ESTATÍSTICA II – Lic. Economia e Finanças
13 de Janeiro de 2017 – Duração: 1h

TESTE I

Nome: _____ N.º: _____

Espaço reservado a classificações							
1.a)	1.b)	2.a)	2.b)	2.c)	3.)	4.)	
						Total	

Perguntas de escolha múltipla: cada resposta certa vale 10 pontos; cada resposta errada vale -2.5 pontos; assinale a resposta escolhida com uma cruz no quadrado adequado.
As cotações das restantes perguntas são indicadas no enunciado.

1. Seja θ a proporção de alunos do 2º ano que concordam com o novo método de avaliação adoptado em certas disciplinas de determinada Universidade. O Conselho Pedagógico afirma que pelo menos 90% dos estudantes estão satisfeitos com o novo método. Para testar a veracidade desta afirmação, foram seleccionados ao acaso 200 alunos do 2º ano tendo 175 concordado com o novo método.

a) **[15]** Teste, ao nível de 5%, a veracidade da afirmação.

RESPOSTA:

Seja X uma v.a. tal que $X = 1$ se o aluno concorda com o novo método de avaliação, e $X = 0$ caso contrário. Portanto $X \sim Bernoulli(\theta)$, onde θ representa a proporção, na população, de alunos que concordam. A hipótese a testar será então:

$$H_0 : \theta \geq 0.9 \quad \text{contra} \quad H_1 : \theta < 0.9$$

Sob H_0,

$$Z = \frac{\bar{X} - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \overset{a}{\sim} N(0,1)$$

e a região crítica será: $W_{5\%} = \{z : z < -1.645\}$. Com $n = 200$ e $\sum x_i = 175$ tem-se $\bar{x} = 175/200 = 0.875$, portanto o valor observado é:

$$z_{\text{obs}} = \frac{0.875 - 0.9}{\sqrt{\frac{0.9(1-0.9)}{200}}} = -1.1785$$

Como $z_{\text{obs}} \notin W_{5\%}$, não se rejeita H_0 a 5%: é de adimitir a veracidade da afirmação.

b) **[20]** No semestre seguinte, numa amostra casual de 250 alunos, 235 alunos concordaram com o novo método de avaliação. Teste, ao nível de 5% se existem diferenças significativas, entre os dois semestres, na proporção de alunos favoráveis ao novo método de avaliação.

RESPOSTA:

Duas populações de Bernoulli: $X_i \sim \text{Bern}(\theta_i)$, com $X_i = 1$ se um aluno do semestre i concorda com o novo método de avaliação, $i = 1, 2$. A hipótese a testar será $H_0 : \theta_1 = \theta_2$ contra $H_1 : \theta_1 \neq \theta_2$. Sob H_0,

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\hat{\theta}(1-\hat{\theta})}} \overset{a}{\sim} \text{N}(0,1) \ , \quad \text{onde} \quad \hat{\theta} = \frac{m\bar{X}_1 + n\bar{X}_2}{m+n}$$

Das duas amostra tem-se: $\bar{x}_1 = 175/200 = 0.875$, com $m = 200$, e $\bar{x}_2 = 235/250 = 0.94$, com $n = 250$. Segue que $\hat{\theta} = (200 \times 0.875 + 250 \times 0.94)/450 = (175 + 235)/450 = 0.911$, portanto o valor observado da estatística-teste é dado por:

$$z_{\text{obs}} = \frac{0.875 - 0.94}{\sqrt{\left(\frac{1}{200} + \frac{1}{250}\right) \times 0.911 \times 0.089}} = -2.4075$$

A região crítica é $W_{5\%} = \{z : |z| > 1.96\}$, portanto $z_{\text{obs}} \in W_{5\%} \Rightarrow$ Rejeita-se H_0 ao nível de 5%. Há evidência de proporções diferentes ente os dois semestres.

2. Admita que a quantidade de sumo de laranja, em litros, por pacote da nova marca *Portukal* é uma variável aleatória com distribuição normal de média μ e variância σ^2. Com o objectivo de testar a afirmação "a quantidade média de sumo da nova marca é igual a 1", foi recolhida uma amostra aleatória de 25 pacotes, tendo-se obtido uma média de 0.98 litros e uma variância corrigida de 0.16.

a) **[15]** Construa um intervalo de confiança a 95% para a quantidade média de sumo por pacote. É de admitir, ao nível de 5%, que a afirmação está correcta? Justifique a sua resposta com base no intervalo obtido.

