

CHƯƠNG 1: KHÁI QUÁT VỀ QUI TRÌNH NGHIÊN CỨU – XỬ LÝ THÔNG TIN VÀ MỘT SỐ LÝ THUYẾT THỐNG KÊ CƠ BẢN SỬ DỤNG TRONG PHÂN TÍCH THÔNG TIN

1. Qui trình của một cuộc nghiên cứu

Thông thường một qui trình nghiên cứu bao gồm 8 bước:

- **Bước 1:** Xác định vấn đề cần nghiên cứu
- **Bước 2:** Xác định loại thông tin cần thu thập và nguồn cung cấp thông tin
- **Bước 3:** Chọn mẫu nghiên cứu
- **Bước 4:** Thiết kế nghiên cứu và xác định phương pháp thu thập thông tin.
- **Bước 5:** Thiết kế bảng câu hỏi
- **Bước 6:** Thu thập dữ liệu
- **Bước 7:** Xử lý, phân tích và diễn giải các dữ liệu đã được xử lý
- **Bước 8:** Trình bày và báo cáo kết quả

2. Xử lý thông tin trong nghiên cứu thực địa

Có hai dạng thông tin nghiên cứu cần thu thập, loại thứ nhất là thông tin thứ cấp và loại thứ hai là thông tin sơ cấp.

- Thông tin thứ cấp là những thông tin đã hiện hữu trên các nguồn tài liệu đã được đăng tải, thông tin này đã được tổ chức thành bảng biểu, đồ thị. Loại thông tin này người nghiên cứu chỉ việc sử dụng và diễn giải theo nhu cầu nghiên cứu của mình mà không cần phải trải qua một quá trình xử lý phức tạp đòi hỏi sự hỗ trợ của các phần mềm phân tích và xử lý thông tin chuyên dụng.
- Thông tin sơ cấp là thông tin chưa hiện hữu, muốn có thông tin này đòi hỏi các nhà nghiên cứu phải thực hiện một qui trình nghiên cứu với nhiều bước đã trình bày ở trên. Trong nghiên cứu thu thập thông tin sơ cấp tồn tại hai dạng nghiên cứu chính yếu nghiên cứu định tính và nghiên cứu định lượng. Thông tin trong nghiên cứu định tính không có ý nghĩa về mặt thống kê, quá trình phân tích và xử lý chỉ dừng ở chỗ tập hợp, phân nhóm những ý kiến quan điểm khác biệt và không đòi hỏi nhiều sự hỗ trợ của các công cụ và kiến thức thống kê. Ngược lại với thông tin nghiên cứu định lượng lại đòi hỏi nhiều kỹ năng và kiến thức phân tích thống kê để tổ chức và phân tích. Phần mềm SPSS là một công cụ hữu hiệu cho việc xử lý và phân tích những thông tin nghiên cứu định lượng này.

Trong nghiên cứu định lượng, dữ liệu ban đầu được thu thập từ hiện trường là dữ liệu thô, chúng ta chưa thể tiến hành phân tích và diễn giải những dữ liệu dạng thô này ngay được mà đòi hỏi phải tiến hành các bước xử lý và phân tích cần thiết từ mã hóa,

kiểm tra, hiệu đính, nhập liệu đến tạo bảng biểu cho dữ liệu và thực hiện các phân tích thống kê tương thích.

Nhiệm vụ tổng quát của việc xử lý – phân tích dữ liệu là chuyển những mẫu dữ liệu quan sát thô mà ta đã tiến hành mã hóa và kiểm tra thành những con số thống kê có ý nghĩa cho việc diễn giải kết quả nghiên cứu. Toàn bộ công việc xử lý – phân tích phức tạp này đòi hỏi cần phải có máy tính và các phần mềm chuyên dụng hỗ trợ.

3. Qui trình xử lý số liệu

Trong một qui trình nghiên cứu định lượng. Việc xử lý dữ liệu bắt đầu từ khi ta nhận được bảng câu hỏi đã được phỏng vấn. Qui trình xử lý số liệu bao gồm các bước sau:

- **Bước 1:** Kiểm tra, hiệu chỉnh các trả lời trên bảng câu hỏi
- **Bước 2:** Mã hóa các câu trả lời trên bảng câu hỏi
- **Bước 3:** Nhập dữ liệu đã được mã hóa vào máy tính
- **Bước 4:** Xác định các lỗi trong cơ sở dữ liệu và làm sạch dữ liệu
- **Bước 5:** Tạo bảng cho dữ liệu và tiến hành các phân tích thống kê

Hai giai đoạn đầu tiên là những bước chuẩn bị cho việc phân tích bằng máy tính sau này. Giai đoạn 3 là nhập các dữ liệu đã được mã hóa vào máy tính. Quá trình nhập liệu này có thể dẫn đến những sai sót do đó một bước kế tiếp phải được thực hiện trước khi tiến hành phân tích dữ liệu là phải làm sạch dữ liệu đã được nhập vào trong máy.

