CHƯƠNG V: LÝ THUYẾT MẪU

Thống kê toán là bộ môn toán học nghiên cứu qui luật của các hiện tượng ngẫu nhiên có tính chất số lớn trên cơ sở thu thập và xử lý số liệu thống kê các kết quả quan sát về những hiện tượng ngẫu nhiên này. Nếu ta thu thập được các số liệu liên quan đến tất cả đối tượng cần nghiên cứu thì ta có thể biết được đối tượng này (phương pháp toàn bộ). Tuy nhiên trong thực tế điều đó không thể thực hiện được vì quy mô của các đối tượng cần nghiên cứu quá lớn hoặc trong quá trình nghiên cứu đối tượng nghiên cứu bị phá hủy. Vì vậy cần lấy mẫu để nghiên cứu.

Chương này giới thiệu về phương pháp lấy mẫu ngẫu nhiên và các thống kê thường gặp của mẫu ngẫu nhiên.

5.1 MẪU NGẪU NHIÊN

5.1.1 Sự cần thiết phải lấy mẫu

Nhiều bài toán trong thực tế dẫn đến nghiên cứu một hay nhiều dấu hiệu định tính hoặc định lượng đặc trưng cho các phần tử của một tập hợp nào đó. Chẳng hạn nếu muốn điều tra thu nhập bình quân của các gia đình ở Hà nội thì tập hợp cần nghiên cứu là các hộ gia đình ở Hà nội, dấu hiệu nghiên cứu là thu nhập của từng gia đình (dấu hiệu định lượng). Một doanh nghiệp muốn nghiên cứu các khách hàng của mình về dấu hiệu định tính có thể là mức độ hài lòng của khách hàng đối với sản phẩm hoặc dịch vụ của doanh nghiệp, còn dấu hiệu định lượng là số lượng sản phẩm của doanh nghiệp mà khách hàng có nhu cầu được đáp ứng. Khi khảo sát một tín hiệu là quá trình ngẫu nhiên người ta tiến hành lấy mẫu tại những thời điểm nào đó và thu được các tín hiệu mẫu.

Để xử lý dấu hiệu cần nghiên cứu đôi khi người ta sử dung phương pháp nghiên cứu toàn bộ, đó là điều tra toàn bộ các phần tử của tập hợp theo dấu hiệu cần nghiên cứu để rút ra các kết luận cần thiết. Tuy nhiên trong thực tế việc áp dụng phương pháp này gặp phải những khó khăn sau:

- Do qui mô của tập hợp cần nghiên cứu quá lớn nên việc nghiên cứu toàn bộ sẽ đòi hỏi nhiều chi phí về vật chất và thời gian, có thể không kiểm soát được dẫn đến bị chồng chéo hoặc bỏ sót.
- Trong nhiều trường hợp không thể nắm được toàn bộ các phần tử của tập hợp cần nghiên cứu, do đó không thể tiến hành toàn bộ được.
- Có thể trong quá trình điều tra sẽ phá hủy đối tượng nghiên cứu …

Vì thế trong thực tế phương pháp nghiên cứu toàn bộ thường chỉ áp dụng đối với các tập hợp có qui mô nhỏ, còn chủ yếu người ta sử dụng phương pháp không toàn bộ mà đặc biệt là phương pháp nghiên cứu chọn mẫu.

5.1.2 Tổng thể nghiên cứu, dấu hiệu nghiên cứu

Toàn bộ tập hợp các phần tử đồng nhất theo một dấu hiệu nghiên cứu định tính hay định lượng nào đó được gọi là tổng thể, ký hiệu $\complement$.

Số lượng các phần tử của tổng thể được gọi là *kích thước của tổng thể*, ký hiệu N. Thường thì kích thước N của tổng thể là hữu hạn, song nếu tổng thể quá lớn hoặc không thể nắm được toàn bộ tổng thể ta có thể giả thiết rằng kích thước của tổng thể là vô hạn.

Mỗi phân tử của tổng thể được gọi là *cá thể*.

Các cá thể của tổng thể được nghiên cứu thông qua các dấu hiệu nghiên cứu. Dấu hiệu nghiên cứu này có thể được định tính hoặc định lượng. Nếu dấu hiệu nghiên cứu có tính định lượng, nghĩa là được thể hiện bằng cách cho tương ứng mỗi cá thể của tổng thể $\complement$ nhận một giá trị thực nào đó thì dấu hiệu này được gọi là một *biến lượng*, ký hiệu X. Bằng cách mô hình hóa ta có thể xem biến lượng X là một biến ngẫu nhiên xác định trên tổng thể $\complement$. Trường hợp dấu hiệu định tính ta chỉ xét các dấu hiệu có thể mã hóa thành biến ngẫu nhiên chỉ nhận hai giá trị 0 và 1, như vậy dấu hiệu định tính X có thể xem là biến ngẫu nhiên có phân bố Bernoulli.

Việc chọn ra từ tổng thể một tập con với n cá thể nào đó gọi là *phép lấy mẫu*. Tập hợp con này được gọi là *một mẫu*, n là kích thước mẫu.

5.1.3 Mô hình hóa mẫu ngẫu nhiên

Ta nói rằng một mẫu là *mẫu ngẫu nhiên* nếu trong phép lấy mẫu đó mỗi cá thể của tổng thể được chọn một cách độc lập và có xác suất được chọn như nhau.

