

GIỚI THIỆU

Dữ liệu

Có ba dạng dữ liệu kinh tế cơ bản: dữ liệu chéo, dữ liệu chuỗi thời gian, và dữ liệu dạng bảng (còn gọi là dữ liệu chéo kết hợp chuỗi thời gian).

Dữ liệu chéo bao gồm các quan sát cho nhiều đơn vị kinh tế tại một thời điểm cho trước. Các đơn vị kinh tế có thể là các cá nhân, các hộ gia đình, các hãng, các tỉnh thành, các quốc gia v.v...

Dữ liệu chuỗi thời gian bao gồm các quan sát trên một đơn vị kinh tế cho trước tại nhiều thời điểm. Ví dụ, chúng ta có thể có các quan sát chuỗi thời gian hàng năm cho chỉ tiêu GDP của một quốc gia từ năm 1960 đến 2002.

Dữ liệu dạng bảng là sự kết hợp giữa các quan sát của các đơn vị kinh tế về một chỉ tiêu nào đó theo thời gian. Ví dụ chúng ta thực hiện điều tra về hộ gia đình cho cùng những hộ gia đình trong vài năm để đánh giá sự thay đổi của những hộ này theo thời gian. Công việc này tạo ra một tập hợp dữ liệu dạng bảng.

Trong khoá học này, chúng ta sẽ tập trung trước nhất vào phân tích dữ liệu chéo. Nếu thời gian cho phép thì chúng ta sẽ giới thiệu thêm những phương pháp cơ bản áp dụng cho phân tích dữ liệu dạng bảng và phân tích chuỗi thời gian.

Dữ liệu có thể được thu thập trên các biến "rời rạc" hay "liên tục".

Một biến rời rạc là biến có một tập hợp các kết quả nhất định có thể đếm được. Ví dụ, số thành viên trong một hộ gia đình là một biến rời rạc.

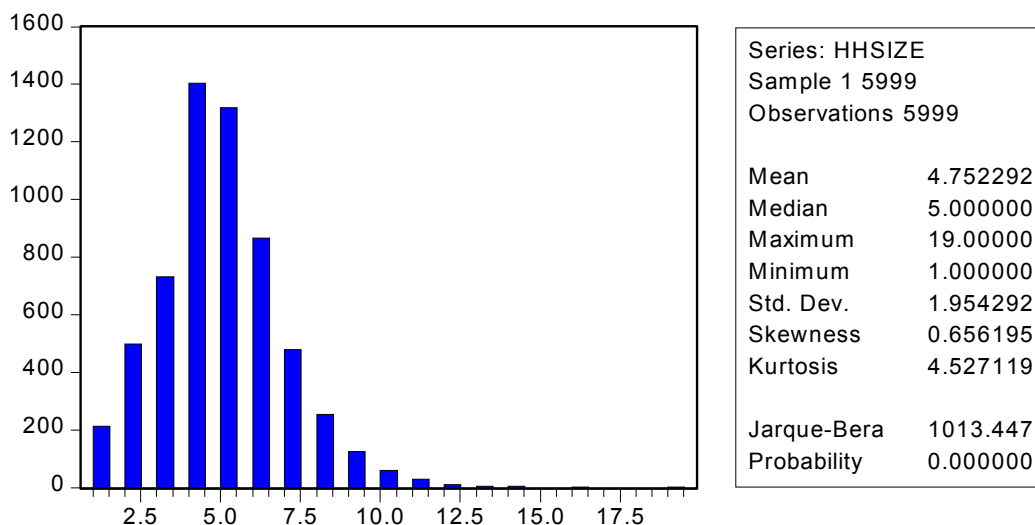
Một biến liên tục là biến có một số vô hạn các kết quả, như là chiều cao của một đứa trẻ.

Nhiều biến kinh tế được đo bằng những đơn vị đủ nhỏ để chúng ta coi chúng như là liên tục, mặc dù thực ra chúng là rời rạc. Ví dụ, xét chỉ tiêu hàng năm của hộ gia đình. Chúng ta có thể đo được chi tiêu đối với sự gia tăng là một đồng, vì thế dữ liệu này là rời rạc. Tuy nhiên, các bước gia tăng một đồng thì tương đối nhỏ so với mức chi tiêu trung bình, vì vậy nói chung chúng ta có thể coi chi tiêu hàng năm như là một biến liên tục.

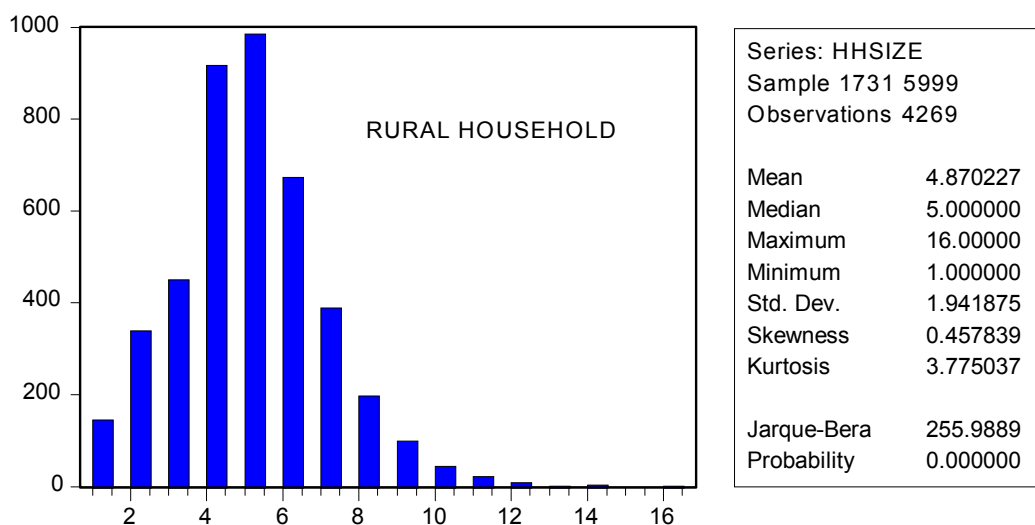
Một số ví dụ tiếp theo :

