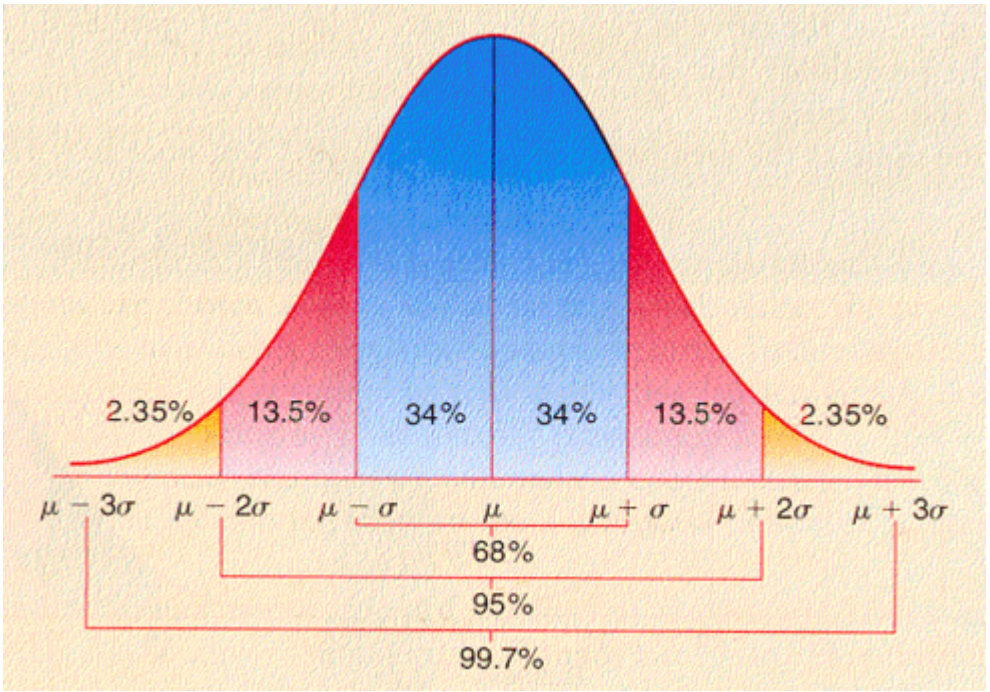
MÃ ĐÌNH TRÊN

BÀI TẬP THỐNG KÊ XÃ HỘI HỌC

VÀ HƯỚNG DẪN THỰC HÀNH TRÊN SPSS



HÀ NỘI
08 - 2015

Mục lục

Chương 1

Làm quen với SPSS

1.1. Về SPSS

SPSS (viết tắt của Statistical Package for the Social Sciences) là một chương trình máy tính phục vụ công tác thống kê. Thế hệ đầu tiên của SPSS được đưa ra từ năm 1968. Thế hệ mới nhất (tính tới năm 2014) là phiên bản 22 cho các hệ điều hành Microsoft Windows, Mac, và Linux. Năm 2009, công ty PASW sở hữu phần mềm này đã được IBM mua lại với giá 1,2 tỷ đô la và tên hiện tại của phần mềm SPSS là "IBM SPSS Statistics".

SPSS là một hệ thống phần mềm thống kê toàn diện được thiết kế để thực hiện tất cả các bước trong các phân tích thống kê từ những thông kê mô tả đến thống kê suy diễn. SPSS cung cấp một giao diện thân thiện giữa người và máy cho phép sử dụng các Menu thả xuống để chọn các lệnh thực hiện.

SPSS được các nhà nghiên cứu sử dụng rộng rãi cho các nghiên cứu trong các lĩnh vực: điều tra xã hội học, tâm lý học, tội phạm học, nghiên cứu kinh tế, nghiên cứu trong y sinh ... Việc sử dụng SPSS làm công cụ giảng dạy và nghiên cứu ở các trường đại học cũng đang dần trở nên phổ biến.

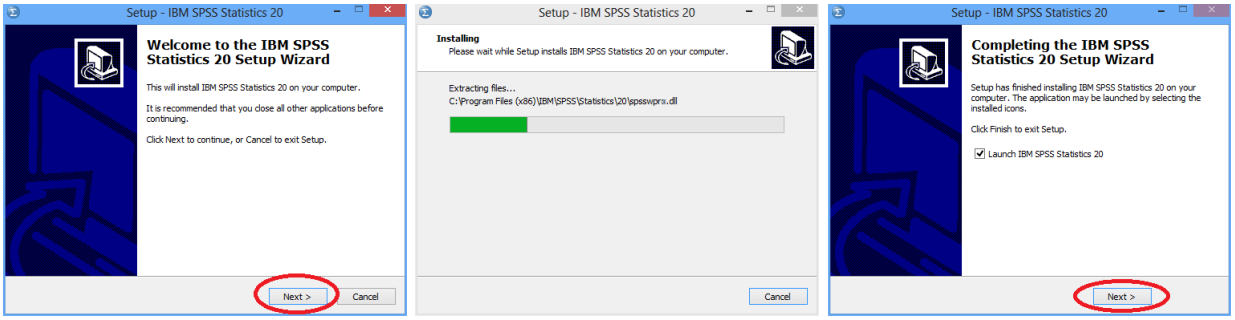
Tài liệu này sử dụng SPSS phiên bản 20 và nó không phải như một tài liệu hướng dẫn sử dụng SPSS một cách toàn diện mà nhằm hướng dẫn sinh viên bước đầu sử dụng SPSS thực hiện những bài toán ăn khớp với nội dung kiến thức của môn thống kê xã hội học.

1.2. Cài đặt và tùy chỉnh đầu tiên

Để đảm bảo việc cài đặt được đúng như hướng dẫn sau, khi cài đặt ta để file cài đặt ngoài màn hình Destop. Thực hiện theo các bước sau:

1. Kích đúp vào file cài đặt. Một hộp thoại hiện ra với câu hỏi "**Do you want to allow the following ...** ". Ta chọn **Yes**.

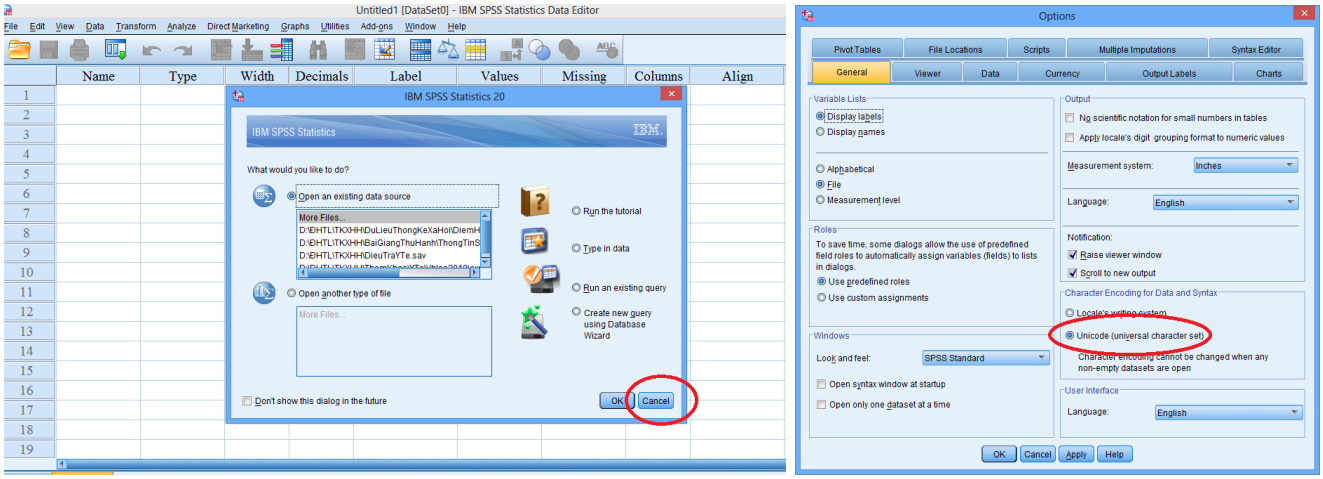
2. Chọn **Next** trong hộp thoại hiện ra, và đợi quá trình cài đặt hoàn thành rồi nhấn **Next** như hình 1.1.



Hình 1.1: Cài đặt SPSS

3. Giao diện khởi động của SPSS sẽ hiện ra như phần bên trái của hình 1.2. Bất cứ khi nào ta khởi động SPSS cũng cho ta một giao diện như vậy.

   Hộp thoại con đầu tiên cho ta lựa chọn mở các file có sẵn. Chúng ta có cách khác để làm việc này và sẽ được đề cập ngay trong các phần dưới đây. Ta nhấn nút **Cancel**.



Hình 1.2: Giao diện đầu tiên và cách tùy chỉnh để sử dụng bộ mã Unicode

4. Trên cửa sổ chính ta chọn **Edit** $\longrightarrow$ **Options** và tích vào lựa chọn khoanh trong phần bên phải của hình 5.4. Tùy chỉnh này cho phép người dùng nhập văn bản Tiếng Việt trên SPSS cũng như mở những file có chứa yếu tố Tiếng Việt. Lưu ý rằng, ta chỉ làm thao tác này một lần, trong những lần khởi động SPSS sau ta không cần điều chỉnh lại nữa.

1.3. Cửa sổ làm việc chính

Mục này nhằm giới thiệu ngắn gọn về 2 cửa sổ làm việc chính cho người mới bắt đầu. Hai cửa sổ làm việc này sẽ hiện ra bất cứ khi nào ta khởi động và thực hiện phân tích trên SPSS: DataSet và Output.

Output cho ta báo cáo hoặc kết quả của những lệnh mà ta thực hiện. Ta có thể điều chỉnh hiển thị của nó và có thể copy sang văn bản word để trình bày. Tuy nhiên ở đây ta không đi sâu vào phần trình bày này mà ta cốt chỉ dùng Output để đọc kết quả. Sau đây, ta tập trung vào cửa sổ DataSet.

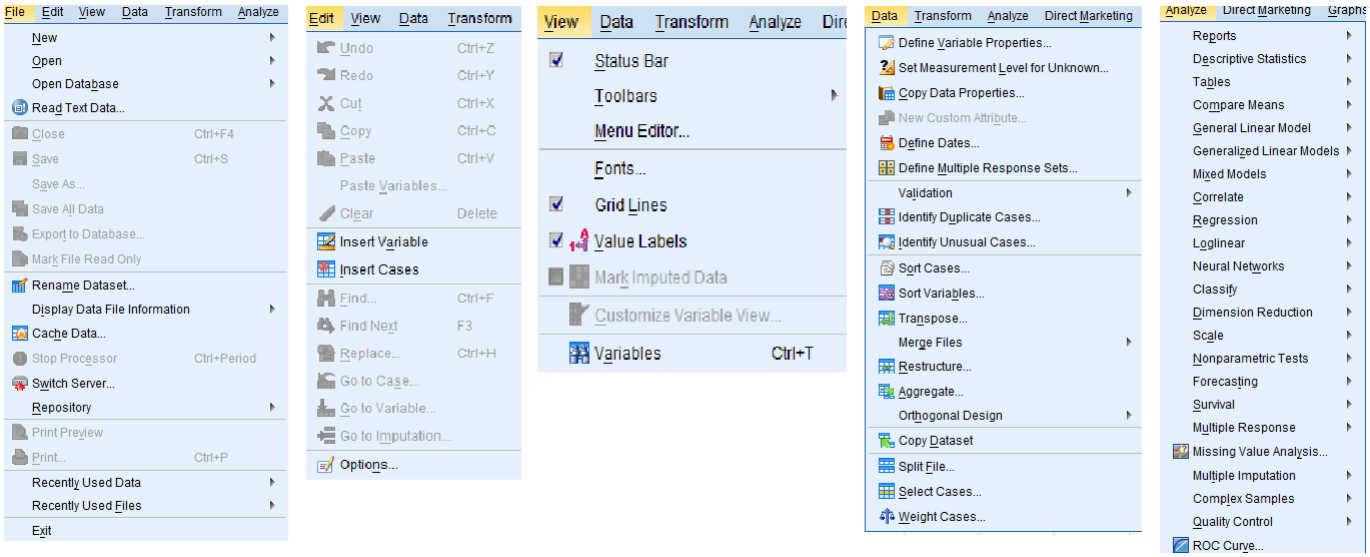
Cửa sổ dữ liệu **DataSet** có hai phần:

- Data View: dùng để nhập và xem dữ liệu đã nhập.

- Variable View: dùng để khai báo biến.

1.3.1. Data View

Mục này giới thiệu chức năng chính của những nút trên thanh Menu là:



Hình 1.3: Hình ghép các lựa chọn của File, Edit, View, Data, Analyze

- **File:** Giúp ta khởi tạo file mới (New), mở file sẵn có (Open), lưu file (Save As), in ...
- **View:** Giúp điều chỉnh hiển thị của thanh công cụ, đổi font (các loại font và cỡ chữ), hiển thị khung dòng hay không (Grid Line), điều chỉnh hiển thị giá trị các biến là nhãn hay giá trị nhập vào (Value Labels), ...
- **Data:** Giúp lựa chọn phục vụ việc biên tập dữ liệu trong đó ta chú ý đến một số lựa chọn hay dùng: sắp xếp thứ tự các quan sát theo trình tự (Sort Cases), sắp xếp các biến theo thứ tự (Sort Variables), nối hai file với nhau (Merge File), tách file (Split File), lựa chọn quan sát thỏa mãn điều kiện mong muốn (Select Cases), ...
- **Transform:** Giúp tính toán (Compute Variable), mã hóa số liệu (Record into ...), thay thế giá trị trống (Replace Missing Values), ...
- **Analyze:** thực hiện các thủ tục thống kê như: tóm tắt dữ liệu bằng bảng tần số (Report), bằng đại lượng thống kê mô tả (Desciptive Statistics), kiểm định về trung bình (Compare Means), kiểm định phi tham số (Nonparametric Test), ...
- **Graphs:** Tạo các biểu đồ và đồ thị.
- Các nút lựa chọn còn lại khi cần sinh viên tự tìm hiểu thêm.

1.3.2. Variable View

Từ cửa sổ **DataSet** nhấp chuột vào **Variable View**, đây là phần mà ta sẽ làm việc đầu tiên khi nhập vào một dữ liệu mới. Chúng ta sẽ khai báo cho các thuộc tính của một biến ở phần này.
Mỗi một dòng dành cho khai báo một biến, bao gồm:

- **Name:** là tên của biến. Lưu ý: không có kí tự đặc biệt và không bắt đầu bởi số, không kết thúc bởi ".".
- **Type:** kiểu của biến (số, kí tự, ngày tháng, ...)
- **Width:** độ rộng tối đa của các giá trị của biến.
- **Decimals:** số lượng số sau dấu phẩy.
- **Label:** gán nhãn cho biến nhằm giải thích cho biến (do tên biến thường đặt ngắn gọn và không được viết bằng tiếng việt nên có thể không thể hiện được hết ý nghĩa).
- **Value:** Thường các dữ liệu thường ở dạng mã hóa, chẳng hạn, với biến giới tính người ta dùng giá trị 1 thay cho biểu hiện Nam, 2 thay cho Nu, khi đó ta cần thực hiện gán thông qua thuộc tính Value. Việc này ngoài tác dụng giải thích ý nghĩa cho giá trị 1 và 0 giúp việc nhập dữ liệu nhanh hơn. Ta có thể thay đổi hiển thị các biểu hiện của giới tính là giá trị mã hóa (0, 1) hoặc là giá trị gốc (Nu, Nam) thông qua nút View nói ở mục 1.3.1.
- **Missing:** Khai báo các giá trị khuyết. Thông thường có 1 biểu hiện của biến không có thông tin, ta để trống (máy sẽ hiểu là System Missing) hoặc gán một giá trị nào đó, nếu ta muốn gán, mục này cho phép ta làm điều đó. Những giá trị mà ta gán ở đây về sau sẽ không tham gia các quá trình tính toán của biến.
- **Columns:** chỉnh độ rộng của cột bởi một số nào đó. Để làm điều này ta cũng có thể kéo thả trực tiếp trên cửa sổ **Data View**.
- **Measure:** chọn thang đo cho biến. Ordinary là thang thứ bậc, Normial là thang định danh, Scale là thang đo tỉ lệ hoặc khoảng.

1.4. Thực hành nhập dữ liệu

1.4.1. Nhập dữ liệu

Nhập dữ liệu trong bảng sau:

Thứ tự	Giới tính	Tuổi	Thể thao
1	1	20	1
2	0	25	4
3	1	24	3
4	1	23	2
5	0	30	3
6	không rõ	23	2
7	0	không rõ	không rõ

Trong đó:

- cột Thứ tự nhập vào là ThuTu, đặt nhãn cho nó là: Thứ tự; kiểu kí tự (String).
- cột Giới tính nhập tên là GioiTinh; đặt nhãn cho nó là: Giới tính; 1 là giá trị của Nam, 0 là giá trị của Nu; số lượng kí tự tối đa là 3, số lượng sau số thập phân là 0, 3 là giá trị khuyết, thang đo định danh.

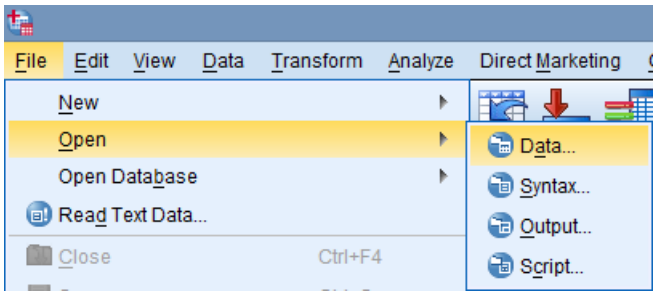
- cột Tuổi nhập tên là Tuoi; đặt nhãn là Tuổi; kiểu số, giá trị khuyết là -1, thang đo tỉ lệ.
- cột Thể thao nhập tên là TheThao; gán nhãn là: Mức yêu thích thể thao; 1 là rất yêu thích, 2 là yêu thích, 3 là bình thường, 4 là không thích; thang đo thứ bậc; giá trị khuyết để dạng System Missing.

Lưu dữ liệu thành các file có tên **ThucHanh.sav**, **ThucHanh.xls**.

1.4.2. Mở dữ liệu từ file sẵn có

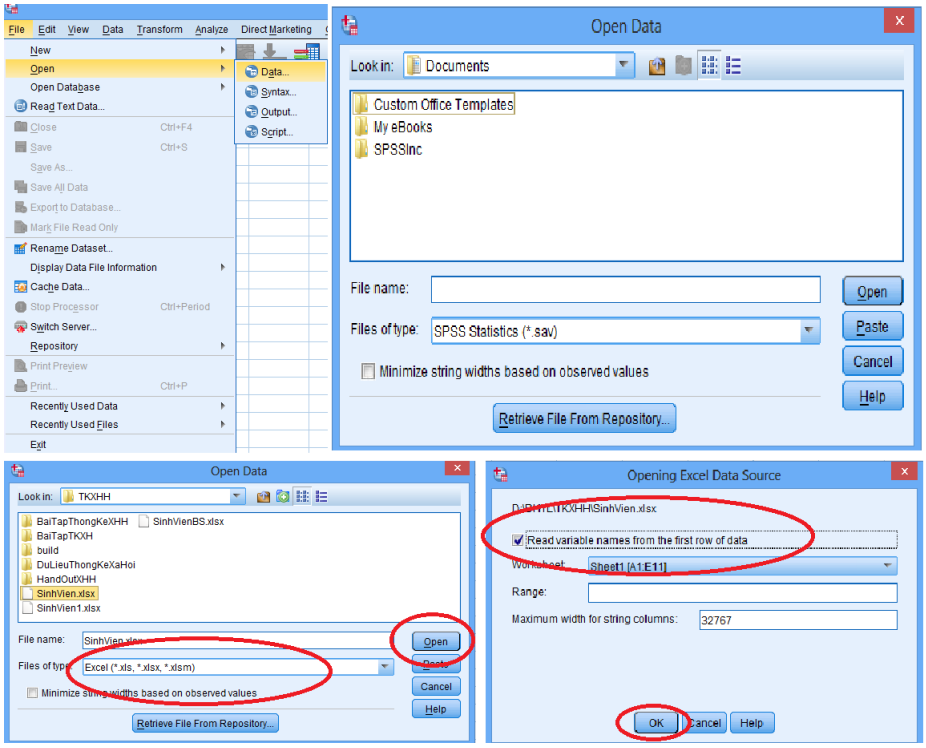
Để đọc dữ liệu từ một file dữ liệu đã sẵn có, nhất là những dữ liệu không ở định dạng .sav ta phải chắc rằng đã thực hiện điều chỉnh trong **Options** của menu **Edit** như đã nói ở hình 5.4.

- Dữ liệu đuôi **.sav** mở bằng cách chọn **File** $\longrightarrow$ **Open** $\longrightarrow$ **Data ...** sau đó chọn ổ chứa và đi đến file cần mở.



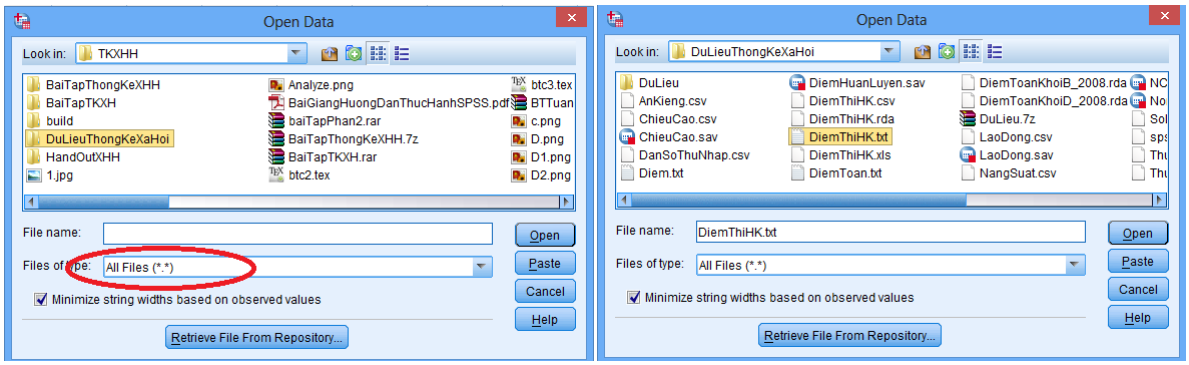
Hình 1.4

- Dữ liệu đuôi **.xls, xlsx** ta làm như sau:



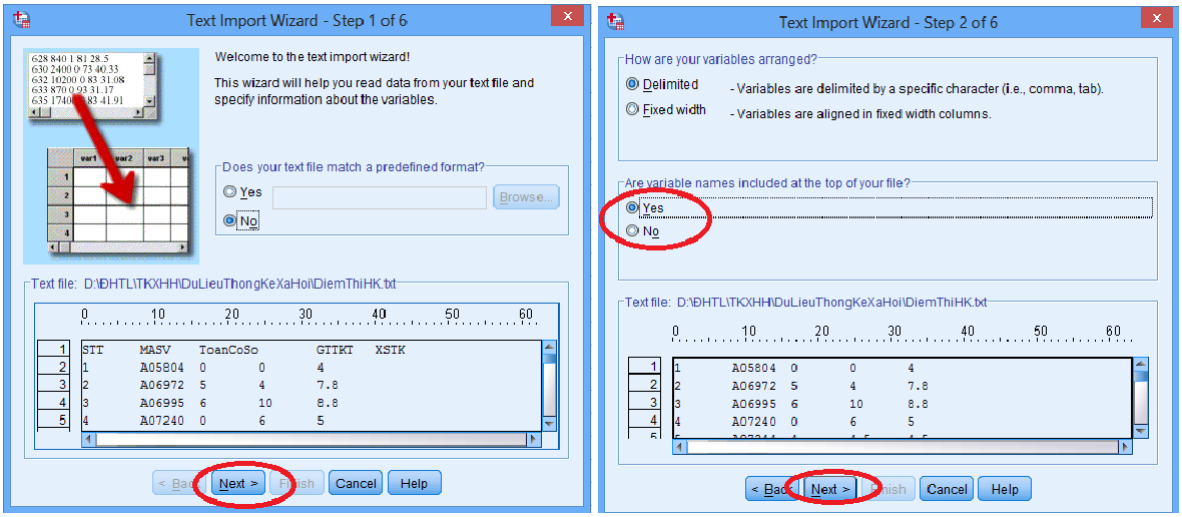
Hình 1.5: Đưa dữ liệu dạng .xls, xlsx, csv vào SPSS - Thứ tự từ trái qua phải, trên xuống dưới

- Dữ liệu đuôi **.txt, .csv ...** ta làm như hình dưới đây.
  - **File** ⟶ **Open** ⟶ **Data ...** sau đó chọn ổ chứa file cần mở.



Hình 1.6: Đưa dữ liệu dạng .txt vào SPSS - Thứ tự từ trái qua phải

  - Mục **File of stype** chọn **All Files** sau đó kích vào file cần mở hiện trong hộp thoại danh sách.
  - Tiếp theo ta phải thực hiện 6 bước với 6 hộp thoại được mở lần lượt. Ở bước 1 ta chọn **Next**, bước 2 chọn **Yes hoặc No** ở khung giữa của hộp thoại và quan sát bản xem trước ở khung cuối của hộp thoại để đạt được yêu cầu mong muốn (Chọn Yes nếu ban đầu dữ liệu đã có tên biến (tên cột)). Các bước tiếp theo chọn **Next** và cuối cùng chọn **Finish**.



Hình 1.7: Tích các lựa chon **Yes/ No** cho phù hợp với dữ liệu ban đầu. Các bước còn lại nhấp **Next**, bước cuối nhấp **Finish**

Lưu ý: khi đọc dữ liệu từ các file không phải dạng .sav ta phải kiểm tra và điều chỉnh lại các thuộc tính của biến (trong cửa sổ **Variable View***) cho phù hợp với biến. Điều này là quan trọng khi ta thực hiện những tính toán, phân tích về sau này.*

Chương 2

Một số thao tác biên tập dữ liệu cơ bản

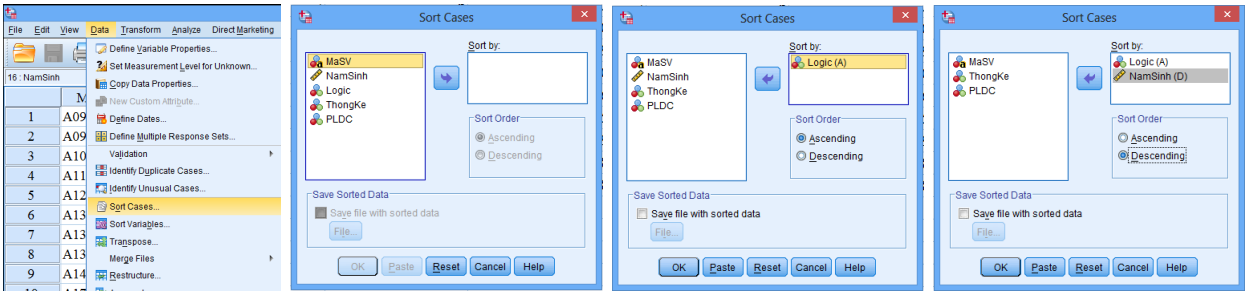
2.1. Sắp xếp, ghép file, lọc dữ liệu

2.1.1. Sắp xếp

Việc sắp xếp dữ liệu theo chiều tăng (hoặc giảm) của các con số hay theo bảng chữ cái của một biến được thực hiện đơn giản nhờ việc bôi vàng một cột (kích vào tên cột) rồi nhấp chuột phải, chọn **Sort Ascending** nếu muốn sắp theo thứ tự tăng dần theo biến được chọn, **Sort Descending** nếu muốn sắp theo thứ tự giảm dần.

Khi muốn sắp xếp ưu tiên sự tăng, giảm theo một số biến, chẳng hạn trong file SinhVien.sav nếu muốn sắp theo thứ tự tăng dần của điểm Logic, sau đó, với những ai điểm Logic giống nhau thì sắp xếp theo thứ tự giảm dần của năm sinh thì ta làm như sau:

- **Bước 1:** Mở file **"SinhVien.sav"**. Vào **Data** $\longrightarrow$ **Sort Cases**. Hộp thoại Sort Cases sẽ hiện ra, bên trái liệt kê các biến có trong tập dữ liệu, bên phải có hai khung là **Sort by** và **Sort Order**.



Hình 2.1: Sort theo nhiều biến - Thứ tự từ trái qua phải

- **Bước 2:** Từ hộp thoại **Sort Cases** chọn biến **Logic**, kích vào mũi tên sang phải để đưa Logic vào khung **Sort by**, sau đó chọn **Sort Order** là **Ascending**.
- **Bước 3:** Làm tương tự bước 2 cho biến **NamSinh**, nhưng chọn **Sort Order** là **Descending**.
- **Bước 4:** Nhấp **Ok**.

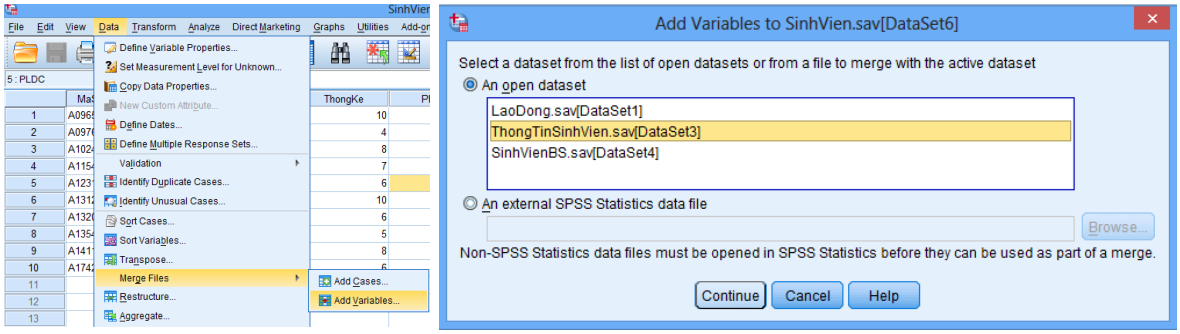
Khi sắp xếp xong như trên, có thể có những người có cùng điểm Logic và cùng năm sinh, khi đó muốn sắp xếp trong nội bộ các nhóm này theo biến khác nữa thì ta lại làm như trên và thêm vào danh sách **Sort by** biến mà ta muốn ưu tiên sắp xếp, tất nhiên ta phải chọn chiều tăng hay giảm cho biến này, ...

2.1.2. Ghép các file

Ghép 2 file với nhau có 2 loại: ghép thêm cột (biến - Variable) và ghép thêm dòng (quan sát - Cases). Để ghép hai file với nhau ta dùng lựa chọn Merge File trong nút Data.