Giả sử các cá thể của tổng thể được nghiên cứu thông qua dấu hiệu X. Với mỗi mẫu ta chỉ cần quan tâm dấu hiệu nghiên cứu X của mỗi cá thể của mẫu.

Chẳng hạn, khi cần nghiên cứu chiều cao trung bình của thanh niên trong một vùng nào đó thì với cá thể A được chọn làm mẫu ta chỉ quan tâm về chiều cao của A, tức là dấu hiệu chiều cao X_A, mà không quan tâm đến các đặc trưng khác của cá thể này.

Vì vậy, mỗi cá thể được chọn khi lấy mẫu có thể đồng nhất với dấu hiệu nghiên cứu X của cá thể đó. Bằng cách đồng nhất mẫu ngẫu nhiên với các dấu hiệu nghiên cứu của mẫu ta có định nghĩa về mẫu ngẫu nhiên như sau:

Mẫu ngẫu nhiên kích thước n là một dãy gồm n biến ngẫu nhiên: $X_1, X_2, ..., X_n$ độc lập cùng phân bố với X, ký hiệu $W = (X_1, X_2, ..., X_n)$, trong đó X_i là dấu hiệu X của phần tử thứ i của mẫu ($i = 1, ..., n$).

Thực hiện một phép thử đối với mẫu ngẫu nhiên W chính là thực hiện một phép thử đối với mỗi thành phần của mẫu. Giả sử X_i nhận giá trị x_i ($i = 1, ..., n$), khi đó các giá trị $x_1, x_2, ..., x_n$ tạo thành một *giá trị cụ thể* của mẫu ngẫu nhiên, hay còn gọi là một thể hiện của mẫu ngẫu nhiên, ký hiệu $w = (x_1, x_2, ..., x_n)$.

Ví dụ 5.1: Gọi X là số chấm của mặt xuất hiện khi tung con xúc xắc cân đối, X là biến ngẫu nhiên có bảng phân bố xác suất sau

X	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

Giả sử tung con xúc xắc 3 lần, gọi X_i là số chấm xuất hiện trong lần tung thứ i $(i=1,2,3)$ thì ta có 3 biến ngẫu nhiên độc lập có cùng quy luật phân bố xác suất với X. Vậy ta có mẫu ngẫu nhiên kích thước 3, $W=(X_1,X_2,X_3)$.

Thực hiện một phép thử đối với mẫu ngẫu nhiên này tức là tung con xúc xắc 3 lần. Giả sử lần thứ nhất được 2 chấm, lần thứ hai được 5 chấm, lần ba được 3 chấm thì $w=(2,5,3)$ là một mẫu cụ thể của mẫu ngẫu nhiên W.

5.2 CÁC PHƯƠNG PHÁP MÔ TẢ GIÁ TRỊ CỦA MẪU NGẪU NHIÊN

5.2.1 Bảng phân bố tần số thực nghiệm

Từ một mẫu cụ thể của mẫu ngẫu nhiên kích thước n của X ta sắp xếp các giá trị của mẫu cụ thể theo thứ tự tăng dần, giả sử giá trị x_i xuất hiện với tần số r_i, $i=1,...,k$

$$x_1<\cdots<x_k;\ r_1+\cdots+r_k=n. \tag{5.1}$$

Khi đó ta có thể biểu diễn mẫu ngẫu nhiên trên qua bảng phân bố tần số thực nghiệm

X	x_1	x_2	…	x_k
Tần số	r_1	r_2	…	r_k

(5.2)

5.2.2 Bảng phân bố tần suất thực nghiệm

Ký hiệu $f_i=\dfrac{r_i}{n}$ và gọi là tần suất của x_i.

Ta có bảng phân bố tần suất thực nghiệm tương ứng của X

X	x_1	x_2	…	x_k
Tần suất	f_1	f_2	…	f_k

(5.3)

Ví dụ 5.2: Lấy một mẫu ngẫu nhiên kích thước 120 ta có bảng phân bố tần số thực nghiệm

X	31	34	35	36	38	40	42	44	$\sum$
Tần số	10	20	30	15	10	10	5	20	120

Bảng phân bố tần suất thực nghiệm tương ứng

X	31	34	35	36	38	40	42	44	$\sum$
Tần suất	2/24	4/24	6/24	3/24	2/24	2/24	1/24	4/24	1

5.2.3 Hàm phân bố thực nghiệm của mẫu

Tương tự công thức xác định hàm phân bố của biến ngẫu nhiên rời rạc (2.12), với mẫu cụ thể của mẫu ngẫu nhiên xác định bởi công thúc (5.1) ta có *hàm phân bố thực nghiệm của mẫu* xác định như sau

$$F_n(x) = \sum_{x_j \le x} f_j \; ; \quad -\infty < x < +\infty \qquad \textbf{(5.4)}$$

Định lý Glivenco chỉ ra rằng hàm phân bố thực nghiệm $F_n(x)$ xấp xỉ với phân bố lý thuyết $F_X(x) = P\{X \le x\}$ khi n đủ lớn.

