Chương III

CÁC SỐ ĐẶC TRƯNG

Những nội dung chính:

- Định nghĩa kì vọng toán và các tính chất của nó.
- Định nghĩa phương sai và các tính chất của nó.
- Các số đặc trưng khác (mô men, mod, trung vị).
- Phân phối điều kiện và kỳ vọng điều kiện.

Những kiến thức chuẩn bị:

- Các kiến thức về giải tích.
- Các kiến thức ở chương I và chương II.

1. KỲ VỌNG TOÁN

1.1. Định nghĩa kì vọng toán

Định nghĩa 3.1. Giả sử biến ngẫu nhiên X có phân phối xác suất:

X	x_1	x_2	....	x_n	....
P[X = x_i]	p_1	p_2	....	p_n	....

với $\sum_{i=1}^{\infty} p_i = 1$.

Nếu $\sum_{i=1}^{\infty} |x_i| p_i < +\infty$ thì gọi tổng $\sum_{i=1}^{\infty} x_i p_i$ là kỳ vọng toán của biến ngẫu nhiên rời rạc X và kí hiệu là: $E(X) = \sum_{i=1}^{\infty} x_i p_i$.

Ví dụ 3.1. Cho phân phối xác suất của biến ngẫu nhiên X là:

X	1	2	3	4	5	6
P[X = x_i]	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tính kỳ vọng của X.

Giải:

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3,5.$$

Ví dụ 3.2. Bắn 3 viên đạn độc lập vào một mục tiêu. Xác suất bắn trúng đích của mỗi viên đạn là 0,5. Gọi X là số viên đạn trúng đích trong 3 viên. Tính kì vọng của X.

Giải:

Giá trị của X có thể nhận là 0, 1, 2, 3.

Coi việc bắn 3 viên đạn độc lập như việc thực hiện 3 phép thử Bernoulli với xác suất $p = \frac{1}{2}$. Theo công thức xác suất nhị thức ta có:

$$P[X = k] = C_3^k \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{3-k} = C_3^k . \frac{1}{8} \quad \text{với} \quad k = 0, 1, 2, 3.$$

Vậy:

X	0	1	2	3
P[X = k]	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$E(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1,5.$$

Ví dụ 3.3. Giả sử biến ngẫu nhiên X có phân phối nhị thức với tham số (n, p). Tính kỳ vọng của X.

Giải:

$$E(X) = \sum_{k=0}^{n} k C_n^k p^k q^{n-k} = \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

$$= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-1-(k-1))!} p^{k-1} q^{n-1-(k-1)}.$$

Đặt $r = k - 1$ ta có:

$$E(X) = np \sum_{r=0}^{n-1} \frac{(n-1)!}{r!(n-1-r)!} p^r q^{n-1-r}$$

$$= np \sum_{r=0}^{n-1} C_{n-1}^r p^r q^{n-1-r}$$

$$= np(p+q)^{n-1} = np.$$

Ví dụ 3.4. Giả sử biến ngẫu nhiên X có phân phối Poisson với tham số $\lambda > 0$. Tính kỳ vọng của X.

Giải:

$$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

Đặt $r = k - 1$ ta có: $E(X) = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^r}{r!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$

Vậy $E(X) = \lambda$.

Ví dụ 3.5. Giả sử biến ngẫu nhiên X có phân phối hình học, nghĩa là:

$$P[X = n] = q^{n-1}p; \quad n = 1, 2, 3,...; \quad q = 1 - p.$$

Tính kỳ vọng của X.

Giải:

$$E(X) = \sum_{n=1}^{\infty} nq^{n-1}p = p \sum_{n=1}^{\infty} nq^{n-1}.$$

Vì $|q|<1$ nên $\sum_{n=0}^{\infty} q^n$ hội tụ (chuỗi số dương). Vậy $\sum_{n=1}^{\infty} nq^{n-1} = \left(\sum_{n=0}^{\infty} q^n\right)'$

(đạo hàm theo biến q).

Ta lại có: $\sum_{n=0}^{\infty} q^n = \frac{1}{1-q}$

và $$\left(\frac{1}{1-q}\right)' = \frac{1}{(1-q)^2} = \frac{1}{p^2}$$

Vậy $E(X) = p \times \frac{1}{p^2} = \frac{1}{p}$.

Định nghĩa 3.2. Giả sử biến ngẫu nhiên X có hàm mật độ là f(x). Nếu $\int_{-\infty}^{+\infty} |x| f(x)dx < +\infty$ thì gọi tích phân $\int_{-\infty}^{+\infty} xf(x)dx$ là kỳ vọng toán của biến ngẫu nhiên X và kí hiệu $E(X) = \int_{-\infty}^{+\infty} xf(x)dx$.

Ví dụ 3.6. Giả sử biến ngẫu nhiên X có hàm mật độ là:

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{với } x > 0,\ \theta > 0 \\ 0 & \text{với } x \le 0 \end{cases}$$

Tìm kỳ vọng của X.

Giải:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{0}^{+\infty} x \frac{1}{\theta} e^{-\frac{x}{\theta}} dx.$$

Tích phân từng phần ta có:

$$E(X) = \frac{1}{\theta}\left[x\left(-\theta e^{-\frac{x}{\theta}}\right)\Big|_0^{+\infty} + \theta \int_0^{+\infty} e^{-\frac{x}{\theta}} dx \right] = \frac{1}{\theta}\left[0 + \theta\left(-\theta e^{-\frac{x}{\theta}}\right)\Big|_0^{+\infty} \right] = \frac{1}{\theta} \times \theta^2 = \theta.$$

Ví dụ 3.7. Giả sử biến ngẫu nhiên X có phân phối chuẩn dạng tổng quát $N(a, \sigma^2)$. Tính kỳ vọng của X.

Giải:

$$E(X) = \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}} dx.$$

Đặt $t = \frac{x-a}{\sigma}$. Ta suy ra $dt = \frac{dx}{\sigma}$.

Vậy $E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (a + \sigma t) e^{-\frac{t^2}{2}} dt$

$$= a \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t e^{-\frac{t^2}{2}} dt.$$

Tích phân thứ hai ở vế phải bằng 0 vì hàm dưới dấu tích phân là hàm lẻ; lại lấy tích phân trên cận đối nhau. Theo tính chất của tích phân xác định, nó bằng 0. Còn tích phân thứ nhất bằng 1 vì $\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ là hàm mật độ chuẩn dạng N(0; 1).

Vậy $E(X) = a$.

1.2. Tính chất của kỳ vọng toán

a. $E(c) = c$ (với c là hằng số).

b. Nếu X, Y có kỳ vọng thì $X \pm Y$ cũng có kỳ vọng và $E(X \pm Y) = E(X) \pm E(Y)$.

Chứng minh

Ta chỉ chứng minh trong trường hợp X, Y là hai biến ngẫu nhiên rời rạc với các giá trị tương ứng $x_1, x_2, ..., x_n, ...$ và $y_1, y_2, ..., y_m, ...$

Đặt $P([X = x_i] \cap [Y = y_j]) = p_{ij}$, $i = 1, 2, ...; j = 1, 2, ...$

Theo định nghĩa kỳ vọng ta có:

$$E(X) = \sum_{i=1}^{\infty} x_i p_i;\ E(Y) = \sum_{j=1}^{\infty} y_j q_j, \text{ trong đó } p_i = \sum_{j=1}^{\infty} p_{ij};\ q_j = \sum_{i=1}^{\infty} p_{ij}.$$

$$\text{Do đó } E(X) + E(Y) = \sum_{i=1}^{\infty} x_i \left(\sum_{j=1}^{\infty} p_{ij} \right) + \sum_{j=1}^{\infty} y_j \left(\sum_{i=1}^{\infty} p_{ij} \right)$$

$$= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} x_i p_{ij} + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} y_j p_{ij} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (x_i + y_j) p_{ij}$$

$$= E(X + Y).$$

Đó là điều phải chứng minh.

c. Nếu X, Y là hai biến ngẫu nhiên độc lập và có kỳ vọng E(X), E(Y) thì XY cũng có kỳ vọng và: E(XY) = E(X).E(Y).

Chứng minh

Ta chứng minh trong trường hợp X, Y là hai biến ngẫu nhiên rời rạc. Theo kết quả của chứng minh tính chất b ta có:

$$E(X).\ E(Y) = \left(\sum_{i=1}^{\infty} x_i p_i \right) \left(\sum_{j=1}^{\infty} y_j q_j \right).$$

Do các chuỗi hội tụ tuyệt đối nên: $E(X).E(Y) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} x_i y_j p_i q_j$

Vì X, Y độc lập nên $p_{ij} = p_i q_j$.

Vậy $E(X).E(Y) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} x_i y_j p_{ij} = E(XY)$. Đó là điều phải chứng minh.

Hệ quả. E(cX) = cE(X), trong đó c là hằng số, X có kỳ vọng E(X).

1. Nếu X > 0 thì $E(X) \geq 0$.

2. Nếu $X \geq Y$ và có E(X), E(Y) thì $E(X) \geq E(Y)$.

3. Giả sử biến ngẫu nhiên X có hàm mật độ là f(x) và biến ngẫu nhiên Y = φ(X) có kỳ vọng Eφ(X). Khi đó $E\varphi(X) = \int_{-\infty}^{+\infty} \varphi(x)f(x)dx$.

Nếu X có phân phối rời rạc $P[X = x_i] = p_i$, i = 1, 2,... thì

$$E\varphi(X) = \sum_{i=1}^{\infty} \varphi(x_i)p_i.$$

Ví dụ 3.8. Giả sử biến ngẫu nhiên X có phân phối xác suất là:

X	0	1
	$\frac{1}{2}$	$\frac{1}{2}$

Tính E(2X + 1), $E(X^3)$, $E(e^X)$.

Giải:

$$E(2X + 1) = 2E(X) + 1.$$

Mà $E(X) = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{1}{2}$.

Vậy $E(2X + 1) = 2 \times \frac{1}{2} + 1 = 2$.

$$E(X^3) = 0^3 \times \frac{1}{2} + 1^3 \times \frac{1}{2} = \frac{1}{2}.$$

$$E(e^x) = e^0 \times \frac{1}{2} + e^1 \times \frac{1}{2} = \frac{1}{2}(1 + e).$$

Ví dụ 3.9. Giả sử biến ngẫu nhiên X có hàm mật độ:

$$f(x) = \begin{cases} e^{-x} & \text{với } x > 0 \\ 0 & \text{với } x < 0 \end{cases}$$

Tính $E(X^3)$.

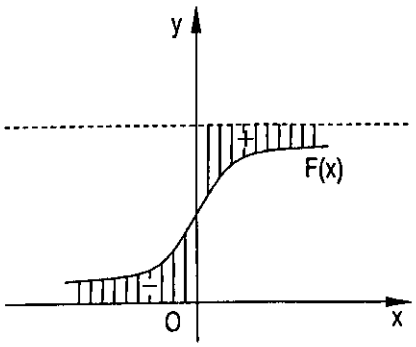
Giải:

$$E(X^3) = \int_0^{+\infty} x^3 e^{-x} dx = \Gamma(4) = 3! = 6.$$

Ý nghĩa của kỳ vọng:

Kì vọng E(X) là trung bình có trọng lượng. Nếu lặp lại độc lập n lần một phép thử để đo đại lượng X. Kết quả của n phép thử đó là $X_1, X_2, ..., X_n$. Dưới một số giả thiết nhất định ta có $\frac{X_1 + X_2 + ... + X_n}{n}$ hội tụ về E(X) khi $n \to \infty$. Vậy với n khá lớn ta có công thức xấp xỉ $\frac{X_1 + X_2 + ... + X_n}{n} \approx E(X)$.

Ý nghĩa hình học của kỳ vọng E(X):



Hình 3.1

$$E(X) = \int_0^{+\infty} (1 - F(x))dx - \int_0^{+\infty} F(x)dx.$$

2. PHƯƠNG SAI

2.1. Định nghĩa 3.3

Gọi phương sai của biến ngẫu nhiên X là $E(X - E(X))^2$ và kí hiệu: $DX = E(X - E(X))^2$.

Ví dụ 3.9. Giả sử biến ngẫu nhiên X có phân phối là:

X	0	1
	$\frac{1}{2}$	$\frac{1}{2}$

Tìm phương sai của X.

Giải:

$$E(X) = 0 \times \frac{1}{4} + 1 \times \frac{3}{4} = \frac{3}{4}.$$

Phương sai của X là:

$$DX = \left(0 - \frac{3}{4}\right)^2 \times \frac{1}{4} + \left(1 - \frac{3}{4}\right)^2 \times \frac{3}{4} = \frac{9}{4^3} + \frac{3}{4^3} = \frac{12}{4^3} = \frac{3}{16}.$$

2.2. Tính chất của phương sai

a. $Dc = 0$ (c là hằng số).

b. $DX = E(X^2) - (EX)^2$.

Thực vậy:

$$DX = E(X - E(X))^2 = E(X^2 - 2XE(X) + E^2(X)) = E(X^2) - (E(X))^2.$$

c. Nếu X, Y độc lập và có phương sai thì: $D(X + Y) = DX + DY$.

Thực vậy:

$$D(X + Y) = E[(X + Y) - E(X + Y)]^2$$

$$= E[(X - E(X)) + (Y - E(Y))]^2$$

$$= E(X - E(X))^2 + E(Y - E(Y))^2 - 2E(X - E(X))(Y - E(Y)).$$

Vì X, Y độc lập nên $E(X - E(X))(Y - E(Y)) = E[X - E(X)].E[Y - E(Y)] = 0$

Vậy $D(X + Y) = D(X) + D(Y)$.

d. $D(cX) = c^2DX$; c là hằng số.

Ví dụ 3.10. Trở về ví dụ 3.3. Tính phương sai của biến ngẫu nhiên X có phân phối nhị thức với tham số (n; p).

Phương sai của X là: $DX = E(X^2) - (E(X))^2$.

Ta biết $E(X) = np$. Bây giờ tính $E(X^2)$.

$$E(X^2) = \sum_{k=0}^{n} k^2 C_n^k p^k q^{n-k} = \sum_{k=0}^{n} k \frac{n!}{(k-1)!(n-k)!} p^k q^{n-k}$$

$$= \sum_{k=0}^{n} \frac{(k-1+1)n!}{(k-1)!(n-k)!} p^k q^{n-k} =$$

$$= n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2} q^{n-2-(k-2)}$$

$$+ np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} q^{n-1-(k-1)}.$$

Đặt $r = k - 2$ và $s = k - 1$ ta có:

$$E(X^2) = n(n-1)p^2 \sum_{r=0}^{n-2} C_{n-2}^r p^r q^{n-2-r} + np \sum_{k=1}^{n} C_{n-1}^s p^s q^{n-1-s}$$

$$= n(n-1)p^2(p+q)^{n-2} + np(p+q)^{n-1}$$

$$= n(n-1)p^2 + np.$$

Phương sai của X là:

$$DX = n(n-1)p^2 + np - (np)^2 = np - np^2 = np(1-p) = npq.$$

Ví dụ 3.11. Trở về ví dụ 3.4. Giả sử biến ngẫu nhiên X có phân phối Poisson với tham số $\lambda > 0$. Tính phương sai của X.

Giải:

Ta biết $EX = \lambda$ (ví dụ 3.4).

Bây giờ ta tính $E(X^2)$.

$$E(X^2) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \left[\sum_{k=1}^{\infty} (k-1+1) \frac{\lambda^k}{(k-1)!} \right]$$

$$= e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}.$$

Đặt $r = k - 2$ và $s = k - 1$ ta có:

$$E(X^2) = e^{-\lambda} \lambda^2 \sum_{r=0}^{\infty} \frac{\lambda^r}{r!} + e^{-\lambda} \lambda \sum_{s=0}^{\infty} \frac{\lambda^s}{s!} = \lambda^2 e^{-\lambda} e^{\lambda} + \lambda e^{-\lambda} e^{\lambda} = \lambda^2 + \lambda.$$

Vậy $DX = E(X^2) - (E(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Ví dụ 3.12. Tính phương sai của biến ngẫu nhiên X có phân phối hình học $P[X = k] = q^{k-1}p$.

Giải:

Ta có: $E(X) = \frac{1}{p}$. (ví dụ 3.5).

$$E(X^2) = \sum_{k=1}^{\infty} k^2 q^{k-1} p = p \sum_{k=1}^{\infty} (k-1+1) k q^{k-1}$$

$$= pq \sum_{k=1}^{\infty} (k-1) k q^{k-1} + p \sum_{k=1}^{\infty} k q^{k-1}.$$

Tổng thứ nhất ở vế phải là đạo hàm hạng 2 của $\sum_{k=1}^{\infty} q^k$, còn tổng thứ hai là đạo hàm hạng nhất của $\sum_{k=1}^{\infty} q^k$.

Mà $\sum_{k=1}^{\infty} q^k = \frac{q}{1-q}$.

Vậy $E(X^2) = pq \times \frac{2}{(1-q)^3} + p \times \frac{1}{(1-q)^2} = \frac{2q}{p^2} + \frac{1}{p}$.

$$\Rightarrow DX = E(X^2) - (EX)^2 = \frac{2q}{p^2} + \frac{1}{p} - \left(\frac{1}{p}\right)^2 = \frac{2q + p - 1}{p^2} = \frac{q}{p^2}.$$

Ví dụ 2.13. Tính phương sai của biến ngẫu nhiên X có phân phối chuẩn dạng $N(a, \sigma^2)$.

Giải:

Ta có $E(X) = a$. Bây giờ ta tính $E(X^2)$.

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}} dx.$$

Đặt $t = \frac{x-a}{\sigma}$. Ta suy ra $dt = \frac{dx}{\sigma}$ và

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma t + a)^2 e^{-\frac{t^2}{2}} dt$$

$$= \frac{1}{\sqrt{2\pi}} \sigma^2 \int_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} 2a\sigma \int_{-\infty}^{+\infty} t e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} a^2 \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt.$$

Tích phân thứ hai bằng 0, vì hàm số dưới dấu tích phân là hàm lẻ và cận của tích phân đối nhau. Tích phân thứ ba bằng a^2 vì $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = 1$ (tính chất của hàm mật độ).

Bây giờ ta tính: $\int_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}} dt = 2 \int_{0}^{+\infty} t^2 e^{-\frac{t^2}{2}} dt$.

Đặt $u = \frac{t^2}{2}$. Ta suy ra $du = tdt$ và $t = \sqrt{2u}$.

$$\int_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}} dt = 2^{3/2} \int_{0}^{+\infty} u^{1/2} e^{-u} du = 2^{3/2} \Gamma\left(\frac{3}{2}\right) = 2^{\frac{3}{2}} \times \frac{1}{2} \Gamma\left(\frac{1}{2}\right) = \sqrt{2\pi}.$$

Vậy $E(X^2) = \frac{1}{\sqrt{2\pi}} \sigma^2 \sqrt{2\pi} + 0 + a^2 = \sigma^2 + a^2$.

Phương sai của X là $DX = E(X^2) - (EX)^2 = \sigma^2 + a^2 - a^2 = \sigma^2$.

Ví dụ 3.14. Giả sử biến ngẫu nhiên X có hàm mật độ là:

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{với } x > 0, \theta > 0. \\ 0 & \text{với } x \le 0 \end{cases}$$

Tính E(X) và DX.

Giải:

$E(X) = \int_0^{+\infty} x \frac{1}{\theta} e^{-\frac{x}{\theta}} dx$. Đặt $t = \frac{x}{\theta}$. Ta suy ra $dt = \frac{dx}{\theta}$.

Vậy $E(X) = \theta \int_0^{+\infty} te^{-t} dt = \theta\Gamma(2) = \theta.1! = \theta.$

Tương tự ta có: $E(X^2) = \int_0^{+\infty} x^2 \frac{1}{\theta} e^{-\frac{x}{\theta}} dx = \theta^2 \int_0^{+\infty} t^2 e^{-t} dt = \theta^2\Gamma(3) = 2\theta^2.$

Phương sai của X là: $DX = E(X^2) - (EX)^2 = 2\theta^2 - \theta^2 = \theta^2$.

Ví dụ 3.15. Giả sử biến ngẫu nhiên X có phân phối đều trên đoạn [0; 1]. Tính kỳ vọng và phương sai của X.

Giải:

Hàm mật độ của X có dạng: $f(x) = \begin{cases} 1 & \text{với } x \in [0;1] \\ 0 & \text{với } x \notin [0;1] \end{cases}$

$$E(X) = \int_0^1 x dx = \frac{1}{2}.$$

$$E(X^2) = \int_0^1 x^2 dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}.$$

$$DX = E(X^2) - (EX)^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}.$$

Ý nghĩa của phương sai:

Về mặt toán học, phương sai cho biết độ sai bình phương trung bình. Mặt khác, phương sai cho biết mức độ phân tán giữa các giá trị của X so với vị trí kỳ vọng E(X).

3. MÔ MEN

2.1. Mô men gốc bậc k

Định nghĩa 3.4. Gọi $E(X^k)$ là mô men gốc bậc k của biến ngẫu nhiên X.

Nếu k = 1 thì E(X) là kỳ vọng.

2.2. Mô men trung tâm bậc k

Định nghĩa 3.5. Gọi $E(X - E(X))^k$ là mô men trung tâm bậc k của biến ngẫu nhiên X.

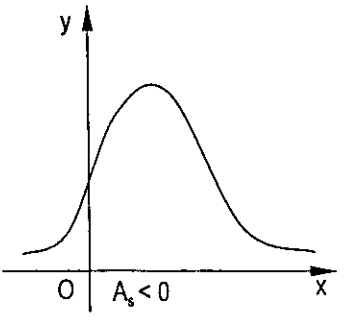
Nếu k = 1 thì E(X – E(X)) = 0.

Nếu k = 2 thì $E(X - EX)^2 = DX$.

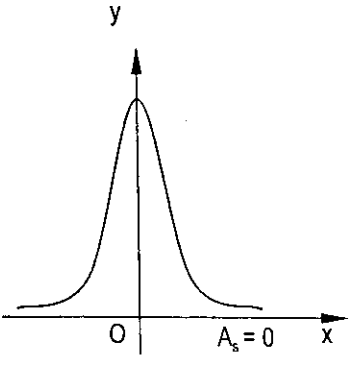
Nếu k = 3 thì $E(X - EX)^3$ sử dụng đo độ lệch trái, phải của đồ thị hàm mật độ.

Ta sử dụng đại lượng sau đây để đo độ lệch trái, phải của đồ thị hàm mật độ.

$$A_s = \begin{cases} \dfrac{1}{\sigma^3}\displaystyle\int_{x_{min}}^{x_{max}} (x - EX)^3 f(x)dx & \text{(Nếu X có hàm mật độ f(x))} \\ \dfrac{1}{\sigma^3}\displaystyle\sum_{i\geq 1}(x_i - EX)^3 p_i & \text{(Nếu X có phân phối rời rạc)} \end{cases}$$



Hình 3.2



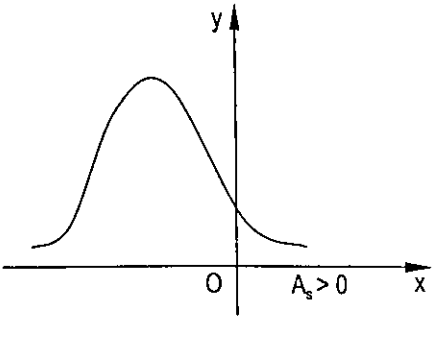
Hình 3.3

Khi $A_s = 0$ phân phối là đối xứng (hình 3.3).

Khi $A_s > 0$ thì phân phối có độ lệch trái (hình 3.4).

Khi $A_s < 0$ thì phân phối có độ lệch phải (hình 3.2).

Nếu k = 4 thì $E(X - EX)^4$ cho biết độ nhọn của đồ thị hàm mật độ.



Hình 3.4

Trong thực tế, người ta dùng đại lượng:

$$\varepsilon_X = \begin{cases} \dfrac{\frac{1}{4}\int_{x_{min}}^{x_{max}} (x-EX)^4 f(x)dx}{\sigma^4} & \text{(Nếu X có hàm mật độ f(x))} \\ \dfrac{\frac{1}{4}\sum_{i=1} (x_i-EX)^4 P_i}{\sigma^4} & \text{(Nếu X có phân phối rời rạc)} \end{cases}$$

4. CÁC SỐ ĐẶC TRƯNG KHÁC

4.1. Mốt (mod)

Định nghĩa 3.6. Mốt (mod) là giá trị của biến ngẫu nhiên X, kí hiệu là x_{mod}, mà tại đó hàm mật độ f(x) đạt cực đại.

Trường hợp X là biến ngẫu nhiên rời rạc, x_{mod} là giá trị, mà xác suất để $X = x_{mod}$ là lớn nhất.

Ví dụ 3.16. Cho phân phối xác suất của biến ngẫu nhiên X là:

X	-1	0	1
p	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

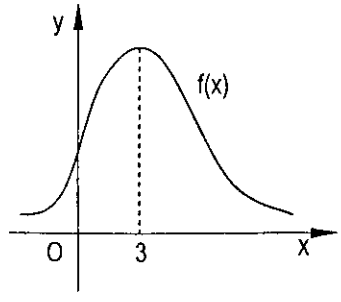
$x_{mod} = 0$.

Ví dụ 3.17. Giả sử biến ngẫu nhiên X có hàm mật độ là

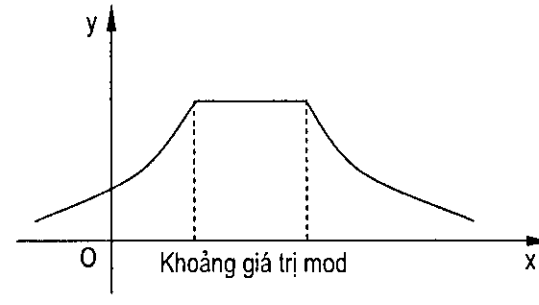
$$f(x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-3)^2}{8}}.$$

$x_{mod} = 3$ vì tại $x = 3$ ta có f(x) đạt cực đại. (Hình 3.5).

Chú ý. Mốt có thể có một giá trị, hoặc cả một khoảng.



Hình 3.5



Hình 3.6

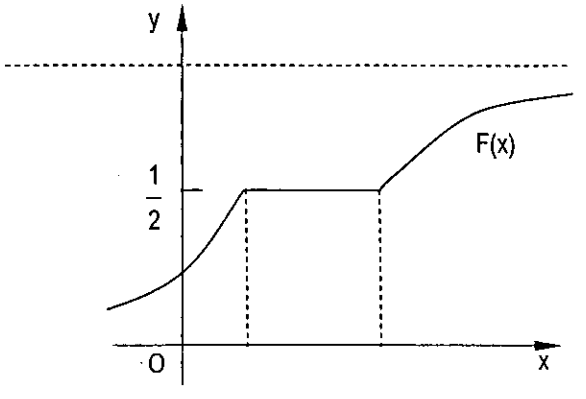
4.2. Trung vị (Median)

Định nghĩa 3.7. Trung vị (Median) là giá trị của biến ngẫu nhiên X, kí hiệu x_{Me}, mà tại đó $F(x_{Me}) = \frac{1}{2}$.

Chú ý 1. Người ta có thể định nghĩa trung vị như sau: Trung vị là giá trị của X, kí hiệu x_{Me} sao cho $P[X \geq x_{Me}] \geq \frac{1}{2} \leq P[X \leq x_{Me}]$.

Ví dụ 3.18. Giả sử biến ngẫu nhiên X có hàm phân phối là:

$$F(x) = \begin{cases} 0 & \text{với } x \leq 0 \\ x & \text{với } 0 < x \leq 1 \\ 1 & \text{với } x > 1 \end{cases}$$



Hình 3.7

Giá trị trung vị x_{Me} suy từ phương trình $F(x_{Me}) = \frac{1}{2}$. Ta suy ra $x_{Me} \approx \frac{1}{2}$.

Chú ý 2. Có trường hợp không có giá trị trung vị, có trường hợp có một trung vị, hoặc có cả một khoảng.

Ví dụ 3.19. Giả sử hàm phân phối của biến ngẫu nhiên X là:

$$F(x) = \begin{cases} 0 & \text{với } x \leq 0 \\ \frac{1}{3} & \text{với } 0 < x \leq 1 \\ 1 & \text{với } x > 1 \end{cases}$$

Trường hợp này không có giá trị trung vị.

Ví dụ 3.20. Giả sử hàm phân phối của biến ngẫu nhiên X là:

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ \frac{1}{2} & \text{với } 0 < x \le 1 \\ 1 & \text{với } x > 1 \end{cases}$$

Trường hợp này có cả một khoảng (0; 1] những giá trị của X mà $F(x) = \frac{1}{2}$.

4.3. Hệ số biến thiên

Định nghĩa 3.8. Tỉ số $\mu = \frac{\sqrt{DX}}{E(X)}$ được gọi là hệ số biến thiên.

Ý nghĩa. Dùng hệ số biến thiên để so sánh độ biến động của các đám đông với nhau.

5. KỲ VỌNG VÀ MA TRẬN TƯƠNG QUAN

5.1. Kỳ vọng toán

Định nghĩa 3.9. Gọi kỳ vọng của vectơ ngẫu nhiên n – chiều $(X_1,..., X_n)$ là 1 vectơ $(EX_1, EX_2,..., EX_n)$ trong đó

$$E(X_i) = \int_{-\infty}^{+\infty} x f_{X_i}(x)dx \,, \; i = \overline{1,n} \,.$$

5.2. Ma trận tương quan

Định nghĩa 3.10. Ma trận tương quan của vectơ ngẫu nhiên n – chiều là ma trận dạng:

$$D = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

trong đó $a_{ij} = E(X_i - EX_i)(X_j - EX_j)$, $a_{ii} = E(X_i - EX_i)^2 = DX_i$, a_{ij} được gọi là hiệp phương sai của X_i, X_j (hoặc gọi là Covariance).

Định nghĩa 3.11. (Hệ số tương quan)

Gọi tỉ số $\dfrac{E(X_i - EX_i)(X_j - EX_j)}{\sqrt{DX_i DX_j}}$ là hệ số tương quan của X_i, X_j và được kí hiệu là $\rho(X_i, X_j)$.

Nếu một trong hai đại lượng X_i, X_j là hằng số thì quy ước $\rho(X_i, X_j) = 0$.

Xét tính chất của hệ số tương quan $-1 \le \rho(X, Y) \le 1$ và $|\rho(X,Y)| = 1$ khi và chỉ khi X và Y là phụ thuộc tuyến tính.

Chứng minh

$$\text{Xét } D\left(\frac{X}{\sqrt{DX}} + \frac{Y}{\sqrt{DY}}\right) = E\left(\frac{X}{\sqrt{DX}} + \frac{Y}{\sqrt{DY}} - E\left(\frac{X}{\sqrt{DX}} + \frac{Y}{\sqrt{DY}}\right)\right)^2$$

$$= \frac{E(X - E(X))^2}{DX} + \frac{E(Y - E(Y))^2}{DY} + 2\frac{E(X - EX)(Y - EY)}{\sqrt{DXDY}}$$

$= 2 + 2\rho(X, Y) \ge 0$ (vì phương sai không âm).

