Probabilidades e Estatística

LEBiol21, LEIC-A, LEQ21; LEIC-T

2º Semestre – 2022/2023
14/07/2023
13:00–15:00

Duração: **120** minutos | **Exame Época Recurso – A**

- As questões de escolha múltipla têm uma e apenas uma resposta correta. Caso assinale uma resposta errada, serão descontados 0.5 valores.
- Se assinalar mais do que uma resposta numa questão de escolha múltipla, o resultado será classificado como errado e serão descontados 0.5 valores.
- As respostas às questões de escolha múltiplas têm de ser assinaladas na folha de instruções previamente distribuída.
- Para as restantes questões, terá de justificar convenientemente as suas respostas.

Pergunta 1 — 2 valores

Uma dada universidade possui 40% de estudantes de áreas TIC e os restantes estudantes são de outras áreas. De entre os estudantes de áreas TIC, 5% têm mais de 25 anos de idade. De entre os estudantes de outras áreas, 2% têm mais de 25 anos de idade.

Tendo sido escolhido ao acaso um estudante da universidade e verificando-se que o mesmo tem mais de 25 anos de idade, calcule a probabilidade de ter sido escolhido um estudante de áreas TIC.

- **Acontecimentos e probabilidades para um estudante da universidade escolhido ao acaso**

Acontecimento	Probabilidade
T = "escolha de um estudante de áreas TIC"	$P(T) = 0.4$
M = "escolha de um estudante com mais de 25 anos de idade"	$P(M) = ?$
	$P(M \mid T) = 0.05$
	$P(M \mid \bar{T}) = 0.02$

- **Cálculo da probabilidade pedida**

  Ao recorrer-se ao teorema de Bayes, temos

$$\begin{aligned}
P(T \mid M) &= \frac{P(M \mid T) \times P(T)}{P(M)} \\
&= \frac{P(M \mid T) \times P(T)}{P(M \mid T) \times P(T) + P(M \mid \bar{T}) \times P(\bar{T})} \\
&= \frac{0.05 \times 0.4}{0.05 \times 0.4 + 0.02 \times (1 - 0.4)} \\
&= \frac{0.02}{0.02 + 0.012} \\
&= 0.625.
\end{aligned}$$

Pergunta 2 — 2 valores

Segundo dados recentes, a percentagem de indíviduos da população portuguesa com sangue do tipo O^- é de 6%.

Qual é a probabilidade de, em 20 dadores de sangue escolhidos ao acaso de tal população, pelo menos um mas não mais de quatro dadores ter este tipo sangue?

- **V.a.**

  $X =$ número de dadores com sangue do tipo O^-, em 20 escolhidos ao acaso

- **Distribuição**

  $X \sim \text{binomial}(n, p)$

  $n = 20$

  $p = P(\text{dador com sangue do tipo } O^-) = 0.06$

- **F.p. de** X

  $P(X = x) = \binom{20}{x} 0.06^x (1-0.06)^{20-x}, \quad x = 0, 1, \ldots, 20$

- **Probabilidade pedida**

$$
\begin{aligned}
P(1 \leq X \leq 4) &= \sum_{x=1}^{4} P(X = x) \\
&= \sum_{x=1}^{4} \binom{20}{x} 0.06^x (1-0.06)^{20-x} \\
&\simeq 0.7043.
\end{aligned}
$$

  Alternativamente, $P(1 \leq X \leq 4) = F_X(4) - F_X(1^-) = F_X(4) - F_X(0) \stackrel{tabelas,calc.}{=} 0.9944 - 0.2901 = 0.7043.$

Pergunta 3	2 valores

Considere que o número semanal de acidentes rodoviários, na localidade A, é uma variável aleatória de Poisson com variância igual a 3, e, na localidade B, é uma variável aleatória de Poisson com valor esperado igual a 5. Suponha que estas duas variáveis aleatórias são independentes.

Determine a probabilidade de o número total de acidentes rodoviários nessas duas localidades numa semana exceder três.

- **V.a.**

  $X =$ número semanal de acidentes rodoviários na localidade A

  $Y =$ número semanal de acidentes rodoviários na localidade B

  $X \sim \text{Poisson}(\lambda_X = 3)$, pois $V(X) = 3 = \lambda_X$
  $\perp\!\!\!\perp$
  $Y \sim \text{Poisson}(\lambda_Y = 5)$, pois $E(Y) = 5 = \lambda_Y$

- **Importante**

  A soma de duas v.a. independentes com distribuição de Poisson possui também distribuição de Poisson: $\quad Z = X + Y \sim \text{Poisson}(\lambda_Z = \lambda_X + \lambda_Y = 3 + 5 = 8)$.

