BÀI 5: CƠ SỞ LÝ THUYẾT MẪU

Mục tiêu

Giới thiệu một số khái niệm cơ bản của Thống kê toán học, cụ thể là những vấn đề liên quan đến cặp phạm trù tổng thể và mẫu, đến các khái niệm thống kê, thống kê của đặc trưng mẫu và phân phối xác suất của thống kê đặc trưng mẫu, xem xét cụ thể các khái niệm đó trong một số trường hợp đặc biệt nhưng thường gặp trong thực hành.

Thời lượng

- 8 tiết

Các kiến thức cần có

- Khái niệm phương pháp mẫu
- Tổng thể nghiên cứu
- Định nghĩa
- Mô tả tổng thể
- Các số đặc trưng của tổng thể
- Mẫu ngẫu nhiên
- Các phương pháp lấy mẫu
- Định nghĩa mẫu ngẫu nhiên
- Mô tả mẫu ngẫu nhiên
- Thống kê (Statistics)
- Định nghĩa
- Các thống kê đặc trưng mẫu
- Mẫu ngẫu nhiên hai chiều
- Khái niệm
- Phương pháp mô tả mẫu
- Thống kê đặc trưng mẫu hai chiều
- Quy luật phân phối xác suất của một số thống kê
- Trường hợp biến ngẫu nhiên gốc có phân phối 0–1
- Trường hợp hai biến ngẫu nhiên gốc có phân phối 0–1
- Trường hợp biến ngẫu nhiên gốc có phân phối chuẩn
- Trường hợp hai biến ngẫu nhiên gốc có phân phối chuẩn

TÌNH HUỐNG KHỞI ĐỘNG BÀI

Tình huống

Điều tra mức thu nhập cá nhân trong một tháng (triệu đồng) ở huyện Đông Anh, ta có bảng số liệu mẫu sau:

Thu nhập	1-2	2-3	3-4	4-5	5-6	6-7
số người	10	8	5	7	3	2

Cần phải tính thu nhập bình quân đầu người và độ chênh lệch thu nhập để xác định mức sống của người dân và mức độ đồng đều về thu nhập trong vùng.

Câu hỏi

1. Thu nhập bình quân đầu người là bao nhiêu?

2. Độ chênh lệch thu nhập là bao nhiêu?

3. Độ chênh lệch bình quân hiệu chỉnh?

5.1. Khái niệm phương pháp mẫu

Bài toán:

Chúng ta cần nghiên cứu tính chất định tính hoặc định lượng của các phần tử trong một tập hợp nào đó. Khi đó ta có hai phương pháp thực hiện nghiên cứu

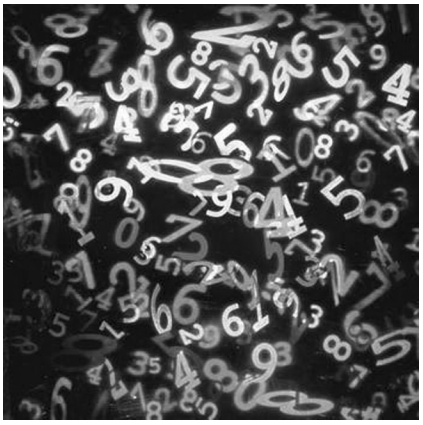
- Nghiên cứu toàn bộ các phần tử của tập hợp và ghi lại các đặc tính cần quan tâm. Khi thực hiện nghiên cứu toàn bộ ta gặp phải những hạn chế sau:
  - Phải trả chi phí lớn về kinh tế và thời gian do số lượng các phần tử trong tập toàn bộ quá lớn.
  - Có thể dẫn tới phá huỷ toàn bộ tập hợp cần nghiên cứu. Ví dụ nghiên cứu thời gian hoạt động của các thiết bị điện tử. Khi áp dụng phương pháp này sẽ dẫn tới phá huỷ toàn bộ các thiết bị điện tử.
  - Có những tập hợp mà ta không thể nghiên cứu được toàn bộ. Ví dụ như trong lĩnh vực khảo cổ học.

  Vậy ta thấy trong đa số các trường hợp nghiên cứu toàn bộ tập hợp là không khả thi.
- Nghiên cứu bộ phận, từ tập hợp nghiên cứu ta lấy ra một tập con và nghiên cứu toàn bộ các phần tử trong tập con đó và từ đó đưa ra kết luận cho các phần tử trong tập hợp nghiên cứu.

Phương pháp nghiên cứu thứ hai gọi là phương pháp nghiên cứu mẫu.

5.2. Tổng thể nghiên cứu

5.2.1. Định nghĩa



Tổng thể (population) là tập hợp các phần tử cần nghiên cứu tính chất định tính hoặc định lượng, số phần tử trong tổng thể gọi là cỡ của tổng thể, ký hiệu là N.

Ví dụ:

- Thu nhập của toàn bộ dân cư của một nước.
- Chất lượng sản phẩm của một nhà máy.
- Nhu cầu tiêu dùng điện của các hộ gia đình.

Khi nghiên cứu tổng thể thì các phần tử có thể có hai loại tính chất định tính hoặc định lượng cần quan tâm, do đó ta có hai loại biến:

- Biến định lượng là các số đo của phần tử;

  Ví dụ: Cân nặng, chiều cao, tuổi, thu nhập,…
- Biến định tính là tính chất nào đó của đối tượng nghiên cứu.

