BÀI 1: NHỮNG VẤN ĐỀ CHUNG CỦA THỐNG KÊ ỨNG DỤNG TRONG NGHIÊN CỨU TÂM LÝ HỌC

1 Khái niệm thống kê

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

Thống kê thường được phân chia thành 2 lĩnh vực:

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

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

2. Mối quan hệ giữa thống kê và nghiên cứu tâm lý học

Thống kê học là một môn khoa học xã hội, bởi vì thống kê nghiên cứu các hiện tượng tâm lý – xã hội nảy sinh trong quá trinh tồn tại và phát triển của đời sống xã hội. Các hiện tượng và quá trình đó thường là:

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

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

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

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

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

* Các hiện tượng tâm lý con người như nhu cầu, động cơ, thị hiếu,

nguyện vọng, thái độ ….của các nhóm xã hội khác nhau.

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

Với đối tượng nghiên cứu của thống kê là các hiện tượng xã hội, phương pháp của thống kê là phân tích, đánh giá bản chất khách quan và tính quy luật của hiện tượng bằng các công cụ toán học… do đó, thống kê toán học được các nhà tâm lý học sử dụng như một phương pháp phân tích những số liệu được thu thập từ các phương pháp nghiên cứu của mình, qua đó giúp nhà nghiên cứu đánh giá được đúng bản chất, quy mô, tính quy luật của các hiện tượng tâm lý nảy sinh trong đời sống văn hóa – xã hội.

3. **Các khái niệm thường dùng trong thống kê**

3.1.1 Khái niệm

Số đo: là việc gán những dữ kiện lượng hoá hay những ký hiệu cho những hiện tượng quan sát

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

3.1.2. Các khái niệm cơ bản thường dùng trong thống kê

Tổng thể thống kê

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

Mẫu

Mẫu là một bộ phận của tổng thể, đảm bảo được tính đại diện và được chọn ra để quan sát và dùng để suy diễn cho toàn bộ tổng thể. Việc chọn mẫu đại diện cho tổng thể không phải dễ dàng, trên thực tế chỉ cố gắng giảm sự sai biệt giữa mẫu và tổng thể chứ không thể khắc phục hoàn toàn.

Tiêu thức thống kê

Các đơn vị tổng thể thường có nhiều đặc điểm khác nhau, song trong thống kê người ta chỉ chọn một số đặc điểm để nghiên cứu, các đặc điểm này người ta gọi là tiêu thức thống kê. Như vậy, tiêu thức thống kê là các đặc điểm của đơn vị tổng thể. Mỗi tiêu thức thống kê đều có giá trị biểu hiện của nó.

Phân loại tiêu thức thông kê dựa vào sự biểu hiện :

+ Tiêu thức thuộc tính : là tiêu thức phản ánh loại hoặc tính chất của đơn vị

+ Tiêu thức số lượng : là đặc trưng của đơn vị tổng thể được biểu hiện bằng con số. Gồm 2 loại:

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

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

Tham số thống kê

Là giá trị quan sát được của tổng thể và dùng để mô tả đặc trưng của hiện tượng nghiên cứu. Ví dụ: trung bình tổng thể, tỷ lệ tổng thể…

Tham số mẫu

Là giá trị tính toán được của một mẫu và được dùng để suy rộng cho tham số tổng thể. Ví dụ: trung bình mẫu, tỷ lệ mẫu…

3.1. Các kiểu thang đo đặc trưng

3.1.1. **Khái niệm**

- Số đo: là việc gán những dữ kiện lượng hoá hay những ký hiệu cho những hiện tượng quan sát
- Thang đo: là tạo ra một thang điểm để đánh giá đặc điểm của đối tượng được nghiên cứu thể hiện qua sự đánh giá, nhận xét.

3.1.2. Các loại thang đo

Thang đo danh nghĩa – thang đo định danh:

Là loại thang đo sử dụng cho các dữ liệu thuộc tính mà các biểu hiện của dữ liệu không có sự hơn kém, khác biệt về thứ bậc. Các con số không có quan hệ hơn kém, không thực hiện được các phép tính đại số.

Ví dụ:

Câu 1: Anh chị cho biết giới tính

1. Nam

2. Nữ

Thang đo thứ bậc:

Là loại thang đo dành cho các dữ liệu thuộc tính. Trường hợp này biểu hiện dữ liệu có sự so sánh, không thực hiện được các phép tính đại số.

Ví dụ:

Câu 1: Mức độ yêu thích sản phẩm quần áo may sẵn nhãn hàng A

1. Rất yêu thích;

2. Có yêu thích nhưng không nhiều;

3. Rất không yêu thích.

Thang đo khoảng:

Là loại thang đo dành cho các dữ liệu số lượng. Đây là loại thang đo được dùng để xếp hạng các đối tượng nghiên cứu nhưng khoảng cách bằng nhau trên thang đo đại diện cho khoảng cách bằng nhau trong đặc điểm của đối tượng. Thang đo này có thể thực hiện phép tính đại số, trừ phép chia (:) không có ý nghĩa

Ví dụ:

Câu 2: Thu nhập trung bình một tháng của anh chị là bao nhiêu

1. Dưới 2 triệu;

2. Từ 2 triệu đến 4 triệu;

3. Từ 4 triệu đến 6 triệu;

4. từ 6 triệu đến 8 triệu…

Thang đo tỷ lệ: l

Là loại thang đo có thể dùng dữ liệu số lượng. Ngoài đặc tính của thang đo khoảng, trong thang đo này phép chia có thể thực hiện.

3.2. Biến số và các kiểu biến số

3.2.1. Khái niệm biến số:

Biến số là một đơn vị thông tin mà nhà nghiên cứu cần tìm hiểu, khai thác. Trong bảng điều tra, biến số là tập hợp những trả lời cho một câu hỏi. Có hai loại biến như sau:

3.2.2. Phân loại

3.2.2.1. Phân loại biến theo số lượng câu trả lời:

Biến một trả lời: Biến dành cho câu hỏi có một trả lời

Câu 1: Nghề nghiệp chính của ông bà/anh chị:

- Trồng trọt □ *- Chăn nuôi/nuôi trồng thuỷ sản* □

- Nghệ nhân/thợ thủ công □ *- Kinh doanh/buôn bán* □

- Chế biến sản phẩm nông nghiệp □*- Công nhân/lao động tự do* □

Với câu hỏi này, người trả lời chỉ có thể chọn 1 phương án đúng với bản thân mình nhất.

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

Câu 10. Anh chị cho biết gia đình mình có những tài sản nào dưới đây?

- Ti vi: □ *- Tủ lạnh:* □ *- Máy móc nông cụ:* □

- Đồ nội thất tốt: □ *- Xe máy:* □ *- Ô tô:* □

- Điều hoà nhiệt độ □*- Máy vi tính kết nối internet:* □

Với câu hỏi này, người trả lời có thể chọn 2 hay nhiều phương án.

3.2.2.2. Phân loại biến theo kiểu dữ liệu:

Có hai loại biến chính là biến định tính và biến định lượng:

Biến định tính (quatitative variables)

Là những biến mà người ta gán cho các giá trị để phân biệt hay phân loại các quan sát. Đây là biến lập nhóm (categorical variables), trị số của chúng được xác định bằng các thang đo định danh hoặc thang đo thứ bậc dưới dạng mã số hoặc chuỗi ngắn.

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.

Ví dụ: Giới tính (nam, nữ); Trình độ học vấn (Mù chữ, tiểu học, trung học, cao đẳng, đại học, trên đại học)…; Thu nhập (thấp, trung bình, khá, cao…)

Biến định lượng (quantitative variables)

Là những biến mà các giá trị của chúng được xác định bằng các thang đo khoảng nên trị số của chúng luôn để dưới dạng số. Biến định lượng cho phép ta thao tác các phép toán trên các giá trị mà nó đại diện.

Ví dụ: Thu nhập (200.000đ; 220.000đ; 211.000đ…), tuổi (15; 17; 19; 18; 16…), số lượng tài sản có trong gia đình: Tivi; tủ lạnh, xe máy…

3.2.2.3. Phân loại theo vị trí ảnh hưởng

Biến độc lập (independent variable)

Biến độc lập là một đặc tính được lựa chọn để nghiên cứu. Biến độc lập được giả thuyết là một biến mà sự biến đổi của nó có ảnh hưởng chi phối hoặc gây ra những biến đổi kéo theo ở một biến khác.

Biến phụ thuộc (dependent variable)

Biến phụ thuộc là một biến mà sự biến đổi của nó chịu sự chi phối (đáp ứng) của 1 biến khác. Một biến được gọi là biến phụ thuộc khi giá trị của nó tuỳ thuộc vào giá trị của biến độc lập. Nó chính là hiệu quả giả định của biến độc lập.

Lưu ý: Việc xác định một biến là độc lập và phụ thuộc thường có tính chất tương đối. Một biến có thể được xem là phụ thuộc trong phạm vi phân tích này lại là độc lập trong phạm vi phân tích khác.

Trong nghiên cứu còn có những yếu tố ảnh hưởng không được kiểm soát (hay không được quan sát một cách có hệ thống) được gọi là các biến bổ trợ.

3.3. Các số đo đặc trưng trong thống kê

3.3.1. Yếu vị (mode)

Yếu vị của một tập hợp các đo lường là trung điểm của khoảng đẳng loại chứa đựng tần số tối đa hay trong trường hợp các biến định tính, nó là tên của loại đo lường có tần số lớn nhất.

3.3.2. 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 xắp 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 hay đầu mút của dãy phân phối.

Trung vị của một tập hợp đo lường là trị số rơi vào chính giữa khi các số đo lường ấy được xếp đặt theo thứ tự độ lớn của chúng

Công thức tính trung vị = 1/2 (N+1)

Nếu trung vị là số lẻ thì lấy giá trị trung bình của thứ hạng đứng trường và sau.

3.33. Trung bình cộng (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.

Trung bình cộng của một tập hợp các số đo lượng là tổng số cộng các đo lường chia cho N (tổng số) của đo lường ấy.

3.3.4.. Phương sai (varian)

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 cũng là một phép đo đánh giá mức độ phân tán hoặc thay đổi của một phân bố điểm. Phương sai chính là bình phương độ lệch chuẩn.

3.3.5. 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.

3.3.6. Tương quan

Tương quan (correlation) là một số đo lường về mối liên hệ giữa hai biến số. Nó có thể là dương (+) hoặc (-) hay = 0.

3.3.7. 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: công ty muốn tìm hiểu xem sở thích của người tiêu dùng về kiểu dáng, màu sắc, mùi vị khác nhau về sản phẩm cuả 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 nhưng 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.

BÀI 2: LÀM VIỆC VỚI PHẦN MỀM SPSS

1. Giới thiệu khái quát về phần mềm SPSS

1.1. SPSS là gì?

SPSS là một sản phẩm phần mềm chuyên ngành thống kê. Lúc đầu được sử dụng cho các máy chủ (máy trung tâm -mainframes) vào những năm 1960s, sau này được sử dụng cho các máy tính cá nhân.

Sản phẩm SPSS được viết tắt từ **S**tatistical **P**roducts for the **S**ocial **S**ervices, có nghĩa là Các sản phẩm Thống kê cho các dịch vụ xã hội. Phiên bản mới nhất là SPSS 17.0.

SPSS là một hệ thống phần mềm thống kê toàn diện được thiết kế để thực hiện tất cả các bước trong các phân tích thống kê từ những thông kê mô tả (liệt kê dữ liệu, lập đồ thị) đến thống kê suy luận (tương quan, hồi quy…)

1.2. Các bộ phận của hệ thống SPSS

SPSS Professional Statistisc: Cung cấp các kỹ thuật để phân tích dữ liệu dạng không thích hợp với mô hình tuyến tính truyền thống.

SPSS Ađvance Statistisc: Tập trung vào các kỹ thuật được dùng trong các thí nghiệm sinh học và phức tạp.

SPSS Tables: Xây dựng một loạt các báo cáo dạng bảng biểu có chất lượng trình bày cao, và phức tạp.

SPSS Trends: Thực hiện các phép dự đoán và phân tích dãy số thời gian phức tạp bao gồm xây dựng các mô hình cho dữ liệu đa biến phi tuyến tính, các mô hình san bằng, và các phương pháp để ước lượng các hàm tự hồi quy.

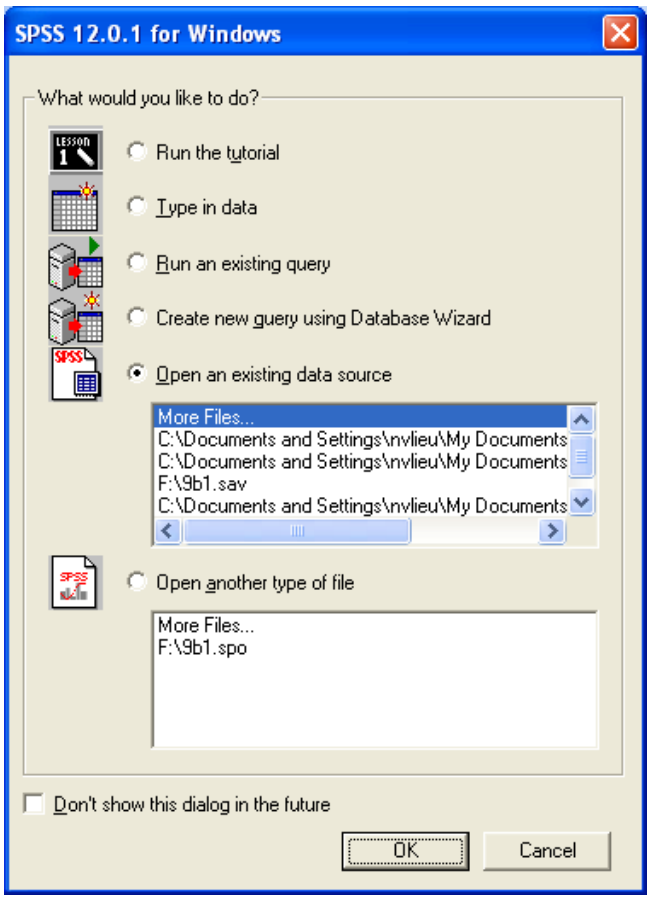
2. Làm việc với SPSS

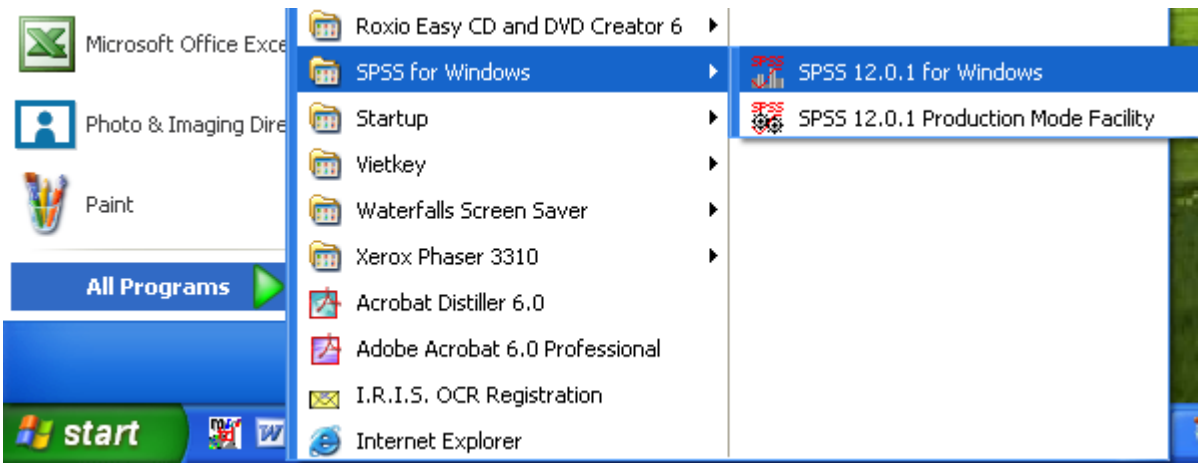
2.1. Khởi động chương trình SPSS trên máy tính

Trên màn hình desktop của Windows nhắp vào biểu tượng

Hoặc mở phím Start, All programs, SPSS for Windows, SPSS 12.0.1 for Windows (hoặc tùy phiên bản cài đặt)

Hình 2.1. Khởi động SPSS





Sẽ xuất hiện cửa sổ SPSS Data Editor và một hộp thoại như sau:

◉ **Run the tutorial**: Chạy chương trình trợ giúp

◉ **Type in data**: Nhập dữ liệu mới

◉ **Run an existing query**: Chạy một truy vấn dữ liệu đã có sẵn

◉ **Create new query using Database Wizard**: Lập một truy vấn dữ liệu sử dụng Database Wizard

◉ **Open an existing data source**: Mở file dữ liệu đã có sẵn

(Chú ý: Hộp thoại này chỉ xuất hiện một lần khi bạn khởi động SPSS)

Bảng chọn {Menu}

Rất nhiều nhiệm vụ bạn muốn tiến hành với *SPSS* bắt đầu với việc lựa chọn các menu {trình đơn}. Từng cửa sổ trong *SPSS* có các menu riêng của nó với các lựa chọn menu thích hợp cho loại cửa sổ đó.

Hai menu Analysis và Graphs là có sẵn đối với mọi loại cửa sổ, làm cho việc tạo các kết xuất mới rất nhanh chóng mà không phải chuyển đổi giữa các cửa sổ.

Thanh công cụ {Toolbars}

Từng cửa sổ SPSS có các thanh công cụ riêng của nó cho phép truy cập nhanh đến các nhiệm vụ thông dụng. Có một số cửa sổ có hơn một thanh công cụ.

Hình 2-2: Thanh công cụ với trợ giúp chỉ dẫn công cụ {ToolTip Help}

Có thể điều khiển thanh công cụ theo nhiều cách:

Hiện hoặc ẩn các biểu tượng

Trình bày thanh công cụ theo phương nằm ngang hay thẳng đứng, gắn bên trên, dưới, trái hoặc phải..

Để ẩn hiện một thanh công cụ

Thực hiện theo cú pháp sau: Từ thanh **Menu/View/Toolbar**

Trong hộp thoại Show Toolbar, chọn thanh công cụ mà bạn muốn mở (hoặc ẩn)

2.2. Làm quen với màn hình SPSS

Phần mềm SPSS có tất cả 4 dạng màn hình:

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

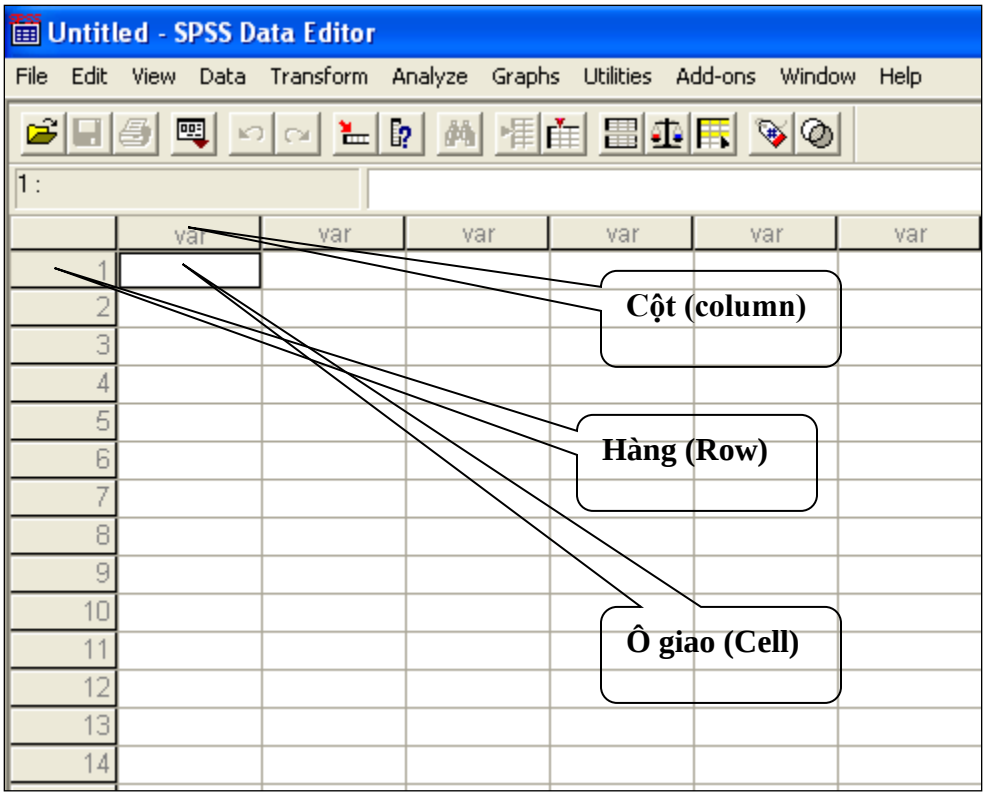
Là nơi lưu trữ dữ liệu nghiên cứu với một cấu trúc cơ sở dữ liệu bao gồm cột, hàng và các ô giao nhau giữa cột và hàng

+ 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)

Hình 2.3. Màn hình quả lý dữ liệu (Data view)



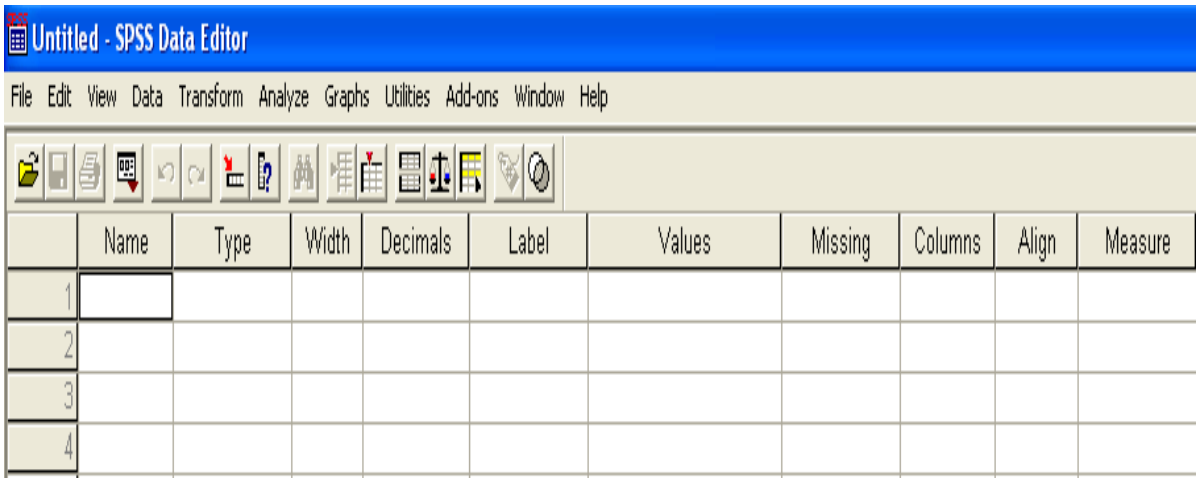
2.2.2 Màn hình quản lý biến (variables view):

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 tắc 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 khuyến (system missing) là một dấu chấm và tự động loại bỏ các giá trị này ra khỏi các phân tích thống kê.
- Kích thướt cột (columns): Cho phép khai báo độ rộng của cột
- Ví 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

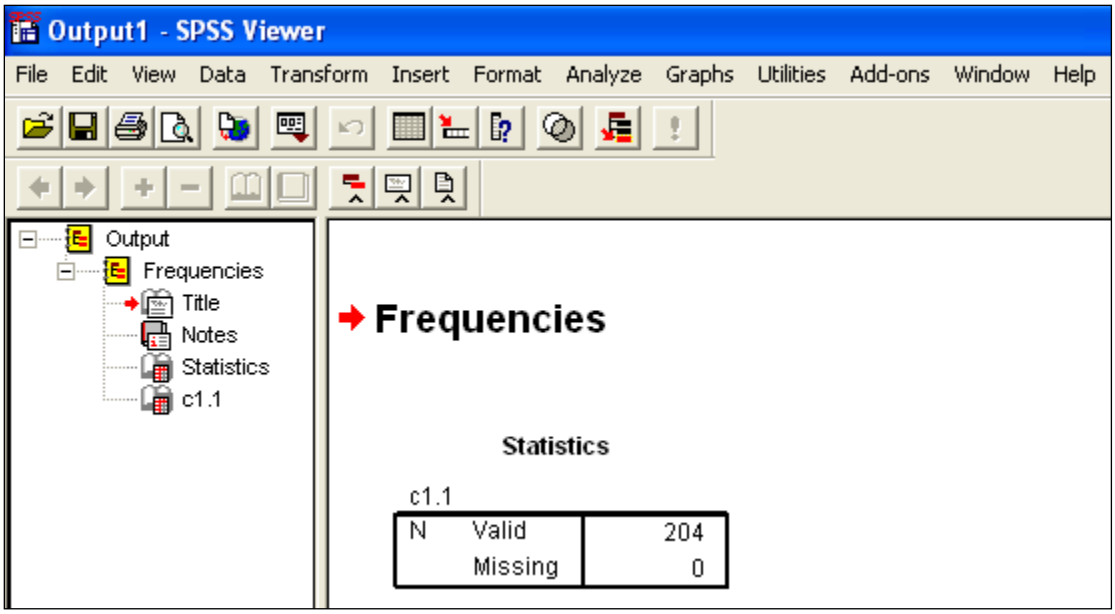
Hình 2.3. Màn hình quản lý biến (Variable view)



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

Các phép phân tích thống kê sẽ cho ra các kết quả như bảng biểu, đồi 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.

Hình 2.4. Màn hình hiển thị kết quả (Output)



2.2.4. 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.

BÀI 3: MÃ HÓA VÀ TẠO ĐỊNH NGHĨA BIẾN CHO DỮ LIỆU

1. Mã hóa dữ liệu

1.1. 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 nhàn 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 bảng câu hỏi

1. 2. Mã hoá dữ liệu

Mã hóa dữ liệu giúp cho việc lưu trữ, quản lý và phân tích số liệu được thuận tiện đồng thời giảm bớt sai sót trong quá trình nhập liệu. Về thực chất mã hóa 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. Thông thường, chúng ta mã hóa những phương án trả lời của khách thể từ dạng chuỗi (ký tự) thành dạng số.

Ví dụ:		
Giới tính:	*nam = 1;*	*Nữ = 2*
Trình độ học vấn:	*Tiểu học = 1:*	*THCS = 2; THPT = 3*
	ĐH/CĐ = 4	*SĐH = 5*
Nghề nghiệp:	*Lực lượng vũ trang*	*= 1*
	Công nhân lao động	*= 2*
	Viên chức	*= 3*
	Buôn bán	*= 4*
	.....	

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.

Ví dụ: Tạo mã code cho câu 1, câu 2, phiếu điều tra tâm lý chọn nghề của học sinh (Phiếu BT1)

Câu 1: Theo em, việc chọn nghề được xem là? (chọn 1 phương án trả lời)

1. Là một công việc quan trọng, cần có những suy nghĩ, cân nhắc chín chắn.:	**Mã code = 1**
2. Là một công việc quan trọng như nhiều việc quan trọng khác như học tập, giao tiếp	**Mã code = 2**
3. Là một công việc buộc phải nghĩ tới khi sắp phải ra trường.	**Mã code = 3**
4. Là một công việc bình thường như bao công việc bình thường khác	**Mã code = 4**

Câu 2: Theo em, học sinh nên tính tới việc lựa chọn nghề nghiệp từ thời điểm nào là thích hợp? (chọn 1 phương án trả lời)

1. Từ trước năm lớp 12	**Mã code = 1**
2. Từ đầu năm lớp 12.	**Mã code = 2**
3. Trước khi làm hồ sơ tuyển sinh.	**Mã code = 3**

Nguyên tắc cần thực hiện khi tạo mã code

- Cần thống nhất cách đặt mã code theo nguyên tắc: Từ nhỏ đến lớn; Từ trái qua phải; Từ trên xuống dưới.
- Nhiều người nhập dữ liệu nhưng phải thống nhất trước bảng mã code trước khi làm.
- Với câu hỏi mở, người nhập dữ liệu cần đọc trước các câu trả lời, tim ra các câu trả lời có cùng 1 ý (có thể khác nhau về diễn đạt), sau đó đưa chúng vào cùng 1 nhóm. Khi nhập sẽ nhập ý kiến chung của nhóm này.
- Với câu hỏi định lượng, không cần phải mã hóa, khi nhập dữ liệu sẽ nhập nguyên giá trị số liệu mà khách thể cung cấp. Trong một số trường hợp cụ thể, nếu giá trị định lượng mà khách thể cung cấp khác nhau về đơn vị tính, nhà nghiên cứu cần quy đổi chúng ra một đơn vị tính duy nhất để nhập.

Ví dụ: Câu 1. Anh chị vui lòng cho biết 1 ngày, anh chị dành ra bao nhiêu thời gian để đọc sách:……
Trong trường hợp này, người trả lời có thể điền thông tin là phút, hoặc giờ…; do đó khi nhập liệu cần quy đổi những câu trả lời theo đơn vị tình "giờ" ra đơn vị tính là "phút" hoặc ngược lại.

