HOÀNG TRỌNG - CHU NGUYỄN MỘNG NGỌC

PHÂN TÍCH DỮ LIỆU NGHIÊN CỨU VỚI

SPSS

ĐẠI HỌC KINH TẾ TP HỒ CHÍ MINH

HOÀNG TRỌNG – CHU NGUYỄN MỘNG NGỌC

PHÂN TÍCH DỮ LIỆU NGHIÊN CỨU VỚI SPSS

(dùng với SPSS các phiên bản 11.5, 13, 14, 15, 16)

Tập 1

NHÀ XUẤT BẢN HỒNG ĐỨC
NĂM 2008

Quyển sách dành cho những bạn đang làm
đề tài nghiên cứu khoa học, khóa luận hay luận văn tốt nghiệp

Hãy đọc kỹ lời nói đầu và xem mục lục
trước khi bạn đi vào nội dung của quyển sách

ĐỊA CHỈ TẢI FILE THỰC HÀNH

Để lấy các file dữ liệu thực hành cùng với sách Phân Tích Dữ Liệu Nghiên Cứu với SPSS, bạn vào một trong các trang web sau để tải file xuống:

Trang web của Khoa Toán – Thống Kê, ĐH Kinh Tế TPHCM (chọn mục Sách và Tài Liệu):
http://www.fos.ueh.edu.vn

Trang web cao học kinh tế:
http://caohockinhte.info/forum/showthread.php?t=3680

Trang web của công ty tư vấn:
http://www.thepathfinder.vn/index.php?option=thongtinnghiencuu&task=view&id=14

Nếu có trục trặc xin vui lòng email đến địa chỉ:

phantichdulieu@yahoo.com.vn

LỜI NÓI ĐẦU

Quyển sách Phân tích Dữ Liệu Nghiên Cứu Với SPSS (Nhà Xuất Bản Thống Kê, 2005) đã ra đời được ba năm. Tác giả đã nhận được nhiều ý kiến góp ý, yêu cầu bổ sung của nhiều bạn đọc. Ý kiến của bạn đọc xoay quanh vấn đề chính.

Một là bổ sung nội dung, ví dụ như: Sử dụng Custom Tables (một số phiên bản SPSS sau này do bạn đọc cài đặt hay do đĩa nguồn cài đặt thiếu, các bạn chỉ có Custom Tables mà không có Basic Tables hay General Tables ...); Vẽ đồ thị trong Excel (vẽ đồ thị trong SPSS không quen và ít tiện lợi như trong Excel); phân tích phương sai hai yếu tố nguyên nhân; Tạo các biến giả và sử dụng biến giả trong hồi qui đối với các biến độc lập định tính, chẩn đoán và tuyến tính hóa các biến nguyên nhân; Lập các bản đồ nhận thức (bản đồ định vị); Gia trọng các quan sát, Ghép trộn dữ liệu... Chúng tôi đã cố gắng bổ sung theo những yêu cầu này. Còn một vài nội dung khác, hiện nay do số lượng bạn đọc có nhu cầu sử dụng còn ít chúng tôi sẽ bổ sung trong lần tái bản sau.

Hai là việc sử dụng quyển sách này với các phiên bản mới hơn của SPSS như 13.0, 15.0 và 16.0. Về phiên bản của SPSS, sau khi khảo sát và sử dụng thử các phiên bản SPSS 13, 15, 16, chúng tôi nhận thấy các giao diện, các lệnh thực hiện hoàn toàn tương tự nhau. Chúng tôi cũng vẫn sử dụng Phiên bản 11.5 và 13.0 vì sự gọn nhẹ, ít lỗi của hai phiên bản chuẩn này. Các phiên bản sau có bổ sung một vài tiện ích mới nhưng hiếm khi được sử dụng đối với người sử dụng thông thường. Bạn đọc yên tâm sử dụng quyển sách này với bất kỳ phiên bản của SPSS từ 11.5 đến 16.0.

Trong lần xuất bản này, chúng tôi tách thành hai tập. Tập 1 phục vụ cho nhu cầu xử lý và phân tích căn bản của các sinh viên bậc cử nhân đang học các môn học liên quan như Thống Kê, Kinh Tế Lượng, Phương Pháp Nghiên Cứu, Phân tích Dữ Liệu. Tập 2 dành cho sinh viên học chuyên ngành muốn đi sâu vào phân tích dữ liệu, học viên cao học , người phân tích dữ liệu chuyên nghiệp.

Khác với lần xuất bản trước, lần xuất bản này chúng tôi không kèm đĩa chứa dữ liệu mẫu với sách vì dung lượng các file này khá nhỏ. Mặt khác nhiều bạn đọc đã yêu cầu chúng tôi gửi file thực hành vì đối khi sách không có đĩa, hay đĩa bị hỏng, hay khi có nhu cầu sử dụng mà đĩa đã thất lạc đâu mất, nhất là các bạn ở xa. Chúng tôi để các file này trong 1 file nén và để trên mạng để bạn đọc ở tất cả mọi nơi đều có thể download xuống. Trong trường hợp các bạn gặp trục trặc về việc tải file xuống hoặc có thắc mắc về việc sử dụng các file này, bạn hãy liên lạc với chúng tôi qua hộp thư điện tử sau đây:

phantichdulieu@yahoo.com.vn

Việc biên soạn một quyển sách nào cũng khó tránh khỏi các sai sót. Mọi sai sót, nếu có, trong quyển sách này hoàn toàn là do người biên soạn. Chúng tôi mong được nhiều ý kiến đóng góp của tất cả các sinh viên, giảng viên và những người làm công tác nghiên cứu để lần tái bản tiếp theo, quyển sách được hoàn chỉnh hơn. Thư góp ý về nội dung quyển sách xin gửi về:

Hoàng Trọng
Khoa Toán – Thống Kê, Đại Học Kinh Tế TPHCM
Số 91 đường 3/2, quận 10, TP Hồ Chí Minh
Email: htrong@ueh.edu.vn

Chu Nguyễn Mộng Ngọc
Email: chunguyenmongngoc@yahoo.com

Xin chân thành cảm ơn và chúc các bạn thành công!

TP Hồ Chí Minh, tháng 09 năm 2008
Tác giả

Hoàng Trọng
Chu Nguyễn Mộng Ngọc

MỤC LỤC

CHƯƠNG IX: TƯƠNG QUAN VÀ HỒI QUI TUYẾN TÍNH

CHƯƠNG MỞ ĐẦU
GIỚI THIỆU PHÂN TÍCH DỮ LIỆU

1. NGHIÊN CỨU VÀ PHÂN TÍCH DỮ LIỆU

Nghiên cứu định lượng một vấn đề kinh tế xã hội thường bao gồm các bước cơ bản sau:

- Xác định vấn đề nghiên cứu
- Thu thập dữ liệu
- Xử lý dữ liệu
- Phân tích dữ liệu
- Báo cáo kết quả

Xác định vấn đề nghiên cứu

Xác định rõ ràng và chính xác vấn đề nghiên cứu là điều kiện đầu tiên để thực hiện tốt một cuộc nghiên cứu. Định nghĩa thật rõ vấn đề nghiên cứu giúp việc thu thập dữ liệu tiến hành nhanh gọn, chính xác. Nếu không thì chúng ta rất khó thu thập dữ liệu một cách hiệu quả vì chúng ta sẽ mất nhiều thời gian để thu thập, xử lý những dữ liệu không cần thiết, trong khi đó những dữ liệu quan trọng rất cần cho phân tích lại không có.

Thu thập dữ liệu

Việc thiết kế quá trình thu thập dữ liệu, ít khi được chú ý đúng mức trong lý thuyết cũng như thực tế nghiên cứu nói chung và phân tích dữ liệu nói riêng. Nhưng nguy cơ này đã bị làm giảm nhẹ đi bởi niềm tin đơn giản rằng khối lượng tính toán nhiều có thể giúp khắc phục những thiếu sót trong thiết kế thu thập dữ liệu. Thiết kế các cách thức thu thập dữ liệu là công việc quan trọng đối với phân tích dữ liệu thống kê. Hai khía cạnh quan trọng của nghiên cứu thống kê là: Tổng thể - tập hợp các phần tử mà chúng ta quan tâm trong nghiên cứu; Mẫu - một tập hợp con của tổng thể.

Chúng ta phải bắt đầu với việc nhấn mạnh vào tầm quan trọng của xác định tổng thể nghiên cứu mà chúng ta muốn suy diễn. Thành công của cuộc nghiên cứu phụ thuộc rất nhiều vào việc xác định

phạm vi tổng thể nghiên cứu và đơn vị điều tra, đơn vị báo cáo. Ví dụ như khi nghiên cứu về sự hội nhập của người nhập cư vào thành phố, thì cần xác định tổng thể nghiên cứu ở đây là gì, như thế nào là người nhập cư, đơn vị điều tra, đơn vị báo cáo trong trường hợp này là cá nhân hay hộ gia đình ... Và sau khi thu thập, đến giai đoạn phân tích dữ liệu thì đơn vị phân tích là cá nhân hay hộ gia đình hay cả hai. Do đó để phân tích tốt thì ngay trong giai đoạn đầu của thiết kế quá trình thu thập dữ liệu phải lường trước các yêu cầu của phân tích để có thể thu thập đủ và đúng các dữ liệu cần cho phân tích.

Cốt lõi của phân tích dữ liệu là suy diễn thống kê, tức là mở rộng hiểu biết của bạn từ một mẫu ngẫu nhiên thành hiểu biết về tổng thể. Trong toán học điều này gọi là suy diễn quy nạp. Tức là tri thức về cái chung xuất phát từ cái riêng. Mục tiêu của suy diễn thống kê là thu được thông tin về tổng thể từ thông tin chứa trong mẫu quan sát. Bởi vì không chỉ thu thập dữ liệu toàn bộ tổng thể là không khả thi, mà mẫu còn là một cách thực tế để thu thập dữ liệu vì hạn chế về thời gian và chi phí trong nghiên cứu.

Dữ liệu có thể được thu thập từ những nguồn có sẵn hay qua quan sát, nghiên cứu thử nghiệm. Trong nghiên cứu thử nghiệm, biến cần nghiên cứu được ghi nhận. Một hay nhiều yếu tố ảnh hưởng trong nghiên cứu được kiểm soát để dữ liệu thu được có thể phản ảnh tác động của các yếu tố ảnh hưởng đến biến cần nghiên cứu. Còn trong nghiên cứu quan sát, không kiểm soát tác động của các yếu tố ảnh hưởng. Điều tra có lẽ là loại nghiên cứu quan sát phổ biến nhất.

Xử lý dữ liệu

Dữ liệu thường được ghi chép thủ công trên các bản ghi chép. Trừ phi số lượng các quan sát ít và số lượng biến ít, thường thì dữ liệu phải được phân tích trên máy tính. Lúc đó dữ liệu phải đi qua 3 bước sau:

- Mã hóa: ngoại trừ một số dữ liệu định lượng (dưới dạng số) thì không cần mã hóa, còn các dữ liệu định tính (không phải dưới dạng số) cần được chuyển đổi thành các con số
- Nhập liệu: Dữ liệu được nhập và lưu trữ bởi ít nhất hai người nhập liệu độc lập khác nhau. Thông thường trong thực tế nhập dữ liệu từ bảng câu hỏi vào máy tính là nhập hai lần.

- Hiệu chỉnh: dữ liệu được kiểm tra bằng cách so sánh hai tập hợp dữ liệu được nhập độc lập với nhau. Lý tưởng thì trong lần nhập thứ hai người nhập liệu là người khác người nhập lần thứ nhất và người này sẽ chú ý phát hiện những sai lệch giữa dữ liệu nhập lần 1 và nhập lần 2. Kiểm tra bằng cách nhập 2 lần bảo đảm mức độ chính xác lên đến 99.8% cho tất cả các lần gõ phím.

Phân tích dữ liệu

Phân tích dữ liệu thống kê chia các phương pháp phân tích dữ liệu thành hai loại: các phương pháp thăm dò và các phương pháp khẳng định. Các phương pháp thăm dò được dùng để khám phá ý nghĩa của dữ liệu bằng các phép tính số học đơn giản và các biểu đồ đơn giản tóm tắt dữ liệu (thống kê mô tả). Các phương pháp khẳng định dùng các ý tưởng trong lý thuyết xác suất để trả lời các vấn đề nghiên cứu cụ thể. Xác suất có vai trò quan trọng trong việc ra quyết định vì nó cung cấp một cơ chế để đo lường, biểu diễn, và phân tích trong những tình huống không có đủ thông tin (không thể biết hết toàn bộ tổng thể) liên quan đến các vấn đề kinh tế xã hội trong tương lai.

Lưu ý rằng phân tích dữ liệu chỉ là một giai đoạn của cả một quá trình nghiên cứu, do đó không thể có phân tích tốt mà không nắm vững toàn bộ quá trình nghiên cứu từ mục tiêu cho đến kết quả cuối cùng muốn đạt được. Không thể có phân tích dữ liệu tốt nếu cơ sở dữ liệu để phân tích không được thiết kế để thu thập tốt, không được xử lý chuẩn bị tốt cho phân tích.

Phân tích dữ liệu không phải chỉ là sử dụng một cách máy móc các kỹ thuật thống kê đơn thuần để có kết luận là chấp nhận hay bác bỏ một giả thuyết, hay xây dựng được một mô hình diễn tả mối liên hệ giữa các yếu tố đang nghiên cứu; mà là một "nghệ thuật" làm cho dữ liệu trở thành những chứng cứ thống kê có cơ sở cho việc hiểu biết, gia tăng tri thức và ra quyết định. Phân tích dữ liệu phải được vận dụng trong mối liên hệ chặt chẽ với các giai đoạn khác của quá trình nghiên cứu ở chỗ người làm công việc phân tích dữ liệu phải tham gia ngay từ đầu vào quá trình thiết kế nghiên cứu, triển khai thu thập dữ liệu và chưa thể kết thúc công việc nếu báo cáo kết quả chưa viết xong. Hoặc người nghiên cứu ngay khi thiết kế nghiên cứu phải hình dung trước những vấn đề quan trọng của phân tích dữ liệu.

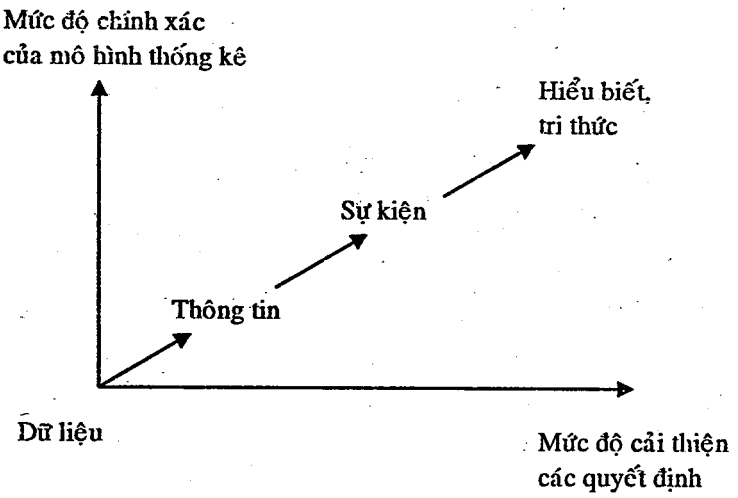
Báo cáo kết quả

Thông qua suy diễn, từ dữ liệu mẫu thu thập được ước lượng, kiểm định và các mô hình phân tích khác sẽ giúp khẳng định các đặc tính của tổng thể. Các kết quả có thể được báo cáo dưới dạng bảng, đồ thị hay các số phần trăm. Vì chỉ nghiên cứu trên một mẫu nhỏ chứ không phải toàn bộ tổng thể, các kết quả báo cáo phải phản ảnh tính không chắc chắn qua việc sử dụng các phát biểu theo kiểu xác suất và khoảng giá trị đưa ra.

Một khía cạnh quan trọng trong nghiên cứu kinh tế xã hội là nghiên cứu để tìm hiểu, đưa ra các quyết định thay đổi cho tương lai. Phán đoán tốt, trực giác và quan tâm đến thực trạng của nền kinh tế, môi trường kinh doanh, môi trường xã hội có thể cho người nghiên cứu một ý tưởng sơ bộ hay "cảm giác" về những gì có thể xảy ra trong tương lai. Tuy nhiên, chuyển từ cảm giác thành con số để có thể sử dụng một cách hiệu quả thì khá khó khăn. Phân tích dữ liệu thống kê giúp các nhà nghiên cứu và quản lý dự đoán thực tế phức tạp của kinh tế và xã hội trong tương lai ít rủi ro hơn. Những người ra quyết định và người quản lý thành công nhất chính là những người có thể hiểu thông tin và sử dụng thông tin hiệu quả.

2. BẢN CHẤT CỦA PHÂN TÍCH DỮ LIỆU

Dữ liệu chỉ là các số liệu thô và bản thân chúng không phải là tri thức. Trình tự đi từ dữ liệu đến tri thức là: từ dữ liệu đến thông tin, từ thông tin đến sự kiện, và cuối cùng là từ sự kiện đến tri thức. Dữ liệu trở thành thông tin khi nó liên quan đến vấn đề nhận thức, kết luận và quyết định của người nghiên cứu. Thông tin trở thành sự kiện khi thông tin hỗ trợ cho việc ra quyết định. Sự kiện trở thành tri thức khi nó được sử dụng để hoàn tất quá trình ra quyết định một cách thành công. Hình sau đây minh họa cho quá trình tư duy thống kê dựa trên dữ liệu trong việc xây dựng các mô hình thống kê để ra các quyết định trong hoàn cảnh không có đầy đủ thông tin như mong muốn.



(Nguồn: Hossein Arsham, Manchester Metropolitan University)

Đó là lý do tại sao chúng ta cần phân tích dữ liệu thống kê. Thống kê xuất phát từ nhu cầu đặt tri thức trên nền tảng chứng cứ có hệ thống. Điều này đòi hỏi phải nghiên cứu các quy luật xác suất, sự phát triển của các thang đo lường các đặc tính của dữ liệu và mối liên hệ ...

Chúng ta sẽ áp dụng các khái niệm và phương pháp thống kê căn bản đã học trong môn học thống kê vào các vấn đề cần giải quyết trong nghiên cứu thực tế. Quyển sách này được thiết kế để đáp ứng nhu cầu phân tích dữ liệu nghiên cứu thống kê trong kinh tế và xã hội, sử dụng một chương trình máy tính được sử dụng rộng rãi là SPSS. Bằng cách sử dụng chương trình máy tính, tất nhiên bạn sẽ thấy là mình phải tự hỏi các câu hỏi liên quan đến dữ liệu và phương pháp định sử dụng, và bạn phải có điều kiện sử dụng các phương pháp để giải quyết thỏa đáng các vấn đề nghiên cứu. Do đó các vấn đề ứng dụng được lấy trong lĩnh vực kinh tế và xã hội.

3. THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU

Để hiểu rõ phân tích dữ liệu thống kê là gì, trước hết phải hiểu thống kê là gì. Thống kê là tập hợp các phương pháp dùng để thu thập, phân tích, trình bày và diễn giải dữ liệu. Các phương pháp thống kê được sử dụng trong nhiều lĩnh vực khác nhau và giúp nhận

.., nghiên cứu, và giải quyết nhiều vấn đề phức tạp. Trong thực tế kinh tế xã hội, những phương pháp này giúp người ra quyết định và nhà quản lý ra các quyết định tốt hơn trong các hoàn cảnh không chắc chắn.

Những khối lượng thông tin thống kê khổng lồ đều sẵn có trong môi trường kinh tế và xã hội ngày nay nhờ những tiến bộ liên tục trong công nghệ thông tin. Để phát triển kinh tế và thay đổi xã hội, các nhà nghiên cứu và quản lý phải có khả năng hiểu được thông tin và sử dụng thông tin một cách hiệu quả. Phân tích dữ liệu cung cấp kinh nghiệm thực hành được đúc kết để giúp đẩy mạnh việc ứng dụng tư duy thống kê và kỹ thuật thống kê nhằm hiểu rõ các hiện tượng nghiên cứu làm cơ sở cho việc ra các quyết định phù hợp.

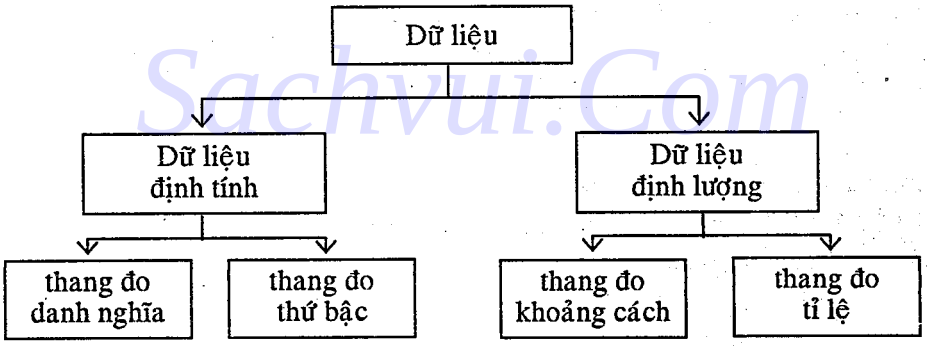
Máy tính đóng vai trò rất quan trọng trong phân tích dữ liệu nghiên cứu. Chương trình máy tính SPSS được sử dụng trong môn học này cung cấp khả năng điều khiển và kiểm soát dữ liệu và rất nhiều các thủ tục phân tích thống kê có thể giúp phân tích từ những tập hợp dữ liệu nhỏ cho đến những dữ liệu rất lớn. Máy tính đóng vai trò hỗ trợ việc tóm tắt dữ liệu, nhưng phân tích dữ liệu thống kê tập trung vào việc diễn giải kết quả để suy diễn và dự đoán.

CHƯƠNG I
PHÂN LOẠI DỮ LIỆU, MÃ HÓA, NHẬP LIỆU VÀ MỘT SỐ XỬ LÝ TRÊN BIẾN

Khi thực hiện nghiên cứu, chúng ta thu thập được khá nhiều dữ liệu từ nhiều nguồn theo nhiều cách thức khác nhau. Trong quyển sách này chúng ta sẽ chủ yếu bàn về cách thức xử lý dữ liệu sơ cấp thu thập qua quan sát, phỏng vấn hay thực nghiệm trong nghiên cứu. Trước khi đi vào nhập liệu và xử lý dữ liệu, người làm công tác xử lý cần hiểu rõ các loại dữ liệu, tính chất của từng loại dữ liệu.

1. PHÂN LOẠI DỮ LIỆU

Dữ liệu nghiên cứu có thể phân chia thành 2 loại chính là dữ liệu định tính và dữ liệu định lượng. Các dữ liệu này được thu thập bằng 4 thang đo cơ bản được thể hiện trên sơ đồ như sau:



Sự khác nhau cơ bản giữa dữ liệu định tính và dữ liệu định lượng là:

- Dữ liệu định tính: loại dữ liệu này phản ánh tính chất, sự hơn kém, ta không tính được trị trung bình của dữ liệu dạng định tính. Có nhiều cách thể hiện các dữ liệu định tính, ví dụ như giới tính của người được phỏng vấn là nam hay nữ; kết quả học tập của sinh viên là giỏi, khá, trung bình hay yếu.
- Dữ liệu định lượng: loại dữ liệu này phản ánh mức độ, mức độ hơn kém, tính được trị trung bình. Nó thể hiện bằng con số thu thập được

ngay trong quá trình điều tra khảo sát, các con số này có thể ở dạng biến thiên liên tục hay rời rạc. Nếu thu thập thông tin về nhiệt độ của từng giờ trong ngày ta có thể có một tập dữ liệu về nhiệt độ ở dạng biến liên tục. Ngược lại, tìm hiểu thông tin về số lượng trẻ em dưới 10 tuổi của từng hộ gia đình trong một khu vực thị trường, kết quả thu được phải là dữ liệu định lượng dạng biến rời rạc.

Cần chú ý rằng các phép toán thống kê dùng cho dữ liệu định tính có những đặc điểm khác với phép toán dùng cho dữ liệu định lượng.

2. CÁC LOẠI THANG ĐO

Thang đo là công cụ dùng để quy ước (mã hóa) các tình trạng hay mức độ của các đơn vị khảo sát theo các đặc trưng được xem xét. Ví dụ như tình trạng hôn nhân gia đình của người trưởng thành, mức độ đồng ý về một vấn đề nào đó. Ngày nay với việc sử dụng máy vi tính thì mã hoá thường được thực hiện bằng ký số thay vì ký tự. Có 4 loại thang đo cơ bản như sau mà theo thứ tự từ trên xuống chúng có khả năng biểu đạt thông tin tăng dần.

2.1. Thang đo danh nghĩa (còn gọi là thang đo định danh hoặc thang đo phân loại) - nominal scale: trong thang đo này các con số chỉ dùng để phân loại các đối tượng, chúng không mang ý nghĩa nào khác. Về thực chất thang đo danh nghĩa là sự phân loại và đặt tên cho các biểu hiện và ấn định cho chúng một ký số tương ứng. Ví dụ: Bạn có thể đặt câu hỏi như sau khi bạn tiến hành một cuộc điều tra:

"Vui lòng cho biết tình trạng hôn nhân của bạn hiện nay?"

Độc thân	☐ 1
Đang có gia đình	☐ 2
Ở goá	☐ 3
Ly thân hoặc ly dị	☐ 4

Thang đo danh nghĩa giúp bạn quy ước các cá nhân trả lời câu hỏi này thành các biểu hiện có thể có của biến "tình trạng hôn nhân". Bạn có thể quy ước đặt Độc thân = 1, Đang có gia đình = 2, Ở goá = 3, và Ly thân hoặc ly dị = 4. Những con số này mang tính định danh vì rõ ràng bạn không thể cộng chúng lại hoặc tính ra giá trị trung bình của "tình trạng hôn nhân". Những phép toán thống kê bạn có thể sử dụng được

cho dạng thang đo danh nghĩa là: đếm, tính tần suất của một biểu hiện nào đó, xác định giá trị mode, thực hiện một số phép kiểm định.

2.2. Thang đo thứ bậc - ordinal scale: lúc này các các con số ở thang đo danh nghĩa được sắp xếp theo một quy ước nào đó về thứ bậc hay sự hơn kém, nhưng ta không biết được khoảng cách giữa chúng. Điều này có nghĩa là bất cứ thang đo thứ bậc nào cũng là thang danh nghĩa nhưng rõ ràng bạn không thể suy ngược lại rằng thang danh nghĩa nào cũng là thang thứ bậc.

Ví dụ: "Bạn hài lòng như thế nào về mùi của sản phẩm Snack Khoai tây chiên mà bạn vừa dùng thử? Bạn thấy hài lòng, bình thường hay không hài lòng?"

Với câu hỏi này bạn có thể quy ước các cá nhân trả lời theo ba biểu hiện của biến "hài lòng". Nghĩa là bạn có thể đo lường người được phỏng vấn theo mức độ họ hài lòng với mùi của sản phẩm. Hãy đặt hài lòng = 3, bình thường = 2, không hài lòng = 1. Vậy là một người có câu trả lời được mã hoá bởi số 3 sẽ có mức độ hài lòng cao hơn người mang số 2 hoặc số 1. Tuy nhiên chúng ta không biết được là người đó hài lòng gấp 2 lần hay gấp 5 lần hoặc chỉ là hài lòng hơn những người ở mức 1 hoặc 2 một chút mà thôi.

Đối với thang đo thứ bậc, khuynh hướng trung tâm có thể xem xét bằng số trung vị và số mốt (mode), còn độ phân tán chỉ đo được bằng khoảng và khoảng tứ trung vị (interquartile range).

2.3. Thang đo khoảng - interval scale: là một dạng đặc biệt của thang đo thứ bậc vì nó cho biết được khoảng cách giữa các thứ bậc. Thông thường thang đo khoảng có dạng là một dãy các chữ số liên tục và đều đặn từ 1 đến 5, từ 1 đến 7 hay từ 1 đến 10. Dãy số này có 2 cực ở hai đầu thể hiện 2 trạng thái đối nghịch nhau. Ví dụ như 1 là rất ghét, 7 là rất thích; 1 là không đồng ý, 7 là rất đồng ý; 1 là rất không hài lòng, 7 là rất hài lòng ...

Ví dụ: Theo anh/chị/ông/bà, tầm quan trọng của các yếu tố sau đây như thế nào đối với cuộc sống của một người?(1=không quan trọng; 7= rất quan trọng)

	Không quan trọng						Rất quan trọng
1. có nhiều tiền	1	2	3	4	5	6	7
2. đạt trình độ học vấn cao	1	2	3	4	5	6	7
3. có địa vị trong xã hội	1	2	3	4	5	6	7
4. có bạn bè tốt	1	2	3	4	5	6	7
5. gia đình ổn định	1	2	3	4	5	6	7
6. có tự do cá nhân	1	2	3	4	5	6	7
7. có sức khỏe tốt	1	2	3	4	5	6	7
8. có nghề nghiệp thích hợp	1	2	3	4	5	6	7
9. có tình yêu	1	2	3	4	5	6	7
10. được mọi người tôn trọng	1	2	3	4	5	6	7
11. sống có ích cho người khác	1	2	3	4	5	6	7
12. được hưởng thụ nhiều thú vui trong cuộc sống	1	2	3	4	5	6	7

Rõ ràng trong việc đo lường thái độ hay ý kiến thì thang đo khoảng cung cấp nhiều thông tin hơn so với thang đo thứ bậc.

Những phép toán thống kê có thể sử dụng thêm cho loại thang đo này so với 2 loại thang đo trước là: tính khoảng biến thiên, số trung bình, độ lệch chuẩn ... Cần chú ý là thang đo khoảng tự nó không có điểm 0 tuyệt đối, do đó bạn chỉ có thể thực hiện được phép tính cộng trừ chứ nếu dùng phép chia thì kết quả không có ý nghĩa (bạn sẽ hiểu rõ điều này hơn ở phần kế tiếp đây).

2.4. Thang đo tỉ lệ - ratio scale: thang đo tỉ lệ có tất cả các đặc tính khoảng cách và thứ tự của thang đo khoảng, ngoài ra điểm 0 trong thang đo khoảng là một trị số "thật" nên ta có thể thực hiện được phép toán chia để tính tỉ lệ nhằm mục đích so sánh.

Ví dụ: " Bạn bao nhiêu tuổi?"

Bạn có thể sử dụng câu hỏi rất phổ biến này làm một ví dụ tiêu biểu cho dạng thang đo tỉ lệ. Các cá nhân được hỏi có các cấp độ khác nhau của biến "tuổi tác", các cấp độ cách đều nhau 1 năm. Các con số thu được từ câu hỏi nêu trên có đặc tính là tính tỉ lệ được (chẳng hạn một người 50 tuổi thì lớn tuổi gấp đôi người 25 tuổi và chỉ bằng 2/3 người 75 tuổi).

Để phân biệt với trường hợp thang đo khoảng, hãy quan sát một cái nhiệt kế bạn sẽ thấy thang đo nhiệt độ dùng đơn vị 1^0C có khoảng cách giống nhau tại bất kỳ điểm nào trên thang, do đó khoảng cách

giữa 33^0C và 34^0C cũng bằng với khoảng cách giữa 0^0C và 1^0C. Tuy nhiên ta chỉ có thể thực hiện được phép cộng và trừ trên thang đo khoảng chứ không làm được phép chia vì thang đo khoảng tự nó không có điểm 0 có ý nghĩa (bạn đừng nhầm lẫn với việc nó có mốc 0^0C). Bạn không thể nói khi thời tiết có nhiệt độ trung bình là 20^0C thì ấm áp gấp 2 lần thời tiết có nhiệt độ trung bình 10^0C. Bạn sẽ thấy điều này thuyết phục hơn nhiều nếu quy đổi 20^0C và 10^0C sang hệ nhiệt độ đo bằng ^{0}F (độ Fahrenheit).

Nói chung với các biến thu thập bằng thang đo khoảng và thang đo tỉ lệ ta có thể đo lường xu hướng trung tâm bằng bảng tần số, biểu đồ tần số, trung bình số học. Còn xu hướng phân tán đo bằng độ lệch chuẩn, phương sai, (khoảng và tứ trung vị ít được sử dụng đến do kém hữu ích hơn). Vì vậy chương trình SPSS gộp chung hai loại thang đo này thành một gọi là Scale Measures (thang đo mức độ).

3. NGUYÊN TẮC MÃ HOÁ VÀ NHẬP LIỆU

Nguyên tắc mã hóa và nhập liệu được mô tả tóm tắt trong 2 bảng 1.1 và 1.2. Một ma trận dữ liệu lưu giữ thông tin về từng người trả lời tại từng biến trong từng "ô" chứa dữ liệu tương ứng (như minh hoạ trong Bảng 1.1), đây là những thông tin về độ tuổi, giới tính, nghề nghiệp của 3 người trả lời đầu tiên trong một tập dữ liệu gồm *n* người, bạn sẽ thấy người trả lời đầu tiên có giới tính nữ, 21 tuổi, là sinh viên.

Bảng 1.1

Các biến

Những người trả lời

	Giới tính	Tuổi	Nghề nghiệp
1	Nữ	21	Sinh viên
2	Nữ	32	Nhân viên văn phòng
3	Nam	53	Về hưu
...	...	...	...
n	Nam	42	Nghề khác

Với mục tiêu nhập dữ liệu thì các dữ liệu ở dạng không phải kí số như "nữ " hay "sinh viên" cần phải được tạo một con số mã hoá, như

minh họa trong Bảng 1.2. Ta quy ước 1 là nam và 2 là nữ (tất nhiên bạn hoàn toàn có thể quy ước ngược lại 1 là nữ và 2 là nam). Những quy ước cho nghề nghiệp bạn có thể dò thấy trên cột 4 ở Bảng 1.2. Có thể suy ra là ma trận dữ liệu này quản lý khoảng 13 loại công việc khác nhau có tên cụ thể, những nghề nghiệp ngoài phạm vi này được gọi chung là "nghề khác".

Bảng 1.2

Các biến

Những người trả lời

	Giới tính	Tuổi	Nghề nghiệp
1	2	21	10
2	2	32	3
3	1	53	11
...	...	...	...
n	1	42	14

Chú ý là ta chỉ mã hóa thông tin thu thập bằng thang đo định tính, còn các thông tin thu thập từ thang đo định lượng đã ở dưới dạng số và có ý nghĩa nên không cần mã hóa.

Bạn nhận diện các cột của ma trận dữ liệu qua các biến đang được đo lường và mã hoá trong cuộc khảo sát, mỗi cột cho mỗi biến. Đi xuống theo cột thì mỗi hàng sẽ là các thông tin của từng người trả lời khác nhau về biến đó. Trong phần lớn trường hợp mỗi biến tương ứng với một câu hỏi cụ thể trong bản câu hỏi. Với câu hỏi chỉ chọn một trả lời, bạn chỉ cần tạo một biến vì từ câu hỏi này, đối với một đối tượng trả lời, về một vấn đề nào đó chẳng hạn như "tuổi" thì chúng ta chỉ có một thông tin (một con số) nhập vào nên chỉ cần 1 ô thuộc 1 cột (hay là 1 biến) là đủ.

Tuy nhiên trong trường hợp câu hỏi có thể chọn nhiều trả lời, cần phải có nhiều biến vì trong trường hợp này, 1 người trả lời sẽ có thể chọn nhiều trả lời phù hợp với tình huống thực tế của họ. Khi họ cung cấp nhiều thông tin như vậy thì chúng ta sẽ có nhiều con số phải nhập vào (các con số này dĩ nhiên cũng do bạn quy ước theo trật tự bạn đặt ra). Do vậy về một vấn đề được hỏi cần phải có nhiều

ô trên nhiều cột (hay nhiều biến) để nhập liệu. Bạn hãy xem một ví dụ về câu hỏi có nhiều trả lời (Multiple Answer MA) như sau:

"Trong vòng 5 năm vừa qua, bạn đã đi nghỉ ở những nơi nào trong các bãi biển sau đây?" (Có thể chọn nhiều trả lời phù hợp với bạn)

Bãi Cháy	☐ 1
Đồ Sơn	☐ 2
Sầm Sơn	☐ 3
Nha Trang	☐ 4
Mũi Né	☐ 5
Vũng Tàu	☐ 6
Khác	☐ 7

Trật tự mà các biến được sắp xếp trong ma trận dữ liệu sẽ đi theo thứ tự chúng được hỏi trong bản câu hỏi. Thật ra với mục đích phân tích thì trật tự này không quan trọng lắm, nhưng trật tự này sẽ tạo cho chúng ta một định hướng trong quá trình nhập dữ liệu, bạn sẽ thấy điều này hữu ích khi cần quay lại bản câu hỏi để tìm hiểu điều gì đó về các biến.

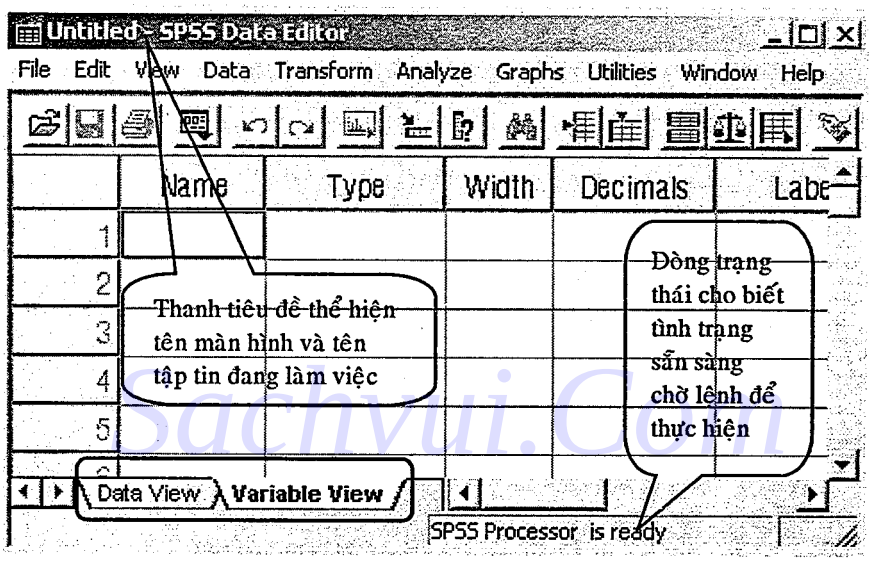
Các hàng trong bản câu hỏi tương ứng với từng người trả lời, mỗi hàng là một người trả lời, điều này có nghĩa là tất cả các thông tin trong một bản câu hỏi được khai thác từ một người trả lời sẽ nằm toàn bộ trên một hàng theo thứ tự của người trả lời đó. Lại nói về trật tự của các hàng, nó cũng không phải là một vấn đề quan trọng đối với mục tiêu phân tích dữ liệu. Các bản câu hỏi thường được nhập theo trật tự chúng được thu thập qua các cuộc phỏng vấn, chúng sẽ được đánh số tương ứng với các con số của hàng. Điều này có nghĩa là nếu bạn có bất cứ thắc mắc gì về một bản câu hỏi cụ thể nào đó, bạn sẽ tìm ra nó ngay.

Với cấu trúc của cột và hàng như vậy thì khi nhập liệu trên trên cửa sổ data của SPSS bạn sẽ nhập từ trái qua phải (theo từng dòng). Xong 1 bản câu hỏi (một dòng) thì chuyển sang bản câu hỏi khác (tức là sang dòng mới).

4. CỬA SỔ LÀM VIỆC CỦA SPSS

Khởi động SPSS for Windows bằng cách nhấp chuột vào biểu tượng của chương trình SPSS for Windows trên Desktop hoặc từ Start menu chọn Programs rồi chọn SPSS for Windows, cửa sổ làm việc đầu tiên là cửa sổ dữ liệu (Data View) sẽ hiện ra.

Hình 1.1



Cửa sổ dữ liệu của SPSS for Windows có 2 loại kiểu nhìn (view)

- Data View: kiểu nhìn dùng để nhập liệu và xem dữ liệu đã nhập
- Variable View: kiểu nhìn dùng để khai báo biến

Để thay đổi giữa hai kiểu nhìn này, ta nhấp chuột chọn tên cửa sổ Data View hay Variable View ở góc dưới bên tay trái của màn hình (phía trên dòng trạng thái), hoặc bạn có thể bấm tổ hợp phím Control+T để chuyển đổi qua lại giữa hai kiểu nhìn của cửa sổ dữ liệu một cách nhanh chóng. SPSS còn cửa sổ thứ hai là cửa sổ kết quả xử lý có tên là Output, sẽ hiện ra khi bạn chạy lệnh xử lý. Cửa sổ này biểu hiện các kết quả do bạn thực hiện như bảng biểu, đồ thị...

Nội dung chủ yếu của các Menu

File: giúp ta khởi tạo file mới, mở các file sẵn có, lưu file, in ấn, thoát ...

Edit: gồm các lựa chọn undo, cắt/dán, tìm kiếm/thay thế, xác lập các mặc định.

View: cho hiện dòng trạng thái, thanh công cụ, chọn font chữ, cho hiện giá trị nhập vào (value) hay nhãn ý nghĩa của các giá trị nhập ...

Data: bao gồm các lựa chọn phục vụ công tác về dữ liệu như chèn thêm biến, tìm nhanh một quan sát trong một tập hợp dữ liệu có nhiều quan sát, xếp thứ tự các quan sát, ghép file, chia file, chọn quan sát...

Transform: gồm các lệnh giúp chuyển đổi dữ liệu, tính toán, mã hóa lại các biến ...

Analyze: thực hiện các thủ tục thống kê như: tóm tắt dữ liệu, lập bảng tổng hợp, so sánh trung bình, phân tích phương sai, tương quan và hồi qui, phân tích phi tham số, phân tích đa biến, ...

Graphs: tạo các biểu đồ và đồ thị

Utilities: tìm hiểu thông tin về các biến, file, ...

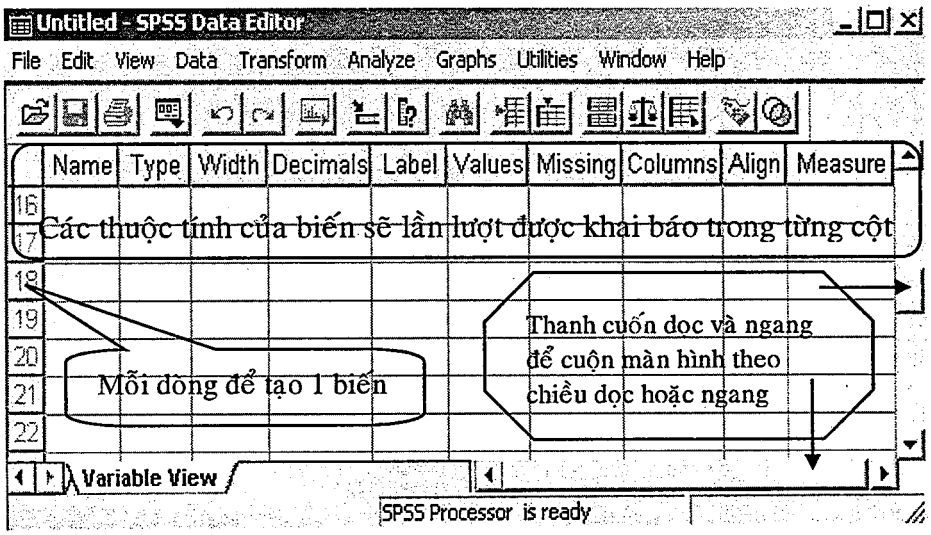
Windows: sắp xếp các cửa sổ làm việc trong SPSS, di chuyển giữa các cửa sổ làm việc ...

5. TẠO TẬP TIN DỮ LIỆU TRONG SPSS FOR WINDOWS

5.1. Khai báo biến

Sau khi khởi động cửa sổ dữ liệu của SPSS, bạn nhấp chuột vào Variable view để chuyển sang màn hình khai báo biến. Màn hình khai báo biến hiện ra như sau:

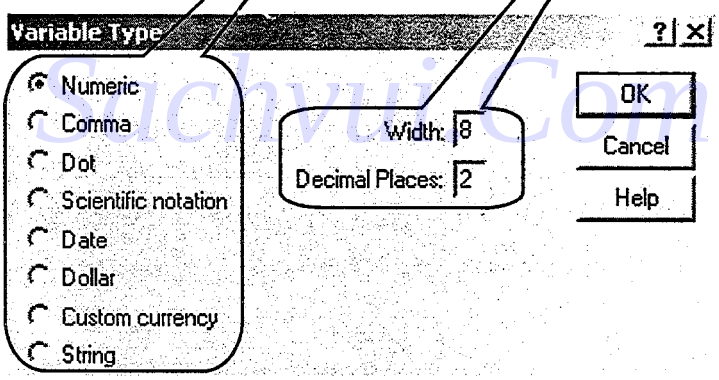
Hình 1.2



Trong màn hình này, mỗi biến bây giờ trên 1 dòng, các cột trong dòng thể hiện các thuộc tính của biến. Đối với từng biến, bạn lần lượt khai báo các thuộc tính như sau:

- **Name**: (tên biến) gõ trực tiếp tên biến vào ô này, tên biến cần đặt có độ dài không quá 8 kí tự hay kí số, không có kí tự đặc biệt và không được bắt đầu bằng một kí số, thông thường ta hay đặt tên biến (Variable name) gần với câu hỏi mà biến đó mô tả, ví dụ với câu hỏi 3 thì ta đặt tên biến đại diện là c3. Sau khi tạo xong tên biến thì chuyển qua ô kế bên phải để khai báo kiểu biến.
- **Type**: (kiểu biến) mặc định chương trình sẽ chọn kiểu định lượng (Numeric). Muốn thay đổi kiểu biến hay thay đổi độ rộng của biến hoặc số lượng số thập phân của biến định lượng, hãy nhấn chuột vào nút ... trong ô Type để mở hộp thoại Type.

Hình 1.3



Sau khi khai báo kiểu biến phù hợp, nhấp nút OK trở về màn hình Variable View và di chuyển sang 2 ô kế tiếp là:

- **Width**: độ rộng của biến là số ký số hay ký tự tối đa có thể nhập
- **Decimals**: số lẻ sau dấu phẩy

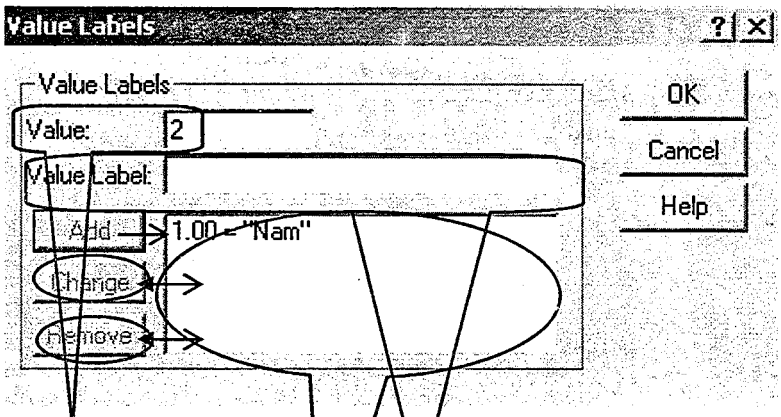
Nếu bạn đã lựa chọn những định dạng về Width và Decimals trong hộp Variable Type thì không cần khai báo mục này trên cửa sổ Variable View nữa. Sau đó nếu muốn thay đổi những định dạng này thì bạn có thể hiệu chỉnh trực tiếp trên cửa sổ Variable View chứ không cần trở lại hộp Variable Type.

- **Label**: đặt nhãn cho biến, nhãn này phải ngắn gọn nhưng có tính giải thích cao, chẳng hạn với câu hỏi về các bãi tắm ở trên ta có thể đặt label là "các bãi biển đã đến trong 5 năm qua" bằng cách

gõ trực tiếp tên nhãn biến vào ô Label. Sau khi đặt nhãn biến bạn sẽ sang ô Value để mã hóa biến tức là gán cho giới tính nữ giá trị 2 và nam giá trị 1 mà ta đã nói ở phần trước.

- **Values**: là thuộc tính quan trọng nhất, lúc này nhấp nút chuột vào nút ... nằm ở phía phải của ô thuộc vị trí cột Value tại dòng của biến ta đang khai báo, hộp thoại khai báo nhãn biến Value Labels sẽ xuất hiện. Trong hộp thoại này ta khai báo những nội dung:

Hình 1.4



* Value: mã hoá các thang đo định tính

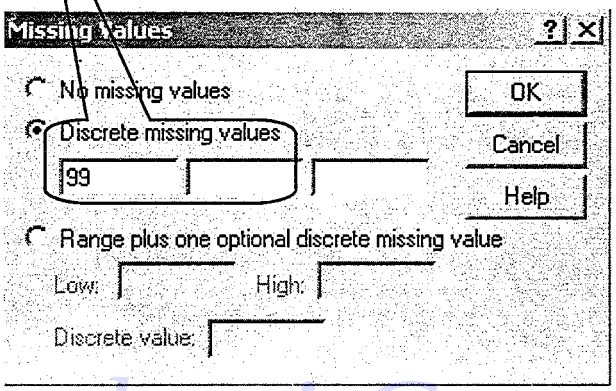
* Value label: nhãn giải thích ý nghĩa của mã số đã nhập

Sau khi khai báo xong thông tin, nhấp nút Add để đưa mã hoá và nhãn giải thích vào danh sách, nếu muốn chỉnh sửa những con số mã hoá và nhãn đã nhập thì chọn sáng đối tượng muốn chỉnh sửa đó và bấm nút Change trả nó lại khung Value và Value Label để thực hiện hiệu chỉnh. Muốn xoá luôn đối tượng khỏi danh sách thì chọn sáng rồi bấm Remove.

- **Missing**: khai báo các loại giá trị khuyết, cách vào hộp thoại này giống như với Values, giả dụ ta gặp tình huống với câu hỏi về trình độ học vấn có những người được điều tra vì lý do tế nhị nào đó đã từ chối trả lời thì trong Value lable ta quy ước giá trị 99 có nhãn là "không trả lời", sau đó sang Missing ta phải khai báo 99 là giá trị khuyết để sau đó khi tính toán các lệnh thống kê ví như tính tần số chẳng hạn máy sẽ loại giá trị khuyết này ra khi tính phần trăm hợp lệ.

Cách đặt con số đại diện cho Missing value là tuỳ tình hình và sự lựa chọn của người xử lý. Ví dụ nếu đặt Missing value cho biến độ tuổi mà ta chọn số 99 sẽ gây nhầm lẫn nếu cuộc điều tra có thể có những người đạt 99 tuổi hoặc hơn nữa, với tình huống này ta nên đặt là –10 hay 999.

Hình 1.5



Trong file *Data thực hành* trong tập hợp dữ liệu dùng kèm với sách, các biến *c29a1* đến *c29c* có khai báo Missing value là 8 hoặc 9 là số người không trả lời (không có ý kiến)

Ngoài ra còn có một loại giá trị khuyết nữa là System Missing, đó là giá trị khuyết của hệ thống, nó được chương trình tự động đặt dấu chấm (·) ở những vị trí không được nhập giá trị. Giá trị System Missing này "vô hình" đối với các lệnh xử lý thống kê của phần mềm SPSS.

- **Columns**: khai báo độ rộng của cột biến khi ta nhập liệu, thường chọn là 8
- **Align**: vi trí dữ liệu được nhập trong cột, thường chọn là Right
- **Measure**: chọn loại thang đo thể hiện dữ liệu với 3 loại chính là Ordinary (thang thứ bậc), Norminal (thang danh nghĩa) và Scale (gồm cả Interval và Ratio tức thang đo khoảng cách và tỉ lệ)

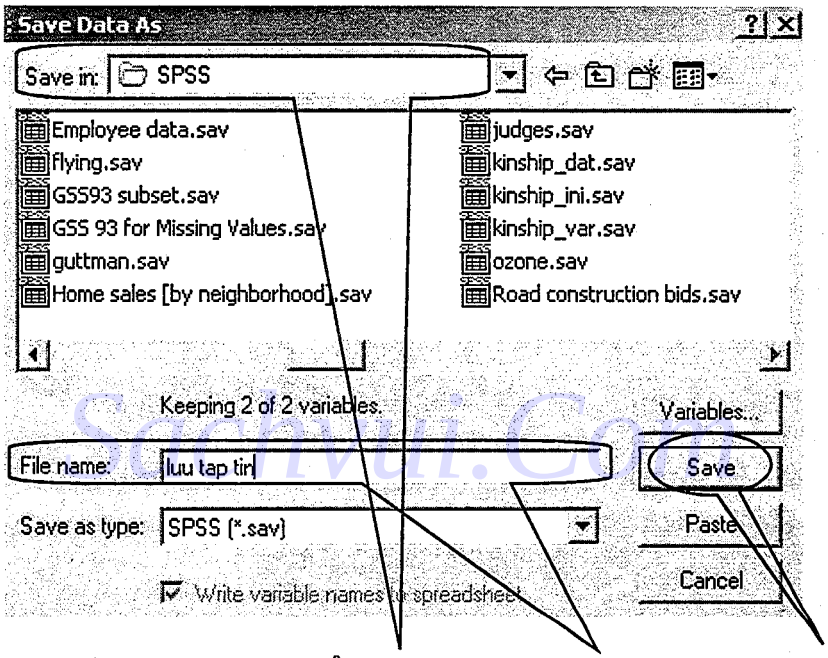
Như vậy là biến bạn cần đã được tạo xong. Nhấn phím Home trở về đầu dòng, xuống dòng dưới để tạo ra biến kế tiếp theo cách tương tự. Đặc biệt, kể từ SPSS phiên bản 10.0 trở về sau, chúng ta có thể copy bất kỳ thuộc tính nào của biến này qua biến khác.

Để copy, chúng ta dùng trỏ chuột để nhấp chọn vùng thuộc tính, ví dụ như Type (kiểu biến) hay Value (mã hóa). Bấm tổ hợp phím Control+C. Rồi dùng trỏ chuột để chọn vùng muốn copy thuộc tính tương ứng vào, và bấm tổ hợp phím Control+V.

5.2. Lưu tập tin dữ liệu

Để lưu lại tập tin dữ liệu, trong màn hình data, từ menu chọn File>Save, hộp thoại sau sẽ xuất hiện

Hình 1.6



Trong hộp thoại này, chọn ổ đĩa, thư mục, đặt tên tập tin và nhấn Save, cách thức cũng như lưu tập tin trên Word hay Excel (phần mở rộng của tên tập tin dữ liệu SPSS mặc định sẽ là sav, chúng ta không cần ghi phần đuôi sav này khi khai báo tên tập tin để lưu)

6. MỘT SỐ XỬ LÝ TRÊN BIẾN

6.1. Mã hoá lại biến (Recode)

Trong thực tế có nhiều tình huống chúng ta cần mã hoá lại biến mà chúng ta đang sử dụng, đó là khi:

- Chúng ta muốn giảm số biểu hiện của một biến định tính xuống chỉ còn 2 hay 3 loại biểu hiện cơ bản. Bạn hãy phân biệt việc mã hoá lại biến với việc mã hoá biến định danh trong quá trình nhập liệu. Giả sử một người nghiên cứu đang xem xét về tiểu sử cá nhân của 200 người di cư lao động đến thành phố Hồ Chí Minh. Với câu hỏi về tên xã, phường, hay thị trấn mà họ đã làm việc, người nghiên cứu có thể nhận được một danh sách với hàng trăm địa danh. Thông thường anh ta sẽ rút gọn danh sách địa danh bằng cách giản lược chúng đi qua quá trình mã hoá ngay trong lúc nhập liệu. Anh ta có thể mã hóa các địa danh theo khu vực địa lý như khu vực Tây Nguyên, Duyên hải Trung Bộ, Đông Nam Bộ, Tây Nam Bộ, ... Mã hoá lại biến thì lại không như vậy, ví dụ với biến *thu nhập gia đình* đã được tạo trong file ví dụ *Data thuc hanh* bạn có thể mã hoá lại biến *tngd* bằng cách gom chung biểu hiện thứ 4 là thu nhập từ 6-10 triệu và biểu hiện thứ 5 là thu nhập trên 10 triệu thành một nhóm chung là thu nhập gia đình trên 6 triệu.
- Chúng ta muốn chuyển một biến định lượng thành một biến định tính. Trường hợp khi các biến định lượng có quá nhiều giá trị thì bảng tần số được lập từ các dữ liệu này sẽ rất dài nên ít có ý nghĩa trong việc tóm tắt và trình bày, thử tưởng tượng bạn có 500 người với 42 độ tuổi từ 18 đến 60, mỗi độ tuổi có một số lượng người khác nhau, nếu liệt kê ra thì bảng tần số của biến *tuoi* sẽ dài đến 42 hàng (xem minh hoạ ở Bảng1.4). Do đó các biến này cần được mã hoá lại để chỉ còn một ít nhóm giá trị giúp việc trình bày ngắn gọn và rõ ràng hơn.

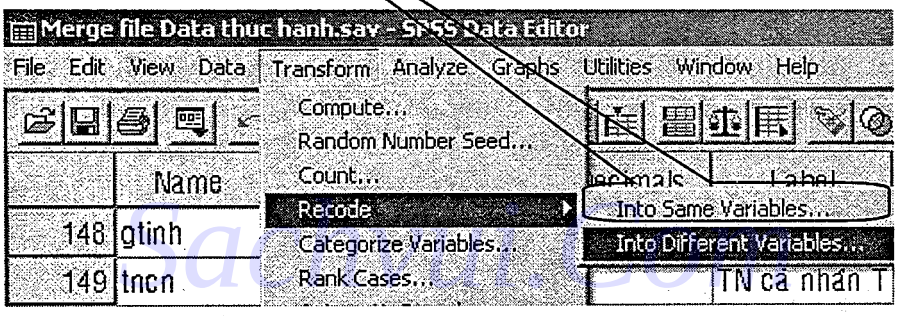
Với tình huống như trên bạn sẽ giải quyết bằng cách phân tổ những người trả lời theo tuổi, với 42 độ tuổi bạn có thể chia thành 4 tổ là (18-25); (26-35); (36-45); (46-60). Tất nhiên bạn cũng có thể chia thành các tổ (18-30); (31-40); (41-50); (51-60). Lựa chọn cách phân tổ nào là tuỳ đánh giá mục tiêu nghiên cứu của bạn, đặc trưng của mẫu nghiên cứu ... miễn sao việc đọc dữ liệu dễ dàng và số người được gom vào các tổ không quá ít, nếu bạn để mỗi tổ ít người quá thì sẽ dẫn đến một khó khăn cho thủ tục kiểm định Chi – bình phương hay những thủ tục tương tự khác mà bạn sẽ gặp sau này.

Lúc này rõ ràng biến tuổi dạng định lượng ban đầu đã được mã hoá thành một biến mới là dạng định tính với 4 biểu hiện. Xem tiếp các chương sau bạn sẽ nhận thấy rõ hơn vai trò của việc mã hoá lại biến.

Quy trình thực hiện việc mã hoá lại biến

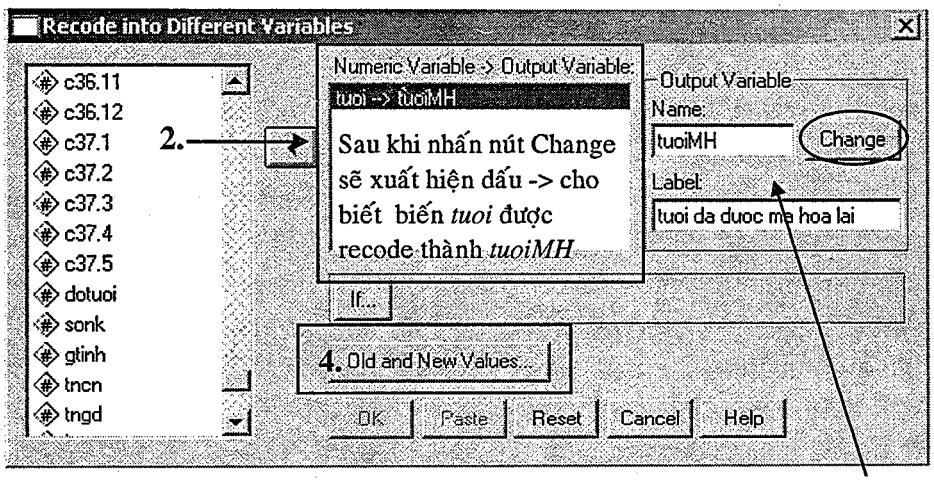
1. Vào menu Transform>Recode>Into Different Variables mở hộp thoại Recode Into Different Variables (Hình 1.8) để lệnh Recode tạo cho bạn một biến mới với các giá trị mã hoá do bạn khai báo trên cơ sở biến gốc, còn biến cũ làm cơ sở mã hoá vẫn được giữ lại. Nhớ là đừng chọn Into Same Variables trừ khi bạn muốn lệnh Recode làm mất đi biến cũ của bạn và tạo ra một biến mới với các biểu hiện vừa được mã hoá trên cơ sở biến cũ.

Hình 1.7



2. Trong hộp thoại Recode Into Different Variables bạn chọn biến muốn recode (ở đây là biến *tuôi*) đưa sang khung Numeric Variable-> Output Variable bằng cách: nhấp chuột tại tên biến muốn recode trong danh sách biến nguồn bên trái và biến đó sẽ được chiếu sáng, sau đó rê con trỏ chuột đến nút mũi tên hướng vào khung Numeric Variable-> Output Variable, nhấp chuột và tên biến này sẽ xuất hiện trong khung Numeric Variable-> Output Variable

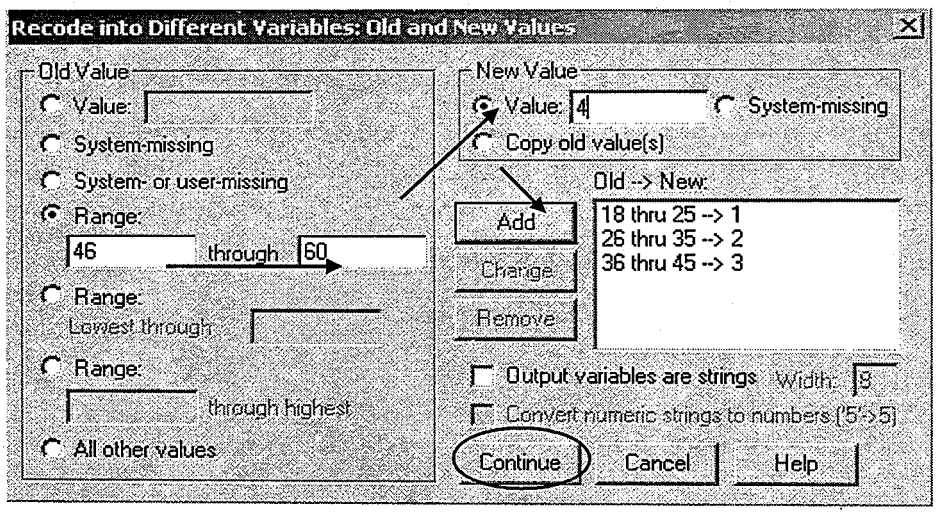
Hình1.8



3. Sau đó sang phần Output Variable đặt tên và nhãn cho biến mới này, ví dụ đặt là *tuoiMH*, đặt nhãn Lable là "tuoi duoc ma hoa lai" sau đó nhấn nút Change để báo cho SPSS biết bạn muốn recode biến *tuôi->tuoiMH,* nhớ đừng quên nút Change nếu không lệnh recode của bạn sẽ không thành công.

4. Nhấp vào nút Old and new value mở tiếp hộp thoại Recode into Diferent Variables: Old and New Values để xác định sự chuyển đổi giữa giá trị cũ và giá trị mới tương ứng.

Hình 1.9



Trong hộp thoại này, lần lượt khai báo phần giá trị cũ (Old Value bên tay trái), tương ứng với từng giá trị mới (New Value bên tay phải), có các loại giá trị cũ có thể được recode như sau:

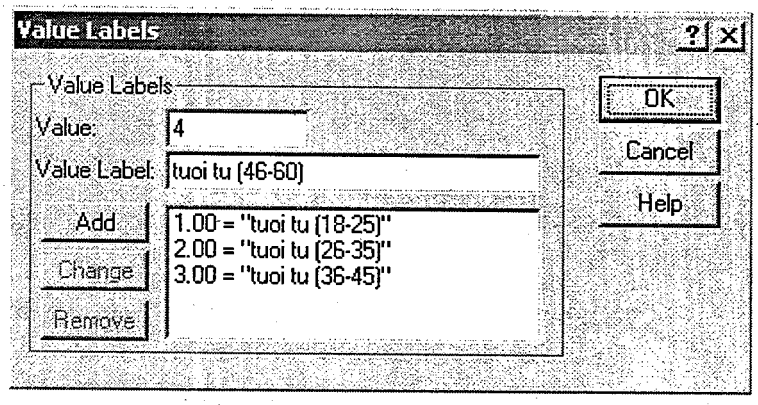
- Value: từng giá trị cũ rời rạc ứng với 1 giá trị mới
- System-missing: giá trị khuyết của hệ thống
- System or user mising: giá trị khuyết của hệ thống hoặc do người sử dụng định nghĩa.
- Range: một khoảng giá trị cũ ứng với một giá trị mới, tình huống này cũng có ba trường hợp nhỏ là khoảng giữa hai giá trị (Range ...through); khoảng từ giá trị nhỏ nhất đến một giá trị xác định được nhập vào (Lowesr through ...Range); khoảng từ một giá trị xác định được nhập vào đến giá trị lớn nhất (Range... through Highest)

Mỗi lần bạn xác định xong một cặp giá trị cũ và chỉ định giá trị mới, nút Add sẽ hiện sáng lên, hãy nhấn vào nút này để đưa cặp giá trị cũ được khai báo và giá trị mới này vào ô Old -> New: (nhớ đừng quên nhấn nút Add sau mỗi lần xác định xong một cặp giá trị cũ – mới)

5. Xác định xong bạn nhấp nút Continue để trở về hộp thoại trước đó và chọn OK để thực hiện lệnh mã hoá lại, lúc đó trên màn hình Variable view xuất hiện một biến mới là *tuoiMH* nằm dưới cùng tức là biến được tạo mới nhất.

6. Trên màn hình Variable View, bạn phải vào thuộc tính Values để gán các nhãn giá trị cho biến vừa tạo, nếu không khai báo các nhãn giá trị thì khi bạn lập bảng tần số cho *tuoiMH,* SPSS sẽ truy xuất ra tần số của các con số 1, 2, 3, 4 mà bạn đã gán chứ không truy xuất các biểu hiện (18-25); (26-35), ... của biến *tuoiMH*. Do đó bạn phải nhớ khai báo Values cho *tuoiMH.*

Hình 1.10



Cuối cùng, bạn hãy so sánh bảng tần số của của biến *tuôiMH* (Bảng 1.3) với bảng tần số của biến *tuôi* ban đầu có tới 42 độ tuổi khác nhau (Bảng 1.4) để thấy sự khác biệt mà lệnh Recode tạo ra.

Bảng 1.3 Bảng tần số của biến *tuôi* đã được mã hoá

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	tuoi tu (18-25)	150	30.0	30.0	30.0
	tuoi tu (26-35)	140	28.0	28.0	58.0
	tuoi tu (36-45)	111	22.2	22.2	80.2
	tuoi tu (46-60)	99	19.8	19.8	100.0
	Total	500	100.0	100.0	

Sau khi được Recode, tuổi của 500 người trong mẫu của ta được gom thành 4 nhóm tuổi chính. Cột Frequency ở bảng trên cho biết số người thuộc về mỗi nhóm tuổi, tổng của cột này luôn phải bằng tổng số người của mẫu nghiên cứu.

Chúng ta sẽ thảo luận chi tiết về cách làm thế nào yêu cầu SPSS tạo ra Bảng 1.3 như trên ở Chương III. Ở đây bạn chỉ cần so sánh và cảm nhận về sự tiện lợi mà thủ tục Recode đã đem lại cho bạn.

Bảng 1.4 Bảng tần số của biến *tuoi* với 42 độ tuổi khác nhau

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	9	1.8	1.8	1.8
	19	8	1.6	1.6	3.4
	20	16	3.2	3.2	6.6
	21	19	3.8	3.8	10.4
	22	22	4.4	4.4	14.8
	23	26	5.2	5.2	20.0
	24	29	5.8	5.8	25.8
	25	21	4.2	4.2	30.0
	26	15	3.0	3.0	33.0
	27	13	2.6	2.6	35.6
	28	19	3.8	3.8	39.4
	29	16	3.2	3.2	42.6
	30	17	3.4	3.4	46.0
	31	10	2.0	2.0	48.0
	32	18	3.6	3.6	51.6
	33	8	1.6	1.6	53.2
	34	9	1.8	1.8	55.0
	35	15	3.0	3.0	58.0
	36	8	1.6	1.6	59.6
	37	6	1.2	1.2	60.8
	38	16	3.2	3.2	64.0
	39	14	2.8	2.8	66.8
	40	18	3.6	3.6	70.4
	41	10	2.0	2.0	72.4
	42	16	3.2	3.2	75.6
	43	7	1.4	1.4	77.0
	44	6	1.2	1.2	78.2
	45	10	2.0	2.0	80.2
	46	12	2.4	2.4	82.6
	47	7	1.4	1.4	84.0
	48	8	1.6	1.6	85.6
	49	8	1.6	1.6	87.2
	50	12	2.4	2.4	89.6
	51	3	.6	.6	90.2
	52	7	1.4	1.4	91.6
	53	4	.8	.8	92.4
	54	10	2.0	2.0	94.4
	55	9	1.8	1.8	96.2
	56	4	.8	.8	97.0
	57	2	.4	.4	97.4
	59	11	2.2	2.2	99.6
	60	2	.4	.4	100.0
	Total	500	100.0	100.0	

6.2. Chuyển một biến dạng Category thành dạng Dichotomy

Biến dạng category là biến phân loại có thể có nhiều trị số mã hóa tượng trưng cho nhiều trạng thái, biểu hiện khác nhau (ví dụ như: đạo Phật, đạo Thiên chúa đạo Tin Lành, đạo Cao Đài...). Biến Dichotomy là biến phân loại chỉ có 2 trị số mã hóa tượng trưng cho 2 trạng thái hay 2 biểu hiện khác nhau (ví dụ như: có tôn giao hay không có tôn giáo; nam hay nữ; đồng ý hay không đồng ý; có đọc báo Tuổi Trẻ hay không đọc báo Tuổi trẻ). Đối với một câu hỏi khảo sát dùng thang đo định danh, trong đó người trả lời có thể chọn nhiều hơn 1 trả lời, có thể mã hóa và nhập liệu theo cả 2 kiểu biến này. Mã hóa và nhập liệu theo kiểu category dễ thực hiện hơn. Tuy nhiên khi phân tích sâu thì kiểu biến dichotomy có nhiều lợi thế hơn. Do đó người ta thường tạo khuôn và nhập liệu theo kiểu category, sau đó khi cần phân tích sâu thì chuyển đổi sang dạng dữ liệu dùng kiểu biến dichotomy.

Có một vài cách để chuyển hoá một biến dạng Category thành dạng Dichotomy, ở đây chúng ta sẽ làm quen với cách sử dụng lệnh Count cuả SPSS. Trường hợp này được chọn vì lệnh Count có thể giúp bạn dễ dàng chuyển biến Category dạng đơn và biến Category dạng câu hỏi có nhiều trả lời thành 1 biến Dichotomy duy nhất.

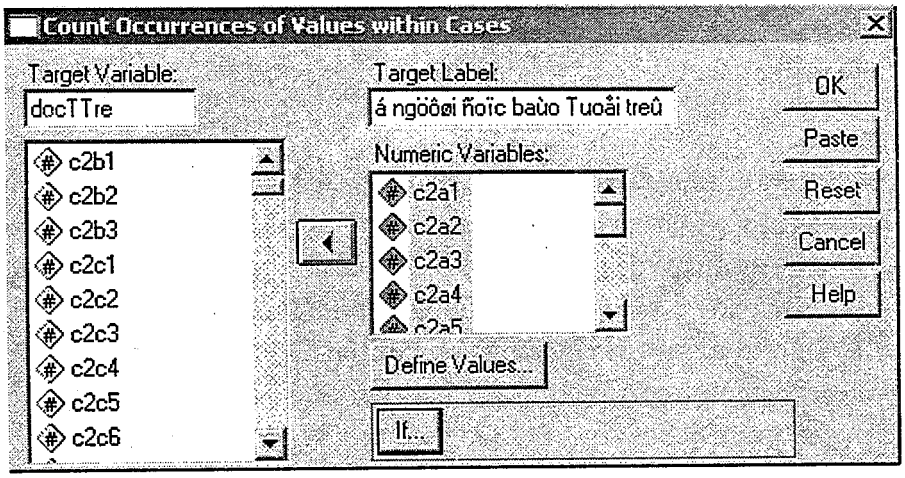
Giả sử bạn gặp câu hỏi nhiều trả lời (MA) trong câu hỏi về các loại báo một người thường đọc, người được hỏi có thể nhắc đến nhiều loại báo khác nhau chứ không chỉ 1 loại báo, thông tin về các loại báo thường đọc sẽ được thể hiện trong 9 biến từ *c2a1 - c2a9* của file ví dụ Data thuc hanh, bạn có nhu cầu muốn biết báo Tuổi trẻ được đọc thường xuyên tới đâu theo thông tin mẫu thu thập được. Nhưng lựa chọn đọc báo Tuổi Trẻ được cung cấp nằm rải rác trong các biến từ *c2a1* đến *c2a9*, làm sao để nhặt chúng ra?

Bạn sẽ tạo 1 biến mới với 2 biểu hiện (biến Dichotomy): biểu hiện 1 với người có đọc báo Tuổi Trẻ, và 0 với người không bao giờ đọc Tuổi Trẻ. Sau đó đếm tần số gặp số 1, bạn sẽ biết được số người đọc báo Tuổi Trẻ và số người không bao giờ đọc.

Cách thực hiện:

1. Vào menu Transform > Count mở hộp thoại Count

Hình 1.11

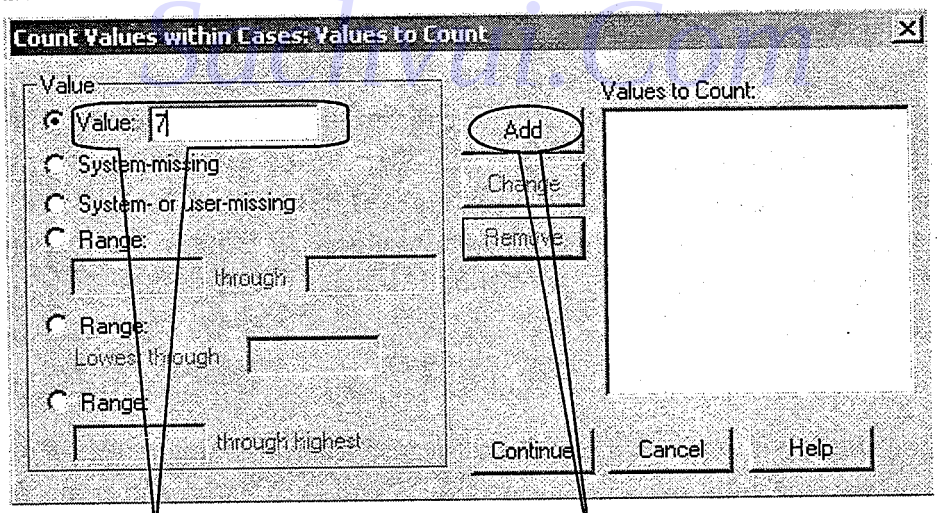


2. Khai báo tên biến Dichotomy muốn tạo trong khung Target Variable và nhãn biến trong khung Target Label.

3. Đưa các biến từ *c2a1* đến *c2a9* vào khung Numeric Variables

4. Nhấp nút Define Values...mở hộp thoại Count Values Within Cases: Values to Count

Hình 1.12



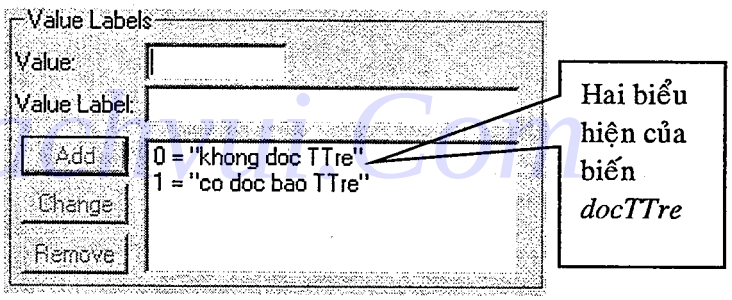
5. Nhập số 7 (là con số được mã hoá cho báo Tuổi Trẻ trong các biến từ *c2a1* đến *c2a9*) vào khung Value rồi bấm nút Add để đưa nó sang khung Values to Count, ở bước này bạn đã yêu cầu SPSS đếm tất cả các trường hợp quan sát dọc theo các biến từ *c2a1* đến *c2a9* để xem có

gặp giá trị 7 không, nếu gặp thì SPSS sẽ gán một con số 1 ở biến *docTTre*, nếu không thì nó để giá trị 0.

6. Bấm Continue trở lại hộp thoại chính và OK.

Danh sách biến trên cửa sổ Variable View của file *Data thuc hanh* có thêm một biến mới tên là *docTTre* được đặt ở vị trí dưới cùng. Biến này nhận giá trị 1 tại những trường hợp mà có chọn đọc báo Tuổi trẻ ở một trong các biến từ *c2a1* đến *c2a9* (trường hợp được nhập liệu là số 7) và 0 tại các trường hợp không có giá trị 7 tại biến bất kỳ nào trong phạm vi 9 biến con kể trên. Bạn phải vào thuộc tính Value của biến *docTTre* để khai báo với SPSS thông tin về 2 biểu hiện của biến Dichotomy vừa được tạo ra như Hình 1.13 dưới đây.

Hình 1.13



Lúc này bạn đã có thể tiến hành các thủ tục thống kê để truy xuất thông tin trong biến Dichotomy *docTTre* mà bạn vừa tạo được, tuy nhiên giờ bạn hãy tạm để biến *docTTre* ở đây, chúng ta sẽ quay lại kiểm chứng nó ở cuối phần Xử lý biến nhiều chọn lựa thuộc Chương III khi mà bạn đã làm quen với hầu hết của thủ tục thống kê giúp mô tả dữ liệu của SPSS.

6.3. Thủ tục Compute để tính toán giá trị biến mới từ biến có sẵn

Lệnh Compute thuộc menu Transform (Transform>Compute...) được sử dụng để tính toán các giá trị mới từ các biến đã có sẵn trong fie dữ liệu, kết quả tính toán của lệnh Compute thường chứa trong một biến mới hoặc chồng lên một biến khác sẵn có là tuỳ thao tác của bạn. Trong các Chương VII và VIII chúng ta sẽ tiếp tục sử dụng lệnh này khá nhiều.

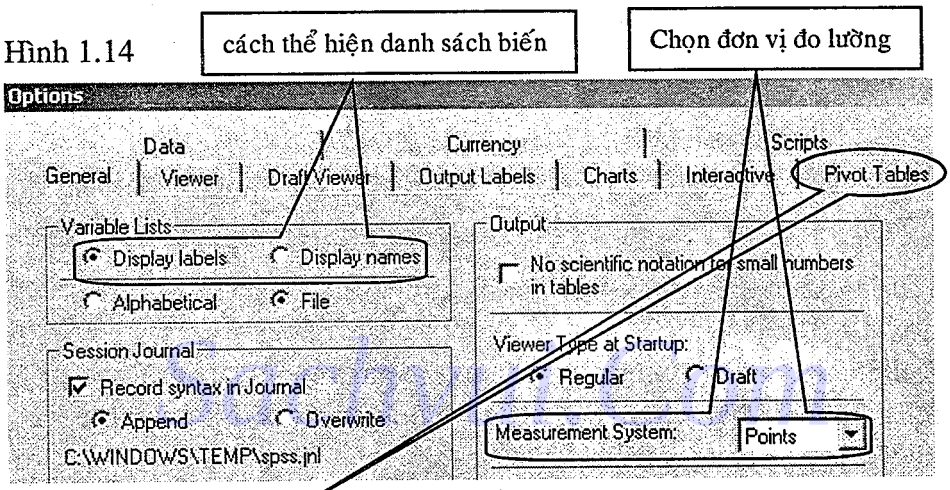
7. THAY ĐỔI MỘT SỐ MẶC ĐỊNH CỦA CHƯƠNG TRÌNH

Để tiện cho việc sử dụng chương trình, bạn cần thay đổi một số mặc định của chương trình trong hộp thoại Options. Từ menu chọn: Edit -> Options. Hộp thoại Options xuất hiện:

Nhấn nút chuột vào dấu mũi tên trong phần Measurement System, chọn 1 trong 3 đơn vị đo lường khoảng cách (thường dùng cm)

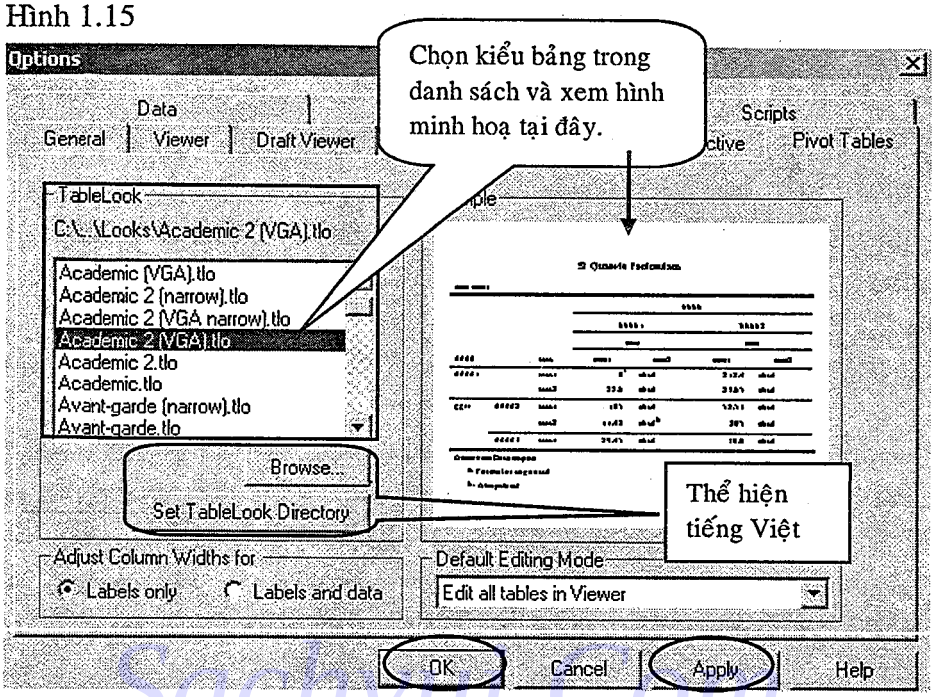
Chọn cách thể hiện danh sách biến trong các hộp thoại lệnh theo kiểu hiện tên biến (Display names) hay nhãn biến (Display lables)

Hình 1.14



Trong thẻ Pivot Table (Hình 1.15) bạn có thể chọn các kiểu định dạng bảng có sẵn mà bạn ưa thích trong danh sách các kiểu bảng biểu bên ô TableLook thay cho kiểu mặc định mà một số người có thể cho rằng trông khá nặng nề hay đơn điệu.

Hình 1.15



Sau khi thực hiện xong các lựa chọn nhấp nút Apply, rồi nhấp nút OK.

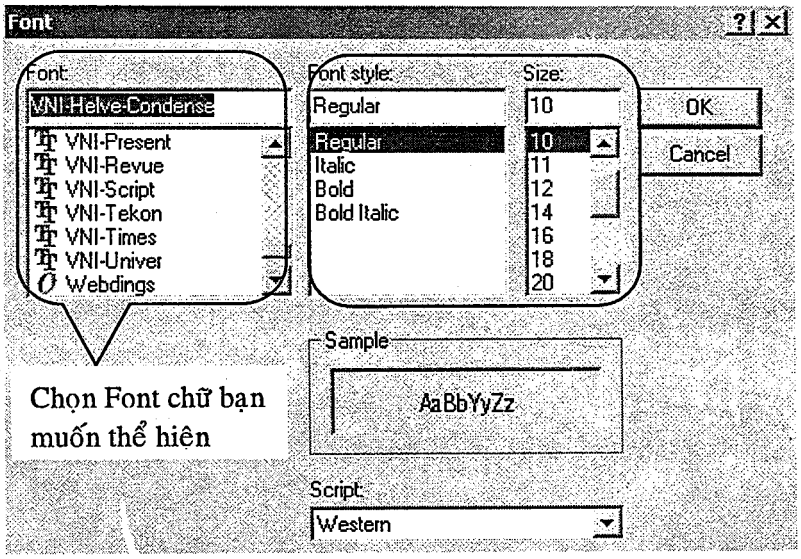
Khi nhấp nút Apply, chương trình có thể đưa ra một hay một số lưu ý nói rằng các điều chỉnh này chỉ có hiệu lực sau khi mở lại file hay khởi động lại chương trình SPSS.

8. THỂ HIỆN TIẾNG VIỆT TRONG SPSS

Để thể hiện tiếng Việt trong SPSS, chúng ta cần chỉnh sang font chữ Việt ở cửa sổ dữ liệu và cửa sổ kết quả. Trong sách này, font VNI-Helve-Condense được sử dụng để làm mẫu.

- Đối với cửa sổ dữ liệu: chọn menu View>Font sẽ mở ra hộp thoại để chỉnh font chữ trên cửa sổ dữ liệu, bạn có thể chọn cả kiểu định dạng chữ ngiêng hay đậm và cỡ chữ mà bạn muốn. Sau khi nhấp nút OK bạn sẽ thấy font chữ trên cửa sổ dữ liệu thay đổi sang kiểu bạn vừa định dạng (ví dụ ở đây là font VNI-Helve Condense) nếu trước đó bạn đã nhập chữ Việt. Còn nếu chưa nhập chữ Việt thì lựa chọn này cũng cho phép bạn sau đó có thể thể hiện tiếng Việt trong quá trình nhập liệu

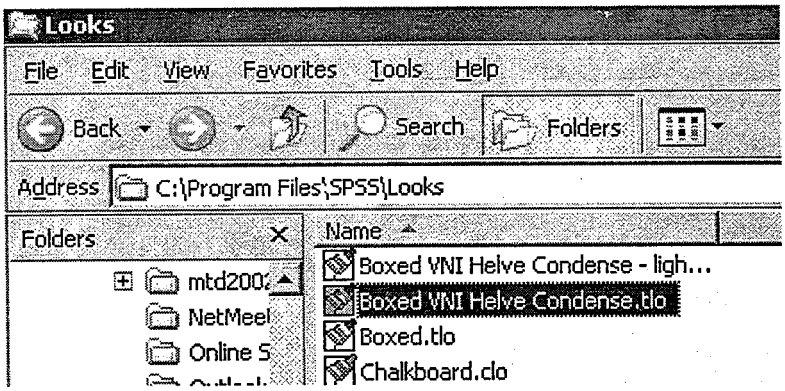
Hình 1.16



- Đối với cửa sổ kết quả xử lý, để cho đơn giản chúng ta sẽ sử dụng file mẫu tiếng Việt có sẵn trong trong đĩa kèm theo sách. Quy trình xác lập tiếng Việt trong cửa sổ kết quả như sau:

1. Chép file ***Boxed VNI Helve Condense.tlo*** trong đĩa vào thư mục Looks của thư mục SPSS bạn đã cài đặt (thông thường có đường dẫn là: C:\Program Files\SPSS\Looks)

Hình 1.17

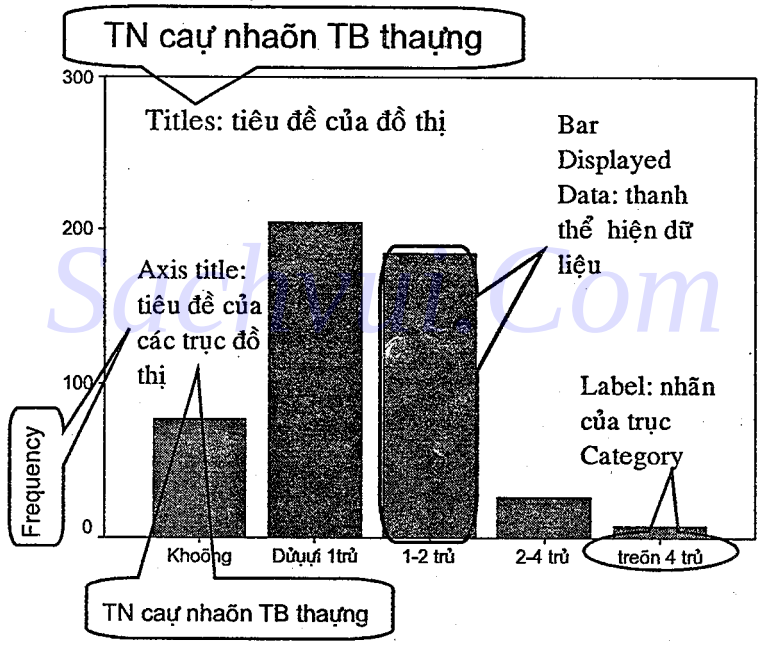


2. Từ menu SPSS chọn Edit -> Options. Trong hộp thoại Options, hãy chọn thẻ (tab) Pivot Tables. Trong phần TableLook, hãy

tìm và chọn sáng tên file *Boxed VNI Helve Condense.tlo*, rồi nhấp nút Set TableLook Directory, nút Apply và cuối cùng là nút OK thì các bảng biểu kết quả xử lý bạn tạo ra đều hiện chữ Việt (tất nhiên là với điều kiện trước đó bạn đã khai báo các biến ở dạng tiếng Việt có dấu).

- Đối với đồ thị, giả dụ bạn đã tạo được đồ thị dạng thanh ngang sau đây trên màn hình kết quả Output của SPSS, hãy xem những thành phần cơ bản thể hiện tiếng Việt trên đồ thị:

Hình 1.18



Để chuyển các tiêu đề thành tiếng Việt bạn hãy tiến hành như sau:

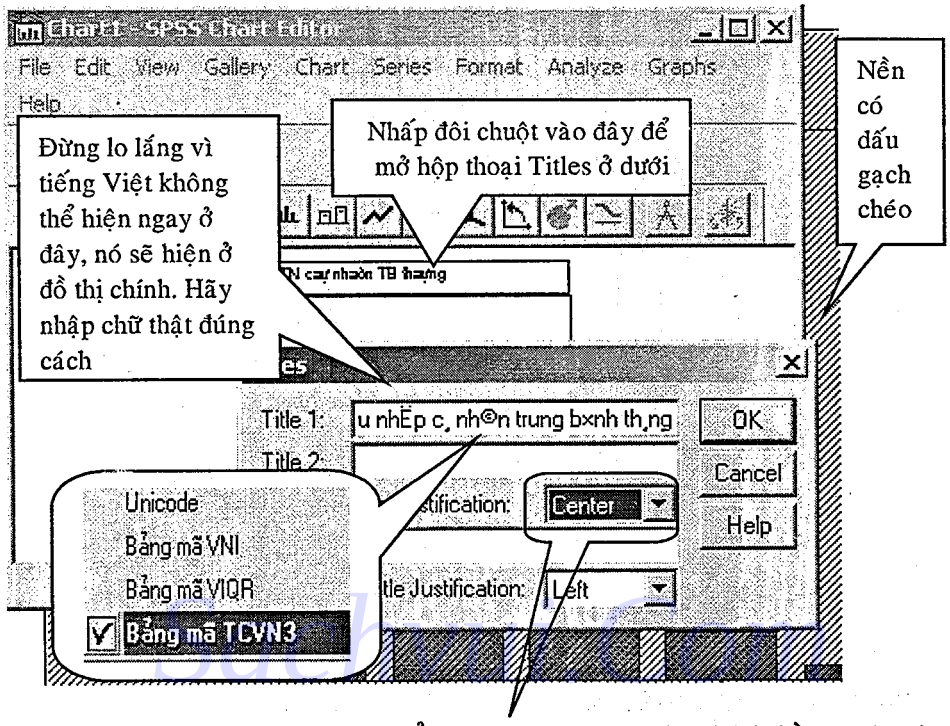
1. Từ menu SPSS chọn Edit -> Options. Trong hộp thoại Options, hãy chọn thẻ Charts, trong phần Current Settings, hãy tìm và chọn font *.VnArialH*, sau đó chọn OK

2. Rê chuột vào đồ thị trên cửa sổ Output, nhấp đôi chuột để mở hộp thoại Chart Editor, lúc này nền đồ thị trên màn hình Output chuyển thành dạng có đấu gạch chéo.

