Probabilidades e Estatística

— TODOS OS CURSOS —

2°semestre – 2019/2020
18/07/2020 – **13:00**

Duração: 60 minutos **2°teste**

Pergunta 1

Admita que a proporção de potássio num dado fertilizante é representada pela variável aleatória X com função de densidade de probabilidade

$$f_X(x) = \begin{cases} \theta\,(1-x)^{\theta-1}, & 0<x<1 \\ 0, & \text{caso contrário,} \end{cases}$$

onde θ é um parâmetro positivo desconhecido. Determine a estimativa de máxima verosimilhança da $P(X \le z)$, baseada na amostra $(x_1,\ldots,x_6)$ proveniente da população X.

Preencha a caixa abaixo com o resultado obtido com, pelo menos, quatro casas decimais.

- **V.a. de interesse, f.d.p.**

  X = proporção de potássio num dado fertilizante

  $$f_X(x) = \begin{cases} \theta\,(1-x)^{\theta-1}, & 0<x<1 \\ 0, & \text{caso contrário,} \end{cases}$$

- **Parâmetro desconhecido**

  $\theta, \quad \theta > 0$

- **Obtenção da estimativa de MV de θ**

  Passo 1 — Função de verosimilhança

$$\begin{aligned} L(\theta \mid \underline{x}) &= f_{\underline{X}}(\underline{x}) \\ &\overset{X_i\, indep}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \\ &\overset{X_i \sim X}{=} \prod_{i=1}^{n} f_X(x_i) \\ &= \prod_{i=1}^{n} \left[\theta\,(1-x_i)^{\theta-1}\right] \\ &= \theta^n \left[\prod_{i=1}^{n}(1-x_i)\right]^{\theta-1}, \quad 0<p<1 \end{aligned}$$

  Passo 2 — Função de log-verosimilhança

$$\ln L(\theta \mid \underline{x}) = n\ln(\theta) + (\theta-1)\sum_{i=1}^{n}\ln(1-x_i)$$

  Passo 3 — Maximização

$$\hat{\theta} \quad : \quad \begin{cases} \left.\frac{d\ln L(\theta|\underline{x})}{dp}\right|_{\theta=\hat{\theta}} = 0 & \text{(ponto de estacionaridade)} \\ \left.\frac{d^2\ln L(\theta|\underline{x})}{d\theta^2}\right|_{\theta=\hat{\theta}} < 0 & \text{(ponto de máximo)} \end{cases}$$

$$\begin{cases} \frac{n}{\hat{\theta}} + \sum_{i=1}^{n}\ln(1-x_i) = 0 \\ -\frac{n}{\hat{\theta}^2} < 0 \quad \text{(prop. verdadeira)} \end{cases}$$

Passo 4 — Estimativa de MV de λ

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln(1-x_i)}, \quad n = 6$$

- **Outro parâmetro desconhecido**

  $h(\theta) = P(X \leq z) = \int_0^z \theta\,(1-x)^{\theta-1}\,dx = -\,(1-x)^{\theta}\big|_0^z = 1-(1-z)^{\theta}, \quad 0 < z < 1$

- **Estimativa de MV de $h(\theta)$**

$$\begin{aligned} \widehat{h(\theta)} &\overset{prop.inv.EMV}{=} h(\hat{\theta}) \\ &= 1-(1-z)^{\hat{\theta}}. \end{aligned}$$

Pergunta 2

Para avaliar os consumos de duas marcas de carros elétricos, foi definido um percurso de 85 km, com pouco tráfego. Durante os testes os carros deviam manter uma velocidade de 100 km/h. Seja X_A (respetivamente, X_B) a variável aleatória que representa o consumo do carro da marca A (respetivamente, B), medido em kWh/100 km. Para realizar estes testes foram escolhidos ao acaso n_A veículos da marca A e n_B veículos da marca B. Considere que todos os testes foram realizados de forma independente e que se obtiveram os seguintes resultados: $\bar{x}_A = \bar{x}_A$, $s_A^2 = s_A^2$, $\bar{x}_B = \bar{x}_B$, $s_B^2 = s_B^2$.

Determine um intervalo aproximado de confiança a $(1-\alpha) \times 100\%$ para a diferença entre os valores esperados dos consumos dos carros elétricos das marcas A e B.