2. Tạo định nghĩa cho biến

Tạo 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 tạo 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. Qui trình định biến này bao gồm các bước sau:

Bước 1: Mở cửa số Variable view

Để tạo/định nghĩa biến, trước hết => mở cửa sổ Variable view. Trong cửa sổ variable view, có những cột sau:

1. Tên biến {Name}
2. Loại dữ liệu {Type}
3. Số lượng con số hoặc chữ {With}
4. Số lượng chữ số thập phân {Decimals}
5. Mô tả biến/nhãn biến {Lable} và nhãn trị số biến {Values}
6. Các trị số khuyết thiếu do người sử dụng thiết lập {Missing}
7. Độ rộng của cột {Width}
8. Căn lề {Align}
9. Thang đo {Measure}

(*xem hình 2.3*)

Bước 2: Đăt tên biến

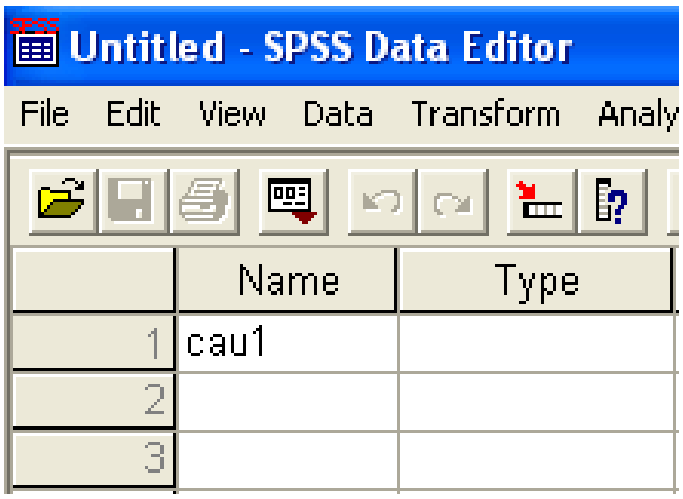
Ta gõ tên biến cần khai báo vào cột đầu tiên trong màn hình Variables view (Nếu ta không gõ tên biến vào thì SPSS sẽ mặc định tên biến này là **Var000001)**. Tên biến được khai báo này sẽ hiển thị trên đầu các cột trong màn hình Data view. Tên biến bị hạn chế về số ký tự hiển thị, do đó cần thiết phải khai báo ngắn gọn và dễ gợi nhớ, thông thường nên đặt theo thứ tự câu hỏi trong bảng câu hỏi như cau1, cau2, cau3, …Có một số qui ước sau đây phải tuân theo khi khai báo tên biến:

Các qui tắc dưới đây được áp dụng cho tên biến:

- Tên phải bắt đầu bằng một chữ. Các ký tự còn lại có thể là bất kỳ chữ nào, bất kỳ số nào, hoặc các biểu tượng như @, #, _, hoặc $.
- Tên biến không được kết thúc bằng một dấu chấm.

- Tránh dùng các tên biến mà kết thúc với một dấu gạch dưới cần (để tránh xung đột với các biến được tự động lập bởi một vài thủ tục)
- Độ dài của tên biến không vượt quá 8 ký tự.
- Dấu cách và các ký tự đặc biệt (ví dụ như !, ?, ‘, và *) không được sử dụng
- Từng tên biến phải đơn chiếc/duy nhất; không được phép trùng lặp. Không được dùng chữ hoa để đặt tên biến. Các tên NEWVAR, NewVar, và newvar được xem là giống nhau.
- □ Các từ khóa sau đây không được dùng làm tên biến: ALL, NE, EQ, TO, LE, LT, BY OR, GT, AND, NOT, GE, WITH

Hình 3.1. Đặt tên cho biến



Bước 3: Xác định kiểu biến (Type)

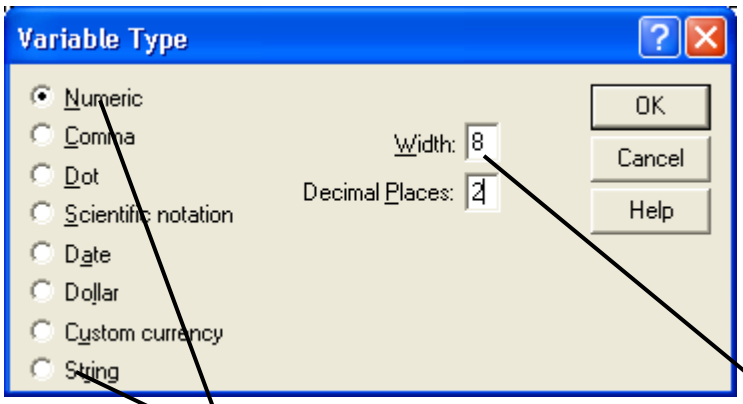
Định ra kiểu biến (Type): Có các dạng biến sau có thể định dạng. Dạng con số (numeric); Dạng tiền tệ; dạng ngày (Date) hoặc dạng chuỗi (String). Ngoài ra phần này cũng cho phép ta định dạng các dạng số được hiển thị khác nhau (*Xem hình 3.2*)

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 dầu phẩy trong ô Decimal.

Hình 3.2: Hộp thoại Variable Type



Lưu ý:

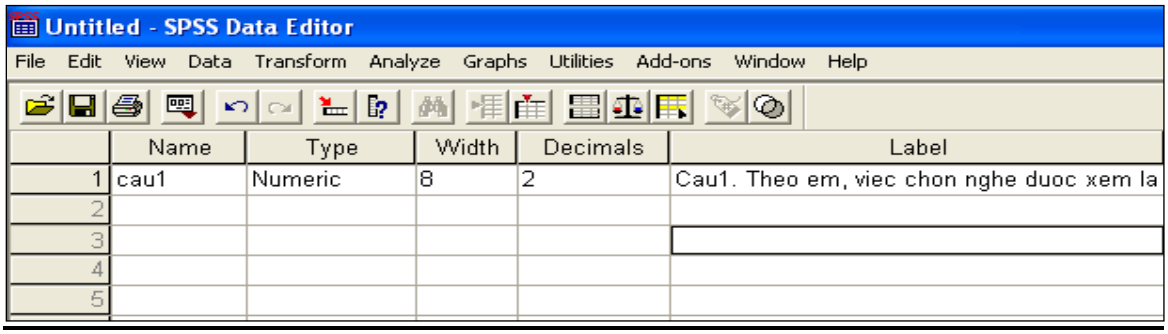
- *Với các câu hỏi đóng (định danh, thứ bậc, định lượng…) chúng ta chọn* ***"Numeric"***
- *Với câu hỏi mở, chúng ta chọn* ***" String"*** *và độ rộng cột* ***Width*** *với giá trị tối đa là 255*

Bước 4: Tạo nhãn biến {Variable Labels}

Tạo 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.

Ở phần này, chúng ta có thể sao chép toàn bộ nội dung câu hỏi hoặc chọn lọc ý chính của câu hỏi để điền vào phần này.

Hình 3.3. Tạo nhãn biến (Label)



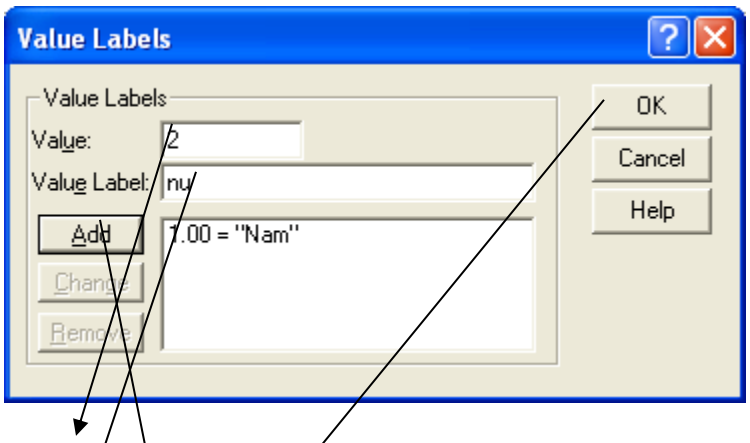
Bước 5: Tạo giá trị cho biến {Value Labels}

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 3-4):

Gán nhãn của giá trị (value lables) có ba thao tác riêng biệt:

Thao tác 1: Gán một nhãn mới:

Hình 3.4. Tạo một nhãn mới



Bước 1: Nhập mã code

Bước 2: Nhập nhãn mã code

Bước 3: Nhấn Add để xác nhận

Bước 4: Nhấn OK để kết thúc

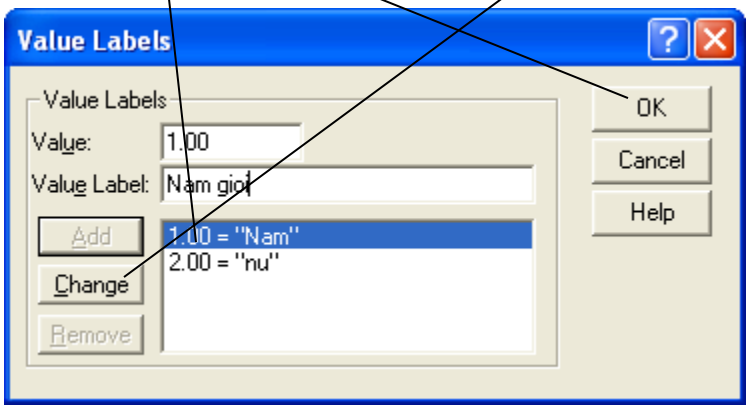
Thao tác 2: Sửa đổi một nhãn biến:

Bước 1: Di vệt sáng đến nhãn cần sửa đổi

Bước 2: Nhập tên nhãn mới, ấn nút Change để thay đổi

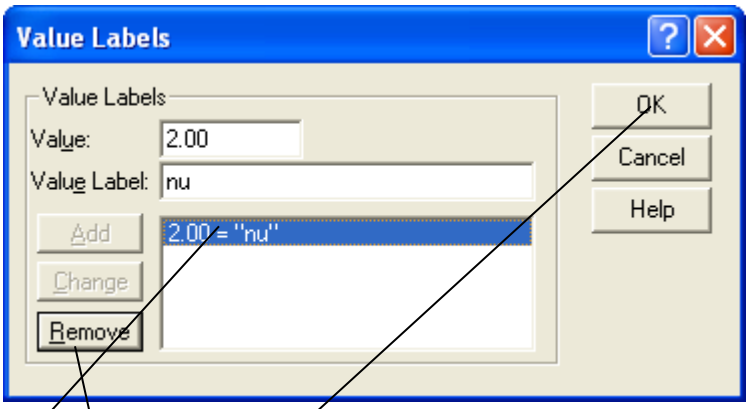
Bước 3: Nhấn OK để xác nhận

Hình 3.5. Sửa đổi giá trị nhãn biến



Thao tác 3: Loại bỏ một nhãn:

Hình 3.6. Loại bỏ một nhãn



Bước 1: Di vệt sáng đến nhãn cần loại bỏ

Bước 2: An nút Remove để loại bỏ

Bước 3. Nhấn OK để xác nhận

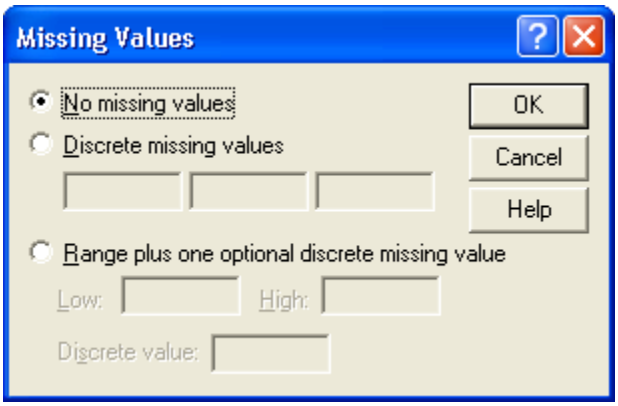
Bước 6: Tạo các giá trị khuyết thiếu {Missing Value}

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 thiếu.

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 3.6)

Hình 3-6: Hộp thoại Missing Values



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

(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 **(**Rang 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 trông ô định ra các giá trị khuyết riêng biệt.

Định kích cỡ cho cột (Colum 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 khái báo này chỉ mang tính chất quản lý không ảnh hưởng đến kết quả phân tích

BÀI 4. NHẬP DỮ LIỆU VÀ HIỆU ĐÍNH DỮ LIỆU CHO PHÂN TÍCH

1. *Nhập dữ liệu cho phần mềm SPSS*

1.1. Nhập dữ liệu cho các dạng câu hỏi đóng và biến số lượng

Nhập dữ liệu là khâu quan trọng, cần bảo đảm một bộ số liệu phản ánh các thông tin xác thực như trong phiếu điều tra

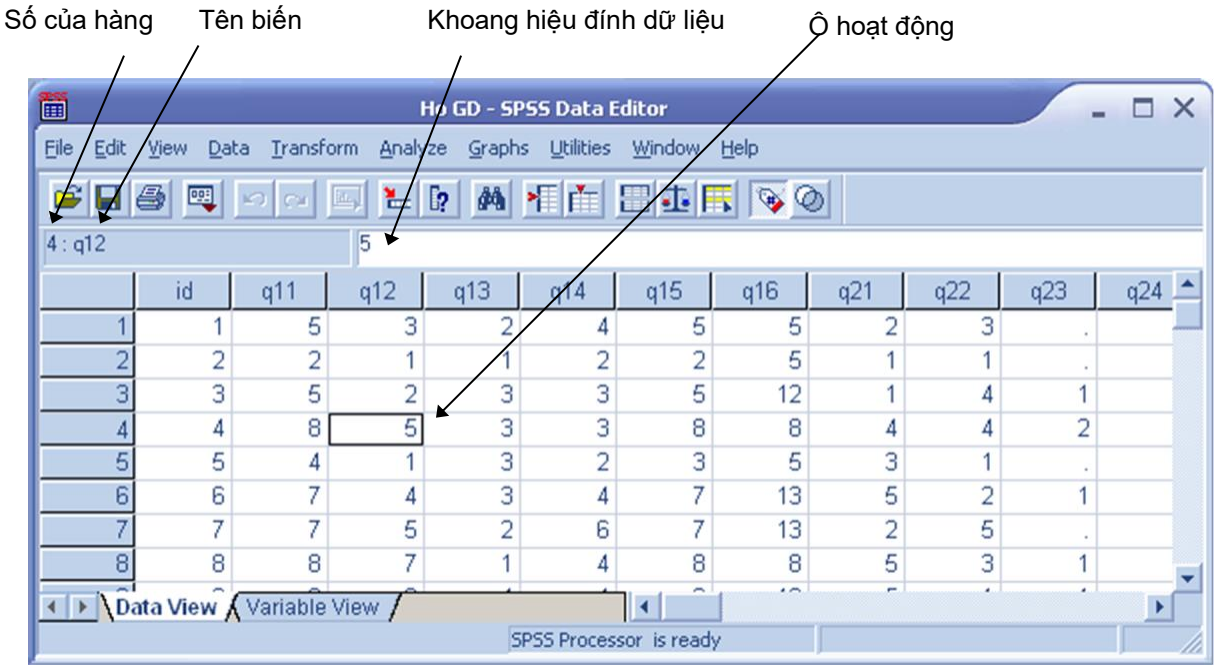
Sau khi đã tạo định nghĩa cho tệp dữ liệu, chúng ta cần nhập những thông tin thu thập qua bảng hỏi dưới dạng đã mã hóa bằng các trị số vào SPSS thông qua màn hình Data View trong cửa sổ Data Editor.

Chúng ta có thể nhập dữ liệu theo bất kỳ trật tự nào. Có thể nhập dữ liệu theo đối tượng hoặc theo biến, hoặc theo khu vực được chọn, hoặc theo từng ô

- Ô hoạt động (ô con trỏ) luôn được làm sáng
- Tên biến và số của hàng của ô hoạt động được thể hiện ở góc cao bên trái của cửa sổ Data Editor.
- Khi bạn chọn một ô và nhập một trị số thì nó sẽ được thể hiện ở khoang hiệu đính dữ liệu nằm ở trên của Data Editor
- Các trị số không được ghi cho đến khi bạn nhấn Enter hoặc chọn ô khác
- Để nhập bất kỳ gì khác một dữ liệu dạng số, trước hết phải định nghĩa loại dữ liệu.

Nếu bạn nhập một trị số vào một cột rỗng, Data Editor tự động tạo ra một biến mới và chỉ định một tên biến.

Hình 4.1: File dữ liệu làm việc trong Data View



- Chọn một ô trong bảng DataView
- Nhập trị số. Trị số này được thể hiện trong khoang hiệu đính dữ liệu ở đỉnh của Data Editor
- Nhấn Enter hoặc chọn một ô khác để ghi trị số này.

1.2. Nhập dữ liệu cho câu hỏi mở

Trong trường hợp chúng ta đã đặt kiểu biến (Type) dưới dạng String, chúng ta chỉ cần nhập câu trả lời nguyên văn hoặc đã rút gọn vào dòng tương ứng với biến đã được tạo từ trước.

Trong trường hợp chúng ta chưa đặt kiểu biến, chúng ta cần:

Nhắp đúp một tên biến ở đỉnh của cột trong bảng Data View hoặc nhắp bảng Variable View

Nhắp núm trong ô Type đối với biến này

Chọn loại dữ liệu trong hộp thoại Variable Type.

Nhắp OK

Nhắp đúp số của hàng hoặc nhắp bảng Data View

Nhập dữ liệu trong hàng đối với biến vừa mới được định nghĩa.

1.3. Những lưu ý tránh sai sót khi nhập dữ liệu.

1. Nhập toàn bộ số liệu hai lần bởi hai người riêng biệt.

2. Nhập toàn bộ số liệu hai lần do một người thực hiện

3. Nhập toàn bộ số liệu một lần, sau đó chọn ngẫu nhiên đơn khoảng 20% bộ số liệu và nhập lần 2. Nếu những sự khác nhau là tối thiểu, dừng lại. Nếu không cần phải cân nhắc đến phương án 2.

4. Nhập toàn bộ số liệu 1 lần, chọn ngẫu nhiên đơn khoảng 20% bộ số liệu, kiểm tra lại bằng mắt. Nếu những sự khác nhau là tối thiểu, dừng lại. Nếu không cần phải cân nhắc đến phương án 2

5. Nhập toàn bộ số liệu một lần, không kiểm tra hai lần. Không có đề nghị gì.

Lưu ý khác: Ngoài việc nhập dữ liệu trực tiếp vào bảng data view, chúng ta có thể nhập dữ liệu bằng phần mềm Excell hoặc một số phần mềm thông dụng khác.

2. Kiểm tra và hiệu đính dữ liệu trong bảng Data View

2.1. Kiểm tra và làm sạch dữ liệu

2.1.1. Các bước kiểm tra làm sạch dữ liệu

Xác định lỗi liên quan đến mã hóa số liệu: sai mã, trùng ID,..

Cách làm:

Cách 1: liệt kê các giá trị của biến - xem bảng phân bố tần số

- Sửa lỗi:
    - Căn cứ vào các thông tin khác
    - Xem lại phiếu gốc
    - Hỏi lại đối tượng được phỏng vấn

Cách 2: Kiểm tra các giá trị bất thường (giá trị quá bé hoặc quá lớn so với các giá trị khác).

Ví dụ: trong biến Tuổi có chứa giá trị 130 → đây có thể là một giá trị bất thường

Cách làm: liệt kê các giá trị của biến, vẽ biểu đồ

Trong một số trường hợp, người ta có thể loại bỏ các giá trị bất thường ra khỏi bộ số liệu, tuy nhiên cũng cần phải cân nhắc rất kỹ trước khi bỏ → *Tại sao?*

Khi thấy giá trị bất thường, cần kiểm tra lại phiếu gốc: nếu thực sự có giá trị đó thì chúng ta vẫn phải đưa nó vào trong các phân tích.

Cách 3: Kiểm tra các giá trị missing

"Missing" là những giá trị trống, biểu hiện bởi dấu "," trong cửa sổ Data view

Hai loại missing:

+ Có thông tin nhưng người nhập liệu lại không nhập vào hoặc người phỏng vấn quên không hỏi hoặc điền → lỗi mất thông tin

+ Thực sự là thông tin đó không có. (Nếu khách thể chưa lập gia đình thì biến số con sẽ không được hỏi)

+ Nếu Missing > 10% đối với mỗi biến, cần xem xét lại

Cách 3. Kiểm tra ***Lỗi nhập liệu:***

Lỗi này thường khó phát hiện nếu chỉ nhập số liệu 1 lần

Phát hiện và chữa những lỗi này: bằng cách Nhập kiểm tra.

Lý tưởng là có hai người độc lập nhập số liệu hai lần riêng rẽ, sau đó so sánh hai bộ số liệu với nhau

Cách 4:

– Kiểm tra tính đồng nhất của thông tin:
– Những câu trả lời không nhất quán cần được xác định và kiểm tra
– Không có một nguyên tắc chung nào cho việc xác định tính không nhất quán, cần phải tùy thuộc vào từng nghiên cứu

Ví dụ: giới tính là "nam" nhưng lại trả lời là "có" cho câu hỏi "đã từng đi khám thai chưa?"

– Tính ko nhất quán có thể do:

• Lỗi mã hóa số liệu hoặc lỗi đánh máy

- Bản thân người trả lời không nhất quán

– Khắc phục lỗi này:

- Căn cứ vào các câu trả lời khác, các thông tin khác
- Lần lại phiếu trả lời gốc để đối chiếu
- Hỏi lại đối tượng được phỏng vấn → khó thực hiện

2.1.2. Một số công cụ để tìm kiếm những giá trị bất thường

2.1.2.1. Kiểm tra bằng công cụ Explore

Công việc đầu tiên rất quan trọng và cần phải thực hiện một cách cẩn thận trước khi đi vào các bước mô tả hay các phân tích thông kê phức tạp sau này là tiến hành xem xét dữ liệu một cách cẩn thận. SPSS cung cấp cho công cụ Explore để xem xét và kiểm tra dữ liệu:

Phát hiện các sai sót

Nhận dạng dữ liệu để tìm phương pháp phân tích thích hợp và chuẩn bị cho việc kiểm tra giả thuyết

Để nhận dạng và phát hiện sai sót trong dữ liệu, ta có ba cách hiển thị dữ liệu như sau

Biểu đồ Histogram

Sơ đồ cành và lá Stem-and-leaf plot

Sơ đồ hộp Boxplot

Để ước lượng các giả định được dùng cho việc kiểm nghiệm các giả thuyết, ta dùng các phép kiểm tra sau:

Kiểm tra levene: Kiểm tra tính đồng đều của phương sai

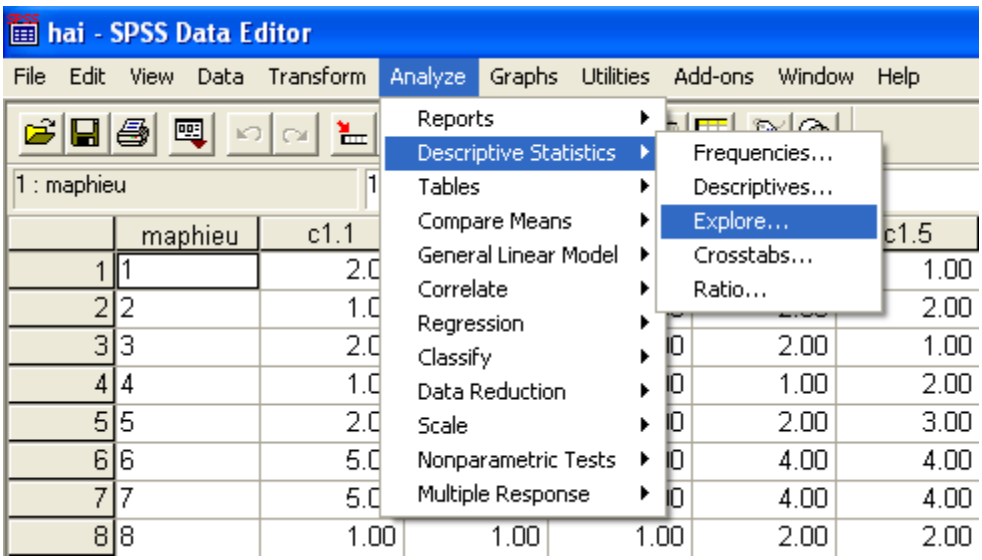
Kiểm tra K-S Lilliefors: Kiểm tra tính chuẩn tắc của tổng thể, xem dữ liệu có được lấy từ một phân bố chuẩn hay không.

Chúng ta thường dùng giá trị trung bình số học để ước lượng độ hội tụ của dữ liệu. Tuy nhiên vì giá trị trung bình bị ảnh hưởng bởi tất cả các giá trị quan sát. Để giảm thiểu những ảnh hưởng của các giá trị bất thường (quá lớn hoặc quá bé), người ta thường loại bỏ các giá trị lớn nhất và các giá trị nhỏ

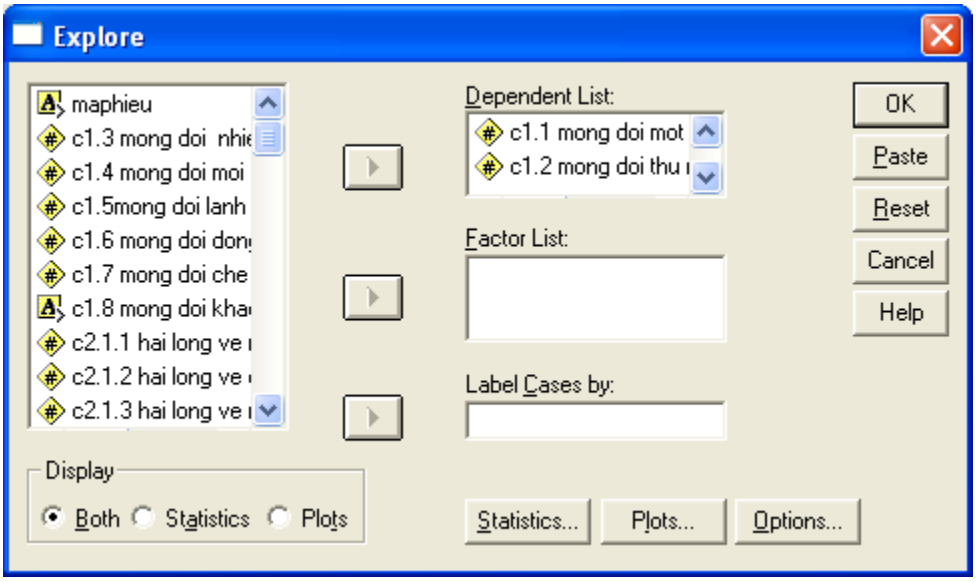
nhất (Outliers) theo cùng một tỷ lệ nào đó. Khi đó gia trị trung bình được gọi là giá trị trung bình giãn lược (Timmed-mean).

Một cách làm khác là gán các trọng số khác nhau cho các giá trị quan sát tùy theo khoảng cách của nó đến giá trị trung bình, càng xa trọng số càng nhỏ. Các trong số này gọi là M-estimators. Có 4 loại trọng số là Huber, Turkey, Hampel, và Andrew. Dựa vào trọng số này ta ước lượng lại giá trị trung bình cho dữ liệu.

Để kiểm tra dữ liệu, chọn trên menu **Analyze/Descriptive Statistic//Explore** để mở hộp thoại Explore như Hình 4.2:



Hình 4.3. Hộp thoại Expore



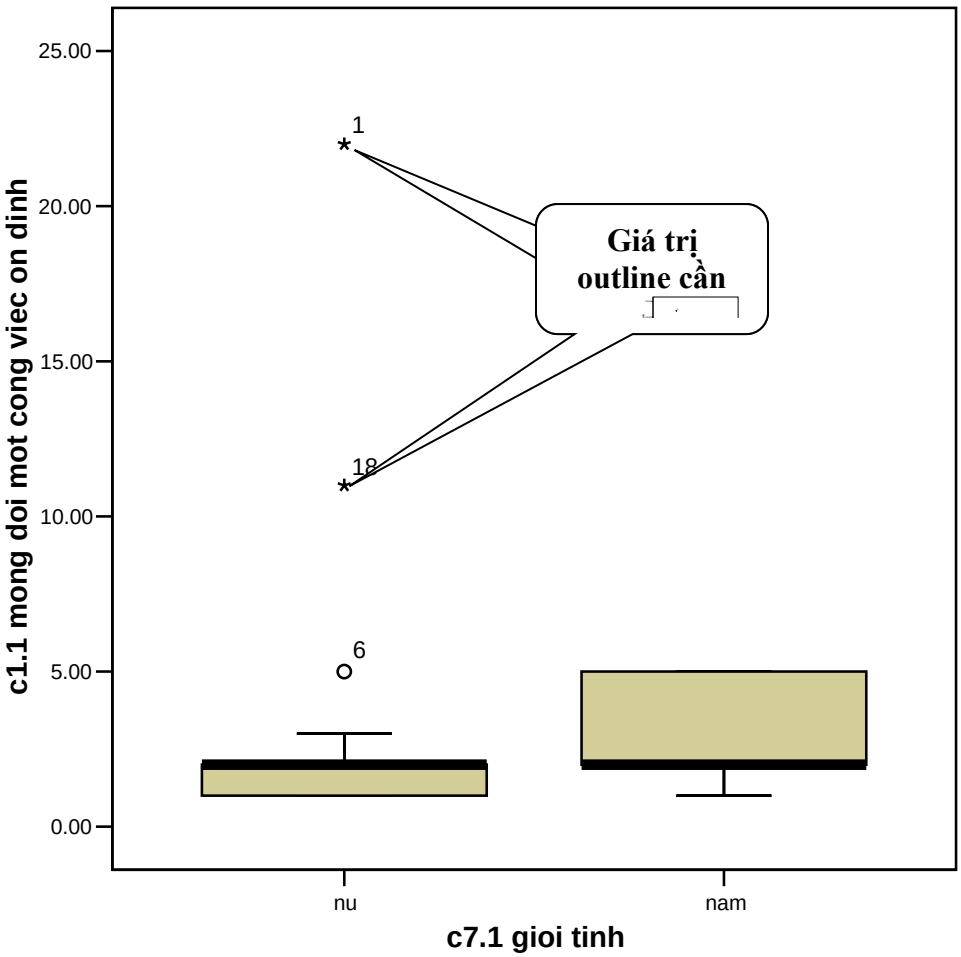
Các biến trong tập dữ liệu xuất hiện trong hộp bên trái. Chọn một hay nhiều biến đưa vào ô **Dependent list,** các biến cần quan sát sẽ được liệt kê trong ô này. Chúng ta cũng có thể tách các quan sát thành các nhóm nhỏ riêng biệt để kiểm tra dựa vào các giá trị của các biến kiểm soát sẽ được đưa vào ô **Factor List**. Ví dụ như kiểm tra biến mức độ đánh giá nói chung dựa vào biến nhãn hiệu đang sử dụng. Có thể lần ra các quan sát này bằng cách gán nhãn cho nó bằng gía trị của một biến nào đó, biến này sẽ được đưa vào trong ô **label cases by**.

