

Mô tả dữ liệu

Trong các bài giảng trước và các bài tập chúng ta đã xem các đồ thị phân phối tần suất của các biến riêng biệt. Các phân phối tần suất như vậy gọi là biểu đồ tần suất.

Nếu ta nghĩ đến một phân phối xác suất hoặc đồ thị của hàm pdf là đại diện phân phối tổng thể, ta có thể nghĩ đến biểu đồ tần suất dựa vào một mẫu của tổng thể đó và xem đó là một cách để “ước lượng” hàm phân phối mà chúng ta quan tâm.

Biểu đồ tần suất là cách biểu diễn trực quan rất tốt cho phân phối của một biến, nhưng biểu đồ dạng này không dễ tóm tắt bằng một câu ngắn gọn, và chúng không thể sử dụng trực tiếp trong các tính toán thống kê.

Các phân phối và các hàm pdf có thể được biểu diễn bằng các công thức toán học: một phân phối được mô tả một cách đầy đủ qua dạng hàm số và tập hợp các tham số của nó (trong nhiều trường hợp, chỉ cần một hoặc hai tham số là đủ). Thông thường, chúng ta quan tâm đến việc ước lượng các tham số chưa biết cho phân phối của một tổng thể nào đó.

Nếu chúng ta nghĩ về sự biểu hiện của một biểu đồ tần suất, thì biểu đồ này có thể được mô tả qua bốn tính chất sau:

- Xu hướng trung tâm là vị trí “giữa” của phân phối.
- Mức độ phân tán là khoảng cách từ một quan sát điển hình đến vị trí “giữa”.
- độ trôi (skewness), là mức độ không đối xứng của biểu đồ tần suất.
- độ nhọn (kurtosis), là độ phẳng hay nhọn của phân phối so với phân phối chuẩn.

Xu hướng trung tâm

Trung bình (của tổng thể và của mẫu)

Trung bình tổng thể: $\mu_X = E[X]$

Trung bình mẫu: $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$

Trung vị (của tổng thể và của mẫu)

Trung vị tổng thể:

Với biến ngẫu nhiên liên tục X , nếu $P(X \leq M_d) = 0.50$ thì M_d là số trung vị.

Trung vị mẫu:

Nếu số phần tử của mẫu là số lẻ thì số trung vị là giá trị “ở giữa” của mẫu.

M Daniel Westbrook

Nếu số phần tử của mẫu là số chẵn thì số trung vị là trung bình cộng của hai giá trị “ở giữa” của mẫu.

Độ phân tán

Phương sai (của tổng thể và của mẫu)

Phương sai tổng thể: $\sigma_X^2 = E[(X - \mu_X)^2]$

Phương sai mẫu: $s_X^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$

hoặc: $\hat{\sigma}_X^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$

Độ lệch chuẩn (của tổng thể và của mẫu)

Độ lệch chuẩn tổng thể: $\sigma_X = \sqrt{\sigma_X^2}$

Độ lệch chuẩn mẫu: $s_X = \sqrt{s_X^2}$

hoặc: $\hat{\sigma}_X = \sqrt{\hat{\sigma}_X^2}$

Tính không đối xứng

Độ trôi (của tổng thể và của mẫu)

Độ trôi tổng thể: $E\left[\left(\frac{X - \mu_X}{\sigma}\right)^3\right]$

Độ trôi mẫu: $S = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{X}}{\hat{\sigma}}\right)^3$

Lưu ý: với phân phối chuẩn, hệ số của độ trôi bằng không.

Độ nhọn (hay độ phẳng)

Độ nhọn (của tổng thể và của mẫu)

$$\text{Độ nhọn tổng thể: } E \left[\left(\frac{X - \mu_X}{\sigma} \right)^4 \right]$$

$$\text{Độ nhọn mẫu: } \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{X}}{\hat{\sigma}} \right)^4$$

Lưu ý: với phân phối chuẩn, hệ số độ nhọn là 3. Nếu độ nhọn lớn hơn 3 thì phân phối này là nhọn (leptokurtic): gầy với các đuôi dài. Nếu độ nhọn nhỏ hơn 3 thì phân phối là phẳng (platykurtic): ít nhọn với các đuôi ngắn.

Thống kê Jarque-Bera

Chúng ta thường quan tâm đến việc kiểm tra xem có phải các giá trị của một biến ngẫu nhiên nào đó được lấy từ một phân phối chuẩn hay không. Nếu đúng như thế thì ta mong đợi là hệ số độ trôi sẽ bằng 0 và hệ số độ nhọn bằng 3. Thống kê Jarque-Bera tính độ sai biệt của các giá trị trong một mẫu nào đó so với kỳ vọng của chúng ta như sau:

$$JB = \frac{(n-1)}{6} \left(S^2 + \frac{(K-3)^2}{4} \right)$$

Nếu dữ liệu của chúng ta tuân theo phân phối chuẩn thì giá trị thống kê này sẽ gần bằng 0; nếu dữ liệu bị trôi hoặc có độ nhọn khác với độ nhọn của phân phối chuẩn thì giá trị thống kê này sẽ là một con số lớn.

