ISEG – Lisbon School of Economics and Management

ESTATÍSTICA II

1º Semestre 2023/2024

Exame – EN (Topicos de Resolução)

27 de Maio de 2024

Duração: **2 horas**

Nome:

Questão	**1**	**2**	**3**	**4**	**5**	Total
Pontuação	20	35	30	40	75	200

- **Justifique** todas as suas respostas. Deverá apresentar todas as fórmulas utilizadas, assim como os respetivos cálculos auxiliares;
- Use **3 folhas separadas de teste**: para as perguntas i) 1-2, ii) 3-4, e iii) 5. Escreva o seu **nome e número** de aluno em todas as folhas do exame que vai entregar.
- **Organize as suas respostas**. Respostas espalhadas e rasuradas na folha de teste, receberão pouco crédito;
- **É permitida a utilização** do formulário da disciplina (sem qualquer anotação), das tabelas estatísticas e de uma calculadora científica. Considere sempre um **nível de significância de 5**%, a não ser que outro nível de significância esteja especificado.
- **Verdadeiro ou Falso:** resposta certa vale 2 pontos; resposta errada vale −1 ponto; ausência de resposta ou resposta dúbia vale 0 pontos.

VERDADEIRO OU FALSO

(20) **1**. Das afirmações seguintes, assinale com uma cruz (**X**) as que são verdadeiras e as que são falsas.

		V	**F**
1	Um estimador é uma variavel aleatoria, função da amostra ($T(X_1, ..., X_n)$), e uma estimativa é um valor concreto assumido pelo estimador para uma dada amostra ($T(x_1, ..., x_n)$).	X	
2	A função de verosimilhança é uma função em θ, fixada a amostra.	X	
3	O erro de 1ª espécie consiste em rejeitar H_0, quando H_0 é verdadeira, enquanto que o erro de 2ª espécie consiste em rejeitar H_0, quando H_0 é falsa.		X
4	Um intervalo aleatório (IA) cobre o parâmetro θ com determinada probabilidade, enquanto que um intervalo de confiança é uma concretização deste IA para uma determinada amostra.	X	
5	O valor-p, tambem chamado nível de significância associado à estatística teste observada, é a probabilidade de obter um valor tão ou mais extremo/desfavorável para a hipótese nula do que o observado para uma amostra obtida.	X	
6	Considere o modelo $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, em que a correlação entre x_1 e x_2 é igual a 0.95. Nesse caso, o modelo sofre de colinearidade perfeita.		X
7	Considere o modelo $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, com $\beta_2 > 0$. Suponha que, em alternativa, foi estimado o modelo $y_i = \delta_0 + \delta_1 x_{1i} + e_i$. No segundo caso o R^2 é menor.	X	
8	Sob as hipóteses do teorema de Gauss-Markov, é possível encontrar outro estimador com variância mais baixa que o estimador OLS.	X	
9	Considere o modelo $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$. Se se rejeitarem as hipóteses $H_0 : \beta_1 = 0$ e $H_0 : \beta_2 = 0$, então de certeza que se rejeita $H_0 : \beta_1 = \beta_2 = 0$.		X
10	Não é possível utilizar o método OLS para estimar directamente o modelo $y_i = \exp(\beta_0 + \beta_1 x_i) \exp(u_i)$.	X	

V.S.F.F.

RESPOSTA ABERTA

2. Seja X uma variável aleatória discreta com função de probabilidade dada por:

$$f_X(x|\theta) = \begin{cases} 0.5 - \theta & (x = 0, \text{ou } 1) \\ \theta & (x = 2, \text{ou } 3), \end{cases}$$

onde θ é um parâmetro desconhecido que verifica $0 < \theta < 0.5$.

(5) (a) Verifique que $E(X) = \sum_x x f(x|\theta) = \frac{1}{2} + 4\theta$

RESPOSTA:

$$E(X) = \sum_{x=0}^{3} x\, f_X(x|\theta) = 0\left(\frac{1}{2} - \theta\right) + 1\left(\frac{1}{2} - \theta\right) + 2\theta + 3\theta = \frac{1}{2} - \theta + 5\theta = \frac{1}{2} + 4\theta$$

(10) (b) Obtenha um estimador para θ usando o método dos momentos.

