HOÀNG TRỌNG - CHU NGUYỄN MỘNG NGỌC

PHÂN TÍCH DỮ LIỆU NGHIÊN CỨU VỚI SPSS

TRƯỜNG ĐẠI HỌC KINH TẾ TP HỒ CHÍ MINH
NHÀ XUẤT BẢN HỒNG ĐỨC

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 2

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

MỤC LỤC

CHƯƠNG XIII: PHÂN TÍCH BIỆT SỐ

CHƯƠNG XIV: PHÂN TÍCH CỤM

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:

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

CHƯƠNG X
HỒI QUI BINARY LOGISTIC

Trong Chương IX ở Tập 1, chúng ta đã nghiên cứu hồi qui tuyến tính để xem xét mối liên hệ tuyến tính giữa biến độc lập và biến phụ thuộc dạng định lượng, tức là mô tả mối quan hệ là dạng đường thẳng; và chúng ta cũng đã phân biệt thuật ngữ tuyến tính trong cụm từ "Hồi qui tuyến tính" là tuyến tính theo các hệ số hồi qui. Với những mối quan hệ có dạng phi tuyến 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". Trong hồi qui tuyến tính với các quan hệ phi tuyến dạng của mối quan hệ giữa các biến độc lập và biến phụ thuộc là phi tuyến nhưng hình thức của các hệ số trong mô hình hồi qui vẫn là tuyến tính.

Tại chương này chúng ta sẽ nghiên cứu một dạng hồi qui khá đặc biệt có tên là hồi qui Binary Logistic. Điểm khá đặc biệt này thể hiện ở ứng dụng chính của hồi qui Binary Logistic.

1. ỨNG DỤNG CỦA HỒI QUI BINARY BINARY LOGISTIC

Hồi qui Binary Logistic sử dụng biến phụ thuộc dạng nhị phân để ước lượng xác suất một sự kiện sẽ xảy ra với những thông tin của biến độc lập mà ta có được.

Có rất nhiều hiện tượng trong tự nhiên chúng ta cần dự đoán khả năng xảy ra một sự kiện nào đó mà ta quan tâm (chính là xác suất xảy ra), ví dụ sản phẩm mới được chấp nhận hay không, người vay trả được nợ hay không, mua hay không mua... Những biến nghiên cứu có 2 biểu hiện như vậy gọi là biến thay phiên (dichotomous), hai biểu hiện này sẽ được mã hóa thành hai giá trị 0 và 1 và ở dưới dạng này gọi là biến nhị phân. Khi biến phụ thuộc ở dạng nhị phân thì không thể phân tích với dạng hồi qui thông thường vì làm như vậy sẽ xâm phạm các giả định, rất dễ thấy là khi biến phụ thuộc chỉ có 2 biểu hiện thì thật không phù hợp khi giả định rằng phần dư có phân phối chuẩn, mà thay vào đó sẽ có phân phối nhị thức, điều này sẽ

làm mất hiệu lực của các kiểm định thống kê trong phép hồi qui thông thường. Một khó khăn khác khi dùng hồi qui tuyến tính thông thường là giá trị dự đoán được của biến phụ thuộc không thể được diễn dịch như xác suất (giá trị ước lượng của biến phụ thuộc trong hồi qui Binary Logistic phải rơi vào khoảng (0;1))

2. MÔ HÌNH BINARY LOGISTIC

Với hồi qui Binary Logistic, thông tin chúng ta cần thu thập về biến phụ thuộc là một sự kiện nào đó có xảy ra hay không, biến phụ thuộc Y lúc này có hai giá trị 0 và 1, với 0 là không xảy ra sự kiện ta quan tâm và 1 là có xảy ra, và tất nhiên là cả thông tin về các biến độc lập X. Từ biến phụ thuộc nhị phân này, một thủ tục sẽ được dùng để dự đoán xác suất sự kiện xảy ra theo quy tắc nếu xác suất được dự đoán lớn hơn 0,5 thì kết quả dự đoán sẽ cho là "có" xảy ra sự kiện, ngược lại thì kết quả dự đoán sẽ là "không". Chúng ta sẽ nghiên cứu mô hình hàm Binary Logistic trong trường hợp đơn giản nhất là khi chỉ có một biến độc lập X

Ta có mô hình hàm Binary Logistic như sau

$$P_i = E(Y = 1/X) = \frac{e^{(B_0+B_1X)}}{1+e^{(B_0+B_1X)}}$$

Trong công thức này $P_i = E(Y=1/X) = P(Y=1)$ gọi là xác suất để sự kiện xảy ra (Y =1) khi biến độc lập X có giá trị cụ thể là X_i. Ký hiệu biểu thức (B_o+B_1X) là z, ta viết lại mô hình hàm Binary Logistic như sau:

$$P(Y=1) = \frac{e^z}{1+e^z}$$

Vậy thì xác suất không xảy ra sự kiện là:

$$P(Y=0) = 1 - P(Y=1) = 1 - \frac{e^z}{1+e^z}$$

Thực hiện phép so sánh giữa xác suất một sự kiện xảy ra với xác suất sự kiện đó không xảy ra, tỷ lệ chênh lệch này có thể được thể hiện trong công thức:

$$\frac{P(Y=1)}{P(Y=0)} = \frac{\frac{e^z}{1+e^z}}{1-\frac{e^z}{1+e^z}}$$

Lấy log cơ số e hai vế của phương trình trên rồi thực hiện biến đổi vế phải ta được kết quả là

$$\log_e[\frac{P(Y=1)}{P(Y=0)}] = \log_e e^z$$

vì $Log_e e^z = z$ nên kết quả cuối cùng là

$$\log_e[\frac{P(Y=1)}{P(Y=0)}] = B_0 + B_1 X$$

Hay viết cách khác: $\log_e[\frac{P_i}{1-P_i}] = B_0 + B_1 X$ (*) là dạng hàm hồi qui Binary Logistic. Và ta có thể mở rộng mô hình Binary Logistic cho 2 hay nhiều biến độc lập X_k.

2.1. Diễn dịch các hệ số hồi qui của mô hình Binary logistic

Tên gọi hồi qui Binary Logistic xuất phát từ quá trình biến đổi lấy logarit của thủ tục này. Sự chuyển hoá này làm cho các hệ số của hồi qui binary logistic có nghĩa khác với hệ số hồi qui trong trường hợp thông thường với các biến phụ thuộc dạng thập phân và trở nên khó diễn dịch ý nghĩa.

Đó là: từ công thức (*) ta hiểu hệ số ước lượng B_1 cho biết khi X_1 tăng 1 đơn vị thì log của tỷ lệ $(P_i/1-P_i)$ tăng B_1 đơn vị.

Tuy nhiên nếu ta chỉ quan tâm đến chiều hướng của tác động thì ta thấy rằng phương trình bên trái của (*) đồng biến với P_i (tức xác suất Y=1) nên nếu hệ số B_i mang dấu dương thì tăng X_i sẽ làm tăng khả năng Y nhận giá trị 1 trong khi hệ số âm làm giảm khả năng này.

Ta có $\frac{\partial P(Y=1/X)}{\partial X} = P(1-P)B_1$ điều này được diễn dịch là tác động biên của X_1 lên xác suất Y nhận giá trị bằng 1 phụ thuộc vào giá trị của X. Tác động biên của X_1 lên khả năng Y =1 xác định với xác suất ban đầu = 0,5

Phương trình Binary Logistic được SPSS chuyển đổi ngược trở lại như sau:

$$\frac{P(Y=1)}{P(Y=0)} = e^{B_o+B_1X}$$

Chương trình SPSS sẽ tự động thực hiện việc tính toán các hệ số cho bạn và cho hiện cả hệ số thật lẫn hệ số đã được chuyển đổi.

Với ví dụ thực tế được trình bày phía sau, các bạn sẽ dễ hình dung cách diễn dịch các hệ số này hơn.

2.2. Độ phù hợp của mô hình

Hồi qui Binary Logistic cũng đòi hỏi ta phải đánh giá độ phù hợp của mô hình. Đo lường độ phù hợp tổng quát của mô hình Binary Logistic được dựa trên chỉ tiêu -2LL (viết tắt của –2 log likelihood), thước đo này có ý nghĩa giống như SSE (Sum of squares of error) nghĩa là càng nhỏ càng tốt. Bạn không cần quan tâm nhiều đến việc –2LL tính toán như thế nào nhưng nhớ rằng quy tắc đánh giá độ phù hợp căn cứ trên –2LL ngược với quy tắc dựa trên hệ số xác định mô hình R^2, nghĩa là giá trị –2LL càng nhỏ càng thể hiện độ phù hợp cao. Giá trị nhỏ nhất của –2LL là 0 (tức là không có sai số) khi đó mô hình có một độ phù hợp hoàn hảo.

Chúng ta cũng còn có thể xác định được mô hình dự đoán tốt đến đâu qua bảng phân loại (Clasification table) do SPSS đưa ra, bảng này sẽ so sánh số trị số thực và trị số dự đoán cho từng biểu hiện và tính tỷ lệ dự đoán đúng sự kiện.

2.3. Kiểm định ý nghĩa của các hệ số

Hồi qui Binary Logistic cũng đòi hỏi kiểm định giả thuyết hệ số hồi qui khác không. Bạn hình dung nếu hệ số hồi qui B_o và B_1 đều bằng 0 thì tỷ lệ chênh lệch giữa các xác suất sẽ bằng 1, tức xác suất để sự kiện xảy ra hay không xảy ra như nhau, lúc đó mô hình hồi qui của chúng ta vô dụng trong việc dự đoán.

Trong hồi qui tuyến tính chúng ta sử dụng kiểm định t để kiểm định giả thuyết Ho: $\beta_k=0$. Còn với hồi qui Binary Logistic, đại lượng Wald Chi Square được sử dụng để kiểm định ý nghĩa thống kê của hệ số

hồi qui tổng thể. Cách thức sử dụng mức ý nghĩa Sig. cho kiểm định Wald cũng theo quy tắc thông thường. Wald Chi Square được tính bằng cách lấy ước lượng của hệ số hồi qui của biến độc lập trong mô hình (hệ số hồi qui mẫu) binary logistic chia cho sai số chuẩn của ước lượng hệ số hồi qui này, sau đó bình phương lên theo công thức sau:

$$\text{Wald Chi - Square} = \left(\frac{\hat{\beta}}{s.e.(\hat{\beta})}\right)^2 = \left(\frac{B}{s.e.(B)}\right)^2$$

2.4. Kiểm định độ phù hợp tổng quát

Ở hồi qui Binary Logistic, tổ hợp liên hệ tuyến tính của toàn bộ các hệ số trong mô hình ngoại trừ hằng số cũng được kiểm định xem có thực sự có ý nghĩa trong việc giải thích cho biến phụ thuộc không. Với hồi qui tuyến tính bội ta dùng thống kê F để kiểm định giả thuyết H_o: $\beta_1 = \beta_2 = ... = \beta_k = 0$, còn với hồi qui Binary Logistic ta dùng kiểm định Chi-bình phương. Bạn sẽ căn cứ vào mức ý nghĩa quan sát mà SPSS đưa ra trong bảng Omnibus Tests of Model Coefficients để quyết định bác bỏ hay chấp nhận Ho.

2.5. Các phương pháp đưa biến độc lập vào mô hình hồi qui Binary Logistic

Với phương pháp hồi qui từng bước (Stepwise), số thống kê được sử dụng cho các biến được đưa vào và dời ra căn cứ trên số thống kê likelihood-ratio (tỷ lệ thích hợp) hay số thống kê Wald.

Bạn cũng có thể chọn một trong các phương pháp thay thế sau

- Enter: đưa vào bắt buộc, các biến trong khối biến độc lập được đưa vào trong một bước
- Forward: Conditional là phương pháp đưa vào dần theo điều kiện. Nó kiểm tra việc loại biến căn cứ trên xác suất của số thống kê Likelihood-ratio dựa trên những ước lượng thông số có điều kiện
- Forward: LR là phương pháp đưa vào dần kiểm tra việc loại biến căn cứ trên xác suất của số thống kê Likelihood-ratio dựa trên ước lượng khả năng xảy ra tối đa (maximum-likelihood estimates)

- Forward:Wald là phương pháp đưa vào dần kiểm tra việc loại biến căn cứ trên xác suất của số thống kê Wald.
- Backwald: Conditional là phương pháp loại trừ dần theo điều kiện. Nó kiểm tra việc loại biến căn cứ trên xác suất của số thống kê Likelihood-ratio dựa trên những ước lượng thông số có điều kiện
- Backwald:LR là phương pháp loại trừ dần kiểm tra loại biến căn cứ trên xác suất của số thống kê Likelihood-ratio dựa trên những ước lượng khả năng xảy ra tối đa.
- Backwald:Wald là phương pháp loại trừ dần kiểm tra loại biến căn cứ trên xác suất của số thống kê Wald.

3. TIẾN HÀNH PHÂN TÍCH HỒI QUI BINARY LOGISTIC VỚI SPSS

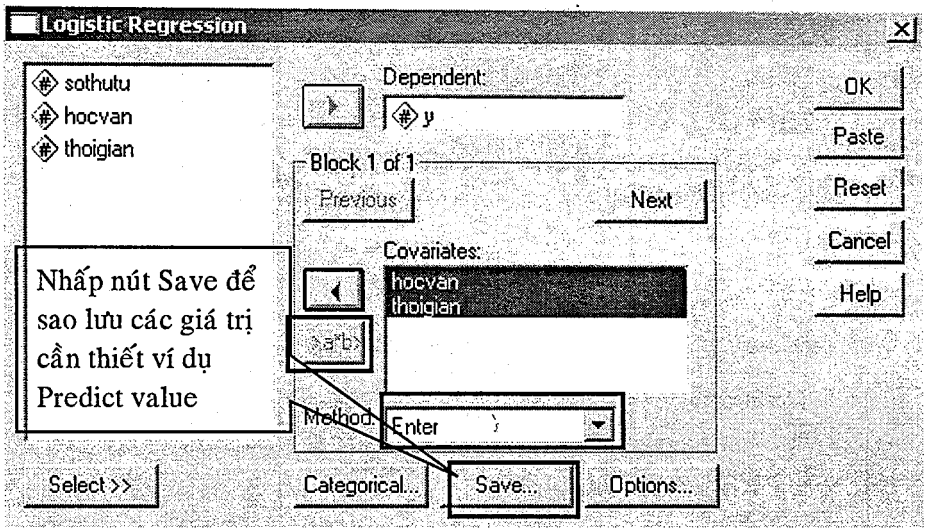
Giả sử chúng ta nghiên cứu về khả năng thu hồi nợ vay của một chương trình xoá đói giảm nghèo, vấn đề ta quan tâm là đối tượng nào nên cho vay và đối tượng nào không nên cho vay dựa trên một lập luận cho rằng khả năng trả nợ có liên hệ với trình độ học vấn của đối tượng vay nợ và thời gian đối tượng đó cư trú trên địa bàn. Do đó biết được 2 thông tin này ta có thể sử dụng mô hình Binary Logistic để dự đoán khả năng đối tượng trả được nợ, từ đó để quyết định có nên cho vay hay không.

Thu thập thông tin về trình độ học vấn và thời gian cư trú của 50 đối tượng đã từng đến vay của chương trình, cùng với kết quả cuối cùng của hợp đồng vay nợ là họ trả được hay không trả được. Các dữ liệu này được lưu với tên file *Binary Logistic* trong tập hợp dữ liệu dùng kèm với sách này, thời gian nhập cư được thu thập theo số tháng. Biến phụ thuộc của ta sẽ có hai giá trị là 1 và 0 đại diện cho hai biểu hiện trả được nợ và không trả được nợ.

3.1. Tiến trình thực hiện

1. Tại của sổ dữ liệu của file *Binary Logistic* bạn chọn menu: Analyze > Regression > Binary Logistic, lựa chọn này mở ra hộp thoại Logistic Regression.

Hình 10.1



2. Chọn biến phụ thuộc (y) đưa sang khung Dependent, nhớ chỉ chọn biến có 2 biểu hiện, nếu biến phụ thuộc bạn chọn không có đúng 2 biểu hiện thì thủ tục này không thực hiện được.

3. Chọn một biến hay một khối biến (block) đưa sang khung Covariate. Nếu muốn tạo biến dạng tương tác thì bạn chọn sáng 2 (hay hơn 2) biến của mối tương tác trong danh sách biến nguồn và nhấp nút >a*b> đưa sang khung Covariate.

4. Trong nút Method bạn chọn các phương pháp đưa biến độc lập vào mô hình, ở đây ta để chế độ mặc định là Enter (xem Hình 10.1)

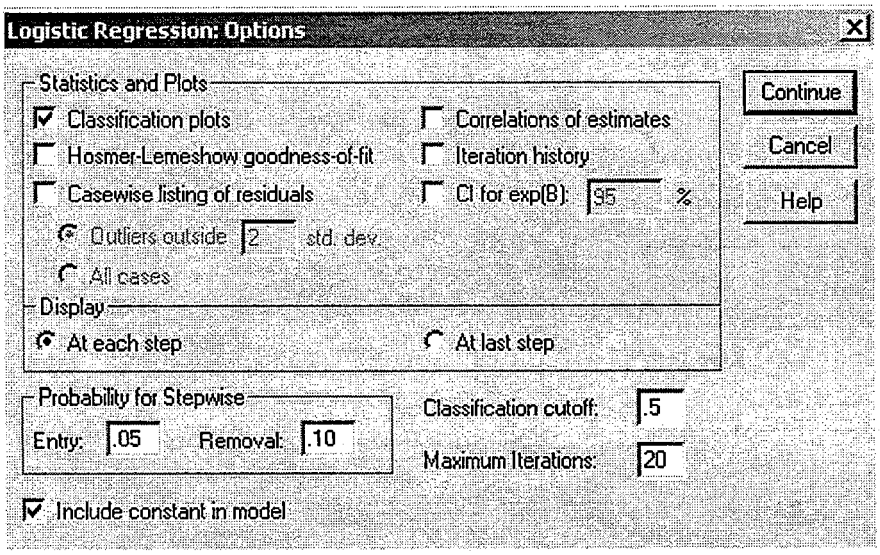
5. Để hiện đồ thị phân loại giá trị thật và giá trị dự báo của biến phụ thuộc, bạn nhấp nút Option để mở hộp thoại Logistic Regression: Options, rồi nhấp chọn Classification plots trong phần Statistics and Plots. Một số tùy chọn khác trên hộp thoại này bạn đọc có thể suy diễn từ hướng dẫn ở phần hồi qui tuyến tính (xem Hình 10.2)

6. Nhấp Continue trở về hộp thoại đầu tiên.

7. Muốn tính được giá trị dự đoán, là xác suất mà một đối tượng sẽ trả nợ ta nhấp Predict value trong hộp thoại save.

8. Sau cùng nhấp OK.

Hình 10.2



Bạn sẽ có hàng loạt bảng kết quả, những bảng đầu tiên thể hiện các thông số thống kê chung về tập tin dữ liệu, nhưng mối quan tâm của chúng ta là các bảng từ Bảng 10.1đến Bảng 10.4 và Hình 10.3

Bảng 10.1 Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	45.121	2	.000
	Block	45.121	2	.000
	Model	45.121	2	.000

Bảng 10.2 Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	23.472	.594	.796

Bảng 10.3 Classification Table(a)

Observed			Predicted		
			tra duoc von vay		Percentage Correct
			khong tra	tra duoc	
Step 1	tra duoc von vay	khong tra	20	2	90.9
		tra duoc	2	26	92.9
	Overall Percentage				92.0

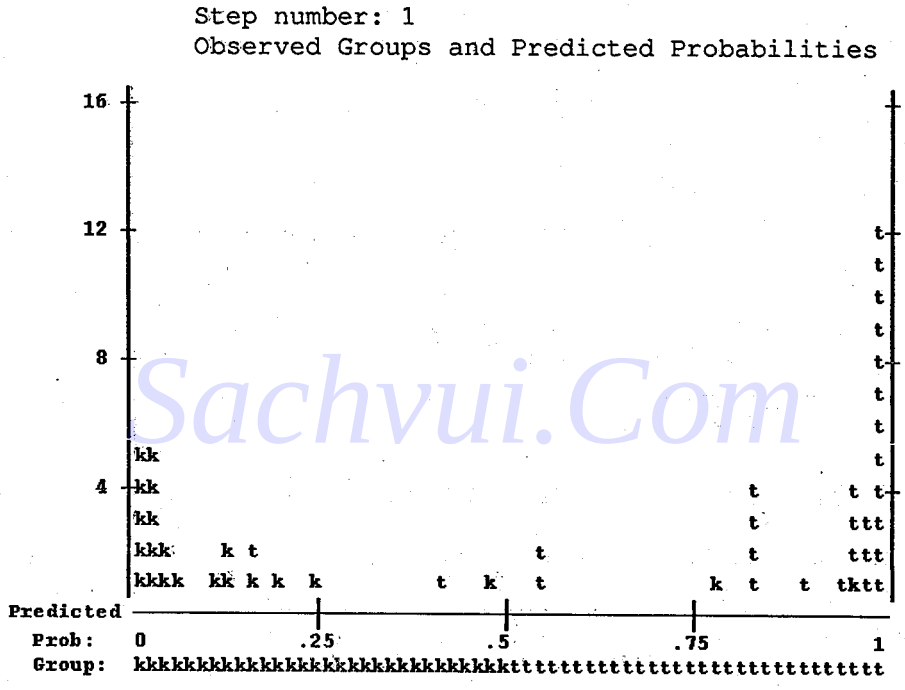
a The cut value is .500

Bảng 10.4 Variables in the Equation

	Step 1(a)		
	HOCVAN	THOIGIAN	Constant
B	.450	.270	-9.003
S.E.	.176	.085	2.556
Wald	6.493	10.143	12.404
df	1	1	1
Sig.	.011	.001	.000
Exp(B)	1.568	1.310	.000

a Variable(s) entered on step 1: HOCVAN, THOIGIAN.

Hình 10.3



3.2. Ý nghĩa của các kết quả

Kết quả kiểm định giả thuyết về độ phù hợp tổng quát ở Bảng 10.1 có mức ý nghĩa quan sát sig. =0,000 nên ta an toàn bác bỏ giả thuyết Ho: $\beta_{hocvan} = \beta_{thoigian} = 0$

Bảng 10.2 cho thấy giá trị của –2LL = 23,472 không cao lắm, như vậy nó thể hiện một độ phù hợp khá tốt của mô hình tổng thể.

Mức độ chính xác của dự báo cũng thể hiện qua bảng Classification Table (Bảng 10.3), bảng này cho thấy trong 22 trường hợp được dự đoán là không trả được nợ (xem theo cột) mô hình đã dự đoán trúng 20 trường hợp (xem theo hàng), vậy tỷ lệ trúng là 91%. Còn với 28 trường hợp thực tế có trả nợ mô hình lại dự đoán sai 2 trường hợp (tức là cho rằng họ không trả) tỷ lệ trúng giờ là 93%. Từ đó ta tính được tỷ lệ dự đoán đúng của toàn bộ mô hình là 92%.

Ở Bảng 10.4, kiểm định Wald về ý nghĩa của các hệ số hồi qui tổng thể của biến học vấn và thời gian cư trú đều có mức ý nghĩa sig. nhỏ hơn 0,05 nên ta an toàn bác bỏ giả thuyết

H_0: $\beta_{hocvan}=0$

H_0: $\beta_{thoigian}=0$

Như vậy các hệ số hồi qui tìm được có ý nghĩa và mô hình của chúng ta sử dụng tốt.

Từ các hệ số hồi qui này ta viết được phương trình

$$\log_e\left[\frac{P(Y=1)}{P(Y=0)}\right] = -9{,}003 + 0{,}45\,hocvan + 0{,}27\,thoigian \quad (**)$$

Có thể diễn dịch ý nghĩa của các hệ số hồi qui Binary Logistic là:

Thời gian cư trú và học vấn đều làm tăng khả năng trả nợ trong đó học vấn có tác động mạnh hơn. Cụ thể tác động biên của học vấn lên khả năng trả được nợ xác định với xác suất ban đầu = 0,5 thì tác động này bằng 0,5(1-0,5)0,45 = 0,1125; còn thời gian cư trú có tác động biên là 0,0675

Đồ thị Histogram ở Hình 10.3 biểu diễn các điểm thực tế và dự báo của biến phụ thuộc Y. Bạn thấy trên trục tung có một điểm phân cách là 0,5, tên của điểm này là Cut Value (trị số phân biệt). Từ 0,5 đến 0 là những trường hợp quan sát không trả nợ và từ 0,5 đến 1 là có trả. Trong phạm vi đồ thị phía trên những quan sát không trả nợ bạn sẽ thấy 2 chữ t lạc giữa những chữ k, đó chính là 2 trường hợp dự báo sai tức là cho rằng không trả mà cuối cùng có trả. Xem xét phía đối xứng qua mốc 0,5 bạn cũng thấy 2 chữ k giữa các chữ t.

4. VẬN DỤNG MÔ HÌNH HỒI QUI BINARY LOGISTIC CHO MỤC ĐÍCH DỰ BÁO

Giả dụ có một đối tượng đến xin vay, bạn hỏi họ các thông tin về thời gian cư trú, trình độ học vấn, thế các giá trị này vào mô hình của hàm Binary Logistic để xem xác suất trả nợ của đối tượng nhỏ hay lớn hơn 0,5 mà quyết định có nên cho vay không. Ví dụ với một người học vấn lớp 9, cư trú liên tục trên địa bàn đã 2 năm thì xác suất trả nợ của họ sẽ là

$$E(Y/X) = \frac{e^{(-9,003+0,45*9+0,27*24)}}{1+e^{(-9,003+0,45*9+0,27*24)}} = \frac{4,6}{1+4,6} = 0,821$$

Mô hình Binary Logistic cho biết khả năng trả nợ của người này tới 82%, như vậy chương trình có thể cho họ vay vì khả năng thu hồi được nợ cao. Nhưng bạn cũng đừng quên, đây là khả năng trả nợ được dự đoán, và dự đoán này có khả năng đúng chỉ có 92%.

5. SO SÁNH GIỮA HỒI QUI TUYẾN TÍNH THÔNG THƯỜNG VÀ HỒI QUI BINARY LOGISTIC

Điểm khác biệt cơ bản giữa hồi qui Binary Logistic và hồi qui bội thông thường là ở chỗ biến phụ thuộc là dạng nhị phân. Hồi qui thông thường đòi hỏi biến phụ thuộc ở dạng định lượng, còn dữ liệu định danh và phân loại chỉ có thể đưa vào làm biến độc lập dưới hình thức các biến giả (biến nhị phân)

Hồi qui Binary Logistic tương tự như hồi qui bội ở các kết quả nhưng nó khác nhau trong cách ước lượng các hệ số, thay vì tối hiểu hoá độ lệch bình phương (thủ tục OLS) như hồi qui tuyến tính thì nó tối đa hoá khả năng một hiện tượng có thể xảy ra với tên gọi ước lượng với khả năng xảy ra tối đa (ước lượng thích hợp cực đại - Maximum Likelihood Estimation) tuy nhiên chúng khá tương đồng về việc kiểm định độ phù hợp của mô hình và kiểm định ý nghĩa của các hệ số hồi qui.

Các thủ tục đưa biến vào ra khỏi mô hình của 2 dạng hồi qui này cũng khá tương tự nhau.

CHƯƠNG XI
ĐÁNH GIÁ ĐỘ TIN CẬY CỦA THANG ĐO

1. GIỚI THIỆU

Khi thực hiện các nghiên cứu định lượng, người nghiên cứu phải sử dụng các loại thang đo lường khác nhau. Hiện tượng kinh tế – xã hội vốn rất phức tạp nên việc lượng hóa các khái niệm nghiên cứu đòi hỏi phải có những thang đo lường được xây dựng công phu và được kiểm tra độ tin cậy trước khi vận dụng. Ví dụ như việc đo lường chất lượng cuộc sống, chất lượng dịch vụ, quan niệm sống ... không thể chỉ sử dụng những thang đo đơn giản (chỉ dùng 1 câu hỏi đo lường – gọi là thang đo một chỉ báo) mà phải sử dụng các thang đo chi tiết hơn (thang đo nhiều chỉ báo) mới có thể nắm bắt được những nội dung phong phú của các khái niệm này. Những chỉ báo khác nhau khi đo lường giúp thể hiện những khía cạnh (chiều kích – dimension) khác nhau của khái niệm muốn đo lường. Chương này sẽ thảo luận về việc xây dựng thang đo nhiều chỉ báo và kiểm tra độ tin cậy của thang đo lường được sử dụng.

2. THANG ĐO NHIỀU CHỈ BÁO

Một trong những hình thức đo lường các khái niệm trừu tượng được sử dụng phổ biến nhất trong nghiên cứu kinh tế xã hội là thang đo do Rennis Likert (1932) giới thiệu. Likert đã đưa ra loại thang đo năm mức độ phổ biến. Câu hỏi điển hình của dạng thang đo Likert này là: "Xin vui lòng đọc kỹ những phát biểu sau. Sau mỗi câu phát biểu, hãy khoanh tròn trả lời thể hiện đúng nhất quan điểm của bạn. Xin cho biết rằng bạn rất đồng ý, đồng ý, thấy bình thường, không đồng ý hay rất không đồng ý với mỗi phát biểu?"

Thang đo 5 mức độ có thể trở thành 3 hoặc 7 mức độ và đồng ý hay không đồng ý, và cũng có thể trở thành chấp nhận hay không chấp nhận, có thiện ý hay phản đối, tuyệt vời hay tồi tệ, nhưng quy tắc là như nhau. Tất cả đều được gọi là thang đo Likert.

Các khái niệm trong nghiên cứu kinh tế xã hội hầu hết đều là mang tính đa khía cạnh (hay còn gọi là đa chiều, đa thành phần). Ví dụ như khái niệm chất lượng dịch vụ ngân hàng, khách hàng có thể cho rằng chất lượng dịch vụ ngân hàng của một ngân hàng cụ thể mà họ giao dịch thể hiện ở chỗ thủ tục thực hiện các dịch vụ ngân hàng rườm rà hay đơn giản, thái độ phục vụ của nhân viên ân cần hay coi thường khách hàng, cơ sở vật chất hiện đại hay đơn giản .... Chúng ta phải hỏi các khách hàng đánh giá của họ về nhiều khía cạnh nêu trên chứ không thể chỉ hỏi bằng một câu hỏi đơn giản.

3. CÁC BƯỚC XÂY DỰNG THANG ĐO LIKERT

Phương pháp của Likert là lên một danh sách các mục có thể đo lường cho một khái niệm và tìm ra những tập hợp các mục hỏi để đo lường tốt các khía cạnh khác nhau của khái niệm. Nếu như khái niệm mang tính đơn khía cạnh thì chỉ cần tìm ra một tập hợp. Nếu khái niệm đó là đa khía cạnh thì cần nhiều tập hợp các mục hỏi. Sau đây là các bước xây dựng và kiểm tra một thang đo Likert.

1- Nhận diện và đặt tên biến mà bạn muốn đo lường. Bạn có thể làm được điều này thông qua kinh nghiệm của bản thân. Giả dụ sau một thời gian quan sát và thăm hỏi những người khách hàng của các ngân hàng, bạn sẽ hình thành những ý niệm về các biến mà bạn muốn đo lường.

2- Lập ra một danh sách các phát biểu hoặc câu hỏi có tính biểu thị. Các ý tưởng cho các câu hỏi biểu thị có thể lấy từ lý thuyết của các môn học marketing, đọc sách báo hoặc từ ý kiến của các chuyên gia. Các câu hỏi biểu thị này cũng có thể lấy từ các thực nghiệm. Nếu bạn muốn xây dựng một công cụ đo lường cho biến "thái độ phục vụ khách hàng", bạn có thể bắt đầu bằng cách hỏi một nhóm khách hàng để "liệt kê những điều liên quan đến vấn đề phục vụ của nhân viên". Bạn có thể xây dựng các câu hỏi hay phát biểu trong thang đo Likert theo các mục trong danh sách này.

   Bạn phải đảm bảo cho các mục hỏi này theo cả hai chiều thuận và nghịch đối với vấn đề đặt ra. Nếu bạn có phát biểu: "Tôi cảm thấy thoải mái khi giao dịch với các nhân viên ngân hàng" thì sau đó bạn cần một câu phát biểu có ý phủ định cho cân bằng như sau" Nhân viên ngân hàng làm cho tôi ngại đến các ngân hàng".

Trong việc soạn các mục hỏi, những chú ý đối với thiết kế bảng câu hỏi cần được tuân thủ : Cần nhớ những người bạn sẽ phỏng vấn là ai và nên sử dụng ngôn ngữ của họ. Thiết kế những câu phát biểu càng ngắn và càng đơn giản càng tốt. Không dùng những câu phủ định hai lần. Không hỏi những câu hỏi có hai ý. Ví dụ "Các nhân viên ngân hàng có thái độ ân cần và tinh thông nghiệp vụ" là một mục hỏi tồi vì một người khách hàng được hỏi có thể đồng ý với cả hai vế của phát biểu, hoặc chỉ đồng ý với một vế và phản đối vế còn lại.

Số lượng các mục hỏi khi bạn xây dựng phải gấp bốn đến năm lần số lượng các mục hỏi bạn sẽ cần trong thang đo cuối cùng. Nếu bạn cần một thang đo với sáu mục bạn phải xây dựng từ 25 đến 30 mục trong lần kiểm tra đầu tiên.

3- Xác định số lượng và loại trả lời. Một vài các loại trả lời phổ biến như là: đồng ý - không đồng ý, ủng hộ - phản đối, hữu ích - vô ích, nhiều - không có, giống tôi-không giống tôi, đúng - không đúng, phù hợp - không phù hợp, luôn luôn - không bao giờ, và v..v. Hầu hết các thang đo của Likert có số lượng lẻ các lựa chọn trả lời như : 3, 5 hoặc 7. Mục đích là để đưa ra cho người trả lời một loạt các lựa chọn trả lời có điểm giữa. Điểm giữa thường mang tính trung lập, ví dụ như không đồng ý cũng không phản đối. Số lựa chọn chẵn buộc người trả lời phải xác định một quan điểm rõ ràng trong khi số lựa chọn lẻ cho phép họ lựa chọn an toàn hơn. Không thể nói cái nào là hay hơn vì cách lựa chọn nào cũng có hệ quả riêng của nó.

4- Kiểm tra toàn bộ các mục hỏi đã khai thác được từ những người trả lời. Lý tưởng thì bạn cần ít nhất 100 người trả lời để kiểm tra các mục hỏi ban đầu. Điều này đảm bảo rằng bạn đã nắm bắt được đầy đủ các khác biệt về trả lời đối với toàn bộ các mục hỏi bạn đề ra. Nếu bạn có thể chọn 100 đến 200 người trả lời một cách ngẫu nhiên, bạn có thể đảm bảo là sự đa dạng của các trả lời trong mẫu này đại diện được cho sự đa dạng trong tổng thể chung mà thực sự đây mới là mục tiêu chính bạn muốn đo lường.

5- Thực hiện một phân tích mục hỏi để tìm ra một tập hợp các mục hỏi tạo nên một thang đo đơn khía cạnh về biến mà bạn muốn đo lường.

6- Sử dụng thang đo mà bạn đã xây dựng được trong nghiên cứu của bạn và tiến hành phân tích lại các mục hỏi lại lần nữa để đảm bảo

rằng thang đo đó là chắc chắn. Nếu làm xong điều này, thì sau đó đi tìm mối quan hệ giữa những điểm số từ thang đo và điểm số từ những biến khác cho các cá nhân trong nghiên cứu của bạn.

3.1. Phân tích các mục hỏi

Đây là chìa khoá để xây dựng thang đo. Mục đích là tìm ra những mục hỏi cần giữ lại và những mục hỏi cần bỏ đi trong rất nhiều mục bạn đưa vào kiểm tra. Tập hợp các mục hỏi mà bạn giữ lại chỉ nên thể hiện một khía cạnh kinh tế xã hội hoặc tâm lý đơn. Nói cách khác, thang đo nên là đơn khía cạnh.

Những trang kế tiếp tóm tắt nguyên tắc xây dựng thang đo đơn khía cạnh. Có ba bước để phân tích các mục hỏi và tìm ra một tập hợp các mục hỏi cấu thành một thang đo đơn khía cạnh: (a) tính điểm các mục (b) kiểm tra mức độ tương quan giữa các mục, và (c) kiểm tra mức độ tương quan giữa tổng điểm của từng người và điểm của từng mục hỏi.

3.2. Xây dựng thang đo đơn khía cạnh

3.2.1 Tính điểm các trả lời

Đầu tiên là chắc chắc rằng tất cả các mục hỏi được ghi số trả lời hợp lý. Giả sử chúng ta đang tìm kiếm các mục hỏi để đo lường mức độ cần thiết của môn học Hành vi người tiêu dùng cho các sinh viên ngành kinh tế học. Sau đây là hai câu hỏi đo lường có thể chọn:

Cần phải đào tạo về Hành vi người tiêu dùng cho tất cả các sinh viên ngành kinh tế học:

1	2	3	4	5
Rất không đồng ý	Không đồng ý	Bình thường	Đồng ý	Rất đồng ý

Các nhà kinh tế học không cần phải được trang bị kiến thức về Hành vi người tiêu dùng:

1	2	3	4	5
Rất không đồng ý	Không đồng ý	Bình thường	Đồng ý	Rất đồng ý

Khi bạn tiến hành đánh số cho các trả lời của người trả lời, bạn cần phải nhớ là số 1 trên mục hỏi đầu tiên chính là số 5 cho mục hỏi thứ

2 và ngược lại. Những người trả lời mà chọn "rất đồng ý" trên mục hỏi đầu tiên thì sẽ ghi 5 điểm trên mục hỏi đó. Những người trả lời chọn "rất đồng ý" cho mục hỏi số 2 thì sẽ ghi 1 điểm. Bạn có thể đặt số lớn hay nhỏ trên thước đo theo hướng nào mà bạn muốn nhưng bạn phải nhất quán. Trong trường hợp này, chúng ta quyết định lấy số lớn hơn (4 hoặc 5) để tượng trưng cho sự cần thiết của môn học và những số nhỏ hơn (1 và 2) tượng trưng cho sự không cần thiết.

3.2.2 Kiểm tra sự tương quan giữa các mục hỏi và tính toán Cronbach alpha

Tiếp theo chúng ta kiểm tra xem các mục hỏi nào đã có đóng góp vào việc đo lường một khái niệm lý thuyết mà ta đang nghiên cứu, và những mục hỏi nào không. Điều này liên quan đến hai phép tính toán: tương quan giữa bản thân các mục hỏi và tương quan của các điểm số của từng mục hỏi với điểm số toàn bộ các mục hỏi cho mỗi người trả lời. Đây là điểm số của 3 người trên 3 mục hỏi

	Mục hỏi 1	Mục hỏi 2	Mục hỏi 3
Người 1	1	3	5
Người 2	5	2	2
Người 3	4	1	3

Để tìm ra sự tương quan giữa nội bộ các mục hỏi, chúng ta cần nhìn vào các cặp có thể tạo thành bởi các cột. Có 3 cặp cột sẽ lập thành bảng ma trận 3 mục hỏi:

Mục hỏi 1	Mục hỏi 2	Mục hỏi 1	Mục hỏi 3	Mục hỏi 2	Mục hỏi 3
1	3	1	5	3	5
5	2	5	2	2	2
4	1	4	3	1	3

Một phép đo lường đơn giản về mức độ các cặp số này giống hoặc không giống nhau là cộng các chênh lệch giữa chúng. Trong cặp cột thứ nhất, hiệu số giữa 1 và 3 là 2, hiệu số giữa 5 và 2 là 3, hiệu số giữa 4 và 1 là 3. Tổng các chênh lệch là 2+3+3 = 8. Đối với mỗi mục hỏi, chênh lệch lớn nhất có thể là 4, ví dụ người nào đó có thể trả lời số 1 cho mục hỏi thứ nhất và số 5 cho mục hỏi thứ 2.

Đối với 3 người trả lời thì tổng của các chênh lệch có thể là: 4 x 3 = 12. Chênh lệch thực sự chính là 8/12=0,67, điều này có nghĩa là giữa 2 mục hỏi này có 0,33 là giống nhau. Giữa mục hỏi 1 và 3 cũng có 0,33 giống và mục hỏi 2 và 3 là 0,67 là giống nhau.

Những mục hỏi đo lường cùng một khái niệm tiềm ẩn thì phải có mối liên quan với những cái còn lại trong nhóm đó. Nếu tôi trả lời "rất đồng ý" đối với câu " Cần phải đào tạo về Hành vi người tiêu dùng cho các sinh viên ngành kinh tế học" thì (nếu tôi giữ thái độ nhất quán và nếu mục hỏi khảo sát quan điểm của tôi được xây dựng hợp lý) tôi phải "rất không đồng ý" với câu "Các nhà kinh tế học không cần phải được trang bị kiến thức về Hành vi người tiêu dùng". Nếu mọi người trả lời "rất đồng ý" đối với câu nói thứ nhất và rất "không đồng ý" đối với câu nói thứ hai thì các mục hỏi là có tương quan hoàn hảo.

Hệ số α của Cronbach là một phép kiểm định thống kê về mức độ chặt chẽ mà các mục hỏi trong thang đo tương quan với nhau. Một trong những phương pháp kiểm tra tính đơn khía cạnh của thang đo được gọi là kiểm định độ tin cậy chia đôi. Nếu một thang đo gồm 10 mục hỏi và là đơn khía cạnh, tất cả những mục hỏi sẽ đo lường các phần khác nhau của cùng một khái niệm cơ bản. Trong trường hợp đó 5 mục hỏi có thể cho ra một số điểm ít hoặc nhiều hơn số điểm của 5 mục hỏi khác giống như sau:

	Điểm số trên mục hỏi 1-5	Điểm số trên mục hỏi 6-10
Người 1	X_1	Y_1
Người 2	X_2	Y_2
Người 3	X_3	Y_3
. . .		
Người N	X_n	Y_n
Tổng cộng	A	B

Có rất nhiều cách để chia đôi một nhóm mục hỏi, và mỗi phần chia đôi sẽ cho bạn một tập hợp khác nhau của các tổng số. Mặc dù vậy, về trung bình, tổng số của tất cả những kiểm định chia đôi có thể sẽ khá giống nhau. Hệ số α của Cronbach có thể kiểm tra điều này.

Công thức của hệ số Cronbach α là

$$\alpha = N\rho/[1 + \rho(N - 1)]$$

Trong đó ρ là hệ số tương quan trung bình giữa các mục hỏi. Ký tự Hy Lạp ρ (đọc là prô) trong công thức tượng trưng cho tương quan trung bình giữa tất cả các cặp mục hỏi được kiểm tra.

Theo quy ước thì một tập hợp các mục hỏi dùng để đo lường được đánh giá là tốt phải có hệ số α lớn hơn hoặc bằng 0,8. Mặc dù vậy cần chú ý rằng nếu bạn có một danh mục quá nhiều các mục hỏi (N là số mục hỏi) thì sẽ có nhiều cơ hội để có được hệ số α cao. Sự tương quan giữa các mục hỏi chỉ là 0.14 đã có được hệ số α = 0,8 trong tập hợp 25 mục hỏi (De Vellis, 1991).

Cuối cùng, bạn muốn đạt được hệ số α lớn hơn hoặc bằng 0,8 cho một danh mục ít các mục hỏi mà các mục hỏi này đi liền với nhau một cách mạch lạc và đo lường cùng một vấn đề. Hệ số α của Cronbach sẽ cho bạn biết các đo lường của bạn có liên kết với nhau hay không nhưng nó sẽ không cho bạn biết mục hỏi nào cần được bỏ đi và mục hỏi nào cần được giữ lại. Để làm được điều này bạn cần phải xác định mục hỏi nào không phân biệt giữa những người cho điểm số lớn và những người cho điểm số nhỏ trong tập hợp toàn bộ các mục hỏi.

3.2.3 Kiểm tra tương quan giữa tổng điểm của từng người và điểm của từng mục hỏi.

Đầu tiên tìm tổng số điểm cho mỗi người. Cộng dồn số điểm của từng người trả lời theo tất cả các mục hỏi. Giả sử rằng có 20 mục hỏi và bạn kiểm tra các mục hỏi đó trên 100 người. Dữ liệu sẽ giống như bảng mô tả sau đây.

Người	Mục hỏi 1	Mục hỏi 2	Mục hỏi 3	...	Mục hỏi 20
1	x	x	x		x
2	x	x	x		x
3	x	x	x		x
.					
.					
.					
100	x	x	x		x

Trong đó x là điểm số của mỗi người cho mỗi mục hỏi. Đối với 20 mục hỏi, điểm số từ 1 đến 5, mỗi người có thể lấy số điểm thấp nhất là 20 (bằng cách ghi điểm 1 clío mỗi mục hỏi) hoặc là cao đến 100 (lấy điểm 5 cho mỗi mục hỏi). Trong thực tế, dĩ nhiên mỗi người trả lời trong một cuộc khảo sát sẽ đạt được một tổng điểm nào đó trong khoảng này. Cách đơn giản để tìm ra những mục hỏi phân biệt tốt những người trả lời là chia những người trả lời thành 2 nhóm, 25% với tổng số điểm cao nhất và 25% với tổng số điểm thấp nhất. Tìm ra những mục hỏi nào mà có mặt trong cả hai nhóm. Những mục hỏi đó không phân biệt được giữa những người trả lời theo khái niệm cần kiểm tra. Ví dụ, các mục hỏi không đạt trong việc phân biệt giữa những người có thiện cảm nhiều đối với phương pháp đào tạo (25% điểm số cao nhất) và những người rất không thiện cảm (25% số điểm thấp nhất) là những mục hỏi không đạt để đo lường, nên loại bỏ chúng đi.

Có thêm một cách để tìm ra những mục hỏi phân biệt tốt giữa những những người trả lời và những mục hỏi không phân biệt tốt. Đó là tương quan giữa tổng số điểm và điểm của từng mục hỏi. Đây là dữ liệu bạn cần cho việc này:

	Tổng điểm	Mục hỏi 1	Mục hỏi 2	Mục hỏi 3
Người 1				
Người 2				
Người 3				
. . .				

Với 20 mục hỏi, tổng điểm cho bạn một ý niệm về quan điểm của mỗi người đối với khái niệm mà bạn đang tiến hành đo lường. Nếu tương quan giữa các mục hỏi là hoàn hảo, thì mọi mục hỏi đóng góp bằng nhau cho sự hiểu biết của chúng ta về quan điểm của mỗi người trả lời. Dĩ nhiên, sẽ có một số mục hỏi thực hiện điều này tốt hơn các mục hỏi khác. Những mục hỏi không đóng góp nhiều sẽ tương quan yếu với tổng số điểm của mỗi người. Hãy giữ lại những mục hỏi có sự tương quan mạnh với tổng số điểm.

Bạn có thể sử dụng một số phần mềm phân tích thống kê như SPSS để tìm ra hệ số tương quan giữa các mục hỏi, hệ số α, và hệ số tương quan giữa tổng điểm và các mục hỏi cho một tập hợp các mục hỏi ban đầu. Mục đích của bạn là loại bỏ các mục hỏi làm giảm sự tương quan giữa các mục hỏi và giữ cho hệ số α lớn hơn hoặc bằng 0,8.

4. TÍNH TOÁN CRONBACH ALPHA VỚI SPSS

Trong phần này chúng ta lấy ví dụ về giá trị dịch vụ đào tạo theo cảm nhận của sinh viên. Theo các lý thuyết tiếp thị thì mức độ hài lòng của sinh viên đối với trường đại học chịu ảnh hưởng của chất lượng đào tạo, giá trị dịch vụ đào tạo và một số yếu tố khác. Người nghiên cứu quan tâm và muốn đo lường xem giá trị dịch vụ hay chất lượng dịch vụ đào tạo cái nào ảnh hưởng mạnh hơn đến mức độ hài lòng của sinh viên. Cho nên cần phải xây dựng thang đo lường đáng tin cậy về khái niệm giá trị dịch vụ đào tạo. Theo các nghiên cứu trước đó thì các nhà nghiên cứu đề xuất giá trị dịch vụ đào tạo gồm có 6 khía cạnh và mỗi khía cạnh bao gồm nhiều mục hỏi. Trong phần này chúng ta chỉ xem xét việc tính toán Cronbach alpha đối với các mục hỏi của 1 khía cạnh của khái niệm giá trị dịch vụ, đó là giá trị chức năng. Câu hỏi liên quan đến đo lường khía cạnh giá trị chức năng của giá trị dịch vụ đào tạo như sau:

Sau đây là những phát biểu liên quan đến việc chọn và học tập của bạn tại trường ĐH Kinh Tế TPHCM. Xin bạn vui lòng trả lời bằng cách khoanh tròn một con số ở từng dòng. Những con số này thể hiện mức độ bạn đồng ý hay không đồng ý đối với các phát biểu theo quy ước như sau:

Rất không đồng ý 1 Không đồng ý 2 Trung lập 3 Đồng ý 4 Rất đồng ý 5

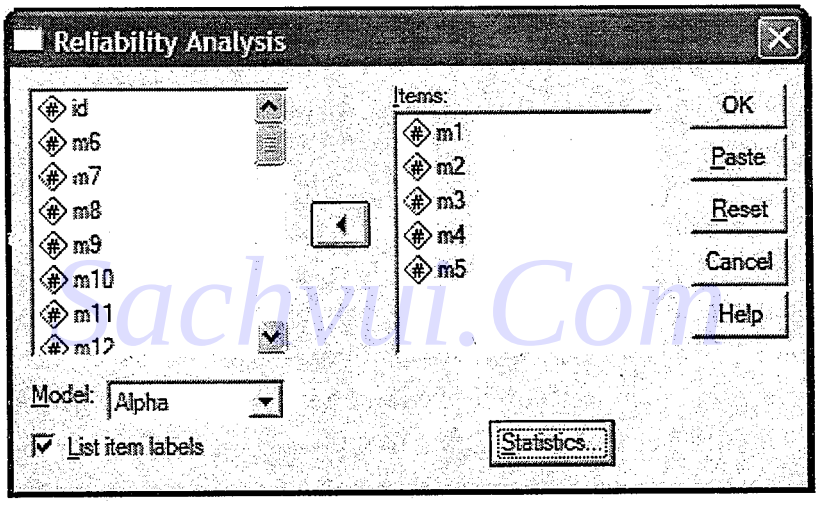
1.	Bằng cấp của trường ĐH Kinh Tế sẽ giúp tôi kiếm được thu nhập cao sau khi ra trường.	1 2 3 4 5
2.	Kiến thức từ trường ĐH Kinh Tế sẽ giúp tôi dễ dàng thăng tiến trong nghề nghiệp.	1 2 3 4 5
3.	Tôi tin rằng các doanh nghiệp rất cần những sinh viên tốt nghiệp từ trường tôi đang học.	1 2 3 4 5
4.	Bằng cấp có được từ trường ĐH Kinh Tế là sự đầu tư tốt của tôi cho tương lai.	1 2 3 4 5
5.	Bằng đại học Kinh Tế bảo đảm việc làm trong tương lai.	1 2 3 4 5

Vấn đề ở đây là thang đo nhiều chỉ báo này có là một thang đo tốt cho một khía cạnh của giá trị dịch vụ đào tạo (giá trị chức năng) không. Để thực hiện việc này chúng ta sẽ tính toán ra đại lượng Cronbach alpha.

Dữ liệu để thực hiện ví dụ này chúng ta sẽ sử dụng file *gia tri dich vu dao tao* trong tập dữ liệu dùng kèm với sách [1].

1. Tại menu của SPSS chọn Analyze > Scale > Reliability Analysis ... , lựa chọn này mở ra hộp thoại Reliability Analysis như Hình 11. 1 sau:

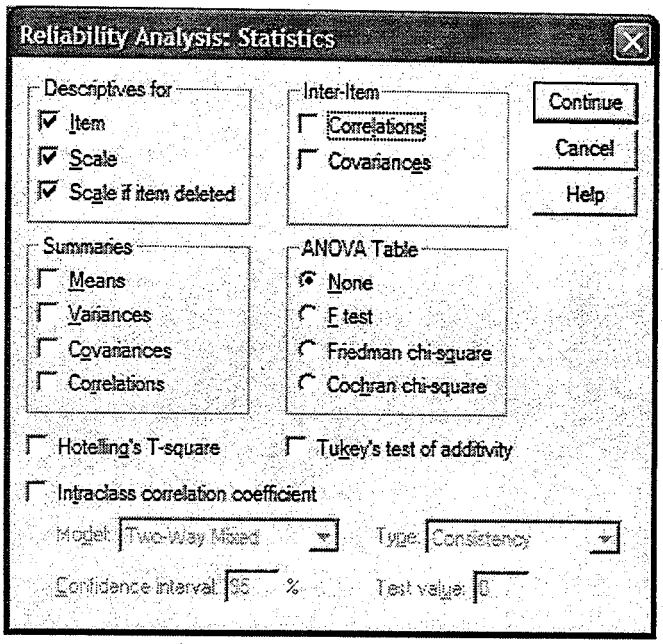
Hình 11.1



Trong hộp thoại này hãy chọn 5 biến đầu tiên từ m1 đến m5 đưa vào ô Items, dùng chuột nhấp chọn List items labels để hiện ra nhãn giải thích tên biến trong kết quả chạy ra. Sau đó nhấp vào nút Statistics để chọn các đại lượng thống kê cần thiết, lệnh này sẽ mở ra hộp thoại như Hình 11.2.

[1] Dữ liệu trích từ Đề tài nghiên cứu khoa học của Hoàng Thị Phương Thảo và Hoàng Trọng, Đại Học Kinh tế TPHCM, năm 2005.

Hình 11.2



Trong hộp thoại Statistics, hãy nhấp chuột để chọn các đại lượng cơ bản nhất là: Item, Scale, và Scale if item deleted như trong hình sau đây. Sau đó nhấp nút Continue trở về hộp thoại đầu tiên rồi nhấp nút OK. Kết quả sẽ xuất hiện trong trong Hình 11.3.

Hình 11.3

```
****** Method 1 (space saver) will be used for this analysis ******

 R E L I A B I L I T Y   A N A L Y S I S  -  S C A L E  (A L P H A)

 1.     M1    Bằng cấp ĐHKT giúp có thu nhập cao
 2.     M2    Kiến thức từ trường ĐHKT giúp thăn
 3.     M3    Các DN cần SV tốt nghiệp trường ĐHKT
 4.     M4    Bằng cấp có được từ ĐHKT là sự đầu tư tốt cho tương lai
 5.     M5    Bằng ĐHKT bảo đảm việc làm trong tương lai

                                Mean        Std Dev     Cases

 1.     M1                     3.1874        .8615      971.0
 2.     M2                     3.4367        .8154      971.0
 3.     M3                     3.7703        .7863      971.0
 4.     M4                     3.5850        .8481      971.0
 5.     M5                     3.2255        .8767      971.0

                                                        N of
Statistics for       Mean     Variance    Std Dev    Variables
     SCALE         17.2049      7.9157     2.8135          5
```

Alpha nếu bỏ đi mục hỏi

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
M1	14.0175	5.2523	.4866	.6299
M2	13.7683	5.5143	.4535	.6446
M3	13.4346	5.7717	.4038	.6646
M4	13.6200	5.2235	.5089	.6202
M5	13.9794	5.4985	.4009	.6678

Reliability Coefficients

N of Cases = 971.0 N of Items = 5

Alpha = .6952

Trong hình trên, Cronbach alpha tính được là 0,6952, gần bằng 0,7. Và nếu nhìn vào cột cuối cùng của đoạn sau của hình thì với các alpha nếu như loại bỏ bớt 1 mục hỏi nào đó (Alpha If item deleted) đều nhỏ hơn thì chúng ta không nên loại bỏ mục hỏi nào.

Nhiều nhà nghiên cứu đồng ý rằng khi Cronbach alpha từ 0,8 trở lên đến gần 1 thì thang đo lường là tốt, từ 0,7 đến gần 0,8 là sử dụng được. Cũng có nhà nghiên cứu đề nghị rằng Cronbach alpha từ 0,6 trở lên là có thể sử dụng được trong trường hợp khái niệm đang đo lường là mới hoặc mới đối với người trả lời trong bối cảnh nghiên cứu (Nunnally [1], 1978; Peterson [2], 1994; Slater [3], 1995).

Thang đo đơn hướng thang đo đa hướng

Việc xây dựng và kiểm tra thang đo dùng trong nghiên cứu hiện nay rất phổ biến, nhất là khi khái niệm hay biến nghiên cứu cần đo lường phức tạp, trừu tượng, có thể được nhiều người hiểu khác

[1] Nunnally, J. (1978), *Psychometric Theory*, New York, McGraw-Hill.

[2] Peterson, R. (1994), "A Meta-Analysis of Cronbach's Coefficient Alpha", *Journal of Consumer Research*, No. 21 Vo.2, pp.38-91.

[3] Slater, S. (1995), "Issues in Conducting Marketing Strategy Research", *Journal of Strategic.*

nhau. Tuy nhiên, khi thực hiện việc đo lường như vậy, người nghiên cứu sẽ có rất nhiều biến quan sát làm cho việc khảo sát liên hệ giữa các biến quan sát trở nên khó khăn. Vả lại việc dùng nhiều câu hỏi đo lường (mục hỏi) để đo lường một khái niệm nghiên cứu hay biến tiềm ẩn nhằm vào việc đo lường chính xác khái niệm nghiên cứu, chứ không hề nhằm tạo ra càng nhiều biến càng tốt. Do đó sau khi thiết lập thang đo và đo lường, chúng ta cần tổng hợp dữ liệu từ các biến quan sát lại thành một hay một vài biến cơ bản để phản ảnh mức độ của khái niệm nghiên cứu chúng ta đang đo trên các đơn vị quan sát.