Từ đó suy ra $\rho(X, Y) \ge -1$.

Tương tự ta xét

$$D\left(\frac{X}{\sqrt{DX}} - \frac{Y}{\sqrt{DY}}\right) = E\left(\frac{X}{\sqrt{DX}} - \frac{Y}{\sqrt{DY}} - E\left(\frac{X}{\sqrt{DX}} - \frac{Y}{\sqrt{DY}}\right)\right)^2$$

$$= 2 - 2\rho(X, Y) \ge 0.$$

Từ đó suy ra $\rho(X, Y) \le 1$.

Vậy $-1 \le \rho(X, Y) \le 1$.

Mệnh đề còn lại độc giả xem như bài tập.

6. PHÂN PHỐI ĐIỀU KIỆN VÀ KÌ VỌNG TOÁN ĐIỀU KIỆN

6.1. Phân phối điều kiện

Giả sử X, Y là hai biến ngẫu nhiên rời rạc nhận các giá trị:

X: $x_1, x_2, \ldots, x_n, \ldots$

Y: $y_1, y_2, \ldots, y_m, \ldots$

Với phân phối đồng thời là $P[X = x_i, Y = y_j] = p_{ij}$, $i = 1, 2, \ldots$; $j = 1, 2, \ldots$

Định nghĩa 3.12. a. Gọi tỉ số $P\left[X = x_i \middle| Y = y_j\right] = \dfrac{p_{ij}}{P[Y = y_j]}$ với $i = 1, 2, \ldots$ là phân phối điều kiện của biến ngẫu nhiên X với điều kiện $Y = y_j$ và gọi $P\left[Y = y_j \middle| X = x_i\right] = \dfrac{p_{ij}}{P[X = x_i]}$, với $j = 1, 2, \ldots$, là phân phối điều kiện của biến ngẫu nhiên Y với điều kiện $X = x_i$.

b. Nếu X, Y có hàm mật độ đồng thời là f(x, y) thì gọi $f(x|y) = \dfrac{f(x,y)}{f_Y(y)}$ là hàm mật độ điều kiện của biến ngẫu nhiên X với điều kiện $Y = y$ và gọi $f(y|x) = \dfrac{f(x,y)}{f_X(x)}$ là hàm mật độ điều kiện của biến ngẫu nhiên Y với điều kiện $X = x$.

Định nghĩa 3.13. a. Nếu (X, Y) có phân phối rời rạc thì gọi $E(X|Y = y_j) = \sum_{i=1}^{\infty} x_i P\left[X = x_i \middle| Y = y_j\right]$ là kỳ vọng toán điều kiện của biến ngẫu nhiên X với điều kiện $Y = y_j$

và gọi $E(Y|X = x_i) = \sum_{j=1}^{\infty} y_j P\left[Y = y_j \middle| X = x_i\right]$ là kỳ vọng toán điều kiện của biến ngẫu nhiên Y với điều kiện $X = x_i$.

b. Nếu (X, Y) có hàm mật độ đồng thời là f(x, y) thì gọi tích phân:

$$E(X|Y = y) = \int_{-\infty}^{+\infty} x f(x|y)dx$$

là kỳ vọng toán điều kiện của biến ngẫu nhiên X với điều kiện Y = y,

và gọi $E(Y|X=x) = \int_{-\infty}^{+\infty} yf(y|x)dy$ là kỳ vọng toán điều kiện của biến ngẫu nhiên Y với điều kiện X = x.

Ví dụ 3.21. Giả sử X, Y là hai biến ngẫu nhiên rời rạc và có phân phối đồng thời:

Y \ X	2	5	8
0,4	0,15	0,30	0,35
0,8	0,05	0,12	0,03

Tìm kỳ vọng và phương sai của X, của Y. Tính ma trận tương quan của X, Y. Tìm hệ số tương quan của X, Y. Tính phân phối điều kiện $P[X|Y=0,4]$ và $P[Y|X=5]$ và kỳ vọng toán điều kiện $E(X|Y=0,4)$ và $E(Y|X=5)$.

Giải:

Phân phối xác suất của biến ngẫu nhiên X là:

X	2	5	8
$P[X=x_i]$	0,20	0,42	0,38

Phân phối của biến ngẫu nhiên Y là:

Y	0,4	0,8
$P[Y=y_j]$	0,80	0,20

Kì vọng của biến ngẫu nhiên X là:

$$E(X) = 2 \times 0,2 + 5 \times 0,42 + 8 \times 0,38 = 5,54.$$

Kì vọng của biến ngẫu nhiên Y là:

$$E(Y) = 0,4 \times 0,80 + 0,8 \times 0,20 = 0,48.$$

$$E(X^2) = 0,2 \times 4 + 25 \times 0,42 + 64 \times 0,38 = 35,62.$$

$$E(Y^2) = 0,16 \times 0,8 + 0,64 \times 0,2 = 0,256.$$

$$DX = E(X^2) - (EX)^2 = 35,6 - 5,54^2 = 4,9084.$$

$$DY = E(Y^2) - (EY)^2 = 2,256 - (0,48)^2 = 0,0256.$$

$$a_{12} = a_{21} = E(XY) - (EX)(EY).$$

$$E(XY) = 2 \times 0,4 \times 0,15 + 5 \times 0,4 \times 0,3 + 8 \times 0,4 \times 0,35 +$$

$$+ 2 \times 0,8 \times 0,05 + 5 \times 0,8 \times 0,12 + 8 \times 0,8 \times 0,03 = 2,592.$$

Vậy $a_{12} = a_{21} = 2,592 - 5,54 \times 0,48 = -0,0672$.

Vậy ma trận tương quan của X, Y là:

$$D = \begin{bmatrix} 4,9084 & -0,0672 \\ -0,0672 & 0,0256 \end{bmatrix}$$

Hệ số tương quan của X, Y là:

$$\rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{DXDY}} = \frac{-0,0672}{\sqrt{4,9084 \times 0,0256}} = -0,1895.$$

Phân phối điều kiện của X với điều kiện Y = 0,4:

$$P\left[X = 2 \middle| Y = 0,4\right] = \frac{0,15}{0,80} = \frac{3}{16};\ P\left[X = 5 \middle| Y = 0,4\right] = \frac{0,3}{0,80} = \frac{3}{8};$$

$$P\left[X = 8 \middle| Y = 0,4\right] = \frac{0,35}{0,80} = \frac{7}{16}.$$

Vậy:

X	2	5	8
$P[X = x \mid Y = 0,4]$	$\frac{3}{16}$	$\frac{3}{8}$	$\frac{7}{16}$

Phân phối điều kiện của Y với điều kiện X = 5.

$$P\left[Y = 0,4 \middle| X = 5\right] = \frac{0,3}{0,42} = \frac{5}{7};\ P\left[Y = 0,8 \middle| X = 5\right] = \frac{0,12}{0,42} = \frac{2}{7}.$$

Vậy:

Y	0,4	0,8
$P\left[Y = y_j \middle| X = 5\right]$	$\frac{5}{7}$	$\frac{2}{7}$

Kỳ vọng toán điều kiện của X với điều kiện Y = 0,4 là:

$$E(X|Y = 0,4) = 2 \times \frac{3}{16} + 5 \times \frac{3}{8} + 8 \times \frac{7}{16} = \frac{23}{4}.$$

Kỳ vọng toán điều kiện của Y với điều kiện X = 5 là:

$$E(Y|X = 5) = 0,4 \times \frac{5}{7} + 0,8 \times \frac{2}{7} = \frac{3,6}{7}.$$

Ví dụ 3.22. Giả sử (X, Y) có hàm mật độ đồng thời là:

$$f(x,y) = \begin{cases} \dfrac{12}{(1+x+y)^5} & \text{với } x > 0, y > 0 \\ 0 & \text{trong trường hợp khác} \end{cases}$$

Tìm hàm mật độ điều kiện $f(x|y)$, $f(y|x)$ và kỳ vọng toán điều kiện $E(X|Y)$, $E(Y|X)$.

Giải:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \begin{cases} \displaystyle\int_0^{+\infty} \frac{12}{(1+x+y)^5} dy & \text{với } x > 0 \\ 0 & \text{với } x \le 0 \end{cases}$$

$$f_X(x) = \begin{cases} \dfrac{3}{(1+x)^4} & \text{với } x > 0 \\ 0 & \text{với } x \le 0 \end{cases}$$

Tương tự $f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx = \begin{cases} \dfrac{3}{(1+y)^4} & \text{với } y > 0 \\ 0 & \text{với } y \le 0 \end{cases}$

Mật độ điều kiện của Y với X = x là:

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \begin{cases} \dfrac{12(1+x)^4}{3(1+x+y)^5} & \text{với } x > 0, y > 0 \\ 0 & \text{trong trường hợp khác} \end{cases}$$

Mật độ điều kiện của X với Y = y là:

$$f(x|y) = \frac{f(x,y)}{f_y(y)} = \begin{cases} \dfrac{12(1+x)^4}{3(1+x+y)^5} & \text{với } x > 0, y > 0 \\ 0 & \text{trong trường hợp khác} \end{cases}$$

Kỳ vọng toán điều kiện của Y với điều kiện X = x là:

$$E(Y|X = x) = \int_0^{+\infty} y \frac{4(1+x)^4}{(1+x+y)^5} dy = \frac{1}{3}(1+x).$$

Kỳ vọng toán điều kiện của X với điều kiện Y = y là:

$$E(X|Y = y) = \int_0^{+\infty} x \frac{4(1+y)^4}{(1+x+y)^5} dx = \frac{1}{3}(1+y).$$

Tính chất của kỳ vọng toán điều kiện:

$$E((c_1Y_1 + c_2Y_2)|X) = c_1E(Y_1|X) + c_2E(Y_2|X);$$

$$E(Y) = E(E(Y|X));$$

$$DY = D(E(Y|X)) + E(D(Y|X)).$$

BÀI TẬP CHƯƠNG III

1. Cho phân phối xác suất của biến ngẫu nhiên X là:

X	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

a. Tính kỳ vọng và phương sai của X.

b. Tính mô men bậc ba của X.

2. Một lô hàng gồm 100 sản phẩm, trong đó có 10 sản phẩm xấu. Ta lấy hú họa từ lô hàng một mẫu ngẫu nhiên (để kiểm tra ngẫu nhiên) gồm 5 sản phẩm.

Tìm kỳ vọng của biến ngẫu nhiên X chỉ số sản phẩm xấu trong mẫu.

3. Giả sử N_t là số hạt phóng xạ trong khoảng thời gian t có phân phối Poisson với tham số $\lambda = \frac{1}{2}$.

Tìm kỳ vọng và phương sai của N_t.

4. Gieo ngẫu nhiên 120 hạt đậu tương. Xác suất nảy mầm của mỗi hạt là 0,6. Gọi X là số hạt không nảy mầm trong 120 hạt.

a. Tính kỳ vọng và phương sai của X.

b. Tính mod và median (trung vị).

5. Tiến hành không hạn chế những phép thử độc lập. Xác suất để sự kiện A thành công trong mỗi phép thử là 0,2. Thực hiện liên tiếp các phép thử cho tới khi biến cố A xuất hiện thì dừng lại. Gọi X là số phép thử cần thiết để lần đầu tiên biến cố A xuất hiện.

Tính kỳ vọng và phương sai của X.

6. Tìm kỳ vọng và phương sai của biến ngẫu nhiên X có phân phối xác suất:

X	–5	2	3	4
P	0,4	0,3	0,1	0,2

7. Tìm phương sai của đại lượng ngẫu nhiên X chỉ số lần xuất hiện biến cố A trong 2 phép thử độc lập nếu như xác suất xuất hiện biến cố A trong mỗi phép thử là như nhau và biết rằng kỳ vọng của X là 1,2.

8. Đại lượng ngẫu nhiên rời rạc X chỉ nhận 2 giá trị x_1 và x_2 mà $x_2 > x_1$. Xác suất để X nhận giá trị x_1 bằng 0,6. Tìm phân phối xác suất của X và các giá trị x_1, x_2 mà X có thể nhận nếu như kỳ vọng và phương sai của X là EX = 1,4 và DX = 0,24.

9. Trong một lô hàng có 500 đơn vị hàng hoá. Tỉ lệ hàng kém phẩm chất là 5%. Lấy ngẫu nhiên 50 đơn vị hàng hoá (lấy ra từng đơn vị và có hoàn lại lô hàng). Tìm kỳ vọng của biến ngẫu nhiên X bằng số hàng hoá kém phẩm chất trong 50 đơn vị chọn ra.

10. Xác suất bắn trúng đích của một khẩu súng là p. Tiến hành bắn liên tiếp trong điều kiện không đổi cho đến khi có k phát trúng đích thì thôi bắn.

Tìm kỳ vọng của số lần bắn cần thiết.

11. Cho biến ngẫu nhiên X có hàm mật độ là:

$$f(x) = \begin{cases} \dfrac{1}{\pi\sqrt{a^2 - x^2}} & \text{với } x \in (-a;a) \\ 0 & \text{với } x \notin (-a;a) \end{cases}$$

Tìm kỳ vọng và phương sai của X.

12. Cho biến ngẫu nhiên X có hàm phân phối là:

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ 1 - \dfrac{x_0^3}{x^3} & \text{với } x > 0 \end{cases} \quad \text{với } x_0 > 0.$$

Tìm kỳ vọng và phương sai của X.

13. Giả sử biến ngẫu nhiên X có hàm mật độ là:

$$f(x) = \begin{cases} 0 & \text{với } x \notin (0;\pi) \\ \dfrac{1}{2}\sin x & \text{với } x \in (0;\pi) \end{cases}$$

Tính kỳ vọng và phương sai của $Y = X^2$.

14. Giả sử T là thời gian một công nhân đến một máy dệt bị đứt sợi để nối có phân phối dạng mũ: $f(x) = \begin{cases} \frac{1}{2}e^{-\frac{x}{2}} & \text{với } x > 0 \\ 0 & \text{với } x \le 0 \end{cases}$

Tìm kỳ vọng và phương sai của T.

15. Giả sử biến ngẫu nhiên X có hàm mật độ là:

$f(x) = ae^{-|x|}$

a. Tìm a.

b. Tìm hàm phân phối của X.

c. Tìm kỳ vọng và phương sai của X.

16. Giả sử biến ngẫu nhiên X có hàm mật độ dạng:

$$f(x) = \begin{cases} \frac{n}{x_0}x^{n-1}e^{-\frac{x^n}{x_0}} & \text{với } x \ge 0 \\ 0 & \text{với } x < 0 \end{cases} \quad \text{với } x_0 > 0.$$

a. Tìm mốt của X.

b. Tính kỳ vọng của X.

17. Giả sử biến ngẫu nhiên X có phân phối đều trên đoạn [a; b], tức là hàm mật độ có dạng: $f(x) = \begin{cases} \frac{1}{b-a} & \text{với } x \in [a;b] \\ 0 & \text{với } x \notin [a;b] \end{cases}$

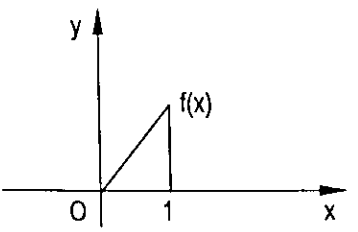
Tìm kỳ vọng và phương sai của X.

18. Cho đồ thị của hàm mật độ của biến ngẫu nhiên X như hình 3.8.

a. Tìm biểu thức của hàm mật độ.

b. Viết hàm phân phối của X.

c. Tìm kỳ vọng và phương sai của X.



Hình 3.8

19. Giả sử biến ngẫu nhiên X nhận các giá trị nguyên dương 1, 2,..., k,...

với xác suất $P[X = k] = \frac{a^k}{(1+a)^{k+1}}$ với $a > 0$, $k = 1, 2,...$

Tính E(X), DX.

20. Chứng minh rằng với X, Y là hai biến ngẫu nhiên bất kì có các kỳ vọng E(X), E(Y) bất đẳng thức sau là đúng:

$$E|XY| < \sqrt{E(X^2)}.\sqrt{E(Y^2)}.$$

LỜI GIẢI – HƯỚNG DẪN – TRẢ LỜI

1. a. $EX = 3,5$; $DX = 2,916$;

b. $EX^3 = 73,5$.

2. $P[X = k] = \dfrac{C_{10}^k \times C_{90}^{5-k}}{C_{100}^5}$; $E(X) = \sum_{k=0}^{5} k \dfrac{C_{10}^k \times C_{90}^{5-k}}{C_{100}^5}$.

3. $P[N_t = k] = \dfrac{\left(\frac{1}{2}\right)^k e^{-\frac{1}{2}}}{k!}$, $k = 0, 1, 2,...$

$E(N_t) = \dfrac{1}{2}$, $D(N_t) = \dfrac{1}{2}$.

4. a. $EX = np = 120 \times 0,6 = 72$.

$DX = npq = 120 \times 0,6 \times 0,4 = 28,8$.

b. $x_{mod} = 72$.

5. $P[X = k] = (0,8)^{k-1} \times 0,2$;

$E(X) = \dfrac{1}{p} = \dfrac{1}{0,2} = 5$;

$DX = \dfrac{q}{p^2} = \dfrac{0,8}{(0,2)^2} = 20$.

6. $E(X) = -0,3$; $E(X^2) = 15,3$; $DX = 15,21$.

7. Xem như tiến hành dãy gồm 2 phép thử Bernoulli với $p(A) = p$.

Ta có: $EX = np = 2p = 1,2$. Suy ra $p = 0,6$, $q = 1 - p = 0,4$.

$DX = np(1 - p) = 2 \times 0,4 \times 0,6 = 0,48$.

8. Ta đưa đến hệ hai phương trình (nhờ công thức tính kỳ vọng và phương sai của X): $\begin{cases} 0,6x_1 + 0,4x_2 = 1,4 \\ 0,6x_1^2 + 0,4x_2^2 = 2,2 \end{cases}$

Giải hệ này ta được $x_1 = 1$, $x_2 = 2$ và phân phối xác suất của X là:

X	1	2
P	0,6	0,4

9. $P[X = k] = C_{500}^{k}(0,05)^k(0,95)^{500-k}$, với $k = 0, 1,..., 25$;

$E(X) = np = 500 \times 0,05 = 25.$

10. $P[X = m] = C_{m-1}^{k-1}p^{k-1}q^{m-k}p = C_{m-1}^{k-1}p^kq^{m-k}$, trong đó X là số lần bắn cần thiết.

$P[X = m] = 0$ với $m < k$.

$$E(X) = \sum_{m=k}^{\infty} mC_{m-1}^{k-1}p^kq^{m-k} = \frac{kp^k}{(1-q)^{k+1}} = \frac{k}{p}.$$

Hoặc có thể lý luận như sau:

Gọi X_1 là số viên đạn cần bắn để lần đầu tiên trúng đích. $E(X_1) = \frac{1}{p}$;

X_2 là số viên đạn cần bắn để lần đầu tiên trúng đích sau viên mà lần thứ nhất trúng đích $E(X_2) = \frac{1}{p}$, $X_1, ..., X_k$ là số viên đạn cần bắn để lần đầu tiên trúng đích sau $k - 1$ viên trúng đích trước đó, $E(X_k) = \frac{1}{p}$.

$$X = X_1 + X_2 + ... + X_k; E(X) = E(X_1) + E(X_2) + ... + E(X_k) = \frac{k}{p}.$$

11. $E(X) = 0$; $D(X) = \frac{a^2}{2}$.

12. $E(X) = \frac{3}{2}x_0$; $DX = \frac{3}{4}x_0^2$.

13. $EY = \frac{\pi^2 - 4}{2}$; $DY = \frac{\pi^4 - 16\pi^2 + 80}{4}$

14. $E(X) = 2$; $DX = 4$.

15. a. $a = \frac{1}{2}$;

b. $F(x) = \frac{1}{2}\int_{-\infty}^{x} e^{-|u|}du = \begin{cases} \frac{1}{2}e^x & \text{với } x \le 0 \\ 1 - \frac{1}{2}e^{-x} & \text{với } x > 0 \end{cases}$;

c. $E(X) = 0$; $DX = 2$.

16. a. $x_{mod} = \left[\frac{(n-1)x_0}{n}\right]^{\frac{1}{n}}$;

b. $E(X) = x_0^{\frac{1}{n}}\Gamma\left(\frac{1}{n}+1\right)$, trong đó: $\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}dx$.

17. $E(X) = \frac{a+b}{2}$; $DX = \frac{(b-a)^2}{12}$.

18. a. Hàm mật độ $f(x) = \begin{cases} 2x & \text{với } x \in (0;1) \\ 0 & \text{với } x \notin (0;1) \end{cases}$;

b. $F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ x^2 & \text{với } 0 < x \le 1 \\ 1 & \text{với } x > 1 \end{cases}$;

c. $E(X) = \frac{2}{3}$; $DX = \frac{1}{18}$.

19. $E(X) = a$; $DX = a^2 + a$.

Chương IV

LUẬT SỐ LỚN VÀ ĐỊNH LÍ GIỚI HẠN TRUNG TÂM

Nội dung chính:

- Luật số lớn dạng Chebyshev và Khinchin.
- Định lí giới hạn trung tâm.

Những kiến thức chuẩn bị:

- Những kiến thức ở chương I, II, III.

1. LUẬT SỐ LỚN

Ở đây chúng ta chỉ trình bày luật số lớn do Chebyshev và Khinchin chứng minh.

Ví dụ 4.1. Gieo n lần một đồng tiền cân đối và đồng chất. Gọi m_A là số lần xuất hiện mặt sấp trong n lần gieo đó. $\frac{m_A}{n}$ được gọi là tần xuất xuất hiện mặt sấp (biến cố A). Người ta thấy rằng nếu số lần gieo càng tăng thì $\frac{m_A}{n}$ càng gần tới $\frac{1}{2}$.

Ví dụ 4.2. Lặp lại n lần đo độc lập đại lượng X trong cùng một điều kiện như nhau. Kết quả các lần đo là $X_1, X_2, ..., X_n$. Thực nghiệm cho ta thấy rằng trung bình số học $\frac{X_1 + X_2 + ... + X_n}{n}$ càng gần kỳ vọng E(X) nếu số lần đo càng tăng ra vô hạn.

Ta có thể đặt vấn đề như sau: Với điều kiện nào thì $\frac{m_A}{n}$ hội tụ theo xác suất tới p = p(A) khi $n \to \infty$, nghĩa là:

$$\lim_{n\to\infty} P\left[\left|\frac{m_A}{n} - p\right| < \varepsilon\right] = 1$$

và $\frac{X_1 + X_2 + \ldots + X_n}{n}$ hội tụ về E(X) theo xác suất khi $n \to \infty$. Các bài toán này được giải nhờ các định lí thuộc loại luật số lớn.

Định nghĩa 4.1. Dãy $(X_n, n \geq 1)$ được gọi là hội tụ theo xác suất về X khi $n \to \infty$ nếu với $\varepsilon > 0$ cho trước tuỳ ý ta có: $\lim_{n\to\infty} P\left[|X_n - X| < \varepsilon\right] = 1$ và kí hiệu $X_n \xrightarrow{P} X$ khi $n \to \infty$.

Định nghĩa 4.2. Dãy các biến ngẫu nhiên $X_1, X_2, \ldots, X_n, \ldots$ được gọi là tuân theo luật (yếu) số lớn nếu với $\varepsilon > 0$ cho trước ta có:

$$\lim_{n\to\infty} P\left[\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n}\right| < \varepsilon\right] = 1.$$

Bất đẳng thức Chebyshev.

Bổ đề. Giả sử biến ngẫu nhiên X có kỳ vọng EX và phương sai DX. Khi đó với $\varepsilon > 0$ cho trước tuỳ ý ta có:

$$P\left[|X - EX| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} DX. \tag{4.1}$$

Chứng minh

Đặt $I_A(\omega) = \begin{cases} 1 & \text{với } \omega \in A \\ 0 & \text{với } \omega \notin A \end{cases}$

Vì $\Omega = \left[\omega : |X - EX| < \varepsilon\right] \cup \left[\omega : |X - EX| \geq \varepsilon\right]$ nên

$$I_\Omega = I_{[|X-EX|<\varepsilon]} + I_{[|X-EX|\geq\varepsilon]}.$$

Ta có: $DX = E(X - EX)^2$

$$= E\left(|X - EX|^2 I_{[|X-EX|<\varepsilon]}\right) + E\left(|X - EX|^2 I_{[|X-EX|\geq\varepsilon]}\right).$$

$$E\left(|X-EX|^2\right) \geq E\left(|X-EX|^2 I_{[|X-EX|\geq\varepsilon]}\right) \geq \varepsilon^2 E I_{[|X-EX|\geq\varepsilon]} \geq \varepsilon^2 P\left[|X-EX| \geq \varepsilon\right]$$

Từ đó suy ra $P\left[|X-EX| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} DX$.

Bổ đề được chứng minh.

Định lí 4.1. (Định lí Chebyshev).

Nếu dãy biến ngẫu nhiên $X_1, X_2, ..., X_n$ độc lập và có phương sai bị chặn bởi cùng một hằng số C, nghĩa là:

$$DX_1 \leq C,\ DX_2 \leq C, ...,\ DX_n \leq C,$$

thì dãy đó tuân theo luật (yếu) số lớn, nghĩa là:

$$\lim_{n\to\infty} P\left[\left|\frac{X_1 + X_2 + ... + X_n}{n} - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| \geq \varepsilon\right] = 0 \qquad (4.2)$$

Chứng minh

Áp dụng bất đẳng thức Chebyshev cho đại lượng ngẫu nhiên $\frac{1}{n}\sum_{i=1}^{n} X_i$, với $\varepsilon > 0$ cho trước tuỳ ý ta có:

$$P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n} DX_i.$$

(Vì $X_1, X_2, ..., X_n$ là độc lập). Theo giả thiết $DX_1 \leq C, DX_2 \leq C, ..., DX_n \leq C$ nên

$$\frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n} DX_i \leq \frac{nC}{\varepsilon^2 n^2} = \frac{C}{\varepsilon^2 n}.$$

Vậy $P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| \geq \varepsilon\right] \leq \frac{C}{\varepsilon^2 n}$. Vế phải $\frac{C}{\varepsilon^2 n} \to 0$ khi $n \to \infty$.

Theo định nghĩa luật (yếu) số lớn thì dãy $X_1, X_2, ..., X_n$ tuân theo luật (yếu) số lớn. Định lí được chứng minh.

Hệ quả 4.1. Giả sử k là số lần xuất hiện biến cố A trong dãy n phép thử Bernoulli và p là xác suất để A xuất hiện trong mỗi phép thử.

Khi đó $\frac{k}{n} \xrightarrow{P} p$ khi $n \to \infty$.

Chứng minh

Đặt X_k là số lần xuất hiện biến cố A trong phép thử thứ k, nghĩa là

X_k	0	1
	1 – p	p

Ta có $k = X_1 + X_2 + ... + X_n$.

Dãy $X_1, X_2, ..., X_n$ độc lập; $EX_k = 0.(1 - p) + 1.p = p$.

$$E(X_k^2) = 0^2.(1 - p) + 1^2.p = p;$$

$$DX_k = E(X_k^2) - (EX_k)^2 = p - p^2;$$

$$DX_k = p(1 - p) \le \frac{1}{4}.$$

Theo định lí Chebyshev ta có:

$$\lim_{n\to\infty} P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| \le \varepsilon\right] = \lim_{n\to\infty} P\left[\left|\frac{k}{n} - p\right| \le \varepsilon\right] = 1.$$

Vậy $\frac{k}{n} \xrightarrow{P} p$ khi $n \to \infty$.

Hệ quả được chứng minh.

(Hệ quả 1 chính là luật (yếu) số lớn dạng Bernoulli).

Định lí 4.2. (Định lí Khinchin) Nếu các đại lượng ngẫu nhiên $X_1, X_2, ..., X_n$ độc lập, có phân phối như nhau và kì vọng bằng a, $EX_i = a$, $i = \overline{1,n}$ thì

$$\frac{X_1 + X_2 + ... + X_n}{n} \xrightarrow{P} a \quad \text{khi} \quad n \to \infty.$$

Chứng minh

Đặt $Y_k = \begin{cases} X_k & \text{nếu } |X_k| < \delta_n \\ 0 & \text{nếu } |X_k| \geq \delta_n \end{cases}$ và $Z_k = \begin{cases} 0 & \text{nếu } |X_k| < \delta_n \\ X_k & \text{nếu } |X_k| \geq \delta_n \end{cases}$.

Ta có $X_k = Y_k + Z_k$, $k = 1, 2, ..., n$.

Vì các đại lượng ngẫu nhiên $X_1, X_2, ..., X_n$ độc lập, có phân phối như nhau và kì vọng bằng a, nên các đại lượng $Y_1, Y_2, ..., Y_n$ cũng độc lập, có phân phối như nhau và cùng kỳ vọng bằng:

$$a_n = EY_k = EX_k I_{A_k}, \quad k = 1, 2, ..., n.$$

trong đó $A_k = [\omega : |X_k| < \delta_n]$, $I_{A_k} = \begin{cases} 1 & \text{với } |X_k| < \delta_n \\ 0 & \text{với } |X_k| \geq \delta_n \end{cases}$.

$$DY_k = E(Y_k^2) - (EY_k)^2 = E(Y_k^2) - a_n^2 < E(Y_k^2)$$

$$= EX_k^2 I_{A_k} < \delta_n E|X_k| I_{A_k} \leq \delta_n E|X_k| = \delta_n b,$$

trong đó $b = E|X_k| < +\infty$.

Vì $a_n \to a$ khi $n \to \infty$, nên với n đủ lớn ta có $|a_n - a| < \varepsilon$ với $\varepsilon > 0$ cho trước tuỳ ý.

Do đó các đại lượng ngẫu nhiên $Y_1, Y_2, ..., Y_n$ thoả mãn các điều kiện của bất đẳng thức Chebyshev. Ta có:

$$P\left[\left|\frac{1}{n}\sum_{k=1}^{n} Y_k - a_n\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} D\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right) = \frac{1}{\varepsilon^2 n^2}\sum_{k=1}^{n} DY_k$$

$$\leq \frac{nb\delta_n}{\varepsilon^2 n^2} = \frac{b\delta_n}{\varepsilon^2 n} = \frac{b\delta}{\varepsilon^2}, \quad \delta = \frac{\delta_n}{n}.$$

Vì $\left|\frac{1}{n}\sum_{k=1}^{n} Y_k - a\right| \le \left|\frac{1}{n}\sum_{k=1}^{n} Y_k - a_n\right| + |a_n - a|$

nên $P\left[\left|\frac{1}{n}\sum_{k=1}^{n} Y_k - a\right| \ge 2\varepsilon\right] \le P\left[\left|\frac{1}{n}\sum_{k=1}^{n} Y_k - a_n\right| \ge \varepsilon\right] + P\left[|a_n - a| \ge \varepsilon\right]$

$$\le \frac{b\delta}{\varepsilon^2} + \varepsilon_1 \quad \forall \varepsilon > 0 \text{ và } \varepsilon_1 > 0 \text{ tuỳ ý.}$$

Mặt khác

$$P[Z_k \ne 0] = P\left[|X_k| \ge \delta_n\right] = EI_{[|X_k| \ge \delta_n]} < E\frac{|X_k|}{\delta_n}$$

$$\le \frac{1}{\delta_n} E|X_k| I_{[|X_k| \ge \delta_n]} \le \frac{b\delta}{\varepsilon^2} + \varepsilon_1 \quad \forall \varepsilon > 0 \text{ và } \varepsilon_1 > 0 \text{ tuỳ ý.}$$

Mặt khác

$$P[Z_k \ne 0] = P\left[|X_k| \ge \delta_n\right] = EI_{[|X_k| \ge \delta_n]} < E\frac{|X_k|}{\delta_n} \le \frac{1}{\delta_n} E|X_k| I_{[|X_k| \ge \delta_n]}.$$

Vì $E|X_k| < +\infty$ nên $E|X_k| I_{[|X_k| \ge \delta_n]}$ tiến tới 0 khi $n \to \infty$.

