INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2019-20

Resoluções de exercícios de testes com a estatística de Pearson

1. Há um total de $N = 141$ pintos na segunda geração, sendo dado no enunciado o número de pintos observados com cada uma das três cores de penas. São igualmente dadas no enunciado as probabilidades associadas a cada cor de penas, a serem verdade os pressupostos genéticos referidos. Assim, o número esperado de pintos de penas brancas seria $E_b = N \times \pi_b = 141 \times \frac{1}{4} = 35.25$, o número esperado de pintos de penas pretas igualmente $E_p = N \times \pi_p = 141 \times \frac{1}{4} = 35.25$ e o número esperado de pintos de penas azuis $E_a = N \times \pi_a = 141 \times \frac{2}{4} = 70.50$. Sintetizando num quadro:

	Brancas	Pretas	Azuis
Observados (O_i)	36	32	73
Esperados (E_i)	35.25	35.25	70.50

Os valores esperados satisfazem claramente o critério de Cochran, pelo que pode admitir-se que, a serem correctas as probabilidades acima referidas, a estatística de Pearson tem distribuição χ^2 com $k - 1 = 3 - 1 = 2$ graus de liberdade. O valor calculado da estatística de Pearson é:

$$X^2_{calc} = \frac{(36 - 35.25)^2}{35.25} + \frac{(32 - 35.25)^2}{35.25} + \frac{(73 - 70.50)^2}{70.50} = 0.4042553 \ .$$

Sabemos que a região crítica do teste χ^2 associado à estatística de Pearson é unilateral direita. Neste caso concreto (e ao nível de significância $\alpha = 0.05$), será a região à direita do valor $\chi^2_{0.05(2)} = 5.991465$. Assim, o valor calculado da estatística não pertence à região crítica, pelo que não se rejeita H_0, ou seja, não se rejeita a validade da teoria genética de recessividade e dominância indicada no enunciado.

2. Se o número de cachos por pé segue uma distribuição Poisson com parâmetro $\lambda = 4$, então a probabilidade de existirem x cachos por pé é dada por:

$$P[X = x] \quad = \quad \mathrm{e}^{-\lambda}\,\frac{\lambda^x}{x!} \quad = \quad \mathrm{e}^{-4}\,\frac{4^x}{x!} \quad (x \in \mathbb{N}_0).$$

Nesse caso, a probabilidade de cada um dos valores de 0 a 8, é dada (tendo em conta que, por convenção, $0! = 1$ e $4^0 = 1$) por:

$$\pi_0 = \mathrm{e}^{-4} \approx 0.01832 \qquad \pi_1 = 4\,\mathrm{e}^{-4} \approx 0.07326 \qquad \pi_2 = 8\,\mathrm{e}^{-4} \approx 0.14653$$

$$\pi_3 = \frac{32}{3}\,\mathrm{e}^{-4} \approx 0.19537 \qquad \pi_4 = \frac{32}{3}\,\mathrm{e}^{-4} \approx 0.19537 \qquad \pi_5 = \frac{128}{15}\,\mathrm{e}^{-4} \approx 0.15629$$

$$\pi_6 = \frac{256}{45}\,\mathrm{e}^{-4} \approx 0.10420 \qquad \pi_7 = \frac{1024}{315}\,\mathrm{e}^{-4} \approx 0.05954 \qquad \pi_8 = \frac{512}{315}\,\mathrm{e}^{-4} \approx 0.02977$$

No `R`, estes valores podem ser obtidos pelo comando `dpois`, da seguinte forma:

```
> dpois(0:8, 4)
[1] 0.01831564 0.07326256 0.14652511 0.19536681 0.19536681 0.15629345 0.10419563
[8] 0.05954036 0.02977018
```

A probabilidade da última classe (número de cachos maior que 8), ou seja, $P[X > 8]$, obtém-se somando as probabilidades anteriores e subtraindo a 1, ou seja, $P[X > 8] = 1 - P[X \leq 8]$. Obter as probabilidades cumulativas corresponde a determinar os valores da função distribuição cumulativa duma Poisson. No R, utiliza-se o comando `ppois`:

```
> ppois(8,4)
[1] 0.9786366
```

Assim, a probabilidade de ter oito ou mais observações é:

$$\pi_{>8} \;=\; 1 - \sum_{i=0}^{8} \pi_i \;=\; 1 - 0.9786366 \approx 0.02136 \;.$$

O cálculo dos valores esperados em cada categoria obtem-se multiplicando o número total de pés de videira observados ($N = 200$) pelas respectivas probabilidades de categoria, obtendo-se:

$$E_0 = 3.663 \quad E_1 = 14.653 \quad E_2 = 29.305 \quad E_3 = 39.073 \quad E_4 = 39.073$$

$$E_5 = 31.259 \quad E_6 = 20.839 \quad E_7 = 11.908 \quad E_8 = 5.954 \quad E_{>8} = 4.273 \;.$$

Como apenas dois dos dez valores esperados são inferiores a 5 e nenhum é inferior a 1, considera-se que a situação é adequada para aproximar a distribuição da estatística de Pearson por uma χ^2. Os graus de liberdade desta χ^2 são 9 (porque há $k = 10$ classes e uma restrição, resultante de fixar o número total de observações em $N = 200$). Vamos proceder ao cálculo da estatística de Pearson, calculando as suas dez parcelas:

