Global Advanced

MBA

THỐNG KÊ TRONG KINH DOANH

Chương trình Đào tạo Thạc sĩ

Quản trị Kinh doanh Quốc tế

TÀI LIỆU THAM KHẢO - LƯU HÀNH NỘI BỘ

8/2009

MỤC LỤC

CHƯƠNG 1

GIỚI THIỆU CHUNG VÀ THU THẬP DỮ LIỆU

1. Thống kê và hoạt động quản trị

Để quản trị hoạt động của doanh nghiệp vấn đề đầu tiên và hết sức quan trọng là phải có đầy đủ thông tin về mọi mặt kể cả thông tin sơ cấp và thứ cấp. Để hoạt động của doanh nghiệp ngày càng tốt hơn cần quan tâm đến ba yếu tố có mối liên hệ chặt chẽ với nhau, đó là: lý luận quản lý, phương pháp thống kê và các biện pháp hành động. Xuất phát từ những nhiệm vụ trong quản lý và đặc điểm tình hình cụ thể của doanh nghiệp cần xác định nhiệm vụ quản lý cụ thể là gì, nhằm vào mục tiêu nào.... Trên cơ sở đó xác định thông tin cần thu thập, các biện pháp thu thập dữ liệu, xử lý và phân tích các dữ liệu đó. Trên cơ sở các phân tích định lượng đó rút ra kết luận về sự tồn tại thực tế của hiện tượng, bản chất và tính quy luật đang tồn tại, từ đó đưa ra các biện pháp quản lý thích hợp với từng hiện tượng trong từng giai đoạn cụ thể...

Trước hết hãy làm quen với một số khái niệm thường dùng trong thống kê:

- ***Tổng thể thống kê*** *(population)*: là hiện tượng kinh tế xã hội số lớn, bao gồm các đơn vị (hoặc phần tử, hiện tượng) cần được quan sát và phân tích. Xác định tổng thể nhằm đưa ra giới hạn về phạm vi nghiên cứu cho người nghiên cứu.

  Căn cứ vào sự nhận biết các đơn vị trong tổng thể có thể phân biệt hai loại: tổng thể bộc lộ và tiềm ẩn. *Tổng thể bộc* lộ là tổng thể có ranh giới rõ ràng, có thể nhận biết hết các đơn vị trong tổng thể (chẳng hạn khi nghiên cứu tình hình sản xuất công nghiệp trên địa bàn Hà Nội ta có tổng thể các doanh nghiệp sản xuất công nghiệp trên địa bàn Hà Nội và đó là tổng thể bộc lộ). *Tổng thể tiềm ẩn* là tổng thể có ranh giới không rõ ràng, hay không thể nhận biết hết các đơn vị trong tổng thể (chẳng hạn tổng thể những người ưa dùng một loại sản phẩm nào đó, hoặc tổng thể những người sẽ sử dụng dịch vụ của Hãng hàng không quốc gia Việt Nam trong năm tới).

  Căn cứ vào mục đích nghiên cứu người ta phân biệt hai loại: tổng thể đồng chất và không đồng chất. *Tổng thể đồng chất* bao gồm các đơn vị giống nhau (hoặc gần giống nhau) về một số đặc điểm chủ yếu có liên quan đến mục đích nghiên cứu. *Tổng thể không đồng chất* bao gồm các đơn vị có những đặc điểm chủ yếu khác nhau.

  Căn cứ vào phạm vi nghiên cứu có thể phân biệt hai loại: Tổng thể chung (bao gồm tất cả các đơn vị thuộc đối tượng nghiên cứu) và tổng thể bộ phận (là một phần của tổng thể chung).

- ***Đơn vị tổng thể***: là các đơn vị (hoặc phần tử) cấu thành nên tổng thể. đơn vị tổng thể là xuất phất điểm của việc nghiên cứu, bởi vì mặt lượng của đơn vị tổng thể là các dữ liệu mà người nghiên cứu cần thu thập.

- ***Tổng thể mẫu*** *(Sample)*: là tổng thể bao gồm một số đơn vị nhất định được chọn ra từ tổng thể chung để tiến hành điều tra thực tế. Các đơn vị này được chọn theo một phương pháp lấy mẫu nào đó.

- ***Quan sát*** *(Observation)*: là cơ sở để thu thập dữ liệu nghiên cứu. Thí dụ trong điều tra chọn mẫu, mỗi đơn vị thuộc mẫu được tiến hành thu thập thông tin được gọi là một quan sát

- ***Tiêu thức thống kê:*** là đặc điểm của từng đơn vị tổng thể được chọn ra để nghiên cứu tuỳ theo mục đích nghiên cứu khác nhau. Có thể phân biệt các loại tiêu thức sau :

  Tiêu thức thực thể : Là loại tiêu thức phản ánh đặc điểm về nội dung của đơn vị tổng thể. Tùy theo cách biểu hiện mà có 3 loại:

  - *Tiêu thức thuộc tính* là tiêu thức phản ánh các thuộc tính của đơn vị tổng thể và không có các biểu hiện trực tiếp bằng con số. Thí dụ: tiêu thức giới tính, nghề nghiệp, dân tộc, thành phần kinh tế....

  - *Tiêu thức số lượng* là tiêu thức phản ánh các đặc điểm về lượng của đơn vị tổng thể và có các biểu hiện trực tiếp bằng con số, các biểu hiện con số của tiêu thức số lượng được gọi là các *lượng biến*. Thí dụ: Số nhân khẩu trong gia đình, tiền lương tháng của mỗi người lao động, Năng suất lao động... Các lượng biến là cơ sở để thực hiện các phép tính thống kê, như: cộng, trừ, nhân, chia, trung bình, tỷ lệ...

  - Tiêu thức chỉ có hai biểu hiện không trùng nhau trên một đơn vị tổng thể được gọi là *tiêu thức thay phiên*. Thí dụ: tiêu thức giới tính chỉ có hai biểu hiện không trùng nhau là *nam* và *nữ* được gọi là tiêu thức thay phiên. Loại tiêu thức này có đặc điểm quan trọng là nếu một đơn vị tổng thể nào đó đã nhận biểu hiện này thì không nhận biểu hiện kia. Đây là loại tiêu thức có nhiều ứng dụng trong thực tế.

- ***Chỉ tiêu thống kê:*** phản ánh đặc điểm của toàn bộ tổng thể trong điều kiện thời gian và địa điểm cụ thể. Chỉ tiêu thống kê là tổng hợp biểu hiện mặt lượng của nhiều đơn vị, hiện tượng cá biệt. Do đó, chỉ tiêu phản ánh những mối quan hệ chung, đặc điểm của số lớn các đơn vị hoặc của tất cả các đơn vị tổng thể. Chỉ tiêu thống kê bao gồm 2 mặt : khái niệm và mức độ của chỉ tiêu. Mặt khái niệm của chỉ tiêu bao gồm các định nghĩa và giới hạn về thực thể, thời gian và không gian; Mức độ của chỉ tiêu là các trị số với các đơn vị tính phù hợp.

  Có hai loại chỉ tiêu thống kê: *chỉ tiêu khối lượng* : biểu hiện quy mô, khối lượng của hiện tượng nghiên cứu và *chỉ tiêu chất lượng* : biểu hiện trình độ phổ biến và mối quan hệ trong tổng thể.

2. Phương pháp thống kê.

Các phương pháp thống kê gồm: Thống kê mô tả và thống kê suy luận (phân tích thống kê).

- *Thống kê mô tả:* Bao gồm thu thập và mô tả dữ liệu. Cụ thể:
  i. Thu thập dữ liệu qua điều tra.
  ii. Biểu diễn dữ liệu bằng bảng và đồ thị thống kê.
  iii. Tính toán các đặc trưng của dữ liệu như trung bình, trung vị... Đặc trưng của tổng thể chung gọi là tham số; đặc trưng của tổng thể mẫu gọi là thống kê.
- *Thống kê suy luận*: bao gồm các phương pháp ước lượng, kiểm định giả thiết thống kê ... Tiêu thức số lượng có từ đó đưa ra các quyết định về tổng thể chung trên cơ sở kết quả từ mẫu điều tra.

3. Các loại dữ liệu và nguồn dữ liệu.

- Có hai loại dữ liệu:

  - Dữ liệu định tính: Là dữ liệu về các tiêu thức thuộc tính (có thể có biểu hiện trực tiếp hoặc gián tiếp).

  - Dữ liệu định lượng: Là dữ liệu về các tiêu thức số lượng. Trong đó các lượng biến có thể là rời rạc (biểu hiện bằng các số nguyên) hoặc liên tục (biểu hiện bằng số thập phân).

- Nguồn dữ liệu: Có hai nguồn dữ liệu chính

  - Nguồn thứ cấp: là những tài liệu đã được thu thập từ trước. Nguồn này bao gồm các tài liệu đã xuất bản (tạp chí, sách báo, kỷ yếu hội thảo, thông tin từ Internet... Nguồn số liệu thứ cấp có thể có do các cơ quan của chính phủ thu thập (Tổng cục thống kê, các bộ..) hoặc từ các nguồn do các tổ chức tư nhân, các tổ chức phi chính phủ... thu thập. Một nguồn số liệu thứ cấp nữa chưa phổ biến ở Việt nam là nguồn thông tin do các công ty chuyên nghiên cứu thị trường thu thập và bán.

  - Nguồn sơ cấp: là nguồn dữ liệu được thu thập lần đầu tiên phục vụ cho mục đích nghiên cứu. Nguồn này bao gồm các tài liệu được thu thập từ các cuộc điều tra, các quan sát, các nghiên cứu hiện trường (thực nghiệm).

4. Các phương pháp chọn mẫu.

Thông thường có hai loại phương pháp chọn mẫu: *Chọn mẫu xác xuất* (các đơn vị được chọn theo xác suất đã biết) và *chọn mẫu phi xác xuất* (chọn mẫu có chủ đích, chọn mẫu định mức, chọn mẫu đoạn). Trên thực tế thường sử dụng phương pháp chọn mẫu xác suất với các phương pháp chọn cụ thể như sau:

- *Mẫu ngẫu nhiên đơn giản*: mỗi đơn vị có cơ hội được chọn vào mẫu như nhau (như gieo con xúc sắc, bắt thăm, quay số). Có thể chọn hoàn lại hoặc không hoàn lại. *Chọn hoàn lại* là một đơn vị được chọn rồi sau khi nghiên cứu lại trả về tổng thể chung và có cơ hội được chọn lại; *chọn không hoàn lại* là một đơn vị được chọn rồi sau khi nghiên cứu không được trả về tổng thể chung và không có khả năng được chọn lại. NgoàI ra còn có thể sử dụng bảng số ngẫu nhiên để chọn mẫu.

- *Mẫu hệ thống*: Trước hết cần sắp xếp các đơn vị của tổng thể chung theo một thứ tự nào đó, như sắp xếp theo thứ tự vần A,B,C ... của tên gọi , theo thứ tự địa dư , theo quy mô từ nhỏ đến lớn v.v... Các đơn vị được lựa chọn từ tổng thể chung theo khoảng cách bằng nhau (k). Khoảng cách k này được tính bằng cách chia số đơn vị của tổng thể chung cho số đơn vị của tổng thể mẫu, (hay nói cách khác tổng thể chung được chia thành k nhóm). Chọn một cách ngẫu nhiên một đơn vị từ nhóm thứ nhất, sau đó cứ mỗi khoảng cách k lại chọn một đơn vị. Thí dụ: Tổng thể chung N= 80 đơn vị, chọn mẫu n=8 đơn vị, vậy k = 80/8 = 10. Nếu trong nhóm đầu tiên ta chọn đơn vị thứ 3 thì các đơn vị được chọn tiếp theo sẽ là đơn vị thứ 13, 23, 33, 43, 53 ...
- *Mẫu phân loại*: Tổng thể chung được phân chia thành hai hoặc nhiều nhóm trên cơ sở một số đặc đIểm chung (các tổ có độ thuần nhất cao). Sau đó chọn các đơn vị đại diện cho từng tổ theo cách chọn ngẫu nhiên đơn giản hoặc máy móc. Số đơn vị được chọn từ mỗi tổ có thể tương ứng với tỷ trọng của tổ đó trong tổng thể chung - gọi là chọn phân loại theo tỷ lệ - hoặc có thể không tương ứng với tỷ trọng đó - gọi là chọn phân loại không theo tỷ lệ. Nếu chọn mẫu phân loại theo tỷ lệ ta được mẫu có kết cấu gần giống với kết cấu của tổng thể chung nên tính đại biểu cao, muốn tính đại biểu của mẫu cao hơn nữa có thể lấy mẫu tối ưu tức là số đơn vị chọn ra ở mỗi tổ không chỉ tỷ lệ với tỷ trọng của tổ đó chiếm trong tổng thể mà còn tương ứng với độ biến thiên ở mỗi tổ theo tiêu thức nghiên cứu.
- *Mẫu chùm* (cluster): Tổng thể chung được chia thành các chùm, mỗi chùm đều đại diện cho tổng thể chung. Sau đó ở mỗi chùm có thể chọn ra một số đơn vị theo cách chọn ngẫu nhiên đơn giản để hợp thành một mẫu. Hoặc có thể tiến hành chọn cả một hoặc một vài chùm nào đó một cách ngẫu ngẫu nhiên nếu số chùm lớn.

5. Phương pháp điều tra.

5.1. Thang đo

5.1.1. Các loại thang đo

- <u>*Thang đo định danh*</u>:

- Là đánh số các biểu hiện cùng loại của tiêu thức.

- Thường dùng với các tiêu thức thuộc tính mà các biểu hiện của nó là một hệ thống các loại khác nhau không theo một trật tự xác định nào như: Giới tính, khu vực địa lý, nghề nghiệp, tôn giáo...

Thí dụ: Với tiêu thức giới tính chỉ có hai loại nam và nữ và không có trật tự nào giữa hai loại này; vì vậy có thể đánh số các biểu hiện Nam là 1 và nữ là 2 hoặc ngược lại.

- Đặc điểm: Các con số không có quan hệ hơn kém, và không được thực hiện tất cả các phép tính, chỉ dùng để mã hóa và đếm tần số.

- <u>*Thang đo thứ bậc*</u>:

- Là thang đo định danh nhưng giữa các biểu hiện của tiêu thức có quan hệ hơn kém.

Thí dụ với tiêu thức *tầng lớp xã hội*, thang đo thứ bậc đơn giản nhất có 3 nấc: "thượng lưu", "trung lưu", "hạ lưu", những loại này lần lượt được đánh số từ 1 đến 3; những số này thể hiện trật tự của các tầng lớp xã hội nhưng nó không phản ánh được khoảng cách giữa các con số. Nhưng cũng có thể dùng thang đo 9 nấc phức tạp hơn để đo bằng cách chia nhỏ mỗi loại trên thành 3 loại: "Thượng lưu bậc cao - Thượng lưu - Thượng lưu bậc thấp" ... Tóm lại, một tiêu thức có bao nhiêu biểu hiện phụ thuộc vào mức độ thay đổi mà chúng ta hy vọng có thể tìm ra trong quá trình điều tra và mức độ chi tiết theo yêu cầu nghiên cứu.

Chú ý: trong một số trường hợp thang đo thứ bậc có thể kết hợp với thang đo định danh để hiểu rõ được khái niệm hơn. Chẳng hạn, để đánh số các biểu hiện của tiêu thức tôn giáo có thể dùng thang đo định danh (1- phật giáo, 2 - thiên chúa giáo, 3 - hồi giáo, 4 - Đạo hin đu, 5 - Đạo cao đài...). Nhưng để biểu hiện hành vi tôn giáo thì có thể dùng thang đo thứ bậc. Chẳng hạn, bạn có dự hoạt động tôn giáo không: 1- hàng ngày, 2- hàng tuần, 3- một vài tuần trong một tháng, 4- hàng tháng, 5- một vài lần trong năm, 6- hàng năm, 7- không bao giờ.

***- Thường dùng** để đo các tiêu thức thuộc tính mà các biểu hiện có quan hệ thứ tự* như đo thái độ đối với một hành vi nào đó (hoàn toàn đồng ý, đồng ý, chưa quyết định, hoàn toàn không đồng ý) hoặc thứ tự chất lượng sản phẩm, huân chương, bậc thợ...

- Đặc điểm: sự chênh lệch giữa các biểu hiện của tiêu thức không nhất thiết phải bằng nhau.

- Hạn chế: Chưa biết được khoảng cách giữa các số thự tự đó gần hay xa bao nhiêu vì vậy không thực hiện được các phép tính cộng trừ nhân chia mà chỉ nói lên đặc trưng chung của tổng thể một cách tương đối căn cứ trên sự giải thích "lớn hơn" hay "nhỏ hơn" mà thôi.

- *Thang đo khoảng*:

- Là thang đo thứ bậc có các khoảng cách đều nhau nhưng không có điểm gốc là 0.

Thí dụ: Sau khi hỏi ý kiến của Ban giám khảo về việc xếp loại các đội trong cuộc thi..., có thể tiếp tục hỏi thêm đội này hơn đội kia bao nhiêu điểm. Như vậy khoảng cách giữa các thứ tự đã được lượng hoá và có thể giải thích được.

- Đặc điểm: Thang đo này có thể thực hiện các phép tính cộng, trừ, tính được các tham số đặc trưng như trung bình, phương sai, tỷ lệ.

- Hạn chế: Là thang đo có khoảng cách hơn kém nhưng chưa có điểm gốc là số 0, nên không so sánh được tỷ lệ giữa các trị số đo.

- *Thang đo tỉ lệ* :

- Là thang đo khoảng với một điểm 0 tuyệt đối (điểm gốc) để có thể so sánh được tỷ lệ giữa các trị số đo. Thí dụ; các đơn vị đo lường vật lý thông thường (kg, mét..), thu nhập, tuổi, số con ...

- Đặc điểm: Thang đo này cho phép thực hiện được tất cả các phép tính với trị số đo.

Theo tuần tự, thang đo sau có chất lượng đo lường cao hơn thang đo trước, đồng thời việc xây dựng thang đo cũng phức tạp hơn. Song không phải lúc nào cũng có thể sử dụng được thang đo hoàn hảo mà phải tuỳ thuộc vào đặc điểm của hiện tượng và tiêu thức nghiên cứu mà sử dụng thang đo cho thích hợp.

5.1.2. Một số cách đặt thang điểm cơ bản

5.1.2.1. Thang điểm điều mục:

Đây là loại thang điểm đơn giản và phù hợp với nhiều hoàn cảnh khác nhau. *Loại này đòi hỏi người được phỏng vấn cho biết thái độ của họ và điều mục đánh giá mà họ lựa chọn*. Các mục được sắp xếp theo một thứ tự nào đó. Loại thang đo này còn được gọi là thang đo Likert.

Ví dụ: Với câu hỏi "Bạn có thoả mãn với công việc mà bạn đang làm hiện nay không" có thể đặt các thang điểm sau:

- Rất thoả mãn.
- Tương đối thoả mãn.
- Không quan tâm.
- Không được thoả mãn lắm.
- Rất khó chịu.

Tuy thang điểm này đơn giản nhưng cần phải quan tâm đến một số vấn đề sau:

+ *Số lượng điều mục*: Cần có sự quyết định số mục lựa chọn tượng trưng cho thái độ của người được phỏng vấn. Chẳng hạn, *thang mục chỉ có 2 mục* đối nhau (đồng ý hay không đồng ý) mang tính chất của thang đo định danh rất khó cho công việc phân tích nhưng có thể thích hợp khi bảng câu hỏi dài hoặc khi trình độ văn hoá của người được hỏi có giới hạn. Mặt khác có thể *sử dụng nhiều điều mục* để giúp cho người được hỏi có nhiều sự lựa chọn rộng rãi và cho phép đo độ nhạy bén hơn.(một số nhà nghiên cứu cho rằng câu hỏi có 5 hoặc 6 mục trả lời là phù hợp hơn cả)

+ *Số điều mục trả lời không nên lẻ* để tránh dẫn đến việc người trả lời có thái độ trung dung với cách chọn câu trả lời ở giữa, và dễ đưa đến câu trả lời không đúng sự thật. Số điều mục chẵn thì người được hỏi bắt buộc phải biểu lộ thái độ của mình.

+ *Không nên đặt câu trả lời lệch về một phía* này hay phía kia làm cho người trả lời khó chọn. Ví dụ: với câu hỏi "đề nghị bạn cho biết tốc độ mau lẹ trong cung cách phục vụ khách hàng" mà sử dụng các điều mục : tuyệt, rất tốt, tốt, trên trung bình, trung bình thì sẽ rất không thích hợp cho người không thích cung cách phục vụ ấy.

5.1.2.2. Thang điểm đánh giá qua hình vẽ:

Thang điểm này sử dụng các hình vẽ để thể hiện thái độ của người được phỏng vấn về một vấn đề nào đó.

Người ta thường sử dụng các loại thang điểm như: thang "hình nhiệt kế" hoặc là các "vẻ mặt" khác nhau nói lên thái độ đồng tình hay không đồng tình về một vấn đề nào đó.

Với loại thang điểm này có thể phân chia nhiều mức khác nhau tuỳ thuộc nội dung và quy mô hiện tượng nghiên cứu.

5.1.2.3. Thang điểm xếp hạng theo thứ tự :

Loại thang điểm này cho phép so sánh các điều mục trả lời trong khi 2 thang điểm trước không thể so sánh được vì ở hai loại thang điểm trên người được hỏi xét đoán không dựa vào căn cứ cụ thể nào cả.

Đối với loại thang điểm xếp hạng theo thứ tự, người được hỏi sắp xếp hạng các mục trả lời theo thứ tự mà họ đánh giá. Ví dụ: Để nâng cao chất lượng học tập của sinh viên, nhà trường đã sử dụng các biện pháp sau đây:

- Điểm danh thường xuyên ở lớp.
- Kiểm tra bài thường xuyên.
- Quy chế thi nghiêm túc.
- Học bổng có nhiều mức theo theo kết quả thi từng học kỳ.
- Cho nhiều chuyên đề nghiên cứu và bài tập lớn.
- Nộp chi phí cao khi phải thi lại.

Hãy xếp thứ tự các biện pháp trên từ phương pháp hiệu quả nhất theo thứ tự từ 1 đến 6.

Loại thang điểm này tuy đơn giản, dễ trả lời song cũng có những nhược điểm là:

- Khó liệt kê được đầy đủ các trường hợp nên dữ liệu thu thập thiếu chính xác.
- Vì nhấn mạnh vào việc xếp thứ tự nên có thể ảnh hưởng đến câu trả lời, đặc biệt là mục thứ nhất và mục chót thường được quan tâm nhiều hơn.
- Khi hỏi để xếp hạng những mục hoàn toàn nằm ngoài ý thích của người được hỏi thì những câu trả lời sẽ không có ý nghĩa
- Thang điểm này không giúp ta xác định được khoảng cách xa gần giữa các mục là bao nhiêu theo sự nhận định của người được hỏi, hoặc là tại sao các mục lại được sắp xếp như vậy.

5.1.2.4. Thang điểm có tổng không đổi:

Thang điểm có tổng không đổi cung cấp một sự nhận thức tổng quát tốt hơn về khoảng cách giữa các điểm trên giải thang điểm.

Đối với thang điểm này, người được hỏi cần chia hoặc xác định một số điểm có tổng không đổi (thường là 100) để biểu thị sự quan trọng tương đối của những đặc điểm được

nghiên cứu. Số điểm được chia cho mỗi điều mục chỉ rõ hạng bậc của nó và đồng thời cũng chỉ rõ số khác biệt giữa các điều mục với nhau. Ví dụ: Cũng vẫn với ví dụ trên, chia 100 điểm cho các biện pháp theo tầm quan trọng của mỗi biện pháp.

Thang điểm này còn một số tồn tại sau:

- Không thể chắc chắn là liệu những kết quả có biểu thị đúng với khoảng cách và tỷ lệ hay không.
- Nếu có quá nhiều đặc điểm thì việc chia điểm cũng gặp khó khăn

5.1.2.5. Thang điểm có ý nghĩa đối nghịch nhau:

Thang điểm này yêu cầu người được hỏi cho biết cảm nghĩ về vấn đề cần được nghiên cứu bằng cách ghi ý kiến trả lời trên một chuỗi tính từ tạo thành từng cặp đối nghịch nhau về ý nghĩa.

Ví dụ: Để hỏi ý kiến đánh giá về một loại bia ta dùng thang điểm 7 vị trí có ý nghĩa đối nghịch nhau như sau:

Cực X								***Cực Y***
(nặng)	1	2	3	4	5	6	7	***(nhẹ)***
	Rất	Khá	Hơi	T/Bình	Hơi	Khá	Rất	

Có nhiều ý kiến khác nhau về số hạng mục đánh giá nên để chẵn hay lẻ. Nếu để lẻ dễ dẫn đếnviệc chọn vị trí giữa vì "vô thưởng vô phạt". Có thể để thang điểm có số hạng mục chẵn.

Cực X							***Cực Y***
(tốt)	1	2	3	4	5	6	***(kém)***
	Rất	Khá	ít	ít	Khá	Rất	

Ngoài các thang điểm cơ bản trên còn có nhiều cách đặt thang điểm nữa, tuỳ theo kỹ thuật của các nhà nghiên cứu. Tuy vậy mỗi thang điểm đều có ưu nhược điểm riêng. Vì vậy, người nghiên cứu phải biết lựa chọn loại thang điểm nào thích hợp nhất, có khả năng đáp ứng tốt nhất những nhu cầu thông tin với chi phí thấp nhất, phương pháp truyền đạt dễ dàng, dễ hiểu, dễ trả lời.

5.2. Các phương pháp điều tra

5.2.1. Phương pháp phỏng vấn

Là phương pháp được sử dụng nhiều nhất trong thu thập thông tin. Căn cứ vào nội dung điều kiện thực tế, người nghiên cứu quyết định dùng phương pháp nào để tiếp xúc với người được phỏng vấn. Có các phương pháp phỏng vấn như sau:

5.2.1.1. Phương pháp Anket:

Phương pháp anket là phương pháp phỏng vấn viết (người hỏi vắng mặt), trong đó sự tiếp xúc thông qua bảng hỏi, người trả lời tự điền câu trả lời vào bảng hỏi. Vì vậy những vấn đề tâm lý trong khi đặt câu hỏi và nguyên tắc tâm lý trong sắp xếp bảng hỏi là để hướng vào

người trả lời. Để thực hiện phương pháp này cần lưu ý đến 2 vấn đề quan trọng nhất là: Thiết kế bảng hỏi và cách phân phát bảng hỏi. Hai vấn đề đó sẽ ảnh hưởng đến chất lượng thông tin, tỷ lệ trả lời và gửi lại bảng hỏi, những yêu cầu quan trọng trong phương pháp anket.

5.2.1.2. Phỏng vấn trực diện

Phương pháp phỏng vấn trực diện thông thường được hiểu là phỏng vấn miệng, còn gọi là "cuộc nói chuyện riêng" hay "trò chuyện có chủ định". Nói chuyện thông thường là cơ sở của phỏng vấn, nó khác với cuộc nói chuyện thông thường ở hai điểm sau đây:

- Mục đích cuộc nói chuyện do chương trình nghiên cứu quy định từ trước.
- Vai trò của người nói chuyện được quy định, thậm chí được "chuẩn hoá"

Cũng vì vậy mà người ta gọi nó là "cuộc tiếp xúc giả tạo" do nguyên nhân từ bên ngoài. Kết quả của phỏng vấn, chất lượng thông tin thu được phần lớn phụ thuộc vào tính chất của việc tiếp xúc, sự giao tiếp chặt chẽ và hiểu biết lẫn nhau giữa người phỏng vấn và người trả lời hay nói cách khác nó mang dấu vết của cuộc tiếp xúc đó.

) Ưu điểm của phỏng vấn trực diện.

+ Việc tiếp xúc trực tiếp tạo ra những điều kiện đặc biệt để hiểu đối tượng sâu sắc hơn. Đây là ưu điểm mà phương pháp anket không thể có được.

+ Do tiếp xúc trực tiếp nên đã đồng thời kết hợp việc phỏng vấn với việc quan sát (từ dáng vẻ bề ngoài, đến những cử chỉ biểu lộ tình cảm, thái độ...).

+ Có thể phát hiện những sai sót và uốn nắn kịp thời ngay.

) Nhược điểm:

+ Tốn kém hơn so với phương pháp anket: tốn hơn về thời gian, chi phí, số cán bộ điều tra.

+ Tổ chức khó khăn hơn: đòi hỏi phải có sự chuẩn bị kỹ càng về điều tra viên, địa điểm, nghi thức gặp gỡ...

+ Không cẩn thận câu trả lời sẽ bị ảnh hưởng bởi thiên kiến của điều tra viên.

)Tính chất của cuộc phỏng vấn trực diện:

+ *Tính một chiều*: Đó là quá trình giao tiếp một chiều do người phỏng vấn điều khiển. Người phỏng vấn phải làm chủ cả quá trình phỏng vấn từ khi mở đầu đến lúc kết thúc. Do tính một chiều và làm chủ đó đòi hỏi người phỏng vấn phải tạo được không khí cởi mở, dễ dàng thổ lộ cho người trả lời.

+ *Tính quy định*: Nội dung và các khả năng xử sự trong cuộc nói chuyện được quy định sẵn trong bảng câu hỏi và kế hoạch phỏng vấn.

+ *Tính giả định*: Nhiều đề tài đặt ra các yêu cầu và tình huống giả định để thu lại những phản ứng khác nhau của người được phỏng vấn.

+ *Tính phi hậu quả:* Cuộc phỏng vấn phải đảm bảo không gây hậu quả cho người được phỏng vấn. Mức phi hậu quả bắt nguồn từ hai lý do:

Thứ nhất là tính giả định của cuộc phỏng vấn. Khi tình huống khác đi, việc nói lại là không nên hoặc không có nghĩa.

Thứ hai là nguyên tắc nặc danh. Yêu cầu giữ bí mật cho người được phỏng vấn, tránh những phiền toái, truy cứu những người trả lới ý này hay ý nọ.

) Các loại phỏng vấn trực diện:

Tuỳ theo mức độ nghiêm ngặt của cách thức tiến hành, có thể chia thành hai loại:

Theo nội dung trình tự tiến hành:

+ *Phỏng vấn tiêu chuẩn hoá: là cuộc phỏng vấn diễn ra theo trình tự với nội dung được vạch sẵn cho mọi người dựa vào một bảng hỏi*. Người phỏng vấn không được thay đổi trình tự các câu hỏi, không có quyền đưa thêm câu hỏi bổ sung hoặc gợi ý thêm các phương án trả lời đã có sẵn trong bảng hỏi. Có thể nói phỏng vấn tiêu chuẩn hoá là phương pháp phỏng vấn theo anket. Hình thức này có *ưu điểm là số liệu có thể so sánh trực tiếp được với nhau, phục vụ việc tổng hợp dễ dàng, phù hợp với việc kiểm định giả thuyết.*

+ *Phỏng vấn bán tiêu chuẩn: là hình thức trung gian giữa phỏng vấn tiêu chuẩn hoá và phỏng vấn tự do.* Cụ thể là ở đây các câu hỏi quyết định được tiêu chuẩn hoá, còn các câu hỏi khác thì có thể phát biểu tuỳ tình hình thực tế. Như vậy sẽ tận dụng được ưu điểm và hạn chế được nhược điểm của mỗi loại.

+ *Phỏng vấn tự do*: *là cuộc phỏng vấn không có những câu hỏi đã định trước và cũng không theo kế hoạch đã định trước, chỉ đưa ra đề tài, người phỏng vấn hoàn toàn tự do tiến hành như một cuộc nói chuyện tự do*. *Ưu điểm* của phương pháp này là người trả lời được tự do tư tưởng, thoải mái trình bày ý kiến quan điểm của mình sâu rộng tuỳ ý, người phỏng vấn chủ động thực hiện mục đích của mình, không bị gò bó. *Khó khăn* là người phỏng vấn phải có trình độ cao, biết duy trì, dẫn dắt câu chuyện đến đích.

+ *Phỏng vấn sâu* (indepth-interview): Khác với phỏng vấn tự do ở chỗ là *ngoài những đề tài nói chuyện chung người ta còn đặt ra trước một số câu hỏi hoặc vấn đề nhất định mà cần phải được trả lời*. *Đặc điểm* của phỏng vấn sâu là không cần nhiều đối tượng điều tra, thậm chí có khi chỉ cần một số ít người để hỏi về những vấn đề sâu kín, tiềm ẩn mà không phải ai trong họ cũng có thể cảm nhận hoặc nói ra. *Khó khăn* ở đây cũng là đòi hỏi người phỏng vấn có trình độ cao, có phương pháp tâm lý và biết cách phân tích tâm lý.

+ *Phỏng vấn định hướng*: là cuộc phỏng vấn dặt mục đích nghiên cứu rõ ràng, những ý kiến về tình hình đã được nêu ra một cách cụ thể; nói cách khác là phỏng vấn tập trung vào một mục tiêu (focused interview).

Theo đối tượng tiếp xúc:

+ *Phỏng vấn cá nhân*: có thể là tất cả các loại phỏng vấn tiêu chuẩn, bán tiêu chuẩn, tự do, phỏng vấn sâu, phỏng vấn định hướng.

+ *Phỏng vấn nhóm*: chỉ thường áp dụng phỏng vấn nhóm tiêu chuẩn (cơ cấu tiêu chuẩn, đồng nhất) và phỏng vấn nhóm tự do (không đồng nhất) trong khi vẫn tuân thủ nội dung phỏng vấn tiêu chuẩn hoặc tự do đã nói trên.

Trên đây là các loại phỏng vấn trực diện khác nhau, việc lựa chọn loại hình nào là tuỳ thuộc mục đích điều tra, số lượng người cần điều tra, số lượng và chất lượng người phỏng vấn, khả năng tài chính cho phép ...

5.2.1.3. Phỏng vấn qua điện thoại

Phỏng vấn qua điện thoại chỉ là phỏng vấn miệng với cá nhân, nhưng người phỏng vấn và người được phỏng vấn không gặp mặt trực tiếp mà thông qua điện thoại.

) Ưu điểm của phỏng vấn qua điện thoại: ngày nay người ta áp dụng phổ biến với những chủ đề rộng rãi hơn và cũng thường là kết hợp với các phương pháp khác.

+ *Tiết kiệm chi phí hơn*: Phỏng vấn qua điện thoại rẻ hơn nhiều so với các cuộc phỏng vấn trực diện (Trong một nghiên cứu so sánh kỹ càng của Trung tâm điều tra khảo sát thuộc trường đại học tổng hợp Michigan cho thấy: một cuộc điều tra 1500 cá nhân, theo cách phỏng vấn trực diện tốn khoảng 84000 USD, theo cách phỏng vấn qua điện thoại tốn 38000 USD. Điều này có nghĩa là mỗi cuộc phỏng vấn trực diện tốn khoảng 55 USD trong khi phỏng vấn qua điện thoại là 23 USD). *Đào tạo, tập huấn chuẩn bị cho nhân viên điều tra trong điều tra trực diện đắt gấp 2 lần so với điều tra qua điện thoại. Giá cho việc đi lại của nhân viên điều tra chiếm gần 20% trong nghiên cứu trực diện trong khi đó không mất phí tổn này trong điều tra qua điện thoại.*

+ *Tiết kiệm thời gian*: Ngồi trong văn phòng gọi điện thoại cho người được phỏng vấn tiết kiệm thời gian và công sức hơn phải ra ngoài phỏng vấn trực diện.

+ *Điều tra qua điện thoại khách quan hơn*: Vì người phỏng vấn không thể thấy người trả lời nên người trả lời có thể sẵn lòng tiết lộ các thông tin riêng tư hơn là phỏng vấn trực diện.

) Nhược điểm của phỏng vấn qua điện thoại: có 3 điều bất lợi đối với phỏng vấn qua điện thoại.

+ *Mất nhiều công sức chọn số điện thoại*: trong số những số điện thoại được chọn ngẫu nhiên chỉ có một số là thành công trong việc vì nhiều lý do khác nhau (không đúng đối tượng, không gặp, không trả lời...)

+ *Giảm hứng thú khi phỏng vấn qua điện thoại:* Khi phỏng vấn qua điện thoại, *người trả lời* dễ dàng trả lời cuộc phỏng vấn hơn so với khi người phỏng vấn đang đứng hay ngồi cạnh họ. Nhưng vì không nhìn thấy nhau nên người phỏng vấn khó định lượng mức độ quan tâm của người trả lời hơn. Do vậy người phỏng vấn cần phải cố giữ cuộc nói chuyện qua điện thoại trôi chảy sao cho người trả lời không có thời gian cân nhắc xem liệu họ có bận quá hay chán quá. Điều này có nghĩa là thu được những thông tin sâu từ câu hỏi mở trong một điều tra qua điện thoại là rất khó.

+ Việc đưa ra các gợi ý hay hỗ trợ thêm bằng quan sát là khó có thể thực hiện được. Chẳng hạn, không thể gợi ý bằng cách đưa ra một danh sách để lựa chọn hay khi cần những phản ứng đối với hình ảnh...

*5.2.1.4. Phỏng vấn qua thư (*kể cả thư điện tử)

Đây là 1 trường hợp đặc biệt của phương pháp anket, trong đó bảng hỏi không được phân phát tận nơi người trả lời mà gửi qua bưu điện hoặc theo dạng thư ngỏ trên mang internet. Phương pháp này đỡ tốn kém hơn nhưng tỷ lệ trả lời rất thấp.

Nếu dùng hình thức gửi thư qua bưu điện, Cơ quan bưu điện có thể giao được bảng câu hỏi đến mọi chỗ có địa chỉ rõ ràng - khả năng nhận được của người trả lời cao. Người trả lời có thể trả lời thoải mái, có suy nghĩ với những câu hỏi dứt khoát hoặc họ cũng trả lời tốt với những câu hỏi cần xem thêm những lưu trữ (của họ) cho thêm chính xác. Việc nặc danh, làm cho người trả lời có thể thổ lộ những vấn đề riêng tư không bao giờ nói với người phỏng vấn. Phí tổn chỉ phải chi trả cho việc ghi địa chỉ, in ấn, cước phí và điều hành. Nhưng phương pháp này cũng có những hạn chế đáng kể, nhất là những bảng trả lời được gửi về nhỏ giọt và nhiều địa chỉ không còn đúng nữa….

5.2.2. Phương pháp quan sát

Là phương pháp thu thập thông tin được chuẩn bị một cách khoa học theo những mục đích đã vạch ra dựa trên việc quan sát đối tượng điều tra. Các phương pháp quan sát :

- *Quan sát lộ diện*: người điều tra trực tiếp tham gia vào quá trình hoạt động của đối tượng được quan sát. Người quan sát có thể đóng vai trò là trung lập, đứng ngoài cuộc để quan sát hoặc có thể chủ động tham gia tích cực cùng với người được quan sát.
- *Quan sát giấu mặt*: thông qua camera, máy ghi âm
- *Quan sát có kết cấu*: việc quan sát được tiến hành dựa trên các tiêu thức được chuẩn bị từ trước.
- *Quan sát không có kết cấu*: là quan sát một cách tự do và người quan sát ghi chép lại tất cả những điều quan sát được.

5.2.3. Phương pháp thực nghiệm

- Được sử dụng nhằm kiểm tra một số nhận định sơ bộ nào đó. Là phương pháp thu thập thông tin được chuẩn bị một cách khoa học theo những mục đích đã vạch ra, có cách quan sát riêng về từng đối tượng và có sự chọn lọc thông tin.
- Nội dung: người điều tra tạo ra một tình huống gần giống với tình huống thực tế rồi quan sát cách ứng xử của những người được điều tra
- Ưu điểm : cho phép ta xác định được hành động của người được quan sát.

- Nhược điểm: Chỉ thấy biểu hiện của hành động mà khó xác định được mục đích, nguyên nhân dẫn đến hành động đó là gì, có nghĩa là khó đi sâu vào bản chất của nó nếu không sử dụng các phương pháp khác.

5.2.4. Phương pháp phân tích thông tin sẵn có

- Là phương pháp thu thập thông tin gián tiếp qua việc phân tích các tài liệu đã có sẵn (các tài liệu của các cơ quan lưu trữ, từ các cuộc nghiên cứu trước, từ các phương tiện thông tin đại chúng...). Thực chất của phương pháp này chỉ đơn thuần là một phương tiện để có một phân tích mới về các dữ liệu được thu thập cho các mục đích khác.

- Ưu điểm : Tiết kiệm về chi phí và nhân công

- Nhược điểm: Các tài liệu này thường ít được phân chia theo các tiêu thức cần nghiên cứu, vì vậy khó có thể tìm được nguyên nhân và các mối liên hệ qua lại giữa các tiêu thức. Hơn nữa các cơ quan thống kê thường chỉ lưu trữ số liệu về các hiện tượng kinh tế, còn các số liệu về các hiện tượng và quá trình xã hội ít được lưu trữ hơn.

5.3. Các loại câu hỏi

*) *Theo công dụng* của câu hỏi chia thành hai nhóm: theo nội dung và theo chức năng kỹ thuật.

- Theo nội dung, chia thành các loại :
  + Câu hỏi sự kiện.
  + Câu hỏi về tri thức
  + Câu hỏi về quan điểm, thái độ, động cơ.
- Theo chức năng kỹ, chia thành các loại:
  + Câu hỏi tâm lý
  + Câu hỏi lọc
  + Câu hỏi kiểm tra

*) *Theo biểu hiện của câu hỏi và câu trả lời*

- Theo biểu hiện của câu trả lời:
  + Câu hỏi đóng
  + Câu hỏi mở
  + Câu hỏi nửa đóng
- Theo biểu hiện của câu hỏi
  + Câu hỏi trực tiếp
  + Câu hỏi gián tiếp

6. Thiết kế phương án điều tra.

Để tổ chức tốt một cuộc điều tra, đòi hỏi phải xây dựng được phương án điều tra thật chi tiết, tỷ mỉ, cụ thể và toàn diện. Đây chính là tài liệu hướng dẫn thực hiện cuộc điều tra, trong đó xác định rõ những bước tiến hành, những vấn đề cần phải giải quyết, cần được hiểu thống nhất trong suốt quá trình thực hiện. Đối với các cuộc điều tra lớn, có liên quan đến nhiều cấp, nhiều ngành, như Tổng điều tra dân số, việc xây dựng phương án điều tra cần có sự phối hợp, bàn bạc thống nhất giữa cơ quan thống kê và các ngành có liên quan và phải được cấp trên có thẩm quyền phê duyệt. Trong điều kiện nền kinh tế thị trường, phương án điều tra thường được được xây dựng dưới dạng "đề xuất kỹ thuật" và "đề xuất tài chính" cho cuộc nghiên cứu. Đây chính là căn cứ để cơ quan có thẩm quyền phê duyệt hoặc là căn cứ để cơ quan chủ quản tiến hành xét chọn thầu theo quy định chung của nhà nước.

Phương án của mỗi cuộc điều tra có thể khác nhau, tùy thuộc điều kiện cụ thể của nó. Nhưng nhìn chung, mỗi phương án điều tra thường gồm những nội dung chủ yếu sau:

- Xác định mục đích điều tra.
- Xác định phạm vi, đối tượng và đơn vị điều tra.
- Xác định nội dung điều tra và thiết lập phiếu điều tra.
- Chọn thời điểm, thời kỳ và thời hạn điều tra.
- Lựa chọn phương pháp điều tra, tổng hợp số liệu và phương pháp tính các chỉ tiêu điều tra.
- Xây dựng phương án tài chính cho cuộc điều tra.
- Lập kế hoạch tổ chức và tiến hành điều tra.

6.1. Xác định mục đích điều tra

Bất kỳ một hiện tượng kinh tế xã hội nào cũng đều có thể được quan sát, xem xét, nghiên cứu trên nhiều mặt, nhiều khía cạnh, nhiều góc độ khác nhau. Nghiên cứu trên mỗi mặt, mỗi khía cạnh khác nhau sẽ cho ta đưa những kết luận khác nhau về hiện tượng và phục vụ những yêu cầu nghiên cứu cũng khác nhau. Vì vậy, trước khi tiến hành điều tra, cần xác định rõ xem cuộc điều tra này nhằm tìm hiểu vấn đề gì, phục vụ cho yêu cầu nghiên cứu nào. Đó chính mục đích của cuộc điều tra.

Mục đích điều tra còn là một trong những căn cứ quan trọng để xác định đối tượng, đơn vị điều tra, xây dựng kế hoạch và nội dung điều tra. Vì vậy, việc xác định đúng, rõ ràng mục đích điều tra sẽ là cơ sở quan trọng cho việc thu thập số liệu ban đầu đầy đủ, hợp lý, đáp ứng yêu cầu nghiên cứu đặt ra.

Căn cứ để xác định mục đích điều tra thường là những nhu cầu thực tế cuộc sống, hoặc những nhu cầu hoàn chỉnh lý luận... Những nhu cầu này được biểu hiện trực tiếp bằng các yêu cầu, đề nghị, mong muốn của cơ quan chủ quản (người sử dụng thông tin).

6.2. Xác định phạm vi, đối tượng và đơn vị điều tra

Xác định đối tượng điều tra là xác định xem những đơn vị tổng thể nào thuộc phạm vi điều tra, cần được thu thập thông tin. Như vậy, khi các đối tượng điều tra được chỉ rõ, cũng có nghĩa là phạm vi nghiên cứu đã được xác định, ranh giới giữa hiện tượng nghiên cứu với các tổng thể khác, hiện tượng khác cũng được phân biệt rõ ràng, tránh được tình trạng trùng lặp hay bỏ xót khi tiến hành điều tra.

Muốn xác định chính xác đối tượng điều tra, một mặt phải dựa vào sự phân tích lý luận, nêu lên những tiêu chuẩn cơ bản phân biệt hiện tượng nghiên cứu với các hiện tượng liên quan, phân biệt đơn vị tổng thể này với các đơn vị tổng thể khác, đồng thời cũng còn phải căn cứ vào vào mục đích nghiên cứu. Trong cuộc Tổng điều tra dân số ngày 1/4/1999 ở nước ta, đối tượng điều tra được xác định là "nhân khẩu thường trú". Điều này, vừa giúp thực hiện tốt các mục đích điều tra đã được nêu rõ trong mục trên, vừa giúp cho quá trình điều tra không bị trùng hay bỏ xót bất kỳ một nhân khẩu nào của nước ta. Tuy nhiên, trong phương án điều tra cũng cần phải có những quy định cụ thể về tiêu chuẩn xác định xác định "nhân khẩu thường trú" để tránh nhầm lẫn.

Đơn vị điều tra là đơn vị thuộc đối tượng điều tra và được điều tra thực tế. Đơn vị điều tra chính là nơi phát sinh các tài liệu ban đầu, điều tra viên cần đến đó để thu thập trong mỗi cuộc điều tra. Như vậy, nếu việc xác định đối tượng điều tra là trả lời câu hỏi "điều tra ai?", thì việc xác định đơn vị điều tra là trả lời câu hỏi "điều tra ở đâu?". Trong một số trường hợp, đơn vị điều tra và đối tượng điều tra có thể trùng nhau. Ví dụ trong cuộc điều tra nghiên cứu tình hình phát triển của các doanh nghiệp công nghiệp nhà nước thành phố Hà Nội, thì cả đối tượng và đơn vị điều tra đều là các doanh nghiệp công nghiệp nhà nước của thành phố. Nhưng cũng có nhiều trường hợp, chúng lại khác nhau. Ví dụ trong Tổng điều tra dân số ở nước ta ngày 1/4/1999, đối tượng điều tra là "nhân khẩu thường trú", còn đơn vị điều tra lại được xác định là các "hộ gia đình" và các "hộ tập thể". Trong các cuộc điều tra chọn mẫu, đơn vị điều tra chỉ bao gồm những đối tượng được chọn vào mẫu.

Cần phân biệt đơn vị điều tra và đơn vị tổng thể. Đơn vị tổng thể là các phần tử, các đơn vị cấu thành hiện tượng, mà qua đó ta có thể xác định được quy mô tổng thể. Việc xác định số đơn vị tổng thể liên quan đến việc lập phương án điều tra, chọn phương pháp điều tra, ước lượng kinh phí điều tra... còn việc xác định số đơn vị điều tra lại liên quan đến việc tổ chức ghi chép, đăng ký tài liệu, phân bổ cán bộ...

6.3. Xác định nội dung điều tra và thiết lập phiếu điều tra

Xác định nội dung điều tra là việc trả lời câu hỏi *"điều tra cái gì?"*. Nội dung điều tra là toàn bộ các đặc điểm cơ bản của từng đối tượng, từng đơn vị điều tra, mà ta cần thu được thông tin. Trong thực tế, các đơn vị của hiện tượng nghiên cứu thường có rất nhiều đặc điểm khác nhau. Tuy nhiên, không thể và cũng không cần thiết phải thu thập toàn bộ các tiêu thức đó, mà chỉ cần những tiêu thức có liên quan đến mục đích nghiên cứu, phục vụ được cho việc nghiên cứu . Vì vậy, bất kỳ cuộc điều tra nào cũng cần phải xác định rõ, cụ thể nội dung điều tra.

Việc xác định nội dung điều tra, cần căn cứ vào các yếu tố sau:

- Mục đích điều tra: Mục đích điều tra chỉ rõ cần thu thập những thông tin nào để đáp ứng yêu cầu của nó. Mục đích điều tra khác nhau, nhu cầu thông tin cũng khác nhau. Mục đích càng nhiều, nội dung điều tra càng phải rộng, càng phải phong phú.
- Đặc điểm của hiện tượng nghiên cứu: Tất cả những hiện tượng mà thống kê nghiên cứu đều tồn tại trong những điều kiện cụ thể về thời gian và không gian. Khi điều kiện này thay đổi, đặc điểm của hiện tượng cũng thay đổi. Khi đó, các biểu hiện của chúng cũng khác nhau. Vì vậy, việc lựa chọn tiêu thức nghiên cứu cũng phải khác nhau.
- Năng lực, trình độ thực tế của đơn vị, của người tổ chức điều tra. Điều này biểu hiện ở khả năng về tài chính, về thời gian, về kinh nghiệm và trình độ tổ chức điều tra. Nếu tất cả các yếu tố này được đảm bảo tốt, có thể mở rộng nội dung điều tra, nhưng vẫn đảm bảo chất lượng của các thông tin thu được. Trường hợp ngược lại, cần kiên quyết loại bỏ những nội dung chưa thực sự cần thiết

Ngoài ra, nội dung điều tra cũng chỉ nên bao gồm những tiêu thức có liên hệ chặt chẽ với nhau, để có thể kiểm tra tính chính xác của những thông tin thu được.

Để có thể thu được những thông tin một cách chính xác và đầy đủ, nội dung của mỗi cuộc điều tra phải được diễn đạt thành những câu hỏi ngắn gọn, rõ ràng, dễ hiểu và mọi người đều hiểu theo một nghĩa thống nhất. Về mặt hình thức, các câu hỏi này có thể được diễn đạt theo hai cách: câu hỏi đóng là các câu hỏi đã có sẵn các phương án trả lời (thang điểm) có thể, người trả lời chỉ cần chọn 1 trong những cách trả lời đã được đưa ra; Câu hỏi mở không có trước những phương án trả lời, người được hỏi tự diễn đạt câu trả lời. Các cuộc điều tra thống kê ít sử dụng loại câu hỏi thứ hai này.

Phiếu điều tra (hay còn gọi là biểu điều tra hay bảng hỏi) là tập hợp các câu hỏi của nội dung điều tra, được sắp xếp theo một trật tự logic nhất định. Tùy theo yêu cầu, nội dung và đối tượng, mỗi cuộc điều tra có thể phải xây dựng nhiều loại phiếu khác nhau. Phiếu điều tra là công cụ để thực hiện cuộc điều tra, nên nó phải phản ánh đầy đủ nội dung điều tra. Việc thiết kế phiếu phải đảm bảo các yêu cầu về mỹ thuật, tiết kiệm và tiện dụng. Về mặt mỹ thuật, phiếu phải được thiết kế đẹp, dễ đọc, có khả năng lôi kéo, duy trì sự quan tâm của người trả lời. Việc sắp xếp các hàng, các cột, bố trí khổ giấy... sao cho phải đảm bảo yêu cầu tiết kiệm, nhưng lại thuận lợi cho việc ghi chép, mã hóa, nhập số liệu và kiểm tra lại sau này.

Thông thường, trong các tài liệu của cuộc điều tra, người ta còn soạn thảo bản giải thích cách ghi phiếu điều tra. Bản giải thích này thường đi kèm theo phiếu điều tra nhằm giúp cho điều tra viên và người trả lời nhận thức thống nhất các câu hỏi được đặt ra, cách thu thập và ghi chép số liệu. Đối với những câu hỏi phức tạp, khó trả lời người ta còn đặt ra những ví dụ cụ thể và những quy định về các trường hợp ngoại lệ...

6.4. Chọn thời điểm, thời kỳ và quyết định thời hạn điều tra.

Các hiện tượng nghiên cứu luôn thay đổi theo thời gian và không gian. Muốn thu thập được chính xác các thông tin về chúng, cần có quy định thống nhất về thời điểm, thời kỳ và thời hạn điều tra.

Thời điểm điều tra là mốc thời gian được quy định thống nhất mà cuộc điều tra phải thu thập thông tin về hiện tượng tồn tại đúng thời điểm đó. Nếu cuộc điều tra được tiến hành vào thời điểm sau đó, thì người trả lời phải hồi tưởng lại để "miêu tả trạng thái của hiện tượng" vào đúng thời điểm điều tra.

Thời kỳ điều tra là khoảng thời gian (tuần, tháng, năm...) được quy định để thu thập số liệu về lượng của hiện tượng được tích lũy trong cả thời ký đó.

Thời hạn điều tra là khoảng thời gian dành cho việc thực hiện nhiệm vụ thu thập số liệu. Thời hạn dài hay ngắn phụ thuộc vào quy mô, tính phức tạp của hiện tượng nghiên cứu và nội dung điều tra, vào khả năng, kinh nghiệm của điều tra viên. Nhìn chung, thời hạn điều tra không nên quá dài, cách quá xa thời điểm điều tra vì có thể làm mất thông tin do người trả lời không nhớ đầy đủ các sự kiện đã xảy ra.

6.5. Lập kế hoạch tổ chức và tiến hành điều tra.

Lập kế hoạch tổ chức và tiến hành điều tra là một vấn đề trọng yếu của điều tra thống kê. Kế hoạch này quy định cụ thể từng bước công việc phải tiến hành trong quá trình từ khâu tổ chức đến triển khai điều tra thực tế. Vì vậy, nó được xây dựng càng chi tiết, tỷ mỉ, rõ ràng, cụ thể thì càng dễ thực thi, chất lượng của cuộc điều tra càng được nâng cao. Tuy nhiên, đây là một công việc phức tạp, đòi hỏi người lập kế hoạch phải có kinh nghiệm và am hiểu tình hình thực tế. Một kế hoạch tổ chức và tiến hành điều tra gồm rất nhiều khâu công việc. Thông thường, nó có thể gồm một số khâu chủ yếu là:

- Thành lập Ban chỉ đạo điều tra và quy định nhiệm vụ cụ thể cho cơ quan điều tra các cấp.
- Chuẩn bị lực lượng cán bộ điều tra, phân công trách nhiệm, địa bàn cho từng cán bộ và tiến hành tập huấn nghiệp vụ cho họ.
- Lựa chọn phương pháp điều tra thích hợp.
- Định các bước tiến hành điều tra.
- Phân chia khu vực và địa bàn điều tra.
- Tổ chức các cuộc hội nghị chuẩn bị.
- Tiến hành điều tra thử để rút kinh nghiệm, nâng cao trình độ nghiệp vụ cho cán bộ điều tra và hoàn thiện phương án điều tra, phiếu điều tra.
- Xây dựng phương án tài chính và chuẩn bị các phương tiện vật chất khác.
- Tuyên truyền mục đích, ý nghĩa của cuộc điều tra.

....

7. Các loại sai số trong điều tra.

Các cuộc điều tra thống kê, dù có cố gắng làm thật tốt vẫn thường gặp những trường hợp mà số liệu điều tra không trùng khớp với số liệu thực tế của hiện tượng nghiên cứu. Người ta gọi là sai số. *Sai số trong điều tra thống kê là chênh lệch giữa trị số thực của hiện tượng nghiên cứu so với trị số của nó mà điều tra thống kê thu được*. Sai số này làm giảm chất lượng của các cuộc điều tra, ảnh hưởng đến kết quả của tổng hợp và phân tích. Do đó, ảnh hưởng đến chất lượng của toàn bộ quá trình nghiên cứu thống kê. Trong các cuộc điều tra thống kê, người ta phải cố gắng áp dụng nhiều biện pháp để hạn chế sai số này.

Tuỳ theo các nguyên nhân dẫn đến sai số mà chia thành ***các loại sai số*** sau:

- *Sai số do đăng ký* xảy ra đối với mọi cuộc điều tra thống kê. Nó phát sinh do việc đăng ký số liệu ban đầu không chính xác. Nguyên nhân gây ra loại sai số này rất đa dạng, có thể do cân đong, đo, đếm sai, tính toán sai, ghi chép sai, do dụng cụ đo lường không chuẩn xác ... Nếu phân chia chi tiết hơn, ta có thể chia loại sai số này thành sai số ngẫu nhiên và sai số có hệ thống, do cố ý, có chủ định của người điều tra và người trả lời. Sai số ngẫu nhiên là những sai số phát sinh một cách tình cờ, không có chủ định, không có bất kỳ một sự sắp đặt trước nào của người điều tra. Nó xảy ra hoàn toàn ngẫu nhiên. Loại sai số này chịu sự chi phối của quy luật số lớn, tức là nếu ta điều tra càng nhiều đơn vị, các sai lệch ngẫu nhiên sẽ có khả năng bù trừ, triệt tiêu nhau làm cho sai số chung càng nhỏ. Sai số có hệ thống, có chủ định thường xảy ra do chủ định của người điều tra, người trả lời hoặc sai số một cách có hệ thống do lỗi của hệ thống đo lường, hệ thống thang đo được thiết kế không chuẩn xác… Loại sai số này không chịu sự chi phối của quy luật số lớn, nên điều tra càng nhiều, khả năng xảy ra sai số sẽ càng lớn.

- *Sai số do tính chất đại biểu* của số đơn vị được chọn trong điều tra chọn mẫu. Các đơn vị được chọn không đảm bảo đại diện cho toàn bộ tổng thể sẽ phát sinh sai số khi suy luận kết quả của mẫu cho tổng thể chung. Nguyên nhân dẫn đến sai số này là do: cỡ mẫu không đủ lớn, cố tình vi phạm nguyên tắc chọn mẫu, và do bản thân nguyên tắc chọn mẫu ngẫu nhiên gây nên.

- *Sai số do đo lường* là sai số do sử dụng thước đo không tốt mà nguyên nhân trực tiếp là do câu hỏi tồi (sử dụng thang đo, triển khai thang điểm không phù hợp…)

Để đảm bảo các kết quả điều tra đạt độ chính xác cao, cần áp dụng một số ***biện pháp để hạn chế sai số*** :

- *Làm tốt công tác chuẩn bị điều tra*: thông thường, trong các cuộc điều tra thống kê, công tác chuẩn bị chiếm vị trí rất quan trọng, nó đòi hỏi một sự đầu tư chất xám khá lớn. Công tác chuẩn bị càng chu đáo, tỷ mỉ, thận trọng và chi tiết, đặc biệt là trong việc thiết lập phương án điều tra, xây dựng phiếu điều tra, lựa chọn và tập huấn cán bộ điều tra càng làm tốt, sai số điều tra càng giảm.

- *Tiến hành kiểm tra có hệ thống toàn bộ cuộc điều tra*: kiểm tra là biện pháp có hiệu quả để sửa chữa, uốn nắn kịp thời các sai lầm có thể mắc phải trong quá trình điều tra. Việc

kiểm tra có thể được tiến hành theo nhiều giai đoạn khác nhau. Trước hết, cần tiến hành kiểm tra ngay từ giai đoạn chuẩn bị xem các khâu cần chuẩn bị đã được đầy đủ, chu đáo chưa. Việc kiểm tra trong giai đoạn thu thập thông tin, việc ghi chép số liệu ban đầu nhằm nâng cao ý thức trách nhiệm của nhân viên điều tra cũng có ý nghĩa quan trọng. Nghiệm thu phiếu điều tra là một khâu kiểm tra có ý nghĩa quyết định. Trong giai đoạn này, người ta cần kiểm tra xem các phiếu điều tra có đầy đủ không; Các câu trả lời, các con số được ghi chép trong từng phiếu có được tính toán đúng, đủ không, có hợp logic không, có mâu thuẫn với nhau không ... Nhìn chung, việc kiểm tra, nghiệm thu phiếu điều tra có tác dụng rất lớn, nhưng nó đòi hỏi người kiểm tra phải có trình độ, kinh nghiệm, hiểu biết thực tế và rất nhạy cảm. Tiếp theo, việc nhập số liệu vào máy cũng cần được kiểm tra thật kỹ lưỡng. Thực tế cho thấy đây cũng là một khâu dễ làm phát sinh sai số. Nhiều cuộc điều tra, người ta yêu cầu nhập hai lần độc lập nhau, để khắc phục những sai sót có thể xảy ra trong quá trình nhập số liệu.

Ngoài ra, trong các cuộc điều tra không toàn bộ, người ta còn tiến hành kiểm tra tính đại diện của các đơn vị được chọn để điều tra.

BÀI TẬP

1.1 Thảo luận và so sánh các khái niệm: Tiêu thức thống kê, chỉ tiêu thống kê, tiêu chí thống kê. Cho thí dụ minh hoạ.

1.2 Phân biệt thống kê mô tả và thống kê suy luận (phân tích)

1.3 Thảo luận và so sánh các thang đo khác nhau

1.4 Năm mùi vị kem được xếp hạng theo sở thích. Cần sử dụng thang đo nào?

1.5 Thang đo đối với màu sắc đai lưng trong môn Karate là gì?

1.6 Một phiếu đăng kí hoàn thuế cá nhân, bên cạnh những nội dung khác, hỏi về các thông tin sau: thu nhập, số nhân khẩu ăn theo, đang sống độc thân hay sống cùng vợ/chồng, các khoản giảm trừ có được phân loại hay không, tiền thuế nộp. Hãy mô tả thước đo của từng biến số, và nêu rõ biến số là biến định tính hay định lượng.

1.7 Một công ty điện lực thực hiện một cuộc điều tra với các câu hỏi về các nội dung sau:

1. Tuổi của chủ hộ gia đình
2. Giới tính của chủ hộ
3. Số lượng nhân khẩu
4. Việc sử dụng điện năng để tạo ra nhiệt lượng (Có hoặc Không)
5. Số lượng thiết bị lớn được sử dụng hằng ngày
6. Thiết bị sưởi trong mùa đông
7. Số giờ đốt nóng bình quân
8. Số ngày đốt nóng bình quân
9. Thu nhập của hộ gia đình
10. Chi phí tiền điện bình quân hằng tháng theo hoá đơn
11. Thứ hạng của công ty điện lực này trong mối quan hệ so sánh với 2 nhà cung cấp điện năng trước đó

Hãy xác định các biến số ẩn dưới 11 nội dung trên như là biến số định lượng hay định tính, và mô tả các thang đo.

1.8 Một thị trấn có 15 gia đình. Nếu bạn phỏng vấn tất cả mọi người trong một gia đình riêng biệt, bạn đang phỏng vấn tổng thể hay một mẫu từ thị trấn đó? Đây có phải là mẫu ngẫu nhiên hay không? Nếu bạn có danh sách tất cả mọi người sống trong thị trấn, và bạn lựa chọn ngẫu nhiên 100 người từ tất cả các gia đình, đây có phải là mẫu ngẫu nhiên không?

