TRƯỜNG ĐẠI HỌC KINH TẾ
VÀ QUẢN TRỊ KINH DOANH THÁI NGUYÊN
TS. Đỗ Anh Tài

# GIÁO TRÌNH PHÂN TÍCH SỐ LIỆU THỐNG KÊ

NHÀ XUẤT BẢN THỐNG KÊ

TRƯỜNG ĐẠI HỌC KINH TẾ
VÀ QUẢN TRỊ KINH DOANH THÁI NGUYÊN
TS. Đỗ Anh Tài

# GIÁO TRÌNH PHÂN TÍCH SỐ LIỆU THỐNG KÊ



NHÀ XUẤT BẢN THỐNG KÊ
Hà Nội, tháng 8/2008

# LỜI NÓI ĐẦU

*Một nghiên cứu tốt, một báo cáo khoa học có căn cứ được người đọc chấp nhận về mặt học thuật đòi hỏi phải có phương pháp tốt, áp dụng các công cụ kỹ thuật để cung cấp các thông tin xác thực. Đặc biệt trong các vấn đề kinh tế - xã hội và khi nghiên cứu số lớn chúng ta cần phải quan tâm đến các công cụ kỹ thuật như thống kê.*

*Thống kê học là một lĩnh vực khá rộng, do vậy trong phạm vi của môn học này tác giả mong muốn trang bị cho người học những kiến thức cơ bản trong việc phân tích số liệu thống kê nhằm mục đích có thể khai thác hiệu quả các thông tin thu thập được, để phục vụ cho công tác nghiên cứu khoa học của các khoa học về kinh tế - xã hội.*

*Với mục đích trang bị kiến thức chuyên sâu cho các sinh viên sau đại học có thể triển khai tốt các nghiên cứu của mình, tác giả mong muốn cuốn sách này sẽ trở thành cẩm nang cho các bạn.*

*Cuốn sách gồm 4 chương chính bao trùm các nội dung từ việc chuẩn bị số liệu cho đến phân tích số liệu và biểu diễn kết quả thành báo cáo khoa học.*

*Lần đầu tiên cuốn sách được giới thiệu đến bạn đọc, mặc dù đã được cập nhật những thông tin mới nhất và hiện đại nhất nhưng cuốn sách khó tránh khỏi những thiếu sót nhất định. Tác giả mong nhận được những ý kiến đóng góp của bạn đọc để lần xuất bản sau cuốn sách được hoàn thiện hơn.*

*Tác giả cũng xin chân thành cảm ơn những ý kiến đóng góp quý báu của bạn đọc, các bạn đồng nghiệp và sinh viên để cuốn sách được xuất bản.*

TÁC GIẢ

# GIỚI THIỆU CHUNG

Do đặc thù khác nhau của việc thu thập và xử lý các số liệu thống kê nên trong khuôn khổ của cuốn sách này chúng tôi mong muốn tập trung vào những vấn đề về các số liệu thống kê phục vụ cho các nghiên cứu về lĩnh vực kinh tế - xã hội.

*Khái niệm phân tích số liệu thống kê:* Là sự kết hợp giữa thống kê, sự tư duy và hiểu biết các vấn đề kinh tế.

*Yêu cầu:* Để có thể nắm vững kiến thức của môn học này đòi hỏi người học phải có kiến thức sâu về thống kê, về kinh tế cũng như những hiểu biết thực tế của vấn đề nghiên cứu. Ngoài ra, cần phải có kiến thức về tin học và các công cụ lượng hoá khác để kết hợp trong nghiên cứu.

Trước khi bước vào nội dung của chương thứ nhất chúng tôi muốn trao đổi sơ lược với các độc giả về tổng quát tiến hành một nghiên cứu trong các vấn đề thuộc về kinh tế - xã hội.

Trong khi tiến hành các nghiên cứu về kinh tế - xã hội có gì khác so với các vấn đề thuộc về khoa học tự nhiên: điều khác cơ bản đó là đối tượng nghiên cứu trong các nghiên cứu kinh tế - xã hội thường là con người hoặc là liên quan đến con người, các mối liên hệ với con người. Ngoài ra, nó còn khác nhau ở cách thức tiến hành, khả năng áp dụng và thời gian cho kết quả, phạm vi tác động v.v...

Thiết kế một nghiên cứu về kinh tế - xã hội cần phải làm những gì? Dưới đây là một đề cương sơ bộ hướng dẫn cho các nghiên cứu thuộc về lĩnh vực khoa học kinh tế - xã hội. Nó sẽ được cụ thể hoá cho từng chương trình nghiên cứu cụ thể.

***I. Vấn đề đặt ra***

A. Trình bày một cách rõ ràng, ngắn ngọn về vấn đề đặt ra, với việc xác định khái niệm cần thiết như thế nào.

B. Chỉ ra vấn đề là sự giới hạn về ranh giới để giải quyết hoặc kiểm tra vấn đề.

C. Mô tả sự cần thiết và ý nghĩa của vấn đề liên quan đến một trong những chỉ tiêu sau:

1. Thời gian.

2. Liên quan đến vấn đề thực tế.

3. Liên quan đến tổng thể rộng lớn hơn.

4. Liên quan đến sự tác động hoặc phản ảnh đến tổng thể.

5. Làm thoả mãn khoảng cách của một nghiên cứu.

6. Cho phép suy rộng ra các hoạt động xã hội hoặc các nguyên lý cơ bản.

7. Làm rõ các khái niệm, mối quan hệ và sự quan trọng.

8. Tìm hiểu phạm vi thực tế của vấn đề trong thực tế.

9. Có thể tạo ra hoặc phát triển những công cụ quan trọng cho việc quan sát và phân tích thông tin.

10. Cung cấp cơ hội và khả năng thu thập thông tin trong thực trạng của việc hạn chế về thời gian.

11. Trình bày khả năng có thể giải thích hoặc phân tích kết quả một cách tốt nhất, có nhiều thông tin nhất dựa trên cơ sở các kỹ thuật phân tích đã có.

***II. Cơ sở lý luận***

A. Trình bày mối quan hệ của vấn đề đến cơ sở lý luận.

B. Sự liên quan của vấn đề nghiên cứu tới các nghiên cứu trước đây.

C. Trình bày các giả cứ lý luận liên quan.

***III. Giả thuyết***

A. Làm rõ các giả thuyết lựa chọn cho việc kiểm định.

B. Thể hiện mức độ ý nghĩa của kiểm định các giả thuyết tới sự tiến bộ của nghiên cứu và lý luận.

C. Định nghĩa các khái niệm hoặc các biến sử dụng (tốt nhất nên ở dạng quan hệ phụ thuộc).

*Ví dụ:* Thu nhập là phần còn lại của doanh thu sau khi trừ đi chi phí trước khi tính công lao động.

1. Các biến độc lập (biến giải thích) và biến phụ thuộc (biến được giải thích) giữa chúng nên được phân biệt rõ.

2. Tỷ lệ trên đó các biến được xác định và đo đạc (định lượng, bán định lượng hay định tính) cần được cụ thể.

D. Miêu tả những lỗi có thể mắc phải và hậu quả của nó.

E. Chú thích các lỗi nghiêm trọng.

***IV. Thiết kế một thí nghiệm hay cuộc điều tra***

A. Trình bày ý tưởng những thiết kế với những quan tâm cụ thể trong việc đáp ứng tính phức tạp của các biến.

B. Mô tả việc lựa chọn một thiết kế để tiến hành.

1. Mô tả các tác nhân kích thích, chủ đề, môi trường và câu hỏi của các mục tiêu, các sự kiện, nhu cầu cần thiết về vật lực.

2. Mô tả làm thế nào để điều khiển được tính phức tạp của các biến.

C. Cụ thể các công cụ dùng để kiểm định thống kê bao gồm cả các bảng giả định cho mỗi một kiểm định.

Trong đó, cần cụ thể mức độ tin cậy mong muốn.

***V. Quá trình chọn mẫu***

A. Mô tả mẫu được lựa chọn trong thí nghiệm hoặc điều tra.

1. Cụ thể tổng thể có liên quan đến giả thuyết nghiên cứu.

2. Giải thích sự xác định của số lượng và kiểu loại mẫu.

B. Cụ thể hoá phương pháp lựa chọn mẫu.

1. Cụ thể hoá mối quan hệ tương đối của sai số ngẫu nhiên và phi ngẫu nhiên.

2. Ước lượng chi phí tương đối của các cỡ và kiểu lấy mẫu khác nhau phù hợp với lý thuyết.

***VI. Phương pháp và cách thức điều tra***

A. Mô tả thước đo của các biến định lượng chỉ ra tính tin cậy và hợp lệ của chúng. Mô tả phương tiện để xác định cho các biến định tính.

B. Các mục bao gồm trong bảng câu hỏi điều tra.

1. Số lượng câu hỏi có thể phỏng vấn người được hỏi.

2. Thời gian có thể cho cuộc phỏng vấn.

3. Lịch trình tiến hành trong thời gian cụ thể.

4. Những kết quả đánh giá, kiểm định trước.

C. Các mục bao gồm trong quá trình điều tra.

1. Các phương tiện thu thập thông tin.

*Ví dụ:* Phỏng vấn trực tiếp, hoặc một phần bằng thư, điện thoại hay các phương tiện khác.

2. Các đặc trưng riêng mà người điều tra viên phải có hoặc cần phải tập huấn cho họ.

D. Mô tả kết quả sử dụng từ các nghiên cứu đại diện hoặc điều tra thử. Trong đó nêu rõ sự quan trọng, các phương tiện để

xử lý tình trạng thông tin kém giá trị, bị loại bỏ hoặc do lỗi người được hỏi.

***VII. Hướng dẫn trong quá trình tiến hành***

A. Chuẩn bị một hướng dẫn trong toàn bộ quá trình tiến hành nghiên cứu, trong đó trình bày cụ thể thời gian và ước tính chi phí.

1. Kế hoạch.

2. Địa điểm nghiên cứu và các kiểm tra trước.

3. Lựa chọn mẫu.

4. Chuẩn bị các trang thiết bị vật chất cho điều tra.

5. Lựa chọn điều tra viên và tiến hành tập huấn.

6. Kế hoạch triển khai thực địa.

7. Chỉnh sửa lại kế hoạch.

8. Thu thập thông tin.

9. Phân tích thông tin (số liệu).

10. Chuẩn bị báo cáo kết quả nghiên cứu.

B. Ước tính số lượng công lao động và chi phí.

***VIII. Phân tích số liệu***

Cụ thể các phương pháp dùng trong phân tích:

1. Sử dụng bảng biểu, các công cụ tính toán, cách thức phân loại, máy tính v.v...

2. Sử dụng các kỹ thuật đồ hoạ.

3. Cụ thể các loại bảng biểu sẽ thiết kế.

***IX. Giải thích kết quả***

Thảo luận kết luận như thế nào sẽ phản hồi cho các giả thuyết đặt ra

***X. In quyển hoặc báo cáo kết quả***

A. Kết quả được viết và in ấn theo yêu cầu của đơn vị đào tạo hoặc nghiên cứu.

B. Lựa chọn kết quả viết bài báo cho các tạp chí khoa học.

Những hướng dẫn trên đây chỉ mang tính chất gợi ý cho những người nghiên cứu, tuỳ từng trường hợp cụ thể mà người nghiên cứu có thể cụ thể hoá hoặc thay đổi theo thực tế yêu cầu.

Phát triển việc phân tích số liệu thống kê thường song song hoặc nâng cao của các vấn đề nghiên cứu khác mà trong đó việc ứng dụng các công cụ thống kê là cần thiết. Bởi vì, phân tích thống kê thường dùng cho những vấn đề quyết định mà việc áp dụng các công cụ thống kê sẽ giúp đưa ra những quyết định đúng đắn hơn trong những điều kiện không biết trước.

# Chương I

# CHUẨN BỊ SỐ LIỆU

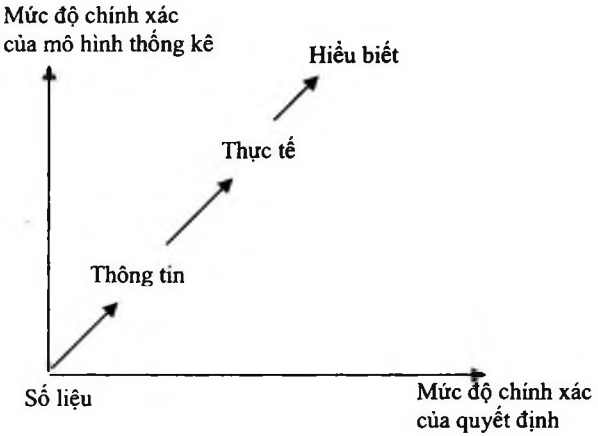
Nội dung của chương này nhằm trang bị cho sinh viên những kiến thức cơ bản về chuẩn bị, điều tra thu thập số liệu phục vụ cho nghiên cứu.

Có được số liệu với chất lượng cao và có độ tin cậy cũng như tính đầy đủ phục vụ cho nghiên cứu là điều hết sức quan trọng, nó quyết định đến kết quả của nghiên cứu đối với mỗi một nhà khoa học kể cả trong lĩnh vực kỹ thuật hay kinh tế - xã hội. Vì vậy, việc thiết kế điều tra như thế nào? Các phương pháp điều tra lựa chọn ra làm sao? Việc chọn mẫu điều tra v.v... sẽ có ảnh hưởng trực tiếp đến kết quả của số liệu mà chúng ta sẽ thu thập được sau này.

Số liệu cho ta biết những gì?

Từ số liệu sẽ cung cấp cho ta những thông tin cần thiết qua đó để vẽ lên được bức tranh thực tế, đây là bức tranh không gian 3 chiều, nó cho ta biết thực tại, quá khứ và cả những điều dự đoán trong tương lai. Từ đó, nó giúp ta xây dựng và phát triển những hiểu biết.

Có bộ số liệu tốt, có được mô hình phân tích thống kê chính xác sẽ giúp ta đưa ra những quyết định chính xác hơn, phù hợp hơn với thực tế.




**Hình 1.1: MỐI QUAN HỆ CỦA VIỆC PHÂN TÍCH SỐ LIỆU VÀ VIỆC RA QUYẾT ĐỊNH**

## 1.1. Thiết kế điều tra nghệ thuật và khoa học

Có thể nói việc thiết kế điều tra vừa mang tính nghệ thuật vừa phải có tính khoa học, điều đó có thể lý giải bởi các lý do như sau:

***Tính khoa học*** được thể hiện ở chỗ khi thiết kế điều tra chúng ta phải sử dụng các nguyên tắc của thống kê mà đã được học như việc chọn mẫu mang tính đại diện về những đặc tính của tổng thể và số lượng của mẫu để có thể suy rộng ra được.

Khi thiết kế điều tra, chúng ta cũng phải dựa vào các nguyên lý kinh tế học như các mối quan hệ khi nghiên cứu

một vấn đề có liên quan đến các vấn đề khác v.v... và một điều cũng hết sức quan trọng đó là vấn đề tâm lý học trong điều tra thu thập số liệu: chúng ta sẽ hỏi ai, với những câu hỏi như thế nào v.v...

Một ví dụ mà chúng ta sẽ thấy rất rõ là khi chúng ta hỏi về tác động của một chương trình trợ giúp mà chúng ta hay chính phủ đang tiến hành với câu hỏi tương tự như "Anh/chị thấy chương trình mà chúng tôi hay chính phủ đang thực hiện như thế nào?" thì thường chúng ta sẽ nhận được câu trả lời ở dạng tốt vì phần lớn người dân mong muốn sẽ được trợ giúp tiếp theo.

Hoặc khi chúng ta tiến hành điều tra mà đối tượng là các nhóm dân tộc khác nhau thì chúng ta cũng cần phải lưu tâm đến phong tục, tập quán của họ để tránh những vấn đề mà phong tục của họ không cho phép hoặc chúng ta sẽ không thể thu thập được thông tin mong muốn.

Do vậy, việc thiết kế điều tra phải hết sức khoa học trong việc vận dụng một cách chặt chẽ những kiến thức đã có về thống kê, tâm lý học và các nguyên lý kinh tế học.

Thiết kế điều tra được coi như một nghệ thuật là vì:

- Thứ nhất, chúng ta không có bất kỳ một sách hướng dẫn chuẩn và tổng quát cho thiết kế điều tra nào mà nó phụ thuộc rất nhiều vào những mục tiêu và câu hỏi nghiên cứu chính mà NHÀ NGHIÊN CỨU phải giải quyết trong nghiên cứu của mình.

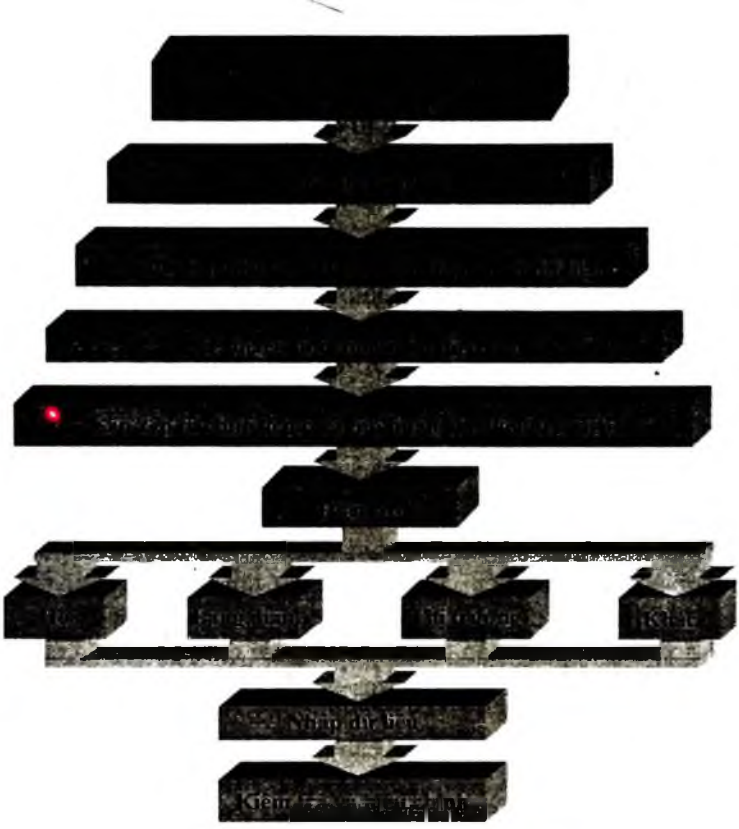
*Ví dụ:* Nó có thể liên quan đến các vấn đề về lao động, việc làm; có thể là các vấn đề về thu nhập và đói nghèo hay các vấn đề về nguồn lực và quản lý các nguồn lực; các số liệu thí nghiệm v.v... nó có thể ở mức độ vi mô nhưng cũng có thể ở cấp độ vĩ mô. Điều này hoàn toàn phụ thuộc vào mục đích nghiên cứu, vấn đề cần nghiên cứu của người làm nghiên cứu.

- Thứ hai, việc thiết kế điều tra tuỳ thuộc vào mức độ nguồn lực và các liên quan khách quan khác mà chúng ta thường đánh đổi giữa các mục tiêu khác nhau.

*Ví dụ:* Đánh đổi giữa kinh phí và số lượng mẫu cần điều tra, giữa sai số chọn mẫu (số lượng mẫu nhiều hay ít) và sai số phi chọn mẫu (số lượng người tham gia điều tra nhiều hay ít với những kinh nghiệm điều tra khác nhau) v.v...

Vì thế, có thể nói rằng thiết kế điều tra là một nghệ thuật mà người làm nghệ thuật ở đây không ai khác hơn là chính người làm nghiên cứu, họ sẽ phải lựa chọn, họ sẽ phải tung hứng một cách nghệ thuật để đáp ứng những yêu cầu do mình đặt ra để được người đọc và những người ứng dụng chấp nhận được kết quả mà mình làm ra.

Thông thường các cuộc điều tra được tiến hành theo một trình tự kế hoạch như hình 1.2:



**Hình 1.2: SƠ ĐỒ KẾ HOẠCH TIẾN HÀNH MỘT CUỘC ĐIỀU TRA THU THẬP SỐ LIỆU PHỤC VỤ NGHIÊN CỨU**

*- Bước 1: Thiết lập các mục tiêu*

Bước đầu tiên trong bất kỳ một cuộc điều tra nào đều là quyết định bạn muốn tìm hiểu vấn đề gì? Mục tiêu của dự án

quyết định bạn sẽ điều tra ai và sẽ hỏi họ điều gì? Nếu các mục tiêu của bạn không rõ ràng, kết quả cũng có thể sẽ không rõ ràng, do vậy, bạn phải luôn làm rõ các mục tiêu của mình để: *thứ nhất,* thu được số liệu như mong muốn; *thứ hai,* không bị thiếu những thông tin cần thiết cho việc nghiên cứu của bạn.

*Ví dụ:* Trong cuộc điều tra 150 doanh nghiệp dệt may ở Việt Nam do Viện Kinh tế học tiến hành năm 2001, mục tiêu chung mà nhóm nghiên cứu năng suất đặt ra là: Xác định những nhân tố ảnh hưởng đến năng suất của các doanh nghiệp dệt may Việt Nam, và dựa trên cơ sở này đưa ra những khuyến nghị đối với các nhà hoạch định chính sách để nâng cao năng suất của doanh nghiệp.

Một ví dụ nữa khi chúng ta tiến hành điều tra về vấn đề đói nghèo, chúng ta cần phải xác định rõ là chúng ta muốn nghiên cứu gì? Nếu chúng ta muốn tìm hiểu về nguyên nhân dẫn đến đói nghèo thì mục đích của chúng ta là điều tra xác định các nhân tố ảnh hưởng tới mức sống của các hộ dân cả tích cực và tiêu cực.

Trên cơ sở xác định được các mục tiêu nghiên cứu, chúng ta sẽ tiến hành lựa chọn vùng nghiên cứu phù hợp và mang tính đại diện và tiến hành bước thứ hai.

*- Bước thứ hai: Là pha chuẩn bị của một điều tra phục vụ nghiên cứu*

Nội dung của bước này là khâu chuẩn bị các nguồn lực như: nhân lực và vật lực phù hợp theo đòi hỏi và nhu cầu của cuộc điều tra trên cả 2 phương diện: số lượng và chất lượng. Trong đó, đặc biệt phải nhấn mạnh đến việc chuẩn bị về tài

chính và vật lực vì nó gần như quyết định đến sự thành công trong quá trình điều tra sau này.

*- Bước thứ 3: Là bước xây dựng phiếu điều tra và cơ sở dữ liệu*

Phiếu điều tra được xây dựng dựa trên mục tiêu của cuộc điều tra cũng như đối tượng và cách thức tiến hành điều tra.

Trong đó, mục tiêu của cuộc điều tra quyết định đến nội dung của phiếu điều tra, các thông tin cần thu thập, thời gian của số liệu cũng như thời gian của cuộc điều tra.

*Ví dụ:* Mục tiêu là điều tra kinh tế hộ nông dân sẽ cần 1 phiếu điều tra khác so với điều tra một doanh nghiệp hay điều tra năng suất lao động khác với điều tra tình hình giảng dạy và học tập trong các trường đại học.

Đối tượng điều tra và cách thức tiến hành điều tra có ảnh hưởng nhiều đến cách thức ra câu hỏi: dạng mở hay dạng có phương án lựa chọn và trong trường hợp nào thì nên để nhiều câu hỏi dạng mở, khi nào thì dùng câu hỏi có gợi ý trước.

Cơ sở dữ liệu (Database) có thể định nghĩa là *phần thông tin thu thập được từ một cuộc điều tra bất kỳ nào đó và được sắp xếp theo một trật tự nhất định để có thể dễ dàng cho việc xử lý số liệu đó thông qua các phần mềm thống kê, cũng như dễ dàng cho việc kiểm tra độ chuẩn xác của thông tin thu lượm được.* Thông thường cơ sở dữ liệu có thể là một bảng tính trong Excel, Lotus hay 1 tệp trong Access và đôi khi chúng ta có thể sử dụng trực tiếp các phần mềm thống kê để nhập dữ liệu như SPSS (Statistical Package for the Social Sciences).

*- Bước thứ 4: Tiếp theo là việc lựa chọn mẫu điều tra về số lượng và cách thức chọn mẫu*

Việc lựa chọn số lượng mẫu điều tra phụ thuộc vào nhiều yếu tố khác nhau như điều kiện tối thiểu của mẫu, tính đại diện của mẫu, khả năng đáp ứng về thời gian, nguồn nhân lực và vật lực (trong đó chủ yếu là nguồn lực tài chính).

Tuy nhiên, mục tiêu của cuộc điều tra đôi khi cũng có ảnh hưởng tới số lượng mẫu, chẳng hạn như mục tiêu là nghiên cứu các trường hợp đặc biệt (case study) thì khi đó số lượng mẫu không cần lớn, còn khi cuộc điều tra để đánh giá chung cho một nhóm đối tượng như nhóm khách hàng hay một vùng như khu vực miền núi chẳng hạn thì số lượng mẫu đòi hỏi phải nhiều hơn và phải đủ để đại diện cho tổng thể nghiên cứu.

Số lượng mẫu cũng phụ thuộc vào yêu cầu của độ chính xác của các thông tin phân tích từ kết quả điều tra, nếu đòi hỏi có độ chính xác cao thì số lượng mẫu cũng sẽ tăng lên.

Tuy nhiên, giữa số lượng và chất lượng thông tin cũng có mâu thuẫn với nhau khi ta đề cập đến sai số phi chọn mẫu ở phần sau. Ngoài ra, số lượng mẫu còn phụ thuộc rất nhiều vào các yếu tố khách quan khác trong quá trình điều tra như thời gian cho phép, kinh phí có thể đáp ứng v.v...

Cách thức chọn mẫu rất quan trọng vì nó sẽ ảnh hưởng tới khả năng đại diện của mẫu cho tổng thể. Cách thức chọn mẫu hoàn toàn phụ thuộc vào tính chất của tổng thể đó là đồng nhất hay có sự khác biệt bên trong của tổng thể mẫu đó.

*Ví dụ:* Khi mẫu đó là đồng đều thì ta có thể lựa chọn mẫu ngẫu nhiên đơn giản, nhưng khi trong tổng thể có nhiều nhóm, lớp khác biệt nhau thì việc lựa chọn mẫu theo phân lớp hay nhiều cấp sẽ đảm bảo tính đại diện cho tổng thể hơn.

*- Bước thứ 5: Tiến hành điều tra*

Việc tiến hành điều tra cũng phải qua nhiều khâu khác nhau như tập huấn cho điều tra viên, điều tra thử và tiến hành điều tra chính thức.

Khâu quan trọng nữa là sắp xếp một cách lô gíc các khâu trong cuộc điều tra cũng như tập huấn một cách kỹ lưỡng cho các điều tra viên, những người sẽ trực tiếp tiến hành việc thu thập số liệu, vì chất lượng của số liệu mà ta có sẽ hoàn toàn phụ thuộc vào những người này. Do vậy, công tác lựa chọn điều tra viên có đủ năng lực, trình độ, trách nhiệm với công việc là hết sức quan trọng. Đặc biệt quan trọng là chúng ta phải lựa chọn được những người có mong muốn tham gia công tác điều tra, nghiên cứu có như vậy họ mới luôn đề cao tinh thần trách nhiệm đối với thông tin mà họ sẽ thu thập.

Việc điều tra thử sẽ giúp cho việc chỉnh sửa phiếu điều tra cho phù hợp để thu được các thông tin cần thiết một cách chính xác nhất, bước này thường được tiến hành trước khi tiến hành điều tra chính thức khoảng 1 tháng để có điều kiện hoàn chỉnh lại phiếu điều tra và bổ sung tập huấn cho cán bộ điều tra nếu cần thiết.

Việc tiến hành điều tra sẽ tuỳ thuộc vào yêu cầu của số liệu mà tiến hành điều tra trong cùng 1 thời điểm hay chia ra nhiều giai đoạn. Tuy nhiên, nếu có nhiều quan sát trong cùng một mẫu thì lên lấy số liệu trong cùng 1 thời điểm để tránh sự khác biệt của số liệu do tác động của thời gian.

Trong quá trình điều tra chúng ta cần phải luôn kiểm tra độ chính xác của thông tin để có thể điều chỉnh kịp thời.

Việc triển khai công tác kiểm tra độ chính xác có thể được sử dụng theo nhiều cách khác nhau, chẳng hạn chúng ta có thể lập ra một đội kiểm tra độc lập với đội đi điều tra. Sau khi kiểm tra nếu phiếu điều tra có những chỗ chưa rõ ràng hoặc thiếu thông tin thì chuyển lại phiếu đó cho điều tra viên. Điều đó có nghĩa là việc kiểm tra này phải tiến hành đồng thời, song song với việc điều tra.

Để tránh những sai sót hoặc thông tin bị thiếu trong quá trình điều tra thì một trong những lưu ý vô cùng quan trọng là các điều tra viên cần phải hoàn chỉnh phiếu điều tra ngay tại hộ hoặc tại địa bàn điều tra để kịp thời bổ sung những thông tin cần thiết.

Một lưu ý nữa là trong quá trình điều tra chúng ta nên thường xuyên tổ chức những buổi họp nhóm điều tra (cả điều tra viên, lãnh đạo và kiểm tra viên) để cùng nhau trao đổi những phát sinh cần phải điều chỉnh trong quá trình điều tra.

*- Bước thứ 6: Nhập dữ liệu vào cơ sở dữ liệu trong máy tính.*

Bước này thuần tuý là khâu kỹ thuật, song nó khá quan trọng và cũng chiếm nhiều thời gian, đòi hỏi người nghiên cứu phải kiên nhẫn và tỉ mỉ, có độ chính xác cao, tránh nhầm lẫn. Người nghiên cứu cũng phải am hiểu về máy tính và các phần mềm sử dụng để xử lý số liệu sau này nhằm xây dựng một cấu trúc cơ sở dữ liệu cho phù hợp.