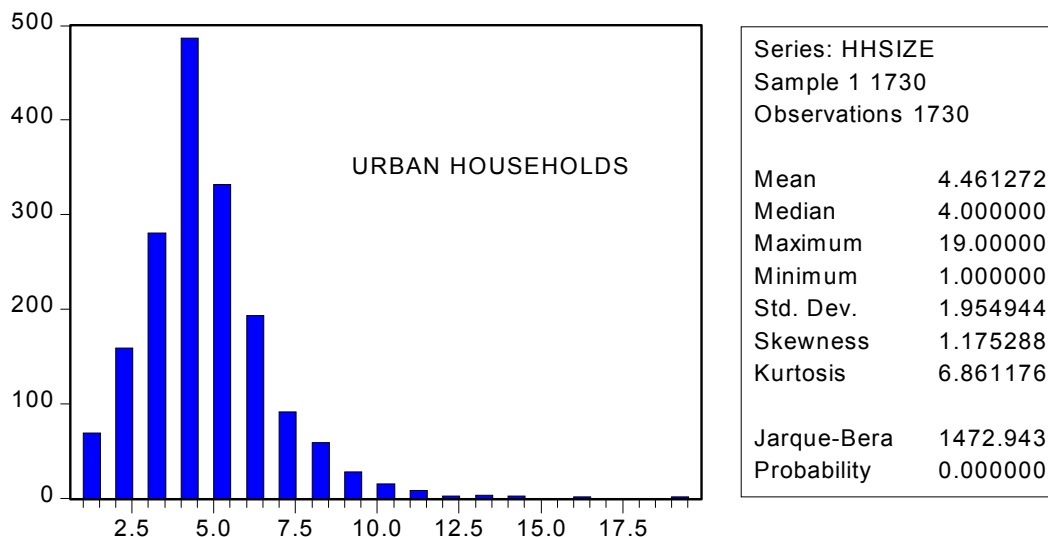
Dưới đây là biểu đồ tần suất và một số thống kê mẫu theo qui mô hộ gia đình của 5999 hộ gia đình đã được điều tra theo VNLSS năm 1998.

Dữ liệu này là dữ liệu chéo đối với năm 1998, trong đó đơn vị chéo là hộ gia đình. Qui mô hộ gia đình là một biến ngẫu nhiên rời rạc.



Chúng ta biết rằng qui mô hộ gia đình có tác động lên sự nghèo khó và tình trạng nghèo khó đó thường xảy ra tại khu vực nông thôn. Chúng ta lại xét chính dữ liệu này một lần nữa nhưng phân theo qui mô hộ gia đình nông thôn và thành thị một cách riêng biệt .





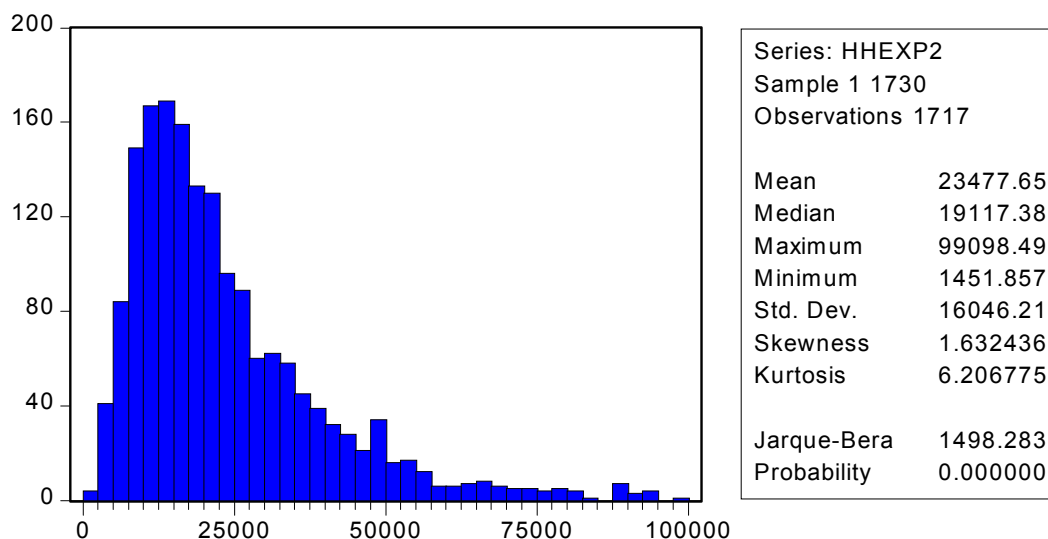
Qui mô hộ gia đình nông thôn có trung vị là 5 còn thành thị có trung vị là 4. Nhà thống kê luôn muốn biết rằng liệu sự sai biệt này có "có ý nghĩa thống kê" hay không.

