thống kê: Những danh mục về các biểu hiện của các tiêu thức đã có ở điều tra thống kê, nhưng không tất cả các biểu hiện của tiêu thức đều đưa vào tổng hợp, mà chỉ chọn các tiêu thức nào có nội dung tổng hợp vừa đủ đáp ứng mục đích nghiên cứu.

Ví dụ: Điều tra dân số, người ta thường tổng hợp theo độ tuổi dưới 1 tuổi, 1-3 tuổi, 4-6 tuổi, 7-11 tuổi, 12-15 tuổi, 16-55 tuổi, 56-100 tuổi, hơn 100 tuổi. Nhưng tên quê quán không nhất thiết phải tổng hợp.

Xác định nội dung tổng hợp thống kê là thống nhất danh mục chính thức về các biểu hiện của các tiêu thức bằng hệ thống các tiêu thức hay chỉ tiêu thống kê cần cho nghiên cứu. Người ta thường dùng phân tổ thống kê để thực hiện các nội dung tổng hợp thống kê.

* Kiểm tra tài liệu dùng để tổng hợp:

Chất lượng của tổng hợp thống kê phụ thuộc vào chất lượng của tài liệu đưa vào tổng hợp. Ở điều tra thống kê người ta đã kiểm tra tài liệu rồi, tuy nhiên trong giai đoạn này vẫn cần kiểm tra lại trước khi tổng hợp.

Nội dung kiểm tra gồm:

- Kiểm tra điển hình: Chọn mẫu các phiếu điều tra để kiểm tra.

- Kiểm tra theo nội dung: Chính xác, đầy đủ, kịp thời và lô gíc.

2. PHÂN TỔ THỐNG KÊ

2.1. Khái niệm, ý nghĩa và tác dụng

a) Khái niệm:

Phân tổ thống kê là căn cứ vào 1 hay một số tiêu thức để tiến hành phân chia các đơn vị của hiện tượng nghiên cứu thành các tổ và tiểu tổ sao cho các đơn vị trong cùng một tổ thì giống nhau về tính chất, ở khác tổ thì khác nhau về tính chất.

Ví dụ: Phân tổ các em sinh viên trong lớp Kinh tế nông nghiệp khoá 50 theo tiêu thức giới tính (bảng 1.3).

Bảng 1.3.

Diễn giải	Số lượng (người)	Tỷ lệ (%)
Tổng	90	100,00
Nam	40	36,67
Nữ	50	63,33

Khi phân tổ thống kê, các đơn vị tổng thể được tập hợp vào một số tổ, giữa các tổ lại có sự khác nhau về tính chất. Còn trong phạm vi mỗi tổ, các đơn vị có cùng (hoặc gần giống nhau) về tính chất theo tiêu thức được dùng làm căn cứ phân tổ.

b) Ý nghĩa:

* Dùng phân tổ để chọn ra các đơn vị điều tra (nhất là trong điều tra chọn mẫu).

* Phân tổ thống kê là phương pháp cơ bản của tổng hợp thống kê.

* Phân tổ thống kê là cơ sở và là một phương pháp phân tích thống kê.

c) Tác dụng của phân tổ thống kê:

Với ý nghĩa của phân tổ đã nêu trên, xuất phát từ yêu cầu của thực tễn xã hội mà phân tổ thống kê có tác dụng sau đây:

* Phân tổ thống kê nghiên cứu các loại hình kinh tế xã hội (*phân tổ phân loại*):

Bất kì một nền kinh tế xã hội nào cũng bao gồm nhiều loại hình kinh tế. Chẳng hạn nền kinh tế Việt Nam hiện tại bao gồm nhiều loại hình kinh tế khác nhau như: kinh tế Nhà nước; kinh tế tập thể; kinh tế tư nhân; kinh tế cá thể; kinh tế hỗn hợp.

Sự vận động và phát triển của nền kinh tế xã hội đó như thế nào, phụ thuộc vào vị trí, vai trò và xu hướng phát triển của từng loại hình kinh tế. Khi nghiên cứu đặc trưng

của nền kinh tế xã hội đó người ta phải nêu rõ: Có bao nhiều loại hình kinh tế? Là những loại hình kinh tế gì? Tỷ trọng mỗi loại hình như thế nào? Mối quan hệ giữa các loại hình? Xu hướng phát triển của các loại hình?

Để đáp ứng yêu cầu nghiên cứu trên, chỉ có thể thực hiện được thông qua phân tổ thống kê.

Ví dụ: Sự phát triển các thành phần kinh tế Việt Nam từ 1995 đến 2003 (bảng 2.3).

Bảng 2.3. Cơ cấu tổng sản phẩm trong nước theo thành phần kinh tế qua các năm

ĐVT: %

Thành phần kinh tế	1995	2000	2001	2002	2003
Kinh tế Nhà nước	40,18	38,53	38,40	38,38	39,08
Kinh tế tập thể	10,06	8,58	8,06	7,99	7,49
Kinh tế tư nhân	7,44	7,31	7,95	8,30	8,23
Kinh tế cá thể	36,02	32,31	31,84	31,57	30,73
Kinh tế có vốn đầu tư nước ngoài	6,30	13,28	13,75	13,76	14,47
Cộng	100,00	100,00	100,00	100,00	100,00

Nguồn: Niên giám thống kê 2003.

Theo bảng 2.3, nền kinh tế Việt Nam từ năm 1995 đến 2003 kinh tế Nhà nước vẫn chiếm tỷ trọng lớn nhất và giữ vai trò chủ đạo. Kinh tế cá thể được chú trọng phát triển, đang cạnh tranh mạnh mẽ với kinh tế Nhà nước. Kinh tế có vốn đầu tư nước ngoài tăng nhanh.

* Phân tổ thống kê nghiên cứu kết cấu nội bộ tổng thể (*phân tổ kết cấu*):

Kết cấu nội bộ tổng thể là tỷ lệ các bộ phận chiếm trong tổng thể và quan hệ tỷ lệ về lượng giữa các bộ phận đó nói lên kết cấu nội bộ tổng thể.

Mỗi hiện tượng kinh tế xã hội hay quá trình kinh tế xã hội đều do cấu thành từ nhiều bộ phận, nhiều nhóm đơn vị có tính chất khác nhau hợp thành. Ví dụ, theo khu vực, dân số của Việt Nam gồm 2 nhóm khác nhau là thành thị và nông thôn. Giữa 2 nhóm có sự khác nhau về tính chất ngành nghề, công việc và cá tính của người dân; tỷ lệ mỗi bộ phận này và quan hệ tỷ lệ giữa 2 nhóm nói lên kết cấu dân số Việt Nam theo khu vực.

Nghiên cứu kết cấu nội bộ tổng thể giúp ta đi sâu nghiên cứu bản chất của hiện tượng, thấy được tầm quan trọng của từng bộ phận trong tổng thể. Nếu nghiên cứu kết cấu nội bộ tổng thể theo thời gian cho ta thấy được xu hướng phát triển của hiện tượng nghiên cứu.

Như vậy, muốn nghiên cứu kết cấu nội bộ tổng thể phải dựa trên cơ sở của phân tổ thống kê.

* Phân tổ thống kê nghiên cứu mối liên hệ ảnh hưởng lẫn nhau giữa các tiêu thức của hiện tượng (*phân tổ phân tích hay liên hệ*):

Các quá trình hay hiện tượng kinh tế - xã hội phát sinh và phát triển không phải ngẫu nhiên, tách rời với các hiện tượng xung quanh mà chúng có liên hệ và phụ thuộc lẫn nhau theo những quy định nhất định. Sự biến động của hiện tượng này sẽ dẫn đến sự biến động của hiện tượng khác và ngược lại mỗi hiện tượng biến động đều do sự tác động của các hiện tượng xung quanh.

Ví dụ: Trẻ em ăn no, đủ chất thì chóng lớn, khoẻ mạnh; lúa thiếu dinh dưỡng, mà tăng lượng phân bón dẫn đến năng suất tăng, giá thành hạ; hàng hoá nhiều thì giá bán hạ.

Nhiệm vụ của thống kê không chỉ nghiên cứu bản chất mà còn nghiên cứu mối liên hệ giữa các hiện tượng kinh tế nói chung và các tiêu thức nói riêng.

Khi nghiên cứu mối quan hệ ảnh hưởng lẫn nhau giữa các hiện tượng, người ta thường chia các tiêu thức thành hai loại: tiêu thức nguyên nhân, tiêu thức kết quả.

+ Tiêu thức nguyên nhân là tiêu thức mà lượng biến của nó thay đổi làm cho lượng biến của tiêu thức khác cũng thay đổi.

+ Tiêu thức kết quả là tiêu thức mà lượng biến của nó có thay đổi do sự biến động của tiêu thức nguyên nhân.

Phân tổ hiện tượng kinh tế xã hội theo một trong hai tiêu thức trên thì biểu hiện về lượng của tiêu thức còn lại sẽ phản ánh mối quan hệ nhân quả mà ta cần nghiên cứu.

Phân tổ thống kê nghiên cứu mối liên hệ ảnh hưởng lẫn nhau giữa các hiện tượng như vậy gọi là phân tổ phân tích hay phân tổ liên hệ.

2.2. Quá trình phân tổ thống kê

Hiện nay do khoa học công nghệ, nhất là công nghệ tin học khá phát triển, người ta đã lập trình và vận dụng được các chương trình máy tính đưa vào ứng dụng trong nghiên cứu và phục vụ sản xuất. Về phân tổ thống kê cũng đã có nhiều chương trình vi tính chuyên cho xử lý số liệu thống kê đã thực hiện, ví dụ IRRISTAT, STATGRAF, SPSS và EXCEl. Nhưng, đó chỉ là công việc đơn thuần mà máy tính thực hiện, còn mục đích phân tổ của chúng ta để làm gì, chia làm bao nhiêu tổ... máy tính không thể thực hiện được. Vì vậy người làm công tác chuyên môn thống kê hoặc vận dụng thống kê làm công cụ quản lý xã hội và kinh tế cần nắm vững, hiểu được những công việc của phân tổ thống kê là gì?

Quá trình phân tổ thống kê bao gồm: Xác định tiêu thức phân tổ; xác định số tổ cần thiết và phạm vi mỗi tổ; xác định các chỉ tiêu giải thích.

a) Tiêu thức phân tổ:

* Khái niệm: Tiêu thức phân tổ là tiêu thức được lựa chọn làm căn cứ để tiến hành phân tổ thống kê.

* Ý nghĩa: Tiêu thức phân tổ phản ánh đúng bản chất của hiện tượng mà mục đích nghiên cứu đề ra. Sở dĩ như vậy là vì mỗi đơn vị tổng thể như chúng ta đã biết bao gồm nhiều tiêu thức khác nhau, tiêu thức nào cũng có thể dùng để phân tổ được, xong mỗi tiêu thức có ý nghĩa khác nhau. Thí dụ: Tổng thể dân số, có thể:

- Phân tổ theo giới tính. Giới tính là tiêu thức phân tổ.

- Phân tổ theo độ tuổi. Độ tuổi là tiêu thức phân tổ.

- Phân tổ theo nghề nghiệp. Nghề nghiệp là tiêu thức phân tổ.

Nhưng, cùng một nguồn tài liệu nếu chọn tiêu thức phân tổ khác nhau có thể đưa đến kết luận khác nhau, hoặc chọn tiêu thức phân tổ không đúng theo mục đích nghiên cứu sẽ dẫn đến nhận xét đánh giá khác nhau về thực tế của hiện tượng.

* Thí dụ: Nghiên cứu kết quả học tập của sinh viên 1 lớp, Trường Đại học Nông nghiệp I Hà Nội năm học 2004 -2005.

- Nếu chọn tiêu thức phân tổ là thời gian tự học thì ta có kết quả như bảng 3.3.

Bảng 3.3. Phân tổ số sinh viên của lớp theo số giờ tự học trong ngày

Số giờ tự học/ngày (giờ)	Số sinh viên (người)	Cơ cấu (%)
0	5	6,25
1	7	8,75
2	15	18,75
3	20	25,00
4	25	31,25
5	8	10,00
Cộng	80	100,00

Kết quả phân tổ ở bảng 3.3 cho biết số sinh viên sử dụng thời gian học ở nhà từ 3 - 4 giờ/ngày chiếm 56,25% chứ chưa cho biết kết quả học tập của sinh viên như thế nào.

- Nếu chọn tiêu thức phân tổ là điểm thi trung bình các môn thi trong năm của 1 sinh viên thì mới thể hiện kết quả học tập của sinh viên (bảng 4.3).

Bảng 4.3. Phân tổ số sinh viên của lớp theo điểm thi trung bình 1 sinh viên

Điểm thi trung bình 1 sinh viên (điểm)	Số sinh viên (người)	Cơ cấu (%)
Dưới 5,0	8	10,00
Từ 5,0 đến 6,9	45	56,25
Từ 7,0 đến 8,9	25	31,25

Kết quả phân tổ ở bảng 4.3 cho biết số sinh viên có điểm thi đạt điểm từ 5 trở lên chiếm 90%, trong đó có 33,75% khá giỏi, chứng tỏ kết quả học tập của lớp này rất tốt.

Từ 9,0 trở lên	2	2,50
Cộng	80	100,00

* Những nguyên tắc để xác định đúng tiêu thức phân tổ:

Thứ nhất: Phải dựa trên cơ sở phân tích lí luận kinh tế – xã hội một cách sâu sắc để chọn ra tiêu thức phản ánh bản chất, phù hợp với mục đích nghiên cứu.

Tiêu thức bản chất là tiêu thức nêu rõ bản chất của hiện tượng, phản ánh đặc trưng cơ bản của hiện tượng trong điều kiện thời gian và địa điểm cụ thể.

Thí dụ: Điểm thi là tiêu thức phản ánh bản chất kết quả học của sinh viên, chứ thời gian tự học chỉ phản ánh một phần nguyên nhân của kết quả học.

Bản chất của hiện tượng có thể được phản ánh qua nhiều tiêu thức khác nhau, vì vậy tuỳ mục đích nghiên cứu mà dùng lí luận kinh tế – xã hội để chọn ra tiêu thức bản chất nhất.

Thứ hai: Phải căn cứ vào điều kiện lịch sử cụ thể của hiện tượng nghiên cứu.

Cùng một hiện tượng nhưng ở các điều kiện lịch sử khác nhau thì tiêu thức phân tổ cũng mang ý nghĩa khác nhau. Nếu chỉ dùng một tiêu thức phân tổ chung cho mọi trường hợp thì tiêu thức đó trong điều kiện lịch sử này có thể giúp ta nghiên cứu chính xác, nhưng ở điều kiện lịch sử khác lại không có tác dụng.

Quay lại với thí dụ về kết quả học tập của sinh viên: Khi sinh viên còn đang học tại trường thì tiêu thức phản ánh đúng đắn nhất kết quả học tập là điểm thi trung bình; khi sinh viên đã làm việc thì điểm thi lại không phản ánh đúng bản chất của kết quả làm việc.

Thứ ba: Tuỳ theo tính chất phức tạp của hiện tượng và mục đích yêu cầu nghiên cứu có thể lựa chọn 1 hay nhiều tiêu thức phân tổ.

- Phân tổ tài liệu theo 1 tiêu thức gọi là phân tổ giản đơn, cách phân tổ này thường dùng nghiên cứu các hiện tượng đơn giản và với 1 mục đích yêu cầu nhất định.

Thí dụ: Phân tổ sinh viên theo giới tính: nam, nữ.

- Phân tổ tài liệu theo từ 2 tiêu thức trở lên kết hợp với nhau gọi là phân tổ kết hợp. Cách phân tổ này thường dùng nghiên cứu các hiện tượng phức tạp và thoả mãn nhu cầu mục đích nghiên cứu.

Thí dụ: Phân tổ sinh viên theo điểm thi trung bình và giới tính.

Phân tổ kết hợp tuy có nhiều ưu điểm, song cũng không nên kết hợp quá nhiều tiêu thức dễ làm cho việc phân tổ trở nên phức tạp, dẫn đến có những sai sót làm giảm mức độ chính xác của tài liệu.

b) Xác định số tổ cần thiết và phạm vi mỗi tổ:

Việc xác định số tổ cần thiết (bao nhiêu tổ) và ranh giới giữa các tổ phụ thuộc vào tiêu thức phân tổ là tiêu thức số lượng hay tiêu thức chất lượng (thuộc tính).

* Tiêu thức thuộc tính: Các tổ được hình thành là do sự khác nhau về thuộc tính, tính chất hay loại hình.

Khi phân tổ theo tiêu thức thuộc tính thì số tổ được hình thành theo 2 xu hướng sau:

- Đơn giản: Có một số trường hợp, việc xác định số tổ và ranh giới giữa các tổ rất đơn giản và rất dễ dàng vì số tổ ít và ranh giới hình thành một cách đương nhiên.

Thí dụ: 1) Phân tổ dân số theo giới tính: Số tổ 2, nam, nữ.

2) Phân tổ diện tích trồng lúa trong năm theo thời vụ gieo trồng: 2 vụ, vụ đông xuân, vụ mùa.

Trong trường hợp này ta coi mỗi loại hình là 1 tổ, số tổ = số loại hình.

- Có những trường hợp phức tạp:

Thí dụ: Phân tổ lao động theo nghề nghiệp. Có rất nhiều nghề như làm bánh kẹo, dệt, thêu ren, làm ruộng, làm gạch…

Phân loại cây trồng: lúa, ngô, khoai, sắn, cải bắp, su hào, cà chua...

Nếu cứ coi mỗi loại hình là 1 tổ thì số tổ sẽ quá nhiều, hơn nữa giữa các loại hình chưa chắc chắn đã khác nhau về chất.

Thí dụ: ngô, khoai, sắn là cây hoa màu dùng làm lương thực.

Trong những trường hợp này, người ta thường ghép một số loại hình nhỏ vào cùng một tổ theo nguyên tắc "*Các loại hình đó phải giống nhau hoặc gần giống nhau về tính chất nào đó hay ý nghĩa kinh tế*".

Thí dụ: 1) Lúa, ngô, khoai, sắn có ý nghĩa đều làm lương thực, xếp vào 1 tổ gọi là cây lương thực.

2) Dệt, thêu, ren... xếp vào công nghiệp dệt.

- Đối với một số phân tổ theo tiêu thức thuộc tính mà dùng cho toàn quốc có quy định chung thống nhất gọi là danh mục phân loại. Phương pháp phân loại là một công trình nghiên cứu khoa học, có tác dụng trong nền kinh tế quốc dân.

Thí dụ: Phân loại ngành kinh tế: Nông, Lâm, Ngư nghiệp, Công nghiệp & tiểu thủ công nghiệp... theo quy định của Tổng cục Thống kê.

* Tiêu thức số lượng:

- Cơ sở để xác định số tổ và phạm vi mỗi tổ là sự khác nhau về lượng biến của tiêu thức phân tổ. Tức là dựa vào sự biểu hiện lượng biến khác nhau mà sắp xếp các đơn vị vào các tổ khác nhau về tính chất.

Dựa trên cơ sở này số tổ và ranh giới giữa các tổ được xác định như sau:

- Nếu lượng biến của tiêu thức phân tổ mà ít, có một số các trị số xác định, khi đó ứng với mỗi trị số lượng biến của tiêu thức phân tổ ta lập 1 tổ.

Thí dụ: Nghiên cứu tình hình sinh đẻ có kế hoạch của một địa phương, có phân tổ số phụ nữ theo số lần sinh con như ở bảng 5.3.

Bảng 5.3. Phân tổ số phụ nữ của địa phương A theo số con của 1 mẹ

Số con của 1 mẹ (con)	Số mẹ (người)	Cơ cấu (%)
0	6	3,51
1	35	20,47
2	82	47,95
3	38	22,22
4	10	5,85
Cộng	171	100,00

- Nếu lượng biến của tiêu thức phân tổ mà nhiều và biến thiên lớn, thí dụ, phân tổ dân số theo độ tuổi, trong trường hợp này ta cần chú ý mối liên hệ giữa lượng biến và tính chất trong phân tổ.

Dùng lí luận để phân tích xem lượng biến tích luỹ đến mức độ nào thì tính chất của nó mới thay đổi làm xuất hiện 1 tổ khác. Như vậy, mỗi tổ sẽ ứng với 1 khoảng trị số lượng biến nhất định của tiêu thức phân tổ, nghĩa là mỗi tổ có 2 giới hạn.

- Giới hạn dưới là lượng biến nhỏ nhất để làm cho tổ đó được hình thành.

- Giới hạn trên là lượng biến lớn nhất của tổ, nếu vượt quá giới hạn trên thì tính chất của hiện tượng thay đổi và chuyển sang tổ khác.

- Mức độ chênh lệch giới hạn trên và giới hạn dưới của mỗi tổ gọi là khoảng cách tổ.

- Tổ đầu và tổ cuối có thể chỉ có 1 giới hạn. Những tổ đó gọi là tổ mở. Việc thành lập các tổ mở trong thống kê rất cần thiết vì nó có tác dụng thu nạp đầy đủ các đơn vị có trị số tiêu thức nhỏ và cực lớn. Trường hợp này gọi là phân tổ có khoảng cách tổ.

Ranh giới giữa các tổ được xác định như sau:

- Trị số lượng biến của tiêu thức phân tổ biến thiên không liên tục thì giới hạn dưới của 1 tổ nào đó là trị số sát với giới hạn trên của tổ trước và giới hạn trên của tổ đó là trị sát với giới hạn dưới của tổ sau.

Thí dụ: Độ tuổi: Lượng biến của nó biến thiên không liên tục.

1 tuổi = 1 năm = 12 tháng; Nếu ta gọi 13 tháng = 1,1 tuổi không có ý nghĩa lắm.

Bảng 6.3. Phân tổ nhân khẩu thực tế thường trú trong hộ gia đình theo nhóm tuổi của cả nước năm 2000

Nhóm tuổi (tuổi)	Số người (triệu người)
Dưới 15	23,41
Từ 15 đến 24	15,23
Từ 25 đến 34	11,69
Từ 35 đến 44	11,67
Từ 45 đến 54	6,83
Từ 55 đến 59	1,94
Từ 60 tuổi trở lên	6,96
Cộng	77,69

Từ bảng 6.3 ta thấy, ở tổ thứ 3 giới hạn dưới của tổ là 25, là trị số nằm sát với giới hạn trên của tổ 2 là 24; giới hạn trên của tổ 3 là 34, là trị số nằm sát với giới hạn dưới của tổ sau.

Nguồn: Thực trạng lao động - việc làm ở Việt Nam năm 2000 (NXB Lao động - Xã hội 2001)

- Trị số lượng biến của tiêu thức phân tổ biến thiên liên tục thì giới hạn dưới của tổ nào đó là trị số trùng với giới hạn trên của tổ trước và giới hạn trên của tổ đó là trị số trùng với giới hạn dưới của tổ sau.

Thí dụ: Giá cả, tiền lương, điểm thi của sinh viên... lượng biến thường biến thiên liên tục (bảng 7.3).

Bảng 7.3. Phân tổ số công nhân ở 1 doanh nghiệp theo tiền lương bình quân 1 người 1 tháng

Tiền lương (1000 đ/tháng)	Số người (người)
Đến 500	20
Từ 500 - 800	30
Từ 800 - 1000	40
Trên 1000	10
Cộng	100

Từ bảng 7.3 ta thấy, ở tổ thứ 3 giới hạn dưới của tổ là 800, là trị số trùng với giới hạn trên của tổ 2; giới hạn trên của tổ 3 là 1000, là trị số trùng với giới hạn dưới của tổ sau.

Chú ý:

- Nếu có đơn vị tổng thể nào đó có trị số lượng biến của tiêu thức phân tổ trùng với giới hạn giữa 2 tổ thì thông thường người ta xếp vào tổ trước (tức là tổ có trị số tiêu thức phân tổ bé hơn).

Thí dụ: Mức lương là 800 thì xếp vào tổ 2 chứ không xếp vào tổ 3.

Nhìn chung khi phân tổ theo tiêu thức số lượng thì khoảng cách giữa các tổ nói chung không bằng nhau vì hiện tượng kinh tế hay quá trình kinh tế xã hội biến thiên thường là không đều đặn, không máy móc cơ học, không phải cứ ứng với một sự thay đổi về lượng như nhau thì tính chất của hiện tượng cũng thay đổi, có khi lượng biến thay đổi khá nhiều mà tính chất của hiện tượng thay đổi chưa rõ rệt lắm (khoảng cách tổ

lớn), còn có khi lượng biến mới thay đổi ít thì tính chất của hiện tượng đã thay đổi (khoảng cách tổ nhỏ).

Thí dụ: Nghiên cứu khả năng tiêu hoá thịt của con người (bảng 8.3).

Bảng 8.3. Mối quan hệ giữa lượng thịt ăn với khả năng tiêu hoá

Lượng thịt ăn bình quân 1 người 1 ngày (g/người)	Tính chất tiêu hoá
50	Tốt
100	Tốt
150	Tốt
200	Tốt
250	t.bình
300	Kém
350	Kém
400	Quá kém

- Trong thực tiễn đối với những hiện tượng mà sự biến đổi về chất đều đặn từ nhỏ đến lớn, thấp đến cao người ta thường và có thể phân tổ có khoảng cách tổ bằng nhau. Khi đó khoảng cách tổ được xác định theo công thức sau:

$$d = \frac{x_{max} - x_{min}}{n}$$

Trong đó:

- d là khoảng cách tổ
- x_{max} và x_{min} là trị số lượng biến lớn nhất và bé nhất của tiêu thức phân tổ
- n là số tổ định chia.

Thí dụ: Năng suất lúa bình quân 1 ha gieo trồng của các hộ trồng lúa trong 1 xã biến động đều đặn từ 30 đến 70 tạ/ha. Nếu định chia thành 5 tổ thì khoảng cách tổ là:

$$d = \frac{x_{max} - x_{min}}{n} = \frac{70 - 30}{5} = 8 \text{ (tạ/ha)}$$

Các tổ được hình thành như sau:

1. Từ 30 đến 38 tạ/ha
2. Từ 38 đến 46 tạ/ha
3. Từ 46 đến 54 tạ/ha
4. Từ 54 đến 62 tạ/ha
5. Từ 62 đến 70 tạ/ha

Tóm lại: Trên đây là lí luận và kỹ thuật về xác định số tổ cần thiết và khoảng cách tổ khi tiến hành phân tổ thống kê. Song cần lưu ý không nên chia số tổ quá nhiều hay quá ít. Trong thực tế người ta đã sử dụng chương trình máy tính để phân tổ.

c) Chỉ tiêu giải thích:

* Khái niệm: Chỉ tiêu giải thích là những chỉ tiêu dùng để nói rõ đặc điểm của các tổ cũng như toàn bộ tổng thể.

