INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2020-21

Resoluções de exercícios de Testes χ^2 para tabelas de contingência

1. No enunciado são indicadas as quatro probabilidades associadas a cada combinação de cor e tipo de superfície, resultantes dos pressupostos genéticos que foram admitidos. Indicando por π_{ij} a probabilidade de se ter a cor i (onde $i=1$ corresponde a amarelo e $i=2$ a verde) e uma superfície de tipo j (onde $j=1$ indica lisa e $j=2$ rugosa), temos $\pi_{11}=9/16$, $\pi_{12}=3/16$, $\pi_{21}=3/16$ e $\pi_{22}=1/16$.

   (a) Uma vez que existem ao todo $N=994$ observações, os valores esperados são, respectivamente, $E_{11}=N\times\pi_{11}=994\times\frac{9}{16}=559.125$, $E_{12}=N\times\pi_{12}=994\times\frac{3}{16}=186.375$, $E_{21}=N\times\pi_{21}=994\times\frac{3}{16}=186.375$ e $E_{22}=N\times\pi_{22}=994\times\frac{1}{16}=62.125$. Resumindo numa única tabela os valores observados e (entre parênteses) esperados ao abrigo dos pressupostos genéticos referidos no enunciado, temos:

	Superfície	
Cor	Lisa	Rugosa
Amarelas	556 (559.125)	184 (186.375)
Verdes	193 (186.375)	61 (62.125)

   Todos os valores esperados são grandes, pelo que não há problemas em admitir que a estatística de Pearson tem distribuição χ^2, neste caso com $ab-1=3$ graus de liberdade. Assim, tem-se:

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=\frac{(556-559.125)^2}{559.125}+\frac{(184-186.375)^2}{186.375}+\frac{(193-186.375)^2}{186.375}+\frac{(61-62.125)^2}{62.125}=0.3036\,.$$

   A região crítica (ao nível $\alpha=0.05$ pedido no enunciado) tem fronteira $\chi^2_{0.05(3)}=7.8147$. Logo, o valor calculado da estatística não pertence à região crítica, pelo que não se rejeita a hipótese nula, isto é, consideram-se admissíveis os pressupostos genéticos de dominância/recessividade e segregação independente das características referidas.

   Para efectuar estes cálculos no R, quando as probabilidades estão completamente especificadas, cria-se um vector de valores observados e outro de *probabilidades* sob H_0, tendo apenas o cuidado de especificar a mesma ordem (por linhas ou por colunas da tabela), quer para os valores observados, quer para as probabilidades. Assim, por exemplo:

```
> Erv.O <- c(556,184,193,61)
> Erv.p <- c(9,3,3,1)/16
> chisq.test(Erv.O, p=Erv.p)
Chi-squared test for given probabilities
data:  Erv.O
X-squared = 0.3036, df = 3, p-value = 0.9594
```

   (b) Agora existem ao todo $N^*=30\times N=30\times 994=29820$ observações (onde o asterisco indica a nova situação desta alínea). Todos os valores esperados são assim 30 vezes maiores do que eram antes, ou seja, $E^*_{ij}=N^*\times\pi_{ij}=30\times N\times\pi_{ij}$, para qualquer i e j. Mas se as proporções observadas em cada célula se mantiveram iguais, é porque os valores observados em cada célula também são 30 vezes maiores do que os observados antes. Assim, $O^*_{ij}=30\times O_{ij}$, para todo o i e j.

O novo valor calculado da estatística de teste é também 30 vezes maior, já que:

$$X^*_{calc} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O^*_{ij} - E^*_{ij})^2}{E^*_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(30\,O_{ij} - 30\,E_{ij})^2}{30\,E_{ij}} = 30 \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 30 \times 0.3036 = 9.108.$$

A região crítica não se alterou, uma vez que os graus de liberdade associados à estatística do teste se mantêm iguais. Mas o valor calculado da estatística alterou-se (é 30 vezes maior) e pertence agora à região crítica para $\alpha = 0.05$, pelo que se rejeita a hipótese nula, isto é, não se consideram admissíveis os pressupostos genéticos de dominância/recessividade e segregação independente das características referidas.

Nota: Todos os valores esperados são maiores do que na alínea anterior, pelo que não há problemas em admitir que a estatística de Pearson tem distribuição χ^2_3.

