INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

3 de Novembro, 2017 **PRIMEIRO TESTE 2017-18** **Uma resolução possível**

I

É dada uma tabela de contingências, sendo os factores de classificação as proveniências ($a = 3$ níveis) e os terrenos ($b = 3$ níveis).

1. Não tendo sido fixados os totais marginais de quaisquer das dimensões (linhas ou colunas) da tabela, as probabilidades marginais podem ser estimadas a partir das frequências relativas marginais de linha e coluna. Assim, a probabilidade de proveniência Trás-os-Montes é estimada por $\hat{\pi}_{3.} = \frac{N_{3.}}{N}$, sendo $N = 1262$ o número total de frutos observados e $N_{3.} = 67 + 140 + 245 = 452$ o número de frutos observados provenientes de Trás-os-Montes. Logo, $\hat{\pi}_{3.} = 0.3581616$. De forma análoga, a probabilidade estimada de um fruto observado ser do Terreno 1 é dada por $\hat{\pi}_{.1} = \frac{N_{.1}}{N} = \frac{85+76+67}{1262} = \frac{228}{1262} = 0.1806656$.

2. Não tendo sido fixados os totais marginais de quaisquer das dimensões da tabela, iremos realizar um teste de independência entre estes factores de classificação. Designando por π_{ij} a probabilidade conjunta dum fruto ser da proveniência i e ter sido observado no terreno j, e as respectivas probabilidades marginais por $\pi_{i.}$ e $\pi_{.j}$, tem-se:

   Hipóteses: $H_0 : \pi_{ij} = \pi_{i.} \times \pi_{.j} \;\; \forall\, i, j \quad vs. \quad H_1 : \exists\, i, j$ tal que $\pi_{ij} \neq \pi_{i.} \times \pi_{.j}$.

   Estatística do Teste: A estatística de Pearson, é dada por $X^2 = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$, sendo O_{ij} o número de observações correspondentes à célula (i, j) e $\hat{E}_{ij}$ o valor esperado estimado correspondente ao abrigo da hipótese nula de independência, que é dado por $\hat{E}_{ij} = \frac{N_{i.} \times N_{.j}}{N}$. A distribuição assintótica desta estatística, caso seja verdade H_0, é $\chi^2_{(a-1)(b-1)}$ com $a, b = 3$. Logo, a distribuição assintótica será χ^2_4.

   Nível de Significância De acordo com o enunciado, escolhem-se dois valores de $\alpha = P[\text{ Erro do tipo I }] = P[\text{ Rej.}H_0 | H_0 \text{ verdade }]$: $\alpha = 0.05$ e $\alpha = 0.01$.

   Região Crítica: (Unilateral direita) Para um nível de significância $\alpha = 0.05$, a regra de rejeição deve ser a de rejeitar H_0 se $X^2_{\text{calc}} > \chi^2_{0.05(4)} = 9.488$. Para um nível de significância $\alpha = 0.01$, a regra de rejeição corresponde a rejeitar H_0 se $X^2_{\text{calc}} > \chi^2_{0.01(4)} = 13.277$.

   Conclusões Como $X^2_{calc} = 10.305$, rejeita-se H_0 (a hipótese de independência) ao nível $\alpha = 0.05$, mas não ao nível $\alpha = 0.01$. Tal facto significa que o valor de prova (*p-value*) tem de estar entre estes dois valores, ou seja: $0.01 < p < 0.05$.

   A validade deste teste depende da validade da distribuição assintótica da estatística do teste. O Critério de Cochran afirma que essa distribuição assintótica é admissível se nenhum valor esperado fôr inferior a 1, e não mais de 20% forem inferiores a 5. No nosso contexto, os valores esperados são estimados por $\hat{E}_{ij} = \frac{N_{i.} \times N_{.j}}{N}$. O menor destes valores esperados estimados corresponde à célula da linha e da coluna com menores totais marginais (ou seja, a (i, j) para a qual $N_{i.}$ é a menor soma de linha e $N_{.j}$ é a menor soma de coluna). Basta olhar para a tabela para verificar que a menor soma de coluna corresponde à coluna 1, que já vimos ser $N_{.1} = 228$.