Vậy với n đủ lớn ta có: $\frac{1}{\delta_n} E|X_k| I_{[|X_k| \ge \delta_n]} \le \frac{\delta}{n}$.

Do đó $P[Z_k \ne 0] \le \frac{\delta}{n}$ và $P\left[\sum_{k=1}^{n} Z_k \ne 0\right] = \sum_{k=1}^{n} P[Z_k \ne 0] < n\frac{\delta}{n} = \delta.$

Từ kết quả trên ta suy ra:

$$P\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_k - a\right| \ge 2\varepsilon\right] \le P\left[\left|\frac{1}{n}\sum_{k=1}^{n} Y_k - a_n\right| \ge 2\varepsilon\right] + P\left[\sum_{k=1}^{n} Z_k \ne 0\right]$$

$$\le \frac{b\delta}{\varepsilon^2} + \varepsilon_1 + \delta.$$

Vì δ, ε_1 là nhỏ tuỳ ý nên ta có:

$$\lim_{n\to\infty} P\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_k - a\right| \geq 2\varepsilon\right] = 0.$$

Định lí được chứng minh.

Ví dụ 4.3. Giả sử $X_1, X_2, ..., X_n$ là dãy các biến ngẫu nhiên độc lập và X_k ($k = \overline{1, n}$) có phân phối xác suất:

X_k	-1	0	1
p	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Chứng minh rằng dãy đó tuân theo luật số lớn.

Giải:

$$EX_k = -1 \times \frac{1}{3} + 0 + 1 \times \frac{1}{3} = 0$$

$$DX_k = E(X_k^2) - (EX_k)^2 = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

Theo định lí Chebyshev ta có: $\dfrac{X_1 + X_2 + ... + X_n}{n} \xrightarrow{P} 0$ khi $n \to \infty$.

2. ĐỊNH LÍ GIỚI HẠN TRUNG TÂM

Ở mục trên ta đã xét điều kiện để $\frac{1}{n}\sum_{k=1}^{n} X_k$ hội tụ khi $n \to \infty$. Trong mục này ta xét phân phối giới hạn của tổng các đại lượng ngẫu nhiên độc lập. Các loại định lí này được gọi là định lí giới hạn trung tâm. Người đầu tiên xét vấn đề này trong trường hợp dãy $X_1, X_2, ..., X_n$ có phân phối:

X_k	0	1
	1 - p	p

là De Moivre (năm 1730), tiếp sau là P.S Laplace (năm 1783). Định lí giới hạn trung tâm tiếp tục được mở rộng bởi A.M Liapunov (năm 1901),

sau đó là Lindeberg (năm 1922), v.v... Do giới hạn của chương trình, chúng tôi chỉ trình bày định lí giới hạn trung tâm dạng Lindeberg mà không chứng minh. Độc giả có thể tham khảo trong [1].

Giả sử dãy biến ngẫu nhiên $X_1, X_2,..., X_n$ là độc lập.

• Điều kiện Lindeberg:

Giả sử với hằng số $\tau > 0$ bất kỳ thực hiện điều kiện:

$$\lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^{n} E\left(X_k - EX_k\right)^2 I_{[|X_k - EX_k| > \tau B_n]} = 0 \tag{4.3}$$

trong đó $B_n^2 = \sum_{k=1}^{n} DX_k$, $I_A(\omega) = \begin{cases} 1 & \text{với } \omega \in A \\ 0 & \text{với } \omega \notin A \end{cases}$.

Định lí 4.3. Nếu dãy biến ngẫu nhiên $X_1, X_2,..., X_n$ độc lập thoả mãn điều kiện Lindeberg (4.3) thì:

$$\lim_{n\to\infty} P\left[\frac{1}{B_n} \sum_{k=1}^{n} (X_k - EX_k) < x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du$$

đều với mọi $x \in \mathbf{R}$.

Hệ quả 4.2. Giả sử $X_1, X_2,..., X_n$ là dãy biến ngẫu nhiên độc lập và có phân phối như nhau và phương sai hữu hạn. Khi đó:

$$\lim_{n\to\infty} P\left[\frac{\sum_{i=1}^{n} (X_i - a)}{\sigma\sqrt{n}} < x\right] = \Phi(x)$$

đều với mọi $x \in \mathbf{R}$,

trong đó $EX_k = a$, $DX_k = \sigma^2$, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du$.

Chứng minh

Ta chỉ cần nghiệm lại điều kiện Lindeberg (4.3).

Theo giả thiết $DX_k = \sigma^2 < +\infty$, k = 1, 2,..., n và $B_n^2 = \sum_{k=1}^{n} DX_k = n\sigma^2$.

Vì $X_1, X_2,..., X_n$ có phân phối như nhau nên $EX_k = a$ với $k = \overline{1, n}$. Do phương sai của X_k là hữu hạn nếu với $\varepsilon > 0$ cho trước tuỳ ý ta có thể chọn n đủ lớn để:

$E(X_k - a)^2 I_{\left[|X_k - a| > \tau\sigma\sqrt{n}\right]}$ đủ bé, chẳng hạn nhỏ hơn $\varepsilon\sigma^2$.

Ta có:

$$\frac{1}{B_n^2}\sum_{k=1}^{n} E(X_k - a)^2 I_{\left[|X_k - a| > \tau B_n\right]} = \frac{1}{n\sigma^2}\sum_{k=1}^{n} E(X_k - a)^2 I_{\left[|X_k - a| > \tau\sigma\sqrt{n}\right]} \le \frac{n\varepsilon\sigma^2}{n\sigma^2} = \varepsilon.$$

Cho $\varepsilon \to 0$, điều kiện Lindeberg thoả mãn, theo định lí 4.3 ta có kết luận của hệ quả.

Hệ quả 4.3. Gọi k là số lần xuất hiện biến cố A trong dãy n phép thử Bernoulli, p là xác suất để biến cố A xuất hiện trong mỗi phép thử, $0 < p < 1$. Khi đó: $\lim_{n\to\infty} P\left[\frac{k - np}{\sqrt{np(1-p)}} < x\right] = \Phi(x)$ đều với mọi $x \in R$.

Chứng minh

Đặt X_k là số lần xuất hiện biến cố A ở phép thử thứ k. Ta có phân phối xác suất của X_k ($k = \overline{1, n}$) là:

X_k	0	1
	1 - p	p

Do giả thiết, n phép thử Bernoulli là độc lập nên $X_1, X_2,..., X_n$ cũng độc lập;

$$k = X_1 + X_2 + ... + X_n.$$

$$EX_k = 0.(1-p) + 1.p = p.$$

$$EX_k^2 = 1^2.\, p + 0^2(1-p) = p.$$

$$DX_k = p - p^2 = p(1-p) \le \frac{1}{4}.$$

Theo hệ quả 4.2 ta có:

$$\lim_{n\to\infty} P\left[\frac{\sum_{i=1}^{n}(X_i - p)}{\sqrt{np(1-p)}} < x\right] = \Phi(x)$$

đều với mọi $x \in \mathbf{R}$.

Vì $k = \sum_{i=1}^{n} X_i$ nên ta có: $\lim_{n\to\infty} P\left[\frac{k-np}{\sqrt{np(1-p)}} < x\right] = \Phi(x).$

Hệ quả được chứng minh.

Từ hệ quả 4.3 ta suy ra: với k_1, k_2 là hai hằng số đã cho ($k_1 < k_2$) thì

$$\lim_{n\to\infty} P[k_1 \le k < k_2] = \frac{1}{\sqrt{2\pi}} \int_{k_1}^{k_2} e^{-\frac{u^2}{2}}\, du.$$

Đây chính là định lí Moivre – Laplace.

Hệ quả 4.4. (Định lí Liapunov). Nếu dãy biến ngẫu nhiên $X_1,\ldots,$ X_n độc lập và đối với $\delta > 0$ sao cho:

$$\lim_{n\to\infty} \frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} E\left|X_k - EX_k\right|^{2+\delta} = 0 \tag{4.4}$$

thì $\lim_{n\to\infty} P\left[\frac{1}{B_n}\sum_{k=1}^{n}(X_k - EX_k) < x\right] = \Phi(x)$

đều với mọi $x \in \mathbf{R}$.

Chứng minh

Ta chỉ việc kiểm tra lại điều kiện Lindeberg.

Ta có:

$$\frac{1}{B_n^2}\sum_{k=1}^{n} E\left|X_k - EX_k\right|^2 I_{\left[|X_k-EX_k|>\tau B_n\right]} \le \frac{1}{\tau^\delta B_n^{2+\delta}}\sum_{k=1}^{n} E\left|X_k - EX_k\right|^{2+\delta} I_{\left[|X_k-EX_k|>\tau B_n\right]}$$

$$\le \frac{1}{\tau^\delta B_n^{2+\delta}}\sum_{k=1}^{n} E\left|X_k - EX_k\right|^{2+\delta}$$

Giả thiết (4.4) thoả mãn thì dẫn đến điều kiện Lindeberg (4.3) thoả mãn. Từ đó suy ra kết quả của hệ quả 4.4 nhờ định lí 4.3.

ƯỚC LƯỢNG XÁC SUẤT

a. Ước lượng gần đúng xác suất $P\left[\left|\frac{k}{n} - p\right| < \varepsilon\right]$.

Ta có $P\left[\left|\frac{k}{n} - p\right| < \varepsilon\right] = P\left[-\varepsilon\sqrt{\frac{n}{pq}} < \frac{k-np}{\sqrt{npq}} < \varepsilon\sqrt{\frac{n}{pq}}\right]$, $q = 1 - p$.

Theo hệ quả 4.3, với n đủ lớn ta có:

$$P\left[\left|\frac{k}{n} - p\right| < \varepsilon\right] = \Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - \Phi\left(-\varepsilon\sqrt{\frac{n}{pq}}\right). \qquad (4.5)$$

Vì hàm phân phối chuẩn $\Phi(x)$ có tính chất đối xứng, nghĩa là:

$$\Phi(-x) = 1 - \Phi(x).$$

Do đó $P\left[\left|\frac{k}{n} - p\right| < \varepsilon\right] \approx 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - 1.$ (4.6)

Ví dụ 4.4. Cho $\varepsilon = 1/100$, $p = 1/2$ và $P\left[\left|\frac{k}{n} - p\right| < \varepsilon\right] = 0{,}95$. Tìm n.

Giải:

Theo (4.6) ta có $2\Phi(x) - 1 = 0{,}95$. Từ đó suy ra $\Phi(x) = 0{,}975$. Tra bảng giá trị của hàm phân phối chuẩn N(0; 1) ta tìm được $x = 1{,}96$.

Mà $x = \varepsilon\sqrt{\frac{n}{pq}} = \frac{1}{100}\sqrt{\frac{n}{\frac{1}{2}\times\frac{1}{2}}} = 1,96$

Từ đó suy ra $\sqrt{n} = 98$, $n = 9604$.

b. Cho k_1, k_2 ($k_1 < k_2$); ước lượng xác suất $P[k_1 \le k < k_2]$.

Từ hệ quả 4.3 với k_1, k_2 đã cho ta có công thức xấp xỉ:

$$P[k_1 \le k < k_2] =$$

$$= P\left[\frac{k_1 - np}{\sqrt{npq}} \le k < \frac{k_2 - np}{\sqrt{npq}}\right] \approx \Phi\left(\frac{k_2 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{k_1 - np}{\sqrt{npq}}\right). \quad (4.7)$$

Ví dụ 4.5. Một hộp chứa 10000 quả cầu đỏ và đen. Xác suất bốc được 1 quả cầu đỏ là 0,4. Một học sinh bốc hú hoạ lần lượt từng quả cầu (sau mỗi lần lấy cầu xong lại hoàn trả hộp cũ). Hãy ước lượng xác suất để trong 1000 cầu lấy ra có số cầu đỏ nằm trong khoảng từ 100 đến 200.

Giải:

Gọi k là số cầu đỏ trong 1000 cầu lấy ra.

Chọn ngẫu nhiên 1000 cầu được xem như thực hiện 1000 phép thử Bernoulli với xác suất chọn 1 cầu đỏ là $p = 0,4$.

Theo công thức (4.7) ta có:

$$P[100 \le k < 200] \approx \Phi\left(\frac{200 - 1000\times 0,4}{\sqrt{1000\times 0,4\times 0,6}}\right) - \Phi\left(\frac{100 - 1000\times 0,4}{\sqrt{1000\times 0,4\times 0,6}}\right)$$

$$\approx \Phi(-13) - \Phi(-19).$$

$$\Phi(-13) = 1 - \Phi(13); \Phi(-10) = 1 - \Phi(19).$$

Tra bảng giá trị của hàm phân phối chuẩn ta có $\Phi(13) \approx \Phi(19) \approx 1$.

Vậy $P[100 \le k < 200] \approx 1 - 1 = 0$.

BÀI TẬP CHƯƠNG IV

1. Dùng bất đẳng thức Chebyshev để ước lượng xác suất của biến cố $\left[|X - EX| < 0,2\right]$ nếu $DX = 0,004$.

2. Tiến hành 10 phép thử độc lập. Biến cố A xuất hiện trong khoảng thời gian (0, T) bằng 0,05. Nhờ bất đẳng thức Chebyshev để ước lượng xác suất sao cho giá trị tuyệt đối của hiệu giữa số lần xuất hiện biến cố A và kì vọng của nó:

   a. Nhỏ hơn 2.

   b. Không nhỏ hơn 2.

3. Chứng minh rằng nếu biến ngẫu nhiên dương X có tồn tại kỳ vọng Ee^{aX} ($a > 0$ là hằng số) thì: $P\left[X \geq \varepsilon\right] \leq \dfrac{Ee^{aX}}{e^{a\varepsilon}}$.

4. Các dãy biến ngẫu nhiên độc lập sau đây có tuân theo luật số lớn không, nếu phân phối xác suất tương ứng là:

   a. $P[X_k = \pm 2^k] = 0,5$; $k = 1, 2,..., n$.

   b. $P[X_k = \pm 2^k] = 2^{-(2k+1)}$, $P[X_k = 0] = 1 - 2^{-2k}$, $k = 1, 2,..., n$.

5. Giả sử rằng dãy các biến ngẫu nhiên độc lập $X_1, X_2,..., X_n$ có phân phối như nhau với hàm mật độ là:

$$f(x) = \begin{cases} \dfrac{1}{\theta} e^{-\frac{x}{\theta}} & \text{với } x > 0,\ \theta > 0 \\ 0 & \text{với } x \leq 0 \end{cases}$$

   Chứng minh rằng $\dfrac{X_1 + X_2 + ... + X_n}{n} \xrightarrow{P} \theta$ khi $n \to \infty$.

6. Xác suất sinh con trai bằng 0,51. Tính gần đúng xác suất sao cho trong 1000 lần sinh (mỗi lần sinh 1 con), số con trai ít hơn số con gái.

7. Gieo một con xúc xắc cân đối và đồng chất 12000 lần. Tính gần đúng xác suất để số lần xuất hiện mặt trên của con xúc xắc có 1 chấm nằm trong khoảng (1900, 2150).

8. Cho biến ngẫu nhiên X có hàm phân phối:

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ x & \text{với } 0 < x \le 1 \\ 1 & \text{với } x > 1 \end{cases}$$

Ước lượng xác suất sao cho trong 100 lần lặp lại độc lập phép đo đại lượng X có tối đa 70 lần nhận giá trị trong khoảng (0,2; 0,7).

LỜI GIẢI – HƯỚNG DẪN – TRẢ LỜI

1. $P[|X - EX| < 0,2] \geq 0,9$.

2. a. $P[|X - 0,5| < 2] \geq 0,88$;

b. $P[|X - 0,5| \geq 2] \geq 0,12$.

3. Hướng dẫn: Chứng minh tương tự như chứng minh bất đẳng thức Chebyshev. (Thay phương sai bằng Ee^{aX}).

4. a. Không tuân theo luật số lớn.

b. Tuân theo luật số lớn.

5. Hướng dẫn: Áp dụng bất đẳng thức Chebyshev.

$EX_k = \theta$, $DX_k = \theta^2$. (bị chặn bởi θ^2).

Do đó X_1, X_2,..., X_n tuân theo luật (yếu) số lớn.

6. $$P[X < 500] \approx \Phi\left(\frac{500 - 510}{\sqrt{1000 \times 0,49 \times 0,51}}\right) \approx 0,2643.$$

7. $$P[1900 < X < 2150] \approx \Phi\left(\frac{2150 - 12000/6}{\sqrt{12000 \times \frac{5}{6} \times \frac{1}{6}}}\right) - \Phi\left(\frac{1900 - 12000/6}{\sqrt{12000 \times \frac{5}{6} \times \frac{1}{6}}}\right)$$

$$\approx \Phi\left(\frac{3\sqrt{3}}{\sqrt{2}}\right) - \Phi(-\sqrt{6}) \approx 0,99.$$

8. $P[0,2 \leq X \leq 0,7] = F(0,7) - F(0,2) = \frac{1}{2}$.

Xem 100 lần đo như là tiến hành 100 phép thử Bernoulli với xác suất $p = \frac{1}{2}$, ta có

$$P[k \leq 70] = P[k < 71] \approx \Phi\left(\frac{71 - 100 \times \frac{1}{2}}{\sqrt{100 \times \frac{1}{2} \times \frac{1}{2}}}\right) \approx \Phi\left(\frac{21}{5}\right) \approx 0,999968.$$

Trong đó $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du.$

Chương V

MỘT SỐ VẤN ĐỀ VỀ THỐNG KÊ TOÁN HỌC

Nội dung chính:

- Hàm phân phối mẫu và các số đặc trưng mẫu.
- Ước lượng điểm và khoảng ước lượng của tham số θ.
- Kiểm định giả thiết về trung bình, xác suất p, so sánh 2 trung bình, so sánh 2 xác suất.
- Tiêu chuẩn χ^2 kiểm định về phân phối đã cho, kiểm định những tính độc lập và tính thuần nhất.
- Hồi quy và tương quan tuyến tính.

Những kiến thức chuẩn bị:

- Khái niệm xác suất, phân phối xác suất, các số đặc trưng.
- Luật số lớn và định lí giới hạn trung tâm.

1. MẪU NGẪU NHIÊN, HÀM PHÂN PHỐI MẪU VÀ CÁC SỐ ĐẶC TRƯNG MẪU

1.1. Mẫu ngẫu nhiên

1.1.1. Khái niệm về mẫu ngẫu nhiên

Giả sử ta cần nghiên cứu một đặc tính nào đó của các phần tử của một tập hợp các đối tượng cùng loại (theo nghĩa các phần tử có một dấu hiệu nào đó về mặt định tính hay định lượng chung). Trên thực tế, do số phần tử của tập hợp đó rất nhiều, ta không thể nghiên cứu tất cả các phần tử của tập hợp đó, nhưng ta lại muốn có một kết luận đủ chính xác

về một đặc tính nào đó cho tất cả những phần tử của tập hợp đã cho. Để giải quyết vấn đề trên ta phải chọn ra một bộ phận gồm các phần tử đại diện của tập hợp đó. Tập hợp các phần tử đại diện đó được gọi là tập mẫu.

Định nghĩa 5.1. Tập hợp mẫu, hay gọi tắt là mẫu, là tập hợp những đối tượng được chọn theo một phân phối xác suất nào đó.

Nếu tập hợp mẫu gồm n phần tử thì n được gọi là kích thước mẫu. Người ta thường kí hiệu mẫu ngẫu nhiên dưới dạng $(X_1, X_2,..., X_n)$. Các X_i là phần tử của mẫu, đôi khi còn gọi là "quan sát" X_i. Về mặt toán học X_i là biến ngẫu nhiên; về mặt thực nghiệm các X_i là kết quả định lượng của phép thử (thí nghiệm). Các giá trị của mẫu được kí hiệu bằng chữ thường $(x_1, x_2,..., x_n)$.

Tập tổng quát là tập những đối tượng mà từ đó ta chọn ra mẫu.

Bằng ngôn ngữ toán học người ta định nghĩa mẫu ngẫu nhiên như sau:

Định nghĩa 5.2. Mẫu ngẫu nhiên là một dãy n biến ngẫu nhiên $(X_1, X_2,..., X_n)$ có phân phối xác xuất $F(x_1, x_2,..., x_n)$

n được gọi là kích thước mẫu.

Về mặt hình học một mẫu $(X_1, X_2,..., X_n)$ được xem như một điểm của không gian n chiều $\mathbf{R}^n$ – Không gian $\mathbf{R}^n$ này được gọi là không gian mẫu.

Chú ý. Thông thường người ta hay xét mẫu ngẫu nhiên là dãy n biến ngẫu nhiên $X_1, X_2,..., X_n$ độc lập và có phân phối F(x) như nhau.

1.1.2. Một số phương pháp chọn mẫu

Tuỳ theo phương pháp thiết lập khác nhau mà ta được các loại mẫu khác nhau. Ta có thể đưa ra 5 loại mẫu thông dụng sau đây; còn các phương pháp khác khi cần thiết độc giả có thể tìm tài liệu viết riêng về các phương pháp đó.

a. Mẫu ngẫu nhiên hoàn lại:

Từ tập hợp tổng quát gồm N phần tử ta chọn ngẫu nhiên 1 phần tử

khảo sát và ghi lại kết quả là X_1 – (Giả sử các phần tử có khả năng chọn như nhau; nghĩa là các xác suất chọn một phần tử là 1/N). Trả phần đó vào tập tổng quát và ta chọn ngẫu nhiên phần tử thứ hai từ tập tổng quát khảo sát và ghi kết quả là X_2. Xác suất chọn được phần tử này vẫn bằng 1/N. Ta lại trả phần tử này vào tập tổng quát. Tiếp tục lặp lại quá trình này đến n lần ta được một mẫu (X_1, X_2,..., X_n). Mẫu này được gọi là mẫu ngẫu nhiên hoàn lại.

b. Mẫu ngẫu nhiên không hoàn lại:

Từ tập hợp tổng quát gồm N phần tử ta chọn ngẫu nhiên một phần tử khảo sát và ghi lại kết quả là X_1. Bỏ phần tử ra ngoài, ta lại chọn ngẫu nhiên phần tử thứ hai từ tập tổng quát khảo sát và ghi lại kết quả X_2. Ta lại bỏ phần tử đó ra ngoài. Tiếp tục quá trình đến lần thứ n ta nhận được một mẫu (X_1, X_2,..., X_n). Mẫu này được gọi là mẫu ngẫu nhiên không hoàn lại.

Hai mẫu ngẫu nhiên hoàn lại và mẫu ngẫu nhiên không hoàn lại được gọi tên chung là mẫu ngẫu nhiên đơn giản.

c. Mẫu được chọn theo phương pháp cơ học:

Trường hợp tập tổng quát có vô hạn phần tử, làm theo hướng phương pháp a), b) rất khó, ta phải tiến hành bằng phương pháp sau đây: Dùng máy để phân tích ngẫu nhiên tập tổng quát thành các tập nhỏ, rồi từ những tập nhỏ đó ta chọn một số phần tử đại diện, rồi hợp nhất lại ta được một mẫu chung.

Ngoài phương pháp đó, ta có thể dùng bảng số ngẫu nhiên để chọn, nghĩa là ta đánh số thứ tự tất cả phần tử của tập tổng quát (thường các số này đã có từ trước). Ta chọn ngẫu nhiên một số trong bảng số ngẫu nhiên sau đó chọn những phần tử của tập tổng quát có số thứ tự trùng với những số ta vừa lấy ra để khảo sát.

Phương pháp chọn mẫu này được gọi là phương pháp cơ học.

d. Mẫu "điển hình" (đặc trưng):

Mẫu điển hình là mẫu mà phần tử của nó không chọn từ toàn bộ tập tổng quát mà từ bộ phận "điển hình" của nó. Tất nhiên từ bộ phận "điển hình" ta chọn mẫu theo cách a) hoặc cách b).

e. Phương pháp phân dãy (series):

Ta chia tập tổng quát thành nhiều dãy. Sau đó từ mỗi dãy ta chọn ra một mẫu con theo cách a) hoặc cách b). Hợp nhất các mẫu con lại ta được mẫu chung.

1.1.3. Sắp xếp số liệu thực nghiệm

Từ mẫu ngẫu nhiên $(X_1, X_2,..., X_n)$ ta thường có hai cách sắp xếp tiện lợi cho việc áp dụng các tiêu chuẩn thống kê.

a. Sắp xếp theo các giá trị khác nhau.

Giả sử mẫu $(X_1, X_2,..., X_n)$ có k quan sát khác nhau là $X_1, X_2,..., X_k$, $k \leq n$ và

X_1 có tần số (số lần xuất hiện) là n_1;

X_2 có tần số là n_2;

.....

X_n có tần số là n_k;

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

Ta có thể viết dưới dạng:

X	X_1	X_2	.....	X_k
Tần số	n_1	n_2	.....	n_k

Ví dụ 5.1. Kiểm tra ngẫu nhiên 50 học sinh bằng 1 bài thi viết. Kết quả được cho dưới bảng sau:

X	2	4	5	6	7	8	9	10
n_i	4	6	20	10	5	2	2	1

b. Sắp xếp dưới dạng khoảng.

Nếu ngẫu nhiên $(X_1, X_2,...., X_n)$ có nhiều quan sát khác nhau, khoảng cách giữa các quan sát không bằng nhau và độ khác nhau rất ít, không thể sắp xếp như ở phần a), ta sẽ sắp xếp dưới dạng khoảng thì việc xử lý sẽ thuận tiện hơn. Ta tiến hành như sau:

Khoảng (x_{min}, x_{max}) chứa toàn bộ các quan sát $X_1, X_2,..., X_n$. Ta chia

khoảng (x_{min}, x_{max}) thành nhiều khoảng nhỏ, độ dài các khoảng đó không nhất thiết phải bằng nhau. Song để tiện tính toán người ta thường chia thành các khoảng có độ dài bằng nhau. Người ta chứng minh được rằng, số khoảng được chọn tối ưu theo công thức (do Sturgen tìm ra):

Số khoảng = 1 + 3,322lgn và độ dài khoảng là:

$$h \approx \frac{x_{max} - x_{min}}{1 + 3.322 \lg n}.$$

Thường chọn mút trái của khoảng đầu tiên bằng $x_{min} - \frac{h}{2} = a_1$ mút kia $a_2 = a_1 + h$, $a_3 = a_2 + h$, $a_3 = a_2 + h$,... tiếp tục làm cho tới lúc mút đầu của khoảng $\leq x_{max}$.

Ví dụ 5.2. Cho mẫu ngẫu nhiên dưới dạng (bảng 5.1)

Bảng 5.1

Khoảng li	Tần số n_i	$\frac{n_i}{h}$
1 – 5	10	2,5
5 – 9	20	5
2 – 13	50	12,5
13 – 17	12	3
17 – 21	8	2

1.2. Hàm phân phối mẫu, đa giác đồ và tổ chức đồ

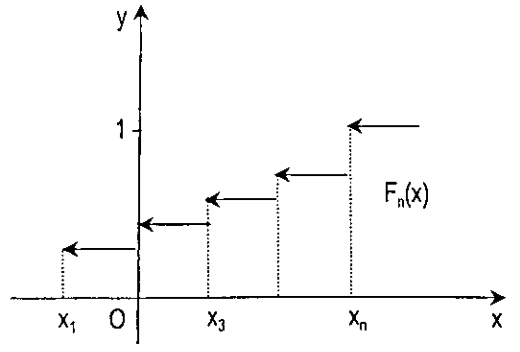
1.2.1. Hàm phân phối mẫu

Giả sử (X_1, X_2,..., X_n) là mẫu ngẫu nhiên độc lập từ phân phối F(x).

Định nghĩa 5.3. Gọi hàm phân phối mẫu (hay hàm phân phối thực nghiệm) là tỉ số $F_n(x) = \frac{m}{n}$, $x \in \mathbf{R}$, trong đó n là kích thước mẫu, m là số các quan sát $X_i < x$.

Ví dụ 5.3. Kiểm tra ngẫu nhiên 20 học sinh trong một lớp. Kết quả ghi bằng điểm như sau:

X	2	3	4	5	6	8
n_i	1	2	2	6	7	2



Hình 5.1

Hàm phân phối mẫu là:

$$F(x) = \frac{m}{n} = \begin{cases} 0 & \text{với } x \le 2 \\ \frac{1}{20} & \text{với } 2 < x \le 3 \\ \frac{3}{20} & \text{với } 3 < x \le 4 \\ \frac{5}{20} & \text{với } 4 < x \le 5 \\ \frac{11}{20} & \text{với } 5 < x \le 6 \\ \frac{18}{20} & \text{với } 6 < x \le 8 \\ 1 & \text{với } x > 8 \end{cases}$$

Tính chất của hàm phân phối mẫu $F_n(x)$:

$0 \le F_n(x) \le 1$ vì $0 \le m \le n$

$F_n(x)$ là hàm đơn điệu tăng.

$F_n(x)$ là hàm liên tục bên trái.

$F_n(x) = 0$ với $x \leq \min(X_1, X_2, ..., X_n)$.

$F_n(x) = 1$ với $x > \max(X_1, X_2, ..., X_n)$.

Người ta chứng minh được rằng:

$F_n(x) \to F(x)$ khi $n \to \infty$ theo nghĩa xác suất (cả theo nghĩa hầu chắc chắn).

Hình ảnh thống kê của hàm phân phối mẫu như hình 5.1.

1.2.2. Đa giác và tổ chức dồ

Mục này chủ yếu dành cho việc dùng đồ thị và biểu đồ để minh hoạ mật độ phân bố của các hiện tượng ngẫu nhiên dựa trên cơ sở mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ đã cho.

a. Đa giác đồ

Giả sử X_1 được lặp lại n_1 lần,

X_2 được lặp lại n_2 lần,

.. ..

X_k được lặp lại n_k lần,

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

Định nghĩa 5.4. Đa giác tần số là đường nối các điểm $(X_1, n_1), (X_2, n_2), ..., (X_k, n_k)$.

