INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

31 de Outubro, 2016 **PRIMEIRO TESTE 2016-17** **Uma resolução possível**

I

Seja X a variável aleatória que conta o número de parcelas onde cada genótipo da casta Tinta Barroca revelou tolerância ao *stress* abiótico. É pedido um teste χ^2 de ajustamento da distribuição de X a uma distribuição Binomial.

1. Uma vez que X conta o número de êxitos em três provas (as 3 parcelas associadas a cada genótipo), o parâmetro m da Binomial, que representa o número total de provas, é naturalmente $m=3$. Quanto ao segundo parâmetro da Binomial, p, que representa a probabilidade (admitida constante) de êxito em cada prova de Bernoulli, no nosso contexto será a probabilidade de haver tolerância em cada parcela. Não é dado no enunciado qualquer valor para p, pelo que será estimado a partir dos dados. Concretamente, e sabendo que o valor esperado duma v.a. X com distribuição $B(m,p)$ é dado pelo produto $E[X]=mp$, podemos estimar p a partir da estimativa de $E[X]$ que é a média amostral $\overline{x}$ do número de parcelas onde houve tolerância. Essa média amostral é dada por $\overline{x}=\frac{(0\times22)+(1\times26)+(2\times17)+(3\times2)}{67}=0.9850746$. Assim, toma-se $\hat{p}=\frac{\overline{x}}{m}=\frac{0.9850746}{3}=0.3283582$.

2. **Hipóteses:** $H_0: X \cap B(m=3,\hat{p}=0.3283582) \quad vs. \quad H_0: X \not\cap B(m=3,\hat{p}=0.3283582)$.

   Estatística do Teste: É a estatística de Pearson, $X^2=\sum_{i=0}^{3}\frac{(O_i-\hat{E}_i)^2}{\hat{E}_i}$, sendo O_i o número de observações correspondentes ao valor $X=i$ e $\hat{E}_i$ os valores esperados ao abrigo da hipótese nula (distribuição Binomial). A distribuição assintótica desta estatística, caso seja verdade H_0, é χ^2_{k-r-1} com $k=4$ (número de categorias para as quais há contagens) e $r=1$ (número de parâmetros que foi necessário especificar para definir H_0). Logo, a distribuição assintótica (cuja validade podemos admitir, tendo em conta o enunciado) será χ^2_2.

   Nível de Significância Vamos escolher $\alpha=\text{P[Erro do tipo I]}=\text{P[Rej.}H_0|H_0 \text{ verdade]}=0.05$.

   Região Crítica: (Unilateral direita) Para um nível de significância $\alpha=0.05$, a regra de rejeição deve ser a de rejeitar H_0 se $\chi^2_{\text{calc}}>\chi^2_{\alpha[k-r-1]}=\chi^2_{0.05(2)}=5.99147$.

   Conclusões Como $X^2_{calc}=1.0894<5.99147$, não se rejeita H_0, pelo que se admite a distribuição Binomial referida no enunciado.

3. Considerando a classe $i=3$, temos que a probabilidade esperada para esse valor de X é, ao abrigo de H_0, dada por $\hat{\pi}_3=P[X=3]=\binom{m=3}{3}\hat{p}^3(1-\hat{p})^{m-3}=\hat{p}^3=0.035403$. Assim, o número esperado de genótipos (de entre o total de $N=67$ observados) para os quais há sempre tolerância ($X=3$) é estimado por $\hat{E}_3=N\times\hat{\pi}_3=67\times0.0354=2.37202$. Esta categoria de contagens contribui com uma parcela de valor $\frac{(O_3-\hat{E}_3)^2}{\hat{E}_3}$ para o valor calculado da estatística do teste, X^2. Tem-se $\frac{(O_3-\hat{E}_3)^2}{\hat{E}_3}=\frac{(2-2.37202)^2}{2.37202}=0.05835$. Rigorosamente falando, o valor estimado que foi calculado ($\hat{E}_3=2.37202$) significa que o critério de Cochran não se verifica, uma vez que mais de 20% das contagens esperadas (pelo menos 25%) são inferiores a 5. No entanto, no enunciado era dito para se admitir a validade da distribuição assintótica, não sendo pedida a discussão do critério de Cochran. Como não se verifica a rejeição de H_0, não faz grande sentido discutir quais as parcelas que mais contribuem para o (pequeno, e não significativo) valor de X^2_{calc}.

