



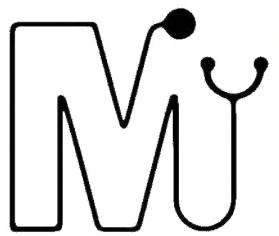
TRƯỜNG ĐẠI HỌC Y DƯỢC HẢI PHÒNG
BỘ MÔN TIN HỌC

GIÁO TRÌNH

TIN HỌC ỨNG DỤNG

Tài liệu lưu hành nội bộ





LỜI NÓI ĐẦU

Tin học được ứng dụng rộng rãi trong tất cả các lĩnh vực khác nhau của xã hội từ khoa học kỹ thuật, y học, kinh tế, công nghệ sản xuất đến khoa học nghệ thuật, ...

Giáo trình *Tin học ứng dụng* được biên soạn dựa trên cơ sở chương trình khung đã được phê duyệt của Trường Đại học Y Dược Hải Phòng.

Giáo trình *Tin học ứng dụng* với việc khai thác và sử dụng phần mềm SPSS 22 sẽ trang bị cho sinh viên những kiến thức và kỹ năng cơ bản về thiết kế cấu trúc bảng lưu dữ liệu, nhập, trình bày, thống kê, phân tích số liệu...

Giáo trình gồm 6 bài được trình bày khoa học từ cơ bản đến nâng cao về việc ứng dụng phần mềm SPSS 22.

Bài 1. Thiết kế và nhập dữ liệu

Bài 2. Các thao tác với biến

Bài 3. Trình bày dữ liệu

Bài 4. So sánh các tỷ lệ và kiểm định tính độc lập

Bài 5. Kiểm định phân phối chuẩn và kiểm định trung bình

Bài 6. Tương quan và hồi quy tuyến tính

Tài liệu được sử dụng trong chương trình chính khóa, phục vụ các đối tượng là sinh viên Đại học, học viên Sau Đại học, cán bộ nghiên cứu và giảng viên.

Mặc dù Giáo trình đã được chỉnh sửa và cập nhật xong không tránh khỏi những khiếm khuyết. Chúng tôi rất mong nhận được các ý kiến đóng góp của đồng nghiệp, học viên và bạn đọc để Giáo trình hoàn thiện hơn.



MỤC LỤC

BÀI 1. THIẾT KẾ VÀ NHẬP DỮ LIỆU.....	7
1.1. Giới thiệu về phần mềm SPSS.....	7
1.2. Các thao tác cơ bản với SPSS 22.....	8
1.2.1. Khởi động	8
1.2.2. Thoát khỏi SPSS	10
1.3. Cơ sở dữ liệu trên SPSS	10
1.4. Khái niệm biến và các thuộc tính của biến.....	10
1.4.1. Biến số (Variable).....	10
1.4.2. Trường hợp/bản ghi (Case).....	12
1.5. Tạo file dữ liệu	12
1.5.1. Cách tạo file dữ liệu.....	12
1.5.2. Lưu file dữ liệu	14
1.6. Một số thao tác với dữ liệu	15
1.6.1. Xem lại dữ liệu	15
1.6.2. Sửa dữ liệu	15
Bài 2. CÁC THAO TÁC VỚI BIẾN	16
2.1. Tính giá trị cho biến.....	16
2.2. Đếm số lần xuất hiện của các giá trị trong từng trường hợp	19
2.3. Mã hóa dữ liệu	20
2.3.1. Mã hóa lại dữ liệu của biến	20
2.3.2. Mã hóa dữ liệu của một biến vào biến mới	22
2.4. Sắp xếp, lọc và tìm kiếm thông tin	23
2.4.1. Sắp xếp dữ liệu	23
2.4.2. Lọc dữ liệu.....	25
2.4.3. Tìm kiếm và thay thế dữ liệu.....	27
2.5. Chia dữ liệu trong tệp thành các nhóm.....	27
2.6. Nối tệp tin dữ liệu	29
2.6.1. Nối thêm hàng	29
2.6.2. Nối thêm biến (Variables)	30
Bài 3. TRÌNH BÀY DỮ LIỆU	32
3.1. Lập bảng tần số các giá trị của biến.....	32
3.2. Tính các đại lượng trong thống kê mô tả.....	34
3.2.1. Lệnh Frequencies.....	34
3.2.2. Lệnh Descriptives	37
3.2.3. Lệnh Explore	38
3.2.4. Lệnh Mean	43
3.3. Lập các bảng tổng hợp nhiều biến	45
3.3.1. Lệnh Crosstabs	45
3.3.2. Lệnh Custom Tables	46
3.3.3. Lệnh Case Summaries	48
3.4. Vẽ biểu đồ trong SPSS	52
3.4.1. Biểu đồ cột - Bar.....	52

3.4.2. Đồ thị dạng đường - Line.....	57
3.4.3. Biểu đồ tròn – Pie	58
3.4.4. Biểu đồ chấm điểm –Scatter/Dot.....	60
3.4.5. Biểu đồ hộp – Boxplot.....	61
3.4.6. Biểu đồ tần suất – Histogram.....	62
Bài 4. SO SÁNH TỶ LỆ VÀ KIỂM ĐỊNH TÍNH ĐỘC LẬP	64
4.1. Cơ sở lý thuyết của Kiểm định Khi bình phương.....	65
4.2. So sánh tỷ lệ.....	66
4.3.1. So sánh các tỷ lệ	66
4.3.2. So sánh tỉ lệ một mẫu với một tỉ lệ lý thuyết.....	68
4.3. Kiểm định tính độc lập	70
4.4. Nguy cơ tương đối (Relative Risk) và Tỷ suất chênh lệch (Odds Ratio).....	74
4.4.1. Cơ sở lý thuyết.....	74
4.4.2. Ví dụ minh họa	75
Bài 5. KIỂM ĐỊNH PHÂN PHỐI CHUẨN & KIỂM ĐỊNH TRUNG BÌNH	77
5.1. Kiểm định phân phối chuẩn.....	77
5.2. Kiểm định trung bình.....	85
5.2.1. So sánh trung bình quan sát và trung bình lý thuyết.....	86
5.2.2. So sánh trung bình quan sát của hai nhóm độc lập.....	88
5.2.3. So sánh đồng thời nhiều trung bình.....	89
5.2.4. So sánh ghép cặp	93
Bài 6. TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH	96
6.1. Tương quan tuyến tính.....	96
6.1.1. Cơ sở lý thuyết.....	96
6.1.2. Tính hтенh hệ số tương quan tuyến tí	97
6.2. Hồi quy tuyến tính	100
TÀI LIỆU THAM KHẢO.....	102
PHỤ LỤC	103

BÀI 1. THIẾT KẾ VÀ NHẬP DỮ LIỆU

Mục tiêu:

- Trình bày và thực hiện được một số thao tác cơ bản trong giao diện của SPSS.
- Trình bày được các loại biến (trường).
- Khai báo, thiết lập và nhập dữ liệu.
- Trình bày một số thao tác với dữ liệu.

1.1. Giới thiệu về phần mềm SPSS

* Khái niệm

SPSS (*viết tắt của Statistical Package for the Social Sciences*) là một phần mềm máy tính phục vụ công tác phân tích thống kê; SPSS được các nhà nghiên cứu sử dụng rộng rãi trong nhiều lĩnh vực.

SPSS được thiết kế để thực hiện tất cả các bước trong các phân tích số liệu từ thống kê mô tả (liệt kê dữ liệu, lập biểu đồ) đến thống kê suy luận (tương quan, hồi quy...).

Phần mềm SPSS có chức năng của một hệ thống *quản lý dữ liệu* và *phân tích thống kê*, với giao diện thân thiện cho người dùng trong môi trường đồ họa và các hộp thoại đơn giản.

Thế hệ đầu tiên của SPSS được đưa ra từ năm 1968 tại Đại học Stanford. Năm 1975, Công ty SPSS Inc thành lập để thương mại hóa phần mềm này. Phần mềm làm việc trên nền tảng MS-DOS (1984), Windows 3.1 (1992). Phiên bản 18 dùng cho các hệ điều hành Windows, Mac, Linux / UNIX. Ngày 28/7/2009, IBM mua lại phần mềm với giá 1,2 tỷ đô la từ công ty PASW (Predictive Analytics SoftWare Statistics). Đến 1/2010, nó trở thành "SPSS: An IBM Company". IBM đã xuất bản các phiên bản SPSS: 19, 20, 21, 22, 23 vào các năm 2010, 2011, 2012, 2013, 2015.

* Chức năng chính của SPSS

- 1 - Nhập và làm sạch dữ liệu.
- 2 - Xử lý, biến đổi và quản lý dữ liệu.
- 3 - Tóm tắt, tổng hợp dữ liệu và trình bày dưới dạng biểu bảng, biểu đồ, bản đồ.
- 4 - Phân tích dữ liệu, tính toán các tham số thống kê và diễn giải kết quả.

* Cấu trúc, tổ chức dữ liệu trong SPSS

SPSS tổ chức các file dưới dạng định dạng riêng (có thể trao đổi – nhập và xuất sang các định dạng khác) và gồm các cấu trúc file như sau:

File dữ liệu: *.sav. File kết quả: *.spv

Các định dạng dữ liệu khác mà SPSS có thể đọc:

.Bảng tính – Excel (*.xls, *.xlsx); Lotus (*.w*); Database – dbase (*.dbf);

.ASCII text (*.txt, *.dat); Các tập tin từ các phần mềm thống kê khác (Stata, SAS).

* Một số ứng dụng chính của SPSS

SPSS có thể là dù để giúp các nhà khoa học thực hiện việc xử lý số liệu nghiên cứu nói chung và ứng dụng trong nghiên cứu các mảng chuyên ngành khác nhau như:

- Nghiên cứu tâm lý học: tâm lý tội phạm, tâm lý học sinh-sinh viên...
- Nghiên cứu xã hội học: ý kiến của người dân trong việc xây dựng lại khu chung cư, thống kê y tế...
- Nghiên cứu thị trường: nghiên cứu và định hướng phát triển sản phẩm, mở rộng thị trường; sự hài lòng của khách hàng...
- Nghiên cứu đa dạng sinh học, trong phát triển nông - lâm nghiệp...

Với SPSS, bạn có thể phân tích được thực trạng, tìm ra nhân tố ảnh hưởng, dự đoán được xu hướng xảy ra tiếp theo, giúp bạn đưa ra các quyết định một cách chính xác, giải quyết các vấn đề một cách nhanh chóng và cải thiện kết quả tốt hơn.

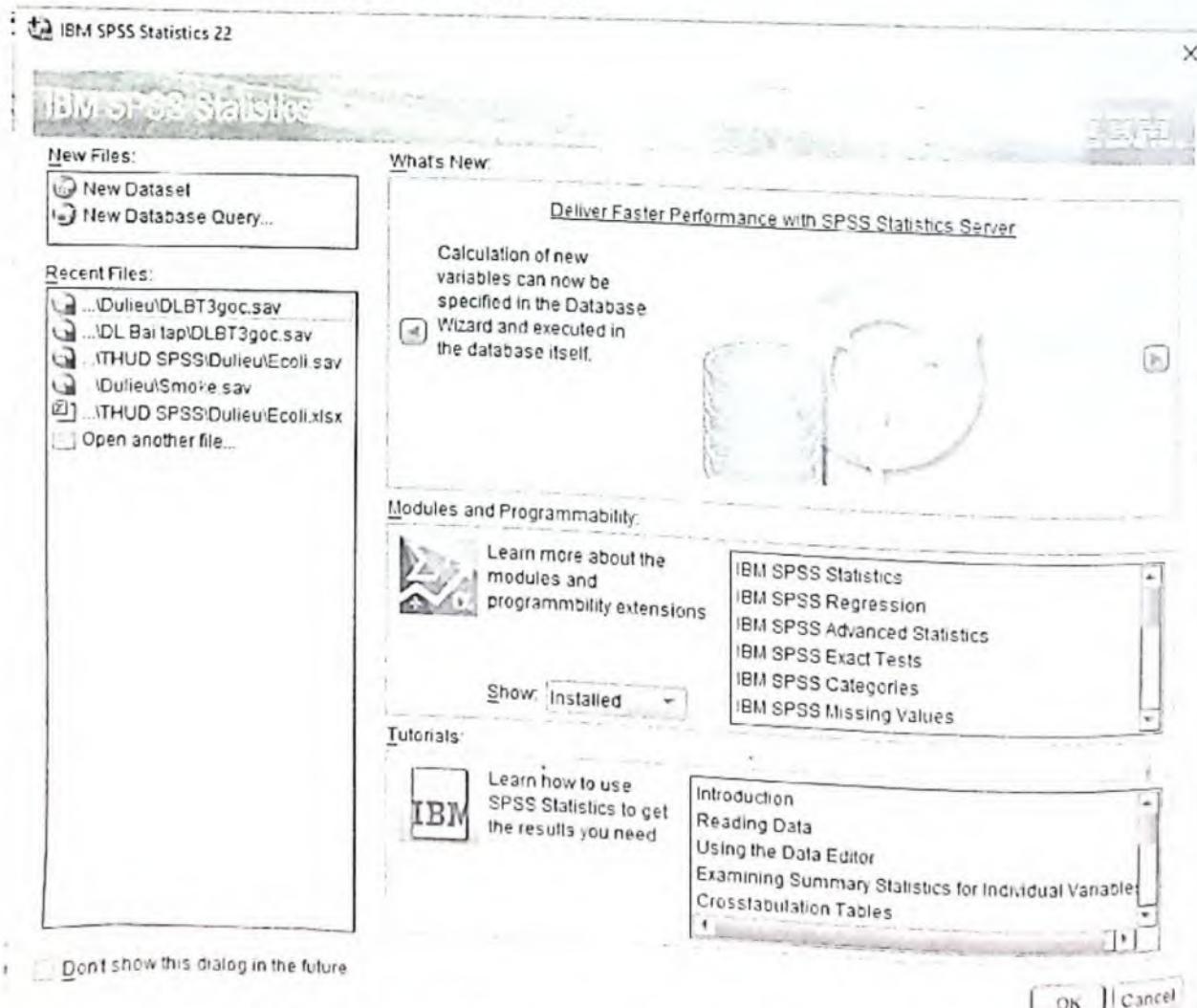
1.2. Các thao tác cơ bản với SPSS 22

1.2.1. Khởi động

Khởi động SPSS bằng cách nháy đúp chuột vào biểu tượng

trên Desktop hay tìm

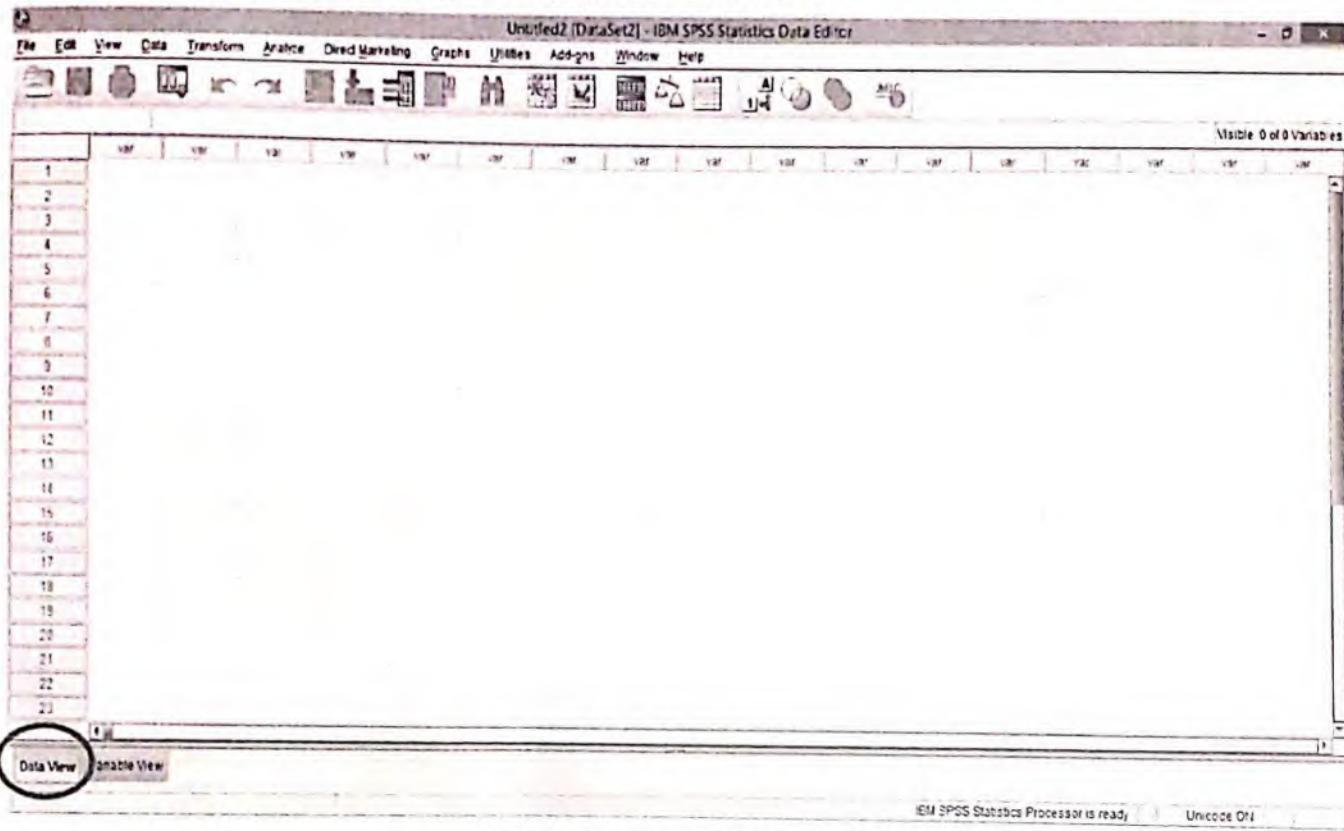
trong All Programs. Xuất hiện cửa sổ - *Hình 1.1*:



Hình 1.1. Cửa sổ khởi động chương trình SPSS 22

- Trong *Recent File* có thể chọn tệp dữ liệu đã được mở trước đó hoặc chọn *Open another file...* để mở tệp dữ liệu khác.
- Chọn **OK** hoặc **Cancel** để trở về màn hình làm việc.

Cửa sổ làm việc của SPSS cũng giống như các phần mềm trên Windows khác bao gồm các menu, các nút lệnh, một số thông báo trạng thái – *Hình 1.2:*



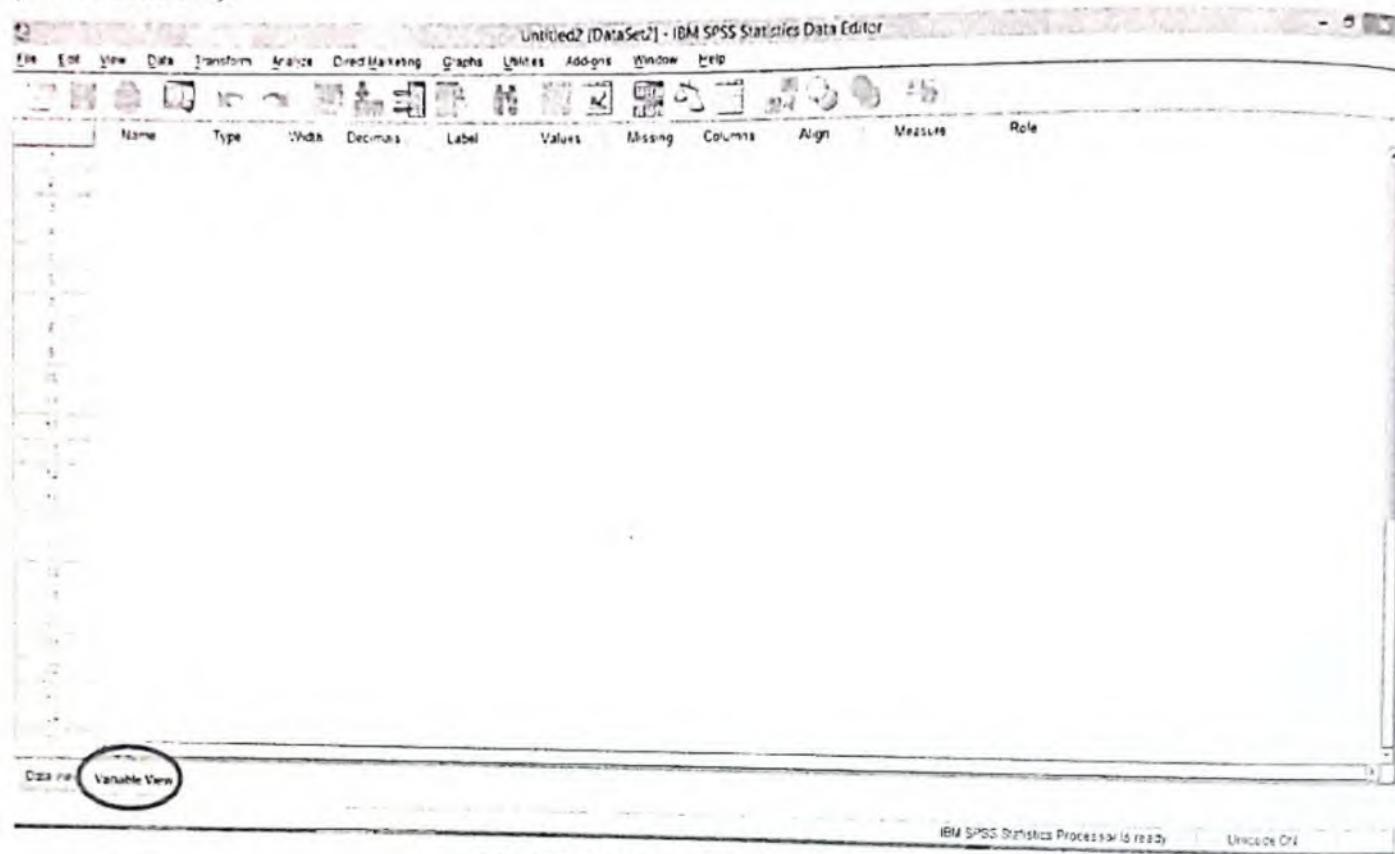
Hình 1.2. Cửa sổ Data View

* Một số menu thường xuyên sử dụng:

- *File* : Tạo file mới, mở file đã có, lưu, in ấn, thoát khỏi SPSS...
- *Edit* : Copy, Cut, Paste, tìm kiếm,...
- *View* : Thay đổi chế độ hiển thị.
- *Transform* : Chuyển đổi dữ liệu, tính toán, mã hóa biến,...
- *Analyze* : Các thống kê cơ bản phân tích dữ liệu.
- *Graphs* : Biểu đồ, đồ thị.
- *Utilities* : Thông tin về các biến, các file, các tiện ích khác.
- *Windows* : Các vấn đề liên quan đến các cửa sổ đang hoạt động, chuyển cửa sổ,...

* Cửa sổ làm việc của SPSS: Data View và Variable View

- **Data View** : Hiển thị các giá trị dữ liệu (Xem Hình 1.2).
- **Variable View** : Hiển thị thông tin khai báo các biến (tên, kiểu, nhãn, giá trị của biến) (Xem Hình 1.3).



Hình 1.3. Cửa sổ Variable View

1.2.2. Thoát khỏi SPSS

Để thoát khỏi SPSS ta thực hiện: Chọn **File** → **Exit**

1.3. Cơ sở dữ liệu trên SPSS

- Khái niệm về cơ sở dữ liệu (CSDL - Database): là tập hợp các dữ liệu có liên quan đến nhau, chứa thông tin của một đối tượng nào đó (trường học, ngân hàng, bệnh viện...) được lưu trữ trên máy tính để đáp ứng nhu cầu khai thác thông tin của nhiều người sử dụng với nhiều mục đích khác nhau.

- Trong SPSS, cơ sở dữ liệu là tập hợp các dữ liệu được tổ chức theo cấu trúc có dạng bảng gồm cột và hàng, mỗi cột như vậy được gọi là một biến (variable) và các hàng (case) để mô tả thông tin của các đối tượng hoặc một lớp các đối tượng. Trong đó, một variable được đặc trưng bởi: Tên (Name), Kiểu dữ liệu (Type), Độ rộng (Width)... và một case là các thông tin liên quan với nhau trên cùng một hàng. Ví dụ: danh sách nhân viên, danh sách hàng hóa, danh sách bệnh nhân...

1.4. Khái niệm biến và các thuộc tính của biến

1.4.1. Biến số (Variable)

Khi chúng ta quan sát/tìm hiểu một vấn đề nào đó, chúng ta thấy có rất nhiều đặc tính mà chúng ta có thể quan sát; chẳng hạn, khi quan sát con người: Họ tên, giới tính, quê quán, chiều

cao, cân nặng, trình độ, ... Chúng ta thường đặt tên cho các đặc tính đó và gọi chúng là các *Biến số* (*Variable*) – thường gọi tắt là *Biến*.

Biến số: là đặc tính của người, sự vật, hiện tượng biến thiên theo các đối tượng khác nhau và điều kiện khác nhau. Biến số do người nghiên cứu lựa chọn phù hợp với mục tiêu nghiên cứu.

* **Phân loại biến:** có 2 loại chính

- **Biến định tính:** là biến miêu tả các giá trị đo lường bằng các chữ, chữ số hay ký hiệu được xếp vào các nhóm khác nhau. Ví dụ: Giới tính là một biến định tính có 2 giá trị là nam và nữ; Trình độ học vấn: mù chữ, tiểu học, trung học, cao đẳng, đại học...; Thu nhập: thấp, trung bình, khá, cao...

- **Biến định lượng:** là biến số miêu tả các giá trị đo lường là các con số. Ví dụ: Chiều cao là một biến số định lượng có thể có những giá trị như 1m, 1,5m, v.v. hoặc số lượng tài sản có trong trường học: bàn, ghế, máy tính...

Khác nhau cơ bản giữa hai biến định tính và biến định lượng:

- Biến định tính: phản ánh tính chất, sự hơn kém, không tính được giá trị trung bình.

- Biến định lượng: phản ánh mức độ, mức độ hơn kém, tính được giá trị trung bình.

* **Các loại thang đo:**

↪ + **Thang đo danh nghĩa** (*Thang đo phân loại*) – *Nominal scale*: Con số chỉ để phân loại các đối tượng, chúng không mang ý nghĩa nào khác, do đó, mọi phép tính đại số giữa chúng đều vô nghĩa. Trong loại thang đo này ta sử dụng biến danh nghĩa (*Nominal Variable*), là biến định tính có từ hai giá trị trở lên, các giá trị không thể biểu diễn bằng số mà thường được biểu diễn bằng các tên gọi (định danh), bàn thân chúng cũng không sắp xếp theo một trật tự từ thấp đến cao.

Ví dụ: Biến phân loại các loại dân tộc khác nhau hoặc các biến nhị phân chỉ nhận một trong hai giá trị như Nam – Nữ, Sốt – Không sốt, Ngô độc – Không ngô độc.

↪ + **Thang đo thứ bậc** - *Ordinal scale*: Các con số dùng để ghi thứ bậc (*Hơn kém*). Với dạng thang đo này chúng ta không thể xác định được mức độ hơn kém nhau giữa các nhóm giá trị. Trong loại thang đo này ta sử dụng biến thứ bậc (*Ordinal Variable*), là biến danh nghĩa có thể sắp xếp thứ tự được.

Ví dụ: Thu nhập: thấp, trung bình, khá, cao..., Học lực của sinh viên: Xuất sắc, Giỏi, Khá, Trung bình, Yếu hoặc Trình độ học vấn: Mù chữ, tiểu học, trung học, cao đẳng, đại học, trên đại học...;

↪ + **Thang đo khoảng cách** (*Interval*): Giống như đặc tính của thang đo thứ tự, tuy nhiên đổi với thang đo khoảng cách cho phép ta biết được khoảng cách giữa các nhóm giá trị.

Ví dụ 1: Chúng ta biết được sự khác biệt giữa 20 và 30 cũng bằng với sự khác biệt giữa 30 và 40 đều ở một khoảng là 10.

Ví dụ 2: Chia Tuổi thành các nhóm: Nhóm 1 gồm những người dưới 20 tuổi, Nhóm 2 từ 20 đến 30 tuổi, Nhóm 3 từ 31 đến 40, ...

Dể thực hiện được điều này chúng ta dựa vào đơn vị đo khoảng cách và giá trị xuất phát để so sánh (hay còn gọi là giá trị 0), cả hai giá trị này đều mang tính trắc lệc. Dạng thang đo

Hình 1.4. Cửa sổ Variable View hiển thị danh sách và thuộc tính các biến

- Sau đó khai báo các biến:

+ *Tên (Name)*: Mỗi tên biến phải là duy nhất trong một tệp. Tên biến có thể dài tới 64 ký tự (byte), các ký tự đầu tiên phải là một chữ hoặc @; ký tự tiếp theo là bất kỳ sự kết hợp giữa chữ cái và số; Tên biến không dấu cách, không trùng với từ khóa. *Ví dụ*: HOTEN, DIACHI, TUOI, PL1,...

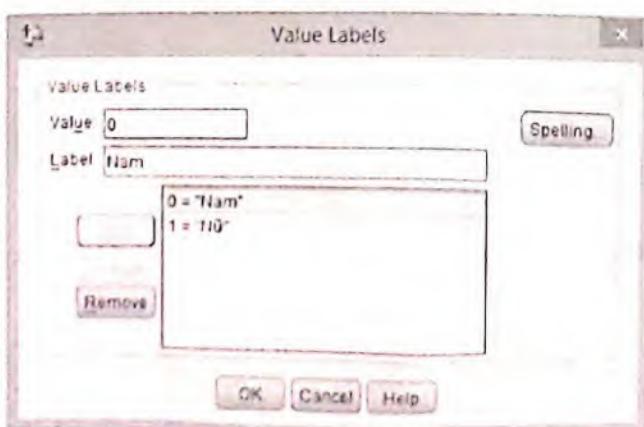
+ *Kiểu dữ liệu (Type)*: Chọn một trong số các kiểu dữ liệu có sẵn của SPSS như *Numeric (số)*, *Date (ngày tháng)*, *String (chuỗi)*, *Comma*, *Dot*,...

+ *Độ rộng (Width)*: là số ký tự nhập được tối đa, chọn phù hợp với kiểu.

+ *Phản thập phân (Decimals)*: quy định chiều dài phản thập phân của biến kiểu số.

+ *Nhãn (Label)*: Thường ngắn gọn, dùng để mô tả, giải thích cho tên biến với 256 ký tự bao gồm cả dấu cách. Khi hiển thị hoặc in bảng thì nhãn sẽ hiển thị thay cho tên biến.

+ *Giá trị của biến (Values)*: Dùng để mô tả các giá trị của biến định tính. Có thể dùng phần này để mã hóa các giá trị cho biến. Nhấp chuột tại ô *Values* tương ứng, nhấp tiếp ở phần , sẽ xuất hiện cửa sổ để định nghĩa các *Values* (xem Hình 1.5). *Ví dụ*: Giới tính 0 là Nam; 1 là Nữ.



Hình 1.5. Cửa sổ Values Labels – mã hóa dữ liệu

Bộ môn Tin học – Trường Đại học Y Dược Hải Phòng

+ **Khuyết thiếu (Missing):** Cho phép khai báo những dữ liệu được xem là *khuyết thiếu*, có hai loại giá trị khuyết: Khuyết thiếu của hệ thống do bản thân khi thu thập số liệu không có, loại thứ hai là số liệu bị nghi ngờ có sai sót khi thu thập. Thông thường gặp dạng lỗi thứ 2 người ta thường điền vào một con số khác biệt để sau này sẽ xem xét lại. Khi xử lý, cả hai loại giá trị khuyết thiếu này sẽ đều bị bỏ qua không xử lý.

+ **Độ rộng của cột (Columns):** độ rộng của cột trên màn hình nhập dữ liệu.

+ **Căn lề (Align):** căn dữ liệu trong ô khi hiển thị trên màn hình nhập dữ liệu.

+ **Thang đo (Measure):**

. **Scale:** là một biến kiểu số. *Ví dụ:* Tuổi, chiều cao, cân nặng...

. **Ordinal:** những giá trị có thứ bậc nhưng khoảng cách giữa các giá trị không rõ.

Ví dụ: Đánh giá hướng dẫn sử dụng thuốc được chia như sau:

0: Không đạt ; 1: Trung bình; 2: Khá, 3: Tốt.

. **Nominal:** những giá trị này không có thứ bậc hơn kém, nó chỉ là qui ước để dễ thống kê.

Ví dụ: Giới tính của một người: 0- Nam, 1-Nữ.

• Bước 2: Nhập dữ liệu – Xem Hình 1.6.

STT	HỌ TÊN	NGÀY SINH	GIOI	DẠCH	TRÌNH ĐỘ	NAM/NỮ	VIỆT NAM	SŘO	THUỐC PHỤ NỮ	VIỆT NAM	THUỐC ĐA
1	1 CAO DUC LUC	31-Dec-1974	0 AN DƯƠNG	YSI	1989	0	2	2	2	2	-
2	2 CHU THI LOAN	01-Aug-1974	1 AN DƯƠNG	BACSI	2001	1	2	1	1	2	-
3	3 DANG THI HONG	03-Feb-1971	1 AN DƯƠNG	BACSI	1989	2	2	2	2	2	-
4	4 DAO THI CANH	01-Feb-1968	1 THUY NGUYEN	BACSI	1981	0	1	0	0	2	-
5	5 DAO TIEU TANH	15-May-1967	1 THUY NGUYEN	BACSI	1990	1	2	1	1	1	-
6	6 DINH QUANG HUNG	24-Nov-1960	0 AN DƯƠNG	BACSI	1999	2	2	2	2	2	-
7	7 DINH THI TRANG	25-Jan-1970	1 THUY NGUYEN	YSI	1999	2	2	1	1	2	-
8	8 DUAN THI TO	01-Feb-1967	1 THUY NGUYEN	YSI	2001	1	2	1	1	2	-
9	9 HOANG THI SANG	13-May-1971	1 THUY NGUYEN	BACSI	1990	0	0	0	0	1	-
10	10 LE CONG JING	02-Sep-1958	0 AN DƯƠNG	BACSI	2002	1	2	1	1	2	-
11	11 LE THI HOAN	19-Aug-1966	1 THUY NGUYEN	BACSI	1989	2	0	1	1	2	-
12	12 LUONG THI NHANH	03-Feb-1972	1 THUY NGUYEN	YSI	1984	2	0	1	1	0	-
13	13 LUU THI QUYEN	19-May-1968	1 AN DƯƠNG	BACSI	1989	0	0	0	0	2	-
14	14 LUU VAN THUY	07-May-1968	0 AN DƯƠNG	BACSI	1990	1	1	1	1	1	-
15	15 NGO VĂN PHONG	07-Nov-1969	0 AN DƯƠNG	YSI	1989	0	2	2	2	1	-
16	16 NGUYEN THI HANH	20-Nov-1962	1 THUY NGUYEN	YSI	2001	1	2	2	2	1	-
17	17 NGUYEN THI LIEN	20-Oct-1967	1 THUY NGUYEN	YSI	2007	1	1	1	1	1	-
18	18 NGUYEN THI QUE	22-Jan-1967	1 AN DƯƠNG	YSI	2001	1	1	1	1	1	-
19	19 NGUYEN THI TOI	17-Dec-1973	1 AN DƯƠNG	BACSI	1997	1	1	2	1	1	-
20	20 PHAM DUY KHATHINH	21-Jul-1963	0 THUY NGUYEN	YSI	1999	0	0	0	0	1	-
21	21 PHAM THE DIEN	28-May-1966	0 AN DƯƠNG	YSI	1995	2	2	2	2	1	-
22	22 PHAM THE HIEN	25-May-1962	1 THUY NGUYEN	YSI	1993	2	1	2	1	1	-

Hình 1.6. Màn hình nhập dữ liệu trên

- Chuyển màn hình làm việc của SPSS về cửa sổ **Data View**.
- Đặt con trỏ vào từng cột nhập, nhập dữ liệu từ bàn phím.
- Nhấn phím **Enter** để kết thúc nhập cho từng ô.

1.5.2. Lưu file dữ liệu

File → Save rồi đặt tên theo đường dẫn hoặc có thể dùng **Save As** để lưu trữ theo đường dẫn khác.

1.6. Một số thao tác với dữ liệu

1.6.1. Xem lại dữ liệu

- Chọn cửa sổ **Data View** hoặc chọn **View → Value Label**: để xem các trường hợp (case).
- Chọn cửa sổ **Variable View**: để xem các biến.

1.6.2. Sửa dữ liệu

1.6.2.1. Sửa trường hợp (case)

- Chọn cửa sổ **Data View**
- Sửa nội dung của trường hợp (case):
 - + Đặt con trỏ vào ô cần sửa, nhấn phím F2.
 - + Nhập nội dung cần thay đổi từ bàn phím, kết thúc bằng phím Enter.
- Thêm một *case* mới:
 - + Chọn vị trí cần thêm
 - + Chọn **Edit → Insert Cases**
- Xóa một *case*:
 - + Chọn *case* cần xóa
 - + Chọn **Edit → Clear** (nhấn phím **Delete** trên bàn phím)

1.6.2.2. Sửa biến (Variable):

- Chọn cửa sổ **Variable View**
- Sửa tên biến:
 - + Đặt con trỏ vào tên biến (name) cần thay đổi, nhấn phím F2
 - + Dánh máy, kết thúc bằng phím Enter
- Sửa kiểu dữ liệu của biến:
 - + Đặt con trỏ vào kiểu dữ liệu của biến (type) cần thay đổi
 - + Chọn kiểu dữ liệu phù hợp, nhấn OK.
- Thêm một biến:
 - + Chọn vị trí cần thêm biến.
 - + Chọn **Edit → Insert Variable**
- Xóa một biến:
 - + Chọn biến cần xóa
 - + Chọn **Edit → Clear** (nhấn phím **Delete**).

Bài 2. CÁC THAO TÁC VỚI BIẾN**Mục tiêu:**

- Trình bày được cách tạo mới, tính toán trên biến.
- Trình bày được các cách mã hóa biến.
- Trình bày được thao tác sắp xếp, lọc và tìm kiếm dữ liệu.
- Trình bày được thao tác phân tách file dữ liệu, nối thêm biến, nối thêm hàng.

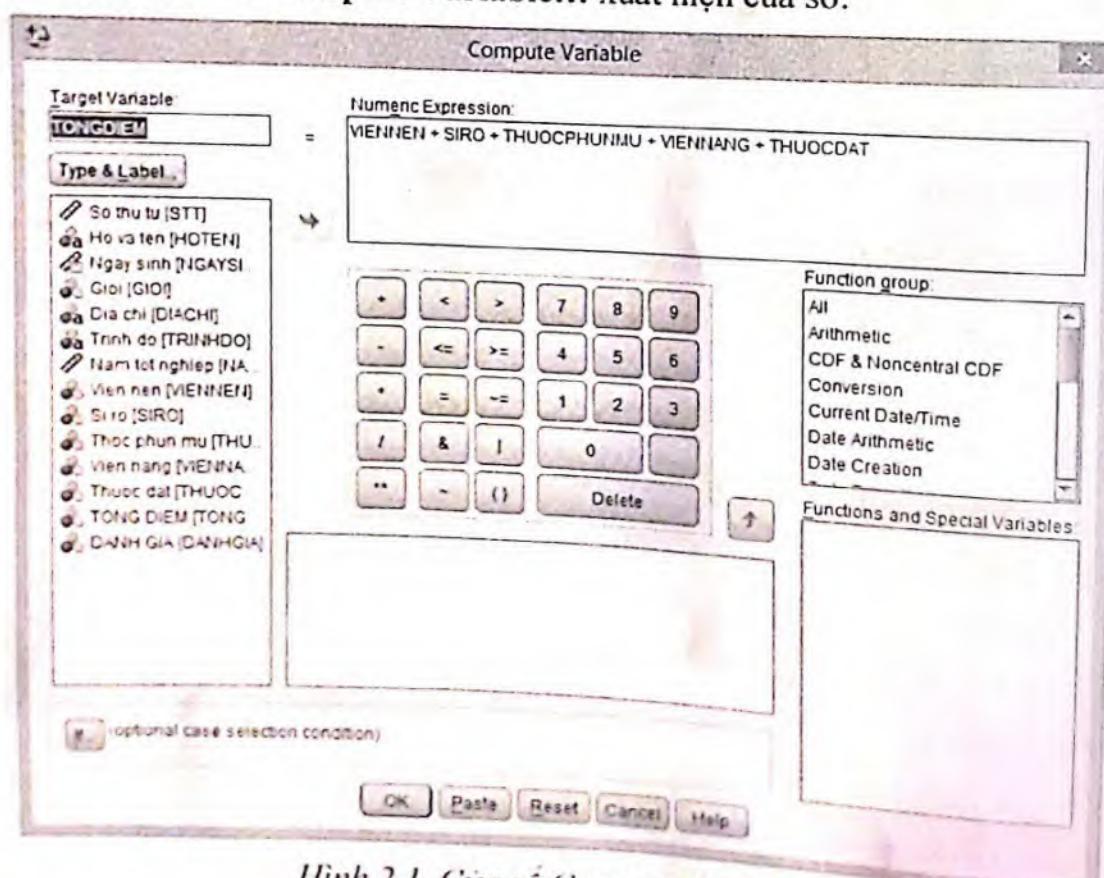
2.1. Tính giá trị cho biến

Sau khi đã thiết kế và nhập dữ liệu, trong quá trình xử lý số liệu, chúng ta cần có một số thao tác như: phát sinh thêm các biến mới, tính toán cho một biến đã được thiết kế, tính giá trị một cách chọn lọc với các bản ghi của dữ liệu trên các điều kiện logic...

Để thực hiện tính giá trị cho biến, ta thực hiện: **Transform → Compute Variable...**

Ví dụ: Tính giá trị cho cột **Tổng điểm** của cách hướng dẫn sử dụng các loại thuốc:

- Chọn **Transform → Compute Variable...** xuất hiện cửa sổ:



Hình 2.1. Cửa sổ Compute Variable

- Khai báo:

- + **Target Variable:** Tên của biến được tính giá trị - **TONGDIEM** (Nó có thể là một biến đã có sẵn hoặc biến mới sẽ được bổ sung vào file dữ liệu đang mở).
- + **Numeric Expression:** xây dựng một biểu thức gán giá trị cho biến, có thể sử dụng các hàm (nếu là ký tự phải để trong dấu ngoặc kép).

- Nhấn **OK**

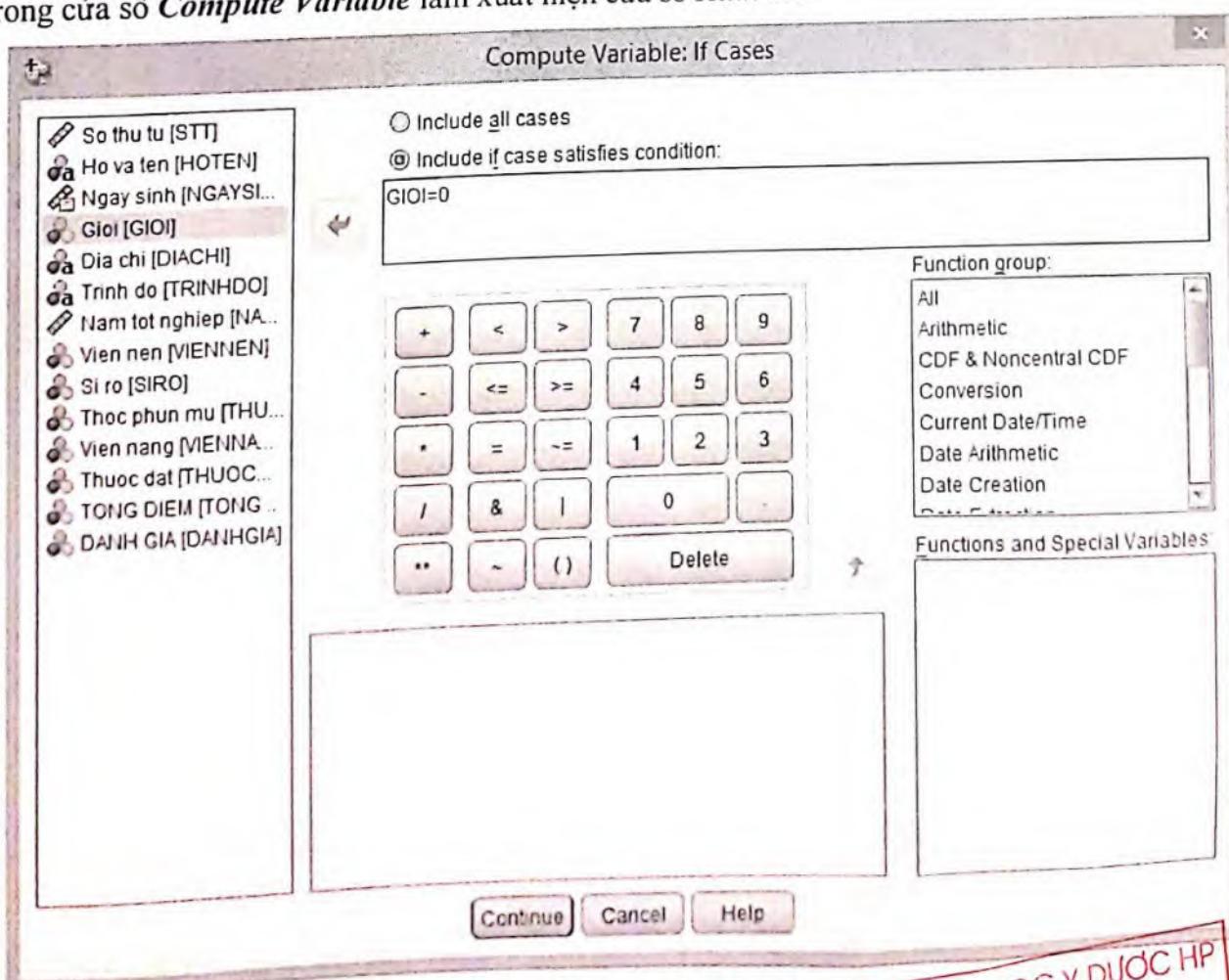
- Chọn màn hình **Data View** xem kết quả - **Hình 2.2**:

*BT1.sav [DataSet 1] - IBM SPSS Statistics Data Editor

STT	HOTTEN	NGAYSINH	GIOI	DIACH	TRINHD	NAMTN	VIENNEN	SIRO	THUOCPHNUMU	VIENNANG	THUOCDAT	TONGDIEM	DANHGIA
1	1 CAO DUC LUC	31-Dec-1974	0 AN DUONG	YSI	1969	0	2	2	2	2	1	7.00	
2	2 CHU THI LOAN	01-Aug-1974	1 AN DUONG	BACSI	2001	1	2	1	2	2	0	6.00	
3	3 DANG THI HONG	03-Feb-1971	1 AN DUONG	BACSI	1969	2	2	2	2	2	1	9.00	
4	4 DAO THI DAITHI	01-Feb-1958	1 THUY NGUYEN	BACSI	1981	0	1	0	2	2	2	5.00	
5	5 DAO THI THANH	15-May-1967	1 THUY NGUYEN	BACSI	1998	1	2	1	1	1	1	6.00	
6	6 DINH QUANG HUNG	24-Nov-1960	0 AN DUONG	BACSI	1998	2	2	2	2	2	1	9.00	
7	7 DINH THI DUNG	25-Jan-1960	1 THUY NGUYEN	YSI	1999	2	2	1	2	1	1	8.00	
8	8 DOAN THI TO	01-Feb-1957	1 THUY NGUYEN	YSI	2001	1	2	1	2	0	0	5.00	
9	9 HOANG THI SANG	13-May-1971	1 THUY NGUYEN	BACSI	1998	0	0	0	1	2	2	3.00	
10	10 LE CONG UNG	02-Sep-1968	0 AN DUONG	BACSI	2002	1	2	1	2	2	2	8.00	
11	11 LE THI HOAN	19-Aug-1966	1 THUY NGUYEN	BACSI	1969	2	0	1	2	0	0	5.00	
12	12 LUONG THI NHINH	03-Feb-1972	1 THUY NGUYEN	YSI	1984	2	0	1	0	2	2	5.00	
13	13 LUU THI QUYEN	19-May-1960	1 AN DUONG	BACSI	1969	0	0	2	1	1	1	4.00	
14	14 LUU VAN THUY	07-May-1968	0 AN DUONG	BACSI	1990	1	0	1	2	1	1	5.00	
15	15 NGO VAN PHANG	07-Nov-1959	0 AN DUONG	YSI	1969	0	2	2	0	0	0	4.00	
16	16 NGUYEN THI HANH	20-Nov-1952	1 THUY NGUYEN	YSI	2001	1	2	2	1	2	2	5.00	
17	17 NGUYEN THI LIEN	20-Oct-1967	1 THUY NGUYEN	YSI	2007	1	1	1	1	1	2	5.00	
18	18 NGUYEN THI QUE	22-Jan-1967	1 AN DUONG	YSI	2001	1	2	1	1	1	0	5.00	
19	19 NGUYEN THI TOI	17-Dec-1973	1 AN DUONG	BACSI	1997	1	1	2	1	2	2	7.00	
20	20 PHAM DUY KHANH	21-Jul-1983	0 THUY NGUYEN	YSI	1999	0	0	1	1	1	1	3.00	
21	21 PHAM THE BINH	28-May-1955	0 AN DUONG	YSI	1995	2	2	2	2	2	2	10.00	
22	22 PHAM THI HIEN	25-May-1960	1 THUY NGUYEN	YSI	1993	2	1	2	2	2	2	9.00	

Hình 2.2. Màn hình hiển thị kết quả sau khi tính giá trị cho biến TONGDIEM

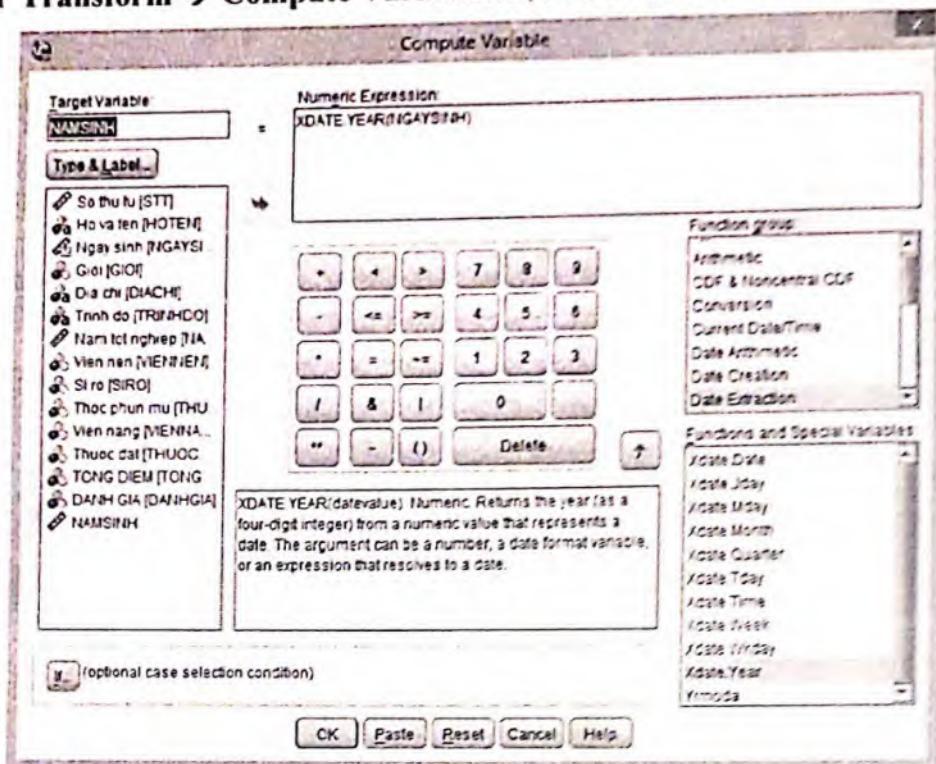
Lưu ý: Với các biểu thức gán có điều kiện, chọn nút trong cửa sổ Compute Variable làm xuất hiện cửa sổ Hình 2.3.



