

Chương Trình Giảng Dạy Kinh tế Fulbright

Học kỳ Thu năm 2015

Các Phương Pháp Phân Tích Định Lượng

GỢI Ý ĐÁP ÁN -- BÀI TẬP 1

THỐNG KÊ MÔ TẢ

Ngày Phát: Thứ ba 06/10/2015

Ngày Nộp: Thứ ba 13/10/2015

Bản in nộp trước 8h20 tại Hộp nộp bài tập trong phòng Lab.

Bản điện tử nộp trước 8h20 tại địa chỉ: <http://www.fetp.edu.vn/vn/tai-nguyen/hoc-vien-hien-tai/>

Câu 1: (40 điểm)

Để chuẩn bị cho việc triển khai chương trình tín dụng vi mô tại một số vùng nông thôn ở Bangladesh, một mẫu ngẫu nhiên gồm 120 quan sát là các hộ gia đình tại khu vực nông thôn đã được điều tra và thu thập các thông tin cơ bản về: xã trong vùng có hộ được khảo sát, tuổi của chủ hộ, giới tính, trình độ học vấn của chủ hộ, số thành viên trong hộ, tài sản của hộ tại thời điểm khảo sát, mức chi tiêu hàng tháng của hộ tại thời điểm khảo sát. Dựa vào thông tin được cung cấp, anh/chị hãy thực hiện các yêu cầu sau đây:

- a. Cho biết biến nào là biến định tính, biến nào là biến định lượng (trong đó biến nào là biến định lượng rời rạc, biến nào là biến định lượng liên tục) và loại thang đo của các biến trên?

Biến	Loại biến	Thang đo
Xã	Định tính	Định danh
Tuổi chủ hộ	Định lượng/Liên tục	Tỷ lệ
Giới tính chủ hộ	Định tính	Định danh
Học vấn chủ hộ	Định tính	Thứ bậc
Số thành viên	Định lượng/Rời rạc	Tỷ lệ
Tài sản hộ	Định lượng/Liên tục	Tỷ lệ
Chi tiêu hộ	Định lượng/Liên tục	Tỷ lệ

- b. Tính tài sản trung bình của hộ phân theo nhóm bằng cấp của chủ hộ. Nhận xét của anh/chị từ kết quả?

- Dùng chức năng Pivot Table trong Excel, kết quả tổng hợp như sau:

Row Labels	Average of Tài sản hộ
Đại học	1,098,275
Phổ thông	1,043,924
Trung học	277,964
Tiểu học	184,728
Không	897,187
Grand Total	715,379

- Nhận xét:

- Xu hướng chung, chủ hộ có học vấn càng cao thì tài sản càng lớn, điều đó cho thấy giáo dục có thể là một trong những lựa chọn để giảm nghèo.
- Nhóm đối tượng tiểu học và trung học, học hành dở dang nên có thể dẫn đến tình trạng không có việc làm và nghề nghiệp.
- Nhóm không có trình độ học vấn lại có tài sản khá cao, điều này có thể giải thích do nhóm đối tượng này vì không có điều kiện đến trường nên chọn con đường lao động để kiếm sống ngay từ đầu. Cần có những thông tin bổ sung để đưa ra kết luận sau cùng.

c. Tính số thành viên trung bình trong mỗi hộ phân theo bằng cấp của chủ hộ. Nhận xét của anh/chị từ kết quả?

- Dùng chức năng Pivot Table trong Excel, kết quả tổng hợp như sau:

Row Labels	Average of Số thành viên
Đại học	4.00
Phổ thông	4.91
Trung học	5.69
Tiểu học	5.83
Không	5.90
Grand Total	5.69

- Nhận xét:

- Số thành viên trong gia đình có thể hiểu đơn giản là biến đại diện gián tiếp cho số con trong gia đình sau khi trừ đi 2 người là bố mẹ (tất nhiên, một số trường hợp hộ có thể có ông, bà, ... ở cùng). Tuy nhiên, xu hướng chung là người có trình độ học vấn càng cao thì số thành viên trong gia đình ít.
- Điều này có thể giải thích là do, càng tiếp xúc với giáo dục thì các hiểu biết về kế hoạch hóa gia đình => số con ít.
- Cần có thông tin sâu hơn để đưa ra kết luận cuối cùng về mối quan hệ giữa học vấn và vấn đề kế hoạch hóa gia đình.