RESPOSTA:

$$E(X) = \overline{X} \Leftrightarrow \frac{1}{2} + 4\theta = \overline{X} \Leftrightarrow$$
$$\Leftrightarrow 4\theta = \overline{X} - \frac{1}{2} \Leftrightarrow \theta = \frac{\overline{X} - \frac{1}{2}}{4}$$

Conclusão: O estimador obtido pelo método dos momentos é:

$$\tilde{\theta} = \frac{\overline{X} - \frac{1}{2}}{4} \tag{1}$$

(10) (c) O estimador obtido na alínea anterior é centrado? E é consistente?

RESPOSTA:

$$E(\tilde{\theta}) = E\left(\frac{\overline{X} - \frac{1}{2}}{4}\right) = \frac{E(\overline{X}) - \frac{1}{2}}{4} = \frac{E(X) - \frac{1}{2}}{4} = \frac{\frac{1}{2} + 4\theta - \frac{1}{2}}{4} = \theta$$

O estimador $\tilde{\theta}$ é centrado.

$$\text{Var}(\tilde{\theta}) = \text{Var}\left(\frac{\overline{X} - \frac{1}{2}}{4}\right) = \frac{1}{16}\,\text{Var}(\overline{X}) = \frac{1}{16}\frac{\sigma_X^2}{n}$$

$$\lim_{n\to\infty} E(\tilde{\theta}) = \lim_{n\to\infty} \theta = \theta$$
$$\lim_{n\to\infty} \text{Var}(\tilde{\theta}) = \lim_{n\to\infty}\left(\frac{1}{16}\frac{\sigma_X^2}{n}\right) = \frac{\sigma_X^2}{16}\lim_{n\to\infty}\left(\frac{1}{n}\right) = 0$$

O estimador $\tilde{\theta}$ é consistênte.

(5) (d) Observou-se a seguinte amostra casual da variável aleatória $X : (0, 1, 2, 3, 1, 0)$. Obtenha uma estimativa para θ.

RESPOSTA:

$$\overline{x} = \frac{0+1+2+3+1+0}{6} = \frac{7}{6}$$

A estimativa é:

$$\tilde{\theta} = \frac{\overline{x} - \frac{1}{2}}{4} = \frac{\frac{7}{6} - \frac{1}{2}}{4} = \frac{1}{6} \qquad (2)$$

(5) (e) Obtenha a expressão da função de verosimilhança para a amostra apresentada na alínea anterior. Não estime o parâmetro.

RESPOSTA:

$$L(\theta) = \prod_{i=1}^{6} f_X(x_i|\theta) \ = f_X(0|\theta)\, f_X(1|\theta)\, f_X(2|\theta)\, f_X(3|\theta)\, f_X(1|\theta)\, f_X(0|\theta) =$$
$$= (0.5-\theta)^4\ \theta^2 \quad (0 < \theta < 0.5)$$

3. Uma universidade está a considerar implementar um novo método de ensino, na disciplina de Estatística. O objectivo é considerar se o novo método resulta num melhoramento significativo das avaliações dos estudantes. Para isso, são escolhidas duas amostras casuais de alunos, com os seguintes resultados do exame (avaliado de 0 a 200):

Amostra A – Avaliação dos estudantes com o método tradicional:
$n = 35, \bar{x} = 152, s^{'} = 55$.
Amostra B – Avaliação dos estudantes com o novo método:
$n = 43, \bar{x} = 163, s^{'} = 57$.

(20) (a) Teste a hipótese de que o novo método permite obter melhores resultados, em média, em relação ao método tradicional, para uma dimensão de 5%. Justifique a sua resposta com recurso à região de rejeição.