Để đảm bảo độ chính xác cao của thông tin có 2 cách làm như sau:

+ *Cách thứ nhất:* chúng ta có thể tiến hành nhập tin hai lần, để kiểm tra cách này tốn nhiều thời gian nhưng đảm bảo ít sai sót trong quá trình nhập tin.

+ *Cách thứ hai:* sau khi nhập xong toàn bộ thông tin cần phải in tất cả ra giấy để kiểm tra, tuy nhiên cách này cũng sẽ gặp một số khó khăn nhất định khi nhiều thông tin và trên định dạng bảng tính như Excel chẳng hạn.

Để dung hoà cho những điểm hạn chế của cả 2 cách chúng ta có thể sử dụng cách lấy ngẫu nhiên một lượng phiếu khoảng 10% để kiểm tra thông qua việc nhập lại những thông tin này rồi kiểm tra với những thông tin của các phiếu đó đã nhập lần 1 nhằm tìm ra những sai sót (nếu có). Nếu có sai sót ta buộc phải kiểm tra lại toàn bộ cơ sở dữ liệu đã nhập trước đây.

Việc kiểm tra này thường được tiến hành ngay sau khi đã nhập được một phần số phiếu điều tra trong tổng số phiếu của cả đợt, nhằm mục đích phát hiện sớm những sai sót để có thể kịp thời điều chỉnh, tránh hiện tượng sau khi nhập toàn bộ phiếu mới kiểm tra, vì khi đó việc điều chỉnh sẽ rất tốn thời

gian và công sức. Nếu việc kiểm tra được tiến hành sau khi đã nhập hết các phiếu mà bắt gặp nhiều sai sót thì việc nhập lại tin hay kiểm tra cũng sẽ không có nhiều ý nghĩa như khi mới bắt đầu tiến hành nhập tin. Việc kiểm tra này cũng phải tiến hành thường xuyên trong quá trình nhập tin vào máy.

*- Bước thứ 7: Kiểm tra và hiệu chỉnh thông tin trong cơ sở dữ liệu hay còn gọi là quá trình làm sạch thông tin*

Đây là khâu cuối cùng trước khi tiến hành xử lý số liệu bằng các phần mềm thống kê và cũng là khâu quan trọng vì nó sẽ quyết định đến độ tin cậy của kết quả phân tích sau này.

Để tiến hành kiểm tra, chúng ta cần phải qua các bước sau:

+ Thứ nhất là kiểm tra xem có các giá trị bất thường hoặc bị thiếu hay không (thông thường thông qua sử dụng một số phần mềm thống kê chuyên dụng như SPSS hay Stata).

+ Sau khi kiểm tra thấy có các giá trị bất thường hoặc bị thiếu đó thì kiểm tra lại xem giữa việc nhập số liệu và phiếu điều tra có chính xác hay không? Nếu quá trình nhập đã chính xác thì câu hỏi đặt ra sẽ là kiểm tra lại trong quá trình điều tra được tiến hành như thế nào? Nếu quá trình điều tra có thể khẳng định được là thông tin hoàn toàn chính xác và đã kiểm tra thông tin ngay khi đi điều tra thì ta có thể chấp nhận được thông tin này. Còn nếu câu trả lời là không thì khi đó ta sẽ phải xử lý theo 2 cách: (1) điều tra lại thông tin đó, nếu điều kiện cho phép hoặc trong trường hợp số lượng thông tin đó lớn; (2) dùng các phương pháp xử lý số liệu bị thiếu và bất

thường trong thống kê nếu lượng số liệu đó so với toàn bộ khối lượng thông tin chiếm tỷ lệ rất nhỏ (thường chỉ một vài phần trăm).

## 1.2. Chọn mẫu

Phần này nhằm trang bị những kiến thức cơ bản cho người nghiên cứu trong việc lựa chọn phương pháp chọn mẫu thích hợp trong nghiên cứu của minh. Với mục tiêu cụ thể của phần này là:

1. Phân biệt giữa tổng thể và mẫu: xác định rõ câu hỏi nghiên cứu và mục tiêu nghiên cứu; lựa chọn các chỉ tiêu lựa chọn và loại trừ.

2. Lựa chọn phương pháp chọn mẫu phù hợp giữa ngẫu nhiên và phi ngẫu nhiên.

3. Hiểu những lập luận về ước lượng sai số.

4. Hiểu những lập luận về xác định cỡ mẫu.

5. Hiểu những lập luận về nguồn sai số trong chọn mẫu.

6. Tính toán trọng số.

### *1.2.1. Chọn mẫu thống kê trong điều tra chọn mẫu*

Trước hết, chúng ta sẽ làm quen với một số khái niệm trong điều tra chọn mẫu cũng như cần phân biệt giữa nhóm đối tượng và mẫu như sau:

- *Tổng thể:* Là một nhóm người, chi tiết hoặc đơn vị đối tượng của nghiên cứu sẽ được điều tra. Tổng thể bao gồm 2

loại là tổng thể lý thuyết và tổng thể có thể tiếp cận được. Trong đó:

+ *Tổng thể lý thuyết:* Là những nhóm đối tượng phù hợp trong nghiên cứu và có thể rộng hơn, bao trùm tổng thể có thể tiếp cận được.

*Ví dụ:* Khi nghiên cứu về hộ nông dân thì tất cả các hộ nông dân là tổng thể lý thuyết.

+ *Tổng thể có thể tiếp cận được:* Là nhóm đối tượng có thể cho phép tiếp cận trong quá trình nghiên cứu và lựa chọn mẫu.

Với ví dụ trên, chúng ta không thể tiếp cận được tất cả các hộ do việc phân bố rộng, do vậy chỉ những hộ ở khu vực nghiên cứu mới cho phép ta có thể tiếp cận được. Đây là nhóm tổng thể có thể tiếp cận được.

- *Tổng điều tra:* Là một cuộc điều tra nhằm thu thập thông tin về mỗi thành viên của tổng thể, do vậy, nó được tiến hành điều tra với tất cả số thành viên có trong tổng thể của nghiên cứu.

*Ví dụ:* Cuộc điều tra tổng thể nông nghiệp, nông thôn năm 1994 hay 2006 của Tổng cục Thống kê là tổng điều tra.

- *Khung chọn mẫu:* Là danh sách những người (từ tổng thể tiếp cận được) để từ đó ta có thể chọn mẫu để điều tra. Danh sách này nên thể hiện toàn diện, hoàn chỉnh và được cập nhật.

*Ví dụ:* Danh sách Đăng ký cử tri, danh sách địa chỉ theo mã bưu điện, niên giám điện thoại, tổng điều tra công nghiệp, tổng điều tra dân số v.v...

Khung để chọn mẫu là danh sách các đơn vị trong tổng thể (hoặc vạn vật) trong đó một số đơn vị này sẽ được chọn để điều tra. Đó có thể là một danh sách thực, một bộ thẻ chỉ số, một bản đồ hoặc dữ liệu lưu trữ trong máy tính. Khung là một tập hợp các tài liệu thực (số liệu tổng điều tra, các bản đồ, các danh sách, các thư mục, các bản ghi) cho phép chúng ta nắm được vạn vật dần dần.

Những vấn đề tiềm tàng với khung chọn mẫu bao gồm: Khung chọn mẫu có thể không chính xác, không đầy đủ hoặc có sự nhân đôi. Do vậy, chúng ta cần phải có chiến lược thay thế ngẫu nhiên trong tầng.

- *Mẫu:* Là một phần danh sách hay nhóm các thành viên đại diện của một tổng thể có được từ các phương pháp lựa chọn khác nhau cho việc thu thập thông tin nghiên cứu.

*Ví dụ:* Cuộc điều tra thứ nhất:

Mục tiêu: Đánh giá thái độ của các bậc cha mẹ liên quan đến chương trình giới thiệu dinh dưỡng cho học sinh cấp II.

Tổng thể: Là toàn thể các bậc cha mẹ có con đang học cấp II tại địa bàn nghiên cứu.

Mẫu: 200 trong tổng số 500 bậc cha mẹ tại xã (lựa chọn ngẫu nhiên).

Cuộc điều tra thứ hai:

Mục tiêu chung: So sánh thói quen đọc sách của các sinh viên một trường đại học.

Tổng thể nghiên cứu: Toàn bộ sinh viên trong trường.

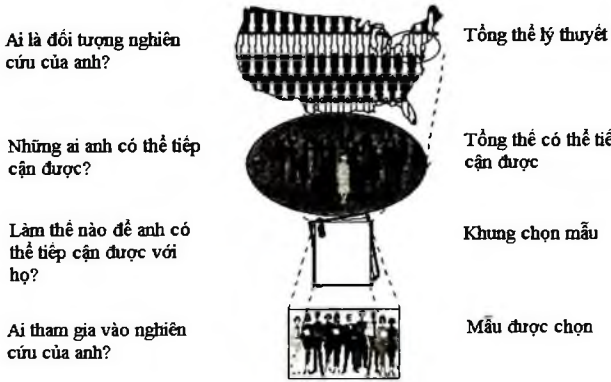
Mẫu: Lựa chọn ngẫu nhiên 300 sinh viên của trường đến mượn sách tại thư viện.

Như vậy, trong cuộc điều tra thứ nhất, 200 bậc cha mẹ sẽ là mẫu điều tra và câu trả lời của họ sẽ mang tính chất đại diện cho toàn bộ các bậc cha mẹ có con đang theo học trường cấp II của khu vực nghiên cứu. Còn trong cuộc điều tra thứ hai, câu trả lời của 300 sinh viên sẽ được xem như là đại diện cho tổng thể sinh viên của trường nếu như việc chọn mẫu của chúng ta là ngẫu nhiên.

Mẫu trong điều tra chọn mẫu hết sức quan trọng vì trong quá trình nghiên cứu chúng ta không có điều kiện để thu thập thông tin từ toàn bộ tổng thể, do vậy, chúng ta phải thu thập được số liệu từ những người càng mang tính đại diện cho nhóm mà chúng ta đang nghiên cứu càng tốt. Thậm chí với một bảng hỏi hoàn hảo (nếu như nó tồn tại), số liệu điều tra của chúng ta sẽ chỉ được coi là có ích nếu người được điều tra đại diện cho toàn bộ tổng thể nghiên cứu của chúng ta.

Một mẫu tốt là một thực thể thu nhỏ của tổng thể, hoàn toàn giống như tổng thể chỉ khác là nhỏ hơn. Một mẫu tốt nhất phải mang tính đại diện cho tổng thể về các đặc trưng quan trọng nhất của tổng thể đó.

Một ví dụ về chọn mẫu không tốt, đó là nếu chúng ta tiến hành một cuộc thăm dò về điều kiện chăm sóc sức khoẻ cho người dân mà lại chỉ tiến hành trên điện thoại, tức là khung chọn mẫu của ta là danh mục điện thoại, thì có nghĩa là chúng ta đã loại bỏ một phần lớn đối tượng người nghèo vì họ không có điện thoại hoặc khu vực vùng sâu, vùng xa nơi điện thoại chưa đến được với họ. Như vậy, kết quả của cuộc điều tra chưa thể phản ánh hết được thực trạng của công tác chăm sóc sức khoẻ người dân như chúng ta đã đặt ra trong mục tiêu ban đầu.




**Hình 1.3: TỔNG THỂ VÀ MẪU**

Một câu hỏi đặt ra ở đây là tại sao ta lại phải chọn mẫu? Sao ta không điều tra toàn bộ tổng thể nghiên cứu? Câu trả lời là:

Chọn mẫu điều tra giúp chúng ta thực hiện nhanh chóng hơn là điều tra tổng thể, chọn mẫu điều tra cũng giúp ta tiết kiệm được kinh phí cho các hoạt động khác như kiểm tra độ chính xác và chất lượng của thông tin thu thập được. Ngoài ra, chọn mẫu điều tra giúp chúng ta tập trung vào những nghiên cứu cụ thể hơn.

*Ví dụ:* Khi chúng ta muốn so sánh giữa những cặp vợ chồng trẻ và già của một nhóm dân tộc nào đó chẳng hạn, thì việc lựa chọn mẫu phân tầng sẽ cho ta tập trung vào vấn đề ta cần, do đó mà nó phù hợp hơn là việc điều tra cả tổng thể mẫu.

Khi tiến hành lựa chọn mẫu điều tra chúng ta cần phải đảm bảo rằng đó là một đại diện của tổng thể. Chúng ta biết không có một mẫu điều tra nào là hoàn hảo, vì nó luôn chứa đựng một sai số hay thành kiến nào đó. Danh mục các tiêu chí sau có thể được sử dụng để đảm bảo mẫu đặc trưng cho tổng thể và mức độ đại diện của mẫu.

Bản thân mẫu nghiên cứu hoàn toàn không có ý nghĩa gì, điều quan trọng của chúng là độ chính xác cho tổng thể mà chúng đại diện hay tấm gương của nhóm mục tiêu nghiên cứu.

*1.2.1.1. Danh mục các tiêu chí cho việc đảm bảo một mẫu có tính đại diện cho tổng thể:*

(1) Mục tiêu điều tra phải rõ ràng

Đây là lý do cho việc triển khai điều tra. Cuộc điều tra được tiến hành nhằm mô tả, so sánh hay tìm hiểu thái độ v.v... Một công ty có thể tiến hành điều tra người lao động trong

công ty của mình hay một trường học có thể điều tra những học sinh đang theo học để tìm hiểu những gì diễn ra trong hiện thực và quá khứ nhằm cải tiến kỹ thuật hay xây dựng những môn học mới cho tương lai.

Có thể xem xét lại ví dụ lần trước khi đánh giá thái độ của các bậc cha mẹ liên quan đến chương trình dinh dưỡng.

Mục tiêu chung: Đánh giá thái độ của các bậc cha mẹ liên quan đến chương trình giới thiệu ăn kiêng và dinh dưỡng cho học sinh cấp II.

Mục tiêu cụ thể: Nhằm mô tả và so sánh thái độ của các bậc cha mẹ ở các độ tuổi khác nhau, nhóm dân tộc khác nhau và với những hiểu biết khác nhau về kiến thức dinh dưỡng tới 3 mức dinh dưỡng được giới thiệu.

Câu hỏi nghiên cứu:

1. Các bậc cha mẹ sẽ có thái độ như thế nào khi được giới thiệu 3 mức dinh dưỡng cho con em họ?

2. Thái độ của các bậc cha mẹ ở những nhóm dân tộc khác nhau như thế nào khi tham gia chương trình này?

3. Liệu những bậc cha mẹ có am hiểu về kiến thức dinh dưỡng sẽ có thái độ khác với những người khác?

Mục tiêu điều tra sẽ là hướng dẫn cho việc triển khai các câu hỏi cụ thể của điều tra, hoặc các mục thông tin cần thu thập trong cuộc điều tra cũng như việc lựa chọn tổng thể và mẫu điều tra.

(2) Chỉ tiêu được lựa chọn phải rõ ràng và xác định được

Một chỉ tiêu cho việc xác định đặc điểm của người được lựa chọn cho cuộc điều tra là người đó có khả năng tham gia. Bên cạnh đó có chỉ tiêu để xác định những đối tượng sẽ không tham gia vào cuộc điều tra. Việc ứng dụng 2 chỉ tiêu này cho phép xác định rõ ràng đối tượng có thể lựa chọn làm mẫu để điều tra. Nếu một ai đó hay một nhóm nào đó không phù hợp với các chỉ tiêu để lựa chọn vào mẫu thì người đó và nhóm đó sẽ không nằm trong mẫu nghiên cứu.

*Ví dụ:* Câu hỏi nghiên cứu là tác động của cụm từ QUITNOW (dừng lại) trong việc giúp đỡ người hút thuốc lá có thể bỏ thuốc lá?

Tổng thể: Những người hút thuốc.

Các chỉ tiêu lựa chọn:

- Có tuổi nằm giữa 18 và 64.

- Hút hơn 1 điếu thuốc lá mỗi ngày.

- Có nhiều vết đen trong phổi khi chụp phim.

Chỉ tiêu loại trừ: Nếu người đó thuộc nhóm cấm chỉ định cho việc sử dụng các chất nicotin.

Kết quả: Cuộc điều tra chỉ tiến hành đối với những người đủ tư cách. Nếu những ai có tuổi ít hơn 18 và nhiều hơn 64 sẽ không được lựa chọn làm mẫu điều tra. Mặc dù, tổng thể nghiên cứu là những người hút thuốc lá, song chỉ tiêu lựa chọn và chỉ tiêu loại trừ sẽ giúp xác định rõ hơn nhóm nghiên cứu "ai là người hút thuốc lá".

Việc ứng dụng các chỉ tiêu trên giúp chúng ta xác định rõ ràng ranh giới cho những người được phỏng vấn ai là phù hợp và có đủ tư cách tham gia vào mẫu đại diện cho tổng thể nghiên cứu.

(3) Lựa chọn phương pháp chọn mẫu khắt khe (chọn mẫu ngẫu nhiên)

Các phương pháp chọn mẫu được chia làm hai nhóm: chọn mẫu ngẫu nhiên và phi ngẫu nhiên.

Lựa chọn ngẫu nhiên cung cấp các thông tin thống kê về tính đại diện mẫu của tổng thể. Trong chọn mẫu ngẫu nhiên, mỗi một cá thể được xác định là có cùng xác suất lựa chọn làm mẫu điều tra.

Lựa chọn phi ngẫu nhiên là sự lựa chọn mẫu phụ thuộc vào đặc tính của tổng thể và nhu cầu của điều tra. Với cách thức này một vài cá thể của tổng thể có cơ hội cao hơn được lựa chọn làm mẫu điều tra, trong khi đó những cá thể khác lại không có cơ hội cao. Như vậy, khả năng ứng dụng những kết quả điều tra nhằm suy rộng cho cả tổng thể có thể không áp dụng được.

*1.2.1.2. Các phương pháp chọn mẫu ngẫu nhiên*

Có nhiều cách lựa chọn mẫu thống kê trong điều tra chọn mẫu, mỗi cách lựa chọn phụ thuộc vào điều kiện cụ thể về tính đại diện, độ tin cậy và đặc trưng cụ thể của tổng thể. Về cơ bản chúng ta có thể phân ra thành các loại mẫu thống kê như sau:

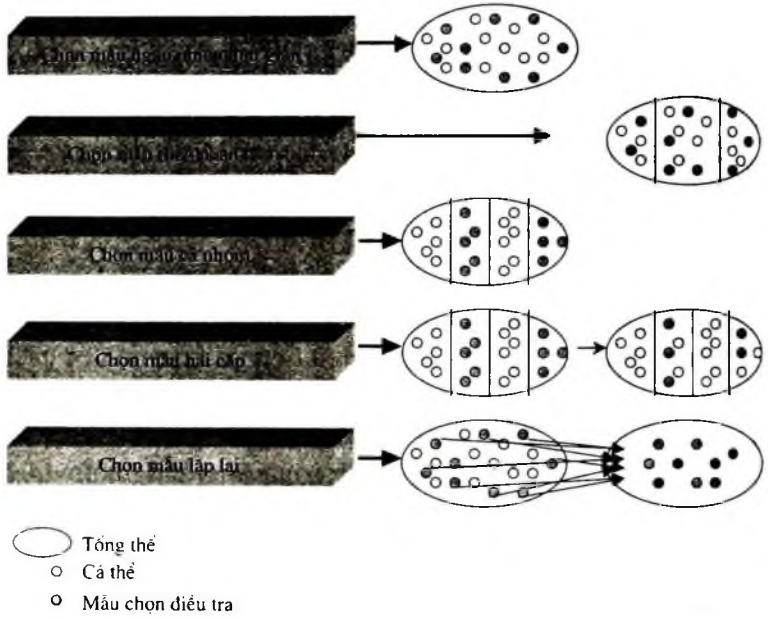
(1) Mẫu ngẫu nhiên đơn giản.

(2) Chọn mẫu theo phân nhóm/tầng.

(3) Chọn mẫu cả nhóm.

(4) Chọn mẫu hai cấp.

(5) Chọn mẫu lặp lại.




**Hình 1.4: CHỌN MẪU NGẪU NHIÊN**

Mỗi cách chọn mẫu lại tiến hành khác nhau, nó tuỳ thuộc vào thực tế tính chất của tổng thể như đã đề cập. Nếu tổng thể là đồng nhất thì việc lựa chọn mẫu theo cách thứ nhất là đơn giản nhất và tính đại diện cao nhất.

Tuy nhiên, khi bên trong tổng thể của chúng ta có sự khác biệt và phân chia thành nhiều nhóm có đặc trưng khác nhau thì phương pháp thứ nhất sẽ không đảm bảo cho ta có một mẫu mang tính đại diện nhất như ta mong muốn. Vì vậy, trong trường hợp này chúng ta cần phải thay đổi cách thức lựa chọn mẫu điều tra cho phù hợp. Lúc này các cách chọn mẫu thứ 2, 3 hoặc 4 sẽ phù hợp hơn, còn việc cụ thể để lựa chọn theo cách nào lại phụ thuộc vào nhiều yếu tố khác như điều kiện thời gian, kinh phí, v.v...

Ngoài ra, cách lựa chọn mẫu thứ 5 là khi chúng ta muốn nghiên cứu sâu hơn vào một vấn đề gì đó hay chúng ta muốn kiểm tra lại thông tin sau khi có kết quả điều tra chúng ta có thể lựa chọn một nhóm nhỏ trong mẫu để tiến hành điều tra lại theo một số chỉ tiêu hẹp nào đó.

Mặc dù chúng ta có nhiều cách để lựa chọn mẫu điều tra khác nhau nhưng vẫn phải bảo đảm nguyên tắc là ở bước cuối cùng trong phương pháp chọn mẫu phải chọn mẫu ngẫu nhiên, khi đó mẫu điều tra mới mang tính đại diện cho tổng thể (tức là để có thể SUY DIỄN về tổng thể dựa trên việc phân tích mẫu điều tra).

* *Mẫu ngẫu nhiên đơn giản:* Được lựa chọn trên cơ sở nguyên tắc là xác suất lựa chọn mẫu là như nhau. Mẫu được lựa chọn theo phương pháp này đáp ứng cao về mặt thống kê và đặc biệt khả năng suy rộng kết quả là rất cao và áp dụng các công cụ toán học trong tính toán dễ dàng hơn.

Bước đầu tiên trong chọn mẫu ngẫu nhiên đơn giản là phải có danh sách của tổng thể để từ đó xác định được mẫu

hay nói cách khác chúng ta phải có khung chọn mẫu. Nếu mẫu muốn đủ đại diện cho tổng thể thì khung chọn mẫu phải chứa đựng tất cả hoặc gần hết tất cả các thành viên trong tổng thể. Trong chọn mẫu ngẫu nhiên đơn giản tất cả các cá thể của tổng thể có cùng một cơ hội để lựa chọn. Các cá thể được lựa chọn cùng một thời gian và hoàn toàn độc lập với nhau. Khi một mẫu đã được lựa chọn sẽ không được phép để trong khung chọn mẫu nữa, hay nói một cách khác là nó không có cơ hội để lựa chọn lại. Chính vì có cùng một cơ hội nên mẫu ngẫu nhiên được xem như là không có sự sai khác.

Một đặc điểm nổi bật của chọn mẫu ngẫu nhiên đơn giản đó là tránh được sự sai lệch của mẫu, tuy nhiên điểm bất lợi là nhiều khi nó không tính toán hay bao hàm được những thành phần của tổng thể mà chúng ta đang quan tâm. Giả sử khi chúng ta muốn tìm hiểu về sự thoả mãn của các khách hàng và giả thiết chúng ta có kết quả của một nghiên cứu trước đây cho rằng những khách hàng già và trẻ thường là các nhóm khác nhau về mức độ thoả mãn của một dịch vụ nào đó. Nếu chúng ta áp dụng phương pháp chọn mẫu ngẫu nhiên đơn giản cho một nghiên cứu mới có thể chúng ta sẽ có mẫu trong đó số lượng cá thể của nhóm khách hàng trẻ không đủ lớn để áp dụng các công cụ thống kê, trong trường hợp này chọn mẫu ngẫu nhiên đơn giản không bao hàm được toàn bộ các thành phần của tổng thể.

Như vậy, nếu các thành phần của tổng thể đa dạng hay nói cách khác một tổng thể nào đó không đồng nhất và với những mục đích nghiên cứu khác nhau do vậy mà phương pháp chọn mẫu ngẫu nhiên đơn giản chỉ phù hợp cho việc áp

dụng trong những tổng thể tương đối nhỏ khi đó tính đồng đều sẽ tăng lên bên trong của tổng thể đó.

Mẫu ngẫu nhiên đơn giản được lựa chọn một cách ngẫu nhiên từ tổng thể. Có nhiều cách khác nhau, ví dụ có thể lựa chọn hoàn toàn ngẫu nhiên không theo một quy luật nào hoặc cũng có thể lựa chọn dựa trên danh sách và có khoảng cách cố định giữa các mẫu.

Việc lựa chọn này phải đảm bảo rằng nó không phụ thuộc vào ý chủ quan của người nghiên cứu.

Đặc điểm của chọn mẫu ngẫu nhiên đơn giản là:

- Sử dụng cơ chế cơ hội để lựa chọn các quan sát.

- Biết xác suất lựa chọn cho từng mẫu.

- Tất cả đều dựa trên khung chọn mẫu.

Có các cách để lấy một mẫu theo hình thức ngẫu nhiên đơn giản đó là:

(a) Hệ thống:

- Đánh số các đơn vị trong tổng thể từ 1 đến N quyết định số lượng n cá thể trong tổng thể (quy mô mẫu) mà chúng ta muốn hoặc cần.

- k = N/n = quy mô khoảng tin cậy lựa chọn ngẫu nhiên một số nguyên nằm giữa 1 và k, sau đó chọn các đơn vị đứng thứ k là mẫu cho thu thập thông tin.

(b) Số ngẫu nhiên:

- Ném đồng xu.

- Sử dụng công cụ Excel bằng công thức rand()

- Hay sử dụng các phần mềm thống kê khác như: Stata bằng câu lệnh lựa chọn mẫu ngẫu nhiên đơn giản: gen randomnr=uniform()

*Ví dụ:* Chọn các hộ gia đình trong điều tra 3.700 hộ gia đình cho dự án của MOLISA/UNDP về đánh giá chương trình xoá đói giảm nghèo.

Các cách này cũng được sử dụng trong các phương pháp chọn mẫu ngẫu nhiên được trình bày dưới đây như phân tầng và nhiều chặng (lặp lại).

(c) Chọn mẫu không tỷ lệ: chọn mẫu với xác suất lựa chọn một đơn vị tỷ lệ với quy mô (pps).

*Ví dụ:* Chọn các tỉnh và xã trong cuộc điều tra 3.700 hộ gia đình cho dự án của MOLISA/UNDP về đánh giá chương trình xoá đói giảm nghèo HERP.

* *Chọn mẫu theo phân tầng (cấp):* Khi tổng thể có thể chia thành nhiều nhóm có đặc điểm tương đối đồng đều trong từng nhóm thì ta có thể sử dụng phương pháp này. Trước hết chúng ta cần chia tổng thể theo các nhóm khác nhau ví dụ theo loại hình dân tộc, hay giới, sau đó trong mỗi nhóm đó ta sẽ tiến hành lấy mẫu ngẫu nhiên để điều tra.

Cơ sở thống kê cho việc phân tầng: Là dựa vào tính chính xác của bất kỳ một ước lượng nào đó có thể được cải thiện bằng cách lựa chọn một thiết kế thích hợp. Người điều tra

thường có hiểu biết về tổng thể đang nghiên cứu trước khi tiến hành điều tra và việc sử dụng những thông tin này có thể cải thiện tính hiệu quả của suy diễn thống kê về những đặc điểm chưa biết của tổng thể.

Thường phương pháp này hay được sử dụng khi trong tổng thể có sự khác biệt về tính chất của các nhóm.

Hai vấn đề tiềm tàng: (1) Biến phân tầng không được biết trước khi điều tra.

*Ví dụ:* Ước lượng sản xuất công nghiệp trong ngành dệt/may. Phân tầng theo mức sản xuất hiện tại. Nhưng mức sản xuất hiện tại không được biết. Sử dụng đại diện: quy mô lao động trong quá khứ.

(2) Các quy mô cơ bản của tổng thể không được biết.

*Ví dụ:* Chỉ có danh sách của các doanh nghiệp với tên và địa chỉ, không có quy mô.

Chọn mẫu phân tầng phức tạp hơn rất nhiều so với chọn mẫu ngẫu nhiên đơn giản, các tầng phải được xác định, sắp xếp đều nhau và nếu như sử dụng nhiều tầng sẽ dẫn đến mẫu lớn, cồng kềnh và tốn kém cho điều tra.

* *Chọn mẫu cả nhóm:* Thường được sử dụng khi tổng thể có sự khác biệt về những đặc trưng như địa: điểm, hay được phân bố thành những cụm có khoảng cách khác nhau xa về mặt địa lý, khi đó nếu chọn mẫu theo phương pháp ngẫu nhiên

đơn giản hay theo phân cấp sẽ dẫn đến tốn kém trong điều tra trong khi kinh phí dành cho nghiên cứu không cho phép.

Để tiến hành chọn mẫu cả nhóm, trước hết chúng ta cần phân chia tổng thể mẫu ra thành nhiều nhóm khác nhau, sau đó ta tiến hành lựa chọn toàn bộ một số nhóm làm mẫu để điều tra.

Theo cách thức này có một điểm rất hạn chế là nếu như mỗi nhóm có số lượng lớn các cá thể thì sẽ dẫn đến việc tốn kém hơn cả về tài chính và thời gian để tiến hành điều tra thông tin.

Chú ý:

- Trong việc chọn mẫu cả nhóm chúng ta không cần danh sách các cá thể (doanh nghiệp/hộ gia đình) ở tất cả các vùng, mà chỉ ở những vùng được chọn.