1.9 Thiết kế một nghiên cứu trên giác độ thống kê về một vấn đề mà anh/chị quan tâm.

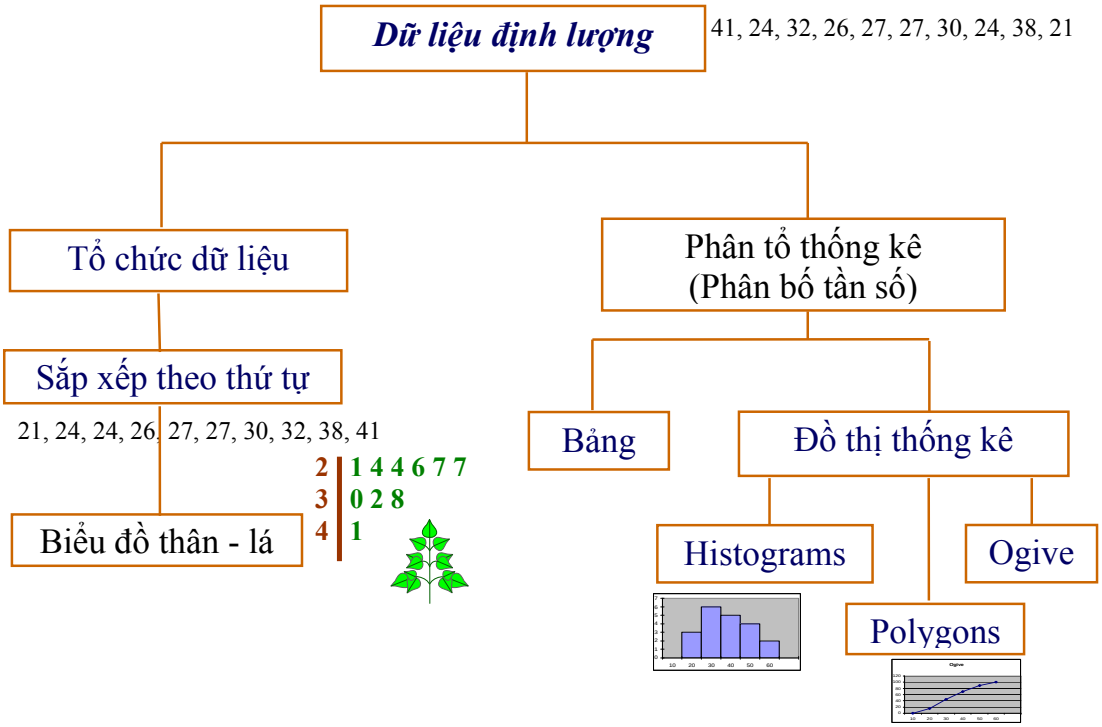
CHƯƠNG 2

TRÌNH BÀY DỮ LIỆU BẰNG BẢNG VÀ ĐỒ THỊ

Như bài 1 đã trình bày, dữ liệu thu thập được qua điều tra thống kê gồm 2 loại: dữ liệu định tính và dữ liệu định lượng. Dữ liệu định tính có biểu hiện là các thuộc tính, dữ liệu định lượng được biểu hiện thông qua các con số. Dữ liệu định lượng gồm 2 loại: rời rạc và liên tục. Việc trình bày các loại dữ liệu trên cần phải lựa chọn các phương pháp thích hợp.

1. Trình bày dữ liệu định lượng.

Các bước trình bày dữ liệu định lượng



1.1. Tổ chức dữ liệu định lượng:

- Sắp xếp số liệu theo thứ tự từ nhỏ đến lớn: bước đầu cho thấy đặc điểm về lượng của hiện tượng (lượng biến lớn nhất, nhỏ nhất, lượng biến phổ biến nhất...), là cơ sở cho việc lập bảng thống kê.

- Biểu hiện dữ liệu bằng biểu đồ thân lá (Stem and leaf): bước đầu cho thấy đặc trưng phân phối của tập hợp dữ liệu. Nội dung cơ bản của phương pháp này là mỗi trị số trong tập hợp dữ liệu được chia thành hai phần: phần thân và phần lá. Tùy theo số chữ số của mỗi trị số trong tập hợp dữ liệu mà chia một hoặc một số chữ số bên trái là phần thân và một hoặc một số chữ số bên phải là phần lá. Để minh họa hãy theo dõi thí dụ dưới đây :

Thí dụ 1:

Chẳng hạn có tập hợp dữ liệu:

26, 27, 13, 12, 58, 17, 53, 46, 21, 24, 44, 24, 41, 43, 35, 27, 38, 30, 37, 32

Đây là dữ liệu thô có được qua điều tra, chúng ta chưa thể đưa ra nhận xét gì về đặc điểm của tập hợp dữ liệu này. Trước hết sắp xếp tập hợp dữ liệu trên theo thứ tự tăng dần và có kết quả sau:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Rất dễ nhận thấy ngay các giá trị nhỏ nhất và lớn nhất trong tập hợp dữ liệu là 12 và 58, có 2 trị số được lặp lại 2 lần là 24 và 27. Để thấy được đặc điểm phân phối của tập hợp dữ liệu này hãy tiến hành thiết kế biểu đồ thân lá. Chúng ta bắt đầu từ việc xác định thân và lá. Mỗi trị số chỉ có hai chữ số, nên việc xác định đơn giản chỉ là chữ số hàng chục là thân và chữ số hàng đơn vị là lá. Như vậy phần thân sẽ gồm 1, 2, 3, 4, 5 và sắp xếp như sau:

Thân	***Lá***
1	
2	
3	
4	
5	

Sau đó lần lượt sắp xếp các trị số của tập hợp dữ liệu vào các bộ phận của phần thân này. Chẳng hạn, số 12 ghi ở phần lá chữ số 2 tương ứng với phần thân là 1… Tuần tự tiến hành như vậy sẽ có biểu đồ thân lá như sau:

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

Chú ý: trên thực tế khi phần thân ít và phần lá ở mỗi thân quá nhiều chúng ta có thể tách thành hai phần dưới và trên: phần thân dưới sẽ có các lá là các chữ số từ 0 đến 4, phần thân trên sẽ có các lá là các chữ số từ 5 đến 9.

Để thấy rõ ý nghĩa ứng dụng của biểu đồ thân lá, hãy xét một tình huống cụ thể như sau:

Thí dụ 2:

Giám đốc 1 công ty tin học dự định trả mức lương 1.900.000 VNĐ/ tháng cho một lập trình viên làm tại công ty với 3 năm kinh nghiệm. Để biết mức lương này đã là thỏa đáng hay chưa, ông ta tổ chức một cuộc điều tra 30 lập trình viên làm cho các công ty cạnh tranh với 3 năm kinh nghiệm. Kết quả điều tra như sau *(đơn vị tính: nghìn đồng/tháng)*:

1.500	1.700	1.600	2.100	1.700
1.800	1.650	1.550	1.900	1.850
1.650	1.700	1.500	1.800	1.600
1.600	1.400	1.850	1.800	1.900
1.500	1.700	1.800	1.750	1.800
1.800	2.100	2.200	2.050	1.800

Số liệu đã sắp xếp:

1.400	1.600	1.700	1.800	1.900
1.500	1.600	1.700	1.800	1.900
1.500	1.650	1.750	1.800	2.050
1.500	1.650	1.800	1.800	2.100
1.550	1.700	1.800	1.850	2.100
1.600	1.700	1.800	1.850	2.200

Qua dữ liệu đã sắp xếp cho thấy mức tiền lương thấp nhất là 1400 ng.đ, mức cao nhất là 2200 ng.đ và mức tiền lương có nhiều người nhất (phổ biến nhất) là 1800 ng.đ. Để có nhận xét đầy đủ hơn về đặc điểm phân phối của các mức tiền lương có thể lập biểu đồ thân lá như sau:

Thân	***Lá*** *(mỗi lá tương ứng với 2 chữ số)*
14	00
15	00 00 00 50
16	00 00 00 50 50
17	00 00 00 00 50
18	00 00 00 00 00 00 00 50 50
19	00 00
20	50
21	00 00
22	00

Biểu đồ thân lá cho thấy mức lương phổ biến nhất là từ 1800 đến dưới 1900 ng.đ; số người có mức lương dưới 1900 ng.đ chiếm đa số và chỉ có 4 người trên mức lương này. Điều đó khẳng định rằng ông giám đốc dự định trả mức lương 1900 ng.đ là mức lương rất thỏa đáng. Với mức lương đó không những đã làm cho các lập trình viên ở Công ty của ông yên tâm và tích cực làm việc mà còn có thể thu hút lao động và cạnh tranh với các Công ty khác.

Như vậy chỉ bằng việc tổ chức dữ liệu hợp lý đã có thể có những kết luận đáng tin cậy làm cơ sở cho việc ra quyết định quản lý.

Qua các thí dụ trên cho thấy biểu đồ thân lá cũng gần giống biểu đồ phân phối (Histograms), ở đây mỗi trị số của phần thân tương ứng với một cột trong biểu đồ phân phối, số lượng lá tương ứng với chiều cao của các cột đó. Tuy nhiên các lá trong biểu đồ thân lá không chỉ cho thấy chiều cao của cột (hay tần số trong phân phối) lớn hay nhỏ mà còn giúp ta quan sát chi tiết từng trị số cụ thể nhằm phân tích sâu sắc và đầy đủ hơn phân phối của cả

tổng thể nói chung và từng bộ phận nói riêng. Điều đó làm cho biểu đồ thân lá có lợi thế hơn khi tóm tắt và trình bày dữ liệu, nhưng chỉ trong trường hợp tập hợp dữ liệu không lớn. Khi quy mô của bộ dữ liệu lên tới hàng trăm, hàng nghìn hoặc nhiều hơn nữa thì sẽ rối mắt và biểu đồ thân lá không còn phù hợp. Lúc này bảng phân bố tần số sẽ thích hợp hơn.

1.2. Lý thuyết phân tổ thống kê.

1.2.1. Một số vấn đề chung về phân tổ thống kê

1.2.1.1. Khái niệm về phân tổ thống kê.

Ta đã biết, hiện tượng và quá trình kinh tế xã hội mà thống kê học nghiên cứu thường rất phức tạp, vì chúng tồn tại và phát triển dưới nhiều loại hình có quy mô và đặc điểm khác nhau. Trong kết cấu nội bộ của hiện tượng nghiên cứu thường bao gồm nhiều tổ, nhiều bộ phận có tính chất khác nhau. Muốn phản ánh được bản chất và quy luật phát triển của hiện tượng, nếu chỉ dựa vào những con số tổng cộng chung chung thì không thể nêu được vấn đề một cách sâu sắc. Phải tìm cách nêu lên được đặc trưng của từng loại hình, của từng bộ phận cấu thành hiện tượng phức tạp, đánh giá tầm quan trọng của mỗi bộ phận, nêu lên mối liên hệ giữa các bộ phận, rồi từ đó nhận thức được các đặc trưng chung của toàn bộ. Yêu cầu nói trên chỉ có thể giải quyết được bằng phương pháp phân tổ thống kê.

Phân tổ thống kê là căn cứ vào một (hay một số) tiêu thức nào đó để tiến hành phân chia các đơn vị của hiện tượng nghiên cứu thành các tổ (và các tiểu tổ) có tính chất khác nhau.

Chẳng hạn, khi nghiên cứu tình hình nhân khẩu, căn cứ vào tiêu thức "giới tính" để chia tổng số nhân khẩu thành hai tổ: nam và nữ; còn căn cứ vào tiêu thức "tuổi" để chia số nhân khẩu này thành nhiều tổ có độ tuổi khác nhau. Hay khi nghiên cứu tình hình sản xuất của các doanh nghiệp công nghiệp, có thể chia tổng số doanh nghiệp thành các nhóm theo các tiêu thức như: thành phần kinh tế, ngành sản xuất, số lượng lao động, giá trị sản xuất công nghiệp...

Khi phân tổ thống kê, trước hết ta *thực hiện được việc nghiên cứu cái chung và cái riêng một cách kết hợp*. Các đơn vị tổng thể được tập hợp lại thành một số tổ (và tiểu tổ): giữa các tổ đều có sự khác nhau rõ rệt về tính chất, còn trong phạm vi mỗi tổ các đơn vị đều có sự giống nhau (hay gần giống nhau) về tính chất theo tiêu thức được dùng làm căn cứ phân tổ.

Phân tổ thống kê là *phương pháp cơ bản để tiến hành tổng hợp thống kê*, vì ta sẽ không thể tiến hành hệ thống hoá một cách khoa học các tài liệu thu được qua điều tra, nếu không áp đụng phương pháp này. Tính chất phức tạp của hiện tượng nghiên cứu đòi hỏi phải tổng hợp theo từng tổ, từng bộ phận. Vì vậy, khi tổng hợp thống kê, trước hết, người ta thường sắp xếp các đơn vị vào từng tổ, từng bộ phận, tính toán các đặc điểm của mỗi tổ hoặc bộ phận, rồi sau đó mới tính các đặc điểm chung của cả tổng thể.

Phân tổ thống kê là *một trong các phương pháp quan trọng của phân tích thống kê, đồng thời là cơ sở để vận dụng các phương pháp phân tích thống kê khác*. Chỉ sau khi đã phân chia tổng thể nghiên cứu thành các tổ có quy mô và đặc điểm khác nhau, việc tính các chỉ tiêu phản ánh mức độ, tình hình biến động, mối liên hệ giữa các hiện tượng mới có ý nghĩa đúng đắn. Nếu việc phân tổ không chính xác, tổng thể được chia thành những bộ phận không đúng với thực tế, thì mọi chỉ tiêu tính ra cũng không giúp ta rút ra được những kết luận đúng đắn. Phương pháp phân tổ được vận dụng phổ biến nhất trong mọi trường hợp nghiên cứu kinh tế, vì không những phương pháp này đơn giản, dễ hiểu mà lại có tác dụng phân tích sâu sắc. Các phương pháp phân tích thống kê khác như: phương pháp số tương đối, phương pháp số bình quân, phương pháp chỉ số, phương pháp bảng cân đối, phương pháp tương quan... thường cũng phải dựa trên các kết quả phân tổ thống kê chính xác.

Phân tổ thống kê còn được *vận dụng ngay trong giai đoạn điều tra thống kê*, nhằm phân tổ đối tượng diều tra thành những bộ phận có đặc điểm tính chất khác nhau từ đó chọn các đơn vị điều điều tra sao cho có tính đại biểu cho tổng thể chung.

Phân tổ thống kê giải quyết những nhiệm vụ nghiên cứu cơ bản sau đây:

Thứ nhất, phân tổ thực hiện việc phân chia các loại hình kinh tế xã hội của hiện tượng nghiên cứu. Hiện tượng kinh tế xã hội mà thống kê học nghiên cứu thường không phải là tổng thể đồng chất, mà là tổng thể bao gồm nhiều đơn vị thuộc các loại hình rất khác nhau, phát triển theo những xu hướng không giống nhau. Vì vậy phương pháp nghiên cứu khoa học là phải nêu lên các đặc trưng riêng biệt của từng loại hình và mối quan hệ giữa các loại hình đó với nhau. Muốn vậy, trước hết phải dựa trên lý luận kinh tế chính trị xã hội để phân biệt các bộ phận khác nhau về tính chất đang tồn tại khách quan trong nội bộ hiện tượng.

Thứ hai, phân tổ có nhiệm vụ biểu hiện kết cấu của hiện tượng nghiên cứu. Ta biết rằng một hiện tượng kinh tế xã hội do nhiều bộ phận, nhiều nhóm đơn vị có tính chất khác nhau hợp thành. Các bộ phận hay nhóm này chiếm những tỷ trọng khác nhau trong tổng thể và nói lên tầm quan trọng của mình trong tổng thể đó. Mặt khác, tỷ trọng của các bộ phận còn nói lên kết cấu của tổng thể theo một tiêu thức nào đó. Muốn nghiên cứu được kết cấu của tổng thể, phải dựa trên cơ sở phân tổ thống kê.

Thứ ba, phân tổ được dùng để biểu hiện mối liên hệ giữa các tiêu thức. Hiện tượng kinh tế xã hội phát sinh và biến động không phải một cách ngẫu nhiên, tách rời với các hiện tượng xung quanh, mà chúng có liên hệ và phụ thuộc lẫn nhau theo những quy luật nhất định. Giữa các tiêu thức mà thống kê nghiên cứu cũng thường có mối liên hệ với nhau: sự thay đổi của tiêu thức này sẽ đưa đến sự thay đổi của tiêu thức kia theo một quy luật nhất định. Tìm hiểu tính chất và trình độ của mối liên hệ giữa các hiện tượng nói chung và giữa các tiêu thức nói riêng là một trong các nhiệm vụ quan trọng của nghiên cứu thống kê. Phân tổ thống kê là một trong các phương pháp có thể giúp ta thực hiện nhiệm vụ nghiên cứu này.

1.2.1.2. Các loại phân tổ thống kê.

Trong thống kê thường có các cách phân loại phân tổ thống kê như sau:

a) Căn cứ vào nhiệm vụ của phân tổ thống kê

Tương ứng với ba nhiệm vụ nói trên của phân tổ thống kê, có ba loại phân tổ : Phân tổ phân loại; phân tổ kết cấu và phân tổ liên hệ

**) Phân tổ phân loại*:

Phân tổ phân loại giúp ta nghiên cứu một cách có phân biệt các loại hình kinh tế xã hội, nêu lên đặc trưng và mối quan hệ giữa chúng với nhau. Từ việc nghiên cứu riêng biệt mỗi loại hình đó, tiến thêm một bước nghiên cứu các đặc trưng của toàn bộ hiện tượng phức tạp, giải thích một cách sâu sắc bản chất và xu hướng phát triển của hiện tượng trong điều kiện thời gian và địa điểm cụ thể.

Tuỳ theo mục đích nghiên cứu, có thể phân loại các đơn vị theo nhiều tiêu thức khác nhau. Chẳng hạn, các doanh nghiệp công nghiệp nước ta có thể được phân loại theo thành phần kinh tế, theo cấp quản lý, theo nhóm, theo ngành, theo quy mô...

**) Phân tổ kết cấu*:

Trong công tác nghiên cứu thống kê, các bảng phân tổ kết cấu được sử dụng rất phổ biến, nhằm mục đích nêu lên *bản chất* của hiện tượng trong điều kiện nhất định và để nghiên cứu *xu hướng phát triển* của hiện tượng qua thời gian. Kết cấu của tổng thể phản ánh một trong các đặc trưng cơ bản của hiện tượng trong điều kiện thời gian và địa điểm cụ thể. Sự thay đổi kết cấu của tổng thể qua thời gian có thể giúp ta thấy được xu hướng phát triển của hiện tượng. Chẳng hạn sự thay đổi kết cấu về tổng sản phẩm trong nước phân theo nhóm ngành (khu vực kinh tế) phản ánh sự chuyển dịch cơ cấu ngành kinh tế trong quá trình phát triển của Việt Nam như sau:

Bảng 2.1. **Cơ cấu tổng sản phẩm trong nước theo nhóm ngành giai đoạn 2000 - 2004**

Đơn vị: %

Tổng sản phẩm trong nước phân theo nhóm ngành	***2000***	***2001***	***2002***	***2003***	***2004***
Nông, lâm nghiệp và thuỷ sản	24,53	23,24	23,03	22,54	21,76
Công nghiệp và xây dựng	36,73	38,13	38,49	39,47	40,09
Dịch vụ	38,74	38,63	38,48	37,99	38,15
Cộng	100,00	100,00	100,00	100,00	100,00

(Nguồn: Niên giám Thống kê 2004 trang 71)

Qua bảng phân tổ kết cấu trên, sự thay đổi về tỷ trọng của 3 nhóm ngành đã nói lên một phần quá trình chuyển dịch cơ cấu ngành ở Việt Nam, cụ thể nhóm ngành công nghiệp và xây dựng chiếm tỷ trọng lớn và đang có xu hướng tăng, nhóm ngành nông, lâm và thuỷ sản chiếm tỷ trọng nhỏ lại đang có xu hướng giảm....

Phân tổ kết cấu giúp ta có thể *so sánh được bản chất* của các hiện tượng cùng loại trong điều kiện không gian khác nhau. Ví dụ, có thể so sánh cơ cấu công nhân của hai nhà máy, cơ cấu giống lúa của hai hợp tác xã. Phân tổ kết cấu còn được vận dụng trong *phân tích thực hiện kế hoạch* để thấy rõ tỷ trọng các bộ phận chưa hoàn thành, hoàn thành và hoàn thành vượt mức kế hoạch. Từ đó có thể đánh giá việc thực hiện kế hoạch; xem xét lại việc đặt kế hoạch như vậy có hợp lý không và có thể tính được khả năng tiềm tàng vượt mức kế hoạch, trên cơ sở kết hợp với các giả thiết khác.

Trong nhiều trường hợp nghiên cứu, phân tổ kết cấu có thể được xác định ngay trên cơ sở phân tổ phân loại, như vậy là hai loại phân tổ này thường kết hợp chặt chẽ với nhau. Mặt khác, ngay cả đối với một tổng thể đồng chất cũng vẫn thường bao gồm các bộ phận khác nhau do nhiều nguyên nhân cụ thể, cho nên vẫn cần phân tổ kết cấu. Như tổng thể công nhân thuộc cùng một nghề trong cùng một doanh nghiệp, số công nhân này vẫn khác nhau về giới tính, về tuổi nghề, về bậc thợ và về nhiều đặc điểm khác. Như vậy là phân tổ kết cấu rất cần thiết đối với bất kỳ công tác nghiên cứu thống kê nào.

**) Phân tổ liên hệ.*

Khi tiến hành phân tổ liên hệ, các tiêu thức có liên hệ với nhau được phân biệt thành hai loại: tiêu thức nguyên nhân và tiêu thức kết quả. *Tiêu thức nguyên nhân* là tiêu thức gây ảnh hưởng; sự biến động của tiêu thức này sẽ dẫn đến sự thay đổi (tăng hoặc giảm) của tiêu thức phụ thuộc mà ta gọi là *tiêu thức kết quả* - một cách có hệ thống. Như vậy, các đơn vị tổng thể trước hết được phân tổ theo một tiêu thức (thường là tiêu thức nguyên nhân), sau đó trong mỗi tổ tiếp tục tính các trị số bình quân của tiêu thức còn lại (thường là tiêu thức kết quả). Quan sát sự biến thiên của hai tiêu thức này có thể giúp ta kết luận về tính chất của mối liên hệ giữa hai tiêu thức. Như trong nhiều doanh nghiệp công nghiệp, ta thường nhận thấy có mối liên hệ giữa năng suất lao động và giá thành đơn vị sản phẩm: năng suất lao động càng tăng thì giá thành đơn vị sản phẩm càng có điều kiện giảm. Nếu ta phân tổ các doanh nghiệp trong cùng một ngành theo năng suất lao động, sau đó từ mỗi tổ tính ra giá thành bình quân đơn vị sản phẩm, thì các kết quả tính toán sẽ cho thấy rõ mối liên hệ giữa năng suất lao động (trong trường hợp này là tiêu thức nguyên nhân) và giá thành đơn vị sản phẩm (trong trường hợp này là tiêu thức kết quả).

Phân tổ liên hệ còn có thể được vận dụng để nghiên cứu mối liên hệ giữa nhiều tiêu thức. Có thể nghiên cứu mối liên hệ giữa năng suất lúa với lượng phân bón, lượng nước tưới, mật độ cấy...; hoặc nghiên cứu mối liên hệ giữa năng suất lao động của công nhân với tuổi nghề, bậc thợ, trình độ trang bị kỹ thuật...

Khi phân tổ liên hệ giữa nhiều tiêu thức (ví dụ, 3 tiêu thức) trước hết tổng thể được phân tổ theo một tiêu thức nguyên nhân, sau đó mỗi tổ lại được chia thành các tiểu tổ theo tiêu thức nguyên nhân thứ hai, cuối cùng tính trị số tổng hoặc bình quân của tiêu thức kết quả cho từng tổ và tiểu tổ đó. Sau đây là thí dụ về mối liên hệ giữa năng suất lao động với trình độ kỹ thuật và tuổi nghề của công nhân trong một doanh nghiệp, được trình bày thành bảng phân tổ kết hợp như sau:

Bảng 2.2. **Mối liên hệ giữa năng suất lao động với trình độ kỹ thuật và tuổi nghề**

Phân tổ công nhân		***Số công nhân***	***Sản lượng cả năm (tấn)***	***Năng suất lao động bình quân năm (tấn)***
Theo trình độ kỹ thuật	***Theo tuổi nghề (năm)***			
Đã được đào tạo kỹ thuật	Dưới 5	15	1125	75
	5 – 10	40	3750	94
	10 – 15	40	4200	105
	15 – 20	15	1725	115
	20 trở lên	10	1200	120
Cả tổ	-	120	12000	100
Chưa được đào tạo kỹ thuật	Dưới 5	10	510	51
	5 – 10	30	2140	71
	10 – 15	20	1580	79
	15 – 20	10	860	86
	20 trở lên	10	910	91
Cả tổ	-	80	6000	75
Chung cho cả doanh nghiệp	-	200	18000	90

b) Căn cứ vào số lượng tiêu thức của phân tổ.

Theo định nghĩa phân tổ thống kê, chúng ta có thể căn cứ vào một hay một số tiêu thức để tiến hành phân tổ. Vì vậy, có thể phân thành hai loại: phân tổ theo một tiêu thức và phân tổ theo nhiều tiêu thức.

**) Phân tổ theo một tiêu thức:* Là tiến hành phân chia các đơn vị thuộc hiện tượng nghiên cứu thành các tổ có tính chất khác nhau trên cơ sở một tiêu thức thống kê hay còn gọi là *phân tổ giản đơn*. Chẳng hạn, theo tiêu thức giới tính, tổng thể dân số được chia thành 2 tổ: Nam và Nữ; hoặc theo tiêu thức thành phần kinh tế, Tổng sản phẩm được chia thành 5 tổ tương ứng với 5 thành phần kinh tế hiện nay...

**) Phân tổ theo nhiều tiêu thức:* Là tiến hành phân chia các đơn vị thuộc hiện tượng nghiên cứu thành các tổ và các tiểu tổ có tính chất khác nhau trên cơ sở nhiều tiêu thức thống kê (từ hai tiêu thức trở lên). Tuỳ thuộc vào mục đích nghiên cứu, đặc điểm của hiện tượng và các tiêu thức phân tổ mà phân tổ theo nhiều tiêu thức được chia thành hai loại: Phân tổ kết hợp và phân tổ nhiều chiều

Phân tổ kết hợp là tiến hành phân tổ lần lượt theo từng tiêu thức một. Các tiêu thức được sắp xếp theo thứ tự phù hợp với mục đích nghiên và đặc điểm của hiện tượng. Thông thường người ta hay phân tổ theo tiêu thức liên quan trực tiếp đến mục đích nghiên cứu và có ít biểu hiện trước. Chẳng hạn, tổng thể Dân số trước hết được phân tổ theo tiêu thức giới tính, sau đó theo tiêu thức độ tuổi và đó là cơ sở để xây dựng tháp dân số, hoặc phân tổ tổng

thể một loại lao động nào đó của một doanh nghiệp theo mức lương và số năm kinh nghiệm... Tuy nhiên theo cách này số tiêu thức phân tổ không nên quá nhiều (thường 2 hoặc 3) vì nếu không sẽ chia tổng thể thành quá nhiều bộ phận nhỏ có thể gây khó khăn cho việc phân tích.

Phân tổ nhiều chiều là cùng một lúc phân tổ theo nhiều tiêu thức khác nhau nhưng có vai trò như nhau trong việc đánh giá hiện tượng. Chẳng hạn, để phản ánh quy mô của một doanh nghiệp có thể biểu hiện qua các tiêu thức: doanh thu, số lượng lao động, tổng vốn... Các tiêu thức này khác nhau về số lượng và đơn vị tính nhưng đều biểu hiện quy mô của doanh nghiệp và việc sắp xếp thứ tự trước sau các tiêu thức này trong phân tổ các doanh nghiệp trong một ngành là không có ý nghĩa. Vì vậy phải cùng một lúc phân tổ theo tất cả các tiêu thức bằng cách đưa các tiêu thức này về một tiêu thức tổng hợp chung gọi là phân tổ nhiều chiều.

1.2.1.3. Tiêu thức phân tổ và chỉ tiêu giải thích.

a) Tiêu thức phân tổ

Tiêu thức phân tổ là tiêu thức được chọn làm căn cứ để tiến hành phân tổ thống kê. Lựa chọn tiêu thức phân tổ là vấn đề quan trọng đầu tiên phải đề ra và giải quyết chính xác. Tuy các đơn vị tổng thể có rất nhiều tiêu thức khác nhau, nhưng chúng ta không thể tuỳ tiện chọn bất kỳ tiêu thức nào làm căn cứ phân tổ.

Tiêu thức phân tổ khác nhau sẽ nói lên những mặt khác nhau của hiện tượng. Có tiêu thức phân tổ nói rõ được bản chất của hiện tượng, nhưng cũng có tiêu thức, nếu được chọn làm căn cứ phân tổ, sẽ không đáp ứng mục đích nghiên cứu, thậm chí còn làm cho ta hiểu sai lệch bản chất của hiện tượng. Bởi vì cũng những tài liệu như nhau mà cách sắp xếp khác nhau, lại đưa đến những kết luận trái ngược hẳn với nhau. Như vậy, việc phân tổ chính xác và khoa học trước hết phụ thuộc vào việc lựa chọn tiêu thức phân tổ.

b*) Các chỉ tiêu giải thích*

Trong phân tổ thống kê, sau khi đã lựa chọn được tiêu thức phân tổ, xác định số tổ cần thiết và khoảng cách tổ, còn phải xác định các chỉ tiêu giải thích để nói rõ đặc trưng của các tổ cũng như của toàn bộ tổng thể. Chẳng hạn, sau khi phân tổ các doanh nghiệp công nghiệp theo khu vực và thành phần kinh tế, có thể đưa ra một số chỉ tiêu giải thích như sau:

Bảng 2.3 **Bảng phân tổ các doanh nghiệp công nghiệp theo khu vực và thành phần kinh tế năm 2003**

Phân tổ các doanh nghiệp theo thành phần kinh tế	*Số doanh nghiệp*	*Số lao động (người)*	*Doanh thu thuần (tỷ đồng)*	*DT thuần b/quân 1 lđ (tr.đ/người)*
1. Khu vực DN Nhà nước Trong đó:	***4.845***	***2.264.942***	***678.735***	***300***

- DN Nhà nước trung ương	1.898	1.463.954	513.509	351
- DN Nhà nước địa phương	2.947	800.988	165.226	206
2. Khu vực DN ngoài Nhà nước	***64.526***	***2.049.891***	***485.104***	***237***
Trong đó:				
- DN tập thể	4.150	160.949	12.705	79
- DN tư nhân	25.653	378.087	104.043	275
- Công ty hợp danh	18	655	10.409	15.892
- Công ty TNHH tư nhân	30.164	1.143.055	270.993	237
- CTy cổ phần có vốn Nhà nước	669	160.879	43.298	269
- CTy cổ phần không có vốn Nhà nước	3.872	206.266	43.656	212
3. Khu vực có vốn đầu tư nước ngoài.	***2.641***	***860.259***	***292.932***	***341***
Trong đó:				
- DN 100% vốn nước ngoài	1.869	687.725	131.158	191
- DN liên doanh với nước ngoài	772	172.534	161.774	938
Chung	**72.012**	**5.175.092**	**1.456.771**	**281**

(Nguồn: Thực trạng doanh nghiệp qua kết quả điều tra năm 2002, 2003, 2004 - Nhà xuất bản Thống kê)

Mỗi chỉ tiêu giải thích có ý nghĩa quan trọng riêng giúp ta thấy rõ các đặc trưng số lượng của từng tổ cũng như của toàn bộ tổng thể, làm căn cứ để so sánh các tổ với nhau và để tính ra hàng loạt chỉ tiêu phân tích khác. Tuy nhiên, cũng không nên đề ra quá nhiều chỉ tiêu, mà phải lựa chọn một số chỉ tiêu nào thích hợp nhất đối với mục đích nghiên cứu.

Muốn xác định các chỉ tiêu giải thích, chủ yếu phải căn cứ vào mục đích nghiên cứu và nhiệm vụ của phân tổ để chọn ra các chỉ tiêu có liên hệ với nhau và bổ sung cho nhau. Mục đích nghiên cứu có thể tiếp cận từ nhiều khía cạnh khác nhau nên chỉ tiêu giải thích chọn ra phải hợp lý mới thoả mãn được mục đích nghiên cứu. Phải chọn các chỉ tiêu giải thích có liên hệ với nhau và bổ sung cho nhau, vì một chỉ tiêu chỉ có thể nói lên biểu hiện số lượng về một mặt nào đó của hiện tượng nghiên cứu, cho nên cần có các chỉ tiêu giải thích bổ sung cho nhau mới giúp cho việc nghiên cứu được sâu sắc.

Cũng cần chú ý tới mối quan hệ nhất định giữa tiêu thức phân tổ với các chỉ tiêu giải thích. Chẳng hạn, khi phân tổ các xí nghiệp theo quy mô, thì các chỉ tiêu giải thích như: số lượng lao động, giá trị tài sản cố định, giá trị sản xuất là những chỉ tiêu giúp ta hiểu rõ thêm về quy mô của xí nghiệp. Trái lại, nếu chọn các chỉ tiêu giải thích như: mức độ hoàn thành kế hoạch, tiền lương bình quân... thì các chỉ tiêu này thường không trực tiếp chịu ảnh hưởng bởi quy mô của xí nghiệp. Các chỉ tiêu giải thích có ý nghĩa quan trọng trong việc so sánh với nhau cần được bố trí gần nhau. Chẳng hạn, nên bố trí chỉ tiêu thực hiện gần chỉ tiêu kế hoạch, chỉ tiêu tương đối gần chỉ tiêu tuyệt đối có liên quan...

1.2.2. Các bước phân tổ thống kê

1.2.2.1 Lựa chọn tiêu thức phân tổ

Lựa chọn tiêu thức phân tổ là bước đầu tiên làm cơ sở để tiến hành phân tổ. Lựa chọn tiêu thức chính xác, phù hợp với mục đích nghiên cứu thì kết quả phân tổ mới thực sự có ích cho việc phân tích đặc điểm và bản chất của hiện tượng. Có thể nêu ra những yêu cầu sau đây về lựa chọn tiêu thức phân tổ:

Thứ nhất, phải dựa trên cơ sở phân tích lý luận một cách sâu sắc để chọn ra tiêu thức bản chất nhất, phù hợp với mục đích nghiên cứu.

Tiêu thức bản chất là tiêu thức nói lên được bản chất của hiện tượng nghiên cứu, phản ánh đặc trưng cơ bản của hiện tượng trong điều kiện thời gian và địa điểm cụ thể. Bản chất của mỗi hiện tượng có thể được phản ánh qua nhiều tiêu thức khác nhau, cho nên phải tuỳ theo mục đích nghiên cứu mà dùng lý luận để chọn ra tiêu thức bản chất nhất. Chẳng hạn, muốn phân tổ các doanh nghiệp sản xuất công nghiệp để biểu hiện quy mô lớn nhỏ, ta phải căn cứ vào thực tế của các doanh nghiệp đó, để xét xem tiêu thức nào có khả năng phản ánh quy mô của chúng như: số lượng lao động, giá trị sản xuất, giá trị thiết bị chủ yếu, diện tích doanh nghiệp... Đối với những doanh nghiệp mà quá trình sản xuất chủ yếu còn dựa vào sức lao động thì có thể chọn tiêu thức "số lượng lao động" để tiến hành phân tổ, vì số lượng lao động nhiều hay ít sẽ nói lên quy mô của doanh nghiệp lớn hay nhỏ. Nhưng đối với doanh nghiệp mà quá trình sản xuất đã được cơ giới hoá hoặc tự động hoá cao, thì muốn biểu hiện quy mô của chúng phải phân tổ theo các tiêu thức như: giá trị sản xuất, giá trị thiết bị sản xuất chủ yếu... Đó là các tiêu thức bản chất nhất, có thể đáp ứng được mục đích nghiên cứu.

Thứ hai, phải căn cứ vào điều kiện lịch sử cụ thể của hiện tượng nghiên cứu để chọn ra tiêu thức phân tổ thích hợp.

Cùng một loại hiện tượng nghiên cứu, nhưng phát sinh trong những điều kiện thời gian và địa điểm khác nhau, thì bản chất có thể thay đổi khác nhau. Vì vậy, 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 này có thể giúp ta nghiên cứu chính xác, nhưng trong điều kiện khác lại không có tác dụng gì cả.

Thứ ba, phải tuỳ theo mục đích nghiên cứu và điều kiện tài liệu thực tế mà quyết định phân tổ hiện tượng theo một hay nhiều tiêu thức.

Nói chung, hiện tượng nghiên cứu thường phức tạp, cho nên việc phân tổ theo một tiêu thức, dù là tiêu thức bản chất nhất cũng chỉ phản ánh được một mặt nào đó của hiện tượng. Nếu phân tổ kết hợp theo nhiều tiêu thức, sẽ phản ánh được nhiều mặt khác nhau của hiện tượng, các mặt này có thể bổ sung cho nhau và giúp cho việc nghiên cứu thêm sâu sắc. Trong nhiều trường hợp phân tổ kết hợp giúp ta nghiên cứu mối liên hệ giữa các tiêu thức. Ví dụ, có thể phân tổ nhân khẩu theo giới tính và theo độ tuổi (kết hợp hai tiêu thức), phân tổ các doanh nghiệp theo nhóm, theo ngành, và theo thành phần kinh tế (kết hợp ba tiêu thức). Tuy nhiên, cũng cần chú ý là không nên phân tổ kết hợp theo quá nhiều tiêu thức (trên ba

tiêu thức) vì làm như vậy số tổ và tiểu tổ sẽ tăng lên nhiều, tổng thể bị chia nhỏ nhiều quá sẽ trở ngại cho việc nghiên cứu. Thường người ta chỉ phân tổ kết hợp theo hai hay ba tiêu thức và nếu thấy cần thiết, có thể lập nhiều bảng phân tổ kết hợp khác nhau.

1.2.2.2 Xác định số tổ và khoảng cách tổ

Sau khi đã chọn được tiêu thức phân tổ thích hợp, vấn đề tiếp theo là xét xem cần phân chia hiện tượng nghiên cứu thành bao nhiêu tổ và căn cứ vào đâu để xác định số tổ cần thiết đó.

Số tổ cần thiết thường được xác định tuỳ theo tiêu thức phân tổ là tiêu thức thuộc tính hay tiêu thức số lượng. Đối với mỗi loại tiêu thức này, vấn đề xác định số tổ cần thiết được giải quyết khác nhau.

a. Phân tổ theo tiêu thức thuộc tính:

Khi phân tổ theo tiêu thức thuộc tính, các tổ được hình thành không phải do sự khác nhau về lượng biến của tiêu thức mà thường do các loại hình khác nhau, tuy nhiên không nhất thiết lúc nào mỗi loại hình cũng phải hình thành nên một tổ.

Trường hợp các loại hình tương đối ít thì mỗi loại hình có thể hình thành nên 1 tổ, như khi phân tổ tổng thể nhân khẩu theo giới tính thì sẽ chia tổng thể đó thành 2 tổ là Nam và Nữ; hoặc phân tổ các doanh nghiệp theo thành phần kinh tế...

Trường hợp số loại hình thực tế nhiều, nếu coi mỗi loại hình là một tổ thì số tổ sẽ quá nhiều, không thể khái quát chung được và cũng không nêu rõ được sự khác nhau giữa các tổ, nên cần ghép những loại hình giống nhau hoặc gần giống nhau vào cùng một tổ. Chẳng hạn khi phân tổ tổng thể nhân khẩu theo nghề nghiệp, phân tổ các loại sản phẩm công, nông nghiệp, phân tổ các mặt hàng theo giá trị sử dụng, phân tổ các ngành kinh tế quốc dân..., số tổ thực tế có thể rất nhiều, có khi tới hàng nghìn, hàng vạn, nếu cứ phân chia tổng thể theo số tổ thực tế đó thì việc phân tổ gặp rất nhiều khó khăn và có thể không giúp gì được cho phân tích thống kê. Trong những trường hợp này phải giải quyết bằng cách ghép nhiều tổ nhỏ lại thành một số tổ lớn, theo nguyên tắc các tổ nhỏ ghép lại với nhau phải giống nhau (hoặc gần giống nhau) về tính chất, về giá trị sử dụng, về loại hình... Yêu cầu của việc ghép nhiều tổ nhỏ thành một số tổ lớn nhằm rút bớt số tổ thực tế quá nhiều, tạo điều kiện cho việc phân tổ được gọn và hợp lý. Trên thực tế, người ta thường tiến hành sắp xếp và trình bày trong những văn bản gọi là bảng phân loại hay bảng danh mục do Nhà nước quy định thống nhất và cố định trong một thời gian tương đối dài, nhằm bảo đảm tính chất so sánh được của tài liệu thống kê.

b. Phân tổ theo tiêu thức số lượng:

Khi phân tổ theo tiêu thức số lượng tuỳ theo lượng biến của tiêu thức thay đổi nhiều hay ít mà cách phân tổ được giải quyết khác nhau. Mặt khác, cũng cần chú ý đến số lượng đơn vị tổng thể nhiều hay ít mà xác định số tổ thích hợp.

- Trường hợp lượng biến của tiêu thức thay đổi ít, tức là sự biến thiên về mặt lượng giữa các đơn vị không chênh lệch nhiều lắm, biến động rời rạc và số lượng các lượng biến ít, như số người trong gia đình, số máy do một công nhân phụ trách... thì ở đây, số tổ có một giới hạn nhất định và thường cứ mỗi lượng biến là cơ sở để hình thành một tổ. Ví dụ: phân tổ công nhân trong một nhà máy dệt theo số máy do mỗi người phụ trách như sau:

Bảng 2.4 **Phân tổ công nhân theo số máy do mỗi người phụ trách**

Số máy dệt mỗi công nhân phụ trách (máy)	***Số công nhân (người)***
11	3
12	7
13	20
14	50
15	35
16	15
Cộng	**130**

Việc phân tổ trên đây rất đơn giản, vì lượng biến của tiêu thức phân tổ (số máy dệt) chỉ thay đổi trong phạm vi từ 11 đến 16 máy. Khi người công nhân đứng thêm được một máy, biểu hiện chất lượng công tác của họ đã thay đổi. Vì vậy, có thể căn cứ vào mỗi lượng biến để thành lập một tổ. Việc phân tổ như trên gọi là *phân tổ không có khoảng cách tổ*.

- Trường hợp lượng biến của tiêu thức biến thiên rất lớn, ta không thể áp dụng cách phân tổ nói trên được, nghĩa là không thể căn cứ vào mỗi lượng biến lập nên một tổ vì làm như vậy số tổ sẽ quá nhiều và không nói rõ sự khác nhau về chất giữa các tổ. Trong trường hợp này ta cần chú ý mối liên hệ giữa lượng và chất trong phân tổ, xét cụ thể xem lượng biến tích luỹ đến một mức độ nào đó thì chất của hiện tượng mới thay đổi và làm nảy sinh ra một tổ khác. Như vậy, mỗi tổ sẽ bao gồm một phạm vi lượng biến, với hai giới hạn: *giới han dưới* là lượng biến nhỏ nhất để làm cho tổ đó được hình thành, và *giới hạn trên* là lượng biến lớn nhất của tổ đó, nếu vượt quá giới hạn đó thì chất của tổ thay đổi và chuyển thành tổ khác. Trị số chênh lệch giữa giới hạn trên và giới hạn dưới của mỗi tổ gọi là *khoảng cách tổ*. Việc phân tổ theo các giới hạn như vậy gọi là *phân tổ có khoảng cách tổ*. Các khoảng cách tổ có thể đều nhau hoặc không đều nhau.

Chẳng hạn, theo tiêu thức "tiền lương" với đơn vị tính nghìn đồng của cán bộ công nhân viên trong một doanh nghiệp, có thể chia thành các tổ có khoảng cách tổ là :

< 500
Từ 500 đến dưới 1.000
Từ 1.000 đến dưới 1.500
Từ 1.500 đến dưới 2.000
Từ 2.000 đến dưới 2.500

Từ 2.500 đến dưới 3.000
Từ 3.000 trở lên

Trong trường hợp trên, lượng biến của tiêu thức tiền lương được sắp xếp thành 7 tổ, các tổ có khoảng cách tổ đều nhau là 500 nghìn đồng, tổ đầu tiên và tổ cuối cùng gọi là tổ có khoảng cách tổ mở.

Hoặc có tiêu thức "số lượng lao động" của 1 doanh nghiệp, có thể chia thành các tổ có khoảng cách tổ là:

1 – 100
101 – 200
201 – 500
501 – 1000
1001 – 3000

Trong trường hợp trên, các tổ có khoảng cách tổ không đều nhau.

Như vậy, cần phân biệt khi nào phân tổ theo khoảng cách tổ đều nhau và khi nào dùng khoảng cách tổ không đều nhau? Nói chung, việc xác định khoảng cách tổ đều nhau hay không đều nhau là phải căn cứ vào đặc điểm của hiện tượng nghiên cứu. Phải bảo đảm các đơn vị được phân phối vào mỗi tổ đều có cùng một tính chất và sự khác nhau về chất giữa các tổ đó. Trong thực tế, sự thay đổi về lượng của các bộ phận trong hiện tượng xã hội thường không diễn biến một cách đều đặn, bởi vì sự khác nhau về chất của chúng cũng không đều nhau, do vậy có nhiều trường hợp nghiên cứu, phải phân tổ theo khoảng cách tổ không đều nhau.

Riêng đối với các hiện tượng tương đối đồng nhất về mặt loại hình kinh tế xã hội và lượng biến trên các đơn vị thay đổi một cách tương đối đều đặn, có thể áp dụng việc phân tổ theo khoảng cách tổ đều nhau. Cách phân tổ này tạo điều kiện dễ dàng cho việc vận dụng các công thức toán học và để trình bày số liệu trên các đồ thị thống kê.

Việc phân tổ theo khoảng cách tổ đều nhau tương đối đơn giản và trị số khoảng cách tổ được xác định theo công thức:

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

Trong đó: d – trị số khoảng cách tổ
x_{max} – lượng biến lớn nhất của tiêu thức phân tổ
x_{min} – lượng biến nhỏ nhất của tiêu thức phân tổ
n – số tổ định chia

Chẳng hạn, năng suất lao động trong một tháng của một doanh nghiệp cao nhất là 300 sản phẩm, thấp nhất là 200 sản phẩm. Chênh lệch là 300 - 200 = 100 sản phẩm. Dự kiến chia tổng thể lao động của doanh nghiệp thành 5 tổ, thì khoảng cách tổ sẽ bằng 100 : 5 = 20 sản phẩm.

Trên đây là lý luận về xác định số tổ cần thiết và khoảng cách tổ đối với các trường hợp phân tổ. Nói chung, khi tiến hành phân tổ cần chú ý sắp xếp làm sao cho số tổ đặt ra không quá nhiều hay quá ít, gây khó khăn cho việc nghiên cứu. Nếu số tổ quá nhiều, tổng thể bị xé lẻ, số đơn vị tổng thể bị phân tán vào nhiều tổ có tính chất giống nhau hoặc gần giống nhau; ngược lại, nếu số tổ quá ít thì các đơn vị có tính chất khác nhau sẽ được phân phối vào cùng một tổ, điều đó làm cho mọi kết luận rút ra sẽ kém chính xác. Mặt khác, cũng cần bảo đảm phân phối cho mỗi tổ một số lượng đơn vị cần thiết. Có như vậy, việc phân tích đặc trưng và mối liên hệ giữa các loại hình mới có ý nghĩa. Tuy nhiên, cũng không nên loại trừ những trường hợp đặc biệt, khi cần phân tổ để vạch rõ những đơn vị điển hình tiên tiến. Các đơn vị này khi mới phát sinh tuy chỉ chiếm một bộ phận nhỏ trong toàn bộ, nhưng lại có ý nghĩa rất lớn đối với việc động viên, thúc đẩy phong trào chung.

1.2.2.3 Phân phối các đơn vị vào từng tổ

Sau khi xác định số tổ và khoảng cách tổ, bước cuối cùng là phân phối các đơn vị vào từng tổ và tính toán trị số của các chỉ tiêu giải thích (nếu có). Việc phân phối các đơn vị vào từng tổ căn cứ vào lượng biến của từng đơn vị tổng thể, vào số tổ và khoảng cách tổ đã xác định ở trên. Số lượng đơn vị của từng tổ nhiều hay ít, phân phối theo dạng nào là cơ sở để biểu hiện và phân tích đặc điểm cơ bản của hiện tượng cũng như tính toán các chỉ tiêu giải thích có liên quan hoặc các chỉ tiêu phản ánh bản chất của hiện tượng.

Các chỉ tiêu giải thích được tính toán cho từng tổ và chung trên cơ sở số lượng các đơn vị trong từng tổ. Tuỳ theo các chỉ tiêu đó là chỉ tiêu tuyệt đối, tương đối hay bình quân mà xác định phương pháp tổng hợp hay tính toán cho phù hợp.

1.2.3. Phân bố tần số và tần số tích lũy

1.2.3.1 Bảng phân bố tần số và tần số tích lũy

Sau khi phân tổ tổng thể theo một tiêu thức số lượng nào đó, các đơn vị tổng thể được phân phối vào trong các tổ và ta sẽ có một phân bố thống kê theo tiêu thức đó và được biểu diễn thành *bảng phân bố tần số và tần số tích lũy*.

Bảng phân bố tần số và tần số tích lũy có nhiều tác dụng trong nghiên cứu thống kê. Người ta thường đùng các bảng phân bố này để khảo sát tình hình phân phối các đơn vị tổng thể theo một tiêu thức nghiên cứu, qua đó thấy được kết cấu của tổng thể và sự biến động kết cấu đó. Thí dụ: để khảo sát đặc điểm phân phối một tổng thể lao động theo mức lương người ta xây dựng một bảng phân bố tần số cho số lao động theo tiêu thức tiền lương.... Bảng phân bố tần số còn được dùng để tính ra nhiều chỉ tiêu nêu lên các đặc trưng của từng tổ và của tổng thể, biểu hiện mối liên hệ giữa các bộ phận hoặc giữa các tiêu thức.

Một bảng phân bố tần số và tần số tích lũy gồm các thành phần chủ yếu sau:

- Thành phần thứ nhất là *lượng biến:* Lượng biến là các trị số nói lên biểu hiện cụ thể của tiêu thức số lượng, thường được ký hiệu là x_i .

Khi phân tổ theo tiêu thức số lượng có lượng biến rời rạc (là lượng biến chỉ có các biểu hiện bằng số nguyên, như: lượng biến của tiêu thức độ tuổi, số học sinh trong một lớp học, số lao động trong 1 doanh nghiệp…) thì bảng phân bố tần số có thể có khoảng cách tổ hoặc không có khoảng cách tổ. Nếu lượng biến của tiêu thức nghiên cứu biến thiên ít và chỉ có một vài trị số (như số nhân khẩu trong một gia đình, số máy dệt do mỗi công nhân phụ trách...) thì dãy số lượng biến không cần có khoảng cách tổ. Nếu lượng biến của dãy số này biến thiên trong phạm vi lớn (như số lao động của các xí nghiệp, số học sinh của các trường học...) thì bảng tần số cần phải có khoảng cách tổ. Trong trường hợp này giới hạn trên của tổ đứng trước và giới hạn dưới của tổ kế tiếp sau có thể khác nhau, chẳng hạn như khi phân tổ các doanh nghiệp theo số lượng lao động, giả sử có các tổ: từ 1 đến 100, từ 101 đến 500, từ 501 đến 1000, từ 1001 đến 2000 người... Ở đây giới hạn trên của tổ đứng trước và giới hạn dưới của tổ đứng liền sau không giống nhau về trị số.

Khi phân tổ theo tiêu thức số lượng có lượng biến liên tục (là lượng biến có thể được biểu hiện bằng những trị số bất kỳ cả số nguyên và số thập phân, như: lượng biến của tiêu thức năng suất thu hoạch lúa (đơn vị tính tạ/ha), tỷ lệ % hoàn thành kế hoạch…) thì bảng phân bố tần số theo tiêu thức nghiên cứu phải có khoảng cách tổ, bởi vì không thể căn cứ vào mỗi lượng biến bất kỳ để xác định một tổ, mà cần phải có một phạm vi lượng biến nhất định. Chẳng hạn như khi phân tổ lao động trong một doanh nghiệp theo mức lương, phân tổ các đơn vị sản xuất theo tỷ lệ % hoàn thành kế hoạch, phân tổ các hợp tác xã theo năng suất thu hoạch... Ở đây, các lượng biến liên tục cho nên phải phân tổ có khoảng cách tổ. Trong trường hợp này, giới hạn trên của tổ đứng trước và giới hạn dưới của tổ kế tiếp có thể giống nhau về trị số. Chẳng hạn, khi phân tổ các doanh nghiệp theo tỷ lệ % hoàn thành kế hoạch, giả sử có các tổ: dưới 80%, từ 80% đến 90%, từ 90% đến 100%, từ 100% đến 110%, từ 110% đến 120%... thì ở đây ta thấy một lượng biến nào đó có thể vừa là giới hạn trên của tổ này, lại vừa có thể là hạn dưới của tổ khác (như trong tổ thứ ba và tổ thứ tư, lượng biến 100% là giới hạn chung của cả 2 tổ). Vấn đề đặt ra là: nếu có một xí nghiệp hoàn thành đúng 100% kế hoạch, thì nên xếp vào tổ thứ ba hay tổ thứ tư? Ta thấy rằng trong một dãy số phân phối có lượng biến liên tục, việc sắp xếp của tổ có giới hạn trùng nhau như trên là hợp lý và cần thiết vì nó bảo đảm không còn một chỗ trống nào giữa các tổ. Mặt khác, cách sắp xếp như trên nói lên rằng mỗi tổ phải bao gồm giới hạn dưới của khoảng cách tổ, là lượng biến tối thiểu để cho tổ đó được hình thành. Như vậy, có nghĩa là doanh nghiệp nào hoàn thành đúng 100% kế hoạch phải được xếp vào tổ thứ tư (từ 100% đến 110%), là tổ bao gồm các doanh nghiệp hoàn thành và hoàn thành vượt mức kế hoạch. Đây chỉ là vấn đề có tính chất quy ước, phù hợp với tính chất của hiện tượng nghiên cứu. Để quy ước thống nhất cho mọi đối tượng sử dụng tài liệu có thể dùng khoảng cách tổ mở để biểu diễn ý nghĩa đó, khi đó cần lưu ý rằng dấu "=" chỉ nằm ở một cận dưới hoặc trên tuỳ theo đặc điểm của hiện tượng ý tưởng của người nghiên cứu.

- Thành phần thứ hai của dãy số lượng biến là *tần số*. Tần số là số đơn vị được phân phối vào trong mỗi tổ, tức là số lần một lượng biến nhận một trị số nhất định trong một tổng thể. Tần số thường được ký hiệu bằng f_i và $\sum f_i$ là tổng tần số hay tổng số đơn vị của tổng thể.

Khi tần số được biểu hiện bằng số tương đối gọi là tần suất, với đơn vị tính là lần hoặc % và ký hiệu bằng d_i ($d_i = f_i / \sum f_i$). Tần suất biểu hiện tỷ trọng của từng tổ trong tổng thể, vì vậy tổng tần suất ($\sum d_i$) sẽ bằng 1 nếu tính theo đơn vị lần và bằng 100 nếu tính theo đơn vị %. Trong phân tích thống kê, tần suất cho phép phân tích đặc điểm cấu thành của tổng thể nghiên cứu và quan sát sự biến động tần suất qua thời gian cho thấy xu hướng biến động về kết cấu của hiện tượng theo tiêu thức đang nghiên cứu. Với tác dụng đó nó thường được sử dụng trong việc phân tích chuyển dịch cơ cấu như phân tích chuyển dịch cơ cấu kinh tế, cơ cấu sản phẩm...

Ngoài hai thành phần trên, người ta thường tính tần số (hoặc tần suất) tích lũy tức là cộng dồn tần số (hoặc tần suất). Tần số tích luỹ (ký hiệu là S_i) cho biết số đơn vị có lượng biến lớn hơn hoặc nhỏ hơn một lượng biến cụ thể nào đó và là cơ sở để xác định một đơn vị đứng ở vị trí nào đó trong dãy số có lượng biến là bao nhiêu.

Trường hợp bảng phân bố tần số có các khoảng cách tổ không bằng nhau thì tần số của các tổ không thể so sánh được với nhau vì các trị số đó phụ thuộc vào trị số khoảng cách tổ. Khi đó người ta thường tính mật độ phân phối - là tỷ số giữa tần số và trị số khoảng cách tổ - và ký hiệu là m_i ($m_i = f_i / h_i$).

Tóm lại, có thể biểu diễn các thành phần của bảng phân bố tần số như sau:

Lượng biến (x_i)	Tần số (f_i)	Tần suất (d_i)	Tần số tích luỹ (S_i)
x_1	f_1	d_1	f_1
x_2	f_2	d_2	$f_1 + f_2$
x_3	f_3	d_3	$f_1 + f_2 + f_3$
...	...	…	….
x_{n-1}	f_{n-1}	d_{n-1}	$f_1 + f_2 + … + f_{n-1}$
x_n	f_n	d_n	$f_1 + f_2 + … + f_{n-1} + f_n$
Cộng	$\sum f_i$	$\sum d_i$ = 1 hoặc 100%	

1.2.3.2. Thí dụ về phân tổ và bảng phân bố tần số

Trở lại thí dụ 2 ở mục 1.1 về mức lương của 30 lập trình viên, có thể phân tổ số người theo theo tiêu thức tiền lương bằng hai cách sau:

- *Phân tổ không có khoảng cách tổ* : Là liệt kê tất cả các mức lương và số người có ở từng mức lương đó. Kết quả như sau:

Mức lương	*Số người*	*Mức lương*	*Số người*
1400	1	1800	7

1500	3	1850	2
1550	1	1900	2
1600	3	2050	1
1650	2	2100	2
1700	4	2200	1
1750	1		

Với tổng thể nghiên cứu chỉ có 30 đơn vị mà có tới 13 mức lương khác nhau tương ứng với 13 tổ, vì vậy kết quả phân tổ trên chỉ dừng lại ở việc nhận xét mức lương nào là phổ biến nhất chứ chưa nêu lên đặc điểm phân phối của số lao động theo mức lương.

- *Phân tổ có khoảng cách tổ*: Với số đơn vị nghiên cứu không nhiều (30), căn cứ vào biểu đồ thân lá đã xây dựng ở trên, có thể chia thành 4 tổ bằng cách ghép 2 trị số của phần thân vào thành một tổ. Cụ thể: Tổ thứ nhất sẽ gồm 2 trị số của phần thân là 14 và 15, tạo nên một tổ có khoảng cách tổ là từ 1400 đến dưới 1600 ng.đ, và tần số người có ở mức lương đó (tương ứng là số lá ở 2 phần thân đó)…. Kết quả có bảng phân bố tần số như sau:

Bảng 2.5 **Phân tổ số lao động theo tiêu thức tiền lương**

Tiền lương *(ng.đ)*	**Trị số giữa** *(ng.đ)*	**Tần số** *(Số người)*	**Tần suất** *(%)*
Từ 1400 đến dưới 1600	1500	5	16,67
Từ 1600 đến dưới 1800	1700	10	33,33
Từ 1800 đến dưới 2000	1900	11	33,67
Từ 2000 đến dưới 2200	2100	4	13,33
Cộng		***30***	***100,00***

Qua bảng phân bố tần số và tần số tích lũy cho thấy: có tới 21/30 người (chiếm 79% trong tổng số) có mức lương nằm trong khoảng từ 1600 ng.đ đến 2000 ng.đ, sự phân bố rất tập trung ở khoảng giữa trong đó phần chiếm tỷ trọng lớn nhất là số người có mức lương từ 1800 ng.đ đến 2000 ng.đ (33,67%). Trên cơ sở bảng phân bố tần số này còn có thể vận dụng các phương pháp thống kê khác.

Cũng với lý thuyết trên còn có thể phân tổ theo hai, ba …tiêu thức để biểu diễn mối liên hệ giữa các tiêu thức. Hãy xét thí dụ sau:

Thí dụ 3: Phân tổ theo hai tiêu thức với tình huống sau:

Ông Giám đốc muốn biết là thực tế lương có được trả theo thâm niên công tác không. Ông ta tổ chức một cuộc điều tra 30 lập trình viên làm cho các công ty cạnh tranh có từ 1 năm đến 10 năm kinh nghiệm. Kết quả điều tra như sau: *(Đơn vị tính: nghìn đồng/tháng)*

Mức lương	**Năm kinh nghiệm**	**Mức lương**	**Năm kinh nghiệm**	**Mức lương**	**Năm kinh nghiệm**
1500	2	2700	4	1800	2

1800	2	2100	3	1800	2
7650	7	1600	2	1750	1
4600	5	1200	2	2050	3
1500	1	5500	4	1700	1
3800	6	2850	3	1850	2
2700	4	1800	2	1600	2
3650	5	2200	2	1900	3
5700	7	3100	4	1800	2
2400	3	2900	3	7800	8

Theo dữ liệu trên, chúng ta phân tổ 30 người theo hai tiêu thức là mức lương và số năm kinh nghiệm. Việc phân tổ được tiến hành lần lượt theo từng tiêu thức năm kinh nghiệm và mức lương, trong đó số năm kinh nghiệm chia thành 4 tổ và mức lương chia thành 7 tổ . Kết quả phân tổ như sau:

Bảng 2.6 **Phân tổ lao động theo số năm kinh nghiệm và mức lương**

Năm KN / Mức lương	1-2	3-4	5-6	≥ 7	Cộng
1000 - 2000	13	1			**14**
2000 - 3000	1	7			**8**
3000 - 4000		1	2		**3**
4000 - 5000			1		**1**
5000 - 6000		1		1	**2**
6000 - 7000					
7000 - 8000				2	**2**
Cộng	**14**	**10**	**3**	**3**	**30**

Qua bảng phân tổ trên có thể nhận thấy tần số phân bố tập trung vào đường chéo của bảng và có thể nhận xét: khi số năm kinh nghiệm tăng thì tiền lương có xu hướng tăng theo hay nói cách khác tiền lương được trả có phụ thuộc vào số năm kinh nghiệm.

1.2.3.3. Đồ thị biểu diễn phân bố tần số và tần số tích lũy

Để biểu diễn phân bố tần số người ta thường dùng biểu đồ tần hình cột (Histogram) và biểu đồ tần số đa giác (Polygon). Đây là 2 cách biểu diễn khác nhau của cùng một dữ liệu. Đặc điểm của Histogram là giữa các cột không có khoảng cách mà là giới hạn giữa 2 tổ, độ cao thấp của các cột biểu thị tần số của mỗi tổ và độ rộng của cột là khoảng cách tổ. trục hoành ghi trị số giữa của các tổ, trục tung biểu diễn tần số của các tổ. Biểu đồ đa giác là một đường gấp khúc nối các điểm giữa đỉnh các cột của histogram.

Đồ thị tần số tích luỹ (Ogive) là đồ thị biểu diễn tần số (hoặc tần suất) cộng dồn của các tổ, đây cũng là một trong các dạng biểu diễn đặc điểm phân phối của dữ liệu và có thể

giúp ta ước lượng số đơn vị (hoặc tỷ lệ % số đơn vị) có lượng biến nhỏ hơn hay lớn hơn một lượng biến cụ thể nào đó.

