Departamento de Matemática da Universidade de Coimbra

Estatística LEI – LECD

Proposta de resolução do Exame de Recurso de Estatística (LEI+LECD) do ano letivo 24/25

I

1. (a) Consideremos os acontecimentos e as probabilidades:
   - $I = \{\text{computadores infetados}\}$,
   - $V = \{\text{computadores em que o antivírus sinaliza a presença do vírus}\}$,
   - Probabilidade de estar infetado: $P(I) = 0.05$
   - Probabilidade de não estar infetado: $P(\overline{I}) = 0.95$
   - Probabilidade de sinalizar a presença de vírus sabendo que está infetado: $P(V/I) = 0.95$
   - Probabilidade de sinalizar a presença de vírus sabendo que não está infetado: $P(V/\overline{I}) = 0.02$

   Pretende-se calcular $P(V)$. Com efeito

$$\begin{aligned} P(V) &= P(V \cap I) + P(V \cap \overline{I}) = P(V/I)P(I) + P(V/\overline{I})P(\overline{I}) \\ &= 0.95 \times 0.05 + 0.02 \times 0.95 \\ &= 0.0475 + 0.019 = 0.0665 \end{aligned}$$

   (b)

$$\begin{aligned} P(I/V) &= \frac{P(I\cap V)}{P(V)} = \frac{P(V/I)P(I)}{P(V)} \\ &= \frac{0.95\times 0.05}{0.0665} = \frac{0.0475}{0.0665} \approx 0.7143. \end{aligned}$$

   (c) Seja X = "número de computadores infetados entre os 15 escolhidos (em mais de 1500)". Temos $X \sim \mathcal{H}(n, M, B)$, com $n = 15$ e $M > 1500$. Uma vez que se tem $n \leq 0.1M$, vem $X \stackrel{\bullet}{\sim} \mathcal{B}(15, p)$, com $p = \frac{B}{M} = P(I) = 0.05$. Considerando $X' \sim \mathcal{B}(15, 0.05)$, obtemos

$$P(X \geq 2) \approx P(X' \geq 2) = 1 - P(X' = 0) - P(X' = 1)$$

   onde

$$P(X' = 0) = (0.95)^{15} \approx 0.4633 \quad \text{e} \quad P(X' = 1) = C_1^{15} \cdot 0.05 \cdot (0.95)^{14} \approx 0.3653.$$

   Então

$$P(X \geq 2) \approx 1 - 0.4633 - 0.3653 = 0.1714.$$

   (d) Seja Y = "número de novos ficheiros corrompidos por dia num computador infetado". Tem-se que $Y \sim \mathcal{P}(4)$ porque $V(X) = 4$. Então

$$P(Y > 6) = 1 - P(Y \leq 6),$$

   e usando a tabela da função de distribuição de Poisson para $\lambda = 4$, obtemos

$$P(Y > 6) = 1 - 0.8893 = 0.1107.$$

2. (a) Se $X \sim N(8, 1)$, então $U = \frac{X-8}{1} \sim \mathcal{N}(0, 1)$. Assim, obtemos

$$\begin{aligned} P(X > 6) &= P\left(\tfrac{X-8}{1} > \tfrac{6-8}{1}\right) = P\left(U > \tfrac{6-8}{1}\right) = P(U > -2) \\ &= P(U \leq 2) = 0.9772, \end{aligned}$$

   usando a tabela da função distribuição da Lei Normal centrada e reduzida ($\mathcal{N}(0, 1)$).