Ví dụ muốn biết những giá trị di thường trong biến mức độ mong đợi về sự ổn định trong công việc giữa giới tính Nam và nữ. Ta gán nhãn cho các quan sát này bằng các giá trị trong biến số bảng câu hỏi. Lúc này nếu có các giá trị dị thường ta dễ dàng lần ra nó bằng số bảng câu hỏi kèm theo.

Ô **Display**, cho phép chúng ta chọn cách hiển thị kết quả, các tham sô thống kê **(Statistic)**, hoặc đồ thị **(Plot)**, SPSS mặc định là hiển thị cả hai

Sử dụng công cụ Statistics cho phép ta lựa chọn các thống kê hiển thị như hộp thoại Hình 4.4:

Hình 4.4.



Biểu đồ cành lá (Stem-and-Leaf Plots)

```
c1.1 mong doi mot cong viec on dinh Stem-and-Leaf Plot for
c7.1= nu

 Frequency    Stem &  Leaf

    12.00        1 .  000000000000
      .00        1 .
    27.00        2 .  000000000000000000000000000
      .00        2 .
     3.00        3 .  000
     3.00 Extremes    (>=5.0)

 Stem width:      1.00
 Each leaf:       1 case(s)
```

Sử dụng công cụ **Plots** (Hình 6-3), để lựa chọn hiển thị dạng đồ thị **(Histogram)**, biểu đồ chỉnh tắc, các phép kiểm tra về phân phối chuẩn, tính đồng đều của phương sai.

Boxplots: Điều kiện để hiển thị của Boxplots là ta phải đang quan sát nhiều hơn một biến phụ thuộc (hiển thị trong ô dependent list).

- **Factor levels together** đưa ra một hiển thị riêng biệt cho mỗi biến phụ thuộc. Trong phạm vi một hiển thị, Boxplots được hiển thị cho mỗi một nhóm được phân ra theo giá trị của biến điều khiển (factor variable). Dependents together đưa ra một hiển thị riêng biệt theo mỗi nhóm được phân theo các giá trị trong biến điều khiển. Trong phạm vi của hiển thị, boxplots được đưa ra lần lượt cho mỗi biến phụ thuộc.
- **Descriptive:** Cho phép lựa chọn hiển thị dạng đồ thị Histogram hay dạng cành lá (stem-and-leaf plots)
- **Normality plots with tests**. Đưa ra các dạng đồ thị về phân phối chuẩn. Đồng thời cung cấp một kiểm nghiệm thống kê Kolmogorov-Smirnov statistic, với mức tin cậy Lilliefors dùng để kiểm nghiện tính chuẩn của phân phối mẫu đang quan sát. Một kiểm nghiệm khác là thống kê Shapiro-Wilk được sử dụng cho mẫu có kích cỡ nhỏ hơn hoặc bằng 50 mẫu.
- **Spread vs. Level with Levene Test**. Cho phép chúng ta kiểm tra tính đồng đều của phương sai giữa các mẫu trong dữ liệu gốc hay dữ liệu đã được biến đổi. Để thực hiện phép thống kê Levene đòi hỏi phải có khai báo biến điều khiển trong khuôn Factor lists, Thông thường ta thường làm việc trên dữ liệu gốc do đó lựa chọn Untransformed trong khung Spread vs Level with Levene test.

 Kiểm nghiệm Kolmogorov-Smirnov (Lilliefors)

Kiểm nghiệm Lilliefors là một dạng kiểm nghiệm Kolmogorov-Smirnov, dùng để kiểm nghiệm tính chuẩn tắc của một mẫu hay hai mẫu. Với giá trị sig. nhỏ hơn mức ý nghĩa (0.05) là kết quả bác bỏ giả thuyết phân phối mẫu là phân phối chuẩn. Phép kiểm nghiệp Shapiro-Wilk chỉ dùng trong những trường hợp số mẫu nhỏ hơn 40.

 Kiểm nghiệm Levene

Trước khi đi vào các kiểm nghiệm trung bình ta cần phải tham khảo một kiểm nghiệm khác mà kết quả của nó là rất quan trọng cho các kiểm nghiệm trung bình sau này. Kiểm nghiệm Levene là phép kiểm nghiệm tính đồng nhất của phương sai. Ở đây ta kiểm nghiệm giả thuyết cho rằng phương sai của giữa các mẫu quan sát là bằng nhau. Kiểm nghiệm cho ta kết quả Sig. nhỏ hơn mức tin cậy (5%) ta kết luận không chấp nhận giả thuyết cho rằng phương sai mẫu thì bằng nhau. Chú ý trong một số kiểm nghiệm như ANOVA, kiểm nghiệm t, … Đòi hỏi phải kiểm nghiệm thông kê Levene trước để xác định tinh cân bằng hay không cân bằng của các phương sai mẫu. Kết quả này sẽ ảnh hưởng đến việc lựa chọn các kiểm nghiệm trung bình khác (Kiểm nghiệp trung bình với phương sai mẫu bằng nhau hoặc kiểm nghiệm trung bình với phương sai mẫu không bằng nhau).

2.1.2.2. Kiểm tra dữ liệu bằng bảng phân bố tần suất cho biến một trả lời (Frequencies)

Công cụ Frequencies sử dụng các tham số thống kê để mô tả cho nhiều loại biến, đây cũng là một công cụ hữu ích để ta khảo sát dữ liệu tìm lỗi cho dữ liệu.

Lập bảng này ngoài việc tóm tắt dữ liệu, nó còn giúp ta phát hiện những sai sót trong dữ liệu như, những giá trị bất thường (quá lớn hay quá nhỏ) có thể làm sai lệch kết quả phân tích thống kê, những giá trị mã hóa bất thường do sai sót việc nhập liệu hay mã hóa

Để tiến hành lập bảng đơn ta chọn công cụ **Analyze/Descriptive Statistic /frequencies** ta có hộp thoại như Hình 4.5

Chuyển biến cần mô tả sang hộp thoại variables, ta có thể lựa chọn nhiều biến cần quan sát cùng một lúc.

Công cụ Charts được dùng để vẽ đồ thị cho dữ liệu, và công cụ Format được sử dụng định ra kiểu hiển thị của dữ liệu, theo thứ tự tăng dần hoặc giãm dần.

Công cụ statistics để truy suất hộp thoại như Hình 4.5. Trong hộp thoại statistics này sẽ bao gồm các công cụ để đo lường các giá trị thống kê của dữ liệu như vị trí tương đối của các nhóm giá trị hay còn gọi là các phân vị, mật độ tập trung và phân tán của dữ liệu, những đặc tính về phân phối của dữ liệu (Distribution)

- **Giá trị bách phân vị (percentile values)**: Được dùng để xác định các ranh giới tương đối của các nhóm từ mẫu quan sát, điều lưu ý là dữ liệu cần quan sát đã được xắp xếp thep thứ tự từ thấp đến cao.

Ta có công cụ phân nhánh dữ liệu thành 4 phần bằng nhau gọi là tứ phân vị (quartiles).

Hoặc ta có thể chia dữ liệu theo các phần bằng nhau cụ thể bằng cách gõ số phần muốn chia vào công cụ cuts points for equal groups.

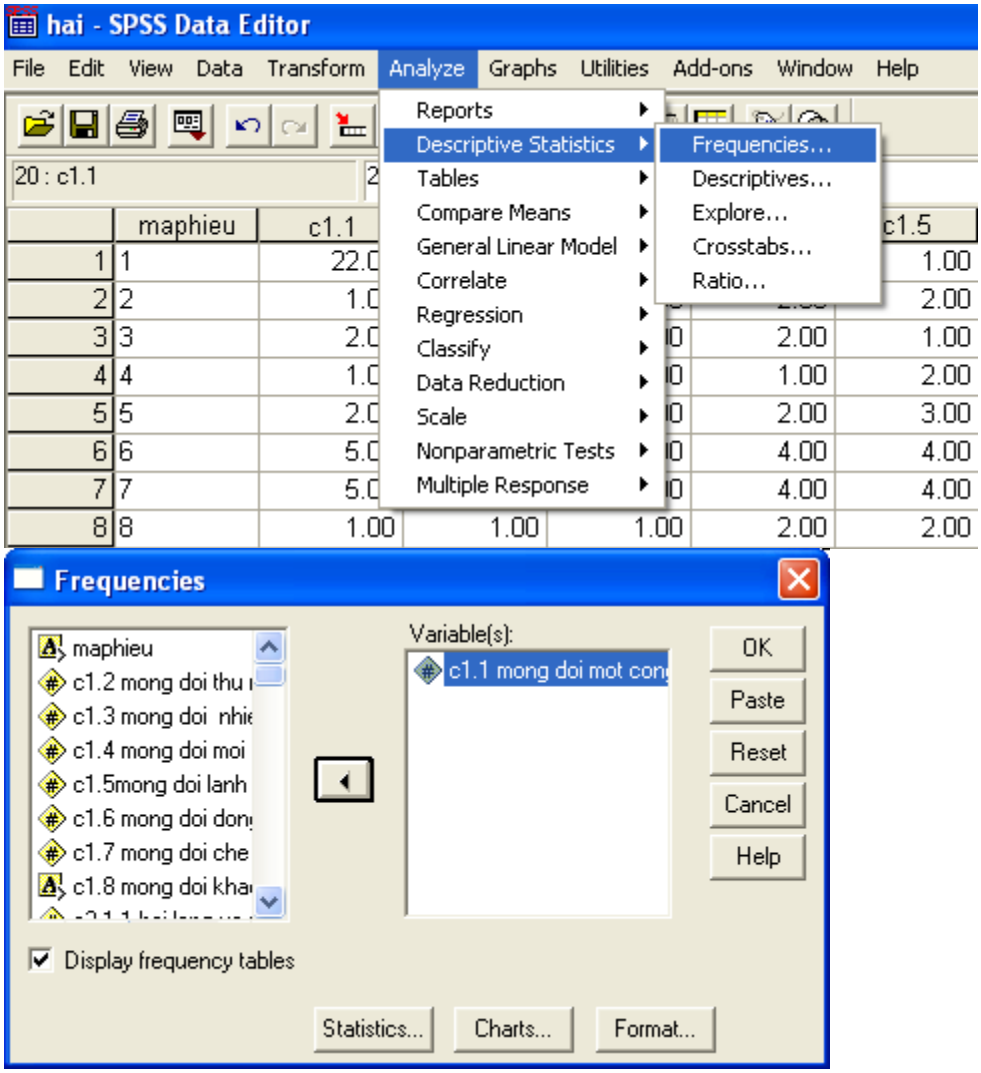
Hoặc ta có thể xem giá trị ở phân nhánh cụ thể nào đó từ công cụ percentile(s).

Sử dụng thanh Add để xác nhận số thứ tự phân vị cần quan sát, sử dụng thanh Remove và Change để loại bỏ hoặc thay đổi sự xác nhận ban đầu.

Ví dụ: Tìm giá trị outline của câu 1.1. Mức độ mong đợi về công việc của người lao động.

Bước 1. Chọn lệnh **Analyze/Descriptive/Frequenies**

Hình 4.5. Hộp thoại Frequencies



Bước 2: Chọn biến c.1.1. mong doi…. bên biến nguồn, dùng nút chuyển, chuyển sang biến đích.

Bước 3. Nhấn OK để kết thúc

Chúng ta có bảng phân phối tần suất sau:

c1.1 mong doi mot cong viec on dinh

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	rat mong doi	13	26.0	26.0	26.0
	mong doi	29	58.0	58.0	84.0
	binh thuong	3	6.0	6.0	90.0
	it mong doi	3	6.0	6.0	96.0
	11.00	1	2.0	2.0	98.0
	22.00	1	2.0	2.0	100.0
	Total	50	100.0	100.0	

Nhìn vào bảng chúng ta thấy ở cột Valid xuất hiện 2 giá trị không gán mã là 11, và 22. Những giá trị này là những giá trị outline, cần phải loại bỏ.

2.1.2.3. Lập bảng mô tả (Descriptive)

Sử dụng **Analyze/Descriptives Statisticts\Descriptives** để mở hộp thoại mô tả thống kê. Đây là một dạng công cụ khác có thể được dùng để tóm tắt dữ liệu và chỉ cho phép thao tác trên dạng dữ liệu định lượng (thang đo khoảng cách và tỷ lệ). Được dùng để thể hiện xu hướng tập trung của dữ liệu (central tendency) thông qua giá trị trung bình của các giá trị trong biến (mean), và mô tả sự phân tán của dữ liệu thông qua phương sai và độ lệch chuẩn. Chuyển các biến cần tóm tắt vào hộp thoại variables và nhấp thanh options để lựa chọn các thông số thống kê cần mô tả, như giá trị trung bình–mean, giá trị tối thiểu, giá trị tối đa, phương sai và độ lệch chuẩn,…

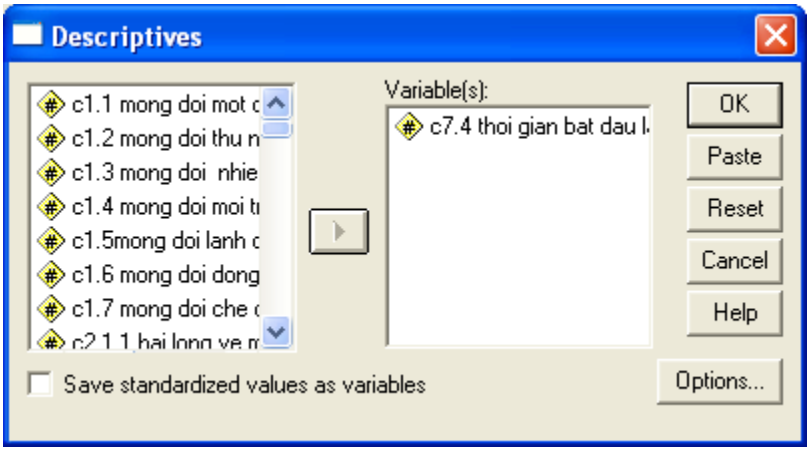
Ví dụ: Tìm giá trị bất thường trong biến 7.4. Thời gian làm việc tại công ty.

Bước 1: **Analyze/Descriptives Statisticts\Descriptives**

Bước 2: Chọn biến c7.4 rồi chuyển từ biến nguồn sang biến đích, nhấn OK để xác nhận.

Bước 3. Đọc bảng **Descriptive Statistics**

Hình 4.6. Hộp thoại Desciptives



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
c7.4 thoi gian bat dau lam viec tai cong ty	50	2	44	3.96	6.612
Valid N (listwise)	50				

Nhìn vào mục Maximum, ta thấy xuất hiện giá trị 44, đây là giá trị bất thường vì hiếm có ai làm việc tại công ty đã 44 năm. Hơn nữa với Mean (điểm trung bình) 3.96 nhỏ hơn độ lệch chuẩn (Std.Deviation) 6.612 cho thấy độ phân tán là rất lớn của tập hợp này là quá lớn. Điều này cho thấy có nhiều giá trị lớn bất thường đã can thiệp làm tăng điểm trung bình lên. Do đó cần phải quay trở lại bảng Data view để sửa chữa.

3. Lập bảng nhiều chiều cho các biến một trả lời (Crosstabs)

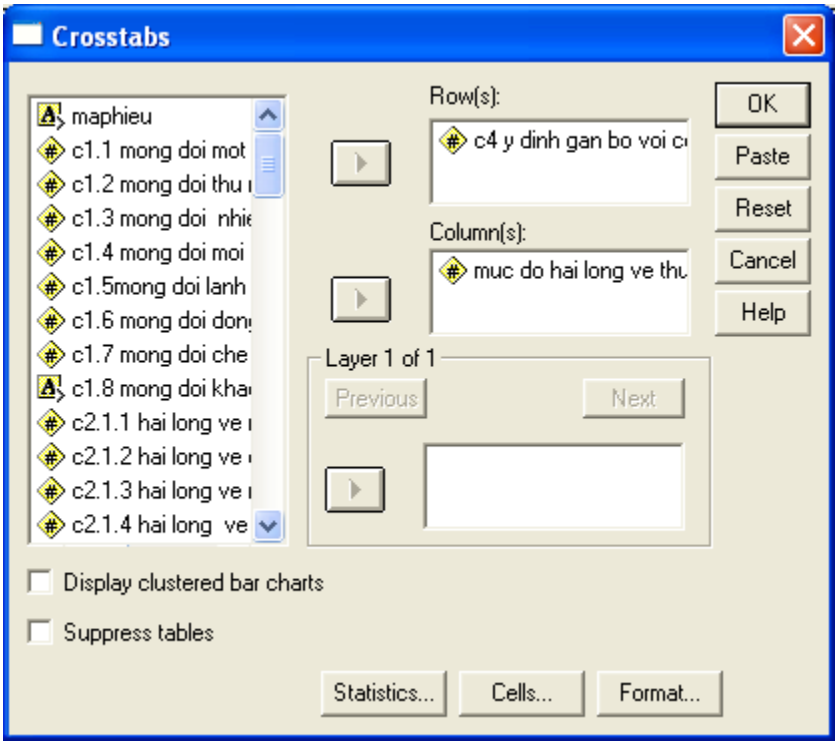
Bảng nhiều chiều là dạng bảng chéo thể hiện tần suất xuất hiện của một biến này trong mối quan hệ với một hay nhiều biến khác. Bảng chéo còn cung cấp nhiều loại kiểm nghiệm thống kê và đo lường mối quan hệ và tương quan giữa các biến trong bảng. Cấu trúc của bảng và loại dữ liệu (loại thang đó) sẽ quyết định loại công cụ nào được sử dụng để đo lường. Ngoài việc thể hiện mối liên hệ giữa các biến. Bảng nhiều chiều còn giúp ta phát hiện những sai sót trong dữ liệu từ việc phát hiện ra những mối quan hệ vô lý và bất thường giữa hai biến.

Ví dụ: Tìm kiếm liệu có người lao động nào không hài lòng về công ty nhưng lại muốn bỏ việc ngay lập tức.

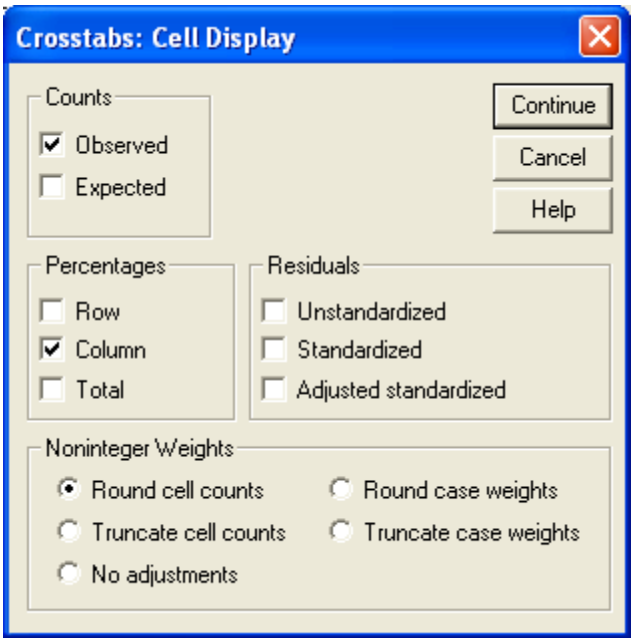
Bước 1: **Analyze/Descriptives Statisticts\Crosstabs**

Bước 2: Chọn biến c4, chuyển vào Row; chọn biến mức độ hài lòng cho vào Column. Bấm bào ô Cells xuất hiện hộp thoại Crosstabs: Cell Display. Tại đây, chọn **Percentages/Column ,** nhấn Continue để quay trở về bảng Crosstabs, nhấn OK để kết thúc.

Hình 4.7. Hộp thoại Crosstabs



4.7. Hộp thoại Crosstabls: Cell Display



Bước 3. Đọc bảng kết quả

c4 y dinh gan bo voi cong ty * muc do hai long ve thu nhap Crosstabulation

			muc do hai long ve thu nhap			
			hai long	nua hai long nua khong hai long	it hai long	Total
c4 y dinh gan bo voi cong ty	**nhanh chong roi bo neu co co hoi**	Count	**4**	4	0	8
		% within muc do hai long ve thu nhap	**20.0%**	16.0%	.0%	16.0%
	tinh toan ve nhung co hoi thay doi	Count	10	12	3	25
		% within muc do hai long ve thu nhap	50.0%	48.0%	60.0%	50.0%
	gan bo them mot thoi gian	Count	6	9	2	17
		% within muc do hai long ve thu nhap	30.0%	36.0%	40.0%	34.0%
Total		Count	20	25	5	50
		% within muc do hai long ve thu nhap	100.0%	100.0%	100.0%	100.0%

Nhìn bảng số liệu, tại cốt Mức độ hài lòng, chúng ta thấy có 20% công nhân hài lòng về công ty nhưng lại muốn nhanh chóng rời bỏ công ty. Điều này gợi cho chúng ta sự vô lý. Do đó cần quay lại phiếu điều tra gốc để kiểm tra lại thông tin cho chính xác.

Lưu ý: Chúng ta chỉ dùng bảng Crosstabl khi:

Dữ liệu sử dụng có phân phối chuẩn, hoặc kích cỡ mẫu phải đủ lớn (n>=30)

Không tồn tại tần suất mong muốn nào của bất kỳ giá trị nào trong bảng chéo nhỏ hơn 5.

2.2. Hiệu đính dữ liệu

Với Data Editor, bạn có thể hiệu đính trị số của dữ liệu trong bảng Data View theo nhiều cách. Bạn có thể:

- Thay đổi trị số của dữ liệu

- Cắt, sao chép, và dán các trị số của dữ liệu
- Thêm vào hoặc xoá các đối tượng
- Thêm vào hoặc xoá các biến
- Thay đổi trật tự của các biến

2.2.1. Thay thế hoặc hiệu đính một trị số của dữ liệu

Để xoá trị số cũ và nhập một trị số mới:

- Trong bảng Data View, nhắp đúp vào ô. Trị số được thể hiện trong khoang hiệu đính dữ liệu.
- Hiệu đính trị số trực tiếp từ ô hoặc trong khoang hiệu đính dữ liệu.
- Nhấn Enter (hoặc chuyển sang ô khác) để ghi trị số mới.

2.2.2. Cắt, sao chép và dán các trị số của dữ liệu

Bạn có thể cắt, sao chép và dán các trị số của từng ô hoặc một nhóm các trị số trong Data Editor. Bạn có thể:

- Chuyển hoặc sao chép trị số của một ô sang một ô khác.
- Chuyển hoặc sao chép trị số của một ô sang một nhóm các ô.
- Chuyển hoặc sao chép trị số của một đối tượng sang cho một nhóm các đối tượng.
- Chuyển hoặc sao chép trị số của một biến sang cho một nhóm các biến.
- Chuyển hoặc sao chép trị số của một nhóm các ô sang cho một nhóm các ô khác.

2.2. 3. Chèn thêm các đối tượng mới

Nhập dữ liệu vào một ô trong một hàng rỗng sẽ tự động tạo ra một đối tượng mới. Data Editor sẽ chèn các trị số khuyết thiếu đối với mọi biến khác cho đối tượng đó. Nếu có bất kể hàng rỗng nào nằm giữa đối tượng mới và các đối tượng đã có sẵn, các hàng rỗng đó cũng trở thành các đối tượng mới với các trị số khuyết thiếu hệ thống đối với mọi biến.

Bạn có thể chèn các đối tượng mới vào giữa các đối tượng đã có sẵn.

Để chèn một đối tượng mới giữa các đối tượng đã có sẵn

- Trong Data View, chọn bất kỳ ô nào trong đối tượng (hàng) nằm dưới vị trí nơi mà bạn muốn chèn đối tượng mới.
- Từ thanh menu chọn

Data/Insert Case

Một hàng mới được chèn vào và mọi mọi biến của đối tượng mới này đều nhận được trị số khuyết thiếu hệ thống.

2.2.4. Chèn một biến mới

Nhập dữ liệu vào một cột rỗng trong bảng Data View hoặc trong một hàng rỗng trong bảng Variable View sẽ tự động tạo ra một biến mới với một tên biến mặc định (tiền tố var và một chuỗi số tuần tự) và một định dạng dữ liệu mặc định (dạng số). Data Editor chèn trị số khuyết thiếu hệ thống cho mọi đối tượng đối với biến mới này. Nếu có bất kỳ cột rỗng nào trong bảng Data View hoặc hàng rỗng nào trong bảng Variable View giữa biến mới và các biến đã có sẵn, thì những cột này (trong bảng Data View) hoặc hàng này (trong bảng Variable View) cũng trở thành biến mới với trị số khuyết thiếu hệ thống cho mọi đối tượng.

Để chèn một biến mới giữa các biến đã có sẵn

- Chọn bất kỳ ô nào trong biến bên phải của (bảng Data View) hoặc dưới (của bảng Variable View) vị trí mà bạn muốn chèn biến mới vào.
- Từ thanh menu chọn

Data/Insert Variable

Một hàng mới được chèn vào với trị số khuyết thiếu hệ thống cho mọi đối tượng.

2.2.5. Để chuyển một biến trong Data Editor

Nếu bạn muốn đặt vị trí biến giữa hai biến đã có sẵn, hãy chèn một biến vào vị trí nơi bạn muốn di chuyển biến đến đó

- Đối với biến bạn muốn chuyển, nhắp tên biến ở đỉnh của cột trong bảng Data View hoặc số hàng trong bảng Variable View. Toàn bộ biến sẽ được làm nổi bật/tô sáng.
- Từ thanh menu chọn

  Edit/Cut
- Nhắp vào tên biến (trong bảng Data View) hoặc số hàng (trong bảng Variable View) nơi bạn muốn di chuyển biến đến. Toàn bộ biến này sẽ được mà nổi bật
- Từ thanh menu chọn

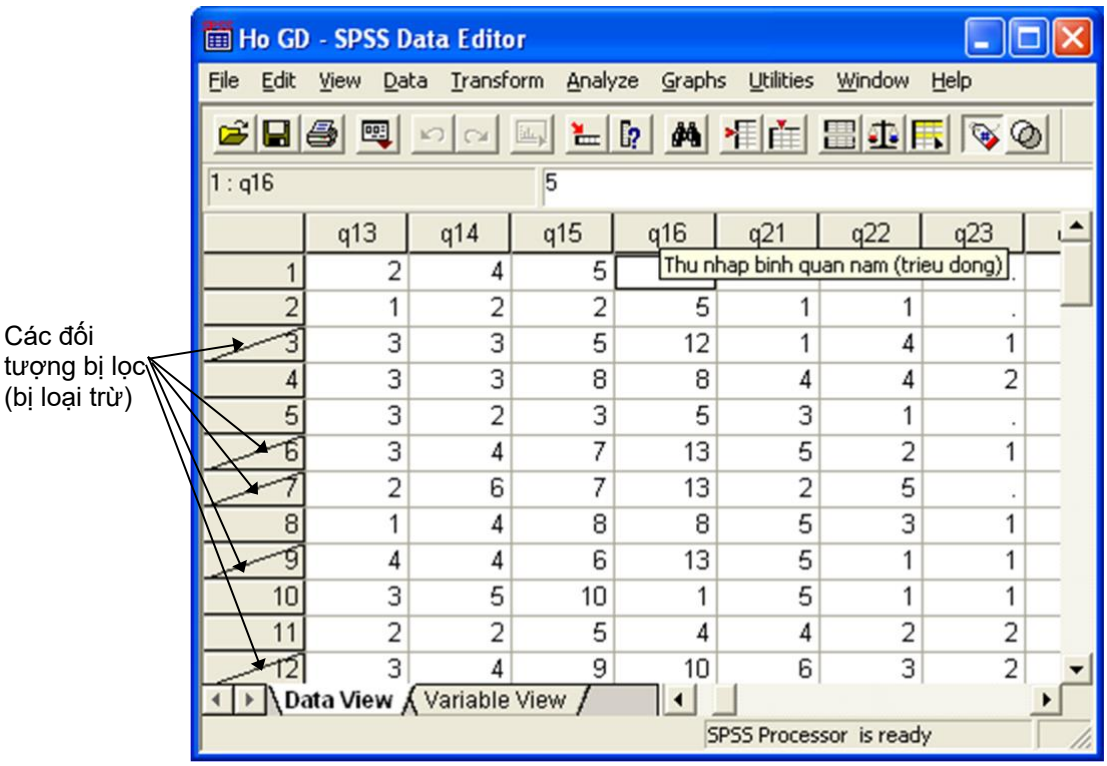
  Edit/Paste

2.2.6. Thay đổi loại dữ liệu

Bạn có thể thay đổi loại dữ liệu cho một biến bất kể lúc nào có sử dụng hộp thoại Variable Type trong bảng Variable View, và Data Editor sẽ cố gắng chuyển đổi các trị số hiện có sang loại mới. Nếu không thể chuyển đổi được thì trị số khuyết thiếu hệ thống sẽ được chỉ định. Các qui tắc chuyển đổi cũng giống như trường hợp dán trị số vào một biến có định dạng khác. Nếu sự thay đổi trong định dạng của dữ liệu có thể gây ra các đặc tả của trị số khuyết thiếu hoặc nhãn trị số, Data Editor thể hiện một hộp cảnh báo và hỏi nếu như bạn muốn tiếp tục với việc thay đổi hay huỷ bỏ nó.

7. Lọc đối tượng trong Data Editor theo yêu cầu

Hình 4.7: Các đối tượng được lọc trong Data Editor



	q13	q14	q15	q16	q21	q22	q23
1	2	4	5				.
2	1	2	2	5	1	1	.
3	3	3	5	12	1	4	1
4	3	3	8	8	4	4	2
5	3	2	3	5	3	1	.
6	3	4	7	13	5	2	1
7	2	6	7	13	2	5	.
8	1	4	8	8	5	3	1
9	4	4	6	13	5	1	1
10	3	5	10	1	5	1	1
11	2	2	5	4	4	2	2
12	3	4	9	10	6	3	2

Nếu bạn chọn một tập hợp phụ các đối tượng nhưng không loại bỏ những đối tượng không được chọn, những đối tượng không được chọn được đánh dấu trong Data Editor với một đoạn thẳng nằm chéo trong các ô số hàng.