Cachos	0	1	2	3	4	5	6	7	8	> 8
O_i	2	20	29	47	54	29	14	4	1	0
E_i	3.663	14.653	29.305	39.073	39.073	31.259	20.839	11.908	5.954	4.273
$\frac{(O_i - E_i)^2}{E_i}$	0.755	1.952	0.003	1.608	5.702	0.163	2.245	5.252	4.122	4.273

A estatística de Pearson calculada vem: $X^2_{calc} = 26.07420$. Como numa distribuição χ^2_9 o valor que deixa à direita uma região de probabilidade $\alpha = 0.05$ é $\chi^2_{0.05(9)} = 16.91898$, *rejeita-se a hipótese nula* e considera-se que o número de cachos por pé *não* segue uma distribuição $P(4)$.

Nota: Esta rejeição não invalida que a distribuição possa ser uma Poisson *de parâmetro diferente*.

No `R`, o valor da estatística do teste pode ser calculado usando a instrução `chisq.test`, com dois parâmetros: o vector dos valores observados, e o vector p das probabilidades π_i de recair em cada classe, caso seja verdade a hipótese nula. No nosso caso, o teste efectua-se da seguinte forma:

```
> Ex1.O <- c(2,20,29,47,54,29,14,4,1,0)
> Ex1.O
 [1]  2 20 29 47 54 29 14  4  1  0
> Ex1.p <- c(dpois(0:8, 4), 1-ppois(8,4))
> Ex1.p
 [1] 0.01831564 0.07326256 0.14652511 0.19536681 0.19536681 0.15629345
 [7] 0.10419563 0.05954036 0.02977018 0.02136343
```

```
> chisq.test(Ex1.O, p=Ex1.p)
Chi-squared test for given probabilities
data:  Ex1.O
X-squared = 26.0742, df = 9, p-value = 0.001987

Warning message:
In chisq.test(Ex1.O, p = Ex1.p) : Chi-squared approximation may be incorrect
```

Aviso: O comando `chisq.test` previne sobre dúvidas relativas à validade da aproximação χ^2, seguindo um critério mais rigoroso do que o adoptado nesta disciplina.

O resultado do comando `chisq.test` inclui o valor de prova (*p-value*) associado ao valor calculado da estatística: no nosso caso $p = 0.001987$. Um valor tão baixo significa que, mesmo um teste com nível de significância $\alpha = 0.01$ ou $\alpha = 0.005$ optar-se-ia por rejeitar H_0, em favor da hipótese de outra distribuição.

3. Dados observados:

```
> fungo <- c(15,20,40,18,7)
> fungo
[1] 15 20 40 18  7
```

Uma distribuição Binomial surge normalmente quando se contam o número de êxitos em m *provas de Bernoulli* (provas independentes com resultado dicotómico: êxito/fracasso), havendo probabilidade p de êxito em cada uma das m provas. Se X é Binomial $B(m, p)$, toma valores inteiros entre 0 e m, sendo a probabilidade do valor i dada por:

$$P[X = i] \quad = \quad \binom{m}{i} p^i \, (1-p)^{m-i}$$

No nosso exemplo, $m = 4$. Para calcular as probabilidades esperadas, caso a distribuição seja Binomial, é preciso estimar o segundo parâmetro da Binomial. Como numa $B(m, p)$ o valor esperado é mp, vamos usar o grau médio de infestação na amostra para estimar esse valor esperado:

$$\overline{x} \quad = \quad (0 \times 15 + 1 \times 20 + 2 \times 40 + 3 \times 18 + 4 \times 7)/100 = 1.82$$

O valor estimado de $mp = 4p$ sugere que o segundo parâmetro da Binomial seja estimado por $\hat{p} = \frac{1.82}{4} = 0.455$, i.e., que a distribuição pelas 5 classes tenha probabilidades dadas por uma B(4,0.455). Calculemos essas probabilidades com o auxílio da função `dbinom` do `R`:

```
> dbinom(0:4,4,0.455)
[1] 0.08822385 0.29461910 0.36894960 0.20534810 0.04285935
```

O *número esperado (estimado) de observações em cada classe* será o número total de observações (N=100) vezes as probabilidades estimadas de cada classe:

$\hat{E}_0 = 8.822$ $\hat{E}_1 = 29.462$ $\hat{E}_2 = 36.895$ $\hat{E}_3 = 20.535$ $\hat{E}_4 = 4.286$

Verificam-se as condições de Cochran para admitir a distribuição assintótica da estatística de Pearson. Tendo em conta que há $k = 5$ classes, e duas restrições (o número total de observações

é N e estimou-se o parâmetro p), teremos uma distribuição assintótica χ^2_{5-2}, ou seja, com 3 graus de liberdade. O cálculo da estatística de Pearson dá:

$$X^2_{calc} = \sum_{i=0}^{4} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} = \frac{(15-8.822)^2}{8.822} + \frac{(20-29.462)^2}{29.462} + \frac{(40-36.895)^2}{36.895} + \frac{(18-20.535)^2}{20.535} + \frac{(7-4.286)^2}{4.286} = 9.657347 \ .$$