RESPOSTA:

Seja X = quantidade de sumo de laranja (litros). Então $X \sim N(\mu, \sigma^2)$. A hipótese a testar é $H_0 : \mu = 1$ contra $H_1 : \mu \neq 1$.

Variável fulcral: $Z = \frac{\bar{X} - \mu}{s'/\sqrt{n}} \sim t(n-1) = t(24)$

$I.C. = \left(\bar{x} \pm t_{\alpha/2} \times \frac{s'}{\sqrt{n}}\right)$, onde $\bar{x} = 0.98$, $s' = \sqrt{0.16} = 0.4$ e $t_{\alpha/2} = t_{0.025} = 2.064$.

$\Rightarrow I.C. = \left(0.98 \pm 2.064 \times \frac{0.4}{\sqrt{25}}\right) = (0.815; 1.145)$

Com uma confiança de 95%, pode afirmar-se que a quantidade média de sumo de laranja da nova marca se situa entre 0.815 e 1.145 litros.

Utilizando esse IC, como o valor nulo $\mu = 1$ pertence ao intervalo, não se rejeita a hipótese H_0 a 5%: é de admitir que a afirmação está correta.

b) Admita que foi construído um segundo intervalo de confiança a 95% para a quantidade média de sumo por pacote com base numa amostra casual da mesma população de dimensão superior a 25. Então:

a amplitude do segundo intervalo é menor	X
a amplitude do segundo intervalo é maior	
os dois intervalos têm a mesma amplitude	
nada se pode concluir porque a amplitude não depende da dimensão da amostra	

c) A estimativa da máxima verosimilhança para a proporção de pacotes com menos de 1 litro de sumo é:

$\Phi\left(\frac{1-\mu}{\sigma/\sqrt{n}}\right)$	
$\Phi\left(\frac{1-\bar{x}}{s}\right)$	X

$\Phi\left(\frac{1-\bar{x}}{s/\sqrt{n}}\right)$	
$\Phi\left(\frac{\bar{x}-\mu}{\sigma}\right)$	

3. Seja $(X_1, X_2, \ldots, X_n)$ uma amostra casual simples retirada de uma população X cuja distribuição depende de um parâmetro θ. Admita que T_1 e T_2 são estimadores centrados para o parâmetro θ, com $\theta > 0$. Então, pode-se concluir que:

T_1 e T_2 são consistentes	
$E(T_1) = E(T_2) = 0$	
$Var(T_1) = Var(T_2)$	
se T_1 é mais eficiente que T_2, então $EQM(T_1) \leq EQM(T_2)$	X

4. [20] Seja $(X_1, X_2, \ldots, X_n)$ uma amostra casual simples retirada de uma população X com média $E(X) = 2\theta$, variância $Var(X) = 2\theta^2$ e função de densidade:

$$f(x;\theta) = \frac{x}{\theta^2}\, exp\left(-\frac{x}{\theta}\right), \qquad x > 0, \quad \theta > 0.$$

Mostre que os estimadores da máxima verosimilhança e do método dos momentos para o parâmetro θ são iguais, e estude a sua consistência.

RESPOSTA:

Estimador pelo método dos momentos:

$$\tilde{\theta} : E(X) = \bar{X} \Longleftrightarrow 2\theta = \bar{X} \Longleftrightarrow \tilde{\theta} = \frac{\bar{X}}{2}$$

Função de verosimilhança $L(\theta)$:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{x_i}{\theta^2}\, exp\left(-\frac{x_i}{\theta}\right) = \frac{1}{\theta^{2n}} \prod_{i=1}^{n} x_i \, \exp\left(-\frac{\sum x_i}{\theta}\right)$$

$$l(\theta) = \ln L(\theta) = -2n \ln \theta + \sum \ln x_i - \frac{\sum x_i}{\theta}$$

