

ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC SƯ PHẠM
Lê Văn Dũng

Giáo trình
THỐNG KÊ TOÁN

NHÀ XUẤT BẢN THÔNG TIN VÀ TRUYỀN THÔNG

MỤC LỤC

Chương 1. Nhắc lại một số nội dung về lý thuyết xác suất	7
1.1. Một số phân phối xác suất quan trọng	7
1.1.1. Phân bố Bernoulli	7
1.1.2. Phân bố nhị thức	7
1.1.3. Phân bố Poisson	8
1.1.4. Phân bố mũ	10
1.1.5. Phân bố đều	11
1.1.6. Phân bố chuẩn	11
1.1.7. Phân bố khi bình phương	14
1.1.8. Phân bố F	14
1.2. Các định lý giới hạn	15
1.2.1. Luật số lớn	15
1.2.2. Định lý giới hạn trung tâm	16
Chương 2. THỐNG KÊ MÔ TẢ	17
2.1. Khái niệm mẫu và tổng thể	17
2.2. Các số đặc trưng của một mẫu số liệu	17
2.2.1. Trung bình, phương sai và độ lệch chuẩn mẫu	17
2.2.2. Trung vị mẫu	19
2.2.3. Hệ số tương quan mẫu	19
2.3. Biểu đồ	19
2.3.1. Biểu đồ phân bố tần số	19
2.3.2. Biểu đồ thân - lá	21
2.3.3. Biểu đồ xác suất chuẩn	22
2.4. Mẫu ngẫu nhiên	23
2.5. Chọn mẫu ngẫu nhiên đơn giản	24
2.5.1. Chọn mẫu từ tổng thể hữu hạn	24
2.5.2. Chọn mẫu từ tổng thể vô hạn	25
2.6. Phân bố của trung bình mẫu	26
Bài tập chương 2	27
Chương 3. ƯỚC LƯỢNG THAM SỐ	31
3.1. Ước lượng điểm	31

3.1.1. Ước lượng điểm và hàm ước lượng	31
3.1.2. Ước lượng không chệch	31
3.1.3. Ước lượng không chệch của kì vọng và phương sai	31
3.1.4. Ước lượng không chệch tỉ lệ	32
3.1.5. Phương pháp ước lượng hợp lí cực đại	32
3.1.6. Phương pháp mô men	33
3.2. Nguyên lí xác suất nhỏ và nguyên lí xác suất lớn	34
3.3. Khoảng tin cậy cho kì vọng	35
3.3.1. $X \sim N(\mu; \sigma^2)$ với σ^2 đã biết	36
3.3.2. $X \sim N(\mu; \sigma^2)$ với σ^2 chưa biết	37
3.4. Khoảng tin cậy cho tỉ lệ	42
Bài tập chương 3	44
Chương 4. KIỂM ĐỊNH GIẢ THUYẾT.....	47
4.1. Khái niệm chung	47
4.1.1. Giả thuyết thống kê	47
4.1.2. Sai lầm loại I và sai lầm loại II	49
4.2. Kiểm định kì vọng của phân phối chuẩn	49
4.2.1. Đã biết phương sai	49
4.2.2. Chưa biết phương sai	54
4.3. So sánh 2 kì vọng	59
4.3.1. Cỡ mẫu lớn	59
4.3.2. Cỡ mẫu nhỏ và hai phương sai bằng nhau	60
4.3.3. Cỡ mẫu nhỏ và hai phương sai không bằng nhau	63
4.4. So sánh cặp	65
4.5. Kiểm định giả thuyết về tỉ lệ	68
4.6. So sánh hai tỉ lệ	71
Bài tập chương 4	72
Chương 5. KIỂM ĐỊNH KHI BÌNH PHƯƠNG	77
5.1. Kiểm định giả thuyết về quy luật phân phối	77
5.1.1. Biến ngẫu nhiên rời rạc	77
5.1.2. Biến ngẫu nhiên liên tục	78
5.2. Kiểm định tính độc lập	80
5.3. Kiểm định phù hợp	84
Bài tập chương 5	85
Chương 6. PHÂN TÍCH PHƯƠNG SAI	89
6.1. Phân tích phương sai một nhân tố	89
6.2. Phân tích phương sai hai nhân tố	93
6.2.1. Phân tích phương sai hai nhân tố không lặp lại	93

6.2.2. Phân tích phương sai hai nhân tố có lặp	97
6.3. Đại cương về bố trí thí nghiệm	103
6.3.1. Một số khái niệm	103
6.3.2. Hai nguyên tắc cơ bản về bố trí thí nghiệm	103
6.3.3. Kỹ thuật ngẫu nhiên hoá	103
6.3.4. Các kiểu bố trí thí nghiệm phổ biến	103
Bài tập chương 6	105
Chương 7. KIỂM ĐỊNH PHI THAM SỐ.....	111
7.1. Kiểm định dấu	111
7.1.1. So sánh hai trung vị	114
7.2. Kiểm định hạng có dấu Wilcoxon	115
7.2.1. Kiểm định trung vị	115
7.2.2. So sánh 2 trung vị	118
7.3. Kiểm định đoạn mạch	119
7.3.1. Kiểm định ngẫu nhiên	119
7.3.2. Kiểm định đoạn mạch kết hợp với trung bình	120
7.4. So sánh hai phân bố xác suất	121
7.4.1. Kiểm định Mann - Whitney	121
7.4.2. Kiểm định đoạn mạch	122
7.5. Phân tích phương sai (Kiểm định Kruskal-Wallis)	122
Bài tập chương 7	123
<i>Bảng phụ lục.....</i>	<i>129</i>
<i>Tài liệu tham khảo</i>	<i>137</i>