Chú ý rằng các thang đơn vị trên hai biểu đồ tần suất này không thể so sánh trực tiếp vì các cỡ mẫu khác nhau. Chúng ta có thể thiết lập lại những biểu đồ tần suất này dưới dạng các tần suất tương đối chứ không phải là tần suất tuyệt đối để đề cập tới vấn đề này.

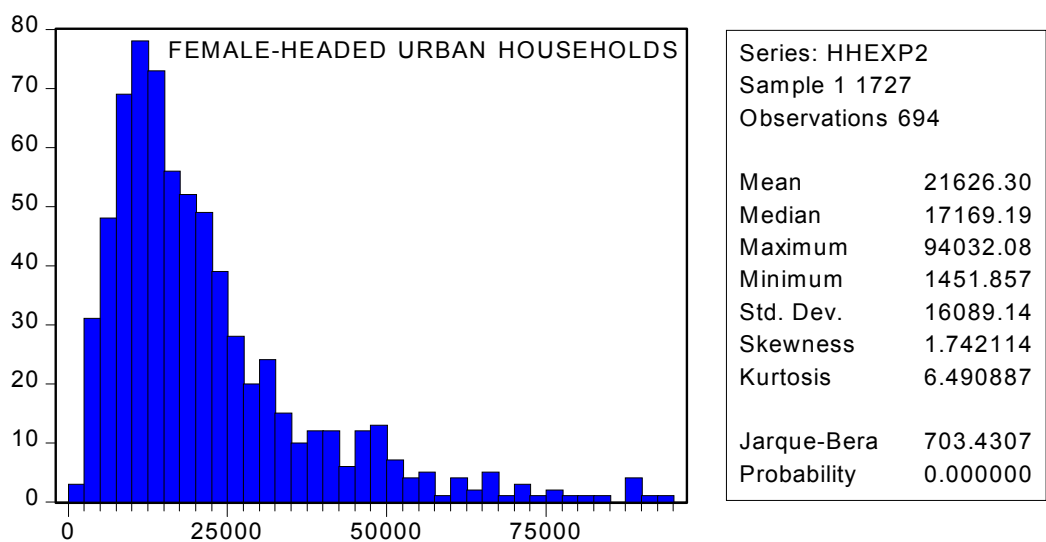
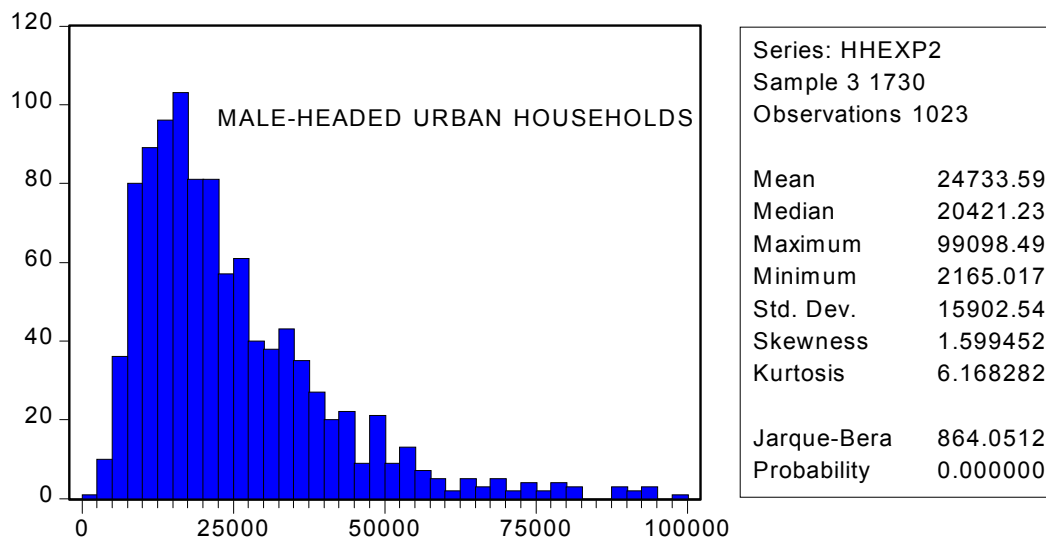
Ý nghĩa thống kê là gì? Ý nghĩa thống kê có nghĩa là chênh lệch giữa một trị thống kê đang quan tâm so với zero là đủ lớn để không thể xuất hiện ngẫu nhiên. Sau này chúng ta sẽ đề cập vấn đề này hơn nữa một cách chính thức.

Bây giờ chúng ta hãy xét một biến ngẫu nhiên liên tục : chi tiêu hàng năm của hộ gia đình thành thị với bước gia tăng 1000 đồng.

Một lần nữa dữ liệu này là chéo và đơn vị dữ liệu chéo là hộ gia đình.



