LỜI GIẢI ĐỀ THI CUỐI HỌC KÌ I LÝ THUYẾT THỐNG KÊ (2023 - 2024)

Bài 1

Đo chỉ số chất béo X (đơn vị: %) trong sữa bò của 125 con bò thuộc một giống bò sữa lai mới của Hà Lan, ta được bảng số liệu sau:

X	3,5	3,8	4,5	5,2	5,6	6,4	6,8
n_i	2	8	35	40	20	15	5

Giả thiết rằng X có phân phối chuẩn.

a) Tìm khoảng tin cậy 99% cho trung bình chỉ số chất béo trong sữa giống bò lai trên.

Lời giải:

Đối tượng khảo sát có cỡ mẫu lớn ($125 > 30$) và chưa biết phương sai tổng thể.

Ta có trung bình mẫu:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{7} n_i x_i = \frac{2.3,5 + 8.3,8 + 35.4,5 + 40.5,2 + 20.5,6 + 15.6,4 + 5.6,8}{125} = 5,1592$$

và phương sai mẫu:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{7} n_i (x_i - \overline{x})^2 \approx 0,6129$$

Độ tin cậy $1 - \alpha = 99\% \Rightarrow \alpha = 0,01$ nên ta có phân vị $z_{1-\frac{\alpha}{2}} = z_{0,995} = 2,58$.

Dung sai:

$$\varepsilon = z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 2,58\sqrt{\frac{0,6129}{125}} \approx 0,1806$$

Vậy khoảng tin cậy 99% cho trung bình chỉ số chất béo trong sữa giống bò lai trên là:

$$\mu \in [\overline{x} - \varepsilon, \overline{x} + \varepsilon] = [5,1592 - 0,1806; 5,1592 + 0,1806] = [4,9786; 5,3398]$$

b) Biết trung bình chỉ số chất béo trong sữa giống bò thuần chủng (giống bò cũ) là 4,65. Việc lai tạo có cho trung bình chỉ số chất béo của sữa bò **tăng lên** hay không, với $\alpha = 1\%$?

Lời giải:

Gọi μ là trung bình chỉ số chất béo trong sữa của giống bò mới.

"Trung bình chỉ số chất béo trong sữa giống bò thuần chủng (giống bò cũ) là 4,65" $\Rightarrow \mu_0 = 4,65$.

"Việc lai tạo có cho trung bình chỉ số chất béo của sữa bò **tăng lên** hay không" $\Rightarrow H_1 : \mu > \mu_0 = 4,65$.

Ta tiến hành kiểm định giả thuyết $\begin{cases} H_0 : \mu = 4,65 \\ H_1 : \mu > 4,65 \end{cases}$

Thống kê kiểm định:

$$z_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{5,1592 - 4,65}{\sqrt{\frac{0,6129}{125}}} = 7,272$$

Ta bác bỏ H_0 nếu $z_0 > z_{1-\alpha} = z_{0,99} = 2,33$. Ta có $z_0 = 7,272 > 2,33 \Rightarrow$ Bác bỏ H_0.

Vậy với mức ý nghĩa 1%, việc lai tạo có cho trung bình chỉ số chất béo của sữa bò tăng lên.

c) Sữa bò được đánh giá là loại 1 nếu chỉ số chất béo nằm trong khoảng từ 4,0 đến 6,0. Có ý kiến cho rằng **ít nhất** 70% lượng sữa bò của giống bò lai mới này thuộc loại 1. Hãy kiểm định ý kiến trên với mức ý nghĩa 5%. Tính p-giá trị.

Lời giải:

Đặt Y là số con bò mà sữa của chúng thuộc loại 1. Ta tính:

$$\widehat{P} = \frac{Y}{n} = \frac{35+40+20}{125} = 0,76$$

Đặt p là tỷ lệ lượng sữa của giống bò mới mà thuộc loại 1.

