BÀI TẬP: LÝ THUYẾT THỐNG KÊ

Lớp: 22KDL

Học kỳ II, năm học 2023 - 2024

Chương 1

Thống kê mô tả và chọn mẫu

Bài tập 1.1. Cho dãy số liệu sau:

121, 119, 117, 121, 120, 120, 118, 124, 123, 139, 120
115, 117, 121, 123, 120, 123 118, 117, 122, 122, 119

(a) Vẽ đồ thị thân và lá (stem & leaf) cho bộ dữ liệu này,

(b) Vẽ đồ thị histogram cho bộ dữ liệu này. Nhận xét về phân bố của dữ liệu (SV có thể dùng R). Bộ dữ liệu này có giá trị nào khác thường không? Giải thích.

Bài tập 1.2. Bảng sau mô tả điểm thi Math và Verbal của 14 sinh viên trong một bài thi SAT của họ

SV	Điểm Verbal	Điểm Math	SV	Điểm Verbal	Điểm Math
1	520	505	8	620	576
2	605	575	9	604	622
3	528	672	10	720	704
4	720	780	11	490	458
5	630	606	12	524	552
6	504	488	13	646	665
7	530	475	14	690	550

(a) Vẽ đồ thị stem & leaf 2 phía (side-by-side stem-and-leaf) của điểm thi Verbal và điểm Math.

(b) Vẽ đồ thị phân tán (scatter plot) biểu diễn điểm thi (Verbal, Math) trên cùng 1 đồ thị (có thể dùng R). Những sinh viên có điểm Verbal cao thì điểm Math có cao không?

Bài tập 1.3. Cho đồ thị stem & leaf:

9	3, 5, 8
8	6, 7, 8, 9, 9, 9
7	0, 1, 2, 2, 4, 5, 5, 6, 7, 8
6	0, 1, 2, 2, 3, 4, 5, 5
5	4, 6, 8

(a) Tính $\bar{x}, s^2, s$ và yếu vị (*mode*),

(b) Tính khoảng tứ phân phân vị $IQR = Q_3 - Q_1$ và vẽ đồ thị boxplot cho bộ dữ liệu trên,

(c) Bộ dữ liệu này có xấp xỉ phân phối chuẩn không?

(d) Trong bộ dữ liệu trên, tỷ lệ phần tử có giá trị nằm trong khoảng $\bar{x} - 2s$ đến $\bar{x} + 2s$ là bao nhiêu? So sánh kết quả với kết quả tìm được từ quy tắc thực nghiệm.

Bài tập 1.4. Cho bộ dữ liệu sau:

31, 38, 39, 39, 42, 42, 45, 47, 48, 48, 48, 52, 52, 53, 54, 55, 57, 59,
60, 61, 64, 64, 66, 66, 67, 68, 68, 69, 71, 71, 74, 75, 77, 79, 79, 79,

(a) Tính $\bar{x}, s^2, s$ và yếu vị (*mode*),

(b) Tính khoảng tứ phân phân vị $IQR = Q_3 - Q_1$ và vẽ đồ thị boxplot cho bộ dữ liệu trên,

(c) Có điểm outlier trong dữ liệu trên không?

(d) Vẽ đồ thị histogram cho bộ dữ liệu này.

Bài tập 1.5. Cho $x_1, x_2, \ldots, x_n$ là các giá trị bất kỳ và $\bar{x} = \sum_{i=1}^n x_i$. Chứng tỏ rằng:

$$\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Bài tập 1.6. Gọi $\bar{X}_n$ và S_n^2 lần lượt là trung bình và phương sai của mẫu $X_1, \ldots, X_n$. Giả sử quan trắc thêm X_{n+1}, gọi $\bar{X}_{n+1}$ và S_{n+1}^2 là trung bình và phương sai của mẫu $X_1, \ldots, X_{n+1}$. Chứng tỏ rằng:

(a)
$$\bar{X}_{n+1} = \frac{X_{n+1} + n\bar{X}}{n+1},$$

(b)
$$nS_{n+1}^2 = (n-1)S_n^2 + \left(\frac{n}{n+1}\right)\left(X_{n+1} - \bar{X}_n\right)^2$$
.

