Chương 6: Xử lý và phân tích dữ liệu

CHƯƠNG 6: XỮ LÝ VÀ PHÂN TÍCH DỮ LIỆU

1. Kiểm tra dữ liệu (Explore)

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

- Phát hiện các sai sót
- Nhận dạng dữ liệu để tìm phương pháp phân tích thích hợp và chuẩn bị cho việc kiểm tra giả thuyết

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

- Biểu đồ Histogram
- Sơ đồ cành và lá Stem-and-leaf plot
- Sơ đồ hộp Boxplot

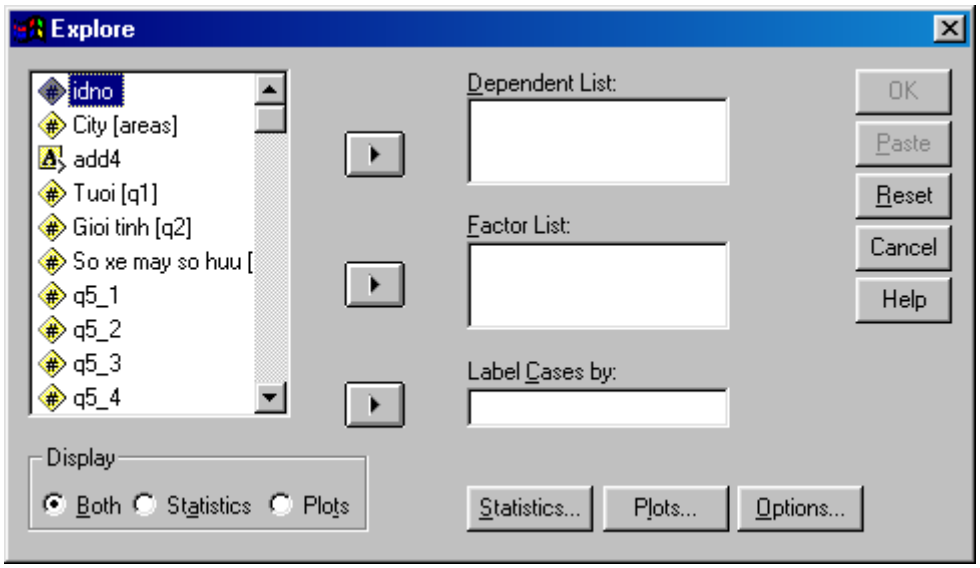
Để ước lượng các giã định được dùng cho việc kiểm nghiệm các giả thuyết, ta dùng các phép kiểm tra sau:

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

Chúng ta thường dùng giá trị trung bình số học để ước lượng độ hội tụ của dữ liệu. Tuy nhiên vì giá trị trung bình bị ảnh hưởng bởi tất cả các giá trị quan sát. Để giảm thiểu những ảnh hưởng của các giá trị bất thường (quá lớn hoặc quá bé), người ta thường loại bỏ các giá trị lớn nhất và các giá trị nhỏ nhất (Outliers) theo cùng một tỷ lệ nào đó. Khi đó giá trị trung bình được gọi là giá trị trung bình giãn lược (Timmed-mean).

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

Để kiểm tra dữ liệu, chọn trên menu **Statistic/Summarize/Explore…** để mở hộp thoại Explore như Hình 6-1:

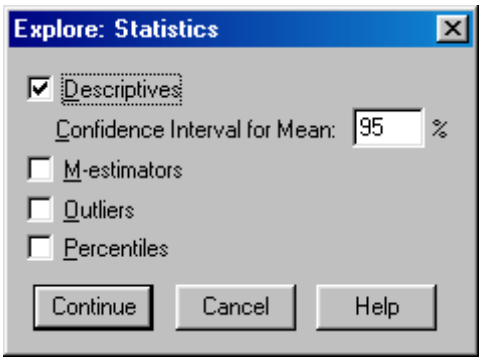


Hình 6-1

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

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

Sử dụng công cụ Statistics cho phép ta lựa chọn các thống kê hiển thị như hộp thoại Hình 6-2:

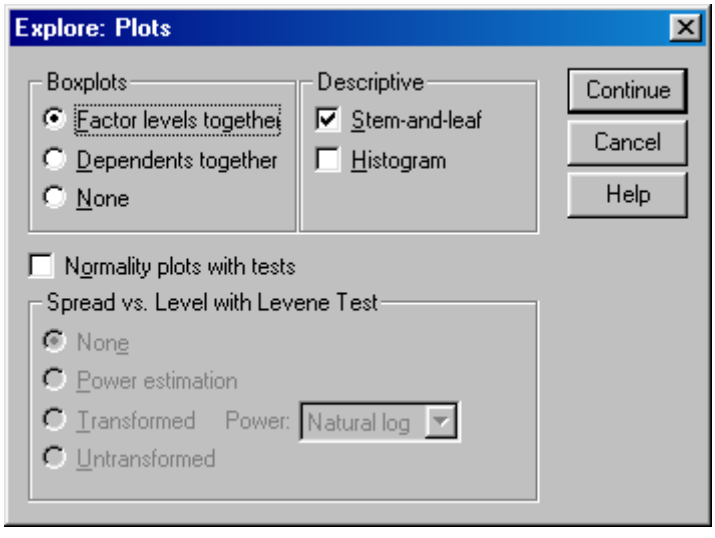


Hình 6-2

- **Descriptives**: Cho phép ta hiển thị các giá trị thống kê như giá trị trung bình, khoảng tin cậy, trung vị, trung bình giãn lược, giá trị nhỏ nhất, lớn nhất, khoảng biến thiên, các bách phân vị
- **M-estimators:** Hiển thị các giá trị trung bình theo 4 loại trọng số
- **Outliers:** Hiển thị các quan sát có 5 giá trị nhỏ nhất và 5 giá trị lớn nhất, gọi là **Extreme Values**
- **Percentiles**: Hiển thị các giá trí bách vị phân

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

Hình 6-3



- **Boxplots:** Điều kiện để hiển thị của Boxplots là ta phải đang quan sát nhiều hơn một biến phụ thuộc (hiển thị trong ô dependent list).
    - **Factor levels together** đưa ra một hiển thị riêng biệt cho mỗi biến phụ thuộc. Trong phạm vi một hiển thị, Boxplots được

hiển thị cho mỗi một nhóm được phân ra theo giá trị của biến điều khiển (factor variable). Dependents together đưa ra một hiển thị riêng biệt theo mỗi nhóm được phân theo các giá trị trong biến điều khiển. Trong phạm vi của hiển thị, boxplots được đưa ra lần lượt cho mỗi biến phụ thuộc

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

- **Kiểm nghiệm Kolmogorov-Smirnov (Lilliefors)**

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

- **Kiểm nghiệm Levene**

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

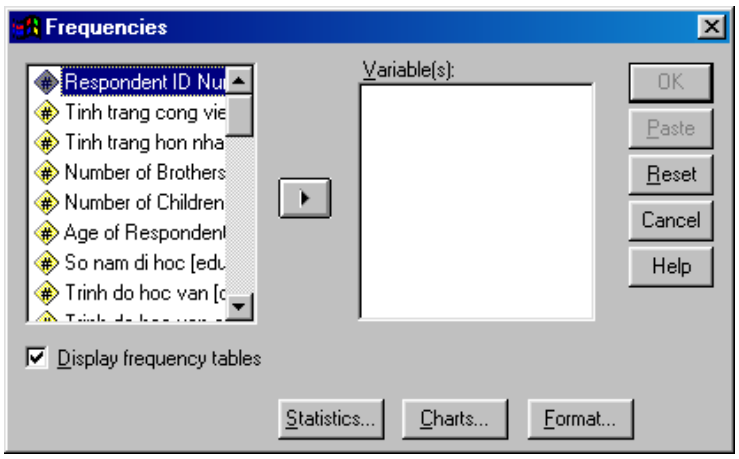
2. Lập bảng phân bố tần suất cho biến một trả lời (Frequencies)

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

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

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

Để tiến hành lập bảng đơn ta chọn công cụ **Statistic/sumarize/frequencies** ta có hộp thoại như Hình 6-4:



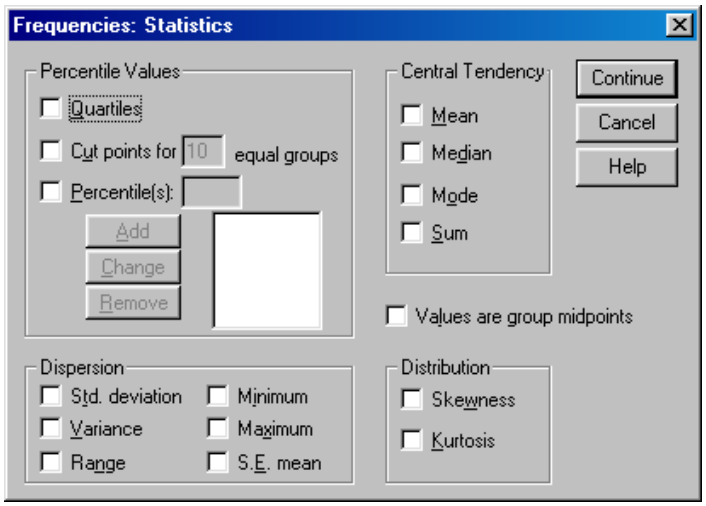
Hình 6-4

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

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

Công cụ statistics để truy suất hộp thoại như Hình 6-5. Trong hộp thoại statistics này sẽ bao gồm các công cụ để đo lường các giá trị thống kê của dữ liệu như vị trí tương đối của các nhóm giá trị hay còn gọi là các phân vị, mật độ tập trung và phân tán của dữ liệu, những đặc tính về phân phối của dữ liệu (Distribution)

Hình 6-5



- **Giá trị bách phân vị (percentile values)**: Được dùng để xác định các ranh giới tương đối của các nhóm từ mẫu quan sát, điều lưu ý là dữ liệu cần quan sát đã được xắp xếp thep thứ tự từ thấp đến cao.
    - Ta có công cụ phân nhánh dữ liệu thành 4 phần bằng nhau gọi là tứ phân vị (quartiles).
    - Hoặc ta có thể chia dữ liệu theo các phần bằng nhau cụ thể bằng cách gõ số phần muốn chia vào công cụ cuts points for equal groups.
    - Hoặc ta có thể xem giá trị ở phân nhánh cụ thể nào đó từ công cụ percentile(s).

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

  Ví dụ như đối với biến chứa các câu trả lời trực tiếp về số tuổi của người trả lời trong một cuộc khảo sát dân số (tuổi người trả lời được ghi trực tiếp từ 18 – 89 tuổi) ta có thể dùng công cụ phân vị dữ liệu để phân các độ tuổi này thành các nhóm nhỏ, ví dụ như ta phân các độ tuổi này bằng phương pháp tứ phân vị (quartiles). Lúc đó tuổi của người trả lời sẽ được phân thành 4 phần sao cho mỗi nhóm tuổi được phân chiếm 25% số lần xuất hiện (tần suất xuất hiện).