Trở lại thí dụ 1(mục 1.1 ở trên), dữ liệu đã được sắp xếp theo thứ tự:

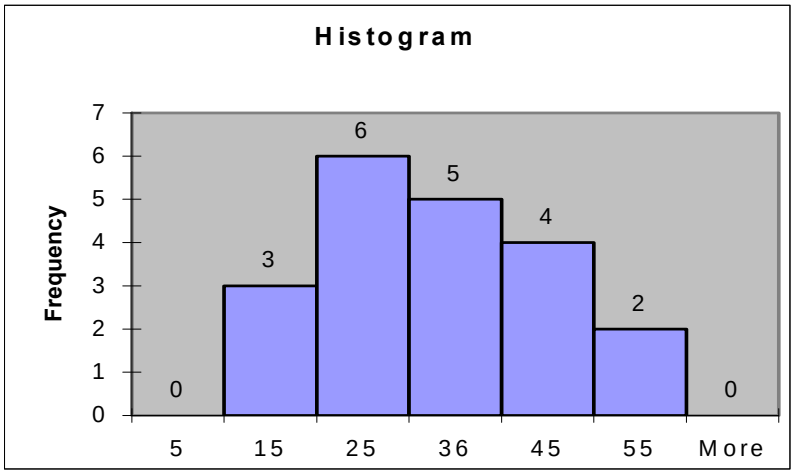
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Trên cơ sở biểu đồ thân lá, bảng phân bố tần số và tần suất như sau:

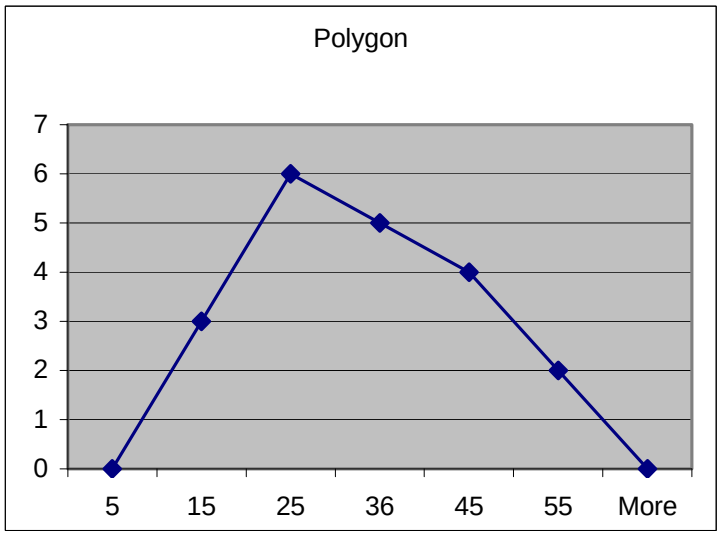
Các tổ	Trị số giữa	Tần số	Tần suất (lần)	Tỷ lệ %
10 - 20	15	3	0,15	15
20 - 30	25	6	0,30	30
30 - 40	35	5	0,25	25
40 - 50	45	4	0,20	20
50 - 60	55	2	0,10	10
Cộng		20	1	100

Biểu diện bảng phân bố tần số và tần số tích lũy trên bằng các đồ thị như sau:

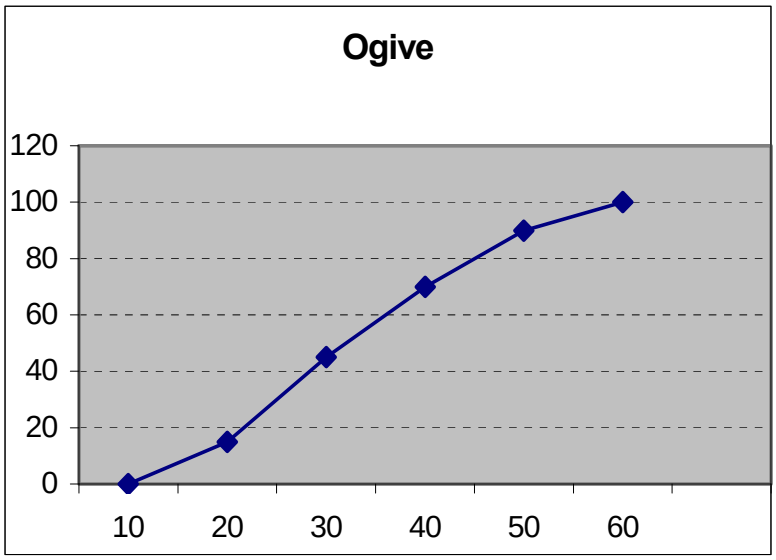
Hình 1: ***Biểu đồ tần hình cột***



Hình 2: ***Biểu đồ tần số đa giác***



Hình 3: ***Đồ thị tần số tích lũy***



1.3. Phân tổ lại.

1.3.1. Khái niệm, ý nghĩa của phân tổ lại.

Trong nghiên cứu thống kê, đôi khi phải tiến hành phân tổ lại các tài liệu thống kê đã được phân tổ. Phân tổ lại là lập ra một số tổ mới trên cơ sở các tổ cũ đã có sẵn từ trước, nhằm đáp ứng một mục đích nghiên cứu nào đó. Phân tổ lại được áp dụng trong các trường hợp sau đây:

- Các tài liệu trước được phân tổ không thống nhất với nhau về số tổ và khoảng cách tổ, làm cho việc so sánh gặp khó khăn.

- Các tài liệu trước được phân thành nhiều tổ nhỏ, mà các tổ này chưa phản ánh rõ được các loại hình kinh tế xã hội. Cần phân tổ lại bằng cách kết hợp nhiều tổ nhỏ ban đầu nhằm nêu rõ các loại hình.

- Các tài liệu phân tổ cũ chưa hợp lý, không phản ánh đúng đắn tình hình thực tế.

Khi tiến hành phân tổ lại, thường vẫn sử dụng tiêu thức phân tổ cũ. Nếu muốn so sánh đối chiếu một vài phân tổ cũ, có thể lấy một trong những phân tổ cũ làm chuẩn, tức là giữ nguyên không thay đổi, còn các phân tổ khác phải được phân tổ lại cho phù hợp. Cũng có trường hợp các phân tổ cũ đều không thoả mãn mục đích nghiên cứu và đều phải được phân tổ lại theo mẫu thống nhất.

1.3.2. Phương pháp phân tổ lại.

Có hai phương pháp phân tổ lại:

1.3.2.1. Lập các tổ mới bằng cách thay đổi các khoảng cách tổ của phân tổ cũ

Với phương pháp này việc thay đổi các khoảng cách tổ được thực hiện bằng cách mở rộng hoặc thu hẹp các khoảng cách tổ cũ. Ta hãy xét thí dụ về tài liệu phân tổ lao động theo thâm niên công tác của hai doanh nghiệp thuộc cùng một ngành sản xuất trong năm 2005, biểu hiện ở bảng sau:

Bảng 2.7 ***Phân tổ lao động theo thâm niên năm 2005***

Doanh nghiệp A			Doanh nghiệp B		
Phân tổ lao động theo thâm niên (năm)	Tỷ lệ % trong tổng số về		Phân tổ lao động theo thâm niên (năm)	Tỷ lệ % trong tổng số về	
	Lao động	Tiền lương		Lao động	Tiền lương
Dưới 2	15	10	Dưới 1	8	5
2 – 5	20	16	1 – 2	10	8
5 – 10	30	30	2 – 5	22	22
10 – 15	20	24	5 – 7	26	27
15 – 20	10	12	7 – 10	20	20
20 trở lên	5	8	10 – 15	8	10
			15 – 20	4	5
			20 trở lên	2	3
Tổng cộng	100	100	Tổng cộng	100	100

Muốn so sánh kết cấu lao động theo thâm niên công tác của hai doanh nghiệp trên đây, cần phải phân tổ lại cho thống nhất. Theo phương pháp thứ nhất, có thể thay đổi các khoảng cách tổ của hai phân tổ nói trên cho phù hợp với mục đích nghiên cứu. Giả sử định phân tổ hai tổng thể lao động trên đây thành 5 tổ có khoảng cách tổ đều nhau: dưới 5 năm, 5 đến 10 năm, 10 đến 15 năm, 15 đến 20 năm, và 20 năm trở lên. Như vậy, đối với doanh nghiệp A chỉ cần kết hợp 2 tổ đầu tiên vào với nhau. Còn đối với doanh nghiệp B, phải kết hợp 3 tổ đầu tiên vào với nhau (dưới 5 năm). Các tổ thứ 4 và thứ 5 cũng được kết hợp với nhau (5 đến 10 năm). Các chỉ tiêu về tỷ lệ phần trăm công nhân và tiền lương của hai xí nghiệp cũng được tính theo cách kết hợp nói trên. Ta có bảng phân tổ lại như sau:

Bảng 2.8 ***Phân tổ lao động theo thâm niên năm 2005***

Phân tổ công nhân theo thâm niên (năm)	Doanh nghiệp A		Doanh nghiệp B	
	Tỷ lệ % trong tổng số về		Tỷ lệ % trong tổng số về	
	Công nhân	Tiền lương	Công nhân	Tiền lương
Dưới 5	35	26	40	35

5 – 10	30	30	46	47
10 – 15	20	24	8	10
15 – 20	10	12	4	5
20 trở lên	5	8	2	3
Tổng cộng	100	100	100	100

Với tài liệu phân tổ lại ta có thể dễ dàng đối chiếu, so sánh và phân tích tình hình lao động của 2 doanh nghiệp trên.

1.3.2.2. Lập các tổ mới theo tỷ trọng của mỗi tổ chiếm trong tổng thể

Phương pháp thứ hai được tiến hành bằng cách xác định các tổ mới theo tỷ trọng mỗi tổ chiếm trong tổng thể. Ta xét thí dụ: Có tài liệu phân tổ các trường học của một tỉnh theo số học sinh như sau:

Bảng 2.9 ***Phân tổ các trường học theo số học sinh***

Phân tổ các trường theo số học sinh (hs)	Tỷ lệ % chiếm trong tổng số về :		
	Số trường	Số giáo viên	Số lớp học
500 trở xuống	4,0	1,8	1,4
501 – 700	6,0	3,2	2,8
701 – 900	15,0	10,1	9,5
901 – 1100	18,0	16,8	16,2
1101 – 1300	27,0	27,2	27,6
1301 – 1500	15,0	16,8	17,7
1501 – 1700	8,0	11,1	11,1
Trên 1700	7,0	15,0	13,7
Cộng	100,0	100,0	100,0

Bây giờ ta cần phân tổ lại số trường học nói trên thành 3 tổ: trường loại nhỏ, trường loại trung bình, trường loại lớn. Theo những tỷ lệ đã được xác định từ trước của tỉnh này, số trường loại nhỏ chiếm 35% tổng số trường, loại trung bình chiếm 50%, còn loại lớn chiếm 15%. Ta sẽ tính toán như sau:

+ *Tổ mới thứ nhất* – trường nhỏ, gồm 35% tổng số trường, sẽ bao gồm toàn bộ số trường của 3 tổ cũ đầu tiên, tức là 4 + 6 + 15 = 25(%), và còn phải lấy thêm 10% của tổ cũ thứ tư nhập vào cho đủ 35%. Từ đó tính:

Tỷ lệ % chiếm trong tổng số giáo viên của tổ mới thứ nhất là:

$$1,8 + 3,2 + 10,1 + \frac{16,8 \times 10}{18} = 24,3\ (\%);$$

Tỷ lệ % chiếm trong tổng số lớp học là:

$$1,4 + 2,8 + 9,5 + \frac{16,2 \times 10}{18} = 22,7\ (\%)$$

+ *Tổ mới thứ hai* – trường trung bình, gồm 50% tổng số trường, sẽ bao gồm 8% số trường còn lại của tổ 4 cũ và các trường của tổ 5 và 6 cũ, tức là 8 + 27 + 15 = 50 (%0. Từ đó tính:

Tỷ lệ % chiếm trong tổng số giáo viên của tổ mới thứ hai là:

$$\frac{16,8 \text{ x } 8}{18} + 27,2 + 16,8 = 51,6\ (\%)$$

Tỷ lệ % chiếm trong tổng số lớp học là:

$$\frac{16,2 \times 8}{18} + 27,6 + 17,7 = 52,5\ (\%)$$

+ *Tổ mới thứ ba* – trường lớn, gồm 15% tổng số trường, sẽ tính theo các tỷ lệ % còn lại:

Tỷ lệ % chiếm trong tổng số giáo viên của tổ mới thứ hai là:

$$11,1 + 13,0 = 24,1\ (\%)$$

Tỷ lệ % chiếm trong tổng số lớp học là:

$$11,1 + 13,7 = 24,8\ (\%)$$

Cuối cùng ta sẽ có kết quả như sau:

Bảng 2.10 ***Phân tổ các trường học theo quy mô***

Phân tổ các trường theo quy mô	Tỷ lệ % chiếm trong tổng số về :		
	Số trường	Số giáo viên	Số lớp học
Loại nhỏ	35,0	24,3	22,7
Loại trung bình	50,0	51,6	52,5
Loại lớn	15,0	24,1	24,8
Cộng	100,0	100,0	100,0

Phương pháp phân tổ lại trên đây tương đối phức tạp, và phải dựa trên một số tính toán giả thiết. Tuy nhiên, trong những điều kiện tài liệu hạn chế, phương pháp này cũng giúp ta nghiên cứu được vấn đề.

1.4. Phân tổ nhiều chiều

1.4.1 Khái niệm, tác dụng của phân tổ nhiều chiều

Phân tổ nhiều chiều là cùng một lúc phân tổ theo nhiều tiêu thức có vai trò như nhau trong việc đánh giá hiện tượng. Phân tổ nhiều chiều có các tác dụng sau:

+ Nghiên cứu kết cấu của tổng thể theo một số tiêu thức cơ bản của mối liên hệ với nhau. Thí dụ: Nghiên cứu kết cấu các doanh nghiệp trong cùng một ngành sản xuất theo giá trị sản xuất, số lượng lao động, giá trị thiết bị sản xuất, giá thành đơn vị sản phẩm, năng suất lao động, lợi nhuận…

+ Dùng phân tổ nhiều chiều để nghiên cứu mối liên hệ giữa nhiều tiêu thức mà khi dùng phân tổ kết hợp không giải quyết được, chẳng hạn như việc sắp xếp thứ tự phân tổ theo tiêu thức nào trước và tiêu thức nào sau là không có ý nghĩa hoặc khi có nhiều tiêu thức nguyên nhân cùng tác động đến một tiêu thức kết quả.

+ Xây dựng tài liệu đồng nhất của thông tin ban đầu để vận dụng các phương pháp thống kê: phân tích tương quan, phân tích phương sai….

+ Trường hợp dựa vào những căn cứ chung ở các phần trên mà vẫn không phân tổ được.

1.4.2 Tiêu thức phân tổ trong phân tổ nhiều chiều

Trong phân tổ nhiều chiều, các tiêu thức nguyên nhân đồng thời làm tiêu thức phân tổ, vì vậy người ta phải đưa các *tiêu thức phân tổ* về dạng một ***tiêu thức tổng hợp*** rồi căn cứ vào tiêu thức tổng hợp này để phân tổ như phân tổ theo 1 tiêu thức.

Nếu gọi các lượng biến của các tiêu thức phân tổ là x_{ij} ($i = \overline{1,n}$, $j = \overline{1,k}$) trong đó: i là thứ tự của lượng biến; j là thứ tự của tiêu thức; tiêu thức tổng hợp được tính như sau:

- Đưa các lượng biến của các tiêu thức (vốn khác nhau) về dạng tỷ lệ bằng cách lấy các lượng biến chia cho số bình quân của các lượng biến của mỗi tiêu thức:

$$P_{ij} = \frac{x_{ij}}{\overline{x_j}} \quad ; \qquad \text{Trong đó: } \overline{x_j} = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

- Sau đó cộng các P_{ij} có cùng thứ tự của tiêu thức, ta có: $\sum_{j=1}^{k} P_{ij}$ hoặc tính bình quân các tỷ số bằng cách lấy $\sum P_{ij}$ chia cho k.:

$$\overline{P_i} = \frac{\sum_{j=1}^{k} P_{ij}}{k}$$

Coi $\sum_{j=1}^{k} P_{ij}$ hoặc $\overline{P_i}$ là tiêu thức phân tổ.

- Ý nghĩa của tiêu thức tổng hợp:

Lượng biến của các tiêu thức khác nhau có trị số (khối lượng) và đơn vị tính toán khác nhau. Khi (nó được) đưa về dạng tỷ số, nó đã xoá bỏ được sự khác nhau đó. Vì vậy, mặc dù các tiêu thức khác nhau, nhưng các tỷ số của nó khi có cùng 1 trị số thì sẽ có vai trò như nhau trong việc biểu hiện tính chất của hiện tượng.

Người ta thường dùng bảng để tính tiêu thức tổng hợp như sau:

Bảng 2.11 ***Bảng tính tiêu thức tổng hợp***

STT của lượng biến	Tiêu thức thứ 1		Tiêu thức thứ 2		Tiêu thức thứ j…		Tiêu thức thứ k		$\sum_{j=1}^{k} P_{ij}$	$\overline{P_i} = \frac{\sum_{j=1}^{k} P_{ij}}{k}$
	x_{i1}	P_{i1}	x_{i2}	P_{i2}	x_{ij}	P_{ij}	x_{ik}	P_{ik}		
1										
2										
3										
.										
.										
.										
n										
$\sum x_{ij}$										
$\overline{x_j}$										

Thí dụ 4 : Có tài liệu của 10 doanh nghiệp cùng 1 ngành sản xuất như nhau

Bảng 2.12. ***Tình hình sản xuất của các doanh nghiệp năm 2005***

Thứ tự doanh nghiệp	Giá trị thiết bị SX chủ yếu (tỷ đồng)	Số lượng lao động	Giá trị sản xuất năm (tỷ đồng)
1	10,6	2730	191
2	0,6	200	10
3	5,9	3000	161
4	0,9	366	12
5	4,7	1210	62
6	3,5	990	30
7	0,8	880	27
8	4,3	960	48
9	7,3	1910	103
10	1,3	854	24

Để nghiên cứu mối quan hệ của 2 tiêu thức giá trị thiết bị sản xuất và số lượng lao động với giá trị sản xuất cần phân tổ các doanh nghiệp trên cùng một lúc theo cả 2 tiêu thức. Trước hết cần đưa 2 tiêu thức này về một tiêu thức tổng hợp với bảng tính sau:

Bảng 2.13 ***Bảng tính tiêu thức tổng hợp từ 2 tiêu thức giá trị thiết bị sản xuất và số lượng lao động***

TT doanh nghiệp	Giá trị TBSX chủ yếu		Số lượng lao động		$\sum_{j=1}^{k} P_{ij}$	$\overline{P_i} = \frac{\sum_{j=1}^{k} P_{ij}}{k}$	Giá trị sản xuất
	x_{i1}	$P_{i1} = \frac{x_{i1}}{\overline{x_1}}$	x_{i2}	$P_{i2} = \frac{x_{i2}}{\overline{x_2}}$			
1	10,6	2,6566	2730	2,0826	4,7392	2,3696	191
2	0,6	0,1503	200	0,1525	0,3028	0,1514	10
3	5,9	1,4786	3000	2,2886	3,7672	1,8836	161
4	0,9	0,2255	366	0,2792	0,5047	0,2524	12
5	4,7	1,1779	1210	0,9231	2,1010	1,0505	62
6	3,5	0,8771	990	0,7552	1,6323	0,8162	30
7	0,8	0,2005	880	0,6774	0,8779	0,4390	27
8	4,3	1,0776	960	0,7323	1,8099	0,9050	48
9	7,3	1,8295	1910	1,4571	3,2866	1,6433	103
10	1,3	0,3258	854	0,6515	0,9773	0,4887	24
$\sum$	39,9	10	13108	10	20	10	
$\overline{x_j}$	3,99		131,08				

Bây giờ có thể dùng $\overline{P_i}$ làm tiêu thức phân tổ

1.4.3. Phương pháp phân tổ nhiều chiều.

Trình tự các bước trong phân tổ nhiều chiều cũng theo các bước của quá trình phân tổ nói chung đã trình bày ở trên. Tuy nhiên vấn đề xác định số tổ và khoảng cách tổ theo tiêu thức tổng hợp phức tạp hơn. Cụ thể, có 2 cách xác định: Thứ nhất, có thể xác định số tổ và khoảng cách tổ như ở trên đã trình bày nhưng tiêu thức phân tổ lúc này là tiêu thức tổng hợp nên khó phân biệt lượng biến tích lũy tới mức nào thì chất mới thay đổi, vì vậy sau khi phân tổ phải kiểm tra tính đồng nhất và bền vững của các tổ. Thứ hai, xác định số tổ và khoảng cách tổ bằng phương pháp toán thông qua các hàm kiểm tra. Các hàm kiểm tra thường được dùng như sau:

a) Hàm kiểm tra tính đồng nhất.

Trong phân tổ thống kê, các đơn vị trong từng tổ phải giống nhau hoặc gần giống nhau về tính chất theo tiêu thức phân tổ vì vậy phải dùng hàm kiểm tra để *kiểm tra tính đồng nhất* của từng tổ. Hàm kiểm tra tính đồng nhất có dạng sau:

$$U(\rho^2) = \frac{n-1}{n(n-L)L} \times \frac{\left[(n-L)\sum_{s=1}^{L} x_s - L\sum_{s=L+1}^{n} x_s\right]^2}{\sum_{s=1}^{n} x_s^2 - \frac{1}{n}\left(\sum_{s=1}^{n} x_s\right)^2}$$

Trong đó: x_s: Lượng biến của tiêu thức trong 1 tổ ($s = \overline{1,n}$)

n: Số đơn vị trong 1 tổ

L: Thứ tự lượng biến trong 1 tổ (đã được sắp xếp theo thứ tự tăng dần)

$\sum_{s=1}^{L} x_s$: Lượng biến tích lũy tiến

$\sum_{s=L+1}^{n} x_s$: Lượng biến tích lũy lùi

Và $\sum_{s=L+1}^{n} x_s = \sum_{s=1}^{n} x_s - \sum_{s=1}^{L} x_s$

Hàm kiểm tra trên được tính cho tất cả các đại lượng ngẫu nhiên và được so sánh với tiêu chuẩn χ^2, cụ thể: $U(\rho^2) \leq \chi^2_{\alpha,m}$

Trong đó: α: mức ý nghĩa

m: bậc tự do

Nếu $U(\rho^2)$ của một trong tất cả các đại lượng ngẫu nhiên không thoả mãn điều kiện trên thì giả thiết về sự đồng nhất của tổng thể *không được thừa nhận* và khi đó, tổng thể kiểm tra được chia làm hai phần: Lấy giá trị $U(\rho^2)$ ***lớn nhất*** làm giới hạn của phần đầu và mỗi phần lại tiếp tục được kiểm tra theo tiêu chuẩn $U(\rho^2)$ (hay mỗi phần lại tính $U(\rho^2)$ riêng để tiến hành kiểm tra. Nếu phần nào đó thoả mãn điều kiện thì coi như tổ đó có tính đồng nhất, nếu không thoả mãn thì lại chia làm 2 và làm tương tự như trên.

Trở lại ví dụ trên, tiêu thức phân tổ bây giờ là tiêu thức tổng hợp $\overline{P_i}$, ta sắp xếp các lượng biến của tiêu thức tổng hợp theo thức tự tăng dần rồi tính toán giá trị của các $U(\rho^2)$ theo bảng sau:

Bảng 2.14 ***Bảng tính hàm kiểm tra tính đồng nhất*** $U(\rho^2)$

STT XN	$\overline{P_i}$	$\Sigma\overline{P_i}$ tích luỹ tiến	$\Sigma\overline{P_i}$ tích luỹ lùi	$\overline{P_i}^2$	L	n-L	$U(\rho^2)$
2	0,1514	0,1514	10	0,02292	1	9	1,4487
4	0,2524	0,4038	9,8486	0,06370	2	8	2,8836
7	0,4390	0,8428	9,5962	0,19272	3	7	4,0129
10	0,4887	1,3315	9,1572	0,23883	4	6	5,3739
6	0,8162	2,1477	8,6685	0,66618	5	5	5,8934
8	0,9050	3,0527	7,8523	0,81902	6	4	6,5548
5	1,0505	4,1032	6,9473	1,10355	7	3	***7,2363***
9	1,6433	5,7465	5,8968	2,70043	8	2	5,7472
3	1,8836	7,6301	4,2535	3,54795	9	1	3,7749
1	2,3696	10	2,3699	5,61500	10	0	-
	10			14,97025			

Với mức ý nghĩa 0,05 tr bảng ta có: $\chi^2_{0,05;1} = 3,841$

Qua bảng trên ta thấy có rất nhiều giá trị $U(\rho^2)$ không thỏa mãn điều kiện $U(\rho^2) \le \chi^2_{\alpha,m}$, vì vậy không thể gộp 10 doanh nghiệp thành 1 tổ.

Lấy $U(\rho^2)$ lớn nhất làm giới hạn tổ đầu, như vậy tổ đầu gồm 7 doanh nghiệp đầu, tổ thứ 2 gồm 3 doanh nghiệp còn lại.

Việc kiểm tra được tiến hành tương tự như trên với 2 tổ đó. Kết quả cuối cùng 10 doanh nghiệp trên được phân thành 3 tổ:

Tổ 1 gồm 4 doanh nghiệp số: 2, 4, 7 và 10
Tổ 2 gồm 3 doanh nghiệp số: 6, 8, 5
Tổ 3 gồm 3 doanh nghiệp số: 9, 3, 1

b) Hàm kiểm tra tính bền vững.

Sau khi dùng hàm kiểm tra để kiểm tra tính đồng nhất của từng tổ, người ta còn phải *kiểm tra sự bền vững* của các tổ đó. Hàm kiểm tra tính bền vững của giới hạn giữa hai tổ gần nhau có dạng sau:

$$U(s_k, s_{k+1}) = \frac{n_k + n_{k+1} - 1}{n_k . n_{k+1} . (n_k + n_{k+1})} \times \frac{\left(n_{k+1} . \sum_{s=1}^{n_k} x_s - n_k . \sum_{s=k+1}^{n_{k+1}} x_s \right)}{\sum_{s=k}^{k+1} x_s^2 - \frac{1}{n_k + n_{k+1}} \left(\sum_{s=k}^{k+1} x_s \right)^2}$$

Trong đó: k: số thứ tự của tổ

n_k: số đơn vị của tổ thứ k

n_{k+1}: số đơn vị của tổ tiếp theo tổ k

$\sum_{s=1}^{n_k} x_s$: Tổng lượng biến của tổ thứ k

$\sum_{s=k+1}^{n_{k+1}} x_s$: Tổng lượng biến của tổ thứ k+1

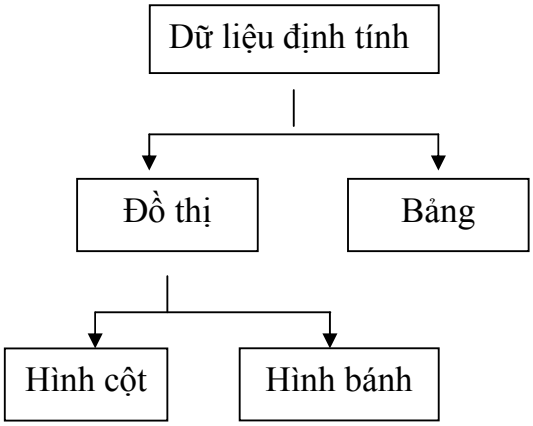
$\sum_{s=k}^{k+1} x_s^2$: Tổng của bình phương các lượng biến trong 2 tổ k và k+1

$\sum_{s=k}^{k+1} x_s$: Tổng các lượng biến trong 2 tổ k và k+1

Sau khi tính toán cũng đem so sánh với χ^2: Nếu $U(s_k, s_{k+1}) \le \chi^2_{\alpha,m}$ thì giới hạn giữa 2 tổ gần nhau ***không bền vững***. Vì vậy, 2 tổ đó nhập làm 1 và coi như ở tổ trên rồi lại tiếp tục tính toán với tổ tiếp theo.

2. Trình bày dữ liệu định tính.

Phương pháp mô tả dữ liệu định tính:



Dữ liệu định tính là dữ liệu về tiêu thức thuộc tính. Có thể trình bày theo một tiêu thức hoặc đồng thời hai, ba… tiêu thức. Kết quả của phân tổ theo tiêu thức thuộc tính gọi là dãy số phân phối theo tiêu thức thuộc tính (hay *dãy số thuộc tính*) phản ánh kết cấu của tổng thể theo một hay một số tiêu thức thuộc tính nào đó. Ví dụ: dãy số phân phối giá trị sản xuất công nghiệp theo ngành và theo thành phần kinh tế, dãy số phân phối các doanh nghiệp công nghiệp theo ngành... Có một số trường hợp tiêu thức thuộc tính chỉ có hai biểu hiện (tiêu thức thay phiên), do đó dãy số phân phối theo tiêu thức này chỉ có 2 tổ, chẳng hạn khi phân tổ tổng thể dân số theo tiêu thức giới tính thì dãy số phân phối chỉ có hai tổ nam và nữ... Để biểu diễn kết quả này có thể dùng bảng và đồ thị thống kê.

2.1 Trình bày dữ liệu định tính theo một tiêu thức:

Để trình bày dữ liệu định tính cũng có thể dùng bảng và đồ thị thống kê, tuy nhiên đặc điểm của các phương pháp biểu diễn này có khác so với việc biểu diễn dữ liệu định lượng.

- *Lập bảng biểu diễn dữ liệu theo một tiêu thức định tính*: đó là bảng giản đơn để tóm tắt dữ liệu, gồm 3 cột: *Thứ nhất*, cột biểu thị tiêu thức thuộc tính, bao gồm tên tiêu thức và các biểu hiện của tiêu thức; *Thứ 2*, cột số lượng phản ánh số lượng tương ứng với từng biểu hiện của tiêu thức; và thứ 3 là cột tỷ lệ % trong tổng số phản ánh kết cấu tổng thể theo tiêu thức thuộc tính đang nghiên cứu.

Chẳng hạn, có tài liệu như sau:

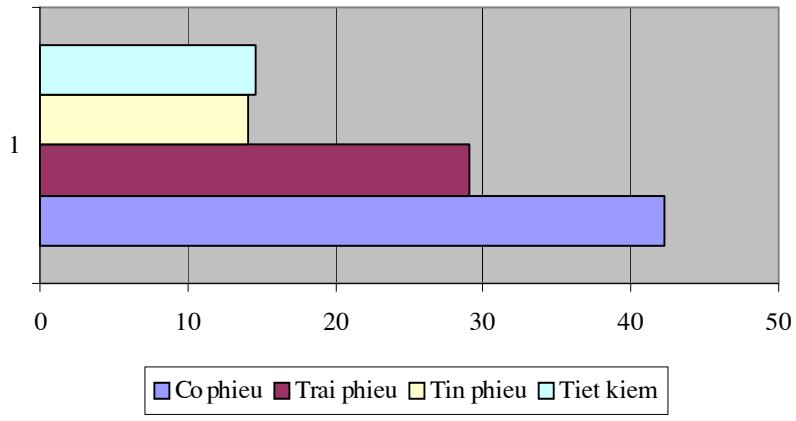
Bảng 2.15 ***Tình hình đầu tư của nhà đầu tư A năm 2007***

Danh mục đầu tư	***Số lượng*** *(1000$)*	***Tỷ lệ %***
Cổ phiếu	46,5	42,27
Trái phiếu	32,0	29,09
Tín phiếu	15,5	14,09
Tiết kiệm	16,0	14,55
Tổng	**110**	**100**

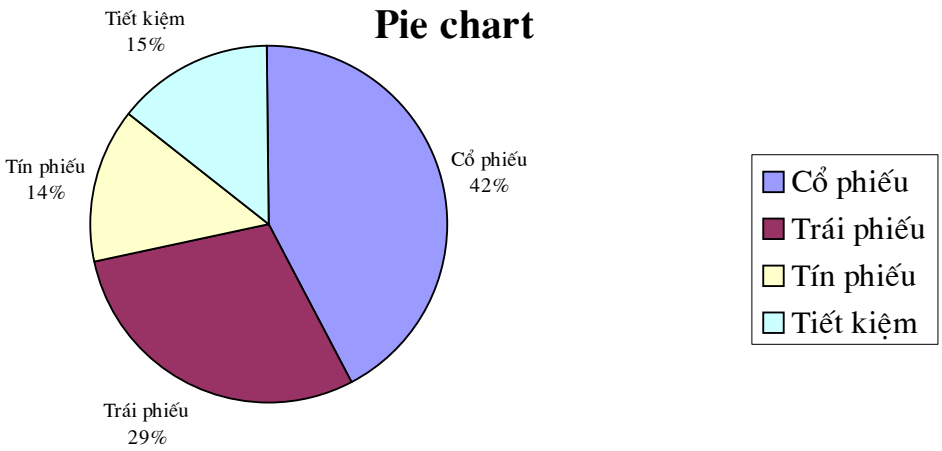
Bảng trên là một tóm tắt dữ liệu về tình hình đầu tư của nhà đầu tư A, trong đó không chỉ phản ánh tổng vốn đầu tư mà còn biểu diễn kết cấu vốn đầu tư theo danh mục đầu tư (một tiêu thức thuộc tính). Kết cấu này được biểu hiện bằng cả số tuyệt đối (số lượng vốn đầu tư) và số tương đối (tỷ lệ %). Qua đó cho thấy, cổ phiếu là loại được đầu tư nhiều nhất là 46,5 ngàn đô la Mỹ chiếm tỷ trọng lớn nhất trong tổng đầu tư (42,27%); tiếp theo là trái phiếu, sau cùng tín phiếu và tiết kiệm gần xấp xỉ nhau chỉ chiếm hơn 14% mỗi loại.....

- *Dùng đồ thị để biểu diễn dữ liệu theo một tiêu thức định tính*: Có thể dùng đồ thị thanh ngang (Bar chart) để biểu hiện số lượng từng tổ và đồ thị hình bánh (Pie chart) để biểu hiện kết cấu của tổng thể. Với dữ liệu trên có thể biểu diễn như sau:

Biểu hiện bằng Bar charts



Biểu hiện bằng Pie chart



2.2 Trình bày dữ liệu định tính theo hai tiêu thức:

- Lập bảng kết hợp để tóm tắt dữ liệu theo 2 tiêu thức định tính (một tiêu thức theo cột và một tiêu thức theo dòng). Chẳng hạn có tài liệu về tình hình đầu tư của các nhà đầu tư A, B, C như sau:

Bảng 2.15 ***Tình hình đầu tư của các nhà đầu tư A, B, C năm 2007***

Đơn vị tính: ngànUSD

Nhà đầu tư / ***Danh mục đầu tư***	***Nhà đầu tư A***	***Nhà đầu tư B***	***Nhà đầu tư C***	***Tổng đầu tư***
Cổ phiếu	46.5	55	27.5	***129***
Trái phiếu	32	44	19	***95***
Tín phiếu	15.5	20	13.5	***49***
Tiết kiệm	16	28	7	***51***
Tổng	**110**	**147**	**67**	**324**

Bảng trên biểu diễn dữ liệu theo 2 tiêu thức: danh mục đầu tư (theo cột) và nhà đầu tư (theo dòng). Qua đó cho thấy tình hình đầu tư nói chung và của từng nhà đầu tư nói riêng theo từng danh mục đầu tư, cụ thể: trong 3 nhà đầu tư, tổng số vốn đầu tư của nhà đầu tư B là lớn nhất và của nhà đầu tư C là nhỏ nhất; nhìn chung các nhà đầu tư đều dành đầu tư vào cổ phiếu nhiều hơn... Tuy nhiên bảng dữ liệu trên chưa so sánh được mức độ quan tâm tới các danh mục đầu tư của các nhà đầu tư (hay kết cấu vốn đầu tư theo danh mục đầu tư). Để phản ánh điều đó, từ bảng 2.15 có thể lập bảng sau:

Bảng 2.16 ***Kết cấu theo danh mục đầu tư của các nhà đầu tư A, B, C năm 2007***

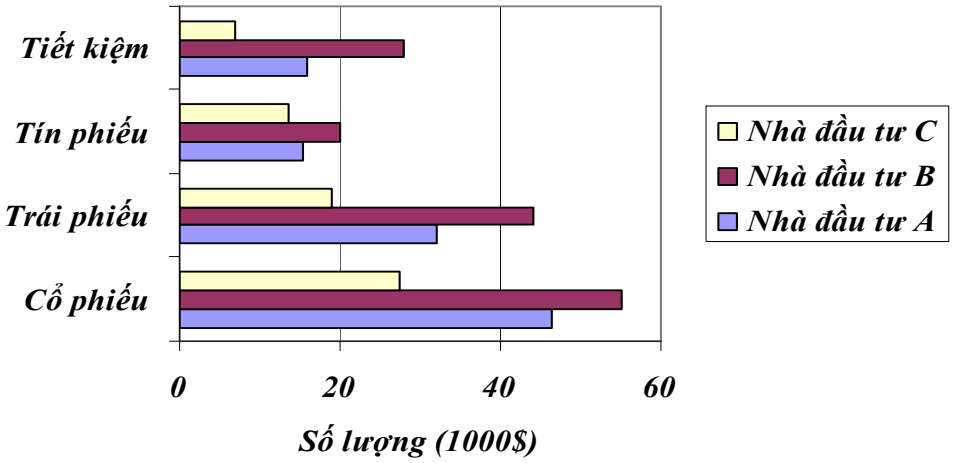
Đơn vị tính: %

Nhà đầu tư / ***Danh mục đầu tư***	***Nhà đầu tư A***	***Nhà đầu tư B***	***Nhà đầu tư C***	***Chungcho 3 nhà đầu tư***
Cổ phiếu	42.27	37.41	41.04	39.81
Trái phiếu	29.09	29.93	28.36	29.32
Tín phiếu	14.09	13.61	20.15	15.12
Tiết kiệm	14.55	19.05	10.45	15.74
Tổng	***100.00***	***100.00***	***100.00***	***100.00***

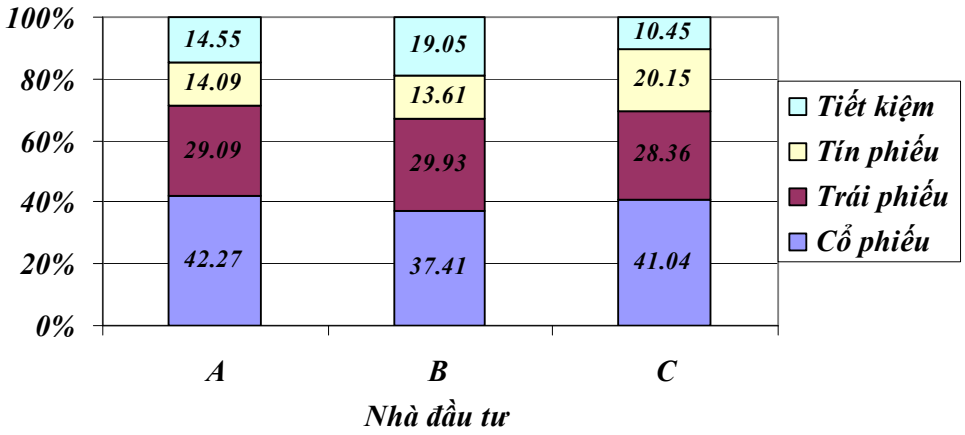
Bảng trên cho thấy: Danh mục cổ phiếu của cả 3 nhà đầu tư đều chiếm tỷ trọng lớn chứng tỏ các nhà đầu tư đều chú trọng vào đầu tư cổ phiếu, trong đó nhà đầu tư A có tỷ lệ cho đầu tư cổ phiếu lớn nhất. Trái phiếu được các nhà đầu tư quan âm như nhau (khoảng trên dưới 29%). Ngoài ra tín phiếu được nhà đầu tư C quan tâm nhiều hơn và với tiết kiệm thì nhà đầu tư B quan tâm nhiều hơn so với 2 nhà đầu tư còn lại...

- *Đồ thị để biểu diễn dữ liệu theo hai tiêu thức định tính*: Có thể dùng biểu đồ nhiều thanh ngang (Side by Side chart) để biểu hiện số lượng và đồ thị hình cột (columns) để biểu hiện kết cấu của tổng thể. Cụ thể:

Biểu đồ thanh ngang để biểu diễn số lượng vốn đầu tư:



Biểu đồ hình cột để biểu diễn kết cấu vốn theo danh mục đầu tư:



3. Kỹ thuật bảng thống kê và đồ thị thống kê

3.1. Bảng thống kê

Sau khi tổng hợp các tài liệu điều tra thống kê, muốn phát huy tác dụng của nó đối với giai đoạn phân tích thống kê, cần thiết phải trình bày kết quả tổng hợp theo một hình thức

thuận lợi nhất cho việc sử dụng sau này. Có thể trình bày các kết quả tổng hợp bằng các hình thức: bảng thống kê, đồ thị thống kê, bài viết...

3.1.1. Ý nghĩa và tác dụng của bảng thống kê

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

Bảng thống kê có nhiều tác dụng quan trọng trong mọi công tác nghiên cứu kinh tế nói chung và trong phân tích thống kê nói riêng. Các tài liệu trong bảng thống kê đã được sắp xếp lại một cách khoa học, nên có thể giúp ta tiến hành mọi việc so sánh đối chiếu, phân tích theo các phương pháp khác nhau, nhằm nêu lên sâu sắc bản chất của hiện tượng nghiên cứu. Nếu biết trình bày và sử dụng thích đáng các bảng thống kê, thì việc chứng minh vấn đề sẽ trở nên rất sinh động, có sức thuyết phục hơn cả những bài văn dài.

3.1.2. Cấu thành bảng thống kê

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

Các hàng ngang, cột dọc phản ánh quy mô của bảng thống kê vì số hàng và cột càng nhiều thì bảng thống kê càng lớn và phức tạp. Các hàng ngang và cột dọc cắt nhau tạo thành các ô dùng để điền các số liệu thống kê vào đó. Các hàng và cột thường được đánh số thứ tự để tiện cho việc sử dụng và trình bày vấn đề.

Tiêu đề của bảng thống kê phản ánh nội dung, ý nghĩa của bảng và của từng chi tiết trong bảng. Trước hết ta có tiêu đề chung, tức là tên gọi chung của bảng thống kê, thường được viết ngắn gọn, dễ hiểu và đặt ở phía trên đầu bảng thống kê. Còn các tiêu đề nhỏ (hay còn gọi là tiêu mục) là tên riêng của mỗi hàng ngang và cột dọc, phản ánh rõ nội dung, ý nghĩa của hàng và cột đó.

Các tài liệu con số, thu thập được do kết quả tổng hợp thống kê, được ghi vào các ô của bảng thống kê, mỗi con số phản ánh một đặc trưng về mặt lượng của hiện tượng nghiên cứu.

b. Về nội dung:

Bảng thống kê gồm 2 phần: phần chủ đề và phần giải thích. *Phần chủ đề* (còn gọi là phần chủ từ) nói lên tổng thể hiện tượng được trình bày trong bảng thống kê, tổng thể này được phân thành những đơn vị nào, bộ phận nào? Nó giải đáp vấn đề: đối tượng nghiên cứu của bảng thống kê là những đơn vị nào, những loại hình gì? Có khi phần chủ đề phản ánh các địa phương hoặc các thời gian nghiên cứu khác nhau của một hiện tượng nào đó. *Phần giải thích* (còn gọi là phần tân từ) gồm các chỉ tiêu giải thích các đặc điểm của đối tượng nghiên cứu, tức là giải thích phần chủ đề của bảng.

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

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

Bảng 2.17 ***Tên bảng thống kê*** *(tiêu đề chung)*

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

3.1.3. Các loại bảng thống kê

Căn cứ theo kết cấu của phần chủ đề, có thể chia làm ba loại bảng thống kê: bảng giản đơn, bảng phân tổ và bảng kết hợp.

a. Bảng giản đơn: là loại bảng thống kê, trong đó phần chủ đề không phân tổ. Trong phần chủ đề của bảng giản đơn có liệt kê các đơn vị tổng thể, tên gọi các địa phương hoặc các thời gian khác nhau của quá trình nghiên cứu. Ví dụ có bảng giản đơn sau:

Bảng 2.18 ***Tình hình sản xuất kinh doanh năm 2005 của các doanh nghiệp thuộc một ngành X***

Tên doanh nghiệp	Số lao động	Giá trị sản xuất (1000đ)	Năng suất lao động bình quân
(a)	(1)	(2)	(3)
Doanh nghiệp A			
Doanh nghiệp B			
Doanh nghiệp C			
...	...	...	...
Cộng	...	...	...

b. Bảng phân tổ:

Bảng phân tổ là loại bảng thống kê, trong đó đối tượng nghiên cứu ghi trong phần chủ đề được phân chia thành các tổ theo một tiêu thức nào đó. Ví dụ, bảng phân tổ các doanh nghiệp công nghiệp theo khu vực và thành phần kinh tế năm 2003 (bảng 3 ở trên). Các bảng phân tổ là kết quả của việc áp dụng phương pháp phân tổ thống kê. Bảng phân tổ cho ta thấy rõ các loại hình kinh tế xã hội tồn tại trong bản thân hiện tượng nghiên cứu, nêu lên kết cấu và biến động kết cấu của hiện tượng; trong nhiều trường hợp còn giúp ta phân tích được mối liên hệ giữa các hiện tượng.

c. *Bảng kết hợp*:

Bảng kết hợp là loại bảng thống kê, trong đó đối tượng nghiên cứu ghi trong phần chủ đề được phân tổ theo hai, ba... tiêu thức kết hợp với nhau. Ví dụ: bảng thống kê công

nhân trong xí nghiệp, được phân tổ theo trình độ kỹ thuật và theo tuổi nghề (bảng 4). Loại bảng kết hợp như trên giúp ta nghiên cứu được sâu sắc bản chất của hiện tượng, đi sâu vào kết cấu nội bộ của hiện tượng, thấy rõ mối quan hệ giữa các tổ, bộ phận của hiện tượng trong quá trình phát triển.

3.1.4. Yêu cầu đối với việc xây dựng bảng thống kê

Một bảng thống kê được xây dựng một cách khoa học sẽ trở nên gọn, rõ, đáp ứng được mục đích nghiên cứu. Việc xây dựng bảng thống kê cần đảm bảo những yêu cầu sau:

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

- *Thứ hai*, các tiêu đề và tiêu mục trong bảng thống kê cần được ghi chính xác, gọn và dễ hiểu. Tiêu đề chung không những nói rõ nội dung chủ yếu của bảng thống kê, mà còn cần chỉ rõ hiện tượng nghiên cứu vào thời gian và địa điểm nào? Nhiều khi ở phần tiêu đề chung còn quy định đơn vị tính toán chung cho các số liệu trong bảng thống kê (nếu đơn vị tính toán không thống nhất cho các số liệu, thì chỉ quy định riêng cho mỗi hàng và cột).

- *Thứ ba*, các hàng và cột thường được ký hiệu bằng chữ hoặc bằng số để tiện cho việc trình bày hoặc giải thích nội dung. Các cột của phần chủ đề thường được ký hiệu bằng các chữ a, b, c... còn các cột của phần giải thích được ký hiệu bằng các số 1, 2, 3... Tuy nhiên, nếu một bảng thống kê chỉ có ít hàng và cột và nội dung các hàng cột đã rõ ràng, dễ hiểu thì không nhất thiết phải dùng ký hiệu.

- *Thứ tư*, các chỉ tiêu giải thích trong bảng thống kê cần được sắp xếp theo thứ tự hợp lý, phù hợp với mục đích nghiên cứu. Giả sử muốn lập một bảng thống kê nêu rõ mối liên hệ giữa mức năng suất lao động và giá trị sản lượng của các xí nghiệp. Như vậy, trước hết trong phần chủ đề ta có thể phân tổ các xí nghiệp theo giá trị sản lượng từ thấp đến cao (phân tổ có khoảng cách tổ). Còn các chỉ tiêu giải thích được bố trí theo thứ tự sau: số xí nghiệp mỗi tổ, giá trị sản lượng của các xí nghiệp trong tổ, số công nhân bình quân trong kỳ của mỗi tổ, năng suất lao động bình quân của mỗi công nhân trong tổ. Nếu bây giờ ta đảo ngược trật tự các chỉ tiêu nói trên, thì việc nhận thức và tính toán sẽ khó khăn hơn.

Trong mỗi bảng thống kê, các chỉ tiêu có ý nghĩa quan trọng trong việc so sánh với nhau thì nên bố trí gần nhau, như chỉ tiêu thực hiện bố trí gần chỉ tiêu kế hoạch, chỉ tiêu tương đối bố trí gần chỉ tiêu tuyệt đối...

- *Thứ năm*, cách ghi các số liệu vào bảng thống kê: các ô trong bảng thống kê đều có ghi số liệu hoặc bằng các ký hiệu quy ước thay thế. Thường dùng các ký hiệu quy ước sau:

+ Nếu hiện tượng không có số liệu đó, thì trong ô sẽ ghi một dấu gạch ngang (-)

+ Nếu số liệu còn thiếu, sau này có thể bổ sung, thì trong ô có ký hiệu 3 chấm (...)

+ Ký hiệu gạch chéo (x) trong một ô nào đó nói lên rằng hiện tượng không có liên quan đến chỉ tiêu đó, nếu viết số liệu vào ô đó sẽ vô nghĩa.

Các số liệu trong cùng một cột, có đơn vị tính toán giống nhau, phải ghi theo trình độ chính xác như nhau (số lẻ đến 0,1 hay 0,01...) đơn vị tính phải ghi thống nhất theo quy định.

Nếu mục đích của bảng thống kê chỉ nhằm nêu lên những nét chung về bản chất hiện tượng, không cần quá chi li số lẻ thì các số liệu trong bảng có thể ghi theo số tròn. Chẳng hạn, các đơn vị đo lường tính lẻ đến kilôgram có thể tính tròn đến tạ, tấn; đơn vị đo lường tính lẻ đến từng mét có thể tính tròn đến kilômét; tiền tệ có thể tính tròn đến nghìn hoặc triệu đồng... Bằng cách tính tròn như vậy, có thể thay những số liệu có 6, 7... chữ số thành những số liệu chỉ có gọn 2, 3... chữ số. Việc tính tròn cũng theo nguyên tắc toán học.

Các số cộng và tổng cộng có thể được ghi ở đầu hoặc ở cuối hàng và cột tuỳ theo mục đích nghiên cứu. Các số này được ghi ở đầu hàng, đầu cột khi ta cần nghiên cứu chủ yếu các đặc trưng của hiện tượng, còn các đặc trưng từng bộ phận chỉ có tác dụng phân tích thêm. Các số cộng và tổng được được ghi ở cuối hàng, cuối cột là khi ta nghiên cứu đi sâu từng tổ, từng bộ phận là chủ yếu.

- *Thứ sáu*, phần ghi chú ở cuối bảng thống kê được dùng để giải thích rõ nội dung của một số chỉ tiêu trong bảng, để nói rõ nguồn số liệu đã được sử dụng trong bảng hoặc các chi tiết cần thiết khác.

3.2. Đồ thị thống kê

3.2.1. Ý nghĩa và tác dụng của đồ thị thống kê

Đồ thị thống kê là các hình vẽ hoặc đường nét hình học dùng để miêu tả có tính chất quy ước các tài liệu thống kê. Khác với các bảng thống kê chỉ dùng con số, các đồ thị thống kê sử dụng con số kết hợp với các hình vẽ, đường nét và mầu sắc để trình bày và phân tích các đặc điểm số lượng của hiện tượng. Vì vậy, người xem không cần mất nhiều công đọc con số mà vẫn nhận thức được vấn đề chủ yếu một cách dễ dàng, nhanh chóng. Mặt khác, các đồ thị thống kê không trình bày chi tiết tỷ mỉ các đặc trưng số lượng của hiện tượng, mà chỉ nêu lên một cách khái quát các đặc điểm chủ yếu về bản chất và xu hướng phát triển cơ bản của hiện tượng. Vì vậy, đồ thị thống kê có tính quần chúng, có sức hấp dẫn và sinh động, làm cho người hiểu biết ít về thống kê vẫn lĩnh hội được vấn đề chủ yếu một các dễ dàng, đồng thời giữ được ấn tượng sâu đối với người đọc.

Các đồ thị thống kê được sử dụng rộng rãi trong mọi công tác nghiên cứu kinh tế, nhằm mục đích hình tượng hóa:

- Sự phát triển của hiện tượng qua thời gian

- Kết cấu và biến động kết cấu của hiện tượng

- Trình độ phổ biến của hiện tượng

- Sự so sánh giữa các mức độ của hiện tượng

- Mối liên hệ giữa các hiện tượng

- Tình hình thực hiện kế hoạch

Ngoài ra, đồ thị thống kê còn được coi là một phương tiện tuyên truyền rất mạnh mẽ, một công cụ dùng để biểu dương các thành tích sản xuất và hoạt động văn hoá xã hội.

3.2.2. Các loại đồ thị thống kê

Trong thống kê thường dùng các loại đồ thị sau đây:

3.2.2.1. Căn cứ vào hình thức biểu hiện: Có thể phân chia đồ thị thống kê thành các loại sau:

- Biểu đồ hình cột.

- Biểu đồ tượng hình.

- Biểu đồ diện tích (vuông, chữ nhật, tròn)

- Biểu đồ ra đa (mạng nhện)

- Đồ thị đường gấp khúc

- Bản đồ thống kê.

3.2.2.2. Căn cứ vào nội dung phản ánh : Có thể phân chia đồ thị thống kê thành các loại sau:

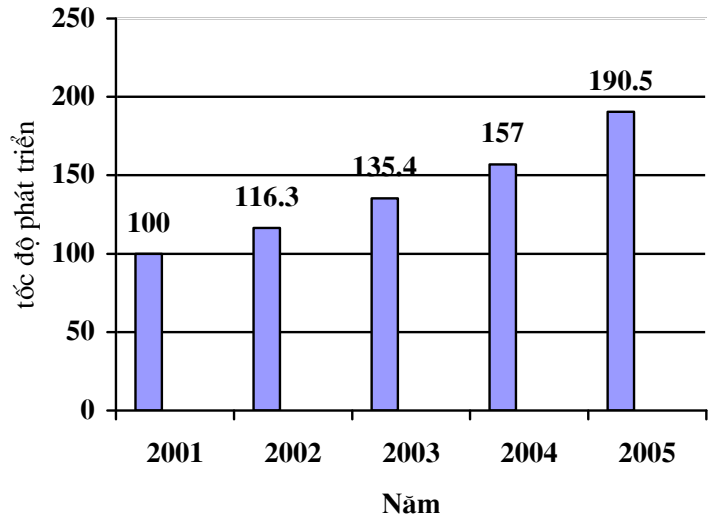
a. Đồ thị phát triển :

Đồ thị này dùng để biểu hiện tình hình phát triển của hiện tượng và so sánh giữa các hiện tượng, có thể dùng các loại biểu đồ hình cột, hình tròn và đồ thị tuyến tính.

Ví dụ: Có tài liệu về tốc độ phát triển giá trị sản xuất công nghiệp tỉnh A từ 2001 đến 2005 như sau (lấy năm 2001 là 100%):

Năm	2001	2002	2003	2004	2005
Tốc độ phát triển (%)	100	116,3	135,4	157,0	190,5

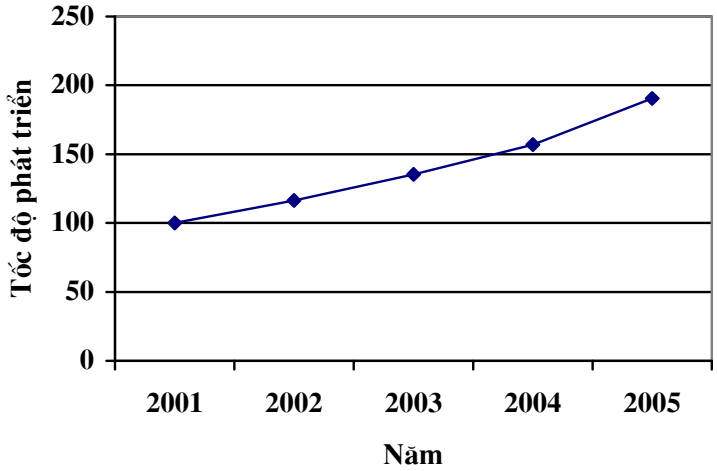
Theo tài liệu trên, có thể vẽ biểu đồ hình cột sau đây:



Hình 1: Biểu đồ về tốc độ phát triển giá trị sản xuất công nghiệp tỉnh A từ 2001 đến 2005

Trong biểu đồ trên, các cột đứng nói lên sự phát triển của sản xuất công nghiệp tỉnh A từ năm 2001 đến 2005. Các cột có bề rộng bằng nhau, còn chiều cao tương ứng với các đại lượng được biểu hiện.

Các đồ thị tuyến tính cũng thường được dùng để biểu hiện sự phát triển của hiện tượng. Theo ví dụ trên, ta vẽ thành đồ thị sau:



Hình 2: *Đồ thị gấp khúc về tốc độ phát triển giá trị sản xuất công nghiệp tỉnh A từ 2001 đến 2005*

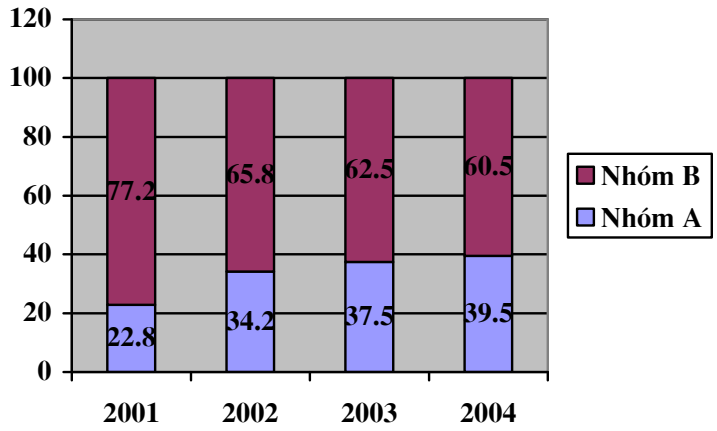
Trên đồ thị tuyến tính, trục hoành thường được dùng để biểu hiện thời gian, còn trục tung biểu hiện các mức độ của chỉ tiêu nghiên cứu. Một chú ý quan trọng khi vẽ loại đồ thị này là phải xác định độ khắc trên các trục toạ độ cho thích hợp, vì độ khắc có ảnh hưởng trực tiếp đến độ dốc của đường gấp khúc. Nếu độ khắc trên trục tung quá nhỏ so với độ khắc trên

trục hoành, đường gấp khúc sẽ vươn dài một cách quá mức, độ dốc của đường sẽ không thấy rõ. Ngược lại, nếu độ khắc trên trục tung quá lớn so với độ khắc trên trục hoành, đường gấp khúc sẽ vươn cao quá mức, độ dốc quá lớn gây cho người xem ấn tượng phóng đại sự phát triển của hiện tượng.

b. Đồ thị kết cấu :

Để biểu hiện kết cấu và biến động kết cấu của hiện tượng, thường dùng các loại biểu đồ hình cột và hình tròn (có chia nhỏ thành các hình quạt)

Ví dụ: Có biểu đồ hình cột về kết cấu giá trị sản xuất của một doanh nghiệp từ năm 2001 đến 2004, với 2 loại sản phẩm A và B như sau:



Hình 3: Biểu đồ tỷ trọng các nhóm sản phẩm A và B trong giá trị sản xuất của doanh nghiệp X từ năm 2001 đến 2004 (với 2 nhóm sản phẩm A và B)

c) Đồ thị liên hệ

Để biểu hiện mối liên hệ giữa 2 tiêu thức, người ta thường dùng đồ thị đường gấp khúc. Trục hoành của đồ thị được dùng để biểu hiện trị số của tiêu thức nguyên nhân (tiêu thức gây ảnh hưởng) ; trục tung của đồ thị được dùng để biểu hiện trị số của tiêu thức kết quả (tiêu thức chịu ảnh hưởng)

3.2.2.3. Những yêu cầu chung đối với việc xây dựng đồ thị thống kê.

Một đồ thị thống kê phải bảo đảm các yêu cầu: chính xác, dễ xem, dễ hiểu và nếu có thể trình bày mỹ thuật. Để đảm bảo những yêu cầu này, ta phải chú ý đến các yếu tố chính của đồ thị, quy mô, các ký hiệu hình học hoặc các hình vẽ, hệ tọa độ, thang và tỷ lệ xích, phần giải thích.

- Quy mô của đồ thị được quyết định bởi chiều dài, chiều cao và quan hệ tỷ lệ giữa hai chiều đó. Quy mô của đồ thị to hay nhỏ còn phải căn cứ vào mục đích sử dụng. Trong các báo cáo phân tích không nên vẽ các đồ thị quá lớn. Quan hệ tỷ lệ giữa chiều cao và chiều dài của đồ thị, thông thường được dùng từ 1:1,33 đến 1:1,5.

- Các ký hiệu hình học hoặc hình vẽ quyết định hình dáng của đồ thị. Các ký hiệu hình học có nhiều loại như: các chấm, các đường thẳng hoặc cong, các hình cột, hình vuông, hình chữ nhật, hình tròn… Các hình vẽ khác trên đồ thị cũng có thể thay đổi nhiều loại tùy tính chất của hiện tượng nghiên cứu. Việc lựa chọn các ký hiệu hình học hoặc hình vẽ của đồ thị là vấn đề quan trọng , vì mỗi hình có khả năng diễn tả riêng. Ví dụ khi cần biểu hiện kết cấu của hiện tượng nghiên cứu, ta có thể vẽ các hình cột (có chia thành nhiều đoạn) hoặc các hình tròn (có chia thành các hình quạt (hoặc hình vuông, hình chữ nhật…) Nhưng người ta thường dùng hình tròn, vì loại này biểu hiện được rõ nhất kết cấu và biến động kết cấu của hiện tượng.

- Hệ tọa độ giúp cho việc xác định chính xác vị trí các ký hiệu hình học trên đồ thị. Các đồ thị thống kê thường dùng hệ tọa độ vuông góc. Trong các bản đồ thống kê, người ta dùng các đường cong để làm căn cứ xác định vị trí các ký hiệu hình học. Các đường cong này có thể là đường biên giới, đường bờ biển, các con sông lớn… Trên hệ tọa độ vuông góc, trục hoành thường được dùng để biểu thị thời gian, trục tung biểu thị trị số của chỉ tiêu. Trong trường hợp phân tích mối liên hệ giữa hai biểu thức, thì biểu thức nguyên nhân được để ở trục hoành, biểu thức kết quả được ghi trên trục tung.

- Thang và tỷ lệ xích giúp cho việc tính chuyển các đại lượng lên đồ thị theo các khoảng cách thích hợp. Người ta thường dùng các thang đường thẳng, được phân bố theo các trục tọa độ. Cũng có khi dùng thang đường cong, ví dụ thang tròn (ở đồ thị hình tròn) được chia thành 360^0. Các thang tỷ lệ có thể có khoảng cách bằng nhau hoặc không bằng nhau. Các thang tỷ lệ có các khoảng cách không bằng nhau (ví dụ thang lôgarit) chỉ dùng để biểu hiện các tốc độ khi khoảng biến thiên của các mức độ quá lớn mà người ta chỉ chú ý đến biến động tương đối của chúng.

- Phần giải thích bao gồm tên đồ thị, các con số và ghi chú đọc theo thang tỷ lệ, các con số bên cạnh từng bộ phận của đồ thị, giải thích các ký hiệu quy ước… cần được ghi rõ, gọn, dễ hiểu.

3.2.2.4. *Nguyên tắc để có đồ thị tốt.*

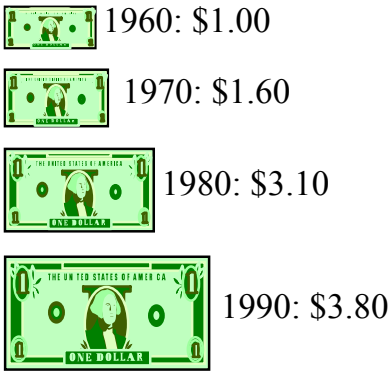
- Phải phản ánh được đầy đủ: nội dung, các tính toán thống kê, các đặc điểm của hiện tượng.
- Thể hiện các ý tưởng một cách rõ ràng, chính xác, hiệu quả.
- Đưa ra một số lượng lớn nhất các ý tưởng với cách thức hiệu quả nhất.
- Luôn bao gồm nhiều khía cạnh khác nhau.
- Phản ánh đúng sự thật dữ liệu.