- **Probabilidade pedida**

$$
P(Z > 3) = 1 - P(Z \leq 3) = 1 - F_Z(3) \stackrel{tabelas,calc.}{=} 1 - 0.0424 = 0.9576.
$$

Pergunta 4	2 valores

Um atleta admite que os tamanhos dos seus passos de corrida são variáveis aleatórias independentes e identicamente distribuídas com valor esperado de 0.97 metros e desvio padrão de 0.1 metros.

Determine um valor aproximado para o quantil de probabilidade 0.1 do tamanho médio de 100 dos seus passos de corrida.

- **V.a.; valor esperado e variância comuns**

  X_i = tamanho (em metro) do passo de corrida i, $\quad i = 1, \ldots, n$

  $n = 100 >> 30$

  $X_i \overset{i.i.d.}{\sim} X, \quad i = 1, \ldots, n$

  $E(X_i) = E(X) = \mu = 0.97$

  $V(X_i) = V(X) = 0.1^2$

- **V.a. de interesse**

  $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ = tamanho médio de n passos de corrida

- **Valor esperado e variância de $\bar{X}_n$**

  $E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) \overset{X_i \sim X}{=} \frac{n\,E(X)}{n} = E(X) = \mu$

  $V(\bar{X}_n) = V\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \overset{X_i\ indep.}{=} \frac{1}{n^2}\sum_{i=1}^n V(X_i) \overset{X_i \sim X}{=} \frac{n\,V(X)}{n^2} = \frac{V(X)}{n} = \frac{\sigma^2}{n}$

- **Distribuição aproximada de $\bar{X}_n$**

  De acordo com o teorema do limite central (TLC),

$$\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{V(\bar{X}_n)}} = \frac{\bar{X}_n - E(X)}{\sqrt{\frac{V(X)}{n}}} = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \overset{a}{\sim} \text{normal}(0,1).$$

- **Quantil pedido (valor aproximado)**

$$\xi \;:\; P(\bar{X}_n \leq \xi) = 0.1 \quad \Leftrightarrow \quad P\left(\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \leq \frac{\xi - \mu}{\sqrt{\sigma^2/n}}\right) = 0.1 \quad \Leftrightarrow \quad \Phi\left(\frac{\xi - \mu}{\sqrt{\sigma^2/n}}\right) \overset{TLC}{\simeq} 0.1$$

$$\frac{\xi - \mu}{\sqrt{\sigma^2/n}} \simeq \Phi^{-1}(0.1) \quad \Leftrightarrow \quad \xi \simeq \mu + \frac{\sigma}{\sqrt{n}} \times \Phi(0.1) \quad \Leftrightarrow \quad \xi \overset{tabelas/calc.}{\simeq} 0.97 + \frac{0.1}{\sqrt{100}} \times (-1.2816)$$

$$\xi \simeq 0.957184.$$

Pergunta 5	2 valores

Seja (X, Y) um par aleatório contínuo com função de densidade de probabilidade conjunta

$$f_{X,Y}(x,y) = \begin{cases} 2, & 0 \leq x \leq 1, 0 \leq y \leq x \\ 0, & \text{caso contrário.} \end{cases}$$

Deduza a função de densidade de probabilidade de X.

- **Par aleatório**

  (X, Y) com a f.d.p. conjunta do enunciado

- **F.d.p. marginal de X**

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy \\ &= \int_0^x 2\,dy = 2y\Big|_0^x\,dx = 2x, \quad 0 \leq x \leq 1. \end{aligned}$$

Pergunta 6	2 valores

Admita que a variável aleatória X, que representa o número de linhas de código sem erros de sintaxe até à deteção da primeira linha de código com erros de tal tipo, possui função de probabilidade

$$P(X=x)=\begin{cases}(1-p)^x\,p, & x=0,1,2,\dots\\ 0, & \text{caso contrário,}\end{cases}$$

onde p é um parâmetro desconhecido.

Deduza o estimador de máxima verosimilhança de p, baseado numa amostra aleatória de dimensão n proveniente de X.

- **V.a. de interesse; f.p.**

  X

  $$P(X=x)=\begin{cases}(1-p)^x\,p, & x=0,1,2,\dots\\ 0, & \text{caso contrário,}\end{cases}$$

- **Parâmetro desconhecido**

  $p \quad (p\in[0,1])$

- **Amostra; amostra aleatória**

  $\underline{x}=(x_1,\dots,x_n)$, amostra de dimensão n proveniente de X.