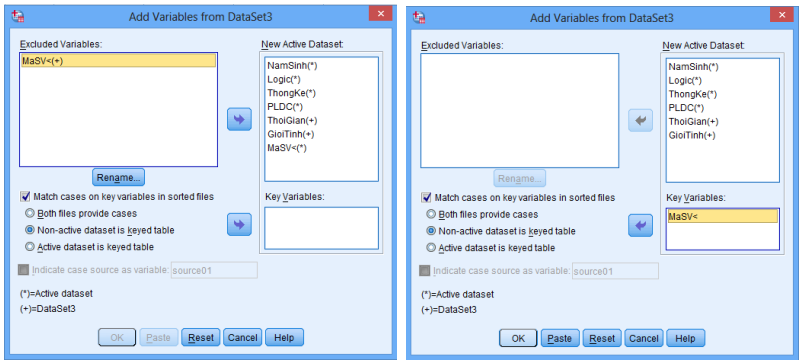
Trước tiên, ta tìm hiểu về ghép thêm cột qua việc thực hành với 2 file cụ thể: **Ghép thêm cột ThoiGian và GioiTinh từ file ThongTinSinhVien.sav vào file SinhVien.sav.** Ở đây ta chọn biến MaSV là khóa để ghép 2 file. (Lưu ý khóa phải có ở cả 2 file và là biến đặc trưng duy nhất cho quan sát, tức là mỗi biểu hiện của biến khóa tương ứng duy nhất với một quan sát).

- Đầu tiên ta mở hai file và *cùng sắp xếp theo chiều tăng (hoặc cùng giảm)* cho cả hai file theo biến **MaSV**.
- Từ cửa sổ DataSet của file SinhVien ta chọn **"Data ⟶ Merge Files ⟶ Add Variables"**
- Làm theo các bước như hình dưới đây. Lưu ý *cách làm này áp dụng cho 2 file đang mở trong SPSS.*



Hình 2.2: Ghép file SinhVien và ThongTinSinhVien - Thứ tự từ trái qua phải

- Ở bước thứ 2 trong hình ta chọn file cần ghép vào file ban đầu, rồi nhấp **"Continue"**
- Ở bước thứ 3 trong hình ta nhấp chọn **"Match cases on key ..."** rồi sau đó nhấp chọn vào biến **MaSV** và chuyển vào khung trống cuối bằng cách nhấp vào mũi tên chuyển.



Hình 2.3: Ghép file SinhVien và ThongTinSinhVien - Thứ tự từ trái qua phải

Tiếp sau ta chọn cách thức ghép, có 3 lựa chọn:

– Khi chọn **"Both files provide cases"** tức là: nối 2 file gộp với nhau. Theo cách này, **file nhận được sẽ bao gồm tất cả các quan sát mà có ở hai file**.

  - Khi chọn **"Non - active dataset is keyed table"** tức là: nối 2 file lấy cột **MaSV trong file SinhVien làm chuẩn**, tức là file nhận được sẽ chỉ có các quan sát như trong file SinhVien trong đó được bổ sung thêm các cột của ThongTinSinhVien. Quan sát nào mà trong SinhVien có nhưng lại không có trong ThongTinSinhVien thì trong các biến thêm vào ứng với quan sát đó sẽ để trống dạng **System Mising**.
  - Còn nếu chọn **"Active dataset is key table"** tức là: nối 2 file lấy cột **MaSV trong file ThongTinSinhVien làm chuẩn**, kết quả sẽ được file có các quan sát như trong file ThongTinSinhVien. Quan sát nào mà trong ThongTinSinhVien có nhưng lại không có trong SinhVien thì trong các biến thêm vào ứng với quan sát đó sẽ để trống dạng System Mising.

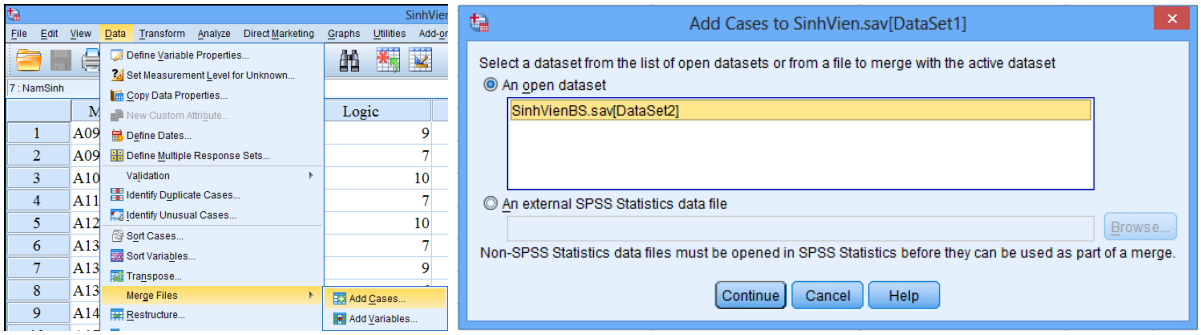
- Cuối cùng nhấn **"Ok"**.

Sau khi thực hiện nối file, file nhận được là file ban đầu có thêm các cột mới, thứ tự các cột được sắp xếp lần lượt hết các cột trong file ban đầu sau đó đến các cột của file ghép vào. Hình trên đây ghép file lấy MaSV trong file ban đầu (SinhVien) làm chuẩn. Các cách ghép còn lại sinh viên tự thực hành.

Sau đây tìm hiểu cách nối thêm quan sát. Nếu là 2 file có các biến như nhau, ta chỉ cần **Sort Variable** cho cả 2 file rồi **Copy - Paste** từ file này vào file kia. Vấn đề chỉ nảy sinh khi ta muốn nối 2 file mà các biến là không như nhau.

Sau đây là hướng dẫn thực hành cho việc nối file **SinhVienBS** vào file **SinhVien** để có được file **SinhVien** mới với những sinh đã có và những sinh viên bổ sung từ **SinhVienBS**. Ta lưu ý rằng ở đây hai file có các biến không giống nhau, chỉ chung các biến MaSV, NamSinh, Logic, ThongKe. Ta thực hành với trợ giúp của hướng dẫn chi tiết và hình minh họa dưới đây.

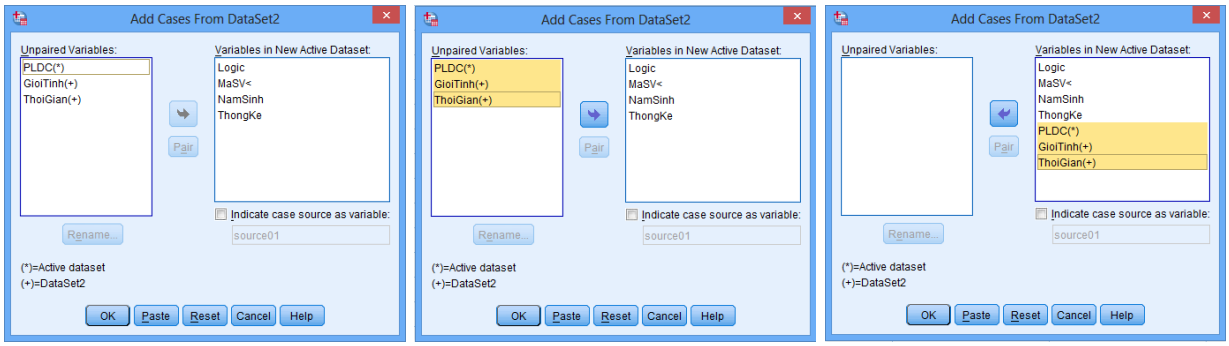
- Đầu tiên ta mở cả 2 file trên SPSS.
- Từ cửa sổ DataSet của file SinhVien ta chọn **"Data ⟶ Merge Files ⟶ Add Cases ... "**
- Chọn file muốn bổ sung vào file ban đầu, và nhấp **"Continue"**.



Hình 2.4: Ghép thêm vào SinhVien những quan sát trong SinhVienBS - Hai bước đầu

- Ở bước thứ 3 trong hình ta thấy hộp thoại xuất hiện có hai khung danh sách, **ban đầu, bên trái là những danh sách biến mà chỉ xuất hiện trong một trong 2 file. Bên phải là danh sách những biến sẽ được hiển thị trong file kết quả mà ta sẽ nhận được**, những biến xuất hiện ở cả hai file sẽ tự động có trong danh sách này từ đầu. Ở bước này ta nhấp chọn

vào những biến bên khung trái mà ta muốn hiển thị trong file kết quả và chuyển sang khung bên phải bởi nút mũi tên, làm tương tự nếu muốn loại bỏ những biến có trong danh sách bên phải. Ở đây chúng tôi chọn file kết quả có tất cả các biến xuất hiện ở 2 file.



Hình 2.5: Ghép thêm vào SinhVien những quan sát trong SinhVienBS - Ba bước cuối

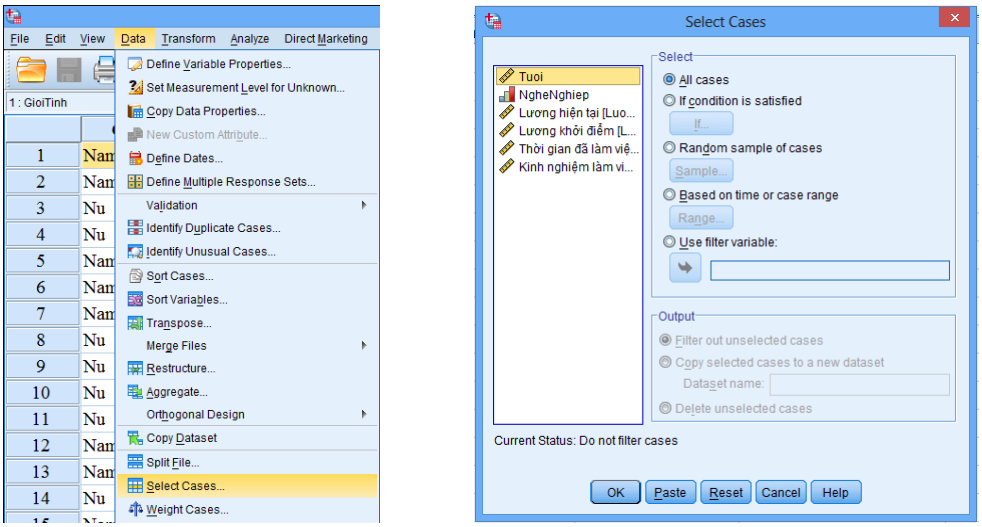
- Cuối cùng nhấp **"Ok"**.

2.1.3. Lọc dữ liệu

Lọc dữ liệu tức là ta lựa chọn ra một số những quan sát từ trong tập dữ liệu ban đầu. Việc lựa chọn có thể là ngẫu nhiên hoặc chọn thỏa mãn yêu cầu nào đó.

Để lọc dữ liệu ta dùng lựa chọn **Select Cases** trong nút **Data**. Hướng dẫn sau đây thực hiện trên file **"LaoDong.sav"**.

- Vào **Data** $\longrightarrow$ **Select Cases** ta được hộp thoại **Select Cases**. Hộp thoại này gồm 3 phần: danh sách các biến ở khung bến trái, nó được dùng khi ta cần lọc dữ liệu thỏa mãn một hệ điều kiện nào đó của các biến; bên phải gồm 2 khung: **Select** (cách thức chọn quan sát) và **Output** (chọn kiểu file đầu ra).



Hình 2.6: Lọc dữ liêu - Hai bước đầu

- Trong khung Select có các lựa chọn:
  - **All cases:** chọn tất cả các quan sát. Việc khi nào dùng lựa chọn này ta sẽ bàn ở các mục bên dưới.

- **If condition is satisfied:** chọn các quan sát thỏa mãn điều kiện nào đó. Đây là bài toán lọc hay gặp nhất. Khi lựa chọn mục này, ta kích tiếp vào nút **If ...** để thiết lập điều kiện.
- **Random sample of cases:** chọn ngẫu nhiên một số lượng quan sát nào đó. Khi lựa chọn mục này, kích tiếp nút **Sample ...** để thiết lập số lượng chọn ngẫu nhiên (theo phần trăm số lượng quan sát của tập dữ liệu hoặc theo số lượng).
- **Based on tim or cases:** chọn các quan sát từ dòng ... đến dòng ... trong dữ liệu ban đầu. Khi lựa chọn mục này, kích tiếp vào nút **Range** để nhập khoảng dòng quan sát muốn lọc ra.
- **Use filter variable:** chọn các quan sát theo một biến lọc cho trước, biến lọc này phải có trong danh sách biến. Biến lọc là một biến mà các biểu hiện của nó chỉ là 0 và 1. Khi lọc theo biến này ta chọn biến lọc và chuyển qua khung bên dưới mục **Use filter variable** bằng mũi tên chuyển. Khi hoàn thành, ta sẽ được tập dữ liệu chỉ lấy các quan sát có biến lọc bằng 1 và quá trình tính toán về sau.

- Trong khung **Output** cho ta lựa chọn kiểu đầu ra của file lọc, có các lựa chọn sau đây:
  - **Filter out unselected cases:** File lọc **vẫn là file ban đầu, nó hiển thị tất cả các quan sát, nhưng thực ra file lọc chỉ gồm những quan sát không không bị gạch ở đầu dòng**. Những quan sát bị gạch không tham gia vào các quá trình tính toán sau này. Sau khi thực hiện các tính toán trên file lọc đôi khi ta muốn trở lại với dữ liệu đầy đủ ban đầu (tức là muốn bỏ những dấu gạch) khi đó ta thực hiên lại thủ tục **Select Cases** với lựa chọn trong **khung Select** là **All cases**.
  - **Copy selected cases to new dataset:** Copy những quan sát được chọn ra một tập dữ liệu mới. Khi dùng lựa chọn này ta phải đặt tên cho tập dữ liệu mới trong khung trống ngay phía dưới.
  - **Deleted unselected cases:** Xóa đi những quan sát không được lựa chọn.

  Ta thấy rằng *tốt nhất nên dùng lựa chọn thứ 2*, tức là ta nên cho những quan sát mà được lựa chọn sang một tập dữ liệu mới. (và không nên dùng lựa chọn cuối cùng.)

Sau đây là hướng dẫn thực hành 3 trong 5 kiểu lọc quan sát.

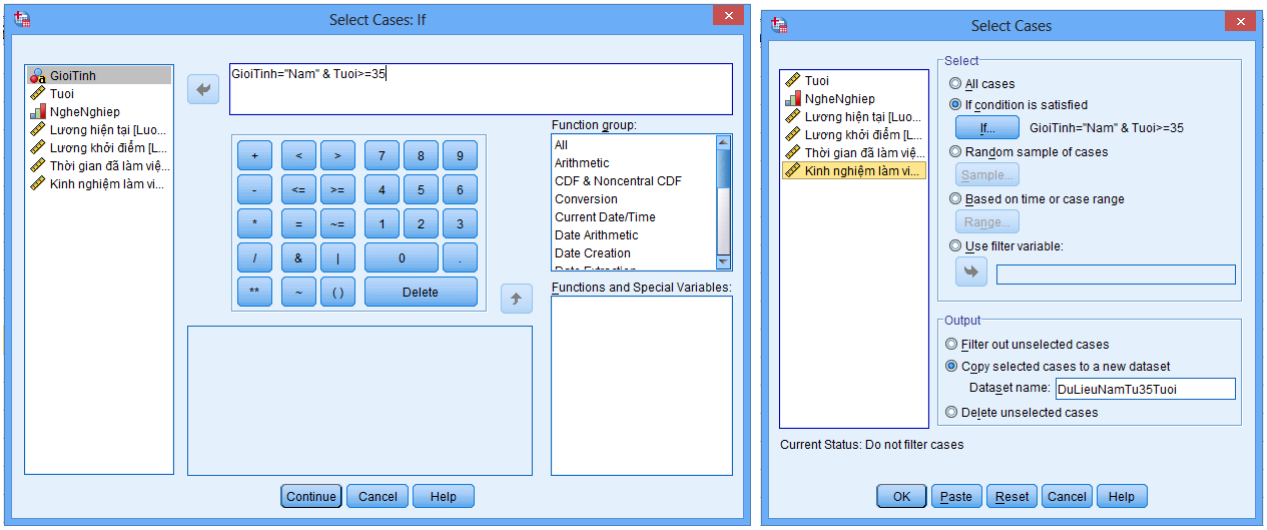
Mở file **LaoDong.sav** (bao gồm thông tin của 474 lao động được điều tra):

1. lọc ra các quan sát là giới tính nam và tuổi từ 35 trở lên, đặt tên là **DuLieuNamTu35Tuoi**.

   Bước đầu tiên ta làm như hình 2.6. Ta tích vào **If condition is satisfiled** chọn **If ...**. Trong hộp thoại hiện ra như hình dưới đây, ta gõ vào khung trống bên phải dòng lệnh: **GioiTinh="Nam"** & **Tuoi>=35**. Nhấp **Continue** ta trở lại với hộp thoại **Select Cases** ở đó hẳn chúng ta đã thấy dòng lệnh vừa gõ được hiển thị mờ mờ bên cạnh nút **If...**.

   Tiếp theo ta chọn kiểu file kết quả, tùy vào mục đích của người dùng có thể chọn các kiểu khác nhau, ở đây chúng ta chọn mục thứ 2: **Copy selected cases to a new dataset** và đặt tên cho file lọc là **DuLieuNamTu35Tuoi**. Cuối cùng nhấp **OK**.

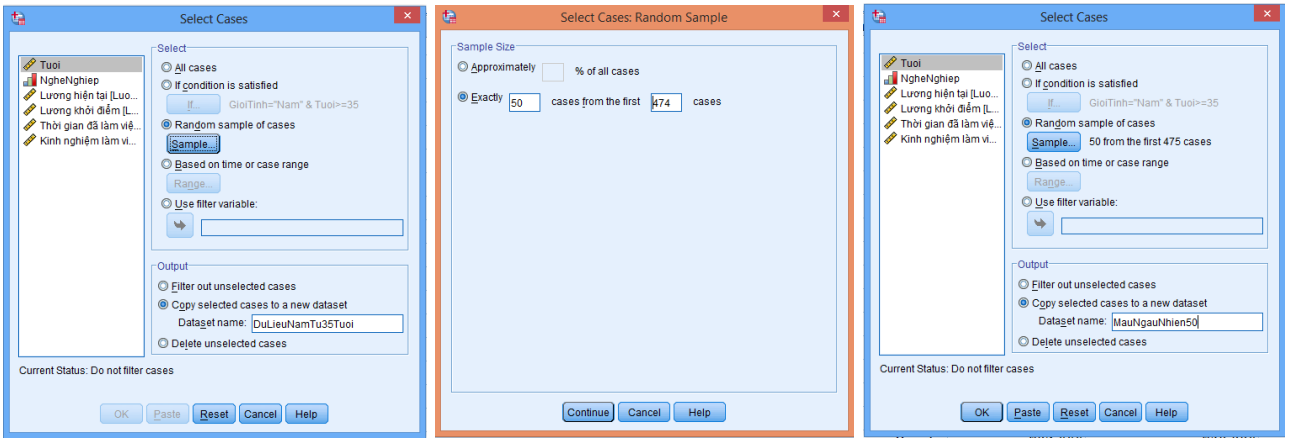
   File kết quả thường không hiển thị thẳng trên màn hình, hãy kiểm tra trong tập hợp các file SPSS đang mở, nó chính là **Untitled...[DuLieuNamTu35Tuoi]**.



Hình 2.7: Lọc dữ liệu thỏa mãn hệ điều kiện nào đó

2. Chọn ra ngẫu nhiên 50 người trong số những người được điều tra.

   Tất nhiên ta cũng thực hiện thao tác ban đầu như trong hình 2.6, sau đó chọn **Random sample of cases**, nhấp tiếp **Sample**. Hộp thoại hiện ra cho phép thực hiện 2 lựa chọn: Xấp xỉ theo phần trăm và chọn chính xác bao nhiêu phần tử trong tập hợp ... dòng đầu tiên. Ở đây ta cần chọn ngẫu nhiên chính xác 50 người từ tập ban đầu, nên các thông số điền như sau (474 là số người được điều tra, chính là số dòng của tập dữ liệu). Nhấp **Continue**.



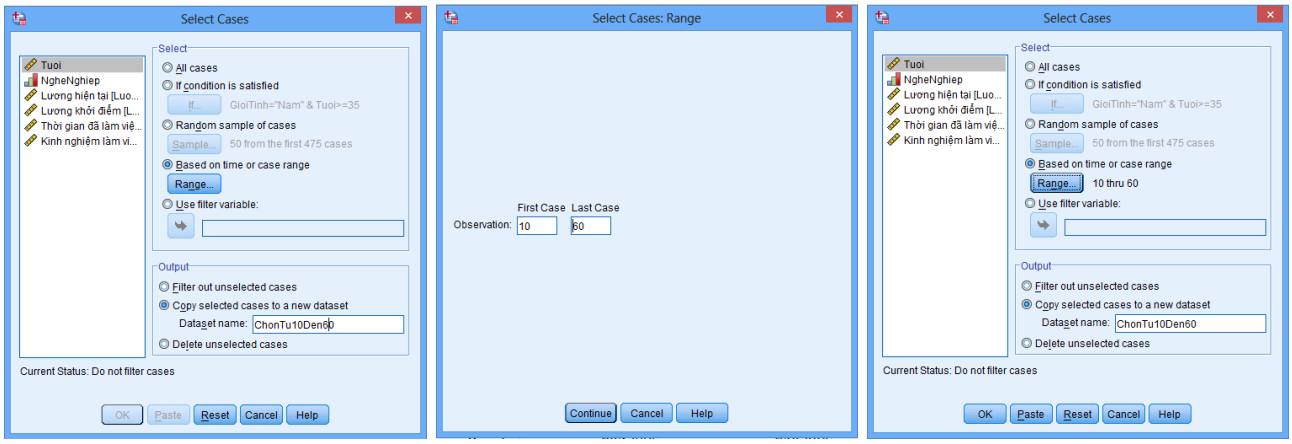
Hình 2.8: Chọn ngẫu nhiên

Trong **Output** đổi tên file đầu ra là **MauNgauNhien50** rồi nhấp **OK**.

3. Chọn ra các quan sát từ dòng 10 đến 60.

   Thực hiện thao tác ban đầu như trong hình 2.6, sau đó chọn **Based on time or case range**. Đổi tên trong **Output** thành **ChonTu10Den60**, sau đó nhấp **Range** điền 10 và 60 vào 2 ô trống.

   Nhấp **Continue**, rồi **OK**.



Hình 2.9: Lọc lấy các phần từ dòng 10 đến 60

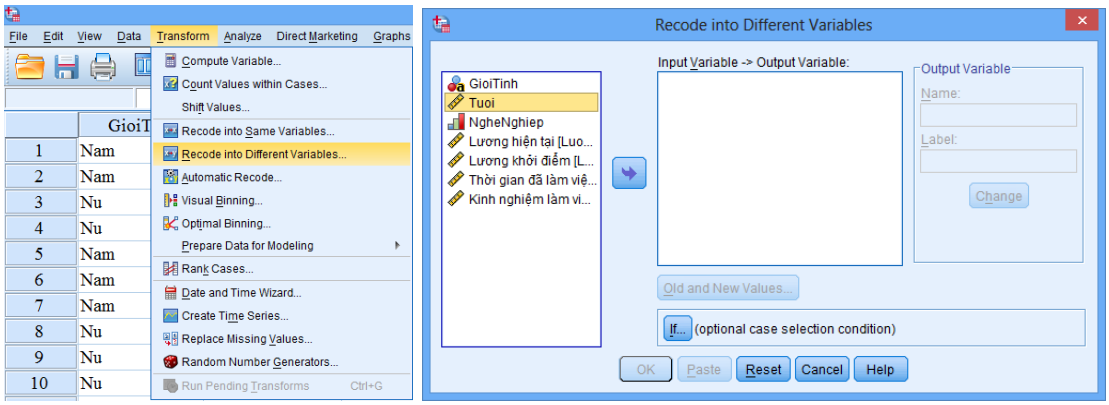
2.2. Tạo biến mã hóa của biến cho trước

Trong quá trình phân tích dữ liệu, nhiều khi vì mục đích nghiên cứu mà ta muốn giảm thiểu số lượng biểu hiện khác nhau của một biến định lượng. Khi đó ta cần mã hóa dữ liệu chuyển biến định lượng thành một biến định tính với ít biểu hiện hơn, tạo thuận lợi cho việc tóm tắt dữ liệu. Việc mã hóa dữ liệu trong SPSS được thực hiện qua lựa chọn **Recode into same Variables** và **Recode into Different Variables** trong nút **Transform**. Khi dùng Recode into Different Variables tức là ta tạo ra một biến mới là mã hóa của biến ban đầu, còn dùng Recode into same Variables tức là tạo ra biến mã hóa và biến này thay thế luôn biến ban đầu.

Sau đây sẽ chỉ hướng dẫn cách dùng **Recode into Different Variables**.

Chẳng hạn trong file **LaoDong.sav** biến tuổi có rất nhiều biểu hiện từ 18 đến 60. Giả sử ta muốn chia nhóm tuổi thành nhóm 1: không quá 25, nhóm 2: từ trên 25 đến 35, nhóm 3: từ trên 35 đến 45, nhóm 4: từ trên 45 thì ta thực hiện như sau:

- Mở file **LaoDong.sav** vào **Transform** $\longrightarrow$ **Recode into Different Variables** ta được hộp thoại mã hóa.



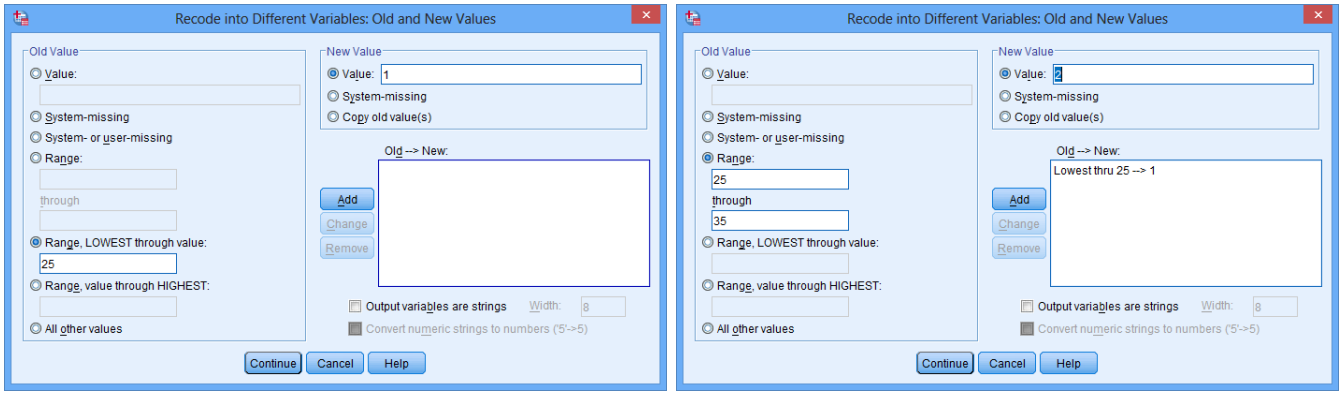
Hình 2.10: Mã hóa biến Tuoi - Bước 1

- Từ danh sách biến chuyển biến Tuoi sang khung trống bên phải.

  Sau đó khai báo cho tên của biến mã hóa trong mục **Name** của khung Output Variable. Nhấn nút **Change** khi khai báo xong như hướng dẫn trong loạt hình 2.11

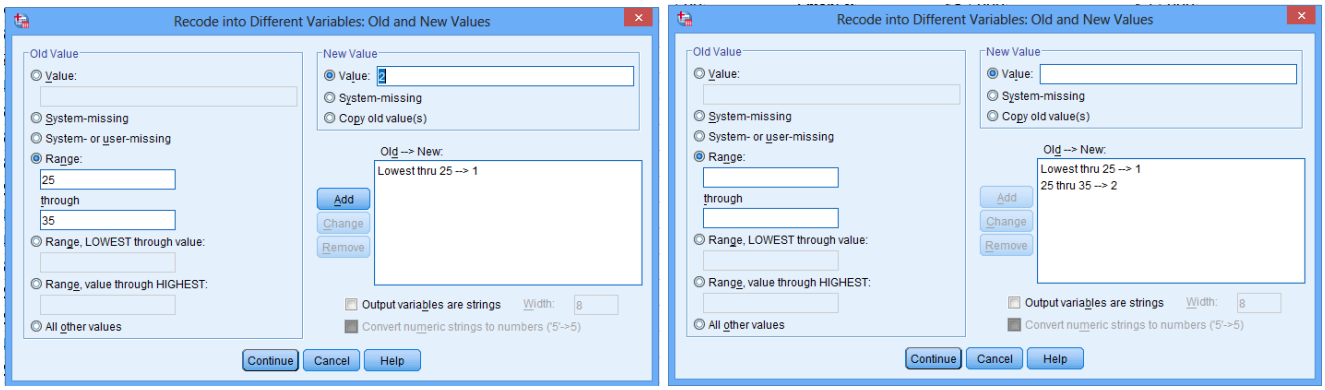
Hình 2.11: Mã hóa biến Tuoi - Bước 2: Đặt tên cho biến mã hóa

- Vào **Old and New Values** hộp thoại xuất hiện và khai báo cho như sau:
  - Từ cửa sổ Old Value tích **Range, LOWEST through value** và điền vào giá trị 25. Bên cửa sổ New Value điền: 1. Nhấp **Add**.



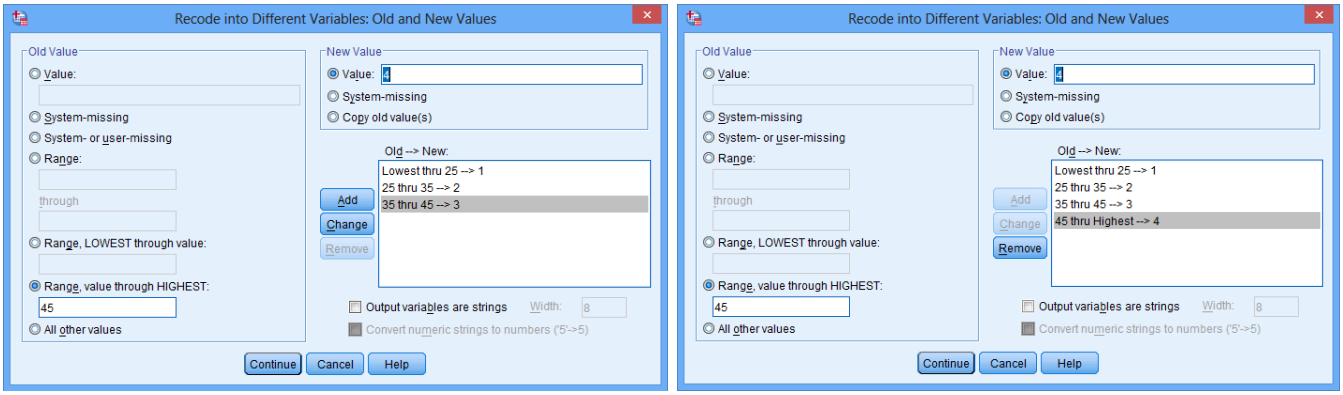
Hình 2.12: Mã hóa biến Tuoi - Bước 3

  - Từ cửa sổ Old Value tích**Range** và điền vào lần lượt hai giá trị là 25 và 35 vào hai ô. Bên cửa sổ New Value điền: 2. Nhấp **Add**



Hình 2.13: Mã hóa biến Tuoi - Bước 4

  - Tương tự như bước trên, từ cửa sổ Old Value tích **Range** và điền vào lần lượt hai giá trị là 35 và 45 vào hai ô. Bên cửa sổ New Value điền: 3. Nhấp **Add**
  - Từ cửa sổ Old Value tích **Range, value through HIGHEST** và điền vào giá trị 45. Bên cửa sổ New Value điền : 4. Nhấp **Add**



Hình 2.14: Mã hóa biến Tuoi - Bước 5

- Nhấp **Continue** và ta trở lại hộp thoại Recode. Ở cuối hộp thoại có nút **If ...** dùng để điều chỉnh khi ta chỉ muốn mã hóa biến trên đối với những quan sát thỏa mãn một điều kiện nào đó. Ở đây ta mã hóa cho mọi quan sát nên ta không cần điều chỉnh mục này.
- Nhấp **OK**.
- Ta trở về với cửa sổ chính, trong phần cửa số khai báo biến **Variable View** ta bổ sung thuộc tính **Value** cho biến **PhanNhomTuoi** mà ta vừa tạo được: giá trị 1 được gán nhãn là <=25, 2 là (25; 35], 3 là (35; 45], 4 là >45.