3.2.2.5. *Một số lỗi trong việc trình bày dữ liệu:*

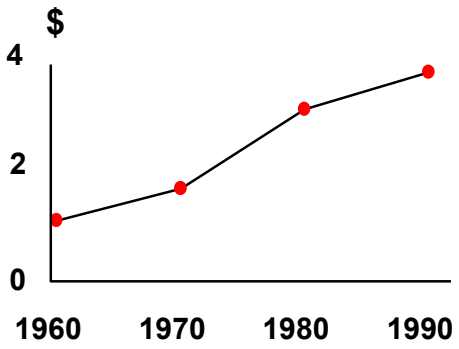
- Sử dụng những hình thức không thích hợp “Chart Junk”

Trình bày tồi

Tiền lương tối thiểu

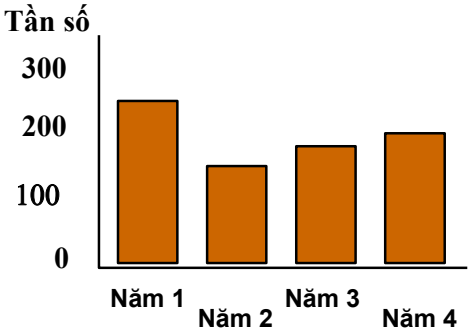


Trình bày tốt

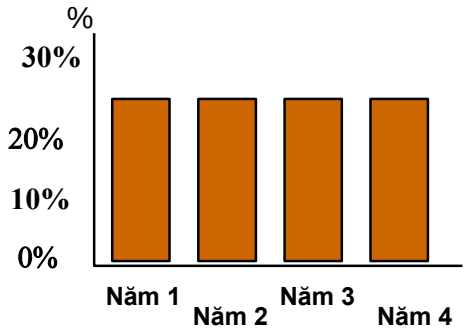


- Không có sự liên quan khi so sánh các nhóm dữ liệu:

Trình bày tồi

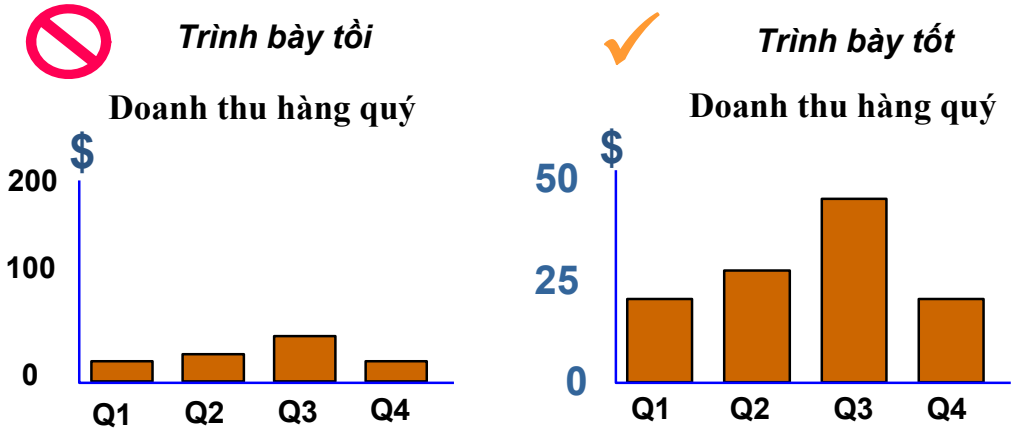


Trình bày tốt

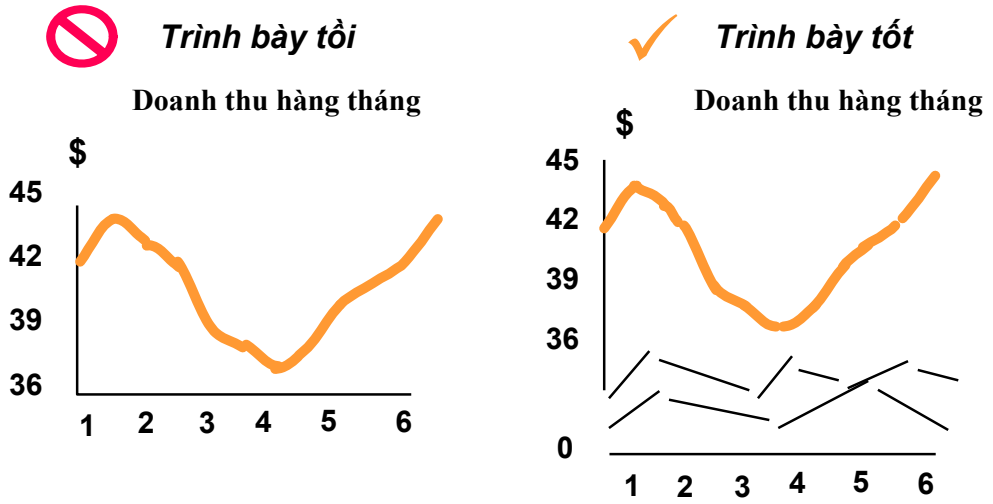


Trục tung biểu hiện số lượng sinh viên

- Nén trục tung qua nhiều: Phân chia khoảng cách quá lớn.



- Không có điểm zero ở trục tung.



Đồ thị doanh thu bán hàng 6 tháng đầu năm

BÀI TẬP

2.1 Tại một toà báo, người ta thu thập thông tin về thời gian cần thiết để hoàn thành trang nhất của tờ báo. Thu thập trong 50 ngày liền và được số liệu sau (đơn vị: phút)

23,8	20,3	23,6	19,0	25,1	25,0	19,5	24,1	24,2	21,8
20,8	22,8	21,9	22,0	20,7	20,9	25,0	22,2	22,8	20,1
25,3	20,7	22,5	21,2	23,8	23,3	20,9	22,9	23,5	19,5
21,3	21,5	23,1	19,9	24,2	24,1	19,8	23,9	22,8	23,9
19,7	24,2	23,8	20,7	23,8	24,3	21,1	20,9	21,6	22,7

Yêu cầu:

a) Xây dựng biểu đồ thân lá; sắp xếp số liệu theo thứ tự từ nhỏ đến lớn

b) Phân số liệu thành 7 tổ với khoảng cách tổ đều nhau. Tính tần số và tần số tích luỹ

c) Vẽ đồ thị tần số và tần số tích luỹ

d) Dựa vào đường cong tần số tích luỹ, hãy ước tính tỷ lệ % của những số báo mà trang nhất được thiết kế trong vòng 24 phút.

2.2 Dưới đây là số liệu về độ tuổi của các bệnh nhân đến khám ở bệnh viện A vào ngày 20/8/2008

32	45	53	60	79	73
73	53	61	48	51	49
62	72	37	70	38	66
52	33	78	45	65	47
64	47	61	75	57	64

Yêu cầu:

a. Xây dựng biểu đồ thân lá

b. Xây dựng bảng phân bố tần số thích hợp

c. Vẽ các đồ thị biểu diễn tần số và tần số tích lũy, tần suất tích lũy

d. Cho nhận xét về phân bố tuổi củ các bệnh nhân nói trên

2.3 Dữ liệu sau là số lượng hành khách trên các chuyến bay của hãng Delta Airlines giữa San Francisco và Seattle trong 33 ngày tháng 4 và đầu tháng 5:

128 121 134 136 136 118 123 109 120 116 125 128 121 129 130 131 127
119 114 134 110 136 134 125 128 123 128 133 132 136 134 129 132

Xây dựng biểu đồ thân lá và vẽ các đồ thị biểu diễn đặc điểm phân bố của về số lượng hành khách ở các chuyến bay nói trên.

2.4 Dữ liệu sau ước tính doanh số bán thiết bị điện trên toàn cầu (tính theo triệu USD). Sử dụng dữ liệu để xây dựng một biểu đồ hình bánh cho doanh số bán thiết bị điện trên toàn cầu của các nhà sản xuất sau:

Electrolux: 5100

General Electric: 4350

Matsushita Electric: 4180
Whirlpool: 3950
Bosch-Siemens: 2200
Philips: 2000
Maytag: 1580

2.5 Các quốc gia có số người tới thăm nước Mỹ nhiều nhất trong năm 1997 như sau (đơn vị tính: Triệu người):
Canada: 15.9
Mexico: 8.9
Nhật Bản: 5.5
Anh: 3.3
Đức: 2.1
Pháp: 1.1
Brazil: 1.0
Hàn Quốc: 1.0
Italy: 0.6
Australia: 0.5
Hãy vẽ bar graph và pie chart.

2.6 Dữ liệu sau đây là sản lượng thép hàng tháng (triệu tấn)
7.0 6.9 8.2 7.8 7.7 7.3 6.8 6.7 8.2 8.4 7.0 6.7 7.5 7.2 7.9 7.6 6.7 6.6 6.3
5.6 7.8 5.5 6.2 5.8 5.8 6.1 6.0 7.3 7.3 7.5 7.2 7.2 7.4 7.6
Vẽ một biểu đồ thân lá (***stem-and-leaf*** display) của dữ liệu trên. Trên cơ sở đó phân tổ với số tổ và khoảng cách tổ thích hợp và cho nhận xét.

2.7 Sự tham gia của công nhân vào hoạt động quản lý là một chương trình mới, thu hút người lao động vào việc ra các quyết định của công ty. Dữ liệu sau là số phần trăm người lao động được thu hút vào chương trình này trong một mẫu các doanh nghiệp. Hãy vẽ một ***stem-and-leaf display*** của dữ liệuvà rút ra kết luận về bộ dữ liệu.
32 33 35 42 43 42 45 46 44 47 48 48 48 49 49 50 37 38 34 51 52
52 47 53 55 56 57 58 63 78

2.8 Dữ liệu sau là bảng giá cả hằng ngày của một cổ phiếu nhất định trong 45 ngày. Xây dựng ***stem-and-leaf display*** của dữ liệu. Bạn có thể kết luận gì về sự phân bố của giá hằng ngày của cổ phiếu trong giai đoạn đó?
10, 11, 10, 11, 11, 12, 12, 13, 14, 16, 15, 11, 18, 19, 20, 15, 14, 14, 22, 25, 27, 23, 22, 26, 27, 29, 28, 31, 32, 30, 32, 34, 33, 38, 41, 40, 42, 53, 52, 47, 37, 23, 11, 32, 23.

2.9 Dữ liệu sau là số ounce bạc trong mỗi tấn quặng ở hai mỏ:
Mỏ A: 34, 32, 35, 37, 41, 42, 43, 45, 46, 45, 48, 49, 51, 52, 53, 60, 73, 76, 85
Mỏ B: 23, 24, 28, 29, 32, 34, 35, 37, 38, 40, 43, 44, 47, 48, 49, 50, 51, 52, 59
Xây dựng một ***stem-and-leaf display*** cho mỗi bộ dữ liệu. So sánh hai ***display*** và rút ra kết luận về dữ liệu đó.

2.10 Có tài liệu sau đây của 50 công nhân đúc bê tông của một phân xưởng bê tông:

STT	Bậc thợ	Tuổi nghề (năm)	Mức độ cơ giới hoá lao động (%)	Năng suất lao động ngày (m^3)	STT	Bậc thợ	Tuổi nghề (năm)	Mức độ cơ giới hoá lao động (%)	Năng suất lao động ngày (m^3)
1	2	2	35	3,0	26	3	4	69	5,0
2	3	3	59	6,5	27	2	3	48	2,5
3	3	2	44	4,8	28	4	7	82	6,8
4	3	4	55	5,7	29	4	6	98	6,6
5	2	2	39	2,8	30	3	5	63	6,3
6	3	3	56	4,7	31	4	10	79	7,9
7	2	3	78	4,2	32	3	5	41	4,6
8	4	3	44	5,3	33	3	4	45	4,2
9	3	2	43	2,0	34	2	5	75	4,8
10	3	5	76	6,5	35	3	4	45	5,8
11	3	4	58	5,1	36	4	3	51	4,9
12	4	2	41	5,5	37	3	4	55	4,3
13	2	2	49	3,0	38	4	8	95	6,4
14	2	3	58	3,6	39	4	10	90	7,0
15	4	6	58	4,5	40	4	9	70	7,1
16	4	7	61	6,7	41	3	6	56	4,4
17	3	5	42	5.6	42	3	5	57	5,1
18	3	3	46	5,2	43	2	3	48	5,0
19	2	2	35	3,2	44	3	8	72	6,1
20	4	4	55	5,4	45	3	6	52	5,9
21	3	2	38	4,5	46	2	4	33	3,8
22	3	3	35	5,5	47	3	2	55	4,6
23	3	2	25	2,5	48	2	2	30	3,4
24	4	8	90	6,2	49	2	4	67	5,5
25	2	4	47	4,1	50	3	3	57	5,9

Hãy phân tổ công nhân để nghiên cứu mối liên hệ:

a) Giữa năng suất lao động và bậc thợ (3 tổ)

b) Giữa năng suất lao động và tuổi nghề (3 tổ)

c) Giữa năng suất lao động và mức độ cơ giới hoá lao động (3 tổ)

Mỗi phân tổ rút ra nhận xét.

2.11 Phân tổ công nhân theo tài liệu bài 2.10 để nghiên cứu mối liên hệ

a) Giữa năng suất lao động với tuổi nghề, mức độ cơ giới hoá lao động

b) Giữa năng suất lao động với tuổi nghề, bậc thợ

c) Giữa năng suất lao động với mức độ cơ giới hoá lao động, bậc thợ

Rút ra nhận xét qua mỗi bản phân tổ.

2.12 Sử dụng phương pháp phân tổ nhiều chiều để nghiên cứu mối liên hệ giữa năng suất lao động với bậc thợ, tuổi nghề, mức độ cơ giới hoá lao động theo tài liệu ở bài 2.10 và rút ra nhận xét?

CHƯƠNG 3

MÔ TẢ DỮ LIỆU ĐỊNH LƯỢNG

Các hiện tượng tồn tại trong những điều kiện thời gian và địa điểm nhất định. Mỗi đặc điểm cơ bản của hiện tượng thường có thể được biểu hiện bằng các mức độ khác nhau.

Các mức độ của hiện tượng kinh tế xã hội... trước hết cho ta một nhận thức cụ thể về quy mô, khối lượng của hiện tượng trong điều kiện lịch sử nhất định. Chẳng hạn, muốn nghiên cứu tình hình sản xuất của một doanh nghiệp trong một thời gian nào đó, trước hết phải tính được số lượng lao động, số máy móc thiết bị, số nguyên vật liệu đưa vào sản xuất, số sản phẩm đã sản xuất ra...

Các mức độ của hiện tượng kinh tế xã hội có thể phản ánh các quan hệ tỷ lệ khác nhau, như quan hệ giữa bộ phận với tổng thể, quan hệ giữa thực tế với kế hoạch, quan hệ giữa kỳ này với kỳ trước, quan hệ giữa hiện tượng này với hiện tượng khác... Như trong việc nghiên cứu tình hình sản xuất nông nghiệp của một địa phương, cần tính tỷ lệ mỗi loại sản phẩm trong toàn bộ giá trị sản xuất nông nghiệp, sản lượng lương thực tính theo đầu người...

Thông qua việc nghiên cứu các mức độ, còn có thể nêu lên đặc điểm chung nhất, đại diện nhất về từng mặt của hiện tượng bao gồm nhiều đơn vị cùng loại. Các mức độ như giá thành bình quân, năng suất lao động bình quân, giá cả bình quân, năng suất thu hoạch bình quân... thường được tính đến trong khi nghiên cứu thống kê.

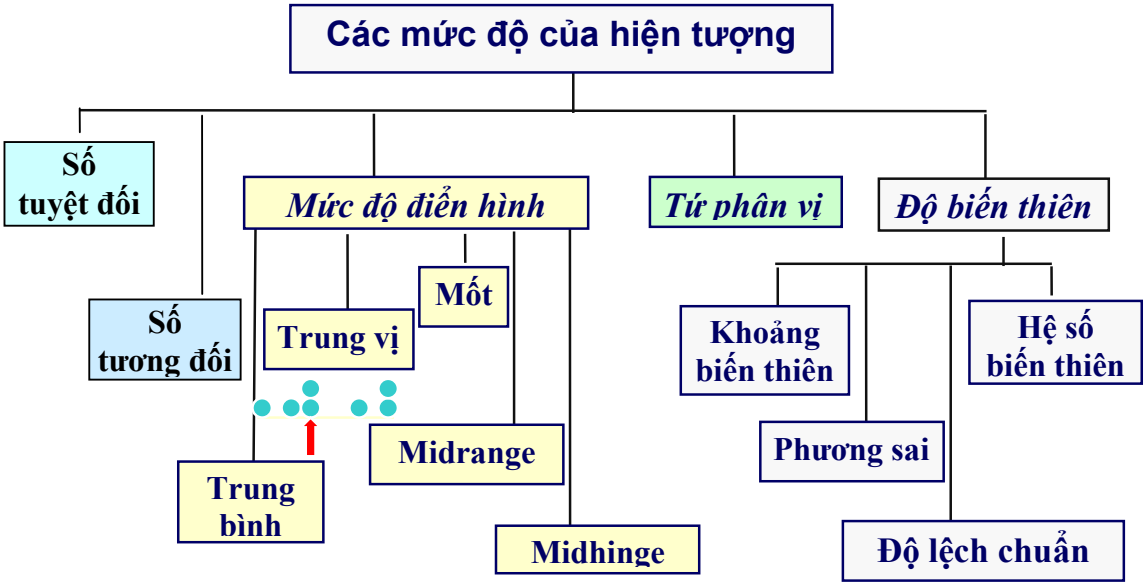
Ngoài ra, các mức độ của hiện tượng nghiên cứu còn giúp ta đánh giá trình độ đồng đều của tổng thể, khảo sát độ biến thiên của tiêu thức, khảo sát tình hình phân phối các đơn vị tổng thể. Đây là những yêu cầu về nhận thức không thể thiếu được trong phân tích thống kê.

Như vậy, việc nghiên cứu các mức độ của hiện tượng kinh tế xã hội là một trong những vấn đề nội dung của phân tích thống kê, nhằm vạch rõ mặt lượng trong mối quan hệ mật thiết với mặt chất của hiện tượng nghiên cứu trong điều kiện thời gian và địa điểm cụ thể. Đây cũng là cơ sở xuất phát của nhiều nội dung phân tích thống kê khác. Trong mọi hoạt động sản xuất, kinh doanh, trong công tác quản lý kinh tế, đều cần thiết nắm được các mức độ của hiện tượng nghiên cứu.

Trong phân tích thống kê các mức độ của hiện tượng bao gồm:

- Số tuyệt đối
- Số tương đối
- Các mức độ điển hình
- Các mức độ đo độ biến thiên của tiêu thức
- Các mức độ và biểu đồ biểu hiện hình dáng phân phối của tổng thể

Tất cả các mức độ đó được thể hiện ở sơ đồ sau



1. Số tuyệt đối trong thống kê

1.1. Khái niệm và ý nghĩa số tuyệt đối

Số tuyệt đối trong thống kê là mức độ biểu hiện quy mô, khối lượng của hiện tượng trong điều kiện thời gian và địa điểm cụ thể.

Số tuyệt đối nói lên số đơn vị của tổng thể hay của bộ phận (số doanh nghiệp, số nông trường, số công nhân, số học sinh, sinh viên...) hoặc các trị số của một tiêu thức nào đó (giá trị sản xuất công nghiệp, tổng chi phí sản xuất, tổng số tiền lương...). Thí dụ: năm 2005, số lao động của doanh nghiệp X là 750 người và doanh thu của doanh nghiệp là 120,5 tỷ đồng. Các con số thống kê trên đều là số tuyệt đối.

Số tuyệt đối có ý nghĩa quan trọng cho mọi công tác nghiên cứu kinh tế, vì thông qua các số tuyệt đối ta sẽ có một nhận thức cụ thể về quy mô, khối lượng thực tế của hiện tượng nghiên cứu. Nhờ các số tuyệt đối, có thể biết cụ thể nguồn tài nguyên, các khả năng tiềm tàng trong nền kinh tế quốc dân, các kết quả phát triển kinh tế, văn hoá, các thành quả lao động mà mọi người đã phấn đấu đạt được. Số tuyệt đối chính xác là sự thật khách quan, có sức thuyết phục không ai có thể phụ nhận được.

Số tuyệt đối là cơ sở đầu tiên để tiến hành phân tích thống kê, đồng thời còn là cơ sở để tính các mức độ khác.

Số tuyệt đối là căn cứ không thể thiếu được trong việc xây dựng các kế hoạch kinh tế quốc dân và chỉ đạo thực hiện kế hoạch.

Do ý nghĩa quan trọng như vậy, thống kê học coi số tuyệt đối là loại chỉ tiêu cơ bản nhất.

1.2. Đặc điểm của số tuyệt đối

Mỗi số tuyệt đối trong thống kê đều bao hàm một nội dung kinh tế xã hội cụ thể trong điều kiện thời gian và địa điểm nhất định. Nó khác với các đại lượng tuyệt đối trong toán học, vì các đại lượng này thường có tính chất trừu tượng, không nhất thiết phải gắn liền với một hiện tượng cụ thể nào. Do đặc điểm nói trên, điều kiện chủ yếu để có số tuyệt đối chính xác là phải xác định được một cách cụ thể, đóng đắn nội dung kinh tế mà chỉ tiêu phản ánh. Thí dụ, muốn tính được tiền lương của lao động phải hiểu rõ bản chất của tiền lương, nội dung của tiền lương bao gồm những khoản mục nào trong tất cả các khoản tiền mà người lao động có thể nhận được tại doanh nghiệp.

Các số tuyệt đối trong thống kê cũng không phải là con số được lựa chọn tuỳ ý mà phải qua điều tra thực tế và tổng hợp một cách khoa học. Cũng có khi còn phải dùng các phương pháp tính toán khác nhau mới có được các số tuyệt đối, như muốn biết số nguyên vật liệu tồn kho cuối kỳ phải lập bảng cân đối đồng thời kết hợp với kiểm kê thực tế.

1.3. Đơn vị tính số tuyệt đối

Các số tuyệt đối trong thống kê đều có đơn vị tính cụ thể. Tuỳ theo tính chất của hiện tượng và mục đích nghiên cứu, số tuyệt đối có thể được tính bằng đơn vị tự nhiên, đơn vị thời gian lao động và đơn vị tiền tệ.

Đơn vị tự nhiên là đơn vị tính toán phù hợp với đặc điểm vật lý của hiện tượng. Các hiện tượng này có thẻ được tính theo chiều dài (mét, kilômét...), theo diện tích (mét vuông, hécta, kilômét vuông...), theo trọng lượng (kilôgam, tạ, tấn...), theo dung tích (lít, mét khối...)... Đơn vị tự nhiên cũng có thể là số đơn vị tổng thể (cái, con, chiếc...), số người, số sự kiện, số trường hợp.

Trong nhiều trường hợp phải dùng đơn vị kép để tính toán, như sản lượng điện tính bằng kilô-oát giờ, khối lượng vận chuyển tính bằng tấn-kilômét. Trong sản xuất những sản phẩm giống nhau về giá trị sử dụng, nhưng khác nhau về kích thước, trọng lượng, công suất..., do đó muốn tổng hợp được những sản phẩm này, người ta dùng đơn vị hiện vật tiêu chuẩn. Thí dụ: máy kéo có công suất tiêu chuẩn là 15 mã lực, đồ hộp có trọng lượng 400gam hay đồ hộp có dung tích 350 cm^3, chất đốt có nhiệt lượng 7000 kilôcalo...

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

Đơn vị tiền tệ, chủ yếu là hai loại đơn vị: đồng Việt Nam và đôla vì nó có thể giúp cho việc tổng hợp, so sánh nhiều sản phẩm có giá trị sử dụng và đơn vị đo lường khác nhau và so sánh quốc tế.

1.4. Các loại số tuyệt đối

Tuỳ theo tính chất của hiện tượng nghiên cứu và khả năng thu thập tài liệu trong những điều kiện thời gian khác nhau, có thể phân biệt hai loại số tuyệt đối sau đây:

- Số tuyệt đối thời kỳ phản ánh quy mô, khối lượng của hiện tượng trong một độ dài thời gian nhất định. Thí dụ: Doanh thu của doanh nghiệp X năm 2005 là 120 tỷ đồng, đó là số tuyệt đối thời kỳ. Nhiều chỉ tiêu khác như: chi phí sản xuất, lượng hàng hoá tiêu thụ... đều là số tuyệt đối thời kỳ, vì đó là kết quả tổng hợp mặt lượng của hiện tượng trong một độ dài thời gian nhất định. Các số tuyệt đối thời kỳ của cùng một chỉ tiêu có thể cộng được với nhau; thời kỳ càng dài thì trị số của nó càng lớn.

- Số tuyệt đối thời điểm phản ánh quy mô, khối lượng của hiện tượng nghiên cứu vào một thời điểm nhất định. Thí dụ: dân số thành phố A vào 0 giờ ngày 1/4/1999 là 2,5 triệu người, đó là số tuyệt đối thời điểm. Nhiều chỉ tiêu khác như: số công nhân ngày đầu tháng, số nguyên vật liệu tồn kho ngày cuối tháng... đều được biểu hiện bằng số tuyệt đối thời điểm. Số tuyệt đối thời điểm chỉ phản ánh tình hình của hiện tượng vào một thời điểm nào đó; trước hoặc sau thời điểm đó, trạng thái của hiện tượng có thể khác. Do đó, muốn có số tuyệt đối thời điểm chính xác, phải quy định thời điểm hợp lý và phải tổ chức điều tra kịp thời.

2. Số tương đối trong thống kê

2.1. Khái niệm và ý nghĩa số tương đối

Số tương đối trong thống kê biểu hiện quan hệ so sánh giữa hai mức độ nào đó của hiện tượng. Đó có thể là kết quả của việc so sánh giữa hai mức độ cùng loại nhưng khác nhau về điều kiện thời gian hoặc không gian, hoặc giữa hai mức độ khác loại nhưng có liên quan với nhau. Trong hai mức độ này, một được chọn làm gốc để so sánh.

Thí dụ: giá trị sản xuất công nghiệp của tỉnh A năm 2005 so với năm 2004 bằng 112% (tăng 12%), còn so với kế hoạch đạt 104,3%; cơ cấu dân số nước Việt Nam năm 2003, nữ chiếm 50,86% và nam chiếm 49,14... Những con số thống kê trên đều là số tương đối.

Trong phân tích thống kê, các số tương đối được sử dụng rộng rãi để nêu lên kết cấu, quan hệ so sánh, trình độ phát triển, trình độ phổ biến... của hiện tượng nghiên cứu trong điều kiện lịch sử nhất định.

Cũng như các số tuyệt đối, số tương đối trong thống kê nói lên mặt lượng trong quan hệ mật thiết với mặt chất của hiện tượng nghiên cứu. Tuy nhiên, trong khi các số tuyệt đối chỉ mới khái quát được về quy mô, khối lượng của hiện tượng, thì các số tương đối tính được bằng các phương pháp so sánh có thể giúp ta đi sâu vào đặc điểm của hiện tượng một cách có phân tích phê phán. Thí dụ, biết giá trị sản xuất nông nghiệp của tỉnh A năm 2005 là 1530 tỷ đồng. Muốn phân tích xem con số đạt được như vậy là nhiều hay ít, đã thỏa mãn được nhu cầu tiêu dùng của xã hội chưa, có hoàn thành kế hoạch không, so với các năm trước hơn hay kém..., cần đem so sánh chỉ tiêu nói trên với nhiều chỉ tiêu khác. Như đem so sánh với cùng chỉ tiêu này năm 2002, ta thấy nó bằng 107,2% (tăng 7,2%); có thể kết luận rằng sản xuất nông nghiệp của tỉnh có tăng lên. Nhưng cũng thời kỳ nói trên, dân số của địa phương đã tăng 7,8%, nghĩa là tăng nhanh hơn tốc độ sản xuất nông nghiệp, có thể nhận định rằng mức sống vật chất của nhân dân còn gặp nhiều khó khăn.

Trong công tác lập kế hoạch và kiểm tra thực hiện kế hoạch, số tương đối cũng giữ vai trò quan trọng. Nhiều chỉ tiêu kế hoạch được đề ra bằng số tương đối, còn khi kiểm tra thực hiện kế hoạch thì ngoài việc tính toán chính xác các số tuyệt đối, bao giờ cũng phải đánh giá trình độ hoàn thành kế hoạch bằng các số tương đối.

Ngoài ra, người ta còn dùng các số tương đối để nêu rõ tình hình thực tế trong khi cần bảo đảm được tính chất bí mật của các số tuyệt đối.

2.2. Đặc điểm và hình thức biểu hiện số tương đối

Các số tương đối trong thống kê không phải là con số thu thập được qua điều tra, mà là kết quả so sánh giữa hai chỉ tiêu thống kê đã có. Bởi vậy, mỗi số tương đối đều phải có gốc dùng để so sánh. Tuỳ theo mục đích nghiên cứu, gốc dùng để so sánh có thể lấy khác nhau: để nêu lên sự phát triển thì gốc được chọn là mức độ kỳ trước, để kiểm tra thực hiện kế hoạch thì gốc được chọn là mức độ kế hoạch, để biểu hiện quan hệ giữa bộ phận với tổng thể thì gốc là mức độ của tổng thể... Như vậy, do khả năng sử dụng gốc so sánh khác nhau, việc tính toán số tương đối khá phong phú.

Hình thức biểu hiện của số tương đối là số lần, số phần trăm (%) hay số phần nghìn (‰). Ba hình thức biểu hiện này căn bản không có gì khác nhau về nội dung, nhưng việc sử dụng hình thức nào là do tính chất của hiện tượng và mục đích nghiên cứu. Số phần trăm thường được dùng trong các trường hợp mức độ đem so sánh với mức độ dùng làm gốc không chênh lệch nhau nhiều lắm. Nếu sự chênh lệch quá lớn, số tương đối thường được biểu hiện bằng số lần; ngược lại số phần nghìn được dùng khi sự chênh lệch quá nhỏ. Ngoài ra, khi dùng số tương đối để nói lên trình độ phổ biến của một hiện tượng nào đó, hình thức biểu hiện có thể là đơn vị kép: người/km^2, sản phẩm/người...

2.3. Các loại số tương đối

Căn cứ theo nội dung mà số tương đối phản ánh, có thể chia thành 5 loại số tương đối sau đây: Số tương đối động thái, số tương đối kế hoạch, số tương đối kết cấu, số tương đối cường độ, số tương đối không gian.

2.3.1. Số tương đối động thái

Số tương đối động thái thường được sử dụng rộng rãi để biểu hiện biến động về mức độ của hiện tượng nghiên cứu qua một thời gian nào đó. Số tương đối này tính được bằng cách so sánh hai mức độ cùng loại của hiện tượng ở hai thời kỳ (hay thời điểm) khác nhau và được biểu hiện bằng số lần hay số phần trăm. Mức độ được đem ra nghiên cứu được gọi là mức độ kỳ nghiên cứu, còn mức độ được dùng làm cơ sở so sánh được gọi là mức độ kỳ gốc . Nếu ký hiệu t là số tương đối động thái, y_1 là mức độ kỳ nghiên cứu, y_0 là mức độ kỳ gốc, ta có công thức tính như sau:

$$t = \frac{y_1}{y_0} \qquad (3.1)$$

Thí dụ: Vốn đầu tư xây dựng của một địa phương năm 2003 là 250 tỷ đồng và năm 2005 là 300 tỷ đồng. Nếu đem so sánh vốn đầu tư xây dựng năm 2005 với năm 2003, ta sẽ có số tương đối động thái:

$$\frac{300}{250} = 1{,}2 \text{ lần (hay } 120\%)$$

Vốn đầu tư xây dựng năm 2005 so với năm 2003 bằng 1,2 lần hay 120%. Trong thực tế số tương đối động thái này thường được gọi là tốc độ phát triển hay chỉ số phát triển.

Theo thí dụ trên, có thể tính cách khác: vốn đầu tư xây dựng năm 2005 tăng 50 tỷ đồng so với năm 2003; đem so sánh mức tăng này với mức kỳ gốc 2003, tính ra bằng 50 : 250 = 0,2 lần hay 20%. Đây cũng là số tương đối vì chỉ tiêu này tính được bằng cách lấy lượng tăng tuyệt đối (tức là hiệu giữa mức độ kỳ nghiên cứu và mức độ kỳ gốc) đem so sánh với mức độ kỳ gốc, người ta thường gọi là tốc độ tăng. Như vậy, tốc độ tăng cũng được kể vào loại số tương đối động thái nói trên.

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

2.3.2. Số tương đối kế hoạch

Số tương đối kế hoạch được dùng để lập và kiểm tra tình hình thực hiện kế hoạch. Có hai loại số tương đối kế hoạch:

- Số tương đối nhiệm vụ kế hoạch là quan hệ tỷ lệ giữa mức độ kỳ kế hoạch (tức là mức độ cần đạt tới của một chỉ tiêu kinh tế nào đó trong kỳ kế hoạch) với mức độ thực tế của chỉ tiêu này đạt được ở trước kỳ kế hoạch hoặc ở một kỳ nào đó được chọn làm gốc so sánh, thường được biểu hiện bằng đơn vị phần trăm. Công thức tính như sau:

$$K_n = \frac{y_K}{y_0} \qquad (3.2)$$

Trong đó: y_k là mức độ kỳ kế hoạch

y_0 là mức độ thực tế ở một kỳ nào đó được chọn làm gốc so sánh

- Số tương đối thực hiện kế hoạch là quan hệ tỷ lệ giữa mức độ thực tế đã đạt được trong kỳ kế hoạch với mức độ kế hoạch đã đề ra về một chỉ tiêu kinh tế nào đó, thường được biểu hiện bằng đơn vị phần trăm. Công thức tính như sau:

$$K_T = \frac{y_1}{y_K} \qquad (3.3)$$

Đối với những chỉ tiêu kinh tế mà kế hoạch dự kiến phải tăng lên mới là chiều hướng tốt, thì số tương đối hoàn thành kế hoạch tính ra trên 100% là vượt kế hoạch, còn dưới 100% là không hoàn thành kế hoạch. Nhưng cũng có một số chỉ tiêu kinh tế mà kế hoạch dự kiến

phải giảm đi mới là chiều hướng tốt (như giá thành, tiêu hao nguyên vật liệu cho một đơn vị sản phẩm...) thì số tương đối hoàn thành kế hoạch tính ra dưới 100% mới là vượt mức, còn trên 100% là không hoàn thành kế hoạch.

Khi tính các số tương đối kế hoạch cũng phải chú ý bảo đảm tính chất có thể so sánh được giữa các mức độ kế hoạch và thực tế về nội dung, phương pháp tính toán.

Giữa các loại số tương đối động thái, số tương đối nhiệm vụ kế hoạch và số tương đối hoàn thành kế hoạch (của cùng một chỉ tiêu) có mối quan hệ với nhau. Nếu đã biết hai loại số tương đối, có thể tính được số tương đối thứ ba. Cụ thể là:

+ Số tương đối động thái bằng tích của số tương đối nhiệm vụ kế hoạch với số tương đối hoàn thành kế hoạch.

$$\frac{y_1}{y_0} = \frac{y_K}{y_0} \text{ x } \frac{y_1}{y_K} \quad \text{hay} \quad t = K_n \times K_T$$

+ Số tương đối hoàn thành kế hoạch bằng tỷ số giữa số tương đối động thái với số tương đối nhiệm vụ kế hoạch.

$$\frac{y_1}{y_K} = \frac{y_1}{y_0} : \frac{y_K}{y_0} \quad \text{hay} \quad K_T = t : K_n$$

+ Số tương đối nhiệm vụ kế hoạch bằng tỷ số giữa số tương đối động thái với số tương đối hoàn thành kế hoạch.

$$\frac{y_K}{y_0} = \frac{y_1}{y_0} : \frac{y_1}{y_K} \quad \text{hay} \quad K_n = t : K_T$$

Các quan hệ toán học trên đây được vận dụng rộng rãi trong các tính toán của thống kê. Thí dụ: kế hoạch của doanh nghiệp tăng năng suất lao động 10% so với kỳ gốc, thực tế năng suất lao động đã tăng 15% so với kỳ gốc. Tỷ lệ hoàn thành kế hoạch tăng năng suất lao động bằng:

$$\frac{115}{110} \text{ x } 100 = 104{,}5\% \text{ (vượt kế hoạch } 4{,}5\%)$$

2.3.3. Số tương đối kết cấu

Số tương đối kết cấu được dùng để xác định tỷ trọng của mỗi bộ phận cấu thành trong một tổng thể. Số tương đối này thường biểu hiện bằng số phần trăm và tính được bằng cách so sánh mức độ của từng bộ phận (tổ) với mức độ của cả tổng thể.

$$\text{Số tương đối kết cấu} = \frac{\text{Mức độ của bộ phận}}{\text{Mức độ của tổng thể}} \text{ x } 100 \qquad (3.4)$$

Thí dụ: Giá trị sản xuất nông nghiệp của tỉnh B năm 2005 là 1600 tỷ đồng, trong đó ngành trồng trọt chiếm 1280 tỷ đồng và ngành chăn nuôi chiếm 320 tỷ đồng. Tính ra các số tương đối kết cấu:

- Tỷ trọng giá trị sản xuất ngành trồng trọt

$$\frac{1280}{1600} \text{ x } 100 = 80\%$$

- Tỷ trọng giá trị sản lượng ngành chăn nuôi

$$\frac{320}{1600} \text{ x } 100 = 20\%$$

Muốn tính các số tương đối kết cấu được chính xác, chủ yếu phải phân biệt rõ các bộ phận có tính chất khác nhau trong tổng thể nghiên cứu. Vì vậy, việc tính số tương đối kết cấu có quan hệ mật thiết với phương pháp phân tổ thống kê.

2.3.4. Số tương đối cường độ

Số tương đối cường độ được dùng để biểu hiện trình độ phổ biến của hiện tượng nghiên cứu trong một điều kiện lịch sử nhất định. Số tương đối này tính được bằng cách so sánh chỉ tiêu của hai hiện tượng khác nhau nhưng có liên quan với nhau. Thí dụ:

$$\text{Mật độ dân số} = \frac{\text{Tổng số dân (người)}}{\text{Diện tích đất đai (km}^2\text{)}} = (\text{đơn vị : người/km}^2)$$

Hệ số sinh của nhân khẩu =

$$\frac{\text{Số trẻ em sinh ra trong năm (người)}}{\text{Số nhân khẩu trung bình trong năm (1000người)}} = (\text{đơn vị : người/1000người})$$

Qua các thí dụ trên, ta thấy hình thức biểu hiện của số tương đối cường độ là đơn vị kép, do đơn vị tính toán của tử số và của mẫu số hợp thành. Vấn đề quan trọng khi tính số tương đối cường độ là phải xét các hiện tượng nào có liên quan với nhau, và khi so sánh thì hiện tượng nào để ở tử số hoặc ở mẫu số. Phải tuỳ theo mục đích nghiên cứu và mối quan hệ giữa hai hiện tượng mà giải quyết vấn đề so sánh cho thích hợp, bảo đảm số tương đối cường độ tính ra có ý nghĩa thực tế.

Số tương đối cường độ được sử dụng rộng rãi để nói lên trình độ phát triển sản xuất, trình độ bảo đảm về mức sống vật chất và văn hoá của nhân dân một nước. Đó là các chỉ tiêu như: GDP bình quân đầu người, sản lượng lương thực hay thực phẩm tính theo đầu người, số bác sĩ và giường bệnh phục vụ cho 1 vạn dân và nhiều chỉ tiêu khác... Số tương đối cường độ còn có thể được dùng để so sánh trình độ phát triển sản xuất giữa các nước khác nhau.

2.3.5. Số tương đối không gian

Là loại số tương đối biểu hiện sự so sánh về mức độ giữa hai bộ phận trong một tổng thể, hoặc giữa hai hiện tượng cùng loại nhưng khác nhau về điều kiện không gian.

Thí dụ: so sánh giá cả một loại hàng hóa giữa hai thị trường, so sánh khối lượng sản phẩm của hai doanh nghiệp trong cùng một ngành, so sánh dân số của hai địa phương..., tác dụng của sự so sánh này nhằm nêu lên ảnh hưởng của các điều kiện khác nhau đối với mức độ của hiện tượng nghiên cứu.

Ngoài ra, còn có thể so sánh các chỉ tiêu cùng loại của hai nước khác nhau trong so sánh quốc tế.

Khi tính các số tương đối so sánh, cũng cần chú ý đến tính chất có thể so sánh được giữa các chỉ tiêu.

2.4. Một số vấn đề vận dụng chung số tương đối và tuyệt đối

a. Khi sử dụng số tương đối và tuyệt đối phải xét đến đặc điểm của hiện tượng nghiên cứu để rút ra kết luận cho đúng

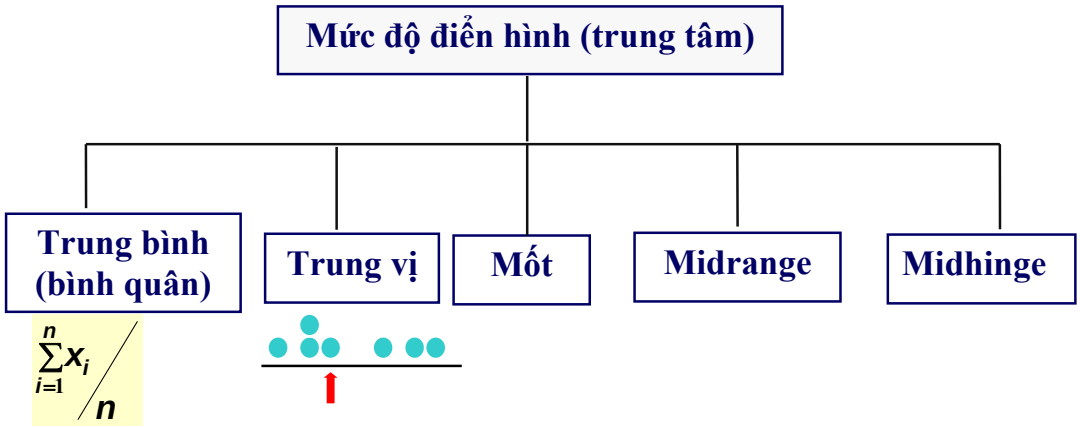
Các hiện tượng kinh tế xã hội khác nhau về nhiều mặt, quan hệ số lượng của chúng có thể thay đổi tuỳ theo điều kiện thời gian và địa điểm cụ thể. Có khi do đặc điểm của hiện tượng luôn luôn thay đổi, cho nên cùng một biểu hiện về mặt lượng nhưng có thể mang ý nghĩa khác nhau. Vì thế, khi so sánh, ta có thể gặp các đơn vị tuy giống nhau về mặt lượng nhưng lại khác nhau về mặt chất, ngược lại, cũng có khi các đơn vị có cùng một tính chất nhưng biểu hiện về mặt lượng có thể khác nhau do nhiều nguyên nhân. Tỷ lệ lao động nữ cao hơn lao động nam trong ngành giáo dục phổ thông và y tế là hợp lý, nhưng cũng tỷ lệ đó trong ngành khai thác than hay ngành vận tải lại là không hợp lý. Như vậy, khi sử dụng số tương đối phải xét đến đặc điểm của hiện tượng thì các kết luận rút ra mới đóng đắn.

b. Phải vận dụng một cách kết hợp các số tương đối với số tuyệt đối

Phần lớn các số tương đối là kết quả so sánh giữa hai số tuyệt đối, do đó, số tuyệt đối là cơ sở bảo đảm tính chất chính xác của số tương đối. Khi phân tích thống kê nếu chỉ dùng các số tương đối thì không nêu lên được tình hình thực tế của hiện tượng. Mặt khác, các nhiệm vụ phân tích thống kê cũng không thể giải quyết được tốt, nếu chỉ dùng các số tuyệt đối. Nếu sử dụng kết hợp giữa các số tương đối và số tuyệt đối thì các quan hệ hơn kém, to nhỏ, nhanh chậm, tốc độ tăng giảm, trình độ phổ biến mới được biểu hiện rõ ràng.

Hơn nữa, ý nghĩa của số tương dối còn phụ thuộc vào trị số tuyệt đối mà nó phản ánh. Thường có những trường hợp tính toán với cùng một số tuyệt đối, nhưng số tương đối tính ra có thể rất khác nhau tuỳ thuộc vào việc lựa chọn kỳ gốc so sánh. Có khi số tương đối tính ra rất lớn, nhưng ý nghĩa của nó không đáng là bao vì số tuyệt đối tương ứng với nó rất nhỏ, ngược lại có khi số tương đối tính ra rất nhỏ nhưng lại có ý nghĩa quan trọng, bởi vì số tuyệt đối ứng với nó có quy mô đáng kể.

3. Các mức độ điển hình trong thống kê



3.1. Số bình quân (Trung bình) trong thống kê

3.1.1. Khái niệm, ý nghĩa số bình quân trong thống kê

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

Việc tính toán số bình quân trong thống kê xuất phát từ tính chất của hiện tượng nghiên cứu. Các tổng thể thống kê bao gồm nhiều đơn vị cấu thành, tuy về cơ bản các đơn vị này có thể cùng một tính chất, nhưng biểu hiện cụ thể về mặt lượng theo các tiêu thức thường chênh lệch nhau. Những chênh lệch này quyết định bởi nhiều nguyên nhân, bên cạnh những nguyên nhân chung tác động đến xu hướng phát triển cơ bản của hiện tượng, còn có những nguyên nhân riêng ảnh hưởng đến mặt lượng của từng đơn vị cá biệt. Điều đó tạo nên cho mỗi đơn vị tổng thể một số đặc điểm riêng, tuy chúng vẫn tồn tại chung trong cùng một tổng thể và cùng mang một số đặc điểm chung nhất. Khi nghiên cứu thống kê ta không thể nêu lên tất cả các đặc điểm riêng biệt, mà cần tìm một mức độ có tính chất đại biểu nhất, có khả năng khái quát đặc điểm chung của cả tổng thể. Mức độ đó chính là số bình quân.

Chẳng hạn, nhiệm vụ nghiên cứu là nêu lên tình hình chung về tiền lương của lao động trong một doanh nghiệp, để phân tích tình hình đời sống, để đối chiếu và biểu hiện mối liên hệ với các chỉ tiêu sản xuất khác và để so sánh với các doanh nghiệp cùng loại. Ta đã biết các mức lương chênh lệch nhau do rất nhiều nguyên nhân, do đó không thể lấy mức lương của một lao động cá biệt nào làm mức lương đại biểu, vì các mức lương cá biệt bị ảnh hưởng bởi các nhân tố ngẫu nhiên và do đó, không giống với mức lương chung của toàn bộ lao động. Cũng không thể căn cứ vào tổng mức tiền lương trong tháng của tất cả lao động, vì số tiền này nhiều hay ít phụ thuộc vào số lượng lao động.

Có thể gạt bỏ được ảnh hưởng của các nhân tố ngẫu nhiên cá biệt cũng như ảnh hưởng của số lượng đơn vị tổng thể, bằng cách tính chỉ tiêu tiền lương bình quân, tức là đem tổng mức tiền lương trong tháng chia cho số lao động. Khi tính toán như vậy, đã coi như là tất cả mọi người cùng có một mức lương như nhau, tức là bằng mức lương bình quân. Thực ra, mức lương bình quân này có thể giống hay không giống với một mức lương cụ thể nào

đó. Tuy vậy, mức lương bình quân tính ra vẫn là một chỉ tiêu có tính chất khái quát, có khả năng đại diện được cho tất cả các mức lương khác nhau của lao động trong doanh nghiệp này trong điều kiện thời gian nhất định.

Như vậy, qua việc tính số bình quân, ta chỉ cần một trị số để nêu lên mức độ chung nhất, phổ biến nhất, có tính chất đại biểu nhất của tiêu thức nghiên cứu, không kể đến chênh lệch thực tế giữa các đơn vị tổng thể. Số bình quân không biểu hiện một mức độ cá biệt, mà là mức độ tính chung cho mỗi đơn vị tổng thể (tiền lương bình quân mỗi công nhân, năng suất lao động bình quân mỗi công nhân, giá thành bình quân mỗi đơn vị sản phẩm...).

Do số bình quân chỉ biểu hiện đặc điểm chung của cả tổng thể nghiên cứu, cho nên các nét riêng biệt có tính chất ngẫu nhiên của từng đơn vị cá biệt bị loại trừ đi. Có nghĩa là số bình quân có đặc điểm san bằng mọi chênh lệch giữa các đơn vị về trị số của tiêu thức nghiên cứu. Nhưng sự san bằng này chỉ có ý nghĩa khi ta tính cho một số khá lớn đơn vị. Nếu số bình quân được tính ra từ một số khá lớn đơn vị cùng loại, nó thực sự trở thành mức độ đại biểu của các đơn vị đó. Còn nếu số đơn vị quá ít, các kết luận rút ra sẽ kém chính xác. Như vậy, việc tính số bình quân là một trường hợp vận dụng định luật số lớn.

Số bình quân có một vị trí và ý nghĩa rất quan trọng trong lý luận và trong công tác nghiên cứu thực tế. Nó được dùng trong mọi công tác nghiên cứu kinh tế, nhằm nêu lên đặc điểm chung của hiện tượng kinh tế xã hội số lớn trong điều kiện thời gian và địa điểm cụ thể. Ta thường gặp các chỉ tiêu như: giá thành bình quân, giá cả bình quân, tốc độ chu chuyển vốn bình quân, năng suất lao động bình quân, năng suất thu hoạch bình quân và rất nhiều chỉ tiêu bình quân khác, là những chỉ tiêu rất cần thiết trong phân tích hoạt động kinh tế. Mác cũng sử dụng các khái niệm bình quân trong nhiều tác phẩm như: lợi nhuận bình quân, giá trị thặng dư bình quân, độ dài ngày lao động bình quân...

Việc sử dụng số bình quân tạo điều kiện để so sánh giữa các hiện tượng không có cùng một quy mô, như so sánh năng suất lao động và tiền lương bình quân của công nhân hai doanh nghiệp, so sánh năng suất thu hoạch lúa giữa hai địa phương... Trong các trường hợp trên, việc so sánh giữa hai số tuyệt đối không thực hiện được hoặc đôi khi không có ý nghĩa.

Số bình quân còn được dùng để nghiên cứu các quá trình biến động qua thời gian, nhất là các quá trình sản xuất. Sự biến động của số bình quân qua thời gian có thể cho ta thấy được xu hướng phát triển cơ bản của hiện tượng số lớn, tức là của đại bộ phận các đơn vị tổng thể, trong khi từng đơn vị cá biệt không thể giúp ta thấy rõ điều đó.

Số bình quân không những chỉ dùng trong công tác thống kê mà còn cả trong công tác kế hoạch. Rất nhiều chỉ tiêu kế hoạch được biểu hiện bằng số bình quân. Khi phân tích thực hiện kế hoạch cũng có thể lấy số bình quân làm cơ sở so sánh, phân biệt các đơn vị tiên tiến và lạc hậu, phát triển các khả năng tiềm tàng trong sản xuất.

Số bình quân chiếm một vị trí quan trọng trong việc vận dụng nhiều phương pháp phân tích thống kê. Các trường hợp phân tích biến động, phân tích mối liên hệ, dự đoán thống kê, điều tra chọn mẫu... đều sử dụng rất nhiều số bình quân trong các công thức tính toán.

3.1.2. Các loại số bình quân

Trên thực tế, có nhiều loại số bình quân, mỗi loại có công thức tính khác nhau. Việc sử dụng loại nào không phải chỉ căn cứ vào mục đích nghiên cứu, ý nghĩa kinh tế của chỉ tiêu bình quân mà còn phải căn cứ vào đặc điểm của hiện tượng và nguồn tài liệu sẵn có để chọn công thức tính toán thích hợp. Thống kê học thường dùng các loại số bình quân sau đây: số bình quân cộng, số bình bình quân nhân, mốt và trung vị.

a. Số bình quân cộng (trung bình cộng)

Số bình quân cộng là số bình quân được tính bằng công thức số trung bình cộng trong toán học. Số bình quân cộng được dùng nhiều nhất trong nghiên cứu thống kê và chịu ảnh hưởng của lượng biến đột xuất. Tùy theo đặc điểm của dữ liệu mà có các trường hợp cụ thể như sau:

- *Số bình quân cộng giản đơn (hay trung bình cộng giản đơn):* được vận dụng khi tính từ tài liệu ban đầu hoặc dãy số phân phối có tần số bằng nhau. Công thức tính như sau:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} \quad \text{hay là} \quad \bar{x} = \frac{\sum x_i}{n} \qquad (3.5)$$

Trong đó: x_i (i = 1, 2,..., n) - các lượng biến

$\bar{x}$ - số bình quân

n - số đơn vị tổng thể

Thí dụ: Tính năng suất lao động bình quân của một tổ công nhân gồm 6 người, trong đó người công nhân thứ nhất đã sản xuất được 50 sản phẩm, người thứ hai: 55, người thứ ba: 60, người thứ tư: 65, người thứ năm: 70 và người thứ sáu: 72 sản phẩm.

Theo công thức trên:

$$\bar{x} = \frac{50 + 55 + 60 + 65 + 70 + 72}{6} = \frac{372}{6} = 62 \text{ sản phẩm}$$

- *Số bình quân cộng gia quyền (hay trung bình cộng gia quyền):* Vận dụng khi các lượng biến có tần số khác nhau. Trong trường hợp này, mỗi lượng biến có thể gặp nhiều lần, muốn tính được số bình quân cộng, trước hết phải đem nhân mỗi lượng biến x_i với tần số tương ứng f_i, rồi mới đem cộng lại và chia cho số đơn vị tổng thể. Trong thống kê, việc nhân các lượng biến x_i với các tần số tương ứng f_i được gọi là gia quyền, còn các tần số được gọi là quyền số.

Công thức số bình quân cộng gia quyền:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + ... + x_n f_n}{f_1 + f_2 + ... + f_n} \quad \text{hay là :} \quad \bar{x} = \frac{\sum x_i f_i}{\sum f_i} \qquad (3.6)$$

Trong đó: x_i (i = 1, 2,..., n) - các lượng biến

$\overline{x}$ - số bình quân

f_i (i = 1, 2,..., n) - các quyền số (tần số)

Thí dụ: Tính năng suất lao động bình quân của công nhân theo tài liệu sau:

Bảng 1

Năng suất lao động (SP) (x_i)	***Số công nhân (f_i)***	***Nhân lượng biến với quyền số (x_if_i)***
50	3	150
55	5	275
60	10	600
65	12	780
70	7	490
72	3	216
Cộng	$\sum f_i = 40$	$\sum x_i f_i = 2511$

Theo công thức (3.6) tính ra:

$$\overline{x} = \frac{(50 \text{ x } 3) + (55 \text{ x } 5) + (60 \text{ x } 10) + (65 \text{ x } 12) + (70 \text{ x } 7) + (72 \text{ x } 3)}{3 + 5 + 10 + 12 + 7 + 3}$$

$$= \frac{150 + 275 + 600 + 780 + 490 + 216}{40} = \frac{2511}{40} = 62{,}8 \text{ sả n phẩm}$$

Qua hai công thức trên, ta thấy số bình quân cộng giản đơn và số bình quân cộng gia quyền khác nhau ở chỗ có hay không có quyền số trong quá trình tính toán. Thực ra, số bình quân cộng giản đơn chỉ là một trường hợp của số bình quân cộng gia quyền, vì khi các quyền số $f_1 = f_2 = f_3 = ... = f_n$, có thể giản đơn đi trong quá trình tính toán. Quyền số của số bình quân có một vai trò quan trọng, bởi vì trị số bình quân không những phụ thuộc vào các lượng biến, mà còn phụ thuộc cả vào quyền số của các lượng biến này (xem hai kết quả tính toán ở trên).

Đôi khi, nguồn tài liệu đã có sẵn các đại lượng $M_i = x_if_i$ thì việc vận dụng công thức số bình quân cộng gia quyền sẽ dễ dàng hơn. Thí dụ, tính năng suất lao động bình quân từ tài liệu sau:

Bảng 2

Tổ	***Số công nhân (f_i)***	***Sản lượng ($M_i = x_if_i$)***
1	3	150
2	5	275
3	10	600

4	12	780
5	7	490
6	3	216
Cộng	$\sum f_i = 40$	$\sum x_i f_i = \sum M_i = 2511$

Dựa theo công thức (3.6), ta có:

$$\overline{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{\sum M_i}{\sum f_i} = \frac{2511}{40} = 62{,}8 \text{ sản phẩm}$$

- Tính số bình quân cộng từ một dãy phân bố tần số có khoảng cách tổ:

Trường hợp này trong mỗi tổ có một phạm vi lượng biến, cho nên cần có một lượng biến đại diện để làm căn cứ tính toán. Người ta thường lấy các trị số giữa làm lượng biến đại diện cho từng tổ, và tính theo công thức:

$$\text{Trị số giữa mỗi tổ} = \frac{x_{min} + x_{max}}{2}$$

Trong đó: x_{min} và x_{max} là giới hạn dưới và giới hạn trên của mỗi khoảng cách tổ. Trị số giữa này được coi là lượng biến (x_i) đại diện của mỗi tổ.

Có thể lấy thí dụ tính toán sau:

Bảng 3

Năng suất lao động (kg)	***Trị số giữa (x_i)***	***Số công nhân (f_i)***	***Nhân trị số giữa với quyền số ($x_i f_i$)***
400 – 500	450	10	4500
500 – 600	550	30	16500
600 – 700	650	45	29250
700 – 800	750	80	60000
800 – 900	850	30	25500
900 – 1000	950	5	4750
Cộng		***200 ($\sum f_i$)***	***140500 ($\sum x_i f_i$)***

Trong bảng trên, trị số giữa của các tổ tính như sau:

$$\text{Tổ thứ nhất: } x_1 = \frac{400 + 500}{2} = 450 \text{ kg}$$

$$\text{Tổ thứ hai: } x_2 = \frac{500 + 600}{2} = 550 \text{ kg} \ldots..$$

Năng suất lao động bình quân được tính theo công thức (3.6):

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{140500}{200} = 702{,}5 \text{ kg}$$

Việc thay thế các phạm vi lượng biến bằng trị số giữa dựa trên cơ sở giả định rằng các lượng biến được phân phối đều đặn trong phạm vi mỗi tổ, và do đó trị số giữa mỗi tổ được coi như số bình quân cộng giản đơn của các đơn vị trong tổ đó. Trong thực tế, sự phân phối đều đặn này ít có, cho nên thường có một sai số nhất định giữa số bình quân của tổ và trị số giữa của tổ, có ảnh hưởng đến tính chất chính xác của số bình quân chung. Những sai số đó lớn hay nhỏ phụ thuộc vào khoảng cách tổ và đặc điểm phân phối của các tổ. Tuy nhiên, dưới tác dụng tính toán của số bình quân chung, các sai số được bù trừ nhau và vẫn cho kết quả sử dụng được.

Trường hợp các khoảng cách tổ được hình thành theo các lượng biến liên tục nhưng không có giới hạn trên và dưới trùng nhau, như: 600 - 699,99; 700 - 799,99; 800 - 899,99... thì trị số giữa tính theo các giới hạn dưới của hai tổ kế tiếp nhau. Thí dụ: $x_1 = \frac{600 + 700}{2}$; $x_2 = \frac{700 + 800}{2}$

Đối với những dãy số lượng biến có khoảng cách tổ mở (tức là tổ thứ nhất và tổ cuối cùng không có giới hạn dưới và giới hạn trên), việc tính trị số giữa của các tổ này phải căn cứ vào các khoảng cách tổ gần chúng nhất mà tính toán cho hợp lý.

- Tính số bình quân chung từ các số bình quân tổ

Trường hợp này thường gặp trong nghiên cứu thống kê, như: tính năng suất lúa bình quân của toàn hợp tác xã trên cơ sở năng suất lúa bình quân của từng loại ruộng, tính năng suất lao động bình quân chung của cả doanh nghiệp trên cơ sở đã có năng suất lao động bình quân của các tỏ, đội sản xuất... Số bình quân chung sẽ là số bình quân cộng gia quyền của các số bình quân tổ, trong đó quyền số là số đơn vị mỗi tổ.

Giả sử có các số bình quân tổ:

$$\bar{x}_1 = \frac{\sum x_1}{n_1}; \quad \bar{x}_2 = \frac{\sum x_2}{n_2}; ...; \quad \bar{x}_n = \frac{\sum x_t}{n_t}$$

Suy ra: $\sum x_1 = \bar{x}_1 n_1; \quad \sum x_2 = \bar{x}_2 n_2; ...; \quad \sum x_t = \bar{x}_t n_t$

Số bình quân cộng chung sẽ bằng:

$$\bar{x} = \frac{\sum x_1 + \sum x_2 + ... + \sum x_t}{n_1 + n_2 + ... + n_t} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + ... + ...\bar{x}_t n_t}{n_1 + n_2 + ... + n_t} = \frac{\sum \bar{x}_i n_i}{\sum n_i} \qquad (3.7)$$

b. Số bình quân điều hoà (trung bình điều hòa)

Số bình quân điều hoà cũng có nội dung kinh tế như số bình quân cộng, tính được bằng cách đem chia tổng các lượng biến của tiêu thức cho số đơn vị tổng thể. Nhưng ở đây vì không có sẵn tài liệu về số đơn vị tổng thể, nên phải dựa vào các tài liệu khác để tính ra.

- Số bình quân điều hoà gia quyền được tính theo công thức:

$$\bar{x} = \frac{M_1 + M_2 + ... + M_n}{\frac{M_1}{x_1} + \frac{M_2}{x_2} + ... + \frac{M_n}{x_n}} = \frac{\sum M_i}{\sum \frac{M_i}{x_i}} \quad \text{hay là : } \bar{x} = \frac{\sum M_i}{\sum \frac{1}{x_i} M_i} \qquad (3.8)$$

Trong đó: x_i (i = 1, 2,..., n) – các lượng biến

$\bar{x}$ - số bình quân

$M_i = x_i f_i$ - tổng các lượng biến của tiêu thức, tức là quyền số của số bình quân điều hoà

Thí dụ: Có tài liệu về năng suất lao động của các tổ công nhân trong một doanh nghiệp như sau:

Bảng 4

Tổ công nhân	***Năng suất lao động mỗi công nhân (tấn) (x_i)***	***Sản lượng (tấn) (M_i)***
I	11	220
II	12	264
III	13	312

Muốn tính được năng suất lao động bình quân (chung cho cả ba tổ) phải lấy tổng sản lượng chia cho tổng số công nhân. ở đây không có tài liệu về số công nhân nhưng dựa vào các tài liệu khác có thể tính ra như sau:

$$\text{Số công nhân tổ I} = \frac{\text{Sản lượng tổ I}}{\text{NSLĐ mỗi CN tổ I}} = \frac{220}{11} = 20 \text{ người}$$

Cũng theo cách tính trên, số công nhân tổ II bằng 22 người và tổ III bằng 24 người. Vì vậy, năng suất lao động bình quân của công nhân toàn doanh nghiệp tính như sau:

$$\text{Năng suất lao động bình quân} = \frac{\text{Tổng sản lượng}}{\text{Tổng số công nhân}} = \frac{\text{Tổng sản lượng}}{\text{Tổng } \frac{\text{Sản lượng mỗi tổ}}{\text{NSLĐ của CN mỗi tổ}}}$$

$$\bar{x} = \frac{\Sigma M_i}{\sum \frac{M_i}{x_i}} = \frac{220 + 264 + 312}{\frac{220}{11} + \frac{264}{12} + \frac{312}{13}} = \frac{796}{20 + 22 + 24} = \frac{796}{66} = 12{,}06 \text{ tấn}$$

- Số bình quân điều hoà giản dơn

Trường hợp các quyền số M_i bằng nhau, tức là khi $M_1 = M_2 = ... = M_n = M$, công thức (3.8) có thể thay đổi như sau:

$$\bar{x} = \frac{\sum M_i}{\sum \frac{M_i}{x}} = \frac{nM}{M\sum \frac{1}{x_i}} = \frac{n}{\sum \frac{1}{x_i}} \qquad (3.9)$$

Công thức (3.9) được gọi là số bình quân điều hoà giản đơn, trong đó n là số lượng biến.

Thí dụ: một nhóm 3 công nhân cùng sản xuất với thời gian lao động như nhau. Người thứ nhất sản xuất một sản phẩm hết 15 phút, người thứ hai 20 phút và người thứ ba là 30 phút. Muốn tính được thời gian hao phí bình quân để sản xuất ra một đơn vị sản phẩm, cần phải đem tổng số thời gian sản xuất chia cho số sản phẩm đã sản xuất ra. ở đây, lượng biến x_i là thời gian hao phí của mỗi công nhân để sản xuất ra một đơn vị sản phẩm, còn thời gian sản xuất của mỗi công nhân bằng nhau, tức là $M_1 = M_2 = M_3$.

