Probabilidades e Estatística

1ºÉpoca/ 2ºTeste — 1ºsemestre – 2002/03

Duração: 3 horas /1 hora e 30 minutos — 18/01/03 – 9 horas

RESOLUÇÃO ABREVIADA

Grupo I

1.

Acontecimento	**Probabilidade**
S=alunos satisfeitos	$P(S) = 1/2$
M=alunos parcialmente satisfeitos	$P(M) = 1/3$
I=alunos insatisfeitos	$P(I) = 1 - P(S) - P(M) = 1 - 1/2 - 1/3 = 1/6$
C=caloiros	
$C\|S$	$P(C\|S) = 0.6$
$C\|M$	$P(C\|M) = 0.4$
$C\|I$	$P(C\|I) = 0.2$

(a) Recorrendo à lei da probabilidade total tem-se

$$\begin{aligned} P(C) &= P(C|S)P(S) + P(C|M)P(M) + P(C|I)P(I) \\ &= 0.6 \times 1/2 + 0.4 \times 1/3 + 0.2 \times 1/6 = 7/15 \simeq 0.467 \end{aligned}$$

(b) Uma vez que $0.467 \simeq P(C) \neq P(C|S) = 0.6$ e de forma equivalente $0.533 \simeq 1 - P(C) = P(\bar{C}) \neq P(\bar{C}|S) = 1 - P(C|S) = 0.4$ pode concluir-se que estar satisfeito com a qualidade da residência depende do facto de o aluno ser ou não caloiro.

2. X=duração (em *horas*) duma componente electrónica, com função densidade de probabilidade

$$f_X(x) = \begin{cases} 100x^{-2}, & x \geq 100 \\ 0, & x < 100 \end{cases}$$

(a) Seja Mo a moda da variável aleatória X. Então Mo deve ser tal que verifica a condição: $f_X(Mo) = \max_x f_X(x)$. Uma vez que a função densidade de probabilidade é decrescente para $x \geq 100$, pode dizer-se que $f_X(100) = \max_{x \geq 100} f_X(x) = \max_x f_X(x)$ *i.e.* a duração modal é de 100h.

A duração mediana, Me, deve ser tal que $P(X \leq Me) = 1/2$. Uma vez que

$$P(X \leq Me) = \int_{-\infty}^{Me} f_X(x)dx = \int_{100}^{Me} \frac{100}{x^2}dx = \left[-\frac{100}{x}\right]_{100}^{Me} = 1 - \frac{100}{Me}$$

tem-se

$$P(X \leq Me) = 1/2 \Leftrightarrow Me = 200$$

logo a duração mediana é de 200 *horas*.

(b)

$$\begin{aligned} P(X<200|X>150) &= \frac{P(150<X<200)}{P(X>150)} \\ &= \frac{P(150<X<200)}{1-P(X\leq 150)} \\ &= \frac{\int_{150}^{200} 100x^{-2}dx}{1-\int_{100}^{150} 100x^{-2}dx} \\ &= \frac{\left[-100x^{-1}\right]_{150}^{200}}{1-\left[-100x^{-1}\right]_{150}^{100}} \\ &= \frac{1/2-1/3}{1-1/3} = 1/4 \end{aligned}$$

(c) Y=número de componentes substituídas (devido a avaria) nas primeiras 150h de funcionamento do equipamento em 30 instaladas.
Admitindo que o tempo de vida de cada componente instalada no equipamento não depende do tempo de vida das restantes, teremos

$$Y \sim Bin(n=30, p=P(X\leq 150)=1/3)$$

Uma vez que a distribuição não está tabelada, $np=10>5$ e $n(1-p)=20>5$ pode recorrer-se à seguinte aproximação

$$Y \overset{a}{\sim} N(\mu=10, \sigma^2=20/3)$$

e com correcção de continuidade obtém-se

$$P(Y\leq 10) \simeq \Phi\left(\frac{10+0.5-10}{\sqrt{20/3}}\right) = \Phi(0.19) = 0.5753$$

