ISEG/UL — ESTATÍSTICA II — 2.º Ano/GES – 2.º Semestre 2024/25

Exame ER — 3 de Junho de 2025 — Duração: 2 horas

Nome: ______________________________ **Número:** __________

Cotação:

Pergunta	1	2	3	4a	4b	5a	5b	6a	6b	7a	7b	8a	8b	9a	9b
Cotação Total	1.5	1.5	2	1	1.5	1	1.5	1	1.5	1	1.5	1	1.5	1	1.5
Cotação Atribuída															

Atenção:

☐ Nas questões de resposta aberta, justifique detalhadamente as respostas.

---

1. Classifique as seguintes afirmações como verdadeiras (V) ou falsas (F).

	V	F
(a) Sejam X_i, $i = 1, \cdots, 4$, variáveis aleatórias independentes e identicamente distribuídas retiradas de uma população com média μ. Então, o estimador $T_1 = \frac{1}{10}X_1 + \frac{4}{10}X_2 + \frac{3}{10}X_3 + \frac{2}{10}X_4$ é enviesado.	☐	☑
(b) Sejam X_i, $i = 1, \cdots, 4$, variáveis aleatórias independentes e identicamente distribuídas retiradas de uma população com média μ e variância σ^2. Então, o estimador $T_1 = \frac{1}{10}X_1 + \frac{4}{10}X_2 + \frac{3}{10}X_3 + \frac{2}{10}X_4$ é mais eficiente que o estimador $T_2 = \frac{2}{10}X_1 + \frac{3}{10}X_2 + \frac{3}{10}X_3 + \frac{2}{10}X_4$.	☐	☑
(c) Num teste de hipótese simples contra hipótese simples tem-se: $P(\text{Rejeitar } H_0 \mid H_0) + P(\text{Não rejeitar } H_0 \mid H_1) = 1$.	☐	☑
(d) Num modelo de regressão $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}^2 + \beta_4 x_{t4} + u$, com $\beta_3 > 0$, temos que Y primeiro diminui e depois aumenta com x_3.	☑	☐
(e) Considere o teste estatístico $H_0 : p = 0.64$ vs $H_1 : p \neq 0.64$, onde p representa uma proporção desconhecida. Assumindo que o teste é realizado com uma amostra de dimensão 500 e que $z_{obs} < -z_{0.025}$, devemos rejeitar a hipótese nula ao nível de significância de 5%.	☑	☐
(f) O modelo de regressão $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4^2 x_{t4} + u_t$, enquadra-se nos modelos lineares ou linearizáveis estudados no curso.	☐	☑

VSFF

2. Escolha a opção correta em cada uma das seguintes questões de escolha múltipla. Cada resposta correta vale 0.75.

a) Considere uma amostra aleatória de 100 observações de uma variável aleatória X com média μ e variância σ^2. Qual das seguintes afirmações é **verdadeira**?

- ☐ O estimador $\bar{X}$ é sempre um estimador enviesado para μ.
- ☐ A variância de $\bar{X}$ é igual a σ^2.
- ☐ A distribuição de $\bar{X}$ é sempre normal, independentemente da distribuição de X.
- ☑ Pelo Teorema Central do Limite, a distribuição de $\bar{X}$ aproxima-se de uma distribuição normal quando n é grande.

b) Considere o modelo de regressão linear,

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i,$$

a estimar com base numa amostra de dimensão n. Pretende-se testar se a variável x_4 tem um impacto positivo no valor esperado da variável y, tudo o resto constante.

Como deveriam ser definidas as hipóteses para este teste?

- ☑ $H_0 : \beta_4 \leq 0 \quad \text{contra} \quad H_1 : \beta_4 > 0$
- ☐ $H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{contra} \quad H_1 : \exists\, \beta_j \neq 0,\ j = 2, 3, 4$
- ☐ $H_0 : \beta_4 = 0 \quad \text{contra} \quad H_1 : \beta_4 \neq 0$
- ☐ $H_0 : \beta_4 \geq 0 \quad \text{contra} \quad H_1 : \beta_4 < 0$

3. Considere a regressão linear múltipla dada por

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i, \quad k > 2$$

e assuma que existem n observações para estimar a regressão. Mostre que o coeficiente de determinação ajustado verifica $\bar{R}^2 \leq R^2$ e que a estatística do teste de nulidade conjunta pode ser escrita como

$$F = \frac{1 + \frac{n-k}{k-1}\bar{R}^2}{1 - \bar{R}^2}.$$

O coeficiente de determinação, R^2 e o coeficiente de determinação ajustado, $\bar{R}^2$, estão ambos definidos no formulário.