Chú ý: các bài tập sau đây sử dụng bảng tra phân phối Student và Chi bình phương để tính xác suất.

Bài tập 1.7. Giả sử $T \sim t(8)$, tính các xác suất sau:

(a) $\mathbb{P}(T \leq 2.896)$,

(b) $\mathbb{P}(T \leq -1.860)$,

(c) Tìm a sao cho $\mathbb{P}(-a < T < a) = 0.99$.

Bài tập 1.8. Giả sử $T \sim t(15)$, tính các xác suất sau:

(a) $\mathbb{P}(T \leq 1.341)$,

(b) $\mathbb{P}(T \leq -2.131)$,

(c) Tìm a sao cho $\mathbb{P}(-a < T < a) = 0.95$.

Bài tập 1.9. Xét X và Y lần lượt là các biến ngẫu nhiên độc lập có phân phối Chi bình phương với 14 và 5 bậc tự do. Tính:

(a) $\mathbb{P}(|X - Y| \leq 11.15)$,

(b) $\mathbb{P}(|X - Y| \geq 3.9)$.

Bài tập 1.10. Xét $X_1, X_2, \ldots, X_{10}$ là mẫu ngẫu nhiên chọn từ một phân phối chuẩn hóa. Tìm a và b sao cho:

$$\mathbb{P}\left(a \leq \sum_{i=1}^{10} X_i^2 \leq b\right) = 0.95$$

Bài tập 1.11. Xét $X_1, X_2, \ldots, X_{10}$ là mẫu ngẫu nhiên chọn từ một phân phối chuẩn có phương sai $\sigma^2 = 0.8$. Tìm a và b sao cho phương sai mẫu S^2 thỏa:

$$\mathbb{P}\left(a \leq S^2 \leq b\right) = 0.90$$

Bài tập 1.12. Xét $X_1, X_2, \ldots, X_{10}$ là mẫu ngẫu nhiên chọn từ một phân phối chuẩn có kỳ vọng bằng 8 và phương sai bằng 4. Tìm a sao cho:

$$\mathbb{P}\left(\sum_{i=1}^{10} \left(\frac{X_i - 8}{2}\right)^2 \leq a\right) = 0.95$$

Bài tập 1.13. Xét $X_1, X_2, \ldots, X_{16}$ là mẫu ngẫu nhiên chọn từ một phân phối chuẩn có kỳ vọng bằng 7 và phương sai bằng 4. Đặt $\bar{X} = \frac{1}{16}\sum_{i=1}^{16} X_i$. Tính các xác suất sau:

a) $\mathbb{P}\left(\sum_{i=1}^{16}(X_i - \bar{X})^2 > 100\right)$

b) $\mathbb{P}\left(\sum_{i=1}^{16}(X_i - 7)^2 > 100\right)$

b) $\mathbb{P}\left(\sum_{i=1}^{16} X_i > 100 \text{ và } \sum_{i=1}^{16}(X_i - \bar{X})^2 > 100\right)$

Chương 2

Ước lượng tham số thống kê

2.1 Bài tập về ước lượng điểm

Bài tập 2.1. Một mẫu ngẫu nhiên $(X_1, X_2, X_3, X_4, X_5)$ được chọn từ một tổng thể có phân phối chuẩn với kỳ vọng μ và phương sai σ^2. Xác định hàm mật độ xác suất đồng thời của mẫu này.

Bài tập 2.2. Biết tuổi thọ của các bóng bán dẫn (transistor) có phân phối mũ với tham số λ. Trong một nhà máy sản xuất linh kiện điện tử, chọn một mẫu ngẫu nhiên gồm n bóng bán dẫn. Xác định hàm mật độ xác suất đồng thời của mẫu này.