3. Rê chuột đến thành phần muốn thể hiện, nhấp đôi chuột để mở trực tiếp hộp thoại hiệu chỉnh thành phần đó, ví dụ bạn

muốn thay đổi tiêu đề của đồ thị thì nhấp đôi chuột vào tiêu đề, hộp thoại Titles mở ra, bạn xoá tên tiều đề cũ và nhập vào tên tiêu đề mới (nhớ chọn Bảng mã TCVN3)

Hình 1.19



Trên hộp Titles bạn có thể chọn vị trí đặt Title chính bằng các chọn lựa ở đây, trong ví dụ này ta chọn Center.

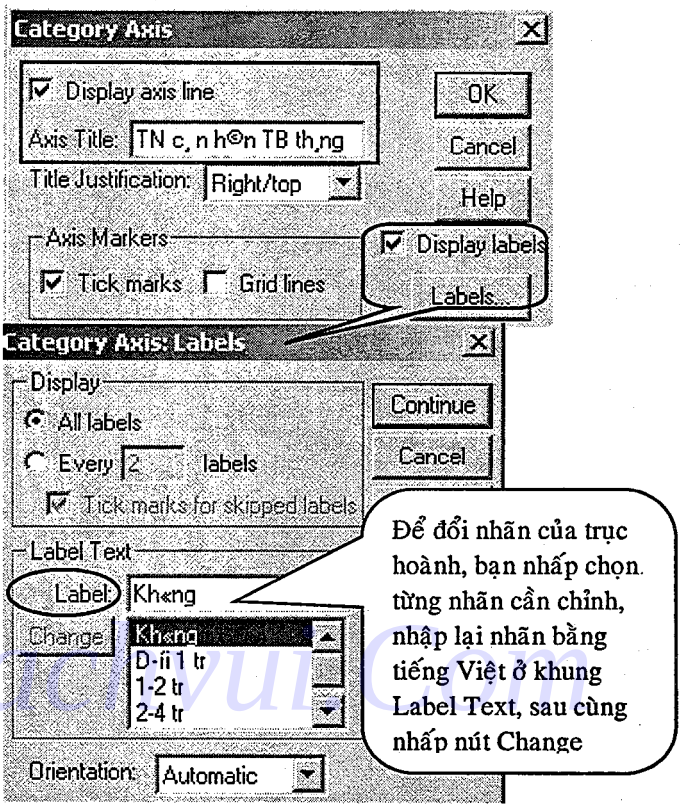
Để hiệu chỉnh tiếng Việt cho thành phần tiêu đề của trục hoành (Category), bạn cũng nhấp chọn tương tự Titles, hộp thoại Category Axis sẽ mở ra (Hình 1.20). Bạn cũng sửa lại tiêu đề và chọn vị trí đặt tiêu đề theo ý muốn, ở ví dụ này bạn thử chọn đặt tiêu đề bên phải (Right/top).

Ngay ở hộp thoại Category Axis, bạn cũng hiệu chỉnh luôn nhãn của trục hoành bằng cách nhấp tiếp nút Labels...để vào hộp thoại Category Axis:Labels (xem hướng dẫn ở Hình 1.20).

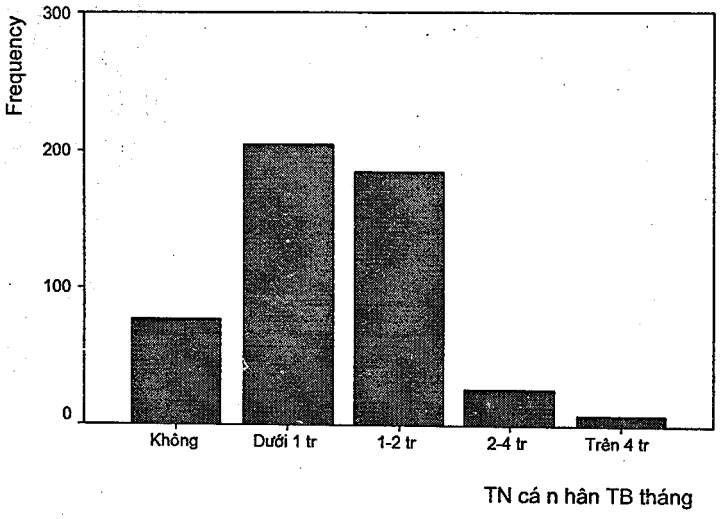
4. Nhớ nhấp nút Continue khi hoàn thành việc hiệu chỉnh.

Đồ thị đã được chỉnh sửa hoàn tất của bạn sẽ trông như Hình 1.21

Hình 1.20



Hình 1.21 Thu nhập cá nhân trung bình tháng



CHƯƠNG II

LÀM SẠCH DỮ LIỆU

1. SỰ CẦN THIẾT

Dữ liệu sau khi nhập xong thường chưa thể đưa ngay vào xử lý và phân tích vì có thể còn nhiều lỗi do:

- Chất lượng của phỏng vấn và đọc soát: phỏng vấn viên hiểu sai câu hỏi và thu thập dữ liệu sai, PVV chọn sai đối tượng phỏng vấn hoặc ghi chép nhầm, người được phỏng vấn trả lời sai ý, người đọc soát chưa phát hiện được ...
- Nhập dữ liệu: sai, sót, thừa

Ví dụ: khi bạn đã quy ước mã hoá 1 đại diện cho nam và 2 đại diện cho nữ sau khi thực hiện lệnh đếm tần số của biến giới tính bạn thu được kết quả như dưới đây thì không còn nghi ngờ gì nữa rằng bạn đã nhầm lẫn khi nhập liệu, tức là thay vì gõ một con số 1 cho giới tính nam thì bạn lại gõ tới hai lần.

Bảng 2.1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nam	248	49.6	49.6	49.6
	Nữ	251	50.2	50.2	99.8
	11	1	.2	.2	100.0
	Total	500	100.0	100.0	

2. CÁC BIỆN PHÁP NGĂN NGỪA

"Phòng bệnh hơn chữa bệnh", vì vậy trước tiên bạn hãy bảo đảm đã thực hiện các biện pháp ngăn ngừa lỗi sau:

- Thiết kế bản câu hỏi rõ ràng, dễ hỏi, dễ trả lời
- Chọn lọc và huấn luyện phỏng vấn viên kỹ lưỡng, điều tra phỏng vấn thử trước khi phỏng vấn thật để hiểu thống nhất, tránh sai sót.
- Các bản câu hỏi sau khi phỏng vấn xong phải được đọc soát kiểm lỗi, chỉnh sửa trước khi nhập.

- Việc mã hoá phải được tiến hành tập trung với một số ít cá nhân phụ trách việc nhập liệu chứ không nên phân tán để tránh bị rối loạn do thiếu thống nhất.

3. CÁC PHƯƠNG PHÁP LÀM SẠCH DỮ LIỆU

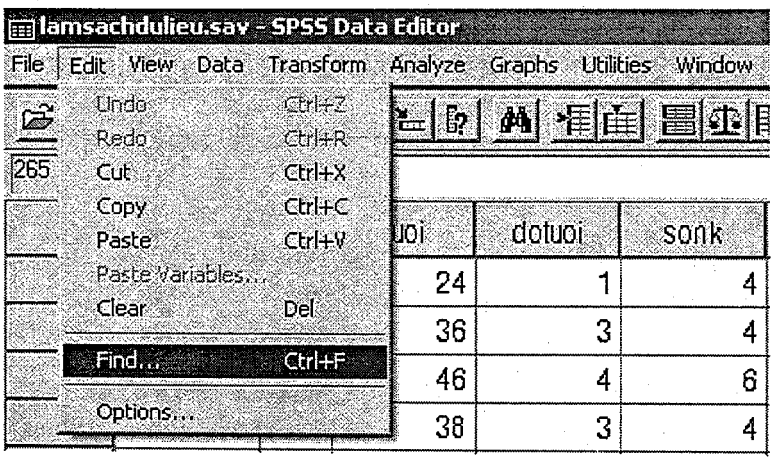
3.1. Dùng bảng tần số

Lập bảng tần số cho tất cả các biến, đọc soát để tìm các giá trị lạ tại các biến như giá trị 11 trong ví dụ trên. Sau đó tại các biến có lỗi bạn dùng lệnh Find để tìm vị trí của giá trị lỗi, rồi chỉnh sửa. Cách tiến hành lập bảng tần số như Bảng 2.1 bạn sẽ làm quen ở Chương III. Còn cách thức dùng thủ tục Find tìm lỗi sẽ tiến hành như sau:

1. Trên cửa sổ Data View bạn chọn toàn bộ cột tương ứng với biến có giá trị bị lỗi (cách chọn giống như bạn vẫn chọn cột trong Excel, tức rê trỏ chuột lên đầu cột, khi con trỏ chuột thành dạng mũi tên màu đen hướng xuống thì click chuột trái một lần, sau khi được chọn toàn bộ cột sẽ bị bôi đen)

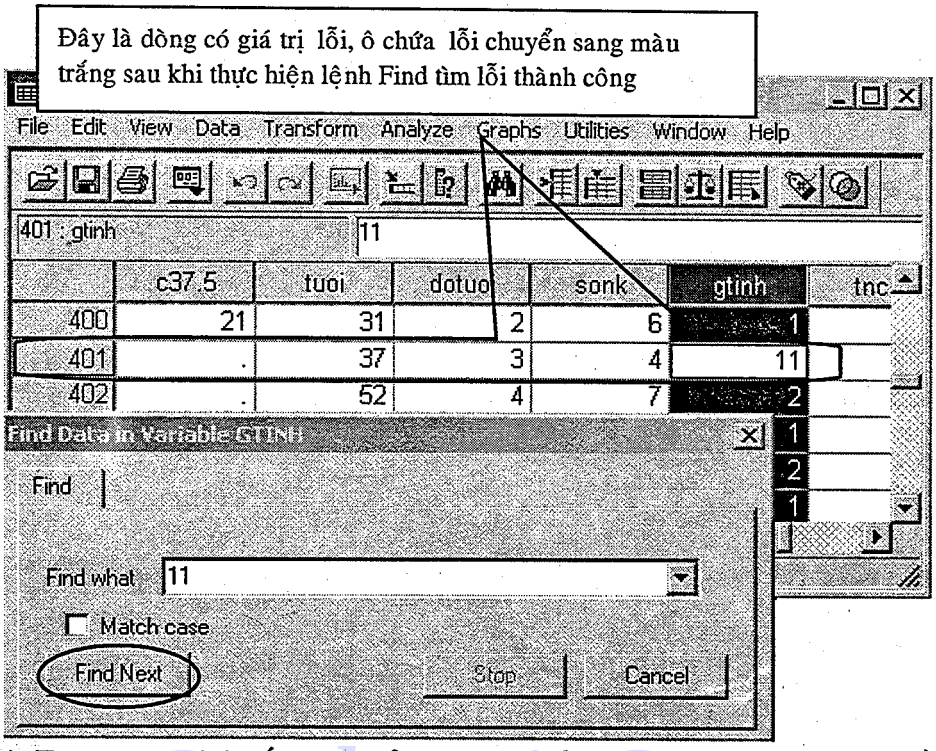
2. Vào menu Edit >Find như hình dưới bạn sẽ mở được cửa sổ Find Data in Variable GTINH ở Hình 2.

Hình 2.1



3. Nhập giá trị 11 vào ô Find what, bấm nút Find next thì vị trí của ô chứa giá trị lỗi 11 trên màn hình dữ liệu sẽ được đổi thành màu trắng để bạn nhận thấy (Hình 2.2)

Hình 2.2



4. Truy ngược lại số thứ tự của hàng đó (ở đây là 401) bạn sẽ tìm về được bản câu hỏi tương ứng.

Diễn giải thì thấy khá dài dòng, nhưng sau khi tiến hành vài lần bạn sẽ phát hiện ra tiến trình tìm lỗi bằng lệnh Find đơn giản hơn nhiều, nhưng với lần đầu bạn hãy đi theo đúng thủ tục ở trên.

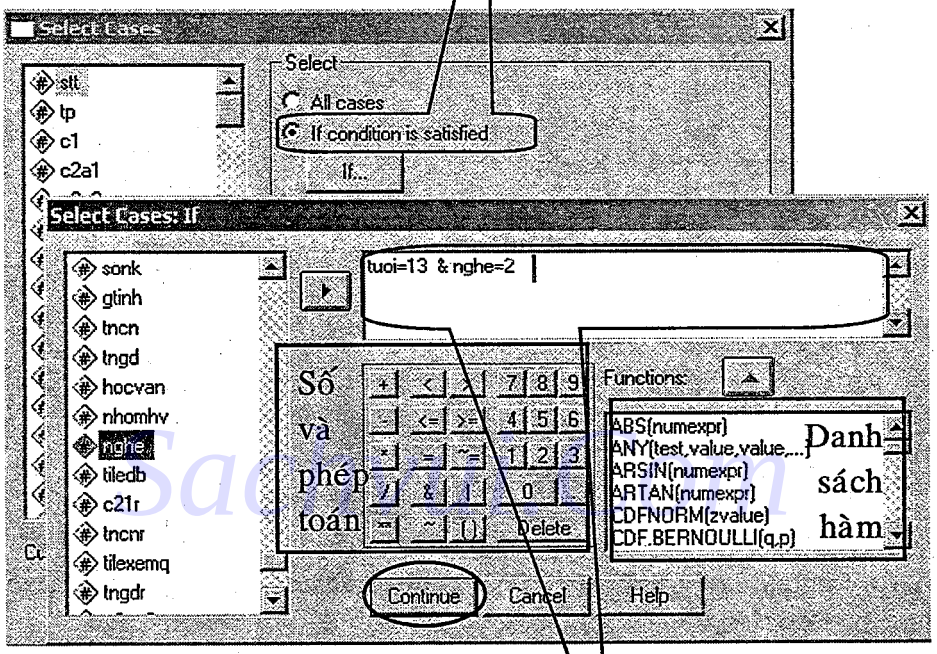
- Ưu điểm của cách tìm lỗi này là đơn giản, dễ thực hiện.
- Nhược điểm là thủ công, phát hiện ít lỗi, chỉ phù hợp với các bản câu hỏi đơn giản.

3.2. Dùng bảng phối hợp hai biến hay ba biến

Bạn lập bảng kết hợp biến (xem nội dung này ở Chương III) rồi dựa vào các quan hệ hợp lý (logic) để phát hiện ra lỗi. Ví dụ như khi lập bảng kết hợp biến tuổi và nghề nghiệp mà bạn phát hiện thấy có trường hợp tuổi chỉ có 13 mà nghề nghiệp ghi là giáo viên tức là một trong hai biến tuổi hoặc nghề nghiệp đã bị nhập sai. Sau khi phát hiện có lỗi, bạn dùng lệnh Select cases để tìm ra trường hợp có lỗi đó.

1. Vào menu Data>Select Cases mở hộp thoại Select Case, trong hộp thoại này bạn lựa chọn mục If condition is satisfied để chỉ định cho SPSS lọc ra trường hợp thỏa điều kiện tuổi = 13 và nghề = giáo viên (trong file ví dụ *lamsachdulieu* nghề giáo viên được mã hoá là 2).

Hình 2.3

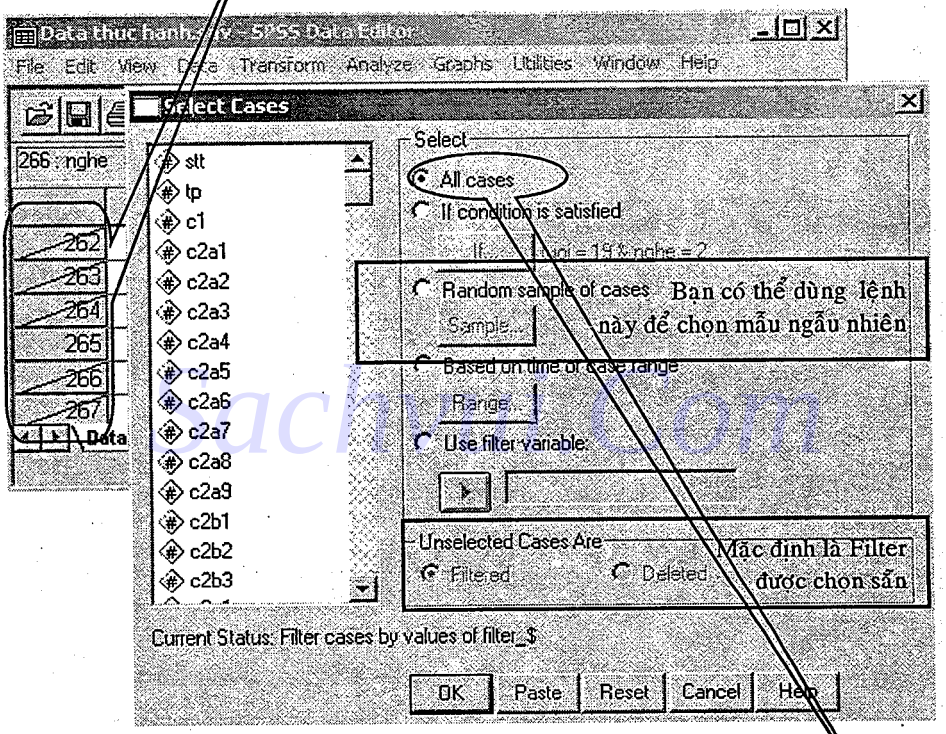


2. Bấm vào nút If... nằm kế dưới lựa chọn tình huống If condition is satisfied trong hộp thoại Select Case để mở tiếp hộp Select Case: If, bạn sẽ xây dựng biểu thức hàm If tại khung nhập hàm theo trình tự: chọn sáng biến *tuoi*, nhấp nút mũi tên đưa nó sang khung nhập hàm bên phải, nhập dấu =, nhập số 13, nhập dấu &, lại chọn sáng biến *nghe*, nhấp nút mũi tên đưa nó sang khung nhập hàm, nhập dấu =, nhập số 2, rồi bấm nút Continue để trở lại hộp thoại Select Case.

3. Trước khi bấm nút OK trên hộp thoại Select Case thì nhớ kiểm tra là trong khung Unselected Case Are phần Filtered đang được chọn chứ không phải phần Deleted, mặc định của chương trình SPSS là Filtered luôn được chọn sẵn. Nếu bạn muốn lọc dữ liệu với mục đích xoá bớt đi những trường hợp dữ liệu không thoả điều kiện lựa chọn thì hãy bấm vào mục Deleted.

Khi lệnh này đã thực hiện, SPSS sẽ tạo cho bạn một biến mới trong danh sách biến đã có sẵn, tên biến mới này là *filter_$*, biến này nhận giá trị 0 tại tất cả các tình huống không thoả mãn và 1 tại tình huống thoả điều kiện của lệnh If tức tình huống có sai sót. Chú ý rằng *filter_$* chỉ là biến tạm, khi bạn thực hiện một lệnh Select Case mới thì biến này sẽ mất đi. Bên cạnh đó các ô đánh dấu hàng sẽ được gạch chéo tại các hàng không được chọn, điều này đồng nghĩa với việc hàng không có dấu gạch chéo ở ô đánh dấu hàng sẽ nhận giá trị 1 tại biến filter_$.

Hình 2.4



Thêm một chú ý nữa là bất cứ khi nào đã thực hiện thành công lệnh Select Case để tìm kiếm hay lựa chọn được những trường hợp cần lọc, bạn phải trở lại hộp thoại Select Case trả lại tình huống All Case. Nếu không các lệnh thống kê sau đó sẽ chỉ được thực hiện với những trường hợp được lọc, và như vậy nghĩa là kết quả không còn chính xác.

Đến đây sẽ nảy sinh một vấn đề nữa là nếu số dòng quá lớn thì việc tìm được dòng chứa giá trị nhập lỗi (tức dòng mang giá trị 1 ở biến Filter hay dòng không bị gạch chéo) cũng khá mất công. Lúc này bạn nhờ đến sự trợ giúp của lệnh Sort Case.

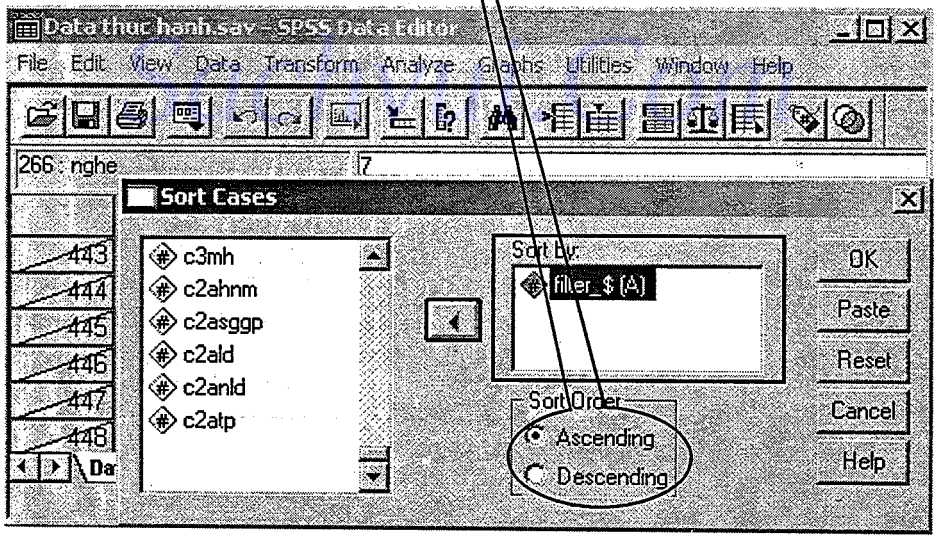
1. Vào menu Data>Sort Case, đưa biến bạn muốn sắp xếp thứ tự (ở tình huống này là biến *filter_$*) vào khung Sort by

2. Chọn hình thức sort tăng dần (Ascending) hay giảm dần (Descending). Nếu sắp xếp tăng dần thì hàng mang giá trị 1 sẽ nằm dưới cùng và ngược lại.

3. Nhấp OK.

Từ số 1 trên biến *filter_$,* tiến hành dò ngược lại số thứ tự của dòng bạn sẽ tìm ra vị trí bản câu hỏi bị lỗi.

Hình 2.5



- Ưu điểm của biện pháp này là phát hiện được nhiều lỗi hơn, phù hợp với các bản câu hỏi phức tạp.
- Nhược điểm là phức tạp, đòi hỏi nhiều thời gian, cần có nhiều kinh nghiệm mới thực hiện được.

3.3. Cách tìm lỗi đơn giản ngay trên cửa sổ dữ liệu (Data View)

Bạn có thể sử dụng lệnh Sort Case vừa nói trên để tìm những lỗi đơn giản ngay trên cửa sổ dữ liệu, ví dụ với tình huống giới tính, chỉ cần chọn lệnh sắp xếp dữ liệu giảm dần, nếu giá trị lớn nhất không phải là 2 mà là một giá trị bất kỳ lớn hơn 2 nghĩa là bạn đã tìm ra lỗi rồi.

Bạn đọc có thể thực tập các phương pháp tìm lỗi trên file *lamsachdulieu* trong tập hợp dữ liệu dùng kèm với sách.

CHƯƠNG III

TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU

1. PHƯƠNG PHÁP VÀ CÔNG CỤ

Có rất nhiều phương pháp và công cụ dùng để tóm tắt và trình bày dữ liệu, trong phần này chúng ta xem xét một số phương pháp cơ bản nhất. Tập hợp dữ liệu dùng để minh họa trong phần này lấy từ một cuộc điều tra nhu cầu người đọc báo của Sài Gòn Tiếp Thị được tiến hành vào tháng 7 năm 2001. Tập hợp dữ liệu này có tên là *Data thuc hanh* được cung cấp trong tập hợp dữ liệu dùng kèm với sách.

Các công cụ cơ bản để tóm tắt và trình bày dữ liệu được trình bày trong phần này là:

- Bảng tần số
- Các đại lượng thống kê mô tả, biểu đồ tần số
- Bảng kết hợp nhiều biến
- Đồ thị.

2. BẢNG TẦN SỐ ĐƠN GIẢN

Ta đếm tần số để biết với tập dữ liệu đang có thì số đối tượng có các biểu hiện nào đó ở một thuộc tính cụ thể là bao nhiêu, nhiều hay ít...

Ví dụ: Lập bảng tần số cuả biến *gtinh* của tập tin dữ liệu *Data thuc hanh*, chú ý biến giới tính trong ví dụ này được mã hoá hai biểu hiện: 1–nam và 2–nữ. Tập tin này có 500 hàng tượng trưng cho 500 phiếu điều tra từ 500 đối tượng, kết quả do SPSS đưa ra gồm 2 bảng như dưới đây, các bảng này cho biết trong 500 người được phỏng vấn có 251 nữ, chiếm tỷ lệ 50,2% mẫu, không có giá trị Missing.

Bảng 3.1 **giới tính**

N	Valid	500
	Missing	0

- Dòng Valid cho biết số quan sát hợp lệ (số người có trả lời)
- Dòng Missing cho biết số quan sát bị thiếu dữ liệu (không trả lời)

Bảng 3.2 **giới tính**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nam	249	49.8	49.8	49.8
	Nữ	251	50.2	50.2	100.0
	Total	500	100.0	100.0	

Ý nghĩa của các cột số liệu trong Bảng 3.2

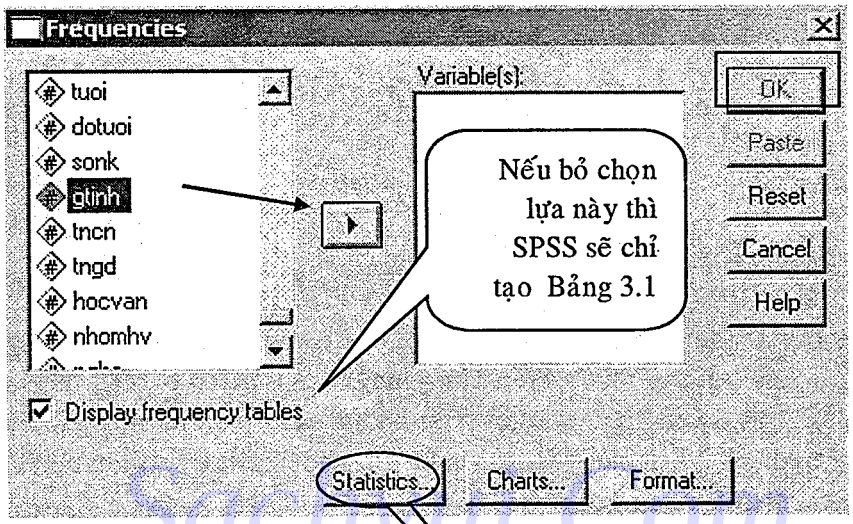
- Cột đầu tiên là các biểu hiện của biến giới tính, như ta đều biết chúng ta có hai biểu hiện là Nam và Nữ, và nếu bạn nhập sai số liệu, ví dụ 11 cho nam thay vì 1 thì chắc chắn ở đây chúng ta sẽ có thêm biểu hiện thứ 3 là 11.
- Cột Frequency là tần số của từng biểu hiện, được tính bằng cách đếm và cộng dồn
- Cột Percent: tần suất tính theo tỉ lệ % bằng cách lấy tần số của mỗi biểu hiện chia cho tổng số quan sát.
- Cột Valid Percent: là phần trăm hợp lệ, tính trên số quan sát có thông tin trả lời, chẳng hạn với câu hỏi về trình độ học vấn có những người được phỏng vấn bỏ khuyết không trả lời thì khi tính phần trăm hợp lệ cho từng loại trình độ bạn phải loại bớt các giá trị khuyết ra bằng cách khai báo Missing value với chương trình SPSS (xem lại phần khai báo giá trị khuyết ở Chương I). Nếu có 8 giá trị khuyết tức 8 người không trả lời câu hỏi về trình độ học vấn thì lúc này phần trăm hợp lệ Valid Percent cũng được tính theo công thức như cột Percent nhưng tính trên mẫu số là tổng số quan sát đã phải trừ đi 8.
- Cột Cumulative Percent: là phần trăm tích luỹ do cộng dồn các phần trăm từ trên xuống, nó cho ta biết có bao nhiêu % đối tượng ta đang khảo sát đang ở mức độ nào đó trở lên.

Bạn có thể thực hiện bảng tần số với tất cả các biến kiểu định tính lẫn định lượng. Trong trường hợp biến định lượng liên tục của bạn có quá nhiều giá trị, ví dụ khi bạn muốn liệt kê tuổi của tất cả các đối tượng được phỏng vấn trong cuộc điều tra này thì bảng tần số sẽ rất dài với những thông tin phân tán, vậy đầu tiên chúng ta phải phân tổ độ tuổi của người trả lời thành một số độ tuổi chính bằng lệnh Recode (xem phần Mã hoá lại biến ở Chương I) rồi mới tính tần số của biến đã được phân tổ này.

Cách thức tiến hành lệnh Frequencies

1. Sau khi mở file *Data thuc hanh*, bạn vào menu Analyze>Descriptive Statistics > Frequencies, hộp thoại Frequencies xuất hiện như Hình 3.1
2. Chọn biến muốn lập bảng tần số (biến *gtinh*) bằng cách nhấp chuột vào tên biến cho biến sáng xanh lên rồi bấm nút có dấu mũi tên hướng sang phải để đưa biến đang chọn vào khung Variable(s).

Hình 3.1



3. Nhấp OK bạn có 2 bảng kết quả thể hiện như Bảng 3.1 và 3.2

Chú ý rằng ngay trong phần này chúng ta cũng có thể chọn được lệnh tính các đại lượng thống kê mô tả cho các biến dạng định lượng, muốn làm được điều đó bạn bấm vào nút Statistics và thực hiện các khai báo cần thiết. Chúng ta sẽ giải quyết tiếp điều này ở phần Lập bảng tần số đồng thời tính toán các đại lượng thống kê mô tả.

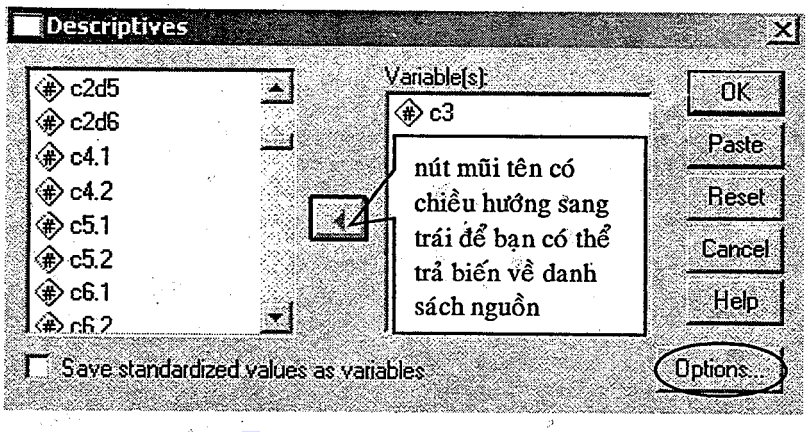
3. CÁC ĐẠI LƯỢNG THỐNG KÊ MÔ TẢ

Các đại lượng thống kê mô tả chỉ được tính đối với các biến định lượng. Nếu tính các đại lượng này đối với các biến định tính thì kết quả sẽ không có ý nghĩa. Thử tưởng tượng với mẫu 500 người gồm 249 nam và 251 nữ, bạn không thể tính được giới tính trung bình của mẫu = (249*1 + 251*2)/500= 1,502 vì giá trị này hoàn toàn vô nghĩa.

Cách thực hiện lệnh tính các đại lượng thống kê mô tả bằng SPSS

1. Vào menu Analyze> Descriptive Statistics > Descriptives..., hộp thoại sau sẽ xuất hiện:

Hình 3.2



2. Chọn một (hay nhiều biến định lượng nếu muốn tính các đại lượng thống kê mô tả cho nhiều biến cùng lúc) ở danh sách biến ở bên trái hộp thoại sau đó nhấp nút có mũi tên hướng sang phải để đưa các biến này vào khung Variable(s). Ở ví dụ này ta chọn biến *c3* có nhãn là "số người thường đọc báo trong nhà"

Chú ý là khi bạn chưa chọn thêm một biến bất kỳ nào khác trong danh sách bên trái thì nút mũi tên sẽ có chiều ngược lại (hướng sang trái) để bạn có thể trả biến vừa đưa sang khung Variable(s) về lại danh sách nếu chẳng may chọn nhầm biến.

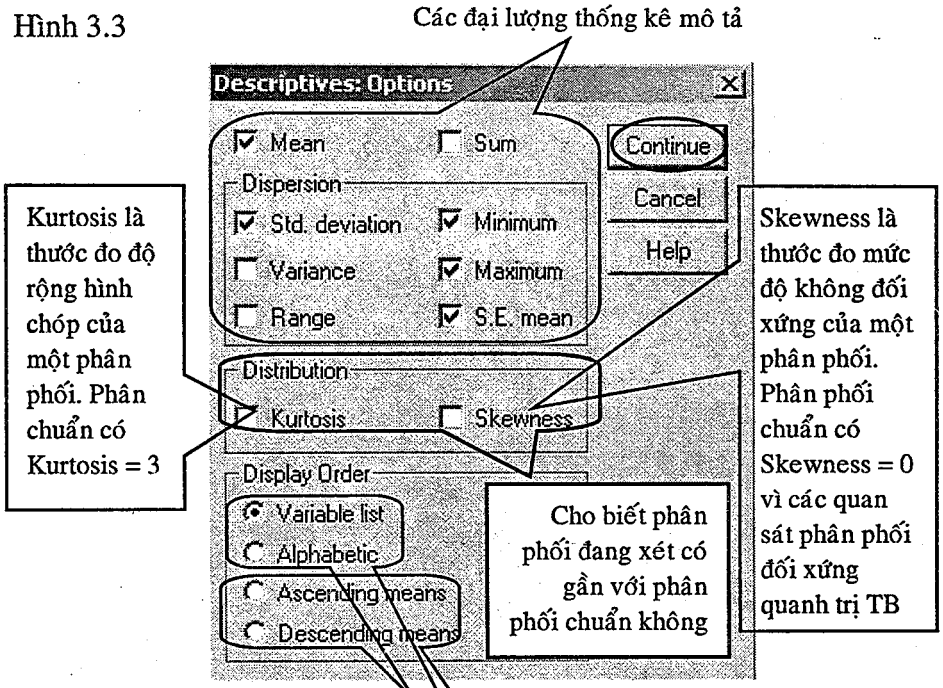
3. Kế tiếp bấm nút Options để vào hộp thoại Desciptives Options, ở đây ta chọn các đại lượng thống kê ta muốn tính toán để mô tả cho biến định lượng đã đưa qua hộp Variable(s) bằng cách nhấp chuột vào các ô vuông cần thiết.

Các đại lượng thống kê mô tả thường được dùng là:

- Mean: trung bình cộng
- Sum: tổng cộng (cộng tất cả các giá trị trong tập dữ liệu quan sát)
- Std. Deviation: độ lệch chuẩn
- Minimum: giá trị nhỏ nhất
- Maximum: giá trị lớn nhất
- SE mean: sai số chuẩn khi ước lượng trị Trung Bình

Các bạn thực hiện các lựa chọn như ở Hình 3.3

Hình 3.3



Trong trường hợp tính toán cho nhiều biến cùng một lúc, bạn có thể chọn 1 trong 4 cách sắp xếp thứ tự kết quả tính toán của các biến này trong bảng kết quả. Thông thường là dùng trật tự tăng dần (Ascending means) hay giảm dần (Descending means) của giá trị trung bình của mỗi biến. Ngoài ra bạn có thể sắp xếp theo thứ tự các biến (Variable list) được đưa lần lượt vào khung Variable(s) khi thực hiện lệnh hoặc theo thứ tự Alphabetic của nhãn biến.

4. Sau đó bấm vào nút Continue để trở lại hộp thoại trước, rồi nhấn nút OK. Bảng kết quả các đại lượng thống kê mô tả của biến *c3* sẽ xuất hiện trên cửa sổ Output.

Bạn hãy thực hiện lệnh Desciptives theo hướng dẫn trên và xem bảng kết quả do SPSS tạo ra, bạn có nhận thấy vấn đề gì không?

Lựa chọn cách thể hiện bảng kết quả

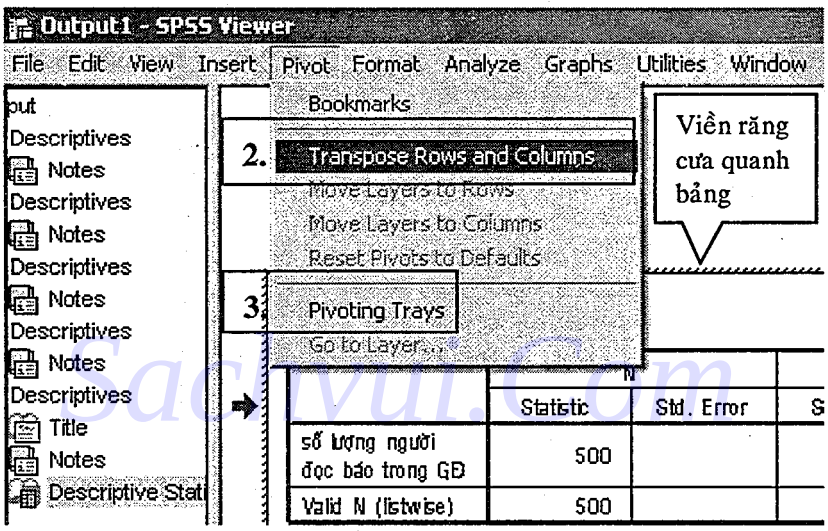
Bạn đã nhận thấy vấn đề rồi chứ? Bề ngang của bảng Descriptive Statistics quá rộng khiến cho bạn khó bao quát hết các số liệu trên bảng, thêm nữa bạn sẽ không thể in hoặc chép bảng trên bề ngang

một tờ A4. Lúc này bạn phải lựa chọn cách thể hiện bảng kết quả sao cho có thể sử dụng thuận lợi nhất.

Bạn tiến hành như sau:

1. Trên cửa sổ Output, đưa con trỏ chuột vào vị trí bảng và nhấp đôi để quanh bảng hiện viền răng cưa (dấu hiệu chuyển sang chế độ chỉnh sửa).

Hình 3.4



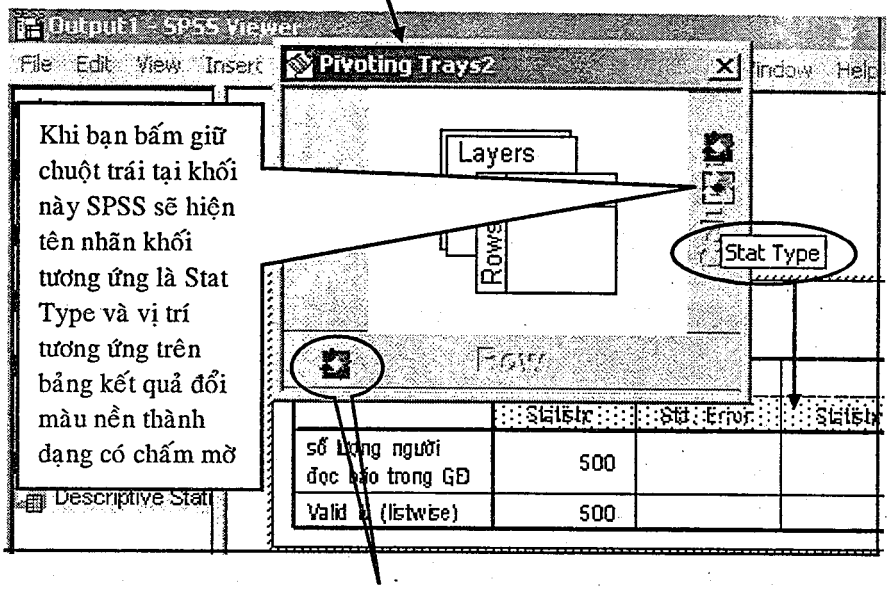
2. Sau đó bạn vào menu Pivot, chọn Transpose Row and Columns để chuyển đổi hàng của bảng thành cột và cột của bảng thành hàng. Bảng tính các đại lượng thống kê mô tả của biến *c3* lúc này sẽ trở nên dễ nhìn hơn như sau:

Bảng 3.3

		số lượng người đọc báo trong GĐ	Valid N (listwise)
N	Statistic	500	500
	Std. Error		
Minimum	Statistic	1	
	Std. Error		
Maximum	Statistic	15	
	Std. Error		
Mean	Statistic	3.47	
	Std. Error	.08	
Std. Deviation	Statistic	1.800	
	Std. Error		

3. Cách chuyển đổi thứ 2 là chọn Pivoting Trays ở dưới Transpose Row and Columns, cửa sổ Pivoting Trays mở ra ở góc trên bên trái của cửa sổ Output

Hình 3.5



Lần lượt rê chuột vào các khối vuông 4 màu trên Pivoting Trays và bấm giữ chuột trái bạn sẽ đọc được các nhãn giới thiệu chúng là đại diện của đối tượng gì trên bảng kết quả mà bạn muốn hiệu chỉnh (xem Hình 3.5). Bạn cũng sẽ thấy vị trí tương ứng trên bảng kết quả đang được chọn hiệu chỉnh (tức là bảng có viền răng cưa xung quanh) của đối tượng mà chúng đại diện vì các đối tượng này bị đổi màu nền. Rê chuột nhấc hai khối Statistics và Stat Type xuống đổi vị trí cho khối Variables, sau khi nhấp biểu tượng X đóng cửa sổ Pivoting Trays lại bạn sẽ thu được kết quả y hệt như Bảng 3.3.

Hãy thử một vài cách sắp xếp các khối vuông này theo ngẫu hứng của bạn, bạn sẽ phát hiện thêm nhiều điều thú vị và hữu ích.

Ý nghĩa các kết quả trên Bảng 3.3

Từ trên đi xuống ta có các đại lượng thống kê sau:

- N là tổng số quan sát tức là cỡ mẫu, ở đây cỡ mẫu của ta là 500 người.

- Minimun là giá trị nhỏ nhất gặp được trong các giá trị của biến *c3*, ở đây là 1 có nghĩa số người đọc báo trong gia đình ít nhất khảo sát được theo mẫu này là 1 người.
- Maximum là giá trị lớn nhất gặp được trong các giá trị của biến *c3*, giá trị lớn nhất của biến *c3* gặp được trong mẫu này là 15 người.
- Mean: trung bình có 3,47 người đọc báo trong mỗi gia đình thuộc mẫu của chúng ta
- Std Error (của mean) là sai số chuẩn khi dùng giá trị trung bình mẫu để ước lượng giá trị trung bình của tổng thể, ta có Std Error = 0,8.
- Std Deviation chính là độ lệch chuẩn cho biết mức độ phân tán của các giá trị của *c3* quanh giá trị trung bình 3,47. Giá trị Std Deviation=1,8. Vậy phương sai trong tình huống này sẽ $=1,8^2$. Bạn hãy chọn thêm tuỳ chọn Variance ở hộp thoại Desciptives Options để kiểm chứng điều này.

4. LẬP BẢNG TẦN SỐ ĐỒNG THỜI TÍNH TOÁN CÁC ĐẠI LƯỢNG THỐNG KÊ MÔ TẢ

Bạn cũng có thể vừa lập bảng tần số vừa đồng thời tính các đại lượng thống kê mô tả, xin nhắc lại một lần nữa là điều này chỉ áp dụng đối với biến định lượng.

Trong thực tế ít khi ta dùng lệnh tính các đại lượng thống kê mô tả riêng lẻ mà ta thường kết hợp vừa lập bảng tần số vừa tính các đại lượng thống kê mô tả.

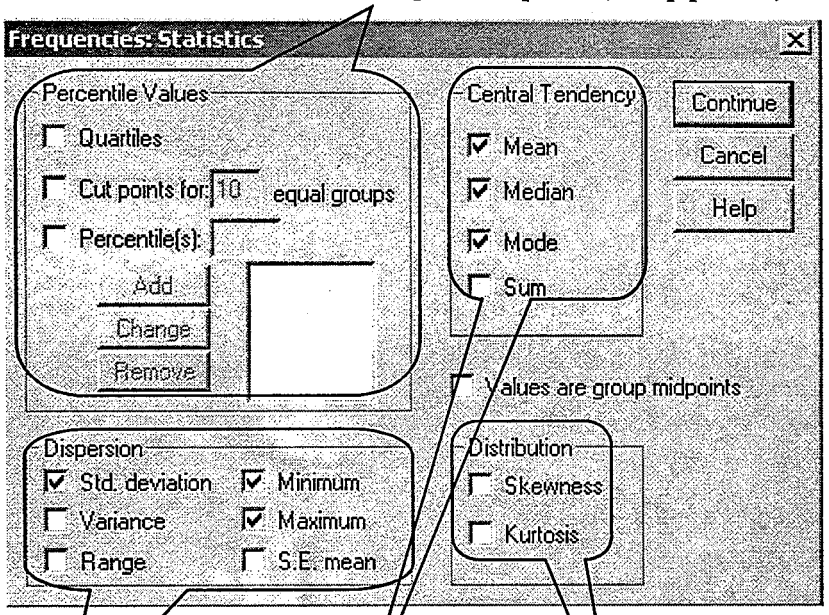
Cách thực hiện

Bạn cũng vào menu Analyze > Descriptive Statistics > Frequencies

1. Trả các biến cũ trong ô Variable(s) về bên trái, chọn biến định lượng (ví dụ chọn *c3*) và đưa biến này vào ô Variable(s).

2. Nhấn nút Statistics... để mở tiếp hộp thoại tính các đại lượng thống kê mô tả như hình sau:

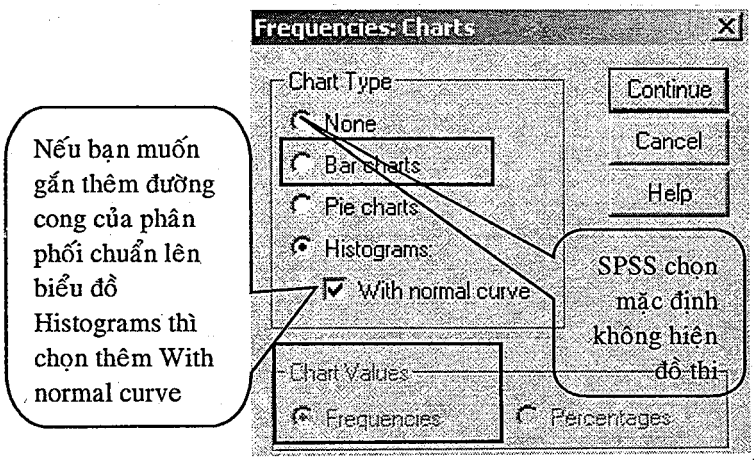
Hình 3.6 Các giá trị tứ phân vị, thập phân vị



3. Trong hộp thoại này, nhấp chuột vào các ô vuông để chọn các đại lượng thống kê cần tính như các đại lượng thống kê đo lường khuynh hướng phân tán, khuynh hướng tập trung, hình dáng phân phối...rồi nhấn nút Continue để trở lại hộp thoại Frequencies.

4. Để vẽ biểu đồ tần số, bấm vào nút Charts... hộp thoại Frequencies: Charts sẽ xuất hiện

Hình 3.7



Trong hộp thoại Charts, nhấp chuột vào các ô để chọn loại biểu đồ cần vẽ. Có thể chọn 1 trong 3 loại biểu đồ sau:

- Bar: biểu đồ dạng thanh (dùng cho biến có các giá trị rời rạc, biến của dữ liệu định tính)
- Pie: biểu đồ hình tròn (hay dùng cho việc mô tả cấu trúc hiện tượng)
- Histograms: biểu đồ phân phối tần số dùng cho biến của dữ liệu liên tục

Sau khi chọn loại biểu đồ (ở ví dụ này bạn sẽ chọn Bar chart để có một đồ thị có cấu trúc sát theo bảng tần số vừa lập được) thì khu vực Chart Value sẽ sáng lên, ở đây bạn lựa chọn vẽ đồ thị theo giá trị tần số (Frequencies), nhấp chuột vào nút Continue để trở về hộp thoại Frequencies.

5. Nhấn nút OK. Kết quả hiện ra như sau:

Bảng 3.4

số lượng người đọc báo trong GĐ

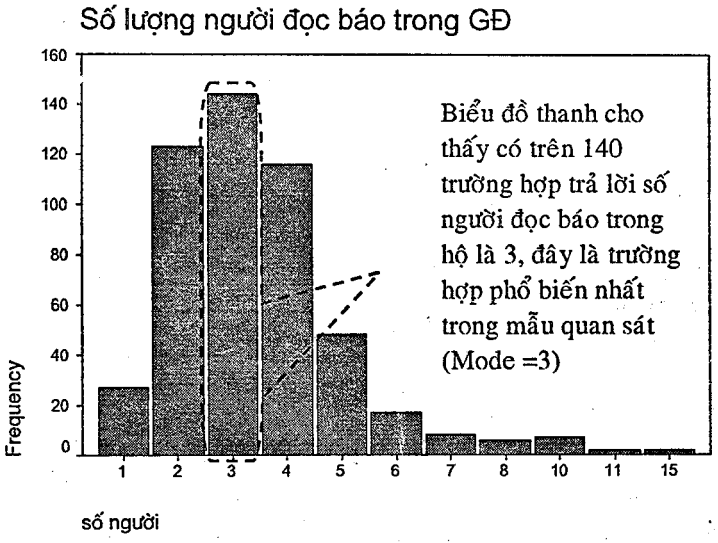
N	Valid	500
	Missing	0
Mean		3.47
Median		3.00
Mode		3
Std. Deviation		1.800
Minimum		1
Maximum		15

Bảng 3.5

số lượng người đọc báo trong GĐ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	27	5.4	5.4	5.4
	2	123	24.6	24.6	30.0
	3	144	28.8	28.8	58.8
	4	116	23.2	23.2	82.0
	5	48	9.6	9.6	91.6
	6	17	3.4	3.4	95.0
	7	8	1.6	1.6	96.6
	8	6	1.2	1.2	97.8
	10	7	1.4	1.4	99.2
	11	2	.4	.4	99.6
	15	2	.4	.4	100.0
	Total	500	100.0	100.0	

Hình 3.8



Trong các đại lượng thống kê mô tả được tính ở Bảng 3.4 ta biết thêm thông tin về giá trị trung vị (Median) của *c3* là 3, có nghĩa là khi số liệu về số người đọc báo trong nhà của mẫu đã được sắp xếp theo thứ tự tăng dần thì có 50% trường hợp nằm dưới giá trị 3 và 50% trường hợp nằm bên trên giá trị 3; và Mode là 3 tức là với mẫu của ta số người đọc báo trong gia đình thường gặp nhất là 3 người .

5. THỐNG KÊ MÔ TẢ VỚI THỦ TỤC EXPLORE

Các thủ tục thống kê mô tả kể trên chỉ hữu dụng cho việc tổng hợp một biến định lượng được đo lường đơn. Giả sử rằng bạn muốn tìm ra những khác biệt trong các đại lượng thống kê mô tả của biến *tuổi* giữa các nhóm đối tượng khác nhau về giới tính hoặc khu vực địa lý, bạn phải nhờ đến thủ tục Explore

Thủ tục Explore sẽ giúp bạn:

- Tính toán các đại lượng thống kê mô tả cho tất cả các trường hợp trong dữ liệu của bạn hoặc cho các nhóm con của chúng (các nhóm con là các nhóm được phân chia bởi các biểu hiện của biến định tính, ví dụ như biến *gtinh* tạo thành hai nhóm là nam và nữ, biến tp tạo thành hai nhóm là Hà nội và Tp HCM trong tình huống này giới tính được gọi là biến nhân tố)

- Nhận diện các giá trị khác biệt. Thủ tục này sẽ giúp bạn lược qua dữ liệu để phát hiện các giá trị bất thường, kiểm tra đối chiếu lại dữ liệu xem đó thực sự là những giá trị khác biệt do ngoại lệ hay thực ra là nhầm lẫn, sơ sót của quá trình thu thập dữ liệu hoặc khi nhập liệu.
- Tính toán các giá trị thập phân vị của phân phối của biến, cũng cho tất cả các trường hợp và cho các nhóm con của chúng.
- Tạo biểu đồ, hình dáng của biểu đồ sẽ cho bạn thấy dữ liệu phân phối như thế nào.

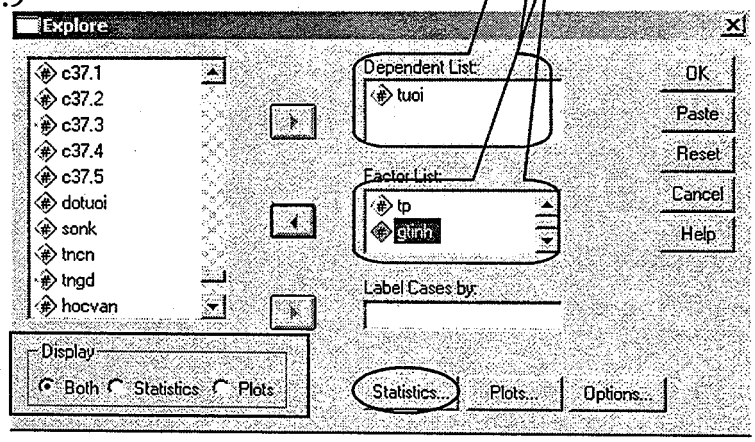
Thực hiện thủ tục Explore

1. Vào menu Analyze>DescriptiveStatistics>Explore mở hộp thoại Explore (chú ý bạn thực hiện lệnh này trên file *chay lenh Explore* chứ không phải file *Data thực hành*)

2. Chọn một hay nhiều biến (tất nhiên là dạng định lượng) trong danh sách biến bên tay trái mà bạn muốn so sánh các khác biệt trong đại lượng thống kê mô tả của chúng theo nhóm và đưa chúng sang khung Dependent List. Ở ví dụ này ta chọn biến *tuoi*.

3. Chọn một hay nhiều biến mà các bạn muốn sử dụng làm điều kiện để phân tách biến định lượng kia ra so sánh. Biến phân tách này phải là dạng Categorical với ít nhóm giá trị thì sự phân tách và so sánh của bạn mới có ý nghĩa. Ở đây ta chọn biến *gtinh*.

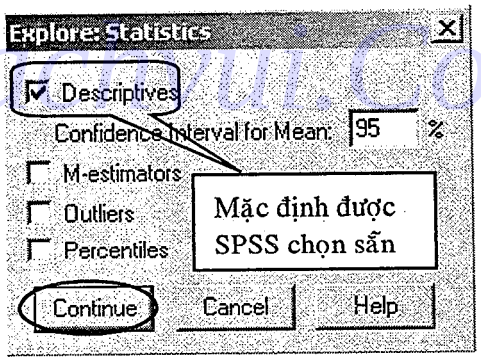
Hình 3.9



4. Nhấp chọn nút Statistics để vào hộp thoại Explore: Statistics. Trong hộp thoại này có những lựa chọn sau:

- Descriptive: mặc định được lựa chọn sẵn, do đó nếu bạn không mở hộp thoại Explore: Statistics để thực hiện bất kỳ lựa chọn nào thì SPSS cũng vẫn tính toán đủ các đại lượng thống kê mô tả cho bạn.
- M-estimators: các số thống kê tương đồng với số trung bình nhưng nó tạo ra những trọng số để cân bằng những quan sát phụ thuộc vào khoảng cách từ chúng đến một điểm trung tâm. Như vậy thực chất M-estimators cũng là một ước lượng cho khuynh hướng tập trung có phân biệt trọng số cho các giá trị khác nhau tuỳ theo vị trí của chúng, ví dụ các giá trị xa được gán trọng số thấp hơn các giá trị gần tâm. Như vậy khi số liệu của bạn có các điểm cực trị hay phân tán nhiều thì M-estimators cho ước lượng tốt hơn trung bình và trung vị.
- Outliers: SPSS thể hiện 5 giá trị lớn nhất và 5 giá trị nhỏ nhất của biến được đưa vào khung Dependent List, bạn có thể định vị những giá trị này vì SPSS chỉ ra cả vị trí của chúng theo hàng.
- Percentile: thể hiện các thập phân vị thứ 5, 10, 25, 50, 75, 90 và 95, trong đó các thập phân vị thứ 25, 50, 75 chính là tứ phân vị.

Hình 3.10

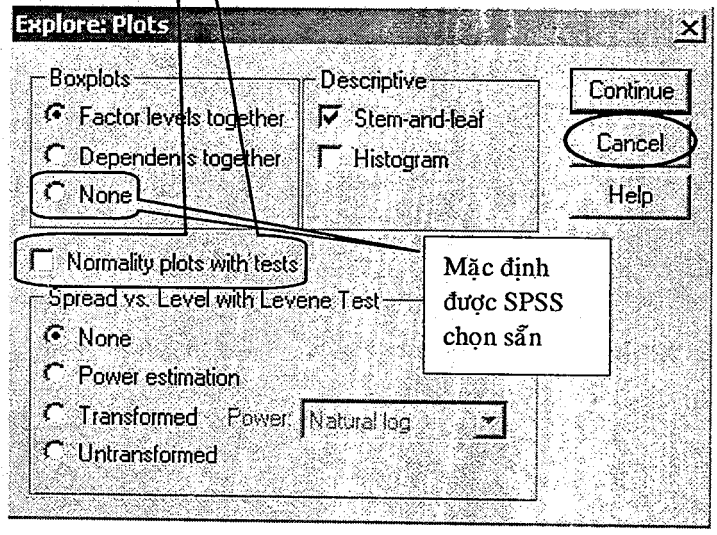


Nếu có mở Statistics thì sau khi hoàn tất việc lựa chọn bạn nhấp Continue để trở lại hộp thoại chính.

5. Nhấp chọn nút Plots để mở hộp thoại Explore: Plots (chú ý rằng để cho nút Plots hiện sáng ở tình trạng sẵn sàng sử dụng thì tại vùng Display ở góc dưới bên trái của hộp thoại chính Explore bạn phải đang lựa chọn Both hoặc Plots chứ không phải Statistics) Trong hộp thoại này bạn có thể lựa chọn các dạng biểu đồ sau: hộp (Boxplots); thân và lá (Stem-and-leaf); Histogram. Cụ thể:

- Khu vực Boxplots ở góc trên bên trái hộp thoại Explore: Plots cho phép bạn sắp xếp lại cách thể hiện các biểu đồ hộp Boxplot hoặc ra lệnh cho SPSS bỏ qua không thể hiện chúng (muốn vậy bạn nhấp vào None). Hai lựa chọn đầu tiên là hai tình huống hoán đổi cho nhau mà bạn sẽ cần đến khi bạn có hơn một biến được đưa vào khung Dependent List và có ít nhất một biến phân tách trong khung Factor List của hộp thoại chính Explore. Bạn hãy thử hai lựa chọn này và tự mình nhận ra sự khác biệt.
- Tại khu vực Descriptive, SPSS mặc định chọn cho bạn dạng biểu đồ thân và lá để mô tả phân phối của biến (bạn có thể chọn thêm dạng biểu đồ Histogram) vì biểu đồ thân và lá cho biết nhiều thông tin hơn biểu đồ Histogram, ngoài ra nó còn cho ta thấy hầu hết các giá trị gốc.
- Normality plots with tests : lựa chọn này sẽ vẽ cho bạn biểu đồ xác xuất chuẩn Q-Q plot giúp kiểm tra biến có phân phối chuẩn không (đây thường là một giả thuyết quan trọng trong nhiều phép kiểm định thống kê). Trong biểu đồ xác suất chuẩn, mỗi giá trị quan sát được vẽ dựa vào giá trị kỳ vọng của nó từ phân phối chuẩn. Trong Chương IX chúng ta sẽ trở lại nghiên cứu kỹ hơn về biểu đồ Q-Q plot.

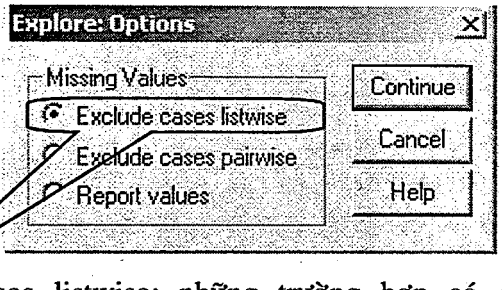
Hình 3.11



Nhấp Continue hoặc Cancel để trở lại hộp thoại trước, (vì nút Continue sẽ không sáng nếu trong hộp thoại này bạn chọn None cho Boxplots và nhấp bỏ lựa chọn mặc định ở biểu đồ thân và lá.)

6. Chọn tiếp nút Options... để lựa chọn cách thức thủ tục Explore xử lý các giá trị Missing. Bạn có các lựa chọn cơ bản là :

Hình 3.12



- Exclude cases listwise: những trường hợp có giá trị bị thiếu (missing) ở bất kỳ một biến nào trong các biến được đưa vào Dependent List hay Factor List ở hộp thoại Explore đều bị bỏ qua trong tất cả các phép tính toán và đồ thị. Đây là lựa chọn mặc định.
- Exclude cases pairwise: khi lựa chọn cách này, mỗi phép toán thống kê hay đồ thị sẽ sử dụng tất cả các trường hợp không có thông tin bị thiếu tại các biến cần cho việc tính toán chúng. Các trường hợp quan sát có giá trị bị thiếu tại một biến phụ thuộc này sẽ vẫn được sử dụng để tính toán các con số thống kê của biến phụ thuộc khác. Tùy chọn này khiến SPSS sử dụng tất cả các dữ liệu có thể dùng được trong các phép tính toán, nhưng các kết quả không hoàn toàn được tính ra trên cùng một tập dữ liệu như nhau.

7. Trở lại hộp thoại Explore, nhấp OK.

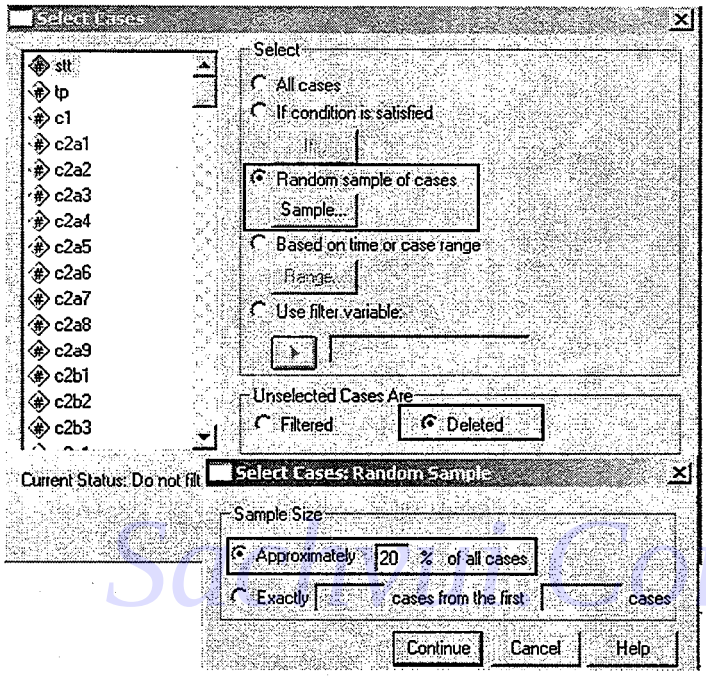
Để các bạn nhận định rõ ràng hơn về các kết quả của lệnh Explore (đặc biệt là các đồ thị) chúng ta nên tiến hành trên một tập dữ liệu không quá nhiều quan sát do đó chúng tôi sẽ chọn ngẫu nhiên 20% số quan sát của bộ dữ liệu Data thuc hanh và lưu lại thành file *chay lenh Explore.*

Trình tự tiến hành chọn 20% số quan sát được làm như sau (chú ý là bạn chỉ tham khảo để biết cách thực hiện chứ không cần chạy lại việc chọn 20% số quan sát ngẫu nhiên này vì chúng tôi đã làm rồi, vả lại bạn chạy lại cũng không thể có đúng các quan sát như trong file *chay lenh Explore* vì mỗi lần làm lệnh mà một tiến trình ngẫu nhiên khác nhau)

Trên cửa sổ lệnh Data/ Select Cases bạn chọn mục Random sample of cases rồi bấm vào nút Sample... mở tiếp cửa sổ thứ hai và tiến hành khai báo như hình (nhập con số 20 vào khung Aproximately...). Sau khi nhấp

Continue trở lại hộp thoại ngoài nhớ là chọn mục Deleted để SPSS xóa những quan sát không được chọn đi.

Hình 3.13



Với file *chay lenh Explore* chúng tôi lập bảng tần số cho giới tính để bạn biết cơ cấu của mẫu này như sau:

Bảng 3.6 **giới tính**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nam	59	50.9	50.9	50.9
	Nữ	57	49.1	49.1	100.0
	Total	116	100.0	100.0	

Sau đó trên file *chay lenh Explore* bạn thực hiện các hướng dẫn như đã nói sẽ được các kết quả sau:

Bảng 3.7 **Descriptives**

	giới tính			Statistic	Std. Error
tuổi	Nam	Mean		35.97	1.521
		95% Confidence Interval for Mean	Lower Bound	32.92	
			Upper Bound	39.01	
		5% Trimmed Mean		35.56	
		Median		32.00	
		Variance		136.447	
		Std. Deviation		11.681	
		Minimum		19	
		Maximum		59	
		Range		40	
		Interquartile Range		20.00	
		Skewness		.471	.311
		Kurtosis		-1.048	.613
	Nữ	Mean		33.67	1.486
		95% Confidence Interval for Mean	Lower Bound	30.69	
			Upper Bound	36.64	
		5% Trimmed Mean		33.16	
		Median		32.00	
		Variance		125.833	
		Std. Deviation		11.218	
		Minimum		18	
		Maximum		59	
		Range		41	
		Interquartile Range		18.50	
		Skewness		.518	.316
		Kurtosis		-.850	.623

tuổi Stem-and-Leaf Plot for
GTINH= Nam

```
 Frequency     Stem &  Leaf
     1.00        1 .  9
     9.00        2 .  112333344
    15.00        2 .  555666677888889
     5.00        3 .  00112
     6.00        3 .  568889
     7.00        4 .  0001234
     5.00        4 .  56678
     7.00        5 .  0012444
     4.00        5 .  5999
  Stem width:    10
  Each leaf:      1 case(s)
```

```
tuổi Stem-and-Leaf Plot for
GTINH= Nữ
 Frequency     Stem &  Leaf
     1.00        1  .  8
    16.00        2  .  0000112222444444
     9.00        2  .  555568899
     8.00        3  .  01223344
     4.00        3  .  6889
     7.00        4  .  0012233
     6.00        4  .  578889
     4.00        5  .  0003
     2.00        5  .  99
Stem width:     10
Each leaf:       1 case(s)
```

- Độ tuổi trung bình của nam giới theo mẫu là 35,97 tuổi còn của nữ giới là 33,67 tuổi. Như vậy trong mẫu của chúng ta các đối tượng nữ nói chung trẻ tuổi hơn nam.
- Khoảng ước lượng với độ tin cậy 95% về tuổi trung bình của tổng thể của nam và nữ lần lượt là (32,92 ; 39,01) và (30.69 ; 36,64)
- Mức độ biến thiên trong tuổi tác của nam lại ít hơn nữ, thể hiện ở độ lệch chuẩn của tuổi tác các đối tượng nam là 11.681 còn nữ là 11.218 tuổi.

Giải thích biểu đồ thân và lá: Trong biểu đồ thân và lá, độ rộng của thân là 10, như vậy các con số ở thân biểu diễn hàng chục và con số ở lá biểu diễn hàng đơn vị, mỗi lá ở trên biểu đồ đại diện cho 1 trường hợp, vì vậy chúng ta có thể đếm trên biểu đồ thân lá của nữ tại hàng đầu tiên có 1 trường hợp 18 tuổi (có 1 lá mang số 8).

6. LẬP BẢNG TỔNG HỢP NHIỀU BIẾN

Sau khi xem xét từng biến một, bước tiếp theo trong quá trình phân tích tập hợp dữ liệu là khảo sát mối liên hệ giữa các cặp kết hợp của các biến mà bạn quan tâm để giải quyết được vấn đề nghiên cứu. Kỹ thuật nào bạn sẽ chọn ở đây phụ thuộc vào tính chất của các biến. Phần tiếp đây sẽ trình bày một cách tổng quát quan hệ giữa các biến với nhau.

6.1. Bảng kết hợp các biến định tính

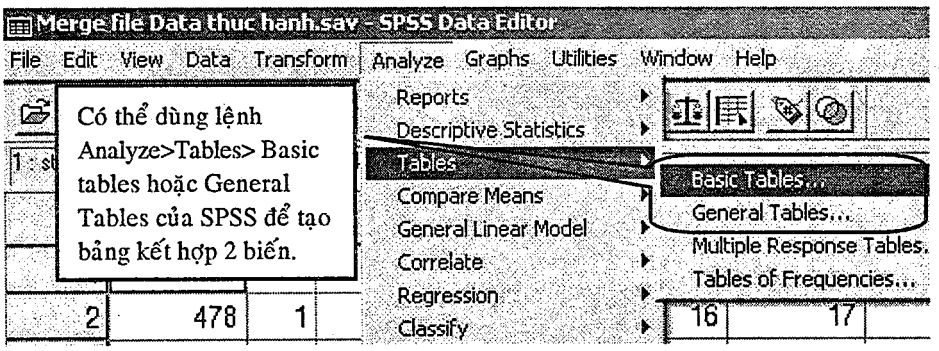
6.1.1 Bảng kết hợp 2 biến định tính

Khi yêu cầu về thông tin đòi hỏi ta phải xem xét tần số hay tần suất của các biểu hiện của một biến định tính theo sự phân loại của một

biến khác (ví dụ ta muốn biết số người trong độ tuổi thanh niên từ 18-25 tuổi của mẫu quan sát có bao nhiêu là nam và bao nhiêu là nữ) thì ta phải lập bảng phân tổ mà hàng là nhóm tuổi và cột là hai loại giới tính để phân tách dữ liệu.

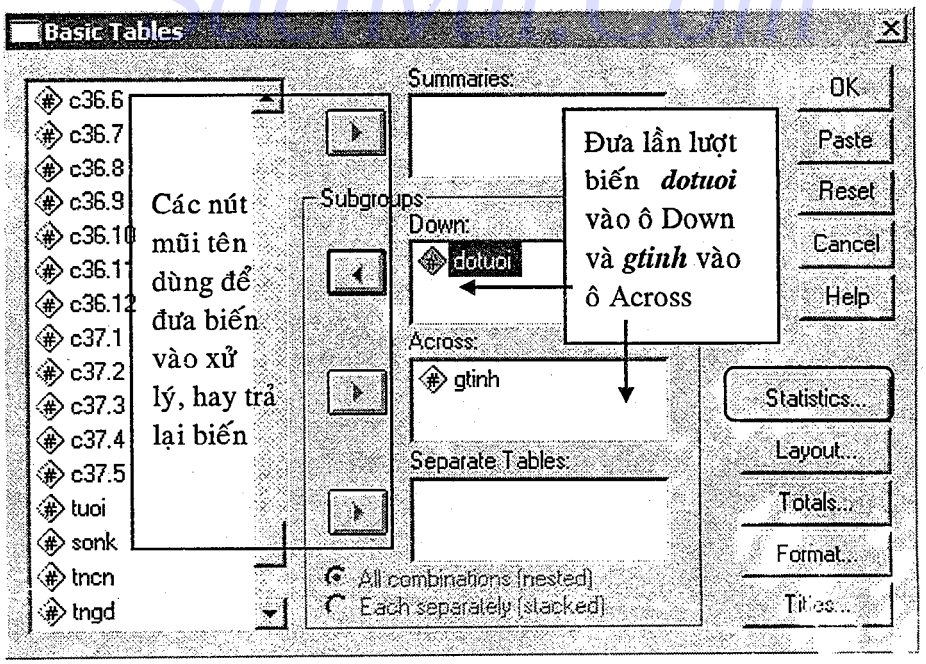
Để lập bảng kết hợp 2 biến định tính, có thể dùng lệnh Analyze > Tables > Basic Tables hoặc General Tables của SPSS.

Hình 3.14



6.1.1.1 Dùng lệnh bảng Basic: bạn sẽ chọn Basic Tables để mở hộp thoại Basic Tables

Hình 3.15



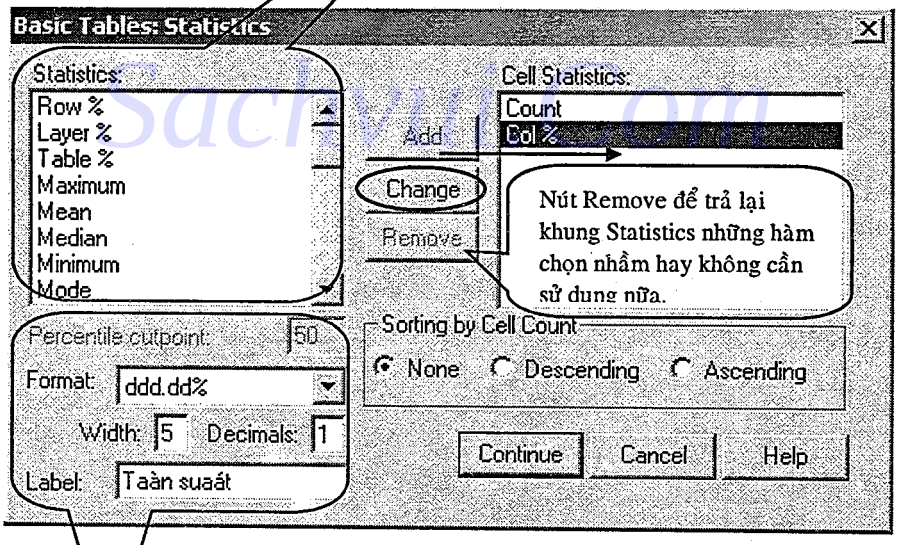
Các lựa chọn trên hộp thoại Basis Tables

- Down: ô chứa biến sẽ nằm trên dòng khi truy xuất ra bảng dữ liệu, ở đây là biến *dotuoi*, nó tạo nên các dòng của bảng.
- Across: ô chứa biến sẽ phân tách các cột, ở đây là biến *gtinh,* tạo nên các cột của bảng.
- Statistics: mở hộp thoại Basis Tables: Statistics chọn các đại lượng thống kê cần thiết và hiệu chỉnh cách định dạng số liệu

Chọn hàm thống kê

Bạn chọn các hàm thống kê trong ô Statistics bên tay trái. Đối với biến định tính, các hàm thường dùng là: Count (tần số), Row% (phần trăm theo dòng), Col % (phần trăm theo cột).

Hình 3.16



Trong ví dụ này, ta lần lượt chọn hàm Count và Hàm Col% rồi nhấp vào nút Add để đưa hàm đang chọn vào ô Cell Statistics bên tay phải.

Chỉnh định dạng số liệu kết quả của từng hàm thống kê đã chọn

Để chỉnh định dạng của các con số tính ra trong bảng, ta chọn tên hàm trong danh sách ô Cell Statistics bên tay phải bằng cách nhấp trỏ chuột vào tên hàm cho hiện dải sáng xanh, trạng thái định dạng của hàm này sẽ nổi đậm lên ở các khung tại vị trí góc dưới bên tay trái của hộp thoại. Ta có thể chỉnh sửa dạng số liệu (Format), số lượng số

thập phân (Decimals), và nhãn (Label) của số thống kê tính ra ở bảng kết quả.

Với ví dụ này ta chọn tên hàm Col%, và format thể hiện là có ký hiệu % phía sau con số.

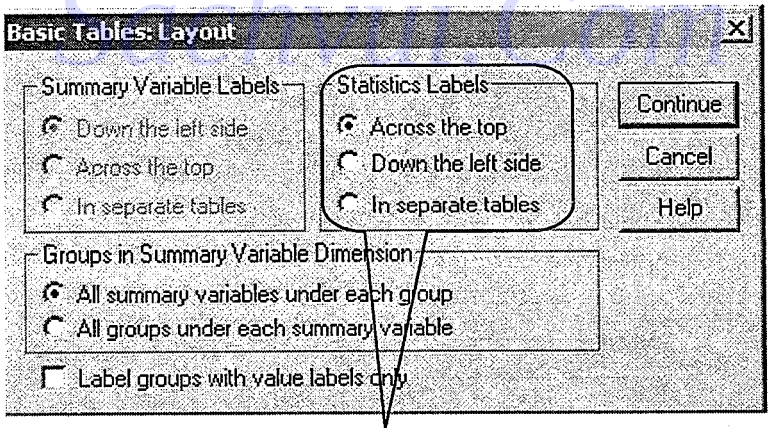
Đưa trỏ chuột đến ô Label, xóa hết Col%, điền nhãn mới là "Tần suất". Ngay lúc này "Tần suất" vẫn chưa thể hiện tiếng Việt đâu, bạn hãy chờ đến lúc bảng được SPSS xuất ra trên cửa sổ Output.

Sau khi chỉnh sửa định dạng, nhớ nhấp trỏ chuột vào nút Change ở giữa hộp thoại, nếu quên thao tác này thì mọi chỉnh sửa của bạn đều thành "công dã tràng". Sau đó nhấp vào nút Continue để trở về hộp thoại Basic Tables ban đầu.

- Layout: sắp xếp các đại lượng tính toán trong bảng số liệu

Trong trường hợp bảng kết quả có nhiều dòng nhiều cột, thì việc sắp xếp số liệu tính toán trở nên cần thiết và quan trọng. Để sắp xếp số liệu, nhấp chuột vào nút Layout trong hộp thoại Basic Tables, hộp thoại Layout xuất hiện như hình sau:

Hình 3.17



Có ba kiểu sắp xếp các số liệu tính ra trong bảng kết quả.

- Across the top: các đại lượng thống kê được sắp xếp theo cột. Kiểu sắp xếp này phù hợp khi bạn chọn hàm Col%. Mặc định SPSS lựa chọn cho bạn là sắp xếp theo cột, cho nên bạn có thể không cần phải chỉnh lại phần này.

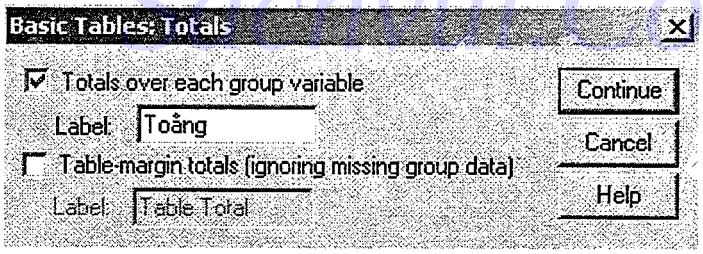
- Down the left side: các đại lượng thống kê được sắp xếp theo dòng. Kiểu sắp xếp này phù hợp khi chọn hàm tính toán là Row%.
- In separate tables: các đại lượng thống kê được sắp xếp trong các bảng riêng, mỗi bảng hiện thị kết quả tính theo một hàm thống kê. Kiểu sắp xếp này phù hợp khi bảng có quy mô lớn, nhiều dòng, nhiều cột tiện cho việc dàn trang in ấn sau này.

Sau khi chọn kiểu sắp xếp số liệu, nhấp nút Continue trở lại hộp thoại ban đầu.

- Totals: chọn tính các tổng dòng và tổng cột bằng cách nhấp nút Totals... trên hộp thoại Basic Tables để mở hộp thoại Totals

Trong hộp thoại này, chọn Total over each group variable để tính dòng cộng và cột cộng của bảng. Nếu cần bạn cũng có thể thay đổi tên của dòng cộng và cột cộng trong ô Label bằng cách xoá chữ Group Total đi và nhập chữ Việt có dấu vào đây.

Hình 3.18



Sau đó nhấp nút Continue trở về hộp thoại ban đầu, và cuối cùng là nhấp nút OK, bảng kết quả sau sẽ xuất hiện:

Bảng 3.8

		giới tính				Tổng	
		Nam		Nữ		Count	Tần suất
		Count	Tần suất	Count	Tần suất		
độ tuổi	18-25	58	23.3%	92	36.7%	150	30.0%
	26-35	71	28.5%	69	27.5%	140	28.0%
	36-45	68	27.3%	43	17.1%	111	22.2%
	46-60	52	20.9%	47	18.7%	99	19.8%
Tổng		249	100.0%	251	100.0%	500	100.0%

Bảng này cho thấy rõ cơ cấu mẫu điều tra về độ tuổi theo từng nhóm giới tính.

Ý nghĩa của các con số trong bảng kết quả

Trong mẫu có tổng cộng 150 người ở độ tuổi thanh niên từ 18-25 tuổi, chiếm 30%. Trong đó có 58 nam và 92 nữ. Số nữ trong độ tuổi thanh niên chiếm 36,7% tổng số nữ, tỷ lệ này cao hơn tỷ lệ nam trong độ tuổi thanh niên trên tổng số nam (23,3%).

Chú ý, ở hộp thoại Basis Tables: Statistics khi chọn hàm thống kê tính toán các chỉ tiêu, bạn lựa chọn phần trăm theo cột hay hàng là tuỳ vào vị trí sắp xếp biến bạn quan tâm nằm ở cột hay dòng và tuỳ thông tin bạn muốn tìm hiểu.

Cũng với ví dụ trên nếu bạn chọn Row% trong hộp thoại Basis Tables: Statistics, bạn sẽ biết được số nam giới trong độ tuổi thanh niên chiếm 38,7% tổng số người thuộc độ tuổi thanh niên của mẫu và nữ giới chiếm 61,3% (xem kết quả sự lựa chọn này ở Bảng 3.9)

Bảng 3.9

		giới tính				Tổng	
		Nam		Nữ		Count	% theo dòng
		Count	% theo hàng	Count	% theo hàng		
độ tuổi	18-25	58	38.7%	92	61.3%	150	100.0%
	26-35	71	50.7%	69	49.3%	140	100.0%
	36-45	68	61.3%	43	38.7%	111	100.0%
	46-60	52	52.5%	47	47.5%	99	100.0%
Tổng		249	49.8%	251	50.2%	500	100.0%

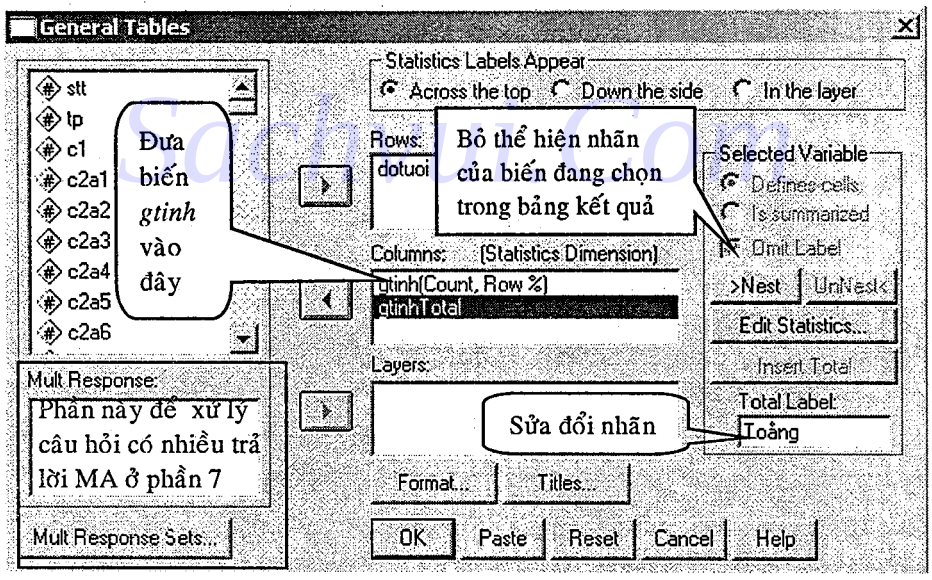
Trong bất kỳ tình huống lựa chọn nào bạn cũng cần nhất quán và tỉnh táo khi đọc số liệu trên bảng vì rất dễ nhầm lẫn. Bạn nhớ kiểm tra các số liệu tổng cộng và phần trăm cuối mỗi dòng và mỗi cột để có cơ sở đối chiếu.

6.1.1.2 Dùng lệnh bảng General: bạn chọn General Tables, hộp thoại sẽ mở ra như Hình 3.19, trong hộp thoại này, bạn sẽ thấy các đối tượng:

- Rows: ô chứa biến dòng (tương tư như ô Down trong Basic Tables)
- Columns: ô chứa biến cột (tương tự như ô Across trong Basic Tables).
- Edit Statistics: chọn hàm thống kê để tính toán và điều chỉnh định dạng số liệu tính ra (tương tự như bảng Basic Tables).

- Statistics Labels Appear: Sắp xếp các đại lượng thống kê tính ra trong bảng (tương tự phần Layout trong bảng Basic Tables).
- Omit Label: bỏ thể hiện nhãn của biến đang chọn trong bảng kết quả (Basic Table không có khả năng này).
- Insert Total: nhấp vào nút này để tính tổng dòng hay tổng cột hoặc cả hai tuỳ theo dạng hàm thống kê mà bạn đã lựa chọn trong Edit Statistics.
- Total Label: rê chuột vào dòng cột hay cột cộng vừa được lệnh Insert Total chèn vào khung Column (Statistic Dimention) và nhấp chọn đối tượng muốn đổi tên (hiện dải sáng xanh) thì nút Total Label sẽ nổi rõ ở chế độ sẵn sàng cho sử dụng, sửa đổi nhãn trong Total Label nếu bạn muốn.

Hình 3.19



Kết quả tính từ bảng General Table được trình bày dưới đây. Kết quả này giống hệt Bảng 3.9, điểm khác biệt là nhãn giới tính (trên tiêu đề cột) không cần thiết đã bị bỏ đi khiến bảng gọn hơn. Đây chính là điểm khác biệt duy nhất khi sử dụng bảng General hay bảng Basic để kết hợp hai biến định tính. Còn những công dụng khác của bảng General bạn sẽ nhận thấy ở phần xử lý biến nhiều chọn lựa.

Bảng 3.10

		Nam		Nữ		Tổng	
		Count	Hàng %	Count	Hàng %	Count	Hàng %
độ tuổi	18-25	58	38.7%	92	61.3%	150	100.0%
	26-35	71	50.7%	69	49.3%	140	100.0%
	36-45	68	61.3%	43	38.7%	111	100.0%
	46-60	52	52.5%	47	47.5%	99	100.0%

6.1.2 Bảng kết hợp 3 biến định tính

Khi bạn muốn biết cụ thể hơn về cơ cấu tuổi của những người được phỏng vấn phân tách theo nhóm giới tính tại từng thành phố thì bạn phải lập bảng phối hợp tới 3 biến.

Bạn cũng có thể lựa chọn dùng bảng Basic hoặc bảng General

6.1.2.1 Dùng lệnh bảng Basic:

Mở hộp thoại Basic Table, cách tiến hành cũng như trường hợp lập bảng Basic kết hợp 2 biến, nhưng bây giờ bạn sẽ lần lượt đưa cả 2 biến *tp* và *gtinh* vào ô Across, đưa biến *dotuoi* vào ô Down.

Khi bạn đưa 2 biến cột vào ô Across thì vì lúc này có tới 2 biến ở vị trí của cột nên phần dưới cùng của hộp thoại Basic Table sẽ nổi rõ lên để cho bạn xác định hai cách phối hợp biến với nhau:

1. All Combinations (nested): hai biến trong cùng ô sẽ phân nhóm lồng ghép trong nhau. Trong ví dụ này, các quan sát sẽ được phân thành 2 nhóm là Hà Nội và TPHCM. Sau đó các quan sát trong mỗi thành phố sẽ được phân chia tiếp theo thành hai nhóm nhỏ hơn là nam và nữ. Nếu ta đặt biến *gtinh* ở trên biến *tp* (bằng cách đưa *gtinh* vào khung Across trước *tp*) thì sự phân tách lồng ghép sẽ ngược lại.

2. Each separately (stacked): hai biến độc lập với nhau. Các quan sát sẽ lần lượt được phân chia vào 2 nhóm Hà Nội và TPHCM, sau đó lại được phân chia vào 2 nhóm là nam và nữ, lúc này bảng kết quả giống như sự ghép nối đơn giản giữa hai bảng Basic của 2 biến định tính : *dotuoi-gtinh* và *dotuoi-tp*

Các xác lập khác trong Statistics hoàn toàn tương tự như phần lập bảng 2 biến, nhưng bạn chỉ nên chọn1 trong 2 dạng hàm hoặc Count hoặc Col% chứ không nên chọn cùng lúc vì quy mô bảng lúc này khá lớn, còn nếu bạn chọn cùng lúc cả Count và Col% thì sau đó bạn phải vào nút Layout... của hộp thoại Basic Table chọn cách trình bày là In separate tables để bạn có thể lật từng lớp của bảng lên xem (xem hình minh hoạ ở dưới). Nếu bạn vẫn chọn các trình bày bảng Across the top thì bảng sẽ dàn hàng ngang tới 12 cột gây khó khăn trong việc xem toàn diện bảng hoặc in ra trên khổ A4.

Cũng có cách khác khắc phục khác cho bảng dàn ngang tới 12 hàng này là ta nhấp đôi chuột cho viền răng cưa xuất hiện quanh bảng sau đó vào menu Pivot chọn lệnh Move layers to Row thì bảng kết quả đầy đủ của cả hai hàm thống kê hiện ra theo trật tự y hệt nhau từ trên xuống dưới với 2 bảng nhỏ kết hợp với nhau theo chiều dọc mà bảng ở trên thể hiện tần số và bảng dưới thể hiện phần trăm (xem lại Lựa chọn cách thể hiện bảng kết quả ở phần 3).

Dưới đây là bảng kết quả khi bạn đã yêu cầu tính cả Count và Col% và chọn cách thể hiện In separate tables. Bạn sẽ xem các lớp của bảng kết quả bằng cách nhấp đôi chuột vào bảng cho xuất hiện viền răng cưa và nhấp chuột vào dấu mũi tên chọn lớp để thể hiện kết quả của lớp bạn muốn xem.

Bảng 3.11

Layers | Count

		thành phố					
		Hà Nội			TPHCM		
		giới tính		Tổng	giới tính		Tổng
		Nam	Nữ		Nam	Nữ	
độ tuổi	18-25	28	40	68	30	52	82
	26-35	33	39	72	38	30	68
	36-45	30	22	52	38	21	59
	46-60	27	31	58	25	16	41
Tổng		118	132	250	131	119	250

Hãy nhấp chuột vào dấu mũi tên chọn lớp ở khung Layer để xem lớp kế tiếp là bảng tần suất.

Bảng 3.12

		thành phố					
		Hà Nội			TPHCM		
		giới tính		Tổng	giới tính		Tổng
		Nam	Nữ		Nam	Nữ	
độ tuổi	18-25	23.7%	30.3%	27.2%	22.9%	43.7%	32.8%
	26-35	28.0%	29.5%	28.8%	29.0%	25.2%	27.2%
	36-45	25.4%	16.7%	20.8%	29.0%	17.6%	23.6%
	46-60	22.9%	23.5%	23.2%	19.1%	13.4%	16.4%
Tổng		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Nếu chọn menu Pivot > Move Layers to Row bạn có tiếp Bảng 3.13. Bạn kiểm tra thử xem Bảng 3.13 có phải là sự kết hợp đơn thuần giữa 2 bảng Count và Col% hay không? Thử xem số lượng nam giới trong mẫu tại Hà Nội là 28 người chiếm bao nhiêu % trong tổng số nam trả lời tại Hà Nội và tại Tp HCM là 30 người chiếm bao nhiêu % trong tổng số nam trả lời tại thành phố HCM? Có phải là 23,7% và 22,9%. Bạn nhớ cẩn thận và chú tâm khi đọc các % theo cột và hàng trên các bảng kết quả mà SPSS tính được.

Bảng 3.13

			thành phố					
			Hà Nội			TPHCM		
			giới tính			giới tính		
			Nam	Nữ	Tổng	Nam	Nữ	Tổng
n	độ tuổi	18-25	28	40	68	30	52	82
		26-35	33	39	72	38	30	68
		36-45	30	22	52	38	21	59
		46-60	27	31	58	25	16	41
	Tổng		118	132	250	131	119	250
Cột %	độ tuổi	18-25	23.7%	30.3%	27.2%	22.9%	43.7%	32.8%
		26-35	28.0%	29.5%	28.8%	29.0%	25.2%	27.2%
		36-45	25.4%	16.7%	20.8%	29.0%	17.6%	23.6%
		46-60	22.9%	23.5%	23.2%	19.1%	13.4%	16.4%
	Tổng		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

6.1.2.2 Dùng Bảng General

Tiến hành lại ví dụ trước, lần này ta dùng bảng General Tables với thao tác như sau:

1. Statistics Labels Appear: chọn In the layers để tách kết quả thành các bảng rời nhau.

2. Đưa *dotuoi* vào khung Rows

3. Lần lượt đưa *tp* và *gtinh* vào khung Columns, do hai biến cột này không ngang cấp nhau nên với hộp thoại General ta phải chọn sáng *gtinh* rồi bấm vào nút >Nest để báo cho SPSS đưa *gtinh* vào phân tách trong *tp*.