II

1. (a) O gráfico da esquerda corresponde a uma relação de tipo hiperbólico entre as variáveis originais x e y. De facto, se admitimos a linearidade entre $y^* = \frac{1}{y}$ e $x^* = x$, temos $\frac{1}{y} = b_0 + b_1\, x$ $\Leftrightarrow y = \frac{1}{b_0 + b_1\, x}$. O gráfico da direita corresponde a admitir que a relação entre as variáveis originais x e y é de tipo potência. De facto, admitir a linearidade entre $y^* = \ln(y)$ e $x^* = \ln(x)$ corresponde a ter:

$$\begin{aligned} \ln(y) = b_0 + b_1\, \ln(x) \quad &\Leftrightarrow \quad \mathrm{e}^{\ln(y)} = \mathrm{e}^{b_0 + b_1 \ln(x)} \\ &\Leftrightarrow \quad y = \underbrace{\mathrm{e}^{b_0}}_{=a} \mathrm{e}^{b_1 \ln(x)} \quad = \quad a\, \mathrm{e}^{ln\, x^{b_1}} \quad = \quad a\, x^{b_1} \ . \end{aligned}$$

   (b) Embora se possa admitir uma tendência linear de fundo nas duas nuvens de pontos, objectivos inferenciais serão melhor atingidos com a transformação linearizante que gerou o gráfico da direita. De facto, no gráfico da esquerda a dispersão dos pontos em torno da tendência linear de fundo parece ter variabilidade que não é homogénea, e vai crescendo à medida que aumenta o número de frutos nos tomateiros. Esta tendência irá reflectir-se na existência dum efeito em forma de funil no gráfico de resíduos contra valores ajustados $\hat{y}_i$, e sugere a violação do pressuposto das variâncias constantes dos erros aleatórios que é parte integrante do modelo de regressão linear. Em contrapartida, o gráfico da direita sugere que a variabilidade dos pontos em torno da recta de regressão é constante, o que estará em consonância com a hipótese de variâncias homogéneas dos erros aleatórios. Por outro lado, na nuvem de pontos da esquerda há relativamente poucos pontos na metade direita do gráfico, o que sugere que esses pontos terão uma influência grande no ajustamento da recta, com elevados valores da distância de Cook. Assim, será mais adequado trabalhar com a transformação linearizante que gerou o gráfico da direita, ou seja, será melhor admitir que a relação entre o peso médio dos frutos dum tomateiro e o número de frutos desse tomateiro, segue uma relação potência (decrescente).

2. (a) O coeficiente de determinação é $R^2 = 0.7603$, o que significa que a regressão linear explica cerca de 76% da variabilidade nos valores observados do log-peso do fruto. Este valor é razoavelmente bom, e é significativamente diferente do valor $\mathcal{R}^2 = 0$ associado ao Modelo Nulo, como se pode verificar através dum teste F de ajustamento global:

   Hipóteses: $H_0 : \mathcal{R}^2 = 0 \quad vs. \quad H_1 : \mathcal{R}^2 > 0.$

   Estatística do Teste: $F = \frac{QMR}{QMRE} = (n-2)\,\frac{R^2}{1-R^2} \ \cap \ F_{(1,n-2)}$, sob H_0.

   Nível de significância: $\alpha = \mathrm{P}[$ Erro do tipo I $] = \mathrm{P}[$ Rej. $H_0 \mid H_0$ verdade $] = 0.05$.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha[1,37]} \ \approx \ 4.10$ (entre os valores tabelados 4.17 e 4.08).