Para efectuar estes cálculos no R, basta dar o seguinte comando:

```
> chisq.test(Erv.O*30, p=Erv.p)
Chi-squared test for given probabilities
data:  Erv.O * 30
X-squared = 9.108, df = 3, p-value = 0.02789
```

2. Tal como no Exercício anterior, as probabilidades resultantes da teoria genética, associadas a cada combinação de comprimento e cor do pêlo são completamente especificados no enunciado. Os valores esperados para cada uma dessas células resulta assim do produto $E_{ij} = N \times \pi_{ij}$ onde $N = 482$ é o número total de cobaias observadas na segunda geração, $i = 1, 2$ indica o comprimento do pêlo (pela ordem de linha da tabela do enunciado) e $j = 1, 2, 3$ indica a cor do pêlo (pela ordem de coluna da tabela do enunciado), sendo π_{ij} a probabilidade da combinação de comprimento e cor do pêlo referidas. Por exemplo, o número esperado de cobaias de pêlo longo e branco será $E_{23} = 482 \times \frac{1}{16} = 30.125$. Eis a tabela com os valores observados e (entre parênteses) os correspondentes valores esperados ao abrigo da teoria genética (que verificam as condições de Cochran):

	Côr		
Pelo	Creme	Amarelo	Branco
Curto	178 (180.750)	93 (90.375)	89 (90.375)
Longo	62 (60.250)	29 (30.125)	31 (30.125)

A estatística de Pearson tem assim o seguinte valor calculado:

$$X^2_{calc} = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(178 - 180.750)^2}{180.750} + ... + \frac{(31 - 30.125)^2}{30.125} = 0.2573\ .$$

O enunciado não especifica qualquer nível de significância, mas optando por $\alpha = 0.05$, rejeita-se H_0 se $X^2_{calc} > \chi^2_{0.05(5)} = 11.0705$. Uma vez que esta desigualdade não se verifica, não se rejeita H_0, sendo admissível que haja segregação independente da cor e comprimento do pêlo das cobaias. No R:

```
> Cob.O <- c(178,93,89,62,29,31)
> Cob.p <- c(6,3,3,2,1,1)/16
> chisq.test(Cob.O, p=Cob.p)
Chi-squared test for given probabilities
data:  Cob.O
X-squared = 0.2573, df = 5, p-value = 0.9984
```

3. (a) O objectivo é o de saber se se pode admitir que a distribuição pelas três categorias de resultados (morte, calo e enraizamento bem sucedido) são idênticas para os quatro tratamentos utilizados na experiência. Dito de outra maneira, queremos saber se a probabilidade de morte é igual, qualquer que seja o nível do factor tratamento (em cujo caso, pode falar-se apenas em π_{Morte}) e, de forma análoga, se há uma única probabilidade de criar calo (π_{Calo}) qualquer que seja o tratamento, e uma única probabilidade de enraizamento (π_{Enraiz}) qualquer que seja o nível do factor Tratamento. Uma forma de explicitar melhor esta hipótese será considerar que $\pi_{j|i}$ indica a probabilidade de, no tratamento i ($i = 1, 2, 3, 4$) o resultado ser j ($j = 1, 2, 3$, associados respectivamente a $Morte, Calo, Enraizamento$), e escrever:

$$H_0 \; : \; \begin{cases} \pi_{Morte|1} = \pi_{Morte|2} = \pi_{Morte|3} = \pi_{Morte|4} & [= \pi_{Morte} = \pi_{.1}] \\ \pi_{Calo|1} = \pi_{Calo|2} = \pi_{Calo|3} = \pi_{Calo|4} & [= \pi_{Calo} = \pi_{.2}] \\ \pi_{Enraiz|1} = \pi_{Enraiz|2} = \pi_{Enraiz|3} = \pi_{Enraiz|4} & [= \pi_{Enraiz} = \pi_{.3}] \end{cases}$$

A hipótese alternativa H_1 será que pelo menos uma das igualdades acima referidas não é verdadeira. A tabela de contingências tem os totais de linha (número de observações para cada um dos quatro tratamentos) fixado à partida pelo experimentador.
Estamos assim perante um *teste de homogeneidade*. A haver uma distribuição comum pelos três tipos de resultados, as probabilidades associadas a cada possível resultado podem ser estimadas a partir das frequências relativas marginais:

$$\begin{aligned} \hat{\pi}_{Morte} = \hat{\pi}_{.1} &= \frac{N_{.1}}{N} = \frac{121}{240} = 0.50417 \\ \hat{\pi}_{Calo} = \hat{\pi}_{.2} &= \frac{N_{.2}}{N} = \frac{83}{240} = 0.34583 \\ \hat{\pi}_{Enraiz} = \hat{\pi}_{.3} &= \frac{N_{.3}}{N} = \frac{36}{240} = 0.15000 \end{aligned}$$