O valor que define uma região crítica unilateral direita, ao nível $\alpha = 0.05$ é $\chi^2_{0.05(3)} = 7.81473$ (que se obtém no `R` com o comando `qchisq(0.95,3)`. Logo, estamos perante um valor que é significativo, isto é, *rejeita-se a hipótese nula de que a distribuição seja Binomial.*

É possível obter o valor calculado da estatística usando o comando `chisq.test` do `R`:

```
> chisq.test(x=fungo, p=dbinom(0:4,4,0.455))
Chi-squared test for given probabilities
data:  fungo
X-squared = 9.6573, [ df = 4, p-value = 0.04661 ]   <----  [IGNORAR]
```

Atenção: os graus de liberdade indicados pelo comando do `R` são sempre $k-1$, não havendo indicação de que houve a estimação dum parâmetro. A introdução da correcção nos graus de liberdade, como indicado acima, é da nossa responsabilidade:

No caso de se utilizar um nível de significância $\alpha = 0.01$, o limiar da região crítica é $\chi^2_{0.01(3)} = 11.3449$ (o valor no `R` obtém-se com o comando `qchisq(0.99,3)`). Neste caso, o valor calculado da estatística ($X^2 = 9.6573$) está fora dessa região crítica, pelo que não se rejeita a hipótese nula. Assim, a evidência disponível é suficiente para rejeitar H_0 caso estejamos dispostos a aceitar uma probabilidade $\alpha = 0.05$ de cometer o erro de tipo I (rejeitar H_0 sendo esta hipótese verdadeira), mas já não é suficiente para rejeitar H_0 se apenas estivermos dispostos a aceitar cometer esse erro com probabilidade $\alpha = 0.01$.

O facto de a conclusão do teste ser diferente, para os níveis de significância $\alpha = 0.01$ e $\alpha = 0.05$ significa que o valor de prova (*p-value*) associado ao valor calculado da estatística do teste está entre estes dois valores. De facto, por definição (e tendo em conta que se trata dum teste com região crítica unilateral direita), esse valor de prova é dado por

$$p = P\left[\chi^2_3 > X^2_{calc}\right] = P\left[\chi^2_3 > 9.6573\right] \ ,$$

onde χ^2_3 representa uma variável aleatória com a distribuição qui-quadrado, com os $k-1-1 = 3$ graus de liberdade acima indicados. Uma vez que sabemos que o valor 11.3449 deixa à sua direita uma região de probabilidade $\alpha = 0.01$ na referida distribuição, e o valor 7.81473 deixa à direita uma região de probabilidade $\alpha = 0.05$, o valor intermédio 9.6573 tem de deixar à sua direita uma região de probabilidade $0.01 < p < 0.05$, numa distribuição χ^2_3. Apenas com base nas tabelas disponíveis nesta disciplina é possível ser um pouco mais concreto, uma vez que as tabelas indicam também o valor que numa distribuição χ^2_3 deixa à direita uma região de probabilidade $\alpha = 0.025$: o valor 9.34840. Assim, com base apenas nas tabelas pode afirmar-se que o valor de prova se situa algures no intervalo $[0.01, 0.025]$.

No programa `R`, é possível usar o comando `pchisq` para calcular com maior precisão o valor de prova. Tal como os restantes comandos congéneres, começados pela letra p, o comando `pchisq`

devolve o valor da função distribuição cumulativa duma distribuição χ^2, para o valor da variável e os graus de liberdade especificados como argumentos. Concretamente, a probabilidade de uma variável com a distribuição χ^2_3 tomar um valor menor ou igual a $X^2_{calc} = 9.6573$ é obtido através do comando `pchisq(9.6573,3)`. O *p-value* é a correspondente probabilidade de estar *à direita* desse valor, e tendo em conta que a probabilidade total para qualquer distribuição é 1, obtém-se o valor de prova desejado da seguinte forma:

```
> 1-pchisq(9.6573,3)
[1] 0.02171551
```

Confirma-se que o valor de prova se localiza entre 0.01 e 0.025, e não surpreende que esteja mais próximo deste último valor.

4. Pede-se para testar se (hipótese H_0) as contagens seguem uma distribuição de Poisson, isto é, tomam valores em $\mathbb{N}_0$ com probabilidade: $P[X = j] = \mathrm{e}^{-\lambda}\frac{\lambda^j}{j!}$.