Chúng ta có thể tò mò là liệu có phải các hộ gia đình mà chủ hộ là nữ nghèo hơn các hộ gia đình mà chủ hộ là nam hay không. Chúng ta có thể nghiên cứu điều này bằng cách tạo ra các biểu đồ tần suất riêng biệt:



Chi tiêu trung bình đối với hộ gia đình mà chủ hộ là nam: 24.734;

Chi tiêu trung bình đối với hộ gia đình mà chủ hộ là nữ : 21.626.

Một lần nữa, chúng ta lại hỏi liệu sự khác biệt trong hai giá trị trung bình này có ý nghĩa thống kê hay không.

Một câu hỏi khác mà chúng ta là liệu các thuộc tính khác của hộ gia đình ngoài thuộc tính giới tính của chủ hộ có giải thích cho khác biệt này hay không. Chúng ta cần các công cụ

ting vi hơn để tìm hiểu những đóng góp của nhiều biến (multiple variables) vào một biến phụ thuộc mà chúng ta đang quan tâm. Chúng ta sẽ dành vài tuần sau nữa để nghiên cứu những điều này.

Các lý thuyết, mô hình và tính ngẫu nhiên

Các trắc nghiệm có kiểm soát

Nếu chúng ta có thể thực hiện các trắc nghiệm có kiểm soát, thì chúng ta sẽ đặt các biến kinh tế (economic agents) vào những thay đổi dưới dạng khuyến khích, cơ hội hay ràng buộc (đó là giữ tất cả mọi yếu tố khác không đổi) và chúng ta sẽ quan sát những thay đổi trong hành vi của những biến kinh tế này. Ví dụ, chúng ta có thể thay đổi thu nhập của hộ gia đình và quan sát các thay đổi trong hành vi chi tiêu hay tiết kiệm. Tuy nhiên, trắc nghiệm với các thay đổi về vị trí hay giới tính của chủ hộ là không thực tế.

Đa số các dữ liệu kinh tế quan tâm không được hình thành bởi các trắc nghiệm có kiểm soát . Sự thực thường là không có gì được giữ không đổi !

Chúng ta có thể nghĩ tới nhiều điều tác động đến chiều hướng chi tiêu của hộ gia đình : số trẻ em, tình trạng ốm đau của một thành viên trong gia đình, vị trí. Không biến nào trong những biến này được kiểm soát trong các biểu đồ tần suất mà chúng ta đã xem xét.

Chúng ta hãy phát triển một ví dụ khác trong ngoại thương. Nhiều quốc gia đã tự do hoá các cơ chế ngoại thương của mình và nhiều nước đã trải qua các giai đoạn chuyển đổi sang nền kinh tế theo định hướng thị trường, hơn nữa các nhà kinh tế vẫn đang say sưa với cuộc tranh luận đáng kể về các mối quan hệ giữa những chính sách và những kết quả kinh tế .

Nếu một quốc gia giảm thuế nhập khẩu trung bình của mình đi 10% thì chúng ta có thể kỳ vọng sự gia tăng trong tăng trưởng kinh tế lớn tới cỡ nào?

Nếu chúng ta có nhiều quốc gia tương tự ở với những điều kiện tương tự, thì chúng ta có thể tiếp cận việc thực hiện một trắc nghiệm có kiểm soát : chúng ta có thể tự do hoá các cơ chế ngoại thương của những quốc gia này ở nhiều mức độ khác nhau và quan sát các kết quả. Có lẽ là chúng ta sẽ có thể nhận dạng một mối quan hệ giữa mức độ tự do hoá và tăng trưởng sau đó.

Tuy nhiên, trong trường hợp các quốc gia mà điều kiện của họ rất khác nhau thì nhiệm vụ nhận dạng mối quan hệ giữa những thay đổi chính sách với kết quả của nó sẽ trở nên rất phức tạp .

Có nhiều ví dụ về những mối quan hệ kinh tế mà chúng ta muốn quan sát và đo lường, nhưng chúng ta phát hiện rằng khả năng trực tiếp làm như vậy của chúng ta bị hạn chế do thực tế là nhiều biến cùng thay đổi.

Các lý thuyết và các mô hình

Thế giới kinh tế rất phức tạp và chúng ta sử dụng lý thuyết kinh tế để cố gắng tổ chức cách thức mà chúng ta khái niệm hoá những mối tương quan giữa các biến. Các mô hình

kinh tế là những trình bày cụ thể, thường là những trình bày dạng toán của các lý thuyết kinh tế. Một lý thuyết hay mô hình kinh tế tốt tập trung vào những nét đặc trưng quan trọng nhất của mối tương quan đang xét và loại bỏ các đặc trưng không quan trọng. Do đó, **các lý thuyết hay mô hình kinh tế là những sự đơn giản hoá có cân nhắc từ thế giới thực.**

Do các mô hình là những sự đơn giản hoá, nên chúng ta không ngạc nhiên là chúng không thích hợp hoàn hảo với dữ liệu. Những sai biệt giữa các mô hình và dữ liệu thường được coi như những yếu tố ngẫu nhiên có các tính chất thống kê đã được xác định rõ.

Ngoài ra, tính ngẫu nhiên có thể đi vào dữ liệu kinh tế vì nó là bản chất vốn có trong quá trình hình thành dữ liệu: mất điện ngẫu nhiên làm cho sản lượng biến động, các sự cố ngẫu nhiên làm cho công nhân bỏ sót công việc, doanh số bán hàng có thể phụ thuộc vào bao nhiêu người đi vào cửa hàng trong một ngày nhất định.