- **Đặc tính phân phối (Distribution):** Có hai đại lượng đo lường những đặc tính của sự phân phối dữ liệu là

  (1) Hệ số đối xứng Skewness (Cs) cho ta biết dạng phân phối của các giá trị quan sát Standard Error of Skewness có thể được sử dụng để kiểm nghiệm tính phân phối chuẩn. Một phân phối Skewness không được xem là phân phối chuẩn khi Statndard error của nó nhỏ hơn –2 hoặc lớn hơn 2. Một giá trị dương lớn của Statndard error cho thấy nhánh của phân phối này dài qua bên phải và ngược lại một trị âm chỉ ra nhánh của phân phối này dài qua bên trái

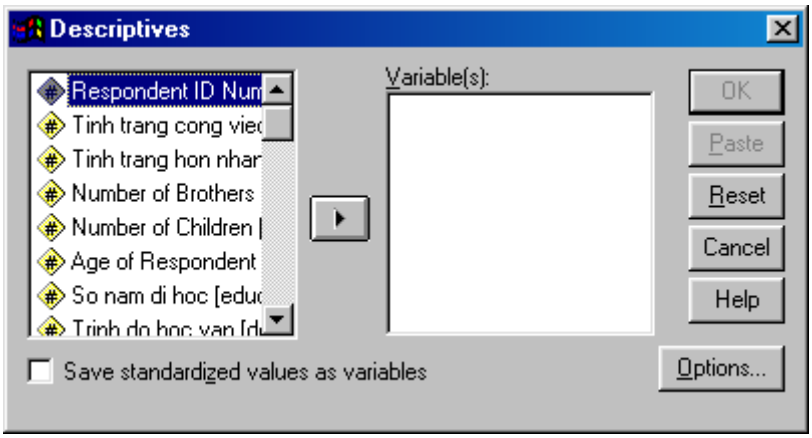
  - Cs = 0: Các quan sát được phân phối một các đối xứng xung quanh giá trị trung bình
  - Cs > 0: Các quan sát tập trung chủ yếu vào các giá trị nhỏ nhất
  - Cs < 0: Các quan sát tập trung chủ yếu vào các giá trị lớn nhất

  (2) Hệ số tập trung Kurtosis (Cc) dùng để so sánh đường cong quan sát với dạng đường cong phân phối chuẩn. Standard Error of Kurtosis có thể được sử dụng để kiểm nghiệm tính phân phối chuẩn. Một phân phối Kurtosis không được xem là phân phối chuẩn khi Statndard error của nó nhỏ hơn –2 hoặc lớn hơn 2. Một giá trị dương lớn của Statndard error cho ta biết hai nhánh của phân phối này dài hơn nhánh của phân phối chuẩn và ngược lại một trị âm chỉ ra hai nhánh của phân phối ngắn hơn phân phối chuẩn

  - Cc > 0: Cho thấy xu hướng tạp trung mạnh của các quan sát xung quanh giá trị trung bình
  - Cc < 0: Cho thấy đường cong có dạng hẹp hơn.

3. Lập bảng mô tả (Descriptive)

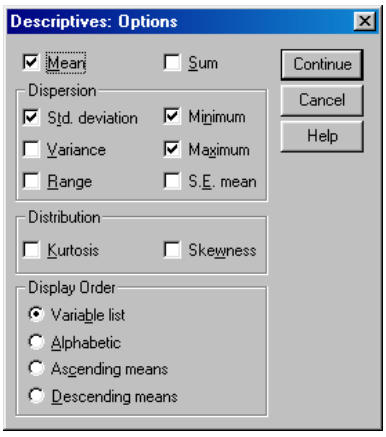
Sử dụng **Statisticts\Summaries\Descriptives** để mở hộp thoại mô tả thống kê như Hình 6-6:



Hình 6-6

Đây là một dạng công cụ khác có thể được dùng để tóm tắc dữ liệu và chỉ cho phép thao tác trên dạng dữ liệu định lượng (thang đo khoảng cách và tỷ lệ). Được dùng để thể hiện xu hướng tập trung của dữ liệu (central tendency) thông qua giá trị trung bình của các giá trị trong biến (mean), và mô tả sự phân tán của dữ liệu thông qua phương sai và độ lệch chuẩn. Chuyển các biến cần tóm tắc vào hộp thoại variables và nhấp thanh options để lựa chọn các thông số thống kê cần mô tả, như giá trị trung bình–mean, giá trị tối thiểu, giá trị tối đa, phương sai và độ lệch chuẩn,… (Hình 6-7)

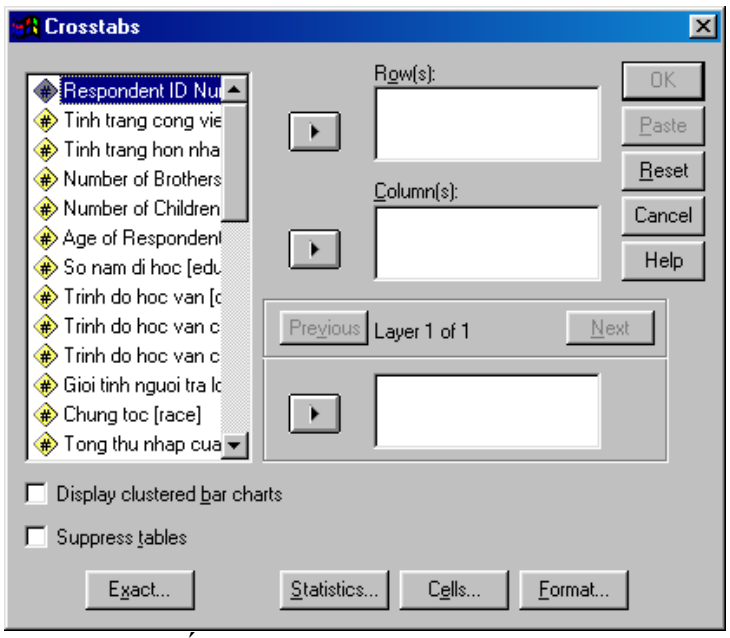
Hình 6-7



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

Bảng nhiều chiều là dạng bảng chéo thể hiện tần suất xuất hiện của một biến này trong mối quan hệ với một hay nhiều biến khác. Bảng chéo còn cung cấp nhiều loại kiểm nghiệm thống kê và đo lường mối quan hệ và tương quan giữa các biến trong bảng. Cấu trúc của bảng và loại dữ liệu (loại thang đó) sẽ quyết định loại công cụ nào được sử dụng để đo lường. Ngoài việc thể hiện mối liên hệ giữa các biến. Bảng nhiều chiều còn giúp ta phát hiện những sai sót trong dữ liệu từ việc phát hiện ra những mối quan hệ vô lý và bất thường giữa hai biến. Chọn trên menu **Statistics/Summaries/Crosstabs** để mở hộp thoại như Hình 6-8:

Hình 6-8



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

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

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

Công cụ Exact cung cấp cho chúng ta hai phương pháp để tính ra mức độ tin cậy cho các phép kiểm nghiệm sử dụng trong bảng chéo, hoặc các phép thử phi tham số (nonparametric). Hai phương pháp này bao gồm phương pháp Exact và

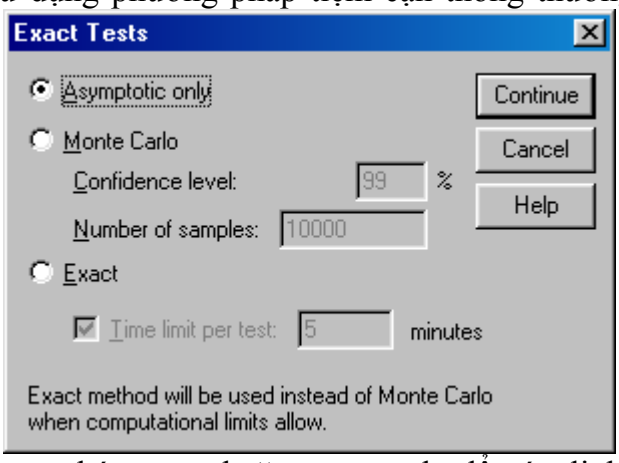
phương pháp Monte Carlo được sử dụng như công cụ để thu được những kết quả chính xác trong trường hợp dữ liệu của chúng ta không đáp ứng được những giả thuyết cần thiết cho một kết quả đáng tin cậy khi sử dụng phương pháp tiệm cận tiêu chuẩn (Standard asymptonic) phương pháp mà kèm theo nó dữ liệu của chúng ta đòi hỏi phải thoả mãn những điều kiện sau:

- Dữ liệu sử dụng có phân phối chuẩn, hoặc kích cở mẫu phải đủ lớn (n>=30)
- Không tồn tại tần suất mong muốn nào của bất kỳ giá trị nào trong bảng chéo nhỏ hơn 5.

Đối với trường hợp dữ liệu không gặp được những yêu cầu như trên. Phương pháp exact hoặc Monte Carlo về độ tin cậy luôn luôn cho ta kết quả đáng tin cậy mà không cần quan tâm đến kích cở mẫu, phân phối của các quan sát cũng như sự cân bằng của dữ liệu (cân bằng về số lượng các giá trị khác nhau trong biến). Chọn công cụ Exact trong hộp thoại Crosstabs ta có hộp thoại con như Hình 6-9.

Hình 6-9

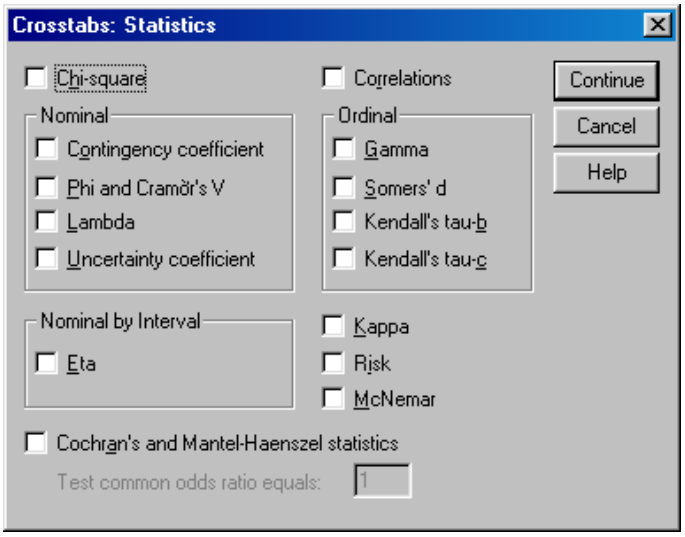
SPSS mặc định là sử dụng phương pháp tiệm cận thông thường (Asymptotic).



Nếu ta sử dụng phương pháp exact hoặc mote carlo để xác định tính độ tin cậy thì cần chú ý các điểm sau:

- Nếu ta lựa chọn phương pháp Monte Carlo, gỏ khoảng tin cậy mong muốn vào công cụ Confidence level, đồng thời cho biết kích cở mẫu được sữ dụng. Sử dụng phương pháp cho ta kết quả nhanh hơn phương pháp exact
- Nếu lựa chọn phương pháp Exact, nhập vào thời gian giới hạn tối đa cho việc tính toán cho mỗi phép thử. Nếu một phép kiểm nghiệm vượt quá thời gian giới hạn tối đa 30 phút, cách tốt hơn nên sử dụng là Moten Carlo

Công cụ Statistics cho phép ta tính các kiểm nghiệm giả thuyết về tính độc lập của các biến, và mối liên hệ giữa các các biến, hệ số tương quan, cũng như đo lường các mối quan hệ đó. (Xem Hình 6-10)



Hình 6-10

❖ Các kiểm nghiệm thống kê – kiểm nghiệm mối quan hệ và tương quan giữa các biến sử dụng trong bảng chéo