Hình 2.3. Cửa sổ Compute Variable: If Cases

Ví dụ: cần tạo và gán giá trị cho biến Năm sinh (NAMSINH), thực hiện như sau:

- Chọn Transform → Compute Variable... (xem Hình 2.4)



Hình 2.4. Cửa sổ minh họa tính giá trị cho biến NAMSINH

- Khai báo:

- + Target Variable ta nhập vào tên biến cần tạo là NAMSINH
- + Tại mục Numeric Expression sử dụng hàm XDATE.YEAR().

- Chọn OK.

- Chọn màn hình DataView xem kết quả - Hình 2.5:

	MOTON	NGAYSINH	SEX	DIENKH	TRIENGANG	NAMTRINH	VIENKHUAN	TRINHCHIEN	VIENKHUAT	THUONGXAT	TRANGTHIEU	DIENKHUA	TUOI
1	1 CAO DƯƠNG LỰC	31-Oct-1974	0	AN DƯƠNG	YES	1993	0	2	2	2	1	7.00	1994
2	1 NGUYỄN THỊ LINH	01-Apr-1974	1	AN DƯƠNG	NO	1993	1	3	3	3	0	6.00	1994
3	1 ĐÀO THỊ HỒNG	10-Feb-1971	1	AN DƯƠNG	NO	1993	2	2	2	2	1	3.00	1994
4	1 ĐÀO THỊ QUYỀN	07-Jun-1968	1	AN DƯƠNG	NO	1993	3	4	3	3	2	3.00	1994
5	1 ĐÀO THỊ THỊ NHẤT	16-May-1967	1	AN DƯƠNG	NO	1993	4	5	4	4	2	3.00	1994
6	1 ĐÀO THỊ QUANG HƯNG	24-Nov-1960	0	AN DƯƠNG	NO	1993	5	6	5	5	1	3.00	1994
7	1 ĐÀO THỊ THỊ NHẤT	26-Jan-1960	1	AN DƯƠNG	NO	1993	6	7	6	6	1	3.00	1994
8	1 ĐÀO THỊ THỊ TÙ	04-Feb-1967	1	AN DƯƠNG	NO	1993	7	8	7	7	0	3.00	1994
9	1 ĐÀO THỊ THỊ KHẢO	13-May-1971	1	AN DƯƠNG	NO	1993	8	9	8	8	0	3.00	1994
10	1 ĐÀO THỊ QUYỀN	02-Aug-1968	1	AN DƯƠNG	NO	1993	9	10	9	9	2	3.00	1994
11	1 ĐÀO THỊ HOA HỒ	15-Aug-1966	1	AN DƯƠNG	NO	1993	10	11	10	10	2	3.00	1994
12	1 ĐÀO THỊ THỊ NHẤT	03-Feb-1962	1	AN DƯƠNG	NO	1993	11	12	11	11	2	3.00	1994
13	1 ĐÀO THỊ QUYỀN	19-May-1964	1	AN DƯƠNG	NO	1993	12	13	12	12	0	3.00	1994
14	1 ĐÀO THỊ KHẨU	07-May-1968	1	AN DƯƠNG	NO	1993	13	14	13	13	0	3.00	1994
15	1 ĐÀO THỊ HỒNG HUÂN	07-Nov-1968	1	AN DƯƠNG	NO	1993	14	15	14	14	1	3.00	1994
16	1 ĐÀO THỊ HỒNG SƠ	25-Nov-1962	1	AN DƯƠNG	NO	1993	15	16	15	15	0	3.00	1994
17	1 ĐÀO THỊ HỒNG SƠ	26-Oct-1967	1	AN DƯƠNG	NO	1993	16	17	16	16	1	3.00	1994
18	1 ĐÀO THỊ HỒNG SƠ	25-Dec-1967	1	AN DƯƠNG	NO	1993	17	18	17	17	1	3.00	1994
19	1 ĐÀO THỊ HỒNG SƠ	17-Oct-1972	1	AN DƯƠNG	NO	1993	18	19	18	18	0	3.00	1994
20	1 ĐÀO THỊ HỒNG SƠ	21-Jun-1968	1	AN DƯƠNG	NO	1993	19	20	19	19	0	3.00	1994
21	1 ĐÀO THỊ HỒNG SƠ	06-May-1965	1	AN DƯƠNG	NO	1993	20	21	20	20	1	3.00	1994
22	1 ĐÀO THỊ HỒNG SƠ	28-May-1964	1	AN DƯƠNG	NO	1993	21	22	21	21	1	3.00	1994
23	1 ĐÀO THỊ HỒNG SƠ	28-Jun-1968	1	AN DƯƠNG	NO	1993	22	23	22	22	0	3.00	1994

Hình 2.5. Kết quả thực hiện tính giá trị cho biến MONTHBUY

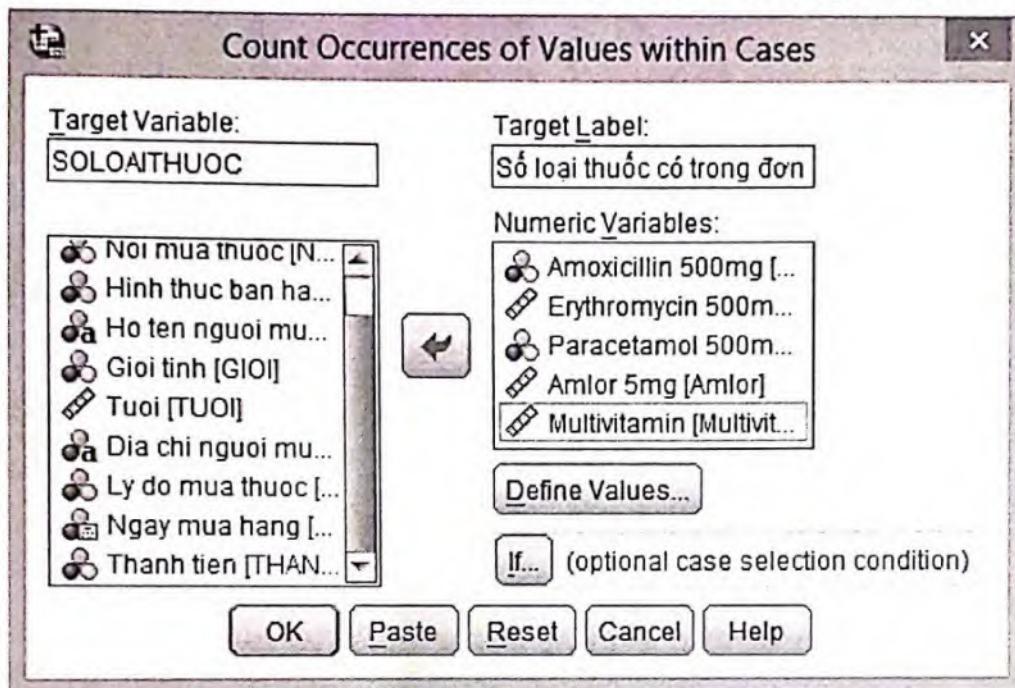
Lưu ý: có thể sử dụng dấu ngoặc, các hàm, các phép toán để tạo ra các biểu thức từ đơn giản đến phức tạp.

2.2. Đếm số lần xuất hiện của các giá trị trong từng trường hợp

Sử dụng Transform → Count Values within Cases

Ví dụ: Trong một cuộc khảo sát tại một số địa điểm bán thuốc, muốn đếm xem trong đơn thuốc, mỗi người mua mấy loại thuốc, thực hiện như sau:

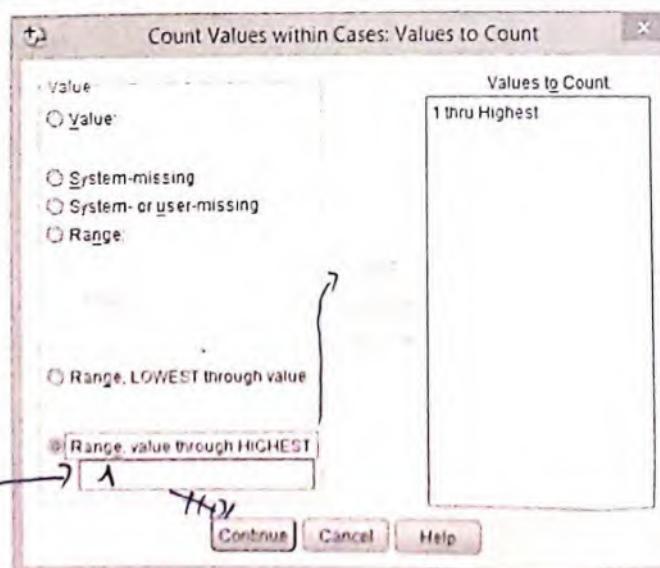
- Chọn Transform → Count Values within Cases, xuất hiện hộp thoại *Hình 2.6*:



Hình 2.6. Hộp thoại Count Occurrences of Values within Cases

- Khai báo:

- + **Target Variable:** Tên của biến được tính giá trị (*Ví dụ: SOLOAITHUOC*).
- + **Target Label:** Nhãn của biến (*Ví dụ: Số loại thuốc có trong đơn*).
- + Chọn các biến để đếm (*Ví dụ: ờ đây muốn đếm số lượng thuốc trong đơn thì chọn các biến về các loại thuốc gồm: Amoxicillin 500mg, Erythromycin 500mg...*).
- + Chọn **Define Values** để định nghĩa các giá trị cần đếm như *Hình 2.7* (chọn đếm các biến có giá trị từ 1 trở lên) – Xem kết quả *Hình 2.7*.



Hình 2.7. Hộp thoại Count Values within Cases: Values to Count

- + Chọn nút **Continue** để quay trở lại hộp thoại *Hình 2.6*.

- Chọn OK và quan sát kết quả - Hình 2.8.

STT	NAMJU	HINH THUC	HOTEN	GIOI	TUOI	DIACHI	BENH	NGAYMUA	Amoxic	Erythromycin	Paracetamol	Amlo	Multivitamin	THAIPITEN	SOLOATHILOC	
1	1	1	0 CAO DUC LUC	0	35	Hàng Bàng	1.00	10/20/12	10	10			30			3
2	2	1	0 CHU TH LOAN	1	41	Ngo Cuyen	1.00	10/20/12		20			30		2	
3	3	2	1 DANG THI HONG	1	45	Le Chan, HP	1.00	12/25/12			30	30	30			2
4	4	1	1 DAO THI CAIHN	1	37	Kien An, HP	2.00	12/25/12			10	30	30			3
5	5	2	1 DAO THI THANH	1	35	An Lao, HP	3.00	11/22/12		20	10					2
6	6	1	1 DINH QUANG HUNG	0	41	Vinh Dao	3.00	11/22/12	20		10					2
7	7	1	1 DINH THI DUONG	1	37	Hàng Bàng	2.00	09/10/12			30	30				2
8	8	2	1 DOAN THI TO	1	34	Ngo Cuyen	4.00	12/06/12		20	10		20			3
9	9	1	1 HOANG THI SAMG	1	20	Le Chan, HP	1.00	11/15/12					30			1
10	10	1	1 LE CONG URG	0	43	Kien An, HP	3.00	11/17/12		20	10					2
11	11	1	1 LE THI HOAN	1	23	An Lao, HP	3.00	11/19/12		20	10					2
12	12	2	0 LUONG THI NHANH	1	48	Vinh Dao	3.00	11/19/12		30	5					2
13	13	1	1 LUU THI QUYEN	1	25	Hàng Bàng	4.00	11/20/12		30	10		60			3
14	14	2	0 LUU VAN THUY	0	43	Ngo Cuyen	4.00	10/16/12	5	30	20	50	60			5
15	15	2	1 NGO VAN PHANG	0	47	Le Chan, HP	2.00	10/20/12		30	10	30	30			4
16	16	2	0 NGUYEN THI HUONG	1	43	Kien An, HP	1.00	10/29/12	10		20		30			3
17	17	2	0 NGUYEN THI LEN	1	45	An Lao, HP	3.00	12/25/12	20	30	10					3
18	18	2	0 NGUYEN THI QUE	1	42	Vinh Dao	3.00	12/25/12	30	30	20	50	10			5
19	19	1	0 NGUYEN THI TOI	1	48	Hàng Bàng	3.00	11/22/12		10	5	10				3
20	20	2	1 PHAM DUY KHOANH	0	41	Ngo Cuyen	4.00	11/22/12		10	10		30			2
21	21	2	1 PHAM THE BINTH	0	35	Le Chan, HP	4.00	10/29/12	10				30			2
22	22	1	1 PHAM THI HIEN	1	43	Kien An, HP	2.00	10/29/12				30				1
23	23	1	0 PHAM THI HOA	1	55	An Lao, HP	4.00	12/25/12	20		10	10	50			4

Hình 2.8. Kết quả thực hiện lệnh

2.3. Mã hóa dữ liệu

Trong quá trình phân tích, nhiều trường hợp ta phải mã hóa lại các giá trị của biến vì một mục đích nào đó. Ta có thể mã hóa lại các giá trị của một biến có sẵn hoặc lập một biến mới để chứa các giá trị được mã hóa lại. Thường áp dụng trong phân nhóm dữ liệu hoặc chuyển một biến định lượng thành một biến định tính.

2.3.1. Mã hóa lại dữ liệu của biến

Là cách mã hóa lại dữ liệu của một biến và dữ liệu sau khi mã hóa được lưu vào chính biến đó, sử dụng chức năng Transform → Recode into Same Variables...

Ví dụ: Khi nhập dữ liệu biến DIACHI có 2 giá trị AN DUONG, THUY NGUYEN – như Hình 2.9. Ta muốn mã hóa lại dữ liệu cho biến DIACHI như sau: AN DUONG =1 và THUY NGUYEN =2.

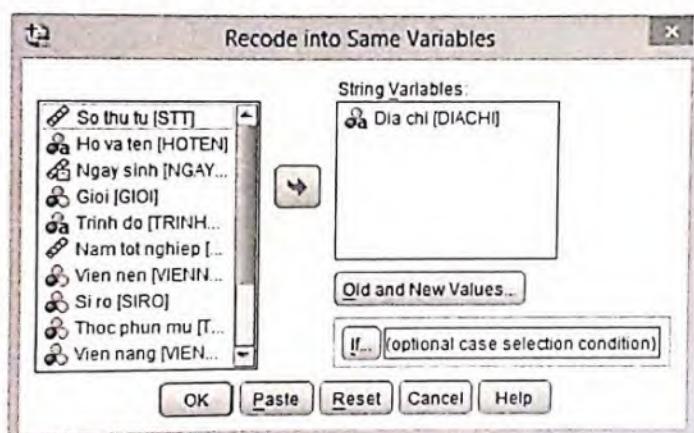
HINH THUC	NGAYMUA	GIO	DIACHI	THUONGDAT	NAMHUU	VIETNAME	SIRD	THUONGDAT	VIETNAME	THUONGDAT	TONGDOU	DAMYRA	THAMKHIEU
1	1/1/2012	11:45:00	0 AN DUONG	1	1969	6	2	2	2	1	1	140	1574
2	1/1/2012	11:45:00	1 AN DUONG	1	2021	1	2	1	2	0	1	125	1575
3	1/1/2012	11:45:00	1 AN DUONG	1	1968	2	2	2	2	1	1	150	1576
4	1/1/2012	11:45:00	1 THUY NGUYEN	1	1961	1	1	5	2	2	1	100	1577
5	1/1/2012	11:45:00	1 THUY NGUYEN	1	1968	1	2	1	1	1	1	120	1578
6	1/1/2012	11:45:00	1 THUY NGUYEN	1	1966	2	2	2	2	1	1	140	1579
7	1/1/2012	11:45:00	1 THUY NGUYEN	1	1969	2	2	2	2	1	1	150	1580
8	1/1/2012	11:45:00	1 THUY NGUYEN	1	1979	2	2	1	2	1	1	100	1581
9	1/1/2012	11:45:00	1 THUY NGUYEN	1	2001	1	2	1	2	1	1	150	1582
10	1/1/2012	11:45:00	1 THUY NGUYEN	1	1994	8	0	0	1	2	1	100	1583
11	1/1/2012	11:45:00	1 AN DUONG	1	2002	1	2	1	2	1	1	100	1584
12	1/1/2012	11:45:00	1 THUY NGUYEN	1	1969	2	0	5	2	1	1	100	1585
13	1/1/2012	11:45:00	1 THUY NGUYEN	1	1984	2	3	1	5	2	1	120	1586
14	1/1/2012	11:45:00	1 AN DUONG	1	1993	0	0	2	1	1	1	100	1587
15	1/1/2012	11:45:00	1 AN DUONG	1	1999	1	0	2	1	1	1	100	1588
16	1/1/2012	11:45:00	1 AN DUONG	1	1999	1	0	2	1	1	1	100	1589
17	1/1/2012	11:45:00	1 THUY NGUYEN	1	2003	1	2	2	1	1	1	100	1590
18	1/1/2012	11:45:00	1 THUY NGUYEN	1	1999	1	1	1	1	1	1	100	1591
19	1/1/2012	11:45:00	1 AN DUONG	1	2004	1	2	1	1	1	1	100	1592
20	1/1/2012	11:45:00	1 AN DUONG	1	1999	1	1	2	1	1	1	100	1593
21	1/1/2012	11:45:00	1 AN DUONG	1	1999	1	0	1	1	1	1	100	1594
22	1/1/2012	11:45:00	1 AN DUONG	1	1999	2	1	2	2	2	2	100	1595
23	1/1/2012	11:45:00	1 AN DUONG	1	2004	1	2	1	1	1	1	100	1596

Hình 2.9. Cửa sổ trước khi mã hóa giá trị biến

Hình 2.10. Kết quả sau khi thực hiện mã hóa vào chính biến đó

Các bước thực hiện:

- Chọn Transform → Recode into Same Variables... và các khai báo như *Hình 2.11*.



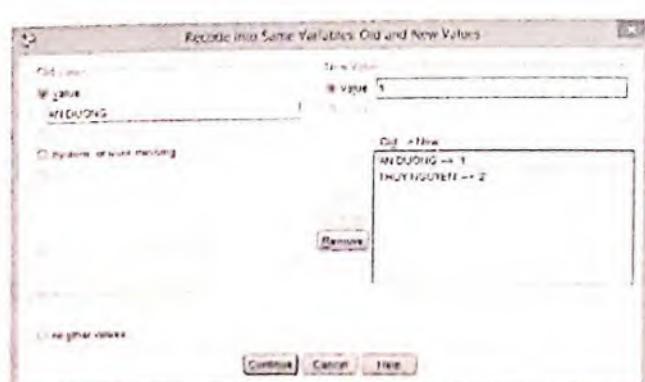
Hình 2.11.

- Nhấn nút Old and New Values để mã hóa các giá trị.

+ Bên phần Old Value chọn những giá trị sẽ được mã hóa vào một nhóm, sau đó nhập giá trị mới đại diện cho cả nhóm đó vào ô Value ở phần New Value.

+ Nhập Add và tiếp tục mã hóa với các giá trị hoặc nhóm giá trị tiếp theo

(*Hình 2.12*).



Hình 2.12.

+ Chọn nút Continue để quay lại hộp thoại *Hình 2.11*.

- Chọn OK và quan sát kết quả - *Hình 2.10*.

2.3.2. Mã hóa dữ liệu của một biến vào biến mới

Là cách mã hóa dữ liệu của một biến đã có sẵn, giá trị sau khi mã hóa được lưu vào một biến mới. Biến chứa giá trị sau khi mã hóa có thể được khai báo trước hoặc trong quá trình mã hóa.

Cách mã hóa này không làm thay đổi dữ liệu của biến ban đầu. Ta có thể mã hóa biến kiểu số hoặc biến kiểu kí tự,... có thể chuyển một biến kiểu số thành biến kiểu kí tự hoặc ngược lại.

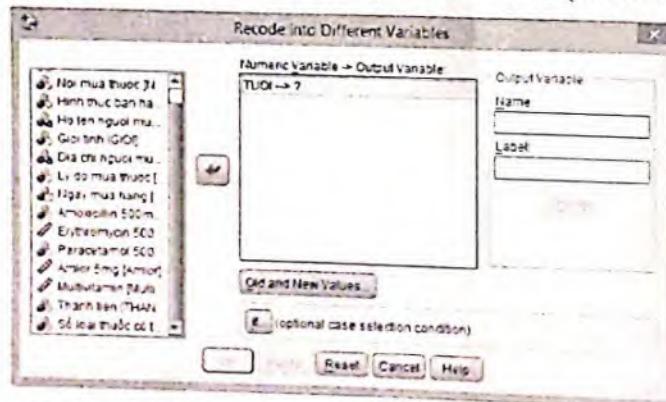
Thường áp dụng cách mã hóa này trong *phân nhóm dữ liệu* hoặc *chuyển đổi một biến định lượng thành một biến định tính*.

Sử dụng **Transform → Recode into Different Variables....**

Ví dụ: Mã hóa dữ liệu của biến **TUOI** (*Tuổi*) thành biến **NHOMTUOI** (*Nhóm tuổi*) theo quy định: Nếu $18 \leq \text{Tuổi} \leq 30$ thì $\text{NHOMTUOI} = 1$
 $31 \leq \text{Tuổi} \leq 40$: $\text{NHOMTUOI} = 2$
 $41 \leq \text{Tuổi} \leq 50$: $\text{NHOMTUOI} = 3$
 $\text{Tuổi} \geq 51$: $\text{NHOMTUOI} = 4$

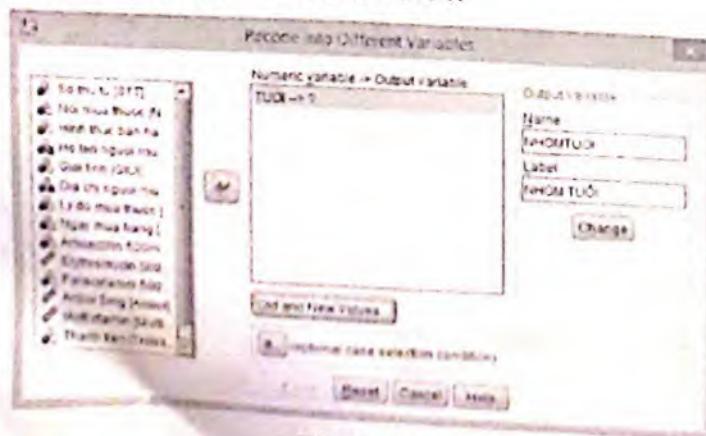
Thực hiện các bước sau đây:

- Chọn **Transform → Recode into Different Variables....**,
- Chọn biến cần mã hóa dữ liệu trong khung bên trái của hộp thoại **Recode into Different Variables** đưa sang khung **Input Variable -> Output Variable** (*Xem Hình 2.13*):



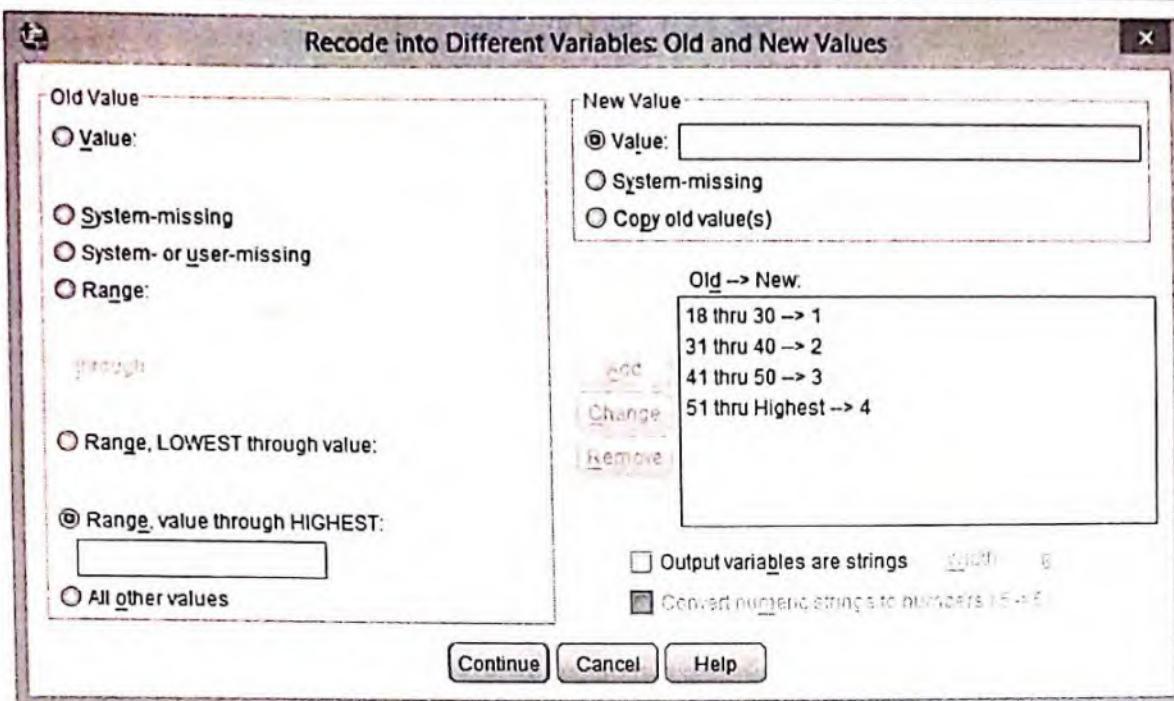
Hình 2.13. Cửa sổ mã hóa dữ liệu sang biến mới

- Khai báo **Tên (Name)**, **Nhãn (Label)** của biến chứa giá trị sau khi mã hóa trong khung **Output Variable** → Nhấp nút **Change** – xem *Hình 2.14*.



Hình 2.14.

- Nhấp nút **Old and New Values** để khai báo quy luật mã hóa – xem *Hình 2.15*.



Hình 2.15.

- Chọn nút **Continue** để quay lại hộp thoại *Hình 2.14*.
- Chọn **OK** và quan sát kết quả *Hình 2.16*.

BT2.sav [DataSet3] - IBM SPSS Statistics Data Editor																	
	STT	NOMUA	HINHTHUC	HOTEN	gioi	TUOI	NHOMTUOI	DIACHI	BENH	NGAYMUA	Amoxicilin	Erythromycin	Paracetamol	Amiloride	Multivitamin	THANHTIEN	SOLDATE
1	1	1	0 CAO DUC LUC	0	35	35	2 Hong Bang	1.00	10/20/12	10	10				30		
2	2	1	0 CHU THI LOAN	1	41	31	3 Ngo Quyen	1.00	10/20/12		20				30		
3	3	2	1 DANG THI HONG	1	45	31	3 Le Chan HP	1.00	12/25/12					30	30		
4	4	1	1 DAO THI DANH	1	37	31	2 Kien An HP	2.00	12/25/12				10	30	30		
5	5	2	1 DAO THI THANH	1	35	31	2 An Lac, HP	3.00	11/22/12		20	10					
6	6	1	1 DINH QUANG HUNG	0	41	31	3 Vinh Bao	3.00	11/22/12	20		10					
7	7	1	1 DINH THI DUNG	1	37	31	2 Hong Bang	2.00	09/10/12					30	30		
8	8	2	1 DOAN THI TO	1	34	31	2 Ngo Quyen	4.00	12/06/12		20	10			20		
9	9	1	1 HOANG THI SANG	1	20	31	1 Le Chan HP	1.00	11/15/12						30		
10	10	1	1 LE CONG UNG	0	40	31	2 Kien An, HP	3.00	11/17/12		20	10					
11	11	1	1 LE THI HOAI	1	20	31	1 An Lac, HP	3.00	11/19/12		20	10					
12	12	2	0 LUONG THI NHINH	1	48	31	3 Vinh Bao	3.00	11/19/12		30	10			60		
13	13	1	1 LUU THI CUYEN	1	25	31	1 Hong Bang	4.00	11/20/12		30	10					
14	14	2	0 LUU VAN THUY	0	40	31	2 Ngo Quyen	4.00	10/16/12	5	30	20	50	60			
15	15	2	1 NGO VAN PHONG	0	47	31	3 Le Chan HP	3.00	10/20/12		30	10	30		30		
16	16	2	0 NGUYEN THI HANH	1	40	31	2 Kien An, HP	1.00	10/20/12	10		20			30		
17	17	2	0 NGUYEN THI LIEN	1	45	31	3 An Lac, HP	3.00	12/25/12	20	30	10					
18	18	2	0 NGUYEN THI QUE	1	42	31	3 Vinh Bao	3.00	12/25/12	30	30	20	50	10			
19	19	1	0 NGUYEN THI TOI	1	48	31	2 Hung Sung	3.00	11/22/12		10	5	10		30		
20	20	2	1 PHAM DUY KHANH	0	41	31	3 Ngo Quyen	4.00	11/22/12		10	10					
21	21	2	1 PHAM THE BINH	0	35	31	2 Le Chan HP	4.00	10/20/12	10					30		
22	22	1	1 PHAM THI HEN	1	40	31	2 Kien An, HP	2.00	10/20/12				30				
23	23	1	0 PHAM THI HOA	1	56	31	4 An Lac, HP	4.00	12/25/12	20		10	10	60			

Hình 2.16. Kết quả mã hóa biến tuổi sang biến nhóm tuổi

Kết quả sẽ tạo thêm biến **NHOMTUOI** (*Nhóm tuổi*) chứa kết quả mã hóa từ giá trị của biến **TUOI**.

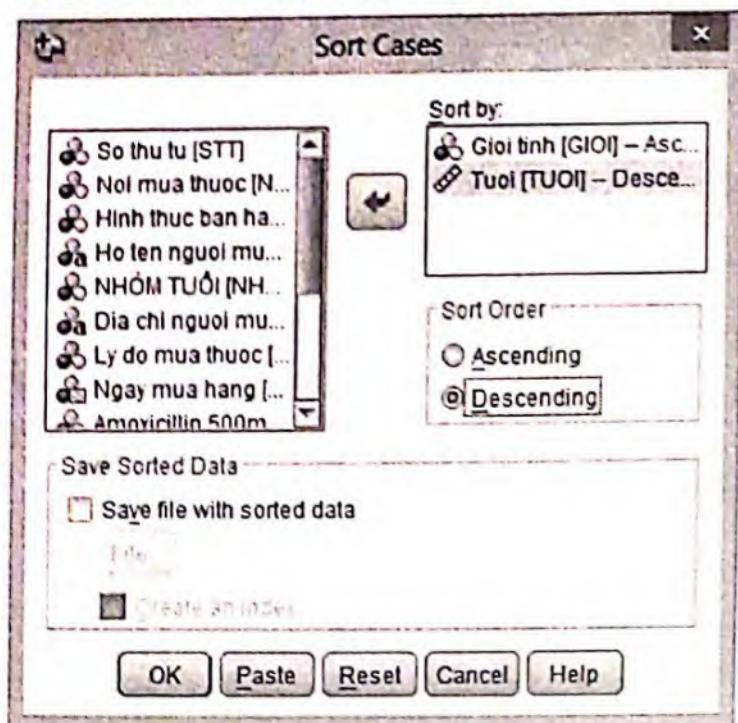
2.4. Sắp xếp, lọc và tìm kiếm thông tin

2.4.1. Sắp xếp dữ liệu

Sắp xếp và tìm kiếm là công việc thường xuyên phải làm để xem xét dữ liệu, tìm các giá trị lỗi, để nhận định sơ bộ về dữ liệu, sửa chữa những sai sót,...

Các bước thực hiện:

- Chọn menu Data → Sort Case, xuất hiện hộp thoại *Hình 2.17*

*Hình 2.17.*

- Khai báo:

+ Sort by: Danh sách các biến cần sắp xếp (Ví dụ: GIOI, TUOI) được chọn trong khung bên trái của hộp thoại *Hình 2.17*.

+ Sort order: Thứ tự sắp xếp Ascending (A-Z; 0-9) hoặc Descending (Z-A; 9-0). Có thể sắp xếp trên nhiều biến.

- Chọn OK để kết thúc lệnh và xem kết quả (*Hình 2.18*).

	STT	HOTEN	HOITU	GIOI	TUOI	NHOMTUOI	DAODA	BENH	NGAYMUA	Amoxicilin	Erythromycin	Paracetamol	Antif.	Multivitamin	THANHTIEN	SOLoan
1	40	1 NGUYEN VAN TAN	0	0	62	4 Kien An HP	2 00	11/22/12					20	20		
2	36	0 TRAN VAN NAM	0	0	60	4 Vinh Bac	3 00	10/20/12	20			10	10			
3	15	2 NGO VAN PHONG	0	0	47	3 Le Chan HP	3 00	10/20/12		30	10	10	30			
4	6	1 DINH QUANG HUNG	0	0	41	3 Vinh Bac	3 00	11/22/12	20			10				
5	20	2 PHAM DUY KHANH	0	0	41	3 Ngo Quyen	4 00	11/22/12		10	10					
6	10	1 LE CONG UNG	0	0	40	2 Kien An HP	3 00	11/17/12		20	10					
7	14	2 LUU VAN THUY	0	0	40	2 Ngo Quyen	4 00	10/16/12	5	30	20	50	50			
8	1	2 CAO DUC LUC	0	0	35	2 Hung Bang	1 00	10/20/12	10	10					30	
9	22	2 PHAM THE BINH	0	0	36	2 Le Chan HP	4 00	10/20/12	10						30	
10	26	2 PHAM VAN HA	0	0	35	2 Ngo Quyen	3 00	11/22/12		20	10					
11	38	1 DANG MAI AN	0	0	32	2 Ngo Quyen	2 00	12/25/12				30	30			
12	20	1 NGUYEN TH LOAN	1	1	64	4 Le Chan HP	2 00	11/22/12				30	30			
13	23	1 PHAM TH HOA	1	1	56	4 An Lac HP	4 00	12/25/12	20		10	10	60			
14	36	1 VU THI TUYET	1	1	53	4 An Lac HP	4 00	10/20/12	20		10	10	60			
15	29	2 TRAN THI HAI	1	1	51	4 An Lac HP	4 00	11/15/12		20	10	20	30			
16	12	2 LUONG THI NHANH	1	1	48	2 Vinh Bac	3 00	11/15/12	20		5		30			
17	19	1 NGUYEN TH TUY	1	1	48	3 Hung Bang	2 00	11/22/12	10	5	10					
18	26	1 PHAM TH MEN	1	1	48	3 Hung Bang	4 00	11/22/12	10		10	30	30			
19	30	1 TU THI TOI	1	1	48	3 Vinh Bac	4 00	11/17/12	20		10	30	30			
20	17	2 NGUYEN TH LIEN	1	1	46	3 An Lac HP	3 00	12/25/12	20	20	10	30	30			
21	3	2 DANG TH HONG	1	1	45	3 Le Chan HP	1 00	12/25/12		20	10					
22	10	2 NGUYEN TH QUYEN	1	1	42	3 Vinh Bac	3 00	12/25/12	20	30	20	50	50	10		
23	27	1 TA THI TAM	1	1	42	3 Le Chan HP	4 00	09/10/12		10	10	30	60			

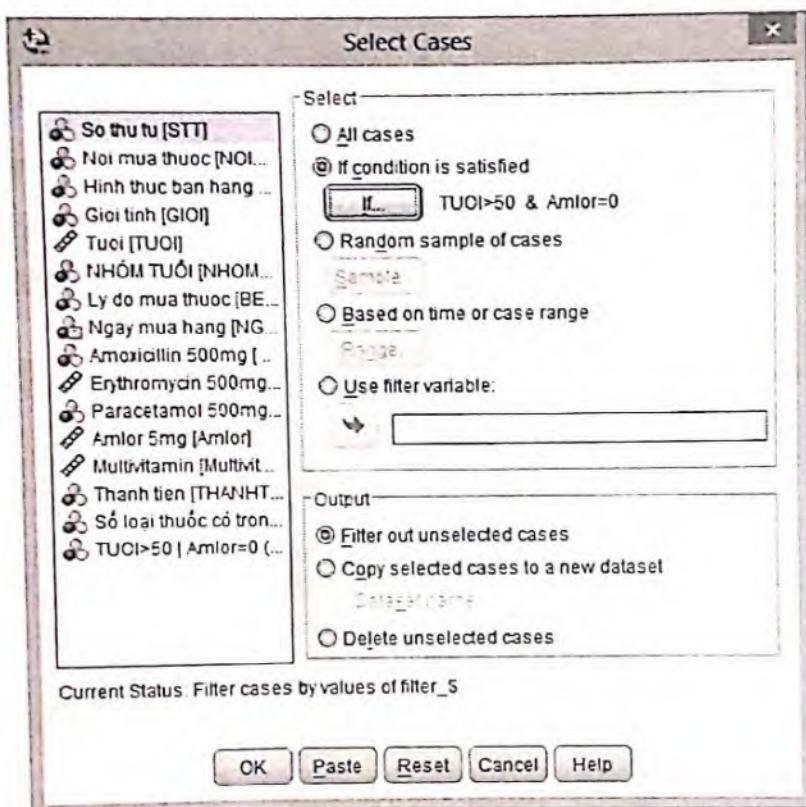
Hình 2.18. Kết quả sắp xếp

Lưu ý: Nên lập một cột STT (*thứ tự*) trước khi sắp xếp để có thể phục hồi bảng số liệu theo thứ tự ban đầu.

2.4.2. Lọc dữ liệu

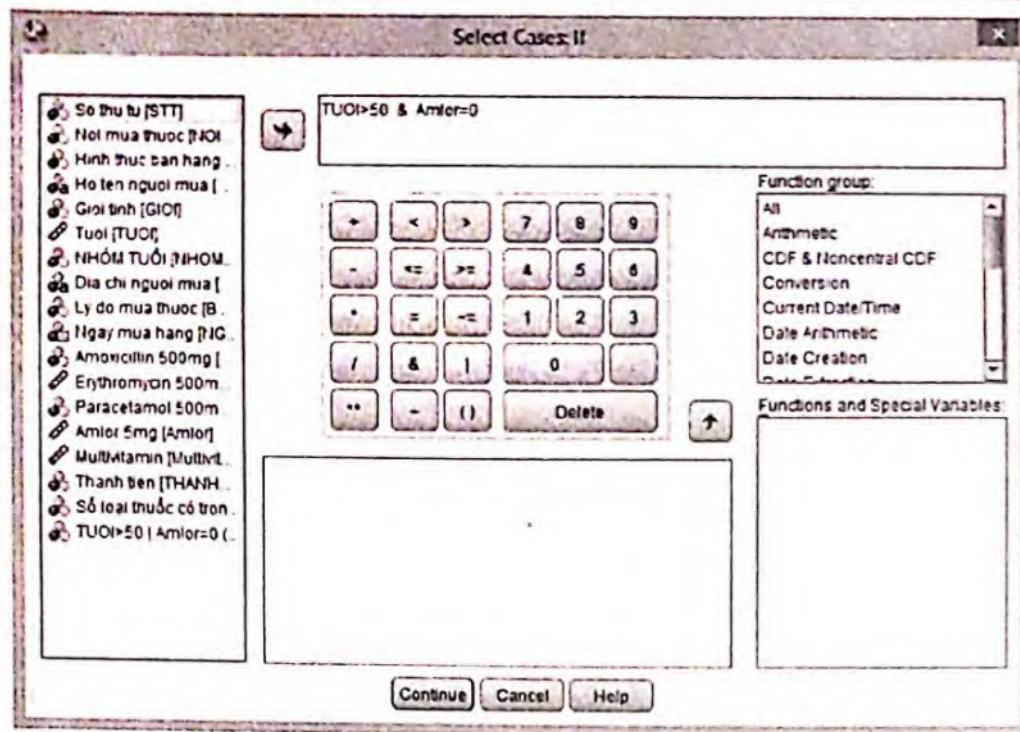
Lọc dữ liệu cho phép chọn ra các trường hợp (*case*) thỏa mãn một hoặc một số điều kiện nào đó. Các trường hợp không thỏa mãn điều kiện lọc sẽ bị bỏ qua (*không tham gia xử lý trong các lệnh thống kê tiếp theo của SPSS*), vì vậy đây cũng là phương án phân nhóm trong xử lý số liệu. Các bước thực hiện:

- Chọn Data → Select Cases... xuất hiện cửa sổ *Hình 2.19*.



Hình 2.19.

- Có nhiều dạng thức lọc nhưng thông dụng là lọc theo điều kiện: Đánh dấu mục **If condition is satisfied** rồi nhấp nút **!** xuất hiện hộp thoại *Hình 2.20* để nhập biểu thức thể hiện điều kiện lọc:



Hình 2.20.

- Nhấp nút **Continue** để quay lại hộp thoại *Hình 2.19* → Nhấp **OK** để quan sát kết quả (*Hình 2.21*).

Ví dụ: Chọn các trường hợp/bản ghi có **TUOI > 50** và không mua thuốc **Amlor 5mg**, thao tác như các bước trên ta được kết quả:

*ST2say [DataSet3] - IBM SPSS Statistics Data Editor														
File	Edit	View	Data	Transform	Analyze	Direct Marketing	Graphs	Utilities	Add-ons	Window	Help			
15	STT	29												
7	14	2	0 LUU VAN THUY	0	40	2 Ngo Quyen	4.00	10/15/12	5	30	20	50	60	
8	1	1	0 CAO DUC LUC	0	35	2 Hong Bang	1.00	10/20/12	10	10	0	30		
9	22	2	1 PHAM THE BINH	0	35	2 Le Chan HP	4.00	10/20/12	10		0	30		
10	26	2	1 PHAM VAN HA	0	35	2 Ngo Quyen	3.00	11/22/12		20	10	0		
11	38	1	1 DANG MAI ANH	0	32	2 Ngo Quyen	2.00	12/25/12			30	30		
12	39	1	1 NGUYEN THI LOAN	1	64	4 Le Chan HP	2.00	11/22/12			30	30		
13	23	1	0 PHAM THI HOA	1	56	4 An Lao HP	4.00	12/25/12	20		10	30		
14	35	1	0 VU TH TUYET	1	53	4 An Lao. HP	4.00	10/20/12	20		10	20	60	
15	23	2	1 TRINH THI HAI	1	51	4 An Lao. HP	4.00	11/15/12		20	10	0	30	
16	12	2	0 LUONG TH NHANH	1	48	3 Vinh Bao	3.00	11/19/12		30	5	0		
17	19	1	0 NGUYEN TH TOI	1	46	3 Hong Bang	3.00	11/22/12		10	5	10		
18	25	1	1 PHAM THI MIEN	1	46	3 Hong Bang	4.00	11/22/12	20		10	30		
19	30	1	1 TU THI TO	1	48	3 Vinh Bao	4.00	11/17/12	20		10	30	20	
20	17	2	0 NGUYEN TH LIEN	1	46	3 An Lao. HP	3.00	12/25/12	20	30	10	0		
21	3	2	1 DANG TH HONG	1	45	3 Le Chan. HP	1.00	12/25/12			30	30		
22	18	2	0 NGUYEN TH QUE	1	42	3 Vinh Bao	3.00	12/25/12	20	30	20	50	10	
23	27	1	0 TA THI TAM	1	42	3 Le Chan. HP	4.00	09/10/12		30	10	30	50	
24	2	1	0 CHU THI LCAN	1	41	3 Ngo Quyen	1.00	10/20/12		20	0	30		
25	24	2	1 PHAM TH HUONG	1	41	3 Vinh Bao	2.00	12/25/12			30			
26	16	2	0 NGUYEN TH HANH	1	46	2 Kien An. HP	1.00	10/20/12	10		20	0	30	
27	22	1	1 PHAM TH HIEN	1	40	2 Kien An. HP	2.00	10/20/12			30			
28	21	1	0 VU TH TANG	1	35	2 Le Chan. HP	1.00	11/20/12			10	0	20	
29	4	1	1 DAO TH DANH	1	37	2 Kien An. HP	2.00	12/25/12			10	30	30	

Hình 2.21

Lưu ý: Sau khi thực hiện lọc dữ liệu xong, cần phải bỏ lọc để dữ liệu trở về trạng thái ban đầu trước khi thực hiện các thao tác tính toán, xử lý... tiếp theo bằng cách chọn **Data → Select Cases**, nhấp nút **Reset, OK**.

