Probabilidades e Estatística E - versão A
Exame de Recurso: 28-06-2014
Duração: 2h30

Nome completo:____

Nº de aluno:____________ **Nº de caderno:**____________

- **Nos grupos 1 a 4, assinale com uma cruz sobre V para verdadeiro ou sobre F para falso, o valor lógico de cada uma das afirmações.**
- **Uma resposta correcta vale 1 valor, uma resposta incorrecta desconta 0.4 valores e uma não resposta nada desconta.**
- **Os grupos 5, 6 e 7 devem ser respondidos no caderno de exame.**

1. Suponha que 10% dos processadores instalados num conjunto de computadores de uma empresa são da marca A e que, de entre estes, 70% sobreaquecem. De entre os que não são da marca A 45%, sobreaquecem. Seleccionado ao acaso um computador desta empresa: (3.0)

   Considere os acontecimentos: A-o processador é da marca A e S-o processador sobreaquece. Temos então que: $P(A) = 0.1; P(\bar{A}) = 0.9; P(S|A) = 0.7; P(S|\bar{A}) = 0.45$.

V F A probabilidade do processador sobreaquecer é 0.475.

R: $P(S)=P(S|A)P(A)+P(S|\bar{A})P(\bar{A})=0.7\times0.1+0.45\times0.9=0.475$ (Teo. da probabilidade total).

V F Se o processador sobreaquecer, a probabilidade de ser da marca A é 0.8526 (arredondado a 4 casas décimais).

R: $P(A|S) = \frac{P(S|A)P(A)}{P(S)} = \frac{0.07}{0.475} = 0.1474$ (Teo. de Bayes).

V F A probabilidade do processador sobreaquecer ou de ser da marca A é 0.505.

R: $P(S \cup A) = P(S) + P(A) - P(S \cap A) = P(S) + P(A) - P(S|A)P(A) = 0.475 + 0.1 - 0.07 = 0.505$ (Teo. probabilidade composta).

2. Considere o vector aleatório (X, Y) com a seguinte função de probabilidade conjunta:

$X \setminus Y$	0	2	4	
0	0.1	0.3	0	0.4
1	0.1	0	0.2	0.3
2	0.1	0.1	0.1	0.3
	0.3	0.4	0.3	1

(3.0)

V F $P(X + Y \leq 1) = 0.8$.

R: $P(X + Y \leq 1) = P(X = 0; Y = 0) + P(X = 1; Y = 0) = 0.2$.

V F As variáveis X e Y são independentes.

R: X e Y serão independentes se, e só se, $\forall x, y, P(X = x; Y = y) = P(X = x)P(Y = y)$ como $P(X = 1; Y = 2) = 0 \neq 0.12 = 0.3 \times 0.4 = P(X = 1)P(Y = 2)$, as v.a. não serão independentes.

V F $Cov(X, Y) = 0.2$.

R: $Cov(X, Y) = E(XY) - E(X)E(Y) = (1 \times 4 \times 0.2 + 2 \times 2 \times 0.1 + 2 \times 4 \times 0.1) - (1 \times 0.3 + 2 \times 0.3) \times (2 \times 0.4 + 4 \times 0.3) = 0.2$.

3. Um fabricante de memórias flash constatou que o peso das memórias tem distribuição Normal, dependendo os seus parâmetros de a memória ser de 16Gb ou 32Gb - caso seja de 16Gb a média é de 45g e o desvio padrão 1g e caso seja de 32Gb a média é de 50g e o desvio padrão é de 2g. Sabendo que as memórias produzidas são 60% de 16Gb e 40% de 32Gb: (3.0)

Considere as v.a. X-peso de uma memória de 16Gb e Y-peso de uma memória de 32Gb. Sabe-se que: $X \sim N(45, 1^2)$ e $Y \sim N(50, 2^2)$.

V F A probabilidade de uma memória de 32Gb pesar menos de 50g é 0.5.

R: $P(Y < 50) = P\left(\frac{Y-50}{2} < \frac{50-50}{2}\right) = \Phi(0) = 0.5.$

V F A probabilidade de uma memória escolhida ao acaso pesar menos de 45g é 0.3025 (arredondado a 4 casas décimais).

R: Considere-se M-peso de uma memória escolhida ao acaso.

$$\begin{aligned}
&P(M < 45) = P(M < 45|\text{"Memória escolhida é de 16Gb"})P(\text{"Memória escolhida é de 16Gb"})+\\
&+ P(M < 45|\text{"Memória escolhida é de 32Gb"})P(\text{"Memória escolhida é de 32Gb"}) =\\
&= P(X < 45)\,0.6 + P(Y < 45)\,0.4 = P\left(\frac{X-45}{1} < \frac{45-45}{1}\right)0.6 + P\left(\frac{Y-45}{2} < \frac{45-50}{2}\right)0.4\\
&= \Phi(0) \times 0.6 + \Phi(-2.5) \times 0.6 = 0.3025
\end{aligned}$$

V F A probabilidade de um conjunto de 2 memórias de 16Gb e 3 de 32Gb ter um peso total superior a 240g é 0.7.

