Estatística II

Licenciatura em Gestão

1º semestre – 2015/2016
14/01/2016

Nome____ **N^o**__________

Espaço reservado a classificações

A utilização do telemóvel, em qualquer circunstância, é motivo suficiente para a anulação da prova.

Perguntas de escolha múltipla: apenas uma opção é correta
cada resposta certa vale **1** valor
cada resposta errada vale **-0.25** valores

Pergunta Verdadeiro/Falso: cada resposta certa vale **0.25** valores
cada resposta errada vale **-0.25** valores

Se necessitar de espaço utilize a última página do enunciado, indicando com clareza a respetiva questão.

É expressamente proíbido destacar as folhas do enunciado.

TESTE I

1. Para cada afirmação assinale se esta é verdadeira **(V)** ou falsa **(F)**. (1.0)

	V	**F**
Num teste estatístico de dimensão $\alpha = 0.05$, rejeita-se H_0 sempre que $t < 0.05$, em que t é o valor observado da estatística de teste.		×
Seja $T = \dfrac{X_1 + X_2 + \sum_{i=1}^{n} X_i}{n+2}$ um estimador para o valor esperado de uma variável aleatória X, $\mu = E(X)$, com base numa amostra casual de dimensão n. Então T é estimador centrado de μ.	×	
Num teste estatístico de dimensão $\alpha = 0.1$ não se rejeita H_0 se o valor-p for 0.09.		×
Uma estatística é uma variável aleatória.	×	

2. Num restaurante foram selecionadas ao acaso 61 faturas com indicação de NIF, tendo-se verificado que a soma dos montantes foi de 1279 euros e a sua variância corrigida $s'^2 = 62$. Assume-se que o montante, em euros, de faturas com NIF tem distribuição normal.

(a) Pretende-se testar, ao nível de significância de 5%, se o montante médio de faturas com NIF é 21 euros. A região crítica do teste estatístico correspondente será: (1.0)

- ☐ $W = \left\{(x_1,\ldots,x_n) : \bar{x} < 21 - 2\sqrt{1.0164}\right\}$
- ☐ $W = \left\{(x_1,\ldots,x_n) : \bar{x} > 21 + 2\sqrt{1.0164}\right\}$
- ☐ $W = \left\{(x_1,\ldots,x_n) : \bar{x} < 21 - 1.645\sqrt{1.0164} \text{ ou } \bar{x} > 21 + 1.645\sqrt{1.0164}\right\}$
- ☒ $W = \left\{(x_1,\ldots,x_n) : \bar{x} < 21 - 2\sqrt{1.0164} \text{ ou } \bar{x} > 21 + 2\sqrt{1.0164}\right\}$

(b) Com base na alínea anterior, o que pode concluir sobre a afirmação "o montante médio das faturas com NIF é 21"? Justifique. (1.0)

X: montante, em euros, de faturas com NIF

$$X \sim Normal(\mu_X, \sigma^2)$$

Pela alínea anterior a região de rejeição para a média da amostra do teste estatístico

$$H_0: \mu_X = 21 \quad \text{contra} \quad H_1: \mu_X \neq 21$$

para um nível de significância de 5% é $W =]-\infty, 18.98367[\cup]23.01633, +\infty[$.
Tem-se $\bar{x} = \dfrac{1279}{61} = 20.96721 \notin W$, logo não se rejeita a hipótese de que o montante médio das faturas com NIF, μ, é 21, ao nível de significância de 5%.

(c) Na mesma loja foram selecionadas ao acaso outras 61 faturas, mas sem indicação de NIF, tendo-se verificado que a soma dos montantes foi de 1324 euros. Assume-se que o montante de faturas sem NIF também tem distribuição normal e que a variância das faturas com ou sem NIF é 50. Usando um teste adequado, comente a veracidade da afirmação "o montante médio das faturas com indicação de NIF é inferior ao das faturas sem NIF". (2.0)

X: montante, em euros, de faturas com NIF $\qquad$ Y : montante, em euros, de faturas sem NIF

$$X \sim Normal(\mu_X, \sigma^2 = 50) \quad \text{e} \quad Y \sim Normal(\mu_Y, \sigma^2 = 50)$$

Pretende-se testar:

$$H_0: \mu_X - \mu_Y \geqslant 0 \quad \text{contra} \quad H_1: \mu_X - \mu_Y < 0$$

Estatística Teste: $T_0 = \dfrac{\bar{X} - \bar{Y}}{\sqrt{2\frac{50}{61}}} \sim Normal(0,1)$