BÀI 5. CÁC PHÉP BIẾN ĐỔI DỮ LIỆU

Trong một trường hợp lý tưởng, dữ liệu ban đầu (thô) của bạn là thích hợp hoàn toàn cho loại phân tích mà bạn muốn tiến hành, và mọi quan hệ giữa các biến là hoặc tuyến tính một cách thích hợp hoặc gần như trực giao. Rất đáng tiếc đây là trường hợp rất hiếm có. Các phân tích sơ bộ có thể bộc lộ các trình tự mã hoá bất tiện hoặc các sai số do mã hoá, hoặc biến đổi dữ liệu có thể bị đòi hỏi để bộ lộ mối quan hệ thực giữa các biến.

Bạn có thể thực hiện các phép biến đổi từ những nhiệm vụ đơn giản, chẳng hạn như thu nhỏ số nhóm/tổ để tiến hành phân tích, hoặc phức tạp hơn như tạo các biến mới dựa trên các phương trình phức tạp và các câu lệnh/khai báo có điều kiện

1.Biến đổi dữ liệu

1.1. Tính toán biến {Compute Variable}

Thủ tục Compute Variable tính toán các trị số của một biến được dựa trên sự biến đổi của một biến khác

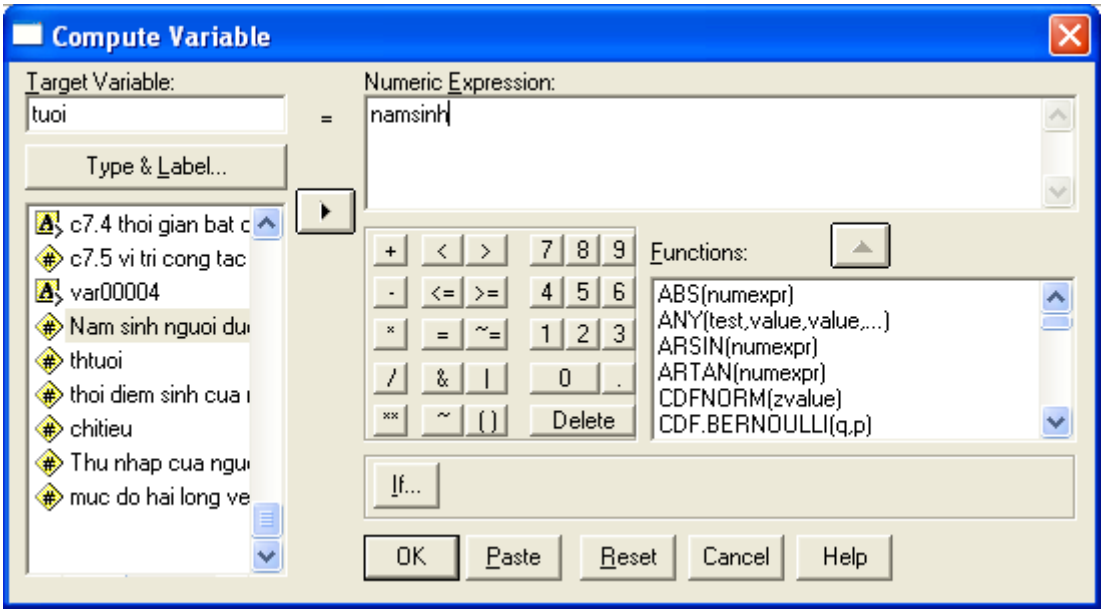
- Bạn có thể tính các trị số cho các biến dạng số hoặc dạng chuỗi (các ký tự chuỗi có dạng số)
- Bạn có thể lập các biến mới hoặc thay thế các trị số của biến đã có. Đối với biến mới, bạn cũng có thể chỉ định loại biến và nhãn biến.
- Bạn có thể tính toán các trị số một cách có chọn lọc đối với các tập hợp con của dữ liệu dựa trên các điều kiện lô-gic.
- Bạn có thể sử dụng trên 70 hàm lập sẵn {built-in}, bao gồm các hàm đại học, các hàm thống kê, các hàm phân bố và các hàm chuỗi.

Để tính toán biến

Ví dụ: Chúng ta muốn chuyển biến năm sinh thành biến tuổi, cách làm như sau:

- Bước 1: Từ thanh menu chọn **Transform/Compute…**

Hình 5.1. Hộp thoại Compute Variable



Bước 2: Đánh tên của biến đích {target variable}. Nó có thể là một biến đã có hoặc một biến mới sẽ được bổ sung vào file dữ liệu làm việc. Trong trường hợp này chúng ta nhập tên biến là **"tuoi".**

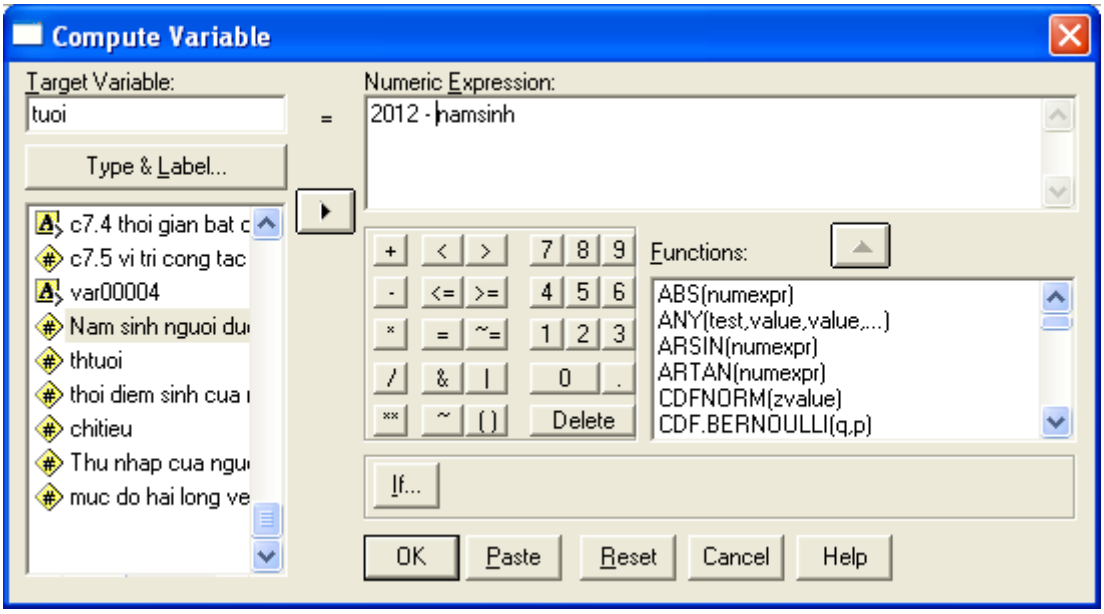
Bước 3: Xây dựng một biểu thức, hoặc dán các bộ phận vào Numeric Expression hoặc gõ trực tiếp vào đó.

- Dán các hàm từ danh sách các hàm {Functions} và nhập các tham số được biểu thị bằng các dấu hỏi
- Các hằng số dạng chuỗi phải được để trong dấu mở đóng ngoặc đơn hoặc ngoặc kép
- Các hằng số dạng số phải được nhập theo định dạng kiểu Hoa Kỳ với dấu chấm (.) là dấu thập phân.

Đối với biến dạng chuỗi mới, bạn còn phải chọn Type&Lable để xác định loại dữ liệu.

Trong trường hợp này, biểu thức hợp lý lấy biến năm hiện tại (2012) trừ đi biến năm sinh.

Hình 5.2. Hộp thoại Compute Variable



Lúc này, ở màn hình Data View xuất hiện 1 biến mới. biến "tuoi" ở cột cuối cùng.

Tính toán biến với tuỳ chọn If Cases

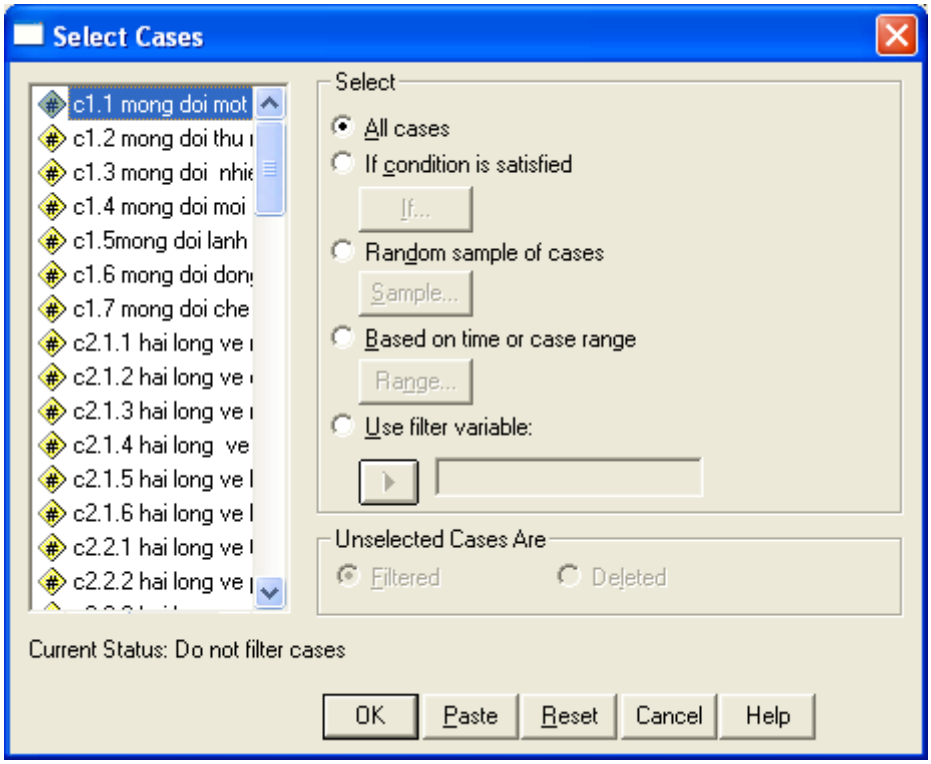
Hộp thoại If Cases cho phép bạn áp dụng phép chuyển đổi dữ liệu đối với các nhóm các đối tượng được chọn lọc, có sử dụng các biểu thức điều kiện. Một biểu thức điều kiện trả lại một trị số đúng hay sai hoặc khuyết thiếu cho từng đối tượng.

- Nếu kết quả của một biểu thức điều kiện là *true* {*đúng*}, phép biến đổi được áp dụng cho đối tượng
- Nếu kết quả của một biểu thức điều kiện là *false* {*sai*} hoặc *missing* {*khuyết thiếu*}, phép biến đổi không được áp dụng cho đối tượng
- Hầu hết các biểu thức điều kiện sử dụng một hoặc một số trong 6 dấu quan hệ (<, >, <= (nhỏ hơn và bằng), >= (bằng và lớn hơn), = và ~= (khác)) trên bảng tính toán.
- Các biểu thức điều kiện có thể bao hàm các tên biến, các hằng số, các phép toán số học, các hàm số và hàm khác, các biến lô-gíc và các thao tác có điều kiện khác

  Ví dụ: Chúng ta chỉ chọn nam giới trong số khách thể điều tra để nghiên cứu sâu hơn. Cách làm như sau:

Bước 1: Từ thanh Menu, chọn **Data/Select case**

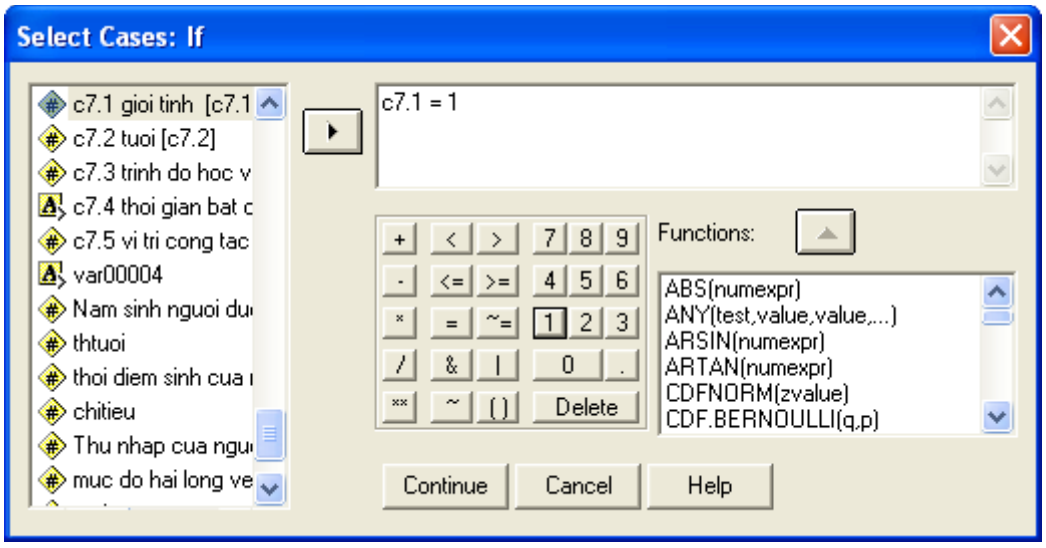
Hình 5.3.



Bước 2: Chọn điều kiện "if condition is satisfied"

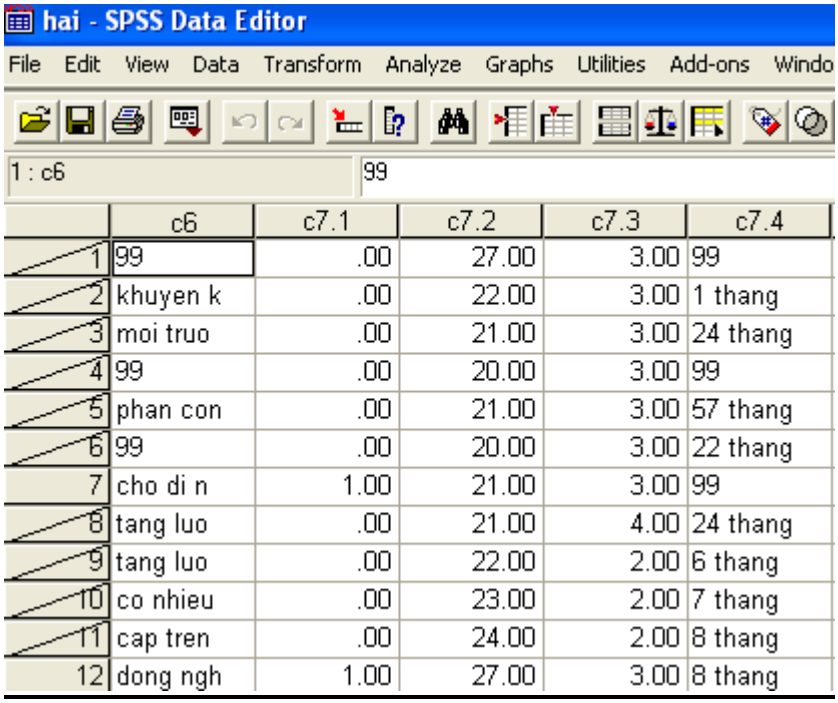
Chọn biến c7.1 (giới tính), đặt điều kiện = 1.

Hình 5.4



Bước 3: Nhấn Continue/ok để kết thúc

Hình 5.5



hai - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Windo

1 : c6 | 99

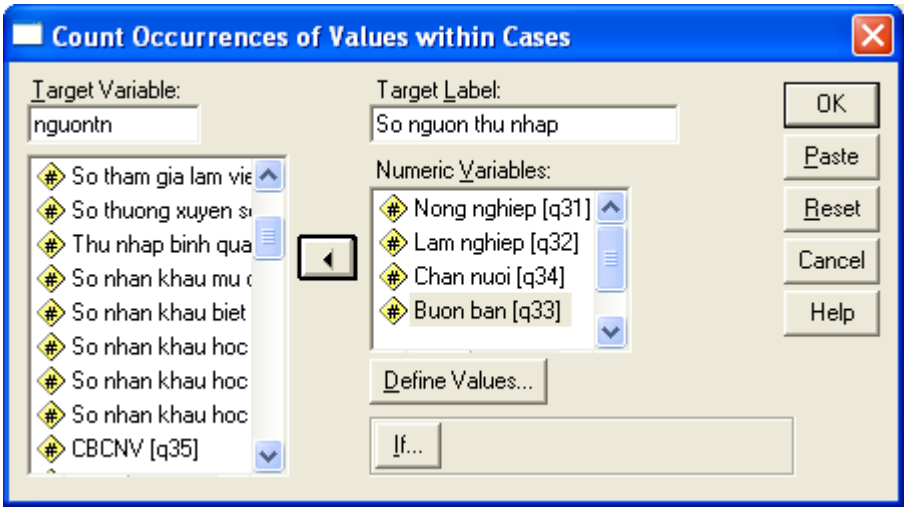
	c6	c7.1	c7.2	c7.3	c7.4
1	99	.00	27.00	3.00	99
2	khuyen k	.00	22.00	3.00	1 thang
3	moi truo	.00	21.00	3.00	24 thang
4	99	.00	20.00	3.00	99
5	phan con	.00	21.00	3.00	57 thang
6	99	.00	20.00	3.00	22 thang
7	cho di n	1.00	21.00	3.00	99
8	tang luo	.00	21.00	4.00	24 thang
9	tang luo	.00	22.00	2.00	6 thang
10	co nhieu	.00	23.00	2.00	7 thang
11	cap tren	.00	24.00	2.00	8 thang
12	dong ngh	1.00	27.00	3.00	8 thang

Lúc này, những trường hợp giới tính là nữ đã bị loại bỏ (bằng những gạch chéo). Ta bắt đầu có thể tính toán theo mục đích nghiên cứu của mình.

1.2. Đếm số lần xảy ra của các trị số trong các đối tượng

Hộp thoại này toạ nên một biến đếm số lần xảy ra của cùng trị số hoặc các trị số trong một danh sách các biến cho từng đối tượng. Ví dụ một cuộc điều tra có thể bao gồm một danh sách các tạp chí với hộp đánh dấu *có/không* để chỉ ra xem loại tạp chí nào mà từng đối tượng điều tra đọc. Bạn có thể đếm số câu trả lời *có* cho từng đối tượng điều tra để tạo ra một biến mới chứa đựng tổng số tạp chí được đọc.

Hình 5.6: Đếm số lần xảy ra của các trị số trong các đối tượng



Để đếm số lần các trị số xảy ra trong các đối tượng

- Từ thanh menu chọn

  Transform/Count…

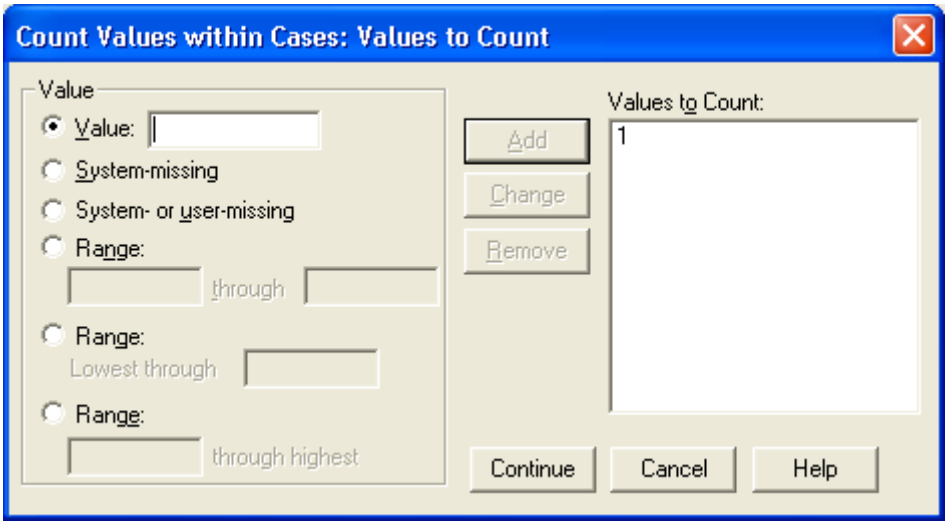
- Chọn một hay hơn một biến cùng loại (dạng số hoặc dạng chuỗi)
- Nhắp Define Variable và xác định loại trị số hoặc các trị số nào sẽ được đếm.
- Không bắt buộc, bạn có thể định nghĩa một tập hợp con các đối tượng để đếm số lần xảy ra của các trị số.

  Hộp thoại If Cases để xác định các tập hợp con giống như được mô tả trong phần Compute Variable.

1.3. Đếm các trị số trong các đối tượng: Các trị số cần đếm

Trị số của biến đích (trong hộp thoại chính) được tăng thêm 1 cho mỗi lần khi một trong những biến được lựa chọn thoả mãn một đặc tả trong Value to Count. Nếu một đối tượng thoả mãn một số mô tả đối với bất kỳ biến nào, biến đích được tăng một số lần tương ứng đối với biến đó. Các đặc tả về trị số có thể bao gồm các trị số riêng biệt, các trị số khuyết thiếu (hệ thống hoặc người sử dụng), và các phạm vi {range}. Các phạm vi bao gồm các điểm cuối của chúng và bất kỳ trị số khuyết thiếu của người sử dụng có độ lớn rơi vào trong phạm vi đó.

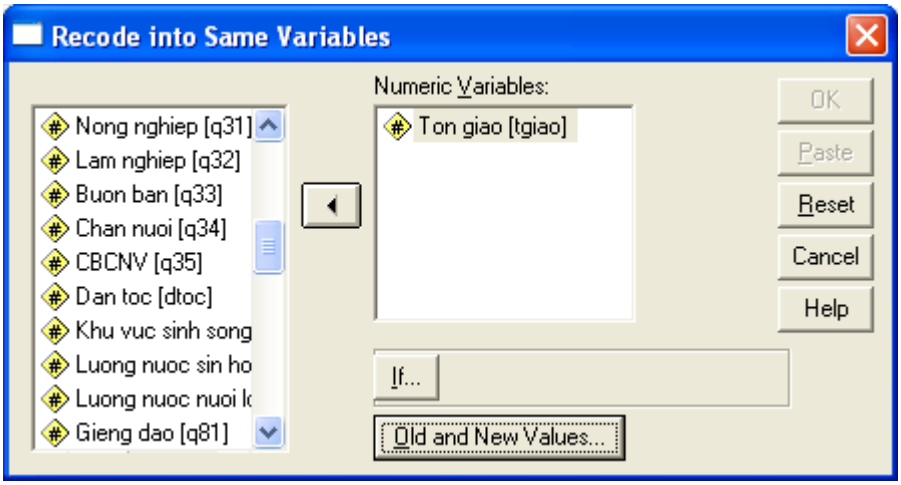
Hình 5.7: Hộp thoại các trị số cần đếm



2. Mã hoá lại dữ liệu

- Bạn có thể biến đổi trị số dữ liệu bằng cách mã hoá lại chúng
- Mã hoá lại dữ liệu ngay trong biến có sẵn (không tạo thành biến mới)
- Mã hoá lại dữ liệu ngay trong biến có sẵn {Recode into Same Variable} gán lại các trị số của biến đang có hoặc cắt giảm bớt các phạm vi của các trị số đang có vào các trị số mới
- Bạn có thể mã hoá các biến dạng số và dạng chuỗi. Nếu bạn chọn nhiều biến, chúng phải có cùng loại. Bạn không thể mã hoá các biến dạng chuỗi và dạng số cùng với nhau.

Hình 5.8: Hộp thoại Recode into Same Variables



2.1. Mã hoá lại dữ liệu ngay trong biến đã có sẵn

 Từ thanh menu chọn

Transform/Recode/Into Same Variables…

- Chọn các biến mà bạn muốn mã hoá, Nếu bạn chọn nhiều biến, chúng phải có cùng dạng (chuỗi hoặc số)
- Nhắp vào Old and New Values và định rõ cách mã hoá lại trị số.

Một cách tuỳ chọn, bạn có thể chọn một nhóm các đối tượng để mã hoá Hộp thoại If Cases để xác định một nhóm các đối tượng cũng giống như đã được mô tả trong mục tính toán biến {Compute Variable}

Hộp thoại Recode into Same Values: Old and NewValues

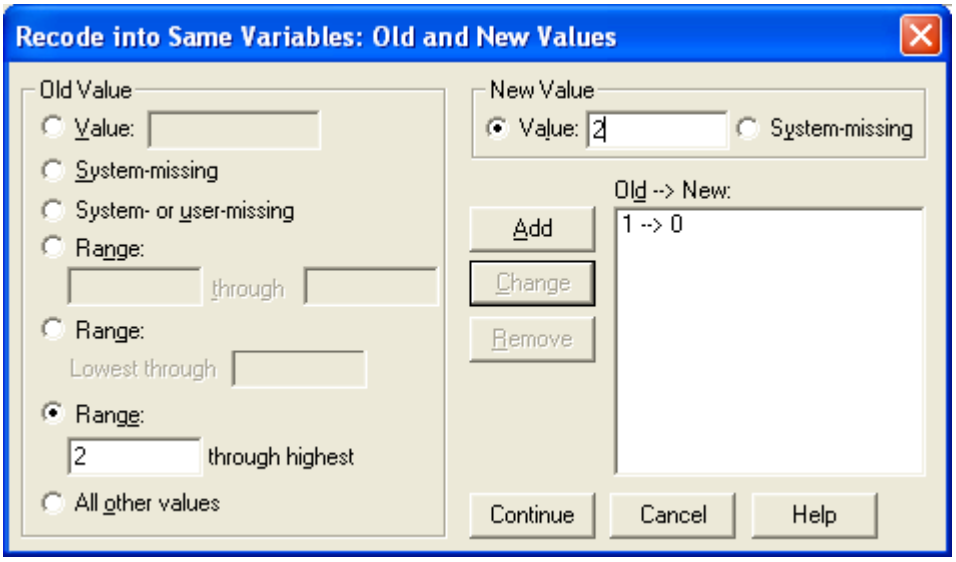
Bạn có thể xác định các trị số để mã hoá trong hộp thoại này. Mọi chỉ định về trị số phải cùng loại dữ liệu (dạng số hay dạng chuỗi) giống như của các biến đã được chọn trong hộp thoại chính.

Old Value. Trị số (hoặc các trị số) bị mã hoá. Bạn có thể mã hoá các trị số đơn, một phạm vi các trị số và các trị số khuyết thiếu. Các trị số khuyết thiếu hệ thống và các phạm vi không thể được chọn đối với các biến dạng chuỗi bởi vì không có khái niệm nào áp dụng cho các biến dạng chuỗi. Các phạm vi bao gồm các điểm cuối của chúng và mọi trị số khuyết thiếu của người sử dụng nằm trong phạm vi này.

New Value. Trị số đơn mà trong nó từng trị số cũ hoặc phạm vi của các trị số được mã hoá. Bạn có thể nhập một trị số hoặc chỉ định trị số khuyết thiếu hệ thống.

Old->New. Danh sách các trị số sẽ được sử dụng để mã hoá biến (hoặc các biến). Bạn có thể bổ sung, thay đổi hoặc loại bỏ các trị số này ra khỏi danh sách. Danh sách được tự động sắp xếp, dựa trên các trị số cũ, sử dụng trật tự sau: các trị số đơn, các trị số khuyết thiếu, các phạm vi và mọi trị số khác. Nếu bạn thay đổi một trị số trong danh sách, thủ tục sẽ tự động sắp xếp lại danh sách, nếu cần thiết, để duy trì trật tự này.

Hình 5-9: Hộp thoại Old and New Values

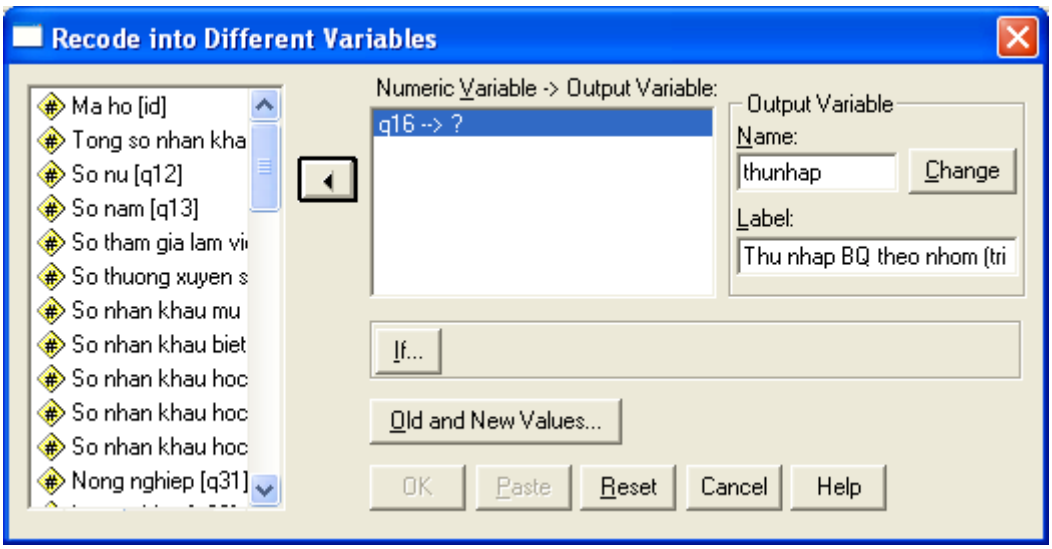


Mã hoá thành biến khác

Thủ tục Recode into Different Variables gán lại các trị số của các biến có sẵn hoặc các phạm vi của các trị số có sẵn vào các trị số mới của một biến mới. Ví dụ bạn có thể mã hoá lương năm của đối tượng điều tra vào một biến mới có các trị số là lương năm nhưng chia theo khoảng.

- Bạn có thể mã hoá các biến dạng số và dạng chuỗi
- Bạn có thể mã hoá các biến dạng số sang dạng chuỗi và ngược lại
- Nếu bạn chọn nhiều biến, chúng phải có cùng loại biến. Bạn không thể cùng một lúc mã hoá lại cả biến dạng số lẫn biến dạng chuỗi được.

