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

Để xử lý dấu hiệu cần nghiên cứu đôi khi người ta sử dụng phương pháp nghiên cứu toàn bộ

Tuy nhiên trong thực tế việc áp dụng phương pháp này gặp phải những khó khăn sau:

- Qui mô của tập hợp cần nghiên cứu quá lớn
- Trong nhiều trường hợp không thể nắm được toàn bộ
- Có thể trong quá trình điều tra sẽ phá hủy đối tượng nghiên cứu

5.1.2. Tổng thể nghiên cứu, dấu hiệu nghiên cứu

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***

Mỗi phân tử của tổng thể được gọi là cá thể

Dấu hiệu nghiên cứu của tổng thể có thể được định tính hoặc định lượng. Bằng cách mô hình hóa ta có thể xem dấu hiệu nghiên cứu là một biến ngẫu nhiên xác định trên tổng thể

Mẫu ngẫu nhiên

Việc chọn ra từ tổng thể một tập con nào đó gọi là ***phép lấy mẫu***. Tập hợp con này được gọi là ***một mẫu***

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

5.1.3 Mô hình hóa mẫu ngẫu nhiên

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 muốn biết 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, và 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: $X_1, X_2, \dots, X_n$ độc lập cùng phân bố với X, ký hiệu

$$W = (X_1, X_2, \dots, X_n)$$

trong đó X_i là dấu hiệu X của phần tử thứ i của mẫu ($i=1, \dots, n$)

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 ($i=1, \dots, n$), khi đó các giá trị $x_1, x_2, \dots, x_n$ tạo thành một giá trị của mẫu ngẫu nhiên, hay còn gọi là một thể hiện của mẫu ngẫu nhiên, ký hiệu

$$w = (x_1, x_2, \dots, x_n)$$

Ví dụ: Gọi X là số chấm của mặt xuất hiện khi tung con xúc xắc cân đối, X là biến ngẫu nhiên nhận các giá trị $1,\ldots,$ n đồng khả năng

Tung con xúc xắc 3 lần và gọi X_i là số nốt 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. Vây 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 nốt, lần thứ hai được 5 nốt lần ba được 3 nốt thì $w=(2,5,3)$ là một mẫu cụ thể của mẫu ngẫu nhiên W.

5.2 CÁC PHƯƠNG PHÁP MÔ TẢ 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, $i=1, \ldots, k$

$$x_1 < \cdots < x_k ; \; r_1 + \cdots + r_k = n$$

Bảng phân bố tần số thực nghiệm

X	x_1	x_2	$\cdots$	x_k
Tần số	r_1	r_2	$\cdots$	r_k

5.2.2 Bảng phân bố tần suất thực nghiệm

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

Bảng phân bố tần suất thực nghiệm của X

X	x_1	x_2	$\cdots$	x_k
Tần suất	f_1	f_2	$\cdots$	f_k

5.2.3 Hàm phân bố thực nghiệm của mẫu

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

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.