Chương 1

NHẮC LẠI MỘT SỐ NỘI DUNG VỀ LÝ THUYẾT XÁC SUẤT

1.1. Một số phân phối xác suất quan trọng

1.1.1. Phân bố Bernoulli

Định nghĩa 1.1. Biến ngẫu nhiên rời rạc X có *phân bố Bernoulli* với tham số p ($0 < p < 1$) nếu X có miền giá trị $X(\Omega) = \{0, 1\}$ và hàm xác suất:

$$p(k) = P(X = k) = \begin{cases} 1 - p & \text{nếu } k = 0 \\ p & \text{nếu } k = 1 \\ 0 & \text{nếu } k \notin \{0, 1\} \end{cases}$$

Kí hiệu: $X \sim \text{Ber}(p)$.

Tính chất 1.1. Nếu $X \sim \text{Ber}(p)$ thì $E(X) = p$ và $V(X) = p(1 - p)$.

1.1.2. Phân bố nhị thức

Định nghĩa 1.2. Biến ngẫu nhiên rời rạc X có *phân bố nhị thức* với tham số n và p ($n \in \mathbb{N}^*$ và $0 < p < 1$) nếu X có miền giá trị $X(\Omega) = \{0, 1, \dots, n\}$ và hàm xác suất:

$$p(k) = C_n^k p^k (1 - p)^{n-k}, k \in X(\Omega).$$

Kí hiệu: $X \sim B(n, p)$.

Tính chất 1.2.

1) Nếu $X \sim B(n, p)$ thì $E(X) = np$ và $V(X) = np(1 - p)$.

2) Nếu $X_1, X_2, \dots, X_n$ là n biến ngẫu nhiên độc lập, cùng phân bố xác suất với $X \sim Ber(p)$ thì biến ngẫu nhiên $T = X_1 + X_2 + \dots + X_n$ có phân bố nhị thức $B(n, p)$.

Ví dụ 1.1. Tỷ lệ phế phẩm của một nhà máy là 12%. Các sản phẩm của nhà máy được đóng gói thành từng hộp, mỗi hộp 20 sản phẩm.

a) Trung bình mỗi hộp chứa bao nhiêu phế phẩm? Tính độ lệch chuẩn số phế phẩm trong mỗi hộp.

b) Một khách hàng mua ngẫu nhiên một hộp sản phẩm. Tính xác suất hộp đó có chứa phế phẩm.

c) Tìm số phế phẩm trong hộp có xác suất lớn nhất.

Giải. Gọi X là số phế phẩm trong mỗi hộp. Khi đó, $X \sim B(20; 0, 12)$.

a) $E(X) = np = 2,4$; $SD(X) = \sqrt{np(1 - p)} \approx 1,45$.

b) $P(X \geq 1) = 1 - P(X = 0) = 1 - 0,88^{20}$.

c) $(n + 1)p = 2,52 \notin \mathbb{Z}$ nên $P(X = k) = C_{20}^k \cdot 0,12^k \cdot 0,88^{20-k}$ đạt giá trị lớn nhất tại $k = 2$. ■

1.1.3. Phân bố Poisson

Định nghĩa 1.3. Biến ngẫu nhiên rời rạc X có phân bố Poisson với tham số λ ($\lambda > 0$) nếu X có miền giá trị $\mathbb{N} = \{0, 1, 2, \dots\}$ và hàm xác suất:

$$p(k) = P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k \in \mathbb{N}$$

Kí hiệu: $X \sim Poi(\lambda)$.

Phân bố Poisson thường gặp thể hiện phân bố số lần xuất hiện 1 biến cố nào đó trong một khoảng thời gian T .

Tính chất 1.3.

1) Nếu $X \sim Poi(\lambda)$ thì $E(X) = \lambda$, $V(X) = \lambda$.

2) Nếu $X_1, X_2, \dots, X_n$ là n biến ngẫu nhiên độc lập, cùng phân bố với $X \sim Poi(\lambda)$ thì biến ngẫu nhiên $T = X_1 + X_2 + \dots + X_n$ có phân bố Poisson $Poi(n\lambda)$.

Ví dụ 1.2. Một gara cho thuê xe ô tô có 2 ô tô loại A. Số đơn đặt hàng ô tô loại này vào ngày cuối tuần có phân bố Poisson với số đơn trung bình 2 đơn/ngày. Tính xác suất trong ngày cuối tuần:

a) có một ô tô loại A được thuê.

b) có 2 ô tô loại A được thuê.

c) gara không đáp ứng nhu cầu thuê ô tô loại này.

Giải. Gọi X là số đơn đặt hàng thuê ô tô ngày cuối tuần của gara. Ta có $X \sim Poi(2)$ (do $E(X) = \lambda = 2$).

a) $P(X = 1) = e^{-2} \frac{2^1}{1!} \approx 0,27$.

- b) $P(X \geq 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1) = 1 - e^{-2} - e^{-2} \frac{2^1}{1!} \approx 0,59$.
 c) $P(X > 2) = 1 - P(X \leq 2) \approx 0,32$. ■

Ví dụ 1.3. Ở một tổng đài bưu điện, số cuộc điện thoại gọi đến xuất hiện là biến ngẫu nhiên có phân bố Poisson với số cuộc điện thoại trung bình là 2 cuộc gọi trong 1 phút. Tính xác suất có đúng 5 cuộc trong khoảng thời gian 1 phút.