Não é tão evidente qual a linha (proveniência) de menor soma, mas rapidamente se verifica que $N_{1.} = 85 + 137 + 186 = 408$, enquanto $N_{2.} = 76 + 112 + 214 = 402$ (tendo sido visto em cima que $N_{3.} = 452$). Logo, a célula com menor valor de $\hat{E}_{ij}$ é a célula $(i, j) = (2, 1)$, para a qual $\hat{E}_{2,1} = \frac{N_{2.1} \times N_{.1}}{N} = \frac{402 \times 228}{1262} = 72.62758 \gg 5$. Assim, todas as células terão valores esperados estimados muito acima do que o necessário para passar o critério de Cochran.

3. A contribuição da célula $(3, 1)$ para o valor da estatística calculada é dada por $\frac{(O_{3,1} - \hat{E}_{3,1})^2}{\hat{E}_{3,1}}$. Ora, $O_{3,1} = 67$ e $\hat{E}_{3,1} = \frac{N_{3.} \times N_{.1}}{N} = \frac{452 \times 228}{1262} = 81.66086$. Logo, $\frac{(O_{3,1} - \hat{E}_{3,1})^2}{\hat{E}_{3,1}} = \frac{(67 - 81.66086)^2}{81.66086} = 2.632116$.

II

1. Na tabela faltam: (i) o valor da estatística T no teste a $H_0 : \beta_1 = 0$, que é dado por $T_{calc} = \frac{b_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.053667}{0.001884} = 28.48567$; (ii) o valor dos graus de liberdade associados ao QMRE, e que são $n - 2 = 238 - 2 = 236$; (iii) o valor dos segundos graus de liberdade na distribuição F do teste de ajustamento global, que é igualmente $n - 2 = 236$; e (iv) o valor da estatística desse mesmo teste, que pode ser calculada como $F_{calc} = (n-2)\frac{R^2}{1-R^2} = 236 \times \frac{0.7746}{1-0.7746} = 811.0275$. Falta calcular um último valor omisso, o valor de s_x^2, onde x indica o número médio de lançamentos por videira. É possível obtê-lo utilizando a expressão do erro padrão (estimado) do estimador do declive da recta, $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\, s_x^2}}$. O enunciado disponibiliza três valores nesta expressão: $\hat{\sigma}_{\hat{\beta}_1} = 0.001884$, $\sqrt{QMRE} = 0.1203$ e $n = 238$. Assim, $s_x^2 = \frac{QMRE}{(n-1)\, \hat{\sigma}^2_{\hat{\beta}_1}} = \frac{0.1203^2}{237 \times (0.001884)^2} = 17.20367$.

2. Há dois aspectos a referir: (i) a discussão do valor do coeficiente de determinação, $R^2 = 0.7746$, que indica que cerca de 77.5% da variabilidade dos valores observados do peso da lenha à poda é explicada pela regressão (um valor bastante elevado); e (ii) o teste F de ajustamento global. Eis este teste:

   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 = $ P[Erro do tipo I] = P[Rej. H_0 | H_0 verdade] = 0.05.

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha[1,236]}$ que, pelas tabelas, é um valor entre os valores tabelados 3.84 e 3.92.

   Conclusões: No enunciado está omisso o valor calculado da estatística F, mas esse valor foi calculado no primeiro ponto, sendo um valor elevadíssimo: $F_{calc} = 811.0275$. Assim, há uma clara rejeição de H_0 e, em conjunto com o valor bastante elevado de R^2, parece adequado usar a recta de regressão para prever o peso da lenha de poda a partir do número médio de lançamentos na videira. **Nota:** Mesmo sem o valor de F_{calc}, seria possível tirar a conclusão do teste, uma vez que o seu valor de prova (*p-value*) está disponível no enunciado e é indistinguível de zero.