Hình 5-10: Hộp thoại Recode into Different Variables



2.2. Mã hoá lại dữ liệu rồi chuyển sang biến mới

- Từ thanh menu chọn

  Transform/Recode/Into Different Variables…

- Chọn các biến mà bạn muốn mã hoá, Nếu bạn chọn nhiều biến, chúng phải có cùng dạng (chuỗi hoặc số)
- Nhập một tên biến mới cho từng biến và nhắp Change.
- Nhắp Old and New Values và định rõ cách mã hoá lại trị số.

  Một cách tuỳ chọn, bạn có thể chọn một nhóm các đối tượng để mã hoá

Hộp thoại Recode into Same Values: Old and NewValues

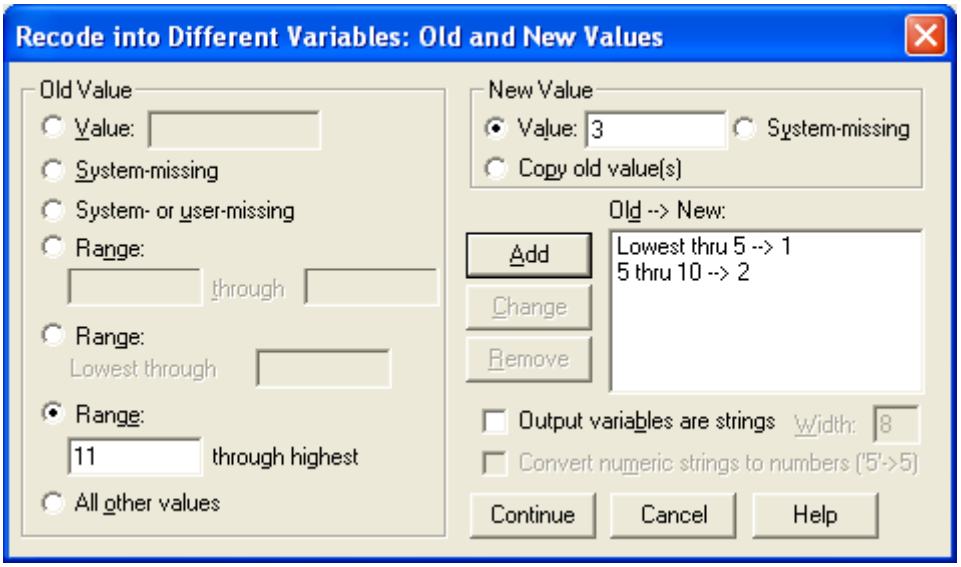
Bạn có thể xác định các trị số để mã hoá trong hộp thoại này. Mọi chỉ định về trị số phải cùng loại dữ liệu (dạng số hay dạng chuỗi) giống như của các biến đã được chọn trong hộp thoại chính.

Old Value. Trị số (hoặc các trị số) bị mã hoá. Bạn có thể mã hoá các trị số đơn, một phạm vi các trị số và các trị số khuyết thiếu. Các trị số khuyết thiếu hệ thống và các phạm vi không thể được chọn đối với các biến dạng chuỗi bởi vì không có khái niệm nào áp dụng cho các biến dạng chuỗi. Các phạm vi bao gồm các điểm cuối của chúng và mọi trị số khuyết thiếu của người sử dụng nằm trong phạm vi này.

New Value. Trị số đơn mà trong nó từng trị số cũ hoặc phạm vi của các trị số được mã hoá. Bạn có thể nhập một trị số hoặc chỉ định trị số khuyết thiếu hệ thống.

Old->New. Danh sách các trị số sẽ được sử dụng để mã hoá biến (hoặc các biến). Bạn có thể bổ sung, thay đổi hoặc loại bỏ các trị số này ra khỏi danh sách. Danh sách được tự động sắp xếp, dựa trên các trị số cũ, sử dụng trật tự sau: các trị số đơn, các trị số khuyết thiếu, các phạm vi và mọi trị số khác. Nếu bạn thay đổi một trị số trong danh sách, thủ tục sẽ tự động sắp xếp lại danh sách, nếu cần thiết, để duy trì trật tự này.

Hình 5.11: Hộp thoại Old and New Values


Recode into Different Variables: Old and New Values
Old Value
Value:
System-missing
System- or user-missing
Range:
through
Range:
Lowest through
Range:
11
through highest
All other values
New Value
Value: 3
System-missing
Copy old value(s)
Old --> New:
Add
Change
Remove
Lowest thru 5 --> 1
5 thru 10 --> 2
Output variables are strings
Width: 8
Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help

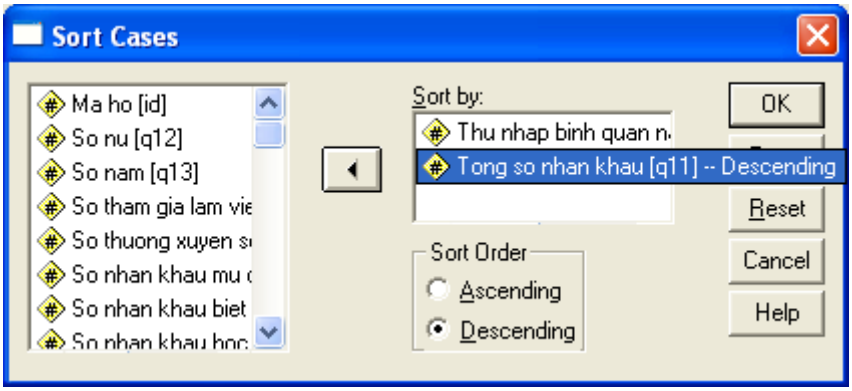
3. Lựa chọn và sắp xếp mẫu theo mục đích sử dụng

3.1. Sắp xếp các đối tượng

Hộp thoại này sắp xếp các đối tượng (các hàng) của file dữ liệu dựa vào các trị số của một hoặc một số biến sắp xếp. Bạn cửa sổ thể sắp xếp các đối tượng theo trật tự tăng dần hoặc giảm dần.

- Nếu bạn chọn nhiều biến sắp xếp, các đối tượng được sắp xếp theo từng biến trong vòng từng nhóm của biến đứng trước trong danh sách Short by. Ví dụ nếu bạn chọn biến *gender* {*giới tính*} là biến sắp xếp thứ nhất và *minority* {*thiểu số*}là biến sắp xếp thứ hai, các đối tượng sẽ được sắp xếp theo phân loại thiểu số trong từng loại giới tính.
- Đối với các biến, các chữ in đứng trước các chữ thường giống nó trong trật tự sắp xếp.

Hình 5-12: *Hộp thoại Sort Cases*



Để sắp xếp các đối tượng

- Từ thanh menu chọn

  Data/Sort Cases …

- Chọn một hoặc một số biến sắp xếp.

3.2. Chọn các đối tượng {Select Cases}

Thủ tục Select Cases cung cấp một số phương pháp khác nhau để chọn một nhóm các đối tượng dựa vào các tiêu chí bao gồm các biến và các

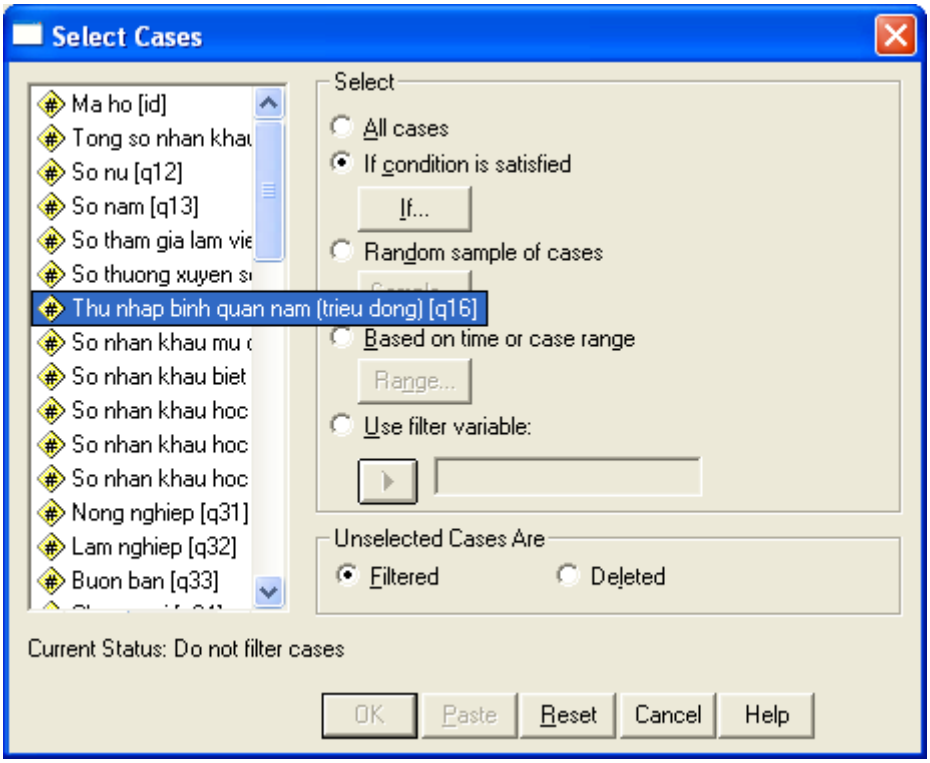
biểu thức phức. Bạn cũng có thể chọn một mẫu ngẫu nhiên các đối tượng. Tiêu chí dùng để định nghĩa một nhóm có thể bao gồm:

- Các trị số biến và các phạm vi/khoảng biến thiên
- Các phạm vi ngày tháng và thời gian
- Các số hàng
- Các biểu thức số học
- Các biểu thức lô-gíc
- Các hàm

Unselected Cases. Bạn có thể lọc hoặc xoá bỏ các đối tượng không đáp ứng tiêu chuẩn lựa chọn. Các đối tượng được lọc vẫn duy trì trong file dữ liệu nhưng bị loại ra khỏi phép phân tích. Thủ tục Select Cases tạo ra một biến lọc, *filter_$*, để chỉ rõ tình trạng lọc. Các đối tượng được chọn có trị số 1; các đối tượng không được chọn (bị lọc) có trị số 0. Các đối tượng bị lọc cũng được đánh dấu bằng một dấu gạch chéo qua số hàng trong cửa sổ Data Editor. Để đóng tình trạng lọc và bao gồm mọi đối tượng trong phép phân tích của ban, hãy chọn All Cases.

Các đối tượng bị xoá bỏ bị loại ra khỏi file dữ liệu và không thể phục hồi lại được nếu bạn lưu file dữ liệu sau khi xoá bỏ các đối tượng.

Hình 5.13: Hộp thoại Select Cases



Để chọn một nhóm các đối tượng

- Từ thanh menu chọn:

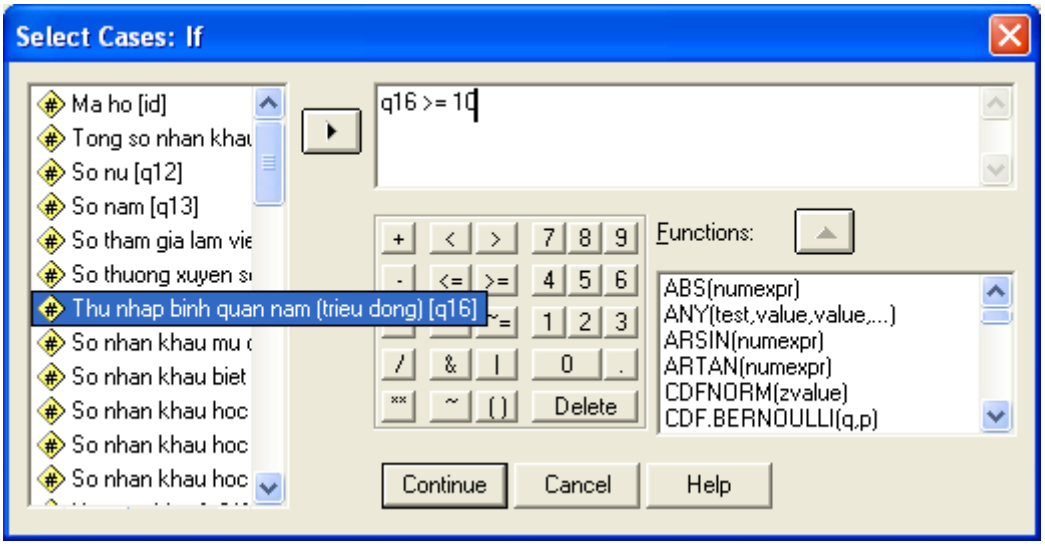
  Data/Select Cases…

- Chọn một trong những phương pháp lựa chọn các đối tượng.
- Định rõ các tiêu chí chọn các đối tượng.

2.1. Select Cases: If

Hộp thoại này cho phép bạn chọn các nhóm đối tượng có sử dụng các biểu thức điều kiện. Một biểu thức điều kiện trả lại một trị số *true* {*đúng*}, *false* {*sai*}, hoặc *missing* {*khuyết thiếu*} cho từng đối tượng.

Hình 6-3: Hộp thoại Select Cases: If



- Nếu kết quả của một biểu thức điều kiện là *true*, đối tượng sẽ được chọn
- Nếu kết quả của một biểu thức điều kiện là *false* hoặc *missing*, đối tượng sẽ không được chọn
- Hầu hết các biểu thức điều kiện sử dụng một hoặc một vài trong số 6 phép tính điều kiện (<, >, =, <=, >=, và ~=) trên bảng tính toán.
- Các biểu thức điều kiện có thể bao gồm các tên biến, hằng số, các phép tính số học, các hàm số và các hàm khác, các biến lô-gic, và các phép tính điều kiện.

2.2. Select Cases: Random Sample

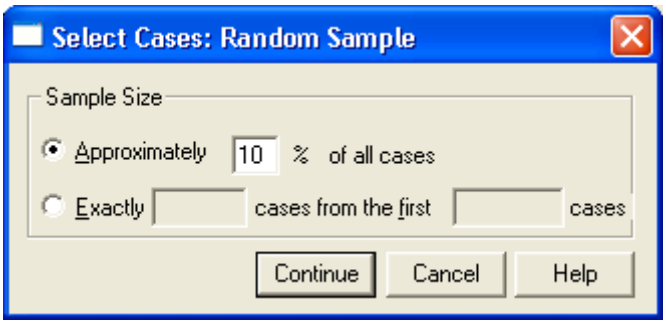
Hộp thoại này cho phép bạn chọn một mẫu ngẫu nhiên dự trên một tỷ lệ thích hợp hoặc một lượng chính xác các đối tượng.

Approximately. Tạo ra một mẫu ngẫu nhiên các các đối tượng gần đúng với một tỷ lệ được xác định trước. Do cách làm này tạo ra một quyết định ngẫu nhiêu giả định độc lập cho từng đối tượng, tỷ lệ các đối tượng được chọn chỉ có thể gần đúng với tỷ lệ được xác định trước. Càng có nhiều đối tượng trong file dữ liệu, tỷ lệ các đối tượng được chọn càng gần đúng với tỷ lệ được xác định trước.

Exactly. Một số lượng đối tượng được xác định bởi người sử dụng. Bạn cũng phải chỉ rõ số các đối tượng để từ đó tạo ra mẫu. Con số thứ hai

cần phải nhỏ hơn hoặc bằng tổng số đối tượng có trong file dữ liệu. Nếu con số này vượt quá tổng số đối tượng có trong file dữ liệu, mẫu sẽ bao gồm một cách tỷ lệ ít đối tượng hơn con số yêu cầu.

Hình 6-4: Hộp thoại Select Cases: Random Sample

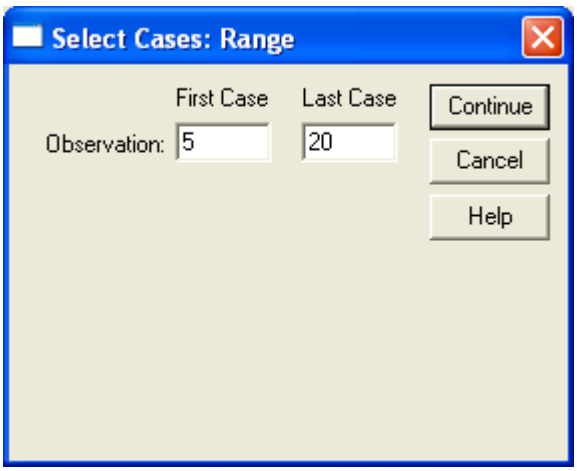


2.3. Select Cases: Range

Hộp thoại này chọn các đối tượng dựa vào một phạm vi số đối tượng hoặc một phạm vi các ngày hoặc thời gian

- Các phạm vi đối tượng được dựa vào số hàng được thể hiện trong cửa sổ Data Editor
- Các phạm vi ngày tháng hoặc thời gian chỉ có sẵn đối với dữ liệu chuỗi thời gian {time series data} với các biến ngày tháng được xác định (menu Data, Define Data).

Hình 6-5: Hộp thoại Select Cases: Range đối với phạm vi các đối tượng (không có các biến ngày tháng được định nghĩa)



BÀI 6. PHÂN TÍCH SỐ LIỆU

1. Lập bàng phân bố tần suất cho một biến trả lời

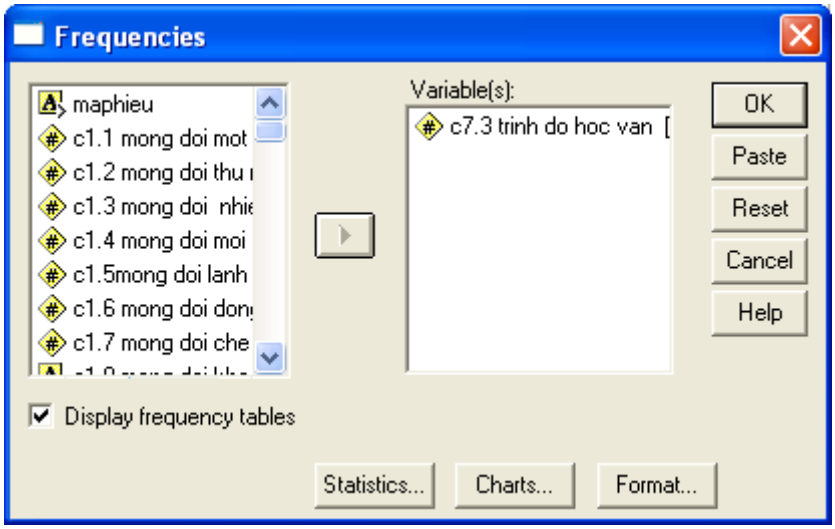
Chúng ta có thể khảo sát dữ liệu thông qua các công cụ như: Tần suất xuất hiện, phần trăm, phần trăm tích lũy. Ngoài ra nó còn cung cấp cho ta các phép đo lường thông kê như độ tập trung (central tendency measurement), độ phân tán (dispersion), tứ phân vị (Quartiles) và các bách phân vị (percentiles), phân phối dữ liệu (distribution).

Lập bảng này ngoài việc tóm tắt dữ liệu, nó còn giúp ta phát hiện những sai sót trong dữ liệu như, những giá trị bất thường (quá lớn hay quá nhỏ) có thể làm sai lệch kết quả phân tích thống kê, những giá trị mã hóa bất thường do sai sót việc nhập liệu hay mã hóa

Để tạo ra kết quả từ thanh menu chúng ta chọn

Analyze/Discriptives statistic/Frequencies...

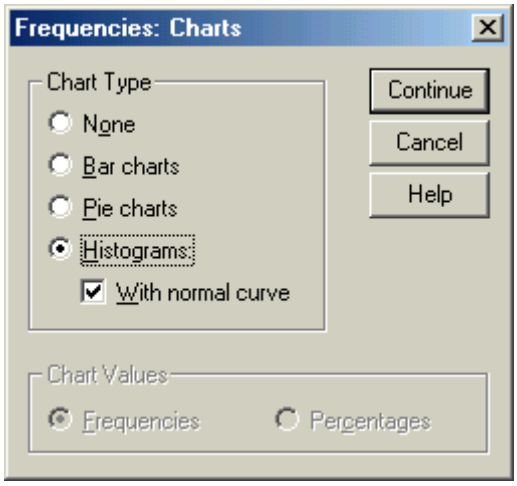
Nhắp reset để phục hồi mặc định của hộp thoại, sau đó chọn:



Hình 6.1: Hộp thoại Frequencies

Sau khi đã chọn xong, nhấn continue, để trở về hộp thoại chính.

Nếu chúng ta muốn sử dụng đồ thị trong phân tích bằng thủ tục frequences, nhấp chuột vào hộp thoại phụ Chart...

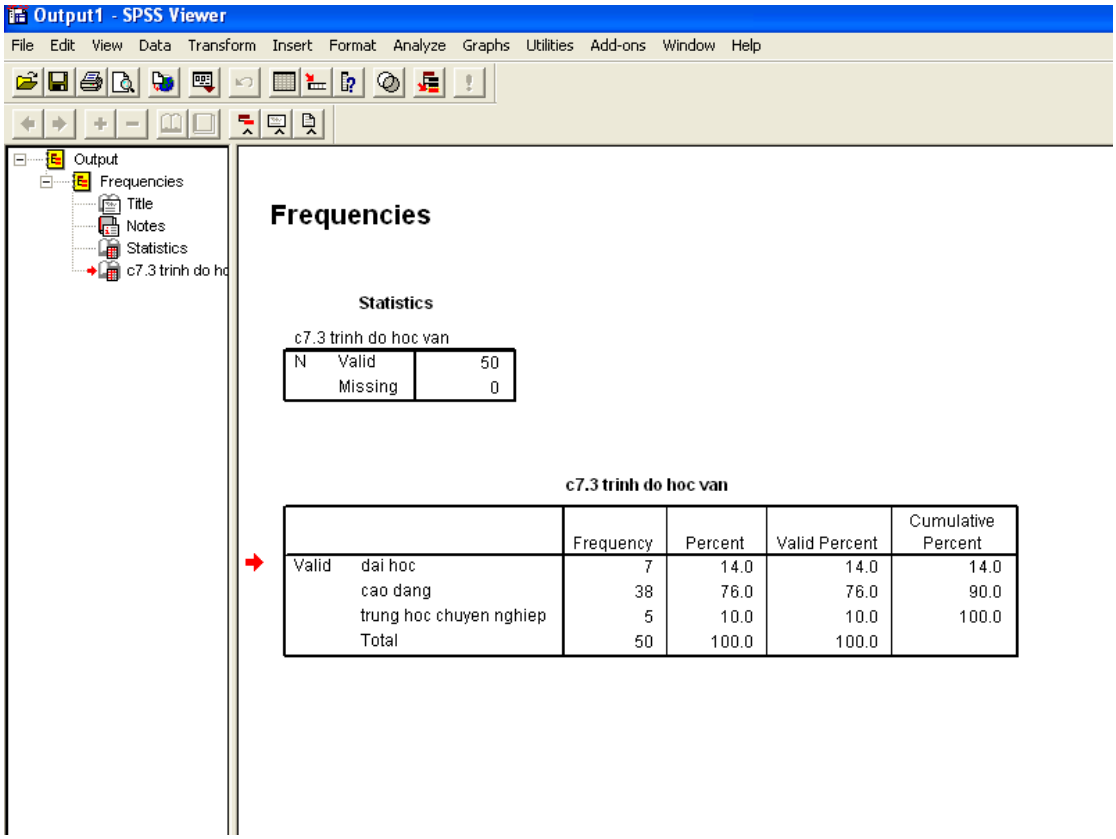


Hình 6.2: Hộp thoại Frequencies: Charts

Chọn kiểu đồ thị, ví dụ: Histogram, sau đó nhấn continue, trở về hộp thoại chính.

Cuối cùng, kết thúc bằng việc nhấn vào nút OK.

Trong bảng kết xuất khi sử dụng thủ tục Frequencies, chúng ta có những thông tin chính sau:



Hình 6.3: Hộp thoại Frequencies: Output – spss viewer

Trong bảng số liệu do SPSS kết xuất, chúng ta thấy có những thông tin sau:

c7.3 trinh do hoc van

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	dai hoc	7	14.0	14.0	14.0
	cao dang	38	76.0	76.0	90.0
	trung hoc chuyen nghiep	5	10.0	10.0	100.0
	Total	50	100.0	100.0	

Cột Valid: SPSS liệt kê những nhãn biến mà ta chọn khi làm thủ tục Frequencies.

Cột Frequency (tần số): Chứa các số đếm (counts) hoặc số lần xuất hiện tại phương án trả lời đó

Cột Percent (phần trăm): Những % này là thống kê hữu ích vì không có những giá trị khuyết thiếu.

Cột Valid percent (Phần trăm trên tổng số các trị số có giá trị): Trong trường hợp có các giá trị khuyết thiếu, hãy sử dụng phần trăm được báo cáo trong cột.

Cột Cumulative Percent (phần trăm cộng dồn):

Lưu ý:

* Đối với các loại biến định danh (phân loại, nhóm tổ không thứ bậc) thì chỉ nên tập trung đọc các số liệu ở cột Frequency, Precent, Valid Precent

* Đối với các loại biến định lượng hoặc biến phân loại nhóm/tổ có thứ bậc thì nên dùng số liệu ở cột Cumulative.

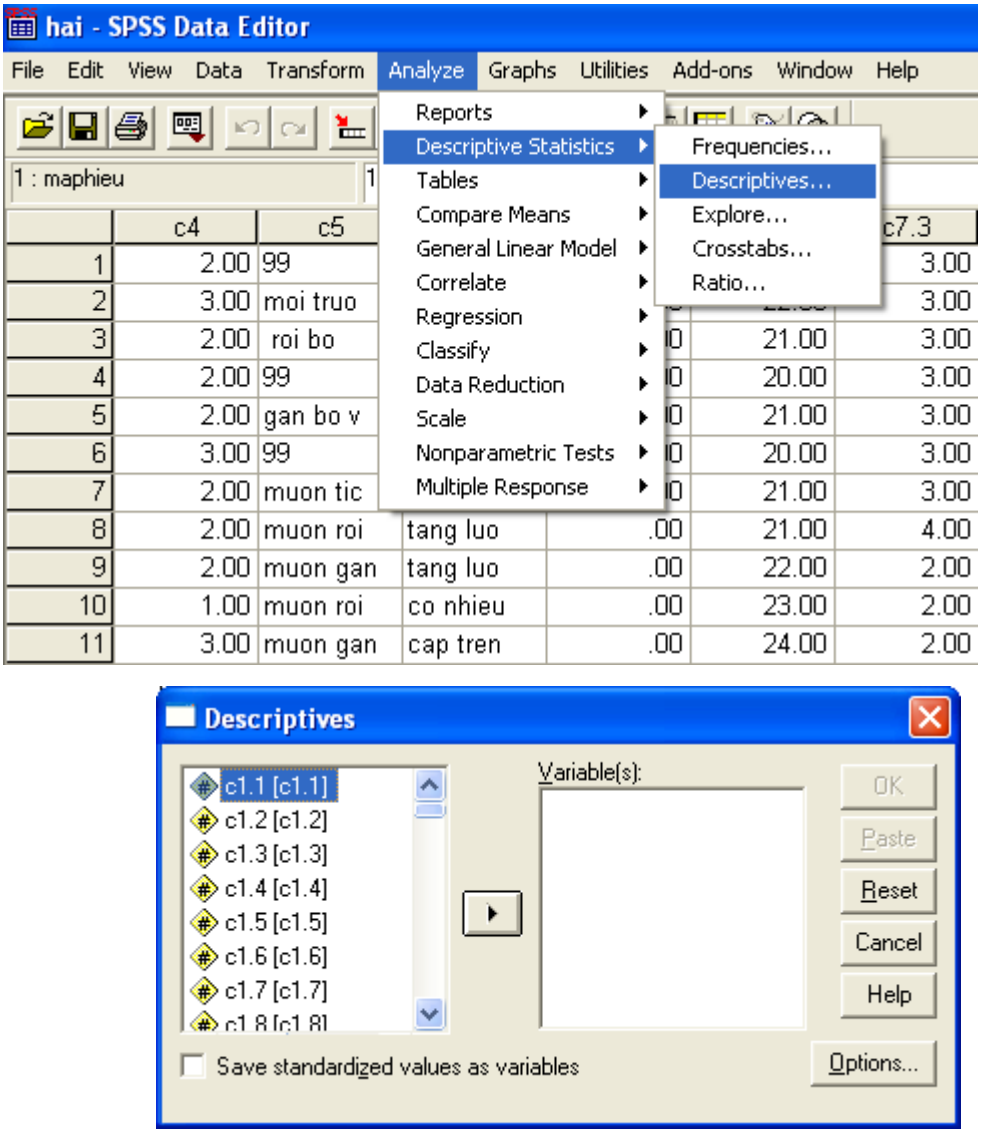
Đối với kiểu số liệu định lượng, chúng ta có thể dùng thêm những tiện ích ở hộp thoại phụ Statistices... để có thể thông tin về những giá trị thống kê như:

Mean: Số trung bình = trung bình trung của tổng thể

Median: Trung vị = giá trị nằm giữa phân phối. Nói dễ hiểu hơn là một nửa trị số trong dãy số là những số lớn hơn trung vị và nửa còn lại nhỏ hơn trung vị

2. Lập bảng mô tả (Descriptive)

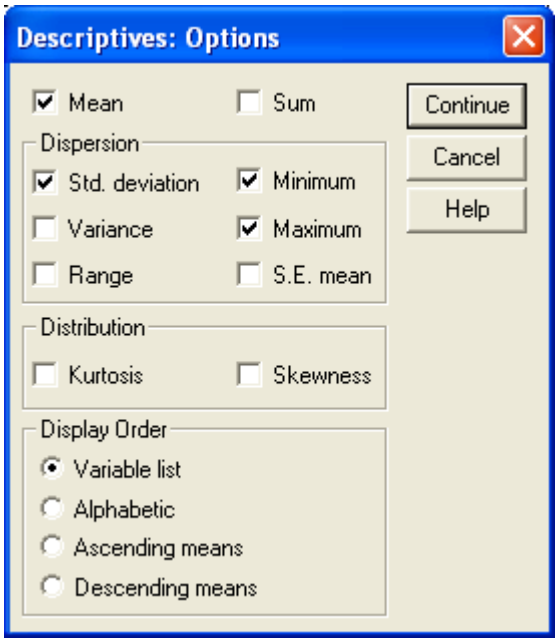
Sử dụng **Statisticts\Summaries\Descriptives** để mở hộp thoại mô tả thống kê