Vì vậy, quá trình tính toán có thể đơn giản và ta có:

$$\bar{x} = \frac{n}{\sum \frac{1}{x_i}} = \frac{3}{\frac{1}{15} + \frac{1}{20} + \frac{1}{30}} = 20 \text{ phút}$$

Qua các thí dụ trên, ta nhận thấy quyền số của số bình quân điều hoà thực ra không phải là một đại lượng giản đơn, mà là tích của 2 nhân tố: lượng biến (x_i) với tần số các lượng biến đó f_i, tức là $M_i = x_i f_i$. Do đó, khi đem chia các quyền số M_i cho các lượng biến x_i, ta tính ra được số đơn vị tổng thể:

$$\frac{M_i}{x_i} = \frac{x_i f_i}{x_i} = f_i$$

Như khi chia sản lượng mỗi tổ cho năng suất lao động mỗi tổ, sẽ được số công nhân tổ đó, chia số thời gian lao động cho số thời gian hao phí để sản xuất một đơn vị sản phẩm, sẽ tính được số sản phẩm.

Như vậy, số bình quân điều hoà thường được vận dụng khi nào không có tài liệu về số đơn vị tổng thể, mà chỉ có tài liệu về các lượng biến và chỉ tiêu về tổng các lượng biến của tiêu thức.

c. Số bình quân nhân (trung bình nhân)

Số bình quân nhân là số bình quân của những đại lượng có quan hệ tích số với nhau. Có hai công thức tính toán như sau:

- Số bình quân nhân giản đơn được tính theo công thức:

$$\bar{x} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} = \sqrt[n]{\prod x_i} \qquad (3.10)$$

Trong đó: x_i (i = 1, 2,..., n) – các lượng biến

$\overline{x}$ - số bình quân

Thí dụ: Có tốc độ phát triển sản xuất của một doanh nghiệp như sau:
Năm 1978 so với năm 1977 bằng 116%
Năm 1979 so với năm 1978 bằng 111%
Năm 1980 so với năm 1979 bằng 112%
Năm 1981 so với năm 1980 bằng 113%
Năm 1982 so với năm 1981 bằng 112%
Năm 1983 so với năm 1982 bằng 111%

Ở đây, các tốc độ phát triển sản xuất (tức là số tương đối động thái) không cộng được với nhau để tính tốc độ phát triển bình quân, vì chúng là các số tương đối có gốc so sánh khác nhau. Nhưng chúng lại có quan hệ tích số với nhau, bởi vì tích của chúng sẽ cho ta một số tương đối động thái mới, nói lên tốc độ phát triển sản xuất của doanh nghiệp trong thời kỳ dài hơn. Vì vậy, muốn tính tốc độ phát triển bình quân hàng năm về sản xuất của doanh nghiệp, trước hết ta phải nhân các tốc độ phát triển sản xuất hàng năm, sau đó khai căn theo công thức (3.10)

Cụ thể là:

$$\overline{x} = \sqrt[6]{1,16 \times 1,11 \times 1,12 \times 1,13 \times 1,12 \times 1,11}$$

Ta có: $\overline{x}$ = 1,125, có nghĩa là tốc độ phát triển sản xuất bình quân hàng năm của doanh nghiệp là 1,125 lần (hay 112,5%)

- Số bình quân nhân gia quyền

Khi các lượng biến (x_i) có các tần số (f_i) khác nhau, ta có công thức số bình quân nhân gia quyền (lúc này f_i là quyền số):

$$\overline{x} = \sqrt[\sum f_i]{x_1^{f_1} \times x_2^{f_2} \times \ldots \times x_n^{f_n}} = \sqrt[\sum f_i]{\prod x_i^{f_i}} \qquad (3.11)$$

Thí dụ: Trong thời gian 10 năm, tốc độ phát triển sản xuất của một doanh nghiệp như sau: có 5 năm phát triển với tốc độ mỗi năm là 110%, có hai năm với tốc độ 125% và ba năm với tốc độ 115%. Để tính tốc độ phát triển sản xuất bình quân hàng năm, ta dùng công thức (3.11):

$$\overline{x} = \sqrt[10]{(1,1)^5 \times (1,25)^2 \text{ x } (1,15)^3}$$

Ta có: $\overline{x}$ = 1,144 (hay 114,4%)

Số bình quân nhân được dùng trong trường hợp các lượng biến có quan hệ tích số với nhau. ứng dụng trong thống kê kinh tế xã hội, công thức số bình quân này thường chỉ dùng để tính các tốc độ phát triển bình quân.

3.2. Mốt (Mode)

Mốt là biểu hiện của một tiêu thức được gặp nhiều nhất trong một tổng thể hay trong một dãy số phân phối. Đối với một dãy số phân phối, mốt là lượng biến có tần số lớn nhất. Trị số của mốt không phụ thuộc vào trị số của tất cả các lượng biến trong dãy số, mà được xác định do sự sắp xếp các lượng biến trong dãy số này.

Thí dụ: Theo tài liệu ở bảng 3.1 ta dễ dàng nhận thấy Mốt về năng suất lao động là 65 sản phẩm vì mức năng suất lao động này có tần số lớn nhất.

Đối với một dãy số phân phối có khoảng cách tổ, muốn tìm mốt trước hết cần xác định tổ có mốt, tức là tổ có tần số lớn nhất. Sau đó, tính trị số gần đúng của mốt theo công thức:

$$M_o = x_{M_o(\min)} + h_{M_o} \frac{f_{M_o} - f_{M_0-1}}{(f_{M_o} - f_{M_0-1}) + (f_{M_o} - f_{M_0+1})} \qquad (3.12)$$

Trong đó: M_0 - ký hiệu của mốt

$x_{M_o(\min)}$ - giới hạn dưới của tổ có mốt

h_{M_o} - trị số khoảng cách tổ có mốt

f_{M_0} - tần số của tổ có mốt

f_{M_0-1} - tần số của tổ đứng trước tổ có mốt

f_{M_0+1} - tần số của tổ đứng sau tổ có mốt

Thí dụ: Theo tài liệu phân tổ công nhân theo năng suất lao động trong bảng 3, trước hết có thể xác định mốt ở vào tổ thứ tư (700 - 800 tấn), vì tổ này có tần số lớn nhất (80 công nhân). Từ đó xác định tiếp:

$$x_{M_o(\min)} = 700;\ h_{M_o} = 100;\ f_{M_0} = 80;\ f_{M_0-1} = 45;\ f_{M_0+1} = 30$$

Thay số liệu vào công thức (3.12):

$$M_0 = 700 + 100\frac{80 - 45}{(80 - 45) + (80 - 30)} = 700 + 100\frac{35}{85} = 700 + 41{,}2$$

$M_0 = 741{,}2$ tấn

Trong trường hợp dãy số lượng biến có khoảng cách tổ không đều nhau, mốt vẫn được tính theo công thức (3.12). Nhưng việc xác định tổ có mốt và tính toán không căn cứ vào tần số lớn nhất, mà căn cứ vào mật độ phân phối (tức là tỷ số giữa các tần số chia cho trị số khoảng cách tổ). Thí dụ:

Bảng 5

Năng suất lao động (tấn)	***Trị số khoảng cách tổ (tấn)***	***Số công nhân***	***Mật độ phân phối***
400 – 450	50	10	0,2

450 – 500	50	15	0,3
500 – 600	100	15	0,15
600 – 800	200	30	0,15
800 – 1200	400	5	0,0125

Như vậy, mốt ở tổ thứ hai là tổ có mật độ phân phối lớn nhất. Tính theo công thức (3.12) ta có:

$$M_0 = 450 + 50\frac{0,3 - 0,2}{(0,3 - 0,2) + (0,3 - 0,15)} = 450 + 50\frac{0,1}{0,25}$$

$M_0 = 450 + 20 = 470$ (tấn)

Trong nghiên cứu thống kê, mốt là mức độ có tác dụng bổ sung hoặc thay thế cho việc tính số bình quân cộng, trong trường hợp việc tính số bình quân này gặp khó khăn, không bảo đảm chính xác hoặc không có ý nghĩa. Mốt có khả năng nêu lên mức độ phổ biến nhất của hiện tượng, đồng thời bản thân nó lại không san bằng, bù trừ chênh lệch giữa các lượng biến. Như khi đăng ký giá cả một mặt hàng trên thị trường, có thể không cần tính theo số bình quân cộng, mà chỉ cần ghi giá phổ biến của mặt hàng trong thời gian đó. Có thể dùng mốt để xác định mức lương phổ biến nhất trong một doanh nghiệp, tìm loại điểm nào của học sinh đạt được nhiều nhất sau một kỳ thi.

Ngoài ra, mốt bảo đảm được ý nghĩa thực tế hơn các tính toán khác, vì nó không chịu ảnh hưởng của tất cả các lượng biến, nhất là các lượng biến đột xuất (quá lớn hay quá nhỏ). Như một mức lương cao đột xuất có thể làm ảnh hưởng đến việc tính số bình quân cộng, nhưng không ảnh hưởng đến mốt. Nhưng cũng vì lý do trên, mốt có nhược điểm là kém nhạy bén đối với sự biến thiên của tiêu thức. Cho nên mốt chỉ được vận dụng đối với một tổng thể tương đối nhiều đơn vị. Mặt khác, nếu dãy số lượng biến có đặc điểm phân phối không bình thường (có quá nhiều điểm tập trung hoặc không có điểm chính tập trung các trị số) thì cũng không nên xác định mốt.

Mốt còn có nhiều tác dụng trong việc tổ chức phục vụ nhu cầu một cách hợp lý. Các tổ chức sản xuất và thương mại cần điều tra và cung ứng đầy đủ các mặt hàng tiêu thụ nhiều nhất, như cỡ giầy, cỡ kiểu quần áo...

Cuối cùng, mốt còn được dùng làm một trong các chỉ tiêu nêu lên đặc trưng của dãy số phân phối (xem phần cuối của chương này).

3.3. Trung vị (Median)

Trung vị là một lượng biến tiêu thức của đơn vị đứng ở vị trí giữa trong một dãy số phân phối. Trung vị phân chia dãy số phân phối thành hai phần (phần trên và phần dưới trung vị), mỗi phần có cùng một số đơn vị tổng thể bằng nhau.

Nếu số đơn vị tổng thể lẻ ($n = 2m + 1$), trung vị sẽ là lượng biến của đơn vị đứng ở vị trí thứ $m + 1$, tức là lượng biến x_{m+1}. Giả sử có mức năng suất lao động của 5 công nhân: 40,

45, 50, 55 và 60 sản phẩm. Số trung vị là mức năng suất lao động của người công nhân thứ 3, tức là Me = 50 sản phẩm.

Nếu số đơn vị tổng thể chẵn (n = 2m), xác định trung vị căn cứ vào lượng biến của hai đơn vị cùng đứng ở vị trí giữa (đơn vị thứ m và m + 1) cộng lại rồi chia đôi, tức là $\frac{x_m + x_{m+1}}{2}$. Thí dụ, có mức năng suất lao động của 6 công nhân: 40, 45, 50, 55, 60 và 65 sản phẩm. Số trung vị bằng $\frac{50+55}{2}$ = 52,5 sản phẩm.

Trong một dãy số phân phối có khoảng cách tổ, muốn tìm trung vị, trước hết phải xác định tổ có trung vị. Đó là tổ có chứa lượng biến của đơn vị ở vị trí giữa trong tổng số các đơn vị của dãy số. Dùng phương pháp cộng dồn các tần số của các tổ thứ nhất, thứ hai, thứ ba... sẽ tìm ra được tần số tích luỹ bằng hoặc vượt một nửa tổng các tần số. Tổ tương ứng với tần số tích luỹ này chính là tổ có trung vị. Sau đó, trị số gần đúng của trung vị được tính theo công thức sau:

$$Me = x_{Me(min)} + h_{Me}\frac{\frac{\sum f}{2} - S_{(Me-1)}}{f_{Me}} \qquad (3.13)$$

Trong đó: Me - ký hiệu số trung vị

$x_{Me(min)}$ - giới hạn dưới của tổ có số trung vị

h_{Me} trị số khoảng cách tổ có số trung vị

$\sum f$ - tổng các tần số của dãy số lượng biến (số đơn vị tổng thể)

$S_{(Me-1)}$ – tổng các tần số của các tổ đứng trước tổ có số trung vị

f_{Me} – tần số của tổ có số trung vị

Lấy thí dụ theo tài liệu trong bảng 3. Tổng số lao động là 200, vậy người ở vị trí giữa là công nhân thứ 100 và 101. Cộng dồn các tần số (xem bảng 6) ta xác định người công nhân thứ 100 và 101 thuộc vào tổ thứ tư và đó chính là tổ có số trung vị.

Bảng 6

Năng suất lao động (kg)	***Số lao động***	***Tần số tích luỹ***
400 – 500	10	10
500 – 600	30	40
600 – 700	45	85
700 – 800	80	165
800 – 900	30	195
900 – 1000	5	200
Cộng	***200***	

Từ đó, tiếp tục xác định các ký hiệu:

$$x_{Me(min)} = 700;\ i_{Me} = 100;\ \sum f = 100;\ S_{(Me-1)} = 85;\ f_{Me} = 80$$

Thay số liệu vào công thức (3.13) tính ra:

$$Me = 700 + 100\frac{\frac{200}{2} - 85}{80} = 700 + 100\frac{15}{80} = 700 + 18{,}75$$

Me = 718,75 tấn

Việc tính số trung vị, chủ yếu căn cứ vào sự sắp xếp theo thứ tự các lượng biến. Số trung vị cũng nêu lên mức độ đại biểu của hiện tượng, mà không san bằng bù trừ chênh lệch giữa các lượng biến. Cho nên nó có khả năng bổ sung hoặc thay thế cho số bình quân cộng, khi ta không có một cách chính xác toàn bộ các lượng biến. Chỉ cần bảo đảm được sự phân phối các đơn vị theo thứ tự lượng biến là có thể tính số trung vị, nhất là đối với các dãy số phân phối có khoảng cách tổ mở và không đều nhau, việc tính số trung vị tỏ ra thuận lợi hơn. Giả sử ta có dãy số phân phối như sau:

Bảng 7

Năng suất lao động (tấn)	Số lao động
Dưới 50	10
50 – 60	30
60 – 85	40
85 – 110	15
110 trở lên	5

Mọi dãy số như trên làm cho việc tính số bình quân cộng phải dựa trên cơ sở giả định rất lớn, nhưng có thể thoạt trông mà xác định ngay rằng số trung vị nằm ở tổ thứ ba và nhanh chóng tính ra Me = 66,25 tấn.

Việc tính số trung vị cũng còn có tác dụng loại trừ ảnh hưởng của những lượng biến đột xuất. Chẳng hạn, một mức lương cao cá biệt trong dãy số lượng biến không làm ảnh hưởng đến việc đánh giá mức lương chung. Vì vậy, ta có thể dùng số trung vị khi tiêu thức nghiên cứu biến thiên quá nhiều, hoặc đối với một dãy số có quá ít đơn vị.

Số trung vị cũng là một trong các chỉ tiêu dùng để nêu lên đặc trưng của một dãy số phân phối (xem phần cuối của chương này).

Một tính chất toán học đáng chú ý của số trung vị là: tổng các độ lệch tuyệt đối giữa các lượng biến với số trung vị là một trị số nhỏ nhất (so với bất kỳ tổng các độ lệch giữa các hiện tượng biến với một đại lượng nào khác - kể cả số bình quân cộng). Tức là:

$$\sum |x_i - Me| = \min \quad hay: \quad \sum |x_i - Me| f_i = \min$$

Tính chất trên đây được ứng dụng trong nhiều công tác kỹ thuật và phục vụ công cộng, như bố trí các nhà câu lạc bộ, nhà trẻ, cửa hàng, ống dẫn nước, trạm đỗ xe, ô tô buýt... sao cho được ở vị trí thuận lợi để có thể phục vụ được nhiều người mà tiết kiệm nhất.

3.4. Midrange: Cũng là một mức độ điển hình, được tính bằng trung bình cộng giữa lượng biến lớn nhất và nhỏ nhất của tiêu thức nghiên cứu, vì vậy chịu ảnh hưởng của lượng biến đột xuất. Công thức tính:

$$\frac{x_{min} + x_{max}}{2} \qquad (3.14)$$

3.5. Điều kiện vận dụng các mức độ điển hình của hiện tượng

Việc dùng các mức độ điển hình như số bình quân, mốt, trung vị có tác dụng tiết kiệm lời, đơn giản hoá sự giải thích đặc điểm của hiện tượng. Nhưng việc lạm dụng các loại số này sẽ dẫn đến việc sử dụng nó có tính chất giả tạo và không có căn cứ khoa học,

Trên thực tế, tuy rằng số bình quân có nhiều tác dụng quan trọng đối với nghiên cứu thống kê, nhưng bản thân nó cũng có nhược điểm đáng chú ý: số bình quân thường mang một ý nghĩa chung rất khái quát cho toàn bộ tổng thể nghiên cứu, vì nó đã san bằng mọi chênh lệch thực tế giữa các đơn vị cá biệt, và tổng thể phức tạp trở thành hết sức đơn giản. Chính đây là chỗ dễ bị lợi dụng trong thống kê. Số bình quân cũng không thể là chỉ tiêu vạn năng, một mức độ "tiêu chuẩn" có tính chất ổn định. Cho nên vấn đề đặt ra là phải biết cách vận dụng một cách khoa học và chính xác số bình quân, phát huy ưu điểm và khắc phục nhược điểm của nó, bảo đảm phân tích thống kê đạt kết quả cao nhất.

Sau đây là các điều kiện vận dụng các mức độ điển hình trong thống kê.

a. Chỉ được tính ra từ tổng thể đồng chất

Tổng thể đồng chất bao gồm nhiều đơn vị, phần tử hoặc hiện tượng có cùng chung một tính chất, thuộc cùng một loại hình kinh tế xã hội, xét theo một tiêu thức nào đó. Thí dụ, một tổng thể công nhân sản xuất công nghiệp phải bao gồm những người lao động trong doanh nghiệp trực tiếp sáng tạo ra sản phẩm công nghiệp hoặc trực tiếp tham gia vào quá trình sản xuất công nghiệp. Đây là một tổng thể đồng chất, mặc dù các công nhân có thể khác nhau về các mặt tuổi tác, giới tính, tuổi nghề, trình độ kỹ thuật, trình độ văn hoá..., nhưng đều có mặt cơ bản giống nhau là cùng tham gia sản xuất sản phẩm công nghiệp trong một doanh nghiệp nhất định.

Các đơn vị trong tổng thể đồng chất có cùng một tính chất, cho nên mới có thể có cùng một lượng tương ứng đại diện cho các đơn vị. Số bình quân tính được từ tổng thể đồng

chấy như vậy mới có đầy đủ ý nghĩa là mức độ đại biểu, có thể thay thế cho các mức độ khác nhau trong tổng thể và mới cho ta một nhận thức đúng đắn về bản chất của hiện tượng. Trái lại, không được tính số bình quân từ tổng thể bao gồm các đơn vị khác nhau về tính chất, phát triển trong các điều kiện khác nhau, vì mức độ này không những không có ý nghĩa thực tế mà còn có khi làm cho ta hiểu sai lệch bản chất của hiện tượng. Người ta gọi đó là những số bình quân giả tạo, không đầy đủ tính chất đại biểu.

b. Cần được vận dụng kết hợp với các số bình quân tổ hoặc dãy số phân phối

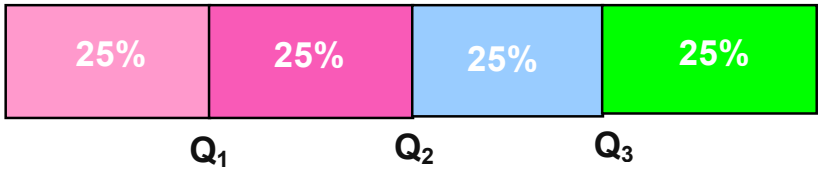
Số bình quân chung chỉ phản ánh đặc trưng chung của toàn bộ tổng thể nghiên cứu, bỏ qua những chênh lệch thực tế giữa các đơn vị tổng thể. Khi cần so sánh tổng thể giữa hai thời gian hoặc địa điểm, bản thân số bình quân chung cũng không thể giải thích được hết nguyên nhân và xu hướng phát triển của hiện tượng.

Mặt khác, nếu ta chỉ xét hiện tượng qua mức độ bình quân, các chênh lệch thực tế coi như bị san bằng, do đó những đơn vị có mức độ cao thấp khác nhau đều bị số bình quân che lấp. Điều đó hạn chế tác dụng của phân tích thống kê, thậm chí nếu không chú ý còn có thể rút ra kết luận sai lệch. Nhiệm vụ nghiên cứu của thống kê là, đi đôi với việc tính số bình quân để tìm hiểu mức đại biểu chung, còn phải nêu được những đơn vị hoặc bộ phận đạt mức độ cao hơn hoặc thấp hơn mức bình quân, tức là vạch ra được những đơn vị tiền tiến và lạc hậu. Điều đó rất cần cho công tác lãnh đạo chung và chỉ đạo riêng, phát hiện các mầm mống mới phát sinh, vạch ra những bộ phận lạc hậu đang kìm hãm sự phát triển chung.

Vì những lý do trên, khi phân tích thống kê ta không thể chỉ thoả mãn với con số bình quân chung, mà cần bổ sung phân tích bằng các số bình quân tổ hoặc dãy số phân phối, tuỳ theo mục đích nghiên cứu. Số bình quân tổ là số bình quân tính riêng cho từng tổ, từng bộ phận cấu thành tổng thể. Nó giúp đi sâu nghiên cứu đặc điểm riêng từng tổ hoặc bộ phận, giải thích được nguyên nhân phát triển chung của hiện tượng. Còn dãy số phân phối giúp đi sâu vào từng đơn vị hoặc bộ phận có mức độ khác nhau. Cũng trên cơ sở dãy số phân phối còn có thể xác định được mức bình quân tiên tiến, là mức bình quân của những đơn vị đã vượt mức bình quân chung.

4. Tứ phân vị

Là các lượng biến đứng ở các vị trí đặc biệt, chia dãy số thành bốn phần bằng nhau. Trong đó Q2 chính là trung vị; có 25% số đơn vị tổng thể có lượng biến nhỏ hơn Q_1; 25% số đơn vị có lượng biến lớn hơn Q_3 và 50% số đơn vị có
lượng biến nằm trong khoảng từ Q_1 đến Q_3.



Vị trí của tứ phân vị thứ i được xác định bằng công thức:

$$\frac{i(n+1)}{4} \qquad (3.15)$$

Thí dụ quan sát 9 đơn vị và có dữ liệu đã được sắp xếp theo thứ tự sau:

11 12 13 16 16 17 18 21 22

Vị trí tại điểm Q_1 là: 1.(9+1)/4 = 2,5 ; từ đó dễ dàng nhận thấy Q_1 = 12,5.

Vị trí tại điểm Q_3 là: 3.(9+1)/4 = 7,5 ; như vậy Q_3 = 19,5

Ngoài ra còn có thể tính ngũ phân vị, thập phân vị, bách phân vị tuỳ theo mục đích nghiên cứu khác nhau.

5. Các mức độ đo độ biến thiên của tiêu thức nghiên cứu

5.1. Ý nghĩa nghiên cứu độ biến thiên của tiêu thức

Số bình quân chỉ nêu lên mức độ đại biểu có tính chất chung nhất của toàn bộ tổng thể nghiên cứu. Mức độ này không phản ánh chênh lệch thực tế giữa các mức độ cá biệt và do đó không chú ý tới từng đơn vị tổng thể. Có khi bản thân nội bộ hiện tượng đã có nhiều thay đổi đáng kể về mặt lượng, nhưng số bình quân tính ra có thể không thay đổi hoặc thay đổi rất ít. Vì vậy, trong phân tích thống kê không nên chỉ hạn chế trong việc nghiên cứu hiện tượng qua các mức độ bình quân, mà cần chú ý quan sát, đánh giá độ biến thiên của tiêu thức.

Việc nghiên cứu độ biến thiên của tiêu thức có nhiều tác dụng quan trọng về mặt lý luận cũng như đối với thực tiễn công tác:

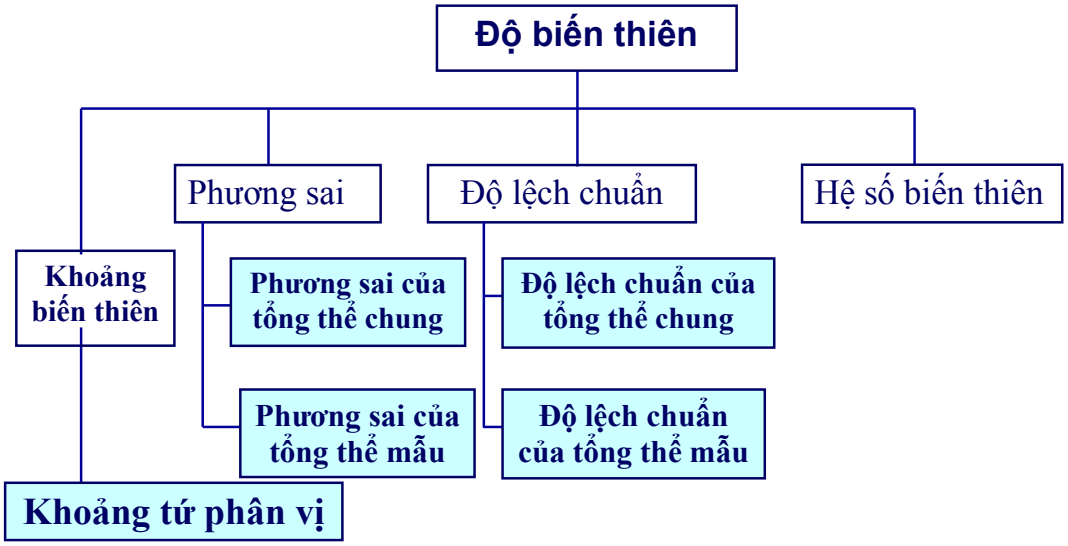
- Các chỉ tiêu đánh giá độ biến thiên của tiêu thức giúp ta xét trình độ đại biểu của số bình quân. Trị số của chỉ tiêu tính ra càng lớn, độ biến thiên của tiêu thức càng nhiều, do đó trình độ đại biểu của số bình quân càng thấp, và ngược lại.

- Quan sát độ biến thiên của tiêu thức trong một dãy số lượng biến, ta sẽ thấy rõ được nhiều đặc trưng của dãy số, như đặc trưng về phân phối, về kết cấu, tính chất đồng đều của tổng thể nghiên cứu.

- Trong phân tích hoàn thành kế hoạch, các chỉ tiêu đánh giá độ biến thiên của tiêu thức giúp ta thấy rõ được chất lượng công tác và nhịp điệu hoàn thành kế hoạch chung cũng như của từng bộ phận, phát hiện khả năng tiềm tàng của các đơn vị.

- Các chỉ tiêu đánh giá độ biến thiên của tiêu thức còn được sử dụng trong nhiều trường hợp nghiên cứu thống kê khác, như: phân tích biến động, phân tích mối liên hệ, dự đoán thống kê...

5.2. Các mức độ đo độ biến thiên của tiêu thức



5.2.1. Khoảng biến thiên

Khoảng biến thiên là độ lệch giữa lượng biến lớn nhất và lượng biến nhỏ nhất của tiêu thức nghiên cứu, biểu hiện bằng công thức:

$$R = x_{max} - x_{min} \qquad (3.16)$$

Trong đó: R - khoảng biến thiên

x_{max}, x_{min} - lượng biến lớn nhất và lượng biến nhỏ nhất của tiêu thức nghiên cứu

Thí dụ: mức năng suất lao động của công nhân hai tổ sản xuất như sau:

Tổ 1: 40, 50, 60, 70, 80 kg

Tổ 2: 58, 59, 60, 61, 62 kg

Mức năng suất lao động bình quân của mỗi tổ đều là 60 kg nhưng thực ra 2 tổ công nhân này không đồng đều về chất lượng, vì năng suất lao động thực tế trong nội bộ tổ 1 chênh lệch nhau rất nhiều so với tổ 2. Để đánh giá trình độ biến thiên của năng suất lao động và qua đó đánh giá tính chất đại biểu của số bình quân, ta hãy tính khoảng biến thiên của 2 tổ:

$$R_1 = 80 - 40 = 40 \text{ kg}$$

$$R_2 = 62 - 58 = 4 \text{ kg}$$

Kết quả cho thấy: R_1 lơn hơn R_2, có nghĩa là độ biến thiên của tiêu thức trong tổ 1 lớn hơn và vì thế tính chất đại biểu của số bình quân tổ 1 thấp hơn.

Khoảng biến thiên là chỉ tiêu đơn giản nhất để đánh giá độ biến thiên của tiêu thức. Chỉ tiêu này nêu lên một cách khái quát nhất độ biến thiên của tiêu thức, khoảng biến thiên càng nhỏ thì tổng thể càng đồng đều, số bình quân càng có tính chất đại biểu cao, và ngược lại.

Nhược điểm của khoảng biến thiên là chỉ phụ thuộc vào 2 lượng biến lớn nhất và nhỏ nhất trong dãy số, mà không xét đến các lượng biến khác, do đó việc nhận định có khi chưa thật hoàn toàn chính xác.

Để khắc phục hạn chế đó có thể sử dụng ***khoảng tứ phân vị (IQR):*** Là chênh lệch giữa tứ phân vị thứ ba và tứ phân vị thứ nhất, nó chỉ tính với 50% các giá trị ở giữa và không chịu ảnh hưởng của lượng biến đột xuất.

$$IQR = Q_3 - Q_1 \qquad (3.17)$$

Khoảng tứ phân vị phản ánh mức độ chênh lệch về lượng biến tiêu thức nghiên cứu của 50% số đơn vị ở giữa của tổng thể đang nghiên cứu.

5.2.2. *Độ lệch tuyệt đối bình quân*:

Độ lệch tuyệt đối bình quân là số bình quân cộng của các độ lệch tuyệt đối giữa các lượng biến với số bình quân cộng của các lượng biến đó. Công thức như sau:

$$\bar{d} = \frac{\sum |x_i - \bar{x}|}{n} \qquad (3.18)$$

$$\bar{d} = \frac{\sum |x_i - \bar{x}| f_i}{\sum f_i} \text{ (trường hợp có quyền số)} \qquad (3.19)$$

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

$\bar{x}$ - số bình quân cộng của các lượng biến x_i

Trị số của độ lệch tuyệt đối bình quân tính ra càng nhỏ thì tiêu thức càng ít biến thiên, tính chất đại biểu của số bình quân càng cao, và ngược lại.

Vẫn lấy thí dụ về năng suất lao động của 2 tổ công nhân, có thể lập bảng tính sau:

Bảng 8

Tổ 1				***Tổ 2***			
x_i	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	x_i	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
40	20	-20	40	58	2	-2	4
50	10	-10	100	59	1	-1	1
60	0	0	0	60	0	0	0
70	10	+10	100	61	1	+1	1

80	20	+20	400	62	2	+2	4
-	60	0	1000	-	6	0	10

từ tài liệu trên, tính ra:

$$\bar{d}_1 = \frac{60}{5} = 12 \text{ kg}$$

$$\bar{d}_2 = \frac{6}{5} = 1{,}2 \text{ kg}$$

Như vậy, các tiêu thức của tổ 1 biến thiên nhiều hơn tổ 2, tính chất đại biểu của số bình quân tổ 1 do đó cũng thấp hơn.

Độ lệch tuyệt đối bình quân có thể phản ánh độ biến thiên của tiêu thức một cách chặt chẽ hơn khoảng biến thiên, vì nó có xét đến tất cả mọi lượng biến trong dãy số. Nhưng khi tính toán chỉ itêu này, ta chỉ xét các trị số tuyệt đối của độ lệch, tức là bỏ qua sự khác nhau thực tế về dấu âm, dương của các độ lệch, cũng vì thế mà việc phân tích bằng các phương pháp toán học gặp nhiều khó khăn. Chỉ tiêu này thường được dùng trong phân tích chất lượng sản phẩm, như xét trình độ đồng đều của sợi dệt trong các nhà máy dệt.

5.2.3. Phương sai

Là một chỉ tiêu rất quan trọng cho biết độ biến thiên xung quanh số trung bình của các lượng biến. Phương sai là số bình quân cộng của bình phương các độ lệch giữa các lượng biến với số bình quân cộng của các lượng biến đó.

Công thức tính:

- Đối với tổng thể chung

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \qquad (3.20)$$

Trong đó: σ^2 là phương sai của tổng thể chung;
μ là trung bình của tổng thể chung;
N là số đơn vị của tổng thể chung

- Đối với tổng thể mẫu

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \qquad (3.21)$$

Trong đó: S^2 là phương sai của tổng thể mẫu;
$\bar{x}$ là trung bình của tổng thể mẫu;
n là số đơn vị của tổng thể mẫu

Cần chú ý: Khi tính phương saicủa tổng thể chung thì mẫu số là N; còn phương sai mẫu thì mẫu số là n-1.

- Trường hợp có quyền số (tính từ dãy số phân phối)

$$S^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i - 1} \qquad (3.22)$$

Phương sai là chỉ tiêu thường dùng để đánh giá độ biến thiên của tiêu thức, khắc phục được những khác nhau về dấu giữa các độ lệch. Phương sai có trị số càng nhỏ thì tổng thể nghiên cứu càng đồng đều, tính chất đại biểu của số bình quân càng cao, và ngược lại. Tuy vậy phương sai có hạn chế là vì bình phương sai số nên khuếch đại trị số và không có đơn vị tính thích hợp.

5.2.4. Độ lệch tiêu chuẩn

Độ lệch tiêu chuẩn là chỉ tiêu quan trọng nhất, bằng căn bậc hai của phương sai (tức là số bình quân toàn phương của bình phương các độ lệch giữa các lượng biến với số bình quân cộng của các lượng biến đó).

Công thức như sau:

- Đối với tổng thể chung

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \qquad (3.23)$$

- Đối với tổng thể mẫu

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \qquad (3.24)$$

Trong đó: σ - độ lệch tiêu chuẩn của tổng thể chung và S là độ tiêu chuẩn của mẫu.

Vẫn lấy thí dụ trên tính ra:

$$\sigma_1 = \sqrt{200} = 14{,}14 \text{ kg}$$

$$\sigma_2 = \sqrt{2} = 1{,}414 \text{ kg}$$

Qua việc so sánh giữa hai độ lệch tiêu chuẩn, các kết luận rút ra cũng giống như các chỉ tiêu trước đã nêu lên.

Độ lệch tiêu chuẩn là chỉ tiêu hoàn thiện nhất và thường dùng nhất trong nghiên cứu thống kê để đánh giá độ biến thiên của tiêu thức. Tuy nhiên, việc tính toán độ lệch tiêu chuẩn phức tạp hơn.

Các chỉ tiêu trên chỉ dùng để so sánh độ biến thiên của các hiện tượng cùng loại và có số trung bình bằng nhau nhưng không được dùng để so sánh biến thiên của các hiện tượng khác loại hoặc các hiện tượng cùng loại nhưng số trung bình không bằng nhau. Khi đó người ta dùng hệ số biến thiên.

5.2.5. Hệ số biến thiên:

Hệ số biến thiên là chỉ tiêu tương đối (%) có được từ sự so sánh giữa độ lệch tiêu chuẩn với số bình quân cộng. Các công thức như sau:

$$CV = \frac{S}{\bar{x}} \times 100\% \qquad (3.25)$$

Trong đó: V - hệ số biến thiên

S - độ lệch tiêu chuẩn

$\bar{x}$ - số bình quân cộng

Hệ số biến thiên được biểu hiện bằng số tương đối, nên có thể dùng để so sánh giữa các tiêu thức khác nhau, như so sánh hệ số biến thiên về năng suất lao động với hệ số biến thiên về tiền lương, hệ số biến thiên của tiền lương với hệ số biến thiên của tỷ lệ hoàn thành định mức sản xuất... Trong khi đó, các chỉ tiêu khác như: khoảng biến thiên, độ lệch tuyệt đối bình quân, độ lệch tiêu chuẩn có đơn vị tính toán giống như đơn vị tính toán của tiêu thức nghiên cứu, nên không thể dùng để so sánh giữa các tiêu thức khác nhau. Ta hãy so sánh độ biến thiên của hai tiêu thức trong thí dụ sau:

Bảng 9

Tiêu thức nghiên cứu	*Số bình quân ($\bar{x}$)*	*Độ lệch tiêu chuẩn (s)*	*Hệ số biến thiên (V%)*
Tiền lương (1000đ)	300	30	10
Năng suất lao động (SP)	110	18	16,4

Từ sự so sánh hai hệ số biến thiên, ta thấy tỷ lệ hoàn thành định mức sản xuất biến thiên nhiều hơn so với tiền lương.

6. Các chỉ tiêu và biểu đồ thể hiện hình dáng của phân phối

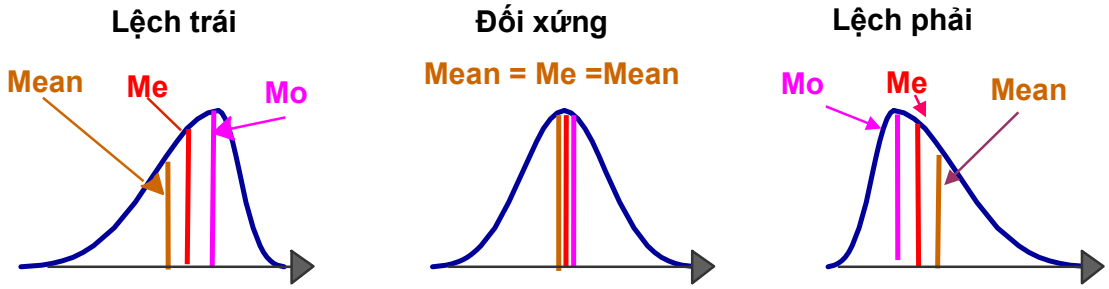
Các chỉ tiêu biểu thị hình dáng của phân phối chủ yếu sử dụng đối với phân phối chuẩn.

6.1. Các chỉ tiêu biểu hiện sự không đối xứng của phân phối

Để biểu hiện sự không đối xứng của phân phối có thể dùng hai cách phổ biến sau:

- So sánh trung bình, Mốt và trung vị, cụ thể:

- Nếu: $\bar{x} = M_0 = M_e$ dãy số có phân phối chuẩn đối xứng
- Nếu: $\bar{x} > M_e = M_0$ dãy số có phân phối chuẩn lệch phải
- Nếu: $M_0 > M_e > \bar{x}$ dãy số có phân phối chuẩn lệch trái



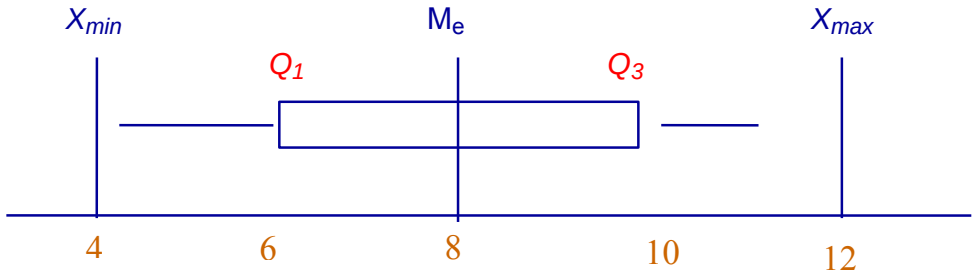
- Hệ số không đối xứng:

$$K_A = \frac{\bar{x} - M_0}{\sigma}$$

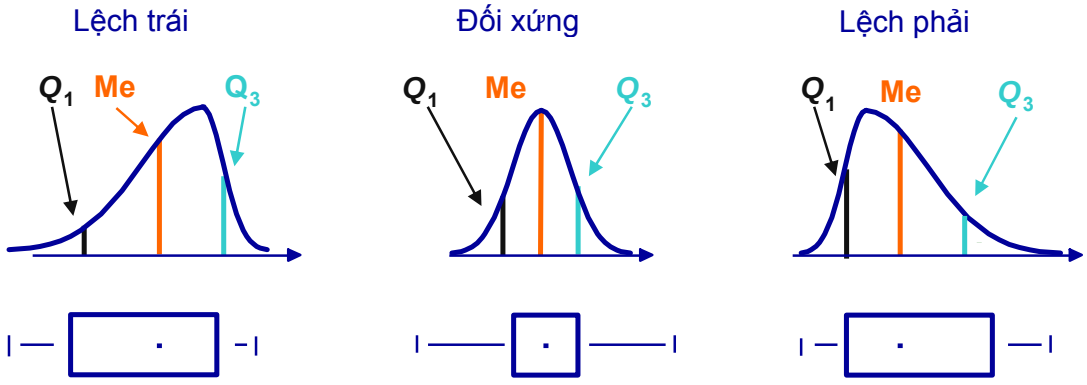
$K_A > 0$ dãy số có phân phối chuẩn lệch phải
$K_A < 0$ dãy số có phân phối chuẩn lệch trái
K_A càng lớn dãy số có phân phối càng không đối xứng
$K_A = 0$ dãy số có phân phối chuẩn đối xứng

6.2. Biểu đồ hộp ria mèo (Box plot)

- Biểu đồ hộp ria mèo:



- Mối quan hệ giữa hộp ria mèo và các chỉ tiêu phản ánh độ đối xứng:



- Tác dụng của hộp ria mèo (*box and whiskers)*:

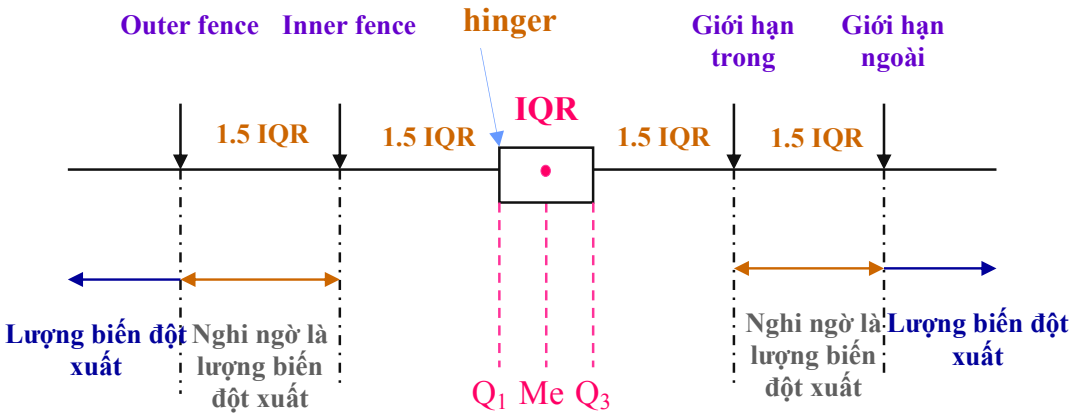
+ Nhận biết vị trí của bộ dữ liệu trên cơ sở Me

+ Nhận biết sự dàn trải của dữ liệu trên cơ sở độ dài của hộp (khoảng tứ + Phân vị IQR) và độ dài của hộp ria mèo

+ Nhận biết độ lệch phân phối của dữ liệu

+ So sánh 2 hay nhiều bộ dữ liệu với cùng 1 thước đo

+ Nhận biết lượng biến đột xuất *(outliers)* và nghi ngờ có thể là lượng biến đột xuất :



Khi có lượng biến đột xuất, trước hết nên loại lượng biến này ra khỏi tập hợp dữ liệu để làm các thống kê suy luận sau đó phân tích nó như 1 trường hợp đặc biệt trong mối quan hệ với tính quy luật chung của cả tập hợp dữ liệu.

BÀI TẬP

3.1 Dưới đây là số liệu về độ tuổi của các bệnh nhân đến khám ở bệnh viện A vào ngày 20/9/2007

32	45	53	60	79	73
73	53	61	48	51	49
62	72	37	70	38	66
52	33	78	45	65	47
64	47	61	75	57	64

Yêu cầu:

a. Xây dựng biểu đồ thân lá và bảng tần số phân bố với các tổ 30 – 40, 40 – 50, 50 – 60, 60 – 70, 70 – 80.

b. Tính tuổi trung bình dựa trên các số liệu ban đầu

c. Tính tuổi trung bình dựa trên bảng tần số phân bố

d. So sánh kết quả tính của câu b và câu c rồi đưa ra nhận xét

3.2 Một nhà nghiên cứu xã hội học đó nghiên cứu tình trạng tội phạm ở một địa phương. Ông TA thu thập được tài liệu và tính được tài liệu sau:

Năm	2000 so 1999	2001 so 2000	2002 so 2001	2003 so 2002	2004 so 2003	2005 so 2004
Tỷ lệ %	96	105	110	103	106	95

Yêu cầu:

a. Tính tốc độ phát triển bình quân về số lượng tội phạm trong các năm 2001 - 2004

b. Tính tốc độ phát triển trung bình về số lượng tội phạm trong các năm 2000 -2005

3.3 Dữ liệu sau là số lượng hành khách trên các chuyến bay của hãng Delta Airlines giữa San Francisco và Seattle trong 33 ngày tháng 4 và đầu tháng 5:

128 121 134 136 136 118 123 109 120 116 125 128 121 129 130 131 127
119 114 134 110 136 134 125 128 123 128 133 132 136 134 129 132

Tìm trung bình, mốt, trung vị, tứ phân vị, khoảng tứ phân vị của dữ liệu trên và giải thích ý nghĩa của các kết quả tính được.

3.4 Dữ liệu sau là giá cả bằng đồng Francs Pháp của một lọ nước hoa Channel No.5 14ml tại cửa hàng miễn thuế ở các sân bay khác nhau trên thế giới. Tìm số trung bình, trung vị, và lượng biến đột xuất *(outliers)*.

Abu Dhabi: 399
Dubai: 570
Bangkok: 616
Seoul: 642

Hong Kong: 616
Singapore: 940
New York: 515
Amsterdam: 540
Frankfurt: 554
Zurich: 562
Paris: 560
Copenhagen: 548
London: 627
Rome: 612

3.5 Giải thích tại sao chúng ta cần các thước đo độ biến thiên và những thông tin mà các thước đo này chứa đựng. Thước đo độ biến thiên nào là quan trọng nhất? Tại sao?

3.6 Sự khác nhau trong tính toán giữa phương sai của mẫu và phương sai của tổng thể?

3.7 Tìm khoảng biến thiên, phương sai và độ lệch chuẩn của bộ dữ liệu trong bài 3.3 (giả định đây là mẫu).

3.8 Các quốc gia có số người tới thăm nước Mỹ nhiều nhất trong năm 2007 như sau (*đơn vị tính: triệu người*):

Canada: 15.9
Mexico: 8.9
Nhật Bản: 5.5
Anh: 3.3
Đức: 2.1
Pháp: 1.1
Brazil: 1.0
Hàn Quốc: 1.0
Italy: 0.6
Australia: 0.5

Tìm số trung bình, trung vị và độ lệch chuẩn. Hãy vẽ bar graph.

3.9 Sự tham gia của công nhân vào hoạt động quản lý là một chương trình mới, thu hút người lao động vào việc ra các quyết định của công ty. Dữ liệu sau là số phần trăm người lao động được thu hút vào chương trình này trong một mẫu các doanh nghiệp. Hãy vẽ một biểu đồ hộp ria mèo *(box plot)* của dữ liệuvà rút ra kết luận về bộ dữ liệu được căn cứ vào *box plot*.

5 32 33 35 42 43 42 45 46 44 47 48 48 48 49 49 50 37 38 34 51 52 52 47 53 55 56 57 58 63 78

3.10 Tham khảo dữ liệu sau về khoảng cách giữa các chỗ ngồi trong khoang hạng nhất (business class) của các hãng hàng không khác nhau. Hãy tìm μ, σ, σ^2, vẽ *box plot,* tìm mode và các *outliers*.

Đặc điểm của khoang hạng nhất

	Khoảng cách giữa các hàng (cm)
EUROPE	
Air France	122
Alitalia	140
British Airways	127
Iberia	107
KLM/Northwest	120
Lufthansa	101
Sabena	122
SAS	132
SwissAir	120
ASIA	
All Nippon Airw	127
Cathay Pacific	127
JAL	127
Korean Air	127
Malaysia Air	116
Singapore Airl	120
Thai Airways	128
Vietnam Airl	140
NORTH AMERICA	
Air Canada	140
American Airl	127
Continental	140
Delta Airlines	130
TWA	157
United	124

3.11 Xác định giới hạn trong (*inner fences)* và giới hạn ngoài (*outer fences)* của một *box plot*; xác định *whiskers* và *hingers*. Phần nào của dữ liệu được trình bày bởi *box*, bởi *whiskers*? Thảo luận về cách thức xử lý các *outliers* - cách nhận biết chúng và làm gì với chúng khi đã nhận biết được. Bạn có thể luôn luôn bỏ một *outlier* hay không? Tại sao?

3.12 Dữ liệu sau là số ounce bạc trong mỗi tấn quặng ở hai mỏ:

Mỏ A: 34, 32, 35, 37, 41, 42, 43, 45, 46, 45, 48, 49, 51, 52, 53, 60, 73, 76, 85

Mỏ B: 23, 24, 28, 29, 32, 34, 35, 37, 38, 40, 43, 44, 47, 48, 49, 50, 51, 52, 59

Vẽ một *box plot* cho mỗi bộ dữ liệu. So sánh hai hai *box plots*. Rút ra kết luận về dữ liệu đó.

3.13 Có tài liệu về năng suất lao động của công nhân một xí nghiệp trong tháng 12 năm 2005 như sau:

NSLĐ (kg)	*Số công nhân*
50 – 54	10
54 – 58	40
58 – 62	80
62 – 66	50
66 – 70	20

Hãy tính:

a. Năng suất lao động trung bình của công nhân trong xí nghiệp
b. Mốt về năng suất lao động
c. Số trung vị về năng suất lao động

Nhận xét về phân phối của công nhân theo NSLĐ

3.14 Có tài liệu về năng suất lao động của công nhân và giá thành đơn vị sản phẩm tại 3 doanh nghiệp thuộc cùng một ngành sản xuất trong tháng 10 năm 2007 như sau:

Doanh nghiệp	*Số lượng lao động*	*NSLĐ bình quân (sản phẩm)*	*Giá thành bình quân 1 sản phẩm (tr.đ)*
Số 1	200	250	20,0
Số 2	300	260	19,5
Số 3	500	280	19,0

Hãy tính:

a. Năng suất lao động trung bình chung cho cả 3 doanh nghiệp

b. Giá thành trung bình một đơn vị sản phẩm tính chung cho cả 3 doanh nghiệp

3.15 Có tài liệu sau đây về một xí nghiệp trong năm 2005

Chỉ tiêu	*Số trung bình* $\bar{x}$	*Độ lệch chuẩn (s)*
Năng suất lao động (kg)	400	60
Giá thành đơn vị sản phẩm (1000đ)	3,8	0,19

Hãy xác định xem trong hai chỉ tiêu nói trên, chỉ tiêu nào có độ phân tán mạnh hơn.

CHƯƠNG 4:

ƯỚC LƯỢNG

1. Một số khái niệm thường dùng trong ước lượng

- Tổng thể chung: là tổng thể bao gồm toàn bộ các đơn vị thuộc đối tượng điều tra. Số đơn vị của tổng thể chung được kí hiệu là N.
- Tổng thể mẫu: là tổng thể bao gồm một số đơn vị nhất định được chọn ra từ tổng thể chung để điều tra thực tế. Số đơn vị tổng thể mẫu được kí hiệu là n.
- Chọn một lần và chọn nhiều lần: việc chọn các đơn vị từ tổng thể chung vào mẫu được thực hiện theo 2 cách
    - *Chọn một lần*: là khi mỗi đơn vị đã được chọn để đăng kí rồi sẽ được xếp riêng ra, không được trả về tổng thể chung, do đó sẽ không có khả năng được chọn lại
    - *Chọn nhiều lần*: là sau khi mỗi đơn vị được chọn ra đăng kí vào mẫu rồi lại được trả về tổng thể chung. Như vậy số đơn vị tổng thể chung không thay đổi trong suốt quá trình chọn mẫu.
- Các tham số của tổng thể chung:
    - Số trung bình

      Công thức: $\mu = \frac{\sum x_i}{N}$

    - Phương sai

      Công thức: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$

    - Tỉ lệ theo một tiêu thức nào đó của tổng thể chung

      Công thức: $p = \frac{N_c}{N}$

- Các tham số của tổng thể mẫu
    - Số trung bình

      Công thức: $\overline{x} = \frac{\sum x_i}{n}$

- Phương sai

Công thức: $S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

- Tỉ lệ theo một tiêu thức nào đó của tổng thể chung

Công thức: $p_s = \frac{n_c}{n}$

- Khái niệm ước lượng: từ sự hiểu biết về các tham số nào đó của mẫu đã điều tra để tính toán thành các tham số tương ứng của tổng thể chung (hay còn gọi là ước lượng các tham số của tổng thể chung từ tham số của tổng thể mẫu). Các tham số của mẫu sẽ được tính toán sau khi tiến hành thu thập tài liệu từ mẫu được chọn ra từ tổng thể chung.
- Ước lượng điểm (Point estimation):
  - Sử dụng số trung bình mẫu để ước lượng cho số bình quân của tổng thể chung.
  - Sử dụng phương sai mẫu để ước lượng cho phương sai của tổng thể chung.
  - Sử dụng tỉ lệ theo một tiêu thức nào đó của tổng thể mẫu để ước lượng cho tỉ lệ của tổng thể chung.
- Ước lượng khoảng (Interval estimation): Là xác định một khoảng giá trị mà tham số của tổng thẻ chung rơi vào đó với xác suất nhất định (không bao giờ chắc chăn bằng 100%).
- Sai số chọn mẫu: do sử dụng phương pháp ước lượng để tính toán tham số cho tổng thể chung nên luôn tồn tại sai số giữa các giá trị của các tham số của mẫu và giá trị thực của các tham số của tổng thể chung.
- Phạm vi sai số chọn mẫu: Là chênh lệch giữa các tham số của tổng thể chung và tổng thể mẫu.

2. Khoảng tin cậy của số trung bình tổng thể chung:

- *Trường hợp đã biết phương sai:*

Giả thiết:

- Đã biết phương sai của tổng thể chung,
- tổng thể chung phân phối chuẩn;
- Nếu không có phân phối chuẩn thì mẫu phải lớn.

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

- *Trường hợp chưa biết phương sai*

Giả thiết:

- Chưa biết phương sai của tổng thể chung,
- Tổng thể chung phải phân phối chuẩn;
- Sử dụng phân vị Student với khoảng tin cậy như sau:

$$\bar{x} - t_{\alpha/2;(n-1)} \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2;(n-1)} \frac{s}{\sqrt{n}}$$

- *Trường hợp tổng thể chung có giới hạn*

  Khi tổng thể chung có giới hạn, mẫu lớn (n/N > .05):

$$\bar{x} - t_{\alpha/2;(n-1)} \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \le \mu \le \bar{x} + t_{\alpha/2;(n-1)} \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

- Các nhân tố ảnh hưởng độ lớn của khoảng tin cậy.
  - Mức ý nghĩa (hay là xác suất của ước lượng).
  - Phương sai của tổng thể chung. Tổng thể chung càng đồng đều khoảng ước lượng càng nhỏ.
  - Độ lớn của mẫu. Mẫu càng lớn khoảng ước lượng càng nhỏ.

3. Khoảng tin cậy của tỷ lệ tổng thể chung:

Giả thiết:

- Theo tiêu thức nghiên cứu, tổng thể có hai loại biểu hiện.
- Tổng thể chung có phân phối nhị thức,
- Phân phối xấp xỉ chuẩn được sử dụng.
- Với mẫu đủ lớn ($n.p \ge 5$ và $n.(1-p) \ge 5$) khoảng tin cậy như sau:

$$p_s - Z_{\alpha/2} \cdot \sqrt{\frac{p_s.(1-p_s)}{n}} \le p \le p_s + Z_{\alpha/2} \cdot \sqrt{\frac{p_s.(1-p_s)}{n}}$$

4. Xác định cỡ mẫu:

- *Yêu cầu của việc xác định cỡ mẫu:*
  - Đảm bảo sai số chọn mẫu là nhỏ nhất;
  - Đảm bảo chi phí là thấp nhất.

  Để dung hoà hai yêu cầu đối lập đó người ta thường căn cứ vào độ chính xác trong ước lượng .

- *Công thức tính cỡ mẫu:*

  - Ước lượng số trung bình:

$$n = \frac{Z^2\sigma^2}{Error^2}$$

Trường hợp với tổng thể chung có giới hạn:

$$n = \frac{n_0 N}{n_0 + (N-1)} \qquad \text{trong đó :} \quad n_0 = \frac{Z^2\sigma^2}{Error^2}$$

- Ước lượng tỉ lệ :

$$n = \frac{Z^2 p(1-p)}{Error^2}$$

Chú ý: Khi xác định cỡ mẫu thường không có phương sai của tổng thể chung, vì vậy có thể giải quyết bằng một trong các cách sau:

- Lấy phương sai lớn nhất trong các lần điều tra trước (nếu có)
- Lấy phương sai của các hiện tượng khác tương tự (nếu có)
- Điều tra thí điểm để tính phương sai
  - Có thể ước lượng độ lệch tiêu chuẩn qua khoảng biến thiên tuỳ theo phân phối của tổng thể. Cụ thể: nếu tổng thể có phân phối chuẩn thì:

$$\sigma = \frac{R}{6} = \frac{x_{max} - x_{min}}{6}$$

- *Các nhân tố ảnh hưởng đến cỡ mẫu:*
  - Phạm vi sai số chọn mẫu cho phép.
  - Độ đồng đều của tổng thể.
  - Yêu cầu về độ tin cậy.

) Trên thực tế việc xác định cỡ mẫu có thể dựa trên các yếu tố sau:

- Dựa trên cảm tính (từ 1% - 5% của tổng thể chung)
- Dựa trên kinh nghiệm
- Dựa trên chi phí
- Dựa trên các phương pháp thống kê (như trên)

BÀI TẬP

4.1 Một đại lý nhà đất cần ước lượng giá trị trung bình của một khu đất với diện tích cho trước ở một vùng nhất định. Đại lý đó tin rằng độ lệch chuẩn của các giá trị của khu đất là σ = \$5500 và các giá trị của khu đất đó xấp xỉ phân phối chuẩn. Một mẫu ngẫu nhiên gồm 16 đơn vị cho trung bình mẫu là \$89673.12. Đưa ra khoảng tin cậy 95% cho giá trị trung bình của tất cả các khu đất loại này.

4.2 Với tình huống trong bài 4.1, giả sử rằng cần tìm khoảng tin cậy 99%. Hãy tính khoảng tin cậy mới và so sánh với khoảng tin cậy 95% trong bài 4.1.

4.3 Một công ty đang xem xét việc lắp đặt một máy fax trong một trong những văn phòng của họ. Trong quá trình ra quyết đinh liệu có nên lắp đặt chiếc máy này hay không, giám đốc công ty muốn ước lượng số tài liệu bình quân cần phải chuyển trong mỗi ngày nếu cái máy đã được lắp đặt. Theo kinh nghiệm của một số văn phòng khác, giám đốc công ty tin rằng độ lệch chuẩn của số tài liệu cần gửi là 32. Người giám đốc cũng tin rằng số lượng tài liệu cần chuyển trong mỗi ngày là một biến ngẫu nhiên phân phối chuẩn. Chiếc máy được kiểm tra qua một mẫu gồm 15 ngày, và kết quả bình quân mẫu là 267. Hãy cho biết khoảng tin cậy 99% cho số lượng tài liệu bình quân được chuyển nếu chiếc máy được lắp đặt.

4.4 Nếu trong tình huống bài tập 4.3, người giám đốc muốn lắp đặt chiếc máy nếu ông ấy có thể tin chắc rằng số tài liệu được chuyển bình quân ngày lớn hơn 245. Liệu kết quả tìm thấy trong bài 4.3 có giải thích việc lắp đặt chiếc máy hay không? Giải thích?

4.5 Một công ty điện thoại muốn ước lượng độ dài bình quân của các cuộc gọi đường dài vào những ngày nghỉ cuối tuần. Một mẫu ngẫu nhiên gồm 50 cuộc gọi cho biết số trung bình $\overline{X}$ = 14.5 phút và độ lệch chuẩn s - 5.6 phút. Hãy cho biết khoảng tin cậy 95% và khoảng tin cậy 90% cho độ dài bình quân của các cuộc gọi đường dài trong những ngày nghỉ cuối tuần.

4.6 Một nhà sản xuất lốp xe muốn ước lượng số dặm đường bình quân một chiếc xe của có thể thực hiện trước khi nó bị hỏng. Một mẫu ngẫu nhiên gồm 32 lốp xe được lựa chọn, các lốp xe được vận hành cho đến khi nó bị hỏng, số dặm đường mỗi lốp xe đi được được ghi lại. Dữ liệu tính theo đơn vị ngàn dặm, như sau:

32, 33, 28, 37, 29, 30, 25, 27, 39, 40, 26, 26, 27, 30, 25, 30, 31, 29, 24, 36, 25, 37, 37, 20, 22, 35, 23, 28, 30, 36, 40, 41.

Hãy đưa ra khoảng tin cậy 99% cho số dặm bình quân có thể thực hiện được với loại lốp xe này.

4.7 Một nhà sản xuất kem dưỡng da muốn xác định tỉ lệ phần trăm số người ở một độ tuổi nhất định có thể có phản ứng tốt với loại kem này. Một mẫu ngẫu nhiên gồm 68 người cho

thấy 42 người đạt kết quả tốt. Hãy đưa ra khoảng tin cậy 99% về tỉ lệ phần trăm số người ở độ tuổi đó có thể được điều trị thành công với loại kem dưỡng da này.

4.8 Một bài báo gần đây mô tả sự thành công của các trường kinh doanh ở Châu Âu và nhu cầu của châu lục này đối với chương trình MBA. Bài báo cho thấy rằng một cuộc điều tra 280 chỗ trong các trường kinh doanh cho kết quả rằng chỉ có 1/7 số chỗ cho chương trình MBA đã có người học. Giả sử rằng các số liệu này là chính xác và mẫu được lựa chọn ngẫu nhiên từ tổng thể, cho biết khoảng tin cậy 90% cho tỉ lệ số chỗ của chương trình MBA có người học ở Châu Âu.

4.9 Một doanh nghiệp dệt có 1000 công nhân , người ta chọn ngẫu nhiên (theo cách chọn không lặp) 100 công nhân để diều tra về năng suất lao động và có kết quả sau đây:

NSLĐ (mét)	***Số công nhân***
Dưới 40	30
40 - 50	33
50 - 60	24
Từ 60 trở lên	13

Hãy tính :

a . Năng suất lao động bình quân chung của công nhân toàn doanh nghiệp với xác suất bằng 0,6826 . Từ đó xác định sản lượng vải của doanh nghiệp .

b . Xác suất để phạm vi sai số chọn mẫu khi suy rộng về năng suất lao động bình quân không vượt quá 1,926 m .

c. Giả sử doanh nghiệp tiến hành một cuộc điều tra chọn mẫu mới để suy rộng về năng suất lao động bình quân . Với xác suất bằng 0,9544 và phạm vi sai số chọn mẫu không vượt quá 2 m , hãy tính số công nhân cần chọn để điều tra theo cách chọn lặp và chọn không lặp.

d. Tỷ lệ chung về số công nhân có năng suất lao động từ 60 m trở lên với xác suất bằng 0,6826 . Có khoảng bao nhiêu công nhân của doanh nghiệp đạt mức năng suất lao động này ? .

e . Xác suất để phạm vi sai số chọn mẫu khi suy rộng về tỷ lệ số công nhân có năng suất lao động từ 60 m trở lên không vượt quá 0, 096 .

h. Giả sử doanh nghiệp tiến hành một cuộc điều tra chọn mẫu mới để suy rộng về tỷ lệ công nhân đạt năng suất lao động từ 60 m trở lên . Với xác suất bằng 0,9544 và phạm vi sai số chọn mẫu không vượt quá 5% , hãy tính số công nhân cần chọn ra để diều tra theo cách chọn lặp và chọn không lặp.