Grupo II

X=número de crianças do sexo feminino
Y=número de crianças do sexo masculino

(a)

$$P(X=x) = \sum_y P(X=x, Y=y) = \begin{cases} 0.38, & x=0,1 \\ 0.20, & x=2 \\ 0.04, & x=3 \\ 0, & c.c. \end{cases}$$

O valor esperado de X é

$$E(X) = \sum_{x=0}^{3} xP(X=x) = (0+1)\times 0.38 + 2\times 0.20 + 3\times 0.04 = 0.90$$

(b)

$$\begin{aligned} E(X|Y=1) &= \sum_{x=0}^{3} xP(X=x|Y=1) = \sum_{x=0}^{3} x\frac{P(X=x,Y=1)}{P(Y=1)} \\ &= 0\times\frac{0.10}{0.38} + 1\times\frac{0.17}{0.38} + 2\times\frac{0.11}{0.38} + 3\times\frac{0}{0.38} \\ &\simeq 1.0263 \end{aligned}$$

(c)

$$\begin{aligned} E(X^2) &= \sum_{x=0}^{3} x^2 P(X=x) = (0^2+1^2)\times 0.38 + 2^2 \times 0.20 + 3^2 \times 0.04 = 1.54 \\ Var(X) &= E(X^2) - E(X)^2 = 1.54 - 0.9^2 = 0.73 \end{aligned}$$

Dada a simetria da tabela pode afirmar-se que X e Y são identicamente distribuídas, logo $E(X) = E(Y)$ e $Var(X) = Var(Y)$.

$$E(XY) = \sum_{x=0}^{3}\sum_{y=0}^{3} xyP(X=x, Y=y) = 1\times 1\times 0.17 + 1\times 2\times 0.11 + 2\times 1\times 0.11 = 0.61$$

Logo

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E(XY)-E(X)E(Y)}{\sqrt{Var(X)Var(Y)}} = \frac{0.61-0.9^2}{\sqrt{0.73^2}} \simeq -0.274$$

X e Y são v.a. dependentes já que $Corr(X,Y) \neq 0$. X e Y têm tendência para variar em sentido contrário porque $Corr(X,Y) < 0$. A dependência linear entre elas é fraca pois o coeficiente de correlação (-0.274) está relativamente próximo de zero.

Grupo III

(a) Seja $(X_1, \ldots, X_n)$ uma amostra aleatória e $(x_1, \ldots, x_n)$ uma sua concretização (amostra). A função verosimilhança é definida por

$$\begin{aligned} \mathcal{L}(\lambda|x_1,\ldots,x_n) &= \prod_{i=1}^{n} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \text{ (Independência)} \\ &= \prod_{i=1}^{n} f_X(x_i) \qquad \text{(Ident. Dist.)} \\ &= \prod_{i=1}^{n} \frac{\lambda}{2} e^{-\lambda|x_i|} = 2^{-n}\lambda^n e^{-\lambda\sum_{i=1}^{n}|x_i|}, \ \ \lambda > 0 \end{aligned}$$

e a função log-verosimilhança por

$$l(\lambda|x_1,\ldots,x_n) = \ln\left(\mathcal{L}(\lambda|x_1,\ldots,x_n)\right) = -n\ln 2 + n\ln\lambda - \lambda\sum_{i=1}^{n}|x_i|$$

A estimativa de máxima verosimilhança de λ, $\hat{\lambda}$, maximiza a função log-verosimilhança, *i.e.* deve ser tal que

$$\begin{cases} \left[\dfrac{dl(\lambda|x_1,\ldots,x_n)}{d\lambda}\right]_{\lambda=\hat{\lambda}} = 0 & \text{(ponto de estacionaridade)} \\ \left[\dfrac{d^2 l(\lambda|x_1,\ldots,x_n)}{d\lambda^2}\right]_{\lambda=\hat{\lambda}} < 0 & \text{(ponto de máximo)} \end{cases}$$

Do mesmo modo

$$\begin{cases} \dfrac{n}{\hat{\lambda}} - \sum\limits_{i=1}^{n}|x_i| = 0 \\ -\dfrac{n}{\hat{\lambda}^2} < 0 \qquad \text{(Prop. Verdadeira)} \end{cases}$$