   Conclusões: No enunciado está omisso o valor calculado da estatística F. Usando a segunda das expressões acima indicadas para essa estatística, tem-se $F_{calc} = 37 \times \frac{0.7603}{1-0.7603} = 117.3596 \gg 4.10$. Logo há uma clara rejeição de H_0, i.e., usar a recta de regressão para prever o log-peso do fruto a partir do log-número de frutos no tomateiro é significativamente melhor do que considerar que esse log-peso do fruto tem apenas variação aleatória, não explicada pelo número de frutos na planta.

   (b) Numa regressão linear simples, o coeficiente de determinação é o quadrado do coeficiente de correlação amostral entre o preditor (no nosso caso `log(nfrutos)`, x^*) e a variável resposta (no nosso caso `log(pesofruto)`, y^*). Logo, o coeficiente de correlação amostral $r_{x^*y^*}$ é

uma das raízes quadradas do coeficiente de determinação, que é indicado na listagem: $R^2 = 0.7603$. Tendo em conta que o declive da recta (que tem sempre o mesmo sinal que o coeficiente de correlação) é negativo, a raíz relevante de R^2 é a raíz negativa: $r_{x^*y^*} = -\sqrt{R^2} = -\sqrt{0.7603} = -0.872$.

(c) O declive da recta de regressão, $b_1 = -0.39386$ é a variação média nos log-pesos dos frutos (variável resposta y^* da recta) associada a aumentar em uma unidade o log-número dos frutos (variável preditora x^* na recta). A transformação utilizada corresponde à transformação linearizante dum modelo potência $y = a\, x^{b_1}$ com (como se viu na alínea 1a) $b_1 = -0.39386$. Assim, a relação ajustada corresponde a dizer que o peso médio dos frutos dum tomateiro é proporcional à potência -0.4 (aproximadamente) do número de frutos ou, de forma equivalente, que o peso médio dos frutos dum tomateiro é inversamente proporcional ao número de frutos elevado à potência $2/5$.

(d) As fórmulas dos parâmetros da recta de regressão podem ser usadas para obter os valores pedidos. De facto, para a recta de regressão relacionando $y^* = \ln(y)$ e $x^* = \ln(x)$, temos $b_0 = \overline{y^*} - b_1\overline{x^*}$. Logo, no nosso caso tem-se $6.32931 = 5.3157 - (-0.39386)\overline{x^*}$, pelo que $\overline{x^*} = \frac{6.32931 - 5.3157}{0.39386} = 2.5735$. Por outro lado, o declive da recta ajustada é dado por $b_1 = \frac{cov_{x^*y^*}}{s^2_{x^*}} = r_{x^*y^*}\frac{s_{y^*}}{s_{x^*}}$, pelo que $s^2_{x^*} = r^2_{x^*y^*}\frac{s^2_{y^*}}{b_1^2}$. No nosso caso, tem-se $s^2_{x^*} = 0.7603\frac{0.155667}{(-0.39386)^2} = 0.76295$.

(e) Um intervalo a $(1-\alpha)\times 100\%$ de confiança para β_1 é dado por:

$$\left]\ b_1 - t_{\frac{\alpha}{2}(n-2)}\ \hat{\sigma}_{\hat{\beta}_1}\ ,\ b_1 + t_{\frac{\alpha}{2}(n-2)}\ \hat{\sigma}_{\hat{\beta}_1}\ \right[\ .$$

Sabemos pela listagem no enunciado que $b_1 = -0.39386$ (valor que tem de ser o ponto central do intervalo de confiança), e $\hat{\sigma}_{\hat{\beta}_1} = 0.03636$. Para um intervalo a 95% de confiança, $\alpha = 0.05$, e $t_{0.025(37)} \approx 2.025$ (entre os valores tabelados 2.042 e 2.021). O IC pedido será então $]-0.4675\,,\ -0.3202\,[$. Assim, podemos afirmar com 95% de confiança que o declive da recta populacional é um dos valores deste intervalo. Tendo em conta a resposta na alínea 1a), pode afirmar-se que o peso médio dos frutos de tomateiros é proporcional ao número de frutos elevado a uma potência neste intervalo ou, de forma equivalente, inversamente proporcional ao número de frutos elevado a uma potência no intervalo $]0.3202, 0.4675[$.