Giải. Gọi X là số cuộc điện thoại gọi đến trong khoảng thời gian 1 phút, theo giả thiết, X có phân bố Poisson. Vì $E(X) = 2$ nên $\lambda = 2$. Do đó:

$$P(X = 5) = e^{-2} \cdot \frac{2^5}{5!} \approx 0,036$$

Định lý 1.1. (Luật biến cố hiếm) Cho $\{X_n; n \geq 1\}$ là dãy biến ngẫu nhiên có phân bố nhị thức $X_n \sim B(n; p_n)$. Nếu tồn tại giới hạn $\lim_{n \rightarrow \infty} np_n = \lambda$ thì:

$$\lim_{n \rightarrow \infty} P(X_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

Chứng minh. Ta có:

$$\begin{aligned} P(X_n = k) &= C_n^k p_n^k (1 - p_n)^{n-k} \\ &= \frac{n(n-1)\dots(n-k+1)}{k!} p_n^k (1 - p_n)^{n-k} \\ &= \frac{(np_n)^k}{k!} \left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) (1 - p_n)^n (1 - p_n)^{-k} \end{aligned}$$

Ta lại có:

$$\lim_{n \rightarrow \infty} \frac{(np_n)^k}{k!} = \frac{\lambda^k}{k!}$$

Đặt $\lambda_n = np_n$. Khi đó:

$$\lim_{n \rightarrow \infty} (1 - p_n)^n = \lim_{n \rightarrow \infty} \left[\left(1 - \frac{\lambda_n}{n}\right)^{\frac{n}{\lambda_n}} \right]^{-\lambda_n} = e^{-\lambda}$$

Các thừa số khác có giới hạn bằng 1. Từ đó ta có điều phải chứng minh. ■

Ứng dụng: Nếu $X \sim B(n; p)$ với n khá lớn và p khá bé thì X có xấp xỉ phân bố Poisson với $\lambda = np$, tức là:

$$P(X = k) = C_n^k p^k (1 - p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

Ví dụ 1.4. Tỷ lệ phế phẩm của một nhà máy là 0,006. Lấy ngẫu nhiên 1.000 sản phẩm của nhà máy, tính xác suất có đúng 9 phế phẩm.

Giải. Gọi X là số phế phẩm trong 1.000 sản phẩm, khi đó $X \sim B(1.000; 0,006)$. Vì $n = 1.000$ khá lớn và $p = 0,006$ khá bé nên ta có thể tính bằng xấp xỉ phân bố Poisson với $\lambda = np = 6$:

$$P(X = 9) \approx e^{-6} \cdot \frac{6^9}{9!} \approx 0,069$$

Ví dụ 1.5. Một xưởng in sách thấy rằng trung bình một cuốn sách 500 trang có chứa 300 lỗi. Tìm xác suất trong một trang:

- Có đúng 2 lỗi.
- Có ít nhất 2 lỗi.

Giải. Gọi p là xác suất một chữ bị lỗi, X là số lỗi trong 1 trang có n chữ. Khi đó $X \sim B(n; p)$ và $E(X) = np = 300/500 = 0,6$. Vì xác suất 1 chữ bị lỗi rất nhỏ và số chữ trong 1 trang rất lớn nên có thể xấp xỉ X bởi phân bố Poisson với $\lambda = 0,6$. Do đó:

- $P(X = 2) = \frac{0,6^2}{2!} e^{-0,6} \approx 0,099$.
- $P(X \geq 2) = 1 - P(X < 2) \approx 1 - 0,549 - 0,359 = 0,122$. ■

1.1.4. Phân bố mũ

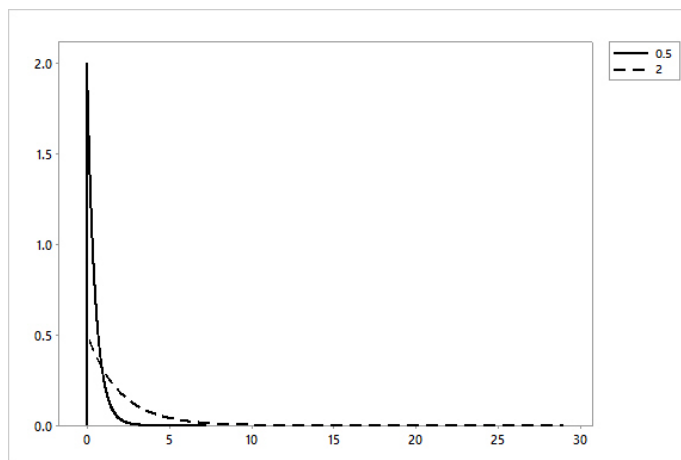
Định nghĩa 1.4. Biến ngẫu nhiên liên tục X có *phân bố mũ* với tham số λ ($\lambda > 0$) nếu có hàm mật độ

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{nếu } x \geq 0 \\ 0 & \text{nếu } x < 0 \end{cases}$$

Kí hiệu: $X \sim \text{Exp}(\lambda)$.

Trong cuộc sống, phân bố mũ thể hiện phân bố thời gian sống của các đối tượng,...

Tính chất 1.4. Nếu $X \sim \text{Exp}(\lambda)$ thì $E(X) = 1/\lambda$, $V(X) = 1/\lambda^2$.



Hình 1.1: Phân bố mũ ($\lambda = 0,5$ và $\lambda = 2$)

Ví dụ 1.6. Giả sử tuổi thọ (X) của một chiếc quạt trong máy tính là một biến ngẫu nhiên phân bố mũ với tuổi thọ trung bình là 3.300 giờ. Tính xác suất:

- Chiếc quạt hỏng trước 10.000 giờ.
- Chiếc quạt có tuổi thọ lớn hơn 7.000 giờ.