Lưu ý quan trọng:

1. Khi ta muốn kết quả biến mã hóa có giá trị là các kí tự (chứ không phải dạng numeric như ví dụ trên đây), ở bước 3, hình 2.11 ta nhấp vào **Output is String** ở dòng bên dưới khung **New Value**.
2. Khi muốn mã hóa từng giá trị của biến cũ thì thay vì sử dụng các nút thay thế theo khoảng (Range) ta điền giá trị muốn thay thế vào khung trống của **Value** bên khung **Old Value**.
3. Khi chuyển các giá trị Missing sang biến mã hóa, sẽ có nhiều sự lựa chọn, nếu ta lựa chọn **Copy Old Values** và biến ban đầu có giá trị missing do người dùng tự định nghĩa thì ta phải bổ sung chú thích lại các giá trị thuộc tính Missing của biến được tạo thành.

Với lựa chọn **Recode into same Variables** sinh viên tự tìm hiểu và thực hành tương tự như đối với **Recode into Different Variables**.

2.3. Một số tính toán cơ bản trên các biến

Đối với các biến định lượng, SPSS cho phép ta thực hiện các phép toán số học, thống kê, lượng giác ... đối với các giá trị của biến trên mỗi quan sát.

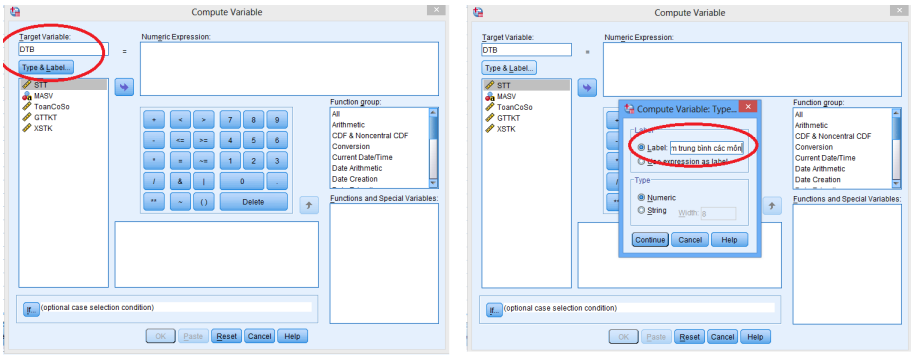
Để làm điều này ta vào **Transform** $\longrightarrow$ **Compute Variable**, hộp thoại hiện ra gồm:

- **Target Variable:** Đặt tên cho biến kết quả.
- **Numeric Expression:** Biểu thức để tính giá trị của biến mới.
- Khung bên trái là danh sách các biến. Các biến này có thể chuyển vào biểu thức tính toán ở mục **Numeric Expression:** bằng cách chọn biến và nhấp vào mũi tên chuyển.

- Khung bên phải là danh sách các hàm: khung phía trên **Function group** là các nhóm hàm, khung bên dưới là các hàm cụ thể. Ở khung trên chọn **All** sẽ cho danh sách tất cả các hàm tính toán trong SPSS ở khung bên dưới, nếu muốn khung dưới hiển thị nhóm hàm riêng biệt thì chọn nhóm hàm trong khung phía trên. Khi ta chọn một hàm để đưa vào biểu thức thì ta nhấp đúp vào hàm đó. Ở khung giữa là bảng tính, ta có thể dùng bàn phím để thay thế nó. Bên dưới bàn tính là giải thích cấu trúc và tham số trong mỗi hàm khi được chọn.

Chẳng hạn, tính trung bình điểm các môn cho danh sách sinh viên trong file **DiemThiHK.csv** ta làm như sau:

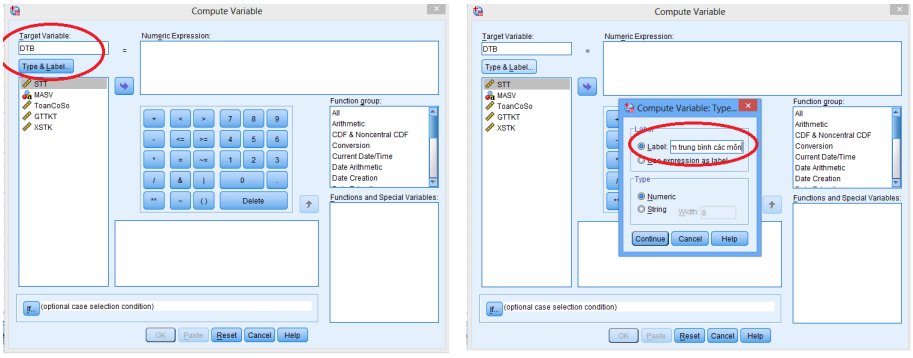
- Mở file **DiemThiHK.csv** trong SPSS. Vào **Transform** $\longrightarrow$ **Compute Variable**.



Hình 2.15: Tính điểm trung bình của các sinh viên trong file DiemThiHK.csv - Bước 1

- Đặt tên biến mới trong **Target Variable** là DTB, nhấp **Type & Label ...** gõ nhãn chú thích là Điểm trung bình các môn, **Type: Numberic**, rồi nhấp **Continue**.

- Trong khung **Function group** tìm và chọn **Statistical**. Sau đó ở khung bên dưới nhấp đúp vào **mean** (tính giá trị trung bình).



Hình 2.16: Tính điểm trung bình của các sinh viên trong file DiemThiHK.csv - Bước 2

- Ở khung trên cùng **Numeric Expression:** ta thay thế các dấu hỏi bởi cách chuyển các biến từ danh sách các biến từ bên trái sang, mỗi tham số của hàm ngăn cách nhau bởi một dấu ",".

- Nhấp **OK**. Kết quả là một biến mới tên là DTB gồm điểm trung bình các môn được chọn của tất cả các quan sát. Lưu ý rằng các quan sát mà có một vài điểm nào đó bị khuyết SPSS sẽ chỉ tính điểm trung bình trên những giá trị (ở đây là điểm) bị khuyết.

Chú ý rằng ở bài toán trên nếu các biến tham gia tính toán không chứa giá trị khuyết thì thay vì dùng hàm mean trong khung **Numeric Expression:** ta có thể sử dụng công thức tổng các biến chia cho số lượng biến như một biểu thức số học bình thường.

2.4. Bài tập

Bài tập 2.1. Trong những biến dưới đây, hãy chỉ ra biến nào là biến định lượng, biến nào là biến định tính. Các biến đó dùng thang đo nào? Biến nào là liên tục, biến nào là rời rạc?

1. Số lỗi đánh máy sai trong các tờ báo.
2. Hóa đơn tiền điện hằng tháng.
3. Số xe máy các gia đình có được.
4. Doanh số lợi nhuận từ việc bán vé số của các tỉnh.
5. Những nơi nghỉ hè mà mọi người ưa thích.
6. Số con trong các gia đình
7. Nhiệt độ trung bình các tháng trong năm
8. Lượng mưa trung bình các tháng trong năm
9. Các câu trả lời của câu hỏi về mức độ yêu thích đối với môn Thông kê xã hội học của một lớp (Với các câu trả lời 1. Rất thích, 2. Thích, 3. Bình thường, 4. Không thích, 5. Rất không thích).

Bài tập 2.2. Bảng dữ liệu sau cho ta thông tin về một số hộ gia đình:

Khu vực	Số lao động	Số phụ thuộc	Tổng thu nhập	Tổng chi tiêu
NongThon	2	3	100	90
ThanhThi	3	2	320	250
MienNui	3	1	100	95
MienNui	2	2	80	100
NongThon	3	3	250	300
ThanhThi	4	4	500	
NongThon	2	3		100
ThanhThi	1	1	120	100
ThanhThi	3	2	280	
MienNui	4	3	200	220

1. Nhập dữ liệu trên vào SPSS. Trong đó những ô trống là không có thông tin. Các tên cột viết không dấu và có chú thích cụ thể trong nhãn.
2. Bằng cách sắp xếp đưa ra thông tin hộ có tổng thu nhập thấp nhất, cao nhất.
3. Mã hóa lại và thay thế biến Khu vực sao cho: NongThon thành 1, MienNui là 2, ThanhThi là 3. Khai báo những giá trị này trong thuộc tính Value của biến. Nhập thêm 2 hộ sau:

Khu vực	Số lao động	Số phụ thuộc	Tổng thu nhập	Tổng chi tiêu
1	2	2	100	85
2	3	1	400	350

4. Lập một biến mới là tổng số thành viên của hộ (tổng của số lao động và số phụ thuộc), đặt tên là SoThanhVien, với nhãn là: số thành viên.
5. Lập biến mới tính thu nhập trên đầu người của mỗi hộ, đặt tên là TNTB với nhãn là Thu nhập trên đầu người.
6. Lập biến mới là tích lũy của các hộ trong năm qua (hiệu của tổng thu nhập và tổng chi tiêu).
7. Tạo biến mã hóa, sao cho những hộ có thu nhập trên đầu người nhỏ hơn 36 được gán là 1, còn lại là 0; trong đó 1 được gán trong thuộc tính Value là HoNgheo, 0 là KhongNgheo.
8. Lọc những hộ ở thành thị ra một bảng dữ liệu riêng, lưu lại thành file: "ThanhThi.xls".
9. Lọc ra những hộ có thu nhập trên đầu người trên 50, lưu lại thành file "ThuNhap50.sav".
10. Lấy ra ngẫu nhiên 3 hộ trong các hộ trên.

Bài tập 2.3. Trong file **DiemThiHK.xls** chứa điểm 3 môn toán của một số sinh viên của trường Thăng Long.

1. Đọc dữ liệu, điều chỉnh lại các thuộc tính của biến cho hợp lí và lưu vào file DiemTHiHK.sav.
2. Lập thêm các cột tính tổng điểm, điểm trung bình của các sinh viên,
3. Tính số người thi lại của mỗi môn, điểm thi cao nhất, thấp nhất của mỗi môn.
4. Sắp xếp lại dữ liệu theo thứ tự tổng điểm từ cao xuống thấp. Sau đó in ra danh sách 5 người có tổng điểm cao nhất.
5. File ThongTin.sav chúa thông tin về mã sinh viên, giới tính của một số sinh viên. Hãy tìm trong file ThongTin.sav giới tính của các sinh viên có trong danh sách DiemTHiHK.xls rồi tạo thêm cột giới tính vào cuối bảng. Lưu lại.
6. Lọc ra danh sách những sinh viên nữ trong danh sách vừa lưu lại ở bước trên và lưu thành file mới tên là SinhVienNu.sav. Từ đây suy ra số sinh viên nữ trong danh sách trên.
7. Lọc ra danh sách những sinh viên mà điểm tất cả các môn đều >= 5. Có bao nhiêu sinh viên như vậy?
8. Lấy ngẫu nhiên 10 người từ danh sách trên.

Bài tập 2.4. Trong file **DanSo.xls** chứa số liệu về tổng thu nhập và dân số của nước ta từ năm 1990 đến năm 2007.

1. Đọc dữ liệu, điều chỉnh lại thuộc tính của các biến và lưu dưới dạng .sav,
2. Thêm vào bảng dữ liệu cột ThuNhapTB tính thu nhập bình quân trên đầu người của nước ta qua các năm và lưu lại số liệu.
3. Hai năm có thu nhập bình quân lớn nhất trong danh sách trên là những năm nào?

Bài tập 2.5. File **LaoDong.sav** chứa thông tin về 1 mẫu ngẫu nhiên gồm 474 nhân viên của 1 công ty.

1. Các biến trong file dữ liệu biến nào là định tính, biến nào là định lượng. Chúng thuộc thang đo nào.
2. Cho thông tin về người có lương cao nhất.
3. Lọc ra thông tin những người làm nghề nghiệp NVVP và lưu lại thành file NVVP.sav.
4. Có bao nhiêu nữ làm QuanLy.
5. Mã hóa biến GioiTinh thành MHGT biết Nam ký hiệu là 1, Nữ ký hiệu là 0.
6. Mã hóa cột Lương thành Luong4N với 4 nhóm lần lượt là: 1. Lương ≤ 30000; 2. Lương trong khoảng $(30000, 50000]$; 3. Lương trong khoảng $(50000, 70000]$; 4. Lương trên 70000.

Chương 3

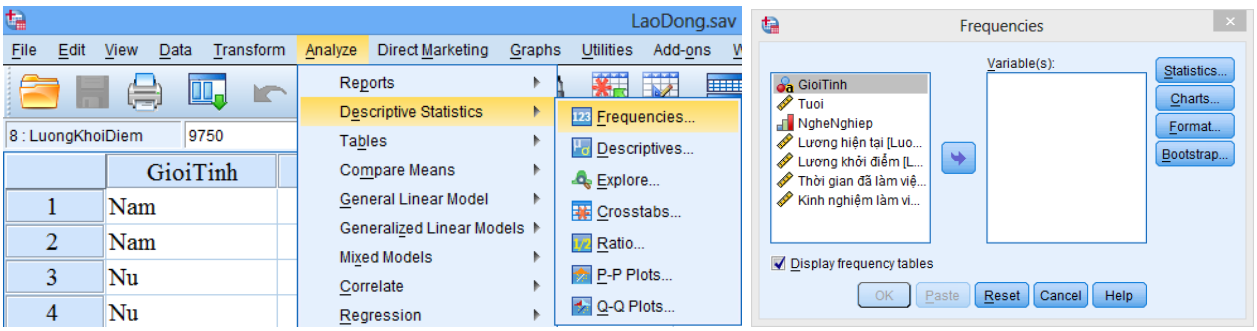
Tóm tắt dữ liệu

Trong chương này ta sẽ tìm hiểu về tóm tắt dữ liệu bằng bảng tần số, bằng các đại lượng thống kê mô tả và minh họa trực quan các tóm tắt đó bởi biểu đồ. Lưu ý rằng khi thực hiện những thao tác để có những thông tin trên, cũng như các phân tích khác mà ta thực hiện từ nay về sau, SPSS sẽ đưa ra kết quả ở cửa sổ **Ouput**. Chúng ta có thể copy từng kết quả riêng lẻ này ra word hoặc lưu lại toàn bộ dưới dạng **.spv** bằng cách từ cửa sổ Output dùng tổ hợp phím **Ctrl + S**.

Trong SPSS, có nhiều thủ tục có thể giúp người dùng có được những thông tin tóm tắt về tập dữ liệu, các thủ tục này dẫn đến cách trình bày bảng biểu khác nhau nhưng thông tin mang lại thì đa phần giống nhau. Ở đây chúng ta chú trọng vào kết quả phân tích nên sẽ chỉ giới thiệu một vài thủ tục dẫn đến kết quả mong muốn.

3.1. Tóm tắt dữ liệu bằng các dạng đơn giản của bảng tần số, biểu đồ và các đại lượng thống kê mô tả

Để lập bảng tần số, vẽ biểu đồ và tính toán các đại lượng thống kê mô tả cho một biến ta vào **Analyze** ⟶ **Desriptive Statistics** ⟶ **Frequencies** . Hộp thoại xuất hiện gồm hai khung: danh sách biến và **Variable(s)** (các biến được chọn).



Hình 3.1: Hộp thoại trong Frequencies

Nhờ hộp thoại Frequencies ta có thể tạo:

- **Bảng tần số:** Để lập bảng tần số của một *biến định tính hoặc biến định lượng ít biểu hiện* ta chỉ cần chọn biến đó từ danh sách biến chuyển vào khung **Variable(s)** bằng mũi tên chuyển. Chọn tích vào **Display frequency tables** để tạo bảng tần số, tần suất và tần số tích lũy. Sau đó nhấp **OK**

Chẳng hạn, hình sau minh họa cách lập bảng tần số, tần suất, tần suất tích lũy cho biến **NgheNghiep**.

Statistics

NgheNghiep

N	Valid	474
	Missing	0

NgheNghiep

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NVVP	363	76.6	76.6	76.6
	CongNhan	27	5.7	5.7	82.3
	QuanLy	84	17.7	17.7	100.0
	Total	474	100.0	100.0	

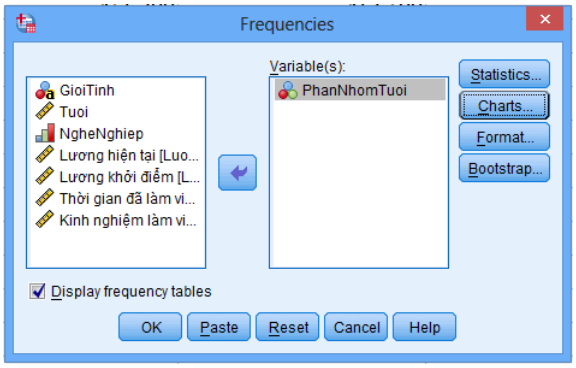
Hình 3.2: Lập bảng tần số, và kết quả trong Output

Trong bảng kết quả ta thấy xuất hiện **Statistics** thống kê những quan sát có (Valid) và quan sát trống (Missing) của biến **NgheNghiep**. Ta thấy biến này không có giá trị trống (Missing).

Bảng còn lại là ghép lần lượt của bảng tần số, tần suất (tính cả Missing), tần suất chỉ tính trên quan sát có dữ liệu, và bảng tần suất tích lũy (không tính Missing). Hai cột giữa giống nhau vì biến NgheNghiep không có quan sát trống. Qua bảng này ta thấy có 363 NVVP chiếm 76.6 %,... Nếu tính gộp cả NVVP và CongNhan thì chiếm tỉ lệ 82.3 %, ...

Với *biến định lượng có nhiều biểu hiện* khi tóm tắt bằng bảng tần số ta phải chia thành các tổ, tức là tạo ta một biến mới bằng cách *mã hóa biến ban đầu. Và ta lập bảng tần số cho biến mới* này như với một biến định tính đã nói ở trên.

Chẳng hạn, để lập bảng tần số cho biến **Tuoi** (biến định lượng) có rất nhiều biểu hiện khác nhau. Đầu tiên ta mã hóa thành biến **PhanNhomTuoi** như đã hướng dẫn trong phần 2.2. Sau đó ta lập bảng tần số cho biến **PhanNhomTuoi** này, như hình sau:



Statistics

PhanNhomTuoi

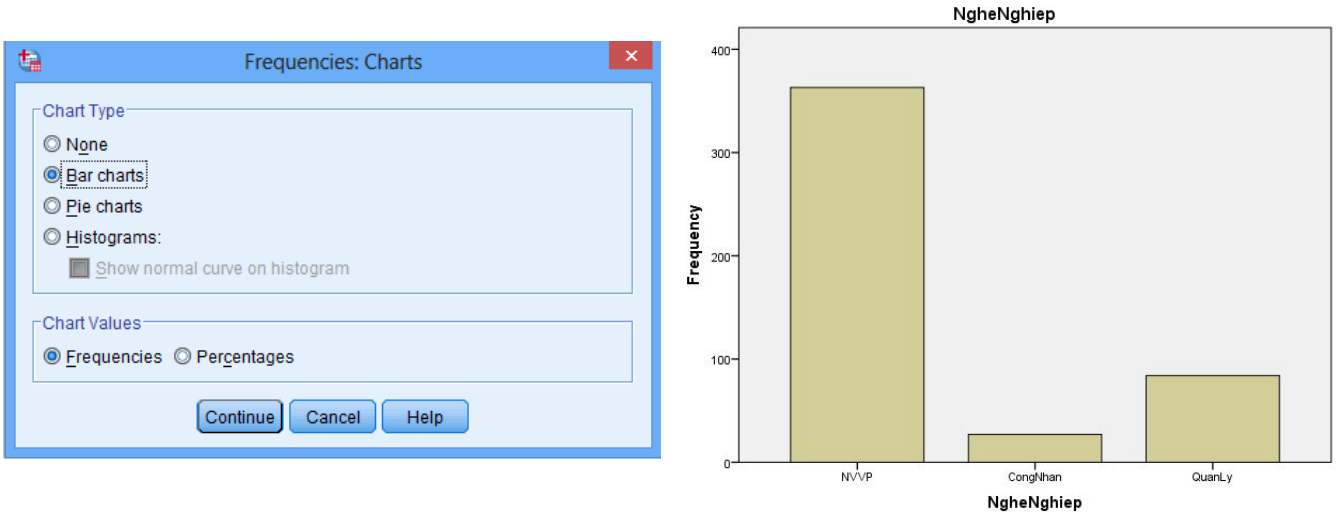
N	Valid	473
	Missing	1

PhanNhomTuoi

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<=25	173	36.5	36.6	36.6
	(25;35]	144	30.4	30.4	67.0
	(35;45]	63	13.3	13.3	80.3
	>45	93	19.6	19.7	100.0
	Total	473	99.8	100.0	
Missing	System	1	.2		
Total		474	100.0		

Hình 3.3: Lập bảng tần số, và kết quả trong Output cho biến PhanNhomTuoi

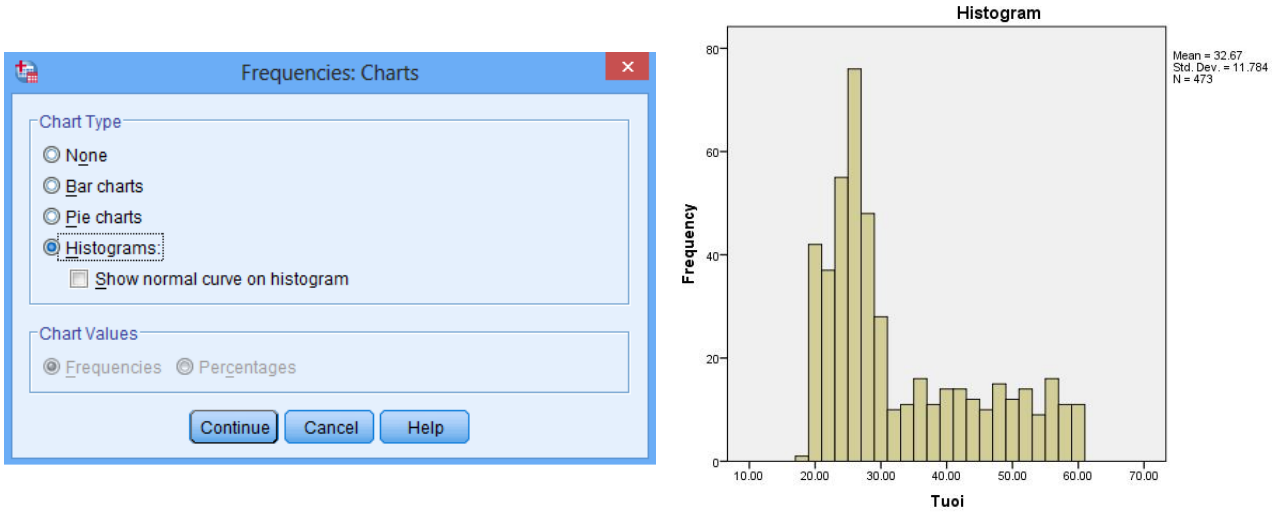
- **Biểu đồ:** Để lập biểu đồ thanh (Bar charts), biểu đồ tròn (Pie charts) và biểu đồ phân phối tần số cho một biến ta nhấp vào nút **Charts ...** và chọn biểu đồ hợp lí:
  - đối với biến định tính ta dùng biểu đồ thanh và tròn.
  - đối với biến định lượng nhiều biểu hiện ta dùng biểu đồ phân phối tần số (Histograms).



Hình 3.4: Lập biểu đồ thanh cho biến NgheNghiep và kết quả trong Output

Chẳng hạn, để lập biểu đồ thanh cho biến **NgheNghiep** ta làm như hình sau:

Để lập biểu đồ phân phối tần số cho biến Tuoi, ta lựa chọn biến này, và trong nút **Charts** ta chọn **Histograms**



Hình 3.5: Lập biểu đồ phân phối tần số cho biến Tuoi, và kết quả trong Output

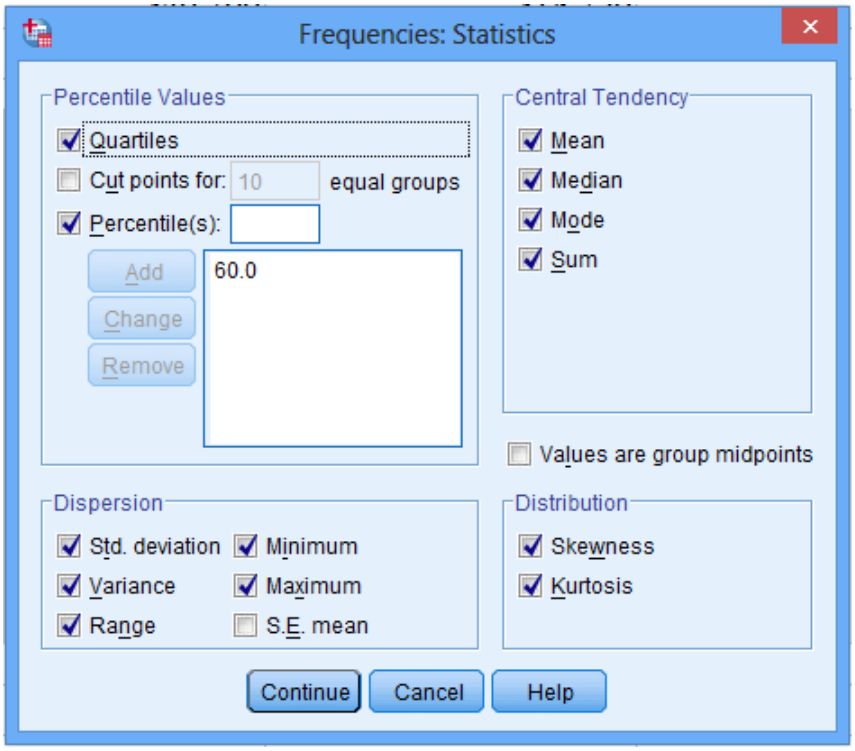
Lưu ý, muốn lập biểu đồ hộp và râu, tán xạ, thân và lá ... ta phải vào các nút menu khác, sẽ được trình bày dưới đây.

- **Các đại lượng thống kê mô tả:** Để tính các đại lượng thống kê mô tả (ở đây chỉ dùng cho biến định lượng), từ hộp thoại **Frequencies** trong hình 3.1 ta chọn biến cần tính, sau đó nhấp nút **Statistics ...** và tích chọn các mục:
  - Tính các đại lượng hướng tâm: trung bình (Mean), trung vị (Median), Mode ta lựa chọn trong khung **Cental Tendency**.
  - Tính các đại lượng đo mức độ phân tán: độ lệch chuẩn (Std. deviation), phương sai (Variance), sai số chuẩn (S.E. mean), khoảng biến thiên (Range), giá trị nhỏ nhất (Minium),

lớn nhất (Maximum) ta lựa chọn trong khung **Dispersion**.

- Tính các phân vị: tứ phân vị (Quartiles), các mức phân vị cách đều (Cut points for ...), phân vị cụ thể (Percentile(s)) ta chọn và điền thông số phân vị muốn tính vào khung **Percentile Values** rồi nhấn **Add**.
- Tính các chỉ số độ nhọn (Kurtosis), hệ số bất đối xứng (độ nghiêng)(Skewness) ta chọn trong mục **Distribution**

Chẳng hạn, ta chọn biến Tuoi, giả sử ta tích các lựa chọn như hình sau, kết quả trong Output ở bên phải.



Statistics

Tuoi

N	Valid	473
	Missing	1
Mean		32.6723
Median		27.0000
Mode		25.00[a]
Std. Deviation		11.78409
Variance		138.865
Skewness		.860
Std. Error of Skewness		.112
Kurtosis		-.570
Std. Error of Kurtosis		.224
Range		42.00
Minimum		18.00
Maximum		60.00
Sum		15454.00
Percentiles	25	24.0000
	50	27.0000
	60	30.0000
	75	41.5000

a. Multiple modes exist. The smallest value is shown

Hình 3.6: Lập biểu đồ phân phối tần số cho biến Tuoi, và kết quả trong Output

Bảng kết quả cho ta thông tin sau:

- Có tất cả 474 quan sát, trong đó 1 quan sát không có thông tin về tuổi.
- Nếu chỉ tính riêng trong 473 người có thông tin về tuổi thì: độ tuổi trung bình là 32.67; trung vị là 27 tuổi, có nghĩa là có không quá 50% số người có tuổi < 27 và có không quá 50 % có tuổi > 27; mode là 25, tức là số người 25 tuổi xuất hiện nhiều nhất, chú thích (a.) cho thấy thực ra dữ liệu về Tuoi có nhiều mode, và 25 là số nhỏ nhất trong các mode đó.
- Độ lệch chuẩn (Std. Deviation) của tuổi là 11.78, nó là căn bậc hai của phương sai (Variance) 138.865; khoảng biến thiên (Range) của tuổi = Maximum − Minimum = 60−18 = 42.
- Tứ phân vị là $Q_1 = 24, Q_2 = 27, Q_3 = 41.5$. Q_2 chính là trung vị nói ở trên. Phân vị 60 là 30 tuổi, nghĩa là có không quá 60% số người có tuổi < 30 và có không quá 40 % có tuổi > 30.
- Hệ số độ nghiêng là 0.86 mô tả cho mức độ lệch của dữ liệu tuổi, càng gần 0 dữ liệu càng cân xứng, càng dương càng lệch phải, càng âm càng lệch trái; hệ số độ nhọn là -0.57 mô

tả cho mức độ tập trung của các quan sát quanh trung tâm của tập dữ liệu trong mối quan hệ với hai đuôi, độ nhọn càng âm thì biểu đồ phân phối tần số của dữ liệu càng "bằng phẳng", càng dương thì hình dáng biểu đồ càng nhọn, dữ liệu rút ra từ tổng thể phân phối chuẩn thì độ nhọn gần bằng 0.