5.2.4 Bảng phân bố ghép lớp

Trong những trường hợp mẫu điều tra có kích thước lớn, hoặc khi các giá trị cụ thể của dấu hiệu X lấy giá trị khác nhau song lại khá gần nhau, người ta thường xác định một số các khoảng $C_1, C_2, ..., C_k$ sao cho mỗi giá trị của dấu hiệu điều tra thuộc vào một khoảng nào đó. Các khoảng này lập thành một phân hoạch của miền giá trị của X.

Việc chọn số khoảng và độ rộng khoảng là tuỳ thuộc vào kinh nghiệm của người nghiên cứu, nhưng nói chung không nên chia quá ít khoảng. Ngoài ra độ rộng các khoảng cũng không nhất thiết phải bằng nhau. Chẳng hạn khi muốn thống kê về tỉ lệ người nghiện thuốc lá thì ta tập trung nhiều vào độ tuổi thanh niên và trung niên.

Ví dụ 5.3: Một mẫu về chiều cao (cm) của 400 cây con được trình bày trong bảng phân bố ghép lớp sau:

Khoảng	Tần số r_i	Tần suất f_i	Độ rộng khoảng l_i	$y_i = r_i / l_i$
$4,5-9,5$	18	0,045	5	3,6
$9,5-11,5$	58	0,145	2	29
$11,5-13,5$	62	0,155	2	31
$13,5-16,5$	72	0,180	3	24
$16,5-19,5$	57	0,1425	3	19
$19,5-22,5$	42	0,105	3	14
$22,5-26,5$	36	0,090	4	9
$26,5-36,5$	55	0,1375	10	5,5

Giá trị $y_i = \dfrac{r_i}{l_i}$ là tần số xuất hiện trong một đơn vị khoảng của khoảng có độ dài l_i.

Nhận xét 5.1:

1) Người ta quy ước đầu mút bên phải của mỗi khoảng thuộc vào khoảng đó mà không thuộc khoảng tiếp theo khi tính tần số của mỗi khoảng.

Trong ví dụ trên ta có các khoảng $[4{,}5;9{,}5]$, $(9{,}5;11{,}5]$, $(11{,}5;13{,}5]$,...

2) Một trong những gợi ý để chọn số khoảng k tối ưu là hãy chọn k nguyên nhỏ nhất sao cho $2^k \geq n$ như sau:

n : kích thước mẫu	33 –64	65 –127	129 –256	257 –512	513 –1024
k : số khoảng	6	7	8	9	10

5.2.5 Biểu diễn bằng biểu đồ

Giả sử dấu hiệu điều tra X có bảng phân bố tần số và tần suất thực nghiệm

X	x_1	x_2	…	x_k
Tần số	r_1	r_2	…	r_k

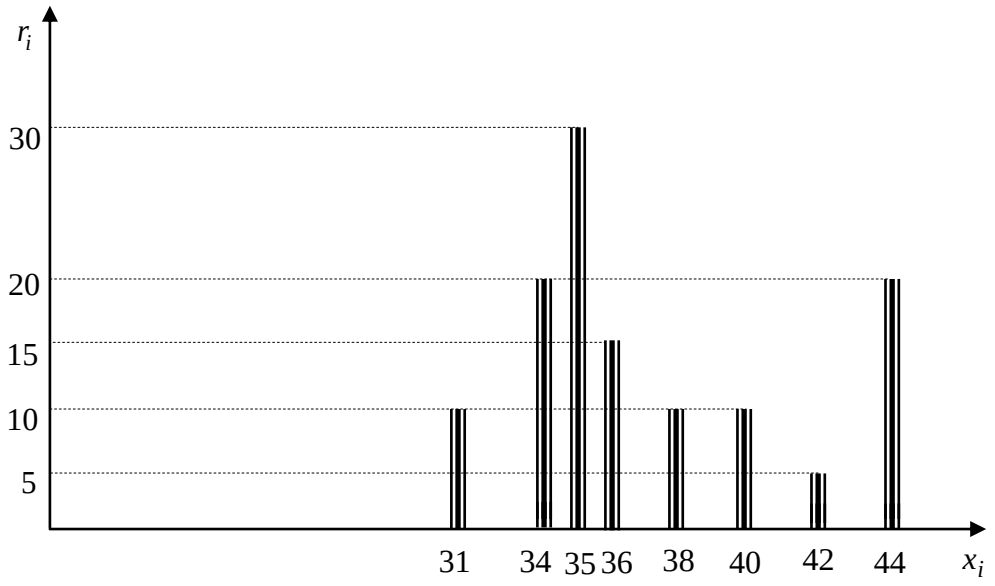
X	x_1	x_2	…	x_k
Tần suất	f_1	f_2	…	f_k

Trong mặt phẳng với hệ trục toạ độ Oxy.