RESPOSTA 3

O coeficiente de determinação R^2 é dado por

$$R^2 = 1 - \frac{VR}{VT}.$$

Uma vez que

$$\bar{R}^2 = 1 - \frac{VR/(n-k)}{VT/(n-1)} = 1 - \frac{VR}{VT} \times \frac{n-1}{n-k} \leq 1 - \frac{VR}{VT} = R^2$$

Sabendo que

$$\bar{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-k}$$

podemos resolver a equação em ordem a R^2, obtendo:

$$R^2 = \frac{(n-1) - (n-k)(1-\bar{R}^2)}{n-1} \Rightarrow R^2 = \frac{(n-k)\bar{R}^2 + k - 1}{n-1}$$

A estatística F é dada por:

$$F = \frac{R^2/k-1}{(1-R^2)/(n-k)}$$

Temos

$$F = \frac{n-k}{k-1} \times \frac{\frac{(n-k)\bar{R}^2+k-1}{n-1}}{1 - \frac{(n-k)\bar{R}^2+k-1}{n-1}} = \frac{n-k}{k-1} \times \frac{(n-k)\bar{R}^2 + k - 1}{n-k-(n-k)\bar{R}^2} = \frac{\frac{n-k}{k-1}\bar{R}^2 + 1}{1 - \bar{R}^2}.$$

4. Uma empresa multinacional decidiu conduzir uma auditoria interna aos pagamentos em falta aos fornecedores de dois departamentos distintos da empresa. Dada a dimensão da empresa recolheram-se duas amostras aleatórias independentes dos pagamentos a efetuar no próximo mês. As **estatísticas amostrais** obtidas relativas ao valor dos pagamentos devidos no próximo mês foram as seguintes:

Departamento	**Média**	**Desvio Padrão Corrigido**	**Tamanho da Amostra**
X	290	16	15
Y	250	11	47

Assuma que os pagamentos a efetuar podem ser modelados por uma distribuição normal e que as variâncias populacionais são desconhecidas mas iguais.

a) Os auditores pretendem realizar o seguinte teste de hipóteses $H_0 : \mu_X = 275$ vs $H_1 : \mu_X \neq 275$. Qual das seguintes hipóteses é **verdadeira**?

☐ **i)** Estatística de teste segue uma $N(0, 1)$ ☐ **ii)** $t_{obs} = 2.145$

☑ **iii)** Estatística de teste segue uma $t_{(14)}$ ☐ **iv)** Estatística de teste segue uma $t_{(15)}$

b) Determine o intervalo de confiança de 95% para $\mu_X - \mu_Y$.

RESPOSTA 4.b) A variável fulcral neste caso é dada por

$$T = \frac{\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}}{\sqrt{\frac{(n_X-1){S'_X}^2+(n_Y-1){S'_Y}^2}{n_X+n_Y-2}}} \sim t_{\underbrace{(n_X + n_Y - 2)}_{60}},$$

onde $n_X = 15$ e $n_Y = 47$. Se

$$SE = \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \times \sqrt{\frac{(n_x - 1){S'_X}^2 + (n_Y - 1){S'_Y}^2}{n_X + n_Y - 2}}$$

o desvio padrão observado associado à diferença das médias é dado por:

$$se = \sqrt{\frac{1}{15} + \frac{1}{47}} \times \sqrt{\frac{14 \times 16^2 + 46 \times 11^2}{15 + 47 - 2}} \approx 3.6622$$

O quantil da distribuição para o intervalo de confiança em causa é $t_{0.025,60} = 2.000$. Consequentemente, o intervalo de confiança aleatório é

$$IC = [\bar{X} - \bar{Y} - t_{0.025,60}SE,\ \bar{X} - \bar{Y} + t_{0.025,60}SE]$$

Concretizando o intervalo aleatório obtemos

$$[(290 - 250) - 2 * 3.6622,\ (290 - 250) + 2 * 3.6622] = [32.68,\ 47.32].$$

5. Um fabricante de produtos eletrónicos quer estimar a variância da duração (em horas) de uma nova linha de baterias. Uma amostra aleatória de 20 baterias forneceu um desvio padrão amostral de $s = 1.3$ horas. Assuma que a duração em horas das baterias segue uma distribuição normal com variância e média desconhecidos.

a) Qual das seguintes afirmações é **verdadeira**?