   O valor do parâmetro λ da distribuição de Poisson não é especificado na hipótese. Uma vez que esse parâmetro é o valor esperado da variável, vamos de novo calcular o valor médio amostral:

$$\overline{x} \quad = \quad (0\times 18 + 1\times 74 + 2\times 139 + 3\times 70 + 4\times 17 + 5\times 2)/320 \quad = \quad 2$$

   Considera-se este como o valor estimado do parâmetro: $\hat{\lambda} = 2$. Caso seja verdade a hipótese de a distribuição ser $P(2)$, têm-se as seguintes probabilidades para cada categoria:

$$\begin{aligned}
P[X=0] &= \mathrm{e}^{-2}\frac{2^0}{0!} = \mathrm{e}^{-2} \approx 0.1353 \qquad (\text{convenciona}-\text{se que } 0! = 1)\\
P[X=1] &= \mathrm{e}^{-2}\frac{2^1}{1!} = 2\,\mathrm{e}^{-2} \approx 0.2707\\
P[X=2] &= \mathrm{e}^{-2}\frac{2^2}{2!} = 2\,\mathrm{e}^{-2} \approx 0.2707\\
P[X=3] &= \mathrm{e}^{-2}\frac{2^3}{3!} = \frac{4}{3}\,\mathrm{e}^{-2} \approx 0.1804\\
P[X=4] &= \mathrm{e}^{-2}\frac{2^4}{4!} = \frac{2}{3}\,\mathrm{e}^{-2} \approx 0.0902\\
P[X\geq 5] &= 1 - P[X\leq 4] = 1 - \mathrm{e}^{-2}\left[1+2+2+\frac{4}{3}+\frac{2}{3}\right] \approx 0.0527
\end{aligned}$$

   Note-se que a última classe foi considerada como uma classe de 5 *ou mais* valores. De facto, *a soma das probabilidades de todas as classes deve dar* 1. Não se observaram mais do que 5 poros por casca, mas tais valores não são impossíveis. Assim, considera-se que a última classe corresponde ao acontecimento "cinco ou mais poros".

   Logo, os valores esperados (estimados) são obtidos multiplicando estas probabilidades estimadas pelo número total de observações ($N = 320$). Tem-se:

Poros	0	1	2	3	4	≥ 5
$\hat{E}_i$	43.307	86.615	86.615	57.743	28.872	16.849
O_i	18	74	139	70	17	2
$\frac{(O_i-\hat{E}_i)^2}{\hat{E}_i}$	14.789	1.837	31.683	2.602	4.881	13.086

A estatística calculada é $X^2_{calc} = 68.87866$. Este valor observado deve ser comparado com uma distribuição χ^2 cujos graus de liberdade são dados pela diferença entre o número de classes, $k = 6$, e o número de restrições que, tal como no exercício anterior, são duas: foi fixado o número total de observações e foi estimado um parâmetro. Logo, tem-se $6 - 1 - 1 = 4$ graus de liberdade. (Todos os valores esperados são superiores a 5, pelo que não é necessário agrupar classes).

O valor fronteira duma região crítica ao nível $\alpha = 0.05$ é $\chi^2_{0.05(4)} = 9.487729$. Assim, *rejeita-se a hipótese nula de que haja uma distribuição de Poisson.* Esta rejeição é muito enfática, como se pode confirmar calculando o valor de prova (*p-value*) associado ao valor calculado da estatística, que é da ordem de 4×10^{-14}:

```
> 1-pchisq(68.87866, 4)
[1] 3.919087e-14
```

Da análise das parcelas individuais da estatística de Pearson observa-se que as que mais contribuem para a rejeição de H_0, isto é, para o elevado valor da estatística, são as parcelas associadas às contagens 0, 2 e 5. No entanto, as parcelas dos valores extremos resultam de haver bastante menos observações do que seria de esperar à luz da hipótese de distribuição Poisson ($O_i \ll \hat{E}_i$), enquanto que o elevado valor da parcela para a contagem 2 resulta de um muito maior número de observações do que seria de esperar ($O_i \gg \hat{E}_i$). Assim, a distribuição observada é muito mais *concentrada* (de menor dispersão) que a Poisson.

5. Pede-se para testar se a distribuição observada das $N = 100$ contagens é compatível com uma distribuição Binomial Negativa com $p = 0.8$ e $m = 5$. Se essa hipótese fôr válida, têm-se as seguintes probabilidades para cada valor (ou classe de valores, no último caso):

$$
\begin{aligned}
P[X = 0] &= (1-0.8)^0\, 0.8^5 \binom{4}{0} = 1 \cdot 0.8^5 \cdot 1 \approx 0.32768 \\
P[X = 1] &= (1-0.8)^1\, 0.8^5 \binom{5}{1} = 0.2 \cdot 0.8^5 \cdot 5 \approx 0.32768 \\
P[X = 2] &= (1-0.8)^2\, 0.8^5 \binom{6}{2} = 0.04 \cdot 0.8^5\, 15 \approx 0.196608 \\
P[X = 3] &= (1-0.8)^3\, 0.8^5 \binom{7}{3} = 0.008 \cdot 0.8^5\, 35 \approx 0.0917504 \\
P[X \geq 4] &= 1 - P[X \leq 3] = 1 - 0.8^5\, [1 + 1 + 0.6 + 0.28] \approx 0.0562816
\end{aligned}
$$

Nota: No `R`, as probabilidade de cada valor individual j numa Binomial Negativa pode ser obtida pela função `dnbinom`. Para o nosso caso, as probabilidades dos valores de 0 a 3 são dadas por:

```
> dnbinom(0:3,5,0.8)
[1] 0.3276800 0.3276800 0.1966080 0.0917504
```

A probabilidade de um valor menor ou igual a j, numa Binomial Negativa, é dada pela função `pnbinom`. A probabilidade de $X \leq 3$ é dada por:

```
> pnbinom(3,5,0.8)
[1] 0.9437184
```

Assim, o acontecimento complementar $X \geq 4$ tem probabilidade $1 - 0.9437184 = 0.0562816$.