Chúng ta sẽ sớm thảo luận chính thức về độ “lớn” của thống kê này.

Mối quan hệ tuyến tính giữa các biến

Hệ số tương quan

Cách đây hai ngày chúng ta đã định nghĩa hệ số tương quan:

$$\rho = \frac{\text{COV}(X, Y)}{\sigma_X \sigma_Y}$$

Hệ số tương quan của mẫu được tính như sau cho các cặp quan sát có thứ tự:

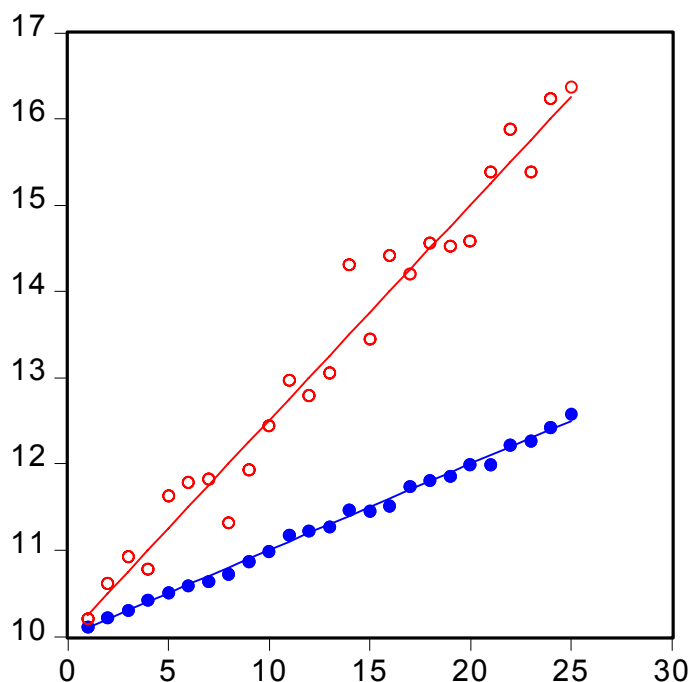
$$r = \frac{s_{XY}}{s_X s_Y} \quad \text{với} \quad s_{XY} = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})$$

Lưu ý rằng mỗi đặc trưng đo lường được thảo luận ở trên là một giá trị vô hướng mà nó có thể được tính toán và sử dụng trong các phép tính sau này.

M Daniel Westbrook

Biểu đồ phân tán

Hãy so sánh hai biểu đồ phân tán sau:



Cặp biến số nào có mối quan hệ “mạnh” hơn?

Hệ số tương quan giữa X và Y cho các chấm xanh là 0.9978.

Hệ số tương quan giữa X và Y cho các chấm đỏ là 0.9848.

Lưu ý rằng cả hai hệ số tương quan đều dương, cho thấy rằng độ dốc của đường thẳng là dương; nếu hệ số tương quan là âm thì các đường thẳng sẽ dốc xuống.

Độ dốc của đường xanh là 0.10

Độ dốc của đường đỏ là 0.25.

Nếu hệ số tương quan bằng 1.0 hoặc -1.0 thì điều này cho chúng ta biết rằng các quan sát sẽ nằm chính xác trên một đường thẳng, nhưng không cho ta biết thông tin gì về độ dốc của đường thẳng.

Để có được thông tin về độ dốc của đường thẳng, chúng ta phải sử dụng phân tích hồi quy.

Hồi quy là một cách mô tả dữ liệu

Chúng ta thường nhìn một đồ thị phân tán và muốn vẽ một đường thẳng “tốt nhất” mà đường thẳng này “thích hợp” với dữ liệu. Điều này có thể thực hiện một cách dễ dàng trong EViews.

M Daniel Westbrook

Đầu tiên, hãy chọn một biến số mà bạn muốn thể hiện trên trục hoành, nhấp chuột để làm đánh dấu biến số. Sau đó, chọn biến mà bạn muốn thể hiện trên trục tung bằng cách giữ phím CTRL rồi nhấp chuột vào vị trí biến này.

Khi cả hai biến quan tâm đã được đánh dấu, nhấp chuột kép lên vùng đó và chọn **Open Group**.

Sau đó nhấp **View / Graph / Scatter / Scatter With Regression** rồi quan sát kết quả.