4. Một số lý thuyết thống kê cơ bản

4.1. Các tham số thống kê đo lường độ tập trung hay hội tụ của dữ liệu (central tendency measurement)

- **Giá trị trung bình (Mean):** Là giá trị trung bình số học của một biến, được tính bằng tổng các giá trị quan sát chia cho số quan sát. Đây là dạng công cụ thường được dùng cho dạng đo khoảng cách và tỷ lệ. Giá trị trung bình có đặc điểm là chịu sự tác động của các giá trị ở mỗi quan sát, do đó đây là thang đo nhạy cảm nhất đối với sự thay đổi của các giá trị quan sát. Giá trị trung bình được tính bằng công thức sau:

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

- **Trung vị (Median):** Là số nằm giữa (nếu lượng quan sát là số lẻ) hoặc là giá trị trung bình của hai quan sát nằm giữa (nếu số lượng quan sát là số chẵn) của một dãy quan sát được sắp xếp theo thứ tự từ nhỏ đến lớn. Đây là dạng công cụ thống kê thường được dùng để đo lường mức độ tập trung của dạng dữ liệu thang đo thứ

tự, nó có đặc điểm là không bị ảnh hưởng của các giá trị đầu mút của dãy phân phối, do đó rất thích hợp để phân tích đối với dữ liệu có sự chênh lệch lớn về giá trị ở hai đầu mút của dãy phân phối.

- **Mode:** Là giá trị có tần suất xuất hiện lớn nhất của một tập hợp các số đo, dạng này thường được dùng đối với dạng dữ liệu thang biểu danh. Giống như trung vị, mode không bị ảnh hưởng bởi giá trị đầu mút của dãy phân phối.

4.2. Các tham số thống kê đo lường mức độ phân tán của dữ liệu (Dispersion),

Khảo sát hai nhóm các con số sau::

Nhóm 1: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Nhóm 2: 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

Ta thấy số kích thước mẫu của hai nhóm này bằng nhau, các giá trị đo lường mức độ tập trung của dữ liệu như mean, media, mode đều bằng nhau và bằng 6. Tuy nhiên hai dữ liệu này hoàn toàn khác nhau. Nhóm 1 các dữ liệu biến đổi nhiều hơn nhóm 2, điều này có nghĩa các giá trị trong nhóm 1 phân tán hơn, các giá trị quan sát nằm xa giá trị trung bình của mẫu hơn là nhóm 2. Đo lường độ phân tán cho biết được những khác biệt giữa hai nhóm dữ liệu. Có một số công cụ đo lường độ phân tán của dữ liệu như:

- **Phương sai (Variance):** Dùng để đo lường mức độ phân tán của một tập các giá trị quan sát xung quanh giá trị trung bình của tập quan sát đó. Phương sai bằng trung bình các bình phương sai lệch giữa các giá trị quan sát đối với giá trị trung bình của các quan sát đó. Người ta dùng phương sai để đo lường tính đại diện của giá trị trung bình tương ứng, các tham số trung bình có phương sai tương ứng càng lớn thì giá trị thông tin hay tính đại diện của giá trị trung bình đó càng nhỏ. Phương sai của mẫu được tính bằng công thức sau:

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

- **Độ lệch chuẩn (Standard deviation):** Một công cụ khác dùng để đo lường độ phân tán của dữ liệu xung quanh giá trị trung bình của nó. Độ lệch chuẩn chính bằng căn bậc hai của phương sai. Vì phương sai là trung bình của các bình phương sai lệch của các giá trị quan sát từ giá trị trung bình, việc khảo sát phương sai thường cho các giá trị rất lớn, do đó sử dụng phương sai sẽ gặp khó khăn trong việc diễn giải kết quả. Sử dụng độ lệch chuẩn sẽ giúp dễ dàng cho việc diễn giải do các kết quả sai biệt đưa ra sát với dữ liệu gốc hơn.

- **Khoảng biến thiên (Range):** Là khoảng cách giữa giá trị quan sát nhỏ nhất đến giá trị quan sát lớn nhất.
- **Sai số trung bình mẫu (Standard Error of Mean)** Được dùng để đo lường sự khác biệt về giá trị trung bình của mẫu nghiên cứu này so với mẫu nghiên cứu khác trong điều kiện có cùng phân phối. Nó có thể được dùng để so sánh giá trị trung bình quan sát với một giá trị ban đầu nào đó (giả thuyết). Và ta có thể kết luận hai giá trị này là khác nhau nếu tỷ số về sự khác biệt đối với standard error of mean nằm ngoài khoảng $(-2,+2)$. Công thức tính sai số trung bình mẫu:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

4.3. Khoảng ước lượng (Confident interval)

Là một ước lượng xác định khoảng giá trị đặc trưng của tổng thể có thể rơi vào. Dựa vào dữ liệu mẫu, với một độ tin cậy cho trước ta có thể xác định được giá trị đại diện cho đám đông có thể nằm trong một khoảng ước lượng nào đó.

Ví dụ gọi x là mức thu nhập trung bình của đám đông cần ước lượng. Với độ tin cậy của khoảng sát nghiên cứu là 95% (nghĩa là các ước lượng sẽ luôn có một lượng sai số chấp nhận là 5%). Dựa vào mẫu quan sát ta có thể xác định được hai giá trị về thu nhập là a và b sao cho xác suất để thu nhập trung bình đám đông x rơi vào khoảng a và b (a, b) là 95%. Lúc này ta có thể diễn giải rằng với độ chính xác là 95% (hay chấp nhận 5% sai số) ta biết được thu nhập trung bình của đám đông nghiên cứu nằm trong khoảng (a, b) .

Công thức tính khoảng ước lượng:

$$E = \bar{X} \pm t_{\alpha, n-1} \frac{S_X}{n}$$

Hoặc: $E = p \pm t_{\alpha, n-1} S_p$

Với p là tỷ lệ % tần suất xuất hiện của một giá trị quan sát

4.4. Kiểm nghiệm giả thuyết (Hypothesis testing)

Bên cạnh việc ước lượng các đặc trưng của tổng thể, các dữ liệu mẫu thu thập được còn được dùng để đánh giá xem một giả thuyết nào đó về tổng thể là đúng hay sai. Ta gọi đó là kiểm nghiệm giả thuyết. Nói cách khác kiểm nghiệm giả thuyết là dựa

vào các thông tin mẫu để đưa ra kết luận bác bỏ hay chấp nhận về giả thuyết của tổng thể

Ví dụ: Sau một thời gian thực hiện các chương trình, biện pháp marketing (quảng cáo, khuyến mãi,...) công ty muốn đánh giá xem thị phần, doanh số có gì thay đổi so với trước không, hay có đạt được mục tiêu đề ra không.