Retomando a resolução, os valores esperados em cada categoria obtêm-se multiplicando as probabilidades por $N = 100$. Temos assim a tabela com valores observados e esperados:

No. de insectos	0	1	2	3	≥ 4
E_i	32.768	32.768	19.661	9.175	5.628
O_i	35	29	21	8	7
$\frac{(O_i - E_i)^2}{E_i}$	0.15203	0.43328	0.09122	0.15049	0.33438

A estatística de Pearson é a soma da última linha, ou seja, $X^2_{calc} = 1.1614$. Este valor deve ser avaliado numa distribuição χ^2 (repare que os valores esperados são todos superiores a 5). Os graus de liberdade desta distribuição são 4 (há $k = 5$ classes e uma única restrição: o número total de observações fixo). Ora, numa χ^2_4, o valor que deixa à direita uma região de probabilidade $\alpha = 0.05$ é $\chi^2_{0.05(4)} = 9.487729$. Logo, é evidente que *não se rejeita a hipótese de a distribuição dos valores pelas classe seguir uma lei Binomial Negativa com* $p = 0.8$ *e* $m = 5$.

Nota: No `R`, aplica-se o comando `chisq.test` aos vectores das observações e das probabilidades de cada possível resultado:

```
> Ex5.O <- c(35,29,21,8,7)
> Ex5.O
[1] 35 29 21  8  7
> Ex5.p <- c( dnbinom(0:3,5,0.8),  1-pnbinom(3,5,0.8))
> Ex5.p
[1] 0.3276800 0.3276800 0.1966080 0.0917504 0.0562816
> chisq.test(Ex5.O,p=Ex5.p)
Chi-squared test for given probabilities
data:  Ex5.O
X-squared = 1.1614, df = 4, p-value = 0.8844
```

6. 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	
Côr	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, pode-se proceder como no caso unidimensional, e criar 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:

```
> Ex6.O <- c(556,184,193,61)
> Ex6.p <- c(9,3,3,1)/16
> chisq.test(Ex6.O, p=Ex6.p)
Chi-squared test for given probabilities
data:  Ex6.O
X-squared = 0.3036, df = 3, p-value = 0.9594
```

Nota: Uma vez que não houve estimação de parâmetros, os graus de liberdade indicados pelo comando do `R` estão correctos.

(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(Ex6.O*30, p=Ex6.p)
Chi-squared test for given probabilities
data:  Ex6.O * 30
X-squared = 9.108, df = 3, p-value = 0.02789
```

7. 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`:

```
> Ex7.O <- c(178,93,89,62,29,31)
> Ex7.p <- c(6,3,3,2,1,1)/16
> chisq.test(Ex7.O, p=Ex7.p)
Chi-squared test for given probabilities
data:  Ex7.O
X-squared = 0.2573, df = 5, p-value = 0.9984
```

8. (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:

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

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.

9. Foram fixados os totais de linha na tabela (isto é, o número de frutos de cada região).

(a) Pede-se um teste de homogeneidade. Neste caso, os valores esperados estimados de cada região i, calibre j deverão ser dados por $\hat{E}_{ij} = N_{i.} \times \hat{\pi}_{.j}$, onde $N_{i.}$ indica o número (fixo) de observações na região i e $\hat{\pi}_{.j}$ indica a probabilidade estimada do calibre j. Estas probabilidades não são pré-especificadas, e serão estimadas pelos totais marginais resultantes de somar cada coluna:

$$\begin{array}{ll} \hat{\pi}_{.1} = \frac{N_{.1}}{N} = \frac{2}{274} \approx 0.0073 & \hat{\pi}_{.2} = \frac{N_{.2}}{N} = \frac{47}{274} \approx 0.1715 \\ \hat{\pi}_{.3} = \frac{N_{.3}}{N} = \frac{82}{274} \approx 0.2993 & \hat{\pi}_{.4} = \frac{N_{.4}}{N} = \frac{72}{274} \approx 0.2628 \\ \hat{\pi}_{.5} = \frac{N_{.5}}{N} = \frac{43}{274} \approx 0.1569 & \hat{\pi}_{.6} = \frac{N_{.6}}{N} = \frac{21}{274} \approx 0.0766 \\ \hat{\pi}_{.7} = \frac{N_{.7}}{N} = \frac{6}{274} \approx 0.0219 & \hat{\pi}_{.8} = \frac{N_{.8}}{N} = \frac{1}{274} \approx 0.0036 \end{array}$$

Multiplicando cada uma destas probabilidades estimadas pelos correspondentes totais de linha ($N_{1.} = 19$, $N_{2.} = 155$ e $N_{3.} = 100$) dá-nos os valores esperados estimados $\hat{E}_{ij}$:

< 40	40-45	45-50	50-55	55-60	60-65	65-70	> 70