- Việc lựa chọn các vùng hoặc các doanh nghiệp/hộ gia đình có thể được thực hiện thông qua việc sử dụng bất kỳ một phương pháp chọn mẫu nào.

* *Chọn mẫu hai cấp:* Là dạng hay được ứng dụng nhất vì nó kết hợp tính tối ưu của cả 3 cách thức chọn mẫu đơn giản, phân cấp và chọn mẫu cả nhóm nêu trên, do vậy nó thường phù hợp nhất, đặc biệt trong điều kiện các nghiên cứu của ta với tổng thể có sự phức tạp như hiện nay.

Để tiến hành chọn mẫu theo cách thức này trước hết tổng thể sẽ được phân chia ra thành nhiều cấp hay nhóm khác nhau, sau đó tiến hành lựa chọn một số nhóm đại diện cho đối

tượng nghiên cứu và để đảm bảo về mặt thời gian chúng ta tiến hành lựa chọn trong các nhóm đó một số cá thể đại diện bằng phương pháp lựa chọn mẫu ngẫu nhiên đơn giản như đã trình bày ở phần trước.

* *Chọn mẫu lặp lại:* Với cách thức chọn mẫu này, trước hết chúng ta tiến hành chọn mẫu ngẫu nhiên đơn giản lần 1, sau đó từ những mẫu đã được lựa chọn đó chúng ta lại tiến hành lựa chọn một số lượng nhỏ (tuỳ theo yêu cầu của công việc) các mẫu một cách cũng hoàn toàn ngẫu nhiên để phục vụ cho mục tiêu nhất định trong nghiên cứu như: kiểm tra lại thông tin hay bổ sung thêm thông tin v.v...

**Bảng 1.1:** NHỮNG THUẬN LỢI KHÓ KHĂN, ƯU NHƯỢC ĐIỂM CÁC PHƯƠNG PHÁP CHỌN MẪU ĐIỀU TRA

| | | |
|---|---|---|
| Ngẫu nhiên đơn giản | Dễ làm, tính khách quan cao.<br>Nhanh dễ làm, trình độ cán bộ đòi hỏi không cao. | Không áp dụng được cho tất cả các trường hợp khi không có tính đồng đều trong tổng thể, chỉ sử dụng cho những mẫu nhỏ. |
| Chọn mẫu phân cấp | Dễ làm, phù hợp với mục tiêu điều tra, thời gian nhanh hơn.<br>Độ chính xác cao, chọn đối tượng theo mục đích điều tra.<br>Tính đại diện cao hơn. | Có thể bị trùng lặp.<br>Phải xác định được tiêu chí phân nhóm trước khi điều tra.<br>Chi phí cao hơn. |
| Chọn mẫu cả nhóm | Độ chính xác cao.<br>Tính đại diện cao, áp dụng cho tổng thể mẫu lớn.<br>Số lượng mẫu nhỏ. | Tốn nhiều thời gian và chi phí.<br>Có thể ảnh hưởng đến kết quả điều tra do đặc thù của các nhóm được chọn. |

*(Tiếp theo)*

| | | |
|---|---|---|
| Chọn mẫu 2 cấp | Dễ làm, tốn ít thời gian và chi phí, độ chính xác cao hơn các phương pháp trên, áp dụng cho các tổng thể mẫu lớn, chia nhỏ theo từng cấp.<br>Tổng hợp ưu điểm của 2 phương pháp trên. | Tổng thể phải lớn.<br>Điều tra viên phải có trình độ cao.<br>Tốn nhiều thời gian. |
| Chọn mẫu lặp lại | Dễ so sánh kết quả điều tra.<br>Tiêu tốn ít thời gian do có sẵn khung chọn mẫu.<br>Có độ chính xác cao, tránh được sai sót của điều tra trước.<br>Dùng để kiểm chứng kết quả các cuộc điều tra lớn. | Tốn chi phí.<br>Dễ bị trùng lặp.<br>Phải có kết quả của cuộc điều tra trước do vậy bị phụ thuộc.<br>Đòi hỏi trình độ cao.<br>Có công cụ đủ mạnh.<br>Tính đại diện không cao. |

### *1.2.2. Chọn mẫu phi thống kê trong điều tra chọn mẫu*

Ngoài cách chọn mẫu thống kê như đã trình bày phần trước chúng ta cũng có thể lựa chọn mẫu phi thống kê theo các cách như sau:

*- Chọn mẫu tiện lợi:*

Chọn mẫu tiện lợi được sử dụng khi chúng ta đơn giản chỉ chặn đường một ai đó trên phố mà họ đang định dừng lại, hoặc khi chúng ta đi quanh một doanh nghiệp, một cửa hàng, một quán ăn, một rạp hát v.v... hỏi những người mà chúng ta gặp xem liệu họ sẽ trả lời câu hỏi của chúng ta hay không. Nói một cách khác, mẫu bao gồm những chủ thể mà nhà nghiên

cứu có thể tiếp cận một cách thuận lợi. Không có sự lựa chọn ngẫu nhiên và khả năng chệch là lớn.

*- Chọn mẫu theo quota:*

Chọn mẫu theo quota thường được sử dụng trong nghiên cứu thị trường. Những người đi phỏng vấn phải tìm kiếm những trường hợp có những đặc tính nhất định. Họ được nhận quota của những nhóm người nhất định để phỏng vấn và quota được tổ chức theo cách làm cho mẫu cuối cùng đại diện cho tổng thể.

Nhược điểm của chọn mẫu theo quota: Người phỏng vấn lựa chọn ai mà họ thích (trong phạm vi tiêu chí như trên) và do vậy có thể lựa chọn những người dễ phỏng vấn nhất và vì thế có thể gây ra chệch mẫu. Ngoài ra, không thể ước lượng được tính chính xác (do mẫu không ngẫu nhiên).

*- Chọn mẫu có mục đích:*

Mẫu có mục đích là mẫu được nhà nghiên cứu chọn một cách chủ quan. Nhà nghiên cứu cố gắng chọn mẫu mà theo họ là mang tính đại diện cho tổng thể và sẽ cố gắng đảm bảo rằng mẫu bao gồm tất cả các khía cạnh của tổng thể nghiên cứu.

*- Mạng lưới hoặc "ném tuyết":*

Với cách tiếp cận này, đầu tiên chúng ta liên hệ với một vài người trả lời tiềm năng và sau đó hỏi xem liệu họ có biết ai đó có cùng đặc tính mà chúng ta đang muốn nghiên cứu hay không.

*- Tự lựa chọn:*

Có lẽ bản thân cụm từ "tự lựa chọn" đã tự nói lên ý nghĩa của nó. Bản thân người được hỏi sẽ tự quyết định xem họ có thích tham gia vào cuộc điều tra hay không và nếu họ không muốn tham gia chúng ta sẽ phải chuyển sang đối tượng khác để hỏi.

*- Chọn mẫu chuyên gia:*

Chọn mẫu chuyên gia liên quan đến chọn một mẫu bao gồm những người đã được biết là có kinh nghiệm và chuyên môn trong một lĩnh vực nào đó. Thường chúng ta thu xếp một mẫu như vậy dưới danh nghĩa là "một nhóm chuyên gia".

Việc tiến hành chọn mẫu phi thống kê tuỳ thuộc vào mục tiêu, đối tượng và yêu cầu của nghiên cứu. Phương pháp này thường được áp dụng trong những điều kiện chúng ta không có khung chọn mẫu cụ thể, chẳng hạn như trong các nghiên cứu thị trường chúng ta không có danh sách khách hàng mua một loại hàng hoá nào đó thì chúng ta buộc phải sử dụng phương pháp chọn mẫu phi thống kê. Phương pháp này cũng có thể áp dụng khi tổng thể của chúng ta có tính đồng nhất cao (mà điều này thường khó diễn ra đối với các vấn đề kinh tế - xã hội), hay trong trường hợp nghiên cứu của chúng ta không cần phải ngoại suy cho tổng thể.

Do vậy, ta có thể thấy điểm hạn chế của phương pháp chọn mẫu phi thống kê là khả năng áp dụng các công cụ thống kê và khả năng suy rộng của kết quả bị hạn chế.

*- Nhóm quan tâm:*

Phương pháp này thường hay được dùng trong nghiên cứu thị trường nhằm tìm hiểu những mặt hàng cụ thể mà xã hội cần và sẽ tiêu dùng. Để nghiên cứu chúng ta thường điều tra 10-20 người cùng mua một mặt hàng nào đó để đại diện cho nhóm những người có cùng sở thích hoặc nhóm khách hàng tiềm năng.

*Ví dụ:* Khi điều tra về nhóm bệnh nhân bị bệnh tiểu đường. 12 bệnh nhân được mời tham dự với những câu hỏi như sau: liệu phiếu điều tra đã bao hàm toàn bộ những câu hỏi cần thiết? Các anh (chị) có thể theo dõi và trả lời một cách dễ dàng các câu hỏi? Mất bao nhiêu thời gian để hoàn chỉnh một phiếu điều tra như vậy? Kết quả của cuộc thảo luận nhóm sẽ giúp ta chỉnh sửa lại phiếu điều tra để tiến hành một cuộc điều tra với quy mô lớn đối với các bệnh nhân tiểu đường.

Phương pháp nhóm quan tâm có kết quả tương đối chính xác vì trong trường hợp nếu nhóm tham gia có mức độ khác biệt lớn với tổng thể, chẳng hạn những người có trình độ cao trong một tổng thể có trình độ ở mức trung bình, thì câu trả lời của họ có thể không ứng dụng được cho tổng thể đó.

Tóm lại, với mỗi phương pháp, cách thức chọn mẫu khác nhau đều có những lợi ích và chứa đựng những vấn đề khác nhau như trình bày tại bảng 1.2 dưới đây:

**Bảng 1.2: LỢI ÍCH VÀ NHỮNG VẤN ĐỀ ĐẶT RA ĐỐI VỚI MỖI MỘT PHƯƠNG PHÁP ĐIỀU TRA ĐƯỢC LỰA CHỌN**

| Phương pháp chọn mẫu | Lợi ích | Vấn đề |
|---|---|---|
| *Chọn mẫu ngẫu nhiên* | | |
| *Chọn mẫu ngẫu nhiên đơn giản* (tất cả các cá thể đều có cùng cơ hội được lựa chọn). | Tiến hành tương đối đơn giản . | Thành viên của các nhóm cá thể khác nhau trong tổng thể có thể không xuất hiện trong mẫu với một tỷ lệ phù hợp. |
| *Chọn mẫu theo phân tầng (cấp)*<br>Tổng thể nghiên cứu được nhóm lại theo các nhóm khác nhau với những chỉ tiêu có ý nghĩa cho nghiên cứu. | Có thể tiến hành phân tích theo từng nhóm (VD: theo giới, tuổi, khu vực v.v...)<br>Mức độ biến động thấp hơn so với cách chọn mẫu ngẫu nhiên giản đơn.<br>Mẫu có tính chất đại diện hơn cho tổng thể. | Phải tính toán số lượng mẫu cho mỗi nhóm.<br>Có thể phải tiêu tốn hơn về mặt thời gian và kinh phí cho việc tiến hành chuẩn bị điều tra. |
| *Chọn mẫu cả nhóm*<br>Tổng thể nghiên cứu có nhiều nhóm khác nhau (về mặt địa lý). Việc lựa chọn sẽ được tiến hành với một số nhóm nhất định. | Thuận lợi trong trường hợp xa nhau về mặt địa lý. | Nếu như mỗi nhóm có số lượng lớn thì tốn kém hơn cả về tài chính và thời gian để tiến hành điều tra thông tin. |
| *Chọn mẫu hai cấp* | Phù hợp nhất trong điều kiện xã hội có sự phức tạp, phân chia theo nhiều nhóm (có những đặc trưng riêng). | |

*(Tiếp theo)*

| Phương pháp chọn mẫu | Lợi ích | Vấn đề |
|---|---|---|
| *Chọn mẫu ngẫu nhiên* | | |
| *Chọn mẫu lặp lại* | Phù hợp trong trường hợp cần kiểm tra độ chính xác công tác điều tra, hay thu thập thêm những thông tin cụ thể khác cho nghiên cứu. | Tốn kém thời gian và kinh phí. |
| *Chọn mẫu phi ngẫu nhiên* | | |
| *Chọn mẫu tiện lợi* | Một phương pháp mang tính thực tế bởi vì việc lựa chọn mẫu luôn có sẵn (VD: một sinh viên trong trường, một bệnh nhân đang trong phòng chờ khám bệnh). | Bởi vì mẫu là một cơ hội và tình nguyện do vậy mà nó đôi khi không giống như những cá thể khác trong tổng thể nghiên cứu. |
| *Chọn mẫu theo quota* (Tổng thể được chia thành từng nhóm nhỏ theo các chỉ tiêu khác nhau như giới, tuổi, v.v...<br>Mẫu sẽ được lựa chọn theo những tỷ lệ nhất định của các nhóm trong tổng thể). | Có thể hiện thực nếu như phần số liệu có sẵn mô tả tỷ lệ của các nhóm. | Các số liệu đã có phải luôn được cập nhật để có tỷ lệ chính xác. |

*(Tiếp theo)*

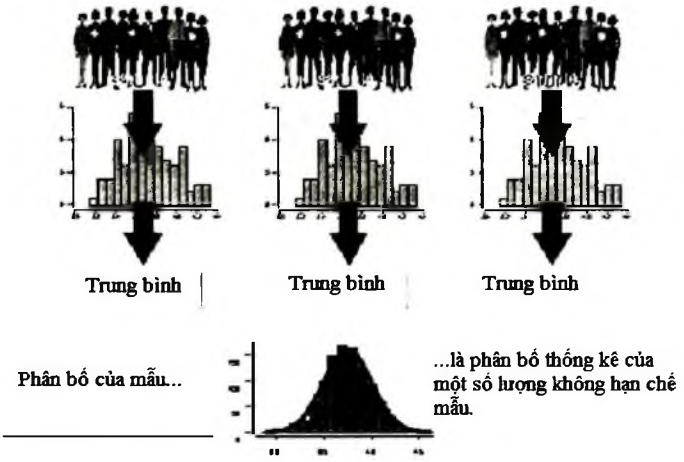
| Phương pháp chọn mẫu | Lợi ích | Vấn đề |
|---|---|---|
| *Chọn mẫu phi ngẫu nhiên* | | |
| *Mạng lưới hoặc 'ném tuyết"*<br>Những người trả lời trước sẽ chỉ định những người tiếp theo trong tổng thể. | Thích hợp trong điều kiện không có khung chọn mẫu. | Việc lựa chọn này có thể dẫn đến những sai sót chọn mẫu.<br>Không thể kiểm tra được ai là người sẽ được tham gia. |
| *Tự lựa chọn* | Phù hợp với các nghiên cứu thuộc dạng thị trường, hay đối với những nhóm khó tiếp cận. | Có thể chứa đựng những sai sót chọn mẫu và tính đại diện. |
| *Chọn mẫu chuyên gia* | Phù hợp cho các nghiên cứu chuyên sâu hay việc tham khảo kinh nghiệm cho những vấn đề lý luận nhà nghiên cứu đưa ra. | |
| *Nhóm quan tâm* | Phù hợp trong việc định hướng cho việc phát triển điều tra. | Phải là những nhóm tương đối nhỏ nhưng mang tính đại diện cho cả tổng thể lớn hơn. |

## 1.3. Quy mô mẫu trong điều tra chọn mẫu

### *1.3.1. Phân phối mẫu*

Chúng ta đi từ thống kê mẫu đến ước lượng tham số tổng thể như thế nào? Một khái niệm trung gian quan trọng mà chúng ta cần phải hiểu là *phân phối mẫu.*

Phân phối của một số vô hạn các mẫu có cùng quy mô như mẫu trong nghiên cứu của chúng ta được gọi là *phân phối mẫu.*




**Hình 1.5: PHÂN PHỐI MẪU**

Tư tưởng chính của thống kê suy rộng là lấy mẫu từ một tổng thể và sau đó sử dụng kết quả phân tích các thông tin từ mẫu này để suy rộng ra cho tổng thể nghiên cứu. Ví dụ, giá trị bình quân (giá trị trung tâm), độ lệch chuẩn (mức độ dao động hay biến động), hoặc là tỷ lệ của một số quan sát/tổng thể về một đặc trưng nào đó. Việc lấy mẫu nghiên cứu sẽ giúp chúng ta tiết kiệm kinh phí, thời gian và cả những công sức phải bỏ ra. Hơn thế nữa, lấy mẫu đôi khi cung cấp các thông tin chính xác cho nghiên cứu hơn là câu trả lời của việc chúng ta cố

gắng điều tra cả tổng thể (sai số phi chọn mẫu), nghiên cứu cẩn thận một mẫu còn hơn là làm không cẩn thận với cả tổng thể.

Chúng ta sẽ xem xét ti mi những đặc điểm của mẫu từ các tổng thể khác nhau. Bởi vì mẫu là một nhóm đối tượng của một tổng thể, giá trị trung bình của mẫu không hoàn toàn chính xác như là của tổng thể. Vì vậy, một điều quan trọng cần phải xem xét đó là mức độ phù hợp của những ước lượng từ mẫu như giá trị bình quân so với tổng thể.

Thông thường trong thực tế, một mẫu rất nhỏ (5-10 quan sát) được lấy ra để kiểm tra cơ chế thu thập thông tin và từ đó thu được thông tin ban đầu cho việc chọn mẫu. Tuy nhiên phục vụ cho việc xác định mức độ phù hợp, chấp nhận được giữa ước lượng của mẫu so với tổng thể chúng ta cần phải xem xét với khoảng 10, 50 hoặc 100 mẫu riêng biệt khác nhau lấy ra từ tổng thể. Liệu sự phù hợp sẽ như thế nào nếu giữa các mẫu nghiên cứu khác nhau? Nếu chúng ta phát hiện rằng kết quả giữa các mẫu gần như giống nhau (và gần chính xác!), vậy chúng ta tin cậy vào một nghiên cứu độc lập hay không? Mặt khác, xem xét kết quả từ các nghiên cứu lặp lại cho một số tiêu chí nào đó cần có độ tin cậy cao hơn, đòi hỏi phải có một mẫu khác với cỡ mẫu lớn hơn.

Một phân phối mẫu được sử dụng để mô tả sự phân bố của những kết quả đầu ra, mà một nghiên cứu có thể thu được từ các mẫu tương tự của một tổng thể. Lưu ý rằng một giá trị bình quân ước lượng từ một mẫu có thể khác với một mẫu khác.

Cần phải hiểu rằng mỗi nghiên cứu thống kê khác nhau có một phân phối mẫu khác nhau, nó phụ thuộc vào những thông tin cụ thể, cỡ mẫu và phân phối của tổng thể. Và chúng ta cần phải lưu ý mối quan hệ giữa cỡ mẫu và phân phối của ước lượng của mẫu. Vì thế, mức độ biến động của phân phối mẫu có thể được thu hẹp lại bằng cách tăng số lượng quan sát của mẫu. Lưu ý khi cỡ mẫu lớn, nhiều phân phối mẫu sẽ tiệm cận với phân phối chuẩn.

Những moment chính của phân phối mẫu:

Moment bậc 1: trung bình của phân phối mẫu - trung bình của các trung bình của một số vô hạn các mẫu - rất gần với trung bình tổng thể - tham số cần quan tâm.

Moment bậc 2: độ lệch chuẩn của phân phối mẫu cho chúng ta biết các mẫu khác nhau có phân phối như thế nào. Trong thống kê, nó được gọi là sai số chuẩn.

$$V(estimate) = \frac{S^2}{n} * \left(1 - \frac{n}{N}\right) \quad (1)$$

Trong đó: n và N - lần lượt là quy mô của mẫu và tổng thể, $S^2$ - phương sai của biến.

Nếu mẫu nhỏ, điều chỉnh tổng thể hữu hạn gần bằng 1. Khi đó, phương sai của các đại lượng ước lượng phụ thuộc vào: (i) số lượng quan sát (n) và (ii) biến thiên của biến $S^2$.

Hàm ý của (1) đối với điều tra tổng thể rất lớn trong đó quy mô mẫu ít hơn 10% của tổng thể: (1) thường bị bỏ qua.

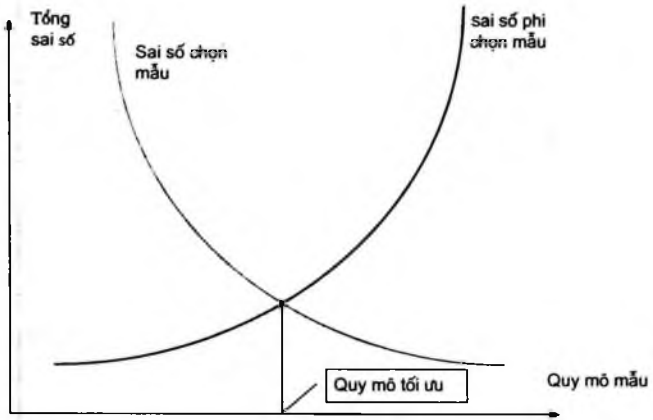
### *1.3.2. Sai số chọn mẫu và phi chọn mẫu*

Một mẫu tốt là một mô hình nhỏ có tính đại diện đầy đủ cho tổng thể. Tuy nhiên, sai số hay sai lầm chọn mẫu là những điều khó tránh khỏi trong bất kỳ một cuộc điều tra chọn mẫu nào.

Giả sử chúng ta muốn làm một nghiên cứu về nhu cầu chăm sóc sức khoẻ tinh thần cho những đứa trẻ vô gia cư. Một vấn đề mà chúng ta cần phải đề cập đến là qua thời gian cần thiết cho cuộc điều tra, nhu cầu có thể sẽ thay đổi bởi vì thực tế lịch sử. Một chính sách sức khoẻ mới có thể ra đời trong thời gian diễn ra cuộc điều tra của chúng ta. Những điều này sẽ dẫn đến những sai sót hay sai lầm khi chọn mẫu và điều tra chọn mẫu. Do vậy, nó đòi hỏi chúng ta phải tính toán đến tất cả những tình huống xảy ra và phải thực sự am hiểu thực tế, lịch sử, các vấn đề xã hội, môi trường đã, đang diễn ra và cả yếu tố thời gian nữa trong bất kỳ một cuộc điều tra nào.

Trong chọn mẫu, sai số chuẩn được gọi là *sai số chọn mẫu*. Sai số chọn mẫu cho ta biết độ chính xác của ước lượng thống kê mà chúng ta tính toán ra được từ mẫu điều tra hay nói cách khác đó là sai số *do việc chọn mẫu* (*vấn đề này* xuất hiện khi một phần của tổng thể được sử dụng để đại diện cho toàn bộ tổng thể và nó có thể đo lường được về mặt toán học).

Tính chính xác nói chung của ước lượng của chúng ta còn phụ thuộc vào *sai số phi chọn mẫu* mà nó có thể xảy ra ở bất kỳ chặng nào của cuộc điều tra.




**Hình 1.6: ĐỒ THỊ SAI SỐ CHỌN MẪU VÀ PHI CHỌN MẪU**

Các sai số chọn mẫu chủ yếu do quá trình chọn mẫu, hẳng hạn như việc áp dụng các phương pháp chọn mẫu phi gẫu nhiên. Sai số chọn mẫu là nguy hiểm vì nó có thể làm ỏng tính tin cậy của cuộc điều tra.

Cách tốt nhất để tránh những sai lầm chọn mẫu là sử dụng ác phương pháp chọn mẫu ngẫu nhiên. Trong những trường ợp không thể, chúng ta phải lựa chọn mẫu phi ngẫu nhiên ·ong những tổng thể ít có sự khác biệt xác suất giữa các cá ıể hoặc tối thiểu là đối với các chỉ tiêu chính của nghiên cứu. )ể xác định được những nhóm trong tổng thể ít có sự khác iệt về những thông tin của các chỉ tiêu chính này chúng ta có

thể thu thập được thông qua những nghiên cứu trước đây hoặc qua số liệu thống kê.

*Ví dụ:* Giả sử chúng ta đang nghiên cứu về nhóm phụ nữ có thu nhập thấp, tham gia một dự án tăng cường sử dụng dịch vụ chăm sóc sức khoẻ trước khi sinh. Nếu không có sự so sánh số liệu chúng ta không thể biết được mức độ sai số của mẫu, mặc dù chúng ta luôn biết là nó tồn tại. Nếu mức độ sử dụng dịch vụ tăng lên chúng ta không thể khẳng định được đó là do tác động của chương trình. Người phụ nữ tham gia chương trình có thể được thúc đẩy tìm kiếm sự chăm sóc hơn so với những người không tham gia. Những thông tin so sánh cần thiết có thể có sẵn trong những ấn phẩm in ấn trước đây. Với cách này, chúng ta có thể có những thông tin cơ bản giống tương tự để xác định những nhóm khác nhau cho việc lựa chọn mẫu.

Các loại sai lầm phi chọn mẫu thường gặp bao gồm:

Sai lầm không quan sát được, nó có nghĩa hoặc là:

- Không bao hàm: tức là có thể không bao hàm một số đơn vị, hoặc một số nhóm của tổng thể điều tra đã xác định trong khung cơ sở chọn mẫu được sử dụng trong thực tế.

- Không trả lời: nghĩa là người được phỏng vấn không cung cấp thông tin và như vậy chúng ta không thu thập được thông tin cần thiết từ một số người được chọn trong mẫu điều tra của chúng ta (thiếu thông tin ngẫu nhiên, thiếu thông tin không ngẫu nhiên và vấn đề chệch mẫu).

Sai lầm quan sát được bao gồm:

- Sai lầm thực địa do các nhân tố bối cảnh, tâm lý và hành vi gây ra, nó chủ yếu phụ thuộc vào người làm nghiên cứu do những suy nghĩ chủ quan của mình chi phối kết quả điều tra.

- Sai lầm văn phòng, sai lầm trong việc biên tập, mã hoá, lập bảng và phân tích số liệu.

Về mặt LÝ THUYẾT chúng ta luôn đặt ra yêu cầu là lựa chọn quy mô mẫu sao cho ước lượng tính toán ra từ mẫu đó có độ CHÍNH XÁC cao nhất song trong THỰC TẾ nó phụ thuộc rất nhiều vào yếu tố:

- Phương pháp luận được lựa chọn bởi vì nó quyết định: (1) Mức độ chính xác mà nghiên cứu yêu cầu (có thể chấp nhận sai số ở mức nào); (2) Ở mức độ mà có sự biến thiên trong tổng thể đối với những đặc điểm chính của nghiên cứu.

- Tỷ lệ trả lời có thể, bản thân nó sẽ phụ thuộc vào phương pháp chọn mẫu được sử dụng, nếu chúng ta áp dụng phương pháp chọn mẫu phù hợp để lựa chọn đối tượng đúng với yêu cầu của nghiên cứu thì chúng ta sẽ có tỷ lệ trả lời cho các câu hỏi cao hơn và ngược lại.

- Thời gian và tiền bạc sẵn có.

- Nguồn nhân lực sẵn có như: nhóm giám sát, điều tra viên, dẫn đường v.v... vì đây là những tiềm ẩn của sai số phi chọn mẫu.

Tuy nhiên, trong thực tế chúng ta cũng gặp không ít các khó khăn khác như trong các cuộc điều tra với mục tiêu đặt ra không phải là một mục tiêu duy nhất mà là đa mục tiêu hay nói cách khác có nhiều ước lượng. Khi đó, chúng ta sẽ khó có thể xác định được một quy mô tối ưu với những thiết kế mẫu phức tạp.

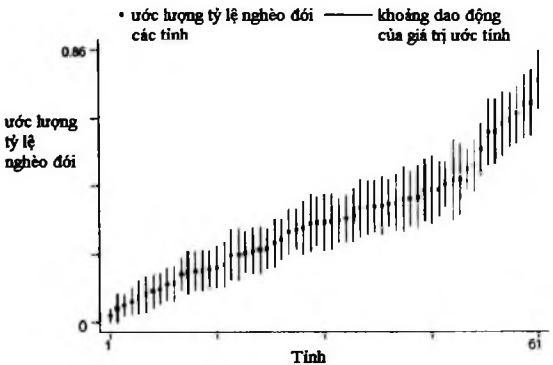
Thông thường thì các nhà nghiên cứu mong muốn có một quy mô mẫu nhỏ bởi vì nó có ưu điểm là có chi phí thấp, tốn ít thời gian hơn và khả năng có 1 sai số phi chọn mẫu thấp là khá lớn, song bên cạnh đó thì một nhược điểm lớn của số mẫu nhỏ đó là sai số mẫu lớn và vì vậy, số mẫu tối ưu là khi chúng ta chấp nhận mức sai số nhất định nào đó mà nó có tổng của sai số phi chọn mẫu và sai số chọn mẫu là nhỏ nhất.

*Ví dụ:* Mẫu 150 doanh nghiệp, phân tầng theo ngành dệt, may và hình thức sở hữu. Như vậy, sẽ chỉ có một vài doanh nghiệp trong mỗi nhóm và điều này dẫn đến rất khó kiểm định xem có khác biệt về mặt thống kê giữa các nhóm hay không.

Nghiên cứu tình huống: quy mô mẫu nhỏ có thể ảnh hưởng đến hoạch định chính sách như thế nào (Hình 1.7).

Dựa vào kết quả phân tích trong biểu đồ trên chúng ta sắp xếp thứ tự các tỉnh theo tỷ lệ nghèo đói và nhận thấy giữa các tỉnh xếp gần nhau, mặc dù có số liệu bình quân là khác nhau, song khi so sánh khoảng dao động của ước lượng tính toán được thì chúng ta thấy nó có thể cùng được xếp vào một nhóm, hay nói cách khác, trong trường hợp này không có sự

khác biệt giữa các tỉnh gần nhau. Trong ví dụ này, nếu chúng ta muốn xem sự khác biệt chỉ trong trường hợp chúng ta lựa chọn các tỉnh ở hai đầu mút của đường đồ thị mà thôi.

