Probabilidades e Estatística

Segundo exame/segundo teste — 2º semestre – 2009/2010

Duração: 180/90 minutos — 03/07/2010 – 9:00 horas

Justifique convenientemente todas as respostas!

Grupo I — 5.0 valores

1. Foram examinados 100 programas de computador a fim de detectar 3 tipos de erros: A, B e C. O resultado foi o seguinte: 20 desses programas tinham erros do tipo A, 10 do tipo B e 5 do tipo C; em 3 havia erros dos tipos A e C, em 6 dos tipos A e B e em 2 dos tipos B e C; havia ainda 2 programas que tinham os 3 tipos de erros. Escolhido ao acaso um programa desses 100, qual é a probabilidade de o programa seleccionado:

(a) ter pelo menos um dos três tipos de erros? (1.5)

Seja A (B,C) o evento um programa tem erros do tipo A (B,C).
$P(A) = 0.2$, $P(B) = 0.1$, $P(C) = 0.05$, $P(A \cap C) = 0.03$, $P(A \cap B) = 0.06$, $P(B \cap C) = 0.02$, $P(A \cap B \cap C) = 0.02$.

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ &\quad - P(B \cap C) + P(A \cap B \cap C) \\ &= 0.2 + 0.1 + 0.05 - 0.03 - 0.06 - 0.02 + 0.02 \\ &= 0.26 \end{aligned}$$

(b) não ter erros do tipo B ou do tipo C, quando tem algum tipo de erro? (1.0)

$$\begin{aligned} P(\bar{B} \cup \bar{C} | A \cup B \cup C) &= \frac{P[(\bar{B} \cup \bar{C}) \cap (A \cup B \cup C)]}{P(A \cup B \cup C)} = \frac{P[(A \cup B \cup C) - (B \cap C)]}{P(A \cup B \cup C)} \\ &= \frac{P(A \cup B \cup C) - P(B \cap C)}{P(A \cup B \cup C)} = 1 - \frac{P(B \cap C)}{P(A \cup B \cup C)} \\ &= 1 - \frac{0.02}{0.26} = \frac{12}{13} \approx 0.9231 \end{aligned}$$

2. O número diário de avarias em determinada máquina é uma variável aleatória X com distribuição de Poisson tal que $P(X = 1) = P(X = 2)$.

(a) Calcule a probabilidade de o número de avarias por dia exceder 3. (1.0)

Se $X \sim \text{Poisson}(\lambda)$ tal que $P(X=1) = P(X=2)$,

$$\mathrm{e}^{-\lambda}\lambda = \frac{\mathrm{e}^{-\lambda}\lambda^2}{2} \;\Rightarrow\; \lambda^2 - 2\lambda = 0 \underset{\lambda>0}{\Rightarrow} \lambda = 2.$$

Logo, $P(X > 3) = 1 - F_X(3) = 1 - \sum_{x=0}^{3} \frac{\mathrm{e}^{-2}2^x}{x!} = 1 - 0.8571 = 0.1429.$

(b) Qual é a mediana do número de dias decorridos até se registar o primeiro dia com pelo menos 4 avarias? (1.5)

Seja Y o número de dias decorridos até se registar o primeiro dia com $X \geq 4$.
Considerando que a ocorrência de um dia com $X \geq 4$ é um ensaio de Bernoulli com probabilidade de sucesso $p = P(X \geq 4) = P(X > 3) = 0.1429$ e que os ensaios são independentes, $Y \sim$ Geométrica(p).
Como $F_Y(y) = \sum_{u=1}^{y} p(1-p)^{u-1}$, $F_Y(1) = 0.1429$, $F_Y(2) = 0.2654$, $F_Y(3) = 0.3704$, $F_Y(4) = 0.4270$, $F_Y(5) = 0.5041$ e $P(Y \geq 5) = 1 - F_Y(4) = 0.5730$.
Assim, a mediana de Y é 5 dias pois $P(X \leq 5) \geq 0.5$ e $P(X \geq 5) \geq 0.5$.