A ser verdade a hipótese de distribuição homogénea nos quatro tratamentos, o valor esperado para cada categoria é dado por: $\hat{E}_{ij} = N_{i.} \times \hat{\pi}_{.j}$. Como os totais de cada linha são todos iguais (60), os valores esperados estimados de cada resultado também vêm iguais nos quatro tratamentos ($i = 1, 2, 3, 4$):

$$\hat{E}_{i1} = 60 \times 0.5041667 = 30.25 \qquad \hat{E}_{i2} = 60 \times 0.3458333 = 20.75 \qquad \hat{E}_{i3} = 60 \times 0.15 = 9 \; .$$

Eis a tabela de valores observados e esperados estimados (estes últimos entre parênteses):

	Resultado			
Tratamento	Morte	Com calo	Enraizamento	Total
Sem incisão/sem boro	26 (30.25)	18 (20.75)	16 (9)	60
Com incisão/sem boro	32 (30.25)	22 (20.75)	6 (9)	60
Sem incisão/com boro	24 (30.25)	24 (20.75)	12 (9)	60
Com incisão/com boro	39 (30.25)	19 (20.75)	2 (9)	60
Total	121	83	36	240

O cálculo do valor da estatística de Pearson produz:

$$\begin{aligned} \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} &= \frac{(26 - 30.25)^2}{30.25} + \frac{(18 - 20.75)^2}{20.75} + \frac{(16 - 9)^2}{9} + ... + \frac{(2 - 9)^2}{9} \\ &= 0.5971074 + 0.3644578 + 5.444444 + 0.1012397 + ... + 5.444444 \\ &= 18.50593 \end{aligned}$$

Não havendo violação do critério de Cochran, o valor calculado da estatística (18.50593) pode ser comparado com a fronteira duma região crítica unilateral direita numa distribuição χ^2_6. Esse valor fronteira, para um nível de significância $\alpha = 0.05$, é $\chi^2_{0.05(6)} = 12.59159$. No R este valor obtém-se através do comando que pede o quantil 0.95 da distribuição χ^2_6:

```
> qchisq(0.95,6)
[1] 12.59159
```

Uma vez que $\chi^2_{calc} > \chi^2_{0.05(6)}$, rejeita-se H_0, ou seja, rejeita-se (ao nível de significância 0.05) a hipótese de haver homogeneidade na distribuição dos resultados do enraizamento, para os quatro tratamentos.

Estes cálculos podem igualmente ser feitos no R. No Exercício 3 do conjunto de exercícios introdutórios, foi já criada a matriz `estacas` com os valores da tabela de contingência:

```
> estacas
      Morte Calo Enraizamento
sI/sB    26   18           16
cI/sB    32   22            6
sI/cB    24   24           12
cI/cB    39   19            2
```

Basta agora invocar o comando `chisq.test` com o nome da matriz como único argumento. De facto, quando o argumento de entrada no comando `chisq.test` é bidimensional, este comando do `R` parte do pressuposto que se pretende efectuar ou um teste de homogeneidade, ou um teste de independência, para os quais (como se viu nas aulas teóricas) os procedimentos de cálculo são idênticos. Repare-se que os graus de liberdade indicados são iguais a $(a-1)(b-1)$, onde a indica o número de linhas da matriz e b o seu número de colunas. Este valor corresponde aos graus de liberdade nos dois testes referidos.

```
> chisq.test(estacas)
Pearson's Chi-squared test
data:  estacas
X-squared = 18.5059, df = 6, p-value = 0.005085
```

Mais uma vez, o valor de prova $p = 0.005085$ indica que para qualquer nível de significância maior do que esse valor, a conclusão do teste seria a rejeição da hipótese de homogeneidade.