Lấy lại ví dụ phân tổ các hộ trồng lúa theo năng suất: Các chỉ tiêu giải thích là diện tích gieo trồng, sản lượng lúa, chi phí... của mỗi nhóm.

* Ý nghĩa: Chỉ tiêu giải thích có vai trò quan trọng trong phân tổ vì:

- Nó nói rõ đặc trưng của từng tổ và toàn bộ tổng thể;

- Nó làm căn cứ để so sánh các tổ với nhau và tính một số chỉ tiêu phân tích khác.

* Cơ sở chọn đúng các chỉ tiêu giải thích

+ Căn cứ vào mục đích nghiên cứu

Ví dụ phân tổ các hộ theo năng suất lúa:

- Nếu mục đích nghiên cứu là ảnh hưởng của các biện pháp canh tác đến năng suất lúa, thì các chỉ tiêu giải thích sẽ là: tổng lượng phân bón, diện tích cấy giống mới, diện tích tưới tiêu chủ động, mật độ cấy...

- Nếu mục đích nghiên cứu là quy mô sản xuất thì các chỉ tiêu giải thích là giá trị sản lượng, diện tích canh tác, lao động, TSCĐ, vốn.

+ Các chỉ tiêu giải thích phải liên quan chặt chẽ đến tiêu thức phân tổ.

Thí dụ: Năng suất lúa là tiêu thức phân tổ, các chỉ tiêu giải thích là diện tích gieo trồng lúa, phân bón đối với lúa...

2.3. Dãy số phân phối

Kết quả của phân tổ thống kê cho chúng ta một dãy số phân phối.

* Khái niệm: Dãy số phân phối là 1 dãy số được lập nên do phân phối các đơn vị tổng thể vào các tổ theo 1 tiêu thức phân tổ nào đó và được sắp xếp theo trình tự biến động của lượng biến tiêu thức phân tổ.

* Các loại dãy số phân phối: Tuỳ theo tiêu thức phân tổ là tiêu thức số lượng hay tiêu thức thuộc tính mà có 2 loại dãy số phân phối.

- Dãy số lượng biến: Là dãy số được hình thành từ việc phân tổ theo tiêu thức số lượng, dãy số này phản ánh kết cấu của tổng thể theo tiêu thức số lượng.

Thí dụ: Phân tổ người lao động theo mức lương.

Một dãy số lượng biến có 2 yếu tố: Lượng biến và tần số.

- Lượng biến là các trị số biểu hiện cụ thể mức độ của tiêu thức số lượng, kí hiệu là x_i.

- Tần số là đơn vị tổng thể được phân phối vào mỗi tổ, kí hiệu là f_i, nếu tần số biểu hiện bằng số tương đối (%) gọi là tần suất, kí hiệu là s_i.

- Nếu lượng biến là 1 trị số xác định (không liên tục), gọi là dãy số phân tổ.

- Nếu lượng biến là 1 khoảng trị số (liên tục), gọi là dãy số có khoảng cách tổ.

Dạng tổng quát của 1 dãy số lượng biến như sau:

Lượng biến	Tần số	Hoặc tần suất
X_1	f_1	$f_1/\sum f_i$
X_2	f_2	$f_2/\sum f_i$
...	...	...
x_n	f_n	$f_n/\sum f_i$
Tổng số	$\sum f_i$	100

- Dãy số thuộc tính là dãy số được hình thành từ phân tổ theo tiêu thức thuộc tính, nó cũng bao gồm cột tần số hay tần suất, còn cột lượng biến thay bằng thuộc tính nào đó của hiện tượng.

Thí dụ: Phân tổ nhân khẩu theo giới tính.

* Mục đích sử dụng dãy số phân phối.

Dãy số phân phối trong thống kê được dùng vào các mục đích sau:

- Nghiên cứu cấu thành tổng thể;

- So sánh dãy số phân phối theo thời gian để nêu lên sự biến đổi của hiện tượng theo thời gian và so sánh giữa 2 hiện tượng cùng loại, cùng thời gian nhưng ở 2 địa điểm khác nhau.

Ở mục đích này khi so sánh cần chú ý:

. Hai dãy số phải phản ánh cùng một hiện tượng;

. Hai dãy số phải phân tổ như nhau;

. Nếu quy mô so sánh khác nhau phải dùng tần suất.

- Tính tổng trị số tiêu thức: Tổng trị số tiêu thức phản ánh quy mô của từng tổ và quy mô của cả tổng thể.

Công thức tính : $\sum_{i=1}^{n} f_i x_i$

Trong trường hợp dãy số có khoảng cách tổ thì x_i là trung bình cộng của 2 giới hạn mỗi tổ.

3. TRÌNH BÀY SỐ LIỆU THỐNG KÊ

3.1. Bảng thống kê

a) Khái niệm, ý nghĩa:

* Khái niệm:

Bảng thống kê là một hình thức trình bày kết quả tổng hợp số liệu thống kê theo từng nội dung riêng biệt nhằm phục vụ cho yêu cầu của quá trình nghiên cứu thống kê.

* Ý nghĩa:

- Phản ánh đặc trưng cơ bản của từng tổ và của cả tổng thể;

- Mô tả mối liên quan mật thiết giữa các số liệu thống kê;

- Làm cơ sở áp dụng các phương pháp phân tích thống kê khác nhau một cách dễ dàng...

b) Kết cấu của bảng thống kê:

+ Về hình thức

- Bảng thống kê bao gồm các hàng ngang và cột dọc, các tiêu đề và các tài liệu con số.

- Hàng ngang cột dọc phản ánh quy mô của bảng thống kê, thường được đánh số thứ tự.

- Ô của bảng dùng để điền số liệu thống kê.

- Tiêu đề của bảng: Phản ánh nội dung của bảng và của từng chỉ tiêu trong bảng. Có 2 loại tiêu đề:

Tiêu đề chung: Tên bảng.

Tiêu đề nhỏ (mục): Tên hàng, cột.

- Các số liệu được ghi vào các ô của bảng, mỗi số liệu phản ánh đặc trưng về mặt lượng của hiện tượng nghiên cứu.

Hình thức của bảng được mô tả qua sơ đồ sau:

Tên bảng:

Tên hàng (Phần chủ đề)	Tên cột (Phần giải thích)						
	1	2	3	4	....	k	Cộng cột
A.							
B.							
C.							

...							
Cộng hàng							

Chú thích của bảng :...

* Về nội dung: chia thành 2 phần: Phần chủ để và phần giải thích.

- Phần chủ để: Nội dung phần chủ đề nhằm nêu rõ tổng thể nghiên cứu được phân thành những bộ phận nào, hoặc mô tả đối tượng nghiên cứu là những đơn vị nào, loại hình gì, tên địa phương hoặc các thời gian nghiên cứu khác nhau. Hay nói cách khác, phân chủ đề thể hiện tiêu thức phân tổ các đơn vị tổng thể thành các tổ. Vị trí của phần này thường để ở bên phải phía dưới của bảng (tên của các hàng- tiêu đề hàng).

- Phần giải thích: Nội dung phần này gồm các chỉ tiêu giải thích về các đặc điểm của đối tượng nghiên cứu (giải thích phần chủ đề của bảng). Vị trí của phần này thường để ở bên trái phía trên của bảng (tên của các cột- tiêu đề cột).

c) Nguyên tắc lập bảng thống kê:

Khi sử dụng bảng thống kê để trình bày các số liệu thống kê cần tôn trọng những vấn đề mang tính nguyên tắc như sau:

- Quy mô của bảng thống kê không nên quá lớn. Nếu bảng thống kê quá lớn (nhiều hàng, cột) có thể tách thành 2 hoặc 3 bảng nhỏ hơn;

- Các tiêu đề, tiêu mục nên ngắn gọn, chính xác và dễ hiểu;

- Các hàng và các cột được ghi kí hiệu và đánh số;

- Các chỉ tiêu giải thích sắp xếp hợp lí;

- Cách ghi số liệu vào bảng thống kê theo quy ước sau:

(-): Không có tài liệu;

(...): Biểu thị số liệu còn thiếu có thể bổ sung;

(x) Biểu thị hiện tượng không có liên quan đến chỉ tiêu đó;

Các đơn vị có cùng 1 đơn vị tính toán giống nhau phải ghi theo mức độ chính xác như nhau (0,1 hay 0,01...) theo nguyên tắc làm tròn số.

- Cuối bảng cần có ghi chú giải thích tài liệu trong bảng như nguồn tài liệu trích, cách tính...

d) Các loại bảng thống kê:

* *Bảng đơn giản*: Bảng thống kê mà phần chủ đề không phân tổ, chỉ liệt kê các đơn vị tổng thể, tên gọi các địa phương hoặc các thời gian khác nhau của quá trình nghiên cứu.

Thí dụ:

Bảng 9.3. Hiện trạng đất đai và dân số trung bình của vùng Tây Nguyên năm 2002

Các tỉnh	Diện tích đất (1000 ha)	Dân số trung bình (1000 người)	Bình quân đất/người (ha/người)
Kon Tum	961,5	339,5	2,83
Gia Lai	1549,6	1064,6	1,46
Đắk Lăk	1959,9	1938,8	1,01
Lâm Đồng	976,5	1064,3	0,92
Cộng	5447,5	4407,2	1,24

Nguồn: Niên giám thống kê 2003

* *Bảng tần số (bảng phân tổ):* Là bảng thống kê mà tổng thể đối tượng nghiên cứu ghi trong phần chủ để được chia thành các tổ theo 1 tiêu thức nào đó.

Bảng phân tổ thường bao gồm 2 cột tính toán là tần số và tần suất. Khi phân tổ theo tiêu thức thuộc tính hay tiêu thức số lượng, người ta thường đếm xem có bao nhiêu đơn vị có cùng một biểu hiện và so với tổng số quan sát thì số đơn vị có cùng biểu hiện này chiếm bao nhiêu phần trăm.

Thí dụ:

Bảng 10.3. Dân số trung bình của Việt Nam phân theo giới tính năm 2003

Giới tính	Tần số (1000 người)	Tần suất (%)
Nam	39.755,4	49,14
Nữ	41.147,0	50,86
Cộng	80.902,4	100,00

Bảng 11.3. Phân tổ số sinh viên của lớp theo số giờ tự học trong ngày

Số giờ tự học/ngày (giờ)	Tần số (người)	Tần suất (%)
0	5	6,25
1	7	8,75
2	15	18,75

3	20	25,00
4	25	31,25
5	8	10,00
Cộng	80	100,00

Bảng tần số có thể được phân tổ theo nhiều tiêu thức, khi đó người ta gọi là bảng tần số có ghép nhóm (có phân tổ) (bảng12.3).

Bảng 12.3. Hiện trạng đất nông nghiệp của Việt Nam năm 2002

Các loại đất	Tần số (1000 ha)	Tần suất (%)
1. Đất trồng cây hàng năm	5977,6	63,55
- Đất trồng lúa	4061,7	43,18
- Đất nương rẫy	642,7	6,83
- Đất trồng cây hàng năm khác	1273,2	13,53
2. Đất vườn tạp	623,2	6,62
3. Đất trồng cây lâu năm	2213,1	23,53
4. Đất đồng cỏ dùng cho chăn unôi	39,5	0,42
5. Đất có mặt nước nuôi trồng thuỷ sản	553,4	5,88
Cộng	9406,8	100,00

Nguồn: Niên giám thống kê 2003.

Bảng phân tổ được dùng để:

- Nêu rõ kết cấu và biến động kết cấu của hiện tượng nghiên cứu;

- Phân tích mối liên hệ giữa các hiện tượng.

* Bảng kết hợp: Là bảng trong đó tổng thể đối tượng nghiên cứu ghi ở phần chủ đề được phân tổ theo 2 tiêu thức trở lên. Bảng kết hợp giúp ta phân tích sâu hơn về đối tượng đang nghiên cứu. Bảng kết hợp thường gặp ở các dạng sau:

- Bảng kết hợp 2 tiêu thức thuộc tính.

Thí dụ:

Bảng 13.3. Số người đủ 15 tuổi trở lên hoạt động kinh tế thường xuyên đã qua các trình độ đào tạo ở Việt Nam năm 2000

Diễn giải	Tổng số		Thành thị		Nông thôn	
	Tần số (Số người)	Tỷ lệ (%)	Tần số (Số người)	Tỷ lệ (%)	Tần số (Số người)	Tỷ lệ (%)
1. Học nghề	22569	25,69	17180	26,38	5389	23,70
2. Trung học chuyên nghiệp	48485	55,18	32718	50,24	15767	69,35
3. Cao đẳng	7602	8,65	6528	10,02	1074	4,72
4. Đại học	9099	10,36	8592	13,19	507	2,23
5. Thạc sĩ	83	0,09	83	0,13		0,00
6. Tiến sĩ	22	0,03	22	0,03		0,00
Cộng	87860	100.00	65123	100,00	22737	100,00
Nguồn: Thực trạng lao động - việc làm ở Việt Nam năm 2000						

Bảng 13.3 cho biết người ta đã kết hợp 2 tiêu thức định tính là trình độ đào tạo và khu vực (thành thị, nông thôn).

- Bảng kết hợp 3 tiêu thức định tính

Thí dụ: Số người lao động phân theo tình trạng việc làm của Hà Nội năm 2000 người ta đã kết hợp 3 tiêu thức định tính như tình trạng việc làm, tuổi quy định và giới tính ở bảng 14.3.

Bảng 14.3. Số lượng lao động phân theo tình trạng việc làm của Hà Nội năm 2000

Diễn giải	Tổng số		Đủ việc làm		Thiếu việc và thất nghiệp	
	Tần số (người)	Tỷ lệ (%)	Tần số (người)	Tỷ lệ (%)	Tần số (người)	Tỷ lệ (%)
1. Trong độ tuổi lao động	1300704	100	894392	68,76	406312	31,24
Nữ	638456	100	450569	70,57	187887	29,43

Nam	662248	100	443823	67,02	218425	32,98
2. Ngoài tuổi quy định	1376585	100	935056	67,93	441529	32,07
Nữ	682719	100	478168	70,04	204551	29,96
Nam	693866	100	456888	65,85	236978	34,15

Nguồn: Thực trạng lao động – việc làm ở Việt Nam năm 2000

- Bảng kết hợp giữa tiêu thức số lượng với tiêu thức thuộc tính

Thí dụ: Số người lao động phân theo tình trạng việc làm của Hà Nội năm 2000 người ta đã kết hợp 3 tiêu thức, trong đó 2 tiêu thức định tính như tình trạng việc làm và giới tính, 1 tiêu thức số lượng là độ tuổi như sau (bảng 15.3).

Bảng 15.3. Số lượng lao động phân theo tình trạng việc làm của Hà Nội năm 2000

Nhóm tuổi (tuổi)	Tổng số		Đủ việc làm		Thiếu việc và thất nghiệp	
	Tần số (người)	Tỷ lệ (%)	Tần số (người)	Tỷ lệ (%)	Tần số (người)	Tỷ lệ (%)
Từ 15 - 24	225517	100	138608	61,46	86909	38,54
Từ 25 - 34	382976	100	283396	74,00	99580	26,00
Từ 35 - 44	408847	100	291292	71,25	117555	28,75
Từ 45 - 54	252854	100	165248	65,35	87606	34,65
Từ 55 - 60	45227	100	26336	58,23	18891	41,77
Trên 60	61148	100	30170	49,34	30978	50,66

Nguồn: Thực trạng lao động – việc làm ở Việt Nam năm 2000

3.2. Biểu đồ và đồ thị thống kê

a) Khái niệm, ý nghĩa:

Biểu đồ và đồ thị thống kê là các hình vẽ, đường nét hình học dùng để mô tả có tính quy ước các số liệu thống kê.

Khác với bảng thống kê, đồ thị hay biểu đồ thống kê sử dụng các số liệu kết hợp với hình vẽ, đường nét hay màu sắc để tóm tắt và trình bày các đặc trưng chủ yếu của hiện tượng nghiên cứu, phản ánh một cách khái quát các đặc điểm về cơ cấu, xu hướng biến động, mối liên hệ, quan hệ so sánh ... của hiện tượng cần nghiên cứu.

Vì dùng các hình vẽ, đường nét và màu sắc để biểu hiện các đặc trưng của hiện tượng nên tài liệu thống kê rất sinh động, có sức hấp dẫn lôi cuốn người đọc, giúp cho người xem nhận thức được những biểu hiện của hiện tượng một cách nhanh chóng, từ đó nhận ra được những nội dung chủ yếu của vấn đề nghiên cứu.

b) Các loại đồ thị thống kê:

* Theo nội dung phản ánh của đồ thị, có thể phân chia đồ thị thành các loại sau đây:

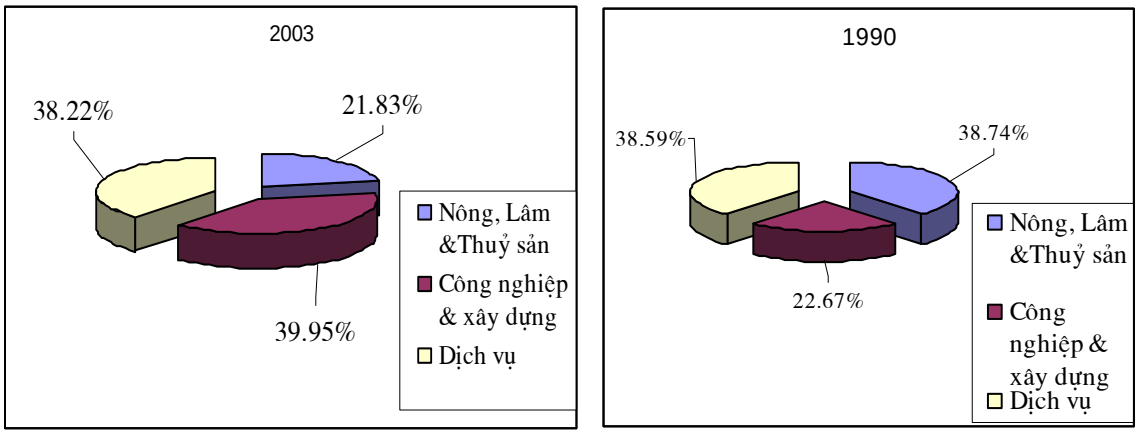
- Đồ thị kết cấu
- Đồ thị xu hướng biến động
- Đồ thị mối liên hệ
- Đồ thị so sánh
- Đồ thị phân phối
- Đồ thị hoàn thành kế hoạch.

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

- Đồ thị hình cột
- Đồ thị hình tròn
- Đồ thị đường gấp khúc
- Đồ thị hình tượng
- Bản đồ thống kê.

* Một số ví dụ về đồ thị thống kê:

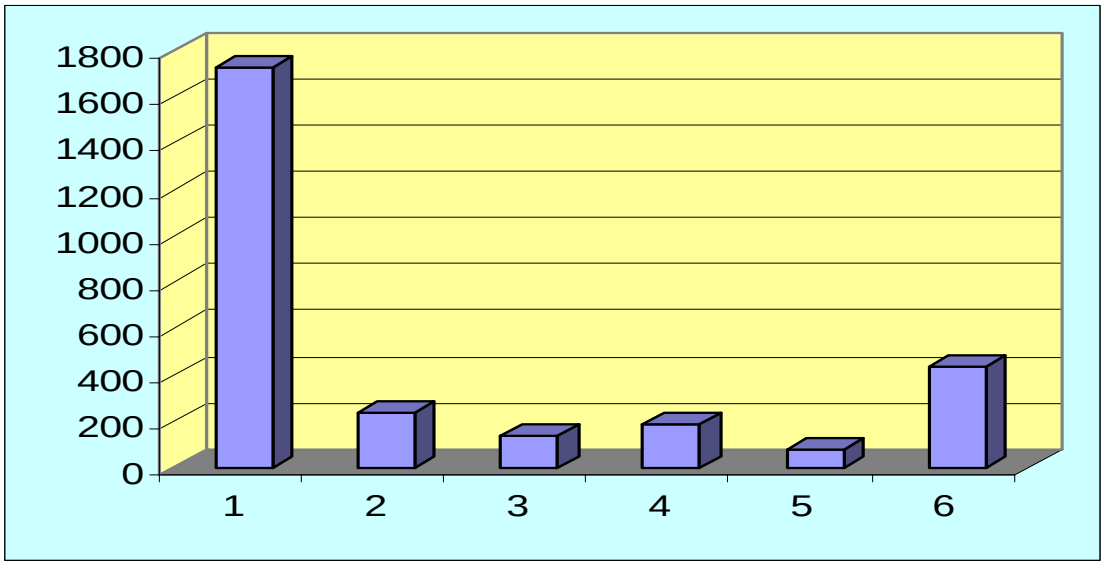
Thí dụ 1: Đồ thị hình tròn thể hiện cơ cấu tổng sản phẩm quốc nội theo giá thực tế phân theo ngành kinh tế (đồ thị 3.1)



Đồ thị 3.1. Cơ cấu tổng sản phẩm quốc nội của Việt Nam qua 2 năm 1990 và 2003 (Niên giám thống kê 2003)

Thí dụ 2: Đồ thị hình cột

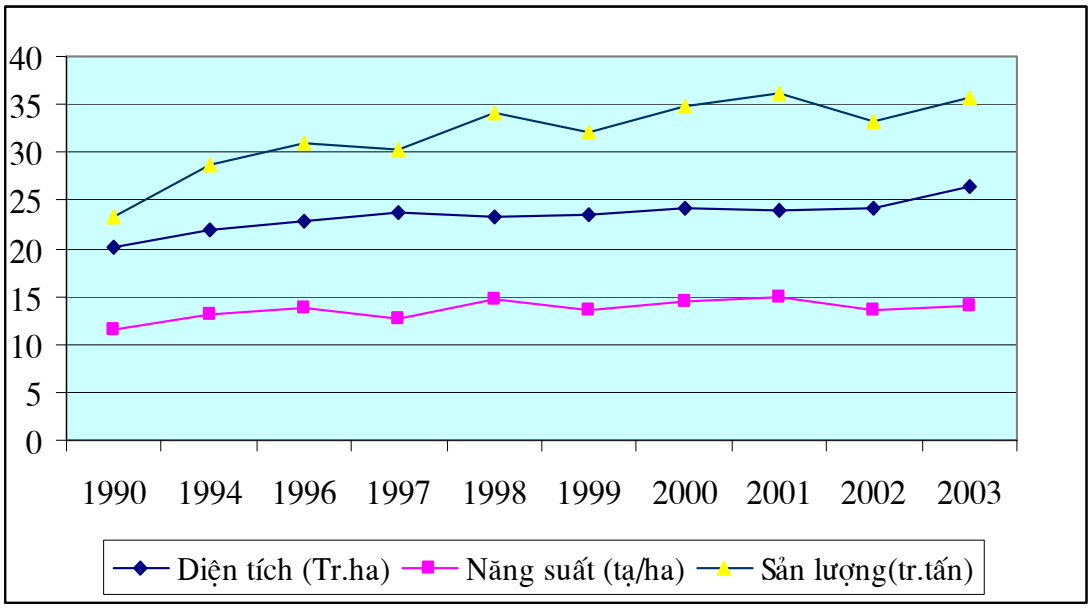
Số hộ vay vốn



Ghi chú: 1. Vay từ ngân hàng NN; 2. Vay từ ngân hàng chính sách
3. Vay từ quỹ tín dụng; 4. Vay từ hội nông dân
5. Vay từ HTX NN; 6. Vay từ nguồn khác

Đồ thị 3.2. Số hộ điều tra vay vốn từ các nguồn vay của Việt Nam năm 2003
(Điều tra hộ nông dân trên 7 vùng. ĐHNNI Hà Nội - 2003)

Thí dụ 3: Đồ thị đường gấp khúc



Đồ thị 3.3. Diện tích, năng suất và sản lượng lạc trên thế giới 1990 - 2003
(*Nguồn*: FAOSTAT, Agricultural Data, 24/5/2004)

Thí dụ 4: Bản đồ thống kê về các vùng sinh thái của Việt Nam


CHINA
FAR NORTH
NORTH WEST
QUANG NINH
HAI PHONG
HA NOI
NINH BINH
THANH HOA
LAOS
NORTHERN DELTA & MIDLANDS
NORTH CENTRAL REGION
PACIFIC OCEAN
HOANG SA
THUA THIEN HUE
QUANG NAM
DA NANG
SOUTH CENTRAL COAST & CENTRAL HIGHLANDS
CAMBODIA
KHANH HOA
LAM DONG
SOUTHEAST REGION
HOCHIMINH CITY
TRUONG SA
BA RIA VUNG TAU
MEKONG DELTA

c) Những vấn đề chú ý khi xây dựng biểu đồ và đồ thị thống kê:

Yêu cầu của một đồ thị hay biểu đồ thống kê là chính xác, đầy đủ, dễ hiểu và thể hiện tính thẩm mỹ. Do đó, khi xây dựng đồ thị thống kê cần chú ý các điểm sau:

* Lựa chọn đồ thị cho phù hợp với nội dung, tính chất của các số liệu cần diễn đạt.

Mỗi loại đồ thị có khả năng diễn đật khác nhau, đồng thời có thể diễn tả nhiều khía cạnh. Vì thế, cần lựa chọn loại đồ thị diễn tả phù hợp nhất, dễ quan sát nhất.

Thí dụ: Khi cần mô tả cơ cấu của hiện tượng thì nên dùng đồ thị hình tròn. Ngược lại khi cần biểu diễn biến động của hiện tượng theo thời gian thì nên dùng đồ thị đường gấp khúc hoặc hình cột... .

* Xác định quy mô của đồ thị cho thích hợp.

Quy mô của đồ thị được thể hiện qua chiều dài, chiều rộng và mối quan hệ tỷ lệ giữa 2 chiều này. Quy mô thích hợp là tuỳ thuộc vào mục đích sử dụng. Thí dụ, dùng đồ thị trong báo cáo phân tích thì không nên dùng đồ thị quá lớn, nhưng dùng vào tuyên truyền, cổ động thì lại không nên dùng đồ thị quá nhỏ.

* Các thanh đo tỷ lệ cần thống nhất và chính xác.

* Cần ghi số liệu, đơn vị tính, thời gian không gian của hiện tượng nghiên cứu sao cho thích hợp với từng loại đồ thị cụ thể. Đặc biệt cần ghi chú rõ các ký hiệu, màu sắc quy ước được dùng trong đồ thị.

* Trong thực tế vẽ đồ thị, người ta thường dùng các phần mềm máy tính. Phần mềm EXCEL được sử dụng khá phổ biến, rộng rãi và rất tiện lợi. Nó có thể liên kết rất tốt với các phần mềm soạn thảo văn bản như Winwords. Vì vậy chúng ta nên sử dụng EXCEL để vẽ đồ thị.