Giải. Theo giả thiết $E(X) = \frac{1}{\lambda} = 3.300$ nên:

$$a) P(X < 10.000) = \int_0^{10.000} \frac{1}{3.300} e^{-x/3.300} dx \approx 0,952.$$

$$b) P(X > 7.000) = 1 - P(X \leq 7.000) = \int_0^{7.000} \frac{1}{3.300} e^{-x/3.300} dx \approx 0,88. \quad \blacksquare$$

1.1.5. Phân bố đều

Định nghĩa 1.5. Biến ngẫu nhiên liên tục X có *phân bố đều* trên đoạn $[a; b]$ ($a < b$) nếu có hàm mật độ xác suất:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{nếu } x \in [a; b] \\ 0 & \text{nếu } x \notin [a; b] \end{cases}$$

Kí hiệu: $X \sim U([a; b])$.

Tính chất 1.5. Nếu $X \sim U([a; b])$ thì $E(X) = \frac{a+b}{2}$, $V(X) = \frac{(b-a)^2}{12}$.

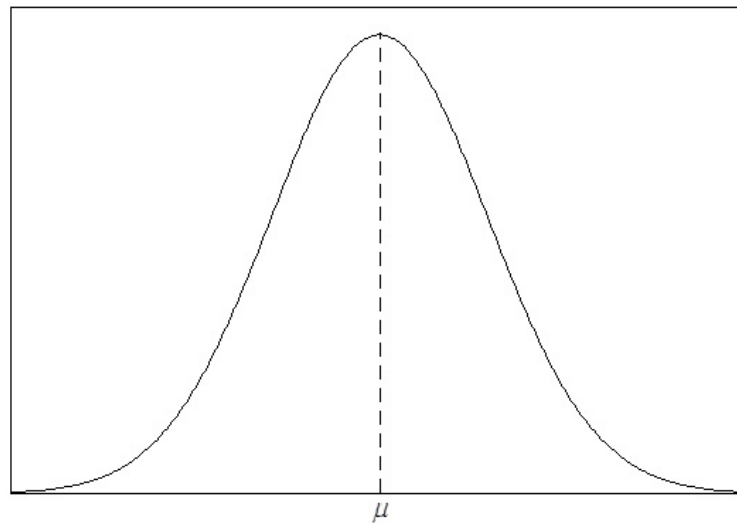
1.1.6. Phân bố chuẩn

Định nghĩa 1.6. Biến ngẫu nhiên liên tục X có *phân bố chuẩn* với tham số μ và σ ($-\infty < \mu < +\infty$, $\sigma > 0$) nếu có hàm mật độ xác suất:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

Kí hiệu $X \sim N(\mu, \sigma^2)$.

Dưới đây là hình dáng đồ thị của hàm mật độ xác suất $f(x)$:



Hình 1.2

Phân bố chuẩn tắc

Biến ngẫu nhiên phân bố chuẩn với $\mu = 0$ và $\sigma = 1$ được gọi là *phân bố chuẩn tắc* và kí hiệu là Z . Khi đó, hàm mật độ xác suất được kí hiệu là $\varphi(x)$,

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Hàm phân bố xác suất được kí hiệu là $\Phi(x)$,

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

Chú ý rằng $\Phi(-x) = 1 - \Phi(x), \forall x \in \mathbb{R}$.

Tính $\Phi(x)$ bằng máy tính Casio

1) CASIO FX570MS:

- Vào Mode tìm SD: Mode $\rightarrow$ Mode $\rightarrow$ 1 (SD);
- Shift $\rightarrow$ 3 (Distr) $\rightarrow$ 1;
- Nhập x .

2) CASIO FX570ES, FX570ES - PLUS, FX570VN - PLUS:

- Vào Mode tìm 1-Var: Mode $\rightarrow$ 3 (Stat) $\rightarrow$ 1 (1-Var) $\rightarrow$ AC;
- Shift $\rightarrow$ 1 (Stat) $\rightarrow$ 7 (Distr) $\rightarrow$ 1;
- Nhập x .

Tính hàm ngược $\Phi^{-1}(y)$ bằng máy tính CASIO FX570VN PLUS

- Mode $\rightarrow$ 3 (DIST) $\rightarrow$ 3;
- Nhập $y \rightarrow = \rightarrow \sigma = 1 = \rightarrow \mu = 0 =$.

Ví dụ 1.7. Tính $\Phi(1,96), \Phi(-1,65)$

Tính chất 1.6. Cho biến ngẫu nhiên $X \sim N(\mu, \sigma^2)$. Khi đó:

1) $E(X) = \mu, V(X) = \sigma^2$.

2) $Z = \frac{X - \mu}{\sigma} \sim N(0; 1)$.

3) Nếu $X_1, X_2, \dots, X_n$ là các biến ngẫu nhiên độc lập, cùng phân bố xác suất với $X \sim N(\mu; \sigma^2)$ thì:

$$T = X_1 + X_2 + \dots + X_n \sim N(n\mu; n\sigma^2)$$

và

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \sim N(\mu; \sigma^2/n)$$

(4) $P(X < a) = P(X \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right)$.

(5) Với $\alpha < \beta$ ta có:

$$P(\alpha < X < \beta) = P(\alpha \leq X \leq \beta) = \Phi\left(\frac{\beta - \mu}{\sigma}\right) - \Phi\left(\frac{\alpha - \mu}{\sigma}\right)$$

Ví dụ 1.8. Cho biến ngẫu nhiên liên tục $X \sim N(1; 4)$. Tính $P(X < 3, 5), P(X > 0), P(0, 5 < X \leq 2, 5)$.