2.4.3. Tìm kiếm và thay thế dữ liệu

Các bước thực hiện:

- Đặt con trỏ ở cột cần tìm (*Con trỏ đặt ở cột nào sẽ tìm trong phạm vi cột đó*).

- Chọn **Edit → Find**, xuất hiện hộp thoại **Find and Replace**.

- Để tìm kiếm, chọn thẻ **Find**, nhập nội dung cần tìm vào khung **Find**.

- Trong trường hợp muốn thay thế nội dung cần tìm bằng một nội dung khác. Chọn thẻ **Replace** và nhập nội dung cần tìm vào khung **Find**, nhập nội dung thay thế vào khung **Replace with**.

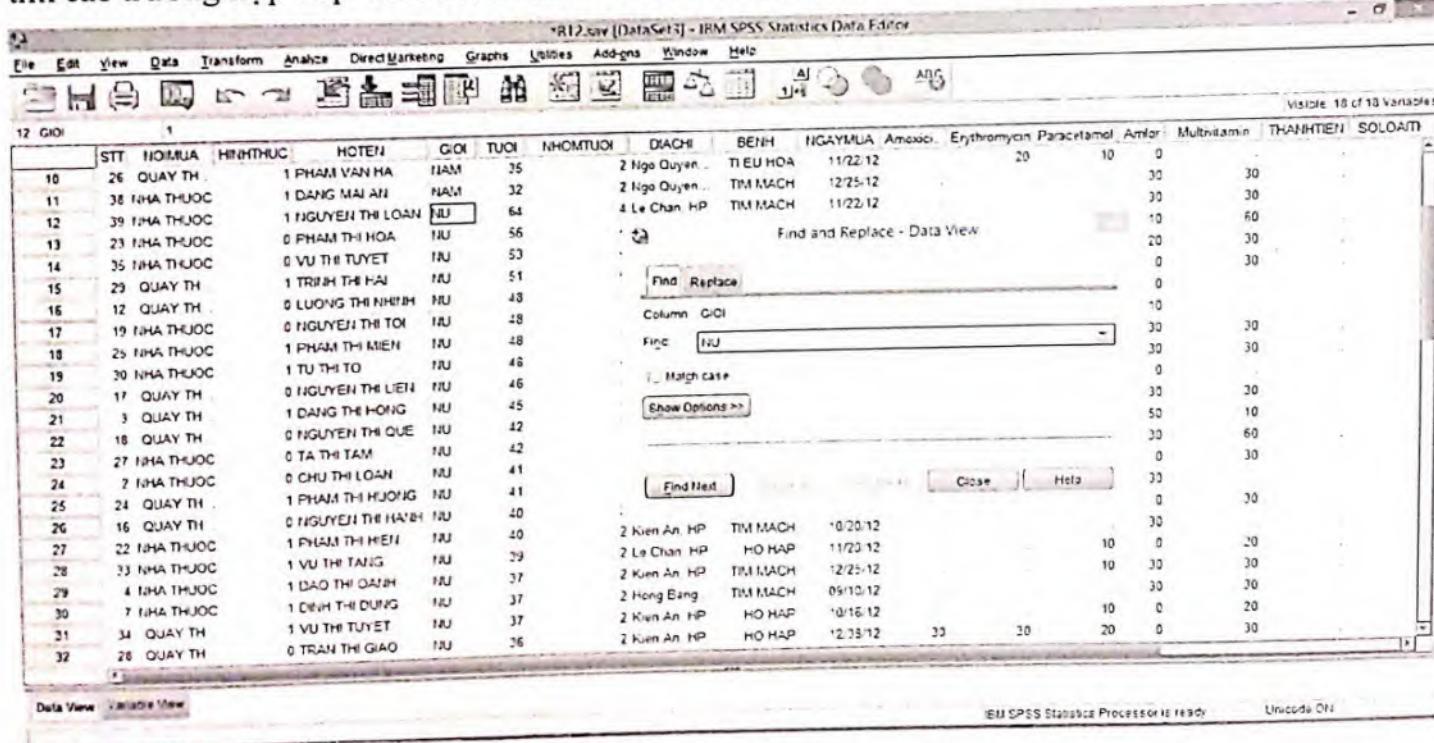
- Nhập **Find Next** để thực hiện việc tìm (hoặc nhấn **Replace** hoặc **ReplaceAll** để thực hiện thay thế).

- Nhấp nút **Close** để kết thúc tìm kiếm.

Ví dụ: Tìm những trường hợp có Giới là Nữ, ta thực hiện các bước sau đây:

- Đặt con trỏ ở cột cần tìm (*cột GIOI*).

- Chọn lệnh và các khai báo *này* Hình 2.22. SPSS sẽ tìm đến trường hợp đầu tiên. Muốn tìm các trường hợp tiếp theo thỏa mãn điều kiện tìm kiếm cần phải sử dụng nút **Find Next**.



Hình 2.22.

2.5. Chia dữ liệu trong tệp thành các nhóm

Chia dữ liệu trong tệp thành các nhóm để phân tích dựa trên các giá trị của một hay nhiều biến nhóm. Nếu muốn nhóm theo nhiều biến, dữ liệu sẽ được nhóm theo thứ tự các biến đưa vào chia nhóm. Chia nhóm dữ liệu trong tệp thường được áp dụng cho các lệnh thống kê mô tả.

Các bước thực hiện:

- Chọn **Data → Split File**, xuất hiện hộp thoại **Split File**.

- Chọn các biến cần phân nhóm trong khung bên trái của hộp thoại **Split File** dưa sang khung **Group Based on**.

- Chọn OK để kết thúc lệnh.

Ví dụ: Chọn giới tính (GIOI) là biến nhóm đầu tiên và lý do mua thuốc (BENH) là biến nhóm thứ hai, các trường hợp sẽ được nhóm lại theo phân loại lý do mua thuốc trong mỗi loại giới tính.

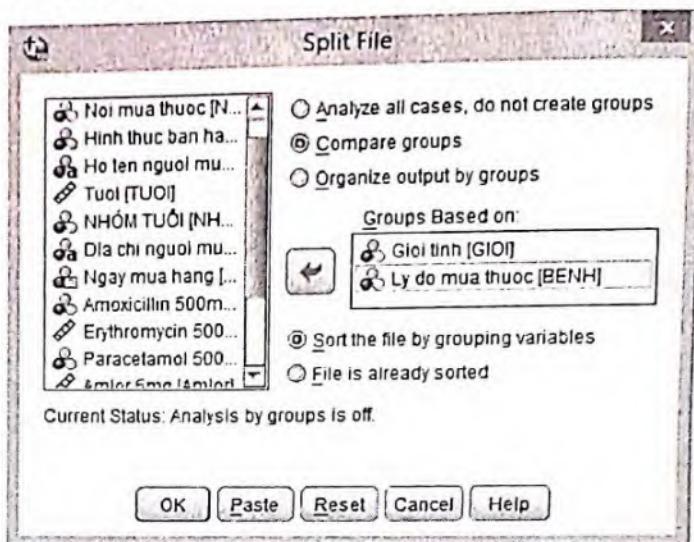
- Chọn lệnh và thực hiện các khai báo như *Hình 2.23*:

+ Biến **GIOI** có thứ tự ưu tiên thứ nhất, nghĩa là dữ liệu sẽ chia thành 2 nhóm Nam và Nữ.

+ **BENH** là biến có thứ tự ưu tiên thứ 2.

Lưu ý: Có thể chọn đến tám biến để phân nhóm.

- Chọn OK để quan sát kết quả - *Hình 2.24*.



Hình 2.23.

*BT2.sav [DataSets] IBM SPSS Statistics Data Editor																
Visible: 18 of 12 Variables																
STT	NOMAL	HINH THUC	HOTEN	GIOI	TUOI	THI HOM TUOI	DIACHI	BENH	NGAYMUA	Amoxicilin	Erythromycin	Paracetamol	Anor	Multivitamin	THANHTIEN	SOLADITH
1	1	1	0 CAO DUC LUC	0	35	2 Hang Bang		1	10/20/12	10	10	0	30			
2	40	1	1 NGUYEN VAN TAN	0	62	4 Kien An, HP		2	11/22/12			30	30			
3	38	1	1 DANG MAU AN	0	32	2 Ngo Cuyen		2	12/25/12			30	30			
4	36	1	0 TRAN VAN NAM	0	60	4 Vinh Bao		3	10/20/12	20		10	10			
5	15	2	1 NGO VAN PHONG	0	47	3 Le Chan, HP		3	10/20/12		30	10	30			
5	5	1	1 DINH QUANG HUNG	0	41	3 Vinh Bao, ..		3	11/22/12	20		10	0			
7	19	1	1 LE CONG UNG	0	40	2 Kien An, HP		3	11/17/12		20	10	0			
8	26	2	1 PHAM VAN HA	0	35	2 Ngo Cuyen..		3	11/22/12		20	10	0			
9	20	2	1 PHAM DUY KHANH	0	41	3 Ngo Cuyen ..		4	11/22/12		10	10	0			
10	14	2	0 LUU VAN THUY	0	40	2 Ngo Cuyen..		4	10/16/12	5	30	20	50		30	
11	22	2	1 PHAM THE BINH	0	35	2 Le Chan, HP		4	10/20/12	10		0		60		
12	3	2	1 DANG THI HONG	1	45	3 Le Chan, HP		1	12/25/12					30		
13	2	1	0 CHU THI LOAN	1	41	3 Ngo Cuyen..		1	10/20/12		20			30		
14	16	2	0 NGUYEN THI HANH	1	40	2 Kien An, HP		1	10/20/12	10		20		30		
15	33	1	1 VU THI TANG	1	39	2 Le Chan, HP		1	11/20/12		10	0		30		
15	34	2	1 VU THI TUYET	1	37	2 Kien An, HP		1	10/16/12			10	0	20		
17	23	2	0 TRAN THI GIAO	1	36	2 Kien An, HP		1	12/03/12	30	30	20	0		20	
18	5	1	1 HOANG THI SANG	1	26	1 Le Chan, HP		1	11/15/12			20	0		30	
19	39	1	1 NGUYEN THI LOAN	1	64	4 Le Chan, HP		2	11/22/12			0		30		
20	24	2	1 PHAM THI HUONG	1	41	3 Vinh Bao		2	12/25/12			30		30		
21	22	1	1 PHAM THI HIEM	1	40	2 Kien An, HP		2	10/20/12			30				
22	4	1	1 DAO THI DANH	1	37	2 Kien An, HP		2	12/25/12			30				
23	1	1 DINH THI DUNG	1	37	2 Hong Bang		2	05/10/12			10	30		30		

Hình 2.24

Các lựa chọn khác:

+ **Analyze all cases, do not create groups:** Phân tích tất cả các trường hợp, không tạo nhóm.

+ **Compare groups:** so sánh giữa các nhóm.

+ **Organize output by groups:** Tổ chức đầu ra của từng nhóm.

+ **Sort the file by grouping variables.** Phân loại tập tin bằng biến nhóm (Nếu các tập tin dữ liệu chưa được sắp xếp theo các giá trị của các biến nhóm).

+ **File is already sorted:** tập đã được sắp xếp.

2.6. Nối tệp tin dữ liệu

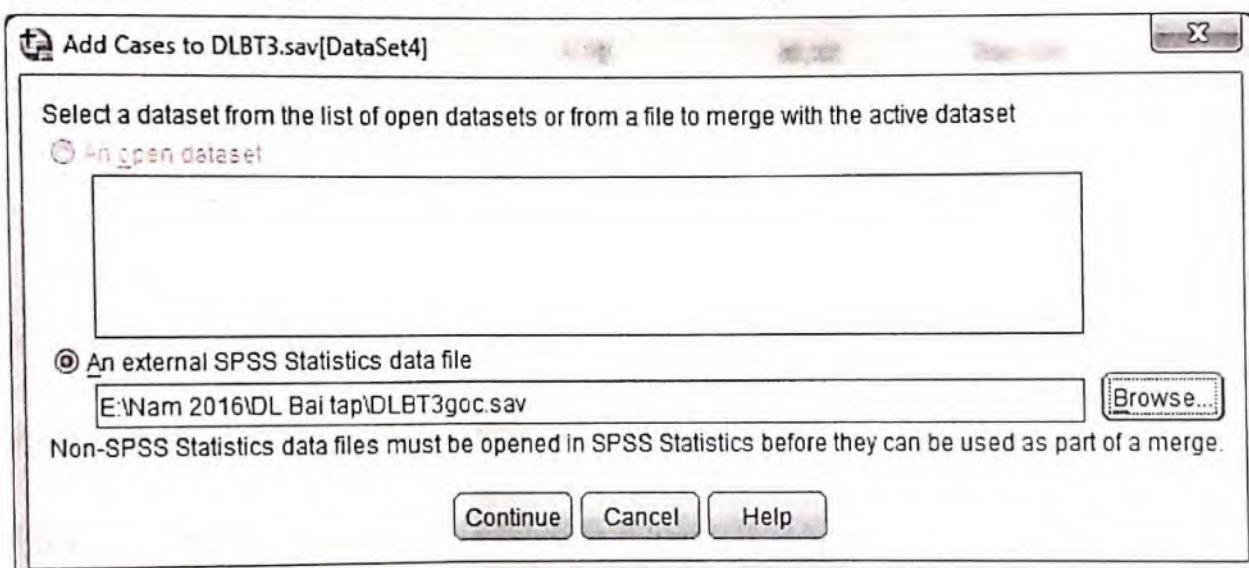
2.6.1. Nối thêm hàng

Nối thêm hàng: được sử dụng khi muốn nối thêm các hàng (**case**) có các biến tương tự nhau nhưng các trường hợp khác nhau. (ví dụ: chia dữ liệu cho hai người cùng nhập sau đó kết nối lại với nhau).

Để thực hiện việc nối thêm hàng, ta phải chuẩn bị 2 tệp dữ liệu **nguồn** và **đích** (*chứa bảng có cùng cấu trúc*).

Cách bước thực hiện:

- Mở tệp dữ liệu cần nối (tệp đích)
- Chọn **Data → Merge files → Add cases...** xuất hiện hộp hội thoại:



Hình 2.25

+ **An open dataset:** nếu tệp nguồn đang mở.

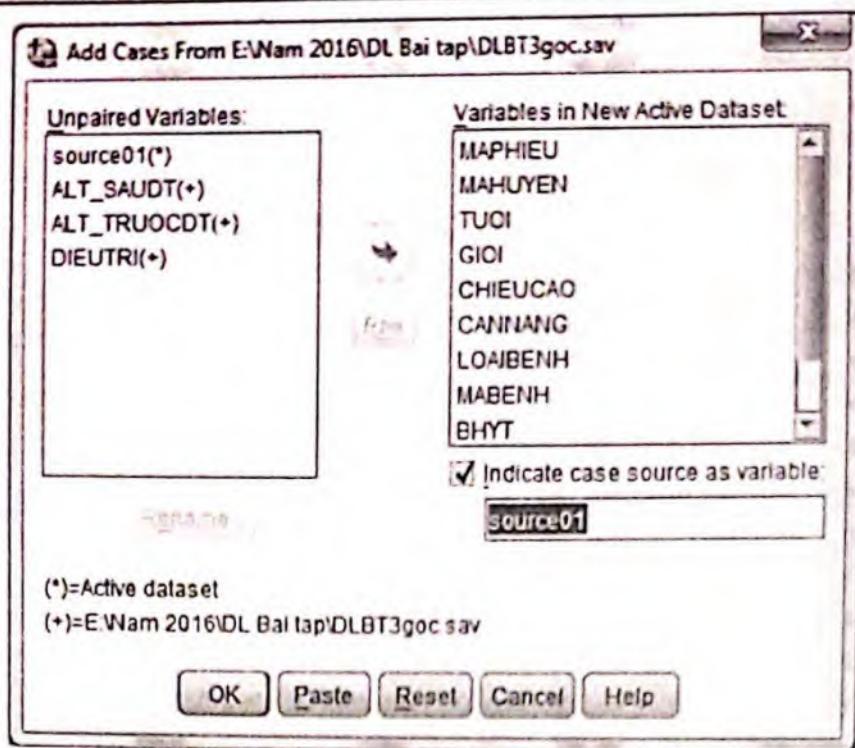
+ **An external SPSS statistics data file:** khai báo đường dẫn tệp nguồn (hoặc nhấn nút **Browse** chọn file cần nối)

- Chọn **Continue**, xuất hiện hộp hội thoại - *Hình 2.26*

+ **Variables in New Active Dataset:** chứa các biến của tệp dữ liệu đích.

+ **Unpaired Variables** (biến lẻ): nếu tệp dữ liệu nối vào có nhiều biến hơn sẽ được hiện thị. (Ta có thể chọn biến trong danh sách ở mục **Unpaired Variables** để đưa sang mục **New Active Dataset** để thêm vào tệp đích, hoặc loại bỏ biến ở tệp nguồn ở mục **New Active Dataset**.)

+ **Indicate case source as variable:** để biết được nguồn kết quả của tệp đích (0: là giá trị của các biến trong tệp đích, 1 là giá trị biến của tệp nguồn).



Hình 2.26

- Chọn **OK** sẽ cho ta kết quả là tệp đích.

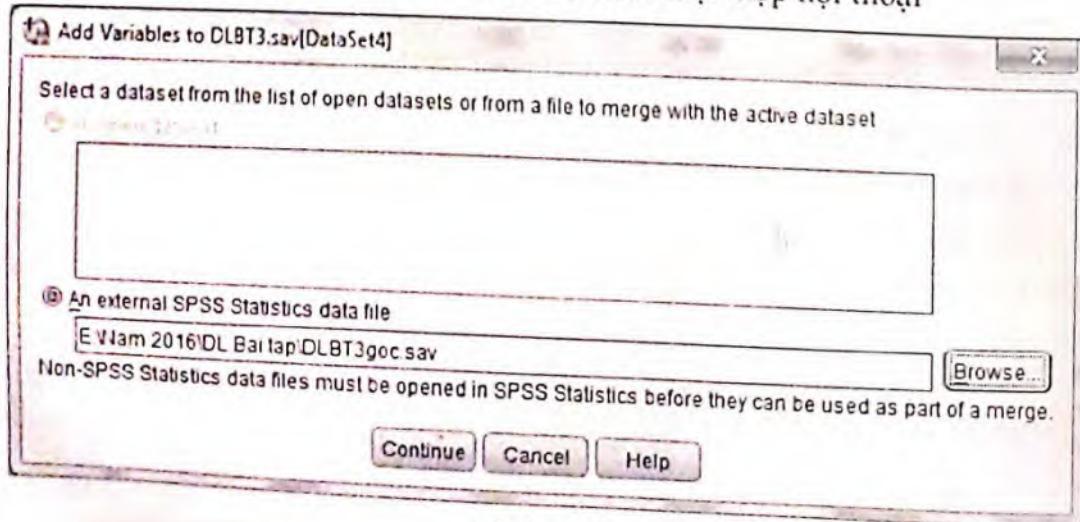
2.6.2. Nối thêm biến (Variables)

Nối thêm biến hay là liên kết các bảng dữ liệu với bảng đang mở, được dùng trong trường hợp các tệp tin dữ liệu có chứa các trường hợp tương tự, nhưng các biến khác nhau thông qua **biến khóa** (*trường khóa*).

Để thực hiện nối thêm biến ta chuẩn bị ít nhất 2 tệp chứa bảng dữ liệu có chung biến khóa và đã được sắp xếp trên biến khóa (biến khóa của 2 tệp phải trùng tên và cùng kiểu dữ liệu).

Cách thức thực hiện:

- Mở tệp dữ liệu cần nối (tệp đích).
- Chọn **Data → Merge files → Add Variables...** xuất hiện hộp hội thoại

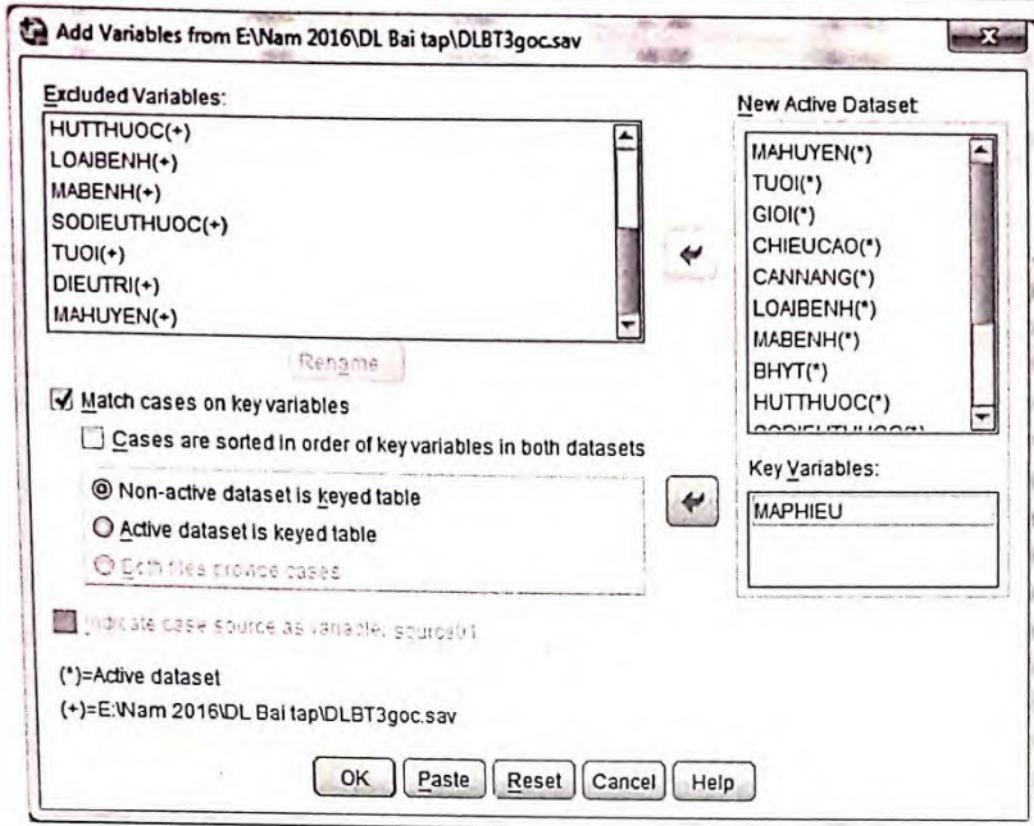


Hình 2.27

+ **An open dataset:** nếu tệp nguồn đang mở.

+ **An external SPSS statistics data file:** khai báo đường dẫn tệp nguồn (hoặc nhấn nút **Browse** chọn tệp nguồn)

- Chọn **Continue**, xuất hiện hộp hội thoại - *Hình 2.28*



Hình 2.28

+ Trong mục **Excluded Variables**: chọn biến làm khóa liên kết; đánh dấu chọn

Match cases on key variables và sử dụng nút để đưa biến khóa vào mục **Key Variables**.

+ **New Active Dataset**: liệt kê các biến sẽ có trong tệp cơ sở dữ liệu sau khi liên kết, các trường đánh dấu (*) là các trường có trong CSDL đích, các trường đánh dấu (+) là các trường có trong CSDL nguồn.

- Chọn **OK** sẽ cho ta kết quả là tệp đích.

BÀI 3. TRÌNH BÀY DỮ LIỆU**Mục tiêu:**

- Lập các bảng tần số theo yêu cầu thống kê.
- Tính các đại lượng thống kê mô tả.
- Vẽ biểu đồ biểu diễn các đại lượng thống kê theo yêu cầu.
- Đọc và phiên giải kết quả thu được.

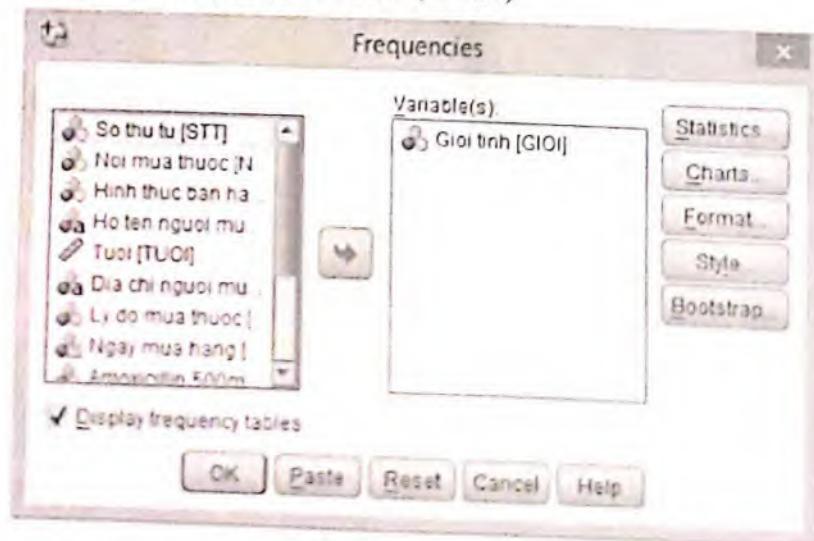
3.1. Lập bảng tần số các giá trị của biến

Một trong các mục tiêu của việc tạo bảng tần số là để phát hiện những giá trị lạ xuất hiện trong một biến nào đó. Chẳng hạn, trong bộ số liệu nghiên cứu có biến **GIOI** (*Giới tính*) chứa hai giá trị 1 (*Nữ*), 0 (*Nam*). Sau khi lập bảng tần số của biến Giới tính ta thấy kết quả có thêm giá trị 11, chứng tỏ dữ liệu của biến **GIOI** bị sai cần phải sửa (*để tìm ra các giá trị sai này, ta có thể sử dụng lệnh Find*).

- Đối với biến định tính: Lệnh **Frequencies** thường dùng để tính số lượng (*Frequency*) và tỷ lệ phần trăm (*Percent*)... các giá trị của một hoặc nhiều biến.
- Đối với biến định lượng: Lệnh **Frequencies** ngoài việc xác định tần số xuất hiện các giá trị của biến còn dùng để tính các đại lượng thống kê mô tả (hay tham số đặc trưng) của biến đó.

Ví dụ 1: Thống kê số lượng và tỷ lệ phần trăm các giá trị của biến **GIOI**.

- Chọn **Analyze → Descriptive Statistics → Frequencies**, xuất hiện hộp thoại - *Hình 3.1*
- Chọn biến cần tính tần số và tỷ lệ phần trăm (**GIOI**)

*Hình 3.1*

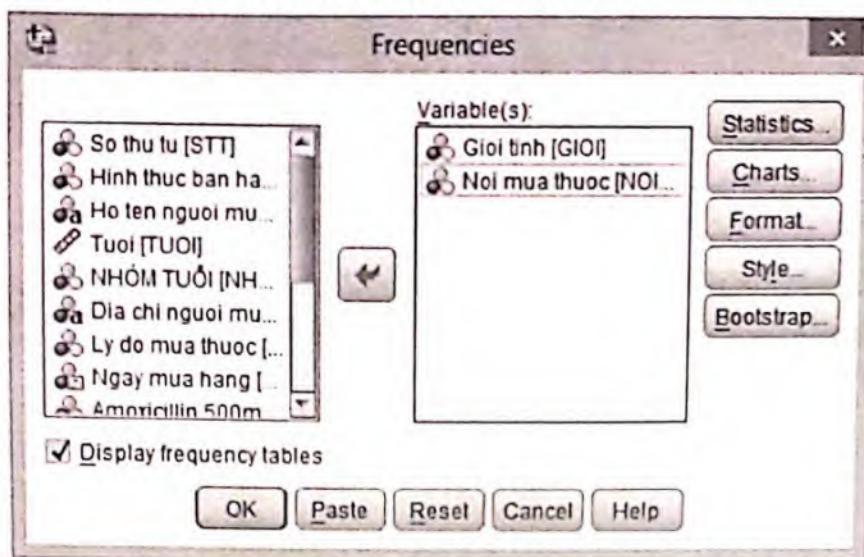
- Chọn **OK** để kết thúc và quan sát kết quả:

Gioi tinh					
		Frequency	Percent	Valid Percent	Cumulative Percent
		Valid	NAM	NU	Total
		11	27.5	27.5	27.5
		29	72.5	72.5	100.0
		40	100.0	100.0	

Bảng kết quả thể hiện số lượng (*Frequency*) và tỷ lệ phần trăm (*Percent*), phần trăm hợp lệ (*Valid Percent*), tỷ lệ cộng dồn (*Cumulative Percent*) các giá trị của biến **GIOI**.

Ví dụ 2: Lập bảng tần số cho nhiều biến định tính: GIOI, NOI MUA.

- Chọn Analyze → Descriptive Statistics → Frequencies.
- Chọn biến cần tính tần số GIOI, NOI MUA - Hình 3.2



Hình 3.2

- Chọn OK để kết thúc và quan sát kết quả:

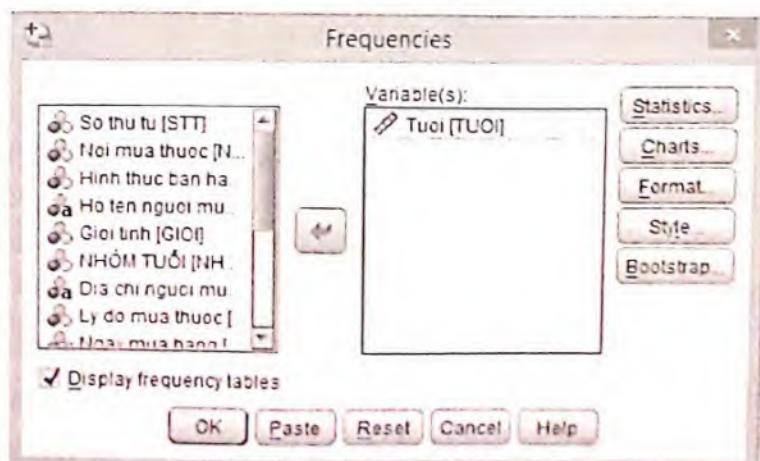
Gioi tinh					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nam	11	27.5	27.5	27.5
	Nu	29	72.5	72.5	100.0
	Total	40	100.0	100.0	

Noi mua thuoc					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nha thuoc	23	57.5	57.5	57.5
	Quay thuoc	17	42.5	42.5	100.0
	Total	40	100.0	100.0	

Ví dụ 3: Lập bảng tần số các giá trị của biến TUOI và cho biết tỷ lệ phần trăm những người dưới 35 tuổi.

- Chọn Analyze → Descriptive Statistics → Frequencies.

- Chọn biến cần tính tần số TUOI
(Hình 3.3)



Hình 3.3

- Nhấp OK và quan sát kết quả

Statistics

Tuoi

N	Valid	40
	Missing	0

Tuoi

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20	2	5.0	5.0	5.0
	25	1	2.5	2.5	7.5
	28	1	2.5	2.5	10.0
	31	1	2.5	2.5	12.5
	32	1	2.5	2.5	15.0
	34	2	5.0	5.0	20.0
	35	4	10.0	10.0	30.0
	36	1	2.5	2.5	32.5
	...				
	60	1	2.5	2.5	95.0
	62	1	2.5	2.5	97.5
	64	1	2.5	2.5	100.0
Total		40	100.0	100.0	

Trong trường hợp này cần đọc giá trị thống kê ở cột **Cumulative Percent** (*Tỷ lệ cộng đồng*) và được kết quả là **20.0%**.

3.2. Tính các đại lượng trong thống kê mô tả

SPSS có thể mô tả biến định lượng bằng cách tính các đại lượng thống kê mô tả (còn gọi là *các tham số đặc trưng của biến đó* như trung bình, trung vị, phương sai, độ lệch chuẩn...);

Để tính các đại lượng thống kê mô tả của biến định lượng ta có thể sử dụng các lệnh: **Frequencies, Descriptives, Explore**.

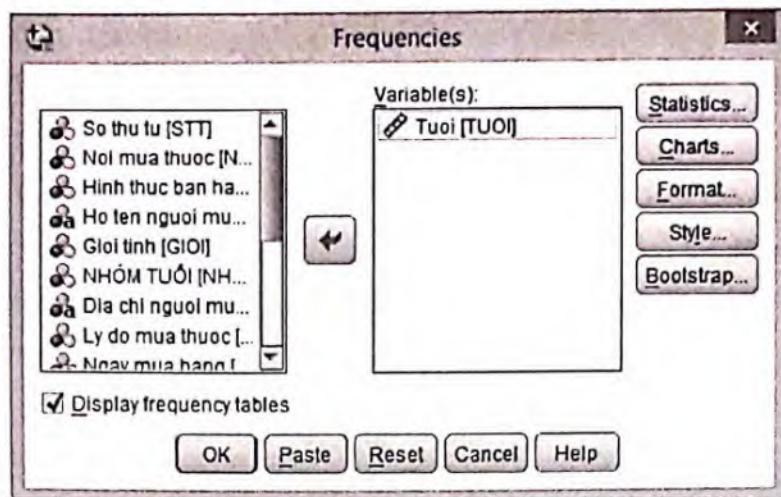
3.2.1. Lệnh Frequencies

Lệnh **Frequencies** có thể dùng để tạo bảng tần số (đối với biến định tính) và bảng các đại lượng thống kê mô tả (đối với biến định lượng) kèm theo các biểu đồ; Tuy nhiên lệnh này không tạo được bảng tần số hoặc bảng tham số đặc trưng nếu muốn đưa vào một biến khác dùng để phân nhóm...

Ví dụ: Tính các đại lượng thống kê mô tả của biến Tuổi, thực hiện như sau

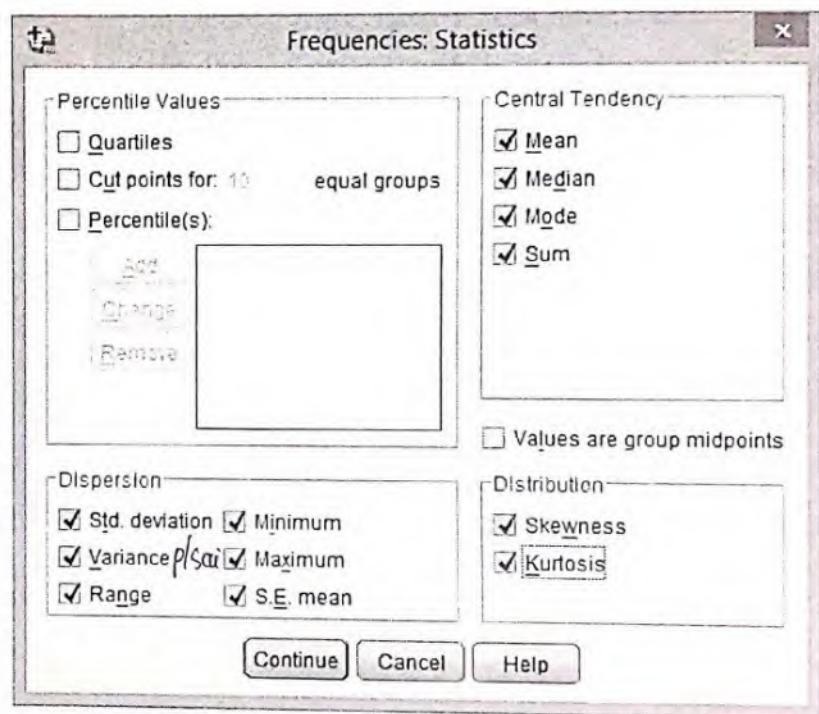
- Chọn **Analyze → Descriptive Statistics → Frequencies**.

- Chọn biến cần tính các đại lượng thống kê mô tả (*Tuổi*) - *Hình 3.4*



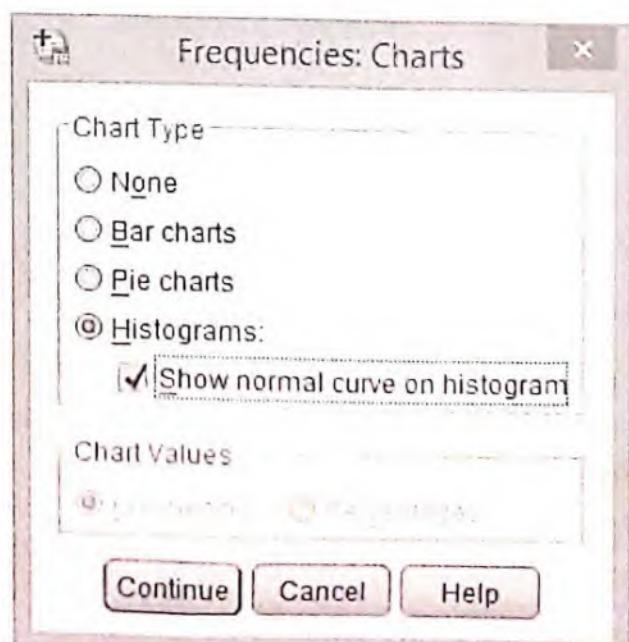
Hình 3.4

- Nhấp chọn nút Statistics để hiện cửa sổ và chọn các tham số đặc trưng cần mô tả. - *Hình 3.5*. Chọn Continue để quay lại hộp thoại Frequencies - *Hình 3.4*



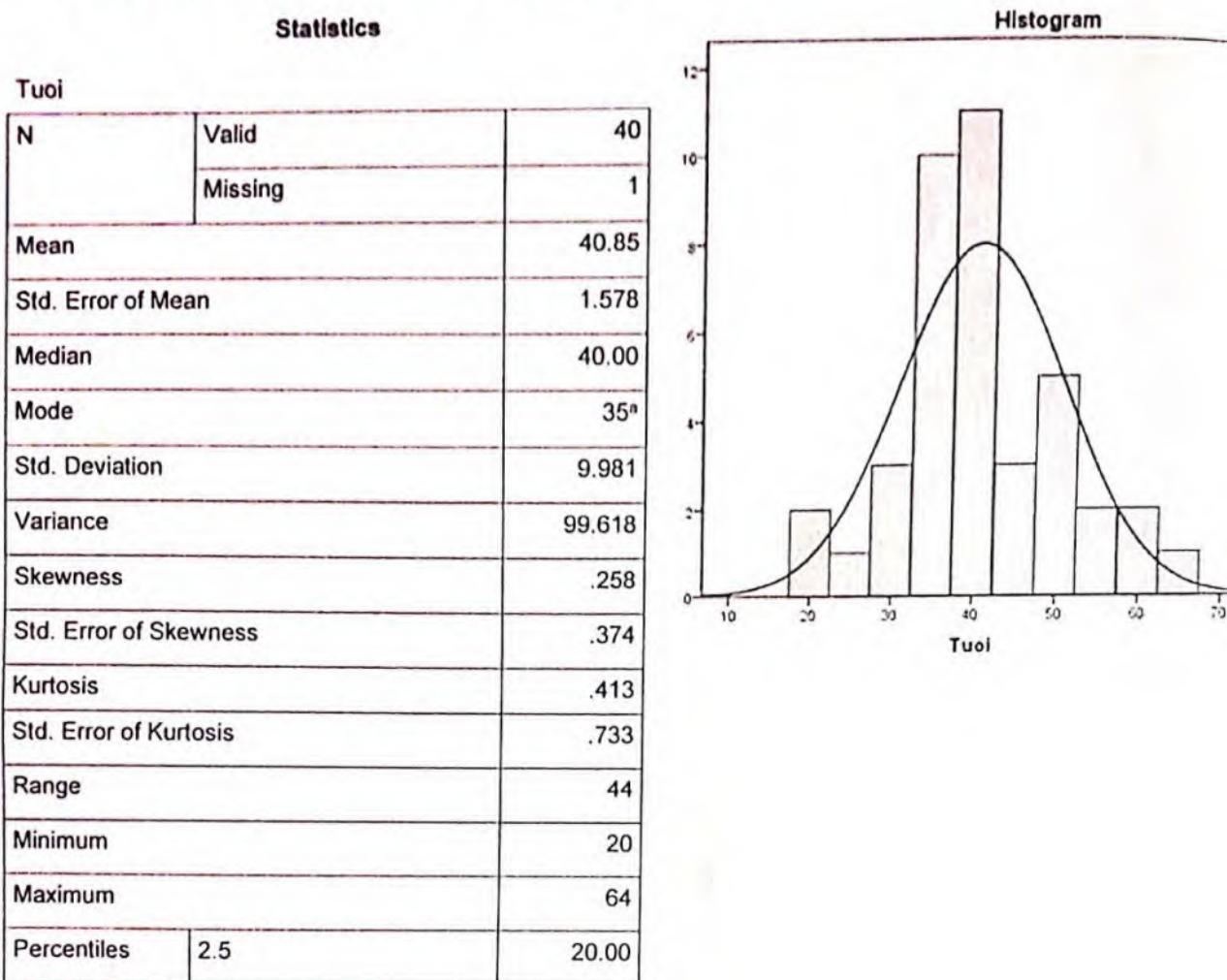
Hình 3.5

- Nhấp chọn nút Charts, xuất hiện cửa sổ Frequencies: Charts - *Hình 3.6*. đánh dấu vào ô Show normal curve on histogram → Chọn Continue để quay lại hộp thoại Frequencies - *Hình 3.4*



Hình 3.6

- Nhập OK để kết thúc lệnh và quan sát kết quả:



Ý nghĩa các tham số trong bảng Statistics:

$$\sum_{i=1}^n x_i$$

- *Mean*: Trung bình cộng mẫu $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$.

- *Sum*: Tổng cộng.

- *Std Deviation*: Độ lệch chuẩn mẫu $Sd = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}}$

- *Max, Min*: Giá trị lớn nhất, giá trị nhỏ nhất.

- *Median*: Trung vị, là giá trị chia số liệu thành hai nhóm có tần số bằng nhau.

- *Mode*: là giá trị có tần số cao nhất (*hay giá trị lặp lại nhiều nhất*).

- *Variance*: Hệ số biến thiên (phương sai), tham số này càng lớn thì độ biến thiên của dữ liệu càng lớn.

- *Std Error Of Mean*: Sai số chuẩn khi ước lượng trung bình $SE = \frac{Sd}{\sqrt{n}}$.

- **Skewness:** Hệ số bất đối xứng, phân phối chuẩn có hệ số này là 0, nếu hệ số này càng khác không thì phân phối càng bất đối xứng, dấu của hệ số này cho biết đồ thị hàm mật độ là lệch trái hay lệch phải.

- **Kurtosis:** Hệ số độ nhọn, phân phối chuẩn có hệ số độ nhọn là 3, nếu lớn quá đồ thị hàm mật độ sẽ quá nhọn, nếu bé quá thì đồ thị sẽ quá tù. γ_3 quá nhọn γ_3 quá tù.

- **Các hệ số như:** Hệ số bất đối xứng và hệ số độ nhọn thường được sử dụng để kiểm định số liệu có tuân theo phân phối chuẩn hay không?

Biểu đồ:

Có nhiều dạng biểu đồ, khi cần vẽ chỉ cần click nút charts... sẽ xuất hiện giao diện như sau:

- **Bar Charts:** Biểu đồ hình cột.

- **Pie Charts:** Biểu đồ hình tròn.

- **Histograms:** Biểu đồ tần số thể hiện hình ảnh hàm mật độ xác suất, nếu đánh dấu mục *With normal curve* thì sẽ có thêm đường cong biều thị hàm mật độ được SPSS thêm vào. Đây là một công cụ trực quan để kiểm định tính chuẩn của một phân phối, nếu nó có dạng chuông úp đối xứng thì phân phối sẽ có dạng là phân phối chuẩn.

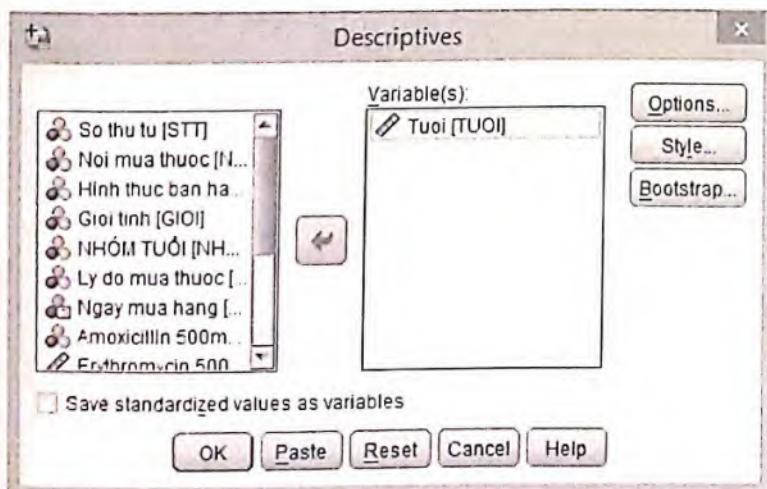
3.2.2. Lệnh Descriptives

Lệnh Descriptives chỉ áp dụng với biến định lượng (*Biến định tính sau khi mô tả cho kết quả nhưng không có ý nghĩa trong thống kê*).

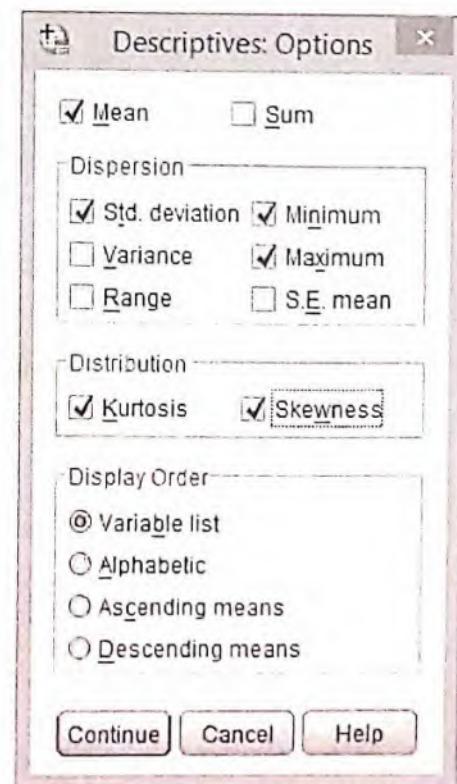
Ví dụ: Tính tham số đặc trưng của biến TUOI (Tuổi)

- Chọn Analyze → Descriptive Statistics → Descriptives, xuất hiện cửa sổ Descriptives

- Chọn biến cần mô tả/tính tham số đặc trưng (*Tuổi*) ở khung bên trái của cửa sổ lệnh Descriptives đưa sang khung Variable(s) – Hình 3.7



Hình 3.7



Hình 3.8

- Chọn nút Options... để mở hộp thoại Descriptives: Options, đánh dấu vào các tham số đặc trưng cần mô tả - *Hình 3.8*. Chọn Continue để quay lại hộp thoại Descriptives.

- Nhập OK để kết thúc lệnh và quan sát kết quả:

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Tuoi	40	44	20	64	1634	40.85	1.578
Valid N (listwise)	40						

	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Tuoi	9.981	99.618	.258	.374	.413	.733
Valid N (listwise)						

3.2.3. Lệnh Explore

Lệnh này sẽ thực hiện các công việc sau:

- Tính các tham số đặc trưng cho các biến theo nhóm (*mỗi nhóm là một giá trị khác nhau có trong biến định tính phân nhóm*).

- Nhận diện các giá trị bất thường trong từng biến, cảnh báo dữ liệu mắc sai số thô.
- Tính toán các phân vị theo các mức.
- Tạo biểu đồ cho từng nhóm, từ đó có cái nhìn khái quát về dữ liệu theo nhóm.