(b) Temos:

- $X \sim N(8,1)$
- $Y \sim \mathcal{N}(10,\sqrt{5})$
- X e Y são v.a. independentes

Uma vez que

$$P(2X-4<Y)=P(2X-Y<4)$$

consideremos $W=2X-Y$. Pela estabilidade da Lei Normal, podemos escrever

$$\begin{array}{rcl} E(W) & = & 2E(X)-E(Y)=16-10=6 \\ \text{e} & & \\ V(W) & = & (2)^2V(X)+(-1)^2V(Y)=4\cdot 1+5=9 \Rightarrow \sigma_W=3 \end{array}$$

pelo que $W \sim N(6,3)$. Então

$$\begin{array}{rcl} P(2X-4<Y) & = & P(W<4)=P\left(U<\frac{4-6}{3}\right)= \\ & = & P(U<-\frac{2}{3})=P(U>\frac{2}{3})=1-P(U\leq\frac{2}{3})\approx 0.2525, \end{array}$$

usando $U \sim \mathcal{N}(0,1)$, a simetria de U e consultando a respetiva tabela.

3. (a) $P(X \leq 0.3)=\int_{-\infty}^{0.3} f(x)\,dx=\int_{0}^{0.3} 2\,(1-x)\,dx=2\left[x-\frac{x^2}{2}\right]_0^{0.3}=2\left(0.3-\frac{0.09}{2}\right)=0.6-0.09=0.51$

(b) i. Temos

$$E(Y)=E(\theta X+1)=\theta\,E(X)+1=\frac{\theta}{3}+1.$$

Logo,

$$\theta=3(E(Y)-1)$$

e, pelo método dos momentos, um estimador de θ é

$$T_n=3\left(\overline{Y}-1\right),$$

onde $\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$, para uma amostra aleatória $(Y_1,Y_2,\ldots,Y_n)$ da variável Y.

ii. Recordemos a seguinte propriedade da lei normal: Se $Y \sim N(m,\sigma)$ então, para quaisquer reais a e b, a variável aleatória $X=aY+b$ segue a lei $N(a\times m,|a|\sigma)$ (ou, mais simplesmente, se Y segue uma lei normal, qualquer transformação afim de Y também segue uma lei normal). Assim, (voltando a este exercício) se Y seguir uma lei normal então, usando a referida propriedade, podemos que concluir que $X=\frac{Y-1}{\theta}$ também segue uma lei normal. Mas a variável X não segue uma lei normal (porque a sua função densidade de probabilidade não é de uma lei normal). Portanto, a variável Y não pode seguir uma lei normal.
Resolução alternativa: O suporte da variável X é o intervalo $[0,1]$, logo o suporte de $Y=\theta\,X+1$ é o intervalo $[1,\theta+1]$, o que não coincide com o suporte de uma lei normal (que é $\mathbb{R}$). Portanto Y não segue uma lei normal.

iii. Queremos um IC para $m=E(Y)$ com g.c. $\beta=0.95$.
A variável fulcral é

$$Z=\sqrt{51}\,\frac{\overline{Y}-m}{S}\;\dot{\sim}\;N(0,1),$$

porque Y não segue uma lei normal, $n=51>30$ e a variância (ou o desvio padrão) de Y é desconhecida.
Um IC para m, com g.c. $\beta=0.95$ (aproximadamente) é

$$\left]\overline{y}-q_{\frac{1+\beta}{2}}\times\frac{s}{\sqrt{51}}\,,\,\overline{y}+q_{\frac{1+\beta}{2}}\times\frac{s}{\sqrt{51}}\right[,$$

onde $\overline{y} = 1.684$, $s = 0.249$ e $q_{\frac{1+\beta}{2}} = q_{0.975} = 1.96$ (da tabela da lei $N(01)$). Portanto, um IC para $m = E(Y)$ com g.c. de 0.95 aproximadamente é

$$]1.6157, 1.7523[\,.$$

iv. Da alínea anterior, sabemos que, com g.c. 0.95 aproximadamente, temos

$$\begin{array}{rcl} 1.6157 \leq & E(Y) & \leq 1.7523 \\ \Leftrightarrow 0.6157 \leq & E(Y)-1 & \leq 0.7523 \\ \Leftrightarrow \ 1.847 \ \leq & \underbrace{3(E(Y)-1)}_{} & \leq \ 2.257 \\ \Leftrightarrow \ 1.847 \ \leq & \theta & \leq \ 2.257 \end{array}$$