4.10 Một thành phố có 500.000 nhân khẩu , người ta chọn ngẫu nhiên 5% tổng số nhân của thành phố theo cách chọn không lặp để điều tra và thu được kết quả sau đây :

+ Nhân khẩu từ 14 tuổi trở xuống có 10000 người .
+ Nhân khẩu từ 60 tuổi trở lên có 2000 người .
+ Nhân khẩu từ 15 tuổi trở lên :

- Có hoạt động khinh tế có 11700 người .
- Không hoạt động kinh tế có 3300 người .

Với xác suất bằng 0,9544 , hãy xác định tỷ lệ và số lượng mỗi loại nhân khẩu ở trên của toàn thành phố

CHƯƠNG 5

KIỂM ĐỊNH

Ở chương 4, chúng ta đã nghiên cứu về điều tra chọn mẫu với mục đích thường là suy rộng trung bình, tỷ lệ theo một tiêu thức nào đó của tổng thể mẫu thành tham số tương ứng của tổng thể chung. Chương tiếp theo sẽ nói về cách sử dụng các thống kê của mẫu để kiểm định giả thiết về tổng thể chung, đó là một vấn đề quan trọng của thống kê. Kiểm định giả thiết bắt đầu từ giả thiết về một tham số của tổng thể chung, sau đó tiến hành chọn mẫu, tính toán các chỉ tiêu mẫu và sử dụng thông tin để xác định xem giả thiết về tham số của tổng thể chung có đúng hay không.

Chẳng hạn, khi đưa ra giả thiết về số trung bình của tổng thể chung bằng một giá trị nào đó, để kiểm tra lại giả thiết đó ta thu thập các số liệu mẫu và xác định sự chênh lệch giữa giá trị giả thiết và giá trị tính được từ mẫu, sau đó đánh giá xem sự chênh lệch đó là có ý nghĩa hay không. Mức chênh lệch càng nhỏ giả thiết của chúng ta càng có khả năng đúng; mức chênh lệch càng lớn, khả năng đúng càng thấp. Nhưng thường thì mức chênh lệch giữa giá trị giả thiết và giá trị thực tế của mẫu không lớn đến mức ta có thể bác bỏ ngay giả thiết ban đầu và cũng không nhỏ đến mức ta có thể chấp nhận ngay giả thiết đó. Do đó, khi tiến hành kiểm định giả thiết (tiến hành những quyết định có ý nghĩa nhất trong cuộc sống thực tế) thì những giải pháp hoàn toàn rõ ràng là những trường hợp ngoại lệ, không phổ biến.

Một thí dụ như sau: Kết cấu của một tổ hợp nhà thi đấu thể thao ở một thành phố do một Công ty thiết kế các công trình kiến trúc lớn CT đảm nhiệm. Theo kết cấu đó cần khoảng 10.000 tấm nhôm dầy 0,15cm. Các tấm nhôm này không được phép dầy hơn 0,15cm vì kết cấu không chịu được trọng lượng thừa đồng thời chúng cũng không được mỏng hơn 0,15cm vì khi đó mái lợp sẽ không đủ độ vững chắc. Do vậy mà CT tiến hành kiểm tra những tấm nhôm rất cẩn thận. CT không muốn phải kiểm tra từng tấm mà chỉ chọn mẫu 100 tấm. Những tấm nhôm trong mẫu có độ dầy trung bình là 0,153cm. Từ kinh nghiệm làm việc với chính người cung cấp tấm lợp này trước kia, CT biết rằng độ lệch tiêu chuẩn về độ dầy của các tấm lợp là 0,015cm. Trên cơ sở các số liệu đó, CT cần đi đến kết luận là 10.000 tấm lợp có thích hợp với công trình không. Phương pháp kiểm định giả thiết sẽ giúp cho CT quyết định cần từ chối hay chấp nhận lô tấm lợp đó.

1. Một số vấn đề chung về kiểm định

1.1. Giả thiết thống kê.

Giả thiết thống kê là giả thiết về một vấn đề nào đó của tổng thể chung. Đó là các giả thiết về dạng của phân phối xác suất; về các tham số như trung bình, tỷ lệ, phương sai; về

tính độc lập.... Thí dụ như: phương pháp điều trị A chữa khỏi 90% bệnh nhân ; tuổi thọ của hai loại bóng đèn A và B là như nhau ; kết quả của 3 phương pháp là khác nhau hay một tổng thể chung nào đó có phân phối chuẩn....

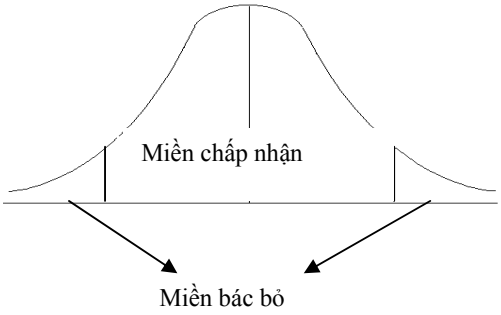
Giả thiết mà ta muốn kiểm định gọi là "giả thiết không" và ký hiệu là H_0. Giả thiết đối lập với nó được gọi là giả thiết đối (hay giả thiết thay thế) và được ký hiệu là H_1. Vấn đề đặt ra là: chúng ta bác bỏ hay chấp nhận một giả thiết bằng cách nào.

Giả thiết thống kê có thể được trình bày dưới nhiều dạng khác nhau. Tuỳ theo dạng của các giả thiết này mà có thể lựa chọn và áp dụng kiểm định hai phía hay kiểm định một phía :

- *Kiểm định 2 phía* là bác bỏ giả thiết H_0 khi tham số đặc trưng của mẫu cao hơn hoặc thấp hơn so với giá trị của giả thiết về tổng thể chung. Kiểm định 2 phía có 2 miền bác bỏ, biểu hiện ở hình 1.1.

Thí dụ: Giả thiết $H_0 : \mu = \mu_0$

Giả thiết $H_1 : \mu \neq \mu_0$

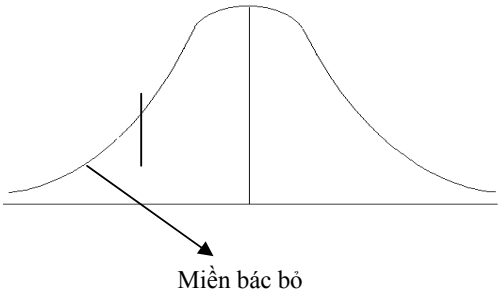


Hình 1.1

- *Kiểm định phía trái* là bác bỏ giả thiết H_0 khi tham số đặc trưng của mẫu nhỏ hơn một cách đáng kể so với giá trị của giả thiết H_0. Miền bác bỏ nằm ở phía trái của đường phân phối, biểu hiện ở hình 1.2

Thí dụ: Giả thiết $H_0 : \mu = \mu_0$

Giả thiết $H_1 : \mu < \mu_0$

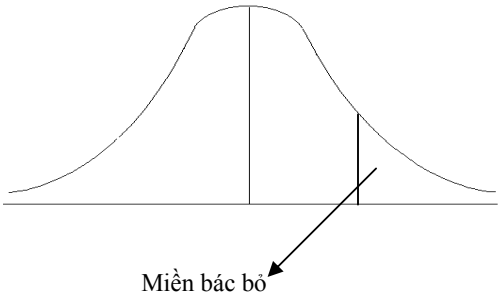


Hình 1.2

- *Kiểm định phía phải* là bác bỏ giả thiết H_0 khi tham số đặc trưng của mẫu lớn hơn một cách đáng kể so với giá trị của giả thiết H_0. Miền bác bỏ nằm ở phía phải của đường phân phối, biểu hiện ở hình 1.3

Thí dụ: Giả thiết $H_0 : \mu = \mu_0$

Giả thiết $H_1 : \mu > \mu_0$



Hình 1.3

1.2. Sai lầm và mức ý nghĩa trong kiểm định.

Trong khi phải lựa chọn giữa hai giả thiết H_0 và H_1 ta có thể mắc phải hai loại sai lầm: *Sai lầm loại 1* là bác bỏ giả thiết H_0 khi nó đúng; ngược lại, thừa nhận H_0 khi nó sai là *sai lầm loại 2*. Một kiểm định thống kê lý tưởng là kiểm định làm cực tiểu cả sai lầm loại 1 và sai lầm loại 2, nhưng không bao giờ tồn tại một kiểm định lý tưởng như vậy. Nếu chúng ta làm giảm sai lầm loại 1 thì sẽ làm tăng sai lầm loại 2 và ngược lại. Có 4 khả năng có thể xảy ra thể hiện trong bảng sau:

Kết luận / Thực tế	Chấp nhận H_0	Bác bỏ H_0 nhận H_1
H_0 đúng	Kết luận đúng	Sai lầm loại 1
H_0 sai	sai lầm loại 2	Kết luận đúng

Xác suất của việc mắc sai lầm loại 1 gọi là *mức ý nghĩa*, được ký hiệu là α. Xác suất mắc sai lầm loại 2 được ký hiệu là β. Trị số $1 - \beta$ được gọi là *lực lượng* của kiểm định. Lực lượng của kiểm định là xác suất bác bỏ H_0 khi H_0 sai. Giữa α và β cũng có mối liên hệ tương tự như mối liên hệ giữa hai loại sai lầm. Xác suất mắc sai lầm loại này có thể giảm đi nếu tăng xác suất mắc sai lầm loại kia. Sử dụng mối liên hệ này để ra quyết định cần chọn mức ý nghĩa thích hợp trên cơ sở xem xét những chi phí mất mát sẽ xảy ra đối với cả hai loại sai lầm.

Chẳng hạn, nếu mắc sai lầm loại 1 thì sẽ phải trả lại lô tấm lợp (ở thí dụ trên) và phải mất chi phí để xử lý lại lô tấm lợp đó mà lẽ ra được chấp nhận. Còn nếu mắc sai lầm loại 2 thì sẽ dẫn đến mất an toàn cho hàng ngàn người tới nhà thi đấu thể thao. Rõ ràng người ta dễ nghiêng về phía sai lầm loại 1 hơn so với sai lầm loại 2, có nghĩa là chọn mức ý nghĩa cho kiểm định cao để có β thấp. Nhưng ngược lại, nếu mắc sai lầm loại 1 sẽ dẫn đến việc phải tháo rời toàn bộ một động cơ hoàn chỉnh tại nhà máy, và mắc sai lầm loại 2 sẽ chỉ dẫn đến phải tiến hành một số sửa chữa bảo hành không đắt lắm, thì nhà sản xuất sẽ nghiêng về phía sai lầm loại 2, thà mắc sai lầm loại 2 còn hơn mắc sai lầm loại 1 và do đó sẽ chọn mức ý nghĩa kiểm định thấp.

Thông thường α được lấy là 0,01 ; 0,02 ; 0,05 hoặc 0,10. Từ mức ý nghĩa kiểm định α có thể xác định miền bác bỏ giả thiết H_0 và miền thừa nhận.

1.3. Tiêu chuẩn kiểm định.

Tiêu chuẩn kiểm định là quy luật phân phối xác suất nào đó được dùng để kiểm định. Trong tập hợp các kiểm định thống kê có cùng mức ý nghĩa α (tức là có xác suất mắc sai lầm loại 1 như nhau), kiểm định nào có xác suất mắc sai lầm loại 2 nhỏ nhất sẽ được xem là "tốt nhất". Vì vậy sau khi chọn mức ý nghĩa của kiểm định, việc tiếp theo là lựa chọn dạng phân phối thích hợp. Tuỳ thuộc vào giả thiết thống kê cần kiểm định mà người ta có thể sử dụng một số quy luật phân phối thông dụng như: quy luật phân phối chuẩn, phân phối T-Student, phân phối χ^2, phân phối Fisher...

1.4. Các bước tiến hành một kiểm định giả thiết thống kê.

Để tiến hành một kiểm định giả thiết thống kê cần thực hiện tuần tự các bước sau:

+ Phát biểu giả thiết H_0 và giả thiết đối H_1.

+ Định rõ mức ý nghĩa α (xác suất mắc sai lầm loại 1)

+ Chọn tiêu chuẩn kiểm định.

+ Tính giá trị của tiêu chuẩn kiểm định từ mẫu quan sát.

+ Kết luận bác bỏ hay chấp nhận H_0 tuỳ theo giá trị của tiêu chuẩn kiểm định rơi vào miền bác bỏ hay chấp nhận. Cụ thể :

- Nếu giá trị của tiêu chuẩn kiểm định thuộc miền bác bỏ: H_0 sai, bác bỏ giả thiết H_0 , thừa nhận H_1.

- Nếu giá trị của tiêu chuẩn kiểm định thuộc miền chấp nhận: Trong trường hợp này không nên hiểu rằng H_0 hoàn toàn đúng mà chỉ nên hiểu rằng qua mẫu cụ thể này chưa đủ cơ sở để bác bỏ H_0, cần nghiên cứu thêm.

2. Kiểm định và so sánh số trung bình

Nội dung phần này đề cập đến một số vấn đề: Kiểm định giả thiết về giá trị trung bình của một tổng thể chung; so sánh hai giá trị trung bình của hai tổng thể chung và so sánh nhiều trung bình thuộc nhiều tổng thể chung.

2.1. Kiểm định giả thiết về giá trị trung bình của một tổng thể chung.

Giả sử lượng biến của tiêu thức X trong tổng thể chung phân phối theo theo quy luật chuẩn với trung bình (kỳ vọng) là μ và phương sai là σ^2. Ký hiệu: N (μ,σ^2).Ta chưa biết μ, nhưng nếu có cơ sở để giả thiết rằng nó bằng μ_0, ta đưa ra giả thiết thống kê $H_0 : \mu = \mu_0$.

Để kiểm định giả thiết này, từ tổng thể chung ta tiến hành điều tra chọn mẫu ngẫu nhiên n đơn vị và tính được trung bình mẫu là $\bar{x}$.

Để chọn tiêu chuẩn kiểm định thích hợp, ta xét các trường hợp sau:

2.1.1 Phương sai của tổng thể chung σ^2 đã biết.

Tiêu chuẩn kiểm định được chọn là thống kê Z :

$$Z = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} \qquad (5.1)$$

Nếu giả thiết H_0 đúng, ta có :

$$Z = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma}$$

Đại lượng Z phân phối theo quy luật chuẩn hoá N(0,1). Từ đó tuỳ thuộc vào dạng của giả thiết đối H_1 mà miền bác bỏ được xây dựng theo các trường hợp sau:

Kiểm định phía phải: Giả thiết H_0: $\mu = \mu_0$

$H_1: \mu > \mu_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5-\alpha}$. Nếu $Z > Z_{0,5-\alpha}$, ta bác bỏ giả thiết H_0, nhận H_1.

Kiểm định phía trái: Giả thiết $H_0: \mu = \mu_0$

$H_1: \mu < \mu_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5-\alpha}$. Nếu $Z < -Z_{0,5-\alpha}$ hay $|Z| > Z_{0,5-\alpha}$; ta bác bỏ giả thiết H_0, nhận H_1.

Kiểm định hai phía: Giả thiết $H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5-\alpha/2}$. Nếu $|Z| > Z_{0,5-\alpha/2}$; ta bác bỏ giả thiết H_0, nhận H_1.

2.1.2 Phương sai của tổng thể chung σ^2 chưa biết, mẫu lớn ($n \geq 30$).

Trong trường hợp này ta vẫn dùng tiêu chuẩn kiểm định như trên, trong đó độ lệch tiêu chuẩn σ được thay bằng độ lệch tiêu chuẩn mẫu.

$$Z = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s} \qquad (5.2)$$

Trong đó : s là độ lệch tiêu chuẩn mẫu

Theo định lý giới hạn trung tâm, đại lượng Z có phân phối xấp xỉ chuẩn, cho dù tổng thể chung có phân phối như thế nào. Và cũng tương tự như trên, tuỳ thuộc vào giả thuyết đối H_1 mà miền bác bỏ được xây dựng theo các trường hợp sau:

Kiểm định phía phải: Giả thiết $H_0: \mu = \mu_0$

$H_1: \mu > \mu_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5-\alpha}$. Nếu $Z > Z_{0,5-\alpha}$, ta bác bỏ giả thiết H_0, nhận H_1.

Kiểm định phía trái: Giả thiết $H_0: \mu = \mu_0$

$H_1: \mu < \mu_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5-\alpha}$. Nếu $Z < -Z_{0,5-\alpha}$ hay $|Z| > Z_{0,5-\alpha}$; ta bác bỏ giả thiết H_0, nhận H_1.

Kiểm định hai phía: Giả thiết $H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5 - \alpha/2}$. Nếu $|Z| > Z_{0,5 - \alpha/2}$; ta bác bỏ giả thiết H_0, nhận H_1 .

Thí dụ 1:

Một công ty có hệ thống máy tính có thể xử lý 1200 hoá đơn trong 1 giờ. Công ty mới nhập một hệ thống máy tính mới. Hệ thống này khi chạy kiểm tra trong 40 giờ cho thấy số hoá đơn được xử lý trung bình trong 1 giờ là 1260 với độ lệch tiêu chuẩn là 215. Với mức ý nghĩa 5% hãy nhận định xem hệ thống mới có tốt hơn hệ thống cũ hay không?

Ta cần kiểm định giả thiết:

$H_0 : \mu = 1200$ (Hệ thống mới tốt bằng hệ thống cũ)

$H_1 : \mu > 1200$ (Hệ thống mới tốt hơn hệ thống cũ)

Đây là bài toán kiểm định giả thiết về giá trị trung bình của tổng thể chung khi chưa biết phương sai tổng thể chung nhưng mẫu lớn, kiểm định phải, tiêu chuẩn kiểm định được chọn là công thức 5.2; kết quả như sau:

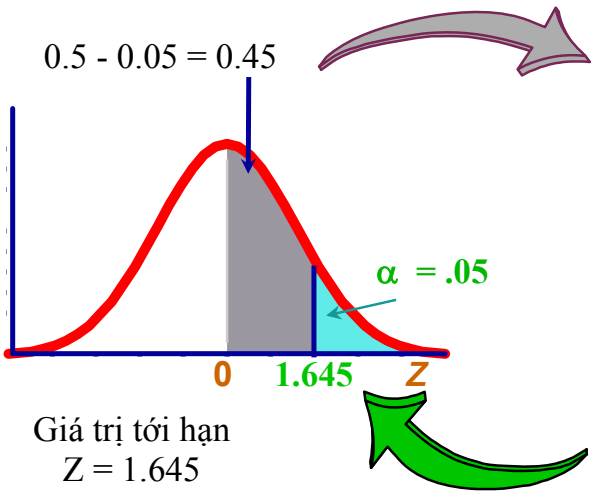
$$Z = \frac{(1260 - 1200)\sqrt{40}}{215} = 1{,}76$$

Tra bảng : $Z_{0,5 - \alpha} = Z_{0,5 - 0,05} = Z_{0,45} = 1{,}64$

Ta thấy : $Z > Z_{0,5 - \alpha}$ nên bác bỏ H_0 và kết luận hệ thống mới tốt hơn hệ thống cũ ở mức ý nghĩa 0,05.

Cách tra bảng :

Z = ? khi α = 0.05



Z	.04	.05	.06
1.6	.4495	.4505	.5515
1.7	.5591	.5599	.5608
1.8	.5671	.5678	.5686
1.9	.5738	.5744	.5750

Thí dụ 2:

Một nhà máy sản xuất săm lốp ô tô tuyên bố rằng tuổi thọ trung bình một chiếc lốp ôtô của họ là 30.000 dặm. Cơ quan giám định chất lượng nghi ngờ lời tuyên bố này đã kiểm

tra 100 chiếc lốp và tìm được trung bình mẫu là 29000 dặm với độ lệch tiêu chuẩn là 5000 dặm. Với mức ý nghĩa 0,05 cơ quan giám định có bác bỏ được lời quảng cáo của nhà máy trên không ?

Trong trường hợp này cơ quan kiểm định nghĩ rằng tuổi thọ trung bình của một chiếc lốp ôtô không phải là 30.000 dặm, giả thiết cần kiểm định là:

$H_0 : \mu = 30000$

$H_1 : \mu < 30000$

Đây là bài toán kiểm định giả thiết về giá trị trung bình của tổng thể chung khi chưa biết phương sai tổng thể chung nhưng mẫu lớn, kiểm định trái, tiêu chuẩn kiểm định được chọn là công thức 5.2; kết quả như sau:

$$\text{Ta có: } Z = \frac{(29000 - 30000)\sqrt{100}}{5000} = -2$$

Tra bảng : $Z_{0,5 - \alpha} = Z_{0,5 - 0,05} = Z_{0,45} = 1,64$

Ta thấy : $Z < - Z_{0,5 - \alpha}$ nên ta bác bỏ H_0 và kết luận quảng cáo của nhà máy là quá sự thật ở mức ý nghĩa 0,05.

Thí dụ 3:

Một nhóm nghiên cứu công bố rằng trung bình một người vào siêu thị A tiêu hết 140 ngàn đồng. Chọn ngẫu nhiên 50 người mua hàng ta tính được số tiền trung bình họ tiêu là 154 ng.đồng với độ lệch tiêu chuẩn là 62 ng.đồng. Với mức ý nghĩa 0,02 hãy kiểm định xem công bố của nhóm nghiên cứu có đúng không?

Ta cần kiểm định giả thiết:

$H_0 : \mu = 140$

$H_1 : \mu \neq 140$

Đây là bài toán kiểm định giả thiết về giá trị trung bình của tổng thể chung khi chưa biết phương sai tổng thể chung nhưng mẫu lớn, kiểm định hai phía, tiêu chuẩn kiểm định được chọn là công thức 5.2; kết quả như sau:

$$\text{Ta có: } Z = \frac{(154 - 140)\sqrt{50}}{62} = 1,59$$

Tra bảng : $Z_{0,5 - \alpha/2} = Z_{0,5 - 0,02/2} = Z_{0,49} = 2,33$

Vì $|Z| < Z_{0,5 - \alpha/2}$ nên có thể kết luận rằng với mẫu đã điều tra chưa đủ cơ sở để bác bỏ H_0 , ta tạm thời chấp nhận rằng báo cáo của nhóm nghiên cứu là đúng.

2.1.3. Phương sai của tổng thể chung σ^2 chưa biết, mẫu nhỏ (n < 30).

Trong trường hợp này tiêu chuẩn kiểm định được chọn là thống kê t :

$$t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s} \qquad (5.3)$$

Người ta đã chứng minh được rằng nếu H_0 đúng thì t sẽ phân phối theo quy luật Student với (n - 1) bậc tự do, s là độ lệch tiêu chuẩn mẫu .

Tuỳ thuộc vào giả thuyết đối H_1 mà miền bác bỏ được xây dựng theo các trường hợp sau:

Kiểm định phía phải: Giả thiết H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng tìm giá trị của $t_{\alpha,(n-1)}$. Nếu $t > t_{\alpha,(n-1)}$, ta bác bỏ giả thiết H_0 .

Kiểm định phía trái: Giả thiết H_0: $\mu = \mu_0$

H_1: $\mu < \mu_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng tìm giá trị của $t_{\alpha,(n-1)}$. Nếu $t < -t_{\alpha,(n-1)}$ hay $|t| > t_{\alpha,(n-1)}$, ta bác bỏ giả thiết H_0 .

Kiểm định hai phía : Giả thiết H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng tìm giá trị của $t_{\alpha/2,(n-1)}$. Nếu $|t| > t_{\alpha/2,(n-1)}$, ta bác bỏ giả thiết H_0 .

Thí dụ 4:

Một bản nghiên cứu thông báo rằng mức tiêu dùng hàng tháng của một sinh viên là 420 nghìn đồng. Để kiểm tra người ta chọn ngẫu nhiên 16 sinh viên và tính được trung bình mỗi tháng họ tiêu 442 nghìn đồng với độ lệch tiêu chuẩn mẫu điều chỉnh là 60 nghìn đồng. Với mức ý nghĩa 5% nhận định xem kết luận của bản thông báo có thấp hơn sự thật hay không?

Ta cần kiểm định giả thiết:

H_0 : $\mu = 420$

H_1 : $\mu > 420$

Ta có : $$t = \frac{(442 - 420)\sqrt{16}}{60} = 1{,}47$$

Tra bảng phân phối Student với 15 bậc tự do ta tìm được $t_{0,05;15} = 1{,}753$.

Vì $t < t_{\alpha,(n-1)}$ do đó không có cơ sở để bác bỏ H_0. Bản thông báo đó được chấp nhận là đúng.

2.2. Kiểm định hai giá trị trung bình của hai tổng thể chung.

Trong phần này ta xét bài toán so sánh hai trung bình của hai tổng thể chung. Đây là vấn đề rất có ý nghĩa của thống kê. Trong thực tế chúng ta luôn phải làm phép so sánh: so sánh chất lượng của hai loại sản phẩm, của hai loại dịch vụ; so sánh hai cơ hội đầu tư; so sánh hai phương pháp dạy học ... Để giải quyết vấn đề trên ta có thể dùng các phương pháp kiểm định thống kê như kiểm định tham số trong các trường hợp hai mẫu độc lập và hai mẫu phụ thuộc ; kiểm định phi tham số.

2.2.1. Kiểm định hai giá trị trung bình của hai tổng thể chung - trường hợp hai mẫu độc lập

Giả sử có hai tổng thể chung: Tổng thể chung thứ nhất có các lượng biến của tiêu thức X_1 phân phối theo quy luật chuẩn N (μ_1, σ_1^2) và tổng thể chung thứ hai có các lượng biến của tiêu thức X_2 phân phối theo quy luật chuẩn N (μ_2, σ_2^2)

Nếu μ_1 và μ_2 chưa biết song có cơ sở để giả thiết rằng giá trị của chúng bằng nhau ta có giả thiết thống kê $H_0 : \mu_1 = \mu_2$.

Để kiểm định giả thiết trên, từ hai tổng thể chung người ta rút ra hai mẫu ngẫu nhiên *độc lập* với kích thước mẫu tương ứng là n_1 và n_2 , từ đó tính các trung bình mẫu là $\overline{x}_1$ và $\overline{x}_2$. Để chọn tiêu chuẩn kiểm định thích hợp, ta xét các trường hợp sau:

a) Đã biết phương sai của 2 tổng thể chung σ_1^2 *và* σ_2^2.

Tiêu chuẩn kiểm định được chọn là:

$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Đại lượng Z phân phối theo quy luật chuẩn hoá N (0, 1). Nếu giả thiết H_0 đúng thì :

$$Z = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \text{cũng có phân phối N (0, 1)} \qquad (5.4)$$

Với mức ý nghĩa của kiểm định α cho trước và tuỳ thuộc vào giả thiết đối H_1 mà ta xây dựng các miền bác bỏ như sau :

<u>*Kiểm định phía phải*</u>: Giả thiết H_0: $\mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5-\alpha}$. Nếu $Z > Z_{0,5-\alpha}$, ta bác bỏ giả thiết H_0.

<u>*Kiểm định phía trái*</u>: Giả thiết $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 < \mu_2$

Nếu $Z < -Z_{0,5-\alpha}$ hay $|Z| > Z_{0,5-\alpha}$; ta bác bỏ giả thiết H_0.

<u>*Kiểm định hai phía*</u>: Giả thiết $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5-\alpha/2}$. Nếu $|Z| > Z_{0,5-\alpha/2}$; ta bác bỏ giả thiết H_0.

b) Chưa biết phương sai của hai tổng thể chung σ_1^2 và σ_2^2, mẫu lớn (n_1 và $n_2 \geq 30$).

Trong trường hợp này ta vẫn dùng thống kê Z làm tiêu chuẩn kiểm định như phần a), trong đó các phương sai σ_1^2 và σ_2^2 được thay bởi các phương sai mẫu.

Như vậy thống kê Z có dạng :

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (5.5)$$

Nếu n_1 và $n_2 \geq 30$ thì theo định lý giới hạn trung tâm, Z có phân phối xấp xỉ chuẩn N (0, 1). Với mức ý nghĩa của kiểm định α cho trước và tuỳ thuộc vào giả thiết đối H_1 mà ta xây dựng các miền bác bỏ như sau :

<u>*Kiểm định phía phải*</u>: Giả thiết $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

Nếu $Z > Z_{0,5-\alpha}$, ta bác bỏ giả thiết H_0.

<u>*Kiểm định phía trái*</u>: Giả thiết $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 < \mu_2$

Nếu $Z < -Z_{0,5-\alpha}$ hay $|Z| > Z_{0,5-\alpha}$; ta bác bỏ giả thiết H_0.

<u>*Kiểm định hai phía*</u>: Giả thiết $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Nếu $|Z| > Z_{0,5-\alpha/2}$; ta bác bỏ giả thiết H_0 .

c) Chưa biết phương sai của hai tổng thể chung σ_1^2 và σ_2^2 , mẫu nhỏ (n_1 và $n_2 < 30$).

Trong trường hợp này tiêu chuẩn kiểm định được chọn là thống kê t :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (5.6)$$

Trong đó : s^2 là giá trị chung của hai phương sai mẫu s_1^2 và s_2^2

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad (5.7)$$

Người ta đã chứng minh được rằng nếu H_0 đúng, cả hai tổng thể chung có phân phối chuẩn thì t sẽ có phân phối Student với $(n_1 + n_2 - 2)$ bậc tự do.

Tuỳ thuộc vào giả thuyết đối H_1 mà miền bác bỏ được xây dựng theo các trường hợp sau:

<u>*Kiểm định phía phải*</u>: Giả thiết $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng tìm giá trị của $t_{\alpha,(n_1+n_2-2)}$.

Nếu $|t| > t_{\alpha,(n_1+n_2-2)}$, ta bác bỏ giả thiết H_0 .

<u>*Kiểm định phía trái*</u>: Giả thiết $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 < \mu_2$

Nếu $t < -t_{\alpha,(n_1+n_2-2)}$ hay $|t| > t_{\alpha,(n_1+n_2-2)}$, ta bác bỏ giả thiết H_0 .

<u>*Kiểm định hai phía*</u> : Giả thiết $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng tìm giá trị của $t_{\alpha/2,(n_1+n_2-2)}$.

Nếu $|t| > t_{\alpha/2,(n_1+n_2-2)}$, ta bác bỏ giả thiết H_0 .

2.2.2. Kiểm định hai giá trị trung bình của hai tổng thể chung - trường hợp hai mẫu phụ thuộc

Trong phần trên hai mẫu được lấy ra một cách độc lập. Tuy nhiên, trong nhiều trường hợp việc chọn các mẫu phụ thuộc, liên hệ với nhau lại có ý nghĩa. Thường việc sử dụng các mẫu phụ thuộc (các mẫu theo cặp) sẽ cho phép phân tích chính xác hơn vì khi đó loại trừ được các yếu tố ngoại vi mà ta không nghiên cứu. Chẳng hạn ta chỉ muốn so sánh năng suất của giống lúa mới với giống lúa cũ và bỏ qua sự khác nhau về các yếu tố khác như phân bón, nước tưới, sâu bọ... thì hai loại giống đó phải được trồng trên hai mảnh của mỗi thửa ruộng và ghi lại sản lượng thu được trên hai mảnh ở các thửa ruộng khác nhau đó...

Với các mẫu phụ thuộc, các bước kiểm định vẫn như trước. Điểm khác nhau chỉ ở chỗ quy mô mẫu phải bằng nhau và kiểm định sự khác nhau theo cặp (hay gọi là phương pháp so sánh từng cặp).

Bài toán tổng quát như sau: Giả sử có hai tổng thể chung: Tổng thể chung thứ nhất có các lượng biến của tiêu thức X_1 phân phối theo quy luật chuẩn N (μ_1, σ_1^2) và tổng thể chung thứ hai có các lượng biến của tiêu thức X_2 phân phối theo quy luật chuẩn N (μ_2, σ_2^2). Muốn so sánh sự khác nhau giữa μ_1 và μ_2 ta xét độ lệch trung bình μ_d . Ta chưa biết μ_d nhưng nếu có cơ sở để giả thiết rằng giá trị của nó bằng μ_0 , ta đưa ra giả thiết thống kê H_0 : $\mu_d = \mu_0$.

Để kiểm định giả thiết trên, từ hai tổng thể chung người ta rút ra hai mẫu phụ thuộc được hình thành bởi các cặp n quan sát độc lập của hai mẫu, từ đó tính $\bar{d}$ là trung bình của các độ lệch giữa các cặp giá trị của hai mẫu d_i. Như vậy ta đưa bài toán so sánh về bài toán kiểm định giả thiết về giá trị trung bình đã xét ở phần I. Tuy nhiên ở đây thường không biết phương sai của các độ lệch của tổng thể chung nên thay bằng phương sai của các độ lệch của tổng thể mẫu S_d^2, và dùng tiêu chuẩn kiểm định t :

$$t = \frac{(\bar{d} - \mu_0)\sqrt{n}}{S_d} \qquad (5.8)$$

Với mức ý nghĩa α cho trước, tuỳ thuộc vào giả thiết đối H_1 mà các miền bác bỏ được xây dựng tương tự như ở phần 1.

Nhận xét: Phương pháp so sánh từng cặp như trên có ưu điểm hơn phương pháp so sánh hai mẫu độc lập ở chỗ:

- Nó không cần giả thiết gì về phương sai của hai tổng thể chung σ_1^2 và σ_2^2

- Nó thường cho kết quả chính xác hơn vì đã bỏ được các nhân tố ngoại lai ảnh hưởng đến giá trị trung bình. Tuy nhiên nhược điểm của nó là việc bố trí thí nghiệm (điều tra) phức tạp hơn, chẳng hạn trong ví dụ trên phương pháp so sánh từng cặp đòi hỏi phải trồng lúa thí nghiệm trên hai mảnh của cùng một thửa ruộng với hai loại giống khác nhau.

Ta xét thí dụ sau để minh hoạ:

Người ta quảng cáo là những người tham gia chương trình luyện tập giảm cân trung bình sẽ giảm trên 17 pound. Một người rất quan tâm đến chương trình này nhưng còn nghi ngờ về lời quảng cáo và đòi có bằng chứng. Người ta đã đồng ý cho anh ta phỏng vấn ngẫu nhiên 10 người để ghi lại cân nặng của họ trước và sau chương trình. Số liệu ghi trong bảng sau (*đvị: Pound*)

Thứ tự người được ĐT	Cân nặng trước chương trình	Cân nặng sau chương trình	Số cân giảm (d_i)	d_i^2
1	189	170	19	361
2	202	179	23	529
3	220	203	17	289
4	207	192	15	225
5	194	172	22	484
6	177	161	16	256
7	193	174	19	361
8	202	187	15	225
9	208	186	22	484
10	233	204	29	841
Cộng	***2025***	***1828***	***197***	***4055***

Anh ta muốn kiểm định lời quảng cáo về mức giảm cân trung bình ít nhất là 17 pound với mức ý nghĩa 5%.

Giải: Ở đây có hai mẫu: một mẫu trước chương trình và một mẫu sau chương trình. Chúng rõ ràng có liên hệ với nhau vì vẫn chính là mười người được điều tra trong hai lần. Điều mà chúng ta thực sự quan tâm không phải là số cân nặng trước hay sau chương trình mà là sự khác nhau về số cân nặng. Nói cách khác, không phải chúng ta có hai mẫu về số cân nặng trước và sau chương trình mà đúng hơn là có một mẫu về số cân nặng giảm được sau chương trình tập luyện.

Như vậy giả thiết cần kiểm định là:

$H_0 : \mu_d = 17$ (Mức giảm cân trung bình là 17 pound)

$H_1 : \mu_d > 17$ (Mức giảm cân trung bình lớn hơn 17 pound)

Với mẫu là 10 người, tiêu chuẩn kiểm định được sử dụng là:

$$t = \frac{(\bar{d} - \mu_0)\sqrt{n}}{S_d}$$

Với số liệu tính toán trong bảng trên ta tính được $\bar{d}$ và s_d như sau:

$$\bar{d} = \frac{\sum d_i}{n} = \frac{197}{10} = 19{,}7$$

$S_d = 4.4$

Vậy : $$t = \frac{(\bar{d} - \mu_0)\sqrt{n}}{S_d} = \frac{(19{,}7 - 17)\sqrt{10}}{4{,}4} = 1{,}94$$

Với mức ý nghĩa 0,05 và bậc tự do là 9, tra bảng ta có $t_{0.05;9} = 1{,}833$. Ta thấy $t > t_{\alpha,(n-1)}$ do đó có thể bác bỏ giả thiết H_0 và kết luận rằng lời quảng cáo cho chương trình tập luyện về số cân giảm là đúng.

2.2.3. Kiểm định phi tham số

Các tiêu chuẩn thống kê để kiểm định sự khác nhau giữa hai trung bình của hai tổng thể chung được trình bày ở trên gọi là kiểm định có tham số. Khi tiến hành các kiểm định này thường phải dựa trên giả thiết quan trọng là tổng thể chung đang xét có phân phối chuẩn và hoặc kích thước mẫu khá lớn. Nếu một trong các điều kiện trên bị vi phạm thì các tiêu chuẩn đó không thể thực hiện được. Trong tình huống như vậy ta phải sử dụng các tiêu chuẩn phi tham số. Tiêu chuẩn này không đòi hỏi phải có các giả thiết về các dạng phân phối của tổng thể chung và dùng trong các phương pháp kiểm định tự do (đối với dạng phân phối), đó là các ***phương pháp kiểm định phi tham số***.

Sau đây là một số phương pháp kiểm định thông dụng để kiểm định sự giống và khác nhau giữa hai trung bình của hai tổng thể (dùng trong hai trường hợp mẫu độc lập và mẫu phụ thuộc).

2.2.3.1. Kiểm định Mann - Whitney.

Kiểm định Mann - Whitney được sử dụng khi chỉ có hai tổng thể nghiên cứu. Kiểm định này cho phép ta xác định xem có phải các mẫu *độc lập* được lấy ra từ cùng một tổng thể chung hoặc từ các tổng thể khác nhau nhưng có chung một phân phối hay không.

Bài toán tổng quát như sau:

Giả sử có hai tổng thể chung X và Y. Phân phối của hai tổng thể này chưa biết và không nhất thiết là phân phối chuẩn. Ta muốn biết liệu hai tổng thể chung này có khác nhau không, giả thiết cần kiểm định là:

H_0: $\mu_1 = \mu_2$ (không có sự khác nhau giữa hai tổng thể chung và do đó có cùng số trung bình)

H_1: $\mu_1 \neq \mu_2$ (có sự khác nhau giữa hai tổng thể chung và chúng có số

trung bình khác nhau)

Để kiểm định giả thiết này, từ tổng thể chung lấy ra 2 mẫu: Mẫu thứ nhất, gồm n_1 đơn vị có các lượng biến (x_1, x_2 ...x_{n1}) lấy ra từ tổng thể chung X. Mẫu thứ hai, gồm n_2 đơn vị có các lượng biến (y_1, y_2 ...y_{n2}) lấy ra từ tổng thể chung Y.

Tiêu chuẩn kiểm định Mann - Whitney được xây dựng như sau:

- Gộp 2 mẫu trên thành 1 mẫu với cỡ mẫu là $(n_1 + n_2)$
- Sắp xếp $(n_1 + n_2)$ lượng biến của 2 mẫu theo thứ tự tăng dần và xác định hạng của mỗi lượng biến đó.
- Tính tổng hạng của các lượng biến thuộc mẫu thứ nhất là R_1 và của mẫu thứ hai là R_2.

Như vậy tổng hạng chung $R = R_1 + R_2 = 1 + 2 + ... + (n_1 + n_2)$.

Người ta đã chứng minh được rằng: nếu H_0 đúng và $n_1, n_2 \geq 10$ thì R_1 có phân phối xấp xỉ chuẩn với trung bình là:

$$\mu_{R_1} = \frac{n_1(n_1 + n_2 + 1)}{2} \qquad (5.9)$$

và phương sai là $\sigma^2_{R_1} = \frac{n_1.n_2(n_1 + n_2 + 1)}{12}$ (5.10)

(Tương tự, ta có R_2 có phân phối xấp xỉ chuẩn với giá trị trung bình là:

$$\mu_{R_2} = \frac{n_2(n_1 + n_2 + 1)}{2} \qquad (5.11)$$

và phương sai là $\sigma^2_{R_2} = \frac{n_1.n_2(n_1 + n_2 + 1)}{12}$) (5.12)

Thông thường chúng ta chọn số nhỏ nhất giữa R_1 và R_2 để tính tiêu chuẩn kiểm định. Giả sử $R_1 < R_2$, khi đó tiêu chuẩn kiểm định được chọn là :

$$Z = \frac{R_1 - \mu_{R_1}}{\sigma_{R_1}} \qquad (5.13)$$

nếu $|Z| > Z_{0,5-\alpha/2}$ ta bác bỏ giả thiết H_0 .

(Nếu thay R_1 bằng R_2 cũng sẽ cho ta cùng một kết luận)

Chú ý: *Nếu trong dãy $(n_1 + n_2)$ các lượng biến của 2 mẫu có những giá trị trùng nhau thì ta quy ước hạng của các lượng biến trùng nhau đó đều được gán giá trị tính bằng trung bình cộng các số thứ tự của các lượng biến đó. Chẳng hạn có 4 lượng biến bằng nhau*

có số thứ tự trong dãy số là 5, 6, 7, 8 thì hạng của 4 lượng biến đó đều được gán giá trị là (5 + 6 + 7 + 8)/ 2 = 6,5 còn lượng biến tiếp theo đó vẫn có hạng là 9 như cũ.

Thí dụ:

Có 1 người lái xe thường xuyên đi lại giữa hai điểm A và B. Có 2 đường nối A và B là đường X và đường Y. Anh ta muốn chọn con đường đi nào mất ít thời gian nhất. Chọn ngẫu nhiên 10 ngày đi trên đường X và 10 ngày đi trên đường Y, anh ta có số liệu sau (thời gian tính bằng phút):

Đường X:	34	28	46	42	56	85	48	25	37	49
Đường Y:	45	49	41	55	39	45	65	50	47	51

Với mức ý nghĩa 5%, hãy nhận định xem có sự khác nhau về thời gian đi lại khi đi theo đường X và đường Y hay không.

Giải: Đầu tiên ta tính được thời gian trung bình đi trên đường X là 45 phút và trên đường Y là 48,5 phút. Tuy nhiên ta không có cơ sở để cho rằng thời gian đi trên đường X và thời gian đi trên đường Y có phân phối chuẩn hay xấp xỉ chuẩn với phương sai bằng nhau. Do đó, việc áp dụng tiêu chuẩn kiểm định Student đã trình bày ở phần trước là không "hợp pháp" (phù hợp) . Vì vậy cần áp dụng phương pháp kiểm định Mann - Whitney.

Trước hết ta lập bảng xếp hạng các số liệu như sau:

Đường	*Thời gian*	*Hạng*	*Đường*	*Thời gian*	*Hạng*
X	25	1	Y	47	11
X	28	2	X	48	12
X	34	3	X	49	13,5
X	37	4	Y	49	13,5
Y	39	5	Y	50	15
Y	41	6	Y	51	16
X	42	7	Y	55	17
Y	43	8	X	56	18
Y	45	9	Y	65	19
X	46	10	X	85	20

Tổng các hạng của đường X là:

$R_1 = 1 + 2 + 3 + 4 + 7 + 10 + 12 + 13,5 + 18 + 20 = 90,5$

Vì n_1 và n_2 đều bằng 10 nên R_1 có phân phối xấp xỉ chuẩn với :

$$\mu_{R_1} = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{10.(10 + 10 + 1)}{2} = 105$$

và phương sai là $\sigma^2_{R_1} = \frac{n_1.n_2(n_1 + n_2 + 1)}{12} = \frac{10 \times 10 \times (10 + 10 + 1)}{12} = 175$

Ta tính tiêu chuẩn kiểm định:

$$Z = \frac{R_1 - \mu_{R_1}}{\sigma_{R_1}} = \frac{90{,}5 - 105}{\sqrt{175}} = -1{,}1$$

Với mức ý nghĩa 0,05, tra bảng ta được $Z_{0,5\ -\ \alpha/2} = 1{,}96$. Như vậy $|Z| < Z_{0,5-\alpha/2}$ do đó ta không có cơ sở bác bỏ giả thiết H_0. Chúng ta tạm thời kết luận rằng thời gian đi giữa 2 con đường X và Y không khác nhau ở mức ý nghĩa 5%.

2.2.3.2. Kiểm định dấu và kiểm định hạng có dấu Wilcoxon

Đây là phương pháp kiểm định phi tham số dùng trong trường hợp 2 mẫu phụ thuộc. Ở phần trên, chúng ta dùng phương pháp so sánh từng cặp, nhưng phương pháp này đòi hỏi một giả thiết quan trọng là các chênh lệch của từng cặp quan sát (d_i) phải có phân phối chuẩn hay xấp xỉ chuẩn. Nếu giả thiết này không được thoả mãn cần sử dụng đến các kiểm định phi tham số. Trong phần này chúng ta sẽ đề cập đến 2 phương pháp kiểm định thông dụng nhất là kiểm định dấu và kiểm định hạng có dấu của Wilcoxon.

a) Kiểm định dấu.

Phương pháp này kiểm định dựa trên cơ sở các dấu âm hoặc dương của các chênh lệch trong từng cặp quan sát chứ không dựa vào giá trị của chúng.

Giả sử có hai tổng thể : chẳng hạn X là hiệu quả của phương pháp thứ nhất và Y là hiệu quả của phương pháp thứ hai tác động lên cùng một đối tượng (hay X và Y phụ thuộc). Ta muốn kiểm định giả thiết H_0 : "Hiệu quả của phương pháp thứ nhất và của phương pháp thứ hai là như nhau".

Để kiểm định giả thiết trên, người ta quan sát n cặp giá trị (x_1, y_1); (x_2, y_2) ... (x_n, y_n). Đặt $d_i = x_i - y_i$. Ta loại bỏ các d_i có giá trị bằng 0 vì chúng không mang lại thông tin gì. Gọi n' là số các d_i có giá trị khác 0 và n^+ là số các d_i mang dấu + . Nếu giả thiết H_0 đúng thì n^+ sẽ có phân phối nhị thức với tham số p = 0,5 và n'. Ta biết rằng nếu $(n'. 0{,}5) > 5$ tức $n' > 10$ thì tần suất $f = n^+/n'$ sẽ có phân phối xấp xỉ chuẩn với kỳ vọng 0,5 và độ lệch tiêu chuẩn là:

$$\sigma_p = \sqrt{\frac{pq}{n'}} = \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{n'}} = \frac{1}{2\sqrt{n'}}$$

Như vậy tiêu chuẩn kiểm định được chọn là:

$$Z = (f - 0{,}5)2\sqrt{n'} = \frac{2n^{+} - n'}{\sqrt{n'}} \qquad (5.14)$$

Đại lượng Z trên sẽ có phân phối chuẩn.

Với mức ý nghĩa α cho trước, tuỳ thuộc giả thiết đối mà ta có các trường hợp: - Kiểm định 2 phía: H_1- "Có sự khác nhau", ta bác bỏ H_0 khi $|Z| < Z_{0,5-\alpha/2}$

- Kiểm định 1 phía: H_1 - "phương pháp thứ nhất hiệu quả hơn phương pháp thứ hai", ta sẽ bác bỏ H_0 khi $Z > Z_{0,5-\alpha}$.

Thí dụ : Một thầy giáo dạy toán cho rằng việc cho học sinh ôn tập 1 tiết cuối kỳ có tác dụng tốt đến kết quả học tập của các em. Một mẫu gồm 21 học sinh được chọn để theo dõi điểm thi của các em trước và sau khi ôn tập. Kết quả thu được ở 3 cột đầu của bảng sau:

Học sinh	***Điểm thi trước***	***Điểm thi sau***	***Hiệu số d_i***	***Dấu của d_i***
(1)	*(2)*	*(3)*	*(4)*	*(5)*
1	22	21	-1	-
2	26	29	3	+
3	17	15	-2	-
4	20	20	0	0
5	28	26	-2	-
6	31	32	1	+
7	23	25	2	+
8	13	14	1	+
9	19	19	0	0
10	25	27	2	+
11	28	27	-1	-
12	24	25	1	+
13	27	27	0	0
14	18	20	2	+
15	20	23	3	+
16	14	16	2	+
17	24	26	2	+
18	15	20	5	+

19	19	20	1	+
20	18	17	-1	-
21	27	19	2	+

Trên cơ sở khảo sát đó, với mức ý nghĩa 5% liệu có thể kết luận rằng sau khi được ôn tập kết quả thi của các em có tốt hơn không?

Giải: Ký hiệu p là tỷ lệ học sinh có điểm thi sau cao hơn điểm thi trước. Ta cần kiểm định giả thiết $H_0 : p = 0,5$

$$H_1 : p > 0,5.$$

Với tài liệu thu được qua điều tra, ta tính được các chênh lệch giữa số điểm thi sau và điểm thi trước khi ôn tập (d_i) và dấu của các chênh lệch đó biểu hiện ở cột 4 và 5 ở bảng trên.

Theo đó ta có : $n' = 18$; $n^+ = 13$. Vậy $f = 13 / 18 = 0,722$. Và:

$$Z = (f - 0,5)2\sqrt{n'} = \frac{2n^+ - n'}{\sqrt{n'}} = \frac{2 \times 13 - 18}{\sqrt{18}} = 1,886$$

Với mức ý nghĩa 0,05 tra bảng ta có $Z_{0,5-\alpha} = 1,64$. Như vậy $Z > Z_{0,5-\alpha}$, ta bác bỏ giả thiết H_0 nghĩa là việc cho học sinh ôn tập có tác dụng nâng cao kết quả học tập của các em.

b) Kiểm định hạng có dấu của Wilcoxon.

Trong khi kiểm định dấu chỉ quan tâm tới dấu của các hiệu số d_i thì kiểm định hạng có dấu của Wilcoxon còn tính đến độ lớn của $|d_i|$. Như vậy kiểm định này sẽ có hiệu quả hơn kiểm định dấu. Các bước thực hiện như sau:

- Xuất phát từ 2 mẫu ta tính các d_i
- Bỏ qua các giá trị $d_i = 0$
- Tính hạng của $|d_i|$ ($d_i \neq 0$)

Gọi: n' là số các giá trị $d_i = 0$

R^+ là tổng các hạng của $|d_i|$ ứng với $d_i > 0$

R^- là tổng các hạng của $|d_i|$ ứng với $d_i < 0$

Người ta đã chứng minh được rằng nếu H_0 đúng thì R^+ và R^- đều có cùng phân phối với kỳ vọng là $\frac{n'(n'+1)}{4}$ và phương sai là $\frac{n'(n'+1)(2n'+1)}{24}$

Nếu $n' \geq 8$ thì R^+ và R^- có phân phối xấp xỉ chuẩn. Như vậy tiêu chuẩn kiểm định được chọn là:

$$Z = \frac{R - n'(n'+1)/4}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}} \qquad (5.15)$$

Đại lượng Z sẽ có phân phối N(0, 1). Trong đó R là R^+ hoặc R^- (thường lấy số nhỏ nhất trong 2 số đó). Giả thiết H_0 sẽ bị bác bỏ ở mức ý nghĩa α nếu $|Z| > Z_{0,5-\alpha/2}$.

Nhận xét về phương pháp phi tham số: Phương pháp phi tham số có những ưu, nhược điểm sau:

Ưu điểm :

- Chúng không đòi hỏi phải có giả thiết là tổng thể chung có phân phối chuẩn hoặc tuân theo một dạng phânphối cụ thể nào đó.

- Nói chung các phương pháp này dễ hiểu và dễ thực hiện. Kiểm định phi tham số có thể được dùng thay thế cho kiểm định tham số bằng cách thay thế các giá trị số bằng các thứ hạng của chúng như đã làm ở trên.

- Đôi khi ngay cả việc sắp xếp theo thứ tự hạng cũng không cần thiết. Thông thường cái cần làm chỉ là mô tả 1 kết quả là "tốt hơn" so với một kết quả khác. Gặp trường hợp đó hoặc khi việc đo lường không được chính xác, không đáp ứng được yêu cầu của kiểm định tham số thì ta có thể sử dụng các phương pháp phi tham số.

Nhược điểm:

- Kiểm định phi tham số bỏ qua một lượng thông tin nhất định chẳng hạn như việc thay giá trị số bằng thứ hạng.

- Kiểm định phi tham số không hiệu quả hay "sắc bén" (nói cách khác là không mạnh) bằng kiểm định tham số. Cần nhớ rằng: Nếu điều kiện cho phép dùng kiểm định tham số được thoả mãn thì ta nên dùng kiểm định có tham số.

2.3. Kiểm định nhiều trung bình thuộc nhiều tổng thể chung

Trong phần 2.2 chúng ta đã xét đến việc so sánh giá trị trung bình của hai tổng thể chung. ở đây chúng ta đề cập đến phương pháp so sánh đồng thời các trung bình của nhiều tổng thể chung (từ 3 trở lên), đó là phương pháp phân tích phương sai (ANOVA). Phân tích phương sai được vận dụng trong các trường hợp như: so sánh việc sử dụng 5 loại ống dẫn khí khác nhau; đánh giá hiệu quả của mỗi phương pháp trong 4 phương pháp học tập khác nhau hoặc so sánh hiệu quả của 4 loại phân bón khác nhau ... Có hai mô hình phân tích

phương sai: phân tích phương sai một nhân tố và phân tích phương sai hai nhân tố. Trong phần này chỉ trình bày phương pháp *phân tích phương sai một nhân tố*.

Giả sử ta có k tổng thể chung $X_1, X_2, ..., X_k$ có phân phối chuẩn, trong đó $X_i \sim N(\mu_i, \sigma_i^2)$. Các giá trị trung bình μ_i chưa biết song có cơ sở giả thiết rằng là chúng bằng nhau, ta có giả thiết cần kiểm định là $H_0: \mu_1 = \mu_2 = ... = \mu_k$.

Trong thống kê vấn đề trên thường được xem xét dưới góc độ sau đây: Giả sử chúng ta quan tâm tới một nhân tố X nào đó. Nhân tố X có thể xem xét ở k mức độ khác nhau. Ký hiệu X_i là hiệu quả của việc tác động của nhân tố X ở mức i . Như vậy μ_i là hiệu quả trung bình của nhân tố X ở mức i. Chúng ta muốn biết khi cho nhân tố X thay đổi ở các mức khác nhau thì điều đó có ảnh hương hay không tới hiệu quả trung bình. Chẳng hạn, chúng ta muốn nghiên cứu ảnh hưởng của giống tới năng suất cây trồng. Nhân tố ở đây là giống, các loại giống khác nhau là các mức của nhân tố. Hiệu quả của giống lên năng suất cây trồng được đo bằng sản lượng của cây trồng. Như vậy X_i chính là sản lượng của giống i và μ_i là sản lượng trung bình của giống i.

Để kiểm định giả thiết này, từ các tổng thể chung các giá trị của X_i người ta rút ra k mẫu ngẫu nhiên, độc lập, với kích thước tương ứng là $n_1, n_2, ..., n_k$. Các số liệu được trình bày thành bảng ở dạng sau:

	Các nhân tố				
	1	2	...j	k	
	x_{11}	x_{21}	$...x_{i1}$	x_{1k}	
	x_{21}	x_{22}	...	x_{2k}	
	$x_{i1}...$	...	...	$x_{ik}...$	
	x_{n11}	x_{n22}	...	x_{nkk}	$n = \sum_{j=1}^{k} n_j$
Tổng số	T_1	T_2	...	T_k	$T = \sum_{j=1}^{k} T_j$
Trung bình	$\overline{x}_1$	$\overline{x}_2$	... $\overline{x}_j$	$\overline{x}_k$	$\overline{x} = T/n$

Các bước phương pháp phân tích phương sai một nhân tố (ANOVA) được tiến hành theo trình tự sau đây:

Bước 1: Tính các trung bình.

+ Trung bình của các mẫu: $$\overline{x}_i = \frac{T_i}{n_i} = \frac{\sum_{j=1}^{n_i} x_{ji}}{n_i} \qquad (5.16)$$

+ Trung bình chung: $$\bar{x} = \frac{T}{n} = \frac{\sum_{j=1}^{k} T_j}{n} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n} \qquad (5.17)$$

Bước 2: Tính các tổng bình phương độ lệch.

+ Tổng bình phương chung, ký hiệu là SST (Total Sum of Squares):

$$SST = \sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{n} \qquad (5.18)$$

+ Tổng bình phương do ảnh hưởng của nhân tố, ký hiệu là SSF (Sum of Squares for Factor):

$$SSF = \sum_{j=1}^{k} (\bar{x}_j - \bar{x})^2 . n_j = \sum_{j=1}^{k} \frac{T_j^2}{n_j} - \frac{T^2}{n} \qquad (5.19)$$

+ Tổng bình phương do sai số, ký hiệu là SSE (Sum of Squares for Error):

$$SSE = \sum_i \sum_j (x_{ij} - \bar{x}_j)^2 = \sum_i \sum_j x_{ij}^2 - \sum_j \frac{T_j^2}{n_j} \qquad (5.20)$$

Từ các công thức trên, ta thấy:

$$SST = SSF + SSE \qquad (5.21)$$

Bước 3: Tính các phương sai tương ứng.

+ Phương sai do ảnh hưởng của nhân tố (hay phương sai giữa các mẫu), ký hiệu là MSF (Mean Square for Factor):

$MSF = \frac{SSF}{k-1}$, trong đó (k - 1) được gọi là bậc tự do của nhân tố.

+ Phương sai do sai số (hay phương sai trong các mẫu), ký hiệu là MSE (Mean Square for Error):

$MSE = \frac{SSE}{n-k}$, trong đó (n - k) được gọi là bậc tự do của sai số.

Bước 4: Kiểm định giả thiết.

Giả thiết H_0: $\mu_1 = \mu_2 = ... = \mu_k$.

H_1: Tồn tại ít nhất 1 cặp $\mu_j \neq \mu_j$ với $j \neq j'$.

Các kết quả nói trên được trình bày trong bảng sau đây và được gọi là bảng ANOVA (Analysis of Variance : Phân tích phương sai).

Nguồn	Tổng bình phương	Bậc tự do	Phương sai (TB bình phương)	Tỷ số F
Nhân tố	SSF	k - 1	MSF	$F = \frac{MSF}{MSE}$
Sai số	SSE	n - k	MSE	
Tổng	SST	n - 1		

Người ta chứng minh được rằng nếu giả thiết H_0 đúng thì tỷ số $F = \frac{MSF}{MSE}$ sẽ có phân phối Fisher với bậc tự do là (k - 1, n - k). Giả thiết H_0 sẽ bị bác bỏ ở mức ý nghĩa α, nếu $F > F_{\alpha, (k-1),(n-k)}$.

Thí dụ:

Điểm thi của 12 sinh viên học các giáo sư A, B, C được cho trong bảng sau (thang điểm 100) :

Giáo sư A	Giáo sư B	Giáo sư C
79	71	82
86	77	68
94	81	70
89	83	76

Với mức ý nghĩa 5%, kiểm định xem liệu điểm thi trung bình của các sinh viên theo học các giáo sư A, B, C có giống nhau không.

Giải: Kết quả tính toán cho ta bảng ANOVA như sau:

Nguồn	Tổng bình phương	Bậc tự do	Phương sai (TB bình phương)	Tỷ số F
Nhân tố	354,67	2	177,34	4,96
Sai số	322	9	35,78	
Tổng	676,67	11		

Với mức ý nghĩa 5%, tra bảng phân phối Fisher với bậc tự do (2, 9) ta tìm được giá trị bằng 4,26. Vì F = 4,96 >4,26 nên ta bác bỏ H_0, nghĩa là điểm thi trung bình của các sinh viên theo học 3 giáo sư nói trên là khác nhau ở mức ý nghĩa 5%.

3. Kiểm định tỷ lệ

Nội dung phần này đề cập đến một số vấn đề: Kiểm định giả thiết về tỷ lệ của một tổng thể chung; so sánh hai tỷ lệ của hai tổng thể chung và so sánh nhiều tỷ lệ thuộc nhiều tổng thể chung.

3.1. Kiểm định giả thiết về tỷ lệ của tổng thể chung.

Giả sử ở tổng thể chung, tỷ lệ theo một tiêu thức A nào đó là p. Nếu p chưa biết song có cơ sở để giả thiết rằng giá trị của nó bằng p_0, ta đưa ra giả thiết:

$$H_0 : p = p_0$$

Để kiểm định giả thiết đó ta lấy mẫu ngẫu nhiên kích thước n và thấy có n_A đơn vị có biểu hiện của tiêu thức A và $(n - n_A)$ đơn vị không có biểu hiện đó. Như vậy ta có tỷ lệ mẫu : $p_s = n_A / n$.

Với n đủ lớn ($n.p_0 \geq 5$ và $n(1- p_0) \geq 5$) ta chọn tiêu chuẩn kiểm định Z:

$$Z = \frac{(p_s - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \qquad (5.22)$$

Tuỳ thuộc vào dạng của giả thiết đối H_1 mà ta có miền bác bỏ được xây dựng theo các trường hợp sau:

Kiểm định phía phải: Giả thiết H_0: $p = p_0$

H_1: $p > p_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5 - \alpha}$. Nếu $Z > Z_{0,5 - \alpha}$, ta bác bỏ giả thiết H_0.

Kiểm định phía trái: Giả thiết H_0: $p = p_0$

H_1: $p < p_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5 - \alpha}$. Nếu $Z < - Z_{0,5 - \alpha}$ hay $|Z| > Z_{0,5 - \alpha}$; ta bác bỏ giả thiết H_0.

Kiểm định hai phía: Giả thiết H_0: $p = p_0$

H_1: $p \neq p_0$

Với mức ý nghĩa của kiểm định α cho trước, ta tra bảng N(0,1) tìm được $Z_{0,5 - \alpha/2}$. Nếu $|Z| > Z_{0,5 - \alpha/2}$; ta bác bỏ giả thiết H_0.

Thí dụ:

Một báo cáo nói rằng 18% gia đình ở thành phố A có máy tính cá nhân ở nhà. Để kiểm tra, người ta chọn ngẫu nhiên 80 gia đình trong thành phố có trẻ em đang đi học và

thấy có 22 gia đình có máy tính.Với mức ý nghĩa $\alpha = 2\%$ hãy kiểm định xem liệu trong các gia đình có trẻ em đang đi học, tỷ lệ gia đình có máy tính có cao hơn tỷ lệ chung không?

Giải: Gọi p là tỷ lệ gia đình có máy tính trong các gia đình có trẻ em đang đi học ở thành phố A.

Ta cần kiểm định giả thiết: H_0: $p = 0,18$

H_1: $p > 0,18$

Ta có $np_0 = 80.0,18 = 14,4 \geq 5$ và $n(1- p_0) = 80.0,82 = 65,6 \geq 5$ do đó điều kiện kiểm định được thoả mãn. Ta tính được tỷ lệ mẫu: $p_s = 22/ 80 = 0,275$ và tiêu chuẩn kiểm định:

$$Z = \frac{(p_s - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} = \frac{(0,275 - 0,18)\sqrt{80}}{\sqrt{0,18(1-0,18)}} = 2,21$$

Tra bảng ta được $Z_{0,5 - \alpha} = Z_{0,5 - 0,02} = 2,05$. Vì $Z > Z_{0,5 - \alpha}$ do đó bác bỏ giả thiết H_0, và kết luận trong các gia đình có trẻ đi học, tỷ lệ gia đình có máy tính cao hơn tỷ lệ chung.

3.2. So sánh hai tỷ lệ của hai tổng thể chung.

Giả sử có hai tổng thể chung, tỷ lệ theo một tiêu thức A nào đó của tổng thể chung thứ nhất là p_1 và của tổng thể chung thứ hai là p_2. Nếu p_1 và p_2 chưa biết, song có cơ sở để giả thiết rằng chúng bằng nhau, ta có giả thiết cần kiểm địnhlà: $H_0 : p_1 = p_2$. Để kiểm định giả thiết này, từ hai tổng thể chung ta rút ra hai mẫu ngẫu nhiên với kích thước tương ứng là n_1 và n_2 ; thấy có n_{1A} và n_{2A} đơn vị có biểu hiện của tiêu thức A.