Nếu tổng hợp các biến quan sát về một khái niệm nghiên cứu lại thành 1 biến tổng hợp để phản ảnh chính xác mức độ của khái niệm nghiên cứu trên các đơn vị khảo sát, thì tập hợp các biến quan sát này sẽ tạo thành một thang đo (nhiều chỉ báo) đơn hướng cho khái niệm nghiên cứu đang đo lường, và khái niệm nghiên cứu này chỉ có một thành phần.

Nếu tổng hợp các lại các biến quan sát về một khái niệm nghiên cứu lại thành nhiều hơn 1 biến tổng hợp để phản ảnh chính xác mức độ của khái niệm nghiên cứu trên các đơn vị khảo sát, thì tập hợp các biến quan sát này sẽ tạo thành một thang đo (nhiều chỉ báo) đa hướng cho khái niệm nghiên cứu đang đo lường, và khái niệm nghiên cứu này có hơn 1 thành phần.

Nói cách khác, tập hợp biến quan sát gốc đo lường khái niệm chỉ có 1 khía cạnh/thành phần (rút trích ra được 1 nhân tố) tạo thành thang đo đơn hướng. Tập hợp biến quan sát gốc đo lường khái niệm có nhiều hơn 1 khía cạnh/thành phần (rút trích ra được nhiều hơn 1 nhân tố) tạo thành thang đo đa hướng.

Chương tiếp theo sẽ trình bày phương pháp phân tích nhân tố, là một kỹ thuật để giảm bớt dữ liệu, giúp chúng ta "rút trích" từ các biến quan sát thành 1 hay một số biến tổng hợp (gọi là nhân tố hay thành phần). Nếu bạn có 30 mục hỏi trong một nhóm các mục hỏi tiềm năng, và các câu trả lời từ một mẫu các đối tượng trả lời cho các mục hỏi đó, phân tích nhân tố sẽ cho phép bạn giảm 30 mục hỏi xuống một tập hợp nhỏ hơn, còn 2 hay 3 hay 4 nhân tố. Mỗi mục hỏi

được tính một tỷ số, được gọi là hệ số tải nhân tố (factor loading). Hệ số này cho bạn biết mỗi mục hỏi "thuộc về" những nhân tố chủ yếu nào.

Những nhà thiết kế thang đo chuyên nghiệp ngày này thường sử dụng phân tích nhân tố để kiểm tra tính đơn hướng (đơn khía cạnh) trong thang đo Likert. Nếu một khái niệm đang đo lường là đơn hướng thì sẽ có một nhân tố trội ẩn dưới tất cả các biến (các mục hỏi) và tất cả các mục hỏi sẽ "tải mạnh" lên nhân tố đó. Các nhà thiết kế thang đo sẽ lấy một số lớn các mục hỏi (ít nhất là 40), hỏi rất nhiều người (ít nhất 200) về các mục hỏi này, tiến hành phân tích các nhân tố, và chọn ra những mục hỏi biến) có hệ số tải lớn tại nhân tố (khái niệm ẩn) mà họ đang cố gắng tìm hiểu.

Chương trình máy tính dễ sử dụng hiện nay làm cho việc phân tích các nhân tố nhẹ đi, hầu hết sự phát triển thang đo trong tương lai sẽ sử dụng phương pháp này và các phương pháp tương tự. Bạn có thể sử dụng các phần mềm thống kê đầy đủ tính năng để phân tích một bảng ma trận các trả lời đối với toàn bộ các mục hỏi của thang đo. Trong chương tiếp theo, chúng ta sẽ dùng SPSS để thực hiện phân tích nhân tố.

CHƯƠNG XII

PHÂN TÍCH NHÂN TỐ

1. KHÁI NIỆM VÀ ỨNG DỤNG

Phân tích nhân tố là tên chung của một nhóm các thủ tục được sử dụng chủ yếu để thu nhỏ và tóm tắt các dữ liệu. Trong nghiên cứu, chúng ta có thể thu thập được một số lượng biến khá lớn và hầu hết các biến này có liên hệ với nhau và số lượng của chúng phải được giảm bớt xuống đến một số lượng mà chúng ta có thể sử dụng được (xem phần xây dựng thang đo trong Chương XI). Liên hệ giữa các nhóm biến có liên hệ qua lại lẫn nhau được xem xét và trình bày dưới dạng một số ít các nhân tố cơ bản. Ví dụ như hình ảnh của siêu thị có thể được đo lường bằng cách yêu cầu những người được phỏng vấn đánh giá các siêu thị về một loạt các chi tiết trên một thang đo ngữ nghĩa khoảng cách (xem ví dụ về thang đo khoảng cách ở Chương I). Những đánh giá về các chi tiết này sẽ được phân tích để xác định các nhân tố hình thành nên hình ảnh của siêu thị.

Trong phân tích phương sai, hồi qui bội và phân tích biệt số (ở chương kế tiếp), một biến được coi là phụ thuộc và các biến khác được coi là biến độc lập hay biến dự đoán. Nhưng trong phân tích nhân tố không có sự phân biệt hai loại như vậy. Mà thay vào đó, phân tích nhân tố là một kỹ thuật phụ thuộc lẫn nhau (interdependence technique) trong đó toàn bộ các mối liên hệ phụ thuộc lẫn nhau sẽ được nghiên cứu.

Phân tích nhân tố được sử dụng trong các trường hợp sau:

- Nhận diện các khía cạnh hay nhân tố giải thích được các liên hệ tương quan trong một tập hợp biến. Ví dụ, chúng ta có thể sử dụng một tập hợp các phát biểu về lối sống để đo lường tiểu sử tâm lý của người tiêu dùng. Sau đó những phát biểu (biến) này được sử dụng trong phân tích nhân tố để nhận diện các yếu tố tâm lý cơ bản.
- Nhận diện một tập hợp gồm một số lượng biến mới tương đối ít không có tương quan với nhau để thay thế tập hợp biến gốc có tương quan với nhau để thực hiện một phân tích đa biến tiếp theo

sau (ví dụ như hồi qui hay phân tích biệt số). Chẳng hạn như sau khi nhận diện các nhân tố thuộc về tâm lý thì ta có thể sử dụng chúng như những biến độc lập để giải thích những khác biệt giữa những người trung thành và những người không trung thành với nhãn hiệu sử dụng.

- Để nhận ra một tập hợp gồm một số ít các biến nổi trội từ một tập hợp nhiều biến để sử dụng trong các phân tích đa biến kế tiếp. Ví dụ như từ một số khá nhiều các câu phát biểu về lối sống (biến) gốc, ta chọn ra được một số ít biến được sử dụng như những biến độc lập để giải thích những khác biệt giữa những nhóm người có hành vi khác nhau.

Phân tích nhân tố có vô số ứng dụng trong các lĩnh vực nghiên cứu kinh tế và xã hội. Trong nghiên cứu xã hội, các khái niệm thường khá trừu tượng và phức tạp, phân tích nhân tố thường được dùng trong quá trình xây dựng thang đo lường các khía cạnh khác nhau của khái niệm nghiên cứu, kiểm tra tính đơn khía cạnh của thang đo lường (xem Chương IX). Trong kinh doanh, phân tích nhân tố có thể được ứng dụng trong nhiều trường hợp:

- Phân tích nhân tố có thể được sử dụng trong phân khúc thị trường để nhận ra các biến quan trọng dùng để phân nhóm người tiêu dùng. Những người mua xe có thể được nhóm theo sự chú trọng tương đối về tính kinh tế, tiện nghi, tính năng, và sang trọng. Và kết quả là có 4 phân khúc: những khách hàng tìm kiếm tính kinh tế, những người tìm kiếm tiện nghi, những người tìm kiếm tính năng và những người tìm kiếm sự sang trọng.
- Trong nghiên cứu sản phẩm, ta có thể sử dụng phân tích nhân tố để xác định các thuộc tính nhãn hiệu có ảnh hưởng đến sự lựa chọn của người tiêu dùng. Ví dụ như các nhãn hiệu kem đánh răng có thể được đánh giá theo khả năng bảo vệ chống sâu răng, trắng răng, mùi vị, hơi thở thơm tho, và giá cả.
- Trong nghiên cứu quảng cáo, phân tích nhân tố có thể dùng để hiểu thói quen sử dụng phương tiện truyền thông của thị trường mục tiêu.
- Trong nghiên cứu định giá, ta có thể sử dụng phân tích nhân tố để nhận ra các đặc trưng của những người nhạy cảm với giá. Ví dụ những người tiêu dùng nhạy cảm với giá có thể là những người có

tính ngăn nắp, có suy nghĩ tiết kiệm và thường ở trong nhà nhiều hơn là đi ra ngoài.

2. MÔ HÌNH PHÂN TÍCH NHÂN TỐ

Về mặt tính toán, phân tích nhân tố hơi giống với phân tích hồi qui bội ở chỗ mỗi biến được biểu diễn như là một kết hợp tuyến tính của các nhân tố cơ bản. Lượng biến thiên của một biến được giải thích bởi những nhân tố chung trong phân tích được gọi là communality. Biến thiên chung của các biến được mô tả bằng một số ít các nhân tố chung (common factor) cộng với một nhân tố đặc trưng (unique factor) cho mỗi biến. Những nhân tố này không bộc lộ rõ ràng. Nếu các biến được chuẩn hóa thì mô hình nhân tố được thể hiện bằng phương trình:

$$X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + ... + A_{im}F_m + V_iU_i$$

trong đó:

X_i : biến thứ i chuẩn hóa

A_{ij} : hệ số hồi qui bội chuẩn hóa của nhân tố j đối với biến i

F : các nhân tố chung

V_i : hệ số hồi qui chuẩn hóa của nhân tố đặc trưng i đối với biến i

U_i : nhân tố đặc trưng của biến i

m : số nhân tố chung

Các nhân tố đặc trưng có tương quan với nhau và với các nhân tố chung. Bản thân các nhân tố chung cũng có thể được diễn tả như những kết hợp tuyến tính của các biến quan sát:

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + ... + W_{ik}X_k$$

trong đó:

F_i : ước lượng trị số của nhân tố thứ i

W_i : quyền số hay trọng số nhân tố (weight or factor score coefficient)

k : số biến

Chúng ta có thể chọn các quyền số hay trọng số nhân tố sao cho nhân tố thứ nhất giải thích được phần biến thiên nhiều nhất trong toàn bộ biến thiên. Sau đó ta chọn một tập hợp các quyền số thứ hai

sao cho nhân tố thứ hai giải thích được phần lớn biến thiên còn lại, và không có tương quan với nhân tố thứ nhất.

Nguyên tắc này được áp dụng như vậy để tiếp tục chọn các quyền số cho các nhân tố tiếp theo. Do vậy các nhân tố được ước lượng sao cho các quyền số của chúng, không giống như các giá trị của các biến gốc, là không có tương quan với nhau. Hơn nữa, nhân tố thứ nhất giải thích được nhiều nhất biến thiên của dữ liệu, nhân tố thứ hai giải thích được nhiều thứ nhì ...

3. CÁC THAM SỐ THỐNG KÊ TRONG PHÂN TÍCH NHÂN TỐ

- Bartlett's test of sphericity: đại lượng Bartlett là một đại lượng thống kê dùng để xem xét giả thuyết các biến không có tương quan trong tổng thể. Nói cách khác, ma trận tương quan tổng thể là một ma trận đồng nhất, mỗi biến tương quan hoàn toàn với chính nó (r=1) nhưng không có tương quan với những biến khác (r=0). Điều kiện cần để áp dụng phân tích nhân tố là các biến phải có tương quan với nhau (các biến đo lường phản ánh những khía cạnh khác nhau của cùng một yếu tố chung). Do đó nếu kiểm định cho thấy không có ý nghĩa thống kê thì không nên áp dụng phân tích nhân tố cho các biến đang xem xét. Lúc đó biến đo lường có thể được xem là các nhân tố thực sự. Giả thuyết không của kiểm định này có thể được mô tả trong trường hợp phân tích nhân tố cho 6 biến quan sát như sau:

	v1	v2	v3	v4	v5	v6
v1	1					
v2	0	1				
v3	0	0	1			
v4	0	0	0	1		
v5	0	0	0	0	1	
V6	0	0	0	0	0	1

- Correlation matrix: cho biết hệ số tương quan giữa tất cả các cặp biến trong phân tích.
- Communality: là lượng biến thiên của một biến được giải thích chung với các biến khác được xem xét trong phân tích. Đây cũng là phần biến thiên được giải thích bởi các nhân tố chung.
- Eigenvalue: đại diện cho phần biến thiên được giải thích bởi mỗi nhân tố.

- Factor loadings (hệ số tải nhân tố): là những hệ số tương quan đơn giữa các biến và các nhân tố.
- Factor matrix (ma trận nhân tố): chứa các hệ số tải nhân tố của tất cả các biến đối với các nhân tố được rút ra
- Factor scores: là các điểm số nhân tố tổng hợp được ước lượng cho từng quan sát trên các nhân tố được rút ra. Còn được gọi là nhân số.
- Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy: là một chỉ số dùng để xem xét sự thích hợp của phân tích nhân tố. Trị số của KMO lớn (giữa 0,5 và 1) là điều kiện đủ để phân tích nhân tố là thích hợp, còn nếu như trị số này nhỏ hơn 0,5 thì phân tích nhân tố có khả năng không thích hợp với các dữ liệu.
- Percentage of variance: phần trăm phương sai toàn bộ được giải thích bởi từng nhân tố. Nghĩa là coi biến thiên là 100% thì giá trị này cho biết phân tích nhân tố cô đọng được bao nhiêu % và bị thất thoát bao nhiêu % .
- Residuals: là các chênh lệch giữa các hệ số tương quan trong ma trận tương quan đầu vào (input correlation matrix) và các hệ số tương quan sau khi phân tích (reproduced correlations) được ước lượng từ ma trận nhân tố (factor matrix).

4. TIẾN HÀNH PHÂN TÍCH NHÂN TỐ

4.1. Xác định vấn đề

Xác định vấn đề nghiên cứu gồm có nhiều bước. Đầu tiên là ta phải nhận diện các mục tiêu của phân tích nhân tố cụ thể là gì. Các biến tham gia vào phân tích nhân tố phải được xác định dựa vào các nghiên cứu trong quá khứ, phân tích lý thuyết, và đánh giá của các nhà nghiên cứu. Một vấn đề quan trọng là các biến này phải được đo lường một cách thích hợp bằng thang đo định lượng (khoảng cách hay tỉ lệ), và cỡ mẫu phải đủ lớn. Thông thường thì số quan sát (cỡ mẫu) ít nhất phải bằng 4 hay 5 lần số biến trong phân tích nhân tố. Trong nhiều tình huống nghiên cứu, quy mô mẫu khá nhỏ và tỉ số này đôi khi cũng khá nhỏ. Trong những trường hợp này thì việc giải thích các kết quả cần phải thận trọng.

Để minh họa, chúng ta sẽ xem xét ví dụ sau, trong đó một nhà nghiên cứu muốn xác định các lợi ích căn bản người tiêu dùng muốn tìm kiếm khi mua một ống kem đánh răng. Mẫu gồm 35 người tiêu dùng có mua kem đánh răng được phỏng vấn . Những người được phỏng vấn cho biết mức độ quan trọng của sáu lợi ích sau trên thang đo bảy điểm (1= không quan trọng chút nào, 7 = rất quan trọng). Các dữ liệu thu thập được nhập vào file *Phan tich nhan to* trong tập hợp dữ liệu dùng kèm sách.

V1 : ngừa sâu răng
V2 : làm trắng răng
V3 : làm khỏe nứu răng
V4 : làm hơi thở thơm tho
V5 : làm sạch cáu răng
V6 : làm răng bóng hơn

4.2. Xây dựng ma trận tương quan

Quá trình phân tích được dựa trên ma trận tương quan của các biến này. Để có thể áp dụng được phân tích nhân tố thì các biến phải có liên hệ với nhau. Trong thực tế thì thường chúng ta luôn có điều này. Nếu hệ số tương quan giữa các biến nhỏ, phân tích nhân tố có thể không thích hợp. Chúng ta trông chờ rằng các biến này có tương quan chặt chẽ với nhau và như vậy sẽ tương quan chặt với cùng một hay nhiều nhân tố.

Chúng ta có thể sử dụng Bartlett's test of sphericity để kiểm định giả thuyết không (Ho) là các biến không có tương quan với nhau trong tổng thể, nói cách khác là ma trận tương quan tổng thể là một ma trận đơn vị trong đó tất cả các giá trị trên đường chéo đều bằng 1 còn các giá trị nằm ngoài đường chéo đều bằng 0. Đại lượng kiểm định này dựa trên sự biến đổi thành đại lượng chi-square từ định thức của ma trận tương quan. Đại lượng này có giá trị càng lớn thì ta càng có nhiều khả năng bác bỏ giả thuyết không này. Nếu giả thuyết Ho không thể bị bác bỏ thì phân tích nhân tố rất có khả năng không thích hợp.

Bảng 12.1 cho thấy ma trận tương quan của các dữ liệu thu thập được. Chúng ta có thể thấy có tương quan giữa các biến V1 (ngừa sâu răng), V3 (khoẻ nứu răng), và V5 (sạch cáu răng). Chúng ta hy

vọng rằng các biến này có tương quan với cùng một (hay một tập hợp) nhân tố. Tương tự như vậy, ta có thể thấy có liên hệ tương đối chặt giữa các biến V2 (làm trắng răng), V4 (hơi thở thơm tho) và V6 (làm răng bóng hơn). Ta cũng hy vọng là ba biến này cùng có tương quan với một hay nhiều nhân tố.

Các kết quả phân tích nhân tố được trình bày trong các Bảng 12.2. Giả thuyết không cho rằng ma trận tương quan tổng thể là một ma trận đơn vị bị bác bỏ theo kết quả kiểm định Bartlett căn cứ trên giá trị sig. (Bảng 12.2b). Vì vậy phân tích nhân tố là phương pháp phù hợp để phân tích ma trận tương quan thể hiện ở Bảng 12.1.

Bảng 12.1 **Correlation Matrix**

	ngua sau rang	lam trang rang	lam khoe nuu rang	lam hoi tho thom tho	lam sach cau rang	lam rang bong hon
ngua sau rang	1.000	.039	.321	.000	.314	-.097
lam trang rang	.039	1.000	-.130	.534	.352	.593
lam khoe nuu rang	.321	-.130	1.000	-.432	.474	.037
lam hoi tho thom tho	.000	.534	-.432	1.000	.077	.345
lam sach cau rang	.314	.352	.474	.077	1.000	.279
lam rang bong hon	-.097	.593	.037	.345	.279	1.000

4.3. Số lượng nhân tố

Chúng ta có thể tính ra một số lượng nhân tố nhiều bằng số biến, nhưng làm như vậy thì không có tác dụng gì cho mục đích tóm tắt thông tin. Để tóm tắt các thông tin chứa đựng trong các biến gốc, chúng ta cần rút ra một số lượng các nhân tố ít hơn số biến. Vấn đề là bao nhiêu nhân tố? Có 5 phương pháp nhằm xác định số lượng nhân tố: xác định từ trước, dựa vào eigenvalue, biểu đồ dốc (scree plot), phần trăm biến thiên giải thích được (percentage of variance), chia đôi mẫu, kiểm định mức ý nghĩa. Ta tìm hiểu cụ thể 2 phương pháp :

- Phương pháp xác định từ trước (Priori determination): đôi khi từ kinh nghiệm và hiểu biết của mình, từ phân tích lý thuyết, hay từ kết quả của các cuộc nghiên cứu trước ... người nghiên cứu biết được có bao nhiêu nhân tố có thể rút ra và như vậy có thể chỉ định trước số lượng nhân tố cần phải rút ra để báo cho chương trình máy

tính. Hầu hết các chương trình máy tính đều cho phép người dùng chỉ định số nhân tố cần phân tích, làm cho việc giải thích kết quả dễ dàng hơn.

- Phương pháp dựa vào eigenvalue (Determination based on eigen value): chỉ có những nhân tố nào có eigenvalue lớn hơn 1 mới được giữ lại trong mô hình phân tích. Đại lượng eigenvalue đại diện cho lượng biến thiên được giải thích bởi nhân tố. Những nhân tố có eigenvalue nhỏ hơn 1 sẽ không có tác dụng tóm tắt thông tin tốt hơn một biến gốc, vì sau khi chuẩn hóa mỗi biến gốc có phương sai là 1.

Chúng ta có thể xác định ý nghĩa thống kê của các eigenvalue riêng biệt và giữ lại những nhân tố nào thực sự có ý nghĩa thống kê. Nhược điểm của cách này là khi quy mô mẫu lớn (hơn 200), có nhiều khả năng sẽ có nhiều nhân tố thỏa mãn mức ý nghĩa thống kê mặc dù trong thực tế có nhiều nhân tố chỉ giải thích được chỉ một phần nhỏ toàn bộ biến thiên.

Trong Bảng 12.2d, chúng ta thấy rằng theo tiêu chuẩn eigenvalue lớn hơn 1 (mặc định của chương trình SPSS) thì chỉ có 2 nhân tố được rút ra. Và từ hiểu biết của bản thân, chúng ta cũng có thể biết được người ta mua kem đánh răng vì hai lý do. Do đó số lượng hai nhân tố là thích hợp. Trong bảng này hàng Cumulative % cho biết 2 nhân tố đầu tiên giải thích được 66,2% biến thiên của dữ liệu. Bảng Communalities cho biết những thông tin có liên quan sau khi số lượng nhân tố mong muốn được rút ra. Nó cho biết các communality của các biến tức là phần biến thiên được giải thích bởi các nhân tố chung.

Việc giải thích kết quả thường được tăng cường bằng cách xoay các nhân tố (Bảng 12.2f Rotated Component Matrix).

Bảng 12.2a **Descriptive Statistics**

	Mean	Std. Deviation	Analysis N
ngua sau rang	6.43	.850	35
lam trang rang	5.63	1.395	35
lam khoe nuu rang	6.00	.970	35
lam hoi tho thom tho	5.40	1.193	35
lam sach cau rang	5.17	1.150	35
lam rang bong hon	4.37	1.629	35

Bảng 12.2b **KMO and Bartlett's Test**

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.584
Bartlett's Test of Sphericity	Approx. Chi-Square	55.275
	df	15
	Sig.	.000

Bảng 12.2c **Communalities**

	Initial	Extraction
ngua sau rang	1.000	.384
lam trang rang	1.000	.793
lam khoe nuu rang	1.000	.780
lam hoi tho thom tho	1.000	.670
lam sach cau rang	1.000	.740
lam rang bong hon	1.000	.604

Extraction Method: Principal Component Analysis.

Bảng 12.2d **Total Variance Explained**

		Component					
		1	2	3	4	5	6
Initial Eigenvalue	Total	2.157	1.813	.912	.490	.350	.278
	% of Variance	35.957	30.214	15.206	8.168	5.829	4.625
	Cumulative %	35.957	66.172	81.378	89.546	95.375	100.000
Extraction Sums of Squared Loading	Total	2.157	1.813				
	% of Variance	35.957	30.214				
	Cumulative %	35.957	66.172				
Rotation Sums of Squared Loading	Total	2.154	1.817				
	% of Variance	35.896	30.276				
	Cumulative %	35.896	66.172				

Extraction Method: Principal Component Analysis.

Bảng 12.2e **Component Matrix(a)**

	Component	
	1	2
ngua sau rang	.050	.618
lam trang rang	.891	-.007
lam khoe nuu rang	-.143	.872
lam hoi tho thom tho	.726	-.377
lam sach cau rang	.462	.726
lam rang bong hon	.775	.050

Extraction Method: Principal Component Analysis.

a 2 components extracted.

Bảng 12.2f **Rotated Component Matrix(a)**

	Component	
	1	2
ngua sau rang	-.014	.620
lam trang rang	.886	.086
lam khoe nuu rang	-.233	.852
lam hoi tho thom tho	.761	-.300
lam sach cau rang	.384	.770
lam rang bong hon	.766	.130

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a Rotation converged in 3 iterations.

Bảng 12.2g **Component Transformation Matrix**

Component	1	2
1	.995	.104
2	-.104	.995

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

Bảng 12.2h **Component Score Coefficient Matrix**

	Component	
	1	2
ngua sau rang	-.012	.341
lam trang rang	.411	.039
lam khoe nuu rang	-.116	.471
lam hoi tho thom tho	.356	-.172
lam sach cau rang	.171	.420
lam rang bong hon	.355	.065

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Component Scores.

Bảng 12.2 i **Component Score Covariance Matrix**

Component	1	2
1	1.000	.000
2	.000	1.000

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Component Scores.

4.4. Xoay các nhân tố

Một phần quan trọng trong bảng kết quả phân tích nhân tố là ma trận nhân tố (Component Matrix). Ma trận nhân tố chứa các hệ số biểu diễn các biến chuẩn hóa bằng các nhân tố (mỗi biến là một đa thức của các nhân tố). Những hệ số này (factor loading) biểu diễn tương quan giữa các nhân tố và các biến. Hệ số này lớn cho biết nhân tố và biến có liên hệ chặt chẽ với nhau. Các hệ số này được dùng để giải thích các nhân tố.

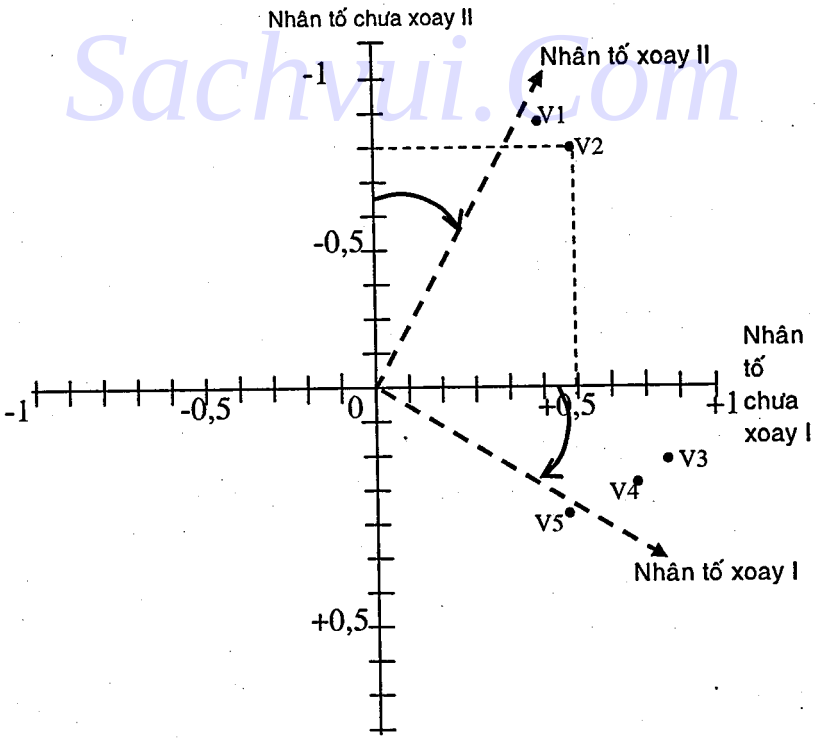
Mặc dù ma trận nhân tố ban đầu hay ma trận nhân tố không xoay này cho thấy được mối quan hệ giữa các nhân tố và từng biến một, nhưng nó ít khi tạo ra những nhân tố có thể giải thích được một cách dễ dàng bởi vì các nhân tố có tương quan với nhiều biến. Ví dụ như trong Bảng 12.2e nhân tố 1 có tương chặt với V2, V4, V6 và cũng khá tương đối với V5 (factor loading = 0,462 xem như gần = 0,5). Làm thế nào để giải thích một cách suôn sẻ nhân tố này? Trong những ma trận như vậy hay phức tạp hơn, việc giải thích các kết quả khá khó khăn. Vì vậy thông qua việc xoay các nhân tố, ma trận nhân tố sẽ trở nên đơn giản hơn và dễ giải thích hơn.

Khi xoay các nhân tố, chúng ta muốn mỗi nhân tố có hệ số khác không (tức là có ý nghĩa) chỉ trong vài biến mà thôi. Tương tự, chúng ta cũng muốn mỗi biến chỉ có hệ số khác không (có ý nghĩa) chỉ với vài nhân tố, hay nếu có thể, chỉ với một nhân tố mà thôi. Nếu nhiều nhân tố có hệ số lớn trong cùng một biến, chúng ta cũng khó mà giải thích được. Việc xoay nhân tố không có ảnh hưởng đến communality và phần trăm của toàn bộ phương sai được giải thích. Tuy nhiên phần trăm phương sai được giải thích bởi từng nhân tố có thay đổi. Phần phương sai giải thích bởi từng nhân tố sẽ được phân phối lại khi xoay nhân tố. Vì vậy các phương pháp xoay khác nhau sẽ nhận diện những nhân tố khác nhau. Có nhiều phương pháp xoay khác nhau như:

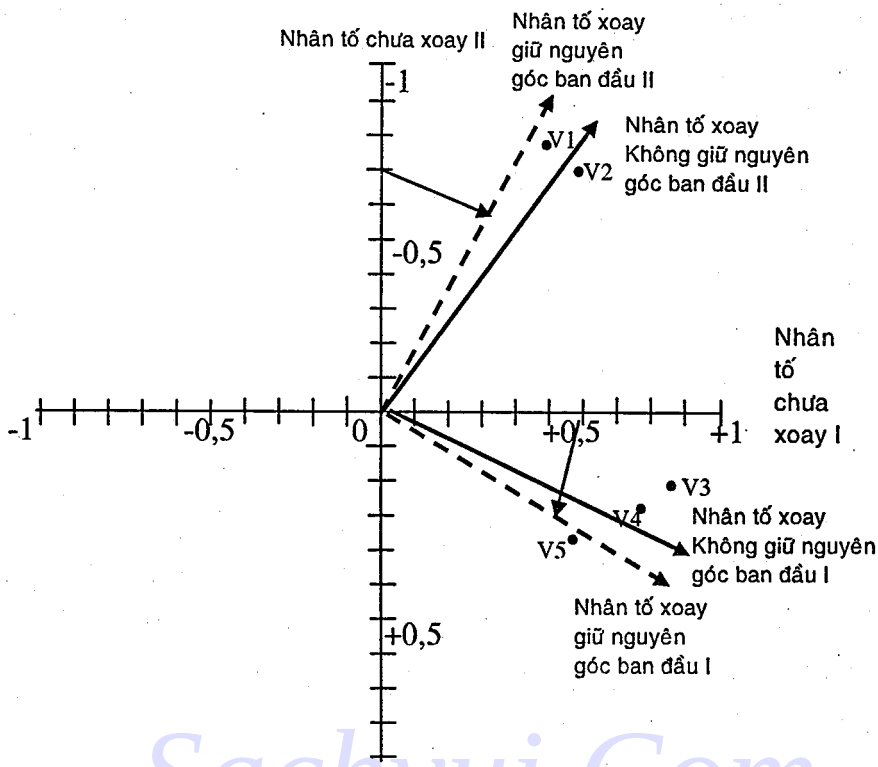
- **Orthogonal rotation**: xoay các nhân tố trong đó vẫn giữ nguyên góc ban đầu giữa các nhân tố.

- **Varimax procedure**: xoay nguyên góc các nhân tố để tối thiểu hóa số lượng biến có hệ số lớn tại cùng một nhân tố, vì vậy sẽ tăng cường khả năng giải thích các nhân tố.
- **Quartimax**: xoay nguyên góc các nhân tố để tối thiểu hóa số nhân tố có hệ số lớn tại cùng một biến, vì vậy sẽ tăng cường khả năng giải thích các biến.
- **Equamax**: xoay các nhân tố để đơn giản hóa việc giải thích cả biến lẫn nhân tố.
- **Oblique** (direct oblimin): xoay các nhân tố mà không giữ nguyên góc ban đầu giữa các nhân tố (tức là có tương quan giữa các nhân tố với nhau). Phương pháp này nên được sử dụng chỉ khi nào các nhân tố trong tổng thể có khả năng tương quan mạnh với nhau.

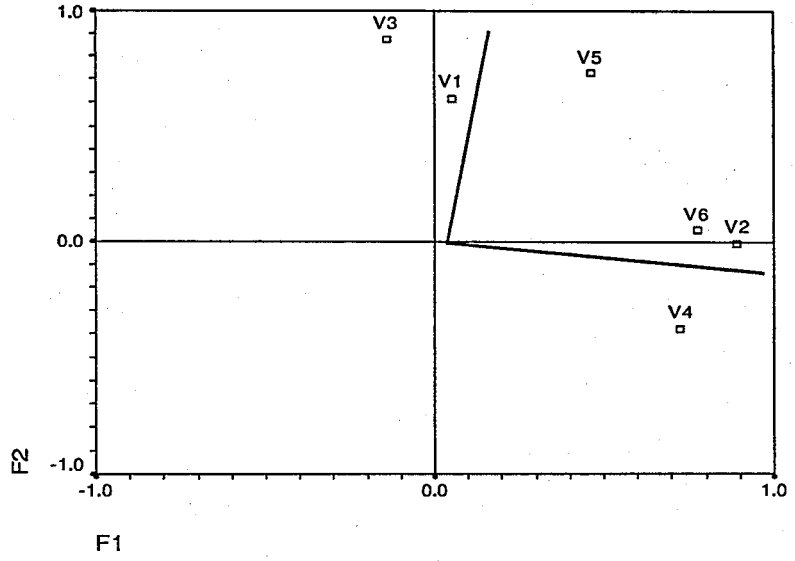
Hình 12.1a: Một ví dụ về xoay giữ nguyên góc ban đầu của các nhân tố



Hình 12.1b: Một ví dụ về xoay không giữ nguyên góc ban đầu của các nhân tố



Hình 12.2: Đồ thị minh họa xoay nhân tố của ví dụ về lợi ích khi mua kem đánh răng



Phương pháp Varimax thường được sử dụng phổ biến nhất. Trong Bảng 12.2f, so sánh ma trận nhân tố sau khi xoay (Rotated Component Matrix) với ma trận trước khi xoay (Bảng 12.2e), chúng ta cũng có thể thấy được xoay nhân tố làm cho việc giải thích dễ dàng hơn. Các biến V1, V2, V3, V4, V6 đều có hệ số lớn (tương quan chặt) với chỉ một trong hai nhân tố, còn tương quan của V5 đối với hai nhân tố này chưa thật rõ ràng dứt khoát. Sau khi xoay, biến V5 có hệ số ở nhân tố 1 nhỏ đi một chút và hệ số của nó ở nhân tố 2 lớn lên một chút. Và chúng ta có thể nói rằng biến V5 chủ yếu có tương quan với nhân tố 2.

4.5. Đặt tên và giải thích các nhân tố

Việc giải thích các nhân tố được thực hiện trên cơ sở nhận ra các biến có hệ số (factor loading) lớn ở cùng một nhân tố. Như vậy nhân tố này có thể được giải thích bằng các biến có hệ số lớn đối với bản thân nó. Trong ma trận nhân tố sau khi xoay trong Bảng 12.2f, nhân tố 1 có hệ số lớn ở các biến V2 (làm trắng răng), V4 (làm hơi thở thơm tho), và V6 (làm răng bóng hơn). Vì vậy nhân tố này có thể được đặt tên là *nhân tố lợi ích giao tiếp xã hội*. Nhân tố 2 có tương quan chặt với các biến V1 (ngừa sâu răng), V3 (làm khỏe nứu răng), và V5 (làm sạch cáu răng). Do đó nhân tố 2 có thể được đặt tên là *nhân tố lợi ích sức khỏe*. Và chúng ta có thể tóm tắt các dữ liệu thu thập được để nói rằng người tiêu dùng dường như tìm kiếm hai loại lợi ích chính khi mua kem đánh răng: lợi ích sức khỏe và lợi ích giao tiếp xã hội.

4.6. Nhân số (factor score)

Sau khi giải thích các nhân tố, nếu cần thì chúng ta có thể tính toán ra các nhân số. Bản thân phân tích nhân tố là một phương pháp độc lập trong phân tích có thể sử dụng một mình. Tuy nhiên nếu mục tiêu của phân tích nhân tố là biến đổi một tập hợp biến gốc thành một tập hợp các biến tổng hợp (nhân tố) có số lượng ít hơn để sử dụng trong các phương pháp phân tích đa biến tiếp theo, thì chúng ta có thể tính toán ra các nhân số (trị số của các biến tổng hợp) cho từng trường hợp quan sát một. Nhân số của nhân tố thứ i bằng:

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + ... + W_{ik}X_k$$

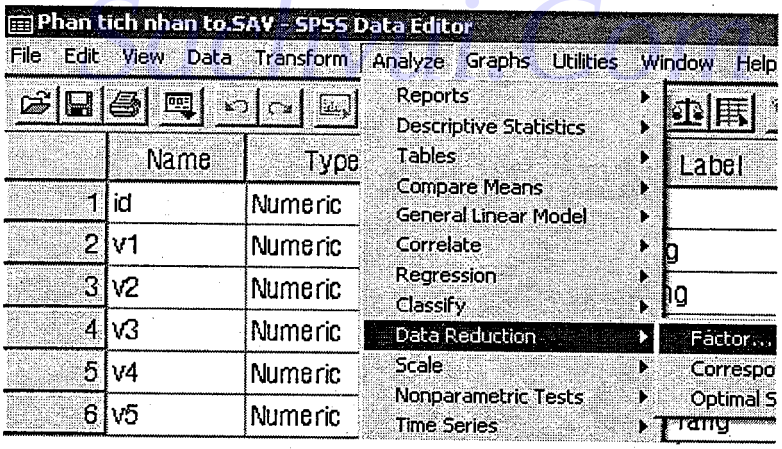
Các hệ số nhân tố W được dùng để kết hợp các biến chuẩn hóa được trình bày trong ma trận hệ số nhân tố (Component Score Coefficient Matrix). Nhờ ma trận này, chúng ta có thể tính ra trị số của các nhân tố (nhân số) dùng thay cho trị số các biến gốc trong các phân tích khác. Ví dụ trong Bảng 12.2h, ta có thể tính ra hai nhân số cho từng quan sát (từng người trả lời) bằng cách nhân giá trị của các biến gốc của một quan sát với các hệ số nhân tố để tính ra các nhân số.

Ví dụ $F_1 = -0{,}012X_1 + 0{,}411X_2 - 0{,}116X_3 + 0{,}356X_4 + 0{,}171X_5 + 0{,}355X_6$

Để thực hiện công việc này một cách tự động, chúng ta có thể ra lệnh cho chương trình máy tính tính toán các nhân số này và lưu các trị số này như những biến mới trong file dữ liệu. Trong ví dụ này thì chúng ta sẽ có thêm hai biến mới được thêm vào file dữ liệu. Chú ý các nhân số được lưu lại bằng lệnh Save của máy sẽ dưới dạng dữ liệu đã chuẩn hóa (đơn vị đo lường độ lệch chuẩn).

5. THỰC HIỆN PHÂN TÍCH NHÂN TỐ VỚI SPSS

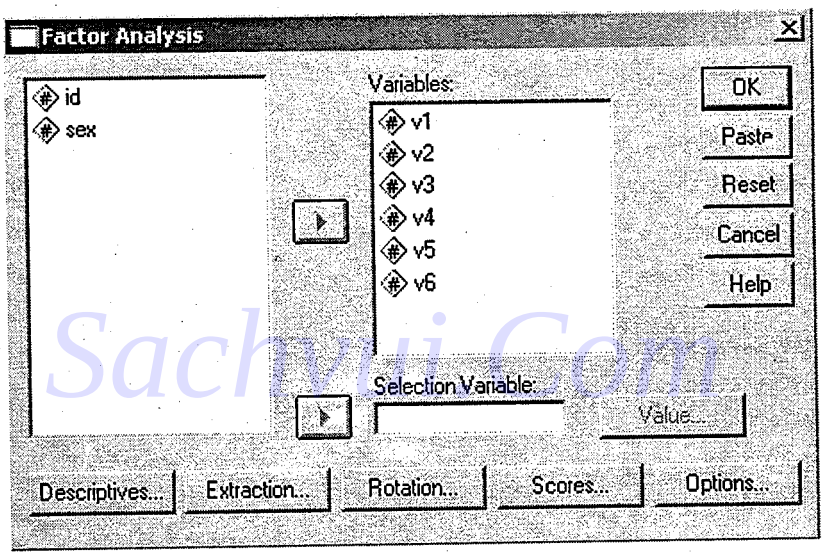
Từ menu ta chọn: Statistics > Data Reduction > Factor như sau đây Hình 12.3



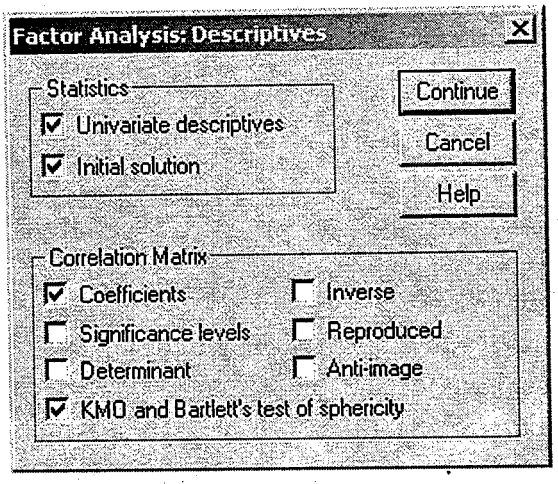
Lệnh này sẽ mở hộp thoại phân tích nhân tố như Hình 12.4. Các biến số trong file dữ liệu sẽ xuất hiện trong ô danh sách biến nguồn ở phía bên trái của hộp thoại. Bạn hãy chọn các biến cần phân tích trong danh sách biến nguồn rồi nhấn chuột vào mũi tên qua phải đưa vào ô Variable.

Tiếp theo bạn nhấn vào nút Descriptives trên hộp thoại để xác định các tham số thống kê mô tả cần tính. Hộp thoại con Descriptives sẽ xuất hiện (Hình 12.5).Trong hộp thoại con này, bạn có thể chọn các thống kê mô tả cho từng biến, phương án nhân tố ban đầu, thực hiện kiểm định Bartlett cũng như tính các ma trận hệ số tương quan . . . Rồi nhấn nút Continue để trở lại hộp thoại phân tích nhân tố ban đầu.

Hình 12.4

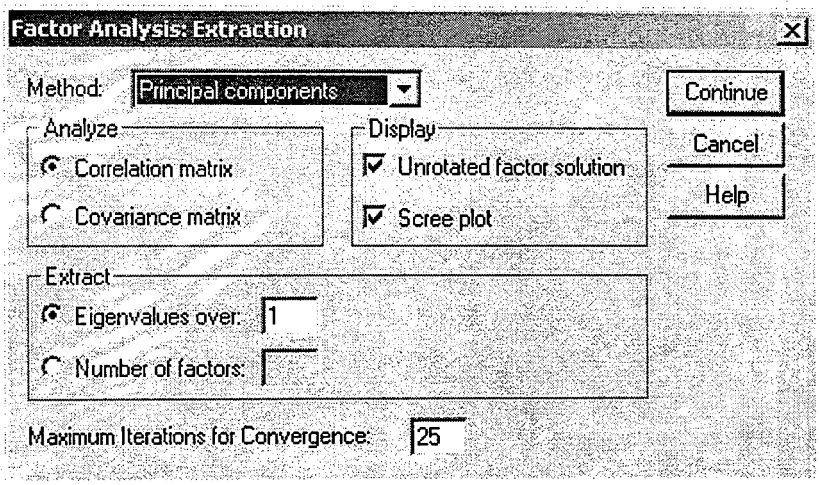


Hình 12.5



Bước tiếp theo là bạn nhấn chuột vào nút Extraction, hộp thoại con sẽ xuất hiện như sau:

Hình 12.6

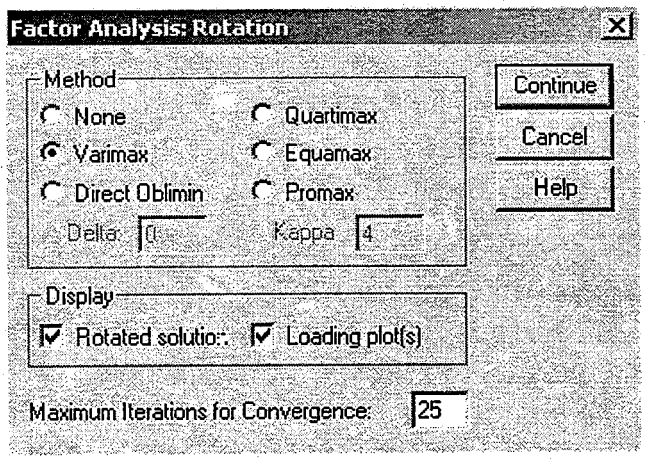


Trong hộp thoại này, bạn chọn:

- phương pháp rút trích các nhân tố (phương pháp mặc định là rút các thành phần chính - Principal components)
- phân tích ma trận tương quan hay hiệp phương sai
- thể hiện phương án nhân tố chưa xoay, và vẽ biểu đồ dốc
- xác định tiêu chuẩn rút trích nhân tố hay số lượng nhân tố cần rút trích

Rồi bạn nhấn nút Continue để trở lại hộp thoại ban đầu. Tiếp theo là bạn nhấn nút Rotation, hộp thoại con sau đây sẽ xuất hiện:

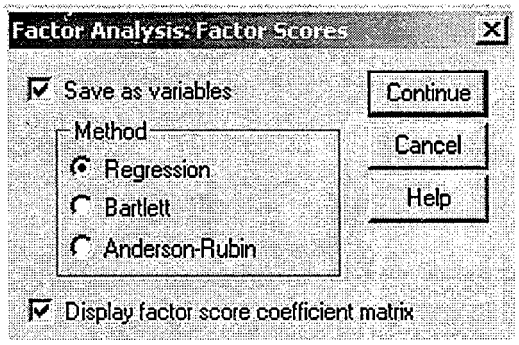
Hình 12.7



Trong hộp thoại này, bạn chỉ định phương pháp xoay các nhân tố; cho thể hiện phương án rút trích nhân tố sau khi xoay; vẽ các nhân tố

(từ 2 nhân tố trở lên). Rồi nhấn nút Continue để trở về hộp thoại ban đầu. Bước tiếp theo là bạn nhấn chuột vào nút Scores trên hộp thoại phân tích nhân tố, hộp thoại sau sẽ xuất hiện:

Hình 12.8



Trong hộp thoại này bạn chọn phương pháp tính nhân số (factor scores). Mặc định của chương trình là phương pháp Regression. Các nhân số (đã được chuyển qua đơn vị đo lường độ lệch chuẩn) của từng quan sát sẽ được chứa trong các biến mới trong file dữ liệu của bạn. Bạn có thể cho hiện bảng trọng số nhân tố. Rồi bạn nhấn nút Continue để trở về hộp thoại ban đầu.

Bước tiếp theo là bạn nhấn nút Options để chỉ định cách thức xử lý các quan sát thiếu dữ liệu. Rồi nhấn nút Continue trở về hộp thoại ban đầu Cuối cùng trong hộp thoại phân tích nhân tố này, nhấn nút OK để thực hiện phân tích.

Sau khi rút trích được các nhân tố và lưu lại thành các biến mới, chúng ta sẽ sử dụng các biến mới này thay cho tập hợp biến gốc để đưa vào các phân tích tiếp theo như kiểm định trung bình, ANOVA, tương quan & hồi qui ...

Trong ví dụ này, chúng ta có thể kiểm định xem lợi ích giao tiếp xã hội khi mua kem đánh răng có khác biệt giữa nam và nữ hay không bằng một kiểm định t đối với mẫu độc lập Nam và Nữ. Bảng kết quả trang sau cho thấy có sự khác biệt có ý nghĩa thống kê: nam quan tâm đến lợi ích giao tiếp xã hội nhiều hơn nữ. Từ kết quả này, các quảng cáo kem đánh răng nhắm vào nam giới thường đề cao lợi ích giao tiếp xã hội với những nhân vật quảng cáo nam rất thành công trong giao tiếp với kem đánh răng đang được quảng cáo.

Hình 12.9:

Group Statistics

	giới tính	N	Mean	Std. Deviation	Std. Error Mean
lợi ích giao tiếp XH	nam	17	.3833755	.87044420	.21111373
	nữ	18	-.3620768	1.000554	.23583278
lợi ích sức khỏe	nam	17	-.2771805	1.155686	.28029505
	nữ	18	.2617816	.77043184	.18159253

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
lợi ích giao tiếp XH	Equal variances assumed	.224	.639	2.346	33	.025	.7454523	.31781402
	Equal variances not assumed			2.355	32.790	.025	.7454523	.31652189
lợi ích sức khỏe	Equal variances assumed	5.726	.023	-1.632	33	.112	-.5389620	.33021613
	Equal variances not assumed			-1.614	27.663	.118	-.5389620	.33397778

CHƯƠNG XIII
PHÂN TÍCH BIỆT SỐ

1. KHÁI NIỆM CĂN BẢN

Phân tích biệt số là một kỹ thuật phân tích dữ liệu khi biến phụ thuộc (biến tiêu chuẩn) là biến phân loại và biến độc lập (biến dự đoán) là biến định lượng (thang đo khoảng cách hay tỉ lệ). Ví dụ, biến phụ thuộc có thể là việc chọn mua một nhãn hiệu máy vi tính (A, B hay C) và biến độc lập có thể là điểm đánh giá các thuộc tính của máy PC trên thang đo Likert 7 điểm. Các mục tiêu của phân tích biệt số là:

- xây dựng các hàm phân tích phân biệt (discriminant functions) hay một hàm tuyến tính kết hợp các biến độc lập sao cho phân biệt rõ nhất các biểu hiện của biến phụ thuộc (biến phụ thuộc trong trường hợp này là biến định tính chỉ có các biểu hiện không có các mức độ)
- nghiên cứu xem có tồn tại sự khác biệt có ý nghĩa giữa các nhóm xét theo các biến độc lập
- xác định những biến độc lập nào là nguyên nhân lớn nhất gây ra những sự khác biệt giữa các nhóm
- phân loại các quan sát vào trong một nhóm nào đó dựa vào các giá trị của các biệt độc lập
- đánh giá tính chính xác của việc phân loại

Phân tích biệt số được sử dụng phổ biến trong nhiều lĩnh vực như tâm lý, xã hội cũng như kinh doanh. Trong kinh doanh, phân tích biệt số có thể được sử dụng trong rất nhiều tình huống nghiên cứu như:

- phân biệt các khách hàng trung thành và những người không trung thành bằng các đặc tính nhân khẩu học, tâm lý hay lối sống.
- phân biệt những người dùng nhiều, dùng trung bình và dùng ít một sản phẩm nào đó qua mức độ họ tiêu thụ các sản phẩm khác.
- phát hiện ra các đặc trưng tâm lý giúp phân biệt giữa những người nhạy cảm với giá và những người không nhạy cảm với giá
- phân biệt khách hàng thường xuyên của các cửa hàng và khách hàng thường xuyên của siêu thị qua quan niệm sống, lối sống . . .

Có hai trường hợp phân tích biệt số là:

- Phân tích biệt số hai nhóm (two-group discriminant analysis): khi biến độc lập chỉ có hai biểu hiện
- Phân tích biệt số bội (multiple discriminant analysis): khi biến độc lập có ba hay nhiều biểu hiện

2. LIÊN HỆ GIỮA PHÂN TÍCH BIỆT SỐ, HỒI QUI VÀ ANOVA

Sự giống nhau và khác nhau giữa 3 phương pháp phân tích biệt số, hồi qui và ANOVA được tóm tắt trong bảng sau:

	ANOVA	Hồi qui	Phân tích biệt số
giống nhau:			
* số biến phụ thuộc	một	một	một
* số biến độc lập	nhiều	nhiều	nhiều
khác nhau:			
* tính chất của biến phụ thuộc	định lượng	định lượng	phân loại
* tính chất của biến độc lập	phân loại	định lượng	định lượng

Trong thực tế phân tích dữ liệu, tùy theo mục tiêu nghiên cứu, tính chất của dữ liệu (do điều kiện thu thập dữ liệu hay do đặc điểm của đối tượng khảo sát) mà người phân tích dữ liệu sử dụng linh hoạt các phương pháp phân tích. Điều cần lưu ý là khi thiết kế nghiên cứu, người nghiên cứu cần hình dung ra các mô hình phân tích để thiết kế thang đo lường thu thập các dữ liệu phù hợp với mô hình phân tích định sử dụng.

3. MÔ HÌNH PHÂN TÍCH BIỆT SỐ

Mô hình phân tích biệt số có dạng tuyến tính như sau:

$$D = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_kX_k$$

trong đó:

D: biệt số

b: hệ số hay trọng số phân biệt

X: biến độc lập

Các nệ số hay trọng số (b) được tính toán sao cho các nhóm có các giá trị của hàm phân biệt (biệt số D) khác nhau càng nhiều càng tốt. Điều này sẽ xảy ra khi tỉ lệ của tổng các độ lệch bình phương của

biệt số giữa các nhóm (between-group sum of squares) so với tổng các độ lệch bình phương của biệt số trong nội bộ các nhóm (within-group sum of squares) đạt cực đại. Và bất cứ kết hợp tuyến tính nào khác của các biến độc lập cũng đều tạo ra những tỉ lệ nhỏ hơn.

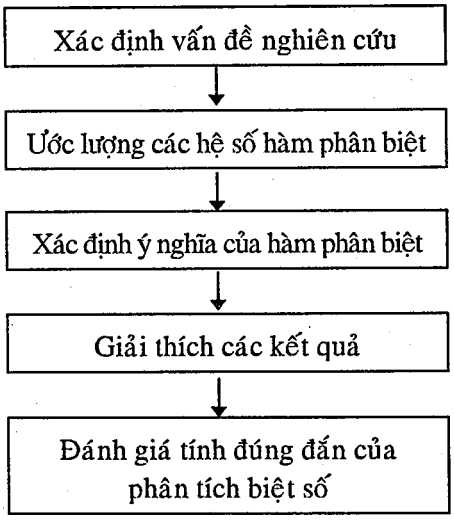
4. CÁC THAM SỐ THỐNG KÊ TRONG PHÂN TÍCH BIỆT SỐ

- Canonical correlation: hệ số tương quan canonical đo lường mức độ liên hệ giữa các biệt số và các nhóm. Nó là một thước đo mối liên hệ giữa hàm phân biệt đơn và tập hợp các biến giả xác định các nhóm.
- Centroid: là trung bình của các giá trị biệt số trong mỗi nhóm. Số centroid bằng với số nhóm vì mỗi nhóm có một centroid.
- Classification matrix: ma trận phân loại (ma trận dự đoán) chứa số quan sát được phân loại đúng và số quan sát phân loại sai. Số quan sát phân loại đúng sẽ nằm trên đường chéo của ma trận, và số quan sát phân loại sai nằm ngoài đường chéo. Tổng của các số nằm trên đường chéo được chia cho tổng số quan sát và được gọi là tỉ lệ đúng (tỉ lệ thành công).
- Discriminant function coefficients: các hệ số hàm phân biệt (chưa chuẩn hóa) là các quyền số (trọng số) của các biến khi các biến được đo lường bằng đơn vị tính nguyên thủy.
- Discriminant scores: các biệt số được tính bằng cách nhân các hệ số không chuẩn hóa được với giá trị của các biến, sau đó lấy tổng của các tích tìm được theo phương trình ở phần trên .
- Eigenvalue: đối với mỗi hàm phân biệt thì eigenvalue là tỉ số giữa tổng các độ lệch bình phương giữa các nhóm và tổng các độ lệch bình phương trong nội bộ nhóm (SSG/SSW, bạn đọc xem lại khái niệm tổng các chênh lệch bình phương giữa các nhóm - SSG và tổng các chênh lệch bình phương trong nội bộ nhóm - SSW trong phần Phân tích phương sai và Tương quan hồi qui tuyến tính trong Tập 1). Eigenvalue càng lớn thì hàm phân biệt càng tốt.
- F value and their significane: giá trị F được tính từ ANOVA một yếu tố, trong đó biến phân loại được sử dụng như biến độc lập, và mỗi biến dự đoán được sử dụng như biến phụ thuộc kiểu định lượng.
- Group means and group standard deviations: Trung bình nhóm và độ lệch chuẩn nhóm được tính cho mỗi biến dự đoán cho mỗi nhóm.

- Pooled within-group correlation matrix: ma trận tương quan nội bộ nhóm chung được tính bằng cách lấy trung bình các ma trận hiệp phương sai riêng cho tất cả các nhóm.
- Standardized discriminant function coefficients: các hệ số hàm phân biệt chuẩn hóa là các hệ số hàm phân biệt được sử dụng như quyền số khi các biến được chuẩn hóa có trung bình là 0 và phương sai là 1.
- Structure correlation (discriminant loadings): tương quan kết cấu (hệ số biệt tải) cho biết các hệ số tương quan đơn giữa các biến dự đoán và hàm phân biệt.
- Total correlation matrix: ma trận tương quan toàn bộ là ma trận tương quan khi các quan sát được coi như xuất phát từ một mẫu duy nhất.
- Wilks'λ : đôi khi được gọi là đại lượng thống kê U (U statistic), Wilks'λ đối với mỗi biến dự đoán là tỉ số giữa tổng các độ lệch bình phương trong nội bộ các nhóm và tổng các độ lệch bình phương toàn bộ. Giá trị của nó nằm trong khoảng từ 0 đến 1. λ lớn (gần 1) cho biết các trung bình nhóm dường như không khác nhau. λ nhỏ (gần 0) cho biết các trung bình nhóm dường như giống nhau.

Các giả định trong phân tích biệt số là: mỗi nhóm là một mẫu của một tổng thể có phân phối chuẩn đa biến và tất cả các tổng thể này có ma trận hiệp phương sai giống nhau (các phương sai bằng nhau).

5. CÁC BƯỚC TIẾN HÀNH PHÂN TÍCH BIỆT SỐ



5.1. Xác định vấn đề nghiên cứu

Bước thứ nhất là xác định vấn đề nghiên cứu bao gồm nhận biết các mục tiêu, biến phụ thuộc (criterion), và các biến độc lập. Biến phụ thuộc phải là biến có hai hay nhiều loại biểu hiện. Khi biến phụ thuộc được đo trên thang khoảng cách hay tỉ lệ, ta phải chuyển đổi về kiểu biến phân loại. Ví dụ thái độ đối với một nhãn hiệu được đo trên thang khoảng cách có 6 điểm có thể được chuyển đổi thành biến phân loại là không thuận lợi (gồm 1,2,3) và thuận lợi (gồm 4,5,6).