- **V.a. de interesse**

  $X_i =$ consumo dos carro da marca $i, \quad i = A, B$

- **Situação**

  X_i v.a. com dist. arbitrária, valor esperado μ_i e variância σ_i^2, $i = A, B$

  $X_A \perp\!\!\!\perp X_B$

  $(\mu_1 - \mu_2)$ DESCONHECIDO

  σ_A^2 e σ_B^2 desconhecidas [não necessariamente iguais]

  n_A e n_B suficientemente grandes

- **Obtenção do IC aproximado para $\mu_A - \mu_B$**

  Passo 1 — Selecção da v.a. fulcral para $\mu_A - \mu_B$

$$Z = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}} \overset{a}{\sim} \text{normal}(0,1)$$

  Passo 2 — Obtenção dos quantis de probabilidade

$$\begin{cases} a_\alpha = \Phi^{-1}(\alpha/2) \\ b_\alpha = \Phi^{-1}(1-\alpha/2) \end{cases}$$

  Passo 3 — Inversão da desigualdade $a_\alpha \leq Z \leq b_\alpha$

$$P(a_\alpha \leq Z \leq b_\alpha) \simeq 1-\alpha$$

$$P\left[a_\alpha \leq \frac{(\bar{X}_A-\bar{X}_B)-(\mu_A-\mu_B)}{\sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}}} \leq b_\alpha\right] \simeq 1-\alpha$$

$$P\left[(\bar{X}_A - \bar{X}_B) - b_\alpha \times \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}} \leq \mu_1 - \mu_2 \leq (\bar{X}_A - \bar{X}_B) - a_\alpha \times \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}\right] \simeq 1-\alpha$$

Passo 4 — Concretização

$$IC_{(1-\alpha)\times 100\%}(\mu_A - \mu_B) \simeq \left[(\bar{x}_A - \bar{x}_B) \pm \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} \right].$$

Pergunta 3

O número de testes de despistagem do vírus SARS-CoV-2 (Covid-19) efetuados por hora por uma institução de saúde segue uma distribuição de Poisson com variância λ desconhecida. Foi contabilizado um total de x_{Sum} testes efetuados em n períodos de uma hora escolhidos ao acaso. Confronte as hipóteses $H_0 : \lambda = \lambda_0$ e $H_1 : \lambda > \lambda_0$, calculando para o efeito o valor-p aproximado.

- **V.a. de interesse**

  X = número de testes de despistagem por hora

- **Situação**

  $X \sim \text{Poisson}(\lambda)$

  $\lambda = E(X) = V(X)$ DESCONHECIDO

  $n > 30$

- **Hipóteses**

  $H_0 : \lambda = \lambda_0$

  $H_1 : \lambda > \lambda_0$

- **Estatística de teste**

  $$T = \frac{\bar{X} - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}} \overset{a}{\sim}_{H_0} \text{normal}(0,1)$$

- **Região de rejeição de** H_0 (para valores da estatística de teste)

  O teste é unilateral superior ($H_1 : \lambda > \lambda_0$), logo a região de rejeição de H_0 é do tipo $W = (c, +\infty)$.

- **Decisão (com base no p-value)**

  $$\begin{aligned} valor-p &= P\left(T > t = \frac{x_{Sum} - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}} \mid H_0 \right) \\ &\simeq 1 - \Phi\left(\frac{x_{Sum}/n - \lambda_0}{\sqrt{\lambda_0/n}} \right) \end{aligned}$$

  Devemos escolher a resposta, das quatro apresentadas em nota de rodapé,[1] que se coaduna com:

  - a não rejeição de H_0 a qualquer n.s. $\alpha_0 \leq valor-p$;
  - a rejeição de H_0 a qualquer n.s. $\alpha_0 > valor-p$.

Pergunta 4

Suspeita-se que o ruído máximo noturno, medido em decibéis (db) em zona residencial próxima de um aeroporto, é uma variável aleatória X que segue uma distribuição normal.

Foram recolhidos os seguintes dados relativos a n medições nocturnas, em dias escolhidos ao acaso:

[1] 1. Rejeita-se para 1%, 5% e 10%. 2. Rejeita-se para 5% e 10% e não se rejeita para 1%. 3. Rejeita-se para 10% e não se rejeita para 1% e 5%. 4. Não se rejeita para 1%, 5% e 10%.

Ruído máximo nocturno	≤ 50	$]50,55]$	$]55,60]$	>60
Frequência absoluta observada	o_1	o_2	o_3	o_4

Avalie a hipótese H_0 de que X possui distribuição normal de valor esperado 55 e desvio padrão σ.

Decida com base no valor-p aproximado e nas frequências absolutas esperadas sob H_0 aproximadas às centésimas.

- **V.a. de interesse**

  X = ruído máximo noturno

- **Hipóteses**

  $H_0 : X \sim \text{normal}(55, \sigma^2)$

  $H_1 : \neg H_0$

- **Estatística de teste**

  $T = \sum_{i=1}^{k} \frac{(O_i - E_i}{E_i} \overset{a}{\sim}_{H_0} \chi_{(k-\beta-1)}$,

  onde: k = no. de classes = 4; β = 0; o_i = freq. abs. observada da classe i; E_i = freq. abs. esperada, sob H_0, da classe i.

- **Região de rejeição de H_0**

  Ao lidarmos com um teste de ajustamento do qui-quadrado, a região de rejeição de H_0 é um intervalo à direita $W = (c, +\infty)$.