Hình 6.4: Hộp thoại Descriptives

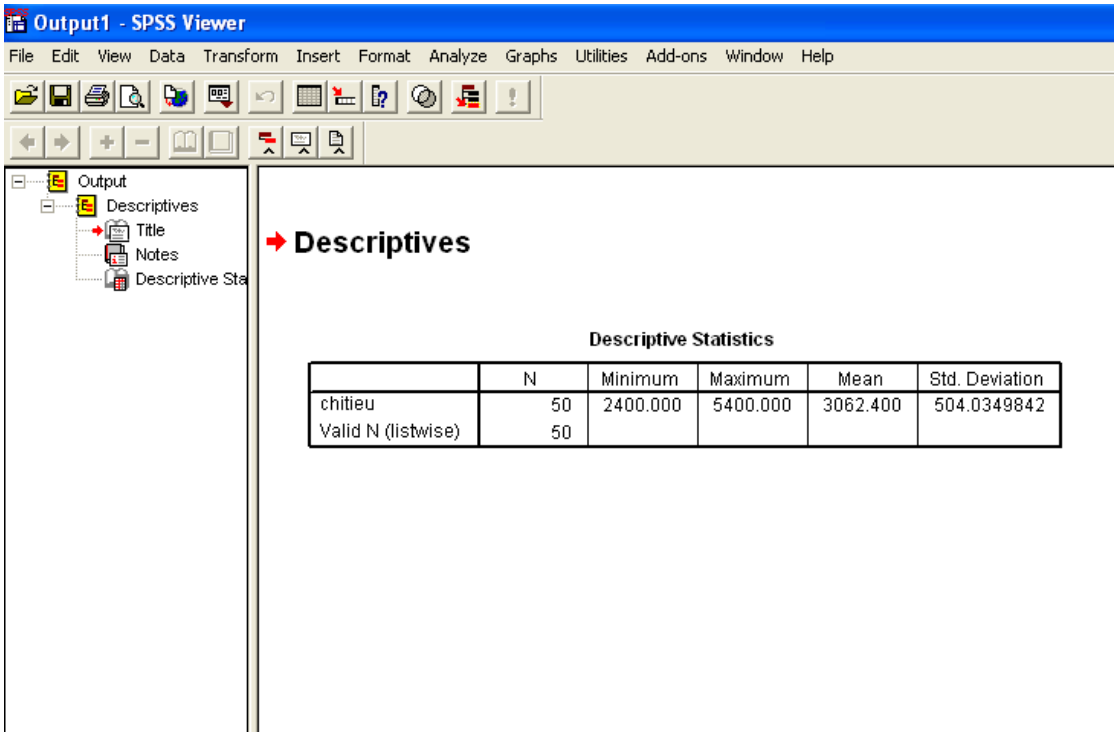
Đây là một dạng công cụ khác có thể được dùng để tóm tắc dữ liệu và chỉ cho phép thao tác trên dạng dữ liệu định lượng (thang đo khoảng cách và tỷ lệ). Được dùng để thể hiện xu hướng tập trung của dữ liệu

(central tendency) thông qua giá trị trung bình của các giá trị trong biến (mean), và mô tả sự phân tán của dữ liệu thông qua phương sai và độ lệch chuẩn. Chuyển các biến cần tóm tắt vào hộp thoại **variables** và nhấp thanh **options** để lựa chọn các thông số thống kê cần mô tả, như giá trị trung bình–mean, giá trị tối thiểu, giá trị tối đa, phương sai và độ lệch chuẩn,…



Hình 6.5. Hộp thoại Descriptives : options

Trong bảng kết xuất (output) Descriptives chúng ta có được kết quả sau :



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
chitieu	50	2400.000000	5400.000000	3062.400000	504.034984209
Valid N (listwise)	50				

Hình 6.6. Bảng Output

Trong bảng, chúng ta có những kết quả sau :

N : Số lượng mẫu

Minimun : Giá trị nhỏ nhất

Maximum : Giá trị lớn nhất

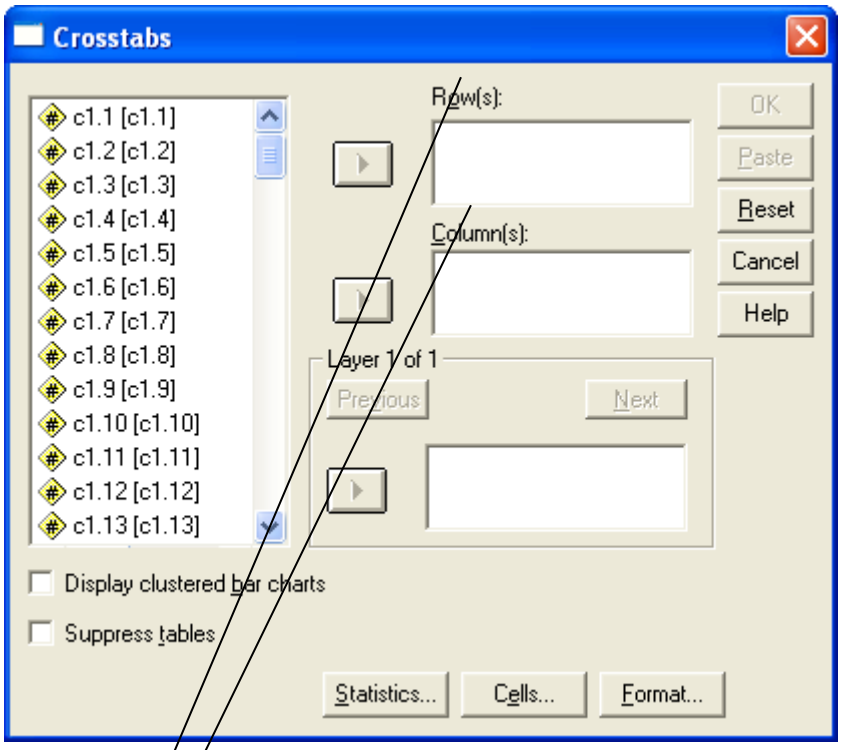
Mean : Giá trị trung bình

Std.Devuation ; Độ lệch chuẩn.

Theo nhưng bảng ví dụ trên, ta thấy mức độ chi tiêu trung bình của người lao động là 3.062.400 đồng Trong đó người chi tiêu thấp nhất là 2,4 triệu, người chi tiêu cao nhất là 5,4 triệu. Độ lệch chuẩn trong chi tiêu là 504 nghìn đồng…

3. Lập bảng tương quan chéo (Crosstabs)

Bảng nhiều chiều là dạng bảng chéo thể hiện tần suất xuất hiện của một biến này trong mối quan hệ với một hay nhiều biến khác. Bảng chéo còn cung cấp nhiều loại kiểm nghiệm thống kê và đo lường mối quan hệ và tương quan giữa các biến trong bảng. Cấu trúc của bảng và loại dữ liệu (loại thang đó) sẽ quyết định loại công cụ nào được sử dụng để đo lường. Ngoài việc thể hiện mối liên hệ giữa các biến. Bảng nhiều chiều còn giúp ta phát hiện những sai sót trong dữ liệu từ việc phát hiện ra những mối quan hệ vô lý và bất thường giữa hai biến. Chọn trên menu **Statistics/Summaries/Crosstabs** để mở hộp thoại như Hình 6-7:



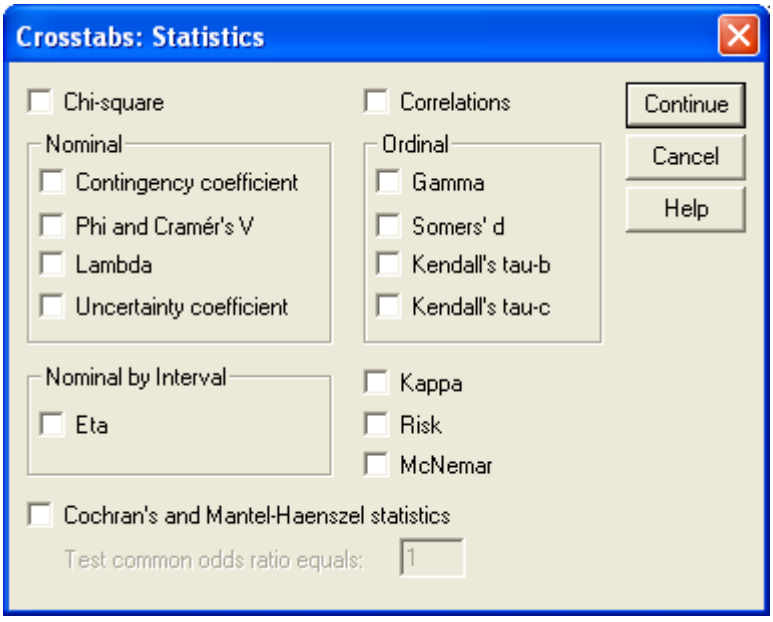
Hình 6.7 : Bảng tương quan chéo (crosstabs)

Các biến trong tập dữ liệu được hiển thị bên hộp bên trái. Chọn các biến hàng đưa váo hộp Row(s) và các biến cột đưa vào hộp Column(s). Thông thường biến phụ thuộc hay biến cần quan sát thường được đưa và hàng (**rows**) và biến độc lập hay biến kiểm soát được đưa và cột (**columns**). Việc lựa chọn các phân tích theo các tỷ lệ phần trăm, % row và % column cũng như % total tuỳ thuộc vào yêu cầu nghiên cứu.

Ngoài ra, chúng ta có thể đưa thêm vào bảng chéo các lớp biến điều khiển (layer) để tạo ra các bảng biến chéo nhiều chiều. Mỗi bảng chéo riêng biệt sẽ được tạo ra ứng với mỗi giá trị của mỗi biến điều khiển. Mỗi lớp điều khiển sẽ chia bảng chéo thành nhiều nhóm nhỏ hơn. Có thể thêm tối đa 8 biến điều khiển, dùng các thanh Next và previous để di chuyển giữa các biến điều khiển này. Việc đưa vào các biến điều khiển này cho phép ta xem xét các mối quan hệ mà lúc ban đầu không thể thấy ngay. Các công cụ thống kê sẽ cho ra các kết quả riêng biết đối với từng giá trị của biến điều khiển.

Công cụ Cells trong hộp thoại cho phép ta tính toán các hệ số đo lường mối quan hệ giữa các biến đó như % hàng, % cột, % Total.

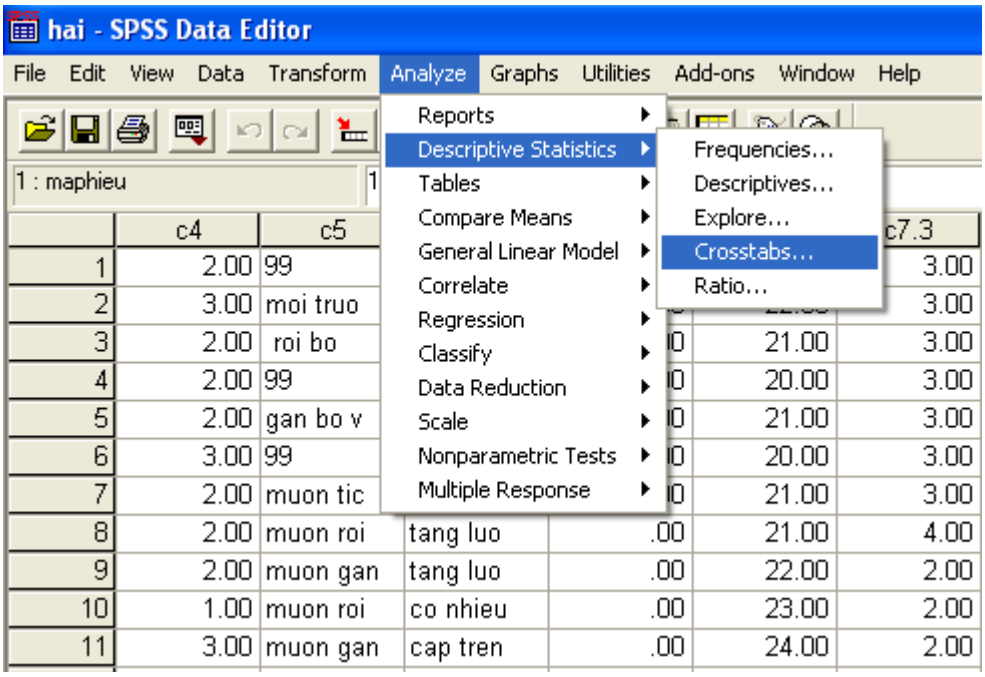
Công cụ Statistics cho phép ta tính các kiểm nghiệm giả thuyết về tính độc lập của các biến, và mối liên hệ giữa các các biến, hệ số tương quan, cũng như đo lường các mối quan hệ đó. (Xem Hình 6- 8)



Hình 6.8: Hộp thoại Crosstabs: Statistic

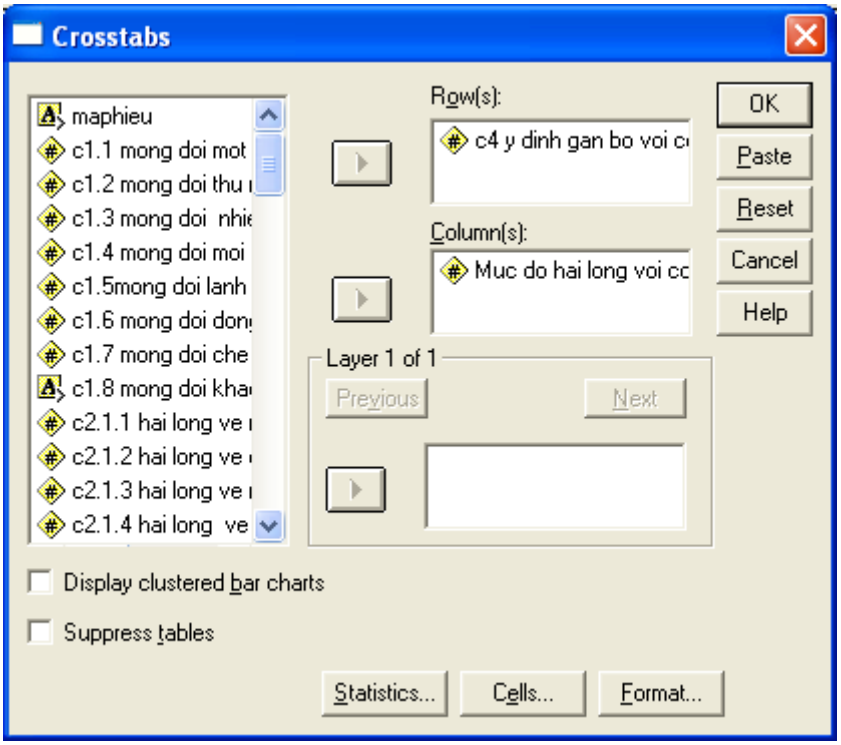
Ví dụ: Chúng ta muốn tìm hiểu liệu có mối quan hệ nào giữa mức độ hài lòng với công việc của người lao động với sự gắn bó của họ với công ty hay không? Chúng ta làm theo những bước sau:

Bước 1: Chọn trên menu **Analyze/Descriptive Statistics/Crosstabs**



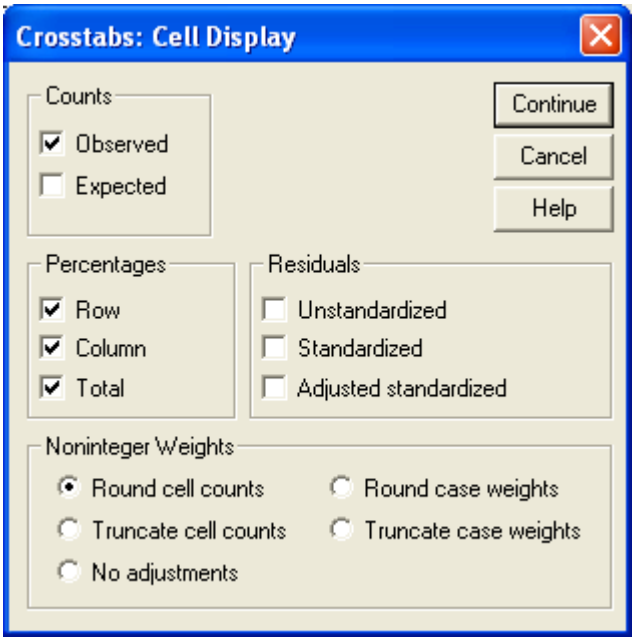
Hình 6.9: Thao tác với Menu/ Analyze/Descriptive Statistics/Crosstabs

Bước 2: Chọn biến Mucdohailong từ biến nguồn sang hộp thoại column (s) bên biến đích. Và chọn biến c4 (ý định gắn bó với công ty) đưa vào hộp thoại Row(s). Lưu ý những biến được xem như là biến độc lập sẽ được đưa vào column; biến phụ thuộc đưa bào Row (s)



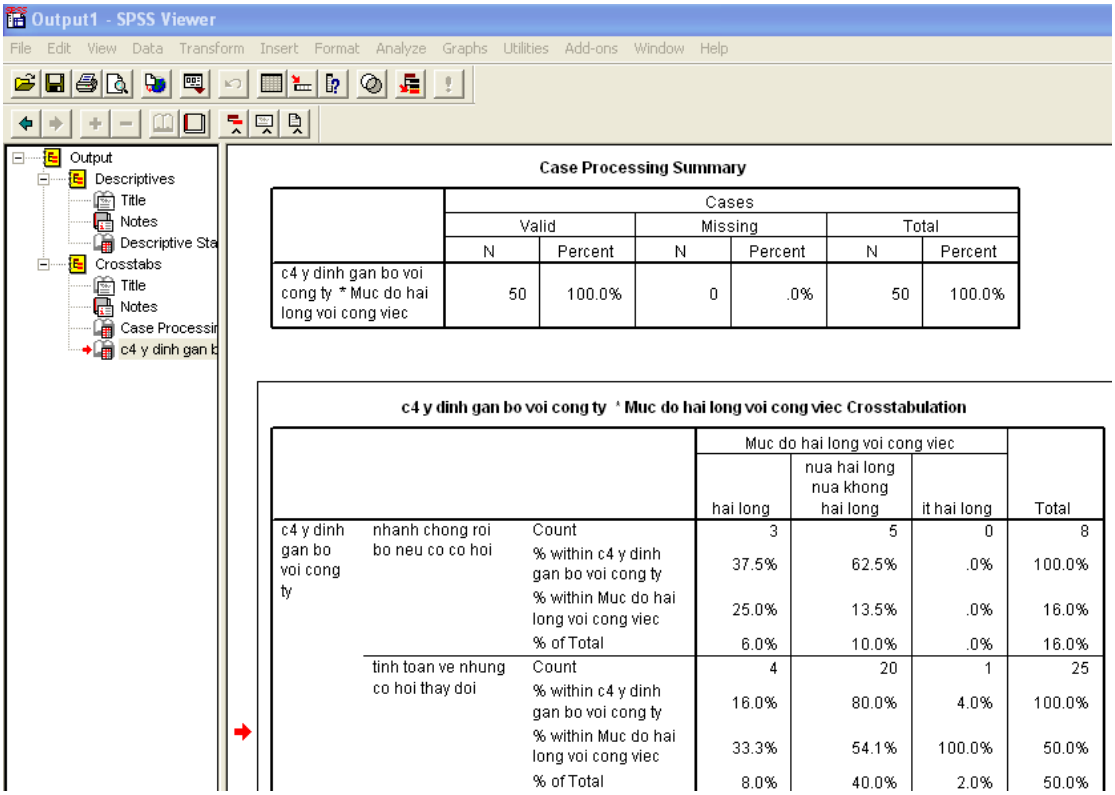
Hình 6.10. Hộp thoại Crosstabs

Bước 3: Bấm ô Cell để đặt các thông số tính toán gồm Row, Column,Total. Nhấn Continue/OK để kết thúc.



Hình 6.11. Hộp thoại Crosstabs: Cell Display

Bước 4. Đọc bảng số liệu bên màn hình Output



Hình 6.11. Màn hình Output

c4 y dinh gan bo voi cong ty * Muc do hai long voi cong viec Crosstabulation

			Muc do hai long voi cong viec			
			hai long	nua hai long nua khong hai long	it hai long	Total
c4 y dinh gan bo voi cong ty	nhanh chong roi bo neu co co hoi	Count	3	5	0	8
		% within c4 y dinh gan bo voi cong ty	37.5%	62.5%	.0%	100.0%
		% within Muc do hai long voi cong viec	25.0%	13.5%	.0%	16.0%
		% of Total	6.0%	10.0%	.0%	16.0%
	tinh toan ve nhung co hoi thay doi	Count	4	20	1	25
		% within c4 y dinh gan bo voi cong ty	16.0%	80.0%	4.0%	100.0%
		% within Muc do hai long voi cong viec	33.3%	54.1%	100.0%	50.0%
		% of Total	8.0%	40.0%	2.0%	50.0%
	gan bo them mot thoi gian	Count	5	12	0	17
		% within c4 y dinh gan bo voi cong ty	29.4%	70.6%	.0%	100.0%
		% within Muc do hai long voi cong viec	41.7%	32.4%	.0%	34.0%
		% of Total	10.0%	24.0%	.0%	34.0%
Total		Count	12	37	1	50
		% within c4 y dinh gan bo voi cong ty	24.0%	74.0%	2.0%	100.0%
		% within Muc do hai long voi cong viec	100.0%	100.0%	100.0%	100.0%
		% of Total	24.0%	74.0%	2.0%	100.0%

Với bảng số liệu dang crosstab, chúng ta chú ý khi đọc số liệu như sau:

***Dòng thứ nhất** (count): Tần suất*

***Dòng thứ hai** (% within c4 y dinh gan bo voi cong ty): Đọc theo chiều ngang: Trong số những người công nhân muốn nhanh chóng rời bỏ công ty thì có 37,5% người hài lòng về công việc, trong khi đó có đến 62,5% người không hài lòng…*

***Dòng thứ ba**: Đọc theo chiều dọc: Trong số những người hài lòng với công việc, chỉ có 25% là muốn rời bỏ công ty ngay, trong khi có đến 41,7% muốn gắn bó thêm 1 thời gian nữa.*

***Dòng thứ 4**: Có 6% người hài lòng với công việc nhưng muốn rời bỏ công ty trong tổng số những người được hỏi.*

4. Các kiểm nghiệm thống kê

4.1. Kiểm nghiệm mối quan hệ và tương quan giữa các biến sử dụng trong bảng chéo

4.1.1. Kiểm nghiệm Chi-square:

Là một công cụ thông kê sử dụng để kiểm nghiệp giả thuyết cho rằng các biến trong hàng và cột thì độc lập với nhau (H_0). Phương pháp kiểm nghiệm này chỉ cho ta biết được liệu một biến này có quan hệ hay không với một biến khác, tuy nhiên phương pháp kiểm nghiệp này không chỉ ra cường độ của mối quan hệ giữa hai biến mạnh hay yếu (nếu có quan hệ), cũng như không chỉ ra hướng thuận hay nghịch của mối quan hệ này (nếu có quan hệ).

Để kiểm nghiệp tính độc lập giữa hai biến cột và hàng, kiểm nghiệp Chi-square sẽ cho ra các kết quả kiểm nghiệp như sau: Pearson chi-square, likelihood-ratio chi-square, and linear-by-linear association chi-square mỗi cái sẽ được sử dụng trong những trường hợp cụ thể.

Để kiểm nghiệm tính độc lập giữa hai biến, người ta sử dụng phân phối ngẫu nhiên Chi bình phương (χ^2) với tham số thống kê **Pearson chi bình phương** để tiến hành so sánh số lượng các trường hợp quan sát được với số lượng các trường hợp mong đợi bằng công thức sau:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Khi kết quả thống kê Chi bình phương (χ^2) đủ lớn (Dựa vào lý thuyết phân phối Chi bình phương với độ tin cậy xác định, kích cở mẫu là n, bậc tự do-degree of freedom là df=(r-1)(c-1)) ta có thể kết luận bác bỏ giả thuyết độc lập giữa hai biến (H_0). Hoặc sử dụng giá trị P (P-value hay Asymtotic Significance) so sánh với mức ý nghĩa (Significance

level) thường là á = 0.05 tương ứng với 95% độ tin cậy, ta có thể kết luận bác bỏ H_0 khi p-value nhỏ hơn hoặc bằng mức ý nghĩa và ngược lại chấp nhận H_0 khi p-value lớn hơn mức ý nghĩa.

Tuy nhiên để việc kiểm nghiệm này là đáng tin cậy thì các số liệu trong bảng chéo giữa hai biến đang khảo sát phải thỏa mãn một số điều kiện nhất định sau:

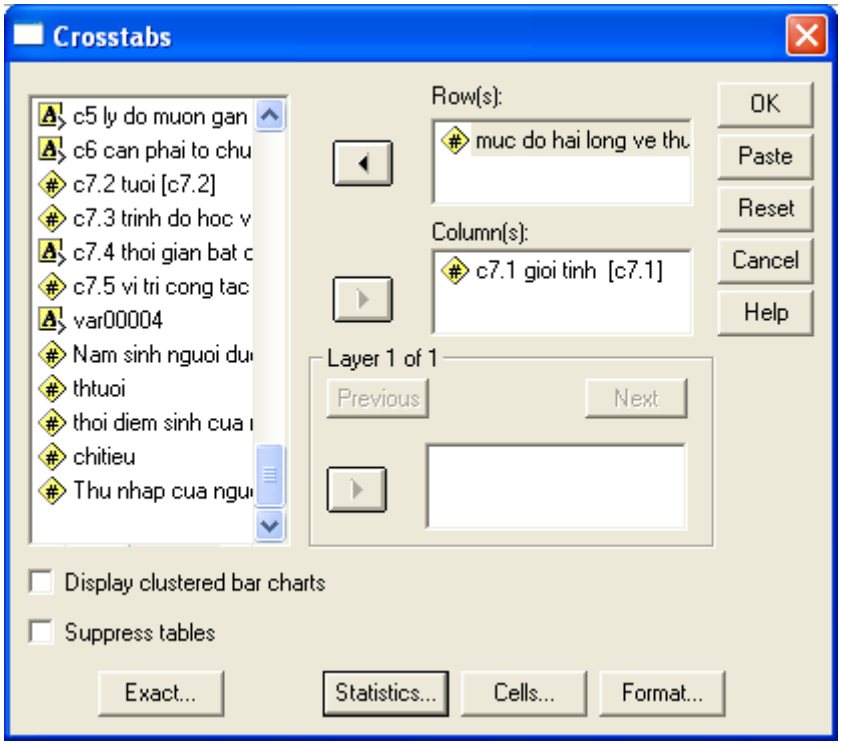
Không tồn tại ở bất kỳ ô giao nhau giữa hai biến có giá trị mong đợi nhỏ hơn 1.

Không vượt quá 20% lượng ô giao nhau giữa hai biến đang khảo sát trong bảng chéo có giá trị nhỏ hơn 5 (đối với bảng 2x2-bảng mà mỗi biến trong bảng chéo chỉ có hai giá trị, phần trăm giới hạn này là 0%)

Nếu không thỏa mãn các điều kiện trên ta phải tiến hành loại bỏ bớt các giá trị trong một biến mà dữ liệu giao nhau của nó là không đáng kể (quá nhỏ)

Để kết luận mối liên hệ giữa hai biến là độc lập hay phụ thuộc vào nhau (có hay không có tương quan) người ta dựa vào Asymptotic Significance với số mẫu đủ lớn hoặc phân phối là phân phối chuẩn. Đây là chỉ số thống kê để đo lường với mức ý nghĩa (thường là 5%) nhằm đưa ra kết luận phản bát hay chấp nhận giả thuyết ban đầu (Hai biến là độc lập với nhau). Ta có thể kết luận giữa hai biến tồn tại một mối quan hệ với nhau khi mà Asym. Sig. nhỏ hơn mức ý nghĩa và ngược lại.

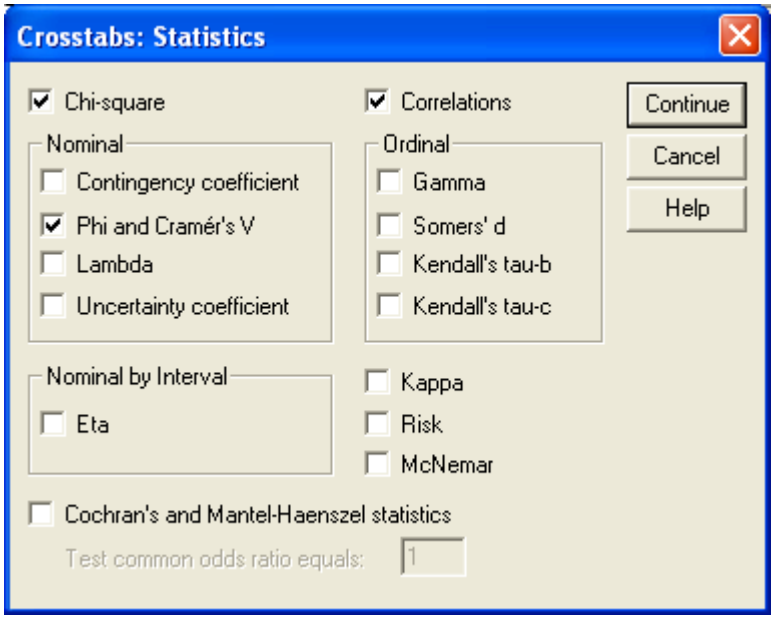
Ví dụ: Tìm mối liên quan giữa giới tính và thu nhập trung bình của công nhân trong công ty A....