(f) Pede-se um intervalo de predição (95%) para um valor de y (`pesofruto`) associado ao valor de `nfrutos` $x = 20$. Com base na recta de regressão entre as variáveis logaritmizadas pode construir-se um intervalo de predição para $y^* = \ln(y)$, que corresponde a um valor logaritmizado $x^* = \ln(20) = 2.995732$. Pelo formulário, sabemos que este intervalo de predição tem extremos: $(b_0 + b_1\, x^*) \pm t_{0.025(n-2)}\ \sqrt{QMRE\ \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x^*})^2}{(n-1)s^2_{x^*}}\right]}$, sendo conhecidos a partir do enunciado os seguintes valores: $b_0 = 6.32931$, $b_1 = -0.39386$, $\sqrt{QMRE} = 0.1958$, $n = 39$, $\overline{x^*} = 2.5735$, $s^2_{x^*} = 0.76295$. Assim, $b_0 + b_1\ \ln(20) = 5.149411$. Já se viu que $t_{0.025(37)} \approx 2.025$. A expressão do erro padrão dá $\sqrt{QMRE\ \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x^*})^2}{(n-1)s^2_{x^*}}\right]} = 0.1988879$. Logo, o intervalo de predição para o log-peso dos frutos dum tomateiro com $x = 20$ frutos é $]\,4.7467\,,\ 5.5522\,[$. Para obter um intervalo para o peso dos frutos (em g) será necessário exponenciar estes extremos, obtendo-se o intervalo de predição $]\ 115.1992\ ,\ 257.7935\ [$. Assim pode afirmar-se que 95% dos tomateiros com 20 frutos terão peso médio dos seus frutos neste intervalo.

III

1. É dado o modelo de regressão linear simples em contexto inferencial.

   (a) Sabemos pelo formulário que a variância do estimador $\hat{\beta}_1$ é $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{(n-1)s_x^2}$. Este valor é estimado por $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{QMRE}{(n-1)s_x^2}$. Sabemos ainda (adaptando o enunciado do Exercício 5d das aulas práticas ao contexto inferencial), que $SQR = \hat{\beta}_1^2(n-1)s_x^2$ e que, numa regressão linear simples, $SQR = QMR$. Logo, tem-se $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{QMRE}{\frac{QMR}{\hat{\beta}_1^2}} = \frac{\hat{\beta}_1^2}{\frac{QMR}{QMRE}} = \frac{\hat{\beta}_1^2}{F}$, onde $F = \frac{QMR}{QMRE}$ é a estatística do teste F de ajustamento global. Assim, o valor calculado de $\sigma^2_{\hat{\beta}_1}$ é $\frac{b_1^2}{F_{calc}}$. **Alternativamente**, e tendo em conta o resultado do Exercício 16, sabemos que o valor calculado da estatística do teste F de ajustamento global é o quadrado do valor da estatística $T = \frac{\hat{\beta}_1 - \beta_{1|H_0}}{\hat{\sigma}_{\hat{\beta}_1}}$ num teste a que $H_0 : \beta_1 = 0$. Mas nesse caso, tem-se $F = T^2 = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2_{\hat{\beta}_1}}$ e, re-arrumando igualdade, obtém-se a expressão indicada no enunciado.

   (b) A expressão do estimador dada no enunciado, $\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$, salienta que $\hat{\beta}_1$ é uma combinação linear das n observações da variável resposta, Y_i. Uma das primeiras conclusões do modelo de regressão linear simples é a de que estas n variáveis aleatórias têm distribuição Normal (mais concretamente, $\mathcal{N}\left(\beta_0 + \beta_1 x_i\,,\,\sigma^2\right)$) e são independentes. Sabe-se ainda que qualquer combinação linear de Normais independentes é ainda Normal, pelo que apenas falta calcular os respectivos parâmetros, ou seja o valor esperado $E[\hat{\beta}_1]$ e a variância $V[\hat{\beta}_1]$. Ora, pelas propriedades dos valores esperados, tem-se:

$$E[\hat{\beta}_1] = E\left[\sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} c_i E[Y_i] = \sum_{i=1}^{n} c_i \underbrace{(\beta_0 + \beta_1 x_i)}_{=E[Y_i]} = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \ .$$