Estando satisfeitas as usuais condições de regularidade, podemos obter o máximo da função através do cálculo das duas primeiras derivadas:

$$\frac{d\, l(\theta)}{d\theta} = 0 \Leftrightarrow -\frac{2n}{\theta} + \frac{\sum x_i}{\theta^2} = 0 \Leftrightarrow \frac{\sum x_i}{\theta} = 2n \Leftrightarrow \hat{\theta} = \frac{\sum x_i}{2n} = \frac{\bar{X}}{2}$$

$$\frac{d^2\, l(\theta)}{d\theta^2} = \frac{2n}{\theta^2} - \frac{2\sum x_i}{\theta^3}; \quad \text{no entorno de } \theta = \hat{\theta}: \quad \left.\frac{d^2\, l(\theta)}{d\theta^2}\right|_{\theta=\hat{\theta}} = -\frac{8n}{\bar{x}^2} < 0$$

Segue que o estimador de máxima verosimilhança para θ é dado por: $\hat{\theta} = \frac{\bar{X}}{2} = \tilde{\theta}$. Tem-se que:

- $E(\tilde{\theta}) = E\left(\frac{\bar{X}}{2}\right) = \frac{1}{2}E(\bar{X}) = \frac{1}{2}E(X) = \frac{1}{2} \times 2\theta = \theta$
- $\mathrm{Var}(\tilde{\theta}) = \mathrm{Var}\left(\frac{\bar{X}}{2}\right) = \frac{1}{4}\mathrm{Var}(\bar{X}) = \frac{1}{4}\frac{\mathrm{Var}(X)}{n} = \frac{1}{4n} \times 2\theta^2 = \frac{\theta^2}{2n}$

Condições suficientes para a consistência do estimador:

1. $\lim_{n\to+\infty} E(\tilde{\theta}) = \lim_{n\to+\infty} \theta = \theta$

2. $\lim_{n\to+\infty} \mathrm{Var}(\tilde{\theta}) = \lim_{n\to+\infty} \frac{\theta^2}{2n} = 0$

Conclusão: o estimador $\tilde{\theta}$ é consistente para θ.

ESTATÍSTICA II – Lic. Economia e Finanças
13 de Janeiro de 2017 – Duração: 1h

TESTE II

Nome: _____ N.º: _____

Espaço reservado a classificações							
5.)	6.a)	6.b)	6.c)	6.d)	6.e)	6.f)	
						Total	

Perguntas de escolha múltipla: cada resposta certa vale 10 pontos; cada resposta errada vale -2.5 pontos; assinale a resposta escolhida com uma cruz no quadrado adequado.
As cotações das restantes perguntas são indicadas no enunciado.

5. Uma imobiliária com agências em todo o país recolheu dados sobre os imóveis vendidos no último ano, com o objetivo de estudar as diferenças entre o preço inicialmente publicitado e o preço final de venda. De uma amostra casual de 500 imóveis, recolheu-se a seguinte informação:

Baixa de preço (Milhares de €)

[0, 10)	**[10, 20)**	**[20, 30)**	**≥30**	**Total**
a	189	139	b	500

onde a e b são dois números inteiros positivos.

Querendo testar se as baixas de preço seguem uma distribuição Exponencial de média 25 mil Euros, podemos afirmar que:

a estatística de teste tem distribuição $\chi^2(2)$ sob H_0	
as frequências esperadas sob H_0 dependem de a e b	
rejeita-se a hipótese sobre a distribuição se a estatística de teste apresentar valores superiores a 7.815	X
não é possível determinar o número de graus de liberdade da estatística de teste sem conhecer a e b, pois se um ou outro forem inferiores a 5 será necessário agrupar classes contíguas	

6. Com base numa amostra casual de 308 empresas, foi estimado o seguinte modelo de regressão linear múltipla, com o objectivo de estudar o nível de investimento em investigação e desenvolvimento:

$$\log(\text{RD}_\text{i}) = \beta_0 + \beta_1 \text{ATIV}_\text{i} + \beta_2 \text{TECMIX}_\text{i} + \beta_3 \text{FUNC.G30}_\text{i} + \beta_4 \text{FUNC.LEQ30}_\text{i} + u_i$$

onde *log* representa o logarítmo natural, e as variáveis têm o seguinte significado:

RD: montante investido anualmente em investigação e desenvolvimento (milhares de Euros);

ATIV: número de anos de atividade da empresa;

TECMIX: índice do nível tecnológico da empresa (valores de 1 = baixo a 6 = elevado);

FUNC.G30: número de funcionários com idade superior a 30 anos;

FUNC.LEQ30: número de funcionários com idade inferior ou igual a 30 anos.