Hoặc công ty muốn tìm hiểu xem sở thích của người tiêu dùng về kiểu dáng, màu sắc, mùi vị khác nhau về sản phẩm của công ty. Họ thích đặc biệt một kiểu dáng nào đó, một màu sắc nào đó, hay các kiểu dáng, màu sắc khác nhau đều được ưa thích như nhau.

Phương pháp kiểm nghiệm giả thuyết sẽ giúp giải quyết những yêu cầu này

Để kiểm nghiệm giả thuyết ta phải xây dựng giả thuyết. Giả thuyết đã hình thành được gọi là giả thuyết H_0 được xem như đúng cho đến khi ta có đủ căn cứ để kết luận khác hơn. Nếu giả thuyết H_0 không đúng thì phải có một giả thuyết nào đó khác H_0 gọi là H_1 là đúng. Một số giả thuyết thường gặp trong phân tích:

_____oO_____

CHƯƠNG 2: GIỚI THIỆU VỀ PHẦN MỀM SPSS

Là phần mềm chuyên dụng xử lý thông tin sơ cấp (thông tin được thu thập trực tiếp từ đối tượng nghiên cứu (người trả lời bảng câu hỏi) thông qua một bảng câu hỏi được thiết kế sẵn.

Thông tin được xử lý là thông tin định lượng (có ý nghĩa về mặt thống kê)

Phần mềm SPSS có tất cả 4 dạng màn hình:

1. Màn hình quản lý dữ liệu (data view):

Là nơi lưu trữ dữ liệu nghiên cứu với một cấu trúc cơ sở dữ liệu bao gồm cột, hàng và các ô giao nhau giữa cột và hàng

- **Cột (Column):** Đại diện cho biến quan sát. Mỗi cột sẽ chứa đựng tất cả các câu trả lời trong một câu hỏi được thiết kế trong bảng câu hỏi
- **Hàng (Row):** Đại diện cho một trường hợp quan sát (người trả lời), Ta phỏng vấn bao nhiêu người (tùy thuộc vào kích thước mẫu) thì ta sẽ có bấy nhiêu hàng. Mỗi hàng chứa đựng tất cả những câu trả lời (thông tin) của một đối tượng nghiên cứu
- **Ô giao nhau giữa cột và hàng (cell):** Chứa đựng một kết quả trả lời tương ứng với câu hỏi cần khảo sát (biến) và một đối tượng trả lời cụ thể (trường hợp quan sát)

2. Màn hình quản lý biến (variables view):

Là nơi quản lý các biến cùng với các thông số liên quan đến biến. Trong màn hình này mỗi hàng trên màn hình quản lý một biến, và mỗi cột thể hiện các thông số liên quan đến biến đó

- **Tên biến (name):** Là tên đại diện cho biến, tên biến này sẽ được hiển thị trên đầu mỗi cột trong màn hình dữ liệu
- **Loại biến (type):** Thể hiện dạng dữ liệu thể hiện trong biến. Dạng số, và dạng chuỗi
- **Số lượng con số hiển thị cho giá trị (Width):** Giá trị dạng số được phép hiển thị bao nhiêu con số.
- **Số lượng con số sau dấu phẩy được hiển thị (Decimals)**
- **Nhãn của biến (label):** Tên biến chỉ được thể hiện tóm tắt bằng ký hiệu, nhãn của biến cho phép nêu rõ hơn về ý nghĩa của biến.
- **Giá trị trong biến (Values):** Cho phép khai báo các giá trị trong biến với ý nghĩa cụ thể (nhãn giá trị)
- **Giá trị khuyết (Missing):** Do thiết kế bảng câu hỏi có một số giá trị chỉ mang tính chất quản lý, không có ý nghĩa phân tích, để loại bỏ các biến này ta cần khai

báo nó như là giá trị khuyết (user missing). SPSS mặc định giá trị khuyết (system missing) là một dấu chấm và tự động loại bỏ các giá trị này ra khỏi các phân tích thống kê.

- **Kích thước cột (columns):** Cho phép khai báo độ rộng của cột
- **Vị trí (align):** Vị trí hiển thị các giá trị trong cột (phải, trái, giữa)
- **Dạng thang đo (measures):** Hiển thị dạng thang đo của giá trị trong biến

3. Màn hình hiển thị kết quả (output):

Các phép phân tích thống kê sẽ cho ra các kết quả như bảng biểu, đồ thị và các kết quả kiểm nghiệm, các kết quả này sẽ được truy xuất ra một màn hình, và được lưu giữ dưới một tập tin khác (có đuôi là .SPO). Màn hình này cho phép ta xem và lưu giữ các kết quả phân tích.

4. Màn hình cú pháp (syntax):

Màn hình này cho phép ta xem và lưu trữ những cú pháp của một lệnh phân tích. Các cú pháp được lưu trữ sẽ được sử dụng lại mà không cần thao tác các lệnh phân tích lại.