CÂU HỎI THẢO LUẬN CHƯƠNG III

1. Thế nào là phân tổ thống kê? Ý nghĩa và cách phân tổ thống kê? Cho ví dụ minh hoạ?
2. Các hình thức trình bày các tài liệu thống kê? Các loại bảng thống kê? Cho ví dụ minh họa?

Chương IV

THỐNG KÊ MỨC ĐỘ CỦA HIỆN TƯỢNG KINH TẾ XÃ HỘI

Các hiện tượng kinh tế xã hội tồn tại trong những điều kiện thời gian và địa điểm nhất định. Mỗi đặc điểm cơ bản của hiện tượng có thể được biểu diễn bằng các mức độ khác nhau. Nghiên cứu các mức độ của hiện tượng kinh tế xã hội là một trong những vấn đề quan trọng của phân tích thống kê vì nó nhằm biểu hiện quy mô, kết cấu và mức độ tập trung hay phân tán của hiện tượng trong những điều kiện thời gian và địa điểm cụ thể.

1. Số tuyệt đối

1.1. Khái niệm và ý nghĩa

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

Ví dụ:

1. Tổng dân số Việt Nam tại 0 giờ ngày 1.4.1999 là 76324753 người.

2. Tổng số sinh viên theo danh sách lớp kế toán 49 A năm học 2005-2006 là 95 người.

Số tuyệt đối là kết quả của điều tra và tổng hợp thống kê. Nó có thể biểu hiện số đơn vị của tổng thể hay từng bộ phận của tổng thể, như số nhân khẩu, số sinh viên... hoặc là trị số của lượng biến theo một tiêu tiêu số lượng nào đó như tổng chi phí sản xuất, tổng doanh thu...

Số tuyệt đối luôn phản ánh một nội dung kinh tế, chính trị trong điều kiện lịch sử nhất định. Nó phản ánh rất cụ thể, chính xác sự thật khách quan không thể phủ nhận được. Ví dụ: Tổng số tiền học bổng của một sinh viên một tháng là 120.000 đồng.

Bằng các số tuyệt đối này có thể xác định một cách cụ thể được nguồn tài nguyên, tài sản, khả năng tiềm tàng, kết quả sản xuất và các thành tựu khác của một doanh nghiệp, một địa phương hay toàn quốc.

Nó còn là căn cứ để tính các chỉ tiêu phân tích khác (số tương đối, số bình quân).

1.2. Tác dụng của số tuyệt đối

- Phục vụ cho công tác quản lý doanh nghiệp, quản lý nhà nước, vì muốn quản lý và kinh doanh được thì trước hết người quản ký phải biết được tình hình cụ thể về mọi mặt. Thí dụ: Biết được tình hình đất đai, lao động, vốn... từ đó mới có kế hoạch sắp xếp sử dụng một cách hợp lý các nguồn lực đó vào kinh doanh và quản lý xã hội.

- Phục vụ cho công tác kế hoạch như lập và kiểm tra thực hiện kế hoạch, các dự án.

- Căn cứ tính toán, so sánh các chỉ tiêu thống kê.

1.3. Các loại số tuyệt đối

a) Số tuyệt đối thời kỳ:

Số tuyệt đối thời kỳ phản ánh quy mô, khối lượng của hiện tượng trong một khoảng thời gian nhất định. Nó hình thành được là nhờ sự tích luỹ về lượng của hiện tượng suốt thời gian nghiên cứu.

Thí dụ: Khối lượng sữa hộp đã chế biến xong của Công ty sữa Hà Nội năm 2005 là 1000 triệu hộp. Tổng doanh thu của doanh nghiệp B năm 2004 là 200 tỷ đồng.

Đặc điểm:

- Phản ánh quá trình của hiện tượng.

- Các số tuyệt đối thời kỳ của một chỉ tiêu có thể cộng được với nhau để được số lượng của thời kỳ lớn hơn.

- Thời kỳ càng dài thì trị số của chỉ tiêu càng lớn.

b) Số tuyệt đối thời điểm:

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

Thí dụ:

- Giá trị hàng tồn kho cuối kỳ của Công ty May 10 năm 2005 là 800 triệu đồng.

- Số dân Việt Nam tại thời điểm 0 giờ ngày 1.4.1999 là 76.324.753 người.

Đặc điểm: Số tuyệt đối thời điểm chỉ phản ánh trạng thái của hiện tượng. Các số tuyệt đối thời điểm của cùng một chỉ tiêu ở các thời điểm khác nhau không cộng lại được với nhau được. Thời điểm khác nhau, trị số của chỉ tiêu cũng khác nhau.

1.4. Đơn vị tính

Số tuyệt đối trong thống kê bao giờ cũng có đơn vị tính cụ thể. Các đơn vị tính của số tuyệt đối như sau:

a) Đơn vị hiện vật:

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

+ Đơn vị đo chiều dài

+ Đơn vị đo diện tích

+ Đơn vị đo trọng lượng

+ Đơn vị đo khối lượng

+ Đơn vị đo dung tích

+ Đơn vị đo thời gian

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

+ Đơn vị đo theo quy ước: huyện, xã, tỉnh...

Các đơn vị hiện vật này phản ánh chính xác giá trị sử dụng của sản phẩm. Tuy nhiên, nó có nhược điểm là không tổng hợp được các sản phẩm khác loại và những công việc có tính chất dịch vụ khác nhau. Để khắc phục một phần nhược điểm này, người ta sử dụng đơn vị hiện vật quy đổi.

b) Đơn vị hiện vật quy đổi:

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

Thí dụ: quy đổi lao động ngoài độ tuổi quy định thành lao động trong tuổi, quy đổi khoai, ngô về lương thực quy thóc.

Cơ sở để xác định hệ số quy đổi là giá trị sử dụng của sản phẩm, đôi khi người ta cũng dùng giá trị sản phẩm để làm cơ sở tính đổi.

Đơn vị tính này có tác dụng dùng để tổng hợp các sản phẩm cùng loại nhưng có quy cách, phẩm chất khác nhau. Song, nó cũng không thể tổng hợp hết được tất cả các loại sản phẩm khác tên, không phản ánh được giá trị sử dụng thực tế nên có tính trìu tượng thiếu cụ thể của đơn vị hiện vật.

c) Đơn vị tiền tệ:

Đơn vị tiền tệ là dùng các loại tiền như Đồng, Đô la, EURO... để biểu hiện giá trị sản phẩm, hoặc dịch vụ.

Đơn vị tiền tệ được sử dụng rộng rãi nhất trong thống kê vì nó có ưu điểm là tổng hợp được nhiều loại sản phẩm có giá trị sử dụng và đơn vị đo lường khác nhau.

Nhược điểm của nó là phụ thuộc vào biến động của giá cả nên không có tính chất so sánh theo thời gian.

Thí dụ: Tổng sản phẩm trong nước theo giá thực tế năm 2003 của Việt Nam là 605.586 tỷ đồng (*Niên giám thống kê 2003*).

Để khắc phục nhược điểm do ảnh hưởng của thay đổi giá cả, người ta dùng giá cố định hoặc chỉ số lạm phát giá cả để loại trừ ảnh hưởng của giá thực tế.

d) Đơn vị thời gian lao động:

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

Thí dụ: Trong công nghiệp may, công nghiệp sản xuất đồ gỗ... đơn vị này dùng nhiều trong định mức thời gian lao động, tính năng suất lao động và quản lý lao động.

2. SỐ TƯƠNG ĐỐI

2.1. Khái niệm và ý nghĩa

a) Khái niệm:

Số tương đối trong thống kê là chỉ tiêu biểu hiện quan hệ so sánh giữa hai lượng tuyệt đối của hiện tượng nghiên cứu. Thường có 2 trường hợp so sánh sau:

- So sánh 2 lượng tuyệt đối của hiện tượng cùng loại nhưng khác nhau về thời gian hoặc không gian. Thí dụ: Doanh thu của Công ty sữa Hà Nội năm 2005 so với năm 2004 là 120%. Doanh thu của Công ty sữa Hà Nội năm 2005 so với kế hoạch năm 2005 là 110 %.

- So sánh 2 lượng tuyệt đối của hai hiện tượng khác loại nhưng có liên quan với nhau. Thí dụ: Mật độ dân số; GDP trung bình 1 đầu người.

Hình thức biểu hiện của số tương đối là số lần, phần trăm (%); phần nghìn (‰), hoặc kết hợp đơn vị tính của 2 chỉ tiêu khi so sánh (kép), ví dụ người/km^2, kg/người.

b) Ý nghĩa:

- Số tương đối là 1 trong những chỉ tiêu phân tích thống kê. Tuỳ theo mục đích nghiên cứu mà nó cho ta biết rõ hơn đặc điểm của hiện tượng, hay bản chất hiện tượng một cách sâu sắc hơn.

- Dùng để giữ bí mật số tuyệt đối.

2.2. Các loại số tương đối

Các số tương đối trong thống kê không phải là do kết quả của điều tra và tổng hợp thống kê mà là do kết quả so sánh 2 số tuyệt đối đã có. Vì vậy mỗi số tương đối đều có gốc so sánh. Tuỳ theo mục đích so sánh mà gốc so sánh được chọn khác nhau. Do đó, khi sử dụng gốc so sánh khác nhau mà có các loại số tương đối sau:

a) Số tương đối kế hoạch:

- Dùng để lập và kiểm tra tình hình thực hiện kế hoạch về một chỉ tiêu nào đó. Có 2 loại số tương đối kế hoạch:

* *Số tương đối nhiệm vụ kế hoạch*: Là tỷ lệ so sánh mức độ kế hoạch với mức độ thực tế của chỉ tiêu ấy ở kì gốc.

- Công thức tính:

$$\text{Số tương đối nhiệm vụ kế hoạch} = \frac{\text{Số tuyệt đối kì kế hoạch}}{\text{Số tuyệt đối kì gốc}} \times 100$$

- Thí dụ: 1 doanh nghiệp có doanh thu thực tế năm 2004 là 600 tỷ đồng; kế hoạch năm 2005 của công ty là 660 tỷ đồng. Thực hiện năm 2005 là 700 tỷ đồng.

Số tương đối nhiệm vụ kế hoạch năm 2005 là:

$$\text{Số tương đối nhiệm vụ kế hoạch doanh thu 2005} = \frac{660}{600} \times 1\ 00 = 110\ \%$$

* *Số tương đối thực hiện kế hoạch*: Là tỉ lệ so sánh giữa mức độ thực tế đạt được trong kì nghiên cứu với mức độ kế hoạch đề ra cùng kì của một chỉ tiêu nào đó.

- Mục đích sử dụng: Xác định mức độ thực hiện nhiệm vụ kế hoạch trong một thời gian nhất định (tháng, quý, năm).

- Công thức tính:

$$\text{Số tương đối thực hiện kế hoạch} = \frac{\text{Số tuyệt đối thực tế đạt được}}{\text{Số tuyệt đối kế hoạch đề ra}} \times 100$$

Lấy lại thí dụ trên ta có:

Số tương đối hoàn thành kế hoạch năm 2005 là:

$$\text{Số tương đối hoàn thành kế hoạch doanh thu 2005} = \frac{700}{660} \times 1\ 00 = 106{,}06\ \%$$

b) Số tương đối động thái:

Số tương đối động thái biểu hiện sự so sánh mức độ của hiện tượng ở 2 thời kì hay 2 thời điểm khác nhau nhằm phản ánh rõ hơn tình hình của hiện tượng ở thời kỳ hay thời điểm nghiên cứu.

- Công thức tính:

$$\text{Số tương đối động thái (\%)} = \frac{\text{Số tuyệt đối kì báo cáo (kì nghiên cứu)}}{\text{Số tuyệt đối kì gốc}} \times 100$$

+ Kì báo cáo là kì đang nghiên cứu.

+ Kì gốc là kì trước dùng làm gốc so sánh.

Ví dụ: Lấy ví dụ trên

$$\text{Số tương đối động thái doanh thu (2005 so với 2004)} = \frac{700}{600} \times 100 = 116{,}67\%$$

Mối quan hệ giữa số tương đối động thái với số tương đối hoàn thành kế hoạch và số tương đối nhiệm vụ kế hoạch là:

$$\text{Số tương đối động thái} = \text{Số tương đối hoàn thành kế hoạch} \times \text{Số tương đối nhiệm vụ kế hoạch}$$

c) Số tương đối kết cấu:

Số tương đối kết cấu là tỷ lệ so sánh giữa số tuyệt đối của từng bộ phận cấu thành nên tổng thể với số tuyệt đối của tổng thể hiện tượng nghiên cứu nhằm nghiên cứu cấu thành của hiện tượng. Nếu kết cấu thay đổi sẽ thấy được nguyên nhân thay đổi bản chất của hiện tượng trong các điều kiện khác nhau.

- Công thức:

$$\text{Số tương đối kết cấu (\%)} = \frac{\text{Số tuyệt đối từng tổ}}{\text{Số tuyệt đối của tổng thể}} \times 100$$

Thí dụ: Lấy lại thí dụ trên, Công ty có 2 phân xưởng. Phân xưởng A doanh thu thực hiện năm 2005 là 300 tỷ đồng, còn lại là doanh thu của phân xưởng B.

$$\text{Số tương đối kết cấu doanh thu phân xưởng A (2005)} = \frac{300}{700} \times 1\,00 = 42{,}86\ \%$$

d) Số tương đối so sánh (số tương đối không gian)***:***

Số tương đối so sánh hay còn gọi là số tương đối không gian là kết quả so sánh giữa hai số tuyệt đối của cùng hiện tượng nhưng khác nhau về không gian, hoặc so sánh giữa 2 bộ phận trong cùng một tổng thể nhằm so sánh điều kiện của hiện tượng ở 2 nơi ta nghiên cứu.

Công thức tính:

$$\text{Số tương đối so sánh (\%)} = \frac{\text{Số tuyệt đối bộ phận A}}{\text{Số tuyệt đối bộ phận B}} \times 100$$

Thí dụ : Lấy lại ví dụ trên, ta so sánh doanh thu của 2 phân xưởng A và B:

$$\text{Số tương đối so sánh doanh thu phân xưởng A so B (2005)} = \frac{300}{400} \times 1\,00 = 75{,}00\%$$

e) Số tương đối cường độ:

Số tương đối cường độ là kết quả so sánh 2 số tuyệt đối của 2 hiện tượng khác loại nhưng có liên quan với nhau nhằm nói lên trình độ phổ biến của hiện tượng. Nó được sử dụng rộng rãi trong thực tế để biểu hiện trình độ phát triển sản xuất, trình độ bảo đảm

mức sống vật chất, văn hoá của dân cư trong một nước hay địa phương. Nó còn dùng để so sánh trình độ phát triển sản xuất và đời sống giữa các quốc gia với nhau.

Công thức tính:

$$\text{Số tương đối cường độ} = \frac{\text{Số tuyệt đối của hiện tượng A}}{\text{Số tuyệt đối của hiện tượng B}}$$

Thí dụ: Mật độ dân số; số bác sĩ trên 1000 dân...

2.3. Nguyên tắc sử dụng số tương đối

Số tương đối trong thống kê là kết quả so sánh giữa 2 số tuyệt đối đã có. Vì vậy, để phát huy được tác dụng của nó trong phân tích thống kê khi sử dụng phải tôn trọng các nguyên tắc sau đây.

* Số tương đối phải được tính ra từ 2 số tuyệt đối có quan hệ với nhau, so sánh có ý nghĩa hay đảm bảo nguyên tắc *"có thể so sánh được"*. Yêu cầu của nguyên tắc này là 2 số tuyệt đối đem so sánh với nhau phải:

- Cùng một chỉ tiêu nghiên cứu (cùng một nội dung kinh tế);
- Phạm vi tính toán thống nhất;
- Phương pháp tính, đơn vị tính thống nhất.

* Kết hợp số tương đối và số tuyệt đối khi phân tích cùng hiện tượng. Trong thực tế trừ một số trường hợp mang tính chất bí mật không được phép công bố số tuyệt đối (bí mật quân sự), người ta thường kết hợp giữa số tuyệt đối và số tương đối để nhận thức bản chất của hiện tượng một cách chính xác.

Thí dụ : Theo số người nhập viện và tử vong, nếu 1 ngày chỉ có 2 ca nhập viện, trong đó 1 ca không cứu chữa được, khi đó ta công bố có 50% ca nhập viện không cứu chữa được, con số này nghe thật khủng khiếp. Song, nếu ta kết hợp với số tuyệt đối mà công bố rằng, có 50% số ca nhập viện tức là 1 ca không cứu chữa được thì sự việc đơn giản hơn.

3. CÁC CHỈ TIÊU ĐO KHUYNH HƯỚNG TẬP TRUNG

3.1. Số trung bình cộng

a) Khái niệm và ý nghĩa:

Một tổng thể thống kê thường bao gồm nhiều đơn vị. Các đơn vị này có bản chất giống nhau nhưng biểu hiện về lượng theo từng tiêu thức ở các đơn vị tổng thể thường khác nhau.

Thí dụ: Tổng dân số Việt Nam, có cùng quốc tịch là Việt Nam nhưng độ tuổi của từng người dân khác nhau. Muốn biết độ tuổi trung bình của tổng thể dân số ở một thời gian nào đó ta dùng số bình quân cộng.

Do đó, khi muốn biểu hiện đặc tính chung của tổng thể theo tiêu thức số lượng nào đó, ta dùng số bình quân cộng.

Số bình quân trong thống kê biểu hiện mức độ đại biểu theo một tiêu thức số lượng nào đó của tổng thể đồng chất bao gồm nhiều đơn vị cùng loại.

Số bình quân cộng trong thống kê thường dùng nhằm:

- Phản ánh mức độ trung bình của hiện tượng;
- So sánh hai tổng thể hiện tượng nghiên cứu cùng loại, không có cùng quy mô;
- Sử dụng trong công tác kế hoạch hoá.

Chú ý: Vì số bình quân mang tính chất đại diện cho tổng thể, nên để số bình quân có tính đại biểu cao thì cần đảm bảo sao cho số đơn vị tổng thể dùng để tính số bình quân phải đủ lớn...

b) Các loại số bình quân:

Số trung bình cộng được tính theo công thức chung là:

$$\text{Số bình quân cộng} = \frac{\text{Tổng trị số lượng biến tiêu thức}}{\text{Tổng số đơn vị tổng thể}}$$

Căn cứ vào nguồn tài liệu có các công thức tính toán số bình quân sau:

* Số bình quân cộng giản đơn: Áp dụng khi lượng biến Xi có các tần số fi bằng nhau hoặc bằng 1.

Thí dụ: 1 nhóm gồm 5 công nhân có mức lương như sau: 500, 650, 800, 950, 1000 (ngàn đồng).

$$\text{Tiền lương bình quân 1 người} = \frac{500 + 650 + 800 + 950 + 1000}{5} = 780 \text{ ngàn đồng}$$

Công thức tổng quát:

$\bar{x} = \frac{\Sigma x_i}{n}$	Trong đó: - $\bar{x}$: Số bình quân - X_i là trị số của đơn vị thứ i (i = 1,2,... n); - n là số đơn vị tổng thể

* Số bình quân cộng gia quyền: Áp dụng khi mỗi lượng biến Xi được gặp nhiều lần, nghĩa là có tần số fi.

Thí dụ : Lấy lại thí dụ trên, ta quan sát tiền lương không phải của 5 người mà của 50 người thể hiện qua bảng 1.4.

Bảng 1.4.

Tiền lương (1000 đồng) Xi	Số công nhân (fi)	Xi fi	Công thức tính
500	5	2500	$\overline{X} = \frac{40200}{50} = 804$ ngàn đồng / người
650	8	5200	
800	20	16000	
950	10	9500	
1000	7	7000	
Cộng	**50**	**40200**	

Mức lương 500 ngàn đồng có 5 công nhân, 800 ngàn đồng có 20 công nhân... Muốn tính mức lương bình quân 1 người 1 tháng thì nhân mức lương với số người cùng mức lương đó, cộng tiền lương của các nhóm với nhau và chia cho toàn bộ số công nhân.

Tiền lương bình quân 1 người/1 tháng là 804 ngàn đồng.

Công thức tổng quát:

$$\overline{x} = \frac{\Sigma x_i f_i}{\Sigma f_i}$$

Trong đó:

$\overline{x}$: Số bình quân

x_i : Lượng biến bình quân của tổ thứ i;

f_i : Số đơn vị của tổ thứ i (f_i còn gọi là tần số hay quyền số).

+ *Một số trường hợp đặc biệt:*

- Không biết f_i (số đơn vị từng tổng thể từng tổ), cho biết tỷ lệ số đơn vị tổng thể từng tổ $\left[\frac{f_i}{\Sigma f_i} = S_i\right]$ (tần suất) thì số bình quân cộng gia quyền được tính theo công thức:

$$\overline{X} = \Sigma x_i S_i$$ Trong đó S_i: Tần suất.

- Lượng biến X_i không phải là một trị số xác định mà một khoảng trị số có 2 giới hạn (trên, dưới):

Tính trị số giữa mỗi tổ $= \frac{X_{i\min} + X_{i\max}}{2} = \overline{X}_i$

Nhân trị số giữa với tần số hoặc tần suất và chia cho tổng số đơn vị tổng thể hoặc cho 100.

Công thức tổng quát:

$$X = \frac{\Sigma\left[\frac{X_{i\max} + X_{i\min}}{2}\right] \times f_i}{\Sigma f_i}$$

Hoặc : nếu tần suất tính theo %

$$\bar{X} = \frac{\Sigma\left[\frac{X_{i\max} + X_{i\min}}{2}\right] \times S_i}{100}$$

Trong đó:

X_{imin}: giới hạn dưới của tổ i

X_{imax}: giới hạn trên của tổ i

- Nếu $f_1 = f_2 = ... = f_n = a$ (hằng số) thì: $\bar{X} = \frac{\Sigma X_i}{n}$

* Số bình quân cộng điều hoà: Áp dụng khi không có tài liệu về số đơn vị tổng thể của mỗi tổ (fi), mà chỉ có tài liệu về các lượng biến Xi và Mi = Xi.fi.

Thí dụ:

Bảng 2.4.

Tiền lương (1000 đồng/người) Xi	Xi fi	(fi = Mi / Xi)	Công thức tính
500	2500	5	$f_i = \frac{M_i}{X_i}$
650	5200	8	
800	16000	20	
950	9500	10	$\bar{X} = \frac{M_1 + M_2 + ... + M_i}{\frac{M_1}{X_1} + \frac{M_2}{X_2} + ... + \frac{M_i}{X_i}} \quad (i = 1...n)$
1000	7000	7	
Cộng	**40200**	**50**	

Từ bảng 2.4, quan sát cột X_i và $x_i.f_i = M_i$, tài liệu chỉ cho chúng ta biết lượng biến của từng tổ và tổng lượng biến toàn tổ.

Cách tính như sau:

- Lấy lượng biến toàn tổ chia cho lượng biến trung bình của tổ, được số đơn vị mỗi tổ.

- Cộng số đơn vị mỗi tổ ta được tổng số đơn vị tổng thể.

- Tổng lượng biến các tổ chia cho tổng số đơn vị tổng thể.

Công thức tổng quát:

$$\bar{x} = \frac{\Sigma M_i}{\Sigma \frac{M_i}{x_i}}$$

Trong đó:

M_i là tổng trị số lượng biến của tổ thứ i

x_i là lượng biến bình quân của tổ thứ i.

- Chú ý: Khi tổng lượng biến của các tổ bằng nhau tức là $M_1 = M_2 = ... = M_n = M$ thì quá trình tính toán sẽ đơn giản hơn như sau:

$$\bar{x} = \frac{n}{\Sigma \frac{1}{x_i}}$$

Trong đó:

n là số tổ (lượng biến);

x_i là lượng biến bình quân tổ thứ i.

Số bình quân tính theo công thức này gọi là *số bình quân điều hoà giản đơn*.

c) Đặc điểm và nguyên tắc sử dụng số bình quân:

Khi tính các số bình quân trong thống kê, chúng ta san bằng mọi chênh lệch lượng biến theo một tiêu thức số lượng nào đó của các đơn vị tổng thể (đơn vị cá biệt) làm cho tổng thể từ phức tạp trở nên khái quát chung. Vì vậy, để sử dụng số bình quân một cách khoa học và chính xác cần phải đảm bảo một số nguyên tắc sau đây:

** Số bình quân chỉ được tính trong một tổng thể đồng chất*

Tổng thể đồng chất là một tổng thể bao gồm những đơn vị tổng thể có chung tính chất, thuộc cùng một loại hình kinh tế xã hội xét theo một tiêu thức nào đó.

Trong một tổng thể đồng chất thì tính chất của các đơn vị tổng thể là giống nhau chỉ khác nhau về lượng cụ thể giữa các đơn vị. Vì vậy, khi tính số bình quân, tức là ta san bằng lượng biến theo tiêu thức số lượng nào đó thì các yếu tố ngẫu nhiên sẽ bù trừ cho nhau và số bình quân sẽ đại diện cho tất cả các mức độ khác nhau trong tổng thể.

Nếu tính trong một tổng thể không đồng chất (tức là các đơn vị tổng thể không những khác nhau về lượng cụ thể mà còn khác nhau về tính chất hay loại hình) ta không thể san bằng lượng biến theo một tiêu thức số lượng nào đó của các đơn vị khác nhau về tính chất được. Khi đó ta chỉ tính được một số bình quân hình thức, giả tạo, không đại biểu cho các mức độ khác nhau của các đơn vị.

Thí dụ: Không thể tính năng suất của lúa + ngô/1 ha gieo trồng được vì đây là tổng thể không đồng chất. Ta chỉ có thể tính năng suất lúa hoặc ngô cho 1 ha gieo trồng lúa hoặc ngô.

** Cần kết hợp giữa số bình quân chung với số bình quân tổ*

Số bình quân chung (tổng thể) che lấp sự chênh lệch lượng biến của các bộ phận cấu thành tổng thể. Vì vậy, nếu chỉ sử dụng số bình quân chung của tổng thể để nghiên

cứu sẽ không thấy được đầy đủ tình hình phát triển giữa các bộ phận của tổng thể hiện tượng đó.

- Thí dụ: Kết quả học tập của 2 sinh viên trong một lớp cùng một học kỳ như sau:

Bảng 3.4.

Môn thi	Sinh viên A		Sinh viên B	
	Điểm thi môn học (Xi)	Số đơn vị học trình (fi)	Điểm thi môn học (Xi)	Số đơn vị học trình (fi)
Toán	5	6	8	6
Anh văn	6	4	6	4
Kinh tế vi mô	5	4	5	4
Triết học	8	3	4	3
Bình quân	5,76	17	6,12	17

Nếu dựa vào điểm trung bình các môn thi để so sánh kết quả học tập của 2 người thì ta có nhận xét sinh viên B có kết quả học tập tốt hơn. Nhưng nếu căn cứ vào điểm thi từng môn thì rõ ràng kết quả học tập của sinh viên A tốt hơn, vì không có môn nào dưới điểm 5, trong đó sinh viên B lại có.