- **Kiểm nghiệp Chi-square:**
  - Là một công cụ thông kê sử dụng để kiểm nghiệp giả thuyết cho rằng các biến trong hàng và cột thì độc lập với nhau (H_0). Phương pháp kiểm nghiệm này chỉ cho ta biết được liệu một biến này có quan hệ hay không với một biến khác, tuy nhiên phương pháp kiểm nghiệp này không chỉ ra cường độ của mối quan hệ giữa hai biến mạnh hay yếu (nếu có quan hệ), cũng như không chỉ ra hướng thuận hay nghịch của mối quan hệ này (nếu có quan hệ).
  - Để kiểm nghiệp tính độc lập giữa hai biến cột và hàng, kiểm nghiệp Chi-square sẽ cho ra các kết quả kiểm nghiệp như sau: Pearson chi-square, likelihood-ratio chi-square, and linear-by-linear association chi-square mỗi cái sẽ được sử dụng trong những trường hợp cụ thể
  - Theo định nghĩa hai biến trong bảng là độc lập với nhau nếu như xác suất sao cho một trường hợp quan sát **(case)** rơi vào một trường hợp cụ thể (ví dụ như giới tính là Nam và đang thất nghiệp) là được tạo ra từ các xác suất biên (xác suất cột và xác suất hàng). Ví dụ ta có xác suất một đối tượng quan sát là thất nghiệp là 35/923. Và xác suất để đối tượng quan sát là Nam giới là 452/923. Do hai biến là độc lập, theo lý thuyết xác suất để một trường hợp quan sát vừa là Nam giới vừa là Thất nghiệp thì xác suất trong trường hợp này phải là (452/923) x (35/923) và bằng 0.018. Xác suất này sẽ được sử dụng để ước lượng (estimate) số lượng các trường hợp quan sát mong đợi trong từng phần giao nhau giữa hai biến trên bảng chéo dưới điều kiện hai biến là độc lập với nhau. Do đó để tính toán được

số lượng quan sát mong đợi là Nam giới và thất nghiệp ta chỉ việc nhân xác suất vừa tìm được với tổng số mẫu quan sát (0.018 x 923). (Xem bảng phía chéo phía dưới)

Tinh trang cong viec * Gioi tinh nguoi tra loi Crosstabulation

			Gioi tinh nguoi tra loi		
			Nam	Nu	Total
Tinh trang cong viec	Lam viec toan thoi gian	Count	379	308	687
		Expected Count	336.4	350.6	687.0
		% of Total	41.1%	33.4%	74.4%
	Lam viec ban thoi gian	Count	32	94	126
		Expected Count	61.7	64.3	126.0
		% of Total	3.5%	10.2%	13.7%
	Tam thoi khong di lam	Count	8	22	30
		Expected Count	14.7	15.3	30.0
		% of Total	.9%	2.4%	3.3%
	That nghiep	Count	25	10	35
		Expected Count	17.1	17.9	35.0
		% of Total	2.7%	1.1%	3.8%
	Khac	Count	8	37	45
		Expected Count	22.0	23.0	45.0
		% of Total	.9%	4.0%	4.9%
Total		Count	452	471	923
		Expected Count	452.0	471.0	923.0
		% of Total	49.0%	51.0%	100.0%

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

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

- Khi kết quả thống kê Chi bình phương (χ^2) đủ lớn (Dựa vào lý thuyết phân phối Chi bình phương với độ tin cậy xác định, kích cở mẫu là n, bật tự do-degree of freedom là df=(r-1)(c-1)) ta có thể kết luận bác bỏ giả thuyết độc lập giữa hai biến (H_0). Hoặc sử dụng giá trị P (P-value hay Asymtotic Significance) so sánh với mức ý nghĩa (Significance level) thường là á = 0.05 tương ứng với 95% độ tin cậy, ta có thể kết luận bác bỏ H_0 khi p-value nhỏ hơn hoặc bằng mức ý nghĩa và ngược lại chấp nhận H_0 khi p-value lớn hơn mức ý nghĩa.
- Tuy nhiên để việc kiểm nghiệm này là đáng tin cậy thì các số liệu trong bảng chéo giữa hai biến đang khảo sát phải thỏa mãn một số điều kiện nhất định sau:

  - Không tồn tại ở bất kỳ ô giao nhau giữa hai biến có giá trị mong đợi nhỏ hơn 1.
  - Không vượt quá 20% lượng ô giao nhau giữa hai biến đang khảo sát trong bảng chéo có giá trị nhỏ hơn 5 (đối với bảng 2x2-bảng mà mỗi biến trong bảng chéo chỉ có hai giá trị, phần trăm giới hạn này là 0%)

- Nếu không thỏa mãn các điều kiện trên ta phải tiến hành loại bỏ bớt các giá trị trong một biến mà dữ liệu giao nhau của nó là không đáng kể (quá nhỏ)
- Để kiểm nghiệm tính độc lập giữa hai biến cột và hàng trong bảng chéo, kiểm nghiệp Chi-square sẽ cho ra các kết quả kiểm nghiệp khác nhau như sau: **Pearson chi-square, likelihood-ratio chi-square**, và **linear-by-linear association chi-square.**
- Thông thường để xác định mối quan hệ giữa hai biến trong bảng chéo, việc sử dụng chỉ số nào để kiểm nghiệm tích độc lập giữa hai biến phụ thuộc vào số lượng cột và hàng trong bảng, số mẫu nghiên cứu, tần suất xuất hiện mong muốn của một giá trị trong biến trong điều kiện của biến khác, dạng đo lường của các biến trong bảng (dạng thang đo). Ta có:
  - Dựa vào các hệ số **Pearson Chi-square** và **Likelihood Ratio** ta có thể kiểm nghiệp mối liên hệ giữa hai biến mà không cần quan tâm đến số lượng hàng và cột trong bảng.
  - Hoặc ta có thể dùng chỉ số **Linear-by-linear association** khi mà các biến trong bảng là biến định lượng.
  - Đối với dạng bảng chéo có hai cột và hai dòng **(2X2 tables)** – mỗi biến trong bảng chỉ có hai giá trị, ta dùng các chỉ số **Yate's corrected chi-square** hay còn gọi là **Continuity Correction** đánh giá mối tương quan giữa hai biến trong bảng.
  - Sử dụng chỉ số **Fisher's exact test** khi mà số mẫu nghiên cứu và các giá trị mong đợi nhỏ, thông thường ta sẽ sử dụng chỉ số này khi mẫu trong bảng nhỏ hơn hoặc bằng 20 hoặc tần suất xuất hiện mong muốn trong một phần giao nhau giữa hai biến trong bảng (cell) nhỏ hơn 5.
- Để kết luận mối liên hệ giữa hai biến là độc lập hay phụ thuộc vào nhau (có hay không có tương quan) người ta dựa vào Asymptotic Significance với số mẫu đủ lớn hoặc phân phối là phân phối chuẩn. Đây là chỉ số thống kê để đo lường với mức ý nghĩa (thường là 5%) nhằm đưa ra kết luận phản bát hay chấp nhận giả thuyết ban đầu (Hai biến là độc lập với nhau). Ta có thể kết luận giữa hai biến tồn tại một mối quan hệ với nhau khi mà Asym. Sig. nhỏ hơn mức ý nghĩa và ngược lại.
- Đối với kiểm nghiệm Chi-square ta chỉ có thể xác định giữa hai biến có hay không tồn tại một mối quan hệ. Tuy nhiên để đo lường cường độ của

các mối quan hệ này đòi hỏi các công cụ thống kê khác sẽ được đề cập sau đây.

- **Correlation:**
  - Dùng để đo lường mối tương quan giữa hai biến thứ tư hoặc khoảng cách. Việc đo lường mối tương quan giữa hai biến thứ tự này chủ yếu dự vào hai hệ số Spearman's correlation coefficient rho và Pearson correlation coefficient. Trong đó:
    - Spearman's rho được dùng để đo lường mối quan hệ giữa hai biến thứ tự (các biến này hầu hết đều được xắp xếp từ thấp nhất đến cao nhất).
    - Khi các biến trong bảng là các biến định lượng ta sử dụng hệ số Pearson correlation coefficient để đo lường mối quan hệ tuyến tính giữa các biến này.
  - Các giá trị của hệ số tương quan biến thiên từ –1 đến 1, dấu cộng hoặc trừ chỉ ra hương tương quan giữa các biến (thuận hay nghịch), giá trị tuyệt đối của chỉ số này cho biết cường độ tương quan giữa hai biến, giá trị này càng lớn mối tương quan càng mạnh.
- **Một số đo lường mối tương quan khác giữa hai biến**
  - **Giữa hai biến định danh:**
    - Để đo lường mối quan hệ giữa hai biến biểu danh. Sử dụng các hệ số **Phi (coefficient)** và **Crắmr's V, Contingency coefficient** để đo lường nếu dựa vào kết quả kiểm nghiệm Chi-bình phương. Ở đây các hệ số này sẽ bằng 0 nếu và chỉ nếu hệ số **Pearson chi bình phương** bằng 0. Do đó người ta sử dụng các thông số này để kiểm nghiệm giả thuyết cho rằng các hệ số này đều bằng 0 - điều này tương đương với giả thuyết độc lập giữa hai biến, hay hai biến không có môi quan hệ với nhau. Ta sẽ từ chối giả thuyết này
    - **Phi:** Chỉ dùng cho dạng bảng 2x2 tables, hệ số phi coefficient này biến thiên từ -1 đến +1. Do đó hệ số này ngoài khả năng chỉ ra mối quan hệ và cường độ của mối quan hệ nó còn chỉ ra hướng của mối quan hệ đó
    - **Cramer's V và Contingency coefficient** (hệ số ngẫu hiên): Được sử dụng cho bảng mà số cột và hàng là bất kỳ, giá trị kiểm nghiệm biến thiên từ 0 đến 1, với giá trị 0 chỉ ra không có mối quan hệ giữa các biến
    - Ngoài ra còn có các hệ số đo lường trực tiếp như **Lambda (symmetric and asymmetric lambdas and Goodman and Kruskal's tau),** và **Uncertainty coefficient.** Là các đo lường không dựa vào giá trị Chi-square để tính toán, và không quan tâm đến tính đối xứng của phân phối chuẩn. Các giá trị của hệ số này cũng biến thiên từ 0 đế 1 và được dùng để đo lường khả năng dự báo của một biến (biến độc lập) đối với một biến khác (biến phụ thuộc). Với giá trị

0 nhận được có ý nghĩa rằng những kiến thức về biến độc lập không giúp ích gì cho việc dự báo những khả năng xảy ra của biến phụ thuộc, và giá trị 1 cho biết khi ta biết được những thông tin về biến độc lập thì nó sẽ giúp ta xác định được một cách hoàn hảo các khả năng xảy ra cho biến phụ thuộc.

- Việc lựa chọn biến nào là biến độc lập và biến nào là biến phụ thuộc tùy thuộc vào vấn đề cụ thể mà ta đang khảo sát
- Hệ số **Asymptotic Std. Error** có thể được dùng để định ra khoảng tin cậy (95%) cho các tham số đo lường (Value +(-) 2*Asymptotic std. Error)

- **Sử dụng Odds Ratio cho bảng hai cột hai hàng (2x2 tables)**
  - Để đo lường mối tương quan giữa hai biến cho loại bảng này người ta có thể sử dụng các kết quả thống kê Yates' corrected chi – bình phương và Fisher's exact test. Các kết quả này được dùng để kiểm nghiệm giả thuyết cho rằng các tỷ lệ giữa các giá trị trong hai biến này là ngang bằng nhau (ví dụ như tỷ lệ người nam đi bảo tàng thì ngang bằng với tỷ lệ người nữ đi bảo tàng), tương tự với các kết quả thống kê chi – bình phương khác ta sẽ từ chối giả thuyết H_0 khi p-value nhỏ hơn mức tin cậy.
  - Ngoài phương pháp trên ta còn có thể sử dụng phương pháp **odds ratio** và **relative risk** để đo lường mối liên hệ giữa hai đặc tính. Thông thường một trong hai đặc tính đó xuất hiện trước (ví dụ như biến chứa đặc tính có hút thuốc hay không) và sau đó là sẽ dẫn đến một đặc tính khác xuất hiện theo sau (ví dụ biến chứa đặc tính có bị bệnh lao phổi hay không). Ta gọi biến chứa đặc tính xuất hiện trước là biến nhân tố (factor) và biến theo sau là biến sự kiện (event). Ta có hai phương pháp tính như sau:

    (1) Relative risk:

		Biến sự kiện			
		Yes	No	Tỷ lệ rủi ro risk	Tỷ lệ rủi ro tương đối Relative risk
Biến nhân tố	Yes	a	b	a/(a+b)	a(c+d)
	No	c	d	c/(c+d)	c(a+b)

Phương pháp này bắt đầu với biến nhân tố và theo sau đó ta đếm số mỗi sự kiện xuất hiện trong mỗi nhóm nhân tố. Tỷ lệ rủi ro được tính riêng biệt cho từng nhóm nhân tố và tỷ lệ rủi ro tương ứng là tỷ số giữa hai tỷ lệ rủi ro của từng nhóm nhân tố

(2) Odds ratio:

		Biến nhân tố			
		Yes	No	odds	Tỷ lệ odds
Biến sự kiện	Yes	a	b	a/b	ad
	No	c	d	c/d	cb

Phương pháp này bắt đầu với biến sự kiện. Với một sự kiện (ví dụ bị bệnh lao phổi) thì tỷ lệ giữa người hút thuốc đối với người không hút thuốc là bao nhiêu, gọi là odd. Sau đó ta lập tỷ lệ các odds này.