☐ **i)** S^2 é centrado para σ^2

☑ **ii)** $\frac{\bar{X}-\mu}{S'/\sqrt{20}}$ é uma variável fulcral para μ

☐ **iii)** $\frac{\bar{X}-\mu}{\sigma/\sqrt{20}}$ é uma variável fulcral para μ

☐ **iv)** Nenhuma das respostas anteriores.

b) Construa um intervalo de confiança de 95% para a variância da população.

RESPOSTA 5.b)

Queremos um intervalo de confiança para σ^2, que é baseado na estatística qui-quadrado:

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

Para $n = 20$, temos uma distribuição do qui-quadrado com $n - 1 = 19$ graus de liberdade. Os quantis para a construção do intervalo de confiança são dados por:

$$\chi^2_{0.975} = 8.907 \quad \text{and} \quad \chi^2_{0.025} = 32.852.$$

O intervalo de confiança para σ^2 a 95% é dado por:

$$\left[\frac{ns^2}{\chi^2_{0.025}}, \frac{ns^2}{\chi^2_{0.975}}\right] = \left[\frac{20 \times 1.3^2}{32.852}, \frac{20 \times 1.3^2}{8.907}\right]$$

Calculando $ns^2 = 20 \times 1.69 = 33.8$, podemos verificar que

$$\left[\frac{33.8}{32.852}, \frac{33.8}{8.907}\right] \approx [1.029,\ 3.795]$$

Portanto, com 95% de confiança, a variância populacional da duração das baterias está entre aproximadamente 1.029 e 3.795 horas^2.

VSFF

6. Um investigador está a estudar os hábitos de leitura dos alunos do ISEG durante o ano letivo. Após uma amostragem aleatória de 61 estudantes, constatou-se que,

$$\frac{1}{61}\sum_{i=1}^{61} x_i = 12 \quad \text{e} \quad \frac{1}{60}\sum_{i=1}^{61}(x_i - \overline{x})^2 = 12.203,$$

onde x_i representa o número de livros lido pelo estudante i, com $i = 1, 2, \cdots, 61$.

a) Considere o teste de hipóteses $H_0 : \mu = 13$ vs $H_1 : \mu < 13$. Qual das seguintes afirmações é **verdadeira**?

- ☐ Devemos rejeitar a hipótese nula.
- ☐ O teste de hipóteses é um teste bilateral.
- ☑ A região crítica, ao nível de significância de 1%, é dada por $\{z : z < -2.326\}$.
- ☐ O valor da estatística observada para o teste de hipóteses é $z_{obs} = 2.2358$.

b) Assuma que o tempo que um aluno demora a ler um livro é modelado por uma distribuição exponencial com média λ. Dos resultados recolhidos percebeu-se que

$$\sum_{i=1}^{61} y_i = 65 \quad \text{and} \quad \sum_{i=1}^{61} y_i^2 = 120$$

onde y_i representa o tempo que o aluno i demora a ler um livro. Calcule a estimativa de maxima verosimilhança para λ.

RESPOSTA 6.b)

Seja Y uma distribuição exponencial com parâmetro λ, i.e.,

$$Y \sim \text{Exp}(\lambda), \quad \text{com função densidade:} \quad f_Y(y;\lambda) = \lambda e^{-\lambda y}, \quad y > 0.$$

Para uma amostra aleatória de 61 alunos, a função de verosimilhança é:

$$L(\lambda;\, y_1, y_2, \ldots, y_{61}) = \prod_{i=1}^{61} \lambda e^{-\lambda y_i} = \lambda^{61} e^{-\lambda \sum_{i=1}^{61} y_i}.$$

A função log-verosimilhança é dado por: $\ell(\lambda;\, y_1, y_2, \ldots, y_{61}) = \log L(\lambda;\, y_1, y_2, \ldots, y_{61}) = 61 \log \lambda - \lambda \sum_{i=1}^{61} y_i$. Derivamos a log-verosimilhança em relação a λ e igualamos a zero:

$$\frac{d\ell}{d\lambda}(\lambda;\, y_1, y_2, \ldots, y_{61}) = \frac{61}{\lambda} - \sum_{i=1}^{61} y_i = 0 \quad \Rightarrow$$