Như vậy, khi so sánh 2 tổng thể cùng loại, cùng quy mô thì phải dùng số bình quân tổ bổ sung cho số bình quân chung.

** Dùng dãy số phân phối bổ sung cho số bình quân chung*

Tổng thể hiện tượng cấu thành bởi các đơn vị tổng thể có lượng biến khác nhau. Có một số đơn vị có lượng biến lớn hơn hoặc nhỏ hơn mức độ điển hình của hiện tượng. Số đơn vị có lượng biến lớn hơn hay nhỏ hơn giữa các tổng thể hiện tượng cùng loại cũng khác nhau. Khi so sánh 2 hiện tượng cùng loại nhưng có kết cấu tổng thể khác nhau, phải dùng dãy số phân phối để giải thích cho mức độ đại biểu của số bình quân chung.

Thí dụ trên: Câu hỏi đặt ra tại sao điểm thi từng môn của sinh viên B thấp hơn sinh viên A mà điểm trung bình của sinh viên B lại cao hơn sinh viên A?

Trả lời câu hỏi này, chúng ta dựa vào kết cấu các học trình theo điểm thi. Sinh viên A có điểm trung bình thấp hơn sinh viên B vì tỷ trọng số đơn vị học trình có điểm cao (điểm 6 và 8) của sinh viên A (41,18%) thấp hơn sinh viên B (58,82%).

Số đơn vị học trình và điểm thi tạo thành 1 dãy số phân phối.

3.2. Số trung vị (Me-Median)

a) Khái niệm:

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

b) Phương pháp xác định số trung vị:

+ Tài liệu không phân tổ: Trước hết cần sắp xếp lượng biến theo thứ tự từ nhỏ đến lớn.

Nếu số lượng biến (n) lẻ thì số trung vị là lượng biến đứng ở vị trí thứ giữa dãy số, tức là ở vị trí thứ $\left(\frac{n+1}{2}\right)$. Khi đó Me được xác định theo công thức:

$Me = X_{(n+1)/2}$; trong đó X là lượng biến đúng ở vị trí $\left(\frac{n+1}{2}\right)$

Thí dụ: Tiền lương tháng của 1 tổ công nhân gồm 5 người như sau:

500; 600; 800; 1000; 1500 thì Me = 800

Nếu n chẵn lẻ thì số trung vị là trung bình cộng lượng biến đứng ở vị trí thứ $\left(\frac{n}{2}\right)$ và ở vị trí thứ $\left(\frac{n+2}{2}\right)$. Khi đó Me được xác định theo công thức:

$$Me = \frac{X_{(n/2)} + X_{(n+2)/2}}{2}$$

Me: Số trung vị

$X_{(n/2)}$: Lượng biến đứng ở vị trí thứ $\left(\frac{n}{2}\right)$

$X_{(n+2)/2}$: Lượng biến đứng ở vị trí thứ $\left(\frac{n+2}{2}\right)$

Thí dụ trên: n = 6 500; 600; 800; 1000; 1500 ; 2000

$$Me = \frac{(800 + 1000)}{2} = 900$$

+ Tài liệu phân tổ

- *Không có khoảng cách tổ*: Ta xác định tổ chứa trung vị.

Thí dụ:

TT	Tuổi	Số người
1	18	12
2	20	20
3	21	30
4	22	50
5	23	18
Cộng	?	130

Tổ chứa số trung vị là tổ lượng biến đứng ở vị trí thứ

$$\frac{\Sigma fi + 1}{2} \quad \text{với} \quad \Sigma fi = n$$

Trong thí dụ này, tổ chứa Me = (130 + 1)/2 = 65,5 đứng ở vị trí thứ 65,2; tức là tổ 4

** Có khoảng cách tổ:*

Để xác định số trung vị, trước tiên ta tìm tổ chứa số trung vị. Tổ chứa số trung vị là tổ có tần số tích luỹ bằng $\left(\frac{\sum fi + 1}{2}\right)$.

Sau đó số trung vị được tính theo công thức:

$$Me = x_{Me(min)} + h_{Me} \frac{\frac{\Sigma fi}{2} - S_{Me-1}}{f_{Me}}$$

Trong đó:

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

h_{Me}: Khoảng cách tổ của tổ chứa số trung vị

Σf_i: Tổng số các tần số

$S_{Me\ -\ 1:}$ Tần số tích luỹ của tổ đứng trước tổ có số trung vị

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

Ví dụ: Tìm số trung vị về khối lượng trứng gà giống theo tài liệu sau:

Khối lượng (g)	Số quả (fi)	Tần số tích luỹ (cộng dồn)

- Căn cứ vào tần số tích luỹ (tần số

80 – 84	10	10
84 – 88	20	30
88- 92	120	150
92 – 96	150	300
96 – 100	400	700
100 – 104	200	900
104 – 108	60	960
108 - 112	40	1000
Cộng	1000	

cộng dồn) tổ có chứa số trung vị là tổ 5 (96 - 100).

Áp dụng công thức trên với:

$x_{Me(min)} = 96$; $h_{Me} = 4$; $\Sigma fi/2 = 500$;

$S_{Me-1} = 300$; $f_{Me} = 400$

Me = 98

** Tính chất của số trung vị*

Tổng độ lệch tuyệt đối giữa các lượng biến với số trung vị là một trị số nhỏ nhất.

$\Sigma | Xi - Me | = \min$ (không phân tổ)

$\Sigma | Xi - Me | fi = \min$ (phân tổ)

Tính chất này được áp dụng nhiều trong công tác kỹ thuật và phục vụ công cộng như xây dựng mạng lưới điện, đường ống dẫn nước, bố trí các trạm đỗ xe công cộng ở vị trí thuận lợi để có thể đạt được hiệu quả cao trong công tác phục vụ.

Trung vị có ưu điểm là không chịu ảnh hưởng của các lượng biến đầu mút trong dãy số lượng biến, dễ hiểu, dễ tính. Song có nhược điểm là không thể dùng để dự đoán vì không chính xác bằng số trung bình. Nó thường được dùng để thay thế hoặc để bổ sung cho trung bình khi cần thiết.

* Chú ý: Khi phân tích các hiện tượng kinh tế - xã hội có nhiều đơn vị quan sát , đôi lúc ta phải xét đến thứ bậc của các đơn vị của tổng thể nghiên cứu trong dãy số phân phối thành các phần bằng nhau: 3 phần, 4 phần, 10 phần. Tuỳ theo vị trí của các đơn vị trong dãy số mà có các tên gọi khác nhau.

- Nếu tổng thể chia thành ba phần đều nhau ta có tam phân vị;
- Nếu tổng thể chia thành bốn phần đều nhau ta có tứ phân vị;
- Nếu tổng thể chia thành 10 phần bằng nhau ta có thập phân vị.

** Ý nghĩa của tứ phân vị, thập phân vị:*

- Tứ phân vị, thập phân vị giúp ta xác định trị số lượng biến của các đơn vị đứng ở các vị trí nhất định trong một dãy số phân phối. Ngoài ra các chỉ tiêu trên còn giúp ta đo lường độ phân tán về lượng biến giữa các đơn vị đó.

3.3. Mốt (Mod- chúng số, kiểu số)

a) Khái niệm:

Mốt là biểu hiện của một lượng biến về tiêu thức nghiên cứu được gặp nhiều nhất trong tổng thể.

Nếu xác định trên đồ thị với trục tung là tần số, trục hoành là lượng biến thì ta có thể nói mốt là hoành độ của điểm có tung độ cao nhất.

b) Phương pháp xác định:

** Trường hợp 1*: Đối với dãy số lượng biến không có khoảng cách tổ thì mốt là lượng biến được gặp nhiều nhất trong dãy số lượng biến.

Thí dụ 2.1: Có tài liệu phân tổ sinh viên trong một lớp học (tiêu thức phân tổ là tuổi).

Tuổi (xi)	Số sinh viên (fi)
22	3
23	5
24	6
25	40
26	12
35	1
Cộng	67

Kí hiệu: Mo là trị số của mốt

=> Mo = 25 vì lượng biến này có tần số lớn nhất (f = 40)

** Trường hợp 2*: Đối với dãy lượng biến có khoảng cách tổ thì mốt là lượng biến mà trên đó chứa mật độ phân phối lớn nhất, tức là xung quanh lượng biến ấy tập trung tần số nhều nhất.

+ Tài liệu phân tổ có khoảng cách đều nhau

Công thức tính:

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

Trong đó:

Mo : Ký hiệu của mốt

$x_{Mo\ (min)}$: Giới hạn dưới của tổ chứa mốt

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

f_{Mo}: Tần số của tổ chức 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

Thí dụ: Có tài liệu phân tổ một loại trái cây theo khối lượng như sau:

Khối lượng (g/quả)	Số quả
80 – 84	10
84 – 88	20
88- 92	120
92 – 96	150
96 – 100	400
100 – 104	200
104 – 108	60
108 - 112	40
Cộng	1000

Yêu cầu xác định mốt của khối lượng quả?

→ Trước hết ta có thể xác định mốt vào tổ thứ 5 (96 – 100) vì tổ này có tần số lớn nhất (400 quả).

Từ đó ta xác định:

$x_{Mo\ (min)} = 96$ $h_{Mo} = 4$

$f_{Mo} = 400$ $f_{Mo-1} = 150$ $f_{Mo+1} = 200$

Từ công thức (2) ta có Mo = 98,2 gam

+ Tài liệu phân tổ có khoảng cách tổ không đều nhau, mốt vẫn được tính theo công thức trên, nhưng cần lưu ý là việc xác định tổ chứa mốt không căn cứ vào tần số mà căn cứ vào mật độ phân phối.

Công thức tính mật độ phân phối như sau:

$$Mi = \frac{hi}{fi}$$

Trong đó:

M_i là mật độ phân phối

fi là tần số

h_i là trị số khoảng cách tổ

** Trường hợp 3*: Số đơn vị của tổng thể nghiên cứu có khuynh hướng tập trung vào một vài lượng biến nhất định, trường hợp này ta có đa mốt.

Thí dụ: Có tài liệu phân tổ sắp xếp cặp vợ chồng theo số con của những người đó như sau:

Số con (xi)	Số cặp vợ chồng (fi)
0	19
1	680
2	750
3	61
4	10
5	6
Cộng	1526

Yêu cầu xác định mốt của tổng thể nghiên cứu.

=> Qua tài liệu phân tổ, ta thấy số đơn vị có khuynh hướng tập trung vào 2 lượng biến (1 con và 2 con) như vậy trường hợp này mốt có 2 trị số là 1 và 2.

c) Ý nghĩa của việc dùng mốt trong thống kê:

- Trong thống kê, mốt là chỉ tiêu có tác dụng bổ sung hoặc thay thế cho việc tính số bình quân số học trong trường hợp việc xác định số trung bình số học gặp khó khăn. Mốt cho ta thấy mức độ phát triển nhất của hiện tượng, mặt khác chỉ tiêu này không chịu ảnh hưởng của các lượng biến giữa các đơn vị tổng thể như số trung bình số học. Chẳng hạn khi nghiên cứu giá cả một mặt hàng nào đó trên thị trường, thông thường người ta không có đủ tài liệu để xác định giá trị trung bình và có thể không cần tính giá trị trung bình mà ta chỉ cần biết giá trị phổ biến nhất của mặt hàng nào đó.

- Mốt còn có tác dụng giúp cho các tổ chức sản xuất và thương nghiệp trong công tác nghiên cứu xem các mặt hàng nào được tiêu thụ nhiều nhất, như cỡ giầy dép, cỡ kiểu quần áo...

- Mốt không phụ thuộc vào giá trị ở hai đầu mút, thậm chí trong trường hợp giá trị ở đầu mút nhỏ và giá trị ở cuối dãy số rất lớn thì giá trị của mốt cũng không bị ảnh hưởng. Mốt có thể tính trong trường hợp lượng biến biến động trong phạm vi rất rộng hoặc rất hẹp.

Tuy mốt có nhiều ưu điểm song mốt cũng không dùng nhiều như trung vị và trung bình, có trường hợp không có mốt và không có giá trị xuất hiện nhiều nhất hoặc có trường hợp có hai hoặc ba mốt ta không thể xác định được giá trị trung tâm chính xác.

3.4. Mối quan hệ giữa số trung bình cộng, số trung vị và mốt

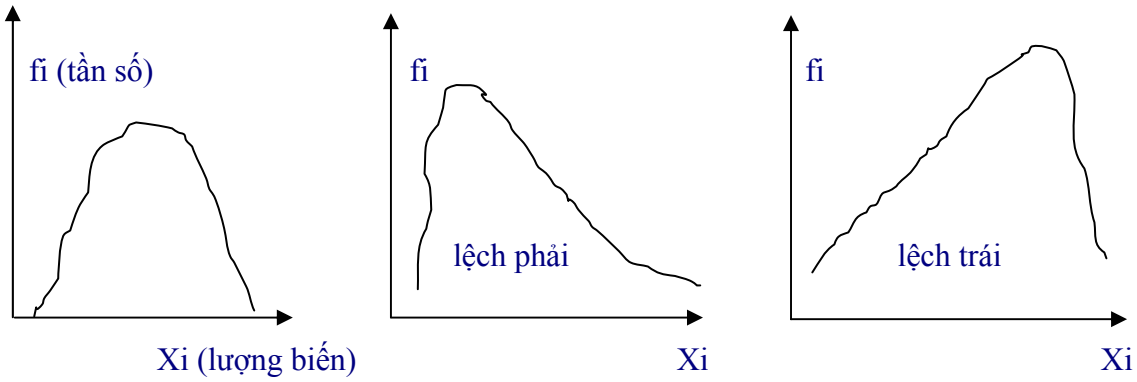
Dựa vào số trung bình, số trung vị và mốt người ta có thể biết được hình dáng phân phối của lượng biến trong tổng thể. Cụ thể là:

- Khi X b/q = Me = Mo thì phân phối đối xứng (phân phối chuẩn);

- Khi X b/q > Me > Mo thì phân phối lệch phải;

- Khi X b/q < Me < Mo thì phân phối lệch trái .

Điều này được thể hiện qua các đồ thị sau:



4. CÁC ĐẶC TRƯNG ĐO LƯỜNG ĐỘ PHÂN TÁN

4.1. Khái niệm

Thí dụ: Ta quan sát độ tuổi của 2 nhóm công nhân, mỗi nhóm gồm 5 người như sau:

Nhóm 1: 20 30 40 50 60 $\bar{x}_1 = 40$ tuổi

Nhóm 2: 38 39 40 41 42 $\bar{x}_2 = 40$ tuổi

Độ tuổi trung bình của 2 nhóm bằng nhau đều bằng 40 tuổi, nhưng ta chưa thể đánh giá chính xác rằng mức độ đồng đều về tuổi tác của 2 nhóm này như thế nào.

Nếu ta quan sát từng lượng biến trong mỗi nhóm ta thấy nhóm 2 lượng biến biến động ít và đồng đều hơn nhóm 1. Có thể nhận định rằng độ tuổi bình quân nhóm 2 đại diện cao hơn nhóm 1. Do đó sự biến động lượng biến tiêu thức có liên quan rất lớn đến mức độ đại biểu của số bình quân.

Sự biến động về lượng biến của các đơn vị tổng thể theo một tiêu thức nào đó gọi là độ phân tán của hiện tượng.

Để đo mức độ phân tán hay mức độ đại biểu của số bình quân người ta đã tính ra một loạt các đặc trưng gọi là các chỉ tiêu đo độ biến động tiêu thức.

4.2. Các chỉ tiêu đo độ biến động tiêu thức

a) Khoảng biến thiên (R) (còn gọi là toàn cự):

Khoảng biến thiên là độ lệch giữa lượng biến lớn nhất và lượng biến nhỏ nhất của tiêu thức nghiên cứu trong tổng thể:

$$R = Xmax - Xmin.$$

Trong đó: Xmax là lượng biến lớn nhất; Xmin là lượng biến nhỏ nhất.

Ý nghĩa: R càng lớn độ biến động tiêu thức càng lớn, tính chất đại biểu của số bình quân càng nhỏ và ngược lại.

Thí dụ: R1 = 60 - 20 = 40 R2 = 42 - 38 = 4

R1 > R2 $\bar{x}_1$ đại diện thấp hơn $\bar{x}_2$

- Ưu điểm: Đơn giản, biểu hiện rõ và cụ thể phạm vi biến động.

- Nhược điểm: Do không xét đến các lượng biến ở giữa nên tính chất phản ánh không đầy đủ, nhiều khi không nêu được tính biến động của tiêu thức.

b) Độ lệch tuyệt đối bình quân $\bar{d}$:

Độ lệch tuyệt đối bình quân là mức chênh lệch bình quân giữa các lượng biến và số bình quân cộng của các lượng biến đó. Vì tổng độ lệch bằng không, nên khi tính toán người ta phải lấy giá trị tuyệt đối của từng độ lệch.

Công thức tính như sau:

$$\bar{d} = \frac{\Sigma|x_i - \bar{x}|}{n} \quad \text{hay} \quad \bar{d} = \frac{\Sigma|x_i - \bar{x}| f_i}{\Sigma f_i}$$

Trong đó:

x_i là lượng biến thứ i

$\bar{x}$ là số bình quân

n (Σf_i) là số đơn vị tổng thể

Ý nghĩa: Độ lệch tuyệt đối bình quân càng nhỏ, độ biến thiên lượng biến càng ít, tính đại biểu của số bình quân càng lớn và ngược lại.

Thí dụ trên:

$\bar{d}_1 = 60/5 = 12$ $\bar{d}_2 = 6/5 = 1,2$

$\bar{d}_1 > \bar{d}_2$ nên $\bar{x}_1$ đại diện $< \bar{x}_2$

- Ưu điểm: Thể hiện biến thiên của lượng biến chặt chẽ, đầy đủ hơn vì nó xét tới sự chênh lệch của tất cả các lượng biến so với số bình quân.

- Nhược điểm: Bỏ qua sự khác nhau thực tế về dấu.

c) Phương sai (δ^2):

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

Công thức tính như sau:

$$\delta^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} \quad \text{hay} \quad \delta^2 = \frac{\Sigma(x_i - \bar{x})^2 f_i}{\Sigma f_i}$$

Trong đó:

x_i là lượng biến thứ i

$\bar{x}$ là số bình quân

n (Σf_i) là số đơn vị tổng thể

Ý nghĩa: Phương sai càng bé thì mức độ biến động tiêu thức ít, tính chất đại biểu số bình quân càng cao và ngược lại.

Phương sai được dùng nhiều nhất trong thực tế vì nó giải quyết được vấn đề về dấu của các độ lệch tuyệt đối.

Thí dụ trên:

Ta lập bảng tính toán như sau:

Diễn giải	Nhóm 1			Nhóm 2		
	Xi	Xi – Xb/q	$(Xi - Xb/q)^2$	Xi	Xi – xb/q	$(Xi - Xb/q)^2$
1	20	-20	400	38	-2	4
2	30	-10	100	39	-1	1
3	40	0	0	40	0	0
4	50	10	100	41	1	1
5	60	20	400	42	2	4
Cộng	**40**		1000	**40**		10
Phương sai	200			2		

$\delta^2_1 = 200$; $\delta^2_2 = 2$. Như vậy $\delta^2_1 > \delta^2_2$ chứng tỏ $\bar{X}_1$ bình quân đại diện thấp hơn X_2 bình quân.

d) Độ lệch chuẩn (δ):

Là căn bậc 2 của phương sai, công thức tính như sau:

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

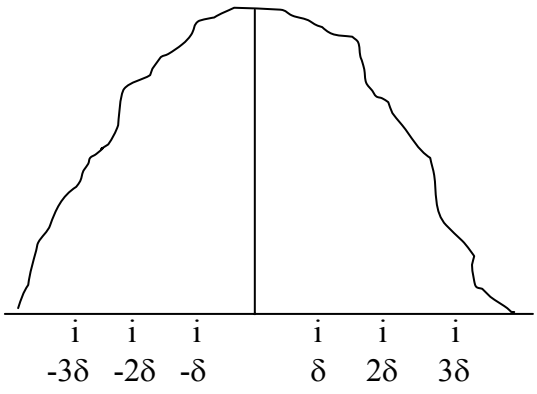
Thí dụ $\delta_1 = 14{,}142$; $\delta_2 = 1{,}4142$; $\delta_1 > \delta_2$; chứng tỏ $\bar{X}_1$ bình quân đại diện thấp hơn X_2 bình quân.

Ý nghĩa của độ lệch chuẩn: Dựa vào độ lệch chuẩn chúng ta biết được độ phân tán của tổng thể. Ngoài ra, nó còn được sử dụng để nhận biết sự phân phối của các lượng biến trong một tổng thể dựa trên quy tắc 3δ (quy tắc thực nghiệm) sau:

Trong một tổng thể, lượng biến của các đơn vị tổng thể có phân phối chuẩn thì:

- Có khoảng 68% giá trị rơi vào khoảng ± δ so với số trung bình;
- Có khoảng 95% giá trị rơi vào khoảng ± 2δ so với số trung bình;
- Có khoảng 99,73% giá trị rơi vào khoảng ± 3δ so với số trung bình;

Điều này được minh hoạ qua đồ thị sau:



Thí dụ: Tiền lương bình quân 1 người trong một doanh nghiệp là 800 ngàn đồng, độ lệch chuẩn về tiền lương là 50 ngàn đồng. Theo quy tắc này ước tính sẽ có:

- 68% số người có mức lương rơi vào khoảng 800 ± 50 (ngàn đồng), tức là từ 750 đến 850 ngàn đồng.
- 95% số người có mức lương rơi vào khoảng 800 ± (2*50) (ngàn đồng), tức là từ 700 đến 900 ngàn đồng.

Hình 1.4. Phân phối các lượng biến trong phân phối chuẩn

e) Hệ số biến động tiêu thức (V- hệ số biến thiên):

Hệ số biến thiên là tỷ số so sánh giữa độ lệch tiêu chuẩn (hoặc độ lệch tuyệt đối bình quân) với số bình quân cộng của các lượng biến.

Công thức:

$$V = \frac{\bar{d}}{\bar{x}} \times 100 \quad \text{hay} \quad V = \frac{\delta}{\bar{x}} \times 100$$

Hệ số biến thiên càng cao, thì độ phân tán của lượng biến càng lớn, tính chất đại diện của số bình quân càng thấp và ngược lại.

Thí dụ trên:

1) Tính theo độ lệch tuyệt đối bình quân $V_1 = 12/40*100 = 30\%$

$V_2 = 1,2/40*100 = 3\%$

2) Tính theo độ lệch chuẩn $V_1 = 14,14/40*100 = 35,35\%$

$V_2 = 1,41/40*100 = 3,52\%$

Chú ý:

- Hệ số biến động của tiêu thức là số tương đối, được dùng để so sánh độ phân tán giữa các hiện tượng có đơn vị tính khác nhau, hoặc giữa các hiện tượng cùng loại nhưng có số trung bình không bằng nhau.

- Trong thực tế, thống kê thực nghiệm đã cho rằng nếu V > 40% tính chất đại biểu của số bình quân thấp.

CÂU HỎI THẢO LUẬN CHƯƠNG IV

1. Các chỉ tiêu phân tích mức độ của hiện tượng kinh tế xã hội? Ý nghĩa, đặc điểm, cách tính và trường hợp vận dụng?
2. Hãy lấy một ví dụ trong thực tiễn về việc sử dụng các chỉ tiêu phân tích mức độ của hiện tượng?

Chương V
ĐIỀU TRA CHỌN MẪU

Như đã trình bày ở chương II để thu thập tài liệu ban đầu, thống kê sử dụng hai hình thức: báo cáo thống kê định kỳ và điều tra chuyên môn. Chế độ báo cáo thống kê định kỳ áp dụng chủ yếu đối với thành phần kinh tế quốc doanh, như các doanh nghiệp Nhà nước. Điều tra chuyên môn áp dụng để thu thập thông tin đối với những hiện tượng và quá trình kinh tế xã hội không thể hoặc không nhất thiết phải thực hiện báo cáo thống kê định kỳ. Điều tra chuyên môn có thể tiến hành trên toàn bộ các đơn vị tổng thể (điều tra toàn bộ) hoặc chỉ tiến hành trên một số đơn vị tổng thể (điều tra không toàn bộ, trong đó điều tra chọn mẫu được áp dụng phổ biến nhất).

1. KHÁI NIỆM VÀ Ý NGHĨA CỦA ĐIỀU TRA CHỌN MẪU

1.1. Khái niệm

** Điều tra chọn mẫu:*

Điều tra chọn mẫu là loại điều tra không toàn bộ. Từ tổng thể hiện tượng cần nghiên cứu người ta chọn ra một số đơn vị mang tính chất đại biểu cho tổng thể để điều tra. Kết quả điều tra được dùng suy rộng cho tổng thể. Các đơn vị được điều tra phải được chọn theo các phương pháp khoa học để đảm bảo tính chất đại biểu cho tổng thể.

Thí dụ: Điều tra tỷ lệ phế phẩm của một hãng sản xuất mì tôm. Người ta thường chọn ra một số gói mì nhất định, xác định tỷ lệ phế phẩm của số gói được chọn (giả sử tỷ lệ phế phẩm của mẫu đã chọn là 2%). Sử dụng kết quả này tính toán và suy rộng thành tỷ lệ phế phẩm của toàn bộ khối lượng mì mà hàng đã sản xuất.

Trong điều tra chọn mẫu, người ta đặc biệt lưu ý tới hai vấn đề cơ bản là:

- Lựa chọn các đơn vị mẫu sao cho đại diện cho toàn bộ tổng thể;

- Sử dụng công thức nào để tính toán và suy rộng cho toàn bộ tổng thể.

* *Tổng thể mẫu*: Là tổng số các đơn vị được chọn ra mang tính chất đại biểu cho tổng thể chung để điều tra.

Kí hiệu: Tổng thể mẫu n, tổng thể chung N

* *Đơn vị mẫu*: Là đơn vị đại biểu cho tổng thể được chọn ra để điều tra.

* *Bình quân mẫu*: Là lượng biến bình quân của các đơn vị mẫu.

Kí hiệu: Bình quân mẫu $\bar{x}$, bình quân chung $\bar{X}$

Số bình quân mẫu cũng được tính theo các công thức của số trung bình cộng trong tổng thể chung.

* *Tỷ lệ mẫu*: Là tỷ lệ của bộ phận có biểu hiện giống nhau về tiêu thức cần nghiên cứu trong tổng thể mẫu.

