Xử lý số liệu thống kê

Assignment 1

Nhóm E:

Trần Tiến Đạt
Nguyễn Thị Ngọc Anh
Nguyễn Thái Hưng Thịnh

Ngày nộp: January 20, 2026

Contents

«««< HEAD =======

Danh sách và đóng góp của thành viên nhóm

Nhóm E bao gồm các thành viên sau đây với các đóng góp cụ thể:

- **Trần Tiến Đạt**: Bài tập 4, Bài tập 11.
- **Nguyễn THị Ngọc Anh**: Bài tập 5, Bài tập 14
- **Nguyễn Thái Hưng Thịnh**: Bài tập 1, Bài tập 3, Bài tập 12

Bài 1

Tính trung bình mẫu $\bar{y}$ của dữ liệu sau: 3, 5, 8, 15, 20, 21, 24. Áp dụng biến đổi logarithm cho dữ liệu này, sau đó tính trung bình mẫu $\bar{y'}$ và trung vị m' của dữ liệu đã biến đổi. Có phải $\log(\bar{y}) = \bar{y'}$ và $\log(m) = m'$ hay không?

- Dữ liệu gốc $Y = \{3, 5, 8, 15, 20, 21, 24\}$.
- Kích thước mẫu $n = 7$.
- Dữ liệu đã được sắp xếp: $y_1 = 3, y_2 = 5, \ldots, y_7 = 24$.

(a) Tính trung bình mẫu $\bar{y}$ của dữ liệu sau: 3, 5, 8, 15, 20, 21, 24.
Lời giải :
Ta có:

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \frac{1}{7}\sum_{i=1}^{7} y_i \\
&= \frac{1}{7}(3 + 5 + 8 + 15 + 20 + 21 + 24) \\
&= \frac{1}{7} \cdot 96 \\
&= \frac{96}{7} \approx 13.7143
\end{aligned}
$$

(b) Tính trung bình mẫu $\bar{y'}$ và trung vị m' của dữ liệu sau khi biến đổi logarithm
Lời giải :
Chúng ta áp dụng phép biến đổi $f(x) = \log(x)$. Tập dữ liệu mới là $Y' = \{\log(y_i)\}_{i=1}^{n} = \{\log(3), \log(5), \log(8), \log(15), \log(20), \log(21), \log(24)\}$

Các giá trị xấp xỉ (làm tròn đến 4 chữ số thập phân):

- $\log(3) \approx 0.4771$
- $\log(5) \approx 0.699$
- $\log(8) \approx 0.9031$
- $\log(15) \approx 1.1761$
- $\log(20) \approx 1.301$
- $\log(21) \approx 1.3222$
- $\log(24) \approx 1.3802$

Tính trung bình mẫu $\bar{y}'$:

$$\begin{aligned}
\bar{y}' &= \frac{1}{n}\sum_{i=1}^{n} y_i' \\
&= \frac{1}{7}\sum_{i=1}^{7} y_i' \\
&= \frac{1}{7}(log(3) + \log(5) + \log(8) + \log(15) + \log(20) + \log(21) + \log(24)) \\
&= \frac{1}{7} \cdot 7.2587 \\
&= \frac{7.2587}{7} \approx 1.037
\end{aligned}$$

Tính trung vị m': Vì $n = 7$ (lẻ), trung vị m' của Y' là giá trị ở vị trí thứ $(7+1)/2 = 4$.

$$m' = \log(y_4) = \log(15)$$

Giá trị xấp xỉ là: $m' \approx 1.176$

(c) Có phải $\log(\bar{y}) = \bar{y'}$ và $\log(m) = m'$ hay không?
Lời giải :
Tính trung vị m: Vì $n = 7$ (lẻ), trung vị m của Y là giá trị ở vị trí thứ $(7+1)/2 = 4$.

$$m = y_4 = 15$$

Xét: $\log(\bar{y}) = \bar{y'} \Rightarrow \log(13.7143) = 1.037 \Rightarrow 1.1371 = 1.037$ (Sai) Vậy $\log(\bar{y}) \neq \bar{y'}$
Xét: $\log(m) = m' \Rightarrow \log(15) = 1.176 \Rightarrow 1.176 = 1.176$ (Đúng) Vậy $\log(m) = m'$

Bài 3

Đặt $\bar{y}$ và m lần lượt là trung bình và trung vị của mẫu $y_1 < y_2 < \cdots < y_n$. Xét f là một hàm số thực.