Định nghĩa 5.5. Đa giác tần suất là đường nối các điểm:

$$(X_1, \frac{n_1}{n}), (X_2, \frac{n_2}{n}), ..., (X_k, \frac{n_k}{n}).$$

Nó mô tả đồ thị hàm mật độ thực nghiệm.

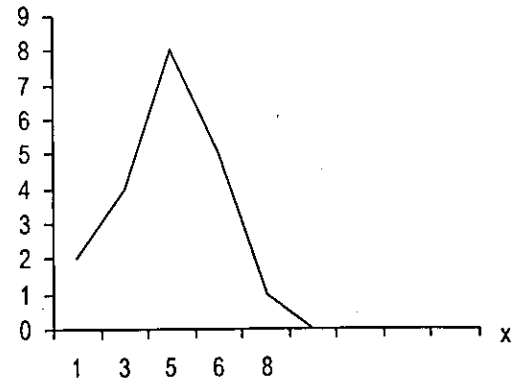
Ví dụ 5.4. Kiểm tra ngẫu nhiên 20 học sinh bằng 1 bài thi môn toán. Kết quả được cho ở bảng sau:

X	1	3	5	6	8
n_i	2	4	8	5	1

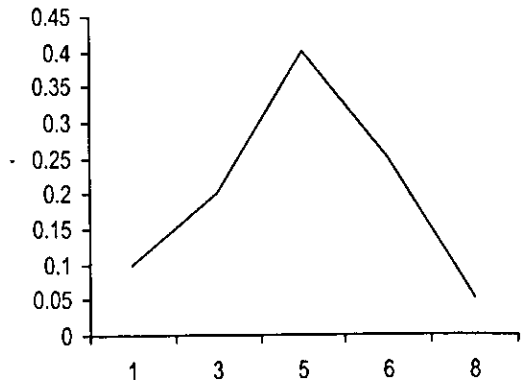
Vẽ đa giác tần số (h.5.2) và đa giác tần suất (h.5.3).

Ta có:

X	1	2	5	6	8
n_i	2	4	8	5	1
$\frac{n_i}{n}$	0,1	0,2	0,4	0,25	0,05



Hình 5.2



Hình 5.3

b. Tổ chức đồ

Dạng biểu đồ này cũng mô tả mật độ phân bố của đại lượng ngẫu nhiên X trên cơ sở mẫu quan sát cho dưới dạng khoảng. Tổ chức đồ tần suất là một hình bậc thang gồm nhiều hình chữ nhật có đáy trùng với

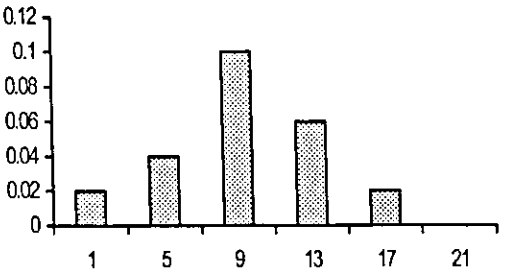
trục hoành, độ dài cạnh đáy hình chữ nhật thứ i là độ dài của khoảng thứ i, còn chiều rộng vuông góc với trục hoành có độ lớn là $\frac{n_i}{nh}$, trong đó h là độ dài khoảng. Diện tích hình chữ nhật thứ i là $\frac{n_i}{nh} \times h = \frac{n_i}{n}$.

Diện tích hình bậc thang là:

$$h \times \frac{n_1}{nh} + h\frac{n_2}{nh} + ... + h\frac{n_k}{nh} = 1.$$

Ví dụ 5.5. Cho mẫu quan sát của đại lượng ngẫu nhiên X dưới dạng bảng 5.2.

STT	Khoảng li h = 4	Tần số n_i	$\frac{n_i}{nh}$
1	1 – 5	40	0,02
2	5 – 9	80	0,04
3	9 – 13	200	0,1
4	13 – 17	120	006
5	17 – 21	50	0,02



Hình 5.4

1.3. Các số đặc trưng mẫu

1.3.1. Trung bình mẫu

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên từ phân phối F(x).

Định nghĩa 5.6. Gọi $\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$ là trung bình mẫu.

Nếu mẫu ngẫu nhiên cho dưới dạng:

X	X_1	X_2	...	X_k
n_i	n_1	n_2	...	n_k

thì trung bình mẫu có dạng

$$\bar{X} = \frac{n_1X_1 + n_2X_2 + ... + n_kX_K}{n_1 + n_2 + ... + n_k}$$

Nếu mẫu ngẫu nhiên cho dưới dạng khoảng thì trung bình mẫu viết dưới dạng $\bar{X} = \frac{X_1^* + X_2^* + ... + X_n^*}{n}$ trong đó $X_i^* = \frac{X_i + X_{i+1}}{2}$

X_i là mút trái của khoảng thứ i, X_{i+1} là mút phải của khoảng đó.

1.3.2. Phương sai mẫu

Có hai công thức ước lượng của phương sai DX:

$$S_n^2(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\left(\sum_{n-1}^{n}X_i^2 - \frac{(\sum_{i=1}^{n}X_i)^2}{n}\right).$$

Và $S_n^{*2}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ $(= \frac{1}{n-1}\sum_{i=1}^{k}n_i(X_i - \bar{X})^2$, nếu mẫu có k giá trị khác nhau $X_1,..., X_k)$.

Ví dụ 5.6. Cho mẫu quan sát của đại lượng ngẫu nhiên X là:

X_i	1	2	3	4
n_i	20	15	10	5

Tính $\bar{X}$, $S_n^2(X)$, $S_n^{*2}(X)$.

Giải:

$$\bar{X} = \frac{20\times1 + 2x15 + 10\times3 + 5\times4}{20+15+10+5} = \frac{100}{50} = 2;$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{20\times 1^2 + 15\times 2^2 + 10\times 3^2 + 5\times 4^2}{50} = \frac{250}{50} = 5;$$

$$S_n^2(X) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2 = 5 - 2^2 = 1;$$

$$S_n^{*2}(X) = \frac{n}{n-1} S_n^2(X) = \frac{50}{49}\times 1 = \frac{50}{49}.$$

1.3.3. Hệ số tương quan mẫu

Cho mẫu quan sát đối với cặp biến ngẫu nhiên (X, Y) là (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n). Hệ số tương quan mẫu của (X, Y) được xác định theo công thức:

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2][n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2]}}$$

Ví dụ 5.7. Cho mẫu ngẫu nhiên đối với cặp biến ngẫu nhiên (X, Y) là:

X	2	3	4	5	6	7	8	9
Y	3	7	8	9	13	15	16	17

Tính hệ số tương quan mẫu r.

Giải:

Từ số liệu trên ta tính được:

$$\sum X_i = 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 = 44;$$

$$\sum Y_i = 3 + 7 + 8 + 9 + 13 + 15 + 16 + 17 = 88;$$

$$\sum X_i^2 = 4 + 9 + 16 + 25 + 36 + 49 + 64 + 81 = 284;$$

$$\sum Y_i^2 = 1142;$$

$$\sum X_i Y_i = 568,\ n = 8.$$

$$r = \frac{8\times 568 - 44\times 88}{\sqrt{[8\times 284 - 44^2][8\times 1142 - 88^2]}} = 0,98.$$

Ví dụ 5.8. Cho mẫu quan sát đối với cặp biến ngẫu nhiên (X, Y) là:

Y \ X	1	2	3
2	10	2	
3	1	8	
4		2	7

Tính hệ số tương quan mẫu của (X, Y).

Giải:

Để tính r trước hết ta lập bảng công phụ để tính tổng $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum Y_i^2$, $\sum X_i Y_i$.

X_i	n_i	$n_i X_i$	$n_i X_i^2$
1	11	11	11
2	12	24	48
3	7	21	93
	30 = n	56 = $\sum X_i$	122 = $\sum X_i^2$

Y_i	n_i	$n_i Y_i$	$n_i Y_i^2$
2	12	11	11
3	9	24	48
4	7	21	93
	30 = n	87 = $\sum Y_i$	273 = $\sum Y_i^2$

X_i	Y_i	n_i	$n_i X_i Y_i$
			20
1	2	10	3
1	3	1	8
2	2	2	48
2	3	8	16
2	4	2	84
3	4	7	179 = $\sum X_i Y_i$

$$n = \frac{30 \times 179 - 56 \times 87}{\sqrt{\left[30 \times 122 - 56^2\right]\left[30 \times 273 - 87^2\right]}} \approx 0,87.$$

1.3.4. Mô men mẫu

a) Mô men gốc mẫu bậc k của biến ngẫu nhiên X được xác định như sau:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

b) Mô men trung tâm mẫu bậc k được xác định bởi đẳng thức:

$$a_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k.$$

1.3.5. Trung vị mẫu

Nếu mẫu ngẫu nhiên cho dưới dạng $(X_1, X_2, ..., X_n)$ thì ta sắp xếp các quan sát X_i, i = 1, 2,..., n theo thứ tự tăng dần:

$$X^{(1)} < X^{(2)} < ... < X^{(q)} < X^{(q+1)} < ... < X^{(n)}.$$

Nếu n là số chẵn, nghĩa là n = 2q thì trung vị bằng:

$$X_{Me} = \frac{X^{(q)} + X(q+1)}{2}.$$

Nếu n là số lẻ, nghĩa là n = 2q – 1 thì số trung vị bằng:

$$X_{Me} = X^{(q)}.$$

Ví dụ 5.9. Cho mẫu quan sát đối với đại lượng ngẫu nhiên X là:

X	1	2	3	4	5	6
n_i	4	6	22	16	36	16

Ta thấy n = 100 là số chẵn. Vậy $X_{Me} = \frac{X(50) + X(51)}{2}$.

Nhìn vào mẫu trên ta thấy có 48 giá trị X_i nhỏ hơn hoặc bằng 4 và 84 giá trị $X_i \le 5$. Vậy từ quan sát thứ 49 trở đi đến 84 có giá trị bằng 5. Do đó $X^{(50)} = X^{(51)} = 5$.

Vậy $X_{Me} = \frac{5+5}{2} = 5.$

• Nếu mẫu quan sát được cho dưới dạng khoảng thì số trung vị được tính theo công thức:

$$X_{Me} = A_{Me} + h\frac{\frac{n}{2} - m_{Me}}{n_{Me}}.$$

trong đó: A_{Me} là đầu mút trái của khoảng trung vị; n_{Me} là số lần xuất hiện khoảng trung vị; m_{Me} là số lần xuất hiện các khoảng trước khoảng trung vị.

Ví dụ 5.10. Cho mẫu quan sát dưới dạng khoảng (Bảng 5.2)

Năng suất hàng năm tính ra % ở dạng khoảng	Số lượng công nhân n_i	Tần suất n_i/n
80 – 90	8	8/117
90 – 100	15	15/117
100 – 110	46	46/117
110 – 120	29	29/117
120 – 130	13	13/117
130 – 140	3	3/117
140 – 150	3	3/117

Ta có h = 10; $A_{Me} = 100$; $n_{me} = 46$; $m_{Me} = 23$; n = 117.

$$\text{Vậy } X_{Me} = 100 + 10 \times \frac{\frac{117}{2} - 23}{46} = 107,8.$$

1.3.6. Mốt mẫu (mod)

Công thức tính mốt (mod) mẫu trong trường hợp mẫu cho dưới dạng khoảng là:

$$X_{mod} = A_{m_0} + h \times \frac{n_{m_0} - n_{m_0-1}}{2n_{m_0} - n_{m_0-1} - n_{m_0+1}}$$

trong đó A_{m_0} là mút trái của khoảng mốt; n_{m_0} là số lần xuất hiện khoảng mốt; n_{m_0-1} số lần xuất hiện của khoảng trước. Khoảng mốt; n_{m_0+1} là số lần xuất hiện của khoảng sau khoảng mốt.

Từ số liệu ở bảng 5.2 ta có:

$$n_{m_0} = 46;\ n_{m_0-1} = 15;\ n_{m_0+1} = 29;\ h = 10.$$

$$A_{m_0} = 100.$$

Vậy $X_{mod} = 100 + 10 \times \dfrac{46-15}{2 \times 46 - 15 - 29} = 106,2.$

• Số liệu cho dưới dạng khác thì giá trị x_{mod} chính là giá trị của x_i mà tần suất xuất hiện giá trị đó lớn nhất.

Ví dụ 5.11. Cho mẫu

X	1	2	3
$\frac{n_i}{n}$	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Ta thấy $x_{mod} = 2$.

2. ƯỚC LƯỢNG THAM SỐ

2.1. Ước lượng điểm

2.1.1. Các khái niệm về ước lượng

Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên từ phân phối $f(x, \theta)$, $\theta \in U$ (θ là tham số).

Trên cơ sở mẫu $(X_1, X_2,..., X_n)$ đã cho ta cần ước lượng tham số θ (hoặc hàm số của tham số θ).

Định nghĩa 5.7. Ước lượng điểm của tham số θ (hoặc hàm số tham số $\tau(\theta)$) là đại lượng ngẫu nhiên $T_n = \varphi(x_1, x_2,..., x_n)$ chỉ phụ thuộc các quan sát $x_1,..., x_n$ và không phụ thuộc tham số θ.

Ví dụ 5.12. Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối chuẩn dạng tổng quát $N(a; \sigma^2)$.

$\overline{X} = \dfrac{X_1 + X_2 ... + X_n}{n}$ là ước lượng điểm của kỳ vọng a.

$S_n^2(X) = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ là ước lượng điểm của phương sai σ^2.

$S_n^{*2}(X) = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ cũng là ước lượng điểm của σ^2.

Trong các loại ước lượng điểm ta thường quan tâm đến hai loại ước lượng sau đây: ước lượng không chệch và ước lượng vững.

Định nghĩa 5.8. Ước lượng T_n của tham số θ (hàm tham số $\tau(\theta)$) được gọi là ước lượng không chệch nếu

$$ET_n = \theta \qquad (ET_n = \tau(\theta)).$$

Trong ví dụ 5.12, $\overline{X}$ là ước lượng không chệch của a vì

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\sum_{i=1}^{n}EX_i = \frac{na}{n} = a.$$

$S_n^{*2}(X) = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ là ước lượng không lệch của σ^2.

Vì:

$$\begin{aligned} ES_n^{*2}(X) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\ &= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}E\left(X_i^2\right) - nE\left(\overline{X}^2\right)\right] \end{aligned}$$

mà $E(X_i^2) = \sigma^2 + a^2$

và $E\left(\overline{X}^2\right)=\frac{1}{n^2}\left(\sum_{i=1}^{n}E\left(X_i^2\right)+\sum_{i\neq j}EX_iEX_j\right)=\frac{\sigma^2+a^2}{n}+\frac{n-1}{n}a^2.$

nên $E\left(S_n^{*2}(X)\right)=\frac{1}{n-1}\left[n\left(\sigma^2+a^2\right)-\frac{n}{n}\left(\sigma^2+a^2+(n-1)a^2\right)\right]$

$$=\frac{n-1}{n-1}.\sigma^2=\sigma^2.$$

$S_n^2(X)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ không là ước lượng không lệch của σ^2.

Thật vậy:

$$E\left(S_n^2(X)\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)=\frac{n-1}{n}E\left(S_n^{*2}(X)\right)=\frac{n-1}{n}\sigma^2\neq\sigma^2.$$

Định nghĩa 5.9. Ước lượng T_n được gọi là ước lượng không chệch với phương sai bé nhất của hàm tham số $\tau(\theta)$ nếu:

1) $E(T_n) = \tau(\theta)$;

2) $D_\theta(T_n) \leq D_\theta(V_n)$, trong đó V_n là ước lượng không chệch bất kỳ của $\tau(\theta)$.

$D_\theta X$ là ký hiệu phương sai của X khi θ đã cho.

Định nghĩa 5.10. Ước lượng $\varphi_n(X_1,..., X_n)$ của tham số θ được gọi là ước lượng vững nếu với $\varepsilon > 0$ cho trước tùy ý ta có:

$$\lim_{n\to\infty}P\left[\left|\varphi_n-\theta\right|<\varepsilon\right]=1.$$

Trong ví dụ 5.12, $\overline{X}$ là ước lượng vững của kỳ vọng a.

Vì dãy $X_1, X_2,..., X_n$ độc lập có phương sai $DX_K = \sigma^2$ (bị chặn ở σ^2) theo định lí Chebyshev ta có:

$$\frac{X_1+X_2+...+X_n}{n}\xrightarrow{P}a \text{ khi } n\to\infty$$

Ngoài ra $S_n^2(X)$, $S_n^{*2}(X)$ cũng là ước lượng vững của σ^2.

2.1.2. Phương pháp hợp lý cực đại

Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối $f(x, \theta)$; $\theta \in U$.

Định nghĩa 5.11. Hàm hợp lý là hàm có dạng:

$$(X; \theta) = f(X_1, \theta)f(X_2, \theta)... f(X_n, \theta).$$

Định nghĩa 5.12. $\hat{\theta} = \hat{\theta}(X_1, X_2,..., X_n)$ được gọi là ước lượng hợp lý cực đại của tham số θ nếu

$$L(X; \hat{\theta}) \geq L(X; \theta), \forall\theta \in U.$$

Từ định nghĩa trên ta rút ra phương pháp tìm ước lượng như sau:

Tìm giá trị $\hat{\theta}$ sao cho hàm $L(X, \theta)$ đạt cực đại tại $\theta = \hat{\theta}$. Ta có thể sử dụng đạo hàm như sau:

1) Trường hợp θ là một số

$$\frac{\partial L(X; \theta)}{\partial\theta} = 0 \tag{5.1}$$

Giải phương trình (5.1) ta tìm được $\hat{\theta}$. Sau đó ta xét dấu của đạo hàm hạng nhất hoặc đạo hàm hạng hai của L theo θ để tìm cực đại của L. (Có thể không phải làm điều này vì người ta chứng minh được rằng với một số giả thiết nhất định, nghiệm của phương trình (5.1) làm cực đại hàm hợp lý).

Nếu $f(x, \theta) > 0$ thì $L(X; \theta) > 0$. Vậy (5.1) tương đương với phương trình

$$\frac{1}{L}.\frac{\partial L}{\partial\theta} = 0 \Leftrightarrow \frac{\partial \ln L}{\partial\theta} = 0.$$

Vì $\ln L(X; \theta) = \sum_{i=1}^{n} \ln f(X_i, \theta)$ nên (5.1) tương đương với

$$\sum_{i=1}^{n} \frac{\partial \ln f(X_i, \theta)}{\partial\theta} = 0. \tag{5.2}$$

Ví dụ 5.13. Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối xác suất với hàm một độ dạng

$$f(x,\theta)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}}, & x>0,\ \theta>0\\ 0 & x\le 0\end{cases}$$

Tìm ước lượng hợp lý cực đại của θ.

Giải:

Ta có thể trực tiếp áp dụng phương trình (5.2). Trước hết ta tính

$$\ln f(X_i,\theta)=\ln\left(\frac{1}{\theta}e^{-\frac{x_i}{\theta}}\right)=-\frac{X_i}{\theta}-\ln\theta;$$

$$\frac{\partial \ln f(X;\theta)}{\partial\theta}=\frac{X_i}{\theta^2}-\frac{1}{\theta}.$$

Thay biểu thức này vào phương trình (5.2) ta có

$$\sum_{i=1}^{n}\left(\frac{X_i}{\theta^2}-\frac{1}{\theta}\right)=0.$$

Từ đó suy ra: $\sum_{i=1}^{n}(X_i-\theta)=0 \Rightarrow n\theta=\sum_{i=1}^{n}X_i \Rightarrow \hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}X_i.$

Ta có thể xét dấu đạo hàm hạng hai của lnL theo θ:

$$\frac{\sigma^2 \ln L}{\partial\theta^2}=\sum_{i=1}^{n}\left(-\frac{X_i}{\theta^3}+\frac{1}{\theta^2}\right).$$

Thay $\theta=\frac{1}{n}\sum_{i=1}^{n}X_i$ ta được

$$\left.\frac{\partial^2 \ln L}{\partial\theta}\right|_{\theta=\frac{1}{n}\sum_{i=1}^{n}X_i}=-\frac{n^3}{\left(\sum_{i=1}^{n}X_i\right)^2}<0.$$

Vậy lnL đạt cực đại tại $\theta=\overline{X}$. Do đó L(X; θ) cũng đạt cực đại tại $\theta=\overline{X}$ tức $\hat{\theta}=\overline{X}$ là ước lượng hợp lý cực đại của θ.

2) Trường hợp $\theta = (\theta_1, \theta_2, ..., \theta_r)$

Làm hoàn toàn tương tự như trường hợp 1.

Ta nhận được hệ phương trình

$$\begin{cases} \dfrac{\partial L}{\partial \theta_1} = 0 \\ .... \\ \dfrac{\partial L}{\partial \theta_r} = 0 \end{cases} \tag{5.3}$$

Giải hệ phương trình (5.3) ta được $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_r)$.

Nếu $f(x, \theta) > 0$ thì $L(X; \theta) > 0$.

Làm tương tự như trường hợp 1 ta nhận được hệ phương trình

$$\begin{cases} \sum\limits_{i=1}^{n} \dfrac{\partial \ln f(X_i, \theta)}{\partial \theta_1} = 0 \\ .... \\ \sum\limits_{i=1}^{n} \dfrac{\partial \ln f(X_i, \theta)}{\partial \theta_r} = 0 \end{cases} \tag{5.4}$$

Ví dụ 5.14. Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối chuẩn dạng tổng quát $N(a; \sigma^2)$.

Tìm ước lượng hợp lý cực đại của (a, σ^2).

Giải:

Trước khi áp dụng hệ phương trình (5.4) ta cần tính:

$$\ln f(X_i; a; \sigma^2) = \ln \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i - a)^2}{2\sigma^2}}$$

$$= -\frac{(X_i - a)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2;$$

$$\frac{\partial \ln f\left(X_i; a, \sigma^2\right)}{\partial a} = \frac{\left(X_i - a\right)}{\sigma^2};$$

$$\frac{\partial \ln f}{\partial\left(\sigma^2\right)} = \frac{\left(X_i - a\right)^2}{2\left(\sigma^2\right)^2} - \frac{1}{2\sigma^2}.$$

Thay các biểu thức này vào hệ phương trình (5.4) ta có

$$\begin{cases} \sum\limits_{i=1}^{n} \dfrac{\left(X_i - a\right)}{\sigma^2} = 0 \\ \sum\limits_{i=1}^{n} \left(\dfrac{\left(X_i - a\right)^2}{2\left(\sigma^2\right)^2} - \dfrac{1}{2\sigma^2} \right) = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{a} = \dfrac{1}{n} \sum\limits_{i=1}^{n} X_i = \overline{X} \\ \hat{\sigma}^2 = \dfrac{1}{n} \sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \end{cases}.$$

Đó là ước lượng phải tìm.

2.2. Ước lượng khoảng

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên từ phân phối $f(x, \theta)$, $\theta \in U$. Ngoài việc tìm ước lượng điểm của θ, ta còn phương pháp tìm ước lượng của θ bằng cách xác định một khoảng mà chứa tham số θ.

Định nghĩa 5.13. Khoảng $(\hat{\theta}_1(X), \hat{\theta}_2(X))$, $(\hat{\theta}_1 < \hat{\theta}_2)$ được gọi là khoảng ước lượng của tham số θ với độ tin cậy $1 - \alpha$ nếu

$$P[\hat{\theta}_1 < \theta < \hat{\theta}_2] = 1 - \alpha$$

Dựa vào định nghĩa này và định lí giới hạn trung tâm ta có thể tìm được khoảng ước lượng của các tham số trong một số phân phối thông dụng.

2.2.1. Khoảng ước lượng của xác suất p trong phân phối nhị thức

Dựa vào định nghĩa khoảng ước lượng và định lí Moivre–Lapcace ta tìm được khoảng ước lượng của xác suất p với độ tin cậy $1 - \alpha$ là:

$$\bar{p} - x_\alpha \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} < p < \bar{p} + x_\alpha \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \qquad (5.5)$$

trong đó x_α tra ở bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 1 - \frac{\alpha}{2}$,

$\hat{p} = \frac{k}{n}$, k là số lần xuất hiện biến cố A trong n phép thử.

Ví dụ 5.15. Gieo 400 hạt giống đậu tương thấy có 5 hạt không nẩy mầm. Hãy tìm khoảng ước lượng của xác suất không nảy mầm p của mỗi hạt với độ tin cậy 0,999.

Giải:

Ta có $1 - \alpha = 0{,}999$. Suy ra $\alpha = 0{,}001$. Vậy $F(x_\alpha) = 1 - \frac{\alpha}{2} = 0{,}9995$.

Tra bảng phân phối chuẩn ta có $x_\alpha = 3{,}3$.

n = 400, k = 5 thay vào công thức của khoảng ước lượng ta được

$$\frac{5}{400} - \frac{3{,}3}{20}\sqrt{\frac{5}{400}\left(1 - \frac{5}{400}\right)} < p < \frac{5}{400} + \frac{3{,}3}{20}\sqrt{\frac{5}{400}\left(1 - \frac{5}{400}\right)}.$$

Rút gọn ta có: $-0{,}0058 < p < 0{,}0308$.

Vì xác suất không âm nên: $0 \le p < 0{,}0308$.

2.2.2. Khoảng ước lượng của kỳ vọng a trong phân phối chuẩn

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối chuẩn $N(a; \sigma^2)$.

Dựa vào định nghĩa khoảng ước lượng và định lí giới hạn trung tâm ta tìm ra khoảng ước lượng của a với độ tin cậy $1 - \alpha$ là:

1. σ đã biết:

$$\bar{X} - x_\alpha \frac{\sigma}{\sqrt{n}} < a < \bar{X} + x_\alpha \frac{\sigma}{\sqrt{n}} \qquad (5.6)$$

x_α tra ở bảng phân phối chuẩn N(0; 1) sao cho: $F(x_\alpha) = 1 - \frac{\alpha}{2}$.

2. σ chưa biết:

$$\bar{X} - t_\alpha \frac{S_n^*(X)}{\sqrt{n}} < a < \bar{X} + t_\alpha \frac{S_n^*(X)}{\sqrt{n}} \qquad (5.7)$$

trong đó: $S_n^{*2}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Nếu n > 30 thì t_α tra ở bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 1 - \frac{\alpha}{2}$.

Nếu n < 30 thì t_α tra ở bảng phân phối Student với n – 1 bậc tự do và mức ý nghĩa α (Bảng tiêu chuẩn 2 phía).

Ví dụ 5.16. Tìm khoảng ước của kỳ vọng a với độ tin cậy 0,95 trong mẫu từ phân phối chuẩn N(a; σ^2), $\sigma^2 = 25$. Cho $\overline{X} = 14$, n = 25.

Giải:

Ở đây $1 - \alpha = 0,95$, ta suy ra $\alpha = 0,05$.

Tra bảng giá trị của hàm phân phối chuẩn N(0; 1) ta được $x_\alpha = 1,96$. Khoảng ước lượng của a với độ tin cậy 0,95 là:

$$14 - 1,96.\frac{5}{\sqrt{25}} < a < 14 + 1,96.\frac{5}{\sqrt{25}}$$

$$\Leftrightarrow 12,04 < a < 15,96.$$

Ví dụ 5.17. Tiến hành kiểm tra ngẫu nhiên 10 học sinh. Kết quả là:

X: 5 4 3 5 6 7 6 2 8 5

Giả sử các quan sát này có phân phối chuẩn dạng tổng quát N(a; σ^2).

Tính khoảng ước lượng của a với độ tin cậy 0,95.

Giải:

Ta suy ra từ số liệu trên các kết quả sau:

$\overline{X} = 5,1$;

$$S_n^{*2} = \frac{1}{9}\Big[3(5-5,1)^2 + (4-5,1)^2 + (3-5,1)^2 + (2-5,1)^2 + \Big(2(6-5,1)^2 + (7-5,1)^2 + (8-5,1)^2\Big)\Big] \approx 3,21.$$

Tra bảng phân phối Student ta tìm được t(5%; 9) = 2,26.

Khoảng ước lượng của a với độ tin cậy 0,95 là

$$5,1-2,26.\sqrt{\frac{3,21}{10}}<a<5,1+2,26.\sqrt{\frac{3,21}{10}}$$

$$\Leftrightarrow 3,71<a<9,39.$$

2.2.3. Khoảng ước lượng của phương sai σ^2 trong mẫu từ phân phối chuẩn

Khoảng ước lượng của phương sai σ^2 với độ tin cậy $1-\alpha$ là

$$\frac{(n-1)S_n^{*2}(X)}{t_2}<\sigma^2<\frac{(n-1)S_n^{*2}(X)}{t_1} \qquad (5.8)$$

trong đó $S_n^{*2}(X)=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, t_1, t_2 tra từ bảng phân phối khi bình phương với $n-1$ bậc tự do sao cho $P[\chi^2>t_1]=1-\frac{\alpha}{2}$ và $P[\chi^2>t_2]=\frac{\alpha}{2}$.

Với số liệu trong ví dụ 5.17, khoảng ước lượng của phương sai với độ tin cậy 0,90 là: $t_1(95\%; 9) = 3,325$, $t_2(5\%, 9) = 16,919$.

$$\frac{9\times3,12}{16,919}<\sigma^2<\frac{9\times3,21}{3,325}$$

$$\Leftrightarrow 1,5<\sigma^2<8,6.$$

3. KIỂM ĐỊNH GIẢ THIẾT

3.1. Thiết lập bài toán

a. Giả thiết thống kê

Giả sử biến ngẫu nhiên X có phân phối F(x) (hoặc $F(x, \theta)$, $\theta \in U$).

Những giả thiết về phân phối F(x) được gọi là giả thiết thống kê và ký hiệu là H_0.

Những giả thiết cũng về phân phối F(x) nhưng khác với giả thiết H_0 được gọi là đối thiết và ký hiệu là K.

Khi phân phối F(x, θ) phụ thuộc vào tham số thì những giả thiết về phân phối F(x, θ) được chuyển sang giả thiết về tham số θ.

Ví dụ 5.18. Giả sử biến ngẫu nhiên X có phân phối chuẩn dạng N(a; 0, 01) – trường hợp này giả thiết về phân phối chuẩn được chuyển sang giả thiết về tham số a.

a) H_0: $a = 5$ với K: $a \neq 5$;

b) H_0: $a = 5$ hoặc $a = 10$ với K: $5 < a < 10$;

c) H_0: $a \leq 5$ với K: $a > 5$.

Nếu tập giả thiết H_0 có một phần tử thì H_0 được gọi là giả thiết đơn, nếu tập giả thiết H_0 có lớn hơn hoặc bằng 2 phần tử thì H_0 được gọi là giả thiết hợp. Tương tự như vậy đối với đối thiết K nếu tập K có một phân tử thì K được gọi là đối thiết đơn. Nếu tập K có lớn hơn hoặc bằng 2 phần tử thì k được gọi là đối thiết hợp.

b. Kiểm định giả thiết thống kê

Kiểm định giả thiết thông kê là việc chọn một trong hai quyết định: chấp nhận giả thiết H_0 hoặc bác bỏ giả thiết H_0.

c. Tiêu chuẩn kiểm định giả thiết

Để có được quyết định chấp nhận hoặc bác bỏ giả thiết H_0 ta phải dựa trên một tiêu chuẩn nào đó. Tiêu chuẩn kiểm định giả thiết được hiểu như sau:

Tiêu chuẩn kiểm định giả thiết là đại lượng ngẫu nhiên Z phụ thuộc vào các quan sát $(X_1, X_2,..., X_n)$, nghĩa là Z xác định trên không gian mẫu R^n, nhờ nó ta có thể kiểm định được giả thiết.

