Capítulo 3 Introdução à Inferência Estatística

Exercícios sobre teste de ajustamento do qui quadrado

Alguns dos exercícios apresentados foram extraídos da UC Estatística e Delineamento e de Murteira, Silva Ribeiro, Andrade e Silva e Pimenta (2008)

Ex1. Galinhas homozigóticas de penas brancas são cruzadas com galos homozigóticos de penas pretas produzindo uma primeira geração em que a conjugação dos dois alelos (penas brancas e penas pretas) produz penas de cor azul. Na segunda geração (F2) e segundo a teoria genética seria de esperar que 1/4 dos pintos tenha penas brancas, 1/4 tenha penas pretas e 1/2 tenha penas azuis. Foi realizada uma experiência como indicado e verificou-se que na segunda geração surgiram 36 pintos de penas brancas, 32 pintos de penas pretas e 73 de penas azuis. Verifique se estes valores observados são compatíveis com a teoria genética. Use um nível de significância $\alpha = 0.05$.

Ex2. Um aspirador é vendido em 5 cores: azul (A), branco (B), castanho (C), encarnado (E) e verde (V). Num estudo de mercado para apreciar a popularidade das diferentes cores, analisou-se uma amostra aleatória de 300 vendas recentes, sendo os resultados obtidos os organizados na seguinte tabela

cores	A	B	C	E	V
nº de vendas	45	50	65	52	88

Poder-se-á dizer que não há preferência por qualquer das cores? Justifique.

Ex3. Num estudo sobre reprodução de porcos observou-se o número de leitões em 200 ninhadas seleccionadas ao acaso:

nº de leitões	6	7	8	9	10	11	12
nº de ninhadas	25	27	31	34	32	29	22

Pretende testar-se a hipótese de o número de leitões por ninhada ter uma distribuição uniforme discreta, em $\{6, \cdots, 12\}$, ao nível de significância de 5%.

Explicite devidamente a estatística de teste e identifique a sua distribuição amostral assintótica e o quantil crítico que está na base da sua decisão.

Ex4. Num processo de controlo de qualidade numa linha de produção de empacotamento de latas de cerveja, foram inspeccionadas 200 *packs* de 6 latas cada e em cada um contou-se o número de latas que não passam o controlo de qualidade. Foram obtidos os seguinte valores:

nº de latas impróprias	0	1	2	3	4	5	6
nº de *packs*	141	48	9	2	0	0	0

Será admissível supor que o n º de latas impróprias siga uma lei Binomial(6,0.05)? Justifique.

Ex5. Pretende-se saber se o número de utilizadores por meia hora de uma dada caixa multibanco segue uma lei de Poisson, de valor médio 2. Recolheu-se uma amostra aleatória de 100 períodos de meia hora, tendo-se obtido os seguintes valores

nº de utentes por meia hora	0	1	2	3	4
frequência absoluta observada	15	25	30	20	10

Poder-se-á aceitar o modelo postulado?

Ex6. Análises estatísticas empíricas levam a admitir que o número de jipes de um novo modelo vendidos mensalmente num stand é bem modelado por uma distribuição de Poisson de parâmetro 2. Com o objectvo de testar essa hipótese recolheu-se uma amostra de 24 meses aquando da saída daquele modelo, tendo-se observado os seguintes resultados:

nº de jipes vendidos	0	1	2	3	4	5	6
nº de meses	4	5	4	5	3	2	1

Com base nesta amostra considera que faz sentido usar o modelo de Poisson de parâmetro 2 para modelar o número de jipes do novo modelo vendidos mensalmente no stand? Justifique.

Ex7. Uma empresa agrícola tem uma estação agronómica experimental onde produz novas variedades de ervilhas. Uma amostra sobre as características das ervilhas resultou em 310 ervilhas amarelas e de casca macia, 109 ervilhas amarelas e de casca dura, 100 ervilhas verdes e de casca macia e 37 ervilhas verdes e de casca dura. Numa experiência semelhante, Mendel, através de um modelo matemático simples, previu que o resultado seria de 56.25% de ervilhas amarelas de casca macia, 18.75% de ervilhas amarelas de casca dura, 18.75% de ervilhas verdes de casca macia e 6.25% de ervilhas verdes de casca dura. Serão os resultados da estação agronómica compatíveis com os resultados de Mendel?