1. a. Có phải $f(\bar{y})$ là trung bình mẫu của dữ liệu $f(y_1), f(y_2), \ldots, f(y_n)$?
2. b. Có phải $f(m)$ là trung vị của dữ liệu $f(y_1), f(y_2), \ldots, f(y_n)$?
3. c. Có hay không bất kỳ điều kiện gì để chắc chắn rằng $f(\bar{y})$ là trung bình mẫu của dữ liệu đã biến đổi?
4. d. Có hay không bất kỳ điều kiện gì để chắc chắn rằng $f(m)$ là trung vị của dữ liệu đã biến đổi?

Gọi $Y = \{y_1, y_2, \ldots, y_n\}$ là tập hợp dữ liệu gốc. Gọi $Z = \{f(y_1), f(y_2), \ldots, f(y_n)\}$ là tập hợp dữ liệu đã qua phép biến đổi f.

(a) a.Có phải $f(\bar{y})$ là trung bình mẫu của dữ liệu $f(y_1), f(y_2), \ldots, f(y_n)$?
Lời giải :

Ta có: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Trung bình mẫu của dữ liệu đã biến đổi Z là $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} f(y_i)$. Câu hỏi yêu cầu chúng ta kiểm tra liệu: $f(\bar{y}) = \bar{z}$? Nói cách khác: $f\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} f(y_i)$

Chứng minh (Phản ví dụ): Chọn $f(x) = x^2$ Giả sử: Mẫu $Y = \{1, 3\}$

- $\bar{y} = \frac{1+3}{2} = 2$
- $f(\bar{y}) = f(2) = 2^2 = 4$
- Xét $Z = \{f(1), f(3)\} = \{1^2, 3^2\} = \{1, 9\}$.
- Trung bình mẫu của Z: $\bar{z} = \frac{1+9}{2} = 5$

Vì $f(\bar{y}) = 4 \neq 5 = \bar{z}$. Nên ta có kết luận: (a): $f(\bar{y})$ không phải là trung bình mẫu của dữ liệu $f(y_1), f(y_2), \ldots, f(y_n)$.

(b) b. Có phải $f(m)$ là trung vị của dữ liệu $f(y_1), f(y_2), \ldots, f(y_n)$? hay $f(m) = m_Z$ luôn đúng với mọi hàm f hay không
Lời giải :

Chứng minh (Phản ví dụ):

TH1: Hàm không đơn điệu, n lẻ (n là kích cỡ mẫu Y và Z) Chọn $f(x) = (x-5)^2$ Giả sử: mẫu $Y = \{1, 2, 10\}$ và $f(m) = m_Z$ là đúng Khi đó: Mẫu đã được sắp xếp $y_1 = 1, y_2 = 2, y_3 = 10$.

- $m = y_2 = 2$
- $f(m) = f(2) = (2-5)^2 = 9$
- Xét $Z = \{f(1), f(2), f(10)\}$:
  - $f(1) = (1-5)^2 = 16$
  - $f(2) = (2-5)^2 = 9$
  - $f(10) = (10-5)^2 = 25$
- Mẫu $Z = \{16, 9, 25\}$. Chúng ta phải sắp xếp lại mẫu này: $Z_{(ordered)} = \{9, 16, 25\}$.
- Trung vị của Z: $m_Z = 16$

Vì $f(m) = 9 \neq 16 = m_Z$ Nên $f(m) \neq m_Z$ trong hàm không đơn điệu

TH2: Hàm đơn điệu, n chẵn (n là kích cỡ mẫu Y và Z) Chọn $f(x) = x^3$ và $f(m) = m_Z$ là đúng Giả sử: $Y = \{1, 3\}$

- $m = \frac{1+3}{2} = 2$
- $f(m) = f(2) = 2^3 = 8$
- Xét $Z = \{f(1), f(3)\} = \{1^3, 3^3\} = \{1, 27\}$. Vì f đơn điệu tăng, Z đã được sắp xếp.