Ước lượng

Các nhà kinh tế sử dụng kinh tế lượng để ước lượng các tham số cho các mối quan hệ tuyến tính giữa các biến. Các tham số này thường tương ứng với các khái niệm kinh tế như hệ số co giãn của cầu, xu hướng tiêu dùng biên, năng suất biên của lao động, suất sinh lợi theo quy mô, v.v.

Một ước lượng (estimator) là một trị thống kê của mẫu được dùng để ước lượng một tham số chưa biết.

Để có thể tin tưởng vào các ước lượng của chúng ta và để phát triển các kiểm định giả thiết về các tham số quan tâm, chúng ta cần thiết lập các tính chất của ước lượng.

Ước lượng “tốt” có những tính chất gì?

Không chệch: $E[\text{tham số}] = \text{giá trị thực của tham số chưa biết}$
Nói theo ngôn ngữ đời thường, tính không chệch tương ứng với sự đúng đắn (accuracy).

Hiệu quả: $\text{VAR}[\text{tham số}]$ nhỏ hơn phương sai của bất kỳ ước lượng không chệch nào khác.
Nói theo ngôn ngữ đời thường, tính hiệu quả tương ứng với sự chính xác (precise).

Phân phối xác suất đã biết: Để có thể phát biểu về xác suất của một ước lượng hay về các giá trị có thể có của một tham số chưa biết, hay để kiểm định giả thiết về một tham số chưa biết, ta cần biết phân phối xác suất của ước lượng.

Tính nhất quán: Trong những trường hợp khó tìm được phân phối chính xác của mẫu, xác định tính nhất quán lại dễ hơn. Nhất quán là một tính chất của mẫu lớn, được định nghĩa như sau:
Gọi $\hat{\theta}$ là ước lượng của tham số chưa biết θ . Nếu

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$

thì ta nói rằng ước lượng đó có tính nhất quán.

Điều này thường được viết là “giới hạn xác suất”: $\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta$

Rất may là trong hầu hết các trường hợp mà chúng ta quan tâm, chúng ta không cần phải lo tìm giới hạn của biểu thức xác suất. Điều kiện đủ để có tính nhất quán là cả độ chệch và phương sai đều tiến về không khi cỡ mẫu tăng lên vô cực.

Chọn mẫu ngẫu nhiên

Trong hầu hết mọi công việc, chúng ta cố gắng chọn các mẫu ngẫu nhiên. Thủ tục chọn mẫu ngẫu nhiên đơn giản đòi hỏi tất cả mọi kết hợp của các phần tử của mẫu đều có khả năng xảy ra như nhau. Các phần tử của một mẫu ngẫu nhiên đơn giản là độc lập thống kê. Chúng ta coi mỗi phần tử của mẫu là một biến ngẫu nhiên, bởi vì trước khi chọn ngẫu nhiên, kết quả của mỗi lần lấy mẫu là chưa biết.

Về trung bình, các mẫu ngẫu nhiên là “đại diện” của các tổng thể mà từ tổng thể đó các mẫu ngẫu nhiên này được lấy ra. Hơn nữa, việc chọn mẫu ngẫu nhiên có một lợi thế về kỹ thuật mà các anh chị đã chứng minh trong Bài tập 2, đó là: phương sai của tổng các biến ngẫu nhiên độc lập thống kê bằng tổng của các phương sai của chúng. Điều này giúp cho việc tính phương sai của nhiều trị thống kê trở nên đơn giản.

Phân phối chọn mẫu

Giả sử có một biến ngẫu nhiên X tuân theo một phân phối xác suất chưa biết, và chúng ta quan tâm đến việc ước lượng giá trị trung bình μ của phân phối đó.

Trực giác mách bảo chúng ta hãy dùng số trung bình \bar{X} của mẫu để ước lượng số trung bình μ mà chúng ta chưa biết.

Phân phối chọn mẫu của trung bình mẫu \bar{X} là gì?

Trước khi bắt đầu, cần xét xem chúng ta muốn nói gì khi đề cập đến nói “phân phối chọn mẫu” Sinh viên thường bị kẹt ở điểm này, vì họ cho rằng chỉ có một giá trị cố định của \bar{X} , và giá trị này chỉ phụ thuộc vào mẫu đã thực sự được lấy ra.

Có hai cách nghĩ về điều này. Cả hai đòi hỏi bạn phải coi \bar{X} là một biến ngẫu nhiên.