Kinh tế lượng cổ điển

Trong kinh tế lượng cổ điển, chúng quan tâm tới dữ liệu mà chúng ta có được coi như là kết quả của một phép lập. Chúng ta sử dụng các phương pháp thống kê để ước lượng những mối quan hệ quan tâm, trong khi đó lại kiểm soát những biến mà chúng ta không thể cố định nó. Ngoài ra, chúng ta sử dụng các phương pháp thống kê để đánh giá tác động của các nguồn gốc tạo ra ngẫu nhiên khác nhau lên sự xác đáng và chuẩn xác trong kết quả của chúng ta.

CÁC BIẾN NGẪU NHIÊN

Một định nghĩa chính thức của biến ngẫu nhiên là : **"một biến ngẫu nhiên là một qui tắc hay một hàm số để gán các giá trị bằng số cho những kết quả của một trắc nghiệm ngẫu nhiên."**

Chúng ta sẽ xem xét hai dạng biến ngẫu nhiên: rời rạc và liên tục. Bảng trực giác, một biến ngẫu nhiên rời rạc có tập hợp các kết quả cơ sở có thể đếm được; một biến ngẫu nhiên liên tục có các kết quả nằm dọc trên một đoạn của trục số thực. Chúng ta sẽ làm việc trước tiên với các biến ngẫu nhiên rời rạc.

Trắc nghiệm: thấy hai súc sắc và tính tổng. Trắc nghiệm ngẫu nhiên bao gồm việc thấy súc sắc này. Nhà nghiên cứu tính xem xuất hiện bao nhiêu chấm trên mặt từng súc sắc và tính chúng. Dựa trên trắc nghiệm này chúng ta có thể xác định nhiều biến ngẫu nhiên.

Gọi X_1 là số các chấm thể hiện trên súc sắc thứ nhất. Những kết quả có thể có của biến ngẫu nhiên X_1 này là $\{1, 2, 3, 4, 5, 6\}$.

Gọi X_2 là số các chấm thể hiện trên súc sắc thứ hai. Những kết quả có thể có của biến ngẫu nhiên X_2 này là $\{1, 2, 3, 4, 5, 6\}$.

Đặt $S = X_1 + X_2$. Những kết quả có thể có của biến ngẫu nhiên này là $\{1, 2, \dots, 12\}$.

Chúng ta thường quan tâm tới xác suất để các biến ngẫu nhiên nhận những giá trị nhất định. Ví dụ, tôi giữ trong tay tôi một con súc sắc và tôi muốn biết xác suất để kết quả của X_1 bằng 6 là gì.

Tôi có thể viết xác suất này là $P(X_1 = 6)$. Nó là gì? $P(X_1 = 6) = 1/6$. Câu trả lời hiển nhiên này được tạo ra bởi một khái niệm được biết là "**nguyên tắc suy luận không đầy đủ (principle of insufficient reason)**."

Nguyên tắc suy luận không đầy đủ áp dụng cho những trường hợp trong đó bản chất vật lý của trắc nghiệm này đề xuất rằng tất cả mọi kết quả có thể có có khả năng xảy ra như nhau. Nếu có K kết quả có thể xảy ra, và nếu chúng có khả năng xảy ra như nhau, thì xác suất của bất cứ kết quả nhất định nào đều là $\left(\frac{1}{K}\right)$.

Bây giờ hãy xét trò chơi được gọi là "thảy súc sắc". Trong trò chơi này, hai súc sắc được thả và tính tổng của số các chấm hiện ra. Đây là các qui tắc:

Nếu $S = (7 \text{ hay } 11)$ thì người chơi thắng.

Nếu $S = (2, 3, \text{ hay } 12)$ thì người chơi thua.

Nếu $S = (4, 5, 6, 8, 9, \text{ hay } 10)$ thì kết quả này được gọi là "điểm" của người chơi và người chơi phải tiếp tục thả cho tới khi cô ta thấy ra được điểm của mình một lần nữa hoặc thấy ra $S = 7$. Nếu cô ta thấy ra điểm của mình, thì cô ta thắng; nếu cô ta thấy ra 7, thì cô ta thua.

Xác suất để một người chơi thả súc sắc thắng là gì? Để tính được giá trị này, chúng ta cần một số công cụ mạnh hơn để tính các xác suất.

Các không gian mẫu và các biến cố (events)

Xét một trắc nghiệm ngẫu nhiên là chúng ta có thể tạo ra một tập hợp rời rạc các kết quả cơ bản

$$S = \{o_1, o_2, \dots, o_k\}$$

Tập hợp S được gọi là không gian mẫu của trắc nghiệm này.