(b) A fim de perceber as causas duma tal rejeição, podemos analisar as parcelas da soma que gera o valor calculado da estatística. As três parcelas de maior valor são a parcela da linha 1, coluna 3 (associada ao enraizamento, no tratamento sem incisão e sem boro), de valor 5.44444; a parcela da linha 4, coluna 3 (enraizamento no tratamento com incisão e com boro), igualmente de valor 5.44444; e a parcela da linha 4, coluna 1 (morte no tratamento com incisão e com boro), de valor 2.530992. Só por si, a soma destas três parcelas já excede a fronteira da região crítica, sendo assim estas combinações de resultados e tratamentos as mais responsáveis pela conclusão de rejeição de H_0. Nos três casos há discrepâncias importantes entre valores esperados e valores observados. No entanto, essas discrepâncias são de sinal diferente. Os enraizamentos observados no tratamento sem incisão, nem boro, são em número muito maior (16) do que o esperado (9). Pelo contrário, os enraizamentos observados no tratamento com incisão e com boro são muito menos (2) do que o esperado (9). Para este último tratamento, as mortes observadas são bastante mais numerosas (39) do que o esperado (30.25). Em suma, pode afirmar-se que a falta de homogeneidade está

sobretudo associada aos dois tratamentos extremos (sem intervenção e com os dois tipos de intervenção), sendo que o enraizamento é mais bem sucedido quando não há qualquer tipo de intervenção nas estacas.

4. (a) Nenhum total marginal foi previamente fixado pelo experimentador: nem o número de observações positivas/negativas , nem o número de genótipos de cada região de proveniência. Assim, as frequências relativas de linhas e colunas estimam probabilidades marginais. Este contexto sugere que se efectue um teste à independência entre os factores "região de proveniência" e "presença/ausência do vírus", baseado na estatística qui-quadrado de Pearson.

   (b) Designe-se por π_{ij} a probabilidade conjunta duma observação recair na célula (i,j); por $\pi_{i.}$ a probabilidade marginal da observação corresponder à região i $(i = 1,2,3,4,5)$; e por $\pi_{.j}$ a probabilidade marginal de ausência $(j = 1)$ ou presença $(j = 2)$ do vírus. Tem-se então:

   Hipóteses: $H_0 : \pi_{ij} = \pi_{i.} \times \pi_{.j}\ ,\ \forall i,j$ [Independência dos factores] vs. $H_1 : \exists\, i,j$ tais que $\pi_{ij} \neq \pi_{i.} \times \pi_{.j}$ [Dependência]

   Estatística do Teste: $\chi^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2_{(a-1)(b-1)}$, sob H_0,

   com $a = 5$, $b = 2$, O_{ij} o número de observações na célula (i,j) e $\hat{E}_{ij} = N\,\hat{\pi}_{ij} = N\,\hat{\pi}_{i.}\,\hat{\pi}_{.j}$ os valores esperados estimados correspondentes, obtidos admitindo a hipótese nula de independência. O cálculo dos valores esperados estimados faz-se a partir das estimativas das probabilidades marginais, que são dadas pelas frequências relativas marginais: $\hat{\pi}_{i.} = \frac{N_{i.}}{N}$ e $\hat{\pi}_{.j} = \frac{N_{.j}}{N}$. A distribuição é apenas assintótica, mas a sua validade pode ser admitida, uma vez que no enunciado é indicado que se verificam as condições de Cochran.

   Nota: Seria, no entanto, fácil verificar as condições de Cochran. O mais pequeno de todos os valores esperados estimados corresponde à linha e coluna de menores totais marginais. Neste caso, trata-se da célula de resultados negativos no Alentejo e Algarve. O número esperado estimado de observações nessa célula é $\hat{E}_{11} = \frac{N_{1.}\times N_{.1}}{N} = \frac{188\times 1482}{3078} = 90.51852$, que é claramente superior a 5.

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

   Região Crítica: (Unilateral direita) Rejeitar H_0 se $\chi^2_{\text{calc}} > \chi^2_{\alpha((a-1)(b-1))} = \chi^2_{0.05(4)} \approx 9.48773$.

   Conclusões: Tendo em conta o valor calculado da estatística do teste, que é dado no enunciado, tem-se $\chi^2_{calc} = 469.1039 \gg 9.48773$. Logo, há uma claríssima rejeição de H_0, pelo que se rejeita a hipótese de independência entre presença do vírus e região de proveniência.