+ Tiêu thức cần nghiên cứu ở đây chỉ có 2 hình thức biểu hiện đối lập nhau (thường gọi là tiêu thức thay phiên).

Ví dụ: Phẩm chất của sản phẩm đồ hộp: sản phẩm đúng quy cách, sản phẩm không đúng quy cách.

Mục đích nghiên cứu là: Tính ra tỷ lệ sản phẩm không đúng quy cách.

Kí hiệu: Tỷ lệ mẫu p, tỷ lệ chung P.

Công thức tính tỷ lệ mẫu: $P = \frac{m}{n}$

Trong đó: m là số đơn vị mẫu có cùng biểu hiện

n: là số đơn vị mẫu.

1.2. Ý nghĩa

Điều tra chọn mẫu là phương pháp điều tra không toàn bộ khoa học nhất, nhằm thu thập các tài liệu ban đầu cần thiết mà báo cáo thống kê định kỳ không thực hiện hay không theo dõi được.

Cơ sở khoa học của điều tra chọn mẫu là lý thuyết xác suất và thống kê toán. Do đó, bằng điều tra chọn mẫu ta có thể biết được các tham số của tổng thể theo một đặc trưng nào đó với một mức độ chính xác, hay mức độ tin cậy tính toán được. Do đó, phương pháp điều tra chọn mẫu hoàn toàn có thể thay thế điều tra toàn bộ trong một số trường hợp. Ngoài ra điều tra chọn mẫu còn kết hợp với điều tra toàn bộ để mở rộng nội dung điều tra, cung cấp nhanh một số tài liệu để đảm bảo kịp thời trong việc chỉ đạo sản xuất.

1.3. Ưu điểm và hạn chế

So với điều tra toàn bộ, điều tra chọn mẫu có các ưu điểm sau:

- Về chi phí: Điều tra chọn mẫu tiết kiệm chi phí hơn.

- Về thời gian: Tiến độ công việc tiến hành nhanh hơn, có thể đáp ứng yêu cầu khẩn cấp của lãnh đạo.

- Về tính chính xác: Với các phương pháp suy rộng khoa học, các kết luận của điều tra chọn mẫu đảm bảo đáng tin cậy.

Tuy nhiên, điều tra chọn mẫu cũng có những hạn chế sau:

- Kết quả suy rộng từ điều tra chọn mẫu cho tổng thể bao giờ cũng có sai số nhất định. Những sai số này có thể trong điều tra toàn bộ không có.

- Đối với nguồn thống kê quan trọng cần nghiên cứu cả tổng thể và từng bộ phận của tổng thể thì điều tra chọn mẫu không thể thay thế được như tổng điều tra dân số; tổng kiểm kê...

Chính vì những hạn chế này mà điều tra toàn bộ thường áp dụng cho những trường hợp sau:

- Đối với những hiện tượng không thể tiến hành điều tra toàn bộ được. Thí dụ điều tra chất lượng sản phẩm, chất lượng công trình...

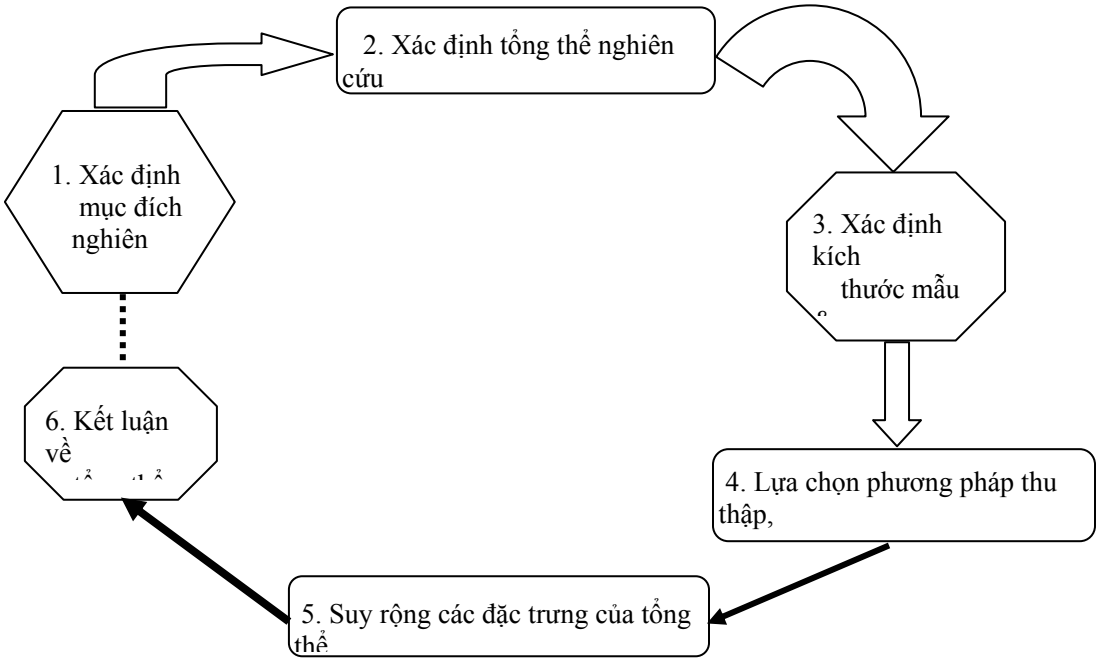
- Phúc tra các kết quả của điều tra toàn bộ;

- Đối với những hiện tượng vừa áp dụng điều tra toàn bộ, vừa áp dụng điều tra không toàn bộ. Đối với những hiện tượng này, người ta thường áp dụng điều tra chọn mẫu với những ưu điểm của nó để kiểm tra chất lượng của điều tra toàn bộ.

2. TRÌNH TỰ TIẾN HÀNH VÀ NỘI DUNG ĐIỀU TRA CHỌN MẪU

2.1. Trình tự tiến hành

Khi tiến hành điều tra chọn mẫu, người ta thường tiến hành theo các bước như sau:



Sơ đồ 5.1. Các bước trong điều tra chọn mẫu

Bước 1: Xác định mục đích điều tra

Do nhu cầu thực tế ta cần thông tin về một hiện tượng nào đó mà không có sẵn và không thể thu thập bằng điều tra toàn bộ được thì ta chọn điều tra chọn mẫu. Xác định mục đích điều tra là nhằm thu thập thông tin gì, phục vụ cho mục đích nghiên cứu nào. Việc xác định rõ mục đích điều tra có ý nghĩa quan trong trong việc lựa chọn số lượng và phương pháp lấy mẫu.

Bước 2: Xác định tổng thể có liên quan

Mẫu được chọn ra phải mang tính chất đại diện cho tổng thể, do đó cần xác định tổng thể nào có chứa mẫu. Xác định tổng thể có liên quan nghĩa là xác định phạm vi, tính chất của tổng thể phù hợp với mục đích nghiên cứu.

Bước 3: Xác định kích thước mẫu và phương pháp chọn mẫu

Số lượng mẫu cần chọn là bao nhiêu? Phương pháp chọn mẫu như thế nào là bước rất quan trọng có liên quan đến kết quả suy rộng cho tổng thể. Nội dung cụ thể của bước này được trình bày chi tiết ở mục sau.

Bước 4: Phương pháp thu thập và tính toán thông tin

Sau khi đã chọn được mẫu đại diện, công việc tiếp theo là thu thập các thông tin của từng đơn vị mẫu. Phương pháp thu thập thông tin của các đơn vị mẫu thường áp dụng như các phương pháp thu thập thông tin đã được trình bày ở chương II (số trung bình mẫu, tỷ lệ mẫu).

Cách xử lý, trình bày và tính toán các đặc trưng của mẫu giống như các phương pháp đã trình bày ở các chương III và IV.

Bước 5: Suy rộng các đặc trưng của tổng thể

Từ các đặc trưng của mẫu như số trung bình mẫu, tỷ lệ mẫu, sử dụng các phương pháp thống kê để suy rộng thành các đặc trưng của tổng thể.

Bước 6: Rút ra kết luận về tổng thể

Nội dung của bước này là xem xét các kết luận rút ra từ kết quả suy rộng trên cơ sở các đặc trưng của mẫu có đáp ứng yêu cầu đặt ra trong mục tiêu nghiên cứu hay không? Nhận xét này cũng cần đối chiếu với nội dung bước 1 xem có phù hợp không?

2.2. Những nội dung cơ bản

Lý thuyết điều tra chọn mẫu là vấn đề khá phức tạp trong lí thuyết thống kê. Nó liên quan nhiều đến lí thuyết xác suất và thống kê toán. Ở đây chỉ trình bày một số nội dung cơ bản của phương pháp này và sử dụng các công thức tính toán mà thống kê toán đã chứng minh.

a) Các cách chọn mẫu:

Việc chọn các đơn vị mẫu điều tra đảm bảo tính khách quan trong điều tra chọn mẫu được tiến hành theo các cách chọn: ngẫu nhiên (hay tuỳ cơ), máy móc, điển hình và cả khối.

* *Chọn ngẫu nhiên (tuỳ cơ):* Là phương pháp chọn mẫu hoàn toàn ngẫu nhiên, trong đó các đơn vị mẫu được chọn bằng cách bốc thăm, quay số hoặc theo bảng số ngẫu nhiên và có thể chọn một lần (không lặp), chọn nhiều lần (chọn có lặp).

+ Chọn 1 lần là sau khi rút ra 1 thăm người ta không bỏ lại vào tổng thể để chọn lần sau. Như vậy, mỗi đơn vị tổng thể chỉ có thể được chọn ra 1 lần và tổng thể mẫu gồm các đơn vị hoàn toàn khác nhau, sẽ đại biểu cho tổng thể cao hơn.

+ Chọn nhiều lần là cách chọn sau khi rút ra 1 thăm người ta ghi lại đơn vị được chọn rồi trả lại cái thăm vào tổng thể cũ. Như vậy, lần sau chọn vẫn có khả năng chọn đúng vào cái thăm đã chọn lần trước. Trong trường hợp này tổng thể mẫu có thể có một số đơn vị được chọn lại nhiều lần và mức độ đại biểu cho tổng thể chung sẽ không cao.

Trong điều tra chọn mẫu ngẫu nhiên người ta thường chọn cách chọn 1 lần.

Phương pháp chọn ngẫu nhiên đơn giản có thể cho kết quả tốt nếu giữa các đơn vị của tổng thể không có khác biệt nhiều. Ngược lại nếu tổng thể các đơn vị khác biệt nhau nhiều quá thì cách chọn này khó đảm bảo tính đại biểu. Hơn nữa, nếu tổng thể quá lớn thì không thể đánh số thăm hay đánh số cho tất cả các đơn vị tổng thể được.

* *Chọn máy móc*: Là phương pháp chọn mẫu hoàn toàn máy móc, nghĩa là cứ sau một khoảng cách nhất định người ta chọn ra một đơn vị mẫu.

Cách chọn này thường được tiến hành như sau:

- Trước hết sắp xếp các đơn vị tổng thể theo trình tự nào đó (thí dụ: tăng dần hoặc giảm dần của lượng biến theo tiêu thức cần nghiên cứu; hoặc theo vần A, B, C...).

- Căn cứ vào trật tự sắp xếp này, sau một khoảng cách nhất định lại chọn ra 1 đơn vị mẫu. Khoảng cách để chọn ra đơn vị mẫu được tính là k = N/n. (N là số đơn vị tổng thể, n là số đơn vị mẫu).

Chú ý: Thông thường đơn vị đầu tiên được chọn là đơn vị có số thứ tự nằm giữa khoảng cách chọn thứ nhất, hoặc nằm chính giữa trật tự sắp xếp nói trên. Đơn vị tiếp theo được chọn bằng cách cộng thêm 1 khoảng cách chọn vào thứ tự của đơn vị chọn trước. Như vậy số đơn vị mẫu đã được phân bố đều theo mức độ biến động của tiêu thức chủ yếu. Vì vậy, tính chất đại biểu của mẫu chọn ra cao hơn so với cách chọn trên.

* *Chọn điển hình tỷ lệ (chọn phân tổ):* Là phương pháp chọn mẫu từ các tổ. Phương pháp này thường được tiến hành như sau:

+ Trước hết phân chia tổng thể thành các tổ căn cứ vào tiêu thức có liên quan chặt chẽ đến mục đích nghiên cứu;

+ Từ mỗi bộ phận hay mỗi tổ chọn ra một số đơn vị mẫu;

+ Số đơn vị mẫu chọn ở mỗi tổ thường tỷ lệ với số đơn vị thuộc mỗi tổ so với tổng thể.

Theo cách chọn này số đơn vị mẫu của từng tổ đã có tính chất đại biểu cao cho từng tổ và tổng thể mẫu, cũng có tính chất đại biểu cao cho tổng thể chung.

Cách chọn này khoa học hơn 2 cách trên nên nó được áp dụng rộng rãi hơn, nhất là đối với hiện tượng cần điều tra có số đơn vị tổng thể lớn không thể chọn theo phương pháp chọn máy móc được. Song, cách chọn này đòi hỏi phải có sẵn các nguồn thông tin về tổng thể và có kiến thức phân tổ.

Phương pháp này phần nào cũng dựa vào những kinh nghiệm phán đoán chủ quan, nên cần phải tuân theo những nguyên tắc chung khi tiến hành phân tổ như:

- Trong mỗi tổ phải đảm bảo tính đồng chất;

- Số tổ không được chia quá ít hoặc quá nhiều;

- Số đơn vị mẫu của từng tổ phải đủ lớn để đảm bảo độ tin cậy cho suy rộng, hay ước lượng.

* *Chọn cả khối*: Là phương pháp tổ chức chọn mẫu, trong đó số đơn vị mẫu được chọn không phải là lẻ tẻ mà cùng một lúc chọn ra một khối đơn vị.

Theo cách chọn này, trước hết tổng thể chung được chia thành các khối, sau đó chọn ngẫu nhiên một số khối để điều tra. Cách chọn này thường áp dụng trong điều tra chất lượng sản phẩm mà khi sản xuất xong, sản phẩm đã được đóng kiện. Mức độ đại biểu thường không cao bằng các cách chọn trên.

b) Sai số bình quân chọn mẫu và phạm vi sai số chọn mẫu:

** Khái niệm về sai số chọn mẫu*

Do cuộc điều tra chọn mẫu chỉ tiến hành ở một số đơn vị tổng thể mà kết quả lại suy rộng ra cho cả tổng thể nên tất yếu nảy sinh sai số (gọi là sai số chọn mẫu).

Vậy sai số chọn mẫu là sự chênh lệch giữa các chỉ tiêu tính được trong điều tra chọn mẫu với các chỉ tiêu tương ứng của tổng thể.

Sai số chọn mẫu phụ thuộc vào các yếu tố sau:

- Số đơn vị mẫu được chọn ra để điều tra.

Nếu mở rộng phạm vi điều tra bằng cách tăng số đơn vị mẫu lên cho tới khi nó bằng số đơn vị tổng thể thì không còn sai số chọn mẫu. Như vậy, sai số chọn mẫu tỷ lệ nghịch với số đơn vị mẫu được chọn để điều tra. Trong thực tế thì số đơn vị mẫu không bao giờ bằng số đơn vị tổng thể.

- Mức độ đồng đều về lượng biến của tiêu thức nghiên cứu ở các đơn vị tổng thể.

Nếu lượng biến của tiêu thức nghiên cứu ở các đơn vị tổng thể xấp xỉ bằng nhau thì khi chọn các đơn vị mẫu để điều tra sẽ tính được lượng biến bình quân của các đơn vị mẫu cũng sẽ xấp xỉ với lượng biến bình quân chung, khi đó sai số chọn mẫu sẽ nhỏ và ngược lại.

Để đo độ đồng đều đó ở chương IV, chúng ta đã nghiên cứu một số các chỉ tiêu (toàn cự, độ lệch tuyệt đối bình quân, phương sai, độ lệch chuẩn và hệ số biến động tiêu thức: R, $\overline{d}$, δ^2, δ, V).

Trong các chỉ tiêu đó, thống kê toán dùng nhiều nhất là phương sai hay độ lệch bình phương bình quân. Chỉ tiêu này được tính theo công thức sau:

Tài liệu không phân tổ	Tài liệu có phân tổ	Dùng tính cho tỷ lệ
$\sigma_x^2 = \frac{\sum(x_i - \bar{x})^2}{n}$	$\sigma_x^2 = \frac{\sum(x_i - \bar{x})^2 f_i}{\sum f_i}$	$\sigma_p^2 = p.q = p.(1-p)$
x_i: Lượng biến của từng đơn vị tổng thể $\bar{x}$: Lượng biến bình quân n: Số đơn vị tổng thể	x_i: Lượng biến từng tổ $\bar{x}$: Lượng biến bình quân f_i: Số đơn vị tổng thể của tổ	P: Tỷ lệ của bộ phận có biểu hiện về tiêu thức cần nghiên cứu q: Tỷ lệ của bộ phận đối lập

- Phương pháp chọn các đơn vị mẫu (phần trên đã trình bày). Các phương pháp chọn mẫu khác nhau, tính đại diện của mẫu chọn ra cũng khác nhau nên có ảnh hưởng đến sai số chọn mẫu.

Sai số chọn mẫu không phải là một trị số cố định. Ngoài các yếu tố chủ quan nói trên , sai số chọn mẫu còn phụ thuộc vào kết cấu mẫu.

Cùng một hiện tượng nếu tiến hành điều tra nhiều lần với các cách chọn mẫu và tổng thể có kết cấu khác nhau sẽ có sai số chọn mẫu khác nhau.

Ví dụ: 1 tổng thể gồm 10 đơn vị ABCDMNPQRV.

Chọn mẫu 3 đơn vị để điều tra.

C1: ABC ta tính được sai số chọn mẫu thứ nhất (s_1);

C2: ABD ta tính được sai số chọn mẫu thứ nhất (s_2);

C1: MNP ta tính được sai số chọn mẫu thứ nhất (s_3);

...

Do đó, muốn tính sai số để đánh giá mức độ chính xác của ước lượng thì phải tính sai số bình quân chọn mẫu.

* *Sai số bình quân chọn mẫu*: Bình quân tất cả các sai số chọn mẫu do việc lựa chọn mẫu có kết cấu thay đổi (còn gọi sai lệch mẫu điển hình).

Thống kê toán đã xác định được công thức tính sai số bình quân chọn mẫu như sau:

Phương pháp chọn	Dùng suy rộng cho số bình quân	Dùng suy rộng cho tỷ lệ
Chọn nhiều lần	$\mu_x = \sqrt{\frac{\sigma^2}{n}}$	$\mu_p = \sqrt{\frac{p(1-p)}{n}}$
Chọn một lần	$\mu_x = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$	$\mu_p = \sqrt{\frac{p(1-p)}{n}\left(1 - \frac{n}{N}\right)}$

Ký kiệu : μ là sai số bình quân chọn mẫu n là số đơn vị mẫu

δ^2 là phương sai N là số đơn vị tổng thể

P là tỷ lệ của tổng thể

Một số lưu ý:

- Giữa chọn một lần và chọn nhiều lần công thức tính sai số bình quân chọn mẫu sai khác nhau một đại lượng (1-n/N). Nếu tổng thể khá lớn thì n/N là quá nhỏ và (1-n/N)→1. Cho nên sự chênh lệch giữa hai công thức này không nhiều, thường khi chọn một lần sai số bình quân chọn mẫu là nhỏ hơn khi chọn nhiều lần. Trong thực tế, người ta thường sử dụng cách chọn một lần để điều tra. Nhưng khi tính sai số để giảm bớt phức tạp trong tính toán, người ta thường dùng công thức chọn nhiều lần.

- Theo lý thuyết σ^2_x và P phải tính từ tổng thể nhưng thực tế σ^2_x hoặc P chưa xác định được. Để giải quyết khó khăn này có thể sử dụng các phương pháp sau đây:

+ Có thể lấy σ^2_x hoặc p của nhiều lần điều tra trước về hiện tượng đó. Nếu trước đó có nhiều lần điều tra thì lấy σ^2_x lớn nhất hoặc p gần 0.5 nhất (nó liên quan đến chọn số đơn vị mẫu phần sau sẽ nhắc lại);

+ Có thể lấy σ^2_x hoặc P của cuộc điều tra tương tự nhưng tiến hành ở nơi khác;

+ Điều tra chọn mẫu thí điểm trong phạm vi hẹp để tính phương sai hoặc tỷ lệ của mẫu thí điểm thay cho phương sai hay P của tổng thể (cách này hiện nay hay làm).

Công thức tính:

$$\sigma_x^2 = \frac{n}{(n-1)}.\sigma_0^2$$

Trong đó: σ^2_x: Phương sai dùng điều tra.

σ^2_0: Phương sai mẫu làm thí điểm

Như trên chúng ta đã biết, sai số bình quân chọn mẫu này không phải là một trị số xác định, nếu ta tiến hành nhiều lần điều tra khác nhau sẽ nhận được các sai số khác nhau và đều dao động quanh μ .

Vì vậy, chúng ta không thể xác định chính xác sai số chọn mẫu cho mỗi lần điều tra mà chỉ có thể dựa vào sai số bình quân chọn mẫu để ước lượng phạm vi sai số. Do đó phạm vi này còn gọi là phạm vi sai số chọn mẫu.

* *Phạm vi sai số chọn mẫu (Δ)*: Là phạm vi chênh lệch giữa các chỉ tiêu của mẫu với các chỉ tiêu tương ứng của tổng thể ứng với độ tin cậy nhất định.

- Thống kê toán đã xác định được công thức tính toán: $\Delta = \pm\ t.\mu$

Trong đó: t: Độ cơ suất (hệ số tin cậy)

μ: Sai số bình quân chọn mẫu.

- Ứng với mỗi trị số của t có một độ tin cậy tương ứng $\Phi(t)$ (hàm xác suất). Quan hệ giữa hệ số tin cậy và độ tin cậy được thể hiện qua hàm tích phân xác suất do nhà toán học Liapunốp xây dựng nên. Với quan hệ này, chúng ta có thể điều chỉnh Δ ứng với độ tin cậy $\Phi(t)$ (hàm xác suất) của tài liệu điều tra.

Hệ số tin cậy (t)	Độ tin cậy Φ(t)
1,0	0,6827
1,5	0,8664
2,0	0,9545
2,5	0,9876
3,0	0,9973

Nếu kết quả điều tra tính được phạm vi sai số chọn mẫu theo công thức $\Delta = \pm\mu$ với độ tin cậy của việc suy rộng tài liệu là 0,6827. Điều này có nghĩa là trong 10000 lần điều tra chỉ có 6827 lần chắc chắn có sai số chọn mẫu không vượt quá $\pm\mu$ (hệ số tin cậy t = 1) còn 3173 lần chắc chắn có sai mẫu vượt quá $\pm\mu$.

Nếu muốn nâng trình độ tin cậy của việc suy rộng tài liệu lên thì hệ số tin cậy cũng phải được nâng lên. Chẳng hạn nếu độ tin cậy là 0,9545 thì hệ số tin cậy t = 2, $\Delta = \pm 2\mu$.

Từ các công thức tính sai số bình quân chọn mẫu, ta suy ra các công thức tính phạm vi sai số chọn mẫu cho các trường hợp cụ thể.

Ví dụ: Trong một doanh nghiệp gồm có 1600 công nhân, người ta tiến hành điều tra chọn mẫu về tình hình tiền lương. Số công nhân được chọn ra là 400 người theo phương pháp chọn ngẫu nhiên đơn thuần có trả lại. Kết quả điều tra cho thấy:

- Tiền lương trung bình của công nhân là 650.000 đồng.

- Độ lệch chuẩn là 80.000 đồng.

Hãy tính:

1, Sai số bình quân chọn mẫu và phạm vi sai số chọn mẫu về tiền lương bình quân với xác suất là 0,997.

2, Nếu cuộc điều tra được tiến hành theo phương pháp chọn ngẫu nhiên đơn thuần (không trả lại) thì sai số bình quân chọn mẫu và phạm vi sai số bình quân chọn mẫu sẽ là bao nhiêu?

Giải:

- Câu 1: $\mu_x = \sqrt{\frac{\delta_x^2}{n}} = 4;\ \Delta = t\mu_x = 12$

- Câu 2: $\mu_x = \sqrt{\frac{\delta_x^2}{n}\left(1 - \frac{n}{N}\right)} = 3{,}46;\ \Delta = t\mu_x = 10{,}39$

c) Số đơn vị mẫu cần chọn:

Như ta đã thấy sai số chọn mẫu tỷ lệ nghịch với đơn vị mẫu chọn để điều tra. Vì vậy, muốn giảm sai số chọn mẫu người ta cần tăng số đơn vị mẫu với khả năng tối đa.

Mặt khác, việc tăng số đơn vị mẫu lên lại liên quan tới những chi phí tốn kém mà kết quả điều tra phải chịu.

Do đó, để đáp ứng yêu cầu đảm bảo kết quả điều tra và giảm bớt tốn kém chi phí người ta chỉ cần xác định số đơn vị mẫu cần thiết theo các điều kiện đã cho để điều tra.

Công thức tính số đơn vị mẫu: Từ công thức tính phạm vi sai số chọn mẫu, ta suy ra công thức tính số đơn vị mẫu cần chọn.

$$\Delta_x = \pm t\sqrt{\frac{\sigma^2}{n}} \rightarrow \Delta_x^2 = \frac{t^2\sigma^2}{n} \rightarrow \frac{t^2\sigma^2}{\Delta_x^2} = n$$

Tương tự chúng ta tính được các công thức xác định số đơn vị mẫu cần thiết cho các trường hợp cụ thể.

Phương pháp chọn	Dùng cho số bình quân	Dùng cho tỷ lệ
Chọn nhiều lần	$n = \frac{t^2\sigma^2}{\Delta_x^2}$	$n = \frac{t^2 p(1-p)}{\Delta_p^2}$
Chọn một lần	$n = \frac{t^2\sigma^2 N}{N\Delta_x^2 + t^2\sigma_x^2}$	$n = \frac{t^2 p(1-p)N}{N\Delta_p^2 + t^2 p(1-p)}$

Thí dụ: Trong cuộc điều tra năng suất sản lượng lúa của một HTX, người ta yêu cầu xác định số đơn vị mẫu cần chọn (mỗi đơn vị mẫu có diện tích gặt là 4 m^2), sao cho phạm vi sai số chọn mẫu của điều tra không vượt quá 0,06 kg/4m^2. Yêu cầu độ tin cậy của việc suy rộng tài liệu là 0,9545, phương sai của lần điều tra trước 0,128.