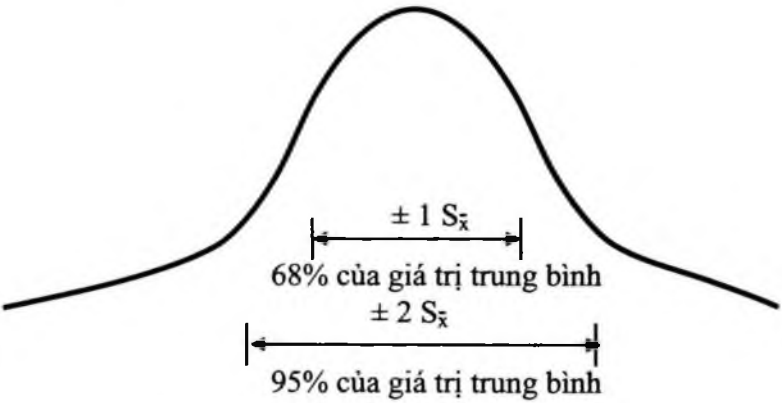



**Hình 1.7: BIỂU ĐỒ GIÁ TRỊ ƯỚC LƯỢNG TỶ LỆ NGHÈO ĐÓI CỦA CÁC TỈNH QUA ĐIỀU TRA CHỌN MẪU**

Tất cả các mẫu đều chứa đựng sai số. Mặc dù, mục tiêu của chúng ta là lựa chọn mẫu như là bản sao thu nhỏ của tổng thể, tuy nhiên nó luôn tồn tại một khoảng cách giữa mẫu và tổng thể.

Mục tiêu của chúng ta là làm sao để có thể có một mẫu mà sai số chọn mẫu là nhỏ nhất. Hay nói cách khác là trong bất kỳ một bước nào của quá trình điều tra đều cần phải hạn chế tối đa sự sai số.

Sai số chọn mẫu như chúng ta đã biết đó là sự khác biệt giữa giá trị trung bình của mẫu và giá trị đúng của tổng thể và thống kê dùng cụm từ mô tả sai số chọn mẫu là sai số chuẩn của giá trị trung bình (Standard error of the mean). Chúng ta cũng cần phân biệt sự khác nhau giữa 2 cụm từ độ lệch chuẩn (Standard deviation) và sai số chuẩn của giá trị bình quân (Standard error of the mean) ở chỗ là độ lệch chuẩn cho thấy sự biến động như thế nào giữa các giá trị cá biệt trong khi đó sai số chuẩn là độ lệch chuẩn của giá trị bình quân trong phân phối mẫu nói lên rằng mức độ biến động có thể mong đợi giữa các giá trị trung bình trong các mẫu sẽ lấy trong tương lai.




**Hình 1.8: PHÂN PHỐI MẪU CỦA GIÁ TRỊ TRUNG BÌNH**

Khi giá trị của sai số chuẩn được tính toán, 68% của giá trị trung bình của một mẫu nào đó sẽ rơi vào khoảng của 1 lần sai số chuẩn của giá trị bình quân đích thực của tổng thể; 95% của giá trị trung bình của một mẫu nào đó sẽ rơi vào khoảng

của 2 lần sai số chuẩn của giá trị bình quân đích thực của tổng thể và 99% của giá trị trung bình của một mẫu nào đó sẽ rơi vào khoảng của 3 lần sai số chuẩn của giá trị bình quân đích thực của tổng thể (hình 1.8)

*Tính toán sai số chuẩn cho chọn mẫu ngẫu nhiên đơn giản:*

Mặc dù, chúng ta đã biết về thuật ngữ sai số chuẩn cũng như hiểu về tính phức tạp của việc chọn mẫu. Tuy nhiên, chúng ta cũng cần phải hiểu về cả thuật ngữ và bản chất.

Công thức tính toán sai số chuẩn (SE) dựa vào sự biến động (Variance) và cỡ của mẫu:

$$SE = \sqrt{Var / n}$$

*Trong đó:*

SE là sai số chuẩn của giá trị bình quân.

Var là độ biến động (Variance - Tổng bình phương độ lệch chuẩn)

n là số lượng mẫu.

Giá trị bình quân được tính

$$\overline{X} = \Sigma x / n$$

*Trong đó:*

$\overline{X}$ là giá trị bình quân.

X là giá trị của từng cá thể.

n là số mẫu.

Trong trường hợp tính toán cho các chỉ tiêu định tính chỉ nhận hai giá trị là 1 (có) và 0 (không), chẳng hạn khi hỏi 100 người thì có 20 người trả lời có hiểu sự khác biệt giữa độ lệch chuẩn và sai số chuẩn còn 80 người trả lời không hiểu về sự khác biệt đó. Khi đó Var của tỷ lệ được tính theo công thức p(1-p).

*Trong đó:* p là tỷ lệ với câu trả lời là có (VD: 20%) và 1-p là tỷ lệ với câu trả lời là không (VD: 80%)

$$SE = \sqrt{p(1-p)/n} \quad SE = \sqrt{0,2*0,8/100} = 0,04$$

Như vậy, nếu ta cộng và trừ tỷ lệ người trả lời có 0,2 với sai số chuẩn của giá trị bình quân 0,04 ta được 0,24 và 0,16. Ta có thể nói là xác suất là 68% (1 lần sai số chuẩn) chắc chắn rằng số liệu thực tế của tổng thể rơi trong khoảng 0,16 - 0,24.

### *1.3.3. Cỡ mẫu*

Ý tưởng của việc chọn mẫu là để minh hoạ và đại diện cho một tổng thể nghiên cứu. Để đạt được điều này tức là nhằm tránh những sai sót chọn mẫu hoặc sai số phi chọn mẫu. Sai số phi chọn mẫu xuất hiện khi chúng ta xác định tổng thể nghiên cứu không chính xác hoặc do các yếu tố chủ quan khác trong quá trình tiến hành điều tra. Sai số chọn mẫu xuất hiện khi chúng ta xác định cỡ mẫu không đủ tính đại diện hoặc thiên lệch. Vì thế, việc chọn mẫu đại diện và đủ lớn là rất quan trọng để đảm bảo tránh những sai số đáng tiếc trong quá trình chọn mẫu. Có nhiều công thức khác nhau dùng để tính toán cỡ mẫu, trong thực tế có nhiều tài liệu giới thiệu cho chúng ta những công cụ này.

Giả sử việc nghiên cứu của chúng ta về một khía cạnh sức khoẻ hay giáo dục nào đó mà chúng ta có phân chia ra các nhóm khác nhau để xem xét câu hỏi đặt ra là mỗi nhóm nên có bao nhiêu quan sát? Để trả lời cho câu hỏi này chúng ta sẽ phải trả lời cho 5 câu hỏi khác đó là:

(1) Giả thiết của chúng ta là gì ($H_o$)?

Giả thiết của chúng ta là không có sự khác biệt giữa giá trị bình quân của hai nhóm.

Giả sử khi nghiên cứu một nhóm thanh niên, giả thiết của chúng ta là không có sự khác biệt giữa mục đích và khát vọng giữa nhóm thanh niên tham gia vào nghiên cứu và nhóm không tham gia.

(2) Mức độ ý nghĩa thống kê (mức α) liên quan đến giả thiết $H_o$ bao hàm giá trị bình quân của tổng thể ($\mu_o$) bằng bao nhiêu?

Lưu ý trong việc kiểm định giả thiết chúng ta thường hay dùng giá trị bình quân của tổng thể hơn là giá trị bình quân mẫu.

Mức độ ý nghĩa được gọi là giá trị α. Giá trị này cho biết xác suất của việc bác bỏ giả thiết $H_o$ khi nó đúng trong thực tế. Thông thường, chúng ta hay lựa chọn với các mức α nhỏ 0,05; 0,01 hoặc 0,001 để tránh việc bác bỏ giả thiết $H_o$ khi nó đúng trong thực tế (và như vậy không có sự khác biệt có ý nghĩa thống kê giữa hai nhóm). Giá trị p là xác suất trong đó một quan sát (hoặc kết quả của một kiểm định thống kê) có cơ

hội nhận được. Nó được tính toán **sau khi** kiểm định thống kê. Nếu giá trị p nhỏ hơn α thì giả thiết $H_o$ sẽ bị bác bỏ.

(3) Vậy cơ hội tìm được sự khác biệt thực sự sẽ như thế nào? Hay hỏi một cách khác (1-β) liên quan đến đối thiết nào?

Khi tìm thấy có sự khác biệt giữa hai nhóm với nhau nhưng trong thực tế không có sự khác biệt, khi đó được gọi là α hoặc kiểu sai số I; khi không có sự khác biệt được tìm ra giữa hai nhóm mặc dù trong thực tế có sự khác biệt giữa chúng, trường hợp này được gọi là β hoặc kiểu sai số II. Mối quan hệ này được thể hiện qua bảng sau:

| | | Thực tế | |
|---|---|---|---|
| | | Tồn tại sự khác biệt | Không tồn tại sự khác biệt |
| Kết luận từ kiểm định giả thiết | Tồn tại sự khác biệt | Đúng | Kiểu I hay sai số α |
| | Không tồn tại sự khác biệt | Kiểu II hay sai số β | Đúng |

(4) Những sự khác biệt gì của các giá trị bình quân tìm được là quan trọng? Ý nghĩa của $\mu_1 - \mu_2$?

Giả sử chúng ta có một nghiên cứu và chúng ta đưa ra 50 điểm tỷ lệ. Bước đầu tiên là chúng ta cần phải đồng ý một mức độ sự khác biệt giữa các giá trị bình quân cả trên hai góc độ thực tế và thống kê. Chúng ta có thể hỏi một nhà chuyên gia 5 điểm đó là sự khác biệt? hay 10 điểm? (sự khác biệt này

đôi khi liên quan đến "hiệu ứng" và cỡ của sự khác biệt là "mức độ hiệu ứng").

(5) Như thế nào là một ước lượng không chệch về độ lệch chuẩn (δ) của tổng thể?

Độ lệch chuẩn là một đo lường của sự dao động của các giá trị thực tế so với giá trị bình quân của mẫu. Hai cách chung để áp dụng độ lệch chuẩn: Thứ nhất, tối thiểu 75% của các giá trị thực tế luôn nằm trong khoảng giữa giá trị bình quân và 2 lần giá trị độ lệch chuẩn (ở mức 95%). Giả sử có 100 mẫu điều tra với giá trị bình quân là 25 và độ lệch chuẩn là 2, như vậy có tối thiểu 75 câu trả lời có giá trị là 25 ± 4. Điều đó có nghĩa là giá trị thực tế của chúng sẽ nằm trong khoảng từ 21 đến 29. Nếu sự phân bố của các giá trị thực tế của các quan sát có dạng hình chuông hay theo phân bố chuẩn thì 68% của các quan sát sẽ nằm trong khoảng giá trị trung bình ± 1 lần độ lệch chuẩn, 95% của các quan sát rơi vào khoảng giá trị trung bình ± 2 lần độ lệch chuẩn và 99% của các quan sát rơi vào khoảng giá trị trung bình ± 3 lần độ lệch chuẩn. Giá trị ước lượng của độ lệch chuẩn có thể sử dụng từ các cuộc điều tra trước đó, tuy nhiên trước khi sử dụng chúng ta cần phải kiểm tra xem liệu phân bố của tổng thể đó có giống với phân bố của tổng thể chúng ta đang nghiên cứu hay không. Hoặc chúng ta có thể tiến hành một điều tra thử ở diện hẹp với khoảng 25 quan sát để ước lượng độ lệch chuẩn từ đó. Cuối cùng, chúng ta có thể nhờ các chuyên gia cung cấp các giá trị cao nhất và thấp nhất như những giá trị cơ bản để có thể tính toán độ lệch chuẩn.

Công thức để tính toán quy mô mẫu điều tra trong nghiên cứu so sánh giữa hai nhóm như sau:

$$\frac{(z_\alpha - z_\beta)\sigma^2}{\mu_1 - \mu_2}$$

*Trong đó:*

$\mu_1 - \mu_2$ là khoảng khác biệt giữa hai nhóm.

$z_\alpha$, $z_\beta$ là các giá trị đối xứng của phân phối chuẩn.

Các giá trị đó được xác định như sau:

$$z_\alpha = \frac{X - \mu_1}{\sigma/\sqrt{n}} \text{ và } z_\beta = \frac{X - \mu_2}{\sigma/\sqrt{n}}$$

Ta hãy xem qua ví dụ sau:

Hai nhóm thanh niên tham gia vào một nghiên cứu sức khoẻ, giáo dục và chất lượng cuộc sống. Một cuộc điều tra được tiến hành nhằm tìm ra mục đích và mong muốn của họ. Giả sử giá trị điểm tối đa mà ta có thể có từ điều tra là 100 điểm. Kiểu sai số I với α = 0,05. Xác suất để bác bỏ có sự khác biệt được xác định ở mức 0,80. Các chuyên gia nghiên cứu trong lĩnh vực này cho biết sự khác biệt trong điểm số giữa nhóm quan sát và nhóm đối chứng nên lớn hơn từ 10 điểm. Sử dụng kết quả một cuộc điều tra trước đây với độ lệch chuẩn là 15 điểm.

Hãy tính toán quy mô mẫu cho từng nhóm?

Trước hết, chúng ta giả thiết rằng mẫu nghiên cứu của chúng ta có phân phối gần hoặc tương đương với phân phối chuẩn. Đường cong phân phối chuẩn có giá trị bình

quân là 0 và độ lệch chuẩn là 1. Giá trị z hai phía ứng với α = 0,05 là 1,96; ứng với α = 0,01 là 2,58; ứng với α = 0,1 là 1,65 và ứng với α = 0,2 là 1,28. Giá trị thấp nhất của z tương ứng với β = -0,84

Áp dụng công thức ta có: $n = 2\left(\frac{\delta * z_{\beta}}{X - \mu}\right)^2$

N= $2\left(\frac{(1,96 + 0,84) * 15}{10}\right)^2$ = 2*(17,64) = 36

Như vậy, tối thiểu chúng ta cần 36 quan sát trong mỗi nhóm để đảm bảo 80% cơ hội nhận được sự khác biệt giữa hai nhóm nghiên cứu với mức khác biệt là 10 điểm số.

Trong thống kê, theo quy luật ngón tay cái chúng ta cần có khoảng 30 quan sát trong mỗi nhóm để đảm bảo ý nghĩa thống kê. Vì vậy, khi chúng ta tăng số nhóm lên chúng ta cũng cần phải tăng số lượng quan sát đảm bảo đủ tối thiểu là 30 quan sát/nhóm.

### *1.3.4. Tính toán trọng số*

Trọng số là gì? Trọng số là một giá trị số sử dụng để điều chỉnh các giá trị thực tế khác về dạng có thể so sánh được với nhau hay có cùng một đơn vị tính.

Dùng trọng số để điều chỉnh vấn đề xác suất chọn mẫu, trong một mẫu ngẫu nhiên (trường hợp chuẩn), mỗi cá nhân đều có cơ hội được lựa chọn như nhau. Tuy nhiên, thủ tục chọn mẫu trong thực tế có thể là hộ gia đình, chứ không phải là cá nhân, có cơ hội như nhau.

Như vậy, các cá nhân ở trong những hộ gia đình có nhiều thành viên hơn cũng chỉ có cơ hội để được chọn như đối với các cá nhân trong các gia đình có số thành viên ít hơn hay nói cách khác cơ hội đối với họ là thấp hơn: họ không được đại diện đúng mức và vì vậy, nên có trọng số để điều chỉnh để họ có tính đại diện trong mẫu điều tra.

Trọng số điều chỉnh sau khi phân tầng, nếu có sự đại diện không đúng mức của một tầng nào đó.

*Ví dụ:* Nam so với nữ do sự không trả lời hoặc do chọn mẫu phân tầng không tỷ lệ.

Ví dụ, nếu tỷ lệ thực tế theo giới tính là 50-50 và nếu chúng ta chọn mẫu có 40 nữ và 60 nam, chúng ta có thể gán cho người trả lời nữ một trọng số là 1,5 để tạo lên sự cân bằng. Trong thực tế, điều này tạo ra 60 nữ và 60 nam. Tuy nhiên, để tránh việc tăng quy mô mẫu một cách nhân tạo từ 100 lên 120, ta cần tính thêm trọng số để điều chỉnh quy mô về 100. Điều này có thể được thực hiện bằng cách tạo thêm trọng số cho cả nữ và nam là 5/6. Mục đích là để tạo trọng số cho các trường hợp hiện tại theo cách làm gia tăng tính đại diện trong mẫu điều chỉnh của những tầng không được đại diện đúng mức trong mẫu.

### 1.4. Phương pháp thu thập số liệu

Có nhiều cách để tạo ra số liệu cho một nghiên cứu, ví dụ như trong các nghiên cứu cơ bản hoặc tự nhiên thông thường chúng ta hay tổ chức hoặc tiến hành các thí nghiệm và qua quá trình tiến hành các thí nghiệm đó người nghiên cứu sẽ

quan sát, đo đạc, ghi chép để có số liệu phục vụ cho việc phân tích của mình. Còn đối với các vấn đề kinh tế - xã hội thì để có số liệu phục vụ nghiên cứu thông thường nhà nghiên cứu sẽ tiến hành điều tra để thu thập thông tin.

Việc điều tra có thể được tiến hành theo nhiều cách khác nhau, như: điều tra các đối tượng (thường là các hộ, các doanh nghiệp, tổ chức hoặc cá nhân) thông qua việc sử dụng các câu hỏi khác nhau tuỳ theo mục tiêu cần nghiên cứu. Các câu hỏi này có thể được chuẩn bị sẵn (bảng câu hỏi chuẩn) hoặc có thể được người điều tra kết hợp với các câu hỏi đặt ra trong quá trình phỏng vấn v.v...

Ngoài ra, người nghiên cứu cũng có thể thu được số liệu thông qua việc tìm hiểu các tài liệu thống kê đã công bố như niên giám thống kê hoặc các báo cáo khoa học. Thường các số liệu này chỉ mang tính tổng quát và minh chứng cho phần nghiên cứu cụ thể của nghiên cứu.

Có nhiều cách phỏng vấn khác nhau như: phỏng vấn trực tiếp, phỏng vấn qua thư, phỏng vấn qua thư điện tử và phỏng vấn qua điện thoại. Mỗi loại hình phỏng vấn khác nhau sẽ đòi hỏi những bảng câu hỏi khác nhau để thu được thông tin mong muốn.

Một trong những vấn đề hết sức quan trọng đó là công cụ trong thu thập thông tin mà chúng ta thường gọi là phiếu điều tra. Kết quả điều tra có đảm bảo tính khoa học hay không, có độ chính xác hay không, có thể phân tích được hay không và thông tin có đủ cho việc phân tích nghiên cứu hay không phụ thuộc rất nhiều vào phiếu điều tra.

Để xây dựng được một phiếu điều tra tốt đòi hỏi người nghiên cứu phải có kinh nghiệm và am hiểu về vấn đề cần nghiên cứu, địa bàn nghiên cứu và đặc biệt là phải kết hợp tốt giữa lý thuyết và thực tế.

### *1.4.1. Phiếu điều tra*

Là công cụ quan trọng trong việc thu thập thông tin cho các nghiên cứu thuộc lĩnh vực kinh tế - xã hội.

Phiếu điều tra có nhiều hình thức khác nhau như trình bày dưới dạng bảng, kết hợp bảng và từng câu hỏi riêng biệt.

Mỗi dạng có những ưu, nhược điểm khác nhau chẳng hạn dưới dạng bảng sẽ làm cho phiếu điều tra có vẻ ngắn hơn và trong một bảng thu được nhiều thông tin khác nhau, tuy nhiên một trong những hạn chế chính đó là dễ để xảy ra tình trạng nhầm lẫn cả trong khi điền số liệu điều tra từ ô này sang dòng kia v.v... hay khi nhập số liệu vào máy tính cũng có thể xảy ra tình trạng nhầm lẫn tương tự.

Trong khi đó dưới dạng các câu hỏi riêng biệt sẽ làm cho phiếu điều tra có vẻ phức tạp hơn, làm cho cả người đi điều tra và người được hỏi cảm thấy ngại khi làm việc với tập phiếu điều tra dày như vậy nhưng nó lại hạn chế được những nhầm lẫn như ở dạng bảng.

Đối với các thông tin định tính, các câu hỏi trong phiếu điều tra có thể được chia làm 2 loại chính:

(1) Câu hỏi dạng đóng hay câu hỏi có gợi ý (hoặc lựa chọn).

(2) Câu hỏi dạng mở là câu hỏi để cho người được hỏi tự do lựa chọn câu trả lời.

Với câu hỏi dạng đóng thì quan trọng nhất là phần gợi ý trả lời hoặc các lựa chọn cho người trả lời vì nó sẽ ảnh hưởng tới kết quả phân tích. Nếu các gợi ý trước không được đầy đủ sẽ làm cho người trả lời khó khăn khi trả lời câu hỏi dạng này hoặc chúng ta sẽ không thu được thông tin chính xác.

Phần gợi ý và lựa chọn có thể được chia ra như sau: gợi ý với các ý trả lời khác nhau không có liên hệ với nhau.

*Ví dụ:* Khi chúng ta tìm hiểu nguyên nhân vay tiền từ ngân hàng của các hộ, chúng ta có các gợi ý như: để đầu tư; để cho các nhu cầu sinh hoạt của gia đình; v.v... và các gợi ý này không có mối liên hệ với nhau.

Các gợi ý của câu hỏi dạng này cũng có thể là phần đánh giá và có tỷ lệ hay khoảng cách nhất định, ví dụ như: Rất tốt, tốt, bình thường, kém v.v... như vậy, trong trường hợp này nó có mối quan hệ với nhau theo 1 tỷ lệ. Đối với câu hỏi dạng này thì việc phân chia khoảng cách có ý nghĩa rất quan trọng, nó có thể bao gồm:

(1) Phân chia theo danh (Nominal scale) thông thường chỉ là 0 và 1 ví dụ như khi phân chia giới A # B.

(2) Phân chia theo thứ tự A>B>C... (Ordinal scale).

(3) Phân chia theo tỷ lệ (Interval scale) đây là hình thức kết hợp giữa dạng phân chia (1) và (2).

*Ví dụ:* Khi đánh giá về một sản phẩm hay dịch vụ nào đó chúng ta có thể đặt câu hỏi: Mức độ thoả mãn của anh chị về dịch vụ hay loại hàng hoá v.v... và đưa ra các gợi ý như sau: Vâng rất thoả mãn; vâng thoả mãn; thoả mãn và không thoả mãn chút nào.

(4) Phân chia theo thang số ví dụ từ 1 đến 5 là mức độ đánh giá của người được hỏi.

Hình thức đặt câu hỏi cho mỗi ý trả lời có ý nghĩa rất quan trọng vì nếu câu hỏi rõ ràng hoặc không hàm ý gì thì việc thu được kết quả mới chính xác.

*Ví dụ:* Khi chúng ta đánh giá một dịch vụ mới thay vì đặt câu hỏi anh/chị đánh giá như thế nào về dịch vụ mới này? Thì chúng ta sẽ thay câu trả lời vâng rất nhiều bởi câu hỏi anh/chị có thích dịch vụ này không?

Trong hầu hết các phiếu điều tra đều có sự kết hợp giữa hai loại câu hỏi dạng đóng và dạng mở, tuy nhiên chúng ta thường hay sử dụng câu hỏi dạng đóng nhiều hơn vì việc xử lý thông tin sau này sẽ dễ hơn cũng như việc thu thập thông tin sẽ dễ dàng hơn, còn câu hỏi dạng mở thường được dùng để minh hoạ, giải thích thêm cho các luận chứng phân tích sau này.

### *1.4.2. Các phương pháp phỏng vấn thu thập thông tin*

Có nhiều cách thức thu thập thông tin khác nhau như: đến hộ phỏng vấn; phỏng vấn qua điện thoại; phỏng vấn qua thư; phỏng vấn qua thư điện tử.

Chúng ta có thể dùng bảng câu hỏi đã chuẩn bị sẵn để hỏi hoặc chúng ta có thể tiến hành cuộc phỏng vấn theo hình thức nói chuyện, thảo luận nhóm v.v...

Mỗi phương pháp thu thập thông tin và công cụ sử dụng để thu thập thông tin sẽ phù hợp với từng yêu cầu cụ thể và mục đích của từng cuộc điều tra, nghiên cứu.

Phương pháp phỏng vấn trực tiếp sẽ giúp ta có cơ hội tiếp cận với đối tượng điều tra, từ đó ta dễ dàng tranh thủ được đối tượng điều tra, có thể trao đổi để họ hiểu về mục đích cuộc điều tra từ đó họ có lòng tin hơn và sẽ cung cấp thông tin đầy đủ cũng như chính xác hơn. Hơn nữa trong phương pháp này thì việc có thể khai thác thêm những thông tin bổ trợ cũng như trao đổi làm rõ những câu hỏi v.v... là điều có thể.

Yêu cầu đối với phương pháp điều tra trực tiếp là người được điều tra phải nắm rõ về bảng câu hỏi để có thể hỏi một cách lô gíc cũng như dẫn dắt người được hỏi cho đúng mục đích yêu cầu cuộc điều tra. Đối với điều tra viên, phải là người có kinh nghiệm tiếp xúc với nhóm đối tượng được điều tra, hiểu tâm lý họ và một điều cũng hết sức quan trọng là bảng hỏi không nên quá nhiều gây tâm lý ngại cho người được hỏi.

Đối với hình thức phỏng vấn qua điện thoại đòi hỏi người điều tra viên phải là người rất có kinh nghiệm, am hiểu vấn đề quan tâm và là người có kinh nghiệm trao đổi qua điện thoại tránh việc hỏi lâu sẽ làm cho người được hỏi khó chịu. Bảng hỏi trong trường hợp này phải hết sức ngắn gọn.

Với hình thức phỏng vấn qua thư hay thư điện tử đòi hỏi việc thiết kế phiếu điều tra ngắn gọn, dễ hiểu để bất cứ ai

thuộc nhóm đối tượng điều tra đều hiểu vấn đề như nhau. Nên sử dụng nhiều câu hỏi dạng lựa chọn hơn là câu hỏi suy luận và diễn giải tránh việc người được hỏi phải sử dụng quá nhiều thời gian cho việc trả lời một câu hỏi.

Việc sử dụng bảng câu hỏi có sự chuẩn bị sẵn sẽ giúp cho quá trình điều tra diễn ra nhanh và đảm bảo kết quả như chúng ta muốn, vì việc điều tra theo hình thức thảo luận đôi khi sẽ kéo dài cuộc phỏng vấn mà vấn đề chúng ta cần thu thập thông tin đôi khi không thể hiện rõ. Tuy nhiên, việc kết hợp cả hai hình thức hỏi theo bảng hỏi và thảo luận là điều tốt nhất nó sẽ giúp cho chúng ta có cả những thông tin cần thiết, bắt buộc và cả những thông tin hỗ trợ cho nghiên cứu của mình.

# Chương II

# CƠ SỞ DỮ LIỆU

Cơ sở dữ liệu là một mẫu thông tin dưới dạng điện tử, nó có thể bao gồm 1 hoặc là nhiều tệp dữ liệu khác nhau. Cơ sở dữ liệu có thể được thể hiện dưới dạng một bảng số liệu gồm nhiều hàng và cột khác nhau trong đó mỗi dòng thể hiện 1 chỉ tiêu nào đó và mỗi cột thể hiện cho 1 quan sát, ví dụ 1 hộ hay một doanh nghiệp. Mỗi một ô trong bảng thể hiện 1 giá trị cụ thể.

Có nhiều phần mềm cho phép xây dựng và quản lý một cơ sở dữ liệu, như: phần mềm MS ACCESS, EXCEL hay LOTUS.

Các thông tin trong cơ sở dữ liệu phải được thể hiện ở dạng số vì các phép xử lý toán học chỉ có thể tiến hành khi thông tin đó đã được lượng hoá, những thông tin về mặt định tính phải được mã hoá trước khi tiến hành các phép xử lý thống kê.

Tất cả các thông tin định tính được mã hoá trong quá trình thu thập hay vào số liệu trong cơ sở dữ liệu phải được ghi lại để tránh nhầm lẫn trong quá trình xử lý tính toán sau này.

Một ví dụ về cơ sở dữ liệu bao gồm nhiều thông tin của nhiều mẫu được quản lý chung trong một tệp tin. Như trong ví

dụ này, số liệu của 1 hộ được thể hiện như là một trang của quyển sách, các trang tiếp theo sẽ là thông tin của các hộ khác theo đúng trật tự như của hộ đầu tiên.