5. Khái quát về phân tích dữ liệu

5.1. Kiểm tra dữ liệu (Data Screening)

Một thực tế luôn luôn gặp phải đối với những người làm công tác phân tích và xử lý số liệu là hầu như không lúc nào mà không gặp những vấn đề đối với dữ liệu trong tay họ, một số xuất hiện do lỗi nhập máy, lỗi mã hóa, hoặc do các lỗi về chọn mẫu và chất lượng phỏng vấn, tất cả những lỗi này thường dẫn đến những khác thường hoặc tính đại diện kém của dữ liệu thu thập.

Trong những cuộc nghiên cứu qui mô lớn, công việc kiểm tra dữ liệu đôi khi còn tốn nhiều công sức và thời gian hơn cả việc phân tích và tóm tắt dữ liệu. Do đó gần như là nhiệm vụ đầu tiên của người phân tích dữ liệu là phải tiến hành kiểm tra dữ liệu nhằm xác định ra các lỗi trong dữ liệu đồng thời kiểm tra xem tính tương thích của dữ liệu như thế nào so với những giả thuyết được yêu cầu cho các phân tích thống kê sau này.

▪ Xác định những giá trị vượt trội (Outliers) và các giá trị lỗi (Roque values)

Có nhiều cách để xác định ra các giá trị vượt trội và giá trị lỗi. Tuy nhiên điều quan trọng là xác định xem các giá trị vượt trội đó có phải là giá trị lỗi hay không hay do sự bất thường trong mẫu nghiên cứu:

- Sử dụng công cụ bảng phân bố tần xuất ngoài việc để đếm số lần xuất hiện của từng giá trị riêng biệt, nó còn giúp ta tìm ra các giá trị lỗi hoặc các giá trị mã hóa sai sót hoặc không mong đợi (ví dụ như biến giới tính chỉ có hai giá trị

mã hóa 1 và 2 tương ứng với giới tính nam và nữ do đó khi khảo sát ta sẽ phát hiện ra các giá trị khác với giá trị mã hóa 1 và 2). Ngoài ra công cụ này còn cho phép ta nhận ra được các giá trị khuyết (Missing values) nhưng lại xuất hiện như là một giá trị hợp lệ (Valid value)

- Đôi khi việc xác định các giá trị vượt trội có thể được xác định một cách tốt hơn khi ta khảo sát hai hay nhiều biến cùng một lúc. Đối với các biến dạng biểu danh (nominal) hoặc thứ tự (ordinal) sử dụng công cụ bảng chéo ta có thể xác định được những sự kết hợp phi lý giữa hai hoặc nhiều biến, ví dụ như một người chưa bao giờ tiêu dùng sản phẩm A nhưng lại tham gia đưa ra những ý kiến mức độ thỏa mãn trong tiêu dùng sản phẩm A.

5.2. Thống kê mô tả (Descriptive Statistics)

Đây có thể được xem là phần cốt lõi và thường gặp nhất trong việc phân tích và xử lý số liệu. Tuy nhiên trước khi bắt tay vào việc mô tả dữ liệu (đo lường độ tập trung hay phân tán, tỷ lệ %, mối quan hệ giữa các biến ...), cần thiết phải nắm được loại biến đang khảo sát (loại thang đo của biến) hay nói cách khác ta phải nắm được ý nghĩa của các giá trị trong biến

Đối với biến định danh hoặc thứ tự (nominal và ordinal) các phép tính toán số học như giá trị trung bình không có ý nghĩa thống kê, đặc biệt đối với biến định danh mọi sự so sánh hơn kém giữa các giá trị trong biến đều vô nghĩa. Ngược lại các biến định lượng như thang đo khoảng cách và thang đo tỷ lệ (Interval và Ratio) thì mọi sự so sánh hay tính toán số học đều có ý nghĩa phân tích thống kê

5.3. Kiểm nghiệm các so sánh trung bình mẫu (Tests for Comparing Means)

Trong phân tích thống kê người ta thường sử dụng các phép kiểm nghiệm kiểm nghiệm các giả thuyết về giá trị trung bình của các biến định lượng, và thống kê cung cấp cho ta các công cụ như kiểm nghiệm t (T-Test) hay kiểm nghiệm Z (Z-test)

▪ Kiểm nghiệm t cho một mẫu, cặp mẫu và hai mẫu ngẫu nhiên độc lập

Ta có ba dạng kiểm nghiệm t cho việc so sánh các giá trị trung bình của mẫu. Việc sử dụng dạng nào tùy thuộc vào vấn đề ta đang tiến hành so sánh cái gì

- Sử dụng kiểm nghiệm t cho hai mẫu ngẫu nhiên độc lập (Independent Samples T Test) là phương pháp nhằm mục đích kiểm nghiệm so sánh giá trị trung bình của một biến riêng biệt theo một nhóm có khác biệt hay không đối với giá trị trung bình của biến riêng biệt đó theo một nhóm khác. Với giả thuyết ban đầu H_0 cho rằng giá trị trung bình của hai nhóm này là bằng nhau. Ví dụ ta kiểm nghiệm thu nhập trung bình (biến thu nhập) theo hai nhóm giới tính là nam và giới tính là nữ (biến giới tính sử dụng để chia các giá trị quan sát trong biến thu nhập thành hai nhóm)

- Công cụ kiểm nghiệm t cho cặp mẫu (Paired-Samples T Test) được sử dụng để kiểm nghiệm có hay không giá trị trung bình của các khác biệt giữa các cặp quan sát là khác giá trị 0. Với giả thuyết ban đầu H_0 cho rằng giá trị trung bình các khác biệt này là bằng 0. Ví dụ như kiểm nghiệm sự khác biệt về điểm thi môn học của hai nhóm sinh viên có tham gia và không có tham gia chương trình phụ đạo ngoài giờ.
- Công cụ kiểm nghiệm t một mẫu (One-Sample T Test) để kiểm nghiệm có hay không giá trị trung bình của một biến là khác biệt với một giá trị giả định từ trước. Với giả thuyết ban đầu H_0 cho rằng giá trị trung bình kiểm nghiệm là bằng với giá trị giả thuyết đưa ra