Grupo II	5.0 valores

1. A espessura de coberturas laminadas para placas de madeira apresenta, segundo o seu fabricante, um valor médio de 0.5 cm e um desvio padrão de 0.05 cm. Duzentas dessas coberturas foram usadas por uma empresa para revestir, por cima e por baixo, 100 placas de madeira.

(a) Determine (aproximadamente) a probabilidade de a espessura total de 100 placas (duplamente) revestidas exceder 11.01 m, supondo que cada placa sem qualquer revestimento apresenta uma espessura constante de 10 cm. (1.5)

Seja X a variável aleatória a representar a espessura (em cm) de uma cobertura laminada genérica com $E(X) = 0.5$ e $Var(X) = 0.05^2$.
Seja Y a espessura (em cm) de 100 placas de madeira (com espessura constante de 10 cm) revestidas por cima e por baixo pelas 200 coberturas. Então,

$$Y = 100 \times 10 + \sum_{i=1}^{200} X_i,$$

onde X_i, $i = 1, \ldots, 200$, são supostamente independentes e identicamente distribuídas (i.i.d.) com distribuição comum de X, pelo que

$$E(Y) = 1000 + \sum_{i=1}^{200} E(X_i) \overset{i.d.}{=} 1000 + 200 \times E(X) = 1100$$
$$Var(Y) \overset{ind.}{=} \sum_{i=1}^{200} Var(X_i) \overset{i.d.}{=} 200 \times Var(X) = 0.5.$$

Pelo Teorema Central do Limite

$$P(Y > 1101) = 1 - F_Y(1101) \approx 1 - F_{N(0,1)}\left(\frac{1101 - 1100}{\sqrt{0.5}}\right) = 1 - F_{N(0,1)}(1.414)$$
$$\approx 1 - 0.9207 = 0.0793$$

(b) Admita que as normas de fabrico especificam que a espessura de cada cobertura deve estar compreendida entre 0.45 cm e 0.55 cm. Supondo uma lei Normal para esta variável aleatória, qual o número médio de coberturas adquiridas pela empresa que não satisfazem aqueles requisitos? (1.5)

Se $X \sim N(\mu, \sigma^2)$ com $\mu = 0.5$ e $\sigma^2 = 0.0025$,

$$q = P(0.45 < X < 0.55) = P\left(\frac{0.45-0.5}{0.05} < \frac{X-\mu}{\sigma} < \frac{0.55-0.5}{0.05}\right) = P\left(-1 < \frac{X-\mu}{\sigma} < 1\right)$$

$$= F_{N(0,1)}(1) - F_{N(0,1)}(-1) = 2F_{N(0,1)}(1) - 1 = 2 \times 0.8413 - 1 = 0.6826.$$

Seja W o número de coberturas com $X \notin [0.45, 0.55]$ em 200 usadas pela empresa. Por a) W traduz o número de sucessos em 200 ensaios de Bernoulli i.i.d., pelo que

$$W \sim \text{Binomial}(n{=}200, p{=}1{-}q{=}0.3174).$$

Logo, $E(W) = np = 200{\times}0.3174 = 63.48$.

2. Seja (X, Y) um par aleatório com a seguinte função de distribuição conjunta

$$F_{X,Y}(x,y) = \begin{cases} \frac{2}{\pi}(1-e^{-x})\arctan y, & x,\ y > 0 \\ 0, & \text{caso contrário.} \end{cases}$$

(a) Especifique a função de distribuição marginal de Y. (1.0)

$$F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y) = \begin{cases} \frac{2}{\pi}\arctan y, & y > 0 \quad \left[\lim_{x\to\infty}(1-e^{-x}) = 1\right] \\ 0, & y \le 0. \end{cases}$$