Hình 2.1: MÔ PHỎNG MỘT CƠ SỞ DỮ LIỆU

Một ví dụ nữa về cơ sở dữ liệu được trình bày dưới dạng bảng trong đó mỗi dòng thể hiện cho một hộ (mẫu) điều tra và mỗi một cột thể hiện cho một chỉ tiêu điều tra (một thông tin), như vậy, trong trường hợp này số lượng mẫu điều tra sẽ quyết định đến số lượng hàng cần phải có trong cơ sở dữ liệu, trong

khi đó số lượng các chỉ tiêu cần điều tra sẽ quyết định số lượng các cột trong cơ sở dữ liệu. Chúng ta cũng có thể thay đổi theo hàng là các chỉ tiêu thông tin thu thập và theo cột là các mẫu điều tra tuỳ theo yêu cầu và cách nào phù hợp hơn cho ta.

| 1. | Thông tin xác định | | | |
|---|---|---|---|---|
| 1.0001 | Tên người được hỏi | Quàng V. Liên | Khổng M. Ngụ | Nguyễn V. A |
| 1.0002 | Huyện | Mai Son | Mai Son | Mai Son |
| 1.0003 | Tên xã | Mường Bon | Hát Lót | Hát Lót |
| 1.0004 | Tên bản | Bản Un | Bắc Quang | Bắc Quang |
| 1.0005 | Dân tộc | Thái | Kinh | Kinh |
| 1.0006 | Ngày phỏng vấn | 25.05.06 | 25.05.06 | 26.05.06 |
| 1.0007 | Hộ số | 1 | 2 | 3 |
| ... | ..... | ..... | ..... | ..... |
| 2.1001 | Số nhân khẩu trong hộ | 6 | 5 | 7 |
| 2.1002 | Số trẻ em dưới 15 tuổi | 2 | 1 | 3 |
| 2.1003 | Chủ hộ (Nam =1; Nữ = 2) | 1 | 1 | 2 |
| 2.1004 | Tuổi chủ hộ | 50 | 48 | 40 |

Quá trình quản lý và nhập số liệu vào máy tính bao gồm hai công đoạn:

(1) Chuẩn bị cơ sở dữ liệu

Là việc chuẩn bị cấu trúc của cơ sở dữ liệu theo một trật tự nhất định sao cho việc quản lý các thông tin khoa học nhất và đảm bảo việc kết xuất dữ liệu sang các phần mềm tính toán khác là có khả thi. Thông thường, chúng ta phải dựa vào kết cấu của phiếu điều tra, số lượng mẫu, số lượng các chỉ tiêu chi tiết trong phiếu điều tra để có thể có được một kết cấu của cơ sở dữ liệu phù hợp.

Trong việc chuẩn bị cấu trúc của cơ sở dữ liệu một trong những vấn đề quan trọng cần phải lưu tâm ngay từ đầu đó là hệ thống mã hoá và các thông tin liên quan. Các phần mềm tính toán thông thường không thể xử lý được các thông tin định tính (cho các câu hỏi mở), do vậy, việc chúng ta phải chuyển các thông tin dạng đó sang dạng định lượng là điều cần thiết và để làm được điều này, chúng ta cần phải xây dựng một hệ thống các mã hoá cho từng câu hỏi và ý trả lời một.

(2) Kết chuyển dữ liệu từ cơ sở dữ liệu sang phần mềm xử lý.

Việc xử lý các thông tin điều tra thường bằng các phần mềm thống kê như phần mềm SPSS hoặc Stata, bên cạnh đó chúng ta cũng có thể sử dụng các công cụ trong Excel và Lotus để tính toán các thông tin cần thiết cho nghiên cứu.

Việc kết chuyển chúng ta có thể làm trực tiếp bằng một số câu lệnh trong các phần mềm đó như đối với SPSS hoặc chúng ta có thể sử dụng các phần mềm cho phép chuyển định dạng của file dữ liệu sang dạng thích hợp cho các phần mềm xử lý thống kê như phần mềm Stat Transfer.

### 2.1. Các dạng cơ sở dữ liệu

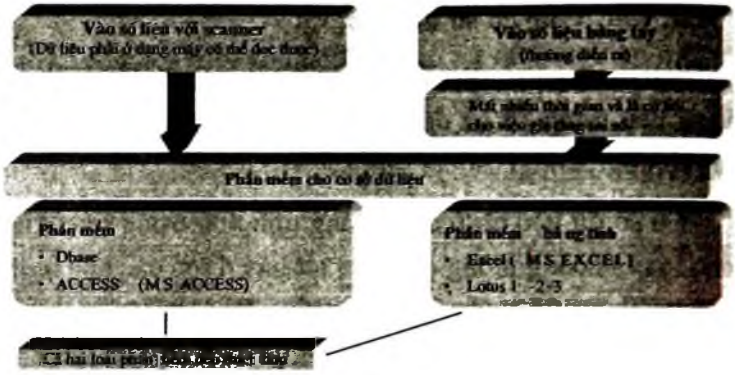
Định dạng cơ sở dữ liệu liên quan mật thiết đến hình thức vào số liệu trong cơ sở dữ liệu đó, tuy nhiên có 2 dạng định dạng chính:

- Cơ sở dữ liệu dưới dạng bảng tính như trong Excel hoặc Lotus: đây là dạng thông dụng và rất hay được các nhà nghiên cứu ứng dụng để quản lý thông tin. Tuy nhiên, một trong những hạn chế của định dạng cơ sở dữ liệu này đó là ta phải thực hiện việc truy nhập dữ liệu một cách thủ công và vì thế mất nhiều thời gian cũng như khả năng để xảy ra nhầm lẫn khá cao hay nói cách khác là nguy cơ tiềm ẩn của sai số phi thống kê cao.

Bảng tính Excel cũng như các chương trình quản lý cơ sở dữ liệu khác (MS Access) đều thích hợp cho việc vào số liệu từ các phiếu điều tra và cũng cho phép kết chuyển số liệu sang các phần mềm xử lý khác trong đó có SPSS hay Stata.

Một số lợi thế của bảng tính Excel là chương trình này sẵn có trong tất cả các máy tính điện tử, đòi hỏi những hiểu biết tối thiểu, có thể tính toán trực tiếp ngay tại bảng tính.

Nhưng bên cạnh đó cũng tồn tại những bất lợi trong việc sử dụng bảng tính Excel, đó là: (1) hạn chế các lệnh trong những tính toán phức tạp (đòi hỏi phải vào các câu lệnh thường xuyên bằng tay); (2) Không thích hợp trong việc tạo ra hàng loạt các báo cáo cho một mẫu.




**Hình 2.2: SƠ ĐỒ CÁC PHƯƠNG THỨC NHẬP TIN VÀO MÁY TÍNH**

- Cơ sở dữ liệu được định dạng lật hay nói cách khác mỗi một hoặc một vài thông tin được thiết kế thành 1 trang, như vậy, phải sử dụng nhiều trang khác nhau như một quyển sách. Đối với định dạng kiểu này chúng ta có thể sử dụng một vài chương trình để nhập thông tin với máy Scanner, như vậy dữ liệu trong phiếu điều tra phải được thể hiện theo đúng quy định nhất định để máy có thể đọc được. Với hình thức cơ sở dữ liệu như vậy rất phù hợp cho trường hợp cuộc điều tra lớn, người ta có thể xây dựng riêng 1 công cụ để đi điều tra và phần mềm riêng cho việc nhập cũng như xử lý dữ liệu (Hình 2.2).

## 2.2. Biểu diễn thông tin thống kê trong cơ sở dữ liệu

Các dữ liệu thống kê có thể được đo đạc dưới dạng chữ, dạng thứ tự hoặc dạng số liên tục và ta có thể quy về 2 dạng chung là định tính và định lượng.

### *2.2.1. Dữ liệu dạng định tính*

Là loại thông tin không được thể hiện ở dạng giá trị số mà những thông tin này được thể hiện phù hợp với một hạng hoặc loại nào đó, ví dụ như giới hoặc nơi sinh. Những thông tin dạng này thi thoảng được gọi là số liệu dạng, loại.

*Ví dụ:* Các câu hỏi cho ta thu được thông tin định tính như sau:

1. Giới của người được trao cho công việc là: (Khoanh tròn vào mục phù hợp)

Nam 1

Nữ 2

2. Mô tả loại ung thư phổi (khoanh tròn mục phù hợp)

Khối nhỏ 1

Khối lớn 2

Dạng sợi 3

Như vậy, các câu hỏi này đã phân loại các câu trả lời. Các câu trả lời là tên của loại đã phân, số liệu thể hiện trong các câu trả lời là thuộc tính và không có giá trị thực. Khi mà thông tin định tính chỉ có 2 sự lựa chọn như câu hỏi 1 về giới hoặc là nam hoặc là nữ thì được gọi là dạng phân đôi. Còn khi có nhiều sự lựa chọn như dạng câu hỏi 2 thì gọi là phân loại.

Các thông tin định tính còn có thể được thể hiện dưới dạng theo thứ tự. Nếu một thứ tự của các thuộc tính tồn tại bên trong

của các thông tin loại thì chúng ta gọi đó là có chứa đựng một sắp xếp theo thứ tự và chúng ta có thể minh hoạ qua ví dụ sau:

Câu hỏi: Mức độ học vấn mà anh đã qua? (lựa chọn một)

Chưa bao giờ hoàn thành chương trình tiểu học 1

Hoàn thành chương trình cấp I nhưng chưa xong THCS2

Hết THCS nhưng chưa xong THPT 3

Hết THPT nhưng không tiếp tục học đại học 4

Câu hỏi: Mức độ thường xuyên anh cảm thấy căng thẳng trong tháng qua? (lựa chọn một)

Luôn luôn 1 Thỉ thoảng 4

Rất thường xuyên 2 Không bao giờ 5

Thường xuyên 3

Việc biểu diễn số liệu dạng định tính được thể hiện dưới hai dạng chính: đó là bằng chữ, thường ít được sử dụng hơn do có nhiều điểm hạn chế như khó có khả năng tính toán, dùng các công cụ thống kê như SPSS hay Stata để tính toán.

*Ví dụ:* Khi hỏi về chất lượng nguồn nước sinh hoạt chúng ta thu được các thông tin kết quả như sau:

- Nước rất sạch
- Nước bình thường
- Nước bẩn

Dạng thứ hai là chúng ta thể hiện các thông tin này theo các mã số do chúng ta tự quy định, dưới dạng này chúng ta sẽ dễ dàng tính toán khi sử dụng các công cụ thống kê chuyên dùng hay bất kỳ một bảng tính nào.

Thông thường, các thông tin định tính khi chúng ta thu thập về để tránh những nhầm lẫn trong quá trình ghi chép do không nhớ mã ký hiệu mà chúng ta đã đặt thì chúng ta nên quy định ghi đầy đủ, rồi sau khi kiểm tra lại mới chuyển sang các mã số tương ứng, như vậy chúng ta vẫn đảm bảo thu thập đầy đủ thông tin mà ít mắc lỗi sai sót nhất.

### *2.2.2. Dữ liệu dạng định lượng*

Dữ liệu dạng định lượng được thể hiện dễ dàng trong cơ sở dữ liệu vì nó đã ở dạng số. Chính vì vậy trong quá trình xử lý thông tin này chúng ta không cần phải chuyển đổi hay mã hoá mà có thể làm trực tiếp ngay.

*Ví dụ:* Khi chúng ta thu thập được thông tin về diện tích đất nông nghiệp của các hộ thì nó sẽ được thể hiện ở dạng số như 1 ha hay 100 $m^2$ v.v...

### *2.2.3. Các chỉ tiêu nghiên cứu*

Một chỉ tiêu là một đặc trưng nghiên cứu có thể đo được, chẳng hạn như trọng lượng là một tiêu chí và một người cân được là 55 kg sẽ có cùng con số trọng lượng trong tiêu chí này. Người ta có thể chia các chỉ tiêu nghiên cứu ra làm hai nhóm: Chỉ tiêu độc lập và chỉ tiêu phụ thuộc.

Chỉ tiêu độc lập hay còn gọi là các chỉ tiêu giải thích hoặ là chỉ tiêu dự báo bởi vì các chỉ tiêu này thường được sử dụn để giải thích hoặc dự báo cho kết quả đầu ra chính là các ch tiêu phụ thuộc. Các chỉ tiêu độc lập hay phụ thuộc có thể đượ xác định thông qua việc nghiên cứu về mục đích và nhón mục tiêu nghiên cứu.

*Ví dụ:* Mục đích: Tìm hiểu chất lượng cuộc sống cho những người già thuộc các nhóm bệnh khác nhau có tiền sử khác nhau.

Nhóm mục tiêu: Những người già trên 65 tuổi có các bệnh già khác nhau và có tiền sử khác nhau.

Các chỉ tiêu độc lập: Tuổi, đặc trưng bệnh và tiền sử trước đó.

Chỉ tiêu phụ thuộc: Chất lượng cuộc sống.

### 2.3. Mã hoá các thông tin trong cơ sở dữ liệu

Mã hoá các thông tin trong cơ sở dữ liệu là vấn đề rấ quan trọng và có ảnh hưởng đến việc xử lý tính toán cũng nh kết quả của việc tính toán đó. Có nhiều vấn đề đòi hỏi chún ta phải mã hoá các thông tin, ở đây chúng ta có thể tạm thò phân ra làm 2 loại:

- Mã hoá cho các dữ liệu mang tính định tính, ví dụ nhu sự đánh giá, tên của các mẫu v.v...

- Mã hoá cho các thông tin định lượng bị thiếu hoặc vượ trội.

Việc mã hoá này phải được thống nhất từ đầu đến cuối của một cơ sở dữ liệu và phải được ghi chú hay chú thích cẩn thận tránh nhầm lẫn đáng tiếc ảnh hưởng đến kết quả phân tích sau này.

Đầu tiên, khi mở một tệp cơ sở dữ liệu chúng ta nhận thấy có các số thứ tự khác nhau theo dòng hoặc cột đó chính là các thông tin cho phép chúng ta đưa ra các nhận dạng về các mẫu điều tra để phân biệt giữa chúng được gọi là mã số của hộ điều tra. Những thông tin nhận dạng thường được thể hiện dưới dạng số và có thể có nhiều hơn một dòng hoặc cột.

*Ví dụ:* Thông tin về vùng, khu vực nghiên cứu thường được thể hiện thành nhiều dòng hoặc cột.

Nếu với mỗi phiếu điều tra có nhiều thông tin không thể thể hiện đủ trong 1 bảng tính thì ở bảng tính tiếp theo cũng phải bao gồm các thông tin nhận dạng để có thể theo dõi dễ dàng và không bị nhầm lẫn.

### *2.3.1. Mã hoá các thông tin định tính*

Máy tính chỉ có thể phân tích số liệu dưới dạng số vì thế những thông tin định tính cần phải được mã hoá trong khi nhập số liệu vào máy để dễ dàng cho việc xử lý sau này.

Những thông tin lựa chọn có/không sẽ được nhập là 1 và 0.

Các thông tin có nhiều sự lựa chọn câu trả lời sẽ được phân thành các nhóm khác nhau:

*Ví dụ:* Khi hỏi về trình độ văn hoá, chúng ta phân ra các hình thức sau: mù chữ, Tiểu học, THCS, THPT, Đại học; khi đó chúng ta sẽ mã hoá theo các số thứ tự từ 0 đến 4 (Mù chữ = 0; .... Đại học = 4).

### *2.3.2. Mã hoá các số liệu bị thiếu và vượt trội*

Các thông tin bị thiếu được hiểu là các thông tin cần thu thập song do một lý do nào đó mà trong phiếu điều tra không thể hiện kết quả của thông tin này mà theo yêu cầu kỹ thuật nó phải có thông tin. Quá trình thông tin bị thiếu có thể do nhiều lý do khác nhau, trong đó được phân ra hai nguyên nhân chính: Thiếu thông tin do người đi điều tra và thiếu thông tin do đối tượng điều tra.

*Ví dụ:* Một trong các hộ điều tra mà kết quả trong phiếu điều tra không thể hiện nhân khẩu của hộ thì đây là thông tin bị thiếu.

Các thông tin vượt trội được hiểu là các thông tin này có giá trị khác so với các giá trị thường gặp hoặc lớn hơn hoặc là nhỏ hơn.

*Ví dụ:* Hầu hết các hộ trong vùng có diện tích đất nông nghiệp là 2 ha song có một hộ có diện tích nông nghiệp lên đến 20 ha. Đây có thể là một thông tin vượt trội.

Tuy nhiên, việc xác định các thông tin vượt trội này còn cần phải có sự kiểm tra thật cẩn thận. Trước khi xác định đây là một thông tin vượt trội chúng ta cần phải kiểm tra lại trong

thực tế, nếu đó là giá trị thực mà trong quá trình điều tra đã kiểm tra kỹ, thì việc chúng ta phải chấp nhận thông tin này là điều đương nhiên. Tuy nhiên, trong nhiều cuộc điều tra mà chúng ta không trực tiếp hoặc không đảm bảo, không tin tưởng rõ vào thông tin vượt trội đó có là sự thật hay không thì chúng ta sẽ liệt thông tin này vào dạng số liệu vượt trội.

Việc xử lý các thông tin bị thiếu và vượt trội được tiến hành như nhau, do vậy trong việc mã hoá cũng sẽ tiến hành tương tự nhau, tức là ta sẽ coi các giá trị vượt trội như là các giá trị bị thiếu trong cơ sở dữ liệu.

Để mã hoá các thông tin bị thiếu và vượt trội chúng ta cần phải tuân theo một số quy định như sau:

*Quy định 1:* Không bao giờ được phép để các ô trống trong trường hợp những số liệu bị thiếu kể cả trong phiếu điều tra và trong cơ sở dữ liệu.

Những số liệu bị khuyết đó phải được mã hoá trong bảng tính bởi các lý do sau đây:

- Một khoảng trống có thể chỉ ra một sai sót nào đó trong quá trình điều tra hoặc là vào số liệu mà chúng ta chưa biết nhưng trong thực tế là nó bị thiếu, do vậy nếu chúng ta không mã hoá sẽ dẫn tới việc chúng ta sẽ bị mất rất nhiều thời gian để kiểm tra lại sai sót đó.

- Một vài phần mềm xử lý không phân biệt giữa khoảng trống và giá trị bằng 0 cho nên nó có thể ngầm định rằng giá trị đó bằng 0 và như vậy là kết quả sẽ bị sai lệch so với thực tế.

*Quy định 2:* Một ô số liệu bị thiếu nên được mã hoá bằng một giá trị âm (VD: -1), nó cũng cho biết lý do của việc thiếu số liệu đó.

Lý do của việc quy định này như sau:

- Nó cho phép loại các chỉ tiêu này trong xử lý bởi các công thức lọc dữ liệu.

- Nó có thể cho phép xử lý các thông tin bị thiếu này tuỳ thuộc vào lý do tại sao bị thiếu dữ liệu, ví dụ như nếu việc thiếu thông tin đó là do hộ không cung cấp thông tin hay hộ không có các thông tin đó để cung cấp cho chúng ta.

### 2.4. Xác định và xử lý các giá trị bị thiếu và vượt trội trong cơ sở dữ liệu

Các thông tin bị thiếu có thể do nhiều nguyên nhân khác nhau như: thiếu do quá trình thu thập thông tin hay thiếu do quá trình nhập thông tin.

Nếu thông tin bị thiếu là do quá trình điều tra, chúng ta cần phải làm rõ việc bị thiếu thông tin này là do người được phỏng vấn (nông dân, doanh nghiệp v.v...) hay do người đi phỏng vấn. Các thông tin bị khuyết này trong thực tế có thể có hai khả năng hoặc là thông tin đó hộ, doanh nghiệp không có để cung cấp cho chúng ta hoặc là có nhưng không cung cấp. Trong trường hợp thứ nhất khu vực trống đó được chấp nhận còn trong trường hợp thứ 2 chúng ta cần có biện pháp để xử lý thông tin bị thiếu này.

Nếu thông tin bị thiếu là do quá trình nhập dữ liệu thì chúng ta cần phải kiểm tra lại và bổ sung thông tin đó ngay trong quá trình kiểm tra.

Các thông tin vượt trội có nhiều dạng khác nhau, về cơ bản chúng ta có thể phân các kiểu giá trị vượt trội như sau:

- Các chỉ tiêu đơn lẻ vượt trội: Sự lệch của các giá trị đơn lẻ.

- Vượt trội của nhiều chỉ tiêu đồng thời: Sự sai lệch vượt trội của mối quan hệ.

Để xác định các giá trị vượt trội chúng ta cần phải so sánh với các tỷ lệ hoặc các giá trị đã có từ trước, điều này đòi hỏi phải có kinh nghiệm hoặc chúng ta cũng có thể sử dụng các công cụ thống kê truyền thống để xác định như dùng các kiểm định theo phân bố chuẩn hay phân tích sai số:

- David-Hartley-Pearson Test: Mối quan hệ của giá trị đến độ lệch chuẩn (chỉ dùng trong trường hợp một biến đơn lẻ).

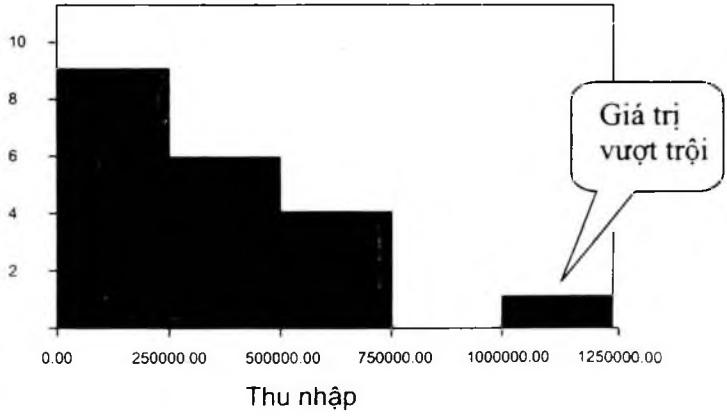
- Grubbs và Dixons R-Statistics: Mối quan hệ của giá trị đến giá trị bình quân (dùng trong trường hợp một giá trị đơn lẻ).

- Phân tích sai số: Từ các mô hình hồi quy (dùng trong trường hợp vượt trội của nhiều chỉ tiêu).

*Ví dụ:* Một số mô tả của sai số theo các hình thức khác nhau bằng cách sử dụng các công cụ thống kê mô tả để xác định:

| Số thứ tự hộ | Thu nhập/tháng | Số thứ tự hộ | Thu nhập/tháng |
|---|---|---|---|
| 1 | 56.400 | 11 | 256.350 |
| 2 | 72.154 | 12 | 302.250 |
| 3 | 85.300 | 13 | 340.466 |
| 4 | 95.700 | 14 | 360.050 |
| 5 | 96.800 | 15 | 380.000 |
| 6 | 112.000 | 16 | 504.813 |
| 7 | 115.331 | 17 | 543.875 |
| 8 | 160.059 | 18 | 575.269 |
| 9 | 185.950 | 19 | 689.375 |
| 10 | 263.800 | 20 | 1.248.563 |

Dùng các biểu đồ, đồ thị: Biểu đồ hình hộp, biểu đồ lá và thân

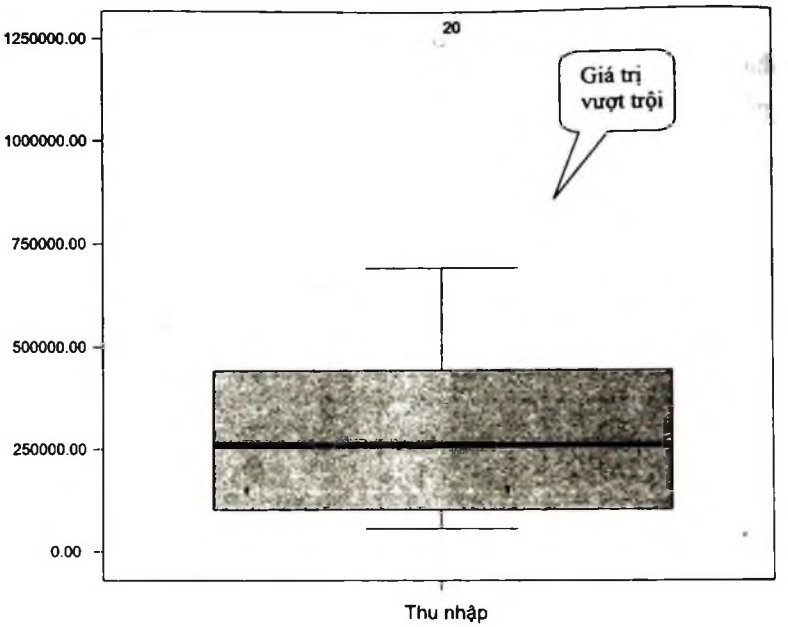


Histogram: Trình diễn số liệu như là một kiểu phân bố, sử dụng phần mềm SPSS qua đó ta có thể dễ dàng xác định được giá trị vượt trội.

```
REVPV Stem-and-Leaf Plot

 Frequency     Stem &  Leaf

     5,00        0 .  57899
     4,00        1 .  1168
     2,00        2 .  66
     4,00        3 .  0468
      ,00        4 .
     3,00        5 .  047
     1,00        6 .  8
     1,00 Extremes    (>=1248563)

 Stem width:     100000
 Each leaf:       1 case(s)
```


Biểu đồ thân lá: Một kiểu trình diễn phân bố số liệu, sử dụng phần mềm SPSS. Giá trị cuối cùng ứng với giá trị Extremes là giá trị vượt trội.

Biểu đồ hình hộp: Một kiểu trình diễn phân bố số liệu, sử dụng phần mềm SPSS. Giá trị thứ 20 nằm quá xa so với giá trị trung vị là giá trị vượt trội.




Xử lý các giá trị vượt trội và bị thiếu:

Việc xử lý các giá trị vượt trội và bị thiếu là như nhau, trước khi chúng ta đưa bất kỳ một thông tin nào vào thay thế cho các vị trí bị thiếu hay vượt trội này chúng ta cần phải tiến hành theo các bước cụ thể như sau:

(1) So sánh, đối chiếu với phiếu điều tra gốc: Nếu không phải lỗi do việc vào số liệu thì chúng ta chuyển sang bước thứ 2.

(2) Kiểm tra tính đúng đắn của thông tin: Nếu như xác suất xác định rằng giá trị vượt trội hoặc bị trống đó có thể là

giá trị đúng thì ta giữ nguyên nó trong cơ sở dữ liệu. Tuy nhiên, chúng ta cần lưu ý là trong trường hợp giá trị bị thiếu đó là đúng thì chúng ta cần phải có những ký hiệu riêng hay mã hoá riêng để tránh hiểu lầm giữa giá trị bị thiếu do không thu được thông tin hay không có thông tin đó.

Trong trường hợp chúng ta xác định được giá trị đó là không đúng thì chúng ta sẽ mã hoá nó như một giá trị bị thiếu để bổ sung.

Trong trường hợp thứ 2 chúng ta cần phải xác định những giá trị thay thế cho nó theo các phương pháp như sau:

(1) Phương pháp thay thế (cho các số liệu đơn lẻ): Phương pháp này có lợi thế là có thể sử dụng các phương pháp chuẩn để tính toán nhưng hạn chế là không tính đến việc gia tăng rủi ro khi sử dụng các giá trị thay thế đó. Để tiến hành theo phương pháp này chúng ta có thể sử dụng các giá trị sau đây để thay thế cho giá trị bị thiếu:

- Dùng giá trị bình quân hay trung vị.

- Lựa chọn một giá trị ngẫu nhiên của các mẫu có thể so sánh được ở ngay trong cuộc điều tra.

- Lựa chọn một giá trị ngẫu nhiên của các mẫu ở trong một cuộc điều tra khác.

- Sử dụng giá trị của mẫu liền kề với nó.

(2) Dùng các giá trị hồi quy (nếu như số liệu bị thiếu có mối quan hệ với nhiều chỉ tiêu khác).

Việc dùng các phương pháp khác nhau hoàn toàn tuỳ thuộc vào thực tế khả năng đáp ứng. Nếu chúng ta đã có các cuộc nghiên cứu trước đây thì có thể sử dụng các thông tin đó (khi không có sự biến động, tác động bởi yếu tố thời gian hoặc các yếu tố khác). Hoặc trong trường hợp mà có mối liên hệ thì chúng ta áp dụng phương pháp hồi quy. Tuy nhiên, chúng ta không lên quá lạm dụng vào việc thay thế các giá trị vượt trội hoặc bị thiếu, điều này chỉ nên diễn ra với một số lượng rất nhỏ các chỉ tiêu và quan sát. Cách tốt nhất để có cơ sở dữ liệu đáng tin cậy là chúng ta điều tra bổ sung.

*Ví dụ:*

| Số thứ tự hộ | Thu nhập/tháng | Thứ tự hộ | Thu nhập/tháng |
|---|---|---|---|
| 1 | 56.400 | 11 | 256.350 |
| 2 | 72.154 | 12 | 302.250 |
| 3 | 85.300 | 13 | 340.466 |
| 4 | 95.700 | 14 | 360.050 |
| 5 | 96.800 | 15 | 380.000 |
| 6 | 112.000 | 16 | 504.813 |
| 7 | 115.331 | 17 | 543.875 |
| 8 | 160.059 | 18 | 575.269 |
| 9 | 185.950 | 19 | 689.375 |
| 10 | 263.800 | 20 | -1 |

Kết quả tính toán giá trị thay thế và tác động:

| Cách thức thay thế | Giá trị đưa vào | Giá trị bình quân | Độ lệch chuẩn |
|---|---|---|---|
| Số liệu ban đầu | 1.248.563 | 322.675 | 237.615 |
| Giá trị bình quân (không bao gồm giá trị vượt trội) | 273.944 | 273.944 | 188.133 |
| Mode | Không thể, vì không có 2 hoặc hơn 2 giá trị bằng nhau | | |
| Trung vị | 263.800 | 273.437 | 188.147 |
| Giá trị gần nhất | 689.375 | 294.716 | 209.817 |
| Giá trị đưa vào bất kỳ, VD: lựa chọn ngẫu nhiên giá trị thứ 6 | 112.000 | 265.847 | 191.586 |
| Giá trị tương ứng từ quan sát khác | Không thể vì không có bất kỳ một phân phối nào tương tự | | |
| Kết quả từ việc chạy hàm hồi quy | 923.046 | 306.399 | 237.615 |
| Hàm hồi quy dựa trên quan hệ giữa số lượng lao động và thu nhập với<br>$R^2 = 0,805$ | | | |

# Chương III
# PHÂN TỔ VÀ KIỂM ĐỊNH THỐNG KÊ

## 3.1. Lý do của việc phân tổ

Việc phân tích và kết quả của nó dựa vào giá trị bình quân của 1 nhóm chỉ có ý nghĩa nếu như giá trị bình quân đó gần với giá trị của các cá thể riêng biệt trong thực tế.