Ta có: $\Phi(t) = 0{,}9545 \rightarrow t = 2$, $\Delta_x = 0{,}06$, $\delta_x^2 = 0{,}128 \rightarrow n = 142$ điểm.

d) Suy rộng tài liệu điều tra:

Kết quả điều tra các đơn vị mẫu tính được $\overleftrightarrow{x}$ và p. Sau khi chúng ta tính được phạm vi sai số chọn mẫu cần suy rộng tài liệu cho tổng thể theo 2 phương pháp sau:

** Phương pháp trực tiếp:*

$$\bar{x} = \overleftrightarrow{x} \pm \Delta_x \qquad P = p \pm \Delta p$$

Thí dụ điều tra năng suất của một HTX, ta tính được $\overleftrightarrow{x} = 32$ tạ/ha, $\Delta_x = \pm 1{,}5$ tạ/ha

$$\Rightarrow 30{,}5 \le \bar{x} \le 33{,}5$$

* *Phương pháp hệ số điều chỉnh*: Phương pháp này dùng để kiểm tra tính chính xác của kết quả điều tra toàn bộ. Thực hiện như sau:

+ Sau khi thực hiện được các cuộc điều tra toàn bộ như điều tra dân số, điều tra gia súc người ta chọn một số mẫu để kiểm tra.

+ Kết quả tính toán ở một số mẫu đó được đem so sánh với kết quả trong điều tra toàn bộ để tính ra hệ số sai số.

+ Dùng hệ số sai số để điều chỉnh kết quả chung của tổng thể.

Thí dụ:

Kết quả điều tra dân số 1/4/1999 của huyện A là 500.000 người, trong đó xã T là 80.800 người.

Người ta chọn xã T điều tra lại thì thấy dân số xã T là 80.816 người.

Số người tính thiếu là 16 người. Vậy hệ số tính thiếu là 16/80800 = 0,0002.

Điều chỉnh lại dân số của cả huyện A = 500000*(1 + 0,0002) = 500100 người.

3. ĐIỀU TRA CHỌN MẪU PHI NGẪU NHIÊN

3.1. Khái niệm, ý nghĩa

Bên cạnh điều tra chọn mẫu ngẫu nhiên trên đây, trong thực tế người ta thường sử dụng điều tra chọn mẫu phi ngẫu nhiên.

*Điều tra chọn mẫu phi ngẫu nhiê*n là phương pháp điều tra mà trong đó việc chọn các đơn vị mẫu đại biểu cho tổng thể để điều tra phụ thuộc nhiều vào sự nhận định chủ quan của người tổ chức điều tra.

Điều tra chọn mẫu phi ngẫu nhiên không hoàn toàn dựa trên cơ sở toán học như điều tra chọn mẫu ngẫu nhiên, mà đòi hỏi phải kết hợp chặt chẽ giữa phân tích lý luận với thực tiễn xã hội.

Điều tra chọn mẫu phi ngẫu nhiên được dùng đối với các hiện tượng mà khi chọn mẫu không thể chọn một cách ngẫu nhiên dựa trên cơ sở toán học được mà phải kết hợp với sự nhận định chủ quan của con người về nhiều đặc điểm để bổ sung thì mới xác định được các đơn vị mang tính đại biểu cao cho tổng thể.

Ví dụ: Điều tra năng suất sản lượng lúa của nước ta.

Thời kỳ 1974→1984: Chúng ta thường dùng phương pháp toán học để xác định số đơn vị mẫu. Song trong thực tế, Tổng cục Thống kê đã giao cho huyện xác định số điểm điều tra cho từng HTX.

Tuỳ theo tình hình biến động về năng suất của từng HTX mà quy định từ 2 đến 6 mẫu Bắc bộ chọn 1 điểm đại diện.

Vậy việc xác định số điểm điều tra như vậy hoàn toàn phụ thuộc vào sự nhận định đánh giá chủ quan của cán bộ huyện.

3.2. Các vấn đề chủ yếu trong điều tra chọn mẫu phi ngẫu nhiên

Trong điều tra chọn mẫu phi ngẫu nhiên, muốn cho chất lượng tài liệu điều tra tốt cần chú ý các vấn đề sau:

- Phân tổ chính xác đối tượng điều tra; bởi vì phân tổ tổng thể giúp chúng ta chọn các đơn vị mẫu có khả năng đại diện cho tổng thể;

- Chọn đơn vị điều tra: Vì số đơn vị mẫu chọn ra dựa vào kinh nghiệm của các chuyên gia hoặc qua bàn bạc phân tích tập thể, nên thông thường nên chọn những đơn vị nào có mức độ phổ biến nhất trong từng nhóm, hay bộ phận, hoặc gần với số trung bình của bộ phận đó.

- Sai số chọn mẫu: Sai số chọn mẫu trong điều tra chọn mẫu phi ngẫu nhiên không thể dựa vào công thức toán học để tính toán mà phải thông qua nhận xét, so sánh để ước lượng. Khi suy rộng kết quả điều tra chọn mẫu phi ngẫu nhiên, người ta sử dụng trực tiếp chứ ít khi suy rộng cho phạm vi toàn bộ tổng thể.

- Huấn luyện cán bộ tham gia điều tra: Trong điều tra chọn mẫu phi ngẫu nhiên, ý kiến chủ quan của con người rất quan trọng. Do đó, người cán bộ điều tra muốn làm tốt công tác điều tra không những có nghiệp vụ tốt mà còn cần phải trung thực, có khả năng vận động quần chúng. Cán bộ điều tra cần được tập huấn và quán triệt ý nghĩa, mục đích, nội dung, phương pháp và kỹ năng để điều tra.

Tóm lại: Điều tra chọn mẫu ngẫu nhiên và phi ngẫu nhiên đều là các phương pháp điều tra chọn mẫu có hiệu quả. Mỗi phương pháp có những mặt ưu và nhược điểm nhất định và thích hợp với từng hiện tượng nghiên cứu. Hai phương pháp này thường hỗ trợ nhau nên trong thực tế, người ta thường kết hợp khéo léo cả hai phương pháp này.

Chương VI
KIỂM ĐỊNH THỐNG KÊ

1. KIỂM ĐỊNH GIẢ THUYẾT

1.1. Khái niệm và các loại giả thuyết

a) Khái niệm:

Trong điều tra chọn mẫu, chúng ta đã xác định được các đặc trưng của mẫu (số bình quân, tỷ lệ). Các đặc trưng này được dùng để ước lượng các đặc trưng của tổng thể. Ngoài ra còn được dùng để kiểm định giả thuyết nào đó của tổng thể.

Thí dụ:

1. Một hãng sản xuất mì tôm cho rằng khối lượng 1 gói mì tôm là 75 g. Để kiểm tra điều này đúng hay sai chúng ta lấy mẫu một số gói mì, cân và tính toán một tiêu chuẩn kiểm định.

2. Một nhà quản lý giáo dục cho rằng cách chấm điểm của các trường đại học là không khác nhau. Để kiểm tra điều này đúng hay sai chúng ta lấy mẫu chấm điểm một số trường sau đó tính toán tiêu chuẩn kiểm định.

Như vậy, việc tìm ra kết luận để bác bỏ hay chấp nhận một giả thuyết nào đó gọi là kiểm định giả thuyết.

b) Các loại giả thuyết:

+ Giả thuyết Ho

Giả sử tổng thể chung có một đặc trưng **a** chưa biết (thí dụ: Số trung bình, tỷ lệ, phương sai). Với giá trị cụ thể $\mathbf{a_o}$ cho trước nào đó, ta cần kiểm định giả thuyết:

Ho: $\mathbf{a = a_o}$ (kiểm định hai phía)

Ho: $\mathbf{a \geq a_o}$ hoặc $\mathbf{a \leq a_o}$ (kiểm định 1 phía).

+ Giả thuyết H_1

Giả thuyết H_1 là kết quả ngược lại của giả thuyết Ho, nghĩa là nếu giả thuyết Ho đúng thì giả thuyết H_1 sai và ngược lại. Vì vậy giả thuyết H_1 được gọi là đối thuyết.

+ Các giả thuyết này thường được thể hiện thành cặp trong kiểm định như sau:

- Kiểm định hai phía Ho : $\mathbf{a = a_o}$; H_1 : $\mathbf{a \neq a_o}$
- Kiểm định 1 phía Ho : $\mathbf{a \geq a_o}$; H_1 : $\mathbf{a < a_o}$

Hoặc Ho : $\mathbf{a \leq a_o}$; H_1 : $\mathbf{a > a_o}$

Thí dụ: Lấy lại thí dụ 1 trên đây, các giả thuyết được viết như sau:

Kiểm định hai phía Ho : $\mathbf{a = 75g}$; H_1 : $\mathbf{a \neq 75g}$

c) Các loại sai lầm trong kiểm định giả thuyết:

Trong kiểm định giả thuyết, do chỉ dựa trên kết quả điều tra mẫu để đưa ra kết luận bác bỏ hay chấp nhận một giả thuyết nào về các đặc trưng của tổng thể, nên thường phạm các sai lầm. Các sai lầm đó là:

- Giả thuyết Ho đúng (tức là $\mathbf{a = a_o}$), nhưng kết quả kiểm định lại kết luận giả thuyết sai (Tức là $\mathbf{a \neq a_o}$), nên ta bác bỏ Ho. Trường hợp này người ta qui ước gọi là **sai lầm loại 1.**

Vậy, sai lầm loại 1 là bác bỏ giả thuyết Ho khi giả thuyết này đúng.

- Giả thuyết Ho sai (tức là $\mathbf{a \neq a_o}$),nhưng kết quả kiểm định lại kết luận giả thuyết đúng (tức là $\mathbf{a = a_o}$), nên ta chấp nhận Ho. Trường hợp này người ta qui ước gọi là **sai lầm loại 2.**

Vậy, sai lầm loại 2 là chấp nhận giả thuyết Ho khi giả thuyết này sai.

Tóm lại: Khi ta bác bỏ một giả thuyết là ta có thể mắc phải sai lầm loại I, còn khi ta chấp nhận một giả thuyết là ta có thể phạm phải sai lầm loại II.

Thực chất sai lầm loại I và sai lầm loại II chỉ mang tính chất tương đối. Nó được xác định khi ta đặt giả thuyết Ho. **Thông thường sai lầm nào gây ra tổn thất lớn hơn người ta sẽ đặt giả thuyết Ho sao cho sai lầm đó là loại 1 và định trước khả năng mắc phải sai lầm loại 1 không vượt qua một số α nào đó ($\alpha = 5\%$), tức là thực hiện kiểm định giả thuyết Ho ở mức ý nghĩa α cho trước**. Có thể xảy ra các trường hợp sau:

- Nếu α càng bé thì khả năng phạm sai lầm loại I càng ít, khi đó xác suất mắc sai lầm loại II sẽ tăng lên. Thí dụ, nếu lấy $\alpha = 0$ thì sẽ không bác bỏ bất kỳ giả thuyết nào, có nghĩa không mắc sai lầm loại I, khi đó xác suất mắc sai lầm loại II sẽ đạt cực đại **($1- \alpha = 1$).**

- Với sai lầm loại I: Nếu quyết định xác suất bác bỏ giả thuyết Ho khi giả thuyết này đúng là α thì xác xuất để chấp nhận nó là **($1- \alpha$).** Người ta gọi α là mức ý nghĩa của kiểm định.

- Với sai lầm loại II: Nếu quyết định xác suất chấp nhận giả thuyết Ho khi giả thuyết này sai là β thì xác xuất để bác bỏ nó là **($1- \beta$).** Người ta gọi β là mức ý nghĩa của kiểm định.

Có thể tóm tắt những quyết định xác suất dựa trên giả thuyết Ho như sau:***Bảng 1.6.***

	Giả thuyết Ho đúng	**Giả thuyết Ho sai**
1. Chấp nhận giả thuyết Ho	Xác suất quyết định đúng: **($1 - \alpha$)**	Xác suất sai lầm loại II : β
2. Bác bỏ giả thuyết Ho	Xác suất sai lầm loại I : α	Xác suất quyết định đúng: **($1 - \beta$)**

Thí dụ: Lấy lại thí dụ 2 trên đây:

Một nhà quản lý giáo dục cho rằng cách chấm điểm của các trường đại học là không khác nhau. Để kiểm tra điều này đúng hay sai chúng ta lấy mẫu chấm điểm một số trường sau đó tính toán tiêu chuẩn kiểm định.

- Trước hết chúng ta chọn giả thuyết Ho: Cách chấm điểm không khác nhau

H_1: Cách chấm điểm khác nhau

- Để thực hiện việc kiểm định giả thuyết, các trường hợp sau đây có thể xảy ra:

Bảng 2.6.

Giả thuyết Ho	Thực tế	Bác bỏ giả thuyết Ho	Chấp nhận giả thuyết Ho
Cách chấm điểm có khác nhau	Cách chấm điểm có khác nhau	Mắc sai lầm loại 1 Xác suất = α	Kết luận đúng Xác suất = **1- β**
	Cách chấm điểm không khác nhau	Kết luận đúng Xác suất = **1- α**	Mắc sai lầm loại II Xác suất = β
Cách chấm điểm không khác nhau	Cách chấm điểm có khác nhau	Kết luận đúng Xác suất = **1- α**	Mắc sai lầm loại II Xác suất = β
	Cách chấm điểm không khác nhau	Mắc sai lầm loại 1 Xác suất = α	Kết luận đúng Xác suất = **1- β**

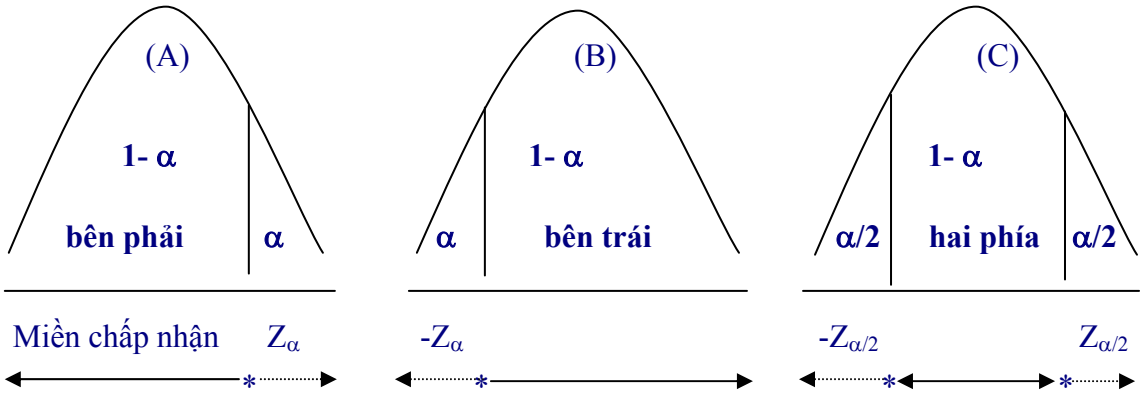
d) Miền bác bỏ và miền xác định trong kiểm định:

- Kiểm định hai phía Ho : $\mathbf{a = a_o}$; H_1 : $\mathbf{a \neq a_o}$; Miền bác bỏ nằm về hai phía của miền chấp nhận (hình C);

- Kiểm định 1 phía Ho : $\mathbf{a \geq a_o}$; H_1 : $\mathbf{a < a_o}$; Gọi là kiểm định bên trái; Miền bác bỏ nằm về phía bên trái của miền chấp nhận (hình B);

Hoặc Ho : $\mathbf{a \leq a_o}$; H_1 : $\mathbf{a > a_o}$; Gọi là kiểm định bên phải; Miền bác bỏ nằm về phía bên phải của miền chấp nhận (hình A).

Điều này được thể hiện qua hình 1.6 như sau:



Hình 1.6. Miền xác định, miền bác bỏ trong kiểm định giả thuyết

Miền xác định ——— Miền bác bỏ ·····

1.2. Các dạng kiểm định giả thuyết thường dùng

1.2.1. Kiểm định giả thuyết về số trung bình của tổng thể

a) Bài toán:

Giả sử một tổng thể có số trung bình là μ chưa biết. Ta cần kiểm định giả thuyết:

Ho: $\mu = \mu_o$ (μ_o cho trước);

H1: $\mu \neq \mu_o$

- Lấy mẫu gồm n quan sát độc lập, thu thập thông tin, tính toán $\overline{X}$. Thực hiện kiểm định giả thuyết Ho ở mức ý nghĩa α cho trước. Ta chia thành 2 trường hợp sau:

+ n ≥ 30 cho biết δ^2 (phương sai), ta tính giá trị kiểm định Z như sau:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\delta}{\sqrt{n}}}$$

Trong đó:

μ_o: Giá trị cụ thể cho trước

$\bar{X}$: Số trung bình của mẫu

δ : Độ lệch chuẩn

n : Số đơn vị mẫu quan sát

Z : Tiêu chuẩn kiểm định (thực nghiệm)

- Dựa vào mức ý nghĩa α cho trước ta tìm $Z_{\alpha/2}$ (**Z** lý thuyết - tra bảng).

- So sánh Z thực nghiệm với Z lý thuyết:

Nếu $|Z| > Z\alpha/2$ ta bác bỏ giả thuyết Ho

Nếu $|Z| \leq Z\alpha/2$ ta chấp nhận giả thuyết Ho

Nếu chưa biết **δ^2 (phương sai),** ta thay $\delta^2 = S^2$ (phương sai hiệu chỉnh của mẫu).

+ n < 30:

- Nếu X tuân theo phân phối chuẩn, biết δ^2 (phương sai), ta làm đúng như trường hợp n ≥ 30 biết δ^2 (phương sai).

- Nếu X tuân theo phân phối chuẩn, chưa biết δ^2 (phương sai), ta tính giá trị kiểm định T.

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

Trong đó:

μ_o: Giá trị cụ thể cho trước

$\bar{X}$: Số trung bình của mẫu

S : Độ lệch chuẩn của mẫu

n : Số đơn vị mẫu quan sát

T : Tiêu chuẩn kiểm định (T- thực nghiệm)

Dựa vào mức ý nghĩa α cho trước ta tìm $\mathbf{T}_{n-1,\alpha/2}$ (**T lý thuyết** - tra bảng phân phối **T- student,** hoặc dùng **hàm TINV (n-1; α/2)** trong EXCEL**.** So sánh T thực nghiệm với T lý thuyết:

Nếu $|T| > \mathbf{T}_{n-1,\alpha/2}$ ta bác bỏ giả thuyết Ho

Nếu $|T| \leq \mathbf{T}_{n-1,\alpha/2}$ ta chấp nhận giả thuyết Ho

Chú ý: Trong tất cả các trường hợp nói trên, nếu giả thuyết đã bị bác bỏ (nghĩa là $\mu \neq \mu_o$), khi đó:

- Nếu **X** (số bình quân của mẫu) $> \mu_o$ ta kết luận $\mu > \mu_o$
- Nếu **X** (số bình quân của mẫu) $< \mu_o$ ta kết luận $\mu < \mu_o$

Bằng cách làm tương tự chúng ta cũng thực hiện cho kiểm định một bên. Chúng ta có thể tóm tắt các trường hợp kiểm định giả thuyết số trung bình của tổng thể như sau:

Bảng 3.6.

N ≥ 30		**N<30**	
Giả thuyết	**Bác bỏ Ho khi**	**Giả thuyết**	**Bác bỏ Ho khi**
Ho: $\mu = \mu_o$ H1: $\mu \neq \mu_o$	$Z > Z\alpha/2$ hoặc $Z < -Z\alpha/2$ Hay $\|Z\| > Z\alpha/2$	Ho: $\mu = \mu_o$ H1: $\mu \neq \mu_o$	$T > \mathbf{T}_{n-1,\alpha/2}$ hoặc $T < -\mathbf{T}_{n-1,\alpha/2}$ Hay $\|T\| > \mathbf{T}_{n-1,\alpha/2}$
Ho: $\mu = \mu_o$ hoặc $\mu \geq \mu_o$ H1: $\mu < \mu_o$	$Z < -Z\alpha$	Ho: $\mu = \mu_o$ hoặc $\mu \geq \mu_o$ H1: $\mu < \mu_o$	$T < -\mathbf{T}_{n-1,\alpha}$
Ho: $\mu = \mu_o$ hoặc $\mu \leq \mu_o$ H1: $\mu > \mu_o$	$Z > Z\alpha$	Ho: $\mu = \mu_o$ hoặc $\mu \leq \mu_o$ H1: $\mu > \mu_o$	$T > \mathbf{T}_{n-1,\alpha/2}$

b) Thí dụ:

Thí dụ 1: Một máy đóng mì gói tự động quy định khối lượng trung bình 1 gói là 75g, độ lệch chuẩn là 15g. Sau một thời gian sử dụng, người ta tiến hành kiểm tra mẫu 80 gói và tính được khối lượng trung bình là 72g. Hãy đánh giá về mức độ chính xác của máy đóng gói này với mức ý nghĩa α = 5%.

Giải:

Gọi μ là khối lượng thực tế 1 gói mì ; μ_o là khối lượng quy định 1 gói mì.

Ta đặt giả thuyết Ho: $\mu = \mu_o$ Đối thuyết H1: $\mu \neq \mu_o$

Kiểm định giả thuyết Ho: n = 80; δ = 15g; α = 5%.

Tính Z thực nghiệm và tra bảng Z lý thuyết:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\delta}{\sqrt{n}}} = \frac{72 - 75}{\frac{15}{\sqrt{80}}} = 1{,}79$$

Z lý thuyết: Z(α/2) = Z(2,5%) = 1,96

Vì $|Z| < Z\alpha/2$; 1,79 < 1,96 nên ta chấp nhận Ho, tức là $\mu = \mu_o$ = **75g**. Như vậy với mức ý nghĩa α = 5% ta có kết luận là khối lượng trung bình 1 gói mì không sai khác với tiêu chuẩn quy định.

Giá trị P (P - value):

Nếu giả sử trong ví dụ trên ta kiểm định giả thuyết Ho: $\mu = \mu_o$ với mức ý nghĩa α = 10% thì ta có cùng kết luận như trên không?

Với α = 10% ta có Zα/2 = Z(5%) = 1,645 < |Z| thực nghiệm =1,79, ta bác bỏ Ho.

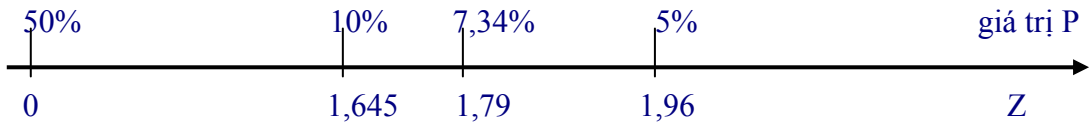
Vậy với mức ý nghĩa α nhỏ nhất nào thì ở đó giả thuyết Ho bị bác bỏ. Mức ý nghĩa nhỏ nhất đó gọi là giá trị P (P - value).

Lấy lại thí dụ trên ta thấy, với giá trị kiểm định thực nghiệm Ho bị bác bỏ |Z| thực nghiệm =1,79, thì giả thuyết Ho bị bác bỏ ở bất cứ giá trị nào của α mà ở đó Zα <1,79.

Tra bảng Z ta có kết quả: φ (1,79) = 0,4633; mà α/2 = 0,5 - 0,4633 = 0,0367

Vậy α = 2 x 0,0367 = 0,0734 hay 7,34%; Nghĩa là giả thuyết Ho sẽ bị bác bỏ ở bất kỳ mức ý nghĩa α nào lớn hơn 7,34%.

Có thể hình dung miền chấp nhận, miền bác bỏ theo giá trị P ở sơ đồ sau:



Hình 2.6. Miền chấp nhận, miền bác bỏ theo giá trị P

Chú ý:

1) Trong thực tế tính giá trị P (**(P - value)** có thể sử dụng hàm **NORMSDIST** trong EXCEL hoặc các phần mềm thống kê.

- Nếu sử dụng hàm **NORMSDIST** trong EXCEL thì thực hiện như sau:

Ta có P - value = P(Z > 1,79) = P(Z <- 1,79)= 1- **NORMSDIST(1,79)= 0,0367269** (tra hàm = **NORMSDIST(1.79)** trong EXCEL).

Từ đó α = 2 x 0,0367 = 0,0734 hay 7,34%.

- Nếu sử dụng các phần mềm thống kê, các kết quả xử lý số liệu bằng máy tính thường luôn thể hiện giá trị P.

2) Nếu quy định trước mức ý nghĩa α, có thể dùng P - value để kết luận theo α. Khi đó nguyên tắc kiểm định như sau:

- P-value <α thì bác bỏ Ho, chấp nhận H1
- P-value ≥ α thì chưa có cơ sở để bác bỏ Ho.

3) Có thể kiểm định giả thuyết Ho theo P-value theo nguyên tắc sau:

- P- value > 0,1 thì thường chấp nhận Ho
- 0,05 < P- value ≤ 0,1 thì cần cân nhắc cẩn thận trước khi bác bỏ Ho (có thể tham khảo thêm tình hình);
- 0,01 < P- value ≤ 0,05 thì nghiêng về hướng bác bỏ Ho nhiều hơn;
- 0,001 < P- value ≤ 0,01 thì ít băn khoăn khi bác bỏ Ho nhều hơn;
- P- value ≤ 0,001 thì có thể yên tâm khi bác bỏ Ho.

Thí dụ 2: **với n <30**

Một nhà sản xuất đèn chiếu X quang cho biết tuổi thọ trung bình của 1 bóng đèn là 100 giờ. Người ta chọn ngẫu nhiên 15 bóng thử nghiệm và cho thấy tuổi thọ trung bình là 99,7 giờ với $S^2 = 0,15$. Giả sử tuổi thọ của bóng đèn tuân theo phân phối chuẩn, hãy đánh giá về tình hình tuổi thọ bóng đèn của nhà máy với mức ý nghĩa α =5 %.

Giải:

- Tuổi thọ trung bình của 1 bóng đèn theo tiêu chuẩn là 100 giờ $\mu_o = 100$;
- Gọi tuổi thọ trung bình của 1 bóng đèn thực tế là μ μ chưa biết
- Đặt giả thuyết Ho: $\mu = \mu_o = 100$; Đối thuyết H1: $\mu \neq \mu_o$
- Kiểm định giả thuyết:

Với n = 15 < 30; $S^2 = 0,15$; $\bar{X} = 99,7$; $\mu_o = 100$; α =5 % ta tính T lý thuyết:

T (n-1; α/2) = T (14; 0.025) = 2,145

Tính T thực nghiệm theo công thức sau:

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{99,7 - 100}{\frac{\sqrt{0,15}}{\sqrt{15}}} = 3$$

Vì $|T| = 3 > T_{n-1, \alpha/2} = 2,145$ nên ta bác bỏ giả thuyết Ho, chấp nhận H1, tức là tuổi thọ trung bình của 1 bóng đèn thực tế khác với qui định (thấp hơn) với mức ý nghĩa là 5%. Trong trường hợp này ta bác bỏ giả thuyết Ho, cũng có nghĩa là khả năng có thể mắc sai lầm loại 1 trong kết luận của mình là 5%.