d. Tính số chủ hộ phân theo giới tính trong mỗi xã khảo sát? Nhận xét của anh/chị từ kết quả tính toán

- Dùng chức năng Pivot Table trong Excel, kết quả tổng hợp như sau:

Count of Giới tính chủ hộ			
Column Labels			
Row Labels	Nam	Nữ	Grand Total
1	27	6	33
2	33	6	39
3	33	5	38
4	9	1	10
Grand Total	102	18	120

- Nhận xét:

- Xu hướng chung là ở tất cả các xã thì chủ hộ đa số là Nam.
- Đây có thể là văn hóa phổ biến ở các nước Á Đông (tương tự như Việt Nam).

e. Tính số chủ hộ phân theo giới tính và phân theo trình độ học vấn của chủ hộ? Nhận xét của anh/chị từ kết quả?

- Dùng chức năng Pivot Table trong Excel, kết quả tổng hợp như sau:

Count of Giới tính chủ hộ			
Column Labels			
Row Labels	Nam	Nữ	Grand Total
Đại học	5	-	5
Phổ thông	11	-	11
Trung học	13	-	13
Tiểu học	19	4	23
Không	54	14	68
Grand Total	102	18	120

- Nhận xét:

- Tỷ lệ chủ hộ được học hành chỉ chiếm khoảng 50%. Khả năng tiếp cận giáo dục ở các vùng này khá thấp.
- Cơ hội học hành của nữ giới gần như là không có, chỉ 4/18 người là chủ hộ đã từng đi học. Và bậc học cao nhất cũng chỉ là tiểu học. Có thể tồn tại bất bình đẳng giới trong khả năng tiếp cận giáo dục ở các vùng này hay không?!

Câu 2: (30 điểm)

Cũng dựa vào dữ liệu đã cung cấp trong file *P01-Data.xlsx*, anh/chị hãy thực hiện các yêu cầu cụ thể sau đây:

a. Hãy cho biết số quan sát (số hộ gia đình) trong mỗi xã được khảo sát?

- Dùng chức năng Pivot Table trong Excel, kết quả tổng hợp như sau:

Row Labels	Count of Tuổi chủ hộ
1	33
2	39
3	38
4	10
Grand Total	20

- Xã 1, 2, 3 có số hộ được khảo sát khá tương đồng (33 – 39 hộ/xã), trong khi xã 4 chỉ có được 10 hộ được khảo sát. Điều này có thể do khả năng tiếp cận các hộ ở xã này khó khăn hoặc bản chất do tổng số hộ trong xã này ít hơn các xã khác.

- b. Hãy lọc ra các quan sát (các hộ gia đình) thuộc xã có số hộ được khảo sát ít nhất sau đó bằng công thức và định nghĩa, anh/chị hãy tính các giá trị trung bình, trung vị, yếu vị, giá trị nhỏ nhất, giá trị lớn nhất và khoảng biến thiên về tuổi của chủ hộ trong nhóm quan sát đã lọc được?

- Kết quả từ câu a, cho thấy xã 4 là xã có số hộ được khảo sát ít nhất.