4. Chọn lựa hàm thống kê cho *gtinh*, bạn hãy chọn Count và Col%

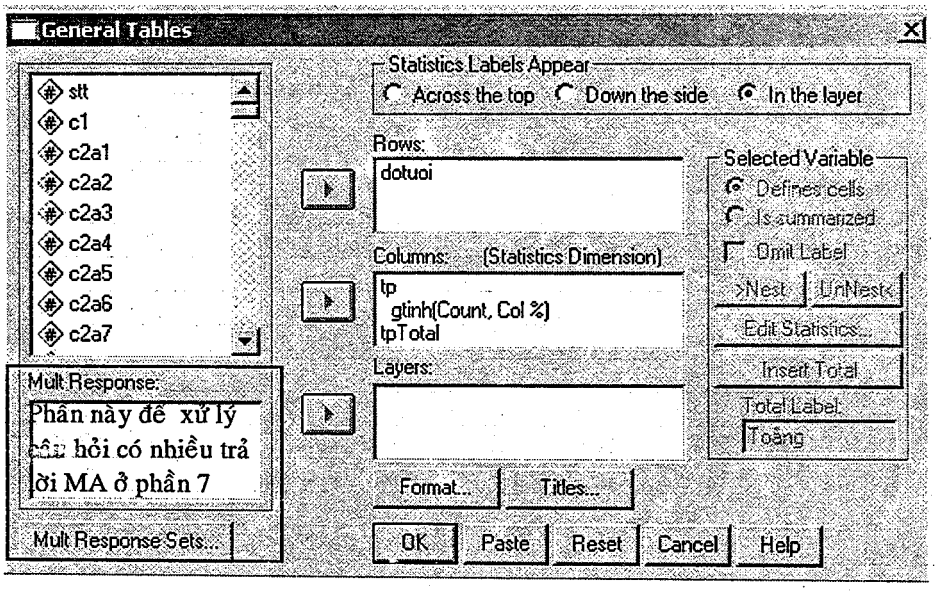
Sau đó lại chọn sáng đối tượng *tp* trong khung Column rồi nhấp nút Insert Total để tính tổng cộng theo hàng, chú ý phải cho *gtinh* thụt vào trước việc chọn Total cho *tp*.

5. Chọn Omit Lable cho cả 2 biến *tp* và *gtinh* để bỏ nhãn đi.

6. Sau cùng là nhấp nút OK

Bạn có thể thấy các lựa chọn cần thiết như trong Hình 3.20 dưới đây và kết quả đầu tiên được SPSS tạo ra ở Bảng 3.14 cho Layer đầu tiên là bảng tính tần số (Count)

Hình 3.20



Bảng 3. 14 Thể hiện kết quả của hàm Count.

		thành phố				
		Hà Nội		TPHCM		
		giới tính		giới tính		
		Nam	Nữ	Nam	Nữ	Tổng
độ tuổi	18-25	28	40	30	52	150
	26-35	33	39	38	30	140
	36-45	30	22	38	21	111
	46-60	27	31	25	16	99

Nhấp đôi chuột vào bảng, vào menu Pivot tại cửa sổ Chart Editor, chọn Move Layers to Rows, bảng Col% sẽ xuất hiện kế tiếp dưới bảng Count. Nhớ đừng chọn Move Layers to Columns vì bảng kết quả sẽ dàn ngang tới 12 cột.

Còn có một cách khác để tạo bảng kết hợp 3 biến định tính là dùng lệnh Crosstable tạo bảng phân tổ kết hợp cho các biến định tính, bảng này còn sử dụng trong kiểm định Chi – Square mà bạn sẽ gặp ở chương sau.

6.2. Bảng kết hợp biến định tính với biến định lượng

6.2.1 Bảng kết hợp 1 biến định tính và 1 biến định lượng.

Cũng với ý tưởng tương tự khi lập bảng kết hợp 2 biến định tính, nhưng ở đây ta không chỉ đếm tần số mà còn tính số trung bình cùng các đại lượng thống kê mô tả khác của một biến định lượng theo sự phân loại của một biến định tính, ví dụ ta cần tính toán số người trung bình thường xuyên đọc báo trong nhà ở từng khu vực Hà Nội hoặc Tp HCM và xét chung cả hai thành phố. Khi đó ta phải lập bảng kết hợp 1 biến định tính và 1 biến định lượng.

Để lập bảng kết hợp 1 biến định lượng và 1 biến định tính, bạn dùng lệnh Analyze > Tables > Basic Tables của SPSS

1. Mở bảng Basic, đưa biến *tp* vào khung Across; biến *c3* vào ô Summaries

Bạn thấy điểm khác biệt chưa, tại sao biến dòng không được đưa vào khung Down mà lại đưa vào Summaries? Vì Summaries chứa các biến định lượng cần tính toán. Bạn không được đưa các biến

định tính vào ô này, vì các kết quả tính ra sẽ không có ý nghĩa (dữ liệu định tính không tóm tắt được bằng cách tính giá trị trung bình). Có thể đưa nhiều biến định lượng có liên quan (trong cùng một câu hỏi) vào để xử lý cùng một lúc.

2. Nhấp nút Stastistics để chọn dạng hàm thống kê. Chọn hàm Mean để tính trung bình, ngoài ra có thể chọn các hàm khác để tính độ lệch chuẩn, Mode, số quan sát có thông tin trả lời (Valid Value Count)... Trong ví dụ này, ta chọn những hàm là Valid Value Count, Mean, Mode, Std. Devitation.

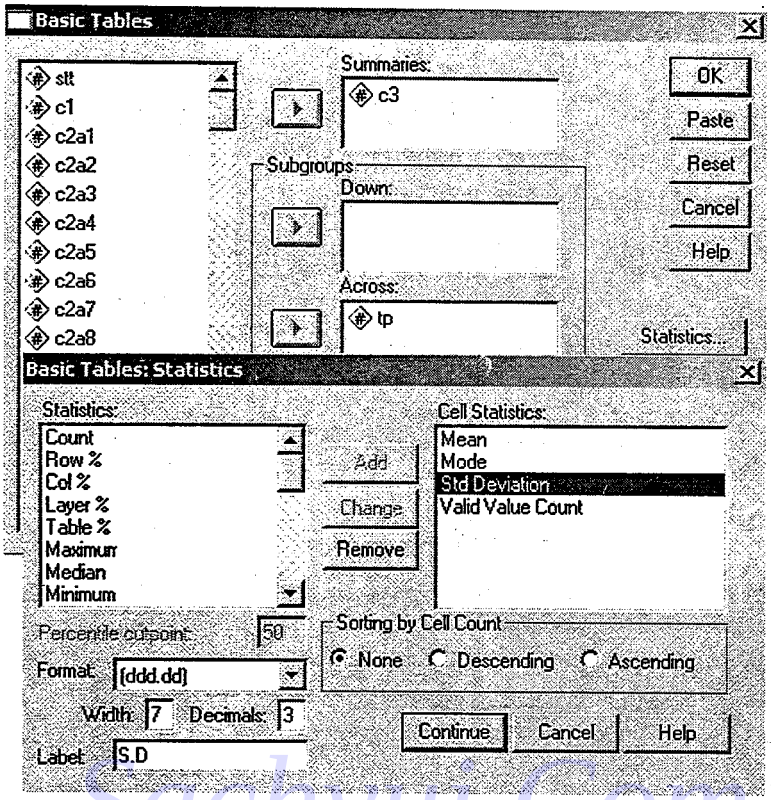
Nếu cần bạn có thể chỉnh lại định dạng của các số liệu tính ra từ các hàm số đã chọn tương tự như phần trước. Giả dụ với đại lượng Mean, nếu bạn không chỉnh lại Decimals là một số khác 0 thì giá trị trung bình sẽ được biểu diễn ở dạng một con số tròn trịa, ví dụ trung bình số người đọc báo trong gia đình tại Hà Nội và Tp HCM trong ví dụ của chúng ta sẽ lần lượt là 3 và 4. Bạn hãy so sánh chúng với kết quả tính ở Bảng 3.15 khi khai báo Decimals =3, thông tin khác biệt đáng kể phải không?

3. Sau khi chỉnh sửa định dạng, nhấp nút Continue, rồi OK

Nếu muốn biết các thông tin về TB; S.D; Mode...của biến *c3* tại cả hai thành phố tức toàn bộ mẫu quan sát thì bạn nhấp thêm nút Totals... của hộp thoại Basic Tables. Phần kết quả bạn nhận được thêm sẽ tương tự như khi bạn dùng lệnh Descriptive để tính các đại lượng thống kê mô tả của biến *c3*.

Kết quả Bảng 3.15 cho thấy trung bình số người đọc báo trong gia đình tại Hà Nội thấp hơn tại Tp HCM (3,116 < 3,832), nhưng số người đọc báo trong gia đình tại Hà Nội lại ít biến động hơn tại Tp HCM thể hiện qua độ lệch chuẩn tại Hà Nội nhỏ hơn tại Tp HCM.

Hình 3.21



Bảng 3.15 Bảng kết hợp 1 biến định lượng và 1 biến định tính

			số người đọc báo trong GĐ
thành phố	Hà Nội	TB	(3.116)
		Mode	2
		S.D	(1.289)
		Valid N	N=250
	TPHCM	TB	(3.832)
		Mode	3
		S.D	(2.139)
		Valid N	N=250

6.2.2 Bảng kết hợp 2 biến định tính và 1 biến định lượng

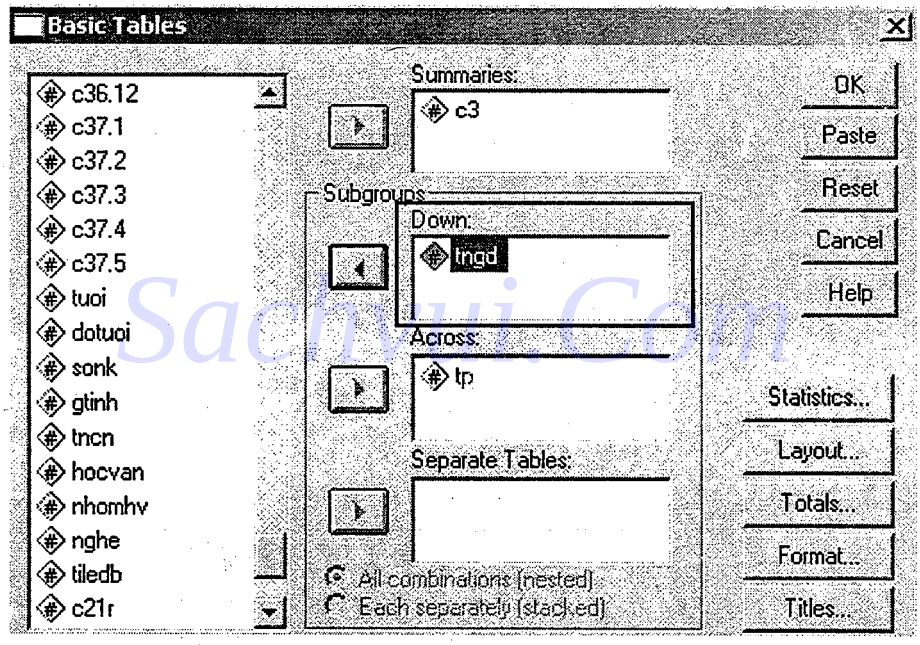
Giờ đây yêu cầu của bạn lại tăng thêm 1 bậc nữa vì bạn muốn biết số lượng người xem báo trong nhà (là 1 biến định lượng) tại từng thành phố được chi tiết theo từng nhóm thu nhập hộ gia đình. Vậy là ở đây bạn cần tới 2 biến định tính là *tp* và *tngd*.

Công cụ của bạn để giải quyết vấn đề này vẫn là bảng Basic, bạn mở lại bảng Basic Table, đưa thêm biến *tngd* vào ô Down, vẫn để biến *c3* trong ô Summaries và *tp* trong ô Across.

Thực hiện một số lựa chọn cần thiết tại Statistics như Mean, Valid Value Count, định dạng thể hiện của các đại lượng này, chọn nút Totals... và đổi Label của nó thành "Tổng".

Sau khi OK, kết quả sẽ xuất hiện như trong Bảng 3.16.

Hình 3.22



Bảng 3.16

		thành phố				Tổng	
		Hà Nội		TPHCM			
		TB	Số QS có thông tin trả lời	TB	Số QS có thông tin trả lời	TB	Số QS có thông tin trả lời
TN hộ TB tháng	Dưới 2trđ	2.74	N=68	3.41	N=46	3.01	N=114
	2-4 trđ	3.26	N=136	3.44	N=100	3.34	N=236
	4-6 trđ	2.91	N=34	3.77	N=69	3.49	N=103
	6-10 trđ	3.78	N=9	4.96	N=28	4.68	N=37
	Trên 10 trđ	5.33	N=3	8.29	N=7	7.40	N=10
Tổng		3.12	N=250	3.83	N=250	3.47	N=500

Trong bảng kết quả này, các số liệu thể hiện số người quan sát và số lượng người đọc báo trong gia đình tại từng thành phố và chi tiết cho từng nhóm thu nhập.

Ví dụ, tại 2 cột của thành phố Hà Nội thuộc dòng thứ nhất (thu nhập dưới 2 trđ), có 68 quan sát và số lượng người đọc báo trung bình trong các hộ này là 2,74 người. Với nhóm có thu nhập hộ gia đình trên 10 triệu/tháng bạn gặp 3 hộ tại Hà Nội với số người đọc báo trung bình trong nhà là 5,33 người và 7 hộ tại Tp HCM với số người đọc báo trung bình trong nhà là 8,29 người. Bạn có thể tính thêm số nhân khẩu trung bình tại từng ô và so sánh số người đọc báo với số nhân khẩu trong hộ.

6.3. Bảng tần số phức tạp với Tables of Frequencies

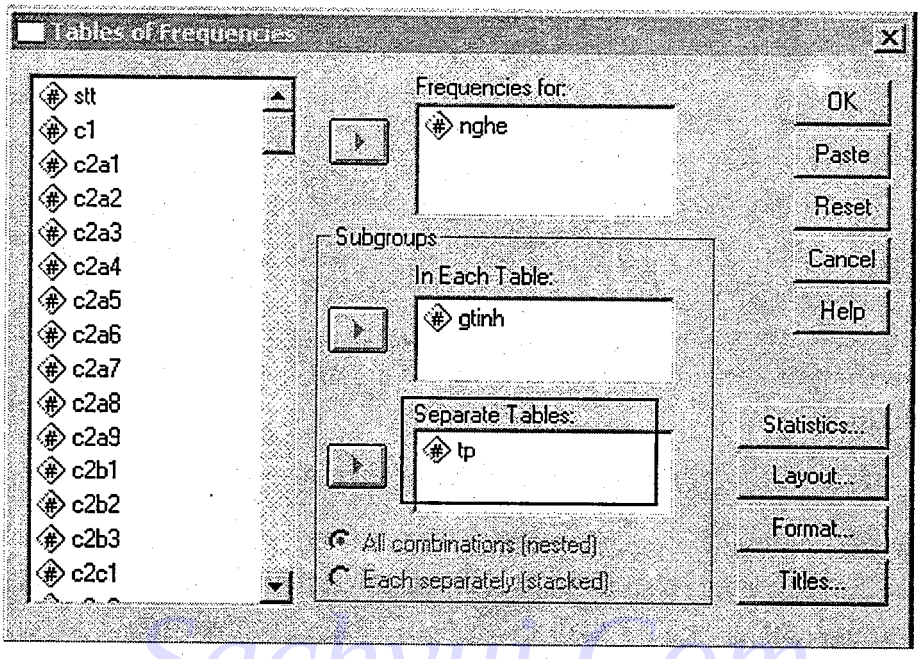
Thủ tục Tables of Frequencies rất hữu ích trong trường hợp khi bạn quan tâm đến việc so sánh hoặc quan sát đồng thời tần số và tần suất của một biến dạng định lượng (có ít trị số) hoặc định tính theo sự phân tách của một biến khác, ví dụ bạn muốn đếm số người cùng làm nghề giáo viên ở cả hai giới tính trong mẫu của bạn, hoặc chi tiết hơn nữa là số người cùng làm nghề giáo viên ở cả hai giới tính tính riêng tại Hà Nội hoặc tại TPHCM.

1. Chọn menu Analyze>Tables>Tables of Frequencies mở hộp thoại dưới Hình 3.23
2. Đưa biến *nghe* vào khung Frequencies for và biến phân tách *gtinh* vào In Each Table.

> Nếu bạn quan tâm sâu hơn đến việc quan sát các thông tin này tại hai khu vực thì đưa biến tp vào Separate Tables. Tên gọi đã có thể giúp cho bạn hình dung kết quả rồi, khi đó bạn sẽ có 2 bảng riêng biệt với kết cấu y hệt nhau cho mỗi khu vực riêng. Công dụng của Separate Tables cũng y hệt như khi bạn tiến hành lệnh Split File với biến *tp*, bạn sẽ gặp lệnh Split File ở mục 7 của chương này.

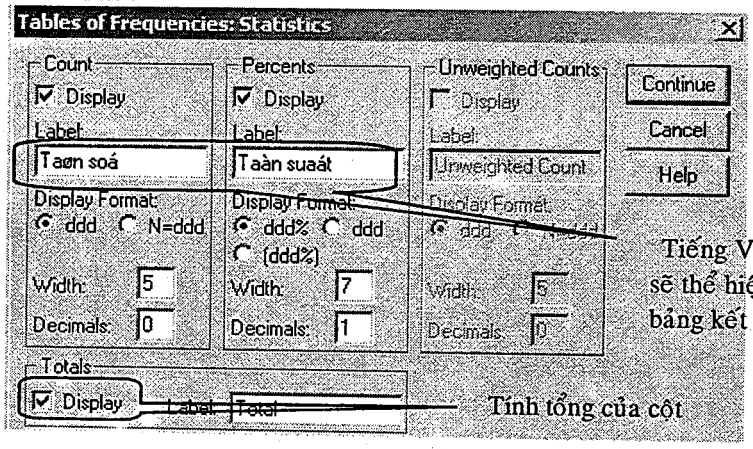
3. Vào Statistics chọn Count và Percent để tính tần số và tần suất và đổi nhãn của chúng nếu bạn muốn. Xem minh hoạ ở Hình 3.24
4. Chọn Display ở dưới cùng hộp thoại để SPSS thể hiện Tổng của mỗi cột.
5. Nhấp Continue, và cuối cùng là OK.

Hình 3.23



Bạn có kết quả đầu tiên như Bảng 3.17 trang dưới đây, bạn có thể chọn xem bảng tần số và tần suất riêng của từng thành phố bằng nút mũi tên ở khung Layers hoặc xem đồng thời cả 2 bảng nhờ menu Pivot > Move Layers to Rows.

Hình 3.24



Bảng 3.17

Layers: thành phố Hà Nội (thành phố Hà Nội / thành phố TPHCM)

	giới tính			
			Nữ	
	Nghề nghiệp		Nghề nghiệp	
	Tần số	Tần suất	Tần số	Tần suất
Công chức	11	4.4%	12	4.8%
Giáo viên	4	1.6%	5	2.0%
NVVP	13	5.2%	11	4.4%
Chủ DN	1	.4%	1	.4%
NV công ty KD	11	4.4%	12	4.8%
Tự KD SP-DV	16	6.4%	16	6.4%
Buôn bán nhỏ	10	4.0%	21	8.4%
CN có tay nghề	12	4.8%	8	3.2%
LĐPT	4	1.6%	3	1.2%
SVHS	15	6.0%	15	6.0%
Về hưu	10	4.0%	9	3.6%
Không LV	1	.4%	7	2.8%
Nghề chuyên môn	4	1.6%	4	1.6%
Nghề khác	6	2.4%	8	3.2%
Total	118	47.4%	132	52.6%

7. XỬ LÝ CÂU HỎI CÓ THỂ CHỌN NHIỀU TRẢ LỜI – Multiple Answer (MA)

Trong thực tế, người nghiên cứu thường dùng các câu hỏi cho phép người trả lời có thể chọn nhiều hơn 1 lựa chọn. Những câu hỏi về nhận biết nhãn hiệu, nhãn hiệu sử dụng, nhãn hiệu đã mua, nhãn hiệu nào có quảng cáo trên TV tối hôm qua ... Ví dụ trong câu hỏi về các loại báo thường đọc của *Data thuc hanh*, người được hỏi có thể nhắc đến nhiều loại báo khác nhau. Lúc đó một biến đơn không thể chứa hết thông tin mà người trả lời cung cấp trong câu hỏi này, vì vậy chúng ta phải tạo nhiều biến cho chỉ 1 ý hỏi.

Trong trường hợp này ta có thể dùng bảng General Tables để gộp các biến trong cùng 1 câu MA lại với nhau nhằm tổng hợp thông tin về câu hỏi ta quan tâm.

Cách thực hiện

1. Từ menu chọn Analyze mở hộp thoại General Table, ở góc trái bên dưới hộp thoại này có một khu vực chúng ta vẫn chưa tìm hiểu ở phần trước, hãy nhấn nút Mult Response Set... trong khu vực này, một hộp thoại kế tiếp sẽ xuất hiện như Hình 3.25 để ta khai báo tập hợp các biến và cách mã hoá các biến đó. Chúng ta sẽ thực hiện ví dụ ghép các biến con của câu 2a là loại báo thường đọc (từ *c2a1-c2a9*)

2. Trong hộp thoại khai báo tập hợp biến của câu hỏi nhiều trả lời: Chọn các biến của câu hỏi 2a - báo thường đọc (biến *c2a1* đến *c2a9*) tại khung Set Definition rồi đưa các biến này vào ô Variables in Set.

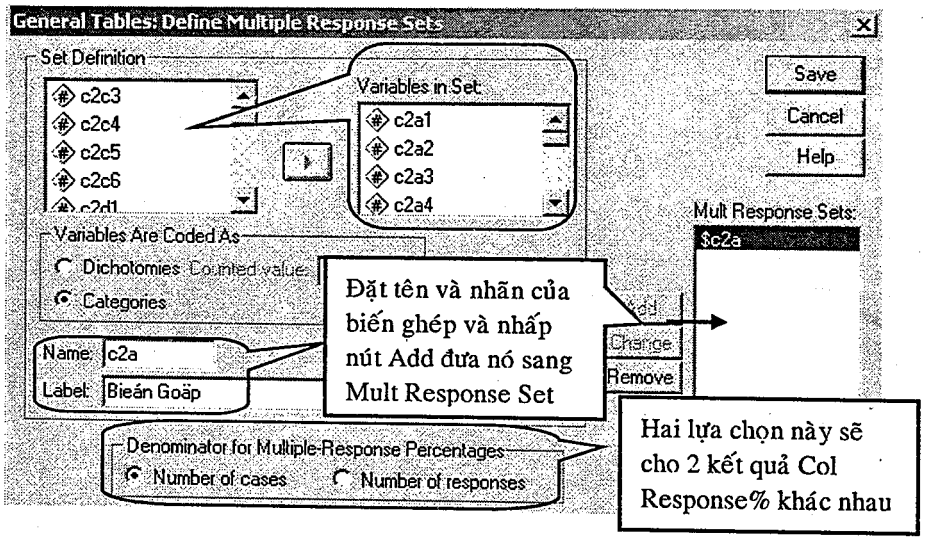
3. Khai báo cách mã hóa các biến

- Chọn Dichotomies nếu biến có 2 biểu hiện (nam hay nữ, mua hay không mua, nhớ hay không nhớ có quảng cáo, có mua hay không mua)
- Chọn Categories nếu biến có nhiều biểu hiện (trong ví dụ này ta chọn Categories).

4. Đặt tên và nhãn của tập biến ghép (biến tổng hợp), ở đây ta đặt *c2a*

5. Nhấn nút Add để xác nhận, biến tổng hợp sẽ được cập nhật vào danh sách ô bên phía phải (Mult Response Set).

Hình 3.25



Sau khi định nghĩa biến *c2a*, ta có thể định nghĩa hàng loạt các biến tương tự khác.

6. Sau khi định nghĩa các biến ghép ứng với các câu MA, nhớ nhấn nút Save để lưu biến vừa ghép và trở lại hộp thoại ban đầu General Tables.

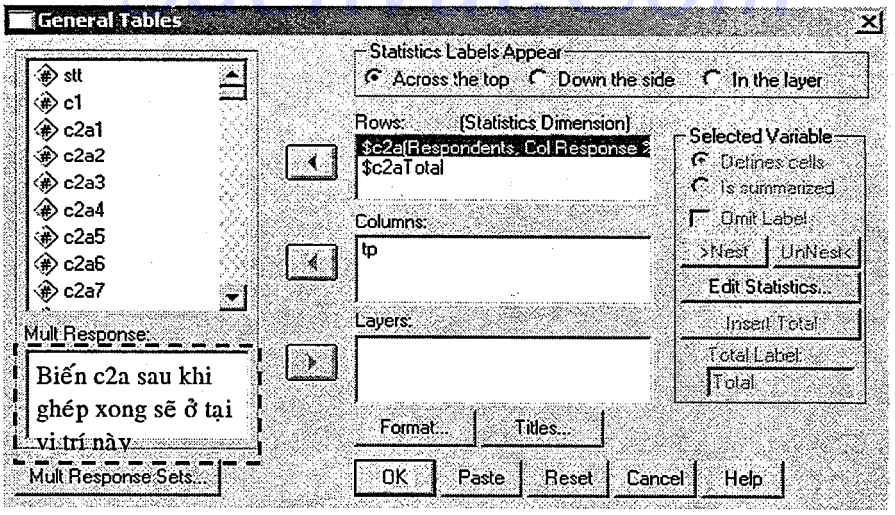
Tạo bảng từ biến vừa ghép.

1. Trong hộp thoại General Tables ta đưa biến ghép đã định nghĩa *c2a* (lúc này nằm tại khung Mul Response) vào ô dòng theo cách thông thường và đưa biến *tp* vào ô cột để xem các tờ báo thường đọc ở từng thành phố.

2. Trong phần Statistics, ta chọn 2 hàm thống kê là Response (số người trả lời) và Col Response % (% người trả lời theo cột). Có thể chỉnh sửa định dạng theo yêu cầu. Nhấp nút Continue để trở về hộp thoại General Tables ban đầu.

3. Sau khi trở về hộp thoại General Tables, chọn sáng tên biến *c2a* rồi chọn nút Insert Total, sao đó nhấp nút OK, kết quả xuất hiện như sau:

Hình 3.26



Bảng 3.18

		thành phố			
		Hà Nội		TPHCM	
		Cases	Col Response %	Cases	Col Response %
Biến Gộp	HN mới	121	48.4%	1	.4%
	SGGP	11	4.4%	59	23.6%
	Lao Động	65	26.0%	14	5.6%
	Người Lao Động	12	4.8%	62	24.8%
	Tiền Phong	87	34.8%	5	2.0%
	Thanh Niên	28	11.2%	57	22.8%
	Tuổi Trẻ	44	17.6%	197	78.8%
	Phụ Nữ VN	97	38.8%	23	9.2%
	Phụ Nữ TPHCM	5	2.0%	77	30.8%
	Thời Báo KTVN	29	11.6%	6	2.4%
	Thời Báo KTSG	10	4.0%	11	4.4%
	SG Tiếp Thị	58	23.2%	101	40.4%
	Thế Giới Phụ Nữ	80	32.0%	84	33.6%
	Tiếp Thị và GĐ	28	11.2%	32	12.8%
	Mua & Bán	47	18.8%	11	4.4%
	An Ninh Thế Giới	193	77.2%	106	42.4%
	An Ninh Thủ Đô	183	73.2%	8	3.2%
	Công An TPHCM	44	17.6%	188	75.2%
	Khác	124	49.6%	74	29.6%
Tổng		250	506.4%	250	446.4

Hãy thử lại ví dụ này nhưng chọn sáng tên biến *tp* và nhấp Insert Total để xem sự khác biệt.

Ý nghĩa của các số liệu trong Bảng 3.18

Tại hàng đầu tiên thuộc phân nhóm Hà Nội, cột Case cho biết khi lần lượt hỏi 250 người tại HN có 121 lần gặp câu trả lời có đọc báo Hà Nội mới (Chú ý là đồng thời một người có thể đọc nhiều báo nên họ chọn nhiều hơn 1 trả lời do đó tổng cột case sẽ không thể = 250)

Cột Col Response % cho ta biết mức độ thường đọc các loại báo, tỷ lệ đọc HN mới là 48,4%, tỷ lệ này được tính trên 250 người được phỏng vấn tại HN, tức là lấy 121 người trả lời có đọc HN Mới chia cho 250.

Đi dọc xuống cột Col Response% của Hà Nội cho ta biết loại báo thường được đọc nhất là An ninh thế giới và An ninh thủ đô. Còn tại Tp HCM báo Tuổi Trẻ được đọc nhiều nhất (78,8%).

Chú ý nếu bên hộp thoại General Tables: Define Multiple Response sets, ở khu vực dưới cùng của hộp thoại ta chọn Number of Response thì khi thực hiện đổ bảng kết quả cột Col Response% sẽ cho thông tin khác. Hãy so sánh cột Col Response% của Bảng 3.18 và 3.19. Phần trăm của cột Col Response% của Hà Nội bây giờ được tính trên mẫu số là tổng của tất cả các trả lời tại Hà Nội (bạn kiểm tra lại bằng máy tính tay xem có phải là 1266 case không?), nó cho biết mỗi loại báo chiếm tỷ lệ bao nhiêu trong tổng số lựa chọn về các loại báo thường đọc tại Hà Nội. Vì thế Col Response% của Hà Nội mới là 9,6%. Hàng cuối cùng Total trong cả hai tình huống ta đã lựa chọn đều là 250 với cột Case nhưng có khác nhau với cột Col Response. Thông tin chúng cung cấp cho ta không có nhiều giá trị?

Bạn có nhớ tại phần chuyển hoá biến chúng ta còn 1 việc chưa làm xong với biến *docTTre* không? Giờ đây là lúc hoàn thành việc đó. Ta sẽ đổ bảng Frequency của biến *docTTre* cho từng thành phố riêng biệt bằng lệnh Frequencies sau đó so sánh với kết quả tại Bảng 3.18

Bảng 3.19

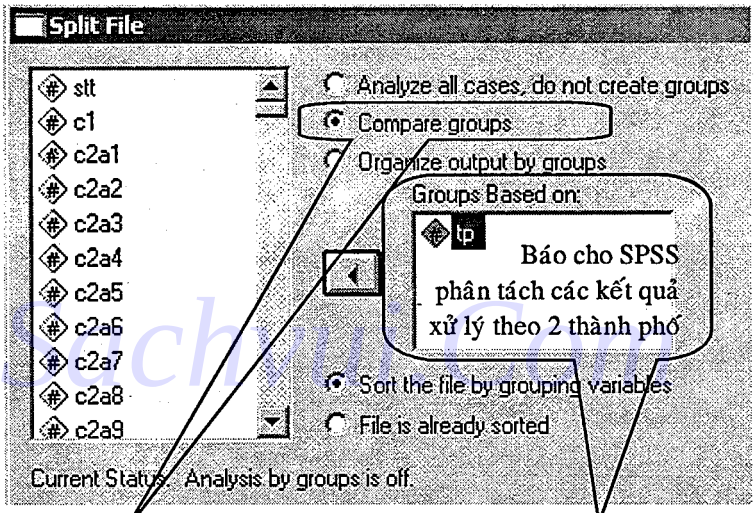
		thành phố			
		Hà Nội		TPHCM	
		Cases	Col Response %	Cases	Col Response %
Bien gop	HN mới	121	9.6%	1	.1%
	SGGP	11	.9%	59	5.3%
	Lao Động	65	5.1%	14	1.3%
	Người Lao Động	12	.9%	62	5.6%
	Tiền Phong	87	6.9%	5	.4%
	Thanh Niên	28	2.2%	57	5.1%
	Tuổi Trẻ	44	3.5%	197	17.7%
	Phụ Nữ VN	97	7.7%	23	2.1%
	Phụ Nữ TPHCM	5	.4%	77	6.9%
	Thời Báo KTVN	29	2.3%	6	.5%
	Thời Báo KTSG	10	.8%	11	1.0%
	SG Tiếp Thị	58	4.6%	101	9.1%
	Thế Giới Phụ Nữ	80	6.3%	84	7.5%
	Tiếp Thị và GĐ	28	2.2%	32	2.9%
	Mua & Bán	47	3.7%	11	1.0%
	An Ninh Thế Giới	193	15.2%	106	9.5%
	An Ninh Thủ Đô	183	14.5%	8	.7%
	Công An TPHCM	44	3.5%	188	16.8%
	Khác	124	9.8%	74	6.6%
Total		250	100.0%	250	100.0%

Trình tự thực hiện:

1. Trước tiên bạn yêu cầu SPSS phân tách kết quả xử lý cho từng thành phố bằng lệnh Split File. Với lệnh Split File, mọi kết quả xử lý của bạn sau đó đều được SPSS chia theo chủ đề bạn đã chọn để tách, ở đây là thành phố.

Vào menu Data> Split File, hộp thoại Split File sẽ trông như dưới đây

Hình 3.27



Chọn Compare groups, đưa biến *tp* sang khung Group Based on và nhấn OK. Bạn sẽ không thấy gì đặc biệt sau khi lệnh Split được OK cho tới khi bạn tiến hành 1 thủ tục thống kê bất kỳ (ví dụ ở đây là lệnh Frequencies).

2. Thực hiện lệnh Frequencies cho biến *docTTre*

3. Sau đó bạn nhớ phải trở lại hộp thoại Split File và chọn Analyze all cases, do not create groues để trả mọi phép phân tích lại bình thường tức là không phân tách theo thành phố. Đừng bao giờ quên bước này một khi đã thực hiện lệnh Split File, nếu không kết quả của các phép phân tích bình thường sau đó sẽ làm bạn bối rối.

Kết quả của lệnh Frequencies là:

Bảng 3.20

thành phố			Frequency	Percent	Valid Percent	Cumulative Percent
Hà Nội	Valid	khong doc TTre	206	82.4	82.4	82.4
		co doc bao TTre	44	17.6	17.6	100.0
		Total	250	100.0	100.0	
TPHCM	Valid	khong doc TTre	53	21.2	21.2	21.2
		co doc bao TTre	197	78.8	78.8	100.0
		Total	250	100.0	100.0	

Bạn có thấy số người có đọc Tuổi Trẻ tại Tp HCM là 197 ngườivới tần suất 78,8% giống như kết quả đã tính được ở Bảng 3.18 không? Bạn còn có thể so sánh thêm nhiều thông tin khác trên 2 bảng này.

Như vậy bạn có thể lọc thông tin trong các biến từ *c2a1* đến *c2a9* bằng hai cách: hoặc là ghép biến, hoặc chuyển hoá biến thành dạng Dichotomy rồi sau đó dùng các lệnh thống kê lọc thông tin trên 2 biến mới được xử lý này. Đây chỉ là một ví dụ đơn giản cho thấy rằng trong thực tế xử lý số liệu, bạn có thể vận dụng sáng tạo theo cách của riêng bạn trên cơ sở các thủ tục khuôn mẫu cố định của SPSS.

8. TRÌNH BÀY KẾT QUẢ BẰNG ĐỒ THỊ.

Đồ thị là một công cụ phân tích thống kê rất hữu ích do tính trực quan và sự hấp dẫn của nó. Đồ thị của SPSS có nhiều tính năng mạnh. Tuy nhiên đồ thị trong SPSS không có khả năng liên kết được với file văn bản Word, hoặc file thuyết trình Power Point (bằng lệnh Paste Special) do đó bạn nên chuyển thông tin sang Excel để vẽ đồ thị nếu muốn tạo liên kết để cập nhật và hiệu chỉnh nhanh trong quá trình làm việc.

8.1. Các loại đồ thị cơ bản của SPSS

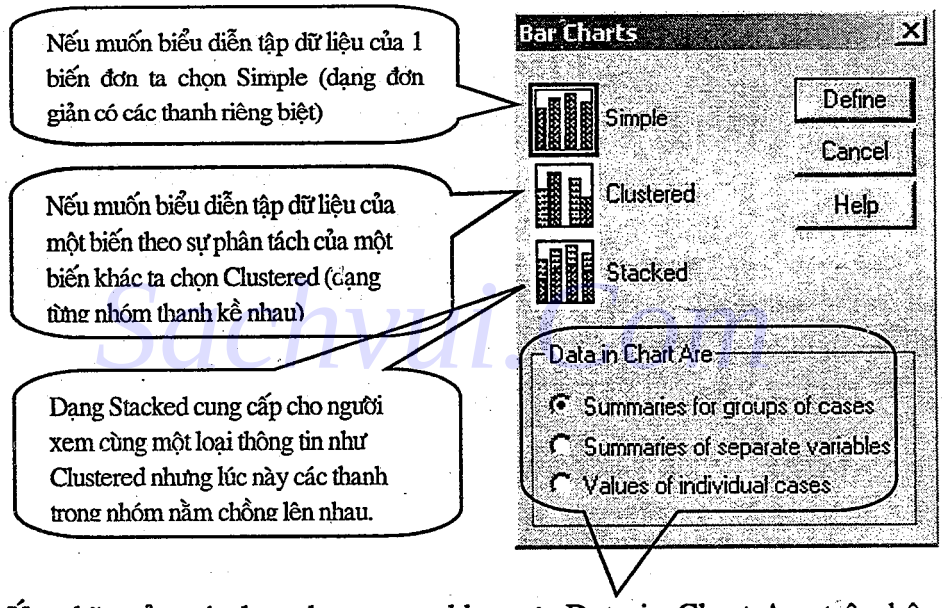
- Biểu đồ thanh Bar hay được sử dụng biểu diễn dữ liệu dưới dạng tần số hay tần suất%
- Biểu đồ hình tròn Pie: thường sử dụng biểu diễn dữ liệu định tính dạng tần số hay % khi chỉ có ít nhóm.
- Đồ thị đường gấp khúc (Line) và diện tích (Area): áp dụng tốt cho dữ liệu định lượng.

Các bước để khởi tạo và hiệu chỉnh 4 loại đồ thị trên với SPSS khá giống nhau, do đó chúng ta sẽ tìm hiểu chi tiết về cách tạo dựng và hiệu chỉnh đồ thị dạng thanh (Bar), bạn đọc sẽ tự mình suy ra cách thực hiện với 3 dạng đồ thị còn lại.

8.1.1 Đồ thị hình thanh (Bar)

Để tạo đồ thị bạn vào Menu Graphs>Bar để mở cửa sổ Bar Charts

Hình 3.28



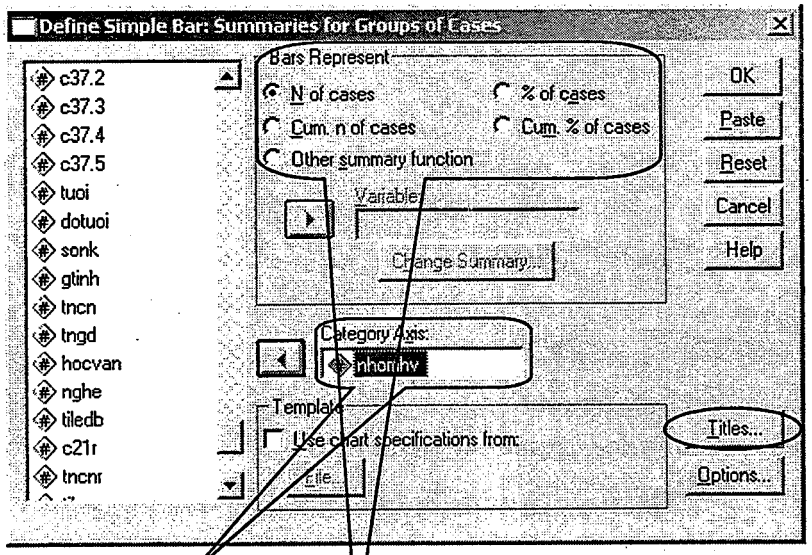
Ý nghĩa của các lựa chọn trong khu vực Data in Chart Are trên hộp thoại đồ thị thanh:

- Summaries for groups of cases: thể hiện 1 con số thống kê tổng hợp cho những nhóm trường hợp khác nhau, giả dụ ở trên bạn đã chọn Clustered thì lúc này mỗi thanh trên đồ thị thể hiện cùng một đại lượng thống kê mà bạn đã chọn tính của từng nhóm
- Summaries of separate variables: thể hiện những con số thống kê tổng hợp cho những biến khác nhau trên cùng một đồ thị.
- Value of indiviual cases: thể hiện giá trị thật của một biến trong từng tình huống cụ thể chứ không thể hiện những con số thống kê tổng hợp. Mỗi thanh đại diện cho trị số tuyệt đối của từng trường hợp. Vậy thì bạn chỉ có thể sử dụng tình huống này để vẽ đồ thị cho một mẫu thật ít quan sát, nếu không bạn sẽ cảm thấy rất rối với đồ thị của mình.

8.1.1.1 Dùng đồ thị thanh biểu diễn tập dữ liệu của 1 biến đơn

1. Trong hộp thoại Bar Charts chọn Simple, tại khu vực Data in Chart Area ta chọn Summaries for groups of cases rồi nhấp Define để mở cửa sổ Define Simple Bar:Summaries for groups of cases

Hình 3.29



2. Đưa biến *nhomhv* sang khung Category Axis,

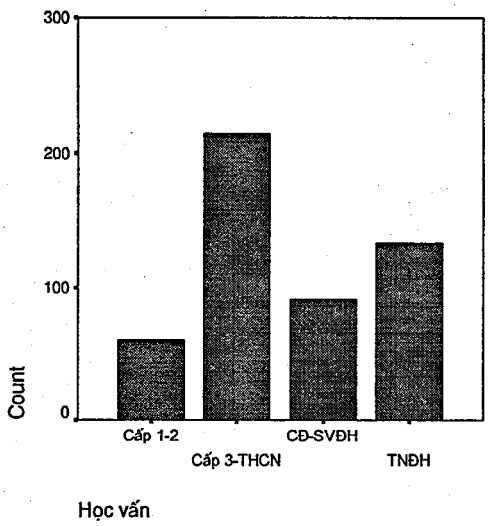
3. Trong khu vực Bar Represent là các lựa chọn về thông tin nào của biến *nhomhv* sẽ được thể hiện trên các thanh, ở đây bạn chọn Number of cases (tần số). Nếu bạn chọn % of cases thì đơn vị của trục tung sẽ là tần suất.

4. Nhấp nút Titles để nhập tên đồ thị, nhớ chuyển sang chế độ gõ chữ Việt với bảng mã TCVN3

5. Sau khi OK ta có được đồ thị hình thanh.

Những nhãn trên đồ thị không hiển thị dạng tiếng Việt sẽ được hiệu chỉnh thành tiếng Việt như đã hướng dẫn ở phần Tiếng Việt trong SPSS tại Chương I hoặc bằng cách nhấp đôi chuột trực tiếp vào đồ thị để mở cửa sổ SPSS Chart Editor, chọn Menu Chart>Axis để mở hộp thoại Axis Selection, ở hộp thoại này nếu chọn Scale bạn sẽ hiệu chỉnh được các đối tượng trên trục tung còn chọn Category bạn sẽ hiệu chỉnh được các đối tượng trên trục hoành (Xem thêm hướng dẫn ở phần hiệu chỉnh đồ thị ở phía sau).

Hình 3.30



Bạn sẽ thấy đồ thị thanh đứng của biến *nhomhv* với bảng tần số của biến *nhomhv* luôn thể hiện cùng 1 thông tin.

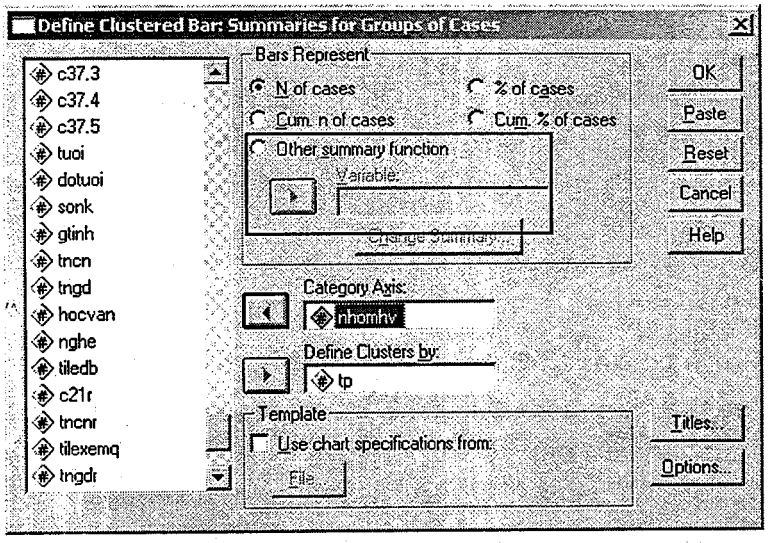
8.1.1.2 Dùng đồ thị thanh diễn tả tập dữ liệu của 1 biến được phân tách theo 1 biến khác

Lần lượt thao tác như trên với những chọn lựa Clustered và Summaries for group of Cases, rồi nhấp chuột vào nút Define bạn sẽ mở được cửa sổ hộp thoại Define Clustered Bar: Summaries for groups of cases.

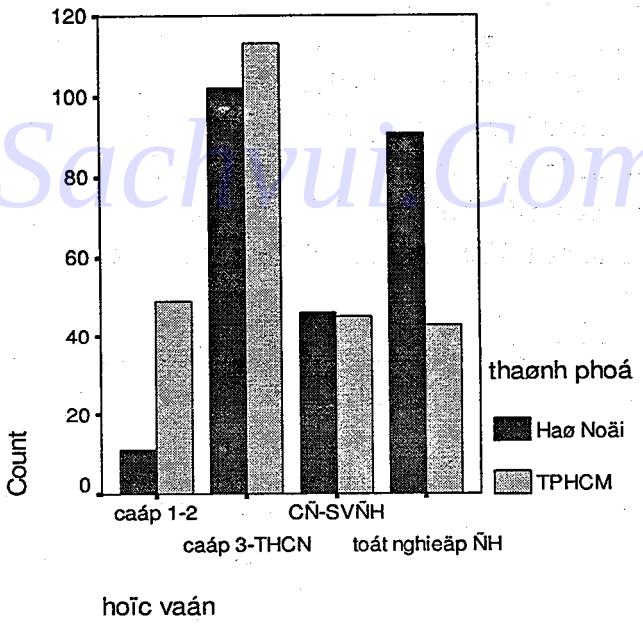
Đưa các biến sang các khu vực như hướng dẫn trong Hình 3.31. Sau khi OK bạn có đồ thị thô như Hình 3.32, chọn các đối tượng bạn muốn chỉnh sửa và hiệu chỉnh chúng theo ý muốn để có đồ thị hoàn chỉnh.

Ở khu vực Bars Represent ở Hình 3.31, khi bạn chọn Other Summary Funtion, khung Variable ở kế nó sẽ sáng lên, bạn đưa biến định lượng muốn tính các đại lượng thống kê mô tả sang, chú ý rằng SPSS mặc định tính đại lượng trung bình cho biến định lượng bạn đưa vào khung Variable, nếu muốn tính các đại lượng thống kê khác bạn vào nút Change Summary... để chọn lựa lại.

Hình 3.31



Hình 3.32



Nhận xét về đồ thị thể hiện mối quan hệ giữa biến *nhomhv* và *tp* tại từng thành phố. Tại mỗi nhóm học vấn trên trục hoành luôn có 2 thanh mang 2 màu nằm kề nhau thể hiện thông tin tại 2 địa điểm. Độ cao của thanh cho ta thông tin về số lượng người tại từng trình độ học vấn.

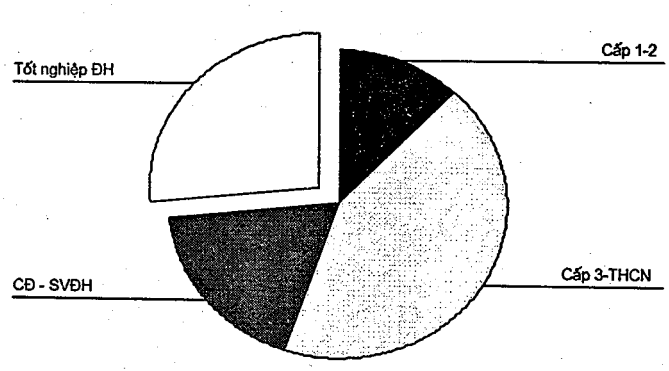
8.1.2 Đồ thị dạng đường và diện tích (Line and Area Chart)

Hai dạng đồ thị này có quan hệ rất gần với đồ thị Bar. Cả 3 dạng đồ thị này đều thể hiện tần số, giá trị của dữ liệu và các số thống kê cho mỗi biểu hiện riêng biệt của một biến. Các hướng dẫn đối với đồ thị Bar đều có thể áp dụng tương tự cho đồ thị Line và Area. Chú ý rằng khi đồ thị Area dạng chồng lên nhau được vẽ trên số liệu tổng hợp của các biến riêng biệt thì các biến này nên được đo lường trên 1 thang đo có thể so sánh được để giúp cho người quan sát nhận định dễ dàng hơn.

8.1.3 Đồ thị hình tròn (Pie)

Đồ thị Pie thể hiện thông tin về kết cấu rất tốt do nó giúp hình thành được cảm nhận về tổng thể và bộ phận của vấn đề nên bạn so sánh được các biểu hiện của biến hay các biến với nhau hoặc các giá trị riêng biệt. Các số thống kê được sử dụng cho đồ thị Pie là tần số, tần suất và tổng cộng. Các bước xây dựng một đồ thị Pie cũng đi theo trình tự như trên, để gia tăng hiệu quả thể hiện của đồ thị, SPSS cho phép bạn chỉnh sửa màu sắc và đặc biệt là bạn có thể tách riêng các "miếng" trên đồ thị ra để thu hút sự chú ý, muốn làm được vậy bạn hãy nhấp đôi chuột vào đồ thị để mở cửa sổ SPSS Chart Editor, chọn miếng bạn muốn tách ra bằng cách nhấp chuột trực tiếp vào nó để đường viền của nó hiện các chấm vuông, sau đó bạn chọn menu Format > Explode Slice miếng đã được chọn sẽ tự động được tách riêng ra.

Hình 3. 33 Đồ thị Pie minh hoạ các nhóm học vấn của mẫu quan sát.

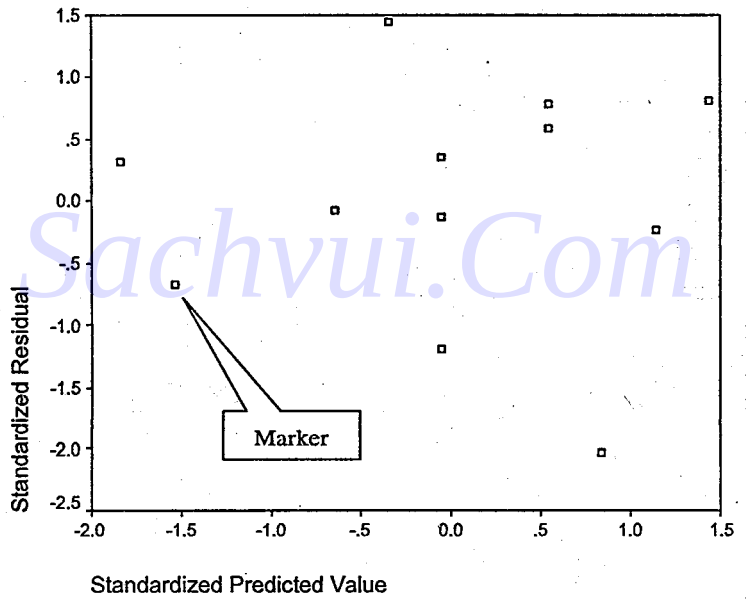


Bạn hãy tự mình kl ám phá thêm rất nhiều tính năng và thể hiện phong phú khác của đồ thị trên SPSS. Ngoài ra ở Chương VII bạn cũng sẽ được làm quen thêm với cách tạo một số loại đồ thị, biểu đồ sử dụng cho những mục đích đặc biệt như Scatter, Q-Q plot...

8.2. Hiệu chỉnh đồ thị trên SPSS

Chúng ta đã biết cách thể hiện chữ Việt trên đồ thị. Trong nội dung này chúng ta sẽ thực hiện một số hiệu chỉnh cơ bản cho đồ thị. Chú ý rằng các hiệu chỉnh này đều thực hiện trên cửa sổ Chart Editor. Chúng ta sẽ thực hiện ví dụ trên đồ thị Scatter, và bạn đọc tự suy ra cách hiệu chỉnh các loại đồ thị khác, bạn sẽ xem đồ thị ban đầu như sau:

Hình 3.34

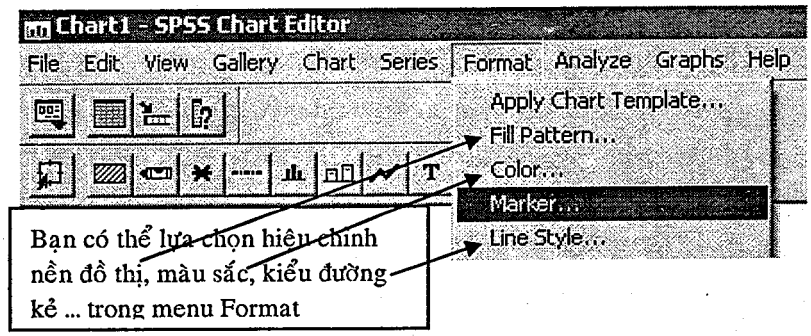


8.2.1 Hiệu chỉnh các điểm phân tán trên đồ thị (Markers)

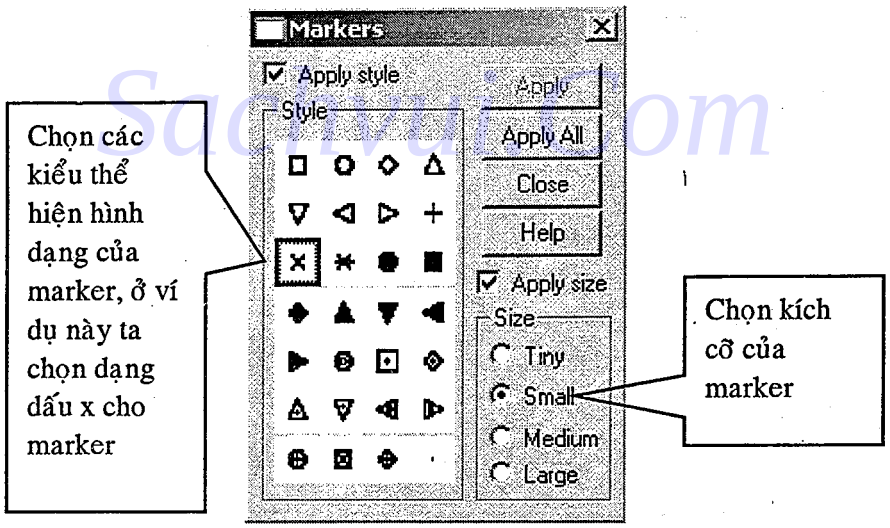
Ttrên cửa sổ Chart Editor, bạn rê chuột trực tiếp vào một marker bất kỳ trên đồ thị rồi nhấp chọn, tất cả các điểm trên đồ thị của bạn sẽ trở thành các chấm đen đậm thể hiện tình trạng được chọn. Bạn vào menu Format>Marker...để mở hộp thoại Marker (Hình 3.35). Hộp thoại Marker mở ra như Hình 3.36

Bạn thực hiện các lựa chọn như ở Hình 3.36 hướng dẫn và nhấp Apply All, rồi Close trên hộp thoại Marker.

Hình 3.35



Hình 3.36



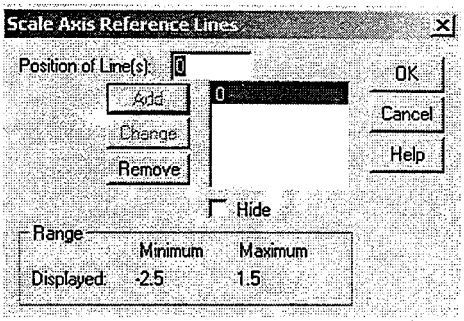
8.2.2 Thể hiện đường kẻ ngang trên đồ thị

Bạn vào menu Chart>Reference Line, sau khi hộp thoại Scale Axis Reference Lines mở ra, bạn có thể khai báo những giá trị trên trục tung mà bạn muốn đường kẻ đi ngang qua trong khung Position of Line(s), nhấp nút Add đưa nó vào danh sách, mặc định SPSS lựa chọn cho bạn đường thẳng đi ngang qua điểm 0 của trục tung.

Bạn cũng có thể Remove hoặc Change những giá trị đã nhập theo cách thông thường mà bạn đã quen từ những phần trước.

Trong khung Range thể hiện giá trị max và min của trục tung.

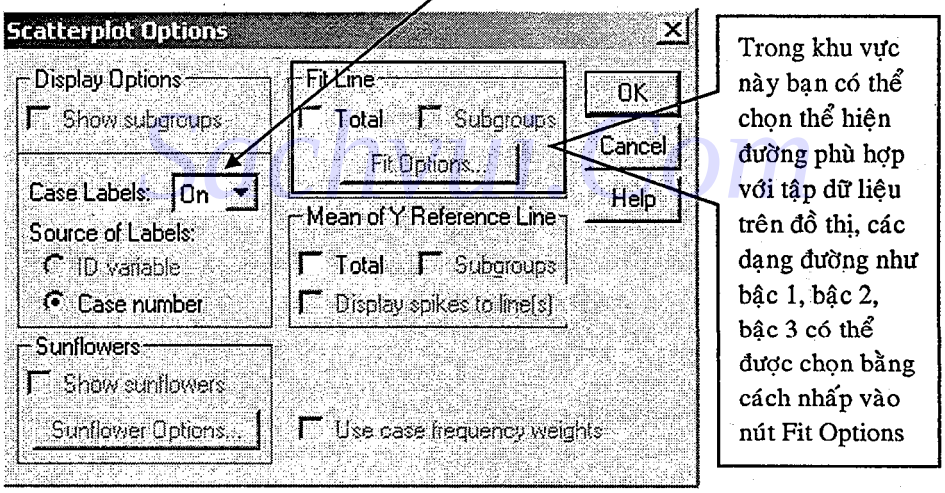
Hình 3.37



8.2.3 Thể hiện trị số tuyệt đối của từng trường hợp

Bạn vào menu Chart>Options... mở hộp thoại Scatterplot Options, trong hộp thoại này bạn chọn tình trạng On tại Case Labels để hiện trị số tuyệt đối của từng điểm trên đồ thị.

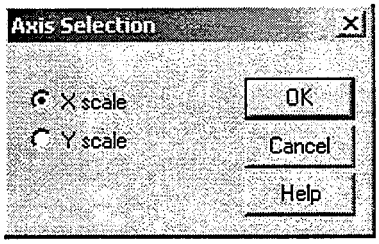
Hình 3.38



8.2.4 Hiệu chỉnh các vấn đề liên quan đến hai trục đồ thị

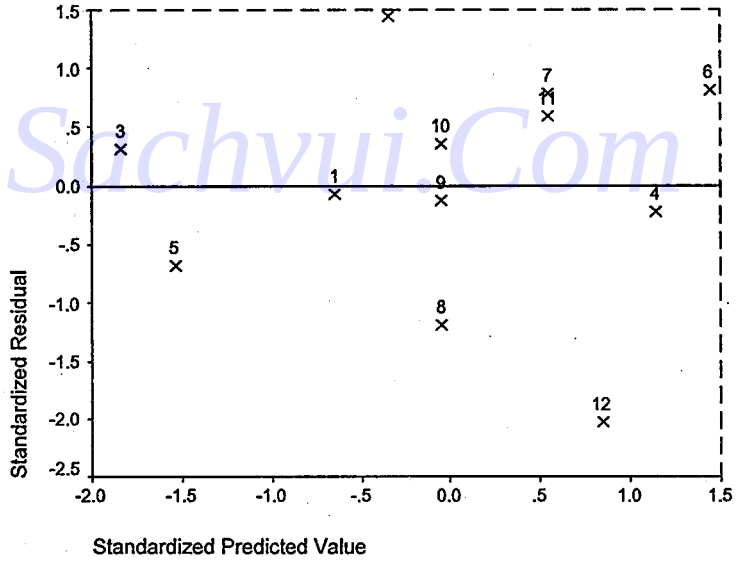
Muốn hiệu chỉnh các vấn đề liên quan đến 2 trục đồ thị như nhãn của trục, đơn vị lớn nhất nhỏ nhất, vị trí tiêu đề... bạn vào menu Chart > Axis... hộp thoại Axis Selection sẽ yêu cầu bạn trả lời là bạn muốn hiệu chỉnh những thông tin trên trục X hay trục Y, chọn xong bạn nhấp OK, hộp thoại hiệu chỉnh trục tương ứng sẽ mở ra cho bạn thực hiện các hiệu chỉnh cần thiết.

Hình 3.39



Cũng trong menu Chart bạn có thể chọn các lựa chọn để hiệu chỉnh Title..., Footnote..., Legend... cho đồ thị, khi hiệu chỉnh các đối tượng này bạn nhớ phải tuân thủ quy tắc thể hiện tiếng Việt trên đồ thị. Sau đây là đồ thị ở Hình 3.34 đã được hiệu chỉnh sau những lựa chọn của chúng ta

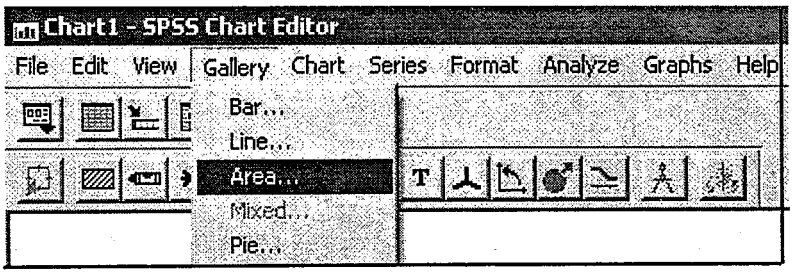
Hình 3.40



8.2.5 Chuyển đổi giữa các loại đồ thị

Giả sử bạn có đồ thị dạng Line, bạn muốn chuyển nó sang dạng Area hay dạng Pie... và ngược lại một cách nhanh chóng mà không phải chọn dữ liệu và vẽ lại từ đầu thì bạn vào menu Gallery và nhấp chọn tên đồ thị mà bạn muốn chuyển đổi sang trong danh sách tên các đồ thị được liệt kê. Vì mỗi loại dữ liệu chỉ phù hợp với một số loại đồ thị nhất định nên sẽ có những tên đồ thị không hiện sáng trong danh sách này. Với ví dụ đồ thị Scatter của chúng ta ở trên các bạn sẽ thấy tên các loại đồ thị trong danh sách Gallery đều ở tình trạng mờ.

Hình 3.41



Chú ý cách nhanh chóng nhất để hiệu chỉnh các đối tượng trên đồ thị là sau khi chuyển đồ thị sang cửa sổ Chart Editor, bạn kê chuột trực tiếp vào đối tượng nào trên đồ thị bạn muốn chỉnh sửa rồi nhấp đôi chuột, hộp thoại liên quan đến viêc chỉnh sửa đối tượng đó sẽ mở ra. Cách này không những nhanh chóng mà còn cứu nguy cho bạn khi bạn không nhớ hết các lệnh menu.

8.3. Lưu đồ thị

Bạn có thể lưu đồ thị được tạo bằng SPSS rồi sau đó xuất ra để in ấn, sử dụng theo nhiều cách:

- Cách thứ nhất là bạn copy và dán đồ thị lên một tài liệu Word, hay Excel và sao lưu tài liệu đó như thông thường.
- Cách thứ hai là lưu chính file Output có chứa đồ thị.
- Cách thứ ba là từ cửa sổ Chart Editor chọn menu File> Export Chart, hộp thoại Export Chart cho phép bạn lưu đồ thị dưới dạng một file ảnh riêng biệt.

8.4. Vẽ đồ thị bằng Excel

Cần phải nói là chức năng vẽ đồ thị trên SPSS ít quen thuộc hơn và ít linh động bằng chức năng vẽ đồ thị trên Excel do đó chúng ta có thể đưa bảng kết quả xử lý từ SPSS qua Excel bằng hai lệnh quen thuộc là: tổ hợp phím Control+C và Control+V, sau đó dùng các lệnh vẽ đồ thị trong Excel để thực hiện.

Lý do nên sử dụng Excel để vẽ đồ thị là:

- việc vẽ đồ thị nhanh và đơn giản,
- có thể tạo liên kết (link - bằng lệnh Paste Special) đồ thị trong Excel với file văn bản Word và file thuyết trình Power Point để việc chỉnh sửa trong Excel được dễ dàng cập nhật tự động trong Word và Power Point.

- Với Excel, từ một đồ thị mẫu, bạn có thể copy thêm một số đồ thị khác và thay đổi nguồn dữ liệu để tạo ra đồ thị có nội dung hoàn toàn mới nhưng hình thức (định dạng màu sắc, font chữ, cấu trúc...) nhất quán với đồ thị gốc ban đầu. Việc này sẽ giúp bạn tiết kiệm rất nhiều thời gian hiệu chỉnh đồ thị.

Các bước thực hiện một đồ thị/ biểu đồ trên Excel

Excel cung cấp cho chúng ta một Wizard gồm 4 bước để thực hiện vẽ đồ thị. Bạn đưa trỏ chuột đến nút Chart Wizard bấm vào và Excel sẽ hướng dẫn chúng ta lần lượt từng bước để hoàn thành đồ thị:

— Bước 1: xác định dạng đồ thị thích hợp

— Bước 2: xác định vùng dữ liệu và cấu trúc dữ liệu (theo hàng hay theo cột)

— Bước 3: xác định các chi tiết của đồ thị: tên đồ thị, tên trục đồ thị, đường lưới toạ độ, vị trí phần chú thích của đồ thị, hiện nhãn giá trị ...

— Bước 4: xác định nơi đặt đồ thị (thông thường là ngay trong cùng trang dữ liệu- cùng sheet)

Với trình tự 4 bước cơ bản này chúng ta sẽ dùng một ví dụ cụ thể để mô tả một số loại đồ thị hay gặp như sau:

8.4.1. Biểu đồ thanh ngang

Biểu đồ thanh ngang đơn giản

Để chuẩn bị dữ liệu bạn dùng lệnh Frequencies (menu Analyze/ Descriptive Statistics) để lập bảng tần số mô tả học vấn của những người tham gia phỏng vấn (biến *nhomhv*). Sau đó chép bảng kết quả từ màn hình Output sang cửa sổ làm việc của Excel bằng cách đưa trỏ chuột đến bảng số liệu kết quả muốn vẽ đồ thị rồi nhấp chuột trái. Bảng số liệu này sẽ được đóng khung (nghĩa là đã được chọn). Sau đó hãy bấm tổ hợp phím Control + C (hay vào menu SPSS chọn Edit -> Copy) để copy bảng kết quả này. Sau đó mở chương trình Excel và đưa bảng kết quả copy vào bảng tính Excel bằng cách nhấn tổ hợp phím Control + V (hay vào menu Excel chọn Edit -> Paste). Kết quả được đưa vào Excel, chúng ta có thể chỉnh sửa định dạng lại (hay sắp xếp thứ tự) như trong hình sau:

Hình 3.42

	A	B	C	D	E	F
1	học vấn					
2			Frequency	Percent	Valid Percent	Cumulative Percent
3	Valid	cấp 1-2	60	12	12	12
4		cấp 3-THCN	215	43	43	55
5		CĐ-SVĐH	91	18.2	18.2	73.2
6		tốt nghiệp ĐH	134	26.8	26.8	100
7		Total	500	100	100	

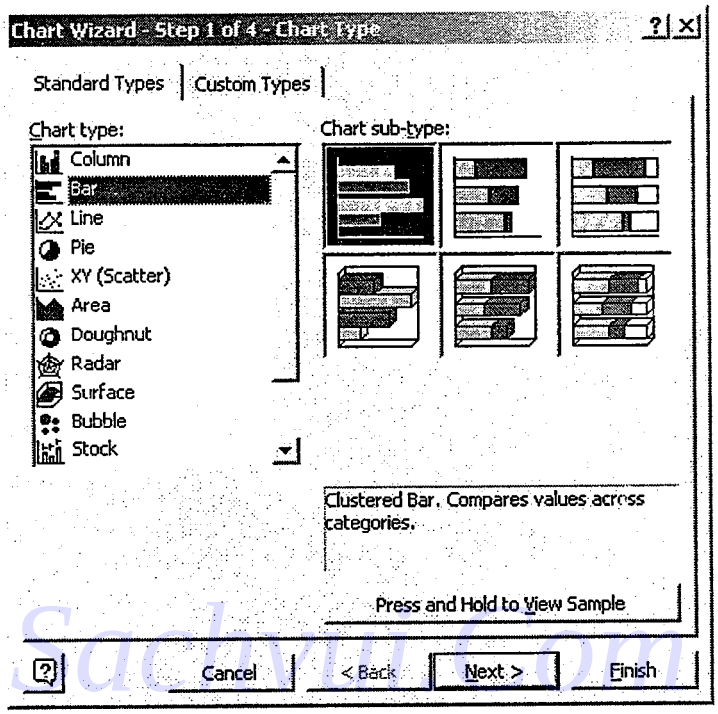
Chúng ta có thể bỏ bớt các thông tin không cần thiết tùy theo loại biểu đồ sử dụng. Trong ví dụ này, chúng ta sẽ vẽ biểu đồ thanh ngang đơn giản nên chỉ giữ lại cột % và xóa các cột khác bằng lệnh xóa cột (Delete column trong Excel). Sau đó đánh dấu vùng dữ liệu trên bảng tính Excel đang làm việc mà bạn muốn thể hiện trên đồ thị bằng cách nhấn kéo thả nút trái chuột quen thuộc trong Windows. Vùng dữ liệu sẽ được tô đậm như trong hình sau:

Hình 3.43

	A	B	C	D
1				
2			Percent	
3		cấp 1-2	12	
4		cấp 3-THCN	43	
5		CĐ-SVĐH	18.2	
6		tốt nghiệp ĐH	26.8	
7		Total	100	

Tiếp theo đưa trỏ chuột bấm vào nút Chart wizard (hay vào menu của Excel chọn Insert -> Chart). Lệnh này mở ra hộp thoại hướng dẫn vẽ đồ thị 4 bước trong Excel như sau:

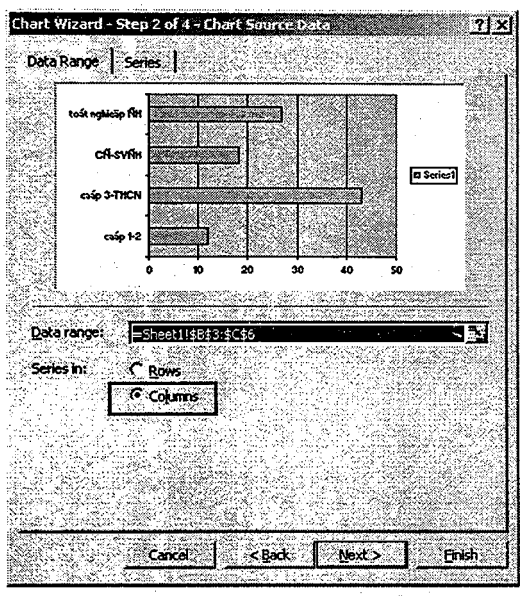
Hình 3.44



Như đã nói ở Bước 1 trong tiến trình vẽ đồ thị phần trên, trong hình cho thấy chúng ta có thể chọn loại biểu đồ phù hợp với số liệu và mục tiêu nghiên cứu. Trong ví dụ này, chúng ta chọn thanh ngang (Bar) ở phần chart type (phía trái của hộp thoại) và chọn kiểu (Chart sub-type) so sánh giữa các nhóm. Sau đó nhấp nút Next

Qua Bước 2 chúng ta có thể theo dõi trong hình bên sau.

Hình 3.45

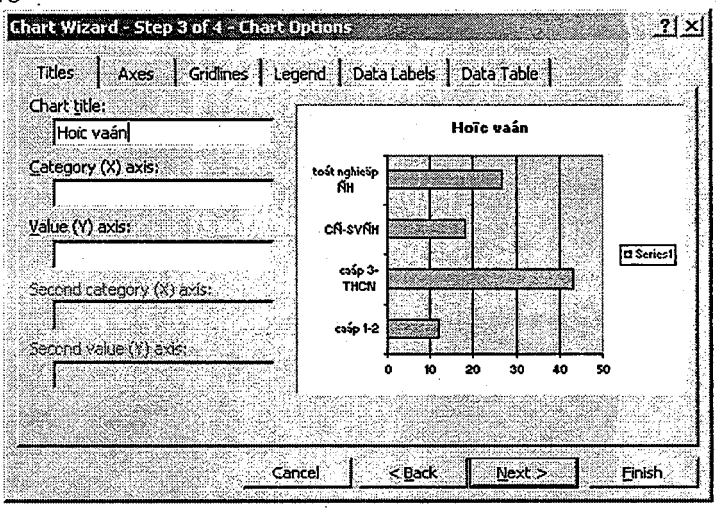


Trong hình chúng ta thấy mặc định đang chọn Columns, số liệu chúng ta đang có cũng là số liệu sắp xếp theo cột nên không cần chỉnh sửa. Nhấp tiếp nút Next để qua Bước 3.

Trong hộp thoại bước 3, chúng ta có thể xác định nhiều thuộc tính của đồ thị như:

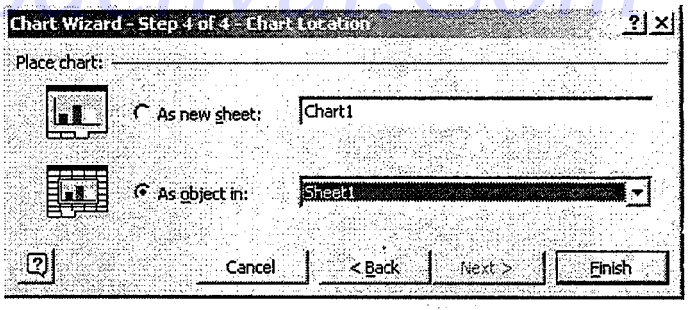
- Titles: tiêu đề biểu đồ, tiêu đề trục hoành và trục tung
- Axes: hiện tên các phân loại của từng biến ở trục hoành và trục tung
- Gridlines: hiện các đường lưới toạ độ của đồ thị
- Legend: cho hiện và vị trí hiện ghi chú của đồ thị (trong đồ thị dạng đơn giản thì không cần thiết)
- Data Labels: cho hiện số liệu ghi chú, tên của từng thanh ngang hay từng thành phần diễn tả số liệu của biểu đồ
- Data Table: cho hiện bảng số liệu chi tiết vẽ nên biểu đồ

Hình 3.46



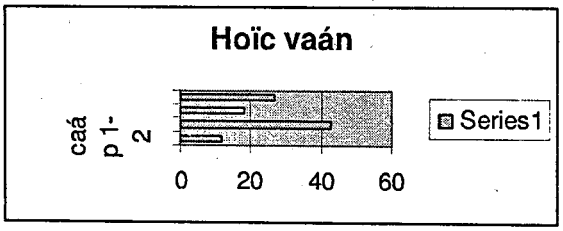
Sau đó nhấp nút Next để qua Bước 4. Trong bước này, chúng ta có thể xác định chỗ để biểu đồ, ở trong cùng worksheet (trang bảng tính) với bảng số liệu, hay để trong một trang bảng tính khác.

Hình 3.47



Đến đây, chúng ta có thể nhấp nút Finish để hiện ra đồ thị kết quả như trong Hình 3.48.

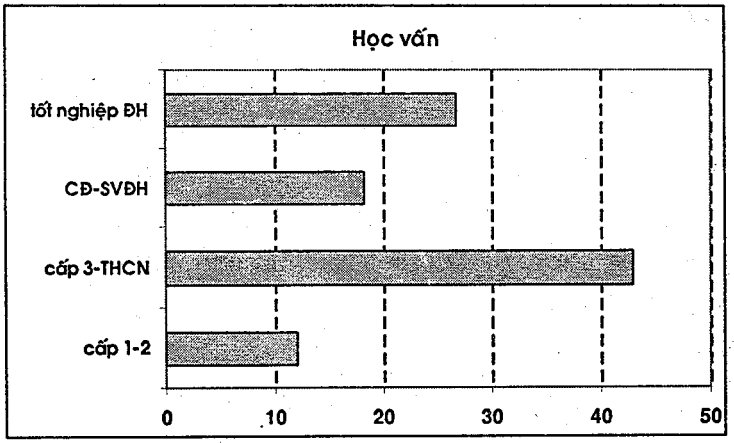
Hình 3.48



Tuy nhiên đồ thị kết quả này cần phải được chỉnh sửa rất nhiều mới được hình thức như mong muốn sau đây. Cách thức hiệu chỉnh đồ thị

theo quy tắc khá đơn giản là muốn chỉnh sửa đối tượng nào trên đồ thị bạn nhấp trái chuột trực tiếp vô đối tượng đó để chọn đối tượng rồi sau đó bấm chuột phải gọi menu tắt để chọn lệnh giúp chỉnh sửa đối tượng.

Hình 3.49

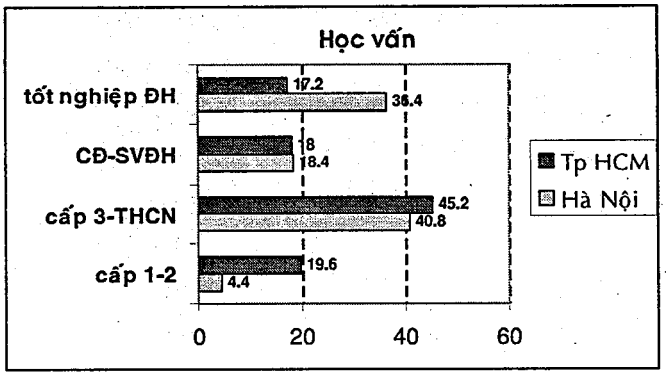


Chú ý là chúng ta có thể sắp xếp các số liệu theo trật tự tăng dần hay giảm dần trước khi vẽ biểu đồ để nhấn mạnh ý muốn biểu đạt.

Biểu đồ thanh ngang dạng kết hợp

Áp dụng khi chúng ta có số liệu tổng hợp của một biến định tính được phân tách theo một biến định tính khác. Ví dụ trình độ học vấn khác nhau như thế nào giữa hai vùng khảo sát thể hiện qua đồ thị sau

Hình 3.50



Để vẽ được đồ thị trên, chúng ta chuẩn bị dữ liệu như sau:

- Dùng lệnh Basic table của SPSS tạo bảng kết hợp hai biến nhomhv (nhóm học vấn) và tp (thành phố). Sau đó chép kết quả sang Excel

định dạng cho dễ coi hơn như sau:

Hình 3.51

	A	B	C	D	E	F
1			thành phố			
2			Hà Nội		TPHCM	
3			Count	Col %	Count	Col %
4	học vấn	cấp 1-2	11	4.4	49	19.6
5		cấp 3-THCN	102	40.8	113	45.2
6		CĐ-SVĐH	46	18.4	45	18
7		tốt nghiệp ĐH	91	36.4	43	17.2

Nhớ kiểm tra các số liệu tổng cộng để bảo đảm không có sai sót nào, khi chỉnh sửa sao chép. Sau đó xóa các thông tin dư đi để chỉ còn lại phần dữ liệu như sau:

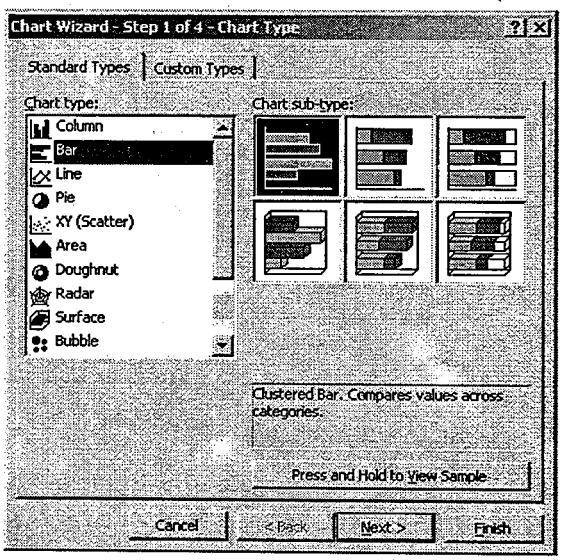
Hình 3.52

	A	B	C	D
1		Col %	Col %	
2	cấp 1-2	4.4	19.6	
3	cấp 3-THCN	40.8	45.2	
4	CĐ-SVĐH	18.4	18	
5	tốt nghiệp ĐH	36.4	17.2	
6				

- Quét chọn dữ liệu, chọn biểu tượng Chart Wizard vào cửa sổ vẽ đồ thị và lần lượt đi qua 4 bước theo thứ tự minh họa bằng hình ảnh sau:

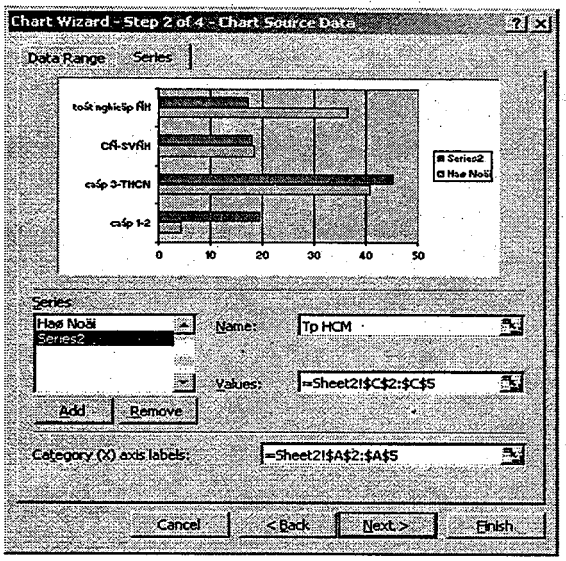
<u>Bước 1</u>:

Hình 3.53



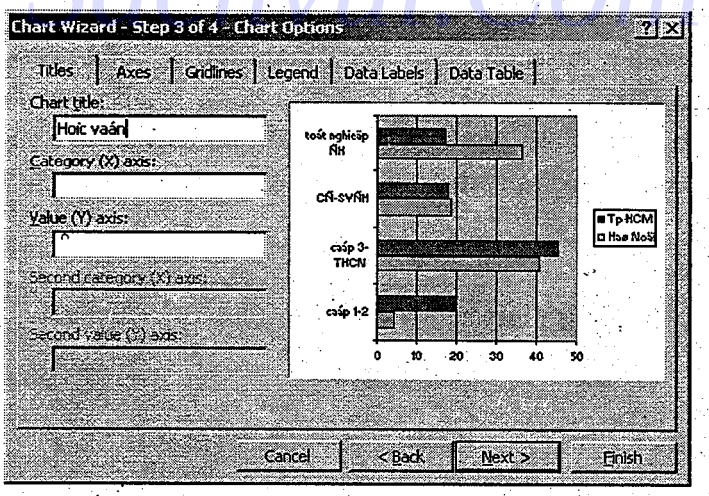
Bước 2: trong bước này các bạn bấm vào Phiếu Series để đổi tên hai chuỗi dữ liệu Series 1 thành Hà Nội và Series 2 thành Tp HCM

Hình 3.54



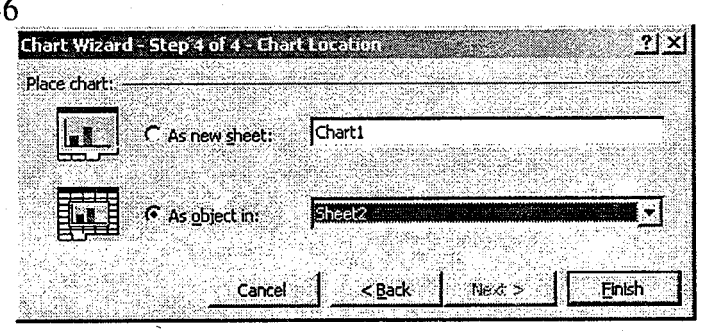
Bước 3:

Hình 3.55



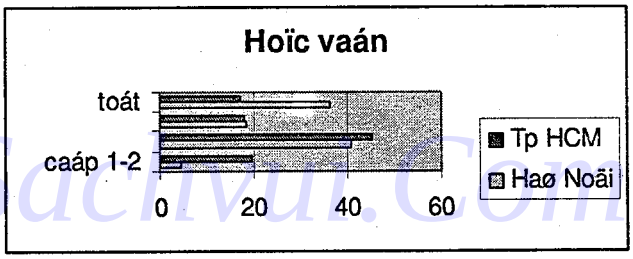
Bước 4:

Hình 3.56



Nhấp nút Finish bạn có kết quả như sau

Hình 3.57

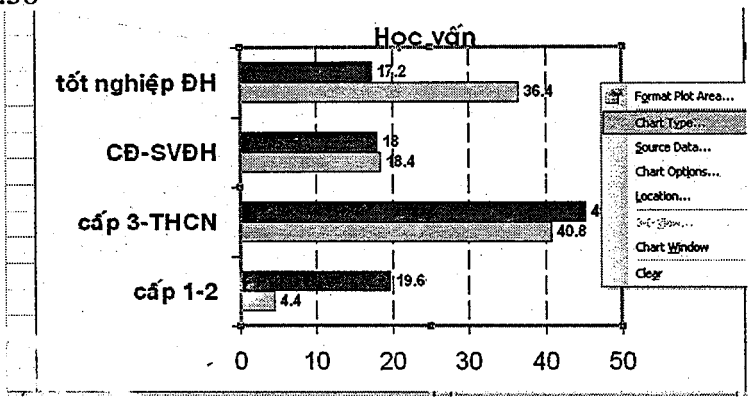


Thực hiện các chỉnh sửa để có đồ thị hoàn chỉnh như đã thấy ban đầu.

8.4.2. Biểu đồ thanh đứng (cột dọc)

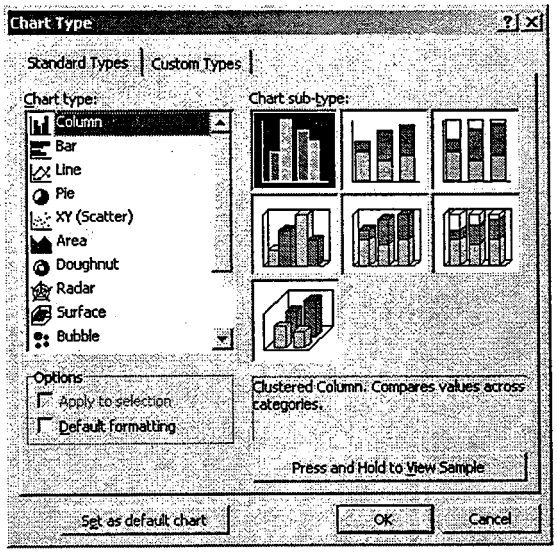
Về nguyên tắc thì cách vẽ biểu đồ thanh đứng không khác biểu đồ thanh ngang chỉ cần trong Bước 1 bạn chọn Chart type là Column, thậm chí từ một biểu đồ thanh ngang sẵn có bạn có thể chọn lại cho nó trở thành biểu đồ thanh đứng bằng cách sau:

Hình 3.58



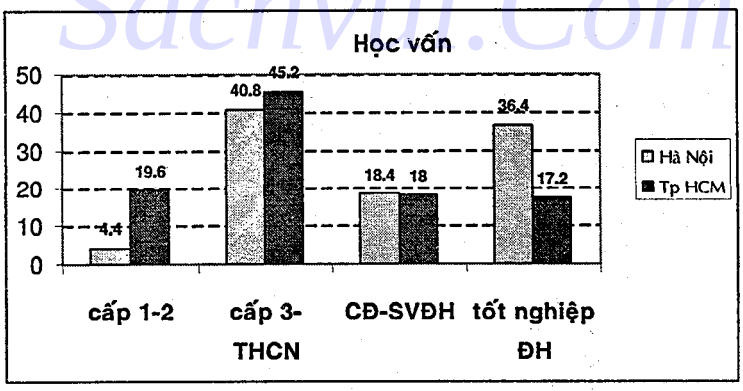
- Bấm chọn toàn bộ biểu đồ rồi bấm chuột phải chọn mục Chart type
- Trong cửa sổ xuất hiện tiếp bạn bấm chọn Column, sau đó chọn OK

Hình 3.59



Biểu đồ sẽ đổi chiều như sau:

Hình 3.60



Dĩ nhiên bạn phải chỉnh sửa lại nhiều chi tiết nữa cho đồ thị đẹp mắt hơn, và có một điểm cần chú ý là nếu các biểu hiện của các nhóm có nội dung quá dài bạn không nên để đồ thị kiểu đứng sẽ làm các nhãn biểu hiện bị xuống dòng liên tục nên khó đọc.

8.4.3. Biểu đồ stack

Chúng ta tiếp cận các vẽ đồ thị này bằng một ví dụ thật đơn giản với mô tả như sau: trong một tình huống nào đó ta có 5 câu hỏi mà mỗi câu đều có các lựa chọn trả lời từ loại 1 đến loại 4 (chẳng hạn Rất không đồng ý-Không đồng ý-Đồng ý-Rất đồng ý). Sau khi thu thập dữ liệu hoàn chỉnh ta lập bảng tổng hợp về tỷ lệ % từng lựa chọn trả lời cho mỗi câu. Nhập bảng dữ liệu này vào Excel theo cấu trúc như hình dưới (nhớ kiểm tra lại để bảo đảm tổng cộng của các cột đúng là 100%).