▪ Phân tích phương sai một chiều (One-Way ANOVA)

Phân tích phương sai là một dạng mở rộng của phương pháp kiểm nghiệm t hai mẫu ngẫu nhiên độc lập (Independent-Samples T Test), và được sử dụng để kiểm nghiệm cho nhiều hơn hai nhóm. Phương pháp phân tích này khảo sát sự biến thiên giữa các trung bình mẫu trong mỗi liên hệ với sự phân tán của các quan sát trong từng mỗi nhóm. Với giả thuyết ban đầu H_0 cho rằng các giá trị trung bình này là bằng nhau.

5.4. Kiểm nghiệm các mối quan hệ (Testing Relationships)

Kiểm nghiệm mối quan hệ giữa hai biến và kiểm nghiệm mối tương quan với cường độ tương quan và chiều của tương quan giữa các biến trong cơ sở dữ liệu

- Trong kiểm nghiệm mối quan hệ giữa hai biến, ta sử dụng kiểm nghiệm Chi-bình phương để kiểm nghiệm giả thuyết ban đầu cho rằng hai biến thể hiện trong bảng chéo (biến cột và biến hàng) là không có mối quan hệ với nhau (độc lập với nhau).
- Trong kiểm nghiệm tương quan giữa các biến ta sử dụng kiểm nghiệm F kiểm nghiệm giả thuyết ban đầu cho rằng giữa các biến đang khảo sát không có tương quan với nhau (hệ số tương quan $R = 0$)

___o0o___

CHƯƠNG 3: CHUẨN BỊ DỮ LIỆU

1. Kiểm tra và hiệu đính dữ liệu

Đây là bước kiểm tra chất lượng thông tin trong bảng câu hỏi nhằm bảo đảm không có bảng câu hỏi nào thiếu hoặc chứa đựng những thông tin sai sót theo yêu cầu thiết kế ban đầu, bước này cần thiết được thực hiện trước khi tiến hành mã hóa và nhập dữ liệu vào máy tính. Người kiểm tra phải bảo đảm tính toàn vẹn và tính chính xác của từng bảng câu hỏi & từng câu trả lời trong bảng câu hỏi. Thông thường bước này nhân viên nghiên cứu sẽ tiến hành kiểm tra những đặc tính sau của bảng câu hỏi:

- **Tính logic của các câu trả lời:** Đôi khi trong bảng câu hỏi, do yêu cầu nghiên cứu sẽ có những đường dẫn, những điều kiện để người trả lời hoặc có thể trả lời tất cả các câu hỏi hoặc có thể bỏ qua một vài câu hỏi nào đó. Kiểm tra tính logic của bảng câu hỏi cho phép nhà nghiên cứu loại bỏ những câu trả lời thừa, cũng như kịp thời bổ xung những phần thiếu trong bảng câu hỏi. Tính logic của câu trả lời còn phụ thuộc vào sự kết dính và liên hệ lẫn nhau giữa các câu hỏi trong một bảng câu hỏi (đôi khi một câu trả lời là có ý nghĩa nếu đứng riêng một mình nó những lại vô nghĩa nếu kết hợp so sánh với các câu trả lời trước hoặc sau nó).
- **Tính đầy đủ của một câu trả lời và của một bảng câu hỏi:** Một bảng câu hỏi chỉ có giá trị nếu như tất cả những câu hỏi theo yêu cầu đều được trả lời đầy đủ. Mỗi câu hỏi trong bảng câu hỏi đều có một ý nghĩa, một giá trị nghiên cứu nhất định, do đó thiếu một câu trả lời nào đó cho một câu hỏi cụ thể nào đó sẽ làm mất đi giá trị của bảng câu hỏi đó.
- **Tính hợp lý và xác thực của các câu trả lời:** Một câu trả lời đầy đủ chưa hẳn là câu trả lời có giá trị, do đó tính chân thực và hợp lý của câu trả lời cũng quyết định đến giá trị của câu trả lời và của bảng câu hỏi, đặc biệt là các câu hỏi chấm điểm, câu hỏi mở và các câu hỏi mang tính logic.

Quá trình kiểm tra, rà soát lại bản câu hỏi là nhằm mục đích kiểm tra, phát hiện, sửa chữa và thông báo kịp thời cho người thu thập dữ liệu tránh những sai sót tiếp theo.

Để xử lý các lỗi trong kiểm tra và hiệu đính, ta có thể lựa chọn cách xử lý như sau tùy thuộc vào mức độ sai sót cụ thể:

- Trả về cho bộ phận thu thập dữ liệu để làm sáng tỏ vấn đề
- Suy luận từ các câu trả lời khác
- Loại bỏ toàn bộ bản câu hỏi

2. Mã hoá dữ liệu

Là quá trình chuyển dịch câu trả lời thực của người trả lời vào từng nhóm, từng mẫu đại diện với các giá trị đại diện tương ứng nhằm làm cho quá trình tóm tắt, phân tích và nhập liệu được dễ dàng và hiệu quả hơn. Có hai dạng mã hóa:

- **Tiền mã hóa:** Là việc mã hóa cho các câu hỏi đóng. Do đặc điểm của các loại câu hỏi này là nhà nghiên cứu đã có sẵn các câu trả lời từ trước, người trả lời chỉ việc lựa chọn câu trả lời nào phù hợp nhất với ý kiến của mình, do đó việc mã hóa cho các câu hỏi này thường được tiến hành từ trước, ở giai đoạn thiết kế bảng câu hỏi.
- **Mã hoá:** Trong bảng câu hỏi ngoài những câu hỏi đóng nêu ở trên, còn những câu hỏi mở, là những câu hỏi mà người trả lời tự do đưa ra câu trả lời theo suy nghĩ và diễn giải của chính họ. Các bảng câu hỏi nhận về thường có những câu trả lời rất khác nhau và rất đa dạng. Do đó công việc mã hóa những câu trả lời này thì cần thiết cho quá trình kiểm tra, nhập liệu, tóm tắt và phân tích sau này.