Đầu tiên, hãy giả sử là trước khi lấy mẫu, bạn liệt kê tất cả các mẫu có thể được lấy, rồi tính xác suất của chúng. Với mỗi mẫu, tính một giá trị cụ thể \bar{X} , và xác suất của giá trị trung bình mẫu này bằng với xác suất chọn ra mẫu tính ra giá trị trung bình mẫu đó. Liệt kê mọi giá trị có thể có của \bar{X} và xác suất của chúng (hoặc là mật độ xác suất của \bar{X}) sẽ tạo nên phân phối chọn mẫu mong muốn.

Cách nghĩ thứ hai về vấn đề này được gọi là “lấy mẫu vô hạn có ý thức”. Nói cách khác, hãy tưởng tượng một thí nghiệm trong đó nhà nghiên cứu chọn một mẫu có số quan sát mong muốn, rồi tính giá trị của \bar{X} cho mẫu đó; rồi chọn một mẫu khác với số quan sát mong muốn và tính giá trị \bar{X} lần thứ hai; v.v. Nhà nghiên cứu tiếp tục quá trình này với một số mẫu vô tận, rồi xây dựng phân phối chọn mẫu bằng một biểu đồ tần suất tương đối.

M Daniel Westbrook

Ý tưởng chính ở đây là có thể có nhiều giá trị của \bar{X} ứng với các xác suất khác nhau. Phân phối chọn mẫu của \bar{X} là phân phối xác suất của biến ngẫu nhiên \bar{X} .

Bài tập 4 sẽ cung cấp một số kinh nghiệm minh họa ý tưởng này.

Phân phối chọn mẫu của trung bình mẫu

Trong câu 2 của bài tập 3, các bạn đã chứng minh rằng $E[\bar{X}] = \mu_x$ và $\text{VAR}[\bar{X}] = \frac{\sigma_x^2}{n}$.

Với các kết quả này, dễ dàng thấy rằng trung bình mẫu có tính chất không chệch và nhất quán. Trung bình mẫu còn có tính chất hiệu quả: nghĩa là không có một ước lượng không chệch nào khác của tổng thể có thể tạo ra phương sai nhỏ hơn.

Để hoàn tất việc tìm ra phân phối chọn mẫu cho trung bình mẫu, chúng ta chỉ cần biết dạng hàm số của phân phối. Có ba trường hợp sau đây:

1. Định lý được nêu trong câu 1 của bài tập 3 đã chứng minh rằng một kết hợp tuyến tính của các biến ngẫu nhiên có phân phối chuẩn thì kết hợp này cũng có phân phối chuẩn. Nếu tổng thể mà từ đó các phần tử trong mẫu ngẫu nhiên của chúng ta được lấy ra có phân phối chuẩn thì chúng ta sẽ có:

$$\bar{X} \sim N(\mu_x, \sigma_{\bar{X}}^2) \text{ với } \sigma_{\bar{X}}^2 = \frac{\sigma_x^2}{n}$$

2. Định lý Giới hạn Trung tâm phát biểu rằng với một số điều kiện không chặt chẽ nhất định, khi qui mô mẫu tăng lên thì phân phối chọn mẫu của trung bình mẫu sẽ tiến đến phân phối chuẩn. Luật số lớn cho rằng khi cỡ mẫu tăng lên thì trung bình của mẫu sẽ gần bằng với trung bình tổng thể, nhưng Luật số lớn lại không cho chúng ta kết quả phân phối.

Cần thận trọng khi áp dụng Định lý Giới hạn Trung tâm. Nếu bạn xem xét phương sai của trung bình mẫu, bạn sẽ nhận thấy ngay rằng nó tiến về không khi cỡ mẫu tăng lên. Các phân phối có phương sai bằng không được gọi là các phân phối cá biệt (degenerate distributions). Để tránh sự cá biệt, phân phối chọn mẫu của trung bình mẫu trong trường hợp mẫu lớn được phát biểu như sau:

$$\frac{\sqrt{n}(\bar{X} - \mu_x)}{\sigma_x} \sim N(0, 1)$$

Trên thực tế, chúng ta chỉ đơn giản nói rằng \bar{X} được phân phối *xấp xỉ* như

$$\bar{X} \sim N(\mu_x, \sigma_{\bar{X}}^2).$$

3. Nếu cỡ mẫu nhỏ và phân phối của tổng thể không là phân phối chuẩn, nhưng nếu biết phân phối của tổng thể thì chúng ta có thể tìm được phân phối chọn mẫu chính xác của \bar{X} . Tuy nhiên, nếu phân phối của tổng thể không phải là phân phối chuẩn, và cũng chưa được biết thì phân phối chọn mẫu của \bar{X} cũng chưa được xác định.

Một câu hỏi rõ rệt là “khi nào thì cỡ mẫu được coi là đủ lớn để sự xấp xỉ với phân phối chuẩn được coi là đáng tin cậy?” Vấn đề này sẽ được nghiên cứu trong Bài tập 4.