RESPOSTA: $\alpha = 5\%$ $H_0 : \mu_B - \mu_A \leq 0$ vs $H_1 : \mu_B - \mu_A > 0$

$$VF : T = \frac{\bar{X}_B - \bar{X}_A - (\mu_B - \mu_A)}{\left(\frac{S_B^2}{n_B} + \frac{S_A^2}{n_A}\right)^{1/2}} \overset{a}{\sim} N(0,1)$$

$RR = \{T : T > q_{5\%}\}$, com $q_{5\%} = 1,645$

$$T_{\text{obs}} = \frac{163 - 152 - 0}{\left(\frac{57^2}{43} + \frac{55^2}{35}\right)^{1/2}} \approx 0.8643 \to 0.86$$

Como $T_{\text{obs}} < q_{5\%}$, não rejeitamos H_0 para $\alpha = 5\%$

(10) (b) Para o mesmo teste da alinea anterior, calcule o valor-p. A conclusão mantém-se? Porquê?

RESPOSTA: valor-p $= P(N(0,1) > 0.86) \approx 1 - 08051 = 0.1949$ Como o valor-p $> 5\%$, não rejeitamos H_0 para $\alpha = 5\%$. A conclusão mantém-se pois valor-p $< \alpha \Leftrightarrow T_{\text{obs}} \in$ RR.

(40) **4.** Um pizzaiolo afirma que os seus clientes apreciam igualmente todas as pizzas vendidas no seu restaurante. Para que fosse possível testar esta teoria, registaram-se as quantidades vendidas dos vários tipos de pizza durante uma semana, tendo-se obtido os seguintes resultados:

Pizza	Margarita	Marinara	Diavola	Quatro Queijos	Cogumelos
Unidades vendidas	200	150	220	235	195

Com base num teste estatístico adequado e considerando uma dimensão de 0.05, concorda com a afirmação do pizzaiolo?

RESPOSTA: $n = 1000$; $j = 1, 2, 3, 4, 5$ (tipos de pizza) $H_0 : \forall_j P_{0j} = \frac{1}{5}$ vs $H_1 : \exists_j P_{0j} \neq \frac{1}{5}$

$$VF : Q = \sum_{j=1}^{5} \frac{(N_j - fe_j)^2}{fe_j} \overset{a}{\sim} \chi^2(4)$$

$RR = \{Q : Q > q_{5\%}\}$, com $q_{5\%} = 9,488$ $\forall_j fe_j = 200$, portanto,

$$Q_{\text{obs}} = \frac{(200-200)^2}{200} + \frac{(150-200)^2}{200} + \cdots + \frac{(195-200)^2}{200} \approx 20,75$$

Como $Q_{\text{obs}} \in \text{RR}$, rejeitamos H_0 para $\alpha = 5\%$.

5. Considere o seguinte modelo para estudar o nível de poupança dos cidadãos portugueses, no ano de 2023:

$$\log(poup_i) = \beta_0 + \beta_1 \log(sal_i) + \beta_2\, dep_i + \beta_3\, esc_i + u_i \tag{3}$$

onde *poup* corresponde ao montante poupado pelo indivíduo i, *sal* corresponde ao salário do indivíduo i, *dep* corresponde ao número de dependentes do indivíduo i e *esc* corresponde ao número de anos de escolaridade do indivíduo i. Notar que a função $\log(\cdot)$ corresponde ao logaritmo natural (base e). Assuma que o erro, u, tem média nula.

Com base numa amostra aleatória de dimensão $n = 101$ e com recurso ao método dos Mínimos Quadrados Ordinários (OLS), foram obtidas as seguintes estimativas (com os erros-padrão dentro de parêntesis).

$$\begin{aligned} \widehat{\log(poup_i)} &= \underset{(0.331)}{-3.115} + \underset{(0.176)}{1.721} \log(sal_i) - \underset{(0.057)}{0.101}\, dep_i - \underset{(0.026)}{0.006}\, esc_i \\ R^2 &= 0.569, \quad SSR = 41.787 \end{aligned} \tag{4}$$

(15) (a) Interprete as estimativas para $\hat{\beta}_1$ e para $\hat{\beta}_2$.