Các biến cố trong trắc nghiệm này được xác định dưới dạng các tập con của S . Ví dụ, chúng ta có thể xác định các biến cố sau:

$$A = \{o_1 \text{ hoặc } o_2\}$$

$$B = \{o_2 \text{ hoặc } o_3 \text{ hoặc } o_5\}$$

Chúng ta cũng có thể định nghĩa các biến cố dưới dạng mối tương quan giữa các biến cố được xác định trước đây. Ví dụ, chúng ta có thể thực hiện trắc nghiệm và xác định biến cố C như là :

$$C = A \text{ hoặc } B$$

Chúng ta cũng có thể xác định biến cố D như là :

$$D = A \text{ và } B$$

Ở đây, chúng ta đòi hỏi các ý nghĩa cụ thể đối với các từ "hoặc" và "và". Theo thuật ngữ của lý thuyết tập hợp, $C = A \text{ hoặc } B$ có nghĩa rằng C là tập hợp của tất cả mọi biến cố cơ bản khi A xảy ra hoặc khi B xảy ra hoặc là khi cả A và B xảy ra. Đây là hợp của các biến cố A và B và được viết chính thức là:

$$C = A \cup B$$

Tiếp theo, theo thuật ngữ của lý thuyết tập hợp, $D = A \text{ và } B$ có nghĩa rằng D là tập hợp của tất cả mọi biến cố cơ bản khi A và B cùng xảy ra. Đó là giao của các biến cố A và B và được viết chính thức là:

$$D = A \cap B$$

Dưới dạng các định nghĩa trước đây của A và B, chúng ta có :

$$C = A \cup B = \{o_1, o_2, o_3, o_5\} \quad \text{và} \quad D = A \cap B = \{o_2\}$$

Các tiên đề xác suất

Nếu chúng ta biết xác suất của những kết quả cơ bản trong trắc nghiệm của chúng ta, chúng ta có thể sử dụng một tập hợp các tiên đề xác suất để tính xác suất của những biến cố khác nhau.

$$P(S) = 1$$

$$0 \leq P(A) \leq 1$$

$$P(C) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Nếu chúng ta trở lại với ví dụ thấy một súc sắc duy nhất thì chúng ta xác định $S = \{1, 2, 3, 4, 5, 6\}$

$$A = \{1, 2\}$$

$$B = \{2, 3, 5\}$$

Sử dụng nguyên lý lý do không đầy đủ để định các xác suất $P(o_i) = 1/6$ chúng ta tính:

$$P(A) = 2/6$$

$$P(B) = 3/6$$

$$P(C) = 5/6 - 1/6 = 4/6$$

Định nghĩa : Chia không gian mẫu là lựa chọn các tập hợp gồm những biến cố loại trừ lẫn nhau, mà hợp của chúng lại hình thành không gian mẫu. Ví dụ, nếu chúng ta xác định $A = \{1, 3, 5\}$ và $B = \{2, 4, 6\}$, thì A và B tạo thành một sự chia (partition) không gian mẫu đối với X_1 . Dễ dàng chứng minh rằng chúng là loại trừ lẫn nhau và rằng hợp của chúng lại bao gồm tất cả các kết quả cơ bản chứa trong không gian mẫu.

Điều gì xảy ra nếu chúng ta có một trắc nghiệm trong đó nguyên tắc suy luận không đầy đủ không áp dụng ? Hãy xét tình huống trong đó chúng ta chơi với một súc sắc "gài bẫy" . Chúng ta có thể làm gì ?

Một khả năng là thực hiện trắc nghiệm này với số lần đủ lớn (n) và tính sự xuất hiện (occurrences) của từng kết quả ($n_i, i = 1, \dots, k$). Sau đó chúng ta có thể ước lượng xác suất của từng kết quả như sau đây:

$$\hat{p} = \left(\frac{n_i}{n} \right)$$

Chúng ta gọi đại lượng $\left(\frac{n_i}{n} \right)$ là tần suất tương đối của kết quả i. Điều này cung cấp cơ sở cho định nghĩa tần suất tương đối của xác suất :

$$P(o_i) = \lim_{n \rightarrow \infty} \left(\frac{n_i}{n} \right)$$

Đây là khái niệm xác suất mà chúng ta sẽ sử dụng trong suốt khoá học này.

Phân phối xác suất : Biến ngẫu nhiên rời rạc.

Xét một tập hợp các biến cố hình thành từ việc tách (partition) một không gian mẫu. Hãy xây dựng một biến ngẫu nhiên rời rạc tương ứng với phân tách này. Khi đó, bảng liệt kê các giá trị của biến ngẫu nhiên này và các xác suất tương ứng là phân phối xác suất của biến ngẫu nhiên rời rạc này.

Ví dụ , thả con súc sắc và xác định các biến cố sau :

$$A = \{1\}$$

$$B = \{2 \text{ hoặc } 3\}$$

$$C = \{4 \text{ hoặc } 5 \text{ hoặc } 6\}$$

Gán biến ngẫu nhiên này như sau đây : $X = 0$ nếu A xảy ra

$$\begin{aligned} X &= 1 \text{ nếu B xảy ra} \\ X &= 2 \text{ nếu C xảy ra} \end{aligned}$$

Phân phối xác suất của X là :

x	P(X = x)
0	1/6
1	2/6
2	3/6

Đồ thị của một phân phối xác suất rời rạc trông giống các cột tại giá trị của từng kết quả ; chiều cao của mỗi cột bằng xác suất đang xét.

Thường thường, chúng ta không thể tính được xác suất của các biến cố khác nhau dựa trên kiến thức *tiên nghiệm*. Trong những trường hợp như vậy, chúng ta có thể lấy một mẫu ngẫu nhiên từ biến ngẫu nhiên đang quan tâm và ước tính các xác suất như nêu trên đây.

Trong trường hợp súc sắc không công bằng, chúng ta sẽ thấy súc sắc nhiều lần và ước tính xác suất của nhiều kết quả khác nhau bằng các tần suất tương đối mà chúng xuất hiện.