Bước thứ hai là chia mẫu quan sát thành hai phần. Phần dùng để ước lượng hàm phân biệt được gọi là *mẫu ước lượng hay mẫu phân tích* (estimation or analysis sample). Phần còn lại được gọi là *mẫu kiểm tra* (holdout or validation sample) dùng để kiểm tra tính đúng đắn của hàm phân biệt. Khi mẫu đủ lớn, ta có thể chia mẫu thành hai phần bằng nhau: một để phân tích và một để kiểm tra. Vai trò của hai nửa mẫu này có thể được thay đổi và việc phân tích được lập lại. Điều này gọi là kiểm tra chéo hai lần (double cross-validation). Thông thường phân phối của số quan sát trong mẫu phân tích và mẫu kiểm tra giống như phân phối trong toàn bộ mẫu.

Ví dụ nếu toàn bộ mẫu gồm 50% người tiêu dùng trung thành và 50% người tiêu dùng không trung thành, thì mẫu phân tích và mẫu kiểm tra cũng gồm 50% trung thành và 50% không trung thành.

Để minh họa phân tích biệt số hai nhóm, chúng ta hãy xem xét một ví dụ. Giả sử chúng ta muốn xác định các đặc trưng nổi bật của các gia đình đã đến một khu nghỉ mát (resort) trong vòng hai năm qua. Các dữ liệu được thu thập từ một mẫu gồm 42 hộ gia đình và lưu trữ trong file *phan tich biet so* trong tập dữ liệu dùng kèm với sách. Trong đó 30 hộ nằm trong mẫu phân tích và 12 hộ còn lại là mẫu kiểm tra, chúng ta chia toàn bộ mẫu quan sát thành mẫu phân tích và mẫu kiểm tra bằng biến *phantich*, biến này nhận giá trị 0 tại mẫu kiểm tra và 1 tại mẫu phân tích (khi tiến hành phân tích biệt số SPSS sẽ yêu cầu chúng ta khai báo các giá trị để xác định mẫu nào được xử lý). Các gia đình đã từng đến khu nghỉ mát này trong vòng hai năm qua có mã là 1, những gia đình không đến khu nghỉ mát này có mã là 2, đặc điểm này được thể hiện trong biến *nghimat*. Cả hai mẫu

phân tích và kiểm tra đều có tỉ lệ đi nghỉ mát tại khu nghỉ mát này bằng nhau. Các dữ liệu được thu thập là thu nhập trung bình tháng của gia đình, thái độ đối với du lịch (được đo trên thang đo 9 điểm), tầm quan trọng của sự gắn bó với kỳ nghỉ của gia đình (thang đo 9 điểm), quy mô hộ gia đình (người), và tuổi của người chủ hộ hay người chi phối chính quyết định của gia đình.

5.2. Ước lượng

Có hai phương pháp ước lượng các hệ số của hàm phân biệt:

- Phương pháp trực tiếp (Enter independents together): ước lượng hàm phân biệt khi tất cả các biến dự đoán được đưa vào cùng một lúc. Trong trường hợp này mỗi biến được đưa vào bất kể khả năng phân biệt của nó. Phương pháp này thích hợp khi dựa vào nghiên cứu trước đó hay mô hình lý thuyết, người nghiên cứu muốn hàm phân biệt được xây dựng trên tất cả các biến dự đoán.
- Phương pháp từng bước (Use stepwise method): các biến dự đoán được đưa vào hàm phân biệt một cách tuần tự dựa vào khả năng phân biệt được các nhóm của chúng. Phương pháp này thích hợp khi nhà nghiên cứu muốn chọn ra một tập con các biến dự đoán để đưa vào phương trình.

Các kết quả phân tích biệt số của các dữ liệu trong file ví dụ bằng SPSS được trình bày trong các Bảng 13.1. Nhìn vào các trung bình và độ lệch chuẩn của hai nhóm có từng đến đây nghỉ mát và không đến (Bảng 13.1a Group Statistics) chúng ta có thể cảm thấy/nhận được kết quả bằng trực giác. Hai nhóm dường như khá cách biệt nhau về thu nhập hơn là về các biến khác, và dường như cách biệt về tầm quan trọng của sự gắn bó với kỳ nghỉ của gia đình nhiều hơn cách biệt về thái độ đối với du lịch. Sự khác biệt về tuổi của chủ gia đình giữa hai nhóm khá nhỏ, và độ lệch chuẩn của tuổi khá lớn.

Ma trận tương quan trong nội bộ các nhóm chung cho thấy tương quan giữa các biến dự đoán khá thấp, như vậy có thể nói rằng hiện tượng cộng tuyến không đáng kể. Mức ý nghĩa của tỉ số F đơn biến cho thấy khi các biến dự đoán được xem xét một cách riêng biệt thì chỉ có thu nhập, tầm quan trọng của kỳ nghỉ và quy mô hộ gia đình có khả năng phân biệt một cách có ý nghĩa khác biệt giữa những người đã đến

nghỉ mát và những người không (bạn xem những thông tin này trong Bảng 13.1b, Tests of Equality of Group Means). Vì trường hợp này chúng ta chỉ có hai nhóm nên chỉ có một hàm phân biệt được ước lượng. Giá trị eigenvalue tương ứng của hàm này là 1,786 (Bảng 13.1e Eigenvalues), và nó chiếm tới 100% phương sai giải thích được nguyên nhân. Hệ số tương quan canonical tương ứng là 0.801. Bình phương của hệ số này, $(0,801)^2 = 0,64$, cho thấy 64% của phương sai biến phụ thuộc (nghỉ mát) được giải thích bởi mô hình này. Bước tiếp theo là xác định mức ý nghĩa.

5.3. Xác định mức ý nghĩa

Chúng ta không được giải thích kết quả phân tích nếu hàm phân biệt được ước lượng không có ý nghĩa về mặt thống kê. Giả thuyết không ở đây là trong tổng thể các trung bình của các hàm phân biệt trong tất cả các nhóm là bằng nhau, giả thuyết này phải được kiểm định xem có ý nghĩa thống kê không. Trong SPSS, kiểm định này được thực hiện dựa trên tiêu chuẩn Wilk λ. Nếu nhiều hàm phân biệt được kiểm định cùng một lúc (trong trường hợp phân tích biệt số bội), thì đại lượng Wilk λ là tích số của các đại lượng λ đơn biến của từng hàm. Mức ý nghĩa được ước lượng dựa trên phép biến đổi sang đại lượng Chi-square của đại lượng λ này. Trong ví dụ này, chúng ta thấy rằng đại lượng Wilk λ của hàm này là 0,359 (Bảng 13.1f), chuyển thành đại lượng Chi- square là 26,13 với 5 bậc tự do. Và mức ý nghĩa quan sát là 0,000 rất nhỏ so với 0,05. Ta đã có đủ cơ sở để bác bỏ giả thuyết không ở trên. Khi giả thuyết không bị bác bỏ, tức là sự phân biệt có ý nghĩa thống kê, chúng ta có thể tiến hành giải thích ý nghĩa kết quả.

5.4. Giải thích kết quả

Việc giải thích kết quả các hệ số của hàm phân biệt và các hệ số khác tương tự như trong trường hợp phân tích hồi qui bội. Trị số của hệ số của một biến dự đoán phụ thuộc vào việc đưa các biến dự đoán khác vào trong hàm phân biệt. Dấu của các hệ số này thì tùy ý, nhưng chúng cho biết biến nào làm cho trị số của hàm phân biệt lớn hay nhỏ và nên gắn chúng với nhóm nào. Sau đây là kết quả phân tích biệt số hai nhóm bằng SPSS cho ví dụ của chúng ta.

Bảng 13.1a **Group Statistics**

resort visit		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
co	thai do dv du lich	5.4000	1.91982	15	15.000
	gan bo voi gia dinh	5.8000	1.82052	15	15.000
	quy mo gia dinh	4.3333	1.23443	15	15.000
	thu nhap	2420.8000	393.22607	15	15.000
	tuoi chu ho	53.7333	8.77062	15	15.000
khong	thai do dv du lich	4.3333	1.95180	15	15.000
	gan bo voi gia dinh	4.0667	2.05171	15	15.000
	quy mo gia dinh	2.8000	.94112	15	15.000
	thu nhap	1676.5333	302.04585	15	15.000
	tuoi chu ho	50.1333	8.27101	15	15.000
Total	thai do dv du lich	4.8667	1.97804	30	30.000
	gan bo voi gia dinh	4.9333	2.09981	30	30.000
	quy mo gia dinh	3.5667	1.33089	30	30.000
	thu nhap	2048.6667	511.80932	30	30.000
	tuoi chu ho	51.9333	8.57395	30	30.000

Bảng 13.1b **Tests of Equality of Group Means**

	Wilks' Lambda	F	df1	df2	Sig.
thu nhap	.453	33.796	1	28	.000
thai do dv du lich	.925	2.277	1	28	.143
gan bo voi gia dinh	.824	5.990	1	28	.021
quy mo gia dinh	.657	14.636	1	28	.001
tuoi chu ho	.954	1.338	1	28	.257

Bảng 13.1c **Pooled Within-Groups Matrices(a)**

		thai do dv du lich	gan bo voi gia dinh	quy mo gia dinh	thu nhap	tuoi chu ho
Covariance	thai do dv du lich	3.748	.317	-.036	134.019	-3.252
	gan bo voi gia dinh	.317	3.762	.150	62.210	.288
	quy mo gia dinh	-.036	.150	1.205	34.200	-.402
	thu nhap	134.019	62.210	34.200	122929.219	-42.781
	tuoi chu ho	-3.252	.288	-.402	-42.781	72.667
Correlation	thai do dv du lich	1.000	.084	-.017	.197	-.197
	gan bo voi gia dinh	.084	1.000	.070	.091	.017
	quy mo gia dinh	-.017	.070	1.000	.089	-.043
	thu nhap	.197	.091	.089	1.000	-.014
	tuoi chu ho	-.197	.017	-.043	-.014	1.000

a The covariance matrix has 28 degrees of freedom.

Bảng 13.1d **Tests of Equality of Group Means**

	Wilks' Lambda	F	df1	df2	Sig.
thai do dv du lich	.925	2.277	1	28	.143
gan bo voi gia dinh	.824	5.990	1	28	.021
quy mo gia dinh	.657	14.636	1	28	.001
thu nhap	.453	33.796	1	28	.000
tuoi chu ho	.954	1.338	1	28	.257

Bảng 13.1e **Eigenvalues**

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.786(a)	100.0	100.0	.801

a First 1 canonical discriminant functions were used in the analysis.

Bảng 13.1f **Wilks' Lambda**

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.359	26.130	5	.000

Bảng 13.1g **Standardized Canonical Discriminant Function Coefficients**

	Function
	1
thai do dv du lich	.096
gan bo voi gia dinh	.233
quy mo gia dinh	.469
thu nhap	.743
tuoi chu ho	.209

Bảng 13.1h **Structure Matrix**

	Function
	1
thu nhap	.822
quy mo gia dinh	.541
gan bo voi gia dinh	.346
thai do dv du lich	.213
tuoi chu ho	.164

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.

Bảng 13.1i **Canonical Discriminant Function Coefficients**

	Function
	1
thai do dv du lich	.050
gan bo voi gia dinh	.120
quy mo gia dinh	.427
thu nhap	.002
tuoi chu ho	.025
(Constant)	-7.975

Unstandardized coefficients

Bảng 13.1k **Functions at Group Centroids**

resort visit	Function
	1
co	1.291
khong	-1.291

Unstandardized canonical discriminant functions evaluated at group means

Bảng 13.1 l trích một phần từ **Casewise Statistics**

Case Number		Actual Group	Predicted Group	Discriminant Scores Function 1
Original	1	1	2(**)	-.172
	2	1	1	2.330
	3	1	1	2.146
	4	1	1	.105
	5	1	1	.571
	.6	1	1	3.427
	7	1	2(**)	-.646
	8	1	1	1.253
	9	1	1	1.516
	10	1	1	2.108
	11	1	1	2.603
	12	1	1	2.610
	13	1	1	1.690
	14	1	2(**)	-.701
	15	1	1	.528
	16	2	2	-1.819
	17	2	2	-2.143
	18	2	2	-1.359
	19	2	2	-.597
	20	2	2	-.905
	21	2	2	-1.750
	22	2	2	-.769
	23	2	2	-.783
	24	2	2	-.522

	25	2	2	-.837
	26	2	2	-1.407
	27	2	2	-.647
	28	2	2	-1.802
	29	2	2	-1.947
	30	2	2	-2.079
	31(u)	1	2(**)	-.242
	32(u)	1	1	2.586
	33(u)	1	1	.875
	34(u)	1	2(**)	-.677
	35(u)	1	1	2.502
	36(u)	1	1	.915
	37(u)	2	2	-1.414
	38(u)	2	2	-.068
	39(u)	2	2	-1.374
	40(u)	2	2	-.629
	41(u)	2	2	-.457
	42(u)	2	2	-1.913

Bảng 13.1m **Classification Results(a,b)**

resort visit				Predicted Group Membership co	khong	Total
Cases Selected	Original	Count	co	12	3	15
			khong	0	15	15
		%	co	80.0	20.0	100.0
			khong	.0	100.0	100.0
Cases Not Selected	Original	Count	co	4	2	6
			khong	0	6	6
		%	co	66.7	33.3	100.0
			khong	.0	100.0	100.0

a 90.0% of selected original grouped cases correctly classified.

b 83.3% of unselected original grouped cases correctly classified.

Tầm quan trọng của các biến được thể hiện qua độ lớn tuyệt đối của hệ số chuẩn hóa của hàm phân biệt. Nói chung, các biến có hệ số chuẩn hóa càng lớn thì càng đóng góp nhiều hơn vào khả năng phân biệt của hàm (Bảng 13.1g Standardized Canonical Discriminant Function Coefficients). Tầm quan trọng của các biến cũng có thể được thể hiện qua các hệ số tương quan kết cấu (Bảng 13.1h Structure Matrix), bảng này được sắp theo thứ tự giảm dần của độ lớn), hay còn gọi là hệ số biệt tải hay trọng số canonical. Những hệ số tương quan đơn giữa từng biến dự đoán và hàm phân biệt này cho biết phần phương sai biến dự đoán này tham gia trong phương trình.

Trong ví dụ này, tương quan giữa các biến dự đoán khá yếu nên chúng ta có thể sử dụng độ lớn của các hệ số chuẩn hóa để nói rằng thu nhập là biến dự đoán quan trọng nhất dùng để phân biệt các nhóm, sau đó là quy mô hộ gia đình, và tầm quan trọng của sự gắn bó với kỳ nghỉ của gia đình. Từ các hệ số tương quan kết cấu, chúng ta cũng có thể rút ra kết luận tương tự. Những hệ số tương quan đơn giữa các biến dự đoán và hàm phân biệt được liệt kê theo độ lớn giảm dần.

Các hệ số hàm phân biệt không chuẩn hóa cũng được thể hiện trong bảng kết quả.(Bảng 13.1i). Chúng ta có thể áp dụng những hệ số này đối với các giá trị thô của các biến trong mẫu kiểm tra để phân biệt các quan sát sẽ thuộc nhóm nào. Biệt số trung bình nhóm(group centroids), giá trị của hàm phân biệt tại trung bình nhóm cũng được thể hiện trong Bảng kết quả11.1k. Nhóm1, những người đã đến nghỉ tại khu nghỉ mát này, có giá trị trung bình nhóm dương, trong khi Nhóm 2 có giá trị âm với độ lớn bằng nhau. Dấu của các hệ số của tất cả các biến dự đoán đều dương cho thấy rằng thu nhập hộ gia đình càng cao, quy mô gia đình càng lớn, sự gắn bó với kỳ nghỉ của gia đình càng quan trọng, thái độ đối với du lịch càng thuận lợi và tuổi chủ gia đình càng cao thì gia đình càng có khả năng đến khu nghỉ mát này. Chúng ta nên xây dựng một bản tiểu sử (profile) của hai nhóm dựa trên 3 biến dự đoán quan trọng nhất là: thu nhập, quy mô gia đình, và tầm quan trọng của kỳ nghỉ. Giá trị các biến này của hai nhóm được thể hiện trong phần đầu của các Bảng 13.1.

5.5. Đánh giá

Như đã đề cập trong phần trên, các dữ liệu được chia thành hai phần: mẫu phân tích dùng để ước lượng hàm phân biệt, và mẫu kiểm tra dùng để xây dựng ma trận (bảng) phân loại. Các hệ số phân biệt (discriminant weights) sau khi được ước lượng từ mẫu phân tích, sẽ được nhân với các giá trị của các biến dự đoán trong mẫu kiểm tra để tính biệt số của từng quan sát trong mẫu kiểm tra. Sau đó các quan sát này được phân vào các nhóm dựa trên biệt số của chúng và dựa trên một nguyên tắc quyết định thích hợp. Trong phân tích biệt số hai nhóm thì các quan sát được phân vào nhóm có centroid gần nhất. Nguyên tắc quyết định thường là tính ra một điểm phân biệt

(cutting point), nếu quan sát có biệt số lớn hơn giá trị này thì sẽ được xếp vào nhóm có centroid lớn, ngược lại nếu biệt số của quan sát nhỏ hơn giá trị này thì sẽ được xếp vào nhóm có centroid nhỏ. Cụ thể là:

- Nếu hai nhóm có quy mô bằng nhau thì điểm phân biệt là trung bình cộng giản đơn của hai centroid của hai nhóm:

$$Z_{CE} = \frac{\bar{Z}_A + \bar{Z}_B}{2}$$

- Nếu hai nhóm có quy mô khác nhau thì điểm phân biệt là trung bình cộng gia quyền của hai centroid của hai nhóm:

$$Z_{CE} = \frac{N_A \bar{Z}_A + N_B \bar{Z}_B}{N_A + N_B}$$

Trong đó:

Z_{CE} : điểm phân biệt

N_A : số quan sát thuộc nhóm A trong mẫu phân tích

N_B : số quan sát thuộc nhóm B trong mẫu phân tích

$\bar{Z}_A$: centroid của nhóm A

$\bar{Z}_B$: centroid của nhóm B

Sau đó, tỉ lệ đúng (hit ratio) hay phần trăm số quan sát được phân loại đúng, được tính bằng cách cộng các con số trên đường chéo của bảng kết quả phân loại (Classification Results) và chia cho tổng số quan sát. Kết quả phân loại tính từ mẫu phân tích luôn luôn tốt hơn kết quả phân loại tính từ mẫu kiểm tra bởi vì hàm phân biệt được ước lượng từ các dữ liệu trong mẫu phân tích.

Trong ví dụ của chúng ta, vì hai centroid của hai nhóm bằng nhau về độ lớn và có dấu ngược nhau (đối xứng nhau qua trị số 0), và quy mô của hai nhóm bằng nhau nên trung bình cộng giản đơn của chúng (bằng 0) sẽ là điểm phân biệt (cutting point) giữa hai nhóm. Các quan sát sẽ được phân biệt theo tiêu chuẩn quyết định này, tức là những quan sát nào có biệt số lớn hơn 0 sẽ được xếp vào nhóm 1 (đã đến nghỉ mát) và những quan sát nào có biệt số nhỏ hơn 0 được xếp vào nhóm 2 (không đến nghỉ mát).

Chúng ta có thể quan sát thấy rõ điều này trong Bảng 13.1l được trích một phần từ bảng Casewise Statistics mà SPSS đưa ra, bảng này liệt kê kết quả phân biệt của từng quan sát một. Trong phần liệt kê chúng ta thấy có ba quan sát trong mẫu phân tích bị phân biệt sai, đó là các quan sát thứ 1, 7 và 14. Theo hàm phân biệt được xây dựng thì 3 quan sát này được xếp vào trong nhóm 2 là nhóm không đến nghỉ mát (vì có biệt số lần lượt là -0,1721, -0,6463 và –0,7007 đều nhỏ hơn 0). Nhưng trong thực tế những người này đã có đến nghỉ mát tại khu nghỉ mát này. Như vậy có 3 quan sát trong mẫu phân tích bị phân biệt sai. Tương tự như vậy ta có thể quan sát kết quả phân biệt đối với mẫu kiểm tra khi sử dụng hàm phân biệt ước lượng từ mẫu phân tích và thấy có 2 quan sát bị phân biệt sai.

Ở Bảng 13.1m Classification Results(a,b), chúng ta có thể thấy kết quả phân loại dựa trên mẫu phân tích. Tỉ lệ phân biệt đúng là (12+15)/30 = 0,90 hay 90%. Chúng ta có thể nghi ngờ là tỉ lệ đúng này bị phóng đại một cách giả tạo vì các dữ liệu dùng để ước lượng cũng được sử dụng để kiểm tra đánh giá. Thực hiện phân tích biệt số trên mẫu kiểm tra độc lập, chúng ta có một tỉ lệ đúng hơi thấp hơn một chút là (4+6)/12 = 0,833 hay 83,3%. Do hai nhóm có quy mô bằng nhau nên nếu chúng ta phân biệt hai nhóm một cách ngẫu nhiên thì tỉ lệ đúng có thể là ½ = 50%. Do đó phân tích biệt số này giúp ta cải thiện được khả năng phân biệt đúng lên khoảng 33,3%. Chỉ cần cải thiện khả năng phân biệt được 25% là thoả mãn, do đó có thể nói mô hình phân tích biệt số này là khá tốt.

6. PHÂN TÍCH BIỆT SỐ BỘI

6.1. Xác định mô hình

Các dữ liệu của file ví dụ được sử dụng lại để minh họa cho phân tích biệt số ba nhóm. Trong cột cuối cùng của hai bảng này, các hộ gia đình được chia thành 3 nhóm theo mức độ chi tiêu cho nghỉ mát của gia đình (nhiều, trung bình và ít). Có 10 hộ gia đình trong mỗi nhóm. Vấn đề ở đây là chúng ta có thể phân biệt các gia đình theo ba mức chi tiêu cho kỳ nghỉ gia đình nhờ vào mức thu nhập hộ gia đình, thái độ đối với du lịch, tầm quan trọng của việc gắn bó với kỳ

nghỉ gia đình, quy mô gia đình, và tuổi của người chủ gia đình hay không?

6.2. Ước lượng

Các Bảng 13.2 trình bày các kết quả ước lượng phân tích biệt số 3 nhóm. Các trung bình nhóm cho thấy dường như thu nhập phân chia các nhóm rõ hơn các biến khác. Có sự khác biệt giữa các nhóm về thái độ đối với du lịch và tầm quan trọng của kỳ nghỉ. Nhóm 1 và nhóm 2 có quy mô gia đình và tuổi của chủ hộ rất gần nhau. Tuổi có độ lệch chuẩn khá lớn so với khả năng phân biệt các nhóm của nó. Ma trận hệ số tương quan nội bộ nhóm chung cho thấy thu nhập có tương quan với tầm quan trọng của kỳ nghỉ và quy mô gia đình. Tuổi có tương quan nghịch với thái độ đối với du lịch. Tuy nhiên, các tương quan này khá lỏng lẻo nên vấn đề đa cộng tuyến không đáng kể. Mức ý nghĩa tương ứng với các tỉ số F đơn biến cho thấy khi các biến dự đoán được xem xét riêng rẽ thì chỉ có thu nhập gia đình và thái độ đối với du lịch là có ý nghĩa trong việc phân biệt các nhóm.

Trong phân tích biệt số bội, nếu có G nhóm thì sẽ có G-1 hàm được ước lượng nếu số biến dự đoán lớn hơn con số này. Tổng quát thì với G nhóm và k biến dự đoán, số hàm phân biệt sẽ bằng con số nhỏ hơn trong hai con số G-1 và k. Hàm thứ nhất có tỉ số giữa tổng các độ lệch giữa các nhóm và tổng các độ lệch trong nội bộ nhóm lớn nhất. Hàm thứ hai, không có tương quan với hàm thứ nhất, có tỉ số này lớn thứ hai, v.v... . Tuy nhiên không phải tất cả các hàm đều có ý nghĩa về mặt thống kê. Vì trong ví dụ này chúng ta có 3 nhóm nên chỉ có hai hàm được ước lượng. Giá trị eigenvalue của hàm thứ nhất là 3,819 và hàm này chiếm tới 93,9% phương sai của dữ liệu. Hàm thứ hai có giá trị eigenvalue là 0,247 và chỉ chiếm 6,1 % của phương sai.

Kết quả phân tích biệt số 3 nhóm bằng SPSS for Windows thể hiện trong các Bảng 13.

Bảng 13.2a **Group Statistics**

chi tieu cho ky nghi GD		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
it	thai do dv du lich	4.5000	1.71594	10	10.000
	gan bo voi gia dinh	4.7000	1.88856	10	10.000
	quy mo gia dinh	3.1000	1.19722	10	10.000
	thu nhap	1542.8000	211.88718	10	10.000
	tuoi chu ho	50.3000	8.09732	10	10.000
trung binh	thai do dv du lich	4.0000	2.35702	10	10.000
	gan bo voi gia dinh	4.2000	2.48551	10	10.000
	quy mo gia dinh	3.4000	1.50555	10	10.000
	thu nhap	2004.4000	240.09220	10	10.000
	tuoi chu ho	49.5000	9.25263	10	10.000
nhieu	thai do dv du lich	6.1000	1.19722	10	10.000
	gan bo voi gia dinh	5.9000	1.66333	10	10.000
	quy mo gia dinh	4.2000	1.13529	10	10.000
	thu nhap	2598.8000	344.57342	10	10.000
	tuoi chu ho	56.0000	7.60117	10	10.000
Total	thai do dv du lich	4.8667	1.97804	30	30.000
	gan bo voi gia dinh	4.9333	2.09981	30	30.000
	quy mo gia dinh	3.5667	1.33089	30	30.000
	thu nhap	2048.6667	511.80932	30	30.000
	tuoi chu ho	51.9333	8.57395	30	30.000

Bảng 13.2b **Tests of Equality of Group Means**

	Wilks' Lambda	F	df1	df2	Sig.
thai do dv du lich	.788	3.634	2	27	.040
gan bo voi gia dinh	.881	1.830	2	27	.180
quy mo gia dinh	.874	1.944	2	27	.163
thu nhap	.262	37.997	2	27	.000
tuoi chu ho	.882	1.804	2	27	.184

Bảng 13.2c **Pooled Within-Groups Matrices**

		thai do dv du lich	gan bo voi gia dinh	quy mo gia dinh	thu nhap	tuoi chu ho
Correlation	thai do dv du lich	1.000	.036	.005	.051	-.340
	gan bo voi gia dinh	.036	1.000	.221	.307	-.013
	quy mo gia dinh	.005	.221	1.000	.380	-.025
	thu nhap	.051	.307	.380	1.000	-.209
	tuoi chu ho	-.340	-.013	-.025	-.209	1.000

Bảng 13.2d **Eigenvalues**

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3.819(a)	93.9	93.9	.890
2	.247(a)	6.1	100.0	.445

a First 2 canonical discriminant functions were used in the analysis.

Cả hai hàm được xem xét một lúc

Bảng 13.2e **Wilks' Lambda**

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.166	44.831	10	.000
2	.802	5.517	4	.238

Hàm thứ nhất đã được lấy ra

Bảng 13.2f **Standardized Canonical Discriminant Function Coefficients**

	Function	
	1	2
thai do dv du lich	.340	.769
gan bo voi gia dinh	-.142	.534
quy mo gia dinh	-.163	.129
thu nhap	1.047	-.421
tuoi chu ho	.495	.524

Bảng 13.2g **Structure Matrix**

	Function	
	1	2
thu nhap	.856(*)	-.278
quy mo gia dinh	.193(*)	.077
thai do dv du lich	.219	.588(*)
gan bo voi gia dinh	.149	.454(*)
tuoi chu ho	.166	.341(*)

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.

* Largest absolute correlation between each variable and any discriminant function

Bảng 13.2h **Canonical Discriminant Function Coefficients**

	Function	
	1	2
thai do dv du lich	.187	.422
gan bo voi gia dinh	-.070	.261
quy mo gia dinh	-.127	.100
thu nhap	.004	-.002
tuoi chu ho	.059	.063
(Constant)	-11.094	-3.792

Unstandardized coefficients

Bảng 13.2i **Functions at Group Centroids**

chi tieu cho ky nghi GD	Function	
	1	2
it	-2.041	.418
trung binh	-.405	-.659
nhieu	2.446	.240

Unstandardized canonical discriminant functions evaluated at group means

Bảng 13.2k **Prior Probabilities for Groups**

chi tieu cho ky nghi GD	Prior	Specified Prior	Effective Prior	Cases Used in Analysis	
				Unweighted	Weighted
it	.333			10	10.000
trung binh	.333			10	10.000
nhieu	.333			10	10.000
Total	1.000			30	30.000

Bảng 13.21 trích từ bảng **Casewise Statistic**

Case Number		Actual Group	Highest Group				
			Predicted Group	P(D>d | G=g)		P(G=g | D=d)	Squared Mahalanobis Distance to Centroid
				p	df		
Original	1	2	2	.582	2	.555	1.082
	2	3	3	.564	2	1.000	1.144
	3	3	3	.841	2	.955	.347
	4	1	1	.661	2	.564	.828
	5	3	2(**)	.198	2	.560	3.241
	6	3	3	.022	2	1.000	7.630
	7	2	2	.493	2	.747	1.413
	8	2	2	.562	2	.959	1.153
	9	3	3	.955	2	.994	.091
	10	3	3	.918	2	.978	.171
	11	3	3	.589	2	.991	1.059
	12	3	3	.472	2	1.000	1.500
	13	2	2	.993	2	.883	.014
	14	3	2(**)	.887	2	.913	.240
	15	3	3	.906	2	.967	.197
	16	1	1	.617	2	.974	.965
	17	1	1	.946	2	.809	.111
	18	2	2	.678	2	.588	.778
	19	2	2	.610	2	.897	.989
	20	2	1(**)	.849	2	.751	.327
	21	1	1	.861	2	.928	.299
	22	2	2	.121	2	.962	4.229
	23	1	2(**)	.768	2	.655	.529
	24	1	1	.213	2	.961	3.092
	25	1	1	.772	2	.954	.517
	26	2	2	.814	2	.666	.411
	27	2	2	.446	2	.698	1.615
	28	1	1	.198	2	.991	3.237
	29	1	1	.336	2	.732	2.182
	30	1	1	.170	2	.512	3.538
	31(u)	2	2	.893	2	.733	.227
	32(u)	3	3	.932	2	.976	.140
	33(u)	2	3(**)	.152	2	.636	3.764
	34(u)	2	2	.424	2	.637	1.714
	35(u)	3	3	.753	2	.943	.568
	36(u)	3	3	.819	2	.936	.398
	37(u)	1	1	.910	2	.923	.189
	38(u)	1	2(**)	.767	2	.780	.530
	39(u)	3	1(**)	.968	2	.896	.065
	40(u)	1	1	.564	2	.958	1.145
	41(u)	2	2	.811	2	.904	.419
	42(u)	1	1	.173	2	.646	3.510

u Unselected case
** Misclassified case

Bảng 13.2 m **Classification Results(a,b)**

chi tieu cho ky nghi GD				Predicted Group Membership			Total
				it	trung binh	nhieu	
Cases Selected	Original	Count	it	9	1	0	10
			trung binh	1	9	0	10
			nhieu	0	2	8	10
		%	it	90.0	10.0	.0	100.0
			trung binh	10.0	90.0	.0	100.0
			nhieu	.0	20.0	80.0	100.0
Cases Not Selected	Original	Count	it	3	1	0	4
			trung binh	0	3	1	4
			nhieu	1	0	3	4
		%	it	75.0	25.0	.0	100.0
			trung binh	.0	75.0	25.0	100.0
			nhieu	25.0	.0	75.0	100.0

a 86.7% of selected original grouped cases correctly classified.
b 75.0% of unselected original grouped cases correctly classified.

6.3. Xác định mức ý nghĩa

Để kiểm định giả thiết không các nhóm có centroid bằng nhau, cả hai hàm phải được xem xét cùng một lúc. Chúng ta có thể kiểm định các giá trị trung bình của các hàm liên tiếp nhau bằng cách trước tiên kiểm định tất cả các trung bình đồng thời. Sau đó loại trừ một hàm và các trung bình của các hàm còn lại được kiểm định trong bước kế tiếp. Trong Bảng 13.2e Wilks' Lambda, trị số của đại lượng Wilk λ là 0,166 tương đương với đại lượng chi-square là 44,831 với 10 bậc tự do, và có mức ý nghĩa quan sát nhỏ hơn mức 0,05. Do đó cả hai hàm này cùng một lúc có khả năng phân biệt 3 nhóm một cách có ý nghĩa. Tuy nhiên khi hàm thứ nhất được lấy ra, Wilk λ của hàm thứ hai là 0,802 không có ý nghĩa ở mức 0,05. Vì vậy hàm thứ hai không có khả năng phân biệt các nhóm một cách có ý nghĩa.

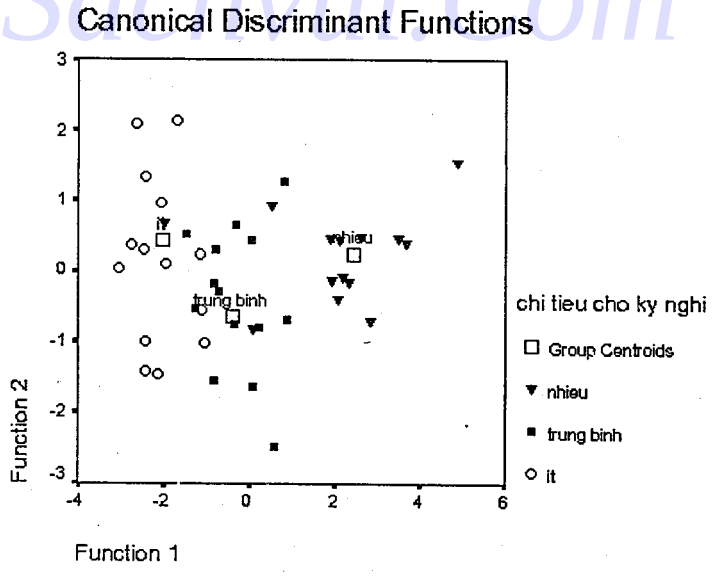
6.4. Giải thích

Việc giải thích các kết quả được thực hiện bằng cách xem xét các hệ số chuẩn hóa của hàm phân biệt, các hệ số tương quan kết cấu, và một số biểu đồ. Các hệ số chuẩn hóa ở Bảng 13.2f cho thấy thu nhập có hệ số lớn trong hàm thứ nhất, trong khi hàm thứ hai có hệ số lớn đối với các biến thái độ đối với du lịch, tầm quan trọng của kỳ nghỉ

và tuổi của chủ gia đình. Điều này cũng được thể hiện qua ma trận kết cấu (Bảng 13.2g Structure Matrix). Để giải thích các hàm, các biến có hệ số lớn trong cùng một hàm được nhóm chung với nhau. Việc phân nhóm này được thể hiện bằng các dấu sao (*). Do đó thu nhập và quy mô gia đình được đánh dấu sao ở hàm thứ nhất vì hai biến này có hệ số trong hàm thứ nhất lớn hơn hệ số trong hàm thứ hai. Hai biến này chủ yếu gắn kết với hàm thứ nhất. Mặt khác, thái độ đối với du lịch, tầm quan trọng của kỳ nghỉ và tuổi của chủ gia đình gắn kết với hàm thứ hai và cũng được đánh dấu sao ở hàm thứ hai.

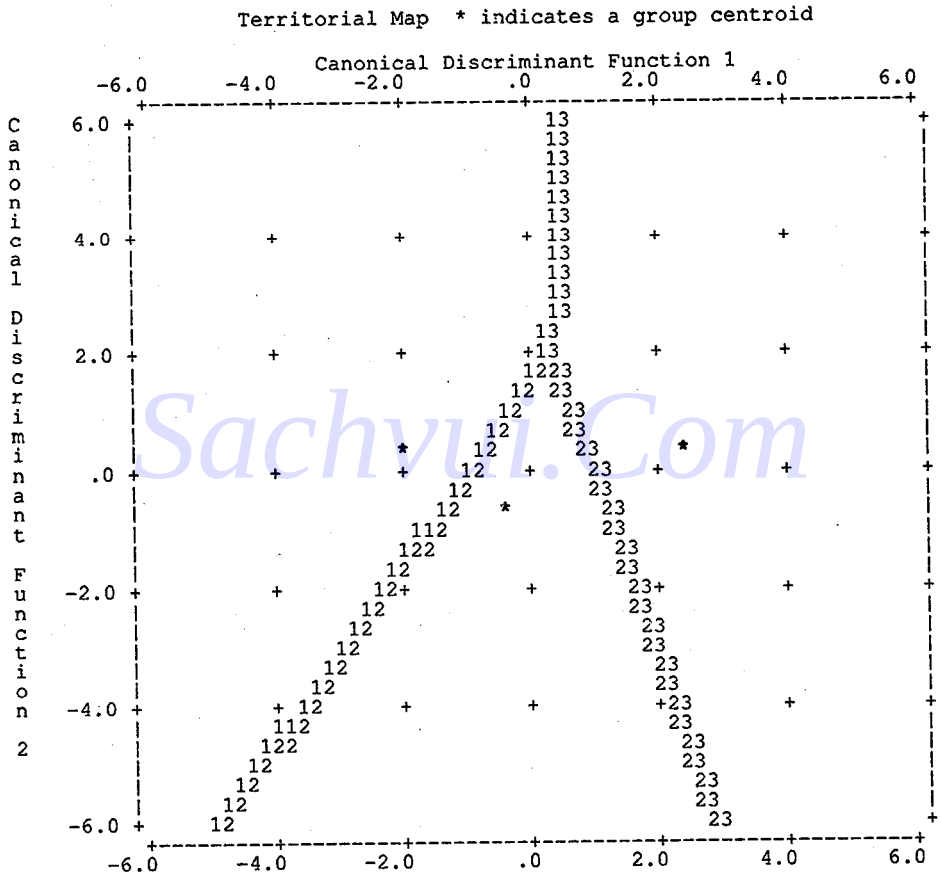
Hình 13.1 là một biểu đồ phân tán của các nhóm diễn tả theo hàm thứ nhất và hàm thứ hai. Chúng ta có thể thấy rằng nhóm 3 có trị số cao nhất theo hàm thứ nhất, và nhóm 1 là thấp nhất. Vì hàm thứ nhất chủ yếu gắn với thu nhập và quy mô gia đình, nên chúng ta có thể nghĩ rằng 3 nhóm có thể được phân biệt theo hai biến này. Những gia đình có thu nhập cao và quy mô lớn thường chi tiêu nhiều cho các kỳ nghỉ, Ngược lại, những gia đình có có thu nhập thấp và quy mô nhỏ thường chi tiêu ít cho kỳ nghỉ. Kết luận này được cũng cố thêm qua các trung bình nhóm về thu nhập và quy mô gia đình.

Hình 13.1



Hình 13.2 Biểu đồ vị trí

```
Symbols used in territorial map
Symbol  Group  Label
------  -----  --------------------
   1        1  it
   2        2  trung binh
   3        3  nhieu
   *           Group centroids
```



Hình 13.1 cũng cho thấy hàm thứ hai có khả năng phân biệt nhóm 1 (trị số hàm thứ hai lớn nhất) và nhóm 2 (trị số hàm thứ hai nhỏ nhất). Hàm này chủ yếu gắn với thái độ đối với du lịch, tầm quan trọng của kỳ nghỉ và tuổi chủ gia đình. Vì các biến này có tương quan thuận với hàm thứ hai trong ma trận kết cấu nên chắc chắn là chúng ta sẽ thấy nhóm 1 hơn nhóm 2 xét về thái độ đối với du lịch, tầm quan trọng của kỳ nghỉ và tuổi chủ gia đình. Quả thật điều này đúng đối

với thái độ đối với du lịch và tầm quan trọng của kỳ nghỉ như các trung bình nhóm của hai biến này trong bảng kết quả thể hiện. Nếu các gia đình trong nhóm 1 có thái độ thuận lợi hơn đối với du lịch và coi sự gắn bó với kỳ nghỉ gia đình quan trọng hơn nhóm 2, thì tại sao họ lại chi tiêu ít hơn. Có lẽ họ muốn chi nhiều cho các kỳ nghỉ gia đình nhưng không thể chi nhiều như vậy vì thu nhập của các gia đình này thấp.

Chúng ta cũng có thể kết luận tương tự như vậy bằng cách xem xét biểu đồ vị trí (territorial map) ở Hình 13.2. Trong biểu đồ vị trí, centroid của mỗi nhóm được biểu diễn bằng một dấu sao (*). Ranh giới của các nhóm được thể hiện bằng các con số tương ứng của các nhóm (1,2,3). Do đó centroid của nhóm 1 bị vây quanh bởi các số 1, nhóm 2 và 3 cũng tương tự như vậy.

6.5. Đánh giá

Các kết quả phân biệt dựa vào mẫu phân tích cho thấy có (9+9+8)/30 = 86,67% các quan sát đã được phân loại đúng. Áp dụng hàm phân biệt này trên mẫu kiểm chứng độc lập thì chúng ta thấy rằng tỉ lệ đúng lại thấp hơn một chút là (3+3+3)/12 = 75%. Vì 3 nhóm có quy mô bằng nhau nên nếu chỉ phân biệt một cách tình cờ thì tỉ lệ đúng kỳ vọng là 1/3 = 33,33%. Vì mức cải thiện kết quả phân biệt chỉ cần 25% là đã có ý nghĩa, cho nên chúng ta có thể nói rằng phân tích biệt số này là tốt.

7. PHÂN TÍCH BIỆT SỐ BỘI THEO PHƯƠNG PHÁP TỪNG BƯỚC (Stepwise discriminant analysis)

Phân tích biệt số bội từng bước cũng tương tự như hồi qui bội từng bước ở chỗ các biến dự đoán được tuần tự đưa vào hàm phân biệt căn cứ vào khả năng phân biệt các nhóm của chúng. Tỉ số F của từng biến dự đoán sẽ được tính qua phân tích phương sai đơn biến trong đó các nhóm được coi như được hình thành từ một biến phân loại và biến dự đoán được coi như biến độc lập. Biến dự đoán có tỉ số F lớn nhất sẽ là biến đầu tiên được chọn đưa vào hàm phân biệt, nếu nó thỏa những điều kiện về mức ý nghĩa và độ chấp nhận. Biến dự đoán thứ hai được chọn căn cứ vào tỉ số F từng phần hay F điều chỉnh lớn nhất, tức là có tính đến biến dự đoán đã đưa vào hàm phân biệt.

Mỗi biến dự đoán được chọn sẽ được kiểm định khả năng ở lại trong hàm trên cơ sở mối liên hệ của nó với các biến dự đoán đã được chọn khác. Quá trình lựa chọn và giữ lại cứ tiếp tục cho đến khi tất cả các biến dự đoán thỏa các tiêu chuẩn mức ý nghĩa để đưa vào và giữ lại đã được đưa hết vào trong hàm phân biệt. Ở mỗi bước, sẽ có nhiều thông số thống kê được tính toán. Ngoài ra, chương trình cũng cho ra một bảng tóm tắt các biến được đưa vào và loại ra khỏi hàm.

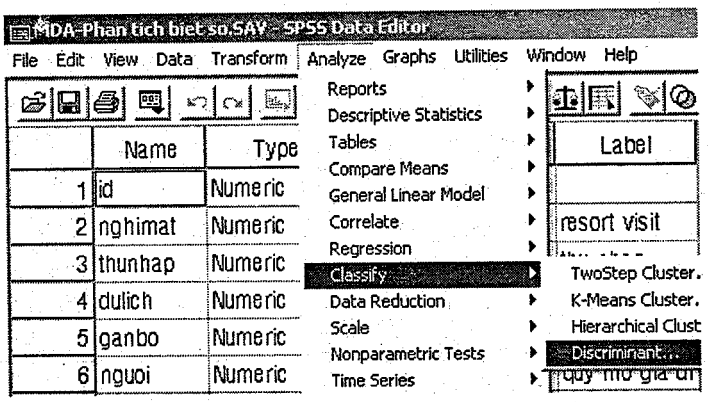
Việc chọn các biến trong thủ tục từng bước này còn phụ thuộc vào tiêu chuẩn tối ưu được chọn. Có 5 phương pháp (tiêu chuẩn) được sử dụng:

- Wilks' lambda: cực tiểu hóa trị số của đại lượng Wilk λ.
- Unexplained variance: cực tiểu hóa tổng các phần phương sai không giải thích được đối với tất cả các nhóm.
- Mahalanobis' distance: tối đa hóa khoảng cách Mahalanobis đối với hai nhóm gần nhất.
- Smallest F ratio: tối đa hóa tỉ số F nhỏ nhất giữa hai nhóm bất kỳ.
- Rao's V: tối đa hóa đại lượng Rao V. Chúng ta có thể chỉ định mức gia tăng tối thiểu của V mà một biến phải thỏa để đưa vào.

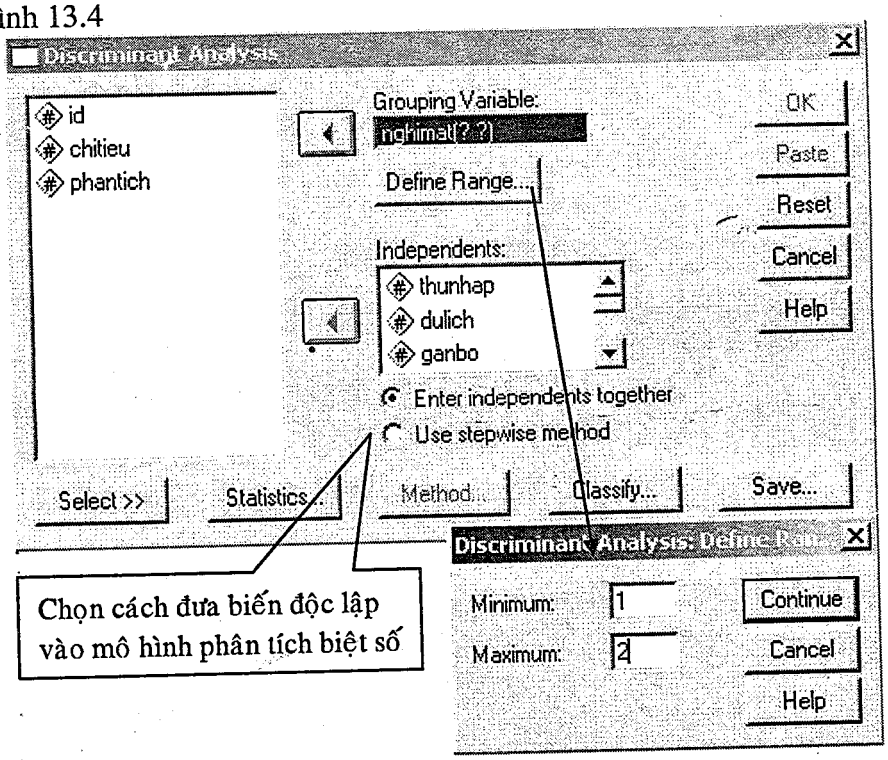
8. THỰC HIỆN PHÂN TÍCH BIỆT SỐ BẰNG SPSS

Từ menu ta chọn Statistics> Classify > Discriminant như Hình 13.3, lệnh này sẽ mở hộp thoại phân tích biệt số ở Hình 13.4.

Hình 13.3



Hình 13.4



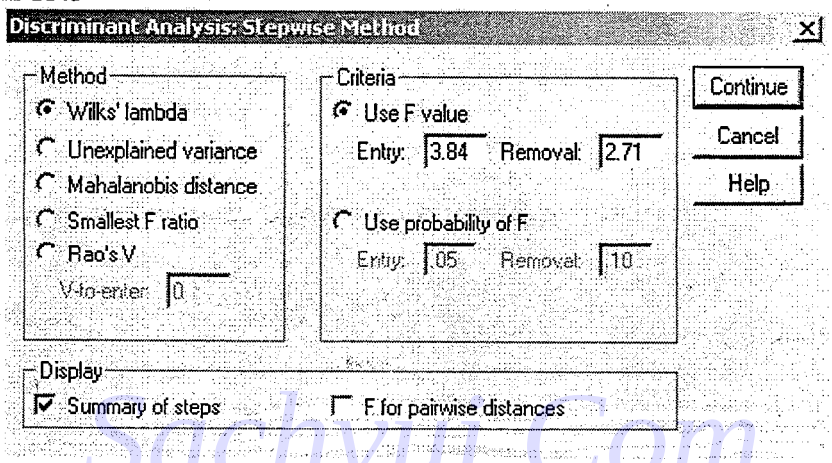
Các biến số trong file dữ liệu sẽ xuất hiện trong ô danh sách biến nguồn ở phía trái của hộp thoại ở Hình 13.4. Bạn hãy chọn một biến phụ thuộc định tính (có 2 hay 3 biểu hiện) bằng cách nhấn chuột vào tên của biến trong ô danh sách biến nguồn, rồi nhấn chuột vào mũi tên qua phải chỉ vào ô Grouping Variable (biến phân nhóm).

Lúc này nút Define Range sẽ nổi rõ lên để báo cho bạn biết cần phải nhấn vào nó để xác định các nhóm mà bạn muốn phân tích biệt số. Nếu có 2 nhóm thì ta sẽ có phân tích biệt số đơn, từ 3 nhóm trở lên ta sẽ có phân tích biệt số bội. Sau khi xác định phạm vi phân tích, nhấn nút Continue để trở về hộp thoại phân tích biệt số ban đầu.

Bước tiếp theo là bạn hãy chọn các biến độc lập định lượng trong danh sách biến nguồn bằng cách trỏ chuột vào các tên biến tương ứng rồi nhấn chuột vào nút mũi tên qua phải chỉ vào ô Independents. Ngay dưới ô chứa các biến độc lập này, bạn có thể chọn cách đưa biến độc lập vào mô hình phân tích biệt số: Enter hay Stepwise.

Cách đưa vào mặc định là Enter. Nếu bạn chọn Stepwise thì nút Method ngay bên dưới sẽ nổi rõ lên để báo cho bạn biết cần phải nhấn chuột vào nó để chọn và phương pháp và tiêu chuẩn đưa biến vào như ở Hình 13.5 dưới đây. Sau khi chọn xong bạn hãy nhấn vào nút Continue để trở lại hộp thoại phân tích biệt số ban đầu.

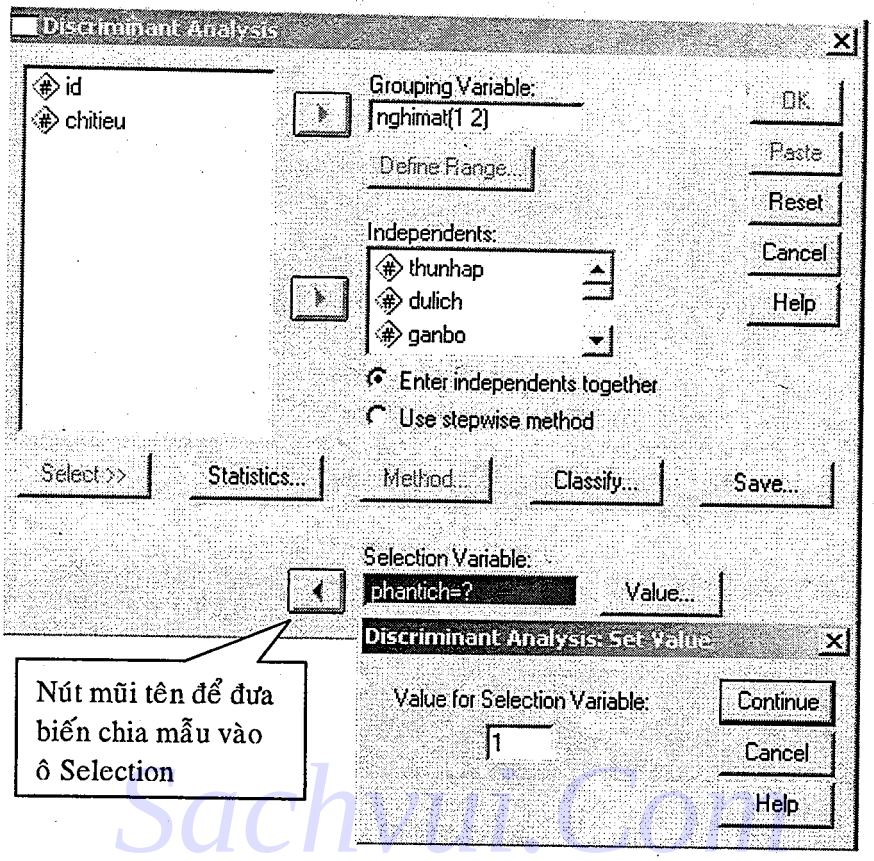
Hình 13.5



Bước tiếp theo là bạn có thể nhấn vào nút Select để xác định các quan sát nào được đưa vào mẫu phân tích, các quan sát nào được đưa vào mẫu kiểu tra. Muốn vậy khi nhập liệu, bạn phải tạo một biến định tính có hai biểu hiện là 0 và 1 (trong ví dụ này, biến chia mẫu này có tên là *phantich* chúng ta đã nói ở đầu). Quan sát nào bạn muốn đưa vào phân tích sẽ có mã là 1, quan sát nào dùng để kiểm tra có mã là 0. Khi nhấn chuột vào nút này thì hộp thoại phân tích biệt số sẽ nở thêm xuống phía dưới như Hình 13.6 sau.

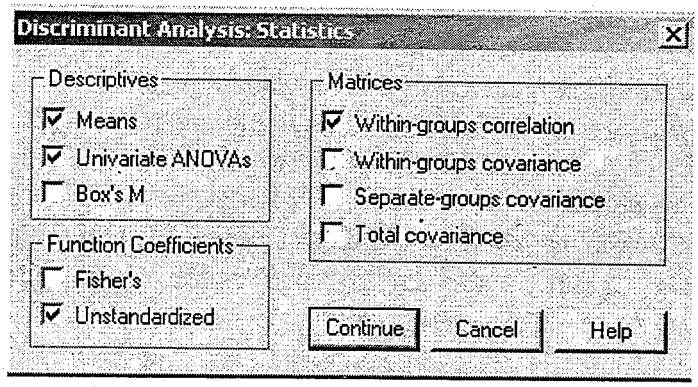
Trong phần nở ra này, bạn phải chọn biến chia mẫu từ danh sách biến nguồn rồi nhấp nút mũi tên để đưa vào ô Selection Variable. Sau đó nhấn nút Value bên cạnh để mở hộp thoại con gõ vào mã của các quan sát mà bạn muốn chọn đưa vào phân tích (trong ví dụ này mã chọn là 1). Rồi nhấn nút Continue để trở lại hộp thoại ban đầu.

Hình 13.6



Bước tiếp theo là bạn nhấn vào nút Statistics trong hộp thoại phân tích biệt số, một hộp thoại con sẽ xuất hiện như sau:

Hình 13.7

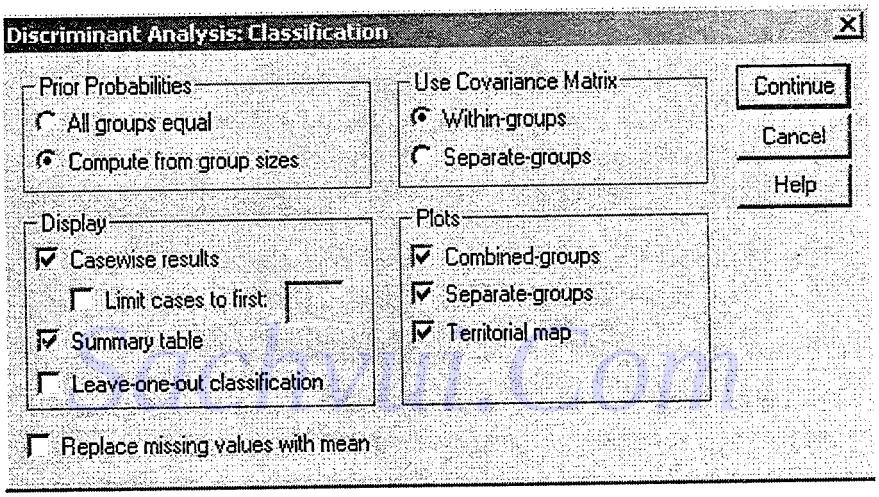


Trong hộp thoại con Statistics này, bạn có thể nhấp chuột vào những

ô chọn tương ứng để báo cho chương trình chạy ra các tham số thống kê mô tả như: trung bình, bảng phân tích phương sai đơn, các ma trận hệ số tương quan và hiệp phương sai, các hệ số của mô hình phân tích biệt số ... Rồi nhấn nút Continue để trở lại hộp thoại ban đầu.