Logo a estimativa de máxima verosimilhança de λ é dada por

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} |x_i|}$$

e o estimador de máxima verosimilhança para λ por

$$EMV(\lambda) = \frac{n}{\sum_{i=1}^{n} |X_i|}$$

Particularizando para a amostra recolhida obtém-se:

$$\hat{\lambda} = \frac{40}{47.5408} \simeq 0.8414$$

(b) Sabendo que $Y = |X| \sim Exp(\lambda)$, a partir da amostra aleatória $(X_1, \ldots, X_n)$ define-se a amostra aleatória $(Y_1, \ldots, Y_n) = (|X_1|, \ldots, |X_n|)$, onde $E(Y_i) = E(Y) = \lambda^{-1} < \infty$ e $Var(Y_i) = Var(Y) = \lambda^{-2} < \infty$, $i = 1, \ldots, n$. De acordo com o Teorema do Limite Central pode afirmar-se que, para n for suficientemente grande,

$$T = \frac{\bar{Y} - E(\bar{Y})}{\sqrt{Var(\bar{Y})}} = \frac{\bar{Y} - E(Y)}{\sqrt{Var(Y)/n}} = \frac{\bar{Y} - \lambda^{-1}}{\sqrt{\frac{1}{n\lambda^2}}} = \left(\lambda\bar{Y} - 1\right)\sqrt{n} \overset{a}{\sim} \mathcal{N}(0,1)$$

(c) *Hipóteses*: $H_0 : \lambda = 1$ *vs* $H_1 : \lambda \neq 1$

Variável fulcral: $T = \left(\lambda\bar{Y} - 1\right)\sqrt{n} \overset{a}{\sim} \mathcal{N}(0,1)$

Estatística do teste: $T_0 = T\,|H_0 = \left(\bar{Y} - 1\right)\sqrt{n} \overset{a}{\sim} \mathcal{N}(0,1)$

Valor observado da estatística do teste: $t_0 = (\frac{47.5408}{40} - 1)\sqrt{40} \simeq 1.19$

Valor-p: $p = P\left(|T_0| \geq 1.19 | H_0\right) \simeq 2\left(1 - \Phi(1.19)\right) = 2(1 - 0.8830) = 0.2340$

Regra de decisão: Rejeitar H_0 para $\alpha \geq 0.2340$ e não rejeitar H_0 para $\alpha < 0.2340$.

Decisão: Não se rejeita H_0 (*i.e.* as observações são consistentes com a hipótese H_0) para os níveis usuais de significância (1%, 5% e 10%).

Grupo IV

X=proventos anuais (em milhares de Euros) das famílias duma comunidade
Y =poupança anual

(a) As estimativas de mínimos quadrados de β_1 e β_0 são iguais a

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{63.7 - 144 \times 3.6/9}{2364 - 144^2/9} = \frac{6.1}{60} = 0.101(6) \simeq 0.1017 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} = \frac{3.6}{9} - \frac{6.1}{60} \times \frac{144}{9} = -1.22(6) \simeq -1.2267
\end{aligned}$$

A recta de regressão linear estimada é:

$$\hat{y}_i = \hat{E}(Y|x_i) = -1.2267 + 0.1017x_i$$

(b) Para determinar o $IC_{95\%}(E(Y|x=x_0))$ recorre-se ao método da v.a. fulcral.

Variável fulcral: $T = \dfrac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\hat{\sigma}^2 \left(\dfrac{1}{n} + \dfrac{(\bar{x} - x_0)^2}{\sum_{i=1}^{9} x_i^2 - n\bar{x}^2} \right)}} \sim t_{(n-2)}$

Quantil de probabilidade: Determina-se a tal que $P(-a \leq T \leq a) = 1 - \alpha$. Como o nível de confiança é $1 - \alpha = 0.95$ tem-se

$$F_{t_{(n-2)}}(a) = 1 - 0.05/2 \Leftrightarrow a = F^{-1}_{t_{(n-2)}}(0.975).$$

Consultando as tabelas de quantis de probabilidade da distribuição $t_{(7)}$ $(n - 2 = 9 - 2 = 7)$ obtém-se $a = 2.365$.