Portanto, um IC que contém θ com g.c. aproximado de 0.95 é

$$]1.847, 2.257[\,.$$

II

1. Uma vez que o valor amostral 102.513 é superior a $q_3 = 94.428$, há que averiguar se este valor é um outlier à direita. Com efeito, qualquer valor amostral x que verifique

$$x > q_3 + 1.5 \times (q_3 - q_1)$$

é um outlier à direita desta amostra. Uma vez que

$$q_3 + 1.5 \times (q_3 - q_1) = 94.428 + 1.5 \times (94.428 - 89.147) = 102.3495 < 102.513$$

prova-se que este valor é um outlier desta amostra.

2. Ter $q_1 = 89.147$ significa que, ordenando a amostra, se encontram aproximadamente 25% dos dados amostrais à esquerda de q_1. (para uma avaliação mais cuidada, ver slide 11, Cap3 – Estatística descritiva). Neste caso, tem-se $q_1 = x_{21:81}$. Ora, o histograma revela menos de 10 valores (em 81) à esquerda de 90, pelo que a vigésima primeira observação (q_1) é superior a 90, logo não pode ser igual a 89.147. Isto incompatibiliza este gráfico com a amostra em estudo. *Outra resolução:* A última classe do histograma termina em 97, o que exclui o valor 102.513, já indicado como valor amostral.

   O Box-plot evidencia uma amplitude entre a terceiro quartil e a mediana ($q_3 - q_2$) muito superior à amplitude entre a mediana e o primeiro quartil ($q_2 - q_1$)(mais do dobro), o que não se verfica nesta amostra, onde estas amplitudes são valores muito próximos ($q_3 - q_2 = 2.65$ e $q_2 - q_1 = 2.631$).

3. Foi realizado o teste de Anderson-Darling com o objetivo de averiguar a compatibilidade desta amostra com a hipótese

$$H_0 : \text{X tem lei normal}$$

   Uma vez que se tem

$$p - valor = 0.3609,$$

   o qual é superior a qualquer nível de significância usualmente considerado $(0.01, \ldots, 0.05)$, não rejeitamos a hipótese H_0. Então, com estes dados, podemos assumir a partir de agora que X tem lei normal.

4. Pretendemos testar as hipóteses

$$H_0 : m = 92 \quad \text{contra} \quad H_1 : m < 92$$

   ao nível de significância 0.05. Uma vez que admitimos que X possui distribuição normal e que nada é dito sobre o desvio padrão de X (da população), relativamente à estatística de teste tem-se:

$$Z = \frac{\overline{X}-92}{S}\sqrt{n} \sim T(n-1) \qquad \text{(admitindo } H_0 \text{ verdadeira).}$$

A região crítica deste teste é

$$RC =]-\infty, -q_{0.95}] =]-\infty, -1.664]$$

onde $q_{0.95} = 1.664$ é o quantil de ordem 0.95 da distribuição $T(80)$. Mais, com esta amostra, obtemos

$$Z_{obs} = \frac{\overline{x} - 92}{s}\sqrt{61} = \frac{91.613 - 92}{4.104}\sqrt{81} = -0.8487.$$

Então, atendendo a que $Z_{obs} \notin RC$, não rejeitamos H_0. Concluímos então que, com estes dados e adotando o nível de significância $\alpha = 0.05$, não temos razões para considerar $m < 92$.

5. O erro de tipo II, que corresponde a não rejeitar H_0 quando esta é falsa.

6. Estes dois outputs são exatamente iguais, exceto no valor de t, o qual corresponde ao valor observado da Estatística de teste, (por nós denotado por Z_{obs}). Concretamente um é negativo e o outro positivo. Como Z_{obs} é negativo, o output associado ao teste da alínea 4 é o da esquerda.