Bước tiếp theo là bạn nhấn chuột vào nút Classify, một hộp thoại con sẽ xuất hiện tiếp như sau:

Hình 13.8

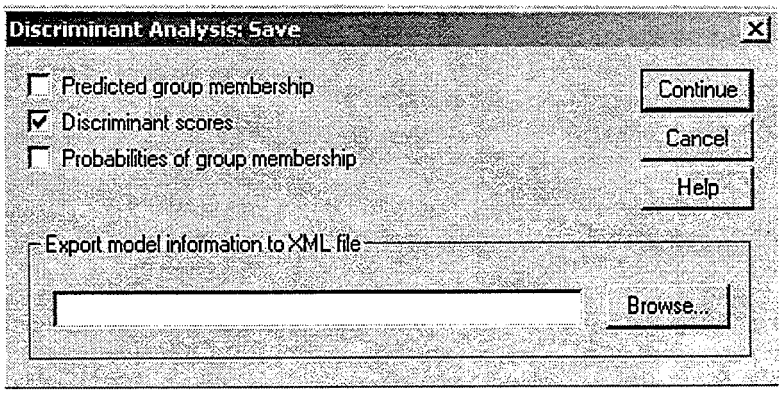


Trong hộp thoại con Classification này, có bốn nhóm tùy chọn:

- Prior probability: xác suất dùng để phân biệt đối tượng thuốc nhóm nào. Có thể định xác suất bằng nhau cho tất cả các nhóm (tổng các xác suất này sẽ bằng 1); hay các xác suất này được tính từ quy mô các nhóm.
- Display:
  - case wise results: cho thể hiện kết quả chi tiết của từng quan sát;
  - Summary table: bảng kết quả phân biệt tóm tắt;
  - leave-one-out classification: phân biệt từng quan sát bằng mô hình phân biệt được xây dựng từ các quan sát khác loại trừ quan sát đang xem xét.
- Use Covariance Matrix:phân biệt các quan sát bằng ma trận hiệp phương sai nội bộ các nhóm trung bình hay bằng ma trận hiệp phương sai các nhóm riêng biệt.
- Plots: vẽ biểu đồ phân tán chung cho các nhóm hay riêng cho từng nhóm, và vẽ biểu đồ vị trí.

Sau đó nhấp chuột tại nút Continue để trở về hộp thoại ban đầu. Bước cuối cùng là nhấp chuột vào nút Save để tạo ra các biến cần thiết trong file dữ liệu để chứa các thông tin cần thiết như: kết quả dự đoán quan sát thuộc nhóm nào; biệt số của từng quan sát; và xác suất phân biệt như hộp thoại dưới. Sau đó nhấp Continue trở về hộp thoại chính rồi nhấp nút OK để thực hiện lệnh phân tích biệt số.

Hình 13.9



9. SO SÁNH PHÂN TÍCH BIỆT SỐ VÀ HỒI QUI BINARY LOGISTIC

Sau khi nghiên cứu các nội dung của phân tích Biệt số, bạn có nhận thấy điểm tương đồng giữa nó với phương pháp hồi qui Binary Logistic? Hai phương pháp này đều áp dụng cho cùng một mục tiêu nhưng mỗi phương pháp có điểm mạnh và yếu riêng. Ưu điểm của hồi qui Binary Logistic là đòi hỏi ít giả định hơn phân tích biệt số, tất nhiên khi những giả định của phân tích biệt số được thoả mãn thì áp dụng hồi qui Logistic vẫn tốt như thường. Logistic còn có lợi thế là mô hình hồi qui nên nó gắn liền với các phép kiểm định rõ ràng dễ hiểu và quen thuộc. Phân tích biệt số và hồi qui Binary Logistic đều cần biến phụ thuộc dạng phân loại. Tuy nhiên hồi qui Binary Logistic chỉ áp dụng được cho tình huống biến phụ thuộc có hai biểu hiện trong khi đó phân tích biệt số phân biệt được nhiều nhóm biểu hiện. Và vì hồi qui Binary Logistic không đòi hỏi những giả định chặt chẽ như phân tích biệt số nên độ tin cậy của nó không cao.

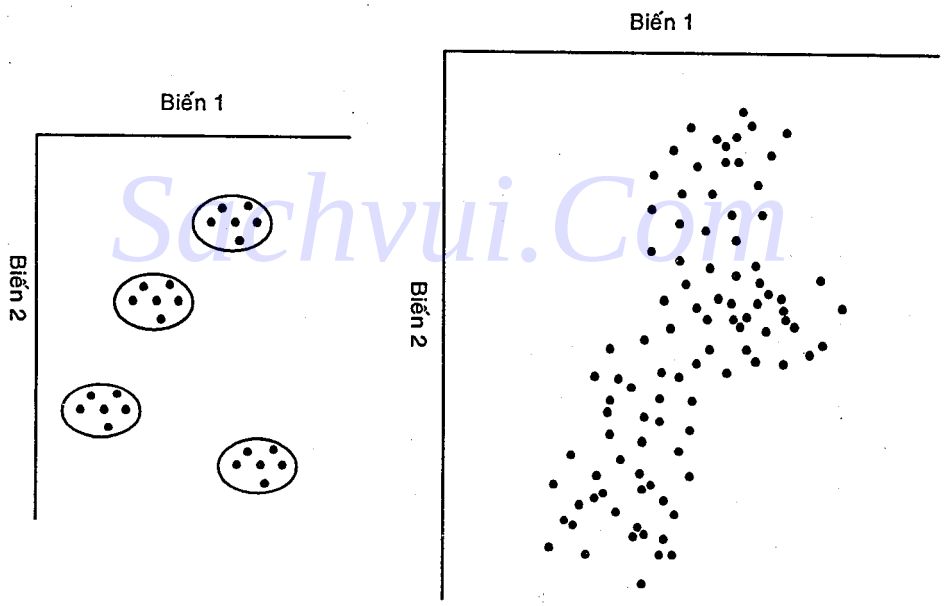
CHƯƠNG XIV

PHÂN TÍCH CỤM

1. KHÁI NIỆM VÀ ỨNG DỤNG

Phân tích cụm là tên của một nhóm các kỹ thuật đa biến có mục tiêu chính là phân loại các đơn vị dựa vào một số các đặc tính của chúng. Các kỹ thuật này nhận diện và phân loại các đối tượng hay các biến sao cho các đối tượng trong cùng một cụm tương tự nhau xét theo các đặc tính được chọn để nghiên cứu. Nội bộ trong các cụm sẽ đồng nhất cao trong khi giữa chúng có sự khác biệt lớn. Vì vậy nếu phân loại thành công thì các đối tượng trong cùng một cụm sẽ nằm gần với nhau và các đối tượng khác cụm sẽ nằm cách xa nhau khi được diễn tả trên đồ thị.

Hình 14.1: Các tình huống phân tích cụm



Phân tích cụm có nhiều tên gọi khác nhau như: phân tích Q, phân tích phân loại, phân loại bằng kỹ thuật định lượng. Có nhiều tên gọi khác nhau như vậy là vì phương pháp phân cụm được ứng dụng trong nhiều lĩnh vực khác nhau như: tâm lý, sinh học, địa lý, xã hội học, kỹ thuật, kinh tế và kinh doanh. Mặc dù có nhiều tên gọi khác nhau nhưng tất cả đều có một đặc điểm chung là phân loại theo các mối

liên hệ tự nhiên. Đặc tính này phản ánh bản chất của tất cả các phép phân cụm.

Cả phân tích cụm và phân tích biệt số đều liên quan đến việc phân loại. Tuy nhiên phân tích biệt số đòi hỏi phải có những hiểu biết trước về các nhóm (có bao nhiêu nhóm) để xây dựng quy tắc phân loại. Ngược lại, trong phân tích cụm, thường không có những thông tin trước về các nhóm hay cụm (sẽ có bao nhiêu nhóm hay cụm). Có bao nhiêu nhóm hay cụm và các nhóm hay cụm này là gì chủ yếu là do dữ liệu thực tế quyết định, không phải hoàn toàn là do ý chí chủ quan.

Trong thực tế, phân tích cụm được ứng dụng trong rất nhiều lĩnh vực khác nhau như: nghiên cứu hành vi, xã hội, tâm lý, kinh doanh. Trong tiếp thị, phân tích cụm được dùng để phân khúc thị trường, tìm hiểu hành vi khách hàng, nhận dạng các cơ hội cho sản phẩm mới, lựa chọn thị trường để thử nghiệm các chiến lược khác nhau, tóm lược dữ liệu lớn để dễ phân tích ...

2. CÁC THUẬT NGỮ VÀ THAM SỐ THỐNG KÊ TRONG PHÂN TÍCH CỤM

- Agglomeration schedule (sơ đồ tích tụ): cung cấp các thông tin về sự kết hợp các đối tượng hay quan sát ở từng giai đoạn tích tụ thành các cụm.
- Cluster centroid (trung bình cụm): là các giá trị trung bình theo các biến của tất cả các quan sát hay các phần tử trong một cụm cụ thể.
- Cluster centers (trung tâm cụm, hạt giống): là điểm khởi đầu để xây dựng cụm. Các cụm được xây dựng dần xung quanh các trung tâm hay hạt giống này.
- Cluster membership (tư cách thành viên): cho biết một đối tượng thuộc cụm nào
- Dendrogram (biểu đồ hình cây): là phương tiện đồ họa để trình bày kết quả phân cụm. Các đường dọc đại diện cho các cụm. Vị trí của các vạch trên thang đo cho biết khoảng cách các cụm được nối vối nhau. Biểu đồ hình cây được xem từ trái qua phải.
- Distances between cluster centers (khoảng cách giữa các hạt giống): các khoảng cách này cho biết khoảng cách giữa từng cặp

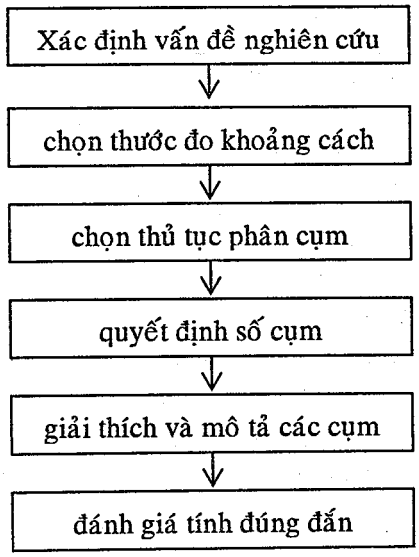
cụm. Các cụm càng rời xa nhau thì càng khác biệt và như vậy càng khó gộp lại với nhau.

- Icicle (biểu đồ cột): là một loại biểu đồ diễn tả kết quả gộp lại thành các cụm. Các cột trong biểu đồ này tương ứng với các đối tượng được phân cụm, và các dòng tương ứng với số cụm. Biểu đồ cột này được xem từ dưới lên.
- Similarity/distance coefficient matrix (ma trận hệ số khoảng cách/ tương đồng): là ma trận chứa khoảng cách giữa từng cặp đối tượng phân cụm hay từng cặp quan sát.

3. TIẾN HÀNH PHÂN TÍCH CỤM

Các bước tiến hành phân tích cụm được trình bày trong Hình 14.2. Bước đầu tiên là xác định các biến số dùng làm cơ sở để phân tích cụm. Sau đó là chọn một thước đo khoảng cách phù hợp. Thước đo khoảng cách cho biết mức độ giống nhau hay khác nhau của các đối tượng được phân cụm. Có nhiều thủ tục phân cụm khác nhau đã được xây dựng và người nghiên cứu phải chọn một thủ tục phù hợp. Số lượng cụm cần thiết là do phán đoán của người nghiên cứu. Các cụm được tạo thành phải được giải thích trên cở sở các biến được sử dụng để phân cụm và được mô tả bằng một số biến quan trọng khác. Cuối cùng là người nghiên cứu phải đánh giá hiệu lực của quy trình phân cụm này.

Hình 14.2: Sơ đồ quy trình tiến hành phân tích cụm



3.1. Xác định vấn đề

Phần quan trọng nhất khi xác định vấn đề phân cụm là việc chọn lựa các biến để phân cụm. Nếu chỉ đưa vào một hay hai biến không có liên quan hay không thích hợp thì cũng sẽ làm nhiễu hay hỏng cả kết quả phân cụm hữu ích. Về cơ bản, nên chọn tập hợp biến có khả năng mô tả được sự giống nhau giữa các đối tượng theo mục đích nghiên cứu. Các biến này có thể được chọn trên cơ sở phân tích lý thuyết, kết quả nghiên cứu trong quá khứ, hay xem xét các giả thuyết có liên quan đã được kiểm định. Trong nghiên cứu thử nghiệm, người nghiên cứu có thể dùng cả phán đoán và trực giác để xác định các biến này.

Để minh họa, chúng ta xem xét ví dụ về phân nhóm người tiêu dùng trên cơ sở thái độ của họ đối với việc đi mua sắm. Dựa vào các nghiên cứu trong quá khứ, có sáu biến thái độ được chọn. Người tiêu dùng được yêu cầu diễn tả mức độ đồng ý đối với các phát biểu sau trên thang đo 7 điểm:

	không đồng ý						đồng ý
mua sắm là một thú vui	1	2	3	4	5	6	7
mua sắm là tốn tiền	1	2	3	4	5	6	7
tôi đi mua sắm kết hợp với ăn uống	1	2	3	4	5	6	7
tôi tìm mua những gì đáng mua nhất khi đi mua sắm	1	2	3	4	5	6	7
tôi không quan tâm đến việc đi mua sắm	1	2	3	4	5	6	7
đi mua sắm giúp tiết kiệm được nhiều nhờ so sánh giá cả	1	2	3	4	5	6	7

Từ các câu hỏi này, ta mã hóa thành 6 biến sau:

V1 mua sắm là một thú vui
V2 mua sắm là tốn tiền
V3 mua sắm kết hợp với ăn uống
V4 cố gắng tìm mua những gì đáng mua nhất khi đi mua sắm
V5 không quan tâm đến việc đi mua sắm
V6 đi mua sắm có thể giúp tiết kiệm được tiền nhờ so sánh nhiều giá cả khác nhau

Trong ví dụ này ta sẽ xem xét một tập dữ liệu đơn giản thu thập từ trả lời của 20 người tiêu dùng về thái độ đối với việc đi mua sắm (xem file Phan tich cum trong tập hợp dữ liệu dùng kèm sách). Trong thực tế, phân tích cụm thường được tiến hành với mẫu có từ 100 quan sát trở lên.

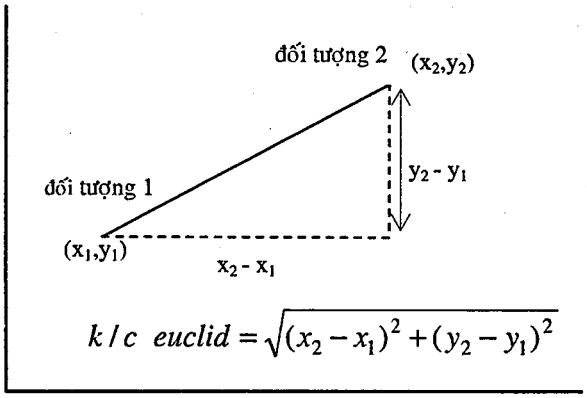
3.2. Chọn lựa thước đo khoảng cách hay thước đo mức độ giống nhau

Vì mục tiêu của phân cụm là nhóm các đối tượng giống nhau lại, cho nên cần phải có một thước đo nào đó để đánh giá mức độ giống nhau hay khác nhau của các đối tượng. Phương pháp thông thường nhất là đo lường mức độ giống nhau bằng khoảng cách giữa hai đối tượng trong một cặp đối tượng. Các đối tượng có khoảng cách giữa chúng nhỏ thì giống nhau hơn là các đối tượng có khoảng cách giữa chúng lớn. Có nhiều cách để tính toán khoảng cách giữa hai đối tượng.

- Thước đo mức độ giống nhau được sử dụng phổ biến nhất là *khoảng cách Euclid* hay khoảng cách Euclid bình phương. Khoảng cách Euclid là căn bậc hai của tổng các độ lệch bình phương của các giá trị trên từng biến của hai đối tượng. Về cơ bản, nó là chiều dài của đường thẳng nối hai đối tượng (Hình 14.3).
- Khoảng cách Manhattan giữa hai đối tượng là tổng các độ lệch tuyệt đối của các giá trị trên từng biến.
- Khoảng cách Chebychev giữa hai đối tượng là chênh lệch tuyệt đối lớn nhất của các giá trị trên từng biến.

Trong ví dụ này ta sử dụng khoảng cách Euclid.

Hình 14.3: Một ví dụ về khoảng cách Euclid giữa hai đối tượng được đo theo hai biến X và Y



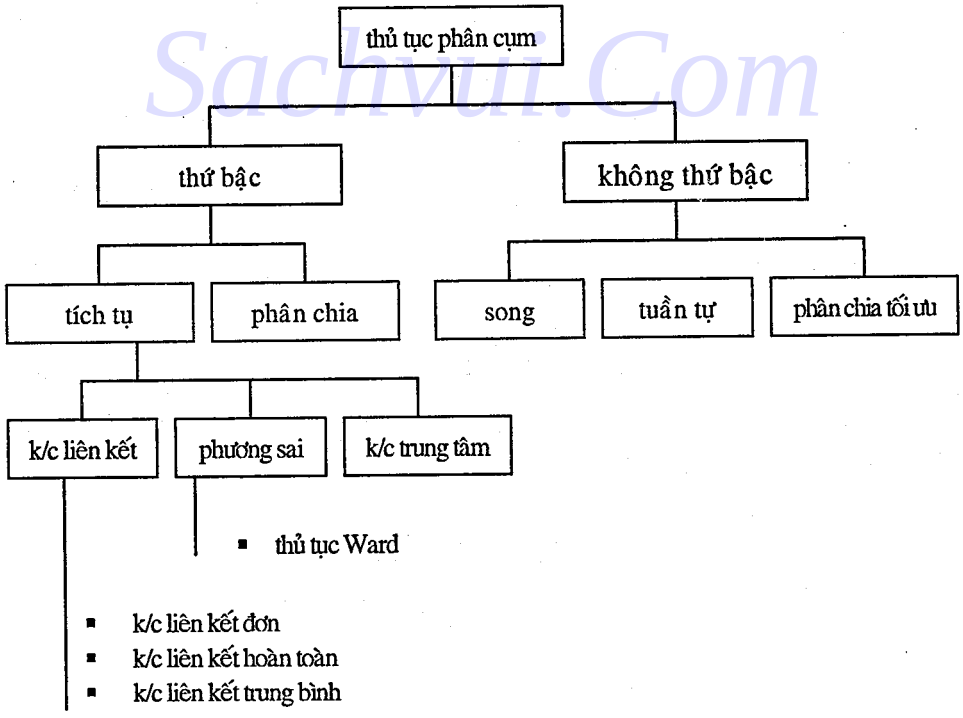
Nếu các biến được đo lường bằng các đơn vị rất khác nhau thì kết quả phân cụm sẽ bị ảnh hưởng bởi các đơn vị đo lường này. Mặc dù việc chuẩn hóa có khả năng loại bỏ ảnh hưởng của sự khác biệt của các đơn vị đo lường, nhưng điều này cũng vẫn còn làm giảm sự khác biệt giữa các nhóm theo những biến có thể giúp phân biệt được tốt nhất các nhóm hay cụm.

Sử dụng các thước đo khoảng cách khác nhau có thể dẫn đến các kết quả phân cụm khác nhau. Do đó bạn nên sử dụng các thước đo khác nhau và so sánh các kết quả nhận được.

3.3. Chọn thủ tục phân cụm

Các thủ tục phân cụm được chia thành hai loại là thủ tục theo thứ bậc và thủ tục không thứ bậc như trong Hình 14.4.

Hình 14.4: Phân loại các thủ tục phân cụm



3.3.1 Phân cụm thứ bậc (hierarchical clustering)

Là thủ tục được xây dựng theo một cấu trúc thứ bậc hay dạng hình cây. Phương pháp này có thể tiến hành theo cách tích tụ lại (agglomerative) hay phân chia ra (divisive).

Phân cụm tích tụ bắt đầu bằng cách mỗi đối tượng là một cụm riêng. Các cụm này được tích tụ cho đến khi tất cả các đối tượng nằm trong một cụm duy nhất. Ngược lại, phân cụm phân chia bắt đầu bằng cách tất cả các đối tượng đều nằm trong một cụm duy nhất. Cụm này được phân ra thành các cụm nhỏ cho đến khi mỗi đối tượng thành một cụm riêng. Chúng ta sẽ nghiên cứu phân cụm tích tụ.

Phân cụm tích tụ thường được sử dụng trong nghiên cứu với các phương pháp: khoảng cách liên kết (linkage method), tổng các độ lệch bình phương hay phương sai (error sums of squares or variance method), và khoảng cách trung tâm (centroid method).

- Các phương pháp phân cụm tích tụ dựa vào khoảng cách liên kết gồm: liên kết đơn (single linkage), liên kết hoàn toàn (complete linkage), liên kết trung bình (average linkage)
  - Phương pháp khoảng cách liên kết đơn dựa vào khoảng cách tối thiểu hay khoảng cách gần nhất. Hai đối tượng được tích tụ (nhập lại) đầu tiên là hai đối tượng có khoảng cách giữa chúng là nhỏ nhất. Tiếp theo là việc nhập lại hai đối tượng có khoảng cách nhỏ thứ nhì, có thể là giữa một đối tượng thứ ba với hai đối tượng đầu tiên trong cụm vừa rồi hay là giữa hai đối tượng mới khác. Ở mỗi giai đoạn, khoảng cách giữa hai cụm là khoảng cách giữa hai đối tượng gần nhau nhất giữa hai cụm (Hình 14.5). Tại một giai đoạn trong quá trình này thì hai cụm được nhập lại là do khoảng cách đơn nhỏ giữa chúng là khoảng cách nhỏ nhất giữa các cặp cụm. Quá trình này tiếp tục cho đến khi tất cả các đối tượng nhập vào một cụm duy nhất. Phương pháp khoảng cách liên kết đơn không đưa ra kết quả tốt nếu các cụm không được định nghĩa đúng đắn.
  - Phương pháp khoảng cách liên kết hoàn toàn: tương tự như phương pháp khoảng cách liên kết đơn, nhưng quá trình tích tụ hay nhập cụm xét trên khoảng cách xa nhất giữa hai cụm.

Khoảng cách xa nhất giữa hai cụm là khoảng cách giữa hai phần tử xa nhất của hai cụm.

- Phương pháp khoảng cách liên kết trung bình: khoảng cách giữa hai cụm là khoảng cách trung bình của tất cả các cặp phần tử giữa hai cụm. Ta có thể thấy rằng phương pháp liên kết trung bình sử dụng thông tin của tất cả các khoảng cách cặp, không chỉ dùng khoảng cách nhỏ nhất hay khoảng cách lớn nhất, nên phương pháp này thường được dùng so với hai phương pháp trên.

Các phương pháp phân cụm tích tụ dựa vào phương sai: cố gắng tối thiểu phương sai trong nội bộ cụm 1. Phương pháp dựa vào phương sai thường dùng nhất là "thủ tục Ward". Theo thủ tục Ward thì ta sẽ tính giá trị trung bình tất cả các biến cho từng cụm một. Sau đó tính khoảng cách Euclid bình phương giữa các phần tử trong cụm với trị trung bình của cụm, rồi lấy tổng tất cả các khoảng cách bình phương này (Hình 14.6). Ở mỗi giai đoạn tích tụ thì hai cụm có phần tăng tổng các khoảng cách bình phương trong nội bộ cụm nếu kết hợp với nhau là nhỏ nhất sẽ được kết hợp.

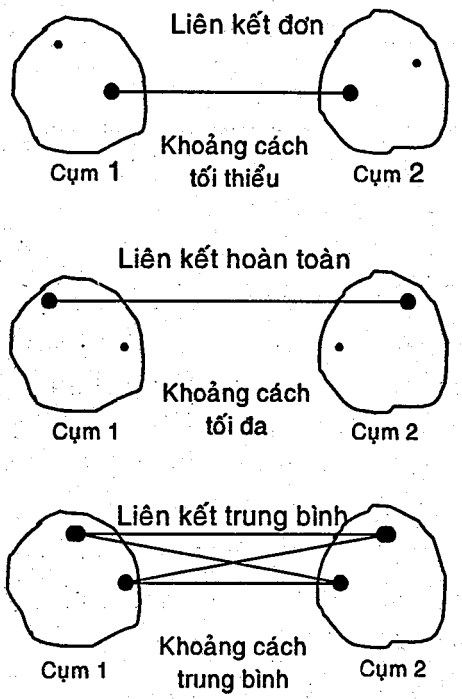
Phương pháp phân cụm tích tụ dựa vào khoảng cách trung tâm: khoảng cách giữa hai cụm được định nghĩa là khoảng cách giữa hai trung tâm của cụm (trung tâm của cụm là trung bình của tất cả các biến) (Hình 14.6). Cứ mỗi lần các đối tượng được nhóm lại thì ta phải tính lại các trung tâm cụm (vì đã có thêm phân tử mới xuất hiện trong cụm).

Trong số các phương pháp phân cụm tích tụ thì phương pháp khoảng cách trung tâm và thủ tục Ward đã được chứng minh là có kết quả tốt hơn các phương pháp kia [2]

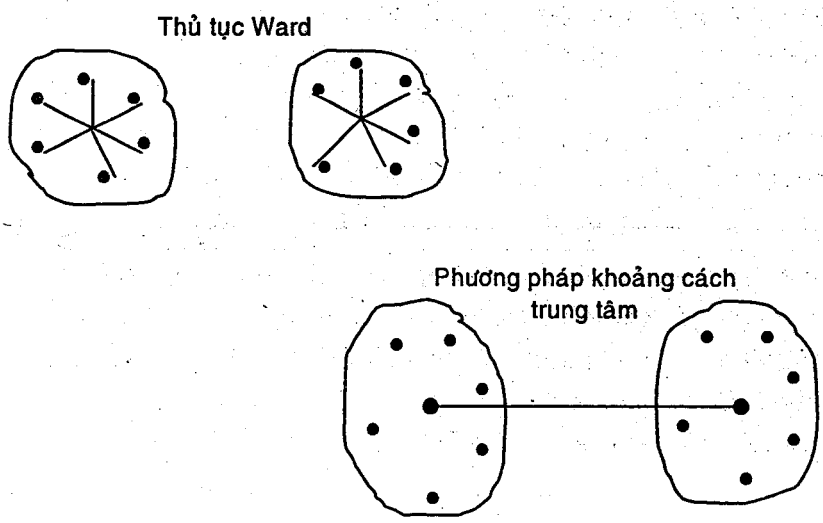
[1] Phương sai trong nội bộ cụm tương tự như phương sai trong nội bộ nhóm. Để rõ hơn về phương sai trong nội bộ nhóm, bạn đọc có thể xem lại phần phân tích phương sai của Tập 1 sách này, hay trong các sách thống kê khác.

[2] G. Milligan, "An Examinaton of the Effect of Six Types of Error Pertubation on Fifteen Clustering Algortthms", *Psychometrika* 45 (September 1980)

Hình 14.5: Các phương pháp phân cụm tích tụ dựa vào các khoảng cách liên kết



Hình 14.6: Phân cụm tích tụ dựa vào phương sai và dựa vào khoảng cách trung tâm



3.3.2 Phân cụm không thứ bậc (Non-hierarchical clustering)

Thường được gọi là phân cụm k-means, gồm có: phương pháp bắt đầu tuần tự (sequential threshold), bắt đầu song song (parallel threshold), phân chia tối ưu (optimizing partitioning).

- phương pháp bắt đầu tuần tự: quá trình bắt đầu từ một hạt giống cụm (trung tâm cụm) được chọn và tất cả các đối tượng cách hạt giống này trong một khoảng cách đã được định trước sẽ nhập vào cụm này. Sau đó chọn tiếp một hạt giống mới và quá trình được tiếp tục với các phần tử còn lại. Một khi một phần tử đã được xếp vào trong một cụm rồi, thì trong những lần tiếp theo ta sẽ không xét đến nó nữa.
- phương pháp bắt đầu song song: tương tự như phương pháp bắt đầu tuần tự nhưng nhiều hạt giống được chọn và quá trình được tiến hành song song. Một phần tử có khoảng cách so với các cụm nhỏ hơn một khoảng cách đã định trước sẽ được nhập vào cụm mà khoảng cách giữa nó và hạt giống của cụm là nhỏ nhất.
- phương pháp phân chia tối ưu: thủ tục này khác với hai phương pháp trên ở chỗ các đối tượng sau khi được phân vào một cụm nào đó có thể sẽ được phân lại vào các cụm khác để thỏa một tiêu chuẩn tối ưu toàn bộ, ví dụ như khoảng cách trong nội bộ cụm.

Hai nhược điểm chính của phân cụm không thứ bậc là ta phải thử xác định trước số cụm, và việc lựa chọn hạt giống của cụm khá tùy ý. Kết quả phân cụm có thể phụ thuộc vào cách chọn trung tâm cụm. Nhiều chương trình phân cụm không thứ bậc chọn k quan sát đầu tiên (k là số cụm được xác định trước) không có giá trị quan sát bị thiếu làm các hạt giống của các cụm. Vì vậy các kết quả phân cụm có thể phụ thuộc vào thứ tự các quan sát trong tập tin dữ liệu.

Ưu điểm của phân cụm không thứ bậc là khối lượng tính toán ít hơn, thời gian thực hiện nhanh hơn, và điều này quan trọng khi số đối tượng hay số quan sát nhiều trong khi khả năng của máy tính hạn chế. Cả hai phương pháp này thường được sử dụng. Đầu tiên nên sử dụng phân cụm thứ bậc để tìm ra kết quả ban đầu, sau đó số cụm và các trung tâm cụm của kết quả này được sử dụng làm thông tin ban đầu để áp dụng phương pháp phân chia tối ưu.

Việc lựa chọn loại thủ tục phân cụm và lựa chọn thước đo khoảng cách có liên hệ với nhau. Khoảng cách Euclid bình phương thường

được sử dụng với phương pháp khoảng cách trung tâm và phương pháp Ward. Các thủ tục phân cụm không thứ bậc cũng sử dụng khoảng Euclid bình phương.

Sau đây chúng ta sẽ sử dụng thủ tục Ward để minh họa cho phân cụm thứ bậc. Phần phân cụm không thứ bậc bạn sẽ gặp ở phần sau. Kết quả phân cụm thứ bậc đối với các dữ liệu trong file *Phan tich cum* được trình bày trong bảng sau.

Bảng 14.1a **Agglomeration Schedule**

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14	16	1.000	0	0	6
2	6	7	2.000	0	0	7
3	2	13	3.500	0	0	15
4	5	11	5.000	0	0	11
5	3	8	6.500	0	0	16
6	10	14	8.167	0	1	9
7	6	12	10.500	2	0	10
8	9	20	13.000	0	0	11
9	4	10	15.583	0	6	12
10	1	6	18.500	0	7	13
11	5	9	23.000	4	8	15
12	4	19	27.750	9	0	17
13	1	17	33.100	10	0	14
14	1	15	41.333	13	0	16
15	2	5	51.833	3	11	18
16	1	3	64.500	14	5	19
17	4	18	79.667	12	0	18
18	2	4	172.667	15	17	19
19	1	2	328.600	16	18	0

Bảng 14.1a là kết quả phân cụm dưới dạng sơ đồ tích tụ cho biết số quan sát hay cụm được kết hợp ở mỗi giai đoạn. Dòng đầu tiên thể hiện giai đoạn 1 có 19 cụm vì người thứ 14 và người thứ 16 được kết hợp trong giai đoạn này (bạn xem trong hai cột thuộc phần các cụm được kết hợp - Cluster combined). Khoảng cách Euclid bình phương giữa hai người này được thể hiện trong cột "hệ số"- coefficient. Cột "Stage Cluster First Appears" cho biết cụm này được tạo thành trong giai đoạn nào, ví dụ như số 1 ở giai đoạn 6 cho biết người thứ 14 được nhóm lại thành cụm đầu tiên trong giai đoạn 1 (hay là cụm

đang chứa người thứ 14 được tạo ra trong giai đoạn 1 gồm có người thứ 14 và người thứ 16) và bây giờ trong giai đoạn 6 cụm này nhập thêm người thứ 10. Cột cuối cùng "Next stage" cho biết ở giai đoạn nào thì có thêm người hay cụm mới được nhập vào với cụm trong dòng này. Ví dụ như trong dòng đầu tiên ở cột cuối cùng ta thấy số 6, có nghĩa là ở giai đoạn 6 thì có thêm người thứ 10 được kết hợp vào với cụm đã có hai người 14 và 16. Bạn có nhận thấy một chuỗi thông tin liên thông với nhau không? Tương tự như vậy, dòng thứ 2 thể hiện giai đoạn 2 có 18 cụm vì người thứ 6 và người thứ 7 được nhập lại với nhau.

Bảng 14.1b **Cluster Membership**

Case	Label	4 Clusters	3 Clusters	2 Clusters
1		1	1	1
2		2	2	2
3		1	1	1
4		3	3	2
5		2	2	2
6		1	1	1
7		1	1	1
8		1	1	1
9		2	2	2
10		3	3	2
11		2	2	2
12		1	1	1
13		2	2	2
14		3	3	2
15		1	1	1
16		3	3	2
17		1	1	1
18		4	3	2
19		3	3	2
20		2	2	2

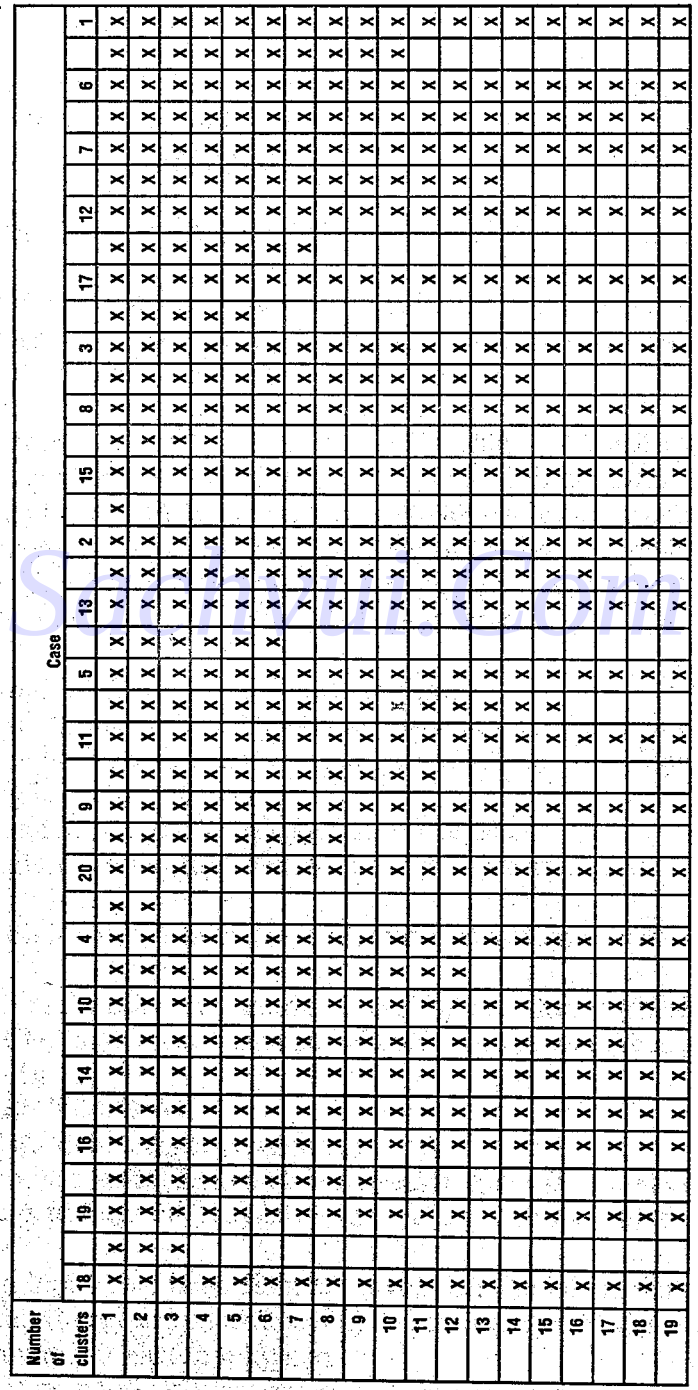
Hình 14.7 là kết quả phân cụm dưới dạng bảng sơ đồ cột, chú ý là bảng sơ đồ phân cụm này phải được đọc từ dưới lên trên, cột có dấu X đại diện cho cụm còn cột khoảng trắng đại diện cho sự tách biệt giữa các cụm. Các con số trên đầu bảng (case) cho biết đối tượng nào được nhóm với đối tượng nào. Trong ví dụ này đối tượng là người tiêu dùng được phỏng vấn được đánh số từ 1 đến 20. Các dòng cho biết số cụm được gom theo thứ tự đi từ dưới lên.

Đầu tiên tất cả các quan sát được xem là các cụm cá thể, vì ta có 20 người nên ta có 20 cụm. Ở bước 1 hai người gần nhau nhất là case 14 và case 16 được kết hợp lại nên ta có 19 cụm thể hiện ở dòng cuối cùng của bảng (ứng với giai đoạn 1). Đếm ngang hàng 19 thấy có 19 cột có dấu X tức 19 cụm được gom từ 20 đối tượng, giữa chúng có cột trắng tách biệt. Trong 19 cột đó có 18 cột dấu X riêng lẻ và một nhóm 3 cột dấu X đi liền đại diện 2 đối tượng 16 và 14 được gom lại. Dòng 18 ứng với giai đoạn tiếp theo là giai đoạn 2 còn có 18 cụm. Ở giai đoạn này người thứ 6 và người thứ 7 được nhóm lại với nhau và ta có 18 cụm, trong đó có 16 cụm chỉ có một người (thể hiện ở 16 cột dấu X đơn) và hai cụm có 2 người (thể hiện ở 2 nhóm có 3 cột X đi liền nhau).

Trong các giai đoạn tiếp theo, một cụm mới có thể được thành lập theo 3 cách:

- hai người được nhóm với nhau,
- một người được nhóm với một cụm có sẵn,
- hai cụm được nhập lại với nhau

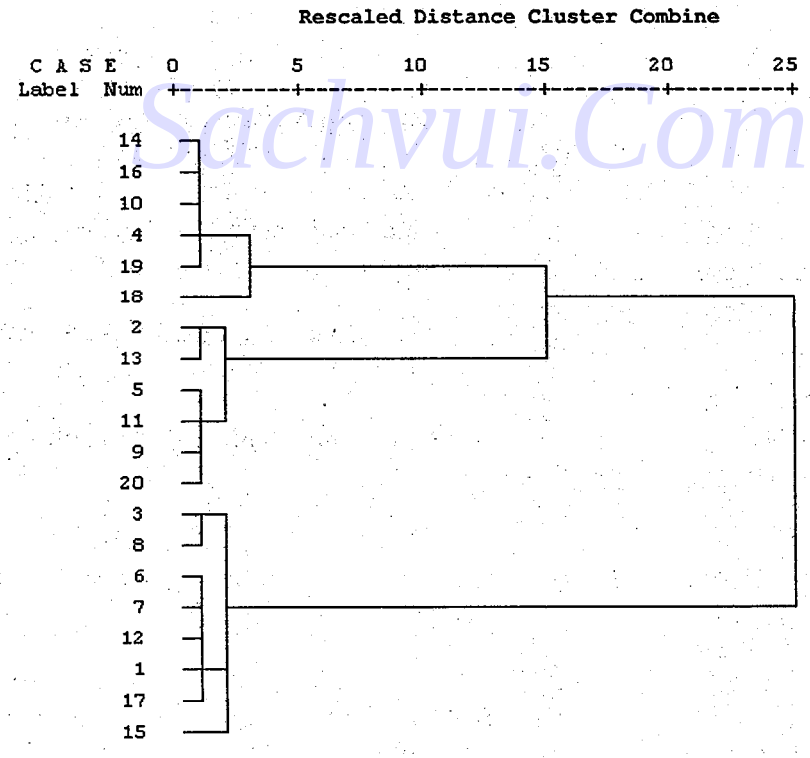
Hình 14.7 Quá trình phân cụm dưới dạng biểu đồ cột sử dụng thủ tục Ward (**Vertical Icicle**)



Hình 14.8 dưới đây là biểu đồ hình cây thể hiện quá trình phân cụm. Biểu đồ hình cây này được đọc từ trái sang phải. Các đường kẻ dọc đại diện các cụm đã được nhập lại với nhau. Vị trí của đường kẻ dọc trên thang đo (rescaled distance cluster combine) cho biết khoảng cách giữa các cụm khi được nhập với nhau. Bởi vì trong những giai đoạn đầu có nhiều khoảng cách có độ lớn bằng nhau nên ta khó quan sát, nhưng trong 2 giai đoạn cuối ta có thể dễ dàng nhận thấy khoảng cách giữa các cụm khi được nhập lại với nhau khá lớn. Thông tin này rất hữu ích khi cần xác định số cụm kết quả.

Khi xác định được số cụm rồi thì ta có thể biết kết quả phân cụm của từng phần tử quan sát. Kết quả này được thể hiện trong bảng sơ đồ cột phân cụm ở Hình 14.7, nhưng ở dưới dạng bảng như trong Bảng 14.1a ta quan sát dễ hơn. Nhìn vào bảng này ta biết được phần tử nào thuộc cụm nào khi kết quả cuối cùng ta chấp nhận có 2, 3 hay 4 cụm.

Hình 14.8: Quá trình phân cụm dưới dạng biểu đồ hình cây sử dụng thủ tục Ward (Dendrogram using Ward Method)



3.4. Quyết định số cụm

Đây là một vấn đề chính trong phân tích cụm. Cho tới nay chưa có những qui y tắc rõ ràng và chắc chắn về việc xác định số cụm. Nói một cách khác là số cụm cần thiết hay hợp lý không phải là một vấn đề hoàn toàn về mặt kỹ thuật, mà nó phụ thuộc vào nhiều yếu tố khác. Sau đây là một số căn cứ:

- phân tích lý thuyết có thể giúp ta xác định được số cụm hợp lý. Ví dụ như khi nhận dạng các phân khúc thị trường các nhà nghiên cứu tiếp thị có thể đã biết trước số phân khúc là 3 hay 4.
- trong phân cụm thứ bậc, ta có thể sử dụng khoảng cách giữa các cụm làm tiêu chuẩn để xác định số cụm. Hai cụm cách nhau khá xa tức là tính chất của chúng tương đối khác nhau nhiều thì không nên nhập lại thành cụm mới. Thông tin về khoảng cách giữa các cụm được thể hiện trong sơ đồ tích tụ hay trong biểu đồ hình cây. Với ví dụ của chúng ta thì từ Bảng 14.1a có thể dễ dàng nhận thấy rằng trong cột "coefficient" khoảng cách giữa các cụm đột ngột tăng lên giữa hai giai đoạn 17 và 18. Tương tự, trong biểu đồ hình cây ở Hình 14.8, ta thấy các cụm được kết hợp với nhau trong hai giai đoạn cuối ở một khoảng cách lớn. Vì vậy dường như phương án 3 cụm là phù hợp.
- trong phân cụm không thứ bậc, khi tỉ số giữa phương sai nội bộ nhóm và phương sai giữa các nhóm có sự thay đổi đột ngột thì ta có thể xác định được số cụm hợp lý. Tăng thêm số cụm thì thường các cụm có tính chất khá giống nhau.
- quy mô tương đối của các cụm cũng có thể được sử dụng để định số cụm. Trong Bảng 14.1, đếm số phần tử trong một cụm ta có thể thấy phương án 3 cụm, mỗi cụm có số phần tử lần lượt là: 8, 6 và 6, là khá hợp lý. Nếu ta chọn phương án 4 cụm thì quy mô của các cụm là: 8, 6, 5, và 1. Phương án này không thích hợp vì có 1 cụm có quy mô quá nhỏ.

3.5. Diễn giải và mô tả các cụm

Để diễn giải và mô tả các cụm ta sẽ xem xét các trung bình cụm (centroid). Các trung bình cụm được tính bình quân từ các giá trị của các đối tượng theo từng biến một. Các trung bình cụm gợi ý cho ta một cái tên cho mỗi cụm. Nếu chương trình máy tính thực hiện phân

tích cụm không đưa ra các thông tin về trung bình cụm, ta có thể dùng phân tích biệt số, hay đơn giản hơn là dùng thủ tục tính trung bình (lệnh Basic Tables trong Chương III ở Tập 1) cho các biến số nghiên cứu theo từng cụm. Bảng 14.2 trình bày các trung bình cụm các biến từ V1 đến V6 của file ví dụ với biến phân nhóm là biến vừa được sao lưu sau quá trình thực hiện thủ tục phân tích cụm cho ví dụ này (để sao lưu được biến này thì trong hộp thoại con Save của hộp thoại phân tích cụm bạn chọn Single solution và khai báo 3 cụm).

Cụm số 1 có trị trung bình lớn đối với biến V1 (đi mua sắm là thú vui), V3 (đi mua sắm kết hợp với ăn uống), và có trị trung bình nhỏ đối với biến V5 (không quan tâm đến việc đi mua sắm). Do đó cụm này có thể được đặt tên là "*nhóm quan tâm và thích thú đi mua sắm*"). Cụm này gồm có các quan sát 1, 3, 6, 7, 8, 12, 15, và 17. Bạn có thể kiểm tra điều này trên biểu đồ hình cây.

Bảng 14.2: Kết quả tính toán trung bình của các biến theo từng cụm

cụm số	đi mua sắm là thú vui	đi mua sắm là tốn tiền	đi mua sắm kết hợp với ăn uống	tìm mua những gì đáng mua nhất	không quan tâm đến việc đi mua sắm	đi mua sắm giúp tiết kiệm được tiền nhờ so sánh giá cả
	V1	V2	V3	V4	V5	V6
1	5.750	3.625	6.000	3.125	1.875	3.875
2	1.667	3.000	1.833	3.500	5.500	3.333
3	3.500	5.833	3.333	6.000	3.500	6.000

Ngược lại, cụm số 2 có trị trung bình thấp đối với biến V1, V3 và có trị trung bình lớn đối với biến V5. Do đó cụm này có thể được đặt tên là "*nhóm thờ ơ với việc đi mua sắm*". Cụm này gồm có các quan sát 2, 5, 9, 11, 13, và 20.

Cụm số 3 có trị trung bình lớn đối với biến V2 (đi mua sắm là tốn tiền), V4 (cố gắng tìm mua những gì đáng mua nhất khi đi mua sắm) và V6 (đi mua sắm có thể giúp tiết kiệm được tiền nhờ so sánh nhiều giá cả khác nhau). Vì vậy cụm này có thể được đặt tên *là "nhóm mua sắm quan tâm đến kinh tế"*. Cụm này gồm có các quan sát 4, 10, 14, 16, 18, và 19.

Thông thường để mô tả các cụm này, ta nên xem xét một số biến khác chưa được sử dụng trong phân tích này như là: các dữ liệu nhân khẩu học (thu nhập, giới tính, tuổi, tình trạng việc làm, tình trạng hôn nhân gia đình, trình độ văn hóa ...), tâm lý, mức độ sử dụng sản phẩm, thói quen sử dụng phương tiện truyền thông, ... để thấy rõ được đặc trưng của từng cụm.

3.6. Đánh giá

Có nhiều cách thẩm định và đánh giá độ tin cậy và tính hợp lý của kết quả phân tích cụm:

- thực hiện phân tích cụm trên cùng một tập hợp dữ liệu nhưng sử dụng các thước đo khoảng cách khác nhau. So sánh các kết quả để xem tính ổn định của các giải pháp
- sử dụng các phương pháp phân cụm khác nhau (thứ bậc và không thứ bậc) và so sánh kết quả
- chia dữ liệu làm hai phần, rồi thực hiện phân tích cụm riêng cho mỗi tập dữ liệu con, sau đó so sánh các trung bình cụm giữa hai tập dữ liệu con này
- bỏ bớt một vài biến rồi thực hiện phân tích cụm trên tập hợp các biến còn lại, sau đó so sánh so sánh kết quả này với kết quả khi sử dụng hết các biến cần thiết
- trong phân tích cụm không thứ bậc, các kết quả có thể phụ thuộc vào thứ tự của các quan sát trong tập hợp dữ liệu, vì vậy nên thực hiện phân tích cụm nhiều lần với nhiều thứ tự khác nhau của các quan sát cho đến khi kết quả ổn định

3.7 Phân tích cụm không thứ bậc

Để minh hoạ ta cũng sẽ thực hiện thủ tục phân cụm không thứ bậc với phương pháp phân chia tối ưu (optimizing partitioning) đối với các dữ liệu trong file ví dụ này. Số cụm ta cần thực hiện là 3 để so sánh với kết quả phân cụm thứ bậc. Các kết quả được trình bày trong Bảng 14.3. Các hạt giống cụm là các giá trị của 3 quan sát thứ 8, 19 và 20 ở file dữ liệu. Các hạt giống cụm này chỉ là những trung tâm cụm tạm thời dùng để phân bổ các quan sát vào các cụm. Mỗi quan sát được phân vào cụm có trung tâm gần nó nhất. Các trung tâm cụm sẽ được cập nhật trong quá trình này cho đến khi tiêu chuẩn phân

chia tối ưu được thỏa mãn. Các trung tâm cụm cuối cùng là các trung bình của các quan sát đối với các biến trong phương án phân cụm đạt được.

Bảng 14.3c (Cluster Membership) cho thấy từng quan sát thuộc về cụm nào và khoảng cách giữa từng quan sát với trung tâm của nó. Ta có thể thấy rằng kết quả phân cụm trong Bảng 14.1b (thứ bậc) và kết quả trong Bảng 14.3 (không thứ bậc) là như nhau: số đối tượng trong các cụm bằng nhau, mỗi cụm đều bao gồm các đối tượng trùng nhau (nhưng chú ý rằng cụm 1 trong Bảng 14.1b được gọi là cụm 3 trong Bảng 14.3, ngược lại cụm 3 trong Bảng 14.1b chính là cụm 1 trong Bảng 14.3). Khoảng cách giữa các cụm trong phương án phân cụm cuối cùng cho thấy các cặp cụm được phân tách rất rõ.

Kiểm định F đối với từng biến của các cụm cũng được trình bày trong bảng kết quả này (F là tỉ số giữa phương sai giữa các cụm - cluster mean squares và phương sai trong nội bộ các cụm - error mean squares). Giả thiết H_0 ở đây là sự khác biệt giữa các cụm chỉ là ngẫu nhiên (về cơ bản thì các cụm giống nhau). F càng lớn có nghĩa là phương sai giữa các cụm lớn so với phương sai trong nội bộ các cụm, tức là các phần tử trong nội bộ cụm khá gần nhau (giống nhau) trong khi các phần tử giữa hai cụm khác nhau thì khá xa nhau (khá khác nhau). Như vậy ta càng có nhiều cơ sở để bác bỏ giả thiết này và kết luận rằng sự khác biệt giữa các cụm là có cơ sở. Thông thường ta sử dụng mức ý nghĩa quan sát suy ngược từ giá trị thống kê F. F càng lớn thì giá trị sig. càng nhỏ, kết quả phân cụm càng có ý nghĩa (các cụm càng khác nhau). Trong thực tế thường sử dụng mốc là 0,05. Nếu sig. lớn ($\geq$ 0,05) thì giả thiết về sự giống nhau giữa các cụm được chấp nhận, ta kết luận là giữa các cụm không có sự khác biệt có ý nghĩa và ta không nên sử dụng kết quả phân cụm này. Giá trị Sig. ở Bảng 14.3f cho phép ta kết luận điều ngược lại.

Kết quả phân cụm không thứ bậc được thể hiện trong các Bảng 14.3.

Bảng 14.3a **Initial Cluster Centers**

	Cluster		
	1	2	3
di mua sam la thu vui	4	2	7
di mua sam la ton tien	6	3	2
ket hop mua sam voi an uong	3	2	6
tim mua nhung gi dang mua nhat khi di mua sam	7	4	4
khong quan tam viec di mua sam	2	7	1
co the tiet kiem nhieu khi so sanh gia ca	7	2	3

Bảng 14.3b **Iteration History(a)**

Iteration	Change in Cluster Centers		
	1	2	3
1	2.154	2.102	2.550
2	.000	.000	.000

a Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 7.746.

Bảng 14.3c **Cluster Membership**

Case Number	Cluster	Distance
1	3	1.414
2	2	1.323
3	3	2.550
4	1	1.404
5	2	1.848
6	3	1.225
7	3	1.500
8	3	2.121
9	2	1.756
10	1	1.143
11	2	1.041
12	3	1.581
13	2	2.598
14	1	1.404
15	3	2.828
16	1	1.624
17	3	2.598
18	1	3.555
19	1	2.154
20	2	2.102

Bảng 14.3d **Final Cluster Centers**

	Cluster		
	1	2	3
di mua sam la thu vui	4	2	6
di mua sam la ton tien	6	3	4
ket hop mua sam voi an uong	3	2	6
tim mua nhung gi dang mua nhat khi di mua sam	6	4	3
khong quan tam viec di mua sam	4	6	2
co the tiet kiem nhieu khi so sanh gia ca	6	3	4

Bảng 14.3e **Distances between Final Cluster Centers**

Cluster	1	2	3
1		5.568	5.698
2	5.568		6.928
3	5.698	6.928	

Bảng 14.3f **ANOVA**

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
di mua sam la thu vui	29.108	2	.608	17	47.888	.000
di mua sam la ton tien	13.546	2	.630	17	21.505	.000
ket hop mua sam voi an uong	31.392	2	.833	17	37.670	.000
tim mua nhung gi dang mua nhat khi di mua sam	15.713	2	.728	17	21.585	.000
khong quan tam viec di mua sam	22.538	2	.816	17	27.614	.000
co the tiet kiem nhieu khi so sanh gia ca	12.171	2	1.071	17	11.363	.001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Bảng 14.3g **Number of Cases in each Cluster**

Cluster	1	6.000
	2	6.000
	3	8.000
Valid		20.000
Missing		.000

4. PHÂN TÍCH CỤM ĐỐI VỚI CÁC BIẾN

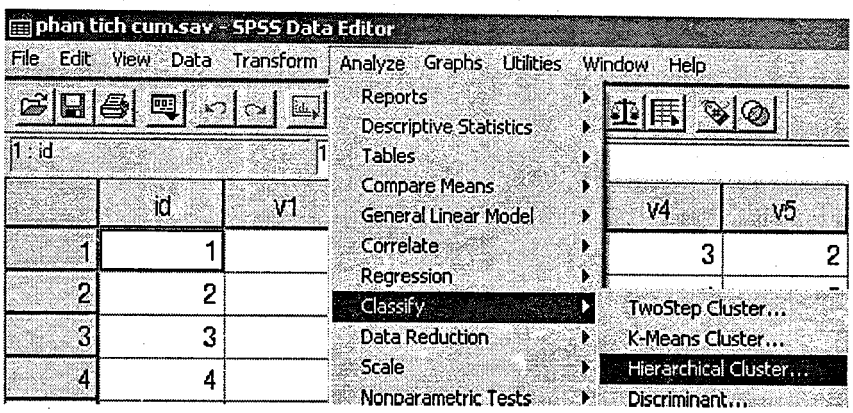
Phân tích cụm đôi khi cũng được dùng để nhóm các biến giống nhau lại. Trong trường hợp này đối tượng hay đơn vị không phải là các quan sát (trong ví dụ ở phần trước, đối tượng phân tích là người) mà là các biến và ta sẽ xem xét khoảng cách giữa các cặp biến. Chẳng hạn như ta có thể dùng hệ số tương quan, chỉ xét độ lớn tuyệt đối hay xét luôn cả dấu, làm thước đo mức độ giống nhau giữa các biến.

Phân tích cụm thứ bậc đối với các biến có thể giúp ta phát hiện hay nhận diện các biến đáng chú ý hay các biến có ảnh hưởng mạnh đến dữ liệu nghiên cứu. Phân cụm còn có thể được sử dụng để giảm số lượng biến nghiên cứu. Một cụm thành phần (cluster component) là một kết hợp tuyến tính của một số biến nhất định. Một số lượng biến lớn có thể được thay thế bằng một số ít các cụm thành phần mà không làm thất thoát nhiều thông tin thu thập được. Tuy nhiên phân tích cụm đối với các biến không đạt hiệu quả cao như trong phân tích nhân tố (phân tích thành phần chính), tức là làm thất thoát thông tin nhiều hơn khi giảm số lượng biến nghiên cứu. Nhưng trong thực tế người ta cũng sử dụng nó vì cụm thành phần dễ giải thích hơn là các nhân tố (thành phần chính) mặc dù các nhân tố này đã được xoay (xem phần xoay nhân tố trong chương Phân tích nhân tố)

5. THỰC HIỆN PHÂN TÍCH CỤM BẰNG SPSS

Từ menu ta chọn Statistics > Classify. Để chọn phân cụm thứ bậc bạn nhấp vào Hierarchical Cluster, còn để chọn phân cụm không thứ bậc bạn chọn lệnh kế trên là K-Means Cluster như trong hình sau:

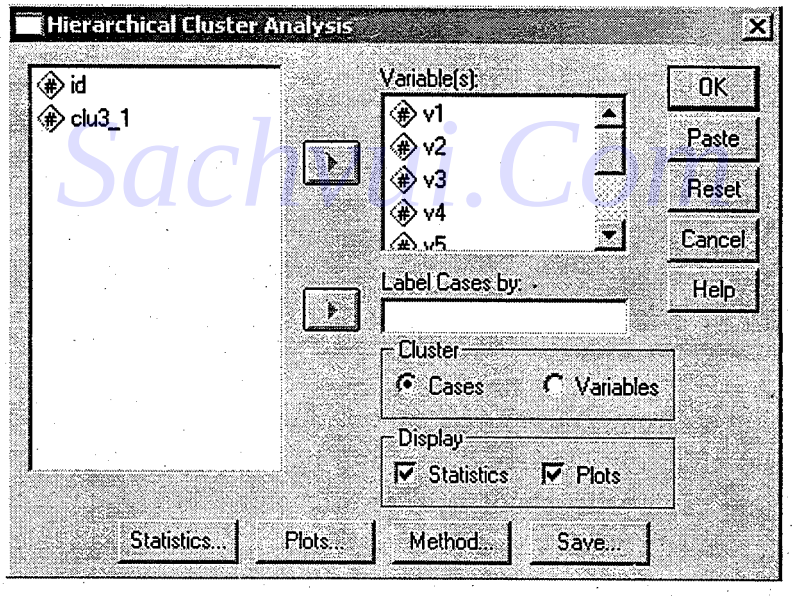
Hình 14.9



1. Phân cụm thứ bậc

Lệnh Hierarchical Cluster sẽ mở hộp thoại phân tích cụm thứ bậc ở Hình 14.10.