Vì Z xác định trên không gian mẫu R^n nên có một bộ phận của không gian mẫu mà khi $(X_1, X_2,..., X_n)$ rơi vào miền đó thì ta bác bỏ giả thiết H_0. Bộ phận này được gọi là miền tiêu chuẩn, ký hiệu là W. Nói cách khác, miền tiêu chuẩn là miền bác bỏ giả thiết.

Đặt $X = (X_1,..., X_n)$. Tiêu chuẩn kiểm định giả thiết thể hiện ở một trong ba dạng:

$$[X \in W] \Leftrightarrow [Z(X) > C_n] \text{ bác bỏ giả thiết } H_0;$$

$$[X \in W] \Leftrightarrow [Z(X) < C_v] \text{ bác bỏ giả thiết } H_0;$$

(C_n, C_v được gọi là điểm tiêu chuẩn).

$$[X \in W] \Leftrightarrow [Z(X) > C_n] \text{ hoặc } [Z(X) < C_v] \text{ bác bỏ giả thiết } H_0.$$

Hai tiêu chuẩn đầu được gọi là tiêu chuẩn một phía. Tiêu chuẩn thứ 3 được gọi là tiêu chuẩn hai phía. (Tiêu chuẩn thứ 3 tổng quát hơn hai tiêu chuẩn đầu).

Muốn tìm tiêu chuẩn Z(X) hoặc miền tiêu chuẩn W ta dựa trên hai loại sai lầm sau:

Sai lầm loại I: Giả thiết H_0 là giả thiết đúng mà bác bỏ thì mắc sai lầm. Sai lầm đó được gọi là sai lầm loại I. Xác suất mắc sai lầm loại I là $P[W/H_0$ là đúng].

Sai lầm loại II: Giả thiết H_0 là giả thiết sai mà chấp nhận giả thiết thì cũng mắc sai lầm. Sai lầm này được gọi là sai lầm loại II. Xác suất mắc sai lầm loại II là $P[R^n - W/H_0$ là sai].

Để tìm tiêu chuẩn kiểm định giả thiết ta phải đồng thời hạn chế tới mức tối thiểu khả năng mắc sai lầm loại I và sai lầm loại II. Song việc làm này rất khó khăn. Trên thực tế người ta thường cho phép được mắc sai lầm loại I ở mức xác suất α nào đó (tùy theo tầm quan trọng của sai lầm loại I) sau đó cực tiểu hóa xác suất mắc sai lầm loại II.

Dưới đây ta sẽ xét một số bài toán kiểm định giả thiết mà có thể ứng dụng nhiều trong các lĩnh vực sinh vật, nông nghiệp, giáo dục, v.v...

3.2. Kiểm định xác suất p trong phân phối nhị thức (tỉ lệ phần trăm)

Bài toán: Giả sử trong dãy n phép thử Bernoulli biến cố A xuất hiện X lần. Gọi $p = p(A)$ là xác suất để A xuất hiện trong mỗi phép thử.

Kiểm định giả thiết $H_0 : p = p_0$ với $K : p \neq p_0$ ở mức α.

Tiêu chuẩn kiểm định giả thiết này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$|Z| = \frac{|X - np_0|}{\sqrt{np_0(1-p_0)}} > x_\alpha. \qquad (5.9)$$

Còn $|Z| < x_\alpha$ thì chấp nhận giả thiết H_0; trong đó x_α tra ở bảng giá trị của hàm phân phối chuẩn sao cho $F(x_\alpha) = 1 - \frac{\alpha}{2}$.

Ví dụ 5.19. Gieo 300 hạt cà phê giống. Kết quả là 261 hạt nảy mầm. Người ta nói rằng tỉ lệ nảy mầm của cà phê là 0,90. Điều nhận định đó có đúng không? Tại sao? Cho mức kiểm định $\alpha = 5\%$.

Giải:

Ta xem việc gieo 300 hạt cà phê như tiến hành 300 phép thử Bernoulli. p là xác suất nảy mầm của mỗi hạt cà phê. Ta kiểm định giả thiết:

$H_0: p = 0,90$ với $K: p \neq 0,90$ (ở đây $X = 261$, $p_0 = 0,90$).

Theo (5.9) ta có: $|Z| = \frac{|261 - 300 \times 0,90|}{\sqrt{300 \times 0,90 \times 0,10}} \approx 1,73.$

Theo giả thiết $\alpha = 0,05$. Ta có $F(x_\alpha) = 1 - \frac{\alpha}{2} = 0,975$. Tra bảng phân phối chuẩn N(0; 1) ta tìm được $x_\alpha = 1,96$.

Ta nhận thấy $|Z| = 1,73 < 1,96$.

Vậy ta chấp nhận giả thiết H_0, nghĩa là tỉ lệ nảy mầm của cà phê là 0,90.

Tiêu chuẩn một phía:

- Nếu $\frac{X}{n} > p_0$ thì ta chuyển về kiểm định giả thiết:

$H_0: p = p_0$ với $K: p > p_0$ ở mức α.

Tiêu chuẩn kiểm định giả thiết này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu: $Z = \dfrac{X - np_0}{\sqrt{np_0(1-p_0)}} > x_\alpha$.

(Nếu $Z < x_\alpha$ thì chấp nhận giả thiết H_0), trong đó x_α tra ở bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 1 - \alpha$.

• Nếu $\dfrac{X}{n} < p_0$ thì ta chuyển về kiểm định giả thiết:

H_0 : $p = p_0$ với K: $p < p_0$ ở mức α.

Tiêu chuẩn kiểm định giả thiết này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu: $Z = \dfrac{np_0 - X}{\sqrt{np_0(1-p_0)}} > x_\alpha$.

(nếu $Z < x_\alpha$ thì chấp nhận giả thiết H_0), trong đó x_α tra ở bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 1 - \alpha$.

Ví dụ 5.20. Trong một lô sản phẩm người ta cho biết tỉ lệ phế phẩm là 0,02. Ta chọn ngẫu nhiên có hoàn lại 480 sản phẩm thấy có 42 phế phẩm. Xét xem tỉ lệ phế phẩm được người ta thông báo có đúng không? Cho mức kiểm định $\alpha = 5\%$.

Giải:

Ở đây $n = 480$, $X = 42$, $\dfrac{X}{n} = \dfrac{42}{480} = 0,875 > 0,02 = p_0$.

Vậy ta đi đến kiểm định giả thiết:

H_0: $p = 0,02$ với K: $p > 0,02$.

$F(x_\alpha) = 1 - \alpha = 0,95$.

Tra bảng phân phối chuẩn N(0; 1) ta tìm được $x_\alpha = 1,65$.

Tính Z: $Z = \dfrac{X - np_0}{\sqrt{np_0(1-p_0)}} = \dfrac{42 - 480 \times 0,02}{\sqrt{480 \times 0,02 \times 0,98}} = \dfrac{32,4}{3,06} = 10,588.$

Ta suy ra $Z > x_\alpha = 1,65$. Vậy giả thiết H_0 bị bác bỏ và ta chấp nhận đối thiết

K : $p > 0,02$, tức tỉ lệ phế phẩm mà được thông báo là không đúng.

3.3. Kiểm định về kì vọng (trung bình) trong mẫu độc lập từ phân phối chuẩn

Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối chuẩn $N(a; \sigma^2)$.

Kiểm định giả thiết $H_0: a = a_0$ với $K: a \neq a_0$ ở mức α.

Tiêu chuẩn kiểm định giả thiết này là:

a. Trường hợp σ đã cho:

Giả thiết H_0 bị bác bỏ ở mức α nếu: $|Z| = \dfrac{|\bar{X} - a_0|\sqrt{n}}{\sigma} > x_\alpha$. (5.10)

(Nếu $|Z| < x_\alpha$ thì chấp nhận H_0) trong đó x_α tra ở bảng phân phối chuẩn $N(0; 1)$ sao cho $F(x_\alpha) = 1 - \dfrac{\alpha}{2}$.

b. Trường hợp σ chưa biết:

Giả thiết H_0 bị bác bỏ ở mức α nếu: $|Z| = \dfrac{|\bar{X} - a_0|\sqrt{n}}{S_n^*(X)} > t_\alpha$. (5.11)

(Nếu $|Z| < t_\alpha$ thì chấp nhận H_0) trong đó t_α tra như sau:

- Nếu $n > 30$ thì t_α tra ở bảng phân phối chuẩn $N(0; 1)$ sao cho:

$$F(t_\alpha) = 1 - \frac{\alpha}{2}.$$

- Nếu $n < 30$ thì t_α tra ở bảng phân phối Student với $n - 1$ bậc tự do và mức ý nghĩa α (bảng tiêu chuẩn 2 phía).

Ví dụ 5.21. Giả sử mẫu ngẫu nhiên $(X_1, X_2,.., X_n)$ từ phân phối chuẩn $N(a; 4)$ và $\bar{X} = 15$; $n = 100$. Hãy kiểm định giả thiết H_0: $a = 16,5$ với K: $a \neq 16,5$ ở mức $\alpha = 5\%$.

Giải:

$\alpha = 0,05$; $F(x_\alpha) = 1 - \dfrac{\alpha}{2} = 0,975$. Tra bảng phân phối chuẩn ta có $x_\alpha = 1,96$.

Tính: $|Z| = \dfrac{|\bar{X} - a_0|\sqrt{n}}{\sigma} = \dfrac{|15 - 16,5|\sqrt{100}}{2} = 7,5.$

Ta thấy $|Z| = 7,5 > x_\alpha = 1,96$. Vậy giả thiết H_0 bị bác bỏ ở mức α, nghĩa là $a \neq 16,5$.

Ví dụ 5.22. Sau một đợt huấn luyện người ta kiểm tra ngẫu nhiên 15 học sinh với kết quả sau:

X : 2	3	7	6	9	7	6	6	1	2	7	8	6	5	6

Giả sử X tuân theo phân phối chuẩn dạng $N(a; \sigma^2)$. Hãy kiểm định giả thiết: $H: a = 6$ với $K: \neq 6$ ở mức $\alpha = 5\%$.

Giải:

Trường hợp này, σ chưa biết, $n = 15 < 30$.

t_α tra ở bảng phân phối Student với 14 bậc tự do ta có $t(5\%; 14) = 2,14$.

$$\bar{X} = \frac{81}{15} = 5,4.$$

$$S_n^{*2}(X) = \frac{1}{14}\Big(2(2-5,4)^2 + (3-5,4)^2 + 3(7-5,4)^2 + 5(6-5,4)^2 +$$

$$+(9-5,4)^2 + (5-5,4)^2 + (1-5,4)^2 + (8-5,4)^2\Big)$$

$$= \frac{77,2}{14} = 5,5.$$

$$\Rightarrow S_n^*(X) = \sqrt{5,5} \approx 2,35.$$

Tính: $|Z| = \dfrac{(5,4-6)\sqrt{15}}{2,35} \approx 0,987.$

Vậy $|Z| < 0,987$. Ta chấp nhận giả thiết H_0, nghĩa là $a = 6$.

Tiêu chuẩn một phía:

• Nếu $\bar{X} > a_0$ thì đưa về kiểm định giả thiết:

$H_0 : a = a_0$ với $K : a > a_0$ ở mức α

Tiêu chuẩn kiểm định giả thiết này là:

a. Trường hợp σ đã biết:

Giả thiết H_0 bị bác bỏ ở mức α nếu: $Z = \frac{(\bar{X} - a_0)\sqrt{n}}{\sigma} > x_\alpha$, (còn $Z < x_\alpha$ thì chấp nhận H_0).

Trong đó, x_α tra ở bảng phân phối chuẩn sao cho: $F(x_\alpha) = 1 - \alpha$.

b. Trường hợp σ chưa biết:

Giả thiết H_0 bị bác bỏ ở mức α nếu: $Z = \frac{(\bar{X} - a_0)\sqrt{n}}{S_n^*(X)} > t_\alpha$ (còn $Z < t_\alpha$ thì chấp nhận H_0).

Trong đó: $S_n^{*2}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Nếu $n > 30$ thì t_α tra ở bảng phân phối chuẩn như ở a).

Nếu $n < 30$ thì t_α tra ở bảng phân phối Student với $n - 1$ bậc tự do và với mức α. (Bảng tiêu chuẩn một phía).

Nếu $\bar{X} < a_0$ ta đưa về bài toán kiểm định giả thiết:

$H_0 : a = a_0$ với $K : a < a_0$ ở mức α.

Tiêu chuẩn kiểm định giả thiết này là:

a. Trường hợp σ đã biết:

Giả thiết H_0 bị bác bỏ ở mức α nếu: $Z = \frac{(a_0 - \bar{X})\sqrt{n}}{\sigma} > x_\alpha$, (còn $Z < x_\alpha$ thì chấp nhận H_0).

Trong đó x_α tra ở bảng phân phối chuẩn N(0; 1) sao cho: $F(x_\alpha) = 1 - \alpha$.

b. Trường hợp σ chưa biết:

Giả thiết H_0 bị bác bỏ ở mức α nếu: $Z = \frac{(a_0 - \bar{X})\sqrt{n}}{S_n^*(X)} > t_\alpha$ (còn $Z < t_\alpha$ thì chấp nhận H_0).

Nếu n > 30 thì t_α tra ở bảng phân phối chuẩn sao cho: $F(x_\alpha) = 1 - \alpha$.

Nếu n < 30 thì t_α tra ở bảng phân phối Student với n – 1 bậc tự do và với mức α. (Bảng tiêu chuẩn một phía).

3.4. So sánh hai trung bình của hai mẫu độc lập có phân phối chuẩn dạng tổng quát

Giả sử $(X_1,..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối chuẩn $N(a_1; \sigma_1^2)$.

Và $(Y_1,..., Y_m)$ là mẫu ngẫu nhiên độc lập từ phân phối chuẩn $N(a_2; \sigma_2^2)$.

Hãy kiểm định giả thiết:

$H_0 : a_1 = a_2$ với $K : a_1 \neq a_2$ ở mức α

Tiêu chuẩn kiểm định giả thiết này là:

a. Trường hợp σ_1; σ_2 đã biết:

Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$|Z| = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > x_\alpha \qquad (5.12)$$

(nếu $|Z| < x_\alpha$ thì chấp nhận H_0), trong đó x_α tra ở bảng phân phối chuẩn $N(0; 1)$ sao cho $F(x_\alpha) = 1 - \frac{\alpha}{2}$.

b. Trường hợp σ_1; σ_2 chưa biết:

Ta phải giả thiết $\sigma_1^2 = \sigma_2^2$.

Giả thiết H_0 bị bác bỏ ở mức α nếu $|Z| = \dfrac{|\bar{X} - \bar{Y}|}{S\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}} > t_\alpha$ (5.13)

(nếu $|Z| < t_\alpha$ thì chấp nhận H_0).

Trong đó: $S^2 = \dfrac{1}{n+m-2}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2\right]$, t_α được tra như sau:

Nếu $n + m > 60$ thì t_α được tra ở bảng phân phối chuẩn như x_α như ở a.

Nếu $n + m < 60$ thì t_α được tra ở bảng phân phối Student với $n + m - 2$ bậc tự do và mức ý nghĩa α. (Bảng tiêu chuẩn 2 phía).

Ví dụ 5.23. Giả sử $(X_1,..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối chuẩn $N(a_1; 80)$ và $(Y_1,..., Y_m)$ là mẫu ngẫu nhiên độc lập từ phân phối chuẩn $N(a_2; 100)$. Giả sử $\bar{X} = 130$; $\bar{Y} = 140$. Hãy kiểm định giả thiết: H_0: $a_1 = a_2$ với K : $a_1 \neq a_2$ ở mức $\alpha = 0,01$.

Giải:

$F(x_\alpha) = 1 - \dfrac{\alpha}{2} = 0,995$. Ta nhận được $x_\alpha = 2,58$. Tính giá trị của $|Z|$:

$$|Z| = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}} = \frac{|130 - 140|}{\sqrt{\dfrac{80}{40} + \dfrac{100}{50}}} = 5.$$

Ta thấy $|Z| = 5 > x_\alpha = 2,58$. Vậy bác bỏ giả thiết H_0, nghĩa là trung bình của hai mẫu khác nhau.

Tiêu chuẩn một phía:

• Nếu $\bar{X} > \bar{Y}$ thì ta chọn K : $a_1 > a_2$ và ta đưa đến bài toán:

Kiểm định giả thiết H_0: $a_1 = a_2$ với K : $a_1 > a_2$ ở mức α.

a. Trường hợp σ_1; σ_2 đã biết:

Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > x_\alpha \qquad (5.14)$$

(nếu $Z < x_\alpha$ thì chấp nhận H_0), trong đó x_α tra ở bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 1 - \alpha$.

b. Trường hợp σ_1; σ_2 chưa biết:

Ta phải giả thiết $\sigma_1^2 = \sigma_2^2$.

Giả thiết H_0 bị bác bỏ ở mức α nếu $Z = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} > t_\alpha$ (5.15)

(nếu $Z < t_\alpha$ thì chấp nhận H_0).

t_α được tra như sau:

Nếu $n + m > 60$ thì t_α được tra ở bảng phân phối chuẩn sao cho: $F(t_\alpha) = 1 - \alpha$.

Nếu $n + m < 60$ thì t_α được tra ở bảng phân phối Student với $n + m - 2$ bậc tự do và mức ý nghĩa α. (Bảng tiêu chuẩn 1 phía).

• Nếu $\bar{X} < \bar{Y}$ thì ta chọn K : $a_1 < a_2$ và ta đưa đến bài toán:

Kiểm định giả thiết H_0: $a_1 = a_2$ với K : $a_1 < a_2$ ở mức α.

a. Trường hợp σ_1; σ_2 đã biết:

Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$Z = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > x_\alpha \qquad (5.16)$$

(nếu $Z < x_\alpha$ thì chấp nhận H_0), trong đó x_α tra ở bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 1 - \alpha$.

b. Trường hợp σ_1; σ_2 chưa biết:

Ta phải giả thiết $\sigma_1^2 = \sigma_2^2$.

Giả thiết H_0 bị bác bỏ ở mức α nếu $Z = \dfrac{\bar{Y} - \bar{X}}{S\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}} > t_\alpha$ (5.17)

(nếu $Z < t_\alpha$ thì chấp nhận H_0).

t_α được tra như trường hợp $\bar{X} > \bar{Y}$.

Chú ý. Khi nhận được kết quả thí nghiệm hai mẫu $(X_1, X_2,..., X_n)$ và $(Y_1, Y_2,..., Y_m)$ từ phân phối chuẩn nhưng không biết gì về phương sai DX, DY. Để so sánh được sự bằng nhau của hai trung bình EX và EY bằng tiêu chuẩn trên, ta phải giải bài toán kiểm định giả thiết về sự bằng nhau của hai phương sai DX và DY. Nếu kiểm tra thấy DX = DY thì ta mới áp dụng tiêu chuẩn trên.

Bây giờ ta kiểm định giả thiết:

H_0 : DX = DY với DX ≠ DY ở mức α

Tiêu chuẩn kiểm định giả thiết này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$Z = \frac{S_n^{*2}(X)}{S_m^{*2}(Y)} > f_{bảng}\left(\frac{\alpha}{2},\ n-1,\ m-1\right) \qquad (5.18)$$

Nếu $Z < f_{bảng}$ thì chấp nhận giả thiết H_0.

Trong đó $f_{bảng}\left(\dfrac{\alpha}{2};\ n-1;\ m-1\right)$ tra ở bảng phân phối F với n – 1, m – 1 bậc tự do và mức α.

Ví dụ 5.24. Để đánh giá chất lượng sản phẩm do 2 nhà máy sản xuất, người ta kiểm tra ngẫu nhiên 10 sản phẩm của nhà máy I và 12 sản phẩm của nhà máy II. Kết quả đo kích thước của các sản phẩm là:

Ở nhà máy I:

X_i	3,4	3,5	3,7	3,9
n_i	2	3	4	1

Ở nhà máy II:

Y_i	3,2	3,4	3,6
M_i	2	2	8

Giả thiết các quan sát X_i, Y_i độc lập và có phân phối chuẩn có $EX = a_1$; $EY = a_2$. Hãy kiểm định giả thiết $H_0 : a_1 = a_2$ với $K : a_1 \neq a_2$ ở mức $\alpha = 0,02$.

Giải:

Từ giả thiết trên ta tính được $\overline{X}$ và $\overline{Y}$ như sau:

$$\overline{X} = \frac{2\times 3,4 + 3\times 3,5 + 4\times 3,7 + 3,9}{10} = 3,6$$

và $$\overline{Y} = \frac{3,2\times 2 + 3,4\times 2 + 3,6\times 8}{12} = 3,5.$$

Phương sai mẫu: $S_n^{*2}(X) = 0,0267$; $S_m^{*2}(Y) = 0,0255$.

Tính Z:

$$Z = \frac{S_n^{*2}(X)}{S_m^{*2}(Y)} = \frac{0,0267}{0,0255} = 1,05.$$

Tra bảng phân phối F ta tìm được $f_{bảng}(0,01; 9; 11) = 4,63$.

Ta thấy $Z = 1,05 < f_{bảng} = 4,63$. Vậy giả thiết $H_0 : DX = DY$ được chấp nhận.

Bây giờ trở về bài toán so sánh hai trung bình EX, EY. Tính:

$$Z = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$= \frac{\overline{X} - \overline{Y}}{\sqrt{(n-1)S_n^{*2}(X) + (m-1)S_n^{*2}(Y)}} \times \sqrt{\frac{nm(n+m-2)}{n+m}}$$

$$= \frac{3,6 - 3,5}{\sqrt{9\times 0,0267 + 11\times 0,0255}} \times \sqrt{\frac{10\times 12(10+12-2)}{10+12}} \approx 1,45.$$

Tra bảng phân phối Student ta tìm được $t_\alpha = 2,53$.

Vì: $|Z| = 1,45 < t_\alpha = 2,53$ nên giả thiết H_0 về sự bằng nhau của hai trung bình được chấp nhận.

3.5. So sánh hai xác suất trong phân phối nhị thức

Bài toán: Xét hai dãy phép thử Bernoulli. Dãy I gồm n phép thử, X là số lần xuất hiện biến cố A trong dãy I; $p_1 = p(A)$ là xác suất để A xuất hiện trong mỗi phép thử trong dãy I. Dãy II gồm m phép thử, Y là số lần xuất hiện A trong dãy II; $p_2 = p(A)$ là xác suất để A xuất hiện trong mỗi phép thử trong dãy II. Hãy so sánh hai xác suất p_1 và p_2 ở mức α.

Để trả lời câu hỏi này, ta đưa về bài toán kiểm định giả thiết:

$H_0 : p_1 = p_2$ với $K : p_1 \neq p_2$ ở mức α.

Tiêu chuẩn kiểm định giả thiết này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$|Z| = \frac{\left|\frac{X}{n} - \frac{Y}{m}\right|}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\left(\frac{X+Y}{n+m}\right)\left(1 - \frac{X+Y}{n+m}\right)}} > x_\alpha. \qquad (5.19)$$

Nếu $|Z| < x_\alpha$ thì chấp nhận giả thiết H_0, trong đó x_α tra ở bảng phân phối chuẩn sao cho $F(x_\alpha) = 1 - \frac{\alpha}{2}$.

Tiêu chuẩn một phía:

• Nếu $\frac{X}{n} < \frac{Y}{m}$ thì ta đưa về bài toán: Kiểm định giả thiết:

$H_0 : p_1 = p_2$ với $K: p_1 < p_2$ ở mức α

Tiêu chuẩn kiểm định giả thiết H_0 này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$Z = \frac{\frac{Y}{m} - \frac{X}{n}}{\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\left(\frac{X+Y}{n+m}\right)\left(1-\frac{X+Y}{n+m}\right)}} > x_\alpha. \qquad (5.20)$$

Nếu $Z < x_\alpha$ thì chấp nhận giả thiết H_0, trong đó x_α tra ở bảng phân phối chuẩn sao cho $F(x_\alpha) = 1 - \alpha$.

• Nếu $\frac{X}{n} > \frac{Y}{m}$ thì ta đưa về bài toán: Kiểm định giả thiết:

H_0: $p_1 = p_2$ với K: $p_1 > p_2$ ở mức α

Tiêu chuẩn kiểm định giả thiết H_0 này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$Z = \frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\left(\frac{X+Y}{n+m}\right)\left(1-\frac{X+Y}{n+m}\right)}} > x_\alpha \qquad (5.21)$$

Nếu $Z < x_\alpha$ thì chấp nhận giả thiết H_0, trong đó x_α tra ở bảng phân phối chuẩn sao cho $F(x_\alpha) = 1 - \alpha$.

Ví dụ 5.25. Có 2 phương pháp gieo hạt. Theo phương pháp I, gieo 100 hạt thấy có 80 hạt nảy mầm. Theo phương pháp II, gieo 125 hạt thấy có 90 hạt nảy mầm. Hãy so sánh hiệu quả của hai phương pháp trên ở mức $\alpha = 5\%$.

Giải:

Ở đây thấy n = 100, X = 80, m = 125, Y = 90, α = 0,05

$F(x_\alpha) = 1 - \frac{\alpha}{2} = 0{,}975.$

Tra ở bảng phân phối chuẩn N(0; 1) ta tìm được $x_\alpha = 1{,}96$.

Ta có:

$$|Z| = \frac{\left|\frac{80}{100} - \frac{90}{125}\right|}{\sqrt{\left(\frac{1}{100}+\frac{1}{125}\right)\left(\frac{80+90}{100+125}\right)\left(1-\frac{80+90}{100+125}\right)}} = 1,387.$$

Ta nhận thấy $|Z| < x_\alpha$. Vậy chấp nhận giả thiết H_0, nghĩa là hiệu quả của hai phương pháp tương đương.

Ví dụ 5.26. Để đánh giá chất lượng sản phẩm do 2 nhà máy sản xuất người ta kiểm tra ngẫu nhiên 100 sản phẩm của nhà máy I thấy có 20 phế phẩm và 150 sản phẩm của nhà máy II thấy có 15 phế phẩm. Hãy so sánh chất lượng sản phẩm của hai nhà máy ở mức $\alpha = 0,05$.

Giải:

Ở đây ta thấy $n = 100, X = 20, m = 150, Y = 15$.

Từ đó suy ra $\frac{X}{n} = \frac{20}{100} = 0,2$; $\frac{Y}{m} = \frac{15}{150} = 0,1$.

Ta thấy $\frac{X}{n} > \frac{Y}{m}$. Từ đó đi đến kiểm định giả thiết:

$H_0: p_1 = p_2$ với $K: p_1 > p_2$ ở mức $\alpha = 0,05$.

Suy ra $F(x_\alpha) = 1 - \alpha = 1 - 0,05 = 0,95$.

Tra bảng phân phối chuẩn ta có $x_\alpha = 1,65$.

Tính:

$$Z = \frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\left(\frac{X+Y}{n+m}\right)\left(1-\frac{X+Y}{n+m}\right)}},$$

$$= \frac{0,2-0,1}{\sqrt{\left(\frac{1}{100}+\frac{1}{150}\right)\left(\frac{20+15}{100+150}\right)\left(1-\frac{20+15}{100+150}\right)}}$$

$\approx 2,232.$

Vì $Z > x_\alpha = 1,65$ nên giả thiết $H_0 : p_1 = p_2$ bị bác bỏ, nghĩa là tỉ lệ phế phẩm của nhà máy I cao hơn tỉ lệ phế phẩm của nhà máy II.

3.6. Tiêu chuẩn Mann – Whitney (1947) kiểm định tính thuần nhất của hai mẫu độc lập

Tiêu chuẩn này kiểm định tính thuần nhất của hai mẫu độc lập $(X_1, X_2,.., X_n)$ và $(Y_1, Y_2,..., Y_m)$ với giả thiết X, Y độc lập và có phân phối liên tục. Tiêu chuẩn này thường được sử dụng trong lĩnh vực nghiên cứu tâm lí, xã hội để so sánh tính trạng của một tính chất của các thành viên của hai mẫu hoặc trong lĩnh vực khoa học giáo dục chẳng hạn, nghiên cứu mức độ phản ứng của học sinh sau khi được học thêm một đơn vị kiến thức hoặc chịu tác động một phương pháp dạy học nào đó.

Bài toán đặt ra như sau: Giả sử X có hàm phân phối $F_1(x)$, Y có hàm phân phối $F_2(x)$ và đều là phân phối liên tục.

Hãy kiểm định giả thiết:

$H_0 : F_1(x) = F_2(x)$ với $K : F_1(x) \neq F_2(x)$ ở mức α.

Có nhiều nhà toán học đã giải bài toán này. Ở đây, chúng tôi chỉ nêu lời giải của Mann – Whitney (1947) trong trường hợp mẫu có kích thước $n > 4$, $m > 4$ và $n + m > 20$.

Để kiểm định giả thiết H_0 với đối thiết K, ta tiến hành theo các bước sau đây:

Bước 1: Dồn 2 mẫu vào một mẫu chung và sắp xếp các quan sát X_i và Y_i theo thứ tự tăng dần.

Giả sử $n < m$ (nếu $n > m$ thì đổi vai trò của X_i cho Y_i). Gọi W là tổng các chỉ số thứ tự của các quan sát X_i trong mẫu I (nếu $n > m$ thì xem mẫu $(Y_1, Y_2,..., Y_m)$ là mẫu I, còn $(X_1, X_2,..., X_n)$ là mẫu II).

Người ta tính được $E(W) = \frac{n(n+m+1)}{2}$; $DW = \frac{nm(n+m+1)}{12}$.

Đặt $U = nm + \frac{n(n+1)}{2} - W$ ta có $EU = \frac{mn}{2}$; $DU = \frac{nm(n+m+1)}{12}$.

Bước 2: Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$|Z| = \frac{\left|U - \frac{nm}{2}\right|}{\sqrt{\frac{nm(n+m+1)}{12}}} > x_\alpha \qquad (5.22)$$

trong đó x_α tra ở bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 1 - \frac{\alpha}{2}$.

Chú ý. Nếu trong mẫu có một số giá trị mẫu trùng nhau, để đảm bảo độ chính xác, người ta hiệu chỉnh lại phương sai DU (hoặc DW) bằng cách trừ đi 1 lượng $\frac{\sum K}{n+m-1}$; $K = \frac{l^3 - l}{12}$, l là số các giá trị trùng nhau cùng một đoạn; $\sum K$ là tổng số trong các đoạn khác nhau. Cụ thể trong biểu thức của $|Z|$ biểu thức căn bậc hai trừ đi một lượng $\frac{\sum K}{n+m-1}$.

Những giá trị trùng nhau được gán cùng một hạng (số thứ tự) bằng số trung bình các hạng (số thứ tự) của chúng.