  Ví dụ: Giới tính, chất lượng, dân tộc, tôn giáo,…

Đối với các biến ta có các cách mã hoá như sau:

- Kỹ thuật mã hoá

- Mã hoá biến định lượng: Ta lấy giá trị của biến định lượng làm mã của biến
- Mã hoá biến định tính: Ta gán tính chất định tính của biến ứng với các số nguyên.

Ví dụ:

Đối tượng là thu nhập của hộ gia đình ta có các mức: Nghèo, trung bình, giàu. Ta mã hoá các biến như sau:

Nghèo → -1; Trung bình → 0; Giàu → 1

Vậy khi nghiên cứu tổng thể ta luôn có thể giả sử là các các phần tử của tổng thể có dấu hiệu định lượng.

5.2.2. Mô tả tổng thể

Cho tổng thể với các phần tử $\{x_1, x_2, \ldots x_N\}$, ta có thể thu gọn bằng cách gộp các giá trị giống nhau lại và biểu diễn như dạng.

x_i	x_1	$x_2 \ldots\ldots\ldots\ldots\ldots x_k$
N_i	N_1	$N_2 \ldots\ldots\ldots\ldots N_k$

trong đó N_i $(i = 1,...,k)$ là số lần giá trị x_i xuất hiện trong tổng thể, ta có

$$N_1 + N_2 + \ldots + N_k = N$$

Đặt $f_1 = \frac{N_i}{N}$ $(i = 1,\ldots,k)$, f_i được gọi là tần suất của x_i trong tổng thể và ta có bảng tần suất.

x_i	x_1	$x_2 \ldots\ldots\ldots\ldots x_k$
f_i	f_2	$f_2 \ldots\ldots\ldots\ldots f_k$

Hiển nhiên ta có: $f_1 + f_2 + \ldots + f_k = 1$

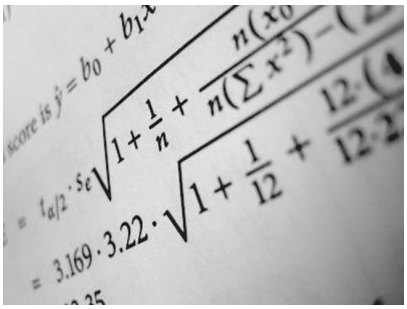
Bảng tần suất giống như một bảng phân phối xác suất của biến ngẫu nhiên, do đó ta có thể đồng nhất tổng thể nghiên cứu với một biến ngẫu nhiên X nào đó với hàm phân phối F. Vậy, thay vì nghiên cứu tổng thể thì ta quy về nghiên cứu biến ngẫu nhiên X.

5.2.3. Các số đặc trưng của tổng thể

- **Trung bình tổng thể**

Trung bình tổng thể là đại lượng ký hiệu là m được xác định bởi:

$$m = \frac{1}{N}\sum_{i=1}^{N} N_i x_i = \sum_{i=1}^{N} f_i x_i$$



Ta thấy m có thể xem là kỳ vọng của biến ngẫu nhiên X.

- **Phương sai tổng thể**

Phương sai tổng thể là đại lượng ký hiệu là s được xác định bởi:

$$s = \frac{1}{N}\sum_{i=1}^{N} (x_i - m)^2 = \sum_{i=1}^{N} f_i x_i^2 - (m)^2.$$

Ta thấy s có thể xem là phương sai của biến ngẫu nhiên X.

5.3. Mẫu ngẫu nhiên

Trong phần trước ta đã biết rằng không thể nghiên cứu cặn kẽ từng phần tử của tổng thể, do đó ta phải nghiên cứu hạn chế trên một nhóm nhỏ được rút ra từ tổng thể gọi là mẫu và từ đó rút ra kết luận cho tổng thể, do vậy ta mong muốn mẫu đại diện tốt nhất cho tổng thể. Nói chung, để có được một mẫu đại diện tốt nhất cho tổng thể người ta thường phải tiến hành xây dựng mẫu theo một quy trình chọn ngẫu nhiên các phần tử của mẫu. Một mẫu như vậy được gọi là *mẫu ngẫu nhiên* (random sample).

5.3.1. Các phương pháp lấy mẫu

Có rất nhiều phương pháp chọn mẫu ngẫu nhiên để thoả mãn tính đại diện tốt nhất cho tổng thể và phù hợp với mục tiêu nghiên cứu. Sau đây ta chỉ nghiên cứu những phương pháp chủ yếu.