Hình 14.10



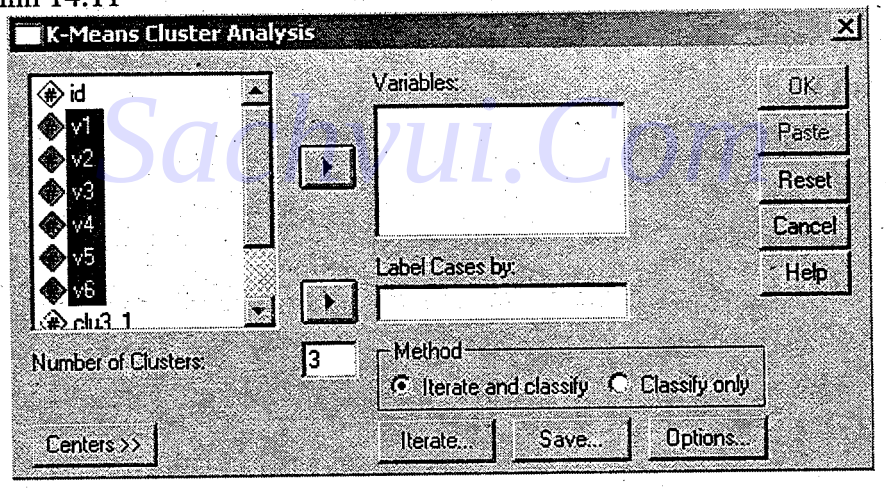
Các biến số trong file dữ liệu sẽ xuất hiện trong ô danh sách biến nguồn ở phía trái của hộp thoại. Bạn hãy chọn các biến cần phân tích bằng cách nhấn chuột vào tên các biến trong danh sách biến nguồn, rồi nhấn chuột vào mũi tên qua phải chỉ vào ô Variable. Lần lượt xác định các yếu tố cần thiết sau để thực hiện phân tích:

- Cluster: phân cụm đối với các quan sát hay đối với các biến số
- Display: cho thể hiện chi tiết các thống kê và vẽ các biểu đồ
- Statistics: tóm tắt kết quả phân tích bằng sơ đồ tích tụ, tính ma trận khoảng cách, số cụm cần tìm
- Plots: vẽ biểu đồ cây, biểu đồ cột
- Method: chỉ định phương pháp phân cụm: liên kết, Ward ...
- Save: tạo các biến mới trong file dữ liệu cho biết từng quan sát cụ thể thuộc cụm kết quả nào.

2. Phân cụm không thứ bậc

Lệnh K-Means Cluster sẽ mở hộp thoại phân tích cụm thứ bậc như Hình 14.11.

Hình 14.11



Các biến số trong file dữ liệu sẽ xuất hiện trong ô danh sách biến nguồn ở phía trái của hộp thoại. Bạn hãy chọn các biến cần phân tích bằng cách nhấp chuột vào tên các biến trong danh sách biến nguồn, rồi nhấn chuột vào mũi tên qua phải chỉ vào ô Variable.

- Number of Clusters: bạn hãy chỉ định số cụm cần thực hiện, số cụm mặc định là 2, ở đây ta chọn 3 cụm.
- Method: khung này thể hiện 2 phương pháp sau:
- Iterate and classify: các hạt giống cụm được cập nhật và thay đổi trong suốt quá trình phân cụm

- Classify only: các hạt giống cụm ban đầu không được cập nhật và được dùng để phân cụm.
- Iterate: xác định số lần cập nhật và tiêu chuẩn dừng (tiêu chuẩn hội tụ) quá trình cập nhật hạt giống cụm
- Save: tạo các biến mới chứa các thông tin cần thiết trên file dữ liệu như: khoảng cách Euclid từ mỗi quan sát đến trung tâm cụm, quan sát thuộc cụm kết quả nào.
- Options: tính các đại lượng thống kê mô tả như: hạt giống cụm, bảng phân tích phương sai, thông tin chi tiết về từng quan sát; cách thức xử lý các quan sát bị thiếu dữ liệu.

Sau khi xác định xong các yếu tố cần thiết, để thực hiện phân tích bạn hãy nhấn chuột vào nút OK trên hộp thoại phân tích cụm thứ bậc.

CHƯƠNG XV

LẬP BẢN ĐỒ NHẬN THỨC VỚI ĐO LƯỜNG ĐA HƯỚNG VÀ PHÂN TÍCH TƯƠNG HỢP

Trong nghiên cứu, nhất là nghiên cứu ứng dụng, người nghiên cứu thường cần tìm hiểu cảm nhận hay nhận thức của các đối tượng mục tiêu về một số đối tượng cần đánh giá. Bản đồ nhận thức có thể áp dụng trong rất nhiều trường hợp như tìm hiểu cảm nhận của học sinh phổ thông về các ngành đào tạo đại học, nhận thức của người đi làm đối với nghề nghiệp khác nhau, cảm nhận của du khách đối với các thành phố biển, cảm nhận của sinh viên đối với các trường đại học có ngành đào tạo tương tự, cảm nhận của những người trẻ 18-25 tuổi, có mức cho tiêu trung bình khá đối với các thương hiệu thời trang thông dụng, nhận thức của các nhà đầu tư cá nhân đối với các ngân hàng thương mại cổ phần chưa niêm yết... Đặc biệt trong lĩnh vực tiếp thị, bản đồ nhận thức là phương pháp thường được sử dụng khi nghiên cứu đo lường định vị các sản phẩm hay thương hiệu. Bản đồ nhận thức là một phương pháp chính quy giúp mô tả trực quan các nhận thức và cảm nhận này.

1. QUY TRÌNH LẬP BẢN ĐỒ NHẬN THỨC

Bước đầu tiên: nhận diện các yếu tố mà đối tượng mục tiêu dựa vào đó cảm nhận về các đối tượng cần đánh giá. Việc khám phá các yếu tố này thường được thực hiện bằng nghiên cứu thăm dò (phỏng vấn sâu, thảo luận nhóm) hay từ kinh nghiệm, sau đó được xác nhận qua nghiên cứu định lượng để nhận diện các yêu tố có liên quan và quan trọng. Ví dụ trong tiếp thị, khách hàng sẽ dựa vào các yếu tố nào để cảm nhận về các thương hiệu? Khách hàng cảm nhận thương hiệu của ta giống với thương hiệu nào nhất? Yếu tố nào tạo ra sự khác biệt chính cho thương hiệu của ta.

Bước tiếp theo: đánh giá vị trí của các đối tượng đánh giá. Ví dụ trong tiếp thị, đánh giá vị trí các thương hiệu để xem chúng ta đã thực hiện tốt đến đâu chiến lược định vị. Nhận ra các yếu tố quan trọng để tạo ra sự khác biệt, phân khúc thị trường nào là hấp dẫn,

nên định vị một thương hiệu mới như thế nào so với các thương hiệu hiện có.

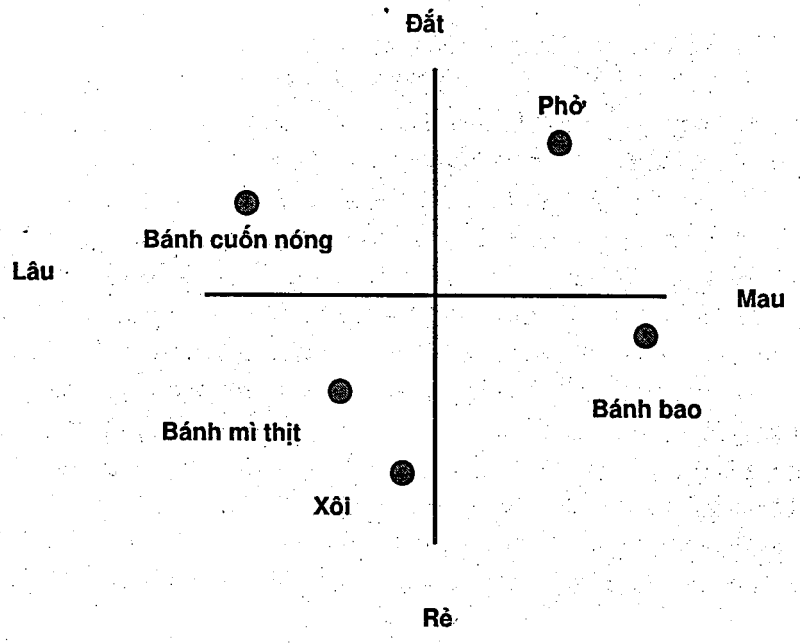
2. CẤU TRÚC VÀ ĐỌC HIỂU BẢN ĐỒ NHẬN THỨC

Bản đồ nhận thức là một cách trình bày các đối tượng trên một không gian Euclid. Nó có 3 đặc tính:

- Khoảng cách giữa hai đối tượng thể hiện "mức độ giống nhau" của 2 đối tượng này theo cảm nhận của khách hàng. Khoảng cách càng nhỏ thể hiện mức độ giống nhau càng nhiều
- Một véc tơ (đoạn thẳng) trên bản đồ biểu thị độ lớn và chiều hướng trong không gian Euclid của các thuộc tính.
- Các trục (hướng) của bản đồ là một tập hợp các véc tơ có thể gợi ra các yếu tố (khía cạnh – dimension) quan trọng chính mô tả cách đối tượng nghiên cứu phân biệt các đối tượng đánh giá như thế nào.

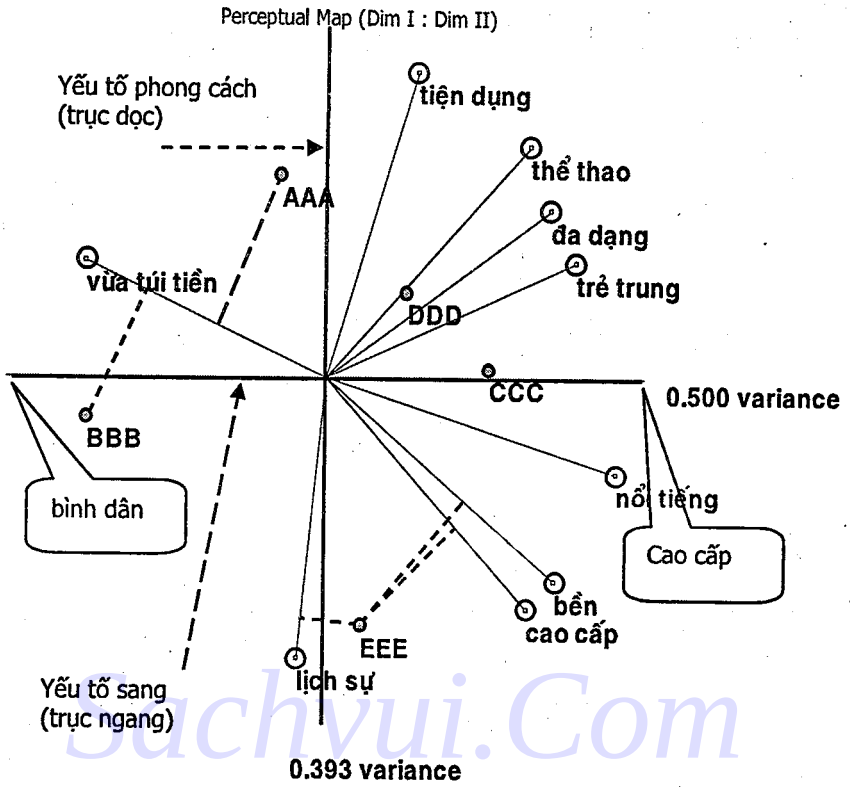
Bản đồ 2 chiều (2 trục) vuông góc thường được sử dụng, và các trục này có thể xoay và có thể không vuông góc. Bản đồ đơn giản nhất có hai trục (theo hai yếu tố cơ bản)

Ví dụ: cảm nhận về các món ăn sáng theo hai yếu tố: đắt-rẻ và có lâu-mau sau khi gọi.



Để đọc hiểu bản đồ nhận thức, hãy xem ví dụ sau về các thương hiệu (TH) thời trang thông dụng

Hình 15.1



Trong hình minh họa trên (kết quả từ chương trình máy tính ME) thì

Cảm nhận về TH theo từng thuộc tính: Khi chiếu từ vị trí của một TH lên véc tơ của một thuộc tính nào đó ta được khoảng cách từ điểm chiếu đó với gốc tọa độ. Khoảng cách này cho biết độ mạnh của TH đó về thuộc tính đang xét. Khoảng cách càng xa gốc toạ độ (khoảng cách so với gốc tọa độ theo hướng của véc tơ thuộc tính) thì thương hiệu càng mạnh về thuộc tính đó. Trong ví dụ hình trên thì BBB có thuộc tính vừa túi tiền hơn AAA. Còn CCC, DDD và EEE không được cảm nhận là vừa túi tiền

Cạnh tranh của các TH: 2 TH càng gần nhau trên bản đồ cho biết cảm nhận của khách hàng về 2 TH này là càng giống nhau trên tất cả thuộc tính đang xem xét, có nghĩa là 2 TH này có chung định vị cảm nhận, càng cạnh tranh trực tiếp với nhau hơn. Hình trên không có trường hợp này.

Trục ngang: có thể coi là tập hợp của các thuộc tính vừa túi tiền, nổi tiếng, cao cấp, bền ... - > thể hiện độ sang của thương hiệu

Trục dọc: có thể được coi là tập hợp của các thuộc tính tiện dụng, thể thao, lịch sự ... -> thể hiện phong cách

Như vậy có thể coi các khách hàng cảm nhận các thương hiệu cạnh tranh theo hai yếu tố quan trọng chính là **mức độ sang trọng** và **phong cách.**

Trong thực tế có thể có nhiều hơn hai yếu tố quan trọng trên 1 bản đồ. Các yếu tố quan trọng này có thể được nhận diện qua phân tích nhân tố đối với các thuộc tính được dùng để đánh giá các thương hiệu.

3. CÁC KỸ THUẬT LẬP BẢN ĐỒ NHẬN THỨC

Có hai kỹ thuật thường dùng để lập bản đồ vị trí thể hiện cảm nhận của khách hàng mục tiêu về các thương hiệu là:

- Multidimensional scaling (MDS): đo lường và thể hiện các đối tượng trong không gian đa chiều hướng hay gọi là đo lường đa hướng
- Correspondence analysis (CA): phân tích và thể hiện sự tương hợp của các đối tượng với các thuộc tính (lý tính và/hay cảm xúc), gọi là phân tích tương hợp.

MDS thường yêu cầu dữ liệu của khảo sát dưới dạng thang đo khoảng cách. Có thể dùng đo lường mức độ giống nhau giữa các đối tượng (similarity-based methods) hay mức độ một đối tượng có một thuộc tính nào đó (attributed based methods). Quyển sách này chỉ trình bày cách thực hiện trên dữ liệu mức độ một đối tượng có một thuộc tính nào đó. Còn CA chỉ cần dữ liệu của khảo sát dưới dạng thang đo danh nghĩa. Các chương trình máy tính thường được sử dụng là SPSS, Excel, XlStat, ME ...

3.1 Kỹ thuật đo lường đa hướng (attribute-based method MDS)

Câu hỏi và thang đo dùng cho MDS dựa trên các thuộc tính mà người trả lời dùng để đánh giá các đối tượng:

Q7a. Đây là những yếu tố Anh/Chị đã đánh giá về độ quan trọng khi chọn mua thời trang thông dụng (casual wear). Bây giờ Anh/Chị vui lòng cho biết theo Anh/Chị từng yếu tố đó phù hợp như thế nào nếu dùng để nói về các thương hiệu thời trang thông dụng. Vui lòng dùng thang điểm từ 1 đến 5, với quy ước:

1	2	3	4	5
Hoàn toàn không phù hợp	Không phù hợp	Bình thường	Phù hợp	Rất phù hợp

Bảng 15.1

		AAA	BBB	CCC	DDD	EEE
[1]	Màu sắc phù hợp với thời trang thông dụng					
[2]	Chất liệu vải phù hợp với thời trang thông dụng					
[3]	Dễ giặt, ủi					
[4]	Có chất lượng may cao					
[5]	Có kiểu dệt mới lạ					
[6]	Có độ bền sản phẩm cao					
[7]	Giá cả hợp lý					
[8]	Có kiểu dáng chững chạc					
[9]	Có kiểu dáng hợp thời trang					
[10]	Nhãn hiệu nổi tiếng					
[11]	Nhân viên bán hàng niềm nở, tận tình,...					
[12]	Được bán tại các cửa hàng thời trang thiết kế bắt mắt					

3.2 Kỹ thuật phân tích tương hợp (CA)

Câu hỏi và thang đo dùng cho CA như sau:

Q18a. **(Showcard)** Vui lòng cho biết thương hiệu mỹ phẩm nào trong danh sách này phù hợp với từng yếu tố mà chúng tôi sẽ đọc lên sau đây.
[PVV lần lượt đọc lên từng yếu tố] [PVV: Chỉ ghi nhận những thương hiệu mà người trả lời biết]. Nhớ xoay vòng các yếu tố

Bảng 15.2

		AAA	BBB	CCC	DDD
[1]	Nhãn hiệu mang tính quốc tế	1	2	3	4
[2]	Mùi hương độc đáo	1	2	3	4
[3]	Mùi hương quyến rũ	1	2	3	4
[4]	Mùi hương tự nhiên	1	2	3	4
[5]	Mùi hương giữ được lâu	1	2	3	4
[6]	Không bị đổi mùi	1	2	3	4
[7]	Kiểu dáng bao bì hấp dẫn	1	2	3	4
[8]	Nhãn hiệu nổi tiếng	1	2	3	4
[9]	Nhân viên tư vấn bán hàng tốt	1	2	3	4
[10]	Hình ảnh cửa hàng ấn tượng	1	2	3	4
[11]	Được giới thiệu/quảng cáo là dành cho những người như tôi mong muốn	1	2	3	4
[12]	Được nhiều người ao ước sở hữu	1	2	3	4
[13]	Được nói đến nhiều trên các tạp chí thời trang	1	2	3	4
[14]	Được giới thượng lưu ưa chuộng	1	2	3	4
[15]	Được bạn trai/chồng/đàn ông ưa thích	1	2	3	4

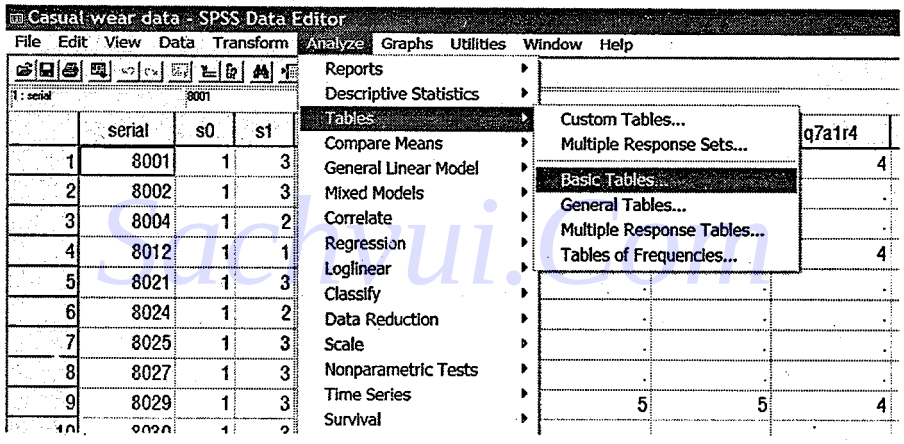
4. SỬ DỤNG SPSS ĐỂ LẬP BẢN ĐỒ VỚI KỸ THUẬT MDS

Trong phần này chúng ta chỉ xem xét kỹ thuật MDS với dữ liệu đánh giá các đối tượng theo các yếu tố trên thang đo khoảng cách. Để minh họa cho phần chạy lệnh MDS, hãy sử dụng file dữ liệu mẫu trong cơ sở dữ liệu dùng với sách có tên là **Casual wear data.sav**

BƯỚC 1: Tính điểm trung bình của từng đối tượng (trong ví dụ này là các thương hiệu thời trang thông dụng) theo các thuộc tính

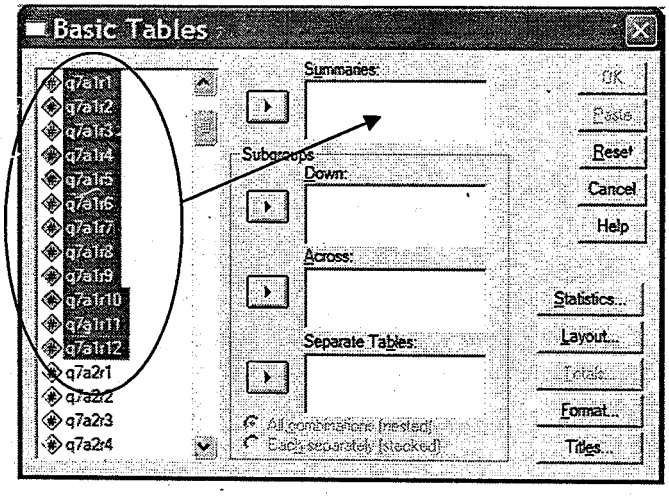
Từ data file, dùng lệnh lập bảng tính trung bình như sau: Analyze > Tables > Basic Tables như trong hình dưới đây:

Hình 15.2



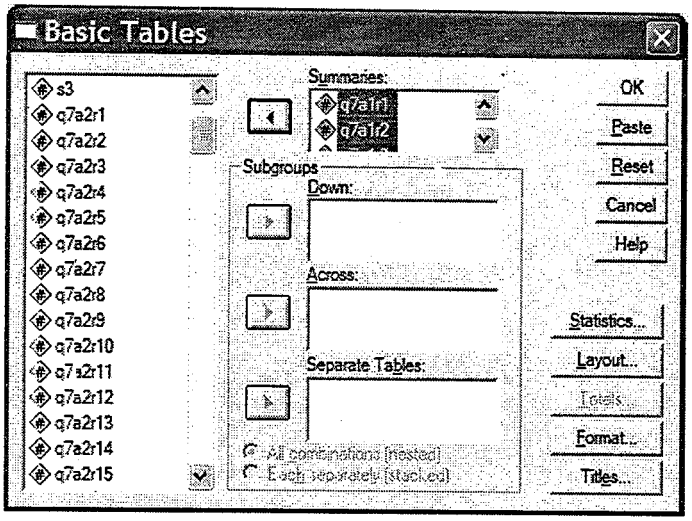
Lệnh này sẽ mở ra hộp thoại sau:

Hình 15.3



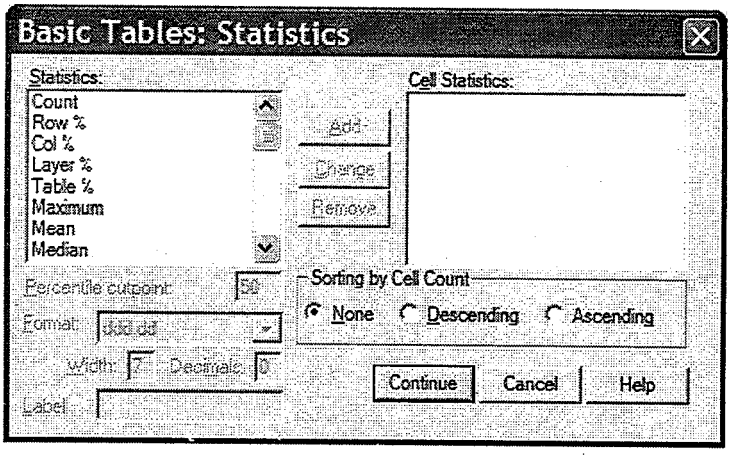
Trong hộp thoại này, chọn các biến chứa điểm đánh giá từng TH theo các thuộc tính (trong ví dụ này có 12 thuộc tính thì sẽ có 12 biến cho từng TH) đưa vào ô Summaries như hình dưới: (Chú ý mỗi lần thực hiện cho 1 TH, và thực hiện nhiều lần)

Hình 15.4



Sau khi đưa biến ứng với các thuộc tính vào ô Summaries, nhấp chuột vào nút Statistics để chọn hàm thống kê. Lệnh này sẽ mở tiếp hộp thoại sau:

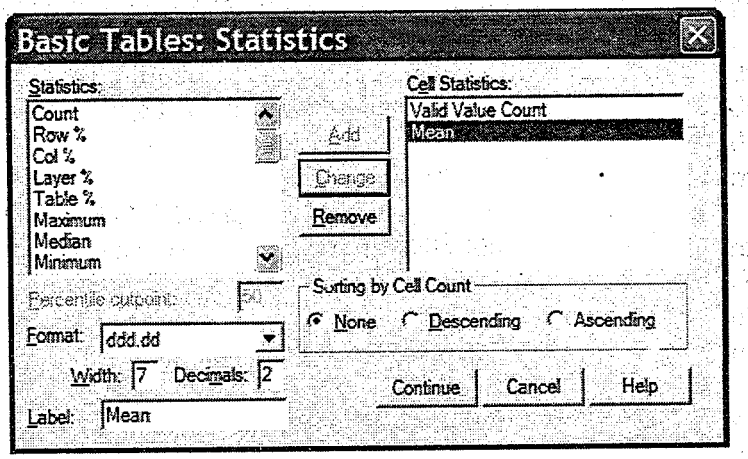
Hình 15.5



Trong hộp thoại này hãy chọn lần lượt 2 hàm là Valid value count và Mean (để đếm số trường hợp đánh giá cho thương hiệu đang tính và

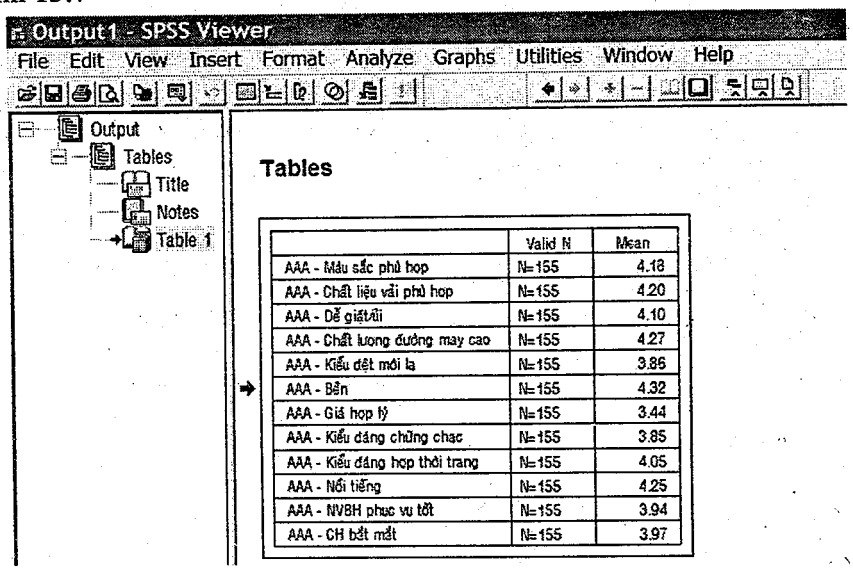
tính ra điểm trung bình ở từng thuộc tính cho TH này) trong ô Statistics rồi nhấp chuột vào nút Add để đưa vào ô bên tay phải (Cell Statistics).

Hình 15.6



Cần chỉnh sửa định dạng của số liệu tính ra bằng cách chọn tên hàm đã chọn trong ô Cell Statistics, và chỉnh sửa định dạng ở góc dưới bên trái của hộp thoại (Phần Format và Label). Sau đó nhấp chuột vào nút Continue để trở lại hộp thoại cấp trên Basic Tables. Cuối cùng là nhấp chuột vào nút OK để thực thi. Kết quả sẽ xuất hiện như sau:

Hình 15.7



	Valid N	Mean
AAA - Màu sắc phù hợp	N=155	4.18
AAA - Chất liệu vải phù hợp	N=155	4.20
AAA - Dễ giặt,ủi	N=155	4.10
AAA - Chất lượng đường may cao	N=155	4.27
AAA - Kiểu dệt mới lạ	N=155	3.86
AAA - Bền	N=155	4.32
AAA - Giá hợp lý	N=155	3.44
AAA - Kiểu dáng chững chạc	N=155	3.85
AAA - Kiểu dáng hợp thời trang	N=155	4.05
AAA - Nổi tiếng	N=155	4.25
AAA - NVBH phục vụ tốt	N=155	3.94
AAA - CH bắt mắt	N=155	3.97

Như vậy chúng ta đã có kết quả tính trung bình của thương hiệu thứ nhất theo các thuộc tính. Thực hiện lại lệnh này cho 4 thương hiệu còn lại, chúng ta sẽ được thêm 4 bảng kết quả tương tự. Sau đó copy bảng kết quả này sang Excel để sắp xếp kết quả tính trung bình theo các thuộc tính cho cả 5 TH như sau:

Hình 15.8

Microsoft Excel - Garment - MDS

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U 100%

A46

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		AAA	BBB	CCC	DDD	EEE						
3	Maøu saéc	4.181	3.807	3.768	4.108	4.245						
4	Chaát lieäu vaûi	4.200	3.895	3.899	4.216	4.245						
5	Deã giaët/uûi	4.103	3.912	3.826	4.081	4.143						
6	CL ñoâng may cao	4.271	3.982	3.942	4.270	4.265						
7	Kieåu deát moûi laï	3.858	3.544	3.478	3.838	4.020						
8	Beàn	4.316	3.895	3.913	4.189	4.367						
9	Giaù hôïp lyù	3.439	4.018	3.754	3.378	3.347						
10	K daùng chöông chaïc	3.845	3.930	3.812	3.892	4.020						
11	K.daùng thôøi trang	4.045	3.754	3.797	4.135	4.041						
12	Noåi tieáng	4.252	3.754	3.725	4.216	4.327						
13	NVBH phuïc vuï toát	3.935	3.772	3.855	3.811	3.898						
14	CH baét maét	3.968	3.596	3.754	3.703	3.918						
15												
16		Maøu saéc	Chaát lieäu	Deã giaët/	Chaát löôn	Kieåu deát	Beàn	Giaù hôïp l	Kieåu daùn	Kieåu daùn	Noåi tieáng	NVBH phu
17	AAA	4.181	4.200	4.103	4.271	3.858	4.316	3.439	3.845	4.045	4.252	3.935
18	BBB	3.807	3.895	3.912	3.982	3.544	3.895	4.018	3.930	3.754	3.754	3.772
19	CCC	3.768	3.899	3.826	3.942	3.478	3.913	3.754	3.812	3.797	3.725	3.855
20	DDD	4.108	4.216	4.081	4.270	3.838	4.189	3.378	3.892	4.135	4.216	3.811
21	EEE	4.245	4.245	4.143	4.265	4.020	4.367	3.347	4.020	4.041	4.327	3.898

Trong hình trên của cửa sổ Excel, phần trên là chúng ta gom 5 bảng kết quả tính trung bình lại thành 1 bảng chung. Sau đó dùng lệnh Copy và Paste Special > Transpose để đảo chiều của bảng dữ liệu trung bình này như phần dưới của cửa sổ Excel. Sau đó trở lại cửa sổ data của SPSS, mở ra file mới và đưa bảng dữ liệu trung bình đã đảo chiều vào trong SPSS. Trong đó chú ý biến đầu tiên là kiểu String để ghi tên Thương hiệu. Các biến còn lại được đặt tên để ghi chú đây là biến chứa dữ liệu điểm đánh giá trung bình theo từng thuộc tính. Sau đó nhớ khai báo các label ứng với các biến thuộc tính (xem 2 hình tiếp theo)

Hình 15.9

Casual wear data - MDS - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : brand AAA

	brand	att1	att2	att3	att4	att5	att6	att7	att8	att9	att10	att11	att12	var
1	AAA	4.18	4.20	4.10	4.27	3.86	4.32	3.44	3.85	4.05	4.25	3.94	3.97	
2	BBB	3.81	3.90	3.91	3.98	3.54	3.90	4.02	3.93	3.75	3.75	3.77	3.60	
3	CCC	3.77	3.90	3.83	3.94	3.48	3.91	3.75	3.81	3.80	3.73	3.86	3.75	
4	DDD	4.11	4.22	4.08	4.27	3.84	4.19	3.38	3.89	4.14	4.22	3.81	3.70	
5	EEE	4.25	4.25	4.14	4.26	4.02	4.37	3.35	4.02	4.04	4.33	3.90	3.92	

Hình 15.10

Casual wear data - MDS - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

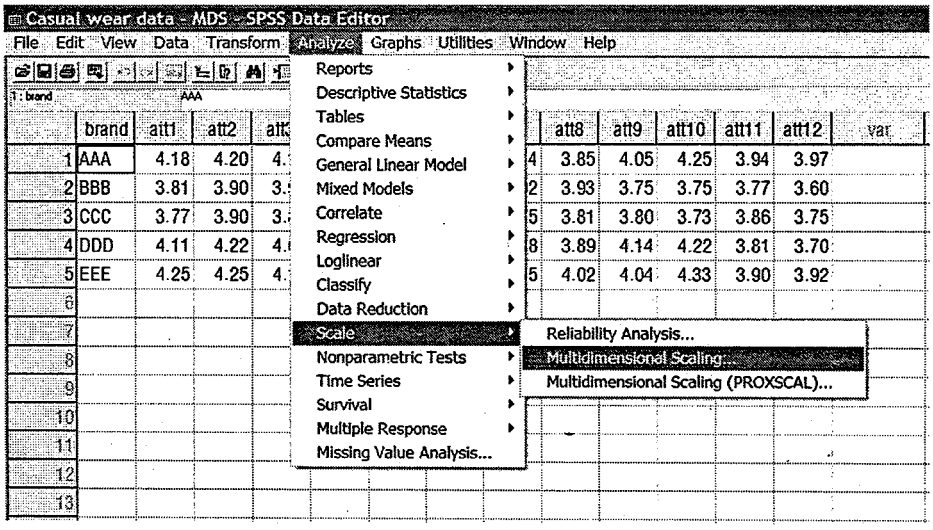
	Name	Type	Width	Deci	Label	Values	Missing
1	brand	String	30	0		None	None
2	att1	Numeric	8	2	Mau sac	None	None
3	att2	Numeric	8	2	Chat lieu	None	None
4	att3	Numeric	8	2	De giat/ui	None	None
5	att4	Numeric	8	2	CL may	None	None
6	att5	Numeric	8	2	Kieu det	None	None
7	att6	Numeric	8	2	Ben	None	None
8	att7	Numeric	8	2	Gia hop ly	None	None
9	att8	Numeric	8	2	K.Dang chung chac	None	None
10	att9	Numeric	8	2	K.Dang thoi trang	None	None
11	att10	Numeric	8	2	Noi tieng	None	None
12	att11	Numeric	8	2	NVBH phuc vu tot	None	None
13	att12	Numeric	8	2	CH bat mat	None	None
14							

Đến đây chúng ta đã chuẩn bị xong data để chạy lệnh MDS

BƯỚC 2: Chạy lệnh MDS để chuyển dữ liệu đánh giá các TH theo các thuộc tính thành các khoảng cách phản ánh mức độ giống nhau trong không gian đa chiều hướng.

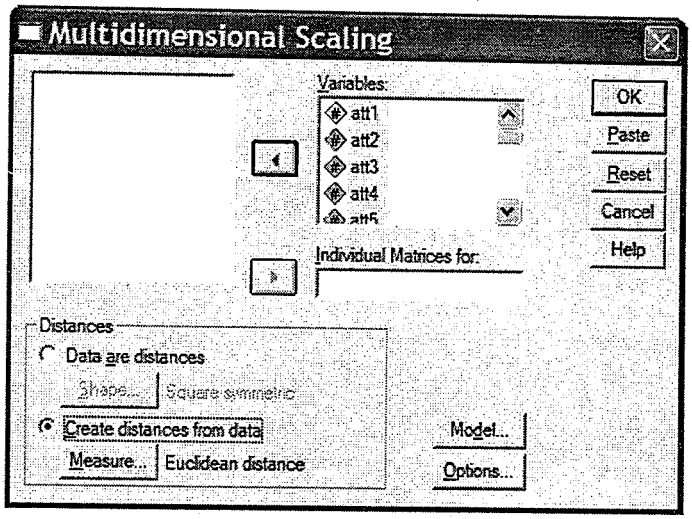
Chọn Analyze > Scale > Multidimensional Scaling như hình dưới

Hình 15.11



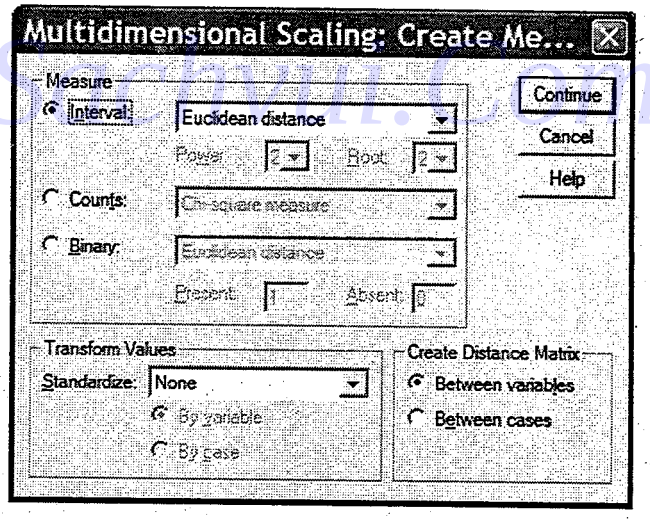
Lệnh này mở ra hộp thoại MDS sau:

Hình 15.12



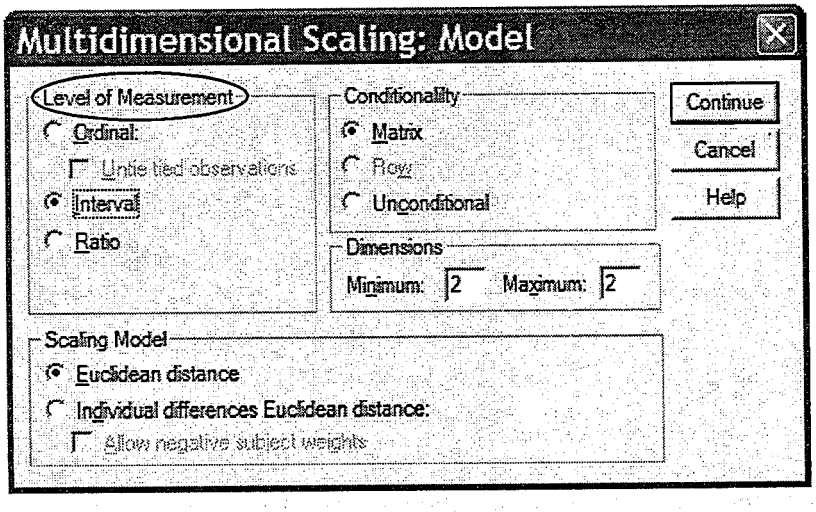
Trong hộp thoại MDS, đưa các biến thuộc tính vào ô Variables. Sau đó nhấp chuột chọn Create distances from data, và nhấp vào nút Measure ngay bên dưới, lệnh này mở ra hộp thoại con như sau:

Hình 15.13



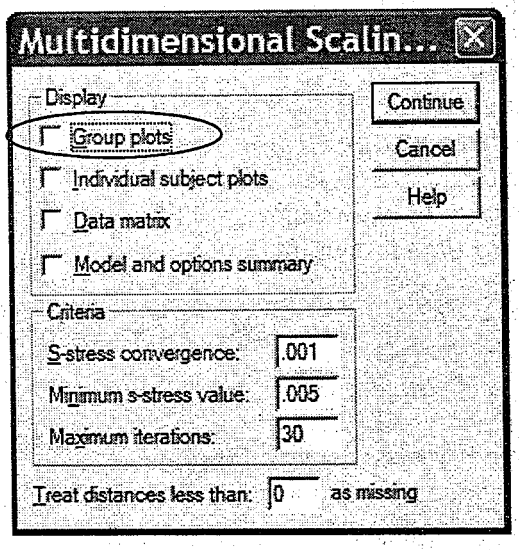
Trong hộp thoại MDS: Create Measure, ở phần góc dưới bên tay phải hãy chọn option Between variables để tính toạ độ của các thuộc tính trong không gian đa hướng, sau đó nhấp chuột vào nút Continue để trở lại hộp thoại MDS. Tiếp theo trong hộp thoại MDS, nhấp chuột vào nút Model và mở ra hộp thoại như sau:

Hình 15.14



Trong hộp thoại MDS: Model, chọn Level of Measurement là Interval (dữ liệu khoảng cách). Số chiều của không gian đa chiều hướng (Dimensions) được xác lập mặc định là 2, hãy giữ nguyên, rồi nhấp chuột vào nút Continue trở lại hộp thoại MDS. Tiếp theo trong hộp thoại MDS, nhấp chuột vào nút Options mở ra hộp thoại sau đây:

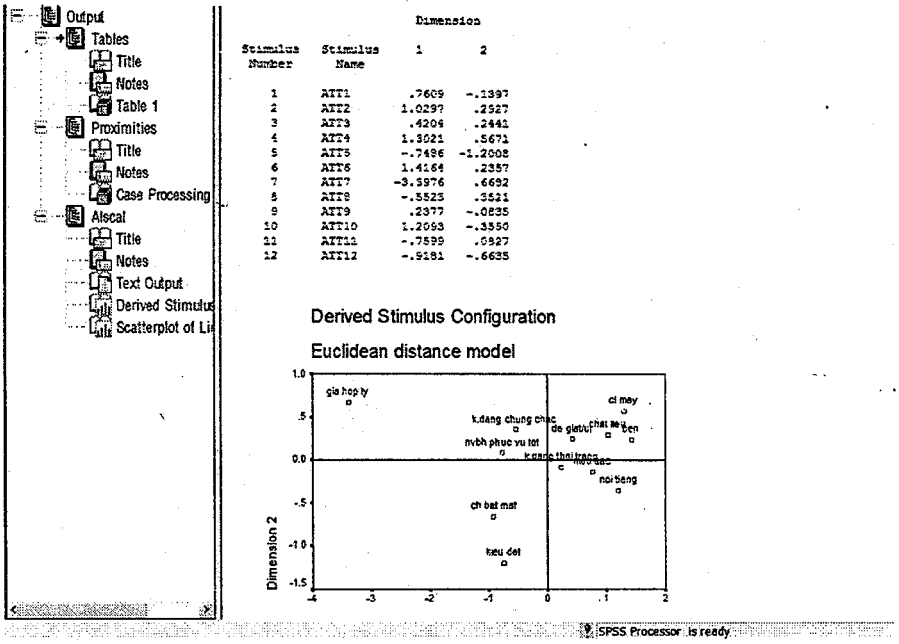
Hình 15.15



Trong hộp thoại MDS: Options, hãy chọn Display Group plots để thể hiện vị trí của các thuộc tính trong bản đồ không gian, sau đó nhấp nút Continue trở về hộp thoại MDS.

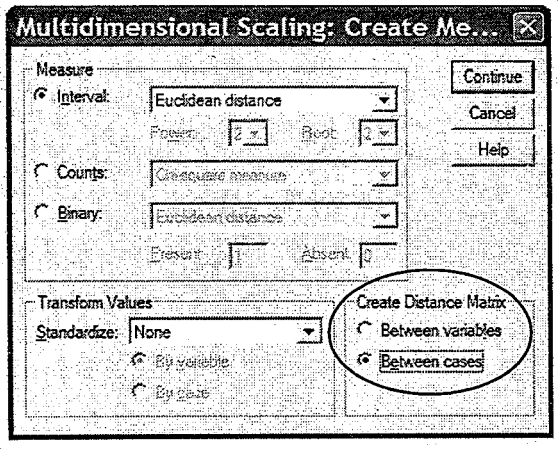
Trong hộp thoại MDS hãy nhấp nút OK và lệnh được thực thi, chúng ta sẽ có kết quả trong cửa sổ output của SPSS như sau:

Hình 15.16



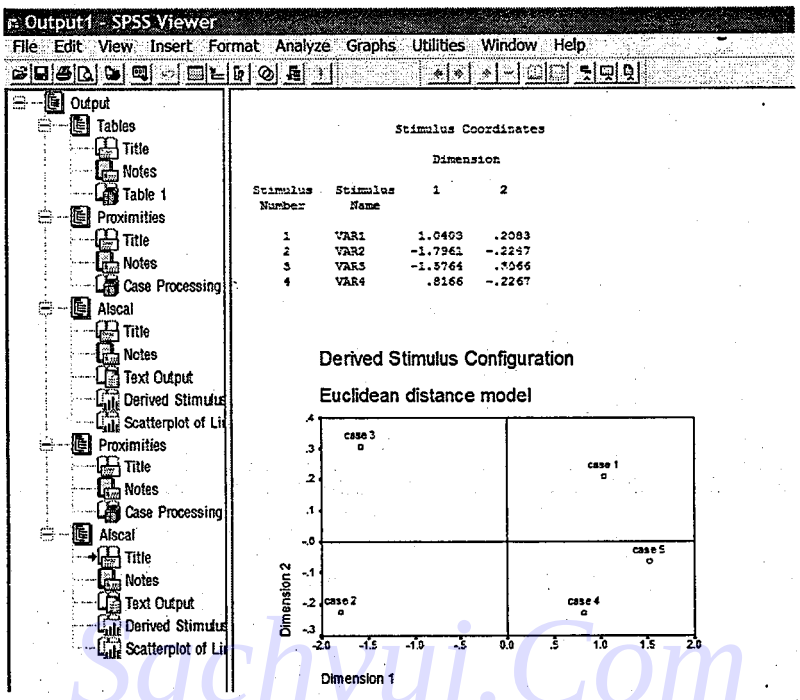
Trong kết quả trên chúng ta chỉ mới có toạ độ và bản đồ vị trí của các thuộc tính trong không gian đa hướng (ở đây là 2 chiều). Chúng ta cần phải thực hiện lại quy trình này để tính tọa độ cho các thương hiệu. Khi thực hiện lại quy trình này chú ý khi tạo ra các tọa độ trong không gian đa chiều hướng bây giờ hãy chọn Between cases thay vì Between variables như lần trước.

Hình 15.17



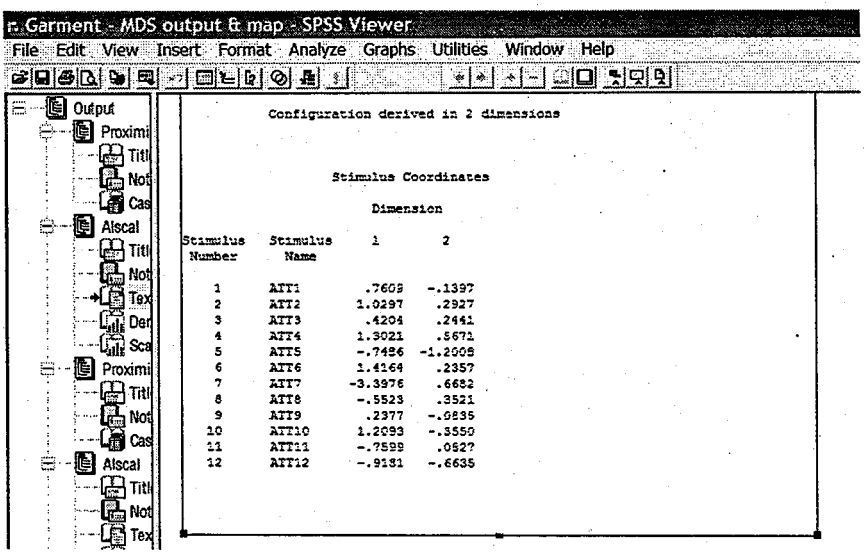
Khi thực hiện lệnh thì kết quả tính toán toạ độ và bản đồ vị trí của các thương hiệu trong không gian đa hướng (ở đây là 2 chiều) xuất hiện như Hình 15.18.

Hình 15.18

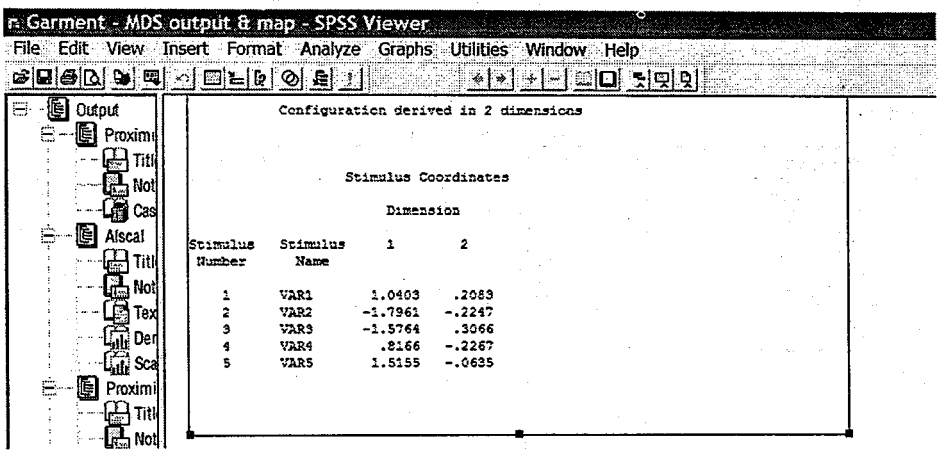


Sau khi có các tọa độ của các thuộc tính và thương hiệu trong không gian đa hướng, chúng ta lần lượt copy các tọa độ (Coordinates) của (1) các thuộc tính và (2) các thương hiệu (theo 2 chiều hướng – 2 dimensions) qua Excel và nối chúng lại với nhau như trong 3 hình sau:

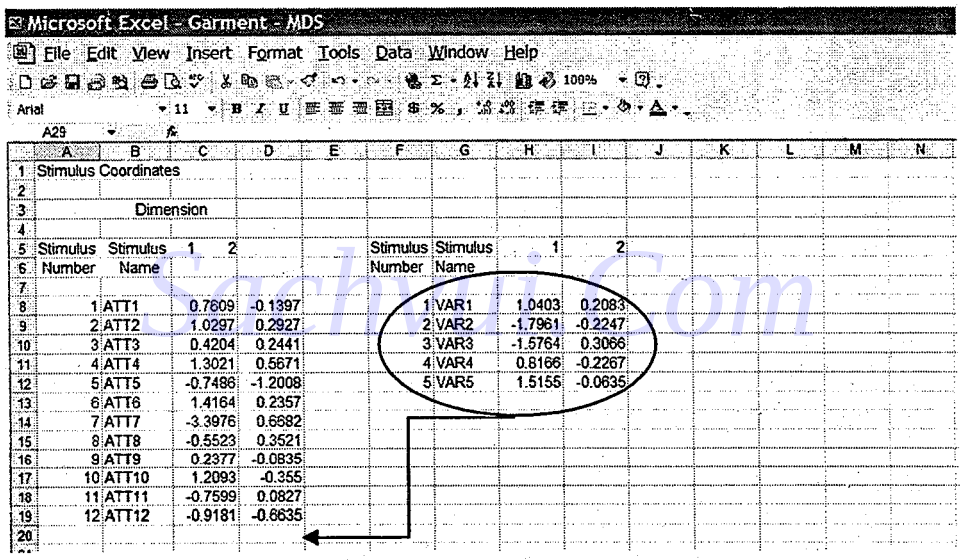
Hình 15.19



Hình 15.20



Hình 15.21



Sau khi chép qua Excel để dễ dàng nối các toạ độ của các thuộc tính và các thương hiệu, chúng ta trở lại màn hình data của SPSS mở ra file mới và chép các tọa độ đã nối lại với nhau vào SPSS như trong hình sau:

Hình 15.22

Casual wear - MDS graph - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : label Mau sac

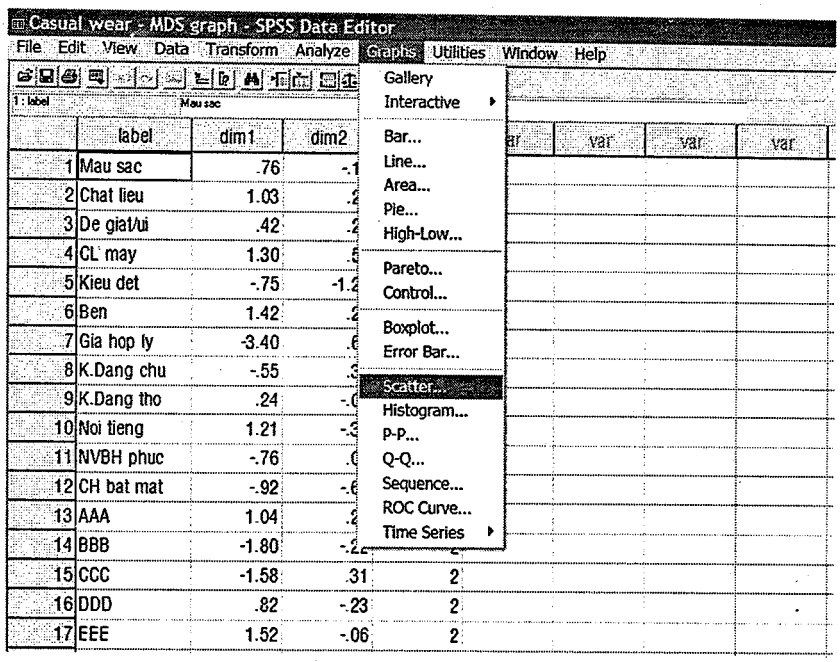
	label	dim1	dim2	object	var	var	var	var
1	Mau sac	.76	-.14	1				
2	Chat lieu	1.03	.29	1				
3	De giat/ui	.42	.24	1				
4	CL may	1.30	.57	1				
5	Kieu det	-.75	-1.20	1				
6	Ben	1.42	.24	1				
7	Gia hop ly	-3.40	.67	1				
8	K.Dang chu	-.55	.35	1				
9	K.Dang tho	.24	-.08	1				
10	Noi tieng	1.21	-.36	1				
11	NVBH phuc	-.76	.08	1				
12	CH bat mat	-.92	-.66	1				
13	AAA	1.04	.21	2				
14	BBB	-1.80	-.22	2				
15	CCC	-1.58	.31	2				
16	DDD	.82	-.23	2				
17	EEE	1.52	-.06	2				

Trước khi dùng công cụ Graph trong SPSS vẽ bản đồ thể hiện cả các thuộc tính và các thương hiệu, cần khai báo biến label ghi chú các thuộc tính và tên TH. Chú ý ghi các label tóm tắt vì SPSS không chấp nhận các label dài. Để tạo định dạng thể hiện trên đồ thị giữa thuộc tính và TH, có thể khai báo thêm biến object để phân biệt giữa hai loại đối tượng này trên đồ thị.