"Có ý kiến cho rằng **ít nhất** 70% lượng sữa bò của giống bò lai mới này thuộc loại 1" $\Rightarrow \begin{cases} p_0 = 0,7 \\ H_0 : p \geq p_0 = 0,7 \end{cases}$

Ta kiểm tra: $n.p_0 = 125.0,7 = 87,5 \geq 5$ và $n(1-p_0) = 37,5 \geq 5$.

Ta tiến hành kiểm định giả thuyết $\begin{cases} H_0 : p \geq 0,7 \\ H_1 : p < 0,7 \end{cases}$

Thống kê kiểm định:

$$z_0 = \frac{\widehat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0,76-0,7}{\sqrt{\frac{0,7.0,3}{125}}} \approx 1,464$$

Ta bác bỏ H_0 nếu $z_0 < -z_{1-\alpha} = -z_{0,95} = -1,64$. Ta có $z_0 = 1,464 > -1,64 \Rightarrow$ Không bác bỏ H_0.

Vậy với mức ý nghĩa 5%, ý kiến trên là đúng.

Ta tính p-giá trị: $p = \Phi(z_0) = \Phi(1,464) = 0,9284$.

Bài 2

Khảo sát đường kính những thanh thép được sản xuất bởi hai máy dập tự động. Hai mẫu ngẫu nhiên $n = 14$ và $m = 16$ được chọn, tính được trung bình mẫu và phương sai mẫu lần lượt là

$$\overline{x} = 8,72; \ s_1^2 = 0,35; \ \overline{y} = 8,68; \ s_2^2 = 0,4.$$

Giả sử dữ liệu chọn từ các tổng thể có phân phối chuẩn và $\sigma_1^2 = \sigma_2^2$. Với mức ý nghĩa 5%, có đủ bằng chứng để khẳng định rằng đường kính trung bình của các thanh thép do hai máy dập tự động này sản xuất là khác nhau hay không?

Lời giải:

Gọi μ_1 và μ_2 lần lượt là đường kính trung bình của các thanh thép do máy dập tự động I và II sản xuất.

"Đường kính trung bình của các thanh thép do hai máy dập tự động này sản xuất là khác nhau"
$\Rightarrow H_1 : \mu_1 - \mu_2 \neq D_0 = 0$.

Ta tiến hành kiểm định giả thuyết $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$ với giả sử rằng $\sigma_1^2 = \sigma_2^2$.

Ta tính phương sai mẫu chung:

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} = \frac{(14-1).0,35 + (16-1).0,4}{14+16-2} \approx 0,377$$

Thống kê kiểm định:

$$t_0 = \frac{\overline{x} - \overline{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}} = \frac{8,72 - 8,68 - 0}{\sqrt{\frac{0,377}{14} + \frac{0,377}{16}}} \approx 0,178$$

Ta bác bỏ H_0 nếu $|t_0| > t^{n+m-2}_{1-\frac{\alpha}{2}} = t^{28}_{0,975} = 2,0484$. Ta có $|t_0| = 0,178 < 2,0484 \Rightarrow$ Không bác bỏ H_0.

Vậy với mức ý nghĩa 5%, ta không có đủ bằng chứng để khẳng định rằng đường kính trung bình của các thanh thép do hai máy dập tự động này sản xuất là khác nhau.

Bài 3

Một nghiên cứu về mối liên hệ giữa tỷ lệ cây xanh X (đơn vị: m^2/người) và nhiệt độ trung bình trong mùa hè Y (đơn vị: ^{o}C) được khảo sát tại 12 thành phố (các thành phố này có cùng kiểu khí hậu địa lý), thu được dữ liệu sau:

$$\sum_{j=1}^{12} x_j = 241; \quad \sum_{j=1}^{12} x_j^2 = 7281; \quad \sum_{j=1}^{12} x_j y_j = 6404; \quad \sum_{j=1}^{12} y_j = 341,5; \quad \sum_{j=1}^{12} y_j^2 = 9813,25$$

a) Tìm phương trình đường thẳng hồi quy tuyến tính bằng phương pháp bình phương nhỏ nhất.

