Hướng dẫn cách đọc kết quả xử lý số liệu điều tra xã hội học bằng chương trình SPSS/PC+

TÔN LƯƠNG CHÍNH

I. Sơ lược về chương trình SPSS/PC+

Ngày nay, các cuộc điều tra Xã hội học mang tính định lượng thường được xử lý trên máy vi tính bằng chương trình (phần mềm) SPSS/PC+. SPSS là tên viết tắt tiếng Anh của Statistical Package For Social Science, nghĩa là Bộ chương trình thống kê cho Khoa học xã hội.

Chương trình SPSS có khả năng tính được tần số, tần xuất, trung vị, mốt, trung bình số học, phương sai, sai số tiêu chuẩn … của từng biến riêng lẻ của toàn mẫu cũng như của từng phân nhóm trong mẫu một cách đồng thời. Ví dụ cùng một lúc ta có thể tín được tuổi kết hôn trung bình của nhóm Nam và Nữ , hoặc tuổi kết hôn của những người ở các độ tuổi khác nhau… SPSS cho phép ta tính được tương quan (mối liên hệ) của hai biến. Điểm ưu việt của chương trình này là ở chỗ: nó không chỉ cung cấp cho chúng ta bảng tương quan giữa hai biến (hoặc ba biến) mà còn cung cấp cho chúng ta hệ số tương quan, mà các chương trình khác như FOX , FOXBASE, LOTUS… không có được. Hệ số tương quan biểu thị mối liên hệ mạnh hay yếu của hai biến. Thông qua hệ số tương quan này chúng ta có thể biết được tác động của biến khác nhau đến một biến nào đó (chẳng hạn xem tác động của biến học vấn và biến mức sống đến biến đánh giá chính sách đối với chính sách đổi mới) thì biến nào mạnh hơn. Hệ số tương quan có đặc tính chung là chúng biến thiên trong khoảng từ 0 đến 1. Khi hệ số tương quan bằng 0 giữa hai biến không có mối liên hệ. Khi nó bằng 1 thì giữa 2 biến có mối liên hệ hàm số (mối liên hệ rất chặt), nghĩa là ta có thể biểu thị mối liên hệ đó bằng một hàm số, chẳng hạn hàm số tuyến tính có dạng:

$$Y = ax + b$$

Trong đó x,y là các biến cần đo, còn a, b là các hằng số. Khi hệ tương quan khác 0 thì ta nói rằng giữa hai biến có mối liên hệ tương quan. Hệ số tương quan càng lớn (càng gần 1) thì mối liên hệ càng chặt. Nói cách khác ảnh hưởng của biến số độc lập đến biến số phụ thuộc càng mạnh. Chẳng hạn hệ số tương quan giữa biến học vấn và biến đánh giá chính sách đổi mới bằng 0.25 còn giữa biến mức sống và biế đánh giá chính sách đổi mới là 0.40 thì lúc đó ta cod thể kết luận được rằng: học vấn và mức sống của người dân có ảnh hưởng đến cách đánh giá của họ về chính sách

đổi mới, song mức sống của những người được hỏi có ảnh hưởng mạnh đến nhận thức của họ về chính sách đổi mới hơn học vấn của họ.

Ngoài ra với hai biến định lượng, chương trình SPSS còn tính được mối liên hệ hồi quy tyến tính, nghĩa là đưa ra được sự phụ thuộc dạng hàm số của hai biến này.

Môt ưu điểm khác của chương trình SPSS là ở chỗ nó có thể tạo ra cho ta các biến mới từ hai hoặc ba biến trở nên đã có sẵn bằng các phép tính số học, logic.

Chương trình SPSS là chương trình sử dụng các công cụ thống kê – toán, bởi vậy các tính toán của nó được thực hiện trên các ngôn ngữ số. Chính vì lẽ đó các thông tin sơ cấp thu được từ các cuộc điều tra Xã hội học bằng phương pháp phỏng vấn tiêu chuẩn (có bảng hỏi) ăng két cần phải được chuyển qua ngôn ngữ số. Công việc chuyển các thông tin bằng lời sang thông tin bằng số được gọi là mã hóa (code). Các thông tin bằng số trong ngôn ngữ máy được gọi là các giá trị (Value) còn thông tin bằng lời ứng với các giá trị đó là tên của giá trị (Value Label). Sau khi mã hóa (chuyển ngôn ngữ lời sang ngôn ngữ số) các thông tin này sẽ được chương trình SPSS /PC+ xử lý và phân tích theo yêu cầu của người nghiên cứu đối với từng câu hỏi (từng biến) cụ thể nhằm đáp ứng mục tiêu đặt ra cho từng câu hỏi phù hợp với dạng thang đo của câu hỏi đó.