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

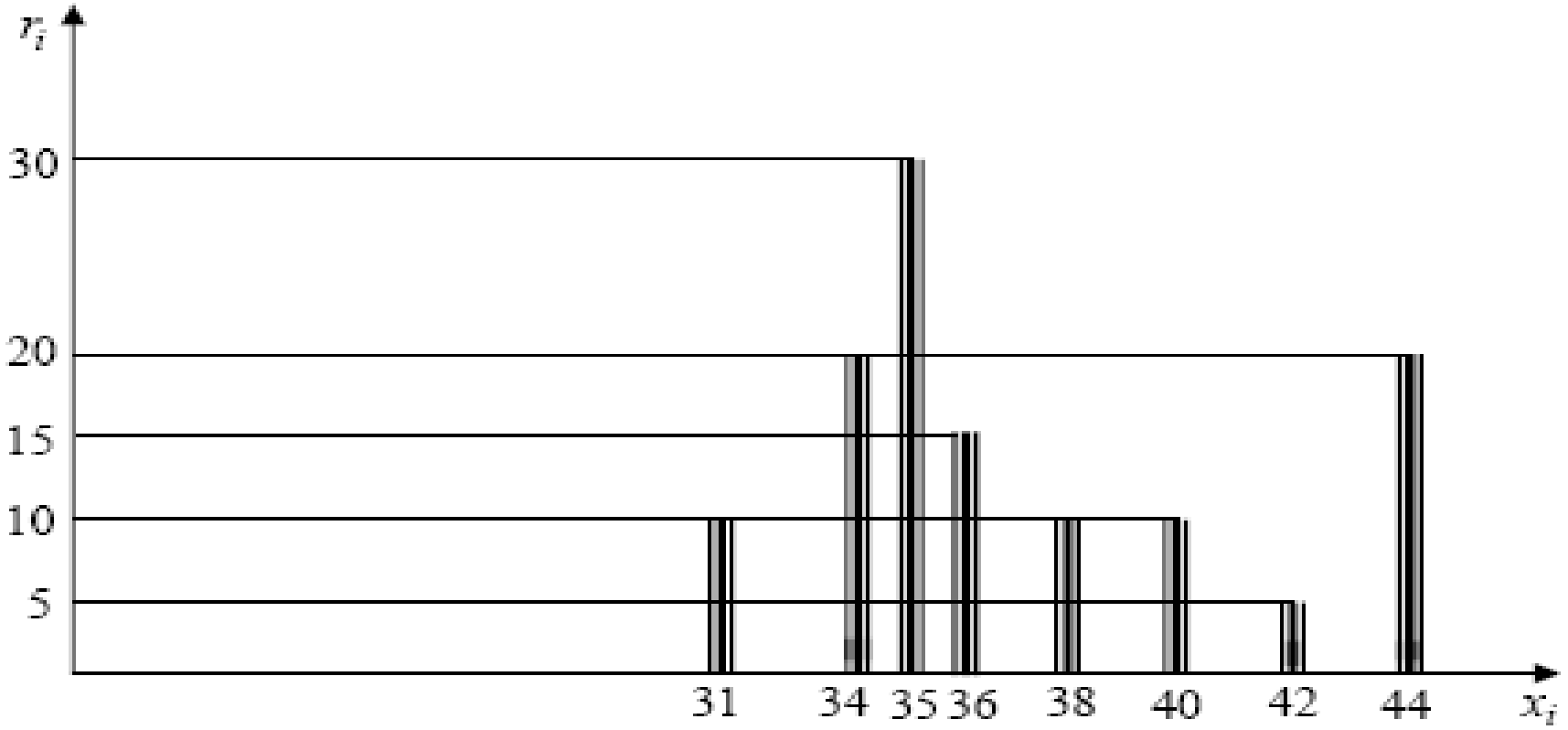
Khoảng	Tần số r_i	Tần suất f_i	Độ rộng khoảng l_i	$y_i = r_i / l_i$
$4,5-9,5$	18	0,045	5	3,6
$9,5-11,5$	58	0,145	2	29
$11,5-13,5$	62	0,155	2	31
$13,5-16,5$	72	0,180	3	24
$16,5-19,5$	57	0,1425	3	19
$19,5-22,5$	42	0,105	3	14
$22,5-26,5$	36	0,090	4	9
$26,5-36,5$	55	0,1375	10	5,5