Lời giải:

Bằng phương pháp bình phương nhỏ nhất, ta tính hệ số góc:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{j=1}^{12} x_j y_j - \frac{1}{n}\sum_{j=1}^{12} x_j \sum_{j=1}^{12} y_j}{\sum_{j=1}^{12} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{12} x_j\right)^2} = \frac{6404 - \frac{1}{12}\cdot 241 \cdot 341,5}{7281 - \frac{1}{12}\cdot 241^2} \approx -0,1862$$

và hệ số chặn:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{1}{n}\sum_{j=1}^{12} y_j + 0,186 \cdot \frac{1}{n}\sum_{j=1}^{12} x_j = \frac{1}{12}\cdot 341,5 + \frac{0,1862}{12}\cdot 241 \approx 32,1979$$

Vậy phương trình đường thẳng hồi quy tuyến tính là:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 32,1979 - 0,1862x$$

b) Nếu một thành phố (cùng kiểu khí hậu này) có tỷ lệ cây xanh là $x_0 = 24$ m^2/người, dựa vào đường thẳng hồi quy tìm được ở câu a, hãy dự đoán nhiệt độ trung bình trong mùa hè của thành phố đó.

Lời giải:

Với tỷ lệ cây xanh là $x_0 = 24$ m^2/người thì nhiệt độ trung bình là:

$$\hat{y} = 32,1938 - 0,1862x_0 = 32,1979 - 0,1862.24 = 27,7291 \ (^o\text{C})$$

c) Tính hệ số xác định R^2 và nhận xét về mối liên hệ tuyến tính giữa tỷ lệ cây xanh và nhiệt độ trung bình trong mùa hè.

Lời giải:

Ta có:

$$SSR = \hat{\beta}_1 S_{xy} = \hat{\beta}_1 \left(\sum_{j=1}^{12} x_j y_j - \frac{1}{n} \sum_{j=1}^{12} x_j \sum_{j=1}^{12} y_j \right) = -0,1862 \left(6404 - \frac{1}{12} \cdot 241 \cdot 341,5 \right) \approx 84,6201$$

$$SST = S_{yy} = \sum_{j=1}^{12} y_j^2 - \frac{1}{n} \left(\sum_{j=1}^{12} y_j \right)^2 = 9813,25 - \frac{1}{12}(341,5)^2 \approx 94,7291$$

Hệ số tương quan: $R^2 = \dfrac{SSR}{SST} = \dfrac{84,6201}{94,7291} = 0,8932$.

Vậy tỷ lệ cây xanh và nhiệt độ trung bình trong mùa hè có mối liên hệ tuyến tính mạnh.

Bài 4

Một bác sĩ dinh dưỡng nghiên cứu một chế độ ăn kiêng và tập thể dục mới để làm giảm lượng đường trong máu của các bệnh nhân bị bệnh tiểu đường. 10 bệnh nhân bị bệnh tiểu đường được chọn để thử nghiệm chương trình này. Bảng số liệu sau cho biết lượng đường trong máu trước và sau khi các bệnh nhân tham gia chương trình:

Trước	268	225	252	192	308	228	246	298	230	185
Sau	106	185	224	110	204	100	212	176	194	205

Số liệu trên có đủ bằng chứng để kết luận rằng chế độ ăn kiêng và tập thể dục mới này có tác dụng làm giảm lượng đường trong máu hay không? Sử dụng mức ý nghĩa $\alpha = 5\%$.

Lời giải:

Đối tượng khảo sát có cỡ mẫu nhỏ $(10 < 30)$.