Thông thường, chương trình SPSS cho ta các dạng kết quả sau:

1- Bảng tần số, tần suất (Frequency) của từng biến riêng lẻ. Khi tính toán tần số (Frequency) chương trình SPSS còn có thể cho ta biết them một số thông số hỗ trợ như: Trung bình số học (Mean), Mốt (Mode): giá trị có tần số cao nhất, Trung vị (Median): giá trị phân chia mẫu thành hai phần bằng nhau theo biến đang tính, Phương sai, Độ lệch chuẩn … Ngoài ra khi tần số máy cũng có thể vẽ được đồ thị phân bố của các giá trị theo kiểu hình cột hoặc đa giác.

2- Bảng giá trị trung bình số học (Mean) của một biến nào đó (thường là các biến đo bằng thang đo định lượng, ví dụ: tiền lương, tuổi, số thóc thu được…) với một biến nào đó. Chẳng hạn với câu hỏi: " Trung bình mỗi năm gia đình ta thu được bao nhiêu thóc". Lúc đó bảng này cho ta không những tiêu chí (biến) khác nhau mà ta cần tính cho nó như: mức sống, quy mô gia đình, nghề nghiệp của gia đình… Bên cạnh đó khi tính trung bình số học Chương trình còn cho ta thệm thông tin về độ lệch tiêu chuẩn (phương sai) theo từng phân nhóm và số người (case) của phân nhóm được đưa vào tính giá trị trung bình số học.

3- Bảng tương quan của hai biến, Bảng này thường có dạng bảng hai chiều với k hàng và l cột. Số lượng hang và cột phụ thuộc vào số lượng các phương án trả lời của các biến mà ta cần tính tương quan. Ví dụ: Biến A có 3 phương án trả lời, còn biến B có 5 phương án trả lời thì số K =3 còn số l = 5. Chương trình SPSS chỉ tính tương quan cho hai biến là hai câu hỏi tuyển (mỗi người chỉ được chọn 1 phương án trả lời trong số phương án vạch ra) chứ không tính được cho các câu hỏi hội. Trong bảng ngoài số lượng tuyệt đối (tần số), còn có số lượng tương đối (phần trăm). Số phần trăm có thể tính được theo hang (Row Percent), theo cột (Collum Percent) hoặc cả hai tùy theo yêu cầu và ý nghĩa của cách tính.

Ở cuối bảng có đưa ra hệ tương quan của hai biến. Có nhiều loại hệ số tương quan khác nhau như: Phi, Cramer, Persson… Tùy theo từng dạng thang đo của các biến mà chúng ta yêu cầu máy tính hệ số thuộc dạng nào cho phù hợp. Thông thường hệ số tương quan Cramer hay được dùng nhất , bởi vì hệ số này dùng được cho tất cả các dạng thang đo, từ thang định tính đến thang đo định lượng.

II. Hướng dẫn sử dụng các kết quả.

1. Bảng tần số có dạng sau:

A1 Thôn điều tra

			Valid	Cum	
Value label	Vale	Frenquency	Percent	Percent	Percent
Xuân Viên	1	106	52.0	52.5	52.5
Mễ Sơn	2	61	50.2	50.2	82.7
Đông Sơn	3	17	8.4	8.4	91.1
Xuân Cầm	4	18	9.9	8.9	100.0
		1	5	MISSING	
TOTAL		203	100.0	100.0	

Mean …….. Std Err…… Median…….

Mode ……. Std Dev …… Variance……

Valid Case 202 Missing Case 1

Giải thích:

Chú ý: Chương trình này được viết bằng tiếng Anh cho nên các phần giải thích bằng tiếng Việt như: Nội dung câu hỏi, các phương án trả lời đều không có dấu, mặt khác trong tiếng Anh quy ước chữ số có khác với tiếng Việt: dấu phẩy thập phân được thay bằng dấu chấm cho nên cần lưu ý khi đọc kết quả. Ví dụ: 34.6 cần hiểu là 34,6.

-Value Label: Tên của giá trị, nghĩa là tên của phương án trả lời trong câu hỏi mà **ta** đang tính. Ví dụ: Tên của thôn trong câu hỏi: "Hiện nay gia đình ta sống ở thôn nào"

-Value: Chữ số, mà ta gán mã cho tên phương án đó (mã hóa) để đưa vào máy tính.