Tendo em conta que $\frac{d^2\ell}{d\lambda^2}(\lambda;\, y_1, y_2, \ldots, y_{61}) = -\frac{61}{\lambda^2} < 0$. Consequentemente,

$$\hat{\lambda}_{MV} = \frac{61}{\sum_{i=1}^{61} Y_i}.$$

A estimativa de máxima verosimilhança é portanto dada por:

$$\hat{\lambda} = \frac{61}{65} = \frac{12}{13} \approx 0.9385.$$

VSFF

7. Um grupo de investigadores está a estudar a presença de mensagens ecológicas em campanhas publicitárias de duas editoras diferentes, uma no Brasil e outra em Portugal. Para isso, foi analisado o conteúdo de anúncios publicitários produzidos por ambas as editoras. Numa amostra aleatória de 270 anúncios publicitários em circulação no Brasil, 56 continham mensagens com preocupação ecológica. Por sua vez, numa amostra aleatória de 203 anúncios de uma editora portuguesa, 52 apresentavam esse tipo de mensagem. Na resolução do exercícios assuma que os dois processos de amostragem são independentes. Represente por X_1 (resp., X_2) a variável aleatória de Bernoulli que vale 1 se um anúncio da agência brasileira (resp., portuguesa) tem uma mensagem ecológica e por p_1 (resp., p_2) a verdadeira proporção de anúncios com uma mensagem ecológica no Brasil (resp. em Portugal).

a) Qual das afirmações relativas aos estimadores e estimativas pontuais para p_1 e p_2 é **falsa**?

- ☑ **i)** $\bar{X}_1$ não é o estimador dos momentos para p_1
- ☐ **ii)** $\bar{X}_1$ é o estimador de máxima verosimilhança para p_1
- ☐ **iii)** $\bar{X}_2$ é o estimador de máxima verosimilhança para p_2
- ☐ **iv)** A estimativa pontual para p_1 dada pelo método dos momentos com a amostra recolhida é de aproximadamente 0.2074.

b) Efectue um teste de hipóteses para avaliar se existe diferença entre as proporções de anúncios com mensagens ecológicas entre os dois países.

RESPOSTA 7.b)

Temos

$$X_1 \sim B(1, p_1) \quad \text{e} \quad X_2 \sim B(1, p_2)$$

Queremos testar:

$$H_0 : p_1 = p_2 \quad \text{vs} \quad H_1 : p_1 \neq p_2$$

Dado que a dimensão da amostra é suficientemente grande, a estatística de teste é dada por

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{a}{\sim} N(0,1), \qquad \text{onde } \hat{p} = \frac{270\,\overline{X}_1 + 203\,\overline{X}_2}{270 + 203}.$$

Dado que $\overline{x}_1 = \frac{56}{270} \approx 0.2074$ e que $\overline{x}_2 = \frac{52}{203} \approx 0.2562$, podemos verificar que

$$\frac{270\,\overline{x}_1 + 203\,\overline{x}_2}{270 + 203} = \frac{56 + 52}{270 + 203} \approx 0.2283 \quad \text{and} \quad \sqrt{0.2283 \cdot (1 - 0.2283)\left(\frac{1}{270} + \frac{1}{203}\right)} \approx 0.03899.$$

Logo, o valor observado de z é:

$$z_{obs} = \frac{0.2074 - 0.2562}{0.03899} \approx -1.25$$

O p-value é dado por $P(|Z| > 1.25) = 2(1 - P(Z < 1.25)) = 2 * (1 - 0.8944) = 0.2112$. Assim, não rejeitamos H_0 a nenhum nível de significância. Não havendo evidência estatística suficiente para concluir que as proporções são diferentes.

São admitidas respostas em que $Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\overline{X}_1(1-\overline{X}_1)}{270} + \frac{\overline{X}_2(1-\overline{X}_2)}{203}}} \overset{a}{\sim} N(0,1)$. Nesse caso $z_{obs} = -1.24$ e a conclusão é a mesma.

VSFF

8. Um economista aeroespacial está a estudar os fatores que determinam o custo total por astronauta em missões espaciais tripuladas. Utilizando dados de 65 missões diferentes, estimou-se a seguinte regressão:

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} - \beta_4 x_{t4} + \beta_5 x_{t5} + u_t$$

onde y representa o custo total por astronauta (em milhares de dólares), x_2 a duração da missão (em dias), x_3 a percentagem da tripulação com formação militar, x_4 a percentagem de tempo passado fora da estação e x_5 a variável artificial que assume o valor 1 se a missão foi liderada por uma agência privada, 0 se liderada por uma agência pública. A regressão estimada e os respetivos erros padrão (entre parêntesis) encontram-se abaixo:

$$\begin{array}{l} \hat{y}_t = 8740 + \underset{(110)}{437 x_{t2}} + \underset{(58)}{115 x_{t3}} - \underset{(107)}{318 x_{t4}} + \underset{(391)}{734 x_{t5}} \end{array}$$

O coeficiente de determinação é $R^2 = 0.47$.

a) Qual das seguintes afirmações é **verdadeira**?