- Cả hai phương pháp này đều có cách kiểm nghiệp kết quả giống nhau. Cả Tỷ lệ Odds và relative risk đều nhận giá trị 1 khi các tỷ lệ này là giống nhau. Và để kiểm nghiệm giã thuyết ban đầu cho rằng các tỷ số này là như nhau (H_0) - từ chối hay chấp nhận ta dựa vào khoảng tin cậy (95%) xem xem giá trị 1 có nằm trong khoảng tin cậy đó hay không. Nếu giá trị 1 không nằm trong khoảng tin cậy 95% ta từ chối giả thuyết H_0, và có thể xem giá trị trong ô (value) là tỷ số diễn giải. Nếu giá trị 1 nằm trong khoảng tin cậy 95%, không cần quan tâm đến các giá trị trong cột value, bởi vì kiểm nghiệm cho ta kết quả chấp nhận giả thuyết hai tỉ lệ odds hoặc relative của hai giá trị là như nhau
- Chú ý phương pháp Odds ratio luôn luôn lấy tỷ số odd ở hàng thứ nhất chia cho hàng thứ hai, và sự kiện cần quan tâm luôn luôn nằm ở cột thứ nhất. Còn đối với phương pháp Relative risk bất cứ cột nào cũng có thể đại diện cho sự kiện cần quan tâm (SPSS sẽ đưa ra các kết quả khác nhau để ước lượng cho mỗi cái

- **Dùng Kappa để đo lường sự đồng ý giữa hai biến trong một bảng có cùng số lượng hàng và cột**

  - **Kappa** dùng để đo lường mức độ đồng ý giữa những đo lường của hai nhóm đánh giá đối với cùng một tiêu chí nào đó. Giá trị 1 chỉ ra sự hoàn toàn đồng ý giữa hai nhóm, giá trị 0 chỉ ra sự đồng ý chỉ là một sự ngẫu hiên.Hoặc ta dùng p-value để kiểm nghiệm giả thuyết ban đầu H_0 cho rằng các giá trị đo lường này là bằng không. **Kappa** chỉ thích ứng với những bảng mà các biến được sử dụng trong bảng có cùng số giá trị trong biến.

- **Đo lường mối tương quan giữa các biến thứ tự và biến định lượng**

  (1) Nominal by Interval: Dùng đo lường mối tương quan giữa biến biểu danh và biến định lượng trong bảng chéo. Sử dụng hệ số **Eta.**

  (2) Correlation: Dùng để đo lường mối tương quan giữa hai biến thứ tư hoặc khoảng cách. Việc đo lường mối tương quan giữa hai biến thứ tự này chủ yếu dự vào hai hệ số **Spearman's correlation coefficient rho** và **Pearson correlation coefficient**. Trong đó **Spearman's rho** được dùng để đo lường mối quan hệ giữa hai biến thứ tự (các biến

này hầu hết đều được xắp xếp từ thấp nhất đến cao nhất). Khi các biến trong bảng là các biến định lượng ta sử dụng hệ số **Pearson correlation coefficient** để đo lường mối quan hệ tuyến tính giữa các biến này. Các giá trị của hệ số tương quan biến thiên từ –1 đến 1, dấu cộng hoặc trừ chỉ ra hương tương quan giữa các biến (thuận hay nghịch), giá trị tuyệt đối của chỉ số này cho biết cường độ tương quan giữa hai biến, giá trị này càng lớn mối tương quan càng mạnh.

(3) Ordinal: Dùng đo lường mối tương quan giữa các biến trong bảng chéo trong đó các biến ở cột và dòng là các biến thứ tự, bao gồm các hệ số sau:

(1) Somers' d: Đo lường mối tương quan phi đối xứng giữa hai biến thứ tự, giá trị biến thiên từ –1 đến 1.

(2) Gamma: Đo lường mối tương quan đối xứng giữa hai biến thứ tự, giá trị biến thiên từ –1 đến 1.

(3) Kendall's tau-b và Kendall's tau-c: Đo lường các mối quan hệ phi tham số giữa hai biến thứ tự, biến thiên từ –1 đế 1

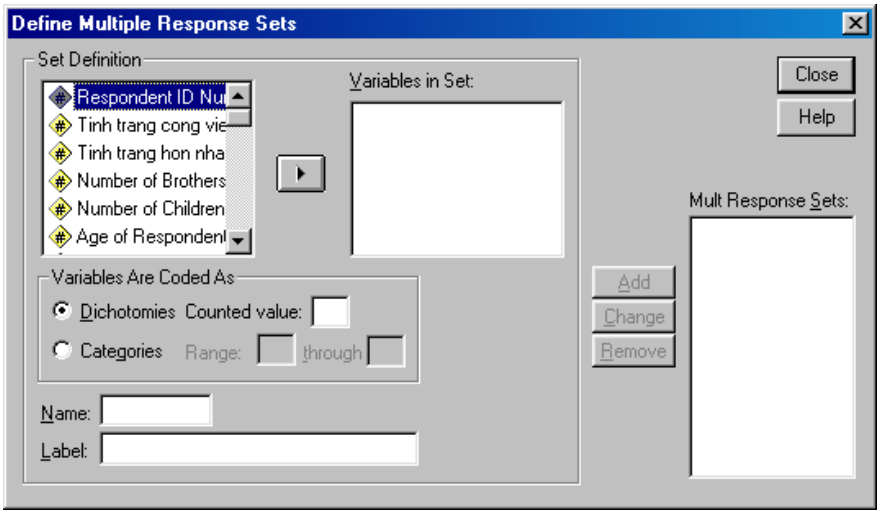
Phần này có thể xem thêm ví dụ trong phần phụ lục

5. Lập bảng cho biến nhiều trả lời:

5.1. Định nghĩa nhóm biến nhiều trả lời (define multi response sets)

Trong câu hỏi nhiều trả lời sẽ bao gồm nhiều biến chứa đựng các trả lời có thể có, những biến này gọi là biến sơ cấp. Do đó để xữ lý, chúng ta phải gộp các biến sơ cấp này thành một biến gộp chứa các biến sơ cấp. Sau đó trong các phân tích thống kê liên quan đến câu hỏi nhiều trả lời, chúng ta sẽ dùng biến gộp này thay thế cho tất cả các biến sơ cấp. Biến gộp chứa đựng toàn bộ các giá trị trong các biến sơ cấp của một câu hỏi nhiều trả lời. Ví dụ như câu hỏi về nhận biết sản phẩm, người tra lời có thể liệt kê ra nhiều nhãn hiệu mà họ biết, do đó ta phải khai báo đủ lượng biến để chứa đựng các nhãn hiệu được liệt kê từ người trả lời, đây là các biến sơ cấp. Tuy nhiên khi xữ lý ta không thể xữ lý riêng biệt các biến này, vì nó không đại diện đầy đủ cho tất cả các nhãn hiệu được nhận biết. Do đó khi tiến hành phân tích câu hỏi nhận biết sản phẩm này ta phải tiến hành gộp các biến sơ cấp thành một biến gộp chứa đựng tất cả các nhãn hiệu được liệt kê.

Để tiến hành gộp các biến sơ cấp này ta chọn menu **Statistics/Multiple Response/Define sets**… để mở hộp thoại **Define Multiple Response Sets** nhu Hình 6-11:



Hình 6-11

Chọn tất cả những biến sơ cấp liên quan đến một câu hỏi nhiều trả lời ở hộp thoại Set Definition bên trái chuyển sang hộp thoại Variables in Set bên phải, ví dụ ta có 10 biến đơn chứa đựng các nhãn hiệu được nhận biết, ta phải chọn tất cả 10 biến này từ hộp thoại Set Definition và chuyển sang hộp thoại Variable in Set. Sau đó chỉ định cách mã hóa các biến đó (dichotomy hay category); dãy giá trị mã hóa (Range …Through) xác định khoảng biến thiên cho các giá trị trong biến gộp; xác định tên và gán nhãn cho biến gộp. Sau đó ấn thanh Add để đưa tên nhóm vừa xác định vào hộp Multi Response Sets. Sau khi tiến hành khai báo biến gộp xong mọi sử lý phân tích các biến nhiều trả lời sẽ được tiến hành trên các biến gộp đã được khai báo trong Multi Response Sets.

Trong khung Variable Are Code As, chúng ta có thể chọn một hay hai mục sau đây tùy theo phương pháp mã hóa:

- **Dichotomies:** Đây là trạng thái mặc định, và chúng ta nhập giá trị cần đếm vào hộp Counted Value. Kết quả chỉ hiển thị duy nhất giá trị đếm vừa khai báo
- **Category:** Mỗi biến sơ cấp có nhiều hơn hai giá trị, và chúng ta nhập các giá trị nhỏ nhất và lớn nhất của dãy giá trị mã hóa vào các ô Range và thourgh (nên khai báo một khoảng cách càng rộng càng tốt)

Chúng ta đặt tên cho nhóm đa biến (tối đa 7 ký tự) và nhãn (tối đa 40 ký tự) vào các hộp Name và Label. Lưu ý là tên của các nhóm đa biến chỉ được sử dụng trong các thủ tục xữ lý biến nhiều trả lời mà thôi. Để loại bỏ và sữa đổi việc định nghĩa một nhóm biến đa trả lời nào đó ta di chuyển vệt sáng đến tên nhóm đó và nhấn thanh remove để loại bỏ và thanh Change để thay đổi.

5.2. Lập bảng cho biến nhiều trả lời

Để tiến hành lập bảng cho các biến nhiều trả lời, ta sử dụng các tên nhóm đa biến đã được định nghĩa bằng công cụ Define Multi Response Sets đã được đề cập ở phần trên sau đó vào Statistics\Multiple response và chọn Frequencies hoặc Crosstabs tùy theo nhu cầu lập bảng một chiều hay đa chiều. Tuy nhiên trong các công cụ Frequencies và Crosstabs sử dụng cho biến nhiều trả lời chỉ mô tả tần suất xuất hiện của các giá trị trong biến gộp và các tỷ lệ % nhưng không có các phương pháp kiểm nghiệm thống kê kèm theo.

6. Custom Table

Ngoài ra khi chúng ta tiến hành lập bảng mô tả thống kê cho kết quả cuối cùng của vấn đề nghiên cứu có thể dùng các công cụ trong **statistics\custom table** để tạo ra các bảng biểu, có thể là bảng một chiều, bảng nhiều chiều hoặc các bảng biểu mô tả thống kê tùy theo yêu cầu của vấn đề nghiên cứu.