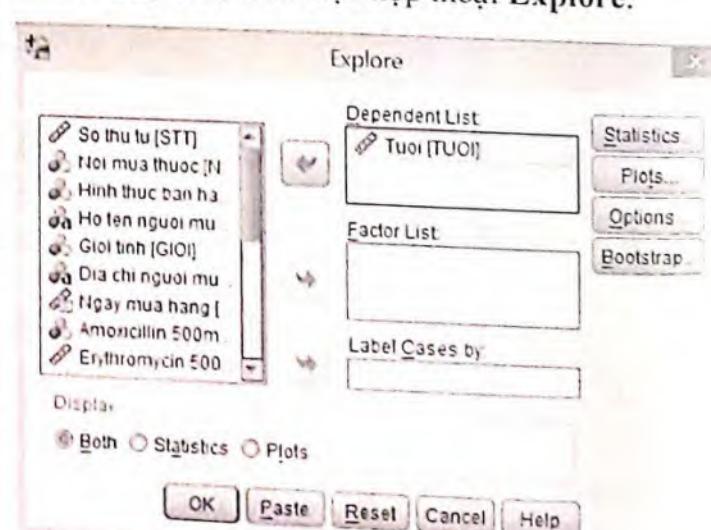
Ví dụ 1: Tính tham số đặc trưng của biến TUOI (Tuổi)

- Chọn Analyze → Descriptive Statistics → Explore, xuất hiện hộp thoại Explore.

- Chọn các biến cần phân tích (*Tuổi*) – *Hình 3.9*

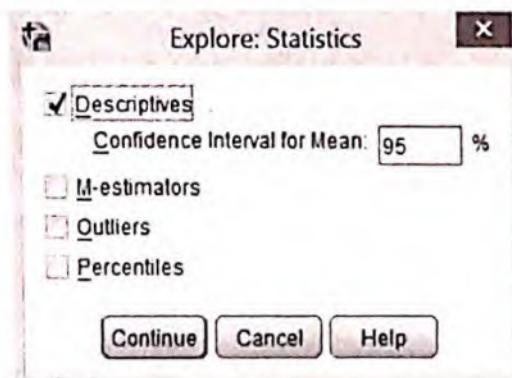
- Mục Display: chọn các chế độ hiển thị kết quả:

- **Statistics:** Chỉ hiển thị các tham số thống kê cơ bản.
- **Plots:** Chỉ hiển thị biểu đồ.
- **Both:** Hiển thị cả tham số thống kê và các biểu đồ.



Hình 3.9

- Chọn nút **Statistics** ở *Hình 3.9* để mở hộp thoại **Explore: Statistics** và đánh dấu mục **Descriptives** như *Hình 3.10*. để đưa ra các tham số thống kê cơ bản. Chọn **Continue** để quay lại hộp thoại **Explore**.



Hình 3.10

- Chọn nút **Plots** ở *Hình 3.9* để mở hộp thoại **Explore: Plots - Hình 3.11**

+ Ở khung **Boxplots**:

- Nếu chọn **None**: không hiển thị biểu đồ hộp.
- Nếu chọn **Factor levels together** hoặc **Dependents together**: dùng để hoán đổi vị trí các biến trong mục **Dependent List** nếu có hai biến trở lên trong mục này.

+ Ở khung **Descriptive**: chọn loại biểu đồ muốn hiển thị.

- **Stem-and-leaf**: biểu đồ thân – lá.
- **Histogram** : Biểu đồ tần suất.

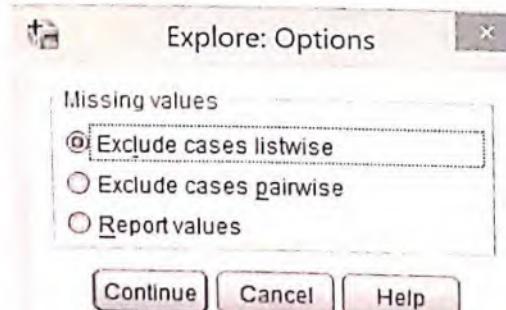


Hình 3.11

+ Nếu đánh dấu mục **Normality plots with tests** sẽ vẽ thêm biểu đồ xác suất chuẩn *Q-Q* để kiểm tra dữ liệu có tuân theo phân phối chuẩn hay không.

→ Chọn **Continue** để quay lại hộp thoại **Explor.**

- Chọn nút **Options** ở *Hình 3.9* để xuất hiện hộp thoại - *Hình 3.12*:



Hình 3.12

+ **Exclude cases listwise**: Các giá trị khuyết thiếu trên một trong các biến của **Dependent List** hay **Factor List** đều bị bỏ qua.

+ **Exclude cases pairwise:** Sử dụng tất cả các giá trị không khuyết thiếu tại các biến cần thiết để tính toán, như vậy SPSS sẽ sử dụng tối đa lượng thông tin được lưu trong dữ liệu để tính toán.

- Chọn OK để kết thúc lệnh.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Tuổi	40	97.6%	1	2.4%	41	100.0%

Descriptives

Tuổi	Mean		Statistic	Std. Error
	85% Confidence Interval for Mean	Lower Bound	37.66	
		Upper Bound	44.04	
5% Trimmed Mean		40.78		
Median		40.00		
Variance		99.618		
Std. Deviation		9.981		
Minimum		20		
Maximum		64		
Range		44		
Interquartile Range		13		
Skewness		.258	.374	
Kurtosis		.413	.733	

Giải thích biểu đồ Stem-and-leaf

```
Tuoi Stem-and-Leaf Plot
Frequency Stem & Leaf
2.002 .00
2.002 .58
4.003 .1244
9.003 .555567779
10.004 .0000111122
7.004 .5678888
2.005 .13
1.005 .6
3.006 .024
Stem width:10
Each leaf:1 case(s)
```

Trong biểu đồ bên *Stem width* là 10 có nghĩa phần thân tính là 10, lá *Leaf* được tinh thành 1. Như vậy, có hai giá trị 20; một giá trị 25, một giá trị 28, một giá trị 31, một giá trị 32, hai giá trị 34....

Ví dụ 2: Tính tuổi trung bình theo giới tính.

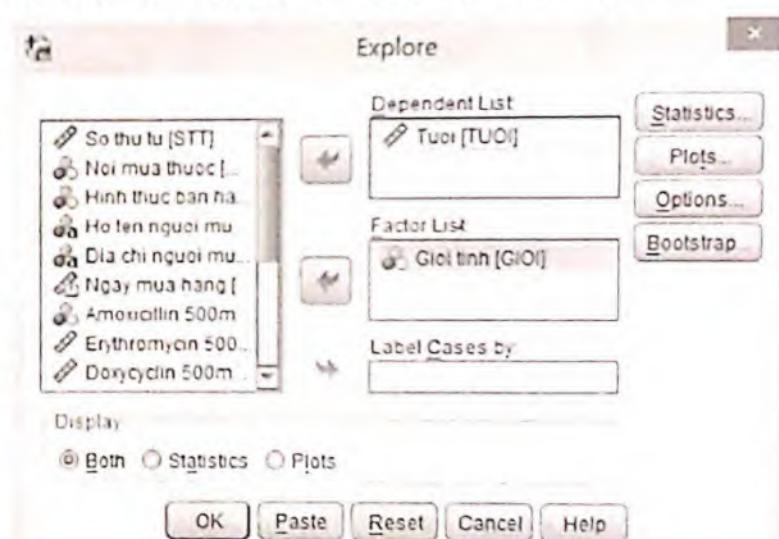
- Chọn **Analyze → Descriptive Statistics → Explore**, xuất hiện hộp thoại **Explore**.

- **Dependent List:** Biến cần phân tích (*Tuổi*) - *Hình 3.13*

- **Factor List:** Biến dùng để phân nhóm (*Giới*) - *Hình 3.13*

- Chọn các nút **Statistics**, **Plots**, **Options** để chọn các mục cần tính toán hoặc hiển thị.

- Nhập **OK** và quan sát kết quả.



Hình 3.13

Case Processing Summary

Gioi tinh		Cases					
		Valid		Missing		Total	
Tuoi	Nam	N	Percent	N	Percent	N	Percent
	Nu	29	100.0%	0	.0%	29	100.0%

Descriptives

	Gioi tinh		Statistic	Std. Error
Tuoi	Nam	Mean	42.55	3.019
		95% Confidence Interval for Mean	Lower Bound	35.82
			Upper Bound	49.27
		5% Trimmed Mean	42.05	
		Median	40.00	
		Variance	100.273	
		Std. Deviation	10.014	
		Minimum	32	
		Maximum	62	
		Range	30	
		Interquartile Range	12	
		Skewness	1.250	.661
		Kurtosis	.549	1.279
	Nu	Mean	40.21	1.870
		95% Confidence Interval for Mean	Lower Bound	36.38
			Upper Bound	44.04
		5% Trimmed Mean	40.15	
		Median	40.00	
		Variance	101.384	
		Std. Deviation	10.069	
		Minimum	20	
		Maximum	64	
		Range	44	
		Interquartile Range	14	
		Skewness	-.017	.434
		Kurtosis	.353	.845

Tests of Normality

Gioi tinh		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Tuoi	Nam	.289	11	.011	.823	11	.019
	Nu	.096	29	.200*	.981	29	.859

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

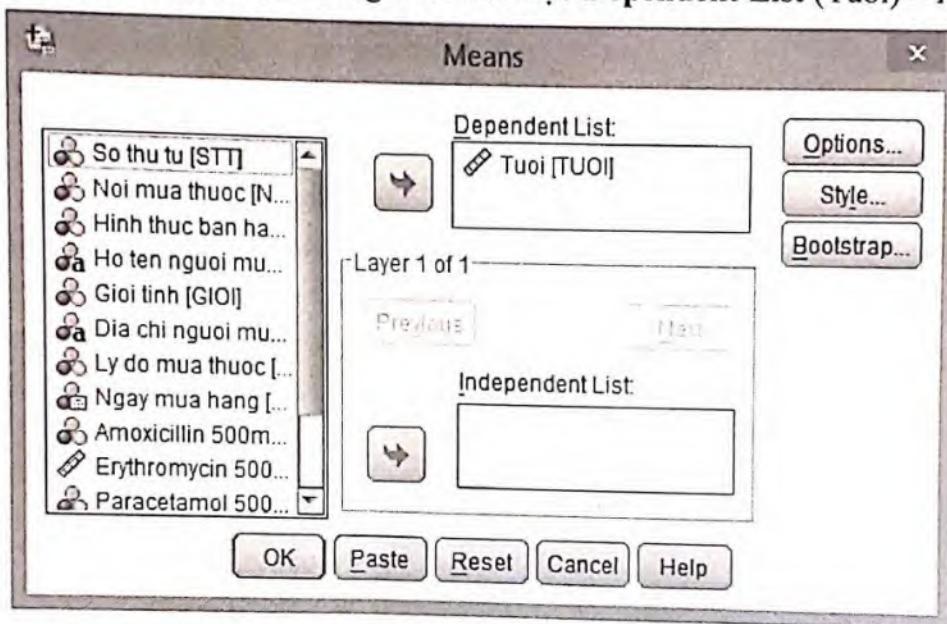
3.2.4. Lệnh Mean

Lệnh này sẽ thực hiện các công việc sau:

- Tính các tham số đặc trưng: Median, Mean, Maximum, Minimum, Std. Error of Mean, Sum, Range, First, Last, Kurtosis, Std. Error of Kurtosis, Skewness, Std. Error of Skewness, Standard Deviation....

Ví dụ 1: Tính tuổi (TUOI) trung bình của những người trong nghiên cứu.

- Chọn Analyze → Compare Means → Mean xuất hiện hộp thoại – *Hình 3.14*
- Chọn biến cần tính giá trị trung bình vào mục Dependent List (Tuoi) – *Hình 3.14*

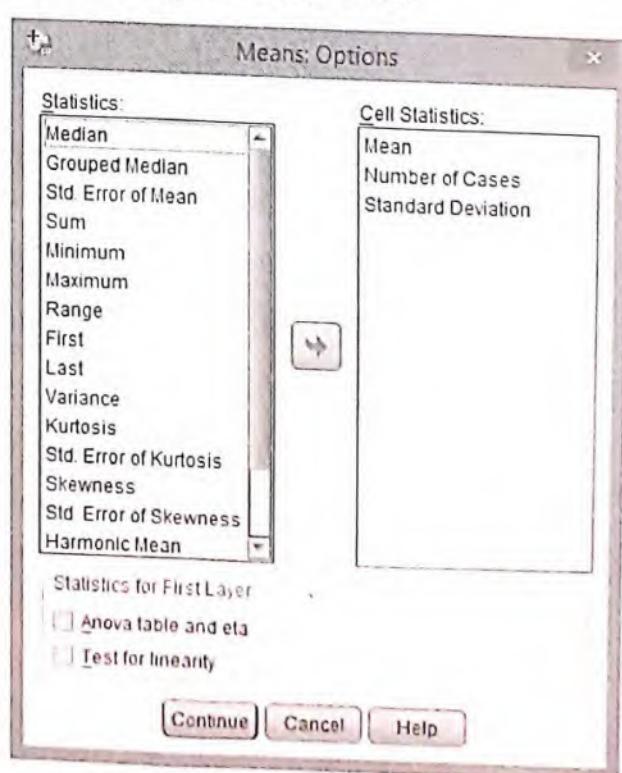


Hình 3.14

- Nhấn Options... để lựa chọn thêm các giá trị cần thống kê – *Hình 3.15..*

Lưu ý: Có thể chọn thêm giá trị thống kê trong mục Statistics For First Layer.

- + Anova table and eta.
- + Test For Linearity.
- Sau đó nhấn Continue để quay lại *Hình 3.14*



Hình 3.15

- Nhấn OK để kết thúc và quan sát kết quả.

Means

Case Processing Summary

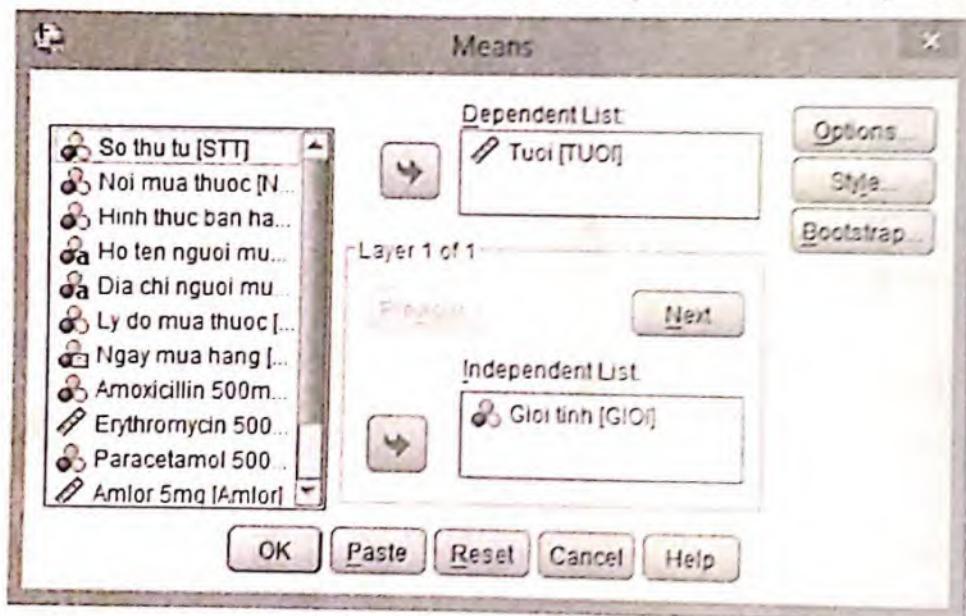
	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Tuoi	40	100.0%	0	0.0%	40	100.0%

Report

Tuoi		
Mean	N	Std Deviation
40.85	40	9.981

Ví dụ 2: Tính tuổi (TUOI) trung bình theo giới (GIOI).

- Chọn Analyze → Compare Means → Mean xuất hiện hộp thoại – *Hình 3.15*
- Chọn biến cần tính giá trị trung bình vào mục Dependent List (Tuoi) – *Hình 3.15*



Hình 3.15

- Nhấn OK và quan sát kết quả.

⇒ Means

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Tuoi * Gioi tinh	40	100.0%	0	0.0%	40	100.0%

Report

Tuoi				
Gioi tinh	Mean	N	Std Dev	Double-click to activate
NAM	42.55	11	10.014	
NU	40.21	29	10.069	
Total	40.85	40	9.981	

3.3. Lập các bảng tổng hợp nhiều biến

Có thể sử dụng nhiều lệnh để tạo ra các bảng thống kê có cấu trúc khác nhau tùy theo nhu cầu của người thực hiện. Trong mỗi lệnh lưu ý các nút cung cấp các lựa chọn hiển thị các tham số thống kê, hình thức thức hiển thị như nút *Statistics, Format, Cells...*

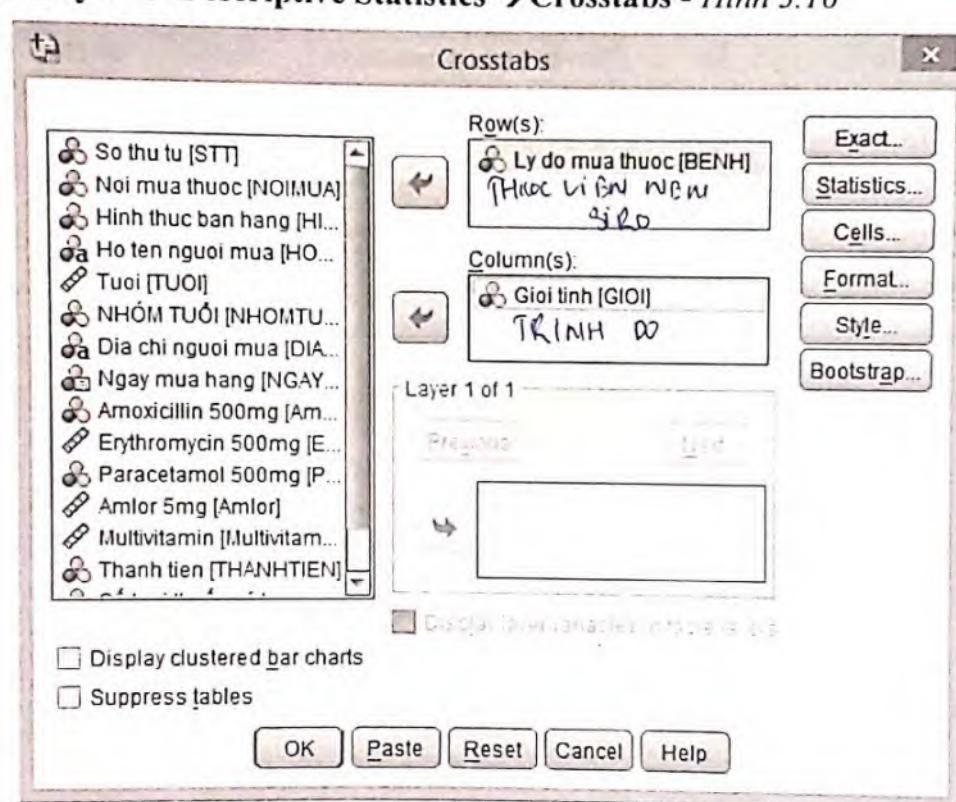
Để tạo các bảng tổng hợp, ngoài lệnh **Explore** có thể sử dụng các lệnh: **Crosstabs, Custom Tables...**

3.3.1. Lệnh Crosstabs

Lệnh **Crosstabs** thường dùng để tạo bảng tổng hợp dữ liệu của các biến định tính.

Ví dụ: Lập bảng thống kê lý do mua thuốc theo Giới tính.

- Chọn **Analyze → Descriptive Statistics → Crosstabs - Hình 3.16**



Hình 3.16

- **Row(s):** Chọn biến

- Một số lựa chọn trên hộp thoại **Crosstabs:**

+ **Statistics...** : Chọn các test thống kê.

+ **Cells...** : Chọn cách hiển thị dữ liệu trong các ô của bảng thống kê *độ rộng Row, column*.

+ **Format...** : Chọn cách sắp xếp dữ liệu trên bảng kết quả.

- Nhập **OK** để kết thúc lệnh và quan sát kết quả :

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Ly do mua thuốc * Gioi tinh	40	100.0%	0	.0%	40	100.0%

Lý do mua thuốc * Giới tính Crosstabulation

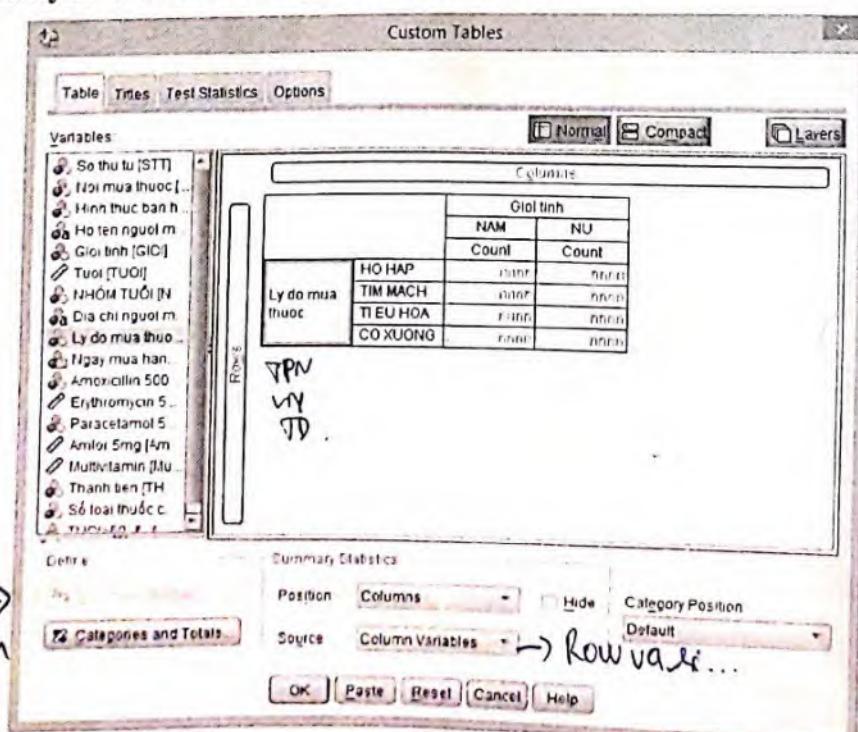
		Giới tính		Total	
		NAM	NU		
Lý do mua thuốc	HO HẠP	Count	1	7	
		% within Lý do mua thuốc	12.5%	87.5%	
		% within Giới tính	9.1%	24.1%	
		% of Total	2.5%	20.0%	
	TIM MACH	Count	2	6	
		% within Lý do mua thuốc	25.0%	75.0%	
		% within Giới tính	18.2%	20.7%	
		% of Total	5.0%	20.0%	
TÌ EU HOA	TI EU HOA	Count	5	6	
		% within Lý do mua thuốc	45.5%	54.5%	
		% within Giới tính	45.5%	20.7%	
		% of Total	12.5%	27.5%	
	CO XUONG KHOP	Count	3	10	
		% within Lý do mua thuốc	23.1%	76.9%	
		% within Giới tính	27.3%	34.5%	
		% of Total	7.5%	32.5%	
Total		Count	11	29	
		% within Lý do mua thuốc	27.5%	72.5%	
		% within Giới tính	100.0%	100.0%	
		% of Total	27.5%	100.0%	

3.3.2. Lệnh Custom Tables

Lệnh Custom Tables để tạo bảng giao

Ví dụ 1: Lập bảng thống kê lý do mua thuốc theo Giới tính.

- Chọn Analyze → Tables → Custom Tables.



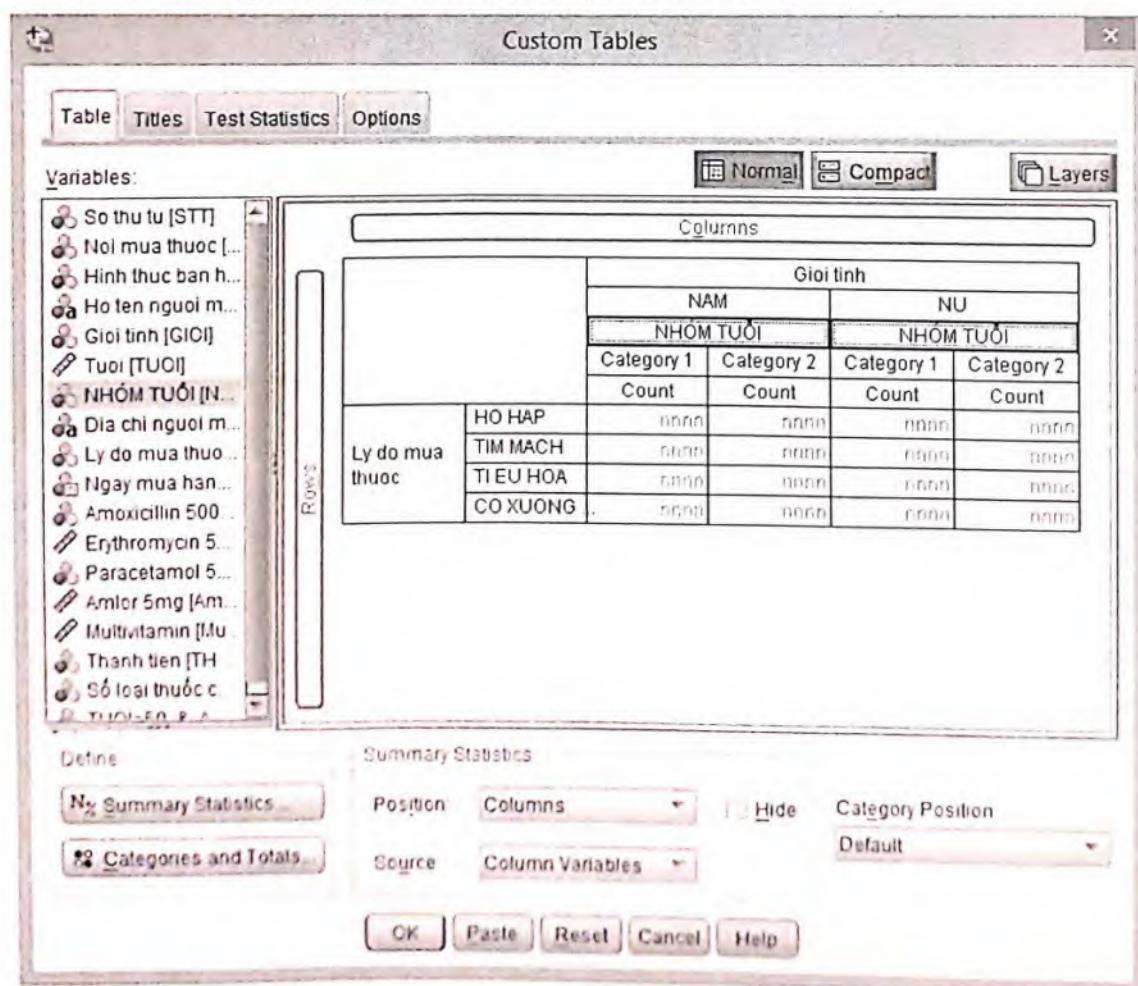
Hình 3.17

- Chọn các biến cần phân tích. (Dùng chuột kéo – thả)
 - + **Columns:** Chọn biến cần phân tích theo cột (*Gioi tinh*) - *Hình 3.17*
 - + **Rows:** Chọn biến cần phân tích theo hàng (*Ly do mua thuoc*) - *Hình 3.17*
- Chọn **OK** và quan sát kết quả.

		Gioi tinh	
		NAM	NU
		Count	Count
Ly do mua thuoc	Ho hap	1	7
	Tim mach	2	6
	Tieu hoa	5	6
	Co xuong khop	3	10

Ví dụ 2: Lập bảng thống kê lý do mua thuốc của từng Nhóm tuổi theo Giới tính.

- Chọn **Analyze → Tables → Custom Tables**.
- Chọn các biến cần phân tích.
 - + **Columns:** Chọn biến (*Gioi tinh*) và (*Nhom tuoi*) - *Hình 3.18*
 - + **Rows:** Chọn biến (*Ly do mua thuoc*) - *Hình 3.18*



Hình 3.18

- Nhập **OK** và quan sát kết quả.

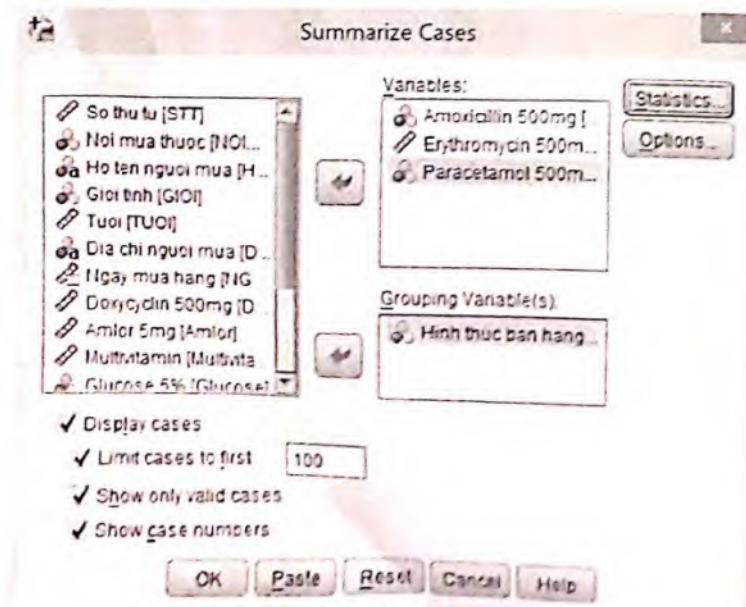
		Giới tính							
		NAM				NU			
		NHÓM TUỔI				NHÓM TUỔI			
		1	2	3	4	1	2	3	4
Ly do mua thuốc	Ho hấp	0	1	0	0	1	4	2	0
	Tim mạch	0	1	0	1	0	4	1	1
	Tiêu hóa	0	2	2	1	1	1	4	0
	Co xương khớp	0	2	1	0	2	2	3	3

3.3.3. Lệnh Case Summaries...

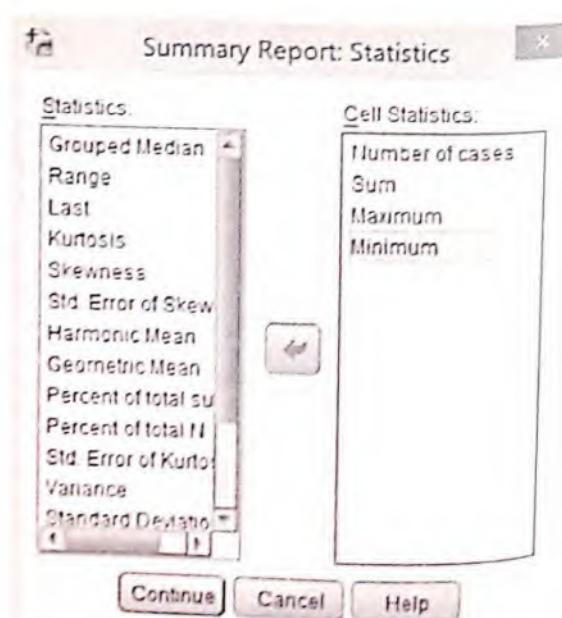
Lệnh này cho phép người dùng tạo ra các bảng báo cáo tóm tắt theo hàng với các đại lượng thống kê tùy ý, người dùng có thể đưa vào các biến để phân nhóm (*Grouping Variable(s)*). Với bộ dữ liệu lớn, có thể chọn n trường hợp đầu tiên để thống kê.

Ví dụ 1: Lập bảng tóm tắt tình hình bán các loại thuốc theo Hình thức bán hàng.

- Chọn Analyze → Reports → Case Summaries, xuất hiện hộp thoại Summarize Cases.
- Chọn các biến cần thống kê ở khung trái của hộp thoại Summarize Cases đưa sang khung Variables – *Hình 3.19*
- Chọn biến để phân nhóm (*Hình thức bán hàng*) đưa sang khung Grouping Variable(s) – *Hình 3.19*
- Tùy chọn trong mục Display cases:
 - + Limit cases to first : Chỉ thống kê với n trường hợp đầu tiên.
 - + Show only valid cases : Chỉ hiển thị những trường hợp có hiệu lực.
 - + Show case numbers : Hiển thị số lượng trường hợp trong mỗi nhóm – nếu có biến phân nhóm.



Hình 3.19



Hình 3.20

- Chọn nút Statistics để hiển thị hộp thoại Summary Report: Statistics → chọn các đại lượng thống kê ở khung Statistics đưa sang Cell Statistics – Hình 3.20. Chọn Continue để trở về hộp thoại Summarize Cases.

- Chọn OK và quan sát kết quả.

Case Summaries^a

			Amoxicillin 500mg	Erythromycin 500mg	Doxycyclin 500mg
Hình thức ban hàng	0	1	.	.	20
		2	.	30	.
		3	.	.	30
		4	.	10	12
		5	20	.	.
		6	20	.	.
		7	10	10	.
		8	.	20	.
		9	.	30	10
		10	5	30	30
		11	20	30	.
		12	30	30	30
		13	30	30	.
		14	10	.	.
1	1	Total	N	8	6
			Sum	145	132
			Minimum	5	10
			Maximum	30	30
		Total	1	.	20
			2	.	20
			3	30	.
			4	.	20
			5	20	.
			6	30	.
			7	30	.
			8	.	20
			9	.	10
			10	.	30
			11	.	20
			12	.	20
			13	20	.
Total	Total	Total	14	10	.
			N	4	7
			Sum	90	150
			Minimum	10	20
		Total	Maximum	30	30
			N	12	13
			Sum	235	282
		Minimum	5	10	10
		Maximum	30	30	30

Case Summaries*

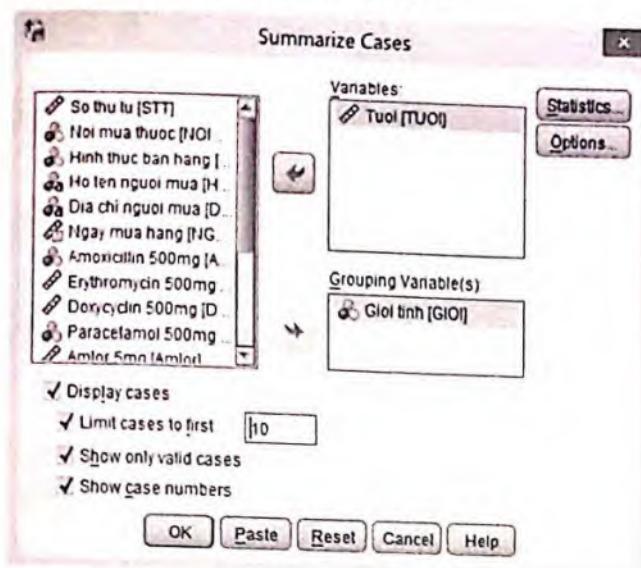
			Amoxicillin 500mg	Erythromycin 500mg	Doxycyclin 500mg
Hình thức ban hàng	0	1	.	.	20
		2	.	30	.
		3	.	.	30
		4	.	10	12
		5	20	.	.
		6	20	.	.
		7	10	10	.
		8	.	20	.
		9	.	30	10
		10	5	30	30
		11	20	30	.
		12	30	30	30
		13	30	30	.
		14	10	.	.
		Total	N	8	9
			Sum	145	220
			Minimum	5	10
			Maximum	30	30
1	1	1	.	.	20
		2	.	.	20
		3	.	30	.
		4	.	.	20
		5	20	.	.
		6	30	.	.
		7	30	.	.
		8	.	.	20
		9	.	10	.
		10	.	.	30
		11	.	.	20
		12	.	.	20
		13	.	20	.
		14	10	.	.
		Total	N	4	3
			Sum	90	60
			Minimum	10	10
			Maximum	30	30
Total	Total	N	12	12	13
		Sum	235	280	282
		Minimum	5	10	10
		Maximum	30	30	30

a. Limited to first 100 cases.

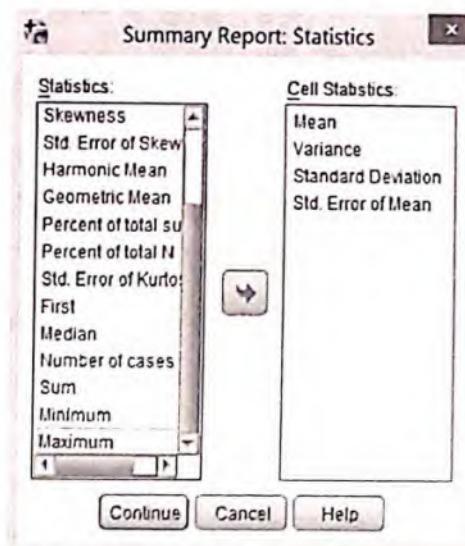
Ví dụ 2: Lập bảng tóm tắt mô tả các đại lượng thống kê mô tả gồm: trung bình, nhỏ nhất, lớn nhất, phương sai, độ lệch chuẩn, sai số chuẩn cho 10 trường hợp đầu tiên của Tuổi theo Giới tính.

- Chọn Analyze→Reports→Case Summaries.

- Chọn biến cần phân tích - Hình 3.21.
- Chọn nút **Statistics**, chọn một số đại lượng như: **Mean**, **Variance**, **Standard Deaviation**, **Std. Error of Mean** - Hình 3.22.



Hình 3.21



Hình 3.22

- Chọn **OK** và quan sát kết quả.

Mô tả tóm tắt 10 trường hợp đầu tiên của biến Tuoi^a

				Case Number	Tuoi
Gioi tinh	Nam	1		1	35
		2		6	41
		3		10	40
		Total	Mean		38.67
			Variance		10.333
Gioi tinh	Nu	Total	Std. Deviation		3.215
			Std. Error of Mean		1.856
			1	2	41
			2	3	45
		Total	3	4	37
			4	5	35
			5	7	37
Gioi tinh	Nu	Total	6	8	34
			7	9	20
			Mean		35.57
		Total	Variance		61.286
			Std. Deviation		7.829
			Std. Error of Mean		2.959

a. Limited to first 10 cases.

3.4. Vẽ biểu đồ trong SPSS

Biểu đồ (hay đồ thị) là một cách thức khác để tóm tắt và trình bày các số liệu. Các biểu đồ thường được sử dụng để thể hiện các đặc tính của bộ số liệu. Tuy biểu đồ thường rất trực quan và dễ hiểu hơn các bảng nhưng lại cung cấp cho chúng ta ít thông tin chi tiết. Cũng giống như các bảng, các biểu đồ cũng cần phải có tiêu đề, các chú giải và các đơn vị đo, ...

Trong SPSS có nhiều loại biểu đồ. Phần này giới thiệu phương pháp tạo ra các biểu đồ và ý nghĩa của từng loại biểu đồ.

Một số loại biểu đồ thông dụng:

- Biểu đồ cột – Bar
- Biểu đồ dạng đường – Line
- Biểu đồ dạng diện tích – Area
- Biểu đồ tròn – Piechart
- Biểu đồ chấm điểm - Scatter/Dot
- Biểu đồ hộp – Boxplot
- Biểu đồ cột liền – Histogram
- ...

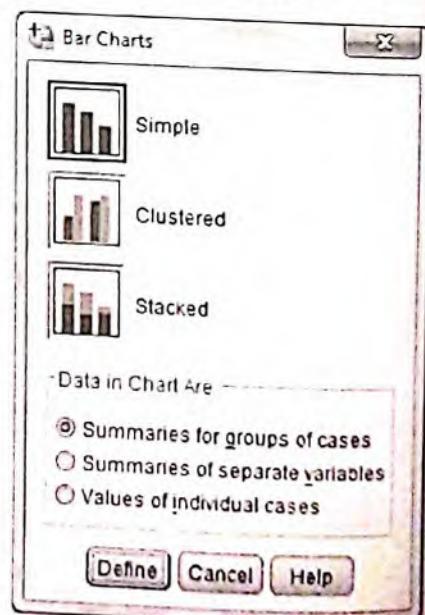
Với mỗi loại biểu đồ, người dùng có thể lựa chọn cách hiển thị dữ liệu của các biến khác nhau. Chẳng hạn, với biểu đồ cột (Bar) người dùng có thể chọn:

- **Simple:** Biểu diễn tập dữ liệu của một biến có n giá trị khác nhau thành n thanh/cột.
- **Clustered:** Biểu diễn tập dữ liệu của một biến được chia nhóm bởi dữ liệu của một biến khác, các thanh trong cùng một nhóm đứng liền kề nhau.
- **Stacked:** Biểu diễn tập dữ liệu của một biến được chia nhóm bởi dữ liệu của một biến khác, các thanh xếp chồng lên nhau.

3.4.1. Biểu đồ cột - Bar

Là kiểu biểu đồ thường được sử dụng để thể hiện phân phối tần số của các số liệu định danh và thứ bậc (biến định tính). Trong biểu đồ cột, các giá trị/nhóm giá trị của một biến số nào đó thường được thể hiện trên trực hoành. Trục tung của biểu đồ thể hiện tần số hoặc tỷ lệ % xuất hiện của các nhóm đó, tương ứng với độ cao của các cột.

- Chọn **Graphs → Legacy Dialogs → Bar**, xuất hiện hộp thoại **Bar Charts** – Hình 3.23.
- Chọn kiểu biểu đồ (*Simple, Clustered, Stacked*).
- Chọn cách thể hiện dữ liệu trong mục **Data in Chart Are:**
 - + **Summaries for groups case:** Thống kê tổng hợp cho các nhóm khác nhau. Giả sử nếu chọn *Clustered* thì mỗi thanh trong một nhóm trên đồ thị sẽ thể hiện cùng một đại lượng thống kê đã chọn.
 - + **Summaries of separate variable:** Thống kê cho từng biến khác nhau trên cùng một đồ thị.
 - + **Values of individual case:** Thể hiện giá trị thật mà không phải con số thống kê tổng hợp.
- Chọn **Define**, xuất hiện cửa sổ **Define ... Bar: Summaries for Groups of Cases** để chọn các biến cần vẽ biểu đồ, ...
- Chọn **OK** để kết thúc lệnh.



Hình 3.23

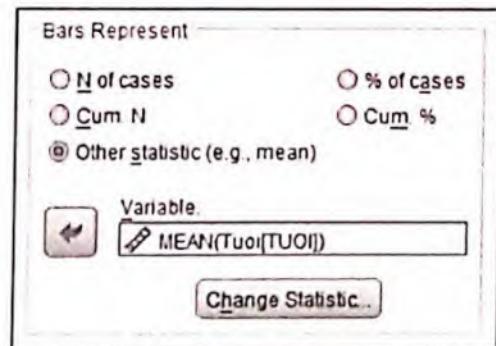
Lưu ý: Cả ba dạng của biểu đồ Bar khi vẽ cần khai báo trong mục Bar Represent – Hình 3.24.

+ **N of case (Number of Case):** Biểu diễn theo số lượng mẫu (trường hợp).

+ **% of case (Percent of case):** Biểu diễn tỷ lệ % các giá trị của biến.

+ **Cum.N (Cumulative Number – Cumulative Frequency):** Biểu diễn tần số tích lũy.

+ **Cum. % (Cumulative Percent):** Biểu diễn theo tỷ lệ cộng dồn.



Hình 3.24

(+ **Other statistics (e.g., mean):** Chọn biến định lượng để thống kê (giá trị trung bình, phương sai, độ lệch chuẩn, ...) → Chọn tham số thống kê trong mục Change Statistics.

* **Biểu diễn dữ liệu của một biến:**

Ví dụ: Biểu đồ biểu diễn Số lượng các loại bệnh.

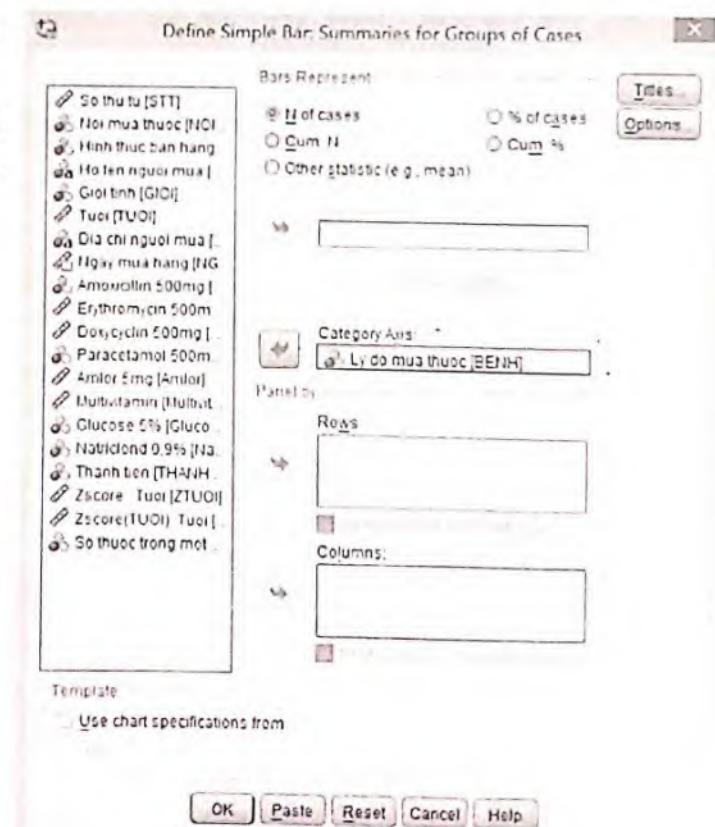
- Chọn **Graphs → Legacy Dialogs → Bar;** Chọn kiểu biểu đồ Simple.

- Trong vùng **Data in Chart Are** chọn **Summaries for groups case.**

- Chọn **Define**, thực hiện khai báo như – *Hình 3.24*:

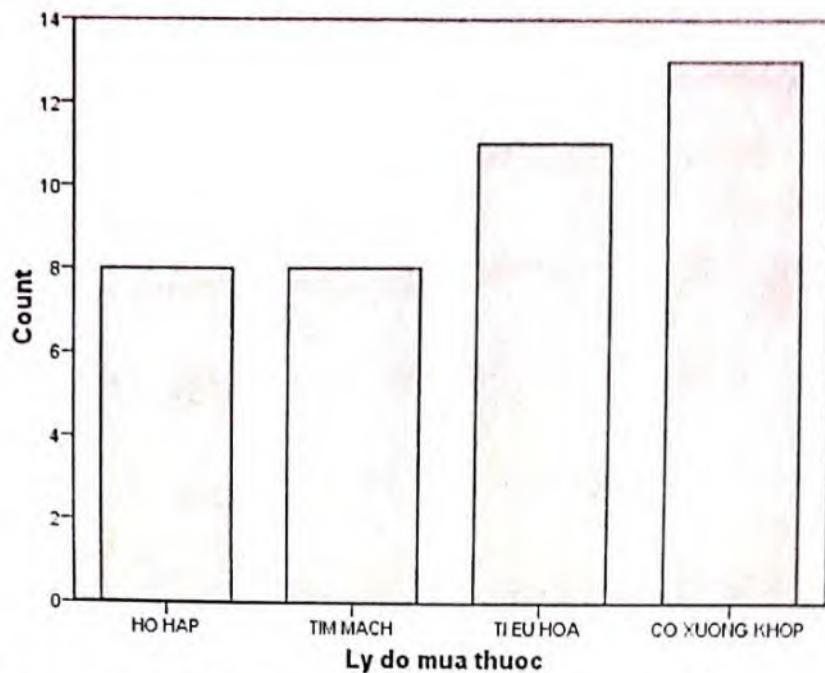
+ **Bars Represent:** Chọn cách biểu diễn số liệu trên biểu đồ.

+ Chọn biến vẽ biểu đồ đưa sang khung **Category Axis** - *Hình 3.25*



Hình 3.25

- Chọn **OK** và quan sát kết quả.