  $\underline{X}=(X_1,\dots,X_n)$, amostra aleatória de dimensão n proveniente de X.

- **Dedução do estimador de MV de** p

  Passo 1 — Função de verosimilhança

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

  Passo 2 — Função de log-verosimilhança

$$\ln L(p|\underline{x}) = n\ln(p)+\ln(1-p)\times\sum_{i=1}^n x_i, \quad p\in[0,1]$$

  Passo 3 — Maximização

  A estimativa de MV de p é representada por $\hat{p}$ e

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

$$\begin{cases}\frac{n}{\hat{p}}-\frac{\sum_{i=1}^n x_i}{1-\hat{p}}=0\\ -\frac{n}{\hat{p}^2}-\frac{\sum_{i=1}^n x_i}{(1-\hat{p})^2}<0 \quad \text{(prop. verdadeira)}\end{cases} \underset{\Leftrightarrow}{\hat{p}\neq 0,1} \begin{cases}n-\hat{p}\left(n+\sum_{i=1}^n x_i\right)=0 \Leftrightarrow \hat{p}=\frac{n}{n+\sum_{i=1}^n x_i}\\ -\end{cases}$$

  Passo 4 — Estimador de MV de p

$$EMV(p) = \frac{n}{n+\sum_{i=1}^n X_i}=(1+\bar{X})^{-1}$$

Pergunta 7	2 valores

Admita que o peso máximo suportado por uma marquesa portátil ergonómica é uma variável aleatória X com distribuição normal. Foram selecionados aleatoriamente 28 dessas marquesas, tendo-se observado uma variância amostral corrigida de $s^2 = 3.8$.

Qual é o intervalo de confiança a 90% para a variância de X?

A: [2.482, 6.060] **B**: [1.599, 2.521] **C**: [2.558, 6.353] **D**: [2.376, 7.042]

- **V.a. de interesse**

  X = peso máximo suportado por uma marquesa portátil ergonómica

- **Situação**

  $X \sim \text{normal}(\mu, \sigma^2)$, $\mu = E(X)$ desconhecido, $\sigma^2 = V(X)$ DESCONHECIDO

- **Obtenção do IC para σ^2**

  Passo 1 — Variável aleatória fulcral para σ^2

$$Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

  Passo 2 — Quantis de probabilidade

  Dado que $n = 28$ e $(1-\alpha) \times 100\% = 90\%$, usaremos os quantis

$$(a_\alpha, b_\alpha) : \begin{cases} P(a_\alpha \le Z \le b_\alpha) = 1-\alpha \\ P(Z < a_\alpha) = P(Z > b_\alpha) = \alpha/2 \end{cases} \Leftrightarrow \begin{cases} a_\alpha = F^{-1}_{\chi^2_{(n-1)}}(\alpha/2) = F^{-1}_{\chi^2_{(27)}}(0.05) \overset{tabela/calc.}{=} 16.15 \\ b_\alpha = F^{-1}_{\chi^2_{(n-1)}}(1-\alpha/2) = F^{-1}_{\chi^2_{(27)}}(0.95) \overset{tabela/calc.}{=} 40.11. \end{cases}$$

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

$$P(a_\alpha \le Z \le b_\alpha) = 1-\alpha$$
$$P\left[a_\alpha \le \tfrac{(n-1)S^2}{\sigma^2} \le b_\alpha\right] = 1-\alpha$$
$$P\left[\tfrac{(n-1)S^2}{b_\alpha} \le \sigma^2 \le \tfrac{(n-1)S^2}{a_\alpha}\right] = 1-\alpha$$

  Passo 4 — Concretização

  Atendendo à expressão geral do IC para σ^2,

$$IC_{(1-\alpha)\times 100\%}(\sigma^2) = \left[\frac{(n-1)\,s^2}{F^{-1}_{\chi^2_{(n-1)}}(1-\alpha/2)}, \frac{(n-1)\,s^2}{F^{-1}_{\chi^2_{(n-1)}}(\alpha/2)}\right],$$

  ao par de quantis acima e ao facto de $s^2 = 3.8$, temos:

$$\begin{aligned} IC_{90\%}(\sigma) &= \left[\frac{(28-1)\times 3.8}{40.11}, \frac{(28-1)\times 3.8}{16.15}\right] \\ &\simeq [2.558, 6.353]. \end{aligned}$$

Pergunta 8	2 valores

Suponha que o comprimento (em mm) de uma peça é uma variável aleatória com distribuição normal com valor esperado μ desconhecido e variância σ^2 desconhecida.