Ex8. Uma determinada cultura encontra-se atacada pelo fungo *F*. Foram seleccionadas aleatoriamente 100 plantas e classificadas quanto à intensidade da infestação. A intensidade da infestação é medida da seguinte forma: a planta é dividida em quatro partes iguais e conta-se o número de partes onde se manifesta a infestação. No quadro figuram o número de plantas classificadas em cada uma das classes de intensidade. Teste se é admissível considerar que uma distribuição Binomial(4, 0.4) é um modelo aceitável para descrever a intensidade do ataque do referido fungo. Utilize o nível de significância 0.05.

grau de ataque	0	1	2	3	4
nº de plantas	15	20	40	18	7

Ex9. Num estudo realizado por uma empresa rodoviária admitiu-se que o número de passageiros que chegam a uma determinada paragem em cada 5 minutos é bem descrito por uma variável com distribuição de Poisson de média 3. Com o objectivo de verificar esta hipótese recolheu-se uma amostra de 25 períodos de 5 minutos, contou-se o número de passageiros que chegam e construi-se o seguinte quadro:

nº de passageiros	0	1	2	3	4	5	6
nº de períodos (5 min)	2	4	6	5	4	3	1

Considera que pode aceitar-se o modelo de Poisson de média 3 para modelar o número de passageiros? Justifique convenientemente.

Ex10. Um padeiro coze, diariamente, três grandes bolos de chocolate e os que não são vendidos, no próprio dia, são oferecidos a uma instituição de caridade. Durante um ano (300 dias úteis) foram observados os seguintes valores:

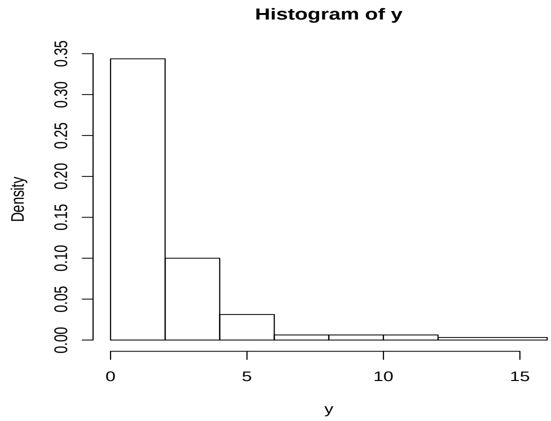
nº de bolos vendidos	0	1	2	3
nº de dias	1	16	55	228

Com base nesta informação, verifique se é estatisticamente aceitável admitir que o número de bolos vendidos diariamente segue uma distribuição Binomial (3, 0.9).

Ex11. Suponha que está a medir os tempos de sobrevivência de uma enzima numa dada solução e se obtiveram os seguintes valores (em horas) que, para facilitar, se apresentam ordenados.

```
0.00  0.04  0.08  0.10  0.11  0.13  0.13  0.15  0.15
0.23  0.23  0.24  0.26  0.29  0.37  0.41  0.42  0.49
0.51  0.53  0.57  0.65  0.67  0.67  0.70  0.75  0.76
0.77  0.82  0.88  0.90  0.98  1.01  1.04  1.06  1.10
1.28  1.32  1.35  1.41  1.42  1.43  1.55  1.58  1.63
1.68  1.72  1.76  1.78  1.80  1.83  1.87  1.88  1.90
1.94  2.07  2.44  2.49  2.58  2.62  2.71  2.77  2.83
2.86  2.90  3.08  3.21  3.39  3.53  3.57  3.90  4.29
4.39  4.57  4.86  5.50  6.49  9.48 11.66 14.12
```

Pretende-se modelar a distribuição do tempo de vida destas enzimas. Numa primeira fase fez-se o histograma das observações, que se apresenta a seguir



Suspeita-se que seja um modelo exponencial de valor médio 3 horas. Será que se tem razão?

Comece por agrupar os dados nas seguintes classes

$$[0,2] \quad]2,4] \quad]4,6] \quad]6,8] \quad]8,10] \quad]10,12] \quad]12,16[$$

e calcule as frequências absolutas.

Teste a hipótese de os tempos de vida destas enzimas estarem de acordo com a suspeita referida.

Nota: Tenha em atenção que no R a exponencial de valor médio β é definida com o parâmetro $\lambda = 1/\beta$.

Tópicos da resolução de alguns exercícios

Ver resolução dos exercícios em R no Anexo abaixo.

Ex1. Dados: $O_i = (36, 32, 73)$, logo $n = 36 + 32 + 73 = 141$ e as probabilidades do modelo a testar são $p_i = (1/4, 1/4, 1/2)$ e número de classes, $k = 3$

Comecemos então por formular as seguintes hipóteses

$\mathbf{H}_0$ **:** $p_1 = 0.25, \quad p_2 = 0.25, \quad p_3 = 0.5$

$\mathbf{H}_{1:}$ pelo menos duas das probabilidades são diferentes do formulado.