Mục đích của mã hóa là tạo nhãn cho các câu trả lời, thường là bằng các con số. Mã hóa còn giúp giảm thiểu số lượng các câu trả lời bằng cách nhóm các câu trả lời vào những nhóm có cùng ý nghĩa. Tiền trình mã hóa có thể được tiến hành như sau:

- Đầu tiên, xác định loại câu trả lời cho những câu hỏi tương ứng. Những câu trả lời này có thể thu thập từ một mẫu các bảng câu hỏi đã hoàn tất, thường là 25% trên tổng số bảng câu hỏi
- Bước tiếp theo là xây dựng một danh sách liệt kê các câu trả lời, các câu trả lời được liệt kê và tiến hành nhóm các câu trả lời theo những nhóm đặc trưng (có cùng ý nghĩa)
- Cuối cùng, những nhóm câu trả lời này được gán cho một nhãn hiệu, một giá trị, thường là một con số cụ thể

_____oOo_____

CHƯƠNG 4: ĐỊNH BIẾN VÀ NHẬP DỮ LIỆU

1. Khái niệm về biến và các giá trị trong biến

Biến là tập hợp những trả lời cho một câu hỏi. Có hai loại biến như sau:

▪ **Phân loại biến theo số lượng câu trả lời:**

- **Biến một trả lời:** Biến dành cho câu hỏi có một trả lời
- **Biến nhiều trả lời:** Các biến dành cho nhiều câu trả lời có thể có trong một câu hỏi nhiều trả lời

Ví dụ như trong bảng câu hỏi có hai câu hỏi sau:

- Câu hỏi 1: Hãy cho biết bạn ở nhóm tuổi nào trong số những nhóm tuổi sau:

<u>Nhóm tuổi</u>	<u>code</u>
Dưới 18	1
19 đến 30	2
31 đến 40	3
41 đến 50	4
Trên 50	5

- Câu hỏi 2: Nói đến điện thoại di động, bạn biết được những nhãn hiệu nào trong danh sách liệt kê dưới đây

<u>Nhãn hiệu</u>	<u>code</u>
Ericson	1
Motorola	2
Nokia	3
Siemens	4
Panasonic	5
....V.V	

Có thể thấy đối với câu hỏi 1, người trả lời chỉ có thể đưa ra một câu trả lời duy nhất về tuổi của mình, do đó biến chứa đựng câu trả lời của câu hỏi 1 là biến một trả lời. Trong khi xem xét câu hỏi 2, người trả lời có thể nêu ra nhiều nhãn hiệu mà họ có biết qua, do đó phải có nhiều biến chứa đựng các trả lời có thể có, ta gọi biến đó là biến nhiều trả lời.

▪ **Phân loại biến theo kiểu dữ liệu:**

Có hai loại biến chính là biến định tính và biến định lượng, đối với biến định tính ta không thể sử dụng các phép toán (cộng, trừ, nhân, chia) để tính toán các giá trị

trên biến đó, ngược lại biến định lượng cho phép ta thao tác các phép toán trên các giá trị mà nó đại diện. Việc xác định dạng biến theo cách này cho phép ta lựa chọn được tham số thống kê tương thích để phân tích.

Để xác định được biến là định lượng hay định tính đòi hỏi phải xác định các giá trị trong biến thuộc dạng thang đo nào trong bốn dạng thang đó sau:

- **Thang đo định danh (Nominal Scale):** Trong dạng thang đo này các con số được sử dụng đơn thuần như một giá trị xác định sự khác biệt cho các câu trả lời, các giá trị quan sát có ý nghĩa khác biệt nhau. Đối với loại thang biểu danh các giá trị số được sử dụng như là ký số nhận dạng và không có giá trị về một thứ tự cao thấp và độ lớn giữa các con số
- **Thang đo thứ tự (Ordinal Scale):** Trong dạng thang đo này dữ liệu được xếp xếp các giá trị quan sát theo một thứ tự cao thấp nhất định, nhưng không diễn tả được độ lớn giữa vị trí cao thấp giữa các con số. Tóm lại thang đo thứ tự bao gồm cả thông tin về biểu danh đồng thời cung cấp luôn mối quan hệ theo thứ tự giữa các giá trị nhưng không đo được khoảng cách giữa các giá trị đó.
- **Thang đo khoảng cách (Interval Scale):** Giống như đặc tính của thang đo thứ tự, tuy nhiên đối với thang đo khoảng cách cho phép ta đo được khoảng cách giữa các giá trị. Tuy nhiên do thang đo khoảng cách không xác định được điểm 0 chung (giống như thang đo nhiệt độ) do đó ta chỉ có thể nói giá trị này lớn hơn giá trị kia bao nhiêu đơn vị nhưng không thể kết luận giá trị này lớn hơn giá trị kia bao nhiêu lần.
- **Thang đo tỷ lệ (ratio):** Đây là thang đo có đủ các đặc tính thứ tự và khoảng cách. Ngoài ra việc xác định ra tỷ số chênh lệch giữa các giá trị là có thể thực hiện do ở thang đo này điểm 0 được xác định một cách có ý nghĩa.