Các loại bảng này cho phép ta tạo ra các bảng biểu đẹp hơn. Tuy nhiên ngoài việc truy suất các giá trị đếm, tỷ lệ phần trăm thì nó không cung cấp thêm cho ta phương pháp kiểm nghiệm thống kê nào khác kèm theo

- Bảng biểu thể hiện tần số xuất hiện (**Tables of frequencies**): Cho phép chúng ta tạo ra những bảng biểu thể hiện tần số xuất hiện của một hay nhiều biến đơn

- Dạng bảng biểu cơ bản (**Basic tables**): Thể hiện các dữ liệu nghiên cứu theo dạng bảng chéo (**cross-tabulation**) giữa hai biến hoặc giữa một biến và một nhóm các biến.
- Dạng bảng đa biến (**Multiple response tables**): Giống như basic tables thể hiện tần suất xuất hiện và bảng chéo, tuy nhiên dạng bảng biểu này cho phép ta xây dựng bảng biểu cho các câu trả lời đa biến
- Dạng bảng biểu tổng hợp (**General tables**): Giống như bảng biểu cơ bản và đa trả lời. Các dữ liệu được thể hiện dưới dạng bảng chéo, tuy nhiên ở dạng bảng biểu này cho phép người phân tích thể hiện mối liên hệ giữa một biến với nhiều biến khác trên cùng một bảng.

7. So sánh các giá trị trung bình

Có nhiều phép kiểm nghiệp được sử dụng trong SPSS:

- Nếu so sánh giá trị trung bình của mẫu với một giá trị cố định nào đó ta sử dụng phép kiểm nghiệm t một mẫu (One-sample t test).
- Nếu so sánh giá trị trung bình của một nhóm các trường hợp quan sát với một nhóm quan sát khác, ta sữ dụng kiểm nghiệm t mẫu độc lập (Independent-sapmles t test).
- Để so sánh giá trị trung bình của hai biến được khảo sát từ cùng một mẫu ta sử dụng kiểm nghiệp t theo từng cặp mẫu (Paired-samples t test).
- Hoặc với trường hợp ta có nhiều hơn hai mẫu độc lập cần kiểm nghiệm trung bình, ta có thể dùng ANOVA một chiều (One-way ANOVA).

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

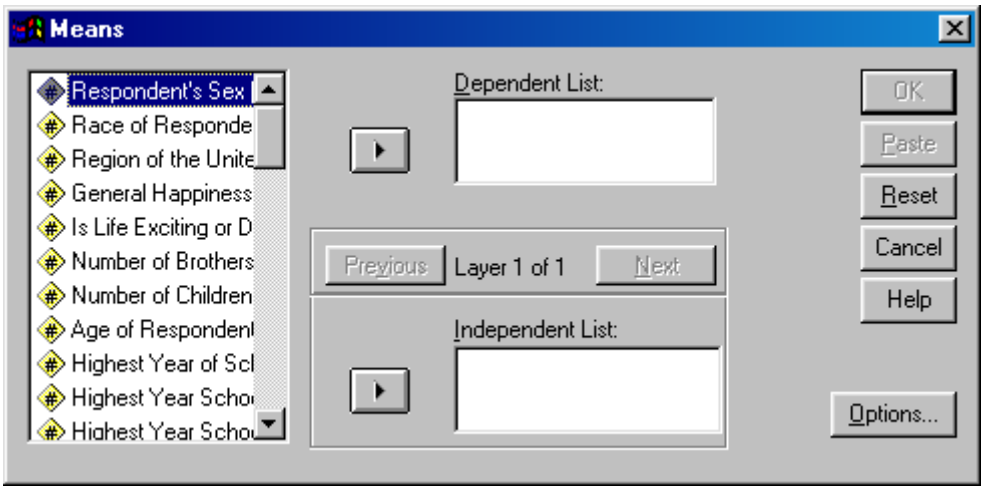
7.1. Means

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

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

phân phối chuẩn và thích hợp cho các biến định lượng với phân phối đối xứng. Các đại lương khác như **Media,** và **range** thì thích hợp cho các biến định lượng mà ta không biết liệu nó có thoả mãn các điều kiện về phân phối chuẩn hay không. Ta có thể lựa chọn **ANOVA** và **eta** để thực hiện việc phân tích sự biến thiên một chiều cho mỗi biến độc lập. **Eta** và **eta bình phương** cho phép đo lường các mối tương quan.

Để thực hiện công cu này ta chọn **Compare Means/Means…. Từ Menus**, ta có hộp thoại như hình 6-12**.**



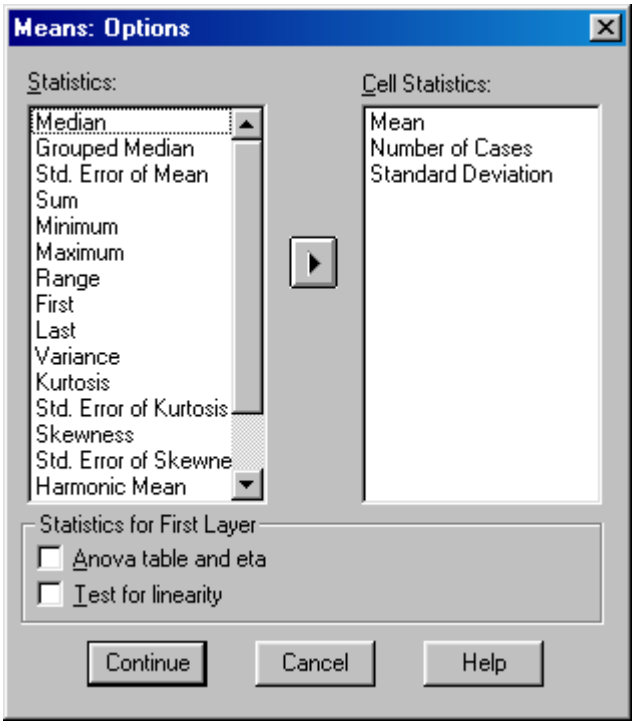
Hình 6-12

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

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

Công cụ **Options** (Hình 6-13). Cho phép ta lựa chọn các đại lượng thống kê cần khảo sát và ANOVA, Eta, và Eta bình phương (sẽ được đề cập chi tiếc về ý nghĩa ở phần sau)

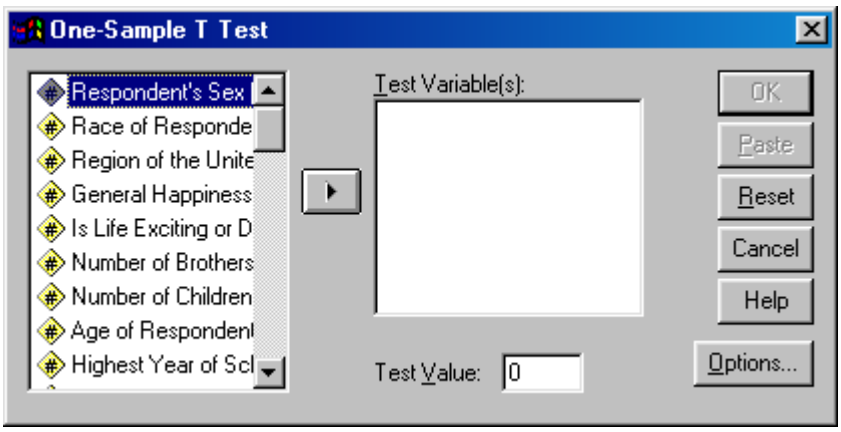
Hình 6-13



7.2. Kiểm nghiệp t-một mẫu

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

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



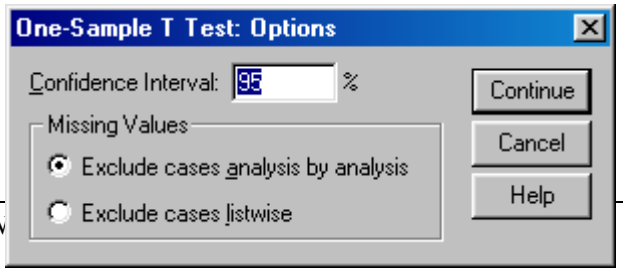
Hình 6-14

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

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

- **Exclude cases analysis by analysis**. Mỗi kiểm nghiệm T sử dụng toàn bộ các trường hợp (cases) chứa đựng giá trị có ý nghĩa đối với biến được kiểm nghiệm. Đặc điểm là kích thương mẫu luôn thay đổi.
- **Exclude cases listwise.** Mỗi kiểm nghiệm T sử dụng chỉ những trường hợp có giá trị đối với toàn bộ tất cả các biến được sử dụng trong bất kỳ kiểm nghiệm T test nào. Kích thước mẫu luôn không đổi

Hình 6-15



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

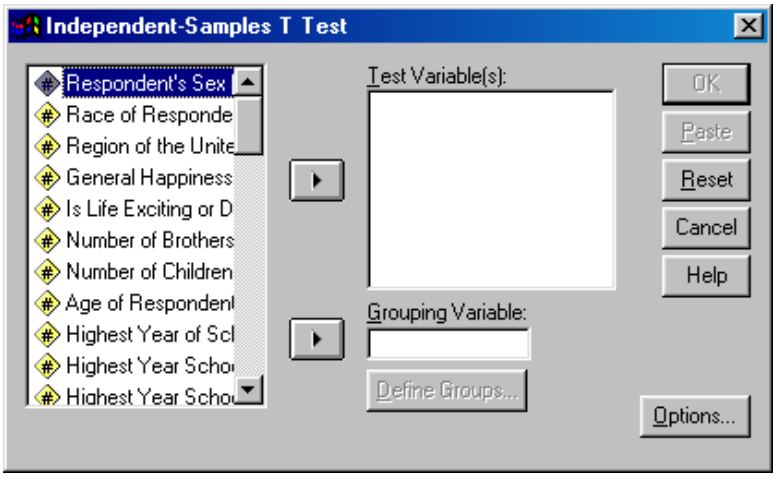
7.3. Kiểm nghiệp t hai mẫu độc lập

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

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

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

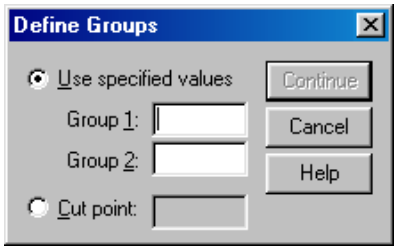
Để thực hiện việc so sánh này ta vào **Compare means\Independent sample t-test….** Từ Menus ta được hộp thoại như hình 6-16:



Hình 6-16

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

Di chuyển vệt tối đến biến dùng để định ra các nhóm cần so sánh với nhau (thường là biến định danh) di chuyển vào hộp thoại **Gouping variable.** Công cụ **Define Groups…** cho phép ta định ra hai nhóm cần so sánh với nhau, như hình 6-17.



Hình 6-17

Có hai cánh định nhóm so sánh:

- Sữ dụng con số cụ thể, nhập hai giá trị đại diện cho hai nhóm cần so sánh trong biến vào ô **group 1 và group 2,** ví dụ so sánh thời gian tự học của hai nhóm sinh viên năm nhất và sinh viên năm cuối nằm trong biến loại sinh viên với 4 nhóm sinh viên được mã hóa như sau sinh viên năm nhất: 1, sinh viên năm hai: 2, sinh viên năm ba: 3, sinh viên năm cuối: 4. Ta nhập giá trị 1 vào Group 1 và nhận giá trị 4 vào group 2. Lúc đó thời gian tự học trung bình sẽ được so sánh giữa hai nhóm sinh viên năm nhất và sinh viên năm cuối.