Tuy nhiên, trong thực tế sự khác biệt giữa các cá thể về một chỉ tiêu nào đó thường khá lớn, vì vậy mục tiêu của việc phân tổ trong nghiên cứu là làm cho sự đồng nhất trong một nhóm và sự khác biệt giữa các nhóm tăng lên.

Dựa vào số lượng các chỉ tiêu và kiểu loại chỉ tiêu dùng để phân tổ, chúng ta có thể phân ra:

- Phân tổ theo 1 chỉ tiêu hay nhiều chỉ tiêu.

- Phân tổ theo chỉ tiêu định lượng hoặc chỉ tiêu định tính.

Chỉ tiêu phân tổ phù hợp sẽ diễn tả được bản chất của các hệ thống nghiên cứu theo vấn đề cần nghiên cứu.

- Các chỉ tiêu định tính: Sẽ phân định được rõ ràng và chỉ cần duy nhất 1 chỉ tiêu.

- Các chỉ tiêu định lượng: Khó khăn hơn trong việc xác định ranh giới giữa các nhóm với nhau.

*Chú ý:* Các chỉ tiêu có sự khác biệt lớn (biến động) giữa các mẫu sẽ có lợi hơn cho việc phân tổ so với các chỉ tiêu ít biến động.

*Ví dụ:* Khi sử dụng chỉ tiêu số nhân khẩu có thể giữa các hộ có sự khác biệt nhưng chỉ 1 hoặc 2 người, trong khi đó chỉ tiêu diện tích đất có thể khác nhau đến vài ha thì chúng ta nên sử dụng chỉ tiêu diện tích đất để phân tổ sẽ có lợi hơn trong việc phân các tổ để nghiên cứu.

### 3.2. Phân tổ thống kê theo các chỉ tiêu định tính

Việc phân tổ theo chỉ tiêu định tính được diễn ra một cách dễ dàng do khi phân tổ chúng ta chỉ có thể sử dụng được một chỉ tiêu duy nhất để phân tổ, việc sử dụng đến chỉ tiêu thứ hai là không cần thiết vì nhiều khi nó sẽ làm cho việc phân tổ trở nên không thể và không có nhiều ý nghĩa.

*Ví dụ:* Khi phân tổ theo loại hình sản xuất, các hộ thuần nông hoàn toàn được phân biệt với các hộ kiêm ngành nghề và do vậy, chúng ta không cần đến các chỉ tiêu khác để có thể phân các hộ thành các nhóm khác nhau.

Việc phân tổ theo chỉ tiêu định tính sẽ giúp chúng ta phân biệt rất rõ ràng một mẫu nào đó sẽ nằm trong tổ nào mà không phải đắn đo về đường ranh giới hay đường biên giữa các tổ, nhóm khi được phân.

*Ví dụ:* Khi phân tổ theo chỉ tiêu dân tộc thì rõ ràng các hộ thuộc nhóm dân tộc này thì không thể thuộc vào dân tộc kia được, do vậy mà ranh giới ở đây được phân biệt rất rạch ròi.

Việc phân tổ theo chỉ tiêu định tính được tiến hành như thế nào? Điều này hoàn toàn dựa vào mục đích nghiên cứu của người nghiên cứu và thực tế của số liệu điều tra.

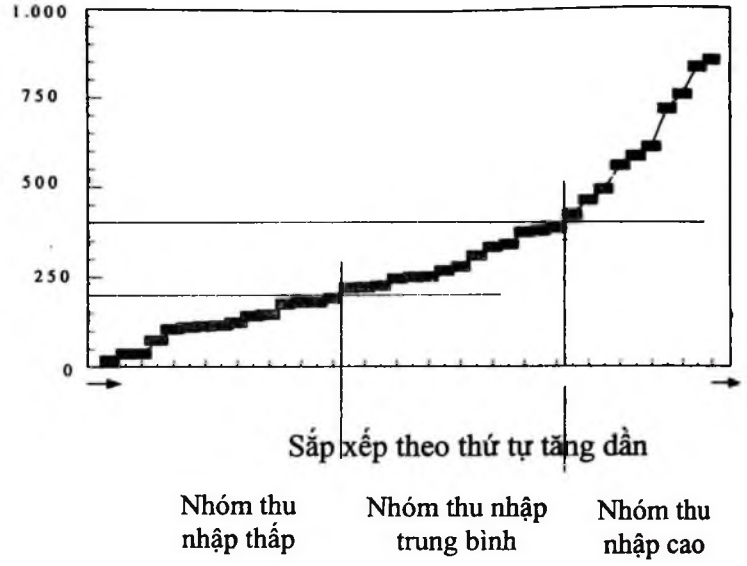
Việc phân tổ này có thể và thường được triển khai trước khi tiến hành điều tra để có thể triển khai việc lựa chọn mẫu theo tiêu chí phân tổ đó với mục đích đảm bảo đủ số lượng mẫu trong mỗi nhóm. Chính vì vậy, trước khi điều tra người nghiên cứu phải có những hiểu biết tối thiểu về khu vực nghiên cứu để hình dung xem liệu trên địa bàn có thể phân ra bao nhiêu nhóm mẫu đặc trưng cho những điểm khác nhau với mục đích tìm hiểu về vấn đề nghiên cứu trên địa bàn.

Trong quá trình phân tổ theo chỉ tiêu định tính cũng có thể được kết hợp với hình thức phân tổ theo chỉ tiêu định lượng.

*Ví dụ:* Sau khi điều tra mẫu theo các khu vực khác nhau (đây là phân tổ theo chỉ tiêu định tính) chúng ta có thể tiếp tục phân tổ theo chỉ tiêu thu nhập hay quy mô diện tích.

### 3.3. Phân tổ thống kê theo các chỉ tiêu định lượng

Các bước tiến hành phân tổ thống kê theo một hoặc nhiều chỉ tiêu định lượng bao gồm:




**Hình 3.1: PHÂN TỔ THEO 1 CHỈ TIÊU ĐỊNH LƯỢNG**

(1) Xác định chỉ tiêu (các chỉ tiêu) để phân tổ các mẫu điều tra.

(2) Sơ bộ xác định ranh giới giữa các nhóm.

(3) Sử dụng thêm chỉ tiêu thứ 2 hoặc thứ 3 trong trường hợp có những mẫu khó xác định rơi vào nhóm nào (trường hợp nằm trên đường biên).

(4) Tính toán hệ số biến động và khoảng cách giữa giá trị trung bình của các nhóm.

(5) Đi đến quyết định về số lượng nhóm và đường ranh giới. Nếu chưa đạt được thì thay đổi ranh giới và quay trở lại bước 4 (hình 3.1).

Để phân tổ theo nhiều chỉ tiêu định lượng chúng ta cần phải xác định được giữa các chỉ tiêu đó không có mối quan hệ hoặc là có nhưng mối quan hệ đó là rất nhỏ không đáng kể.

Tính được khoảng cách của sự khác biệt giữa các mẫu dựa vào đó chúng ta sẽ xác định được số lượng nhóm cần thiết cho nghiên cứu.

*Ví dụ:* Chúng ta có bảng số liệu của các hộ điều tra và theo 3 tiêu chí chúng ta muốn dùng để phân tổ các hộ theo các nhóm khác nhau về 3 tiêu chí này.

| Hộ số | Bò | Ha | De |
|---|---|---|---|
| 1 | 1 | 2 | 1 |
| 2 | 2 | 3 | 3 |
| 3 | 3 | 2 | 1 |
| 4 | 5 | 4 | 7 |

Tính toán hệ số đo lường khoảng cách khác biệt giữa các mẫu. Để tính toán khoảng cách biệt giữa hộ số 1 và hộ thứ k ta sử dụng công thức sau:

$$d^2_{kl} = \Sigma(v_n k - v_n l)^2$$

Trong đó: $v_n$ trình bày chỉ tiêu phân tổ.

*Ví dụ:* $d^2$ giữa trường hợp 1 và 5 theo số liệu đã cho ở ví dụ trước:

$$d^2 = (1-6)^2 + (2-7)^2 + (1-6)^2 = 75$$

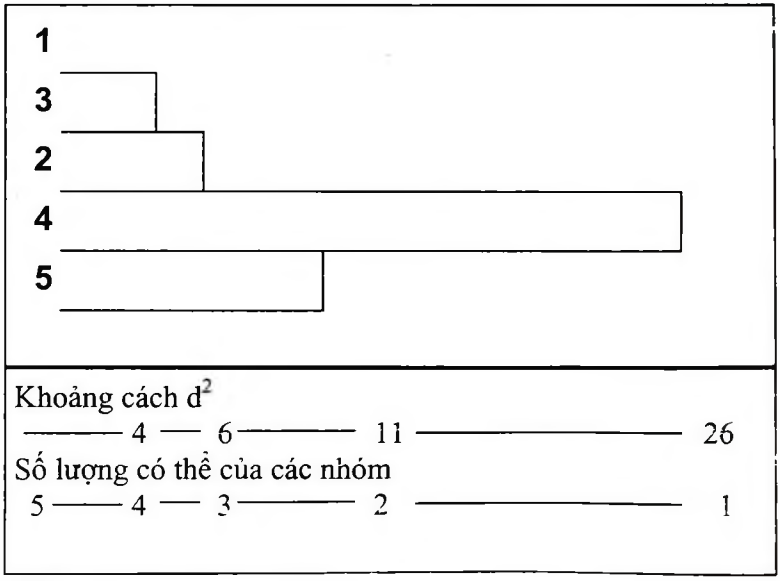
Ma trận khoảng cách, từ ví dụ trước ta xây dựng được ma trận khoảng cách sau:

| Hộ số | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | | | | |
| 2 | 6 | 0 | | | |
| 3 | 4 | 6 | 0 | | |
| 4 | 56 | 26 | 44 | 0 | |
| 5 | 75 | 41 | 59 | 11 | 0 |

Thuật toán phân nhóm: mối quan hệ đơn lẻ.

Là sự kết hợp các trường hợp theo khoảng cách ngắn nhất theo *chương trình phân nhánh.*

Từ kết quả tính toán tại ma trận khoảng cách ta có thể vẽ được đồ thị khoảng cách như sau:



Khoảng cách $d^2$ đánh giá cho sự đồng nhất, khoảng cách càng nhỏ thì độ đồng nhất trong cùng một nhóm càng cao.

Như vậy, theo cách phân tổ thống kê này chúng ta sẽ nhận được những nhóm khác nhau có đặc trưng như sau: giữa các nhóm có sự khác biệt lớn và trong cùng một nhóm có sự khác biệt ít nhất hay nói cách khác là có sự đồng nhất cao nhất.

Tóm lại: Hình thức phân tổ theo chỉ tiêu định lượng thường được tiến hành sau khi thu thập thông tin từ các quan sát. Chúng ta cũng cần lưu ý là giữa các chỉ tiêu dùng để phân tổ thống kê cần nhất thiết phải không có mối quan hệ tương quan với nhau và số lượng các tổ không nên quá nhiều (hơn 5 nhóm) và cũng không nên quá ít (ít hơn 3 nhóm) vì việc đó làm cho các nghiên cứu so sánh ít có ý nghĩa hơn hoặc là quá phức tạp, hơn nữa việc trình bày các báo cáo sẽ khó khăn và không đẹp mắt.

### 3.4. Kiểm định thống kê

Trong phân tổ thống kê việc so sánh để tìm hiểu các đặc trưng của các nhóm sau khi đã phân tổ là cần thiết và là một trong những mục đích chính của việc phân tổ. Chính vì vậy, để cho việc so sánh có căn cứ khoa học và có tính thuyết phục, việc kiểm định ý nghĩa thống kê của sự sai khác là cần thiết (vì thông thường ta hay sử dụng mẫu để nghiên cứu).

Trong kiểm định, chúng ta cần phải lưu ý với hai loại chỉ tiêu khác nhau: định tính và định lượng thì việc kiểm định cũng sẽ phải sử dụng các công cụ khác nhau.

### *3.4.1. Đối với các chỉ tiêu định tính*

Đối với các chỉ tiêu định tính việc phân tích kết quả thường thể hiện dưới dạng phần trăm hoặc tỷ lệ.

*Ví dụ:* Tỷ lệ hộ có nhà kiên cố/tổng số hộ; hay tỷ lệ hộ nghèo/tổng số hộ; v.v... đây là cách thức biểu diễn của các chỉ tiêu định tính trong phân tích.

Vì vậy, trong so sánh để có thể kiểm định ý nghĩa thống kê của sự sai khác chúng ta sử dụng bảng chéo và phân tích ngẫu nhiên:

*Ví dụ* một bảng chéo

| Y | X | | |
|---|---|---|---|
| | Nhóm 1 | Nhóm 2 | Tổng Y |
| Nhiều hơn 5 bò | 80 (40%) | 120 (60%) | 200 |
| 5 hoặc ít hơn 5 bò | 70 (70%) | 30 (30%) | 100 |
| Tổng của X | 150 | 150 | 300 |

Đối với các chỉ tiêu định tính việc kiểm định sẽ được tiến hành bằng sử dụng bảng chéo với kiểm định Chi-square.

### *3.4.2. Đối với các chỉ tiêu định lượng*

Sử dụng kiểm định T-Student cho các chỉ tiêu tuân theo luật phân bố chuẩn.

Trong trường hợp chúng ta chỉ kiểm định giả thuyết cho giá trị bình quân của hai nhóm độc lập có phân phối mẫu tuân theo

luật phân bố chuẩn chúng ta sẽ dùng kiểm định tại Independent-samples T-test. Đối với kiểm định này, chúng ta thường so sánh giữa hai nhóm mà bất kỳ sự khác biệt là do yếu tố chúng ta quan tâm (hoặc có hoặc không) chứ không phải do các yếu tố khác.

*Ví dụ:* Sẽ không phù hợp trong trường hợp chúng ta so sánh sự khác biệt về thu nhập giữa nhóm nam giới và nữ giới mà sử dụng kiểm định này vì rằng một người nào đó không phải ngẫu nhiên phân thành nam hay nữ. Trong trường hợp này, chúng ta phải lưu ý sự khác biệt ở các tiêu chí khác mà nó không che giấu hay làm nổi rõ sự khác biệt có ý nghĩa thống kê của các giá trị bình quân. Sự khác biệt của giá trị thu nhập bình quân có thể bị tác động bởi các yếu tố như trình độ học vấn chứ không phải bởi chỉ tiêu giới tính.

Trong trường hợp chúng ta có nhiều hơn hai nhóm độc lập với nhau có phân phối mẫu tuân theo luật phân phối chuẩn, chúng ta sử dụng công cụ One-Way ANOVA để kiểm định. Kiểm định ANOVA được dùng để kiểm định cho giả thuyết có nhiều nhóm với giá trị bình quân là như nhau. Kiểm định ANOVA được tiến hành bằng cách kiểm tra tỷ lệ của sự biến động giữa hai điều kiện và biến động trong cùng một điều kiện.

*Ví dụ:* Giả sử chúng ta có hai nhóm bệnh nhân khác nhau: một nhóm được chữa theo phương pháp riêng và một nhóm chữa theo phương pháp chung. Kiểm định ANOVA sẽ so sánh sự biến động mà ta quan sát được giữa hai nhóm bệnh nhân với sự thay đổi bên trong của từng nhóm bệnh nhân.

Trong trường hợp nếu phân bố của mẫu không phải là phân bố chuẩn thì chúng ta sử dụng kiểm định phi tham số (Nonparametric-test) với 2 dạng kiểm định khác nhau: Kiểm định KRUSKAL-WALLIS sử dụng trong trường hợp so sánh nhiều hơn hai nhóm độc lập với nhau và Friedman sử dụng trong trường hợp có hơn hai nhóm phụ thuộc lẫn nhau; hoặc ta dùng kiểm định MANN-WHITNEY trong trường hợp so sánh hai nhóm là độc lập và kiểm định Wilcoxon sử dụng trong trường hợp so sánh hai nhóm là phụ thuộc với nhau.

Tuy nhiên, trong cả 2 trường hợp kiểm định giữa các nhóm mà phân phối của mẫu tuân theo phân bố chuẩn hoặc không tuân theo phân phối chuẩn chúng ta đều có thể sử dụng kiểm định phi tham số cho việc kiểm định sự sai khác có ý nghĩa thống kê được.

| Chi phí (1000 đ) | | Nhóm hộ | | | Xác suất của sự khác biệt Có ý nghĩa thống kê | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | tất cả[1] | 1-3[2] | 1-2[3] |
| | | | | | % | % | % |
| Củi đốt | μ | 285.60 | 438.00 | 583.20 | 97 | 99 | 87 |
| | ($Cl_{90}$ ± | 55.17 | 129.83 | 138.02) | | | |
| | Trung vị | 252.00 | 342.00 | 504.00 | | | |
| Than | μ | 242.15 | 1,100.00 | 88.72 | 100 | 58 | 100 |
| | ($Cl_{90}$ ± | 183.74 | 456.14 | 60.77) | | | |
| | Trung vị | 0.00 | 1,.001.00 | 0.00 | | | |
| Dầu | μ | 52.40 | 0.00 | 64.80 | 45 | 19 | 74 |
| | ($Cl_{90}$ ± | 72.09 | 0.00 | 72.82) | | | |
| | Trung vị | 0.00 | 0.00 | 0.00 | | | |

*(Tiếp theo)*

| Chi phí (1000 đ) | | Nhóm hộ | | | Xác suất của sự khác biệt Có ý nghĩa thống kê | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | tất cả[1] | 1-3[2] | 1-2[3] |
| | | | | | % | % | % |
| Ga | μ | 179.67 | 178.22 | 181.70 | 14 | 16 | 44 |
| | ($Cl_{90}$ ± | 81.32 | 169.06 | 80.51) | | | |
| | Trung vị | 144.00 | 145.25 | 75.00 | | | |
| Điện | μ | 254.67 | 665.78 | 490.50 | 79 | 82 | 88 |
| | ($Cl_{90}$ ± | 115.00 | 419.50 | 277.58) | | | |
| | Trung vị | 250.00 | 500.00 | 400.00 | | | |
| Tổng số | μ | 1,050.48 | 2,404.22 | 1,495.92 | 99 | 93 | 100 |
| | ($Cl_{90}$ ± | 317.66 | 494.20 | 336.63) | | | |
| | Trung vị | 810.00 | 2,464.00 | 1,150.40 | | | |

*Ghi chú:*
1 - Sự khác biệt của cả 3 nhóm theo Kruskal-Wallis
2 và 3 - Kiểm định sự khác biệt giữa nhóm 1 và 2 và giữa nhóm 1 và 3 theo Mann-Whitney

### *3.4.3. Ý nghĩa và sự giải thích của giá trị xác suất P (P-values) (số liệu nói lên điều gì?)*

Giá trị P phụ thuộc trực tiếp vào mẫu nghiên cứu, nhằm cung cấp độ chắc chắn của kết quả kiểm định, cho phép đưa ra kết luận là bác bỏ hay không bác bỏ giả thuyết đưa ra. Nếu giả thuyết $H_o$ (giả thuyết không có sự khác biệt giữa các nhóm) là

đúng và sự biến động ngẫu nhiên là lý do cho sự khác biệt của các mẫu, khi đó giá trị P là mức đo mà căn cứ vào đó chúng ta có thể đưa ra quyết định chấp thuận hay không. Bảng 3.1 dưới đây sẽ cung cấp sự giải thích ý nghĩa của giá trị P (P-values):

**Bảng 3.1: MỨC Ý NGHĨA CỦA GIÁ TRỊ P**

| P-value | Giải thích |
|---|---|
| P< 0,01 | Căn cứ mạnh để bác bỏ $H_o$ |
| 0,01≤ P < 0,05 | Mức độ vừa phải trong việc bác bỏ $H_o$ |
| 0,05≤ P < 0,10 | Mức độ yếu trong việc bác bỏ $H_o$ |
| 0,10≤ P | Rất yếu hoặc không có căn cứ để bác bỏ $H_o$ |

Sự giải thích trên được áp dụng rộng rãi, nhiều nhà nghiên cứu đã áp dụng nó trong việc kiểm định ý nghĩa thống kê của các giả thuyết đưa ra trong nghiên cứu của mình.

Thông thường, cho một mẫu cho trước có phân phối đồng dạng. Chúng ta có thể minh hoạ giá trị $P(p \leq x) = x$, có nghĩa là p <0,05 tương đương với $\alpha = 0,05$. Khi giá trị P được minh hoạ cho một bộ số liệu nào đó thì có nghĩa là bộ số liệu đó được lấy ra một cách ngẫu nhiên từ một tổng thể mô tả bởi một kiểm định thống kê.

Giá trị P là căn cứ để chúng ta bác bỏ giả thuyết $H_o$, giá trị này càng nhỏ thì ta càng có căn cứ để bác bỏ giả thuyết hơn. Người ta cũng có thể sử dụng giá trị P làm mức độ ý nghĩa thống kê của việc bác bỏ giả thuyết, trong trường hợp này giá trị P phải nhỏ hơn một ngưỡng giá trị (thường thì 0,05, thi thoảng cao hơn 0,1 hoặc thấp hơn 0,01) thì chúng ta bác bỏ giả thuyết $H_o$.

## Chương IV

# PHÂN TÍCH SỐ LIỆU ĐIỀU TRA VÀ BIỂU DIỄN KẾT QUẢ

Nội dung của chương này nhằm cung cấp cho người đọc những kiến thức cơ bản về thống kê khi áp dụng vào trong phân tích các kết quả xử lý số liệu điều tra. Với mục đích là sau khi ứng dụng chúng ta không những chỉ nêu ra được những điều chính xác mà nhà nghiên cứu cần phải có mà còn có thể giải thích những điều mà số liệu có thể trình bày cho ta.

Thống kê là toán học của việc tổ chức và giải thích các thông tin dạng số. Kết quả của những phân tích thống kê là các mô tả, các so sánh, các dự báo, các mối quan hệ v.v...

*Ví dụ:* Một cuộc điều tra 160 người để tìm hiểu xem số lượng sách và loại sách mà họ đọc. Về mặt thống kê điều tra làm các công việc sau:

- Mô tả yếu tố căn bản của những người được phỏng vấn.

- Mô tả câu trả lời cho các câu hỏi.

- Xác định nếu có sự tồn tại sự liên quan giữa số lượng sách họ đọc và việc đi du lịch trong năm qua.

- So sánh lượng sách mà người nam giới và nữ giới đọc trong năm qua.

- Tìm hiểu vấn đề về giới, trình độ học vấn hoặc thu nhập liên quan đến việc đọc sách của người được phỏng vấn.

Kết quả được minh hoạ như sau:

1. Những yếu tố căn bản của người phỏng vấn: Trong 160 người được hỏi có 77 (48,1%) là nam giới, trong đó 72 (48%) thu nhập hàng năm hơn 20 triệu và có ít nhất 2 năm đi làm.

2. Trả lời cho các câu hỏi: Với câu hỏi bao nhiêu sách anh/chị đọc trong 1 năm qua và liệu họ thích đọc tiểu thuyết hơn hay các sách khác. Trung bình một người tốt nghiệp đại học đọc hơn 10 cuốn sách một năm, với mức dao động từ ít nhất 2 cuốn đến 50 cuốn. Họ thích đọc các sách không phải là tiểu thuyết hơn các sách tiểu thuyết.

3. Mối quan hệ giữa đi du lịch và đọc sách: Người ta được hỏi mức độ thường xuyên đi du lịch trong năm qua. Tần suất đi du lịch có thể được so sánh với mức độ đọc sách của họ, người đi du lịch nhiều sẽ đọc sách nhiều hơn.

4. So sánh: Tỷ lệ phần trăm của nam và nữ có đọc hơn 5 cuốn sách được so sánh với nhau. Kết quả, tỷ lệ nữ thích đọc sách cao hơn hẳn và có ý nghĩa thống kê khi so sánh với nam giới.

5. Dự báo tần suất: Học vấn và thu nhập được tìm thấy là các yếu tố quan trọng trong việc dự báo tần suất/mức độ đọc sách. Người có trình độ học vấn và có thu nhập cao hơn sẽ đọc sách nhiều hơn.

Như vậy, trong ví dụ này chúng ta thấy những kết quả của việc phân tích thể hiện dưới dạng tần số hay tỷ lệ phần trăm, kết quả cũng được thể hiện dưới dạng giá trị bình quân và độ dao động. Trong câu hỏi thứ 3 việc phân tích mối quan hệ được tiến hành giữa 2 tiêu chí đi du lịch và đọc sách mà một cách để có thể phân tích mối liên hệ này là phân tích tương quan (correlation).

Ở kết quả thứ 4 cho thấy sự so sánh giữa hai nhóm nam và nữ. Từ "Có ý nghĩa thống kê" ở đây được dùng nhằm chỉ ra rằng chúng có giá trị về mặt thống kê (tức là nó sẽ có kết quả tương tự như vậy khi ta lựa chọn một mẫu khác để nghiên cứu) chứ không phải do cơ hội (may mắn) mà có sự khác biệt này.

Ở kết quả thứ 5, việc phân tích nhằm tìm ra xu hướng và cũng trả lời cho câu hỏi tiêu chí nào có liên quan đến việc đọc sách, chẳng hạn như thu nhập có làm cho mức độ đọc sách khác nhau không? Hay trình độ văn hoá v.v... điều này cũng giúp ta vẽ ra được những xu hướng hay xu thế trong tương lai khi ta tác động vào một yếu tố này thì kết quả sẽ ra sao.

Phân tích số liệu thống kê phân nhóm các phương pháp phân tích số liệu ra làm hai nhóm: Nhóm các phương pháp thăm dò và nhóm các phương pháp khẳng định. Nhóm các phương pháp thăm dò thường được sử dụng để phát hiện, khám phá những điều mà số liệu có thể nói cho ta biết bằng cách sử dụng số học cơ bản và dễ dàng giúp ta vẽ lên những bức tranh tóm lược từ những con số. Nhóm các phương pháp khẳng định sử dụng ý tưởng từ lý thuyết xác suất bằng các phép thử để trả lời cho những câu hỏi cụ thể. Xác suất có ý

nghĩa quan trọng trong việc đưa ra quyết định, bởi vì nó cung cấp một cơ chế cho việc đo lường, biểu diễn và phân tích những điều không biết trước có liên quan đến các sự kiện trong tương lai.

Vậy, phương pháp nào giúp ta mô tả, tổng hợp, so sánh và dự đoán? Để trả lời cho câu hỏi này trước hết chúng ta sẽ phải trả lời cho các câu hỏi sau: số liệu ta thu được ở dạng định tính hay định lượng? Bao nhiêu tiêu chí là độc lập và phụ thuộc? Việc tiến hành điều tra thống kê của chúng ta có phù hợp cho việc áp dụng các phương pháp thống kê hay không? (các câu hỏi này đã được chúng ta trả lời trong những chương trước).

### 4.1. Mô tả thống kê

Mô tả thống kê là cách thức miêu tả số liệu dưới dạng số trung bình, trung vị hay mode. Những con số này thể hiện giá trị trung tâm của các phân phối. Thông thường trong một phân phối bao gồm nhiều các giá trị (chẳng hạn như: điểm số cho một giá trị số nào đó như số năm công tác và tuổi đời) của một biến số nào đó, như thái độ, hiểu biết, tình trạng sức khoẻ, thu nhập v.v...

Ngoài ra, mô tả thống kê còn sử dụng các mức độ biến động như độ lệch chuẩn để minh hoạ

Giá trị trung bình: Là bình quân số học của các quan sát. Ký hiệu của giá trị bình quân là $\overline{X}$. Giá trị này được tính bằng cách lấy tổng các giá trị của từng quan sát chia cho số lượng các quan sát và công thức toán học được thể hiện như sau:

$$\sum X / n$$

*Trong đó:*

$\sum X$ là tổng các giá trị của từng quan sát.

n là số lượng các quan sát trong nghiên cứu.

Giá trị trung vị (median): Là giá trị của quan sát có vị trí được sắp xếp ở giữa theo cách sắp xếp giá trị từ nhỏ đến lớn hoặc ngược lại. Vì thế, một nửa số quan sát sẽ có giá trị nhỏ hơn giá trị trung vị và một nửa có giá trị lớn hơn giá trị trung vị. Giá trị trung vị đôi khi được xem như là một quan sát đặc biệt. Để xác định giá trị trung vị chúng ta làm như sau:

(1) Sắp xếp lại các quan sát theo giá trị từ nhỏ nhất đến lớn nhất (hoặc ngược lại).

(2) Tìm quan sát có vị trí trung tâm. Trong trường hợp số quan sát là lẻ ta sẽ có một quan sát ứng với giá trị trung vị đó, còn trong trường hợp số quan sát là chẵn ta sẽ lấy giá trị bình quân của 2 quan sát đứng giữa làm giá trị trung vị.

Giá trị trung vị sẽ không bị tác động nhiều trong trường hợp mà có giá trị vượt trội như đối với giá trị bình quân, vì thế trong trường hợp có một vài giá trị vượt trội chúng ta nên dùng giá trị trung vị để thay thế cho giá trị bình quân trong phân tích.

Giá trị mode: Giá trị mode là giá trị của những quan sát được xuất hiện nhiều nhất. Nó thường được sử dụng khi ta muốn tìm hiểu giá trị thường gặp nhất.