Giải.

$$P(X < 3, 5) = \Phi\left(\frac{3,5 - 1}{2}\right) = \Phi(1, 25) = 0,8944;$$

$$P(X > 0) = 1 - P(X \leq 0) = 1 - \Phi(-0,5) = \Phi(0,5) = 0,6915;$$

$$P(0,5 < X \leq 2,5) = \Phi(0,75) - \Phi(-0,25) = 0,3721. \quad \blacksquare$$

Ví dụ 1.9. Giả sử số đo chiều dài của một sợi dây kim loại do một máy tự động cắt ra là một biến ngẫu nhiên chuẩn với $\mu = 10\text{mm}$, $\sigma^2 = 4\text{mm}^2$.

- Tính xác suất lấy ra được một sợi dây có chiều dài lớn hơn 13mm.
- Tìm tỉ lệ sợi dây do máy cắt ra có chiều dài từ 8,5mm đến 12,5mm.

Giải.

$$\text{a) } P(X > 13) = 1 - P(X \leq 13) = 1 - \Phi(1,5) = 0,067.$$

$$\text{b) } P(8,5 \leq X \leq 12,5) = \Phi(1,25) - \Phi(-0,75) = 0,668. \quad \blacksquare$$

Ví dụ 1.10. Đường kính của một trục trong ổ đĩa quang là một biến ngẫu nhiên chuẩn với đường kính trung bình là 0,2508inch và độ lệch chuẩn 0,0005inch. Thông số kỹ thuật ghi trên trục là $0,25 \pm 0,0015\text{inch}$. Tìm tỉ lệ trục có đường kính phù hợp với thông số kỹ thuật.

Giải. Gọi X là đường kính của trục ổ đĩa quang, ta có:

$$X \sim N(0,2508; 0,0005^2)$$

$$\begin{aligned} P(0,25 - 0,0015 \leq X \leq 0,25 + 0,0015) &= \Phi(1,4) - \Phi(-4,6) \\ &= 0,919 \end{aligned} \quad \blacksquare$$

Ví dụ 1.11. Tỉ lệ lợi nhuận X (%) của một dự án đầu tư được xem là một biến ngẫu nhiên có phân bố chuẩn. Theo đánh giá của ủy ban đầu tư thì với xác suất 0,1587 cho tỉ lệ lợi nhuận cao hơn 20% và với xác suất 0,0228 cho tỉ lệ lợi nhuận lớn hơn 25%. Tìm xác suất $P(X > 0)$.

Giải. Gọi X là lãi suất đầu tư vào 1 dự án trong 1 năm, khi đó $X \sim N(\mu; \sigma^2)$. Từ giả thiết ta có:

$$\begin{aligned} \begin{cases} P(X > 0,2) = 0,1587 \\ P(X > 0,25) = 0,0228 \end{cases} &\Leftrightarrow \begin{cases} 1 - \Phi\left(\frac{0,2 - \mu}{\sigma}\right) = 0,1587 \\ 1 - \Phi\left(\frac{0,25 - \mu}{\sigma}\right) = 0,0228 \end{cases} \\ &\Leftrightarrow \begin{cases} \Phi\left(\frac{0,2 - \mu}{\sigma}\right) = 0,8413 = \Phi(1) \\ \Phi\left(\frac{0,25 - \mu}{\sigma}\right) = 0,9772 = \Phi(2) \end{cases} \Leftrightarrow \begin{cases} \mu = 0,15 \\ \sigma = 0,05 \end{cases} \end{aligned}$$

Vì vậy:

$$P(X > 0) = 1 - \Phi\left(\frac{0 - 0,15}{0,05}\right) = \Phi(3) = 0,9987 \quad \blacksquare$$

Ví dụ 1.12. Chiều cao X (mét) của nam thanh niên trưởng thành ở quốc gia A tuân theo quy luật phân bố chuẩn $N(\mu; 0,1^2)$. Chọn ngẫu nhiên 100 nam thanh niên

của quốc gia A. Tính xác suất sai số tuyệt đối giữa chiều cao trung bình của 100 nam thanh niên được chọn với μ không vượt quá 0,03.

Giải.

Gọi X_k là chiều cao của nam thanh niên thứ k ($k = 1, 2, \dots, 100$). Khi đó:

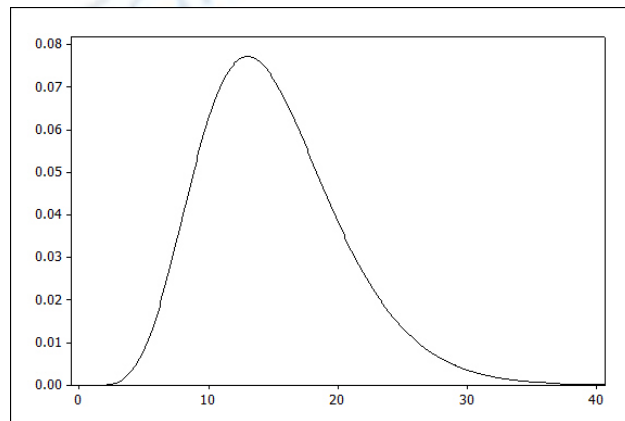
$$\bar{X} = \frac{X_1 + X_2 + \dots + X_{100}}{100}$$

là chiều cao trung bình của 100 nam thanh niên được chọn. Vì $\bar{X} \sim N(\mu; 0,01^2)$ nên ta có:

$$P(|\bar{X} - \mu| \leq 0,03) = 2\Phi(3) - 1 = 0,9974$$

Như vậy, khi chọn ngẫu nhiên 100 nam thanh niên thì hầu như chắc chắn rằng chiều cao trung bình của 100 nam thanh niên đó rơi vào đoạn $[\mu - 0,03; \mu + 0,03]$. ■

1.1.7. Phân bố khi bình phương



Hình 1.3: Phân bố khi bình phương ($n = 15$)

Định nghĩa 1.7. Biến ngẫu nhiên liên tục X có phân bố *khi bình phương* n bậc tự do nếu có hàm mật độ xác suất:

$$f_n(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{nếu } x > 0 \\ 0 & \text{nếu } x \leq 0 \end{cases}$$

Kí hiệu $X \sim \chi_n^2$.