☐ A variável x_2 não é estatisticamente significativa ao nível de significância de 5%.
☐ A variável x_3 é estatisticamente significativa ao nível de significância de 1%.
☑ A variável x_4 é estatisticamente significativa ao nível de significância de 5%.
☐ A variável x_5 é estatisticamente significativa ao nível de significância de 1%.

b) Teste a significância global da regressão.

RESPOSTA 8.b)

Queremos testar se a regressão é globalmente significativa, ou seja:

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \quad \text{vs} \quad H_1 : \text{Pelo menos um } \beta_j \neq 0$$

Usamos o teste F, com a estatística:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F(4, 60)$$

A estatística observada é dada por

$$f_{obs} = \frac{0.47/4}{(1-0.47)/60} = \frac{0.1175}{0.008833} \approx 13.30$$

O quantil da distribuição F que verifica $P(F > f_{4,60,0.05}) = 0.05$ é dado por $f_{4,60,0.05} = 2.53$. Assim a região crítica é dada por $W = \{f : f > 2.53\}$. Como $13.30 > 2.53$, temos que $f_{obs} \in W$ e consequentemente rejeitamos H_0. Concluímos que a regressão é globalmente significativa ao nível de 5%.

VSFF

9. Um economista da saúde pretende investigar se os determinantes da despesa pública em saúde por habitante diferem consoante a orientação política dos governos regionais. A seguinte regressão foi estimada com dados de 65 regiões administrativas:

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + \beta_6 x_{t4} x_{t5} + u_t,$$

onde y representa a despesa pública em saúde por habitante (em milhares de euros), x_2 o rendimento médio per capita (em milhares de euros), x_3 a percentagem da população com grau universitário, x_4 a percentagem da população em zonas rurais e x_5 é uma variável binária que assume valor 1 se o governo regional é de esquerda, e 0 se é de direita. Para testar se os coeficientes da regressão diferem entre regiões governadas pela esquerda e pela direita, foram estimadas separadamente diferentes regressões:

- Soma dos quadrados dos resíduos para regiões de direita $\left(n_1 = 27 : \sum \hat{u}_i^2 = 13520\right)$
- Soma dos quadrados dos resíduos para regiões de esquerda $\left(n_2 = 25 : \sum \hat{u}_i^2 = 11960\right)$
- Soma dos quadrados dos resíduos para todas as regiões $\left(n = 52 : \sum \hat{u}_i^2 = 26120\right)$

$$R^2 = 0.7991 \quad \text{e} \quad \sum_{t=1}^{643} (y_t - \bar{y})^2 = 66.18.$$

a) Qual das seguintes regressões representa a regressão a ser estimada para as regiões administrativas governadas pela esquerda?

☐ $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} - (\beta_4 + \beta_6) x_{t4} + u_t$
☐ $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} - \beta_4 x_{t4} + u_t$
☑ $y_t = (\beta_1 + \beta_5) + \beta_2 x_{t2} + \beta_3 x_{t3} + (\beta_4 + \beta_6) x_{t4} + u_t,$
☐ $y_t = (\beta_1 + \beta_5) + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + u_t,$

b) Teste se há evidência de mudança de estrutura entre os dois grupo de regiões. Considere um nível de significância de 5%.

RESPOSTA 9.b)

H_0 : Os coeficientes são iguais para regiões de esquerda e direita vs H_1 : Pelo menos um coeficiente difere.

Estatística do teste de Chow:

$$F = \frac{(VR_0 - VR_1)/k}{VR_1/(n-2k)} \sim F(6, 40)$$

A variação residual quando H_0 não se verifica depois das regressões estimadas é de $13520 + 11960 = 25480$. Logo, a estatística observada neste caso é dada por

$$f_{obs} = \frac{(26120 - 25480)/6}{25480/40} \approx 0.1674$$

O quantil da distribuição F que verifica $P(F > f_{6,40,0.05}) = 0.05$ é $f_{6,40,0.05} =\approx 2.34$. Como $f_{obs} = 0.1674 < 2.34$, não rejeitamos H_0. Logo, não parece haver evidência de mudança de estrutura entre regiões de esquerda e de direita ao nível de significância de 5%.

FIM