Hình 3.61

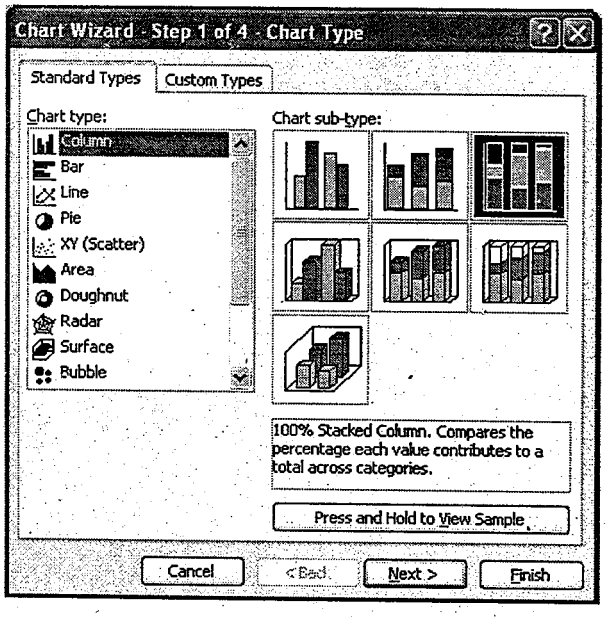
	A	B	C	D	E	F
1		Cau 1	Cau 2	Cau 3	Cau 4	Cau 5
2	Loai 1	19	30	20	25	25
3	Loai 2	36	15	37	15	25
4	Loai 3	31	20	19	23	25
5	Loai 4	14	35	24	37	25
6						

Trình tự vẽ đồ thị Stack như sau:

- Rê chuột quét chọn khối toàn bộ dữ liệu từ A1 đến F5

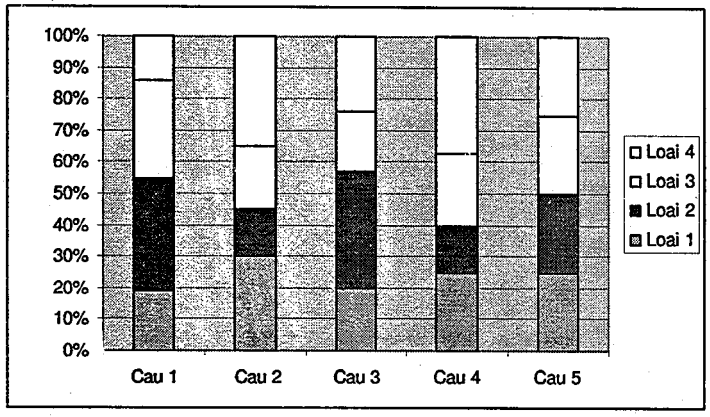
- Nhấp chuột vào biểu tượng Chart Wizard  trên cửa sổ làm việc của Excel để mở cửa sổ sau

Hình 3.62



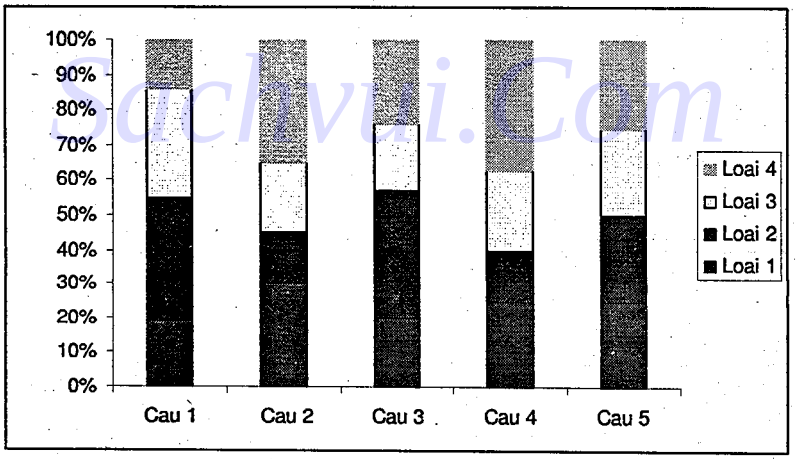
- Bấm chọn kiểu đồ thị cột đứng (Column) trong Chart Type và chọn kiểu Stacked Column trong phần Chart sub-type bên tay phải. Sau đó nhấn nút Finish. Bạn được đồ thị với thể hiện như sau

Hình 3.63



Bạn có thể làm các hiệu đính theo ý thích để được đồ thị sống động hơn

Hình 3.64



8.4.4. Đồ thị hình tròn

Nguyên tắc của đồ thị hình tròn đã được mô tả bên phần hướng dẫn vẽ đồ thị của SPSS, đồ thị hình tròn của Excel có một số thể hiện hấp dẫn hơn ví dụ đồ thị hình tròn kiểu 3-D.

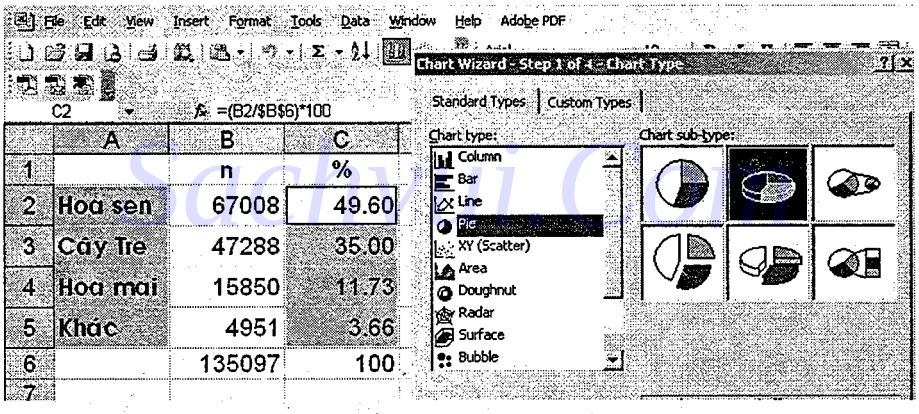
Để minh họa cho phần này chúng ta sẽ dùng một ví dụ về kết quả bình chọn của bạn đọc Tuổi trẻ Online về quốc hoa của VN, số liệu được tổng hợp và nhập lên Excel như hình sau:

Hình 3.65

	A	B	C
1		n	%
2	Hoa sen	67008	49.60
3	Cây Tre	47288	35.00
4	Hoa mai	15850	11.73
5	Khác	4951	3.66
6		135097	100

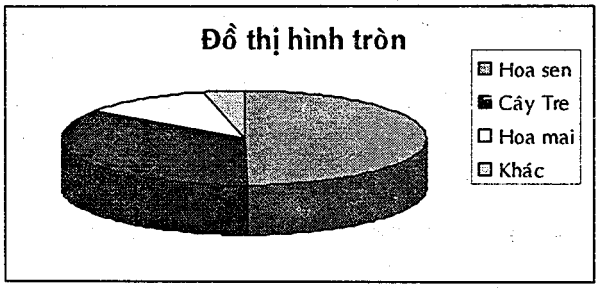
Vì chúng ta chỉ dùng thông tin tần suất để vẽ nên chúng ta có thể quét chọn dữ liệu cách quãng nhau bằng cách đè thêm phím Ctrl khi chọn, sau đó bấm nút Chart Wizard để chọn loại đồ thị như hình dưới.

Hình 3.66



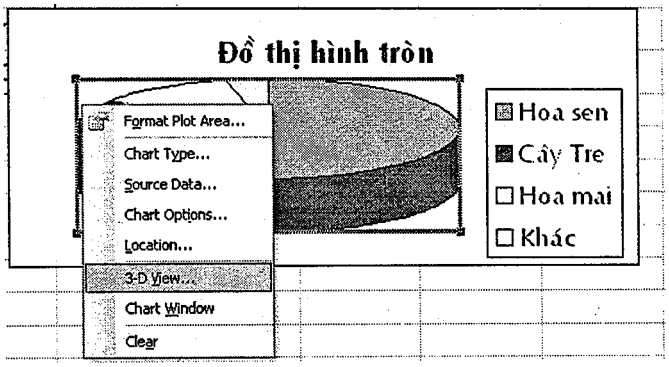
Ba bước sau chúng ta cũng tiến hành theo trình tự đã biết rồi nhấp nút Finish, chỉnh sửa để được kết quả sau.

Hình 3.67



Excel còn có thể giúp bạn tăng độ nổi của đồ thị 3_D nếu bạn vào menu phải như hướng dẫn trong hình dưới đây.

Hình 3.68

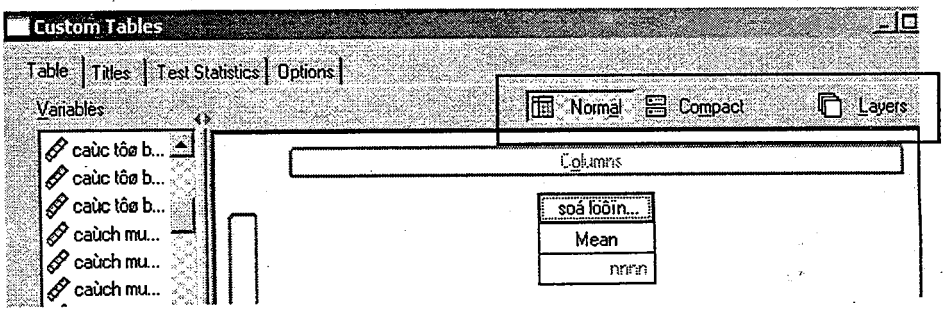


9. BẢNG TÙY BIẾN (Custom tables)

Trong một số trường hợp, chương trình SPSS đã cài đặt không có đủ lệnh chạy bảng, nên có thể bạn sẽ không thấy hiển thị menu General Table hoặc Basic Table mà chỉ có thấy Custom Tables thì bạn cũng có thể xoay sở thực hiện các yêu cầu tóm tắt và trình bày dữ liệu theo hướng dẫn sau đây.

Khi mở cửa sổ Custom Tables sẽ có một cửa sổ cảnh báo rằng để sử dụng tối ưu lệnh này thì Value lable của các biến phân loại phải được định nghĩa đầy đủ với các cấp độ đo lường được đặt đúng cách. Có thể bấm nút OK để bỏ qua cửa sổ này hoặc vào Define Variable Properties để khai báo cho các thuộc tính các biến. Nếu muốn cửa sổ này không nhảy ra làm phiền bạn nữa hãy click chuột vào lựa chọn Don't show this message in the future. Khi bấm Ok để mở cửa sổ ra, trên hàng ngang có 3 loại bảng để lựa chọn là bảng bình thường, bảng kết hợp và bảng phân lớp.

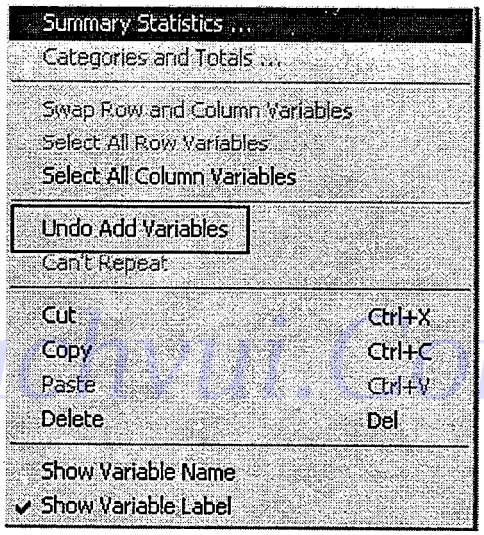
Hình 3.69



Giả sử bạn muốn tính các đại lượng thống kê mô tả cho một biến định lượng như biến c3 bạn lần lượt thao tác theo các hướng dẫn sau. Rê chuột trái mang biến c3 sang phần **Columns** thì thả chuột trái ra mặc định có các số liệu và cấu trúc như hình trên.

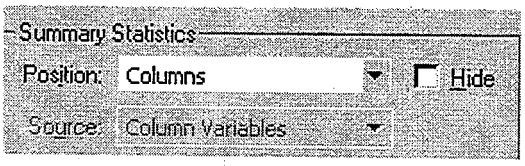
Nếu bấm vào Undo Add Variables thì biến vừa chọn đưa sang xử lý sẽ được trả lại danh sách biến bên tay trái.

Hình 3.70



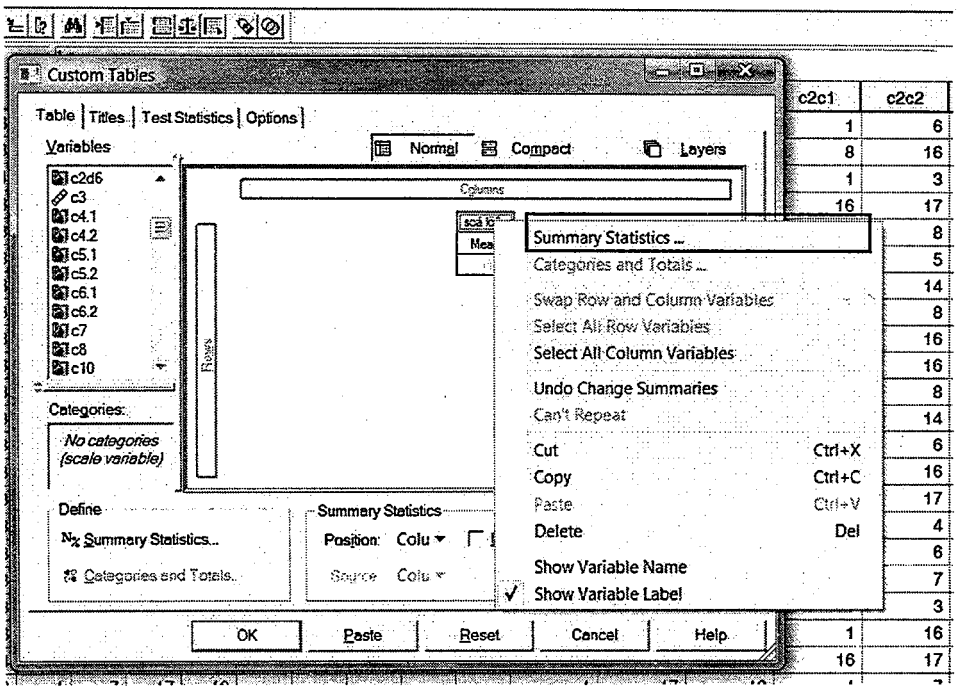
Bấm vào mục Position (ở phía dưới hộp thoại) để lựa chọn sắp xếp ở cột hay hàng các tham số thống kê (như Trung bình, Phương sai ...)

Hình 3.71



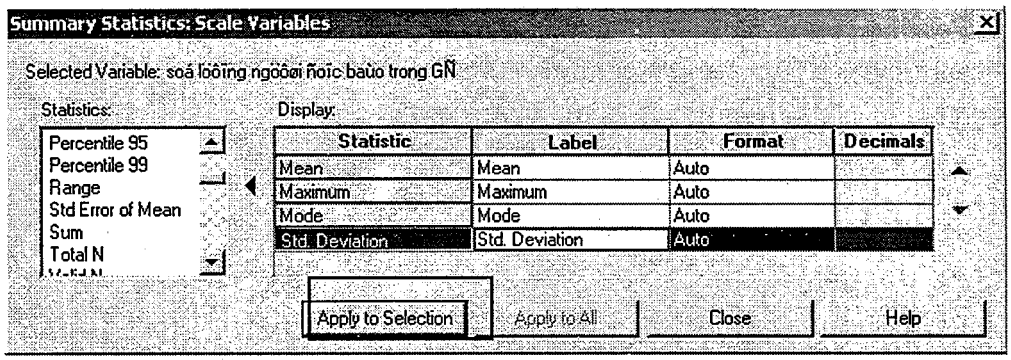
Ngoài hàm tính toán mặc định là Mean (trung bình), muốn lựa chọn tính thêm các tham số thống kê khác bạn nhấp chuột vào Nhãn biến (số lượng nhân khẩu của biến c3 đang hiện ở trên tên hàm mặc định Mean) rồi nhấp chuột phải một lần sẽ xuất hiện menu sau:

Hình 3.72



Nhấp chọn mục đầu tiên Summary Statistics ... hộp thoại Summary Statistics ứng với biến định lượng c3 mở ra. Bạn lần lượt chọn tên các hàm thống kê muốn tính trong danh sách hàm trong khung phía bên trái và bấm vào nút mũi tên chuyển hàm đã chọn qua danh sách hàm sẽ tính trong bảng bên phải như minh họa trong hình sau.

Hình 3.73



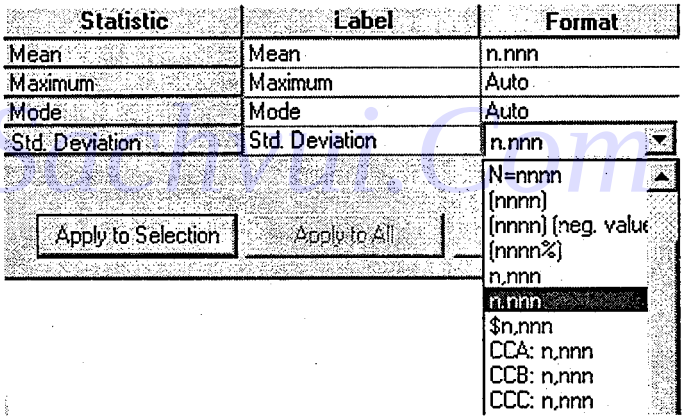
Sau khi nhấn nút Apply to Selection, bạn trở lại màn hình chính và nhấn tiếp nút Ok bạn được một bảng số liệu đơn giản như sau

Bảng 3.21

số lượng người đọc báo trong GĐ			
Mean	Maximum	Mode	Std. Deviation
3	15	3	2

Nếu có sự so sánh với các số liệu thống kê mô tả về biến c3 đã thực hiện ở trước bạn sẽ thấy rằng giá trị trung bình và độ lệch chuẩn đều bị làm tròn số do đó muốn nhận được giá trị thực với hai số lẻ sau dấu phẩy bạn hãy lặp lại lệnh cũ cho đến bước chọn lệnh Summary Statistics thì tiến hành thêm các lựa chọn sau đây.

Hình 3.74



Ở đây chúng ta chọn kiểu format là n.nnn (ngoài ra bạn đọc có thể tự chọn các kiểu thể hiện khác như (nnnn) tức là số liệu dược đặt trong ngoặc đơn; hay N = nnnn tức số liệu được đặt sau chữ N và dấu =, ... để xem các dạng format phong phú ra sao).

Sau đó bạn nhớ sang khung Decimal chọn lựa số lẻ sau dấu phẩy theo ý mình, ví dụ 3 số. sau đó nhấp nút apply và rồi sau cùng là nút OK bạn được kết quả sau

Bảng 3.22

số lượng người đọc báo trong GĐ			
Mean	Maximum	Mode	Std. Deviation
3,474	15	3	1,800

Trong trường hợp bạn muốn lập bảng kết hợp một biến định tính và một biến định lượng như số người đọc báo (*c3*) tại hai thành phố (*tp*) như bảng dưới đây bạn sẽ thực hiện như sau.

Bảng 3.23

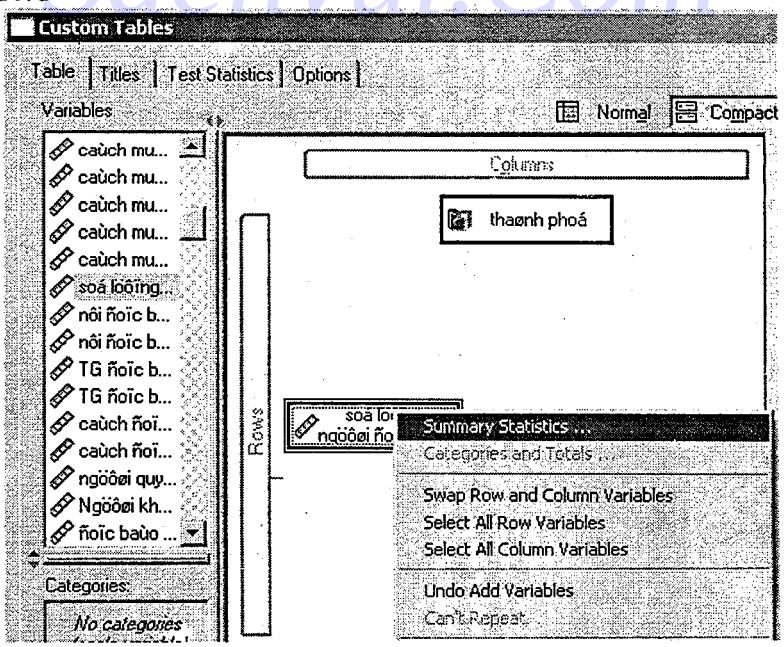
	thành phố			
	Hà Nội		TPHCM	
	Mean	Std. Deviation	Mean	Std. Deviation
số lượng người đọc báo trong GĐ	3,12	1.29	3,83	2.14

Trước tiên bạn phải về cửa sổ Variable View trong *Data thuc hanh* vào phần khai báo Measure của biến tp kiểm tra xem nó có được chọn là Nominal hay không, nếu không thì bạn chọn lại cho đúng cách. Sau đó lại vào mục Custom Tables thực hiện các thao tác sau:

Rê biến tp (biến định tính) bỏ vào khu vực Columns và c3 (biến định lượng) bỏ vào khu vực Rows

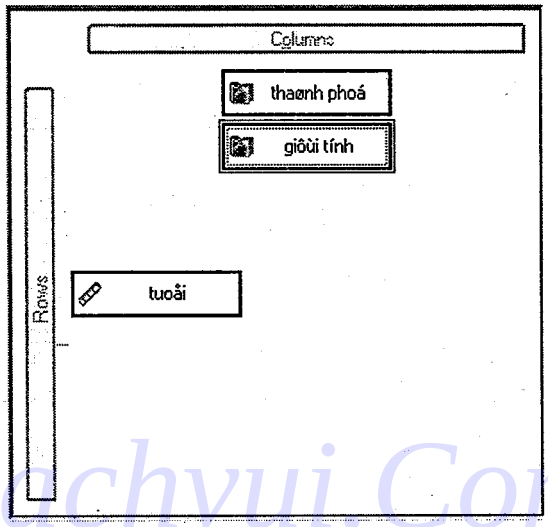
Chọn biến c3 rồi bấm chuột phải trên nó để mở menu tắt như hình dưới sau đó chọn Summary ststistics rồi chọn các tham số thống kê quan tâm và các thuộc tính đi kèm như cách làm phần trên đã hướng dẫn.

Hình 3.75



Nếu muốn lập bảng phức tạp hơn nữa như bảng mô tả các tham số thống kê về tuổi của hai giới nam nữ tại hai thành phố riêng biệt là Hà Nội và Tp HCM chúng ta rê dữ liệu để sang khung của cửa sổ Custom Tables như sắp đặt sau.

Hình 3.76



Bảng số liệu thu được có bề ngang rất lớn, như dưới đây

Bảng 3.24

	thaønh phoá											
	Haø Noäi						TPHCM					
	giôùi tính						giôùi tính					
	Nam			Nöõ			Nam			Nöõ		
	Mean	Count	td. Deviatio	Mean	Count	td. Deviatio	Mean	Count	td. Deviatio	Mean	Count	td. Deviatio
tuoåi	36	118	12	35	132	12	35	131	10	31	119	11

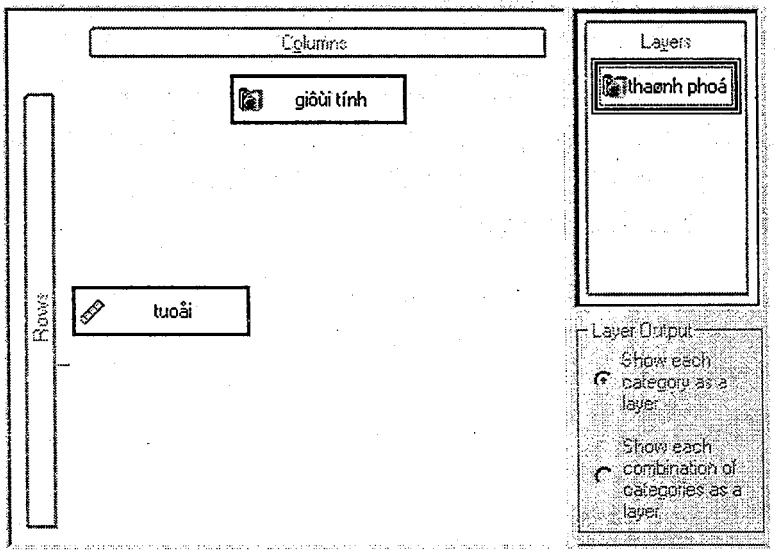
Muốn cho dễ nhìn chúng ta dùng Pivot table đảo chiều bảng lại như sau (phải thêm một số chỉnh sửa thủ công cho bảng đẹp mắt hơn).

Bảng 3.25

tuổi	thành phố	Hà Nội	giới tính	Nam	Mean	36,11
					Std. Deviation	11,60
				Nữ	Mean	34,66
					Std. Deviation	11,68
		TPHCM	giới tính	Nam	Mean	35,37
					Std. Deviation	10,21
				Nữ	Mean	31,13
					Std. Deviation	10,84

Có một cách thức hai để có bảng gọn hơn là cho hiển thị theo lớp, muốn vậy trong cửa sổ Custom table làm như hình sau, trong hình mô tả cách bạn nhặt biến tp bỏ sang khung Layers, các tùy chọn khác không có gì thay đổi

Hình 3.77



Bấm nút OK bạn có kết quả như sau:

Bảng 3.26

thành phố TPHCM

	giới tính			
	Nam		Nữ	
	Mean	Std. Deviation	Mean	Std. Deviation
tuổi	35,37	10,21	31,13	10,84

Nhấp chuột trái hai lần trên bảng này cho hiện viền răng cưa thì đồng thời bạn sẽ thấy dấu mũi tên cho phép chọn bảng ở lớp thứ hai là bảng số liệu tại Tp HCM.

Bảng 3.27

Layer	thaønh phoá TPHCM			
	thaønh phoá Haø Noäi		tính	
	thaønh phoá TPHCM		Nöõ	
	Mean	Std. Deviation	Mean	Std. Deviation
tuoåi	35,37	10,21	31,13	10,84

Trên cơ sở những hướng dẫn này các bạn có thể tự mình tìm hiểu thêm những tính năng khác của Custom Tables và bạn sẽ thấy cũng khá nhiều công dụng đáng kể.

CHƯƠNG IV
KIỂM ĐỊNH MỐI LIÊN HỆ GIỮA HAI BIẾN ĐỊNH TÍNH

Ở Chương III, khi chúng ta lập các bảng kết hợp 2 biến định tính, chúng ta mới chỉ mô tả những mối quan hệ mà chúng ta nhận thấy trong mẫu. Ví dụ ở bảng kết hợp 2 biến định tính được lập bằng lệnh Basic Tables dưới đây (Bảng 4.1) ta nhận thấy, với mẫu quan sát 500 người, nhóm người trả lời có trình độ học vấn cao có xu hướng đọc báo theo cách "xem lướt qua các đề mục và đọc các mục ưa thích trước". Nhóm người đọc có trình độ học vấn cấp 1-2 lại "chỉ đọc các trang mục ưa thích, ít đọc các trang khác" hoặc hay tìm đọc các tin đáng chú ý trước. Như vậy có lẽ trình độ học vấn có tác động đến cách đọc báo, nói cách khác, phải chăng có mối liên hệ giữa trình độ học vấn và cách đọc báo của người đọc.

Bảng 4.1

		học vấn				Tổng
		cấp 1-2	cấp 3-THCN	CĐ-SVĐH	Tnghiệp ĐH	
Đọc theo thứ tự từ trang đầu đến trang cuối	n	18	77	18	35	148
	Cột %	30.0%	35.8%	19.8%	26.1%	29.6%
Xem lướt qua các đề mục, đọc các mục ưa thích trước	n	23	95	57	77	252
	Cột %	38.3%	44.2%	62.6%	57.5%	50.4%
Chỉ đọc các trang mục ưa thích, ít đọc các trang khác	n	9	20	9	8	46
	Cột %	15.0%	9.3%	9.9%	6.0%	9.2%
Xem các tin đáng chú ý trên trang 1 và tìm đọc trước	n	10	23	7	14	54
	Cột %	16.7%	10.7%	7.7%	10.4%	10.8%
Tổng	n	60	215	91	134	500
	Cột %	100%	100%	100%	100%	100%

Nhưng mục tiêu nghiên cứu của chúng ta không phải là các mẫu mà là tổng thể, để biết được kết quả trên mẫu có đủ mạnh để thuyết phục chúng ta rằng nó cũng đúng với tổng thể hay không, chúng ta phải tìm bằng chứng thống kê, do đó chúng ta thực hiện các phép kiểm định.

Với ví dụ ở Bảng 4.1 chúng ta sẽ tiến hành kiểm định về mối liên hệ giữa hai biến "cách đọc các tờ báo nói chung" (*c6.1*) và "trình độ học vấn" (*nhomhv*) xem thử có tồn tại mối liên hệ giữa hai biến này trong tổng thể không.

Trong các phép kiểm định thì kiểm định về mối liên hệ giữa các biến là một phép kiểm định hay được sử dụng trong phân tích thống kê, kiểm định này còn gọi là kiểm định tính độc lập. Trong chương này chúng ta sẽ phân biệt 2 tình huống kiểm định mối liên hệ giữa 2 biến định tính như sau:

- Trường hợp dữ liệu định danh-định danh hoặc định danh-thứ bậc (chú ý rằng biến định lượng rời rạc với vài trị số cũng có thể xem như biến định tính)
- Trường hợp dữ liệu thứ bậc-thứ bậc

1. KIỂM ĐỊNH MỐI LIÊN HỆ GIỮA HAI BIẾN ĐỊNH DANH - ĐỊNH DANH HOẶC ĐỊNH DANH - THỨ BẬC

Khi hai yếu tố nghiên cứu đều là biến định danh hay một định danh - một thứ bậc thì kiểm định Chi-bình phương (χ^2) được sử dụng rất phổ biến. Kiểm định Chi-bình phương sẽ cho bạn biết có tồn tại mối liên hệ giữa hai biến trong tổng thể hay không. Tuy nhiên Chi-bình phương không cho biết độ mạnh của mối liên hệ giữa hai biến.

Bạn sử dụng χ^2 để kiểm định giả thuyết H_0 (Null Hypothesis) là: "không có mối liên hệ giữa hai biến" (hay "hai biến độc lập với nhau"). Ví dụ bạn giả định rằng không có mối liên hệ giữa hai biến *c6.1* (định danh) và *nhomhv* (thứ bậc) trong Bảng 4.1. Với thủ thuật dùng giả thuyết H_0, bạn sẽ cố gắng để chứng minh rằng bạn đã sai lầm, sự thật là có tồn tại mối liên hệ giữa "cách đọc các tờ báo nói chung" và "trình độ học vấn", có nghĩa là bạn cố gắng tìm bằng chứng để có thể bác bỏ H_0. Chính vì vậy bạn khi tiếp cận với vấn đề bằng cách dùng giả thuyết H_0 bạn cần hết sức cảnh giác vì lúc này rõ ràng bạn đang rất háo hức tìm kiếm những mối liên hệ, và bạn tiến hành kiểm định với ý tưởng mong muốn thấy kết quả như đã dự định trước rằng thực ra có mối liên hệ giữa hai biến trong tổng thể, do đó bạn sẽ có thể tìm cách biện hộ để bác bỏ giả thuyết Ho.

1.1 Tóm tắt lý thuyết Kiểm định Chi-bình phương

1. Đặt giả thuyết thống kê

Giả thuyết không H_o: hai biến độc lập với nhau

Giả thuyết đối H_1 : hai biến có liên hệ với nhau

2. Tính toán đại lượng $\chi 2$

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Trong đó:

χ^2 : đại lượngChi-bình phương dùng để kiểm định

O_{ij} : đại diện cho số trường hợp được quan sát trong một ô cụ thể của bảng chéo (tần số quan sát)

E_{ij} : đại diện cho số trường hợp mà bạn mong đợi gặp trong những ô của bảng chéo đó nếu không có mối liên hệ giữa 2 biến trong bảng (tần số mong đợi)

c : số cột của bảng

r : số hàng của bảng

E_{ij} được tính theo công thức sau: $E_{ij} = \frac{R_i \times C_j}{n}$

R_i: tổng số quan sát của hàng thứ i

C_j: tổng số quan sát của cột thứ j

Từ công thức tính χ^2 có thể thấy ngay là $\chi^2 = 0$ khi tất cả các tần số quan sát bằng với các tần số mong đợi, nghĩa là lúc này không có mối liên hệ nào giữa các biến. Mặc dù Chi-bình phương có thể = 0, nó không bao giờ nhận giá trị âm. O khác biệt E càng nhiều, thì giá trị χ^2 tính được càng lớn, nghĩa là lúc này có khả năng có mối liên hệ giữa 2 biến.

3. Tìm giá trị giới hạn $\chi^2_{(r-1)(c-1),\alpha}$

Đại lượng kiểm định này có phân phối Chi-bình phương nên bạn tra bảng phân phối χ^2 để tìm được giá trị giới hạn với mức ý nghĩa α và số bậc tự do df =(r – 1).(c -1). Mức ý nghĩa α là khả năng tối đa cho phép phạm phải sai lầm loại I trong kiểm định, tức khả năng bạn bác bỏ Ho mặc dù thực tế Ho đúng. Nếu cho α = 5% nghĩa là khi thực hiện kiểm định bạn chấp nhận một khả năng phạm sai lầm loại I tối đa là 5%, từ đó độ tin cậy được của kiểm định của bạn là (1-α) = 95%.

4. Tiêu chuẩn quyết định là so sánh giá trị giới hạn và đại lượng χ^2:

Bác bỏ giả thuyết H_o nếu : $\chi^2 \rangle \chi^2_{(r-1)(c-1),\alpha}$

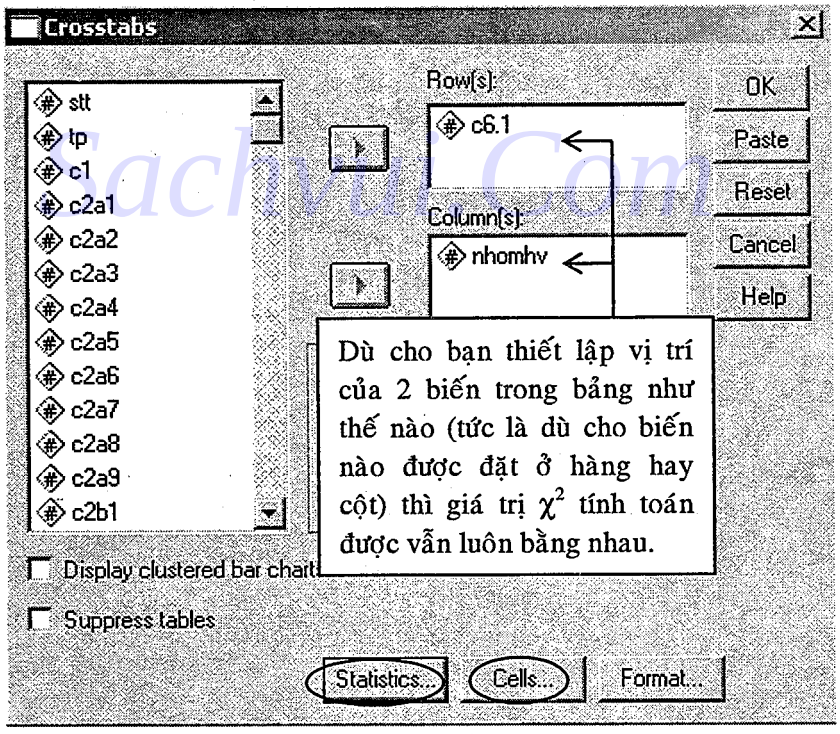
Chấp nhận giả thuyết H_o nếu : $\chi^2 \leq \chi^2_{(r-1)(c-1),\alpha}$

1.2 Sử dụng SPSS thực hiện kiểm định Chi-bình phương

Để nghiên cứu mối liên hệ giữa trình độ học vấn và cách đọc các tờ báo của người đọc bằng kiểm định Chi-bình phương, ta lập bảng chéo (Crosstabs) để tìm hiểu mối quan hệ này.

Từ menu, chọn Analyze > Descriptive Statistics > Crosstabs ...Lệnh này mở ra hộp thoại Crosstabs như hình sau:

Hình 4.1



Trong hộp thoại này, nếu đưa biến thứ bậc *nhomhv* (nhóm học vấn) vào ô cột và biến định danh *c6.1* (cách đọc các tờ báo) ô dòng, nhấn nút OK, ta được một bảng kết hợp của 2 biến trên mà trong các ô là tần số quan sát giống như bảng 2 biến được lập bằng lệnh Basic Tables (như vậy bạn đã biết thêm một cách thứ 2 để lập bảng kết hợp 2 biến định tính là dùng lệnh Crosstabs)

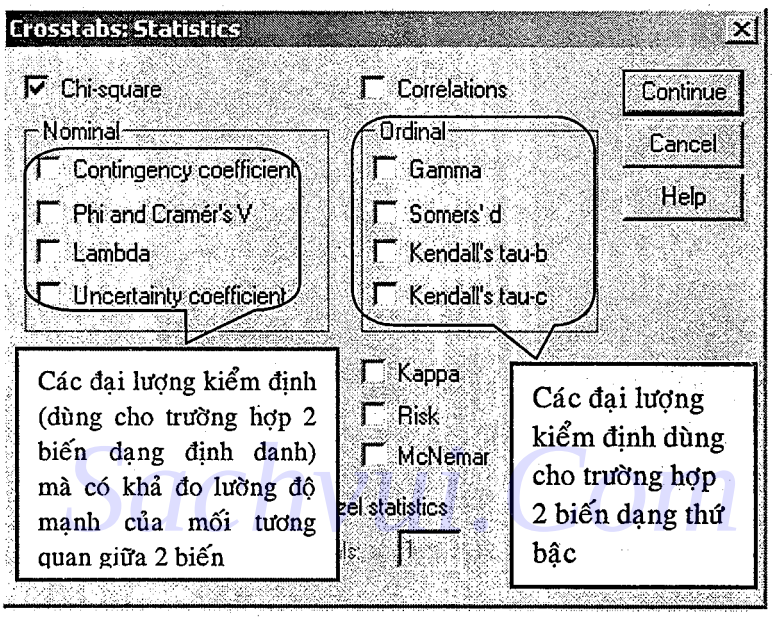
Để kiểm định giả thiết về mối liên hệ giữa *nhomhv* và *c6.1*, ta đặt giả thuyết H_o như sau

H_o: Học vấn **không** có liên hệ với cách đọc báo

(Cách đọc báo **không** chịu ảnh hưởng của học vấn)

Bạn mở lại hộp thoại Crosstab, từ trong hộp thoại Crosstabs nhấn nút Statistics..., hộp thoại Crosstab: Statistics như sau xuất hiện:

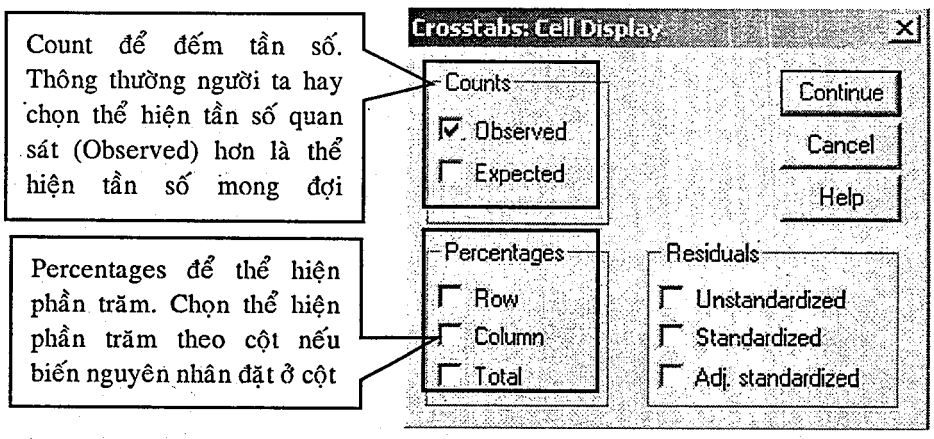
Hình 4.2



Trong hộp thoại này ta đánh dấu chọn đại lượng Chi-square rồi nhấp nút Continue để trở về hộp thoại Crosstabs. Trong hộp thoại Crosstabs nhấn tiếp nút Cells... để mở hộp thoại Crosstabs: Cell Display nhằm xác định các đại lượng thống kê thể hiện trong từng ô của bảng chéo. (Hình 4.3)

Lựa chọn xong bạn nhấn Continue, cuối cùng là OK. Kết quả do lệnh Crosstabs đưa ra gồm 3 bảng: bảng đầu tiên thể hiện những thông tin tổng hợp, thứ 2 là bảng chéo kết hợp 2 biến mà trong các ô thể hiện đại lượng thống kê bạn đã chọn ở hộp thoại Crosstabs: Cell Display và cuối cùng là bảng tóm lược kết quả kiểm định χ^2 (Bảng 4.2; 4.3 và 4.4)

Hình 4.3



Bạn sẽ đọc kết quả kiểm định ở dòng đầu tiên Pearson Chi-Square của Bảng 4.4. Tra bảng Chi-bình phương tìm giá trị giới hạn ở bậc tự do 9 và mức ý nghĩa 0,05 (vì bạn đã chọn độ tin cậy của kiểm định này là 95%) rồi so sánh giá trị Chi-bình phương tính toán được 20,238 với giá trị giới hạn này: $\chi^2_{(4-1).(4-1);\,0,05} = 16,9190 < 20,238$.

Theo tiêu chuẩn quyết định, chúng ta sẽ bác bỏ giả thuyết H_0 và kết luận rằng Học vấn có ảnh hưởng đến cách đọc báo của người đọc.

Bảng 4.2 **Case Processing Summary**

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
cách đọc các tờ báo nói chung * học vấn	500	100.0%	0	.0%	500	100.0%

Bảng 4. 3 **cách đọc các từ báo nói chung * học vấn Crosstabulation**

		học vấn				Total
		cấp 1-2	cấp 3-THCN	CĐ-SVĐH	Tnghiệp ĐH	
Đọc theo thứ tự từ trang đầu đến trang cuối	Count	18	77	18	35	148
	% within học vấn	30.0%	35.8%	19.8%	26.1%	29.6%
Xem lướt qua các đề mục, đọc các mục ưa thích trước	Count	23	95	57	77	252
	% within học vấn	38.3%	44.2%	62.6%	57.5%	50.4%
Chỉ đọc các trang mục ưa thích, ít đọc các trang khác	Count	9	20	9	8	46
	% within học vấn	15.0%	9.3%	9.9%	6.0%	9.2%
Xem các tin đáng chú ý trên trang 1 và tìm đọc trước	Count	10	23	7	14	54
	% within học vấn	16.7%	10.7%	7.7%	10.4%	10.8%
Total	Count	60	215	91	134	500
	% within học vấn	100%	100%	100%	100%	100%

Bảng 4.4 **Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	20.238(a)	9	.017
Continuity Correction	.		
Likelihood Ratio	20.134	9	.017
Linear-by-Linear Association	.138	1	.710
N of Valid Cases	500		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.52.

Có một nguyên tắc khác hay được sử dụng trong kiểm định giả thuyết là dùng giá trị p-value. P-value là xác suất bạn sẽ phạm sai lầm loại I- nghĩa là xác suất loại bỏ giả thuyết H_o với những thông tin bạn tính toán được, như vậy nó có cùng ý nghĩa với mức ý nghĩa α. Xác suất này càng cao cho thấy hậu quả của việc phạm sai lầm khi loại bỏ giả thuyết Ho càng nghiêm trọng (và ngược lại) như vậy quy tắc chung là không bác bỏ Ho nếu p-value quá lớn. Với quy tắc này bạn không cần phải mất công tra bảng tìm giá trị tới hạn mà chỉ cần xem xét độ lớn của p-value rồi ra quyết định như sau:

- Nếu p-value < 0,1 thì kiểm định của bạn có ý nghĩa với độ tin cậy 90% (khái niệm "có ý nghĩa" được hiểu là giả thuyết Ho có thể bị bác bỏ với độ tin cậy 90%)

- Nếu p-value < 0,05 thì kiểm định có ý nghĩa với độ tin cậy 95% (khái niệm có ý nghĩa được hiểu là giả thuyết Ho có thể bị bác bỏ với độ tin cậy 95%). Đây là điều kiện thường được sử dụng.
- Nếu p-value < 0,01 thì kiểm định có ý nghĩa với độ tin cậy 99% (khái niệm có ý nghĩa được hiểu là giả thuyết Ho có thể bị bác bỏ với độ tin cậy 99%).

SPSS gọi p-value là Sig. (viết tắt từ Observed significance level là mức ý nghĩa quan sát). Lúc này thay vì bạn phải tra bảng Chi-bình phương để tìm giá trị tới hạn rồi so sánh giá trị Chi-bình phương tính toán với giá trị này thì SPSS đã tính ngược lại mức ý nghĩa quan sát Sig. ứng với giá trị Chi-bình phương tính toán được 20,238. Ở đây Sig.= 0,017 (hay 1,7%).

Từ quy tắc của p-value, bạn sẽ quyết định theo nguyên tắc:

- Chấp nhận Ho nếu sig. > α, vì nếu ta bác bỏ Ho thì khả năng phạm sai lầm của ta sẽ lớn hơn mức ý nghĩa cho phép
- Bác bỏ Ho nếu sig. < α vì lúc này xác suất phạm sai lầm nếu bác bỏ Ho nhỏ hơn mức cho phép nên có thể an toàn khi bác bỏ Ho

Vì ở đây Sig. < α nên ta bác bỏ giả thuyết Ho. Ta kết luận rằng với tập dữ liệu mẫu, có đủ bằng chứng để nói rằng trình độ học vấn có liên hệ với cách đọc báo. Chúng ta có thể dựa vào các tỉ lệ phần trăm theo cột trong bảng chéo (Bảng 4.3) để mô tả sự liên hệ hay sự khác biệt về cách đọc báo giữa các nhóm học vấn. Quy ước chung là tính phần trăm từ trên xuống, đọc và so sánh theo hàng ngang. Để thấy được các % này bạn chọn mục Percentage Column trong hộp thoại Crosstabs: Cell Display.

Giải thích về các đại lượng trên Bảng 4.4

- Kiểm định Chi-bình phương chỉ có ý nghĩa khi số quan sát đủ lớn, nếu có quá 20% số ô trong bảng chéo có tần số lý thuyết nhỏ hơn 5 thì giá trị χ^2 nói chung không còn đáng tin cậy. Lúc này bạn phải nghĩ đến biện pháp gom các biểu hiện trên các biến lại để tăng số quan sát trong mỗi nhóm lên, phần Recode ở Chương I có hướng dẫn cách thực hiện điều này. Cuối bảng Chi-square Tests (Bảng 4.4) SPSS luôn đưa ra 1 dòng thông báo cho bạn biết % số ô có tần suất mong đợi dưới 5 của bảng, trong ví dụ của chúng ta không có ô

nào có tần suất mong đợi dưới 5 nên ta có thể tin tưởng vào độ chính xác của kiểm định.

- Continuity Correction là một dạng biến thể của Pearson Chi-Square để sử dụng cho những bảng dạng 2x2, tức là bảng kết hợp của 2 biến mà mỗi biến đều chỉ cơ 2 biểu hiện.
- Likelihood Ratio là một số thống kê tương tự Pearson Chi-Square, với những cỡ mẫu lớn kết quả của 2 số thống kê này rất gần nhau.
- Linear-by-Linear Association đo lường mối liên hệ tuyến tính giữa 2 biến, số thống kê này chỉ hữu dụng khi biến hàng và cột được sắp trật tự từ nhỏ nhất đến lớn nhất, còn nếu không bạn hãy bỏ qua nó.
- Ngoài ra còn một số thống kê nữa không được thể hiện ở Bảng 4.4 là kết quả kiểm định Fisher's Exact. Fisher's Exact test rất phù hợp cho dạng bảng 2x2 với tình huống các tần số mong đợi tại các ô nhỏ hơn 5.Vì thế khi bạn lập bảng 2x2 thì SPSS mới cung cấp thông tin của kiểm định Fisher's Exact cùng với các kết quả của các kiểm định khác.

1.3 Một số đại lượng thống kê khác về mối liên hệ giữa 2 biến định danh

Kiểm định Chi-bình phương được sử dụng phổ biến nhất trong kiểm định mối liên hệ giữa 2 biến định danh- định danh hay định danh-thứ bậc. Tuy nhiên nó không cho biết độ mạnh của mối liên hệ đó, lúc đó bạn sẽ phải nhờ đến Lambda, Cramer V, hay hệ số liên hợp (Coefficient of contigency) do Pearson đề xuất. Các đại lượng này cũng được cung cấp trong SPSS (xem Hình 4.2) và được tóm tắt dưới đây. Trong khuôn khổ quyển sách phân tích dữ liệu, chúng tôi không đi sâu vào công thức, mà chỉ giới thiệu chúng vì việc sử dụng chúng khá đơn giản.

1.3.1 Cramer V

Cramer V được tính dựa trên Chi-bình phương và là một kiểm định trực tiếp mối liên hệ của 2 biến. Khi biết χ^2 của một bảng, bạn có thể tính toán Cramer V một cách dễ dàng từ công thức:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} \qquad 0 \leq V < 1$$

- k là số hàng hoặc số cột trong bảng, ta chọn k nào nhỏ hơn, ví dụ nếu bạn có 3 hàng và 4 cột thì k = 3 và (k–1) = 2. Bài toán của ta có số hàng bằng số cột và = 4 nên k = (4-1)=3
- N là số quan sát trong mẫu

Cramer V cho biết độ mạnh của mối liên hệ giữa các biến định danh.

1.3.2 Hệ số liên hợp (Coefficient of contigency)

Cũng là một chỉ số đánh giá mức độ tương quan giữa 2 biến, công thức tính của nó là

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \qquad \text{(N là quy mô mẫu)}$$

C = 0 khi giữa 2 biến không có quan hệ và $0 \leq C \leq 1$.

1.3.3. Lambda

Lambda (viết tắt là *L* hay λ) là một phép đo lường liên hệ của các biến định danh cung cấp cho người tiến hành những cảm nhận dễ dàng và khả năng giải thích trực tiếp.

Lambda cho biết liệu các trị số của một biến có xu hướng tập trung quanh một số trị số nào đó của biến kia không. Nếu có thì khi biết được trị số của biến độc lập ta có thể dự đoán được trị số của biến phụ thuộc. Chẳng hạn như nếu biết được đối tượng có học vấn cấp 3 thì có thể dự đoán cách thức đọc báo của đối tượng tốt hơn là dự đoán mà không biết gì về học vấn của người này. Ví dụ một λ bằng 0,33 cho thấy rằng nếu bạn biết các giá trị của biến độc lập, bạn có thể dự đoán được giá trị của biến phụ thuộc tốt hơn 33% khi bạn không biết gì về biến độc lập. Công thức của Lambda như sau:

$$\lambda = \frac{\text{sai số trước} - \text{sai số sau}}{\text{sai số trước}}$$

Trong đó sai số trước là số các sai lầm có thể phạm phải khi dự đoán các trị số của biến phụ thuộc mà không xem xét đến biến độc lập. Sai số sau là số các sai lầm có thể phạm phải khi dự đoán các trị số của biến phụ thuộc có xem xét đến biến độc lập. Dự báo tốt nhất cho các trị số của biến phụ thuộc khi không xem xét đến trị số của biến độc lập là số mode của biến phụ thuộc. Dự báo tốt nhất cho các trị số của biến phụ thuộc khi có xem xét trị số của biến độc lập là số mode của từng phân nhóm của biến độc lập

Tuy nhiên có một số vấn đề sau về lambda

- Số thống kê λ phụ thuộc vào vị trí bạn thiết lập biến phụ thuộc trong hàng hay cột của bảng.
- Không có cách để kiểm định một giá trị Lambda
- Lambda có thể = 0 (chỉ ra không có mối liên hệ giữa các biến), ngay cả khi có mối liên hệ rõ ràng và mạnh giữa các biến. Điều này đặc biệt hay xảy ra cho bảng 2x2 với hơn 50% quan sát trên biến độc lập được chứa trong các ô cho cùng một biểu hiện của biến phụ thuộc.

2. KIỂM ĐỊNH MỐI LIÊN HỆ GIỮA 2 BIẾN THỨ BẬC

Trong trường hợp hai yếu tố nghiên cứu là hai biến thu thập từ thang đo thứ bậc, thay vì dùng đại lượng Chi-bình phương, chúng ta có thể dùng một trong các đại lượng: tau-b của Kendall, d của Somer, gamma của Goodman và Kruskal. Các đại lượng này giúp phát hiện ra mối liên hệ tốt hơn Chi-bình phương

Ví dụ: chúng ta cần nghiên cứu mối liên hệ giữa tuổi tác và mức độ quan tâm đối với chủ đề gia đình trên báo Sài Gòn Tiếp Thị. Cả hai yếu tố này đều là dữ liệu thứ bậc vì nó được phân hạng như sau:

- Độ tuổi: (18-25) tuổi; (26-35) tuổi; (36-45) tuổi; (46-60) tuổi.
- Mức độ quan tâm đến chủ đề gia đình: quan tâm nhất, quan tâm thứ nhì, quan tâm thứ ba.

Trước tiên ta lập bảng Crosstabs biểu diễn mối quan hệ giữa tuổi tác và mức độ quan tâm đến chủ đề gia đình trên báo SGTT. Kết quả thể hiện trong Bảng 4.5.

Bảng 4.5 **Gia đình * độ tuổi Crosstabulation**

	độ tuổi				Total
	18-25	26-35	36-45	46-60	
Quan tâm nhất	8	7	5	2	22
Quan tâm nhì	21	16	13	8	58
Quan tâm ba	15	14	10	8	47
Total	44	37	28	18	127

Để thực hiện kiểm định mối liên hệ trong tình huống này ta đặt giả thuyết H_0 : Tuổi tác không có liên hệ với mức độ quan tâm đến chủ đề gia đình trên báo SGTT (hay mức độ quan tâm đến chủ đề gia đình không khác nhau giữa các nhóm tuổi)

2.1. Gamma của Goodman và Kruskal

Nếu 2 biến thứ bậc này có mối liên hệ chặt chẽ, ta có thể tin rằng tất cả những người đọc có thứ hạng cao về tuổi tác (tức trẻ tuổi hơn) sẽ quan tâm nhiều hơn đến chủ đề gia đình trên báo SGTT, mọi người đọc có thứ hạng tuổi tác thấp hơn (già hơn) thì thứ hạng của mức độ quan tâm cũng thấp... Tất nhiên thực tế sẽ không hoàn toàn diễn ra giống hệt như vậy, nhưng từ những cặp kết hợp có thể hình thành giữa các bậc của 2 biến trong Bảng 4.5 bạn có thể xác định được một thước đo về mối liên hệ của 2 biến thứ bậc này, thước đo đó là gamma. Gamma là một thước đo phổ biến và dễ cảm nhận, vì trị số của nó nằm trong khoảng từ –1 (liên hệ nghịch hoàn toàn) đến +1 (liên hệ thuận hoàn toàn), giá trị 0 ở trung tâm đại diện cho sự độc lập hoàn toàn giữa hai biến.

Một lần nữa bạn đừng quên gamma ta tính được là dựa trên thông tin của mẫu, nên nó thể hiện độ mạnh của một mối liên hệ có thể chỉ có trong mẫu. Để chắc điều đó đúng với tổng thể chúng ta phải kiểm định ý nghĩa của gamma

Chúng ta xuất phát với giả thuyết H_0 rằng gamma của tổng thể chung = 0, nghĩa là thật sự không có mối liên hệ giữa các biến thứ bậc trong tổng thể chúng ta đang nghiên cứu. Nếu kết quả kiểm định cho phép chúng ta bác bỏ giả thiết H_0, thì chúng ta có thể kết luận rằng 2 biến thứ bậc của chúng ta có mối liên hệ và giá trị gamma của mẫu mà ta tính được chắc chắn xấp xỉ giá trị gamma của tổng thể chung (ký hiệu γ là Gamma của tổng thể chung).

2.2. tau-b của Kendall (τ_b)

Gamma được ưa thích vì là một số thống kê dễ hiểu, dễ dàng giải thích và đánh giá bằng cách sử dụng bảng phân phối z. Nhưng nhiều nhà nghiên cứu không thích gamma vì nó có xu hướng thổi phồng mối liên hệ giữa các biến bởi những dữ liệu nó bỏ qua không xét đến trong quá trình tính toán (do quy tắc tính của gamma). Trong khi đó tau-b sử dụng hầu hết dữ liệu nên sẽ gần như luôn luôn nhỏ hơn gamma vì vậy nó

đáng tin cậy hơn khi đo lường mối liên hệ. Tuy nhiên rất khó kiểm tra mức ý nghĩa của tau-b trong lĩnh vực nghiên cứu, công thức tính nó lại khó và không sẵn bảng để bạn có thể tra cứu.

Nói chung, khi tính toán với bất kỳ bảng chéo nào thì các chỉ số tau-b của Kendall, d của Somer, gamma của Goodman và Kruskal có trị số tương tự nhau (tức là tất cả đều = 0 khi hai biến không liên hệ, và cùng có dấu giống nhau khi chỉ chiều hướng của mối tương quan) nhưng không giống hệt nhau theo cách chúng tóm lược nội dung của bảng chéo, chỉ số tau-b thích hợp hơn cho những bảng cân đối tức là có số hàng bằng số cột còn tau-c thích hợp cho những bảng không cân đối, còn trị số gamma thường cao hơn các số thống kê khác nên có thể dẫn ta đến sai lầm là ước lượng quá cao về độ mạnh của mối liên hệ.

Trong phạm vi cuốn sách này chúng ta không tìm hiểu sâu hơn về bản chất của gamma cũng như một số thước đo mối kết hợp khác cho dữ liệu thứ bậc mà ta sẽ tập trung vào kết quả do SPSS đưa ra (bạn đọc có thể tìm hiểu chi tiết về các đại lượng này trong các sách thống kê hay sách về phương pháp nghiên cứu).

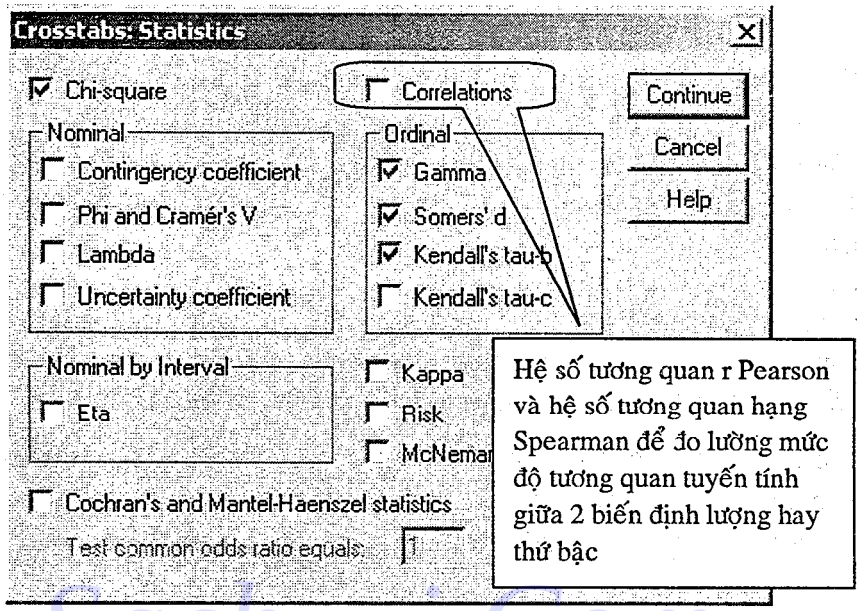
2.3. Vận dụng SPSS để thực hiện kiểm định

Để kiểm định giả thiết Ho đề ra ở trên, mở lại hộp thoại Crosstabs. Trong hộp thoại này đưa biến *c19.3* (mức độ quan tâm đến chủ đề gia đình) vào ô Row và biến *dotuoi* (nhóm tuổi) vào ô Column. Rồi chọn nút Statistics...

Trong hộp thoại Statistics, ta chọn các đại lượng kiểm định như Hình 4.4. Sau đó nhấp Continue trở về hộp thoại Crosstab và nhấp OK.

Chú ý là trong các bảng kết quả kiểm định SPSS đưa ra cũng có bảng Chi-square test (Bảng 4.8) và bạn có thể so sánh độ mạnh của các loại kiểm định cho trường hợp kiểm định thực tế của bạn.

Hình 4.4



Kết quả kiểm định xuất hiện (từ Bảng 4.6 đến Bảng 4.10). Trong Bảng 4.10, chúng ta thấy gamma = 0,077. SPSS tra ngược bảng giá trị z (bảng này liệt kê phần diện tích ở dưới đường cong chuẩn được xác định bởi những giá trị chuẩn hoá z khác nhau) để tìm phần diện tích nằm dưới đường cong chuẩn giữa trung bình và giá trị z, từ đó nó suy ra phần diện tích dưới đường cong mà chính là giá trị p-value.

Với mức ý nghĩa Sig. = 0,498 > 0,05 ta không thể bác bỏ giả thuyết Ho. Có thể kết luận rằng với dữ liệu mẫu ta có thì không đủ bằng chứng thống kê cho thấy tuổi tác có liên quan đến mức độ quan tâm đến chủ đề gia đình. Như vậy có lẽ ở bất kỳ độ tuổi nào người đọc cũng có những mối quan tâm nhất định đối với chủ đề gia đình. Dùng tau-b ta cũng đi đến kết luận tương tự.

Bảng 4.6 **Case Processing Summary**

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gia đình * độ tuổi	127	25.4%	373	74.6%	500	100.0%

Bảng 4.7 **Gia đình * độ tuổi Crosstabulation**

		độ tuổi				Total
		18-25	26-35	36-45	46-60	
Quan tâm nhất	Count	8	7	5	2	22
	% within độ tuổi	18.2%	18.9%	17.9%	11.1%	17.3%
Quan tâm nhì	Count	21	16	13	8	58
	% within độ tuổi	47.7%	43.2%	46.4%	44.4%	45.7%
Quan tâm ba	Count	15	14	10	8	47
	% within độ tuổi	34.1%	37.8%	35.7%	44.4%	37.0%
Total	Count	44	37	28	18	127
	% within độ tuổi	100%	100%	100%	100%	100%

Bảng 4.8 **Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.966(a)	6	.987
Continuity Correction			
Likelihood Ratio			.985
Linear-by-Linear Association			.474
N of Valid Cases			

Giá trị sig. từ kiểm định χ^2 trong trường hợp này lớn hơn giá trị sig. mà gamma đưa ra rất nhiều.

a 2 cells (16.7%) have expected count less than 5. The minimum expected count is 3.12.

Bảng 4.9 **Directional Measures**

			Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	.051	.076	.678	.498
		Gia đình Dependent	.048	.070	.678	.498
		độ tuổi Dependent	.056	.082	.678	.498

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

Bảng 4.10 **Symmetric Measures**

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	.052	.076	.678	.498
	Gamma	.077	.113	.678	.498
N of Valid Cases		127			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

Một lần nữa bạn hãy nhớ rằng đừng để các biểu hiện trong mẫu đánh lừa bạn, chỉ qua kiểm định thống kê bạn mới có thể có kết luận về tổng thể. Và kết luận cuối cùng của bạn cũng chỉ là một kết luận thống kê dựa trên những dữ liệu mẫu bạn thu thập được với những nguồn sai số mà bạn không thể kiểm soát hoàn toàn (bởi vì vậy chúng ta luôn tiến hành kiểm định với một độ tin cậy nhất định). Hoài nghi một cách lành mạnh các kết quả thống kê là một thói quen tốt vì nó khuyến khích chúng ta không ngừng tìm tòi, chứng minh và có được những phát hiện mới.

CHƯƠNG V

PHÂN TÍCH LIÊN HỆ GIỮA BIẾN NGUYÊN NHÂN ĐỊNH TÍNH VÀ BIẾN KẾT QUẢ ĐỊNH LƯỢNG:

KIỂM ĐỊNH TRUNG BÌNH TỔNG THỂ

Trong thống kê có các phép kiểm định về trị trung bình của tổng thể phổ biến sau:

- Nếu muốn so sánh trị trung bình của một tổng thể với một giá trị cụ thể nào đó ta sẽ thực hiện phép kiểm định giả thuyết về trung bình của tổng thể. Trong SPSS có thể sử dụng lệnh One-Sample T-Test để thực hiện kiểm định này (vào menu Analyze > Compare Means > One-Sample T-Test).
- Nếu muốn so sánh hai trị trung bình của hai nhóm tổng thể riêng biệt ta thực hiện phép kiểm định giả thuyết về sự bằng nhau của 2 trung bình tổng thể dựa trên hai mẫu độc lập rút từ hai tổng thể này. SPSS sử dụnglệnh Independent-Samples T-Test thuộc menu Analyze > Compare Means để thực hiện kiểm định này cho bạn.
- Nếu muốn so sánh hai trị trung bình cuả hai nhóm tổng thể riêng biệt có đặc điểm là mỗi phần tử quan sát trong tổng thể này có sự tương đồng theo cặp với 1 phần tử ở tổng thể bên kia ta sử dụng kiểm định giả thuyết về sự bằng nhau của hai trung bình tổng thể dựa trên dữ liệu mẫu rút từ hai tổng thể theo cách phối hợp từng cặp. Ta tiến hành kiểm định này bằng lệnh Analyze > Compare Means > Paired-Samples T-Test của SPSS.
- Nếu muốn mở rộng sự so sánh cho trị trung bình của nhiều nhóm tổng thể độc lập ta sử dụng phương pháp kiểm định giả thuyết về sự bằng nhau của trung bình nhiều tổng thể. Phương pháp kiểm định này có tên gọi phổ biến là phân tích phương sai (ANOVA). Ta có thể sử dụng lệnh One-way ANOVA cũng thuộc Analyze > Compare Means của SPSS để tiến hành kiểm định này.

1. KIỂM ĐỊNH GIẢ THUYẾT VỀ TRỊ TRUNG BÌNH CỦA MỘT TỔNG THỂ

Ví dụ: có người cho rằng tuổi trung bình của độc giả báo Sài Gòn tiếp thị (SGTT) là 30 tuổi, với dữ liệu có được là file *Data Thuc hanh* bạn sẽ làm như thế nào để kiểm định giả thuyết này?

Trình tự thực hiện:

1. Đặt giả thuyết Ho: Tuổi trung bình của độc giả báo SGTT =30

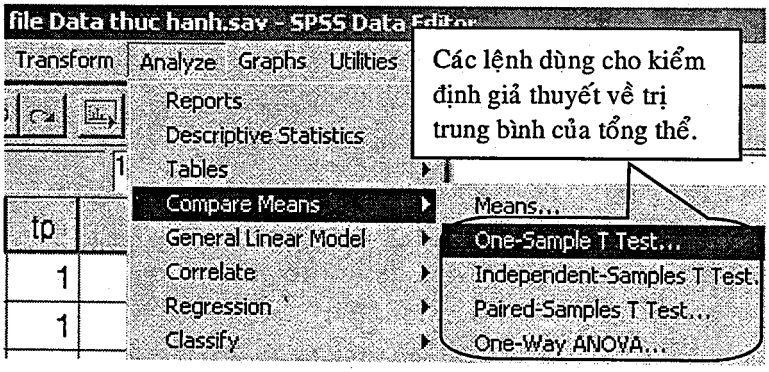
2. Bạn dùng lệnh Count để chuyển các biến Category từ *c2a.1* đến *c2a.9* thành biến Dichotomy tên là *docSGTT* với biểu hiện 1 là người có đọc SGTT và 0 là người không đọc SGTT mà đọc các báo khác.

3. Dùng lệnh Select Case lọc ra các trường hợp mà *docSGTT* nhận giá trị 1 để các lệnh thống kê sau đó của SPSS chỉ thực hiện trên các trường hợp này, tức là các trường hợp người đọc là độc giả của SGTT.

Với file thực hành của chúng ta, mẫu các độc giả của SGTT sẽ gồm 159 người.

4. Để thực hiện kiểm định giả thuyết về tuổi trung bình của độc giả SGTT bạn vào Analyze > Compare Means > One-Sample T-Test

Hình 5.1

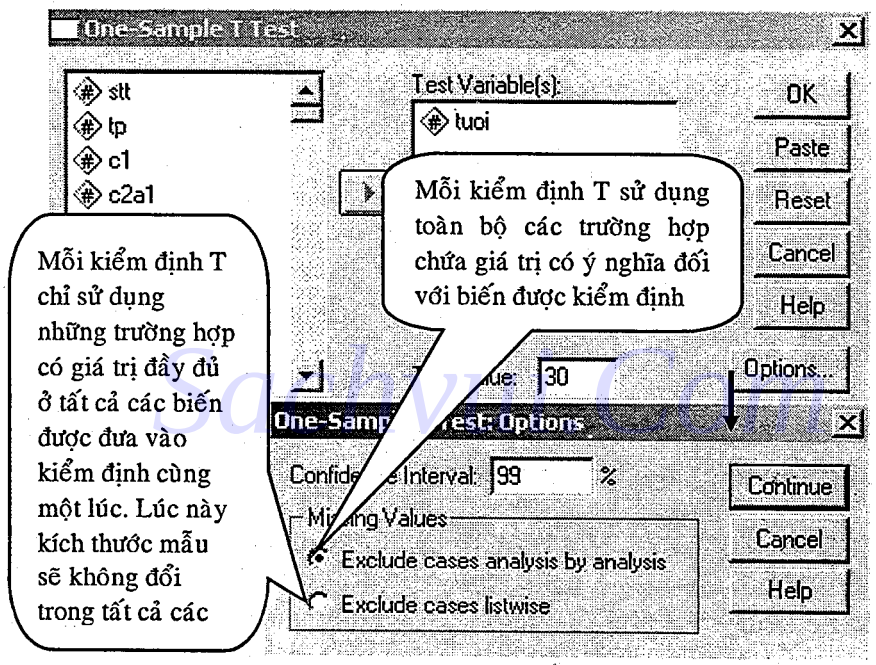


5. Khi mở được hộp thoại One-Sample T-Test (Hình 5.2) bạn đưa biến *tuoi* vào khung Test Variable, khai báo Test Value là 30 (đó là giá trị bạn muốn so sánh với tuổi trung bình của những độc giả SGTT)

6. Bạn nhấp nút Options...để chọn độ tin cậy cho ước lượng khoảng cho trung bình tổng thể, SPSS mặc định chọn cho bạn 95%, bạn có thể tăng hoặc giảm độ tin cậy của phép ước lượng theo ý bạn, ở đây ta thử chọn độ tin cậy 99% (một độ tin cậy khá cao). Bạn luôn nhớ rằng độ tin cậy hay mức ý nghĩa của bài toán thống kê hoàn toàn là do bạn lựa chọn, SPSS không chịu trách nhiệm về việc bạn quyết định lựa chọn hay bác bỏ giả thuyết Ho căn cứ trên mức ý nghĩa (hay độ tin cậy) bạn đã chọn.

7. Nhấn nút Contiune trở lại hộp thoại trước rồi OK. SPPS tạo ra cho bạn các bảng kết quả là Bảng 5.1 và 5.2 dưới đây.

Hình 5.2



Bảng 5.1 **One-Sample Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
tuổi	159	32.79	10.328	.819

Bảng 5.2 **One-Sample Test**

	Test Value = 30					
	t	df	Sig. (2-tailed)	Mean Difference	99% Confidence Interval of the Difference	
					Lower	Upper
tuổi	3.402	158	.001	2.79	.65	4.92

Theo mẫu của chúng ta, tuổi trung bình của các độc giả SGTT là 32,79 tuổi. Giá trị của kiểm định t về tuổi trung bình của độc giả SGTT là 3,402 ứng với mức ý nghĩa quan sát 0,001; nhỏ hơn rất nhiều so với mức ý nghĩa 0,01. Như vậy là nếu bạn bác bỏ giả thuyết Ho về tuổi trung bình bạn có nguy cơ phạm sai lầm rất thấp, và thấp dưới mức ý nghĩa bạn đã chọn cho kiểm định này vì vậy bạn có thể yên tâm bác bỏ giả thuyết Ho rằng tuổi trung bình của độc giả báo SGTT là 30 tuổi. Căn cứ trung bình mẫu và kết quả kiểm định vừa rồi, có thể nói rằng tuổi trung bình của những người thường xuyên đọc báo Sài Gòn Tiếp Thị là trên 30 tuổi.

2. KIỂM ĐỊNH GIẢ THUYẾT VỀ SỰ BẰNG NHAU GIỮA HAI TRUNG BÌNH TỔNG THỂ

2.1 Kiểm định giả thuyết về trị trung bình của hai tổng thể – trường hợp mẫu độc lập (Independent-samples T-test)

Trong nhiều trường hợp bạn cần so sánh trị trung bình về một chỉ tiêu nghiên cứu nào đó giữa hai đối tượng bạn quan tâm. Bạn có 2 biến tham gia trong một phép kiểm định trung bình :1 biến định lượng (tất nhiên là biến định lượng) dạng khoảng cách hay tỉ lệ để tính trung bình, và1 biến định tính dùng để chia nhóm ra so sánh.

Ví dụ: so sánh giữa 2 thành phố Hà Nội và TPHCM về số nhân khẩu trung bình của hộ gia đình. Với ví dụ này, bạn có biến định lượng là *sonk* và biến định tính là *tp* và bạn sẽ sử dụng phép kiểm định sự bằng nhau về trị trung bình của hai tổng thể từ thông tin của 2 mẫu độc lập, từ đây về sau gọi tắt là kiểm định trung bình 2 mẫu độc lập (Independent-samples T-test)

Đối với phép kiểm định Independent-samples T-test, có một nguyên tắc mà trên thực tế hầu như không thể đạt được một cách tuyệt đối là bất kỳ một sự khác biệt nào về trị trung bình tìm được từ kết quả kiểm định là do sự khác biệt từ chính nội tại của mẫu thử chứ không phải do các nguyên nhân khác.

Giả dụ nếu bạn muốn so sánh sự bằng nhau của trị trung bình về thu nhập cá nhân giữa 2 nhóm giới tính trong mẫu của chúng ta, bạn sẽ không thể sử dụng phương pháp kiểm định Independent-samples T-test vì nguyên tắc trên không được tuân thủ, rõ ràng là thu nhập của một cá nhân trước tiên chịu ảnh hưởng lớn của bằng cấp, tính chất công việc, trình độ ngoại ngữ, chức vụ đảm nhiệm chứ không chỉ chịu ảnh hưởng của riêng yếu tố giới tính (nếu cho rằng ta đã chứng minh được thu nhập có bị giới tính tác động).

Khi có nhiều yếu tố gây ra sự khác biệt giữa 2 trị trung bình về thu nhập mà ta quan sát được, thì một biện pháp để làm giảm tối đa ảnh hưởng của các yếu tố riêng lẻ khác là ta chọn 2 nhóm mẫu kiểm định như thế nào cho có sự tương đồng hoàn toàn trong từng cặp quan sát về các yếu tố có khả năng tác động đến vấn đề ta muốn kiểm tra có sự khác biệt hay không. Cách lấy mẫu này gọi là lấy mẫu từng cặp (paired-samples design) khi mỗi đối tượng ở mẫu này có một đối tượng tương ứng trong mẫu kia, với loại mẫu này bạn sẽ áp dụng phép kiểm định sự bằng nhau giữa 2 trung bình dựa trên mẫu phối hợp từng cặp Paired-samples T-test (chúng ta sẽ thảo luận ở mục 2.2 của chương này). Còn với tình huống so sánh số nhân khẩu trung bình của hộ gia đình tại 2 thành phố nói trên, tất cả các quan sát cho từng mẫu xem như được chọn ngẫu nhiên độc lập từ 2 tổng thể có phân phối chuẩn và phương sai bằng nhau, ta gọi cách chọn mẫu này là independent-samples design.

Nếu kiểm định giả thuyết về trị trung bình của hai mẫu độc lập vi phạm giả định là hai mẫu được lấy ngẫu nhiên từ hai tổng thể có phân phối chuẩn thì chúng ta sẽ phải thực hiện một phương pháp kiểm định khác tên là Mann-Whitney sẽ được thảo luận ở Chương VI.

Còn trước khi thực hiện kiểm định trung bình ta cần phải thực hiện một kiểm định khác mà kết quả của nó ảnh hưởng rất quan trọng đến kiểm định trung bình, đó là kiểm định sự bằng nhau của hai phương sai tổng thể. Về mặt trực quan bạn dễ dàng nhận thấy rằng so sánh hai tổng thể có trị trung bình bằng nhau nhưng mức độ phân tán (hay đồng đều) hoàn toàn khác nhau là khập khiễng. Và vì phương sai diễn tả mức độ đồng đều hoặc không đồng đều của dữ liệu quan sát nên bạn phải tiến hành kiểm định sự bằng nhau về phương sai, kiểm định này có tên là Levene test.

Levene test được tiến hành với giả thuyết Ho rằng phương sai của 2 tổng thể bằng nhau, nếu kết quả kiểm định cho bạn mức ý nghĩa quan sát nhỏ hơn 0,05 bạn có thể bác bỏ giả thuyết Ho. Kết quả của việc bác bỏ hay chấp nhận giả thuyết Ho sẽ ảnh hưởng đến việc bạn lựa chọn tiếp loại kiểm định giả thuyết về sự bằng nhau giữa hai trung bình tổng thể nào: kiểm định trung bình với phương sai bằng nhau hay kiểm định trung bình với phương sai khác nhau.

Đừng lo lắng, SPSS tự động thực hiện Levene test cho bạn trước khi thực hiện kiểm định trung bình.

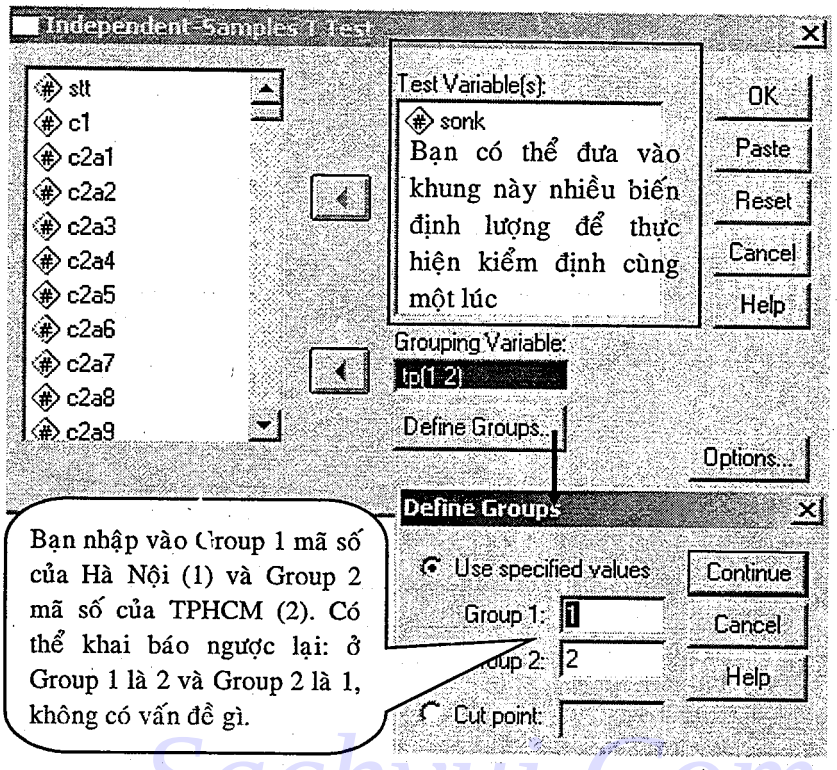
Tiến hành kiểm định Independent-samples T-test

Để tiến hành kiểm định trung bình cho ví dụ của chúng ta bạn xuất phát từ giả thuyết Ho rằng: quy mô gia đình trung bình tại 2 thành phố là như nhau.

1. Bạn hãy chọn Analyze>Compare Means>Independent-samples T Test. Lệnh này sẽ mở ra hộp thoại Independent-Samples T Test (kiểm định trung bình với mẫu độc lập).

2. Trong hộp thoại này, bạn lần lượt chọn biến định lượng muốn kiểm định trị trung bình (biến *sonnk*) đưa vào danh sách các biến cần kiểm định Test Variable(s). Sau đó chọn biến định tính chia số quan sát thành hai nhóm mẫu để so sánh giữa hai nhóm này với nhau (biến *tp*) đưa vào Grouping Variable.

Hình 5.3



3. Sau khi chọn biến để phân nhóm, bạn phải nhấn nút Define Groups... để chỉ định hai nhóm cần so sánh với nhau. Trong hộp thoại Define Groups gõ mã số của 2 nhóm bạn muốn so sánh. Trong trường hợp biến định tính có nhiều hơn 2 biểu hiện thì có thể chọn riêng 2 mã số đại diện cho 2 nhóm bạn quan tâm đến sự khác biệt về trung bình để nhập vào. Ví dụ bạn phân loại các sinh viên trong trường theo 4 nhóm về khoá học, thứ tự mã hoá trong biến khoá học là: 1 cho sinh viên năm đầu, 2 cho sinh viên năm 2, 3 cho sinh viên năm 3 và 4 cho sinh viên năm cuối. Nếu muốn so sánh thời gian tự học của sinh viên năm đầu và năm cuối, bạn sẽ nhập 1 vào Group 1 và 4 vào Group 2

4. Nhấn nút Continue để trở về hộp thoại chính rồi nhấn tiếp nút OK, bảng kết quả sẽ xuất hiện.

Bảng 5.3 **Group Statistics**

	thành phố	N	Mean	Std. Deviation	Std. Error Mean
số nhân khẩu trong hộ	Hà Nội	250	4.29	1.667	.105
	TPHCM	250	5.33	2.715	.172

Bảng 5.4 **Independent Samples Test**

			số nhân khẩu trong hộ	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		33.587	
	Sig.		.000	
t-test for Equality of Means	t		-5.142	-5.142
	df		498	413.356
	Sig. (2-tailed)		.000	.000
	Mean Difference		-1.04	-1.04
	Std. Error Difference		.201	.201
	95% Confidence Interval of the Difference	Lower	-1.432	-1.432
		Upper	-.640	-.640

Dựa vào kết quả kiểm định sự bằng nhau của 2 phương sai, ta sẽ xem kết quả kiểm định t.

- Nếu giá trị Sig. trong kiểm định Levene < 0,05 thì phương sai giữa 2 thành phố khác nhau, ta sẽ sử dụng kết quả kiểm định t ở phần Equal variances not assumed.
- Ngược lại nếu giá trị Sig. trong kiểm định Levene >= 0,05 thì phương sai giữa 2 thành phố không khác nhau, ta sẽ sử dụng kết quả kiểm định t ở phần Equal variances assumed.

Với kết quả kiểm định sự bằng nhau của 2 phương sai ở Bảng 5.4 ta bác bỏ giả thuyết Ho về sự bằng nhau của 2 phương sai do đó chúng ta sẽ sử dụng kết quả ở phần Equal variances not assumed cho kiểm định t.

- Nếu giá trị Sig. trong kiểm định t < 0,05 thì ta kết luận có sự khác biệt có ý nghĩa về trung bình giữa 2 thành phố.

- Nếu giá trị sig. trong kiểm định t >= 0,05 thì ta kết luận chưa có sự khác biệt có ý nghĩa về trị trung bình giữa 2 thành phố.

Trong ví dụ này, căn cứ vào giá trị Sig. ta có thể bác bỏ giả thuyết Ho và kết luận số nhân khẩu trung bình trong hộ gia đình ở TPHCM lớn hơn một cách có ý nghĩa thống kê so với Hà Nội (dựa vào giá trị trung bình mẫu ở Bảng 5.3).

2.2 Kiểm định trị trung bình của hai mẫu phụ thuộc hay mẫu phối hợp từng cặp. (Paired-samples T-test)

Đây là loại kiểm định dùng cho 2 nhóm tổng thể có liên hệ với nhau. Dữ liệu của mẫu thu thập ở dạng thang đo định lượng khoảng cách hoặc tỉ lệ. Quá trình kiểm định sẽ bắt đầu với việc tính toán chênh lệch giá trị trên từng cặp quan sát bằng phép trừ, sau đó kiểm nghiệm xem chênh lệch trung bình của tổng thể có khác 0 không, nếu không khác 0 tức là không có khác biệt. Lợi thế của phép kiểm định mẫu phối hợp từng cặp là nó loại trừ được những yếu tố tác động bên ngoài vào nhóm thử.

Phương pháp kiểm định này rất thích hợp với dạng thử nghiệm trước và sau, một thử nghiệm rất hay gặp trong nghiên cứu.

Ví dụ khi công ty chế biến thực phẩm của bạn khảo sát sự đánh giá của người tiêu dùng về loại đậu phộng chế biến sẵn vừa được cải tiến thành phần nước bột áo, bạn phải tổ chức cho dùng thử sản phẩm trên cùng một nhóm người mới có thể thu được những thông tin xác thực về sự đánh giá mùi vị, độ ngon. Nếu bạn tiến hành so sánh giữa 2 nhóm người dùng thử khác nhau thì sự đánh giá khác biệt có thể do những tác nhân khác gây ra như sự khác biệt về khẩu vị, nhận thức, kinh nghiệm...

Bạn sẽ tìm ra kết quả của thử nghiệm bằng cách yêu cầu người dùng thử cho điểm từng loại sản phẩm họ đã thử, càng ngon thì cho điểm càng cao. Sau đó áp dụng phương pháp kiểm định trị trung bình của hai mẫu phối hợp từng cặp trên 2 nhóm mẫu là điểm đánh giá của nhóm người tham gia thử nghiệm cho loại đậu phụng chưa cải tiến và đã cải tiến bạn sẽ có được kết luận là sản phẩm đã cải tiến hay chưa

cải tiến được ưa thích hơn. Cũng có khi cả 2 loại sản phẩm được đánh giá như nhau, tức là phương pháp cải tiến không thu được kết quả gì.

Điều kiện để áp dụng Paired-Samples t test là kích cỡ 2 mẫu so sánh phải bằng nhau (điều này là tất nhiên vì chúng ta phải lấy mẫu theo cặp); và chênh lệch giữa các giá trị của 2 mẫu phải có phân phối chuẩn hoặc cỡ mẫu phải đủ lớn để xem như xấp xỉ phân phối chuẩn. Nếu điều kiện về phân phối chuẩn không được thoả mãn hay cỡ mẫu không đủ lớn thì chúng ta phải nhờ đến một số loại kiểm định phi tham số khác (Bạn sẽ tìm hiểm những loại kiểm định này ở chương kế tiếp).

Với giá trị trung bình ($\overline{d}$) và độ lệch chuẩn (s_d) của các chênh kệch mẫu, kiểm định giả thuyết về sự khác nhau của 2 trị trung bình của tổng thể thực hiện như sau:

Ho: không có sự khác nhau về 2 trị trung bình tổng thể (tức là khác biệt giữa hai trung bình là bằng 0)

Trị kiểm định tính theo công thức như sau, trong đó n là số cặp quan sát trong mẫu

$$t = \frac{\overline{d}}{s_d / \sqrt{n}}$$

Quy tắc quyết định là bác bỏ giả thuyết Ho ở mức ý nghĩa α nếu

$-t_{n-1;\alpha/2} < t < t_{n-1;\alpha/2}$ ở đây t_{n-1} có phân phối Student với n-1 bậc tự do

Thực hiện kiểm định với SPSS

Điểm đánh giá của người dùng thử về 2 loại sản phẩm đậu phộng trước và sau cải tiến được thu thập trên thang điểm 10 và trình bày ở Bảng 5.5.

Bảng 5.5

STT	Trước cải tiến	Sau cải tiến	STT	Trước cải tiến	Sau cải tiến
1	7	8	11	7	9
2	8	9	12	7	5
3	6	5	13	8	9
4	8	9	14	9	10
5	7	8	15	7	7
6	7	9	16	7	9
7	7	7	17	8	7
8	6	7	18	7	9
9	8	7	19	6	6
10	6	8	20	8	8

Như vậy mẫu hiện có 40 quan sát, với 20 cặp sánh đôi, dữ liệu trên được nhập vào SPSS cũng theo kiểu từng cặp tương ứng nhau trong file ví dụ *Dauphong* trong tập hợp dữ liệu dùng kèm với sách, trong file này 2 mẫu được lưu trữ trong 2 biến có tên *spban* và *spthu*. Cột thứ nhất chứa điểm đánh giá về sản phẩm trước cải tiến trong biến *spban*, cột thứ hai chứa điểm đánh giá về sản phẩm sau cải tiến (sản phẩm thử nghiệm) trong biến *spthu*. Bạn tiến hành theo các bước sau:

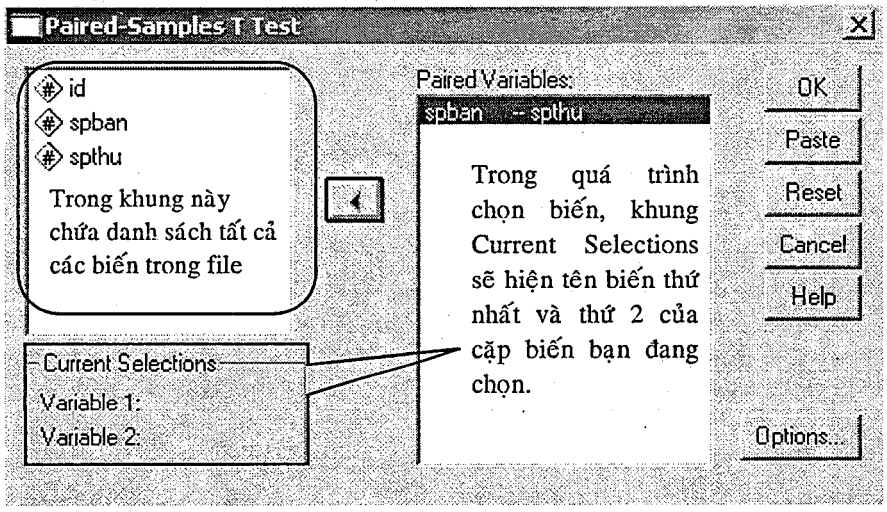
1. Vào menu Analyze > Compare Means mở ra hộp thoại Paired-Samples T Test

2. Trong hộp thoại này, chọn hai biến chứa các giá trị của hai mẫu quan sát trong danh sách biến nguồn ở phía bên trái hộp thoại đưa vào khung Paired Variables để so sánh. Nếu file của bạn có nhiều biến, bạn sẽ chọn cặp biến muốn so sánh bằng cách nhấp biến 1 rồi giữ phím Ctrl nhấp tiếp biến 2.

3. Nếu cần có thể nhấp nút Options để chỉnh lại độ tin cậy cho ước lượng khoảng chênh lệch giữa 2 trung bình. Trong hộp thoại này, hãy gõ độ tin cậy bạn cần vào ô Confidence Interval. Bấm nút continue để trở lại hộp thoại trước

4. Sau đó bấm OK, bảng kết quả hiện ra (từ Bảng 5.6 đến Bảng 5.8)

Hình 5.5



Bảng 5.6 Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	san pham dang ban	7.20	20	.834	.186
	san pham cai tien	7.80	20	1.399	.313

Bảng 5.7 Paired Samples Correlation

		N	Correlation	Sig.
Pair 1	san pham dang ban & san pham cai tien	20	.533	.016

Số cặp quan sát

Bảng 5.8 Paired Samples Test

			Pair 1
			san pham dang ban - san pham cai tien
Paired Differences	Mean		-.60
	Std. Deviation		1.188
	Std. Error Mean		.266
	95% Confidence Interval of the Difference	Lower	-1.16
		Upper	-.04
t			-2.259
df			19
Sig. (2-tailed)			.036

Bảng 5.7 thể hiện mối tương quan tuyến tính giữa 2 biến đại diện cho 2 mẫu. Trong Bảng 5.8, với mức ý nghĩa quan sát 2 phía Sig. (2 tailed) = 0,036 < 0,05 có thể kết luận rằng có sự chênh lệnh có ý nghĩa thống kê về sự đánh giá của người tiêu dùng đối với sản phẩm đậu phụng trước và sau khi cải tiến. Cụ thể sản phẩm sau khi cải tiến được đánh giá cao hơn sản phẩm trước cải tiến, trung bình là khoảng 0,6 điểm.

Để những nghiên cứu như thế này bảo đảm tính khách quan và độ chính xác cao người ta thường sử dụng dạng "blind test" để người tham gia không biết lần thử nào là sản phẩm cải tiến và lần thử nào là sản phẩm đang tiêu thụ.

Nên chú ý rằng Sig. (2 tailed) = 0,036 là tổng diện tích dưới đường cong hình chuông của phân phối t ứng với giá trị t = 2,259 trong trường hợp kiểm định 2 đuôi, nếu đây là trường hợp kiểm định 1 đuôi diện tích ứng với giá trị t này sẽ = 0,018 (chính là = 0,036/2). Như vậy trong các kiểm định 2 đuôi mà chúng ta đã thực hiện trên SPPS tới nay, khi áp dụng quy tắc so sánh Sig. (2-tailed) với giá trị α các bạn sẽ so mức ý nghĩa quan sát mà SPSS đưa ra với chính α chứ không so với $\alpha/2$.

CHƯƠNG VI

PHÂN TÍCH LIÊN HỆ GIỮA BIẾN NGUYÊN NHÂN ĐỊNH TÍNH VÀ BIẾN KẾT QUẢ ĐỊNH LƯỢNG: PHÂN TÍCH PHƯƠNG SAI

1. PHÂN TÍCH PHƯƠNG SAI MỘT YẾU TỐ (ANOVA)

1.1 Khái niệm và vận dụng

Khi sử dụng kiểm định t đối với hai mẫu độc lập, trong trường hợp biến phân loại của bạn có 3 nhóm, chúng ta vẫn có thể thực hiện được kiểm định bằng lệnh Independent-Samples T-test với 3 cặp so sánh (1-2; 1-3; 2-3). Lần lượt lặp lại lệnh Independent-Samples T-test 3 lần, mỗi lần như vậy bạn lần lượt khai báo từng cặp so sánh bằng mã số tương ứng của nhóm trong Grouping Variable. Quy tắc là nếu biến phân loại của bạn có k nhóm thì số cặp trung bình cần so sánh là = k!/[2!(k-2)!]

Chú ý rằng mỗi lần tiến hành kiểm định giả thuyết trung bình bằng nhau cho từng cặp như vậy ta chấp nhận một khả năng phạm sai lầm là 5% (hoặc nhiều hơn hay ít hơn tuỳ độ tin cậy ta mong muốn), như vậy khi làm kiểm định nhiều lần khả năng sai lầm sẽ tăng lên theo số lần làm kiểm định.