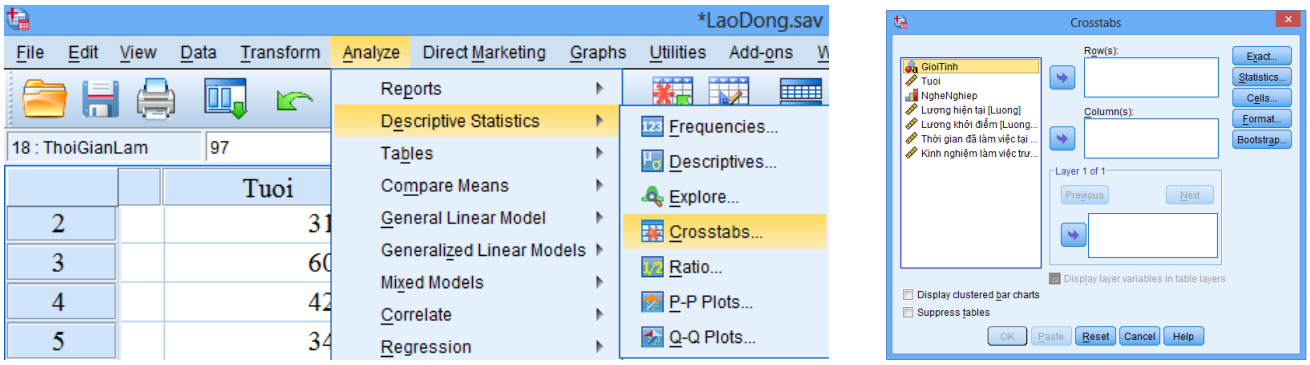
3.2. Bảng tần số chéo, biểu đồ theo nhóm, thống kê mô tả theo nhóm

SPSS có thể giúp tóm tắt dữ liệu dễ dàng một nhóm các giá trị của biến thuộc cùng một lớp đối tượng (được phân chia qua một biến khác). Chẳng hạn tính các đai lượng thống kê môt tả của biến tuổi theo nhóm nam và nhóm nữ riêng biệt.

3.2.1. Tạo bảng tần số chéo nhờ Crosstabs

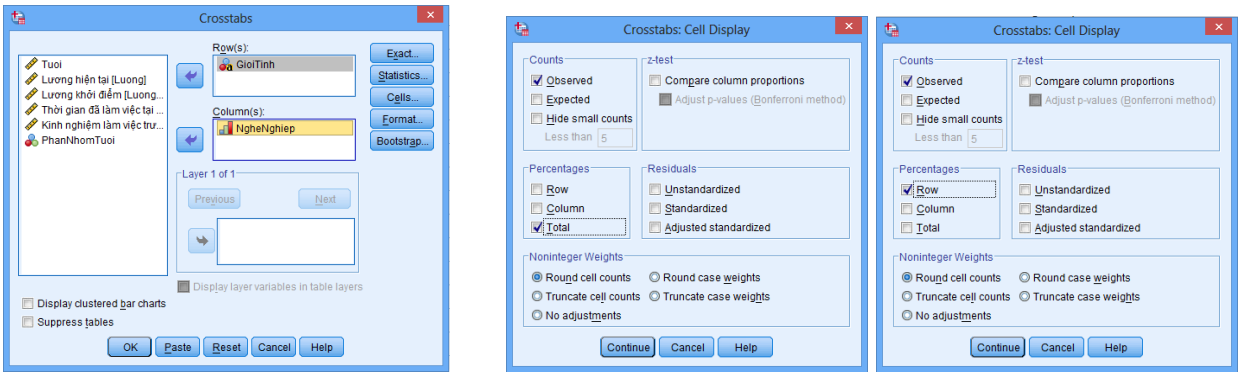
Để lập bảng tần số chéo giữa 2 biến định tính cách đơn giản nhất là ta vào **Analyze** ⟶ **Descriptive Satistics** ⟶ **Crosstabs** chọn biến dòng và cột trong bảng tần số và nhấp **OK**. Toàn bộ thông tin về tỉ lệ phần trăm sẽ được tích hợp trong bảng tần số. Chẳng hạn, để lập bảng tần số (cũng là tần suất, tần suất theo cột, theo hàng) cho giới tính (hàng) và nghề nghiệp (cột) trong file **LaoDong.sav** ta làm như sau:

- Vào **Analyze** ⟶ **Descriptive Satistics** ⟶ **Crosstabs**



Hình 3.7: Lập bảng tần số chéo nhờ Crosstabs

- Chọn biến **GioiTinh** và chuyển vào ô **Row(s)**, chọn biến **NgheNghiep** và chuyển vào ô **Column(s)**.



Hình 3.8: Lập bảng tần số chéo cho GioiTinh và NgheNghiep (hình trái); Lựa chọn Cells hai hình bên phải (chỉ làm theo 1 trong 2 hình này)

(Nói thêm, ở **Row(s)** và **Column(s)** có thể chọn nhiều biến. Khi đó SPSS sẽ ghép để tạo ra đủ các bảng tần số chéo ghép mỗi biến ở mục **Row(s)** với mỗi biến trong **Column(s)**.)

- (Bước này không cần làm nếu chỉ cần bảng tần số) Nhấp vào nút **Cells** trong danh sách lựa chọn bên phải của hộp thoại. Nhấp chọn trong khung **Percentages** (Row, Column, Total - lần lượt là hiển thị tần suất theo dòng, cột, và trên tổng số).

  – Nếu trong hộp thoại **Cell Display** ta chọn hiển thị 2 thông tin là **Observeb** và **Total**, nhấp **Continue** rồi nhấp **OK**, kết quả trong Output sẽ như sau: Qua bảng này ta thấy,

GioiTinh * NgheNghiep Crosstabulation

			NgheNghiep			
			NVVP	CongNhan	QuanLy	Total
GioiTinh	Nam	Count	157	27	74	258
		% of Total	33.1%	5.7%	15.6%	54.4%
	Nu	Count	206	0	10	216
		% of Total	43.5%	0.0%	2.1%	45.6%
Total		Count	363	27	84	474
		% of Total	76.6%	5.7%	17.7%	100.0%

Hình 3.9

chẳng hạn, có 157 NVVP giới tính nam, chiếm 33.1 % tổng số người được điều tra.

– Nếu trong hộp thoại **Cell Display** ta chọn 2 hiển thị 2 thông tin là **Observeb** và **Row**, nhấp **Continue** rồi nhấp **OK**, kết quả trong Output sẽ như sau: Qua bảng này ta thấy

GioiTinh * NgheNghiep Crosstabulation

			NgheNghiep			
			NVVP	CongNhan	QuanLy	Total
GioiTinh	Nam	Count	157	27	74	258
		% within GioiTinh	60.9%	10.5%	28.7%	100.0%
	Nu	Count	206	0	10	216
		% within GioiTinh	95.4%	0.0%	4.6%	100.0%
Total		Count	363	27	84	474
		% within GioiTinh	76.6%	5.7%	17.7%	100.0%

Hình 3.10

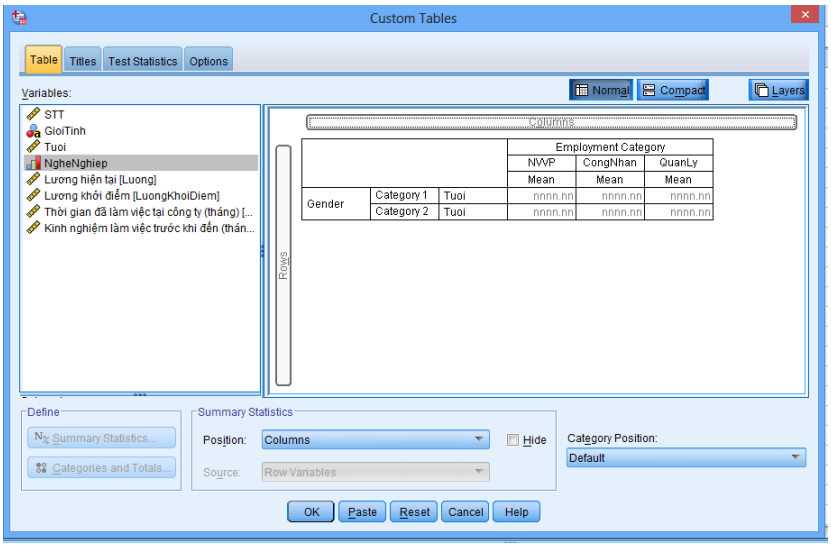
tổng % các dòng đều là 100%. Ngoài thông tin về tần số giống ở bảng phần trên thì các tần suất cho ta biết rằng: nếu chỉ xét trong nhóm nam thì có 60.9% là NVVP, 10.5% là CongNhan, 28.7% là QuanLy; nếu chỉ xét trong nhóm nữ thì có 95.4% là NVVP, 0% là CongNhan, 4.6% là QuanLy. Như vậy, tỉ lệ lao động là NVVP trong nhóm nữ là rất cao.

3.2.2. Tạo bảng tần số chéo với Custom Tables

Đây là một cách khác để lập một bảng tần số như trên, và còn hơn thế, nó có thể giúp lập một bảng tần số chéo nhiều tầng, nhiều lớp. Để sử dụng chức năng này ta vào mục **Analyze** ⟶ **Tables** ⟶

Custom Tables, nhấp **OK** trong hộp thoại con xuất hiện sẽ dẫn ta tới hộp thoại để lập bảng tần số chéo.

Chẳng hạn, làm các bước trên cho file **LaoDong.sav**, ta có hộp thoại sau cùng như sau:

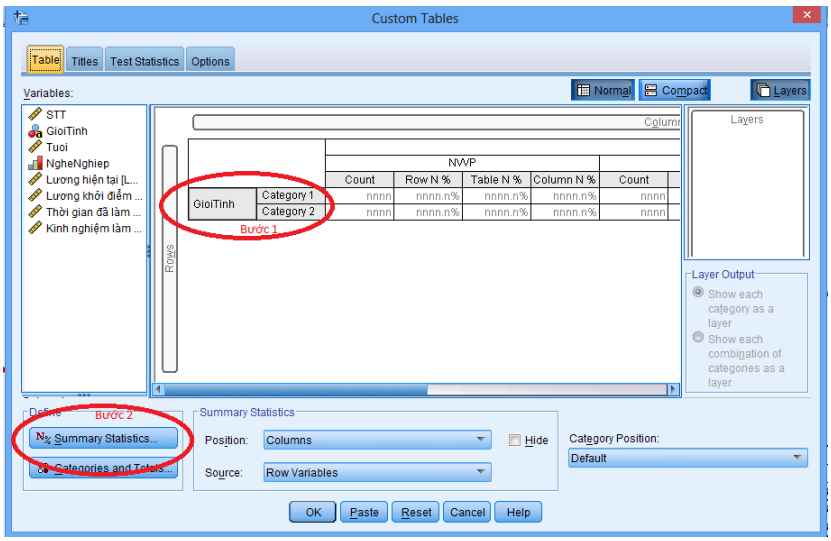


Hình 3.11: Hộp thoại Custom Tables

Để lập bảng tần số chéo của giới tính và nghề nghiệp ta nhấp chuột vào **GioiTinh**, kéo và thả vào thanh **Rows** ở khung bên phải. Kéo biến **NgheNghiep** vào thanh **Columns**. Sau đó nhấp **OK**. Bảng tần số chéo được tạo trong cửa sổ **Ouput**.

Để lập bảng tần số chéo của mức độ yêu nghề và giới tính trong đó mức yêu nghề lại được chia tiếp theo các nhóm nghề ta làm nhấp chuột vào NgheNghiep, kéo vào thanh **Rows**, kéo tiếp **MucYeuNghe** vào thanh **Rows**. Kéo biến GioiTinh vào thanh **Columns**. Sau đó nhấp **OK**. Bảng tần số chéo nhiều tầng được tạo trong cửa sổ Ouput.

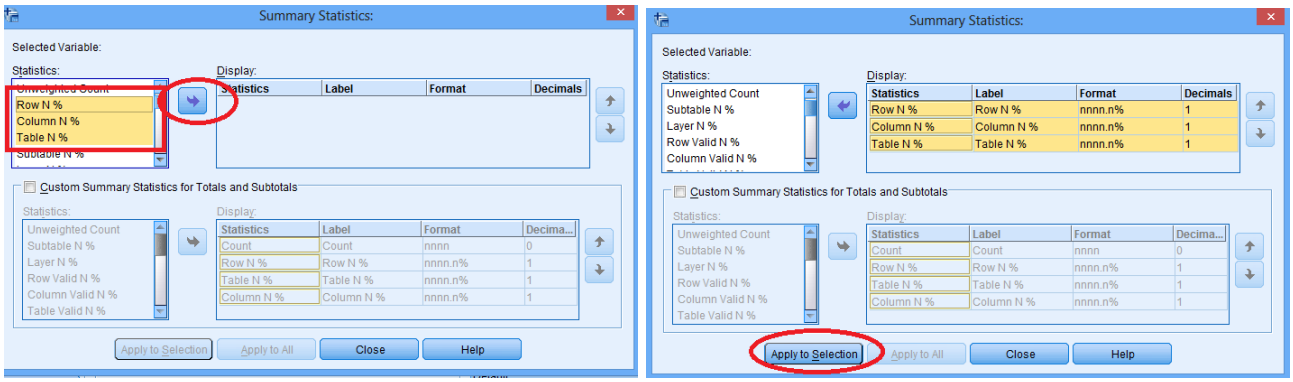
Nếu ta muốn thêm các thông số về tần suất (trên toàn bộ, trong nội bộ dòng, nội bộ cột) thì trước khi nhấp **OK** ở trên, ta nhấp vào GioiTinh trong bảng tần số xem trước. Sau đó nhấp vào **N % Summary Statistic** như sau:



Hình 3.12

Trong hộp thoại hiện ra sẽ có 3 khung, trong đó 2 khung phía trên giúp ta chọn danh sách những

thống kê muốn có trong bảng (chọn bằng nút mũi tên chuyển), sau đó nhấp **Apply Selection**. Hộp thoại đóng lại.



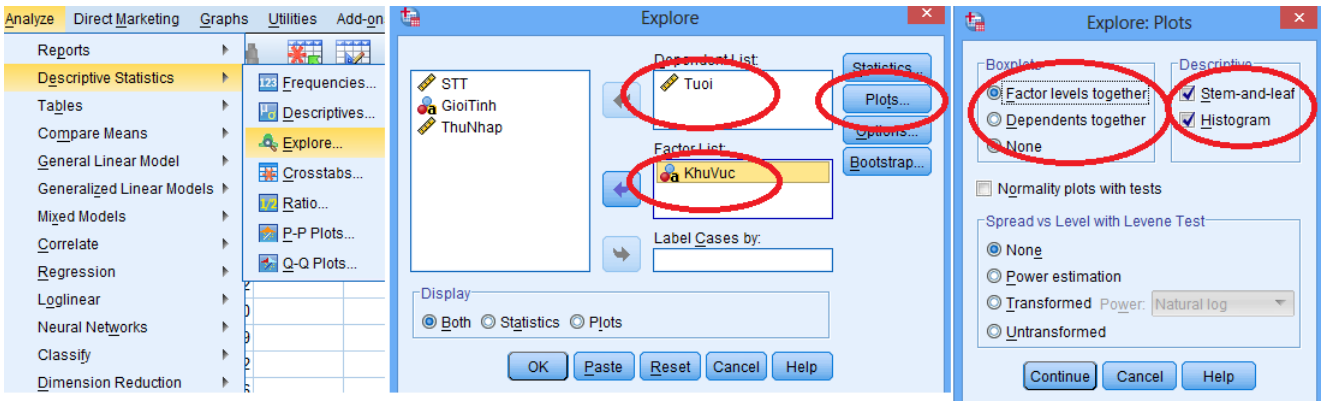
Hình 3.13: Các hộp thoại Summary Satistics khi ta lựa chọn từng đại lượng thống kê

Để hoàn thành ta nhấp **OK** và xem kết quả ở **Output**.

3.2.3. Biểu đồ theo nhóm, biểu đồ hộp và râu, thân và lá

Dọc file **SoLieu.csv**. Chẳng hạn, để lập biểu đồ hộp và râu, thân và lá, phân phối tần số của biến tuổi theo khu vực ta vào mục **Analyze** $\longrightarrow$ **Desciptive Statistics** $\longrightarrow$ **Explore**. Hộp thoại **Explore** xuất hiện gồm khung danh sách biến, khung biến phụ thuộc (Dependent List), khung phân loại (Factor List).

- Ta chuyển biến Tuoi qua khung **Dependent List** và KhuVuc qua khung **Factor List**.
- Mục **Display** cho ta lựa chọn hiển thị: Statistics (hiển thị thống kê), Plots (hiển thị biểu đồ), Both (hiển thị cả hai). Ta lựa chọn Both.
- Nhấp vào nút **Plots** được hộp thoại **Explore: Plots** xuất hiện. Trong hộp thoại này ta lựa chọn những biểu đồ mà ta muốn vẽ: boxplot là biểu đồ hộp và râu, Stem - and - leaf là biểu đồ thân và lá, Histogram là biểu đồ phân phối tần số.



Hình 3.14: Các hộp thoại trong Explore

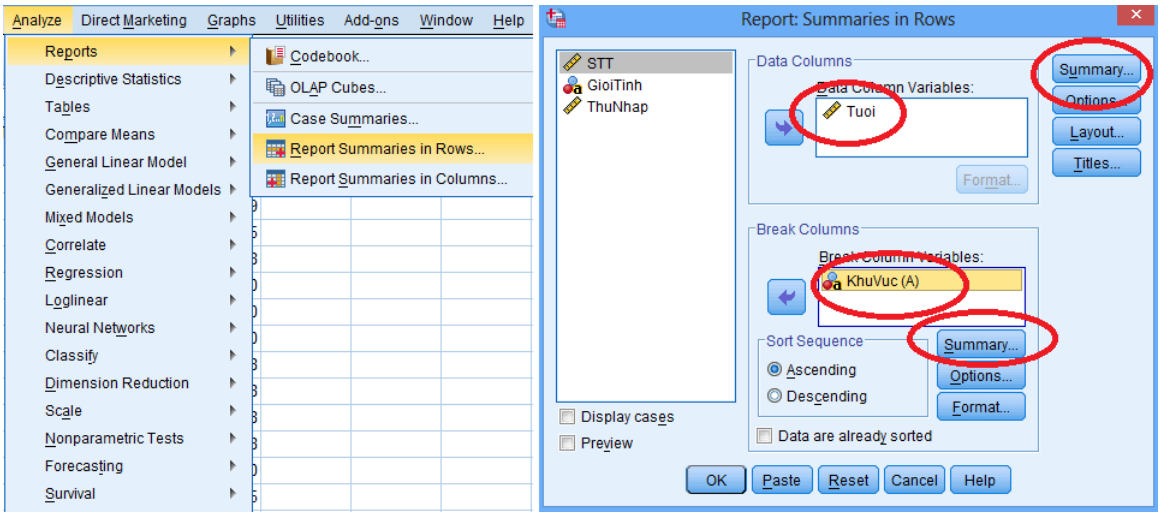
- Khi chọn xong nhấp **Continue**, sau đó **OK**. Và ta được kết quả trong cửa sổ Output.

Trong **Output** cho ta một loạt các biểu đồ về tuổi phân theo từng khu vực. Ngoài ra, do ta chọn hiển thị cả hai (both) nên trong bảng kết quả có cả những thông tin về các đại lượng thống kê mô tả (có thể chưa được đầy đủ như ý muốn) theo từng nhóm.

Lưu ý: Nếu trong khung **Factor List** ta không chọn biến phân loại thì phân tích vẫn được thực hiện cho biến Tuoi mà không phân nhóm. **Điều đó có nghĩa là qua đây ta có thể lập được biểu đồ hộp và râu, biểu đồ thân và lá cho một biến định lượng bất kì (không phân nhóm)**.

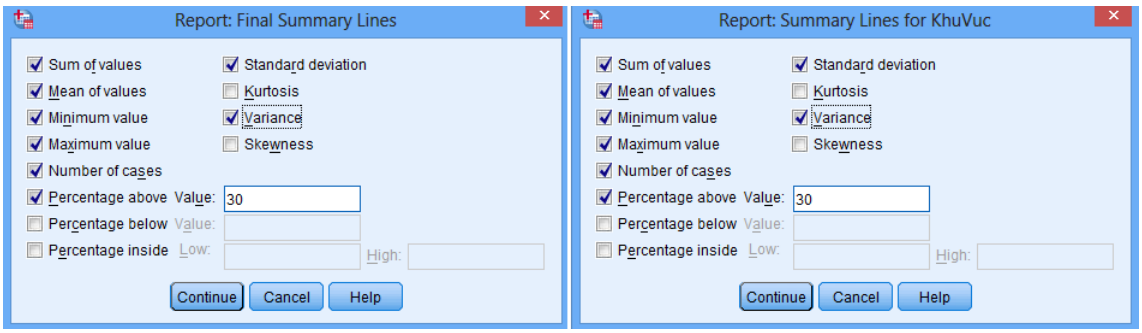
3.2.4. Phân tích tổng quan theo nhóm

Trong mục trên ta thấy rằng, khi ta lập biểu đồ theo nhóm cho một biến ta cũng có thể có được luôn thống kê mô tả theo nhóm. Ngoài cách trên ta còn có cách khác nữa để không những có được các đại lượng thống kê mô tả theo nhóm mà còn có cả đại lượng tổng quan khác nữa. Theo cách này, ta vào **Analyze** $\longrightarrow$ **Report Summaries in Row**



Hình 3.15: Các hộp thoại trong Report: Summaries in Rows

- Ta chuyển biến Tuoi qua khung **Data Column Variables**, biến KhuVuc qua khung **Break Column Variables**.
- Nhấp vào **Summary** để chọn các đại lượng thống kê muốn tính: Có hai nút như vậy, một dành cho biến tuổi (nút phía trên góc trái). Một nút dành cho các đại lượng muốn tính theo khu vực (nút bên dưới khung **Break Column Variables**).



Hình 3.16: Các hộp thoại Summary: bên trái là của biến Tuoi, bên phải là biến Tuoi phân nhóm theo KhuVuc

Lưu ý: *Trong hai hộp thoại Summary, có thêm lựa chọn tính phần trăm các giá trị lớn hơn, nhỏ một số nào đó, hoặc nằm trong một khoảng nào đó*. Lựa chọn này chúng ta cũng hay dùng về sau. Ở đây, trong hộp thoại thứ nhất chúng tôi điền giá trị 30 để tính tỉ lệ phần trăm các giá trị của biến tuổi > 30, hộp thoại thứ 2 cũng điền 30 để tính tỉ lệ phần trăm các giá trị > 30 xét trong nội bộ mỗi nhóm.

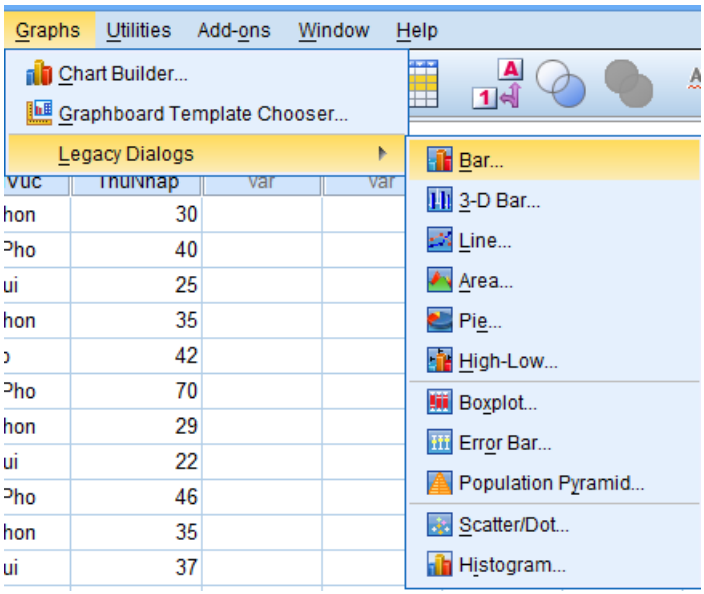
- Sau khi lựa chọn xong nhấp **Continue** và sau đó nhấp **OK**. Kết quả sẽ được hiển thị trong cửa sổ Output, bao gồm: Các đại lượng thống kê theo nhóm (khu vực) và các thống kê của biến (tuổi) không phân nhóm.

3.2.5. Lập các biểu đồ bằng nút menu Graph

Ở mục trên, chúng ta đã thực hành lập đa số các loại biểu đồ hay dùng. Ta thấy một bất tiện trong tất cả các cách lập biểu đồ ở trên là manh mún và thiếu tính định hướng, tức là phải vào bên trong mỗi thủ tục ta mới "phát hiện" ra các lựa chọn tạo biểu đồ.

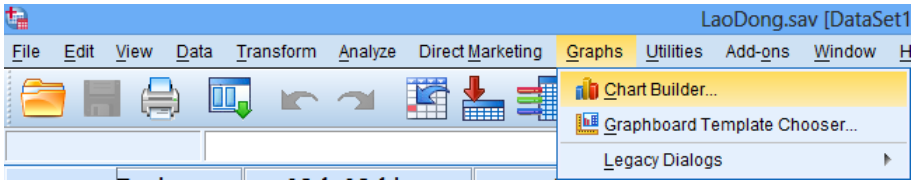
Nút **Graph** trên thanh menu chính giúp ta định nghĩa lập được hầu hết các biểu đồ (trừ biểu đồ thân và lá). Có hai cách sau để tạo biểu đồ từ nút này:

- Cách thứ nhất ta dùng các thiết kế có sẵn bằng cách vào **Graphs** $\longrightarrow$ **Legacy Dialogs** và lựa chọn biểu đồ muốn lập:



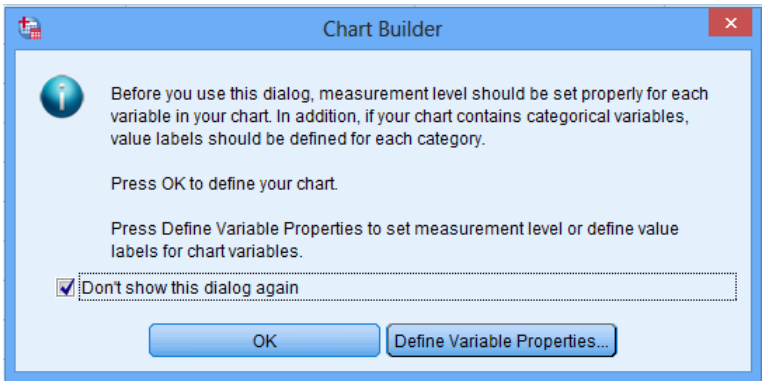
Hình 3.17: Lập biểu đồ bằng Legacy Dialogs

- Cách thứ hai, người dùng tự thiết kế biểu đồ, bằng cách vào **Graphs** $\longrightarrow$ **Chart Builder ...**,



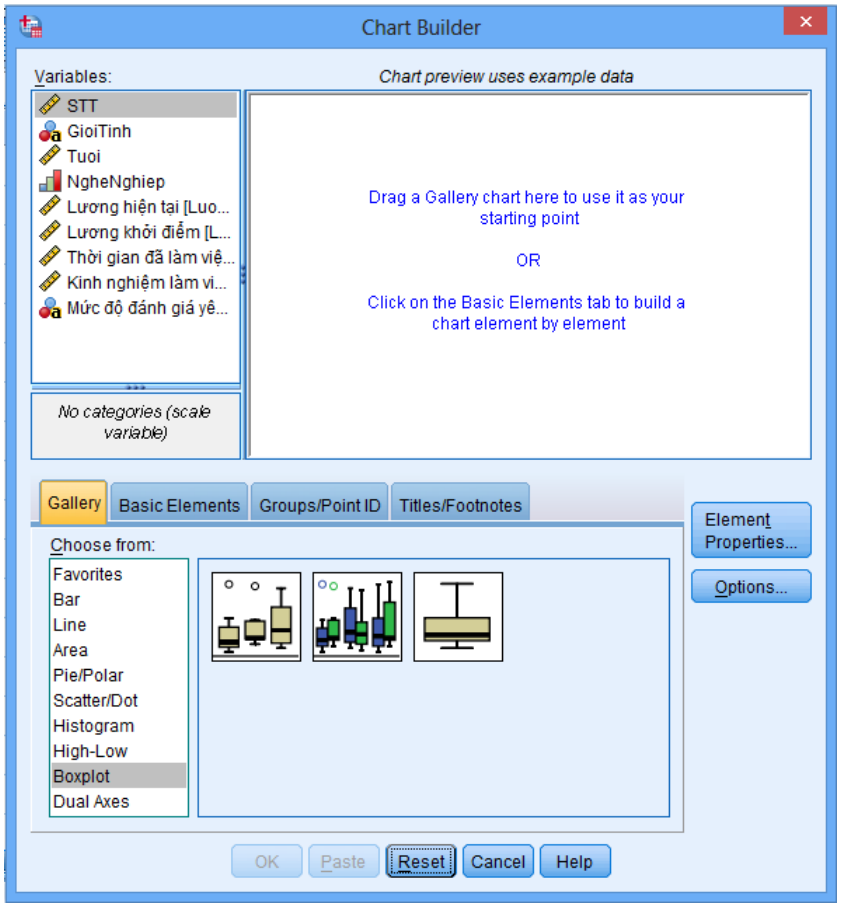
Hình 3.18: Lập biểu đồ bằng Chart Builder

Một hộp thoại con hiện ra, có thể tích chọn không hiển thị lại trong lần sau, sau đó ta nhấp **OK**:



Hình 3.19

Hộp thoại lựa chọn và xây dựng biểu đồ hiện ra

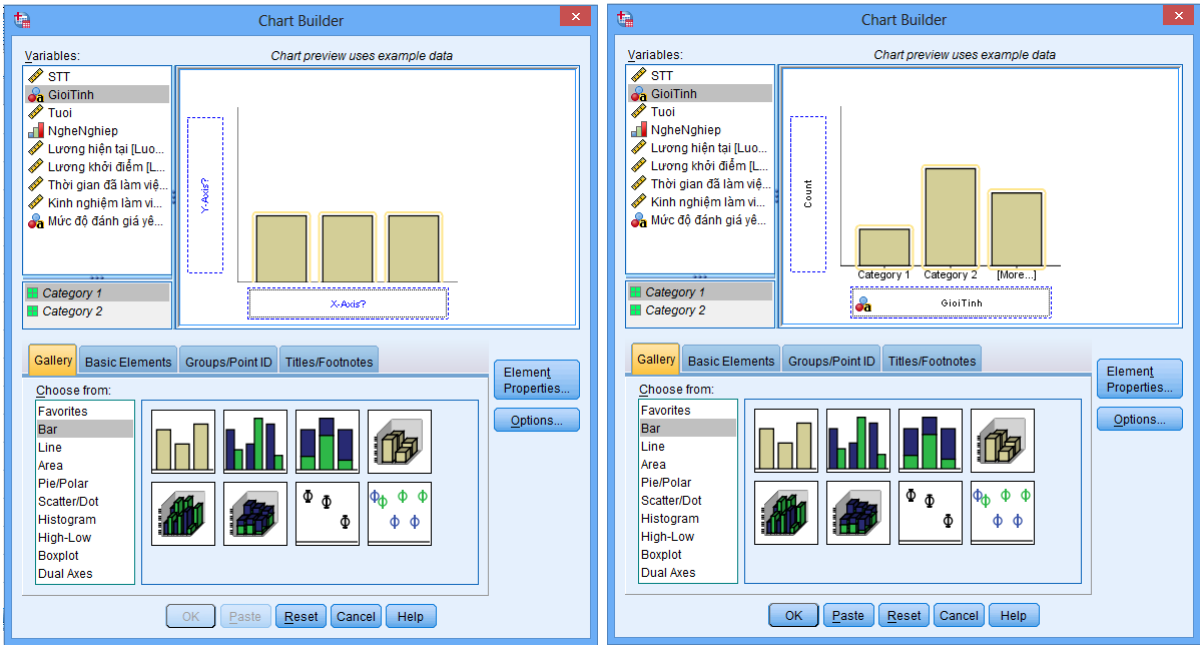


Hình 3.20: Hộp thoại tạo biểu đồ

Sau đây là một vài ví dụ minh họa lập biểu đồ theo cách thứ 2.