Đồ thị hàm mật độ $f_n(x)$ của phân bố χ_n^2 có dạng như Hình 1.3.

1.1.8. Phân bố F

Định nghĩa 1.8. Cho W và Y là hai biến ngẫu nhiên độc lập có phân bố khi bình phương với bậc tự do lần lượt là m và n . Khi đó tỉ số:

$$F = \frac{W/m}{Y/n}$$

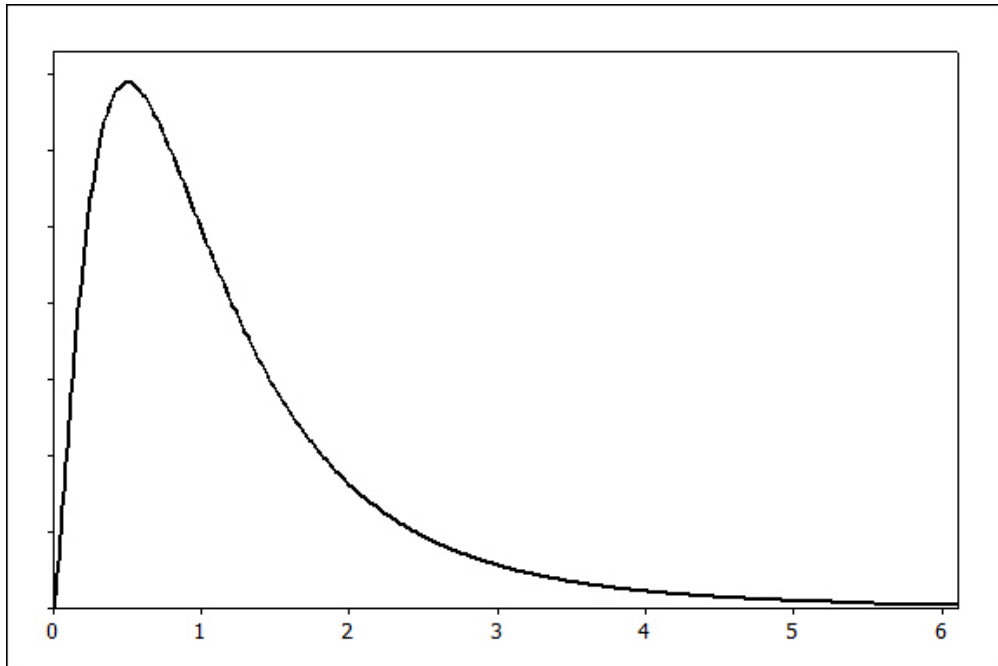
có hàm mật độ xác suất là:

$$f_{m,n}(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2}) \left(\frac{m}{n}\right)^{m/2} x^{(m/2)-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2}) \left[\frac{m}{n}x + 1\right]^{(m+n)/2}} & \text{nếu } x > 0 \\ 0 & \text{nếu } x \leq 0 \end{cases}$$

và được gọi là *phân bố F* với hai bậc tự do m và n .

Kí hiệu $F \sim F_{m,n}$.

Đồ thị hàm mật độ $f_{m,n}(x)$ của phân bố F có dạng như Hình 1.4.



Hình 1.4: Phân bố F ($m = 5, n = 10$)

1.2. Các định lí giới hạn

1.2.1. Luật số lớn

Định lý 1.2. (*Law of Large numbers*) Cho $\{X_n, n \geq 1\}$ là dãy biến ngẫu nhiên độc lập, cùng phân phối xác suất với kì vọng chung $= \mu$ và phương sai chung σ^2 . Khi đó, với mọi $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{k=1}^n X_k - \mu\right| \leq \varepsilon\right) = 1.$$

Nói cách khác, với n đủ lớn ta có

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \approx E(X).$$

1.2.2. Định lý giới hạn trung tâm

Định lý 1.3. (*Law of Large numbers*) Cho $\{X_n, n \geq 1\}$ là dãy biến ngẫu nhiên độc lập, cùng phân phối xác suất với kì vọng chung $= \mu$ và phương sai chung σ^2 . Khi đó

$$\lim_{n \rightarrow \infty} P\left(\frac{S - n\mu}{\sqrt{n}\sigma} < x\right) = \Phi(x) \quad x \in \mathbb{R},$$

trong đó $S = X_1 + X_2 + \dots + X_n$.

Nói cách khác, với n đủ lớn ta có

$$S = X_1 + X_2 + \dots + X_n \approx N(n\mu; n\sigma^2),$$

và

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \approx N(\mu; \sigma^2/n).$$

Ví dụ 1.13. Một thang máy vận chuyển hàng hóa lớn có thể vận chuyển mỗi lần tối đa 9800 kg. Giả sử mỗi lô hàng có 49 thùng hàng cần được vận chuyển. Kinh nghiệm cho thấy trọng lượng mỗi thùng hàng là biến ngẫu nhiên có kì vọng $\mu = 205$ kg và độ lệch chuẩn $\sigma = 15$ kg. Tính xác suất thang máy có thể vẫn chuyển được một lô hàng được chọn ngẫu nhiên gồm 49 thùng hàng như trên?