- Cách chọn mẫu ngẫu nhiên đơn giản
  - Chọn mẫu ngẫu nhiên có hoàn lại: Từ tổng thể ta rút ngẫu nhiên một phần tử và ghi lại các đặc trưng cần quan tâm, sau đó trả lại phần tử đó về tổng thể và làm tương tự ở các lần tiếp theo cho tới khi ta được một mẫu cỡ n.
  - Chọn mẫu ngẫu nhiên không hoàn lại: Làm tương tự như trên, chỉ khác là sau mỗi lần rút các phần tử ta loại phần tử đó ra khỏi tổng thể.
- Chọn mẫu phân cấp

  Ở những tổng thể lớn có thể có những yêu cầu phải chọn một mẫu phân cấp chẳng hạn như điều tra phân tích mức sống của dân cư trong nước thường có những yêu cầu kết luận cho các vùng, các miền.
  - Mẫu phân cấp đơn giản có thể được thành lập như sau: Chia tổng thể ra thành k tổng thể bộ phận và ta thực hiện cách lấy mẫu ngẫu nhiên đơn giản trên mỗi tổng thể thành phần rồi tổng hợp lại để có mẫu của toàn bộ tổng thể.

Ta cũng có thể tiến hành lấy mẫu phân cấp theo những quy trình phức tạp hơn. Chẳng hạn như sau khi chia tổng thể ra thành k tổng thể bộ phận, ta chọn ngẫu nhiên trong số k tổng thể bộ phận đó ra m tổng thể rồi tiếp tục thực hiện lấy mẫu ngẫu nhiên trên từng tổng thể được chọn để tổng hợp thành mẫu của toàn bộ tổng thể.

5.3.2. Định nghĩa mẫu ngẫu nhiên

Một mẫu ngẫu nhiên cỡ n của biến ngẫu nhiên X là một bộ n các biến ngẫu nhiên $X_1, X_2, \ldots X_n$ độc lập và có cùng phân phối với biến ngẫu nhiên X, trong đó mỗi X_k là một *quan sát* về biến ngẫu nhiên X.

Ta ký hiệu x_k là kết quả quan sát được ở lần thứ k, tức là quan sát X_k nhận giá trị x_k (k = 1,2,…, n). Khi đó bộ giá trị $(x_1, x_2, \ldots, x_n)$ gọi là giá trị cụ thể của mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$.

Ví dụ 1:

Khi gieo con xúc xắc 5 lần ta được một mẫu ngẫu nhiên $(X_1, X_2, X_3, X_4, X_5)$ trong một lần lấy mẫu nào đó, chẳng hạn ta được giá trị của mẫu là (3, 5, 2, 3, 1).

Ví dụ 2:

Nghiên cứu thời gian hoạt động của các thiết bị điện tử do một công ty sản xuất, ta lấy ngẫu nhiên n thiết bị, khi đó ta được một mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$, theo dõi thời gian hoạt động của n thiết bị điện tử này ta được các giá trị mẫu là $(x_1, x_2, \ldots, x_n)$.

5.3.3. Mô tả mẫu ngẫu nhiên

Cho biến ngẫu nhiên X và một mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ với các giá trị mẫu $(x_1, x_2, \ldots, x_n)$. Để mô tả mẫu ngẫu nhiên ta có hai cách như sau:

- Biểu đồ tần suất

  Ta có thể thu gọn bằng cách gộp các giá trị giống nhau trong mẫu và biểu diễn dưới dạng bảng sau:

x_i	x_1	x_2	…	x_n
n_i	n_1	n_2	…	n_k

trong đó n_i là số lần giá trị x_i xuất hiện trong mẫu. Ta có:

$$n_1 + n_2 + \ldots + n_k = n.$$

Ví dụ:

Giá trị mẫu quan sát là (5; 1; 8; 5; 3; 8; 9; 7; 5; 1; 8; 3), cỡ mẫu n = 12, số liệu được thu gọn lại có dạng:

x_i	1	3	5	7	8	9
n_i	2	2	3	1	3	1

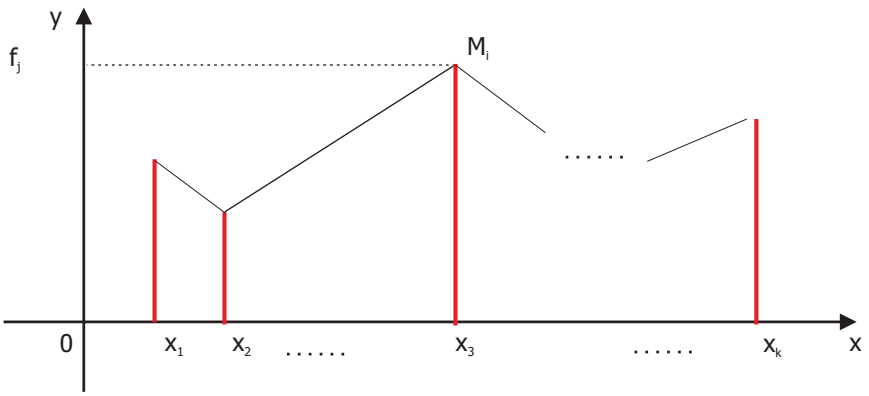
Đặt $f_i = \frac{n_i}{n}$ và gọi đó là tần suất của x_i trong mẫu, khi đó ta có bảng biểu diễn tần suất mẫu.

x_i	x_1	x_2	…	x_n
n_i	f_1	f_2	…	f_k

Ta có:

$$f_1 + f_2 + \ldots + f_k = (n_1 + n_2 + \ldots + n_k)/n = 1.$$

Trên trục tọa độ 0xy ta biểu diễn các điểm $M_i(x_i, f_i)$ và nối các điểm M_i với nhau ta được một biểu đồ tần suất trong Hình 1.