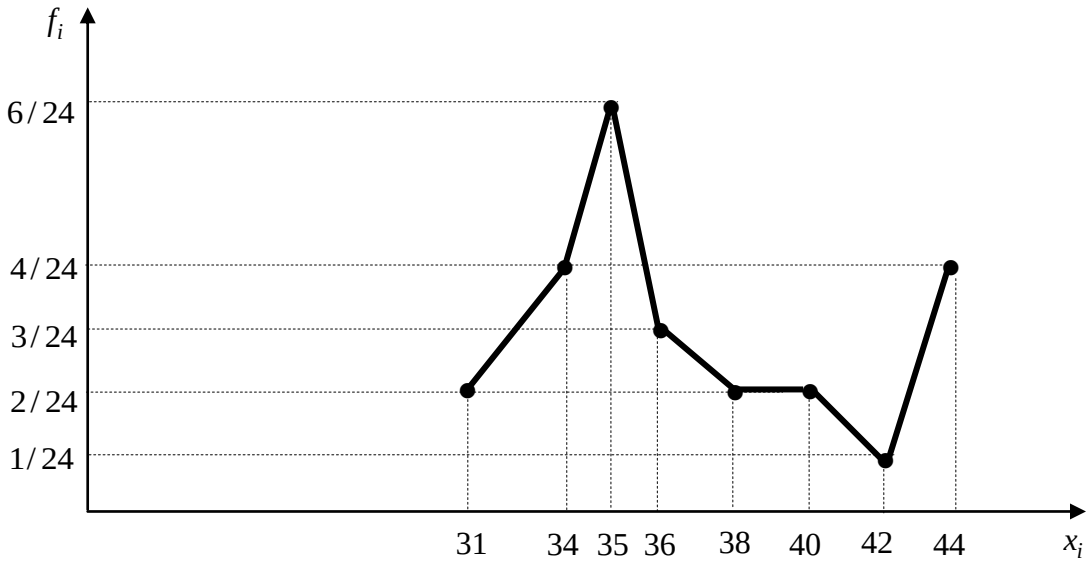
❖ Nối điểm trên trục hoành có toạ độ $(x_i,0)$ với điểm có toạ độ (x_i,r_i); $i=1,...,k$ ta được *biểu đồ tần số hình gậy*.

❖ Nối lần lượt điểm có toạ độ $(x_i,\ f_i)$ với điểm có toạ độ $(x_{i+1},\ f_{i+1})$; $i=1,...,k-1$ ta được *biểu đồ đa giác tần suất*.

Bảng phân bố tần số và tần suất thực nghiệm trong ví dụ 5.2 có biểu đồ tần số hình gậy và đa giác tần suất



Hình 5.1: *Biểu đồ tần số hình gậy*

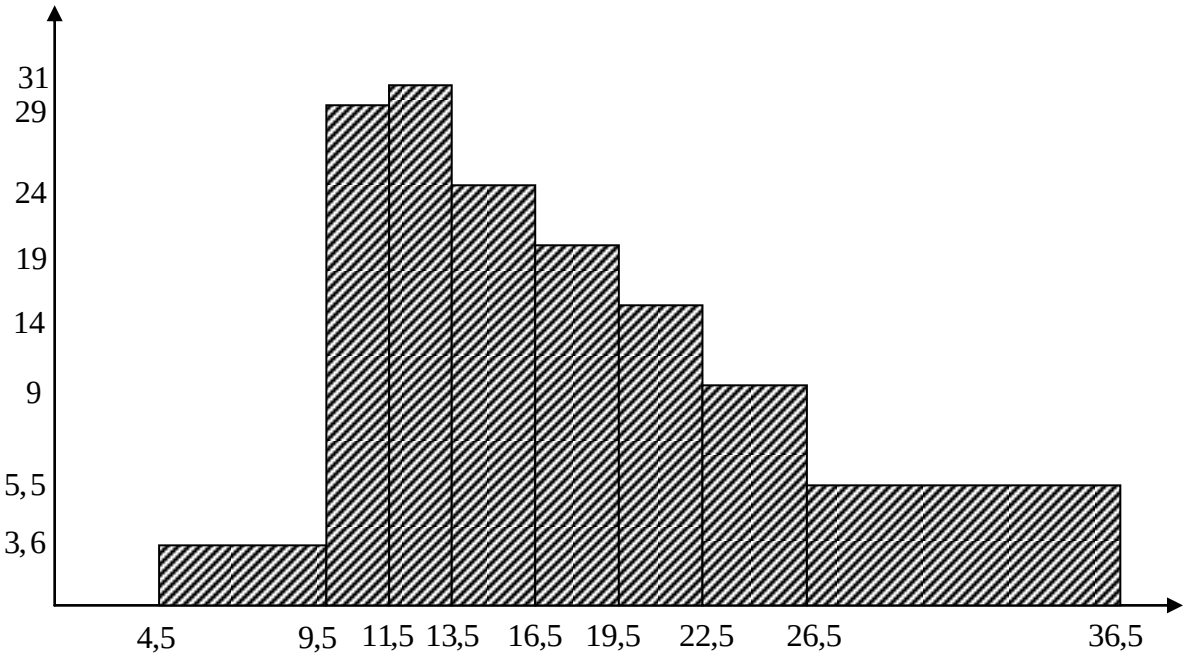


*Hình 5.*2: Biểu đồ đa giác tần suất

5.2.6 Tổ chức đồ (histogram)

Đối với bảng phân bố ghép lớp, người ta thường dùng tổ chức đồ để biểu diễn.

Trong mặt phẳng với hệ toạ độ Oxy, trên trục hoành ta chia các khoảng C_i có độ rộng l_i. Với mỗi khoảng C_i ta dựng hình chữ nhật có chiều cao $y_i = \frac{r_i}{l_i}$ (đối với tổ chức đồ tần số), hay $y_i = \frac{f_i}{l_i}$ (đối với tổ chức đồ tần suất).