7. Trata-se de um teste para a variância de uma variável aleatória com lei normal, a realizar com $\alpha = 0.025$. Sendo a hipótese alternativa unilateral à direita, a região crítica do teste é: $RC = [q_{1-\alpha}, +\infty[$, onde $q_{1-\alpha} = q_{0.975}$ denota o quantil de ordem 0.975 da lei $\chi^2(80)$. Como $q_{0.975} = 106.6$, então

$$RC = [106.6, +\infty[.$$

III

1. Pretende-se realizar um teste de ajustamento do Qui-Quadrado para verificar as probabilidades propostas pelo modelo mendeliano para cada um dos 4 genótipos indicados. Seja X uma variável aleatória que indica o genótipo de uma planta escolhida ao acaso e considere-se as categorias $A_1 = \{A\}$, $A_2 = \{B\}$, $A_3 = \{C\}$ e $A_4 = \{D\}$. As hipóteses do teste são

   $H_0 : P(X \in A_i) = 0.1 * i, 1 \leq i \leq 4;$ $\qquad H_1 : P(X \in A_i) \neq 0.1 * i$, para algum $i \in \{1, \ldots, 4\}$.

   Calculemos os valores esparados em cada categoria: $e_i = n * P(X \in A_i | sob\ H_0) = 10 * i, 1 \leq i \leq 4$, uma vez que $n = 100$. Desta forma, $e_2 = 20$ e $e_4 = 40$ confirmando-se $\sum_{i=1}^{4} e_i = 100$. Como $e_i > 5, 1 \leq i \leq 4$, não há necessidade de agrupar classes e $k = 4$. Falta calcular na tabela o valor $\frac{(n_3-e_3)^2}{e_3} = \frac{(28-30)^2}{30} = \frac{2}{15}$.

   A estatística de teste será $\sum_{i=1}^{4} \frac{(N_i-e_i)^2}{e_i} \overset{\bullet}{\sim} \chi^2(k-1)$, com $k - 1 = 3$, sob H_0 (pois já verificamos que $e_i > 5, \forall i$).

   A região crítica deste teste é

   $$RC =]q_{0.95}, \infty] =]7.81, +\infty[$$

   onde $q_{0.95} = 7.81$ é o quantil de ordem 0.95 da distribuição $\chi^2(3)$. Mais, com esta amostra, obtemos

   $$Z_{obs} = \sum_{i=1}^{4} \frac{(n_i - e_i)^2}{e_i} = 0.1 + 1.25 + \frac{2}{15} + 0.1 = \frac{19}{12} = 1.583.$$

   Então, atendendo a que $Z_{obs} \notin RC$, aceitamos H_0. Concluímos então que, com estes dados e adotando o nível de significância $\alpha = 0.05$, temos razões para aceitar o modelo mendeliano para as probabilidades dos genótipos.

2. (a) Coeficiente de determinação: $R^2 = 0.9267$
   Interpretação: 92.67% da variação nos valores de Y são explicados pelo modelo de regressão linear.

   (b) O gráfico QQ-plot indica que os quantis observados nos resíduos estão próximos dos quantis teóricos associados a uma lei normal uma vez que os pontos estão próximos da reta, o que nos sugere que a distribuição dos resíduos seja normal.
   Esta hipótese é confirmada através do teste de Shapiro-Wilk cujo p-valor é 0.5738 que é maior

dos que os níveis de significância habituais $(0.01, ..., 0.05)$. Este teste de normalidade é adequado para este caso uma vez que os pontos mais extremos desse gráfico não apresentam um afastamento significativo da reta.
Finalmente, a estimativa para a média dos resíduos ε é zero (assegurado pelo *software* R) e a estimativa para o desvio-padrão de ε é 0.5672 (Residual standard error).

(c) Pretende-se encontrar o valor de x que produza a estimativa $\widehat{y} = 11$ para $Y(x)$. Deste modo, devemos encontrar x tal que $\widehat{y} = \widehat{a} \times x + \widehat{b} = 11$, onde $\widehat{a} = 2.44982$ e $\widehat{b} = 1.29602$ são as estimativas dos parâmetros do modelo de regressão linear fornecidos pelo *software* R. Assim, obtemos $x = 3.96$.