Hình 6.12. Hộp thoại Crosstabs

Chọn biến giới tính (c7.1. gioi tinh) đưa vào column; biến Mức độ hài lòng (mudohailong) vào row(s).

Nhấn nút ***Statictics...***



Hình 6.13. Hộp thoại Crosstabs: statistics

Chọn ***Chi-square/.../continue/ok***

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5.667 (a)	2	.059
Likelihood Ratio	3.899	2	.142
Linear-by-Linear Association	3.320	1	.068
N of Valid Cases	50		

a 4 cells (66.7%) have expected count less than 5. The minimum expected count is .50.

Khi :

+ Asymp. Sig. (2-sided) (p.value) (mức độ ý nghĩa) > 0,05: Không có mối liên hệ giữa hai biến

+ Asymp. Sig. (2-sided) (p.value) (mức độ ý nghĩa) < 0,05: Có mối liên hệ giữa hai biến.

4.1.2. Kiểm nghiệm tương quan (Correlation:)

Phân tích mối tương quan : Dùng để đo lường về mối liên hệ giữa hai biến số. Nó có thể là dương (+) hoặc âm (-) hay 0.

- Spearman's rho được dùng để đo lường mối quan hệ giữa hai biến thứ tự (các biến này hầu hết đều được xắp xếp từ thấp nhất đến cao nhất).
- Khi các biến trong bảng là các biến định lượng ta sử dụng hệ số Pearson correlation coefficient để đo lường mối quan hệ tuyến tính giữa các biến này.

Hệ số tương quan cho ta biết chiều hướng âm hay dương và độ mạnh (strength) của mối tương quan. Nếu r dương, điều đó có nghĩa là khi giá trị một biến tăng lên thì giá trị của biến kia cũng tăng lên theo một chiều hướng. Ngược lại nếu r âm thì giá trị của biến kia thay đổi theo chiều hướng ngược lại. Trị tuyệt đối của r nói lên độ mạnh của sự tương quan theo chiều

thuận hoặc nghịch. Trị tuyệt đối tối đa của r là 1.00. Khi không có tương quan nào giữa hai biến, trị số r = 0.

Đánh giá mức độ tương quan theo các mức sau đây:

Từ 0,80 đến 1: tương quan cao, đáng tin cậy

Từ 0,60 đến 0,79: tương quan vừa phải và đáng kể

Từ 0,40 đến 0,59: tạm được

Từ 0,20 đến 0,39: tương quan ít

Từ 0,00 đến 0,19: tương quan không đáng kể hay tương quan do may rủi.

Các giá trị của hệ số tương quan biến thiên từ –1 đến 1, dấu cộng hoặc trừ chỉ ra hương tương quan giữa các biến (thuận hay nghịch), giá trị tuyệt đối của chỉ số này cho biết cường độ tương quan giữa hai biến, giá trị này càng lớn mối tương quan càng mạnh.

Ví dụ: Xác định mối tương quan giữa giới tính và mức độ hài lòng với thu nhập.

Lập lại thao tác như làm với kiểm định Khi bình phương (Chi-Square Tests), sau đó có kết quả

Symmetric Measures

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Nominal by Nominal	Phi	.337			.059
	Cramer's V	.337			.059
Interval by Interval	Pearson's R	.260	.162	1.868	.068(c)
Ordinal by Ordinal	Spearman Correlation	.231	.152	1.644	.107(c)
N of Valid Cases		50			

a Not assuming the null hypothesis.

b Using the asymptotic standard error assuming the null hypothesis.
c Based on normal approximation.

Đọc kết quả:

Với Approx. Sig (p.value) = 0.59 > 0,05: 2 biến không có mối liên hệ với ý nghĩa thống kê, do đó hệ số tương quan Speaman/Pearson không có ý nghĩa. Điều này cho thấy giữa giới tính và mức độ hài lòng với thu nhập không có mối liên hệ có ý nghĩa thống kê.

Với Approx. Sig (p.value) < 0,05: 2 biến có mối liên hệ với ý nghĩa thống kê, do đó cần đọc hệ số tương quan Speaman/Pearson để thấy được độ mạnh của mối liên hệ.

4.1.3. Một số đo lường mối tương quan khác giữa hai biến

Giữa hai biến định danh:

Để đo lường mối quan hệ giữa hai biến biểu danh. Sử dụng các hệ số **Phi (coefficient)** và **Cramer's V, Contingency coefficient** để đo lường nếu dựa vào kết quả kiểm nghiệm Chi-bình phương. Ở đây các hệ số này sẽ bằng 0 nếu và chỉ nếu hệ số **Pearson chi bình phương** bằng 0. Do đó người ta sử dụng các thông số này để kiểm nghiệm giả thuyết cho rằng các hệ số này đều bằng 0 - điều này tương đương với giả thuyết độc lập giữa hai biến, hay hai biến không có môí quan hệ với nhau. Ta sẽ từ chối giả thuyết này.

Phi: Chỉ dùng cho dạng bảng 2x2 tables, hệ số phi coefficient này biến thiên từ -1 đến +1. Do đó hệ số này ngoài khả năng chỉ ra mối quan hệ và cường độ của mối quan hệ nó còn chỉ ra hướng của mối quan hệ đó

Cramer's V và Contingency coefficient (hệ số ngẫu hiên): Được sử dụng cho bảng mà số cột và hàng là bất kỳ, giá trị kiểm nghiệm biến thiên từ 0 đến 1, với giá trị 0 chỉ ra không có mối quan hệ giữa các biến

5. So sánh các giá trị trung bình

Có nhiều phép kiểm nghiệp được sử dụng trong SPSS:

Nếu so sánh giá trị trung bình của mẫu với một giá trị cố định nào đó ta sử dụng phép kiểm nghiệm t một mẫu (One-sample t test).

Nếu so sánh giá trị trung bình của một nhóm các trường hợp quan sát với một nhóm quan sát khác, ta sử dụng kiểm nghiệm t mẫu độc lập (Independent-sapmles t test).

Để so sánh giá trị trung bình của hai biến được khảo sát từ cùng một mẫu ta sử dụng kiểm nghiệp t theo từng cặp mẫu (Paired-samples t test).

Hoặc với trường hợp ta có nhiều hơn hai mẫu độc lập cần kiểm nghiệm trung bình, ta có thể dùng ANOVA một chiều (One-way ANOVA).

Với các trường hợp trên, hoặc các biến được kiểm nghiệm trung bình đòi hỏi phải là các biến định lượng và phân phối phải là phân phối ngẫu nhiên hay mẫu nghiên cứu phải đủ lớn. Tuy nhiên với những trường hợp biến quan sát là biến định lượng (nhưng là biến thang đó thứ tự) hoặc số lượng mẫu không đủ lớn hoặc không thỏa mãn điều kiện phân phối chuẩn ta có thể tiến hành kiểm nghiệp bằng công cụ Wilcoxon signed rank test trong kiểm nghiệm phi tham số

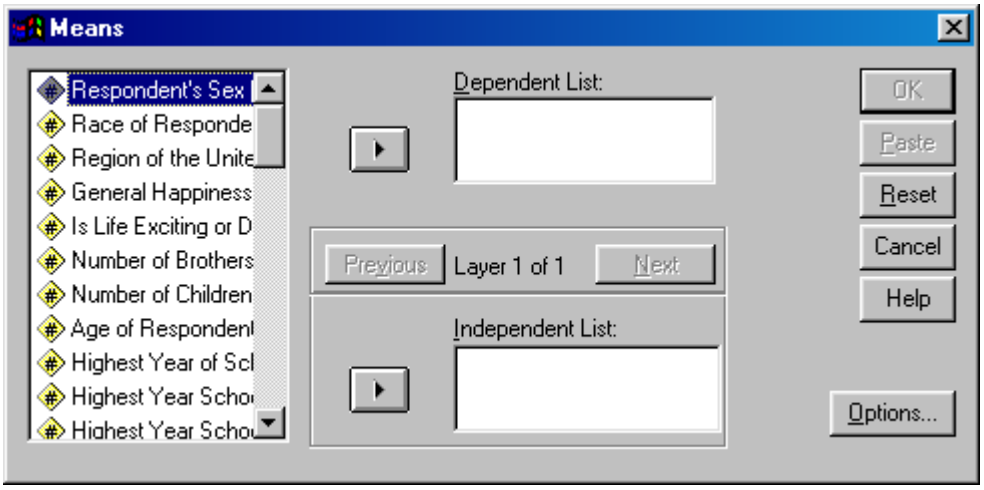
5.1. Means

Công cụ **Means** dùng để tính toán các giá trị trung bình và đưa các tham số thống kê liên quan cho một biến phụ thuộc trong phạm vi các nhóm của một hay nhiều biến độc lập. Ta có thể lựa chọn các công cụ kèm theo như phân tích ANOVA một chiều, eta, và các kiểm nghiệm tuyến tính. Ví dụ ta có thể đo lường mức độ đánh giá trung bình về một show quảng cáo của ba nhóm tiêu dùng khác nhau, công nhân, sinh viên và công chức. Công cụ này sẽ cho ta một bảng chéo thể hiện sự đánh giá của ba nhóm người này về show quảng cáo được xem.

*Các biến phụ thuộc trong bảng **Means** phải là biến định lượng và các biến độc lập thường là các biến định danh. Các đại lượng thống kê được sử dụng tùy thuộc vào dạng dữ liệu. Như **mean** và **stadard deviation** thì dựa trên lý thuyết phân phối chuẩn và thích hợp cho các biến định lượng*

với phân phối đối xứng. Các đại lương khác như ***Media,*** *và* ***range*** *thì thích hợp cho các biến định lượng mà ta không biết liệu nó có thoả mãn các điều kiện về phân phối chuẩn hay không.*

Để thực hiện công cu này ta chọn **Compare Means/Means…. Từ Menus**, ta có hộp thoại như hình 7-10**.**



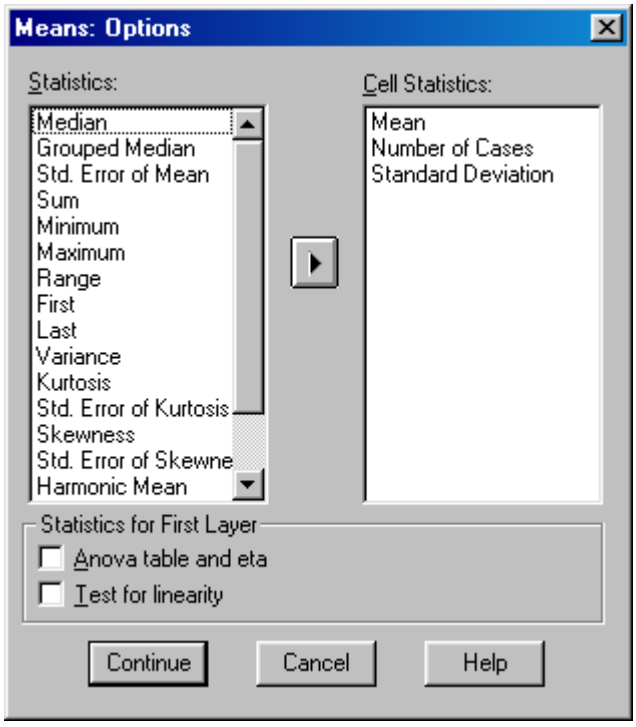
Hình 6-14. Hộp thoại Mean

Có thể chọn một hay nhiều biến phụ thuộc. Di chuyển vệt đen đến biến chứa đựng các giá trị định lượng mà ta cần quan sát giá trị trung đó trong phạm vi các nhóm trong biến độc lập, sử dụng mủi tên chuyển biến đã chọn vào hộp thoại **dependent list.** Có hai cách để lựa chọn biến độc lập, là biến mà dựa váo các giá trị trong nó mà ta phân chia các gia tri trung bình của biến phụ thuộc thành những nhóm nhỏ.

Lựa chọn một hoặc nhiều biến độc lập. Lúc này các kết quả cũng như các đại lượng thống kê kèm theo sẽ được thể hiện trên các bản riêng biệt cho mỗi biến độc lập

Lựa chọn biến độc lập theo lớp, mỗi biến độc lập trong một lớp, lúc này các kết quả và đại lượng thống kê được thể hiện trên chung một bảng.

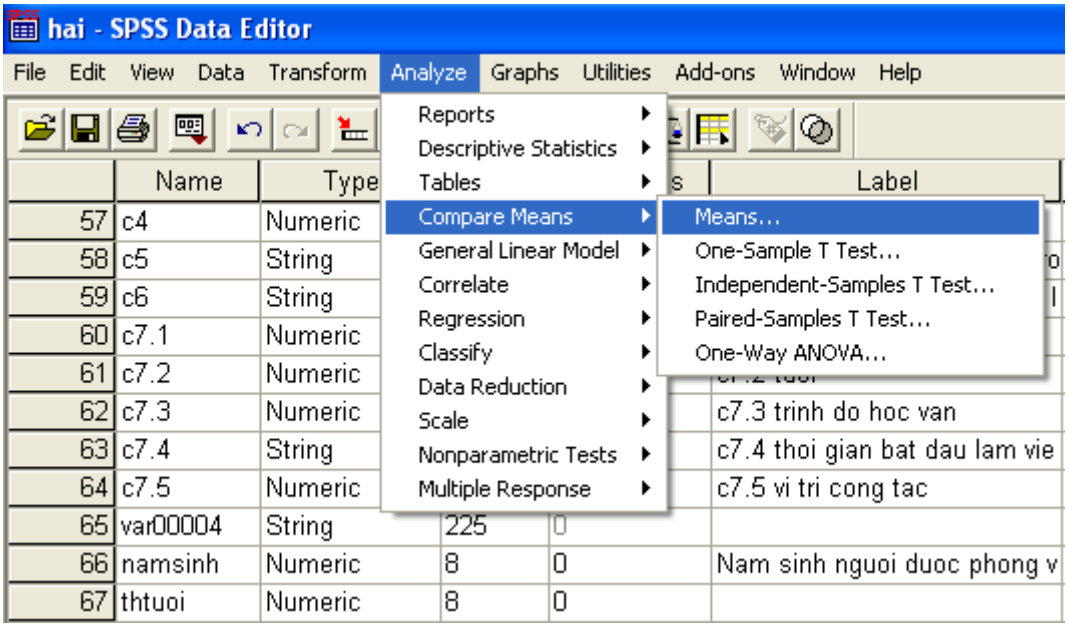
Công cụ **Options** (Hình 7-11). Cho phép ta lựa chọn các đại lượng thống kê cần khảo sát và ANOVA, Eta, và Eta bình phương (sẽ được đề cập chi tiếc về ý nghĩa ở phần sau)



Hình 6.15. Hộp thoại Means: Options

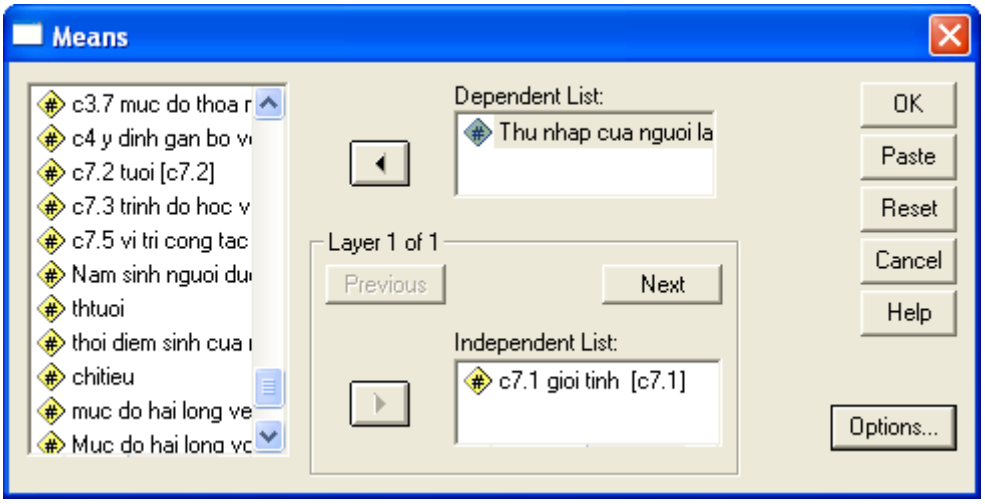
Ví dụ: Chúng ta cần xác định mối quan hệ giữa giới tính và mức thu nhập trung bình của người lao động trong công ty. Các thao tác thực hiện như sau:

Bước 1: Compare Means/Means



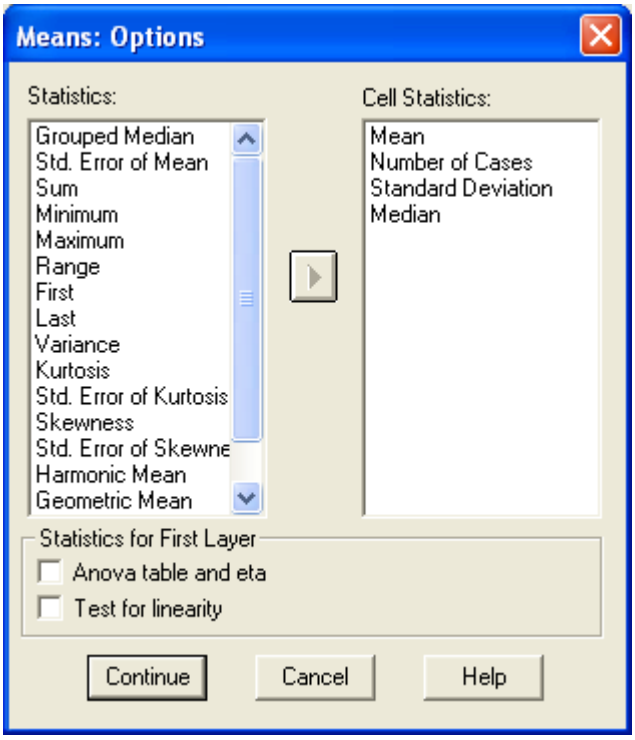
Hình 6.16: Thao tác với Compare Means

Bước 2: Chúng ta cần phải xác định trước biến nào là biến độc lập và biến nào là biến phụ thuộc. Trong trường hợp cụ thể này, chúng ta thấy giới tính có thể là 1 yếu tố có ảnh hưởng đến thu nhập. Do đó chúng ta chọn biến "thunhap" đưa vào hộp Dependent List; biến "gioitinh" vào hộp Independent List.



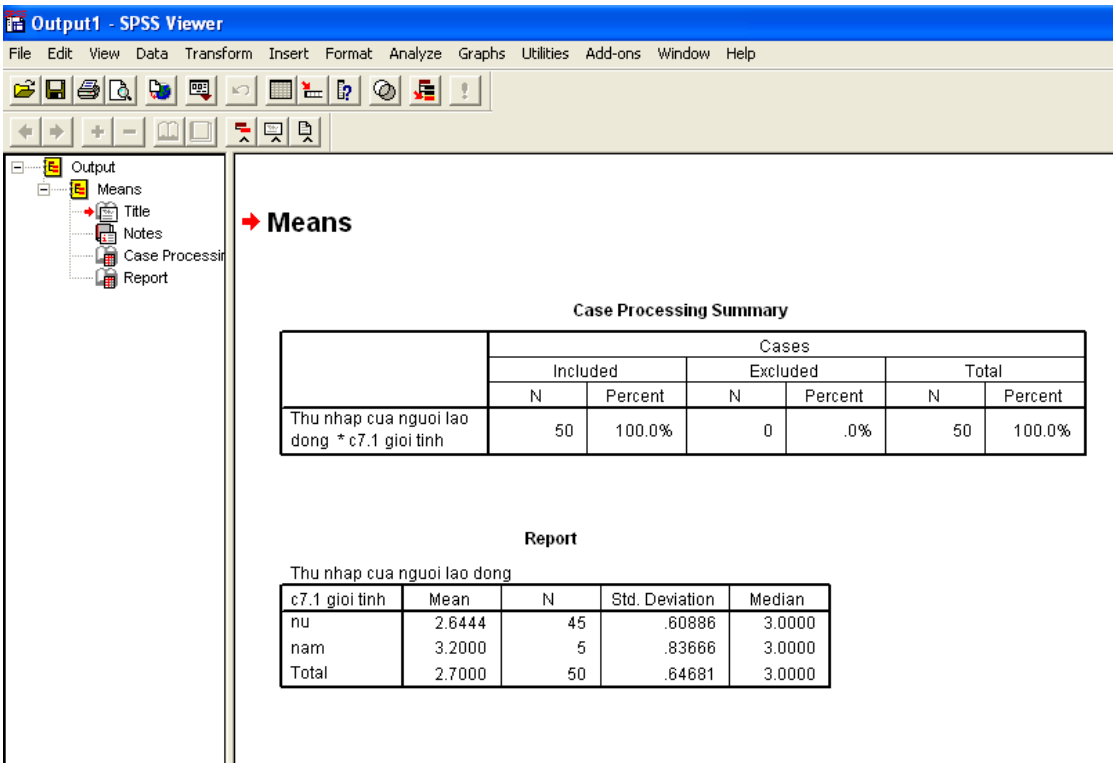
Hình 6.17. Hộp thoại Means

Bước 3: Bấm Opitions đến chọn các đại lượng thống kê rồi nhất Continue/OK để kết thúc.



Hình 6.18. Hộp thoại Means: Options

Bước 4: Đọc kết quả trên màn hình Output



Report

Thu nhap cua nguoi lao dong

c7.1 gioi tinh	Mean	N	Std. Deviation	Median

nu	2.6444	45	.60886	3.0000
nam	3.2000	5	.83666	3.0000
Total	2.7000	50	.64681	3.0000

Hình 6.19. Màn hình Output

N : Số lượng mẫu

Mean : Giá trị trung bình

Std.Devuation ; Độ lệch chuẩn.

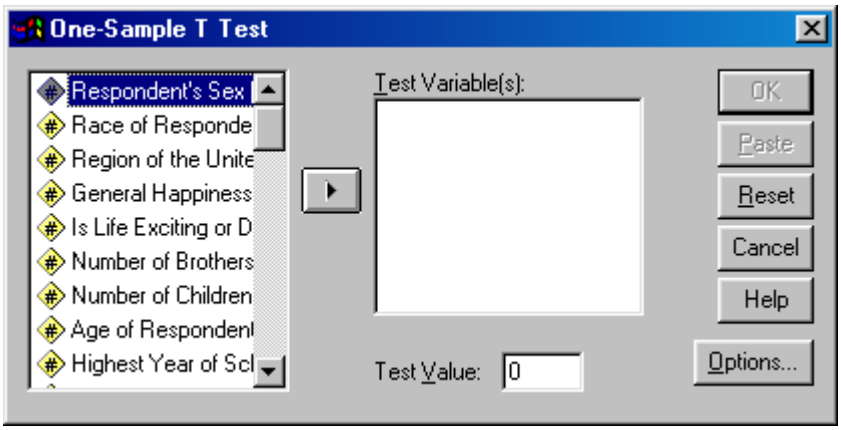
Median : Trung vị

Như vậy, nhìn vào bảng trung bình, chúng ta thấy thu nhập trung bình của nữ giới là 2,64 triệu, trong khi đó thu nhập trung bình của nam là 3,2 triệu. Thu nhập trung bình của cả hai giới là 2,7 triệu. Như vậy thu nhập của nữ công nhân thấp hơn thu nhập của nam công nhân và thấp hơn cả thu nhập trung bình của cả hai giới.

6.2. Kiểm nghiệp t-một mẫu

Phương pháp kiểm nghiệp một mẫu được dùng để kiểm định có hay không sự khác biệt của giá trị trung bình của một biến đơn với một giá trị cụ thể, với giả thuyết ban đầu cho rằng giá trị trung bình cần kiểm nghiệm thì bằng với một con số cụ thể nào đó. Ví dụ một nhà nghiên cứu có thể kiểm định có hay không sự khác biệt giữa chỉ số IQ trung bình của một nhóm sinh viên với chỉ số cụ thể là 100 ở độ tinh cậy là 95%. Phương pháp kiểm nghiệm này dùng cho biến dạng thang đo khoảng cách hay tỉ lệ. Ta sẽ loại bỏ giả thuyết ban đầu khi kiểm nghiệm chó ta chỉ số **Sig.** nhỏ hơn mức tinh cậy (0.05).

Từ Menus ta chọn **Compare Mean\One-Sample T Test…** ta có hộp thoại như hình 6-14



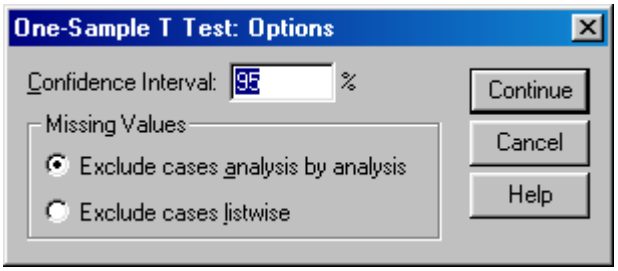
Hình 6.20: Hộp thoại One – sample T-Test

Lựa chọn biến cần so sánh bằng cách di chuyển vệt đen và chuyển đến vào hộp thoại **Test Variable(s),** nhập giá trị cần so sánh vào hộp thoại **Test Value.**

Chọn công cụ **Options** (hình 6-15) để xác định độ tin cậy cho kiểm nghiệm, mặc định là 95% và cách xử lý đối với các giá trị khuyết, Khi kiểm nghiệp các biến ta sẽ gặp một vài giá trị khuyết trong các biến đó, vấn đề ở đây là ta loại bỏ các giá trị khuyết đó trong kiểm nghiệm hay bao hàm luôn tất cả.

Exclude cases analysis by analysis. Mỗi kiểm nghiệm T sử dụng toàn bộ các trường hợp (cases) chứa đựng giá trị có ý nghĩa đối với biến được kiểm nghiệm. Đặc điểm là kích thương mẫu luôn thay đổi.

Exclude cases listwise. Mỗi kiểm nghiệm T sử dụng chỉ những trường hợp có giá trị đối với toàn bộ tất cả các biến được sử dụng trong bất kỳ kiểm nghiệm T test nào. Kích thước mẫu luôn không đổi.



HÌnh 6.21. Hộp thoại One – sample T-Test: Options

Điều kiện để tiến hành một kiểm nghiệm t một mẫu đòi hỏi dữ liệu phải đáp ứng giả định sau: dữ liệu phải là phân phối chuẩn, hoặc kích thước mẫu phải đủ lớn để được xem là xấp xỉ phân phối chuẩn.

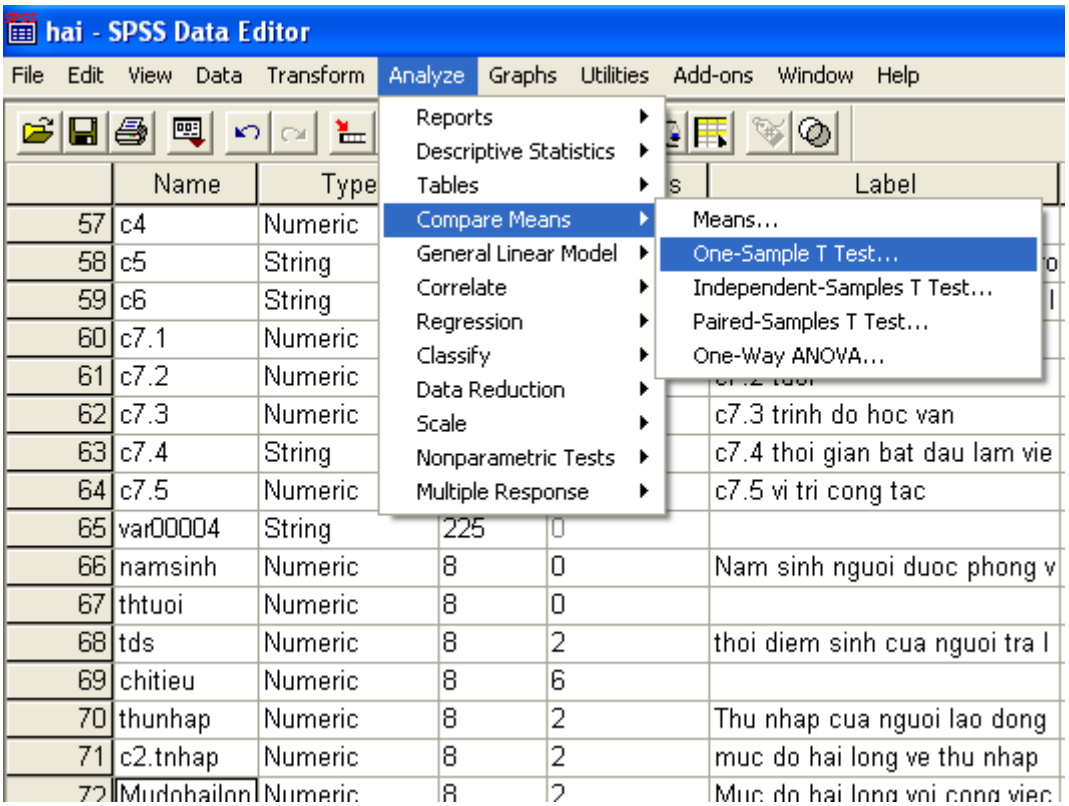
Ví dụ: Chúng ta đã biết thu nhập trung bình của công nhân công ty là 2,7 triệu đồng/tháng (xem hình 6.19), chúng ta có giả thuyết rằng thu nhập này cao hơn thu nhập trung bình của người lao động trong các doanh nghiệp nhà nước khác hiện tại là 2,6 triệu đồng/tháng. Khi đó, giả thuyết của bài toán là:

H_0: $\mu = \mu_0 = 2,6$ triệu đồng/tháng

H_1 : $\mu \neq \mu_0 = 2,6$ triệu đồng/tháng

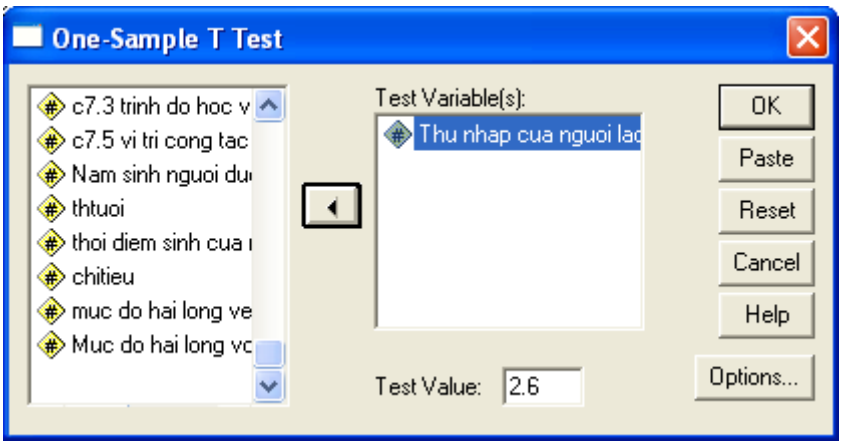
Cách thực hiện như sau:

Bước 1:



Hình 6.22. Thao tác kiểm định One – sample T-Test

Bước 2: Chọn biến "thunhap" bên biến nguồn sang biết đích, chọn giá trị 2,6 đưa vào ô Test Value, nhấn OK để kết thúc. (hộp thoại *One – sample T-Test: Options luôn mặc định mức độ ý nghĩa quan sát là 95%)*



Hình 6.23.