Hình 1: Trình bày mẫu bằng biểu đố tần suất

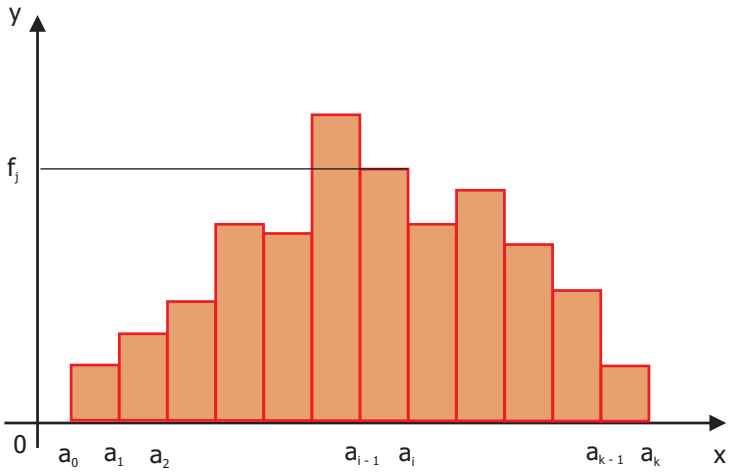
- Tổ chức đồ (biểu đồ tần số)

  Chia miền giá trị của mẫu thành k khoảng $(a_0; a_1]$, $(a_1; a_2]$, ... , $(a_{k-1}; a_k]$, ký hiệu n_i là số các giá trị mẫu rơi vào khoảng $(a_{i-1}; a_i]$, (i=1,2,..,k). Ta biểu diễn mẫu dưới dạng:

Khoảng	$[a_0 - a_1]$	$[a_1 - a_2]$	...	$[a_{k-1} - a_k]$
n_i	n_1	n_2	...	n_k

$$n_1 + n_2 + \ldots + n_k = n$$

n_i là số giá trị mẫu rơi vào khoảng $(a_{i-1}; a_i]$. Trong mặt phẳng Oxy, trên trục Ox biểu diễn các khoảng $(a_{i-1}; a_i]$, trên trục Oy biểu diễn các giá trị $y_i = n_i /(n.h_i)$, trong đó h_i là độ dài khoảng $(a_{i-1}; a_i]$, i =1,2,..k . Ta dựng các hình chữ nhật có chiều cao là y_i và độ dài đáy là h_i. Hình được tạo bởi các hình chữ nhật trên được gọi là tổ chức đồ (biểu đồ tần số).



Hình 2: Tổ chức đồ

5.4. Thống kê (Statistics)

Cho biến ngẫu nhiên X với mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ và giá trị mẫu $(x_1, x_2, \ldots, x_n)$.

5.4.1. Định nghĩa.

Thống kê là một hàm của các quan sát trong mẫu ngẫu nhiên, ký hiệu là $G(X_1, X_2, \ldots, X_n)$. Khi mẫu nhận giá trị cụ thể $(x_1, x_2, \ldots, x_n)$ thì thống kê G nhận giá trị g được xác định bởi

$$g = G(x_1, x_2, \ldots, x_n) .$$

Ví dụ:

- Thống kê:

$$\overline{X} = G(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

được gọi là *trung bình mẫu*. Giá trị cụ thể của $\overline{X}$ là:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

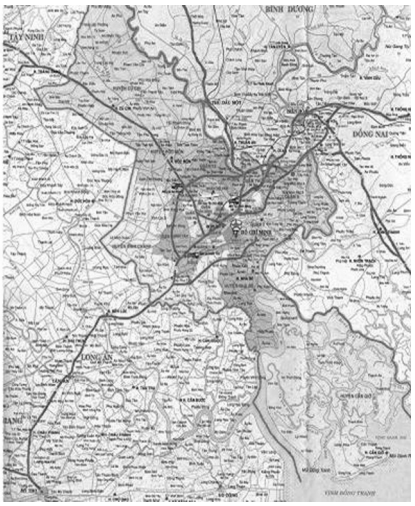
- Thống kê:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2 .$$

được gọi là *phương sai mẫu*. Giá trị cụ thể của S^2 :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 .$$

5.4.2. Các thống kê đặc trưng mẫu



Ngoài hai thống kê thường gặp là kỳ vọng và phương sai đã nêu trên đây, ta còn có nhiều thống kê đặc trưng của mẫu khác nữa. Có thể kể thêm một số thống kê khác dưới đây:

- Phương sai mẫu hiệu chỉnh

  Định nghĩa: Thống kê

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_1 - \overline{X}\right)^2 = \frac{n}{n-1} S^2$$

  được gọi là *phương sai mẫu hiệu chỉnh*.

- Độ lệch chuẩn mẫu và độ lệch chuẩn mẫu hiệu chỉnh

  Định nghĩa: Thống kê

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

  được gọi là *độ lệch chuẩn mẫu*.

Định nghĩa: Thống kê

$$S' = \sqrt{S'^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

được gọi là *độ lệch chuẩn mẫu hiệu chỉnh*.

- Cách tính các giá trị thống kê đặc trưng mẫu.