   A conclusão do teste era de esperar, tendo em conta que a nível global, as 3078 observações repartem-se em partes quase iguais entre presenças e ausências, mas há regiões (e.g., Vinhos Verdes) em que mais de dois terços dos genótipos registam a presença do vírus, enquanto que noutras regiões (e.g., Douro) mais de dois terços dos genótipos estão livres do vírus. Não surpreende, pois, a conclusão que a presença/ausência do vírus não é independente da região.

   (c) Para calcular a contribuição da região Douro para a estatística do teste, é necessário calcular as duas parcelas correspondentes, ou seja as parcelas da estatística de teste associadas às

células $(3,1)$ e $(3,2)$. Comecemos por determinar o valor esperado estimado nestas células:

$$\hat{E}_{31} = \frac{N_{3.} \times N_{.1}}{N} = \frac{909 \times 1482}{3078} = 437.6667$$
$$\hat{E}_{32} = \frac{N_{3.} \times N_{.2}}{N} = \frac{909 \times 1596}{3078} = 471.3333$$

Assim, as duas parcelas na expressão da estatística χ^2 que correspondem à região Douro somam:

$$\frac{(O_{31} - \hat{E}_{31})^2}{\hat{E}_{31}} + \frac{(O_{32} - \hat{E}_{32})^2}{\hat{E}_{32}} = 105.9446 + 98.37716 = 204.3218 \ .$$

A região do Douro contribui com $\frac{204.3218}{469.1039} \times 100\% = 43.56\%$ do valor total da estatística do teste, sendo por si só amplamente suficiente para que se rejeite a hipótese nula de independência. Como referido na discussão da alínea anterior, este facto não surpreende, uma vez que a região do Douro tem um comportamento bastante distinto do comportamento global, no que respeita à presença do vírus do enrolamento foliar tipo 3.

5. É apresentada uma tabela de contingência (tabela de contagens de dupla entrada), com contagens relativas ao cruzamento de dois factores: o factor proveniência (factor A, com $a=6$ categorias) e o factor adaptabilidade (factor B, com $b=2$ categorias).

   (a) Nesta tabela foram pré-fixados os totais marginais de linha ($N_{i.} = 150, \ \forall i$), sendo perguntado se a distribuição destas 150 contagens pelas duas categorias do outro factor se pode considerar homogénea. Tendo em conta que apenas existem duas categorias no factor adaptabilidade, as hipóteses deste teste de homogeneidade podem ser escritas apenas em termos da probabilidade de morte em cada linha. Designe-se por π_i a probabilidade duma observação corresponder a uma árvore da proveniência i morrer antes dos dois anos. Tem-se então:

   Hipóteses: $H_0 : \pi_i = \pi \ , \ \forall i$ [Homogeneidade] vs. $H_1 : \exists \ i, j$ tais que $\pi_i \neq \pi_j$ [Heterogeneidade]

   Estatística do Teste: $\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2_{(a-1)(b-1)}$, sob H_0,

   com $a=6$, $b=2$, O_{ij} o número de observações na célula (i,j) e $\hat{E}_{ij}$ os valores esperados estimados correspondentes, obtidos admitindo a hipótese nula de homogeneidade. A distribuição é apenas assintótica, mas a sua validade pode ser admitida caso se verifiquem as condições de Cochran (nenhuma célula ter valor esperado $\hat{E}_{ij}$ inferior a 1 e não mais de 20% das células terem valor esperado inferior a 5). Neste caso não é difícil verificar a validade das condições de Cochran. De facto, aplicando a fórmula de cálculo do número esperado estimado de observações num teste de homogeneidade ao nosso contexto, tem-se: $\hat{E}_{ij} = \frac{N_{i.} \times N_{.j}}{N} = \frac{150 \times N_{.j}}{900} = \frac{N_{.j}}{6}$. Assim, o número esperado estimado de observações é igual para todas as células duma mesma coluna. O menor valor esperado tem de corresponder à coluna com menos observações, ou seja, à primeira coluna ($j=1$), correspondente ao número de pinheiros mortos ao fim de 2 anos. Tem-se, para qualquer proveniência i, $\hat{E}_{i1} = 35.5$. Assim, nenhum valor esperado estimado é inferior a 5, pelo que se verificam as condições de Cochran e é admissível a aproximação assintótica na distribuição da estatística de Pearson.