- **Frequências absolutas esperadas sob H_0**

$$
\begin{aligned}
E_1 &= n \times P(X \leq 50 \mid H_0) = n \times \Phi\left(\frac{50-55}{\sigma}\right) \\
E_4 &= n - \sum_{i=1}^{3} E_i
\end{aligned}
$$

- **Decisão (com base no p-value)**

$$
valor-p \quad = \quad P\left(T > t = \frac{(o_i - E_i)^2}{E_i} \mid H_0\right) \simeq 1 - F_{\chi^2_{(k-1)}}\left(\sum_{i=1}^{k} \frac{(o_i - E_i)^2}{E_i}\right)
$$

  Devemos escolher a resposta, das quatro apresentadas em nota de rodapé,[2] que se coaduna com:

  - a não rejeição de H_0 a qualquer n.s. $\alpha_0 \leq valor-p$;
  - a rejeição de H_0 a qualquer n.s. $\alpha_0 > valor-p$.

Pergunta 5

Num estudo para avaliar a relação entre a densidade populacional (Y, em centenas de habitantes por quilómetro quadrado) das áreas residenciais de determinada cidade e a distância (x, em quilómetro) ao centro da mesma, recolheu-se uma amostra casual de dimensão n que conduziu aos seguintes valores:

$$
\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i, \quad \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2, \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i, \quad \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i^2, \quad \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i y_i.
$$

Os responsáveis por este estudo consideraram um modelo de regressão linear simples e defendem a conjetura $H_0 : \beta_0 = \beta_{0,0}$, ao passo que os membros de uma instituição de urbanismo local defendem a hipótese $H_1 : \beta_0 \neq \beta_{0,0}$.

Confronte estas duas hipóteses e decida com base no valor-p.

---

[2] 1. Rejeita-se para 1%, 5% e 10%. 2. Rejeita-se para 5% e 10% e não se rejeita para 1%. 3. Rejeita-se para 10% e não se rejeita para 1% e 5%. 4. Não se rejeita para 1%, 5% e 10%.

- **Hipóteses de trabalho**

  $\epsilon_i \stackrel{i.i.d.}{\sim} \text{normal}(0,\sigma^2),\ i=1,\ldots,n$

- **Hipóteses**

  $H_0 : \beta_0 = \beta_{0,0}$

  $H_1 : \beta_0 \neq \beta_{0,0}$

- **Estatística de teste**

$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \times \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2}\right)}} \sim_{H_0} t_{(n-2)}$$

- **Região de rejeição de H_0 (para valores da estatística de teste)**

  Dado que o teste é bilateral, a região de rejeição de H_0 é a reunião de intervalos $W = (-\infty,-c) \cup (c,+\infty)$.

- **Valor observado da estatística de teste**

$$t = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\left(\frac{\sum_{i=1}^n x_i}{n}\right)^2}\right)}}$$

  onde

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&= \bar{y} - \frac{\sum_{i=1}^n x_i y_i - n \times \frac{\sum_{i=1}^n x_i}{n} \times \frac{\sum_{i=1}^n y_i}{n}}{\sum_{i=1}^n x_i^2 - n \times \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2} \times \bar{x} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\left(\frac{\sum_{i=1}^n x_i}{n}\right)\left(\frac{\sum_{i=1}^n y_i}{n}\right)}{\sum_{i=1}^n x_i^2 - n\left(\frac{\sum_{i=1}^n x_i}{n}\right)^2} \\
\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right) - (\hat{\beta}_1)^2\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right)\right] \\
&= \frac{1}{n-2}\left\{\left[\sum_{i=1}^n y_i^2 - n\left(\frac{\sum_{i=1}^n y_i}{n}\right)^2\right] - (\hat{\beta}_1)^2\left[\sum_{i=1}^n x_i^2 - n\left(\frac{\sum_{i=1}^n x_i}{n}\right)^2\right]\right\}
\end{aligned}$$

  [Convinha que t fosse calculado com base nas fórmulas acima que tiram partido do valor de n e das somas que se encontram no enunciado do problema.]

- **Decisão (com base no valor-p)**

  $valor - p = 2 \times \left[1 - F_{t_{(n-2)}}(|t|)\right]$

  Devemos escolher a resposta, das quatro apresentadas em nota de rodapé,[3]

---

[3] 1. Rejeita-se para 1%, 5% e 10%. 2. Rejeita-se para 5% e 10% e não se rejeita para 1%. 3. Rejeita-se para 10% e não se rejeita para 1% e 5%. 4. Não se rejeita para 1%, 5% e 10%. que se coaduna com:

- a não rejeição de H_0 a qualquer n.s. $\alpha_0 \leq valor - p$;
- a rejeição de H_0 a qualquer n.s. $\alpha_0 > valor - p$.