Tendo em conta os resultados da estimação do modelo na Equação 1 e das restantes regressões auxiliares apresentadas no Anexo, responda as seguintes questões:

a) **[15]** Supondo satisfeitas as hipóteses do modelo de regressão linear, teste ao nível de 5% a significância estatística individual dos parâmetros β_1 e β_2 da **Equação 1** e interprete as suas estimativas.

RESPOSTA:

$\hat{\beta}_1 = 0.005876$ (semi-elasticidade): estima-se que, por cada ano adicional de atividade da empresa, o montante investido em I&D aumenta, em média, aproximadamente 0.59%, mantendo constantes as restantes variáveis.

$\hat{\beta}_2 = -0.039838$ (semi-elasticidade): estima-se que, quando o índice de mix tecnológico aumenta uma unidade, o montante investido em I&D diminui, em média, aproximadamente 3.98%, mantendo constantes as restantes variáveis.

Testes de significância individual:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0, \quad j = 1, 2$$

Sob H_0, a estatística de teste é: $T_j = \hat{\beta}_j / se\left(\hat{\beta}_j\right) \sim t(303)$.

Pelo output da Equação 1, verifica-se que o valor-p associado aos testes é, respetivamente, $p_1 = 0.01$ e $p_2 = 0.008$, rejeitando-se H_0 aos níveis de significância habituais. Conclui-se assim que ambas as variáveis são estatisticamente significativas.

b) Pelos resultados da **Equação 2** podemos concluir que

não há evidência de heterocedasticidade nos erros do modelo da Equação 1	X
a forma funcional apresenta erros de especificação	
não há evidência de má especificação na forma funcional	
a variância dos erros depende negativamente do número de funcionários com mais de 30 anos	

c) **[20]** Será que a evidência estatística é favorável à hipótese de que o efeito de mais um funcionário com idade superior a 30 anos é compensado por mais um funcionário com idade inferior ou igual a 30 anos? Justifique através de um teste estatístico adequado.

RESPOSTA:

A hipótese a testar é $H_0 : \beta_3 + \beta_4 = 0$ contra $H_1 : \beta_3 + \beta_4 \neq 0$. Seja $\theta = \beta_3 + \beta_4$

$\Leftrightarrow \beta_4 = \theta - \beta_3$. A hipótese a testar será então: $H_0 : \theta = 0$ contra $H_0 : \theta \neq 0$.

Substituindo no modelo da Equação 1, obtém-se o modelo reparametrizado:

$$\log(\text{RD}) = \beta_0 + \beta_1 \text{ATIV} + \beta_2 \text{TECMIX} + \beta_3 \text{FUNC.G30} + (\theta - \beta_3)\text{FUNC.LEQ30} + u$$

$$\log(\text{RD}) = \beta_0 + \beta_1 \text{ATIV} + \beta_2 \text{TECMIX} + \beta_3 (\text{FUNC.G30} - \text{FUNC.LEQ30}) + \theta\, \text{FUNC.LEQ30} + u$$

que corresponde à Equação 3 do anexo. Dos resultados obtém-se $\hat{\theta} = -0.019375$, com um valor-p de 0.696, ou seja não se rejeita $H_0 : \theta = 0$. É de admitir que os dois efeitos compensam-se.

d) **[15]** Avalie a hipótese de que o número de funcionários da empresa não influencia a quantidade investida anualmente em investigação e desenvolvimento. Justifique através de um teste de hipótese adequado.

RESPOSTA:

Teste de restrições de exclusão. Testa-se a hipótese

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

Sob H_0,

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n-k-1)} \sim F(q, n-k-1) = F(2, 303)$$

Do output, $SSR_R = 104.0938$ e $SSR_{UR} = 100.1171$, portanto

$$F_{obs} = \frac{(104.0938 - 100.1171)/2}{100.1171/303} = 6.0176$$

com valor crítico a 5% dado por $F(2, 303; 0.05) = 3$.

Rejeita-se H_0 já que o valor observado é superior ao valor crítico: há evidência de que β_3 e β_4 são, em conjunto, estatisticamente significativos. Não é de admitir que o número de funcionários não influencia as despesas em I&D.

e) Uma empresa acaba de contratar 3 novos funcionários, dos quais apenas 1 tem idade inferior ou igual a 30 anos. Segundo a **Equação 1**, qual é aproximadamente a alteração percentual esperada no valor do investimento em investigação e desenvolvimento, mantendo tudo o resto constante?