   O enunciado afirma que o segundo somatório tem valor 1. Quanto ao primeiro somatório, facilmente se conclui (tendo em conta a expressão dos coeficientes c_i dada no formulário e ainda o Exercício 3a)) que:

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{(n-1)s_x^2} = \frac{1}{(n-1)s_x^2} \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})}_{=0} = 0 \ .$$

   Logo, $E[\hat{\beta}_1] = \beta_1$. Tendo em conta as propriedades da variância, a independência dos Y_i's e o facto acima referido de $V[Y_i] = \sigma^2$, $\forall\, i$, tem-se:

$$V[\hat{\beta}_1] = V\left[\sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} V[c_i Y_i] = \sum_{i=1}^{n} c_i^2 \underbrace{V[Y_i]}_{=\sigma^2} = \sigma^2 \sum_{i=1}^{n} c_i^2 \ .$$

   Mas

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{(n-1)s_x^2}\right)^2 = \frac{1}{[(n-1)s_x^2]^2} \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})^2}_{=(n-1)s_x^2} = \frac{1}{(n-1)s_x^2} \ .$$

   Logo, $V[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)s_x^2}$, completando assim a demonstração. Este resultado tem a seguinte interpretação intuitiva: caso fossem seleccionadas todas as possíveis amostras aleatórias

de dimensão n (com os n valores x_i fixados pelo experimentador), e para cada uma fosse calculada a correspondente estimativa b_1 do declive da recta, o diagrama de frequências dos valores do declive resultantes seria dado pela curva Gaussiana $\mathcal{N}\left(\beta_1\,,\,\frac{\sigma^2}{(n-1)s_x^2}\right)$. Assinale-se ainda que o estimador $\hat{\beta}_1$ é um estimador centrado (não enviesado) e que a sua variância diminui com o aumento da dimensão da amostra (n) e da variância amostral dos x_i (s_x^2).
NOTA: Esta demonstração foi feita nas aulas teóricas, e está afixada na página *web* da disciplina, na secção dos materiais de apoio relativos às aulas teóricas.

2. Num contexto apenas descritivo, demonstremos dois importantes resultados correspondentes ao coeficiente de determinação duma regressão linear simples.

   (a) Por definição, $R^2 = \frac{SQR}{SQT}$. Tendo em conta que $SQT = (n-1)s_y^2$ e que (Exercício 5d)) $SQR = b_1^2(n-1)s_x^2$, tem-se $R^2 = \frac{b_1^2\, s_x^2}{s_y^2}$. Mas, por definição, $b_1^2 = \left(\frac{cov_{xy}}{s_x^2}\right)^2$. Substituindo, vem $R^2 = \frac{cov_{xy}^2}{s_x^2\, s_y^2} = \left(\frac{cov_{xy}}{s_x\, s_y}\right)^2 = (r_{xy})^2$.

   (b) Os valores ajustados $\hat{y}_i$ são dados por uma mesma transformação linear (afim) dos valores do preditor: $\hat{y}_i = b_0 + b_1\, x_i$. São conhecidas as propriedades destas transformações sobre a covariância e a variância. Assim,

$$r_{y\hat{y}}^2 = \left(\frac{cov_{y\hat{y}}}{s_y s_{\hat{y}}}\right)^2 = \frac{cov_{y,b_0+b_1x}^2}{s_y^2\, s_{b_0+b_1x}^2} = \frac{(b_1 cov_{y,x})^2}{s_y^2\, b_1^2 s_x^2} = \frac{\cancel{b_1^2} cov_{xy}^2}{\cancel{b_1^2} s_x^2\, s_y^2} = r_{xy}^2 = R^2.$$

   Assim, o coeficiente de determinação duma regressão linear simples é também o quadrado do coeficiente de correlação linear entre os valores observados e os valores ajustados de y. Esta propriedade estende-se às regressões lineares múltiplas, embora seja necessário adaptar a justificação.