Bước 3: Đọc kết quả

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Thu nhap cua nguoi lao dong	50	2.7000	.64681	.09147

One-Sample Test

	Test Value = 2.6					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Thu nhap cua nguoi lao dong	1.093	49	.280	.10000	-.0838	.2838

Trong kết quả thu được chúng ta thấy được giá trị trung bình, độ lệch chuẩn, sai số chuẩn. Ngoài ra với t = 1.093, p.value = 0,280 > 0,05 cho phép chúng ta kết luận rằng không có cơ sở để bác bỏ giả thuyết H_0 (hay nói cách khác là chấp nhận giả thuyết Ho và bác bỏ giả thuyết H1), điều này có chưa có có sở thống kê để khẳng định thu nhập của người lao động trong công ty là 2,7 triệu là cao hơn thu nhập trung bình của người lao động trong các doanh nghiệp nhà nước.

Điều này cũng giống trường hợp 1 sinh viên A có điểm học lực là 7,2 và 1 sinh viên B có học lựa 7,4. Sự khác biệt này chưa đủ kết luận Sinh viên A học kém hơn sinh viên B, vì trong bảng xếp loại, họ cùng có kết quả học tập ở mức khá.

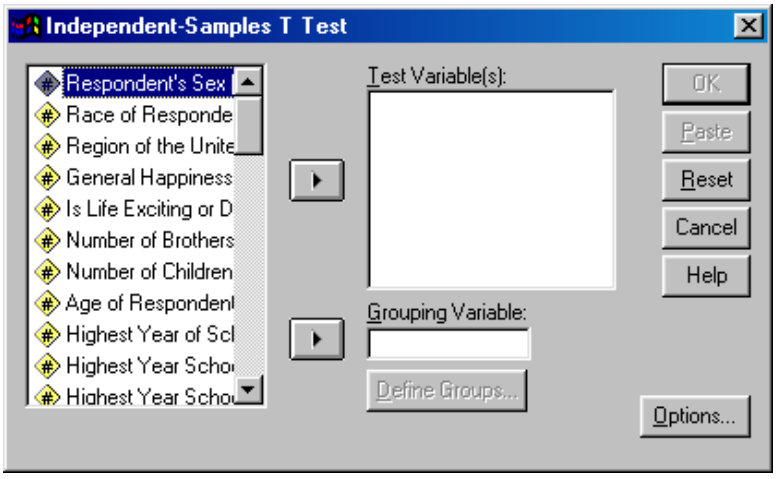
6.3. Kiểm nghiệp t hai mẫu độc lập

Kiểm nghiệm này dùng cho hai mẫu độc lập, dạng dữ liệu là dạng thang đo khoảng cách hoặc tỷ lệ

Đối với dạng kiểm nghiệm này, các chủ thể cần kiểm nghiệm phải được ấn định một cách ngẫu nhiên cho hai nhóm dữ liệu cần nghiên cứu sao cho bất kỳ một khác biệt nào từ kết quả nghiên cứu là do sự tác động của chính nhóm thử đó, chứ không phải do các yếu tố khác. Ví dụ như ta không thể dùng phương pháp này để so sánh thu nhập của nam và nữ bởi vì thu nhập còn bị ảnh hướng lớn bời trình độ học vấn và nghề nghiệp. Hoặc để đánh giá tác động của một chương trình quảng cáo ta lựa chọn ra hai nhóm khách hàng độc lập, nhóm đã xem qua chương trình quảng cáo và nhóm chưa xem qua chương trình quảng cáo để đánh giá mức độ ưa thích của sản phẩm đã được quảng cáo. Ở đây ngoài công cụ thử là việc xem quảng cáo hoặc không xem, nhà nghiên cứu phải bảo đảm không tồn tại yếu tố nào đáng kể tác động đến sự đánh giá về sản phẩm, như giới tính, sự tiêu dùng, trình độ, … Tóm lại để đánh giá giá trị trung bình (về đánh giá sự ưa thích, thu nhập, chi tiêu, …) của hai nhóm độc lập nghĩa là các phản ứng thu được của nhóm này không bị ảnh hưởng bởi nhóm kia và ngoài các tác nhân cần đánh giá cần phải chú ý đến các tác động khác có thể làm thay đổi sự phản ứng thu nhận được giữa hai nhóm.

Các dữ liệu cần so sánh nằm trong cùng một biến định lượng. Để so sánh ta tiến hành nhóm các giá trị thành hai nhóm để tiến hành so sánh. Giả thuyết ban đầu cần kiểm nghiệm là giá trị trung bình của một biến nào đó thì bằng nhau giữa hai nhóm mẫu và chúng ta sẽ từ chối giả thuyết này khi mà chỉ số **Sig.** nhỏ hơn mức ý nghĩa (thường là 0.05)

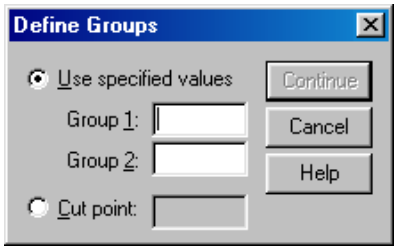
Để thực hiện việc so sánh này ta vào **Compare means\Independent sample t-test….** Từ Menus ta được hộp thoại như hình 6.24:



Hình 6.24

Di chuyển vệt tối vào biến định lượng mà ta cần so sánh giá trị trung bình, chọn bằng cách nhấn nút mũi tên để chuyển biến định lượng đó vào hộp thoại **Test variable(s).** Ta có thể chọn nhiều biến định lượng để so sánh.

Di chuyển vệt tối đến biến dùng để định ra các nhóm cần so sánh với nhau (thường là biến định danh) di chuyển vào hộp thoại **Gouping variable.** Công cụ **Define Groups…** cho phép ta định ra hai nhóm cần so



sánh với nhau, như hình 6.25.

Hình 6.25

Có hai cánh định nhóm so sánh:

Sử dụng con số cụ thể, nhập hai giá trị đại diện cho hai nhóm cần so sánh trong biến vào ô **group 1 và group 2,** ví dụ so sánh thời gian tự học của hai nhóm sinh viên năm nhất và sinh viên năm cuối nằm trong biến loại

sinh viên với 4 nhóm sinh viên được mã hóa như sau sinh viên năm nhất: 1, sinh viên năm hai: 2, sinh viên năm ba: 3, sinh viên năm cuối: 4. Ta nhập giá trị 1 vào Group 1 và nhận giá trị 4 vào group 2. Lúc đó thời gian tự học trung bình sẽ được so sánh giữa hai nhóm sinh viên năm nhất và sinh viên năm cuối.

Cách thứ hai là sử dụng **Cut point,** nhập giá tri phân cách các giá trị trong biến thành hai nhóm. Toàn bộ các trường hợp có giá trị (con số mã hóa) nhỏ hơn giá trị được nhập vào trong **cut point** sẽ định ra một nhóm, và toàn bộ các trường hợp có giá trị mã hóa lớn hơn hoặc bằng giá trị trong **Cut point** sẽ tạo ra một nhóm khác. Ví dụ ta muốn so sánh thời gian tự học của sinh viên hai năm đầu và sinh viên hai năm cuối, ta nhập giá trị 3 (là giá trị mã hóa của nhóm sinh viên năm thứ ba) và **cut point** lúc đó ta tạo được hai nhóm sinh viên bao gồm, sinh viên hai năm đầu (sinh viên năm thứ nhất và sinh viên năm thứ hai) và nhóm sinh viên hai năm cuối (sinh viên năm ba và sinh viên năm cuối) và sẽ tiến hành so sánh số thời gian tự học trung bình trên hai nhóm sinh viên này.

Đối với công cụ **Options** có thao tác và ý nghĩa giống công cụ **Options** đã đề cập trong phần Kiểm nghiệp t một mẫu đã đề cập ở phần trước.

Các giả định phải được thỏa mãn khi dùng kiểm nghiệm t cho hai mẫu độc lập:

Đối với kiểm nghiệm t cho hai mẫu có phương sai bằng nhau (có thể kiểm định giả định này bằng thống kê **Levene**), các quan sát phải độc lập, được lấy ngẫu nhiên từ tổng thể có phân phối chuẩn với phương sai đám đông bằng nhau.

Đối với kiểm nghiệm t cho hai mẫu có phương sai không bằng nhau, các quan sát phải độc lập, được lấy ngẫu nhiên từ tổng thể có phân phối chuẩn.

Công thức tính t:

Với phương sai hợp nhất	Với phương sai riêng biệt
$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$

Với:

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Với x_i: Giá trị trung bình của nhóm i

n_i: Số các quan sát trong nhóm i

S_i: Phương sai mẫu trong nhóm i

Bật tự do trong kiểmnghiệm phương sai hợp nhất bằng

$df = (n_1 + n_2 - 2)$

Bật tự do trong kiểmnghiệm phương sai riêng biệt bằng:

$$df = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

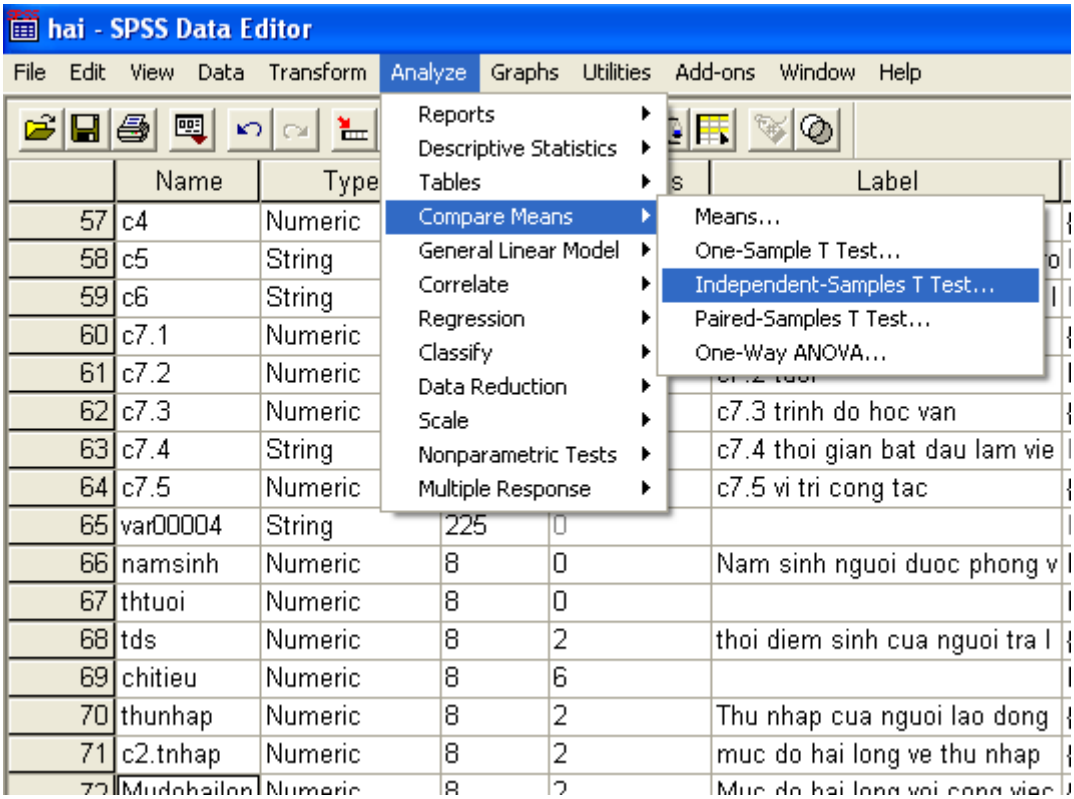
Ví dụ: So sánh thu nhập trung bình giữa nam và nữ có khác nhau hay không? Giả thuyết của bài toán là:

H_0 = Thu nhập của 2 giới là bằng nhau

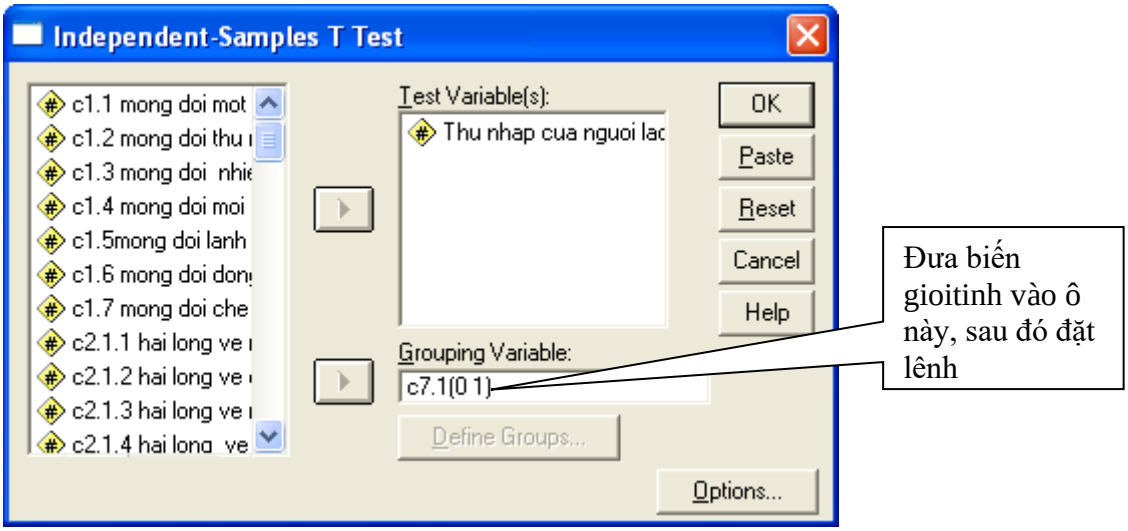
H_1 = Thu nhập của 2 giới là khác nhau

Cách làm như sau:

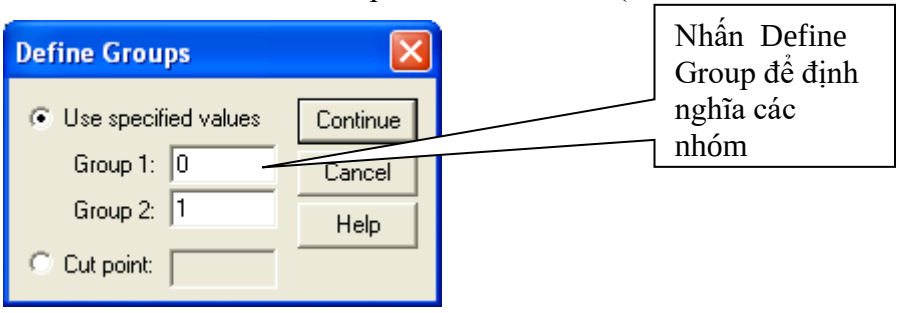
Bước 1:



Bước 2:



Sau đó đặt lệnh Define Group: 0 = nam; 1 = nữ (theo mã code ban đầu)



Group Statistics

	c7.1 gioi tinh	N	Mean	Std. Deviation	Std. Error Mean
Thu nhap cua nguoi lao dong	nu	28	3077.1429	622.43760	117.62965
	nam	22	3043.6364	306.60267	65.36791

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Thu nhap cua nguoi lao dong	Equal variances assumed	4.192	.046	.231	48	.818	33.50649	145.00779	-258.05122	325.06420
	Equal variances not assumed			.249	41.199	.805	33.50649	134.57228	-238.22783	305.24082

Khi Sig trong kiểm định phương sai <0,05 thì dùng kết quả kiểm định t ở dòng thứ 2

Giá trị t kiểm định

P.value của giá trị t

Khi kiểm định Levene's (giả thiết Ho: phương sai của 2 mẫu bằng nhau; giả thiết H1: phương sai hai mẫu không bằng nhau) cho phép kiểm định phương sai hai mẫu có bằng nhau hay không. Nếu sig của F <0,05 ta bác bỏ giả thuyết Ho và chấp nhận giả thuyết H1, có nghĩa phương sai 2 mẫu không bằng nhau do đó ta phải tham chiếu t ở dòng thứ 2. Ngược lại, nếu Sig của F >0,05 thì ta dùng t ở dòng thứ nhất.

Đối với kiểm định t ta nhận thấy là t = 0,29 và p.value = 0,805>0,05 do đó chấp nhận giả thiết Ho và bác bỏ giả thuyết H1. Có nghĩa không có sự chênh lệch thu nhập của 2 giới nam và nữ trong công ty.

6.4. Kiểm nghiệm t theo từng cặp mẫu

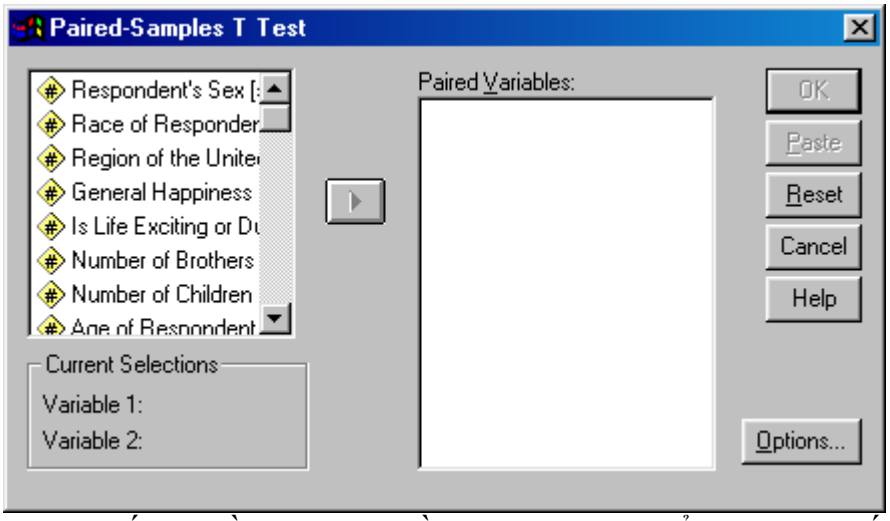
Đây là dạng kiểm nghiệp dùng cho hai biến trong cùng một mẫu có liên hệ với nhau, dữ liệu dạng thang đó khoảng cách hoặc tỷ lệ. Nó tính toán sự khác biệt giữa các giá trị của hai biến cho mỗi trường hợp và kiểm

nghiệm xem giá trị trung bình các khác biệt có khác 0 hay không. Giả thuyết ban đầu được đưa ra là giá trị trung bình của các khác biệt là bằng 0. Và ta sẽ loại bỏ giả thuyết này trong trường hợp kiểm nghiệm cho kết quả **Sig.** nhỏ hơn mức ý nghĩa (0.05)

Lợi điểm của việc sử dụng kiểm nghiệm T theo từng cặp là ta loại trừ được những yếu tố tác động bên ngoài vào nhóm thử. Ví dụ ta khảo sát sự ưa thích của hai loại nước hoa chuẩn bị tung ra thị trường. Kết quả kiểm nghiệp trên cùng một nhóm mẫu sẽ cho những thông tin xác thực hơn về sự ưa thích của mùi vị hai loại nước hoa này, đồng thời tập trung vào sự khác biệt tư nhiên của hai loại nước hoa này. Nếu ta tiến hành so sánh giữa hai nhóm mẫu độc lập với nhau sẽ cho ra những kết quả khác biệt do những tác nhân khác với bản thân sự khác biệt của hai loại nước hoa này như sự khác biệt về con người, về nhận thức, về kinh nghiệm cũng như các yếu tô bên ngoài khác. Phương pháp này thích ứng cho việc kiểm nghiệm sản phẩm. Phương pháp này kiểm nghiệm giả thuyết cho rằng sự khác biệt giữa hai trung bình mẫu là bằng không. Ta từ chối giả thuyết này khi mức ý nghĩa của ta (significante) là nhỏ hơn mức ý nghĩa (thường là 5%).

Điều kiện yêu cầu cho loại kiểm nghiệm này là kích cở hai mẫu so sánh phải bằng nhau. Các quang sát cho mỗi bên so sánh phải được thực hiện trong cùng những điều kiện giống nhau. Các khác biệt từ giá trị trung bình của hai mẫu phải là phân phối chuẩn hoặc số lượng mẫu đủ lớn để xấp xỉ là phân phối chuẩn. Phương sai của mỗi biến là ngang bằng hoặc không ngang bằng (có thể kiểm nghiệm qua phép kiểm nghiệm phương sai Levene).

Để thực hiện việc so sánh này ta vào **Compare means\Paired-samples t-test….** Từ Menus ta được hộp thoại như hình:



Chọn hai biến ta cần so sánh bằng cách di chuyển vệt đen đến lần lược hai biến cần quan sát, di chuyển biến cần quan sát vào hộp thoại **Paired Variables** bằng nút mũi tên. **Paired-samples t test** còn cho ta kết quả về mối tương quan giữa hai biến đang quan sát. Cho biết liệu hai biến này có tương quan với nhau hay không, độ tương quan và chiều tương quan (thể hiện ở bảng **Paired samples correlation**).

Các giả định phải được thỏa mãn khi dùng kiểm nghiệm cặp mẫu là các quan sát ở mỗi cặp phải được thực hiện trong cùng một điều kiện. Những khác biệt giá trị trung bình phải có phân phối chuẩn. Phương sai của mỗi biến có thể ngang bằng hoặc không.

Đối với kiểm nghiệm t các cặp mẫu, SPSS sẽ tính toán giá trị khác biệt giữa hai bến trong từng quan sát và tiến hành kiểm nghiệm giá trị trung bình các khác biệt đó có bằng 0 hay không

Trong kiểm nghiệm hai mẫu độc lập đã đề cập ở phần trước SPSS chia các giá trị của một biến đơn thành hai nhóm dựa trên một biến kiểm soát và sau đó tiến hành so sánh trung bình trong biến đơn giữa hai nhóm đó với nhau. Đối với kiểm nghiệm cặp, giá trị trung bình các giá trị trong hai biến được so sánh với nhau. Kiểm nghiệm loại này được sử dụng để kiểm nghiệm xem trung bình của hai đo lường là khác biệt hay ngang bằng nhau, hay nói

cách khác kiểm nghiệm xem có hay không trung bình của các giá trị khác biệt giữa hai biến trên mỗi trường hợp quan sát là khác 0

Để tiến hành kiểm nghiệm t theo cặp đòi hỏi hai biến trong kiểm nghiệm phải bằng nhau về số lượng mẫu quan sát và có cùng kiểu đo lường và đơn vị đo lường

Công thức tin giá trị kiểm nghiệm t theo cặp được tính như sau:

$$t = \frac{\text{Trung bình các sai biệt giữa hai biến kiểm nghiệm}}{\frac{SD}{\sqrt{n}}}$$

VớiSD: Độ lệch tiêu chuẩn của các sai biệt

n : Số lượng các quan sát (mẫu)

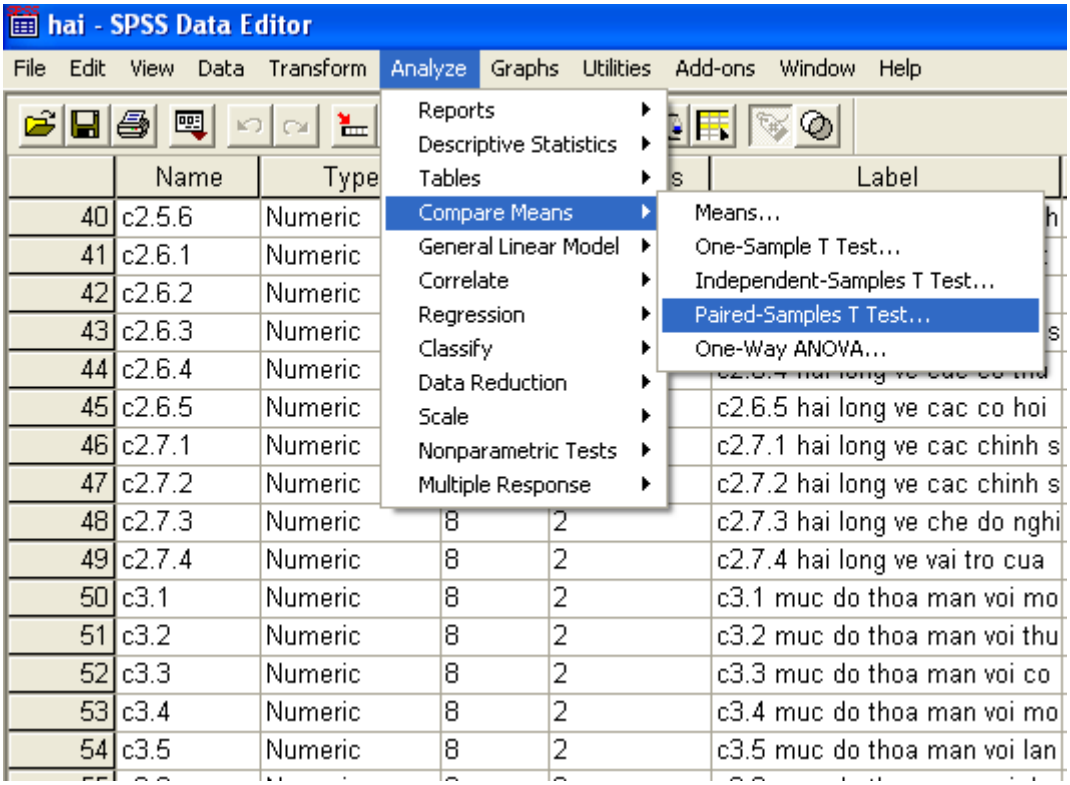
Ví dụ: Chúng ta xem xét liệu có sự khác biệt nào giữa mong đợi và mức độ thỏa mãn mong đợi của người công nhân trước và sau khi vào làm việc ở công ty hay không?

Giả thuyết bài toán là:

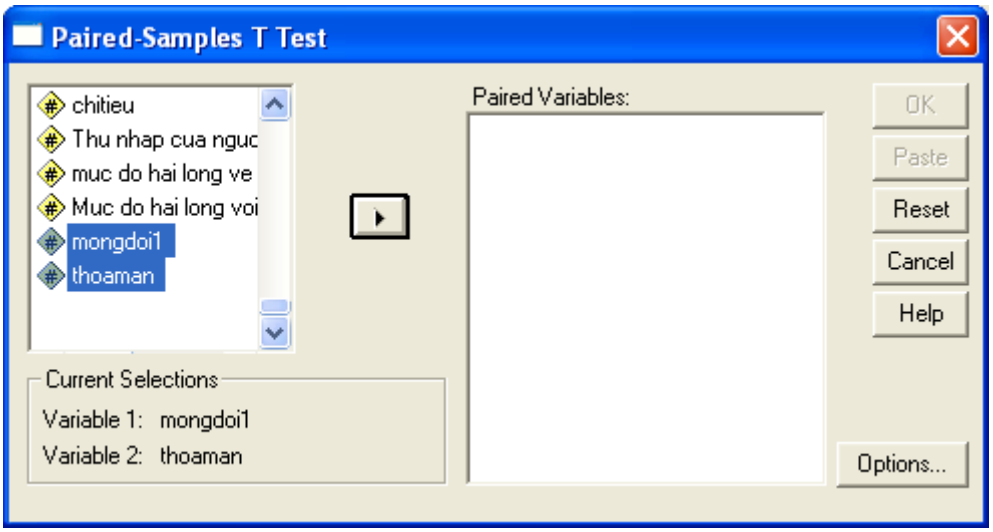
H0 = Không có sự khác biệt giữa mong đợi và mức độ thỏa mãn mong đợi trước và sau khi vào làm việc tại cty

H1 = Có sự khác biệt

Cách làm như sau:



Chọn 2 biến để so sánh: biến "mongdoi1" (mong đợi trước khi vào công ty); biến "thoaman" (mức độ thỏa mãn mong đợi sau khi vào công ty). Chuyển 2 biến này từ biến nguồn sang biến đích. Nhấn OK để kết thúc.



Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	mongdoi1	2.1400	50	.70133	.09918

Điểm trung bình mong đợi và thỏa mãn trước và sau khi vào làm việc

thoaman	2.6714	50	.56502	.07991

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	mongdoi1 & thoaman	50	-.191	.185

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	mongdoi1 - thoaman	-.53143	.98088	.13872	-.81019	-.25266	-3.831	49	.000

Sự khác biệt giữa mức độ mong đợi và mức độ thỏa mãn

Sự khác biệt giữa mức độ mong đợi và mức độ thỏa mãn

Đọc kết quả: Với sự chênh lệch giữa giá trị mong đợi và mức độ thỏa mãn với các điều kiện của công ty là – 0,53, với p.value <0,05 cho phép bác bỏ giả thuyết Ho và chấp nhận giả thuyết H1. Có nghĩa mức độ mong đợi (kỳ vọng) của người lao động trước khi vào công ty là cao hơn so với những gì họ được thỏa mãn.