Thực tế, với những trường hợp như vậy, chúng ta sử dụng phân tích phương sai (Analysis of variance–ANOVA) vì nó tiến hành kiểm định tất cả các nhóm mẫu cùng một lúc với khả năng phạm sai lầm chỉ là 5%.

Có thể nói phân tích phương sai là sự mở rộng của kiểm định t, vì phương pháp này giúp ta so sánh trị trung bình của 3 nhóm trở lên. Kỹ thuật phân tích phương sai được dùng để kiểm định giả thiết các tổng thể nhóm (tổng thể bộ phận) có trị trung bình bằng nhau. Kỹ thuật này dựa trên cơ sở tính toán mức độ biến thiên trong nội bộ các nhóm và biến thiên giữa các trung bình nhóm. Dựa trên hai ước lượng này của mức độ biến thiên ta có thể rút ra kết luận về mức độ khác nhau giữa các trung bình nhóm.

SPSS có hai thủ tục phân tích phương sai: ANOVA một yếu tố và ANOVA nhiều yếu tố. Phân tích phương sai một yếu tố sử dụng khi chúng ta chỉ sử dụng 1 biến yếu tố để phân loại các quan sát thành các nhóm khác nhau. Trong trường hợp căn cứ vào 2 hay nhiều biến yếu tố để phân chia các nhóm thì ta phải dùng đến thủ tục ANOVA nhiều yếu tố. Trong phạm vi quyển sách này, chúng ta chỉ xem xét đến phân tích phương sai hai yếu tố (Two - Way ANOVA)

Trong ví dụ về nghiên cứu bạn đọc ở file *Data Thuc hanh*, giả sử chúng ta cần so sánh có khác biệt hay không về mức độ đánh giá tầm quan trọng của yếu tố "có tự do cá nhân" đối với cuộc sống của một con người giữa những nhóm người có học vấn khác nhau. Với thang đo định lượng dạng khoảng cách 7 mức độ (với 1: không quan trọng đến 7: rất quan trọng), chúng ta có biến *c36.6* là biến định lượng cần nghiên cứu. Các học vấn khác nhau được đo lường bằng biến *nhomhv*. Ở đây ta có 4 mức học vấn là: 1: cấp 1-2; 2: cấp 3- THCN; 3: CĐ-SVĐH; 4: Tốt nghiệp ĐH.

Vấn đề nghiên cứu ở đây là mức độ quan trọng của yếu tố "có tự do cá nhân" có khác biệt nhau không giữa 4 nhóm người có trình độ học vấn khác nhau. Ta đặt giả thuyết:

H_0: Không có khác biệt về sự đánh giá tầm quan trọng của yếu tố "có tự do cá nhân" giữa các nhóm trình độ học vấn.

(Hay là "Trình độ học vấn không có liên hệ với sự đánh giá về tầm quan trọng của yếu tố tự do cá nhân đối với cuộc sống")

1.2 Tóm tắt lý thuyết phân tích phương sai một yếu tố (One-Way ANOVA)

Có một số giả định sau đối với phân tích phương sai một yếu tố:

- Các nhóm so sánh phải độc lập và được chọn một cách ngẫu nhiên.
- Các nhóm so sánh phải có phân phối chuẩn hoặc cỡ mẫu phải đủ lớn để được xem như tiệm cận phân phối chuẩn.
- Phương sai của các nhóm so sánh phải đồng nhất

Nếu giả định tổng thể có phân phối chuẩn với phương sai bằng nhau không đáp ứng được thì kiểm định phi tham số Kruskal-Wallis sẽ là

một giải pháp thay thế hữu hiệu cho ANOVA. Bạn có thể tìm hiểu về kiểm định này ở Chương sau.

Một cách tổng quát, giả sử từ một biến phân loại ta chia tổng thể mẫu thành k nhóm độc lập gồm $n_1, n_2, ..., n_k$ quan sát tương ứng trong từng nhóm. n là số quan sát của tổng thể mẫu.

Ta ký hiệu:

- x_{ij} : giá trị của biến định lượng đang nghiên cứu tại quan sát thứ j thuộc nhóm i
- $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ là các trung bình nhóm, và $\mu_1, \mu_2, ... \mu_k$ là các trung bình thực của các tổng thể nhóm mà từ đó ta rút ra được các mẫu tương ứng.
- $\bar{x}$ là trung bình chung của các tất cả các nhóm theo biến định lượng đang nghiên cứu tức trung bình tính chung cho mẫu không phân tách thành nhóm.

Giả thiết Ho cần kiểm định là trung bình thực (trung bình tổng thể) của k nhóm này bằng nhau:

Ho: $\mu_1 = \mu_2 = ... = \mu_k$

(Nghĩa là không có sự khác biệt giữa các trung bình của các nhóm được phân loại theo biến định tính).

Ta có thể tính toán đại lượng kiểm định theo trình tự sau:

Tổng các chênh lệch bình phương (sum of squares) được xác định như sau:

1. Tổng các chênh lệch bình phương trong nội bộ nhóm (Within-groups sum of squares): phản ảnh biến thiên ngẫu nhiên do ảnh hưởng của các yếu tố khác không xem xét ở đây.

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

2. Tổng các chênh lệch bình phương giữa các nhóm (Between-groups sum of squares): phản ảnh biến thiên của biến định lượng đang nghiên cứu do tác động của biến phân loại xem xét.

$$SSG = \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2$$

3. Tổng các chênh lệch bình phương toàn bộ (Total sum of squares): phản ảnh toàn bộ biến thiên của biến định lượng đang nghiên cứu.

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

Bằng các biến đổi toán học chúng ta có: SST= SSW + SSG

Các chênh lệch bình phương bình quân hay còn gọi là chênh lệch quân phương (mean squares) được xác định như sau:

1. Phương sai trong nội bộ các nhóm (Within-groups mean squares)

$$MSW = \frac{SSW}{n-k}$$

2. Phương sai giữa các nhóm (Between-groups mean squares)

$$MSG = \frac{SSG}{k-1}$$

Nguyên tắc quyết định với mức ý nghĩa α là:

Bác bỏ Ho nếu: $\frac{MSG}{MSW} \rangle F_{k-1,n-k,\alpha}$

trong đó $F_{k-1,n-k,\alpha}$ là giá trị sao cho $P(F_{k-1,n-k} \rangle F_{k-1,n-k,\alpha}) = \alpha$

$F_{k-1,n-k}$ có phân phối F với bậc tự do của tử số là (k-1) và bậc tự do của mẫu số là (n-k).

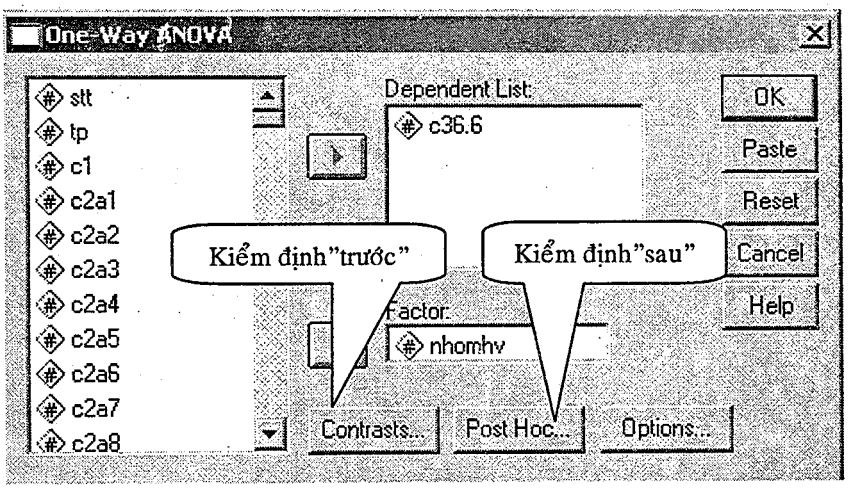
Nếu kết quả kiểm định dẫn đến việc bác bỏ Ho thì ta phải làm tiếp phân tích sâu (thủ tục Post Hoc) để xác định trung bình của nhóm nào khác với nhóm nào, tức là tìm xem sự khác biệt xảy ra ở đâu, và xác định hướng cũng như độ lớn của khác biệt.

1.3 Thực hiện phân tích phương sai một yếu tố với SPSS

Chúng ta có thể thực hiện phân tích ANOVA 1 yếu tố trên SPSS:

1. Từ menu chọn: Analyze > Compare Means > One-Way ANOVA. Lệnh này sẽ mở ra hộp thoại sau:

Hình 6.1



2. Trong hộp thoại này:

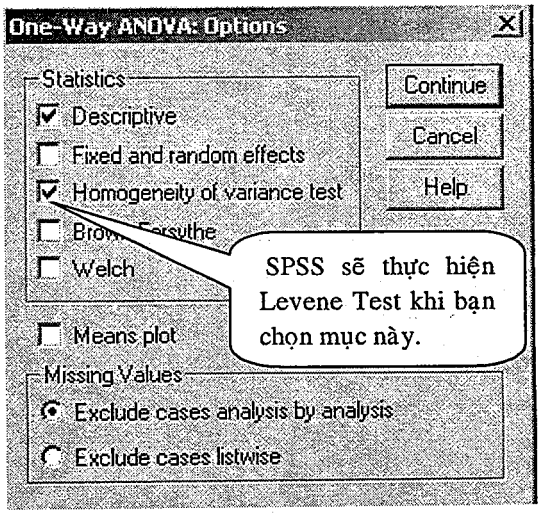
- đưa biến định lượng vào khung Dependent List
- biến phân loại xác định các đối tượng (nhóm) cần so sánh vào khung Factor

3. Chọn nút Options...Lệnh này mở ra hộp thoại Options (Hình 6.2)

- Descriptive để tính các đại lượng thống kê mô tả chi tiết cho từng nhóm được phân tách để so sánh.
- Homogeneity-of-variance để kiểm định sự bằng nhau của các phương sai nhóm.

Sau đó nhấp nút Continue trở về hộp thoại ban đầu và nhấp nút OK.

Hình 6.2



Về quy trình của thủ tục ANOVA thì đến đây bạn có thể chọn OK trong hộp thoại One-Way ANOVA để xem kết quả kiểm định ANOVA. Nếu Ho được chấp nhận thì công việc đã xong. Nếu Ho bị bác bỏ bạn phải trở lại hộp thoại One Way ANOVA thực hiện tiếp thủ tục kiểm định nhằm xác định cụ thể trung bình của nhóm nào khác với nhóm nào, tức là tìm xem sự khác biệt giữa các nhóm xảy ra ở đâu. Bạn có thể thực hiện liên tục thủ tục ANOVA và thủ tục tìm kiếm sự khác biệt một lần trong quá trình tiến hành lệnh. Nếu bạn bác bỏ giả thuyết Ho của phân tích ANOVA, bạn có thể xem tiếp kết quả tìm kiếm sự khác biệt do SPSS đưa ra, còn nếu bạn không thể bác bỏ giả thuyết Ho thì cũng không có ảnh hưởng gì. Ở đây ta làm đúng trình tự là đọc kết quả ANOVA rồi mới tiến hành tiếp việc tìm kiếm.

1.4 Đọc kết quả phân tích phương sai của SPSS

Thủ tục ANOVA của SPSS cho ra các bảng kết quả từ Bảng 6.1 đến 6.3 như ở trang sau:

Bảng kết quả đầu tiên cho thấy các đại lượng thống kê mô tả cho từng nhóm và cho toàn bộ mẫu nghiên cứu. Bảng kết quả thứ nhì cho biết kết quả kiểm định phương sai. Với mức ý nghĩa 0,257 có thể nói phương sai của sự đánh giá tầm quan trọng của yếu tố "có tự do cá nhân" giữa 4 nhóm học vấn không khác nhau một cách có ý nghĩa thống kê. Như vậy kết quả phân tích ANOVA có thể sử dụng tốt.

Bảng 6.1 **Descriptives**

Có tự do cá nhân

		cấp 1-2	cấp 3-THCN	CĐ-SVĐH	tốt nghiệp ĐH	Total
N		60	215	91	134	500
Mean		5.43	5.18	5.36	4.97	5.19
Std. Deviation		1.294	1.391	1.502	1.392	1.406
Std. Error		.167	.095	.157	.120	.063
95% Confidence Interval for Mean	Lower Bound	5.10	4.99	5.05	4.73	5.06
	Upper Bound	5.77	5.37	5.68	5.21	5.31
Minimum		2	1	2	1	1
Maximum		7	7	7	7	7

Bảng 6.2 **Test of Homogeneity of Variances**

Có tự do cá nhân

Levene Statistic	df1	df2	Sig.
1.351	3	496	.257

Bảng 6.3 **ANOVA**

Có tự do cá nhân

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	12.756	3	4.252	2.166	.091
Within Groups	973.572	496	1.963		
Total	986.328	499			

Bảng kết quả thứ ba trình bày kết quả phân tích ANOVA. Với mức ý nghĩa quan sát Sig.= 0,091 nếu bạn chấp nhận độ tin cậy của phép kiểm định này là 90% (mức ý nghĩa = 0,1) thì có thể nói có sự khác biệt có ý nghĩa thống kê về sự đánh giá tầm quan trọng của yếu tố "có tự do cá nhân" giữa 4 nhóm người có trình độ học vấn khác nhau. Nhìn vào Bảng thống kê mô tả 6.1, chúng ta có thể thấy mức độ quan trọng có vẻ được đánh giá giảm dần khi học vấn càng cao.

1.5. Xác định chỗ khác biệt (phân tích sâu ANOVA)

Việc kế tiếp chúng ta phải làm là tìm xem sự đánh giá này là khác biệt giữa những nhóm học vấn nào.

Có 2 phương pháp để xác định sự khác biệt trong các trị trung bình nhóm xảy ra ở đâu, đó là kiểm định "trước" và kiểm định "sau"

- Kiểm định "trước" có thể định nghĩa như là kiểm định các giả định về sự khác nhau của các trung bình nhóm trước khi thực hiện phân tích ANOVA. Lúc này ta thực hiện việc xác định những chênh lệch cụ thể giữa các trung bình nhóm theo phán đoán chủ định của ta. Kiểu kiểm định này được thực hiện trong hộp thoại Contrasts với tên gọi là kiểm định Priori Contrasts (Hình 6.1)
- Kiểm định sau có thể định nghĩa như là kiểm định các giả định về sự khác nhau của các trung bình nhóm sau khi đã thực hiện phân tích ANOVA. Trong phạm vi cuốn sách này ta chọn phương pháp kiểm định thứ 2 như một cách tiếp cận gần với phương pháp nghiên cứu thực. Kiểm định đó được thực hiện trong hộp thoại Post Hoc.

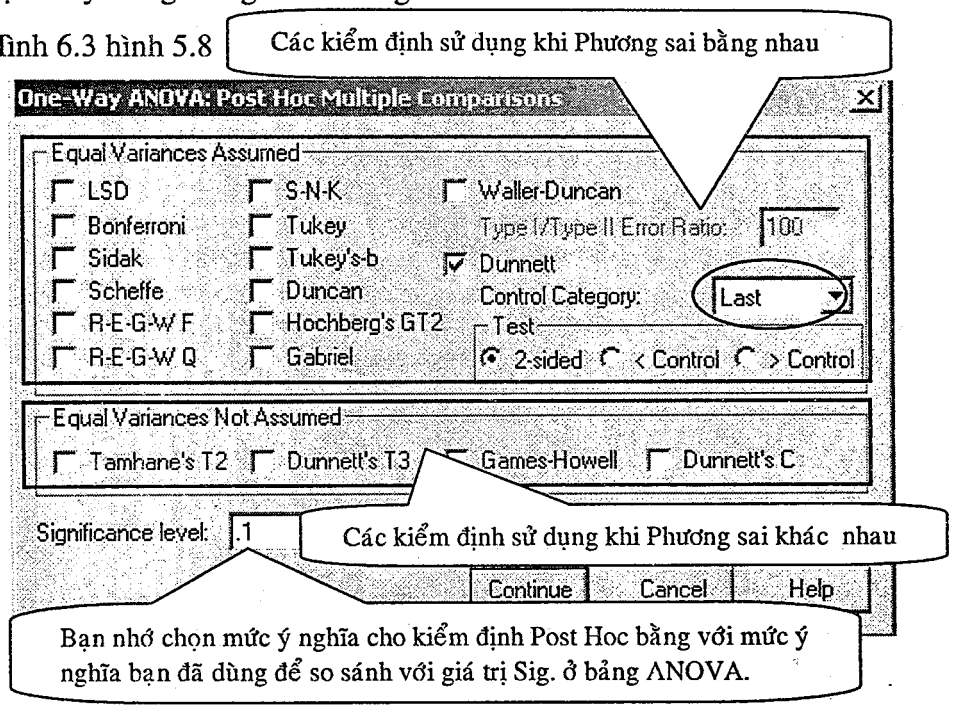
Bằng cách nhấp nút Post Hoc bạn mở hộp thoại One-Way ANOVA: Post Hoc Multiple Comparisons (Hình 6.3), trong hộp thoại này ta có các phương pháp kiểm định thống kê sau để so sánh các trị trung bình của các nhóm:

- LSD: phép kiểm định này chính là việc dùng kiểm định t lần lượt cho từng cặp trung bình nhóm mà ta đã nhắc đến ở trên, do vậy nhược điểm của nó là độ tin cậy không cao vì nó làm gia tăng mức độ phạm sai lầm (mức ý nghĩa) tương ứng với việc so sánh nhiều nhóm cùng một lúc.
- Bonferroni: tiến hành giống quy tắc của LSD nhưng điều chỉnh được mức ý nghĩa khi tiến hành so sánh bội dựa trên số lần tiến hành so sánh. Nó là một trong những thủ tục kiểm định đơn giản nhất và hay được sử dụng cho mục tiêu này.
- Tukey: cũng được sử dụng phổ biến cho việc tìm kiếm các trung bình nhóm khác biệt. Nó sử dụng bảng phân phối Studentizze range distribution. Tukey hiệu quả hơn Bonferroni khi số lượng các cặp trung bình cần so sánh khá nhiều.
- Scheffe: phương pháp này kém nhạy hơn trong việc so sánh các trị trung bình của các cặp, nó đòi hỏi phải có sự khác biệt lớn giữa các trị trung bình so với các thủ tục so sánh bội khác để bảo đảm có sự khác biệt thật sự, nhưng vì thế nó lại đưa ra một kết quả kiểm định thận trọng hơn.
- R-E-G-W: thực hiện 2 bước kiểm định, đầu tiên nó tiến hành kiểm định lại toàn bộ các trị trung bình nhóm xem có bằng nhau không; nếu không bằng thì bước kế tiếp nó sẽ kiểm định để tìm các nhóm nào khác biệt thật sự với nhau về trị trung bình. Nhưng kiểm định này không phù hợp khi kích cỡ các nhóm mẫu không bằng nhau.
- Dunnett: là thủ tục cho phép chọn so sánh các trị trung bình của các nhóm mẫu còn lại với một trị trung bình của một nhóm mẫu cụ thể nào đó được chọn ra so sánh (nhóm điều khiển), SPSS mặc định chọn nhóm cuối (Last) để làm nhóm điều khiển.
- Trong trường hợp phương sai giữa các đối tượng cần so sánh khác nhau, người ta hay chọn Tamhane's T2 (kiểm định t từng cặp trường hợp phương sai khác nhau).

Ở ví dụ này ta thử chọn dạng kiểm định Dunnett với lựa chọn mặc định là nhóm cuối cùng. Sau đó nhấp Continue trở lại hộp thoại ban

đầu, rồi chọn OK, kết quả kiểm định sự khác biệt sẽ xuất hiện như bạn thấy trong Bảng 5.12. Bảng 6.4

Hình 6.3 hình 5.8



Bảng 6.4 **Multiple Comparisons**

Dependent Variable: Có tự do cá nhân

Dunnett t (2-sided)

(J) học vấn			(I) học vấn		
			cấp 1-2	cấp 3-THCN	CĐ-SVĐH
tốt nghiệp ĐH	Mean Difference (I-J)		.46(*)	.21	.39
	Std. Error		.218	.154	.190
	Sig.		.090	.391	.104
	90% Confidence Interval	Lower Bound	.01	-.11	.00
		Upper Bound	.92	.53	.79

* The mean difference is significant at the .1 level.

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

Bảng kết quả cuối cùng cho thấy kết quả kiểm định t cho từng cặp 2 nhóm (tốt nghiệp ĐH với Cấp 1-2; tốt nghiệp ĐH với cấp 3-THCN; tốt nghiệp ĐH với CĐ-SVĐH). Chúng ta có thể thấy chỉ có sự khác biệt có ý nghĩa giữa nhóm có trình độ học vấn cấp 1-2 và nhóm đã

tốt nghiệp ĐH vì mức ý nghĩa quan sát ở kiểm định chênh lệch trung bình cặp ở này < 0,1 là mức ý nghĩa ta đã chọn cho kiểm định này.

Chú ý rằng dấu (*) chỉ sự khác biệt phát hiện được ở cặp 2 nhóm tương ứng với dòng và cột của ô chứa dấu (*). Ở các cặp khác, giá trị Sig. đều lớn hơn 0,1 nên không có dấu (*). Như vậy nếu một bảng kết quả kiểm định sự khác biệt không có dấu (*) nào nghĩa là không có cặp trị trung bình khác biệt nào được tìm thấy.

Nếu bạn tiến hành lại kiểm định này với thủ tục kiểm định Bonfferoni hoặc Tuykey bạn sẽ thu được 1 bảng kết quả không có dấu (*) nào, đó là vì mỗi loại kiểm định có những giới hạn chặt chẽ khác nhau. Tốt nhất là bạn nên tiến hành cùng lúc một vài kiểm định và quyết định bác bỏ giả thuyết Ho hay không tuỳ vào mức độ chấp nhận mạo hiểm của bạn, chất lượng mẫu bạn có, và bạn thoả mãn được đến đâu các giả định thống kê ...

Với kết quả phân tích ANOVA trên, nếu bạn đề ra một độ tin cậy cao hơn (0,05 chẳng hạn) bạn sẽ có kết luận khác. Một lần nữa, bạn nhớ rằng trong thống kê, chọn mức ý nghĩa bao nhiêu còn phụ thuộc khá nhiều vào mục đích của kiểm định của bạn là gì, với một cuộc nghiên cứu khám phá, bạn nên hài lòng với mức ý nghĩa 0,1. Nhưng trong đánh giá các mặt ảnh hưởng của một phương pháp điều trị y học, bạn cần đòi hỏi một mức ý nghĩa tới 0,001 tức độ tin cậy 99%.

2. PHÂN TÍCH PHƯƠNG SAI HAI YẾU TỐ (Two -way anova)

Giả thuyết cho kiểm định Anova hai (chiều) yếu tố là như sau

Trước khi tiến hành kiểm định ANOVA hai chiều phải đảm bảo các giả định sau được thỏa mãn (các giả định này cũng giống ANOVA một yếu tố)

- tổng thể có phân phối chuẩn: tổng thể mà từ đó mẫu nghiên cứu của chúng ta được chọn phải có phân phối chuẩn
- "phương sai đồng đều" phương sai của các nhóm nghiên cứu phải đều nhau

Trong hai giả định thì giả định thứ 2 cần được chú ý trước tiên vì nếu bạn không đạt được giả định này tức là kiểm định của bạn được đánh giá tại một mức ý nghĩa lớn hơn mức ý nghĩa bạn dự định lúc ban đầu. Ví dụ thay vì mức ý nghĩa 0,05 như dự định thì bây giờ kiểm định của bạn chỉ thực sự có ý nghĩa tại mức $\alpha = 0,1$. Bởi vì sự vi phạm giả định phương sai

bằng nhau đã bóp méo hình dạng của phân phối F khiến cho giá trị tới hạn F không còn tương ứng với điểm 5% nữa.

Để minh họa cho lệnh ANOVA hai chiều chúng ta dùng lại ví dụ đã trình bày trong Sách Thống kê ứng dụng (2007) là ảnh hưởng của mức độ yêu thích ngành học & thời gian tự học đến kết quả học tập của sinh viên.

Các giả thuyết H_0 đặt ra cho bài toán này như sau:

1. Điểm trung bình học tập (ĐTB) của sinh viên có thời gian tự học khác nhau đều bằng nhau.
2. ĐTB của sinh viên có mức độ yêu thích ngành đang học khác nhau đều bằng nhau.
3. Không có ảnh hưởng tương tác giữa thời gian tự học và mức độ yêu thích ngành đang học của sinh viên. Nói một cách cụ thể, ảnh hưởng của thời gian tự học đến ĐTB là như nhau đối với các nhóm sinh viên có mức độ yêu thích ngành đang học khác nhau; và ảnh hưởng của mức độ yêu thích ngành đang học đến ĐTB là như nhau đối với các nhóm sinh viên có thời gian tự học khác nhau.

Cách tiến hành

- Đầu tiên ta khởi tạo dữ liệu trên SPSS như cấu trúc sau

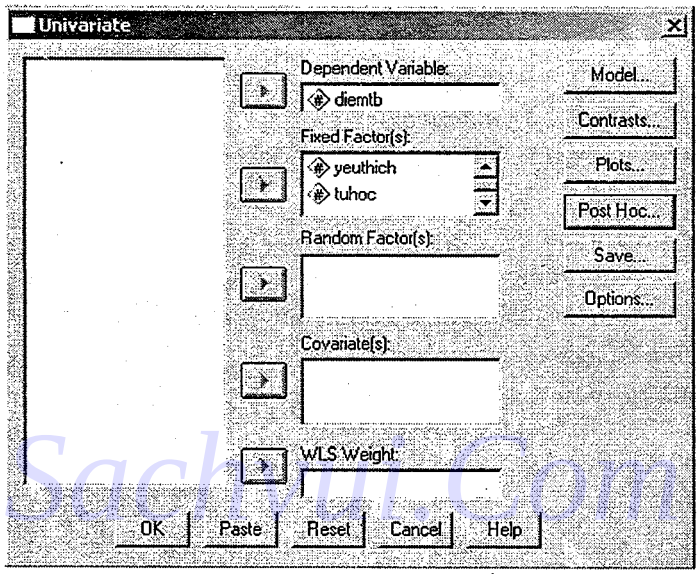
Hình 6.4

	diemtb	yeuthich	tuhoc
1	5,8	1	1
2	6,2	1	1
3	5,4	1	1
4	6,0	1	1
5	5,2	1	1
6	5,3	1	1
7	5,4	1	1
8	5,6	2	1
9	6,2	2	1
10	5,7	2	1
11	5,5	2	1
12	6,1	2	1
13	6,0	2	1
14	5,2	2	1
15	6,4	3	1

- Sau đó vào menu Analyze chọn General Liner Model chọn Univariate để mở hộp thoại thực hiện ANOVA hai chiều.

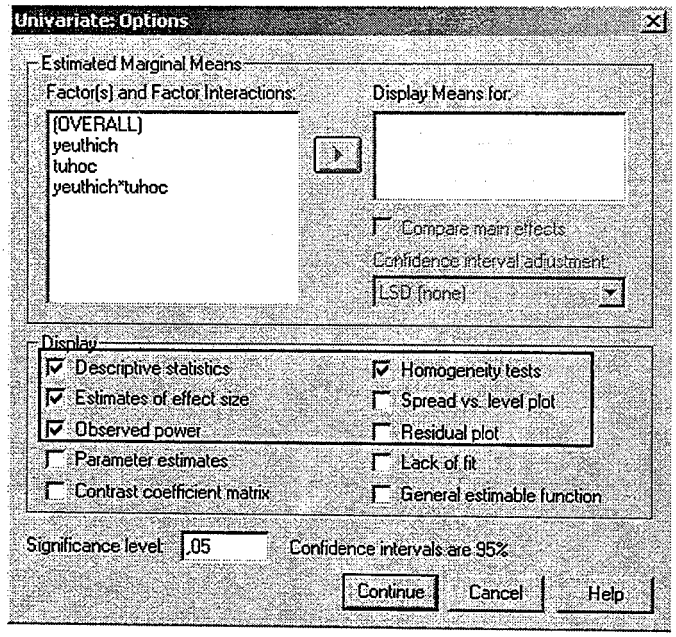
- Đưa biến ***diemtb*** sang khung Dependent variable và hai biến ***yeuthich*** và ***tuhoc*** sang khung Fixed Factor (chọn sáng hai biến đưa qua đồng thời).

Hình 6.5



- Rồi vào Options chọn hiện hành các mục sau

Hình 6.6



Sau đó nhấp Continue rồi về màn hình chính nhấp OK được các kết quả.

Bảng 6.5 **Between-Subjects Factors**

		Value Label	N
YEUTHICH	1	khong thich lam	21
	2	thich	21
	3	Rat thich	21
TUHOC	1	Tu hoc it	21
	2	Tu hoc TB	21
	3	Tu hoc nhieu	21

Bảng 6.6 **Descriptive Statistics**

Dependent Variable: DIEMTB

YEUTHICH	TUHOC	Mean	Std. Deviation	N
khong thich lam	Tu hoc it	5,614	,3848	7
	Tu hoc TB	6,043	,2637	7
	Tu hoc nhieu	6,129	,2928	7
	Total	5,929	,3797	21
thich	Tu hoc it	5,757	,3599	7
	Tu hoc TB	6,400	,2944	7
	Tu hoc nhieu	6,914	,2545	7
	Total	6,357	,5653	21
Rat thich	Tu hoc it	5,729	,5155	7
	Tu hoc TB	6,757	,3735	7
	Tu hoc nhieu	7,357	,3599	7
	Total	6,614	,7970	21
Total	Tu hoc it	5,700	,4087	21
	Tu hoc TB	6,400	,4219	21
	Tu hoc nhieu	6,800	,5958	21
	Total	6,300	,6602	63

Bảng 6.7 **Levene's Test of Equality of Error Variances(a)**

Dependent Variable: DIEMTB

F	df1	df2	Sig.
1,491	8	54	,182

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
a Design: Intercept+YEUTHICH+TUHOC+YEUTHICH * TUHOC

Bảng 6.8

Tests of Between-Subjects Effects

Dependent Variable: DIEMTB

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Corrected Model	20,306[b]	8	2,538	20,414	,000	,752	163,310	1,000
Intercept	2500,470	1	2500,470	20110,163	,000	,997	20110,163	1,000
YEUTHICH	5,040	2	2,520	20,267	,000	,429	40,534	1,000
TUHOC	13,020	2	6,510	52,357	,000	,660	104,714	1,000
YEUTHICH * TUHOC	2,246	4	,561	4,515	,003	,251	18,061	,922
Error	6,714	54	,124					
Total	2527,490	63						
Corrected Total	27,020	62						

a. Computed using alpha = ,05

b. R Squared = ,752 (Adjusted R Squared = ,715)

Hai bảng kết quả đầu tiên cho ta thấy các số liệu thống kê mô tả về mẫu dữ liệu của chúng ta. Bảng thứ ba là kết quả kiểm định sự bằng nhau của phương sai các nhóm. Kết quả kiểm định Levene cho thấy giả định phương sai bằng nhau đã không bị vi phạm. Cuối cùng là bảng kết quả kiểm định chính, bảng này cho thấy hai nhân tố chính cũng như sự tương tác giữa chúng đều có ảnh hưởng đến kết quả học tập (tại mức ý nghĩa 5% các p-value đều bé hơn mức ý nghĩa nên ta bác bỏ cả ba giả thuyết H_0 đã đặt ra). Do đó chúng ta phải tiếp tục tiến hành phân tích sâu ANOVA, bằng kiểm định Tukey chúng ta có các bảng kết quả sau cho từng tình huống kiểm định

1.Xác định các cặp trung bình tổng thể khác nhau theo yếu tố nghiên cứu thứ nhất là thời gian tự học

Bảng 6.9 **Multiple Comparisons**

Dependent Variable: DIEMTB
Tukey HSD

(I) TUHOC	(J) TUHOC	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-,700(*)	,1088	,000	-,962	-,438
	3	-1,100(*)	,1088	,000	-1,362	-,838
2	1	,700(*)	,1088	,000	,438	,962
	3	-,400(*)	,1088	,002	-,662	-,138
3	1	1,100(*)	,1088	,000	,838	1,362
	2	,400(*)	,1088	,002	,138	,662

Based on observed means.

* The mean difference is significant at the ,05 level.

Bảng 6.10 **DIEMTB**

Tukey HSD

TUHOC	N	Subset 1	Subset 2	Subset 3
1	21	5,700		
2	21		6,400	
3	21			6,800
Sig.		1,000	1,000	1,000

Means for groups in homogeneous subsets are displayed. Based on Type III Sum of Squares The error term is Mean Square(Error) = ,124.
a Uses Harmonic Mean Sample Size = 21,000.
b Alpha = ,05.

Tất cả các giá trị sig tính toán được đều bé hơn 0,05 rất nhiều nên ta có thể bác bỏ các giả thuyết H_0 và kết luận rằng sinh viên mà có thời gian tự học khác nhau sẽ có kết quả học tập khác nhau.

2. Xác định các cặp trung bình tổng thể khác nhau theo yếu tố thứ hai là mức độ ưa thích ngành học

Bảng 6.11 **Multiple Comparisons**

Dependent Variable: DIEMTB
Tukey HSD

(I) YEUTHICH	(J) YEUTHICH	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-,429(*)	,1088	,001	-,691	-,166
	3	-,686(*)	,1088	,000	-,948	-,423
2	1	,429(*)	,1088	,001	,166	,691
	3	-,257	,1088	,056	-,519	,005
3	1	,686(*)	,1088	,000	,423	,948
	2	,257	,1088	,056	-,005	,519

Based on observed means.
* The mean difference is significant at the ,05 level.

Bảng 6.12 **DIEMTB**

Tukey HSD

YEUTHICH	N	Subset 1	Subset 2
1	21	5,929	
2	21		6,357
3	21		6,614
Sig.		1,000	,056

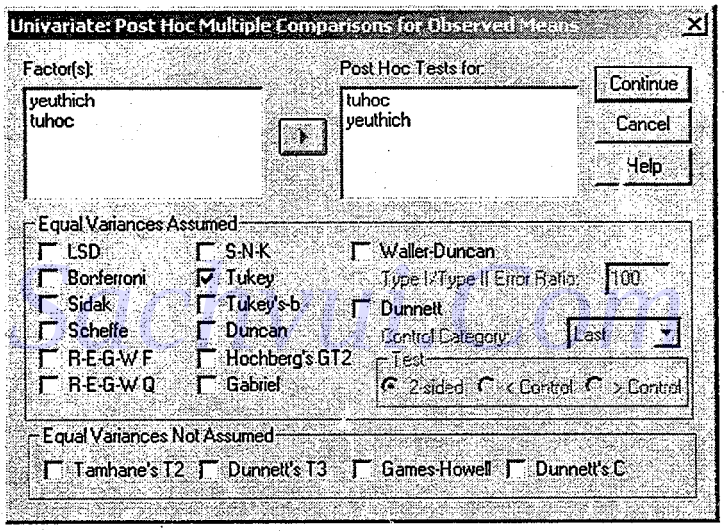
Means for groups in homogeneous subsets are displayed. Based on Type III Sum of Squares The error term is Mean Square(Error) = ,124.
a Uses Harmonic Mean Sample Size = 21,000.
b Alpha = ,05.

Giá trị sig cho ta kết luận là các sinh viên có mức độ yêu thích ngành học nhiều hoặc rất nhiều thì không có khác biệt nhau về kết quả học tập vì giá trị sig cho chênh lệch giữa nhóm 2 và 3 lớn hơn mức ý nghĩa 0,05.

Cách thức tiến hành kiểm định Tukey trên SPSS

Các bạn đi lại trình tự kiểm định ANOVA, trên cửa sổ Univariate, chọn mục Post Hoc để mở cửa sổ sau

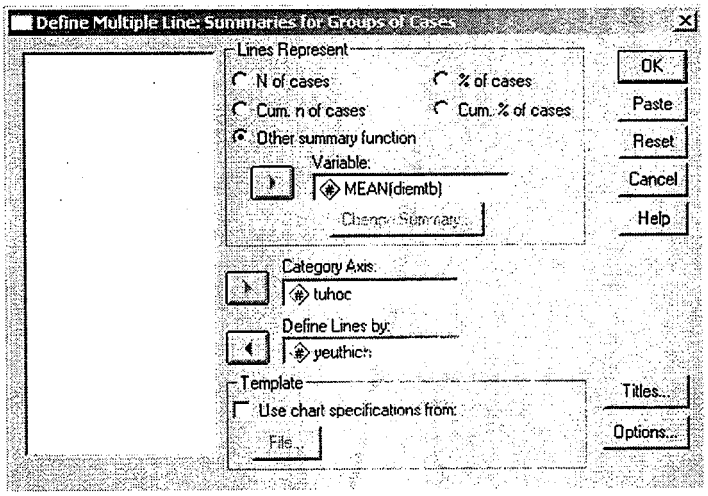
Hình 6.7



Trong kết quả chính của phân tích ANOVA ta thấy YEUTHICH*TUHOC có giá trị sig = 0,003 < 0,05 nghĩa là tác động của thời gian tự học đến kết quả học tập phụ thuộc vào mức độ yêu thích môn học của sinh viên. Điều này khuyến khích chúng ta tiến hành một phân tích tác động đơn, sự phân tích này có thể được trình bày rất dễ hiểu bằng cách vẽ đồ thị các giá trị điểm trung bình theo quá trình như sau.

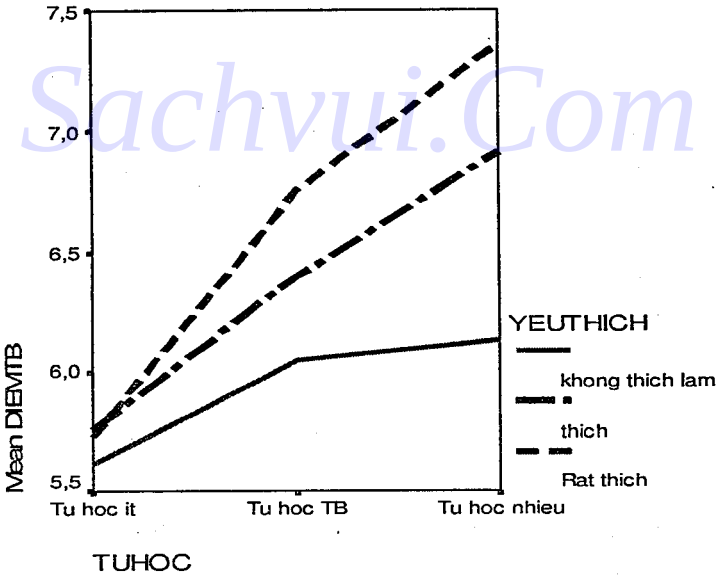
Vào menu Graph chọn Line rồi chọn mục Multiple (nhớ chọn Summaries for Groups of case) rồi nhấn nút Define mở hộp thoại kế tiếp. Trên hộp thoại này bạn thực hiện các thao tác như trong hình sau đây (chú ý nên bỏ vào khung Define Lines by factor nào ít biểu hiện hơn trong hai factor của nghiên cứu)

Hình 6.8



Sau khi nhấp nút OK có kết quả sau

Hình 6.9



Đồ thị chứng tỏ theo mức độ gia tăng của thời gian tự học điểm trung bình học tập cũng tăng nhưng điểm của các sinh viên rất yêu thích ngành học tăng cao nhất và thuyết phục nhất.

SPSS có thể thực hiện hai loại kiểm định ANOVA hai chiều, ngoài cách chúng ta vừa nghiên cứu ở trên (gọi là Two-Way Between) thì SPSS còn có một dạng chạy ANOVA hai chiều khác là vào lệnh General Liner Model nhưng chọn lệnh phụ là Repeat measures để vào hộp thoại Repeat measures Define Factor(s) thực hiện lệnh ANOVA hai chiều (gọi là Two-Way Repeat).

CHƯƠNG VII
KIỂM ĐỊNH PHI THAM SỐ

Hầu hết các kiểm định thống kê bạn sử dụng trong chương trước đều đòi hỏi những giả định khá chặt chẽ về phân phối chuẩn của tổng thể mà từ đó mẫu được chọn ra. Ví dụ như trong phần kiểm định phương sai chúng ta đã biết rằng mỗi nhóm quan sát phải là 1 mẫu ngẫu nhiên độc lập được rút từ một tổng thể có phân phối chuẩn và các phương sai của các tổng thể cần so sánh này cần bằng nhau. Chúng ta cũng đều biết rằng khi các giả định không hoàn toàn được thoả mãn thì kết quả kiểm định sẽ không thuyết phục. Các kiểm định đã đề cập ở Chương V được gọi chung là kiểm định tham số (Parametric test) vì chúng gắn liền với một tham số nào đó của tổng thể.

Nhưng trong phân tích dữ liệu, không phải lúc nào bạn cũng gặp được các tình huống thoả mãn hoàn toàn các giả định cần thiết này, đặc biệt khi bạn chỉ có các mẫu nhỏ. Lúc này bạn phải nhờ đến những thủ tục đòi hỏi những giả định ít nghiêm ngặt hơn về phân phối dữ liệu, những thủ tục này được gọi là kiểm định với phân phối bất kỳ hay còn gọi là kiểm định phi tham số (Nonparametric test)

Nhược điểm của kiểm định phi tham số là khả năng tìm ra được những sai biệt thật sự của chúng kém hơn trong những trường hợp mà các giả định của thủ tục kiểm định có tham số (Parametric test) được thoả mãn. Nói cách khác Nonparametric test không mạnh như những kiểm định có tham số vì nó bỏ qua một số thông tin có giá trị. Như trong kiểm định dấu (Sign test) sắp thảo luận ở phần sau đây bạn sẽ thấy giá trị thực của dữ liệu bị thay thế bởi hạng của chúng vì vậy bạn đã lãng phí một lượng lớn thông tin.

Như vậy kiểm định phi tham số chỉ hữu dụng cho những trường hợp chúng ta không thể sử dụng các kiểm định tham số như với mẫu nhỏ thì vi phạm giả định về phân phối chuẩn (mẫu lớn sẽ được coi như tiệm cận chuẩn). Các kiểm định phi tham số cũng hữu dụng khi mẫu có những giá trị quan sát bất thường (Outlier) vì những giá trị nằm xa trung tâm này sẽ không gây ảnh hưởng lớn đến kết quả như khi

chúng được sử dụng trong các thủ tục kiểm định có tham số căn cứ trên những số thống kê dễ bị ảnh hưởng như trung bình (vì gắn liền với những tham số nên chúng mới có tên là kiểm định tham số). Kiểm định phi tham số cũng phù hợp trong các trường hợp dữ liệu hiện có của chúng ta là loại dữ liệu định danh (nominal) hay dữ liệu thứ bậc (ordinal), hoặc khi các dữ liệu khoảng cách (interval) không có phân phối chuẩn một cách rõ ràng. Ta có thể xác định các mức ý nghĩa đối với các kiểm định phi tham số bất chấp hình dạng phân phối của tổng thể bởi vì các kiểm định phi tham số dựa vào hạng của dữ liệu.

Trong nội dung chương này chúng ta sẽ tìm hiểu về các loại kiểm định phi tham số sau đây như sự thay thế cho một số kiểm định tham số đã gặp.

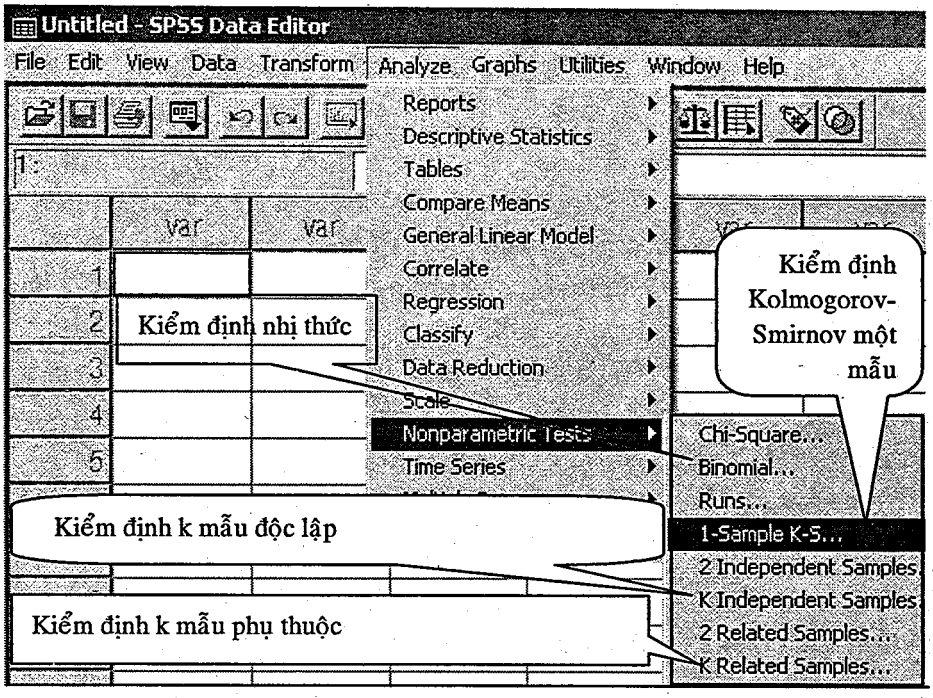
1. Kiểm định dấu (Sign test) 2. Kiểm định dấu và hạng Wilcoxon (Wilcoxon test) 3. Kiểm định McNemar	⟺ Kiểm định sự bằng nhau của 2 trị trung bình trong trường hợp mẫu phối hợp từng cặp (Paired-Samples t test)
4. Kiểm định Mann-Whitney	⟺ Independent-Samples t test
5. Kiểm định Kruskal-Wallis	⟺ One-way ANOVA

Ngoài ra chúng ta còn thảo luận thêm về kiểm định Kolmogorov-Smirnov và kiểm định Chi-bình phương để kiểm định giả thuyết về phân phối của tổng thể. Chú ý rằng kiển định Chi- bình phương chúng ta nghiên cứu ở Chương IV cũng là một dạng kiểm định phi tham số cho tình huống kiểm định giả thuyết về mối liên hệ.

Bạn đọc muốn tìm hiểu kỹ về bản chất của các kiểm định phi tham số này có thể tham khảo thêm các tài liệu về thống kê.

Các lệnh để thực hiện kiểm định phi tham số của SPSS đều thuộc menu Analyze > Nonparametric test (xem hình sau)

Hình 7.1



1. KIỂM ĐỊNH DẤU (SIGN TEST) VÀ KIỂM ĐỊNH MCNEMAR

Kiểm định dấu là một thủ tục phi tham số đơn giản nhất được sử dụng cho hai mẫu liên hệ trong tình huống so sánh sự khác nhau của trị trung bình của 2 tổng thể mà không cần giả thiết nào về hình dạng của hai phân phối này.

Trong phần này chúng ta sẽ xem xét lại ví dụ minh họa về cải tiến sản phẩm đậu phộng rang vì chúng ta có thể nghi ngờ rằng mẫu không thoả mãn giả định về phân phối chuẩn của tổng thể các chênh lệch, do đó kết quả kiểm định của chúng ta ở phần Paired- samples t test chưa chắc chính xác.

Giả thiết không (H_0) đặt ra ở đây là không có khuynh hướng thích sản phẩm loại này hơn so với sản phẩm loại kia trong toàn bộ người tiêu dùng.

1.1 Trình tự tiến hành kiểm định dấu

So sánh từng cặp điểm đánh giá của từng người đối với 2 tình huống, cụ thể ở đây bạn sẽ lấy điểm người đó đánh giá sản phẩm sau cải tiến trừ đi điểm chính người đó đánh giá sản phẩm trước cải tiến. Nhưng bạn không quan tâm đến độ lớn chênh lệch mà chỉ cần ghi lại dấu của phép trừ (dấu + khi điểm cho sản phẩm sau cải tiến cao hơn; dấu – khi điểm cho sản phẩm trước cải tiến cao hơn), những trường hợp cho điểm ngang nhau thì bỏ qua không xét (đúng như tên gọi của phép kiểm định này là kiểm định dấu, nó chỉ quan tâm đến dấu). Nếu đi theo giả thuyết Ho của bài toán rằng trong tổng thể người tiêu dùng không có khuynh hướng thích sản phẩm loại này hơn so với sản phẩm loại kia thì xét về dấu khả năng gặp một dấu + hay gặp một dấu – là ngang nhau, tức là khả năng này = 0,5. Giả thuyết Ho của bài toán kiểm định của chúng ta lúc này cụ thể hoá thành:

H_0: Xác suất để một người tiêu dùng nào đó đánh giá sản phẩm sau cải tiến cao hơn sản phẩm trước cải tiến (hoặc ngược lại) là 0,5.

Chúng ta thực hiện phép trừ trên mẫu và ghi nhận lại các dấu như sau

Bảng 7.1

STT	Trước	Sau	Δ	Dấu	STT	Trước	Sau	Δ	Dấu
1	7	8	+1	+	11	7	9	+2	+
2	8	9	+1	+	12	7	5	-2	-
3	6	5	-1	-	13	8	9	+1	+
4	8	9	+1	+	14	9	10	+1	+
5	7	8	+1	+	15	7	7	0	0
6	7	9	+2	+	16	7	9	+2	+
7	7	7	0	0	17	8	7	-1	-
8	6	7	+1	+	18	7	9	+2	+
9	8	7	-1	-	19	6	6	0	0
10	6	8	+2	+	20	8	8	0	0

Giá trị kiểm định T chính bằng số lượng các dấu cộng đếm được (ở Bảng 7.1 có 12 dấu + và 4 dấu -) sẽ được đem so với các giá trị giới hạn được tính ra từ phân phối nhị thức. Với các kiểm định phi tham số bạn cũng vẫn có thể sử dụng quy tắc p-value. Ở đây chúng ta

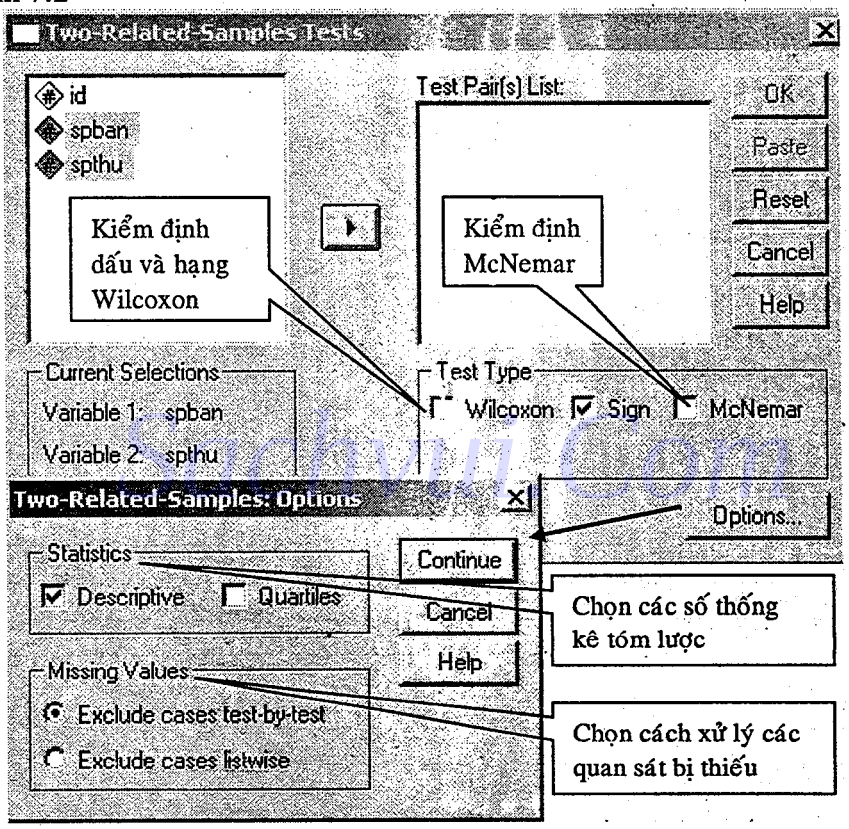
không đi sâu vào công thức tính toán và sẽ áp dụng quy tắc so sánh mức ý nghĩa quan sát.

Chú ý rằng khi mẫu lớn chúng ta sẽ thực hiện kiểm định dấu dựa trên phân phối chuẩn thay cho phân phối nhị thức.

1.2 Thực hiện kiểm định dấu bằng SPSS

1. Bạn vào menu chọn lệnh Analyze> Nonparametric Tests >2 Related Samples, lệnh này sẽ mở ra hộp thoại như trong hình dưới đây

Hình 7.2



2. Trong hộp thoại này, hãy chọn theo thứ tự biến *spban* rồi biến *spthu*, và bấm nút mũi tên đưa vào khung Test Pair.

Trong phần Test Type, chúng ta có thể chọn một trong ba loại kiểm định sau:

- Wilcoxon: kiểm định dấu và hạng Wilcoxon (sẽ được thảo luận ở phần kế tiếp)

- Sign: kiểm định dấu. Trong ví dụ này chúng ta nhấp vào ô vuông trước chữ Sign để chọn phương pháp kiểm định dấu.
- McNemar: kiểm định McNemar có thể được coi như là kiểm định dấu áp dụng cho các biến thay phiên (Dichotomous) là biến chỉ có hai nhóm hay hai biểu hiện, ví dụ: giới tính nam hay nữ, quyết định mua hay không mua, thái độ thích hay không thích... Nếu biến có nhiều hơn hai nhóm hay 2 biểu hiện, bạn có thể mã hóa lại biến để tạo thành biến chỉ có hai biểu hiện. Nó được tính toán như kiểm định Chi-bình phương thông thường. Với 2 biến Dichotmous và số lượng mẫu lớn, nếu bạn bình phương giá trị z trong kết quả của kiểm định dấu và hạng bạn sẽ được giá trị χ^2 của kiểm định McNemar, nhưng mức ý nghĩa quan sát cho 2 số thống kê của 2 tình huống thì như nhau.

  Chú ý rằng McNemar được tính toán như kiểm định Chi-bình phương thông thường. Nhưng tại sao chúng ta không sử dụng McNemar cho bảng chéo Crosstabs? Đơn giản là vì McNemar kiểm định giả thuyết khác với giả thuyết của bảng chéo.

3. Vào nút Options... chọn các tuỳ chọn giúp bạn tính toán các thông số thống kê tóm lược tùy ý hay chọn cách xử lý các quan sát thiếu giá trị tương tự như các kiểm định trước.

4. Cuối cùng nhấp nút OK. Kết quả sẽ xuất hiện.

Bảng 7.2 **Descriptive Statistics**

	N	Mean	Std. Deviation	Minimum	Maximum
san pham dang ban	20	7.20	.834	6	9
san pham cai tien	20	7.80	1.399	5	10

Bảng 7.3 **Frequencies**

		N
san pham cai tien - san pham dang ban	Negative Differences(a)	4
	Positive Differences(b)	12
	Ties(c)	4
	Total	20

a san pham cai tien < san pham dang ban
b san pham cai tien > san pham dang ban
c san pham cai tien = san pham dang ban

Bảng 7.4 **Test Statistics(b)**

	san pham cai tien - san pham dang ban
Exact Sig. (2-tailed)	.077(a)

a Binomial distribution used.
b Sign Test

Ở Bảng 7.3 dòng thứ nhất liệt kê số lượng các dấu - và dòng thứ hai là số các dấu +, dòng thứ ba là số trường hợp bằng điểm nhau bị bỏ qua không xét, kết quả không khác bảng tính tay của chúng ta.

Mức ý nghĩa quan sát khi kiểm định hai phía là 0,077 chưa đủ để ta chấp nhận giả thiết phân phối của điểm đánh giá cho sản phẩm trước cải tiến và điểm đánh giá cho sản phẩm cải tiến là không giống nhau (với độ tin cậy 95%). Như vậy theo kết quả của kiểm định này thì người tiêu dùng nói chung chưa có khuynh hướng thích loại sản phẩm cải tiến hơn.

Bạn thấy là với mức ý nghĩa ôn hoà 0,05 mà ta đã chọn, ở Pair-Samples t Test ta có kết luận là có sự khác biệt giữa cách đánh giá của người tiêu dùng về sản phẩm đậu phộng rang trước và sau khi cải tiến, cụ thể khác biệt trung bình là 0,6 điểm; nhưng với Sign test ta lại không kết luận được như vậy. Như vậy là trong tình huống này kiểm định tham số đã thể hiện ưu thế vượt trội của nó so với kiểm định phi tham số khi mà các giả định thống kê được thoả mãn nhờ nó tận dụng triệt để thông tin trên mẫu. Vậy có thực là giả định về phân phối chuẩn của tổng thể các chênh lệch ở thử nghiệm về đậu phộng cải tiến được thoả mãn không? Chúng ta sẽ kiểm chứng điều này ở phần kiểm định Kolmogorov-Smirnov một mẫu.

2. KIỂM ĐỊNH DẤU VÀ HẠNG WILCOXON (WILCOXON SIGNED-RANK TEST)

Kiểm định dấu cho mẫu phối hợp từng cặp mà chúng ta vừa thảo luận ở phần trên chỉ xét chiều hướng (- hay +) của chênh lệch giữa các cặp quan sát và bỏ qua độ lớn của các chênh lệch này nên mặc dù đơn giản nhưng nó không mạnh. Kiểm định dấu và hạng Wilcoxon khắc phục nhược điểm này vì nó sử dụng luôn cả thông tin về độ lớn

của các chênh lệch với giả thuyết rằng phân phối của 2 tổng thể là giống nhau.

Đặt giả thuyết Ho: Hai trị trung bình của 2 tổng thể là như nhau

Để thực hiện kiểm định dấu và hạng Wilcoxon, các chênh lệch được xếp hạng theo độ lớn không tính đến dấu của chúng. Trong trường hợp điểm cân bằng (tức được đánh giá bằng điểm nhau) thì hạng của chúng được tính bình quân. Sau đó ta tính tổng các hạng đối với các chênh lệch dương và đối với các chênh lệch âm. Giá trị tổng hạng nào nhỏ nhất sẽ là giá trị T của kiểm định. Quy tắc quyết định là ở mức ý nghĩa α, bác bỏ giả thuyết Ho nếu $T < T_\alpha$ với T_α là giá trị tra bảng phân phối của kiểm định Wilcoxon.

Trong trường hợp mẫu lớn cũng có thể dùng phân phối chuẩn thay cho phân phối của kiểm định Wilcoxon.

Chúng ta sẽ thực hiện lại ví dụ về sản phẩm đậu phộng rang cải tiến. Bạn hãy tham khảo thủ tục tính tay trước khi xem cách thực hiện bằng SPSS.

2.1 Trình tự tiến hành kiểm định dấu và hạng Wilcoxon

Bảng 7.5

QS	Trước	Sau	Δ	|Δ|	Dấu	Thứ hạng	Hạng	Hạng (+)	Hạng (-)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
7	7	7	0	0	0				
15	7	7	0	0	0				
19	6	6	0	0	0				
20	8	8	0	0	0				
1	7	8	+1	1	+	1	5,5	5,5	
2	8	9	+1	1	+	2	5,5	5,5	
3	6	5	-1	1	-	3			5,5
4	8	9	+1	1	+	4	5,5	5,5	
5	7	8	+1	1	+	5	5,5	5,5	
8	6	7	+1	1	+	6	5,5	5,5	
9	8	7	-1	1	-	7			5,5
13	8	9	+1	1	+	8	5,5	5,5	
14	9	10	+1	1	+	9	5,5	5,5	
17	8	7	-1	1	-	10			5,5
6	7	9	+2	2	+	11	13,5	13,5	
10	6	8	+2	2	+	12	13,5	13,5	

11	7	9	+2	2	+	13	13,5	13,5	
12	7	5	-2	2	-	14			13,5
16	7	9	+2	2	+	15	13,5	13,5	
18	7	9	+2	2	+	16	13,5	13,5	
Tổng hạng								106	30
Hạng TB								8,83	7,50

Chú ý là lúc này thứ tự các quan sát ở cột (1) bị đảo lộn vì nó được sắp lại trật tự theo trật tự tăng dần của giá trị tuyệt đối của các chênh lệch.

Cột (4): Tính các chênh lệch, các trường hợp chênh lệch bằng 0 được bỏ qua.

Cột (5): Sắp xếp giá trị tuyệt đối của các chênh lệch theo thứ tự tăng dần, dấu của chúng vẫn được bảo lưu ở cột (6)

Cột (7): Xếp hạng cho các giá trị tuyệt đối ở cột (5) theo thứ tự đếm từ trên xuống. Hạng theo thứ tự này thể hiện ở cột (7).

Từ hạng 1 đến hạng 10 đều mang giá trị tuyệt đối của chênh lệch là 1, do đó hạng thật sự của chúng ngang nhau và được tính bằng hạng trung bình của các hạng theo thứ tự, tức là bằng (1+2+..+10)/10 = 5,5

Tính tương tự cho các hạng còn lại và thể hiện kết quả sắp hạng theo độ lớn không tính đến dấu trên cột (8)

Tách riêng các hạng - và + để tính tổng cộng hạng riêng cho các chênh lệch dương và âm ở cột (9) và (10). Ta có tổng hạng + là 106, và tổng hạng – là 30.

Tra bảng phân phối của kiểm định Wilcoxon với n=16 (vì không xét bốn cặp có chênh lệch = 0) thì $T_{0,05} = 36$. Căn cứ quy tắc kiểm định ta bác bỏ Ho. Như vậy có thể thấy nhờ sử dụng nhiều thông tin hơn mà kiểm định dấu và hạng Wilcoxon mạnh hơn và đã giúp kết luận được đậu phộng rang cải tiến đã được ưa thích hơn (số lượng chênh lệch dương nhiều hơn hẳn số lượng chênh lệch âm)

2.2 Thực hiện kiểm định dấu và hạng Wilcoxon trên SPSS

Để thực hiện kiểm định dấu và hạng Wilcoxon, chúng ta thực hiện lệnh giống như kiểm định dấu, nhưng trong phần chọn loại kiểm định Test Type ta chọn loại kiểm định là Wilcoxon (xem trong Hình 7.2)

Bên dưới là những bảng kết quả của kiểm định dấu và hạng Wilcoxon với dữ liệu trong file *Dauphong*. Hạng trung bình của 4

chênh lệch âm là 7,5 và hạng trung bình của 12 chênh lệch dương là 8,83. Và có 4 quan sát có giá trị bằng nhau đối với cả hai biến. Mức ý nghĩa quan sát của kiểm định này là 0,042. Do đó ta bác bỏ giả thiết Ho. Kết luận rút ra theo kiểm định dấu và hạng Wilcoxon là người tiêu dùng nói chung có khuynh hướng thích sản phẩm cải tiến hơn. Một lần nữa rõ ràng ví dụ này đã chứng tỏ kiểm định dấu và hạng Wilcoxon mạnh hơn kiểm định dấu.

Bảng 7.6 **Ranks**

		N	Mean Rank	Sum of Ranks
san pham cai tien - san pham dang ban	Negative Ranks	4(a)	7.50	30.00
	Positive Ranks	12(b)	8.83	106.00
	Ties	4(c)		
	Total	20		

a san pham cai tien < san pham dang ban
b san pham cai tien > san pham dang ban
c san pham cai tien = san pham dang ban

Bảng 7.7 **Test Statistics(b)**

	san pham cai tien - san pham dang ban
Z	-2.034(a)
Asymp. Sig. (2-tailed)	.042

a Based on negative ranks.
b Wilcoxon Signed Ranks Test

3. KIỂM ĐỊNH MANN-WHITNEY 2 MẪU ĐỘC LẬP

Kiểm định Mann-Whitney là phép kiểm định phổ biến nhất để kiểm định giả thuyết về sự bằng nhau của trung bình 2 mẫu độc lập (Independent Samples design) khi các giả định không thoả mãn. Giống các kiểm định phi tham số, kiểm định Mann-Whitney cũng không yêu cầu các giả định về hình dạng của phân phối đang xem xét.

Tương ứng với giả thuyết không: 2 trung bình của 2 nhóm tổng thể bằng nhau của Independent Samples t test, trong tình huống hình dạng của phân phối không xác định của kiểm định phi tham số Mann-Whitney ta có thể suy ra rằng hình dạng của 2 phân phối tổng thể phải giống nhau, điều này cũng có nghĩa là phương sai tổng thể cho 2 nhóm phải như nhau. Tóm lại không cần biết hình dáng phân

phối của 2 nhóm này là gì nhưng chúng phải có cùng hình dáng với nhau để các đại lượng thể hiện độ tập trung cũng như phân tán của chúng không khác nhau.

Như vậy Mann-Whitney được dùng để kiểm định giả thuyết về sự giống nhau của 2 phân phối tổng thể hay nói cách khác là hai mẫu độc lập phải xuất phát từ hai tổng thể có phân phối giống nhau.

Kiểm định này không đòi hỏi biến nghiên cứu phải là biến khoảng cách, mà chỉ cần biến xếp hạng là đủ. Trong phần này chúng ta sẽ xem xét một ví dụ thực tế về khảo sát tuổi thọ của các bóng đèn tròn hiệu A và hiệu B. Có 7 bóng hiệu A và 5 bóng hiệu B được khảo sát và ghi nhận tuổi thọ của chúng, kết quả quan sát được trình bày trong Bảng 7.8 (file *Bongden* trong tập hợp dữ liệu dùng kèm với sách).

Giả thuyết Ho của chúng ta là phân phối tuổi thọ của hai loại bóng đèn A và B giống nhau. Bản chất thực sự của giả thuyết này là tuổi thọ trung bình của 2 loại bóng đèn như nhau.

3.1 Trình tự thực hiện kiểm định Mann-Whitney 2 mẫu độc lập

Giống như kiểm định Wilcoxon, để thực hiện kiểm định trước hết các quan sát từ cả hai mẫu được kết hợp với nhau và xếp hạng từ giá trị nhỏ nhất đến giá trị lớn nhất. Những trường hợp đồng hạng thì được thay thế bằng hạng trung bình. Kết quả xếp hạng cũng được thể hiện trong Bảng 7.8.

Bảng 7.8 thể hiện tuổi thọ của 12 bóng đèn được quan sát (giờ)

QS	hiệu A	hạng	QS	hiệu B	hạng
1	2400	**3,5**			
2	3800	11	8	3900	12
3	2300	2	9	3200	9
4	2600	5	10	2900	8
5	2400	**3,5**	11	3400	10
6	2800	7	12	2700	6
7	2100	1			
Tổng hạng		33			45
Hạng TB		4,71			9

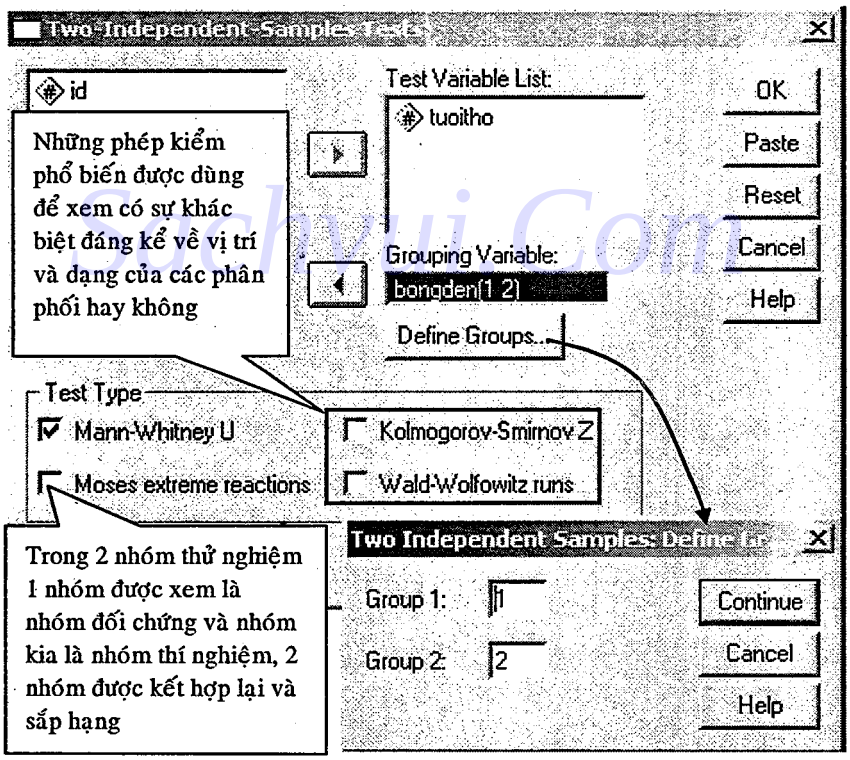
Đại lượng thống kê dùng để kiểm định hai phân phối giống nhau được tính từ tổng hạng của từng nhóm (loại bóng đèn). Nếu các

nhóm có phân phối giống nhau, thì phân phối các hạng của mẫu quan sát cũng phải giống nhau. Nếu một nhóm nào đó có hạng nhỏ hơn thì chúng ta có cơ sở nghi ngờ rằng hai phân phối này khác nhau.

3.2 Thực hiện kiểm định Mann-Whitney hai mẫu độc lập trên SPSS

1. Từ menu chọn Analyze > Nonparametric Tests > 2 Independent Samples. Lệnh này mở ra một hộp thoại 2-Independent-Sample Test như hình sau:

Hình 7.2



2. Các biến có trong file dữ liệu sẽ hiện trong danh sách biến nguồn. Bạn hãy chọn một hay nhiều biến cần kiểm định đưa vào khung Test Variable List và một biến phân nhóm đưa vào ô Grouping Variables, biến phân nhóm sẽ chia các quan sát thành 2 nhóm. SPSS sẽ hỏi chúng ta 2 nhóm muốn so sánh. Hãy nhấp chuột vào nút Define Groups... để xác định hai nhóm cần so sánh, hộp thoại nhỏ sẽ mở ra và hãy gõ mã số của từng nhóm vào hai ô như hình trên.

File ví dụ *Bongden* của chúng ta quản lý 3 loại bóng đèn thuộc 3 hiệu A, B và C lần lượt được mã hoá là 1, 2 và 3. Nhưng mối quan tâm của chúng ta là tuổi thọ trung bình của 2 loại bóng đèn hiệu A và B do đó trong hộp thoại Define Group ta nhập 1 và 2.

3. Sau khi xác định các giá trị của mã nhóm, bạn hãy nhấp chuột tại nút Continue trở về, nhớ chọn loại kiểm định Mann-Whitney U, và nhấp tiếp nút OK để thực hiện kiểm định.

Những bảng dưới thể hiện kết quả kiểm định Mann-Whitney đối với các dữ liệu trong ví dụ của chúng ta.

Bảng 7.9 **Ranks**

loai bong den		N	Mean Rank	Sum of Ranks
tuoi tho bong den (gio)	bong den hieu A	7	4.71	33.00
	bong den hieu B	5	9.00	45.00
	Total	12		

Bảng 7.10 **Test Statistics(b)**

	tuoi tho bong den (gio)
Mann-Whitney U	5.000
Wilcoxon W	33.000
Z	-2.034
Asymp. Sig. (2-tailed)	.042
Exact Sig. [2*(1-tailed Sig.)]	.048(a)

a Not corrected for ties.
b Grouping Variable: loai bong den

Trong Bảng 7.9, hai dòng đầu thể hiện thông tin về tổng hạng và hạng trung bình của mỗi nhóm, hãy so sánh với bảng tính tay 6.8. Các mức ý nghĩa của Mann-Whitney U và Wilcoxon W như nhau, ta có thể tính được bằng cách chuẩn hóa chúng (tính bằng đơn vị lệch chuẩn - Z score). Nếu toàn bộ cỡ mẫu nhỏ hơn 30 thì mức xác suất chính xác (Exact Sig) dựa trên phân phối của U và W sẽ được thể hiện. Trong Bảng 7.10, mức ý nghĩa quan sát của ví dụ này là 0,042 và mức ý nghĩa quan sát chính xác là 0,048 (do cỡ mẫu nhỏ hơn 30). Vì mức ý nghĩa này khá nhỏ (và nhỏ hơn 5%) nên giả thiết tuổi thọ bóng đèn của hai hiệu A và B có phân phối giống nhau bị bác bỏ, và như vậy tuổi thọ trung bình của chúng khác nhau.

Một lần nữa, kiểm định Mann-Whitney chỉ đòi hỏi mẫu là ngẫu nhiên và các giá trị có thể sắp thứ tự. Những giả định này, đặc biệt là giả định ngẫu nhiên, không nên coi nhẹ, nhưng chúng ít nghiêm ngặt hơn trường hợp kiểm định trung bình t hai mẫu. Kiểm định t đòi hỏi các quan sát phải được rút ra từ các tổng thể có phân phối chuẩn với phương sai bằng nhau. Kiểm định Mann-Whitney được sử dụng thay thế cho kiểm định t khi các giả định trong kiểm định t không được thỏa mãn. Nếu thỏa các giả định cần thiết thì kiểm định t mạnh hơn kiểm định Mann-Whitney bởi vì kiểm định t sử dụng nhiều thông tin từ các dữ liệu hơn. Kiểm định Mann-Whitney thay thế các giá trị quan sát thực tế bằng các hạng và như vậy sẽ loại bỏ bớt các thông tin hữu ích.

4. KIỂM ĐỊNH KRUSKAL-WALLIS

Kiểm định Mann-Whitney được sử dụng để xem xét sự khác biệt về phân phối giữa 2 tổng thể từ các dữ liệu của hai mẫu độc lập. Để kiểm định sự khác biệt về phân phối giữa ba (hay nhiều hơn ba) tổng thể từ các dữ liệu mẫu của chúng, chúng ta sử dụng kiểm định Mann-Whitney mở rộng, có tên là kiểm định Kruskal-Wallis. Với bản chất này, kiểm định Kruskal-Wallis cũng là phương pháp kiểm định giả thuyết trị trung bình của nhiều nhóm tổng thể bằng nhau hay chính là phương pháp phân tích phương sai một yếu tố mà không đòi hỏi bất kỳ giả định nào về phân phối chuẩn của tổng thể. Khi chỉ có 2 nhóm tổng thể muốn so sánh trị trung bình, kiểm định Kruskal-Wallis tương tự kiểm định Mann-Whitney đã đề cập ở trước.

Thủ tục tính toán kiểm định Kruskal-Wallis cũng tương tự như thủ tục kiểm định Mann-Whitney. Tất cả các quan sát của ba nhóm được gộp lại với nhau để xếp hạng. Sau đó hạng của các quan sát trong từng nhóm được cộng lại, và đại lượng thống kê Kruskal-Wallis H được tính từ các tổng hạng này. Đại lượng H này xấp xỉ một phân phối Chi-bình phương với giả thuyết Ho là cả ba nhóm có phân phối giống nhau.

Ví dụ minh họa cho phần này sẽ mở rộng ví dụ trong kiểm định Mann-Whitney, ta sẽ xem xét xem phân phối của tuổi thọ bóng đèn của ba hiệu A, B và C có giống nhau không. Các giá trị quan sát được

trình bày trong Bảng 7.11, bảng này thể hiện thêm thông tin về 5 loại bóng đèn hiệu C (xem file *Bong den* trong tập hợp dữ liệu dùng kèm với sách)

Bảng 7.11 Tuổi thọ của 17 bóng đèn thuộc 3 hiệu được quan sát (giờ)

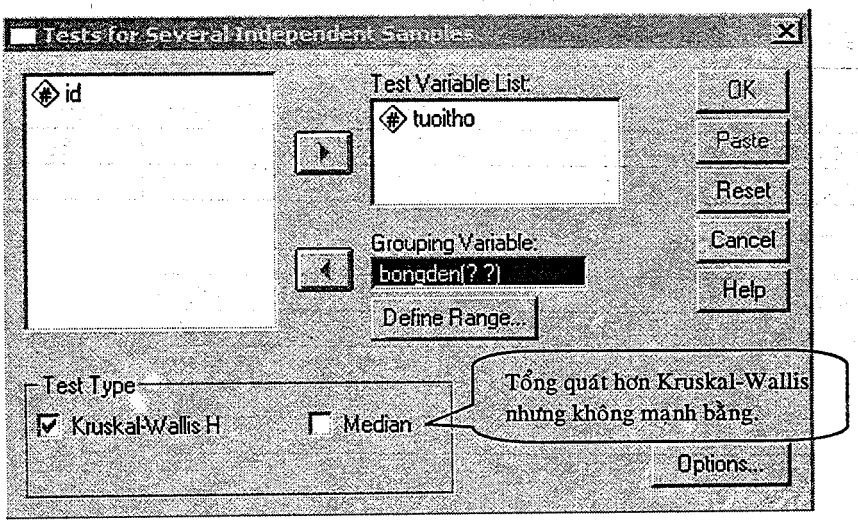
QS	hiệu A	QS	hiệu B	QS	hiệu C
1	2400	8	3900	13	3000
2	3800	9	3200	14	2100
3	2300	10	2900	15	1800
4	2600	11	3400	16	2200
5	2400	12	2700	17	2600
6	2800				
7	2100				

Giả thuyết Ho là không có sự khác biệt về tuổi thọ trung bình của 3 loại bóng đèn. Để thực hiện kiểm định 3 (hay nhiều hơn 3) mẫu độc lập, tiến hành như sau

1. Từ menu chọn Analyze>Nonparametric Tests>K Independent Samples... Lệnh này sẽ mở ra hộp thoại Test for Several Independent Samples như trong Hình 7.3:

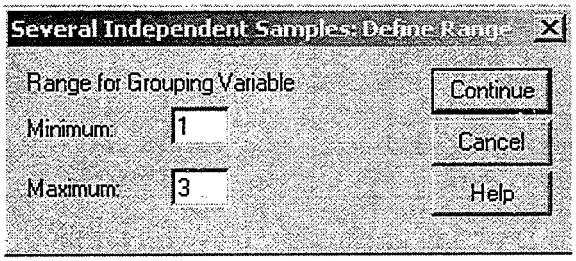
2. Các biến trong file dữ liệu sẽ hiện trong danh sách biến nguồn. Bạn hãy chọn một hay nhiều biến cần kiểm định đưa vào ô Test Variable List và một biến phân nhóm đưa vào ô Grouping Variables.

Hình 7.3



3. SPSS sẽ hỏi bạn cần so sánh khoảng nhóm nào. Hãy nhấp chuột vào nút Define Groups để xác định khoảng cho ba nhóm cần so sánh của ví dụ này, hộp thoại nhỏ Define Range sẽ mở ra và hãy gõ mã số 1 và mã số 3 vào hai ô như hình bên dưới. Bạn hãy xem lại phần hướng dẫn thực hiện phân tích One-Way ANOVA ở Chương V về phần này vì chúng tương tự nhau.

Hình 7.4



4. Sau khi xác định các giá trị của biến nhóm, bạn hãy nhấp chuột tại nút Continue để trở về hộp thoại chính, nhớ chọn loại kiểm định Kruskal-Wallis H, và nhấp tiếp nút OK để thực hiện kiểm định.

Kết quả phân tích phương sai một yếu tố Kruskal-Wallis được thể hiện trong từ Bảng 7.12 đến Bảng 7.14 . Kết quả Bảng 7.13 cho thấy nhóm bóng đèn B có hạng trung bình lớn nhất. Bảng 7.14 cho giá trị thống kê Chi-bình phương cho kiểm định Kruskal-Wallis là 6,455. Mức ý nghĩa quan sát là 0,04, do đó ta có thể kết luận rằng tuổi thọ bóng đèn của ba hiệu này khác nhau.

Bảng 7.12 **Descriptive Statistics**

	N	Mean	Std. Deviation	Minimum	Maximum
tuoi tho bong den (gio)	17	2717.65	595.016	1800	3900
loai bong den	17	1.88	.857	1	3

Bảng 7.13 **Ranks**

loai bong den		N	Mean Rank
tuoi tho bong den (gio)	bong den hieu A	7	8.00
	bong den hieu B	5	13.60
	bong den hieu C	5	5.80
	Total	17	

Bảng 7.14 **Test Statistics(a,b)**

	tuoi tho bong den (gio)
Chi-Square	6.455
df	2
Asymp. Sig.	.040

a Kruskal Wallis Test
b Grouping Variable: loai bong den

Bạn có thể thử lại kiểm định Kruskal- Wallis với ví dụ ở phần One-Way ANOVA rồi so sánh kết quả ở hai tình huống kiểm định này.

Đến đây chắc bạn sẽ băn khoăn là bạn sẽ phải làm thế nào nếu không chắc mình nên sử dụng kiểm định tham số hay kiểm định phi tham số cho một tình huống cụ thể. Nếu vậy, bạn cứ tiến hành cả 2 cách, nếu cả hai cho cùng một kết luận thì không còn gì phải lo lắng. Nhưng nếu kết luận mà hai cách kiểm định đưa ra khác nhau thì sao? Lúc đó bạn phải cố gắng tìm hiểu xem tại sao lại có sự khác nhau đó. Ví dụ có một vài quan sát bất thường trong dữ liệu không vì nếu có chúng có thể ảnh hưởng đến trung bình và từ đó tác động lớn đến kết quả kiểm định của bạn. Lúc đó bạn phải khảo sát các bất thường này cẩn thận để quyết định cách xử lý đối với chúng. Nếu vấn đề nằm ở chỗ phân phối không chuẩn của dữ liệu, hãy xem xét xem bạn có thể chuyển hoá dữ liệu để hình dạng phân phối của chúng gần hơn với các giả định của kiểm định tham số không, từ đó bạn có thể sử dụng một trong những thủ tục kiểm định tham số mạnh hơn cho dữ liệu đã chuyển hoá.

5. KIỂM ĐỊNH CHI-BÌNH PHƯƠNG MỘT MẪU

Kiểm định Chi-bình phương được sử dụng khá phổ biến đối với các biến định tính (phân loại). Chúng ta đã xem xét việc sử dụng bảng chéo và kiểm định Chi-bình phương để xem xét sự liên hệ của một biến định tính này với một biến định tính khác, ví dụ trình độ học vấn có ảnh hưởng (tức là có liên quan) đến sự đánh giá của một người nào đó về tầm quan trọng của các yếu tố tinh thần trong cuộc sống không... Kiểm định Chi-bình phương còn được vận dụng để giải quyết nhiều yêu cầu nghiên cứu khác nữa. Trong phần này chúng ta sẽ sử dụng kiểm định Chi-bình phương để xem xét dữ liệu của chúng

ta phù hợp (thích hợp) đến mức độ nào với giả thuyết về phân phối của tổng thể

Ví dụ: Một công ty muốn nghiên cứu các vụ tai nạn lao động có xảy ra như nhau vào các ngày làm việc trong tuần không hay là nó có xu hướng tăng cao vào các ngày thứ Hai và các ngày cuối tuần. Ta lập luận rằng nếu giả thuyết cho rằng "các vụ tai nạn xảy ra với xác suất như nhau trong 6 ngày làm việc của tuần" là đúng thì xác suất xảy ra tai nạn của mỗi ngày phải bằng nhau và bằng 1/6. Với tổng số 32 vụ tai nạn lao động công ty đó thu thập được trong vòng 5 năm qua tại các nhà máy của công ty, số lượng các vụ tai nạn trong từng ngày phải bằng nhau và =1/6 x 32=5,3 vụ.

Các vụ tai nạn được mã hóa khi nhập liệu vào file có tên *Tai nan* (trong trong tập hợp dữ liệu dùng kèm với sách) như sau: Những vụ xảy ra vào thứ Hai có mã là 1, thứ Ba mã là 2,.., ngày thứ Bảy mã là 6. Bảng 7.15 cho thấy tần số xảy ra tai nạn tại các ngày trong tuần của 32 vụ tai nạn, dường như các vụ tai nạn xảy ra không đồng đều giữa 6 ngày làm việc trong tuần (xem cột Frequency và Percent)

Để kiểm định giả thuyết Ho: khả năng xảy ra tai nạn lao động vào các ngày làm việc trong tuần như nhau, ta sử dụng kiểm định Chi-bình phương một mẫu.

Bảng 7.15

		Frequency	Percent	Valid Percent	Cumulative Percent
	hai	7	21.9	21.9	21.9
	ba	3	9.4	9.4	31.3
	tu	3	9.4	9.4	40.6
	nam	2	6.3	6.3	46.9
	sau	5	15.6	15.6	62.5
	bay	12	37.5	37.5	100.0
	Total	32	100.0	100.0	

Trước tiên các dữ liệu được phân thành các nhóm, trong ví dụ này là phân theo các ngày xảy ra tai nạn, sau đó ta tính tần số lý thuyết hay còn gọi là tần số kỳ vọng (Expected frequency) xảy ra tai nạn tại các ngày trong tuần. Tần số lý thuyết là tần số xảy ra nếu giả thuyết đã

cho là đúng (trong bài toán kiểm định này, về mặt thống kê giả thuyết H_0 cần kiểm định là: xác suất xảy ra tai nạn của các ngày trong tuần là bằng nhau). Tần số lý thuyết đó chính là 5,3 vụ tai nạn đã tính ở trên. Đại lượng thống kê Chi-bình phương được tính như công thức sau:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Trong đó O_i là tần số quan sát thực tế của ngày thứ i, E_i là tần số lý thuyết của ngày thứ i, và k là số ngày trong tuần.

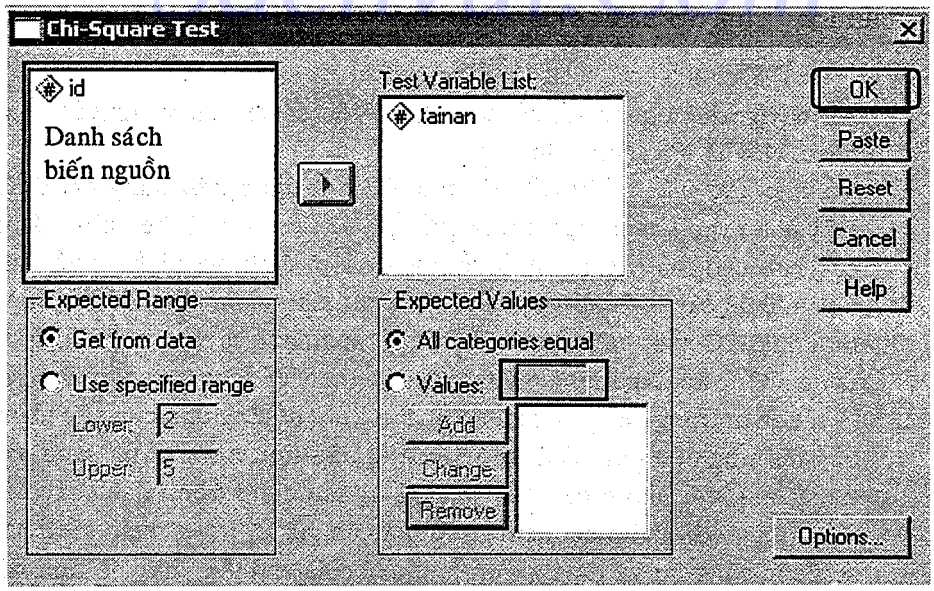
Tra bảng Chi-bình phương để tìm giá trị giới hạn và so sánh với χ^2. Đại lượng χ^2 sẽ lớn nếu các tần số quan sát và tần số lý thuyết khác nhau nhiều nên giả thuyết Ho sẽ có khả năng bị bác bỏ nếu χ^2 lớn.

Thực hiện kiểm định Chi-bình phương một mẫu bằng SPSS

1. Từ menu chọn Analyze>Nonparametric Tests>Chi-Square...

Lệnh này mở ra một hộp thoại Chi-Square Test như trong hình sau:

Hình 7.5



2. Các biến trong file dữ liệu sẽ xuất hiện trong danh sách biến nguồn. Hãy chọn một hay nhiều biến muốn kiểm định bằng cách nhấp chuột tại tên biến đó (biến đó sẽ được chiếu sáng) và đưa nó sang khung Test Variable List.

3. Nếu tiếp theo bạn nhấp chuột vào nút OK thì SPSS sẽ thực hiện kiểm định Chi-bình phương với những lựa chọn mặc định là tần số lý thuyết cho từng nhóm trong biến nghiên cứu sẽ bằng nhau và tất cả các nhóm có mặt đều được nghiên cứu tính đồng đều. Mỗi biến nghiên cứu được đưa vào Test Variable List sẽ có một kết quả kiểm định riêng biệt.

Ngoài ra, chúng ta có thể lựa chọn một số thuộc tính sau cho phép kiểm định:

Phần Expected Range: có một trong hai cách chọn các nhóm của biến quan sát như sau:

- Get from data: mỗi biểu hiện của biến định tính được định nghĩa như một nhóm để kiểm định tính đồng đều. Đây là lựa chọn mặc định.
- Use specified range: chỉ kiểm định trong khoảng giới hạn dưới và giới hạn trên đã ấn định trong ô Lower và Upper. Các quan sát có các giá trị nằm ngoài phạm vi này sẽ không được nghiên cứu. Ví dụ nếu bạn đã chỉ định giới hạn dưới là 2 và giới hạn trên là 5 thì kiểm định chỉ nghiên cứu tính đồng đều của các ngày từ thứ Ba đến thứ Sáu.

Phần Expected Values: có thể chọn một trong hai cách xác định tần số lý thuyết sau:

- All categories equal: tất cả các nhóm đều có tần số lý thuyết bằng nhau. Đây là lựa chọn mặc định.
- Values: xác suất lý thuyết của các nhóm là do người dùng chỉ định. Hãy nhập vào khung Value một giá trị tỷ lệ (lớn hơn 0) theo phán đoán của bạn cho mỗi nhóm của biến được kiểm định, và nhấp chuột vào nút Add để cập nhật nó vào danh sách. Mỗi lần nhấn nút Add thì giá trị mới nhập sẽ xuất hiện ở vị trí kế tiếp trong danh sách các giá trị tỷ lệ được nhập vào. Nhớ rằng thứ tự của các giá trị tỷ lệ được nhập vào này là rất quan trọng, vì chúng tương ứng với thứ tự tăng dần của các nhóm của biến nghiên cứu. Giá trị được Add đầu tiên trong danh sách này tương ứng với nhóm đầu tiên và giá trị sau cùng trong danh sách sẽ tương ứng với nhóm cuối cùng. Tất cả các giá trị trong danh sách này sẽ được cộng lại và từng giá trị lại được chia cho tổng này để tính xác suất của tần số lý thuyết trong từng nhóm tương ứng. Ví dụ một danh sách các giá trị nhập vào gồm 2,1,1,1 (với khai báo ở Use specified

range là Lower:1 và Upper: 4) sẽ cho biết xác suất lý thuyết của bốn nhóm thứ Hai, thứ Ba, thứ Tư và Thứ Năm của biến ta đang nghiên cứu lần lượt là 2/5,1/5,1/5 và 1/5.

Để loại bỏ một giá trị, hãy chọn nó và nhấn nút Remove, Để thay đổi một giá trị thì chọn nó và nhập vào Value giá trị mới rồi nhấn nút Change.

Dưới đây là những bảng kết quả kiểm định Chi-bình phương một mẫu

Bảng 7.16 **Descriptive Statistics**

	N	Mean	Std. Deviation	Minimum	Maximum
thu	32	3.97	2.055	1	6

Bảng 7.17

	Observed N	Expected N	Residual
hai	7	5.3	1.7
ba	3	5.3	-2.3
tu	3	5.3	-2.3
nam	2	5.3	-3.3
sau	5	5.3	-.3
bay	12	5.3	6.7
Total	32		

Bảng 7.18 **Test Statistics**

	thu
Chi-Square(a)	13.000
df	5
Asymp. Sig.	.023

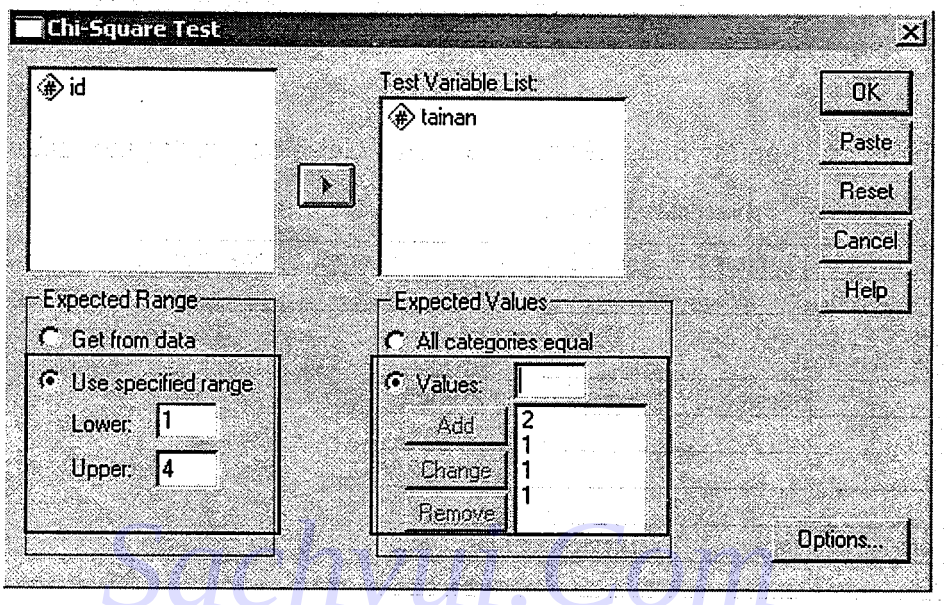
a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 5.3.