Hình 5.3: Tổ chức đồ tần số của mẫu ghép lớp của ví dụ 5.3

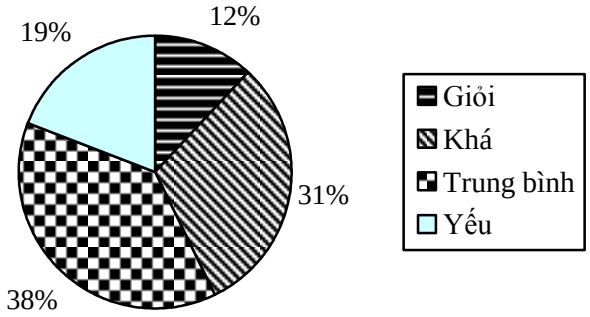
Chú ý rằng diện tích giới hạn bởi tổ chức đồ bằng tần số xuất hiện. Chẳng hạn số cây con nằm trong khoảng (12; 25] chính là diện tích của tổ chức đồ giới hạn bởi đường thẳng $x = 12$ và $x = 25$.

$$(13{,}5-12)\times 31+(16{,}5-13{,}5)\times 24+(19{,}5-16{,}5)\times 19+(22{,}5-19{,}5)\times 14+(25-22{,}5)\times 9=240$$

Vậy có 240 cây con có chiều cao từ $12cm$ đến $25cm$.

Khi dấu hiệu điều tra có tính chất định tính thì người ta thường mô tả các số liệu mẫu bằng biểu đồ hình bánh xe. Đó là hình tròn được chia thành những góc có diện tích tỷ lệ với các tần số tương ứng của mẫu.

Ví dụ 5.4: Tổng kết kết quả học tập của sinh viên Học viện CNBCVT trong năm 2005 được số liệu sau:



5.3 THỐNG KÊ VÀ CÁC ĐẶC TRƯNG CỦA MẪU NGẪU NHIÊN

5.3.1 Định nghĩa thống kê

Một thống kê của mẫu là một hàm của các biến ngẫu nhiên thành phần của mẫu. Thống kê của mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$ *có dạng:*

$$T = T(X_1, X_2, ..., X_n) \qquad \textbf{(5.5)}$$

Như vậy thống kê T cũng là một biến ngẫu nhiên tuân theo một quy luật phân bố xác suất nhất định và có các tham số đặc trưng như kỳ vọng $\mathrm{E}T$ phương sai $\mathrm{D}T$ …(xem mục 3.5 chương 3). Mặt khác, khi mẫu ngẫu nhiên nhận một giá trị cụ thể $w = (x_1, x_2, ..., x_n)$ thì T cũng nhận một giá trị cụ thể còn gọi là giá trị quan sát của thống kê

$$T_{qs} = T(x_1, x_2, ..., x_n)$$

Các thống kê cùng với quy luật phân bố xác suất của chúng là cơ sở để suy rộng các thông tin của mẫu cho dấu hiệu nghiên cứu của tổng thể.

5.3.2 Trung bình mẫu

Trung bình mẫu của mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$ của biến ngẫu nhiên gốc X được định nghĩa và ký hiệu

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \textbf{(5.6)}$$

Giá trị trung bình mẫu cụ thể của mẫu ngẫu nhiên cụ thể $w = (x_1, x_2, ..., x_n)$ là

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.7}$$

Giả sử dấu hiệu nghiên cứu X có kỳ vọng và phương sai hữu hạn, áp dụng các công thức tính kỳ vọng và phương sai của tổng các biến ngẫu nhiên độc lập (2.47), (2.48), (2.58) ta có

$$\mathrm{E}\left(\overline{X}\right) = \mathrm{E}\,X\ ;\ \ \mathrm{D}\left(\overline{X}\right) = \frac{\mathrm{D}\,X}{n}. \tag{5.8}$$

5.3.3 Phương sai mẫu

- Phương sai mẫu S^2:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - \left(\overline{X}\right)^2 \tag{5.9}$$

$$\mathrm{E}\left[S^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left((X_i - \mu) + (\mu - \overline{X})\right)^2\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\left(\left(X_i - \mu\right)^2 + \left(\overline{X} - \mu\right)^2 - 2\left(X_i - \mu\right)(\overline{X} - \mu)\right)\right]$$

$$= \frac{1}{n}\mathrm{E}\left[\left(\sum_{i=1}^{n}\left(X_i - \mu\right)^2\right) + n\left(\overline{X} - \mu\right)^2 - 2n(\overline{X} - \mu)(\overline{X} - \mu)\right]$$

$$= \frac{1}{n}\mathrm{E}\left[\left(\sum_{i=1}^{n}\left(X_i - \mu\right)^2\right) - n\left(\overline{X} - \mu\right)^2\right] = \frac{1}{n}\left(n\,\mathrm{D}\,X - n\frac{\mathrm{D}\,X}{n}\right) = \frac{n-1}{n}\mathrm{D}\,X\,.$$

Để kỳ vọng của phương sai mẫu trùng với phương sai của biến ngẫu nhiên gốc ta cần hiệu chỉnh như sau.