(b) Calcule a probabilidade de X tomar valores menores do que 3 sabendo que $Y = 1$. (1.0)

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y) = \begin{cases} (1-e^{-x}), & x > 0 \quad \left[\lim_{y\to\infty}\arctan y = \frac{\pi}{2}\right] \\ 0, & x \le 0. \end{cases}$$

Como $F_{X,Y}(x,y) = F_X(x)F_Y(y)$, $\forall\,(x,y)$, X e Y são variáveis aleatórias independentes

$$\therefore \quad P(X < 3|Y = 1) = F_{X|Y=1}(3) \overset{ind.}{=} F_X(3) = 1 - \mathrm{e}^{-3} \approx 0.9502$$

Nota: Se precisar, recorde que $(\arctan y)' = 1/(1+y^2)$.

Resolução alternativa:

a) $f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\,\partial y}F_{X,Y}(x,y) = \begin{cases} \frac{2}{\pi}\frac{\partial}{\partial x}(1-\mathrm{e}^{-x})\frac{\partial}{\partial y}\arctan y = \frac{2}{\pi}\frac{1}{1+y^2}\mathrm{e}^{-x}, & x,\ y > 0 \\ 0, & \text{c.c.} \end{cases}$

$f_Y(y) = \int_0^\infty f_{X,Y}(x,y)dx = \begin{cases} \frac{2}{\pi}\frac{1}{1+y^2}\int_0^\infty \mathrm{e}^{-x}dx = \frac{2}{\pi}\frac{1}{1+y^2}[-\mathrm{e}^{-x}]|_0^\infty = \frac{2}{\pi}\frac{1}{1+y^2}, & y > 0 \\ 0, & y \le 0. \end{cases}$

$F_Y(y) = \int_0^y f_Y(v)dv = \begin{cases} \frac{2}{\pi}\int_0^y \frac{1}{1+v^2}dv = \frac{2}{\pi}[\arctan v]|_0^y = \frac{2}{\pi}\arctan y, & y > 0 \\ 0, & y \le 0. \end{cases}$

b) $f_X(x) = \int_0^\infty f_{X,Y}(x,y)dy = \begin{cases} \mathrm{e}^{-x}\frac{2}{\pi}\int_0^\infty \frac{1}{1+y^2}dy = \mathrm{e}^{-x}\frac{2}{\pi}[\arctan y]|_0^\infty = \mathrm{e}^{-x}, & x > 0 \\ 0, & x \le 0. \end{cases}$

Como $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, $\forall\,(x,y)$, X e Y são variáveis aleatórias independentes

$$P(X < 3|Y{=}1) = F_{X|Y=1}(3) \overset{ind.}{=} F_X(3) = \int_0^3 \mathrm{e}^{-u}du = [-\mathrm{e}^{-u}]|_0^3 = 1{-}\mathrm{e}^{-3} \approx 0.9502$$

1. Num estudo sobre a eficiência da execução de certo tipo de tarefas numa empresa fizeram-se n medições $x_1, \ldots, x_n$ de uma variável aleatória X supostamente modelada pela função densidade de probabilidade

$$f_X(x|\theta) = \begin{cases} \theta^{-1} x^{\frac{1-\theta}{\theta}}, & 0 < x < 1 \\ 0, & \text{caso contrário} \end{cases}$$

em que $\theta > 0$. Supondo que as medições efectuadas são uma concretização de uma correspondente amostra aleatória $X_1, \ldots, X_n$:

(a) Mostre que $T = n^{-1} \sum_{i=1}^{n} \log X_i^{-1}$ é o estimador de máxima verosimilhança (e.m.v.) de θ e determine a estimativa de máxima verosimilhança da função $\frac{1-\theta}{\theta}$. (1.5)

Função de verosimilhança:

$$L(\theta|x_1, \ldots, x_n) \stackrel{i.i.d.}{=} \prod_{i=1}^{n} f_X(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\theta}\, x_i^{\frac{1-\theta}{\theta}} = \theta^{-n} \left[\prod_{i=1}^{n} x_i\right]^{\frac{1-\theta}{\theta}}$$

Seja $L_\theta \equiv \log L(\theta|x_1, \ldots, x_n) = -n \log \theta + \frac{1-\theta}{\theta} \sum_{i=1}^{n} \log x_i$.