R: Considerem-se X_i-peso da i-ésima memória de 16Gb ($i = 1, 2$) e Y_j-peso da j-ésima memória de 32Gb ($j = 1, 2, 3$), então o peso do total do conjunto das 5 memórias é

$$T = X_1 + X_2 + Y_1 + Y_2 + Y_3 \sim N(240, 14).$$

Vindo que

$$P(T > 240) = 1 - P\left(\frac{T-240}{\sqrt{14}} \leq \frac{240-240}{\sqrt{14}}\right) = \Phi(0) = 0.5.$$

4. Considere uma amostra aleatória $(X_1, X_2, ..., X_n)$ de uma população com valor médio μ e variância σ^2. (3.0)

V F $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ não é estimador consistente da média populacional.
R: Como $EQM(\overline{X}) = \frac{\sigma^2}{n} \to 0$ quando $n \to \infty$, então $\overline{X}$ é estimador consistente para a média populacional.

V F Enquanto estimadores de μ, $\hat{\theta}_1 = \frac{2X_1+3X_2+5X_3}{10}$ é mais eficiente que $\hat{\theta}_2 = \frac{X_1+X_n}{2}$.

R: Como $EQM(\hat{\theta}_1) = \frac{38}{100}\sigma^2 < \frac{\sigma^2}{2} = EQM(\hat{\theta}_2)$, então $\hat{\theta}_1$ é mais eficiente que $\hat{\theta}_2$.

V F $\left(\overline{X}\right)^2$ é estimador centrado de μ^2.

R: Por $E(\overline{X}^2) = V(\overline{X}) + (E(\overline{X}))^2 = \frac{\sigma^2}{2} + \mu^2 \neq \mu^2$, conclui-se que $\left(\overline{X}\right)^2$ não é estimador centrado de μ^2.

Justifique detalhadamente as suas respostas

5. A quantidade de tempo, em segundos, que um computador demora a processar um determinado conjunto de informação é uma v.a. com a seguinte função densidade probabilidade:

$$f(x) = \begin{cases} ke^{-\frac{x}{10}}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

(3.0)

(a) Determine o valor da constante k.

R: Por $f(x)$ ser função densidade de probabilidade sabe-se que:

$$\forall x \in \mathbb{R}, f(x) \geq 0 \quad \text{e} \quad \int_{-\infty}^{+\infty} f(x)dx = 1.$$

Como,

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{+\infty} ke^{-\frac{x}{10}}dx = 0 - 10k\left[e^{-\frac{x}{10}}\right]_0^{+\infty} = -10k(0-1) = 10k$$

teremos

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \Leftrightarrow 10k = 1 \Leftrightarrow k = \frac{1}{10}.$$

(b) Determine a função distribuição de X.

R: A função distribuição, $F(x)$, é por definição,

$$\forall x \in \mathbb{R}, F(x) = P(X \leq x).$$

Então teremos que:

$$\begin{cases} x < 0, & F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du = \int_{-\infty}^{x} 0du = 0 \\ \\ x \geq 0, & F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du = \int_{-\infty}^{0} 0du + \int_{0}^{x} \frac{e^{-\frac{u}{10}}}{10}du = 0 - \left[e^{-\frac{u}{10}}\right]_0^x = 1 - e^{-\frac{x}{10}} \end{cases}$$

Vindo assim,

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\frac{x}{10}}, & x \geq 0 \end{cases}$$

(c) Qual a probabilidade de um computador demorar mais de 10 segundos a processar um conjunto de informação?

R: Queremos,

$$P(X > 10) = 1 - P(X \leq 10) = 1 - F(10) = 1 - (1 - e^{-\frac{10}{10}}) = e^{-1}.$$

6. Sabe-se que numa determinada loja, a procura de videojogos por dia e em euros é uma v.a. normalmente distribuída com desvio padrão 40€. O gerente afirma que a procura média é superior a 200€ e numa amostra aleatória constituída por 9 dias seleccionados ao acaso verificou-se que $\overline{x} = 216$. (2.5)

   (a) Teste, ao nível de significância de 5%, a afirmação feita pelo gerente.

   R: Pretende-se realizar um teste de hipóteses unilateral direito para o valor médio da procura (em euros) de videojogos.

   - Hipóteses:
   $$H_0 : \mu \leq 200 \text{ vs } H_1 : \mu > 200$$
   - Estatística de teste:
   $$Z = \frac{\overline{X} - 200}{\sigma/\sqrt{n}} \overset{sob\ H_0}{\sim} N(0,1).$$
   Uma vez que a população é normal e o desvio padrão populacional é conhecido.
   - Região de rejeição para $\alpha = 5\%$:
   $$R_{5\%} \equiv]z_{0.05}, \infty[\equiv]1.645, +\infty[$$
   - Regra de decisão: Rejeitar H_0 se $z_{obs} \in R_{5\%}$. Como $z_{obs} = \frac{216-200}{40/3} = 1.2 \notin R_{5\%}$, pelo que não se rejeita H_0, ou seja, os dados analisados não suportam a afirmação do gerente, para um nível de significância de 5%.