RESPOSTA:

- $\hat{\beta}_1$: estima-se que um aumento de 1% no salário provoque um aumento médio de 1.721% na poupança, ceteris paribus;
- $\hat{\beta}_2$: estima-se que um aumento de uma unidade no número de dependentes (ter mais um filho) provoque uma redução média de 10.1% na poupança, ceteris paribus.

(15) (b) Teste a significância individual das variaveis log(*sal*) e *dep*, ao nível de significância $\alpha = 0.01$. [Nota: se não encontrar o valor certo nas tabelas estatísticas, utilizar o valor mais próximo]

RESPOSTA:

- Hipóteses em teste:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases} , j = 1, 2$$

- Estatística de teste:

$$t_j = \frac{\hat{\beta}_j}{\text{se}\left(\hat{\beta}_j\right)} \sim t(n-k-1)$$

- Região de Rejeição:

$$W_j = \left\{ t_j : t_j < -t_{\alpha/2} \vee t_j > t_{\alpha/2} \right\}$$

com $t_{\alpha/2} : \mathrm{P}\left(t > t_{\alpha/2}\right) = \alpha/2$

- Neste caso teremos:

$$t_{1,obs} = \frac{1.721}{0.176} \approx 9.778 \qquad t_{2,obs} = \frac{-0.101}{0.057} \approx 1.772$$

Para $n = 101$, $k = 3$ e $\alpha = 1\%$ teremos $t_{0.005}(97) = 2.627$ (valor crítico exato) ou $2.626 < t_{0.005}(97) < 2.632$ (das tabelas estatísticas). Assim, como $t_{1,obs} \in W_1$ ou $|t_{1,obs}| > t_{0.005}$, rejeitamos $H_0 : \beta_1 = 0$, logo a variável log(*sal*) é estatísticamente significativa ao nível de 1%. Por outro lado, como $t_{2,obs} \notin W_2$ ou $|t_{2,obs}| < t_{0.005}$, não rejeitamos $H_0 : \beta_1 = 0$, logo a variável *dep* não é estatísticamente significativa ao nível de 1%.

(15) (c) Considere a hipótese $H_0 : \beta_3 = -\beta_2$ contra a alternativa bilateral. Como testaria esta hipótese, reparametrizando o modelo definido acima? Identifique a estatística de teste e a respetiva distribuição.

RESPOSTA:

- Pretende-se testar as hipóteses:

$$\begin{cases} H_0 : \beta_3 = -\beta_2 \\ H_1 : \beta_3 \neq -\beta_2 \end{cases} \Leftrightarrow \begin{cases} H_0 : \beta_3 + \beta_2 = 0 \\ H_1 : \beta_3 + \beta_2 \neq 0 \end{cases} \Leftrightarrow \begin{cases} H_0 : \theta = 0 \\ H_1 : \theta \neq 0 \end{cases}$$

  com $\theta = \beta_3 + \beta_2$.

- Estatística de teste:

$$t = \frac{\hat{\beta}_3 + \hat{\beta}_2}{\text{se}\left(\hat{\beta}_3 + \hat{\beta}_2\right)} = \frac{\hat{\theta}}{\text{se}\left(\hat{\theta}\right)} \sim t\left(n-k-1\right)$$

- Por forma a obter-se diretamente $\text{se}\left(\hat{\theta}\right)$, poderá recorrer-se à estimação de uma equação auxiliar de teste, que consiste numa reparametrização do modelo original. Assim, sabendo que $\theta = \beta_3 + \beta_2 \Leftrightarrow \beta_3 = \theta - \beta_2$, e impondo esta restrição no modelo original, teriamos:

$$\begin{aligned} \log\left(poup_i\right) &= \beta_0 + \beta_1 \log\left(sal_i\right) + \beta_2\, dep_i + \beta_3\, esc_i + u_i \\ &= \beta_0 + \beta_1 \log\left(sal_i\right) + \beta_2\, dep_i + \left(\theta - \beta_2\right)\, esc_i + u_i \\ &= \beta_0 + \beta_1 \log\left(sal_i\right) + \beta_2\, dep_i + \theta\, esc_i - \beta_2\, esc_i + u_i \\ &= \beta_0 + \beta_1 \log\left(sal_i\right) + \beta_2\left(dep_i - esc_i\right) + \theta\, esc_i + u_i \end{aligned}$$