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

5.2.5 Biểu diễn bằng biểu đồ

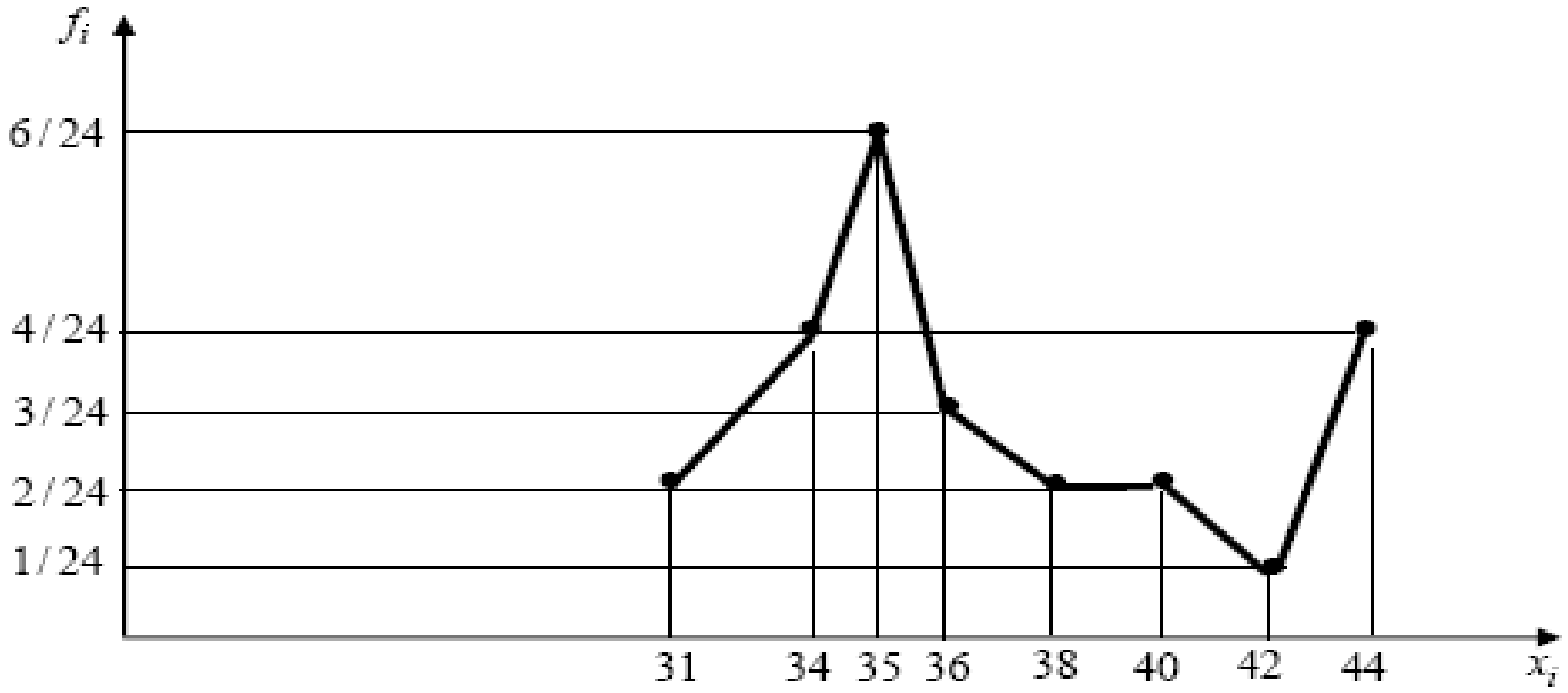
Biểu đồ tần số hình gậy

X	31	34	35	36	38	41	42	44	$\sum$
Tần số	10	20	30	15	10	10	5	20	120



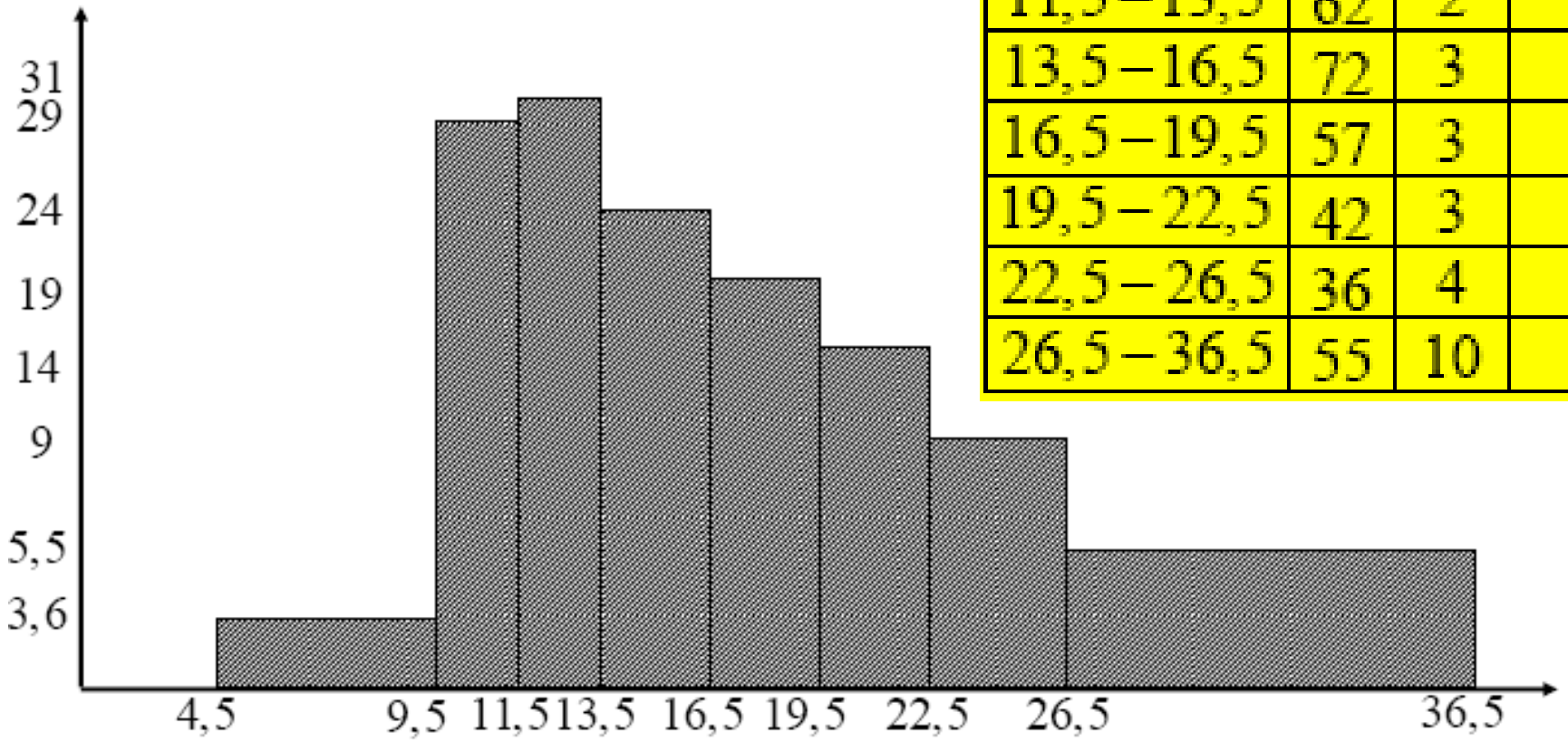
Biểu đồ đa giác tần suất

X	31	34	35	36	38	41	42	44	$\sum$
Tần suất	2/24	4/24	6/24	3/24	2/24	2/24	1/24	4/24	1