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

Região Crítica: (Unilateral direita) Rejeitar H_0 se $\chi^2_{\text{calc}} > \chi^2_{\alpha(5\times1)} = \chi^2_{0.05(5)} = 11.070$.

Conclusão: Tendo em conta o valor calculado da estatística do teste, que é dado no enunciado, tem-se $\chi^2_{calc} = 51.2762 \gg 11.070$. Logo, rejeita-se H_0, ou seja, rejeita-se a hipótese da probabilidade de mortalidade ser igual para todas as proveniências.

(b) Já se viu que o valor esperado estimado para qualquer célula na primeira coluna é $\hat{E}_{i1} = 35.5$. Analogamente, todas as células da segunda coluna (nível de adaptabilidade $j=2$, indicando sobrevivência) têm igual valor esperado estimado: $\hat{E}_{i2} = \frac{N_{i.} \times N_{.2}}{N} = \frac{150\times687}{900} = \frac{687}{6} = 114.5$. Logo, a contribuição de qualquer proveniência $i=1,...,6$ para o valor calculado da estatística de Pearson é dado pela soma de duas parcelas do tipo:

$$\frac{(O_{i1} - \hat{E}_{i1})^2}{\hat{E}_{i1}} + \frac{(O_{i2} - \hat{E}_{i2})^2}{\hat{E}_{i2}} = \frac{(O_{i1} - 35.5)^2}{35.5} + \frac{(O_{i2} - 114.5)^2}{114.5}.$$

A identificação de qual a proveniência i que dará o maior valor à soma destas duas parcelas é fácil. Trata-se de identificar a proveniência cujos valores observados sejam, simultâneamento, mais distantes de 35.5 na coluna de pinheiros mortos, e de 114.5 na coluna de pinheiros sobreviventes (a mesma localidade tem de satisfazer estes dois critérios já que a soma, quer dos pinheiros observados, quer esperados, em cada linha tem de ser sempre igual a $N_{i.}=150$). Uma inspecção visual da tabela confirma tratar-se da Turquia, cujos pinheiros registaram uma mortalidade muito mais elevada do que as restantes (quase o dobro do valor esperado estimado) e para a qual o valor da soma destas duas parcelas é:

$$\frac{(67 - 35.5)^2}{35.5} + \frac{(83 - 114.5)^2}{114.5} = 37.01594 \ .$$

Este valor corresponde a mais de 72% do valor calculado da estatística ($X^2_{calc} = 51.2762$), o que confirma tratar-se da proveniência que (de longe) mais contribui para o valor de X^2_{calc} e, por conseguinte, para a rejeição da hipótese nula de homogeneidade.

6. Tem-se uma tabela de contingências de dimensão 4×3, estando as linhas ($i=1,2,3,4$) associadas a variedades e as colunas ($j=1,2,3$) a possíveis resultados da experiência da susceptibilidade aos fungos. O experimentador fixou o número $N_{i.}$ de experiências em cada variedade (sempre $N_{i.} = 600$), pelo que apenas os totais de coluna $N_{.j}$ são livres.

(a) O problema colocado corresponde a um teste de homogeneidade, ou seja, procura-se saber se a probabilidade de cada possível resultado da experiência (não germinados; germinados sem apressório; ou germinados com apressório) é igual, qualquer que seja a variedade, sendo esta hipótese de homogeneidade colocada como hipótese nula. Tendo sido fixados os totais das linhas, não faria sentido usar um teste de independência.

Ao abrigo da hipótese nula de homogeneidade, as probabilidades marginais de coluna são comuns, e serão designadas $\pi_{.j}$. Essas probabilidades são estimadas pelas frequências relativas marginais de coluna, ou seja $\hat{\pi}_{.j} = \frac{N_{.j}}{N}$, e são, respectivamente:

$$\hat{\pi}_{.1} = \frac{N_{.1}}{N} = \frac{1274}{2400} = 0.5308333 \qquad ; \qquad \hat{\pi}_{.2} = \frac{N_{.2}}{N} = \frac{836}{2400} = 0.34833333$$