Valor observado da estatística teste: $\bar{x} = 20.96721$ e $\bar{y} = \dfrac{1324}{61} = 21.70492$, logo $t_{obs} = \dfrac{20.96721 - 21.70492}{1.280369} = -0.5761698$
Valor-p: $P(T_0 < t_{obs}) = 1 - \Phi(-0.58) = 1 - 0.7190 = 0.281 \Rightarrow$ valor-p = 28%
Logo, não se rejeita a hipótese H_0. Há evidência a favor de H_0, de que o montante médio das faturas com indicação de NIF é superior ou igual ao das faturas sem indicação de NIF.
Alternativamente, a região de rejeição do teste ao nível de significância de 5% é $W =]-\infty, -1.645[$ e $t_{obs} \notin W$, logo não se rejeita H_0 ao nível de significância de 5%.

3. Para inferir sobre a probabilidade θ de cancelamento de reservas num hotel por parte dos seus clientes, o gestor do hotel decidiu analisar os processos relativos a 230 reservas passadas, escolhidas ao acaso da sua vasta carteira, registando-se para cada reserva se a mesma tinha sido cancelada ou não.

(a) Mostre que o estimador do método dos momentos para θ é centrado. (1.0)

$$X = \begin{cases} 1, & \text{reserva cancelada} \\ 0, & \text{caso contrário} \end{cases} \qquad \text{logo} \qquad X \sim Bernoulli(\theta) \qquad \text{e} \qquad E(X) = \theta.$$

Pelo método dos momentos $E(X) = \bar{X}$, logo o estimador de θ pelo método dos momentos é $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Tem-se

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) \underbrace{=}_{ident.dist.} \frac{1}{n}\sum_{i=1}^{n} E(X) = \frac{nE(X)}{n} = E(X) = \theta$$

logo, o estimador do método dos momentos para θ, $\bar{X}$, é centrado.

(b) De acordo com a desigualdade de Fréchet-Cramér-Rao, o estimador T da alínea anterior é o mais eficiente se: (1.0)

☐ $Var(T) = 0$ ☒ $Var(T) = \frac{\theta(1-\theta)}{230}$ ☐ $Var(T) < \frac{\theta(1-\theta)}{230}$ ☐ $Var(T) > \frac{\theta(1-\theta)}{230}$

(c) Sabendo que, das 230 reservas analisadas, 29 foram canceladas, construa um intervalo com nível de confiança aproximadamente 95% para θ. (1.5)

Variável Fulcral: $\frac{\bar{X}-\theta}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}} \overset{a}{\sim} Normal(0,1)$

Logo, sendo $\bar{x} = \frac{29}{230} = 0.126087$, o intervalo de confiança aproximadamente a 95% é dado por

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} = 0.126087 \pm 1.96\sqrt{\frac{0.126087(1-0.126087)}{230}} = (0.0831866, 0.1689874)$$

(d) Qual deveria ser a dimensão da amostra casual por forma a garantir que o "erro" cometido no intervalo da alínea anterior é inferior a 4%? (1.5)

Pretende-se determinar n tal que $P(|\bar{X}-\theta|<0.04)\geqslant 0.95$:

$$0.95 \leqslant P(|\bar{X}-\theta|<0.04)=P(-0.04<\bar{X}-\theta<0.04)=P\left(-\frac{0.04}{\sqrt{\frac{\theta(1-\theta)}{n}}}<\frac{\bar{X}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}<\frac{0.04}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right)$$

$$\underbrace{\Leftrightarrow}_{\frac{\bar{X}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}\sim N(0,1)} \quad \Phi\left(\frac{0.04}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right)-\Phi\left(-\frac{0.04}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right)\geqslant 0.95 \Leftrightarrow \Phi\left(\frac{0.04}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right)-\left(1-\Phi\left(\frac{0.04}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right)\right)\geqslant 0.95$$

$$\Leftrightarrow \quad \Phi\left(\frac{0.04}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right)\geqslant \frac{1.95}{2}=0.975 \Leftrightarrow \frac{0.04}{\sqrt{\frac{\theta(1-\theta)}{n}}}\geqslant 1.96 \Leftrightarrow \left(\frac{0.04}{1.96}\right)^2\geqslant \frac{\theta(1-\theta)}{n}$$

$$\Leftrightarrow \quad n\geqslant \left(\frac{1.96}{0.04}\right)^2\theta(1-\theta)$$

Considerando o valor máximo que a variância de X ($V(X)=\theta(1-\theta)$) pode tomar, obtém-se

$$n\geqslant \left(\frac{1.96}{0.04}\right)^2\frac{1}{4}=600.25$$

Logo, $n=601$.