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

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

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

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

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

Công thức tính t:

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

Với:

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

Với x_i: Giá trị trung bình của nhóm i
n_i: Số các quan sát trong nhóm i
S_i: Phương sai mẫu trong nhóm i

Bật tự do trong kiểmnghiệm phương sai hợp nhất bằng

$df = (n_1 + n_2 - 2)$

Bật tự do trong kiểmnghiệm phương sai riêng biệt bằng:

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

7.4. Kiểm nghiệm t theo từng cặp mẫu

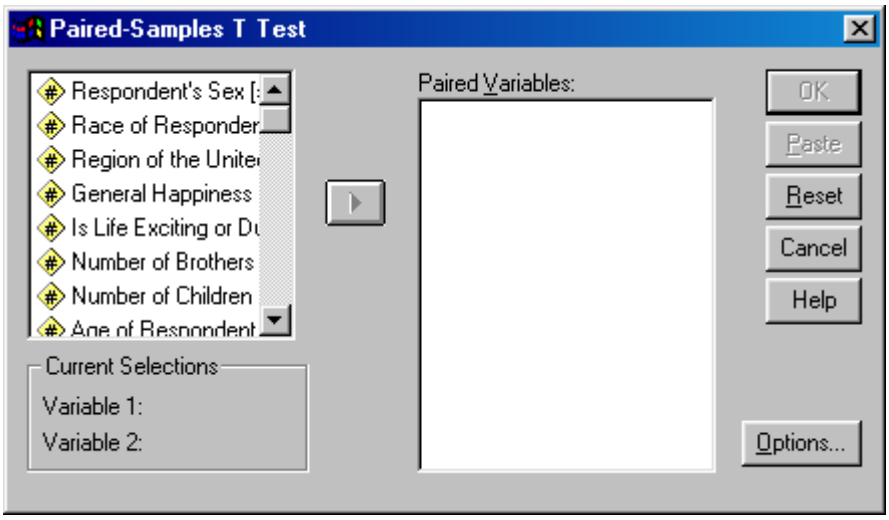
Đây là dạng kiểm nghiệp dùng cho hai biến trong cùng một mẫu có liên hệ với nhau, dữ liệu dạng thang đó khoảng cách hoặc tỷ lệ. Nó tính toán sự khác biệt giữa các giá trị của hai biến cho mỗi trường hợp và kiểm nghiệm xem giá trị trung bình các khác biệt có khác 0 hay không. Giả thuyết ban đầu được đưa ra là giá trị trung bình của các khác biệt là bằng 0. Và ta sẽ loại bỏ giả thuyết này trong trường hợp kiểm nghiệm cho kết quả **Sig.** nhỏ hơn mức ý nghĩa (0.05)

Lợi điểm của việc sử dụng kiểm nghiệm T theo từng cặp là ta loại trừ được những yếu tố tác động bên ngoài vào nhóm thử. Ví dụ ta khảo sát sự ưa thích của hai loại nước hoa chuẩn bị tung ra thị trường. Kết quả kiểm nghiệp trên cùng một nhóm mẫu sẽ cho những thông tin xác thực hơn về sự ưa thích của mùi vị hai loại nước hoa này, đồng thời tập trung vào sự khác biệt tư nhiên của hai loại nước hoa này. Nếu ta tiến hành so sánh giữa hai nhóm mẫu độc lập với nhau sẽ cho ra những kết quả khác biệt do những tác nhân khác với bản thân sự khác biệt

của hai loại nước hoa này như sự khác biệt về con người, về nhận thức, về kinh nghiệm cũng như các yếu tô bên ngoài khác. Phương pháp này thích ứng cho việc kiểm nghiệm sản phẩm. Phương pháp này kiểm nghiệm giả thuyết cho rằng sự khác biệt giữa hai trung bình mẫu là bằng không. Ta từ chối giả thuyết này khi mức ý nghĩa của ta (significante) là nhỏ hơn mức ý nghĩa (thường là 5%).

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

Để thực hiện việc so sánh này ta vào **Compare means\Paired-samples t-test….** Từ Menus ta được hộp thoại như hình 6-17:



Hình 6-17

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

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

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

Trong kiểm nghiệm hai mẫu độc lập đã đề cập ở phần trước SPSS chia các giá trị của một biến đơn thành hai nhóm dựa trên một biến kiểm soát và sau đó tiến hành so sánh trung bình trong biến đơn giữa hai nhóm đó với nhau. Đối với kiểm nghiệm cặp, giá trị trung bình các giá trị trong hai biến được so sánh với nhau. Kiểm nghiệm loại này được sử dụng để kiểm nghiệm xem trung bình của hai đo lường là khác biệt hay ngang bằng nhau, hay nói cách khác kiểm nghiệm xem có hay không trung bình của các giá trị khác biệt giữa hai biến trên mỗi trường hợp quan sát là khác 0

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

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

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

Với SD: Độ lệch tiêu chuẩn của các sai biệt

n : Số lượng các quan sát (mẫu)

8.5. Phân tích phương sai một chiều (One way ANOVA)

Các phép so sánh đề cập ở phần trên chỉ cho phép ta so sánh trung bình hai tổng thể dựa trên mẫu từng cặp phối hợp hoặc hai mẫu độc lập. Trong phần này phương pháp kiểm định sẽ mở rộng cho trường hợp so sánh trung bình của nhiều tổng thể được xây dựng trên việc xem xét các biến thiên (phương sai) của các giá trị quan sát trong nội bộ từng nhóm (mẫu) và giữa các nhóm (mẫu) với nhau. Ở đây ta đề cập đến phân tích phương sai một yếu tố là trường hợp chỉ có một yếu tố (biến kiểm soát) được xem xét nhằm xác định ảnh hưởng của nó đến một yếu tố khác. Yếu tố được xem xét ảnh hưởng được dùng để phân loại các quan sát thành các nhóm nhỏ khác nhau.

Một cách tổng quát, giả sử ta có k nhóm (mẫu) n_1, n_2, …, n_k quan sát được chọn ngẫu nhiên độc lập từ k tổng thể (n_1, n_2, …, n_k có thể khác nhau về kích thước). Gọi μ_1, μ_2, …, μ_k là các trung bình của k tổng thể, x_{ij} là quan sát thứ j của nhóm thứ i. Ta có thể mô tả các quan sát của k nhóm như sau:

Nhóm			
1	2	…	K
X_{11}	X_{21}	…	X_{K1}
X_{12}	X_{22}	…	X_{K2}
…	…	…	…
X_{1n1}	X_{2n2}	…	X_{KnK}

Với giả định các tổng thể có phân phối chuẩn, có phương sai bằng nhau, các sai số là độc lập với nhau, phân tích phương sai một yếu tố kiểm nghiệm giả thuyết ban đầu như sau: H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$. Ta thấy ở đây là việc so sánh giữa các giá trị trung bình, vậy phân tích phương sai nghe như là một sai sót. Tuy nhiên việc phân tích phương sai ở đây dựa trên thông số thống kê F, với F là tỷ số giữa biến thiên giữa trung bình các nhóm trên biến thiên giữa các quan sát trong nội bộ nhóm:

$$F = \frac{\text{Biến thiên giữa trung bình các nhóm}}{\text{Biến thiên giữa các giá trị quan sát trong nội bộ nhóm}}$$

Nếu các giá trị trung bình của các nhóm khác biệt nhau nhiều, đặc biệt trong mối quan hệ với sự biến thiên của nội bộ từng nhóm, giá trị F thu được sẽ lớn và khi đó giả thuyết H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$. sẽ bị từ chối. Và nếu ta quan sát việc phân tích phương sai một yếu tố cho hai nhóm thì kết quả thống kê F tính được sẽ chính bằng bình phương kết quả thống kê t trong kiểm nghiệm t cho hai mẫu độc lập

- **Các bước phân tích phương sai một yếu tố để kiểm nghiệm sự ngang bằng giữa các giá trị trung bình của k tổng thể**

Phân tích phương sai một yếu tố để kiểm nghiệm giả thuyết H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ được tiến hành thông qua các bước sau:

Bước 1: Tính giá trị trung bình x_i cho từng nhóm và x chung cho tất cả các nhóm

$$\bar{x}_i = \frac{\sum_{i=1}^{n_i} x_{ij}}{n_i} (i = 1,2,...,k)$$

$$\bar{x} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}}{n}$$

Hoặc

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \ddot{x}_i}{n} (n = \sum_{i=1}^{k} n_i)$$

Bước 2: Tính các đại lượng thể hiện sự biến thiên trong nội bộ từng nhóm (SSW) và giữa các nhóm (SSG)

Gọi SS là đại lượng thể hiện sự biến thiên trong nội bộ từng nhóm, ta có:

Ta co tổng cộng các biến thiên trong nội bộ từng nhóm là:

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

$$SSW = SS_1 + SS_2 + ... + SS_k = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Nói một cách đơn giản SSW là tổng bình phương các chênh lệch giữa từng quan sát với trung bình của nhóm mà quan sát đó thuộc về (within-groups sum of squares). SSW là những biến thiên không do yếu tố kiểm soát (yếu tố dùng để phân chia các nhóm) gây ra.

Đại lượng thể hiện sự biến thiên giữa các nhóm (between-groups sum of squares) được tính bằng công thức:

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

SSG thể hiện sự biến thiên do sự khác nhau giữa các nhóm, tức là biến thiên do yếu tố kiểm soát gây ra

Gọi STT là tổng bình phương các chênh lệch giữa từng quan sát với trung bình của tất cả các quan sát ta có:

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

Đã chứng mính được rằng SST = SSW + SSG và công thức này chính là cơ sở của phương pháp phân tích phương sai một yếu tố với biến thiên của các quan sát so với giá trị trung bình là tổng cộng của biến thiên được giải thích bởi yếu tố kiểm soát (SSG) và biến thiên do các yếu tố khác ngoài yếu tố kiểm soát là SSW

Bước 3: Tính các ước lượng cho phương sai chung của k tổng thể, MSW và MSG, bằng cách cia SSW và SSG cho số bật tự do tương ứng, ta có:

$$MSW = \frac{SSW}{n-k} \quad \text{(Within-groups mean square)}$$

$$MSG = \frac{SSG}{k-1} \quad \text{(Between-groups mean square)}$$

Tỷ số này được dùng để kiểm nghiệm giả thuyết H_0. Nếu H_0 đúng, nghĩa là trung bình của k tổng thể bằng nhau thì tỷ số MSG/MSW sẽ gần với giá trị 1. Ngược lại, khi các trung bình của k tổng thể không bằng nhau, thì MSG lớn hơn MSW, do vậy tỷ số MSG/MSW sẽ lớn hơn 1. Mức độ lớn hơn bao nhiêu thì được xem là "đủ lớn" (tuỳ thuộc vào độ tin cậy) để ta có thể bác bỏ H_0. Bước 4 với việc tính ra gia trị kiểm định F sẽ lý giải điều này

Bước 4: Tính giá trị kiểm định F:

$$F = \frac{MSG}{MSW}$$

Ta sẽ bác bỏ H_0 ở mức ý nghĩa $\propto$ (thường là 0.05), nếu giá trị p-value nhỏ hơn mức ý nghĩa, tương ứng với tỷ số F=MSG/MSW lớn hơn $F_{k-1, n-k, \propto}$, với $F_{k-1, n-k, \propto}$ có phân phối F với k-1 và n-k bật tự do tương ứng ở tử và mẫu số.