Ví dụ 5.27. Để đánh giá hiệu quả của việc cải tiến phương pháp giảng dạy người ta tiến hành dạy thí điểm trên 2 lớp A và B. Lớp A được dạy theo phương pháp mới, lớp B được dạy theo phương pháp cũ. Sau khi dạy hết một phần chương trình người ta kiểm tra ngẫu nhiên 18 em ở lớp A và 20 em ở lớp B bằng một bài thi. Kết quả thu được như sau:

Lớp A: 14, 13, 11, 10, 12, 9, 15, 14, 16, 18, 17, 14, 12, 13, 16, 15, 17, 13.

Lớp B: 7, 8, 10, 11, 13, 6, 8, 7, 15, 16, 6, 11, 12, 17, 13, 15, 12, 14, 11, 10.

Hãy so sánh hai phương pháp ở mức $\alpha = 0{,}01$.

Giải:

Ta dồn 2 mẫu thành một mẫu chung và sắp xếp theo thứ tự tăng dần:

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
6	6	7	7	8	8	9	10	10	10	11	11	11	11

(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)
12	12	12	12	13	13	13	13	13	14	14	14	14

(28)	(29)	(30)	(31)	(32)	(33)	(34)	(35)	(36)	(37)	(38)
15	15	15	15	16	16	16	17	17	17	18

Ta thấy:

6 có hạng 1,5; 7 có hạng $\frac{3+4}{2} = 3,5$; 8 có hạng $\frac{5+6}{2} = 5,5$; 9 có hạng 7; 10 có hạng $\frac{8+9+10}{3} = 9$; 11 có hạng $\frac{11+12+13+14}{4} = 12,5$; 12 có hạng $\frac{15+16+17+18}{4} = 16,5$; 13 có hạng $\frac{19+20+21+22+23}{5} = 21$; 14 có hạng $\frac{24+25+26+27}{4} = 25,5$; 15 có hạng $\frac{28+29+30+31}{4} = 29,5$; 16 có hạng $\frac{32+33+34}{3} = 33$; 17 có hạng $\frac{35+36+37}{3} = 36$; 18 có hạng 38.

Ta có W = 7 + 9 + 12,5 + 16,5 + 16,5 + 21 + 21 + 21 + 25,5 + 25,5 + 25,5 + 29,5 + 29,5 + 33 + 33 + 36 + 36 + 38 = 436; n = 18; m = 20 (n < m).

$$U = 18 \times 20 + \frac{18 \times 19}{2} - 436 = -57,5;$$

$$k_1 = k_2 = k_3 = \frac{2^3 - 2}{12} = 0,5;\ k_4 = k_{10} = k_{11} = \frac{3^3 - 3}{12} = 2;$$

$$k_5 = k_6 = k_8 = k_9 = \frac{4^3 - 4}{12} = 5;\ k_7 = \frac{5^3 - 5}{12} = 10;$$

$$\sum_{i=1}^{11} k_i = 1,5 + 6 + 20 + 10 = 37,5;$$

$$\sqrt{\frac{nm(n+m+1)}{12} - \frac{\sum K}{n+m-1}} = \sqrt{\frac{18 \times 20(18+20+1)}{12} - \frac{37,5}{37}}$$

$$\approx 34,19;$$

$$|Z| = \frac{\left|U - \frac{mn}{2}\right|}{\sqrt{\frac{nm(n+m+1)}{12} - \frac{\sum k}{n+m+1}}} = \frac{237,5}{34,19} \approx 6,946;$$

$$\alpha = 0,01;\ F(x_\alpha) = 1 - \frac{\alpha}{2} = 1 - 0,005 = 0,995.$$

Tra bảng phân phối chuẩn N(0, 1) ta tìm được $x_\alpha \approx 2,33$.

Vì $|Z| = 6,946 > x_\alpha = 2,33$ nên giả thiết H_0 bị bác bỏ, nghĩa là trình độ học lực của hai lớp là không tương đương. Lớp A có nhiều điểm trội hơn điểm của lớp B nên trình độ chung của học sinh lớp A tốt hơn trình độ chung của học sinh lớp B.

3.7. Tiêu chuẩn Wilcoxon kiểm định tính thuần nhất của hai mẫu phụ thuộc

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối $F_1(x)$ và $(Y_1, Y_2, ..., Y_n)$ là mẫu ngẫu nhiên độc lập từ phân phối $F_2(x)$.

Giả sử X, Y là phụ thuộc và có phân phối liên tục và phân phối của nó có tính chất đối xứng, nghĩa là:

$$P[X < c - x] = P[X \geq c + x].$$

Ví dụ. Biến ngẫu nhiên X đặc trưng cho tính trạng của tính chất nào đấy của một tập các đối tượng trong lần đo thứ nhất, Y là đặc trưng cũng tính trạng của tính chất đó ở tập này trong lần đo thứ 2.

Đặt $D_i = Y_i - X_i$. Các cặp (X_i, Y_i) mà $X_i = Y_i$ thì không được kể tới. Gọi m là số cặp (X_i, Y_i) mà $X_i \neq Y_i$ $(m < n)$. Ta sắp xếp các $|D_i|$, $i = \overline{1,m}$ theo thứ tự tăng dần theo độ lớn. $|D_i|$ nhỏ nhất được gán hạng 1, hạng 2 cho số sau đó,... và số cuối cùng là hạng m. Nếu có một số $|D_i|$ liên tiếp bằng nhau thì những số đó được gán cùng một hạng là số trung bình các hạng của chúng.

Sau đó mỗi hạng được mang dấu "+" nếu nó ứng với $D_i > 0$ và dấu "–" nếu nó ứng với $D_i < 0$.

Kiểm định giả thiết:

$H_0 : P[D_i > 0] = P[D_i < 0]$ (tức là số trung vị bằng 0) ở mức α.

$K : P[D_i > 0] \neq P[D_i < 0]$.

Gọi R_i là hạng của D_i và đặt $T = \sum_{D_i > 0} R_i$ (T là tổng của các hạng của $D_i > 0$).

Ta có thể viết bài toán dưới dạng sau:

Kiểm định giả thiết:

H_0 : số trung vị của $D_i = 0$;

K : số trung vị của $D_i \neq 0$ ở mức α

a. Trường hợp m < 20

Tiêu chuẩn kiểm định giả thiết như sau:

Giả thiết H_0 bị bác bỏ ở mức α nếu:

$$T < W_{\frac{\alpha}{2}} \text{ hoặc } T > W_{1-\frac{\alpha}{2}} \tag{5.23}$$

trong đó $W_{\frac{\alpha}{2}}$, $W_{1-\frac{\alpha}{2}}$, được tra từ bảng tiêu chuẩn Wilcoxon.

– Nếu X_i có xu thế trội hơn Y_i thì đối thiết K được chọn như sau:

K : số trung vị của $D_i < 0$,

Giả thiết H_0 bị bác bỏ ở mức α nếu $T < W_\alpha$. (5.24)

– Nếu Y_i có xu thế trội hơn X_i thì đối thiết K được chọn như sau:

K : số trung vị của $D_i > 0$,

Giả thiết H_0 bị bác bỏ ở mức α nếu $T > W_{1-\alpha}$. (5.25)

b. Trường hợp m > 20

Bảng giá trị giới hạn của thống kê tiêu chuẩn của Wilcoxon dựa trên phân phối nhị thức. Vì thế với m đủ lớn ta có thể thay xấp xỉ bằng phân phối chuẩn.

• Tiêu chuẩn 2 phía:

Kiểm định giả thiết:

H_0 : số trung vị của $D_i = 0$; K : số trung vị của $D_i \neq 0$.

Ta chọn $W_\alpha = \frac{m(m+1)}{2} + x_{\frac{\alpha}{2}} \sqrt{\frac{m(m+1)(2m+1)}{24}}$ (5.26)

và $W_{1-\alpha} = \frac{m(m+1)}{2} - W_\alpha$, trong đó $x_{\frac{\alpha}{2}}$ tra ở bảng phân phối chuẩn N(0; 1) sao cho: $F(x_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.

Giả thiết H_0 bị bác bỏ ở mức α nếu $T < W_\alpha$ hoặc $T > W_{1-\alpha}$.

• Tiêu chuẩn 1 phía:

Kiểm định giả thiết: H_0 : số trung vị của $D_i \leq 0$; K : số trung vị của $D_i > 0$.

Tiêu chuẩn kiểm định giả thiết này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu: $T > W_{1-\alpha}$ (5.27)

Nếu X_i có xu thế trội hơn Y_i.

Bài toán kiểm định giả thiết:

H_0: số trung vị của $D_i \geq 0$; K: số trung vị của $D_i < 0$.

Tiêu chuẩn kiểm định giả thiết này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu $T < W_\alpha$. (5.28)

Ví dụ 5.28. Để đánh giá việc tuyên truyền vận động sinh đẻ có kế hoạch, người ta phỏng vấn ngẫu nhiên 12 người bằng 1 câu hỏi. Sau đó mở một đợt huấn luyện và người ta lại kiểm tra lần 2 cũng trên 12 người đó. Kết quả cho theo thang điểm bậc 10 như sau:

STT	1	2	3	4	5	6	7	8	9	10	11	12
Điểm kiểm tra đợt I	3	8	5	4	2	1	6	7	1	3	4	2
Điểm kiểm tra đợt II	4	8	6	2	6	3	6	10	6	10	7	8
Hiệu số điểm	1	0	1	–2	4	2	0	3	5	7	3	6
Hạng R_i	1,5		1,5	–3,5	7	3,5		5,5	8	10	5,5	9

Giả thiết H_0 : "đợt tuyên truyền không làm thay đổi nhận thức của con người về kế hoạch hóa gia đình".

Ta có: $T = 1,5 + 1,5 + 7 + 3,5 + 5,5 + 8 + 10 + 5,5 + 9 = 51,5$; $m = 10$; $\alpha = 0,05$.

Tra bảng Wilcoxon ta tìm được $W_{1-\alpha} = 44$, $W_\alpha = 11$.

Vì $T > W_{1-\alpha}$ nên giả thiết H_0 bị bác bỏ ở mức $\alpha = 0,05$, nghĩa là việc tuyên truyền có ý nghĩa làm thay đổi nhận thức về kế hoạch hoá gia đình.

Ví dụ 5.29. Vào đầu năm, giáo viên kiểm tra trình độ ngữ pháp của học sinh bằng một bài Chính tả, sau đó giáo viên luyện tập cho học sinh một thời gian. Hết đợt luyện tập, giáo viên tổ chức kiểm tra lần II.

Kết quả 2 lần kiểm tra như sau:

STT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Số lỗi lần I	7	8	6	4	2	5	3	10	8	7	6	10	3	1	5	8	7	6	4	10	9	0	5	7	8	3	5
Số lỗi lần II	0	2	2	5	4	3	1	4	5	2	4	5	3	4	1	0	4	4	5	6	7	3	6	2	2	4	9

Trên cơ sở số liệu trên, việc luyện tập cho học sinh của giáo viên có hiệu quả không? Cho mức kiểm định $\alpha = 0,05$.

Giải:

STT	1	2	3	4	5	6	7	8	9	10	11	12	13
Số lỗi lần I	7	8	6	4	2	5	3	10	8	7	6	10	3
Số lỗi lần II	0	2	2	5	4	3	1	4	5	2	4	5	3
Hiệu số	–7	–6	–4	1	2	–2	–2	–6	–3	–5	–2	–5	0
Hạng R_i	–25	–23	–16,5	2,5	7,5	–7,5	–7,5	–23	–12,5	–20	–7,5	–20	

STT	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Số lỗi lần I	1	5	8	7	6	4	10	9	0	5	7	8	3	5
Số lỗi lần II	4	1	0	4	4	5	6	7	3	6	2	2	4	9
Hiệu số	3	–4	–8	–3	–2	1	–4	–2	3	1	–5	–6	1	4
Hạng R_i	12,5	–16,5	–26	–12,5	–7,5	2,5	–16,5	–7,5	12,5	2,5	–20	–23	2,5	16,5

Kiểm định giả thiết H_0:

Số trung vị $D_i \geq 0$; K : Số trung vị $D_i < 0$.

Ta có:

$T = 2,5 + 7,5 + 12,5 + 2,5 + 12,5 + 2,5 + 2,5 + 16,5 = 59$; $m = 27 - 1 = 26$, $\alpha = 0,05$; $F(x_\alpha) = 1 - \alpha = 1 - 0,05 = 0,95$.

Tra bảng phân phối chuẩn ta tìm được $x_\alpha = 1,64$.

$$W_\alpha = \frac{26(26+1)}{4} - 1,64\sqrt{\frac{26(26+1)(52+1)}{24}} = 110,93.$$

Vậy $T < W_\alpha$. Giả thiết H_0 bị bác bỏ, nghĩa là sự luyện tập của giáo viên có hiệu quả tốt.

3.8. Tiêu chuẩn χ^2 (khi bình phương) kiểm định về phân phối đã cho

Trong mục 5.3.2, ta đã nghiên cứu bài toán kiểm định về xác suất trong phân phối nhị thức. Bây giờ ta xét vấn đề tổng quát hơn là kiểm định về phân phối p_i, i = 1,..., s, mà các phân phối xác suất p_i là những con số hoặc chúng phụ thuộc vào các tham số θ (ví dụ như phân phối chuẩn $N(a, \sigma^2)$, phân phối Poisson, phân phối đa thức,...).

Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên từ phân phối f(x) (hoặc $f(x, \theta)$). Xét khoảng (a; b) trên trục số sao cho tất cả các quan sát X_1, $X_2,..., X_n$ đều thuộc khoảng (a; b). Chia khoảng (a; b) thành s khoảng không nhất thiết phải bằng nhau: $l_1, l_2,..., l_s$. Giả sử rằng quan sát X_i rơi vào khoảng l_k, k = 1, 2,..., s với xác suất p_k.

Gọi n_k là số các quan sát X_i rơi vào khoảng l_k.

Hãy kiểm định giả thiết:

$H_0 : p_1 = p_1^0,..., p_s = p_s^0$ ở mức α; $K : p_1 \neq p_1^0,..., p_s \neq p_s^0$.

Trong đó p_1^0, $p_2^0,..., p_s^0$ có thể là các con số hoặc những phân phối đã cho, chẳng hạn phân phối nhị thức, phân phối chuẩn,....

Để kiểm định giả thiết H_0 ở trên ta xét 2 trường hợp:

a. Trường hợp các p_i, i = 1,.., s là những con số:

Tiêu chuẩn kiểm định giả thiết này là:

Giả thiết H_0 bị bác bỏ ở mức α nếu $Z = \sum_{i=1}^{s} \frac{(n_i - np_i^0)^2}{np_i^0} > C_\alpha$ (5.29)

(Nếu $Z < C_\alpha$ thì chấp nhận giả thiết H_0).

Trong đó C_α tra ở bảng phân phối χ^2 với s – 1 bậc tự do và mức α.

b. Trường hợp các $p_i = p_i(\theta)$, i = 1, 2,..., s, phụ thuộc tham số θ với $\theta = (\theta_1, \theta_2,..., \theta_r)$:

Bước 1

Tìm ước lượng tham số $\theta = (\theta_1, \theta_2,..., \theta_r)$ là $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2,..., \hat{\theta}_r)$. (Có thể dùng phương pháp hợp lý cực đại hoặc phương pháp mô men,...).

Tính $\hat{p}_i^0 = p_i^0(\hat{\theta})$.

Bước 2

Giả thiết H_0 bị bác bỏ ở mức α nếu: $Z = \sum_{i=1}^{s} \frac{(n_i - n\hat{p}_i^0)^2}{n\hat{p}_i^0} > C'_\alpha$ (5.30)

(Nếu $Z < C'_\alpha$ thì chấp nhận H_0).

Trong đó C'_α tra ở bảng phân phối χ^2 với s – 1 – r bậc tự do và mức ý nghĩa α.

Ví dụ 5.30. Gieo một con xúc xắc 6000 lần. Số lần xuất hiện các mặt 1, 2, 3, 4, 5, 6 được cho ở bảng sau đây:

Số chấm	1	2	3	4	5	6
Số lần gieo	1045	942	1016	1064	988	945

Có thể coi con xúc xắc đó là xúc xắc thật không? (Nghĩa là xác suất để mỗi mặt xuất hiện bằng $\frac{1}{6}$). Cho $\alpha = 0,05$.

Giải:

Để giải bài toán này ta giả sử giả thiết H_0 là "con xúc xắc đó là con xúc xắc thật".

Con xúc xắc thật, nghĩa là cân đối và đồng chất, nên mỗi mặt xuất hiện có khả năng như nhau với xác suất đều bằng $\frac{1}{6}$.

Vậy ta kiểm định giả thiết H_0:

$$p_1 = \frac{1}{6},\ p_2 = \frac{1}{6},\ldots,\ p_6 = \frac{1}{6};\ K:\ p_i \neq \frac{1}{6} \text{ với } i = 1, 2,\ldots, 6.$$

Ta có: $N = 6000$; $n_1 = 1045$; $n_2 = 942$; $n_3 = 1016$; $n_4 = 1064$; $n_5 = 988$; $n_6 = 945$;

$$Z = \sum_{i=1}^{6} \frac{(n_i - np_i^0)^2}{np_i^0} = \frac{(1045 - 6000\times\frac{1}{6})^2}{6000\times\frac{1}{6}} + \frac{(942 - 6000\times\frac{1}{6})^2}{6000\times\frac{1}{6}} +$$

$$+ \frac{(1016 - 6000\times\frac{1}{6})^2}{6000\times\frac{1}{6}} + \frac{(1064 - 6000\times\frac{1}{6})^2}{6000\times\frac{1}{6}} +$$

$$+ \frac{(988 - 6000\times\frac{1}{6})^2}{6000\times\frac{1}{6}} + \frac{(945 - 6000\times\frac{1}{6})^2}{6000\times\frac{1}{6}} = 12,91.$$

Tra bảng phân phối khi bình phương ta tìm được $C(5\%; 5) = 11,07$.

Vì $Z > C_\alpha$. nên giả thiết H_0 bị bác bỏ ở mức $\alpha = 0,05$, nghĩa là con xúc xắc đó là giả.

Ví dụ 5.31. Người ta điều tra ngẫu nhiên 1600 gia đình có 4 con. Kết quả điều tra cho ở bảng sau đây:

Số con trai trong một gia đình	0	1	2	3	4
Số gia đình	111	367	576	428	118

Hãy kiểm định sự đúng đắn của tập hợp hai giả thiết sau đây ở mức $\alpha = 0,05$.

H'_0 = "Xác suất để một trẻ em là con trai bằng 0,5".

H''_0 = "Giới tính các trẻ em trong cùng một gia đình có tính độc lập".

Nếu ta quyết định loại bỏ tập hợp hai giả thiết H'_0, H''_0 thì hãy kiểm định sự đúng đắn của giả thiết H''_0 ở mức $\alpha = 0,05$.

Giải:

a) Ta hãy kiểm định sự đúng đắn của hai giả thiết H'_0, H''_0. Gọi p_i, $(i = 0, 1, 2, 3, 4)$ là xác suất để một gia đình có 4 con thì có i con trai. Ta nhận thấy H'_0 và H''_0 tương đương với giả thiết là bốn lần sinh độc lập (mỗi lần sinh 1 con) và trong mỗi lần sinh xác suất sinh con trai là 0,5. Vậy ta có thể xem việc sinh 4 con như tiến hành dãy 4 phép thử Bernoulli với xác suất $p = 0,5$. Theo công thức xác suất nhị thức, ta có:

$$P_4(k) = C_4^k(\frac{1}{2})^k\left(\frac{1}{2}\right)^{4-k} = C_4^k.\frac{1}{2^4}.$$

Vậy giả thiết H'_0 và H''_0 tương đương với giả thiết H_0 như sau:

$$H_0: p_0 = \frac{1}{16}, p_1 = \frac{1}{4}, p_2 = \frac{3}{8}, p_3 = \frac{1}{4}, p_4 = \frac{1}{16}.$$

Áp dụng công thức (5.29), ta có:

$$Z = \frac{(111-100)^2}{100} + \frac{(367-400)^2}{400} + \frac{(576-600)^2}{600} +$$

$$+\frac{(428-400)^2}{400} + \frac{(118-100)^2}{100} = 10,09$$

Tra bảng phân phối khi bình phương ta tìm được C(5%; 4) = 9,49.

Vì $Z > C_\alpha$ nên giả thiết H_0 bị bác bỏ, nghĩa là tập hợp hai giả thiết H'_0 và H''_0 bị bác bỏ.

b) Kiểm định sự đúng đắn của giả thiết H''_0. Trước hết, ta giả sử xác suất sinh con trai là θ. Dưới giả thiết H''_0 thì xác suất để trong gia đình 4 con có k con trai bằng $p_k = C_4^k \theta^k (1-\theta)^{4-k}$, k = 0, 1, 2, 3, 4.

Ta hãy kiểm định giả thiết H''_0: $p_k = C_4^k 0^k (1-\theta)^{4-k}$, k = 0, 1, 2, 3, 4.

Để kiểm định giả thiết này ta hãy ước lượng θ.

Ta có thể tính $\hat{\theta}$ như sau:

Tính tỉ số giữa số con trai và tổng số con trong 1600 gia đình:

$$\hat{\theta} = \frac{1\times 367 + 2\times 576 + 3\times 428 + 4\times 118}{6400} \approx 0{,}51.$$

Thay $\hat{\theta} = 0{,}51$ vào biểu thức của p_k ta được:

$$\hat{p}_0 = 0{,}057;\ \hat{p}_1 = 0{,}23;\ \hat{p}_2 = 0{,}37;\ \hat{p}_3 = 0{,}26;\ \hat{p}_4 = 0{,}07.$$

Tính giá trị của Z:

$$Z = \sum_{i=0}^{4} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

$$= \frac{(111-96)^2}{96} + \frac{(367-384)^2}{384} + \frac{(576-592)^2}{592}$$

$$+ \frac{(428-416)^2}{416} + \frac{(118-112)^2}{112} \approx 4{,}18.$$

Tra bảng phân phối khi bình phương ta được C'(5%; 3) = 7,81.

Vì $Z < C'_\alpha$ nên giả thiết H''_0 được chấp nhận, nghĩa là giới tính của các trẻ em trong cùng một gia đình có tính độc lập.

3.9. Tiêu chuẩn khi bình phương kiểm định tính độc lập và tính thuần nhất

3.9.1. Kiểm định tính độc lập

Ta xét n phép thử độc lập. Trong mỗi phép thử có một và chỉ một trong các biến có A_1, A_2,..., A_v và một trong các biến cố B_1, B_2,... B_w xảy ra.

Đặt $P(A_iB_j) = p_{ij}$.

Ta lập bảng xác suất của các biến cố A_i B_j:

B / A	B_1	B_2	....	B_w	Tổng
A_1	p_{11}	p_{12}	...	p_{1w}	p_{10}
A_2	p_{21}	p_{22}	...	p_{2w}	p_{20}
⋮	...	....	...	...	...
A_v	p_{v1}	p_{v2}	...	p_{vw}	p_{v0}
Tổng	p_{01}	p_{02}	...	p_{0w}	1

Trong đó: $P(A_i) = p_{i0} = \sum_{j=1}^{w} p_{ij}$, $P(B_f) = p_{0j} = \sum_{i=1}^{v} p_{ij}$.

Hãy kiểm định giả thiết:

$$H_0: p_{ij} = p_{i0}.p_{0j};\ K: p_{ij} \neq p_{i0}\,p_{ij} \text{ ở mức } \alpha.$$

Gọi X_{ij} là số lần xuất hiện biến cố tích A_iB_j trong n phép thử. Ta có bảng quan sát sau:

B / A	B_1	B_2	....	B_w	Tổng
A_1	X_{11}	X_{12}	...	X_{1w}	X_{10}
A_2	X_{21}	X_{22}	...	X_{2w}	X_{20}
⋮	...	....	...	...	...
A_v	X_{v1}	X_{v2}	...	X_{vw}	X_{v0}
Tổng	X_{01}	X_{02}	...	X_{0w}	n

Tiêu chuẩn kiểm định giả thiết này là :

Giả thiết H_0 bị bác bỏ ở mức α nếu

$$Z = n \times \sum_{i=1}^{v} \sum_{j=1}^{w} \frac{(X_{ij} - \frac{X_{i0}X_{0j}}{n})^2}{X_{0i}X_{0j}} > C_\alpha \qquad (5.31)$$

Trong đó C_α tra trong bảng phân phối khi bình phương với $(v-1)(w-1)$ bậc tự do và mức ý nghĩa α. (Còn $Z < C_\alpha$ thì chấp nhận giả thiết H_0).

Ví dụ 5.32. Để nghiên cứu sự phụ thuộc giữa việc phân chia nhóm học sinh để giảng dạy với tình trạng kiến thức của học sinh về bộ môn thể dục, người ta chia 30 học sinh thành 3 nhóm. Mỗi nhóm 10 em: nhóm thứ nhất gồm những em có năng khiếu đặc biệt; nhóm thứ hai gồm những em khá; nhóm thứ 3 gồm những em trung bình. Kết quả kiểm tra sau một đợt huấn luyện như sau:

	Trung bình	Khá	Giỏi	Tổng
Nhóm I	3	3	4	10
Nhóm II	4	4	2	10
Nhóm III	2	8		10
Tổng	9	15	6	N = 30

Hãy kiểm định giả thiết:

H_0: "Sự phân nhóm để dạy độc lập với tình trạng kiến thức của học sinh" ở mức $\alpha = 0,05$.

Giải:

Tra bảng phân phối khi bình phương ta tìm được $C(5\%; 4) = 9,49$.

Tính giá trị của Z:

$$Z = n \times \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{\left(X_{ij} - \frac{X_{i0}X_{0j}}{n}\right)^2}{X_{i0}X_{0j}}$$

$$= 30 \times \left[\frac{\left(3 - \frac{10\times 9}{30}\right)^2}{10\times 9} + \frac{\left(3 - \frac{10\times 15}{30}\right)^2}{10\times 15} + \frac{\left(4 - \frac{10\times 6}{30}\right)^2}{10\times 6} + \right.$$

$$+ \frac{\left(4 - \frac{10\times 9}{30}\right)^2}{10\times 9} + \frac{\left(4 - \frac{10\times 15}{30}\right)^2}{10\times 15} + \frac{\left(2 - \frac{10\times 6}{30}\right)^2}{10\times 6} +$$

$$\left. + \frac{\left(2 - \frac{10\times 9}{30}\right)^2}{10\times 9} + \frac{\left(8 - \frac{10\times 15}{30}\right)^2}{10\times 15} + \frac{\left(0 - \frac{10\times 6}{30}\right)^2}{10\times 6} \right] \approx 7{,}4.$$

Ta thấy $Z = 7{,}4 < C_\alpha = 9{,}49$. Ta chấp nhận giả thiết H_0, nghĩa là việc phân chia nhóm để dạy không ảnh hưởng gì đến tình trạng kiến thức của học sinh về môn thể dục.

• Xét trường hợp $v = w = 2$:

Ta có bảng giá trị quan sát:

A \ B	B_1	B_2
A_1	a	b
A_2	c	d

Khi đó:

$$\text{Nếu } Z = \frac{n \times (ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} > C_\alpha \qquad (5.32)$$

thì bác bỏ giả thiết; Nếu $Z < C_\alpha$ thì chấp nhận giả thiết H_0. Trong đó C_α tra ở bảng phân phối khi bình phương 1 bậc tự do và với mức ý nghĩa α.

Chú ý. Các tiêu chuẩn (5.31), (5.32) áp dụng tốt trong trường hợp mẫu có kích thước lớn, vì phân phối của Z có phân phối tiệm cận là phân phối khi bình phương.

Ta còn yêu cầu $\frac{X_{i0}X_{0j}}{n}$ phải lớn hơn hoặc bằng 5.

Trong trường hợp bảng vuông 2×2 thì số quan sát phải vượt quá 10.

Ví dụ 5.32. Xét đám ốc sên rừng. Đặc tính A về mầu của vỏ có 2 biến dạng: vàng (A_1) và hồng (A_2).

Đặc tính B là số vạch trên vỏ có từ 0 đến 5. Quan sát 169 con ốc sên ta ghi được kết quả như sau:

Số vạch		0	1	2	3	4	5
Màu	Vàng	35	6	13	32	4	25
	Hồng	14	2	12	16	0	10

Có thể chấp nhận là trong đám ốc sên đó thì màu vỏ và số vạch trên vỏ độc lập với nhau không? Cho mức kiểm định $\alpha = 0,05$.

Giải:

Số vỏ ốc sên có 1 hay 4 vạch quá ít để có thể sử dụng tiêu chuẩn khi bình phương vì $\frac{X_{i0}X_{0j}}{n}$ ít nhất cũng phải bằng 5. Vì vậy ta phải gộp số ốc có 1 vạch và 2 vạch vào nhau, 3 vạch và 4 vạch vào nhau. Ta có bảng số liệu mới:

Số vạch		0	1 hay 2	3 hay 4	5	Tổng
Màu	Vàng	35	19	36	25	115
	Hồng	14	14	16	10	54
Tổng		49	33	52	35	169

Ở đây $v = 2$, $w = 4$. Số bậc tự do $(v - 1)(w - 1) = 3$.

Tra bảng phân phối χ^2 ta tìm được $C(5\%; 3) = 7,81$.

Tính giá trị của Z:

$$Z = 169 \times \left[\frac{\left(35 - \frac{115 \times 49}{169}\right)^2}{115 \times 49} + \frac{\left(19 - \frac{115 \times 33}{169}\right)^2}{115 \times 33} + \right.$$

$$+ \frac{\left(36 - \frac{115 \times 52}{169}\right)^2}{115 \times 52} + \frac{\left(25 - \frac{115 \times 35}{169}\right)^2}{115 \times 35} + \frac{\left(14 - \frac{49 \times 54}{169}\right)^2}{49 \times 54} +$$

$$\left. + \frac{\left(14 - \frac{33 \times 54}{169}\right)^2}{33 \times 54} + \frac{\left(16 - \frac{52 \times 54}{169}\right)^2}{52 \times 54} + \frac{\left(10 - \frac{35 \times 54}{169}\right)^2}{35 \times 54} \right]$$

$= 2,13.$

Ta có $Z < C(5\%; 3)$.

Vậy chấp nhận giả thiết trên, nghĩa là màu trên vỏ ốc sên và số các vạch trên vỏ của nó là đặc tính độc lập với nhau.

3.9.2. Kiểm định tính thuần nhất

Xét một đám đông Q các cá thể có các biến dạng $A_1, A_2,..., A_v$.

Giả sử đám đông Q được chia thành các đám đông nhỏ $B_1, B_2,... B_w$.

w đám đông nhỏ $B_1, B_2,..., B_w$ của đám đông Q được gọi là thuần nhất đối với đặc tính A_i nào đó nếu việc khảo sát đặc tính đó không cho phép ta phân biệt được ở các đám đông nhỏ, nghĩa là xác suất để cá thể có đặc tính A_i ở các đám đông $B_1, B_2,..., B_w$ là như nhau.