4.3%	
7.0%	X
1.7%	
10.4%	

f) **[20]** Um investigador suspeita que o número de anos de atividade possa ter um efeito quadrático. Se isso for verdade, qual as consequências sobre os estimadores OLS dos parâmetros da **Equação 1**? O investigador reestimou o modelo, acrescentando como regressor adicional YFIT^2, onde YFIT são os valores ajustados da **Equação 1**. Sabendo que obteve para o correspondente coeficiente um valor de -1.40115, com standard error igual a 0.892599, será que podemos concluir que o investigador tinha razão?

RESPOSTA:

Se o número de anos de atividade tiver um efeito quadrático, então a especificação da Equação 1 não estaria correta, tendo omitido uma variável relevante (ATIV^2), que está correlacionada com um dos regressores do modelo (ATIV). Portanto, o erro da Equação 1 estaria correlacionado com os regressores, provocando enviesamento do estimador OLS.

Teste Reset de especificação da forma funcional. Pretende-se testar:

$$H_0 : \text{Especificação correcta} \qquad \textit{contra} \qquad H_1 : \text{Presença de erros de especificação}$$

Regressão auxiliar:

$$\log(\text{RD}) = \beta_0 + \beta_1 \text{ATIV} + \beta_2 \text{TECMIX} + \beta_3 \text{FUNC.G30} + \beta_4 \text{FUNC.LEQ30} + \delta\ \text{YFIT}^2 + u$$

Com base na regressão auxiliar:

$$H_0' : \delta = 0 \quad \text{contra} \quad H_1' : \delta \neq 0$$

Sob H_0, a estatística teste é $T = \hat{\delta}/se(\hat{\delta}) \overset{a}{\sim} N(0,1)$ e a região crítica é dada por $W_{5\%} = \{t : |t| > 1.96\}$.

Pelo dados, o valor observado da estatística é $t_{\text{obs}} = (-1.40115)/0.892599 = -1.56974 \notin W_{5\%}$: não há evidência de má especificação, e portanto o número de anos de atividade não parece ter um efeito quadrático: o investigador não tem razão.

Anexo EN Janeiro 2017

Equação 1

Dependent Variable: log(RD)
Included observations: 308

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.468145	0.168897	38.30	0.000
ATIV	0.005876	0.002257	2.60	0.010
TECMIX	-0.039838	0.014879	-2.68	0.008
FUNC.G30	0.089042	0.025668	3.47	0.001
FUNC.LEQ30	-0.108417	0.058574	-1.85	0.065

R-squared	0.0817	Sum squared resid	100.1171
Adjusted R-squared	0.0696	F-statistic	6.74
		Prob(F-statistic)	0.000

Equação 2

Dependent Variable: RESID^2
Included observations: 308

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.306369	0.119608	2.56	0.011
ATIV	0.000125	0.001599	0.08	0.938
TECMIX	0.009147	0.010537	0.87	0.386
FUNC.G30	-0.019068	0.018178	-1.05	0.295
FUNC.LEQ30	0.034835	0.041480	0.84	0.402

R-squared	0.0069	Sum squared resid	50.20971
Adjusted R-squared	0.0063	F-statistic	0.52
		Prob(F-statistic)	0.719

Nota: a variável RESID representa os resíduos da Equação 1

Equação 3

Dependent Variable: log(RD)
Included observations: 308

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.468145	0.168897	38.30	0.000
ATIV	0.005876	0.002257	2.60	0.010
TECMIX	-0.039838	0.014879	-2.68	0.008
FUNC.G30 - FUNC.LEQ30	0.089042	0.025668	3.47	0.001
FUNC.LEQ30	-0.019375	0.049590	-0.39	0.696

R-squared	0.0817	Sum squared resid	100.1171
Adjusted R-squared	0.0696	F-statistic	6.74
		Prob(F-statistic)	0.000

Equação 4

Dependent Variable: log(RD)
Included observations: 308

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.752736	0.142166	47.5	0.000
ATIV	0.004994	0.002166	2.31	0.022
TECMIX	-0.043528	0.015082	-2.89	0.004

R-squared	0.0452	Sum squared resid	104.0938
Adjusted R-squared	0.0389	F-statistic	7.22
		Prob(F-statistic)	0.000