   (b) Resolva a alínea anterior usando o valor-p (ou p-value).

   R: Como estamos perante um teste unilateral direito e como $z_{obs} = 1.2$, teremos:
   $$p - value = P(Z > 1.2|H_0 \text{ verdade}) = 1 - P(Z \leq 1.2|H_0 \text{ verdade}) = 1 - \Phi(1.2) = 0.1151.$$
   Como $p - value = 0.1151 > 0.05 = \alpha$, não se rejeita H_0 para um nível de significância de $\alpha = 5\%$, concluindo-se que os dados analisados não suportam a afirmação do gerente.

7. Pretende-se modelar a relação entre a dimensão final de ficheiros (Y) e a sua dimensão original (x), quando são comprimidos usando um software específico. Para tal registaram-se, para 10 ficheiros, as suas dimensões em KB, antes e depois da compressão.

i	1	2	3	4	5	6	7	8	9	10
x_i	100	250	500	750	1000	1250	1500	1750	2000	2250
Y_i	4	9	15	16	20	46	54	59	64	70

(2.5)

$$\overline{x} = 1135, \quad \overline{Y} = 35.7, \quad S_{xx} = 4940250, \quad S_{xY} = 164205, \quad S_{YY} = 5742.1.$$

Utilize sempre 4 casas décimais nos seus cálculos.

   (a) Ajuste um modelo de regressão linear simples aos dados e determine o coeficiente de determinação.

   R: Dado o modelo de regressão linear simples $Y = \beta_0 + \beta_1 + \varepsilon$ onde $\varepsilon \sim N(0, \sigma^2)$, queremos começar por estimar os coeficientes de regressão β_0 e β_1. Como
   $$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.0332, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x} = -1.982,$$
   teremos a reta ajustada,
   $$\hat{Y} = -1.982 + 0.0332x.$$
   Para o coeficiente de determinação, teremos
   $$R^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{YY}} = 0.9483.$$

(b) Construa, **de forma detalhada**, um intervalo de confiança a 95% para o verdadeiro declive da recta de regressão.

R: Queremos construir um intervalo de confiança a 95% para o verdadeiro declive da recta de regressão, β_1.

- Variável Pivot:
$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}.$$
- Escolha de c_1 e c_2 tais que $P(c_1 \leq T \leq c_2) = 95\%$:
Tendo em atenção que $T \sim t_{n-2}$, que é simétrica em relação à origem, basta escolher $c_1 = -c_2$ com c_2 tal que
$$P(T \leq c_2) = 0.975 \Leftrightarrow F_{t_{n-2}}(c_2) = 0.975 \Leftrightarrow c_2 = F^{-1}_{t_{n-2}}(0.975) \Leftrightarrow c_2 = t_{n-2;0.025}.$$
- Determinar os extremos do intervalo de confiança:
Com probabilidade 95%, sabemos que,
$$\begin{aligned} & c_1 \leq T \leq c_2 \Leftrightarrow -t_{n-2;0.025} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \leq t_{n-2;0.025} \\ & \Leftrightarrow -t_{n-2;0.025}\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \leq \hat{\beta}_1 - \beta_1 \leq t_{n-2;0.025}\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \\ & \Leftrightarrow \hat{\beta}_1 - t_{n-2;0.025}\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2;0.025}\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \end{aligned}$$
Vindo,
$$IC_{95\%}(\beta_1) \equiv \left[\hat{\beta}_1 - t_{n-2;0.025}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}; \hat{\beta}_1 + t_{n-2;0.025}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right].$$
- Concretização do intervalo de confiança:
Como
$$\hat{\beta}_1 = 0.0332, n = 10, \hat{\sigma} = \sqrt{\frac{S_{YY} - \hat{\beta}_1^2.S_{xx}}{n-2}} = 6.0906 \text{ e } t_{8;0.025} = 2.31$$
virá,
$$IC_{95\%}(\beta_1) \equiv \left[0.0332 - 2.31\frac{6.0906}{\sqrt{4940250}}; 0.0332 + 2.31\frac{6.0906}{\sqrt{4940250}}\right] \equiv [0.0269; 0.0395].$$

(c) Para um ficheiro de dimensão inicial 1100KB qual será a sua dimensão prevista, após a compressão?

R: A dimensão final prevista para um ficheiro de dimensão inicial de 1100KB, será obtida através da reta estimada na alínea a). Assim sendo, para uma dimensão inicial $x = 1100$ teremos
$$\hat{Y}|_{x=1100} = -1.982 + 0.0332 \times 1100 = 34.538.$$