Vậy bài toán kiểm định tính thuần nhất được viết dưới dạng:

$$H_0: \begin{cases} p_{11} = p_{12} = ... = p_{1w} \\ p_{21} = p_{22} = ... = p_{2w} \\ ..... \\ p_{v1} = p_{v2} = ... = p_{vw} \end{cases}$$

K : $p_{i1} \neq p_{i2} \neq ... \neq p_{iw}$, $i = \overline{1,v}$ ở mức α.

Lời giải của bài toán này giống như lời giải của bài toán kiểm định tính độc lập. Do đó, tiêu chuẩn (5.31), (5.32) cũng dùng để kiểm định giả thiết H_0.

Ví dụ 5.33. Có hai phương pháp đều tự được tiến hành trên hai nhóm bệnh nhân I và II. Một số bệnh nhân thì bệnh nặng thêm, còn một số bệnh nhân thì thuyên giảm và khỏi. Kết quả điều trị được cho ở bảng như sau:

	Tình trạng bệnh	Nặng thêm	Nhẹ đi
Nhóm	I	270	220
	II	120	210

So sánh hiệu quả của hai phương pháp điều trị ở mức $\alpha = 0,05$.

Giải:

Giả thiết H_0 "Hiệu quả của hai phương pháp đó là như nhau"

Ở đây $v = w = 2$ ta áp dụng công thức (5.32).

Tra ở bảng phân phối khi bình phương $C(5\%; 1) = 3,84$.

Tính Z: $Z = \dfrac{820(270 \times 210 - 120 \times 220)^2}{490 \times 330 \times 390 \times 430} = 27,76.$

Vì $Z > C_\alpha$ nên giả thiết H_0 bị bác bỏ, nghĩa là hai phương pháp điều trị đó có hiệu quả khác nhau.

Ngoài ra tỉ lệ khỏi bệnh (thực nghiệm) ở nhóm II cao hơn ở nhóm I. Ta kết luận hiệu quả của hai phương pháp điều trị ở nhóm II tốt hơn ở nhóm I.

4. HỒI QUY VÀ TƯƠNG QUAN

4.1. Hồi quy một biến

4.1.1. Hồi quy tuyến tính một biến

Định nghĩa 5.14. Gọi kỳ vọng toán điều kiện E(Y/X) là hàm hồi quy của biến ngẫu nhiên Y theo biến ngẫu nhiên X.

Tương tự, gọi kỳ vọng toán điều kiện E(X/Y) là hàm hồi quy của biến ngẫu nhiên X theo biến ngẫu nhiên Y.

Định nghĩa 5.15. Hàm hồi quy của biến ngẫu nhiên Y theo biến ngẫu nhiên X được gọi là tuyến tính nếu:

$E(Y/X) = aX + b$ với a, b là hằng số.

Hàm hồi quy của biến ngẫu nhiên X theo biến ngẫu nhiên Y được gọi là tuyến tính nếu:

$E(X/Y) = cY + d$ với c, d là hằng số.

Phương pháp tìm hồi quy

Trước hết ta chứng minh mệnh đề sau:

Mệnh đề. $E(Y - \varphi(X))^2$ đạt cực tiểu khi $\varphi(X) = E(Y/X)$.

Chứng minh

$$E(Y - \varphi(X))^2 = E[Y - E(Y/X) - (\varphi(X) - E(Y/X))]^2$$

$$= E(Y - E(Y/X))^2 + E(\varphi(X) - E(Y/X))^2$$

$$- 2\,E(Y - E(Y/X))(\varphi(X) - E(Y/X)).$$

Vì $E(Y - E(Y/X))(\varphi(X) - E(Y/X)) = E(Y/X) - E(Y/X) = 0$

nên $E(Y - \varphi(X))^2 = E(Y - E(Y/X))^2 + E(\varphi(X) - E(Y/X))^2$.

Vế phải là tổng của hai số hạng dương. Nó nhỏ nhất khi một số phải triệt tiêu. Ta chọn $\varphi(X) = E(Y/X)$ thì vế phải sẽ nhỏ nhất nghĩa là vế trái cũng bé nhất. Đó là điều phải chứng minh.

Phương pháp tìm hàm $\varphi(X)$ sao cho $E(Y - \varphi(X))^2$ đạt cực tiểu được gọi là phương pháp bình phương bé nhất.

Áp dụng phương pháp bình phương bé nhất để tìm hàm hồi quy tuyến tính.

Ta phải tìm hằng số a, b sao cho $E(Y - aX - b)^2 \to \min$.

Muốn vậy ta tìm các đạo hàm riêng cấp 1 của $E(Y - aX - b)^2$ theo ẩn a và ẩn b và cho chúng bằng 0, nghĩa là:

$$\begin{cases} \dfrac{\partial}{\partial a}E(Y-aX-b)^2=0 \\ \dfrac{\partial}{\partial b}E(Y-aX-b)^2=0 \end{cases} \Leftrightarrow \begin{cases} E(Y-aX-b)X=0 \\ E(Y-aX-b)\quad =0 \end{cases}$$

$$\Leftrightarrow \begin{cases} aE(X^2)+bEX=EXY \\ aEX+b\qquad =EY \end{cases}$$

$$\Leftrightarrow \begin{cases} \dfrac{\partial}{\partial a}E(Y-aX-b)^2=0 \\ \dfrac{\partial}{\partial b}E(Y-aX-b)^2=0 \end{cases}$$

$$\Leftrightarrow \begin{cases} E(Y-aX-b)X=0 \\ E(Y-aX-b)\quad =0 \end{cases}$$

$$\Leftrightarrow \begin{cases} aE(X^2)+bEX=EXY \\ aEX+b\qquad =EY \end{cases}$$

$$\Leftrightarrow \begin{cases} a=\dfrac{E(XY)-EXEY}{E(X^2)-(EX)^2} \\ b=EY-aEX \end{cases}$$

Hàm hồi quy tuyến tính tìm được là: $y = ax + b$.

Tương tự, ta cũng có thể tìm hàm hồi quy tuyến tính của X theo Y và kết quả là:

$$c=\frac{E(XY)-EXEY}{E(Y^2)-(EY)^2};$$

$$d = EX - cEY.$$

Phương trình đường hồi quy tuyến tính là $x = cy + d$.

Khi cho mẫu quan sát (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n) đối với vectơ ngẫu nhiên (X, Y) thì hàm hồi quy tuyến tính mẫu được tính như sau:

$$\hat{a}=\frac{\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n}X^2_i-(\bar{X})^2}=\frac{n\sum_{i=1}^{n}X_iY_i-(\sum_{i=1}^{n}X_i)(\sum_{i=1}^{n}Y_i)}{n\sum_{i=1}^{n}X^2_i-(\sum_{i=1}^{n}X_i)^2},$$

$$\hat{b} = \bar{Y} - \hat{a}\bar{X}$$

và hàm hồi quy tuyến tính mẫu của Y theo X là $y = \hat{a}x + \hat{b}$.

Tương tự, ta cũng tìm được:

$$\hat{c} = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n\sum_{i=1}^{n} Y^2{}_i - (\sum_{i=1}^{n} Y_i)^2},$$

$$\hat{d} = \bar{X} - \hat{c}\bar{Y}$$

và hàm hồi quy của X theo Y là $x = \hat{c}y + \hat{d}$.

Ví dụ 5.34. Với số liệu trong ví dụ 5.7. Tìm hàm hồi quy tuyến tính mẫu của Y theo X và của X theo Y.

Giải:

Trong ví du 5.7 ta đã tính được n = 8.

$$\sum_{i=1}^{n} X_i = 44; \sum_{i=1}^{n} X_i{}^2 = 284; \sum_{i=1}^{n} Y_i = 88; \sum_{i=1}^{n} Y_i^2 = 1142; \sum_{i=1}^{n} X_i Y_i = 568.$$

Ta sẽ tính $\hat{a}$; $\hat{b}$; $\hat{c}$; $\hat{d}$.

$$\hat{a} = \frac{8 \times 568 - 44 \times 88}{8 \times 284 - 44^2} = \frac{672}{336} = 2;$$

$$\hat{b} = 11 - 2 \times 5{,}5 = 0.$$

Vậy hàm hồi quy tuyến tính mẫu của Y theo X là $y = 2x$.

Tương tự:

$$\hat{c} = \frac{8 \times 568 - 44 \times 88}{8 \times 1142 - 88^2} = \frac{672}{1392} \approx 0{,}4827;$$

$$\hat{d} = 5{,}5 - 0{,}4827 \times 11.$$

Vậy hàm hồi quy tuyến tính mẫu của X theo Y là: $x = 0{,}4827y + 0{,}19$.

4.1.2. Hồi quy không tuyến tính

Ta chỉ xét dạng hồi quy:

$$\hat{y} = E(Y/X) = a_k X^k + a_{k-1} X^{k-1} + ... + a_1 X + a_0$$

Khi cho mẫu quan sát (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n) đối với vectơ ngẫu nhiên (X, Y), ta có thể tìm trực tiếp $a_0, a_1, ..., a_k$ bằng phương pháp bình phương bé nhất.

Tìm $a_0, ..., a_k$ sao cho:

$Q = \sum_{i=1}^{n} (Y_i - a_k X_i^k - a_{k-1} X_i^{k-1} - ... - a_1 X_i - a_0)^2$ đạt cực tiểu. Lấy đạo hàm riêng cấp một của Q theo $a_0, a_1, ..., a_k$ và cho chúng bằng 0 ta được hệ phương trình.

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} X_i + ... + a_k \sum_{i=1}^{n} X_i^k = \sum_{i=1}^{n} Y_i \\ ..... \\ a_0 \sum_{i=1}^{n} X_i^k + a_1 \sum_{i=1}^{n} X_i^{k+1} + ... + a_k \sum_{i=1}^{n} X_i^{2k} = \sum_{i=1}^{n} X_i^k Y_i \end{cases} \quad (5.33)$$

Giải hệ phương trình (5.33) ta nhận được $a_0, a_1, ..., a_k$ và hàm hồi quy mẫu của Y theo X là: $y = a_k x^k + a_{k-1} x^{k-1} + ... + a_1 x + a_0$.

Ví dụ 5.35. Cho mẫu quan sát đối với vectơ ngẫu nhiên (X, Y) là

Y \ X	1	2	3	4
5				1
4		3	6	6
3	3	8	6	3
2	5	4	3	
1	2			

Tìm hàm hồi quy mẫu của Y theo X dạng $y = a_0 + a_1 x + a_2 x^2$.

Giải:

Từ hệ (5.33) ta suy ra:

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} X_i + a_2 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i. \\ a_0 \sum_{i=1}^{n} X_i + a_1 \sum_{i=1}^{n} X_i^2 + a_2 \sum_{i=1}^{n} X_i^3 = \sum_{i=1}^{n} X_i Y_i \\ a_0 \sum_{i=1}^{n} X_i^2 + a_1 \sum_{i=1}^{n} X_i^3 + a_2 \sum_{i=1}^{n} X_i^4 = \sum_{i=1}^{n} X_i^2 Y_i \end{cases}$$

Lập bảng để tính các tổng

$$\sum X_i; \sum X_i^2; \sum X_i^3; \sum X_i^4; \sum Y_i; \sum X_i Y_i; \sum X_i^2 Y_i.$$

Y \ X	1	2	3	4	n_y	$n_y Y$	$n_{xy} XY$	$n_{xy} X^2 Y$
5				1	1	5	20	80
4		3	6	6	15	60	192	648
3	3	8	6	3	20	60	147	411
2	5	4	3		12	24	44	96
1	2				2	2	2	2
n_x	10	15	15	10	50	151	405	1237
$n_x X$	10	30	45	40	125			
$n_x X^2$	10	60	135	160	365			
$n_x X^3$	10	120	405	640	1175			
$n_x X^4$	10	240	1215	2560	4025			

Thay kết quả ở bảng phụ vào hệ phương trình trên ta có:

$$50a_0 + 125a_1 + 365a_2 = 151.$$

$$125a_0 + 365a_1 + 1175a_2 = 405.$$

$$365a_0 + 1175a_1 + 4025a_2 = 1237.$$

Giải hệ này ta tìm được:

$$\hat{a}_0 = 1,407; \hat{a}_1 = 0,815; \hat{a}_2 = -0,058.$$

Phương trình hồi quy mẫu của Y theo X là:

$$Y = 1,047 + 0,815x - 0,058x^2.$$

4.2. Hệ số tương quan và tỉ số tương quan

Như ở phần xác suất ta đã định nghĩa hệ số tương quan của X, Y là:

$$\rho(X, Y) = \frac{E(X - EX)(Y - EY)}{\sqrt{DXDY}}.$$

Khi cho mẫu quan sát (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n) đối với (X, Y), hệ số tương quan mẫu được tính theo công thức:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}}{S_n(X)S_n(Y)} = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}}.$$

Định nghĩa 5.16. Nếu DY > 0 thì căn bậc hai của tỉ số $\frac{D(E(Y/X))}{DY}$ được gọi là tỉ số tương quan của Y đối với X, kí hiệu là $K_X(Y)$:

$$K_X(Y) = \sqrt{\frac{D(E(Y/X))}{DY}}.$$

Người ta chứng minh được rằng: $0 < K_X(Y) < 1$.

Nếu $K_X(Y) = 0$ thì X và Y là không tương quan. Trong lĩnh vực dự báo thì người ta gọi hàm hồi quy E(Y/X) là hàm dự báo của Y theo X.

Độ sai bình phương trung bình giữa hai giá trị của Y với giá trị dự báo, nghĩa là $\delta = E(Y - E(Y/X))^2$ được gọi là độ sai dự báo. Từ tính chất của kỳ vọng điều kiện và từ đẳng thức:

$$DY = E(Y - E(Y/X))^2 + D(E(Y/X))$$

Ta suy ra: $\delta = DY\,(1 - K_X^2(Y))$. (5.34)

Nếu hàm dự báo E(Y/X) là tuyến tính, nghĩa là E(Y/X) = aX + b thì $\rho(X, Y) = K_X(Y)$ và như vậy: $\delta = DY\,(1 - \rho^2(X, Y))$. (5.35)

Khi cho mẫu quan sát (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n) đối với vectơ ngẫu nhiên (X, Y), ta có ước lượng độ sai dự báo tuyến tính của Y theo X là:

$$\hat{\delta} = S_n^2(Y)(1 - r^2) \tag{5.36}$$

Trở lại ví dụ 5.7 và ví dụ 5.34 ta tính độ sai dự báo khi ta thay Y bằng hồi quy mẫu: $\hat{y} = \hat{a}X + \hat{b}$ là:

$$\hat{\delta} = S_n^2(Y)(1 - r^2) = 5{,}25(1 - 0{,}98^2) = 0{,}2079.$$

4.3. Khoảng ước lượng của hệ số hồi quy và hệ số tương quan

Người ta tính được khoảng ước lượng của hệ số hồi quy a với độ tin cậy $1 - \alpha$ là:

$$\hat{a} - x_\alpha \frac{S_n(X)}{S_n(Y)} \times \frac{1 - r^2}{\sqrt{n}} < a < \hat{a} + x_\alpha \frac{S_n(X)}{S_n(Y)} \times \frac{1 - r^2}{\sqrt{n}} \tag{5.37}$$

trong đó x_α tra ở bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 1 - \frac{\alpha}{2}$.

Khoảng ước lượng của hệ số tương quan $\rho(X, Y)$ với độ tin cậy $1 - \alpha$ là:

$$r - x_\alpha \times \frac{1 - r^2}{\sqrt{n}} < \rho(X, Y) < r + x_\alpha \times \frac{1 - r^2}{\sqrt{n}} \tag{5.38}$$

trong đó x_α tra trong bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 1 - \frac{\alpha}{2}$.

Ví dụ 5.36. Cho mẫu quan sát đối với vectơ ngẫu nhiên (X, Y) là:

X	0	1	1	3	3	4
Y	3	3	4	4	5	5

Tìm hệ số tương quan mẫu r.

Tìm hàm hồi quy tuyến tính mẫu của Y theo X.

Tính độ sai dự báo tuyến tính của Y theo X.

Tính khoảng ước lượng của hệ số hồi quy và hệ số tương quan với độ tin cậy 95%.

Giải:

$$\sum_{i=1}^{n} X_i = 0 + 1 + 1 + 3 + 3 + 4 = 12;\ \sum_{i=1}^{n} Y_i = 24;$$

$$\sum_{i=1}^{n} X_i^2 = 0^2 + 1^2 + 1^2 + 3^2 + 3^2 + 4^2 = 36;\ \sum_{i=1}^{n} Y_i^2 = 100;$$

$$\sum_{i=1}^{n} X_i Y_i = 0 \times 3 + 1 \times 3 + 1 \times 4 + 3 \times 4 + 3 \times 5 + 4 \times 5 = 54.$$

Hệ số tương quan mẫu:

$$r = \frac{6 \times 54 - 12 \times 24}{\sqrt{[6 \times 36 - 12^2][6 \times 100 - 24^2]}} = \frac{36}{\sqrt{24 \times 72}} = \frac{36}{41,57} \approx 0,866;$$

$$S_n^2(X) = \frac{36}{6} - 2^2 = 2;\ S_n^2(Y) = \frac{100}{6} - 4^2 = 0,66;$$

$$\hat{a} = \frac{36}{72} = 0,5;\ \hat{b} = \bar{Y} - \hat{a}\bar{X} = 4 - 2 \times 0,5 = 3.$$

Hàm hồi quy tuyến tính mẫu của Y theo X là: $y = 0,5x + 3$.

Độ sai dự báo: $\hat{\delta} = S_n^2(Y)(1 - r^2) = 0,66(1 - 0,866^2) = 0,165$.

Khoảng ước lượng của hệ số hồi quy a với độ tin cậy 95%:

$x_\alpha = 1,96$ (tra bảng phân phối chuẩn N(0; 1) sao cho $F(x_\alpha) = 0,975$).

$$\hat{a} - x_\alpha \frac{S_n(X)}{S_n(Y)} \times \frac{1 - r^2}{\sqrt{n}} < a < \hat{a} + x_\alpha \frac{S_n(X)}{S_n(Y)} \times \frac{1 - r^2}{\sqrt{n}}$$

$$\Leftrightarrow 0,5 - 1,96 \frac{\sqrt{2}}{\sqrt{0,66}} \times \frac{1 - 0,866^2}{\sqrt{6}} < a < 0,5 + 1,96 \frac{\sqrt{2}}{\sqrt{0,66}} \times \frac{1 - 0,866^2}{\sqrt{6}}$$

$$\Leftrightarrow 0,5 - 0,35 < a < 0,5 + 0,35$$

$$\Leftrightarrow 0,15 < a < 0,85.$$

Khoảng ước lượng của hệ số tương quan $\rho(X, Y)$ với độ tin cậy 95%.

$$r - x_\alpha \frac{1-r^2}{\sqrt{n}} < \rho(X,Y) < r + x_\alpha \frac{1-r^2}{\sqrt{n}}$$

$$0,866 - 1,96 \times \frac{1-0,866^2}{\sqrt{6}} < \rho(X,Y) < 0,866 + 1,96 \times \frac{1-0,866^2}{\sqrt{6}}$$

$$\Leftrightarrow 0,866 - 0,200 < \rho(X,Y) < 0,866 + 0,200$$

$$\Leftrightarrow 0,666 < \rho(X,Y) < 1,066.$$

BÀI TẬP CHƯƠNG V

1. Tìm hàm phân phối mẫu biết rằng mẫu ngẫu nhiên đã cho là:

a.

X_i	1	4	6
n_i	10	15	25

b.

X_i	2	5	7	8
n_i	10	3	2	4

c.

X_i	4	7	8
n_i	5	2	3

Cũng với số liệu đã cho ở trên, tính trung bình mẫu và phương sai mẫu tương ứng.

2. Xây dựng các đa giác tần suất tương ứng với các mẫu ngẫu nhiên sau:

a.

X_i	1	4	5	7
n_i	20	10	14	6

b.

X_i	2	3	5	6
n_i	10	15	5	20

c.

X_i	15	20	25	30	35
n_i	10	15	30	20	25

3. Xây dựng tổ chức đồ tần suất tương ứng đối với các mẫu ngẫu nhiên đã cho sau đây:

a.

STT	Khoảng li	Tần số n_i
1	2 – 7	5
2	7 – 12	10
3	12 – 17	25
4	17 – 22	6
5	22 – 27	4

b.

STT	Khoảng li	Tần số n_i
1	10 – 15	2
2	15 – 20	4
3	20 – 25	8
4	25 – 30	4
5	30 – 35	2

4. Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối Poisson với tham số $\lambda > 0$, nghĩa là $dP[X = x_i] = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$, $x_i = 0, 1, 2,...$

Tìm ước lượng hợp lí cực đại của λ.

5. Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối:

$$f(x_i, p) = p^{x_i}(1-p)^{1-x_i} \text{ với } x_i = 0,1.$$

Tìm ước lượng hợp lí cực đại của tham số p.

6. Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối với hàm mật độ là:

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{với } x > 0,\ \theta > 0 \\ 0 & \text{với } x \le 0 \end{cases}$$

Tìm ước lượng hợp lí cực đại của θ.

7. Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên độc lập từ phân phối chuẩn dạng $N(a, \sigma^2)$, nghĩa là hàm mật độ dạng:

$$f(x, a, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Tìm ước lượng hợp lí cực đại của a, σ^2.

8. Giả sử biến ngẫu nhiên X là sản lượng được tính ra tạ/ha của loại lúa đã cho trong một miền xác định có phân phối chuẩn và gọi a và σ^2 là kỳ vọng và phương sai của X. Hãy tìm ước lượng điểm của a và σ^2.

Tìm khoảng ước lượng của a, σ^2 với độ tin cậy 90%. Biết rằng kết quả thu được trên 10 mảnh đất là:

Mảnh	1	2	3	4	5	6	7	8	9	10
Sản lượng	51	48	56	57	44	52	54	60	46	47

9. Độ cao X của trẻ em là đại lượng ngẫu nhiên tuân theo phân phối chuẩn dạng $N(a; \sigma^2)$. Tìm ước lượng điểm của a, σ^2. Tìm khoảng ước lượng của a, σ^2 với độ tin cậy 95% nếu như đo ngẫu nhiên 10 em có kết quả như sau:

1	2	3	4	5	6	7	8	9	10
1,5	1,55	1,49	1,51	1,50	1,52	1,45	1,60	1,50	1,46

10. Gieo ngẫu nhiên 60 hạt đậu tương thấy có 15 hạt không nảy mầm. Tìm ước lượng điểm của xác suất nảy mầm p và tìm khoảng ước lượng của p với độ tin cậy 95%.

11. Bắn ngẫu nhiên liên tiếp 10000 viên đạn độc lập vào một mục tiêu thấy có 7000 viên trúng đích. Tìm ước lượng điểm của xác suất bắn trúng đích p của mỗi viên đạn và tìm khoảng ước lượng của p với độ tin cậy 95%.

12. Trong một đợt kiểm tra sức khoẻ của trẻ em ở các nhà trẻ, người ta khám ngẫu nhiên 100 cháu thấy có 20 cháu có triệu chứng còi xương do suy dinh dưỡng. Gọi p là xác suất để một trẻ em mắc

bệnh còi xương. Hãy kiểm định giả thiết:

H_0: p = 0,15 với K : p ≠ 0,15 ở mức α = 5%.

13. Tỉ lệ phế phẩm trong một lô sản phẩm theo người ta cho biết là 0,02. Kiểm tra ngẫu nhiên 480 sản phẩm từ lô hàng thấy có 12 phế phẩm. Hỏi tỉ lệ phế phẩm công bố ở trên có đúng không? Tại sao? Cho α = 5%.

14. Người ta cân ngẫu nhiên 10 trẻ em 2 tuổi. Kết quả như sau:

Trọng lượng X_i	12,3	12,5	12,8	13,0	13,5
Tần số n_i	1	2	4	2	1

Giả sử các quan sát X_i tuân theo luật chuẩn $N(a; \sigma^2)$. Hãy kiểm định giả thiết:

H_0: a =12 với K: a ≠ 12 ở mức α = 5%.

15. Sau một đợt huấn luyện quân sự người ta kiểm tra ngẫu nhiên 70 học sinh. Kết quả như sau:

X_i	5	6	7	8	9	10
n_i	5	10	15	20	12	8

Giả sử các X_i tuân theo luật phân phối chuẩn $N(a; \sigma^2)$. Hãy kiểm định giả thiết:

H_0: a = 8 với K : a ≠ 8 ở mức α = 5%.

16. Để kiểm tra chất lượng sản phẩm do 2 nhà máy sản xuất người ta kiểm tra ngẫu nhiên 10 sản phẩm do nhà máy I và 12 sản phẩm do nhà máy II sản xuất. Kết quả như sau:

Kích thước X_i	3,4	3,5	3,7	3,9
Tần số n_i	2	3	4	1

Kích thước Y_i	3,2	3,4	3,6
Tần số m_i	2	2	8

Giả sử X, Y có phân phối chuẩn dạng tổng quát và DX = DY. Hãy kiểm định giả thiết:

H_0: EX = EY với K : EX ≠ EY ở mức α = 5%.

17. Người ta muốn so sánh trọng lượng óc ở những người trên và dưới 50 tuổi ta xét các kết quả ghi trong bảng sau:

(Các trọng lượng được nhóm thành các lớp cách nhau 50 gam, mỗi lớp được xác định bởi trung điểm của nó).

Tuổi	Trọng lượng (g)						
	1175	1225	1275	1325	1375	1425	1475
Trên 50 tuổi	6	15	27	25	28	18	8
Dưới 50 tuổi	15	36	42	50	54	44	24

Ta có thể cho là trọng lượng trung bình của óc người trên 50 tuổi và người dưới 50 tuổi là như nhau không? Cho $\alpha = 5\%$.

18. Người ta điều tra ngẫu nhiên 250 người ở xã A thấy có 140 nữ và 160 người ở xã B thấy có 80 nữ. Hãy so sánh tỉ lệ nữ ở 2 xã với mức $\alpha = 5\%$.

19. Người ta gieo đậu tương bằng 2 phương pháp:

Theo phương pháp A, gieo 180 hạt có 150 hạt nảy mầm.

Theo phương pháp B, gieo 256 hạt có 160 hạt nảy mầm.

Hãy so sánh hiệu quả của 2 phương pháp ở mức $\alpha = 0{,}05\%$.

20. Người ta nói rằng những người ăn một lượng lớn cam quýt có số lần mắc bệnh cảm cúm ít hơn người bình thường. Qua 5 năm thực nghiệm trên 12 người, kết quả ghi được ở bảng sau:

5 năm ăn bình thường	10	7	12	8	14	19	12	15	10	8	12	14
5 năm ăn lượng lớn cam, quýt	2	3	2	2	5	2	2	2	10	10	7	3

Theo anh (chị) điều nhận định trên có đúng không? Cho mức kiểm định $\alpha = 1\%$.

21. Để so sánh chất lượng lốp xe, người ta đo số kilômét đi được của từng loại lốp. Kết quả thử nghiệm trên 2 loại lốp là như sau:

Loại Y	41	50	33	59	46	54	58	53	54	55	59
Loại X	38	54	30	35	36	50	52	45	47	46	40

Hãy kiểm định giả thiết:

H_0: "Loại lốp X có chất lượng như loại lốp Y", ở mức $\alpha = 5\%$.

22. Cho 2 mẫu độc lập X và Y như sau:

X	3	5	7	8	11	15	19	20	22	24	28
Y	4	2	6	9	10	12	16	17	21	26	27

X	31	50	57	60	75	79	80	85	100	98		
Y	29	45	48	65	67	68	52	53	25	43	44	50

Giả sử X có phân phối $F_1(x)$, Y có phân phối $F_2(y)$ và phân phối đó liên tục. Hãy kiểm định giả thiết:

$H_0: F_1(x) = F_2(x)$ với $K: F_1(x) \neq F_2(x)$ ở mức $\alpha = 5\%$.

23. Để dự báo sâu bệnh của cánh đồng ngô, người ta kiểm tra ngẫu nhiên 500 hốc, mỗi hốc có 2 cây. Kết quả kiểm tra như sau:

Số cây bị bệnh	2 cây bị bệnh	1 cây bị bệnh	0 cây bị bệnh
Số hốc	73	185	242

Tình hình sâu bệnh của ngô như thế đã cần báo động chưa? (Cho mức kiểm định $\alpha = 0{,}05\%$).

24. Người ta chọn ra 200 sản phẩm của một máy tiện và kiểm tra kích thước của các sản phẩm đó với độ chính xác đến 1mm. Bảng sau đây cho độ lệch của các kích thước đo được với kích thước lấy làm chuẩn sau khi đã phân thành các khoảng. Sử dụng tiêu chuẩn χ^2 hãy kiểm định giả thiết H_0: "Phân bố của độ lệch của kích thước các sản phẩm với kích thước chuẩn có phân bố chuẩn $N(a; \sigma^2)$. (Cho mức kiểm định $\alpha = 5\%$).

STT	Khoảng li	Tần số n_i
1	–20 – 15	7
2	–15 – 10	11
3	–10 – 5	15
4	–5 0	24
5	0 5.	49

STT	Khoảng li	Tần số n_i
6	5 – 10	41
7	10 – 15	26
8	15 – 20	17
9	20 – 25	7
10	25 – 30	3

25. Trong kì thi cử nhân, hai nhóm thí sinh I và II đã đạt được những kết quả sau:

	I	II
Đỗ	92	119
Trượt	127	108

Hãy kiểm định giả thiết:

H_0: "Các kết quả thi của thí sinh trong 2 nhóm là tương đương" ở mức $\alpha = 5\%$.

26. Để khảo sát xem màu của mắt và tóc có phải là các đặc tính độc lập hay không, người ta đã quan sát một mẫu gồm 3200 người và có được kết quả:

Mắt	Tóc		
Xanh da trời	vàng	nâu	đen và hung
	872	380	112
Xanh lá cây, nâu	500	815	521

Hãy kiểm định giả thiết: H_0: "Màu tóc và màu mắt là độc lập với nhau" ở mức $\alpha = 5\%$.

27. Nghiên cứu tình hình lao động nông nghiệp ở một nước (X) người ta lấy số liệu ở 2 năm 1969 và 1979 như sau:

Năm	1969	1979
Số lao động nông nghiệp	28300	7840
Số lao động khác	10000	4260

Có thể nói gì về số lao động nông nghiệp ở 2 thời điểm trên, mức kiểm định $\alpha = 5\%$.

28. Tìm hàm hồi quy tuyến tính của Y theo X và tính hệ số tương quan mẫu của X, Y. Tính độ sai dự báo của Y đối với X. Tìm khoảng ước lượng của hệ số tương quan và hệ số hồi quy với độ tin cậy 95% trên cơ sở các mẫu quan sát sau:

a.

X	2	3	7	6	5
Y	6	6	15	13	10

b.