BƯỚC 3: Vẽ bản đồ vị trí và hiệu chỉnh bản đồ

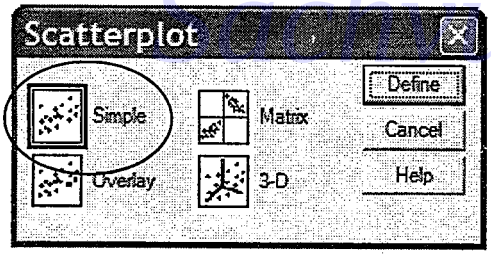
Trong cửa sổ data của SPSS đang chứa dữ liệu tọa độ của các thuộc tính và TH, từ menu chọn: Graphs > Scatter ... như hình sau:

Hình 15.23



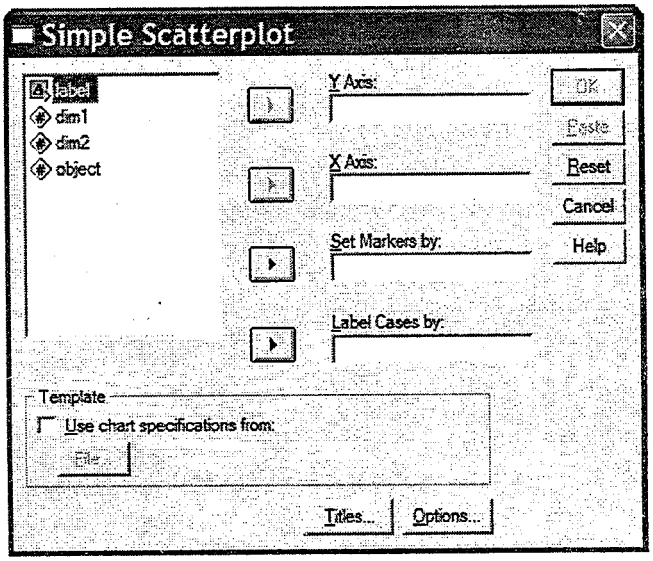
Lệnh này mở ra hộp thoại Scatterplot sau:

Hình 15.24



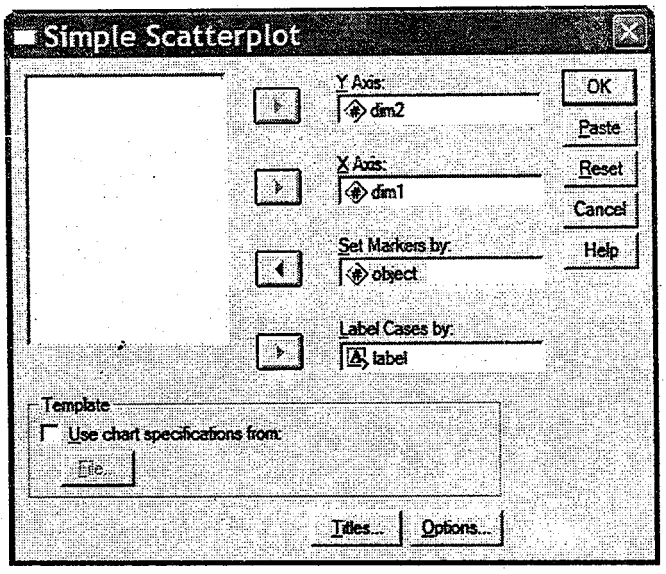
Trong hộp thoại Scatterplot hãy chọn loại Simple và mở ra hộp thoại Simple Scatterplot như sau:

Hình 15.25



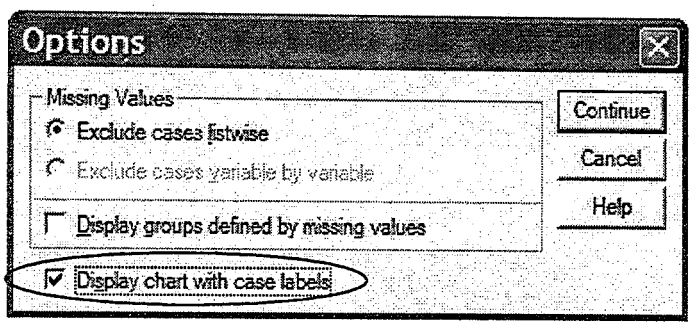
Trong hộp thoại Simple Scatterplot lần lượt đưa các biến trong tập tin dữ liệu chứa các toạ độ vào như hình dưới:

Hình 15.26



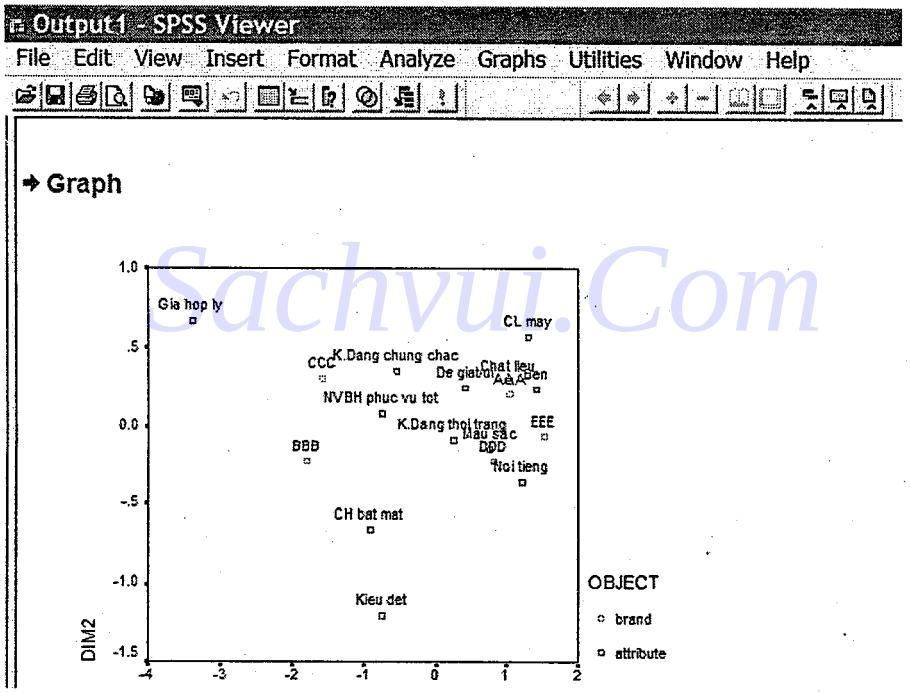
Tiếp theo nhấp chuột vào nút Options... xuất hiện hộp thoại Options và chọn vào mục Display chart with case labels như sau:

Hình 15.27



Sau đó nhấp nút Continue trở về hộp thoại Simple Scatterplot rồi nhấp nút OK, bản đồ vị trí sẽ xuất hiện trong cửa sổ output như sau:

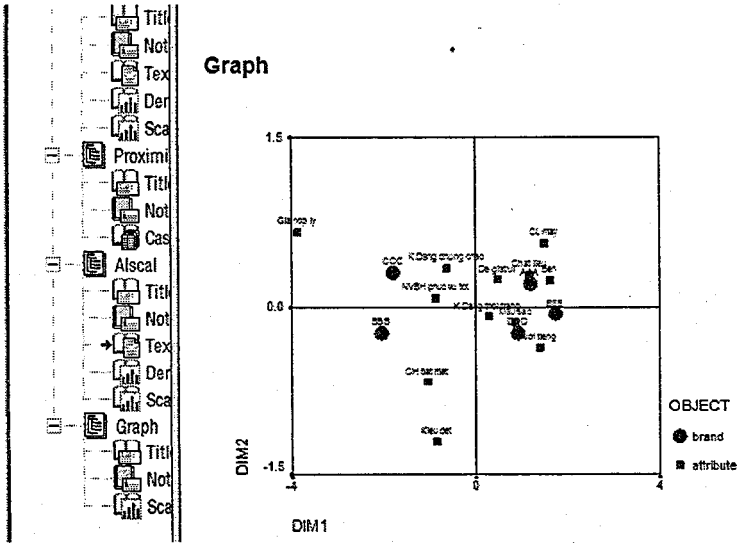
Hình 15.28



Để bản đồ vị trí dễ xem hơn, chúng ta nhấp nhanh chuột hai lần vào bản đồ vị trí để chỉnh sửa Font chữ, màu sắc, cách quy ước các đối tượng trên bản đồ ...

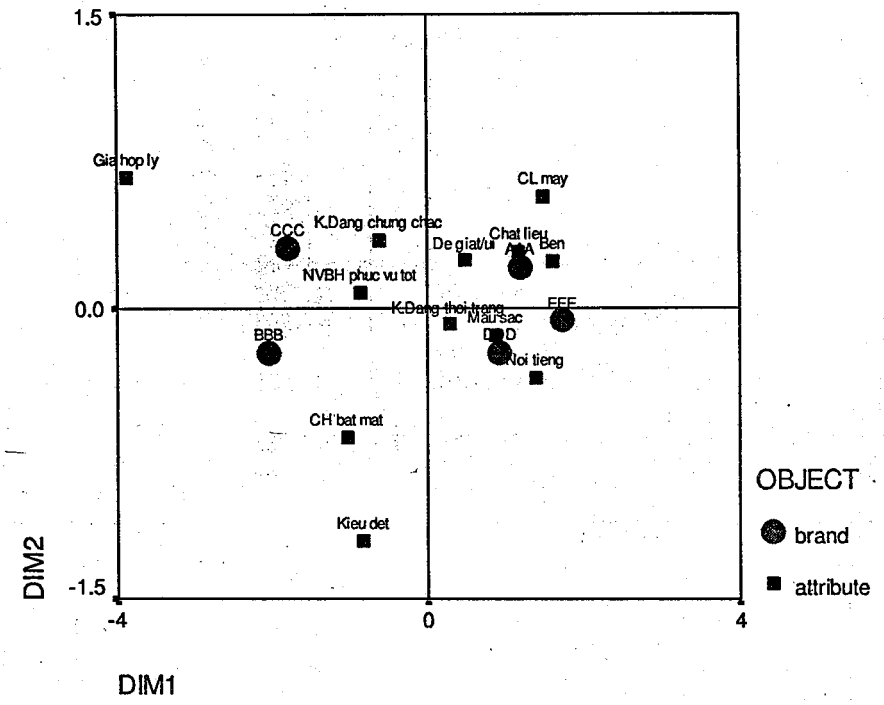
Sau một số thay đổi định dạng, bản đồ thể hiện vị trí các thương hiệu xuất hiện như trong hình minh họa dưới đây:

Hình 15.29



Có thể copy bản đồ vị trí này vào các chương trình Word hay PowerPoint, phóng to hay thu nhỏ lại để trình bày.

Hình 15.30



5. SỬ DỤNG SPSS ĐỂ LẬP BẢN ĐỒ VỚI KỸ THUẬT TƯƠNG HỢP (Correspondence Analysis)

Để minh họa cho phần chạy lệnh CA với dữ liệu đánh giá các đối tượng theo các yếu tố bằng thang đo danh nghĩa, hãy sử dụng file dữ liệu mẫu trong cơ sở dữ liệu dùng với sách có tên là **Data My pham.sav** (trong ví dụ này là các thương hiệu nước hoa)

BƯỚC 1: Tính tần số (số người trả lời đồng ý thương hiệu có sở hữu các thuộc tính)

Câu hỏi đánh giá các thương hiệu theo các thuộc tính được tạo khuôn và nhập liệu vào SPSS như được mô tả trong hai hình dưới đây (chú ý mỗi thuộc tính có 4 biến vì có 4 TH được đánh giá)

Hình 15.31

Data My pham - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : q18aa_1 2

	q18aa_1	q18aa_2	q18aa_3	q18aa_4	q18ab_1	q18ab_2	q18ab_3	q18ab_4	q18ac_1	q18ac_2
1	2	3	.	.	2	3	.	.	2	3
2	1	3	4	.	1	3	.	.	3	.
3	2	.	.	.	9	.	.	.	9	.
4	1	2	3	.	1	2	3	.	9	.
5	1	.	.	.	2	.	.	.	1	.
6	2	.	.	.	2	.	.	.	2	.
7	2	.	.	.	2	.	.	.	3	.
8	1	2	.	.	1	2	3	.	1	2

Hình 15.32

Data My pham - SPSS Data Editor

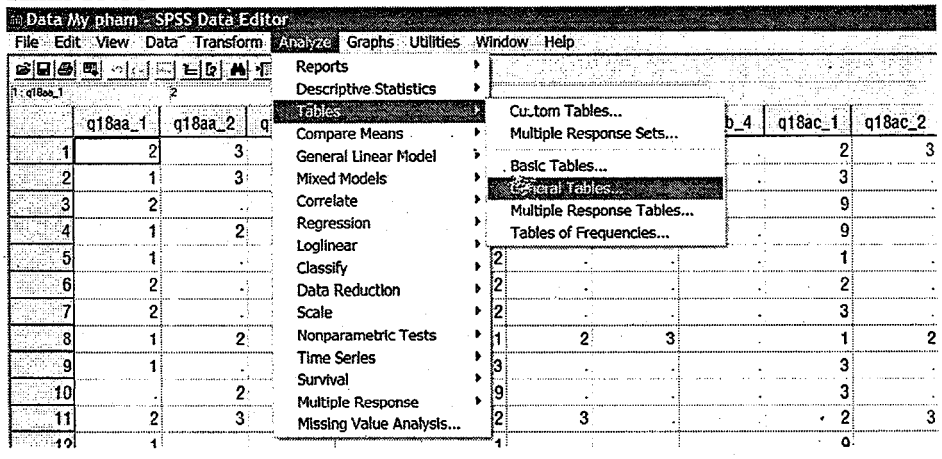
File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing
7	q18aa_1	Numeric	11	0	Tên nhãn hiệu mang tính quốc tế	{1, AAA}...	9
8	q18aa_2	Numeric	11	0		None	9
9	q18aa_3	Numeric	11	0		None	9
10	q18aa_4	Numeric	11	0		None	9
11	q18ab_1	Numeric	11	0	Mùi hương độc đáo	{1, AAA}...	9
12	q18ab_2	Numeric	11	0		None	9
13	q18ab_3	Numeric	11	0		None	9
14	q18ab_4	Numeric	11	0		None	9
15	q18ac_1	Numeric	11	0	Mùi hương quyến rũ	{1, AAA}...	9
16	q18ac_2	Numeric	11	0		None	9
17	q18ac_3	Numeric	11	0		None	9
18	q18ac_4	Numeric	11	0		None	9

(Chú ý: cách tạo khuôn và nhập liệu có thể khác nhau tùy theo cách đặt câu hỏi và cách tạo khuôn nhập liệu của người làm công tác dữ liệu. Cách làm trên chỉ là 1 ví dụ).

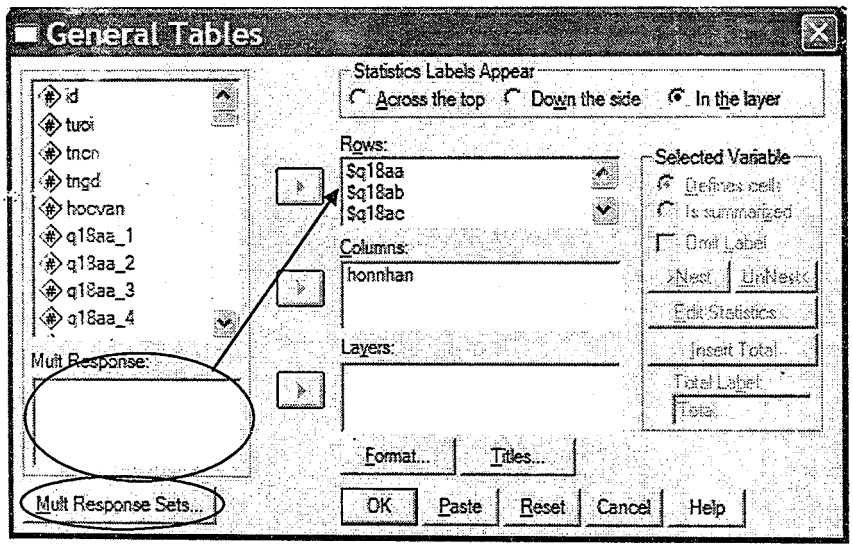
Từ menu ra lệnh Analyze > General Tables như hình sau:

Hình 15.33



Lệnh này mở ra hộp thoại General Tables. Trong hộp thoại này ghép các biến của cùng một thuộc tính lại với nhau (nhấp nút Mult Response Sets ... Xem lại chi tiết trong Chương 3). Do có 15 thuộc tính nên sẽ có 15 biến ghép trong ô Multi Response. Sau đó đưa các biến đã ghép vào ô Rows để tính tần số và biết mỗi thương hiệu có bao nhiêu người đồng ý là có thuộc tính nào như hình dưới đây:

Hình 15.34



Sau đó nhấp chuột vào nút OK, kết quả sẽ xuất hiện trong cửa sổ Output của SPSS như sau:

Hình 15.35

My pham - CA output & map - SPSS Viewer

File Edit View Insert Format Analyze Graphs Utilities Window Help

Quốc tế	AAA	67
	BBB	83
	CCC	52
	DDD	57
Mùi độc dác	AAA	44
	BBB	72
	CCC	55
	DDD	30
Mùi quyến rũ	AAA	38
	BBB	75
	CCC	63
	DDD	31
Mùi tự nhiên	AAA	43
	BBB	51

Trong bảng kết quả trên ta thấy Thương hiệu BBB có 83 người đồng ý là NH quốc tế, 72 người đồng ý là có mùi độc đáo, có 75 người đồng ý là có mùi quyến rũ ...

BƯỚC 2: Sắp xếp lại kết quả tính tần số và chuẩn bị dữ liệu để thực hiện Correspondence Analysis

Để tiện cho việc trình bày và hình dung cũng như tính toán tiếp theo chúng ta sẽ đưa kết quả này qua Excel để sắp xếp lại.

Hình 15.36

Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

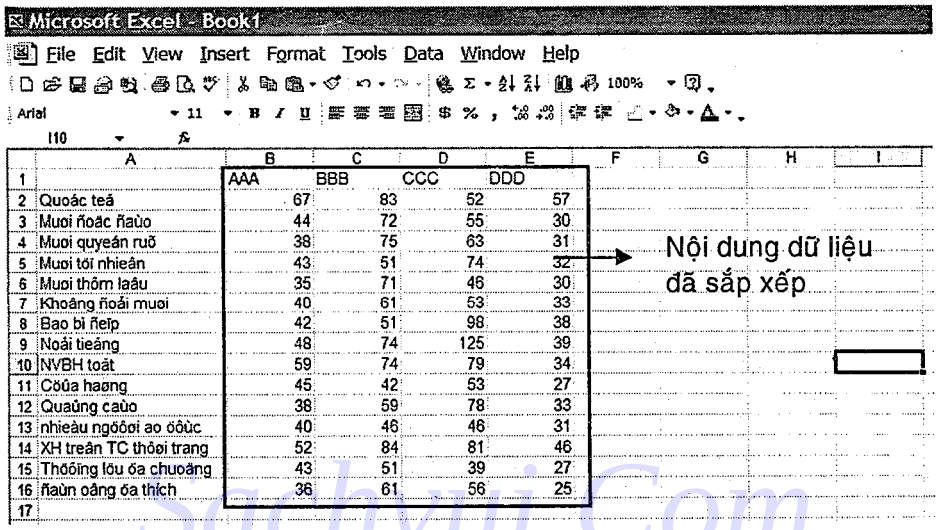
Arial 10 B I U 100%

A1

	A	B	C	D	E	F	G	H	I	J	K
1											
2	Quoác teá	AAA	67								
3		BBB	83								
4		CCC	52								
5		DDD	57								
6	Muøi ñoác	AAA	44								
7		BBB	72								
8		CCC	55								
9		DDD	30								
10	Muøi quye	AAA	38								
11		BBB	75								
12		CCC	63								
13		DDD	31								
14	Muøi töï nh	AAA	43								
15		BBB	51								
16		CCC	74								

Trong Excel dùng các lệnh Copy và Paste Special >Transpose để sắp xếp lại dữ liệu trong đó các dòng là thuộc tính, các cột là tên thương hiệu. Kết quả sắp xếp được thể hiện như trong hình sau:

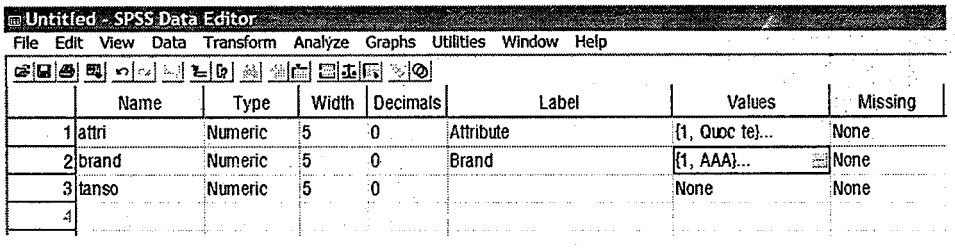
Hình 15.37



	A	B	C	D	E	F	G	H	I
1		AAA	BBB	CCC	DDD				
2	Quoác teá	67	83	52	57				
3	Muoi ñoác ñaùo	44	72	55	30				
4	Muoi quyeán ruõ	38	75	63	31				
5	Muoi töi nhieân	43	51	74	32				
6	Muoi thôm laâu	35	71	46	30				
7	Khoáng ñoái muoi	40	61	53	33				
8	Bao bì ñeïp	42	51	98	38				
9	Noái tieáng	48	74	125	39				
10	NVBH toát	59	74	79	34				
11	Cöûa haøng	45	42	53	27				
12	Quaûng caùo	38	59	78	33				
13	nhieàu ngöôøi ao öôùc	40	46	46	31				
14	XH treân TC thôøi trang	52	84	81	46				
15	Thöôøng löu öa chuoäng	43	51	39	27				
16	ñaùn oâng öa thích	36	61	56	25				
17									

Với kết quả sắp xếp trên chúng ta thấy các số liệu tổng hợp về thuộc tính của từng NH rõ ràng hơn (dễ vẽ biểu đồ ziczac). Sau đó chúng ta đưa lại các kết quả này vào trở lại cửa số data của SPSS. Chúng ta phải mở ra một file data SPSS mới và khai báo 3 biến: attri (thuộc tính, có 15 mã ứng với 15 thuộc tính), brand (thương hiệu, có 4 mã ứng với 4 thương hiệu), và tần số trong cửa số data ở chế độ Variable view như trong hình dưới:

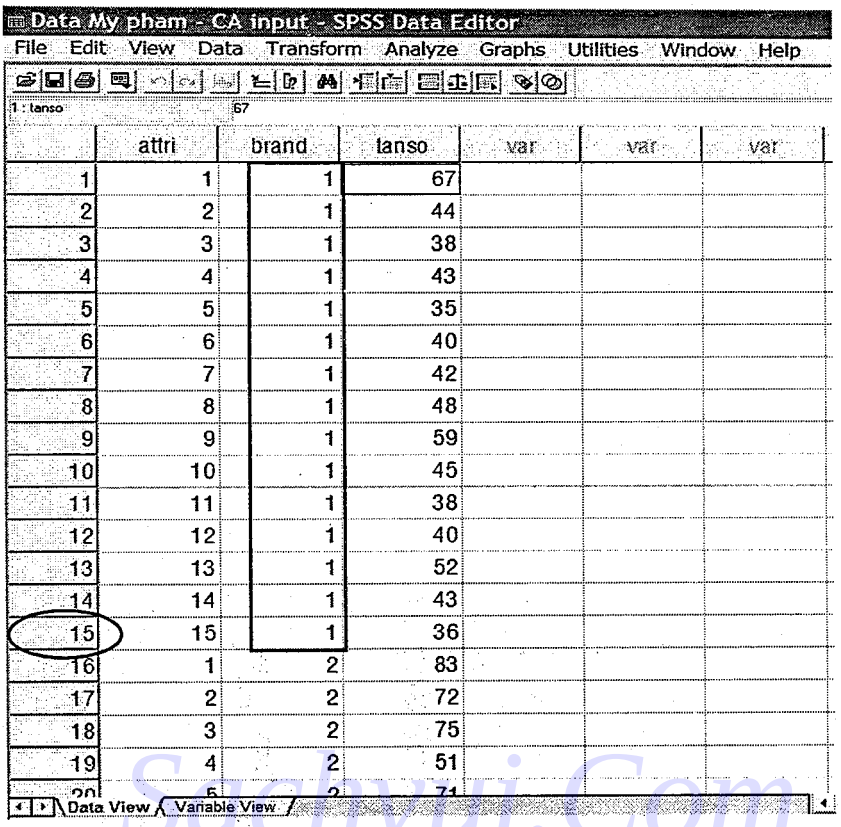
Hình 15.38



	Name	Type	Width	Decimals	Label	Values	Missing
1	attri	Numeric	5	0	Attribute	{1, Quoc te}...	None
2	brand	Numeric	5	0	Brand	{1, AAA}...	None
3	tanso	Numeric	5	0		None	None
4							

Trả cửa sổ data của SPSS về chế độ Data view và copy nội dung dữ liệu (tần số) đã sắp xếp từ Excel vào cửa sổ Data của SPSS ở cột biến tanso. Mỗi lần copy cho một thương hiệu và nối đuôi liên tiếp nhau. Như trong hình sau:

Hình 15.40

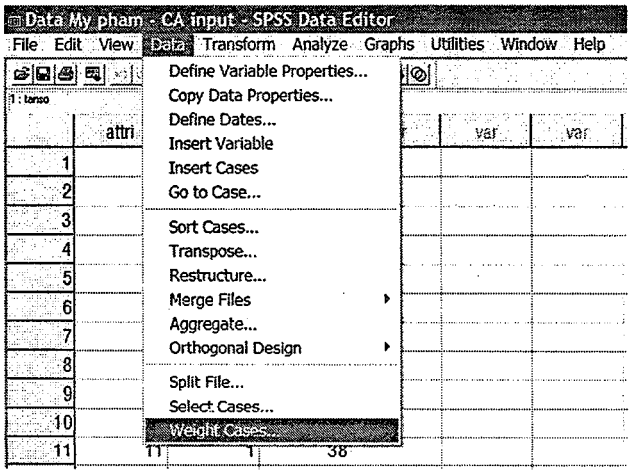


	attri	brand	tanso	var	var	var
1	1	1	67			
2	2	1	44			
3	3	1	38			
4	4	1	43			
5	5	1	35			
6	6	1	40			
7	7	1	42			
8	8	1	48			
9	9	1	59			
10	10	1	45			
11	11	1	38			
12	12	1	40			
13	13	1	52			
14	14	1	43			
15	15	1	36			
16	1	2	83			
17	2	2	72			
18	3	2	75			
19	4	2	51			

Trong hình trên, chúng ta có thể thấy 15 tần số ứng với 15 thuộc tính của thương hiệu AAA được copy vào lại cửa sổ Data, tương ứng thì chúng ta phải nhập 15 mã của 15 thuộc tính này trong cột biến attri và nhập số 1 (mã số của thương hiệu AAA) vào 15 ô tương ứng trong cột biến brand và sẽ tiếp tục như vậy cho 3 thương hiệu còn lại.

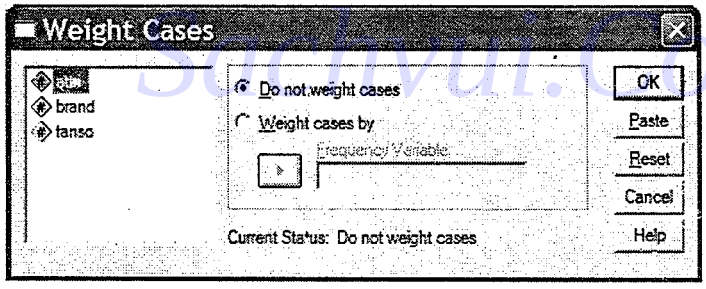
Tiếp theo chúng ta chuẩn bị dữ liệu để chạy lệnh bằng cách ra lệnh Data > Weight cases... như trong hình sau:

Hình 15.41



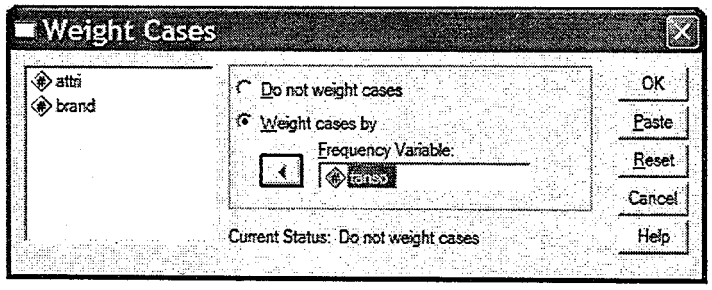
Lệnh này mở ra hộp thoại Weight Cases như sau:

Hình 15.42



Trong hộp thoại này, chọn mục Weight cases by rồi chọn biến tanso đưa vào ô Frequency variable như sau:

Hình 15.43



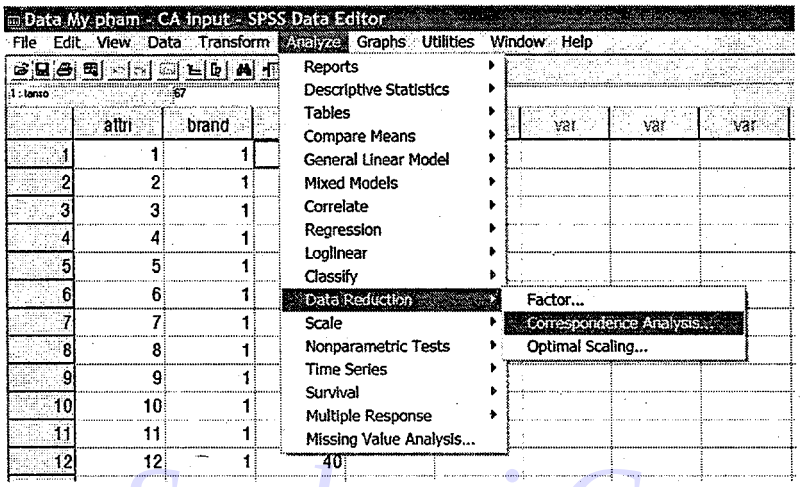
Sau đó nhấp nút OK.

(Trên đây chỉ là một cách đơn giản, nhưng hơi thủ công, để chuẩn bị dữ liệu thực hiện Correspondence Analysis. Trong thực tế còn một số cách thực hiện khác, nhưng phức tạp hơn)

BƯỚC 3: Thực hiện Correspondence Analysis và hiệu chỉnh bản đồ

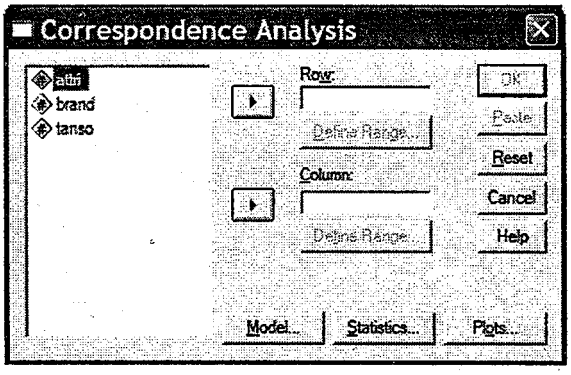
Bây giờ chúng ta sẽ chạy lệnh để lập bản đồ vị trí, từ menu chọn Amalyze> Data Reduction > Correspondence Analysis như sau:

Hình 15.44



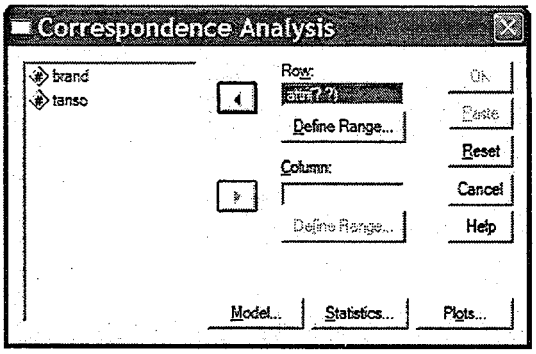
Lệnh này mở ra hộp thoại Correspondence Analysis sau:

Hình 15.45



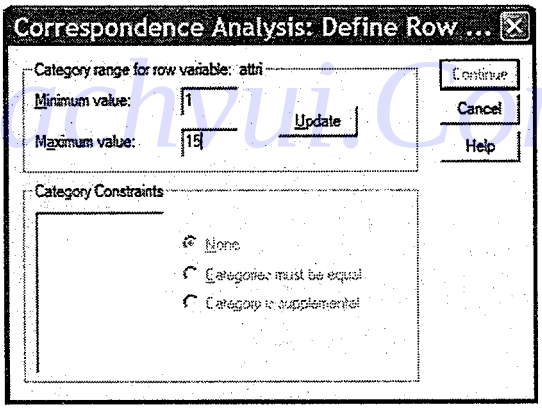
Đưa biến attri vào ô Rows, sẽ xuất hiện hai dấu ? và nút Define Range... nổi đậm lên.

Hình 15.46



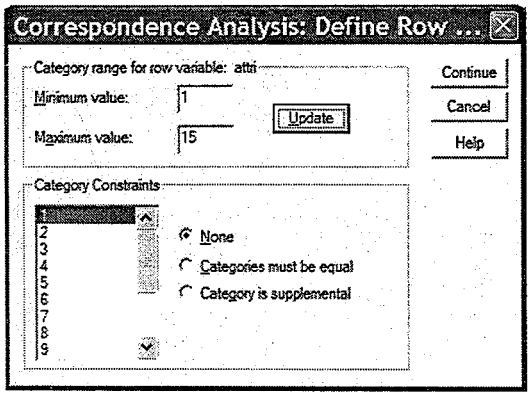
Nhấp chuột vào nút Define Range... để mở ra hộp thoại con và xác định số lượng thuộc tính muốn đưa vào phân tích bằng cách khai báo các giá trị mã của các thuộc tính nhỏ nhất và lớn nhất trong các ô Minimum và Maximum:

Hình 15.47



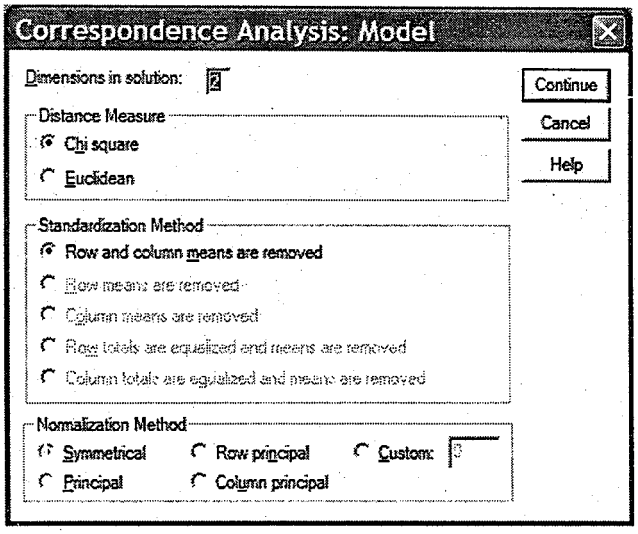
Sau đó nhấp nút Update

Hình 15.48



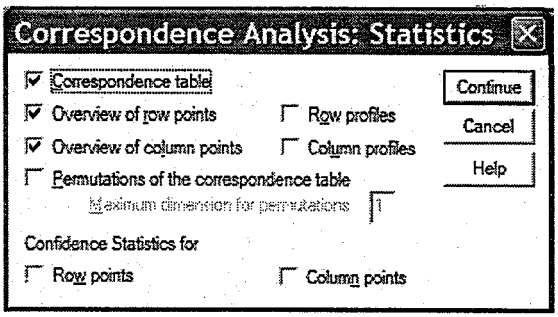
Rồi nhấp nút Continue trở lại hộp thoại Correspondence Analysis. Tiếp theo nhấp vào nút model và mở ra hộp thoại Model như sau:

Hình 15.49



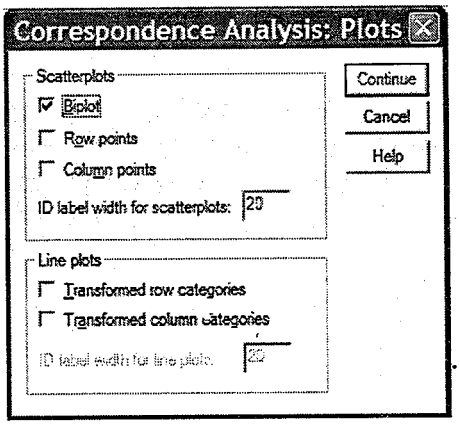
Trong hộp thoại model, số lượng chiều hướng trong bản đồ mặc định là 2 (vì đây là không gian được nhiều người sử dụng vì tính trực quan của nó), nếu cần thì có thể chỉnh lại thành 3. Trong ví dụ này chúng ta vẫn giữ nguyên. Rồi nhấp nút Continue trở về hộp thoại Correspondence Analysis. Trong hộp thoại này nhấp nút Statistics... mở ra hộp thoại sau:

Hình 15.50



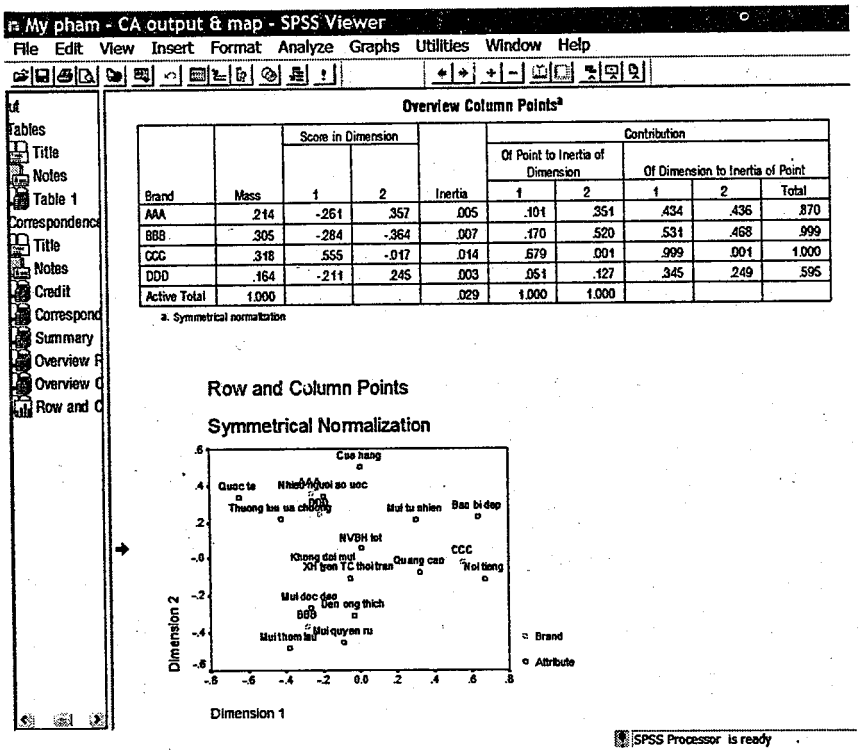
Trong hộp thoại Correspondence Analysis: Statistics có thể chọn thêm các kết quả thống kê trung gian như Row profiles (tỉ lệ dòng) và Column profiles (tỉ lệ cột). Sau đó nhấp nút Continue trở về hộp thoại Correspondence Analysis. Trong hộp thoại này nhấp nút Plots... và mở ra hộp thoại sau:

Hình 15.51



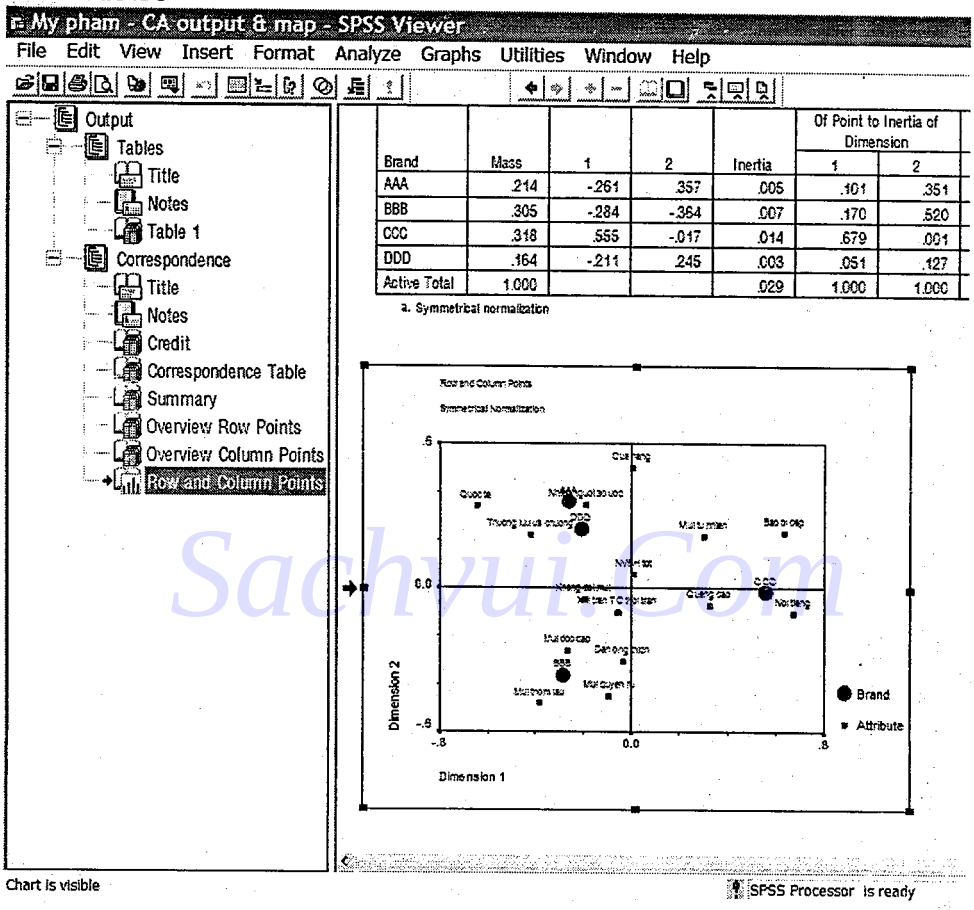
Trong hộp thoại Correspondence Analysis: Plots mặc định là chọn Biplot để vẽ đồ thị theo các cặp chiều hướng (nếu chọn 2 chiều hướng khi xác định Model thì chỉ có 1 bản đồ vị trí xuất hiện) và có thể chọn thêm Row points và Column points để vẽ bản đồ đơn hướng. Sau đó nhấp nút Continue trở về hộp thoại Correspondence Analysis, rồi nhấp nút OK, kết quả phân tích và bản đồ vị trí xuất hiện như sau:

Hình 15.52



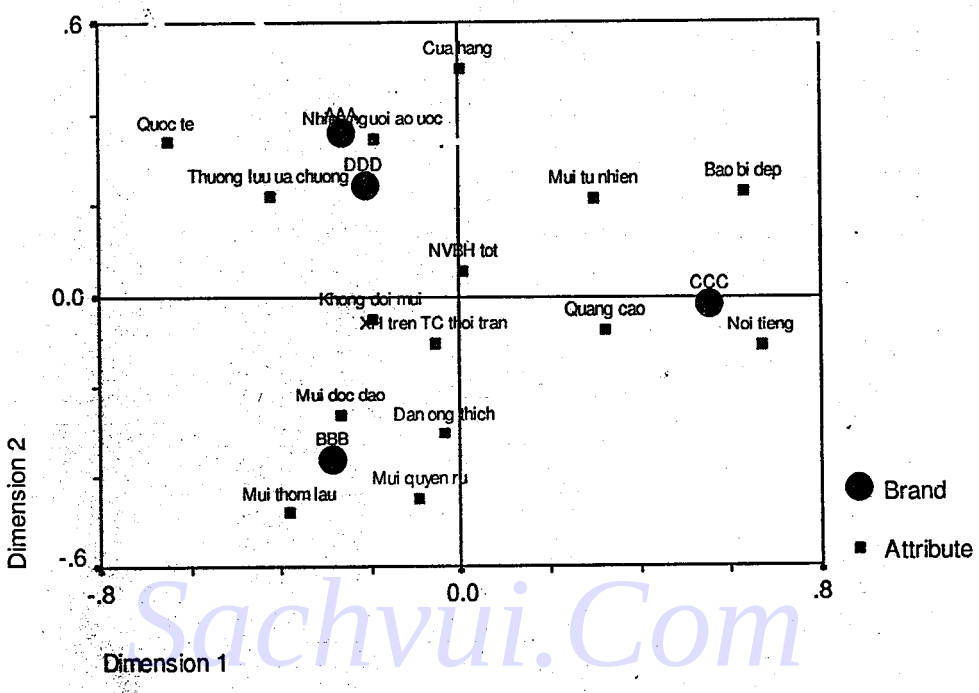
Để bản đồ vị trí dễ xem hơn, chúng ta nhấp nhanh chuột hai lần (Double click) vào bản đồ vị trí để chỉnh sửa Font chữ, màu sắc, cách quy ước các đối tượng trên bản đồ ... Sau một số thay đổi định dạng, bản đồ thể hiện vị trí các thương hiệu xuất hiện như trong hình dưới đây:

Hình 15.53



Sau khi điều chỉnh định dạng của bản đồ, có thể copy bản đồ vị trí này vào các chương trình Word hay PowerPoint. Trong các chương trình này bản đồ vị trí là 1 đối tượng (object), lúc này bạn chỉ có thể phóng to hay thu nhỏ lại để trình bày mà thôi.

Hình 15.54



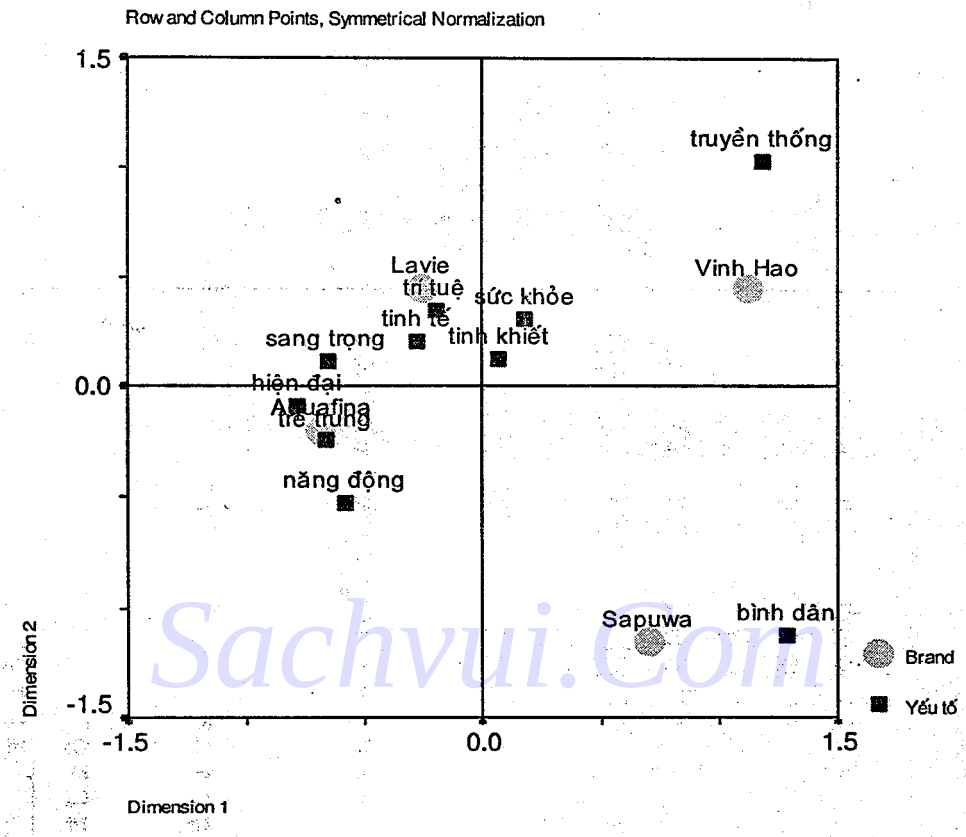
Để tiện thực hành thêm, bạn hãy dùng các dữ liệu đã được tổng hợp về cảm nhận của NTD (mẫu khảo sát là 100 người) về tính cách của các thương hiệu nước uống đóng chai (các con số trong bảng thể hiện số người đồng ý từng thương hiệu có thuộc tính tương ứng), hãy vẽ bản đồ hình ảnh các thương hiệu này và đọc hiểu ý nghĩa của bản đồ.

Bảng 15.3

		Brand				Group Total
		Lavie	Aquafina	Vinh Hao	Sapuwa	
Yếu tố	hiện đại	35	55	2	7	99
	truyền thống	24	6	53	5	88
	tinh tế	34	24	8	8	74
	bình dân	7	11	39	45	102
	sang trọng	30	43	6	4	83
	trẻ trung	24	50	5	8	87
	năng động	15	46	5	10	76
	sức khỏe	42	30	29	13	114
	trí tuệ	24	13	6	5	48
	tinh khiết	53	44	31	20	148
Group Total		288	322	184	125	919

Hình bên dưới là bản đồ kết quả để bạn đọc kiểm tra lại kết quả đã phân tích của mình.

Hình 15.55: Bản đồ thể hiện cảm nhận của NTD về cá tính của các NH nước uống đóng chai (dùng kỹ thuật CA)



CHƯƠNG XVI

CÁC TIỆN ÍCH (UTILITIES)

Trong thực tế, khi xử lý và phân tích, người nghiên cứu đôi khi có nhu cầu gia trọng các quan sát, ghép trộn file hay thay đổi cấu trúc của file dữ liệu ... Chương trình máy tính đã cung cấp tất cả các lệnh cần thiết để phục vụ các yêu cầu này của người phân tích. Chương này trình bày cá tiện ích này sau đó là các tiện ích về in ấn cũng như cài đặt.

1. GIA TRỌNG CÁC QUAN SÁT (Weighting cases)

Gia trọng các quan sát cần thiết khi cơ cấu mẫu khảo sát khác với cơ cấu của tổng thể nghiên cứu trong khi người phân tích cần kết quả tổng hợp chung của mẫu phản ảnh đúng đắn đặc điểm/tình hình của toàn bộ tổng thể.

Thử lấy 1 ví dụ đơn giản. Giả sử chúng ta coi Hà Nội và TPHCM (2 đô thị lớn nhất Việt Nam) là 1 tổng thể cần nghiên cứu về việc đọc báo của người dân ở đô thị lớn. Để cho ví dụ đơn giản, chúng ta coi như dân số Hà Nội là 3 triệu người và dân số TPHCM là 6 triệu người. Giả sử người nghiên cứu lấy ra 1 mẫu khảo sát 500 người, nếu chúng ta chỉ quan tâm đến kết quả khảo sát riêng của từng thành phố thì phân bổ mẫu vào 2 địa bàn khảo sát không quan trọng lắm, ví dụ như Hà Nội 250 người, TPHCM 250 người (phân bổ đều); hay là Hà Nội 167 người, TPHCM 333 người (phân bổ theo tỉ lệ tam suất). Nếu chúng ta chọn cách phân bổ đều mỗi thành phố là 250 người thì kết quả khảo sát tổng hợp cho từng thành phố không có gì bàn cãi, tuy nhiên kết quả tổng hợp chung cho cả hai thành phố sẽ bị lệch lạc. Nếu chúng ta chọn cách phân bổ theo tỉ lệ thì kết quả tổng hợp chung cho hai thành phố từ dữ liệu khảo sát sẽ phản ánh đúng đắn tình hình chung của hai thành phố. Tuy nhiên, lúc này quy mô mẫu khảo sát ở Hà Nội hơi nhỏ, nên nếu cần phân tích sâu thêm riêng cho Hà Nội thì sẽ gặp khó khăn. Để khắc phục điểm yếu của từng cách xách định tỉ lệ mẫu, thông thường người nghiên cứu sẽ phân bổ mẫu đều (trong trường hợp quy mô toàn bộ mẫu không lớn do giới hạn bởi thời gian và ngân sách) hay bảo đảm quy mô tối

thiểu của từng địa bàn khảo sát và có chênh lệch giữa các địa bàn ví dụ như mẫu 200 tại Hà Nội và mẫu 300 tại TPHCM. Sau đó khi tổng hợp và xử lý, người phân tích sẽ gia trọng các quan sát sao cho cơ cấu của mẫu khảo sát giống với cơ cấu của tổng thể nghiên cứu.

Trong file dữ liệu Data thuc hanh.sav, chúng ta đã thấy mẫu khảo sát thực tế ở Hà Nội là 250 và ở TPHCM là 250, trong khi dân số ở TPHCM gấp đôi ở Hà Nội, 6 triệu so với 3 triệu dân (giả sử số người thương xuyên đọc báo ở TPHCM cũng gấp đôi số người thường xuyên đọc báo ở Hà Nội). Để tính ra kết quả chung không bị lệch lạc do cách phân bổ mẫu đều này, chúng ta cần gán trọng số khác nhau cho các quan sát ở Hà Nội và ở TPHCM.

Công thức tính trọng số chung là [1]

$$w_i = \frac{\text{Tỉ trọng của nhóm trong tổng thể}}{\text{Tỉ trọng của nhóm trong mẫu}} = \frac{N_i/N}{n_i/n}$$

N: quy mô tổng thể nghiên cứu; trong ví dụ này là 3+6 = 9 triệu dân

N_i: quy mô của từng nhóm cần gia trọng; trong ví dụ này là N_1 ở Hà Nội là 3 triệu dân, N_2 ở TPHCM là 6 triệu dân.

n: quy mô mẫu khảo sát; trong ví dụ này là 500 người

n_i: quy mô mẫu khảo sát của từng nhóm; trong ví dụ này là n_1 ở Hà Nội là 250 người, n_2 ở TPHCM là 250 người.

Trọng số của Hà Nội sẽ là:

$$w_1 = \frac{N_1/N}{n_1/n} = \frac{3.000.000/9.000.000}{250/500} = \frac{1/3}{1/2} = 2/3 = 0,67$$

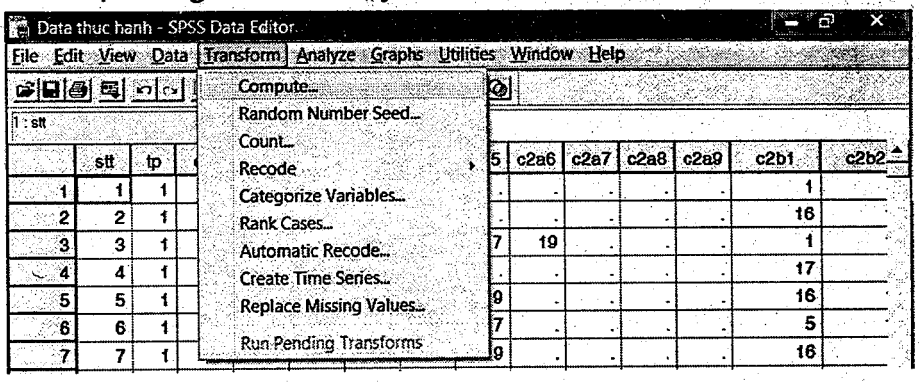
Trọng số của TPHCM sẽ là:

$$w_1 = \frac{N_2/N}{n_2/n} = \frac{6.000.000/9.000.000}{250/500} = \frac{2/3}{1/2} = 4/3 = 1,33$$

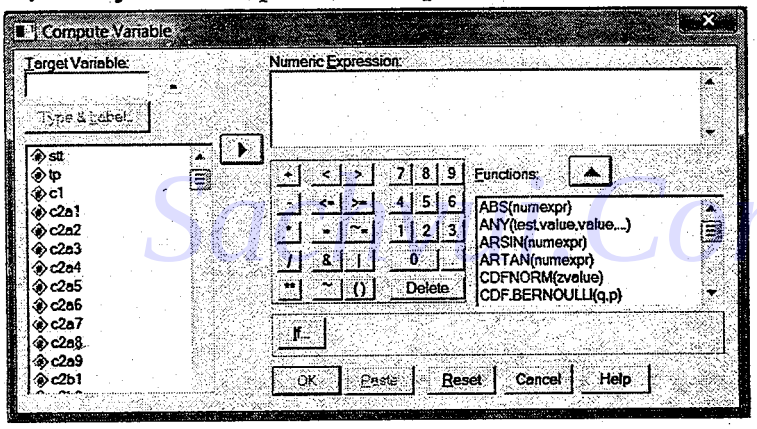
[1] Xem thêm ở Chương 2 sách Thống Kê Ứng Dụng trong Kinh Tế Xã Hội cùng tác giả.

Sau khi tính được các trọng số theo công thức tổng quát trên, chúng ta sẽ dùng lệnh để gán các trọng số này vào các quan sát trong file dữ liệu.

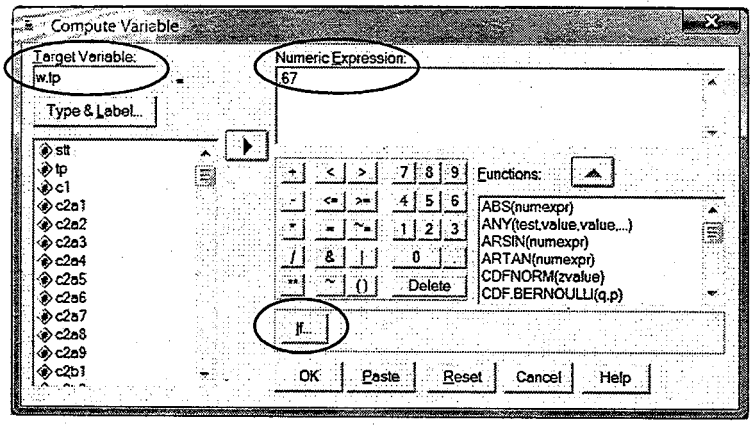
Mở file Data thuc hanh.sav, chọn lệnh Transform > Compute... như minh họa trong hình dưới đây:



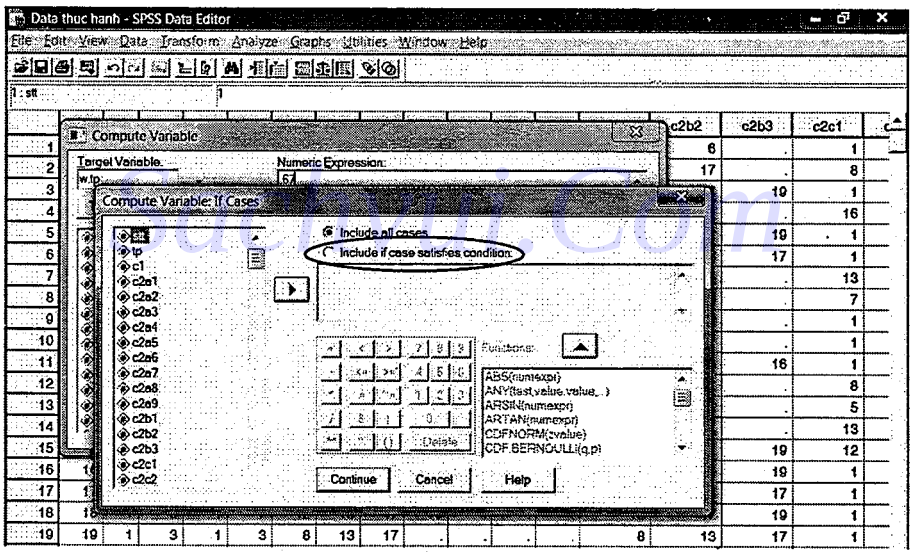
Lệnh này mở ra hộp thoại Compute variable sau:



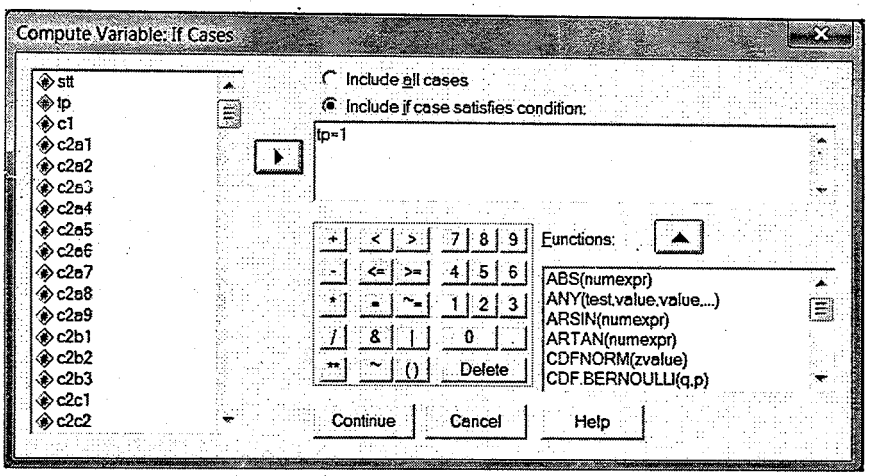
Gõ tên biến trọng số cần tạo là **w.tp** vào ô Target Variable, và ghi trị số của trọng số cho Hà Nội là 0.67 vào ô Numberic Expression. Chúng ta chỉ gán trọng số 0.67 này cho các trường hợp thỏa điều kiện là quan sát ở Hà Nội, do đó hãy chọn nút If...