1. Lập biểu đồ thanh cho biến GioiTinh: Trong khung **Gallery** chọn **Bar**, khung bên phải hiện ra các lựa chọn, ta chọn dạng đầu tiên, nhấp đúp vào hình. Một hình dạng biểu đồ được hiện lên trên khung lớn phía trên.

   Nhấp vào **GioiTinh** kéo và thả vào khung **X - Axis?**



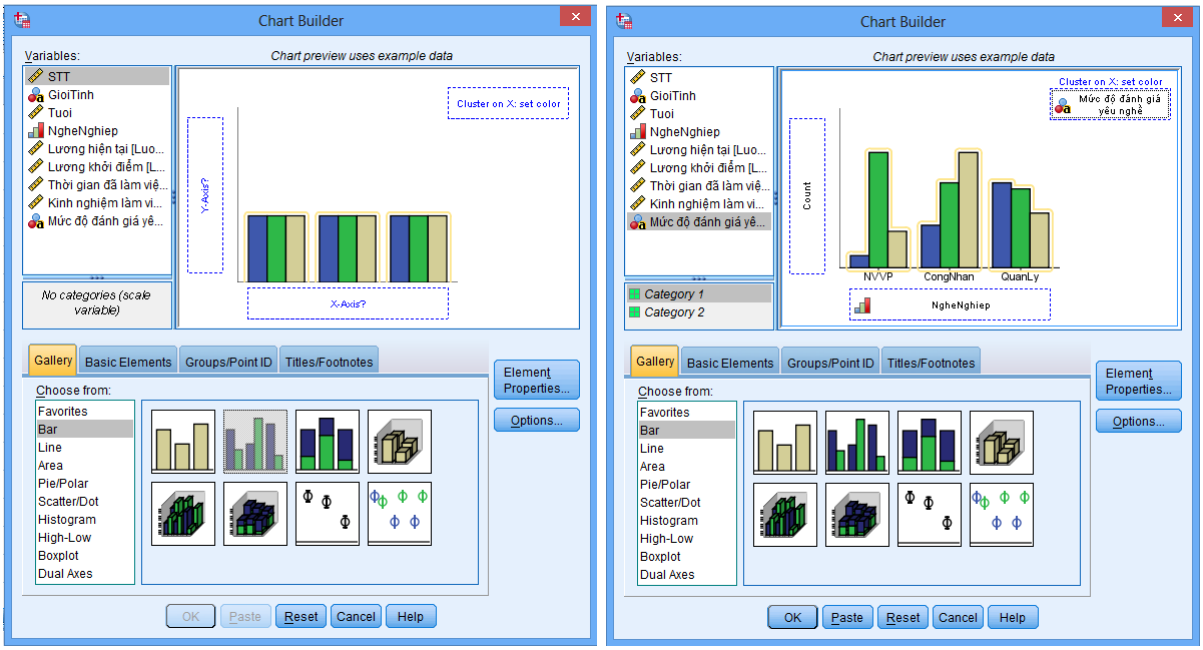
Hình 3.21: Tạo biểu đồ thanh cho GioiTinh

Cuối cùng nhấp **OK** ta được biểu đồ thanh của biến giới tính trong cửa sổ Output.

2. Lập biểu đồ thanh của **MucYeuNghe** theo **NgheNghiep**

   Trong khung **Gallery** chọn **Bar**, khung bên phải hiện ra các lựa chọn, ta chọn dạng thứ 2, nhấp đúp vào hình. Một hình dạng biểu đồ được hiện lên trên khung lớn phía trên. Lưu ý khung hình lớn có 2 khung nhỏ: X - Axis, Y - Axis và Cluster on X.

   Ta nhấp vào NgheNghiep, kéo và thả vào **X - Axis?**, nhấp MucYeuNghe (Mức độ đánh giá yêu nghề) kéo vào **Cluster on X: set color**

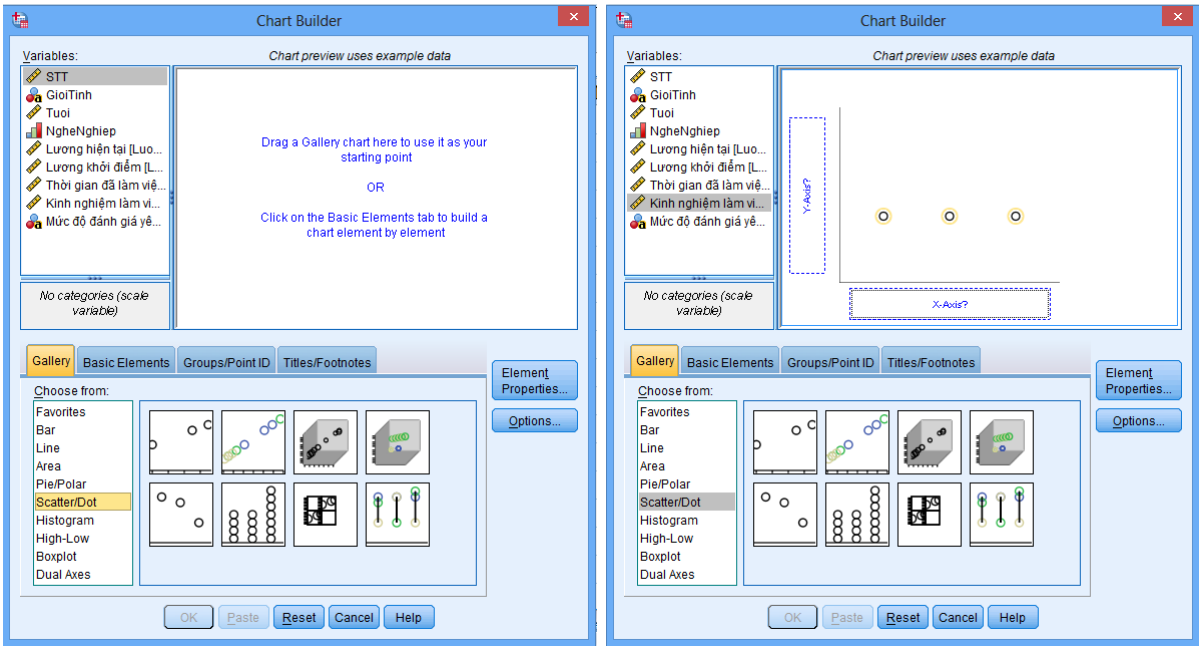


Hình 3.22: Tạo biểu đồ thanh cho MucYeuNghe theo NgheNghiep

Cuối cùng nhấp **OK** và xem kết quả ở Output.

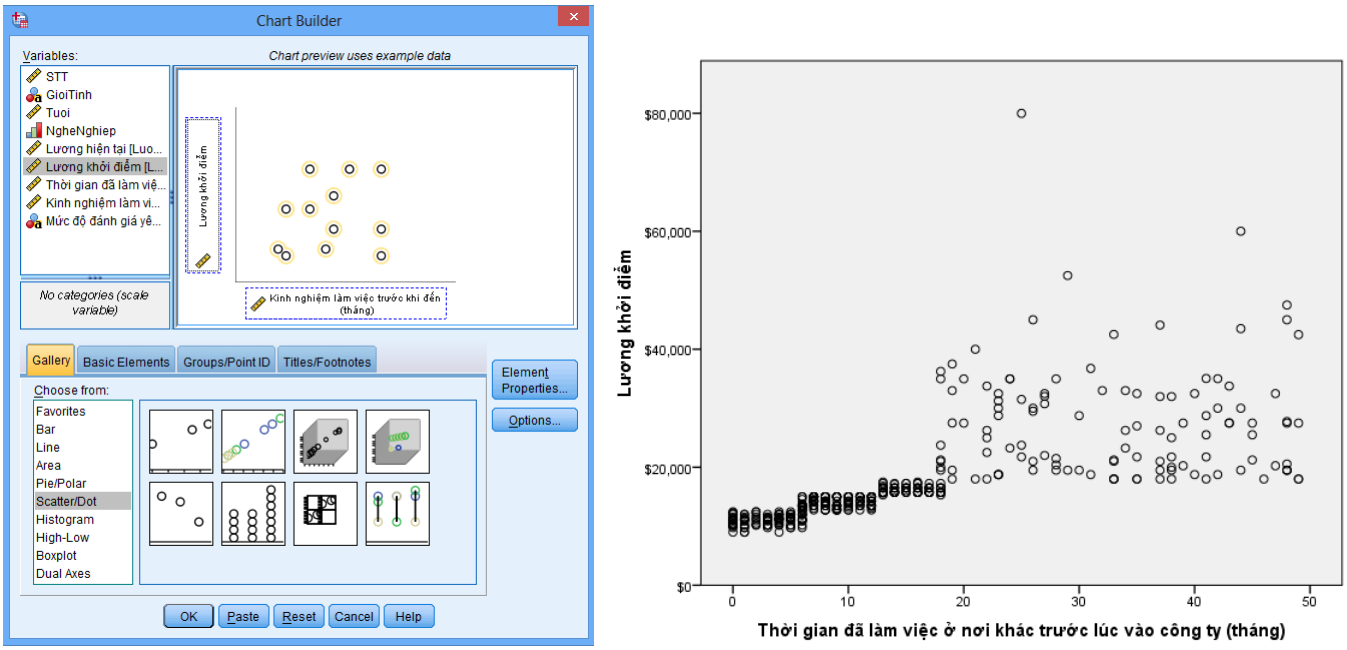
3. Lập biểu đồ tán xạ mô tả mối quan hệ giữa kinh nghiệm làm việc trước khi đến và lương khởi điểm.

   Trong khung **Gallery** chọn **Scatter/Dot**, khung bên phải hiện ra các lựa chọn, ta chọn dạng đầu tiên, nhấp đúp vào hình. Một hình dạng biểu đồ được hiện lên trên khung lớn phía trên. Lưu ý khung hình lớn có 3 khung nhỏ: X - Axis và Y - Axis.



Hình 3.23: Tạo biểu đồ tán xạ

Ta kéo Kinh nghiệm làm việc vào X - Axis, Lương khởi điểm và Y - Axis. Nhấp **OK**. Và ta được kết quả là biểu đồ như hình dưới đây:



Hình 3.24: Biểu đồ tán xạ

Qua biểu đồ ta thấy đối với tập dữ liệu này: kinh nghiệm làm việc trước khi chuyển đến có ảnh hưởng đến lương khởi điểm: kinh nghiệm làm việc nhiều hơn lương khởi điểm có phần cao hơn.

Trên đây là một số ví dụ về lập biểu đồ bằng **Graph**. Việc sử dụng có lẽ không phải là khó, đơn giản chỉ là chọn lựa biểu đồ và kéo thả các biến sao cho phù hợp với nhu cầu cần lập. Ngoài những biểu đồ trên, chức năng **Chart Builder** còn cho phép lập nhiều biểu đồ khác như: biểu đồ tròn (Pie/Polar), hộp và râu (Boxplot), phân phối tần số (Histogram), đa giác tần số (Line), ... mỗi biểu đồ lại bao gồm nhiều sự lựa chọn rất phong phú. Sau cùng, nếu muốn điều chỉnh chi tiết hơn có thể tìm hiểu ở hộp thoại **Element Properties** cho phép điều chỉnh từng chi tiết trên biểu đồ ...

3.3. Bài tập

Bài tập 3.1. File **LaoDong.sav** chứa thông tin về 474 lao động.

1. Xác định loại dữ liệu (định tính hay định lượng) và thang đo mỗi cột trong file dữ liệu.
2. Lập bảng tần số cho giới tính. Hãy tìm mode cho cột đó. Giá trị đó cho chúng ta thông tin gì?
3. Biểu đồ gì mô tả thông tin về phân phối tần số cho cột dân tộc. Hãy vẽ biểu đồ đó. Thông tin từ biểu đó là gì?
4. Lập bảng tần số chéo cho giới tính và dân tộc. Tính tỷ lệ phần trăm của từng giới tính theo dân tộc, từng dân tộc theo giới tính
5. Vẽ biểu đồ thanh của giới tính theo từng dân tộc.
6. Tính các số đo hướng tâm trung bình, trung vị, mode cho cột lương. Nêu ý nghĩa của giá trị trung vị.
7. Bạn dùng biểu đồ gì mô tả phân phối tần số cho cột lương? Hãy vẽ biểu đồ đó và nhận xét.
8. Tính tứ phân vị, phân vị thứ 90 cho cột lương hiện tại. Nêu ý nghĩa các con số đó. Vẽ biểu đồ hộp và râu cho biến này và cho nhận xét.
9. Hãy tóm tắt các đại lượng thống kê mô tả của lương khởi điểm theo từng nhóm nghề nghiệp. Nhóm nào có lương khởi điểm trung bình cao nhất?
10. Tính xem có bao nhiêu phần trăm lao động được điều tra có lương hiện tại lớn hơn 80000, bao nhiêu nằm trong khoảng từ 40000 đến 60000? Tính các tỉ lệ này theo mỗi nhóm nghề nghiệp.
11. Phân tổ cột lương hiện tại thành các mức: 1. ≤ 25000; 2.$(2500 - 45000]$; 3. Trên 4500. Và lập bảng tần số cho cách phân tổ đó. Cho biết tỉ lệ những lao động có lương không quá 45000.

Bài tập 3.2. Dùng file **Toi pham xa hoi.sav** (ghi lại thông tin về số tội phạm ở một số địa phương.

1. Tính tứ phân vị cho cột tội phạm. Nêu ý nghĩa của những con số đó.
2. Vẽ biểu đồ hộp và râu cho cột tội phạm. Nhận xét.
3. Hãy khảo sát (qua việc tính các đại lương thống kê mô tả) về số án mạng ở các địa phương.
4. Tóm tắt các số đo thống kê của tội phạm theo các địa phương mà có dân số ở đô thị lớn hơn 80% (hướng dẫn: lọc ra dữ liệu chỉ gồm các địa phương có dân số đô thị $> 80\%$ và tính toán trên dữ liệu này).

5. Tính tỷ lệ phần trăm các địa phương có số tội phạm lớn hơn giá trị trung bình.

Bài tập 3.3. File **ThamDoBenhVien.sav** ghi lại thông tin về mức độ hài lòng của các bệnh nhân đối với một cơ sở y tế. Hãy dùng file dữ liệu đó để trả lời các câu hỏi sau:

1. Hãy lập bảng tần số, tần suất cho các biến giới tính, tình trạng hôn nhân, phương tiện đi đến cơ sở y tế. Bảng đó cho chúng ta thông tin gì?
2. Lập biểu đồ thanh, tròn cho các biến GioiTinh, PhuongTien, KhuVuc. Nhận xét.
3. Tính các số đo hướng tâm cho cột điểm hài lòng: trung bình, trung vị, mode. Thông tin có được từ các giá trị đó?
4. Tính tứ phân vị cho cột điểm hài lòng. Nêu ý nghĩa của các giá trị đó.
5. Hãy tính các giá trị trung bình, trung vị, độ lệch chuẩn của điểm hài lòng theo mỗi nhóm: nhóm bệnh nhân đến lần đầu và nhóm đến nhiều hơn một lần. Nhận xét.
6. Vẽ biểu đồ hộp và râu cho điểm hài lòng. Nhận xét.
7. Vẽ biểu đồ tán xạ mô tả mối quan hệ giữa tuổi và điểm hài lòng. Nhận xét.
8. Vẽ biểu đồ tán xạ mô tả mối quan hệ giữa nhiệt độ cơ thể bệnh nhân khi nhập viện với điểm hài lòng. Nhận xét.
9. Tính tỉ lệ bệnh nhân có nhiệt độ trong khoảng 38°C đến 40°C (Lưu ý, độ C = (độ F - 32)/1.8).

Bài tập 3.4. Trong file dữ liệu có tên là **SoLieu.csv** chứa một số thông tin cá nhân của 100 người về giới tính (GioiTinh), tuổi (Tuoi), khu vực sống (KhuVuc) và tổng thu nhập (đơn vị triệu VND) trong năm qua (ThuNhap). Hãy lấy file dữ liệu và thực hiện các yêu cầu sau:

1. Lập bảng tần số, tần suất chéo cho GioiTinh và KhuVuc. Từ bảng đó thu được thông tin gì?
2. Trong số nữ được điều tra, hãy tính tỉ lệ nữ sống ở thành phố.
3. Tiến hành phân tổ cột dữ liệu về tuổi thành các tổ với các điểm chia là 20; 30; 40; 50; 60 (tạo ra biến mới là mã hóa của tuổi, tên là PhanToTuoi). Lập bảng tần số cho biến này và tính tỉ lệ những người được điều tra có độ tuổi không vượt quá 50.
4. Tiến hành phân tổ cột dữ liệu về thu nhập thành các tổ với các điểm chia là 20; 40; 60; 80; 100 (tạo ra biến mới là mã hóa của thu nhập, tên là PhanToTN). Lập bảng tần suất tích lũy và cho biết có bao nhiêu phần trăm người được điều tra có thu nhập trên 60.
5. Bằng cách lập bảng tần suất chéo giữa hai biến phân tổ nói trên, hãy tính tỉ lệ những người có thu nhập hơn 80 triệu nằm từ độ tuổi từ trên 40 đến 50.

Chương 4

Xác suất và biến ngẫu nhiên

4.1. Xác suất căn bản

***Ví dụ* 4.1.1.** *Tại một xã ở vùng cao phía bắc có 60 % hộ gia đình có xe máy, 80 % hộ gia đình có ti vi, trong đó có 50% các hộ là có cả xe máy và ti vi. Chọn ngẫu nhiên một hộ ở xã trên, tính xác suất để hộ đó có ít nhất ti vi hoặc xe máy.*

Lời giải: Ta gọi A là biến cố "hộ được chọn có tivi", B là biến cố "hộ được chọn có xe máy". Khi đó AB là biến cố "hộ được chọn có cả ti vi và xe máy".

Theo khảo sát ta có $P(A) = 0.8, P(B) = 0.6, P(AB) = 0.5$. Ta cần tính $P(A + B)$.

Theo công thức cộng xác suất ta có $P(A + B) = P(A) + P(B) - P(AB) = 0.8 + 0.6 - 0.5 = 0.9$. Vậy xác suất để hộ được chọn có ít nhất ti vi hoặc xe máy là 0.9.

***Ví dụ* 4.1.2.** *Giả sử một cuộc khảo sát về mức độ hài lòng của 500 người dân (phân thành 2 nhóm: nhóm đến lần đầu và nhóm đã đến nhiều hơn một lần) với thái độ cán bộ của Ủy ban nhân dân tại một quận cho ta bảng sau:*

	Hài lòng	*Không hài lòng lắm*	*Bức xúc*
Lần đầu	*112*	*95*	*4*
Nhiều lần	*100*	*180*	*9*

Chọn ngẫu nhiên một trong số 500 người trên.

1. *Tính xác suất của biến cố: người được chọn đến ủy ban quận lần đầu và hài lòng với thái độ của các cán bộ.*
2. *Tính xác suất để người được chọn "không hài lòng lắm" với thái độ của các cán bộ quận.*
3. *Tính xác suất để người được chọn không hài lòng với thái độ của các cán bộ quận.*
4. *Trong các biến cố "Người được chọn hài lòng", "Người được chọn bức xúc", "Người được chọn từng đến nhiều lần" có những cặp biến cố nào xung khắc? Vì sao?*
5. *Biến cố "người được chọn hài lòng" và biến cố "người được chọn đến ủy ban quận lần đầu" có độc lập nhau hay không?*

Lời giải:

1. Trong số 500 người được khảo sát, có 112 người đến lần đầu và hài lòng với thái độ của các cán bộ tại đây, nên xác suất để người được chọn đến lần đầu và hài lòng là $\frac{112}{500} = 0.224$.

2. Trong 500 người có $95 + 180 = 275$ người không hài lòng lắm với thái độ của cán bộ, vậy xác suất để người được chọn thuộc nhóm không hài lòng lắm là $\frac{275}{500} = 0.55$.

3. Những người "không hài lòng" bao gồm những người "không hài lòng lắm" và những người "bức xúc". Có tất cả 288 người như vậy trong nhóm được điều tra. Vậy xác suất để người được chọn không hài lòng là $\frac{288}{500} = 0.576$.

4. Hai biến cố "Người được chọn hài lòng" và "Người được chọn bức xúc" xung khắc nhau, vì không có ai thuộc cả hai nhóm "Hài lòng" và "Bức xúc".

   Hai biến cố "Người được chọn hài lòng", "Người được chọn từng đến nhiều lần" không xung khắc, vì có 100 người có cả hai tính chất này, tức là biến cố "Người được chọn từng đến nhiều lần và hài lòng" là khác rỗng.

5. Goi A là biến cố "người được chọn hài lòng" và B là biến cố "người được chọn đến ủy ban quận lần đầu". Ta có

$$P(A) = \frac{212}{500} = 0.424, P(B) = \frac{112 + 95 + 4}{500} = 0.422, P(AB) = \frac{112}{500} = 0.224$$

   Ta tính được $P(A).P(B) = 0.178928 \neq 0.224$. Do đó A và B không độc lập.

4.2. Biến ngẫu nhiên

***Ví dụ* 4.2.1.** *Một thống kê về đời sống tâm lí trên 400 hộ gia đình tại một quận cho ta bảng sau đây về số người có vấn đề về tâm lí của các gia đình:*

Số người có vấn đề về tâm lí	*0*	*1*	*2*
Số gia đình	*348*	*43*	*9*

1. *Chọn ngẫu nhiên 1 gia đình trong số 400 gia đình trên. Gọi X là số người trong gia đình có vấn đề về tâm lí. Lập bảng phân phối xác suất của X.*

2. *Tính $P(X \geq 1), E(X), VX$.*

Lời giải:

1. Bảng phân phối xác suất của X là:

X	0	1	2
P	0.87	0.1075	0.0225

2. Từ bảng phân phối xác suất của X ta có $P(X \geq 1) = P(X = 1) + P(X = 2) = 0.1075 + 0.0225 = 0.13$.

   $E(X) = 0.87 \times 0 + 0.1075 \times 1 + 0.0225 \times 2 = 0.1525$.

   $V(X) = 0.87 \times (0 - 0.1525)^2 + 0.1075 \times (1 - 0.1525)^2 + 0.0225 \times (2 - 0.1525)^2 = 0.1742438$.

***Ví dụ* 4.2.2.** *Cho X là biến ngẫu nhiên có phân phối nhị thức với $n = 18, p = 0.6$. Tính:*

1. *$P(X = 10)$*

2. *$P(X \leq 9), P(X > 8), P(X \geq 10), P(X < 15), P(5 < X \leq 15)$.*

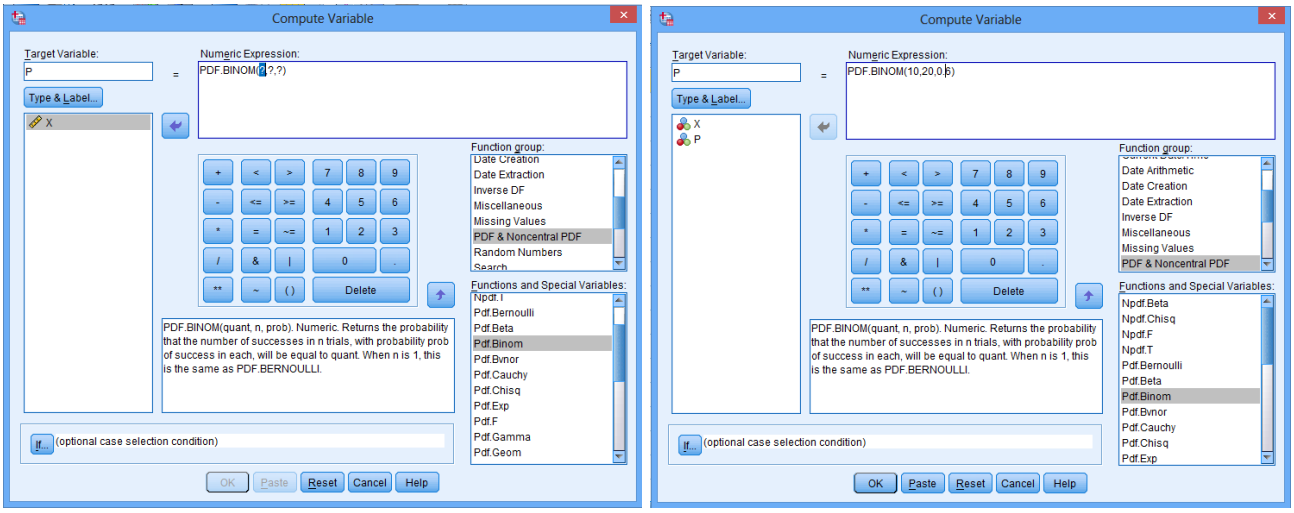
3. *Tìm giá trị của X có xác suất lớn nhất.*

4. *Tính* $E(X), V(X)$.

Lời giải:

1. Ta có thể tính các xác suất trên bằng công thức xác suất của biến ngẫu nhiên nhị thức. Ở đây chúng ta sẽ tìm hiểu cách tính các xác suất trên bằng SPSS. Sau đây là cách tính $P(X = 10)$.

   - Khởi động SPSS và tạo ra một biến tên là X (tên tùy thích).
   - Điền cho X một giá trị nào đó.
   - Trên thanh Menu vào **Transform** $\longrightarrow$ **Compute Variable...** hộp thoại **Compute Variable** hiện ra. Trong hộp thoại này ta điền **P** vào khung **Tanget Variable** (chỉ là để đặt tên), trong khung bên cạnh bảng tính "dò" đến hàm **PDF & Noncentral PDF**. Khi đó khung bên dưới hiện ra các hàm tính xác suất của một số phân phối phổ biến, ta chọn nhấp đúp vào **PDF. Binom**. Khung chính sẽ hiện **PDF.BINOM(?,?,?)**.



Hình 4.1: Lựa chọn hàm tính xác suất của biến ngẫu nhiên nhị thức

   - Ta thay thế 3 dấu ? bởi lần lượt: 10, 18, 0.6. Nhấp **OK** và ta có kết quả ở cửa sổ **Data View** cho ta $P(X = 10) = 0.1734$ (chỉnh thuộc tính của P trong cửa sổ Variable View để được nhiều số sau dấu "," hơn, mặc định ban đầu là 2 số).

     Lưu ý rằng các thao tác "dò hàm" có thể được thay thế bằng cách gõ trực tiếp dòng lệnh **PDF.BINOM(10,18,0.6)** *vào thẳng khung chính.*

   - Để tính $P(X \leq 9)$, ta làm tương tự như trên nhưng ta dò tới **CDF & Noncentral PDF** ở khung bên phải bàn tính, sau đó chọn **CDF. Binom** ở bên khung dưới. Câu lệnh trong khung chính hiện ra **CDF.BINOM(?,?,?)** ta thay thành **CDF.BINOM(9,18,0.6)**.

     Ta cũng có thể gõ trực tiếp vào khung hình chính câu lệnh **CDF.BINOM(9,18,0.6)**.

     Sau cùng nhấp **OK**. Kết quả ta được $P(X \leq 9) = 0.2631588$.

     Để tính $P(X > 8)$ ta phân tích: $P(X > 8) = 1 - P(X \leq 8)$. Do đó trong hộp thoại **Compute Variable** ta gõ câu lệnh 1−**CDF.BINOM(8,18,0.6)**. Kết quả cho ta $P(X > 8) = 0.86528585$.

Để tính $P(X \geq 10)$ ta phân tích: $P(X \geq 10) = 1 - P(X \leq 9)$ (Do X lấy giá trị thuộc tập $\mathbb{N}$). Do đó trong hộp thoại **Compute Variable** ta gõ câu lệnh $1-$**CDF.BINOM(9,18,0.6)**. Kết quả cho ta $P(X \geq 10) = 0.736841$.

Để tính $P(X < 15)$ ta phân tích: $P(X < 15) = P(X \leq 14)$ (Do X lấy giá trị thuộc tập $\mathbb{N}$).

Để tính $P(5 < X \leq 15)$ ta phân tích: $P(5 < X \leq 15) = P(X \leq 15) - P(X \leq 5)$ (Do X lấy giá trị thuộc tập $\mathbb{N}$). Do đó trong hộp thoại **Compute Variable** ta gõ câu lệnh **CDF.BINOM(15,18,0.6)-CDF.BINOM(5,18,0.6)**. Kết quả cho ta 0.986023.

2. Để xem xác suất X bằng bao nhiêu là lớn nhất, ta có thể làm như sau, đầu tiên ta nhập giá trị cho cột biến X từ 0 đến 18. Sau đó vào **Transform** $\longrightarrow$ **Compute Variable...**, trong hộp thoại **Compute Variable** hiện ra ta gõ vào khung chính lệnh **PDF.BINOM(X,18,0.6)**. Nhấp **OK**. Kết quả trong **Data Variable** cho ta danh sách xác suất ứng với các giá trị của cột X. Ta **sort** cột **P** theo chiều giảm dần, dòng đầu tiên sẽ là giá trị của X ứng với giá trị lớn nhất của p và xác suất lớn nhất tương ứng: $X = 11, P = 0.18916$

3. Ta có $E(X) = n \times p = 18 \times 0.6 = 10.8, V(X) = n \times p \times (1-p) = 4.32$.

***Ví dụ* 4.2.3.** *Việt Nam nằm trong top 20 thế giới về tỉ lệ người dùng Internet. Tính đến 31/3/2012 có 34% dân số ở nước ta dùng Internet. Giả sử rằng vào thời điểm đó:*

1. *ta chọn ngẫu nhiên 2 người. Gọi X là số người dùng Internet trong 2 người này. Hãy lập bảng phân phối xác suất cho X. Tính kì vọng và phương sai của X.*

2. *ta chọn 10 người. Gọi Y là số người dùng Internet trong số được chọn. Hãy lập bảng phân phối xác suất của Y, tính số người dùng trung bình trong 10 người đó và cho biết khả năng có bao nhiêu người dùng là lớn nhất.*

Lời giải:

1. Ta thấy rằng X chỉ có thể nhận các giá trị: 0, 1, 2. Do số lượng người ta chọn (2 người) là rất nhỏ so với tổng thể (khoảng 80 triệu) nên xác suất người đầu dùng internet là 0.34, xác suất người sau dùng internet sau khi đã chọn người đầu vẫn có thể coi là 0.34 (việc bỏ 1 người ra khỏi gần 80 triệu người có thể coi là không làm ảnh hưởng đến xác suất chọn được người dùng internet). Do vậy, việc chọn 2 người ta coi như 2 thực hiện 2 phép thử, khả năng mỗi lần thử người được chọn có dùng internet đều là 0.34, do đó X tuân theo phân phối nhị thức với $n = 2, p = 0.34$.