Ta có độ sai khác d_i giữa mỗi cặp trong quan trắc là:

Trước	268	225	252	192	308	228	246	298	230	185
Sau	106	185	224	110	204	100	212	176	194	205
d_i	162	40	28	82	104	128	34	122	36	-20

Ta tính trung bình mẫu:

$$\overline{d} = \frac{1}{n} \sum_{i=1}^{n} d_i = \frac{1}{10} \sum_{i=1}^{10} d_i = 71,6$$

và phương sai mẫu:

$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^{n} (d_i - \overline{d})^2 = \frac{1}{9} \sum_{i=1}^{10} (d_i - 71,6)^2 = 3224,71$$

Ta tiến hành kiểm định giả thuyết:

$$\begin{cases} H_0 : \mu_d \leq d_0 = 0 & \text{(chế độ ăn kiêng và tập thể dục mới không có tác dụng)} \\ H_1 : \mu_d > d_0 = 0 & \text{(chế độ ăn kiêng và tập thể dục mới có tác dụng)} \end{cases}$$

Thống kê kiểm định:

$$t_0 = \frac{\overline{d} - d_0}{s_d/\sqrt{n}} = \frac{71,6 - 0}{\sqrt{\frac{3224,71}{10}}} \approx 3,9871$$

Ta bác bỏ H_0 nếu $t_0 > t_{1-\alpha}^{n-1} = t_{0,95}^{9} = 1,8331$. Ta có $t_0 = 3,9871 > 1,8331 \Rightarrow$ Bác bỏ H_0.

Vậy với mức ý nghĩa 5%, chế độ ăn kiêng và tập thể dục mới này có tác dụng làm giảm lượng đường trong máu.

Bài 5

Cho một tổng thể tuân theo phân phối xác suất với trung bình μ và phương sai σ^2 với $\sigma > 0$. Xét mẫu ngẫu nhiên (X_1, X_2, X_3) được chọn từ tổng thể này. Với $0 < a < 1$, đặt:

$$M_a = \frac{1}{2}[X_1 + aX_2 + (1-a)X_3]$$

a) Chứng tỏ M_a là một ước lượng không chệch của μ.

Lời giải:

Vì (X_1, X_2, X_3) được chọn từ tổng thể trên nên:

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) = \mathbb{E}(X_3) = \mu, \quad \text{Var}(X_1) = \text{Var}(X_2) = \text{Var}(X_3) = \sigma^2$$

Ta có:

$$\begin{aligned}
\mathbb{E}(M_a) &= \mathbb{E}\left[\frac{1}{2}(X_1 + aX_2 + (1-a)X_3)\right] \\
&= \frac{1}{2}\left[\mathbb{E}(X_1) + a\mathbb{E}(X_2) + (1-a)\mathbb{E}(X_3)\right] \\
&= \frac{1}{2}\left[\mu + a\mu + (1-a)\mu\right] \\
&= \frac{1}{2} \cdot 2\mu = \mu
\end{aligned}$$

$\Rightarrow \text{bias}(M_a) = \mathbb{E}(M_a) - \mu = \mu - \mu = 0$

Vậy M_a là một ước lượng không chệch cho μ.

b) Tính phương sai của M_a.

Lời giải:

Phương sai của M_a:

$$\begin{aligned}
\text{Var}(M_a) &= \text{Var}\left[\frac{1}{2}(X_1 + aX_2 + (1-a)X_3)\right] \\
&= \frac{1}{4}\left[\text{Var}(X_1) + a^2\text{Var}(X_2) + (1-a)^2\text{Var}(X_3)\right] \\
&= \frac{1}{4}\left[\sigma^2 + a^2\sigma^2 + (1-2a+a^2)\sigma^2\right] \\
&= \frac{1}{4}\left(2\sigma^2 - 2a\sigma^2 + 2a^2\sigma^2\right) \\
&= \frac{1-a+a^2}{2}\sigma^2
\end{aligned}$$

c) Trung bình bình phương sai số (MSE) của M_a được cho bởi $\text{MSE}(M_a, \mu) = \mathbb{E}\left[(M_a - \mu)^2\right]$. Tìm giá trị của a sao cho M_a là ước lượng tốt nhất theo nghĩa cực tiểu hóa trung bình bình phương sai số.