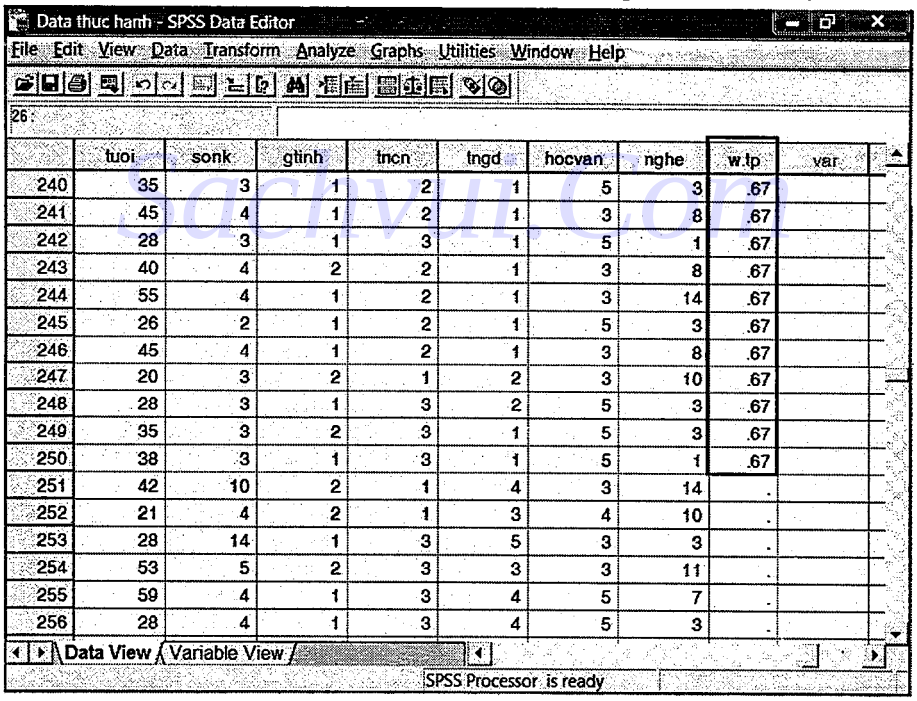
Khi chọn nút If... thì hộp thoại con If Cases sẽ xuất hiện như hình dưới:



Trong hộp thoại If cases, hãy chọn điều kiện để gán trọng số này bằng cách chọn Include if case satisfies condition. Sau đó chọn biến **tp** (thành phố) trong danh sách biến bên tay trái rồi bấm nút mũi tên qua phải đưa biến tp vào khung chứa biến bên tay phải. Sau đó gõ dấu = và trị số 1 (mã số của Hà Nội) như hình bên dưới.

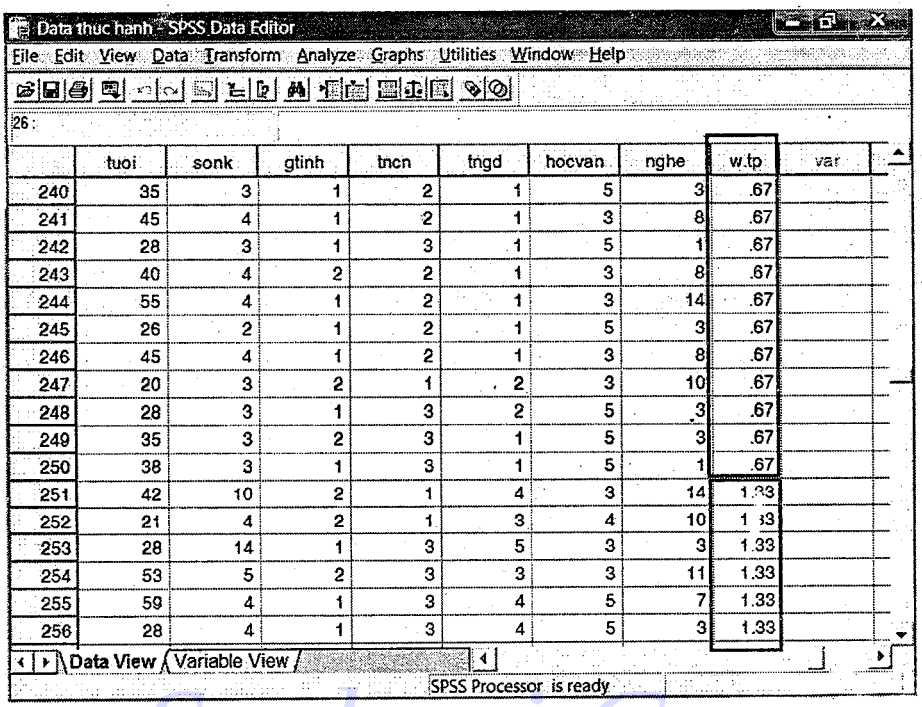


Tiếp theo nhấn nút Continue trong hộp thoại If cases để trở ra hộp thoại Compute variable, và nhấn tiếp nút OK. Lệnh sẽ được thực hiện và biến mới w.tp có trị số 0.67 tại tất cả những dòng nào thỏa điều kiện tp=1 (là thành phố Hà Nội) như trong hình minh họa:



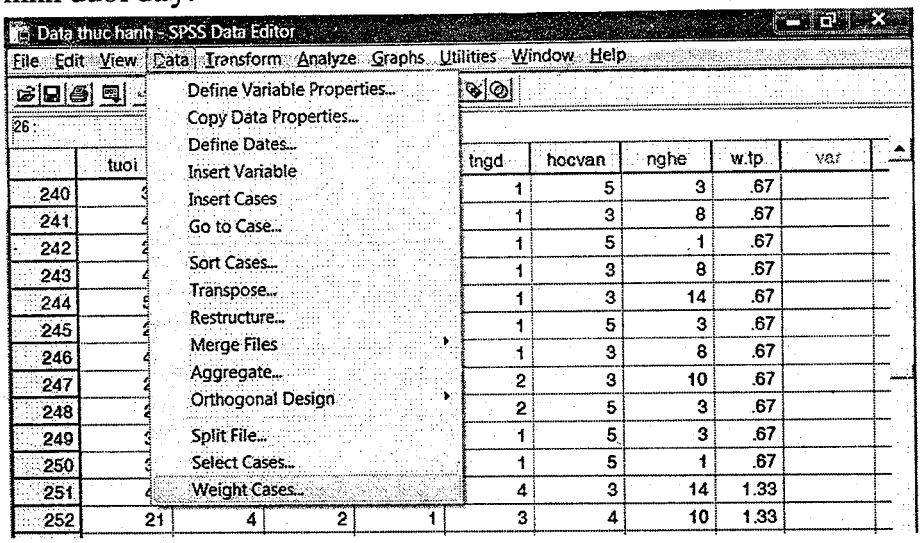
	tuoi	sonk	gtinh	tncn	tngd	hocvan	nghe	w.tp	var
240	35	3	1	2	1	5	3	.67	
241	45	4	1	2	1	3	8	.67	
242	28	3	1	3	1	5	1	.67	
243	40	4	2	2	1	3	8	.67	
244	55	4	1	2	1	3	14	.67	
245	26	2	1	2	1	5	3	.67	
246	45	4	1	2	1	3	8	.67	
247	20	3	2	1	2	3	10	.67	
248	28	3	1	3	2	5	3	.67	
249	35	3	2	3	1	5	3	.67	
250	38	3	1	3	1	5	1	.67	
251	42	10	2	1	4	3	14	.	
252	21	4	2	1	3	4	10	.	
253	28	14	1	3	5	3	3	.	
254	53	5	2	3	3	3	11	.	
255	59	4	1	3	4	5	7	.	
256	28	4	1	3	4	5	3	.	

Tiếp tục, chúng ta sẽ gán trọng số 1.33 cho các dòng dữ liệu nào thỏa điều kiện tp=2 tương tự như trên. Kết quả chúng ta đã có biến trọng số w.tp có đủ hai trị số ứng với 2 thành phố đã tính.

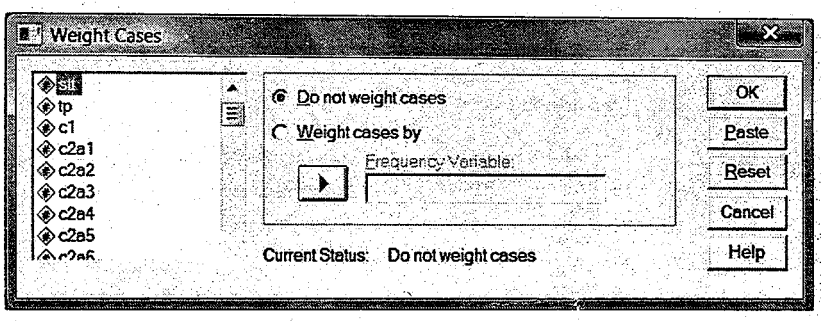


	tuoi	sonk	gtinh	tncn	tngd	hocvan	nghe	w.tp	var
240	35	3	1	2	1	5	3	.67	
241	45	4	1	2	1	3	8	.67	
242	28	3	1	3	1	5	1	.67	
243	40	4	2	2	1	3	8	.67	
244	55	4	1	2	1	3	14	.67	
245	26	2	1	2	1	5	3	.67	
246	45	4	1	2	1	3	8	.67	
247	20	3	2	1	2	3	10	.67	
248	28	3	1	3	2	5	3	.67	
249	35	3	2	3	1	5	3	.67	
250	38	3	1	3	1	5	1	.67	
251	42	10	2	1	4	3	14	1.33	
252	21	4	2	1	3	4	10	1.33	
253	28	14	1	3	5	3	3	1.33	
254	53	5	2	3	3	3	11	1.33	
255	59	4	1	3	4	5	7	1.33	
256	28	4	1	3	4	5	3	1.33	

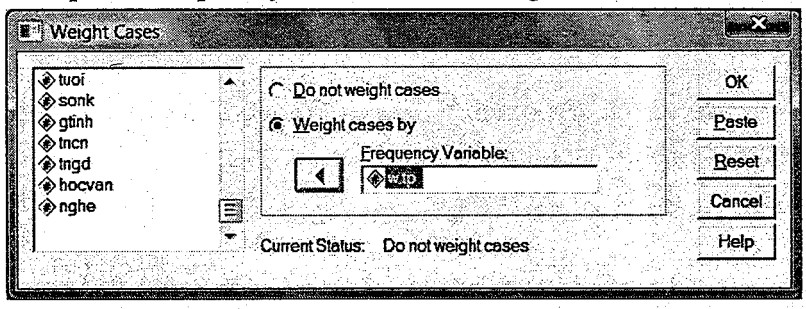
Sau khi tạo được biến trọng số w.tp, chúng ta phải báo cho chương trình biết là các kết quả sắp xử lý tiếp theo phải được gia trọng theo trọng số này. Chọn lệnh Data > Weight cases như minh họa trong hình dưới đây:



Lệnh này mở ra hộp thoại Weight cases như sau:



Trong hộp thoại này, nhấn chọn Weight cases by, rồi dùng chuột nhấp chọn biến w.tp ở cuối danh sách biến bên tay trái của hộp thoại đưa qua ô Frequency Variable như trong hình dưới:



Sau đó nhấp nút OK trên hộp thoại để lệnh thực hiện. Sau khi chạy lệnh, chương trình sẽ hiện lưu ý trên dòng trạng thái ở phía dưới cuối góc phải là Weight On.

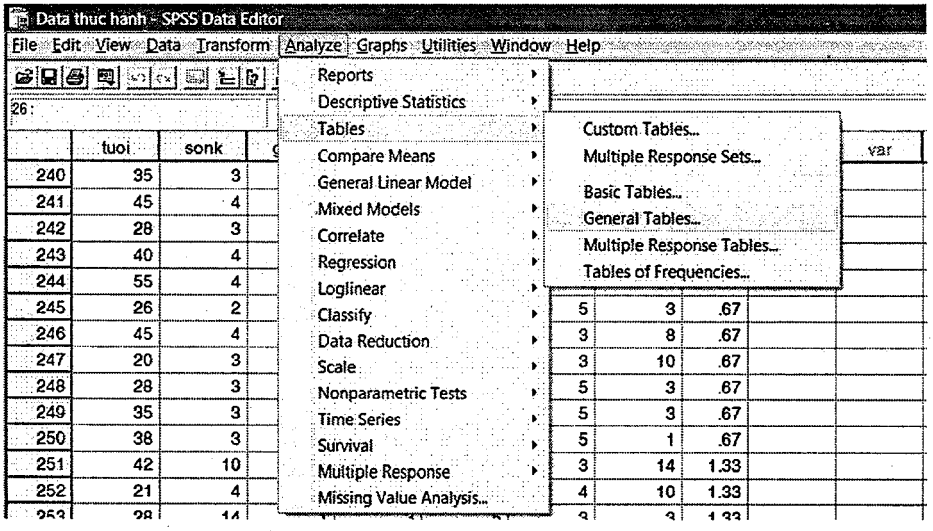
251	42	10	2	1	4	3	14	1.33				
252	21	4	2	1	3	4	10	1.33				
253	28	14	1	3	5	3	3	1.33				
254	53	5	2	3	3	3	11	1.33				
255	59	4	1	3	4	5	7	1.33				
256	28	4	1	3	4	5	3	1.33				
257	22	3	1	3	2	3	6	1.33				
258	40	4	1	3	3	5	1	1.33				
259	50	6	1	3	4	5	1	1.33				
260	54	7	1	3	4	3	6	1.33				
261	22	4	2	3	4	3	3	1.33				
262	18	4	2	1	3	4	10	1.33				
263	29	5	1	3	3	4	1	1.33				

Data View | Variable View

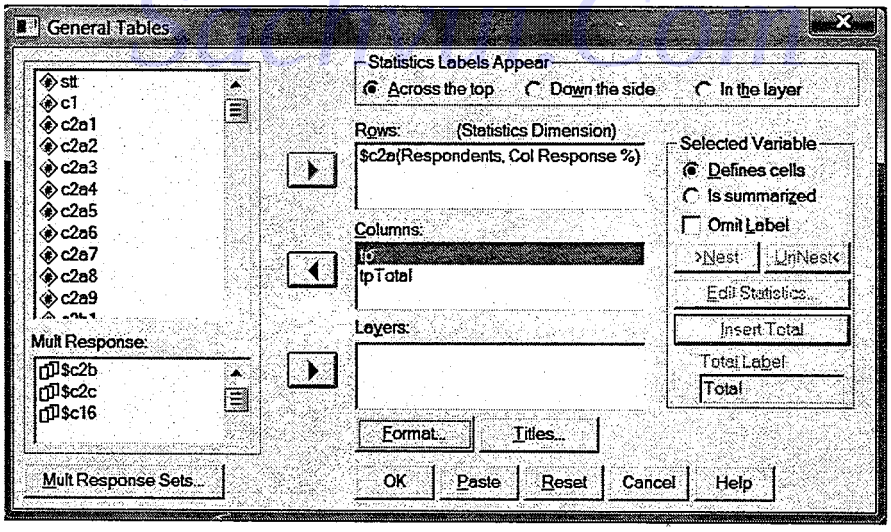
SPSS Processor is ready — Weight On

Tới đây chúng ta đã có dữ liệu đã gia trọng sẵn sàng cho việc tổng hợp kết quả chung cho 2 thành phố.

Để xem thử hiệu quả của lệnh Weighting cases, chúng ta tính tỉ lệ đọc từng loại báo của từng thành phố và tính chung cho 2 thành phố bằng cách dùng General Tables:



Chọn biến tp đưa vào ô Columns và Insert Total, chọn biến ghép $c2a trong khung Mult Responsese đưa vào ô Rows, sau đó nhấp vào nút Edit Statistics để chọn hàm thống kê tính % theo cột như hình dưới:



Lệnh này cho chúng ta kết quả thống kê bên dưới. Trong đó tỉ lệ đọc báo Tuổi Trẻ chung là 58,3%.

		thành phố				Total	
		Hà Nội		TPHCM			
		Cases	Col Response %	Cases	Col Response %	Cases	Col Response %
BÁO THƯỜNG ĐỌC	HN mới	81	48.4%	1	.4%	82	16.5%
	SGGP	7	4.4%	78	23.6%	86	17.2%
	Lao Động	44	26.0%	19	5.6%	62	12.4%
	Người Lao Động	8	4.8%	82	24.8%	90	18.1%
	Tiền Phong	58	34.8%	7	2.0%	65	13.0%
	Thanh Niên	19	11.2%	76	22.8%	95	18.9%
	Tuổi Trẻ	29	17.6%	262	78.8%	291	58.3%
	Phụ Nữ VN	65	38.8%	31	9.2%	96	19.1%
	Phụ Nữ TPHCM	3	2.0%	102	30.8%	106	21.2%
	Thời Báo KTVN	19	11.6%	8	2.4%	27	5.5%
	Thời Báo KTSG	7	4.0%	15	4.4%	21	4.3%
	SG Tiếp Thị	39	23.2%	134	40.4%	173	34.6%
	Thế Giới Phụ Nữ	54	32.0%	112	33.6%	165	33.1%
	Tiếp Thị và GĐ	19	11.2%	43	12.8%	61	12.3%
	Mua & Bán	31	18.8%	15	4.4%	46	9.2%
	An Ninh Thế Giới	129	77.2%	141	42.4%	270	54.1%
	An Ninh Thủ Đô	123	73.2%	11	3.2%	133	26.7%
	Công An TPHCM	29	17.6%	250	75.2%	280	55.9%
	Khác	83	49.6%	98	29.6%	182	36.3%

Nếu chúng ta không gia trọng các quan sát, thì kết quả như hình dưới. Trong hình này chúng ta thấy tỉ lệ đọc báo Tuổi Trẻ chung 2 thành phố là 48,2%, thấp hơn khi so với có gia trọng. Đó là do Tỉ lệ đọc báo Tuổi Trẻ không cao ở Hà Nội, mà Hà Nội lại chiếm tới 50% số lượng mẫu trong khi thực tế dân số và số người đọc báo tại Hà Nội chỉ bằng 1/2 của TPHCM (giả sử gần đúng) làm giảm tỉ lệ tỉ lệ đọc báo Tuổi Trẻ chung của hai thành phố.

		thành phố				Total	
		Hà Nội		TPHCM			
		Cases	Col Response %	Cases	Col Response %	Cases	Col Response %
BÁO THƯỜNG ĐỌC	HN mới	121	48.4%	1	.4%	122	24.4%
	SGGP	11	4.4%	59	23.6%	70	14.0%
	Lao Động	65	26.0%	14	5.6%	79	15.8%
	Người Lao Động	12	4.8%	62	24.8%	74	14.8%
	Tiền Phong	87	34.8%	5	2.0%	92	18.4%
	Thanh Niên	28	11.2%	57	22.8%	85	17.0%
	Tuổi Trẻ	44	17.6%	197	78.8%	241	48.2%
	Phụ Nữ VN	97	38.8%	23	9.2%	120	24.0%
	Phụ Nữ TPHCM	5	2.0%	77	30.8%	82	16.4%
	Thời Báo KTVN	29	11.6%	6	2.4%	35	7.0%
	Thời Báo KTSG	10	4.0%	11	4.4%	21	4.2%
	SG Tiếp Thị	58	23.2%	101	40.4%	159	31.8%
	Thế Giới Phụ Nữ	80	32.0%	84	33.6%	164	32.8%
	Tiếp Thị và GĐ	28	11.2%	32	12.8%	60	12.0%
	Mua & Bán	47	18.8%	11	4.4%	58	11.6%
	An Ninh Thế Giới	193	77.2%	106	42.4%	299	59.8%
	An Ninh Thủ Đô	183	73.2%	8	3.2%	191	38.2%
	Công An TPHCM	44	17.6%	188	75.2%	232	46.4%
	Khác	124	49.6%	74	29.6%	198	39.6%

Chú ý rằng sau khi gia trọng để phân tích theo từng thành phố và chung cho 2 thành phố, khi không cần gia trọng nữa, hãy vào lại menu chọn Data > Weight cases và chọn tiếp Do not Weight cases để trở lại trạng thái bình thường (tức là không gia trọng nữa) lúc đó trên dòng trạng thái sẽ mất lưu ý Weight On.

2. THAY ĐỔI CẤU TRÚC DỮ LIỆU (Restructure Data)

Thay đổi cấu trúc của dữ liệu để có thể thực hiện được một số xử lý và phân tích. Thông thường trong thực tế khi người nghiên cứu cần lập bảng kết hợp (bảng chéo) giữa 2 biến mà cả 2 biến này đều là biến ghép (Multiple response) thì cần phải thay đổi cấu trúc dữ liệu.

Trở lại ví dụ trong tập 1 về Khảo sát người đọc báo, trong bảng câu hỏi ở phần Phụ lục, ở câu 2c và 2d, chúng ta có câu hỏi về các tờ báo thường mua và cách mua các tờ báo thường mua như thế nào. Khi xử lý, chúng ta đã ghép các biến của câu 2c lại với nhau thành

biến ghép $c2c và ghép các biến của câu 2d lại với nhau thành biến ghép $c2d. Khi chúng ta cần biết từng tờ báo được mua theo những cách nào. Nếu lập bảng kết hợp trực tiếp giữa 2 biến ghép $c2c và $c2d thì kết quả sẽ bị sai vì với cấu trúc hiện tại của file dữ liệu, máy tính không thể thực hiện lệnh count phức tạp được. Trong trường hợp này chúng ta sẽ chuyển data hiện có từ nhiều cột biến cho câu 2c và nhiều cột biến cho câu 2d về dạng data chỉ có 1 cột biến duy nhất cho 2c và 1 cột biến cho 2d, rồi sau đó lập bảng kết hợp giữa 2 biến này một cách bình thường. Để các bạn dễ hình dung, chúng ta xem sơ đồ trước khi thực hiện chạy lệnh. Chúng ta đang có 6 biến c2c1 đến c2c6 (báo thường mua) và 6 biến c2d1 đến c2d6 (cách mua báo), tổng cộng là 12 cột, mỗi cột 500 ô dữ liệu.

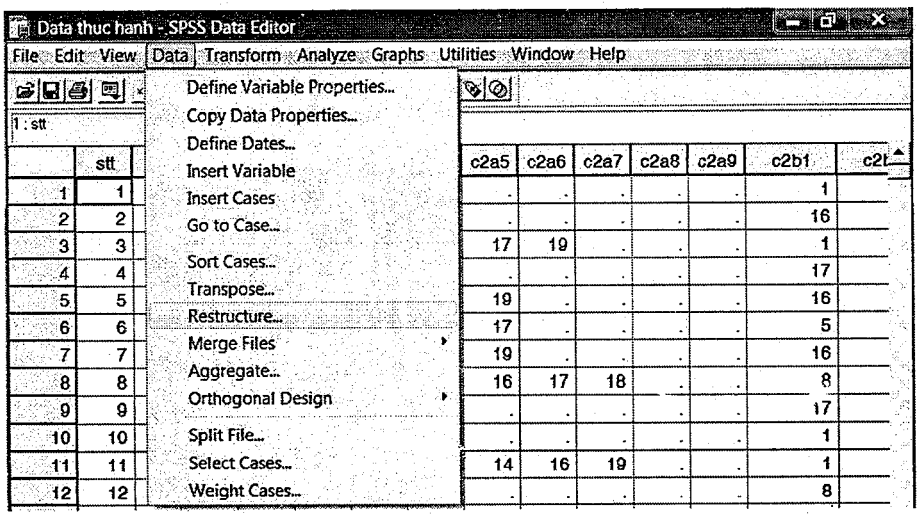
c2c1	c2c2	c2c3	c2c4	c2c5	c2c6	c2d1	c2d2	c2d3	c2d4	c2d5	c2d6
1	2	3	4	5	6	7	8	9	10	11	12

c2c1	c2d1
1	7
2	8
3	9
4	10
5	11
6	12

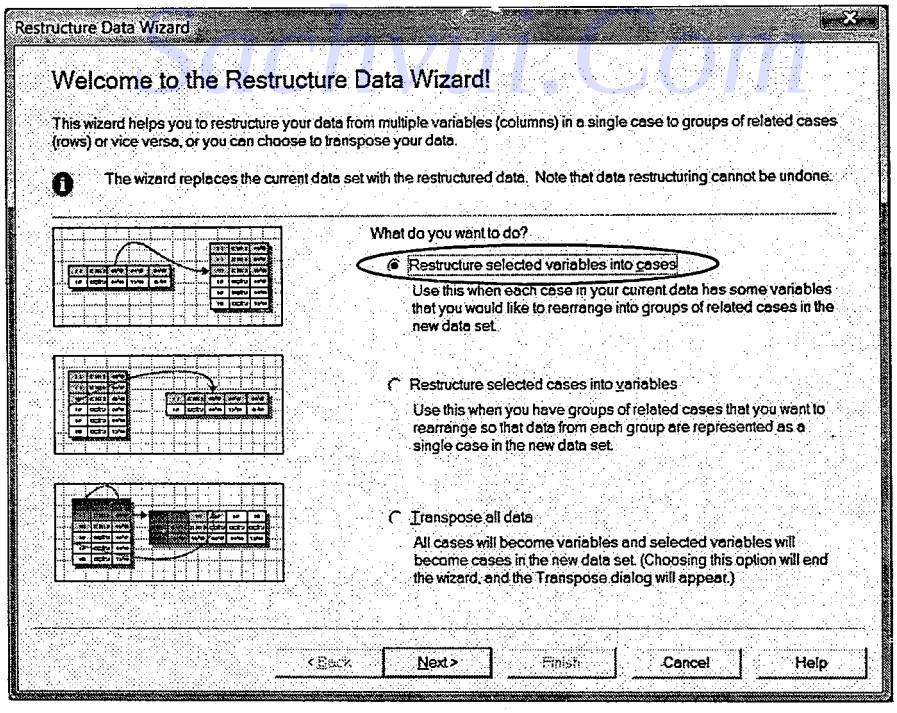
Chúng ta sẽ đem 500 ô dữ liệu của biến c2c xuống nối vào 500 ô dữ liệu của c2c1, 500 ô dữ liệu của c2c3 nối tiếp theo, ... cứ như vậy chúng ta chỉ còn 1 cột biến c2c bây giờ lên tới 500 x6 = 3000 ô (3000 dòng).

Làm tương tự như vậy đối với các dữ liệu của 6 biến c2d, chúng ta cũng sẽ có 1 cột biến c2d như yêu cầu. Trong hình kế bên, chúng ta tưởng tượng mỗi một khối đánh số từ 1 đến 12 gồm có 500 ô dữ liệu thì sẽ thấy 6000 ô dữ liệu được chuyển từ cấu trúc 12 cột biến x 500 dòng về thành cấu trúc 2 cột biến x 3000

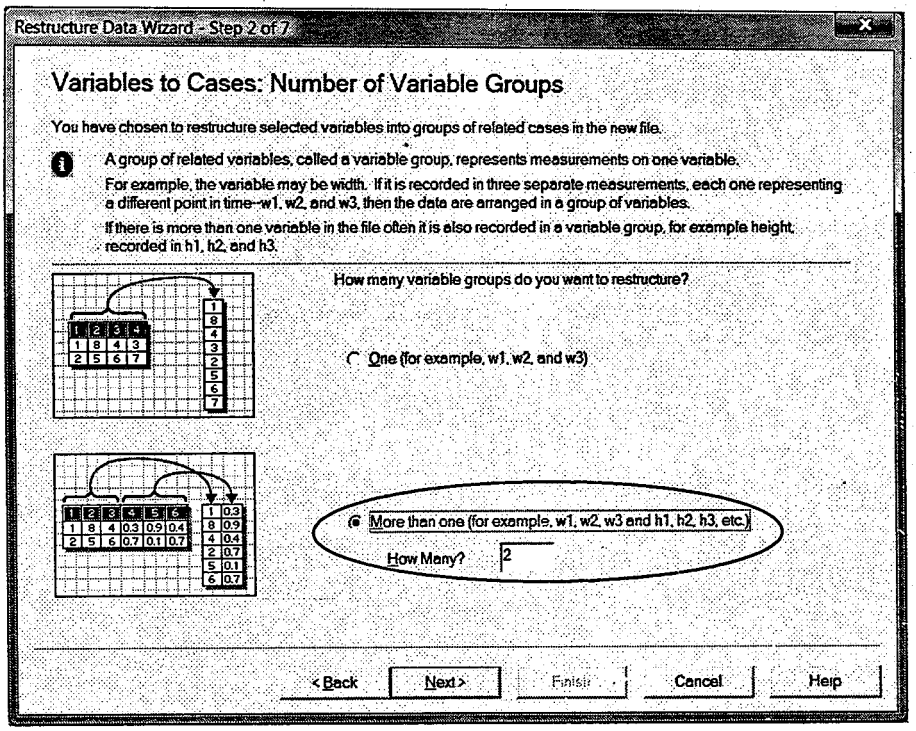
Để thực hành việc chạy lệnh, hãy mở file Data thuc hanh.sav. Từ Menu chọn Data > Restructure như hình sau đây:



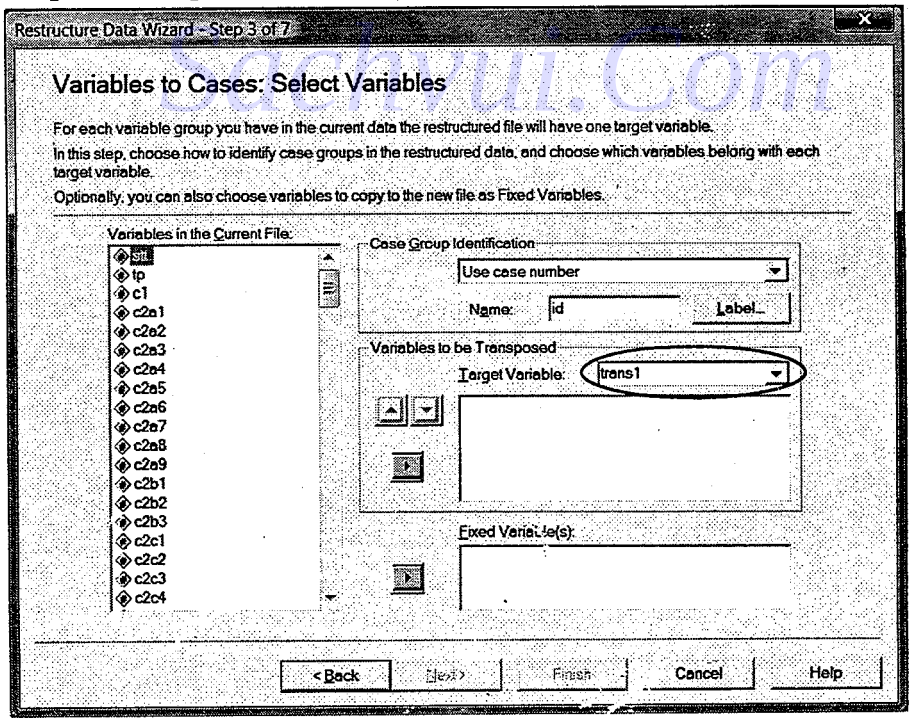
Khi chọn mục Restructure, hộp thoại Restructure Data Wizard của trình hướng dẫn xuất hiện để giúp bạn. Trong hộp thoại này, chọn mục đầu tiên là Restructure selected variables into cases như hình dưới.



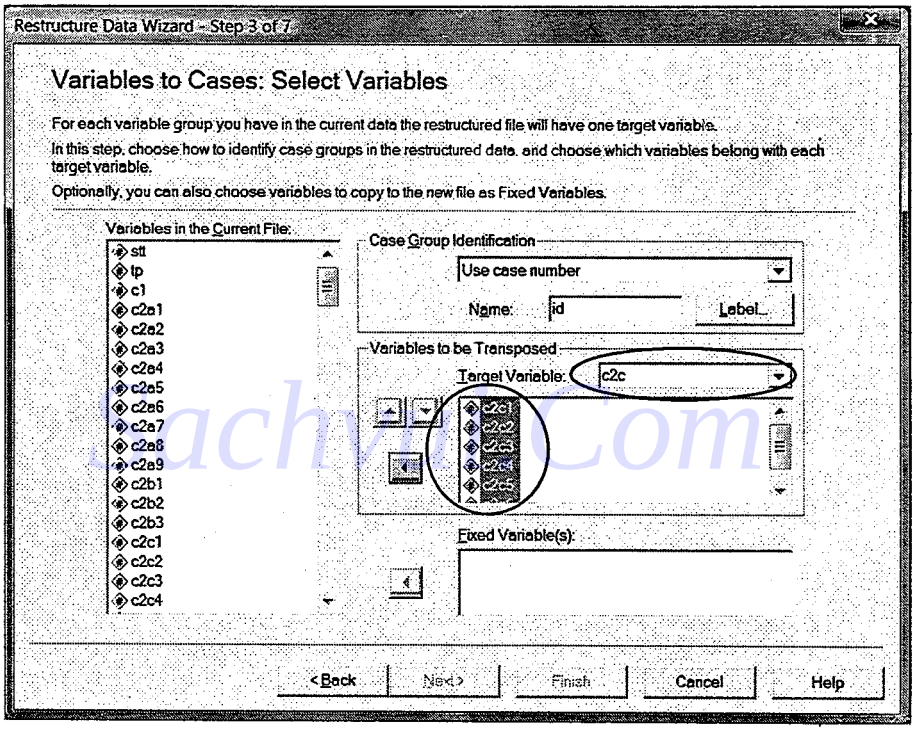
Sau đó nhấp vào nút Next để chuyển qua bước 2. Trong hộp thoại hướng dẫn ở bước 2, hãy chọn mục thứ hai là More than one (vì chúng ta sẽ cấu trúc lại dữ liệu của 2 tập biến c2c và c2d) và khai báo là 2 trong ô kế bên mục How Many?



Tiếp theo nhấp vào nút Next để chuyển qua bước 3 có màn hình sau

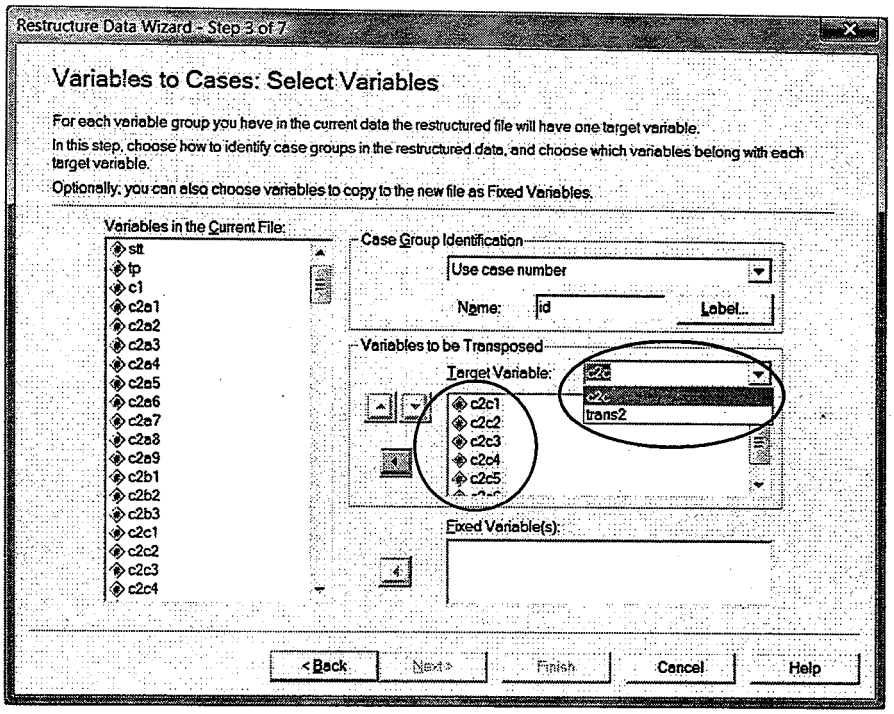


Trong bước 3 này, tại ô Target Variable, có sẵn 2 biến kết quả chúng ta muốn tạo ra có tên mặc định là Trans1 và Trans2. Hãy chọn Trans1 và gõ tên mới theo quy ước của chúng ta là c2c, rồi trong trong sách biến trong khung Variables in the Current File bên tay trái, hãy chọn đủ 6 biến c2c1, c2c2, c2c3, c2c4, c2c5, c2c6 và nhấp vào nút mũi tên qua phải để đưa các biến này vào khung bên dưới Target Variable như mô tả trong hình dưới.

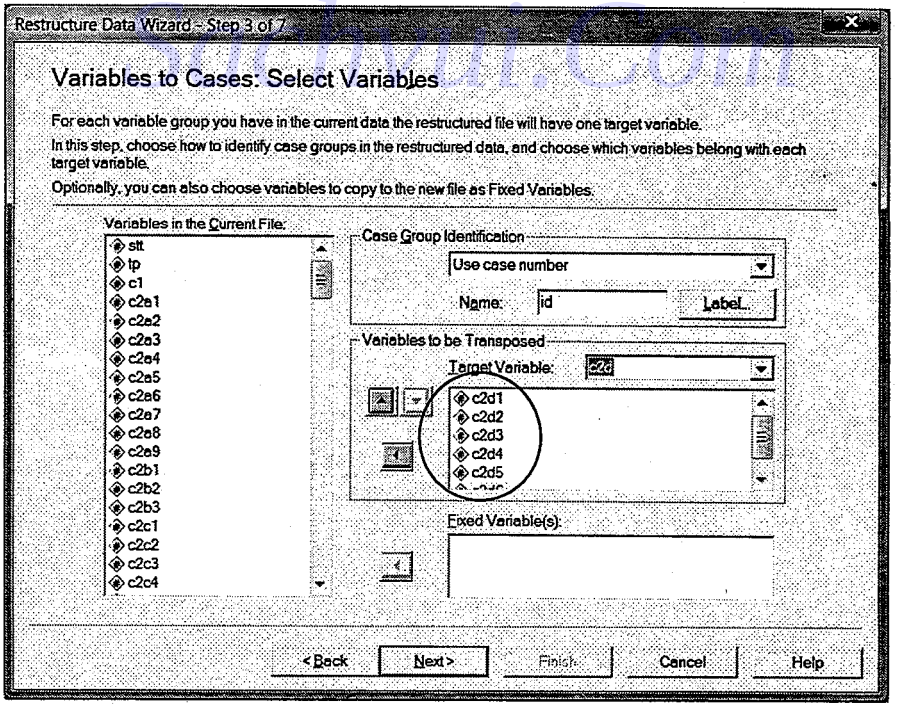


Như vậy chúng ta đã báo cho chương trình biết là hãy nối các dữ liệu của 6 biến trong khung (c2c1, c2c2, c2c3, c2c4, c2c5, c2c6) thành 1 biến mới có tên là c2c.

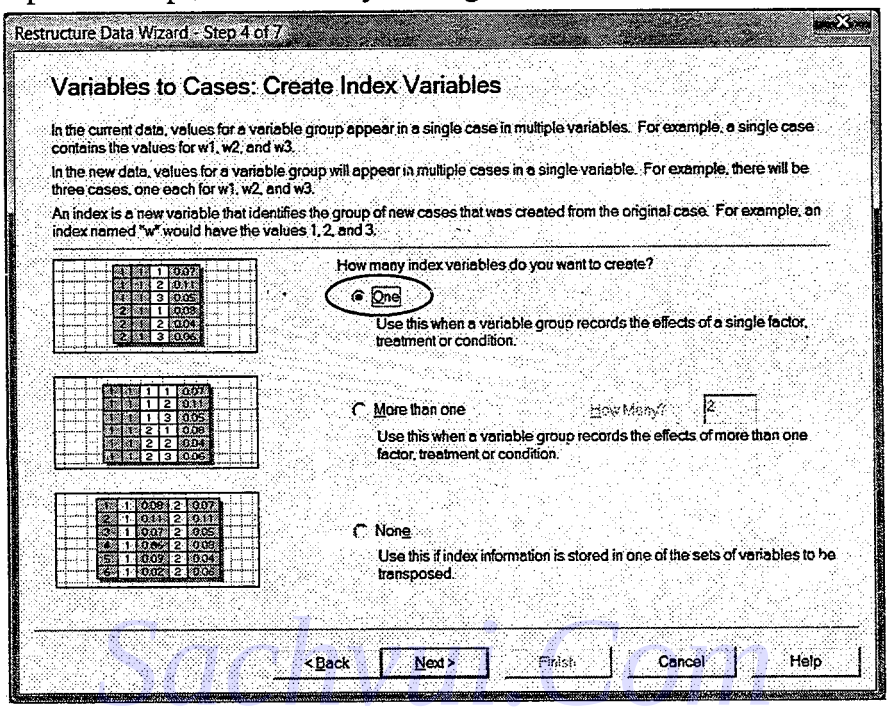
Tiếp theo, cũng trong bước này, làm tương tự cho tập biến thứ nhì. Bạn hãy nhấp vào nút mũi tên cuối ô chứa c2c để hiện ra danh sách các biến kết quả thì bạn sẽ thấy tên biến mặc định trans2 xuất hiện như hình sau:



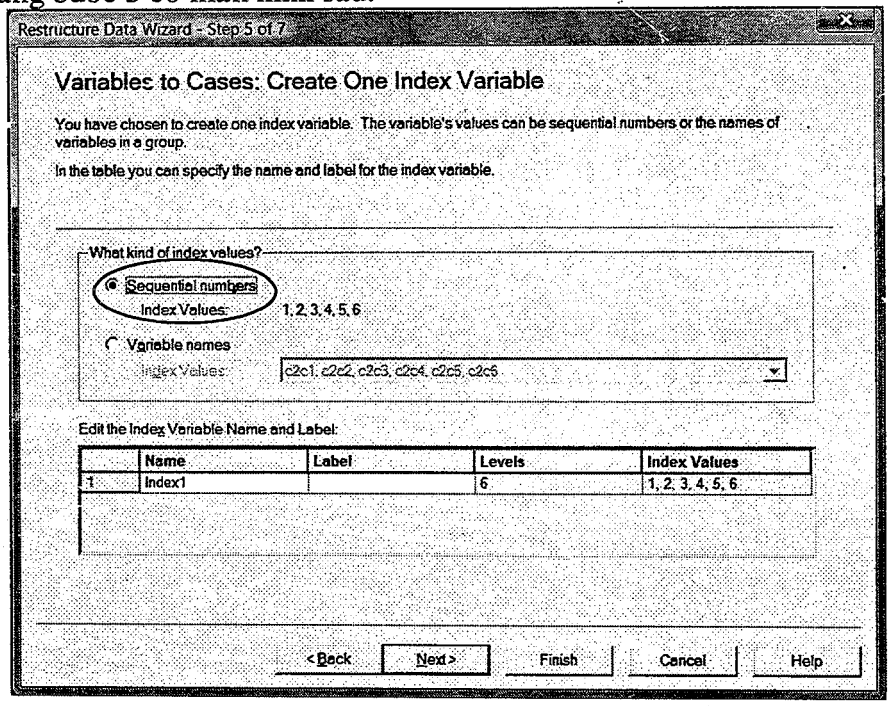
Nhấp vào tên Trans2 gõ vào tên biến mới là c2d, làm tương tự đối với các biến c2d1, c2d2, c2d3, c2d4, c2d5, c2d6 như trong hình sau



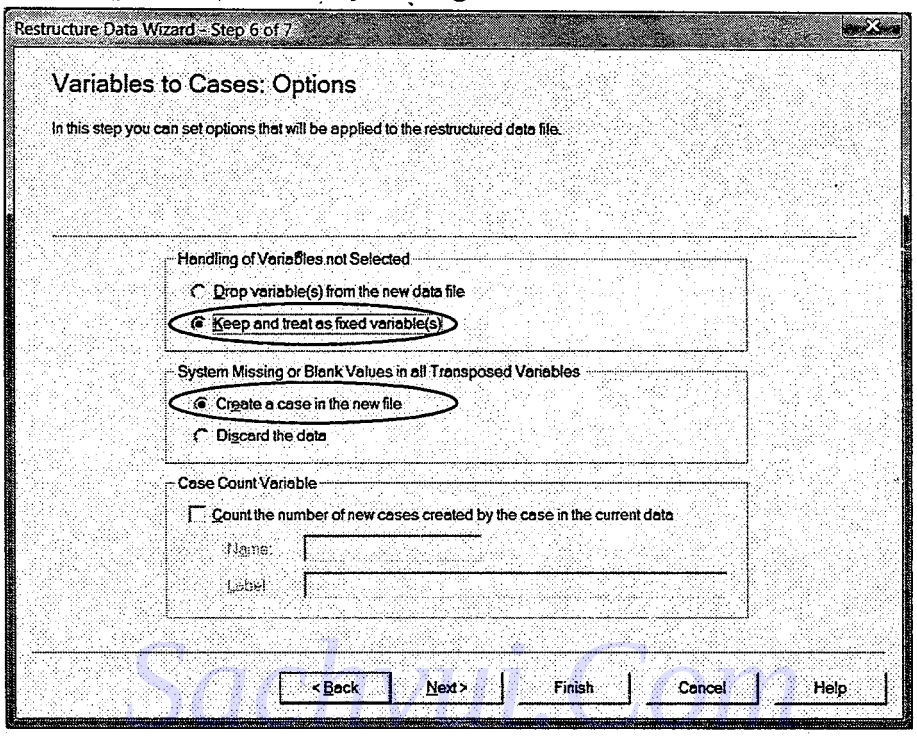
Tiếp theo nhấp nút Next chuyển sang bước 4 có màn hình sau:



Để mặc định là tạo ra 1 biến chỉ mục rồi nhấp tiếp nút Next chuyển sang bước 5 có màn hình sau.



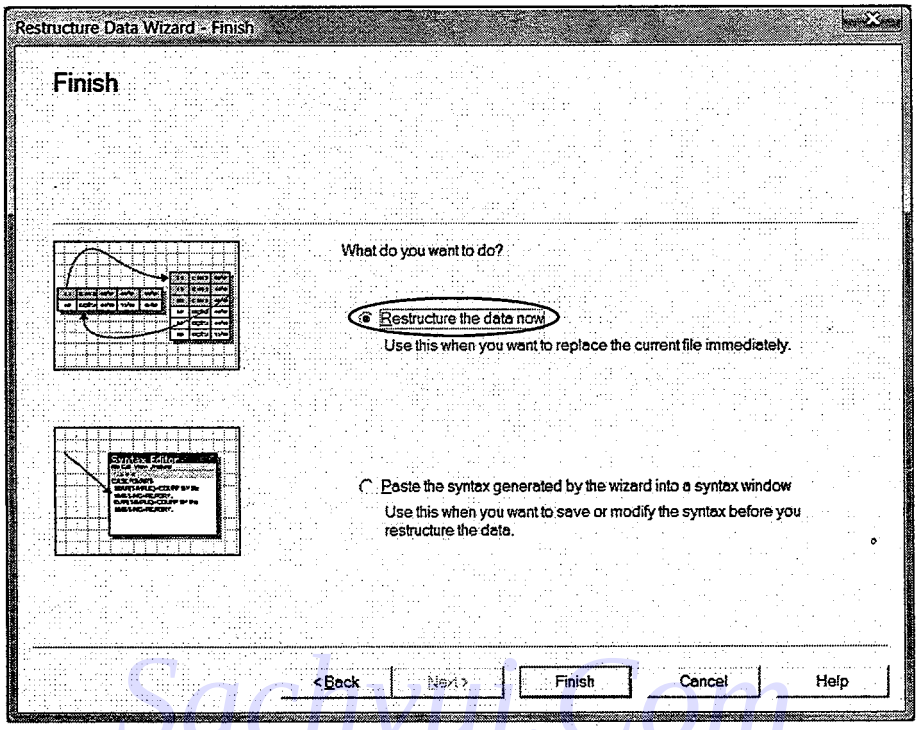
Trong màn hình ở bước 5, hãy để lựa chọn mặc định của chương trình là Sequential numbers (tạo ra biến chỉ mục với các trị số tuần tự từ 1 đến 6 vì chúng ta restructure từ 6 biến về thành 1 biến). Tiếp theo nhấp nút Next để chuyển sang bước 6.



Trong màn hình ở bước 6, hãy để các lựa chọn mặc định của chương trình là:

- Handling variable(s) not Selected: Keep and treat as fixed variables (coi các biến còn lại không nằm trong tập biến c2c1 ... c2c6 và c2d1 ... c2d6 là biến cố định và điền lập lại các dữ liệu này cho các dòng dữ liệu mới tạo ra)
- System Missing ot Blank Values in all Transposed Variables: Create a case in the new file (các ô chứa các giá trị khuyết hay để trống trong tập hợp biến cần cấu trúc lại cũng được coi như các giá trị bình thường và cũng tạo ra những dòng dữ liệu tương ứng, do đó số dòng dữ liệu của file mới tạo ra từ lệnh Restructure sẽ giống như tính toán lý thuyết, trong ví dụ này là từ 500 dòng ban đầu thành 3.000 dòng dữ liệu, để dễ kiểm tra)

Sau đó nhấp nút Next chuyển sang màn hình bước cuối.



Trong bước cuối này để mặc định là thực hiện lệnh ngay (Restructure the data now) rồi nhấp nút Next, lệnh được thực hiện ngay. 2 cột biến mới c2c và c2d xuất hiện ở phía cuối, và có tới 3000 dòng dữ liệu trong file mới này. Bạn hãy save lại file mới này và đặt tên, ví dụ như Data thuc hanh – restructure 2c-2d.

Untitled - SPSS Data Editor

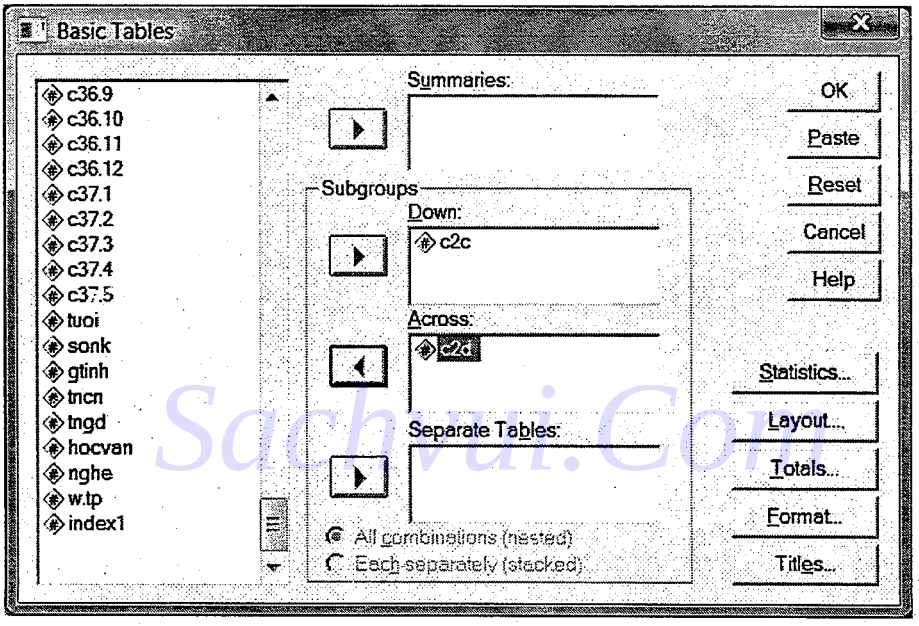
File Edit View Data Transform Analyze Graphs Utilities Window Help

3000 : c2d

	tncn	tngd	hocvan	nghe	w.tp	index1	c2c	c2d	var
2991	3	2	3	14	1.33	3	13	1	
2992	3	2	3	14	1.33	4	18	1	
2993	3	2	3	14	1.33	5	.	.	
2994	3	2	3	14	1.33	6	.	.	
2995	4	2	2	8	1.33	1	12	3	
2996	4	2	2	8	1.33	2	18	1	
2997	4	2	2	8	1.33	3	.	.	
2998	4	2	2	8	1.33	4	.	.	
2999	4	2	2	8	1.33	5	.	.	
3000	4	2	2	8	1.33	6	.	.	
3001									
3002									
3003									

Sau khi có được dữ liệu đã cấu trúc lại, bây giờ chúng tá có dữ liệu trong đó báo thường mua chỉ có 1 cột và nơi mua báo tương ứng chỉ có 1 cột.

Để biết được tỉ lệ mua đặt trước, mua tại các sạp báo trên trục đường đi và mua tại các sạp báo gần nhà (khu dân cư), chúng ta có thể dùng lệnh đơn giản như Custom Tables hay Basic Tables để lập bảng kết hợp tính ra các kết quả này. Trong ví dụ này chúng ta sử dụng lệnh Basic Tables như mô trả trong hình dưới (Bạn đọc có thể dùng Lệnh Custom Tables hay General Tables để có kết quả tương tự).



Sau khi chọn biến mới c2c (báo thường mua) đưa vào ô tạo biến dòng và đưa biến mới c2d (cách mua báo) vào ô chứa biến cột; vào Statistics chọn hàm Count và Row Percent; vào Total để hiện cột cộng và dòng cộng; vào Layout để tách bảng tần số riêng và bảng % riêng. Nhấp nút OK bạn sẽ được bảng kết quả.

Để hiện đồng thời 2 bảng tần số và % cùng một lúc, hãy nhấp chuột nhanh vào bảng kết quả để chuyển sang chế độ hiệu chỉnh bảng. Vào tiếp menu Pivot, chọn Move Layers to Rows, bạn sẽ thấy bảng kết quả như ở trang sau (Bạn đọc có thể tham khảo lại hướng dẫn những thao tác này trong phần lập bảng ở Tập 1 của sách này).

			cách mua báo					Group Total
			Đặt báo	Sạp báo tiện đường	Sạp báo gần r hà	không biết	hàng rong	
Count	các tờ báo GĐ thường mua	HN mới	32	16	39	1	3	91
		SGGP	19	4	10	2		35
		Lao Động	14	12	22		1	49
		Người Lao Động	9	14	16	4		43
		Tiền Phong	4	12	34			54
		Thanh Niên	15	8	24	3	1	51
		Tuổi Trẻ	54	47	73	16	2	192
		Phụ Nữ VN	10	26	41	2	7	86
		Phụ Nữ TPHCM	19	18	16	6		59
		Thời Báo KTVN	4	8	6	1		19
		Thời Báo KTSG		2	3	1		6
		SG Tiếp Thị	19	24	52	11		106
		Thế Giới Phụ Nữ	23	38	69	4	2	136
		Tiếp Thị và GĐ	9	7	19	1	1	37
		Mua & Bán	3	6	23		7	39
		An Ninh Thế Giới	29	73	102	9	16	229
		An Ninh Thủ Đô	9	50	75	2	16	152
		Công An TPHCM	47	40	62	14	3	166
		Khác	28	39	70	10	15	162
	Group Total		347	444	756	87	78	1712
Row %	các tờ báo GĐ thường mua	HN mới	35.2%	17.6%	42.9%	1.1%	3.3%	100.0%
		SGGP	54.3%	11.4%	28.6%	5.7%		100.0%
		Lao Động	28.6%	24.5%	44.9%		2.0%	100.0%
		Người Lao Động	20.9%	32.6%	37.2%	9.3%		100.0%
		Tiền Phong	7.4%	22.2%	63.0%		7.4%	100.0%
		Thanh Niên	29.4%	15.7%	47.1%	5.9%	2.0%	100.0%
		Tuổi Trẻ	28.1%	24.5%	38.0%	8.3%	1.0%	100.0%
		Phụ Nữ VN	11.6%	30.2%	47.7%	2.3%	8.1%	100.0%
		Phụ Nữ TPHCM	32.2%	30.5%	27.1%	10.2%		100.0%
		Thời Báo KTVN	21.1%	42.1%	31.6%	5.3%		100.0%
		Thời Báo KTSG		33.3%	50.0%	16.7%		100.0%
		SG Tiếp Thị	17.9%	22.6%	49.1%	10.4%		100.0%
		Thế Giới Phụ Nữ	16.9%	27.9%	50.7%	2.9%	1.5%	100.0%
		Tiếp Thị và GĐ	24.3%	18.9%	51.4%	2.7%	2.7%	100.0%
		Mua & Bán	7.7%	15.4%	59.0%		17.9%	100.0%
		An Ninh Thế Giới	12.7%	31.9%	44.5%	3.9%	7.0%	100.0%
		An Ninh Thủ Đô	5.9%	32.9%	49.3%	1.3%	10.5%	100.0%
		Công An TPHCM	28.3%	24.1%	37.3%	8.4%	1.8%	100.0%
		Khác	17.3%	24.1%	43.2%	6.2%	9.3%	100.0%
	Group Total		20.3%	25.9%	44.2%	5.1%	4.6%	100.0%

3. GHÉP TRỘN 2 FILE DỮ LIỆU (Merge Files – Add Variables)

Ghép trộn dữ liệu từ 2 file lại với nhau cần thiết trong một số tình huống người nghiên cứu có 2 đơn vị phân tích trong một cuộc khảo sát. Ví dụ như khi khảo sát các hộ gia đình, vừa có thông tin chung của hộ gia đình, vừa có thông tin của từng nhân khẩu trong hộ. Thường thì các dữ liệu của hộ được nhập vào file riêng và dữ liệu của từng nhân khẩu được nhập vào file riêng. Do mỗi hộ thường có nhiều hơn 1 nhân khẩu nên số dòng dữ liệu trong file nhân khẩu nhiều hơn số dòng dữ liệu trong file hộ. Người nghiên cứu có thể thực hiện phân tích theo đơn vị hộ trên file chứa dữ liệu hộ và phân tích theo đơn vị nhân khẩu (người) trên file chứa dữ liệu nhân khẩu. Trong một số trường hợp, người nghiên cứu cần phân tích kết hợp giữa dữ liệu của hộ và dữ liệu của cá nhân thì trước khi phân tích người nghiên cứu cần trộn 2 dữ liệu này lại với nhau. Chẳng hạn như cần phân tích xem quy mô hộ gia đình, thu nhập trung bình 1 người trong hộ (dữ liệu của hộ) có liên quan đến mức độ thường xuyên đi ăn ngoài của những người lớn trưởng thành không (dữ liệu của từng người)? Học vấn của người có học vấn cao nhất trong hộ có liên quan đến kết quả học tập của các học sinh phổ thông đang đi học không? Thu nhập trung bình của những người có đi làm trong hộ có liên quan đến quan niệm về nguyên nhân của nghèo đói...