- Phương sai mẫu có hiệu chỉnh S^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2\right) - \frac{n}{n-1}\left(\overline{X}\right)^2 \tag{5.10}$$

- Trường hợp biến ngẫu nhiên gốc X có kỳ vọng xác định $\mathrm{E}\,X = \mu$ thì phương sai mẫu được chọn là S^{*2}

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 \tag{5.11}$$

Áp dụng công thức tính kỳ vọng (2.47), (2.48) và (5.8) ta có:

$$\mathrm{E}\,S^2 = \mathrm{D}\,X \ \text{ và } \ \mathrm{E}\,S^{*2} = \mathrm{D}\,X \tag{5.12}$$

5.3.4 Độ lệch tiêu chuẩn mẫu

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \tag{5.13}$$

5.3.5 Tần suất mẫu

Trường hợp cần nghiên cứu một dấu hiệu định tính A nào đó mà mỗi cá thể của tổng thể có thể có hoặc không, giả sử p là tần suất có dấu hiệu A của tổng thể.

Nếu cá thể có dấu hiệu A ta cho nhận giá trị 1, trường hợp ngược lại ta cho nhận giá trị 0. Lúc đó dấu hiệu nghiên cứu có thể xem là biến ngẫu nhiên X có phân bố Bernoulli tham số p có kỳ vọng $\mathrm{E}X = p$ và phương sai $\mathrm{D}X = p(1-p)$ (công thức (2.17)).

Lấy mẫu ngẫu nhiên: $W = (X_1, X_2, ..., X_n)$, $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên độc lập có cùng phân bố Bernoulli với tham số p. Tần số xuất hiện dấu hiệu A của mẫu là

$$r = X_1 + X_2 + \cdots + X_n. \tag{5.14}$$

Tần suất mẫu

$$f = \frac{r}{n} = \overline{X} \tag{5.15}$$

Như vậy tần suất mẫu là trung bình mẫu của biến ngẫu nhiên X có phân bố Bernoulli tham số p.

Tương tự công thức (5.13) ta có công thức tính kỳ vọng và phương sai của tần suất mẫu:

$$\mathrm{E}(f) = p; \quad \mathrm{D}(f) = \frac{p(1-p)}{n} \tag{5.16}$$

5.3.6 Cách tính giá trị cụ thể của trung bình mẫu và phương sai mẫu $\overline{x}$, s^2

1. Nếu mẫu chỉ nhận các giá trị $x_1, x_2, ..., x_k$ với tần số tương ứng $r_1, r_2, ..., r_k$ thì giá trị trung bình mẫu và phương sai mẫu cụ thể được tính theo công thức

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} r_i x_i, \quad \sum_{i=1}^{k} r_i = n \tag{5.17}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} r_i\left(x_i - \overline{x}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} r_i x_i^2 - \frac{\left(\sum_{i=1}^{k} r_i x_i\right)^2}{n}\right) \tag{5.18}$$

2. Nếu giá trị của mẫu cụ thể được cho dưới dạng bảng phân bố ghép lớp với các khoảng $C_1, ..., C_m$ và tần số của C_i là r_i thì giá trị trung bình mẫu và phương sai mẫu được tính như trên, trong đó

x_i là trung điểm của khoảng C_i. **(5.19)**

3. Mẫu thu gọn: nếu các giá trị của mẫu cụ thể x_i không gọn (quá lớn hoặc quá bé hoặc phân tán) ta có thể thu gọn mẫu bằng cách đổi biến:

$$u_i = \frac{x_i - a}{h} \Rightarrow x_i = hu_i + a \Rightarrow \bar{x} = h\bar{u} + a;\quad s^2 = h^2 s_u^2 \qquad \textbf{(5.20)}$$

trong đó

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{k} r_i u_i,\ s_u^2 = \frac{1}{n-1}\sum_{i=1}^{k} r_i\left(u_i - \bar{u}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} r_i u_i^2 - \frac{\left(\sum_{i=1}^{k} r_i u_i\right)^2}{n}\right) \qquad \textbf{(5.21)}$$

Thật vậy: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} r_i x_i = \frac{1}{n}\sum_{i=1}^{k} r_i\left(hu_i + a\right) = \frac{h}{n}\sum_{i=1}^{k} r_i u_i + \left(\frac{1}{n}\sum_{i=1}^{k} r_i\right)a = h\bar{u} + a$.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} r_i\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{k} r_i\left(hu_i + a - h\bar{u} - a\right)^2 = \frac{h^2}{n-1}\sum_{i=1}^{k} r_i\left(u_i - \bar{u}\right)^2 = h^2 s_u^2.$$

Các số a và h được chọn phù hợp sao cho $\bar{u}$, s_u^2 tính dễ dàng hơn. Thông thường ta chọn a là điểm giữa của các giá trị x_i.

Ví dụ 5.4: Giá trị trung bình mẫu và phương sai mẫu của mẫu ở ví dụ 5.3.