  Cho mẫu ngẫu nhiên thu gọn

x_i	x_1	x_2	...	x_n
n_i	n_1	n_2	...	n_k

Nếu mẫu cho dưới dạng khoảng, ta chọn mỗi khoảng điểm đại diện

$$x_i = \frac{a_{i-1} + a_i}{2}, \quad i = 1, 2, ..., k,$$

khi đó ta có mẫu thu gọn. Để thuận tiện trong việc tính toán các giá trị thống kê đặc trưng với mẫu cụ thể, ta lập một bảng tính như sau:

Khoảng giá trị mẫu	x_i	n_i	$n_i.x_i$	$n_i.x_i^2$
$a_0 - a_1$	x_1	n_1	n_1x_1	$n_1x_1^2$
$a_1 - a_2$	x_2	n_2	n_2x_2	$n_2x_2^2$
⋮	⋮	⋮	⋮	⋮
$a_{i-1} - a_i$	x_i	n_i	n_ix_i	$n_ix_i^2$
⋮	⋮	⋮	⋮	⋮
$a_{k-1} - a_k$	x_k	n_k	n_kx_k	$n_kx_k^2$
$\sum$		n	A	B

Ta có:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{A}{n},$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - (\overline{x})^2 = \frac{B}{n} - (\overline{x})^2,$$

$$s'^2 = \frac{n}{n-1}s^2,$$

$$s = \sqrt{s^2} \text{ và } s' = \sqrt{s'^2}.$$

Ví dụ:

Điều tra mức thu nhập cá nhân trong một tháng (triệu đồng), ta có bảng số liệu mẫu sau:

Thu nhập	1-2	2-3	3-4	4-5	5-6	6-7
số người	10	8	5	7	3	2

Tính các giá trị đặc trưng mẫu: $\bar{x}$, s^2, s'^2, s, s'

Ta lập bảng tính:

Khoảng thu nhập	x_i	n_i	$n_i.x_i$	$n_i.x_i^2$
1-2	1,5	10	15	22,5
2-3	2,5	8	20	50
3-4	3,5	5	17,5	61,25
4-5	4,5	7	31,5	141,75
5-6	5,5	3	16,5	90,75
6-7	6,5	2	13	84,5
$\sum$		n = 35	113,5	450,75

Từ đó,

$$\bar{x} = 113,5/35 = 3,243.$$

$$s^2 = 450,75/35 - (3,243)^2 = 2,363.$$

$$s = \sqrt{2,363} = 1,537$$

$$s'^2 = \frac{n}{n-1}s^2 = \frac{35}{34}2,363 = 2,43$$

$$s' = \sqrt{2,43} = 1,559.$$

5.5. Mẫu ngẫu nhiên hai chiều

Trong phần trước ta đã xét tổng thể với một dấu hiệu định tính hoặc định lượng và ta đã đồng nhất tổng thể nghiên cứu như là một biến ngẫu nhiên X nào đó. Trong phần này ta mở rộng xét tổng thể nghiên cứu với hai dấu hiệu định tính hoặc

định lượng. Ví dụ khi xét tới tổng thể nghiên cứu trong xã hội thì ta xét tới dấu hiệu chiều cao và dấu hiệu cân nặng. Cả hai dấu hiệu này cùng xuất hiện trên mỗi phần tử của tổng thể nghiên cứu. Tương tự như phần trước ta cũng sẽ đồng nhất tổng thể nghiên cứu với biến ngẫu nhiên hai chiều (X, Y).

5.5.1. Khái niệm

Một mẫu ngẫu nhiên cỡ n của véc tơ ngẫu nhiên (X, Y) là một tập các véc tơ ngẫu nhiên $(X_1,Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ độc lập và có cùng phân phối với biến ngẫu nhiên (X, Y), trong đó véc tơ (X_i, Y_i) là quan sát thứ i về véc tơ ngẫu nhiên (X, Y).

Ký hiệu (x_i, y_i) là giá trị của mẫu (X_i, Y_i) $(i = 1,2,..,n)$. Khi đó bộ giá trị $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ được gọi là giá trị cụ thể của mẫu ngẫu nhiên $(X_1,Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

Ví dụ 1:

Lấy mẫu điều tra thu nhập và tiêu dùng (triệu đồng/tháng) của 10 hộ gia đình ta thu được giá trị mẫu: (2; 1,4), (2; 1,5), (3; 1,8), (4; 1,8), (2; 1,5), (4; 3,5), (7; 5,5), (3; 1,4), (4; 3,5), (5; 3,7).