   Bằng cách tính xác suất ở ví dụ 4.2.2 ta tính được $P(X = 0) = 0.4356, P(X = 1) = 0.4488, P(X = 2) = 0.1156$.

2. Tương tự, ta có Y tuân theo phân phối nhị thức với $n = 10, p = 0.34$. Bảng phân phối xác suất của Y là

Y	0	1	2	3	4	5	6	7	8	9	10
P	0.0157	0.0808	0.1873	0.2573	0.232	0.143389	0.0616	0.0181	0.0035	0.0004	0.00002

Số người dùng internet trung bình trong 10 người chính là trung bình của X là $E(X) = 10 \times 0.34 = 3.4$.

Từ bảng phân phối xác suất ta thấy rằng xác suất để $Y = 3$ là lớn nhất. Tức là khả năng có 3 người dùng internet là cao nhất.

Như vậy, qua đây ta thấy rằng số người dùng trung bình và số người dùng internet có khả năng cao nhất trong những người được chọn là hai giá trị khác nhau (nhiều sinh viên nhầm lẫn 2 đại lượng này là một!).

***Ví dụ* 4.2.4.** *Cho X là biến ngẫu nhiên có phân phối chuẩn với trung bình 100, độ lệch chuẩn là 8. Hãy tính:*

1. $P(X \leq 80), P(X \geq 110), P(85 < X < 120)$.
2. *Tìm x_0 sao cho $P(X < x_0) = 0.4$.*
3. *Tìm x_1 sao cho $P(X \geq x_1) = 0.1$.*
4. *Tính kì vọng và phương sai của X.*

Lời giải:

1. Việc tính xác suất $P(X \leq x_0)$ của biến ngẫu nhiên tuân theo phân phối chuẩn bằng SPSS ta làm tương tự như ví dụ 4.2.2. Lưu ý là hàm gõ vào khung chính là: **CDF.NORMAL(?,?,?)** trong đó 3 dấu hỏi lần lượt là: x_0, trung bình (=100), độ lệch chuẩn (=8).

   Ta có $P(X \leq 80)$ được tính qua lệnh **CDF.NORMAL(80,100,8)**và được kết quả là 0.006209665.

   Lưu ý rằng X là biến ngẫu nhiên liên tục nên $P(X < x_0) = P(X \leq x_0), P(X > x_0) = P(X \geq x_0)$. Do đó ta có:

   - $P(X \geq 110) = 1 - P(X < 110) =$ **1− CDF.NORMAL(110,100,8)**$= 0.10564977$.
   - $P(85 < X < 120) = P(X < 120) - P(X \leq 85) =$ **CDF.NORMAL(120,100,8)**− **CDF.NORMAL(85,100,8)**$= 0.96339397$.

2. Trong SPSS giá trị x_0 sao cho $P(X < x_0) = p$ với X tuân theo phân phối chuẩn được tính tương tự như tính xác suất của phân phối chuẩn ở phần trên với dòng lệnh: **IDF.NORMAL(?,?,?)** trong đó 3 dấu hỏi lần lượt là: p, giá trị trung bình, độ lệch chuẩn của X.

   Để tìm x_0 thỏa mãn yêu cầu đề bài, trong hộp thoại **Compute Variable** ta gõ **IDF.NORMAL (0.4, 100, 8)** vào khung chính, nhấp **OK**. Kết quả được $x_0 = 97.97$.

3. Ta có $P(X \geq x_1) = 0.1 \Leftrightarrow 1 - P(X < x_1) = 0.1 \Leftrightarrow P(X < x_1) = 0.9$. Do đó $x_1 =$ **IDF.NORMAL(0.9,100,8)** $= 110.25$.

4. $E(X) = 100, V(X) = 8^2 = 64$.

***Ví dụ* 4.2.5.** *Giả sử chiều cao của nữ sinh các trường đại học ở một nước là biến ngẫu nhiên tuân theo phân phối chuẩn với trung bình là 1.6 m, độ lệch chuẩn là 0.2 m.*

1. *Tính tỉ lệ nữ sinh cao hơn 1.7 m.*
2. *Tính tỉ lệ nữ có chiều cao khoảng* $[1.5; 1.8]$
3. *Tìm x_0 để tỉ lệ cao hơn x_0 là 10 %.*

Lời giải: Chon ngẫu nhiên một nữ sinh viên ở nước trên, gọi X là chiều cao của sinh viên đó. Ta có X tuân theo phân phối chuẩn với trung bình 1.6m và độ lệch chuẩn là 0.2 m.

1. Tỉ lệ nữ sinh viên cao hơn 1.7 m là $P(X > 1.7)$. Làm tương tự như ví dụ 4.2.4 ta có $P(X > 1.7) = 1-$CDF.NORMAL(1.7,1.6,0.2)$= 0.3085$. Vậy có khoảng 31 % sinh viên nữ cao hơn 1.7 m.

2. Tỉ lệ nữ sinh có chiều cao từ 1.5 m đến 1.8 m là $P(1.5 \leq X \leq 1.8) = P(X \leq 1.8) - P(X < 1.5)$ =CDF.NORMAL(1.8,1.6,0.2)$-$CDF.NORMAL(1.5,1.6,0.2) $= 0.5328$. Vậy có khoảng 53 % nữ sinh có chiều cao từ 1.5 m đến 1.8 m.

3. Ta cần tìm x_0 để $P(X > x_0) = 0.1 \Leftrightarrow P(X \leq x_0) = 0.9$. Ta có $x_0 =$ IDF.NORMAL(0.9,1.6,0.2) $= 1.8563$. Vậy nhóm 10 % những sinh viên cao nhất có chiều cao từ 1.86 m trở lên.

4.3. Bài tập

4.3.1. Bài tập phần xác suất

Bài tập 4.1. Tại một tòa nhà chung cư có 350 cư dân là nam giới, 300 là nữ giới. Nếu ta chọn phỏng vấn ngẫu nhiên một cư dân tại chung cư trên thì xác suất để người đó là nam giới là bao nhiêu?

Bài tập 4.2. Giả sử ở Việt Nam có 65 triệu công dân đủ tuổi để sở hữu ô tô. Biết rằng hiện tại có 16 triệu chiếc ô tô đã được đăng kí. Giả sử rằng mỗi ô tô được sở hữu bởi một công dân nào đó. Chọn ngẫu nhiên một công dân đủ tuổi sở hữu ô tô ở nước ta, tính xác suất để người đó không sở hữu ô tô.

Bài tập 4.3. Theo điều tra (số liệu năm 2006), mức độ hài lòng về cuộc sống vợ chồng tăng lên theo mức sống của hộ gia đình, trình độ học vấn và giảm dần theo số năm chung sống. Bất hoà về ứng xử và khó khăn về kinh tế là hai nguyên nhân chính khiến các cặp vợ chồng không hài lòng về hôn nhân của mình, trong số đó bất hoà về ứng xử: 45.3%, khó khăn về kinh tế: 43.4% trong đó 10 % là do cả hai nguyên nhân. Chọn ngẫu nhiên một cặp vợ chồng trong những cặp không hài lòng về hôn nhân. Tính xác suất để nguyên nhân gây ra điều này là do khó khăn kinh tế hoặc bất hòa ứng xử.

Bài tập 4.4. Theo điều tra tại một quận về mức độ quan tâm của cha mẹ với con cái đang học cấp 3, tỷ lệ cha mẹ biết về bạn thân của con là 75 %, biết về nơi con cái thường đến chơi là 60 % trong đó có 50 % bậc cha mẹ biết cả bạn bè thân và nơi con thường đến. Chọn ngẫu nhiên một phụ huynh có con học THPT ở quận trên, tính xác suất để phụ huynh đó biết bạn thân của con hoặc nơi con mình thường đến.

Bài tập 4.5. Một mẫu gồm 2000 người lớn được hỏi họ đã từng mua hàng qua mạng Internet chưa. Bảng sau cho thấy kết quả trả lời của họ.

	Đã từng mua	Chưa bao giờ
Nam	300	900
Nữ	200	600

Giả sử chọn ngẫu nhiên một người từ 2000 người trên. Tính xác suất để người được chọn:

1. Chưa bao giờ mua hàng qua mạng Internet hoặc là nữ giới.

2. Là nam giới hoặc đã từng mua hàng qua mạng Internet.

3. Đã từng hoặc chưa bao giờ mua hàng qua mạng Internet.

4. Các biến cố "nam" và "nữ" có xung khắc không? Còn các biến cố "đã từng mua hàng" và "nữ" thì sao? Giải thích tại sao.

5. Chưa bao giờ mua hàng qua Internet.

6. Là một nam giới.

7. Đã từng mua hàng qua Internet và biết rằng người này là nam.

8. Là nữ và biết rằng người này chưa từng mua hàng qua Internet.

Bài tập 4.6. Một nhóm gồm 2000 người lớn được chọn ngẫu nhiên để hỏi ý kiến họ ủng hộ hay chống đối việc nhân bản vô tính. Sau đây là kết quả trả lời của họ:

	Ủng hộ	Chống đối	Không có ý kiến
Nam	395	405	100
Nữ	300	680	120

Nếu chọn ngẫu nhiên một người từ 2000 người này, tính xác suất để người này:

1. Ủng hộ việc nhân bản vô tính.

2. Chống đối việc nhân bản vô tính.

3. Ủng hộ việc nhân bản vô tính và biết rằng người được hỏi là nữ.

4. Là nam và biết rằng người được hỏi là người không có ý kiến.

5. Các biến cố "nam" và "ủng hộ" có loại trừ lẫn nhau không? Hỏi tương tự với các biến cố "ủng hộ" và "chống đối"? Giải thích tại sao.

6. Các biến cố "nữ" và "không có ý kiến" là độc lập nhau phải không? Giải thích tại sao.

4.3.2. Bài tập phần biến ngẫu nhiên

Bài tập 4.7. Một văn phòng về người tiêu thụ thực hiện một cuộc điều tra khảo sát trên tất cả 2500 gia đình sống tại một thành phố nhỏ nhằm thu thập dữ liệu về số ti vi các gia đình sở hữu. Bảng sau liệt kê phân phối tần suất từ dữ liệu thu được:

Số ti vi gia đình có được	0	1	2	3	4
Số gia đình	120	970	730	410	270

1. Lập bảng phân phối xác suất đối với số ti vi các gia đình sở hữu.

2. Chọn ngẫu nhiên một gia đình từ thành phố này. Gọi X là số ti vi mà gia đình này sở hữu. Tính xác suất $P(X = 1), P(X > 2), P(X \leq 1), P(1 \leq X \leq 3)$.

3. Tính kỳ vọng, phương sai của X. Nêu ý nghĩa của các con số này.

Bài tập 4.8. Điều tra thống kê của We are Social ở nước ta vào tháng 1 năm 2014 cho thấy tỉ lệ người dùng internet trên di động chiếm 34 % dân số, trong số những người dùng di động có 58 % người sử dụng ứng dụng mạng xã hội trên điện thoại.

1. Giả sử tại thời điểm trên ta chọn ngẫu nhiên 3 người dân bất kì, gọi X là số người dùng internet trên di động trong số 3 người đó. X tuân theo phân phối gì? Hãy lập bảng phân phối xác suất cho X. Tính kì vọng và phương sai của X.

2. Giả sử tại thời điểm trên ta chọn ngẫu nhiên 8 người dùng điện thoại di động. Trong số 8 người này:

   (a) Tính xác suất để có đúng 4 người truy cập mạng xã hội từ điện thoại.

   (b) Tính xác suất để có ít nhất 3 người truy cập mạng xã hội từ điện thoại.

   (c) Tính xác suất để có từ 2 đến 6 người truy cập mạng xã hội từ điện thoại.

   (d) Tính trung bình số người dùng điện thoại truy cập mạng xã hội.

   (e) Khả năng có bao nhiêu người truy cập mạng xã hội từ điện thoại là lớn nhất?

Bài tập 4.9. Theo ý kiến nhiều chuyên gia, tỉ lệ người dân hài lòng với các dịch vụ công ở nước ta là 40 %. Giả sử con số này là đúng trên thực tế. Chọn ngẫu nhiên 20 công dân ở nước ta, gọi X là số người hài lòng với dịch vụ công trong số 20 người đó.

1. Phép thử trên có phải là phép thử nhị thức? Tại sao?

2. Trung bình có bao nhiêu người hài lòng với các dịch vụ công? Tính xác suất xảy ra trường hợp đó. Đó có phải trường hợp có khả năng xảy ra lớn nhất không?

3. Tính xác suất để có ít nhất 10 người hài lòng với dịch vụ công.

4. Tính xác suất để có ít nhất 10 người không hài lòng với các dịch vụ công.

Bài tập 4.10. Một cuộc khảo sát mức độ hiểu biết về HIV (năm 2009) ở thanh thiếu niên nước ta cho thấy có tới 26 % người được hỏi tin rằng HIV có thể lây truyền qua đường muỗi đốt. Giả sử ta chọn ngẫu nhiên 15 người trong danh sách được phỏng vấn và gọi X là số người cho rằng HIV lây qua đường muỗi đốt trong 15 người trên.

1. Trung bình có bao nhiêu người trong 15 người trên cho rằng HIV lây qua đường muỗi đốt.

2. Tính EX, VX.

3. Số người cho rằng HIV lây qua đường muỗi đốt có khả năng nhất là bao nhiêu?

4. Xác suất để hơn nửa số người trong số được chọn cho rằng HIV lây qua đường muỗi đốt là bao nhiêu?

Bài tập 4.11. Cho X là biến ngẫu nhiên phân phối chuẩn với tham số $\mu = 8, \sigma^2 = 25$, tính

1. P(X $<$ 5) 2. P(2 $<$ X $<$ 6) 3. P(X $>$ 15)

Bài tập 4.12. Chỉ số IQ của trẻ em trong độ tuổi 12 - 15 tuân theo phân phối chuẩn $N(85; 25)$.

1. Chỉ số IQ trung bình của trẻ ở độ tuổi này là bao nhiêu?
2. Khả năng chọn được một trẻ lứa tuổi này rất thông minh (có chỉ số IQ$\geqslant$ 90) là bao nhiêu?
3. Xét trong 100 học sinh khối lớp 7 tại một trường:
   (a) Gọi Y là số học sinh rất thông minh trong khối trên. Hỏi Y tuân theo phân phối gì?
   (b) Hỏi trung bình có bao nhiêu học sinh rất thông minh?
   (c) Tìm xác xuất để có đúng 20 học sinh rất thông minh trong khối.

Bài tập 4.13. Giả sử chiều cao của nam thanh niên Việt Nam trong độ tuổi 22-26 tuân theo phân phối chuẩn với trung bình (năm 2010) là 1.64m với độ lệch chuẩn 0.08m.

1. Xác suất để chọn được một nam thanh niên ở độ tuổi trên có chiều cao từ 1.5m đến 1.7m là bao nhiêu?
2. Tính tỉ lệ nam thanh niên có chiều cao dưới 1.4m.
3. 10 % nam thanh niên cao nhất có chiều cao từ bao nhiêu trở lên.

Bài tập 4.14. Đời sống của một loại máy tính xách tay có phân phối chuẩn với trung bình là 3 năm và độ lệch chuẩn là 0.06 năm. Chọn một máy bất kì thuộc loại này.

1. Tính xác suất máy được chọn hỏng sau 2.5 năm.
2. Nếu muốn tỉ lệ bảo hành 5% thì thời gian bảo hành là bao lâu?

Bài tập 4.15. Một nhà tâm lý học thực hiện một kiểm tra về mức độ căng thẳng đối với các bệnh nhân đi chữa răng trong khi họ ngồi chờ ở phòng đợi. Cuộc kiểm tra cho kết quả là mức độ căng thẳng (có thang đo từ 1 đến 10) đối với các bệnh nhân chờ lấy tủy răng có phân phối xấp xỉ chuẩn với trung bình là 7.59 và độ lệch chuẩn là 0.73.

1. Hãy cho biết số phần trăm các bệnh nhân như vậy có mức độ căng thẳng dưới 6.0?
2. Tính xác suất để một bệnh nhân cần phải lấy tủy răng khi ngồi chờ ở phòng đợi có mức căng thẳng nằm giữa 7.0 và 8.0?
3. Nhà tâm lý học đề nghị rằng đối với các bệnh nhân mà có độ căng thẳng là 9.0 hoặc cao hơn thì nên được điều trị giảm căng thẳng trước khi thực hiện việc lấy tủy răng. Nếu như đề nghị này được chấp thuận thì hỏi có bao nhiêu phần trăm bệnh nhân cần được điều trị giảm căng thẳng?

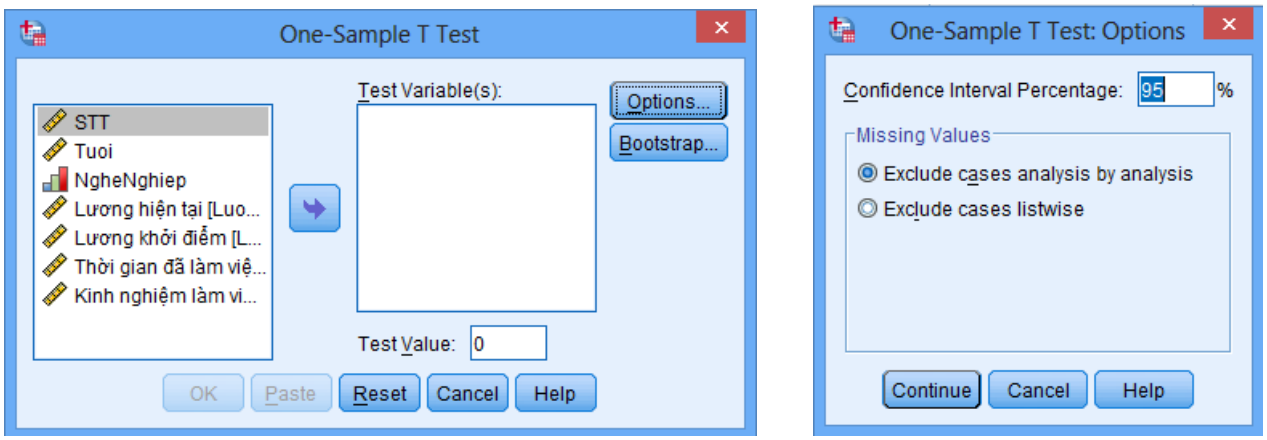
Bài tập 4.16. Bộ phận quản lý của một siêu thị muốn đưa ra một chính sách vận động nhằm lôi kéo khách hàng đến mua sắm nhiều hơn tại siêu thị. Chính sách này là sẽ tặng một món quà miễn phí cho những khách hàng nào mua hàng vượt quá một mức tiền nào đó trong mỗi lần ghé vào siêu thị. Bộ phận quản lý hy vọng rằng khi chính sách tặng quà được quảng bá thì số tiền mua sắm của khách hàng trong mỗi lần ghé vào siêu thị sẽ có phân phối chuẩn với trung bình là 300 (nghìn) và độ lệch chuẩn là 86 (nghìn). Nếu bộ phận quản lý chỉ muốn tặng quà nhiều nhất là 10% khách hàng thì số tiền mua sắm của khách hàng trong mỗi lần ghé vào siêu thị nên tối thiểu là bao nhiêu?

Chương 5

Ước lượng và kiểm định giả thuyết

5.1. Ước lượng và kiểm định trung bình một tổng thể với một số

Để ước lượng điểm và ước lượng khoảng với độ tin cậy α cho trung bình của một tổng thể từ một mẫu điều tra ngẫu nhiên ta vào **Analyze** $\longrightarrow$ **Compare Means** $\longrightarrow$ **One - Sample T Test**.



Hình 5.1: Các hộp thoại ước lượng khoảng và kiểm định trung bình một tổng thể

Trong hộp thoại trên, khi ước lượng hay kiểm định cho trung bình một tổng thể nào đó ta chọn biến tương ứng vào khung **Test Variable**, và:

- Đối với bài toán ước lượng thì ta chọn giá trị cho mục **Test Value** là **0**, sau đó điều chỉnh độ tin cậy ở hộp thoại **One - Sample T Test: Options** thông qua nút **Options**. Mặc định ban đầu là 95 (%).
- Để thực hiện thủ tục kiểm định so sánh trung bình một tổng thể với một số nào đó ta điền giá trị cần so sánh vào **Test Value** (mặc định ban đầu là 0).

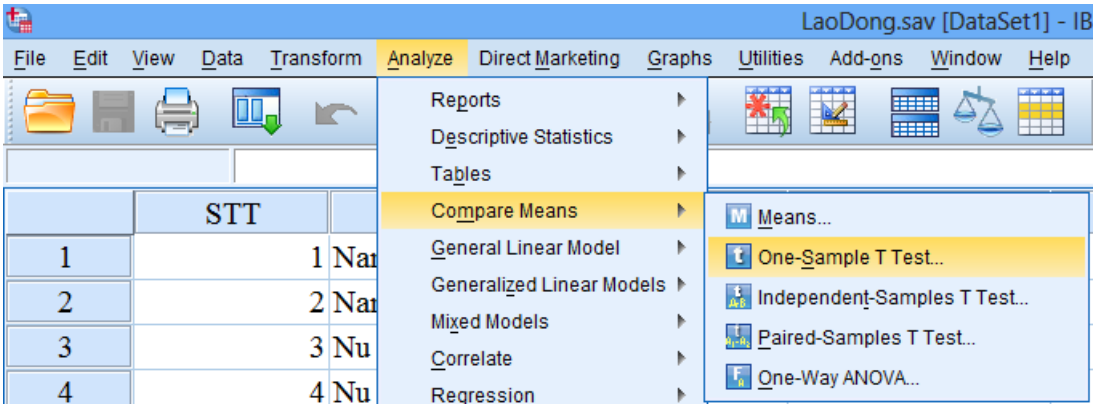
5.1.1. Ước lượng trung bình một tổng thể

***Ví dụ* 5.1.1.** *File* **LaoDong.sav** *là thông tin về 473 lao động được điều tra ngẫu nhiên tại quận tên là A. Tìm một ước lượng điểm và một khoảng tin cậy 90% cho độ tuổi trung bình của người lao động ở quận trên.*

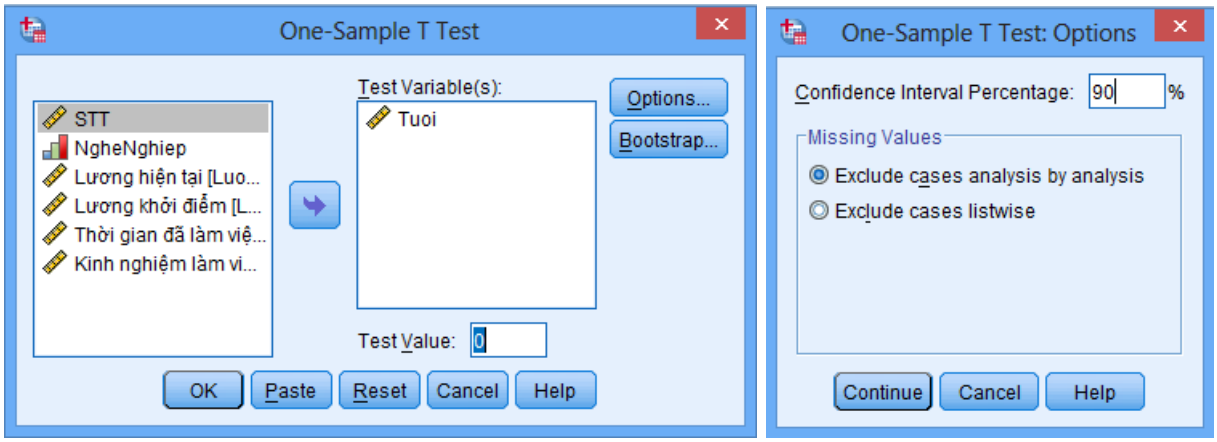
Lời giải:

Để tìm khoảng tin cậy 90% cho độ tuổi trung bình của những người lao động ta làm theo các bước như hình sau:

Bước 1



Bước 2: Trong hộp thoại hiện ra điền 0 vào **Test Value**, nhấp vào **Options** thay số 95 (%) thành 90 (%)



Hình 5.2: Tìm khoảng tin cậy 90% cho độ tuổi trung bình của những người lao động

Lần lượt nhấp **Continue** và **OK**, kết quả ở cửa sổ Output cho ta 2 bảng:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Tuoi	473	32.6723	11.78409	.54183

Hình 5.3: Bảng thống kê biến về mẫu Tuoi

- Bảng đầu tiên (hình 5.3) cho một số thống kê về biến Tuoi, trong đó có ước lượng điểm (Mean) của tuổi lao động ở quận A là 32.67.
- Bảng thứ 2 (hình 5.4), phần **90 % Confidence Interval of Difference** là khoảng ước lượng từ 31.8 đến 33.7 (tuổi). (Thực ra nó là ước lượng cho hiệu của trung bình tổng thể với **Test Value** (ở đây ban đầu chọn là 0))

Tóm lại hai bảng cho ta kết luận: ước lượng điểm của tuổi người lao động quận A là 32.67 tuổi, ước lượng khoảng tin cậy 90 % cho độ tuổi lao động trung bình ở quận trên là từ 31.8 đến 33.7 (tuổi).

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	90% Confidence Interval of the Difference	
					Lower	Upper
Tuoi	60.300	472	.000	32.67230	31.7793	33.5653

Hình 5.4: Kết quả ước lượng với khoảng tin cậy 90% cho độ tuổi trung bình của người lao động

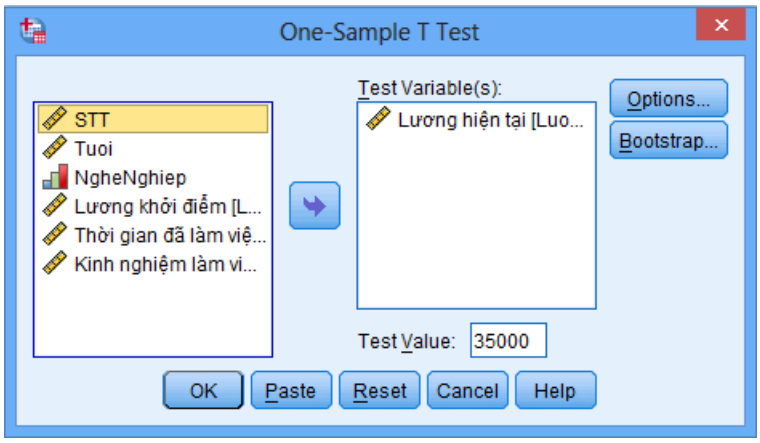
5.1.2. Kiểm định trung bình một tổng thể với một số

***Ví dụ* 5.1.2.** *Có một tờ báo địa phương cho rằng lương trung bình của lao động ở quận A (đã nói ở ví dụ 5.1.1) là 35000$ một năm. Ở mức ý nghĩa 0.05 khẳng định của bài báo có đúng không?*

Lời giải:

Gọi μ là lương trung bình của tổng thể lao động ở quận A.

- Cặp giả thuyết: $H_0 : \mu = 35000 \qquad H_1 : \mu \neq 35000.$
- Các bước tiến hành kiểm định làm như phần ước lượng nói trên nhưng thay giá trị 0 của Test Value bởi giá trị cần so sánh là 35000; phần độ tin cậy không cần thay đổi.



One-Sample Test

	Test Value = 35000					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Lương hiện tại	-.740	473	.460	-$580.432	-$2,121.60	$960.73

Hình 5.5: Kiểm định so sánh μ với 35 000 và bảng kết quả

Kết quả trong **Output** cho ta: **sig.(2-tailed)** là 0.46 đây chính là **P-giá trị** của bài toán 2 bên. Ta có P-giá trị $= 0.46 > 0.05$ nên chấp nhận giả thiết H_0, tức là ở mức ý nghĩa 0.05 ta thấy lương trung bình của tổng thể lao động ở quận A có thể coi là 35000.

***Ví dụ* 5.1.3.** *Tờ báo nói trên cũng cho rằng lương khởi điểm trung bình của các lao động quận A lớn hơn 18000 $. Ở mức ý nghĩa 5 % với dữ liệu ta có, lời khẳng định đó chấp nhận được không?*

Lời giải: Gọi μ là lương khởi điểm trung bình của tổng thể lao động ở quận A.

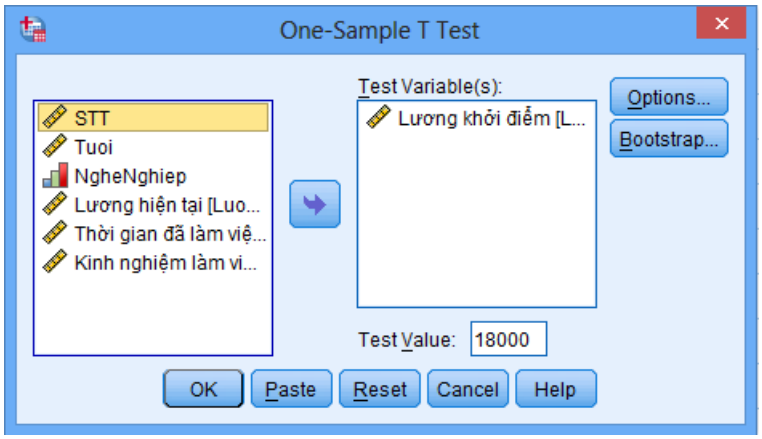
- $H_0 : \mu \leq 18000 \qquad H_1 : \mu > 18000.$

  Đây là bài toán kiểm định bên phải. Cách tìm P-giá trị của bài toán này như sau:

  – Nếu giá trị kiểm định **t** tính ra âm thì P-giá trị $= 1 - \dfrac{\text{sig.(2 tailed)}}{2}$.

  – Nếu giá trị kiểm định **t** tính ra dương thì thì P-giá trị $= \dfrac{\text{sig.(2 tailed)}}{2}$.

- Tiến hành các bước trên SPSS như ở phần trên trong đó giá trị của **Test Value** là 18000.