- Dùng chức năng Advance Filter với vùng điều kiện là “Xã 4”, kết quả lọc và đã được sắp xếp theo thứ tự tuổi của chủ hộ tăng dần:

STT	Xã	Tuổi chủ hộ
1	4	30
2	4	30
3	4	32
4	4	36
5	4	37
6	4	45
7	4	50
8	4	55
9	4	62
10	4	69

- Dùng định nghĩa và công thức:

- Trung bình: $\bar{x} = \frac{x_1 + x_2 + \dots + x_{10}}{10} = 44,6$ (tuổi)
- Trung vị: số quan sát là số chẵn, trung vị là giá trị trung bình của 2 số ở giữa (tức quan sát thứ $n/2$ và $n/2 + 1$ là quan sát thứ 5 và 6). Vậy trung vị là: $(37+45)/2 = 41$ (tuổi)
- Yếu vị: là giá trị có tần suất xuất hiện nhiều nhất. Trong trường hợp tập dữ liệu ít quan sát như trường hợp này, dễ dàng thấy chỉ có 1 yếu vị là 30 tuổi với 2 lần lặp lại.
- Giá trị nhỏ nhất là giá trị đầu tiên sau khi đã sắp xếp tuổi theo thứ tự: 30 (tuổi)
- Giá trị nhỏ nhất là giá trị cuối cùng sau khi đã sắp xếp tuổi theo thứ tự: 69 (tuổi)
- Khoảng biến thiên là $\text{Range} = \text{Max} - \text{Min} = 69 - 30 = 39$ (tuổi)

- c. Hãy lập lại câu b ở trên nhưng thay vì sử dụng định nghĩa và công thức, anh/chị hãy kiểm tra lại bằng các hàm tính toán trong excel và so sánh kết quả với câu b.

- Sử dụng từng lệnh trong excel hoặc dùng công cụ Descriptive Statistics trong Excel cho dữ liệu của xã 4 cho kết quả:

- Dùng câu lệnh cụ thể:

Average (Trung bình)	44.6
Median (Trung vị)	41
Mode (Yếu vị)	30
Min (Nhỏ nhất)	30
Max (Lớn nhất)	69
Range (Khoảng biến thiên)	39

- Dùng Descriptive Statistics (các kết quả tương ứng ở các ô tô đỏ):

Tuổi chủ hộ	
Mean	44.6
Standard Error	4.41
Median	41
Mode	30
Standard Deviation	13.95
Sample Variance	194.71
Kurtosis	-0.98
Skewness	0.60
Range	39
Minimum	30
Maximum	69
Sum	446
Count	10

- d. Dựa vào 120 quan sát trong mẫu dữ liệu thu thập anh chị hãy tính trung bình, trung vị và chi tiêu của các hộ trong nhóm quan sát đã lọc. Kết luận của anh/chị về độ lệch xiên trong trường hợp này là gì? Giải thích?

- Dùng lệnh cụ thể trong Excel:

- Trung bình_{Chi tiêu} = Average() = 7.983 (ĐVT)
- Trung vị_{Chi tiêu} = Median () = 6.809 (ĐVT)

=> Kết quả Trung bình > Trung vị => phân phối về chi tiêu của hộ gia đình trong mẫu quan sát bị lệch xiên về bên phải.

- e. Anh chị hãy tính độ lệch chuẩn về chi tiêu trung bình của các hộ trong mẫu khảo sát. Theo anh chị, trong số 120 quan sát của mẫu thu thập, có bao nhiêu hộ có mức chi tiêu được xem là bất thường một cách tương đối so với các hộ khác? Giải thích?

- Anh chị có thể dùng định nghĩa để tính từng bước trong Excel, hoặc tính nhanh bằng câu lệnh (giả định trong trường hợp này là độ lệch chuẩn về chi tiêu của mẫu là các hộ gia đình được đại diện):

- STDEV () hoặc STDEV.S () = 5.002 (ĐVT)

- Xác định giá trị bất thường hay còn gọi là giá trị dị biệt (outliers) thì có nhiều tiêu chí và quan điểm khác nhau, trong trường hợp này giả định chi tiêu hộ gia đình sẽ tuân theo phân phối chuẩn, do vậy, giá trị outliers được xác định là nằm ngoài vùng 3 độ lệch chuẩn so với giá trị trung bình (có thể áp dụng tiêu chí 2 độ lệch chuẩn tùy theo quan điểm).