Intervalo de Confiança Aleatório: Partindo de $P(-a \leq T \leq a) = 1 - \alpha$ obtém-se

$$P\left[(\hat{\beta}_0 + \hat{\beta}_1 x_0) - a \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{9} x_i^2 - n\bar{x}^2} \right)} \leq \beta_0 + \beta_1 x_0 \right.$$
$$\left. \leq (\hat{\beta}_0 + \hat{\beta}_1 x_0) + a \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{9} x_i^2 - n\bar{x}^2} \right)} \right] = 1 - \alpha$$

e pode escrever-se

$$ICA_{95\%}(\beta_0 + \beta_1 x_0) = \left[(\hat{\beta}_0 + \hat{\beta}_1 x_0) - a \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{9} x_i^2 - n\bar{x}^2} \right)}, (\hat{\beta}_0 + \hat{\beta}_1 x_0) + a \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{9} x_i^2 - n\bar{x}^2} \right)} \right]$$

Intervalo de Confiança: A estimativa de σ^2 é dada por

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2} \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1^2 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \right) \\ &= \frac{1}{7} \left(2.08 - 3.6^2/9 - (0.1017)^2 \left(2364 - 144^2/9 \right) \right) \simeq 2.8 \times 10^{-3} \end{aligned}$$

e como $x_0 = 20$ tem-se

$$\begin{aligned} IC_{95\%}(\beta_0 + \beta_1 x_0) &\simeq \left[-1.2267 + 0.1017 \times 20 - 2.365 \sqrt{2.8 \times 10^{-3} \left(\frac{1}{9} + \frac{(144/9 - 20)^2}{2364 - 144^2/9} \right)}, \right. \\ &\quad \left. -1.2267 - 0.1017 \times 20 + 2.365 \sqrt{2.8 \times 10^{-3} \left(\frac{1}{9} + \frac{(144/9 - 20)^2}{2364 - 144^2/9} \right)} \right] \\ &\simeq [0.7304, 0.8842] \end{aligned}$$

(c) *Hipóteses*: $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 \neq 0$

Variável fulcral: $T = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}$

Estatística do teste: $T_0 = T\,|H_0 = \dfrac{\hat{\beta}_1 - 0}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}$.

Região de rejeição: Tratando-se de um teste bilateral a região de rejeição é do tipo $(-\infty, -a) \cup (a, +\infty)$.

Determinação do quantil: Determina-se a tal que $P(|T_0| > a|H_0) = \alpha$. Como o nível de significância é $\alpha = 0.01$ tem-se

$$F_{t_{(n-2)}}(a) = 1 - 0.01/2 \Leftrightarrow a = F^{-1}_{t_{(n-2)}}(0.995).$$

Consultando as tabelas dos quantis de probabilidade da distribuição $t_{(7)}$ $(n-2 = 9-2 = 7)$ obtém-se $a = 3.499$.

Regra de decisão: Rejeitar H_0 se $|t_0| \geq 3.499$ e não rejeitar caso contrário, onde t_0 é o valor observado da estatística de teste.

Decisão: $t_0 \simeq \dfrac{0.1017}{\sqrt{\dfrac{2.8 \times 10^{-3}}{2364 - 144^2/9}}} = 14.8874 > 3.499$ logo deve rejeitar-se H_0 (a hipótese de X não explicar parte da variabilidade de Y) ao nível de significância de 1%.

(d) O coeficiente de determinação é igual a

$$\begin{aligned} R^2 &= \frac{\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)} \\ &= \frac{(63.7 - 144 \times 3.6/9)^2}{(2364 - 144^2/9)(2.08 - 3.6^2/9)} = \frac{6.1^2}{60 \times 0.64} \simeq 0.9690. \end{aligned}$$

Pode então afirmar-se que cerca de 96.9% da variabilidade observada de Y é explicada pelo modelo, pelo que a recta estimada se ajusta bem aos dados.