Tính các tỷ lệ mẫu $p_{s_1} = \frac{n_{1_A}}{n_1}$ và $p_{s_2} = \frac{n_{2_A}}{n_2}$.

Khi n_1 và n_2 khá lớn (n_1p_{s1} ; $n_1(1- p_{s1})$; n_2p_{s2} ;$n_2(1- p_{s2}) \geq 5$) thì Z phân phối xấp xỉ chuẩn N(0, 1). Nếu giả thiết H_0 đúng thì tiêu chuẩn kiểm định có dạng:

$$Z = \frac{p_{s_1} - p_{s_2}}{\sqrt{p_s(1-p_s)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (5.23)$$

Trong đó : p_s là tỷ lệ chung của cả hai mẫu và được tính bằng:

$$p_s = \frac{n_1p_{s_1} + n_2p_{s_2}}{n_1 + n_2} = \frac{n_{1A} + n_{2A}}{n_1 + n_2} \qquad (5.24)$$

Đại lượng Z vẫn có phân phối xấp xỉ chuẩn N(0, 1).

Với mức ý nghĩa α cho trước, tuỳ thuộc vào dạng của giả thiết đối H_1 mà ta có miền bác bỏ được xây dựng theo các trường hợp tương tự như trên.

Kiểm định phía phải: Giả thiết H_0: $p_1 = p_2$

H_1: $p_1 > p_2$

Nếu $Z > Z_{0,5-\alpha}$, ta bác bỏ giả thiết H_0 .

Kiểm định phía trái: Giả thiết H_0: $p_1 = p_2$

H_1: $p_1 < p_2$

Nếu $Z < -Z_{0,5-\alpha}$ hay $|Z| > Z_{0,5-\alpha}$; ta bác bỏ giả thiết H_0 .

Kiểm định hai phía : Giả thiết H_0: $p_1 = p_2$

H_1: $p_1 \neq p_2$

Nếu $|Z| > Z_{0,5-\alpha/2}$; ta bác bỏ giả thiết H_0.

Thí dụ:

Công ty nước giải khát Côca - Côla đang nghiên cứu việc đưa vào một công thức mới để cải tiến sản phẩm của mình. Với công thức cũ khi cho 500 người dùng thử thì có 120 người ưa thích nó. Với công thức mới khi cho 1000 người khác dùng thử thì có 300 người tỏ ra ưa thích nó. Hãy kiểm định xem liệu công thức mới đưa vào có làm tăng tỷ lệ những người ưa thích Côca hay không với mức ý nghĩa là 2%?

Giải: Gọi p_1 là tỷ lệ những người ưa thích Côca với công thức mới, p_2 là tỷ lệ những người ưa thích Côca với công thức cũ. Ta cần kiểm định giả thiết:

H_0: $p_1 = p_2$

H_1: $p_1 > p_2$

Với các số liệu đã có ta tính được:

$P_{s1} = n_{1A}/ n_1 = 300/ 1000 = 0,3$; $p_{s2} = n_{2A}/ n_2 = 120/ 500 = 0,24$

Và tỷ lệ chung: $p_s = \dfrac{300+120}{500+1000} = \dfrac{420}{1500} = 0,28$

Trong trường hợp này n_1 và n_2 đủ lớn, tiêu chuẩn kiểm được tính như sau:

$$Z = \frac{p_{s_1} - p_{s_2}}{\sqrt{p_s(1-p_s)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0,3-0,24}{\sqrt{0,28(1-0,28)\left(\frac{1}{1000}+\frac{1}{500}\right)}} = \frac{0,06}{0,025} = 2,4$$

Vì $Z = 2,4 > Z_{0,5-\alpha} = 2,06$, nên ta bác bỏ giả thiết H_0 .

Kết luận: Tỷ lệ những người ưa thích Côca với công thức mới cao hơn tỷ lệ những người ưa thích Côca với công thức cũ. Như vậy Công ty có thể quyết định sử dụng công thức mới để tăng thị phần của mình.

3.3. Kiểm định nhiều tỷ lệ thuộc nhiều tổng thể chung

Trong phần trên ta đã sử dụng tiêu chuẩn kiểm định Z (phân phối chuẩn) để so sánh hai tỷ lệ của hai tổng thể chung. Để kiểm định ba (hay nhiều hơn) tỷ lệ người ta sử dụng tiêu chuẩn kiểm định là phân phối Khi bình phương.

Chúng ta xuất phát từ thí dụ sau: Để nghiên cứu tỷ lệ phụ nữ có từ 3 con trở lên ở 3 địa phương A, B, C xem có khác nhau không, từ mỗi địa phương người ta chọn ngẫu nhiên một số phụ nữ, kết quả như sau:

Địa phương (j) / Số con (i)	A	B	C	Tổng dòng i
Từ 2 con trở xuống	140	240	60	440
Hơn 2 con	60	160	60	280
Tổng cột j	200	400	120	720

Bảng trên gọi là bảng ngẫu nhiên 2 dòng (i = 1, 2) và 3 cột (j = 1, 2, 3). Gọi tỷ lệ phụ nữ có hơn 2 con của A, B, C lần lượt là p_1, p_2, p_3. Ta cần kiểm định giả thiết $H_0 : p_1 = p_2 = p_3$ (tỷ lệ ở 3 địa phương là như nhau)

$H_1 : p_1 \neq p_2 \neq p_3$ (tỷ lệ ở 3 địa phương khác nhau)

Các bước được tiến hành như sau:

+ Gọi n_{ij} là số phụ nữ có số con i ở địa phương j (ví dụ: số phụ nữ có từ 2 con trở xuống của địa phương A là $n_{11} = 140$ ...). $\hat{n}_{ij}$ là *tần số thực nghiệm* (do điều tra)

+ Từ bảng ngẫu nhiên, tính *tần số lý thuyết* như sau:

$$\hat{n}_{ij} = \frac{\text{Tổng của dòng i} \times \text{Tổng của cột j}}{n} \; ; \quad \text{với } n = \sum_i \sum_j n_{ij}$$

+ Tính tiêu chuẩn χ^2 :

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

χ^2 là một đại lượng ngẫu nhiên (vì giá trị của nó thay đổi từ mẫu này qua mẫu khác) và có số bậc tự do (df) = (số dòng - 1).(số cột - 1).

Ta có nhận xét rằng nếu giả thiết H_0 đúng thì n_{ij} càng gần với $\hat{n}_{ij}$ và do đó giá trị χ^2 nhỏ và ngược lại.

Với mức ý nghĩa α cho trước, tra bảng tìm $\chi^2_{\alpha,df}$, nếu $\chi^2 > \chi^2_{\alpha,df}$ ta bác bỏ giả thiết H_0 .

Trở lại thí dụ trên, ta tính χ^2 qua bảng tính sau:

Địa phương(j) / Số con (i)	A	B	C	Tổng dòng i
Từ 2 con trở xuống	140 122,222 2,586	240 244,44 0,081	60 73,333 2,424	440
Hơn 2 con	60 77,777 4,063	160 155,555 0,127	60 46,666 3,81	280
Tổng cột j	200	400	120	720

(Chú thích : Trong mỗi ô có 3 dòng : Dòng 1là tần số thực nghiệm

Dòng 2 là tần số lý thuyết

Dòng 3 là giá trị χ^2 của mỗi ô).

Ta tính được tiêu chuẩn χ^2 như sau:

$$\chi^2 = 2,586 + 0,081 + 2,424 + 4,063 + 0,127 + 3,81 = 13,091$$

Với mức ý nghĩa 0,05 và bậc tự do df = (2 - 1).(3 - 1) = 2, tra bảng ta có giá trị của $\chi^2_{\alpha,df}$ = $\chi^2_{0,05;\ 2}$ = 5,991. Vì $\chi^2 > \chi^2_{\alpha,df}$ nên bác bỏ giả thiết H_0, nhận H_1 và kết luận tỷ lệ nữ có từ 2 con trở lên ở 3 địa phương là thực sự khác nhau (hay nói cách khác : tỷ lệ nữ có từ 2 con trở lên phụ thuộc vào từng địa phương).

Những điểm cần chú ý khi sử dụng tiêu chuẩn χ^2 :

+ *Sử dụng cỡ mẫu lớn*: Nếu cỡ mẫu nhỏ thì giá trị χ^2 quá lớn dẫn đến loại bỏ quá nhiều giả thiết cần kiểm định. Cần tuân theo một nguyên tắc chung là không nên sử dụng tần số lý thuyết nhỏ hơn 5 đơn vị trong 1 ô của bảng phân phối ngẫu nhiên.

+ *Sử dụng cẩn thận số liệu thu thập*: Số liệu thu thập được phải đúng, nếu có nghi ngờ phải kiểm tra lại cách thu thập hoặc phương pháp tính toán, đo lường hoặc cả hai. Khi giá trị $\chi^2 = 0$, phải thận trọng đặt câu hỏi không có sự chênh lệch tuyệt đối giữa tần số thực nghiệm và tần số lý thuyết? Và kiểm tra lại số liệu.

BÀI TẬP

5.1. Một nhà sản xuất ô tô thay thế một động cơ khác cho các ô tô có tỉ lệ miles-per-gallon bình quân *(số dặm đường/ 1 gallon nhiên liệu)* là 31.5 trên đường cao tốc. Nhà sản xuất muốn kiểm tra xem liệu động cơ mới có làm thay đổi tỉ lệ miles-per-gallon của mô hình ô tô đó hay không. Một mẫu ngẫu nhiên gồm 100 cuộc chạy thử nghiệm cho $\bar{x}$=29.8 dặm/gallon và s=6.6 dặm/gallon. Với mức ý nghĩa là 0.05, liệu tỉ lệ miles-per-gallon trên đường cao tốc của các ô tô dùng động cơ mới có khác so với các ô tô dùng động cơ cũ hay không?

5.2. Một loại thuốc chữa bệnh chứa bình quân 247 parts per million (ppm) của một loại hoá chất xác định. Nếu mức độ tập trung lớn hơn 247 ppm, loại thuốc này có thể gây ra một số phản ứng phụ; nếu mức độ tập trung nhỏ hơn 247 ppm, loại thuốc này có thể sẽ không có hiệu quả. Nhà sản xuất muốn kiểm tra xem liệu mức độ tập trung bình quân trong một lô hàng lớn có đạt mức 247 ppm yêu cầu hay không. Một mẫu ngẫu nhiên gồm 60 đơn vị được kiểm nghiệm và người ta thấy rằng trung bình mẫu là 250 ppm và độ lệch chuẩn của mẫu là 12 ppm. Hãy kiểm định rằng mức độ tập trung bình quân trong toàn bộ lô hàng là 247 ppm với mức ý nghĩa α=0.05. Thực hiện điều đó với α=0.01. Kết luận của bạn như thế nào? Bạn có quyết định gì đối với lô hàng này? Nếu lô hàng đã được bảo đảm rằng nó chứa đựng mức độ tập trung bình quân là 247 ppm, quyết định của bạn sẽ như thế nào căn cứ vào việc kiểm định giả thiết thống kê?

5.3. Một cuộc nghiên cứu được thực hiện nhằm xác định sự thoả mãn của khách hàng trong thị trường ô tô Canada sau khi đã có một số thay đổi về dịch vụ khách hàng. Giả sử rằng trước khi thay đổi, mức độ thoả mãn bình quân của khách hàng là 77 (trong thang điểm 0-100). Một bảng câu hỏi điều tra được gửi tới một mẫu ngẫu nhiên gồm 350 người dân, những người này đã mua xe mới sau khi có sự thay đổi dịch vụ khách hàng, và mức độ thoả mãn bình quân của mẫu này là $\bar{x}$=84, độ lệch chuẩn mẫu tìm được là s=28. Hãy lựa chọn mức ý nghĩa α và xác định liệu rằng có bằng chứng thống kê chứng tỏ có sự thay đổi trong mức độ thoả mãn của khách hàng hay không? Nếu bạn xác định rằng có sự thay đổi, hãy cho biết, theo bạn, sự thoả mãn của khách hàng được cải thiện hay giảm sút?

5.4. Một công ty dịch vụ đầu tư cho rằng mức thu nhập hằng năm bình quân của các cổ phiếu trong một ngành là 11.5%. Một nhà đầu tư muốn kiểm định xem đánh giá này có đúng hay không. Nhà đầu tư chọn một mẫu ngẫu nhiên gồm 50 cổ phiếu trong ngành. Ông ta thấy rằng mức thu nhập hằng năm bình quân của mẫu là 10.8%, và độ lệch chuẩn mẫu là 3.4%. Liệu nhà đầu tư có đủ chứng cớ để bác bỏ nhận định của công ty đầu tư đó hay không? (sử dụng α=0.05)

5.5. Tara Pearl tìm thấy một cơ hội kinh doanh đồ nội thất trị giá nhiều triệu đô la trong lĩnh vực sản xuất futon. Giá bán lẻ của futon thay đổi giữa các cửa hàng khác nhau, và người ta nhận thấy mức giá bình quân là $210 cho một futon đôi. Để kiểm định giả thiết này, giám đốc marketing của Tara lựa chọn một mẫu ngẫu nhiên gồm 120 cửa hàng và thấy rằng mức

giá bình quân là $225 và độ lệch chuẩn là $82. Hãy thực hiện kiểm định với α=0.05 và α=0.01.

5.6. Người ta biết rằng số ngày bình quân các khách du lịch ở trong các khách sạn của Hồng Kông là 3.4 đêm. Một nhà phân tích ngành du lịch muốn kiểm định liệu những thay đổi gần trong đặc thù ngành du lịch Hồng Kông có làm thay đổi mức bình quân này hay không. Nhà phân tích có một mẫu ngẫu nhiên về số đêm các khách du lịch ở lại các khách sạn của Hồng Kông:

5, 4, 3, 2, 1, 1, 5, 7, 8, 4, 3, 3, 2, 5, 7, 1, 3, 1, 1, 5, 3, 4, 2, 2, 2, 6, 1, 7

Hãy thực hiện kiểm định. Sử dụng mức ý nghĩa 0.05.

5.7. Một cuộc điều tra các trường kinh doanh cho thấy rằng 16% tổng số chỗ trong các trường này đang trống. Một công ty dịch vụ muốn kiểm định liệu nhận định này có đúng hay không. Họ thu thập thông tin trên một mẫu ngẫu nhiên gồm 300 chỗ lựa chọn từ các trường đại học khác nhau trong nước. Kết quả cho thấy rằng 51 trong 300 chỗ được điều tra là chỗ trống. Hãy thực hiện kiểm định. Sử dụng α=0.05.

5.8. Giả sử Công ty Goodyear Tire nắm giữ 42% thị trường bánh xe ô tô của Mỹ. Những thay đổi gần đây trong hoạt động của công ty, đặc biệt là việc đa dạng hoá lĩnh vực kinh doanh, cũng như sự thay đổi trong chính sách cạnh tranh của công ty đã đòi hỏi công ty cần kiểm định căn cứ của nhận định cho rằng nó vẫn kiểm soát 42% thị trường. Một mẫu ngẫu nhiên gồm 550 ô tô trên đường cho thấy có 219 xe có bánh xe của Goodyear. Hãy thực hiện kiểm định. Sử dụng α=0.01.

5.9. Thị phần của một công ty rất nhạy cảm với cả mức độ quảng cáo của nó và mức độ quảng cáo của các đối thủ cạnh tranh. Một công ty có thị phần là 56% muốn kiểm định liệu con số đó có còn giá trị trong bối cảnh có những chiến dịch quảng cáo mới của các đối thủ cạnh tranh và sự tăng cường hoạt động quảng cáo của chính nó. Một mẫu ngẫu nhiên gồm 500 người tiêu dùng cho thấy rằng có 298 người đang sử dụng sản phẩm của công ty. Với mức ý nghĩa là 0.01, liệu đó có phải là bằng chứng để kết luận rằng thị phần của công ty không còn là 56%?

5.10. LINC là một phần mềm do công ty Burroughs phát triển nên. Chương trình sẽ tự động viết một số mã (coding) mà người lập trình thường phải làm. Người ta cho rằng LINC sẽ giúp tiết kiệm thời gian lập trình và cho phép người lập trình làm việc hiệu quả hơn. Trong một cuộc kiểm tra, 45 người lập trình (nhóm 1) được yêu cầu viết một chương trình mà không dùng LINC và sau đó chạy chương trình cho đến khi nó có thể chạy mà không bị lỗi. Thời gian từ khi bắt đầu cho đến khi kết thúc được ghi lại. Nhóm 2 bao gồm 32 người lập trình cũng được yêu cầu viết chương trình với sự hỗ trợ của LINC. Trước khi thu thập dữ liệu, người ta người ta quyết định thực hiện one-tailed test để chứng tỏ rằng phần mềm này làm giảm thời gian lập trình bình quân. Kết quả thu thập dữ liệu là $\bar{x}_1$=26 phút, $\bar{x}_2$=21

phút, s_1=8 phút, s_2=6 phút. Thực hiện kiểm định và đưa ra kết luận. Liệu LINC có hiệu quả trong việc giảm thời gian lập trình hay không?

5.11. Ikarus, một hãng sản xuất xe buýt của Hungari vừa bị mất thị trường CIS quan trọng của họ. Hiện nay, công ty đang thử nghiệm một loại động cơ mới cho xe buýt của họ. Họ đã tiến hành thu thập các mẫu ngẫu nhiên sau về số dặm trên mỗi gallon đối với loại động cơ cũ và mới:

Đ/cơ cũ: 8, 9, 7.5, 8.5, 6, 9, 9, 10, 7, 8.5, 6, 10, 9, 8, 9, 5, 9.5, 10, 8.

Đ/cơ mới: 10, 9, 9, 6, 9, 11, 11, 8, 9, 6.5, 7, 9, 10, 8, 9, 10, 9, 12, 11.5, 10, 7, 10, 8.5

Liệu đây có phải là bằng chứng khẳng định động cơ mới có hiệu quả kinh tế hơn động cơ cũ?

5.12. Một công ty có hệ thống máy tính có thể xử lý 1200 hoá đơn trong 1 giờ. Công ty mới nhập một hệ thống máy tính mới. Hệ thống này khi chạy kiểm tra trong 40 giờ cho thấy số hoá đơn được xử lý trung bình trong 1 giờ là 1260 với độ lệch tiêu chuẩn là 215. Với mức ý nghĩa 5% hãy nhận định xem hệ thống mới có tốt hơn hệ thống cũ hay không?

5.13. Một nhà máy sản xuất săm lốp ô tô tuyên bố rằng tuổi thọ trung bình một chiếc lốp ôtô của họ là 30000 dặm. Cơ quan giám định chất lượng nghi ngờ lời tuyên bố này đã kiểm tra 100 chiếc lốp và tìm được trung bình mẫu là 29000 dặm với độ lệch tiêu chuẩn là 5000 dặm. Với mức ý nghĩa 0,05 cơ quan giám định có bác bỏ được lời quảng cáo của nhà máy trên không ?

5.14. Một nhóm nghiên cứu công bố rằng trung bình một người vào siêu thị A tiêu hết 140 ngàn đồng. Chọn ngẫu nhiên 50 người mua hàng ta tính được số tiền trung bình họ tiêu là 154 ng.đồng với độ lệch tiêu chuẩn là 62 ng.đồng. Với mức ý nghĩa 0,02 hãy kiểm định xem công bố của nhóm nghiên cứu có đúng không?

5.15. Một bản nghiên cứu thông báo rằng mức tiêu dùng hàng tháng của một sinh viên là 420 nghìn đồng. Để kiểm tra người ta chọn ngẫu nhiên 16 sinh viên và tính được trung bình mỗi tháng họ tiêu 442 nghìn đồng với độ lệch tiêu chuẩn mẫu điều chỉnh là 60 nghìn đồng. Với mức ý nghĩa 5% nhận định xem kết luận của bản thông báo có thấp hơn sự thật hay không?

5.16 Tỷ lệ khách tiêu dùng 1 loại sản phẩm ở địa phương A là 60%. Sau chiến dịch quảng cáo người ta cho rằng tỷ lệ đã tăng lên. Để kiểm tra ý kiến này người ta phỏng vấn ngẫu nhiên 400 người và thấy có 250 người tiêu dùng loại sản phẩm đó. Với mức ý nghĩa 0,05 hãy kết luận về ý kiến trên.

5.17 Người ta đưa ra giả thiết là thu nhập bình quân đầu người của địa phương A là không vượt quá 150 ng.đ/tháng và độ lệch tiêu chuẩn về thu nhập là 20ng.đ. Để kiểm định giả thiết trên người ta chọn ngẫu nhiên 100 hộ và tính được thu nhập bình quân của một người một tháng là 153 ng.đ . Với mức ý nghĩa 0,05, hãy kiểm tra giả thiết trên.

5.18 Tuổi thọ trung bình của một loại bóng đèn theo quy định là 2000 giờ và độ lệch tiêu chuẩn là 36 gì. Nghi ngờ về tuổi thọ của lô bóng đèn mới sản xuất không đạt theo quy định, người ta lấy mẫu ngẫu nhiên kích thước n = 25 và kiểm tra thì thấy tuổi thọ trung bình là 1975 giờ. Với mức ý nghĩa của kiểm định là 0,01 hãy kiểm định điều nghi ngờ trên.

5.19 Một máy sản xuất bi, theo tiêu chuẩn kỹ thuật thì đường kính trung bình là 5 mm và độ lệch tiêu chuẩn là 0,025 mm. Nghi ngờ về độ chính xác của những viên bi được sản xuất ra không đảm bảo tiêu chuẩn trên, người ta chọn ngẫu nhiên 100 viên bi vừa được sản xuất ra và tính được đường kính trung bình là 4,995 mm. Hãy kiểm định về điều nghi ngờ trên với mức ý nghĩa là 0,01.

5.20 Theo quy định, trọng lượng trung bình các bao gạo trong kho là 50kg. Nghi ngờ gạo bị đóng thiếu, người ta chọn ngẫu nhiên 25 bao đem cân và thu được các kết quả sau:

Trọng lượng (kg)	***Số bao***
48,0 - 48,5	2
48,5 - 49,0	5
49,0 - 49,5	10
49,5 - 50,0	6
50,0 - 50,5	2

Với mức ý nghĩa 0,05 hãy kết luận điều nghi ngờ trên.

5.21 Tại 1 doanh nghiệp người ta xây dựng hai phương án sản xuất một loại sản phẩm. Để đánh giá xem chi phí trung bình theo hai phương án ấy có khác nhau hay không người ta tiến hành sản xuất thử và thu được các kết quả sau:(đvị: ng.đ)

Phương án 1	25	32	35	38	35	
Phương án 2	20	27	25	29	23	26

Chi phí theo cả hai phương án trên phân phối theo quy luật chuẩn với $\sigma = 1,5$; $\sigma = 1,2$. Với mức ý nghĩa 0,05 hãy rút ra kết luận về hai phương án trên.

5.22 Có 2 doanh nghiệp cùng sản xuất một loại sản phẩm. Người ta nghi ngờ rằng năng suất lao động bình quân của 2 DN đó khác nhau thực sự và chọn ngẫu nhiên từ mỗi DN một số công nhân để điều tra năng suất của họ. Gọi số công nhân được chọn ra ở DN thứ nhất là nhóm 1 (8 người), của DN thứ hai là nhóm 2 (10 người), ta có kết quả điều tra như sau: (sản phẩm)

Nhóm 1	29	27	24	30	28	22	32	26		
Nhóm 2	23	22	32	25	29	24	27	31	30	26

Với mức ý nghĩa 0,05 hãy rút ra kết luận.

5.23 Công ty nước giải khát Côca - Côla đang nghiên cứu việc đưa vào một công thức mới để cải tiến sản phẩm của mình. Với công thức cũ khi cho 500 người dùng thử thì có 120 người ưa thích nó. Với công thức mới khi cho 1000 người khác dùng thử thì có 300 người tỏ ra ưa thích nó. Hãy kiểm định xem liệu công thức mới đưa vào có làm tăng tỷ lệ những người ưa thích Côca hay không với mức ý nghĩa là 2%?

5.24 Một bà quản lý muốn biết xem hàng hóa của mình có được bán rộng rãi trong cả nước không nên đã làm một cuộc điều tra. Bà ta chia nước thành 4 vùng, trong từng vùng chọn một mẫu ngẫu nhiên 100 người tiêu dùng để điều tra. Kết quả như sau:

	Đông Bắc	Tây Nam	Đông Nam	Tây Bắc	Tổng cộng
Mua hàng	40	55	45	50	190
Không mua hàng	60	45	55	50	210
Cộng	100	100	100	100	400

Yêu cầu:

a. Lập bảng ngẫu nhiên dự đoán và thực nghiệm cho vấn đề này
b. Tính λ^2
c. Kiểm định H_1 và H_0
d. Cho mức ý nghĩa là 0,05. Giả thuyết thay thế có bị bác bỏ không?

5.25 Một chủ toà báo muốn biết độc giả của mình nhiều hay ít có quan hệ đến trình độ học vấn của họ hay không. Ông ta tổ chức một cuộc điều tra những người lớn ở trong vùng theo hai nội dung: trình độ học vấn và việc thường xuyên đọc báo. Kết quả như bảng sau:

Trình độ học vấn / Việc thường xuyên đọc báo	Sau Đại học	Đại học	Tốt nghiệp PTTH	Chưa tốt nghiệp PTTH	Tổng số
Không đọc	7	14	13	16	50
Thỉnh thoảng đọc	13	17	7	7	44
Chỉ đọc buổi sáng hoặc chiều	39	41	10	5	95
Đọc cả hai buổi	22	23	8	12	65
Tổng cộng	81	95	38	40	254

Với mức ý nghĩa 0,1, việc thường xuyên đọc báo có phải do trình độ học vấn quyết định không?

CHƯƠNG 6

KIỂM SOÁT QUÁ TRÌNH BẰNG THỐNG KÊ

1. Thế nào là kiểm soát quá trình bằng thống kê

Những công cụ thống kê như là bảng thống kê, biểu đồ Pareto, biểu đồ tần số, biểu đồ quan hệ… là một số trong rất nhiều công cụ quản lý chất lượng. Việc làm đó rất quan trọng để đánh giá sản phẩm và chọn quyết định đúng đắn. Nếu nhận thấy sản phẩm phù hợp thì tốt nhưng nếu sản phẩm không phù hợp thì đã muộn rồi: tình trạng không có chất lượng đã xảy ra.

Tránh cho tình trạng không có chất lượng xảy ra là mục đích của mọi "chính sách sản xuất không sai sót". Để thực hiện chính sách đó thì chúng ta phải giải quyết ba vấn đề:

- Chọn một quy trình sản xuất có khả năng tạo ra toàn là những sản phẩm phù hợp với quy định kỹ thuật.

- Một khi đã chọn quy trình có khả năng thì việc điều khiển nó để sản xuất một sản phẩm hay một dịch vụ phải luôn luôn phù hợp với quy định kỹ thuật.

- Trong khi sản xuất thì cần cải tiến liên tục quy trình và sản phẩm.

Biểu đồ kiểm soát là một công cụ có thể giải quyết những vấn đề này. Công cụ này là một loại bảng thống kê với một hay hai biểu đồ. Người ta gọi nó là phương pháp điều khiển quy trình bằng thống kê hay là SPC (Statistical Process Control).

2. Các loại khác biệt trong quá trình

Điều mà ai cũng biết là hai sản phẩm không bao giờ giống nhau một cách tuyệt đối cả. Sản phẩm có thể khác nhau nhiều, khác nhau ít, nhưng thế nào cũng khác nhau vì đó là một định luật của thiên nhiên. Nếu chỉ khác biệt rất ít thì chúng ta có thể coi chúng là giống nhau. Nhưng nếu khác biệt nhiều thì chúng ta bắt buộc phải coi chúng là khác nhau.

Những khác biệt có thể chia ra làm 3 loại:

- *Khác biệt trong cùng một đơn vị*: một chiếc bánh có nơi ngọt ít và có nơi ngọt nhiều, tình trạng sần sùi của cùng một tờ giấy khác nhau từ điểm đo này đến điểm đo khác.

- *Khác biệt giữa hai đơn vị được sản xuất theo cùng một quy trình*: khi thời tiết biến đổi đôi chút là hai thước vải cùng một loại có trọng lượng khác nhau, gió thổi hay không thổi khi chúng ta mở cửa lò là hai mẻ bánh mì sẽ ngon dòn khác nhau.

- *Khác biệt một cách chu kỳ*: chúng ta nhận thấy nồng độ mỡ trong sữa chua khác nhau nếu vắt bò buổi sáng hay vắt bò buổi chiều, độ chua nước cốt những quả dừa khác nhau tùy mùa hái quả.

Những khác biệt đó thể hiện bằng sự phân bố của những số liệu. Định luật phân bố thống kê thông thường được xác định bởi trung bình và khoảng biến thiên. Vì thế thông thường chúng ta dùng trung bình hay khoảng biến thiên hay cả hai thông số đó để theo dõi những biến động trong quy trình sản xuất. Chỉ cần một trong hai thông số đó biến đổi một cách đáng kể là chúng ta có thể nói rằng quy trình đã biến động.

3. Biểu đồ kiểm soát

3.1. Tác dụng của biểu đồ kiểm soát:

Biểu đồ kiểm soát là biểu đồ mô tả ghi nhận sự thay đổi của quá trình dựa trên cơ sở mối quan hệ giữa các tham số đo xu hướng trung tâm và độ biến thiên của quá trình. Nó có tác dụng sau:

+ Căn cứ vào biểu đồ đó cho phép xác định vấn đề cần thay đổi, cần cải tiến.

+ Căn cứ vào biểu đồ cho phép nhận dạng quá trình hoạt động ổn định hay không ổn định, trên cơ sở phân biệt các nguyên nhân ảnh hưng đến sự biến thiên của quá trình.

3.2. Đặc điểm phân tích biểu đồ

+ Là để đánh giá xem quá trình có nằm trong phạm vi kiểm soát hay không, khi quá trình được đánh giá là nằm trong phạm vi kiểm soát, có nghĩa là những biến thiên theo đặc tính chất lượng sản phẩm là tương đối ổn định và nằm trong giới hạn đã được thiết lập.

+ Có 2 loại nguyên nhân gây nên biến thiên của quá trình bao gồm:

- Nguyên nhân chung là những nhân tố vốn có trong điều kiện bình thường và phản ánh bản chất của quá tình.
- Nguyên nhan đặc biệt hình thành do những yếu tố bất thường ngoài hệ thống.

+ Kiểm soát quá trình dựa vào biểu đồ được tuân thủ theo nguyên tắc cơ bản của kiểm định giả thiết.

3.3. Các bước lập và phân tích biểu đồ kiểm soát

+ *Bước 1* : Điều tra thu thập số liệu. Lập phiếu kiểm tra, ghi chép số liệu vào phiếu kiểm tra.

+ *Bước 2* : Tính giá trị trung bình để vẽ đường đường trung tâm CL (Central line). Đường trung tâm được xác định dựa trên cơ sở vận dụng tham số trung tâm (trung bình).

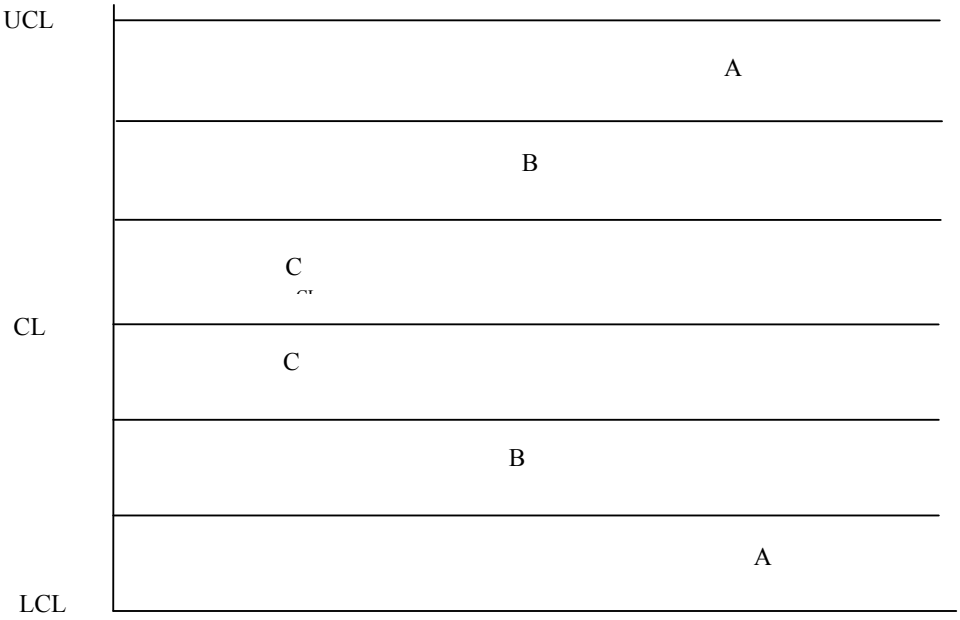
+ *Bước 3* : Tính giá trị và vẽ các đường giới hạn kiểm soát dưới (LCL) và giới hạn kiểm soát trên (UCL).

+ *Bước 4* : Vẽ biểu đồ kiểm soát theo các vùng kiểm soát:

- Vùng A tương ứng với phạm vi chênh lệch ± 3 lần độ lệch chuẩn tính từ đường trung tâm (tương ứng với xác suất 99%)

- Vùng B tương ứng với phạm vi chênh lệch ± 2 lần độ lệch chuẩn tính từ đường trung tâm với xác suất 95%

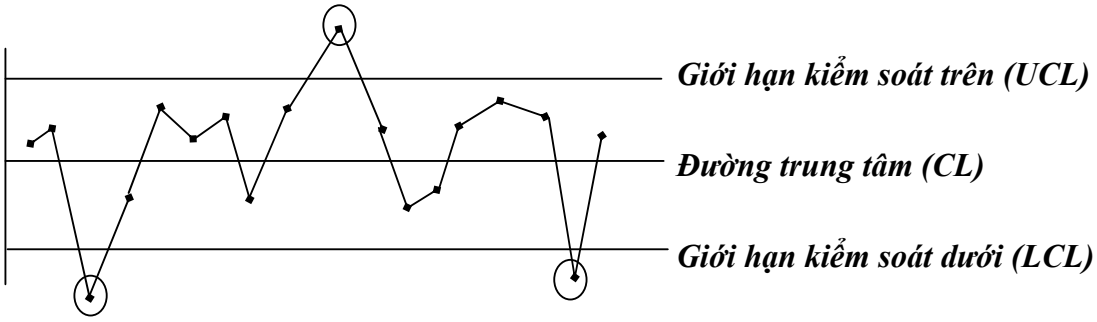
- Vùng C tương ứng với phạm vi ±1 lần độ lệch chuẩn tính từ đường trung tâm với xác suất 68%



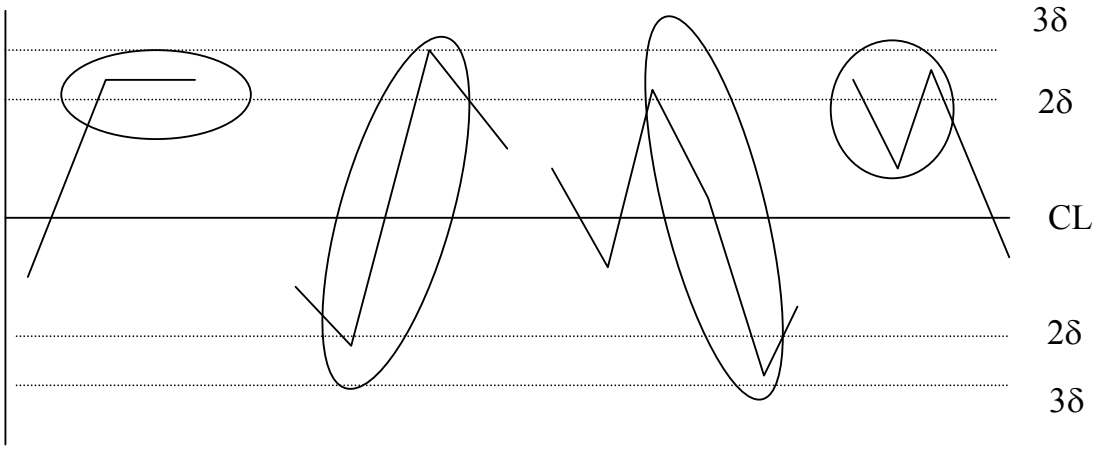
Chú ý : Thông thường giá trị của đường giới hạn kiểm soát dưới và trên được tính trên cơ sở sử dụng các hằng số tính sẵn cho từng loại biểu đồ.

3.4. Phân tích biểu đồ kiểm soát: Dựa trên cơ sở vận dụng quy tắc ngoài vùng kiểm soát với 5 quy tắc cụ thể sau:

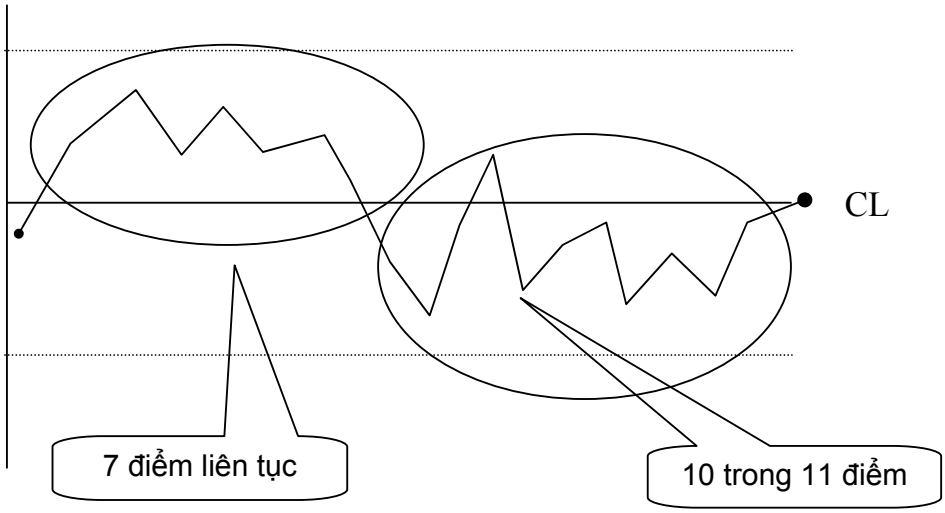
- Bất kỳ giá trị nào nằm ngoài giới hạn kiểm soát



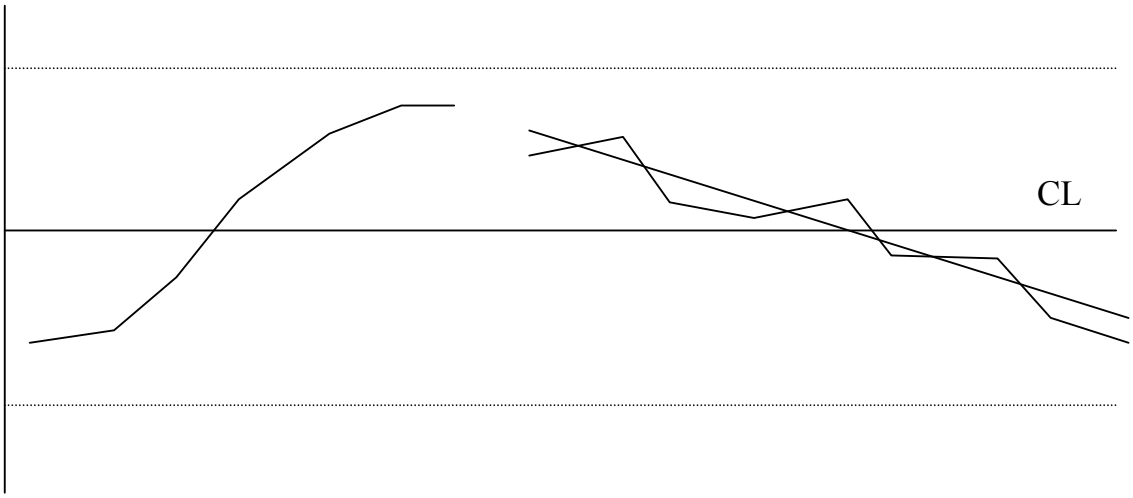
- Bất kỳ 2/3 điểm liên tiếp rơi vào vùng A cùng 1 phía của đường tâm hoặc nằm ngoài vùng 2δ



- Bất kỳ 4/5 điểm liên tiếp rơi vào vùng B cùng 1 phía của đường tâm
- Bất kỳ ≥8 điểm liên tiếp nằm cùng 1 phía của đường tâm



- Bất kỳ ≥8 điểm liên tiếp có xu hướng tăng hoặc giảm



3.5. Các loại biểu đồ kiểm soát:

+ Biểu đồ kiểm soát biến số: dựa trên số liệu theo tiêu thức số lượng: đặc tính kỹ thuật của sản phẩm

Bao gồm các biểu đồ:

- Biểu đồ trung bình: ví dụ trọng lượng trung bình…
- Biểu đồ độ lệch chuẩn (s) và biểu đồ khoảng biến thiên (R)
- Biểu đồ kết hợp trung bình và khoảng biến thiên ($\overline{X}$ - R), biểu đồ kết hợp trung bình và độ lệch chuẩn ($\overline{X}$ - s)

+ Biểu đồ kiểm soát thuộc tính: căn cứ vào dữ liệu theo tiêu thức thuộc tính: biểu đồ kiểm soát tỷ lệ (biểu đồ p) và b/đồ kiểm soát sai sót (biểu đồ c).

Sau đây trình bày một số dạng đơn giản, cơ bản và thường dùng:

4. Biểu đồ kiểm soát trung bình (Biểu đồ $\overline{X}$)

- Tác dụng: sử dụng để kiểm soát đặc tính chất lượng có thể đo lường được bằng trị số cụ thể.

- Phương pháp lập biểu đồ kiểm soát trung bình: dựa trên giả thiết giá trị trung bình mẫu $\overline{X}$ có đặc điểm phân phối chuẩn công thức tổng quát để xác định giới hạn kiểm soát.

- Các giới hạn kiểm soát:

$$LCL = \mu - z.\sigma_{\overline{x}}$$

$$UCL = \mu + z.\sigma_{\overline{x}}$$

$$(z = 3)$$

4.1. Trường hợp đã biết trung bình (μ)và độ lệch chuẩn (σ) của quá trình:

Đường trung tâm của biểu đồ: CL xác định tại giá trị μ

$$\text{Với } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Ta có 2 giới hạn kiểm soát:

Giới hạn kiểm soát dưới: $LCL = \mu - 3\frac{\sigma}{\sqrt{n}}$

Giới hạn kiểm soát trên: $UCL = \mu + 3\frac{\sigma}{\sqrt{n}}$

4.2. Trường hợp chưa biết trung bình (μ)và độ lệch chuẩn (σ) của quá trình

- Căn cứ vào dữ liệu mẫu, xác định trung bình của các trung bình mẫu và sử dụng làm ước lượng cho trung bình của quá trình

$$\overline{\overline{X}} = \frac{\sum \overline{X}_i}{k} \quad \text{k là số mẫu}$$

- Xác định trung bình của các độ lệch chuẩn mẫu

$$\bar{s} = \frac{\Sigma s}{k}$$

Trong đó s là các độ lệch chuẩn mẫu: $s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$

Trung bình của độ lệch chuẩn $\bar{s}$ là ước lượng chệch của độ lệch chuẩn của quá trình nên phải sử dụng hằng số C_4 để điều chỉnh và thiết lập giới hạn kiểm soát. Các giới hạn kiểm soát được xác định như sau:

$$LCL = \overline{\overline{X}} - 3\frac{\bar{s}}{C_4\sqrt{n}}$$

$$UCL = \overline{\overline{X}} + 3\frac{\bar{s}}{C_4\sqrt{n}}$$

Xác định hằng số C_4 bằng cách tra bảng. Trong bảng tính sẵn các hằng số có thể tra giá trị A_3 vì : $A_3 = \frac{3}{C_4\sqrt{n}}$

Khi đó: $CLC = \overline{\overline{X}} - A_3.\bar{s}$

$$UCL = \overline{\overline{X}} + A_3.\overline{s}$$

5. Biểu đồ kiểm soát độ lệch chuẩn của quá trình (Biểu đồ s)

5.1. Trường hợp biết độ lệch chuẩn của quá trình (σ)

- Đường trung tâm CL: $E(s) = C_5\sigma$

Kỳ vọng độ lệch chuẩn mẫu không đảm bảo là không chệch so với độ lệch chuẩn của quá trình do vậy là cơ sở để xác định đường trung tâm của quá trình và xác định các độ lệch chuẩn của quá trình

$$\sigma_s = C_5\sigma$$

- Các giới hạn kiểm soát:

$$LCL = C_4\sigma - 3C_5\sigma$$

$$UCL = C_4\sigma + 3C_5\sigma$$

5.2. Trường hợp chưa biết độ lệch chuẩn của quá trình (σ)

Xác định trung bình độ lệch chuẩn của các mẫu: $\overline{s} = \frac{\Sigma s}{k}$

Đó là căn cứ để thiết lập biểu đồ kiểm soát:

- Đường trung tâm CL:

$$CL = \overline{s}$$

- Các giới hạn kiểm soát:

$$LCL = \overline{s} - 3\frac{C_5\overline{s}}{C_4} = \left(1 - 3\frac{C_5}{C_4}\right)\overline{s} = B_3\overline{s}$$

$$UCL = \overline{s} + 3\frac{C_5\overline{s}}{C_4} = \left(1 + 3\frac{C_5}{C_4}\right)\overline{s} = B_4\overline{s}$$

6. Biểu đồ kiểm soát khoảng biến thiên (Biểu đồ R)

- Đường trung tâm CL:

$$CL = R = \frac{\sum R}{k}$$

- Các giới hạn kiểm soát:

$$\overline{R} \pm 3\frac{d_3}{d_2}\overline{R} = \left(1 \pm 3\frac{d_3}{d_2}\right)\overline{R}$$

Mà $1-3\frac{d_3}{d_2}=D_3$ và $1-3\frac{d_3}{d_2}=D_4$

Vì vậy: $LCL = D_3\overline{R}$

$UCL = D_4\overline{R}$

7. Biểu đồ kiểm soát tỷ lệ (Biểu đồ p)

7.1. Tác dụng

+ Vận dụng để lập biểu đồ kiểm soát phân tích chất lượng quá trình đối với các tiêu thức thuộc tính.

+ Thông thường kiểm soát: tỷ lệ phế phẩm, tỷ lệ sản phẩm không hợp chuẩn…

7.2. Cách xây dựng

- *Trường hợp đã biết tỷ lệ của quá trình (p):*

Đường trung tâm được xác định bởi giá trị tỷ lệ của quá trình.

$$CL = p$$

Do $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$ nên các đường giới hạn kiểm soát được xác định bằng:

$$p \pm 3\sigma_p = p \pm 3\sqrt{\frac{p(1-p)}{n}}$$

- *Trường hợp chưa biết tỷ lệ của quá trình (p)*

Với giả định các mẫu gần đây được lấy từ 1 quá trình ổn định.

+ Đường trung tâm CL:

$$CL = \overline{p_s} = \frac{\sum p_s}{k}$$

Trong đó: p_s là tỷ lệ từng mẫu

+ Các đường giới hạn kiểm soát được xác định bằng:

$$\overline{p}_s \pm 3\sqrt{\frac{\overline{p_s}(1-\overline{p_s})}{n}}$$

CÁC HẰNG SỐ CỦA BIỂU ĐỒ KIỂM SOÁT

n	A_2	A_3	B_3	B_4	C_5	C_4	d_2	d_3	D_3	D_4
2	1.88	2.659	0	3.267	0.6028	0.7979	1.128	0.853	0	3.267
3	1.023	1.954	0	2.568	0.4633	0.8862	1.693	0.888	0	2.574
4	0.729	1.628	0	2.266	0.3889	0.9213	2.509	0.880	0	2.282
5	0.577	1.427	0	2.089	0.3412	0.9400	2.326	0.864	0	2.114
6	0.483	1.287	0.029	1.970	0.3076	0.9515	2.534	0.848	0	2.004
7	0.419	1.182	0.113	1.882	0.2820	0.9594	2.704	0.833	0.076	1.924
8	0.373	1.000	0.179	1.815	0.2622	0.9650	2.847	0.820	0.136	1.864
9	0.337	1.032	0.232	1.761	0.2459	0.9693	2.970	0.808	0.184	1.816
10	0.300	0.075	0.276	1.716	0.2321	0.9727	3.078	0.797	0.223	1.777
11	0.285	0.927	0.313	1.679		0.9754	3.173	0.787	0.258	1.744
12	0.266	0.886	0.346	1.646		0.9776	3.258	0.778	0.283	1.717
13	0.294	0.850	0.374	1.618		0.9794	3.336	0.770	0.307	1.693
14	0.235	0.817	0.399	1.594		0.9810	3.407	0.763	0.328	1.672
15	0.223	0.789	0.421	1.572		0.9823	3.472	0.756	0.347	1.653
16	0.212	0.763	0.440	1.552		0.9835	3.532	0.750	0.363	1.637
17	0.203	0.739	0.458	1.534		0.9845	3.588	0.744	0.378	1.622
18	0.194	0.718	0.475	1.518		0.9854	3.640	0.739	0.391	1.608
19	0.187	0.698	0.490	1.503		0.9862	3.689	0.734	0.403	1.597
20	0.180	0.680	0.504	1.490		0.9869	3.735	0.729	0.415	1.585

BÀI TẬP

6.1 Một nhà máy sản xuất loại vòng bi dùng cho một loại đông cơ có đường kính lồng trục là 5cm với độ lệch chuẩn là 0,04cm. Cán bộ phụ trách quy trình sản xuất đã điều tra trong 10 ngày các mẫu gồm 5 vòng bi và xác định được kết quả như sau:

Mẫu số	***Đường kính đo được của các vòng bi (cm)***				
	1	***2***	***3***	***4***	***5***
1	5.02	5.01	4.94	4.99	4.96
2	5.01	5.03	5.07	4.95	4.96
3	4.99	5.00	4.93	4.92	4.99
4	5.03	4.91	5.01	4.98	4.89
5	4.95	4.92	5.03	5.05	5.01
6	4.97	5.06	5.06	4.96	5.03
7	5.05	5.01	5.10	4.96	4.99
8	5.09	5.10	5.00	4.99	5.08
9	5.14	5.10	4.99	5.08	5.09
10	5.01	4.98	5.08	5.07	4.99

a. Xác định các giới hạn kiểm soát, vẽ biểu đồ trung bình và phân tích tình trạng của quá trình

b. Lập biểu đồ khoảng biến thiên và phân tích biến động của quá trình

6.2. Một công ty sản xuất kem đánh răng muốn khắc phục vấn đề về tình trạng rò rỉ các tuýp kem đánh răng. Công ty thường đóng thành từng thùng chứa 100 tuýp kem đánh răng. Công ty đã chọn kiểm tra 10 thùng và xác định được số lượng tuýp bị rò rỉ như sau:

Mẫu	***Số bị rò rỉ***	Mẫu	***Số bị rò rỉ***
1	***4***	6	***6***
2	***8***	7	***10***
3	***12***	8	***9***
4	***11***	9	***5***
5	***12***	10	***8***

a. Loại biểu đồ kiểm soát nào có thể sử dụng thích hợp để phân tích biến động quá trình dựa vào dữ liệu trên? Tại sao?

b. Lập biểu đồ và phân tích tình trạng của quá trình

6.3 Một công ty dệt thực hiện kiểm soát chất lượng sản xuất của một loại vải. Hàng ngày, kiểm tra viên thực hiện đếm số lỗi xuất hiện. Trong giai đoạn ba tuần lễ kiểm tra viên kiểm tra từ 15 cuộn vải và ghi được kết quả như sau:

Mẫu	Số sai sót	***Mẫu***	Số sai sót	***Mẫu***	Số sai sót
1	12	***6***	11	***11***	12
2	8	***7***	9	***12***	10
3	16	***8***	14	***13***	14
4	14	***9***	13	***14***	17
5	10	***10***	15	***15***	15

Công ty tin rằng khoảng chừng 99% các sai sót là do các biến thiên ngẫu nhiên trong quá trình dệt gây ra, chỉ có 1% biến thiên là do nguyên nhân không ngẫu nhiên. Hãy lập biểu đồ kiểm soát và phân tích tình trạng của quá trình trên

CHƯƠNG 7

HỒI QUY VÀ TƯƠNG QUAN

Trên thực tế, chúng ta thường xuyên phải đưa ra những quyết định trong quản trị kinh doanh. Đó là những quyết định về một vấn đề mà nó có mối quan hệ tới nhiều yếu tố xung quanh và chịu ảnh hưởng bởi sự tác động của những yếu tố đó. Chẳng hạn: Một công ty quyết định tăng chi phí quảng cáo và muốn dự đoán mức doanh thu tương ứng; hoặc một giám đốc bán hàng muốn dự đoán mức bán hàng của từng nhân viên dựa trên số khách hàng từng nhân viên hiện đang có. Người vượt mức sẽ được thưởng và người không đạt sẽ được đào tạo thêm… Một trong những phương pháp đáp ứng được yêu cầu đó là phương pháp phân tích hồi quy và tương quan. Nội dung chương này sẽ đề cập đến các nội dung chủ yếu sau:

- Mối liên hệ giữa các hiện tượng kinh tế xã hội và nhiệm vụ của phương pháp hồi quy tương quan
- Xác định mô hình hồi quy tuyến tính đơn.
- Đánh giá cường độ của mối liên hệ.
- Ước lượng các giá trị trong tương lai dựa vào mô hình hồi quy.
- Mô hình Hồi quy bội

1. Mối liên hệ giữa các hiện tượng và phương pháp hồi quy tương quan.

1.1. Liên hệ hàm số và liên hệ tương quan

Các hiện tượng tồn tại trong mối liên hệ phụ thuộc lẫn nhau. Phương pháp phân tích hồi quy và tương quan là một trong những phương pháp thường được sử dụng trong thống kê để nghiên cứu mối liên hệ phụ thuộc đó. Khi nghiên cứu mối liên hệ phụ thuộc, nếu xét theo mức độ chặt chẽ của mối liên hệ, có thể phân thành hai loại : liên hệ hàm số và liên hệ tương quan .

a. Liên hệ hàm số

Liên hệ hàm số là mối liên hệ hoàn toàn chặt chẽ (khi hiện tượng này thay đổi có tác dụng quyết định đến sự thay đổi của hiện tượng có liên quan theo một tỷ lệ nhất định) giữa tiêu thức nguyên nhân - ký hiệu là x và tiêu thức kết quả - ký hiệu là y. Dạng tổng quát của liên hệ hàm số : y= f(x) . Điều đó có nghĩa là cứ mỗi giá trị của tiêu thức nguyên nhân sẽ có một giá trị tương ứng của tiêu thức kết quả. Mối liên hệ này có thể thấy được không những ở toàn bộ tổng thể, mà cả trên từng đơn vị cá biệt. Liên hệ hàm số thường gặp khi nghiên cứu các hiện tượng tự nhiên như trong vật lý, hoá học, v.v…Chẳng hạn: S = v.t (quãng đường bằng vận tốc nhân với thời gian).

b. Liên hệ tương quan

Liên hệ tương quan là mối liên hệ không hoàn toàn chặt chẽ giữa tiêu thức nguyên nhân (biến độc lập) và tiêu thức kết quả (bến phụ thuộc): cứ mỗi giá trị của tiêu thức nguyên nhân sẽ có nhiều giá trị tương ứng của tiêu thức kết quả. Thí dụ: mối liên hệ giữa số lượng sản phẩm và giá thành đơn vị sản phẩm. Không phải khi khối lượng sản phẩm tăng lên thì giá thành đơn vị sản sẽ giảm theo một tỷ lệ tương ứng. Cũng như mối liên hệ giữa số lượng phân bón và năng suất cây trồng, mối liên hệ giữa vốn đầu tư và kết quả sản xuất v.v... Các mối liên hệ này là các mối liên hệ không hoàn toàn chặt chẽ, không được biểu hiện một cách rõ ràng trên từng đơn vị cá biệt. Do đó, để phản ảnh mối liên hệ tương quan thì phải nghiên cứu hiện tượng số lớn - tức là thu thập tài liệu về tiêu thức nguyên nhân và tiêu thức kết quả của nhiều đơn vị.

Liên hệ tương quan thường gặp khi nghiên cứu các hiện tượng kinh tế - xã hội.

1.2 Nhiệm vụ của phân tích hồi quy và tương quan

Phân tích hồi quy và tương quan giải quyết hai nhiệm vụ cơ bản sau đây:

1.2.1. Xác định mô hình hồi quy phản ánh mối liên hệ

Căn cứ vào nhiệm vụ nghiên cứu cụ thể để chọn ra một, hai, ba, v.v.. tiêu thức nguyên nhân và một tiêu thức kết quả. Các tiêu thức nguyên nhân được chọn là các tiêu thức có ảnh hưởng lớn đến tiêu thức kết quả. Để giải quyết vấn đề này đòi hỏi phải có sự phân tích một cách sâu sắc bản chất của mối liên hệ trong điều kiện lịch sử cụ thể. Đây là vấn đề trước tiên quyết định sự thành công của nghiên cứu hồi quy.

Từ đó có thể xây dựng mô hình hồi quy giữa một tiêu thức nguyên nhân và một tiêu thức kết quả và được gọi mô hình hồi quy đơn. Mô hình hồi quy đơn có thể là mô hình tuyến tính (mô hình đường thẳng) hoặc mô hình phi tuyến tính (mô hình đường cong). Việc xác định dạng cụ thể mô hình hồi quy đơn có thể dựa vào đồ thị kết hợp với kinh nghiệm nghiên cứu.

Hoặc có thể xây dựng mô hình hồi quy giữa hai, ba, v.v... tiêu thức nguyên nhân và một tiêu thức kết quả. Mô hình này thường được xây dựng dưới dạng tuyến tính và được gọi là mô hình hồi quy tuyến tính bội.

Các bước tiến hành để giải quyết nhiệm vụ thứ nhất như sau:

- Giải thích sự tồn tại thực tế và bản chất của mối liên hệ bằng phân tích lý luận (đặt vấn đề).

- Thăm dò (mô tả) mối liên hệ bằng các phương pháp thống kê.

- Xác định phương trình hồi quy.

- Giải thích ý nghĩa của các tham số.

1.2.2. *Đánh giá mức độ chặt chẽ của mối liên hệ tương quan*

Việc đánh giá mức độ chặt chẽ của mối liên hệ tương quan được thực hiện thông qua việc tính toán hệ số tương quan, tỷ số tương quan, hệ số tương quan bội, hệ số tương quan riêng phần. Dựa vào kết quả tính toán có thể kết luận về mức độ chặt chẽ của mối liên hệ, giúp cho việc nhận thức hiện tượng được sâu sắc, từ đó đề ra những giải pháp cụ thể.

1.2.3. Đánh giá sự phù hợp của mô hình

Đánh giá sự phù hợp của mô hình qua hệ số xác định. Đồng thời còn giúp ta quyết định xem có thể sử dụng mô hình đã có để dự đoán hay không.

1.3. Ý nghĩa phân tích hồi quy và tương quan

- Phương pháp phân tích hồi quy và tương quan là phương pháp thường được sử dụng trong thống kê để nghiên cứu mối liên hệ giữa các hiện tượng, như mối liên hệ giữa các yếu tố đầu vào của quá trình sản xuất với kết quả sản xuất, mối liên hệ giữa thu nhập và tiêu dùng, mối liên hệ giữa phát triển kinh tế và phát triển xã hội, v.v…

- Phương pháp phân tích hồi quy và tương quan còn được vận dụng trong một số phương pháp nghiên cứu thống kê khác như phân tích dãy số thời gian, dự đoán thống kê, v.v…

2. Xác định mô hình hồi quy tuyến tính đơn

(Liên hệ tương quan tuyến tính giữa hai tiêu thức số lượng)

2.1. Giải thích mô hình tuyến tính

- Mô hình hồi quy tuyến tính của tổng thể chung.

$$Y_{\mu x} = \beta_0 + \beta_1 X_i$$

- Mô hình tuyến tính của tổng thể mẫu.

$$\hat{Y}_i = b_0 + b_1 X_i$$

Trong đó: b_0 là hệ số chặn của Y được dùng để ước lượng β_0

b_1 là độ dốc (hệ số hồi quy) dùng để ước lượng β_1

2.2. Tính toán các tham số

- Dùng phương pháp bình phương nhỏ nhất: Tối thiểu hoá tổng bình phương các độ lệch giữa giá trị thực tế và giá trị dự đoán của biến phụ thuộc.

$$\sum (y_i - \hat{y})^2 = \min$$

- Tính toán các tham số:

+ Giải hệ phương trình:

$$\begin{cases} \sum y = b_0.n + b_1.\sum x \\ \sum xy = b_0.\sum x + b_1.\sum x^2 \end{cases}$$

+ Tính trực tiếp:

$$b_1 = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sigma_x^2}$$

$$b_0 = \overline{y} - b_1.\overline{x}$$

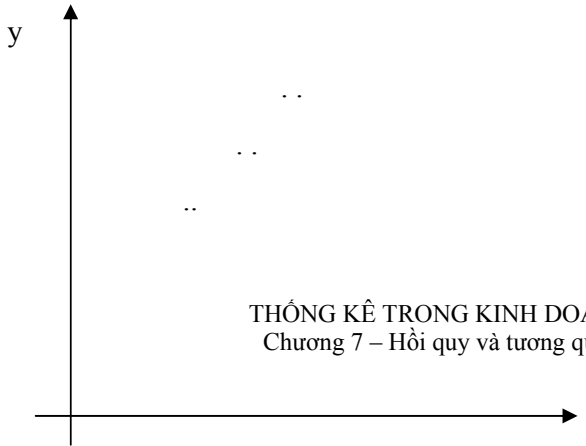
Thí dụ: Có tài liệu về số lao động và giá trị sản xuất (GO) của mười doanh nghiệp công nghiệp như sau:

Lao động (người)	GO (tỷ đ)
60	9,25
78	8,73
90	10,62
115	13,64
126	10,93
169	14,31
198	22,10
226	19,17
250	25,20
300	27,50

Trong mối liên hệ giữa số lượng lao động và giá trị sản xuất thì số lượng lao động là tiêu thức nguyên nhân - ký hiệu là x, giá trị sản xuất là tiêu thức kết quả - ký hiệu là y.

Tài liệu trên cho thấy nhìn chung cùng với sự tăng lên của số lượng lao động thì giá trị sản xuất cũng tăng lên, nhưng cũng có trường hợp không hẳn như vậy - như doanh nghiệp thứ hai so với doanh nghiệp thứ nhất : số lao động nhiều hơn nhưng giá trị sản xuất lại thấp hơn. Điều này chứng tỏ giữa số lượng lao động và giá trị sản xuất có mối liên hệ không hoàn toàn chặt chẽ - tức là liên hệ tương quan.

Có thể dùng đồ thị để biểu hiện mối liên hệ trên với trục hoành là số lao động (x), trục tung là giá trị sản xuất (y) như sau :



. .

. .

0 x

Trên đồ thị có mười chấm, mỗi chấm biểu hiện số lao động và giá trị sản xuất của từng doanh nghiệp. Các chấm trên đồ thị tạo thành một băng đường thẳng, từ đó có thể xây dựng mô hình hồi quy tuyến tính như sau :

$$\hat{y}_x = b_0 + b_1 x$$

Trong đó :

$\hat{y}_x$ là giá trị của tiêu thức kết quả được tính từ mô hình hồi quy.

b_0 là hệ số tự do , phản ảnh $\hat{y}_x$ không phụ thuộc vào x .

b_1 là hệ số góc , phản ảnh sự thay đổi của $\hat{y}_x$ khi x tăng một đơn vị.