Bài tập 2.3. Một mẫu ngẫu nhiên (X_1, X_2, X_3, X_4) được chọn từ một tổng thể có phân phối đều trên đoạn $(5, 10)$. Xác định hàm mật độ xác suất đồng thời của mẫu này.

Bài tập 2.4. Xét $(X_1, \ldots, X_n)$ là một mẫu ngẫu nhiên của một biến ngẫu nhiên có phân phối Poisson với tham số λ.

(a) Tìm ước lượng hợp lý cực đại (MLE) cho λ.

(b) Chứng tỏ ước lượng MLE tìm được là ước lượng không chệch.

Bài tập 2.5. Xét $(X_1, \ldots, X_n)$ là một mẫu ngẫu nhiên của một biến ngẫu nhiên có phân phối nhị thức với tham số m và p, với p không biết.

(a) Tìm ước lượng hợp lý cực đại (MLE) cho p.

(b) Chứng tỏ ước lượng MLE tìm được là ước lượng không chệch.

Bài tập 2.6. Xét $(X_1, \ldots, X_n)$ là một mẫu ngẫu nhiên của một biến ngẫu nhiên có hàm mật độ xác suất

$$f(x) = \frac{1}{\theta^2} x e^{-x/\theta},\ 0 \leq x < \infty,\ 0 < \theta < \infty$$

(a) Tìm ước lượng hợp lý cực đại (MLE) cho θ.

(b) Chứng tỏ ước lượng MLE tìm được là ước lượng không chệch.

Bài tập 2.7. Xét $(X_1, \ldots, X_n)$ là một mẫu ngẫu nhiên của một biến ngẫu nhiên có hàm mật độ xác suất

$$f(x) = \frac{1}{\theta},\ 0 \leq x \leq \theta,\ \theta > 0$$

(a) Tìm ước lượng moment cho tham số θ.

(b) Tìm ước lượng hợp lý cực đại (MLE) cho θ.

(c) Kiểm tra tính không chệch của các ước lượng tìm được ở câu a) và b). Nên sử dụng ước lượng nào? Tại sao?

Bài tập 2.8. Xét $X_1, \ldots, X_n$, $n > 3$, là một mẫu ngẫu nhiên chọn từ tổng thể có phân phối chuẩn với kỳ vọng μ và phương sai σ^2. Xét 3 ước lượng sau của μ:

$$\hat{\mu}_1 = \frac{1}{3}(X_1 + X_2 + X_3)$$
$$\hat{\mu}_2 = \frac{1}{8}X_1 + \frac{3}{4(n-2)}(X_2 + \cdots + X_{n-1}) + \frac{1}{8}X_n$$
$$\hat{\mu}_3 = \bar{X}$$

(a) Chứng tỏ 3 ước lượng trên đều là ước lượng không chệch.

(b) Tính Eff $(\hat{\mu}_2, \hat{\mu}_1)$, Eff $(\hat{\mu}_3, \hat{\mu}_1)$ và Eff $(\hat{\mu}_3, \hat{\mu}_2)$. Ước lượng nào trong 3 ước lượng đó hiệu quả hơn?

Bài tập 2.9. Xét $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên độc lập sao cho X_j có hàm mật độ xác suất

$$f_j(x; \mu) = \frac{j}{\sqrt{\pi}} e^{-(jx-\mu)^2},\ x \in \mathbb{R},$$

với $j = 1, \ldots, n$. Tìm ước lượng hợp lý cực đại (MLE) cho μ.