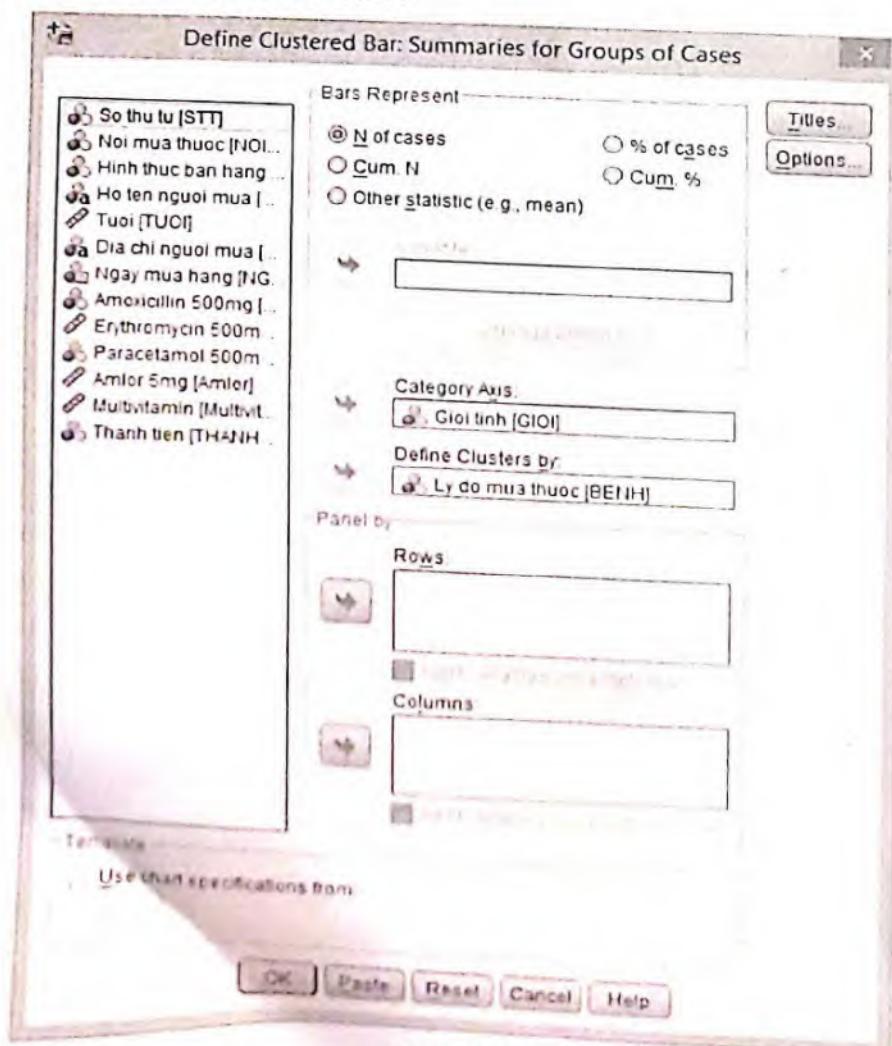


* Biểu diễn dữ liệu một biến được phân nhóm theo một biến khác

Ví dụ 1: Vẽ biểu đồ biểu diễn số lượng Bệnh theo Giới.

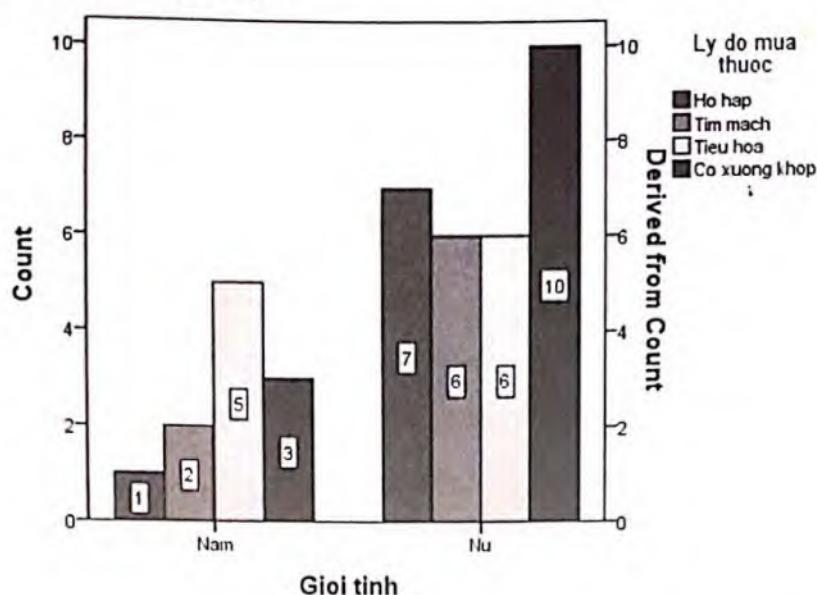
Dạng biểu đồ này thường được dùng để so sánh giữa các nhóm về số lượng, trung bình, tỷ lệ %, cực đại, cực tiểu...

- Chọn **Graphs → Legacy Dialogs → Bar;** Chọn dạng biểu đồ **Clustered - Hình 3.23**
- Trong mục **Data in Chart Are** đánh dấu **Summaries for Group of Cases - Hình 3.23**
- Chọn **Define** và khai báo - *Hình 3.27.*



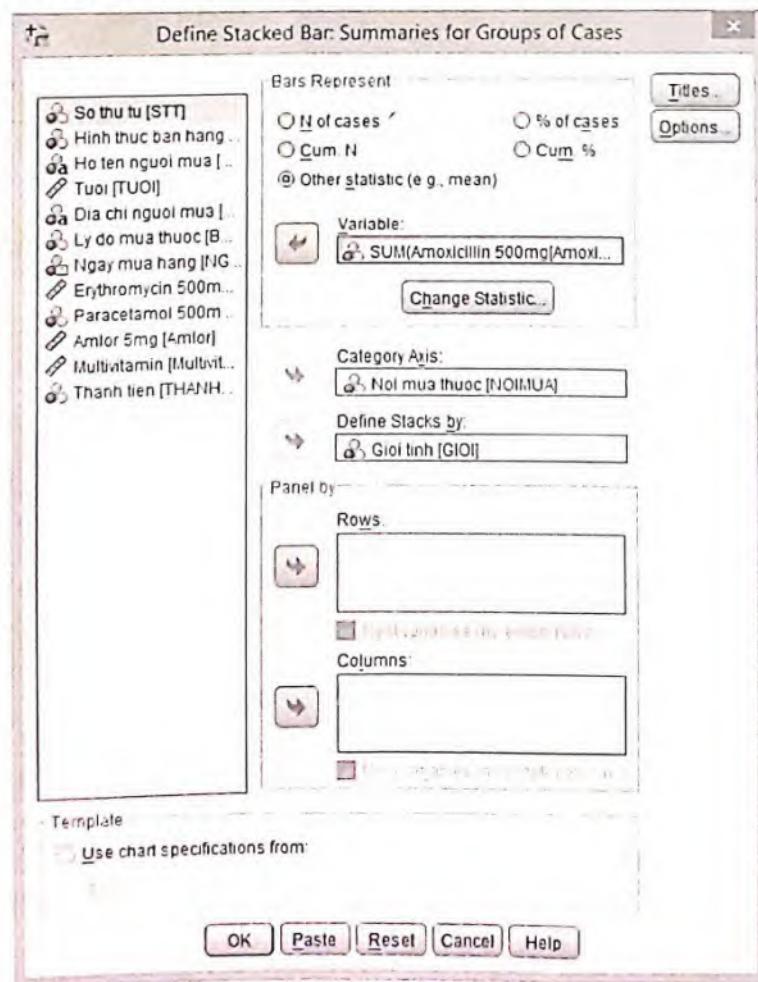
Hình 3.27

- Chọn OK và quan sát kết quả:



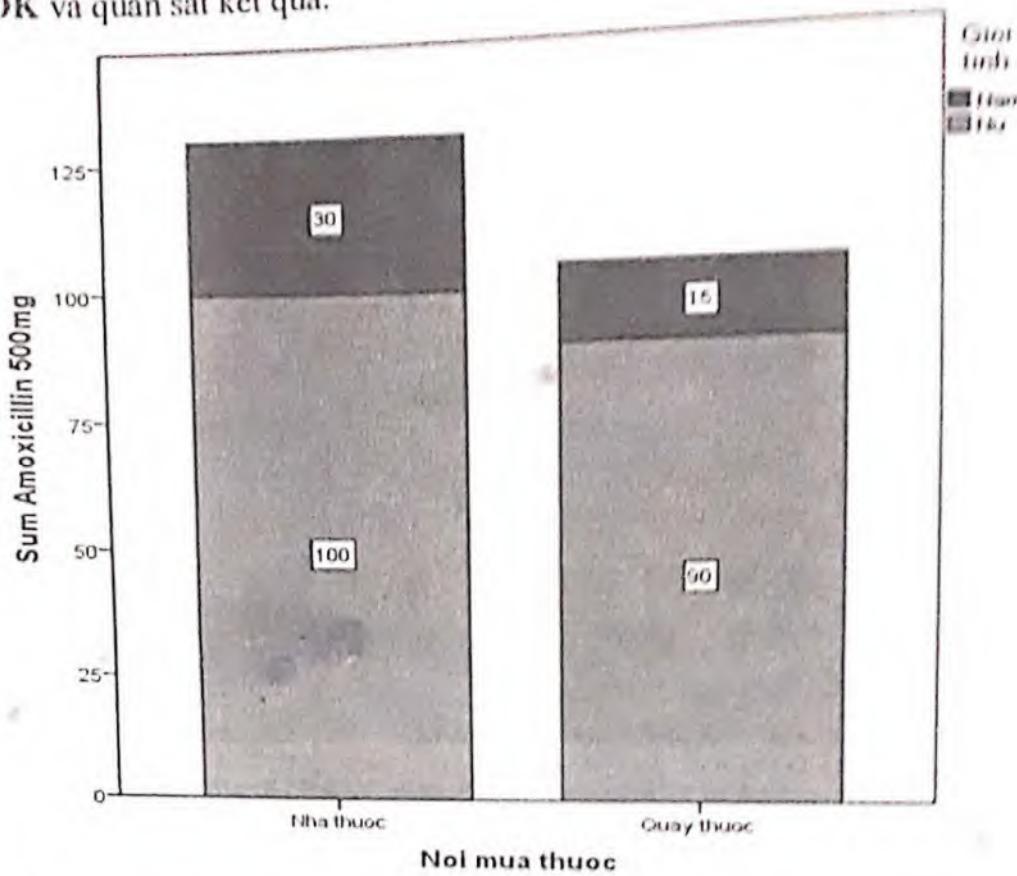
Ví dụ 2: Vẽ biểu đồ biểu diễn tổng số Amoxiellin được bán cho nam và nữ theo nơi bán thuốc.

- Chọn **Graphs → Legacy Dialogs → Bar** ; Chọn dạng biểu đồ Stacked - *Hình 3.23*
- Trong mục **Data in Chart Are** đánh dấu **Summaries for Group of Cases** - *Hình 3.23*
- Chọn **Define** và khai báo *Hình 3.28*

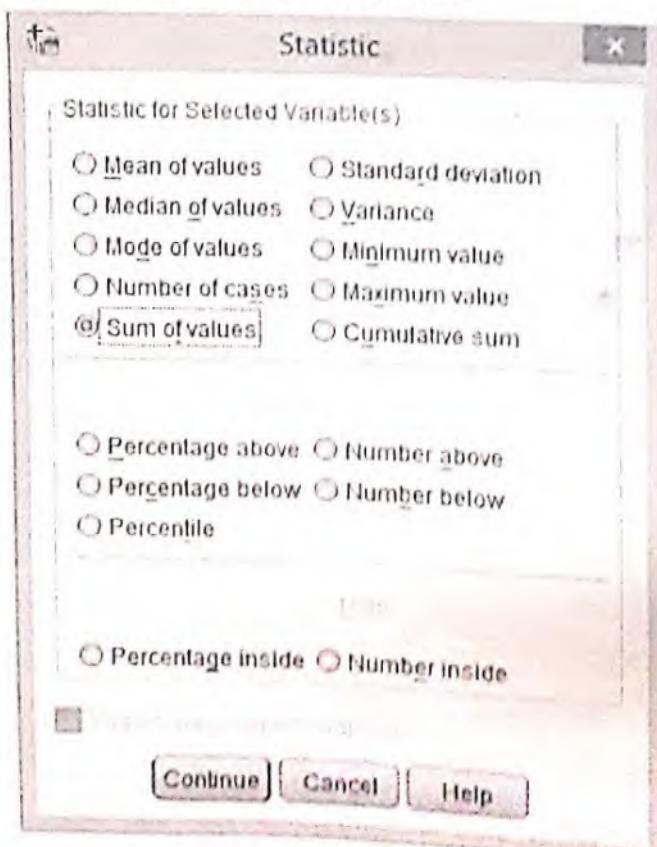


Hình 3.28

- Nhập OK và quan sát kết quả.



Lưu ý: Chọn Change Statistic: để hiển thị các giá trị thống kê theo yêu cầu (số lượng, trung bình, tỷ lệ %, cực đại, cực tiểu...) - Hình 3.29



Hình 3.29

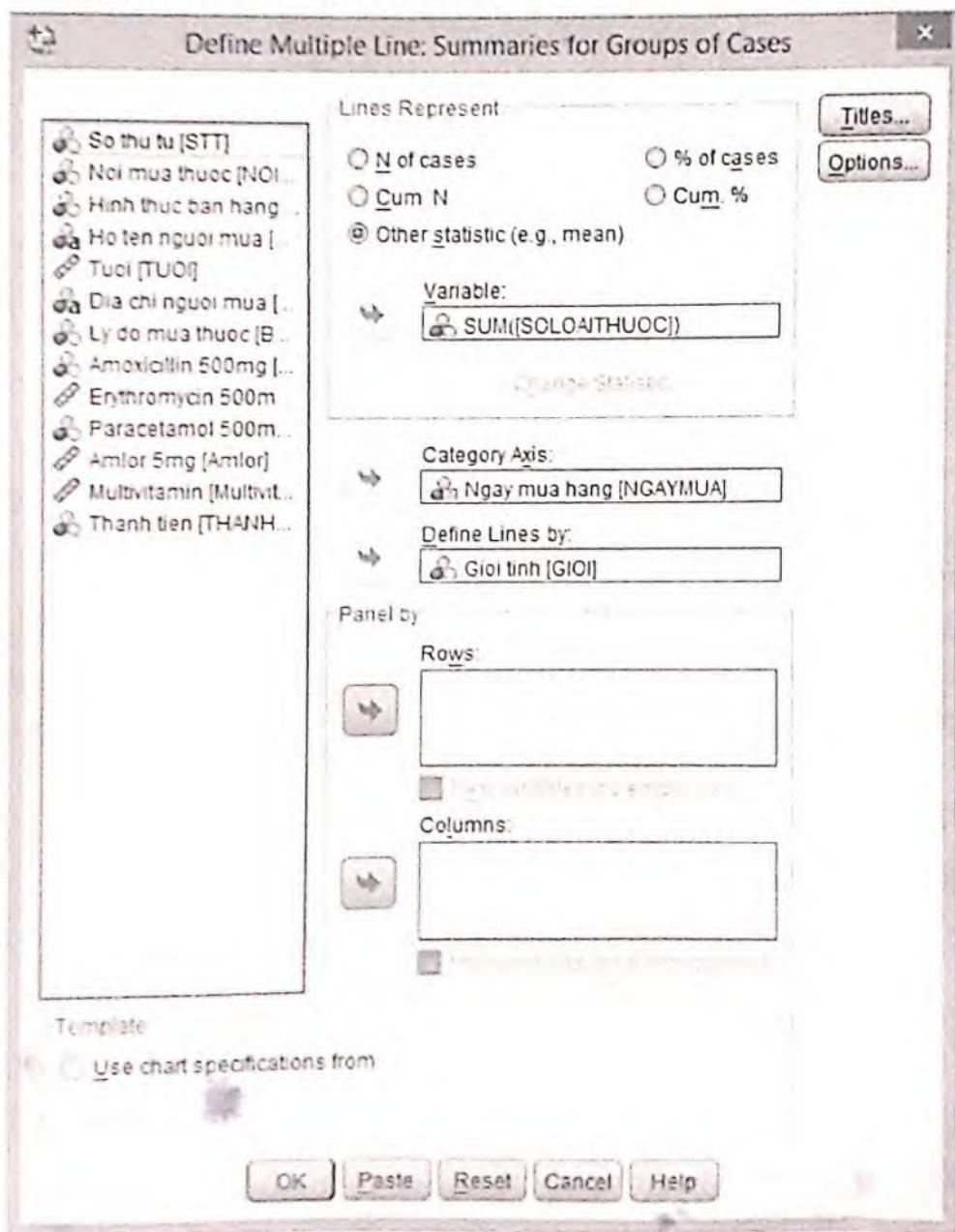
3.4.2. Đồ thị dạng đường - Line

Là kiểu biểu đồ dùng để mô tả mối quan hệ giữa hai biến liên tục hoặc sự thay đổi theo thời gian, mỗi một trục của biểu đồ sẽ biểu hiện một biến. Nói cách khác, biểu đồ **Line** dùng để biểu diễn cho các biến định lượng. Trên cùng một biểu đồ có thể vẽ nhiều đường cùng lúc.

Ví dụ: Biểu diễn tổng số loại thuốc mua của nam, nữ theo thời gian.

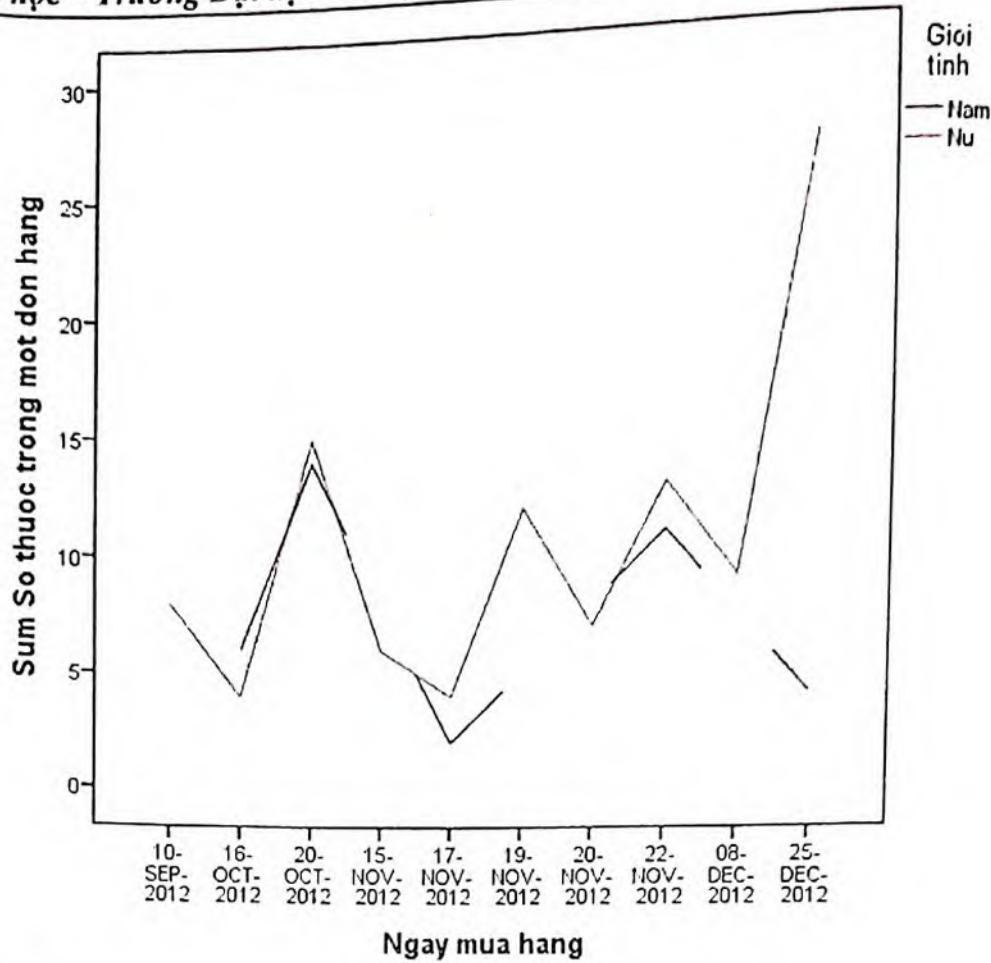
- Chọn **Graphs → Legacy Dialogs → Line**

- Chọn **Multiple**, khai báo như - *Hình 3.30*.



Hình 3.30

- Nhập **OK** và quan sát biểu đồ.

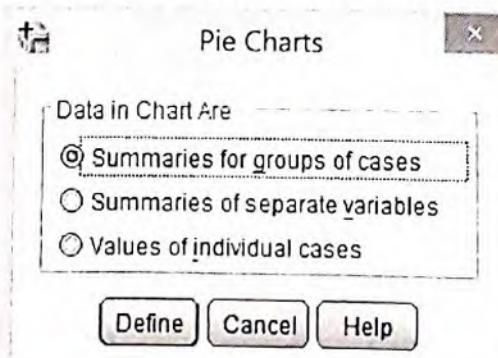


3.4.3. Biểu đồ tròn – Pie

Biểu đồ tròn hay biểu đồ bánh, biểu đồ quạt, ... Là kiểu biểu đồ dùng để biểu diễn sự phân chia toàn thể các quan sát (giá trị) của một biến số nào đó ra thành từng nhóm khác nhau, mỗi nhóm là một phần của biểu đồ. Biểu đồ này thường dùng để biểu diễn tỷ lệ % các giá trị khác nhau của một biến số. Thường biểu diễn dữ liệu cho biến định tính.

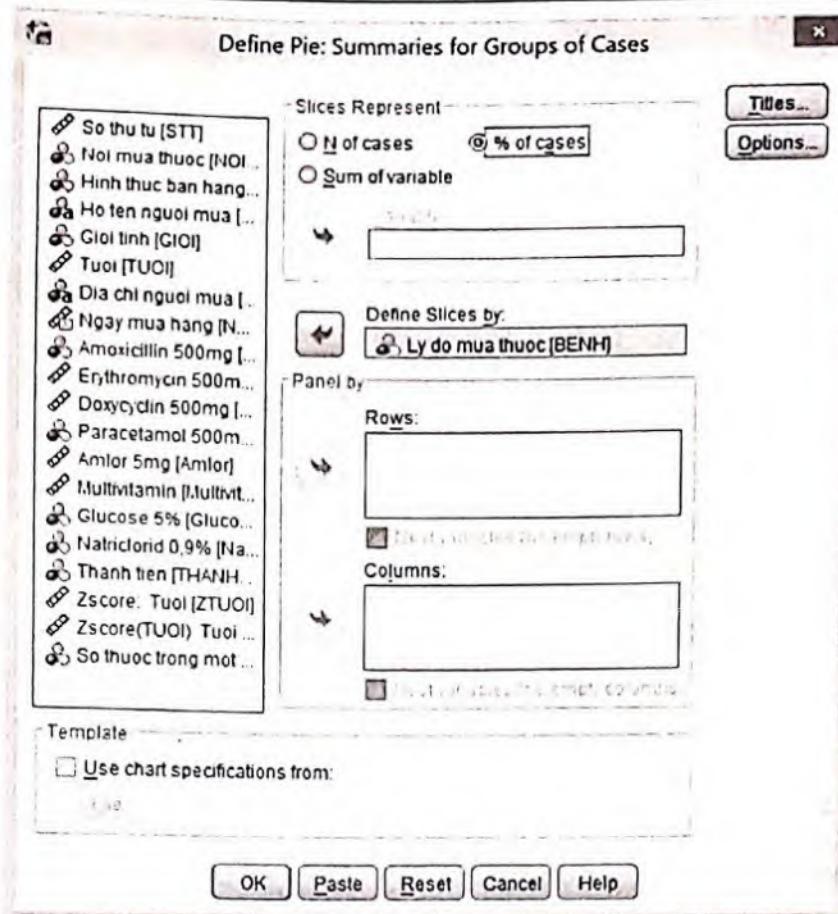
Ví dụ: Vẽ biểu đồ biểu diễn tỷ lệ % các Bệnh.

- Chọn **Graph → Legacy Dialogs → Pie**, xuất hiện hộp thoại - *Hình 3.31*.



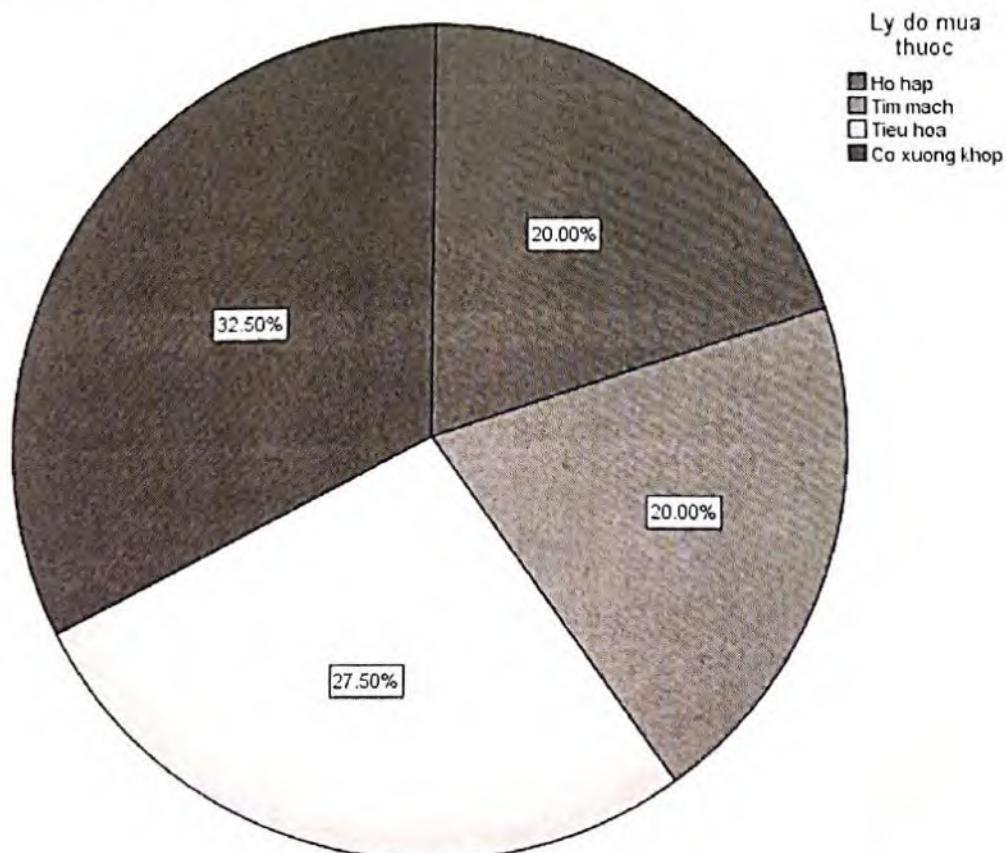
Hình 3.31

- Chọn **Summaries for groups of case** → chọn **Define** - *Hình 3.31*
- Thực hiện các khai báo như *Hình 3.32*



Hình 3.32

- Chọn **OK** và quan sát kết quả.



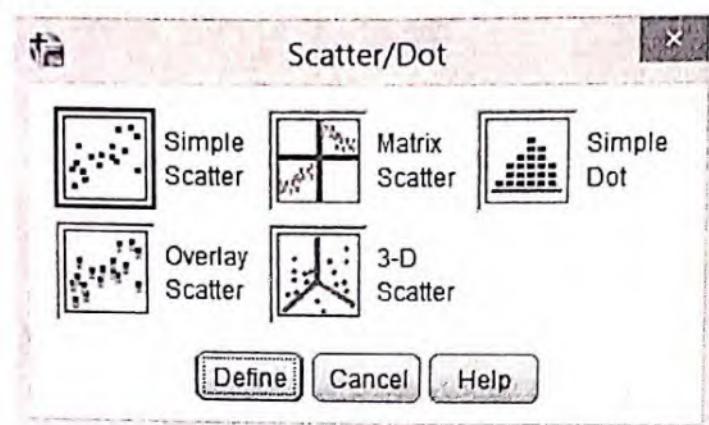
3.4.4. Biểu đồ chấm điểm –Scatter/Dot

Dùng để biểu diễn mối tương quan giữa hai biến định lượng. Nhìn vào biểu đồ có thể thấy được chiều hướng của mối tương quan.

Ví dụ: Vẽ biểu đồ biểu diễn mối tương quan giữa mache đậm và tuổi.

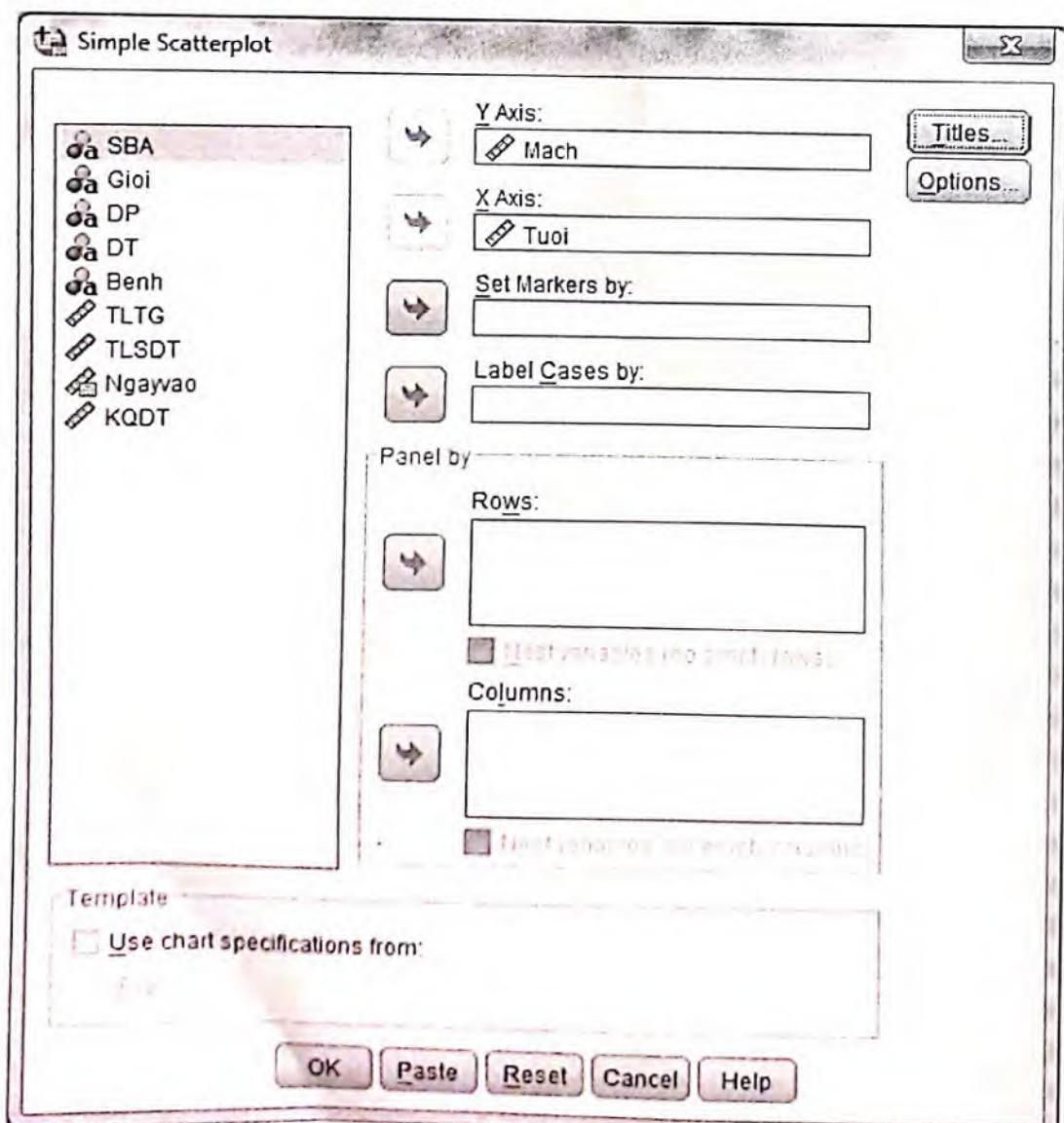
Các bước thực hiện:

- Chọn **Graph → Legacy Dialogs → Scatter/Dot**, xuất hiện hộp thoại - *Hình 3.33*, chọn kiểu biểu đồ **Simple Scatter**.



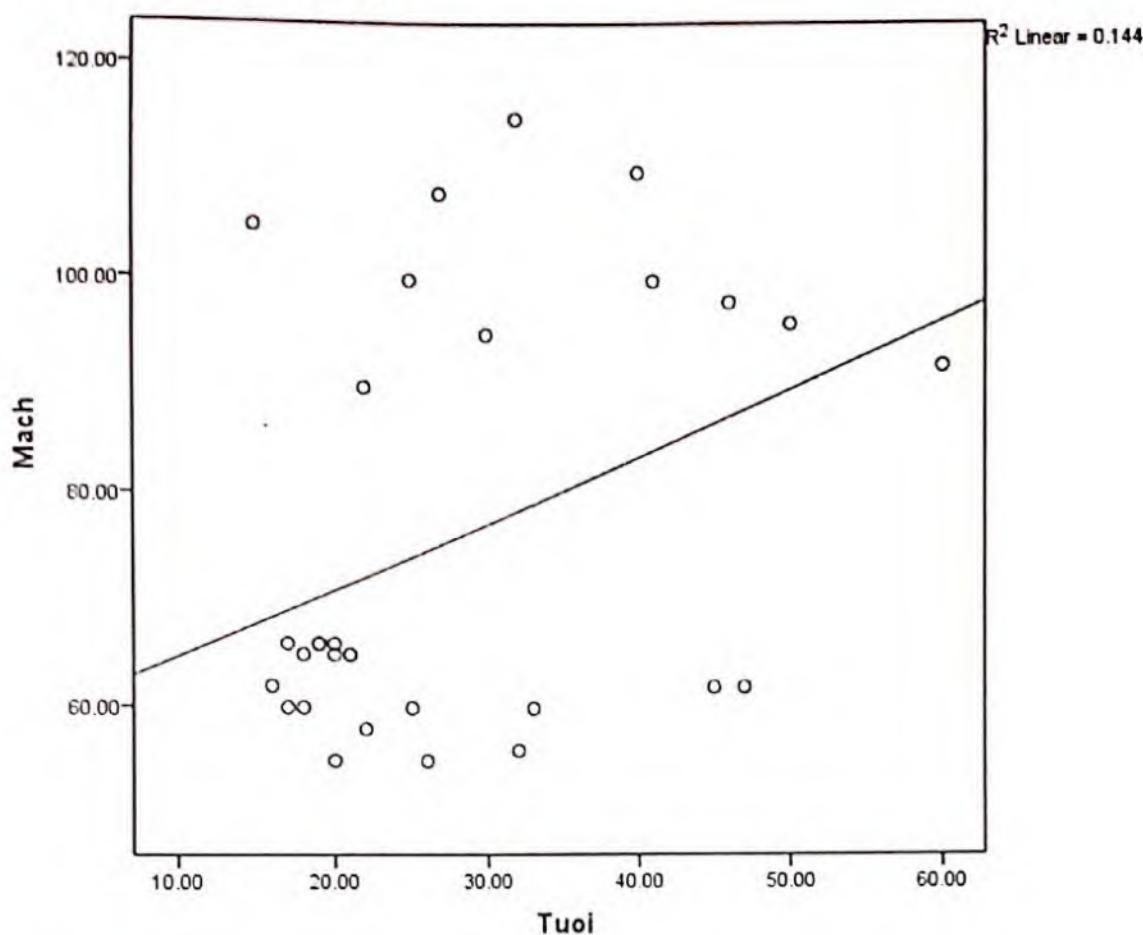
Hình 3.33.

- Thực hiện các khai báo như *Hình 3.34*.



Hình 3.34.

- Chọn OK và quan sát kết quả.



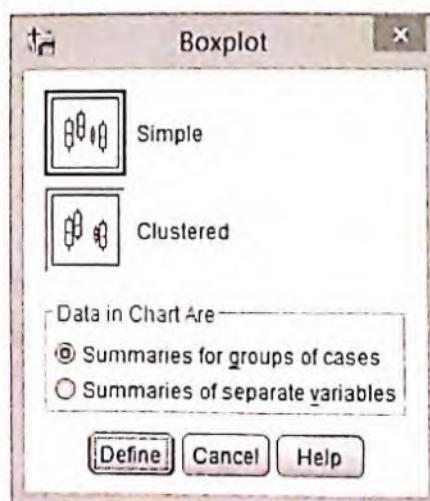
3.4.5. Biểu đồ hộp – Boxplot

Là biểu đồ có dạng hộp và 2 râu, phản ánh tính đối xứng và sự phân tán của tập dữ liệu; Áp dụng cho biến định lượng. Biểu đồ này phù hợp với cả phân bố chuẩn và không chuẩn. Đoạn thẳng trong hộp cho biết giá trị trung vị của tập dữ liệu, hai cạnh còn lại (song song với nó) cho biết giá trị tứ phân vị thứ nhất và thứ 3. Hai râu nối tới giá trị nhỏ nhất và lớn nhất.

Ví dụ: Vẽ biểu đồ biểu diễn Tuổi theo Giới tính.

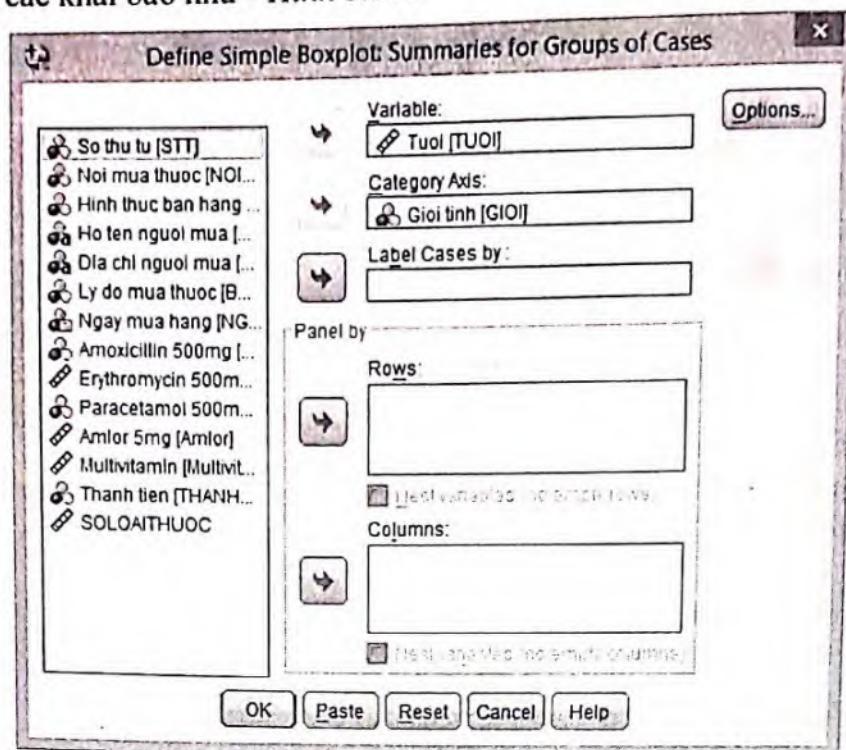
Các bước thực hiện:

- Chọn **Graph → Legacy Dialogs → Boxplot**, xuất hiện hộp thoại - *Hình 3.35*, chọn kiểu đồ loại biểu đồ **Simple** → Chọn **Define**.



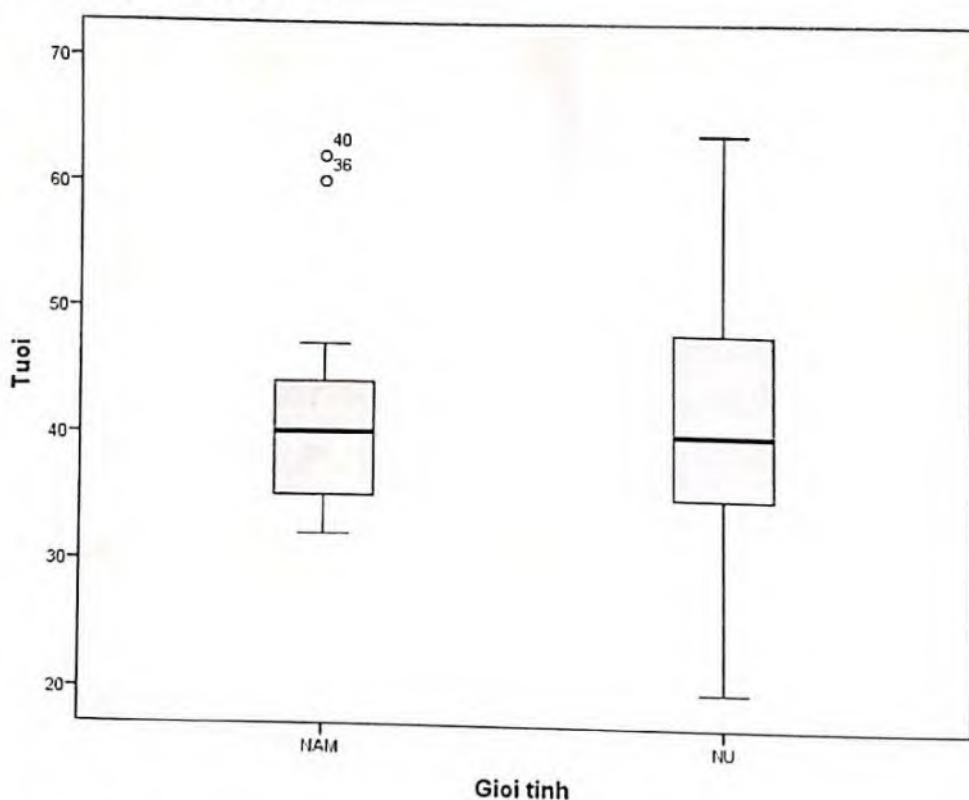
Hình 3.35.

- Thực hiện các khai báo như - Hình 3.36.



Hình 3.36.

- Nhấp **OK** và quan sát kết quả.

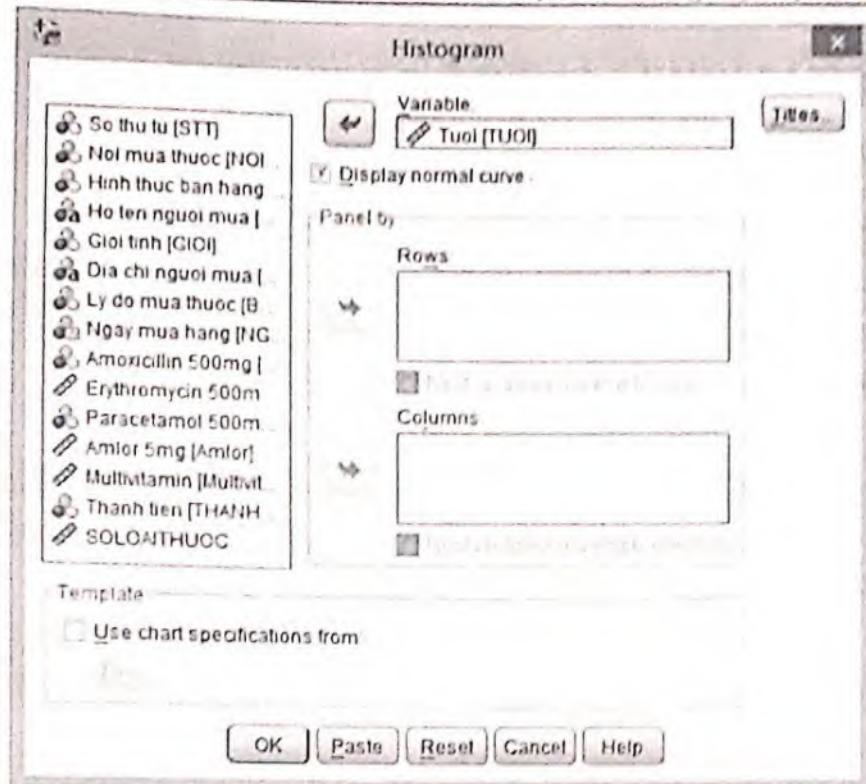


3.4.6. Biểu đồ tần suất – Histogram

Hay biểu đồ liên tục: Là kiểu biểu đồ biểu diễn dữ liệu cho các biến định lượng liên tục nhằm biểu diễn phân phối các giá trị của tập số liệu. Trục hoành của biểu đồ thể hiện giới hạn thực của các khoảng số liệu khác nhau, trực tung thể hiện tần số hoặc tần số tương đối của các giá trị quan sát trong các khoảng khác nhau.

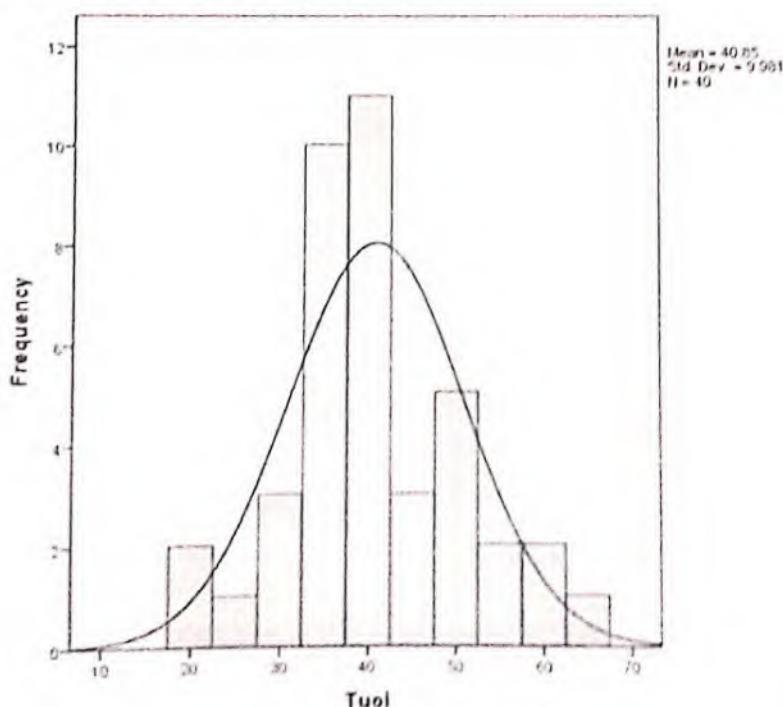
Ví dụ: Vẽ biểu đồ biểu diễn dữ liệu của Tuổi. Các bước thực hiện:

- Chọn **Graph → Legacy Dialogs → Histogram**. Thực hiện các khai báo như - Hình 3.37



Hình 3.37

- Chọn **OK** và quan sát kết quả.



☞ **Lưu ý:** Chọn ô **Display normal curve** để hiển thị đường cong phân bố.

Bài 4. SO SÁNH TỶ LỆ VÀ KIỂM ĐỊNH TÍNH ĐỘC LẬP

Mục tiêu:

- Trình bày được cơ sở lý thuyết của kiểm định khi bình thường.
- Trình bày các bước tiến hành khi so sánh các tỷ lệ và kiểm định tính độc lập.
- Đọc và phiên giải được ý nghĩa các tham số thống kê trong các bài toán so sánh các tỷ lệ và kiểm định tính độc lập.

Dữ liệu định tính có hai dạng là *Định danh (Nominal)* và *Thứ bậc (Ordinal)*. Trong các bài trước ta đã dùng các bảng tần số, biểu đồ cột tần suất và tỷ lệ để mô tả phân bố các giá trị của một biến định tính.

Bài học này sẽ nghiên cứu về: so sánh các tỷ lệ, có hay không có mối liên hệ giữa hai biến (*thường sử dụng kiểm định Chi-Square*), nguy cơ tương đối và tỉ suất chênh của dữ liệu định tính trong bộ số liệu nghiên cứu.

Trong bài này ta sử dụng bộ số liệu về khuẩn Ecoli (Ecoli.sav) với cỡ mẫu đủ lớn (359 bài ghi). Bảng dưới đây cho biết cấu trúc và ý nghĩa các cột.