Lời giải:

Ta có:

$$\begin{aligned}
\mathrm{MSE}(M_a, \mu) &= \mathbb{E}\left[(M_a - \mu)^2\right] \\
&= \mathbb{E}\left(M_a^2 - 2M_a\mu + \mu^2\right) \\
&= \mathbb{E}(M_a^2) - 2\mu\mathbb{E}(M_a) + \mu^2 \\
&= \mathrm{Var}(M_a) + [\mathbb{E}(M_a)]^2 - 2\mu\mathbb{E}(M_a) + \mu^2 \\
&= \frac{1-a+a^2}{2}\sigma^2 + \mu^2 - 2\mu\mu + \mu^2 \\
&= \frac{1-a+a^2}{2}\sigma^2
\end{aligned}$$

Để cực tiểu hóa MSE, ta cần tìm giá trị a sao cho $f(a) = 1 - a + a^2$ đạt GTNN.

Ta tính $f'(a) = 2a - 1$ và cho $f(a) = 0 \Rightarrow a = \dfrac{1}{2}$

Ta tiếp tục tính $f''\left(\dfrac{1}{2}\right) = 2 > 0$

Vậy với $a = \dfrac{1}{2}$ thì M_a là ước lượng tốt nhất.

Bài 6

Cho hai tổng thể độc lập với nhau, trong đó:

- tổng-thể-1 có trung bình μ_1 và phương sai σ_1^2 với $\sigma_1 > 0$. Mẫu ngẫu nhiên với cỡ mẫu n được chọn từ tổng-thể-1 có trung bình mẫu là $\overline{X}$ và phương sai mẫu là S_1^2.

- tổng-thể-2 có trung bình μ_2 và phương sai σ_2^2 với $\sigma_2 > 0$. Mẫu ngẫu nhiên với cỡ mẫu m được chọn từ tổng-thể-2 có trung bình mẫu là $\overline{Y}$ và phương sai mẫu là S_2^2.

a) Biết sai số chuẩn (standard error) của ước lượng $\hat{\theta}$ được cho bởi

$$\mathrm{SE}(\hat{\theta}) := \sqrt{\mathrm{Var}(\hat{\theta})}$$

Tính sai số chuẩn $\mathrm{SE}(\overline{X} - \overline{Y})$.

Lời giải:

Với X và Y độc lập, trước hết ta tính:

$$\begin{aligned}
\mathrm{Var}(\overline{X} - \overline{Y}) &= \mathrm{Var}(\overline{X}) + \mathrm{Var}(\overline{Y}) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) + \frac{1}{m^2}\mathrm{Var}\left(\sum_{i=1}^{m} Y_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}(Y_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_1^2 + \frac{1}{m^2}\sum_{i=1}^{m}\sigma_2^2 \\
&= \frac{n\sigma_1^2}{n^2} + \frac{m\sigma_2^2}{m^2} = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}
\end{aligned}$$

Vậy ta có $\mathrm{SE}(\overline{X} - \overline{Y}) = \sqrt{\mathrm{Var}(\overline{X} - \overline{Y})} = \sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}$

b) Giả sử hai tổng thể có cùng phương sai, nghĩa là $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Chứng minh rằng:

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n-1+m-1}, \quad n, m \geq 2$$

là một ước lượng không chệch của σ^2.

Lời giải:

Ta có $\mathbb{E}(S_1^2) = \sigma_1^2 = \sigma^2$, $\mathbb{E}(S_2^2) = \sigma_2^2 = \sigma^2$ và:

$$\begin{aligned}
\mathbb{E}(S_p^2) = \mathbb{E}\left[\frac{(n-1)S_1^2 + (m-1)S_2^2}{n-1+m-1}\right] &= \frac{(n-1)\mathbb{E}(S_1^2) + (m-1)\mathbb{E}(S_2^2)}{n+m-2} \\
&= \frac{(n-1)\sigma^2 + (m-1)\sigma^2}{n+m-2} = \frac{(n-1+m-1)\sigma^2}{n+m-2} \\
&= \frac{(n+m-2)\sigma^2}{n+m-2} = \sigma^2
\end{aligned}$$

$\Rightarrow \text{bias}(S_p^2) = \mathbb{E}(S_p^2) - \sigma^2 = \sigma^2 - \sigma^2 = 0$

Vậy S_p^2 là một ước lượng không chệch cho σ^2.