3. A recta de regressão populacional tem equação $y = \beta_0 + \beta_1 x$ e passa na origem se $\beta_0 = 0$. Assim, é pedido um intervalo de confiança para β_0. Sabemos que a respectiva expressão, ao nível de confiança $(1 - \alpha) \times 100\%$, é $\left] \, b_0 - t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} \, , \, b_0 + t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} \, \right[$. Das três quantidades necessárias para calcular este IC, duas estão no enunciado: $b_0 = 0.001227$ e o erro padrão correspondente, $\hat{\sigma}_{\hat{\beta}_0} = 0.013809$. O valor da distribuição t-Student, usando um grau

de confiança de 95%, será $t_{0.025(236)}$. Na tabela apenas estão disponíveis os valores para 120 graus de liberdade (1.97993) e o valor indicado como associado a infinitos graus de liberdade (1.96234 e que corresponde ao valor da distribuição limite da t-Student, ou seja, da Normal reduzida). Usando o valor intermédio 1.97 no cálculo, o intervalo a 95% de confiança para β_0 vem: $]-0.02598\ ,\ 0.02843\ [$. Estes são os valores admissíveis para β_0, que têm as mesmas unidades de medida que as da variável resposta Y, no nosso caso kg. Este intervalo contém o valor zero, pelo que $\beta_0=0$ é um valor admissível. Assim, é admissível considerar que a recta de regressão populacional passa na origem. Repare-se que, nesse caso, a relação entre Y e X é uma relação de proporcionalidade simples: $y = \beta_1\, x$.

4. Pede-se um intervalo de predição (a $(1-\alpha)\times 100\%$) para um valor individual de Y (peso) quando o preditor toma o valor x, intervalo cuja expressão é dada no formulário:

$$\left]\ (b_0+b_1x) - t_{\frac{\alpha}{2};n-2}\cdot\sqrt{QMRE\cdot\left[1+\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\ ,\ (b_0+b_1x) + t_{\frac{\alpha}{2};n-2}\cdot\sqrt{QMRE\cdot\left[1+\frac{1}{n}+\frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]}\ \right[\ .$$

No enunciado são disponibilizados quase todos estes valores: $b_0 = 0.001227$, $b_1 = 0.053667$, $\sqrt{QMRE}=0.1203$, $n=238$, $\overline{x}=6.049$, $s_x^2=17$ (em substituição do verdadeiro valor, que em todo o caso era muito próximo deste). Também já vimos nas alíneas anteriores que $t_{0.025(236)} \approx 1.97$. Usando estes valores, temos, com $x=10$, o intervalo de predição (95%) $]\ 0.300\ ,\ 0.776\ [$. Assim, numa parcela de terreno com um número médio de 10 lançamentos por videira, o peso médio de lenha de poda estaria (95%) entre 0.300 e 0.776 kg.

5. Neste ponto considera-se o modelo linear entre `log(peso)` e `log(Nlancamentos)`.

   (a) A afirmação não é válida, devido à transformação da variável resposta `peso`. O valor $R^2=0.7746$ do modelo inicial é a proporção da variabilidade *dos pesos* observados explicada por essa regressão. O valor $R^2=0.8049$ do modelo agora ajustado diz respeito à proporção da variância dos *log-pesos* explicada pela nova regressão. Não existe uma relação directa entra a variância dos pesos e a variância dos log-pesos que permita fazer a afirmação do enunciado.

   (b) A relação linear entre *os logaritmos* das duas variáveis observadas 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:

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

   Assim, a curva potência ajustada à relação original entre `peso` (y) e `Nlancamentos` (x) é dada por $y = e^{-2.96320}\,x^{1.00085} = 0.05165336\,x^{1.00085}$. Como sabemos das aulas teóricas, este tipo de relação potência corresponde a admitir que ambas as variáveis são funções duma terceira variável t (que, neste contexto se poderia supôr ser o comprimento das videiras, ou o tempo) e que as respectivas taxas de variação relativas são proporcionais, sendo a constante de proporcionalidade dada pelo declive da recta na relação linear entre as variáveis logaritmizadas ou, de forma equivalente, pela potência na relação entre as variáveis originais. Assim, a relação que se admite existir na população entre as taxas de variação relativas das variáveis y e x é $\frac{y'(t)}{y(t)}=\beta_1\frac{x'(t)}{x(t)}$. A relação estimada (uma vez que $b_1=1.00085$) é, aproximadamente, uma relação de igualdade entre as duas taxas de variação relativas.