	Name	Type	Width	Decimals	Label	Values
1	CaseID	Numeric	4	0	Chỉ số ca phỏng vấn	None
2	DateofInterview	Date	11	0	Ngày phỏng vấn	None
3	Sex	String	8	0	Giới tính	{F-Female, N Nữ}...
4	DOB	Date	11	0	Ngày sinh	None
5	Age	Numeric	3	0	Tuổi	None
6	State	String	2	0	Nơi ở	None
7	Occupation	String	21	0	Nghề nghiệp	None
8	ILL	Numeric	1	0	Nhiễm khuẩn	{0, Không nhiễm khuẩn...}
9	Antibiotics	Numeric	1	0	Kháng sinh	{0, Không dùng kháng...}
10	Hospitalized	Numeric	1	0	Nhập viện	{0, Không nhập viện}...
11	FeverTemp	Numeric	5	1	Nhiệt độ sốt	None
12	Fever	String	5	0	Sốt	{False, Không}...
13	Headache	String	5	0	Đau đầu	{False, Không}...
14	Vomiting	String	5	0	Nôn mửa	{False, Không}...
15	PoorFeeding	String	5	0	Kém ăn	{False, Không}...
16	Chills	String	5	0	Ớn lạnh	{False, Không}...
17	Nausea	String	5	0	Buồn nôn	{False, Không}...

Trong bộ số liệu về Ecoli, biến **Nhomtuoi** (*Nhóm tuổi*), **Mucdosot** (*Mức độ sốt*) là dữ liệu kiểu số được tạo theo các quy tắc sau:

Tuổi	Nhóm tuổi
8 – 35	1
36 – 60	2
61 – 91	3

Nhiệt độ sốt	Mức độ sốt
≤ 103.1	1
> 103.2	2

4.1. Cơ sở lý thuyết của Kiểm định Khi bình phương

1) Xây dựng giả thuyết H_0 và đối thuyết H_1

2) Tính đại lượng Khi bình phương quan sát

- Lập bảng với các tần số quan sát là O_{ij} (i : chỉ số hàng, j chỉ số cột).

$$\text{- Tính Chi-Square quan sát } \chi^2_{QS} = \sum_{i=1, j=1}^{n, m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Trong đó:

+ χ^2_{QS} : đại lượng Chi-Square quan sát dùng để kiểm định.

+ O_{ij} : là tần số quan sát (*Observed*) - Tần số quan sát thực tế trong bảng n hàng, m cột tương ứng với các mức khác nhau của biến 1 và các mức khác nhau của biến 2 (*bảng n × m ô vuông*).

+ E_{ij} : là tần số mong đợi (*Expected*) - Tần số quan sát lý thuyết có được trên cơ sở công nhận giả thiết H_0 là đúng. $E_{ij} = (\text{Tổng hàng} \times \text{Tổng cột}) / \text{Tổng chung}$.

3) Xác định Khi bình phương lý thuyết

- Bậc tự do $df = (\text{số hàng} - 1) \times (\text{số cột} - 1)$

- Với mức độ tin cậy thường là 90%, 95% hoặc 99% tương ứng với mức ý nghĩa cho trước α là 0.1, 0.05 hoặc 0.01

- $\chi^2_{LT} = \chi^2_{(n-1) \times (m-1)}$ với bậc tự do df tương ứng với mức ý nghĩa cho trước α được tra trong Bảng giá trị χ^2_{LT} .

4) So sánh χ^2_{QS} với χ^2_{LT} để kết luận

- Nếu $\chi^2_{QS} \leq \chi^2_{LT}$: chấp nhận H_0 . $\Leftrightarrow sig. \geq 0.05$

- Nếu $\chi^2_{QS} > \chi^2_{LT}$: bác bỏ H_0 . $\Leftrightarrow sig. < 0.05$

Lưu ý:

1) Test Khi bình phương chỉ áp dụng khi:

- Cỡ mẫu đủ lớn (với bảng 2×2 cỡ mẫu phải là 40).

- Các giá trị mong đợi (*Expected*) không được nhỏ hơn 5.

- Với các bảng lớn hơn 2×2 cho phép dưới 20% số ô có giá trị < 5 nhưng không được bằng 0. Thường có thông báo cho biết có bao nhiêu ô có giá trị < 5 ở dưới bảng Chi-Square Tests.

2) Trường hợp không đủ điều kiện áp dụng Test Khi bình phương có thể xử lý như sau:

- Sử dụng test chính xác của Fisher

- Với bảng có trên 2 dòng (hoặc 2 cột) có thể gộp chung lại với nhau để tăng giá trị mong đợi.

4.2. So sánh tỷ lệ

Việc so sánh các tỷ lệ không yêu cầu dữ liệu đầu vào phải tuân theo luật phân phối chuẩn; tuy nhiên hệ thống yêu cầu là mẫu phải được chọn một cách ngẫu nhiên. Trong SPSS người ta đưa dữ liệu cần so sánh tỷ lệ vào một biến nhị phân (*chi nhận 2 giá trị*) hoặc 1 biến có thứ bậc nào đó. Trường hợp biến có nhiều giá trị thì khi so sánh tỷ lệ SPSS sẽ yêu cầu đưa vào 1 giá trị để phân biến này thành hai nhóm rồi so sánh tỷ lệ giữa hai nhóm với nhau.

Sử dụng phương pháp χ^2 khi muốn so sánh các tỷ lệ trong một nghiên cứu; cụ thể là xét xem có hay không có sự khác nhau giữa các tỷ lệ khi có tác động của một nhân tố đến một đặc tính nào đó.

4.3.1. So sánh các tỷ lệ

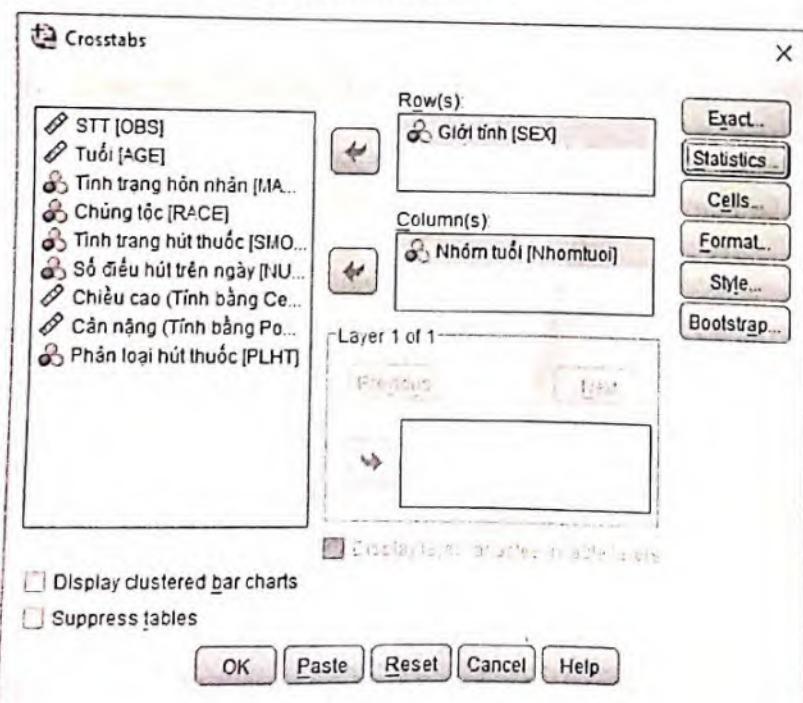
Khi ta cần so sánh tỷ lệ Nam-Nữ ở các nhóm tuổi xem có sự khác nhau hay không với độ tin cậy 95%. Áp dụng kiểm định Khi bình phương như sau:

1) Xây dựng giả thuyết và đối thuyết

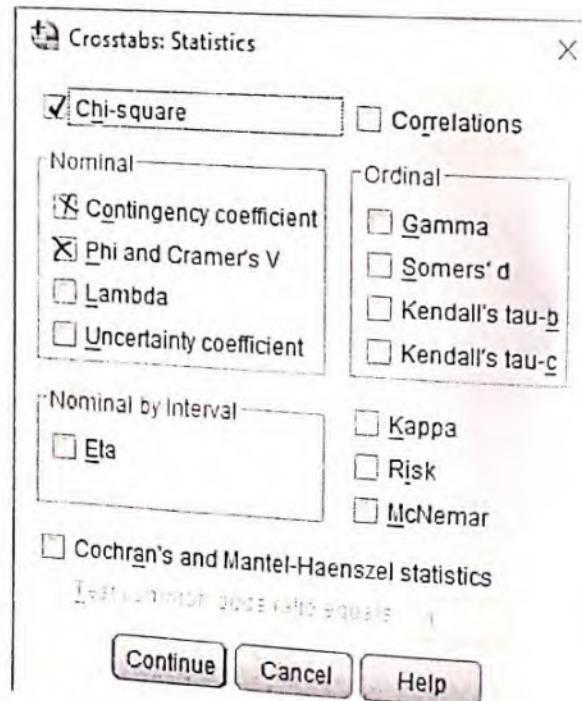
- Giả thuyết H_0 : Tỷ lệ Nam-Nữ giữa các nhóm tuổi là như nhau
- Đối thuyết H_1 : Tỷ lệ Nam-Nữ giữa các nhóm tuổi là khác nhau

2) Tính giá trị χ^2_{QS} (bằng lệnh Crosstabs)

- Chọn Analyze → Descriptive Statistics → Crosstabs... xuất hiện hộp thoại Crosstabs, rồi đưa vào các biến như *Hình 4.1*.



Hình 4.1.



Hình 4.2.

- Chọn nút Statistics... và đánh dấu mục Chi-square như *Hình 4.2*, nhấn Continue trở về cửa sổ Crosstabs.
- Chọn nút Cell và đánh dấu mục Expected, Chọn Continue trở về cửa sổ Crosstabs.
- Nhấn OK và quan sát các kết quả từ các bảng sau:

Giới tính * Nhóm tuổi Crosstabulation

Giới tính	Nhóm tuổi	Nhóm tuổi			Total
		1	2	3	
Nữ	Count	102	71	13	186
	Expected Count	96.9	75.1	14.0	186.0
Nam	Count	85	74	14	173
	Expected Count	90.1	69.9	13.0	173.0
Total	Count	187	145	27	359
	Expected Count	187.0	145.0	27.0	359.0

1 → đồng

2 + 2 → đồng

Chi-Square Tests .

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.175 ^a	2	.556
Likelihood Ratio	1.176	2	.555
N of Valid Cases	359		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 13.01.

Lưu ý: Kết quả kiểm định được đọc trong dòng đầu tiên của bảng Chi-Square Test với bảng $n \times m$ và ở dòng thứ 2 với bảng 2×2 .

Từ các bảng kết quả trên ta thấy bảng thống kê có dạng $2 \times 3 \rightarrow$ Đọc kết quả dòng 1: $\chi^2_{QS} = 1.175$, mức ý nghĩa (Asymp. Sig. (2-sided)) = $0.556 > 0.05$.

3) Xác định giá trị của χ^2_{LT}

Với bậc tự do $df = 2$, độ tin cậy 95% ($\alpha = 0.05$) $\rightarrow \chi^2_{LT} = 5.99$ (tra trong Bảng giá trị χ^2_{LT})

4) Kết luận

Ta thấy $\chi^2_{QS} = 1.175 < \chi^2_{LT} = 5.99 \rightarrow$ chấp nhận H_0 , nghĩa là: các tỷ lệ Nam – Nữ ở các Nhóm tuổi là chưa có sự khác biệt (tỷ lệ Nam-Nữ ở các nhóm tuổi là như nhau); điều này hoàn toàn phù hợp với Asymp.Sig.(2-sided) = $0.556 > 0.05$.

* Lưu ý:

Trong bảng kết quả còn có một số tham số khác nữa (nếu muốn tính các tham số này này thì đánh dấu vào mục Phi and Cramer V và Contingency Coefficient)

- Test Fisher hay kiểm định Chi-Square được dùng phổ biến nhất nhưng không cho biết độ mạnh của mối liên hệ giữa hai biến định danh, cần dựa thêm vào một số test khác nếu muốn biết điều này, cụ thể:

- Test Cramer V: Được tính dựa trên Chi-Square theo công thức:

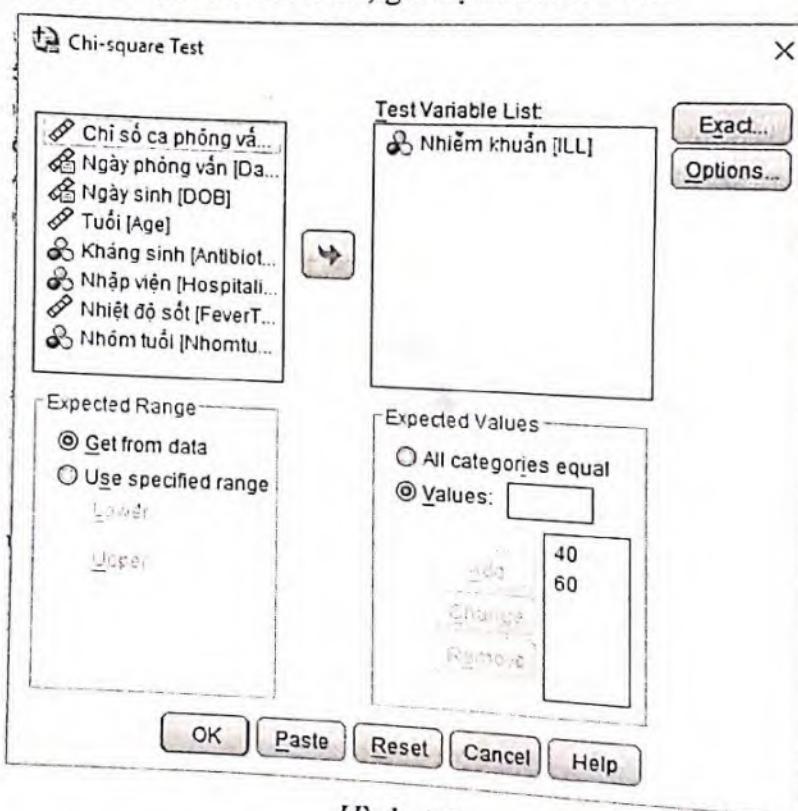
$$V = \sqrt{\frac{\chi^2}{N(k-1)}} \quad 0 \leq V \leq 1. \quad \text{Ở đây, } N \text{ là tổng số quan sát (cỡ mẫu)} \text{ còn } k \text{ là số bé nhất trong số hàng và số cột. Nếu } V \text{ càng gần } 1 \text{ thì mối liên hệ là càng mạnh.}$$

- **Hệ số liên hợp (Contingency Coefficient):** là tham số đánh giá mức độ tương quan giữa hai biến. $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$ $0 \leq C \leq 1$. Khi $C = 0$ hai biến không có quan hệ, C càng lớn mức độ quan hệ sẽ càng mạnh.

4.3.2. So sánh tỷ lệ một mẫu với một tỷ lệ lý thuyết

Khi muốn so sánh tỷ lệ không nhiễm khuẩn, nhiễm khuẩn trong nghiên cứu này với một tỷ lệ cho trước là 40%, 60%. Ta sử dụng **Chi-Square Test**.

- Giả thuyết H_0 : Tỷ lệ không nhiễm khuẩn, nhiễm khuẩn là 40%, 60%.
- Đối thuyết H_1 : Tỷ lệ không nhiễm khuẩn, nhiễm khuẩn khác 40%, 60%.
- Chọn Analyze → Nonparametric Tests → Legacy Dialogs → Chi-square... xuất hiện hộp thoại Chi-Square Test, rồi đưa vào các biến, giá trị như **Hình 4.3**.



Hình 4.3

- Nhấn **OK** và quan sát các kết quả từ các bảng sau:

Nhiễm khuẩn			
	Observed N	Expected N	Residual
Không nhiễm khuẩn	83	143.6	-60.6
Có nhiễm khuẩn	276	215.4	60.6
Total	359		

Test Statistics	
Chi-Square	Nhiễm khuẩn
Df	42.623 ^a
Asymp. Sig.	1
	.000

a. 0 cells (0.0%) have expected frequencies less than 5.
The minimum expected cell frequency is 143.6.

- Từ bảng kết quả trên cho thấy tần số quan sát của những người không nhiễm khuẩn, nhiễm khuẩn là 83 và 276 trong khi tần số mong đợi là 143.6 và 215.4 theo tỉ lệ 40%, 60%.
- Giá trị Khi bình phương = 42.623 với mức ý nghĩa Asymp. Sig. = 0.000 (<0.05). Như vậy ta có cơ sở bác bỏ giả thiết H_0 . $\chi^2_{QS} > \chi^2_{L1} = 5,84$

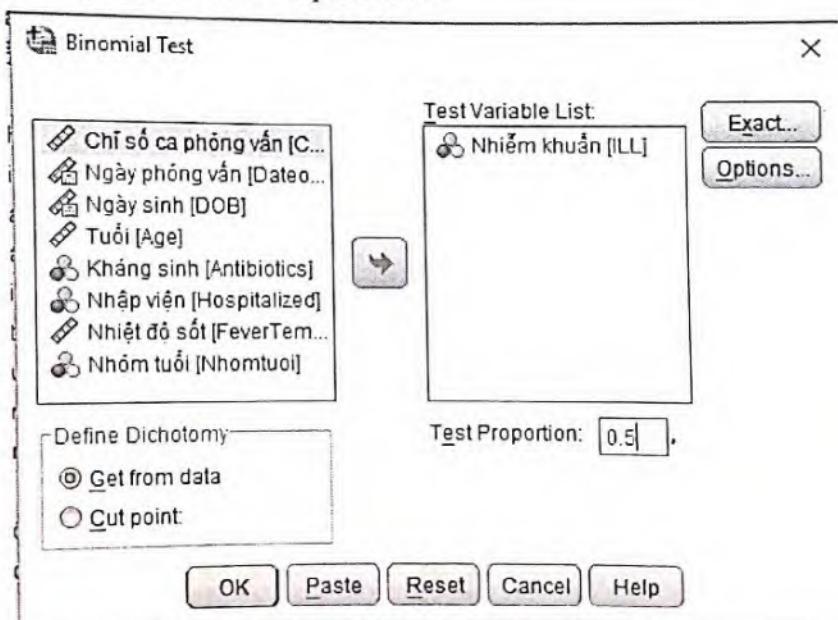
Kết luận: Tỷ lệ những người không nhiễm khuẩn, nhiễm khuẩn là khác 40%, 60%.

Lưu ý:

Hypo Value dùng để thay đổi tỷ lệ cân so sánh, nhưng phải đảm bảo tổng là 100% cho các hiện tượng xảy ra trong biến. Tuy nhiên nếu có 2 hiện tượng xảy ra trong một biến người ta thường hay so sánh tỉ lệ 2 hiện tượng có nhau không tức là so sánh với 50%.

Ngoài cách sử dụng Chi-Square Test ta có thể sử dụng Biomial Test như sau:

- Chọn Analyze → Nonparameteric Test → Legacy Dialogs → Biomail..., xuất hiện hộp thoại Hình 4.4. Trong cửa sổ này mục Test Variable List đưa vào biến Nhiễm khuẩn. Ở chế độ mặc định Test Proportion do hệ thống tự đặt là 0.50. Có nghĩa chúng ta so sánh tỷ lệ những người bị nhiễm khuẩn và không nhiễm khuẩn với 0.5(50%). Ta có thể thay đổi tỷ lệ kiểm định này bằng cách nhập trực tiếp vào ô Test Proportion.



Hình 4.4.

- Nhấp OK và quan sát kết quả:

Binomial Test

	Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Nhiễm khuẩn	Group 1 Có nhiễm khuẩn	276	.77	.50	.000
	Group 2 Không nhiễm khuẩn	83	.23		
	Total	359	1.00		

Từ bảng kết quả trên cho thấy tỉ lệ quan sát của người bị nhiễm khuẩn là $276/359=0.77$, người không bị nhiễm khuẩn là $83/359=0.23$.

- Tỷ lệ người bị nhiễm khuẩn, không bị nhiễm khuẩn so với 50% có Sig = 0.000 (<0.05)
- bác bỏ giả thiết H_0

Kết luận: Tỷ lệ người bị nhiễm khuẩn, không bị nhiễm khuẩn trong mẫu là khác nhau.

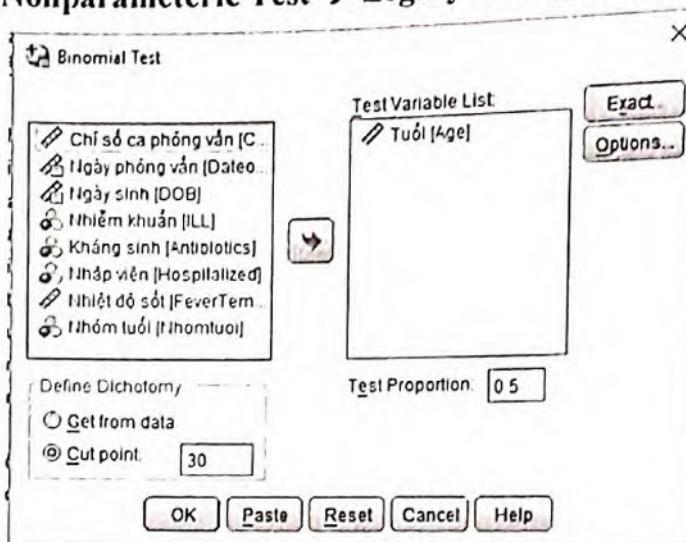
Lưu ý:

Nếu muốn phân biến thành hai nhóm bằng một *điểm cắt* cần đánh dấu mục *Cut point*. Để chia biến Tuổi thành hai nhóm: một nhóm có Tuổi > 30, và một nhóm có Tuổi ≤ 30 ta khai báo *Cut Point* là 30 (xem Hình 4.5).

Ví dụ khi cần so sánh tỷ lệ những người có Tuổi > 30 và Tuổi ≤ 30 với tỷ lệ 50%.

- Giả thuyết H_0 : Tỷ lệ những người có Tuổi > 30 và Tuổi ≤ 30 nghiên cứu này là 50%
- Đối thuyết H_1 : Tỷ lệ những người có Tuổi > 30 và Tuổi ≤ 30 trong nghiên cứu này khác 50%

- Chọn Analyze → Nonparametric Test → Legacy Dialogs → Biomial



Hình 4.5.

- Khai điểm *Cut point* = 30 → Nhập OK và quan sát kết quả:

Binomial Test						
	Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)	
Tuổi	Group 1 <= 30	149	.42	.50	.002	
	Group 2 > 30	210	.58			
	Total	359	1.00			

- Kết quả *Exact Sig. (2-tailed)* = 0.002 (< 0.05) → bác bỏ giả thiết H_0

Kết luận: Tỷ lệ tuổi của hai nhóm dưới 30 và trên 30 là khác nhau có ý nghĩa thống kê.

4.3. Kiểm định tính độc lập

Phương pháp Khi bình phương dùng để kiểm định xem có mối liên hệ giữa hai yếu tố (biến định tính) đang nghiên cứu trong một tổng thể hay không? Ví dụ khi cần nghiên cứu mối liên hệ giữa:

- **Giới tính** (Nam, Nữ) và **Tình trạng nhiễm khuẩn** (Có nhiễm khuẩn, Không nhiễm khuẩn)
- **Nhóm tuổi** (1-Trẻ, 2-Trung niên, 3-Già) và **Tình trạng hút thuốc** (Có hút, Không hút)
- **Tiêm chủng** (Có tiêm chủng, Không tiêm chủng) và **Tình trạng mắc bệnh** (Có mắc bệnh, không mắc bệnh)

- Nhóm tuổi (1-Trẻ, 2-Trung niên, 3-Già) và Mức độ sốt (1- Sốt nhẹ, 2- Sốt cao, 3- Sốt rất cao)

- Giới tính (Nam, Nữ) và Trình độ học vấn (Cao đẳng, Đại học, Sau đại học)

Lưu ý:

Trong trường hợp kiểm định giữa hai biến có thứ bậc, thay vì dùng đại lượng *Chi-Square*. có thể sử dụng các đại lượng khác tốt hơn như: *Kendall's Tau-b* (thích hợp khi số hàng = số cột), *Kendall's Tau-c* (thích hợp khi số hàng \neq số cột). Kiểm định mối liên hệ giữa hai biến thứ bậc thường là xác định độ mạnh (chặt chẽ) mối liên hệ tuyến tính của hai biến đó.

Ví dụ 1: Kiểm định xem Giới tính và Nhiễm khuẩn có ảnh hưởng lẫn nhau hay không?

1) Xây dựng giả thuyết và đối thuyết

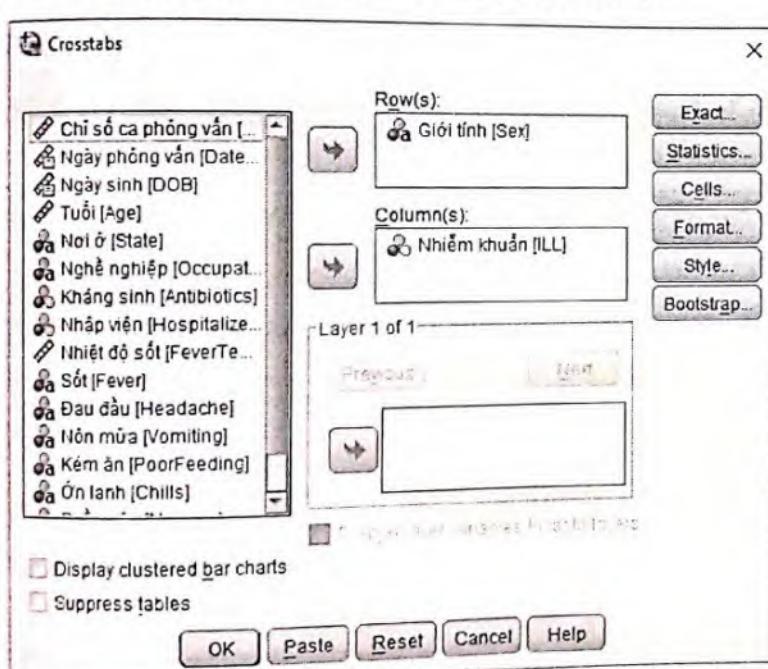
- H_0 : Giới tính và Nhiễm khuẩn là hai biến độc lập.

- H_1 : Giới tính và Nhiễm khuẩn là hai biến phụ thuộc.

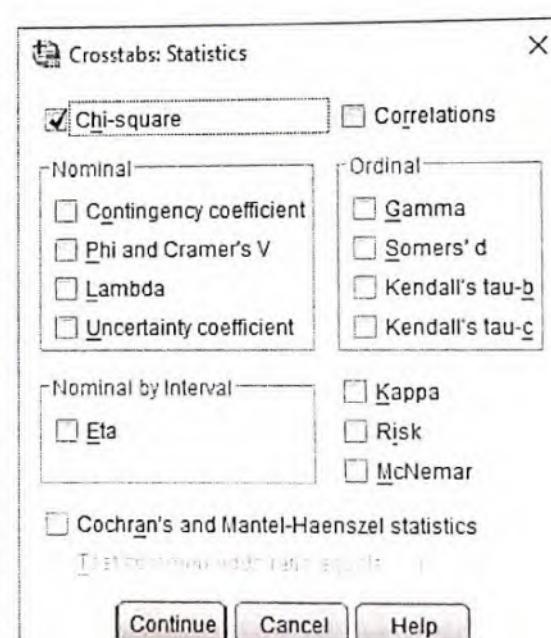
2) Tính giá trị χ^2_{QS} (bằng lệnh Crosstabs)

- Chọn Analyze → Descriptive Statistics → Crosstabs... xuất hiện hộp thoại Crosstabs.

Chọn các biến đưa vào tính toán như *Hình 4.6*.



Hình 4.6



Hình 4.7

- Chọn nút **Statistics...** và đánh dấu mục **Chi-square** như hộp thoại - *Hình 4.7*, chọn **Continue** trở về cửa sổ **Crosstabs**.

- Chọn **OK** và quan sát kết quả trong các bảng sau:

Giới tính * Nhiễm khuẩn Crosstabulation

Count		Nhiễm khuẩn		Total
		Không nhiễm khuẩn	Có nhiễm khuẩn	
Giới tính	Nữ	39	147	186
	Nam	44	129	173
Total	83	276	359	

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.006 ^a	1	.316		
Continuity Correction ^b	.770	1	.380		
Likelihood Ratio	1.005	1	.316		
Fisher's Exact Test				.320	.190
N of Valid Cases	359				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 40.00.

b. Computed only for a 2x2 table

Ta thấy bảng thống kê có dạng $2 \times 2 \rightarrow$ Đọc kết quả dòng 2: $\chi^2_{QS} = 0.77$, mức ý nghĩa (*Asymp. Sig. (2-sided)*) = 0.38

3) Xác định giá trị của χ^2_{LT}

Với bậc tự do = 1, độ tin cậy 95% ($\alpha = 0.05$) $\rightarrow \chi^2_{LT} = 3.84$ (tra trong Bảng giá trị χ^2_{LT})

4) Kết luận

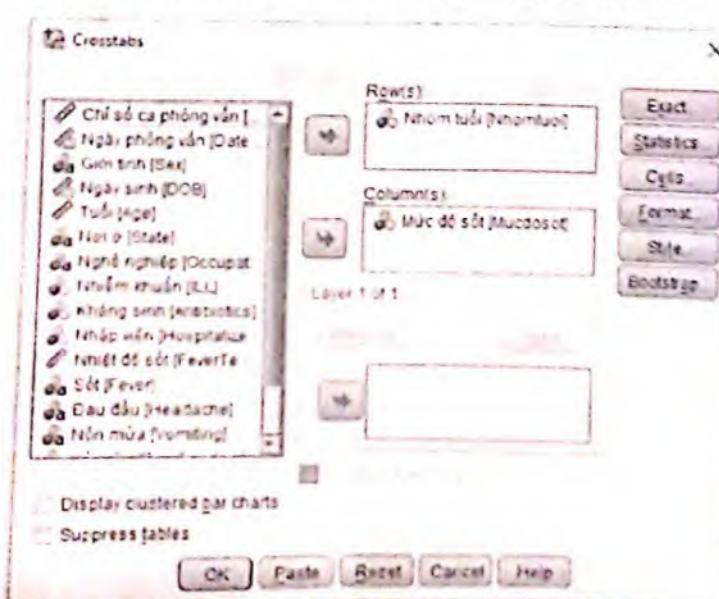
Ta thấy $\chi^2_{QS} = 0.77 < \chi^2_{LT} = 3.84 \rightarrow$ Chấp nhận giả thuyết H_0 , nghĩa là Giới tính và Nhiễm khuẩn là hai biến độc lập. Kết luận này cũng hoàn toàn phù hợp với *Asymp.Sig.(2-sided)* = 0.38 > 0.05 .

Ví dụ 2: Kiểm định mối quan hệ giữa hai biến **Nhomtuoi** (*Nhóm tuổi*) và **Mucedosot** (*Mức độ sốt*) - là hai biến thứ bậc.

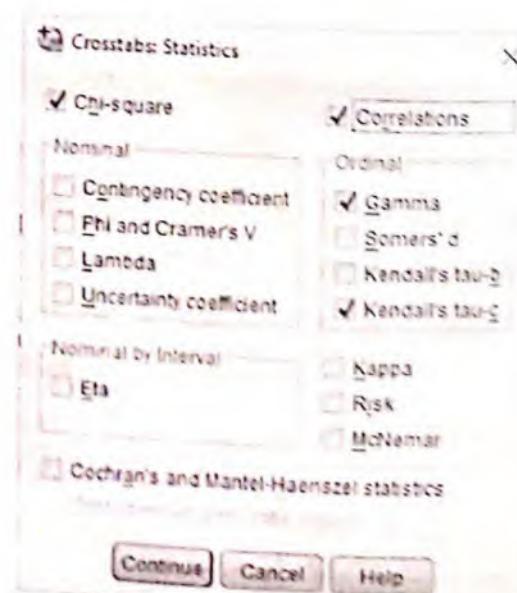
Xét mối quan hệ giữa hai biến thứ bậc Nhomtuoi và Mucedosot, ta thực hiện:

- Giả thuyết H_0 : Nhóm tuổi và Mức độ sốt là hai biến độc lập.
- Đổi thuyết H_1 : Nhóm tuổi và Mức độ sốt là hai biến phụ thuộc.
- Chọn **Analyze** \rightarrow **Descriptive Statistics** \rightarrow **Crosstabs...** xuất hiện hộp thoại Crosstabs.

Chọn các biến đưa vào tính toán - xem *Hình 4.8*.



Hình 4.8.



Hình 4.9.

- Chọn nút **Statistics...** và đánh dấu mục **Chi-square, Correlations, Gamma, Kendall's tau-c** như trong hộp thoại **Hình 4.9**.

- Chọn nút **Continue** để quay lại hộp thoại **Crosstabs** → Chọn **OK** và quan sát kết quả:

		Mức độ sорт		Total
		1	2	
Nhóm tuổi	1	75	32	107
	2	52	37	89
	3	15	1	16
Total		142	70	212

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.597 ^a	2	.014
Likelihood Ratio	10.072	2	.006
Linear-by-Linear Association	.049	1	.825
N of Valid Cases	212		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.28.

Symmetric Measures

		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-c	.018	.064	.281	.779
	Gamma	.036	.129	.281	.779
	Spearman Correlation	.019	.066	.270	.787 ^c
Interval by Interval	Pearson's R	-.015	.063	-.220	.826 ^c
N of Valid Cases		212			

- Từ các bảng kết quả trên, ta thấy bảng thống kê có dạng 3×2 → Đọc kết quả dòng 1 trong bảng **Chi-Square Tests**: $\chi^2_{QS} = 8.597$, mức ý nghĩa (*Asymp. Sig. (2-sided)*) = 0.014. Chỉ số **Gamma** trong bảng **Symmetric Measures** là $0.036 > 0$.

- Với bậc tự do $df = 2$, độ tin cậy 95% ($\alpha = 0.05$) → $\chi^2_{LR} = 5.99$ (tra trong Bảng giá trị χ^2_{LR})

- Ta thấy $\chi^2_{QS} = 8.597 > \chi^2_{LR} = 5.99 \rightarrow$ Bác bỏ giả thuyết H_0 .

Kết luận: Nhóm tuổi và Mức độ sорт là hai biến phụ thuộc. Điều này hoàn toàn phù hợp với *Asymp. Sig. (2-sided)* = 0.014 < 0.05 .

Bên cạnh đó, Chỉ số **Gamma** = $0.036 > 0$ và Kendall's tau-c (hệ số cấp bậc tương quan) = $0.018 > 0$ đều báo hiệu có sự tương quan thuận chiều giữa hai biến này (nếu **Gamma** = 0 thì không tương quan) → Hai biến có quan hệ với nhau và tương quan thuận chiều.

4.4. Nguy cơ tương đối (Relative Risk) và Tỷ suất chênh lệch (Odds Ratio)

4.4.1. Cơ sở lý thuyết

Cả RR (Relative Risk) và OR (Odds Ratio) đều là hai chỉ số thống kê rất phổ biến và có ích trong dịch tễ học, vì cả hai chỉ số đều được dùng cho kiểm định mối liên hệ giữa một yếu tố nguy cơ (phơi nhiễm) và một kết cục (bệnh, chết, hồi phục...). Các biến này đều là biến nhị phân.

Giả sử có bảng tần số quan sát của yếu tố nguy cơ và bệnh như sau:

	Bệnh	Không bệnh
Có phơi nhiễm	a	b
Không phơi nhiễm	c	d

• Các công thức tính toán

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)}$$

$$RR = \frac{\text{Tỷ lệ phát sinh bệnh trong nhóm có phơi nhiễm (tiếp xúc với nguy cơ)}}{\text{Tỷ lệ phát sinh bệnh trong nhóm không phơi nhiễm (không tiếp xúc với nguy cơ)}}$$

a, c nhỏ $RR \approx OR$

a, c lớn $RR < OR$.

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$OR = \frac{\text{Tỷ lệ giữa số ca bệnh có phơi nhiễm với không bệnh có phơi nhiễm}}{\text{Tỷ lệ giữa số ca bệnh không phơi nhiễm với không bệnh không phơi nhiễm}}$$

Trong trường hợp a rất nhỏ và c rất nhỏ tức là tỷ lệ phát sinh bệnh trong quần thể rất thấp $\rightarrow a+b$ sẽ gần bằng b và $c+d$ sẽ gần bằng d. Khi đó RR sẽ tiến đến gần bằng OR. Trường hợp ngược lại, a và b lớn thì RR luôn luôn nhỏ hơn OR. Nói cách khác, nếu tỉ lệ mắc bệnh thấp, thì OR gần bằng với RR. Nhưng nếu tỉ lệ mắc bệnh cao (chẳng hạn như trên 10%) thì chỉ số OR cũng cao hơn chỉ số RR.

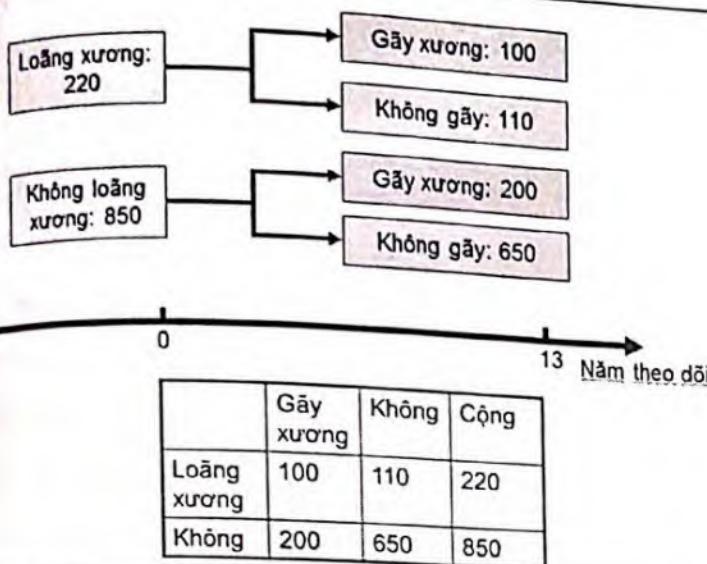
RR là tỉ số của hai tỉ lệ, nếu nói tỉ lệ mắc bệnh 3%, có nghĩa 3 trong 100 người mắc bệnh. Nếu RR = 2, có thể nói rằng tỉ lệ tăng gấp 2 lần. OR là tỉ số của hai tỉ lệ. Tỷ lệ phản ánh “khả năng” mắc bệnh. Tỷ lệ bằng 2 có nghĩa là khả năng mắc bệnh cao hơn khả năng không mắc bệnh 2 lần. Đơn vị của RR là tỉ lệ mắc bệnh cho nên chúng ta có thể nói rằng nhóm phơi nhiễm có tỉ lệ mắc bệnh cao/thấp hơn nhóm đối chứng. Nhưng với OR chỉ có thể phát biểu rằng “khả năng” mắc bệnh của nhóm phơi nhiễm cao/thấp hơn nhóm đối chứng.

• Ý nghĩa RR và OR [4]

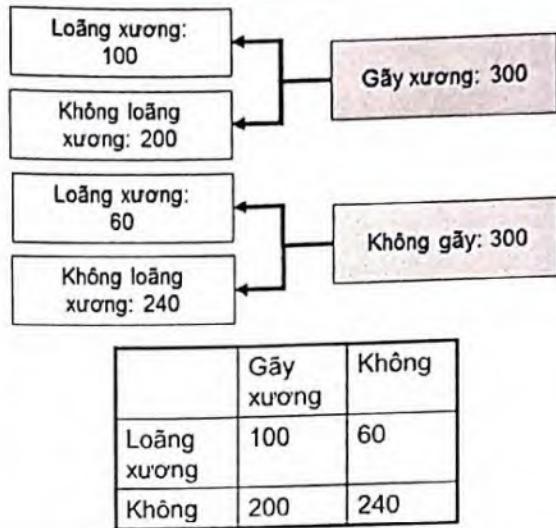
RR	OR
• $RR > 1$: yếu tố nguy cơ làm tăng khả năng mắc bệnh.	• $OR > 1$: khả năng mắc bệnh cao hơn khả năng không mắc bệnh.
• $RR = 1$ không có mối liên hệ nào giữa yếu tố nguy cơ và khả năng mắc bệnh.	• $OR = 1$ khả năng mắc bệnh tương đương với khả năng không mắc bệnh.
• $RR < 1$: yếu tố nguy cơ làm giảm khả năng mắc bệnh	• $OR < 1$: khả năng mắc bệnh thấp hơn khả năng không mắc bệnh.

- Phân biệt giữa RR và OR [5]

RR: Nghiên cứu theo thời gian
(prospective/longitudinal study)



OR: Nghiên cứu bệnh chứng
(case-control study)



- Tỷ lệ phát sinh gãy xương trong nhóm bị loãng xương: $I_{lx} = 100/220 = 0.454$

- Tỷ lệ phát sinh gãy xương trong nhóm không bị loãng xương: $I_{kly} = 200/850 = 0.236$

$$RR = I_{lx}/I_{kly} = 0.454/0.236 = 2.13$$

hoặc

$$O_{lx} = 100/110 = 0.909$$

$$O_{kly} = 200/650 = 0.308$$

$$OR = O_{lx}/O_{kly} = 0.909/0.308 = 2.95$$

- Tỷ lệ giữa đối tượng gãy xương và không gãy xương trong *nhóm bị loãng xương*: $O_{lx} = 100/60 = 1.667$

- Tỷ lệ giữa đối tượng gãy xương và không gãy xương trong *nhóm không bị loãng xương*: $O_{kly} = 200/240 = 0.833$

$$\text{Odds ratio} = O_{lx}/O_{kly} = 1.667/0.833 = 2.00$$

RR là chỉ số cần biết và có thể giải thích dễ dàng, trực tiếp nói lên nguy cơ mắc bệnh tăng hay giảm hoặc không tăng không giảm. Đối với nghiên cứu xuôi theo thời gian: có thể tính được cả RR và OR. Các nghiên cứu bệnh chứng (nghiên cứu tại một thời điểm) chỉ có thể tính được OR.

Kết luận:



Nếu trong khoảng tin cậy 95% của OR không chứa 1 thì sự khác nhau giữa 2 tỉ lệ có ý nghĩa thống kê (yếu tố nguy cơ có ảnh hưởng đến tình trạng bệnh); nếu trong khoảng tin cậy 95% của OR có chứa 1 thì sự khác nhau giữa 2 tỉ lệ chưa có ý nghĩa thống kê (yếu tố nguy cơ chưa ảnh hưởng đến tình trạng bệnh).

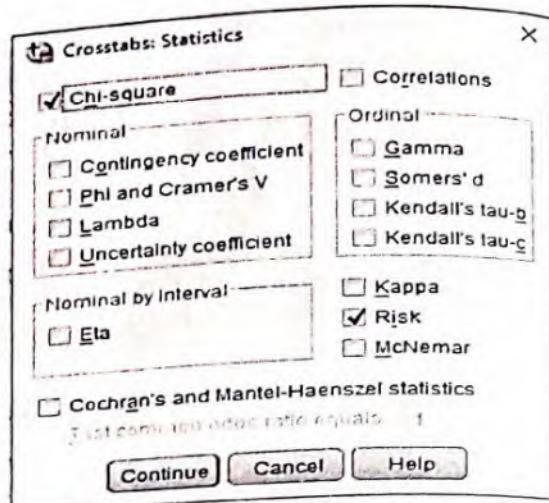
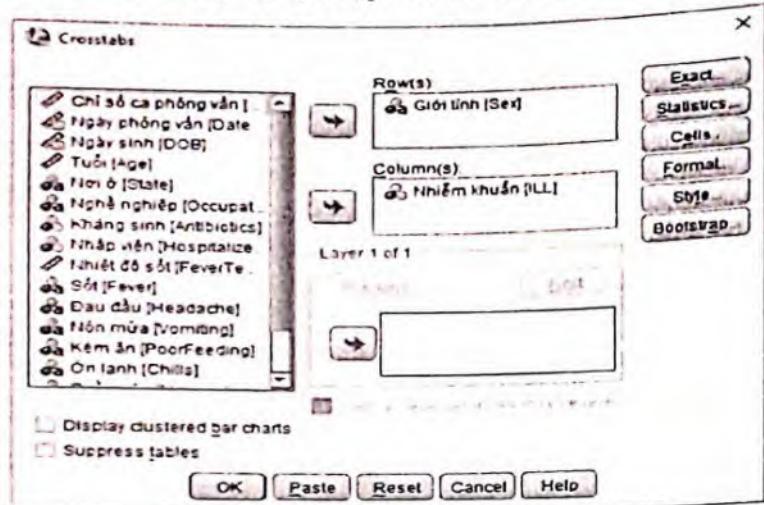
4.4.2. Ví dụ minh họa

Xét xem ảnh hưởng của yếu tố Giới tính (Sex) tới Nhiễm khuẩn (ILL)?

- Chọn Analyze → Descriptive Statistics → Crosstabs; Chọn các biến cần xem xét như Hình 4.10.

Bộ môn Tin học – Trường Đại học Y Dược Hải Phòng

- Chọn nút Statistics → Đánh dấu Test Chi-square và mục Risk như Hình 4.11 → Nhấn Continue để quay lại hộp thoại Hình 4.11



Hình 4.11.

Hình 4.10.

- Nhập OK và quan sát các bảng kết quả:

Giới tính * Nhiễm khuẩn Crosstabulation

Count

		Nhiễm khuẩn		Total
		Không nhiễm khuẩn	Có nhiễm khuẩn	
Giới tính	Nữ	39	147	186
	Nam	44	129	173
Total	83	276	359	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.006 ^a	1	.316		
Continuity Correction ^b	.770	1	.380		
Likelihood Ratio	1.005	1	.316		
Fisher's Exact Test				.320	.190
N of Valid Cases	359				

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Giới tính (Nữ / Nam)	.778	.476	1.272
For cohort Nhiễm khuẩn = Không nhiễm khuẩn	.824	.565	1.203
For cohort Nhiễm khuẩn = Có nhiễm khuẩn	1.060	.945	1.188
N of Valid Cases	359		

Phần mềm SPSS chỉ tính OR. Giá trị này cho ở dòng đầu tiên của bảng Risk Estimate.

Từ bảng Risk Estimate, ta có OR = 0.778, khoảng tin cậy 95% của OR = 0.476 - 1.272.

Điều này chứng tỏ yếu tố Giới tính chưa thực sự ảnh hưởng đến Nhiễm khuẩn.

Kết luận này cũng phù hợp với kết quả đọc trong bảng Chi-Square Tests:

Bài 5. KIỂM ĐỊNH PHÂN PHỐI CHUẨN & KIỂM ĐỊNH TRUNG BÌNH**Mục tiêu:**

- Trình bày được ý nghĩa và các bước kiểm định phân phối chuẩn, kiểm định trung bình.
- Sử dụng được các test thống kê để kiểm định trung bình.
- Đọc và phiên giải được ý nghĩa các tham số thống kê trong các bài toán kiểm định.

5.1. Kiểm định phân phối chuẩn

Phân phối chuẩn (*Normal distribution*) hay Kiểm định phân phối chuẩn là đánh giá xem các giá trị của một biến định lượng nào đó có phải là một phân phối chuẩn hay không để chọn các Test thống kê phù hợp (*Test có tham số* hoặc *Test phi tham số*) trong các bài toán kiểm định trung bình.