- Frenquency: Tần số, nghĩa là số người trả lời theo các phương án khác nhau của câu hỏi này. Ví dụ: Số người của thôn Xuân Viên là 106 người

- Percent (Phần trăm): Tỷ lệ phần trăm số người trả lời của mỗi phương án (Value) trên tổng số người được điều tra, kể cả những người không trả lời câu hỏi này. Ví dụ trong số 203 người được điều tra có 106 người của thôn Xuân Viên, chiếm 52,0%.

- **Valid Percent** (Tỷ lệ phần trăm thực): tỷ lệ phần trăm số người trả lời của mỗi phương án (value) trên tổng số người trả lời câu hỏi này. Ví dụ ở bảng trên do có một người không trả lời câu hỏi: "sống ở thôn nào" nên tỷ lệ phần trăm thực (Valid Percent) của số người ở thôn Xuân Viên (Value = 1) là 52,3%.

- Cum Percent (Tần suất tích lũy hay phần trăm tích lũy): Tổng phần trăm của các phương án trả lời.

- **Missing:** Số đơn vị thiếu vắng không quan sát được (không trả lời câu hỏi này).

- **Total:** Tổng số theo các mục (cột). Ví dụ: Total theo cột Frequency là tổng số người được điều tra, bằng 203.

- **Mean:** Trung bình số học của biến đang tính (thường chỉ được tính cho các biến đo bằng thang đo định lượng như: thu thập, tuổi đời, số con....).

- **Std Err:** Sai số tiêu chuẩn.

- **Median:** Trung vị, điểm phân chia tập hợp mẫu (số người được điều tra) thành hai phần bằng nhau, nghĩa là có 50% số người được điều tra nằm ở dưới trung vị, còn 50% người còn lại có giá trị của biến (câu hỏi) đang tính lớn hơn trung vị.

- **Mode:** Giá trị của biến (phương án trả lời của câu hỏi) có tần số cao nhất. Mode cũng như Median là một dạng trung bình của các biến định tính.

- **Std Dev:** Độ lệch chuẩn (có tài liệu thống kê gọi là độ lệch chuẩn quân phương), nói lên sự phân tán các giá trị của biến được tính so với giá trị trung bình. Ví dụ: 68% người trả lời có giá trị nằm trong khoảng từ trung bình số học (mean) trừ đi độ lệch tiêu chuẩn (Std Dev) đến trung bình số học cộng với độ lệch tiêu chuẩn.

- **Variance:** Phương sai, biểu thị độ phân tán của các giá trị quanh trung binh số học. Phương sai bằng bình phương của độ lệch tiêu chuẩn. (Để hiểu rõ hơn ý nghĩa của các con số này xin xem thêm tài liệu thống kê toán)

- **Valid Cases:** Số người trả lời câu hỏi này

- **Missing Cases:** Số người không trả lời câu hỏi này

Tổng của Valid Cases và Missing Cases bằng số người được điều tra.

2. Bảng tương quan (Crosstabulation)

Biểu thị mối liên hệ của hai biến (hai câu hỏi hay hai dấu hiệu) với nhau. Thông thường là một biến độc lập và một biến phụ thuộc. Bảng tương quan có dạng sau:

Crosstabulation A5 Nội dung (tên) của biến A5 (câu hỏi A5)

by B1 Nội dung (tên) của biến B1 (câu hỏi B1)

Count Nội dung (tên) của các phương án trả lời của biến B1

B1	Col Pct	1	2	3	Row total
A5 Nội dung (tên) của các phương án trả lời	1	3 27.3	2 10.0		5 14.7
	2	5 45.5	1 5.0		6 17.6
	3	3	17	3	23

		27.3	85.0	100.0	67.6
Câu A5	Column	11	20	3	34
Total		34.4	58.8	8.8	100.0

Statistic	Value		Significance
Cramer''s V	43876		

Number of Missing Observations = 168

Giải thích:

Mở đầu bảng có chữ Crosstabulation (tương quan), tên biến thứ nhất (thường được biểu thị bằng chữ cái với số chữ viết tắt tên biến, ví dụ chữ **A1**) tiếp đó là giải thích đầy đủ tên biến (nội dung câu hỏi thứ nhất).