   (c) Na pergunta 3 deste Grupo, já se viu que o modelo que admite a linearidade entre as variáveis originais `peso` (Y) e `Nlancamentos` (X) pode ser admitido como um modelo de proporcionalidade directa entre Y e X. Uma conclusão análoga num modelo potência corresponde a admitir que na relação populacional $Y = \alpha\, x^{\beta}$ se tem $\beta = 1$. Como já se viu na alínea anterior, isso corresponde a admitir que o declive da recta de regressão relacionando as variáveis logaritmizadas é igual a 1. Assim, pode estudar-se a admissibilidade dessa hipótese através dum teste a β_1:

Hipóteses: $H_0 : \beta_1 = 1 \quad vs. \quad H_1 : \beta_1 \neq 1.$

Estatística do Teste: $T = \frac{\hat{\beta}_1 - \beta_{1|H0}}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-2}$, sob H_0.

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

Região Crítica: (Bilateral) Rejeitar H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}(n-2)} = t_{0.025(236)}$ que, como já se viu, é um valor próximo de 1.97.

Conclusões: Tem-se $T_{calc} = \frac{b_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}}$. No enunciado são dados os valores de $b_1 = 1.00085$ e $\hat{\sigma}_{\hat{\beta}_1} = 0.03208$. Logo, $T_{calc} = \frac{1.00085-1}{0.03208} = 0.02649626$, um valor que está muito longe de pertencer à Região Crítica, pelo que, de forma muito clara, indica a não rejeição de H_0 (ao nível $\alpha = 0.05$ usado).

Assim, também este modelo sugere que é admissível considerar que o peso médio da lenha à poda é (aproximadamente) proporcional ao número de lançamentos. Ambos os modelos ajustam relações aproximadamente iguais a $y = 0.05\, x$, ou seja, que aproximam o peso, em kg, da lenha de poda dividindo por 20 o número de lançamentos da videira.

III

1. Pede-se para fazer com β_0 o que foi feito nas aulas teóricas para o parâmetro β_1, e que se encontra na página *web* da disciplina, na subsecção de materiais de apoio relativos às aulas teóricos com a designação *Demonstrações de resultados teóricos*. Sabemos pelo enunciado que $\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \cap t_{n-2}$. Designando por $t_{\alpha/2(n-2)}$ o valor que, numa distribuição t_{n-2} deixa à sua direita uma região de probabilidade $\alpha/2$, e uma vez que o simétrico desse valor, $-t_{\alpha/2(n-2)}$, será (dada a simetria da distribuição t-Student em torno de zero) o valor que deixa à sua *esquerda* uma área $\alpha/2$, pode-se escrever a seguinte equação:

$$P\left[-t_{\alpha/2(n-2)} \;<\; \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \;<\; t_{\alpha/2(n-2)} \right] \;=\; 1 - \alpha$$

Substituindo a dupla desigualdade por outras duplas desigualdades equivalentes não altera a probabilidade $1 - \alpha$. Vamos efectuar essas substituições com o objectivo de deixar o parâmetro para o qual se pretende construir o intervalo de confiança (β_0) sozinho no meio duma dupla desigualdade. Tem-se (primeiro multiplicando a dupla desigualdade por $\hat{\sigma}_{\hat{\beta}_0}$, depois por -1 e finalmente somando $\hat{\beta}_0$):

$$\begin{array}{lrcccl} & -t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} & < & \hat{\beta}_0 - \beta_0 & < & t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} \\ \Leftrightarrow & t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} & > & \beta_0 - \hat{\beta}_0 & > & -t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} \\ \Leftrightarrow & \hat{\beta}_0 - t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} & < & \beta_0 & < & \hat{\beta}_0 + t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} \; . \end{array}$$

Assim, a probabilidade de o verdadeiro valor da ordenada na origem β_0 da recta populacional estar contido entre os dois extremos indicados é $1 - \alpha$. O intervalo de confiança a $(1 - \alpha) \times 100\%$

para β_0 resulta de substituir os valores dos estimadores $\hat{\beta}_0$ e de $\hat{\sigma}_{\hat{\beta}_0}$ pelas suas estimativas amostrais, obtendo-se a fórmula conhecida:

$$\left] \, b_0 - t_{\alpha/2\,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} \quad , \quad b_0 + t_{\alpha/2\,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0} \, \right[\; .$$