Bài tập 2.10. Xét họ các hàm mật độ xác suất sau:

$$f(x; \theta) = \begin{cases} 3e^{-3(x-\theta)} & \text{khi } x > \theta \\ 0 & x \leq \theta \end{cases}.$$

Xét $X_1, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ tổng thể với phân phối $f(x; \theta)$.

a) Tìm ước lượng moment cho θ. Ta gọi ước lượng là này $\tilde{\theta}$. Tính độ chệch và phương sai của $\tilde{\theta}$.

b) Tìm ước lượng hợp lý cực đại $\hat{\theta}$ của θ.

c) Tìm hàm phân phối xác suất tích lũy của $\hat{\theta}$. Tính $\mathbb{E}(\hat{\theta})$ và $\mathbb{V}ar(\hat{\theta})$.

Bài tập 2.11. Xét $X_1, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ tổng thể có phân phối hình học với tham số p, $0 \leq p \leq 1$. Giả sử rằng phân phối tiên nghiệm của p là phân phối Beta với hai tham số $\alpha = 4$ và $\beta = 4$.

(a) Xác định phân phối hậu nghiệm của p.

(b) Tìm ước lượng Bayes cho p.

Bài tập 2.12. Xét $X_1, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ tổng thể có phân phối Poisson với tham số λ, $\lambda > 0$. Giả sử rằng phân phối tiên nghiệm của λ là phân phối Gamma(α, β).

(a) Xác định phân phối hậu nghiệm của λ và chứng tỏ rằng phân phối đó là Gamma$(\sum_{i=1}^n x_i + 1, n+1)$.

(b) Tìm ước lượng Bayes cho λ.

2.2 Bài tập về Khoảng tin cậy

Bài tập 2.13. Xét $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. Khoảng $(-\infty, U(X_1, \dots, X_n)]$ được gọi là khoảng tin cậy trên (upper - confidence bound) (hoặc khoảng tin cậy một phía bên trái) cho kỳ vọng μ với độ tin cậy $100(1-\alpha)\%$ nếu

$$\mathbb{P}\left(\mu \leq U(X_1, \dots, X_n)\right) = 1 - \alpha.$$

Giả sử σ^2 biết, chứng tỏ rằng KTC trên với độ tin cậy $100(1-\alpha)\%$ cho μ có dạng

$$\mu \leq \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Bài tập 2.14. Xét $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. Khoảng $[L(X_1, \dots, X_n), +\infty)$ được gọi là khoảng tin cậy dưới (lower - confidence bound) (hoặc khoảng tin cậy một phía bên phải) cho kỳ vọng μ với độ tin cậy $100(1-\alpha)\%$ nếu

$$\mathbb{P}\left(\mu \geq L(X_1, \dots, X_n)\right) = 1 - \alpha.$$

Giả sử σ^2 biết, chứng tỏ rằng KTC trên với độ tin cậy $100(1-\alpha)\%$ cho μ có dạng

$$\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu.$$

Bài tập 2.15. Thực hiện yêu cầu của bài tập 2.13 và 2.14 trong trường hợp σ^2 không biết.

Bài tập 2.16. Gọi p là tỷ lệ phần tử thỏa một tính chất $\mathcal{A}$ mà ta quan tâm trong một tổng thể. Khảo sát n phần tử thấy có X phần tử thỏa tính chất $\mathcal{A}$. Với độ tin cậy $100(1-\alpha)\%$, hãy xây dựng khoảng tin cậy trên và dưới (một phía bên trái và bên phải cho tỷ lệ p).

Bài tập 2.17. Xét $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$, σ^2 không biết. Gọi X_{n+1} là một quan trắc trong tương lai mà ta muốn dự đoán.

a) Hãy chỉ ra một ước lượng điểm cho X_{n+1}.

b) Gọi $U = X_{n+1} - \bar{X}$. Tín kỳ vọng và phương sai của U.

c) Từ các câu trên, hãy xây dựng khoảng tin cậy (đối xứng) $100(1-\alpha)\%$ cho giá trị dự báo X_{n+1}.

Bài tập 2.18. Xét $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ và $Y_1, \dots, Y_m \overset{i.i.d}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ là hai mẫu ngẫu nhiên độc lập. Hãy xây dựng khoảng tin cậy đối xứng với độ tin cậy $100(1-\alpha)\%$ cho $\mu_X - \mu_Y$, giả sử σ_X^2 và σ_Y^2 đã biết.