5.2.6 Tổ chức đồ (histogram)

Khoảng	r_i	l_i	$y_i = r_i / l_i$
4,5 – 9,5	18	5	3,6
9,5 – 11,5	58	2	29
11,5 – 13,5	62	2	31
13,5 – 16,5	72	3	24
16,5 – 19,5	57	3	19
19,5 – 22,5	42	3	14
22,5 – 26,5	36	4	9
26,5 – 36,5	55	10	5,5



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 $W=(X_1, X_2, \ldots, X_n)$ có dạng

$$T = T(X_1, X_2, \ldots, X_n)$$

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 $\mathrm{D}T$ …

Với một giá trị cụ thể $w=(x_1, x_2, \ldots, x_n)$ của mẫu thì T cũng nhận một giá trị cụ thể gọi là giá trị quan sát được của thống kê

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

5.3.2 Trung bình mẫu

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

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

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

Kỳ vọng, phương sai của trung bình mẫu biến ngẫu nhiên gốc X

$$\mathrm{E}\left(\overline{X}\right) = \mathrm{E}X \qquad \mathrm{D}\left(\overline{X}\right) = \frac{\mathrm{D}X}{n}$$

5.3.3 Phương sai mẫu

1. Phương sai mẫu S^2

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

2. Phương sai mẫu có hiệu chỉnh S^2

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

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

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

5.3.4 Độ lệch tiêu chuẩn mẫu

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

5.3.5 Tần suất mẫu

Biến ngẫu nhiên gốc X có phân bố Bernoulli tham số p là xác suất xuất hiện biến cố A

Lấy mẫu ngẫu nhiên $W=(X_1, X_2, \ldots, X_n)$. Tần số xuất hiện dấu hiệu A của mẫu là

$$r = X_1 + X_2 + \cdots + X_n$$

Tần suất mẫu $f = \frac{r}{n} = \overline{X}$

5.3.6 Cách tính giá trị cụ thể của trung bình mẫu và phương sai mẫu

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

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

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

2. Nếu giá trị của mẫu cụ thể được cho dưới dạng bảng phân bố ghép lớp với các khoảng $C_1, \dots, C_m$ thì giá trị x_i trong thức trên là trung điểm của khoảng C_i

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:

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

Thật vậy

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

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

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

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

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

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

5.4 PHÂN BỐ XÁC SUẤT CỦA MỘT SỐ THỐNG KÊ MẪU

5.4.1 Trường hợp biến ngẫu nhiên gốc có phân bố chuẩn

Giả sử biến ngẫu nhiên gốc X có phân bố chuẩn $\mathbf{N}(\mu;\sigma^2)$. Các tham số này có thể đã biết hoặc chưa biết.

Từ tổng thể rút ra một mẫu ngẫu nhiên $W=(X_1, X_2, \ldots, X_n)$

Các biến ngẫu nhiên thành phần $X_1, X_2, \ldots, X_n$ độc lập và có cùng phân bố chuẩn như biến ngẫu nhiên gốc X

Từ tính chất: 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 $\mathrm{E}\left(\overline{X}\right)=\mu, \mathrm{D}\left(\overline{X}\right)=\frac{\sigma^2}{n}$

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

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

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

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

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

5.4.2 Phân bố của tần suất mẫu

Giả sử biến ngẫu nhiên gốc của tổng thể có phân bố Bernoulli tham số p

Từ tổng thể rút ra một mẫu ngẫu nhiên $W=(X_1, X_2, \ldots, X_n)$

Tần suất mẫu $f = \dfrac{X_1 + \cdots + X_n}{n}$ là một biến ngẫu nhiên có kỳ vọng và phương sai $\mathrm{E}(f) = p; \mathrm{D}(f) = \dfrac{pq}{n}$

Áp dụng Định lý Moivre-Laplace ta có

$$\text{Với mọi } x \in 3\,,\ \lim_{n\to\infty} P\left\{\frac{(f-p)\sqrt{n}}{\sqrt{pq}} \le x\right\} = \Phi(x)$$

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 $npq > 20$

Vậy có thể coi

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