Chúng ta cũng có thể làm tương tự khi có dữ liệu về công ty (ví dụ như 40 công ty) và có dữ liệu về những nhân viên làm việc trong các công ty này (ví dụ như khảo sát 30 người trong mỗi công ty). Việc ghép file này cần thiết khi người nghiên cứu muốn tìm hiểu tương quan giữa hiệu quả kinh doanh của công ty (dữ liệu thứ cấp thu thập từ tài liệu có sẵn của từng công ty) và mức độ hài lòng của các nhân viên về công ty (dữ liệu sơ cấp thu thập từ nhiều nhân viên làm việc trong cùng 1 công ty).

Phần này sẽ hướng dẫn các bạn ghép 2 file lại với nhau qua 1 ví dụ cụ thể. Giả sử chúng ta có 1 file chứa thông tin của các nhân khẩu trong từng hộ gia đình (Hokhau_NK.sas) đã khảo sát và 1 file chứa thông tin chung của hộ (Hokhau_Ho.sas). File nhân khẩu có 45 dòng dữ liệu ứng với 45 nhân khẩu trong khi file hộ chỉ có 10 dòng dữ liệu

ứng với 10 hộ đã khảo sát (số thứ tự hộ được đánh số từ 451 đến 460). Như vậy mỗi hộ thường có nhiều hơn 1 nhân khẩu. Các file dữ liệu này nằm trong bộ dữ liệu thực hành với sách mà bạn đã tải xuống.

3.1. Trộn ghép dữ liệu của đơn vị bậc cao vào dữ liệu của đơn vị bậc thấp

Trộn ghép dữ liệu của đơn vị bậc cao vào dữ liệu của đơn vị bậc thấp thường diễn ra dưới dạng dữ liệu của đơn vị bậc cao là yếu tố nguyên nhân ảnh hưởng đến yếu tố kết quả trong dữ liệu bậc thấp. Ví dụ như hình ảnh uy tín của công ty/tổ chức đối với người tiêu dùng hay xã hội ảnh hưởng như thế nào đến sự gắn bó đối với công ty/tổ chức của nhân viên.

Trong trường hợp nghiên cứu các doanh nghiệp, thì các doanh nghiệp là đơn vị bậc cao, còn những người làm trong các doanh nghiệp đó là đơn vị bậc thấp. Khi nghiên cứu các trường học thì trường học là đơn vị bậc cao còn giáo viên hay học sinh - sinh viên trong các trường đó là đơn vị bậc thấp. Khi nghiên cứu các hộ gia đình thì hộ gia đình là đơn vị bậc cao còn các nhân khẩu trong các hộ là đơn vị bậc thấp.

Trong phần này chúng ta sẽ lấy 1 ví dụ về khảo sát các hộ gia đình và các nhân khẩu trong hộ được trích ra từ một cuộc khảo sát trong năm 2005 do Trung Tâm Xã Hội Học (Viện Khoa Học Xã Hội Vùng Nam Bộ) thực hiện. Đầu tiên hãy mở file chi tiết từng nhân khẩu ra (Hokhau_NK). Sau khi mở đúng file thành công, bạn sẽ thấy trên màn hình như sau:

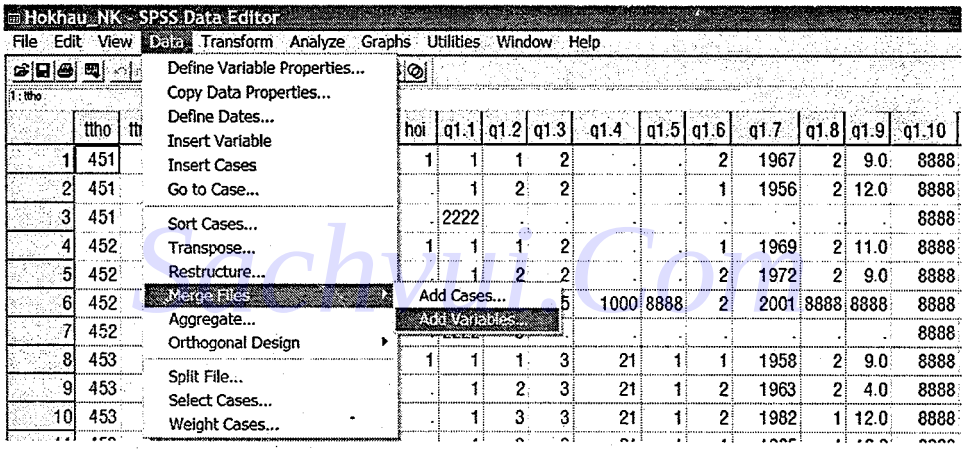
Hokhau_NK - SPSS Data Editor

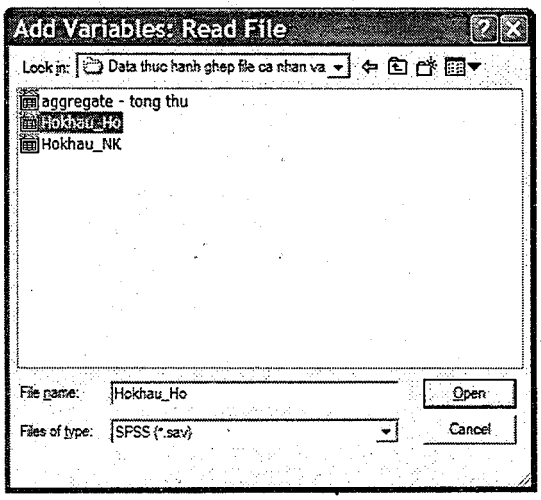
File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : ttho 451

	ttho	ttnk	ten	hoi	q1.1	q1.2	q1.3	q1.4	q1.5	q1.6	q1.7	q1.8	q1.9	q1.10
1	451	1	Ng Thúy Vân	1	1	1	2	.	.	2	1967	2	9.0	8888
2	451	2	Hồ Kim Thúc	.	1	2	2	.	.	1	1956	2	12.0	8888
3	451	3	Ng Nhứt Thiên	.	2222	.	.	.	.	.	.	.	.	8888
4	452	1	Ng Hữu Đức	1	1	1	2	.	.	1	1969	2	11.0	8888
5	452	2	Ng T Thanh Hương	.	1	2	2	.	.	2	1972	2	9.0	8888
6	452	3	Ng Thanh Ngân	.	1	3	5	1000	8888	2	2001	8888	8888	8888
7	452	4	Ng Đức Thanh Xuân	.	2222	3	.	.	.	.	.	.	.	8888
8	453	1	Trần Hữu Vương	1	1	1	3	21	1	1	1958	2	9.0	8888
9	453	2	Ng T Ngư	.	1	2	3	21	1	2	1963	2	4.0	8888
10	453	3	Trần T Nguyệt	.	1	3	3	21	1	2	1982	1	12.0	8888

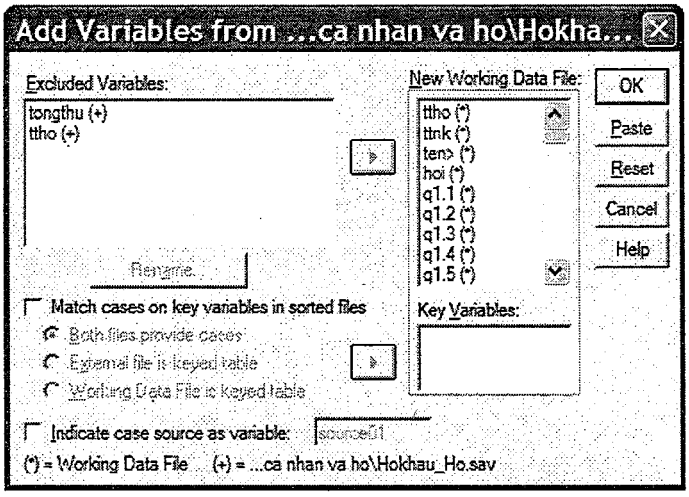
Trong hình trên chúng ta thấy hộ có số thứ tự hộ (ttho) 451 có 3 nhân khẩu, hộ tiếp theo có số thứ tự hộ là 452 có 4 nhân khẩu. Sau đó bạn chọn lệnh Data > Merge Files như hình dưới đây:





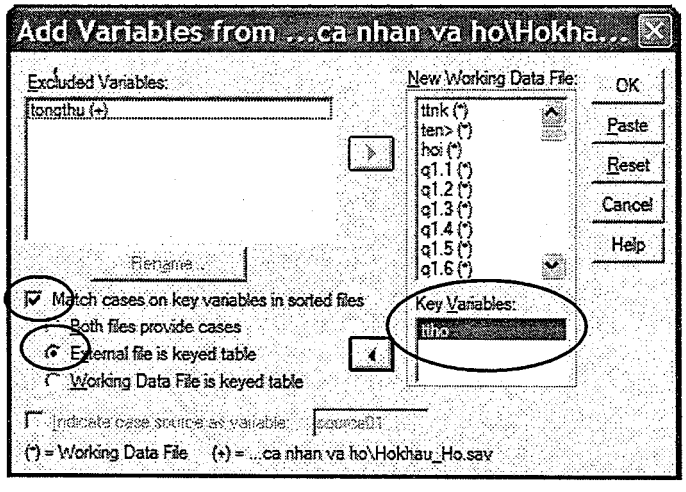
Bạn hãy chọn option Add Variables như hình trên, hộp thoại Add Variables sẽ xuất hiệnnhư hình bên cạnh.

Trong hộp thoại này chọn thư mục, tên file chứa thông tin của hộ (Hokhau_Ho.sas), nhấp nút Open, hộp thoại sau đây sẽ hiện ra:

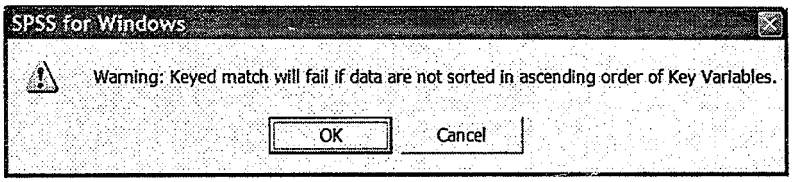


Trong hộp thoại này chọn Match cases on key variable in sorted files: Ghép biến của file hộ vào file cá nhân, trong đó các biến của file hộ được đưa vào file cá nhân theo cách các dòng dữ liệu của file hộ sẽ điền vào các dòng tương ứng (match cases) của file cá nhân theo biến chủ. (Key Variables) với điều kiện các file này đã được xếp thứ tự theo biến chủ.

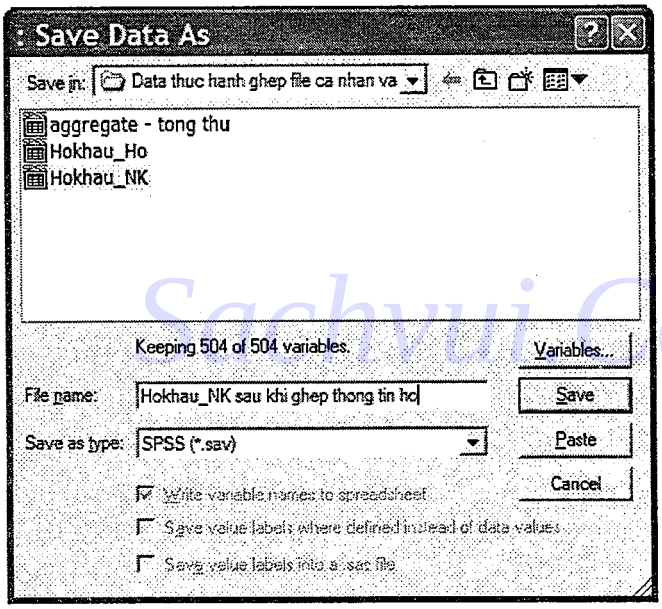
Trong ví dụ này, hãy chọn mục External file is keyed table (File đang được mở ra là file hộ chứa dữ liệu sẽ được dùng để điền vào file cá nhân đã mở trước đó) chọn biến ttho (số thứ tự hộ) trong ô Excluded Variables rồi nhấn vào nút mũi tên qua phải đưa biến này vào ô Key Variables như hình dưới:



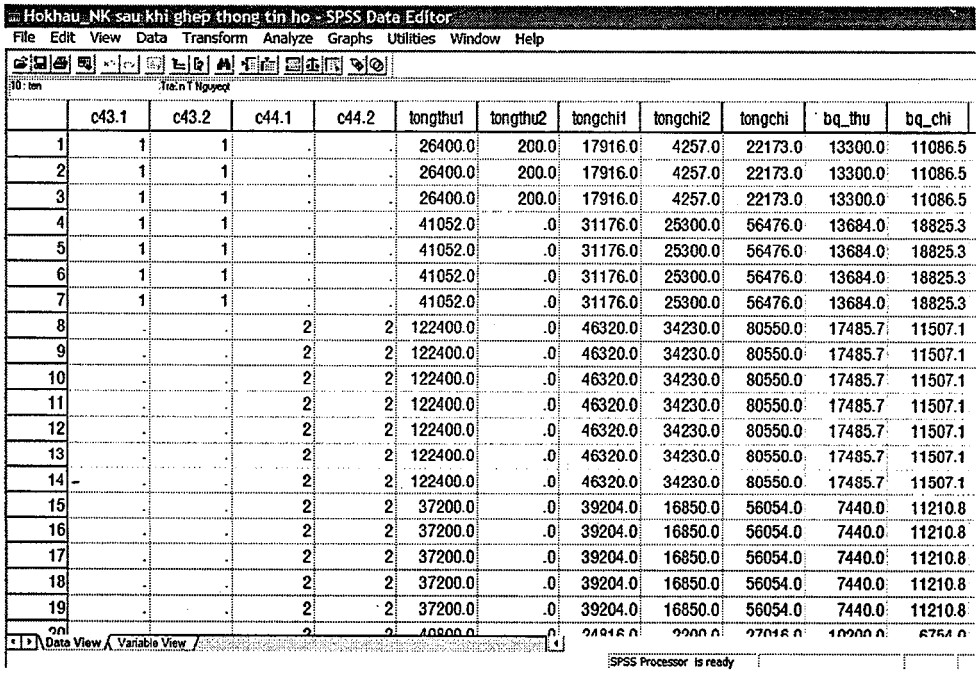
Rồi nhấn nút OK thì thông báo cảnh báo sau sẽ hiện ra lưu ý rằng lệnh này sẽ không có kết quả khi biến chủ không được sắp xếp theo thứ tự tăng dần.



Cuối cùng nhấn nút OK sẽ được 1 file mới chứa thông tin của từng cá nhân có thêm thông tin của hộ tương ứng. Số biến trong file cá nhân ban đầu là 63 biến, sau khi ghép xong lên tới 504 biến. Cần save lại và đặt tên file mới như gợi ý dưới đây:



Sau khi Save xong chúng ta đã có file chứa thông tin cá nhân cùng với những thông tin của hộ gia đình tương ứng sẵn sàng cho phân tích.



Hokhau_NK sau khi ghep thong tin ho - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	c43.1	c43.2	c44.1	c44.2	tongthu1	tongthu2	tongchi1	tongchi2	tongchi	bq_thu	bq_chi
1	1	1	.	.	26400.0	200.0	17916.0	4257.0	22173.0	13300.0	11086.5
2	1	1	.	.	26400.0	200.0	17916.0	4257.0	22173.0	13300.0	11086.5
3	1	1	.	.	26400.0	200.0	17916.0	4257.0	22173.0	13300.0	11086.5
4	1	1	.	.	41052.0	.0	31176.0	25300.0	56476.0	13684.0	18825.3
5	1	1	.	.	41052.0	.0	31176.0	25300.0	56476.0	13684.0	18825.3
6	1	1	.	.	41052.0	.0	31176.0	25300.0	56476.0	13684.0	18825.3
7	1	1	.	.	41052.0	.0	31176.0	25300.0	56476.0	13684.0	18825.3
8	.	.	2	2	122400.0	.0	46320.0	34230.0	80550.0	17485.7	11507.1
9	.	.	2	2	122400.0	.0	46320.0	34230.0	80550.0	17485.7	11507.1
10	.	.	2	2	122400.0	.0	46320.0	34230.0	80550.0	17485.7	11507.1
11	.	.	2	2	122400.0	.0	46320.0	34230.0	80550.0	17485.7	11507.1
12	.	.	2	2	122400.0	.0	46320.0	34230.0	80550.0	17485.7	11507.1
13	.	.	2	2	122400.0	.0	46320.0	34230.0	80550.0	17485.7	11507.1
14	.	.	2	2	122400.0	.0	46320.0	34230.0	80550.0	17485.7	11507.1
15	.	.	2	2	37200.0	.0	39204.0	16850.0	56054.0	7440.0	11210.8
16	.	.	2	2	37200.0	.0	39204.0	16850.0	56054.0	7440.0	11210.8
17	.	.	2	2	37200.0	.0	39204.0	16850.0	56054.0	7440.0	11210.8
18	.	.	2	2	37200.0	.0	39204.0	16850.0	56054.0	7440.0	11210.8
19	.	.	2	2	37200.0	.0	39204.0	16850.0	56054.0	7440.0	11210.8

Data View | Variable View

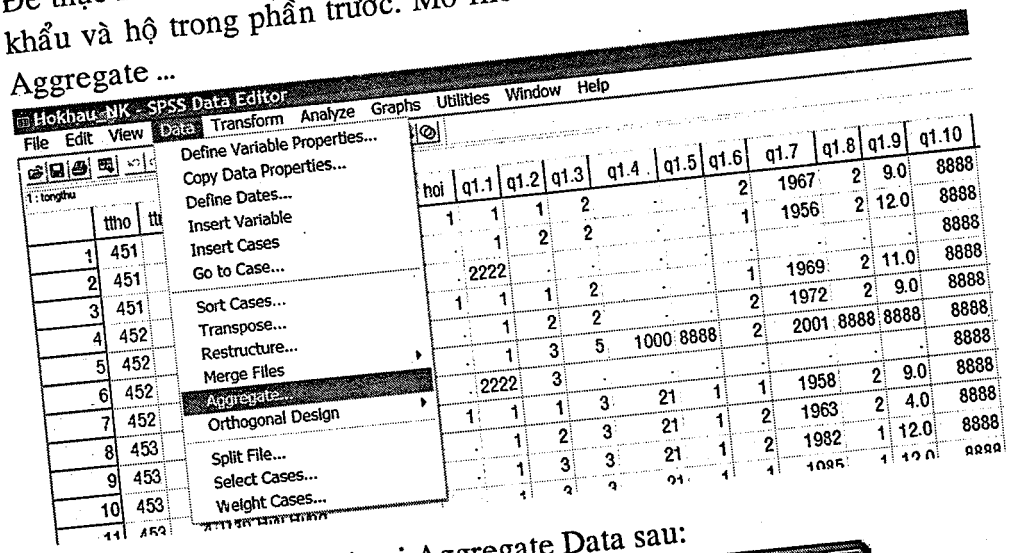
SPSS Processor is ready

3.2. Tổng hợp dữ liệu của các đơn vị bậc thấp trong cùng một đơn vị bậc cao thành dữ liệu đại diện và ghép vào dữ liệu của các đơn vị bậc cao

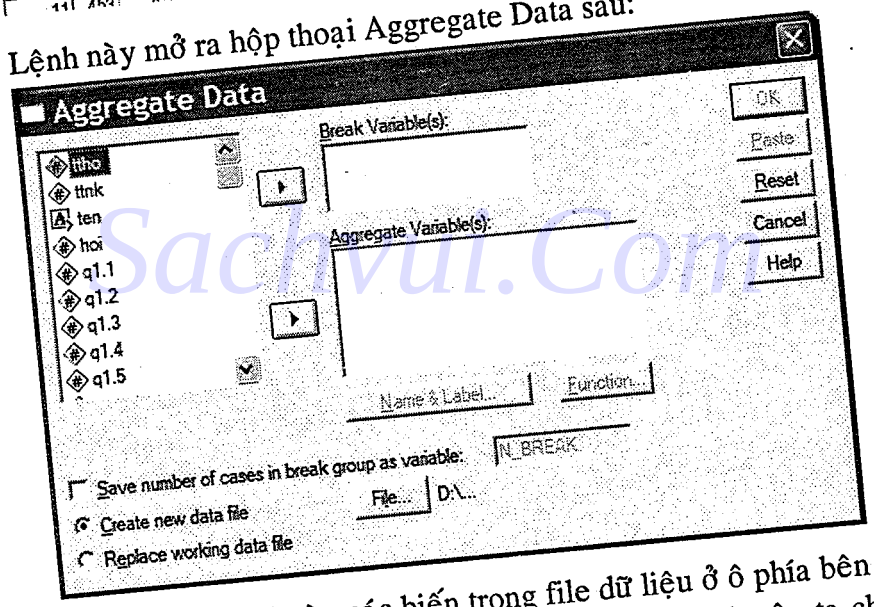
Ngược lại với phần trước, chúng ta cũng có thể ghép (trộn) dữ liệu sơ cấp từ khảo sát thực tế (trong 1 file riêng) với dữ liệu thứ cấp có sẵn (trong 1 file riêng). Thay vì trộn ghép dữ liệu của hộ vào dữ liệu của cá nhân như phần trên đã trình bày, chúng ta có thể ghép dữ liệu tổng hợp của các cá nhân trong cùng 1 hộ lại thành dữ liệu đại diện cho hộ để ghép vào file hộ. Ví dụ như tổng hợp dữ liệu từ các nhân khẩu trong nội bộ 1 hộ để có tỉ lệ người trong tuổi lao động đang có việc làm của hộ rồi ghép ngược trở lại vào file dữ liệu của hộ.

Tương tự, giả sử chúng ta có dữ liệu khảo sát các các cổ đông của từng công ty, thì cũng có thể tổng hợp thành dữ liệu đại diện cho từng công ty như tỉ lệ cổ đông có sở hữu cổ phần của các doanh nghiệp cùng ngành và ghép vào file chứa dữ liệu của các công ty.

Để thực hành, chúng ta sử dụng lại hai file dữ liệu của ví dụ về nhân khẩu và hộ trong phần trước. Mở file cá nhân ra, chọn lệnh Data > Aggregate ...

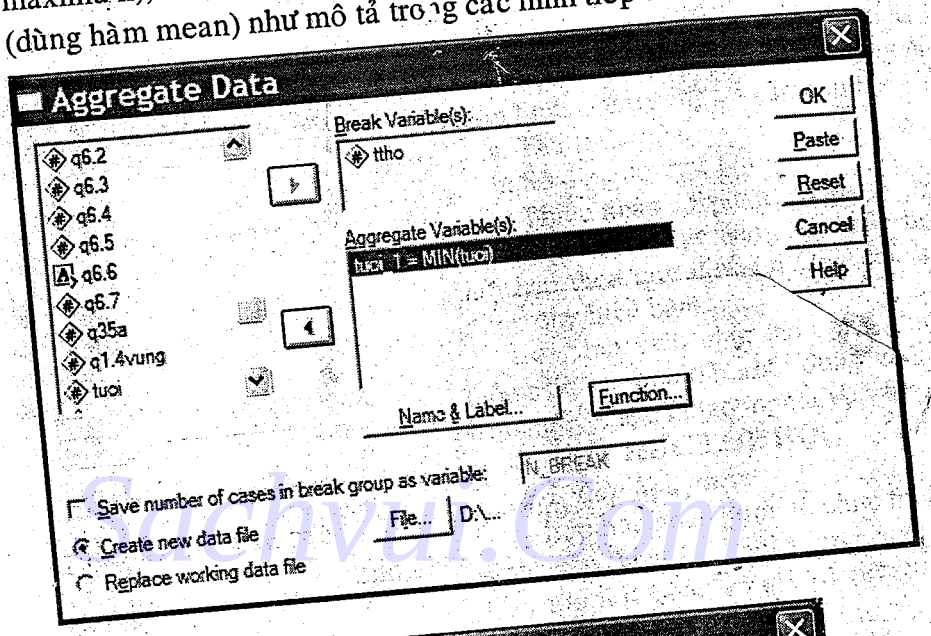


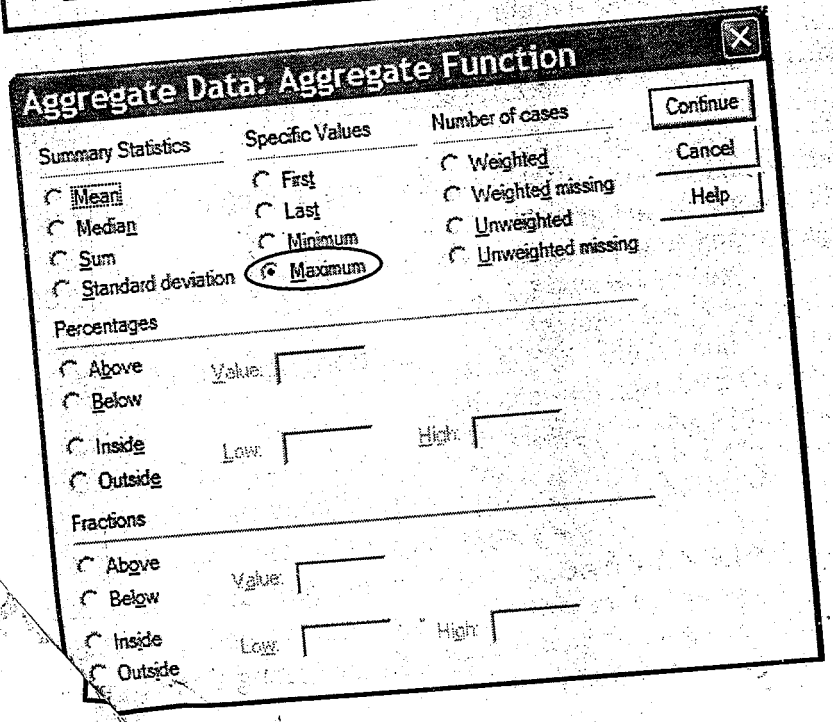
Lệnh này mở ra hộp thoại Aggregate Data sau:



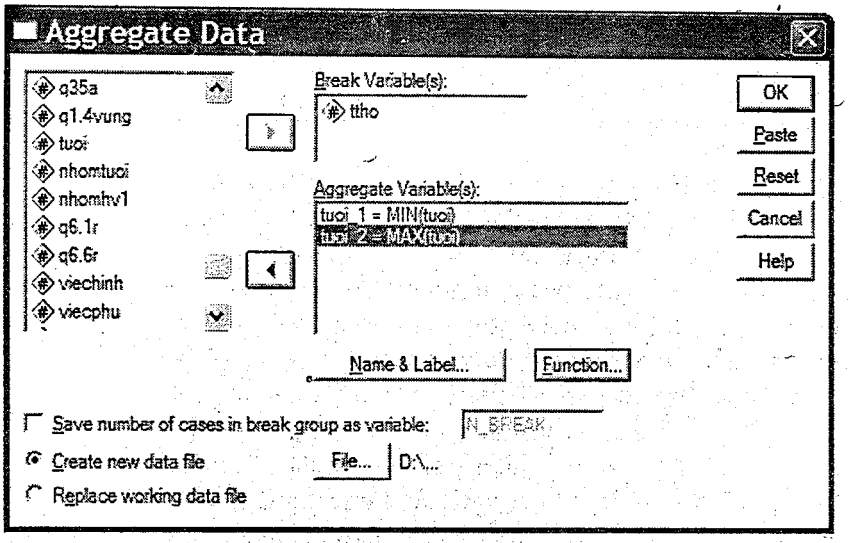
Trong hộp thoại này các biến trong file dữ liệu ở ô phía bên tay trái, do chúng ta muốn tổng hợp dữ liệu theo từng hộ nên ta chọn biến ttho đưa vào ô Break Variable(s) để báo cho chương trình biết là chỉ tính toán tổng hợp trong phạm vi từng hộ (tính toán cho các cases có chung mã số ttho). Rồi tùy theo nhu cầu tính toán mà chọn biến đưa vào ô Aggregate Variables (tổng hợp dữ liệu theo biến nào). Trong ví dụ này hãy chọn biến tongth (tổng thu nhập). Khi đưa biến vào thì hàm mặc định là MEAN (tính trung bình).

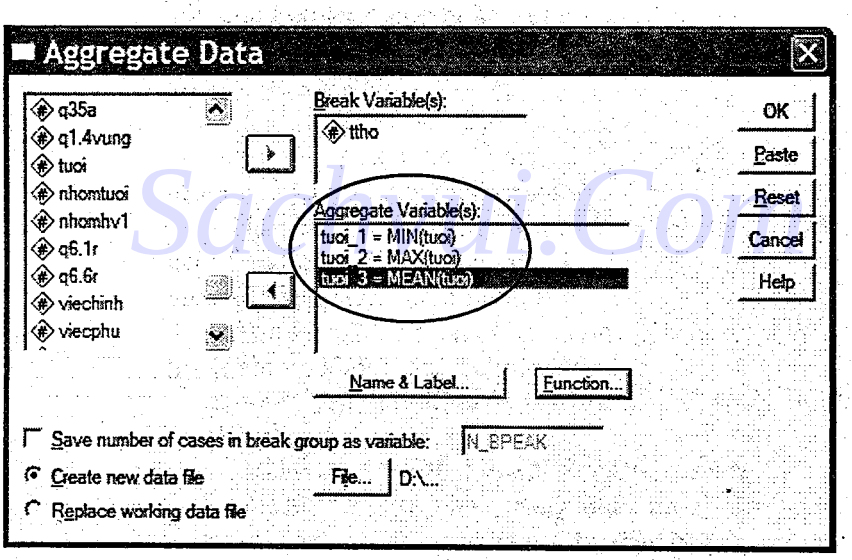
Trong ví dụ này chúng ta tiếp tục tổng hợp tổng thu để biết người thu nhập thấp nhất trong hộ cụ thể là bao nhiêu (dùng hàm minimum), người thu nhập cao nhất trong hộ cụ thể là bao nhiêu (dùng hàm maximum), thu nhập trung bình của 1 người trong hộ là bao nhiêu (dùng hàm mean) như mô tả trong các hình tiếp theo:



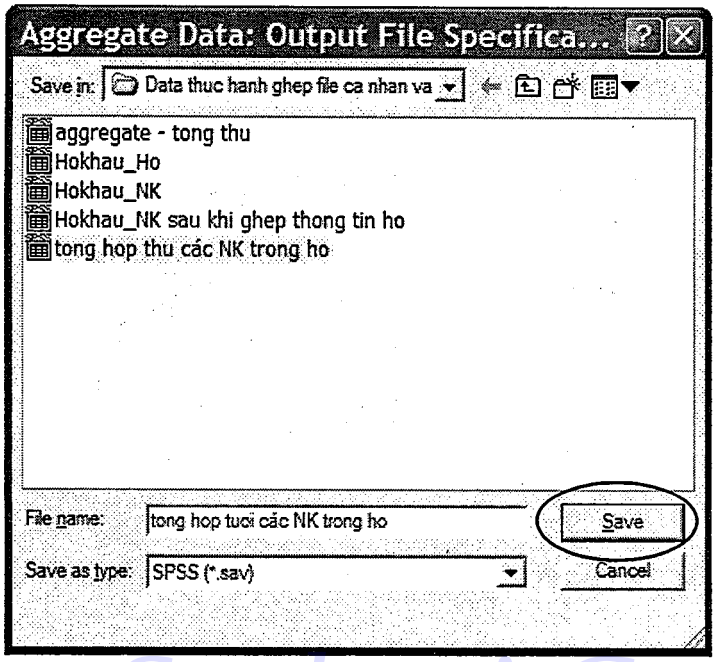


Ví dụ
thành v
nút Con
Data, để
Creat new





Sau đó chọn mục Create new data file và khai báo thư mục và tên file rồi nhấn nút Save.



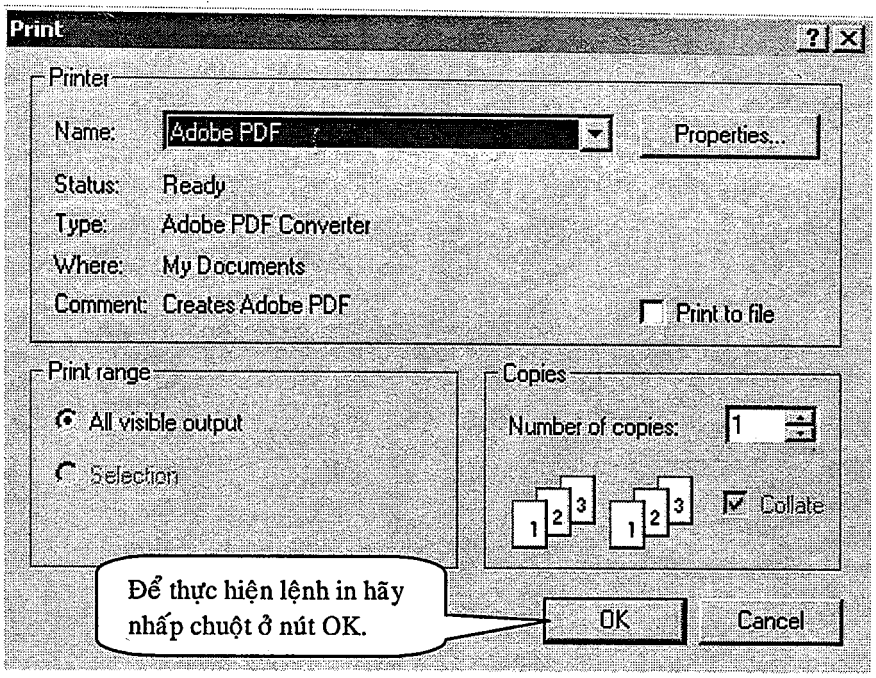
Sau đó vào thư mục đã lưu, tìm tên file và mở file ra để xem kết quả. Từ các biến mới tạo ra trong file kết quả này, bạn có thể copy và dán vào file chứa dữ liệu hộ (hoặc dùng lệnh Merge Files để ghép các biến mới này vào) và thực hiện các phân tích tiếp theo mà bạn cần.

4. CÁCH THỨC IN ẤN

4.1. Cách thức in một tập tin kết quả

Bạn có thể in toàn bộ các tập tin hay chỉ một phần của tập tin kết quả bằng cách từ menu chọn File > Print. Lệnh này mở ra hộp thoại như Hình 16.1. Tên của loại máy in hiển thị ở đây.

Hình 16.1



Khi in tập tin kết quả ta có các tuỳ chọn sau:

- Phần Print Range
  - *All* visible output: mặc định in toàn bộ kết quả nhìn thấy được (không in những thành phần ẩn của output)
  - *Selection* (chọn): chỉ in đối với phần đã được chọn từ trước trong tập tin
- Phần Copies (Số bản in): mặc định chỉ in một bản. Nếu muốn in bao nhiêu bản thì ta gõ số bản cần in vào (hoặc nhấp vào nút tăng giảm) khung Number of Copies.

4.2. Cách thức in một tập tin dữ liệu

Để in một tập tin dữ liệu, các bước sẽ được tiến hành giống như khi in một tập tin kết quả. Ta cũng có thể in toàn bộ tập tin hay chỉ một vùng trong tập tin dữ liệu được chọn.

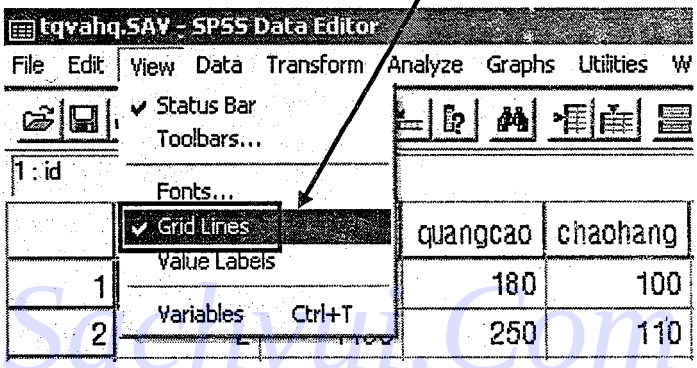
- Phần Print Range
  - *All* (tất cả): mặc định in toàn bộ tập tin
  - *Selection* (chọn): chỉ in đối với phần được chọn trong tập tin
  - *Page (trang):chỉ định in từ trang mấy đến trang mấy*

- Phần Copies (Số bản in): như trường hợp trên

Các tuỳ chọn phụ trợ cho kết quả in tập tin dữ liệu

- Bỏ các đường lưới trên màn hình dữ liệu
  Theo mặc định, các đường kẻ ngang dọc được hiển thị trên màn hình cũng sẽ được in ra. Để tắt, mở chế độ lưới (Grid lines) bạn chọn menu View rồi nhấp bỏ dấu chọn trước Grid Lines. Thao tác ngược lại nếu bạn lại muốn hiện đường lưới. (xem hình)

Hình 16.2



- Hiện nhãn biến Value Labels

  Theo mặc định, các nhãn biến không hiển thị trên màn hình. Để tắt, mở chế độ hiển thị Value Labels, bạn làm như cách tắt mở lưới, tuỳ chọn này nằm ngay dưới tuỳ chọn về Grid Line.

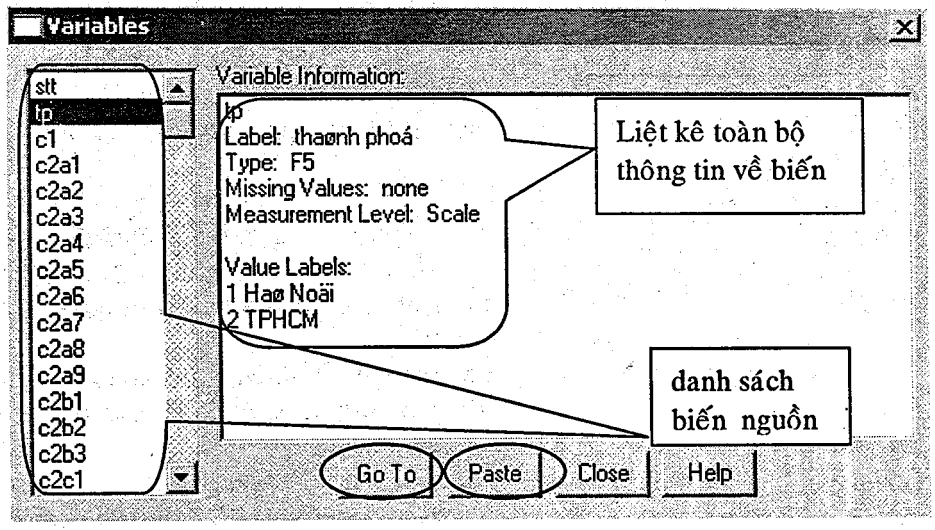
  Khi mở chế độ Value Labels, tất cả các giá trị trên data view sẽ xuất hiện dưới dạng nhãn. Nếu nhãn này dài hơn chiều rộng đã khai báo của ô biến thì khi hiển thị sẽ bị che mất phần dài hơn này.

5. XEM CÁC THÔNG TIN VỀ BIẾN

Để xem thông tin của từng biến, sao chép và dán tên các biến vào cú pháp lệnh, hoặc đi tìm các biến cụ thể trên màn hình để sửa chữa dữ liệu, ta chọn menu Utilities > Variables ...

Giả dụ ta đang ở tập tin có tên *Data thuchanh*, lệnh trên mở ra hộp thoại các biến như sau:

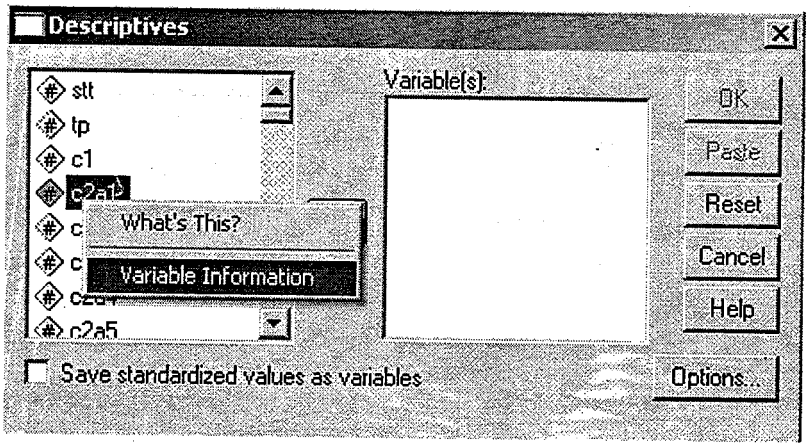
Hình 16.3



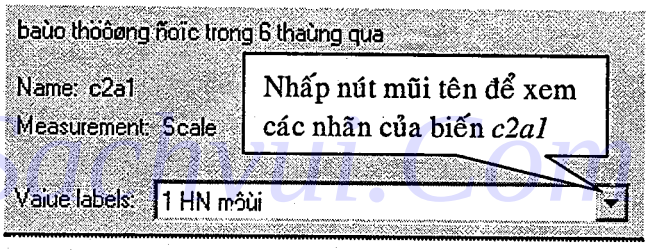
- Muốn biết thông tin về biến nào bạn nhấp chuột vào tên biến đó trong danh sách biến nguồn bên trái thì các thông tin về biến sẽ hiện ở khung bên phải, bao gồm: nhãn của biến, dạng dữ liệu, định nghĩa các giá trị thiếu (missing value) và các nhãn giá trị của biến.
- Nhấp nút Goto để nhảy trực tiếp đến vị trí của biến đang chọn trên cửa sổ Data Vienhằm thực hiện hiệu chỉnh.
- Nhấp nút Paste để dán tên biến vào màn hình cú pháp lệnh.
- Nhấp nút Close để đóng hộp thoại.

Có 1 cách nhanh gọn để xem thông tin về biến khi bạn đang thao tác trong các cửa sổ hộp thoại lệnh bất kỳ của SPSS. Giả dụ bạn đang thực hiện lệnh Descriptives, sau khi mở cửa sổ hộp thoại, bạn có nhu cầu cần xác định lại thông tin về biến *c2a1* mà bạn ngại quay lại màn hình khai báo biến để kiểm tra, lúc này bạn nhấp chọn tên biến trong danh sách biến nguồn của hộp thoại rồi nhấp chuột phải, sẽ có một menu tắt mở ra với 2 lựa chọn (xem Hình 16.4), bạn nhấp tiếp Variable Information, các thông tin về biến *c2a1* sẽ xuất hiện kế tiếp như trong Hình 16.5

Hình 16.4



Hình 16.5



Muốn trở lại trạng thái cũ của hộp thoại Descriptives bạn chỉ cần nhấp chuột trái một lần vào vị trí bất kỳ trên màn hình.

6. XEM THÔNG TIN VỀ TẬP TIN

Để biết thông tin về tập tin, bạn cũng dùng menu Utilities, bạn nhấp File Info...Các thông tin về tập tin sẽ hiển thị trên cửa sổ kết quả:

- tên biến
- mô tả nhãn biến
- định dạng kiểu in và chữ viết
- chiều rộng tối đa và số số lẻ cho phép của biến
- mô tả các nhãn giá trị (nếu có) của các biến.

7. TRAO ĐỔI THÔNG TIN VỚI CÁC ỨNG DỤNG KHÁC

Bạn có thể trao đổi các kết quả của SPSS cho các ứng dụng khác ví dụ cung cấp bảng biểu, đồ thị cho văn bản Word.

Để dán bảng biểu tạo được bằng SPSS lên văn bản Word bạn nhấp chuột trực tiếp vào bảng và quanh nó hiện đường khung chọn, nhấp chuột phải, trong menu lệnh tắt, bạn sẽ có 2 lựa chọn copy:

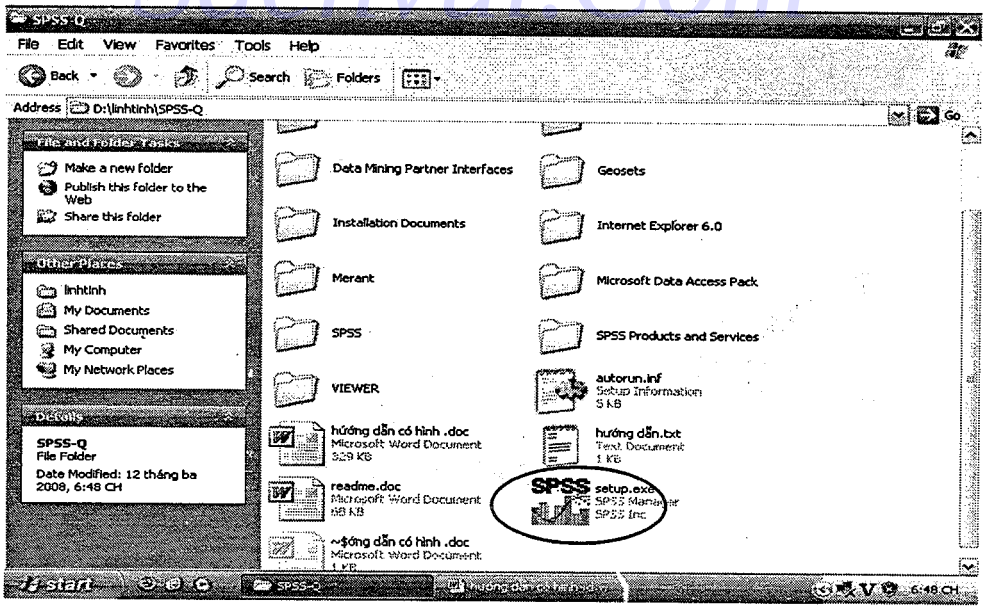
- Nếu chọn Copy Object (tổ hợp phím Control + K) thì sau khi dán lên văn bản Word, bảng của bạn sẽ là một picture và bạn chỉ có thể hiệu chỉnh kích cỡ, vị trí... như với một bức tranh.
- Nếu chọn Copy (tổ hợp phím Control + C) thì khi bạn dán bảng lên văn bản Word, bạn có thể thực hiện mọi hiệu chỉnh trên bảng của bạn như với một đối tượng Table bình thường của Word.

Bạn cũng có thể dán đồ thị của SPSS cho tập tin văn bản Word nhưng lúc này bạn không tạo liên kết trực tiếp giữa đồ thị trên văn bản Word với đồ thị trên output của SPSS được nên mọi thay đổi của đồ thị từ SPSS sẽ không được cập nhật tự động vào tập tin văn bản trên Word. Bạn tiến hành giống cách chép bảng.

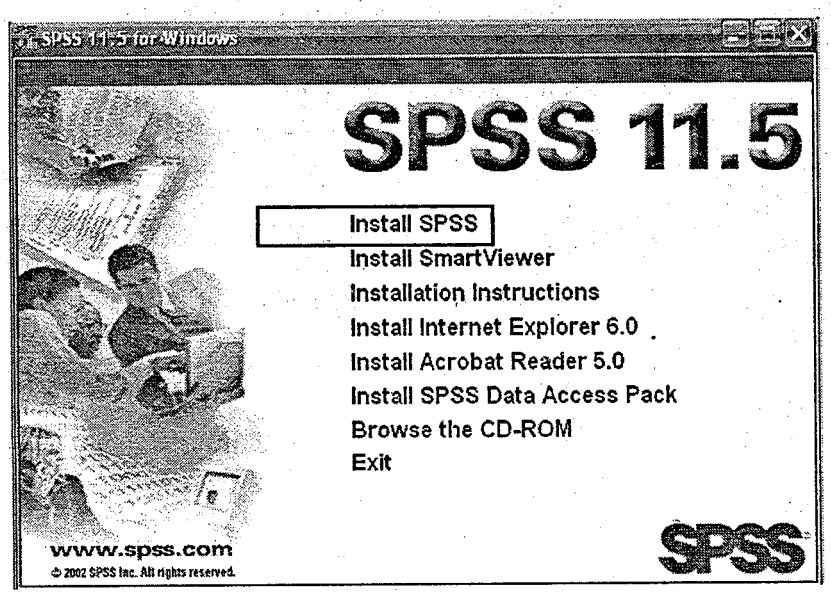
8. CÀI ĐẶT SPSS

Bước 1: sau khi để đĩa cài đặt chương trình vào ổ CD, bạn vào thư mục SPSS-Q, tìm biểu tượng file setup.exe bấm vào đó màn hình sẽ xuất hiện: bạn chọn install (cài đặt) spss (Hình 16.6 và 16.7)

Hình 16.6

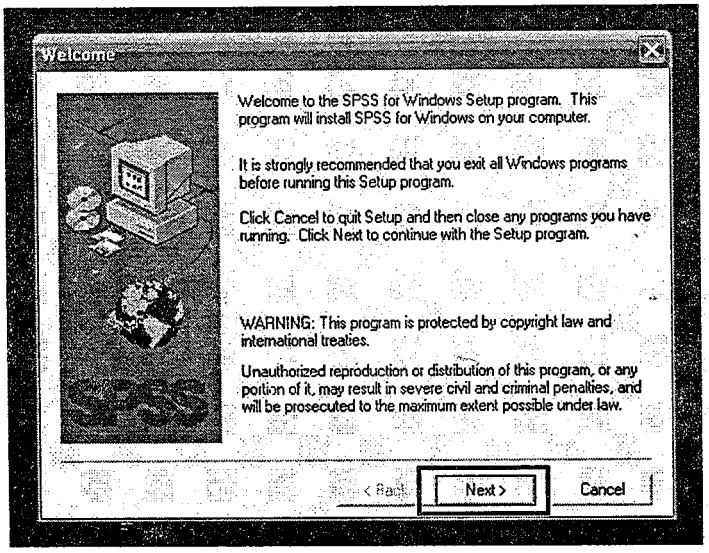


Hình 16.7



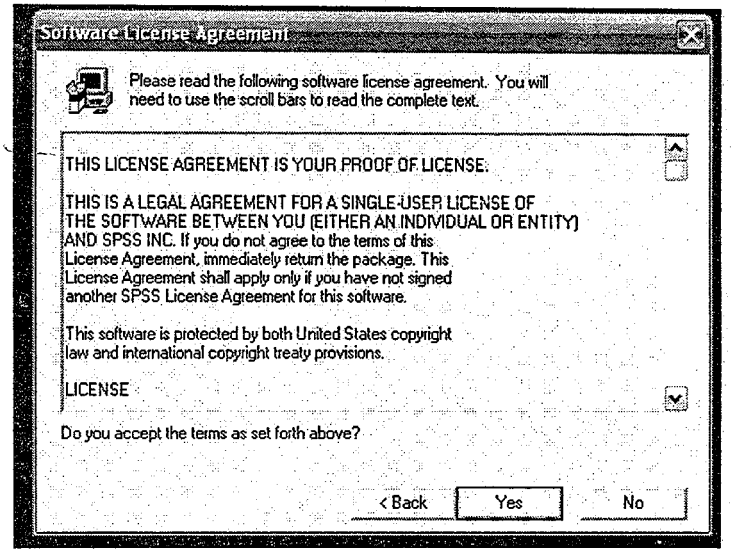
Bước 2: Trên màn hình hiện ra tiếp theo, bạn chọn Next

Hình 16.8



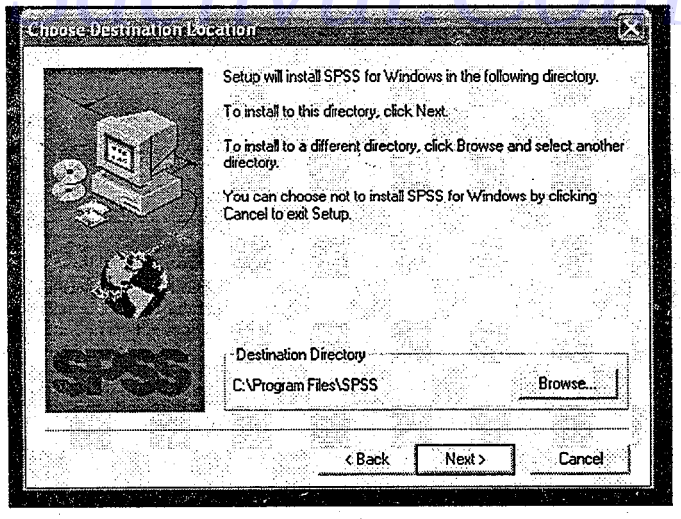
Bước 3: Trên màn hình sau bạn bấm nút Yes

Hình 16.9



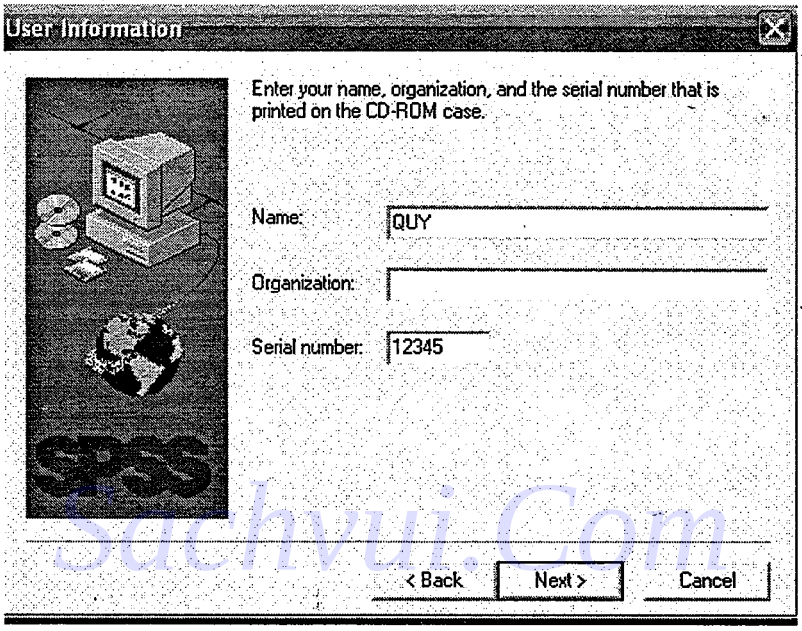
Bước 4: Chọn ổ đĩa xong bạn bấm Next

Hình 16.10



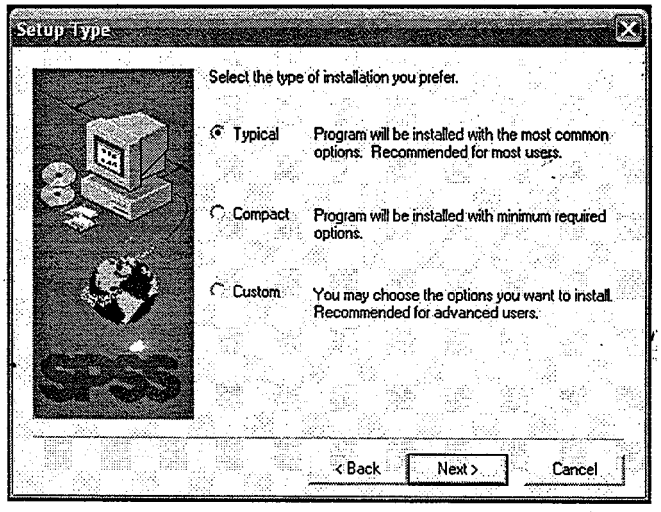
Bước 5: Bạn nhập tên của bạn hay tên nào bạn muốn, còn số series thì bạn nhập số được hướng dẫn trong đĩa cài đặt.

Hình 16.11



Bước 6: Chọn loại cài đặt thông thường là Typical

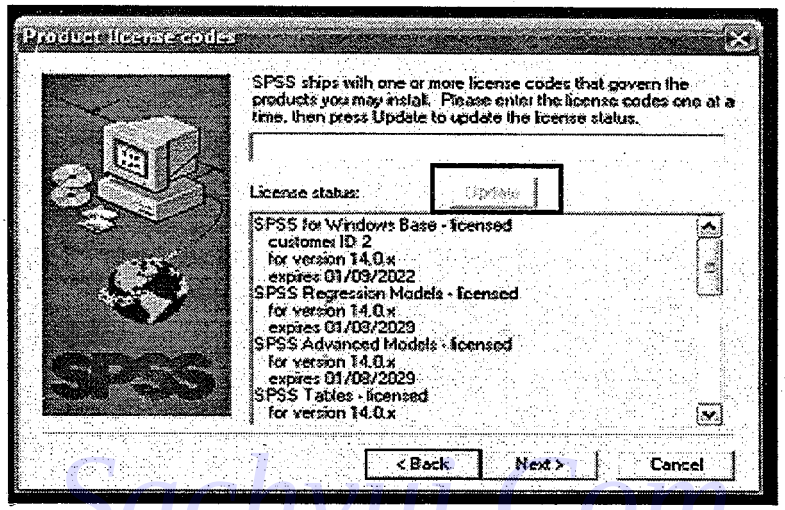
Hình 16.12



Bước 7: Chọn Personal installation trên màn hình xuất hiện kế theo, xong rồi bấm next

Bước 8 : màn hình kế tiếp sẽ yêu cầu nhập 2 license code bạn được cung cấp. Lần lượt nhập license code thứ nhất vào khung màu trắng, và nhấn nút update sát dưới khung. Sau đó nhập tiếp license code thứ hai lần thứ 2 và nhấn update lần nữa. Đảm bảo nhập thật chính xác.

Hình 16.13



Xong một hay nhiều lần nhập license code (tùy theo đĩa cài đặt cung cấp mấy số license), bạn chờ máy cài đặt (vài phút cho đến vài chục phút tùy cấu hình máy và mức độ cài đặt nhiều hay ít thành phần của SPSS) rồi trả lời một số thông tin là bạn có thể cài đặt thành công.

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.

PHÂN TÍCH DỮ LIỆU NGHIÊN CỨU VỚI SPSS

Tập II

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/T2-46-24/HĐ, cấp ngày 20/08/2008.
In xong và nộp lưu chiểu quý 3 năm 2008.