$$\hat{\pi}_{.3} = \frac{N_{.3}}{N} = \frac{290}{2400} = 0.12083333 \ .$$

Os valores esperados estimados são dados pelo produto destas probabilidades estimadas com os totais de linha, ou seja, $\hat{E}_{ij} = N_{i.} \times \hat{\pi}_{.j} = \frac{N_{i.} \times N_{.j}}{N}$. Como neste problema todos os totais de linha são iguais ($N_{i.} = 600$) cada coluna tem sempre o mesmo número esperado de observações (traduzindo a homogeneidade de distribuições que é a hipótese nula ao abrigo da qual se determinam estes valores esperados estimados). Assim, tem-se, para qualquer linha $i = 1, 2, 3, 4$:

$$\hat{E}_{i1} = 600 \times 0.53083333 = 318.5 \quad ; \quad \hat{E}_{i2} = 600 \times 0.34833333 = 209.0$$

$$\hat{E}_{i3} = 600 \times 0.12083333 = 72.5 \ .$$

(b) **Hipóteses:** Represente-se por $\pi_{j|i}$ a probabilidade do resultado $j = 1, 2, 3$, condicional a se ter a variedade $i = 1, 2, 3, 4$. As hipóteses em confronto são:

Hipótese Nula (H_0, hipótese de homogeneidade): $\begin{cases} \pi_{1|1} = \pi_{1|2} = \pi_{1|3} = \pi_{1|4} \\ \pi_{2|1} = \pi_{2|2} = \pi_{2|3} = \pi_{2|4} \\ \pi_{3|1} = \pi_{3|2} = \pi_{3|3} = \pi_{3|4} \end{cases}$

Hipótese Alternativa (H_1, hipótese de heterogeneidade): pelo menos uma das desigualdades em H_0 não se verifica.

Estatística do Teste: É a estatística de Pearson, $X^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$, sendo $a = 4$, $b = 3$, O_{ij} o número de observações na célula (i, j) e $\hat{E}_{ij}$ os valores esperados ao abrigo da hipótese de homogeneidade. A distribuição assintótica desta estatística, caso seja verdade H_0, é $\chi^2_{(a-1)(b-1)}$ sendo $(a-1)(b-1) = 6$ graus de liberdade. É inteiramente legítimo admitir a validade desta distribuição assintótica, uma vez que o menor dos valores esperados estimados para qualquer célula da tabela é $\hat{E}_{i3} = 72.5$ (para qualquer i), pelo que estamos bem acima do limiar de 5 (e por maioria de razão do limiar 1) referido nas condições de Cochran.

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

Região Crítica: (Unilateral direita) Rejeitar H_0 se $\chi^2_{\text{calc}} > \chi^2_{\alpha[(a-1)(b-1)]} = \chi^2_{0.05(6)} = 12.592$.

Conclusões: Como $\chi^2_{calc} = 259.7168 \gg 12.592$, tem-se uma clara rejeição da hipótese nula, ou seja, conclui-se pela rejeição da hipótese de homogeneidade.

(c) Pede-se o valor, na estatística do teste, da parcela correspondente à variedade Galega ($i = 1$) e resultado "não germinado"($j = 1$). Essa parcela é dada por:

$$\frac{(O_{11} - \hat{E}_{11})^2}{\hat{E}_{11}} = \frac{(197 - 318.5)\ 2}{318.5} = 46.35 \ .$$

Trata-se dum valor elevado, que corresponde a cerca de um quinto do valor calculado da estatística de teste, $\chi^2_{calc} = 259.7168$, e que só por si levaria à rejeição de H_0. Neste caso a hipótese de homogeneidade levaria a esperar um número muito superior de contagens nesta célula do que aquelas que foram efectivamente observadas. A variedade Galega é a única na qual o resultado "não germinado" não é o resultado mais frequente. Mais de dois terços dos esporos desta variedade germinaram.

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

(a) 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 $\hat{\pi}_{.1} = \frac{N_{.1}}{N} = \frac{85+76+67}{1262} = \frac{228}{1262} = 0.1806656$.

(b) 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: $H_0 : \pi_{ij} = \pi_{i.} \times \pi_{.j} \quad \forall\, i, j \quad vs. \quad H_1 : \exists\, i, j$ 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 O enunciado pede dois valores de $\alpha = \text{P[Erro do tipo I]} = \text{P[Rej.}H_0 | H_0 \text{ verdade]}$: $\alpha = 0.05$ 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.

(c) 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, $\frac{(O_{3,1} - \hat{E}_{3,1})^2}{\hat{E}_{3,1}} = \frac{(67 - 81.66086)^2}{81.66086} = 2.632116$.