- Trung vị của Z: $m_Z = \frac{1+27}{2} = 14$

Vì $f(m) = 8 \neq 14 = m_Z$ Nên $f(m) \neq m_Z$ trong hàm đơn điệu

Từ TH1 và TH2, ta có kết luận (b): $f(m)$ không phải là trung vị của dữ liệu $f(y_1), f(y_2), \ldots, f(y_n)$.

(c) c. Có hay không bất kỳ điều kiện gì để chắc chắn rằng $f(\bar{y})$ là trung bình mẫu ủa dữ liệu đã biến đổi?

Lời giải :

Có. Chúng ta cần tìm điều kiện để $f(\bar{y}) = \bar{z}$, tức là:

$$f\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} f(y_i)$$

Theo Bất đẳng thức Jensen, đẳng thức xảy ra khi và chỉ khi f là một hàm affine. Hay $f(x) = a \cdot x + b$ với $a, b \in \mathbb{R}$ là các hằng số.

Ta sẽ chứng minh điều kiện này:

- Vế trái (LHS):

$$\begin{aligned} f(\bar{y}) &= f\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\ &= a\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + b \end{aligned}$$

- Vế phải (RHS):

$$\begin{aligned} \bar{z} &= \frac{1}{n}\sum_{i=1}^{n} f(y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}(ay_i + b) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}(ay_i) + \sum_{i=1}^{n} b\right) \\ &= \frac{1}{n}\left(a\sum_{i=1}^{n} y_i + nb\right) \\ &= a\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + b \end{aligned}$$

Vì LHS = RHS, nên điều kiện được thỏa mãn.

Kết luận (c): Có, điều kiện là f phải là một hàm affine.

(d) d. Có hay không bất kỳ điều kiện gì để chắc chắn rằng $f(m)$ là trung vị của dữ liệu đã biến đổi?

Lời giải :

Chứng minh:

TH1:Trường hợp tổng quát, đúng cho mọi n $(nlkchthcmu)$ Điều kiện để $f(m) = m_Z$ đúng cho mọi n (cả chẵn và lẻ) là f phải là một hàm affine ($f(x) = ax + b$).

Khi đó, ta có: $f(m) = f\left(\frac{y_k+y_{k+1}}{2}\right) = a\left(\frac{y_k+y_{k+1}}{2}\right) + b$.

- Nếu $a \geq 0$, f không giảm, mẫu Z được sắp xếp là $f(y_1), \ldots, f(y_n)$. $m_Z = \frac{f(y_k)+f(y_{k+1})}{2} = \frac{(ay_k+b)+(ay_{k+1}+b)}{2} = a\left(\frac{y_k+y_{k+1}}{2}\right) + b$.

- Nếu $a < 0$, f không tăng, mẫu Z được sắp xếp ngược lại: $f(y_n) \leq \cdots \leq f(y_1)$. Các phần tử giữa là $f(y_{k+1})$ và $f(y_k)$. $m_Z = \frac{f(y_{k+1})+f(y_k)}{2} = \frac{(ay_{k+1}+b)+(ay_k+b)}{2} = a\left(\frac{y_k+y_{k+1}}{2}\right) + b$.

Trong mọi trường hợp, $f(m) = m_Z$ khi f là affine.

TH2: Trường hợp đặc biệt, n lẻ $(nlkchthcmu)$ Nếu chúng ta được đảm bảo rằng kích thước mẫu n là lẻ ($n = 2k+1$), thì một điều kiện yếu hơn (ít nghiêm ngặt hơn) là đủ: f chỉ cần là một hàm đơn điệu (monotonic).