Bảng 7.18 cho thấy kết quả kiểm định từ các dữ liệu thu thập được. Mức ý nghĩa quan sát = 0,023 là khá nhỏ, do đó ta có bằng chứng để bác bỏ giả thiết khả năng xảy ra tai nạn các ngày trong tuần như nhau. Tai nạn có nhiều khả năng xảy ra vào ngày đầu tuần và nhất là 2 ngày cuối tuần, do đó công ty nên áp dụng các biện pháp đặc biệt để đề phòng tai nạn lao động vào những ngày này.

Bạn có thể tiến hành lại ví dụ này nhưng với những lựa chọn như Hình 7.6 dưới đây và xem thử mức ý nghĩa quan sát lúc này có phải là 0,919. Nếu vậy thì chúng ta có cơ sở để kết luận rằng khả năng

xảy ra tai nạn lao động ở ngày thứ Hai gấp đôi 3 ngày làm việc còn lại là thứ Ba, Tư, Năm không?

Hình 7.6



6. KIỂM ĐỊNH KOLMOGOROV-SMIRNOV MỘT MẪU

Kolmogorov-Smirnov test được sử dụng để kiểm định giả thuyết phân phối của dữ liệu có phù hợp với phân phối lý thuyết. Nó tiến hành xét các sai lệch tuyệt đối lớn nhất giữa 2 đường phân phối tích luỹ thực nghiệm và lý thuyết, sai lệch tuyệt đối càng lớn, giả thuyết Ho càng dễ bị bác bỏ.

Chúng ta biết rằng giả định về phân phối chuẩn của tổng thể là một giả định rất quan trọng đối với nhiều phép kiểm định tham số, có một số phương pháp thường được sử dụng để khảo sát tổng thể có phân phối chuẩn hay không như dựng đồ thị Q-Q Plot mà ta sẽ gặp lại ở Chương IX, hay tiến hành kiểm định Jacque-Berra, kiểm định Kolmogorov-Smirnov ... Không thể phủ nhận Q-Q Plot là một phương pháp khảo sát trực quan rất tốt nhưng nó lại không căn cứ trên một giả thuyết thống kê được kiểm định chặt chẽ do đó độ thuyết phục không cao. Trong tình huống cần đến một phương pháp kiểm định

chuẩn tắc cho giả thuyết biến có phân phối chuẩn chúng ta có thể sử dụng kiểm định One-sample Kolmogorov-Smirnov của SPSS.

Ví dụ: Cuối phần kiểm định dấu, chúng ta có đặt ra một câu hỏi là "Có thực là giả định về phân phối chuẩn của tổng thể các chênh lệch ở thử nghiệm về đậu phộng cải tiến được thoả mãn không?" Bây giờ chúng ta sẽ kiểm chứng điều này bằng kiểm định Kolmogorov-Smirnov.

Lúc này biến bạn cần kiểm định giả thuyết về phân phối chuẩn của tổng thể chính là Δ của Bảng 7.1, nó đại diện cho chênh lệch giữa điểm số ưa thích của sản phẩm sau khi cải tiến và sản phẩm trước khi cải tiến. Để tính được các giá trị của Δ như ở Bảng 7.1 trên SPSS chúng ta sẽ sử dụng lệnh Compute đã giới thiệu qua ở Chương I.

Lệnh Compute

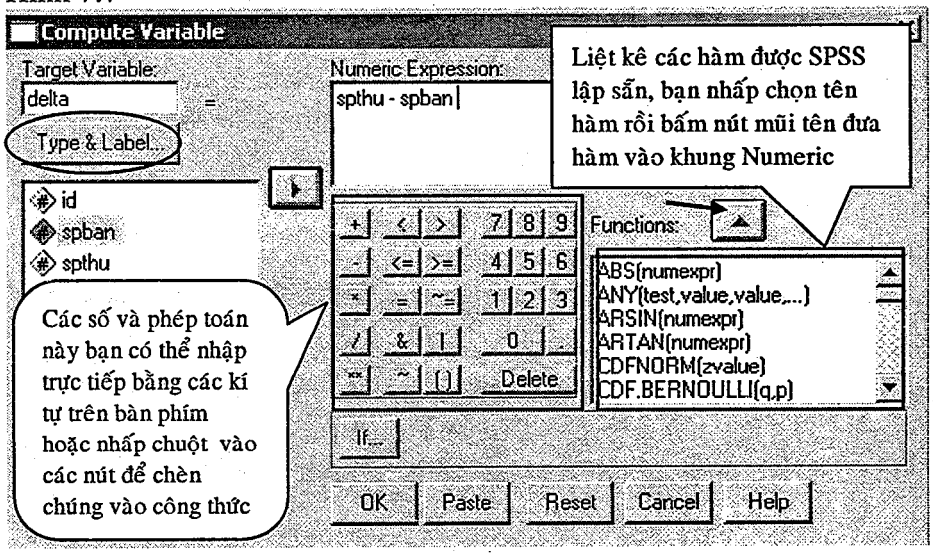
Bạn hãy mở lại file *Dauphong*, chọn menu Transform > Compute... để mở hộp thoại Compute Variable (Hình 7.7). Bạn khai báo tên biến cần tính giá trị và công thức tính giá trị của biến theo trình tự như sau:

1.Nhập tên biến mới vào khung Target Variable, ví dụ delta, nhớ là ở đây bạn cũng phải tuân thủ các quy tắc về cách đặt tên biến.

2. Thành lập công thức theo các bước:

- Chọn sáng tên *spthu* rồi nhấp nút mũi tên đưa nó sang khung Numeric Expression
- Nhấp nút mang dấu – ở khu vực các số và phép toán để chèn phép tính trừ.
- Chọn sáng tên *spban* rồi nhấp nút mũi tên đưa nó sang khung Numeric Expression

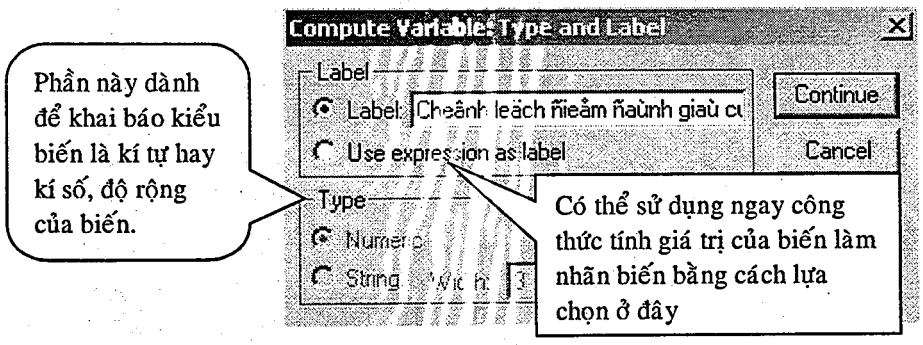
Hình 7.7



3. Bạn có thể đặt nhãn cho biến mới này bằng cách nhấp nút Type&Label... bạn khai báo nhãn bằng tiếng Việt ở chế độ gõ tiếng Việt trong khung Label, sau đó nhãn tiếng Việt của biến sẽ được thể hiện ở cửa sổ Variable View của file.

4. Nhấn OK, biến mới *delta* sẽ xuất hiện ở cuối danh sách biến ở cửa sổ Variable View, ở cửa sổ Data View nó sẽ nhận những giá trị y hệt giá trị của △ ở Bảng 7.1

Hình 7.8

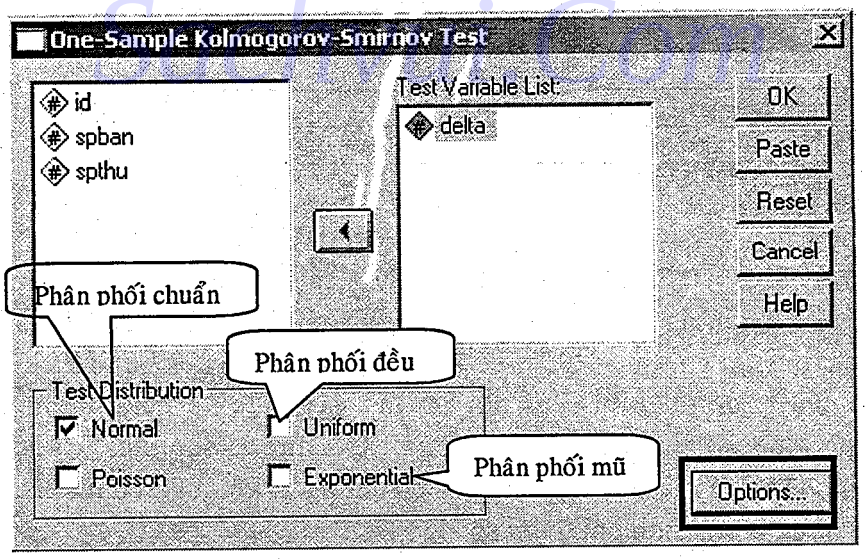


Sau khi tạo được biến *delta* chúng ta trở lại với mục tiêu chính của phần này là thực hiện kiểm định One-sample Kolmogorov-Smirnov cho giả thuyết Ho: Tổng thể của biến *delta* có phân phối chuẩn

Trình tự kiểm định One-sample Kolmogorov-Smirnov như sau

1. Bạn vào menu 1-Sample K-S mở hộp thoại One-Sample Kolmogorov-Smirnov Test

Hình 7.9



2. Đưa biến cần kiểm định (*delta*) vào khung Test Variable List

3. Nhấp chọn tên phân phối mà giả thuyết của bạn muốn kiểm định (ở đây là Normal) trong khung Test Distribution.

4. Nếu cần thì bạn vào nút Option... yêu cầu tính các đại lượng thống kê mô tả và cách xử lý các quan sát bị thiếu.

5. Trở lại hộp thoại chính, nhấp OK

Sau đây là kết quả kiểm định một mẫu Kolmogorov-Smirnov cho biến *delta*

Bảng 7.19 **Descriptive Statistics**

	N	Mean	Std. Deviation	Minimum	Maximum
Chênh lệch điểm đánh giá của người td	20	.6000	1.18766	-2.00	2.00

Bảng 7.20 **One-Sample Kolmogorov-Smirnov Test**

		Chênh lệch điểm đánh giá của người td
N		20
Normal Parameters(a,b)	Mean	.6000
	Std. Deviation	1.18766
Most Extreme Differences	Absolute	.232
	Positive	.119
	Negative	-.232
Kolmogorov-Smirnov Z		1.037
Asymp. Sig. (2-tailed)		.232

a Test distribution is Normal.
b Calculated from data.

Giá trị Sig. = 0,23 ở Bảng 7.20 lớn hơn cả mức ý nghĩa 0,1 nên giả thuyết Ho không thể bị bác bỏ ngay cả với độ tin cậy chỉ 90%, do đó ta chấp nhận giả thuyết Ho. Có thể kết luận rằng các chênh lệch tổng thể của ta có phân phối chuẩn và như vậy việc áp dụng kiểm định trung bình 2 hai mẫu phối hợp từng cặp Paired –samples T test cho tình huống này là phù hợp, nó sẽ cho chúng ta kết luận chính xác hơn khi dùng kiểm định phi tham số sign test.

CHƯƠNG VIII
KIỂM ĐỊNH TỶ LỆ

Thủ tục kiểm định tỷ lệ tổng thể là một thủ tục đơn giản tuy nhiên việc tiến hành kiểm định này trên SPSS cần chú ý:

Biến đưa vào kiểm định cần phải là một biến nhị phân (chẳng hạn giới tính chỉ có hai tình huống lựa chọn), còn nếu là biến trên 2 lựa chọn, ví dụ ngành học trong một trường ĐH hay tình trạng tôn giáo... thì cần phải khai báo điểm cắt đối với biến đó để chia dữ liệu thành hai nhóm (quy tắc là các quan sát có giá trị nhỏ hơn hoặc bằng giá trị điểm cắt thì xếp vào nhóm 1, còn lại là nhóm 2, và kiểm định được mặc định tiến hành liên quan đến thông tin của nhóm 1), cách tốt nhất khi muốn tiến hành kiểm định tỷ lệ tổng thể trên SPSS là nên dùng các thủ tục phụ như Recode chẳng hạn để chuyển biến nhiều lựa chọn thành biến chỉ có 2 lựa chọn (ví dụ một người theo hay không theo tôn giáo mà ta đang muốn tiến hành kiểm định tỷ lệ số người theo tôn giáo đó...)

Kiểm định này không đòi hỏi giả định về dạng phân phối của biến đang xét xong số liệu phải được chọn mẫu ngẫu nhiên. Để đạt được điều kiện này chúng tôi sẽ cung cấp cho bạn 1 phần bộ số liệu trong nghiên cứu "Giá trị cảm nhận về dịch vụ đào tạo của sinh viên Khoa Kinh tế - ĐH Nha Trang" của Chu Nguyễn Mộng Ngọc, bộ dữ liệu gồm thông tin thu thập trên 490 sinh viên, bộ số liệu này được chúng ta giả định như một tổng thể được đặt tên là file *kiem dinh ty le_goc*

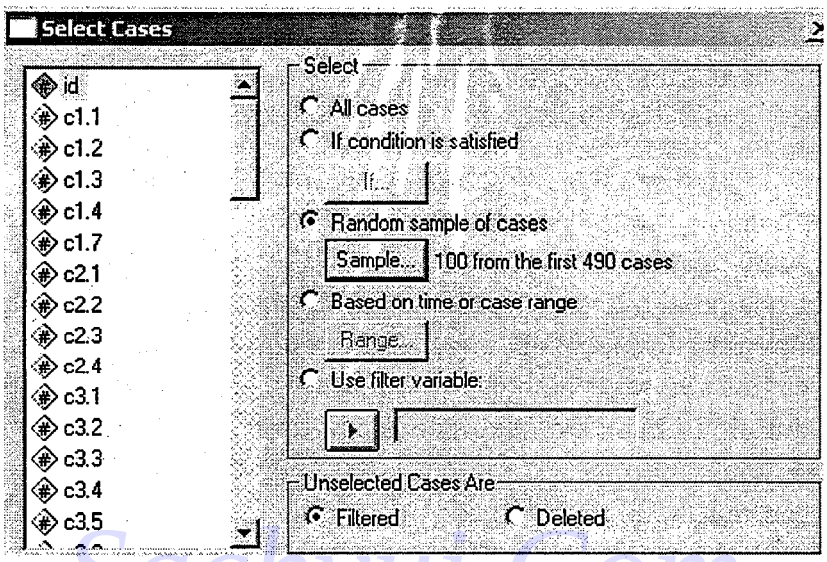
Bảng 8.1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nam	180	36,7	36,7	36,7
	Nu	310	63,3	63,3	100,0
	Total	490	100,0	100,0	

Chúng ta dùng lệnh Select case với lựa chọn Random Sample of cases để chọn một mẫu (xem như là ngẫu nhiên) 100 sinh viên từ tổng thể này rồi lưu thành file Kiem dinh ty le_mau (chú ý là khi bạn dùng lại bộ dữ liệu kiem dinh ty le_goc để thực hành, bạn sẽ chọn ra

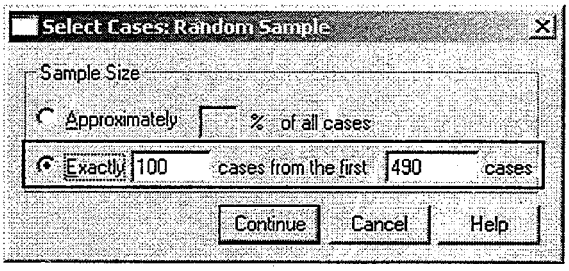
một mẫu với cấu trúc khác của chúng tôi vì tính ngẫu nhiên của việc chọn mẫu).

Hình 8.1



Cách tiến hành là bạn bấm vào mục Random sample of cases rồi bấm vào nút Sample...để tiến hành các khai báo cần thiết trong hộp thoại sau:

Hình 8.2



Trên bộ số liệu mẫu đại diện chúng ta xác định được cấu trúc mẫu về giới tính như sau (nam được mã hóa là 1 và nữ là 2)

Bảng 8.2 **Gioi tinh cua sinh vien tra loi phong van**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	**Nam**	**39**	39,0	39,0	39,0
	Nu	61	61,0	61,0	100,0
	Total	**100**	100,0	100,0	

Để kiểm định cặp giả thuyết sau ta lần lượt thực hiện các bước :

H_0: tỷ lệ sinh viên nam trong tổng thể ≥ 40%

H_1: tỷ lệ sinh viên nam trong tổng thể < 40%

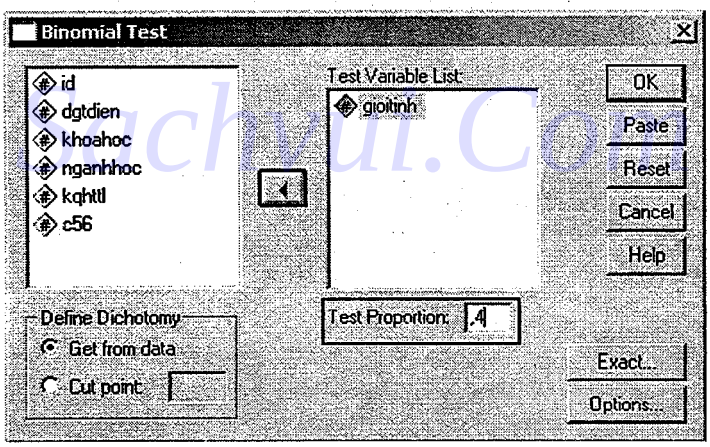
(dựa vào tỉ lệ mẫu mà đặt H_1, chú ý là bạn xem như không biết gì về cấu trúc tổng thể giới tính của file Kiem dinh ty le_goc),

Vào Menu Analyze/Noparametric Test/Binomial mở cửa sổ Binomial test

- Đưa biến gioitinh sang khung Test Variable List
- Giá trị về tỷ lệ tổng thể muốn kiểm định trong khung Test Proportion (mặc định là 0,5) sửa thành 0,4.

Nếu biến của chúng ta là nhị phân thì mặc định để lựa chọn Get from Data với giới tính nam được chọn là nhóm 1 ở dữ liệu này (bảng kết quả kiểm định sẽ chỉ cho bạn thấy điều đó khi các kết quả kiểm định chạy ra nằm ngang hàng với nhóm nam).

Hình 8.3



Bảng 8.3 **Binomial Test**

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Gioi tinh cua sinh vien tra loi phong van	Group 1	Nam	39	,4	,4	,462(a,b)
	Group 2	Nu	61	,6		
	Total		100	1,0		

a Alternative hypothesis states that the proportion of cases in the first group < ,4.
b Based on Z Approximation.

Giá trị sig = 0,462 cho ta biết không thể bác bỏ giả thuyết H_0 nghĩa là không thể nói tỷ lệ nam giới của tổng thể này dưới 40%.

Xét một khía cạnh khác, là tình huống nếu biến định tính của ta có nhiều phân loại thì. Cũng trên tập dữ liệu này kết quả chạy bảng tần số trên tập dữ liệu được xem là tổng thể cho biết tỷ lệ sinh viên học chuyên ngành Kinh tế Thủy sản là 25,7%. Còn kết quả thống kê mô tả trên dữ liệu mẫu như sau

Bảng 8.4 **Nganh hoc**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	**Kinh te Thuy san**	**26**	**26,0**	26,0	26,0
	Ke toan Doanh nghiep	26	26,0	26,0	52,0
	Quan tri Kinh doanh	28	28,0	28,0	80,0
	Thuong mai	20	20,0	20,0	100,0
	Total	100	100,0	100,0	

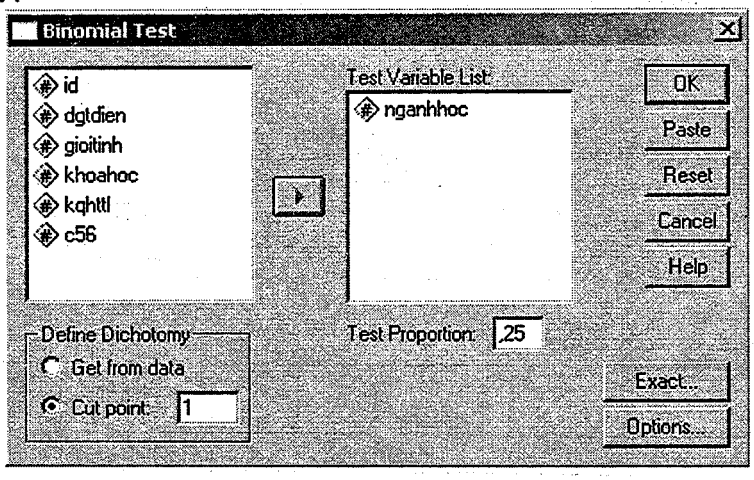
Ta muốn kiểm định giả thuyết

H_0: tỷ lệ sinh viên học ngành KTTS trong tổng thể ≤ 25%

H_1: tỷ lệ sinh viên học ngành KTTS trong tổng thể > 25%

Ta chạy lại lệnh Binomial test nhưng các lựa chọn thực hiện như sau:

Hình 8.4



Chý ý là biến Nganhhoc được mã hóa tới 4 lựa chọn trong đó KTTS mang giá trị 1 nên bạn phải nhập giá trị 1 vào ô Cut point để dữ liệu được cắt thành 2 phần là sinh viên theo hoặc không theo chuyên ngành KTTS.

Bảng 8.5 **Binomial Test**

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Nganh hoc	Group 1	<= 1	26	,26	,25	,447(a)
	Group 2	> 1	74	,74		
	Total		100	1,00		

a Based on Z Approximation.

Các thông tin nằm ngang hàng nhóm mang giá trị <=1 tức là nhóm KTTS. Chúng ta thấy rằng không thể bác bỏ H_0 vì giá trị Sig quá lớn, như vậy xem như tỷ lệ tổng thể là một giá trị gần 0,25

Tuy nhiên cũng với chính thông tin kiểm định này nhưng nếu bạn đặt lại giả thuyết là:

H_0: tỷ lệ sinh viên học ngành KTTS trong tổng thể ≤ 10%

H_1: tỷ lệ sinh viên học ngành KTTS trong tổng thể > 10%

Giả thuyết đặt lại thì khi chạy kiểm định trên SPSS bạn phải thay đổi giá trị nhập vào khung Test Proportion.

Kết quả như sau:

Bảng 8.6 **Binomial Test**

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Nganh hoc	Group 1	<= 1	26	,3	,1	,000(a)
	Group 2	> 1	74	,7		
	Total		100	1,0		

a Based on Z Approximation.

Bạn đi đến kết luận bác bỏ H_0 vì rõ ràng 10% là quá khác xa tỷ lệ thật của tổng thể.

CHƯƠNG IX

TƯƠNG QUAN VÀ HỒI QUI TUYẾN TÍNH

Chương IV đã đề cập đến việc đo lường mối liên hệ giữa hai biến định tính, nhưng trong thực tế, chúng ta thường gặp nhiều tình huống phải đo lường ảnh hưởng hay liên hệ giữa một biến định lượng này với một biến định lượng khác. Trong chương này chúng ta sẽ xem xét mối liên hệ giữa hai hay nhiều biến định lượng, sử dụng hai phương pháp tương quan và hồi qui.

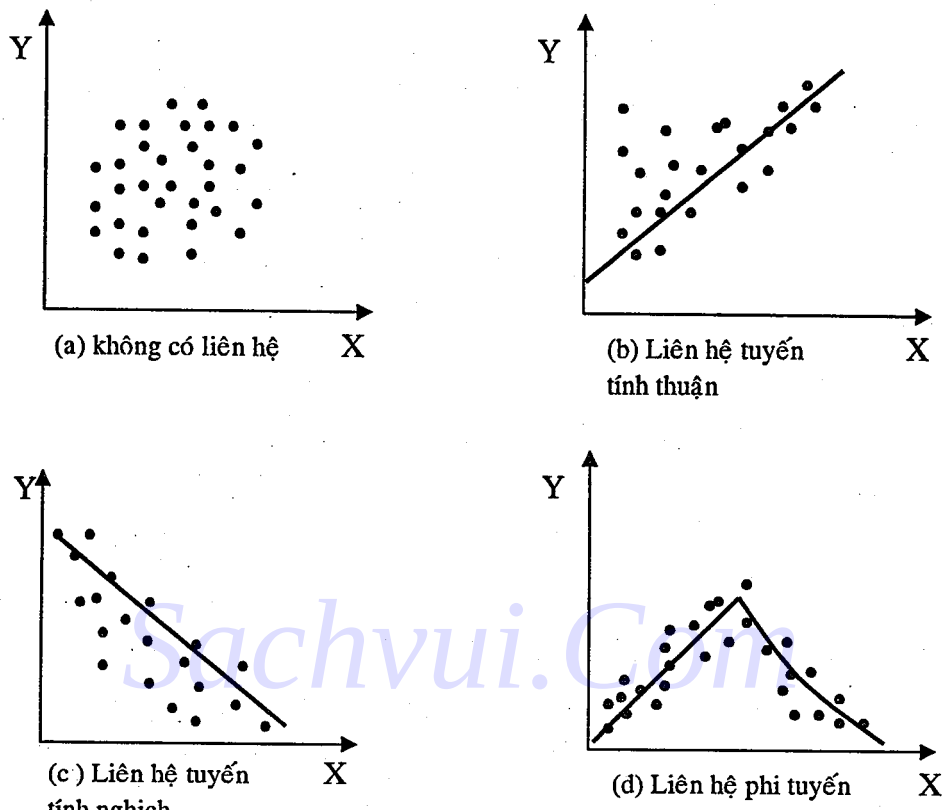
1. TƯƠNG QUAN TUYẾN TÍNH

Ở phần này chúng ta sẽ xem xét cách đo lường mối quan hệ giữa 2 biến định lượng qua ví dụ về mối quan hệ giữa doanh số bán hàng (y) với chi phí chào hàng (x). Công ty chúng ta nghiên cứu có 24 khu vực bán hàng trong toàn quốc. Các số liệu về doanh số bán, chi phí quảng cáo và chi phí chào hàng của công ty trong năm 1996 tại 12 khu vực bán hàng của công ty được thu thập trong file *tqvahq*

Một đồ thị phân tán là công cụ hữu ích có thể cho chúng ta thấy nhiều loại liên hệ giữa hai biến ta đang khảo sát. Một số dạng liên hệ thường gặp giữa hai biến định lượng được biểu diễn ở Hình 9.1. Như bạn thấy có 4 dạng liên hệ giữa 2 biến:

- Trong Hình 9.1a các chấm đại diện cho các cặp giá trị thực tế quan sát được (X;Y) phân tán ngẫu nhiên, và không có mối liên hệ giữa hai biến này.
- Trong Hình 9.1b thì mối liên hệ đó gần như là tuyến tính và thuận chiều.
- Đồ thị phân tán thứ ba thể hiện mối liên hệ tuyến tính và có chiều nghịch
- Cuối cùng ở Hình 9.1d mối liên hệ giữa 2 biến rõ ràng rất mạnh vì những chấm không nằm phân tán ngẫu nhiên nhưng nó cũng chỉ rõ rằng mối liên hệ đó là phi tuyến, vì đường thẳng không đi theo một hướng duy nhất.

Hình 9.1 Bốn biểu đồ thể hiện dạng chung của liên hệ giữa 2 biến

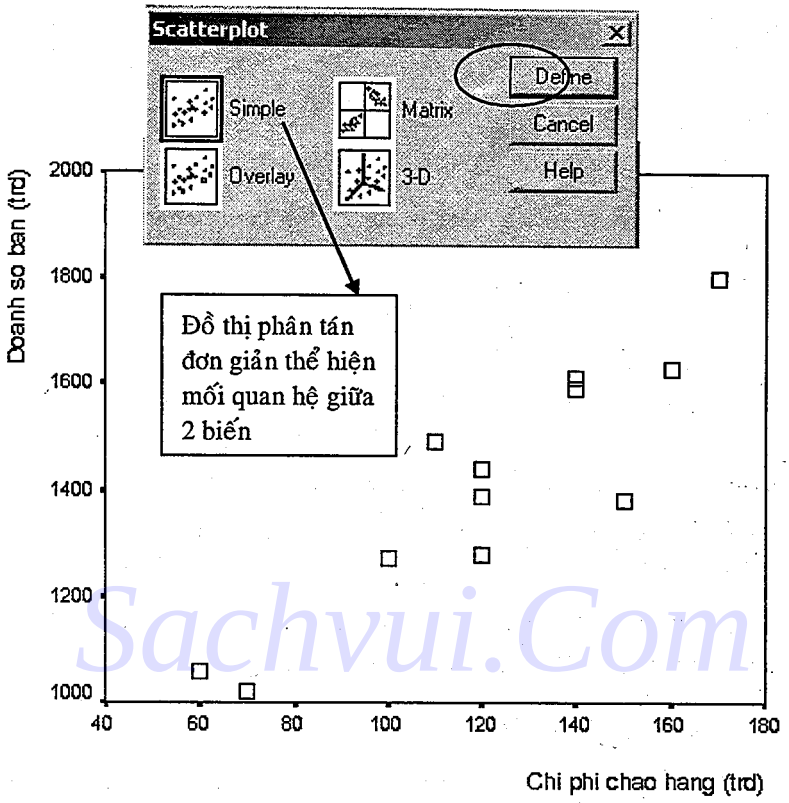


Với ý tưởng này, chúng ta dùng đồ thị Scatter của SPSS biểu diễn các số liệu về doanh số bán và chi phí chào hàng tại12 khu vực bán hàng của công ty (Hình 9.2). Từ đồ thị này ta có thể thấy giữa doanh số và chi phí chào hàng có mối liên hệ thuận. Tức là chi phí chào hàng càng tăng thì doanh số bán càng cao. Ta hình dung mối liên hệ này có thể được biểu diễn bằng một đường thẳng vì các điểm quan sát (các chấm vuông) ít hay nhiều đều có vẻ tập trung theo một đường thẳng vô hình. Theo trực quan có thể kết luận mối liên hệ giữa doanh số và chi phí chào hàng là tuyến tính thuận, tuyến tính chỉ dạng đường thẳng còn thuận chỉ sự tăng giảm cùng chiều của doanh số và chi phí chào hàng.

Để vẽ được đồ thị Scatter bạn chọn menu Graphs > Scatter để mở hộp thoại Scatterplot (xem hình), nhấp chọn Simple trong hộp thoại

này rồi nhấp nút Define để vào hộp thoại kế tiếp là Simple Scatterplot, thực hiện theo trình tự thông thường như đã hướng dẫn ở phần vẽ đồ thị thuộc Chương III bạn sẽ có đồ thị phân tán.

Hình 9.2



1.1 Hệ số tương quan đơn r (Pearson Correlation Coefficient)

1.1.1 Tính toán r

Người ta sử dụng một số thống kê có tên là hệ số tương quan Pearson (ký hiệu là r) để lượng hóa mức độ chặt chẽ của mối liên hệ tuyến tính giữa 2 biến định lượng. Nhìn chung r được sử dụng để kiểm tra liên hệ giữa những biến định lượng (khoảng cách hay tỷ lệ). Công thức của r như sau:

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{(N-1)S_X S_Y} \qquad (-1 \le r \le +1)$$

Trong đó N là số quan sát và Sx và Sy là độ lệch chuẩn của từng biến X và Y.

Trị tuyệt đối của r cho biết mức độ chặt chẽ của mối liên hệ tuyến tính. Giá trị tuyệt đối của r tiến gần đến 1 khi hai biến có mối tương quan tuyến tính chặt chẽ (Khi tất cả các điểm phân tán xếp thành một đường thẳng thì trị tuyệt đối của r =1).

Khi đường thẳng dốc lên như Hình 9.1b thì r mang giá trị dương và khi đường thẳng dốc xuống như Hình 9.1c thì r mang giá trị âm

Giá trị r = 0 chỉ ra rằng hai biến không có mối liên hệ tuyến tính. "Không có mối liên hệ tuyến tính" cần được phân biệt 2 tình huống :

- Không có mối quan hệ giữa 2 biến như Hình 9.1a.
- Hai biến có thể có liên hệ chặt chẽ với nhau nhưng hệ số tương quan vẫn sẽ nhỏ gần bằng 0 nếu như dạng của mối liên hệ này không phải là tuyến tính (ta gọi là phi tuyến) như Hình 9.1d. Dạng liên hệ này yêu cầu một số thống kê đặc biệt gọi là eta (η) để kiểm định mối quan hệ (phi tuyến) Vấn đề này sẽ không được thảo luận trong phạm vi quyển sách này.

Xem tiếp phần hướng dẫn cách thực hiện trên SPSS bạn sẽ biết phải làm như thế nào để tính được r.

1.1.2 Một số đặc điểm của r

- Giá trị của r cho biết không có mối liên hệ tuyến tính giữa 2 biến chưa hẳn có nghĩa là 2 biến đó không có mối liên hệ. Do đó hệ số tương quan tuyến tính chỉ nên được sử dụng để biểu thị mức độ chặt chẽ của liên hệ tương quan tuyến tính.
- Ngoài ra cần phải cẩn thận xem xét đồng thời hệ số tương quan và cả đồ thị bởi vì hệ số tương quan có thể có cùng một giá trị trong khi hình dạng của mối quan hệ lại rất khác nhau.
- Một lỗi thông thường khi giải thích hệ số tương quan tuyến tính là cứ cho rằng có liên hệ tương quan có nghĩa là lúc nào cũng có mối liên hệ nhân quả. Doanh số có thể tương quan chặt chẽ với số lượng nhân viên chào hàng trong khu vực bán hàng. Nhưng tăng chào hàng không phải lúc nào cũng tăng cao được doanh số, sự gia tăng doanh số sau khi thực hiện chiến lược đẩy mạnh chào

hàng của bạn có thể là do tác động của chiến dịch chào hàng, nhưng cũng có thể do các yếu tố khác mà bạn quên chú ý đến như mùa vụ, tác động của chiến dịch quảng cáo... chứ có thể việc đẩy mạnh chào hàng cuả bạn chẳng có tác động gì đến doanh số, hoặc mức độ tác động không mạnh như hệ số r thể hiện.

- Bạn cần cảnh giác với những mối quan hệ gọi là tương quan giả. 2 biến định lượng có thể có hệ số tương quan rất r cao nhưng thực tế lại chẳng có quan hệ gì, giá trị r cao tính được chỉ do một sự ngẫu nhiên trong mẫu và là một sản phẩm chế tác của kỹ thuật thống kê đang được bạn sử dụng. Giá trị r cao đó có thể hoá ra là vô nghĩa khi bạn kiểm định độ phù hợp tổng thể của nó. Ví dụ khi bạn tìm ra được mối tương quan giữa số lượng trường học và số lượng quán bia trong thành phố, bạn có thể ngờ rằng đây chỉ là kết quả do tình cờ, hoặc cả 2 hiện tượng này đều là kết quả của việc dân số thành phố đang tăng rất nhanh.
- Hệ số tương quan là một thước đo mang tính đối xứng, bởi vì nếu ta thay đổi vai trò của hai biến X và Y cho nhau trong công thức thì kết quả vẫn không thay đổi. Hệ số tương quan tuyến tính không có đơn vị đo lường, và nó không bị ảnh hưởng bởi những phép biến đổi tuyến tính như cộng trừ nhân hoặc chia tất cả các giá trị của một biến bởi một hằng số.

1.1.3 Kiểm định giả thuyết về hệ số tương quan tuyến tính r

Cũng như với Gamma, ta có thể kiểm tra hệ số r có giá trị cao tính được ở trong mẫu có phản ánh một hiệp biến thiên thật sự trong tổng thể không hay chỉ do tình cờ. Trong trường hợp kiểm định giả thuyết về hệ số tương quan cuả tổng thể (kí hiệu ρ), giả thuyết không là hệ số tương quan thật trong tổng thể ta quan tâm bằng 0. Nói cách khác, không có mối liên hệ nào giữa hai biến. Ho: $\rho = 0$

Để kiểm định giả thiết này, chúng ta cần một số giả định về phân phối chung của cả hai biến. Giả định thông thường là các mẫu ngẫu nhiên độc lập được lấy ra từ một tổng thể trong đó cả hai biến đều có phân phối chuẩn. Nếu điều kiện này thỏa mãn thì kiểm định giả thiết hệ số tương quan của tổng thể bằng 0 được dựa vào thống kê sau:

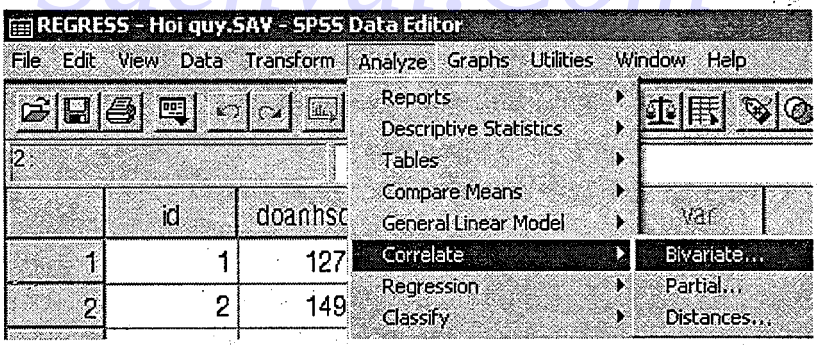
$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

Như trên đã nói, N là số quan sát. Nếu ρ = 0, thì ta có phân phối t-Student với N-2 bậc tự do. Chúng ta có thể sử dụng kiểm định một hay hai phía. Nếu không biết trước gì về chiều hướng của mối liên hệ (thuận hay nghịch) ta nên sử dụng kiểm định hai phía. Tức là giả thiết hệ số tương quan bằng 0 bị bác bỏ đối với cả hai giá trị dương quá lớn hoặc âm quá nhỏ của t. Nếu chiều hướng của mối liên hệ có thể xác định trước được thì giả thiết chỉ bị bác bỏ khi giá trị t đủ lớn theo hướng đã xác định.

Cách thực hiện tính r bằng SPSS

Để tính được r của mẫu và thực hiện kiểm định giả thuyết về hệ số tương quan tuyến tính của tổng thể chúng ta sử dụng lệnh Correlate > Bivariate của menu Analyze. Xem Hình 9.3 ở dưới để biết cách vào lệnh này

Hình 9.3



Sau khi bạn chọn Bivariate... hộp thoại Bivariate Correlations (hộp thoại tương quan hai biến) xuất hiện như Hình 9.4. Thủ tục tương quan hai biến của SPSS sẽ tính toán hệ số tương quan hạng Pearson, hệ số rho-Spearman và tau-*b* Kendall tùy theo lựa chọn của bạn với các mức ý nghĩa tương ứng. Trình tự thao tác trong hộp thoại này như sau:

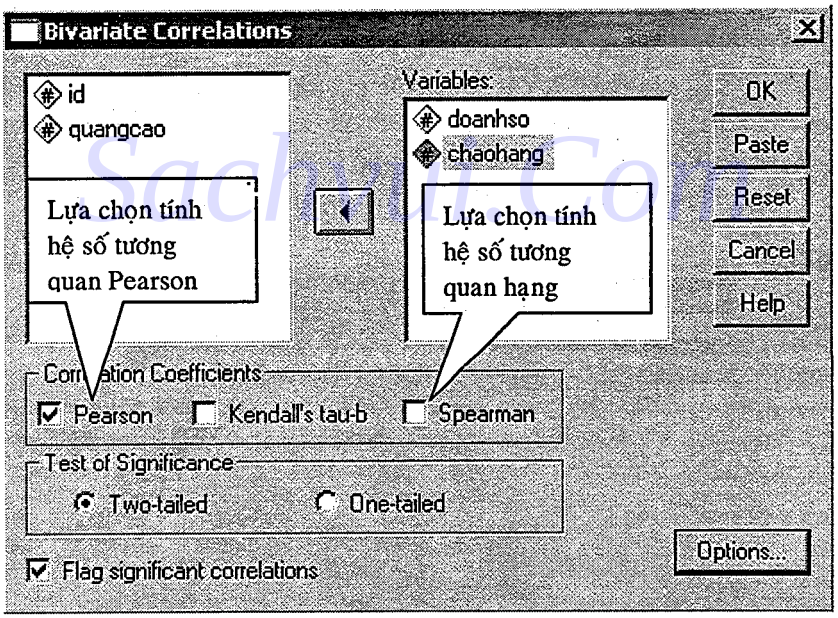
1. Các biến trong file dữ liệu của bạn sẽ xuất hiện trong khung danh sách biến nguồn. Bạn có thể chọn hai hay nhiều hơn hai biến để đưa vào khung Variables yêu cầu phân tích. Nếu tất cả các trường hợp quan sát đều có giá trị bị thiếu đối với một hay cả hai biến được chọn, hoặc nếu tất cả các trường hợp đều có cùng một giá trị trên một biến (không

có biến thiên), thì hệ số tương quan không thể tính được và được SPSS thể hiện bằng một dấu chấm (.).

2. Trong phần Correlation Coefficients (hệ số tương quan) bạn lựa chọn ít nhất là một loại hệ số trong các loại hệ số sau:

- Pearson: hệ số này là được chọn mặc định. Bảng kết quả sẽ thể hiện một ma trận vuông gồm các hệ số tương quan. Tương quan của một biến nào đó với chính nó sẽ có hệ số tương quan là 1 và bạn có thể thấy chúng trên đường chéo của ma trận. Mỗi biến sẽ xuất hiện hai lần trong ma trận với hệ số tương quan y hệt nhau trong hai tam giác trên và dưới đối xứng nhau qua đường chéo của ma trận.
- Kendall's tau-b: cũng là một loại hệ số tương quan hạng, chúng ta đã xem ở Chương IV.
- Spearman: Spearman rho là một loại hệ số tương quan hạng sẽ được nghiên cứu ở phần 1.2

Hình 9.4



3. Trong phần Test of significance (kiểm định mức ý nghĩa). Bạn có thể chọn trong hai loại kiểm định sau:

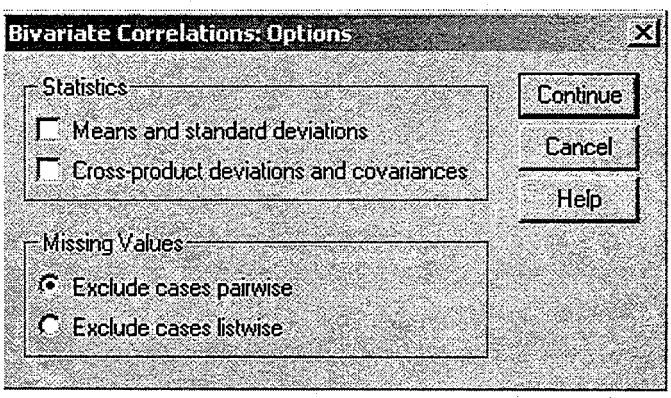
- Two-tail (kiểm định hai phía). Loại kiểm định này được sử dụng trong trường hợp chiều hướng của mối liên hệ tuyến tính không thể xác định trước được. Đây là loại kiểm định mặc định.

- One-tail (kiểm định một phía). Loại kiểm định này được sử dụng khi bạn có thể xác định trước chiều hướng của mối liên hệ giữa hai biến trong phần phân tích.

4. Để thu được các thông số thống kê tùy chọn thêm trong tương quan Pearson hay giải quyết các trường hợp quan sát thiếu số liệu, bạn hãy nhấp chuột vào nút Options... trong hộp thoại Bivariate Correlations. Lệnh này sẽ mở ra một hộp thoại tùy chọn trong tương quan hai biến như trong Hình 9.5. các lựa chọn trong hộp thoại này bao gồm:

- Statistics (các thông số thống kê): phần này sẽ không hiện sáng để chọn nếu ở hộp thoại trước bạn chọn hệ số Spearman, bạn có thể chọn thêm một hay cả hai loại thông số sau:
  - Means and standard deviations (trung bình và độ lệch chuẩn) cho biết giá trị trung bình và độ lệch chuẩn của từng biến. Ngoài ra SPSS còn cho biết số lượng quan sát có đầy đủ số liệu tham gia vào quá trình tính toán. Các giá trị bị thiếu được giải quyết trên từng biến một.
  - Cross-product deviations and covariances (tổng các tích mômen chéo và hiệp phương sai): cho biết tổng các tích mômen chéo và hiệp phương sai của từng cặp biến. Tổng các tích mômen chéo bằng tổng các tích của từng cặp độ lệch giữa giá trị quan sát với giá trị trung bình của hai biến. Đây chính là giá trị của tử số của hệ số tương quan đơn tuyến tính Pearson. Hiệp phương sai là một thước đo mối quan hệ giữa hai biến chưa chuẩn hóa, nó bằng tổng các tích mômen chéo chia cho N-1.

Hình 9.5



- Missing values (các số liệu bị thiếu, sót): bạn có thể chọn một trong hai cách giải quyết sau:
  - Exclude cases pairwise: các trường hợp quan sát bị thiếu mất giá trị ở một hay cả hai biến của cặp biến đang dùng để tính hệ số tương quan giữa 2 biến đó sẽ bị loại ra khỏi quá trình tính toán phân tích. Còn với

tình huống tính toán hệ số tương quan giữa cặp biến khác mà các quan sát đó không thiếu giá trị thì chúng vẫn được sử dụng. Đây là lựa chọn mặc định. Bởi vì mỗi hệ số tương quan được tính dựa trên tất cả các quan sát có đầy đủ các giá trị trên cặp biến đang tính cho nên các thông tin có sẵn được sử dụng tối đa trong các tính toán. Lựa chọn này có thể đưa ra một tập hợp các hệ số tương quan được tính toán dựa trên số lượng quan sát khác nhau.

- Exclude cases listwise: các trường hợp quan sát bị thiếu mất giá trị ở bất kỳ biến nào cũng sẽ bị loại ra khỏi toàn bộ các phân tích tính toán (và như vậy tất cả các hệ số tương quan đều tính toán từ cùng một số lượng các quan sát)

5. Nhấp Continue để trở về hộp thoại trước và nhấp OK.

Vì trong hộp thoại thủ tục tương quan 2 biến của ví dụ của chúng ta bạn chỉ để những lựa chọn mặc định nên bạn có 1 bảng kết quả là Bảng 9.1 thể hiện hệ số tương quan tuyến tính giữa doanh số bán và chi phí chào hàng. Số liệu trong các ô Pearson Correlation là các hệ số tương quan. Hệ số tương quan giữa chi phí chào hàng với chính nó là 1, giữa doanh số và chào hàng là 0,905. Giá trị này cho thấy rằng giữa doanh số bán và chi phí chào hàng có mối liên hệ thuận khá chặt chẽ (như đã được biểu diễn trong Hình7.1).

Bảng 9.1 **Correlations**

		Chi phi chao hang (trd)	Doanh so ban (trd)
Chi phi chao hang (trd)	Pearson Correlation	1	.905(**)
	Sig. (2-tailed)	.	.000
	N	12	12
Doanh so ban (trd)	Pearson Correlation	.905(**)	1
	Sig. (2-tailed)	.000	.
	N	12	12

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

Trong SPSS, bạn có thể kiểm định các giả thuyết ở mức ý nghĩa nhỏ hơn 0,05 (SPSS phân biệt bằng cách đánh một dấu sao * ở cạnh giá trị thống kê tính được trên mẫu) và ở mức ý nghĩa nhỏ hơn 0,01 (phân biệt bằng hai dấu sao **).

Từ Bảng 9.1 ta có thể thấy khả năng để hệ số tương quan tính được từ mẫu là 0,905 trong khi trên thực tế không có mối liên hệ tuyến tính nào trong tổng thể giữa doanh số và chi phí chào hàng là 0,000 nhỏ hơn 0,01. Như vậy nếu ta sử dụng mức ý nghĩa1% (tức là xác suất chấp nhận giả thuyết sai là 1%) thì giả thuyết hệ số tương quan của tổng thể bằng không bị bác bỏ.

1.2. Hệ số tương quan hạng (Rank correlation coefficient)

Hệ số tương quan Pearson chỉ phù hợp trong trường hợp các dữ liệu thu thập được ở dạng thang đo định lượng, giống như các số liệu về doanh số và chi phí trong ví dụ của chúng ta. Giả định cần để kiểm định giả thuyết về hệ số tương quan tuyến tính ρ là tổng thể có phân phối chuẩn. Đối với các dữ liệu không thỏa mãn được giả định về phân phối chuẩn này, thì ta có một thước đo liên hệ tuyến tính khác giữa hai biến, đó là hệ số tương quan hạng Spearman.

Hệ số tương quan hạng cũng giống hệ số tương quan Pearson nhưng được tính dựa vào các hạng của dữ liệu chứ không dựa vào giá trị thực của quan sát. Nếu dữ liệu nguyên thủy của mỗi biến không có các mức độ bằng nhau thì dữ liệu của từng biến trước hết được xếp hạng, và sau đó hệ số tương quan Pearson giữa các hạng của hai biến được tính toán. Giống như hệ số tương quan Pearson, hệ số tương quan hạng Spearman chạy từ -1 đến + 1 , trong đó -1 và +1 cho thấy mối liên hệ hoàn toàn tuyến tính giữa hạng của hai biến. Vì vậy phần giải thích cũng tương tự r nhưng bản chất liên hệ ở đây là liên hệ giữa các hạng, không phải liên hệ giữa các giá trị.

Bảng 9.2 cho thấy ma trận hệ số tương quan hạng giữa doanh số bán và chi phí chào hàng. Chúng ta có thể thấy rằng các hệ số tương quan này có dấu tương tự như hệ số tương quan Pearson trong Bảng 9.1 nhưng độ lớn của nó không bằng, ở Chương Kiểm định phi tham số chúng ta đã biết rằng các số thống kê dựa trên hạng của dữ liệu được sử dụng trong tình huống giả định về phân phối không được thoả mãn thường không mạnh như những số thống kê sử dụng trong tình huống thông thường khi giả định được thoả mãn, trường hợp này không là ngoại lệ.

Bảng 9.2 **Correlations**

			Doanh so ban (trd)	Chi phi chao hang (trd)
Spearman's rho	Doanh so ban (trd)	Correlation Coefficient	1.000	.822(**)
		Sig. (2-tailed)	.	.001
		N	12	12
	Chi phi chao hang (trd)	Correlation Coefficient	.822(**)	1.000
		Sig. (2-tailed)	.001	.
		N	12	12

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

Bạn xem cách thức tính toán hệ số tương quan hạng Spearman bằng SPSS ở phần hệ số tương quan Pearson.

2. HỒI QUI TUYẾN TÍNH

Nếu kết luận được là 2 biến có liên hệ tương quan tuyến tính chặt chẽ với nhau qua hệ số tương quan r, đồng thời giả định rằng chúng ta đã cân nhắc kỹ bản chất của mối liên hệ tiềm ẩn giữa 2 biến, và xem như đã xác định đúng hướng của một mối quan hệ nhân quả có thật giữa chúng (bởi vì có quan hệ tương quan tuyến tính chưa chắc đã có quan hệ nhân quả) thì ta có thể mô hình hoá mối quan hệ nhân quả của chúng bằng mô hình hồi qui tuyến tính trong đó một biến được gọi là biến phụ thuộc (hay biến được giải thích - Y) và biến kia là biến độc lập (hay biến giải thích -X). Mô hình này sẽ mô tả hình thức của mối liên hệ và qua đó giúp ta dự đoán được mức độ của biến phụ thuộc (với độ chính xác trong một phạm vi giới hạn) khi biết trước giá trị của biến độc lập.

Nhắc đến hệ số tương quan tuyến tính chúng ta phải khẳng định lại rằng trong phân tích hồi qui, các biến không có tính chất đối xứng như phân tích tương quan. Trong phân tích tương quan không có sự phân biệt giữa 2 biến, còn với phân tích hồi qui chúng ta ngầm giả định là X gây ra Y, bạn ước lượng biến Y trên cơ sở đã biết các biến X. Biến độc lập X thì bạn đã biết giá trị, còn biến phụ thuộc Y là một biến ngẫu nhiên, chúng ta thừa nhận còn có vô vàn nhân tố khác tác động đến nó ngoài biến độc lập X mà ta đề cập, do những tác động

không liệt kê hết được này mà ứng với mỗi giá trị của biến độc lập có thể có nhiều giá trị khác nhau của biến phụ thuộc.

Trong nội dung chương này, chúng ta sẽ thảo luận về mô hình hồi qui tuyến tính để xem xét quan hệ tuyến tính giữa X và Y tức dạng của mối quan hệ là đường thẳng. Nếu hình dáng của mối quan hệ là phi tuyến (parabol, hyperbol, mũ ...) thì chúng ta phải sử dụng một dạng hồi qui tuyến tính khác một chút có tên gọi là "Hồi qui tuyến tính với các quan hệ phi tuyến". Như vậy thì chúng ta lại phải làm sáng tỏ tiếp thuật ngữ tuyến tính trong cụm từ "Hồi qui tuyến tính" là tuyến tính theo các hệ số hồi qui (nếu chưa rõ hệ số hồi qui là gì, bạn sẽ làm quen ở phần kế tiếp) chứ thuật ngữ này không chỉ mối quan hệ tuyến tính trong các biến độc lập và phụ thuộc, dạng của mối quan hệ giữa các biến có thể là phi tuyến, có thể là tuyến tính, nhưng hình thức của các hệ số trong mô hình hồi qui tuyến tính luôn là tuyến tính.

Chúng ta tạm thời giới hạn nghiên cứu của chúng ta trong mô hình hồi qui tuyến tính mô tả mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc, mặc dù rõ ràng điều đó không hoàn toàn thực tế và thuyết phục lắm đối với đủ kiểu quan hệ đa dạng phong phú trong thế giới thực, và càng không thực tế và thuyết phục hơn khi chúng ta sẽ xuất phát từ mô hình hồi qui tuyến tính mô tả mối quan hệ tuyến tính giữa chỉ 1 biến độc lập và 1 biến phụ thuộc, bởi chúng ta đã thừa nhận có rất nhiều nhân tố khác tác động đến biến phụ thuộc ngoài biến độc lập X mà ta đề cập. Nhưng sự hiểu biết những vấn đề cơ bản của mô hình hồi qui tuyến tính đơn biến là nền tảng cho việc tìm hiểu những mô hình phức tạp hơn, cũng như việc chúng ta nghiên cứu về thị trường cạnh tranh hoàn hảo trong môn Kinh tế học, tuy phi thực tế nhưng nó cung cấp cho bạn những kiến thức cơ sở cho việc nghiên cứu tiếp các hình thức thị trường có tính chất phức tạp hơn.

Sau khi có được những kiến thức cơ sở, chúng ta sẽ bàn đến việc mô tả mối liên hệ tương quan tuyến tính giữa một biến phụ thuộc và nhiều biến độc lập (một tình huống nghiên cứu rất thật trong thực tế chúng ta hay gặp phải), tức là ta cố gắng khám phá càng nhiều càng tốt các nhân tố tác động đến biến phụ thuộc nhằm giúp ta dự đoán được tốt nhất mức độ của biến phụ thuộc. Rồi sau đó ta phát triển tiếp hồi qui tuyến tính với các quan hệ đường cong.

2.1. Hồi qui đơn tuyến tính

2.1.1 Xây dựng phương trình của mô hình hồi qui đơn tuyến tính từ dữ liệu mẫu

Trước khi xem xét mô hình thể hiện liên hệ tương quan tuyến tính giữa một biến phụ thuộc và nhiều biến độc lập. Chúng ta xem xét mối liên hệ tuyến tính giữa một biến phụ thuộc và một biến độc lập. Mô hình được xây dựng từ dữ liệu mẫu có dạng $\hat{Y}_i = B_0 + B_1 * X_i$

Trong đó

- Xi là trị quan sát thứ i của biến độc lập
- $\hat{Y}_i$ là giá trị dự đoán (hay giá trị lý thuyết) thứ i của biến phụ thuộc, dấu mũ đại diện cho giá trị dự đoán.
- B_0 và B_1 là hệ số hồi qui ta đã nhắc đến ở trên, phương pháp được dùng để xác định B_0 và B_1 là phương pháp bình phương nhỏ nhất thông thường (Ordinary least square - OLS). Tại sao OLS được sử dụng, bạn sẽ trả lời được qua xem xét ví dụ thực tế sau.

Để ví dụ chúng ta sẽ nhắc lại mối quan hệ giữa doanh số bán hàng và chi phí chào hàng đã nghiên cứu ở phần Hệ số tương quan r.

Đồ thị phân tán giữa 2 biến là một gợi ý cho chúng ta loại hàm số toán học thích hợp để mô tả và tóm tắt các dữ liệu quan sát. Chúng ta sẽ sử dụng loại đồ thị rải điểm Scatter để thể hiện mối quan hệ giữa 2 biến, nếu các điểm phân tán có xu hướng tạo thành một đường thẳng thì mô hình hồi qui đơn tuyến tính là một lựa chọn có nhiều khả năng (bạn xem Hình 9.2, các điểm phân tán được biểu diễn bằng các chấm vuông). Hệ số tương quan tuyến tính giữa 2 biến có giá trị cao cũng là một chỉ dẫn tốt cho tình huống này. Trong phần Hệ số tương quan chúng ta đã xác định là giữa doanh số bán hàng và chi phí chào hàng có mối quan hệ tương quan tuyến tính rất mạnh qua cả hệ số r cũng như đồ thị phân tán.

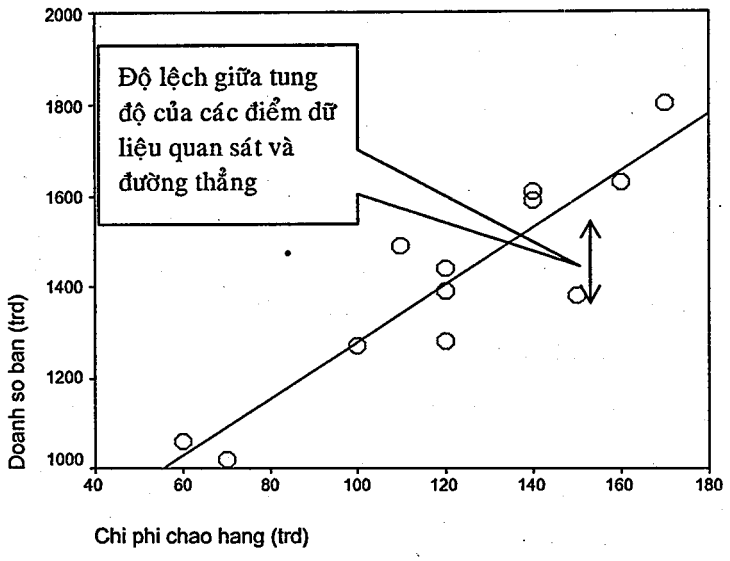
Từ gợi ý này chúng ta sẽ xây dựng một mô hình hồi qui tuyến tính đơn biến mô tả mối quan hệ giữa chúng trong đó doanh số là biến phụ thuộc và chi phí chào hàng là biến độc lập. Phương trình của đường thẳng có dạng:

$$\text{Doanh số dự đoán} = B_o + B_1 * \text{Chi phí chào hàng}$$

Trong phương trình này, độ dốc B_1 là lượng tăng giảm (triệu đồng) của doanh số điều chỉnh (còn gọi là doanh số dự đoán hay doanh số lý thuyết theo mô hình) do lượng tăng giảm của chi phí chào hàng. Hằng số B_o (tung độ của vị trí tại đó đường thẳng cắt trục tung) là giá trị doanh số lý thuyết khi chí phí chào hàng bằng 0 (tức là khi không tổ chức hoạt động chào hàng). Ở phần Kiểm định giả thuyết về hệ số hồi qui bạn sẽ thấy phát biểu như thế này về B_o là không hoàn toàn chính xác, nhưng ở đây chúng ta sẽ tạm thời chấp nhận điều này.

Tuy nhiên từ đồ thị rải điểm bạn cũng đã thấy tất cả các điểm dữ liệu quan sát không phải nằm hoàn toàn trên cùng một đường thẳng, chúng chỉ có vẻ tập trung xung quanh một đường thẳng mà thôi. Nên chúng ta có thể kẻ nhiều đường thẳng xuyên qua các điểm này chứ không phải chỉ một đường duy nhất, vấn đề là ta phải chọn ra đường thẳng nào mô tả sát nhất xu hướng này. Phương pháp bình phương nhỏ nhất OLS sẽ tìm ra được đường thẳng đó dựa trên nguyên tắc nó cực tiểu hóa tổng các độ lệch bình phương giữa tung độ của các điểm dữ liệu quan sát và đường thẳng. Hình 9.6 cho thấy đường thẳng tìm được bằng phương pháp OLS được kẻ ngay trên đồ thị phân tán. Chú ý là đồ thị phân tán này và đồ thị phân tán ở Hình 9.1 là một.

Hình 9.6 Đường hồi qui của doanh số và chi phí chào hàng



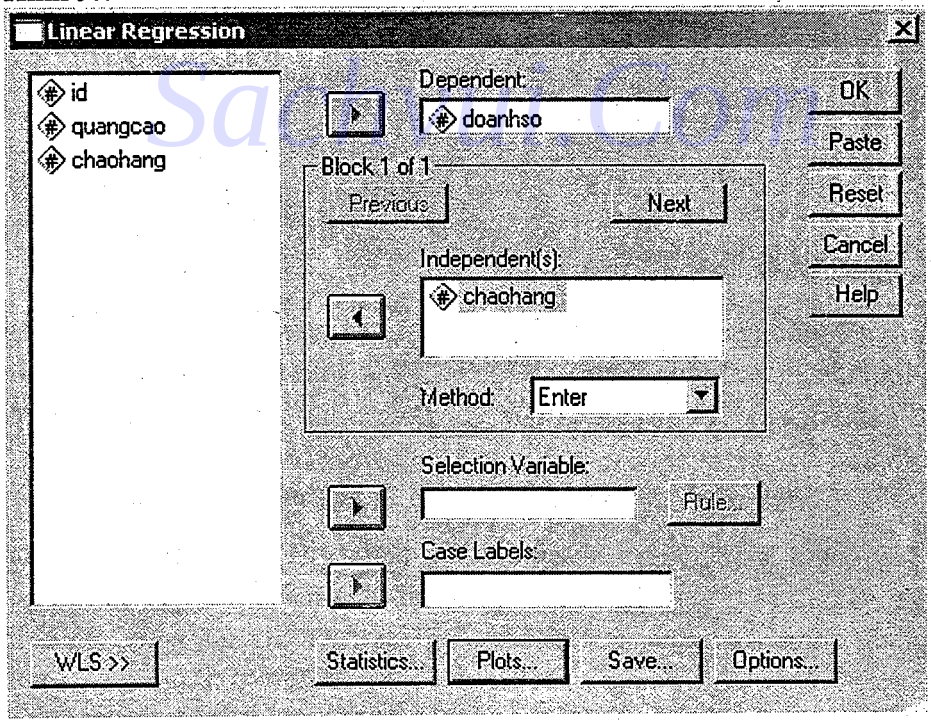
Cách thức xây dựng mô hình hồi qui đơn tuyến tính bằng SPSS

Bạn có thể sử dụng SPSS để tính toán ra đường thẳng này bằng cách chọn menu Analyze > Regression > Linear... bạn mở được hộp thoại Linear Regression. Đây là hộp thoại chính để bạn thực hiện các vấn đề có liên quan đến phương trình hồi qui tuyến tính, bạn sẽ thấy có rất nhiều lựa chọn trong các hộp thoại con mà chúng ta sẽ lần lượt làm sáng tỏ ở các phần kế tiếp (và được hệ thống lại cuối chương này) còn bây giờ bạn chỉ thực hiện những thao tác cơ bản nhất là

1. Đưa biến phụ thuộc (doanhso) vào khung Dependent
2. Đưa biến độc lập (chaohang) vào Independent(s)
3. Nhấp OK tức là bạn chấp nhận những lựa chọn mặc định của SPSS

Bạn xem các thao tác lựa chọn cơ bản nói trên được thể hiện ở Hình 9.7. Giờ hãy xem những lựa chọn mặc định của SPSS đối với hộp thoại Liner Regression sẽ cho bạn những thông tin gì trong các bảng từ 7.3 đến 7.6.

Hình 9.7



Bảng 9.3 **Variables Entered/Removed(b)**

Model	Variables Entered	Variables Removed	Method
1	Chi phi chao hang (trd)(a)	.	Enter

a All requested variables entered.

b Dependent Variable: Doanh so ban (trd)

Bảng 9.4 **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.905(a)	.818	.800	103.741

a Predictors: (Constant), Chi phi chao hang (trd)

Bảng 9.5 **ANOVA(b)**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	484845.373	1	484845.373	45.051	.000(a)
	Residual	107621.294	10	10762.129		
	Total	592466.667	11			

a Predictors: (Constant), Chi phi chao hang (trd)

b Dependent Variable: Doanh so ban (trd)

Bảng 9.6 **Coefficients(a)**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	651.523	117.384		5.550	.000
	Chi phi chao hang (trd)	6.261	.933	.905	6.712	.000

a Dependent Variable: Doanh so ban (trd)

Bảng 9.6 cung cấp cho chúng ta thông tin về hệ số hồi qui mà phương pháp OLS ước lượng được, độ dốc và hằng số được thể hiện trong cột B của bảng kết quả. Ta viết được phương trình hồi qui tuyến tính đơn biến:

Doanh số dự đoán = 651,523 + 6,261 Chi phí chào hàng*

Đây chính là phương trình của đường thẳng mà bạn thấy trong Hình 9.6. Ngoài ra trong bảng này tại cột thứ 5 còn thể hiện hệ số hồi qui chuẩn hóa (Standardized regression coefficient) được ký hiệu là Beta. Thực chất thì hệ số beta là độ dốc của đường thẳng mà ta vừa tìm được theo phương pháp bình phương bé nhất khi cả hai biến X và Y được biểu diễn bằng thang đo chuẩn hóa (Z scores). Hệ số hồi qui

chuẩn hóa sẽ được xem xét chi tiết tại mục Hệ Số Beta trong phần Hồi qui tuyến tính bội của chương này.

2.1.2 Các giả định đối với phân tích hồi qui tuyến tính.

Phân tích hồi qui không phải chỉ là việc mô tả các dữ liệu quan sát được. Từ các kết quả quan sát được trong mẫu, bạn phải suy rộng kết luận cho mối liên hệ giữa các biến trong tổng thể. Liên hệ giữa doanh số và chi phí chào hàng sẽ như thế nào đối với cả 24 khu vực bán hàng của công ty, chứ không phải chỉ ở trong 12 khu vực bán hàng được quan sát. Sự chấp nhận và diễn dịch kết quả hồi qui không thể tách rời các giả định cần thiết và những chuẩn đoán về sự vi phạm các giả định đó. Nếu các giả định bị vi phạm thì các kết quả ước lượng được không đáng tin cậy nữa. Ví dụ độ chính xác của kết luận bạn rút ra từ hàm hồi qui là "khi chi phí chào hàng tăng 1 đơn vị tiền thì doanh số tăng trung bình 6,261 đơn vị tiền" còn phụ thuộc vào mức độ thoả mãn những giả định, nếu những giả định này được thoả mãn thì sau đó bạn có thể tin tưởng vào sự diễn dịch kết quả của mình.

Sự suy rộng các kết quả của mẫu cho các giá trị của tổng thể phải trên cơ sở các giả định cần thiết sau:

- Phân phối chuẩn và phương sai bằng nhau: đối với bất kỳ giá trị nào của biến độc lập X, thì phân phối của biến phụ thuộc Y là phân phối chuẩn với trung bình của Y tại một giá trị X cụ thể là $\mu(Y/X)$ và phương sai không đổi σ^2 (xem Hình 9.8). Giả thiết này cho rằng không phải tất cả các khu vực bán hàng có cùng chi phí chào hàng bằng nhau cũng có doanh số bằng nhau. Mà thay vì vậy, sẽ có một phân phối chuẩn của doanh số tại mỗi mức chi phí chào hàng. Mặc dù các phân phối này có trung bình khác nhau, chúng đều có phương sai bằng nhau.
- Độc lập: các giá trị Y độc lập thống kê đối với nhau, tức là quan sát này không bị ảnh hưởng bởi các quan sát khác. Trong ví dụ đang xem là doanh số bán hàng của khu vực này sẽ không ảnh hưởng đến doanh số bán hàng của khu vực khác. Về mặt tiếp thị thì điều này có nghĩa là sự ưa chuộng và mua của dân cư ở một khu vực này không ảnh hưởng đến sự ưa chuộng và mua của dân cư một khu vực khác (chẳng hạn như người Đà Nẵng chấp nhận mua hay không

không phải là do người Sài Gòn đã mua hay không mua nhãn hiệu đó). Mặt khác đó là sự phân chia khu vực bán hàng theo khu vực địa lý là rõ ràng, không có sự chồng chéo cạnh tranh giữa chính lực lượng bán hàng của hai khu vực kế cận nào đó.

- Tuyến tính: tất cả các giá trị trung bình μ(Y/X) đều nằm trên một đường thẳng - đường hồi qui của tổng thể. Đây là đường thẳng được kẻ trong Hình 9.8. Nói cách khác, giả định này cho rằng mô hình hồi qui tuyến tính ta lựa chọn là đúng nên các giá trị Y trung bình ước lượng được từ mô hình tại một giá trị cụ thể của X đều nằm trên đường hồi qui tổng thể.
- Khi chỉ có một biến độc lập, thì mô hình hồi qui tuyến tính tổng thể của chúng ta có thể được mô tả bằng : $Y_i = \beta_0 + \beta_1 * X + e_i$.

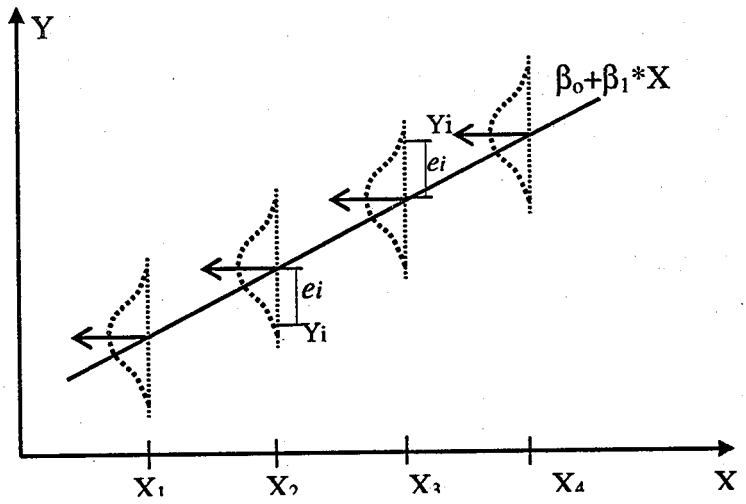
  Các hệ số độ dốc và hằng số của tổng thể được ký hiệu là β_1 và β_0. Thành phần e_i được gọi là sai số thực, là chênh lệch giữa giá trị thực Y_i quan sát được và giá trị dự báo (trung bình của các giá trị của biến Y tại điểm Xi) tức là

$$e_i = (Y_i - \hat{Y}_i) = Y_i - (\beta_0 + \beta_1 * X)$$

  Bởi vì mô hình hồi qui mà ta xây dựng vẫn còn bỏ qua những nhân tố khác có tác động đến Yi , các dữ liệu thu được về X và Y vẫn có các sai số đo lường ta không thể kiểm soát được, dạng của mối quan hệ giữa X và Y chưa chắc là tuyến tính ... nên vẫn còn các sai lệch giữa giá trị thực tế và giá trị lý thuyết của biến phụ thuộc, các sai lệch đó được thể hiện trong e_i. Vì trong sai lệch có chứa tác động của các yếu tố ảnh hưởng khác chưa nghiên cứu nên SPSS quy ước gọi sai số này là Residual (phần dư), nên từ đây về sau chúng ta cũng gọi e_i là phần dư. Trong phân tích hồi qui phần dư e_i được cho là biến ngẫu nhiên, độc lập, có phân phối chuẩn với trung bình bằng 0 và phương sai không đổi σ^2 (Hình 9.8) nếu như thật sự mô hình hồi qui tuyến tính phù hợp với các dữ liệu quan sát. Giả định về phân phối của phần dư rất quan trọng nên bạn sẽ thấy các kiểm tra vi phạm giả định hồi qui chủ yếu xoay quanh phần dư.

  Nếu mô hình hồi qui mẫu $Y_i = B_0 + B_1 + E_i$ phù hợp với dữ liệu thì phần dư quan sát được E_i trên dữ liệu mẫu (được coi là ước lượng của sai số thực e_i) phải có những đặc trưng tương tự.

Hình 9.8 Mô tả các giả định cần thiết trong phân tích hồi qui



Độ lớn của các phần dư sẽ được đánh giá dễ dàng hơn dưới dạng tương đối. Nên người ta có thể điều chỉnh phần dư theo hai phương pháp. Phương pháp thứ nhất là chuẩn hóa bằng cách chia phần dư của quan sát thứ i cho độ lệch chuẩn của các phần dư trong mẫu quan sát. Phép chia này cho ra kết quả là phần dư chuẩn hóa (standardized residuals) được tính bằng đơn vị độ lệch chuẩn lớn hơn hay nhỏ hơn trị trung bình. Các phần dư chuẩn hóa có trung bình là 0 và độ lệch chuẩn là 1. Phương pháp thứ hai là Student hóa (studentize) phần dư bằng cách chia phần dư cho ước lượng độ lệch chuẩn thay đổi từ điểm này qua điểm khác theo khoảng cách từ X_i đến trung bình $\bar{X}$. Thông thường các phần dư chuẩn hóa và student hóa có giá trị gần bằng nhau nhưng không phải lúc nào cũng vậy. Phần dư student hóa phản ảnh chính xác hơn khác biệt phương sai của sai số thực của các điểm quan sát.

2.1.3 Độ chính xác khi ước lượng các tham số của tổng thể từ các hệ số hồi qui mẫu

Một giá trị thống kê tính toán từ mẫu cho chúng ta một ước lượng điểm về tham số chưa biết của tổng thể. Một ước lượng điểm có thể được xem như là phỏng đoán tốt nhất về giá trị của tổng thể. Ta không biết đ·ược các tham số β_0 và β_1 của tổng thể, nên ta phải ước

lượng chúng từ các hệ số hồi qui tính toán được bằng phương pháp OLS từ mẫu. Các hệ số B_0 và B_1 được dùng để ước lượng các tham số này của tổng thể.

Tuy nhiên, độ dốc và hằng số (giao điểm giữa trục tung và đường thẳng) tính từ một mẫu cụ thể sẽ khác với các giá trị của tổng thể và khác nhau với độ dốc và hằng số tính từ các mẫu khác bởi vì mẫu của chúng ta suy cho cùng cũng chỉ là một mẫu được chọn ngẫu nhiên từ tổng thể, do đó hai tham số này cũng có một phân phối mẫu. Khi các giả đinh cần thiết để thực hiện hồi qui tuyến tính được thỏa mãn thì phân phối của B_0 và B_1 là phân phối chuẩn với trung bình chính là giá trị β_0 và β_1 của tổng thể.

Sai số chuẩn của B_0 là:

$$\sigma_{B_0} = \sigma\sqrt{\frac{1}{N} + \frac{\bar{X}^2}{(N-1)S_X^2}}$$

Trong đó S_X^2 là phương sai mẫu của biến độc lập.

Sai số chuẩn của B_1 là:

$$\sigma_{B_1} = \frac{\sigma}{\sqrt{(N-1)S_X^2}}$$

và bởi vì ta không biết được phương sai tổng thể của sai số σ^2 nên ta phải ước lượng. Ước lượng thông thường của σ^2 là:

$$S^2 = \frac{\sum_{i=1}^{N}(Y_i - B_0 - B_1X)^2}{N-2}$$

Căn bậc hai của S^2 được gọi là sai số chuẩn của ước lượng, hay độ lệch chuẩn của phần dư.

Các ước lượng sai số chuẩn của độ dốc B_1 và hằng số B_0 được thể hiện trong cột thứ 4 (cột Std. Error) trong Bảng 9.6

2.1.4 Đánh giá độ phù hợp của mô hình

Một công việc quan trọng của bất kỳ thủ tục thống kê xây dựng mô hình từ dữ liệu nào cũng đều là chứng minh sự phù hợp của mô hình. Như chúng ta đã thảo luận ở trên, hầu như không có đường thẳng nào

có thể phù hợp hoàn toàn với tập dữ liệu, vẫn luôn có sự sai lệch giữa các giá trị dự báo được cho ra bởi đường thẳng và các giá trị thực tế (thể hiện qua phần dư). Để biết mô hình hồi qui tuyến tính đã xây dựng trên dữ liệu mẫu phù hợp đến mức độ nào với dữ liệu thì chúng ta cần dùng một thước đo nào đó về độ phù hợp của nó.

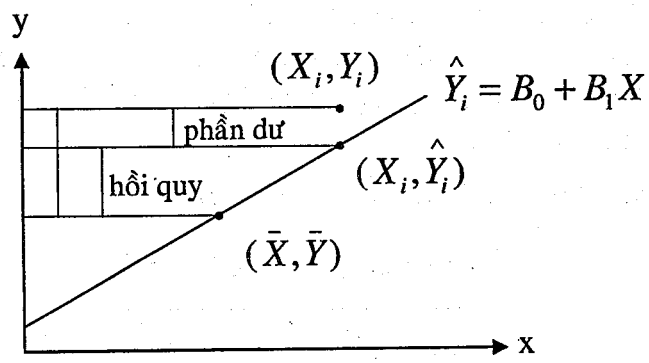
Một thước đo sự phù hợp của mô hình tuyến tính thường dùng là hệ số xác định R^2 (coefficient of determination). Công thức tính R^2 xuất phát từ ý tưởng: toàn bộ biến thiên quan sát được của biến phụ thuộc được chia thành hai phần – phần biến thiên do hồi qui (được ghi nhãn là Regression ở dòng 1 của Bảng 9.5) và phần biến thiên không do hồi qui (được ghi nhãn là Residual ở dòng 2 của Bảng 9.5) hay còn gọi là phần dư. Chúng ta sẽ diễn dịch cụ thể bản chất của từng phần như sau.

Đối với tập các giá trị mẫu của biến phụ thuộc Y, tất nhiên chúng ta có thể tính được đại lượng thống kê như trung bình, phương sai... Thông thường khi không có đường hồi qui, bạn có thể sử dụng giá trị trung bình $\bar{Y}$ để ước lượng cho biến phụ thuộc như một thói quen. Khi sử dụng trung bình để ước lượng thì ước lượng của bạn sẽ có một khoảng sai lệch (chúng ta quy ước gọi nó là sai lệch toàn bộ) là $(Y_i - \bar{Y})$. Nếu bạn chèn thêm một đại lượng $+\hat{Y}_i$ và một đại lượng $-\hat{Y}_i$ vào giữa $(Y_i - \bar{Y})$ bạn có thể viết lại công thức như sau :

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \quad (*)$$

Ý nghĩa của công thức trên có thể trình bày như Hình 9.9

Hình 9.9



Với một giá trị Yi cụ thể quan sát được, chênh lệch (khoảng cách) giữa Y_i và $\bar{Y}$ gồm hai phần

1. $(Y_i - \hat{Y}_i)$ là khoảng cách từ giá trị quan sát Y_i đến giá trị dự đoán từ mô hình (giá trị cho từ phương trình đường thẳng). Ở phần Các giả định đối với phân tích hồi qui tuyến tính ta đã biết đó được gọi là phần dư (Residual) còn lại sau khi hồi qui. Tất nhiên phần dư sẽ bằng 0 nếu đường hồi qui đi ngang qua chính giá trị quan sát Y_i, một tình huống thể hiện sự phù hợp lý tưởng của mô hình đã xây dựng được.
2. Thành phần thứ hai $(\hat{Y}_i - \bar{Y})$ là khoảng cách từ đường hồi qui đến trung bình của các giá trị Yi. Ở trên đã biết $\bar{Y}$ là giá trị được dùng dự đoán khi không có đường hồi qui. Khi xây dựng được đường hồi qui thì nguyên nhân biến động của biến phụ thuộc được giải thích bằng đường hồi qui (Regression) nên sự dự đoán được cải thiện hơn hẳn khi dùng giá trị trung bình để dự đoán. Do đó khoảng cách này chính là mức độ cải thiện đó.

Các biến đổi toán học cũng đã chứng minh được từ (*) rằng:

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 \quad (**)$$

Thành phần thứ nhất của vế phải của phương trình trên được gọi là tổng các độ lệch bình phương phần dư (Residual sum of squares). Trong Bảng 9.5 của chúng ta số liệu về nó được ghi ở cột thứ 3 dòng thứ 2 và thành phần thứ hai là tổng các độ lệch bình phương giải thích được từ hồi qui (Regression sum of squares) được thể hiện ở cột 3 dòng 1. Tổng của chúng (vế trái) gọi là tổng các độ lệch bình phương toàn bộ (Total sum of squares), giá trị này được ghi cùng cột với 2 giá trị trên tại dòng cuối cùng, bạn có thể thực hiện phép cộng để kiểm tra lại điều này.

Chia 2 vế (**) cho $\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ rồi chuyển vế ta có

$$\frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

Tỷ số bên trái dấu = chính là R^2, có thể thấy khi điểm quan sát nằm càng gần đường hồi qui (tức đường hồi qui được ước lượng càng phù hợp) thì (Y_i - $\hat{Y}_i$) càng nhỏ nên R^2 càng gần 1. Từ đó người ta sử dụng R^2 làm thông số đo lường độ thích hợp của đường hồi qui theo quy tắc R^2 càng gần 1 thì mô hình đã xây dựng càng thích hợp, R^2 càng gần 0 mô hình càng kém phù hợp với tập dữ liệu mẫu.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}$$

$$0 \leq R^2 \leq 1.$$

Trong bảng kết quả 7.4, R^2 được ghi nhãn là R Square (cột thứ 3).

Trong trường hợp mô hình hồi qui tuyến tính của bạn chỉ có một biến độc lập (đơn biến) thì hệ số này là bình phương của hệ số tương quan giữa hai biến X và Y, bạn hãy thử bình phương hệ số r tính được giữa doanh số với chi phí chào hàng và so sánh nó với giá trị Rsquare ở cột thứ 3 Bảng 9.4 xem có bằng nhau không. Nếu không có liên hệ tuyến tính giữa hai biến độc lập và phụ thuộc thì R^2 bằng 0 là do r = 0.

R^2 còn một ý nghĩa khác, nó là hệ số đo lường mối tương quan giữa giá trị quan sát Yi của biến phụ thuộc và giá trị dự đoán $\hat{Y}_i$. Nếu bạn tính được các giá trị dự đoán $\hat{Y}_i$ của biến phụ thuộc từ đường thẳng hồi qui phù hợp, tính hệ số tương quan đơn giữa Yi và $\hat{Y}_i$ rồi đem bình phương hệ số đó, kết quả sẽ bằng R^2. Quy tắc này vẫn đúng trong trường hợp phương trình hồi qui nhiều hơn 1 biến giải thích (với điều kiện phương trình đó phải có hằng số tung độ gốc B_0)

Rõ ràng nếu tất cả các doanh số thực tế quan sát được đều nằm ngay trên đường thẳng hồi qui tuyến tính thì hệ số tương quan giữa giá trị thực tế và giá trị dự báo r = 1 và do đó R^2 cũng bằng 1. R^2 = 1 thể hiện mô hình hồi qui tuyến tính ta xây dựng được phù hợp 100% với tập dữ liệu mẫu. Đây là một tình huống gần như không tưởng, mô hình có tốt đến mấy cũng không thể đạt được giá trị R^2 này vì ta đã

biết còn có các nhân tố tác động khác mà ta không thể nhận biết hết được, hoặc có nhận biết cũng khó có thể mô hình hoá nó được, nếu mô hình hoá được cũng chưa chắc đã thu thập được dữ liệu về nó.

Với ví dụ của chúng ta, $R^2 = 0,818$ nghĩa là mô hình hồi qui tuyến tính đã xây dựng phù hợp với tập dữ liệu đến mức 81,8%. Hay hơn 80% khác biệt của các mức doanh số quan sát có thể được giải thích bởi sự khác biệt về chi phí chào hàng.

2.1.5 Kiểm định các giả thuyết

2.1.5.1 Kiểm định giả thuyết về độ phù hợp của mô hình (Phân tích phương sai)

Xây dựng xong một mô hình hồi qui tuyến tính, vấn đề quan tâm đầu tiên của bạn phải là xem xét độ phù hợp của mô hình đối với tập dữ liệu qua giá trị R square. Nhưng nhớ rằng sự phù hợp đó mới chỉ thể hiện giữa mô hình bạn xây dựng được với tập dữ liệu mẫu. Kết có thể mô hình hồi qui tuyến tính mẫu với các hệ số bạn đã tìm được bằng phương pháp OLS không có giá trị gì khi suy diễn cho mô hình thực của tổng thể. Để kiểm định độ phù hợp của mô hình hồi qui tổng thể chúng ta đặt giả thuyết hệ số Rsquare của tổng thể = 0. Nếu sau khi tiến hành bài toán kiểm định chúng ta có đủ bằng chứng bác bỏ giả thuyết Ho: $R^2_{pop} = 0$ thì đây là thành công bước đầu của mô hình hồi qui tuyến tính của chúng ta, bạn nhớ là chỉ bước đầu thôi đấy.

Đại lượng F được sử dụng cho kiểm định này. Nếu xác suất F nhỏ thì giả thiết $R^2_{pop} = 0$ bị bác bỏ. Trong ví dụ của chúng ta thì giá trị F được tính theo công thức sau

$$F = \frac{\dfrac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{p}}{\dfrac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N-p-1}}$$

Nếu các giả định hồi qui cần thiết được thỏa mãn thì F sẽ có phân phối theo đại lượng thống kê F với p bậc tự do ở tử số và (N-p-1) bậc tự do ở mẫu. Bạn dùng máy tính tay ráp các giá trị tương ứng trong Bảng 9.5 xem có phải F = 45,05 không

Các số liệu để tính F thường được lấy từ bảng phân tích phương sai ANOVA. Bảng phân tích phương sai ANOVA của SPSS thể hiện bậc tự do của hai thành phần tổng các độ lệch ở cột df, ở ví dụ này bậc tự do của tử là 1 (vì p là số biến độc lập trong mô hình =1) và của mẫu là (12-1-1) = 10

Cột Mean Square là độ lệch bình phương bình quân hay độ lệch quân phương bằng tổng các độ lệch bình phương chia cho bậc tự do tương ứng (df). F được tính trực tiếp từ tỉ số giữa độ lệch quân phương hồi qui và độ lệch quân phương phần dư

$$F = \frac{mean\,square\,regression}{mean\,square\,residual} = \frac{phuong\ sai\ cua\ Y\ duoc\ giai\ thich\ duoc\ boi\ X}{phuong\ sai\ cua\ Y\ duoc\ giai\ thich\ boi\ cac\ yeu\ to\ khac}$$

Giá trị F thể hiện ở cột áp chót của Bảng 9.5 tương ứng với mức ý nghĩa quan sát được là 0,0001 ở cột cuối cùng. Ta an toàn bác bỏ giả thuyết Ho và kết luận mô hình hồi qui tuyến tính xây dựng được phù hợp với tổng thể.

2.1.5.2 Kiểm định giả thiết về ý nghĩa của hệ số hồi qui

Một giả thiết thường được kiểm định là độ dốc của mô hình tổng thể (β_1) bằng 0. Vì sao chúng ta lại phải thực hiện kiểm định này? Cũng cùng ý tưởng với tình huống kiểm định R square, mặc dù mô hình hồi qui tuyến tính mẫu ta xây dựng được có giá trị hệ số độ dốc $B_1 \neq 0$, nhưng ta chưa thể chắc hệ số độ dốc của mô hình tổng thể đã khác 0, vì vậy ta phải làm kiểm định để có kết luận về β_1. Giả thuyết dùng để kiểm định giả thiết này là Ho: $\beta_1 = 0$. Ta kỳ vọng giả thuyết này sẽ bị bác bỏ vì nếu $\beta_1 = 0$ nghĩa là Y độc lập với X hay X chẳng có ảnh hưởng gì đến Y, nghĩa là mối quan hệ tương quan tuyến tính ta nhận thấy ở mẫu chỉ xảy ra do ngẫu nhiên chứ không phải bản chất, mô hình hồi qui tuyến tính ta đã xây dựng được dựa trên một mối quan hệ "giả" giữa 2 biến.

Trị thống kê dùng để kiểm định giả thuyết là

$$t = \frac{B_1}{S_{B_1}}$$

Khi các giả định cần thiết được thỏa mãn thì phân phối của đại lượng thống kê này là Student với N-2 bậc tự do.

Trị thống kê dùng để kiểm định giả thiết hằng số (β_0) bằng không là:

$$t = \frac{B_o}{S_{B_o}}$$

Phân phối của đại lượng thống kê này là Student với N-2 bậc tự do.

Các giá trị thống kê t và mức ý nghĩa hai phía quan sát được của kiểm định t đối với giả thuyết về các hệ số hồi qui thể hiện trong hai cột cuối cùng của Bảng 9.6. Bạn thấy mức ý nghĩa quan sát được đối với hệ số độ dốc của doanh số = 0,000 chứng tỏ rằng giả thuyết Ho: $\beta_1 = 0$ có thể bị bác bỏ với độ tin cậy rất cao (99%). Giả thuyết Ho: $\beta_1 = 0$ cũng đồng nghĩa với giả thuyết doanh số và chi phí chào hàng không có liên hệ tuyến tính.

Chú ý

- Khi chỉ có một biến độc lập trong mô hình thì giả thiết R^2 tổng thể bằng không cũng đồng nghĩa với giả thiết độ dốc tổng thể bằng 0. Nên với mô hình hồi qui đơn chỉ có một biến giải thích thì kiểm định F không cần thiết mà chỉ cần tiến hành kiểm định ý nghĩa của hệ số độ dốc là đủ.
- Nếu kiểm định giả thuyết H_0: $\beta_0 = 0$ đối với hằng số của mô hình cho ta mức ý nghĩa quan sát sig. > mức ý nghĩa ta chọn cho kiểm định là 5%, ta không thể bác bỏ Ho nhưng thay vào đó ta có thể phát biểu rằng "xét về mặt thống kê β_0 không lớn hơn 0 với mức ý nghĩa 5%". Thực ra vì mô hình với một biến độc lập của ta không bao giờ đầy đủ để giải thích toàn bộ biến thiên của doanh số nên hằng số B_0 bao hàm trong nó ảnh hưởng trung bình của những biến bị bỏ sót và cả ảnh hưởng của việc mô hình hoá quan hệ của X và Y ở dạng tuyến tính có thể không thực tế bằng một dạng quan hệ phi tuyến nào đó. Chính những ảnh hưởng trên sẽ khiến cho bạn không nên diễn dịch β_0 là doanh số khi khi chi phí chào hàng X = 0, còn có những cơ chế tiềm ẩn khác khiến cho sự diễn dịch của bạn không chắc đúng. Và do đó việc kiểm định giả thuyết về hằng số thường được bỏ qua.
- Đến các phần sau bạn sẽ nhận thấy rằng khi kiểm định t cho bạn thấy một trong các hệ số hồi qui (ví dụ hệ số hồi qui β_k) của mô hình hồi qui tuyến tính đa biến (p>1) không có ý nghĩa thì chưa chắc

là biến độc lập X_k không có ảnh hưởng gì đến Y hoặc X_k không hề quan trọng. Bạn chỉ có thể chắc rằng "với tập dữ liệu mẫu và mô hình được bạn mô tả" thì không có bằng chứng nào cho thấy β_k khác 0. Kí hiệu k ở đây là để chỉ tình huống biến thứ k của mô hình hồi qui tuyến tính đa biến.