Chú ý:

1. Trong thực tế chúng ta cũng có thể tìm giá trị P (P-value) bằng cách dùng hàm TDIST trên EXCEL với cấu tạo lệnh như sau:

$$= \mathbf{TDIST\ (T_{tn}, n-1, 1)}$$

Trong đó: T_{tn}: Giá trị T thực nghiệm

n: Số mẫu quan sát

1: 1 phía

Lấy lại thí dụ trên:

P- value = P(T>3) = P(T<-3) = TDIST(3,14,1) = 0,004776

$\alpha/2$ = 0,004776 suy ra α = 2 x 0,004776 = 0,009552 = 0,95%

Kết luận: Giả thuyết Ho bị bác bỏ ở bất kỳ mức ý nghĩa α nào lớn hơn 0,95% (α > 0,95%).

1.2.2. Kiểm định giả thuyết về tỷ lệ của tổng thể

a) Bài toán:

- Giả sử một tổng thể được chia thành 2 loại với tính chất khác nhau. Tỷ lệ số phân tử có tính chất A là p (P thực nghiệm chưa biết). Ta cần kiểm định giả thuyết:

Ho: **$P=P_o$ (P_o cho trước)**;

H1: **$P \neq P_o$**

- Lấy mẫu gồm n quan sát độc lập, thu thập thông tin, tính toán tỷ lệ mẫu p. Thực hiện kiểm định giả thuyết Ho ở mức ý nghĩa α cho trước. Với **n ≥ 40**; tỷ lệ mẫu p có phân phối chuẩn, kiểm định giả thuyết P thực hiện như sau:

+ Đặt giả thuyết

- Kiểm định hai phía Ho : **$P = P_o$** ; H_1 : **$P \neq P_o$**

- Kiểm định 1 phía Ho : **$P \geq P_o$** ; H_1 : **$P < P_o$**

Hoặc Ho : **$P \leq P_o$** ; H_1 : **$P > P_o$**

- Tính giá trị kiểm định Z (Z thực nghiệm) theo công thức:

$$Z = \frac{\phi - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

Trong đó: $P_\circ$: Giá trị cụ thể cho trước

ϕ : Tỷ lệ của mẫu

n : Số đơn vị mẫu quan sát

Quy tắc kiểm định được tóm tắt như sau:

Giả thuyết	**Bác bỏ Ho khi**
Ho : **$P = P_o$** H_1 : **$P \neq P_o$**	$Z > Z_{\alpha/2}$ hoặc $Z < -Z_{\alpha/2}$ hay $\lvert Z \rvert > Z_{\alpha/2}$

Ho : $\mathbf{P \geq P_o}$	H_1 : $\mathbf{P < P_o}$	$Z <- Z_\alpha$
Ho : $\mathbf{P \leq P_o}$	H_1 : $\mathbf{P > P_o}$	$Z > Z_\alpha$

Tìm $Z_{\alpha/2}$ bằng cách tra bảng hoặc dùng hàm NORMSINV với α hoặc $\alpha/2$ trong EXCEL.

Chú ý:

+ Nếu $|Z| \leq Z_{\alpha/2}$ ta chấp nhận giả thuyết Ho, coi P= Po

+ Nếu $|Z| > Z_{\alpha/2}$ ta bác bỏ giả thuyết Ho, coi $P \neq Po$ và khi đó :

- Nếu ϕ (tỷ lệ mẫu) > Po ta xem P >Po
- Nếu ϕ (tỷ lệ mẫu) < Po ta xem P <Po.

b) Thí dụ:

Nhà máy sữa VINAMILK sản xuất sữa chua theo công nghệ cũ thì tỷ lệ sữa loại 1 đạt là 0,2. Nhà máy áp dụng công nghệ mới của Pháp từ năm 2005. Để có nhận xét về chất lượng sản phẩm áp dụng theo công nghệ mới, người ta tiến hành điều tra 500 hộp cho thấy có 150 hộp đạt chất lượng loại 1. Với mức ý nghĩa α =1%, hãy kiểm định chất lượng sản phẩm do áp dụng công nghệ mới.

Giải:

Ta có Po = 0,2; gọi chất lượng sản phẩm do áp dụng công nghệ mới là P (P chưa biết).

Đặt giả thuyết **Ho: P = Po = 0,2**; **H1: P ≠ Po ≠ 0,2**.

Kiểm định giả thuyết Ho:

- Tính ϕ (tỷ lệ mẫu) = 150/500 = 0,3; n = 500
- Tính Z lý thuyết: $\mathbf{Z_{\alpha/2} = Z_{0.005} = 2,58}$
- Tính Z kiểm định với Po = 0,2; ϕ (tỷ lệ mẫu) = 0,3.

$$Z = \frac{\phi - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0,3 - 0,2}{\sqrt{\frac{0,2(1-0,2)}{500}}} = 5,59$$

Như vậy, $|Z| = 5,59 > Z_{\alpha/2} = 2,58$ nên ta bác bỏ Ho, nghĩa là $\mathbf{P \neq Po \neq 0.2}$. Do ϕ (tỷ lệ mẫu) = 0,3 >Po = 0,2 nên P > Po. áp dụng công nghệ mới chất lượng sản phẩm loại 1 cao hơn phương pháp cũ.

1.2.3. Kiểm định giả thuyết về sự khác nhau giữa 2 số trung bình của 2 tổng thể

a) Lấy mẫu từng cặp:

+ Bài toán

Giả sử ta có **n quan** sát về một tiêu thức nào đó cần so sánh (theo hai thời gian, không gian hoặc kỳ thực hiện với kế hoạch …). Như vậy, **n quan sát sẽ được lấy mẫu theo từng cặp phối hợp từ 2 tổng thể X và Y** như sau:

Quan sát	X	Y	X-Y
1	X_1	Y_1	X_1- Y_1
2	X_2	Y_2	X_2 -Y_2
3	X_3	Y_3	X_3 –Y_3
.	.	.	.
.	.	.	.
.	.	.	.
n	Xn	Yn	Xn -Yn
Trung bình	μ_x	μ_y	Ď
Phương sai	δ^2_x	δ^2_y	S^2_d
Độ lệch chuẩn	δ_x	δ_y	Sd

μ_x : Trung bình của tổng thể X

μ_y : Trung bình của tổng thể Y

Ď : Trung bình của tổng thể sai lệch X - Y

Sd : Độ lệch chuẩn của tổng thể X-Y

Giả sử tổng thể các sai lệch giữa X và Y (X-Y) có phân phối chuẩn. Ta cần kiểm định giả thuyết sau:

Ho: $\mu_x - \mu_y = Do$ (Do là giá trị cho trước

Do = 0)

H1: $\mu_x - \mu_y \neq Do$

Hay:

+ Nguyên tắc kiểm định

- Tính giá trị t kiểm định

Trong đó:

$$T = \frac{Ď - D_o}{\frac{Sd}{\sqrt{n}}}$$

D_o : Giá trị cụ thể cho trước

Ď: Trung bình của tổng thể sai lệch (X - Y)

n: Số đơn vị mẫu quan sát

T: Tiêu chuẩn kiểm định (T thực nghiệm)

Sd: Độ lệch chuẩn của tổng thể sai lệch (X - Y)

- Tìm T lý thuyết với bậc tự do là n-1; $\alpha/2$. Ta có thể tra bảng phân phối Student với n-1 và $\alpha/2$; hoặc tìm hàm TINV(n-1, α).

- Quy tắc kiểm định được tóm tắt như sau:

Giả thuyết	Bác bỏ Ho khi
Ho : $\mu_x - \mu_y = Do$ **H1 : $\mu_x - \mu_y \neq Do$**	**$T > T_{n-1,\alpha/2}$** hoặc **$T < -T_{n-1,\alpha/2}$** Hay **$\|T\| > T_{n-1,\alpha/2}$**
Ho : $\mu_x - \mu_y = Do$ hoặc **$\mu_x - \mu_y \geq D_o$;** **H_1 : $\mu_x - \mu_y < Do$**	**$T < -T_{n-1.,\alpha}$**

Ho : $\mu_x - \mu_y = Do$ hoặc **$\mu_x - \mu_y \leq D_o$;** **$H_1 : \mu_x - \mu_y > Do$**	**$T > T_{n-1,\alpha}$**

- So sánh T thực nghiệm với T lý thuyết"

Nếu $|T| \leq T_{n-1,\alpha/2}$ ta chấp nhận giả thuyết Ho,

Nếu $|T| > T_{n-1,\alpha/2}$ ta bác bỏ giả thuyết Ho và khi đó:

- Nếu Ď > Do thì $\mu_x - \mu_y > 0$
- Nếu Ď < Do thì $\mu_x - \mu_y < 0$

+ Thí dụ: Công ty VINAMILK áp dụng công nghệ mới trong chế biến sữa chua. Hãy kiểm định xem năng suất lao động của công nhân sau khi sử dụng công nghệ mới với công nghệ cũ có khác nhau không với mức ý nghĩa là 5% ?

Giải: Lấy mẫu 10 công nhân trong Công ty, thu thập số liệu về năng suất lao động của 10 công nhân này trước và sau khi áp dụng công nghệ mới. Kết quả điều tra thể hiện ở bảng 4.6.

Bảng 4.6. Năng suất lao động (NSLĐ) của 10 công nhân điều tra

Thứ tự công nhân quan sát	NSLĐ (kg/ngày)		X - Y
	Trước khi X	Sau khi Y	
1	50	52	-2
2	48	46	2
3	45	50	-5
4	60	65	-5
5	70	78	-8
6	62	61	1
7	55	58	-3
8	62	70	-8
9	58	67	-9
10	53	65	-12
Trung bình	56,30	61,20	-4,90
Phương sai	57,57	97,07	20,10
Độ lệch chuẩn	7,59	9,85	4,4833

μ_x NSLĐ trung bình của 10 công nhân theo công nghệ cũ = 56,30

μ_y NSLĐ trung bình của 10 công nhân theo công nghệ mới = 61,20

Ď : Trung bình của tổng thể sai lệch X – Y = 4,9

Sd : Độ lệch chuẩn của tổng thể X - Y = 4,4833

Ta cần kiểm định giả thuyết sau:

Ho: $\mu_x - \mu_y = Do = 0$

H1: $\mu_x - \mu_y \neq Do \neq 0$

Tính T kiểm định:

$$T = \frac{Ď - D_o}{\frac{Sd}{\sqrt{n}}} = \frac{4,9 - 0}{\frac{4,4833}{\sqrt{10}}} = \frac{4,9}{1,4177} = 3,456$$

- Tìm T lý thuyết với bậc tự do là 9; $\alpha = 0,025$: Ta tìm hàm TINV(9, 0,05)= 2,262;

Như vậy, $|T|$ kiểm định = 3,456 >T lý thuyết = 2,262 ta bác bỏ Ho, nghĩa là năng suất lao động của công nhân sau khi áp dụng công nghệ mới khác với công nghệ cũ.

Vì Ď = 4,9 > Do nên $\mu_x - \mu_y > 0$, nghĩa là ở mức ý nghĩa 5% áp dụng công nghệ mới đã làm tăng năng suất so với công nghệ cũ.

b) Trường hợp lấy mẫu độc lập:

+ Bài toán:

Giả sử ta có n_x và n_y là số đơn vị mẫu được chọn ngẫu nhiên, độc lập từ hai tổng thể X và Y có phân phối chuẩn, thể hiện ở bảng sau:

Quan sát	X	Y
1	X1	Y1
2	X2	Y2
3	X3	Y3
.	.	.
.	.	.
N	Xn	Yn
Số quan sát	n_x	n_y
Trung bình mẫu	x	ŷ
Trung bình	μ_x	μ_y
Phương sai	δ^2_x	δ^2_y
Độ lệch chuẩn	δ_x	δ_y

μ_x Trung bình của tổng thể X

μ_y Trung bình của tổng thể Y

$\hat{x}$, $\hat{y}$ là trung bình của 2 mẫu chọn ngẫu nhiên từ 2 tổng thể X ; Y

δ^2_x và δ^2_y là phương sai của tổng thể X và Y

Với mức ý nghĩa α, cần kiểm định giả thuyết sau:

Ho: $\mu_x - \mu_y = Do$ (Do là giá trị cho trước Do=0)

H1: $\mu_x - \mu_y \neq Do$

Hay:

Ho: $\mu_x - \mu_y = 0$; H1: $\mu_x - \mu_y \neq 0$

+ Nguyên tắc kiểm định: Có 2 trường hợp xảy ra

1) Nếu $n_x, n_y \geq 30$, với X, Y tuân theo phân phối chuẩn và $\delta^2_x \neq \delta^2_y$

Tính tiêu chuẩn kiểm định Z (Z thực nghiệm):

$$Z = \frac{\hat{x} - \hat{y} - D_o}{\sqrt{\frac{\delta^2_x}{n_x} + \frac{\delta^2_y}{n_y}}}$$

Trong đó:

D_o : Giá trị cụ thể cho trước (Do =0)

$\hat{x}$, $\hat{y}$: Trung bình của 2 mẫu

δ^2_x và δ^2_y : Phương sai của tổng thể X và Y

n_x , n_y : Số đơn vị mẫu quan sát của tổng thể X và Y

Z: Tiêu chuẩn kiểm định (Z thực nghiệm)

- Tìm Z lý thuyết:

Tìm $Z_{\alpha/2}$ bằng cách tra bảng hoặc dùng hàm NORMSINV với $\alpha/2$ trong EXCEL.

Quy tắc kiểm định được tóm tắt như sau:

Giả thuyết	Bác bỏ Ho khi
Ho : $\mu_x - \mu_y = Do$ **H1 : $\mu_x - \mu_y \neq Do$**	**$Z > Z_{\alpha/2}$** hoặc **$Z < -Z_{\alpha/2}$** hay **$\|Z\| > Z_{\alpha/2}$**
Ho : $\mu_x - \mu_y = Do$ hoặc **$\mu_x - \mu_y \geq D_o$** ; **$H_1 : \mu_x - \mu_y < Do$**	**$Z < -Z_{\alpha}$**
Ho : $\mu_x - \mu_y = Do$ hoặc **$\mu_x - \mu_y \leq D_o$** ; **$H_1 : \mu_x - \mu_y > Do$**	**$Z > Z_{\alpha}$**

Chú ý:

+ Nếu $|Z| \leq Z_{\alpha/2}$ ta chấp nhận giả thuyết Ho, coi **$\mu_x - \mu_y = Do$**

+ Nếu $|Z| > Z_{\alpha/2}$ ta bác bỏ giả thuyết Ho, coi **$\mu_x - \mu_y \neq Do$** và khi đó :

Nếu $\hat{x} > \hat{y}$ ta xem $\mu_x > \mu_y$

Nếu $\hat{x} < \hat{y}$ ta xem $\mu_x < \mu_y$

+ Nếu chưa biết phương sai của tổng thể, mà số đơn vị mẫu lớn ($n_x, n_y \geq 30$) ta vẫn dùng công thức trên để tính Z kiểm định, thay phương sai tổng thể bằng phương sai mẫu ($\delta^2_x = s^2_x$ và $\delta^2_y = s^2_y$).

Thí dụ: Một trại chăn nuôi gà tiến hành thí nghiệm sử dụng 2 loại thức ăn A và B trên cùng một giống. Sau một thời gian thử nghiệm cho ăn, người ta điều tra 50 con nuôi bằng thức ăn A và 40 con nuôi bằng thức ăn B thu được các số liệu sau:

Bảng 5.6. Một số chỉ tiêu của 2 mẫu thí nghiệm cho ăn 2 loại thức ăn A và B

Diễn giải	ĐVT	Thức ăn A	Thức ăn B
1. Số đơn vị mẫu quan sát	con	50	40
2. Khối lượng trung bình 1 con	Kg/con	2,2	1,2
3. Độ lệch chuẩn	Kg/con	1,25	1,02

Yêu cầu: Anh, chị hãy cho biết khối lượng trung bình 1 con sử dụng ở 2 loại thức ăn sau thời gian nuôi có khác nhau không với mức ý nghĩa là 5%?

Giải:

- Gọi μ_x và μ_y là khối lượng trung bình 1 con sau khi nuôi sử dụng thức ăn A và B;

- Đặt giả thuyết: **Ho : $\mu_x - \mu_y = 0$**

H1 : $\mu_x - \mu_y \neq 0$

- Tính tiêu chuẩn kiểm định Z:

$$Z = \frac{\hat{x} - \hat{y} - D_o}{\sqrt{\frac{\delta^2_x}{n_x} + \frac{\delta^2_y}{n_y}}} = \frac{\mathbf{2,2 - 1,2 - 0}}{\sqrt{\frac{1,25^2}{50} + \frac{1,02^2}{40}}} = \frac{\mathbf{1}}{0,2392} = 4,179$$

- Tìm Z lý thuyết qua hàm NORMSINV với $\alpha = 0,025$ trong EXCEL ta được Z lý thuyết = 1,96.

- $|Z| = 4,179 > Z_{\alpha/2} = 1,96$ ta bác bỏ giả thuyết Ho, coi $\mu_x - \mu_y \neq 0$.

Vì $\hat{x}$ =2,2 kg/con > $\hat{y}$ = 1,2 kg/con nên ta xem $\mu_x > \mu_y$, chứng tỏ khối lượng trung bình 1 con nuôi bằng thức ăn A lớn hơn nuôi bằng thức ăn B.

2) Nếu $n_x, n_y < 30$ với X; Y đều tuân theo phân phối chuẩn và $\delta^2_x = \delta^2_y$

Với mức ý nghĩa α, Ta cần kiểm định giả thuyết sau:

Ho: $\mu_x - \mu_y = D_o$ (Do là giá trị cho trước Do = 0)

H1: $\mu_x - \mu_y \neq D_o$

Hay:

Ho: $\mu_x - \mu_y = 0$; H1: $\mu_x - \mu_y \neq 0$

- Tính tiêu chuẩn kiểm định T:

$$T = \frac{\hat{x} - \hat{y} - \mathbf{D_o}}{\sqrt{s^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

Trong đó:

$\mathbf{D}_\circ$: Giá trị cụ thể cho trước (Do = 0)

$\hat{x}$, $\hat{y}$: Trung bình của 2 mẫu

n_x, n_y: Số đơn vị mẫu quan sát của tổng thể X và Y

T: Tiêu chuẩn kiểm định (T thực nghiệm)

s^2 được tính theo công thức sau:

$$s^2 = \frac{(n_x-1)\,s^2_x + (n_y-1)s^2_y}{(n_x + n_y - 2)}$$

- Tìm T lý thuyết:

Từ α cho trước, tra bảng phân phối student với bậc tự do là $(n_x + n_y - 2)$ để tìm $T(n_x + n_y - 2; \alpha/2)$, hoặc tra hàm TINV $((n_x + n_y - 2; \alpha)$ trong EXCEL;

- Quy tắc kiểm định được tóm tắt như sau:

Giả thuyết	Bác bỏ Ho khi
Ho : $\mu_x - \mu_y = Do$ **H1 : $\mu_x - \mu_y \neq Do$**	**$T > T_{nx+ny-2;\alpha/2}$** hoặc **$T < -T_{nx+ny-2;\alpha/2}$** hay **$\lvert T \rvert > T_{nx+ny-2;\alpha/2}$**
Ho : $\mu_x - \mu_y = Do$ hoặc **$\mu_x - \mu_y \geq D_o$** **$H_1 : \mu_x - \mu_y < Do$**	**$T < -T_{nx+ny-2;\alpha}$**
Ho : $\mu_x - \mu_y = Do$ hoặc **$\mu_x - \mu_y \leq D_o$** **$H_1 : \mu_x - \mu_y > Do$**	**$T > T_{nx+ny-2;\alpha}$**

- So sánh T thực nghiệm với T lý thuyết:

Nếu $\lvert T \rvert \leq T_{(nx+ny-2;\alpha/2)}$ ta chấp nhận giả thuyết Ho.

Nếu $\lvert T \rvert > T_{(nx+ny-2;\alpha/2)}$ ta bác bỏ giả thuyết Ho và khi đó:

Nếu $\hat{x} > \hat{y}$ ta xem $\mu_x > \mu_y$

Nếu $\hat{x} < \hat{y}$ ta xem $\mu_x < \mu_y$

Thí dụ: (Lấy lại ví dụ trên)

Một trại chăn nuôi gà tiến hành thí nghiệm sử dụng 2 loại thức ăn A và B trên cùng một giống. Sau một thời gian thử nghiệm cho ăn, người ta điều tra 20 con nuôi bằng thức ăn A và 15 con nuôi bằng thức ăn B thu được các số liệu sau:

Bảng 6.6. Một số chỉ tiêu của 2 mẫu thí nghiệm cho ăn 2 loại thức ăn A và B

Diễn giải	ĐVT	Thức ăn A	Thức ăn B
1. Số đơn vị mẫu quan sát	Con	20	15
2. Khối lượng trung bình 1 con	Kg/con	2,2	1,2
3. Độ lệch chuẩn	Kg/con	1,25	1,02

Yêu cầu: Anh chị hãy cho biết khối lượng trung bình 1 con sử dụng ở 2 loại thức ăn sau thời gian nuôi có khác nhau không với mức ý nghĩa là 5%?

Giải:

- Gọi μ_x và μ_y là khối lượng trung bình 1 con sau khi nuôi sử dụng thức ăn A và B;

- Đặt giả thuyết: **Ho : $\mu_x - \mu_y = 0$**

H1 : $\mu_x - \mu_y \neq 0$

- Vì số mẫu quan sát n_x, $n_y < 30$, ta giả định phương sai của 2 tổng thể bằng nhau.
- Tính tiêu chuẩn kiểm định T:

$$T = \frac{\hat{x} - \hat{y} - D_o}{\sqrt{s^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} = \frac{2,2 - 1,2 - 0}{\sqrt{1,34\left(\frac{1}{20} + \frac{1}{15}\right)}} = \frac{1}{0,1564} = 6,39$$

s^2 được tính theo công thức sau:

$$s^2 = \frac{(n_x-1)\,s^2_x + (n_y-1)s^2_y}{(n_x + n_y - 2)} = \frac{(20-1)1,25^2 + (15-1)1,02^2}{(20+15-2)} = \frac{44,2531}{33} = 1,34$$

- Tìm T lý thuyết:

Tra hàm TINV với bậc tự do là 33; $\alpha = 0,05$ ta được **T lý thuyết = 2,03.**

Như vậy $|T| = 6,39 > \mathbf{T}_{(nx + ny -2;\, \alpha/2)} = 2,03$ ta bác bỏ giả thuyết Ho.

Vì x = 2,2 kg/con > $\hat{y}$ = 1,2 kg/con nên ta xem $\mu_x > \mu_y$, chứng tỏ khối lượng trung bình 1 con nuôi bằng thức ăn A lớn hơn nuôi bằng thức ăn B.

1.2.4. Kiểm định giả thuyết về sự bằng nhau giữa 2 phương sai của 2 tổng thể:

a) Bài toán

Giả sử ta có n_x và n_y là số đơn vị mẫu được chọn ngẫu nhiên, độc lập từ hai tổng thể X và Y có phân phối chuẩn , thể hiện ở bảng sau:

Quan sát	X	Y
1	X1	Y1
2	X2	Y2
3	X3	Y3
.	.	.
.	.	.
n	Xn	Yn
Số quan sát	n_x	n_y
Trung bình mẫu	x	ŷ

μ_x : Trung bình của tổng thể X

μ_y : Trung bình của tổng thể Y

$\hat{x}$, $\hat{y}$: Trung bình của 2 mẫu chọn ngẫu nhiên từ 2 tổng thể X ; Y

δ^2_x và δ^2_y : Phương sai của tổng thể X và Y

s^2_x và s^2_y : Phương sai của 2 mẫu n_x và n_y

Với mức ý nghĩa α ta cần kiểm định giả thuyết sau:

Ho : $\delta^2_x = \delta^2_y$

H1 : $\delta^2_x \neq \delta^2_y$

Trung bình	μ_x	μ_y
Phương sai	δ^2_x	δ^2_y
Phương sai mẫu	s^2_x	s^2_y

b) Nguyên tắc kiểm định

- Tính tiêu chuẩn kiểm định F (F kiểm định):

$$F = \frac{s_y^2}{s_x^2}$$ Với giả thiết $s^2_x > s^2_y$ hoặc ngược lại.

- Tìm F lý thuyết:

Ta tra bảng FISHER – SNEDECOR với n_x-1 và n_y-1 bậc tự do ; $\alpha/2$

$\mathbf{F}_{(nx-1;\ ny-1;\ \alpha/2)}$; hoặc tìm hàm FINV ($n_x$-1 ; n_y-1; $\alpha/2$).

- Quy tắc kiểm định được tóm tắt như sau:

Giả thuyết	Bác bỏ Ho khi
Ho : $\delta^2_x = \delta^2_y$ **H1 : $\delta^2_x \neq \delta^2_y$**	**$F > F_{(nx-1;\ ny-1;\ \alpha/2)}$** hoặc **$F < -F_{(nx-1;\ ny-1;\ \alpha/2)}$** hay **$\lvert T \rvert > F_{(nx-1;\ ny-1;\ \alpha/2)}$**
Ho : $\delta^2_x = \delta^2_y$ hoặc $\delta^2_x \leq \delta^2_y$; **$H_1 : \delta^2_x > \delta^2_y$**	**$F > F_{(nx-1;\ ny-1;\ \alpha)}$**

- So sánh F thực nghiệm với F lý thuyết:

Nếu $|F| > \mathbf{F}_{(nx-1;\ ny-1;\ \alpha/2)}$ ta bác bỏ giả thuyết Ho,

Nếu $|F| \leq \mathbf{F}_{(nx-1;\ ny-1;\ \alpha/2)}$ ta chấp nhận giả thuyết Ho.

Trong trường hợp bác bỏ giả thuyết Ho:

Nếu $s^2_x > s^2_y$ ta xem $\delta^2_x > \delta^2_y$

Nếu $s^2_x < s^2_y$ ta xem $\delta^2_x < \delta^2_y$.

Thí dụ: Công ty chè Phú Đa sử dụng 2 máy đóng gói chè đen xuất khẩu. Để kiểm tra mức độ chính xác của 2 máy này, người ta chọn ra 20 túi sản phẩm từ máy thứ nhất, và 15 túi sản phẩm từ máy thứ hai. Tính toán phương sai về khối lượng trung bình 1 túi cho thấy ở máy 1 là 17 gam/túi, máy 2 là 26 gam/túi. Với mức ý nghĩa là 5% hãy cho biết độ chính xác của 2 máy có như nhau không?