Nếu $n = 2k + 1$, thì $m = y_{k+1}$. Do đó $f(m) = f(y_{k+1})$.

- Trường hợp 1: f đơn điệu không giảm. Vì $y_1 \leq \cdots \leq y_n$, chúng ta có $f(y_1) \leq \cdots \leq f(y_n)$. Mẫu Z đã được sắp xếp. Trung vị m_Z là phần tử thứ $(k+1)$ của Z, tức là $m_Z = f(y_{k+1})$. Vậy $f(m) = m_Z$.

- Trường hợp 2: f đơn điệu không tăng. Vì $y_1 \leq \cdots \leq y_n$, chúng ta có $f(y_1) \geq \cdots \geq f(y_n)$. Mẫu Z được sắp xếp theo thứ tự ngược lại: $f(y_n) \leq \cdots \leq f(y_{k+1}) \leq \cdots \leq f(y_1)$. Trung vị m_Z vẫn là phần tử chính giữa (thứ $k+1$ trong dãy đã sắp xếp), tức là $m_Z = f(y_{k+1})$. Vậy $f(m) = m_Z$.

Từ TH1 và TH2: Kết luận (d): Có điều kiện để chắc chắn rằng $f(m)$ là trung vị của dữ liệu đã biến đổi.

»»»> 96cff8b9f888f6e3eaf38b17df1e9842471c0ff6

Bài 4

Xét hai bộ dữ liệu $x_1 < x_2 < x_3 < ... < x_n$ và $y_1 < y_2 < y_3 < ... < y_n$, có trung bình mẫu tương ứng là $\bar{x}$, $\bar{y}$ và trung vị lần lượt là m_x và m_y. Đặt $w_i = x_i + y_i$.

(a) Chứng minh hoặc đưa ra phản chứng rằng: $\bar{x} + \bar{y}$ là trung bình mẫu của $w_1, w_2, ..., w_n$

Lời giải :

Trước hết, ta xét trung bình mẫu của $w_1, w_2, ..., w_n$:

$$
\begin{aligned}
\bar{w} &= \frac{1}{n}\sum_{i=1}^{n} w_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \bar{x} + \bar{y}
\end{aligned}
$$

Vậy suy ra $\bar{w} = \bar{x} + \bar{y}$ là trung bình mẫu của $w_1, w_2, ..., w_n$.

(b) Chứng minh hoặc đưa ra phản chứng rằng: $m_x + m_y$ là trung vị của $w_1, w_2, \ldots, w_n$

Lời giải :

Ta sẽ chứng minh mệnh đề trên là đúng, trước hết ta sẽ xét với trường hợp n là số lẻ, tức rằng ta có trung vị của $\{x\}_n$ và $\{y\}_n$ lần lượt là:

$$
\begin{aligned}
m_x &= x_{\frac{n+1}{2}} \\
m_y &= y_{\frac{n+1}{2}}
\end{aligned}
$$

Khi đó với dãy w ta có:

$$
\begin{aligned}
m_w &= w_{\frac{n+1}{2}} \\
&= x_{\frac{n+1}{2}} + y_{\frac{n+1}{2}} \\
&= m_x + m_y
\end{aligned} \tag{1}
$$

Với trường hợp n là số chẵn, ta có trung vị của $\{x\}_n$ và $\{y\}_n$ lần lượt là:

$$
\begin{aligned}
m_x &= \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) \\
m_y &= \frac{1}{2}(y_{\frac{n}{2}} + y_{\frac{n}{2}+1})
\end{aligned}
$$

Khi đó với dãy w ta có:

$$
\begin{aligned}
m_w &= \frac{1}{2}(w_{\frac{n}{2}} + w_{\frac{n}{2}+1}) \\
&= \frac{1}{2}((x_{\frac{n}{2}} + y_{\frac{n}{2}}) + (x_{\frac{n}{2}+1} + y_{\frac{n}{2}+1})) \\
&= \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) + \frac{1}{2}(y_{\frac{n}{2}} + y_{\frac{n}{2}+1}) \\
&= m_x + m_y
\end{aligned} \tag{2}
$$

Từ (2) và (1) ta suy ra $m_x + m_y$ là trung vị của $w_1, w_2, \ldots, w_n$ với mọi n.