Khoảng	tần số r_i	x_i	$u_i = \frac{x_i - 20}{5}$	$r_i u_i$	$r_i u_i^2$
$4{,}5 - 9{,}5$	18	7	$-2{,}6$	$-46{,}8$	121,68
$9{,}5 - 11{,}5$	58	10,5	$-1{,}9$	$-110{,}2$	209,38
$11{,}5 - 13{,}5$	62	12,5	$-1{,}5$	-93	139,5
$13{,}5 - 16{,}5$	72	15	-1	-72	72
$16{,}9 - 19{,}5$	57	18	$-0{,}4$	$-22{,}8$	9,12
$19{,}5 - 22{,}5$	42	21	0,2	8,4	1,68
$22{,}5 - 26{,}5$	36	24,5	0,9	32,4	29,16
$26{,}5 - 36{,}5$	55	31,5	2,3	126,5	290,95
$\sum$	400			$-177{,}5$	873,47

$$\bar{x} = 5\bar{u} + 20 = 5 \times \frac{-177{,}5}{400} + 20 = 17{,}78.\ s_u^2 = \frac{1}{399} \times \left(873{,}47 - \frac{(-177{,}5)^2}{400}\right) = 1{,}9917$$

$$s^2 = 5^2 \times s_u^2 = 49{,}79 \Rightarrow s = \sqrt{49{,}79} = 7{,}056.$$

5.4 PHÂN BỐ XÁC SUẤT CỦA MỘT SỐ THỐNG KÊ ĐẶC TRƯNG MẪU

5.4.1 Trường hợp biến ngẫu nhiên gốc có phân bố chuẩn

Giả sử dấu hiệu nghiên cứu trong tổng thể có thể xem như một biến ngẫu nhiên X có phân bố chuẩn với kỳ vọng $EX = \mu$ và phương sai $DX = \sigma^2$. Các tham số này có thể đã biết hoặc chưa biết. Từ tổng thể rút ra một mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

Các biến ngẫu nhiên thành phần $X_1, X_2, ..., X_n$ độc lập có cùng quy luật phân bố chuẩn $\mathbf{N}(\mu, \sigma^2)$ như X. Theo định lý 3.10 mọi tổ hợp tuyến tính của các biến ngẫu nhiên có phân bố chuẩn là biến ngẫu nhiên có phân bố chuẩn. Vì vậy ta có các kết quả sau:

5.4.1.1 Phân bố của thống kê trung bình mẫu

Trung bình mẫu $\overline{X}$ có phân bố chuẩn với kỳ vọng $E(\overline{X}) = \mu$ và phương sai $D(\overline{X}) = \frac{\sigma^2}{n}$, áp dụng công thức (2.44) suy ra thống kê sau có phân bố chuẩn tắc $\mathbf{N}(0;1)$:

$$U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim \mathbf{N}(0;1) \qquad \textbf{(5.22)}$$

5.4.1.2 Phân bố của thống kê phương sai mẫu S^{*2}.

Từ công thức (5.18) ta có: $nS^{*2} = \sum_{i=1}^{n} (X_i - \mu)^2$ và $\frac{nS^{*2}}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$.

Vì các biến ngẫu nhiên X_i độc lập nên các biến ngẫu nhiên $\frac{X_i - \mu}{\sigma}$ cũng độc lập. Mặt khác theo công thức (2.44) thì $\frac{X_i - \mu}{\sigma} \sim \mathbf{N}(0;1)$. Do đó thống kê $\chi^2 = \frac{nS^{*2}}{\sigma^2}$ có phân bố "khi bình phương" n bậc tự do (công thức 2.51)

$$\chi^2 = \frac{nS^{*2}}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n) \qquad \textbf{(5.23)}$$

5.4.1.3 Phân bố của thống kê phương sai mẫu S^2.

Tương tự công thức 5.23, thống kê $\frac{(n-1)S^2}{\sigma^2}$ có phân bố "khi bình phương" $n-1$ bậc tự do

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \sim \chi^2(n-1) \qquad \textbf{(5.24)}$$

Áp dụng công thức (2.55) với biến ngẫu nhiên U từ công thức (5.22) và $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ từ công thức (5.24), thì $T = \frac{U}{\sqrt{\chi^2/(n-1)}}$ có phân bố Student $n-1$ bậc tự do.

Vậy

$$T = \frac{(\overline{X} - \mu)\sqrt{n}}{S} = \frac{\frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{U}{\sqrt{\chi^2/(n-1)}} \sim \mathbf{T}(n-1). \qquad \textbf{(5.25)}$$

Phân bố Student $\mathbf{T}(n)$ hội tụ khá nhanh về phân bố chuẩn tắc $\mathbf{N}(0;1)$, do đó trong thực tế khi $n \geq 30$ ta có thể xem thống kê T xấp xỉ $\mathbf{N}(0;1)$.

5.4.2 Phân bố của tần suất mẫu

Giả sử trong tổng thể dấu hiệu nghiên cứu có thể xem như biến ngẫu nhiên có phân bố Bernoulli tham số p. Từ tổng thể rút ra một mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

Từ công thức (5.14), (5.15), (5.16) ta biết rằng tần suất mẫu $f = \frac{X_1 + \cdots + X_n}{n}$ là biến ngẫu nhiên có kỳ vọng và phương sai: $\mathrm{E}(f) = p$; $\mathrm{D}(f) = \frac{pq}{n}$.