2. Esta pergunta corresponde ao Exercício 19 da Regressão Linear Simples. Sabemos que $\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$. Nesta expressão, os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ são variáveis aleatórias (os seus valores variam ao longo do universo de amostras), enquanto que o valor x fixado para a variável preditora é não aleatório. Tem-se então, usando as propriedades das (co)variâncias envolvendo produtos de constantes e variáveis aleatórias, bem como as expressões das variâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$ e da respectiva covariância (todas disponíveis no formulário):

$$\begin{aligned} V[\hat{\mu}_{Y|x}] &= V[\hat{\beta}_0 + \hat{\beta}_1 x] = V[\hat{\beta}_0] + V[\hat{\beta}_1 x] + 2\,Cov(\hat{\beta}_0, \hat{\beta}_1 x) \\ &= V[\hat{\beta}_0] + x^2 \cdot V[\hat{\beta}_1] + 2x \cdot Cov(\hat{\beta}_0, \hat{\beta}_1) \\ &= \underbrace{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2} \right]}_{=V[\hat{\beta}_0]} + x^2 \cdot \underbrace{\frac{\sigma^2}{(n-1)s_x^2}}_{=V[\hat{\beta}_1]} + 2x \cdot \underbrace{\frac{-\sigma^2 \overline{x}}{(n-1)s_x^2}}_{=Cov[\hat{\beta}_0,\hat{\beta}_1]} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2 + x^2 - 2\overline{x}\,x}{(n-1)\,s_x^2} \right] = \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\,s_x^2} \right] \; . \end{aligned}$$

3. Viu-se nas aulas teóricas que uma relação exponencial entre Y e X corresponde a admitir que y é função de x e que a sua taxa de variação relativa é constante, ou seja, corresponde a admitir a equação diferencial $\frac{y'(x)}{y(x)} = b$. De facto, primitivando ambos os lados da equação em ordem a x, e acrescentando a constante de primitivação do lado direito, obtém-se a equação $\ln(y) = b\,x + C$, ou seja, uma relação linear entre $ln(y)$ e x. Tomando exponenciais para isolar y, resulta $y = e^{bx+C} = e^C \cdot e^{bx}$. Esta equação é do tipo exponencial, como pedido.

4. A covariância amostral cov_{xY} é o numerador do estimador $\hat{\beta}_1$ do declive da recta de regressão. Mais concretamente, tem-se $\hat{\beta}_1 = \frac{cov_{xY}}{s_x^2} \Leftrightarrow cov_{xY} = \hat{\beta}_1\, s_x^2$. Nesta expressão, $\hat{\beta}_1$ é uma variável aleatória e s_x^2 é não aleatória. Assim, a covariância é uma transformação linear (afim) de $\hat{\beta}_1$ (recorde-se que uma transformação linear afim duma variável aleatória X é uma quantidade $a + bX$, onde a e b são constantes). Sabemos que a distribuição de $\hat{\beta}_1$ sob o Modelo Linear é Normal, e que qualquer transformação linear (afim) duma quantidade com distribuição Normal continua a ter distribuição Normal. Assim, está justificada a Normalidade de cov_{xY}. Falta identificar os respectivos parâmetros, ou seja, $E[cov_{xY}]$ e $V[cov_{xY}]$. Usando as propriedades dos valores esperados e variâncias, dadas nas aulas, temos:

$$E[cov_{xY}] = E[\hat{\beta}_1\, s_x^2] = s_x^2\, E[\hat{\beta}_1] = s_x^2\, \beta_1 \; ,$$

já que $\hat{\beta}_1$ é um estimador centrado. Analogamente,

$$V[cov_{xY}] = V[\hat{\beta}_1\, s_x^2] = (s_x^2)^2\, V[\hat{\beta}_1] = (s_x^2)^{\cancel{2}} \frac{\sigma^2}{(n-1)\,\cancel{s_x^2}} = s_x^2\, \frac{\sigma^2}{n-1} \; ,$$

como se pedia para mostrar.