```
Bombarral 0.139  3.259  5.686  4.993  2.982  1.456 0.416 0.069
Alentejo  1.131 26.588 46.387 40.730 24.325 11.880 3.394 0.566
Setubal   0.730 17.153 29.927 26.277 15.693  7.664 2.190 0.365
```

No entanto, vários destes valores esperados estimados são inferiores a 1, pelo que se torna necessário agregar classes a fim de poder admitir a distribuição assintótica da estatística de Pearson. A natureza da tabela de contingências justifica que quaisquer agregações digam respeito a colunas inteiras. Da observação dos valores esperados conclui-se que se torna *necessário agregar as duas primeiras colunas, bem como as últimas três colunas.* Assim, obtem-se a seguinte tabela, com os valores observados O_{ij} e (entre parenteses) os valores esperados estimados $\hat{E}_{ij}$ (com $i = 1, 2, 3$ e $j = 1, 2, 3, 4, 5$):

	Calibre				
Região	≤ 45	$45 - 50^+$	$50 - 55^+$	$55 - 60^+$	> 60
Bombarral	0 (3.398)	0 (5.686)	3 (4.993)	4 (2.982)	12 (1.942)
Alentejo	7 (27.719)	37 (46.387)	59 (40.730)	36 (24.325)	16 (15.839)
Setúbal	42 (17.883)	45 (29.927)	10 (26.277)	3 (15.693)	0 (10.219)

Os valores esperados estimados estão (quase) de acordo com o critério de Cochran. Cada uma das 15 parcelas $(O_{ij} - \hat{E}_{ij})^2/\hat{E}_{ij}$ da estatística de Pearson é dada por:

	Calibre				
Região	≤ 45	$45 - 50^+$	$50 - 55^+$	$55 - 60^+$	> 60
Bombarral	3.39781	5.686131	0.7953323	0.3477249	52.10701937
Alentejo	15.48672	1.899529	8.1953392	5.6037372	0.00162804
Setubal	32.52321	7.591641	10.0829278	10.2669190	10.21897810

A soma destas 15 parcelas é $X^2_{calc} = 164.2046$. Este valor deve ser comparado com um quantil duma χ^2_8, já que $(3-1)(5-1) = 8$. Considerando $\alpha = 0.05$, tem-se $\chi^2_{0.05(8)} = 15.50731$ pelo que, de forma clara, rejeita-se a hipótese de uma distribuição homogénea dos frutos pelas cinco classes de calibre, em cada uma das regiões analisadas.

(b) Da análise da tabela com as parcelas individuais da estatística X^2_{calc}, dada na alínea anterior, ressalta que, embora várias classes contribuem de forma importante para a rejeição da hipótese de homogeneidade, são sobretudo os valores associados às classes $(1, 5)$ (calibre > 60 no Bombarral) e $(3, 1)$ (calibres ≤ 45 em Setúbal) que mais contribuem para a rejeição. Nos dois casos, trata-se duma *sobre-representação* das classes, relativamente ao que seria de esperar ao abrigo da hipótese de homogeneidade ($O_{ij} \gg \hat{E}_{ij}$). A inspecção da tabela de valores observadas (quer da original, quer da agregada) ajuda a compreender que se trata dum reflexo do facto de que os frutos do Bombarral tendem a ser de maior dimensão, e os de Setúbal de menor dimensão, do que os das outras proveniências.

Um teste de homogeneidade pode ser efectuado, no `R`, a partir duma matriz de contingências e através do comando `chisq.test`. Admitindo que a matriz de contingências agregada, acima referida, já foi criada, tem-se:

```
> calibre.agreg
          <= 45   45-50 50-55 55-60 > 60
Bombarral     0       0     3     4   12
Alentejo      7      37    59    36   16
```

```
Setubal      42      45    10     3    0
> chisq.test(calibre.agreg)
Pearson's Chi-squared test
data:  calibre.agreg
X-squared = 164.2046, df = 8, p-value < 2.2e-16
Warning message:
In chisq.test(calibre.agreg) : Chi-squared approximation may be incorrect
```

Mais uma vez, o alerta resulta do facto de algumas classes terem valor esperado estimado inferior a 5, um critério mais rigoroso do que aquele que admitimos na disciplina.

10. O *package* `MASS` é um de muitos módulos adicionais que podem ser incorporados numa sessão do R. Desde que os módulos tenham já sido instalados no sistema onde estamos a funcionar (no caso do módulo `MASS` essa instalação é feita aquando da instalação inicial no R), basta utilizar o comando `library` para carregar o módulo para a sessão de trabalho: `library(MASS)`.

    Neste módulo existe a matriz `caith`, descrita no enunciado. Uma vez que apenas se fixou o número de pessoas observadas, e não o número total de observações para cada um dos atributos físicos (côr de cabelo e côr de olhos), não há totais marginais pré-fixados e faz sentido um teste de independência. Os dados e o resultado do teste podem ser obtidos no R da seguinte forma:

```
> caith
       fair red medium dark black
blue    326  38    241  110     3
light   688 116    584  188     4
medium  343  84    909  412    26
dark     98  48    403  681    85