Bài tập 2.19. Xét $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ và $Y_1, \dots, Y_m \overset{i.i.d}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ là hai mẫu ngẫu nhiên độc lập. Giả sử σ_X^2 và σ_Y^2 không biết, và $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

a) Chứng tỏ rằng

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2},$$

là một ước lượng không chệch cho σ^2, trong đó S_X^2 và S_Y^2 lần lượt là các ước lượng cho σ_X^2 và σ_Y^2.

b) Xây dựng khoảng tin cậy đối xứng với độ tin cậy $100(1-\alpha)\%$ cho $\mu_X - \mu_Y$.

Bài tập 2.20. Khảo sát thời gian tự học hàng ngày (giờ) của sinh viên một trường đại học, kết quả cho bởi bảng sau

Thời gian	$1-3$	$3-5$	$5-7$	$7-9$	$9-11$
Số sinh viên	3	5	8	6	2

Biết rằng thời gian tự học hàng ngày của sinh viên tuân theo phân phối chuẩn.

(a) Lập khoảng tin cậy 99% cho thời gian tự học trung bình của sinh viên trường đại học này.

(b) Lập khoảng tin cậy 99% bên trái cho thời gian tự học trung bình của sinh viên trường đại học này.

(c) Với độ tin cậy 99%, nếu muốn sai số ước lượng ≤ 0.5 (giờ) thì cần khảo sát thêm bao nhiêu sinh viên?

(d) Lập khoảng tin cậy 90% cho tỷ lệ sinh viên có thời gian tự học trên 5 giờ mỗi ngày.

Bài tập 2.21. Trong một hồ nuôi một loại cá, người ta bắt lên 100 con cá, đánh dấu bằng sơn lên chúng và thả lại xuống hồ. Sau đó, người ta bắt lên 200 con cá thấy có 60 con cá có dấu sơn.

(a) Với độ tin cậy 95%, hãy ước lượng số lượng cá trong hồ.

(b) Lập khoảng tin cậy 95% bên phải cho tỷ lệ cá trong hồ.

(c) Cân trọng lượng 50 con cá bắt được, kết quả cho bởi bảng sau

Trọng lượng (kg)	$0,5$	$1,0$	$1,5$	$2,0$	$2,5$	$3,0$
Số cá	6	8	10	12	9	5

Lập khoảng tin cậy 99% cho trọng lượng trung bình của cá trong hồ. Nếu muốn sai số ước lượng bằng $0,1$ kg mà vẫn giữ nguyên cỡ mẫu khảo sát thì độ tin cậy sẽ bằng bao nhiêu?

Bài tập 2.22. Khảo sát về độ pH của nước ở các hồ bơi trong một thành phố, chọn ngẫu nhiên 20 hồ bơi để đo độ pH và được bảng số liệu sau:

Độ pH	$4,5-5,5$	$5,5-6,5$	$6,5-7,5$	$7,5-8,5$	$8,5-9,5$
Số hồ bơi	2	5	7	3	3

Biết rằng độ pH của nước hồ bơi có phân phối chuẩn với độ lệch tiêu chuẩn $\sigma = 1,30$.

(a) Lập khoảng tin cậy 98% cho độ pH trung bình của nước hồ bơi trong thành phố.

(b) Nếu muốn sai số ước lượng $\leq 0,05$ thì phải khảo sát ít nhất bao nhiêu hồ bơi?

(c) Độ pH của nước hồ bơi từ 6 đến 8 thì đạt tiêu chuẩn, lập khoảng tin cậy 95% cho tỷ lệ hồ bơi đạt tiêu chuẩn trong thành phố.

(d) Lập khoảng tin cậy 95% bên trái cho độ pH trung bình của nước hồ bơi trong thành phố.