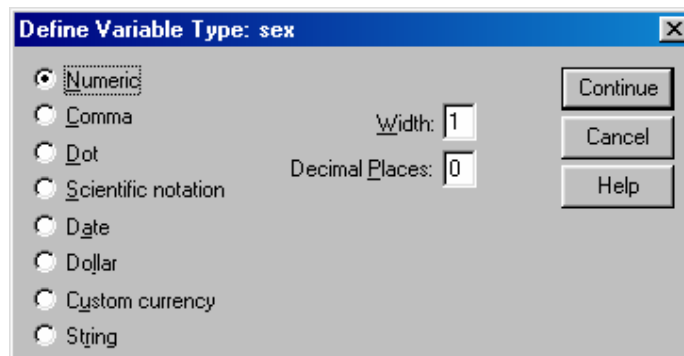
Từ bốn dạng thang đo trên ta phân ra hai loại biến. Biến định tính là biến chứa các giá trị quan sát ở dạng thang đo biểu danh và thứ tự. Còn biến định lượng là biến chứa các giá trị có dạng thang đo khoảng cách và tỷ lệ.

2. Phương pháp định biến trên SPSS (Define Variable)

Định biến trong màn hình quản lý biến (variables view). Công việc định biến này có thể được thực hiện trước khi tiến hành nhập dữ liệu vào trong máy

Mục đích của việc định biến là gán nhãn và các thông số cho các biến và gán ý nghĩa cho các giá trị trong biến. Sau khi được mã hóa các dữ liệu sẽ được đại diện bằng những con số và các con số này có ý nghĩa khác nhau tùy theo câu trả lời thu thập được. Để các con số này có thể nhập vào máy tính và có thể quản lý cũng như có ý nghĩa trong SPSS, ta phải tiến hành định biến cho dữ liệu. Quy trình định biến này bao gồm các bước sau:

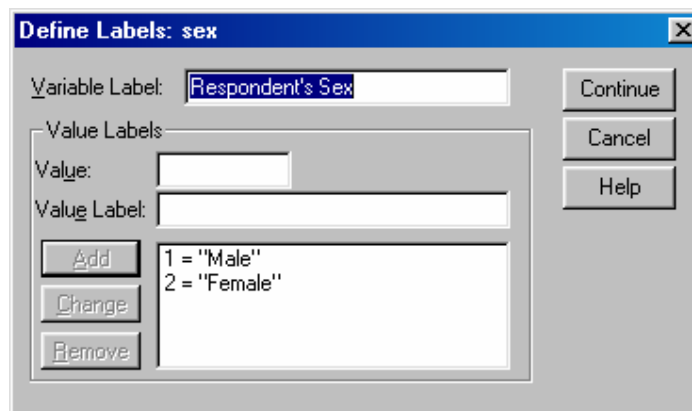
- **Gán tên cho biến (Name):** Ta gõ tên biến cần khai báo vào cột đầu tiên trong màn hình Variables view (Nếu ta không gõ tên biến vào thì SPSS sẽ mặc định tên biến này là **Var000001**). Tên biến được khai báo này sẽ hiển thị trên đầu các cột trong màn hình Data view. Tên biến bị hạn chế về số ký tự hiển thị, do đó cần thiết phải khai báo ngắn gọn và dễ gợi nhớ, thông thường nên đặt theo thứ tự câu hỏi trong bảng câu hỏi như q1, q3, q4a, ...Có một số qui ước sau đây phải tuân theo khi khai báo tên biến:
 - Bắt đầu bằng một chữ cái và không bắt đầu bằng dấu chấm(.).
 - Tên biến không được qua 8 ký tự
 - Không được chứa khoảng trắng và các ký tự đặc biệt như (!), (?), (*).
 - Các từ khóa sau đây không được dùng làm tên biến: ALL, NE, EQ, TO, LE, LT, BY OR, GT, AND, NOT, GE, WITH
- **Định ra kiểu biến (Type):** Có các dạng biến sau có thể định dạng. Dạng con số (numeric); Dạng tiền tệ; dạng ngày (Date) hoặc dạng chuỗi (String). Ngoài ra phần này cũng cho phép ta định dạng các dạng số được hiển thị khác nhau (Xem hình 4-1)



Hình 4-1

Tùy thuộc vào yêu cầu của dữ liệu, mà ta sẽ định loại biến cho biến, SPSS mặc định loại biến là kiểu số (numeric); ngoài ra còn có thể khai báo các kiểu hiển thị số khác nhau như kiểu số có dấu phẩy (Comma) hay dấu chấm (Dot) ngăn cách giữa các khoảng cách hàng ngàn của con số; cách hiển thị theo các ký hiệu khoa học (Scientific notation); Hiển thị ngày, dollar và các kiểu tiền tệ khác; cuối cùng là cách hiển thị dạng chuỗi.