$L'_\theta = -n\frac{1}{\theta} - \frac{1}{\theta^2} \sum_{i=1}^{n} \log x_i = 0 \ \Rightarrow \ n\theta = -\sum_{i=1}^{n} \log x_i \ \Rightarrow \ \theta = \frac{1}{n} \sum_{i=1}^{n} \log x_i^{-1}$

$L''_\theta = n\frac{1}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^{n} \log x_i^{-1} \ \Rightarrow \ L''_\theta|_{\theta = \frac{1}{n}\sum_{i=1}^{n} \log x_i^{-1}} = -\frac{n^3}{(\sum_{i=1}^{n} \log x_i^{-1})^2} < 0$ para $n \geq 1$.

$$\therefore \qquad \hat{\theta} = n^{-1} \sum_{i=1}^{n} \log X_i^{-1} \text{ é o e.m.v. de } \theta$$

Pela propriedade de invariância dos e.m.v.,

$$g(\hat{\theta}) = \frac{1-\hat{\theta}}{\hat{\theta}} = \frac{n - \sum_{i=1}^{n} \log X_i^{-1}}{\sum_{i=1}^{n} \log X_i^{-1}} \text{ é o e.m.v. de } g(\theta) = \frac{1-\theta}{\theta},$$

cuja estimativa é dada por $\frac{n - \sum_{i=1}^{n} \log x_i^{-1}}{\sum_{i=1}^{n} \log x_i^{-1}}$.

(b) Sabendo que $Y_i = -\log X_i,\ i = 1, \ldots, n$, têm distribuições exponenciais independentes de valor esperado θ, determine o erro quadrático médio do e.m.v. de θ. (1.0)

Se $Y_i = \log X_i^{-1} \sim$ Exponencial$(\frac{1}{\theta})$, $E(Y_i) = \theta$ e $Var(Y_i) = \theta^2$, $i = 1, \ldots, n$ (independentes). O erro quadrático médio de $\hat{\theta} = n^{-1} \sum_{i=1}^{n} Y_i$ é

$$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 = \theta^2/n,$$

visto que $Var(\hat{\theta}) \stackrel{ind.}{=} n^{-2} \sum_{i=1}^{n} Var(Y_i) \stackrel{i.d.}{=} \frac{\theta^2}{n}$ e $E(\hat{\theta}) = n^{-1} \sum_{i=1}^{n} E(Y_i) \stackrel{i.d.}{=} \theta$.

2. Um engenheiro mecânico desenvolveu um novo tipo de pára-choques que julga ser mais resistente ao impacto. Um fabricante de automóveis usou 100 desses pára-choques em acidentes controlados, tendo a percentagem de acidentes que não resultaram em estragos visíveis nos pára-choques sido igual a 70%.

(a) Construa um intervalo de confiança aproximado a 90% para a probabilidade (p) de um acidente similar aos descritos não resultar em estragos visíveis no pára-choques. (1.0)

Seja $(X_1, \dots, X_n)$ uma amostra aleatória de $X \sim \text{Bernoulli}(p)$, onde

$$X = \begin{cases} 1, & \text{não há estrago visível no pára-choques} \\ 0, & \text{caso contrário,} \end{cases}$$

com $E(X) = p$ e $Var(X) = p(1-p)$. Para $n = 100$, $\bar{x} = 0.7$.
Método da variável fulcral (quantidade pivotal):
i) Varíável fulcral: Como a amostra aleatória é grande ($n = 100$), pelo Teorema Central do Limite e pelo facto de $\bar{X} \to p$ quando $n \to \infty$