Để tìm b_0 và b_1 cần tính Σx , Σy , Σxy , Σx^2 bằng cách lập bảng sau :

x	y	xy	x^2	y^2
60	9,25	555,00	3600	85,5625
78	8,73	680,94	6084	76,2129
90	10,62	955,80	8100	112,7844
115	13,64	1568,60	13225	186,0496
126	10,93	1377,18	15876	119,4649
169	14,31	2418,39	28561	204,7761
198	22,10	4375,80	39204	488,4100
226	19,17	4332,42	51076	367,4889
250	25,20	6300,00	62500	635,0400
300	27,50	8250,00	90000	756,2500
Σx=1612	Σy=161,45	Σxy= 30814,13	Σx^2 = 318226	Σy^2 = 3032,039

Thay số liệu vào hệ phương trình trên :

$$161,45 = 10\, b_0 + 1612\, b_1$$
$$30814,13 = 1612\, b_0 + 318226\, b_1$$

Giải hệ phương trình, sẽ được:

$$b_0 = 2,927\,, \qquad b_1 = 0,082$$

Mô hình hồi quy tuyến tính phản ảnh mối liên hệ giữa số lượng lao động và giá trị sản xuất là :

$$\hat{y}_x = 2,927 + 0,082\, x$$

Hoặc có thể tính b_0 và b_1 theo công thức :

$$b_1 = \frac{\overline{xy} - \overline{x} * \overline{y}}{\sigma_x^2}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

Với : $\overline{xy} = (\Sigma xy)/n = 30814,13/10 = 3081,413$

$\overline{x} = (\Sigma x)/n = 1612/10 = 161,2$

$\overline{y} = (\Sigma y)/n = 161,45/10 = 16,145$

$\sigma_x^2 = \overline{x}^2 - (\overline{x})^2 = (318226/10) - 161,2^2 = 5837,16$

Từ đó tính được :

$$b_1 = \frac{3081,413 - 161,2 * 16,145}{5837,16} = 0,082$$

$b_0 = 16,145 - 0,082 * 161,2 = 2,927$

Thí dụ trên đây nhằm trình bày phương pháp xây dựng mô hình hồi quy nên số lượng đơn vị được nghiên cứu không nhiều. Trong thực tế, số lượng đơn vị được nghiên cứu có thể hàng trăm đơn vị, khi đó các chấm trên đồ thị sẽ rất nhiều và tạo thành như một " đám mây ". Nhiều kinh nghiệm nghiên cứu cho thấy nếu "đám mây" có dạng hình elíp hoặc hình bình hành thì có thể xây dựng mô hình hồi quy tuyến tính.

2.3. Giải thích ý nghĩa các tham số

- Tham số b_0 : Phản ánh ảnh hưởng của tất cả các nhân tố khác ngoài nhân tố đang nghiên cứu tới biến kết quả.

- Tham số b_1 : Phản ánh ảnh hưởng của nhân tố đang nghiên cứu tới biến kết quả. Cụ thể mỗi khi biến giải thích thay đổi (tăng lên) 1 đơn vị thì biến kết quả thay đổi (tăng lên) b_1 đơn vị.

Trở lại thí dụ trên:

b_0=2,927 , nói lên các nguyên nhân khác, ngoài x, ảnh hưởng đến GO.

b_1 = 0,082 , nói lên khi thêm một lao động thì GO tăng bình quân 0,082 tỷ đồng

2.4. Kiểm định hệ số hồi quy

- Dùng tiêu chuẩn kiểm định T-Student để kiểm định hệ số hồi quy β_1 với ý nghĩa "liệu thực sự có mối liên hệ tuyến tính giữa x và y hay không?".
- Cặp giả thiết không và giả thiết đối là:

  $H_0 : \beta_1 = 0$ (không có mối liên hệ tuyến tính)

  $H_1 : \beta_1 \neq 0$ (có mối liên hệ tuyến tính)

  - Tiêu chuẩn kiểm định:

$$t = \frac{b_1 - \beta_1}{S_{b1}} \qquad \text{Trong đó: } S_{b1} = \frac{S_{yx}}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})}}$$

Với n-2 bậc tự do.

2.5. Sai số chuẩn của mô hình

Dùng trong dự đoán các giá trị tương lai.

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

(Độ lệch chuẩn của sự biến thiên của các quan sát xung quanh đường hồi quy).

3. Đánh giá cường độ của mối liên hệ, sự phù hợp của mô hình

3.1. Đánh giá cường độ của mối liên hệ

Hệ số tương quan: là chỉ tiêu đánh giá trình độ chặt chẽ của mối liên hệ tương quan tuyến tính giữa hai tiêu thức.

- Công thức:

$$r = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sigma_x.\sigma_y}$$

$$r = b\frac{\sigma_x}{\sigma_y}$$

Từ thí dụ trên:

$$r = \frac{3081,413 - 161,2 * 16,145}{\sqrt{5837,16 * 42,54}} = 0,961$$

Hoặc:

$$r = 0,082 \frac{\sqrt{5837,16}}{\sqrt{42,54}} = 0,961$$

- Tính chất của hệ số tương quan:
  - r nằm trong khoảng $[-1;1]$, tức là : $-1 \le r \le 1$. Cụ thể:
  - nếu r = 1 (hoặc r = -1) : giữa x và y có mối liên hệ hàm số.
  - nếu r = 0 : giữa x và y không có mối liên hệ tương quan tuyến tính.
  - nếu $r \to 1$ (hoặc $r \to -1$) giữa x và y có mối liên hệ càng chặt chẽ.
  - nếu r dương : giữa x và y có mối liên hệ thuận, nếu r âm : giữa x và y có mối liên hệ nghịch.

Trong ví dụ trên, r = 0,961 cho thấy : mối liên hệ giữa số lượng lao động và giá trị sản xuất rất chặt chẽ và đây là mối liên hệ thuận.

3.2. Đánh giá sự phù hợp của mô hình

Dùng hệ số xác định: r^2

Phản ánh tỷ lệ % sự thay đổi của Y được giải thích bởi mô hình (hay bởi sự thay đổi của X).

4. Ước lượng giá trị trong tương lai dựa vào mô hình hồi quy

4.1. Khoảng tin cậy của dự đoán

- Ước lượng khoảng tin cậy cho μ_{yx} (trung bình của tổng thể chung với một giá trị cá biệt X_i nào đó):

$$\hat{Y}_i \pm t_{\alpha/2;n-2} \cdot S_{yx} \cdot \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- Ước lượng khoảng tin cậy cho từng giá trị riêng biệt của Y với mỗi giá trị cá biệt Xi :

$$\hat{Y}_i \pm t_{\alpha/2;n-2} \cdot S_{yx} \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

4.2. Các nhân tố ảnh hưởng đến khoảng tin cậy

- Độ tin cậy (1 - α)

- Quy mô mẫu.

5. Mô hình hồi quy bội

- Mô hình hồi quy bội biểu diễn mối liên hệ giữa một biến phụ thuộc (biến kết quả) với hai hay nhiều biến độc lập (hay biến giải thích, biến nguyên nhân) bằng một hàm tuyến tính.

- Mô hình hồi quy bội của tổng thể chung

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

- Mô hình hồi quy bội của tổng thể mẫu

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_p X_{pi}$$

- Các tham số của mô hình cũng được xác định bằng phương pháp bình phương nhỏ nhất (SPSS)

- Có thể dùng mô hình để dự đoán giá trị của biến phụ thuộc khi biết các giá trị trong tương lai của các biến độc lập.

- Hệ số xác định bội r^2 (= SSR/ SST) cũng dùng để đánh giá sự phù hợp của mô hình.

- Kiểm định mô hình:

- Kiểm định mức ý nghĩa chung: Dùng tiêu chuẩn kiểm định F để kiểm định sự phụ thuộc của Y với tất cả các biến độc lập X_i.

  Cặp giả thiết:

  $H_0 : \beta_1 = \beta_2 = \bullet\bullet\bullet = \beta_p = 0$ (không có mối liên hệ tuyến tính)

  H_1 : Tồn tại ít nhất một $\beta_i \neq 0$ (có ít nhất một biến độc lập ảnh hưởng tới Y)

- Kiểm định các hệ số hồi quy: Dùng kiểm định t để xem liệu có mối liên hệ thực sự giữa từng biến độc lập X_i với biến phụ thuộc Y hay không.

  Cặp giả thiết:

  $H_0 : \beta_i = 0$ (không có mối liên hệ tuyến tính)

  $H_1 : \beta_i \neq 0$ (có mối liên hệ tuyến tính giữa X_i và Y)

- ***Xây dựng mô hình:***

- Lựa chọn biến giải thích (nguyên nhân): Vấn đề đặt ra khi xây dựng mô hình hồi quy bội là chọn bao bao nhiêu biến giải thích. Về lý thuyết, có thể nói rằng: nếu số biến giải thích được chọn ra càng nhiều thì càng phản ánh một cách đầy đủ mối liên hệ, song việc thu thập tài liệu và tính toán càng trở nên phức tạp. Do vậy chỉ nên chọn những biến có tác

động lớn, dễ giải thích và không hoặc ít có liên hệ với nhau (tránh hiện tượng đa cộng tuyến).

- Dùng phương pháp hồi quy từng bước (stepwise) để lựa chọn mô hình tốt nhất (SPSS).
- Lựa chọn mô hình tốt nhất: Là mô hình có r^2 lớn nhất, và sai số của mô hình nhỏ nhất.

BÀI TẬP

7.1. Mô hình thống kê là gì?

7.2. Các bước xây dựng mô hình?

7.3. Các giả thiết của mô hình hồi quy tuyến tính đơn? (simple linear regression)

7.4. Xác định các tham số của mô hình hồi quy tuyến tính đơn?

7.5. Các tác dụng của mô hình hồi quy

7.6. Mục đích và ý nghĩa của sai số trong hồi quy?

7.7. Đưa ra ví dụ về các tình huống kinh doanh mà bạn cho rằng có một mối quan hệ đường thẳng giữa hai biến số. Tác dụng của mô hình hồi quy trong từng trường hợp là gì?

7.8. Hãy giải thích những ưu điểm của phương pháp bình phương nhỏ nhất? Cho biết cách thực hiện.

7.9. Gần đây, một nhóm nghiên cứu đã tập trung vào vấn đề dự đoán thị phần của nhà sản xuất bằng cách sử dụng thông tin về chất lượng sản phẩm của họ. Giả sử rằng các số liệu sau là thị phần đã có tính theo đơn vị phần trăm (%) (Y) và chất lượng sản phẩm theo thang điểm 0-100 được xác định bởi một quy trình định giá khách quan (X).

X: 27, 39, 73, 66, 33, 43, 47, 55, 60, 68, 70, 75, 82.

Y: 2, 3, 10, 9, 4, 6, 5, 8, 7, 9, 10, 13, 12.

Hãy ước lượng mối quan hệ hồi quy tuyến tính đơn giữa thị phần và chất lượng sản phẩm.

7.10. Số liệu sau so sánh chỉ số Standard & Poor 500 và tỷ giá đồng đô la Mỹ so với Mark Đức từ tháng 12/1995 đến tháng 6/1997. Có mối quan hệ tuyến tính giữa hai biến số hay không? Bạn có thể nói rằng một biến là nguyên nhân của biến kia hay không?

Tháng	Chỉ số Standard & Poor 500	Tỷ giá đồng đô la Mỹ so với Mark Đức
12/95	610	110
1/96	620	111
2	660	109
3	640	109
4	640	108
5	670	107
6	665	107
7	640	107
8	670	108
9	690	107
10	725	108
11	745	107
12	740	105

1/97	760	104
2	785	104
3	810	104
4	760	104
5	840	103
6	900	102

Hãy tính toán hệ số tương quan.

7.11. Đối với tình huống trong bài tập 6.9, tìm sai số chuẩn (standard errors) của các ước lượng trong các tham số hồi quy (regression parameters). Đồng thời đưa ra khoảng tin cậy 95% cho hệ số hồi quy thực tế. Liệu rằng 0 có phải là giá trị đáng tin cậy cho hệ số góc hồi quy thực tế (true regression slope) với độ tin cậy 95% hay không?

7.12. Sự khác nhau cơ bản giữa phân tích tương quan (correlation) và phân tích hồi quy là gì?

7.13. Dữ liệu sau là giá so sánh của vàng và đồng trong giai đoạn 10 năm. Giả sử giá so sánh này thu được từ một mẫu ngẫu nhiên của một tổng thể các giá trị có thể có. Kiểm định sự tồn tại của tương quan tuyến tính giữa hai giá trị so sánh của hai kim loại này.

Vàng: 76, 62, 70, 59, 52, 53, 53, 56, 57, 56

Đồng: 80, 68, 73, 63, 65, 68, 65, 63, 65, 66

7.14. Một phân tích hồi quy giữa hiệu quả sử dụng nhiên liệu (X) và doanh thu bán các loại máy bay khác nhau (Y) của một công ty chứa đựng các kết quả sau: $b_1 = 2.435$, $s(b_1) = 1.567$ và $n = 12$. Bạn có cho rằng tồn tại mối quan hệ tuyết tính giữa doanh số bán máy bay của công ty và hiệu quả sự dụng nhiên liệu của máy bay?

7.15. Với tình huống trong bài 6.9, hãy kiểm định sự tồn tại của mối quan hệ tuyến tính giữa hai biến số.

7.16. Kết quả một cuộc nghiên cứu được đăng trên tạp chí Phân tích Tài chính bao gồm một phân tích hồi quy tuyến tính đơn giữa mức chi cho quỹ hưu trí (Y) và lợi nhuận của doanh nghiệp. ***Hệ số xác định*** là $r^2 = 0.02$. (kích thước mẫu là 515)

a. Bạn có sử dụng mô hình hồi quy để dự báo mức chi cho quỹ hưu trí hay không?

b. Mô hình có giải thích nhiều lắm sự biến đổi của mức chi cho quỹ hưu trí theo mức lợi nhuận hay không?

c. Theo bạn, kết quả hồi quy đó có đủ giá trị để báo cáo hay không? Giải thích?

7.17. Trong vài năm gần đây, Mita, một nhà sản xuất máy copy đã chi thêm một khoản tiền vào việc quảng cáo trên đài và truyền hình. Một nhà phân tích của công ty Mita muốn ước lượng hồi quy tuyến tính đơn giữa doanh số bán máy copy với chi phí quảng cáo. Kết quả hồi quy bao gồm: SSE = 12,745 và SSR = 87,691.

Xác định coefficient of determination của hồi quy này. Bạn có cho rằng mô hình này có thể là một công cụ hữu ích để dự đoán doanh số dựa trên chi phí quảng cáo? Giải thích?

7.18. Một người muốn xem xét ảnh hưởng của diện tích (feet vuông) và khoảng cách từ trung tâm thành phố (dặm) tới giá trị của các ngôi nhà (ngàn đô la) tại một vùng nhất định. 9 ngôi nhà được lựa chọn ngẫu nhiên và dữ liệu thu được như sau:

Y (giá trị): 345, 238, 452, 422, 328, 375, 660, 466, 290

X_1 (diện tích): 1650, 1870, 2230, 1740, 1900, 2000, 3200, 1860, 1230

X_2 (khoảng cách): 3.5, 0.5, 1.5, 4.5, 1.8, 0.1, 3.4, 3.0, 1.0

Tính toán các ước lượng của các hệ số hồi quy và giải thích ý nghĩa của chúng.

CHƯƠNG 8

PHÂN TÍCH DÃY SỐ THỜI GIAN

Mặt lượng của hiện tượng thường xuyên biến động qua thời gian, việc nghiên cứu sự biến động này được thực hiện trên cơ sở phân tích dãy số thời gian. Qua dãy số thời gian có thể phân tích đặc điểm biến động của hiện tượng qua thời gian, phân tích tính quy luật của sự phát triển hiện tượng bằng các mô hình. Trên cơ sở nhận thức đặc điểm và tính quy luật biến động của hiện tượng có thể thực hiện các dự đoán cho mức độ của hiện tượng trong tương lai. Có rất nhiều các phương pháp phân tích và dự đoán khác nhau được sử dụng với dãy số thời gian, trong phạm vi chương này đề cập đến một số phương pháp cơ bản, phổ biến, hiệu quả và được trình bày thành các nội dung sau :

- Khái niệm chung về dãy số thời gian
- Phân tích đặc điểm biến động của hiện tượng qua thời gian
- Phân tích các thành phần của dãy số thời gian
- Dự đoán dựa trên cơ sở dãy số thời gian.

1. Khái niệm chung về dãy số thời gian.

1.1 Khái niệm dãy số thời gian

Dãy số thời gian là dãy các số liệu thống kê của hiện tượng nghiên cứu được sắp xếp theo thứ tự thời gian .

Thí dụ 1: Có tài liệu về giá trị sản xuất (GO) của doanh nghiệp A qua một số năm như sau. Dãy số thời gian này phản ánh GO của doanh nghiệp từ năm 2003 đến năm 2007:

Năm	2002	2003	2004	2005	2006	2007
GO (tỷ đồng)	10,0	12,5	15,4	17,6	20,2	22,9

Thí dụ 2: Có tài liệu về giá trị hàng hóa tồn kho của cửa hàng B vào các ngày đầu của 4 tháng đầu năm 2007 như sau:

Thời gian	1 - 1	1 - 2	1 - 3	1 - 4
Lượng hàng hoá tồn kho (Trđ)	356	364	370	352

Dãy số trên phản ánh giá trị hàng hóa tồn kho tại ngày đầu mỗi tháng năm 2007, các ngày khác trong tháng thì giá trị hàng hoá tồn kho có thể thay đổi do việc xuất, nhập hàng hoá thường xảy ra trong quá trình kinh doanh.

Qua hai thí dụ trên cho thấy một dãy số thời gian gồm hai thành phần: Thời gian và chỉ tiêu về hiện tượng nghiên cứu.

- Thời gian có thể là ngày, tuần, tháng, quý, năm… Độ dài giữa hai thời gian liền nhau gọi là khoảng cách thời gian. Dãy số thời gian ở trên có khoảng cách thời gian là một năm.

- Chỉ tiêu về hiện tượng nghiên cứu gồm tên chỉ tiêu và trị số của chỉ tiêu với đơn vị tính thích hợp. Các trị số của chỉ tiêu có thể được biểu hiện bằng số tuyệt đối, số tương đối , số bình quân và được gọi là các mức độ của dãy số (y_1, y_2, …. Y_n)

1.2 Các loại dãy số thời gian

Tùy theo hình thức biểu hiện của các mức độ trong dãy số thời gian mà có thể phân loại như sau:

- ***Dãy số tuyệt đối***: Là dãy mà các mức độ được biểu hiện bằng số tuyệt đối. Tuỳ theo ý nghĩa phản ánh của các mức độ mà dãy số tuyệt đối được chia ra làm hai loại:

+ *Dãy số thời kỳ*: Dãy số thời kỳ là dãy số mà các mức độ là những số tuyệt đối thời kỳ, phản ánh quy mô (khối lượng) của hiện tượng trong từng khoảng thời gian nhất định. Thí dụ 1 ở trên là một dãy số thời kỳ, mỗi mức độ của dãy số phản ánh kết quả sản xuất của doanh nghiệp trong khoảng thời gian từng năm. Từng mức độ của dãy số có sự tích luỹ về lượng qua thời gian do đó có thể cộng dồn các mức độ qua thời gian để có mức độ trong khoảng thời gian dài hơn.

+*Dãy số thời điểm*: Dãy số thời điểm là dãy số mà các mức độ là những số tuyệt đối thời điểm, phản ánh quy mô (khối lượng) của hiện tượng tại những thời điểm nhất định hay nó phản ánh trạng thái của hiện tượng tại thời điểm đó (thí dụ 2). Các mức độ của dãy số thời điểm không phải là sự cộng dồn của các mức độ trước đó (sẽ không có ý nghĩa nếu cộng các mức độ liền nhau).

- ***Dãy số tương đối***: Dãy số mà các mức độ biểu hiện bằng số tương đối. Chẳng hạn dãy số của chỉ tiêu tốc độ phát triển doanh thu của một doanh nghiệp hoặc cơ cấu kinh tế thay đổi theo thời gian,...

- ***Dãy số bình quân***: Là dãy số mà các mức độ của nó biểu hiện bằng số bình quân. Chẳng hạn dãy số của chỉ tiêu năng suất lao động qua thời gian, thu nhập bình quân đầu người….

1.3. Tác dụng của dãy số thời gian

- Cho phép thống kê phân tích và nhận thức được các *đặc điểm* về sự biến động của hiện tượng qua thời gian

- Cho phép nhận thức về *xu hướng và tính quy luật* của sự phát triển hiện tượng, trong đó bao gồm cả việc phân tích các thành phần của dãy số thời gian.

- Dựa trên cơ sở những phân tích đặc điểm và tính quy luật ở trên có thể *dự đoán các mức độ* của hiện tượng trong tương lai (trong thống kê gọi là dự đoán có điều kiện).

1.4. Yêu cầu chung khi xây dựng dãy số thời gian

Để phân tích dãy số thời gian được chính xác thì yêu cầu cơ bản khi xây dựng dãy số thời gian là phải đảm bảo tính chất có thể so sánh được giữa các mức độ trong dãy số. Cụ thể :

- Nội dung và phương pháp tính chỉ tiêu qua thời gian phải thống nhất.

- Phạm vi hiện tượng nghiên cứu qua thời gian phải nhất trí.

- Các khoảng cách thời gian trong dãy số nên bằng nhau, nhất là đối với dãy số thời kỳ thì phải bằng nhau.

Trong thực tế, do những nguyên nhân khác nhau, các yêu cầu trên có thể bị vi phạm, khi đó đòi hỏi có sự chỉnh lý phù hợp để tiến hành phân tích.

2. Phân tích đặc điểm biến động của hiện tượng qua thời gian

Các chỉ tiêu sau đây thường được sử dụng để phân tích những đặc điểm biến động của hiện tượng qua thời gian.

2.1. Mức độ bình quân theo thời gian:

Chỉ tiêu này phản ánh mức độ đại diện cho các mức độ tuyệt đối của dãy số thời gian. Tuỳ theo dãy số thời kỳ hay dãy số thời điểm mà công thức tính khác nhau.

- *Đối với dãy số thời kỳ*, mức độ bình quân qua thời gian được tính theo công thức sau đây :

$$\bar{y} = \frac{y_1 + y_2 + ... + y_{n-1} + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Trong đó y_i (i = 1, 2, ... , n) là các mức độ của dãy số thời kỳ .

Từ thí dụ 1, ta có :

$$\bar{y} = \frac{10,0 + 12,5 + 15,4 + 17,6 + 20,2 + 22,9}{6} = 16,433 \text{ tỷ đồng}$$

Như vậy, giá trị sản xuất bình quân hàng năm của doanh nghiệp từ 2002 đến 2007 đạt 16,433 tỷ đồng.

- *Đối với dãy số thời điểm:* Có 3 trường hợp

  + *Trường hợp dãy số biến đổi tương đối đều đặn*: áp dụng khi biến động của các mức độ trong dãy số thời điểm là tương đối ***đồng đều*** và có số liệu ở đầu kỳ và cuối kỳ.

$$\bar{y} = \frac{y_{DK} + y_{CK}}{2}$$

  + *Trường hợp khoảng cách thời gian bằng nhau*: áp dụng khi biến động của các mức độ trong dãy số thời điểm là ***không đồng đều*** và có số liệu tại các thời điểm có khoảng cách thời gian bằng nhau.

Trở lại thí dụ 2 ở trên, để tính giá trị hàng hoá tồn kho bình quân của từng tháng, cần phải giả thiết: sự biến động về giá trị hàng hoá tồn kho của các ngày trong tháng xảy ra tương đối đều đặn. Từ đó, dựa vào giá trị hàng hoá tồn kho của ngày đầu tháng và ngày cuối tháng - tức của đầu tháng sau, để tính giá trị hàng hoá tồn kho bình quân của tháng . Giá trị hàng hoá tồn kho bình quân của từng tháng được tính như sau :

Tháng 1- 2007 : $\bar{y}_1 = \frac{356+364}{2} = 360$ triệu đồng

Tháng 2- 2007 : $\bar{y}_2 = \frac{364+370}{2} = 367$ triệu đồng

Tháng 3 - 2007 : $\bar{y}_3 = \frac{370+352}{2} = 361$ triệu đồng

Giá trị hàng hoá tồn kho bình quân của quý I năm 2004 (ký hiệu $\bar{y}_I$) tính được bằng cách bình quân cộng giá trị hàng hoá tồn kho bình quân của tháng1, tháng 2, tháng 3 năm 2007 . Tức là :

$$\bar{y}_I = \frac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3}{3} = \frac{360+367+361}{3} = \frac{\frac{356}{2} + 364 + 370 + \frac{352}{2}}{4-1}$$

$$= 362{,}666 \text{ triệu đồng.}$$

Từ đó, công thức để tính mức độ bình quân qua thời gian từ dãy số thời điểm có các khoảng cách thời gian bằng nhau là:

$$\bar{y} = \frac{\frac{y_1}{2} + y_2 + ... + y_{n-1} + \frac{y_n}{2}}{n-1}$$

+ *Trường hợp khoảng cách thời gian không bằng nhau*: sử dụng khi biến động của các mức độ trong dãy số là ***không đồng đều*** và ***khoảng cách thời gian không bằng nhau***. Tính theo công thức bình quân cộng gia quyền:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i}$$

Trong đó:

y_i – các mức độ của dãy số thời gian

t_i - độ dài thời gian có các mức độ y_i tương ứng.

Thí dụ 3: Có tài liệu về số lượng lao động của một doanh nghiệp trong tháng 4/2007 như sau:

Ngày 1- 4 có 400 người

Ngày 10 - 4 nhận thêm 5 người

Ngày 15 - 4 nhận thêm 3 người

Ngày 21 - 4 cho thôi việc 2 người và từ đó cho đến hết tháng 4 năm 2007 số lao động không thay đổi .

Yêu cầu tính số lao động bình quân của tháng 4 - 2004. Bảng sau đây được lập ra để tính toán :

Thời gian	y_i	t_i	$y_i t_i$
- Ngày 1/4 có 400 công nhân	400	9	3600
- Ngày 10/4 thêm 3 người	403	5	2015
- Ngày 15/4 thêm 2 người	405	6	2430
- Ngày 21/4 thôi việc 4 người	401	10	4010
		30	**12055**

$$\bar{y} = \frac{\sum y_i t_i}{\sum t_i} = \frac{12055}{30} = 401.8(ng)$$

2.2. Lượng tăng (giảm) tuyệt đối

Chỉ tiêu này phản ánh sự biến động về mức độ tuyệt đối giữa hai thời gian. Tuỳ theo mục đích nghiên cứu, có thể tính các chỉ tiêu về lượng tăng (giảm) tuyệt đối sau đây :

- *Lượng tăng (giảm) tuyệt đối liên hoàn*: Phản ánh sự biến động về mức độ tuyệt đối giữa hai thời gian liền nhau và được tính theo công thức sau đây :

$$\delta_i = y_i - y_{i-1} \quad (\text{với } i = 2, 3, \ldots, n)$$

Trong đó :

δ_i : Lượng tăng (hoặc giảm) tuyệt đối liên hoàn ở thời gian i so với thời gian đứng liền trước đó là $i-1$

y_i : Mức độ tuyệt đối ở thời gian i

y_{i-1} : Mức độ tuyệt đối ở thời gian $i-1$

Nếu $y_i > y_{i-1}$ thì $\delta_i > 0$: phản ánh quy mô hiện tượng tăng , ngược lại nếu $y_i < y_{i-1}$ thì $\delta_i < 0$: phản ánh quy mô hiện tượng giảm .

Từ số liệu ở thí dụ 1, ta có:

$\delta_2 = y_2 - y_1$ = 12,5 tỷ đồng - 10,0 tỷ đồng = 2,5 tỷ đồng

$\delta_3 = y_3 - y_2$ = 15,4 tỷ đồng - 12,5 tỷ đồng = 2,9 tỷ đồng

$\delta_4 = y_4 - y_3$ = 17,6 tỷ đồng - 15,4 tỷ đồng = 2,2 tỷ đồng

$\delta_5 = y_5 - y_4$ = 20,2 tỷ đồng - 17,6 tỷ đồng = 2,6 tỷ đồng

$\delta_6 = y_6 - y_5 = 22,9$ tỷ đồng $- 20,2$ tỷ đồng $= 2,7$ tỷ đồng

Như vậy, năm sau so với năm trước giá trị sản xuất của doanh nghiệp đều tăng lên.

- *Lượng tăng (giảm) tuyệt đối định gốc* : Phản ánh sự biến động về mức độ tuyệt đối trong những khoảng thời gian dài và được tính theo công thức sau đây :

$$\Delta_i = y_i - y_1 \qquad (\text{với } i = 2, 3, \ldots, n)$$

Trong đó : Δ_i : Lượng tăng (giảm) tuyệt đối định gốc ở thời gian i so với thời gian đầu của dãy số .

y_i : Mức độ tuyệt đối ở thời gian i .

y_1 : Mức độ tuyệt đối ở thời gian đầu.

Từ số liệu ở bảng 1 :

$\Delta_2 = y_2 - y_1 = 12,5 - 10,0 = 2,5$ (tỷ đồng)

$\Delta_3 = y_3 - y_1 = 15,4 - 10,0 = 5,4$ (tỷ đồng)

$\Delta_4 = y_4 - y_1 = 17,6 - 10,0 = 7,6$ (tỷ đồng)

$\Delta_5 = y_5 - y_1 = 20,2 - 10,0 = 10,2$ (tỷ đồng)

$\Delta_6 = y_6 - y_5 = 22,9 - 10,0 = 12,9$ (tỷ đồng)

Dễ dàng nhận thấy mối liên hệ:

$$\Delta_i = \sum \delta_i \; ; \; \Delta_n = \sum_{i=2}^{n} \delta_i = y_n - y_1$$

Từ thí dụ trên : $2,5 + 2,9 + 2,2 + 2,6 + 2,7 = 12,9$ (tỷ đồng)

- *Lượng tăng (giảm) tuyệt đối bình quân* : Phản ánh mức độ đại diện của các lượng tăng (giảm) tuyệt đối liên hoàn và được tính theo công thức sau đây:

$$\bar{\delta} = \frac{\sum_{i=2}^{n} \delta_i}{n-1} = \frac{\Delta_n}{n-1} = \frac{y_n - y_1}{n-1}$$

Trong thí dụ trên:

$$\bar{\delta} = \frac{22,9 - 10,0}{6-1} = 2,58 \text{ tỷ đồng}$$

Tức là: trong giai đoạn từ năm 2002 đến năm 2007 , giá trị sản xuất của doanh nghiệp hàng năm đã tăng bình quân là 2,58 tỷ đồng.

2.3. Tốc độ phát triển

Chỉ tiêu này phản ánh tốc độ và xu hướng biến động của hiện tượng nghiên cứu qua thời gian. Tuỳ theo mục đích nghiên cứu, có thể tính các tốc độ phát triển sau đây :

- *Tốc độ phát triển liên hoàn* : Phản ánh tốc độ và xu hướng biến động của hiện tượng ở thời gian sau so với thời gian liền trước đó và được tính theo công thức sau đây :

$$t_i = \frac{y_i}{y_{i-1}} \cdot 100 \qquad (\text{với } i = 2, 3, \ldots, n)$$

Trong đó : t_i là tốc độ phát triển liên hoàn thời gian i so với thời gian i -1 và có thể biểu hiện bằng lần hoặc % .

Từ thí dụ 1, ta có :

$$t_2 = \frac{y_2}{y_1} = \frac{12,5}{10,0} = 1,250 \text{ lần hay } 125,0\%$$

$$t_3 = \frac{y_3}{y_2} = \frac{15,4}{12,5} = 1,232 \text{ lần hay } 123,2\%$$

$$t_4 = \frac{y_4}{y_3} = \frac{17,6}{15,4} = 1,143 \text{ lần hay } 114,3\%$$

$$t_5 = \frac{y_5}{y_4} = \frac{20,2}{17,6} = 1,148 \text{ lần hay } 114,8\%$$

$$t_6 = \frac{y_6}{y_5} = \frac{22,9}{20,2} = 1,134 \text{ lần hay } 113,4\%$$

- *Tốc độ phát triển định gốc* : Phản ánh tốc độ và xu hướng biến động của hiện tượng ở thời gian những khoảng thời gian dài và được tính theo công thức sau đây :

$$T_i = \frac{y_i}{y_1} \cdot 100 \qquad (\text{với } i = 2, 3, \ldots, n)$$

Trong đó : T_i : Tốc độ phát triển định gốc thời gian i so với mức độ đầu của dãy số và có thể biểu hiện bằng lần hoặc % .

Từ thí dụ 1, ta có :

$$T_2 = \frac{y_2}{y_1} = \frac{12,5}{10,0} = 1,25 \text{ lần hay } 125\%$$

$$T_3 = \frac{y_3}{y_1} = \frac{15,4}{10,0} = 1,54 \text{ lần hay } 154\%$$

$$T_4 = \frac{y_4}{y_1} = \frac{17,6}{10,0} = 1,76 \text{ lần hay } 176\%$$

$$T_5 = \frac{y_5}{y_1} = \frac{20,2}{10,0} = 2,02 \text{ lần hay } 202\%$$

$$T_6 = \frac{y_6}{y_1} = \frac{22,9}{10,0} = 2,29 \text{ lần hay } 229\%$$

Giữa tốc độ phát triển liên hoàn và tốc độ phát triển định gốc có các mối quan hệ sau đây :

- Thứ nhất: Tích các tốc độ phát triển liên hoàn bằng tốc độ phát triển định gốc, tức là :

$$T_i = \Pi t_i;\ T_n = \prod_{i=2}^{n} t_i$$

- Thứ hai : Thương của tốc độ phát triển định gốc ở thời gian i với tốc độ phát triển định gốc ở thời gian i -1 bằng tốc độ phát triển liên hoàn giữa hai thời gian đó , tức là :

$$\frac{T_i}{T_{i-1}} = t_i \qquad (\text{với } i = 2, 3, \dots, n)$$

- *Tốc độ phát triển bình quân*: Phản ánh mức độ đại diện của các tốc độ phát triển liên hoàn. Từ mối quan hệ thứ nhất giữa các tốc độ phát triển liên hoàn và tốc độ phát triển định gốc nên tốc độ phát triển bình quân được tính theo công thức số bình quân nhân:

$$\bar{t} = \sqrt[n-1]{t_2.t_3...t_n} = \sqrt[n-1]{\prod_{i=2}^{n} t_i} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

Từ thí dụ 1, ta có :

$$\bar{t} = \sqrt[6-1]{\frac{22,9}{10,0}} = \sqrt[5]{2,29} = 1,18 \text{ lần hay } 118\%$$

Tức là: tốc độ phát triển bình quân hàng năm về giá trị sản xuất của doanh nghiệp bằng 1,18 lần hay118% .

Từ công thức tính tốc độ phát triển bình quân cho thấy : chỉ nên tính chỉ tiêu này đối với những hiện tượng biến động theo một xu hướng nhất định .

2.4. *Tốc độ tăng(giảm)*

Chỉ tiêu này phản ánh qua thời gian, hiện tượng đã tăng (giảm) bao nhiêu lần hoặc bao nhiêu phần trăm, hay phản ánh nhịp điệu biến động qua thời gian.Tuỳ theo mục đích nghiên cứu, có thể tính các tốc độ tăng (giảm) sau đây:

- *Tốc độ tăng (giảm) liên hoàn* : phản ánh tốc độ tăng (giảm) ở thời gian i so với thời gian i -1 và được tính theo công thức sau đây :

$$a_i = \frac{y_i - y_{i-1}}{y_{i-1}} \cdot 100 = t_i - 100\%$$

Tức là: Tốc độ tăng (giảm) liên hoàn bằng tốc độ phát triển liên hoàn trừ 1 (biểu hiện bằng lần) và trừ 100% (nếu tốc độ phát triển liên hoàn biểu hiện bằng phần trăm).

Từ các kết quả ở mục 2.3, ta có :

$a_2 = t_2 - 1 = 1,250 - 1 = 0,25$ lần hay 25%

$a_3 = t_3 - 1 = 1,232 - 1 = 0,232$ lần hay 23,2%

v.v ...

- *Tốc độ tăng (giảm) định gốc* : phản ánh tốc độ tăng (giảm) ở thời gian i so với mức độ đầu trong dãy số và được tính theo công thức sau đây :

$$A_i = \frac{y_i - y_1}{y_1} \cdot 100 = T_i - 100\,\%$$

Tức là : Tốc độ tăng (giảm) định gốc bằng tốc độ phát triển định gốc trừ 1 (nếu biểu hiện bằng đơn vị lần) và trừ 100% (nếu tốc độ phát triển liên hoàn biểu hiện bằng phần trăm).

Từ các kết quả ở mục 3, ta có :

$A_2 = T_2 - 1 = 1,25 - 1 = 0,25$ lần hay 25%

$A_3 = T_3 - 1 = 1,54 - 1 = 0,54$ lần hay 54%

v.v...

- *Tốc độ tăng (giảm) bình quân*: Phản ánh tốc độ tăng (giảm) đại diện cho các tốc độ tăng (giảm) liên hoàn và được tính theo công thức sau đây :

$\bar{a} = \bar{t} - 1$ (nếu $\bar{t}$ biểu hiện bằng lần)

Hoặc : $\bar{a} = \bar{t}\,(\%) - 100$ (nếu $\bar{t}$ biểu hiện bằng %)

Từ kết quả mục 2.3. , ta có :

$\bar{a} = 1,18 - 1 = 0,18$ lần hay 18%

Tức là: tốc độ tăng bình quân hàng năm về giá trị sản xuất của doanh nghiệp bằng 18%.

2.5. Giá trị tuyệt đối của 1% tăng (giảm)

Chỉ tiêu này phản ánh cứ 1% tăng (giảm) của tốc độ tăng (giảm) liên hoàn thì tương ứng với một giá trị cụ thể là bao nhiêu và tính được bằng cách chia lượng tăng (giảm) tuyệt đối liên hoàn cho tốc độ tăng (giảm) liên hoàn, tức là :

$$g_i = \frac{\delta_i}{a_i(\%)} = \frac{y_i - y_{i-1}}{\frac{y_i - y_{i-1}}{y_{i-1}} \cdot 100} = \frac{y_{i-1}}{100}$$

Từ thí dụ 1, ta có :

$g_2 = \frac{y_1}{100} = \frac{10,0}{100} = 0,1$ tỷ đồng - tức là cứ 1% tăng lên cuả năm 2003 so với năm 2002 thì tương ứng 0,1 tỷ đồng.

$g_3 = \frac{y_2}{100} = \frac{12,5}{100} = 0,125$ tỷ đồng - tức là cứ 1% tăng lên cuả năm 2004 so với năm 2003 thì tương ứng 0,125 tỷ đồng.

v.v...

Chỉ tiêu này không tính đối với tốc độ tăng (giảm) định gốc vì luôn là một số không đổi và bằng $\frac{y_1}{100}$.

Trên đây là năm chỉ tiêu thường được sử dụng để phân tích đặc điểm biến động của hiện tượng qua thời gian. Mỗi một chỉ tiêu có nội dung và ý nghĩa riêng, song giữa các chỉ tiêu có mối liên hệ với nhau nhằm giúp cho việc phân tích được đầy đủ và sâu sắc.

3. Phân tích các thành phần của dãy số thời gian

Các thành phần của dãy số thời gian được biểu hiện bằng sơ đồ sau :

3.1. Thành phần xu thế

Sự biến động về mặt lượng của hiện tượng qua thời gian chịu sự tác động của nhiều yếu tố và có thể chia thành hai nhóm: nhóm yếu tố chủ yếu và nhóm yếu tố ngẫu nhiên.

Với sự tác động của nhóm các yếu tố chủ yếu sẽ xác lập xu thế (xu hướng) phát triển của hiện tượng. Xu thế phát triển thường được hiểu là chiều hướng tiến triển chung kéo dài theo thời gian, phản ánh tính quy luật của sự phát triển của hiện tượng.

Với sự tác động của nhóm các yếu tố ngẫu nhiên sẽ làm cho sự biến động về mặt lượng của hiện tượng lệch khỏi xu hướng chung. Vì vậy, cần sử dụng những phương pháp phù hợp, trong một chừng mực nhất định, nhằm loại bỏ dần sự tác động của các yếu tố ngẫu nhiên và phản ánh xu thế phát triển của hiện tượng.

Thành phần xu thế có thể có ở tất cả các dãy số với các loại thời gian khác nhau như tháng, quý, năm... Sau đây sẽ đề cập đến một số phương pháp thường được sử dụng để biểu hiện xu thế phát triển của hiện tượng.

3.1.1. Mở rộng khoảng cách thời gian

Phương pháp này được sử dụng đối với dãy số thời kỳ có khoảng cách thời gian tương đối ngắn và có nhiều mức độ mà qua đó chưa phản ánh xu hướng phát triển của hiện tượng.

Nội dung của phương pháp này là ghép một số thời gian liền nhau vào thành khoảng thời gian dài hơn, chẳng hạn ghép 3 tháng vào thành quý gọi là mở rộng khoảng cách từ tháng sang quý.

Thí dụ 4: Có tài liệu về sản lượng hàng tháng năm 2007 của một doanh nghiệp như sau:

Tháng	*Sản lượng (1000 tấn)*	*Tháng*	*Sản lượng (1000 tấn)*
1	40,4	7	40,8
2	36,8	8	44,8
3	40,6	9	49,4
4	38,0	10	48,9
5	42,2	11	46,2
6	48,5	12	42,2

Dãy số thời gian ở trên cho thấy sản lượng của các tháng khi tăng, khi giảm không phản ánh rõ xu hướng biến động. Có thể mở rộng khoảng cách thời gian từ tháng sang quý bằng cách cộng sản lượng của tháng 1, tháng 2 và tháng 3 sẽ được sản lượng của quý I ; cộng sản lượng của tháng 4, tháng 5 và tháng 6 sẽ được sản lượng của quý II v.v.. và được kết quả sau đây :

Qúy	*Sản lượng (1000 tấn)*
I	117,8
II	128,7
III	135,0
IV	137,3

Bảng trên cho thấy sản lượng của doanh nghiệp tăng dần từ quý I đến quý IV năm 2007.

Phương pháp mở rộng khoảng cách thời gian đơn giản, dễ làm nhưng có hạn chế lớn là số lượng các mức độ trong dãy số mất đi quá nhiều. Như vậy đôi khi không chỉ làm mất ảnh hưởng của các nhân tố ngẫu nhiên mà làm giảm đi ảnh hưởng của cả các nhân tố cơ bản đến sự biến động của hiện tượng. Nhất là đối với các dãy số theo tháng của hiện tượng có biến động thời vụ thì không thể vận dụng phương pháp này vì sẽ làm mất tính thời vụ.

3.1.2. Số bình quân trượt (di động)

Đây là phương pháp sử dụng để san bằng dãy số có nhiều biến động ngẫu nhiên. Số bình quân trượt (còn gọi số bình quân di động) là số bình quân cộng của một nhóm

nhất định các mức độ dãy số thời gian tính được bằng cách loại dần các mức độ đầu, đồng thời thêm vào các mức độ tiếp theo, sao cho số lượng các mức độ tham gia tính số bình quân không thay đổi. Số lượng các mức độ tham gia tính số bình quân thường là lẻ (3, 5, 7)

Giả sử có dãy số thời gian : $y_1, y_2, y_3, ..., y_n$

Nếu tính số bình quân trượt cho nhóm ba mức độ, sẽ có các số bình quân trượt như sau:

$$\bar{y}_2 = \frac{y_1 + y_2 + y_3}{3}$$

$$\bar{y}_3 = \frac{y_2 + y_3 + y_4}{3}$$

....

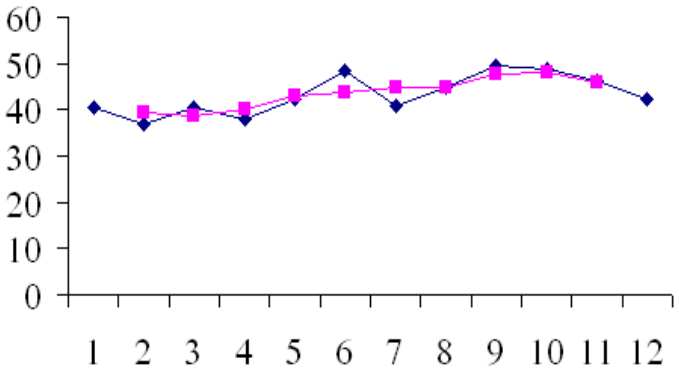
$$\bar{y}_{n-1} = \frac{y_{n-2} + y_{n-1} + y_n}{3}$$

Từ đó, sẽ có dãy số mới gồm các số bình quân trượt $\bar{y}_2, \bar{y}_3, \ldots, \bar{y}_{n-1}$

Từ bảng số liệu ở thí dụ 4, tính số bình quân trượt cho nhóm ba mức độ, sẽ có kết quả sau:

Tháng	y_i	$\bar{y}_i$	***Tháng***	y_i	$\bar{y}_i$
1	40,4	-	7	40,8	44,7
2	36,8	39,3	8	44,8	45,0
3	40,6	38,5	9	49,4	47,7
4	38,0	40,3	10	48,9	48,2
5	42,2	42,9	11	46,4	45,8
6	48,5	43,8	12	42,2	-

Biểu diễn hai dãy số này lên đồ thị với trục hoành là các tháng và trục tung là các mức độ của hai dãy số: dãy số thực tế (y_i) và dãy số bình quân trượt ($\bar{y}_i$). So sánh hai đồ thị cho thấy đồ thị của dãy bình quân trượt " phẳng" hơn đồ thị của dãy số thực tế vì ảnh hưởng của các yếu tố ngẫu nhiên - qua tính bình quân trượt - phần nào đã bị san bằng .



Việc chọn bao nhiêu mức độ để tính số bình quân trượt đòi hỏi phải dựa vào đặc điểm biến động và số lượng mức độ của dãy số thời gian . Nếu sự biến động tương đối đều đặn và số lượng mức độ của dãy số không nhiều thì có thể tính số bình quân trượt với ba mức độ. Nếu có sự biến động lớn và dãy số có nhiều mức độ thì có thể tính số bình quân trượt với bốn, năm mức độ... Số bình quân trượt càng được tính từ nhiều mức độ thì càng có tác dụng san bằng ảnh hưởng của các yếu tố ngẫu nhiên, nhưng đồng thời làm cho số lượng các mức độ của dãy số bình quân trượt càng giảm, do đó ảnh hưởng đến việc biểu hiện xu hướng phát triển của hiện tượng.

Phương pháp này làm trơn nhẵn sự biến động thực tế nên cũng không dùng với dãy số có biến động thời vụ mà thường dùng với dãy số theo năm.

3.1.3. San bằng mũ

Là phương pháp để san bằng dãy số và dự đoán ngắn hạn. San bằng mũ thực chất là bình quân trượt có trọng số. Trọng số được chọn theo ý chủ quan và nằm trong khoảng từ 0 đến 1. Trọng số lớn hơn thường tính cho các mức độ đứng sau.

$$E_i = W.Y_i + (1 - W).E_{i-1}$$

3.1.4. Hàm xu thế

Ở phương pháp này, sự biến động các mức độ trong dãy số thời gian được biểu hiện bằng một hàm số gọi là hàm xu thế. Nếu gọi t là thứ tự thời gian, dạng tổng quát của hàm xu thế là :

$$\hat{y}_t = f(t) \qquad \text{với } t = 1, 2, 3, \ldots, n$$

Sau đây là một số dạng hàm xu thế thường sử dụng :

a. <u>Hàm xu thế tuyến tính</u> :

Hàm xu thế tuyến tính dược sử dụng khi các lượng tăng (giảm) tuyệt đối liên hoàn xấp xỉ nhau.

$$\hat{y}_t = b_0 + b_1 t$$

Sử dụng phương pháp bình phương nhỏ nhất sẽ có hệ phương trình sau đây để tìm giá trị của các hệ số b_0 và b_1 :

$$\Sigma y = n b_0 + b_1 \Sigma t$$
$$\Sigma ty = b_0 \Sigma t + b_1 \Sigma t^2$$

Hoặc có thể tính b_0, b_1 theo các công thức sau đây :

$$b_1 = \frac{\overline{ty} - \overline{t}\,\overline{y}}{\sigma_t^2}$$
$$b_0 = \overline{y} - b_1 \overline{t}$$

b. *Hàm xu thế pa- ra - bôn* :

Hàm xu thế pa- ra - bôn được sử dụng trong trường hợp các mức độ của dãy số tăng dần theo thời gian, đạt cực đại, sau đó lại giảm dần theo thời gian; hoặc giảm dần theo thời gian, đạt cực tiểu, sau đó lại tăng dần theo thời gian. Dạng tổng quát của hàm xu thế pa- ra - bôn như sau :

$$\hat{y}_t = b_0 + b_1 t + b_2 t^2$$

Sử dụng phương pháp bình phương nhỏ nhất sẽ có hệ phương trình sau đây để tìm giá trị của các hệ số b_0, b_1 và b_2 :

$$\Sigma y = n b_0 + b_1 \Sigma t + b_2 \Sigma t^2$$
$$\Sigma ty = b_0 \Sigma t + b_1 \Sigma t^2 + b_2 \Sigma t^3$$
$$\Sigma t^2 y = b_0 \Sigma t^2 + b_1 \Sigma t^3 + b_2 \Sigma t^4$$

c. *Hàm xu thế hy-pe-bôn*:

Hàm xu thê hy- pe- bôn được sử dụng khi các mức độ của hiện tượng giảm dần theo thời gian. Dạng tổng quát của hàm xu thế hy-pe-bôn như sau:

$$\hat{y}_t = b_0 + \frac{b_1}{t}$$

Sử dụng phương pháp bình phương nhỏ nhất sẽ có hệ phương trình sau đây để tìm giá trị của các hệ số b_0, b_1 :

$$\sum y = n b_0 + b_1 \sum \frac{1}{t}$$
$$\sum \frac{y}{t} = b_0 \sum \frac{1}{t} + b_1 \sum \frac{1}{t^2}$$

d. *Hàm xu thế hàm mũ* :

Hàm xu thế dạng mũ được sử dụng khi các tốc độ phát triển liên hoàn xấp xỉ nhau

$$\hat{y}_t = b_0 \, b_1^t$$

Sử dụng phương pháp bình phương nhỏ nhất sẽ có hệ phương trình sau đây để tìm giá trị của các hệ số b_0, b_1 :

$$\Sigma \ln y = n \ln b_0 + \ln b_1 \Sigma t$$

$$\Sigma t \ln y = \ln b_0 \Sigma t + \ln b_1 \Sigma t^2$$

Giải hệ phương trình trên sẽ được $\ln b_0$, $\ln b_1$; tra đối ln sẽ được b_0, b_1 .

Để xác định đúng dạng hàm xu thế phù hợp, đòi hỏi phải phân tích đặc điểm biến động của hiện qua thời gian, dựa vào đồ thị và một số tiêu chuẩn khác như sai số chuẩn của mô hình- ký hiệu SE :

$$SE = \sqrt{\frac{\sum (y_t - \hat{y}_t)^2}{n - p}}$$

Trong đó :

y_t : mức độ thực tế của hiện tượng ở thời gian t .

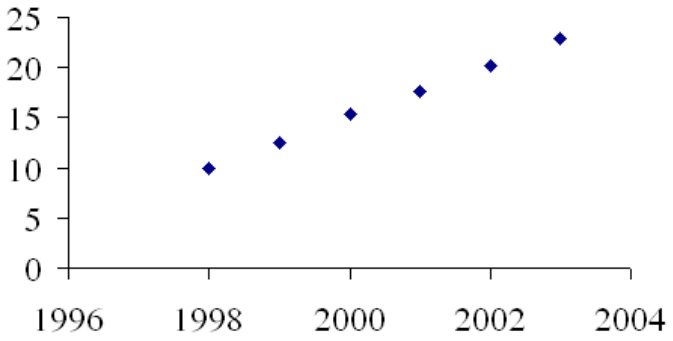
$\hat{y}_t$: mức độ của hiện tượng ở thời gian t được tính từ hàm xu thế .

n : số lượng các mức độ của dãy số thời gian .

p : số lượng các hệ số của hàm xu thế .

Nếu trên đồ thị biểu hiện mức độ thực tế của hiện tượng qua thời gian có thể xây dựng một số hàm xu thế thì chọn hàm xu thế nào có sai số chuẩn của mô hình nhỏ nhất.

Trở lại thí dụ 1, biểu diễn trên đồ thị với trục hoành là thứ tự thời gian, trục tung là các mức độ dãy số :



Trên đồ thị cho thấy các điểm được phân bố hầu như nằm trên một đường thẳng. Mặt khác, ở mục 2.2 cho thấy các lượng tăng tuyệt đối liên hoàn xấp xỉ nhau. Do đó có thể sử dụng hàm xu thế tuyến tính để biểu hiện giá trị sản xuất của doanh nghiệp :

$$\hat{y}_t = b_0 + b_1 t$$

Dựa vào hệ phương trình tìm các hệ số b_0, b_1 để lập bảng tính toán sau đây :

Năm	*Y*	*t*	*ty*	t^2
1999	10,0	1	10,0	1
2000	12,5	2	25,0	4
2001	15,4	3	46,2	9
2002	17,6	4	70,4	16
2003	20,2	5	101,0	25
2004	22,9	6	137,4	36
Tổng	98,6	21	390	91

Thay số liệu vào hệ phương trình :

$$98,6 = 6\ b_0 + 21\ b_1$$

$$390 = 21\ b_0 + 91\ b_1$$

Giải ra, sẽ được : $b_0 = 7,452$, $b_1 = 2,566$. Do đó hàm xu thế tuyến tính biểu hiện giá trị sản xuất của doanh nghiệp có dạng cụ thể như sau :

$$\hat{y}_t = 7,452 + 2,566\ t$$

Hoặc có thể tính :

$$b_1 = \frac{\frac{390}{6} - \frac{21}{6} * \frac{98,6}{6}}{\frac{91}{6} - (\frac{21}{6})^2} = 2,566$$

$$b_0 = \frac{98,6}{6} - 2,566\frac{21}{6} = 7,452$$

3.1.5. Tự tương quan

- Các mức độ trong dãy số thời gian có liên hệ tương quan với nhau.
- Có thể có các mô hình tự tương quan bậc 1, bậc 2, …, bậc p
- Mô hình tổng quát có dạng:

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + ... + A_p Y_{i-p} + \delta_i$$

Trong đó δ_i là sai số ngẫu nhiên.

3.2. Thành phần thời vụ

Biến động thời vụ là sự biến động của hiện tượng có tính chất lặp đi lặp lại trong từng thời gian nhất định của năm. Thí dụ: Sản xuất nông nghiệp phụ thuộc vào thời vụ. Trong các

ngành khác như công nghiệp, xây dựng, giao thông vận tải, dịch vụ, du lịch v.v… ít nhiều đều có biến động thời vụ .

Nguyên nhân gây ra biến động thời vụ là do ảnh hưởng của điều kiện tự nhiên và phong tục, tập quán sinh hoạt. Biến động thời vụ làm cho hiện tượng lúc thì mở rộng, khẩn trương, khi thì thu hẹp, nhàn rỗi. Nghiên cứu biến động thời vụ nhằm đề ra những biện pháp phù hợp, kịp thời hạn chế ảnh hưởng của biến động thời vụ đối với sản xuất và sinh hoạt của xã hội.

Phương pháp thường được sử dụng để biểu hiện biến động thời vụ là tính các chỉ số thời vụ. Tài liệu được sử dụng để tính các chỉ số thời vụ thường là tài liệu hàng tháng hoặc hàng quý của ít nhất ba đến năm năm.

Công thức tính chỉ số thời vụ:

Trường hợp dãy số ổn định: $I_i = \frac{\bar{y}_i}{\bar{y}_0} \cdot 100$

Trường hợp dãy số có xu thế: $I_i = \frac{\sum_{j=1}^{m} \frac{y_{ij}}{\hat{y}_{ij}}}{n} \cdot 100$

Y_{ij} là mức độ tháng (quý) thứ i năm j (i = 1,n ; j = 1,m)

Thí dụ : Có tài liệu về doanh thu (đơn vị tính: tỷ đông) hàng quý của năm năm ở một doanh nghiệp như sau :

Quý / Năm	I	II	III	IV
2000	14,85	16,22	16,62	18,86
2001	16,06	17,01	17,53	19,92
2002	17,04	18,22	18,50	20,85
2003	18,03	19,30	19,66	22,18
2004	18,85	19,97	20,20	22,86

Từ tài liệu trên, ta tính :

- Doanh thu bình quân từng quý :

Quý I : $\bar{y}_I = \frac{14,85 + 16,06 + 17,04 + 18,03 + 18,85}{5} = 16,966$ tỷ đồng

QuýII : $\bar{y}_{II} = \frac{16,22 + 17,01 + 18,22 + 19,30 + 19,97}{5} = 18,144$ tỷ đồng

$$\text{QuýIII}: \bar{y}_{III} = \frac{16,62 + 17,53 + 18,50 + 19,66 + 20,20}{5} = 18,502 \text{ tỷ đồng}$$

$$\text{QuýIV}: \bar{y}_{IV} = \frac{18,86 + 19,92 + 20,85 + 22,18 + 22,86}{5} = 20,934 \text{ tỷ đồng}$$

- Doanh thu bình quân một quý tính chung cho năm năm :

$$\bar{y}_0 = \frac{16,966 + 18,144 + 18,502 + 20,934}{4} = 18,6365 \text{ tỷ đồng}$$

Chỉ số thời vụ có thể được biểu hiện bằng lần hoặc bằng %. Nếu $I_i < 1$ (hoặc 100%) thì sự biến động của hiện tượng ở thời gian i giảm, ngược lại, nếu $I_i > 1$ (hoặc 100%) thì sự biến động của hiện tượng ở thời gian i tăng.

$$I_I = \frac{16,966}{18,6365} = 0,9104 \text{ hay } 91,04\%$$

$$I_{II} = \frac{18,144}{18,6365} = 0,9736 \text{ hay } 97,36\%$$

$$I_{III} = \frac{18,502}{18,6365} = 0,9928 \text{ hay } 99,28\%$$

$$I_{IV} = \frac{20,934}{18,6365} = 1,1233 \text{ hay } 112,33\%$$

Như vậy, doanh thu giảm mạnh ở quý I, rồi đến quý II , quý III và tăng lên ở quý IV.

3.3. Thành phần ngẫu nhiên

Biến động không theo tính hệ thống, mang tính ngẫu nhiên, không dự đoán trước được và không mang tính lặp lại.

Mô hình tổng hợp các nhân tố (không kể thành phần chu kỳ)

+ Số liệu theo năm:

Dạng nhân: $Y_i = T_i.\varepsilon_t$

Dạng cộng: $Y_i = T_i + \varepsilon_t$

+ Số liệu theo tháng (quý)

Dạng nhân: $Y_i = T_i.S_i.\varepsilon_t$

Dạng cộng: $Y_i = T_i + S_i + \varepsilon_t$

Dạng kết hợp: $Y_i = T_i.S_i + \varepsilon_t$

4. Dự đoán dựa vào dãy số thời gian.

Nói chung trên cơ sở dãy số thời gian, một số phương pháp dự đoán ngắn hạn, đơn giản thường dùng trong thống kê gồm :

- Trường hợp dãy số không có xu thế: phương pháp bình quân trượt, san bằng mũ.
- Trường hợp dãy số có xu thế: dự đoán dựa trên lượng tăng (giảm) tuyệt đối bình quân, tốc độ phát triển bình quân và ngoại suy hàm xu thế.
- Trường hợp dãy số có cả thành phần xu thế và thời vụ

4.1. Các phương pháp dự đoán trong trường hợp dãy số có xu thế.

4.1.1. Dự đoán dựa vào lượng tăng (giảm) tuyệt đối bình quân:

+ Điều kiện vận dụng: khi dãy số có lượng tăng giảm tuyệt đối liên hoàn xấp xỉ bằng nhau.

+ Mô hình dự đoán có dạng :

$$\hat{Y} = Y_n + \bar{\delta}.L$$

Trong đó : L là thời hạn dự đoán.

4.1.2. Dự đoán dựa vào tốc độ phát triển bình quân

+ Điều kiện vận dụng: khi dãy số có tốc độ phát triển liên hoàn xấp xỉ bằng nhau.

+ Mô hình dự đoán.

$$\hat{Y} = Y_n \times (\bar{t})^L$$

4.1.3. Dự đoán bằng ngoại suy hàm xu thế

Mô hình dự đoán:

$$\hat{y}_{n+L} - t_{\alpha/2,(n-2)}.S_p \le \hat{Y} \le \hat{y}_{n+L} + t_{\alpha/2,(n-2)}.S_p$$

Trong đó: $S_p = S_{yt}.\sqrt{1+\frac{1}{n}+\frac{3(n+2L-1)^2}{n(n^2-1)}}$

S_{yt} Là sai số của mô hình, chẳng hạn nếu là mô hình tuyến tính thì:

$$S_{yt} = \sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n-2}} = \sqrt{\frac{\Sigma y^2 - b_0\Sigma y - b_1\Sigma ty}{n-2}}$$

Nguyên tắc lựa chọn mô hình để dự đoán: Chọn mô hình đơn giản, có sai số nhỏ nhất...

BÀI TẬP

8.1. Các ưu điểm và nhược điểm của phân tích xu thế? Trong trường hợp nào bạn nên sử dụng phương pháp dự báo này?

8.2. Dữ liệu sau là thị phần bình quân trong quý I của các năm từ 2000-2007 của các hãng xe ô tô Nhật Bản trên thị trường ô tô:

25, 27.5, 27, 25.5, 26.4, 27.3, 28.2, 30.1

Hãy thực hiện phân thích xu thế để dự báo giá trị của năm tiếp sau.

8.3. Dữ liệu sau là số độc giả của một tờ báo địa phương (nghìn người):

Năm:	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Số độc giả:	53	65	74	85	92	105	120	128	144	158	179	195

Hãy thực hiện một hàm xu thế cho dữ liệu trên, dự báo tổng số độc giả của năm 2008 và 2009 với độ tin cậy 95%.

8.4. Liệu phương pháp phân tích xu thế có phải là một công cụ dự báo hữu ích đối với doanh số bán áo bơi hay không? Giải thích?

8.5. Lợi nhuận của một doanh nghiệp thay đổi theo chu kì kinh doanh trong vài năm. Liệu phương pháp phân tích xu thế có phải là công cụ dự báo hữu ích đối với lợi nhuận của doanh nghiệp hay không? Giải thích?

8.6. Hãy giải thích sự khác nhau giữa hai khái niệm thay đổi theo thời vụ và thay đổi theo chu kì?

8.7. Sau đây là chỉ số Dow Jones hàng tháng đối với giá cả các hàng hoá tiêu dùng từ 6/2006 đến 6/2007:

142, 137, 143, 142, 149, 143, 151, 150, 151, 146, 144, 145, 147

Hãy xây dựng một mô hình san bằng mũ, sử dụng w = 0.6, và dự báo chỉ số hàng hoá tiêu dùng vào tháng 7/2007. Thử nghiệm với các giá trị khác của w.

8.8. Sau đây là chỉ số sản lượng sản xuất công nghiệp của Nigeria trong các năm 1994-2007.

Năm	Chỉ số
1994	175
1995	190
1996	132
1997	96
1998	100
1999	78
2000	131
2001	135
2002	154
2003	163

2004	178
2005	170
2006	145
2007	133

a. Năm gốc được sử dụng ở đây là năm nào?

b. Hãy thay đổi năm gốc là năm 2003.

c. Điều gì đã xảy ra đối với sản lượng công nghiệp của Nigeria từ năm 2006-2007.

d. Mô tả các xu hướng trong sản lượng công nghiệp từ năm 1994-2007.

8.9 Có tài liệu về số khách du lịch quốc tế của một đơn vị kinh doanh du lịch như sau:

Năm / Tháng	2000	2001	2002	2003	2004	2005
1	462	493	516	490	482	580
2	481	510	522	536	566	602
3	470	502	547	503	550	614
4	458	431	526	516	542	608
5	393	473	498	490	560	585
6	355	402	455	422	414	573
7	314	337	413	440	383	507
8	286	306	380	410	372	466
9	370	280	310	366	351	377
10	358	314	292	372	344	370
11	430	460	363	241	303	299
12	471	490	470	388	460	310

Yêu cầu:

- Tính chỉ số thời vụ và cho nhận xét
- Xác định hàm xu thế biểu hiện biến động lượng khách du lịch qua các năm
- Dự đoán số lượng khách du lịch cho 2 năm tiếp theo và theo từng tháng.