Dòng thứ hai có chữ By, tiếp đó là chữ viết tắt của biến thứ hai (ví dụ B1) và giải thích đầy đủ của biến thứ hai (nội dung câu hỏi thứ hai). Chú ý: Giải thích đầy đủ tên biến không được dài quá 48 ký tự cho nên đối với những câu hỏi dài chỉ có thể viết được các chữ đầu. Giải thích các phương án trả lời không được dài quá 16 ký tự (kể cả khoảng cách giữa hai từ), dài quá số quy định ở trên máy sẽ tự động cắt. Toàn bộ hai dòng đầu này có ý nghĩa là bảng tương quan của biến số nào (câu hỏi nào) với biến (cái hỏi) nào.

Tiếp đến là bảng số liệu tương quan của hai biến số. Các phương án trả lời của biến thứ hai được bố trí theo hàng ngang của dòng đầu (có mũi tên chỉ vào, ví dụ B1→) còn các phương án trả lời của biến thứ nhất được bố trí trong cột đầu. Trong phần các phương án trả lời có nội dung các phương án trả lời và chữ số mã của nó. Trên cùng của cột đầu có ghi chữ **Count,** nghĩa là cách tính phần trăm như thế nào. Nếu dưới chữ **Count** có chữ:

- **Row. Pct:** Tính phần trăm theo hàng, nghĩa là bằng số ở ô đó chia cho tổng số các số trong hàng cùng với ô đó nhân với 100.
- **Col. Pct:** Tính phần trăm theo cột.
- Hoặc có cả hai chữ **Row.Pct và Col.Pct** tức là tính phần trăm cho cả hàng lẫn cột.

Trong mỗi ô của bảng tương quan có hai hoặc ba chữ số. Chữ số ở trên (số nguyên) là tần số, chữ số thứ hai (số thập phân) là tỷ lệ phần trăm, chữ số thứ ba (nếu có) cũng là chữ số thập phân, là tỷ lệ phần trăm.

Cột cuối cùng bên tay phải của bảng có chữ **Row Total** nghĩa là cách trả lời các phương án của tất cả mẫu đối với biến **A1.**

Dòng cuối cùng của cột đầu có chữ **Column Total,** cách trả lời các phương án đối với câu B1 của tất cả các mẫu (cả cuộc điều tra).

Tiếp theo dưới bảng có các dòng chữ sau:

Dòng thứ nhất:

- **Statistic:** Tên hệ số tương quan
- **Value:** Giá trị của hệ số tương quan

Dòng thứ hai: Dưới chữ Statistic là tên hệ số tương quan mà ta tính, chẳng hạn: Crammer, Phi, Student.... dưới chữ Value là giá trị của hệ số tương quan.

Dòng cuối cùng của bảng này là chữ:

Number ò Missing Observation = ...Nghĩa là tổng số người không có thông tin một trong hai biến. Ví dụ: câu A1 có 3 người không trả lời, còn câu B1 có 5 người không trả lời thì Number ò Missing Observation bằng 8.

3. Bảng tính giá trị trung bình số học

Bảng trung bình số học của một biến nào đó tính cho các nhóm nhỏ trong mẫu (theo cách phân chia nào đó toàn mẫu thành nhóm nhỏ) thường có dạng sau đây:

Summaries of A2 Tên biến cần tính trung bình

By levels of A1 Tên biến lấy làm chuẩn để phân chia thành nhóm nhỏ.

Variable	Value Label	Mean	Std Dev	Cases
For Entire Population		52.6190	55.0671	126
A1	1 (tên các nhóm	52.6190	22.9662	32

A1	3 nhỏ của tiêu	49.7727	14.3093	22
A1	4 chuẩn	40.3448	18.2860	29
A1	5 phân chia	156.8000	263.4838	5
A1	6 A1)	48.5238	15.6257	21
A1	7	41.1765	15.9501	17

Total cases = 202

Missing Cases = 76 or 37.6 PCT

Ở đây cột Variable cột nêu lên biến (câu hỏi, dấu hiệu) làm tiêu chuẩn phân chia mẫu thành các nhóm nhỏ. Cột value labels là cột tên các nhóm nhỏ, còn cột Mean là giá trị trung bình của từng nhóm nhỏ, cột Std Dev là độ lệch tiêu chuẩn của nhóm và cuối cùng cột Cases là số người được tính trong mỗi nhóm.

Dưới dòng chữ tiếng Anh đó có dòng chữ For Entire Population nghĩa là trung bình cho toàn mẫu.

Total cases: Số người được điều tra

Missing cases: Tổng số người không trả lời của hai câu là bao nhiêu và chiếm bao nhiêu phần trăm.