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

ii) $P(a < W < b) = 0.9 \;\Rightarrow\; b = -a = F^{-1}_{N(0,1)}(0.95) = 1.64$
iii)

$$\begin{aligned} 0.9 = P(-1.64 < W < 1.64) &= P(-1.64 < \frac{\bar{X} - p}{\sqrt{\bar{X}(1-\bar{X})/n}} < 1.64) \\ &= P(\bar{X} - 1.64\sqrt{\bar{X}(1-\bar{X})/n} < p < \bar{X} + 1.64\sqrt{\bar{X}(1-\bar{X})/n}) \end{aligned}$$

iv) Intervalo aleatório de confiança (aproximado) a 90% para p:

$$IAC_a(p, 0.9) = (\bar{X} \pm 1.64\sqrt{\bar{X}(1-\bar{X})/n})$$

Intervalo de confiança (aproximado) a 90% para p:

$$\begin{aligned} IC_a(p, 0.9) &= (\bar{x} \pm 1.64\sqrt{\bar{x}(1-\bar{x})/n}) \\ &= (0.7 \pm 1.64\sqrt{0.7(1-0.7)/100}) = (0.7 \pm 0.075) \\ &= (0.625, 0.775) \end{aligned}$$

(b) O fabricante adoptará o novo tipo de pára-choques se a probabilidade p não for inferior a 0.75 e, por isso, entendeu testar $H_0 : p \geq 0.75$ contra $H_1 : p < 0.75$. Para que níveis de significância o fabricante não deverá adoptar o novo tipo de pára-choques? (1.5)

Teste de hipóteses:
i) Hipóteses: $H_0 : p \geq 0.75$ versus $H_1 : p < 0.75$.
ii) Estatística do teste e valor observado:

$$T = \frac{\bar{X} - 0.75}{\sqrt{0.75(1-0.75)/100}} \underset{p=0.75}{\overset{a}{\sim}} N(0,1)$$

$$t = \frac{0.7 - 0.75}{\sqrt{0.75(1-0.75)/100}} = -1.155$$

iii) Valor-p: $P(T \leq -1.155 | p = 0.75) = 1 - F_{N(0,1)}(1.155) = 1 - 0.8759 = 0.1241$
iv) Conclusão: Rejeita-se H_0 (i.e., o fabricante não deverá adoptar o novo tipo de pára-choques) para $\alpha > 0.1241$ e aceita-se caso contrário. Dado que os níveis de significância usuais são ≤ 0.10, conclui-se que não há evidência para rejeitar o novo tipo de pára-choques.

Grupo IV	5.0 valores

1. (a) Prove que a estatística do teste de ajustamento do qui-quadrado de Pearson satisfaz a (0.5) igualdade

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{k}{n} \left(\sum_{i=1}^{k} O_i^2 \right) - n$$

quando as k classes consideradas são equiprováveis sob a hipótese nula.

Se as k classes do teste de ajustamento são equiprováveis sob a hipótese nula, $E_i = n\frac{1}{k}$ e, portanto,

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(O_i - \frac{n}{k})^2}{\frac{n}{k}} = \frac{k}{n} \sum_{i=1}^{k} \left[O_i^2 - 2O_i\frac{n}{k} + \frac{n^2}{k^2} \right]$$
$$= \frac{k}{n} \sum_{i=1}^{k} O_i^2 - 2\sum_{i=1}^{k} O_i + n = \frac{k}{n} \sum_{i=1}^{k} O_i^2 - n$$

(b) Use o resultado da alínea anterior e os seguintes registos das frequências (o_i) de inícios de 1088 partos de mulheres grávidas em cada um de 6 períodos de 4 horas (2.0)

i	1	2	3	4	5	6
o_i	214	203	153	134	292	92

para averiguar se o início do parto se distribui uniformemente ao longo do dia, ao nível de significância de 5%.