TESTE II

1. Para cada afirmação assinale se esta é verdadeira **(V)** ou falsa **(F)**. (1.0)

	V	**F**
Se num teste de ajustamento que envolve 2 parâmetros desconhecidos e onde não foi necessário agrupar classes se tem, sob H_0, $Q \overset{a}{\sim} \chi^2_{(9)}$, onde Q é a estatística teste, então o número de classes é 11.		×
No MRL a variável residual u é uma variável aleatória não observável porque depende dos coeficientes do modelo, que são desconhecidos.	×	
O teste de permanência de estrutura é um teste que permite detetar a presença de autocorrelação nos erros do modelo.		×
Uma das hipóteses do MRL é a heterocedasticidade condicionada.		×

2. Para testar se o desempenho profissional de determinada categoria de trabalhadores depende do género inquiriu-se uma amostra casual simples de 1000 trabalhadores tendo-se obtido: (2.0)

	género	
desempenho	masculino	feminino
baixo	170	120
médio	260	180
alto	170	100

Teste, ao nível de significância de 5%, a hipótese de independência entre o desempenho profissional e o género.

Pretende-se testar

$$H_0: p_{ij} = p_{i\cdot} \times p_{\cdot j} \forall i,j \qquad \text{contra} \qquad H_1: \exists (i,j): p_{ij} \neq p_{i\cdot} \times p_{\cdot j}$$

Estatística Teste (sob H_0): $Q = \sum_i \sum_j \frac{N_{ij} - n\hat{p}_{ij}}{n\hat{p}_{ij}} \overset{a}{\sim} \chi^2_{(2)}$ (número de graus de liberdade=$(3-1)\times(2-1)=2$)

Logo, a região crítica, ao nível de significância de 5% é $W = \{q: \ q > 5.991\}$.

A tabela da frequência esperada em cada classe, sob a validade de H_0 é dada por:

	género		
desempenho	masculino	feminino	$\hat{p}_{i\cdot}$
baixo	174	116	0.29
médio	264	176	0.44
alto	162	108	0.27
$\hat{p}_{\cdot j}$	0.6	0.4	1

O valor observado da estatística teste é

$$q_{obs} = \frac{(170-174)^2}{174} + \frac{(120-116)^2}{116} + \frac{(260-264)^2}{264} + \frac{(180-176)^2}{176} + \frac{(170-162)^2}{162} + \frac{(100-108)^2}{108} = 1.369 < 5.991$$

Logo $q_{obs} = 1.369 \notin W$, pelo que não se rejeita a hipótese ao nível de significância de 5%, ou seja há evidência de independência entre o desempenho profissional e o género.

3. Para analisar o comportamento das vendas ($vendas$) das lojas de material de telecomunicação com base nas despesas em publicidade ($desp$) e na superfície da loja ($area$) em m^2, foi especificado o seguinte modelo

$$lvendas = \beta_1 + \beta_2 ldesp + \beta_3 area + \beta_4 cc + u$$

Os resultados obtidos encontram-se abaixo, onde o prefixo l representa o logaritmo natural e cc é uma variável artificial que assume o valor 1 se a loja se situa num centro comercial e o valor 0 caso contrário.

Regression Statistics	
Multiple R	0.6952
R Square	0.4833
Adjusted R Square	0.4744
Standard Error	0.2660
Observations	179

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	11.5849	3.8616	54.5590	0.0000
Residual	175	12.3863	0.0708		
Total	178	23.9712			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	9.9147	0.1395	71.0640	0.0000
ldesp	0.9620	0.2515	3.8256	0.0002
area	0.0062	0.0010	6.4936	0.0000
cc	0.1568	0.0649	2.4174	0.0167

(a) Interprete as estimativas de β_3 e de β_4 e analise a significância estatística da variável artificial, considerando um nível de significância de 5%. (1.5)

$b_3 = 0.0062$: Estima-se que, se a área aumenta 1 m^2, as vendas aumentam em média aproximadamente 0.62%, mantendo constantes as restantes variáveis.

$b_4 = 0.1568$: Estima-se que as vendas de uma loja situada num centro comercial são, em média, cerca de 15.68% superiores relativamente às vendas de uma loja situada fora de um centro comercial, para os mesmos valores das restantes variáveis.