5.5.2. Phương pháp mô tả mẫu

Cho mẫu ngẫu nhiên hai chiều với các giá trị mẫu là $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. Khi đó ta có thể biểu diễn mẫu dưới hai dạng như sau:

Dạng 1: Lập một bảng hai dòng theo dạng:

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n

Dạng 2: Thu gọn mẫu và biểu diễn dưới dạng bảng chữ nhật:

y_i / x_i	y_1	y_2	... y_j ...	y_h	a
x_1	n_{11}	n_{12}	... n_{1j} ...	n_{1h}	a_1
x_2	n_{21}	n_{22}	... n_{2j} ...	n_{2h}	a_2
⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1}	n_{i2}	... n_{ij} ...	n_{ih}	a_i
⋮	⋮	⋮	⋮	⋮	⋮
x_k	n_{k1}	n_{k2}	... n_{kj} ...	n_{kh}	a_k
b	b_1	b_2	... b_j ...	b_h	$\sum\sum = n$

trong đó n_{ij} là số lần xuất hiện cặp (x_i, y_j) trong mẫu, a_i là số lần xuất hiện x_i trong mẫu, b_j là số lần xuất hiện y_j trong mẫu. Ta có:

$$\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} = n\,,\ a_i = \sum_{j=1}^{h} n_{ij}\ ,\ b_j = \sum_{i=\ell}^{k} n_{ij}$$

Ví dụ 2: Ta xét ví dụ 1. Mẫu có thể thu gọn và biểu diễn dưới dạng:

y_j / x_i	1,4	1,5	1,8	3,5	3,7	5,5	*a*
2	1	2	0	0	0	0	3
3	1	0	1	0	0	0	2
4	0	0	1	2	0	0	3
5	0	0	0	0	1	0	1
7	0	0	0	0	0	1	1
b	2	2	2	2	1	1	$\sum\sum = 10$

CHÚ Ý

Ta cũng có thể phân khoảng giá trị mẫu đối với từng thành phần của mẫu ngẫu nhiên hai chiều. Khi đó các thành phần được xử lý tương tự như đối với mẫu ngẫu nhiên một chiều.

5.5.3. Thống kê đặc trưng mẫu hai chiều

- Trung bình mẫu

  Định nghĩa:

  Véc tơ ngẫu nhiên hai chiều $(\overline{X}, \overline{Y})$ gọi là trung bình mẫu của véc tơ ngẫu nhiên (X, Y), trong đó $\overline{X}$ và $\overline{Y}$ là các trung bình mẫu của biến ngẫu nhiên thành phần X và Y.

  Giá trị thống kê mẫu của mẫu ngẫu nhiên hai chiều là $(\overline{x}, \overline{y})$.

- Hệ số tương quan mẫu

  Định nghĩa:

  Hệ số tương quan mẫu của mẫu ngẫu nhiên hai chiều ký hiệu là R được xác định bởi:

$$R = \frac{\overline{XY} - (\overline{X})(\overline{Y})}{S_X S_Y}$$

  trong đó

$$\overline{XY} = \frac{1}{n}\sum_{k=1}^{n} X_k Y_k$$

Giá trị của hệ số tương quan mẫu đối với mẫu cụ thể $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ là: $r = \dfrac{\overline{xy} - (\overline{x})(\overline{y})}{s_X s_Y}$,

với $$\overline{xy} = \frac{1}{n}\sum_{l=1}^{n} x_l y_l = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} x_i y_j$$

$$\overline{x} = \frac{1}{n}\sum_{l=1}^{n} x_l = \frac{1}{n}\sum_{i=1}^{k} a_i x_i \quad , \quad \overline{y} = \frac{1}{n}\sum_{l=1}^{n} y_l = \frac{1}{n}\sum_{j=1}^{h} b_j y_j$$

$$s_X = \sqrt{\frac{1}{n}\sum_{l=1}^{n} x_l^2 - (\overline{x})} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} a_i x_i^2 - (\overline{x})} \quad ,$$

$$s_Y = \sqrt{\frac{1}{n}\sum_{l=1}^{n} y_l^2 - (\overline{y})} = \sqrt{\frac{1}{n}\sum_{j=1}^{h} b_j y_j^2 - (\overline{y})} \, .$$

Nếu mẫu biểu diễn dưới dạng 1 ta sử dụng dấu bằng thứ nhất. Nếu mẫu biểu diễn dưới dạng 2 ta sử dụng dấu bằng thứ hai trong công thức trên.

5.6. Quy luật phân phối xác suất của một số thống kê

Trong mục này ta sẽ xác định quy luật phân phối xác suất của một số thống kê mẫu. Phân phối mẫu của một thống kê phụ thuộc vào phân phối của biến ngẫu nhiên gốc, cỡ của mẫu và phương pháp lựa chọn mẫu. Phần này giới thiệu phân phối mẫu của một số thống kê quan trong có nhiều ứng dụng trong các bài tiếp theo.

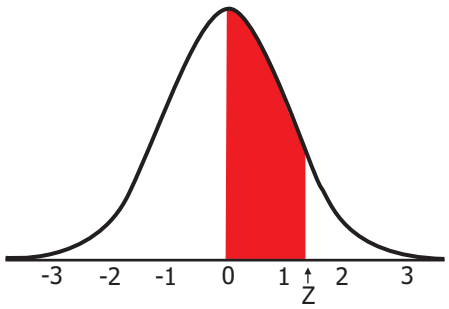
Định nghĩa:

Phân phối xác suất của một thống kê được gọi là phân phối mẫu.

Ví dụ: Phân phối xác suất của $\overline{X}$ được gọi là phân phối mẫu của thống kê trung bình mẫu.

5.6.1. Trường hợp biến ngẫu nhiên gốc có phân phối 0–1

Cho biến ngẫu nhiên X có quy luật phân bố 0–1 với tham số p. Xét mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ rút ra từ X. Dựa trên cơ sở lý thuyết xác suất (Mục §2, bài 3), ta có ngay định lý sau:



Định lý 0: Thống kê $n.\overline{X}$ có quy luật phân phối nhị thức $B(n, p)$.