Seja X o período diário de início de parto.
Teste de hipóteses:
i) Hipóteses: $H_0 : X \sim \text{Uniforme}\{1, 2, 3, 4, 5, 6\}$ versus $H_1 : X \nsim \text{Uniforme}\{1, \ldots, 6\}$.
ii) Estatística do teste e valor observado:

$$T = \frac{k}{n} \sum_{i=1}^{k} O_i^2 - n \underset{H_0}{\overset{a}{\sim}} \chi^2_{(k-m-1)} \text{ com } k = 6, m = 0$$

$$t = \frac{6}{1088} \sum_{i=1}^{k} o_i^2 - 1088 = \frac{6 \times 222098}{1088} - 1088 = 136.805$$

iii) Região crítica a 5%: $RC_{5\%} = (b, \infty)$, onde $b = F^{-1}_{\chi^2_{(5)}}(0.95) = 11.07$
iv) Conclusão: Como t está bem dentro da $RC_{5\%}$, há fortes evidências de que o período de início de parto não se distribui uniformemente ao longo do dia, pelo menos para os níveis de significância de 5% ou superiores.

2. Para relacionar o tempo de montagem (x) com o tempo de reparação (Y) das avarias mais frequentes em certa marca de aparelhos de ar condicionado, foram feitas 52 observações, tendo-se registado:

$$\sum_i x_i = 1040; \ \sum_i x_i^2 = 21320; \ \sum_i y_i = 832; \ \sum_i y_i^2 = 26000; \ \sum_i x_i y_i = 18720.$$

Admitindo a validade do modelo de regressão linear simples com as suposições usuais:

(a) Obtenha a estimativa de mínimos quadrados da recta de regressão linear de Y sobre x. (1.0)

Modelo de regressão linear simples: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, onde $\epsilon_i \sim N(0, \sigma^2)$, $i=1,\ldots,n$ independentes. Recta de regressão: $E(Y|x) = \beta_0 + \beta_1 x$.
Estimadores de mínimos quadrados de β_1 e β_0:

$$\hat{\beta}_1 = \frac{\sum_i \sum_i x_i Y_i - n\bar{x}\bar{Y}}{\sum_i x_i^2 - n\bar{x}^2} \quad \text{e} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

cujas estimativas são dadas por

$$\hat{\beta}_1 = \frac{18720 - 52\times 20\times 16}{21320 - 52\times 20^2} = \frac{2080}{520} = 4 \quad \text{e} \quad \hat{\beta}_0 = 16 - 4\times 20 = -64$$

Portanto, a estimativa da recta de regressão é $\hat{E}(Y|x) = -64 + 4x$.

(b) Averigue se há evidência de que o tempo de reparação não depende do tempo de montagem. Calcule o valor-p. (1.5)

Teste de hipóteses:
i) Hipóteses: $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.
ii) Estatística do teste e valor observado:

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{\sum_i x_i^2 - n\bar{x}^2}}} \overset{\beta_1=0}{\sim} t_{50}$$

$$t = \frac{4}{\sqrt{87.36/520}} = 9.76,$$

$\hat{\sigma}^2 = \frac{1}{n-2}[(\sum_i y_i^2 - n\bar{y}^2) - \hat{\beta}_1^2(\sum_i x_i^2 - n\bar{x}^2)] = \frac{1}{50}(26000 - 52\times 16^2 - 4^2\times 520) = 87.36.$
iii) Valor-p: $P(|T| \geq 9.76|H_0) < 0.001$ pois $F^{-1}_{t_{(50)}}(0.995) = 3.496$.
iv) Conclusão: Rejeita-se H_0 para $\alpha > 0.001$, pelo menos. Ou seja, há evidência de que o tempo de reparação depende do tempo de montagem para os níveis de significância usuais ($1\% \leq \alpha \leq 10\%$).