- Trung bình_{Chi tiêu} = Average() = 7.983 (ĐVT)

- Hộ chi tiêu thấp nhất là Min = 1.193 (ĐVT)

(Vị trí tương đối so với giá trị trung bình là $(1.193 - 7.893)/5.002 = -1,36$)

- Hộ chi tiêu cao nhất là Max = 27.776 (ĐVT)

(Vị trí tương đối so với giá trị trung bình là $(27.776 - 7.893)/5.002 = 3.96$)

=> Như vậy, các outliers chỉ nằm bên phải giá trị trung bình vì giá trị thấp nhấp bên trái giá trị trung bình chỉ các $|-1,36| = 1,36$ lần độ lệch chuẩn.

- Ta có: $\bar{x} + 3 * \sigma_x = 7.893 + 3 * 5.002 = 22.989$ (ĐVT)

=> Dùng hàm countif với điều kiện chi tiêu lớn hơn hoặc bằng 22.988 (ĐVT) cho thấy có 2 điểm dị biệt. (Cụ thể là 23.549 và 27.766 ĐVT). Số điểm dị biệt sẽ khác 2 nếu anh chị dùng tiêu chí 2 độ lệch chuẩn hoặc dùng tiêu chí khác.

Câu 3: (30 điểm)

Tiêu tục sử dụng dữ liệu *P01-Data.xlsx* để làm bài, anh chị thực hiện các yêu cầu cụ thể sau:

- Lọc ra các quan sát thuộc xã 1 và xã 2, sau đó tính toán các giá trị trung bình, trung vị, giá trị nhỏ nhất, giá trị lớn nhất, khoảng biến thiên về tài sản của hộ gia đình trong từng xã. Anh chị hãy cho biết, xã nào có tài sản của hộ biến thiên nhiều hơn, vì sao?
 - Dùng Advanced Filter cho biến Tài sản với 2 vùng điều kiện là “Xã 1” và “Xã 2”
 - Dùng Descriptive Statistics để tính các trị thống kê theo yêu cầu.
 - Để so sánh độ biến thiên của 2 tập dữ liệu dùng Hệ số biến thiên – CV.

=> Kết quả:

Tài sản hộ - Xã 1		Tài sản hộ - Xã 2	
Mean	833,591	Mean	696,106
Standard Error	405,382	Standard Error	315,158
Median	130,050	Median	75,822
Mode	#N/A	Mode	#N/A
Standard Deviation	2,328,741	Standard Deviation	1,968,158
Sample Variance	5,423,034,769,252	Sample Variance	3,873,646,358,729
Kurtosis	13	Kurtosis	14
Skewness	4	Skewness	4
Range	9,941,879	Range	9,470,792
Minimum	5,340	Minimum	2,750
Maximum	9,947,219	Maximum	9,473,542
Sum	27,508,513	Sum	27,148,124
Count	33	Count	39
CV = SD/MEAN	2.79	CV = SD/MEAN	2.83

Vì Xã 1 có hệ số biến thiên (2,79) nhỏ hơn Xã 2 (2,83) nên xã 2 biến thiên hơn. Đây là một bằng chứng cho thấy, mặc dù các tiêu chí như độ lệch chuẩn, Range là dùng để đo lường sự biến thiên nhưng không phải là tiêu chí quan trọng nhất để đánh giá độ biến thiên của tập dữ liệu.

- Anh chị kỳ vọng như thế nào về mối quan hệ giữa tài sản và chi tiêu hộ gia đình? Nêu ngắn gọn lý thuyết nào ủng hộ lập luận của anh chị?
 - Kỳ vọng: Biến tài sản giả định được xem như biến thay thế cho biến thu nhập, khi đó, một hộ có thu nhập (hay tài sản càng cao) thì chi tiêu hộ gia đình đó sẽ cao.
 - Lý thuyết: Khuynh hướng chi tiêu trung bình sẽ tăng lên khi thu nhập của người đó tăng (Keynes)

- c. Anh chị hãy tính đồng phương sai của mẫu (gồm tất cả 120 quan sát) giữa tài sản và chi tiêu của hộ gia đình, kết luận của anh/chị từ kết quả tính toán?