### 4.2. Phân phối mẫu: lệch và đối xứng

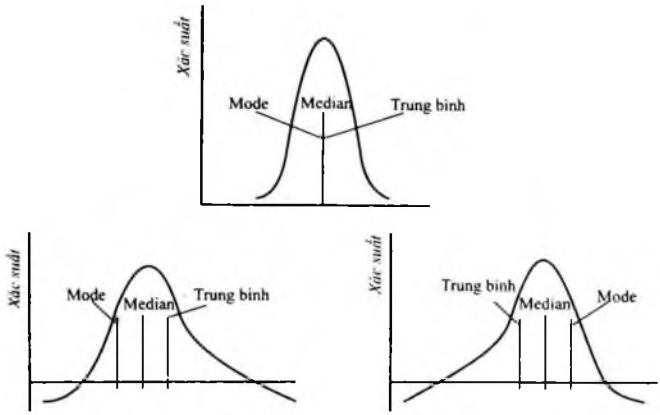
Phân phối mẫu rất quan trọng trong việc kiểm định thống kê. Nó được hình thành và tính toán từ việc lấy mẫu lặp lại. Ví dụ sau đây sẽ làm rõ khái niệm về phân phối mẫu:

Giả sử chúng ta đã biết về trọng lượng trung bình và độ lệch chuẩn của nhóm nam thanh niên nước ta (Trọng lượng trung bình là 60kg, độ lệch chuẩn là 5,4kg). Nếu bây giờ chúng ta lựa chọn ngẫu nhiên 40 nam giới và xác định trọng lượng cơ thể của họ, chúng ta sẽ có một giá trị bình quân riêng, giá trị này gần với bình quân chung của tổng thể. Lặp lại nhiều lần lấy mẫu tương tự, kết quả là chúng ta sẽ có một đồ thị phân bố của các giá trị bình quân có thể tiệm cận với sự kiện chuẩn.

Có một vài phân phối mẫu thường gặp mà chúng ta rất hay sử dụng trong tính toán thống kê và kiểm định thống kê đó là:

- Phân phối chuẩn.
- Phân phối Chi-square.
- Phân phối Student.
- Phân phối Fishery.

Khi ta có một phân phối mà trong đó có một vài giá trị quan sát vượt trội về một phía nào đó - một vài giá trị rất nhỏ hoặc một vài giá trị rất lớn, khi đó chúng ta sẽ có dạng phân phối lệch. Còn đối với một phân phối cân đối là no có cùng dạng ở cả hai phía của giá trị bình quân.




**Hình 4.1: CÁC DẠNG PHÂN PHỐI MẪU**

Bảng tổng hợp cho việc sử dụng các giá trị trung bình, trung vị và mode:

(1) Sử dụng giá trị trung bình khi:

- Phân phối mẫu gần với phân phối cân đối.

- Đối với các giá trị dạng số.

(2) Sử dụng giá trị trung vị khi:

- Đang làm việc với các giá trị cho điểm (chỉ tiêu định tính cho điểm).

- Phân phối mẫu có dạng lệch.

- Có số liệu dạng thứ tự.

(3) Sử dụng mode khi:

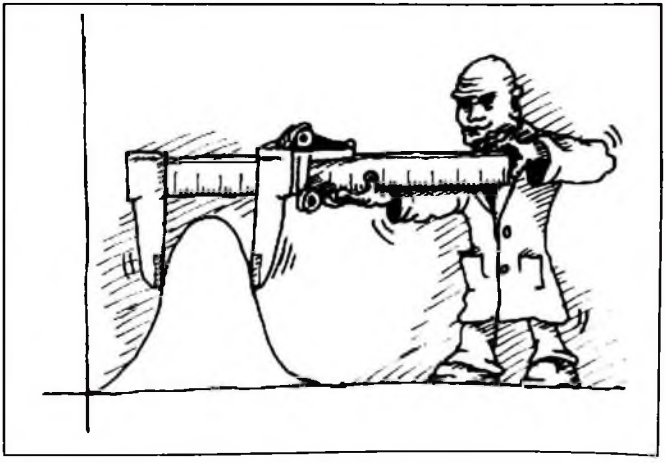
- Khi phân phối có hai hoặc nhiều hơn hai đỉnh.

- Chúng ta muốn chỉ ra giá trị, đặc trưng hoặc chỉ tiêu thịnh hành.

### 4.3. Đo lường sự biến động

Giả sử ta đang nghiên cứu hỏi mọi người về chất lượng bữa ăn của một nhà hàng. Kết quả cho thấy giá trị bình quân mọi người đánh giá là 3,5 trong khoảng từ 1 (kém) đến 5 (rất ngon). Câu hỏi đặt ra ở đây là liệu tất cả mọi người trả lời đều gần với 3 hay không hay là một số người cho điểm là 1 và phần còn lại là 5 điểm? Để xác định khoảng dao động chúng ta có thể sử dụng khoảng biến thiên, độ lệch chuẩn.

Khoảng biến thiên là sự khác biệt giữa giá trị lớn nhất và giá trị nhỏ nhất. Còn giá trị độ lệch chuẩn là đo lường mức độ dao động của số liệu quanh giá trị bình quân mà đây chính là bản chất của nhiều kiểm định thống kê.



Độ lệch chuẩn được xác định qua công thức $\sqrt{\sum(X-\overline{X})^2/(n-1)}$, bình phương của độ lệch chuẩn là hương sai (variance). Tuy nhiên, giá trị phương sai ít khi ược sử dụng là vì:

- Liên quan đến phân bố của các quan sát, tối thiểu 75% ủa các quan sát sẽ nằm trong khoảng giá trị trung bình ± 2 ần độ lệch chuẩn.

- Kiểm tra xem liệu dạng phân phối có phải dạng cân đối ay dạng chuẩn khi đó: 68% của các quan sát nằm trong hoảng giá trị bình quân ± 1 lần giá trị độ lệch chuẩn; 95% ủa các quan sát nằm trong khoảng giá trị bình quân ± 2 lần giá trị độ lệch chuẩn; 99% của các quan sát nằm trong khoảng giá trị bình quân ± 3 lần giá trị độ lệch chuẩn.

### 4.4. Tương quan và mối quan hệ

Tương quan chỉ ra mối quan hệ qua lại giữa các nhân tố với nhau trong quan hệ tương quan không có chứa quan hệ nhân quả, tức là A tác động đến B và ngược lại từ B ta cũng có thể suy ra được A.

Quan hệ tương quan khác với quan hệ hồi quy ở chỗ quan hệ hồi quy là quan hệ nhân quả trong đó một yếu tố chịu sự tác động của một hoặc một vài yếu tố khác.

#### *4.4.1. Số liệu dạng định lượng*

Một mối quan hệ có thể tồn tại giữa hoặc trong các biến. giả sử ta có một phiếu điều tra với những câu hỏi khác nhau như sau:

1. Người đọc của tạp chí này có tiềm lực tài chính cao?

2. Hai người được hỏi sẽ đồng ý về những gì họ thấy?

3. Những người lập trình giỏi phải là những người c trình độ tiếng Anh tốt?

Những câu hỏi dạng này là câu hỏi về mối quan hệ. Kh chúng ta liên quan đến các câu hỏi dạng liên hệ như thế nà tức là chúng ta cần phải phân tích mối quan hệ. Khi số liệ của cả 2 biến đều là định lượng chúng ta có thể sử dụng hệ s tương quan r (Correlation coefficient). Hệ số tương quan da động trong khoảng từ -1 đến +1.

Xem xét hai biến X và Y. Giả sử X là biến độc lập và Y là biến phụ thuộc, nếu ta tìm thấy hệ số tương quan giữa X v Y là +1 có nghĩa là khi giá trị của X tăng lên một lượng thì gi trị của Y cũng tăng lên một lượng tương ứng. Ngược lại vớ hệ số tương quan r = -1 khi X tăng lên 1 đơn vị thì Y sẽ giảr đi 1 đơn vị tương ứng. Trong trường hợp R = 0 có nghĩa l giữa X và Y không có mối quan hệ nào với nhau.

Hệ số tương quan được tính bởi công thức sau:

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sqrt{\sum (Y - \overline{Y})^2}}}$$

Hệ số tương quan chỉ đo lường mối quan hệ tuyến tính do vậy trong những trường hợp không phải là quan hệ tuyế tính thì việc tính toán hệ số tương quan sẽ không có ý nghĩa Để biết được có phải là quan hệ tuyến tính hay không chún

ta có thể dùng đồ thị để thể hiện mối quan hệ giữa hai yếu tố mà chúng ta đang quan tâm.

Mức độ tương quan, theo quy luật ngón tay cái ta có:

Nếu r từ 0 tới +0,25 hoặc -0,25 tương đương với tương quan yếu hoặc không có tương quan với nhau.

Từ +0,26 đến +0,50 (hoặc từ -0,26 đến -0,50) mức tương quan trung bình.

Từ +0,51 đến +0,75 (hoặc từ -0,51 đến -0,75) tương quan khá.

Trên 0,75 (hoặc nhỏ hơn -0,75) tương quan chặt.

Tuy nhiên, đối với một số khoa học xã hội mà mối tương quan có r từ 0,26 đến 0,50 đã cho là có tương quan khá, đặc biệt trong mối quan hệ đa bội.

### *4.4.2. Dữ liệu dạng định tính*

Đối với dữ liệu định tính việc sử dụng tương quan có sắp xếp Spearman để phân tích, đôi khi có tài liệu sử dụng Spearman's rho. Công cụ này thường được sử dụng mô tả mối quan hệ giữa hai yếu tố định tính hoặc giữa yếu tố định tính và định lượng.

Công cụ này cũng được sử dụng với các số liệu dạng định lượng trong trường hợp phân bố của số liệu không tuân theo luật phân bố chuẩn hay thuộc loại phân bố lệch với các giá trị vượt trội. Trong thực tế nếu giá trị trung vị được sử dụng thay

thế cho giá trị bình quân thì việc sử dụng Spearman's rho đúng đắn nhất.

Ký hiệu của hệ số tương quan trong phân tích tương qu Spearman's rho là $r_s$. Để tính toán $r_s$ chúng ta cần phải sắp x số liệu từ bé đến lớn hoặc ngược lại. Hệ số $r_s$ từ -1 đến +1 cho chúng ta thấy mối quan hệ của sự sắp xếp hơn là m quan hệ của các con số.

### *4.4.3. Phương pháp hồi quy và phân tích nhân tố*

#### *4.4.3.1. Lý thuyết chung*

Việc áp dụng các công cụ toán học trong kinh tế và c khoa học xã hội rất phổ biến và có ý nghĩa trong nghiên cú nó giúp chúng ta đưa ra những quyết định đúng đắn.

Trong các phương pháp định lượng áp dụng trong kinh - xã hội chúng ta có thể dùng các hàm hồi quy trong kinh lượng để nghiên cứu các mối quan hệ, các tác động và ph tích dự báo.

#### *4.4.3.2. Các hàm cơ bản*

Các hàm sử dụng trong phân tích hồi quy phân ra 2 lc chính là: (1) Hàm hồi quy dạng tuyến tính và (2) Hàm hồi q dạng phi tuyến.

Ở đây, sẽ trình bày một số dạng hàm đơn giản thường g trong ứng dụng thực tiễn.

(1) Hàm hồi quy tuyến tính:

Mô hình hồi quy tuyến tính được hiểu theo nghĩa tuyến tính đối với các tham số. Nó có thể tuyến tính hoặc không tuyến tính đối với các biến. Dạng của hàm hồi quy là một vấn đề quan trọng, một trong những nhân tố có tính chất quyết định đối với kết quả nghiên cứu. Tuy vậy, vấn đề dạng của hàm hồi quy lại không có một cơ sở lý thuyết đủ mạnh để có thể khẳng định dạng của hàm hồi quy là dạng này mà không phải là dạng khác. Hay nói một cách khác dạng hàm của mô hình hồi quy là một vấn đề thực nghiệm.

Một trong những phương pháp thường được dùng là biểu diễn các số liệu lên hệ toạ độ. Nếu như đồ thị chỉ ra quan hệ giữa hai biến là tuyến tính thì dạng hàm của mô hình là tuyến tính, nếu quan hệ được chỉ ra là hàm bậc 2 (phi tuyến) thì dạng hàm của mô hình được chọn một cách tương ứng. Phương pháp này được sử dụng trong mô hình hồi quy giản đơn. Nó sẽ là không hữu ích nếu chúng ta có mô hình hồi quy bội.

Về mặt toán học hàm tuyến tính thường được biểu diễn như sau:

$$Y = a + bX$$

Đây là hàm tuyến tính đơn, trong đó Y là biến phụ thuộc hay còn gọi là biến được giải thích; X là biến độc lập hay biến giải thích; a và b là các tham số mô tả hàm.

Chúng ta cũng có thể gặp hàm hồi quy tuyến tính với nhiều biến giải thích như:

$$Y = a + b_1X_1 + b_2X_2 + .... + b_nX_n$$

Trong mô hình hồi quy tuyến tính để ước lượng được cá tham số của mô hình người ta thường hay dùng phương phá bình phương bé nhất (OLS) để ước lượng, vì phương phá này sẽ cho ta những ước lượng không chệch tốt nhất.

Khi dùng phương pháp OLS chúng ta được:

$$b = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \quad \text{và} \quad a = \overline{Y} - b\overline{X}$$

Trong đó $y_i = Y_i - \overline{Y}$ và $x_i = X_i - \overline{X}$

(2) Hàm có hệ số co giãn không đổi (hàm Cobb Douglas)

Hàm Cobb - Douglas có dạng $Y = aX^b$

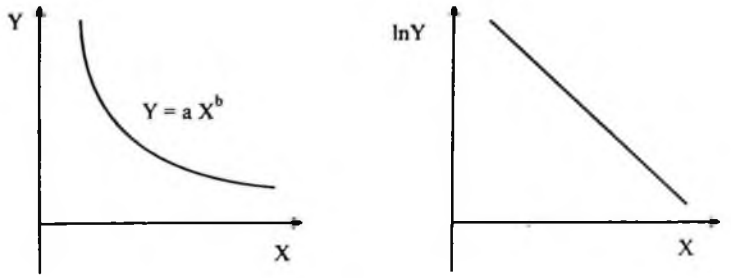
Hàm này là phi tuyến đối với X và phi tuyến đối với thar số b. Tuy nhiên, có thể biến đổi về dạng tuyến tính đối vớ tham số. Lấy ln hai vế, ta có:

$\ln Y = \ln a + b \ln X$

Đặt $a' = \ln a$; $Y' = \ln Y$; $X' = \ln X$

Ta có: $Y' = a' + bX'$

Đây là mô hình tuyến tính giản đơn đã biết. Ta có thể mini hoạ hàm ban đầu và hàm sau khi biến đổi qua hình 4.2 sau:




**Hình 4.2: HÀM TUYẾN TÍNH VÀ HÀM PHI TUYẾN**

Hàm Cobb - Douglas có thể mở rộng cho trường hợp có nhiều biến giải thích:

$$Y = aX_1^{b_1} X_2^{b_2} .....X_n^{b_n}$$

Bằng phép biến đổi: $\ln Y = \ln a + b_1 \ln X_1 + b_2 \ln X_2 + ..... + b_n \ln X_n$ chúng ta dễ dàng có hàm tuyến tính đối với các tham số. Trong hàm Cobb - Douglas, hệ số co giãn của Y đối với $X_i$ bằng $b_i$.

3) Hàm có dạng $Y_t = b(1+r)^t$

Hàm có dạng $Y_t = b(1+r)^t$, trong đó t là thời gian. Hàm này thường dùng để đo sự tăng trưởng của yếu tố $Y_t$ theo thời gian, r là tỷ lệ tăng trưởng.

Ở năm (thời kỳ) $t = 0$, ta có $Y_0 = b$, do đó $Y_t = Y_0(1+r)^t$

Biến đổi hàm về dạng tuyến tính đối với tham số:

$\ln Y_t = \ln Y_0 + t \ln(1+r)$

Đặt $Y_t' = \ln Y_t$ ; $a = \ln Y_0$ ; $b = \ln(1 + r)$

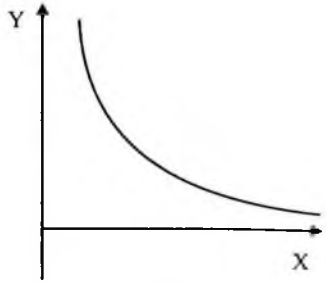
Khi đó: $Y_t' = a + b\,t$

Dễ dàng ước lượng được hàm này và từ đó tìm được Y và r.

4) Hàm dạng Hypecbol $Y = a + \frac{b}{X}$

Hàm này là phi tuyến đối với X, nhưng tuyến tính đối vớ các tham số. Trường hợp a,b > 0, khi đó đồ thị có dạng qua chiều cong xuống dưới, trường hợp này có mức tiệm cậ dưới, dù có tăng đến đâu Y không thể nhỏ hơn b.

Hàm này thường được dùng khi phân tích chi phí trun bình của đầu tư để sản xuất ra một sản phẩm.

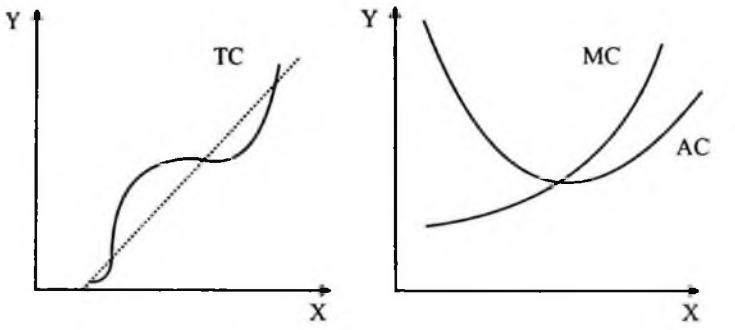



**Hình 4.3: HÀM DẠNG HYPECBOL**

(5) Hàm có dạng đa thức: $Y = a + b_1X + b_2X^2 + b_3X^3$

Hàm này thường được sử dụng để nghiên cứu quan h giữa chi phí đầu vào của đất đai và số lượng sản phẩm được

sản xuất ra. Chẳng hạn Y - tổng chi phí, X - số lượng sản phẩm được sản xuất ra. Nếu như xây dựng được hàm này thì ta dễ dàng tìm được chi phí trung gian và chi phí biên.




**Hình 4.4: HÀM ĐA THỨC**

TC: Tổng chi phí; MC: Chi phí biên; AC: Chi phí trung bình

Với mô hình này, chúng ta có thể biết được khi nào chúng ta có mức đầu tư trên một đơn vị diện tích để đạt được hiệu quả đầu ra là cao nhất.

Trên đây đã trình bày một số dạng mô hình hồi quy. Thực tế còn rất nhiều dạng khác. Trong thực tế, để vận dụng mô hình này hay mô hình khác trước hết phải hiểu được mối quan hệ giữa các biến, tính chất của các mô hình (các dạng hàm) muốn vận dụng.

*4.4.3.3. Hàm sản xuất*

Là hàm hồi quy trong đó biểu hiện mối quan hệ giữa đ vào và kết quả đầu ra của một quá trình sản xuất.

*Ví dụ:* Mối quan hệ giữa lượng phân bón trên một đơn vị diện tích nào đó với sản lượng sản phẩm đầu ra của một loại cây trồng là một hàm sản xuất.

Trong phân tích hàm sản xuất người ta thường dùng c dạng hàm tuyến tính, Cobb-Douglas hay các hàm đa thức.

Có thể mô tả hàm sản xuất bằng ngôn ngữ toán học:

$Y = f(X_1, X_2, X_3, ..., X_n)$

*Trong đó:*

Y là đầu ra (có thể là sản lượng một loại sản phẩm n đó).

$X_1, X_2, X_3, ..., X_n$ là các yếu tố đầu vào. Trong số c yếu tố đầu vào có thể định lượng được như phân bó thuốc trừ sâu v.v..., hoặc các yếu tố đầu vào không t định lượng được (hay là yếu tố định tính) như quản lý

Hàm sản xuất còn được gọi là hàm đáp ứng vì khi tăn hay giảm các yếu tố sử dụng nó sẽ có các đáp ứng của đầu r Các kiểu đáp ứng điển hình là đáp ứng tuyệt đối và đáp ứn tương đối. Ví dụ, khi sử dụng hàm sản xuất dạng tuyến tín thì đáp ứng là tuyệt đối vì khi đó nếu ta sử dụng tăng hoặ giảm một lượng yếu tố đầu vào $X_n$ nào đó thì đáp ứng của yế tố đầu ra sẽ tương ứng với tham số $b_n$ của biến số đó. Tron khi đó nếu áp dụng hàm Cobb-Douglas thì đáp ứng là tươn

đối vì khi yếu tố đầu vào $X_n$ nào đó thay đổi 1% thì kết quả đầu ra sẽ đáp ứng là $b_n$%.

Các giả định trong phân tích hàm sản xuất:

(1) Hàm sản xuất được xây dựng với các giá trị không âm của đầu vào và đầu ra $Y \geq 0$ và $X \geq 0$.

(2) Quan hệ đầu vào, đầu ra được đánh giá riêng biệt, liên tục mà với chúng sẽ có đạo hàm riêng bậc một và bậc hai của đầu ra theo yếu tố đầu vào tức là $\partial y/ \partial x_1$ và $\partial^2 y/ \partial x_1^2$ không triệt tiêu.

(3) Hàm sản xuất được đặc trưng bởi:

- Sản phẩm cận biên giảm dần với mọi kết hợp yếu tố - sản phẩm.

- Tỷ lệ thay thế cận biên giữa hai yếu tố bất kỳ giảm dần.

- Tỷ lệ chuyển đổi giữa hai sản phẩm bất kỳ giảm dần.

Trình tự phân tích hàm sản xuất:

(1) Xác định mô hình kinh tế: Mục đích của giai đoạn này là chỉ ra một mô hình kinh tế thích hợp nhằm thể hiện hàm sản xuất bằng quan hệ toán học. Đây là giai đoạn chuyển đổi từ các giả thuyết sang công cụ toán học. Các công việc chủ yếu trong giai đoạn này bao gồm: Chọn dạng mô hình; Chọn các biến đưa vào mô hình.

(2) Phân loại các yếu tố đầu vào, đầu ra: Liên quan đến các yếu tố đầu vào như đất đai, các loại đầu tư trên đất kể cả quản lý v.v...

(3) Thu thập số liệu: Có hai nguồn số liệu là số liệu tl nghiệm và số liệu điều tra.

(4) Các kỹ thuật kinh tế lượng trong ước lượng các tha số của hàm sản xuất. Bằng phương pháp bình phương nh nhất đã trình bày phần trên (OLS).

(5) Đánh giá và phân tích kết quả của hàm sản xuất từ đ rút ra các vấn đề đáng quan tâm. Mục đích của việc đánh g một hàm sản xuất là để rút ra các đại lượng có ý nghĩa kinh khác nhau như độ co giãn sản xuất, mức thay đổi tuyệt đ v.v... để từ đó có thể đưa ra những quyết định phù hợp ch việc tác động vào quá trình sản xuất nhằm mang lại lợi ích v hiệu quả cao nhất cho nhà sản xuất thông qua việc sử dụng đ đai.

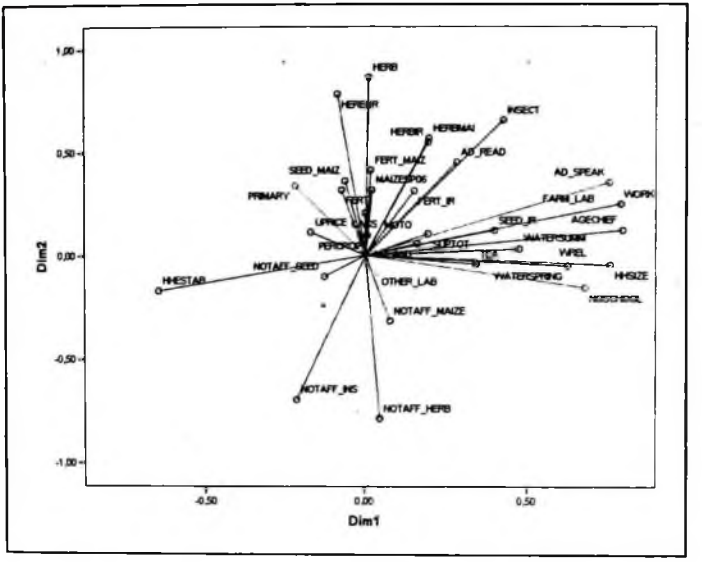
*4.4.3.4. Phân tích nhân tố*

Phân tích nhân tố là kỹ thuật cho việc giảm số liệu tức giải thích sự biến động trong việc thu thập của các sự biế động liên tục bằng một số lượng nhỏ các yếu tố (nhân tố). Vấ đề chính của phân tích nhân tố mang nặng tính chủ quan tron việc giải thích các kết quả.

Bản chất của việc phân tích nhân tố là việc khai tha thông tin trên cơ sở giảm bớt đi các hướng khác nhau của s liệu. Nhìn chung phương pháp này khai thác các mối quan l giữa một số biến ngẫu nhiên và bớt đi các biến ngẫu nhiê không có quan hệ, đây là cách chuyển đổi từ bộ số liệu gố sang bộ số liệu gần với nó.

Phương pháp này nhanh chóng xác định những yếu tố/nhân tố chính hoặc nhóm nhân tố chính có tính kiểm soát toàn bộ hệ thống mà ta đang nghiên cứu. Phương pháp này có thể trình bày kết quả ở dạng đồ thị, do đó ta có thể thấy rõ mối quan hệ giữa các quan sát với nhau.

Hình 4.5 dưới đây thể hiện một ví dụ trong sử dụng công cụ SPSS để phân tích nhân tố. Trong ví dụ này, chúng ta sẽ sử dụng kết quả của một điều tra về kinh tế hộ nông dân để phân tích các loại hình hộ nông dân khác nhau về phương thức kiếm sống.




Hình 4.5: ĐỒ THỊ PHÂN TÍCH CÁC NHÂN TỐ

Trên đồ thị chúng ta có thể thấy một số xu hướng/hướ phân lập rõ của các nhân tố. Trên cơ sở sự phân bố xu hướ của các nhân tố có cùng hướng ta sẽ nhóm lại thành nhữ nhóm riêng và sử dụng các công cụ phân tích nhóm để phâ tích. Như vậy, thay vì việc chúng ta phân tích đầy đủ các c tiêu/nhân tố thì chúng ta sẽ lựa chọn ra một số nhân tố có tí đại diện và có ảnh hưởng lớn đến xu hướng biến động và phâ chia các nhóm để phân tích.

## 4.5. Cách thức trình bày kết quả phân tích số liệ thống kê trong báo cáo khoa học

Kết quả phân tích số liệu thống kê có thể được trình bà dưới dạng bảng, biểu đồ, đồ thị. Bởi vì, chỉ có một số lượn nhỏ các mẫu được quan sát trong tổng thể nên kết quả củ việc phân tích mẫu cần phải phản ánh được sự không chắ chắn qua xác suất và sự biến động, hay nói cách khác trong việc thể hiện kết quả cần phải thể hiện được hai moment là g trị bình quân/trung vị/mode và sự biến động.

Đối với một báo cáo khoa học cách thức trình bày kết q phân tích phải hết sức linh hoạt giữa bảng biểu và đồ thị ( tránh sự nhàm chán và để thể hiện tốt nhất những ý tưởng n số liệu có thể thông tin cho ta biết.

### *4.5.1. Danh mục*

Dưới dạng danh mục ta có thể minh hoạ các kết quả phâ tích được của số liệu thống kê.

*Ví dụ:*

- Không có sự khác biệt giữa nam giới và nữ giới.
- Không có sự khác biệt giữa nhóm người nghèo và nhóm khác.
- Nhóm thanh niên hoàn toàn thoả mãn với dịch vụ.

Danh mục khá đơn giản cho việc tiến hành và rất hữu dụng trong việc báo cáo kết quả. Tuy nhiên, việc đơn giản đó đòi hỏi phải có những lời giải thích đi kèm.

Để giúp cho việc sử dụng danh mục được hiệu quả cần phải lưu ý:

(1) Chỉ sử dụng một vài từ để thể hiện ý chính, không nên sử dụng cả đoạn văn dài.

(2) Đảm bảo sự thống nhất trước sau, tránh việc khi thì sử dụng một vài từ khi thì sử dụng cả câu.

(3) Để khoảng trống giữa các mục của danh mục để dễ cho người đọc.

(4) Sử dụng cùng loại các dấu hay gạch đầu dòng cho các mục.

(5) Trong cùng một bản báo cáo nên sử dụng cùng một loại ký hiệu.

### *4.5.2. Biểu đồ*

Biểu đồ là phương pháp trình bày số liệu nhằm giúp người đọc dễ hiểu. Trong biểu đồ số liệu được trình bày dưới

dạng số. tuyệt đối hoặc số tương đối. Khi sử dụng biểu đ chúng ta có thể đưa ra những kết luận nhanh chóng thay nhìn vào số liệu.

Có nhiều dạng biểu đồ có thể sử dụng kết hợp trong m báo cáo khoa học của số liệu thống kê như: dạng bánh, dạ cột, hay dạng đồ thị, dạng mạng nhện v.v...

Khi sử dụng biểu đồ cần lưu ý một số điểm sau:

(1) Giải thích ngắn gọn các phần của biểu đồ.

(2) Đưa nguồn số liệu.

(3) Không nên sử dụng quá nhiều phần chia trong m biểu đồ.

(4) Làm nổi bật các phần mong muốn bằng các màu s khác nhau.

(5) Đối với dạng cột và đồ thị cần phải lưu ý về gốc toạ độ

(6) Sử dụng dạng cột để so sánh giữa các nhóm hoặc sánh sự biến động số liệu qua thời gian.

(7) Dạng đồ thị thường được dùng với số liệu qua th gian.

### *4.5.3. Dạng bảng*

Dạng bảng thường được dùng để thể hiện kết quả của m hoặc một vài chỉ tiêu nghiên cứu tương ứng với các tiêu c khác nhau (bảng chéo).

*Ví dụ:* Số liệu diện tích đất của các nhóm hộ khác nhau (nghèo, trung bình và khá).

Một số điều lưu ý khi sử dụng bảng để minh hoạ kết quả:

(1) Đầu mục của các cột cần phải xác định là các tiêu chí quan trọng cho việc so sánh.

(2) Không nên có quá nhiều cột hay dòng trong một bảng.