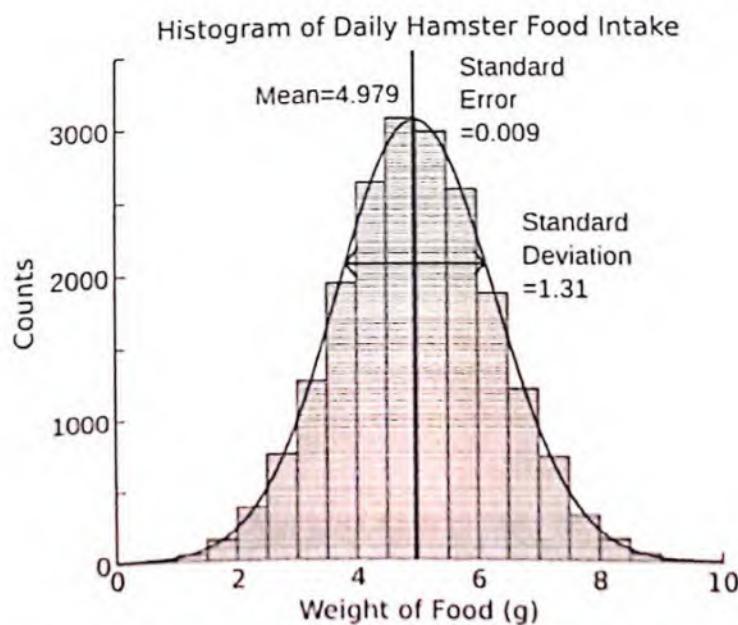
- **Cách đánh giá một phân phối chuẩn trong SPSS**

- Đơn giản nhất là xem biểu đồ với đường cong chuẩn (*Histograms with normal curve*) với dạng hình chuông úp đối xứng có tần số cao nhất nằm chính giữa (đỉnh chuông) và các tần số thấp dần nằm ở 2 bên. Giá trị trung bình (*mean*) và trung vị (*median*) gần bằng nhau và độ lệch – hệ số bất đối xứng (*skewness*) gần bằng zero (gần bằng không).

- Vẽ biểu đồ xác suất chuẩn (*normal Q-Q plot*). Phân phối được gọi là chuẩn khi biểu đồ xác suất này có quan hệ tuyến tính (đường thẳng).

- Dùng phép kiểm định Kolmogorov-Smirnov khi cỡ mẫu lớn hơn 50 hoặc phép kiểm định Shapiro-Wilk khi cỡ mẫu nhỏ hơn 50. Các đại lượng được gọi là phân phối chuẩn khi mức ý nghĩa (*Sig.*) lớn hơn 0,05.

Tuy nhiên, chúng ta thường kiểm định phân phối chuẩn bằng cách thứ 3 vì độ chính xác của nó cao hơn do sử dụng các Test thống kê.

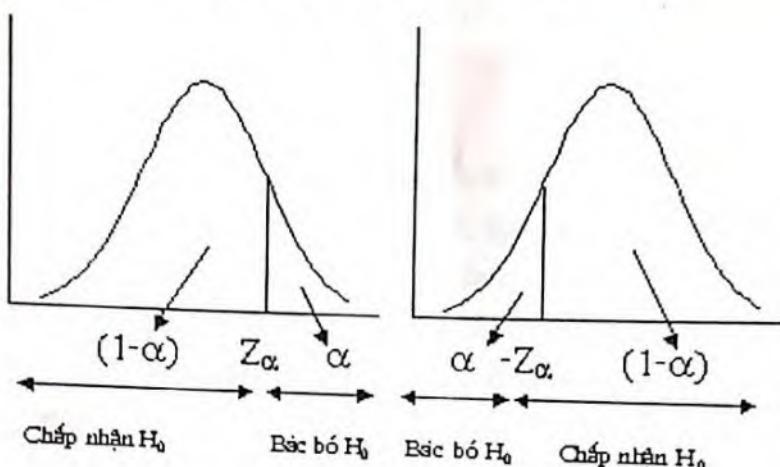


Hình 5.1a: Hình minh họa đại lượng tuân theo luật phân phối chuẩn

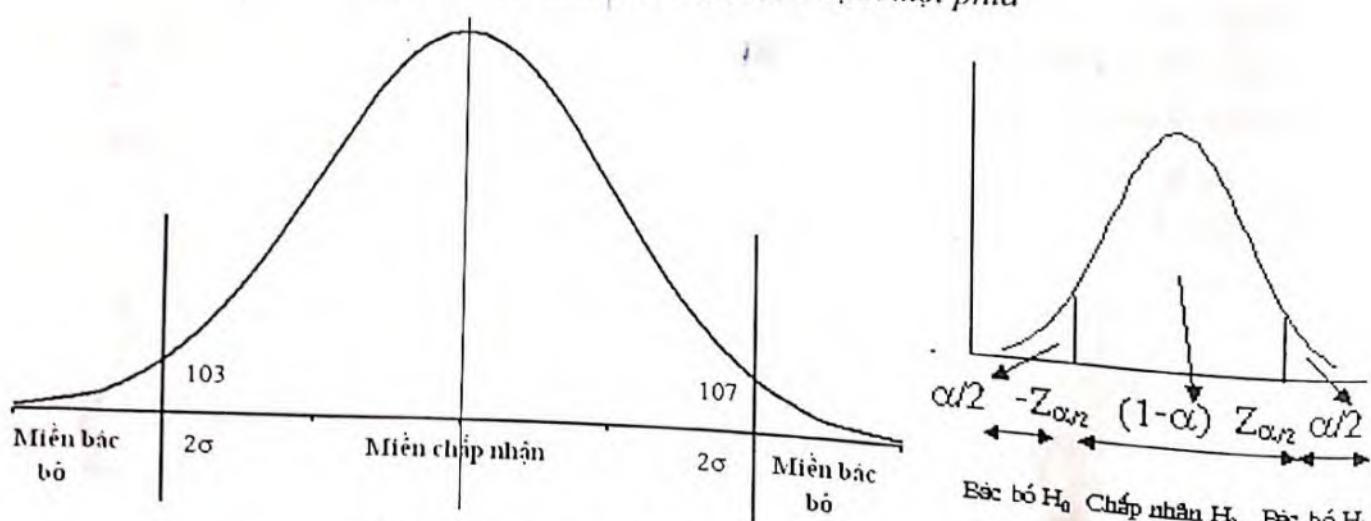
Kiểm định ý nghĩa một phía (1-tailed) và hai phía (2-tailed)

Kiểm định một phía hoặc hai phía được quyết định do việc chúng ta đặt giả thuyết thống kê; Nếu giả thuyết chỉ xác định xem hai giá trị có *khác nhau* hay không và không quan tâm đến giá trị nào lớn hơn là kiểm định hai phía. **Nếu kiểm định liên quan đến việc xác định xem giá trị A có thực sự lớn/nhỏ hơn giá trị B hay không thì lúc đó là kiểm định một phía** (Xem *Hình 5.1b, Hình 5.1c.*

Việc xác định rõ kiểm định một phía hay hai phía rất quan trọng vì nó giúp chúng ta xác định chính xác miền bác bỏ hay chấp nhận giả thuyết (H_0) và giúp ta xác định việc so sánh chính xác giá trị thống kê và giá trị tra bảng để đi đến kết luận chính xác.



Hình 5.1b: Hình minh họa kiểm định một phía



Hình 5.1c: Hình minh họa kiểm định hai phía

• Ví dụ

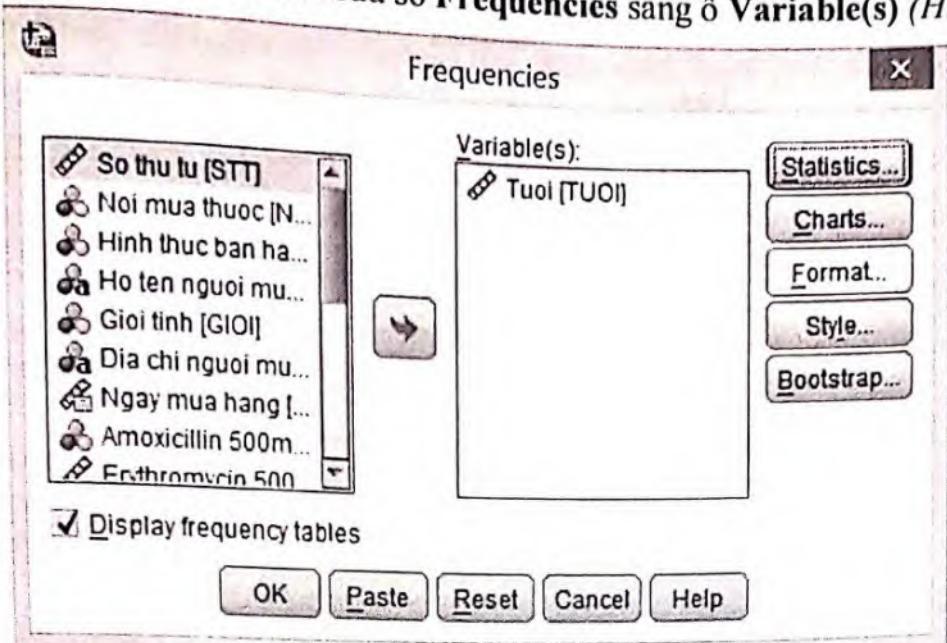
Trong bộ số liệu điều tra về tình hình mua và sử dụng thuốc kháng sinh, vitamin cơ bản có 40 bản ghi, trong đó Tuổi (TUOI) là biến định lượng. Ở ví dụ 1 sau đây sẽ kiểm tra xem Tuổi có phải là một phân phối chuẩn hay không.

Ví dụ 1: Kiểm định Tuổi có phải là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn hay không. Nói cách khác, Kiểm định Tuổi có phải là một phân phối chuẩn hay không?

Cách 1: Sử dụng biểu đồ đường cong chuẩn (*Histograms with normal curve*) và giá trị trung bình, trung vị, hệ số bất đối xứng.

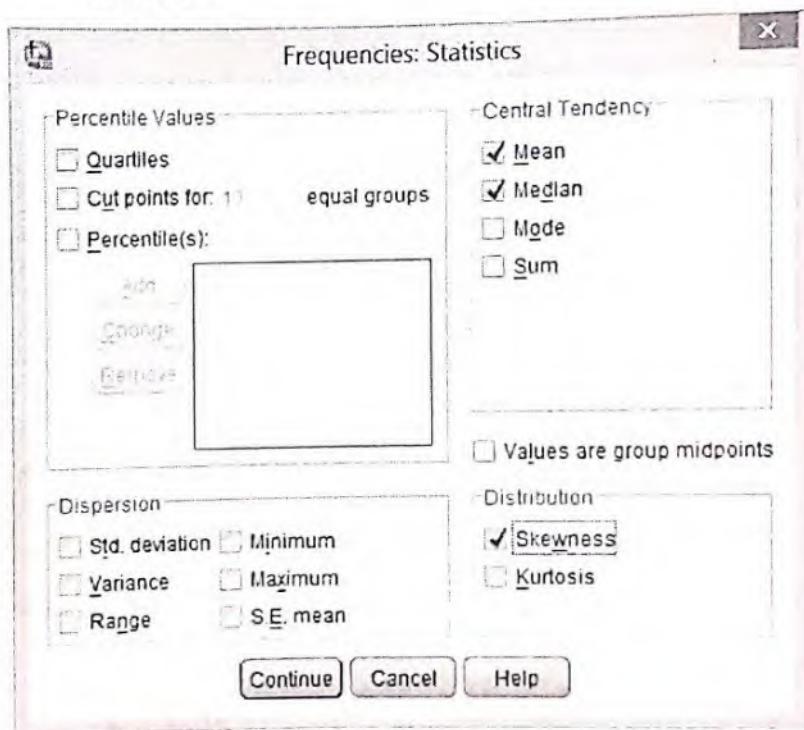
- Chọn **Analyze → Descriptive Statistics → Frequencies**, xuất hiện hộp thoại

- Chọn biến TUOI từ ô bên trái cửa sổ Frequencies sang ô Variable(s) (Hình 5.2).



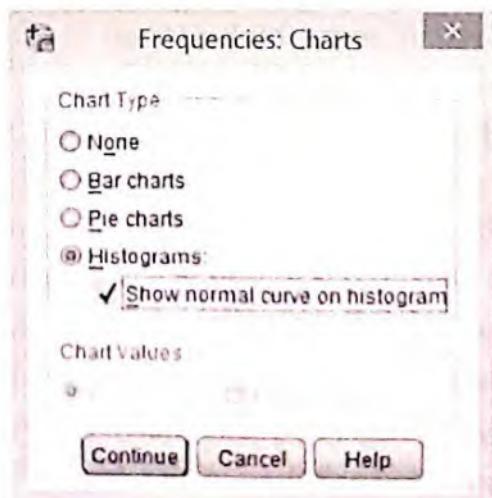
Hình 5.2

- Nhấp chọn nút Statistics... và đánh dấu (✓) vào các mục Mean, Median (trong khung Central Tendency), Skewness (trong khung Distribution) – Hình 5.3 → Nhấp nút Continue.



Hình 5.3

- Nhấp chọn nút Charts (trên Hình 5.2), đánh dấu mục Histograms và Show normal curve on histogram (Hình 5.4) → Chọn Continue.



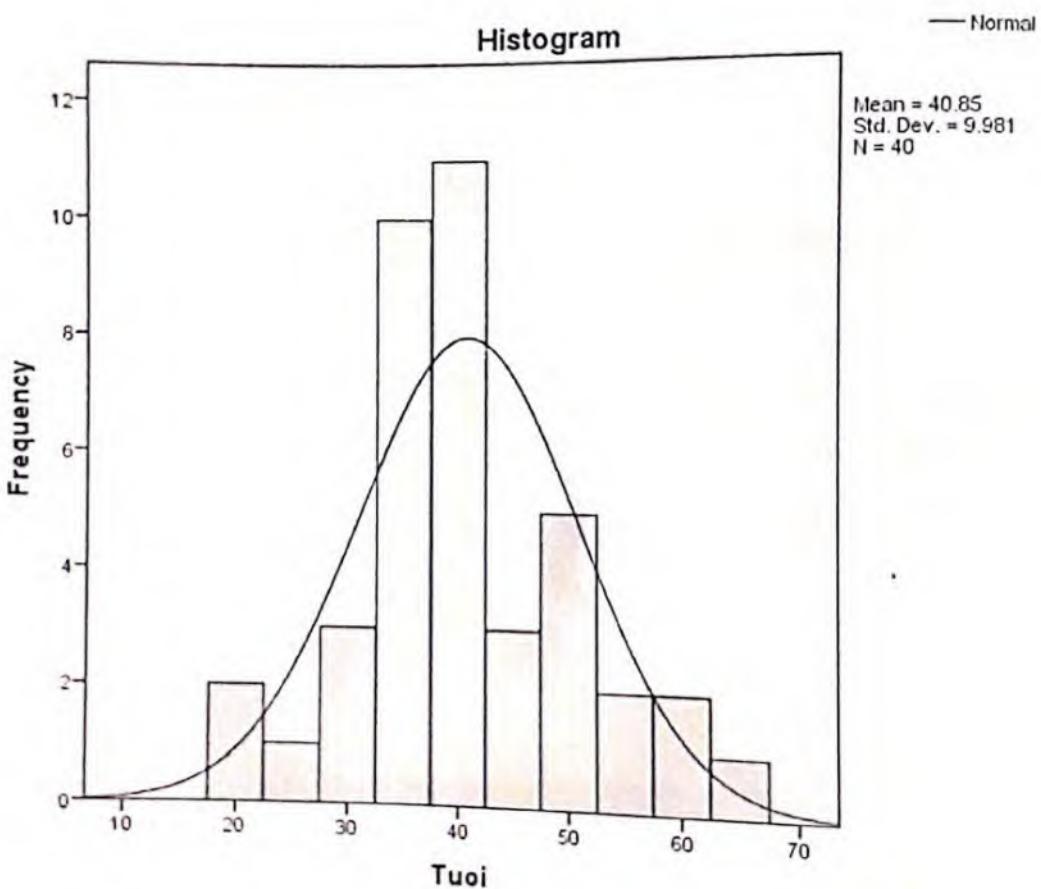
Hình 5.4

- Chọn OK, quan sát kết quả Hình 5.5 và biểu đồ Histogram

Statistics

Tuoi		
N	Valid	40
	Missing	0
	Mean	40.85
	Median	40.00
	Skewness	.258
	Std. Error of Skewness	.374

Hình 5.5



Hình 5.6

Từ kết quả Hình 5.5 ta thấy:

- Giá trị trung bình (Mean) = 40.85, trung vị (Median) = 40.00, hệ số bất đối xứng (Skewness) = 0.258.

- Trong phân phối này, trị số trung bình và trung vị gần bằng nhau, hệ số bất đối xứng dao động từ -1 đến +1.

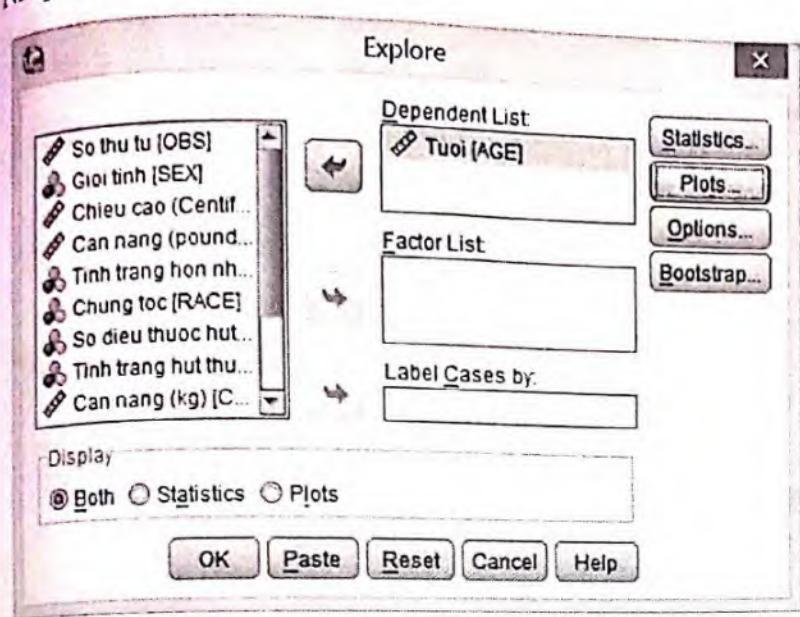
- Bên cạnh đó biểu đồ phân phối với đường cong chuẩn có dạng chuông (Hình 5.6), có giá trị trung bình là 40.85 và số liệu phân phối tương đối đều sang hai bên.

Do đó, Tuổi có thể là một phân phối chuẩn. Tuy nhiên để đánh giá Tuổi có phải là phân phối chuẩn thực sự không ta nên kiểm định bằng các Test cụ thể (Cách 2).

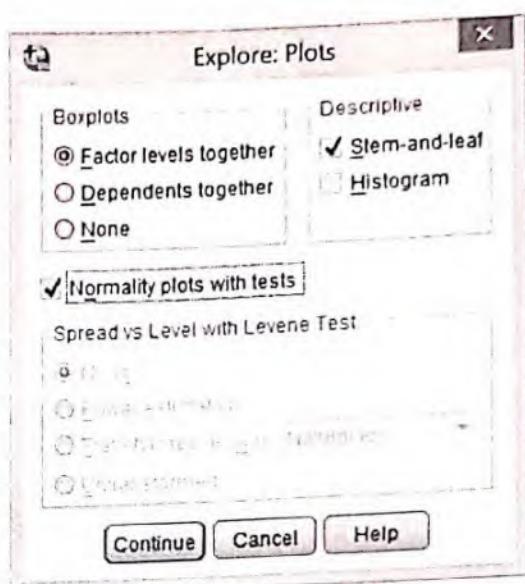
Cách 2: Sử dụng phép kiểm định Shapiro-Wilk (vì cỡ mẫu nhỏ hơn 50) kết hợp với biểu đồ xác suất chuẩn.

- Chọn Analyze → Descriptive Statistics → Explore, xuất hiện hộp thoại Explore.

- Chọn biến định lượng cần kiểm định (biến TUOI) – Hình 5.7.
- Chọn nút Plots, đánh dấu mục Normality plots with tests và khai báo như Hình 5.8 → Nhập Continue.



Hình 5.7



Hình 5.8

- Chọn OK.

SPSS sẽ hiển thị kết quả gồm nhiều mục trong bảng Descriptives và Tests of Normality như sau:

Descriptives

		Statistic	Std. Error
Tuoi	Mean	40.85	1.578
	95% Confidence Interval for Mean		
	Lower Bound	37.66	
	Upper Bound	44.04	
	5% Trimmed Mean	40.78	
	Median	40.00	
	Variance	99.618	
	Std. Deviation	9.981	
	Minimum	20	
	Maximum	64	
	Range	44	
	Interquartile Range	13	
	Skewness	.258	.374
	Kurtosis	.413	.733

Tests of Normality

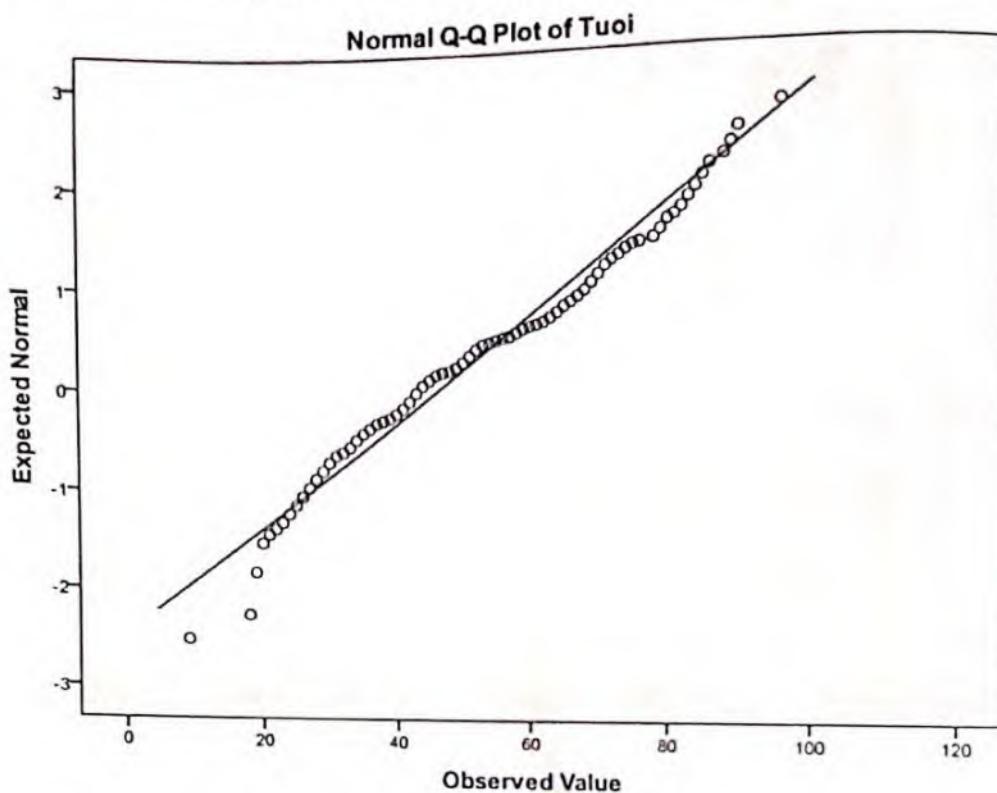
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
Tuoi	.083	337	.000	.965	337	.000

^aSignificance Correction

Từ bảng trên ta thấy giá trị trung bình (Mean) = 40.85, trung vị (Median) = 40.00 gần bằng nhau, hệ số bất đối xứng (Skewness) = 0.258; Test Shapiro-Wilk có Sig. < 0.05.

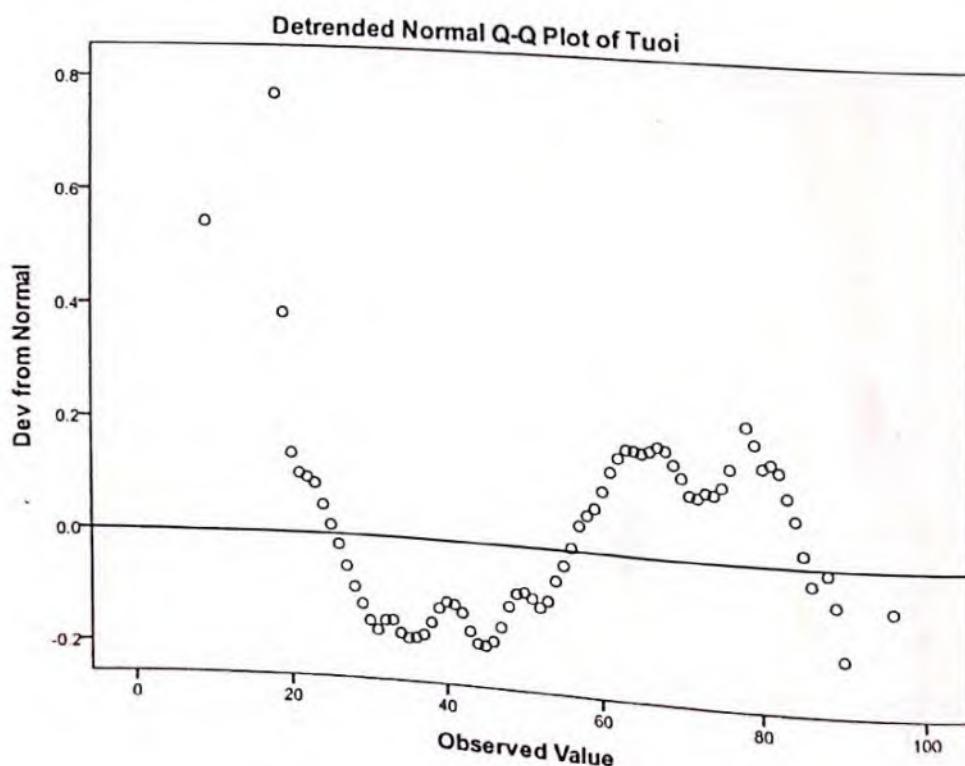
Vậy, Tuổi là đại lượng ngẫu nhiên không phân phối chuẩn.

Quan sát biểu đồ Normal Q-Q Plot (Hình 5.8) và Detrended Normal Q-Q Plot (Hình 5.9).



Hình 5.8

Trong biểu đồ Normal Q-Q Plot, đường thẳng biểu thị cho phân phối chuẩn lý thuyết còn các điểm chấm biểu thị các giá trị thực tế của Tuổi. Nếu chúng càng gần nhau thì khả năng phân phối thực tế sẽ là phân phối chuẩn càng lớn.



Hình 5.9

Biểu đồ xác suất chuẩn Detrended Normal Q-Q Plot, đã được khử xu thế: Hệ thống so sánh với đường song song với trục hoành đi qua trung bình 0.

Ví dụ 2: Khảo sát men ALT (U/L) của 30 người bình thường, kết quả được nhập vào SPSS như bảng bên dưới. Hãy cho biết ALT có phải là một phân phối chuẩn hay không?

Cách 1: Sử dụng biểu đồ đường cong chuẩn (*Histograms with normal curve*) và giá trị trung bình, trung vị và hệ số bát đối xứng.

- Chọn **Analyze → Descriptive Statistics → Frequencies**, xuất hiện hộp thoại **Frequencies**.

- Chọn biến **ALT** là biến cần kiểm tra phân phối chuẩn đưa sang ô **Variable(s)**.

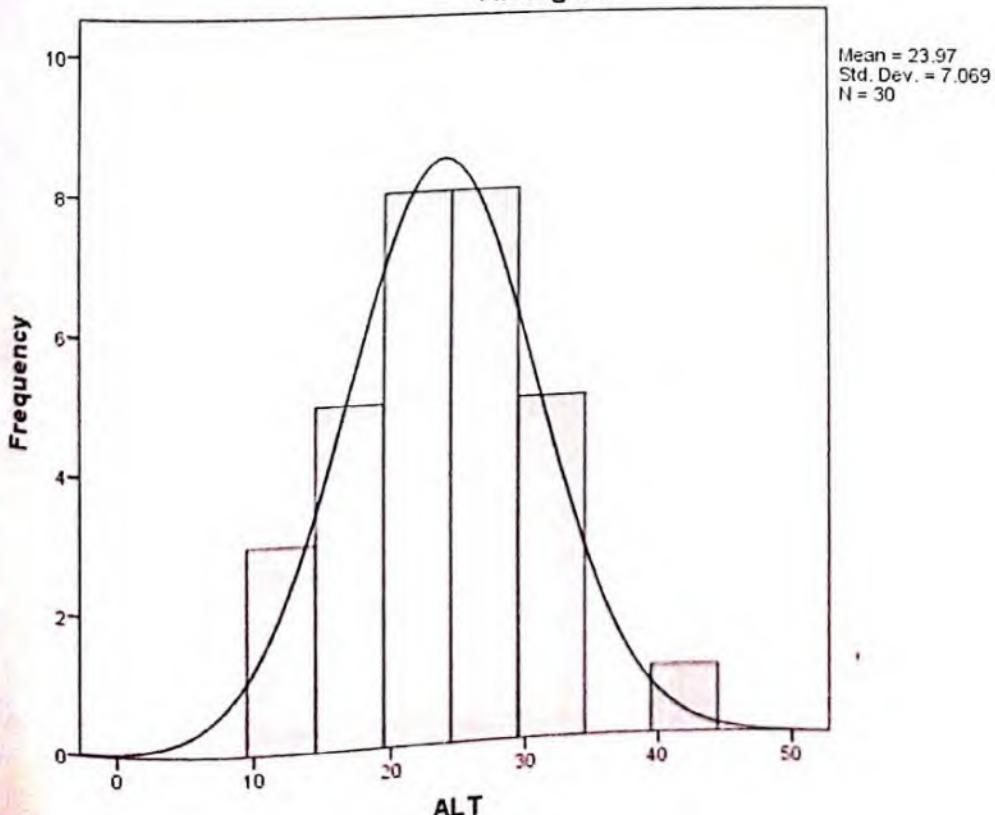
- Nhấp chọn các nút : **Statistics, Chart** và đánh dấu tích vào các mục cần thiết như trong cách 1 của ví dụ 1.

- Chọn **OK** và quan sát bảng **Statistics** và biểu đồ *Hình 5.10*.

Statistics		
ALT		
N	Valid	30
	Missing	10
Mean		23.97
Median		24.00
Mode		26
Skewness		.533
Std. Error of Skewness		.427

Hình 5.10

Histogram



Hình 5.10

MaKS	ALT
1	12
2	13
3	14
4	15
5	16
6	17
7	18
8	19
9	20
10	21
11	22
12	23
13	24
14	25
15	26
16	27
17	28
18	29
19	30
20	31
21	32
22	33
23	34
24	26
25	22
26	23
27	24
28	25
29	26

Từ kết quả trên ta có: Giá trị trung bình (*Mean*) = 23.97 và Trung vị (*Median*) = 24 gần bằng nhau; Hệ số bất đối xứng (*Skewness*) = 0.533 dao động từ -1 đến +1;

Biểu đồ phân phối với đường cong chuẩn có dạng chuông, có giá trị trung bình là 23.97 và số liệu phân phối tương đối đều sang hai bên. Do đó, **ALT** có thể là một phân phối chuẩn.

Cách 2: Sử dụng test Shapiro-Wilk.

Thực hiện các bước chọn lệnh và đưa biến tương tự như trong cách 2 của ví dụ 1 ta nhận được kết quả trong bảng **Descriptives** và biểu đồ - *Hình 5.11, Hình 5.12*:

Descriptives

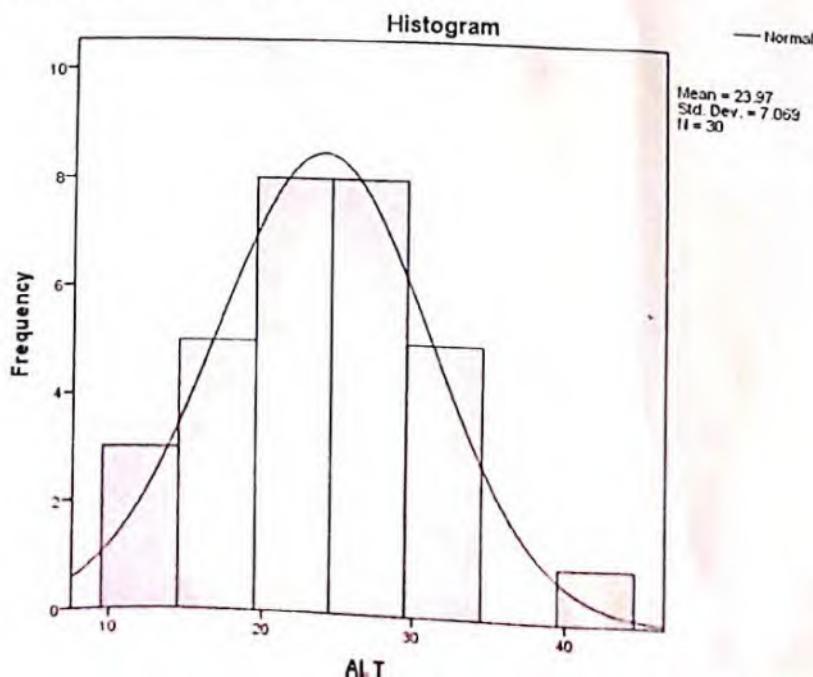
		Statistic	Std. Error
ALT	Mean	23.97	1.291
	95% Confidence Interval for Mean	Lower Bound	21.33
		Upper Bound	26.61
	5% Trimmed Mean	23.69	
	Median	24.00	
	Variance	49.964	
	Std. Deviation	7.069	
	Minimum	12	
	Maximum	44	
	Range	32	
	Interquartile Range	10	
	Skewness	.533	.427
	Kurtosis	.856	.833

Tests of Normality

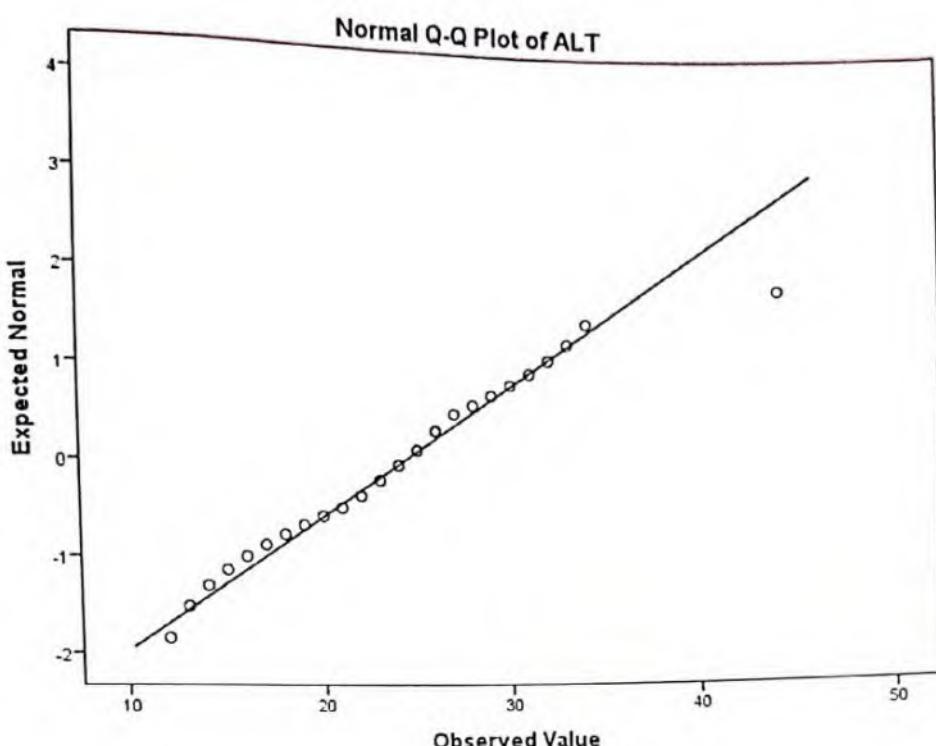
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
ALT	.087	30	.200*	.971	30	.571

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



Hình 5.11



Hình 5.12

Từ bảng kết quả thống kê ở trên ta thấy:

Test Shapiro-Wilk có Sig. = 0.571 > 0.05. Do đó ta khẳng định ALT là một phân phối chuẩn.

Như vậy: Để khẳng định một đại lượng nào đó có phải là một phân phối chuẩn hay không nên sử dụng các Test thống kê để kiểm định.

5.2. Kiểm định trung bình

Trong phần này sẽ thực hiện kiểm định trung bình như: So sánh trung bình quan sát và trung bình lý thuyết, So sánh hai hoặc nhiều trung bình, So sánh ghép cặp. Đây là một số bài toán đặc trưng thường gặp trong trong nghiên cứu y học.

Có 2 loại test được sử dụng trong thống kê: Test có tham số và Test phi tham số. Việc sử dụng loại Test nào phụ thuộc vào đại lượng cần kiểm định có phải là một phân phối chuẩn hay không. **Sử dụng Test có tham số khi đại lượng cần kiểm định là một phân phối chuẩn và Test phi tham số nếu đại lượng cần kiểm định không phải là một phân phối chuẩn.**

Trong các bài toán kiểm định trung bình nói riêng hay biến định lượng (hoặc các biến số liên tục) nói chung thường sử dụng T-Test (hay phép kiểm T) để kiểm định trung bình của biến định lượng khi biến đó là một phân phối chuẩn.

Với các bộ số liệu nghiên cứu có cỡ mẫu nhỏ (dưới 30), trước khi kiểm định trung bình cần thực hiện kiểm định phân phối chuẩn đối với biến đó.

Với các bộ số liệu nghiên cứu có cỡ mẫu lớn (hàng trăm trường hợp) người dùng có thể bỏ qua bước kiểm định chuẩn trước khi so sánh các giá trị trung bình. Tuy nhiên, để cẩn thận, chúng tôi vẫn khuyến khích các nhà phân tích nên kiểm định phân phối chuẩn trước khi kiểm định trung bình.

Để áp dụng cho các ví dụ về kiểm định trung bình, chúng ta sử dụng bộ số liệu về hút thuốc (Smoke.Sav) với cỡ mẫu đủ lớn: 270 bản ghi (case). Bảng sau đây cho biết cấu trúc và ý nghĩa các biến.

Name	Type	Width	Decimals	Label	Values	Missing
OBS	Numeric	3	0	Số thứ tự	None	None
SEX	Numeric	1	0	Giới tính	{0, Nu}...	None
HEIGHT	Numeric	5	2	Chiều cao (Centileet)	None	None
WEIGHT	Numeric	4	2	Cân nặng (pound)	None	None
MARITAL	Numeric	2	0	Tình trạng hôn nhân	None	None
RACE	Numeric	1	0	Chủng tộc	None	None
AGE	Numeric	2	0	Tuổi	None	None
NUMCIGAR	Numeric	2	0	Số điếu thuốc hút một ngày	None	None
SMOKE	Numeric	1	0	Tình trạng hút thuốc	{1, Hút thuốc}	None
Cannang	Numeric	5	2	Cân nặng (kg)	None	None
Chiềucao	Numeric	7	2	Chiều cao (m)	None	None

Với bài toán kiểm định (so sánh) trung bình ta sử dụng các bước sau đây:

1) Xác định giả thuyết, đối thuyết H_1 .

- Giả thuyết H_0 : Các đối tượng cần so sánh là như nhau.
- Đối thuyết H_1 : Các đối tượng cần so sánh khác nhau.

2) Tính tqs.

3) Tính t_α căn cứ vào độ tin cậy và bậc tự do:

- Tính bậc tự do df: $df = n_1 + n_2 - 2$

- Xác định α (tùy theo độ tin cậy):

 - + Với độ tin cậy 90% : $\alpha = 0.10$

 - + Với độ tin cậy 95% : $\alpha = 0.05$

 - + Với độ tin cậy 99% : $\alpha = 0.01$

4) So sánh tqs với t_α và rút ra kết luận:

- Nếu $t_{qs} < t_\alpha$: Chấp nhận giả thuyết H_0 .

- Nếu $t_{qs} > t_\alpha$: Bác bỏ giả thuyết H_0 .

5.2.1) So sánh trung bình quan sát và trung bình lý thuyết

Ví dụ: So sánh Số điếu thuốc hút trung bình một ngày trong nghiên cứu này với 10.

1) Xác định giả thuyết, đối thuyết H_1 .

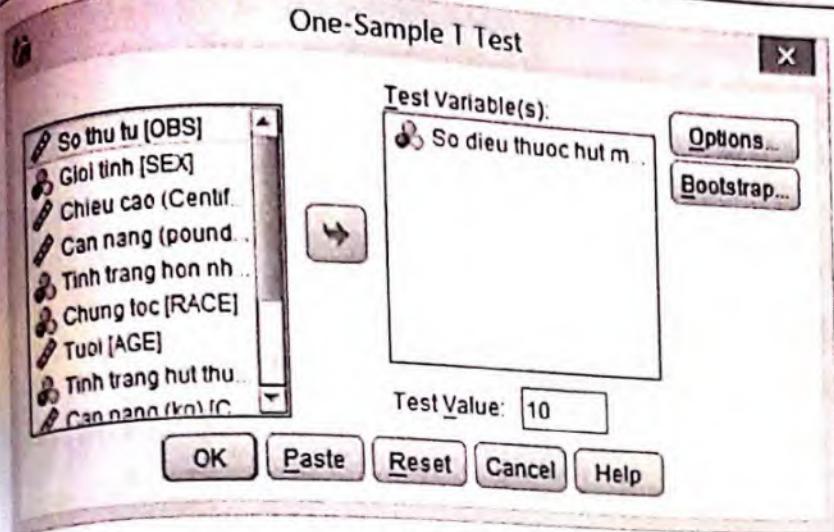
- Giả thuyết H_0 : Số điếu thuốc hút trung bình một ngày trong nghiên cứu này là 10.
- Đối thuyết H_1 : Số điếu thuốc hút trung bình một ngày trong nghiên cứu này khác 10.

2) Tính tqs bằng lệnh One-Sample T Test

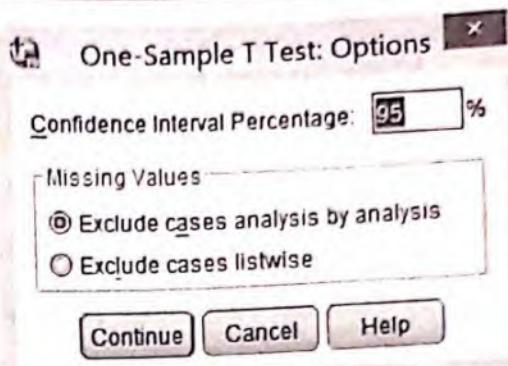
- Chọn Analyze → Compare Means → One-Sample T Test...

- Chọn biến cần kiểm định đưa sang khung Test Variable(s),

- Trong mục Test Value: Nhập giá trị cần kiểm định, ở đây là 10 (xem Hình 5.13):



Hình 5.13



Hình 5.14

- Chọn **Options** khai báo **Confidence Interval Percentage** (*khoảng tin cậy*) = 95 (xem Hình 5.14) → Nhập **Continue** để quay lại hộp thoại Hình 5.13.

- Nhập **OK** và quan sát các bảng kết quả Hình 5.15

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
So dieu thuoc tren ngay	72	17.06	9.185	1.082

One-Sample Test

	Test Value = 10					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
So dieu thuoc tren ngay	6.518	71	.000	7.056	4.90	9.21

Hình 5.15

Từ các bảng kết quả trên ta thấy:

- N: số người hút thuốc là 72, Mean: trung bình số điếu thuốc hút 1 ngày của 72 người là 17.06, Std. Deviation: Độ lệch chuẩn là 9.185, Std. Error Mean: Sai số chuẩn = 1.082.
- Giá trị t-test là 6.518 \rightarrow $t_{qs} = 6.518$; df: Bậc tự do = $(n - 1)$, Mức ý nghĩa của kiểm định 2 phía (Sig. (2-tailed)) là $p < 0.05$;
- Sự khác biệt giữa 2 giá trị so sánh (Mean Difference: khác biệt giữa trung bình mẫu và trung bình lý thuyết) là 7.056. Khoảng tin cậy 95% của sai số là 4.90– 9.21

3) Tính t_α

Bậc tự do $df = 72 - 1 = 71$, với khoảng tin cậy 95% ta có $\alpha = 0.05$;

Tra bảng phân phối Student (t) tìm được $t_\alpha = 1.9945$;

Tra bảng phân phối Student (t) tìm được $t_\alpha = 1.9945$;

4) Kết luận

Vì $t_{qs} > t_\alpha$ nên bác bỏ giả thiết H_0 .

Vậy Số điếu thuốc hút trung bình một ngày trong nghiên cứu này khác 10. Sự khác biệt giữa trung bình số điếu thuốc hút trong một ngày là có ý nghĩa thống kê với mức ý nghĩa Sig. (2-tailed) = 0.0000 < 0.05.

5.2.2. So sánh trung bình quan sát của hai nhóm độc lập

Ví dụ: So sánh trung bình số điếu thuốc hút trong một ngày của nam và nữ.

1) Xác định giả thuyết, đối thuyết H₁.

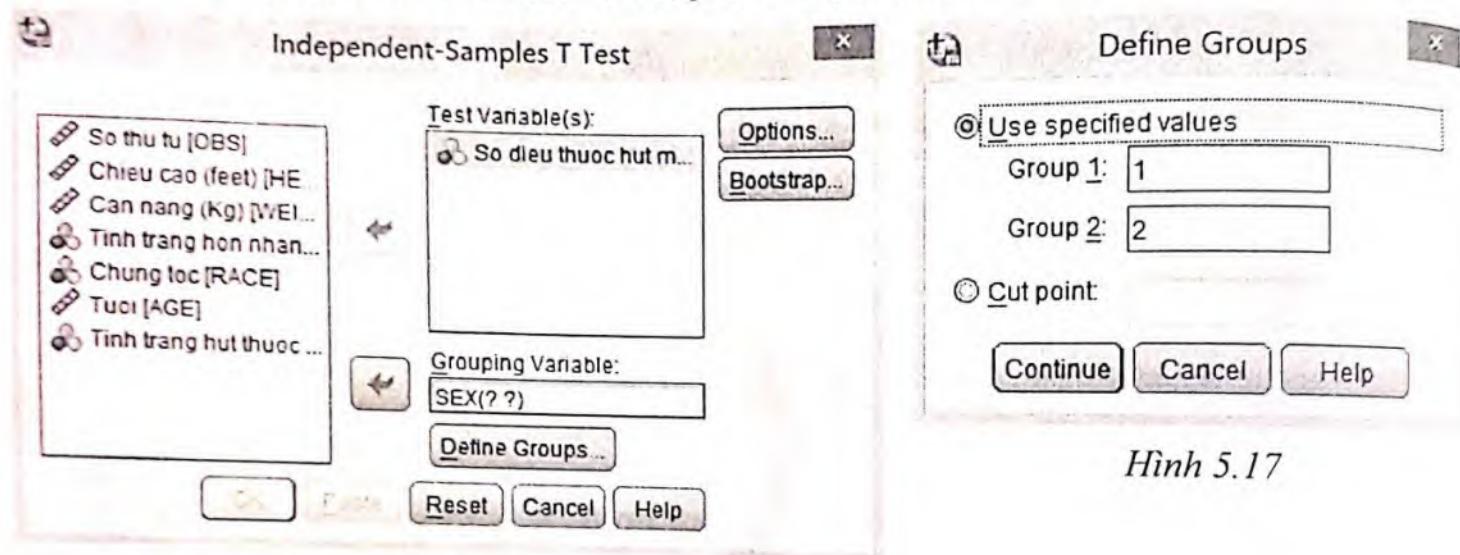
- Giả thiết H₀: trung bình số điếu thuốc hút trên ngày của nam và nữ là như nhau.
- Đối thuyết: trung bình số điếu thuốc hút trên ngày của nam và nữ là khác nhau.

2) Tính tqs bằng lệnh Independent-Samples T Test

- Chọn Analyze → Compare Means → Independent-Samples T Test.