Bây giờ, hãy tưởng tượng rằng chúng ta quan tâm tới phân phối xác suất của một biến ngẫu nhiên rời rạc, chẳng hạn qui mô hộ gia đình tại Việt nam . Sẽ tốn kém khi điều tra tất cả mọi hộ gia đình với mục đích tính tỉ trọng thực tế của các hộ gia đình thuộc ứng với từng qui mô hộ gia đình, vì thế chúng ta chỉ rút một mẫu ngẫu nhiên và tính tần suất tương đối cho qui mô hộ gia đình cho mẫu của chúng ta.

Nếu chúng ta vẽ đồ thị phân phối tần suất này (hay phân phối tần suất tương đối), thì chúng ta có một hình ảnh về phân phối xác suất "ước lượng" của biến ngẫu nhiên này.

Tính chất của các biến ngẫu nhiên

Hãy ghi nhận rằng cả bảng liệt kê bao gồm phân phối xác suất lẫn sự trình bày bằng đồ thị của nó đều chứa nhiều thông tin không ở dạng tổng hợp. Chúng ta sẽ sử dụng các mô hình toán học để mô tả triệt để những phân phối xác suất của các biến ngẫu nhiên, nhưng chúng ta có thể tổng hợp một số nét đặc trưng chính của chúng bằng cách sử dụng khái niệm *kỳ vọng toán học*.

Kỳ vọng toán học (còn được gọi là giá trị kỳ vọng hay giá trị trung bình) của một biến ngẫu nhiên là một thước đo xu hướng trung tâm của biến ngẫu nhiên đó. Giá trị trung bình của một biến ngẫu nhiên rời rạc được xác định là :

$$\mu = E[X] = \sum_i x_i P(x_i) \quad i = 1, \dots, k$$

Giá trị kỳ vọng là một trung bình có trọng số, trong đó những trọng số tương ứng với các xác suất được gán với mỗi giá trị có thể có của biến ngẫu nhiên này.

Nếu chúng ta tính giá trị kỳ vọng của X_1 , là kết quả đối với một lần thấy duy nhất của một súc sắc, chúng ta có tính toán sau:

$$\mu = E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \dots + 6 \times \frac{1}{6} = 3,5$$

Điều thú vị là giá trị trung bình này không phải là một giá trị có thể xảy ra với bất cứ lần thấy cho trước nào, mà lại có ý nghĩa là các kết quả có thể xảy ra xoay quanh giá trị trung bình này. Điều này có nghĩa là nếu chúng ta thấy một súc sắc nhiều lần và tính trung bình các kết quả, thì chúng ta sẽ nhận được một số gần bằng 3,5.

Chúng ta sẽ khám phá rằng điều hữu ích là tính một số đo tổng hợp cho độ lệch của các kết quả nhất định so với giá trị trung bình. Tính toán này cũng là một giá trị kỳ vọng: nó là giá trị kỳ vọng của bình phương độ lệch giữa giá trị kết quả với giá trị trung bình này. Chúng ta sử dụng bình phương độ lệch vì độ lệch kỳ vọng luôn là zero: vì các độ lệch dương và âm triệt tiêu lẫn nhau. Phương sai được xác định như là:

$$\sigma^2 = \text{VAR}[X] = E[(X - \mu)^2] = \sum_i (x_i - \mu)^2 P(x_i)$$

Vì nhiều mục đích, chúng ta muốn đo sự phân tán có cùng đơn vị đo với giá trị trung bình. Số đo này được gọi là Độ lệch chuẩn: nó là căn bậc hai dương của phương sai:

$$\sigma = \sqrt{\sigma^2}$$

Chúng ta thường làm việc với các phép biến đổi các biến ngẫu nhiên. Ví dụ, chúng ta có thể nhân giá trị hàng xuất khẩu tính bằng nội tệ với tỉ giá hối đoái để có được giá trị hàng xuất khẩu bằng đô la.

Dễ dàng thiết lập những tính chất của các phép biến đổi biến ngẫu nhiên bằng cách sử dụng các phép tính đại số cơ bản. Chúng là nội dung chính của Bài tập 1.

Mật độ xác suất : Biến ngẫu nhiên liên tục .

Nếu chúng ta nghĩ về tiếp cận tần suất tương đối tới xác suất, và chúng ta tưởng tượng việc lựa chọn một quan sát ngẫu nhiên, dường như rõ ràng là xác suất của việc thu được *chính xác* một giá trị nhất định phải là zero. Mặt khác, nếu chúng ta đặt vấn đề dưới dạng khoảng, thì việc xác định xác suất này là đơn giản.

Hãy tưởng tượng rằng đang mưa và rằng Anh/Chị đặt một thước đo trên mặt đất. Xác suất để hạt mưa sau sẽ rơi vào giữa 0 và 10 cm là gì? Xác suất để hạt mưa sau sẽ rơi vào giữa 10 và 20 cm là gì?