Kết quả phân tích phương sai một yếu tố thường được thể hiện dưới dạng bảng sau:

Biến thiên (Variance)	Tổng các chệnh lệch bình phương (Sum of squares)	Bật tự do (df)	Trung bình các chênh lệch bình phương-Phương sai (Mean square)	Giá trị kiểm định	P-value Sig.
Giữa các nhóm (Between Groups)	SSG	k-1	MSG=SSG/k-1	F=MSG/ MSW	
Trong nội bộ nhóm (Within Groups)	SSW	n-k	MSW=SSW/n-k		
Tổng công (Total)	SST	n-1			

- **So sánh từng cặp trung bình tổng thể**

Một khi đã quyết định được sự khác biệt tồn tại giữa các giá trị trung bình-bác bỏ H_0, hiển nhiên nảy sinh câu hỏi tiếp theo là trung bình những tổng thể nào là khác nhau, tổng thể nào có trung bình lơn hơn hoặc nhỏ hơn. Để trả lời các câu hỏi này SPSS cung cấp các kiểm nghiệm **post hoc range** và **pairwise multiple comparisons** có thể quyết định được những giá trị trung bình nào là khác biệt. **Range tests** xác định ra những nhóm giá trị trung bình đồng nhất không tồn tại sự khác biệt giữa các giá trị trung bình này. Kiểm nghiệm **Pairwise multiple comparisons** kiểm nghiệm sự khác biệt giữa các cặp giá trị trung bình và đưa ra một ma trận đánh dấu hoa thị chỉ những nhóm giá trị trung bình có khác biệt đáng kể ở mức độ tin cậy là 5%

Đối với giả thuyết cân bằng về phương sai được chấp nhận (thông qua kiểm nghiệm Levene) ta có các phương pháp kiểm nghiệm thống kê sau để so sánh các trung bình mẫu:

- **The least significant difference (LSD)** là phép kiểm nghiệm tương đương với việc sử dụng phương pháp kiểm nghiệm t riêng biệt cho toàn bộ các cặp trong biến. Yếu điểm của phương pháp này là nó không chỉnh lý độ tin cậy cho tương thich với việc kiểm nghiệm cho nhiều so sánh cùng một lúc. Do đó dẫn đến độ tin cậy không cao. Các kiểm nghiệm khác sẽ được tham khảo sau đây loại bỏ được yếu điểm này bằng cách điều chỉnh độ tin cậy cho một so sánh nhiều thành phần.
- Phương pháp kiểm nghiệp **Bonferroni** và **Tukey's honestly significant difference** thì được sử dụng cho hầu hết các kiểm nghiệm so sánh đa bội. Kiểm nghiệm **Sidak's t test** cũng được sử dụng tương tư như phương pháp **Bonferroni** tuy nhiên nó cung cấp những giới hạn chặt chẻ hơn. Khi tiến hành kiểm nghiệm một số lượng lớn các cặp trung bình **Tukey's honestly significant difference test** sẽ có tác động mạnh hơn là **Bonferroni test**. Và ngược lại **Bonferroni** thì thích hợp hơn cho các kiểm nghiệm có số lượng cặp so sánh ít.
- **Hochberg's GT2** thì giống như **Tukey's honestly significant difference test** nhưng thông thường **Tukey's test** có tác dụng tốt hơn. **Gabriel's pairwise comparisons test** thì giống như **Hochberg's GT2** nhưng nó thường được sử dụng hơn khi kích cở giữa các mẫu kiểm nghiệm có sự sai biệt lớn
- Phương pháp kiểm nghiệm **Dunnett's pairwise** thì được dùng để so sánh các giá trị trung bình của các mẫu với một gía trị trung bình cụ thể được lấy từ trong tập các mẫu so sánh. Thông thường mặc định nhóm mẫu cuối cùng làm nhóm kiển soát, hoặc ta có thể lựa chọn nhóm đâu tiêu làm nhóm kiểm soát, lúc đó các giá trị trung bình của các nhóm tong biến độc lập sẽ được so sánh với giá trị trung bình của nhóm đầu tiên hoặc nhóm sau cùng của biến độc lập
- **Ryan, Einot, Gabriel, and Welsch (R-E-G-W)** đưa ra hai bước kiểm nghiệm. Đầu tiên tiến hành kiểm nghiệm có hay không toàn bộ các giá trị trung bình là ngang bằng nhau hay không. Nếu toàn bộ các giá trị trung bình là không ngang bằng nhau sau đó bước thứ hai sẽ kiểm nghiệm sự khác biệt giữa các nhóm nhỏ với nhau, để tìm ra những nhóm nào thật sự khác biệt và không khác biệt về giá trị trung bình. Tuy nhiên việc kiểm nghiệm này không nên thực hiện đối với trường hợp kích cở mẫu trong các nhóm không ngang bằng nhau
- Thông thường khi kích thước mẫu không ngang bằng giữa các nhóm. **Bonferroni** và **Scheffé** là hai phương pháp kiểm nghiệm được lựa chọn hơn là phương pháp **Tukey**
    - **Duncan's multiple range test, Student-Newman-Keuls (S-N-K), and Tukey's b** cũng tương tự tuy nhiên nó ít khi được sử dụng như các phương pháp trên.

- Kiểm nghiệm **Waller-Duncan t** được sử dụng khi kích thước mẫu là không bằng nhau
- Phương pháp kiểm nghiệm **Scheffé** cho phép sự kết hợp tuyến tính của những giá trị trung bình sẽ được kiểm nghiệm, không chỉ là so sánh giữa các cặp. Chính vì vậy kết quả của kiểm nghiệm **Scheffé** thì thường thận trọng hơn các phương pháp kiểm nghiệm khác , nó đòi hỏi một sự khác biệt lớn giữa các giá trị trung bình quan sát được để bảo đảm tính thật sự khác biệt của phép kiểm nghiệm

Đối với trường hợp giả thuyết về sự cân bằng phương sai giữa các mẫu không được chấp nhận ta sẽ sử dụng các phương pháp kiểm nghiệm sau để tiến hành so sánh giá trị trung bình giữa các nhóm:**Tamhane's T2**, **Dunnett's T3**, **Games-Howell**, **Dunnett's C**

Ví dụ như trong nông nghiệp người ta muốn biết ngũ cốc sẽ phát triển như thế nào khi sử dụng các loại phân bón khác nhau. Nhà nghiên cứu muốn biết liệu tất cả các loại phân bón trên thì có ảnh hưởng ngang bằng đến sự phát triển của ngu cốc hay một vài loại phân bón sẽ có tác dụng tốt hơn một vài loại khác. Để kiểm nghiệm điều này người ta dùng ANOVA đề kiểm nghiệm tốc độ phát triển trung bình (có thể là trong lượng ngũ cốc thu hoạch, chiều cao của cây, số lượng trái trung bình thu hoạch được, …) đây chính là các giá trị trung bình được sử dụng trong thống kê.

ANOVA thông thường kiểm nghiệm trên một số lượng mẫu lớn hơn hai, nếu số lượng mẫu bằng 2 ta có thể dùng một phương pháp tương đối đơn giãn hơn là kiểm nghiệm t hai mẫu như đã đề cập ở phần trên. ANOVA được sử dụng rộng rãi trong thực tế bởi vì ta sẽ gặp rất nhiều trường hợp đòi hỏi ta phải kiểm nghiệm nhiều mẫu trong cùng một lúc. Chú ý nếu ta kiểm nghiệp theo từng cặp lần lượt bằng phương pháp kiểm nghiệm t hai mẫu mỗi lần kiểm nghiệm độ sai lệch sẽ là 5% (tuỳ thuộc vào mức tin cậy mà ta mong muốn). Do đó khi kiểm nghiệm tất cả các cặp mẫu lần lượt tỷ lệ sai sót sẽ tăng lên theo mỗi lần. Do đó ANOVA sẽ cho phep1 ta kiểm nghiệm tất cả các mẫu trong cùng một mức độ sai sót là 5% và kiểm nghiệm trong một lần

Để thực hiện kiểm nghiệm ANOVA, dữ liệu đòi hỏi phải thỏa mãn một số giả thuyết sau:

- Các mẫu kiểm nghiệm phải độc lập và mang tính ngẫu nhiên
- Các mẫu sử dụng trong kiểm nghiệm phải có phân phối chuẩn hoặc kích thước mẫu đủ lớn để được xem là gần như phân phối chuẩn.
- Phương sai của các mẫu thì phải ngang băng nhau (có thể kiểm nghiệp điều này bằng phép kiểm nghiệm phương sai Levene.

Nếu như các mẫu nghiên cứu của ta không thỏa mãn điều kiện trên ta cò thể dùng phép kiểm nghiện phi tham số (nonparametric) như như phép kiểm nghiệm **Kruskal-Wallis**

Ví dụ minh họa:

Các nhà chế biến và phân phối Coffee ở thị trường Hoa Kỳ đang đối mặt với một tình hình bất ổn về giá của hạt Coffee. Trong một năm giá của hạt coffee trội xụt từ $1.40 một pound (0.373 kg) lên $2.50/pound rồi sau đó lại tụt xuống $2.03/pound. Người ta xác định sự bất ổn về giá coffee này là do tình hình hoạt động của bản thân các nhà chế biến và phân phối coffee và một yếu tố khác rất quan trọng là vấn đề hạn hán ở Brazil, bởi vì Brazil sản xuất ra 30% sản lượng coffee trên thế giới, do đó thị trường coffee rất nhạy cảm với những biến chuyển về thời tiết (nguy cơ hạn hán) ở Brazil.

Để tạo ra một sự ổn định cho hoạt động của mình một nhà phân phối Coffee muốn loại bỏ mặt hàng Coffee Brazil ra khỏi cơ cấu hàng hóa của mình. Tuy hiên trước khi thực hiện quyết định này còn có một cân nhắc là liệu loại bỏ mặt hàng Coffee Brazil thì có làm giảm doanh số của công ty hay không. Vì vậy công ty thuê một công ty nghiên cứu Marketing tiến hành kiểm nghiệm thông kê về sự ưa thích mùi vị coffee của khách hành tiêu dùng Coffee trên thị trường. Công ty tiến hành khảo sát dựa trên ba nhóm khách hàng được lựa chọn ngẫu nhiên bao gồm nhóm khách hàng chuyên tiêu dùng Coffee Brazil, Nhóm khách hàng chuyên tiêu dùng Coffee Colombia và nhóm khách hàng tiêu dùng Coffee Châu Phi (đây là 3 loại Coffee được tiêu dùng chủ yếu của công ty). Chú ý công ty loại trừ những nhóm khách hành vừa tiêu dùng nhiều loại coffee khác nhau, để bảo đảm tính độc lập của các mẫu được chọn, và do nghiên cứu về mùi vị nên đòi hỏi chọn những khách hàng có gu tiêu dùng riêng biệt. Ở đây công ty muốn xác định xem liệu có sự khác biệt về sự mức độ ưa thích đối với ba loại coffee (Sẽ cho khách hành thử ba loại coffee và khảo sát sự đánh giá về mức độ ưa thích của ba loại Coffee) hay có sự khác nhau và khác nhau này như thế nào ở bao loại Coffe và ở ba nhóm khách hàng.

Dựa vào kết quả phân tích ANOVA sẽ cho ta biết liệu mức độ ưa thích trung bình của ba nhóm khách hàng trên là giống nhau hay khác nhau đối với từng loại coffee. Sau đó dùng phương pháp kiểm nghiệp **Post Hoc** để xác định những khác biệt của từng nhóm khách hàng về loại coffee đã thử.

Sau khi dùng ANOVA khảo sát sự khác biệt giữa các mẫu. Nếu ta có đủ cơ sở để kết luận là không có sự khác biệt giữa các mẫu. Ta có thể kết thúc công việc (việc loại bỏ coffee brazil không gây ảnh hưởng đến doanh số, người tiêu dùng có thể chuyển sang coffee comlobia hoặc châu Phi một cách dễ dàng). Tuy nhiên khi ta loại bỏ giã thuyết về sự ngang bằng giữa các nhóm. Ta phải xác định tiếp sự khác biệt như thế nào giữa các mẫu kiểm nghiệm. Chúng ta cần phải xác định hướng và độ lớn của các khác biệt này bằng cách lần lượt so sánh sự khác biệt giữa các mẫu với nhau (người tiêu dùng coffee brazil có thể thích coffee comlombia hơn coffe châu Phi, hoặc người tiêu dùng coffee brazil đánh giá coffee brazil ngang bằng với coffee colombia, trong khi mức độ ưa thích coffee châu Phi thì thấp hơn do đó để giảm thiểu sự mất doanh số bán coffee brazil khi loại bỏ mặt hàng công ty nên tăng lượng coffee comlombia tiêu thụ trên thị

trường) các công cụ thống kê trong kiểm nghiệp **Post Hoc** cho phép ta thực hiện công việc này.