  A estimação OLS da regressão de $\log\left(poup\right)$ sobre $\log\left(sal\right)$, $dep - esc$ (o regressor transformado) e esc, permitiria obter diretamente a estimativa para $\hat{\theta}$ e o respetivo $\text{se}\left(\hat{\theta}\right)$.

(15) (d) Foi adicionado o termo quadrático, esc^2, ao modelo da equação 4. A nova equação estimada é a seguinte:

$$\begin{aligned} \widehat{\log\left(poup_i\right)} &= \underset{(0.975)}{-4.020} + \underset{(0.181)}{1.764} \log\left(sal_i\right) - \underset{(0.057)}{0.100} dep_i - \underset{(0.145)}{0.135} esc_i + \underset{(0.0057)}{0.0056} esc_i^2 \\ R^2 &= 0.573, \quad SSR = 41.368 \end{aligned} \tag{5}$$

Deduza o efeito parcial da escolaridade sobre a log(poupança). A partir de que nível de escolaridade muda o sinal deste efeito?

RESPOSTA:

- Efeito parcial:

$$\frac{\partial \log\left(poup\right)}{\partial\, esc} = \beta_3 + 2\beta_4 esc_i \Longrightarrow \frac{\partial\, \widehat{\log\left(poup\right)}}{\partial\, esc} = -0.135 + 0.0112\, esc_i$$

- Ponto de viragem do efeito parcial:

$$\begin{aligned} &-0.135 + 0.0112\, esc^* = 0 \Leftrightarrow 0.0112\, esc^* = 0.135 \\ &\Leftrightarrow esc^* = \frac{0.135}{0.0112} \Leftrightarrow esc^* \approx 12.054 \text{ anos} \end{aligned}$$

(15) (e) Explique, de forma sucinta, como procederia para testar, no modelo da equação 5, a hipótese de que a escolaridade não tem qualquer efeito sobre a poupança. Apresente as hipóteses em teste, a estatística de teste e a respetiva distribuição. Obtenha ainda, se possível, o valor crítico associado a este teste.

RESPOSTA:

- Hipóteses em teste:

$$\begin{cases} H_0 : \beta_3 = \beta_4 = 0 \\ H_1 : \exists_j^1 \beta_j \neq 0,\ j = 3,4 \end{cases}$$

- Estatística de teste:

$$F = \frac{\mathrm{SSR_R} - \mathrm{SSR_{UR}}}{\mathrm{SSR_{UR}}} \times \frac{n-k-1}{q} \sim F\left(q, n-k-1\right)$$

OU

$$F = \frac{R^2_{\mathrm{UR}} - R^2_{\mathrm{R}}}{1 - R^2_{\mathrm{UR}}} \times \frac{n-k-1}{q} \sim F\left(q, n-k-1\right)$$

- Região de Rejeição:

$$W = \{f : f > f_\alpha\}$$

com $f_\alpha : \mathrm{P}\left(F > f_\alpha\right) = \alpha$ Para $n = 101$, $k = 4$, $q = 2$ e $\alpha = 5\%$ teremos $f_{0.05}\left(2, 96\right) = 3.09$ (valor crítico exato) ou $3.07 < f_{0.05}\left(2, 96\right) < 3.15$ (das tabelas estatísticas). Se $f_{obs} \in W$ ou $f_{obs} > f_{0.05}$, rejeitamos H_0, logo existiria evidência estatística que a variável escolaridade tem efeito sobre a poupança, ao nível de 5%, caso contrario, não rejeitamos H_0.

Fim