One-Sample Test

	Test Value = 18000					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Lương khởi điểm	-2.722	473	.007	-$983.914	-$1,694.28	-$273.55

Hình 5.6: Kiểm định μ với 35 000 và bảng kết quả

Do giá trị **t** $= 2.722 < 0$ và đây là bài toán kiểm định bên phải nên P-giá trị $= 1 - \dfrac{\text{sig.(2 tailed)}}{2} = 1 - \dfrac{0.07}{2} > 0.05$ nên chấp nhận H_0. Như vậy, ở mức ý nghĩa 0.05 lương khởi điểm trung bình của lao động quận A không thể lớn hơn 18000 $ như lời khẳng định của tờ báo.

***Ví dụ* 5.1.4.** *Trong báo cáo của một tổ chức phi chính phủ điều tra ở quận A có chi tiết rằng lương khởi điểm trung bình của lao động ở quận A ít nhất là 16000$. Ở mức ý nghĩa 10 % khẳng định này của báo này có thể chấp nhận được không?*

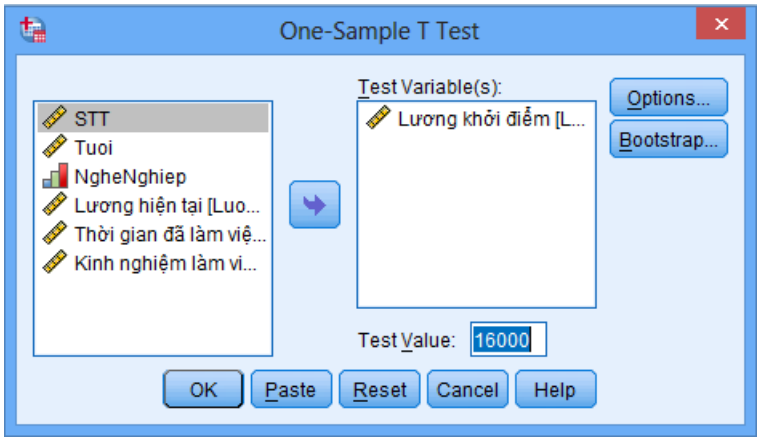
Lời giải: Gọi μ là lương khởi điểm trung bình của tổng thể lao động ở quận A.

- Cặp giả thuyết: $H_0 : \mu \geq 16000 \qquad H_1 : \mu < 16000.$

  Đây là bài toán kiểm định bên trái. Cách tìm P-giá trị của bài toán này như sau:

  – Nếu giá trị kiểm định **t** tính ra âm thì P-giá trị $= \dfrac{\text{sig.(2 tailed)}}{2}$.

  – Nếu giá trị kiểm định **t** tính ra dương thì thì P-giá trị $= 1 - \dfrac{\text{sig.(2 tailed)}}{2}$.



One-Sample Test

	Test Value = 16000					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Lương khởi điểm	2.811	473	.005	$1,016.086	$305.72	$1,726.45

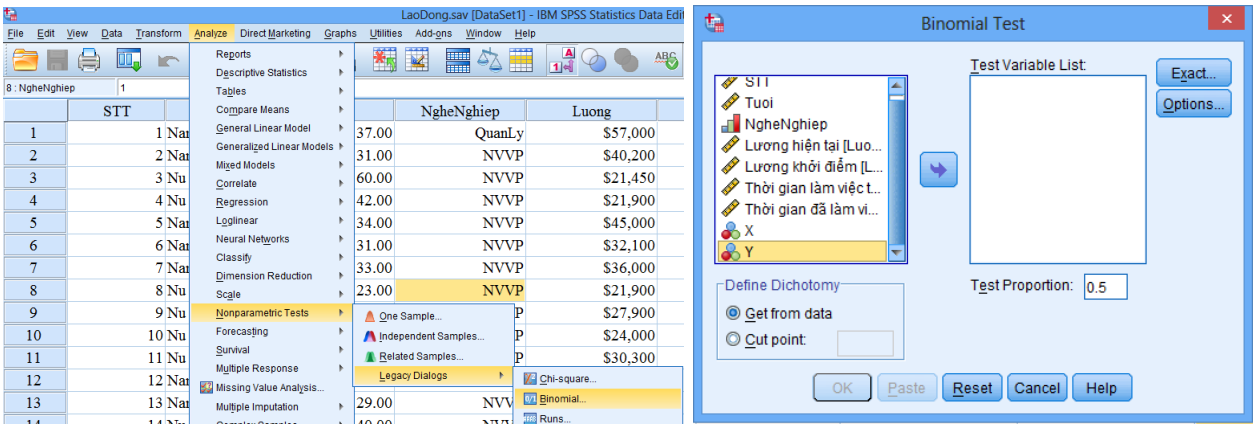
Hình 5.7: Kiểm đinh lương khởi điểm trung bình của lao động quận A với 16000$

- Làm các bước như phần trên, mục **Test Value** điền giá trị 16000.

Kết quả cho thấy giá trị kiểm định $\mathbf{t} = 2.811 > 0$ do đó $\mathbf{P\text{-giá trị}} = 1 - \frac{\text{sig.(2 tailed)}}{2} = 1 - \frac{0.05}{2} > 0.1$ nên chấp nhận H_0, tức là bác bỏ $m < 38$. Vậy ở mức ý nghĩa 0.05 ta thấy rằng lương khởi điểm trung bình của các lao động quận A không thể nhỏ hơn 16 000 $ như báo cáo.

5.2. Kiểm định tỉ lệ một tổng thể với một số

Để tiến hành thủ tục cho bài toán kiểm định tỉ lệ một tổng thể với một số trong SPSS ta vào mục **Analyze** $\longrightarrow$ **Nonparametric Test** $\longrightarrow$ **Legacy Dialogs** $\longrightarrow$ **Binomial**



Hình 5.8: Hộp thoại Binomial Test

Ta chọn **Get from data** khi biến đã cho ở dạng số nhị phân (chỉ có 2 giá trị), chọn **Cut point** khi biến định lượng có nhiều biểu hiện và ta có điểm cắt. Lưu ý rằng chỉ biến dạng **Nummeric** mới có trong danh sách biến để tiến hành thủ tục kiểm định này. Do đó một biến là **String** thì ta phải

mã hóa nó thành biến **Numeric** nhị phân. Còn khung **Test Proportion** điền giá trị p_0 mà ta cần so sánh.

Khi tiến hành thủ tục này, kết quả trong **Output** chỉ cho **sig.(2-tailed)** khi $p_0 = 0.5$ còn khi $p_0 \neq 0.5$ thì cho **sig.(1-tailed)**, đây là kết quả p-giá trị của bài toán một bên. Sau đây là hướng dẫn chi tiết việc tính p - giá trị các bài toán.

Trước hết, ta có các bài toán kiểm định tỉ lệ p của tổng thể với một số p_0:

- Bài toán hai bên: $H_0 : p = p_0 \qquad H_1 : p \neq p_0$.
- Bài toán bên phải: $H_0 : p \leq p_0 \qquad H_1 : p > p_0$.
- Bài toán bên trái: $H_0 : p \geq p_0 \qquad H_1 : p < p_0$.

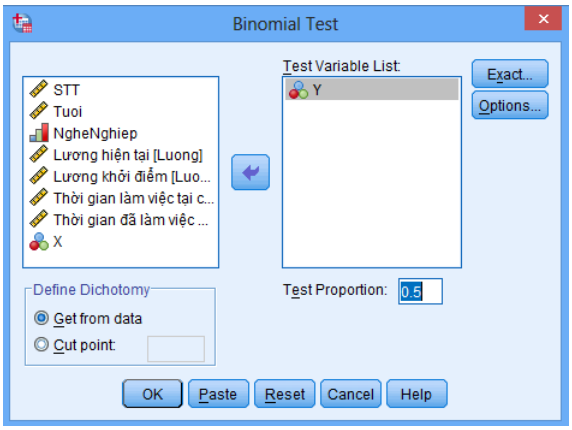
Ta có p-giá trị của bài toán bên trái + p-giá trị bài toán bên phải = 1 và có các trường hợp sau:

- Trường hợp 1: khi $p_0 = 0.5$, kết quả ở **Output** cho ta **sig.(2-tailed)**.
- Trường hợp 2: khi $p_0 \neq 0.5$, kết quả trong **Output** cho ta **sig.(1-tailed)**.Lưu ý, sig.(1-tailed) này là p-giá trị của bài toán bên trái nếu **Observed Prop** $< p_0$ (hoặc có chú thích ngay bên dưới bảng), là của bài toán bên phải nếu **Observed Prop** $> p_0$ (hoặc dưới bảng không có chú thích gì).

***Ví dụ* 5.2.1.** *Ta trở lại với mẫu dữ liệu điều tra* **LaoDong.sav** *tại quận A. Có khẳng định cho rằng tỉ lệ lao động nam ở quận A chiếm đúng 50 %. Kiểm định khẳng định trên ở mức ý nghĩa 1%.*

Lời giải: Gọi p là tỉ lệ lao động nam ở quận A.

$H_0 : p = 0.5 \qquad H_1 : p \neq 0.5$



Hình 5.9: Kiểm định tỉ lệ với biến Y - là biến nhị phân, chọn **Get from data**

Ta cần xử lí dữ liệu trước khi tiến hành thủ tục kiểm định như sau:

1. Tạo biến Y là mã hóa biến GioiTinh với Nam được gán thành 1, Nu được gán thành 2.
2. Sort biến Y theo chiều tăng dần (làm điều này là do ta muốn biểu hiện 1 (tương ứng với giá trị "Nam" trong GioiTinh) lên trước tiên).

Sau đó trong hộp thoại **Binomial Test** ta điền như hình 5.9. Kết quả trong Output như sau:

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Y	Group 1	1.00	258	.54	.50	.060
	Group 2	2.00	216	.46		
	Total		474	1.00		

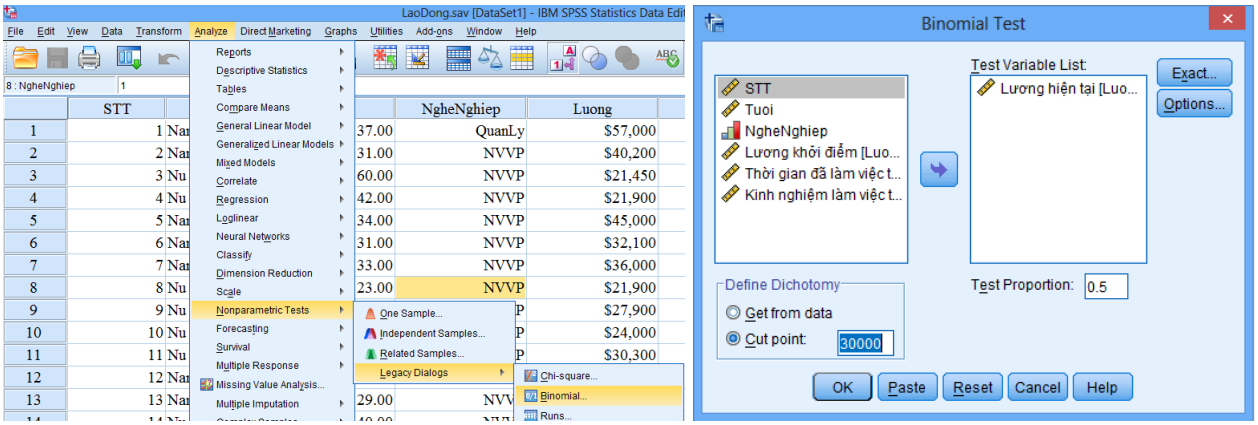
Kết quả này cho ta p - giá trị = **Sig.(2-tailed)** = 0.06> 0.01 nên chấp nhận H_0. Do vậy, ở mức ý nghĩa 0.01, tỉ lệ lao động nam ở quận A là 50 %.

Lưu ý: SPSS sẽ cho biểu hiện xuất hiện hiện ở dòng đầu tiên của biến Y thành Group 1, do đó để đảm bảo cho nhóm Nam là nhóm được so sánh tỉ lệ ta phải sort biến Y theo chiều tăng dần (để cho 1 lên dòng đầu tiên).

***Ví dụ* 5.2.2.** *Vẫn là mẫu dữ liệu điều tra* **LaoDong.sav** *tại quận A. Ở mức ý nghĩa 0.05 có thể chấp nhận rằng tỉ lệ lao động ở quận A hưởng lương* $\leq$ *30000 có chiếm đúng 50 % hay không?*

Lời giải: Gọi p là tỉ lệ lao động hưởng lương $\leq$ 30000.

- $H_0 : p = 0.5$ $H_1 : p \neq 0.5$. Đây là bài toán 2 bên.
- Ta làm như sau, ở hộp thoại hiện ra ta tích vào **Cut point** và điền vào số 30000.



Hình 5.10: Kiểm định tỉ lệ với điểm cắt

Nhấp **OK** và kết quả như sau:

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Lương hiện tại	Group 1	<= 30000	263	.55	.50	.019
	Group 2	> 30000	211	.45		
	Total		474	1.00		

Hình 5.11: Kết quả của bài tỉ lệ lao động hưởng lương $\leq$ 30000 với 0.5

Từ bảng kết quả ta thấy **P-giá trị** $= 0.019 < 0.05$ nên bác bỏ H_0, tức là tỉ lệ lao động hưởng lương ≤ 30000 là khác 50 %.

Nhận xét: Trong kết quả mà SPSS phân tích ta thấy dữ liệu luôn được chia thành 2 nhóm: một nhóm có giá trị biến $\leq$ điểm cắt và nhóm kia bao gồm các giá trị lớn hơn. Và nhóm thứ nhất luôn ở group 1, nhóm thứ 2 (lớn hơn điểm cắt) luôn ở group 2. Điều này đôi khi không thực sự phù hợp với bài toán mà chúng ta đặt ra. Chẳng hạn: ta cần kiểm định tỉ lệ lao động lương > 30000 và muốn nó ở group 1 (là phần để so sánh với p_0), ... Điều này có thể khắc phục bằng cách sau:

- Bước đầu tiên ta mã hóa lại biến đã cho thành một biến khác chỉ có 2 biểu hiện tương ứng với 2 nhóm: nhóm quan tâm ứng với 1 nhóm còn lại ứng với 2. Chẳng hạn ở ví dụ trên những lao động có lương ≤ 30000 có giá trị ở biến mã hóa là 1, phần còn lại là 2.
- Bước thứ 2 ta sort lại biến mã hóa theo chiều tăng.
- Bước thứ 3 ta tiến hành kiểm định Binom với biến mã hóa.

***Ví dụ* 5.2.3.** *Trong khảo sát tại quận A, người ta cũng quan tâm đến sự luân chuyển và ổn định của lao động ở quận A, một trong những thông tin có thể tham khảo là thời gian gắn bó của người lao động với công ty, thời gian đã từng làm ở nơi khác trước khi đến công ty. Ở mức ý nghĩa 5% hãy kiểm định xem:*

1. *tỉ lệ lao động gắn bó trên 2 năm (24 tháng) đối với chỗ làm hiện tại năm có hơn 40 % không?*
2. *tỉ lệ lao động từng đi làm dưới 1 năm trước khi vào công ty hiện tại có lớn hơn 60 % không?*

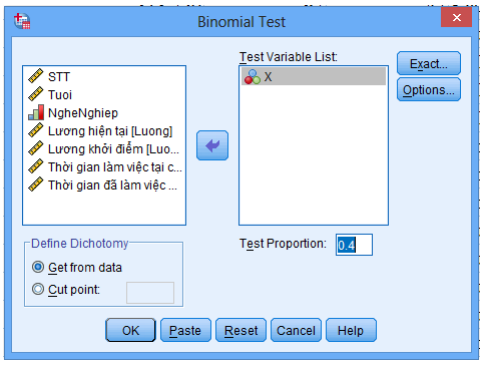
Lời giải:

1. Gọi p là tỉ lệ lao động có thời gian làm tại công ty trên 24 tháng.

   $H_0 : p \leq 0.4$ $H_1 : p > 0.4$ đây là bài toán kiểm định bên phải.

   - Ta lập biến X là mã hóa của biến ThoiGianLam, gán 1 cho những giá trị từ 24 trở lên, những giá trị còn lại gán là 2.
   - Sort lại X theo chiều tăng dần.

   Trong hộp thoại Binomial ta chuyển biến X sang vào điền giá trị 0.4 cho **Test Proportion** như hình sau:



   Kết quả trong bảng là **Sig.(1-tailed)** < 0.001 cho bài toán bên phải, trùng khớp với bài toán ta đang xét, nên có p - giá trị $< 0.001 < 0.05$ nên bác bỏ H_0. Tức là tỉ lệ lao động gắn bó hơn 2 năm với công ty hiện tại là lớn hơn 40 %.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
X	Group 1	1.00	336	.7	.4	.000
	Group 2	2.00	138	.3		
	Total		474	1.0		

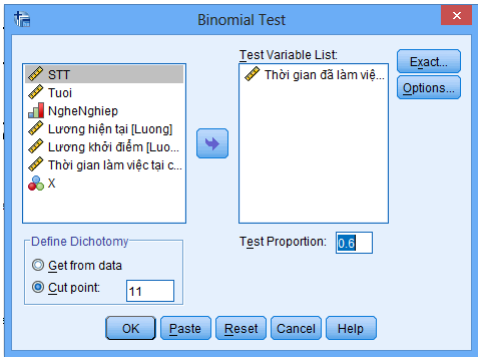
Hình 5.12: Kết quả kiểm định không có chú thích gì thêm: là kết quả của bài toán bên phải

2. Gọi p là tỉ lệ lao động có kinh nghiệm làm việc dưới 12 tháng trước khi vào công ty hiện tại.

   $H_0 : p \leq 0.6 \qquad H_1 : p > 0.6$

   Ta lưu ý rằng kinh nghiệm làm việc trước khi đến là các số nguyên nên < 12, tức là ≤ 11. Tỉ lệ ta quan tâm là tỉ lệ lao động có KinhNghiem ≤ 11 nên ta chỉ cần chọn 11 làm điểm cắt, không nhất thiết tạo biến mã hóa mới.

   Trong hộp thoại **Binomial Test** ta điền như sau:



Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Thời gian đã làm việc ở nơi khác trước lúc vào công ty (tháng)	Group 1	<= 11	244	.5	.6	.000[a]
	Group 2	> 11	230	.5		
	Total		474	1.0		

a. Alternative hypothesis states that the proportion of cases in the first group < .6.

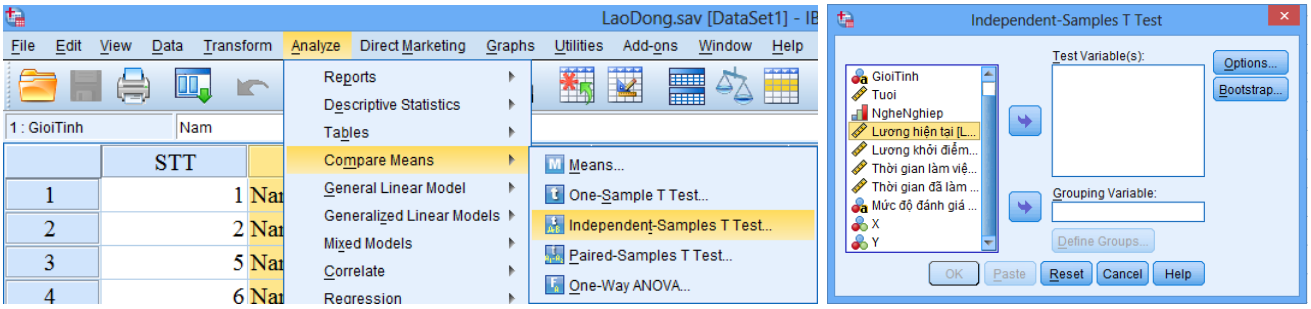
Hình 5.13: Kết quả kiểm định có chú thích: là kết quả của bài toán bên bên trái

Kết quả cho ta **Sig.(1-tailed)** < 0.001 của bài toán bên trái, bài toán ta đang xét lại là bài toán bên phải, do đó, p - giá trị $> 1 - 0.001 = 0.999 > 0.05$. Do đó chấp nhận H_0. Tức là tỉ lệ lao động có kinh nghiệm dưới 1 năm trước khi đến công ty không quá 60 %.

5.3. Kiểm định trung bình hai tổng thể

Kiểm định trung bình hai tổng thể trong SPSS chia thành 2 trường hợp: mẫu chọn theo đôi và mẫu chọn độc lập.

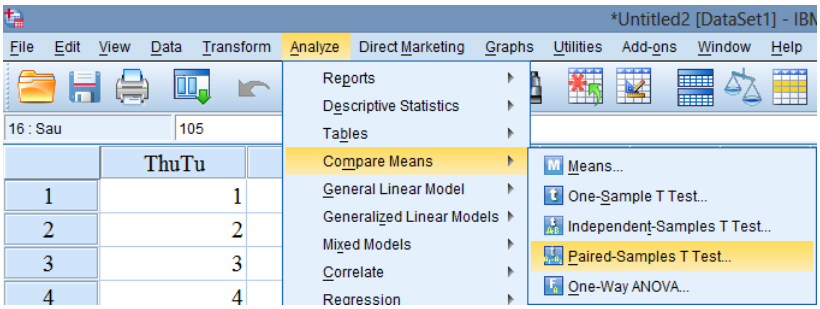
- Trường hợp mẫu chọn độc lập ta vào: **Analyze** $\longrightarrow$ **Compare Means** $\longrightarrow$ **Independent Samples T Test**:



Hình 5.14: Hộp thoại Independent - Samples T Test

Trong hộp thoại **Independent - Samples T Test** ta chuyển biến (đinh lượng) cần kiểm định vào khung **Test Variable**. Chuyển biến phân nhóm (cũng là định lượng) vào **Grouping Variable**. Nút **Define Groups** giúp ta chọn 2 nhóm để kiểm định.

- Trường hợp mẫu chọn theo đôi ta vào: **Analyze** $\longrightarrow$ **Compare Means** $\longrightarrow$ **Paired Samples T Test**



5.3.1. Các ví dụ

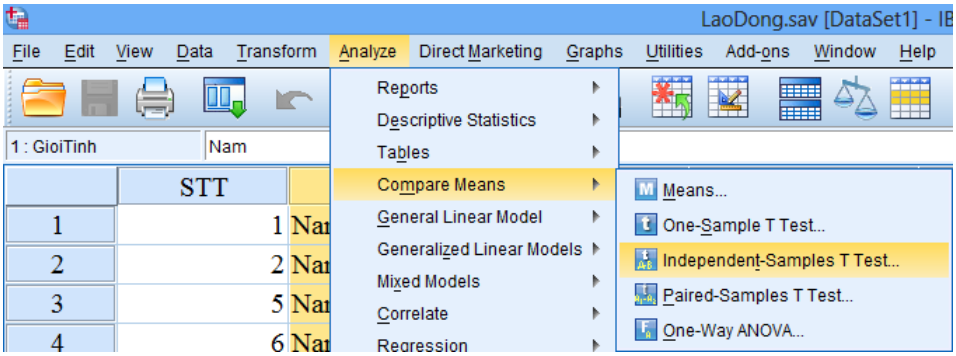
***Ví dụ* 5.3.1.** *Người ta cho rằng thu nhập trung bình của nam giới tại quận A lớn hơn thu nhập của nữ giới. Với dữ liệu điều tra ngẫu nhiên* **LaoDong.sav**, *tại mức ý nghĩa 5% có thể chấp nhận được quan điểm đó không?*

Gọi m_1, m_2 lần lượt là lương trung bình của nam và nữ ở quận A.

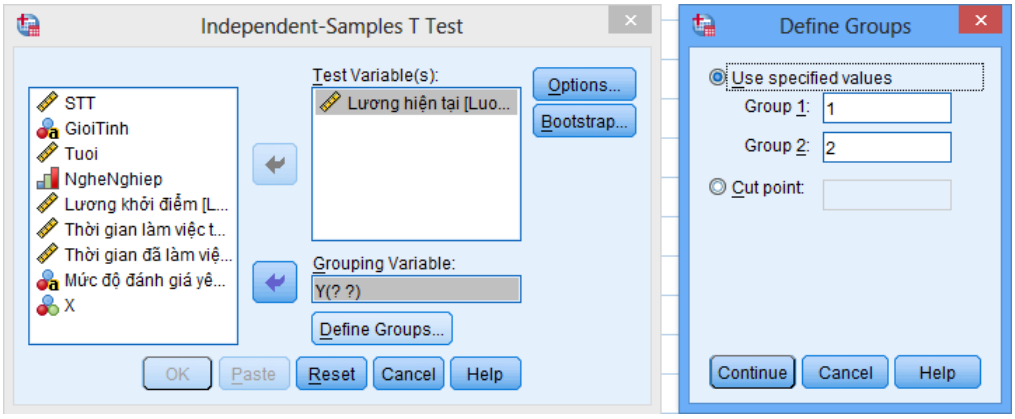
Cặp giả thiết: $H_0 : m_1 \leq m_2 \qquad H_1 : m_1 > m_2$.

Do biến dùng để phân nhóm giới tính là định tính nên trước khi tiến hành thủ tục kiểm định ta phải tạo ra biến định lượng **Y** là mã hóa của biến giới tính, với nam được mã hóa là 1, nữ là 2.

Ta tiến hành các thao tác kiểm định trọng SPSS như hình sau:



Trong hộp thoại xuất hiện, ta chọn biến **Lương hiện tại** vào khung **Test Variable**, **Y** vào khung **Grouping Variable**. Nhấp vào nút **Define Group** để chọn những nhóm so sánh với nhau (chỉ được chọn 2), ở đây là nhóm 1 (nam) và nhóm 2 (nữ). Nhấp **Continue** rồi **OK**.



Trong cửa sổ **Output** cho ta bảng sau:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Lương hiện tại	Equal variances assumed	119.669	.000	10.945	472	.000	$15,409.862	$1,407.906	$12,643.322	$18,176.401
	Equal variances not assumed			11.688	344.262	.000	$15,409.862	$1,318.400	$12,816.728	$18,002.996

Trong bảng kết quả có hai phần:

- **Levene's Test for Equality of Variances** là kết quả kiểm định của bài toán so sánh phương lương hai tổng thể: tổng thể lao động nam và tổng thể nữ ở quận A.
- **t-test for Equality of Means** là kết quả của bài toán kiểm định so sánh lương trung bình của hai tổng thể.

Từ phần thứ nhất ta thấy p - giá trị cho bài toán so sánh hai phương sai là **Sig.** $< 0.001 < 0.05$ nên ta bác bỏ giả thiết hai phương sai bằng nhau. Do đó ta sẽ lấy kết quả cho bài toán kiểm định so sánh trung bình theo dòng thứ 2 (tương ứng với **Equal variances not assumed**). Vậy **sig. (2-tailed)** của bài toán ta xét là < 0.001.

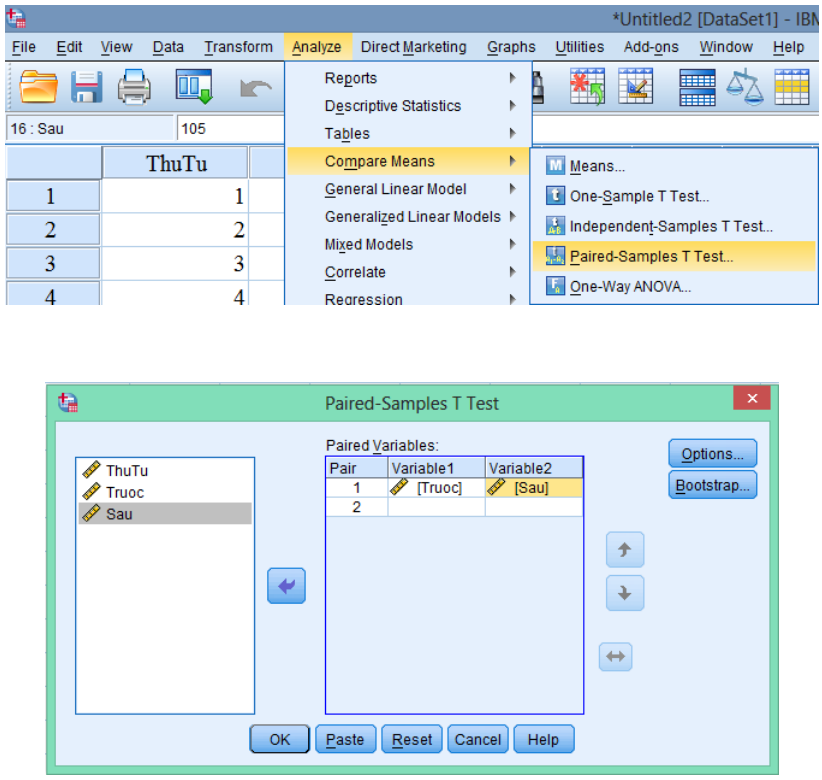
Do là bài toán ta xét là bài toán bên phải, có $t = 11.668 > 0$ nên P-giá trị $< \dfrac{0.001}{2} = 0.0005 < 0.05$ nên bác bỏ H_0, tức là ở mức ý nghĩa 0.05 ta chấp nhận quan điểm cho rằng lương tủng bình của tổng thể lao động nam giới là lớn hơn.

***Ví dụ* 5.3.2.** *Để xem việc treo thưởng có tác dụng làm tăng năng suất lao động của công nhân hay không người ta tiến hành thử nghiệm đo năng suất lao động một tháng có hứa treo thưởng cho công nhân nào đạt được mức năng suất cao hơn so với trung bình các tháng. Dữ liệu trong file* **LuongThuong.txt**, *trong đó Truoc là số sản phẩm làm được trong tháng chưa có treo thưởng, Sau là số sản phẩm làm trong tháng có treo thưởng.*
Ở mức ý nghĩa 0.05 qua số liệu đó có thể nói rằng việc treo thưởng làm tăng năng suất lao động trung bình của các công nhân hay không?

Gọi m_1, m_2 lần lượt là số sản phẩm trung bình một tháng của công nhân trước và sau khi treo thưởng.

- Cặp giả thiết $H_0 : m_1 = m_2 \qquad H_1 : m_1 < m_2$

- Lưu ý rằng mẫu được chọn ở đây là theo đôi nên quá trình thực hiện bài toán kiểm định này trong SPSS qua các bước như hình sau:



Hình 5.15: Kéo từng biến Truoc và Sau vào cặp biến thứ nhất

Trong hộp thoại hiện ra chọn biến (bằng cách kéo) Truoc và Sau vào cặp biến thứ nhất. Nhấp **OK**.

Kết quả trong Output cho ta ba bảng, trong đó ta quan tâm đến bảng cuối cùng.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Truoc - Sau	-3.640	8.780	1.242	-6.135	-1.145	-2.931	49	.005

Từ bảng này cho ta sig.(2-tailed)=0.005, vì bài toán của ta là bài toán kiểm định bên trái, giá trị $t = -2.931 < 0$ nên P-giá trị $= \frac{1}{2}$ sig.(2-tailed)<0.05 nên bác bỏ H_0.

Vậy ở mức ý nghĩa 0.05, việc treo thưởng có tác dụng làm tăng số sản phẩm trung bình của các công nhân.