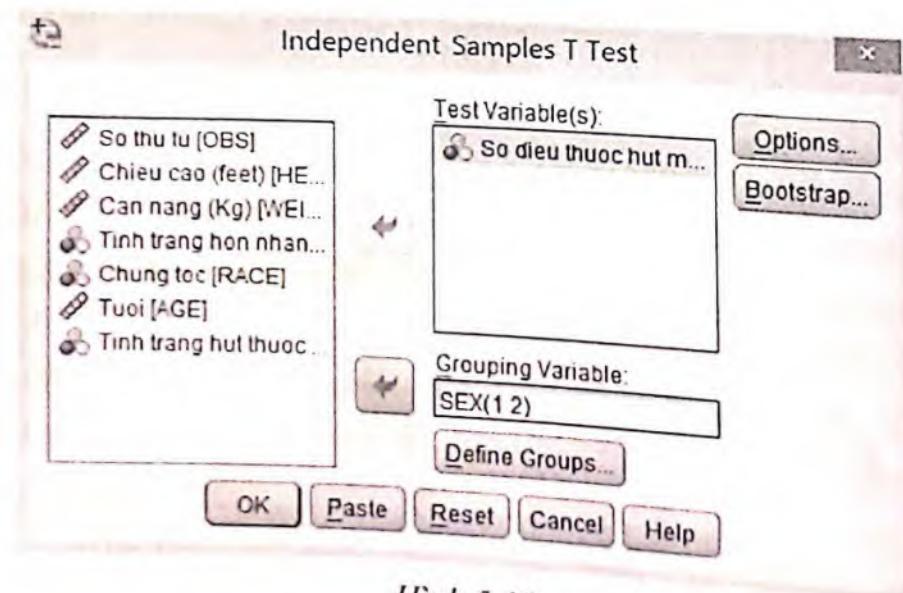
- Chọn các biến Numcigar và Sex đưa vào các khung Test Variable(s) và Grouping Variable như *Hình 5.16*.

- Chọn Define Group... để định nghĩa nhóm trong biến phân nhóm SEX – *Hình 5.17*, nhấn Continue được kết quả như *Hình 5.18* → Nhập OK và quan sát kết quả.



Hình 5.17

Hình 5.16



Hình 5.18

Group Statistics

	Gioi tinh	N	Mean	Std. Deviation	Std. Error Mean
So dieu thuoc tren ngay	Nam	30	18.90	10.138	1.851
	Nu	42	15.74	8.314	1.238

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	
So dieu thuoc hut tren ngay	Equal variances assumed	.322	.572	1.451	70	.151	3.162	2.179	-1.183	7.507
	Equal variances not assumed			1.404	54.636	.166	3.162	2.252	-1.352	7.676

Đọc kết quả:

- Bảng Group Statistics:

- + Số người nam hút thuốc là: 30, số người nữ hút thuốc là: 42
- + Trung bình số điếu thuốc của nam là 18.90; của nữ là 15.74

- Bảng Independent Samples Test: Để nhận xét được hai trung bình này đã khác nhau có ý nghĩa thống kê hay chưa ta phải dựa vào bảng kết quả phân tích Independent Samples Test, đọc các giá trị của t và p như sau:

+ Cột Levene's Test for Equality of Variances: Kiểm định tính bằng nhau (đồng nhất) phương sai của hai nhóm.

. Nếu $\text{Sig.} > 0.05$: 2 mẫu có phương sai bằng nhau: Chọn t ở hàng trên (**Equal variances assumed**).

. Nếu $\text{Sig.} < 0.05$: 2 mẫu có phương sai không bằng nhau: Chọn t ở hàng dưới (**Equal variances not assumed**).

+ Vì $\text{Sig.} = 0.572 > 0.05$ nên hai mẫu có phương sai bằng nhau; Vậy ta đọc t ở hàng **Equal variances assumed** $\rightarrow t = 1.451$ hay $t_{QS} = 1.451$

3) Tính t_α

$$df = (n_1 + n_2 - 2) = 70$$

Tra bảng có: $t_\alpha = 1.9945$ (với khoảng tin cậy = 95%)

4) Kết luận

Vì $t_{QS} < t_\alpha$: Chấp nhận giả thuyết H_0 .

Vậy: Với độ tin cậy 95% thì trung bình số điếu thuốc hút trong một ngày của Nam và nữ là như nhau.

5.2.3. So sánh đồng thời nhiều trung bình

Đây là một bài toán mở rộng của bài toán kiểm định T-test với hai nhóm độc lập, khi số mẫu lớn hơn 2 thì việc so sánh giá trị trung bình của các mẫu đó được gọi là so sánh nhiều trung bình. Phương pháp này giúp chúng ta xác định xem phương sai của các mẫu có bằng nhau hay không hay tồn tại những khác biệt nào đó giữa các mẫu quan sát. Sau đó ta cần chỉ ra cụ thể có những cặp nào có sự khác biệt.

Trong bài toán này chúng ta sử dụng phép phân tích phương sai một nhân tố (One-Way Anova) để so sánh đồng thời nhiều giá trị trung bình.

Việc sử dụng One-Way Anova yêu cầu:

- Các nhóm so sánh phải độc lập và được chọn một cách ngẫu nhiên.
- Các nhóm so sánh phải là một phân phối chuẩn hoặc cỡ mẫu phải đủ lớn ($n > 30$) để có thể xem là tiệm cận chuẩn.

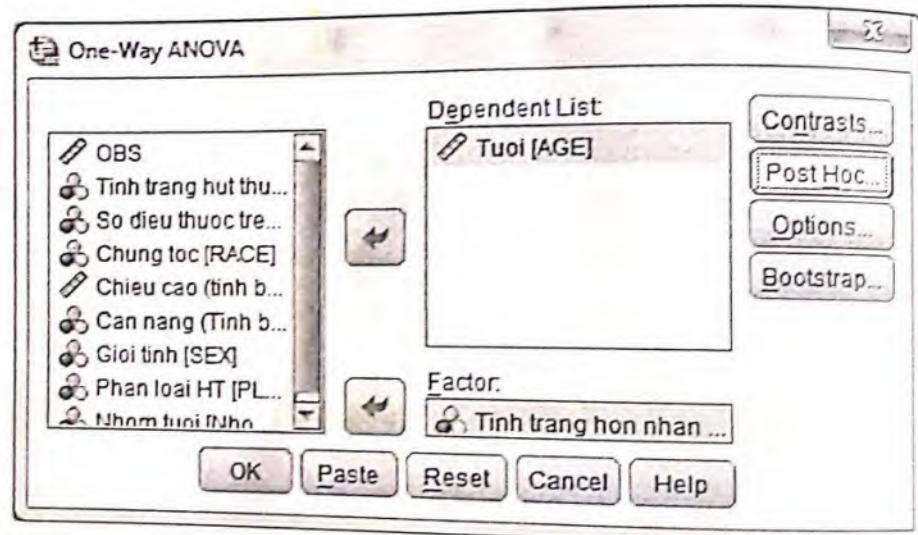
- Phương sai các nhóm phải đồng nhất.

Ví dụ: So sánh Tuổi trung bình của các Tình trạng hôn nhân.

- Giả thuyết H_0 : Tuổi trung bình của các tình trạng hôn nhân là như nhau.
- Giả thuyết H_1 : Tuổi trung bình của các tình trạng hôn nhân khác nhau (Có thể có sự khác biệt về Tuổi trung bình của một số tình trạng hôn nhân).

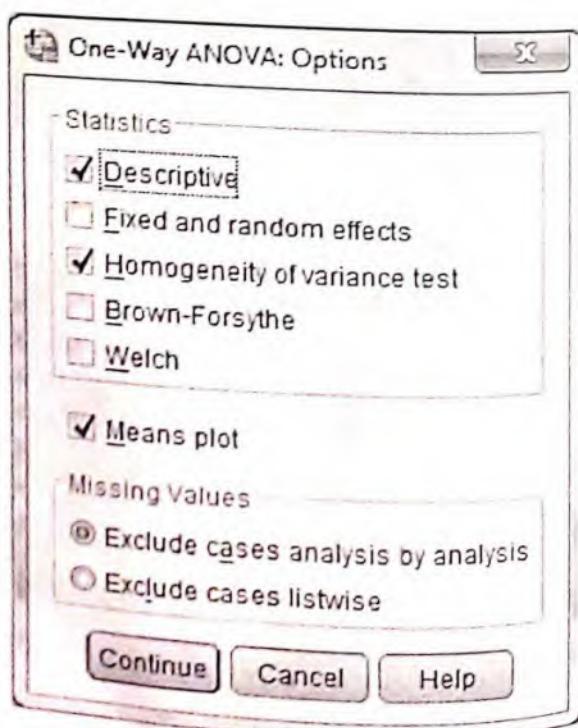
- Chọn Analyze → Compare Means → One-Way ANOVA, xuất hiện hộp thoại One-Way ANOVA – Hình 5.19

- Đưa biến định lượng cần so sánh trung bình vào mục **Dependent List**, biến phân loại đưa vào mục **Factor** – Hình 5.19



Hình 5.19

- Chọn Options... đánh dấu mục **Descriptive** để SPSS đưa ra bảng các tham số thống kê mô tả cho biến định lượng. Đánh dấu mục **Homogeneity of variance test** (Hình 5.20) để kiểm tra sự bằng nhau của phương sai → Nhấp Continue.



- Nhấp OK và quan sát kết quả – Hình 5.21

Hình 5.20

Descriptives

Tuổi	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	146	44.73	15.236	1.261	42.23	47.22	19	83
2	39	41.38	11.231	1.798	37.74	45.03	22	71
3	41	69.95	15.495	2.420	65.06	74.84	9	96
4	5	35.20	10.354	4.630	22.34	48.06	26	52
5	39	28.26	14.493	2.321	23.56	32.95	18	83
Total	270	45.52	18.679	1.137	43.28	47.76	9	96

Test of Homogeneity of Variances

Tuổi

Levene Statistic	df1	df2	Sig.
1.984	4	265	.097

ANOVA

Tuổi

Tổng bình phương
giữa các nhóm

bptb giữa các
nhóm

Phân phối F.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	37386.997	4	9346.749	43.865	.000
Within Groups	56466.410	265	213.081		
Total	93853.407	269			

Hình 5.21

Với các bảng kết quả thống kê Hình 5.21 ta thấy:

- Bảng Descriptives cho biết: Số lượng (N), tuổi trung bình (Mean), độ lệch chuẩn (Std. Deviation), sai số chuẩn (Std. Error) ... của từng tình trạng hôn nhân.

- Bảng Test of Homogeneity of Variances: cho kết quả Sig=0.097 > 0.05 → Phương sai của các nhóm bằng nhau → Có thể dùng kiểm định ANOVA.

- Bảng ANOVA cho biết:

+ Tổng bình phương giữa các nhóm (Sum of Squares - Between Groups) = 37386.997, bậc tự do df = (5 - 1) = 4. Bình phương trung bình giữa các nhóm (Mean Square) = 9346.749;

+ Tổng bình phương toàn bộ mẫu (Sum of Squares - Within Groups) = 9346.749; bậc tự do df = 270 - 5 = 265; Bình phương trung bình toàn bộ mẫu = 56466.410/265 = 213.081.

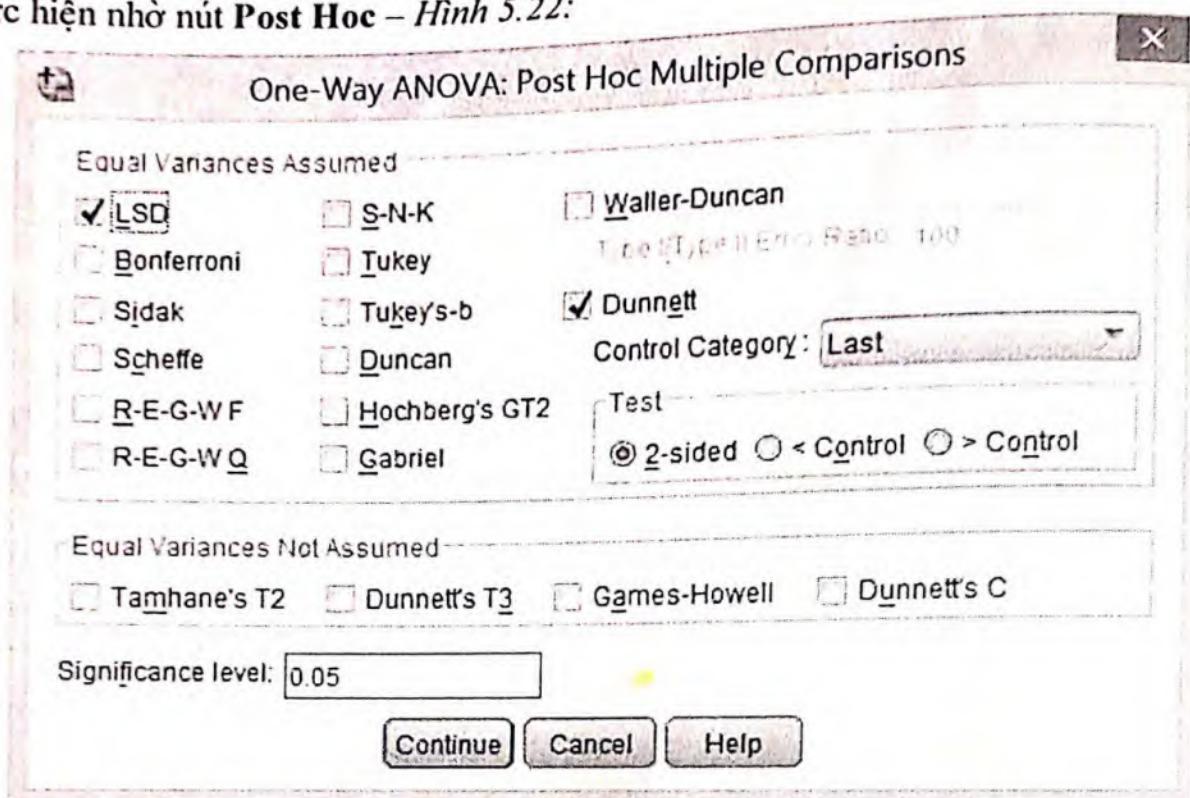
+ Phân phối F (Tỷ số giữa bình phương trung bình giữa các tình trạng hôn nhân/trung bình bình phương trong nội bộ mỗi tình trạng hôn nhân) = 43.865.

Ý nghĩa thống kê (Sig.) < 0.05 → Có sự khác biệt về tuổi trung bình giữa các tình trạng hôn nhân. Vấn đề đặt ra tiếp theo là cần xác định chính xác sự khác biệt xảy ra ở những nhóm nào? Có hai phương pháp tìm ra sự khác biệt đó, đó là kiểm định trước và kiểm định sau.

Bộ môn Tin học – Trường Đại học Y Dược Hải Phòng

- **Kiểm định trước:** nhấp nút **Contrasts**, kiểm định này không bàn đến ở đây và hiện nay ít được sử dụng.

- **Kiểm định sau:** là kiểm định được thực hiện sau khi thực hiện kiểm định ANOVA, nó được thực hiện nhờ nút **Post Hoc – Hình 5.22**:



Hình 5.22

- **LSD:** So sánh tự động từng cặp.
- **Bonferroni:** Tiến hành như LSD nhưng có hiệu chỉnh.
- **Tukey:** Phương pháp này thường được sử dụng, nhất là khi có nhiều nhóm con. Nó dựa trên phân phối *Studentized range distribution*.
- **Dunnett:** So sánh trung bình của các nhóm còn lại với một nhóm điều khiển, ở chế độ mặc định nó là nhóm cuối.
- Nếu phương sai các nhóm khác nhau người ta thường sử dụng *Tamhane's T2*.

Trong ví dụ này ta chọn kiểm định **LSD** để so sánh tự động giữa các cặp và **Dunnett** → Chọn **Continue**.

- Nhấp **OK** và quan sát kết quả *Hình 5.23*

Nhìn vào bảng thống kê (*Hình 5.23*) ta thấy thực sự có sự khác biệt về tuổi trung bình giữa một số cặp tình trạng hôn nhân (các cặp có $Sig. < 0.05$).

Theo **LSD** có sự khác nhau về tuổi giữa các cặp 1-3, 1-5, 2-3, 2-5, 3-1, 3-2, 3-4, 3-5.

Dependent Variable: Tuoi

Multiple Comparisons

	(I) Tình trạng hon nhan	(J) Tình trạng hon nhan	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1	2	3.341	2.631	.205	-1.84	8.52
		3	-25.225*	2.580	.000	-30.31	-20.15
		4	9.526	6.639	.153	-3.55	22.60
		5	16.470*	2.631	.000	11.29	21.65
	2	1	-3.341	2.631	.205	-8.52	1.84
		3	-28.567*	3.265	.000	-35.00	-22.14
		4	6.185	6.934	.373	-7.47	19.84
		5	13.128*	3.306	.000	6.62	19.64
		1	25.225*	2.580	.000	20.15	30.31
	3	2	28.567*	3.265	.000	22.14	35.00
		4	34.751*	6.915	.000	21.14	48.37
		5	41.695*	3.265	.000	35.27	48.12
		1	-9.526	6.639	.153	-22.60	3.55
		2	-6.185	6.934	.373	-19.84	7.47
	4	3	-34.751*	6.915	.000	-48.37	-21.14
		5	6.944	6.934	.318	-6.71	20.60
		1	-16.470*	2.631	.000	-21.65	-11.29
		2	-13.128*	3.306	.000	-19.64	-6.62
		3	-41.695*	3.265	.000	-48.12	-35.27
	5	4	-6.944	6.934	.318	-20.60	6.71
		1	16.470*	2.631	.000	9.99	22.95
		2	13.128*	3.306	.000	4.99	21.26
		3	41.695*	3.265	.000	33.66	49.73
		5	6.944	9.634	.717	-10.12	24.01

*. The mean difference is significant at the 0.05 level.

b. Dunnett t-tests treat one group as a control, and compare all other groups against it.

Hình 5.23

5.2.4. So sánh ghép cặp

Mục đích của việc sử dụng nghiên cứu ghép cặp là để loại bỏ tối đa những nguồn biến thiên, sai khác của những biến số mà chúng ta không quan tâm bằng cách ghép thành các cặp đối tượng càng giống nhau về nhiều mặt (biến số) càng tốt. Trong trường hợp này ta vẫn sử dụng kiểm định t nhưng cho dãy dữ liệu theo từng cặp.

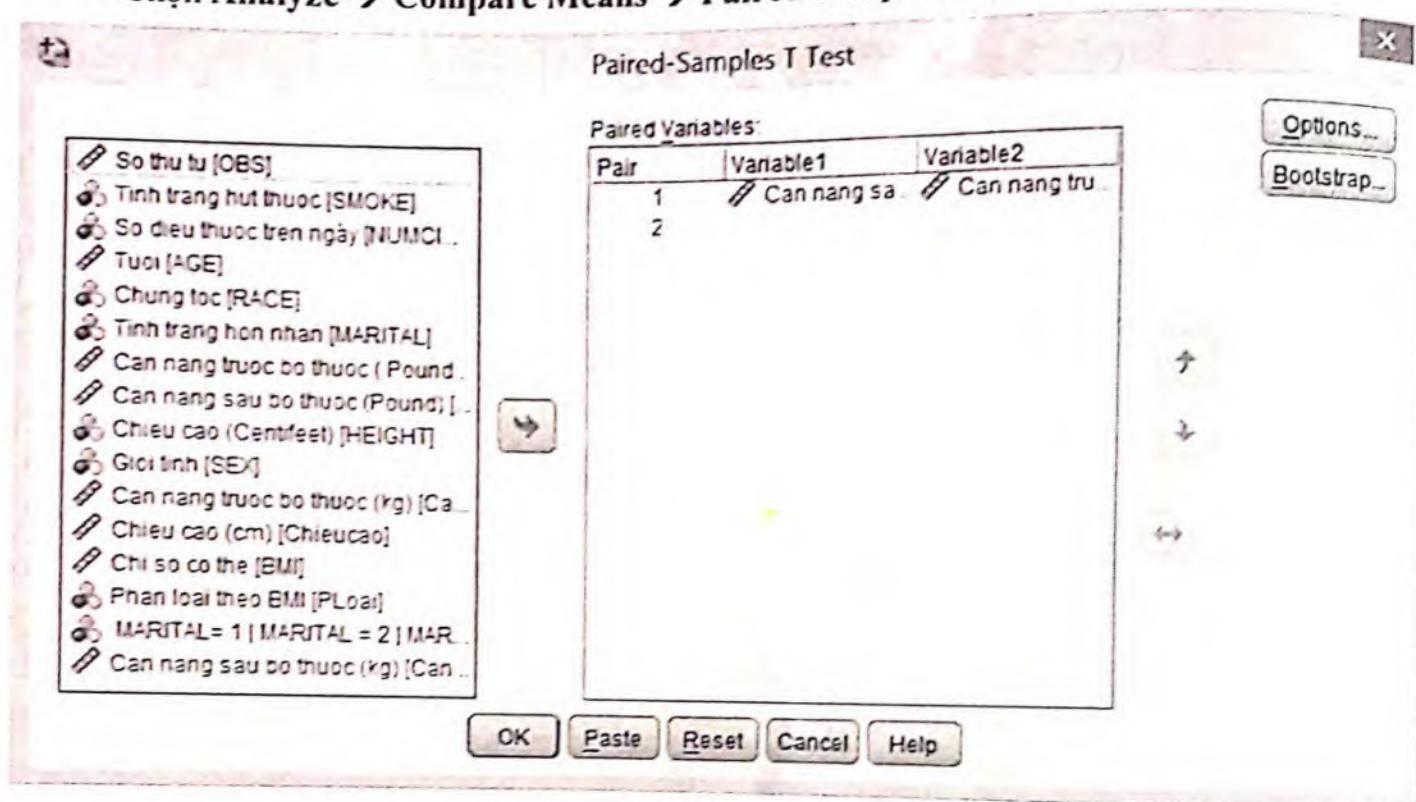
Kiểm định t ghép cặp thường được sử dụng để kiểm định xem trung bình sự khác biệt giữa một cặp biến số đo lường trên mỗi cá thể có bằng 0 hay không.

Trong trường hợp này các biến không độc lập mà có sự phụ thuộc nhau, không thể tiến hành như với hai mẫu độc lập.

Để thực hiện điều này người ta dùng biến $D = X - Y$ để đánh giá hiệu số chênh lệch rồi kiểm định giả thiết $M(D) = 0$; đối với $M(D) \neq 0$. Khi đó cần tính: $T = \frac{\bar{D}}{S_D} \sqrt{n}$. Đại lượng T tuân theo phân phối Student n-1 bậc tự do.

Ví dụ: So sánh trọng lượng trước và sau khi bỏ thuốc của những người hút thuốc lá.

- Chọn Analyze → Compare Means → Paired Samples T-Test



Hình 5.24

- Khai báo:

+ **Variable1** là biến cân nặng sau khi bỏ thuốc lá.

+ **Variable2** là biến cân nặng trước khi bỏ thuốc lá – xem Hình 5.24.

- Nhập OK và quan sát các bảng kết quả Hình 5.25.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Cân nặng sau bỏ thuốc (Kg)	70.7560	73	9.94535	1.16420
	Cân nặng trước bỏ thuốc (kg)	67.2164	73	12.23248	1.43170

		Paired Samples Test								
		Paired Differences								
Pair 1	Can nang sau bo thuoc (Kg) - Can nang truoc bo thuoc (kg)	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)	
					Lower	Upper				
		3.35959	7.15379	.83729	1.69049	5.02869	4.012	72	.000	

Paired Samples Correlations										
				N	Correlation			Sig.		
Pair 1 Can nang sau bo thuoc (Kg) & Can nang truoc bo thuoc (kg)				73	.811			.000		

Hình 5.25

Trong bảng trên ta thấy:

- Hiệu số chênh lệch trung bình $D = 3.35959$, Độ lệch chuẩn của $D = 7.15379$;
- Khoảng tin cậy 95% của D là $(1.69049 - 5.02869)$.
- Giá trị $t = 4.012$, bậc tự do $df = 72$,
- Mức ý nghĩa $p = \text{Sig. (2-tailed)} = 0.000 < 0.05$.

Kết luận: Có sự khác biệt giữa trọng lượng trước bỏ thuốc và sau bỏ thuốc.

Hình số 5.25. Kết quả $D =$

Bài 6. TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH

Mục tiêu:

- Trình bày được cơ sở lý thuyết và ý nghĩa của hệ số tương quan tuyến tính.
- Trình bày, tính và phiên giải được hệ số tương quan.
- Trình bày và vận dụng được thuật toán kiểm định giả thuyết cho hệ số tương quan.
- Tính và phiên giải được ý nghĩa các đại lượng biểu diễn phương trình hồi quy tuyến tính.

6.1. Tương quan tuyến tính

Phân tích tương quan là việc phân tích để đo lường độ lớn của mối quan hệ giữa các biến số với nhau. Khi chúng ta đo lường mối tương quan của một bộ số liệu chúng ta quan tâm đến mức độ của mối liên hệ giữa các biến với nhau. Trong nội dung này chúng ta sẽ xem xét mối tương quan của hai biến định lượng.

6.1.1. Cơ sở lý thuyết

Nếu X và Y là 2 biến định tính thì tương quan giữa chúng biểu thị qua tính toán nguy cơ tương đối (RR) và tỷ suất chênh (OR). Nếu X và Y là 2 biến định lượng thì tương quan giữa chúng biểu thị qua *hệ số tương quan* (r) và *hồi quy tuyến tính*.

$$\text{Hệ số tương quan tuyến tính: } r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{\sqrt{\left\{ n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 \right\} \left\{ n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2 \right\}}}$$

Hệ số tương quan đánh giá mức độ tương quan giữa hai đại lượng X và Y với mức ý nghĩa p.

Giá trị $|r| \leq 1$; nếu $|r|$ càng gần 1 thì tương quan càng chặt chẽ, $r > 0$ thì mối tương quan là đồng biến, $r < 0$ thì mối tương quan là nghịch biến, $r = 0$ là không có mối tương quan.

Quy ước:

- | | | |
|------|---------------------------|--|
| ○ | $0 \leq r < 0.3$: | x và y không tương quan tuyến tính |
| ○ | $0.3 \leq r \leq 0.6$: | x và y có tương quan tuyến tính ↗ |
| đánh | $0.6 < r \leq 1$: | x và y có tương quan tuyến tính chặt chẽ |

* Kiểm định giả thuyết cho giá trị r

- Sử dụng test thống kê $t = r \sqrt{\frac{n-2}{1-r^2}}$ (1) để đánh giá hai biến có tương quan tuyến tính hay không.

- Đại lượng t có phân phối *t-Student* với $n-2$ bậc tự do (nếu X và Y là độc lập và đều là các phân phối chuẩn). Tra bảng phân vị của phân phối Student $n-2$ bậc tự do có thể tìm được giá trị

Kiểm định giả thuyết cho giá trị r hay còn gọi là kiểm định mối tương quan tuyến tính giữa hai biến định lượng; Để kiểm định giả thuyết cho giá trị r , thực hiện theo các bước sau:

1) Xây dựng giả thuyết:

- Giả thuyết H_0 : X và Y không có tương quan tuyến tính.
- Đổi thuyết H_1 : X và Y có tương quan tuyến tính.

2) Tính t_{QS} , t_α :

- t_{QS} được tính theo công thức (1) ở trên;
- t_α được tra trong bảng t – Student với bậc tự do = $n - 2$ và độ tin cậy α .

3) So sánh t_{QS} và t_α và rút ra kết luận

- Nếu $t_{QS} < t_\alpha$: Chấp nhận giả thuyết H_0 . (Với độ tin cậy 95% có $Sig. > 0.05$)
- Nếu $t_{QS} > t_\alpha$: bác bỏ giả thuyết H_0 . (Với độ tin cậy 95% có $Sig. < 0.05$)

Chú ý: Yếu tố cần quan tâm đầu tiên là giá trị sig . Giá trị sig nhỏ hơn 0.05 thì hệ số tương quan r mới có ý nghĩa thống kê, giá trị sig lớn hơn 0.05 nghĩa là r có lớn nhỏ thế nào cũng không liên quan gì bởi vì nó không có ý nghĩa, hay nói cách khác không có tương quan giữa 2 biến.

Để xác định hệ số tương quan tuyến tính của hai biến định lượng, có thể sử dụng lệnh **Bivariate** hoặc **Linear**.

6.1.2. Tính htent h²e số tương quan tuyến tí

6.1.2.1. Lệnh Bivariate

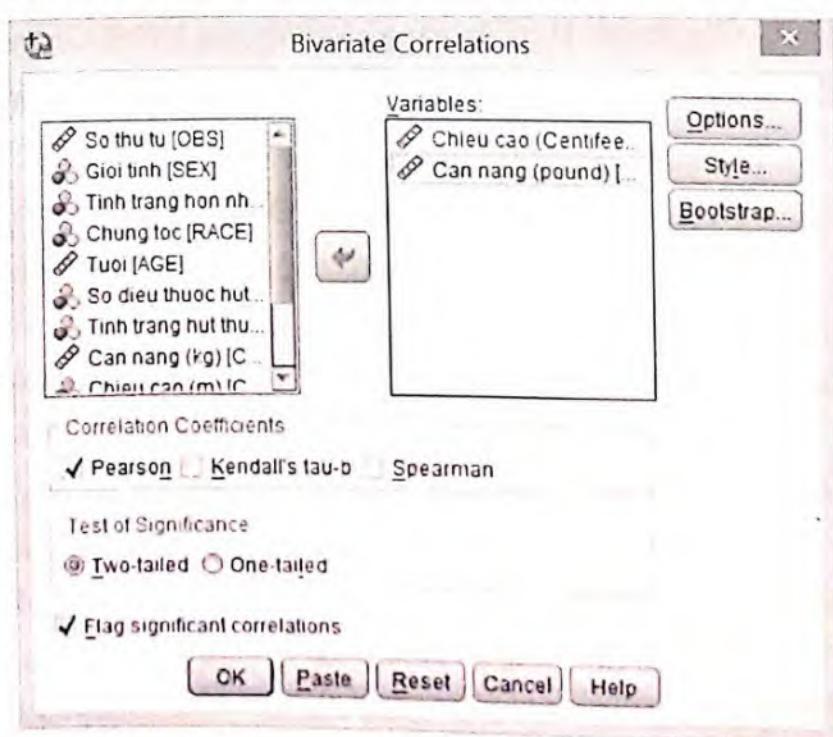
Để tính hệ số tương quan tuyến tính của hai biến định lượng Cân nặng và Chiều cao chúng ta có thể sử dụng lệnh **Bivariate** theo ví dụ sau đây:

Ví dụ: Tính hệ số tương quan tuyến tính của Cân nặng và Chiều cao.

- Chọn **Analyze → Correlate → Bivariate**, đưa vào các biến cần đánh giá – *Hình 6.1*

- Trong *Hình 6.1* có 3 dạng công thức tính hệ số tương quan tuyến tính thường sử dụng: Hệ số theo Pearson là hệ số đã dẫn ở trên; các hệ số Kendall's tau-b, Spearman là các hệ số tương quan hay dùng trong kiểm định phi tham số.

- Trong mục *Test of Significance* có hai lựa chọn *Two-tailed* (kiểm định hai phía) hoặc *One-tailed* (kiểm định một phía). Thông thường ta chọn tiêu chuẩn kiểm định hai phía rồi.



Hình 6.1

	Mean	Std. Deviation	N
Cân nặng (kg)	67.2050	11.79136	270
Chiều cao (m)	1.5573	.10126	270

Correlations

		Chiều cao (m)	Cân nặng (kg)
Chiều cao (m)	Pearson Correlation	1	.544**
	Sig. (2-tailed)		.000
	N	270	270
Cân nặng (kg)	Pearson Correlation	.544**	1
	Sig. (2-tailed)	.000	
	N	270	270

**. Correlation is significant at the 0.01 level (2-tailed).

Hình 6.2

Theo kết quả Hình 6.2, hệ số tương quan giữa hai biến: $r = 0.544$.

Ngoài ra, trong mục giá trị *Sig. (2-tailed)* $= 0.000 < 0.001$ cho biết hai biến Cân nặng và Chiều cao có tương quan tuyến tính ở mức trung bình.

6.1.2.2. Lệnh Linear.

Để kiểm định mối tương quan giữa Cân nặng và Chiều cao chúng ta có thể sử dụng lệnh **Linear** theo ví dụ sau đây:

Ví dụ: Kiểm định mối tương quan giữa Cân nặng và Chiều cao. Thực hiện theo các bước:

1) Xây dựng giả thuyết.

- Giả thuyết H_0 : Cân nặng và chiều cao không có tương quan tuyến tính.
- Đối thuyết H_1 : Cân nặng và chiều cao có tương quan tuyến tính.

2) Tính giá trị của t_{qs} , t_α .

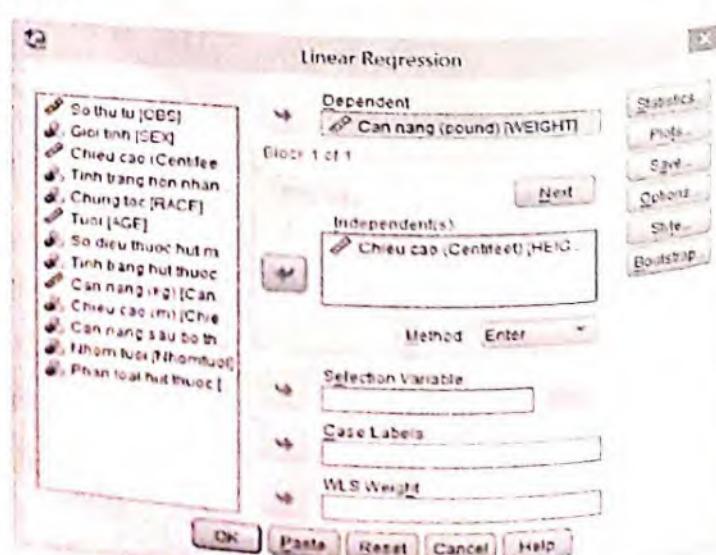
- Chọn **Analyze → Regression → Linear**

- Chọn các biến cần kiểm định tương quan tuyến tính từ khung bên trái của cửa sổ **Linear Regression**:

+ Mục **Dependent**: đưa vào biến *phụ thuộc*.

+ Mục **Independent(s)**: đưa vào các biến *độc lập* – Hình 6.3

- Chọn **OK** và quan sát kết quả –
Hình 6.4



Hình 6.3

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	Chieu cao (m) ^b		Enter

a. Dependent Variable: Can nang (kg)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.544	.295	.293	9.91552

a. Predictors: (Constant), Chieu cao (m)

ANOVA^a

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	11051.641	1	11051.641	112.408	.000 ^b
	Residual	26349.090	268	98.317		
	Total	37400.731	269			

a. Dependent Variable: Can nang (kg)

b. Predictors: (Constant), Chieu cao (m)

Coefficients^a

Model		Unstandardized Coefficients		Beta	t	Sig.
		B	Std. Error			
1	(Constant)	b -31.372	9.317		-3.367	.001
	Chieu cao (m)	a 63.300	5.970	.544	10.602	.000

a. Dependent Variable: Can nang (kg)

Hình 6.4

đ/q8.

Từ các bảng kết quả Hình 6.4 ta có:

- $t = 10.602$ đây chính là giá trị của tqs;
- Bậc tự do: $df = 270 - 2 = 268$;
- Khoảng tin cậy 95% $\rightarrow \alpha = 0.05$; Tra bảng Student (t) được $t_{\alpha} = 1.96$
- Hệ số tương quan giữa hai biến: $r = 0.544$, $R^2 = 0.295$.
- Mức ý nghĩa (p): $Sig. = 0.000 < 0.05$

3) Kết luận

Ta thấy: $t_{qs} > t_{\alpha}$ do đó bác bỏ giả thuyết H_0 . Vậy cân nặng và chiều cao có tương quan tuyến tính. $r = 0.544$ nên Cân nặng và Chiều cao là hai biến có tương quan tuyến tính ở mức trung bình.

6.2. Hồi quy tuyến tính

Phân tích hồi quy rất tiện dụng trong việc khẳng định mối liên hệ giữa 2 biến số, mục tiêu cuối cùng của phương pháp này là **dự đoán hoặc ước lượng** giá trị của một biến số từ các giá trị của một hay nhiều biến số khác.

Trong mô hình hồi quy tuyến tính cơ bản bao giờ cũng liên quan đến hai biến X và Y. Trong đó, X là **biến độc lập**, Y là **biến phụ thuộc**. Các giá trị của X bao giờ cũng được kiểm soát bởi người nghiên cứu do đó các giá trị của X luôn được người nghiên cứu lựa chọn; trên cơ sở các giá trị đã được chọn của X thì sẽ xác định được các giá trị của Y.

Trong trường hợp X, Y có tương quan tuyến tính, chúng ta có thể biểu diễn mối tương quan đó bằng phương trình đường thẳng có dạng $Y = aX + b$. Đường biểu diễn Y theo X gọi là **đường hồi quy**.

Sử dụng phương pháp bình phương bé nhất chúng ta có các công thức sau đây:

$$Y = aX + b$$

$$\text{trong đó: } a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}; \quad b = \bar{Y} - a \bar{X}$$

Ví dụ: Viết phương trình hồi quy tuyến tính biểu diễn mối liên hệ giữa cân nặng và chiều cao.

Trong các bước tiến hành kiểm định mối tương quan tuyến tính giữa 2 biến cân nặng và chiều cao trong ví dụ của mục 6.1.3 ở trên ngoài việc xác định được giá trị của hệ số tương quan, chúng ta cũng xác định được các giá trị a và b của phương trình hồi quy tuyến tính trong bảng **Coefficients - Hình 6.4**.

Cụ thể: hệ số tương quan tuyến tính $r = 0.544$, mức ý nghĩa $p < 0.0001$

Biến phụ thuộc là Cân nặng (Y), biến độc lập là Chiều cao (X) ;

Hằng số (constant) $b = -31.372$, hệ số $a = 63.300$;

Ngoài ra, $b = -31.372$ còn được hiểu là **số chặn (Intercept)**, $a = 63.300$ là **độ dốc** của đường hồi quy.

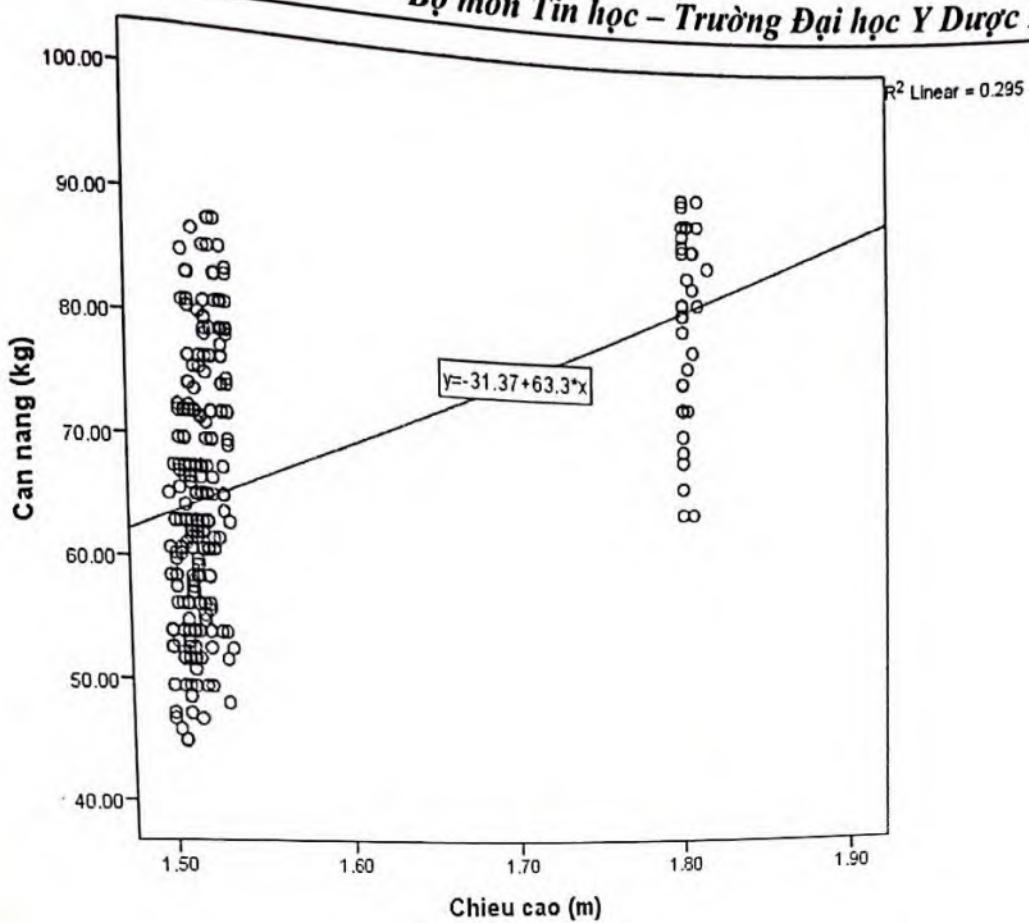
Ta có phương trình hồi quy: $Y = aX + b$

$$Y = 63.300 * X - 31.372$$

$$\text{Hay: } Cân\ nặng = 63.300 * Chiều\ cao - 31.372$$

- Cũng trong bảng này còn thể hiện hệ số Beta là hệ số hồi quy sau khi chuẩn hóa. Nó chính là **độ dốc** của đường thẳng hồi quy khi cả X và Y đã được chuẩn hóa lại theo thang đo chuẩn.

- Vẽ biểu đồ **Scatter/Dot** biểu diễn mối quan hệ giữa chiều cao và cân nặng – **Hình 6.5**.



Hình 6.5

Lưu ý: Khi thay đổi vị trí khai báo biến phụ thuộc và biến độc lập thì hệ số tương quan tuyến tính (r) không thay đổi nhưng các hệ số, hằng số sẽ có giá trị khác, dẫn đến ý nghĩa của phương trình hồi quy cũng thay đổi. Do đó, khi áp dụng các bước của thuật toán kiểm định giả thuyết cho giá trị r và viết phương trình hồi quy cần xác định rõ biến độc lập và biến phụ thuộc.

TÀI LIỆU THAM KHẢO

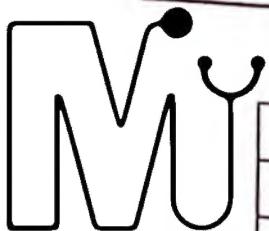
- [1] Lê Cự Linh, *Thống kê y tế công cộng – Phần thống kê cơ bản*, Nhà xuất bản Y học Hà Nội, 2009.
- [2] Phạm Việt Cường, *Thống kê y tế công cộng – Phần phân tích số liệu*, Nhà xuất bản Y học Hà Nội, 2009.
- [3] Nguyễn Ngọc Rạng, *Thiết kế nghiên cứu và thống kê y học*, Nhà xuất bản Y học TP Hồ Chí Minh, 2012.
- [4] Nguyễn Minh Tuấn, Hà Trọng Quang, *Giáo trình Xử lý dữ liệu nghiên cứu với SPSS for Windows*, Trường Đại học Công nghiệp TP Hồ Chí Minh, 2008.
- [5] Nguyễn Văn Tuấn, *Lâm sàng thống kê - Ý nghĩa của odds ratio và relative risk*, Chương trình huấn luyện y khoa - YKHOA.NET Training www.ykhoa.net/baigiang/lamsangthongke/lstk14_oddratio.pdf
- [6] Nguyễn Văn Tuấn - Nguyễn Đình Nguyên, *Nguy cơ tương đối (RR) và Odds ratio (OR)*. <https://tqhien.files.wordpress.com/2011/10/phan-biet-rr-va-or.ppt>
- [7] Nghiêm Văn Thiệp, Lê Hồng Vượng, *Giáo trình Tin học đại cương*, NXB LĐXH, 2006.
- [8] Phạm Thế Quê, *Công nghệ máy tính*, NXB Thông tin & Truyền Thông, 2009.
- [9] Lê Thuận, Thanh Tâm, Quang Huy, *Microsoft Office 2010*, NXB Thông tin & Truyền Thông, 2010.
- [10] Microsoft, *Giáo trình hướng dẫn sử dụng Word 2010*, 2010.
- [11] Trí Việt, Hà Thành, *Tin Học Văn Phòng 2010 - Tự Học Microsoft Word 2010*, NXB: Văn hoá Thông tin, 2010.
- [12] IIG Việt Nam, *Microsoft Office Word 2010*, NXB Tổng Hợp TP.HCM, 2010.
- [13] IIG Việt Nam, *Sử Dụng Windows 7 Và Microsoft Office 2010*, NXB Tổng Hợp TP.HCM, 2010.
- [14] <https://www.microsoft.com>



PHỤ LỤC

Bảng Phân vị mức α của phân phối χ^2 với bậc tự do tương ứng

Bậc tự do	$\alpha=0.9950$	$\alpha=0.9900$	$\alpha=0.9750$	$\alpha=0.9500$	$\alpha=0.0500$	$\alpha=0.0250$	$\alpha=0.0100$	$\alpha=0.0050$
1	0.0000	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2103	10.5966
3	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449	12.8382
4	0.2070	0.2971	0.4844	0.7107	9.4877	11.1433	13.2767	14.8603
5	0.4117	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119	18.5476
7	0.9893	1.2390	1.6899	2.1673	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	15.5073	17.5345	20.0902	21.9550
9	1.7349	2.0879	2.7004	3.3251	16.9190	19.0228	21.6660	23.5894
10	2.1559	2.5582	3.2470	3.9403	18.3070	20.4832	23.2093	25.1882
11	2.6032	3.0535	3.8157	4.5748	19.6751	21.9200	24.7250	26.7568
12	3.0738	3.5706	4.4038	5.2260	21.0261	23.3367	26.2170	28.2995
13	3.5650	4.1069	5.0088	5.8919	22.3620	24.7356	27.6882	29.8195
14	4.0747	4.6604	5.6287	6.5706	23.6848	26.1189	29.1412	31.3193
15	4.6009	5.2293	6.2621	7.2609	24.9958	27.4884	30.5779	32.8013
16	5.1422	5.8122	6.9077	7.9616	26.2962	28.8454	31.9999	34.2672
17	5.6972	6.4078	7.5642	8.6718	27.5871	30.1910	33.4087	35.7185
18	6.2648	7.0149	8.2307	9.3905	28.8693	31.5264	34.8053	37.1565
19	6.8440	7.6327	8.9065	10.1170	30.1435	32.8523	36.1909	38.5823
20	7.4338	8.2604	9.5908	10.8508	31.4104	34.1696	37.5662	39.9968
21	8.0337	8.8972	10.2829	11.5913	32.6706	35.4789	38.9322	41.4011
22	8.6427	9.5425	10.9823	12.3380	33.9244	36.7807	40.2894	42.7957
23	9.2604	10.1957	11.6886	13.0905	35.1725	38.0756	41.6384	44.1813
24	9.8862	10.8564	12.4012	13.8484	36.4150	39.3641	42.9798	45.5585
25	10.5197	11.5240	13.1197	14.6114	37.6525	40.6465	44.3141	46.9279
26	11.1602	12.1981	13.8439	15.3792	38.8851	41.9232	45.6417	48.2899
27	11.8076	12.8785	14.5734	16.1514	40.1133	43.1945	46.9629	49.6449
28	12.4613	13.5647	15.3079	16.9279	41.3371	44.4608	48.2782	50.9934
29	13.1211	14.2565	16.0471	17.7084	42.5570	45.7223	49.5879	52.3356
30	13.7867	14.9535	16.7908	18.4927	43.7730	46.9792	50.8922	53.6720
40	20.7065	22.1643	24.4330	26.5093	55.7585	59.3417	63.6907	66.7660
50	27.9907	29.7067	32.3574	34.7643	67.5048	71.4202	76.1539	79.4900
60	35.5345	37.4849	40.4817	43.1880	79.0819	83.2977	88.3794	91.9517
70	43.2752	45.4417	48.7576	51.7393	90.5312	95.0232	100.4252	104.2149
80	51.1719	53.5401	57.1532	60.3915	101.8795	106.6286	112.3288	116.3211
90	59.1963	61.7541	65.6466	69.1260	113.1453	118.1359	124.1163	128.2989



Bảng Phân vị mức α của phân phối T - Student với bậc tự do tương ứng

Bậc tự do	$\alpha=0.100$	$\alpha=0.050$	$\alpha=0.025$	$\alpha=0.010$	$\alpha=0.005$	$\alpha=0.001$
1	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145
4	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076
7	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853
8	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247
12	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296
13	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520
14	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874
15	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328
16	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852
40	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317
120	1.2886	1.6577	1.9799	2.3578	2.6174	3.1595
∞	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902