Y / X	1	3	5
2	10	2	
3	5	5	
4		8	10

29. Tìm hàm hồi quy bậc hai mẫu: $y = ax^2 + bx + c$, của Y theo X và tính hệ số tương quan mẫu r của X, Y dựa trên mẫu quan sát sau:

X / Y	2	3	5
25	20		
45		30	
110			48

LỜI GIẢI – HƯỚNG DẪN – TRẢ LỜI

4. Ước lượng $\hat{\lambda} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

5. Ước lượng của p là: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{m}{n}$.

6. Ước lượng của θ là: $\hat{0} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

7. Ước lượng của a và σ^2 là:

$$\hat{a} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;\ S_n^2(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

8. Trung bình mẫu: $\bar{X} = 51{,}50$.

Phương sai mẫu: $S_n^2(X) = 24{,}55$; $S_n(X) = 4{,}985$;

$S_n^{*2}(X) = 27{,}611$; $S_n^*(X) = 5{,}255$.

Khoảng ước lượng của a là: $a = 51{,}5 \pm 3{,}04$.

9. Trung bình mẫu; $\bar{X} = 1{,}508$.

Phương sai mẫu: $S_n^2(X) = 0{,}001656$; $S_n(X) = 0{,}040694$;

$S_n^{*2}(X) = 0{,}001840$; $S_n^*(X) = 0{,}042895$.

Khoảng ước lượng của a là: $a = 1{,}508 \pm 0{,}0306$.

10. Ước lượng điểm của p là: $\hat{p} = \frac{45}{60} = 0{,}75$.

Khoảng ước lượng của p với độ tin cậy 95%: $p = 0{,}75 \pm 0{,}11$.

11. Ước lượng điểm của p là $\hat{p} = 0{,}7$.

Khoảng ước lượng của p với độ tin cậy 95%: $p = 0{,}7 \pm 0{,}00898$.

12. $n = 100$; $p_0 = 0{,}15$; $X = 20$; $x_\alpha = 1{,}96$; $Z = 1{,}40$.

Chấp nhận giả thiết H_0.

13. $n = 480$; $X = 12$; $p_0 = 0,02$; $x_\alpha = 1,96$.

$Z = 0,78$. Chấp nhận giả thiết H_0, nghĩa là tỉ lệ phế phẩm 0,02 là đúng.

14. $S_n^{*2}(X) = 9,111111$; $S_n^*(X) = 0,3333$; $t(5\%; 9) = 2,26$; $|Z| = 7,09$.

Bác bỏ giả thiết H_0.

15. $S_n^{*2}(X) = 1,9569749$; $S_n^*(X) = 1,409521$ $t(5\%; 5) = 2,57$; $d|Z| = 0,5446$.

Chấp nhận giả thiết H_0.

16. $\overline{X} = 3,6$; $\overline{Y} = 3,5$; $t(5\%; 20) = 2,09$; $|Z| = 1,45$

Chấp nhận giả thiết H_0.

17. Chấp nhận giả thiết H_0, nghĩa là trọng lượng của óc người trên 50 tuổi và dưới 50 tuổi là như nhau.

18. $x_\alpha = 1,65$; $|Z| = 10,99$. Bác bỏ giả thiết H_0, tức là số nữ ở xã A nhiều hơn số nữ ở xã B.

19. $x_\alpha = 1,66$; $Z = 6,87$. Bác bỏ giả thiết H_0, tức là phương pháp A có hiệu quả cao hơn phương pháp B.

20. Bác bỏ giả thiết H_0, nghĩa là những người ăn số lượng cam quýt lớn ít bị cảm cúm hơn những người ăn số lượng cam quýt bình thường.

21. Bác bỏ giả thiết H_0, nghĩa là chất lượng 2 loại lốp (X) và (Y) là khác nhau.

22. $m = 23$; $n = 21$; $U = nm + \dfrac{n(n+1)}{2} - w = 208,5$;

$$x_\alpha = 1,96;\ |Z| = \frac{\left|U - \dfrac{nm}{2}\right|}{\sqrt{\dfrac{nm(n+m+1)}{12}}} = 0,775 < x_\alpha = 1,96.$$

Chấp nhận giả thiết H_0.

23. Bác bỏ giả thiết H_0, nghĩa là bệnh đang lây lan, nên cần báo động để ngăn chặn sự phát triển của sâu bệnh của ngô.

24. $\hat{a} \approx \frac{1}{n}\sum_{i=1}^{n} n_i x_i^*$; x_i^* là tâm điểm của khoảng l_i.

$\hat{\sigma}^2 \approx \frac{1}{n}\sum_{i=1}^{n} n_i x_i^{*2} - \hat{a} = 94,26 nm^2$; $\hat{\sigma} = 9,71$.

$C(0,95; 6) = 12,59$ (số bậc tự do $s - 1 - r = 9 - 2 - 1 = 6$).

$Z = 7,09 < C_\alpha = 12,59$. Chấp nhận giả thiết H_0.

25. Bác bỏ giả thiết H_0 vì $v = w = 2$; $C_\alpha = 3,84$; $Z = 4,85 > C_\alpha = 3,84$.

Vậy kết quả thi của 2 nhóm khác nhau, và nhóm II có kết quả tốt hơn nhóm I.

26. $C(5\%; 2) = 6$, $v = 2$, $w = 3$, $Z = 463,95$. Bác bỏ giả thiết H_0, nghĩa là màu của tóc và màu của mắt có sự liên quan với nhau; nó mang đặc tính di truyền.

27. $n = 50846$; $C_\alpha = 3,84$; $Z = 378,37 > C_\alpha$. Bác bỏ giả thiết H_0, tức là tỉ lệ lao động nông nghiệp ở 2 thời điểm 1969 và 1979 là khác nhau; ở năm 1969 có cao hơn năm 1979.

28. a. $\bar{X} = 4,6$; $\bar{Y} = 9,9$; $r = 0,99$.

Hàm hồi quy tuyến tính mẫu của Y theo X là:

$y = 2,07x + 0,28$ ($\hat{a} = 2,07$).

$S_n^2(Y) = 13,2$; $S_n^2(X) = 3,44$. Độ sai dự báo: $a = 2,07 \pm 0,0342$.

Khoảng ước lượng của hệ số tương quan $\rho(X, Y)$ với độ tin cậy 95%:

$\rho(X, Y) = 0,99 \pm 0,0174$.

b. $\bar{X} = 2,75$; $\bar{Y} = 3,15$; $r = 0,78$; $\hat{a} = 0,43$.

Hàm hồi quy tuyến tính mẫu của Y theo X:

$y = 0,43x + 1,98$.

Độ sai dự báo $\hat{\delta} = 0,9545$.

Khoảng ước lượng của hệ số hồi quy a với độ tin cậy 95% là: $a = 0,43 \pm 0,222$.

Khoảng ước lượng của hệ số tương quan ρ với độ tin cậy 95% là:

$\rho = 0,78 \pm 0,189.$

29. $y = 2,94x^2 + 7,27x - 1,25.$

Hệ số tương quan mẫu $r = \dfrac{444000}{\sqrt{85560 \times 13260000}} \approx \dfrac{4440}{10647} \approx 0,417.$

PHỤ LỤC

Bảng 1. Giá trị hàm $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{u^2}{2}} du$

x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)
0,00	0,0000	0,30	0,1179	0,60	0,2257	0,90	0,3159
0,01	0,0040	0,31	0,1217	0,61	0,2291	0,91	0,3186
0,02	0,0080	0,32	0,1255	0,62	0,2324	0,92	0,3212
0,03	0,0120	0,33	0,1293	0,63	0,2357	0,93	0,3238
0,04	0,0160	0,34	0,1331	0,64	0,2389	0,94	0,3264
0,05	0,0199	0,35	0,1368	0,65	0,2422	0,95	0,3289
0,06	0,0239	0,36	0,1406	0,66	0,2454	0,96	0,3315
0,07	0,0279	0,37	0,1443	0,67	0,2486	0,97	0,3340
0,08	0,0319	0,38	0,1480	0,68	0,2517	0,98	0,3365
0,09	0,0359	0,39	0,1517	0,69	0,2549	0,99	0,3389
0,10	0,0398	0,40	0,1554	0,70	0,2580	1,00	0,3413
0,11	0,0438	0,41	0,1591	0,71	0,2611	1,01	0,3438
0,12	0,0478	0,42	0,1628	0,72	0,2642	1,02	0,3461
0,13	0,0517	0,43	0,1664	0,73	0,2673	1,03	0,3485
0,14	0,0557	0,44	0,1700	0,74	0,2703	1,04	0,3508
0,15	0,0596	0,45	0,1736	0,75	0,2743	1,05	0,3631
0,16	0,0636	0,46	0,1772	0,76	0,2764	1,06	0,3554
0,17	0,0675	0,47	0,1808	0,77	0,2794	1,07	0,3577
0,18	0,0714	0,48	0,1844	0,78	0,2823	1,08	0,3599
0,19	0,0753	0,49	0,1879	0,79	0,2852	1,09	0,3621
0,20	0,0793	0,50	0,1915	0,80	0,2881	1,10	0,3643
0,21	0,0832	0,51	0,1950	0,81	0,2910	1,11	0,3665
0,22	0,0871	0,52	0,1985	0,82	0,2939	1,12	0,3686
0,23	0,0910	0,53	0,2019	0,83	0,2967	1,13	0,3708
0,24	0,0948	0,54	0,2054	0,84	0,2995	1,14	0,3729
0,25	0,0987	0,55	0,2088	0,85	0,3023	1,15	0,3749
0,26	0,1026	0,56	0,2123	0,86	0,3051	1,16	0,3770
0,27	0,1064	0,57	0,2157	0,87	0,3078	1,17	0,3790
0,28	0,1103	0,58	0,2190	0,88	0,3106	1,18	0,3810
0,29	0,1141	0,59	0,2224	0,89	0,3133	1,19	0,3830
1,20	0,3849	1,60	0,4452	2,00	0,4772	2,80	0,4974
1,21	0,3869	1,61	0,4463	2,02	0,4783	2,82	0,4976
1,22	0,3883	1,62	0,4474	2,04	0,4793	2,84	0,4977
1,23	0,3907	1,63	0,4484	2,06	0,4803	2,86	0,4979
1,24	0,3925	1,64	0,4495	2,08	0,4812	2,88	0,4980

Bảng 1. Giá trị hàm $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{u^2}{2}} du$ (tiếp theo)

x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)
1,25	0,3944	1,65	0,4505	2,10	0,4821	2,90	0,4981
1,26	0,3962	1,66	0,4515	2,12	0,4830	2,92	0,4982
1,27	0,3980	1,67	0,4525	2,14	0,4838	2,94	0,4984
1,28	0,3997	1,68	0,4535	2,16	0,4846	2,96	0,4985
1,29	0,4015	1,69	0,4545	2,18	0,4854	2,98	0,4986
1,30	0,4032	1,70	0,4554	2,20	0,4861	3,00	0,49865
1,31	0,4049	1,71	0,4564	2,22	0,4868	3,20	0,49931
1,32	0,4066	1,72	0,4573	2,24	0,4875	3,40	0,49966
1,33	0,4082	1,73	0,4582	2,26	0,4881	3,60	0,499841
1,34	0,4099	1,74	0,4591	2,28	0,4887	3,80	0,499928
1,35	0,4115	1,75	0,4599	2,30	0,4893	4,00	0,499968
1,36	0,4131	1,76	0,4608	2,32	0,4898	4,50	0,499997
1,37	0,4147	1,77	0,4616	2,34	0,4904	5,00	0,499997
1,38	0,4162	1,78	0,4625	2,36	0,4909		
1,39	0,4177	1,79	0,4633	2,38	0,4913		
1,40	0,4192	1,80	0,4641	2,40	0,1918		
1,41	0,4207	1,81	0,4649	2,42	0,4922		
1,42	0,4222	1,82	0,4656	2,44	0,4927		
1,43	0,4236	1,83	0,4664	2,46	0,4931		
1,44	0,4251	1,84	0,4671	2,48	0,4934		
1,45	0,4265	1,85	0,4678	2,50	0,4938		
1,46	0,4279	1,86	0,4686	2,52	0,4941		
1,47	0,4292	1,87	0,4693	2,54	0,4945		
1,48	0,4306	1,88	0,4699	2,56	0,4948		
1,49	0,4319	1,89	0,4706	2,58	0,4951		
1,50	0,4332	1,90	0,4713	2,60	0,4953		
1,51	0,4345	1,91	0,4719	2,62	0,4956		
1,52	0,4357	1,92	0,4726	2,64	0,4959		
1,53	0,4370	1,93	0,4732	2,66	0,4961		
1,54	0,4382	1,94	0,4738	2,68	0,4963		
1,55	0,4394	1,95	0,4744	2,70	0,4965		
1,56	0,4406	1,96	0,4750	2,72	0,4967		
1,57	0,4418	1,97	0,4756	2,74	0,4969		
1,58	0,4429	1,98	0,4762	2,76	0,4971		
1,59	0,4441	1,99	0,4767	2,78	0,4973		

Bảng 2. Phân phối Student $P[|T| > t_\alpha] = \alpha$.

Số bậc tự do	Mức ý nghĩa α (tiêu chuẩn hai phía)					
k	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,3	6,97	9,92	22,33	3,16
3	2,33	3,18	4,54	5,84	10,22	1,29
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,79
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,71	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	266	3,23	3,46
120	1,66	1,98	2,36	262	3,17	3,37
∞	1,64	1,96	2,33	258	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005

Mức ý nghĩa α (tiêu chuẩn một phía).

Bảng 3: Phân phối khi bình phương với n bậc tự do $P[X > C_\alpha] = \alpha$

Bậc tự do n	Xác suất															
	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
1	0,0001	0,0006	0,0009	0,016	0,064	0,148	0,455	1,07	1,64	2,7	3,84	5,5	6,6	7,9	9,5	10,83
2	0,020	0,040	0,103	0,211	0,446	0,173	1,386	2,41	3,22	4,6	6,0	7,8	9,2	11,6	12,4	13,8
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,66	4,64	6,3	7,8	9,8	11,3	12,8	14,8	16,3
4	0,30	0,43	0,71	0,06	1,65	2,19	3,36	4,9	6,0	7,8	9,5	11,7	13,3	14,9	16,9	18,5
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,1	7,3	9,2	11,1	13,4	15,1	16,3	18,9	20,5
6	0,187	1,13	1,63	2,20	3,07	3,83	5,35	7,2	8,6	10,6	12,6	15,0	16,8	18,6	20,7	22,5
7	1,24	1,56	2,17	2,83	3,82	4,67	6,34	8,4	9,8	12,0	14,1	16,6	18,5	20,3	22,6	24,3
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,5	11,0	13,4	15,5	18,2	20,1	21,9	24,3	21,6
9	2,09	2,53	3,32	4,17	5,38	6,39	8,35	10,7	12,2	14,7	16,9	19,7	21,7	23,6	26,1	27,9
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8	13,4	16,0	18,3	21,2	23,2	25,2	27,7	29,6
11	3,1	3,6	4,6	5,6	7,0	8,1	10,3	12,9	14,6	17,3	19,7	22,6	24,7	26,8	29,4	31,3
12	3,6	4,2	5,2	6,3	7,8	9,0	11,3	14,0	15,8	18,5	21,0	24,1	26,2	28,3	31,0	32,9
13	4,1	4,8	5,9	7,0	8,6	9,9	12,3	15,1	17,0	19,8	22,4	25,5	27,7	29,8	32,5	34,5
14	4,7	5,4	6,6	7,8	9,5	10,8	13,3	16,2	18,2	21,1	23,7	26,9	29,1	31,0	34,0	36,1
15	5,2	6,0	7,3	8,5	10,3	11,7	14,3	17,3	19,3	22,3	25,0	28,3	30,6	32,5	35,5	37,7
16	5,8	6,6	8,0	9,3	11,2	12,6	15,3	18,4	20,5	23,5	26,3	29,6	32,0	34,0	37,0	39,2
17	6,4	7,3	8,7	10,1	12,0	13,5	16,3	19,5	21,6	24,8	27,6	31,0	33,4	35,5	38,5	40,8
18	7,0	7,9	9,4	10,9	12,9	14,4	17,3	20,6	22,8	26,0	28,9	32,3	34,8	37,0	40,0	42,3
19	7,6	8,6	10,1	11,7	13,7	15,4	18,3	21,7	23,9	27,2	30,1	33,7	36,2	38,5	41,5	43,8
20	8,3	9,2	10,9	12,4	14,6	16,3	19,3	22,8	25,0	28,4	31,4	35,0	37,6	40,0	43,0	45,3
21	8,9	9,9	11,6	13,2	15,4	17,2	20,3	23,9	26,2	29,6	32,7	36,3	38,9	41,5	44,5	46,8
22	9,5	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3	30,8	33,9	37,7	40,3	42,0	46,0	48,3
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0	28,4	32,0	35,2	39,0	41,6	44,0	47,5	49,7
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1	29,6	33,	36,4	40,3	43,0	45,5	48,5	51,2
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,1	30,7	34,4	37,7	41,6	44,3	47,0	50,0	52,6
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,3	31,8	35,6	38,9	42,9	45,6	48,0	51,5	54,1
27	12,9	14,1	16,2	18,1	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0	49,5	53,0	55,5
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3	51,0	54,5	56,9
29	14,3	15,6	17,7	19,8	22,5	24,6	28,3	32,5	35,1	39,1	42,6	46,7	49,6	52,5	56,0	58,3
30	15,0	16,3	18,5	20,6	23,4	25,5	29,3	33,5	36,3	40,3	43,8	48,0	50,9	54,0	57,5	59,7

Bảng 4: Phân phối F (Fisher-Snedecor) mức ý nghĩa $\alpha = 0,01$.

K_2 \ K_1	1	2	3	4	5	6	7	8	9	10	11	12
1	4052	4999	5403	5621	5764	5889	5928	5981	6022	6056	6082	6106
2	98,49	99,01	99,17	99,25	99,30	99,33	99,37	99,36	99,38	99,40	99,41	99,42
3	34,12	30,81	29,48	27,71	28,34	27,91	27,67	27,49	27,34	27,23	27,13	27,05
4	21,20	18,00	16,69	15,98	15,52	14,21	14,98	14,80	14,66	14,55	14,45	14,37
5	16,26	13,27	12,06	11,30	10,97	10,67	10,45	10,27	10,15	10,55	9,96	9,89
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87.	7,79	7,72
7	12,25	9,55	8,45	7,89	7,46	7,19	7,00	6,48	6,71	6,62	6,54	6,47
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74	5,67
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,31	5,26	5,18	5,11
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71
11	9,85	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46	4,40
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4016
13	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55
20	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45	3,37	3,30	3,23
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,25	3,17	3,09	3,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,06	2,98	2,90	2,84
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,88	2,80	2,73	2,66
50	7,17	5,06	4,20	3,72	3,41	3,18	3,02	2,88	2,78	2,70	2,62	2,56
70	7,01	4,92	4,08	3,60	3,29	3,07	2,91	2,77	2,67	2,59	2,51	2,45
100	6,90	4,82	3,98	3,51	3,20	2,99	2,82	2,69	2,59	2,51	2,43	2,36
200	6,76	4,71	3,88	3,41	3,11	2,90	2,73	2,60	2,50	2,41	2,34	2,28
400	6,70	4,66	3,83	3,36	3,06	2,85	2,69	2,55	2,46	2,37	2,29	2,23
∞	6,64	4,60	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,24	2,18

Phân phối F (Fisher-Snedecor) mức ý nghĩa $\alpha = 0,01$ (tiếp bảng 4)

K_2 \ K_1	14	16	20	24	30	40	50	75	100	200	500	∞
1	6142	6169	6208	6234	6258	6286	6302	6323	6334	6352	6361	6366
2	99,43	99,44	99,45	99,46	99,47	99,48	99,48	99,49	99,49	99,49	99,50	99,50
3	26,92	26,83	26,69	26,60	26,50	26,41	26,35	26,27	36,23	26,18	26,14	26,12
4	14,24	14,15	14,02	13,93	13,83	13,74	13,69	13,61	13,57	13,52	13,48	13,46
5	9,77	9,68	9,55	9,47	9,38	9,29	9,24	9,17	9,13	9,07	9,04	9,02
6	7,60	7,52	7,39	7,31	7,23	7,14	7,09	7,02	6,99	6,94	6,90	6,88
7	6,35	6,27	6,15	6,07	5,98	5,90	5,85	5,78	5,75	5,70	5,67	5,65
8	5,56	5,48	5,36	5,28	5,20	5,11	5,11	5,06	5,00	4,96	4,88	4,86
9	5,00	4,92	4,80	4,73	4,64	4,56	4,51	4,45	4,41	4,36	4,33	4,31
10	4,60	4,52	4,41	4,33	4,25	4,17	4,12	4,05	4,01	3,96	3,93	3,91
11	4,29	4,21	4,10	4,02	3,94	3,86	3,80	3,74	3,70	3,66	3,62	3,66
12	4,05	3,98	3,86	3,78	3,70	3,61	3,56	3,49	3,46	3,41	3,38	3,36
13	3,70	3,62	3,51	3,43	3,34	3,26	3,21	3,14	3,11	3,06	3,02	3,00
16	3,45	3,37	3,25	3,18	3,01	3,01	2,96	2,89	2,86	2,80	2,77	2,75
20	3,13	3,05	2,94	2,86	2,77	2,69	2,63	2,56	2,53	2,47	2,44	2,42
24	2,93	3,85	2,74	2,66	2,58	2,49	2,44	2,36	2,33	2,27	2,23	2,21
30	2,74	2,66	2,55	2,47	2,38	2,29	2,24	2,16	2,13	2,07	2,03	2,01
40	2,56	2,49	2,37	2,29	2,20	2,11	2,05	1,97	1,94	1,88	1,84	1,81
50	2,46	2,39	2,26	2,18	2,10	2,00	1,94	1,86	1,82	1,76	1,71	1,68
70	2,35	2,28	2,15	2,07	1,98	1,88	1,82	1,74	1,69	1,62	1,56	1,53
100	2,26	2,19	2,06	1,98	1,89	1,79	1,73	1,64	1,59	1,51	1,46	1,43
200	2,17	2,09	1,97	1,88	1,79	1,69	1,62	1,53	1,48	1,39	1,33	1,28
400	2,12	2,04	1,92	1,84	1,74	1,64	1,57	1,47	1,42	1,32	1,24	1,19
∞	2,07	1,99	1,87	1,79	1,69	1,59	1,52	1,41	1,36	1,25	1,15	1,09

Bảng 5: Phân phối F (Fisher-Snedecor) mức ý nghĩa $\alpha = 0,05$

K_2 \ K_1	1	2	3	4	5	6	7	8	9	10	11
1	161	200	216	225	230	234	237	239	241	242	243
2	18,5	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40
3	10,13	9,55	9,28	9,12	9,01	8,44	8,88	8,84	8,81	8,78	8,76
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70
6	5,99	5,14	4,76	4,53	4,39	4,21	4,15	4,10	4,10	4,06	4,03
7	5,59	4,74	4,35	4,12	3,97	3,79	3,73	3,68	3,68	3,63	3,60
8	5,32	4,46	4,07	3,84	3,69	3,50	3,44	3,39	3,39	3,34	3,31
9	5,12	4,26	3,86	3,63	3,48	3,29	3,23	3,18	3,18	3,13	3,10
10	4,96	4,10	3,71	3,48	3,22	3,14	3,07	3,02	3,02	2,97	2,94
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72
13	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,26	2,21	2,16	2,12
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,04
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	2,98
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,01	1,97	1,93
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,93	1,88
200	3,89	3,04	2,65	2,41	2,26	214	2,05	1,98	1,92	1,87	1,83
400	3,86	3,02	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,81
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,79

Bảng 5: Phân phối F (Fisher-Snedecor) mức ý nghĩa $\alpha = 0,05$ (tiếp theo)

K_2 \ K_1	12	13	16	20	30	40	50	75	100	200	500	∞
1	244	245	246	248	250	251	252	253	253	254	254	254
2	19,41	19,42	19,43	19,44	19,46	19,47	19,47	19,48	19,49	19,49	19,50	19,50
3	8,74	8,71	8,66	8,66	8,62	8,60	8,58	8,57	8,59	8,54	8,54	8,53
4	5,91	5,87	5,84	5,80	5,74	5,71	5,70	5,68	5,66	5,65	5,64	5,63
5	4,68	4,64	4,60	4,56	4,50	4,46	4,44	4,42	4,40	4,38	4,38	1,36
6	4,00	3,96	3,92	3,87	3,81	3,77	3,75	3,72	3,71	3,69	3,68	3,67
7	3,57	3,52	3,49	3,44	3,38	3,34	3,32	3,29	3,28	3,25	3,24	2,33
8	3,28	3,23	3,20	3,15	3,08	3,05	3,03	3,00	2,98	2,96	2,94	2,93
9	3,07	3,02	2,98	2,93	2,86	2,82	2,80	2,77	2,76	2,73	2,72	3,71
10	2,91	2,86	2,82	2,77	2,70	2,67	2,64	2,61	2,59	2,56	2,55	2,54
11	2,79	2,74	2,70	2,65	2,57	2,53	2,50	2,47	2,45	2,42	2,41	2,40
12	2,69	2,64	2,60	2,54	2,46	2,42	2,40	2,36	2,35	2,32	2,30	2,30
13	2,53	2,48	2,44	2,39	2,31	2,27	2,24	2,21	2,19	2,16	2,14	2,13
16	2,42	2,37	2,33	2,28	2,20	2,16	2,13	2,09	2,07	2,04	2,02	2,01
20	2,98	2,23	2,18	2,12	2,04	1,99	1,96	1,92	1,90	1,87	1,85	1,84
24	2,18	2,13	2,09	2,02	1,94	1,89	1,86	1,82	1,80	1,76	1,74	1,73
30	2,09	2,04	1,99	1,93	1,84	1,79	1,76	1,72	1,69	1,66	1,64	1,62
40	2,00	1,95	1,90	1,84	1,74	1,69	1,66	1,61	1,59	1,55	1,53	1,51
50	1,95	1,90	1,85	1,78	1,69	1,63	1,60	1,55	1,52	1,18	1,46	1,44
70	1,89	1,84	1,79	1,72	1,62	1,46	1,53	1,47	1,45	1,40	1,37	1,35
100	1,85	1,79	1,75	1,68	1,57	1,51	1,48	1,42	1,39	1,34	1,30	1,28
200	1,80	1,74	1,69	1,62	1,52	1,45	1,42	1,35	1,32	1,26	1,22	1,19
400	1,78	1,72	1,67	1,60	1,49	1,42	1,38	1,32	1,28	1,22	1,16	1,13
∞	1,75	1,69	1,64	1,57	1,46	1,40	1,35	1,28	1,24	1,17	1,11	1,00

Bảng 6: Giá trị tiêu chuẩn Wilcoxon với $n \leq 20$ (so sánh cặp đôi).

Tiêu chuẩn một phía.

n	$\alpha = 0,05$		$\alpha = 0,025$		$\alpha = 0,01$		$\alpha = 0,005$	
	W_α	$W_{1-\alpha}$	W_α	$W_{1-\alpha}$	W_α	$W_{1-\alpha}$	W_α	$W_{1-\alpha}$
6	3	17	1	20	0	21	0	21
7	4	24	3	25	0	21	0	21
8	6	30	4	32	2	34	1	35
9	9	36	6	39	4	41	2	43
10	11	14	9	36	6	49	4	51
11	14	52	11	55	8	58	6	60
12	18	60	14	64	10	68	8	70
13	22	69	18	73	13	78	10	81
14	26	79	22	83	16	89	13	92
15	31	89	16	94	20	100	16	104
16	36	100	30	106	24	112	20	116
17	42	111	35	118	28	125	24	129
18	48	123	41	130	33	138	28	143
19	54	136	47	143	38	152	33	157
20	61	149	53	157	44	166	38	172

Tiêu chuẩn hai phía:

$\alpha = 0,1$		$\alpha = 0,05$		$\alpha = 0,01$	
$W_{\frac{\alpha}{2}}$	$W_{1-\frac{\alpha}{2}}$	$W_{\frac{\alpha}{2}}$	$W_{1-\frac{\alpha}{2}}$	$W_{\frac{\alpha}{2}}$	$W_{1-\frac{\alpha}{2}}$

Kích thước mẫu		Q				Kích thước mẫu		Q			
n_1	n_2	0,005	0,01	0,025	0,05	n_1	n_2	0,005	0,01	0,025	0,05
6	6	23	24	26	28	7	20	52	56	62	67
	7	24	25	27	30		21	53	58	64	69
	8	25	27	29	31		22	55	59	66	72
	9	26	28	31	33		23	57	61	68	74
	10	27	29	32	35		24	58	63	70	76
	11	28	30	34	37		25	60	64	72	78
	12	30	32	35	38		8	43	45	49	51
	13	31	33	37	40		9	45	47	51	54
	14	32	34	38	42		10	47	49	53	56
	15	33	36	40	44		11	49	51	55	59
	16	34	37	42	46		12	51	53	58	62
	17	36	39	43	47		13	53	56	60	64
	18	37	40	45	49		14	54	58	62	67
	19	38	41	46	51		15	56	60	65	69
	20	39	43	48	53		16	58	62	67	72
	21	40	44	50	55		17	60	64	70	75
	22	42	45	51	57		18	62	66	72	77
	23	43	47	53	58		19	64	68	74	80
	24	44	48	54	60		20	66	70	77	83
	25	45	50	56	62		21	68	72	79	85
	7	32	34	56	39		22	70	74	81	88
	8	34	35	38	41		23	71	76	84	90
	9	35	37	40	43		24	73	78	86	93
	10	37	39	42	45		25	75	81	89	96
	11	38	40	44	47		9	56	59	62	66
	12	40	42	46	49		10	58	61	65	69
	13	41	44	48	52		11	61	63	68	72
	14	43	45	50	54		12	63	66	71	75
	15	44	47	52	56		13	65	68	73	78
	16	46	49	54	58		14	67	71	76	81
	17	47	50	56	61		15	69	73	79	85
	18	49	52	58	63		16	72	76	82	87
	19	50	54	60	65						

TÀI LIỆU THAM KHẢO

1. Phạm Văn Kiều.
   Lí thuyết xác suất và thống kê toán học.
   Nhà xuất bản Đại học Sư phạm. Hà Nội (1993).
2. Phạm Văn Kiều.
   Giáo trình lí thuyết xác suất và thống kê toán.
   (Dùng cho ngành sinh vật, nông nghiệp, kinh tế, v.v...).
   Nhà xuất bản Khoa học và Kỹ thuật (1998).
3. **A. Rényi.** ***Probability theory.***
   Hungarian Academy of Sciences, Budapest (1970).
4. V. I Gnedenko.
   A course of probability theory.
   Moscow (1962).
5. Seymour Lipschutz, Marc Lars Lipson.
   Theory and Problems of Probability.
   Second Edition New York, San Francisco, Washington (2000).
6. Hoàng Hữu Như, Nguyễn Văn Hữu.
   Bài tập lí thuyết xác suất và thống kê toán.
   Nhà xuất bản Đại học và Trung học chuyên nghiệp (1976).
7. Phạm Văn Kiều, Lê Thiên Hương.
   Xác suất thống kê.
   Nhà xuất bản Giáo dục (2001).