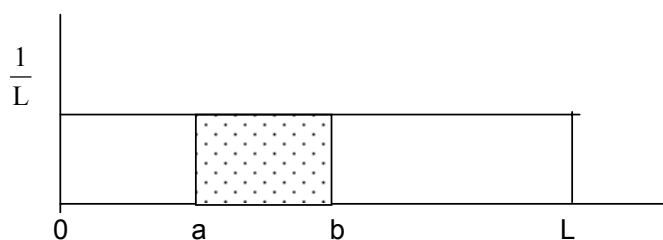
Chúng ta có thể chia thước đo này thành 10 bước với khoảng cách là 10 cm mỗi bước, sau đó chúng ta có thể sử dụng nguyên tắc suy luận không đầy đủ (principle of insufficient reason) để gán các xác suất. Xác suất để một hạt mưa rơi vào bất cứ khoảng cụ thể nào sẽ

bằng $1/k$, trong đó k là số các khoảng trong thước. Trong trường hợp này, việc tính xác suất để một hạt mưa rơi vào một khoảng có bất cứ độ dài cụ thể nào thì thật là đơn giản.

Bây giờ giả sử rằng chúng ta muốn tạo ra một trình bày đồ thị cho xác suất giọt mưa này. Chỉ ra rằng không gian mẫu của biến ngẫu nhiên liên tục X được cho bởi khoảng $[0, L]$. Mật độ xác suất đều được định nghĩa là

$$\frac{1}{L - 0}$$

Vẽ đồ thị mật độ xác suất đều này lên trục X như sau :



Xác suất để $(a \leq X \leq b)$ được tính là

$$\frac{b - a}{L}$$

Vùng tô đậm là xác suất cần tính.

Hàm mật độ xác suất là một hàm số $f(x)$ có các tính chất sau :

- 1) $f(x) \geq 0$
- 2) Diện tích nằm dưới pdf trong một khoảng đã cho chính là xác suất để biến ngẫu nhiên này rơi vào khoảng đã cho.

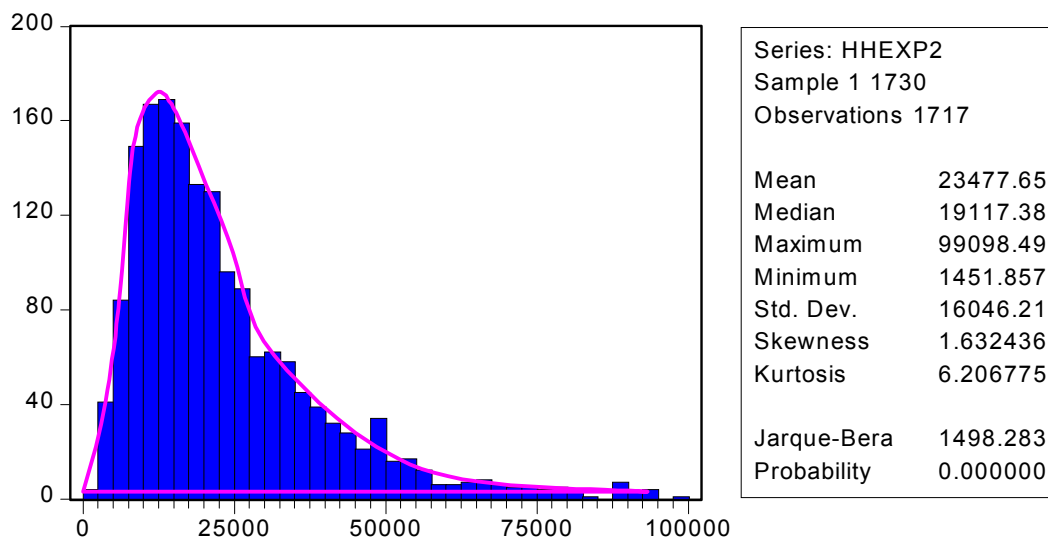
$$P(a \leq x \leq b) = \int_a^b f(x) dx$$

- 3) $\int_s f(x) dx = 1$

Giả sử chúng ta muốn biết về phân phối xác suất của một biến ngẫu nhiên liên tục như chi tiêu hộ gia đình. Chúng ta không có đủ ngân sách để điều tra tất cả mọi hộ gia đình, vì vậy

chúng ta chọn một mẫu ngẫu nhiên. Do xác suất để một thành phần (element) trong mẫu ngẫu nhiên của chúng ta rơi vào một khoảng cho trước được cho bởi mật độ xác suất chưa biết này, nên chúng ta kỳ vọng tần suất tương đối trong mẫu của chúng ta phản ánh xác suất thực sự kết hợp trong một khoảng cụ thể.

Một lần nữa hãy xét các biểu đồ tần suất về chi tiêu hộ gia đình. Chúng ta có thể coi những giá trị thống kê này là "các ước lượng" của pdf thực sự chưa biết.



Các kỳ vọng của các biến ngẫu nhiên liên tục

Các kỳ vọng (các giá trị trung bình và phương sai) của các biến ngẫu nhiên liên tục tương tự như kỳ vọng của các biến ngẫu nhiên rời rạc. Chỉ có một khác biệt là dấu tích phân thay bằng dấu tích phân.

$$\mu = E[X] = \int_S x f(x) dx$$

Chúng ta sẽ không có thời gian để nghiên cứu về tích phân. Chúng ta sẽ chỉ ghi nhận rằng những tính chất toán học đối với các kỳ vọng của biến ngẫu nhiên rời rạc được minh họa trong Bài tập 1 đúng với các kỳ vọng của biến ngẫu nhiên liên tục.