Giải. Đặt X_i = trọng lượng thùng hàng thứ i . Ta có $E(X_i) = \mu = 205$, $Var(X_i) = \sigma^2 = 15^2$ với $i = 1, 2, \dots, 49$. Vì vậy

$$P(S = X_1 + \dots + X_{49} < 9800) \approx \Phi\left(\frac{9800 - 49 \cdot 205}{\sqrt{49 \cdot 15}}\right) = \Phi(-2.33) = 0.0099.$$

■

Chương 2

THỐNG KÊ MÔ TẢ

2.1. Khái niệm mẫu và tổng thể

Giả sử ta cần nghiên cứu tính chất X nào đó của các phần tử trong tập hợp Ω mà Ω có số phần tử khá lớn ($|\Omega|$ có thể bằng vô cùng). Khi đó ta khó có thể nghiên cứu được tính chất X trên tất cả các phần tử. Phương pháp thống kê là chọn ngẫu nhiên một số lượng hữu hạn n phần tử để nghiên cứu. Trên cơ sở kết quả nghiên cứu của n phần tử này sẽ đưa ra kết luận cho toàn bộ tổng thể. Ta đưa ra các khái niệm sau:

- 1) *Tổng thể* là tập hợp tất cả các phần tử của Ω mà ta cần nghiên cứu tính chất X .
- 2) *Mẫu* là một tập con n phần tử của tổng thể được chọn ngẫu nhiên để nghiên cứu. n được gọi là *kích thước mẫu* (hoặc *cỡ mẫu*).
- 3) Nếu mỗi phần tử của tổng thể có tính chất X là một số thực thì với phương pháp chọn mẫu ngẫu nhiên, ta có X là biến ngẫu nhiên, tập các giá trị của X trong mẫu được gọi là *mẫu số liệu*.

Ví dụ 2.1. X là chiều cao của thanh niên Việt Nam 22 tuổi hiện nay. Khi đó:

- Tổng thể là tập hợp toàn bộ thanh niên Việt Nam 22 tuổi.
- Vì số lượng thanh niên 22 tuổi trên cả nước rất lớn nên ta không thể điều tra hết được mà chỉ chọn ra 1 tập hợp con để điều tra. Tập hợp con được chọn ra đó được gọi là một mẫu, số phần tử của mẫu là kích thước mẫu, tập tất cả các giá trị chiều cao của các cá thể trong mẫu là mẫu số liệu.

2.2. Các số đặc trưng của một mẫu số liệu

2.2.1. Trung bình mẫu, phương sai mẫu và độ lệch chuẩn mẫu

Cho $\{x_1, x_2, \dots, x_n\}$ là mẫu số liệu của biến ngẫu nhiên X .

1) **Trung bình mẫu**, kí hiệu là $\bar{x}$, được tính theo công thức:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

2) **Phương sai mẫu**, kí hiệu là s^2 , được tính theo công thức:

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

3) Độ lệch chuẩn mẫu.

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

Ví dụ 2.2. Giả sử ta có mẫu số liệu về chiều cao (mét) của 10 sinh viên một trường đại học như sau:

1,75 1,69 1,73 1,77 1,68
1,73 1,77 1,70 1,74 1,71

Tìm trung bình mẫu, phương sai mẫu và độ lệch chuẩn mẫu.

Chú ý 2.1.

1) Mẫu số liệu cho dạng bảng phân bố tần số rời rạc:

X	x_1	x_2	...	x_m
n_i	n_1	n_2	...	n_m

- Kích thước mẫu: $n = n_1 + n_2 + \dots + n_m$.

- Trung bình mẫu: $\bar{x} = \frac{1}{n} \sum_{i=1}^m n_i x_i$.

- Phương sai mẫu: $s^2 = \frac{1}{n-1} \left[\sum_{i=1}^m n_i x_i^2 - n\bar{x}^2 \right]$.

2) Mẫu số liệu cho dạng bảng phân bố tần số liên tục:

X	$a_0 - a_1$	$a_1 - a_2$	...	$a_{m-1} - a_m$
n_i	n_1	n_2	...	n_m

Trong đó $a_{k-1} - a_k = [a_{k-1}; a_k)$. Đặt $x_k = \frac{a_{k-1} + a_k}{2}$ ta được:

X	x_1	x_2	...	x_m
n_i	n_1	n_2	...	n_m

3) Tính $\bar{x}$ và s bằng máy tính CASIO FX570VN PLUS.

- Mode $\rightarrow 3 \rightarrow 1$;

- Bật/tắt tần số: Shift $\rightarrow$ SETUP $\rightarrow$ REPLAY $\rightarrow 4$ (Stat);

- Nhập số liệu, kết thúc nhập: bấm AC;

- Lấy $\bar{x}$: Shift $\rightarrow 1 \rightarrow 4 \rightarrow 2 \rightarrow =$;

- Lấy s : Shift $\rightarrow 1 \rightarrow 4 \rightarrow 4 \rightarrow =$.

Ví dụ 2.3. Doanh thu X (triệu đồng) trong 100 ngày được chọn ngẫu nhiên của 1 cửa hàng cho bởi bảng sau:

19,0 - 19,4	19,4 - 19,8	19,8 - 20,2	20,2 - 20,6	20,6 - 21,0
15	25	30	20	10

Tìm trung bình mẫu và độ lệch chuẩn mẫu.