Áp dụng định lý 4.6 (Định lý Moivre-Laplace) và công thức 4.12 ta có:

$$\text{Với mọi } x \in \mathrm{R},\ \lim_{n\to\infty} P\left\{\frac{(f-p)\sqrt{n}}{\sqrt{pq}} < x\right\} = \Phi(x). \qquad \textbf{(5.26)}$$

Như vậy có thể xấp xỉ thống kê $U = \frac{(f-p)\sqrt{n}}{\sqrt{pq}}$ với phân bố chuẩn tắc $\mathbf{N}(0;1)$ khi n đủ lớn. Người ta thấy rằng xấp xỉ là tốt khi $np > 5$ và $nq > 5$ hoặc $npq > 20$.

$$U = \frac{(f-p)\sqrt{n}}{\sqrt{pq}} \sim \mathbf{N}(0;1) \text{ khi } \begin{cases} np > 5 \\ nq > 5 \end{cases} \text{ hoặc } npq > 20. \qquad \textbf{(5.27)}$$

TÓM TẮT

Mẫu ngẫu nhiên kích thước n của biến ngẫu nhiên gốc X là véc tơ ngẫu nhiên $W = (X_1, ..., X_n)$, trong đó các biến ngẫu nhiên thành phần $X_1, ..., X_n$ độc lập và có cùng phân bố với biến ngẫu nhiên gốc X. Mỗi thống kê là một hàm của mẫu $T = T(X_1, ..., X_n)$.

Các thống kê thường gặp: trung bình mẫu $\overline{X}$, phương sai mẫu có hiệu chỉnh S^2 và tần suất mẫu $f = \overline{X}$ của dấu hiệu định tính được đặc trưng bởi biến ngẫu nhiên gốc X có phân bố Bernoulli.

Áp dụng công thức (5.17), (5.18) để tính giá trị cụ thể của trung bình mẫu và phương sai mẫu. Trường hợp giá trị cụ thể của không gọn ta sử dụng công thức (5.20) tính theo mẫu rút gọn.

Nếu biến ngẫu nhiên gốc X có phân bố chuẩn $\mathbf{N}(\mu, \sigma^2)$ thì trung bình mẫu $\overline{X}$ cũng có phân bố chuẩn $\mathbf{N}\left(\mu, \sigma^2 / n\right)$, vậy $U = \dfrac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim \mathbf{N}(0;1)$.

Thống kê $\dfrac{(n-1)S^2}{\sigma^2}$ có phân bố "khi bình phương" $n-1$ bậc tự do, trong đó S^2 là phương sai mẫu có hiệu chỉnh, do đó $\dfrac{(\overline{X} - \mu)\sqrt{n}}{S} \sim \mathbf{T}(n-1)$.

Áp dụng định lý Moivre-Laplace cho tần suất mẫu ta có thể xấp xỉ thống kê của tần suất mẫu với phân bố chuẩn tắc $U = \dfrac{(f - p)\sqrt{n}}{\sqrt{pq}} \sim \mathbf{N}(0;1)$.

CÂU HỎI ÔN TẬP VÀ BÀI TẬP

5.1 Mẫu ngẫu nhiên kích thước n về dấu hiệu nghiên cứu X là một dãy gồm n biến ngẫu nhiên: $X_1, X_2, ..., X_n$ độc lập cùng phân bố với X.

Đúng ☐ Sai ☐.

5.2 Một thống kê của mẫu ngẫu nhiên là con số cụ thể về dấu hiệu nghiên cứu.

Đúng ☐ Sai ☐.

5.3 Trung bình mẫu của dấu hiệu nghiên cứu có phân bố chuẩn cũng có phân bố chuẩn.

Đúng ☐ Sai ☐.

5.4 Một thống kê của mẫu là một hàm của các biến ngẫu nhiên thành phần của mẫu do đó cũng là một biến ngẫu nhiên .

Đúng ☐ Sai ☐.

5.5 Từ tổng thể có dấu hiệu nghiên cứu X có bảng phân bố xác suất sau

X	0	1
P	0,5	0,5

lập mẫu ngẫu nhiên kích thước $n = 10$. Tính xác suất để trung bình mẫu của mẫu ngẫu nhiên này nhận giá trị 0,5.

5.6 Giả sử biến ngẫu nhiên gốc có phân bố chuẩn $\mathbf{N}(20;1)$. Chọn mẫu ngẫu nhiên kích thước $n = 100$. Hãy tính xác suất để trung bình mẫu $\overline{X}$ nằm trong khoảng: $19{,}8 < \overline{X} < 20{,}2$.

5.7 Một mẫu cụ thể của biến ngẫu nhiên X như sau:

$$2\ ;\ 3\ ;\ 2\ ;\ 4\ ;\ 1\ ;\ 4\ ;\ 2\ ;\ 2\ ;\ 3\ ;\ 1\ \ (n=10).$$

a) Lập bảng phân bố tần suất.

b) Xây dựng hàm phân bố thực nghiệm.

Tính $\overline{x}$, s^2, s.

5.8 Hãy tính giá trị trung bình mẫu $\overline{x}$ và phương sai mẫu s^2 của mẫu cụ thể có bảng phân bố tần số thực nghiệm sau

x_i	21	24	25	26	28	32	34
r_i	10	20	30	15	10	10	5

5.9 Hãy tính giá trị trung bình mẫu $\overline{x}$ và phương sai mẫu s^2 của mẫu cụ thể có bảng phân bố tần số thực nghiệm sau

x_i	18,6	19,0	19,4	19,8	20,2	20,6
r_i	4	6	30	40	18	2