> chisq.test(caith)
Pearson's Chi-squared test
data:  caith
X-squared = 1240.039, df = 12, p-value < 2.2e-16
```

    NOTA: Repare-se que a forma de solicitar ao R um teste de independência é em tudo idêntica à forma de solicitar um teste de homogeneidade. Este facto não surpreende uma vez que, como se viu nas aulas teóricas, a forma de calcular a estatística do teste, a distribuição associada à estatística e a natureza da região crítica correspondente são iguais nos dois casos. Cabe ao utilizador saber distinguir os dois contextos, as respectivas hipóteses e interpretação dos resultados.

    Neste caso, há uma clara rejeição da hipótese nula de independência entre côr de olhos e côr de cabelo, como se pode constatar pelo baixíssimo valor de prova (*p-value*) associado ao valor calculado da estatística (inferior a 2.2×10^{-16}, ou seja, inferior à precisão da máquina e portanto indistinguível de zero). Repare-se que o limiar duma região crítica ao nível de significância $\alpha = 0.05$ seria, neste caso, o valor $\chi^2_{0.05(12)} = 21.02607$, enquanto que a estatística do teste tem valor calculado muito mais elevado: $\chi^2_{calc} = 1240.039$. Este resultado não surpreende, dada a natureza do problema. É natural que algumas côres de olhos e de cabelo apareçam associadas com muito maior frequência, enquanto que outras sejam raras.

    Alternativamente, este exercício pode ser resolvido calculando o valor da estatística de Pearson ao abrigo da hipótese de referência parcela a parcela. Para tal, será necessário começar por

estimar os valores esperados, através da expressão $\hat{E}_{ij} = N \times \hat{\pi}_{i.} \times \hat{\pi}_{.j} = \frac{N_{i.} \cdot N_{.j}}{N}$, onde $N_{i.}$ são as frequências absolutas marginais das linhas $i = 1, 2, 3, 4$; $N_{.j}$ as frequências absolutas marginais das colunas $j = 1, 2, 3, 4, 5$; e N o número total de observações. Estas frequências marginais podem ser calculadas no R com o auxílio do comando `apply` (ou com uma máquina de calcular), sendo, respectivamente para linhas e colunas:

```
> apply(caith,1,sum)
  blue  light medium   dark
   718   1580   1774   1315
```

```
> apply(caith,2,sum)
  fair    red medium   dark  black
  1455    286   2137   1391    118
```

Assim, o valor esperado da célula $(1, 1)$ é $\hat{E}_{11} = \frac{718 \times 1455}{5387} = 193.9280$ (o número total de observações, devolvido pelo comando `sum(caith)`, é 5387). De forma análoga se calculam os restantes valores esperados estimados, obtendo-se a seguinte tabela de $\hat{E}_{ij}$s:

```
           fair      red   medium     dark    black
blue   193.9280 38.11918 284.8275 185.3978 15.72749
light  426.7496 83.88342 626.7793 407.9785 34.60924
medium 479.1479 94.18303 703.7383 458.0720 38.85873
dark   355.1745 69.81437 521.6549 339.5517 28.80453
```

Para criar esta tabela de valores esperados no `R`, pode-se começar por guardar o vector de somas de linha (isto é, o vector dos $N_{i.}$) e o vector de somas de coluna (isto é, o vector dos $N_{.j}$), respectivamente:

```
> caith.sl <- apply(caith,1,sum)
> caith.sc <- apply(caith,2,sum)
```

Depois, utiliza-se a função do `R` `outer` que, dados dois vectores, aplica uma função, neste caso a função produto (“`*`”), a cada possível par de elementos (um de cada vector), criando todos os produtos do tipo $N_{i.} \times N_{.j}$. Finalmente, divide-se estes produtos pela dimensão da amostra (N), que se obtem somando todos os elementos da tabela. Os comandos indicados estão em baixo:

```
> caith.E <- outer(caith.sl , caith.sc, "*")/sum(caith)
> caith.E
           fair      red   medium     dark    black
blue   193.9280 38.11918 284.8275 185.3978 15.72749
light  426.7496 83.88342 626.7793 407.9785 34.60924
medium 479.1479 94.18303 703.7383 458.0720 38.85873
dark   355.1745 69.81437 521.6549 339.5517 28.80453
```

Comparando esta tabela de valores esperados com a tabela dos valores observados, não é difícil de constatar a grande disparidade entre os mesmos. Por exemplo, a combinação louro/olhos azuis tem $O_{11} = 326$ observações, quando pela hipótese de independência seria de esperar um valor bastante mais baixo: $\hat{E}_{11} = 193.9280$. A parcela da estatística de Pearson correspondente a esta célula é $\frac{(O_{11} - \hat{E}_{11})^2}{\hat{E}_{11}} = 89.94587$ o que, só por si, já seria suficiente para colocar χ^2_{calc} na região crítica (de rejeição).

A tabela da totalidade das parcelas $\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$ da estatística de Pearson (arredondadas a quatro casas decimais) é dada por:

```
> round((caith-caith.E)^2/caith.E, d=4)
          fair     red  medium     dark    black
blue    89.9459  0.0004  6.7439  30.6629  10.2997
light  159.9340 12.2965  2.9198 118.6105  27.0715
medium  38.6859  1.1010 59.8694   4.6338   4.2551
dark   186.2147  6.8162 26.9891 343.3555 109.6331
```

Como se pode verificar, várias associações ou combinações raras são, só por si, responsáveis pela rejeição da hipótese de independência.

11. Um teste de independência corresponde, neste caso, à pior das situações possíveis: a classificação no terreno e a classificação com base nas imagens de satélite não terem qualquer correspondência[1]. É de desejar que haja uma claríssima rejeição desta hipótese, dado o contexto do problema. A hipótese de independência gera, como valores esperados estimados em cada célula, os valores $\hat{E}_{ij} = N\,\hat{\pi}_{i.}\,\hat{\pi}_{.j} = \frac{N_{i.}\times N_{.j}}{N}$. Tem-se $N_{1.} = 35$, $N_{2.} = 40$, $N_{3.} = 25$, $N_{.1} = 31$, $N_{.2} = 42$, $N_{.3} = 27$ e $N = 100$. Logo,

$$\begin{array}{lll} \hat{E}_{11} = \frac{35\times 31}{100} = 10.85 & \hat{E}_{12} = \frac{35\times 42}{100} = 14.7 & \hat{E}_{13} = \frac{35\times 27}{100} = 9.45 \\ \hat{E}_{21} = \frac{40\times 31}{100} = 12.4 & \hat{E}_{22} = \frac{40\times 42}{100} = 16.8 & \hat{E}_{23} = \frac{40\times 27}{100} = 10.8 \\ \hat{E}_{31} = \frac{25\times 31}{100} = 7.75 & \hat{E}_{32} = \frac{25\times 42}{100} = 10.5 & \hat{E}_{33} = \frac{25\times 27}{100} = 6.75 \end{array}$$

A estatística de Pearson toma assim o valor:

$$\begin{aligned} X^2_{calc} &= \frac{(16-10.85)^2}{10.85} + \frac{(15-14.7)^2}{14.7} + \frac{(4-9.45)^2}{9.45} + \frac{(15-12.4)^2}{12.4} + \frac{(22-16.8)^2}{16.8} + \frac{(3-10.8)^2}{10.8} \\ &\quad + \frac{(0-7.75)^2}{7.75} + \frac{(5-10.5)^2}{10.5} + \frac{(20-6.75)^2}{6.75} \\ &= 2.4445 + 0.0061 + 3.1431 + 0.5452 + 1.6095 + 5.6333 + 7.7500 + 2.8810 + 26.0093 \\ &= 50.02194 \end{aligned}$$

Nos testes de independência, e como resultado da estimação das distribuições marginais, há $(a-1)(b-1)$ graus de liberdade, onde a designa o número de linhas e b o número de colunas. No nosso caso, são 4 graus de liberdade. Usando $\alpha = 0.05$, tem-se $\chi^2_{0.05(4)} = 9.4877$, pelo que a rejeição da hipótese de independência é clara.

Da análise das parcelas da estatística X^2_{calc} é possível constatar que é sobretudo a última parcela (correspondente à classe $(3,3)$) que é responsável pelo grande valor da estatística, e em menor medida, também as parcelas correspondentes às classes $(3,1)$ e $(2,3)$. Inspeccionando de novo a tabela dos valores observados constata-se que, como seria de esperar, há uma grande concentração de parcelas no canto inferior direito da tabela, sendo que a maioria das parcelas classificadas como sendo de regadio ao abrigo duma técnica, também o são ao abrigo da outra técnica. No

[1] Na realidade, a hipótese "ideal" de interesse seria a hipótese de que cada parcela era classificada da mesma forma pelas duas técnicas, ou seja, corresponderia a dizer que as probabilidades de classificação na classe j através do satélite, *condicionais* à classificação na classe i por inspecção directa, só tomariam dois valores: 1 quando $i = j$ e 0 se $i \neq j$. Para evitar mais casos específicos de aplicação da estatística de Pearson (para mais, com problemas na validade da distribuição assintótica da estatística de teste), procede-se com o teste de independência.

entanto, o mesmo não se pode dizer das parcelas de sequeiro/não cultivadas. Para estas duas classes, a situação em pouco se distingue da que seria de esperar ao abrigo da hipótese de independência. Admitindo que a classificação feita por inspecção directa no terreno é a verdadeira, pode afirmar-se que a classificação por satélite consegue separar parcelas de regadio, mas não consegue distinguir entre culturas de sequeiro e parcelas não cultivadas.

A instrução `chisq.test` também permite, de forma expedita, efectuar no `R` um teste de independência, a partir duma matriz de contingências. Eis como proceder, começando por criar a matriz de contingências e atribuir nomes às suas colunas e linhas:

```
> sat <- matrix(nrow=3,ncol=3,c(16,15,0,15,22,5,4,3,20))
> colnames(sat) <- c("N","S","R")
> rownames(sat) <- c("N","S","R")
> sat
   N  S  R
N 16 15  4
S 15 22  3
R  0  5 20

> chisq.test(sat)
Pearson's Chi-squared test
data:  sat
X-squared = 50.0219, df = 4, p-value = 3.573e-10
```

O facto de, quer um teste de independência, quer um teste de homogeneidade, se obterem da mesma forma no `R` reflecte o facto de que, sendo embora testes diferentes para contextos diferentes, ambos terem a mesma expressão da estatística de Pearson e os mesmos graus de liberdade associados, como se viu nas aulas teóricas.