Teste a hipótese $H_0 : \mu = 100$ contra $H_1 : \mu < 100$, num dia em que foram recolhidas casualmente 4 peças e a média e a variância corrigida amostrais foram de $\bar{x} = 101.0$ e $s^2 = 2.56$. Decida com base no valor-p.

- **V.a. de interesse**

  X = comprimento de uma peça metálica

- **Situação**

  $X \sim \text{normal}(\mu, \sigma^2)$

  μ DESCONHECIDO

  σ desconhecido

- **Hipóteses**

  $H_0 : \mu = \mu_0 = 100$ vs. $H_1 : \mu < \mu_0$

- **Estatística de teste**

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim_{H_0} t_{(n-1)}$$

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

  O teste é unilateral inferior $(H_1 : \mu < \mu_0)$, logo a região de rejeição de H_0 é tipo $(-\infty, c)$.

  Decisão (com base no valor-p)

  Atendendo a que o valor observado da estatística de teste e o valor-p aproximado são iguais a

$$\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \\
&= \frac{101.0 - 100}{\frac{\sqrt{2.56}}{\sqrt{4}}} \\
&\simeq 1.25 \\
valor-p &= P(T < t \mid H_0) \\
&= F_{t_{(n-1)}}(t) \\
&= F_{t_{(3)}}(1.25) \\
&\stackrel{tabelas/calc.}{=} 0.85
\end{aligned}$$

  é suposto:

  - não rejeitar H_0 a qualquer n.s. $\alpha_0 \le valor - p = 85\%$, nomeadamente ao níveis usuais de significância $(1\%, 5\%, 10\%)$[;
  - rejeitar H_0 a qualquer n.s. $\alpha_0 > valor - p = 85\%$].

Pergunta 9	2 valores

O engenheiro que gere uma linha de apoio *online* pretende averiguar se o tempo de assistência (em minutos) possui função de distribuição dada por $F_0(x) = 1 - e^{-(x/2)^2}$, para $x > 0$ (hipótese H_0).

Recorra ao teste de ajustamento do qui-quadrado e à tabela de frequências seguinte para testar H_0 ao nível de significância de 1%.

Tempo de assistência	$]0, 2/3]$	$]2/3, 2]$	$]2, 3]$	$]3, +\infty[$
Frequência absoluta observada	13	32	13	6
Frequência absoluta esperada sob H_0	6.73	33.73	16.80	6.74

- **V.a. de interesse**

  X = tempo de assistência (em minutos)

- **Hipóteses**

  $H_0 : F_X(x) = P(X \le x) = F_0(x) = 1 - e^{-(x/2)^2}, \quad x > 0$

  $H_1 : F_X(x) \ne F_0(x), \quad$ para algum $x > 0$

- **Nível de significância**

  $\alpha_0 = 1\%$

- **Estatística de teste**

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{a}{\sim}_{H_0} \chi^2_{(k-1)},$$

  onde: k = no. de classes = 4; O_i = frequência absoluta observável da classe i; E_i = frequência absoluta esperada, sob H_0, da classe i.

- [**Frequências absolutas esperadas sob** H_0 — De acordo com a tabela facultada, as frequências absolutas esperadas sob H_0 aproximadas às centésimas são: $E_1 \simeq 6.73$; $E_2 \simeq 33.73$; $E_3 \simeq 16.80$; $E_4 \simeq 6.74$. Não é necessário fazer qualquer agrupamento de classes uma vez que se verifica $E_i \ge 5$, em pelo menos 80% das classes, e que $E_i \ge 1$, para todo o i. Caso fosse preciso efectuar agrupamento de classes, os valores de k e $c = F^{-1}_{\chi^2_{(k-1)}}(1-\alpha_0)$ teriam que ser recalculados...]