Bài 5

Giả sử rằng ta có dữ liệu ngẫu nhiên $X_1, X_2, \dots, X_n$ là độc lập cùng phân phối Poisson $P(\lambda)$. Ta biết rằng $\bar{X}$ là ước lượng không chệch của λ.
(a) Sử dụng định lý giới hạn trung tâm, hãy xác định công thức cho khoảng tin cậy 95% cho λ.
Lời giải :

Với $X_i \sim \text{Poisson}(\lambda)$, ta có

$$\mathbb{E}[X_i] = \lambda, \qquad \text{Var}(X_i) = \lambda.$$

Khi đó, trung bình mẫu

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

là ước lượng không chệch của λ.

Từ định lý giới hạn trung tâm, ta có:

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \xrightarrow{d} \mathcal{N}(0,1),$$

tức là khi n đủ lớn, phân phối của $\bar{X}$ xấp xỉ chuẩn với trung bình λ và phương sai $\dfrac{\lambda}{n}$:

$$\bar{X} \approx \mathcal{N}\left(\lambda,\ \frac{\lambda}{n}\right).$$

Do đó,

$$P\left(-1.96 \le \frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \le 1.96\right) \approx 0.95,$$

hay tương đương,

$$P\left(\bar{X} - 1.96\sqrt{\frac{\lambda}{n}} \le \lambda \le \bar{X} + 1.96\sqrt{\frac{\lambda}{n}}\right) \approx 0.95.$$

Do tham số λ chưa biết nên trong thực tế ta không thể tính chính xác sai số chuẩn $\sqrt{\lambda/n}$. Vì $\bar{X}$ là ước lượng hợp lý và không chệch của λ, ta có thể thay λ bằng $\bar{X}$ trong biểu thức này để thu được ước lượng xấp xỉ của sai số chuẩn. Khi đó, khoảng tin cậy 95% (theo xấp xỉ Wald) cho λ được viết là:

$$\boxed{\lambda \in \left[\bar{X} - 1.96\sqrt{\frac{\bar{X}}{n}},\ \bar{X} + 1.96\sqrt{\frac{\bar{X}}{n}}\right]}.$$

(b) Áp dụng công thức trong ý (a), hãy xác định khoảng tin cậy cho λ theo dữ liệu $4, 6, 7, 9, 10, 13$.
Lời giải :

Với dữ liệu đã cho, ta có $n = 6$ và

$$\bar{x} = \frac{4+6+7+9+10+13}{6} = \frac{49}{6} \approx 8.167.$$

Độ tin cậy:

$$1-\alpha = 0.95 \;\Rightarrow\; \alpha = 0.05 \;\Rightarrow\; 1-\frac{\alpha}{2} = 0.975 \;\Rightarrow\; z_{1-\alpha/2} = z_{0.975} = 1.96.$$

Biên độ tin cậy:

$$\Delta \;=\; z_{0.975}\sqrt{\frac{\bar{x}}{n}} = 1.96\sqrt{\frac{49/6}{6}} = 1.96\sqrt{\frac{49}{36}} = 1.96\cdot\frac{7}{6} = \frac{343}{150} \approx 2.287.$$

Khi đó, khoảng tin cậy 95% cho λ là

$$\lambda \in [\,8.167 - 2.287,\; 8.167 + 2.287\,] = [\,5.880,\; 10.454\,].$$

Kết luận: Khoảng tin cậy 95% cho tham số λ của phân phối Poisson dựa trên mẫu đã cho là

$$\boxed{\lambda \in (5.880,\; 10.454)}.$$

haha