- Dùng hàm Covariance.s () hoặc tính tay theo các bước:

B1: Tính $(X_i - \bar{X})$	
B2: Tính $(Y_i - \bar{Y})$	
B3: Tổng các Tích của B1*B2	6.92777E+11
Bước 4: Kết quả đồng phương sai là B3/(n-1)	5,821,652,225

- Nhận xét:

- Đồng phương sai của mẫu $> 0 \Rightarrow$ Tài sản và chi tiêu hộ gia đình có quan hệ đồng biến.
- Nghĩa là hộ có thu nhập càng cao thì chi tiêu càng cao

- d. Anh chị hãy tính hệ số tương quan giữa tài sản và chi tiêu hộ gia đình trong mẫu gồm 120 quan sát? Kết luận của anh chị từ kết quả tính toán?

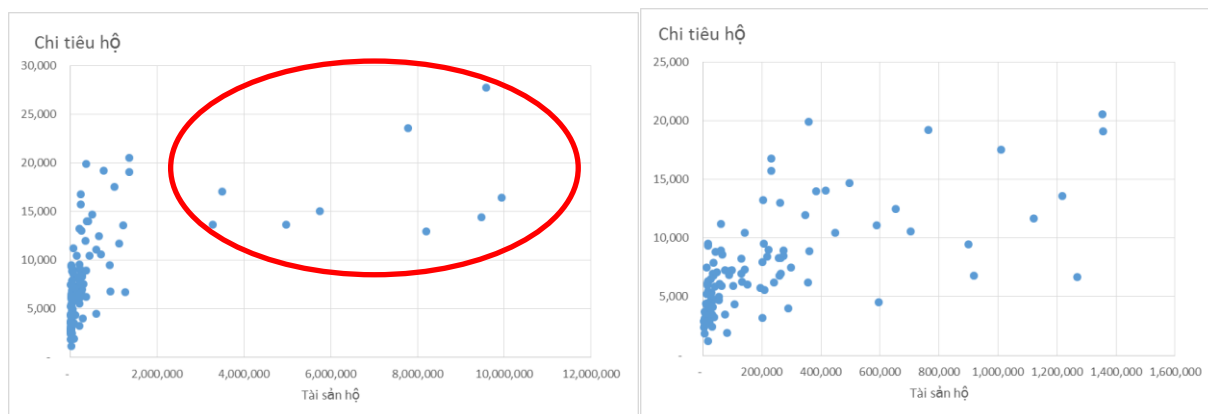
- Dùng hàm CORREL () hoặc tính tay.

- Kết quả CORREL $(X_i, Y_i) = 0,60$

- Nhận xét:

- Hệ số tương quan của mẫu dương cho thấy quan hệ đồng biến giữa tài sản và chi tiêu hộ gia đình. Nghĩa là hộ có thu nhập càng cao thì chi tiêu càng cao.
- Hệ số tương quan chỉ là 0,6 cho thấy mối tương quan giữa chi tiêu và thu nhập trong trường hợp này chỉ ở mức trung bình hoặc trung bình khá.

- e. Anh chị hãy vẽ đồ thị phân phối điểm (Scatter Chart) giữa tài sản và chi tiêu hộ gia đình trong mẫu gồm 120 quan sát. Kết luận của anh/chị từ đồ thị là gì? Nhận xét chung của anh chị từ các kết quả tính toán ở câu c, d, e?



- Đồ thị bên trái vẫn cho thấy xu hướng dốc lên, tức có quan hệ đồng biến giữa tài sản và chi tiêu của hộ (hộ có tài sản càng cao thì chi tiêu càng cao) nhưng xu hướng này không rõ ràng, và phần lớn các hộ là có tài sản dưới 2.000.000 (ĐVT).

- Loại các hộ có tài sản lớn hơn 2.000.000 (ĐVT) cho đồ thị bên phải có xu hướng rõ ràng hơn.

- Kết luận chung là vẫn cho thấy mối quan hệ đồng biến giữa tài sản và chi tiêu, tuy nhiên mối tương quan này chỉ ở mức chung bình.