Giải. Đưa về bảng tần số rời rạc:

X	19,2	19,6	20,0	20,4	20,8
n_i	15	25	30	20	10

Áp dụng Chú ý 1 ta tính được $\bar{x} = 19,94$ và $s \approx 0,48$. ■

2.2.2. Trung vị mẫu

Sắp xếp mẫu số liệu theo thứ tự tăng dần, giả sử $x_1 \leq x_2 \leq \dots \leq x_n$. *Trung vị mẫu*, kí hiệu $\tilde{x}$, xác định bởi:

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{nếu } n \text{ lẻ,} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{nếu } n \text{ chẵn} \end{cases}$$

2.2.3. Hệ số tương quan mẫu

Cho $\{(x_1, y_1); (x_2, y_2); \dots; (x_n, y_n)\}$ là mẫu hai chiều của vectơ ngẫu nhiên (X, Y) . *Hệ số tương quan mẫu* được xác định bởi:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

2.3. Biểu đồ

2.3.1. Biểu đồ phân bố tần số (Histogram)

Cho $(x_1, x_2, \dots, x_n)$ là mẫu số liệu của biến ngẫu nhiên X .

Trường hợp 1: X là biến ngẫu nhiên rời rạc.

Lập bảng phân bố tần số rời rạc của số liệu đã cho như sau:

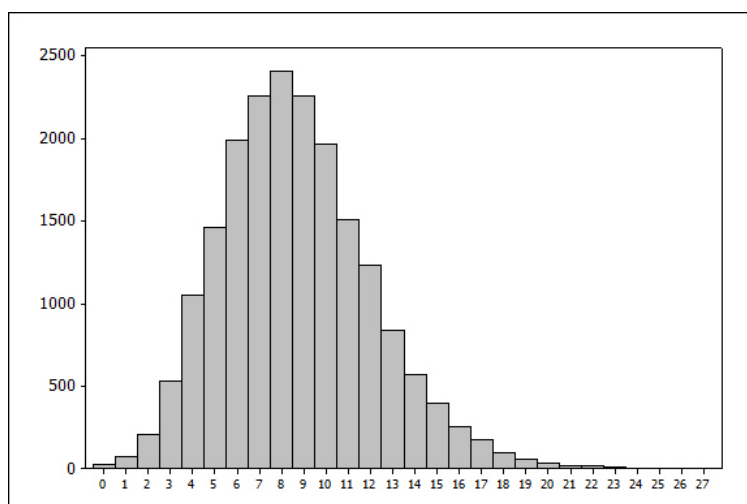
X	x_1^*	x_2^*	...	x_m^*
n_i	n_1	n_2	...	n_m

Sử dụng hệ trục tọa độ Descartes vuông góc để vẽ biểu đồ với trục hoành là các giá trị $x_1^*, x_2^*, \dots, x_m^*$, trục tung là tần số (hoặc tần số tương đối).

Ví dụ 2.4. Trong một cuộc thi game online có 27 màn được tổ chức có 19.383 game thủ tham gia. Kết quả cho bởi bảng sau:

Vượt qua màn	Số game thủ	Tần số tương đối	Vượt qua màn	Số game thủ	Tần số tương đối
0	20	0,0010	14	569	0,0294
1	72	0,0037	15	393	0,0203
2	209	0,0108	16	253	0,0131
3	527	0,0272	17	171	0,0088
4	1048	0,0541	18	97	0,0050
5	1457	0,0752	19	53	0,0027
6	1988	0,1026	20	31	0,0016
7	2256	0,1164	21	19	0,0010

8	2403	0,1240	22	13	0,0007
9	2256	0,1164	23	5	0,0003
10	1967	0,1015	24	1	0,0001
11	1509	0,0779	25	0	0,0000
12	1230	0,0635	26	1	0,0001
13	834	0,0430	27	1	0,0001



Hình 2.1: Biểu đồ tần số của cuộc thi game online

Trường hợp 2: X là biến ngẫu nhiên liên tục.

Lập bảng phân bố tần số liên tục:

X	$[a_0; a_1)$	$[a_1; a_2)$	...	$[a_{m-1}; a_m)$
n_i	n_1	n_2	...	n_m

Trong đó, số khoảng cần chia tốt nhất là từ 5 đến 20 khoảng, có thể chọn xấp xỉ bằng $\sqrt{n}$ (hoặc $1 + \log_2(n)$). Nếu ta chia dữ liệu thành m khoảng thì độ dài mỗi khoảng xấp xỉ $(\max\{x_k\} - \min\{x_k\})/m$.

Sử dụng hệ trục tọa độ Descartes vuông góc với trục hoành là các khoảng giá trị $[a_{k-1}; a_k)$, trục tung là tần số (hoặc tần số tương đối).

Ví dụ 2.5. Đo chiều dài (mm) của 50 con bọ cánh cứng ở một khu vực, nhà sinh học thu được kết quả sau:

4,1	4,35	4,45	4,5	4,55	4,55	4,6	4,65	4,7	4,75
4,2	4,35	4,45	4,5	4,55	4,6	4,6	4,65	4,7	4,75
4,3	4,35	4,45	4,5	4,55	4,6	4,6	4,65	4,7	4,75
4,3	4,4	4,45	4,5	4,55	4,6	4,65	4,65	4,725	4,78
4,35	4,4	4,45	4,525	4,55	4,6	4,65	4,7	4,75	4,95