5.4. Bài tập

5.4.1. Ước lượng và kiểm định trung bình một tổng thể với một số

Bài tập 5.1. Trong một báo cáo có đoạn: một điều tra trên 300 nữ sinh đại học cho ta ước lượng khoảng tin cậy 90% về chiều cao của nữ sinh đại học ở Việt Nam là từ 1.5m đến 1.7m. Bạn hiểu thế nào về khoảng [1.5m ; 1.7m] này? Có phải mọi nữ sinh đại học đều có chiều cao trong khoảng này không?

Bài tập 5.2. Điều tra về đời sống y tế, giáo dục của người dân được thực hiện bởi một tổ chức phi chính phủ trên mẫu 500 hộ gia đình vùng miền núi phía bắc cho ta số liệu trong file **DoiSong.sav** trong đó các chi phí được tính với đơn vị nghìn đồng.

1. Ước lương điểm và ước lượng khoảng tin cậy 90% cho trung bình của chi tiêu cho chăm sóc y tế trong năm vừa qua (cột YPhi) của tổng thể hộ dân miền núi phí bắc.
2. Ước lương điểm và ước lượng khoảng tin cậy 95% cho trung bình chi tiêu giáo dục (cột HocPhi) năm qua của tổng thể các hộ dân miền núi phí bắc.

Bài tập 5.3. Vừa qua các phương tiên thông tin đại chúng có đưa tin kế quả thống kê của các tỉnh cho thấy có ít nhất 80% người dân hài lòng với dịch vụ công. Một tổ chức muốn kiểm định khẳng định trên.

1. Hãy thiết lập cặp giả thuyết cho bài toán.
2. Giải thích các sai lầm có thể mắc phải.

Bài tập 5.4. Ta trở lại với dữ liệu **DoiSong.sav** ở trên. Tại mức ý nghĩa 0.05

1. Có thể chấp nhận trung bình tổng chi phí cho giáo dục của các hộ dân miền núi phía bắc là 200 (nghìn) hay không?
2. Có thể khẳng định rằng chi phí trung bình cho chăm sóc sức khỏe của các hộ dân miền núi phía bắc là nhỏ hơn 500 (nghìn) hay không?
3. Có thể khẳng định rằng trung bình chi phí đi lại (cột PhiDiLai) trong năm của các hộ dân miền núi phía bắc là không quá 500 (nghìn) hay không?
4. Có thể khẳng định rằng trung bình tổng chi phí cho các ngày lễ trong năm (cột PhiLe) của các hộ dân miền núi phía bắc là ít nhất 200 (nghìn) hay không?

Bài tập 5.5. Một công ty sản xuất ngũ cốc cho rằng tỷ lệ chất béo trung bình trong 100mg sản phẩm của họ nhỏ hơn 1.5mg. Nghiên cứu trên 15 mẫu, mỗi mẫu gồm 100mg ngũ cốc thì thu được só liệu về lượng chất béo (số liệu có trong file **HamLuongBeo.sav**. Tại mức ý nghĩa 5% bạn kết luận gì về công bố của công ty?

5.4.2. Kiểm định tỉ lệ một tổng thể với một số

Bài tập 5.6. Ta trở lại với dữ liệu điều tra trong file **DoiSong.sav**. Chọn mức ý nghĩa 5%.

1. Có thể coi tỉ lệ hộ nghèo ở vùng này là thấp hơn 25% không?

2. Biết rằng trong dữ liệu những hộ có chi phí giáo dục bằng 0 là những hộ không có thành viên nào đi học trong năm qua. Kiểm định xem tỉ lệ số hộ không có người đi học trong năm qua tại vùng trên có lớn hơn 25 % không?

3. Kiểm định xem tỉ lệ số hộ có chi tiêu mua sắm định kì (PhiDinhKi) không quá 250 nghìn có chiếm đúng 50% không.

4. Kiểm định cho khẳng định tỉ lệ số hộ có không quá 2 người chiếm không đến 25 %.

5. Lọc ra những người thuộc hộ nghèo. Kiểm định cho khẳng định có ít hơn 40 % những gia đình hộ nghèo có từ 5 người trở lên.

6. Lọc ra những người ở khu vực thành phố. Có phải tỉ lệ hộ nghèo ở đây nhỏ hơn 15 % không?

5.4.3. Kiểm định trung bình hai tổng thể

Bài tập 5.7. Với dữ liệu trong **DoiSong.sav** hãy kiểm định ở mức ý nghĩa 5% cho các khẳng định sau cho tổng thể các hộ ở miền núi phía bắc:

1. Chi phí giáo dục trung bình trong năm qua của các hộ ở thành phố là cao hơn so với nông thôn.

2. Chi phí tiền thuốc (cột PhiThuoc) trung bình của những hộ ở nông thôn và thành phố là như nhau.

3. Chi phí sinh hoạt gia đình (PhiAnUong) trung bình của các hộ nghèo là thấp hơn so với các hộ không thuộc hộ nghèo.

4. Trung bình tiền điện nước tháng qua của các hộ nghèo thấp hơn các hộ không thuộc diện nghèo.

5. Chi phí chăm sóc y tế trung bình trong năm qua của các hộ nghèo và các hộ không thuộc diện nghèo là như nhau.

6. Tạo biến mã hóa phân loại các hộ theo số người trong gia đình thành 2 nhóm: nhóm có số thành viên nhỏ hơn 4 và nhóm có số thành viên từ 4 trở lên. So sánh xem chi phí đi lại trung bình của các hộ của hai nhóm trên có như nhau không.

Bài tập 5.8. Với dữ liệu trong **DoiSong.sav** hãy kiểm định ở mức ý nghĩa 5% cho các khẳng định sau cho tổng thể các hộ ở miền núi phía bắc:

1. Chi phí trung bình cho chăm sóc y tế và giáo dục trong năm qua là như nhau.

2. Chi phí đi lại trung bình trong năm lớn hơn chi phí trung bình dành cho mua sắm định kì (PhiDinhKi).

Bài tập 5.9. Tại một văn phòng công ty lớn người ta muốn cất nhắc một trong hai phó phòng lên làm trưởng phòng. Ban giám đốc công ty thu thập đánh giá của nhân viên bộ phận đó để có thông tin tham khảo. Dữ liệu trong file **DiemDanhGia.csv** chứa điểm đánh giá của 40 nhân viên được chọn ngẫu nhiên về 2 phó phòng An và Bình. Hãy kiểm định tại mức ý nghĩa 5% khẳng định cho rằng trung bình điểm đánh giá của An cao hơn Bình.

Bài tập 5.10. Một spa công bố sau 2 tuần tập luyện tại spa của họ vòng eo sẽ giảm. Số liệu trong file VongeEo.sav ghi lại vòng eo của một số người tham gia luyện tập tại spa này trong thời gian hai tuần trước và sau luyện tập. Hãy kiểm định công bố của spa.

Chương 6

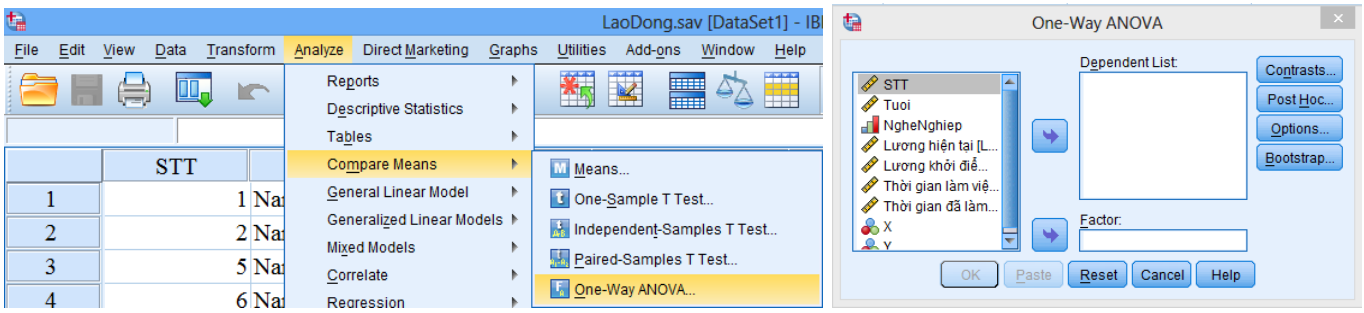
Phân tích phương sai

Để so sánh nhiều trung bình tổng thể (từ 3 tổng thể trở lên) ta dùng phương pháp phân tích phương sai. Ở đây ta chỉ thực hành với phân tích phương sai một nhân tố.

Ta cần các giả định sau:

- Các tổng thể tuân theo phân phối chuẩn có phương sai bằng nhau.
- Các mẫu chọn ra là độc lập.

Để tiến hành phân tích phưng sai một nhân tố trong SPSS ta vào **Analyze** $\longrightarrow$ **Compare Means** $\longrightarrow$ **One Way ANOVA**:



Hình 6.1: Kiểm định so sánh trung bình nhiều tổng thể

Trong hộp thoại xuất hiện ta chọn biến định lượng cần so sánh vào khung **Dependent List** và biến phân loại vào khung **Factor**. Chú ý rằng biến phân loại phải là biến định lượng.

6.1. Ví dụ

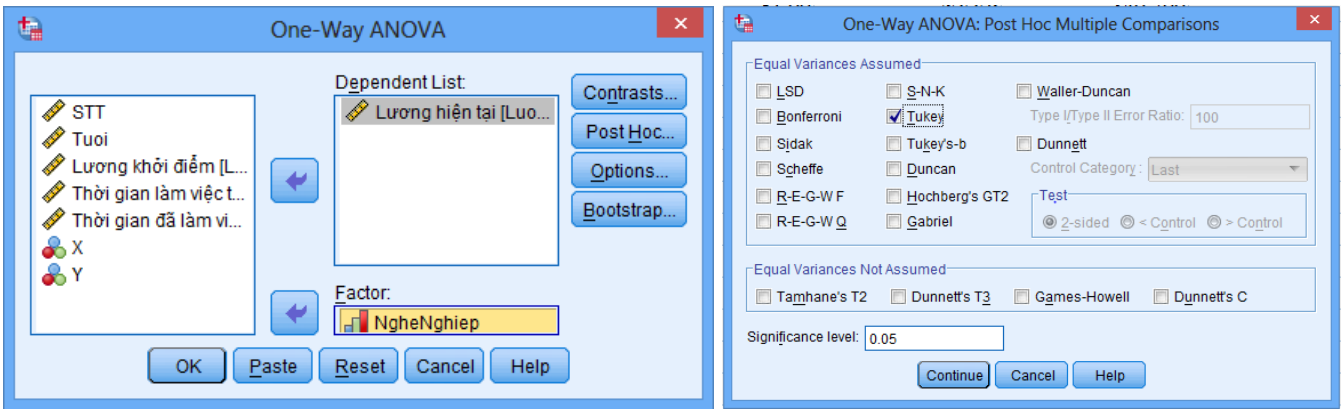
***Ví dụ* 6.1.1.** *Giả sử rằng lương của người công nhân, NVVP và quản lí ở quận A đều tuân theo phân phối chuẩn có phương sai bằng nhau.*

Hãy kiểm định tại mức ý nghĩa 5% xem lương trung bình của người lao động những nhóm nghề nghiệp khác nhau ở quận A có như nhau không? Trong trường hợp khác nhau, hãy cho biết lương trung bình nhóm nào là cao nhất.

Gọi m_1, m_2, m_3 lần lượt là tiền lương trung bình của 3 tổng thể công nhân, NVVP và quản lí.

- $H_0 : m_1 = m_2 = m_3$ $\qquad$ H_1 : Có i, j để $m_i \neq m_j$.

- Sau khi vào thủ tục như hình 6.1, trong hộp thoại **One - Way ANOVA** hiện ra ta chuyển biến Lương hiện tại vào khung phía trên, NgheNghiep vào khung **Factor**. Nhấp vào nút **Post Hoc** để chọn kiểm định phân tích sâu **Tukey**.



Hình 6.2: Nhấp vào **Post Hoc** để chọn **Tukey**

Cuối cùng nhấp **Continue** rồi **OK**. Kết quả trong Output như sau: Trong bảng phân tích so

ANOVA

Lương hiện tại

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	89438483926	2	44719241963	434.481	.000
Within Groups	48478011510	471	102925714.5		
Total	1.379E+11	473			

Kết quả của bài toán so sánh trung bình

Multiple Comparisons

Dependent Variable: Lương hiện tại
Tukey HSD

(I) NgheNghiep	(J) NgheNghiep	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
NVVP	CongNhan	-$3,100.349	$2,023.760	.277	-$7,858.50	$1,657.80
	QuanLy	-$36,139.258*	$1,228.352	.000	-$39,027.29	-$33,251.22
CongNhan	NVVP	$3,100.349	$2,023.760	.277	-$1,657.80	$7,858.50
	QuanLy	-$33,038.909*	$2,244.409	.000	-$38,315.84	-$27,761.98
QuanLy	NVVP	$36,139.258*	$1,228.352	.000	$33,251.22	$39,027.29
	CongNhan	$33,038.909*	$2,244.409	.000	$27,761.98	$38,315.84

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

Kết quả phân tích sâu Tukey

sánh trung bình ta có p - giá trị của bài toán là **Sig.** $< 0.001 < 0.05$ nên bác bỏ H_0. Như vậy lương trung bình của các nhóm nghề tại quận A là không như nhau.

Để biết nhóm nghề nào có lương cao nhất ta dùng kết quả của bảng kết quả phân tích sâu Tukey, qua bảng này ta thấy p - giá trị cho bài toán so sánh lương trung bình nhóm nghề NVVP và nhóm QuanLy là nhỏ hơn $0.001 < 0.05$ nên lương trung bình hai nhóm này là khác nhau, và do hiệu trung

bình âm nên có thể coi lương nhóm nghề QuanLy cao hơn so với NVVP. Tương tự lương nhóm quản lý cao hơn công nhân. So sánh lương trung bình hai nhóm CongNhan và NVVP cho ta p - giá trị $0.277 > 0.05$ nên ta chấp nhận hai nhóm nghề này có lương trung bình bằng nhau.

Tóm lại nhóm nghề quản lý có lương cao nhất trong 3 nhóm.

6.2. Bài tập

Bài tập 6.1. Người ta so sánh tác dụng của việc tập huấn kết hợp với việc luyện tập thể dục so với việc chỉ luyện tập thể dục hoặc chỉ được chăm sóc y tế thông thường trong việc hoạt động thể lực của các bệnh nhân vừa được phẫu thuật tim. Đánh giá kết quả là điểm tự tin được cho theo một thang điểm nhất định, điểm càng cao càng thể hiện sự tự tin vào sức khỏe bản thân. Người ta chọn ngẫu nhiên một số người tham gia vào 3 nhóm, nhóm 1 tham gia tập huấn kết hợp với việc luyện tập thể dục, nhóm 2 chỉ tập huấn, nhóm 3 chỉ chăm sóc y tế. Kết quả được cho trong dữ liệu **DanhGia.xls**. Tại mức ý nghĩa 5%, hãy kiểm định xem có sự khác biệt giữa 3 nhóm hay không? Giả sử rằng các giả thiết của bài toán phân tích phương sai một yếu tố đều được thỏa mãn.

Bài tập 6.2. Trong file **DieuTraDienThoai.xls** là thông tin của 2000 người được điều tra một cách ngẫu nhiên. Trong đó TienDT là số tiền điện thoại tháng vừa qua của họ. HonNhan là biến mã hóa tình trạng hôn nhân thành dạng số (1 - chưa có gia đình, 2 - đang có gia đình, 3 - ly hôn).

1. Ở mức ý nghĩa 5% hãy xét xem tiền điện thoại trung bình một tháng của 3 tổng thể chưa có gia đình, đang có gia đình và ly hôn có bằng nhau không? Trong trường hợp khác nhau hãy thực hiện phân tích sâu để biết nhóm nhiều nhất, ít nhất. Giả sử tiền điện thoại một tháng bất kì của 3 tổng thể này đều tuân theo phân phối chuẩn có phương sai bằng nhau. Việc chọn mẫu là độc lập.

2. Tại mức ý nghĩa 0.05 hãy xét xem trung bình tiền điện thoại tháng vừa qua của các thuê bao của những nhà mạng khác nhau có bằng nhau không. Giả sử các giả thiết của bài toán phân tích phương sai đều được thỏa mãn.

Bài tập 6.3. Nghiên cứu điểm hài lòng hôn nhân của các nhóm khác nhau trong mối quan hệ với thời gian chung sống (1 - chưa đến 5 năm, 2 - từ 5 năm đến 10 năm, 3 - trên 10 năm) và với điều kiện kinh tế (1 - nghèo, 2 - trung bình, 3 - khá giả) cho ta kết quả ở file **HaiLongTrongHonNhan.csv**. Giả sử điểm hài lòng của mỗi nhóm tuân theo phân phối chuẩn có phương sai bằng nhau, mẫu được chọn ngẫu nhiên độc lập. Hãy kiểm định ở mức ý nghĩa 0.05:

1. điểm hài lòng trung bình của các tổng thể chia theo điều kiện kinh tế nói trên có như nhau không. Trong trường hợp khác nhau thì nhóm nào có điểm hài lòng tốt nhất.

2. điểm hài lòng trung bình của các tổng thể chia theo thời gian gian chung sống nói trên có như nhau không. Trong trường hợp khác nhau thì nhóm nào có điểm hài lòng tốt nhất.

Chương 7

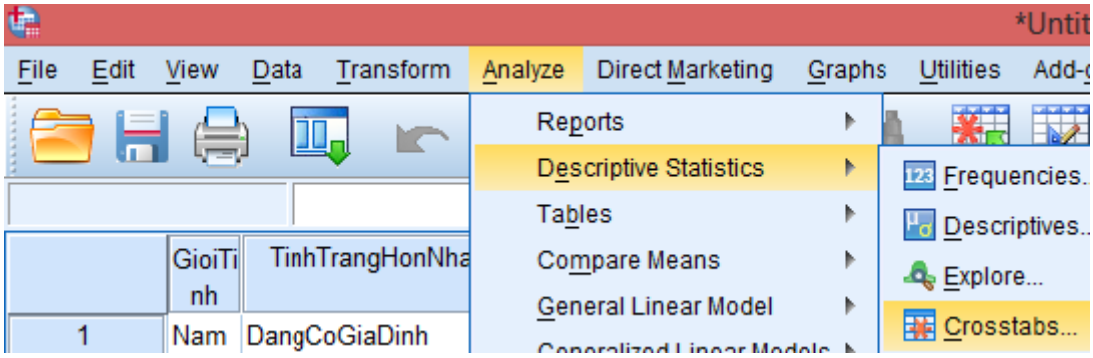
Kiểm định tính độc lập và so sánh tỉ lệ hai tổng thể

Để kiểm định tính độc lập hay phụ thuộc của hai yếu tố ta thiết lập bài toán:

H_0: Hai yếu tố độc lập.

H_1: Hai yếu tố là có mối liên hệ với nhau.

Trong SPSS để làm thủ tục kiểm định này ta vào **Analyze** $\longrightarrow$ **Descriptive Statistic** $\longrightarrow$ **Crosstabs**.



Hình 7.1

Trong hộp thoại **Crosstabs** ta chọn 2 biến định tính lần lượt vào **Row(s)** và **Column(s)**, sau đó nhấp vào tùy chọn **Statistics** tích chọn **Chi-square**. Nhấp **Continue**, nhấp **OK**

7.1. Ví dụ

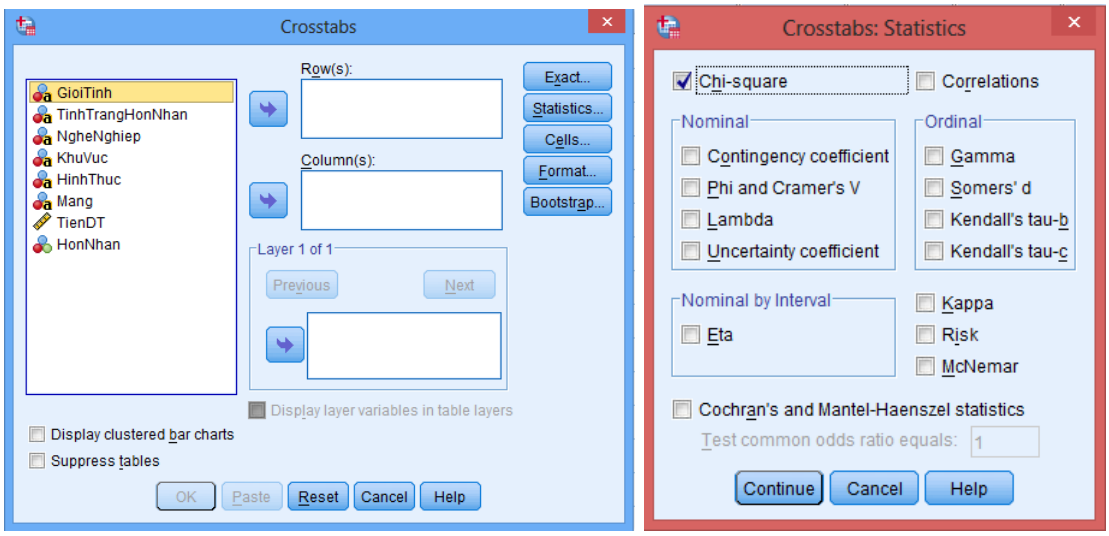
***Ví dụ* 7.1.1.** *Trong file "DieuTraDienThoai.xls" là thông tin của 2000 người được điều tra một cách ngẫu nhiên. Trong đó TienDT là số tiền điện thoại tháng 1 của họ. HonNhan là biến mã hóa tình trạng hôn nhân thành dạng số (1 - chưa có gia đình, 2 - đang có gia đình, 3 - ly hôn).*

1. *Ở mức ý nghĩa 5% hỏi có mối liên hệ giữa tình trạng hôn nhân và hình thức trả tiền điện thoại không?*

2. *Ở mức ý nghĩa 5% hỏi yếu tố giới tính và hình thức thuê bao có mối liên nhau không?*

   Từ đó có thể cho rằng tỉ lệ nam trong những người dùng hình thức trả trước và tỉ lệ nam trong những người dùng hình thức trả sau là như nhau không?

Lời giải

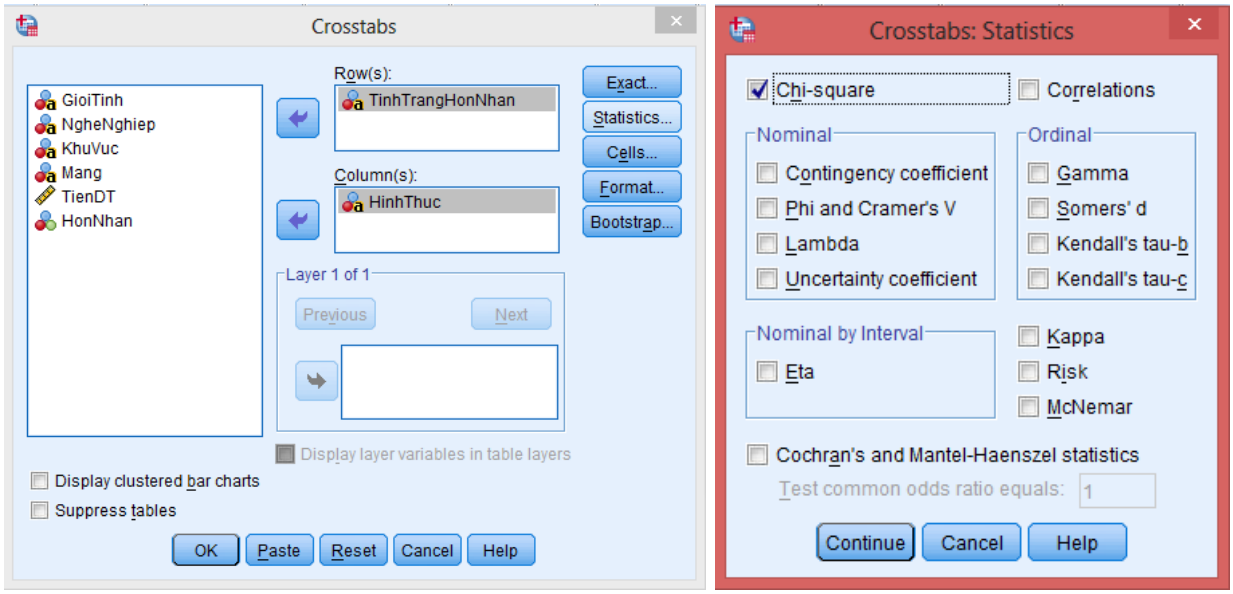


Hình 7.2: Kiểm định Chi - square kiểm chứng tính độc lập của hai biến định tính

1. H_0: Hai yếu tố độc lập.

   H_1: Hai yếu tố là có mối liên hệ với nhau.

   Đầu tiên, thao tác như hình 7.1, trong hộp thoại hiện ra như hình 7.4 ta chọn hai yếu tố TinhTrangHonNhan và HinhThuc vào Row(s) và Column(s). Tiếp sau đó nhấp vào tùy chọn **Statistics** tích chọn **Chi-square**. Nhấp **Continue**, nhấp **OK**



Hình 7.3: Nhấp **Statistics** tích chọn **Chi-square**

Kết quả trong Output cho ta 3 bảng, bảng cuối cùng là kết quả của kiểm định mối liên hệ giữa 2 yếu tố.

Trong bảng này (hình 7.4) cho ta **Sig.(2-tailed)**$= 0.176 > 0.05$ nên chấp nhận H_0. Tức là ở mức ý nghĩa 0.05 hai yếu tố nói trên là độc lập nhau.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3.479[a]	2	.176
Likelihood Ratio	3.606	2	.165
N of Valid Cases	2000		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 60.50.

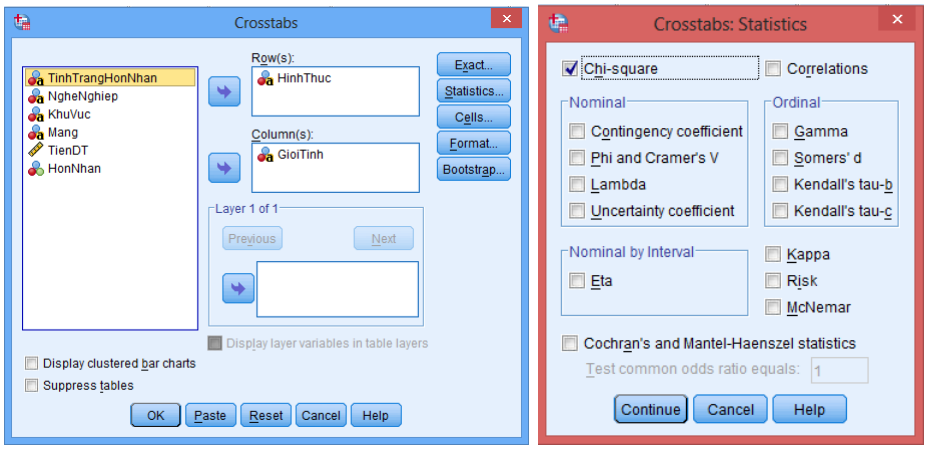
Hình 7.4: Kết quả của bài toán kiểm định

2. Cặp giải thiết:

H_0 : Yếu tố giới tính và hình thức thanh toán điện thoai là độc lập nhau.

H_1 : Hai yếu tố trên có mối liên hệ với nhau.

Quá trình và kết quả kiểm định như sau: Từ bảng kết quả ta có p - giá trị = 0.38 (=0.48 khi



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.770[a]	1	.380		
Continuity Correction[b]	.686	1	.408		
Likelihood Ratio	.769	1	.381		
Fisher's Exact Test				.404	.204
N of Valid Cases	2000				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 263.07.

b. Computed only for a 2x2 table

Hình 7.5: Kết quả của bài toán kiểm định

đã làm liên tục) > 0.05 nên chấp nhận H_0.

Vậy hai yếu tố giới tính và hình thức thanh toán tiền điện thoại là độc lập với nhau.

Từ đó suy ra tỉ lệ nam giới trong tổng thể những người sử dụng hình thức trả trước và tỉ lệ nam trong tổng thể người dùng hình thức trả sau là như nhau.

7.2. Bài tập

Bài tập 7.1. Dùng file **KhaoSatLuongNhanVien.sav** để kiểm định các bài toán sau tại mức ý nghĩa 5%:

1. có mối liên hệ giữa giới tính và loại nghề không?
2. Có mối liên hệ giữa giới tính và dân tộc không? Từ đó kết luận xem có thể cho rằng tỷ lệ nam giới trong từng dân tộc có như nhau không?

Bài tập 7.2. Để xem có mối liên hệ giữa thời gian chung sống và sự hài lòng với cuộc hôn nhân của các bà vợ hay không người ta tiến hành phỏng vấn 106 bà vợ và được kết quả trong file **HonNhan.csv**. Trong đó ThoiGian là thời gian chung số với nhau cho đến nay: 1 - chưa đến 5 năm, 2 - từ 5 năm đến 10 năm, 3 - trên 10 năm. HaiLong là mức độ hài lòng của họ. Ở mức ý nghĩa 0.05 có mối liên hệ giữa thời gian chung sống và mức hài lòng của các bà vợ hay không.

Bài tập 7.3. Trong file **DieuTraDienThoai.xls** là thông tin của 2000 người được điều tra một cách ngẫu nhiên. Trong đó TienDT là số tiền điện thoại tháng 1 họ. HonNhan là biến mã hóa tình trạng hôn nhân thành dạng số (1 - chưa có gia đình, 2 - đang có gia đình, 3 - ly hôn).

1. Ở mức ý nghĩa 5%, yếu tố nghề nghiệp và yếu tố mạng điện thoại có độc lập nhay không?
2. Tại mức ý nghĩa 5%, tỉ lệ những người ly hôn trong tổng thể người dùng hình thức thanh toán trả trước và tỉ lệ này trong tổng thể người dùng trả sau có khác nhau không?
3. Tại mức ý nghĩa 5%, ở miền bắc tỉ lệ dùng hình thức thanh toán trả sau và tỉ lệ dùng hình thức thanh toán trả trước có như nhau không?
4. Tại mức ý nghĩa 5%, tỉ lệ dùng hơn 100 (nghìn) tiền điện thoại ở hai tổng thể: tổng thể thuê bao trả trước, tổng thể thuê bao trả sau có như nhau không? (Hướng dẫn: tạo biến định tính mới là mã hóa của biến TienDT thành 2 nhóm: nhóm > 100 và nhóm ≤ 100 sau đó dùng kiểm định Chi - square).

Tài liệu tham khảo

[1] Đào Hữu Hồ, *Thống kê xã hội học*, Nhà xuất bản Giáo dục 2007.

[2] Nguyễn Hữu Tân, *Thống kê xã hội học I*, Đại học Đà Lạt, 2006.

[3] Một số thông tin và số liệu thống kê được lấy từ các website:

- http://www.gopfp.gov.vn
- http://wearesocial.net
- http://vi.wikipedia.org