Với định lý trên, khi cỡ mẫu n đủ nhỏ, ta dễ dàng tính toán để xác định được phân phối xác suất của kỳ vọng mẫu $\overline{X}$ cho trường hợp biến ngẫu nhiên

gốc có phân phối 0–1. Tuy nhiên, tính toán này không phải đơn giản nếu cỡ mẫu lớn. Trong các trường hợp như vậy, ta có thể dựa vào các định lý giới hạn để tính xấp xỉ các phân phối xác suất của thống kê cần quan tâm. Cụ thể, từ Định lý Moivre-Laplace (xem Bài 4) ta có:

Định lý 1:

Thống kê: $U = \dfrac{\overline{X} - p}{\sqrt{p(p-1)}}\sqrt{n}$ *có quy luật phân phối xấp xỉ phân phối chuẩn tắc* $N(0,1)$ *khi n đủ lớn.*

CHÚ Ý

Thống kê U cũng có thể viết lại dưới dạng

$$U = \frac{f - p}{\sqrt{p(p-1)}}\sqrt{n}$$

trong đó $f = k/n$, với k là số lần mẫu nhận giá trị 1. Nếu ta có một biến cố A với xác suất p thì n là số lần thực hiện phép thử, k là số lần A xuất hiện và f là tần số xuất hiện biến cố A.

5.6.2. Trường hợp hai biến ngẫu nhiên gốc có phân phối 0–1

Cho hai biến ngẫu nhiên độc lập X và Y có cùng phân phối 0–1 với hai tham số tương ứng là p_1 và p_2. Xét hai mẫu ngẫu nhiên $(X_1, X_2,\ldots,X_n)$ và $(Y_1, Y_2,\ldots,Y_m)$ rút ra từ X và Y

Định lý 2:

Thống kê: $U = \dfrac{\overline{X} - \overline{Y} - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n} + \dfrac{p_2(1-p_2)}{m}}}$ *có quy luật xấp xỉ phân phối chuẩn* $N(0,1)$ *khi n đủ lớn*

CHÚ Ý

Thống kê U có thể viết dưới dạng:

$$U = \frac{f_1 - f_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n} + \dfrac{p_2(1-p_2)}{m}}}$$

trong đó $f_1 = k_1/n,\ f_2 = k_2/m,$ với k_1 là số lần mẫu ngẫu nhiên của X nhận giá trị 1, k_2 là số lần mẫu ngẫu nhiên của Y nhận giá trị 1. Nếu ta có hai biến cố A và B thì k_1 và k_2 là số lần biến cố A và B xuất hiện trong n phép thử về biến cố A và m phép thử về biến cố *B*, f_1, f_2 là các tần suất tương ứng của hai biến cố.

Nếu ta có hai biến cố A và B thì k_1 và k_2 là số lần biến cố A và B xuất hiện trong n phép thử về biến cố A và m phép thử về biến cố B, f_1, f_2 là các tần suất tương ứng của hai biến cố.

5.6.3. Trường hợp biến ngẫu nhiên gốc có phân phối chuẩn

Cho mẫu ngẫu nhiên $(X_1, X_2, \ldots X_n)$ được rút ra từ biến ngẫu nhiên X có quy luật phân phối chuẩn $N(\mu, \sigma^2)$. Từ tính chất của phân phối chuẩn, ta có:

Định lý 3:

Thống kê trung bình mẫu $\overline{X}$ *có phân phối chuẩn* $N\left(\mu_{\overline{X}}, \sigma^2_{\overline{X}}\right)$, *trong đó:*

$$\mu_{\overline{X}} = \frac{\mu+\mu+\ldots+\mu}{n} = \mu \quad ; \quad \sigma^2_{\overline{X}} = \frac{\sigma^2+\sigma^2+\ldots+\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Từ đó ta thấy thống kê $U = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ *có quy luật phân phối chuẩn* $N(0,1)$.

CHÚ Ý

Nếu mẫu ngẫu nhiên được rút ra từ biến ngẫu nhiên X chưa biết dạng phân phối xác suất thì theo Định lý giới hạn trung tâm, phân phối của trung bình mẫu cũng xấp xỉ phân phối chuẩn với trung bình $\mu_{\overline{X}}$ và phương sai $\sigma^2_{\overline{X}}$ được xác định như trên khi mà cỡ mẫu n đủ lớn.

Định lý 4:

- *Thống kê* $T = \frac{\overline{X}-\mu}{S'}\sqrt{n}$ *có quy luật phân bố Student với* n–1 *bậc tự do.*
- *Thống kê* $\chi^2 = \frac{(n-1)S'^2}{\sigma^2}$ *có quy luật phân phối khi–bình phương với n–1 bậc tự do.*

5.6.4. Trường hợp hai biến ngẫu nhiên gốc có phân phối chuẩn

Cho hai biến ngẫu nhiên X có phân phối chuẩn $N(\mu_1, \sigma_1^2)$, Y có phân phối chuẩn $N(\mu_2, \sigma_2^2)$, X độc lập với Y. Xét mẫu ngẫu nhiên $(X_1, X_2, \ldots X_n)$ rút ra từ X và mẫu ngẫu nhiên $(Y_1, Y_2, \ldots Y_m)$ rút ra từ Y.