(3) Nên sử dụng các ký hiệu để minh hoạ cho mức độ tin cậy/ý nghĩa thống kê.

(4) Nên có nguồn trích dẫn của số liệu trong bảng.

(5) Trong cùng một bảng các số liệu cần phải có sự liên hệ với nhau.

### *4.5.4. Viết báo cáo kết quả phân tích*

Một báo cáo phân tích số liệu thống kê đầy đủ là báo cáo giải thích rõ các thông tin. Việc viết báo cáo như thế nào phụ thuộc vào việc báo cáo đó nộp cho ai? Ai là người sẽ đọc nó? Dưới đây là danh mục các phần trong báo cáo mà khi viết ta cần lưu ý:

(1) Tên của báo cáo, chúng tôi, nơi và ngày viết. Việc đưa ra một cái tên cho bản báo cáo đòi hỏi phải hết sức ngắn gọn, nhưng rõ ràng.

(2) Trong phần giới thiệu, cần phải chỉ ra được vấn đề đặt ra cần phải giải quyết và những câu hỏi phải trả lời thông qua nghiên cứu vấn đề, những giả thuyết cần kiểm tra.

(3) Danh mục các đặc trưng của cuộc điều tra: Kiểu điều tra; công cụ sử dụng trong điều tra; bao nhiêu câu hỏi; bao nhiêu mẫu/quan sát; các vấn đề liên quan khác.

(4) Giải thích phương pháp điều tra: Thiết kế điều tra; lự chọn mẫu; phân tích.

(5) Các kết quả liên quan đến các câu hỏi đặt ra, vấn đ nghiên cứu.

(6) Các kết luận.

(7) Các kiến nghị rút ra từ nghiên cứu.

Đối với phần phân tích các số liệu cần phải rõ ràng và cá đầu của đoạn văn đầu hết sức quan trọng vì nó sẽ thể hi toàn bộ nội dung chính của mục đó. Phần mở đầu cần ph viết ngắn gọn, sử dụng những từ ngữ thông dụng để bất kỳ đọc cũng có thể hiểu được nội dung mà ta đang mong muố làm và đạt được. Phần nội dung chính cần lưu ý sử dụng c từ ngữ học thuật để đảm bảo cho những người có chuyên mô trong lĩnh vực mà ta đang nghiên cứu thấy được trình độ c người viết.

Tóm lại: Việc phân tích số liệu là để hiểu được quá kh hiện tại nhằm mục đích phục vụ cho tương lai do vậy một l giải thích, phân tích rõ ràng sẽ giúp cho người đọc thấy đư điều gì sẽ có thể diễn ra trong tương lai. Một nhà kinh doa giỏi, một nhà quản lý tốt là người có thể hiểu được các thô tin và sử dụng chúng một cách có hiệu quả.

### 4.6. Một số phương pháp thường dùng kết hợp trơ phân tích số liệu thống kê

Nhiều phương pháp khác nhau đã được áp dụng và có t áp dụng được trong nghiên cứu và phân tích trong các nghi

cứu về kinh tế - xã hội, theo các tiêu thức khác nhau chúng ta có thể chia các phương pháp phân tích số liệu thành: (1) Nhóm phương pháp định lượng; (2) Nhóm các phương pháp định tính hoặc theo phạm vi ứng dụng có thể chia ra làm 2 nhóm: 1) Nhóm 1: Phương pháp nghiên cứu vĩ mô như GIS, ảnh viễn thám; (2) Nhóm 2: Các phương pháp định lượng và áp dụng các mô hình toán học như mô hình kinh tế lượng, mô hình bài toán tối ưu. Ngoài ra, mô hình đa nhân tố được ứng dụng một cách rộng rãi trong rất nhiều lĩnh vực nghiên cứu trong đó đặc biệt đối với các vấn đề thuộc về kinh tế - xã hội khi đối tượng của chúng ta nhiều và có tính phức tạp v.v...

Trong mục này, chúng ta sẽ lần lượt làm quen với một số phương pháp cũng như khả năng ứng dụng của chúng.

### *4.6.1. Phương pháp quy hoạch tuyến tính*

Bài toán quy hoạch tuyến tính - bài toán tối ưu sẽ giúp chúng ta cân đối tối ưu hoá việc sử dụng các yếu tố hay nguồn lực giữa các mục đích khác nhau theo một mục tiêu chung nào đó, đồng thời thoả mãn theo một số yêu cầu của chúng ta đặt ra hoặc buộc phải tuân theo. Lý thuyết trò chơi là một phương pháp cũng được áp dụng trong nhiều lĩnh vực, trong đó có cả việc lựa chọn quyết định trong điều kiện hoàn toàn bất định hay chúng ta có thể quy về bài toán quy hoạch tuyến tính.

Trong thực tế đời sống sản xuất có rất nhiều vấn đề đặt ra đòi hỏi chúng ta phải giải quyết mang tính hệ thống. Điều đó có nghĩa là ta phải nhìn nhận vấn đề đó một cách toàn diện

trong sự gắn kết và mối quan hệ giữa các nhân tố bên trc cũng như giữa các nhân tố bên trong và bên ngoài của sự v hiện tượng đó. Để đáp ứng yêu cầu đó, phương pháp c hoạch tuyến tính sẽ giúp ta một công cụ hữu ích cho các v đề mang tính hệ thống như vậy. Vì vậy trong phần này chú ta sẽ nghiên cứu bài toán quy hoạch tuyến tính áp dụng trc thực tế đời sống kinh tế.

Tuỳ thuộc vào từng vấn đề cụ thể, mô hình sẽ có nhữ đặc điểm khác nhau chi tiết. Nhưng có thể mô tả nội dung bản của nó như sau:

Một đơn vị sản xuất có thể bố trí nhiều loại hình sản x khác nhau. Mỗi loại hình đòi hỏi nhu cầu các yếu tố đầu v khác nhau. Chúng ta có một số yếu tố đầu vào của sản x nào đó với trữ lượng cho trước, sử dụng kết hợp những yếu này để cùng sản xuất nhiều loại sản phẩm đầu ra khác nl với định mức tiêu hao nguyên vật liệu khác nhau.

Ta phải lập phương án sử dụng các tài nguyên đó sao tối ưu hoá một chỉ tiêu nào đó chẳng hạn như lãi, doanh hoặc chi phí v.v...

Để xây dựng bài toán, chúng ta giả thiết có m loại yếu đầu vào với trữ lượng lần lượt là $b_1, b_2, \ldots, b_m$ cùng tham vào các quá trình công nghệ khác nhau để tạo ra n loại phẩm khác nhau. Hệ số hao phí yếu tố đầu vào thứ i ($i = \overline{1}$ tính cho 1 đơn vị sản phẩm thứ j là $a_{ij}$ ($j = \overline{1,n}$). Giá mỗi vị sản phẩm thứ j là $C_j$ ($j = \overline{1,n}$), chúng ta phải lập phương sản xuất sao cho tổng doanh thu là lớn nhất.

Rõ ràng phương án sản xuất được xác định bởi số lượng sản phẩm mỗi loại cần sản xuất ra nên chúng ta đặt biến số là $x_j$ ($j = \overline{1,n}$) biểu diễn số lượng sản phẩm j được sản xuất. Hiển nhiên $x_j \geq 0$ ($j = \overline{1,n}$). Phương án sản xuất chỉ có thể chấp nhận được khi tổng số yếu tố đầu vào không vượt quá số lượng hiện có. Nghĩa là phương án phải thoả mãn các ràng buộc về yếu tố sản xuất:

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n = \sum_{j=1}^{n} a_{ij}x_j \leq b_1 (i = 1, m)$$

Gọi F(x) là tổng doanh thu ứng với phương án X = ($x_1$, $x_2$,..., $x_n$) thì F(x) được tính bằng biểu thức:

$$F(x) = C_1x_1 + C_2x_2 + \ldots + C_nx_n = \sum_{j=1}^{n} C_jx_j$$

Theo yêu cầu thì F(x) phải được tối đa hoá, do vậy chúng ta sẽ có bài toán như sau:

$$F(x) = \sum_{j=1}^{n} C_jx_j \Rightarrow \max$$

$$\sum_{j=1}^{n} a_{ij}x_j = (\leq;\geq) b_i (i = \overline{1,m})$$

$$x_j \geq 0 (j = \overline{1,n})$$

Việc giải bài toán này sẽ cho chúng ta một phương án sử dụng các nguồn lực đảm bảo tối đa yếu tố doanh thu nhưng

đồng thời lại đảm bảo thoả mãn sự hạn chế của các yếu nguồn lực khác.

Ngoài vấn đề kinh tế đang quan tâm, chúng ta còn có t đưa thêm vào đó các yếu tố đòi hỏi về môi trường hay mức thiểu phải đạt được về các lợi ích xã hội.

### *4.6.2. Mô hình đa nhân tố (MAS) và ứng dụng tro nghiên cứu các vấn đề kinh tế - xã hội*

Mô hình đa nhân tố (Multi-Agent-Systems) giúp ta có kết hợp nhiều nhân tố khác nhau trong mô hình và dưới nh cấp độ khác nhau, hay nói một cách khác đó là nghiên cứu thống trong mối quan hệ rộng hơn với các hệ thống bao tr nó.

*Ví dụ:* Phát triển kinh tế bền vững đòi hỏi phải xem xét tỉ m dưới 3 góc độ: Kinh tế (thu nhập, vốn, tích luỹ v.v...); xã hộ (lao động có việc làm, tệ nạn v.v...) và môi trường (xói mòi đất, tỷ lệ che phủ, bỏ hoang đất v.v...).

*Mục đích của mô hình đa nhân tố là cố gắng tìm hiểu vận hành của các quá trình độc lập.* Một nhân tố được c vào mô hình như là một chương trình máy tính và được m tả như một quá trình tự động hoá bởi vì nó được trang bị l năng đáp ứng với sự thay đổi của môi trường. Một mô hình nhân tố được tạo thành bởi các nhân tố trong cùng một t gian. Chẳng hạn, các nhân tố xuất hiện trong cùng một t gian và sử dụng chung một số nguồn lực đồng thời lại những quan hệ qua lại với nhau. Vấn đề chính trong mô h

đa nhân tố là việc tổ chức phối hợp giữa các nhân tố với nhau. Vì thế nghiên cứu đa nhân tố tức là nghiên cứu:

*Quá trình ra quyết định*: Cơ chế ra quyết định như thế nào của các nhân tố? Mối quan hệ giữa quan điểm, sự thể hiện và hành động của chúng như thế nào?

*Kiểm tra*: Mối quan hệ thứ bậc tồn tại giữa các nhân tố? Sự đồng nhất giữa chúng?

*Liên lạc*: Những thông tin mà các nhân tố cung cấp cho nhau? Những cú pháp mà chúng ra lệnh?

Mô hình đa nhân tố có thể được áp dụng như một sự thông minh nhân tạo. Nó có thể giải quyết được vấn đề bằng cách chia vấn đề ra làm nhiều phần nhỏ như là một nhân tố có liên quan đến nhau và được tổ chức đồng thời. Quá trình này đặc biệt được dùng để điều kiển một quá trình công nghiệp.

Mô hình đa nhân tố được ứng dụng rộng rãi trong nhiều lĩnh vực trong đó có kinh tế - xã hội.

*MAS và khoa học xã hội*

MAS được phát triển nhanh chóng trong các khoa học xã hội. Mô phỏng xã hội là một mục tiêu của khoa học số. Trong các khoa học xã hội việc ứng dụng mô hình đa nhân tố để mô phỏng các sự kiện có liên quan mật thiết đến xu hướng vận động của xã hội được gọi là "cá nhân", trong đó mỗi một cá nhân được xem như là một thành tố của xã hội. Sự trùng lặp, hoặc tiếp cận từ dưới lên có đặc trưng như một mô hình đa nhân tố. Tuy nhiên, sự đồng hoá các cá thể trong xã hội, trong

một mô hình đa nhân tố có thể bị sai lệch. Thực tế cho thấy mô hình đa nhân tố rất phù hợp cho các tổ chức xã hội được xem như là một nhân tố với những tiêu chuẩn và vai trò của riêng họ. Một nhân tố bị hạn chế bởi vai trò của nó thể hiện trong nhóm, chẳng hạn số lượng của các vị trí trong môi trường động.

Một số quan điểm cho việc xây dựng mô hình đa nhân tố:

(1) Mỗi một cá thể tạo ra lịch sử, vận động bằng cách thu lượm các giá trị và vai trò.

(2) Các giá trị và vai trò tích luỹ được tiến triển bởi sự tác động qua lại giữa các cá thể và các nhóm với nhau.

(3) Các cá thể hoặc là giống nhau hoặc là tương đương nhau nhưng chúng có vai trò và tình trạng xã hội riêng.

*MAS và sự tác động qua lại giữa các tổ chức xã hội và nguồn lực*

Vấn đề đặt ra trong mô hình hoá việc quản lý nguồn lực đó là làm thế nào để mô phỏng sự tương tác giữa các nhóm nhân tố và sự thay đổi nguồn lực. Có nhiều cách tiếp cận đến vấn đề này. Cách thứ nhất đó là mô hình hoá quản lý hệ thống xã hội. Ở đây, chúng ta xem xét mối quan hệ giữa con người và nguồn lực, trong thực tế chúng ta có thể coi con người tác động đến các nguồn lực. Các nhân tố sẽ trao đổi thông tin trong một mạng hoặc có thể coi đó là mạng liên hệ và điều này có thể liên hệ đến mô hình đa nhân tố. Trong trường hợp

này, nó có thể chuyển đổi thông tin và dịch vụ cũng như hợp đồng và các thoả thuận giữa các nhân tố.

*Ví dụ:* Trường hợp của một hệ thống tưới tiêu, các nông dân có thể truyền tải các thông điệp đến những người khác, do vậy họ có thể biết được mức nước hiện nay ở các điểm khác nhau.

Bản thân các nhân tố cũng được hỏi để trao đổi các dịch vụ và địa chỉ. Trong trường hợp này nó thể hiện sự phát triển của hệ thống có thể tác động đến cấu trúc và sự chuyển động của mạng xã hội.

Cách tiếp cận thứ hai, dựa trên hiểu biết hoặc sự thể hiện, qua đó xác định nhân tố và nguồn lực tác động qua lại. Mỗi một nhân tố phát triển và sau đó có các hoạt động trên cơ sở sự trình bày cá nhân chúng với các nguồn lực. Trong các hoạt động đó các nhân tố sẽ chuyển giao các nguồn lực cho các nhân tố khác. Mô hình này trình bày sự tương tác như điều mà trong kinh tế gọi là quan hệ bên ngoài. Chúng ta quan tâm đến quản lý các nguồn tài nguyên có thể tái tạo thông qua việc kiểm tra quá trình thể hiện của các hoạt động có thể tác động đến các nguồn lực. Kết quả các nguồn lực có thể thoả mãn hoặc không thể thoả mãn cho các nhân tố, điều này có thể minh hoạ như sự sắp xếp qua môi trường.

Phương pháp này cung cấp cơ sở cho việc áp dụng mô hình đa nhân tố vào xử lý các vấn đề liên quan giữa các vấn đề xã hội và nguồn lực.

### *4.6.3. Ứng dụng GIS trong phân tích, nghiên cứu kinh tế - xã hội*

*4.6.3.1. GIS là gì?*

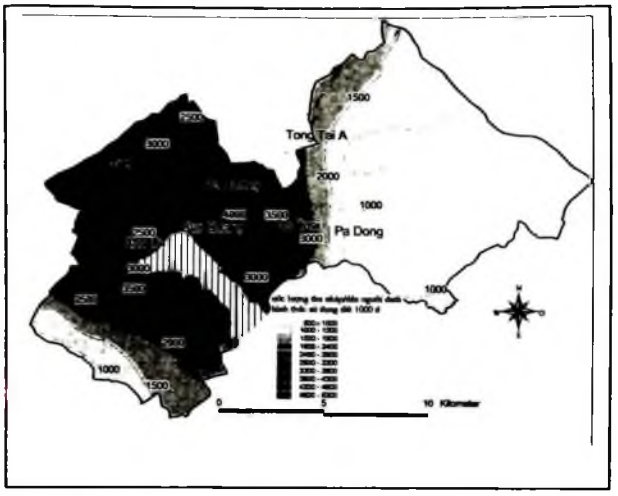
GIS (Geographic Information System) là hệ thống thông tin địa lý, nó tạo ra sự khác biệt trong quản lý thông tin so với các dạng quản lý thông tin khác như hệ cơ sở dữ liệu, bảng tính ở chỗ công cụ này xử lý thông tin khoảng cách.

GIS có khả năng tạo ra những chồng ghép của các thông tin ở cùng một vị trí, kết hợp chúng, phân tích chúng và cuối cùng là vẽ ra các bản đồ. Do vậy, trong phân tích kinh tế GIS sẽ là một công cụ hữu ích cho việc nghiên cứu ở tầm vĩ mô.

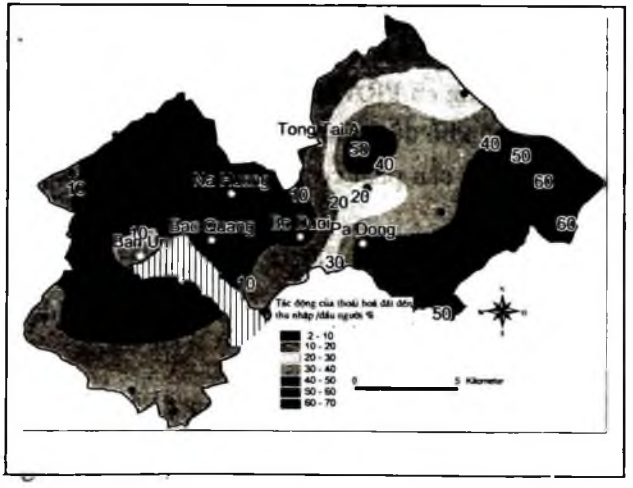
*4.6.3.2. Khả năng ứng dụng của GIS*

Ngày nay, GIS không chỉ ứng dụng để xây dựng các bản đồ về hiện trạng sử dụng đất, chất đất, v.v... mà nó còn được ứng dụng trong các nghiên cứu kinh tế - xã hội. Thông qua việc kết hợp các thông tin kinh tế cho phép xây dựng lên các bản đồ sử dụng đất hợp lý mang lại nguồn lợi kinh tế lớn hơn cho các khu vực, hoặc bản đồ quan hệ giữa sử dụng đất và phát triển kinh tế cho một khu vực nào đó.

*Ví dụ:* Bằng việc sử dụng công cụ GIS và thu thập số liệu theo phương pháp phỏng vấn người có hiểu biết của khu vực nghiên cứu xây dựng bản đồ ước lượng thu nhập theo hình thức sử dụng đất hiện tại hay tác động của việc thoái hoá đất đến thu nhập/đầu người ở Mai Sơn, Sơn La.




Nguồn: Lentes, 2004



Đây chỉ là một vài ví dụ minh hoạ cho việc ứng dụng công nghệ GIS trong nghiên cứu kinh tế đất, để có thể vận dụng thành công phương pháp này đòi hỏi người học phải nắm vững kiến thức về kinh tế và kiến thức về GIS.

*4.6.3.3. Trình tự các bước khi sử dụng công cụ GIS trong nghiên cứu.*

Khi ứng dụng công nghệ GIS trong nghiên cứu các vấn đề về kinh tế - xã hội thông thường chúng ta cần phải thực hiện theo các nội dung sau:

(1) Xây dựng bản đồ nền số hoá khu vực nghiên cứu.

(2) Thu thập các thông tin kinh tế, xã hội và thông tin liên quan đến vấn đề cần nghiên cứu thông qua việc sử dụng các phương pháp thu thập thông tin khác nhau như: điều tra phỏng vấn những người am hiểu trong toàn bộ khu vực (được phân bố đều), sử dụng công cụ PRA hay sử dụng bảng câu hỏi điều tra các hộ được phân bố đều trong toàn bộ khu vực. Tất cả các thông tin thu thập đó cần phải có vị trí địa lý cụ thể để số hoá đưa vào bản đồ.

(3) Số hoá các thông tin kinh tế - xã hội đã thu thập được thông qua việc sử dụng các phần mềm xử lý bản đồ như Arcview hay Mapinfor v.v...

(4) Chồng ghép các bản đồ khác nhau để làm rõ vấn đề đang quan tâm.

(5) Phân tích và đưa ra những kết luận.

Trong quá trình tiến hành các nội dung trên không nhất thiết phải tuân thủ theo một trật tự cố định như vậy, ví dụ chúng ta có thể đồng thời cùng tiến hành các nội dung 1 và 2 nhưng các nội dung còn lại không thể thay đổi trật tự tiến hành được.

Đồng thời việc sử dụng công cụ GIS đòi hỏi người học phải kết hợp với nhiều phương pháp khác như thu thập số liệu, kinh tế lượng v.v... hay nói một cách khác là việc vận dụng uyển chuyển và nhuần nhuyễn các phương pháp như trên đã trình bày một cách đồng thời sẽ giúp ích trong việc phân tích và nghiên cứu.

# BÀI TẬP THỰC HÀNH

Phần này giúp cho sinh viên thực hành các kỹ năng lý thuyết đã được cung cấp trong phần trước của cuốn sách. Với yêu cầu sinh viên phải thực hành các bài tập đưa ra và sử dụng phần mềm xử lý thống kê SPSS.

Phần này được thiết kế thành 4 bài tập thực hành lớn bao hàm toàn bộ các thông tin đã được học.

**Bài 1: Chuẩn bị cho việc thu thập thông tin thống kê**

Hãy đưa ra một vấn đề nghiên cứu và qua đó cho biết nên vận dụng phương pháp chọn mẫu nào phù hợp cho nghiên cứu của mình.

Yêu cầu của bài thực hành này là sinh viên sẽ biết cách đưa ra vấn đề nghiên cứu mà mình quan tâm cũng như phù hợp với thực tiễn yêu cầu; bên cạnh đó sinh viên cần nắm rõ và có khả năng vận dụng các phương pháp chọn mẫu phù hợp với mục đích nghiên cứu.

Cách thức tiến hành: Sinh viên làm việc theo nhóm, sau đó sẽ trình bày ý tưởng của mình cho các thành viên khác nghe và đóng góp ý kiến.

Công cụ: Bảng viết phấn, bảng giấy lật, máy tính và máy chiếu.

**Bài 2: Thu thập thông tin và xây dựng cơ sở dữ liệu**

Hãy xây dựng một mẫu phiếu điều tra cho vấn đề mà anh/chị quan tâm, trên cơ sở mẫu phiếu điều tra đó hãy chuẩn bị một cơ sở dữ liệu cho việc quản lý thông tin sau khi thu thập trên Excel.

Yêu cầu của bài thực hành này là sinh viên phải xây dựng được một mẫu phiếu điều tra phù hợp với mục đích điều tra.

Phương pháp tiến hành: Sinh viên làm việc độc lập, sau đó sẽ thảo luận và trao đổi để có một phiếu điều tra hoàn chỉnh.

Công cụ: Máy tính máy chiếu, bảng viết phấn, bảng giấy lật.

**Bài 3: Xử lý thông tin**

Một cơ sở dữ liệu cho trước, anh chị hãy sử dụng phần mềm SPSS để xử lý các số liệu thống kê trên. Kết xuất thông tin đó thành các bảng biểu kết quả phân tích.

Yêu cầu của bài thực hành là giúp sinh viên nắm bắt kỹ năng xây dựng cơ sở dữ liệu; sử dụng tốt phần mềm xử lý thống kê, cụ thể là phần mềm SPSS.

Cách thức tiến hành: Sinh viên sẽ được thực hành trên máy tính với phần mềm SPSS.

Công cụ: Máy tính với phần mềm SPSS.

**Bài 4: Trình bày kết quả phân tích số liệu thống kê**

Trên cơ sở các xử lý trên hãy trình bày kết quả phân tích dưới dạng một báo cáo khoa học.

Yêu cầu của bài thực hành thứ 4 là sinh viên cần phải nắm được cách thức trình bày một báo cáo khoa học từ những kết quả phân tích số liệu thống kê.

Cách thức tiến hành: Sinh viên làm việc theo từng nhóm 3-5 sinh viên, các nhóm sẽ trình bày kết quả của nhóm để các nhóm khác góp ý.

Công cụ: Máy tính và máy chiếu, bảng, giấy lật.

# TÀI LIỆU THAM KHẢO

Arlene Fink, 1995: The survey Kit - SAGE Puliscations, London.

D.A. Lind et. al., 2003: Basic Statistics for Business and Economics - McGraw-Hill Higher Education.

DeAngelis, D. L. and L. J. e. Gross, 1992: Individual-based models and approaches in ecology, Chapman and Hall.

Gilbert, N., 1995: Emergence in social simulation. Artificial societies. The computer simulation of social life. R. c. a. N. Gilbert, UCL Press: 144-156.

Nguyễn Quang Dong, 2002: Kinh tế lượng - Chương trình nâng cao - Nhà xuất bản Khoa học và Kỹ thuật.

Nguyễn Quang Dong, 2003: Bài giảng kinh tế lượng - Nhà xuất bản Thống kê.

Paul Newbold et. al., 2003: Statistics for business and economics - Prentice Hall.

Peter Lentis, 2003: The contribution of GIS and remote sensing to farming systems research on Micro and regional scale in Northern Vietnam Margraf Verlag - Kanalstr. 21 P.O Box 1205.

P. J. Riedel, 2003: Research and Reflect - a course in scientific writing handout - Hohenheim.

Weiss, G., Ed. 1999. Multiagent Systems: a Modern Approach to Distributed Artificial Intelligence, MIT Press.

# MỤC LỤC


Lời nói đầu 3

Giới thiệu chung 5

**Chương I: CHUẨN BỊ SỐ LIỆU** 13

**1.1. Thiết kế điều tra nghệ thuật và khoa học** 14

**1.2. Chọn mẫu** 25

1.2.1. Chọn mẫu thống kê trong điều tra chọn mẫu 25

1.2.2. Chọn mẫu phi thống kê trong điều tra chọn mẫu 42

**1.3. Quy mô mẫu trong điều tra chọn mẫu** 48

1.3.1. Phân phối mẫu 48

1.3.2. Sai số chọn mẫu và phi chọn mẫu 52

1.3.3. Cỡ mẫu 60

1.3.4. Tính toán trọng số 65

**1.4. Phương pháp thu thập số liệu** 66

1.4.1. Phiếu điều tra 68

1.4.2. Các phương pháp phỏng vấn thu thập thông tin 70

**Chương II: CƠ SỞ DỮ LIỆU** 73

**2.1. Các dạng cơ sở dữ liệu** 77

**2.2. Biểu diễn thông tin thống kê trong cơ sở dữ liệu** 78

2.2.1. Dữ liệu dạng định tính 79

2.2.2. Dữ liệu dạng định lượng 81

2.2.3. Các chỉ tiêu nghiên cứu 81

**2.3. Mã hoá các thông tin trong cơ sở dữ liệu** 82
2.3.1. Mã hoá các thông tin định tính 83
2.3.2. Mã hoá các số liệu bị thiếu và vượt trội 84
**2.4. Xác định và xử lý các giá trị bị thiếu và vượt trội trong cơ sở dữ liệu** 86

**Chương III: PHÂN TỔ VÀ KIỂM ĐỊNH THỐNG KÊ** 95
**3.1. Lý do của việc phân tổ** 95
**3.2. Phân tổ thống kê theo các chỉ tiêu định tính** 96
**3.3. Phân tổ thống kê theo các chỉ tiêu định lượng** 97
**3.4. Kiểm định thống kê** 101
3.4.1. Đối với các chỉ tiêu định tính 102
3.4.2. Đối với các chỉ tiêu định lượng 102
3.4.3. Ý nghĩa và sự giải thích của giá trị xác suất P (P-values) (số liệu nói lên điều gì?) 105

**Chương IV: PHÂN TÍCH SỐ LIỆU ĐIỀU TRA VÀ BIỂU DIỄN KẾT QUẢ** 107
**4.1. Mô tả thống kê** 110
**4.2. Phân phối mẫu: lệch và đối xứng** 112
**4.3. Đo lường sự biến động** 114
**4.4. Tương quan và mối quan hệ** 115
4.4.1. Số liệu dạng định lượng 115
4.4.2. Dữ liệu dạng định tính 117
4.4.3. Phương pháp hồi quy và phân tích nhân tố 118

**4.5. Cách thức trình bày kết quả phân tích số liệu thống kê trong báo cáo khoa học** 128

4.5.1. Danh mục 128

4.5.2. Biểu đồ 129

4.5.3. Dạng bảng 130

4.5.4. Viết báo cáo kết quả phân tích 131

**4.6. Một số phương pháp thường dùng kết hợp trong phân tích số liệu thống kê** 132

4.6.1. Phương pháp quy hoạch tuyến tính 133

4.6.2. Mô hình đa nhân tố (MAS) và ứng dụng trong nghiên cứu các vấn đề kinh tế - xã hội 136

4.6.3. Ứng dụng GIS trong phân tích, nghiên cứu kinh tế - xã hội 140

**Bài tập thực hành** 144

**Tài liệu tham khảo** 147

# GIÁO TRÌNH
# PHÂN TÍCH SỐ LIỆU THỐNG KÊ

**Chịu trách nhiệm xuất bản:**
TS. TRẦN HỮU THỰC

**Biên tập:**
ĐỖ VĂN CHIẾN

**Trình bày, bìa:**
TRẦN KIÊN - THÙY DƯƠNG

**Sửa bản in:**
PHÒNG XUẤT BẢN - CÔNG TY CP KIẾN THỨC VÀNG

In 100 cuốn khổ 14,5cm x 20,5cm tại Công ty Cổ phần Kiến Thức Vàng
Giấy phép xuất bản số 85-2008/CXB/324-134/TK
Do Cục Xuất bản cấp ngày 17/01/2008
In xong và nộp lưu chiểu tháng 8 năm 2008.