Giải:

Gọi δ^2_x là phương sai đo sự biến động về khối lượng sản phẩm trung bình 1 túi đóng gói từ máy 1; δ^2_y là phương sai đo sự biến động về khối lượng sản phẩm trung bình 1 túi đóng gói từ máy 2.

- Đặt giả thuyết:

Ho : $\delta^2_x = \delta^2_y$

H1 : $\delta^2_x \neq \delta^2_y$

- Tính tiêu chuẩn kiểm định F :

$$F = \frac{s_y^2}{s_x^2} = \frac{26}{17} = 1{,}529 \quad (s_y^2 > s_x^2)$$

- Tìm F lý thuyết:

Tìm hàm **FINV (n_x-1 ; n_y-1; α/2) = FINV (14,19,0,025) = 2,65**

- Do $|F| = 1{,}529 \leq F_{nx-1;\ ny-1;\ \alpha/2} = 2{,}65$ ta chấp nhận giả thuyết Ho, nghĩa là mức độ chính xác của 2 máy đóng gói là như nhau.

1.2.5. Kiểm định giả thuyết về sự bằng nhau giữa 2 tỷ lệ của 2 tổng thể:

a) Bài toán

Giả sử ta có n_x và n_y là số đơn vị mẫu được chọn ngẫu nhiên, độc lập từ hai tổng thể X và Y có phân phối chuẩn , thể hiện ở bảng sau:

Quan sát	X	Y
1	X1	Y1
2	X2	Y2
3	X3	Y3
.	.	.
.	.	.
n	Xn	Yn
Số quan sát	n_x	n_y
Trung bình mẫu	$\hat{x}$	$\hat{y}$
Trung bình	μ_x	μ_y
Tỷ lệ của tổng thể	P_x	P_y
Tỷ lệ của mẫu	$\jmath_x$	$\jmath_y$

μ_x : Trung bình của tổng thể X

μ_y : Trung bình của tổng thể Y

$\hat{x}$, $\hat{y}$: Trung bình của 2 mẫu chọn ngẫu nhiên từ 2 tổng thể X ; Y

P_x; P_y : Tỷ lệ của các đơn vị có cùng một tính chất trong tổng thể X và Y

$\jmath_x$; $\jmath_y$: Tỷ lệ của các đơn vị có cùng một tính chất trong tổng thể mẫu n_x và n_y

Với mức ý nghĩa α, ta cần kiểm định giả thuyết sau:

Ho : $P_x - P_y = 0$

H1 : $P_x - P_y \neq 0$

b) Nguyên tắc kiểm định

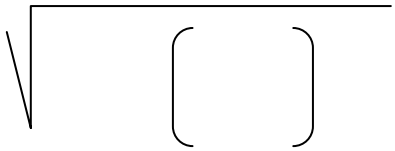
- Tính tiêu chuẩn kiểm định Z (Z kiểm định) với n_x và $n_y \geq 40$

$$Z = \frac{\jmath_x - \jmath_y}{\jmath_0 (1 - \jmath_0)\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$$

Trong đó:

$\jmath_0$ được tính theo công thức sau:

$$\jmath_0 = \frac{n_x \jmath_x + n_y \jmath_y}{(n_x + n_y)}$$



- Tìm Z lý thuyết:

Tìm $Z_{\alpha/2}$ bằng cách tra bảng hoặc dùng hàm NORMSINV với $\alpha/2$ trong EXCEL.

Quy tắc kiểm định được tóm tắt như sau:

Giả thuyết	Bác bỏ Ho khi
Ho : $P_x - P_y = 0$ **H1 : $P_x - P_y \neq 0$**	**$Z > Z_{\alpha/2}$** hoặc **$Z < -Z_{\alpha/2}$** hay **$\|Z\| > Z_{\alpha/2}$**
Ho : $P_x - P_y = 0$ hoặc **$P_x - P_y \geq 0$** **H1 : $P_x - P_y < 0$**	**$Z < -Z_{\alpha}$**
Ho : $P_x - P_y = 0$ hoặc **$P_x - P_y \leq 0$** **H1 : $P_x - P_y > 0$**	**$Z > Z_{\alpha}$**

Chú ý:

+ Nếu $|Z| \leq Z_{\alpha/2}$ ta chấp nhận giả thuyết Ho,

+ Nếu $|Z| > Z_{\alpha/2}$ ta bác bỏ giả thuyết Ho và khi đó:

Nếu $\jmath_x > \jmath_y$ ta xem $P_x > P_y$

Nếu $\jmath_x < \jmath_y$ ta xem $P_x < P_y$

Thí dụ: Để kiểm tra chất lượng sản phẩm đúng quy cách của 2 phân xưởng, Công ty chè Phú Đa tiến hành kiểm tra ngẫu nhiên 200 gói sản phẩm ở phân xưởng A, và 220 gói sản phẩm của phân xưởng B. Kết quả kiểm tra cho thấy số gói sản phẩm sai hỏng của phân xưởng A là 20 gói, phân xưởng B là 5 gói. Với mức ý nghĩa là 1% hãy cho biết tỷ lệ sai hỏng của 2 phân xưởng có như nhau không?

Giải: Gọi tỷ lệ sai hỏng sản phẩm của phân xưởng A là **P_x** ; của phân xưởng B là **P_y**

Đặt giả thuyết: **Ho: $P_x - P_y = 0$** và **H1: $P_x - P_y \neq 0$**

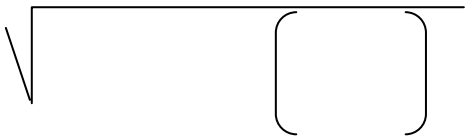
- Tính tiêu chuẩn kiểm định Z với **$\jmath_x = 20/200 = 0,1$**; **$\jmath_y = 5/220 = 0,0227$**

$$Z = \frac{\jmath_x - \jmath_y}{\sqrt{\jmath_0(1-\jmath_0)\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

Trong đó: $\jmath_0$ được tính theo công thức sau:

$$\jmath_0 = \frac{n_x\jmath_x + n_y\jmath_y}{(n_x + n_y)} = \frac{20 + 5}{200 + 220} = 0,0595$$

$$Z = \frac{0,1 - 0,0227}{0,0595(1-0,0595)\left(\frac{1}{200} + \frac{1}{220}\right)} = \frac{0,0773}{0,0231} = \mathbf{3,34}$$



- Tìm Z lý thuyết ($\mathbf{Z_{\alpha/2}}$= $Z_{0,005}$). Tìm hàm NORMSINV với $\alpha/2 = 0,005$ trong EXCEL ta được Z lý thuyết = **2,58.**

$|Z| = 3,34 > Z_{\alpha/2} = 2,58$ ta bác bỏ giả thuyết Ho, nghĩa là $\mathbf{P_x - P_y \neq 0.}$

Vì $\mathbf{\rfloor_x = 0,1 > \rfloor_y = 0,0227}$ ta xem $\mathbf{P_x > P_y,}$ nghĩa là tỷ lệ sai hỏng của phân xưởng A lớn hơn phân xưởng B.

2. PHÂN TÍCH PHƯƠNG SAI

Mục tiêu của phân tích phương sai là so sánh trung bình của nhiều nhóm dựa trên các số trung bình mẫu và thông qua kiểm định giả thuyết để kết luận về sự bằng nhau của các số trung bình này.

Trong nghiên cứu, phân tích phương sai được dùng như là một công cụ để xem xét ảnh hưởng của một hay một số yếu tố nguyên nhân (định tính) đến một yếu tố kết quả kia (định lượng).

Thí dụ: Nghiên cứu ảnh hưởng của phương pháp chấm điểm đến kết quả học tập của sinh viên. Nghiên cứu ảnh hưởng của bậc thợ tới năng suất lao động. Nghiên cứu ảnh hưởng của loại lò, loại chất đốt đến chi phí chất đốt (kg/h) để sấy vải khô.

2.1. Phân tích phương sai một yếu tố

a) Bài toán:

Phân tích phương sai một yếu tố là phân tích ảnh hưởng của một yếu tố nguyên nhân (thường là yếu tố định tính) đến một yếu tố kết quả (thường là yếu tố định lượng) đang nghiên cứu.

Giả sử chúng ta cần so sánh số trung bình của k tổng thể độc lập. Người ta lấy k mẫu có số quan sát là n_1; n_2… n_k; tuân theo phân phối chuẩn. Trung bình của các tổng thể được ký hiệu là μ_1; μ_2 …. μ_k thì mô hình phân tích phương sai một yếu tố ảnh hưởng được mô tả dưới dạng kiểm định giả thuyết có dạng như sau:

Ho: $\mu_1 = \mu_2 = \ldots = \mu_k$

H1: Tồn tại ít nhất 1 cặp có $\mu_1 \neq \mu_2$; $\mu_2 \neq \mu_k$

Để kiểm định ta đưa ra 2 giả thiết sau:

1) Mỗi mẫu tuân theo phân phối chuẩn $N(\mu, \sigma^2)$

2) Ta lấy k mẫu độc lập từ k tổng thể. Mỗi mẫu được quan sát n_j lần.

b) Các bước tiến hành:

Bước 1: Tính các trung bình mẫu và trung bình chung của k mẫu

Ta lập bảng tính toán như sau:

TT	k mẫu quan sát				
	1	2	3	…	k
1	X_{11}	X_{12}	X_{13}	…	X_{1k}
2	X_{21}	X_{22}	X_{23}	…	X_{2k}
3	X_{31}	X_{32}	X_{33}	…	X_{3k}
…					
…					
J	X_{j1}	X_{j2}	X_{j3}	…	X_{jk}
Trung bình mẫu	$\overline{x1}$	$\overline{x2}$			

Trung bình mẫu $\overline{x1}$; $\overline{x2}$ ... $\overline{xk}$ được tính theo công thức

$$\overline{xi} = \frac{\sum_{j=1}^{ni} Xij}{n_i} \quad (i = 1,2...k)$$

Trung bình chung của k mẫu được tính theo công thức ⇒

$$\overline{x} = \frac{\sum_{i=1}^{k} ni\ \overline{xi}}{\sum_{i=1}^{k} ni}$$

Bước 2: Tính các tổng độ lệch bình phương

Ở bước này cần tính tổng các độ lệch bình phương trong nội bộ nhóm (nội bộ từng mẫu - SSW) và tổng các độ lệch bình phương giữa các nhóm (SSB).

- Tổng các độ lệch bình phương trong nội bộ nhóm (nội bộ từng mẫu - SSW) được tính theo công thức sau:

Nhóm 1	Nhóm 2	Nhóm k
$SS_1 = \sum_{j=1}^{n1} (Xj1 - \overline{x1})^2$	$SS_2 = \sum_{j=1}^{n2} (Xj2 - \overline{x2})^2$	$SS_k = \sum_{j=1}^{nk} (Xjk - \overline{xk})^2$
$SSW = SS_1 + SS_2 + ...+ SS_k = \sum_{i=1}^{k} \sum_{ij=11}^{ni} (Xij - \overline{xi})^2$		

- Tổng các độ lệch bình phương giữa các nhóm (SSG) được tính như sau:

$$SSB = \sum_{i=1}^{k} ni\,(\overline{xi} - \overline{x})^2$$

- Tổng các độ lệch bình phương của toàn bộ tổng thể (SST) bằng tổng các độ lệch bình phương trong nội bộ nhóm (nội bộ từng mẫu) SSW cộng với tổng các độ lệch bình phương giữa các nhóm SSB.

Cụ thể theo công thức sau:

$$SST = SSW + SSB = \sum_{i=1}^{k} \sum_{j=1}^{ni} (Xij - \overline{x})^2$$

Như vậy, toàn bộ biến thiên của yếu tố kết quả (SST) được phân tích thành 2 phần: phần biến thiên do yếu tố nguyên nhân đang nghiên cứu (SSW); phần biến thiên còn lại do yếu tố khác không nghiên cứu ở đây (MSB). Nếu phần biến thiên do yếu tố nguyên nhân đang nghiên cứu tạo ra càng nhiều so với phần biến thiên do yếu tố khác tạo ra, thì ta càng có cơ sở để bác bỏ Ho và đi đến kết luận yếu tố nguyên nhân có ảnh hưởng có ý nghĩa đến yếu tố kết quả.

Bước 3: Tính các phương sai (phương sai của nội bộ nhóm và phương sai giữa các nhóm)

Ta ký hiệu k là số nhóm (mẫu); n là tổng số quan sát của các nhóm thì các phương sai được tính theo công thức sau:

$MSW = \dfrac{SSW}{n - k}$	$MSB = \dfrac{SSB}{k - 1}$

Bước 4: Kiểm định giả thuyết

- Tính tiêu chuẩn kiểm định F (F thực nghiệm)

$$F = \frac{MSB}{MSW}$$

Trong đó:

MSB : Phương sai giữa các nhóm

MSW : Phương sai trong nội bộ nhóm

- Tìm F lý thuyết (F tiêu chuẩn = F (k-1; n-k; α)):

F lý thuyết là giá trị giới hạn tra từ bảng phân phối F với k-1 bậc tự do của phương sai ở tử số và ; n-k bậc tự do của phương sai ở mẫu số với mức ý nghĩa α. F lý thuyết có thể tra qua hàm FINV(α, k-1, n-1) trong EXCEL.

- Nếu F thực nghiệm > F lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể không bằng nhau.

Bảng phân tích phương sai 1 yếu tố khi sử dụng máy tính (phần mềm EXCEL hoặc SPSS) tóm tắt như sau:

Bảng gốc bằng tiếng Anh

Source of variation	*Sum of squares (SS)*	*Degree of freedom (df)*	*Mean squares (MS)*	*F- ratio*
Between - groups	SSB	(k-1)	MSB	$F = \frac{MSB}{MSW}$
Within - groups	SSW	(n-k)	MSW	
Total	SST	(n-1)		

Bảng phân tích phương sai tổng quát dịch ra tiếng việt – ANOVA

Nguồn biến động	Tổng độ lệch bình phương (SS)	Bậc tự do (df)	Phương sai (MS)	F- Tỷ số
Giữa các mẫu	SSB	(k-1)	MSB	$F = \frac{MSB}{MSW}$
Trong nội bộ các mẫu	SSW	(n-k)	MSW	
Tổng số	SST	(n-1)		

c) Thí dụ:

Có tài liệu về cách cho điểm môn Lý thuyết thống kê của 3 giáo sư như sau (điểm tối đa là 100). Hãy cho biết cách chấm điểm của 3 giáo sư có sai khác nhau không?

TT	**A**	**B**	**C**
1	82	74	79
2	86	82	79
3	79	78	77
4	83	75	78
5	85	76	82
6	84	77	79

Giải:

- Đặt giả thuyết Ho: Cách chấm điểm của 3 giáo sư không sai khác nhau

Ho: $\mu_1 = \mu_2 = \ldots = \mu_k$;

H1: Tồn tại ít nhất 1 cặp có $\mu_1 \neq \mu_2$; $\mu_2 \neq \mu_k$

- Từ kết quả lấy mẫu của 3 nhóm ta tính các độ lệch bình phương thể hiện qua bảng sau:

					SS_1	SS_2	SS_3	
TT	A	B	C	Chung (Xbq)	$(X_1j - \overline{x1})^2$	$(X_2j - \overline{x2})^2$	$(X_3j - \overline{x3})^2$	**Cộng**
1	82	74	79		1,36	9,00	0,00	
2	86	82	79		8,03	25,00	0,00	
3	79	78	77		17,36	1,00	4,00	
4	83	75	78		0,03	4,00	1,00	
5	85	76	82		3,36	1,00	9,00	
6	84	77	79		0,69	0,00	0,00	
Trung bình	$\overline{x1}$ = **83,17**	$\overline{x2}$ = **77,00**	$\overline{x3}$ = **79,00**	$\overline{x}$ = **79,72**				
P.sai (σ_i^2)	6,17	8,00	2,80	11,98				
Cộng					30,83	40,00	14,00	SSW=84,83
$(\overline{xi} - \overline{x})^2 nj$	71,185	44,463	3,130					SSB=118,78

$SSW = SS_1 + SS_2 + SS_3 = 84,83$

$SSB = \sum_{i=1}^{k} ni\,(\overline{xi} - \overline{x})^2 = 118,78$

- Tính các phương sai:

$MSW = \frac{SSW}{n-k} = \frac{84,83}{15} = 5,66$	$MSB = \frac{SSB}{k-1} = \frac{118,78}{2} = 59,39$

- Tính F thực nghiệm:

$$F = \frac{MSB}{MSW} = \frac{59,39}{5,66} = 10,5$$

- Tra bảng F lý thuyết (F (0.05; 2; 15)) = 3,68

- So sánh F thực nghiệm với F lý thuyết ta thấy: F thực nghiệm > F lý thuyết bác bỏ Ho, nghĩa là cách cho điểm của 3 giáo sư có khác nhau.

Sử dụng kết quả của máy tính, phần mềm EXCEL chúng ta cũng có kết quả tương tự (bảng sau).

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
A	6	499	**83,17**	6,17
B	6	462	**77,00**	8,0
C	6	474	**79,00**	2,8

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	118,78	2	59,39	10,50	0,00	3,68
Within Groups	84,83	15	5,66			
Total	203,61	17				

2.2. Phân tích phương sai 2 yếu tố

Phân tích phương sai 2 yếu tố nhằm xem xét cùng lúc hai yếu tố nguyên nhân (dưới dạng dữ liệu định tính) ảnh hưởng đến yếu tố kết quả (dưới dạng dữ liệu định lượng) đang nghiên cứu.

Thí dụ: Nghiên cứu ảnh hưởng của loại chất đốt và loại lò sấy đến tỷ lệ vải loại 1 sấy khô. Phân tích phương sai 2 yếu tố giúp chúng ta đưa thêm yếu tố nguyên nhân vào phân tích làm cho kết quả nghiên cứu càng có giá trị.

a) Bài toán:

Giả sử ta nghiên cứu ảnh hưởng của 2 yếu tố nguyên nhân định tính đến một yếu tố kết quả định lượng nào đó. Ta lấy mẫu không lặp lại, sau đó các đơn vị mẫu của yếu tố nguyên nhân thứ nhất sắp xếp thành K nhóm (cột), các đơn vị mẫu của yếu tố nguyên nhân thứ hai sắp xếp thành H khối (hàng). Như vậy, ta có bảng kết hợp 2 yếu tố nguyên nhân gồm K cột và H hàng và (K x H) ô dữ liệu. Tổng số mẫu quan sát là n = (K x H). Dạng tổng quát như ở bảng 6.6.

Bảng 6.6. Sắp xếp các mẫu quan sát của phân tích phương sai 2 yếu tố không lặp

Hàng (khối)	Cột (nhóm)			
	1	2	...	K
1	X_{11}	X_{21}	X_{31}	X_{K1}
2	X_{12}	X_{22}	X_{32}	X_{K2}
...	...	...	...	...
...	...	...	...	...
H	X_{1H}	X_{2H}	X_{3H}	X_{KH}

Mô hình phân tích phương sai hai yếu tố ảnh hưởng được mô tả dưới dạng kiểm định giả thuyết bao gồm 2 phần :

(1) Kiểm định giả thuyết cho số trung bình của K tổng thể, tương ứng với K nhóm mẫu là bằng nhau;

(2) Kiểm định giả thuyết cho số trung bình của H tổng thể, tương ứng với H khối mẫu là bằng nhau;

Để kiểm định ta đưa ra 2 giả thiết sau:

1) Mỗi mẫu tuân theo phân phối chuẩn $N(\mu, \sigma^2)$

2) Ta lấy K mẫu độc lập từ K tổng thể, H mẫu độc lập từ H tổng thể. Mỗi mẫu được quan sát 1 lần không lặp.

b) Các bước tiến hành:

Bước 1: Tính các số trung bình

Trung bình riêng của từng nhóm (K cột)	**Trung bình riêng của từng khối (H hàng)**	**Trung bình chung của toàn bộ mẫu quan sát**
$\bar{x}_i = \frac{\sum_{j=1}^{H} X_{ij}}{H}$ (i = 1,2...K)	$\bar{x}_j = \frac{\sum_{i=1}^{K} X_{ij}}{K}$ (j = 1,2...H)	$\bar{x} = \frac{\sum_{i=1}^{K}\sum_{J=1}^{H} X_{ij}}{n} = \frac{\sum_{i=1}^{K} \bar{x}_i}{K} = \frac{\sum_{j=1}^{H} \bar{x}_j}{H}$

Bước 2. Tính tổng các độ lệch bình phương

Diễn giải	**Công thức tính**
1. Tổng các độ lệch bình phương chung (SST) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của tất cả các yếu tố*	$SST = \sum_{i=1}^{K}\sum_{J=1}^{H}(Xij - \bar{x})^2$

2. Tổng các độ lệch bình phương giữa các nhóm (SSK) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân thứ nhất (xếp theo cột)*	$SSK = H \sum_{i=1}^{K} (\overline{xi} - \overline{x})^2$
3.Tổng các độ lệch bình phương giữa các nhóm (SSH) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân thứ hai (xếp theo hàng)*	$SSH = K \sum_{J=1}^{H} (\overline{xj} - \overline{x})^2$
4. Tổng các độ lệch bình phương phần dư (ERROR) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân khác không nghiên cứu*	SSE = SST- SSK- SSH

Bước 3. Tính các phương sai

Diễn giải	Công thức
1. Phương sai giữa các nhóm (cột) (MSK)	$MSK = \frac{SSK}{K - 1}$
2. Phương sai giữa các khối (hàng) (MSH)	$MSH = \frac{SSH}{H - 1}$
3. Phương sai phần dư (MSE)	$MSE = \frac{SSE}{(K - 1)(H - 1)}$

Bước 4. Kiểm định giả thuyết

- Tính tiêu chuẩn kiểm định F (F thực nghiệm)

$F_1 = \frac{MSK}{MSE}$	Trong đó: MSK là phương sai giữa các nhóm (cột) MSE là phương sai phần dư F_1 dùng kiểm định cho yếu tố nguyên nhân thứ nhất

$F_2 = \frac{MSH}{MSE}$	Trong đó: MSH là phương sai giữa các khối (hàng) MSE là phương sai phần dư F_2 dùng kiểm định cho yếu tố nguyên nhân thứ hai

- Tìm F lý thuyết cho 2 yếu tố nguyên nhân.

- Yếu tố nguyên nhân thứ nhất: (F tiêu chuẩn = F (k-1; (k-1)(h-1), α) là giá trị giới hạn

tra từ bảng phân phối F với k-1 bậc tự do của phương sai ở tử số và (k-1)(h-1) bậc tự do của phương sai ở mẫu số với mức ý nghĩa α.

F lý thuyết có thể tra qua hàm FINV(α, k-1, (k-1)(h-1)) trong EXCEL.

- Yếu tố nguyên nhân thứ hai: (F tiêu chuẩn = F (h-1; (k-1)(h-1), α) là giá trị giới hạn tra từ bảng phân phối F với h-1 bậc tự do của phương sai ở tử số và (k-1)(h-1) bậc tự do của phương sai ở mẫu số với mức ý nghĩa α.

F lý thuyết có thể tra qua hàm FINV(α, h-1, (k-1)(h-1)) trong EXCEL.

- Nếu F_1 thực nghiệm > F_1 lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể nhóm (cột) không bằng nhau.

- Nếu F_2 thực nghiệm > F_2 lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể khối (hàng) không bằng nhau.

Bảng phân tích phương sai 2 yếu tố khi sử dụng máy tính (phần mềm EXCEL hoặc SPSS) tóm tắt như sau:

Bảng gốc bằng tiếng Anh

Source of variation	*Sum of squares(SS)*	*Degree of freedom(df)*	*Mean squares(MS)*	*F- ratio*
Rows	SSH	(h-1)	MSH	F_1
Columns	SSK	(k-1))	MSK	F_2
Error	SSE	(k-1))(h-1)	MSE	
Total	SST	(n-1)		

Bảng phân tích phương sai tổng quát dịch ra tiếng Việt – ANOVA

Nguồn biến động	*Tổng độ lệch bình phương (SS)*	*Bậc tự do (df)*	*Phương sai (MS)*	*F- Tỷ số*
Giữa các hàng	SSH	(h-1)	MSH	F_1
Giữa các cột	SSK	(k -1)	MSK	F_2
Phần dư	SSE	(k -1) (h-1)	MSE	
Tổng số	SST	(n-1)		

c) Ví dụ:

Có tài liệu về giá bán đậu tương của các tỉnh qua 2 năm như sau (đồng/kg)

Tỉnh	**2003**	**2004**
Sơn La	4440	4247,7
Hà Tây	4850	4294,3
Đắc Lắc	4400	4284,3
Đồng Nai	4500	4314,3

Yêu cầu: Sử dụng kết quả phân tích phương sai so sánh giá bán đậu tương qua 2 năm và giữa 4 tỉnh?

Giải: Sử dụng phân tích phương sai (ANOVA) 2 yếu tố lấy mẫu không lặp trong EXCEL cho kết quả sau:

ANOVA: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Sơn La	2	8687,7	4343,85	18489,645
Hà Tây	2	9144,3	4572,15	154401,245
Đắc Lắc	2	8684,3	4342,15	6693,245
Đồng Nai	2	8814,3	4407,15	17242,245
2003	4	18190,0	4547,50	42358,333
2004	4	17140,6	4285,15	778,89

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F thực nghiệm*	*P-value*	*F crit*
Rows	70240,34	3	23413,45	1,1871	0,4456	9,2766
Columns	137655	1	137655,04	6,9791	0,0775	10,1280

Error	59171,34	3	19723,78			
Total	**267066,7**	**7**				

Từ kết quả phân tích ANOVA ở bảng trên cho thấy:

- Xét theo hàng: So sánh giá bán đậu tương bình quân giữa các tỉnh với giả thuyết là

Ho: Giá bán trung bình đậu tương giữa các tỉnh không sai khác nhau; F thực nghiệm = 1,18; F lý thuyết = 9,27. Như vậy, F thực nghiệm < F lý thuyết, ta chấp nhận Ho với xác suất có ý nghĩa là 55, 44%.

- Xét theo cột: So sánh giá bán đậu tương bình quân giữa các năm với giả thuyết là

Ho: Giá bán trung bình đậu tương giữa các năm không sai khác nhau; F thực nghiệm = 6,97; F lý thuyết = 10,12. Như vậy, F thực nghiệm < F lý thuyết, ta chấp nhận Ho với xác suất có ý nghĩa là 92,25%.

CÂU HỎI THẢO LUẬN CHƯƠNG VI

1. Thế nào là kiểm định giả thuyết? Các bước tiến hành kiểm định giả thuyết? Cho ví dụ?
2. Phân tích phương sai là gì? Các bước tiến hành? Cho ví dụ trong ngành để áp dụng phân tích phương sai phân tích ảnh hưởng của 2 yếu tố nguyên nhân đến 1 yếu tố kết quả?