- **Região de rejeição de** H_0 (para valores de T)

  Lidamos com um teste de ajustamento, donde a região de rejeição de H_0 é o intervalo à direita $W = (c, +\infty)$, onde

$$c = F^{-1}_{\chi^2_{(k-\beta-1)}}(1-\alpha_0) = F^{-1}_{\chi^2_{(4-1)}}(1-0.01) \overset{tabela/calc.}{=} 11.34.$$

- **Decisão**

i	Classe i	Freq. abs. obs. o_i	Freq. abs. esp. sob H_0 E_i	Parcelas valor obs. estat. teste $\frac{(o_i-E_i)^2}{E_i}$
1	$]0,2/3]$	13	6.73	$\frac{(13-6.73)^2}{6.73} \simeq 5.841$
2	$]2/3,2]$	32	33.73	0.089
3	$]2,3]$	13	16.80	0.860
4	$]3,+\infty[$	6	6.74	0.081
		$\sum_{i=1}^k o_i = n = 64$	$\sum_{i=1}^k E_i = n = 64$	$t = \sum_{i=1}^k \frac{(o_i-E_i)^2}{E_i} \simeq 6.871$

  Uma vez que $t \simeq 6.871 \not\in W = (11.34, +\infty)$, não devemos rejeitar H_0 ao n.s. de $\alpha_0 = 1\%$ [nem a qualquer outro n.s. inferior a α_0].

Pergunta 10	2 valores

Uma engenheira biomédica está a investigar, em pacientes hemiplégicos, a relação entre o tempo de realização do teste *timed up and go* (x, em segundos) e a frequência cardíaca (Y, em batimentos por minuto) após a realização deste teste. Numa amostra de 10 pacientes, obtiveram-se os seguintes resultados:

$$\sum_{i=1}^{10} x_i = 202, \quad \sum_{i=1}^{10} x_i^2 = 4328, \quad \sum_{i=1}^{10} y_i = 801, \quad \sum_{i=1}^{10} y_i^2 = 65343, \quad \sum_{i=1}^{10} x_i y_i = 16581.$$

Admita que as variáveis x e Y estão relacionadas de acordo com o modelo de regressão linear simples, $Y = \beta_0 + \beta_1 x + \epsilon$.

Estime $E(Y \mid x = x_0 = \bar{x})$. Calcule o coeficiente de determinação e interprete o seu valor.

- **[Modelo de RLS**

  Y = frequência cardíaca (v.a. resposta)

  x = tempo de realização do teste (variável explicativa)

  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$

  $E(\epsilon_i) = 0,\ V(\epsilon_i) = \sigma^2,\ i = 1, \ldots, n, \quad cov(\epsilon_i, \epsilon_j) = 0,\ i \neq j,\ i, j = 1, \ldots, n$]

- **Estimativas de MQ dos parâmetros desconhecidos β_0, β_1 e $E(Y \mid x = x_0 = \bar{x})$**

  Uma vez que

  - $n = 10$
  - $\sum_{i=1}^n x_i = 202$
    $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{202}{10} = 20.2$
    $\sum_{i=1}^n x_i^2 = 4\,328$
    $\sum_{i=1}^n x_i^2 - n\,\bar{x}^2 = 4\,328 - 10 \times 20.2^2 = 247.6$
  - $\sum_{i=1}^n y_i = 801$
    $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{801}{10} = 80.1$
    [$\sum_{i=1}^n y_i^2 = 65\,343$
    $\sum_{i=1}^n y_i^2 - n\,\bar{y}^2 = 65\,343 - 10 \times 80.1^2 = 1\,182.9$]
  - $\sum_{i=1}^n x_i y_i = 16\,581$
    $\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y} = 16\,581 - 10 \times 20.2 \times 80.1 = 400.8.$

  as estimativas de MQ dos parâmetros desconhecidos β_1, β_0 e $\beta_0 + \beta_1 \times x_0$ são dadas por:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2} = \frac{400.8}{247.6} \simeq 1.618740. \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \simeq 80.1 - (1.618740) \times 20.2 \simeq 47.401454. \\
\hat{E}(Y \mid x = x_0 = \bar{x}) &= \hat{\beta}_0 + \hat{\beta}_1 \times \bar{x} = (\bar{y} - \hat{\beta}_1 \times \bar{x}) + \hat{\beta}_1 \times \bar{x} = \bar{y} = 80.1.
\end{aligned}$$

- **Cálculo do coeficiente de determinação**

$$\begin{aligned}
r^2 &= \frac{\left(\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right)} \\
&= \frac{400.8^2}{247.6 \times 1\,182.9} \\
&\simeq 0.548475.
\end{aligned}$$

- **Interpretação coeficiente de determinação**

  Cerca de 55% da variação total da variável resposta Y é explicada pela variável x, através do modelo de regressão linear simples ajustado, donde possamos afirmar que a recta estimada parece ajustar-se relativamente bem ao conjunto de dados.