- **Xác định số lượng con số hiển thị cho giá trị (Width) và số lượng con số sau dấu phẩy hiển thị (Decimals):** Khai báo bề rộng của con số (hàng đơn vị, hàng trăm, hàng triệu, ...) trong ô Width, Và khai báo số con số thập phân sau dấu phẩy trong ô Decimal.
- **Gán nhãn cho biến (Variable Label):** Đặt tên nhãn cho biến một cách đầy đủ hơn, tên biến này sẽ hiển thị ý nghĩa của biến trên các kết quả phân tích trong màn hình kết quả (output), công cụ này giúp ta hiểu được ý nghĩa của biến đang khảo sát dễ dàng hơn trong quá trình phân tích.
- **Định tên cho các giá trị trong biến (Value labels):** Trong quá trình mã hóa dữ liệu ta đã gán các giá trị trong biến thành các con số đại diện, Nhưng để cho quá trình đọc và phân tích các kết quả nghiên cứu dễ dàng hơn ta phải gán các con số này các ý nghĩa như nó mà nó đang đại diện, công cụ định lại nhãn cho giá trị cho phép ta thực hiện điều này (Xem hình 4-2):

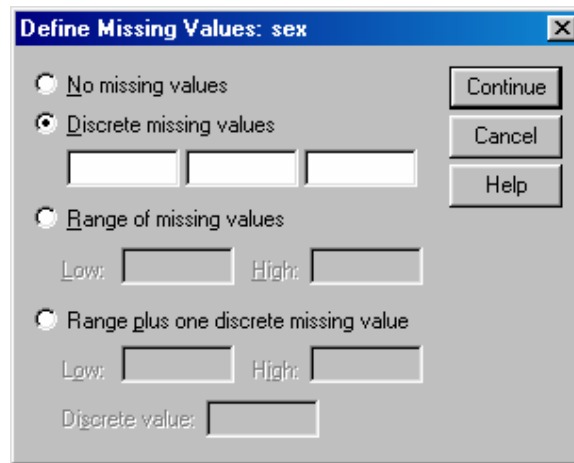


Hình 4-2

Gán nhãn của giá trị (value labels) có ba thao tác:

- Gán một nhãn mới:
 - Nhập giá trị vào hộp thoại Value
 - Nhập nhãn của giá trị vào hộp thoại Value Label
 - Ấn nút Add để xác định nhãn đó
- Sửa đổi một nhãn:

- Di vệt sáng đến nhãn cần sửa đổi
- Nhập tên nhãn mới, ấn nút Change để thay đổi
- Loại bỏ một nhãn:
 - Di vệt sáng đến nhãn cần loại bỏ
 - Ấn nút Remove để loại bỏ
- **Định nghĩa các giá trị khuyết (Missing Values):** Được dùng để định ra các giá trị cụ thể cho các giá trị mà ta muốn loại bỏ ra khỏi các phân tích và xử lý thống kê sau này hay còn gọi là các giá trị khuyết. Ví dụ trong câu hỏi về thu nhập, sẽ có một số trường hợp từ chối trả lời tương ứng với giá trị mã hóa là 99. Trong quá trình phân tích để loại bỏ tất cả các trường hợp này ra khỏi các xử lý thống kê, ta phải tiến hành khai báo giá trị 99 là giá trị khuyết trong phần giá trị khuyết (Missing values). (Xem hình 4-3)



Hình 4-3

SPSS mặc định là không có khai báo giá trị khuyết. Có ba cách để khai báo các giá trị khuyết

- (1) khai báo bằng 3 giá trị rời rạc (Discrete missing values)
- (2) Khai báo một chuỗi liên tục các giá trị (Range of missing values)
- (3) Khai báo một chuỗi các giá trị khuyết và một giá trị khuyết riêng biệt (Range plus one discrete missing value)

Đối với dữ liệu dạng chuỗi. Toàn bộ các giá trị vô dụng hoặc trống đều được xem là có nghĩa. Để định nghĩa các giá trị vô nghĩa và các giá trị trống là giá trị khuyết ta phải nhập vào một khoảng trống vào trong ô định ra các giá trị khuyết riêng biệt

- **Định kích cỡ cho cột (Column format):** Định ra chiều rộng của cột đang khai báo biến

- **Định ra vị trí hiển thị các giá trị (align):** Vị trí hiển thị các giá trị trong cột (phải, trái, giữa)
- **Định ra dạng thang đo mà biến thể hiện (measurement):** Tùy thuộc vào dạng thang đo được sử dụng trong biến mà ta khai báo trong công cụ measurement, chú ý khai báo scale được dùng chung cho dạng thang đo khoảng cách và thang đo tỷ lệ. Việc khai báo này chỉ mang tính chất quản lý không ảnh hưởng đến kết quả phân tích

3. Nhập dữ liệu

Dữ liệu cần nhập sẽ được nhập vào trong màn hình Data views. Màn hình này thể hiện ra một ma trận thông tin bao gồm: cột và hàng, và ô giao nhau giữa cột và hàng. (Xem hình 2-1)

Dữ liệu được nhập theo trình tự sau:

- Khai báo tên biến chứa đựng thông tin cần nhập vào thanh bên trên mỗi cột (tên mặc định của các cột này trong SPSS là var00001, ..., var0000x). Phần này đã được đề cập chi tiết trong phần định biến.
- Chọn ô cần nhập dữ liệu, là phần giao nhau giữa cột và hàng. Ô cần nhập sẽ có khung viền chung quanh báo cho người nhập biết đó là ô đang hoạt động, tên biến và số hiệu hàng được hiện ở góc trái của cửa sổ.
- Gõ giá trị cần nhập vào khung đã chọn, giá trị này được hiện trong thanh sửa đổi (cell editor) nằm ở trên cửa sổ. Chú ý khi nhập dữ liệu phải bảo đảm đúng với kiểu biến đã được định nghĩa. Thông thường các kiểu biến được khai báo là dạng chuỗi (ngắn tối đa 8 ký tự) hoặc dạng số, nhằm bảo đảm tính tương thích cho việc phân tích sau này.

Ta cũng có thể nhập liệu từ các phần mềm khác như Excel, Fox, ... và sau đó chuyển vào trong SPSS.

_____oOo_____