2.1.6 Dự đoán bằng mô hình hồi qui

Một trong những ứng dụng cụ thể của mô hình hồi qui tuyến tính là để dự báo, chúng ta có thể dự đoán giá trị trung bình Y trong trường hợp biết được các giá trị cụ thể của X (ký hiệu là X_0) hoặc tiên đoán giá trị của Y trong một trường hợp cụ thể tại giá trị X_0. Ví dụ, chúng ta có thể tiên đoán doanh số trung bình của tất cả các khu vực bán hàng có chi phí chào hàng là 120 triệu đồng hay tiên đoán doanh số của một khu vực cụ thể có chi phí chào hàng là 120 triệu đồng.
Trong cả hai trường hợp, giá trị dự đoán bạn tính được đều như nhau:

$\hat{Y}_0 = B_0 + B_1 X_0 = 651{,}523 + (6{,}261 \text{ x } 120) = 1{,}402.9$ triệu đồng

Nhưng sai số chuẩn trong hai trường hợp này lại khác nhau.

2.1.6.1 Dự đoán giá trị trung bình

Ở trên β_0 và β_1 được ước lượng với một mức độ sai số, giá trị dự báo cũng có sai số. Sai số chuẩn khi dự đoán trị trung bình tại X_0 là

$$S_{\hat{Y}} = S\sqrt{\frac{1}{N} + \frac{(X_0 - \bar{X})^2}{(N-1)S_X^2}}$$

Công thức này cho thấy sai số chuẩn nhỏ nhất khi X_0 bằng $\bar{X}$, khi X càng xa giá trị trung bình thì sai số càng lớn. Vì vậy dự đoán trị trung bình của Y tương ứng một giá trị cụ thể của X đối với các giá trị X nằm ở trung tâm sẽ tốt hơn đối với các giá trị X quá lớn hoặc quá nhỏ. Xem Hình 9.10 bạn sẽ thấy đường màu có đậm hơn là đường nối các giá trị dự báo trung bình Y ứng với mỗi giá trị X_i, hai đường đứt nét hai bên tạo ra khoảng tin cậy 95% cho các giá trị dự báo trung bình tại mỗi mức chi phí chào hàng X_i được xác định theo công thức

$$\hat{Y} \pm t_{(1-\alpha/2;N-2)} * S_{\hat{Y}}$$

Khoảng tin cậy rộng ra khi đi xa khu vực trung tâm của đồ thị thể hiện mức độ sai số càng lớn (mức độ tin cậy giảm đi) khi X_i càng xa giá trị trung bình của nó.

Cách thực hiện bằng SPSS

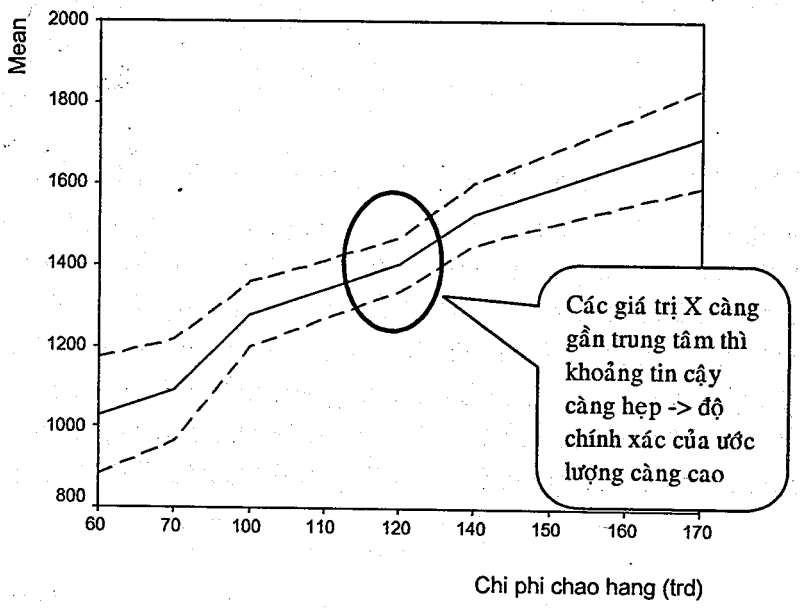
Để thu được các giá trị dự báo trung bình ứng với các giá trị của biến độc lập Xi và khoảng tin cậy 95% của chúng để vẽ được đồ thị bằng SPSS như Hình 9.10 bạn sẽ thao tác theo hướng dẫn ở cuối phần Dự đoán một giá trị riêng lẻ

2.1.6.2 Dự đoán một giá trị riêng lẻ

Như đã nói, mặc dù dự đoán của một giá trị Y tại X_0 cũng bằng với giá trị dự đoán trị trung bình Y tại X_0, nhưng sai số trong hai trường hợp này lại khác nhau. Khi dự đoán cho một quan sát riêng lẻ, bạn gặp hai nguồn sai số sau:

- Giá trị dự đoán riêng lẻ này có thể khác với trung bình của Y tại X_0
- Ước lượng trung bình tổng thể tại X_0 có thể khác với trung bình thực sự của tổng thể tại X_0 tức là giá trị ước lượng từ đường hồi qui mẫu có thể sai lệch so với giá trị thật do đường hồi qui mẫu không thực sự đại diện.

Hình 9.10



Khi ước lượng giá trị trung bình, chỉ có thành phần sai số thứ hai ở trên được xem xét. Còn khi dự báo một giá trị riêng lẻ, phương sai dự báo bằng phương sai khi dự báo giá trị trung bình cộng với phương sai của các giá trị Yi tại một giá trị X_0 cụ thể này. Biểu hiện bằng công thức như sau:

$$S^2_{ind\hat{Y}} = S^2_{\hat{Y}} + S^2 = S^2 * (1 + \frac{1}{N} + \frac{(X_0 - \overline{X})^2}{(N-1)S^2_x})$$

Nếu cỡ mẫu lớn thì 1/N và $\frac{(X_0 - \bar{X})^2}{(N-1)S^2_X}$ sẽ không đáng kể (rất nhỏ) và sai số chuẩn chỉ còn là S.

Khoảng tin cậy khi dự báo cho một quan sát riêng lẻ được tính theo công thức giống khoảng tin cậy khi dự báo giá trị trung bình nhưng lấy $S_{ind\hat{Y}}$ thay cho $S_{\hat{Y}}$

$$\hat{Y} \pm t_{(1-\alpha/2;N-2)} * S_{ind\hat{Y}}$$

2.1.6.3 Cách thực hiện bằng SPSS

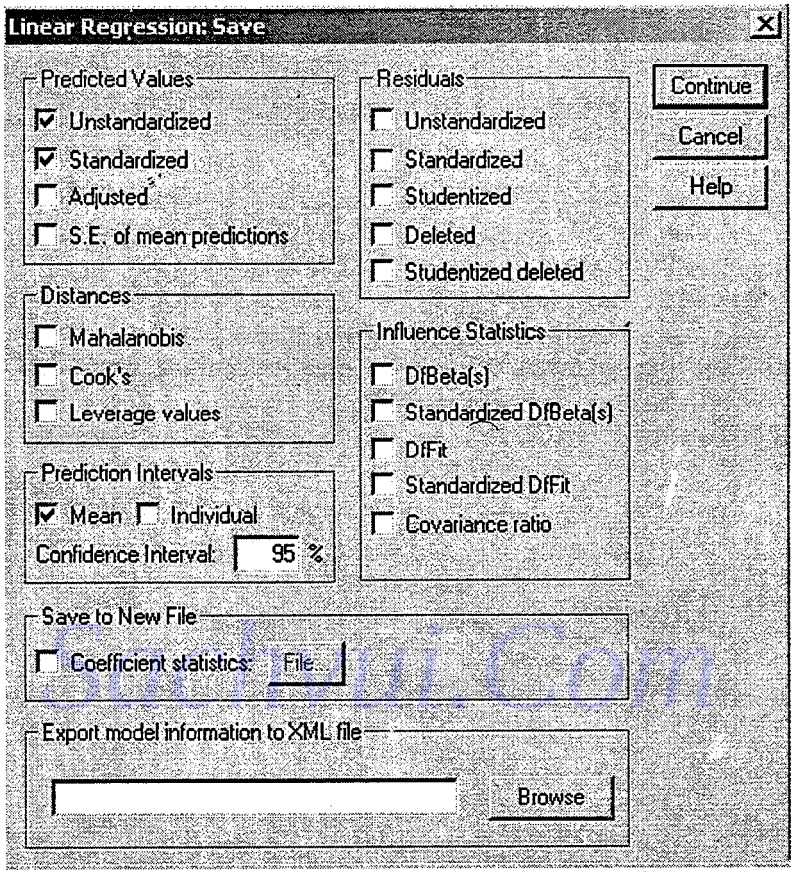
Để thu được các giá trị dự báo riêng lẻ ứng với các giá trị của biến độc lập Xi bằng SPSS, trong quá trình thực hiện lệnh lập mô hình hồi qui trong hộp thoại Linear Regression, sau khi đưa các biến trong mô hình vào đúng vị trí như đã hướng dẫn ở phần Xây dựng mô hình, bạn nhấp vào nút Save mở hộp thoại như Hình 9.11 và nhấp vào các lựa chọn như trong hình thể hiện. Trở lại hộp thoại cũ rồi nhấp OK.

Trong danh sách các biến của file *tqvahq* sẽ xuất hiện thêm các biến có tên: Unstandardized Predicted Value, 95% L CI for DOANHSO mean, 95% U CI for DOANHSO mean.

Nếu bạn đưa đồng thời cả 3 biến này lên đồ thị Line (chọn menu Graph > Line...) để quan sát chúng theo chi phí chào hàng (trong hộp thoại Line Charts nhớ chọn Multiple và tại Data in Chart là Summaries of separate Variables) bạn sẽ được đồ thị trong Hình 9.10.

Unstandardized Predicted Value cũng chính là giá trị dự báo riêng lẻ ứng với từng giá trị của biến độc lập.

Hình 9.11



2.1.7 Dò tìm sự vi phạm các giả định cần thiết trong hồi qui tuyến tính

2.1.7.1 Giả định liên hệ tuyến tính

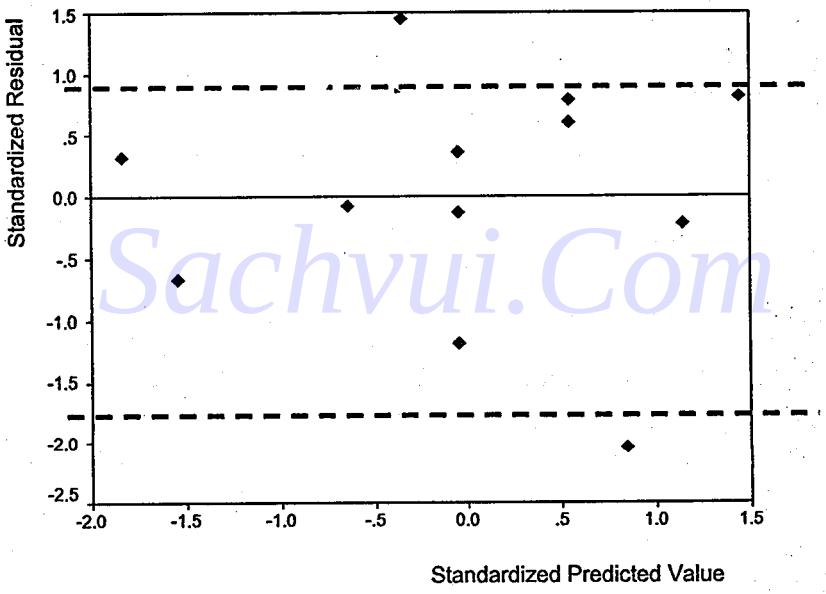
Đối với hồi qui tuyến tính hai biến, biểu đồ phân tán giữa 2 biến (Scatter) là một phương tiện tốt để đánh giá mức độ đường thẳng phù hợp với dữ liệu quan sát.

Còn có một phương pháp khác là vẽ đồ thị phân tán giữa các phần dư và giá trị dự đoán mà mô hình hồi qui tuyến tính cho ra. Người ta hay vẽ biểu đồ phân tán giữa 2 giá trị này đã được chuẩn hoá (Standardized) với phần dư trên trục tung và giá trị dự đoán trên trục hoành. Nếu giả định liên hệ tuyến tính và phương sai bằng nhau được thỏa mãn, thì ta sẽ không nhận thấy có liên hệ gì giữa các giá trị dự đoán và phần dư, chúng sẽ phân tán rất ngẫu nhiên.

Chúng ta nên xem xét thật kỹ tất cả những kiểu biến thiên mà ta quan sát được. Giả dụ sau khi chúng ta tìm đường hồi qui tuyến tính cho các dữ liệu và mô tả phần dư cùng giá trị dự đoán lên đồ thị mà thấy phần dư của chúng thay đổi theo một trật tự nào đó (có thể là cong dạng bậc 2 Parabol, cong dạng bậc 3 Cubic...) thì mô hình hồi qui tuyến tính mô tả quan hệ đường thẳng là không phù hợp với các dữ liệu này. Sự thay đổi có hệ thống giữa các giá trị dự đoán và phần dư chứng tỏ rằng giả định có quan hệ tuyến tính đã bị vi phạm.

Nếu giả định tuyến tính được thỏa mãn (đúng) thì phần dư phải phân tán ngẫu nhiên trong một vùng xung quanh đường đi qua tung độ 0 như trong Hình 9.12 chứ không tạo thành một hình dạng nào.

Hình 9.12



Chúng ta có thể vẽ đồ thị phân tán của Standardized residual và Standardized predicted valuenhư Hình 9.12 bằng cách sao lưu giá trị dự đoán chuẩn hóa (Standardized predicted value) và phần dư chuẩn hóa (Standardized residual) trong hộp thoại Linear Regression khi tiến hành thủ tục hồi qui (nhấp vào nút save mở hộp thoại Hình 9.11 rồi nhấp các chọn lựa), sau đó vẽ biểu đồ phân tán (Scatter trong menu Graph) cho 2 giá trị này.

2.1.7.2 Giả định phương sai của sai số không đổi

Chúng ta cũng có thể sử dụng công cụ hữu dụng là đồ thị ở trên để kiểm tra xem giả định phương sai của sai số không đổi có bị vi phạm không. Nếu độ lớn của phần dư tăng hoặc giảm cùng với các giá trị dự đoán (hay giá trị của biến độc lập mà ta nghi ngờ gây ra hiện tượng phương sai thay đổi đối với mô hình hồi qui tuyến tính bội), chúng ta nên nghi ngờ giả định phương sai cuả sai số không đổi đã bị vi phạm.Với ví dụ mô hình doanh số theo chi phí chào hàng của chúng ta trên Hình 9.12, hai đường gạch đứt nét được bổ sung vào đồ thị giúp bạn nhìn rõ hơn sự thay đổi của phần dư, nếu phương sai không đổi thì các phần dư phải phân tán ngẫu nhiên quanh trục 0 (tức quanh giá trị trung bình của phần dư) trong một phạm vi không đổi.

Hiện tượng "Phương sai thay đổi" ("Heteroskedasticity") này gây ra khá nhiều hậu quả tại hại đối với mô hình ước lượng bằng phương pháp OLS. Nó làm cho các ước lượng của các hệ số hồi qui không chệch nhưng không hiệu quả (tức là không phải là ước lượng phù hợp nhất), ước lượng của các phương sai bị chệch làm kiểm định các giả thuyết mất hiệu lực khiến bạn đánh giá nhầm về chất lượng của mô hình hồi qui tuyến tính (vì vậy ở phần kiểm định giả thuyết về hệ số hồi qui chúng ta chưa nên vội vã kết luận quá chắc chắn về những gì kết quả kiểm định đưa ra).

Nếu bạn ưa chuộng một thủ tục kiểm định giả thuyết chuẩn tắc, bạn có thể sử dụng kiểm định White, kiểm định Glesjer... nhưng những kiểm định này chỉ phù hợp khi cỡ mẫu của bạn lớn. Với cỡ mẫu nhỏ ở ví dụ này chúng ta sẽ sử dụng một loại kiểm định khá đơn giản là kiểm định tương quan hạng Spearman. Giả thuyết đặt ra cho kiểm định là Phương sai của sai số thay đổi, nếu giả thuyết này đúng thì hệ số tương quan hạng tổng thể giữa phần dư và biến độc lập sẽ khác 0.

Cách thực hiện bằng SPSS

1. Bạn thực hiện lệnh hồi qui và sao lưu phần dư như đã thảo luận ở trên
2. Lấy giá trị tuyệt đối của phần dư bằng lệnh Compute, nếu chưa rõ cách thức tiến hành lệnh Compute bạn trở lại xem phần kiểm định Kolmogrov một mẫu ở cuối Chương Kiểm định phi tham số. Bạn sẽ khai báo tên biến mới là *ABScuare* và công thức tính toán biến cho lệnh Compute như Hình 9.13, chú ý là lệnh tính trị tuyệt đối nằm ngay hàng

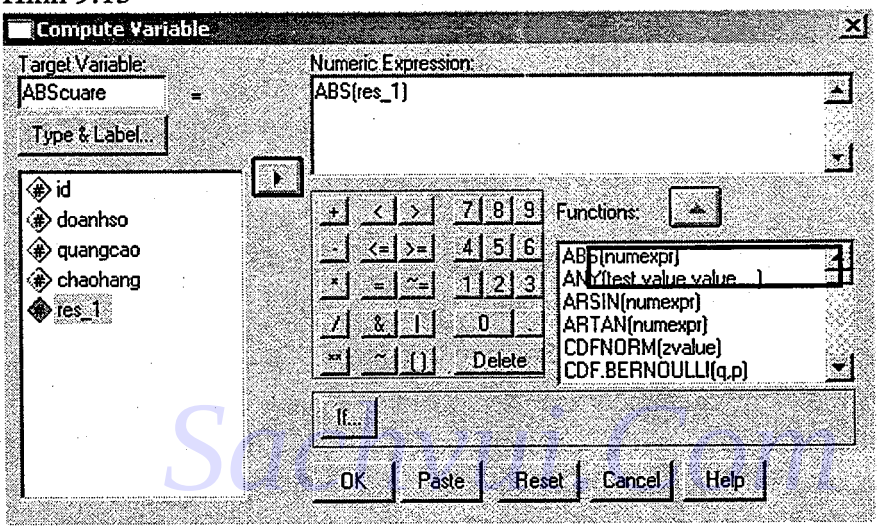
đầu trong danh sách các hàm ở khung Functions.

3. Thực hiện kiểm định tương quan hạng Spearman cho 2 biến *chaohang* và *ABScuare*

Giả thuyết Ho là: hệ số tương quan hạng của tổng thể bằng 0

Nếu kết quả kiểm định không bác bỏ giả thuyết Ho, đây là một tin tốt vì bạn có thể kết luận phương sai của sai số không thay đổi, ngược lại nếu giá trị Sig. của kiểm định nhỏ hơn mức ý nghĩa chúng ta phải chấp nhận giả thuyết phương sai của sai số thay đổi.

Hình 9.13



Kết quả kiểm định Spearman cho mối tương quan giữa 2 biến *chaohang* và *ABScuare.*

Bảng 9.7 **Correlations**

			Chi phi chao hang (trd)	ABSCUARE
Spearman's rho	Chi phi chao hang (trd)	Correlation Coefficient	1.000	.282
		Sig. (2-tailed)	.	.374
		N	12	12
	ABSCUARE	Correlation Coefficient	.282	1.000
		Sig. (2-tailed)	.374	.
		N	12	12

Kết quả kiểm định cho thấy chúng ta không thể bác bỏ giả thuyết Ho: hệ số tương quan hạng của tổng thể bằng 0, như vậy giả thuyết

phương sai của sai số thay đổi bị bác bỏ trong ví dụ của chúng ta.

Nếu mô hình hồi qui có nhiều biến giải thích thì hệ số tương quan hạng có thể tính giữa trị tuyệt đối của phần dư với từng biến riêng và áp dụng quy tắc kết luận như trên.

2.1.7.3 Giả định về phân phối chuẩn của phần dư

Phần dư có thể không tuân theo phân phối chuẩn vì những lý do như: sử dụng sai mô hình, phương sai không phải là hằng số, số lượng các phần dư không đủ nhiều để phân tích ... Vì vậy chúng ta nên thực hiện nhiều cách khảo sát khác nhau. Một cách khảo sát đơn giản nhất là xây dựng biểu đồ tần số của các phần dư như trong Hình 9.14.

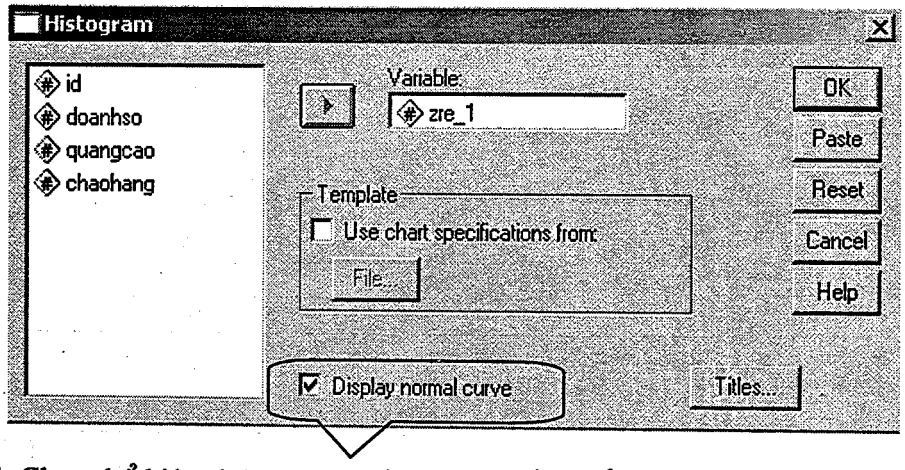
Cách xây dựng biểu đồ tần số Histogram để khảo sát phân phối của phần dư

1. Bạn thực hiện lệnh hồi qui và sao lưu phần dư chuẩn hoá (Standardized) như đã thảo luận ở trên. Bạn sẽ có thêm một biến tên *zre_1* được SPSS cập nhật vào danh sách biến của file.

2. Vào menu Graphs>Histogram... mở hộp thoại Histogram

3. Đưa biến mới tạo được là *zre_1* vào khung Variable(zre là viết tắt của phần dư chuẩn hoá và 1 là của hồi qui lần thứ nhất; theo thứ tự bạn làm đi làm lại thủ tục hồi qui SPSS sẽ đặt các con số thứ tự tương ứng)

Hình 9.14



4. Chọn thể hiện đường cong của phân phối chuẩn lên trên đồ thị.

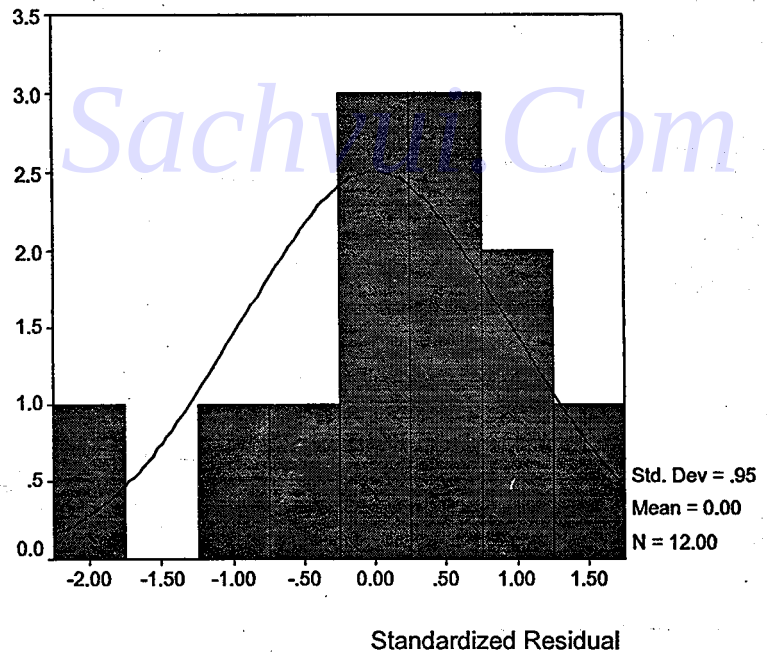
5. Nhấp OK bạn có đồ thị Hình 9.15

Như vậy bạn đã biết thêm một cách vẽ đồ thị Histogram khác, dĩ nhiên bạn cũng có thể vẽ đồ thị này bằng lệnh thống kê mô tả mà chúng ta đã nghiên cứu ở Chương III

Hình 9.15 cho thấy một đường cong phân phối chuẩn được đặt chồng lên biểu đồ tần số. Thật không hợp lý khi chúng ta kỳ vọng rằng các phần dư quan sát có phân phối hoàn toàn chuẩn vì luôn luôn có những chênh lệch do lấy mẫu. Ngay cả khi các sai số có phân phối chuẩn trong tổng thể đi nữa thì phần dư trong mẫu quan sát cũng chỉ xấp xỉ chuẩn mà thôi. Trong ví dụ này, có thể nói phân phối phần dư xấp xỉ chuẩn (trung bình Mean = 0 và độ lệch chuẩn Std. Dev. = 0,95 tức là gần bằng 1). Do đó có thể kết luận rằng giả thiết phân phối chuẩn không bị vi phạm.

Nếu bạn có một mẫu lớn, phân phối của phần dư có thể xem như tiệm cận chuẩn.

Hình 9.15 Biểu đồ tần số của phần dư chuẩn hóa.



Cách xây dựng biểu đồ tần số Q-Q plot để khảo sát phân phối của phần dư

Mặc dù bạn có thể khảo sát biểu đồ Histogram và cả biểu đồ thân lá

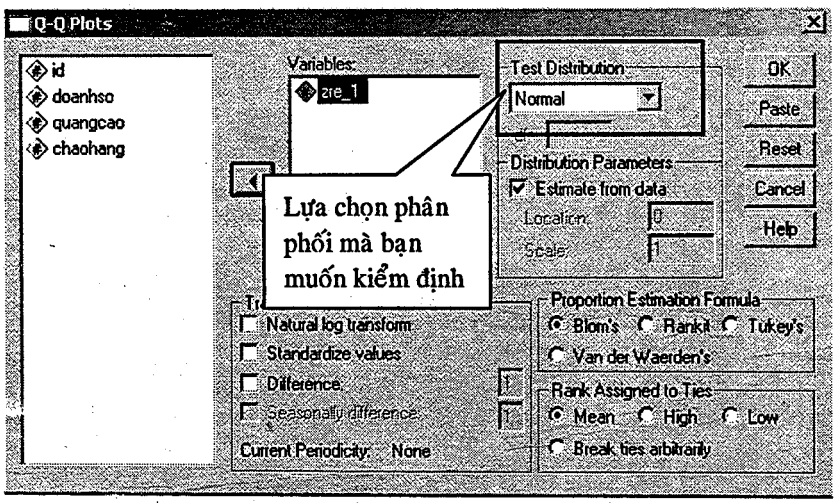
để xem phần dư có phân phối chuẩn hay không nhưng có một dạng biểu đồ đặc biệt hơn có thể giúp bạn tiếp cận nhanh chóng vấn đề này là biểu đồ Q-Q plot. Q-Q plot thể hiện những giá trị của các điểm phân vị của phân phối của biến theo các phân vị của phân phối chuẩn. Những giá trị kỳ vọng này tạo thành một đường chéo. Các điểm quan sát thực tế sẽ tập trung sát đường chéo nếu dữ liệu có phân phối chuẩn.

Bạn vẽ Q-Q plot bằng menu Graphs > Q-Q plot, trong hộp thoại này bạn sẽ khai báo đơn giản như sau

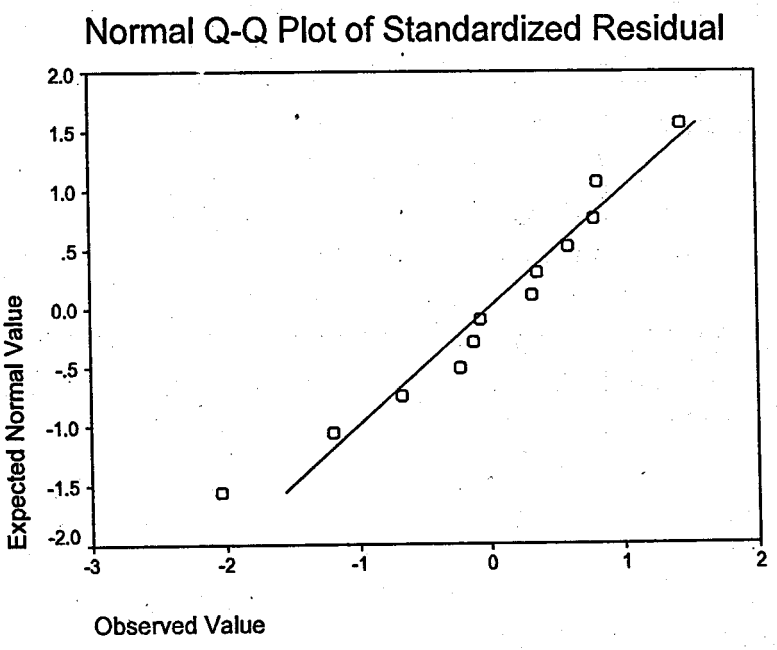
1. Đưa biến muốn kiểm tra phân phối chuẩn vào khung Variables
2. Trong Test Distribution chọn Normal
3. Các lựa chọn khác cứ để ở chế độ mặc định.
4. Nhấp OK

Xem Hình 9.16 để biết các lựa chọn cho Q-Q plot và Hình 9.17 là kết quả, đồ thị cho thấy các chấm phân tán sát với đường chéo, phân phối phần dư có thể xem như chuẩn.

Hình 9.16



Hình 9.17



Một cách khác để so sánh phân phối phần dư quan sát với phân phối chuẩn kỳ vọng tương tự Q-Q plot là vẽ cả hai phân phối tích lũy lên biểu đồ P-P Plot. Bằng cách quan sát mức độ các điểm thực tế phân tán xung quanh đường thẳng kỳ vọng, ta có thể so sánh hai phân phối này và kết luận phân phối phần dư có gần chuẩn không.

Cách thực hiện vẽ P-P plot bằng lệnh menu Graphs>P-P plot cũng tương tự Q-Q Plot. Kết quả cũng sẽ cho thấy các điểm quan sát không phân tán quá xa đường thẳng kỳ vọng, nên ta có thể kết luận là giả thiết phân phối chuẩn không bị vi phạm.

Bạn cũng đã biết một thủ tục kiểm định chuẩn tắc cho tình huống này, đó là kiểm định Kolmogrov một mẫu. Bạn hãy thử lại kiểm định này cho phần dư ở đây để củng cố hơn kết luận về phân phối chuẩn của phần dư.

2.1.7.4 Giả định về tính độc lập của sai số (không có tương quan giữa các phần dư)

Giả định về sai số thực e_i cho nó là biến ngẫu nhiên, độc lập, có phân phối chuẩn với trung bình bằng 0 và phương sai không đổi σ^2. "Độc lập" ở đây ẩn ý rằng giữa các phần dư không có mối tương quan. Khi bạn có dữ liệu được thu thập và ghi chép một cách tuần tự theo chuỗi thời gian thì giả định này càng dễ bị vi phạm. Ngay cả khi thời gian không phải là một biến trong mô hình, nó cũng có thể ảnh hưởng đến phần dư, bạn xem ví dụ sau, nếu chúng ta nghiên cứu thời gian sống sau phẫu thuật của bệnh nhân phụ thuộc vào mức độ phức tạp của cuộc phẫu thuật, lượng máu truyền, lượng thuốc sử dụng... thì ngoài những yếu tố này, có thể sự khéo tay của bác sĩ phẫu thuật đã gia tăng qua mỗi cuộc giải phẫu và do đó có thể thời gian sống của bệnh nhân chịu ảnh hưởng bởi số ca phẫu thuật bác sĩ đã thực hiện trước đó. Nếu bạn đưa lên đồ thị phần dư chuẩn hóa tương ứng với thứ tự bệnh nhân được phẫu thuật bạn sẽ thấy những bệnh nhân đầu có thời gian sống ngắn hơn những bệnh nhân về sau. Tính chất này được gọi là tương quan chuỗi (Serial Correlation). Nếu trình tự bệnh nhân theo thời gian và phần dư độc lập với nhau thì chúng ta sẽ không thấy một kiểu biến thiên rõ ràng nào trên đồ thị.

Ở phần trước chúng ta đã nhắc đến một số lý do dẫn đến sự tồn tại phần dư e_i đó là các biến có ảnh hưởng không được đưa hết vào mô hình do giới hạn và mục tiêu của nghiên cứu, chọn dạng tuyến tính cho mối quan hệ lẽ ra là phi tuyến, sai số trong đo lường các biến... các lý do này có thể dẫn đến vấn đề tương quan chuỗi trong sai số và tương quan chuỗi cũng gây ra những tác động sai lệch nghiêm trọng đến mô hình hồi qui tuyến tính như hiện tượng phương sai thay đổi.

Đại lượng thống kê Durbin-Watson (d) có thể dùng để kiểm định tương quan của các sai số kề nhau (Tương quan chuỗi bậc nhất). Giả thuyết khi tiến hành kiểm định này là:

Ho: hệ số tương quan tổng thể của các phần dư = 0

(d) được định nghĩa như sau: $$d = \frac{\sum_{i=2}^{N}(E_i - E_{i-1})^2}{\sum_{i=2}^{N}(E_i^2)}$$

Đại lượng d có giá trị biến thiên trong khoảng từ 0 đến 4. Nếu các phần dư không có tương quan chuỗi bậc nhất với nhau, giá trị d sẽ gần bằng 2. Giá trị d thấp (và nhỏ hơn 2) có nghĩa là các phần dư gần nhau có tương quan thuận. Giá trị d lớn hơn 2 (và gần 4) có nghĩa là các phần dư có tương quan nghịch.

Chúng ta có thể tra bảng thống kê Durbin-Watson để tìm các giới hạn d_L và d_U với N là số quan sát của mẫu và k là số biến độc lập trong mô hình để kiểm định mức ý nghĩa theo quy tắc quyết định

Có tự tương quan thuận chiều (dương)	*Miền không có kết luận*	*Chấpnhận giả thuyết không có tự tương quan chuỗi bậc nhất*	*Miền không có kết luận*	*Có tự tương quan ngược chiều (âm)*
0	d_L	d_U **2**	$4-d_U$	$4-d_L$ 4

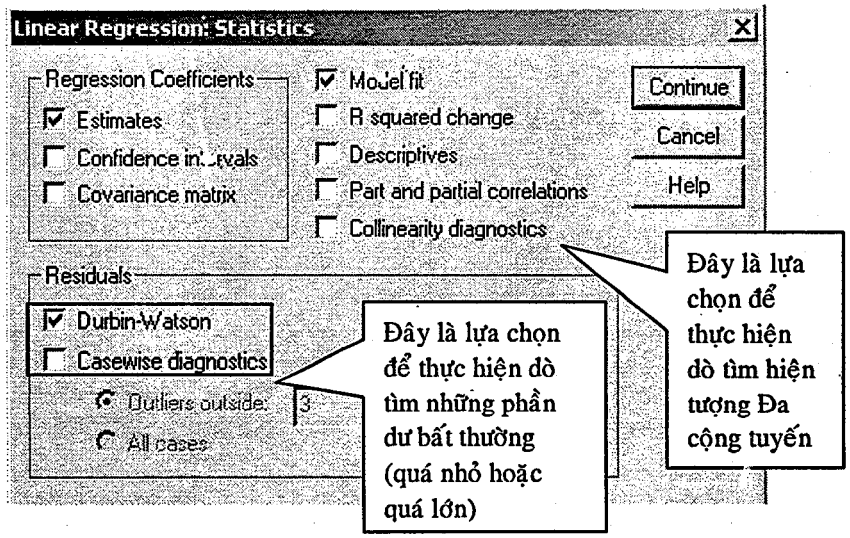
Cách thực hiện kiểm định Durbin-Watson với SPSS

1. Bạn thực hiện lại thủ tục xây dựng hồi qui như đã biết ở phần trên

2. Nhấp nút Statistics... để mở hộp thoại Linear Regression: Statistics như Hình 9.13. Trong hộp thoại này ngoài những lựa chọn mặc định bạn chọn thêm Durbin-Watson ở khung Residuals

3. Nhấp nút Continue trở về hộp thoại cũ rồi nhấp OK.

Hình 9.18



Sau khi thực hiện tiến trình này, trong bảng Model Summary, SPSS cung cấp thêm cho bạn thông tin về trị kiểm định d của Durbin-Watson ở cột cuối cùng, bạn thử so sánh Bảng 9.8 dưới đây với Bảng 9.4 xem.

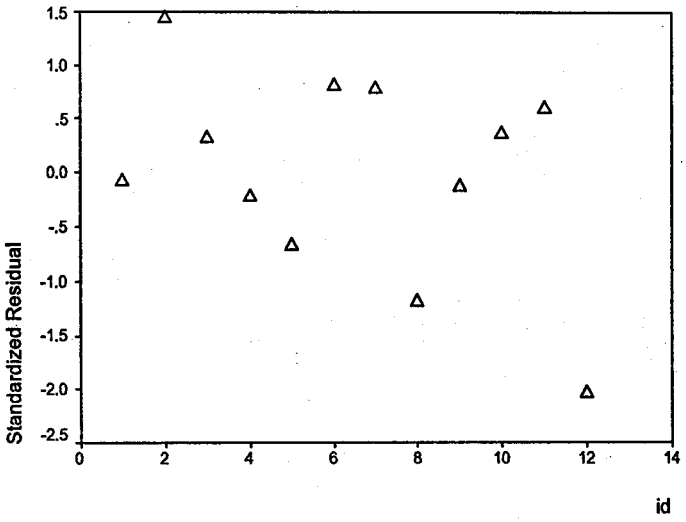
Bảng 9.8 **Model Summary(b)**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.905(a)	.818	.800	103.741	1.848

a Predictors: (Constant), Chi phi chao hang (trd)
b Dependent Variable: Doanh so ban (trd)

Giá trị d tra bảng Durbin-Watson với 1 biến độc lập và 12 quan sát là (d_L = 0,97;d_U =1,3), giá trị d tính được rơi vào miền chấp nhận giả thuyết không có tương quan chuỗi bậc nhất. Đồ thị thể hiện tuần tự phần dư theo thứ tự quan sát cũng khẳng định điều này vì chúng không mô tả một quy luật nào trong mối quan hệ giữa các phần dư. Bạn có thể vẽ đồ thị phân tán phần dư chuẩn hoá theo thứ tự quan sát (biến *id*) để kiểm chứng điều này.

Hình 9.19



2.1.7.5 Giả định không có mối tương quan giữa các biến độc lập (đo lường Đa cộng tuyến)

Cộng tuyến là trạng thái trong đó các biến độc lập có tương quan chặt chẽ với nhau. Vấn đề của hiện tượng cộng tuyến là chúng cung cấp cho mô hình những thông tin rất giống nhau, và rất khó tách rời ảnh hưởng của từng biến một đến biến phụ thuộc. Hiệu ứng khác của sự tương quan khá chặt giữa các biến độc lập là nó làm tăng độ lệch chuẩn của các hệ số hồi qui và làm giảm trị thống kê t của kiểm định ý nghĩa của chúng nên các hệ số có khuynh hướng kém ý nghĩa hơn khi không có đa cộng tuyến trong khi hệ số xác định R square vẫn khá cao. Chính vì vậy ngay khi kiểm định giả thuyết hệ số hồi qui tổng thể bằng 0 không thể bị bác bỏ bạn cũng chớ vội kết luận trước khi tiến hành tất cả các dò tìm vi phạm giả định.

Vì là vấn đề gây ra do sự tương quan chặt chẽ giữa các biến độc lập nên chúng ta không thể gặp vấn đề Đa cộng tuyến đối với một mô hình hồi qui tuyến tính đơn. Chính vì vậy chúng ta sẽ thảo luận lại về cách chuẩn đoán Đa cộng tuyến ở phần Mô hình hồi qui tuyến tính bội. Nhưng bạn có thể biết là Đa cộng tuyến được SPSS chuẩn đoán bằng lựa chọn Collinearity Diagnostic trong hộp thoại Linear Regression: Statistics (xem hướng dẫn ở Hình 9.18)

2.2 Mô hình hồi qui tuyến tính bội

2.2.1 Mô hình hồi qui tuyến tính bội, kí hiệu và các giả định

Mô hình hồi qui bội mở rộng mô hình hồi qui hai biến bằng cách thêm vào một số biến độc lập để giải thích tốt hơn cho biến phụ thuộc. Mô hình có dạng như sau:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + e_i$$

Ký hiệu X_{pi} biểu hiện giá trị của biến độc lập thứ p tại quan sát thứ i.

Các hệ số βk được gọi là hệ số hồi qui riêng phần (Partial regression coefficients), vì sao chúng được gọi như vậy bạn sẽ được giải thích ở phần sau.

Thành phần e_i là một biến độc lập ngẫu nhiên có phân phối chuẩn với trung bình là 0 và phương sai không đổi σ^2.

Mô hình hồi qui tuyến tính bội giả định rằng biến phụ thuộc có phân phối chuẩn đối với bất kỳ kết hợp nào của các biến độc lập trong mô hình. Ví dụ như nếu chiều cao của đứa trẻ là biến phụ thuộc, còn tuổi của đứa trẻ và chiều cao của mẹ đứa trẻ là biến độc lập, thì mô hình này cho rằng đối với bất kỳ kết hợp nào giữa tuổi của đứa trẻ và chiều cao của người mẹ thì chiều cao của trẻ có phân phối chuẩn và mặc dù trị trung bình của các phân phối này khác nhau, tất cả đều có chung một phương sai.

Một giả định quan trọng đối với mô hình hồi qui tuyến tính là không có biến giải thích nào có thể được biểu thị dưới dạng tổ hợp tuyến tính với những biến giải thích còn lại. Nếu tồn tại một quan hệ tuyến tính như vậy, khi đó xảy ra hiện tượng cộng tuyến.

Trong phần hồi qui bội này, ví dụ của chúng ta sẽ là một nghiên cứu xem mức độ tiêu thụ xăng - *ttxang* (km/lít) của các loại xe ôtô phụ thuộc như thế nào vào trọng lượng xe- *nang* (kg), công suất xe - *maluc* (hp- mã lực), số máy - *may* (số cylinder), và dung tích động cơ của xe - *lit* (lít). Trong nghiên cứu này chúng ta có 50 quan sát được trình bày trong file *hoiquyboi*

2.2.2 Xây dựng mô hình

2.2.2.1 Xem xét ma trận hệ số tương quan

Bước đầu tiên khi tiến hành phân tích hồi qui tuyến tính bội cũng là xem xét các mối tương quan tuyến tính giữa tất cả các biến. Ở mô hình hồi qui tuyến tính đơn ta chỉ cần xem mối quan hệ giữa biến độc lập với biến 1 phụ thuộc còn ở đây có nhiều biến nên ta phải xem xét tổng quát mối quan hệ giữa từng biến độc lập với biến phụ thuộc và chính giữa các biến độc lập với nhau. . Chúng ta xây dựng ma trận tương quan giữa tất cả các biến cho mục đích này.

Chúng ta cũng sử dụng lệnh Correlation > Biariate thuộc menu Analyze để tính toán ma trận hệ số tương quan nhưng lúc này ta đưa tất cả các biến có trong file *hoiquyboi* sang khung Variables. Sau khi nhấp OK, SPSS đưa ra một bảng ma trận hệ số tương quan như sau

Bảng 9.9 **Correlations**

		mục tiêu thu xang (km/lit)	cong suat dong co (HP)	trong luong xe (kg)	so may (cylinder)	dung tich dong co (lit)
Pearson Correlation	mục tiêu thu xang (km/lit)	1	-.788	-.858	-.681	-.777
	cong suat dong co (HP)	-.788	1	.786	.752	.818
	trong luong xe (kg)	-.858	.786	1	.802	.901
	so may (cylinder)	-.681	.752	.802	1	.941
	dung tich dong co (lit)	-.777	.818	.901	.941	1
Sig. (2-tailed)	mục tiêu thu xang (km/lit)	.	.000	.000	.000	.000
	cong suat dong co (HP)	.000	.	.000	.000	.000
	trong luong xe (kg)	.000	.000	.	.000	.000
	so may (cylinder)	.000	.000	.000	.	.000
	dung tich dong co (lit)	.000	.000	.000	.000	.

Ma trận này cho biết tương quan giữa biến phụ thuộc *ttxang* với từng biến độc lập, cũng như tương quan giữa các biến độc lập với nhau. Bạn hãy chú ý đến bất cứ liên hệ tương quan qua lại chặt chẽ nào giữa các biến độc lập bởi vì những tương quan như vậy có thể ảnh

hưởng lớn đến kết quả của phân tích hồi qui bội, ví dụ như gây ra hiện tượng đa cộng tuyến vừa nhắc đến ở trên.

Bạn chú ý đến các số 1 trên đường chéo, rất đơn giản, đây là hệ số tương quan tính được giữa 1 biến với chính nó. Chúng ta chỉ cần quan tâm đến phần tam giác phía dưới hay phía trên đường chéo này, vì chúng đối xứng nhau qua đường chéo.

Hệ số tương quan giữa *ttxang* và các biến độc lập còn lại đều cao (thấp nhất cũng là 0,68). Sơ bộ ta có thể kết luận các biến độc lập này có thể đưa vào mô hình để giải thích cho *ttxang*. Nhưng hệ số tương quan giữa các biến độc lập với nhau cũng cao (thấp nhất cũng tới 0,75) điều này sẽ khiến chúng ta phải xem xét lại thật kỹ vai trò của các biến độc lập trên trong mô hình hồi qui tuyến tính bội ta xây dựng được.

Sử dụng SPSS xây dựng mô hình

Bảng 9.10 trình bày bảng kết quả lệnh Regression của SPSS khi tất cả 4 biến độc lập *nang* (kg), *maluc* (hp- mã lực), *may* (số cylinder), và *lit* (lít) được đưa vào phương trình hồi qui bội để giải thích cho mức tiêu thụ xăng. Cách thức tiến hành giống hệt trường hợp hồi qui tuyến tính đơn nhưng lúc này khi tiến hành lệnh bạn đưa tất cả các biến có vai trò là biến độc lập vào khung Independent Variable của hộp thoại Linear Regression.

Kết quả SPSS tạo ra cũng gồm các bảng chính như đã nghiên cứu ở hồi qui tuyến tính đơn, chúng gồm các nội dung từ Bảng 9.10 đến 7.12. Ở đây chúng ta sẽ tách các bảng ra để đi theo minh hoạ cho tiến trình khảo sát sơ bộ mô hình đa biến của chúng ta.

2.2.2.2 Đánh giá độ phù hợp của mô hình hồi qui tuyến tính bội

Hệ số xác định R^2 đã được chứng minh là hàm không giảm theo số biến độc lập được đưa vào mô hình, bạn càng đưa thêm biến độc lập vào mô hình thì R^2 càng tăng, tuy nhiên điều này cũng được chứng minh rằng không phải phương trình càng có nhiều biến sẽ càng phù hợp hơn với dữ liệu (tức là tốt hơn). Như vậy R square có khuynh hướng là một ước lượng lạc quan của thước đo sự phù hợp của mô hình đối với dữ liệu trong trường hợp có hơn 1 biến giải thích trong

mô hình. Mô hình thường không phù hợp với dữ liệu thực tế như giá trị R^2 thể hiện.

Trong tình huống này R^2 điều chỉnh (Adjusted R square thể hiện ở cột thứ 4 của Bảng 9.10) từ R^2 được sử dụng để phản ảnh sát hơn mức độ phù hợp của mô hình hồi qui tuyến tính đa biến. R^2 điều chỉnh không nhất thiết tăng lên khi nhiều biến được thêm vào phương trình, nó là thước đo sự phù hợp được sử dụng cho tình huống hồi qui tuyến tính đa biến vì nó không phụ thuộc vào độ lệch phóng đại của R^2. R^2 điều chỉnh được tính như sau:

$$R_a^2 = R^2 - \frac{p\,(1-R^2)}{N-p-1}$$

trong đó p là số biến độc lập trong phương trình (trong tình huống mô hình hồi qui đơn biến thì p = 1)

Bảng 9.10 **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.881(a)	.776	.757	1.77383

a Predictors: (Constant), dung tich dong co (lit), cong suat dong co (HP), trong luong xe (kg), so may (cylinder)

So sánh 2 giá trị R square và Adjusted R square ở bảng trên bạn sẽ thấy Adjusted R square nhỏ hơn, dùng nó đánh giá độ phù hợp của mô hình sẽ an toàn hơn vì nó không thổi phồng mức độ phù hợp của mô hình.

2.2.2.3 Kiểm định độ phù hợp của mô hình

Kiểm định *F* sử dụng trong bảng phân tích phương sai vẫn là một phép kiểm định giả thuyết về độ phù hợp của mô hình hồi qui tuyến tính tổng thể . Ý tưởng của kiểm định này về mối quan hệ tuyến tính giữa biến phụ thuộc Y và biến độc lập cũng tương tự như ở hồi qui tuyến tính đơn biến, nhưng ở đây nó xem biến phụ thuộc có liên hệ tuyến tính với toàn bộ tập hợp các biến độc lập hay không. Giả thuyết Ho là $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

Nếu giả thuyết Ho bị bác bỏ chúng ta kết luận là kết hợp của các biến hiện có trong mô hình có thể giải thích được thay đổi của Y, điều này cũng có nghĩa là mô hình ta xây dựng phù hợp với tập dữ

liệu. Như vậy sau khi chạy ra mô hình từ SPSS thì nhiệm vụ đầu tiên là bạn phải xem giả thuyết Ho của kiểm định F có bị bác bỏ không.

Trị thống kê F được tính từ giá trị R square của mô hình đầy đủ, giá trị sig. rất nhỏ cho thấy ta sẽ an toàn khi bác bỏ giả thuyết Ho cho rằng tất cả các hệ số hồi qui bằng 0 (ngoại trừ hằng số), mô hình hồi qui tuyến tính bội của ta phù hợp với tập dữ liệu và có thể sử dụng được.

Bảng 9.11 **ANOVA(b)**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	491.837	4	122.959	39.078	.000(a)
	Residual	141.592	45	3.146		
	Total	633.429	49			

a Predictors: (Constant), dung tich dong co (lit), cong suat dong co (HP), trong luong xe (kg), so may (cylinder)

b Dependent Variable: muc tieu thu xang (km/lit)

2.2.2.4 Ý nghĩa các hệ số hồi qui riêng phần trong mô hình

Các hệ số hồi qui của từng biến độc lập trong mô hình hồi qui tuyến tính bội (Bảng 9.12) tương tự như các hệ số trong phương trình hồi qui chỉ có một biến độc lập nhưng trong hồi qui bội các hệ số này được gọi là hệ số hồi qui riêng phần

Bảng 9.12 Các thông số thống kê của từng biến trong phương trình

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	26.234	2.319		11.312	.000
	cong suat dong co (HP)	-.046	.016	-.348	-2.783	.008
	trong luong xe (kg)	-.009	.002	-.722	-4.161	.000
	so may (cylinder)	.244	.536	.100	.455	.651
	dung tich dong co (lit)	.178	.882	.063	.202	.841

a Dependent Variable: muc tieu thu xang (km/lit)

Ý nghĩa của hệ số hồi qui riêng phần là : β_k đo lường sự thay đổi trong giá trị trung bình Y khi X_k thay đổi một đơn vị, giữ các biến độc lập còn lại không đổi. Nói một cách khác, nó cho biết ảnh hưởng

"thuần" của các thay đổi một đơn vị trong X_k đối với giá trị trung bình của biến phụ thuộc Y khi loại trừ ảnh hưởng của các biến độc lập khác. Trong hồi qui tuyến tính bội, để đánh giá đóng góp "thật sự" của một biến đối với thay đổi trong Y thì bằng cách nào đó chúng ta phải "kiểm soát" được ảnh hưởng của các biến khác.

Như ví dụ này B_{nang} đo lường thay đổi trong giá trị trung bình của *ttxang* khi biến *nang* thay đổi một đơn vị, giữ *maluc, may, lit* không đổi. Có nghĩa là, nó cho biết 1 kg tăng lên của trọng lượng xe sẽ làm cho trung bình số km xe chạy được trên mỗi lít xăng giảm đi 0,009 (tức là xăng tiêu hao nhiều hơn) trong điều kiện *maluc, may, lit* không thay đổi.

Các hệ số hồi qui riêng phần của tổng thể cũng cần được thực hiện kiểm định giả thuyết Ho: $\beta_k = 0$, chúng ta thấy mặc dù R square khá cao nhưng giá trị sig. của 2 biến *may* và *lit* lại cho thấy nó không có ý nghĩa trong mô hình. Tìm hiểu các phần sau bạn đọc sẽ khám phá được lý do của vấn đề này.

Hệ số beta

Ở phần trên ta đã nhắc đến hệ số beta, vì độ lớn của các hệ số phụ thuộc vào đơn vị đo lường của các biến nên chỉ khi nào tất cả các biến độc lập đều có cùng đơn vị đo lường thì các hệ số của chúng mới có thể so sánh trực tiếp với nhau. Một cách để làm cho các hệ số hồi qui có thể so sánh được với nhau là tính trọng số beta, đó là hệ số của biến độc lập khi tất cả dữ liệu trên các biến được biểu diễn bằng đơn vị đo lường độ lệch chuẩn. Hệ số beta được tính trực tiếp từ hệ số hồi qui như sau:

$$beta_k = B_k\left(\frac{S_k}{S_Y}\right)$$

Trong đó S_k là độ lệch chuẩn của biến độc lập thứ k.

Phương trình thể hiện mức tiêu thụ xăng dự đoán theo tất cả các biến độc lập là:

ttxang=26,234 - 0,046(maluc)+0,244(may) - 0,009(nang)+ 0,178(lit)

2.2.3 Xác định tầm quan trọng của các biến trong mô hình

Trong hồi qui bội với nhiều biến độc lập chúng ta có thể muốn xác định với các biến ta đã đưa vào mô hình, biến nào có vai trò quan trọng hơn trong việc dự đoán giá trị lý thuyết của Y hay chúng quan trọng như nhau. Giả dụ chúng ta sẽ muốn biết công suất máy có vai trò quan trọng hơn trọng lượng xe trong việc dự đoán mức tiêu thụ xăng hay không? đây là 2 biến có tương quan tuyến tính mạnh nhất với mức tiêu thụ xăng.

Có 2 vấn đề quan tâm khi xác định tầm quan trọng tương đối của từng biến độc lập trong một mô hình hồi qui tuyến tính bội:

- tầm quan trọng của công suất máy và trọng lượng xe khi mỗi biến được sử dụng riêng biệt để dự đoán mức tiêu thụ xăng.
- tầm quan trọng của công suất máy và trọng lượng xe khi chúng được sử dụng cùng với những biến khác trong phương trình hồi qui để dự đoán mức tiêu thụ xăng.

Vấn đề thứ nhất được giải đáp bằng cách nhìn vào các hệ số tương quan giữa mức tiêu thụ xăng và các biến độc lập. Trị tuyệt đối của hệ số tương quan càng lớn thì liên hệ tuyến tính càng mạnh. Bảng 9.9 cho thấy trọng lượng xe tương quan với mức tiêu thụ xăng mạnh hơn công suất máy (r = 0,858 và r = 0,788). Do đó chúng ta có thể gán hệ số quan trọng (trọng số) lớn hơn cho trọng lượng xe khi dự đoán mức tiêu thụ xăng.

Vấn đề thứ hai phức tạp hơn. Khi tất cả các biến độc lập cùng tương quan với nhau (bạn hãy nhìn bảng ma trận hệ số tương quan, hệ số r của giữa các biến độc lập đều rất lớn) thì ảnh hưởng của mỗi biến đến biến phụ thuộc rất khó đánh giá. Ảnh hưởng đó bây giờ còn phụ thuộc vào các biến độc lập khác trong phương trình chứ không thể tách riêng, tức là bạn khó có thể đạt được điều kiện giữ các biến khác không đổi khi đọc ý nghĩa của hệ số hồi qui riêng phần của từng biến độc lập. Và ảnh hưởng này làm hệ số hồi qui riêng phần của một biến độc lập thay đổi cả về độ lớn. Ví dụ như hệ số hồi qui (B) của công suất máy khi phương trình hồi qui chỉ có một biến này là -0,104 trong khi đó khi cả bốn biến cùng được đưa vào phương trình thì B của công suất máy là -0,0459.

Bạn thử chạy lại mô hình hồi qui tuyến tính đơn với chỉ một biến giải thích maluc xem có phải B_{maluc} = -0,104.

Như vậy dùng các hệ số hồi qui B sẽ không thích hợp để giải thích tầm quan trọng tương đối của các biến. Và dùng hệ số Beta cũng vậy, cũng giống như các hệ số *B* là nó tùy thuộc vào các biến độc lập khác trong phương trình. Chúng cũng có thể bị ảnh hưởng bởi iên hệ tương quan giữa các biến độc lập và chúng không phản ảnh chính xác tầm quan trọng của các biến độc lập.

Để xác định tầm quan trọng của các biến khi chúng được sử dụng cùng với những biến khác trong mô hình ta dùng hệ số tương quan từng phần và tương quan riêng (Part and partial correltions)

Ý tưởng để đánh giá tầm quan trọng tương đối của các biến độc lập ở đây là xem xét mức độ tăng của R square khi một biến giải thích được đưa vào phương trình trong khi phương trình đã chứa sẵn các biến độc lập khác. Mức độ tăng này được tính bằng hiệu của R square với $R^2_{(k)}$ là bình phương hệ số tương quan bội khi tất cả các biến độc lập đã ở trong phương trình ngoại trừ biến thứ k. Công thức tính mức độ tăng này $R^2_{change} = R^2 - R^2_{(k)}$

Mức độ thay đổi của R^2 do một biến gây ra mà lớn chứng tỏ biến này cung cấp những thông tin độc nhất về biến phụ thuộc mà các biến độc lập khác trong phương trình không có được..

Căn bậc hai của mức độ gia tăng này $\sqrt{R^2_{change}}$ gọi là **hệ số tương quan từng phần** (Part correlation correlations). Nó chính là tương quan giữa Y và X_k khi ảnh hưởng tuyến tính của các biến độc lập khác đối với biến độc lập X_k bị loại bỏ. Nếu tất cả các biến độc lập không có tương quan với nhau thì mức độ thay đổi của R^2 khi một biến độc lập được đưa thêm vào phương trình đơn giản chỉ là bình phương của hệ số tương quan giữa biến độc lập này và biến phụ thuộc. Nếu mức độ thay đổi của R^2 khi đưa vào biến độc lập này mà lớn hơn mức độ thay đổi của R^2 khi đưa vào biến độc lập khác thì biến độc lập kể trước có vai trò quan trọng hơn.

Mức độ thay đổi của R^2 sẽ thể hiện ở cột R Square Change của bảng Model Summary nếu bạn chọn mục R Squared Change của hộp thoại Statistic... khi tiến hành lệnh hồi qui tuyến tính trên SPSS. Nó cho thấy nếu thêm 1 biến độc lập vào phương trình có chứa sẵn những biến khác thì sẽ làm cho R^2 thay đổi bao nhiêu.

Nhưng giá trị này chỉ cho biết R^2 tăng bao nhiêu khi một biến được thêm vào phương trình hồi qui chứ nó không chỉ ra được tỉ lệ của phần biến thiên mà một mình biến đó có thể giải thích được. Và nếu hầu hết các biến thiên trong Y đã được giải thích bởi các biến độc lập khác có trong mô hình thì những biến còn lại được đưa thêm vào chỉ có thể thay đổi R^2 một chút mà thôi.

Một hệ số có thể đo lường được khả năng giải thích biến thiên của biến phụ thuộc do ảnh hưởng của một biến độc lập là:

$$P_{r_k^2} = \frac{R^2 - R^2_{(k)}}{1 - R^2_{(k)}}$$

Trong công thức này, tử số chính là mức độ thay đổi của R^2 do một biến độc lập gây ra, còn mẫu số là phần biến thiên không giải thích được nguyên nhân khi tất cả các biến độc lập được đưa vào mô hình ngoại trừ biến thứ k. vậy có phải $P_{r_k^2}$ là khả năng giải thích được cho biến phụ thuộc trong mô hình của một biến độc lập X_k.

Căn bậc 2 của $P_{r_k^2}$ gọi là **hệ số tương quan riêng** (Partial correlation coefficient). Nó có thể được giải thích như là tương quan giữa biến độc lập thứ k và biến phụ thuộc Y khi ảnh hưởng tuyến tính của các biến độc lập khác đối với cả Y và X_k bị loại bỏ. Bởi vì mẫu số của công thức tính $P_{r_k^2}$ luôn luôn nhỏ hơn hay bằng 1, nên hệ số tương quan riêng $P_{r_k^2}$ không bao giờ lớn hơn trị tuyệt đối của hệ số tương quan từng phần (cũng là trị tuyệt đối chênh lệch của tử số của công thức tính $r^2_{r_k^2}$).

Để tính được các Hệ số tương quan từng phần và tương quan riêng, trong hộp thoại Statistics bạn chọn mục Part and partial correlations khi chạy lệnh hồi qui của SPSS.

2.2.4 Lựa chọn biến cho mô hình

Sự lựa chọn biến để đưa vào mô hình phần nào có tính chất chủ quan. Một số biến có thể không được đưa vào mà ta không lường được vai trò quan trọng của nó trong khi đó có một vài biến được sử dụng có thể lại không phải là biến quyết định cho biến thiên của biến phụ thuộc. Do vậy chúng ta sẽ áp dụng biện pháp xem xét những kết quả khi ta đưa vào hoặc bỏ ra các biến trong phương trình hồi qui để quyết định.

Ta sử dụng ví dụ với bốn biến độc lập của file *hoiquyboi* để dự đoán mức tiêu thụ xăng.

2.2.4.1 Xem xét các tác động của việc đưa vào và bỏ ra các biến

Bước đầu tiên (Model 1) trong Bảng 9.13 cho thấy các thống kê tóm tắt khi dung tích động cơ (dung tích các cylinders) là biến độc lập duy nhất trong phương trình.

Hãy xem bước thứ hai (Model 2) trong Bảng 9.14, một biến khác là trọng lượng xe được thêm vào. Giá trị R square Change trong cột 6 Bảng 9.14 là mức độ thay đổi của R^2 khi trọng lượng xe được đưa thêm vào. R^2 khi chỉ có dung tích động cơ là 0,604 nên

$$R^2_{change} = 0,736 - 0,60 = 0,132$$

Ở đây ta cũng muốn chắc giá trị thực của R square Change trong tổng thể ($R^2_{change_{(pop)}}$) cũng khác 0 hay không. Nghi ngờ này được kiểm định bằng giả thuyết :

$$\text{Ho: } R^2_{change_{(pop)}} = 0$$

Đại lượng kiểm định F được tính theo công thức

$$F_{change} = \frac{R^2_{change}(N-p-1)}{q(1-R^2)} = \frac{(0,1323)(50-2-1)}{1(1-0,7365)} = 23,604$$

Đại lượng này gọi là kiểm định F riêng (partial F test). Trong đó N là số quan sát, p là tổng số biến độc lập trong phương trình, và q là số biến được đưa vào phương trình trong bước này. Với giả thuyết cho rằng mức độ thay đổi bằng 0, ta có thể tính mức ý nghĩa của giá trị FCh từ phân phối F với q và (N-p-1) bậc tự do. Mức ý nghĩa quan sát của kiểm định này được cho luôn trong Bảng 9.14. Với mức ý nghĩa quan sát đó chúng ta an toàn bác bỏ giả thuyết Ho.

Bảng 9.13 **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.777(a)	.604	.596	2.28552

a Predictors: (Constant), dung tich dong co (lit)

Bảng 9.13b **ANOVA(b)**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	382.696	1	382.696	73.263	.000(a)
	Residual	250.733	48	5.224		
	Total	633.429	49			

a Predictors: (Constant), dung tich dong co (lit)

b Dependent Variable: muc tieu thu xang (km/lit)

Bảng 9.14 **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.777(a)	.604	.596	2.28552	.604	73.263	1	48	.000
2	.858(b)	.736	.725	1.88448	.132	23.604	1	47	.000

a Predictors: (Constant), dung tich dong co (lit)

b Predictors: (Constant), dung tich dong co (lit), trong luong xe (kg)

Cách thực hiện bằng SPSS

Bạn có thể xây dựng được các Bảng 9.13 và 7.14 (và cả Bảng 9.16) bằng cách sau:

1. Mở hộp thoại Linear Regression, đưa biến *ttxang* vào khung Dependent và *lit* vào khung Independent(s)

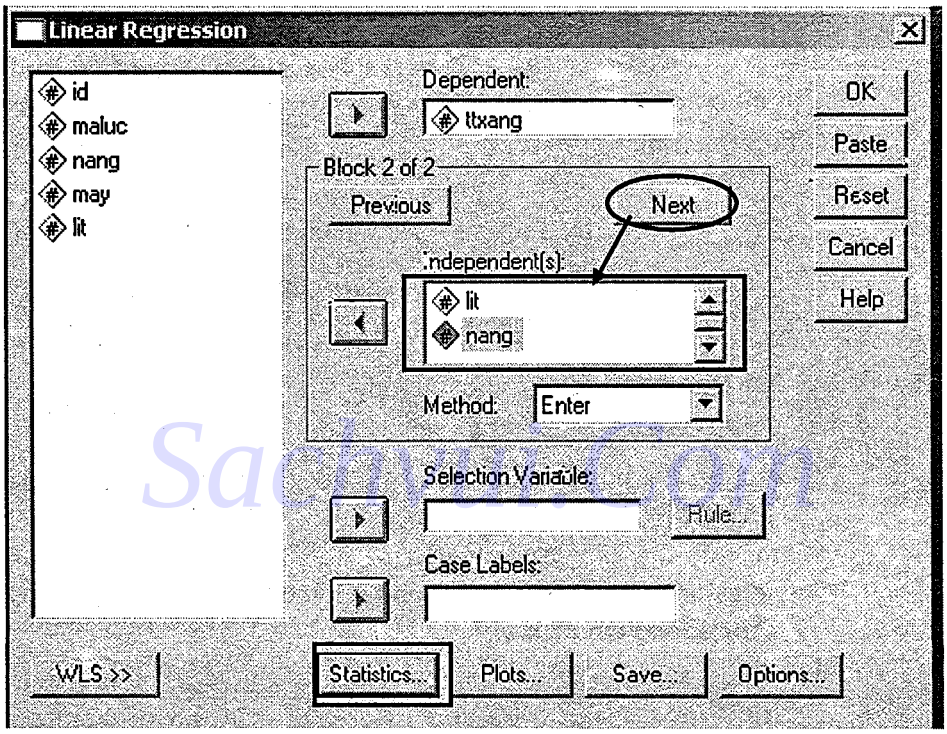
2. Kế tiếp bạn nhấp nút Next ở ngay trên khung Independent(s) của hộp thoại

3. Bạn tiến hành lại việc đưa biến độc lập vào khung Independent(s), nhưng lần này đưa đồng loạt 2 biến *lit* và *nang*.

4. Bạn nhấp vào nút Statistic... và nhấp chọn tiếp tùy chọn R squared Change trong hộp thoại Statistic.

5. Trở lại hộp thoại cũ và OK, bạn sẽ được các kết quả cần thiế

Hình 9.20



Chú ý

- Giả thuyết mức độ thay đổi thực của R^2 bằng 0 cũng có thể được phát biểu dưới hình thức kiểm định ý nghĩa của hệ số hồi qui β của biến vừa được đưa thêm vào mô hình. Khi chỉ có 1 biến độc lập Xk được đưa vào phương trình trong một bước thì giả thuyết mức độ thay đổi của R^2 bằng 0 tương đương với giả thuyết $\beta_k = 0$. Giá trị FCh để kiểm định mức độ thay đổi của R^2 lúc này chính là bình phương của giá trị t khi kiểm định giả

thuyết hệ số $\beta_k = 0$ của mô hình bao gồm biến Xk vừa được thêm vào. Bạn đọc thử chạy hồi qui mô hình *ttxang* theo 2 biến *nang* va *lit* rồi lấy giá trị t tương ứng với hệ số hồi qui của biến *nang* mà bình phương lên xem thử có đúng bằng 23,604 không. Và bạn có thể kiểm tra kết quả chạy hồi qui của mình bằng cách so với Bảng 9.15 tại khu vực Model 2.

- Khi có q biến độc lập cùng được đưa vào phương trình trong một bước, kiểm định giả thuyết mức độ thay đổi thực của R^2 bằng 0 tương đương với kiểm định đồng thời các hệ số của tất cả q biến độc lập vừa đưa vào bằng 0. Ví dụ, nếu trọng lượng xe và công suất động cơ cùng một bước được đưa vào trong phương trình có chứa dung tích máy, thì kiểm định F đối với mức độ thay đổi của R^2 sẽ giống như kiểm định F đối với giả thuyết Ho: $\beta_{nang} = \beta_{hp} = 0$. Kiểm định này giúp bạn trả lời câu hỏi q biến độc lập được đưa vào (hay không được đưa vào) có ảnh hưởng liên kết có ý nghĩa đối với Y không).

- Việc đưa trọng lượng xe vào phương trình cùng với dung tích động cơ còn có ảnh hưởng khác ngoài việc làm thay đổi R^2.
  - Độ lớn của hệ số hồi qui của biến dung tích động cơ từ Model 1 sang Model 2 là từ -2,189 sang -0,064. Điều này được hiểu là mức độ tốn xăng trung bình dưới tác động của dung tích động cơ giảm đi khi đi qua 2 mô hình (trong điều kiện trong lượng xe không đổi ở mô hình 2).
  - Giá trị sig. của phép kiểm định ý nghĩa của hệ số hồi qui đứng trước biến dung tích động cơ ở Model 2 là 0,896 khiến cho chúng ta không thể bác bỏ giả thuyết cho rằng trong tổng thể biến *lit* không hề có liên hệ gì với biến *ttxang*, mặc dù ở Model 1 giữa chúng thể hiện một mối quan hệ thật thuyết phục.

Lý do của hai khác biệt này là gì ?

Sự khác nhau trong thay đổi mức độ tiêu thụ xăng này là do tương quan giữa 2 biến độc lập dung tích động cơ và trọng lượng xe. Khi các biến độc lập có tương quan chặt với nhau được đưa vào phương trình hồi qui thì kết quả có thể bất thường. Hồi qui chung

có thể có ý nghĩa trong khi đó không có hệ số hồi qui của biến nào lại có ý nghĩa. Thậm tệ hơn nữa là dấu của hệ số hồi qui có thể đảo ngược. Tương quan chặt chẽ giữa các biến độc lập sẽ thổi phồng phương sai của các ước lượng, làm cho các hệ số hồi qui đơn không tin cậy được, không làm cho mô hình phù hợp hơn. Điều này đã được nói đến ở phần dò tìm các vi phạm giả định.

Bảng 9.15 **Coefficients(a)**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18.162	.730		24.877	.000
	dung tich dong co (lit)	-2.189	.256	-.777	-8.559	.000
	trong luong xe (kg)					
2	(Constant)	25.676	1.660		15.471	.000
	dung tich dong co (lit)	-.064	.485	-.023	-.132	.896
	trong luong xe (kg)	-.011	.002	-.838	-4.858	.000

a Dependent Variable: muc tieu thu xang (km/lit)

2.2.4.2 Xem xét các thông số thống kê của các biến không ở trong phương trình

Khi tiến hành khảo sát việc đưa vào và bỏ ra các biến độc lập trong phương trình, chúng ta có thể xem xét điều gì sẽ xảy ra nếu biến được đưa vào phương trình trong bước tiếp theo. Các thống kê mô tả biến cho ví dụ của chúng ta là biến trọng lượng xe được trình bày trong Bảng 9.16. Bảng này thể hiện các hệ số của các biến không ở trong phương trình. Cột Beta In là hệ số hồi qui chuẩn hóa nếu biến được đưa vào trong phương trình ở bước tiếp theo. Giá trị t và mức ý nghĩa tương ứng là để kiểm định hệ số hồi qui của biến được đưa thêm vào có bằng 0 không (nên nhớ rằng kiểm định t và kiểm định F là tương đương). Từ các giá trị thống kê này, chúng ta có thể xác định biến nào nên được đưa vào tiếp theo trong danh sách một vào biến không có trong mô hình. Quá trình này được đề cập chi tiết trong phần Các thủ tục chọn biến ở phía sau.

Bảng 9.16 **Excluded Variables(b)**

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	trong luong xe (kg)	-.838(a)	-4.858	.000	-.578	.189

a Predictors in the Model: (Constant), dung tich dong co (lit)
b Dependent Variable: muc tieu thu xang (km/lit)

2.2.4.3 Số lượng biến độc lập "tối ưu" cho mô hình hồi qui bội

Mặc dù việc thêm biến độc lập vào làm tăng R^2, nhưng nó chưa hẳn làm giảm sai số chuẩn của ước lượng. Mỗi lần một biến được đưa vào phương trình, tổng các độ lệch bình phương phần dư (Residual Sum of Square) sẽ mất đi một bậc tự do vì bậc tự do của nó bằng (N-p-1) với p là số biến độc lập có trong mô hình nên thêm một biến độc lập thì p tăng thêm 1 đơn vị khiến hiệu số giảm đi 1 đơn vị. Và tổng các độ lệch bình phương hồi qui (Regression Sum of Square) sẽ thêm một bậc tự do (bậc tự do của nó = p là số biến độc lập có trong mô hình).

Sai số chuẩn có thể sẽ tăng khi mức độ giảm của tổng các độ lệch bình phương phần dư rất nhỏ và không đủ bù đắp cho việc mất đi một bậc tự do của nó. Còn giá trị F để kiểm định hồi qui toàn bộ có thể giảm khi tổng các độ lệch bình phương hồi qui không tăng nhanh như mức độ tăng số bậc tự do hồi qui. Bạn đọc sẽ hiểu rõ điều này hơn nếu nhìn vào công thức tính của các đại lượng trên mà đã trình bày ở phần hồi qui đơn

Tóm lại là đưa nhiều biến độc lập vào mô hình hồi qui không phải lúc nào cũng tốt vì những lý do sau (trừ phi chúng có liên hệ rất mạnh với biến phụ thuộc):

- Mức độ tăng R^2 quan sát không hẳn phản ảnh mô hình hồi qui càng phù hợp hơn với tổng thể.
- Đưa vào các biến không thích đáng sẽ làm tăng các sai số chuẩn của tất cả các ước lượng mà không cải thiện được khả năng dự đoán.
- Mô hình có nhiều biến thì khó mà giải thích và khó hiểu hơn một mô hình ít biến.

Các vấn đề liên quan đến việc lựa chọn biến có thể được SPSS giải quyết một cách tự động và nhanh chóng qua các Thủ tục chọn biến mà chúng ta sẽ thảo luận ở 2.2.6

2.2.5. Dò tìm các vi phạm giả định cần thiết

Thủ tục kiểm tra vi phạm các giả định cần thiết trong hồi qui đơn biến đã nghiên cứu trong phần Dò tìm các vi phạm các giả định cần thiết trong hồi qui tuyến tính cũng được áp dụng trong hồi qui bội do các giả định đề ra là áp dụng chung cho hồi qui tuyến tính, ví dụ các phần dư cũng nên được vẽ đồ thị cùng với giá trị dự đoán cũng như với từng biến độc lập; phân phối của phần dư cũng cần được kiểm tra xem có phân phối chuẩn không...

Có một tình huống vi phạm giả định xảy ra riêng với hồi qui tuyến tính bội đó là hiện tượng Cộng tuyến ta đã nhắc đến nhiều lần ở trên, sự dò tìm vi phạm giả định này được gọi tên là Đo lường đa cộng tuyến (Collinearity Diagnostics)

Các công cụ chẩn đoán giúp ta phát hiện sự tồn tại của cộng tuyến trong dữ liệu và đánh giá mức độ cộng tuyến làm thoái hóa các tham số được ước lượng là:

- Độ chấp nhận của biến (Tolerance) thường được sử dụng để đo lường hiện tượng cộng tuyến. Độ chấp nhận của biến Xk được định nghĩa là $1-R_k^2$, trong đó R_k^2 là hệ số tương quan bội khi biến độc lập Xk được dự đoán từ các biến độc lập khác. Tức R_k^2 thể hiện khả năng của các biến độc lập (ngoại trừ X_k) trong mô hình của ta giải thích được cho biến động của Xk khi ta xây dựng mô hình hồi qui tuyến tính Xk theo tất cả các biến độc lập khác. Cũng như một tình huống hồi qui tuyến tính bình thường, R square này lớn thể hiện một độ phù hợp tốt của tổ hợp kết hợp tuyến tính của các biến trong mô hình. Mà R_k^2 lớn thì $1-R_k^2$ nhỏ. Vậy quy tắc là nếu độ chấp nhận của một biến nhỏ, thì nó gần như là một kết hợp tuyến tính của các biến độc lập khá, và đó là dấu hiệu của Đa cộng tuyến.
- Hệ số phóng đại phương sai (Variance inflation factor - VIF), có liên hệ gần với độ chấp nhận. Thực tế nó là nghịch đảo của độ chấp

nhận, tức là đối với biến Xk thì $VIF = \frac{1}{1-R_k^2}$. Khi Tolerance nhỏ thì VIF lớn, quy tắc là khi VIF vượt quá 10, đó là dấu hiệu của Đa cộng tuyến.

- Một biện pháp dò tìm bước đầu cũng khá có hiệu quả là xem xét các hệ số tương quan tuyến tính giữa các biến giải thích trong ma trận hệ số tương quan tuyến tính mà ta đã xây dựng từ đầu. Trong ma trận này bạn thấy biến *lit* có hệ số r lớn ở tất cả các biến, nó thể hiện một mối tương quan mạnh với 3 biến độc lập còn lại của mô hình.
- Ngoài ra một mô hình hồi qui tuyến tính có giá trị lớn của hệ số xác định R^2 mà có một vài hệ số hồi qui riêng lẻ không có ý nghĩa hoặc thậm chí có dấu ngược với lý thuyết hoặc cơ chế thông thường của mối quan hệ giữa biến độc lập và biến phụ thuộc cũng là một dấu hiệu chỉ báo có thể của hiện tượng cộng tuyến.

Ở phần Dò tìm các vi phạm giả định ở Hồi qui đơn chúng ta đã biết cách yêu cầu SPSS tính các đại lượng chuẩn đoán Đa cộng tuyến. Bạn thực hiện một mô hình hồi qui biến *ttxang* theo tất cả các biến độc lập của file *hoiquyboi* với lựa chọn Collinearity Diagnostics trong hộp thoại Statistics, Bảng 9.17 dưới đây là tóm tắt những thông số liên quan đến các hệ số hồi qui.

Bảng 9.17 **Coefficients(a)**

		(Constant)	cong suat dong co (HP)	trong luong xe (kg)	so may (cylinder)	dung tich dong co (lit)
Unstandardized Coefficients	B	26.234	-.046	-.009	.244	.178
	Std. Error	2.319	.016	.002	.536	.882
Standardized Coefficients	Beta		-.348	-.722	.100	.063
t		11.312	-2.783	-4.161	.455	.202
Sig.		.000	.008	.000	.651	.841
Collinearity Statistic	Tolerance		.318	.165	.102	.051
	VIF		3.141	6.067	9.763	19.748

a Dependent Variable: muc tieu thu xang (km/lit)

Độ chấp nhận (Tolerance) thấp và VIF rất lớn của biến *lit* khẳng định nó đã gây ra hiện tượng cộng tuyến với các biến độc lập khác, ma trận hệ số tương quan cũng khẳng định điều này.

2.2.6 Các thủ tục chọn biến

Chúng ta có thể xây dựng nhiều mô hình hồi qui từ cùng một tập biến. Ví dụ ta có thể xây dựng 7 phương trình từ 3 biến độc lập, gồm: 3 phương trình có một biến độc lập, 3 phương trình có hai biến, và 1 phương trình có 3 biến. Khi số biến tăng lên thì số mô hình có khả năng xây dựng được cũng tăng lên (với 10 biến độc lập sẽ có 1.023 khả năng).

Mặc dù có nhiều thủ tục tính toán tất cả các phương trình khả năng, nhưng có một số không được sử dụng thường xuyên. Ba thủ tục phổ biến mà ta sẽ nghiên cứu là: đưa vào dần (forward selection), loại trừ dần (backward elimination), và hồi qui từng bước (stepwise regression). Bạn nhớ rằng không có thủ tục chọn biến nào là "tốt nhất": chúng chỉ đơn giản là nhận ra các biến độc lập có khả năng dự đoán tốt cho biến phụ thuộc trong bộ dữ liệu mẫu. Mặt khác, bạn cũng cần chú ý không nên loại trừ các biến độc lập có tiềm năng thích đáng ra khỏi mô hình mặc dù các thủ tục này không nhận diện được nó căn cứ trên các điều kiện tiến hành thủ tục.

2.2.6.1 Phương pháp đưa vào dần (forward selection)

Trong phương pháp đưa vào dần, biến đầu tiên được xem xét để đưa vào phương trình là biến có tương quan thuận hay nghịch lớn nhất với biến phụ thuộc. Sau đó kiểm định F đối với giả thiết hệ số của biến được đưa vào bằng 0 sẽ được chương trình tính. Để xác định biến này (và mỗi biến tiếp theo) được đưa vào, giá trị thống kê F sẽ được so sánh với một giá trị chuẩn được thiết lập trước. Chúng ta có thể chỉ định một trong hai tiêu chuẩn trong SPSS.

- Tiêu chuẩn thứ nhất là giá trị nhỏ nhất của thống kê F mà một biến phải đạt được để được đưa vào, gọi là F vào (F-to-enter) ký hiệu trong chương trình là FIN, có giá trị mặc định là 3,84.
- Tiêu chuẩn thứ hai là xác suất tương ứng của thống kê F, gọi là xác suất F vào (probability of F-to-enter) ký hiệu trong chương trình là PIN, có giá trị mặc định là 0,05. Trong trường hợp này, một biến đi vào

phương trình chỉ khi nào xác suất tương ứng với giá trị thống kê F của nó tính ra nhỏ hơn hay bằng xác suất mặc định là 0,05 hay một giá trị mà ta đã chỉ định. SPSS mặc định dùng tiêu chuẩn PIN.

Hai tiêu chuẩn PIN và FIN không hẳn là tương đương nhau. Khi các biến được đưa vào phương trình, số bậc tự do tương ứng với Residual Sum of Square giảm trong khi số bậc tự do hồi qui tăng. Vì vậy, một giá trị F cố định có những mức ý nghĩa khác nhau tùy thuộc vào số lượng biến hiện đang ở trong phương trình. Đối với quy mô mẫu lớn, sự khác biệt này không đáng kể.

Nếu biến đầu tiên được chọn để đưa vào thỏa mãn được tiêu chuẩn vào, thì phương pháp đưa vào dần sẽ tiếp tục. Nếu không thủ tục này sẽ chấm dứt, không có biến nào được đưa vào phương trình, vì biến đầu tiên được chọn để thực hiện thủ tục đã là biến có tương quan chặt nhất với biến phụ thuộc rồi. Một khi có một biến được đưa vào thì ta sẽ xem xét tiếp các thông số thống kê của các biến không ở trong phương trình để chọn biến kế tiếp (bảng Exclude Varibles). Các tương quan riêng giữa biến phụ thuộc và từng biến độc lập không ở trong phương trình được điều chỉnh theo biến ở trong phương trình (tức là loại bỏ ảnh hưởng liên hệ tuyến tính của biến độc lập trong phương trình). Biến nào có tương quan riêng lớn nhất là ứng viên tiếp theo. Chọn biến có trị tuyệt đối của tương quan riêng lớn nhất là tương đương với việc chọn biến có giá trị F lớn nhất. Nếu tiêu chuẩn được thỏa mãn thì biến này được đưa vào phương trình và thủ tục tiếp tục lập lại. Thủ tục này dừng khi không còn biến nào khác thỏa mãn tiêu chuẩn vào nữa.

Trong bảng kết quả, SPSS thường thể hiện giá trị thống kê t và xác suất tương ứng. Những xác suất t này tương đương với các xác suất F. ta có thể tính F bằng cách bình phương giá trị t, bởi vì $t^2 = F$.

2.2.6.2 Phương pháp loại trừ dần (backward elimination)

Trong khi phương pháp đưa vào dần khởi đầu với phương trình không có biến nào cả và tuần tự đưa các biến vào theo tiêu chuẩn vào, thì phương pháp loại trừ dần khởi đầu với tất cả các biến đều ở trong phương trình và sau đó tuần tự loại trừ chúng bằng tiêu chuẩn loại trừ (removal criteria). Có hai tiêu chuẩn loại trừ (tiêu chuẩn ra) trong SPSS.

- Tiêu chuẩn thứ nhất là giá trị F tối thiểu biến mà thống kê F của biến độc lập đó phải đạt được để ở lại trong phương trình, gọi là F ra (F-to-remove), ký hiệu trong chương trình là FOUT. Các biến độc lập có giá trị F nhỏ hơn FOUT sẽ bị loại ra khỏi phương trình.
- Tiêu chuẩn thứ hai là xác suất tối đatương ứng với F ra (probability of F-to-remove) mà một biến không được vượt quá để được ở lại trong mô hình, ký hiệu trong chương trình là POUT.

Giá trị mặc định của FOUT là 2,71 và giá trị mặc định của POUT là 0,10. Tiêu chuẩn sử dụng mặc định là FOUT. Đầu tiên tất cả các biến độc lập đều được đưa vào mô hình, biến có hệ số tương quan từng phần nhỏ nhất sẽ được kiểm tra đầu tiên. Nếu giá trị thống kê F của biến đó nhỏ hơn FOUT thì nó sẽ bị loại khỏi mô hình (nhớ là có thể tính F bằng cách bình phương giá trị t, bởi vì $t^2 = F$), và phương trình sẽ được tính toán lại mà không có biến độc lập vừa loại. Trên các bảng giá trị thống kê mới SPSS sẽ lặp lại thủ tục trên, cho đến khi nào giá trị F của biến có hệ số tương quan từng phần nhỏ nhất lớn hơn FOUT thì quá trình này dừng lại. Bạn sẽ khai báo các tuỳ chọn FOUT, FIN, PIN... trong nút Option của hộp thoại Linear Regression của SPSS

2.2.6.3 Phương pháp chọn từng bước (stepwise selection)

Chọn biến độc lập từng bước thực ra là một kết hợp của thủ tục đưa vào dần và loại trừ dần và nó có lẽ là phương pháp được sử dụng thông thường nhất. Biến thứ nhất được chọn theo cách giống như chọn dần từng bước. Nếu biến này không thỏa điều kiện vào (FIN hoặc PIN) thì thủ tục này sẽ chấm dứt và không có biến độc lập nào trong phương trình. Nếu nó thỏa tiêu chuẩn, thì biến thứ hai được chọn căn cứ vào tương quan riêng cao nhất. Nếu biến thứ hai thỏa tiêu chuẩn vào nó cũng sẽ đi vào phương trình.

Sau khi biến thứ nhất được đưa vào, thủ tục chọn từng bước khác với đưa vào dần ở chỗ: biến thứ nhất được xem xét xem có nên loại bỏ nó ra khỏi phương trình căn cứ vào tiêu chuẩn ra (FOUT hặc POUT) giống như thủ tục loại trừ dần. Trong bước kế tiếp, các biến không ở trong phương trình được xem xét để đưa vào. Sau mỗi bước, các biến ở trong phương trình lại được xem xét để loại trừ ra. Các biến được loại trừ ra cho đến khi không còn biến nào thỏa điều kiện ra nữa.

Để ngăn chặn hiện tượng một biến bị đưa vào và loại ra lập đi lập lại, PIN phải nhỏ hơn POUT (hay FIN lớn hơn FOUT). Thủ tục chọn biến này sẽ chấm dứt khi không còn biến nào thỏa tiêu chuẩn vào và ra nữa.

Chú ý:

- Ba thủ tục vừa kể trên không phải lúc nào cũng cho ra cùng một phương trình. Mô hình được chọn của bất kỳ phương pháp nào cũng nên được nghiên cứu cẩn thận xem có vi phạm giả thiết nào không. Chúng ta nên xây dựng nhiều mô hình có thể chấp nhận được và sau đó chọn ra một mô hình trên cơ sở phù hợp thực tế nhất.
- Trong SPSS ta dùng t thay cho F vì hai kiểm định này tương đương.
- SPSS còn có phương pháp Enter, SPSS xử lý tất cả các biến bạn đưa vào một lần, đưa ra các thông số thống kê liên quan đến các biến, người xử lý sẽ có điều kiện tự mình đánh giá việc nên loại biến nào ra, đưa biến nào vào. Việc đưa vào và loại ra các biến đôi khi không nên tiến hành cứng nhắc theo các điều kiện mà còn phải căn cứ trên vai trò của biến độc lập đối với biến phụ thuộc căn cứ trên cơ chế hành vi tiềm ẩn giữa các biến, độ tin cậy của dữ liệu đã thu thập, khả năng giải thích, khả năng thu thập dữ liệu của biến dễ dàng. SPSS mặc định sử dụng phương pháp Enter, cũng theo mặc định thì các quan sát bị thiếu mất giá trị ở bất kỳ biến nào cũng sẽ không được sử dụng trong phân tích này.
- Sử dụng thủ tục chọn biến nào còn phụ thuộc vào tính chất của cuộc nghiên cứu – nghiên cứu của bạn là loại nghiên cứu diễn dịch hay quy nạp. Hay nói một cách khác là bạn đang muốn chứng minh tính đúng đắn của một mô hình lý thuyết trong một bối cảnh nghiên cứu cụ thể hay bạn đang muốn tìm ra một mô hình mới.
- Nên chú ý rằng một biến được đưa vào mô hình mà không có ý nghĩa thống kê trong việc giải thích biến thiên của biến phụ thuộc hay toàn bộ mô hình hồi qui không có ý nghĩa thống kê cũng là một kết quả của nghiên cứu.

Chúng ta áp dụng phương pháp chọn từng bước cho ví dụ trong file *hoiquyboi* của chúng ta. Kết quả cuối cùng của phương pháp chọn từng bước gồm các bảng từ 9.18 đến 9.21

Bảng 9.18 **Model Summary**

		Model	
		1	2
R		.858(a)	.878(b)
R Square		.736	.770
Adjusted R Square		.731	.761
Std. Error of the Estimate		1.86509	1.75942
Change Statistics	R Square Change	.736	.034
	F Change	134.095	6.939
	df1	1	1
	df2	48	47
	Sig. F Change	.000	.011

a Predictors: (Constant), trong luong xe (kg)

b Predictors: (Constant), trong luong xe (kg), cong suat dong co (HP)

Bảng 9.19 **ANOVA(c)**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	466.457	1	466.457	134.095	.000(a)
	Residual	166.971	48	3.479		
	Total	633.429	49			
2	Regression	487.937	2	243.968	78.812	.000(b)
	Residual	145.492	47	3.096		
	Total	633.429	49			

a Predictors: (Constant), trong luong xe (kg)

b Predictors: (Constant), trong luong xe (kg), cong suat dong co (HP)

c Dependent Variable: muc tieu thu xang (km/lit)

Bảng 9.20 **Coefficients(a)**

Model			(Constant)	trong luong xe (kg)	cong suat dong co (HP)
1	Unstandardized Coefficients	B	25.828	-.011	
		Std. Error	1.176	.001	
	Standardized Coefficients	Beta		-.858	
	t		21.965	-11.580	
	Sig.		.000	.000	
	Correlations	Zero-order		-.858	
		Partial		-.858	
		Part		-.858	

2	Unstandardized Coefficients	B	25.778	-.008	-.039
		Std. Error	1.109	.001	.015
	Standardized Coefficients	Beta		-.624	-.298
	t		23.235	-5.523	-2.634
	Sig.		.000	.000	.011
	Correlations	Zero-order		-.858	-.788
		Partial		-.627	-.359
		Part		-.386	-.184

a Dependent Variable: muc tieu thu xang (km/lit)

Bảng 9.21 **Excluded Variables(c)**

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	cong suat dong co (HP)	-.298(a)	-2.634	.011	-.359	.383
	so may (cylinder)	.021(a)	.165	.870	.024	.356
	dung tich dong co (lit)	-.023(a)	-.132	.896	-.019	.189
2	cong suat dong co (HP)					
	so may (cylinder)	.137(b)	1.106	.274	.161	.318
	dung tich dong co (lit)	.181(b)	1.025	.311	.149	.157

a Predictors in the Model: (Constant), trong luong xe (kg)

b Predictors in the Model: (Constant), trong luong xe (kg), cong suat dong co (HP)

c Dependent Variable: muc tieu thu xang (km/lit)

Giải thích phương trình

Phương trình hồi qui bội được phương pháp Stepwise ước lượng trên (Model 2) cho thấy trọng lượng xe và công suất động cơ là hai biến dự đoán tốt nhất cho mức tiêu thụ xăng. Trọng lượng xe càng nặng thì mức tiêu hao xăng càng lớn, công suất động cơ càng cao thì xe càng tiêu thụ nhiều xăng hơn (hệ số hồi qui âm chứng tỏ khi xe càng nặng và công suất càng cao thì số km xe chạy được trên 1 lít xăng càng giảm). Dung tích động cơ và số cylinder cũng có liên hệ với mức tiêu thụ xăng, nhưng khi trọng lượng và công suất được đưa vào phương trình thì ảnh hưởng của dung tích động cơ và số cylinder không còn nổi bật nữa. Và từ Bảng 9.20 ta có phương trình dự đoán mức tiêu thụ xăng là:

$$ttxang = 25,778 - 0,008\ nang - 0,039\ maluc$$

Trong đó:

- *ttxang* là mức tiêu thụ xăng tính bằng km/lít
- *nang* là trọng lượng xe tính bằng kg
- *maluc* là công suất động cơ tính bằng mã lực

2.3. Các lựa chọn trong hộp thoại Linear Regression của SPSS

Cách thức để thực hiện một phân tích hồi qui tuyến tính đã được thảo luận từ các phần trước, ở đây chúng ta sẽ hệ thống lại các lựa chọn có thể gặp trên hộp thoại Liner Regression ở Hình 9.21.