Định lý 5: *Thống kê* $U = \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}$ *có quy luật phân phối chuẩn* $N(0,1)$.

Định lý 6: *Thống kê* $T = \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{nS_X^2+mS_Y^2}}\sqrt{\frac{nm(n+m-2)}{n+m}}$ *có quy luật phân phối Student với* n + m – 2 *bậc tự do. Trong đó* S_X^2 *là phương sai mẫu của biến ngẫu nhiên* X, S_Y^2 *là phương sai mẫu của biến ngẫu nhiên* Y.

Các thống kê được đề cập đến trên đây sẽ được sử dụng trong phần ước lượng và kiểm định giả thuyết trong các bài tiếp theo.

TÓM LƯỢC CUỐI BÀI

Để học tập bài 5 các bạn cần nắm vững các vấn đề sau: Khái niệm phương pháp mẫu, các thống kê đặc trưng mẫu và cách tính các giá trị thống kê đặc trưng mẫu, mẫu ngẫu nhiên hai chiều và cách tính các giá trị thống kê mẫu ngẫu nhiên hai chiều. Quy luật phân phối xác suất của một số thống kê. Đặc biệt các bạn cần phải nắm vững về các thống kê đặc trưng mẫu và cách tính giá trị của các thống kê này với mẫu cụ thể, ngoài ra cần nắm vững các quy luật phân phối xác suất của một số thống kê mẫu.

BÀI TẬP

1. Theo dõi thời gian hoàn thành sản phẩm của 50 công nhân ta có bảng số liệu sau

Thời gian	12-14	14-16	16-18	18-20	20-22	22-24	24-26	26-28
Số công nhân	4	10	1	12	14	2	6	1

Xác định trung bình mẫu, phương sai, phương sai hiệu chỉnh, độ lệch chuẩn và độ lệch chuẩn hiệu chỉnh của mẫu.

2. Trong một trại chăn nuôi lợn khi thử nghiệm một loại thức ăn mới, sau ba tháng người ta cân thử một số con lợn và thu được số liệu sau:

Trọng lượng(kg)	65	67	68	69	70	71	73
Số con	1	4	3	6	7	2	2

a. Tìm tần suất của những con lợn có trọng lượng ít hơn 68 kg.

b. Hãy tính trung bình mẫu, phương sai, phương sai mẫu hiệu chỉnh.

3. Điều tra sản lượng sản xuất thép hàng tháng của một công ty thép (đơn vị: nghìn tấn) ta có bảng số liệu sau:

7,0 6,9 7,8 7,7 7,3 6,8 6,7 8,2 8,4 7,0 6,7 7,5 7,2 7,9 6,7 7,8

7,5 6,6 7,8 7,5 7,2 7,6 .

a. Hãy thu gọn số liệu mẫu trên.

b. Hãy tính tần suất của những tháng mà có số sản lượng thép lớn hơn 7,5 tấn.

c. Hãy tính trung bình, phương sai, phương sai mẫu hiệu chỉnh.

4. Điều tra mức thu nhập X hàng năm (100$/năm) và số tiền Y chi cho các nhu cầu xa xỉ phẩm ($/tháng) ta được số liệu mẫu:

X	23	17	34	56	49	31	26	80	65	40	26
Y	10	50	120	225	90	60	55	340	170	25	80

Hãy tính hệ số tương quan mẫu giữa X và Y từ đó có kết luận về mối quan hệ giữ X và Y?

5. Ký hiệu f_1 là tần suất của những người có mức thu nhập trên 3400$ trong bài tập 4. Hãy tính giá trị của thống kê:

$$U = \frac{(f_1 - 0,6)}{\sqrt{f_1(1-f_1)}}\sqrt{n}\ .$$

6. Ký hiệu f_2 là tần suất của những người có mức chi dùng cho nhu cầu xa xỉ dưới 60$ trong

bài tập 4. Hãy tính giá trị của thống kê:

$$U = \frac{(f_1 - 0,2)}{\sqrt{f_2(1-f_2)}}\sqrt{n}\,.$$

7. Cho biến cố A, gọi X là biến ngẫu nhiên nhận giá trị 1 nếu biến cố A xuất hiện, X nhận giá trị 0 nếu biến cố X không xuất hiện. Lấy một mẫu ngẫu nhiên cỡ 15 ta thu được giá trị mẫu: (0; 0; 1; 0; 1; 1; 1; 0; 0 ;1; 1; 0; 0; 0; 1).

a. Hãy thu gọn số liệu mẫu? số lần thực hiện phép thử về biến cố A là bao nhiêu?

b. Tính tần suất xuất hiện biến cố A?

8. Điều tra thu nhập của một nhóm công nhân, ta thu được số liệu sau

Thu nhập (triệu)	15-17	17-19	19-21	21-23	23-25	25-27
số công nhân	1	3	4	12	3	2

Hãy tính tần suất của những công nhân trong mẫu điều tra có mức thu nhập trong khoảng 19 đến 25 triệu.