Para analisar a significância estatística da variável artificial testa-se:

$$H_0: \beta_4 = 0 \qquad \text{contra} \qquad H_1: \beta_4 \neq 0$$

Sob H_0, a estatística teste é $\dfrac{b_4}{s.e.(b_4)} \sim t_{(175)}$ e a região de rejeição, ao nível de significância de 5%, é (aproximadamente) $W = \{t : |t| > 1.96\}$.
Como $t_{obs} = 2.4174 \in W$, rejeita-se H_0 ao nível de significância de 5%. Logo, a variável cc é relevante na explicação de lvendas.

(b) Construa um intervalo de confiança a 95% para a elasticidade das vendas relativamente às despesas em publicidade e teste se é de admitir que um aumento de 2% nas despesas em publicidade origine um aumento, em média, de 3% nas vendas, mantendo constantes as restantes variáveis. (2.0)

Variável Fulcral: $T = \dfrac{b_2 - \beta_2}{s.e.(b_2)} \sim t_{(175)}$ Logo, o intervalo de confiância 95% para β_2 é:

$$b_2 \pm 1.96 s.e.(b_2) = 0.962 \pm 1.96 \times 0.2515 = (0.4691, 1.4549)$$

Pretende-se testar

$$H_0 : \beta_2 = 1.5 \qquad \text{contra} \qquad H_1 : \beta_2 \neq 1.5$$

Estatística teste (sob H_0): $T = \dfrac{b_2 - 1.5}{s.e.(b_2)} \sim t_{(175)}$

Com base no intervalo de confiânça a 95% para β_2, rejeita-se H_0 ao nível de significância de 5% pois $1.5 \notin IC_{(95\%)}$. Portanto, não há evidência estatística que suporte a afirmação.

(c) Interprete o coeficiente de determinação R^2 e teste a significância global do modelo. (1.5)

$R^2 = 0.4833$**:** 48.33% da variação total de lvendas é explicada pelo modelo.

Pretende-se testar

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \qquad \text{contra} \qquad H_1 : \exists \beta_j \neq 0, \;\; j = 2,3,4$$

Estatística teste: $\dfrac{(VR_0 - VR_1)/3}{VR_1/(179-4)} = \dfrac{R^2/(4-1)}{(1-R^2)/(179-4)} \sim F_{(3,175)}$

Como $F_{obs} = 54.559$ com valor-p=0, rejeita-se H_0. β_2, β_3e β_4 são, em conjunto, estatísticamente significativos.

(d) Para testar a nulidade conjunta de β_3 e β_4 foi estimado o modelo $lvendas = \beta_1 + \beta_2 ldesp + u$, tendo-se obtido $VR = 16.959$, onde VR é a variação residual. Então, sendo F a estatística de teste adequada, tem-se (1.0)

- ☐ Sob H_0, $F \sim F_{(2,177)}$ e $F_{obs} = 32.67$
- ☐ Sob H_0, $F \sim F_{(2,176)}$ e $F_{obs} = 32.48$
- ☒ Sob H_0, $F \sim F_{(2,175)}$ e $F_{obs} = 32.30$
- ☐ Sob H_0, $F \sim F_{(2,174)}$ e $F_{obs} = 32.11$

(e) Com base na mesma amostra e definindo $ncc = 1 - cc$, foi estimado o modelo (1.0)

$$lvendas = \alpha_1 + \alpha_2 ldesp + \alpha_3 area + \alpha_4 ncc + u$$

Apresente, justificando, a equação ajustada desse modelo.

Substituindo no modelo inicial $cc = 1 - ncc$, vem

$$\begin{aligned} & lvendas = \beta_1 + \beta_2 ldesp + \beta_3 area + \beta_4(1 - ncc) + u \\ \Leftrightarrow\; & lvendas = \underbrace{(\beta_1 + \beta_4)}_{\alpha_1} + \underbrace{\beta_2}_{\alpha_2} ldesp + \underbrace{\beta_3}_{\alpha_3} area \underbrace{-\beta_4}_{\alpha_4} ncc + u \end{aligned}$$

Logo

$$\begin{aligned} \widehat{lvendas} &= 9.9147 + 0.1568 + 0.962 ldesp + 0.0062 area - 0.1568 ncc \\ &= 10.0715 + 0.962 ldesp + 0.0062 area - 0.1568 ncc \end{aligned}$$