1. Dependent: là khung chứa biến phụ thuộc, bạn chỉ có thể đưa duy nhất một biến vào đây

2. Independent(s): khung chứa block (1 hoặc nhiều hơn 1) biến độc lập

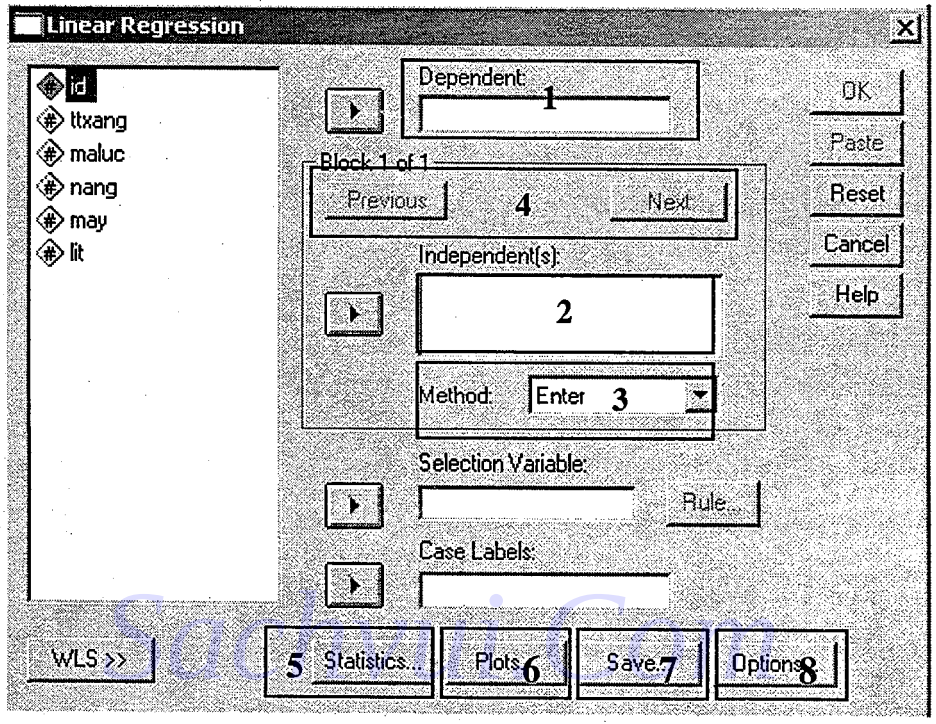
3. Nút Method (phương pháp): ở đây bạn có thể chọn các phương pháp khác nhau để xây dựng mô hình hồi qui. Chúng ta có thể chọn một trong năm phương pháp sau:

- **Enter** (đưa vào một lượt): đây là phương pháp mặc định của chương trình. Các biến trong khối sẽ được đưa vào mô hình cùng một lúc.
- **Remove** (loại bỏ một lượt): các biến trong khối sẽ được rút ra chỉ trong một bước.
- **Backward**: đã giải thích ở phần Các thủ tục chọn biến
- **Forward** đã giải thích ở phần Các thủ tục chọn biến
- **Stepwise** đã giải thích ở phần Các thủ tục chọn biến

Chú ý:

- Với Các thủ tục chọn biến kể trên trong SPSS, biến phải vượt qua được tiêu chuẩn chấp nhận (Tolerance criterion) mới được đưa vào phương trình, cho dù chúng ta có chỉ định phương pháp đưa biến vào nào đi nữa. Độ chấp nhận (Tolerance) mặc định là 0,0001. Một biến cũng không được đưa vào phương trình nếu nó làm cho độ chấp nhận của một biến đã được đưa vào mô hình xuống dưới tiêu chuẩn chấp nhận.
- Đối với phương pháp từng bước, số bước tối đa để đưa các biến vào phương trình là bằng hai lần số biến độc lập. Đối với phương pháp rút ra dần và loại bỏ dần thì số bước tối đa bằng với số biến thỏa mãn tiêu chuẩn đưa vào và rút ra.

Hình 9.21



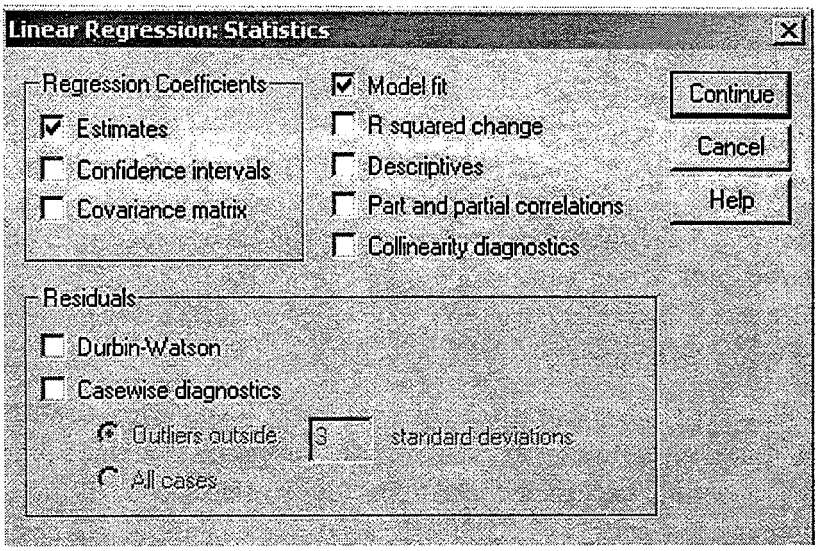
4. Chúng ta có thể chỉ định các phương pháp xử lý khác nhau cho các block khác nhau của các biến độc lập trong cùng một lần xử lý lệnh Linear Regression.Ví dụ, chúng ta có thể đưa một khối biến vào mô hình hồi qui bằng phương pháp Enter và đưa khối biến thứ hai vào mô hình bằng phương pháp từng bước. Để đưa khối biến thứ hai vào mô hình, chúng ta hãy nhấp chuột tại nút Next, sau khi đưa khối biến vào khung Independent(s) xong chúng ta hãy nhấp nút Method chọn một phương pháp đưa biến vào từng bước.

Để di chuyển tới lui giữa các khối biến độc lập, chúng ta sử dụng các nút Previous và Next. Chúng ta có thể chỉ định lên tới 9 khối biến khác nhau.

5. Chọn các thông số thống kê hồi qui tuyến tính

Để điều khiển thể hiện các kết quả tính ra, chúng ta hãy nhấp chuột vào nút Statistics... trong hộp thoại Linear Regression. Lệnh này sẽ mở ra hộp thoại Linear Regression: Statistics như trong hình sau:

Hình 9.22



Các tuỳ chọn về các thông số thống kê trên hộp thoại này

- **Estimates** (các ước lượng): cho hiện các hệ số hồi qui và các đo lường có liên quan, mục này được lựa chọn mặc định thể hiện trong bảng kết quả.
  - Đối với các biến được đưa vào phương trình, các thông số thống kê được thể hiện gồm: hệ số hồi qui *B*, sai số chuẩn của *B*, hệ số beta chuẩn hóa, giá trị thống kê t ứng với *B*, và mức ý nghĩa hai phía của t.
  - Đối với các biến không được đưa vào phương trình, các thông số thống kê được thể hiện gồm: hệ số beta nếu biến này được đưa vào phương trình, giá trị t ứng với beta, xác suất t, tương quan từng phần với biến phụ thuộc có kiểm soát các biến đã được đưa vào trong phương trình. và độ sai số tối thiểu.
- **Confidence interval** (khoảng tin cậy): cho thể hiện khoảng tin cậy 95% của từng hệ số hồi qui không chuẩn hóa.
- **Covariance matrix** (ma trận hiệp phương sai): thể hiện ma trận phương sai-hiệp phương sai của các hệ số hồi qui không chuẩn hóa. Các hiệp phương sai sẽ nằm bên dưới đường chéo, và các phương sai sẽ nằm trên đường chéo của ma trận.
- **Model fit** (các thống kê đánh giá sự phù hợp của mô hình): như R, R^2, R^2 điều chỉnh, và sai số chuẩn. Ngoài ra, bảng ANOVA sẽ thể hiện số

bậc tự do, tổng các độ lệch bình phương, độ lệch bình phương bình quân, giá trị thống kê F, và xác suất F quan sát được. Các thống kê đánh giá sự phù hợp của mô hình cũng được SPSS thể hiện theo mặc định.

- **Descriptives** (các thống kê mô tả): các trị trung bình, độ lệch chuẩn, và ma trận tương quan với các xác suất kiểm định một phía.
- **R squared change**: chúng ta đã thảo luận cách sử dụng tuỳ chọn này cho mục đích xem xét mức độ tăng của R square khi một biến giải thích được đưa vào phương trình trong khi phương trình đã chứa sẵn các biến độc lập khác
- **Collinearity diagnostics** (chuẩn đoán hiện tượng cộng tuyến): tuỳ chọn này sẽ thể hiện hệ số phóng đại phương sai (Variance inflation factor - VIF), các giá trị đặc trưng (eigenvalues) của ma trận tích mômen chéo, các chỉ số điều kiện, và các tỉ lệ của các bộ phận phương sai. Tùy chọn này cũng sẽ thể hiện độ chấp nhận của các biến trong phương trình, của các biến không được được đưa vào phương trình, độ chấp nhận của một biến nếu như nó được đưa vào trong phương trình ở bước tiếp theo.
- **Durbin-Watson**: thể hiện thống kê kiểm định Durbin-Watson; và cũng thể hiện các thống kê tóm tắt của các phần dư không chuẩn hóa và chuẩn hoá cũng như các giá trị dự đoán.

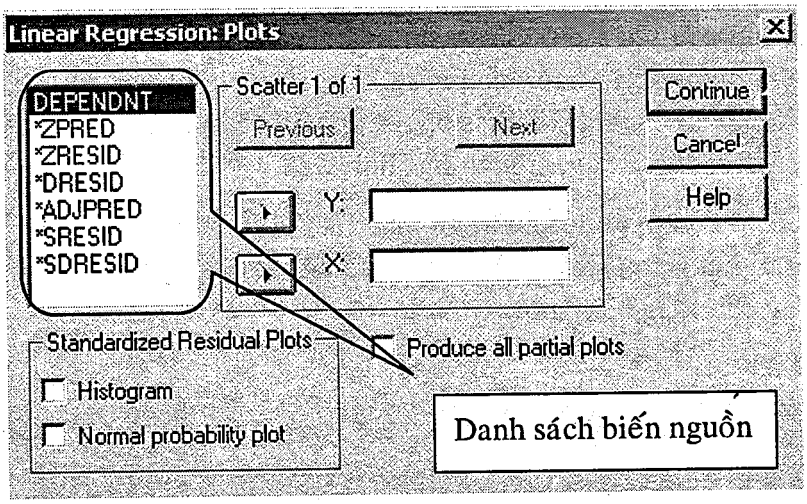
6. Vẽ đồ thị hồi qui tuyến tính

Để vẽ các dạng đồ thị liên quan đến mô hình hồi qui tuyến tính, bạn nhấp vào nút Plot... trong hộp thoại Linear Regression. Lệnh này sẽ mở ra hộp thoại Linear Regression Plots như trong Hình 9.22.

Biến phụ thuộc, các dạng của biến phần dư (residual) và các dạng của biến dự đoán sẽ xuất hiện trong danh sách biến nguồn ở góc trên bên trái, ý nghĩa của chúng là:

- **ZPRED* : giá trị dự đoán (hay lý thuyết) chuẩn hóa
- **ZRESID* : phần dư chuẩn hóa
- **DRESID* : phần dư loại bỏ quan sát đang xem xét.
- **ADJPRED* : giá trị dự đoán điều chỉnh
- **SRESID* : phần dư student hóa
- **SDRESID* : phần dư loại bỏ quan sát đang xem xét được student hóa

Hình 9.22



Bạn hãy chọn một biến trong danh sách biến nguồn cho trục tung đưa vào khung Y và một biến cho trục hoành đưa vào khung X. Để thực hiện thêm nhiều đồ thị như vậy một lúc, hãy nhấp chuột tại nút Next và lập lại việc xác định biến cho trục tung và biến cho trục hoành. Chúng ta có thể chỉ định được đến 9 đồ thị cùng một lúc. Tất cả các đồ thị đều được chuẩn hóa.

Phần dưới của hộp thoại còn có các lựa chọn sau

Standardized Residual Plots (Đồ thị phần dư chuẩn hóa): chúng ta có thể chọn một trong các loại sau:

- **Histogram** (biểu đồ tần số): thể hiện biểu đồ tần số của các phần dư chuẩn hóa, và đường cong chuẩn mặc định được đặt chồng lên trong biểu đồ này.
- **Normal probability plot** (biểu đồ xác suất chuẩn): đồ thị so sánh xác suất chuẩn của các phần dư chuẩn hóa với một phân phối chuẩn.

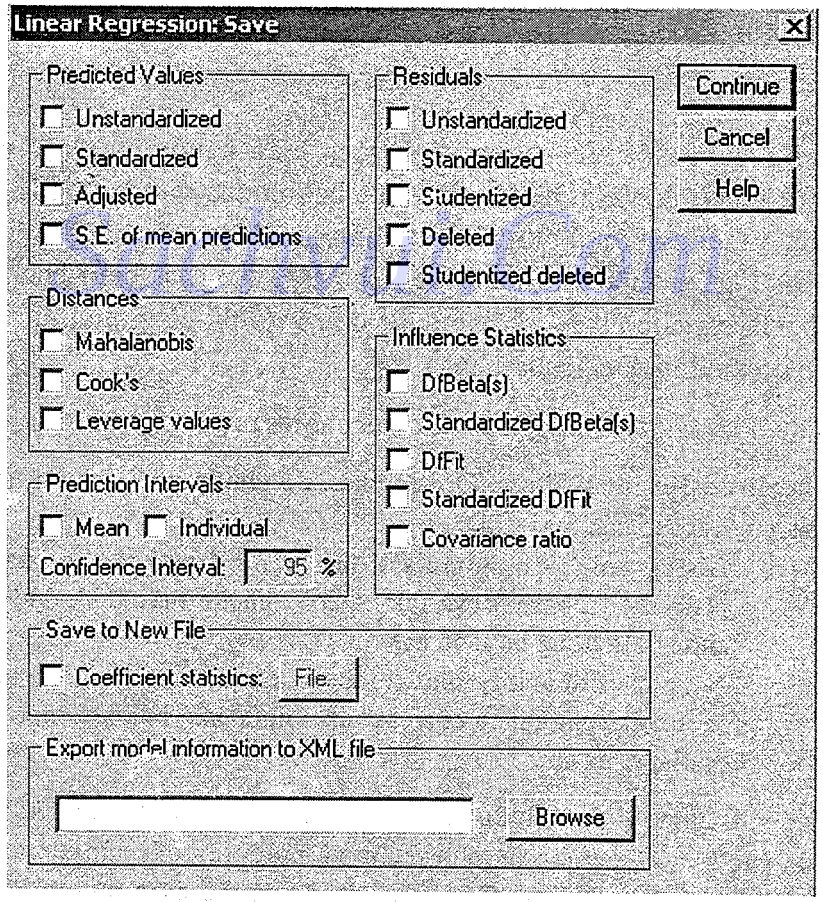
Produce all partial plots (biểu đồ phân tán từng phần): vẽ biểu đồ phân tán của các phần dư của biến phụ thuộc và một biến độc lập khi cả hai biến này được hồi qui riêng. Các biểu đồ này sẽ thể hiện từng biến độc lập trong phương trình theo thứ tự giảm dần của sai số chuẩn của hệ số hồi qui. tất cả các biểu đồ này được chuẩn hóa. Phải có ít nhất hai biến độc lập trong phương trình thì chúng ta mới có thể thực hiện biểu đồ từng phần này được.

7. Sao lưu các biến mới trong phân tích hồi qui tuyến tính

Để sao lưu các phần dư, giá trị dự đoán, hoặc các thông số có liên quan như những biến mới, chúng ta hãy nhấp chuột vào nút Save... trong hộp thoại Linear Regression. Lệnh này sẽ mở ra một hộp thoại Linear Regression: Save như trong Hình 9.23.

SPSS sẽ tự động gán tên biến mới cho bất cứ thông số nào chúng ta muốn sao lưu. Trong kết quả chạy ra sẽ có một bảng cho biết tên và nội dung của từng biến mới.

Hình 9.23



Trong hộp thoại [Linear Regression: Save] có các nội dung sau

Predicted values (các giá trị dự đoán): chúng ta có thể chọn một hay nhiều các thông số sau:

- Unstandardized: các giá trị dự đoán không chuẩn hóa.

- Standardized: các giá trị dự đoán chuẩn hóa.
- Adjusted: các giá trị dự đoán điều chỉnh.
- S.E. of mean predictions: sai số chuẩn của các giá trị dự đoán

Distances (khoảng cách): các điểm quan sát có trị số của biến độc lập rất lớn hay rất nhỏ có thể có ảnh hưởng mạnh đến kết quả phân tích nên chúng ta cần nhận diện chúng. Khoảng cách từ trị số quan sát đến trị trung bình của biến độc lập có thể đo bằng các loại khoảng cách sau:

- Mahalanobis
- Cook
- Leverage values

Prediction intervals (khoảng dự đoán): chúng ta có thể chọn một hay cả hai loại dự đoán sau:

- Mean: giới hạn trên và dưới của khoảng dự đoán của trị trung bình.
- Individual: giới hạn trên và dưới của khoảng dự đoán cho từng quan sát.
- Ngoài ra ta có thể chọn khoảng tin cậy mặc định cho trị trung bình và từng quan sát là 95%. Chúng ta có thể thay đổi mặc định này bằng cách gõ vào khung Confidence interval một giá trị lớn hơn 0 và nhỏ hơn 100. Ví dụ chúng ta muốn khoảng tin cậy là 99% thì chúng ta gõ vào 99.

Residuals (phần dư): chúng ta có thể chọn một hay nhiều loại sau:

- Unstandardized: phần dư không chuẩn hóa.
- Standardized: phần dư chuẩn hóa.
- Studentized: phần dư student hóa.
- Deleted: phần dư loại bỏ quan sát đang xem xét
- Studentized deleted: phần dư loại bỏ quan sát đang xem xét được student hóa

Influence Statistics (các thông số thống kê ảnh hưởng): chúng ta có thể chọn một hay nhiều thông số sau:

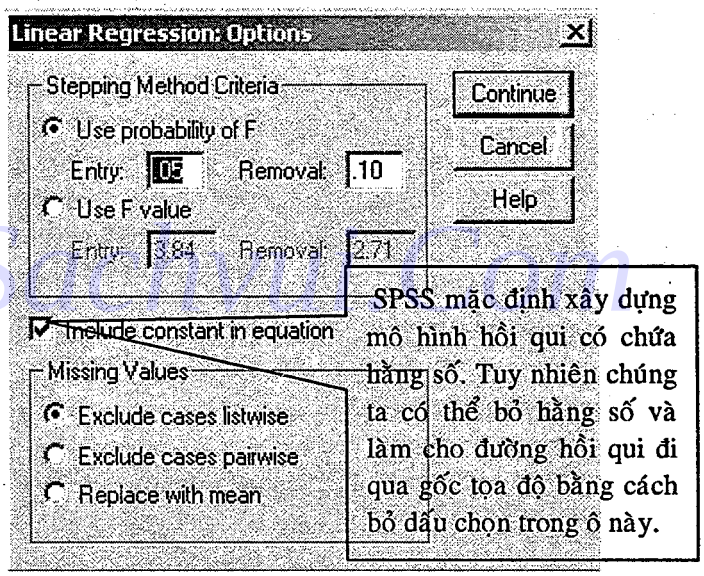
- DfBeta(s): phản ánh sự thay đổi của hệ số hồi qui khi loại bỏ một quan sát nào đó ra khỏi quá trình tính toán. SPSS sẽ tính lại tất cả các hệ số hồi qui trong phương trình, kể cả hằng số.
- Standardized DfBeta(s): tính các DfBeta chuẩn hóa.
- DfFit: phản ánh sự thay đổi của các giá trị dự đoán khi loại bỏ một quan sát đang xem xét ra khỏi quá trình tính toán.
- Standardized DfFit: tính các DfFit chuẩn hóa.
- Covariance ratio: tính tỉ số giữa các thành phần của ma trận phương sai-hiệp phương sai khi quan sát đang xem xét bị loại ra khỏi qua trình tính toán với các thành phần của ma trận phương sai-hiệp phương sai khi tất cả các quan sát được đưa vào tính toán. Nếu tỉ số này tiến đến 1 thì

quan sát đang xem xét không ảnh hưởng nhiều đến ma trận hiệp phương sai.

8. Các tùy chọn trong hồi qui tuyến tính

Để điều chỉnh các tiêu chuẩn biến vào hoặc ra khỏi mô hình hồi qui, hay điều chỉnh việc xử lý các quan sát thiếu dữ kiện, chúng ta hãy nhấp chuột tại nút Options... trong hộp thoại Linear Regression. Lệnh này sẽ mở ra hộp thoại Linear Regression: Options như trong hình Hình 9.24.

Hình 9.24



Các lựa chọn bạn có thể thực hiện trong hộp thoại này bao gồm:

Stepping Method Criteria: tiêu chuẩn vào hay ra áp dụng cho các phương pháp đưa biến vào dần, loại trừ dần và từng bước. Chúng ta có thể chọn :

- Use probability of F: sử dụng xác suất F vào (PIN) và xác suất F ra (POUT) làm tiêu chuẩn đưa biến vào và ra. Đây là tiêu chuẩn mặc định. Giá trị vào mặc định là 0,05 và giá trị ra mặc định là 0,10. Để thay đổi các giá trị mặc định này, chúng ta hãy nhập các xác suất F mới vào. Cả hai giá trị phải lớn hơn 0 và nhỏ hơn hay bằng 1, và PIN phải nhỏ hơn POUT.
- Use F value: sử dụng các giá trị F làm tiêu chuẩn vào và ra. Giá trị vào mặc định (FIN) là 3,84. Giá trị ra mặc định (FOUT) là 2,71. Để

thay đổi các giá trị mặc định này, chúng ta hãy nhập các giá trị F mới vào. Cả hai giá trị phải lớn hơn 0 và giá trị FIN phải lớn hơn giá trị FOUT.

Missing values: chúng ta có thể chọn một trong ba cách xử lý các quan sát thiếu dữ kiện sau:

- Exclude cases listwise: chỉ có những quan sát nào có đầy đủ giá trị của tất cả các biến mới được đưa vào phân tích. Đây là cách xử lý mặc định.
- Exclude cases pairwise: các quan sát có đầy đủ dữ liệu đối với một cặp biến đang nghiên cứu liên hệ sẽ được sử dụng để tính hệ số tương quan làm cơ sở cho phân tích hồi qui. Bậc tự do được tính trên N nhỏ nhất của các cặp biến.
- Replace with mean: thay thế các dữ kiện thiếu bằng trung bình của biến và tất cả các quan sát đều được sử dụng để tính toán.

3. HỒI QUI VỚI QUAN HỆ PHI TUYẾN

Bên cạnh hồi qui tuyến tính cũng có một số trường hợp mà ở đó mối liên hệ giữa hai biến X và Y là dạng đường cong, thay vì là quan hệ tuyến tính. Ví dụ, nhu cầu sử dụng điện tăng theo số mũ và tỉ lệ này thuận chiều với sự phát triển dân số ở một số khu vực hay các nhà quảng cáo tin rằng quan hệ lợi tức giảm dần sẽ xảy ra giữa doanh số bán hàng và quảng cáo nếu để quảng cáo phát triển quá.

Chúng ta cũng nhanh chóng nhận thấy rằng các mô hình quan hệ phi tuyến tính sẽ phức tạp hơn các mô hình chỉ thể hiện quan hệ tuyến tính giữa các biến. Mặc dù các mô hình phức tạp đôi khi cũng cần thiết, những người sử dụng cũng nên thận trọng khi dùng nó vì nhiều lí do. Trước tiên, rõ ràng người ta sử dụng những phương tiện trợ giúp đưa ra quyết định mà họ nắm rõ và không muốn sử dụng những phương tiện mà họ không hiểu, bởi vậy, một mô hình càng phức tạp thì càng ít được sử dụng. Thứ hai, cách tính "vừa đủ" trong khoa học cho thấy rằng sử dụng mô hình đơn giản nhất có thể thường phù hợp với số liệu, vì các mô hình phức tạp thông thường không thể hiện các hiện tượng tiềm ẩn trong số liệu ngay từ ban đầu.

Để minh họa cho nội dung hồi qui phi tuyến chúng ta sử dụng File ví dụ *World95.sav* được cung cấp sẵn trong đĩa phần mềm SPSS để Nghiên cứu các yếu tố ảnh hưởng đến tuổi thọ trung bình của phụ nữ tại các quốc gia trên thế giới. Vì chúng tôi đã chỉnh sửa file này theo kiểu giải nghĩa một số biến có tên bằng tiếng Anh nên chúng ta đặt lại tên là *World95_Viet.sav*

Đầu tiên chúng ta liệt kê một số biến được cho là có ảnh hưởng đến tuổi thọ của phụ nữ (biến này được đặt tên là Lifeexpf) :

- Tỉ lệ dân sống ở đô thị (biến này được đặt tên là Urban)
- Mật độ dân số (biến này được đặt tên là Density)
- Tỉ lệ dân biết chữ (biến này được đặt tên là Literacy)
- GDP bình quân đầu người (biến này được đặt tên là Gdp_cap)
- Lượng calori một người tiêu thụ mỗi ngày (biến này được đặt tên là Calories)
- Số con trungbình của một phụ nữ (biến này được đặt tên là Fertility)
- Tỉ lệ phụ nữ biết chữ (biến này được đặt tên là Lit_fema)

Bạn đọc có thể tìm thấy các biến này với tên gọi tương ứng trong file *World95_Viet.sav.*

Kế đến chúng ta xem xét mối quan hệ tương quan đơn tuyến tính giữa các biến độc lập liệt kê trên đây với biến tuổi thọ trung bình của phụ nữ (Lifeexpf) và quan hệ tương quan giữa chính các biến độc lập đó với nhau để loại bỏ những biến không có ý nghĩa trong việc giải thích, đánh giá hoặc tính toán tuổi thọ trung bình của phụ nữ, hoặc loại bỏ các biến có quan hệ quá chặt chẽ với nhau khiến dẫn đến hiện tượng cộng tuyến trong mô hình. Ma trận hệ số tương quan đơn Person của phân tích trên đước trình bày như sau:

Bảng 9.22

	Mật độ dân số (người/ km2)	Tỉ lệ dân sống ở vùng đô thị (%)	Tỉ lệ dân biết chữ (%)	GDP tính trên đầu người (USD)	Calori nạp hàng ngày TB 1 người	Số con TB của 1 phụ nữ	Tỉ lệ nữ giới biết chữ (%)
Mật độ dân số (người/km2)	1	,223	,031	,201	,067	-,162	,029
Tỉ lệ dân sống ở vùng đô thị (%)	,223	1	,650	,605	,692	-,619	,612
Tuổi thọ TB phụ nữ	**,128**	**,743**	**,865**	**,642**	**,775**	**-,838**	**,819**
Tỉ lệ dân biết chữ (%)	,031	,650	1	,552	,682	-,866	,973
GDP tính trên đầu người (USD)	,201	,605	,552	1	,751	-,583	,429
Calori nạp hàng ngày TB 1 người	,067	,692	,682	,751	1	-,696	,548
Số con TB của 1 phụ nữ	-,162	-,619	-,866	-,583	-,696	1	-,839
Tỉ lệ nữ giới biết chữ (%)	,029	,612	,973	,429	,548	-,839	1

Ma trận hệ số tương quan cho thấy các biến giải thích được liệt kê trên đều có quan hệ tương quan chặt chẽ với Tuổi thọ trung bình của người phụ nữ. Ngoại trừ trường hợp mật độ dân cư, nệ số tương quan giữa mật độ dân cư và tuổi thọ của phụ nữ rất thấp (0,128), như vậy chúng có khả năng giải thích cho tuổi thọ TB của phụ nữ nếu được đưa vào một mô hình hồi qui phù hợp.

Bên cạnh đó bạn đọc có thể chạy ra bảng đầy đủ các giá trị p-value của kiểm định t (2- đuôi) đối với các hệ số tương quan đơn của tổng thể giữa biến Lifeexpf và các biến giải thích được liệt kê, các số này đều rất nhỏ nên ta có thể kết luận các hệ số tương quan tổng thể đều có ý nghĩa thống kê ở mức 0,05 hoặc 0,01; ngoại trừ biến Mật độ dân cư Density có p-value rất cao khiến ta không thể bác bỏ giả thuyết hệ số tương quan tổng thể giữa biến Mật độ dân cư Density và Tuổi thọ trung bình của người phụ nữ bằng 0, do vậy ta hoàn toàn có thể loại bỏ biến Density khỏi mô hình hồi qui.

Cũng trên ma trận này ta thấy mối quan hệ tương quan tuyến tính giữa 2 biến Tỉ lệ dân biết đọc và Tỉ lệ phụ nữ biết đọc khá mạnh, điều này có thể giải thích được là do Tỉ lệ dân biết đọc đã bao hàm trong nó Tỉ lệ phụ nữ biết đọc do vậy ta xem xét và loại bỏ bớt một biến khỏi mô hình hồi qui dự kiến. Vậy ta loại biến nào ?

Hệ số tương quan tuyến tính giữa Tỉ lệ dân biết đọc với Tuổi thọ TB của phụ nữ lớn hơn hệ số giữa Tỉ lệ phụ nữ biết đọc và Tuổi thọ TB của phụ nữ (0,865 > 0,819), như vậy khả năng giải thích của yếu tố Tỉ lệ dân biết đọc đối với Tuổi thọ TB của phụ nữ có vẻ mạnh hơn Tỉ lệ phụ nữ biết đọc. Như vậy nên duy trì biến Tỉ lệ dân biết đọc và loại bớt biến Tỉ lệ phụ nữ biết đọc.

Hệ số tương quan tuyến tính giữa biến Gdp_cap và biến Calories cũng khá cao (tới 0,751) thể hiện một quan hệ tương quan tuyến tính chặt chẽ giữa hai yếu tố này, rõ ràng gdp_cap cao không chỉ quyết định đến chất lượng Calories mà còn ảnh hưởng đến nhiều vấn đề khác có liên quan trực tiếp đến tuổi thọ của con người, đặc biệt là phụ nữ, ví dụ mức độ được chăm sóc y tế, chất lượng cuộc sống tinh thần, khả năng tự bảo vệ mình... tuy nhiên trong quá trình khảo sát để tìm mô hình phù hợp ta nên đưa vào mô hình cả hai biến để giải thích đầy đủ hơn vấn đề ta quan tâm, sau đó sẽ căn cứ vào các tiêu chuẩn kiểm định thực tế của mô hình mà quyết định có loại bỏ biến Calories hay không.

Kết luận

Như vậy sau bước khảo sát này ta giữ lại 5 biến sau để xem xét

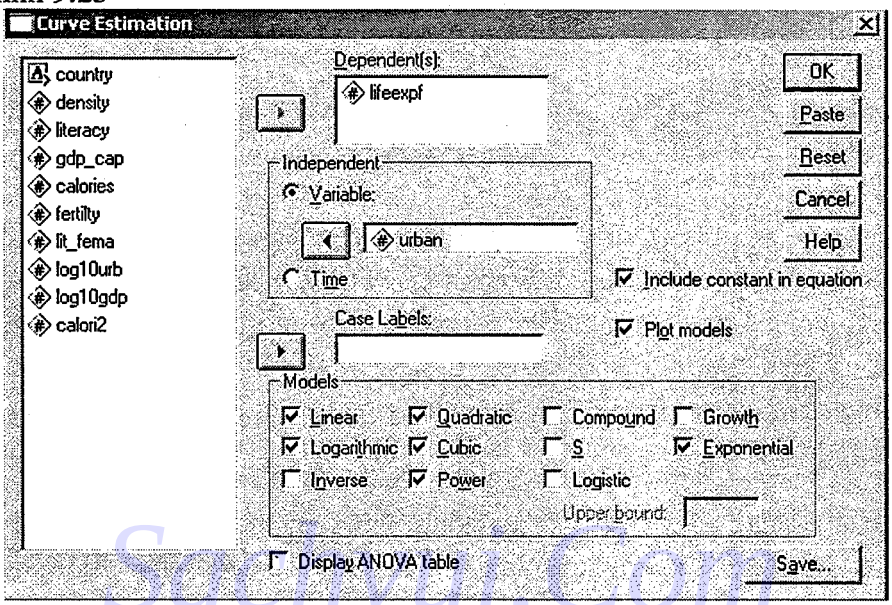
- Tỉ lệ dân sống ở đô thị (biến này đặt tên là Urban)
- Tỉ lệ dân biết chữ (biến này đặt tên là Literacy)
- GDP bình quân đầu người (biến này đặt tên là Gdp_cap)
- Lượng calori một người tiêu thụ mỗi ngày (biến này đặt tên là Calories)
- Số con trungbình của một phụ nữ (biến này đặt tên là Fertility)

Ta tiến hành khảo sát riêng dạng hàm hồi qui phù hợp của từng biến độc lập đối với biến phụ thuộc là Tuổi thọ TB của phụ nữ và

1. Biến Urban

Để khảo sát các dạng hàm hồi qui có thể giữa Urban và Lifeexpf ta có thể tiến hành theo hướng dẫn của hình vẽ sau sẽ được các kết quả về hệ số phù hợp R^2, giá trị p-value của phép kiểm định độ phù hợp của toàn bộ mô hình của từng mô hình như sau

Hình 9.25



Kết quả xuất hiện một loạt như sau

```
Independent:  URBAN
   Dependent Mth    Rsq  d.f.        F  Sigf
   LIFEEXPF LIN  ,553   106   131,00  ,000
   LIFEEXPF LOG  ,563   106   136,70  ,000
   LIFEEXPF QUA  ,586   105    74,43  ,000
   LIFEEXPF CUB  ,587   104    49,36  ,000
   LIFEEXPF POW  ,559   106   134,17  ,000
   LIFEEXPF EXP  ,533   106   121,12  ,000
```

Ta thấy giá trị R^2 của 3 mô hình bậc hai, bậc ba và mô hình log là cao nhất thể hiện khả năng giải thích của ba dạng mô hình này mạnh nhất cho mối quan hệ Urban và Lifeexp (trong đó giá trị R^2 của mô hình bậc hai và bậc ba gần như bằng nhau) nhưng để đơn giản hoá mô hình và tránh hiện tượng cộng tuyến có thể có giữa các biến urban bậc một, hai và ba, ta chọn dạng mô hình log-tuyến tính cho quan hệ giữa urban và Lifeexp với log cơ số 10 của biến Urban.

Thực hiện tương tự với các biến còn lại, lần lượt được các kết quả sau.

2. Biến Literacy

Khảo sát các dạng hàm hồi qui có thể giữa Literacy và Lifeexpf cho ta các kết quả về hệ số phù hợp R^2, giá trị p-value của phép kiểm định độ phù hợp của toàn bộ mô hình của từng mô hình như sau:

Dependent	Mth	Rsq	d.f.	F	Sigf
LIFEEXPF	**LIN**	**.749**	105	313.26	.000
LIFEEXPF	LOG	.699	105	244.39	.000
LIFEEXPF	QUA	.750	104	156.40	.000
LIFEEXPF	CUB	.756	103	106.13	.000
LIFEEXPF	POW	.685	105	228.59	.000
LIFEEXPF	EXP	.724	105	275.74	.000

Ta thấy giá trị R^2 của ba mô hình: tuyến tính, bậc hai và bậc ba không chênh lệch nhiều; bên cạnh đó kết quả khảo sát phân phối của phần dư của ba mô hình trên không cho thấy mô hình nào tạo ra phần dư có chất lượng tốt nổi trội nên để đơn giản hoá mô hình ta chọn dạng mô hình tuyến tính cho quan hệ giữa Literacy và Lifeexp.

3. Biến Gdp_cap

Khảo sát các dạng hàm hồi qui có thể giữa Gdp_cap và lifeexpf ta có các kết quả về hệ số phù hợp R^2, giá trị p-value của phép kiểm định độ phù hợp của toàn bộ mô hình của từng mô hình như sau

Dependent	Mth	Rsq	d.f.	F	Sigf
LIFEEXPF	LIN	.412	107	75.11	.000
LIFEEXPF	**LOG**	**.691**	107	238.93	.000
LIFEEXPF	QUA	.544	106	63.35	.000
LIFEEXPF	CUB	.604	105	53.32	.000
LIFEEXPF	POW	.652	107	200.32	.000
LIFEEXPF	EXP	.364	107	61.26	.000

Giá trị R^2 của mô hình Logarithmic lớn vượt trội so với các dạng mô hình khác (69.1%) ta chọn dạng mô hình log-tuyến tính với log cơ số 10 cho biến Gdp_ca, như vậy ta chuyển biến gdp_cap thành log(gdp_cap)

4. Biến Calories

Khảo sát các dạng hàm hồi qui có thể giữa Calories và lifeexpf ta có các kết quả về hệ số phù hợp R^2, giá trị p-value của phép kiểm định độ phù hợp của toàn bộ mô hình của từng mô hình như sau:

Dependent	Mth	Rsq	d.f.	F	Sigf
LIFEEXPF	LIN	.601	73	110.05	.000
LIFEEXPF	LOG	.631	73	125.07	.000
LIFEEXPF	**QUA**	**.667**	72	72.00	.000
LIFEEXPF	CUB	.668	72	72.35	.000
LIFEEXPF	POW	.602	73	110.49	.000
LIFEEXPF	EXP	.569	73	96.56	.000

Căn cứ trên R^2 ta chọn dạng mô hình bậc hai cho quan hệ giữa Calori với Lifeexp.

Như vậy ta cần tạo ra thêm một biến Calori ^{2} (Calori bình phương) để sử dụng cùng với biến Calori.

5. Biến Fertility

Khảo sát các dạng hàm hồi qui có thể giữa Fertility và Lifeexpf ta có các kết quả về hệ số phù hợp R^2, giá trị p-value của phép kiểm định độ phù hợp của toàn bộ mô hình của từng mô hình như sau

Dependent	Mth	Rsq	d.f.	F	Sigf
LIFEEXPF	**LIN**	**.702**	105	246.79	.000
LIFEEXPF	LOG	.667	105	210.11	.000
LIFEEXPF	QUA	.702	104	122.47	.000
LIFEEXPF	CUB	.702	103	80.90	.000
LIFEEXPF	POW	.629	105	177.64	.000
LIFEEXPF	EXP	.674	105	217.31	.000

Giá trị R^2 của mô hình tuyến tính lớn vượt trội nên ta chọn dạng mô hình tuyến tính cho quan hệ giữa Fertility và lifeexpf là phù hợp.

Kết luận

Ta xây dựng một mô hình hồi qui bội với các biến có khả năng giải thích cho tuổi thọ trung bình của phụ nữ (lifeexpf)như sau :

- Log(urrban)
- Literacy
- Log(gdp_cap)
- Calori
- Calori ^{2}
- Fertility

Như vậy mô hình hồi qui của chúng ta vừa có các biến tuyến tính vừa có các biến phi tuyến, ta thực hiện các khảo sát trên các mô hình có thể kết hợp các biến trên để tìm ra mô hình phù hợp nhất.

Đưa toàn bộ 6 biến vừa xây dựng trên vào một mô hình tổng quát nhất (ta dùng phương pháp Enter cho mọi bước khảo sát để tự nhận xét và chọn lựa các biến).

Bảng 9.23 **Variables Entered/Removed(b)**

Model	Variables Entered	Method
1	CALORI2, LOG10URB, Số con TB của 1 phụ nữ, Tỉ lệ dân biết chữ (%), LOG10GDP, Calori nạp hàng ngày TB 1 người(a)	Enter

a All requested variables entered.
b Dependent Variable: Tuổi thọ TB phụ nữ

Bảng 9.24 **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,935(a)	,873	,862	4,253

a Predictors: (Constant), CALORI2, LOG10URB, Số con TB của 1 phụ nữ, Tỉ lệ dân biết chữ (%), LOG10GDP, Calori nạp hàng ngày TB 1 người

Bảng 9.25 **ANOVA(b)**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8355,495	6	1392,582	76,985	,000(a)
	Residual	1211,965	67	18,089		
	Total	9567,459	73			

a Predictors: (Constant), CALORI2, LOG10URB, Số con TB của 1 phụ nữ, Tỉ lệ dân biết chữ (%), LOG10GDP, Calori nạp hàng ngày TB 1 người
b Dependent Variable: Tuổi thọ TB phụ nữ

Bảng 9.26

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3,284	15,515		,212	,833		
	Tỉ lệ dân biết chữ (%)	,149	,047	,301	3,163	,002	,208	4,804
	Calori nạp hàng ngày TB 1 người	,026	,012	1,282	2,191	,033	,005	182,643
	Số con TB của 1 phụ nữ	-1,455	,531	-,247	-2,742	,008	,233	4,294
	LOG10URB	5,472	3,017	,136	1,814	,074	,338	2,960
	LOG10GDP	3,489	1,700	,203	2,052	,044	,193	5,169
	CALORI2	-4,17E-06	,000	-1,150	-2,018	,048	,006	171,675

a. Dependent Variable: Tuổi thọ TB phụ nữ

Xem xét thông tin sơ bộ trên mô hình đầu tiên ta thấy có xảy ra hiện tượng cộng tuyến ở hai biến calori và calori^2 nên ta loại bỏ biến calori và chạy lại được mô hình thứ hai với các thông tin sau đây

Bảng 9.27 **Variables Entered/Removed(b)**

Model	Variables Entered	Method
1	CALORI2, LOG10URB, Số con TB của 1 phụ nữ, Tỉ lệ dân biết chữ (%), LOG10GDP(a)	Enter

a All requested variables entered.

b Dependent Variable: Tuổi thọ TB phụ nữ

Bảng 9.28 **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,930(a)	,864	,854	4,369

a Predictors: (Constant), CALORI2, LOG10URB, Số con TB của 1 phụ nữ, Tỉ lệ dân biết chữ (%), LOG10GDP

Bảng 9.29 **ANOVA(b)**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8269,435	5	1653,887	86,643	,000(a)
	Residual	1298,025	68	19,089		
	Total	9567,459	73			

a Predictors: (Constant), CALORI2, LOG10URB, Số con TB của 1 phụ nữ, Tỉ lệ dân biết chữ (%), LOG10GDP

b Dependent Variable: Tuổi thọ TB phụ nữ

Bảng 9.30

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	34,348	6,324		5,432	,000		
	Tỉ lệ dân biết chữ (%)	,169	,048	,342	3,559	,001	,216	4,623
	Số con TB của 1 phụ nữ	-1,371	,544	-,233	-2,522	,014	,234	4,271
	LOG10URB	8,314	2,795	,206	2,975	,004	,415	2,408
	LOG10GDP	3,183	1,740	,185	1,829	,072	,195	5,134
	CALORI2	2,899E-07	,000	,080	,948	,347	,281	3,560

a. Dependent Variable: Tuổi thọ TB phụ nữ

Mô hình mới có hiện tượng đa cộng tuyến không còn rõ rệt (VIF nhỏ) nhưng hệ số hồi qui đứng trước Calori2 không có ý nghĩa, hiện tượng này nhắc cho chúng ta nhớ đến mối quan hệ cộng tuyến tiềm ẩn giữa hai biến calories và gdp_gap mà ta đã khảo sát. Như vậy ta quyết định loại bỏ luôn biến $Calori^2$ khỏi mô hình.

Chạy lại mô hình với 4 biến chính là log(urrban), literacy, log(gdp_cap) và fertility. Kết quả thu được như sau:

Bảng 9.31 **Variables Entered/Removed(b)**

Model	Variables Entered	Method
1	LOG10GDP, Số con TB của 1 phụ nữ, LOG10URB, Tỉ lệ dân biết chữ (%)(a)	Enter

a All requested variables entered.

b Dependent Variable: Tuổi thọ TB phụ nữ

Bảng 9.32 **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,933(a)	,870	,865	3,929

a Predictors: (Constant), LOG10GDP, Số con TB của 1 phụ nữ, LOG10URB, Tỉ lệ dân biết chữ (%)

Bảng 9.33 **ANOVA(b)**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10351,984	4	2587,996	167,652	,000(a)
	Residual	1543,673	100	15,437		
	Total	11895,657	104			

a Predictors: (Constant), LOG10GDP, Số con TB của 1 phụ nữ, LOG10URB, Tỉ lệ dân biết chữ (%)

b Dependent Variable: Tuổi thọ TB phụ nữ

Bảng 9.34

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	33,767	5,031		6,712	,000		
	Tỉ lệ dân biết chữ (%)	,143	,036	,308	3,996	,000	,218	4,585
	Số con TB của 1 phụ nữ	-1,459	,419	-,258	-3,479	,001	,236	4,236
	LOG10URB	8,772	2,337	,213	3,753	,000	,403	2,481
	LOG10GDP	4,486	1,079	,263	4,159	,000	,324	3,084

a. Dependent Variable: Tuổi thọ TB phụ nữ

Rõ ràng giá trị R^2 điều chỉnh lúc này đã tăng trở lại và còn khả quan hơn so với mô hình đầu tiên, bạn thử chạy kiểm định Durbin – Watson cũng sẽ khẳng định không có tự tương quan.

Giá trị Sig trên bảng ANOVA cho phép ta bác bỏ giả thuyết H_0: mô hình hồi qui xây dựng được không phù hợp với tổng thể. Như vậy toàn bộ mô hình hồi qui có thể nghĩa giải thích cho biến động của lieexpf.

Các giá trị Sig. và VIF của các biến của mô hình cuối cùng đều cho ta những thông tin rất tốt. Mô hình không còn hiện tượng cộng tuyến, hệ số hồi qui của các biến riêng lẻ đều có ý nghĩa với độ tin cậy trên 95%. Ta có thể chấp nhận sử dụng mô hình này cho việc dự báo tuổi thọ TB của phụ nữ, đánh giá các yếu tố tác động đến tuổi thọ của nữ để đề ra các chính sách vĩ mô phù hợp. Mô hình tổng quát có dạng sau

$$\widehat{Lifexpf_i} = 33,767 + 8,772 logUrban_i + 0,143 literacy_i + 4,486 logGdp\_cap_i - 1,459 fertility_i + e_i$$

4. HỒI QUI VỚI BIẾN ĐỘC LẬP ĐỊNH TÍNH (BIẾN GIẢ)

Bạn đã quen với việc sử dụng các biến độc lập là các biến định lượng. Tuy nhiên, bạn cũng sẽ gặp phải nhiều tình huống mà trong đó bạn cần sử dụng các biến độc lập định tính.

Nếu một biến là định danh, và các mã số được code cho các phân loại, bạn đã biết rằng không nên thực hiện phép tính toán sử dụng

những số liệu này vì kết quả đạt được sẽ không có ý nghĩa. Tuy nhiên, chúng ta vẫn muốn sử dụng tình trạng hôn nhân, giới tính, hoặc vị trí địa lí như một biến độc lập trong mô hình hồi quy vậy thì những biến này được kết hợp trong một phân tích hồi quy đa biến như thế nào? Câu trả lời nằm ở việc sử dụng biến giả (hoặc biến chỉ định) gọi là biến dummy.

Biến giả là một biến được đặt giá trị tương đương với 0 hoặc 1, phụ thuộc vào việc liệu các quan sát có chứa các tính chất được quan tâm hay không.

Ví dụ, xem xét biến giới tính có thể giữ hai giá trị: Nam giới hoặc nữ giới. Giới tính có thể được chuyển đổi thành biến giả X_1 như sau:

X_1= 1 nếu đó là nữ

X_1= 0 nếu là nam

Do đó, tập hợp dữ liệu về giới tính bao gồm nam và nữ giờ sẽ chỉ có các giá trị bằng 0 và 1 tương ứng trên X_1. Ta cũng chú ý rằng không có sự khác biệt nào trong việc mã hóa giới tính ngược lại là 1 nếu là nam và 0 nếu là nữ.

Nếu một biến phân loại có hơn hai tình huống đối lập, ta cần tạo nhiều biến giả. Ví dụ tình trạng hôn nhân, với các kết quả có thể có sau:

Chưa kết hôn bao giờ
Đã có gia đình
Đã li hôn
Góa

Trong trường hợp này, tình trạng hôn nhân có bốn giá trị. Để giải thích cho tất cả các khả năng, bạn cần tạo ra 3 biến giả, ít hơn 1 so với số lượng các tình huống có thể có cho biến nguyên thủy. Chúng có thể được mã hóa như sau:

X_1 = 1 nếu chưa kết hôn bao giờ, 0 nếu ngược lại

X_2= 1 nếu đã có gia đình, 0 nếu ngược lại

X_3 = 1 nếu đã li hôn, 0 nếu ngược lại

Chú ý rằng chúng ta không cần đến biến thứ 4 vì chúng ta sẽ biết một người là góa phụ nếu x_1= 0, x_2 = 0 và x_3= 0. Nếu một người

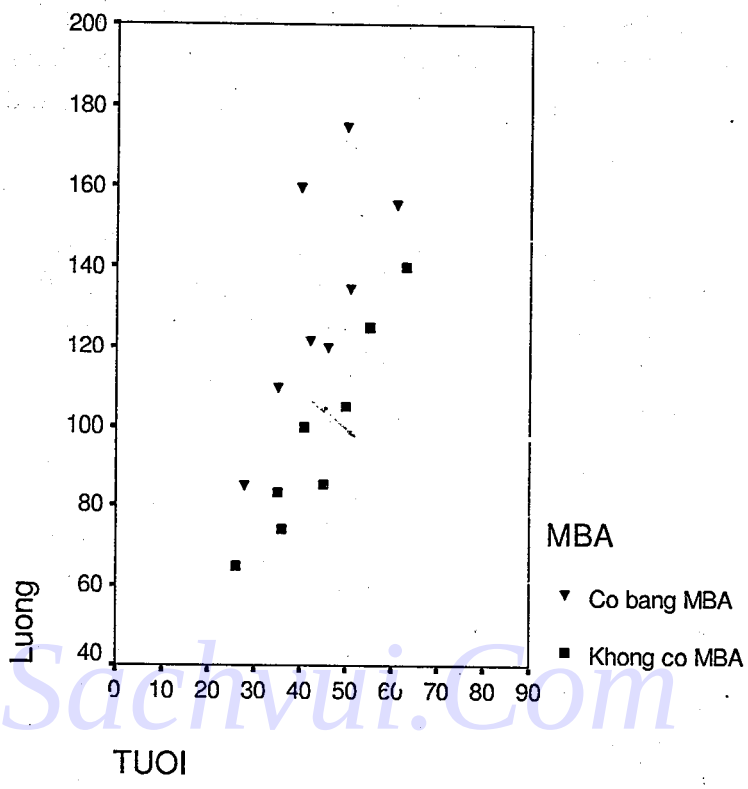
không độc thân, đã có gia đình, hay li hôn, thì họ ắt hẳn đã góa vợ hoặc chồng. Cần nhớ là luôn sử dụng ít hơn một biến giả so với số lượng phân loại. Lí do toán học là tính đa cộng tuyến hoàn hảo sẽ xuất hiện và ước lượng bình phương nhỏ nhất không thể đạt được nếu số biến giả bằng với số biến phân loại.

Ví dụ : người ta thu thập số liệu trên các nhân viên tuổi từ 24 đến 60 đang làm việc trong ngành kinh doanh về lương hàng tháng (Y-100.000 đồng) và số tuổi (X-tuổi). Mục tiêu ở đây là để xây dựng một mô hình để giải thích cho những biến động trong lương hằng tháng của nhân viên. Mặc dù tuổi và lương hằng năm có tương quan đáng kể với nhau (r = 0,686) ở mức ý nghĩa 0,05, nhưng hệ số xác định R^2 chỉ có 47% (bạn đọc tự chạy hồi qui để kiểm chứng các số liệu này). Bởi vậy, chúng ta có thể tìm những biến độc lập khác có thể giúp ta giải thích rõ hơn sự biến động trong lương hàng năm.

Giả sử chúng ta có thể chọn biến độc lập là các đối tượng chọn vào mẫu có bằng MBA hay không. Hình 9.26 cho thấy đồ thị rải điểm cho cùng một dữ liệu. Với số liệu có MBA được thể hiện bằng các tam giác và hình vuông là của không có MBA. Để kếp hợp thông tin về việc có MBA không ta tạo một biến mới, X_2, là một biến giả được mã hóa như sau : $X_2 = 1$ nếu có bằng MBA, 0 nếu không có.

Dữ liệu toàn bộ nằm trong file *Dummy*.

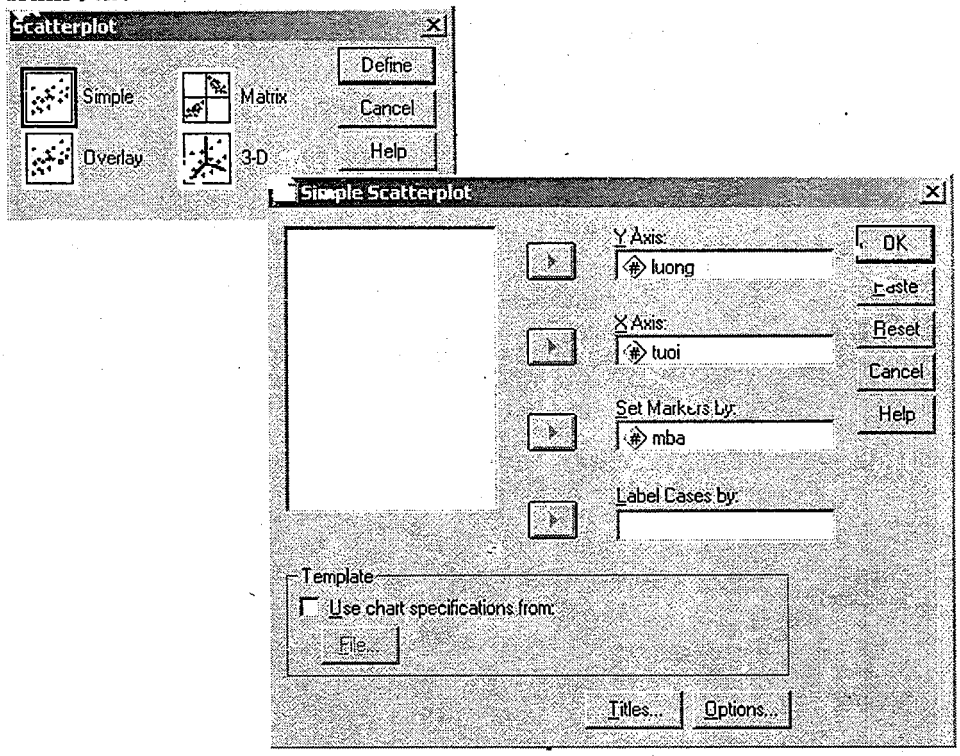
Hình 9.26



Muốn đưa bộ dữ liệu lên đồ thị rải điểm như hình Hình 9.26 bạn thực hiện các bước như hướng dẫn như Hình 9.27 sau đây.

Nhìn trên đồ thị rải điểm cảm nhận của chúng ta là mức lương của các nhà quản lý có bằng MBA ở trên một "mặt bằng" cao hơn.

Hình 9.27



Chúng ta sẽ phát triển một mô hình hồi qui đa biến với biến giả X_2 được kết hợp thêm với X_1 như 1 biến độc lập bình thường.

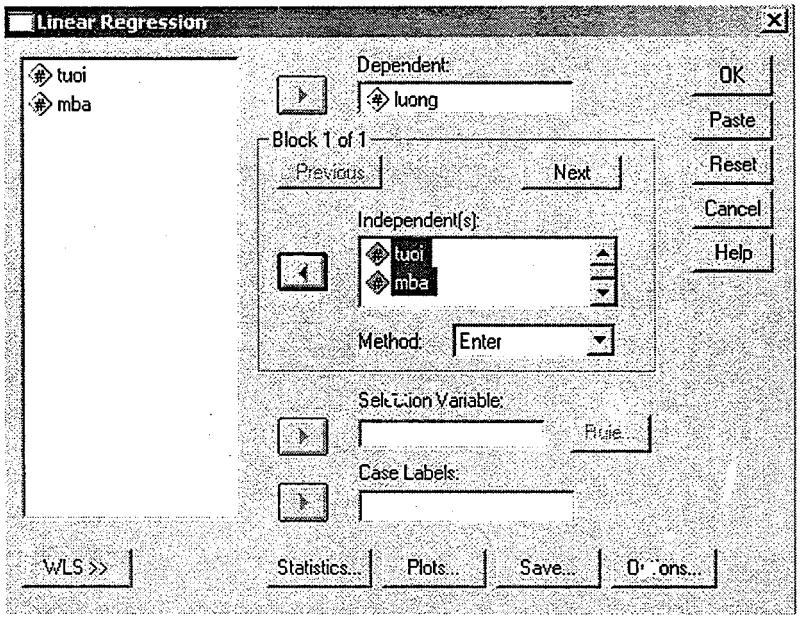
Mô hình hồi qui đa biến hai biến của ta có dạng như sau;

$$Y = \hat{a}_0 + \hat{a}_1X_1 + \hat{a}_2X_2 + e$$

Chạy hồi qui theo hướng dẫn Hình 9.28, bạn sẽ có được phương trình hồi quy sau đây như một ước lượng của mẫu.

$$\hat{Y} = 6{,}974 + 2{,}055X_1 + 35{,}236X_2$$

Hình 9.28



Bạn có thể kiểm tra lại qua các bảng kết quả sau :

Bảng 9.35 **Variables Entered/Removed(b)**

Model	Variables Entered	Variables Removed	Method
1	MBA, TUOI(a)	.	Enter

a All requested variables entered.
b Dependent Variable: Luong

Bảng 9.36 **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,885(a)	,783	,750	16,259

a Predictors: (Constant), MBA, TUOI

Bảng 9.37 **ANOVA(b)**

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	12423,434	2	6211,717	23,498	,000(a)
Residual	3436,566	13	264,351		
Total	15860,000	15			

a Predictors: (Constant), MBA, TUOI
b Dependent Variable: Luong

Bảng 9.38 **Coefficients(a)**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6,974	18,080		,386	,706
	TUOI	2,055	,391	,679	5,259	,000
	MBA	35,236	8,130	,560	4,334	,001

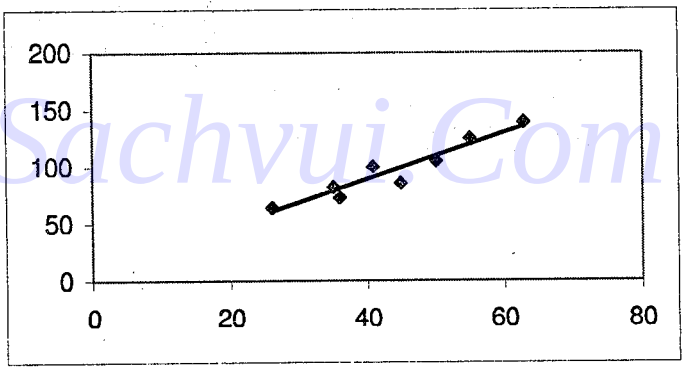
a Dependent Variable: Luong

Vì biến giả X_2 được mã hóa thành 0 hoặc 1 phụ thuộc vào địa vụ mức độ, kết hợp nó vào một mô hình hồi qui cũng giống như có hai đường hồi qui đơn có cùng độ nghiên, nhưng khác mặt bằng. ví dụ, khi $X_2 = 0$, phương trình hồi qui là:

$$\hat{Y} = 6,974 + 2,055X_1 + 35,236(0) = 6,974 + 2,055X_1$$

Đường này được thể hiện ở Hình 9.29.

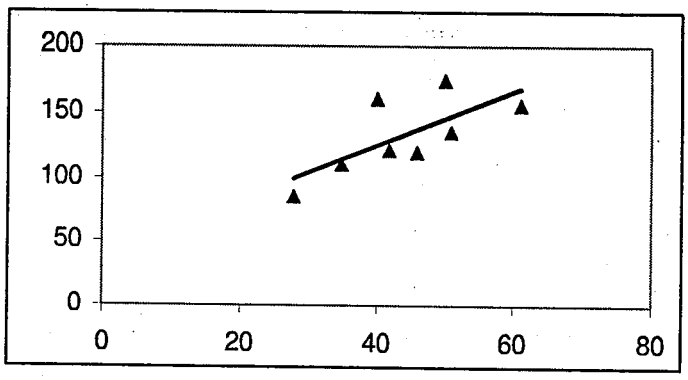
Hình 9.29



Tuy nhiên, khi $X_2 = 1$ (giám đốc điều hành có bằng MBA) phương trình hồi qui là:

$$\hat{Y} = 6,974 + 2,055X_1 + 35,236(1) = 42,210 + 2,055X_1$$

Hình 9.30



Đường hồi qui này được thể hiện trong hình trên.

Như bạn thấy, kết hợp biến giả ảnh hưởng mặt phẳng hồi qui. Trong trường hợp này, mặt phẳng cho các nhân viên có bằng MBA cao hơn so với mặt phẳng những người không có bằng MBA là 35,236 (100.000 đồng). Chúng ta hiểu ý nghĩa hệ số hồi qui đứng trước biến giả như sau: dựa trên dữ liệu này, và với tuổi (X_1) không đổi, chúng ta ước tính rằng các nhân viên có bằng MBA trung bình có lương hằng tháng cao hơn 35,236 (100.000 đồng) so với lương của các đồng nghiệp không có bằng MBA.

TÀI LIỆU THAM KHẢO

1. Aczel, D. Amir , *Complete Business Statistics*, Irwin, 1993.
2. Arsham, Hossein, *Statistical Data Analysis: Prove it with Data*, Manchester Metropolitan University, 2004
3. Berenson M. L., Levine D. M., Krehbiel T. C., *Basic Business Statistics*, 9th ed. Pearson Prentice Hall, 2004.
4. Bernard, H. Rusell, *Research Methods in Anthropology, Qualitative and Quantitative Approaches*, 2nd ed., AltaMira Press, 1995.
5. Cooper D. R., Schindler P.S., *Business Research Methods*, McGraw-Hill, 2003.
6. Croxton F. E., Cowden D. J., Klein S. *General Applied Statistics*, Prentice Hall of India, New Dehli, 1988.
7. Dutka, Alan, *AMA Handbook for Customer Satisfaction*, NTC Business Books, 1994.
8. Endruweit G., Trommsdorff G., *Từ Điển Xã Hội Học*, bản tiếng Việt, NXB Thế Giới, 2002.
9. Groebner, D.F., Shannon P.W., Fry P.C., and Smith K.D. (2005), *Business Statistics, A Decision Making Appoach*, Updated 6th ed., Peason Prentice Hall.
10. Gujarati, Damonda, Basic Econometrics, 4th ed., McGraw Hill, 2003.
11. Gupta, Vijay, *SPSS for Beginners*, Vijay Gupta Publication, 1999.
12. Hair Jr., J. F., Anderson, R. E., Tatham, R. L., Black, W. C. , *Multivariate Data Analysis with Readings*, 3rd ed., Macmillan Publishing Company, 1992.
13. Hà Văn Sơn và tập thể tác giả, *Giáo Trình Lý Thuyết Thống Kê*, ĐH Kinh Tế, NXB Thống Kê, 2004.
14. Hoàng Trọng, *Phân Tích Dữ liệu Đa Biến, Ứng Dụng Trong Kinh Tế và Kinh Doanh*, NXB Thống Kê, 1999.
15. Hoàng Trọng, *Xử Lý Dữ Liệu Nghiên Cứu với SPSS for Windows*, NXB Thống Kê, 2002.

16. Hoàng Trọng, Chu Nguyễn Mộng Ngọc, *Thống Kê Ứng Dụng trong Kinh tế Xã hội*, NXB Thống Kê, 2007.

17. Holbert, N. Bruce , Speece W. Mark , *Practical Marketing Research, An Integrated Global Perspective*, Prentice Hall, 1993.

18. Malhotra, K. Naresh, *Marketing Research, An Applied Oriented*, 2nd ed., Prentice Hall International Inc., 1996.

19. Neuman, William Lawrence, *Social Research Methods, Qualitative and Quantitative Approaches*, Allyn & Bacon, 2000.

20. Norusis, J. Marija, *SPSS 12.0, Guide to Data Analysis*, Prentice Hall.

21. Norusis, J. Marija, *SPSS for Windows, Base System User's Guide*, SPSS Inc., 1993.

22. Phạm Văn Quyết, Nguyễn Quý Thanh, *Phương Pháp Nghiên Cứu Xã Hội Học*, NXB Đại Học Quốc Gia Hà Nội.

23. Robert M. Worcester, John Downham, *Consumer Market Research Handbook*, 3rd ed., 1986, ESOMAR.

24. Saunders M. NK., Lewis P., Thornhill A., *Research Methods for Business Students*, Pitman Publishing, 1997.

25. Sirkin, R. Mark, *Statistics for the Social Sciences*, 2nd ed., Sage Publications, 1999.

26. *SPSS Base 8.0 User's Guide*, SPSS Inc., 1998.

27. Trần Bá Nhẫn, Đinh Thái Hoàng, *Thống Kê Ứng Dụng trong quản trị, kinh doanh và nghiên cứu kinh tế*, Đại Học Kinh Tế, 2003.

28. Trần Chung Ngọc, Trần Văn Tươi, *Thống Kê Căn Bản*, Phân khoa Khoa học xã hội, ĐH Vạn Hạnh, 1974.

29. Trần Xuân Kiêm, Nguyễn Văn Thi, *Nghiên Cứu Tiếp Thị*, NXB Thống Kê, 2001.

30. Võ Văn Huy, Võ Thị Lan, Hoàng Trọng, *Ứng dụng SPSS For Windows để xử lý và phân tích dữ kiện nghiên cứu*, NXB Khoa Học và Kỹ Thuật, 1997.

PHỤC LỤC
BẢNG CÂU HỎI PHỎNG VẤN (tập tin data thuc hanh)

PVV	Đọc soát	Mã số

Kính chào ông/bà/anh/chị, chúng tôi đang thực hiện một cuộc thăm dò về thói quen đọc và mua báo; hôm nay đến đây để xin ý kiến của ông/bà/anh/chị. Mục đích của cuộc thăm dò này là tìm hiểu về một phần cuộc sống tinh thần của người dân cũng như giúp cho các cơ quan báo chí hiểu rõ hơn nhu cầu đọc và mua báo, trên cơ sở đó có phương hướng cải tiến tờ báo, phục vụ tốt hơn nhu cầu của bạn đọc.

Giờ bắt đầu phỏng vấn:_____

A. ĐỌC VÀ MUA BÁO

1. Trong vòng 1 năm qua anh/chị/ông/bà có thường xuyên đọc báo không?
(SHOWCARD - CHỈ CHỌN 1 TRẢ LỜI)

1.	hầu như không đọc báo	Ngưng → ghi chép vào bảng thống kê
2.	thỉnh thoảng (tuần 1-2 tờ)	Ngưng → ghi chép vào bảng thống kê
3.	thường xuyên (tuần 3-7 tờ)	Tiếp tục
4.	rất thường xuyên (trên 7 tờ/ tuần)	Tiếp tục

2a. Trong vòng 6 tháng qua anh/chị/ông/bà <u>thường đọc</u> các tờ báo tiếng Việt nào? (CÓ THỂ CHỌN NHIỀU TRẢ LỜI)

2b. Trong các tờ báo kể trên, anh/chị/ông/bà <u>thích đọc</u> các tờ báo nào nhất? (CHỈ CHỌN TỐI ĐA 3 TRẢ LỜI)

2c. Trong vòng 6 tháng qua các thành viên gia đình anh/chị/ông/bà <u>thường mua</u> các tờ báo tiếng Việt nào?

2d. Các thành viên trong gia đình anh/chị/ông/bà <u>thường mua</u> các tờ báo này <u>như thế nào</u>? (HỎI LẦN LƯỢT TỪNG LOẠI BÁO ĐÃ MUA TRONG CÂU 2c) *(đặt báo định kỳ: ghi số 1; mua tại sạp báo tiện đường đi làm, đi học: ghi số 2; mua tại sạp báo gần nhà: ghi số 3; không biết: ghi số 4)*

	2a	2b	2c	2d	GHI CHÚ
Hà Nội Mới	1	1	1		DÙNG SHOWCARD CHO 4 CÂU: 2a, 2b, 2c, 2d
Sài Gòn Giải Phóng	2	2	2		
Lao Động	3	3	3		
Người Lao động	4	4	4		
Tiền Phong	5	5	5		
Thanh Niên	6	6	6		
Tuổi Trẻ	7	7	7		
Phụ Nữ Việt Nam	8	8	8		
Phụ Nữ TP HCM	9	9	9		
Thời Báo Kinh Tế Việt Nam	10	10	10		

Thời Báo Kinh Tế Sài Gòn	11	11	11		HỎI VÀ GHI CHÉP THEO CỘT
Sài Gòn Tiếp Thị	12	12	12		
Thế Giới Phụ Nữ	13	13	13		
Tiếp Thị và Gia Đình	14	14	14		
Mua & Bán	15	15	15		
An Ninh Thế Giới	16	16	16		
An Ninh Thủ đô	17	17	17		
Công An TPHCM	18	18	18		
Khác (kể tên).....	19	19	19		

3. Trong gia đình anh/chị/ông/bà, số lượng người đọc báo thường xuyên là bao nhiêu người?

Ghi một con số cụ thể : _____ người

4. Anh/chị/ông/bà thường đọc báo ở đâu? (SHOWCARD - CHỈ CHỌN TỐI ĐA 2 TRẢ LỜI)

1. nhà
2. cơ quan, văn phòng, nơi làm việc
3. nơi bán hàng
4. nơi khác (ghi cụ thể) .....

5. Anh/chị/ông/bà thường đọc báo vào những lúc nào? (SHOWCARD - CHỈ CHỌN TỐI ĐA 2 TRẢ LỜI)

1. sáng sớm/trước giờ làm việc
2. trong giờ làm việc
3. lúc rãnh rỗi
4. lúc khác (ghi cụ thể) .....

6. Anh/chị/ông/bà thường đọc báo như thế nào? (SHOWCARD - CHỈ CHỌN 1 TRẢ LỜI CHO 1 CỘT)

	Các tờ báo nói chung	tờ báo thường đọc nhất
Đọc theo thứ tự từ trang đầu đến trang cuối	1	1
Xem lướt qua các đề mục, đọc những trang mục ưa thích trước rồi đến các trang mục khác sau	2	2
Chỉ đọc các trang mục ưa thích, ít khi đọc các trang mục khác	3	3
Xem các tin đáng chú ý trên trang nhất rồi tìm đến đọc trước các bài có quan tâm chú ý	4	4

7. Trong gia đình anh/chị/ông/bà, ai thường là người quyết định việc mua báo? (CHỈ CHỌN 1 TRẢ LỜI)

1. bản thân
2. người khác

8. Nếu là người khác quyết định việc mua báo, xin cho biết đó là ai trong gia đình? (CHỈ CHỌN 1 TRẢ LỜI)

ông/bà	1	người con gái	4
người cha	2	người con trai	5
người mẹ	3	người khác	6

9. Với các báo thường đọc, anh/chị/ông/bà thường đọc từ

1. báo cơ quan
2. gia đình tự mua
3. mượn từ bạn bè, người thân, hàng xóm
4. khác .....

B. ĐỌC BÁO SÀI GÒN TIẾP THỊ

XEM LẠI Q2a. NẾU ĐÃ CÓ ĐỌC SGTT THÌ HỎI Q10. NẾU KHÔNG, HỎI Q11

10. Trong vòng 6 tháng qua, anh/chị/ông/bà có đọc thường xuyên báo SGTT không? (SHOWCARD)

1. mỗi tháng đọc 1-2 số báo — tiếp câu 11
2. gần như đọc hàng tuần — tiếp câu 12
3. không bỏ sót số báo nào — tiếp câu 12

11. Xin vui lòng cho biết lý do vì sao anh/chị/ông/bà không đọc, hay không đọc báo SGTT thường xuyên?

_____

=> tiếp câu 36

12. Trong gia đình, ngoài anh/chị/ông/bà, còn bao nhiêu người nữa đọc báo SGTT ít nhất 2 số báo/1 tháng trong vòng 6 tháng qua?

Ghi một con số cụ thể _____ người

13. Anh/chị/ông/bà thường đọc báo SGTT vào ngày nào trong tuần? (SHOWCARD – CHỈ CHỌN 1 TRẢ LỜI)

1. thứ Năm
2. thứ Sáu
3. thứ Bảy
4. Chủ Nhật
5. ngày khác trong tuần

14. Cách thức Anh/chị/ông/bà đọc báo SGTT như thế nào? (SHOWCARD - CHỈ CHỌN 1 TRẢ LỜI)

1. thường đọc 1 lần hết tờ báo
2. thường đọc 2-3 lần, mỗi lần đọc 1 phần
3. thường đọc 1 lần, sau đó xem lại một vài trang mục ưa thích ở những lần khác

15. Anh/chị/ông/bà thường đọc báo SGTT theo thứ tự nội dung như thế nào? (SHOWCARD - CHỈ CHỌN 1 TRẢ LỜI)

đọc theo thứ tự từ trang đầu đến trang cuối	1
xem lướt qua các đề mục, đọc những trang mục ưa thích trước rồi đến các trang mục khác sau	2
chỉ đọc các trang mục ưa thích, ít khi đọc các trang mục khác	3
xem các tin đáng chú ý trên trang nhất rồi tìm đến đọc trước các bài có quan tâm chú ý	4

16. Anh/chị/ông/bà thường đọc các trang mục nào của báo SGTT? (SHOWCARD - CÓ THỂ CHỌN NHIỀU TRẢ LỜI)

17. Trong các trang mục của báo SGTT thường đọc, Anh/chị/ông/bà thích đọc các trang mục nào nhất? (SHOWCARD - CHỌN TỐI ĐA 5 TRẢ LỜI)

18. Trong 6 tháng đầu năm 2001, theo Anh/chị/ông/bà thì báo SGTT đã nổ lực tập trung nhiều vào trang mục nào? (SHOWCARD – CHỈ CHỌN 1 TRẢ LỜI)

Trang mục	16	17	18	Trang mục	16	17	18
Bạn đọc	1	1	1	ĐB sông Cửu Long	12	12	12
Nhà đất	2	2	2	Kinh doanh tiếp thị	13	13	13
Dịch vụ	3	3	3	Phóng sự ảnh	14	14	14
Tin học	4	4	4	Chuyển động thị trường	15	15	15
Mua sắm - dịch vụ	5	5	5	Vấn đề	16	16	16
Mua sắm	6	6	6	Cẩm nang siêu thị	17	17	17
Ẩm thực & đời sống	7	7	7	Quà tặng bạn đọc	18	18	18
Giải trí	8	8	8	Quảng cáo	19	19	19
Gia đình	9	9	9	Trang Hà Nội và MB	20	20	20
Dành cho đàn ông	10	10	10	Không nhớ, không để ý	21	21	21
Thế giới tiêu dùng	11	11	11				

19. Hãy xếp hạng các chủ đề sau đây trên báo SGTT tùy theo mức độ quan tâm của Anh/chị/ông/bà đối với từng loại chủ đề? (SHOWCARD – CHỦ ĐỀ NÀO QUAN TÂM NHẤT THÌ GHI SỐ 1, QUAN TÂM THỨ NHÌ THÌ GHI SỐ 2, QUAN TÂM THỨ BA THÌ GHI SỐ 3)

thông tin thị trường _____
mua sắm _____
gia đình _____

20. Theo Anh/chị/ông/bà, báo SGTT có tác dụng: (SHOWCARD - CÓ THỂ CHỌN NHIỀU TRẢ LỜI)

1. giải trí
2. tư vấn, hướng dẫn tiêu dùng
3. cung cấp tài liệu nghiên cứu và tham khảo
4. tác dụng khác .....

21. Anh/chị/ông/bà xem (đọc) các trang quảng cáo trên báo SGTT như thế nào? (SHOWCARD - CHỈ CHỌN 1 TRẢ LỜI)

1. thường xem (đọc) hết các trang quảng cáo
2. thường xem lướt qua và chỉ đọc một số một số quảng cáo có quan tâm
3. ít khi xem (đọc) các trang quảng cáo
4. hầu như không xem (đọc) các trang quảng cáo

22. Mục đích xem (đọc) quảng cáo trên báo SGTT của anh/chị/ông/bà là gì? (SHOWCARD – CÓ THỂ CHỌN NHIỀU TRẢ LỜI)

1. Tìm kiếm thông tin để mua sắm
2. Tìm cơ hội mua hàng khuyến mãi
3. Xem giới thiệu công ty và SP mới
4. Phục vụ cho học tập, nghiên cứu
5. Để giải trí
6. Mục đích khác .....

23. Nếu gia đình anh/chị/ông/bà có mua báo SGTT, thì số lượng người đọc báo SGTT trong gia đình trung bình là bao nhiêu người (kể cả anh/chị/ông/bà)? Trong đó số người thường xuyên xem các trang quảng cáo là bao nhiêu người?

Số người đọc : _____; số người xem quảng cáo: _____

24. Nếu nơi làm việc của anh/chị/ông/bà có mua báo SGTT (hoặc do một người nào đó mua đem vào), thì số lượng người đọc 1 tờ báo SGTT trung bình là bao nhiêu người ? Trong đó số người thường xuyên xem các trang quảng cáo là bao nhiêu người?

Số người đọc : _____; số người xem quảng cáo: _____

25. Anh/chị/ông/bà có thường bàn luận với những người khác về những nội dung trên tờ SGTT?

1. Thường xuyên 2. Thỉnh thoảng 3. Ít Khi 4. Không

26. Những người Anh/chị/ông/bà thường bàn luận về nội dung trên báo SGTT thường là những người nào?

(SHOWCARD - CÓ THỂ CHỌN NHIỀU TRẢ LỜI)

thành viên trong gia đình	1	bà con	5
bạn bè cùng nơi làm việc	2	láng giềng	6
bạn học	3	người khác	7
bạn thân	4		

27. Anh/chị/ông/bà thường để các số báo SGTT mới trong vòng 1 tháng ở đâu trong nhà?

(SHOWCARD - CHỈ CHỌN 1 TRẢ LỜI)

phòng khách	1	phòng ăn	5
phòng sinh hoạt gia đình	2	nơi khác trong nhà	6
phòng học hay làm việc ở nhà	3	không lưu trữ	7
phòng ngủ	4	gia đình không mua	8

28. Anh/chị/ông/bà giữ gìn tờ báo SGTT như thế nào? (SHOWCARD - CHỈ CHỌN 1 TRẢ LỜI)

1. không giữ lại (không phải báo ở nhà mua)
2. không quan tâm đến việc giữ gìn tờ báo SGTT
3. để đại đâu đó
4. có giữ gìn, sắp xếp thứ tự để có thể dễ dàng tham khảo về sau
5. cắt và lưu trữ một số thông tin quan trọng

E. QUAN NIỆM VỀ CUỘC SỐNG VÀ GIẢ

36. Theo anh/chị/ông/bà, rồ
với cuộc sống của

Đ
Ng
Chơ
Hát
Chơi
Lướt In
Chơi bi-

F. THÔN

Họ tên: _____

Địa chỉ : _____

Tuổi: _____

Giới tính

PHÂN TÍCH DỮ LIỆU NGHIÊN CỨU VỚI SPSS

Tập I

HOÀNG TRỌNG - CHU NGUYỄN MỘNG NGỌC

NHÀ XUẤT BẢN HỒNG ĐỨC
111 Lê Thánh Tôn - Q.1 - TP.HCM
ĐT : 08.8244.534

Chịu trách nhiệm xuất bản : Hoàng Chí Dũng
Biên tập : Hoàng Trọng
Thiết kế bìa : Vũ Xuân Khanh

In 2.000 cuốn khổ 16x24 cm tại Nhà In THÀNH CÔNG
Giấy phép xuất bản số 323-2008/CXB/T1-46-24/HĐ, cấp ngày
In xong và nộp lưu chiểu quý 3 năm 2008.

23. Nếu gia đình anh/chị/ông/bà có mua báo SGTT, thì số lượng người đọc báo SGTT trong gia đình trung bình là bao nhiêu người (kể cả anh/chị/ông/bà)? Trong đó số người thường xuyên xem các trang quảng cáo là bao nhiêu người?

Số người đọc : _____; số người xem quảng cáo: _____

24. Nếu nơi làm việc của anh/chị/ông/bà có mua báo SGTT (hoặc do một người nào đó mua đem vào), thì số lượng người đọc 1 tờ báo SGTT trung bình là bao nhiêu người ? Trong đó số người thường xuyên xem các trang quảng cáo là bao nhiêu người?

Số người đọc : _____; số người xem quảng cáo: _____

25. Anh/chị/ông/bà có thường bàn luận với những người khác về những nội dung trên tờ SGTT?

1. Thường xuyên 2. Thỉnh thoảng 3. Ít Khi 4. Không

26. Những người Anh/chị/ông/bà thường bàn luận về nội dung trên báo SGTT thường là những người nào?

(SHOWCARD - CÓ THỂ CHỌN NHIỀU TRẢ LỜI)

thành viên trong gia đình	1	bà con	5
bạn bè cùng nơi làm việc	2	láng giềng	6
bạn học	3	người khác	7
bạn thân	4		

27. Anh/chị/ông/bà thường để các số báo SGTT mới trong vòng 1 tháng ở đâu trong nhà?

(SHOWCARD - CHỈ CHỌN 1 TRẢ LỜI)

phòng khách	1	phòng ăn	5
phòng sinh hoạt gia đình	2	nơi khác trong nhà	6
phòng học hay làm việc ở nhà	3	không lưu trữ	7
phòng ngủ	4	gia đình không mua	8

28. Anh/chị/ông/bà giữ gìn tờ báo SGTT như thế nào? (SHOWCARD - CHỈ CHỌN 1 TRẢ LỜI)

1. không giữ lại (không phải báo ở nhà mua)
2. không quan tâm đến việc giữ gìn tờ báo SGTT
3. để đại đâu đó
4. có giữ gìn, sắp xếp thứ tự để có thể dễ dàng tham khảo về sau
5. cắt và lưu trữ một số thông tin quan trọng

C. ĐÁNH GIÁ VÀ GÓP Ý CHO BÁO SGTT

29. Anh/chị/ông/bà đánh giá các mặt sau đây của tờ báo SGTT như thế nào?
(SHOWCARD)

A. Về nội dung	Hoàn toàn không hài lòng	Không hài lòng	Được	Hài lòng	Rất hài lòng	Không ý kiến
1. Tính xác thực của thông tin	1	2	3	4	5	8
2. Tính thời sự, cập nhật	1	2	3	4	5	8
3. Tính bổ ích	1	2	3	4	5	8
4. Tính phân tích	1	2	3	4	5	8
5. Tính thực tế	1	2	3	4	5	8
6. Tính đúc kết, hướng dẫn	1	2	3	4	5	8
7. Tính mới, đột phá	1	2	3	4	5	8
B. Hình thức						
1. Trình bày bìa	1	2	3	4	5	8
2. Ngôn ngữ thể hiện	1	2	3	4	5	8
3. Hình ảnh	1	2	3	4	5	8
4. Chất lượng in	1	2	3	4	5	8
5. Sắp xếp trang mục	1	2	3	4	5	8
6. Trang trí	1	2	3	4	5	8
C. Đánh giá chung	1	2	3	4	5	8

30a. Theo Anh/chị/ông/bà, báo SGTT nên tăng thêm diện tích mặt báo của những trang mục nào?

30b. Theo Anh/chị/ông/bà, báo SGTT nên giảm bớt diện tích mặt báo của những trang mục nào?

30c. Theo Anh/chị/ông/bà, báo SGTT nên tập trung cải tiến những trang mục nào?

30d. Theo Anh/chị/ông/bà, báo SGTT nên bỏ bớt những trang mục nào?

	30a Tăng	30b Giảm	30c Cải tiến	30d Bỏ	GHI CHÚ
Bạn đọc	1	1	1	1	DÙNG SHOWCARD CHO 4 CÂU 30a, 30b, 30c, 30d HỎI VÀ GHI CHÉP THEO CỘT
Nhà đất	2	2	2	2	
Dịch vụ	3	3	3	3	
Tin học	4	4	4	4	
Mua sắm - dịch vụ	5	5	5	5	
Mua sắm	6	6	6	6	
Ẩm thực & đời sống	7	7	7	7	
Giải trí	8	8	8	8	
Gia đình	9	9	9	9	
Dành cho đàn ông	10	10	10	10	
Thế giới tiêu dùng	11	11	11	11	
ĐB sông Cửu Long	12	12	12	12	

Kinh doanh tiếp thị	13	13	13	13	
Phóng sự ảnh	14	14	14	14	
Chuyển động thị trường	15	15	15	15	
Vấn đề	16	16	16	16	
Cẩm nang siêu thị	17	17	17	17	
Quà tặng bạn đọc	18	18	18	18	
Quảng cáo	19	19	19	19	
Trang Hà Nội và MB	20	20	20	20	
Không cần điều chỉnh	21	21	21	21	

31. Theo Anh/chị/ông/bà, báo SGTT nên mở thêm những trang mục nào? (tối đa 2 đề nghị)

1.
2.

D. SẢN PHẨM VÀ HOẠT ĐỘNG KHÁC CỦA BÁO SGTT

32. Trong 6 tháng qua, ngoài tờ báo SGTT, gia đình Anh/chị/ông/bà có mua và đọc các sản phẩm báo chí nào khác của SGTT không? (SHOWCARD)
33. Mức độ hài lòng của Anh/chị/ông/bà đối với các sản phẩm báo chí này của SGTT như thế nào? (SHOWCARD)

	Đọc và mua		Mức độ hài lòng					
	đọc	mua	Hoàn toàn không hài lòng	Không hài lòng	Được	Hài lòng	Rất hài lòng	Không ý kiến
Cẩm nang tiêu dùng	1	1	1	2	3	4	5	8
Cẩm nang kỹ thuật	2	2	1	2	3	4	5	8
Cẩm nang mua sắm	3	3	1	2	3	4	5	8

34. Trong 6 tháng qua, Anh/chị/ông/bà có tham gia các hoạt động ngoài báo của SGTT / SGTT tham gia tổ chức không? (SHOWCARD)
35. Mức độ hài lòng của Anh/chị/ông/bà đối với các hoạt động này như thế nào? (SHOWCARD)

	Tham gia	Mức độ hài lòng					
		Hoàn toàn không hài lòng	Không hài lòng	Được	Hài lòng	Rất hài lòng	Không ý kiến
Hội chợ HVN CLC	1	1	2	3	4	5	8
Hướng dẫn tiêu dùng	2	1	2	3	4	5	8
Hàng VNCLC trên HTV	3	1	2	3	4	5	8

E. QUAN NIỆM VỀ CUỘC SỐNG VÀ GIẢI TRÍ

36. Theo anh/chị/ông/bà, tầm quan trọng của các yếu tố sau đây như thế nào đối với cuộc sống của một người? (SHOWCARD)

	Không quan trọng						Rất quan trọng
1. có nhiều tiền	1	2	3	4	5	6	7
2. đạt trình độ học vấn cao	1	2	3	4	5	6	7
3. có địa vị trong xã hội	1	2	3	4	5	6	7
4. có bạn bè tốt	1	2	3	4	5	6	7
5. gia đình ổn định	1	2	3	4	5	6	7
6. có tự do cá nhân	1	2	3	4	5	6	7
7. có sức khỏe tốt	1	2	3	4	5	6	7
8. có nghề nghiệp thích hợp	1	2	3	4	5	6	7
9. có tình yêu	1	2	3	4	5	6	7
10. được mọi người tôn trọng	1	2	3	4	5	6	7
11. sống có ích cho người khác	1	2	3	4	5	6	7
12. được hưởng thụ nhiều thú vui trong cuộc sống	1	2	3	4	5	6	7

37. Trong vòng một năm qua, trong những lúc rãnh rỗi, anh/chị/ông/bà thường làm gì? (SHOWCARD - CHỈ CHỌN TỐI ĐA 5 TRẢ LỜI)

Xem TV	1	Nói chuyện với bạn bè	13
Nghe radio	2	Chơi thể thao	14
Đọc báo, tạp chí	3	Đi du lịch, dã ngoại	15
Xem băng video	4	Dạo công viên, khu vui chơi	16
Đi xem phim ở rạp	5	Đi uống cà phê, trà/chè	17
Đi xem ca nhạc, kịch	6	Chơi cây cảnh, động vật cảnh	18
Nghe nhạc tại nhà	7	Nấu ăn	19
Chơi nhạc (tự chơi, với bạn bè)	8	Đi ăn uống	20
Hát Karaoke	9	Uống bia, rượu (nhậu)	21
Chơi video game, game vi tính	10	Đi mua sắm	22
Lướt Internet, chat ...	11	Khác	
Chơi bi-da	12	.....	

F. THÔNG TIN CÁ NHÂN

Họ tên: _____,điện thoại: _____

Địa chỉ : _____

Tuổi: _____; Số người trong hộ gia đình : _____

Giới tính 1. nam 2. nữ

Thu nhập CN (TB tháng) 1. Không 2. dưới 1trđ 3. 1-2 trđ 4. 2-4 trđ 5. trên 4 trđ

Thu nhập GĐ (TBtháng) 1. dưới 2 trđ 2. 2-4 trđ 3. 4-6 trđ 4. 6-10 trđ

Học vấn

1. cấp 1 2. Cấp 2 3. cấp 3-THCN 4. CĐ -SV ĐH 5. Tốt nghiệp ĐH 6. Sau ĐH

Nghề nghiệp

1. Công chức	2. Giáo viên	3. Nhân viên văn phòng
4. Chủ doanh nghiệp	5. Nhân viên KD	6. Tự kinh doanh SP-DV
7. Buôn bán nhỏ	8. CN có tay nghề	9. Lao động đơn giản
10. SV-học sinh	11. Về hưu	12. Không làm việc

13. Nghề chuyên môn (bác sĩ, kiến trúc sư, luật sư, nhạc sĩ, nghệ sĩ, ...)

14. Nghề khác _____

Giờ kết thúc phỏng vấn: _____

XIN CHÂN THÀNH CẢM ƠN SỰ HỢP TÁC CỦA ANH/CHỊ/ÔNG/BÀ !

Có thể sẽ có một người nào đó của toà soạn báo SGTT sẽ tới hỏi/ gọi điện thoại hỏi thăm là vào ngày giờ này, tôi có đến phỏng vấn hay không, xin anh/chị/ông/bà xác nhận là có, rất cám ơn (tặng quà).

Ngày..... tháng 7 năm 2001

Phỏng vấn viên

PHÂN TÍCH DỮ LIỆU NGHIÊN CỨU VỚI SPSS

Tập I

HOÀNG TRỌNG - CHU NGUYỄN MỘNG NGỌC

NHÀ XUẤT BẢN HỒNG ĐỨC
111 Lê Thánh Tôn - Q.1 - TP.HCM
ĐT : 08.8244.534

Chịu trách nhiệm xuất bản : Hoàng Chí Dũng
Biên tập : Hoàng Trọng
Thiết kế bìa : Vũ Xuân Khanh

In 2.000 cuốn khổ 16x24 cm tại Nhà In THÀNH CÔNG
Giấy phép xuất bản số 323-2008/CXB/T1-46-24/HĐ, cấp ngày 20/08/2008.
In xong và nộp lưu chiểu quý 3 năm 2008.