Phân tích phương sai một chiều là tiến trình phân tích phương sai một chiều cho một biến định lượng phụ thuộc với một yếu tố đơn lẻ hay còn gọi là biến độc lập. Phân tích phương sai (ANOVA) được dùng để kiểm nghiệm giả thuyết cho rằng tất cả các giá trị trung bình đều ngang bằng nhau. Kỷ thuật này là một dạng mở rộng của kiểm nghiệm T hai mẫu.

Để xác định sự khác biệt giữa các giá trị trung bình chúng ta có thể muốn biết những giá trị trung bình nào là khác biệt. Một khi đã quyết định được sự khác biệt tồn tại giữa các giá trị trung bình, các kiểm nghiệm **post hoc range** và **pairwise multiple comparisons** có thể quyết định được những giá trị trung bình nào là khác biệt. **Range tests** xác định ra những nhóm giá trị trung bình đồng nhất không tồn tại sự khác biệt giữa các giá trị trung bình này. Kiểm nghiệm **Pairwise multiple comparisons** kiểm nghiệm sự khác biệt giữa các cặp giá trị trung bình và đưa ra một ma trận đánh dấu hoa thị chỉ những nhóm giá trị trung bình có khác biệt đáng kể ở mức độ tin cậy là 5%

Đối với giả thuyết cân bằng về phương sai được chấp nhận (thông qua kiểm nghiệm Levene) ta có các phương pháp kiểm nghiệm thống kê sau để so sánh các trung bình mẫu:

- **The least significant difference (LSD)** là phép kiểm nghiệm tương đương với việc sử dụng phương pháp kiểm nghiệm t riêng biệt cho toàn bộ các cặp trong biến. Yếu điểm của phương pháp này là nó không chỉnh lý độ tin cậy cho tương thich với việc kiểm nghiệm cho nhiều so sánh cùng một lúc. Do đó dẫn đến độ tin cậy không cao. Các kiểm nghiệm khác sẽ được tham khảo sau đây loại bỏ được yếu điểm này bằng cách điều chỉnh độ tin cậy cho một so sánh nhiều thành phần.

- Phương pháp kiểm nghiệp **Bonferroni** và **Tukey's honestly significant difference** thì được sử dụng cho hầu hết các kiểm nghiệm so sánh đa bội. Kiểm nghiệm **Sidak's t test** cũng được sử dụng tương tư như phương pháp **Bonferroni** tuy nhiên nó cung cấp những giới hạn chặt chẻ hơn. Khi tiến hành kiểm nghiệm một số lượng lớn các cặp trung bình **Tukey's honestly significant difference test** sẽ có tác động mạnh hơn là **Bonferroni test**. Và ngược lại **Bonferroni** thì thích hợp hơn cho các kiểm nghiệm có số lượng cặp so sánh ít.

- **Hochberg's GT2** thì giống như **Tukey's honestly significant difference test** nhưng thông thường **Tukey's test** có tác dụng tốt hơn. **Gabriel's pairwise comparisons test** thì giống như **Hochberg's GT2** nhưng nó thường được sử dụng hơn khi kích cở giữa các mẫu kiểm nghiệm có sự sai biệt lớn

- Phương pháp kiểm nghiệm **Dunnett's pairwise** thì được dùng để so sánh các giá trị trung bình của các mẫu với một gía trị trung bình cụ thể được lấy từ

trong tập các mẫu so sánh. Thông thường mặc định nhóm mẫu cuối cùng làm nhóm kiển soát, hoặc ta có thể lựa chọn nhóm đâu tiêu làm nhóm kiểm soát, lúc đó các giá trị trung bình của các nhóm tong biến độc lập sẽ được so sánh với giá trị trung bình của nhóm đầu tiên hoặc nhóm sau cùng của biến độc lập

- **Ryan, Einot, Gabriel, and Welsch (R-E-G-W)** đưa ra hai bước kiểm nghiệm. Đầu tiên tiến hành kiểm nghiệm có hay không toàn bộ các giá trị trung bình là ngang bằng nhau hay không. Nếu toàn bộ các giá trị trung bình là không ngang bằng nhau sau đó bước thứ hai sẽ kiểm nghiệm sự khác biệt giữa các nhóm nhỏ với nhau, để tìm ra những nhóm nào thật sự khác biệt và không khác biệt về giá trị trung bình. Tuy nhiên việc kiểm nghiệm này không nên thực hiện đối với trường hợp kích cở mẫu trong các nhóm không ngang bằng nhau

- Thông thường khi kích thước mẫu không ngang bằng giữa các nhóm. **Bonferroni** và **Scheffé** là hai phương pháp kiểm nghiệm được lựa chọn hơn là phương pháp **Tukey**

- **Duncan's multiple range test, Student-Newman-Keuls (S-N-K), and Tukey's b** cũng tương tự tuy nhiên nó ít khi được sử dụng như các phương pháp trên.

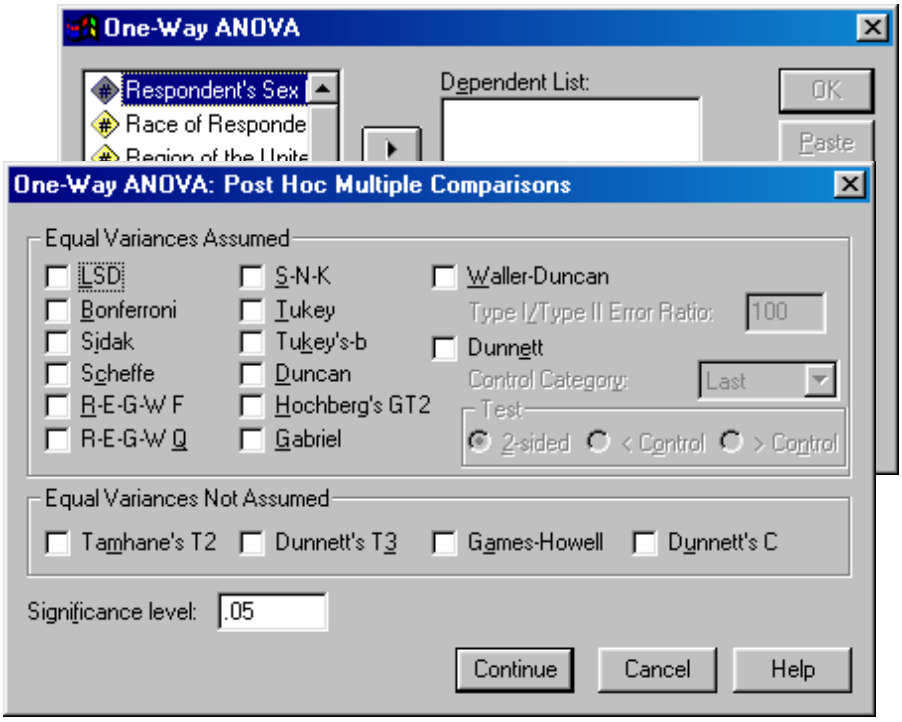
- Kiểm nghiệm **Waller-Duncan t** được sử dụng khi kích thước mẫu là không bằng nhau

- Phương pháp kiểm nghiệm **Scheffé** cho phép sự kết hợp tuyến tính của những giá trị trung bình sẽ được kiểm nghiệm, không chỉ là so sánh giữa các cặp. Chính vì vậy kết quả của kiểm nghiệm **Scheffé** thì thường thận trọng hơn các phương pháp kiểm nghiệm khác , nó đòi hỏi một sự khác biệt lớn giữa các giá trị trung bình quan sát được để bảo đảm tính thật sự khác biệt của phép kiểm nghiệm

Đối với trường hợp giả thuyết về sự cân bằng phương sai giữa các mẫu không được chấp nhận ta sẽ sử dụng các phương pháp kiểm nghiệm sau để tiến hành so sánh giá trị trung bình giữa các nhóm:**Tamhane's T2**, **Dunnett's T3**, **Games-Howell**, **Dunnett's C**

Để thực hiện phép kiểm nghiệm ANOVA ta vào **Comapre means\One-Way ANOVA…** từ thanh menus để truy xuất ra hộp thoại như hình 6-18. Di chuyển vệt tối đến các biến định lượng cần so sánh, chuyển sang hộp thoại **Dependent List.** Lựa biến kiểm soát tức là biến độc lập (yêu cầu phải có ba giá trị trở lên trong biến kiểm soát này) chuyển biến kiểm soát vào hộp thoại **Factor**, Biến kiểm soát này cho phép ta phân các giá trị trung bình theo từng nhóm để kiểm nghiệm. Thao tác đến đây cho phép ta đưa ra kết luận liệu các trung bình của các nhóm có bằng nhau hay không.

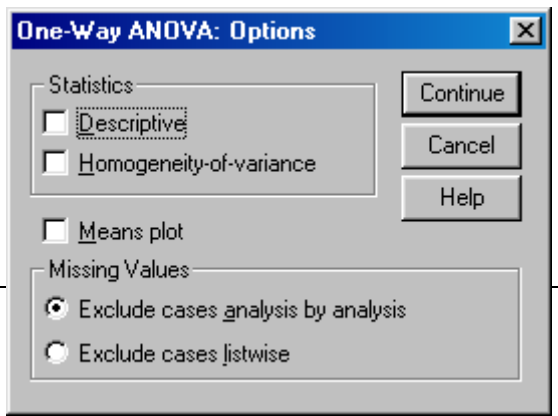
Hình 6-18



Để tiến hành kiểm nghiệp so sánh sự khác biệt giữa các nhóm với nhau ta lựa chọn công cụ **Post Hoc** ta có được hộp thoại như hình 6-19 và lựa chọn các phương pháp kiểm nghiệm thích hợp

Hình 6-19

Lựa chọn công cụ **Options** cho ta hộp thoại như hình 6-20. Để xác định loại loại thông kê mô tả (**Descriptive**) và tính đồng nhất của phương sai, công cụ để tính hệ số thống kê **Levene** để kiểm nghiệm sự ngang bằng về phương sai giữa các nhóm (việc tính toán này quyết định đến sự lưa chon phương pháp kiểm nghiệp trong phần **Post Hoc.** Công cụ **Means Plot** dùng để hiển thị đồ thị về giá tri



trung bình của các nhóm. Công cụ **Missing Values** dùng để kiểm soát giá trị khuyết.

Hình 6-20

- **Exclude cases analysis by analysis**: Những trường hợp có giá trị khuyết ở trong biến phụ thuộc và cả biến kiểm soát sẽ không được đưa vào trong kiểm nghiệm. Ngoài ra những trường hợp có giá trị quan sát nằm bên ngoài chuỗi đã xác định cho biến kiểm soát cũng không được sử dụng
- **Exclude cases listwise**. Những trường hợp có giá trị khuyết Cases trong biến điều khiển hoặc bất kỳ biến phụ thuộc nào được đưa ra hoặc không đưa ra kiểm nghiệm đều bị loại trừ ra khỏi quá trình kiểm nghiệm phân tích .

Các giả định phải được thỏa mãn khi dùng phân tích ANOVA một chiều

- Các mẫu dữ liệu phải độc lập, ngẫu nhiên và được lấy ra từ một tổng thể phân phối chuẩn
- Trong tổng thể các phương sai của các mẫu dữ liệu phải bằng nhau (điều này sẽ được kiểm nghiệm qua thông kê **Levene's homogeneity-of-variance**.

Xem thêm vị dụ trong phần phu lục