$\alpha = 0.05$

	Brancas	Pretas	Azuis
O_i	36	32	73
p_{i0}	0.25	0.25	0.5
$E_i = np_{i0}$	35.25	35.25	70.5

As frequências esperados satisfazem o critério de Cochran pelo que, sob a hipótese nula, pode admitir-se que a estatística de Pearson, $X^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i}$ tem, assintoticamente, distribuição $\chi^2_{(2)}$.

$$X^2_{cal} = \frac{(36-35.25)^2}{35.25} + \frac{(32-35.25)^2}{35.25} + \frac{(73-70.50)^2}{70.5} = 0.4043 \quad \text{e} \quad \chi^2_{0.05(2)} = 5.99.$$

Resposta: $X^2_{cal} < \chi^2_{0.05(2)}$, logo não se rejeita $\mathbf{H}_0$ a um nível de significância de 5%, portanto não há razões para duvidar que os dados estejam de acordo com a teoria genética.

Ex2. Dados: $n = 300$; $O_i = (45, 50, 65, 52, 88)$ e as probabilidades do modelo a testar (todas as cores terem igual preferência) são $p_i = 1/5 \quad \forall i = 1, \ldots, 5$

Comecemos então por formular as seguintes hipóteses

$\mathbf{H}_0$ **:** $p_1 = p_2 = p_3 = p_4 = p_5 = 1/5$

$\mathbf{H}_{1:}$ pelo menos duas das probabilidades são diferentes do formulado.

$\alpha = 0.05$

cores	A	B	C	E	V
O_i	45	50	65	52	88
p_{i0}	0.2	0.2	0.2	0.2	0.2
$E_i = np_{i0}$	60	60	60	60	60

As frequências esperados satisfazem o critério de Cochran pelo que, sob a hipótese nula, pode admitir-se que a estatística de Pearson, $X^2 = \sum_{i=1}^{5} \dfrac{(O_i - E_i)^2}{E_i}$ tem, assintoticamente, distribuição $\chi^2_{(4)}$.

$$X^2_{cal} = \frac{(45-60)^2}{60} + \frac{(50-60)^2}{60} + \dots \frac{(88-60)^2}{60} = 19.9667 \quad \text{e} \quad \chi^2_{0.05(4)} = 9.48773.$$

$X^2_{cal} > \chi^2_{0.05(4)}$, logo rejeita-se $\mathbf{H}_0$ a um nível de significância de 5%, portanto podemos afirmar que a preferência pelas cores não é a mesma, verificando-se que o **verde** colhe mais preferências – verifique que é a cor que contribui com o valor mais elevado para o resultado da Estatística de Teste, portanto o valor observado é muito diferente do valor esperado (sob a hipótese nula).

Ex3. Dados: $n = 200$; $O_i = (25, 27, 31, 34, 32, 29, 22)$ e as probabilidades do modelo a testar (qualquer nº de leitões em $[6, 12]$ ter igual probabilidade) são $p_i = 1/7 \quad \forall i = 1, \dots, 7$

Comecemos então por formular as seguintes hipóteses

$\mathbf{H}_0$ **:** $p_i = 1/7 \quad \forall i = 1, \dots, 7$

$\mathbf{H}_{1:}$ pelo menos duas das probabilidades são diferentes do formulado.

$\alpha = 0.05$

nº de leitões	6	7	8	9	10	11	12
O_i	25	27	31	34	32	29	22
p_{i0}	1/7	1/7	1/7	1/7	1/7	1/7	1/7
$E_i = np_{i0}$	28.57143	28.57143	28.57143	$\cdots$	$\cdots$	$\cdots$	28.57143

As frequências esperados satisfazem o critério de Cochran pelo que, sob a hipótese nula, pode admitir-se que a estatística de Pearson, $X^2 = \sum_{i=1}^{7} \dfrac{(O_i - E_i)^2}{E_i}$ tem, assintoticamente, distribuição $\chi^2_{(6)}$.

$$X^2_{cal} = \frac{(25-28.57143)^2}{28.57143} + \frac{(27-28.57143)^2}{28.57143} + \dots + \frac{(25-28.57143)^2}{28.57143} = 3.7 \quad \text{e}$$
$\chi^2_{0.05(6)} = 12.59159.$

$X^2_{cal} < \chi^2_{0.05(6)}$, logo não se rejeita $\mathbf{H}_0$ a um nível de significância de 5%, portanto não temos motivos para duvidar que o número de leitões por ninhada seja uma uniforme discreta em $\{6, \dots, 12\}$.

Ex4. Dados: $n = 200$; $O_i = (141, 48, 9, 2, 0, 0, 0)$ e o modelo a testar é que X, “número de latas impróprias”, segue uma lei Binomial(6, 0.05)

Comecemos então por formular as seguintes hipóteses

$\mathbf{H}_0$ **:** $X \frown Binomial(6, 0.05)$ *vs* $\mathbf{H}_{1:}$ $X \not\frown Binomial(6, 0.05)$

$\alpha = 0.05$

Temos então os seguintes dados, onde $p_{i0} = P[X = x_i | H_0]$:

x_i	0	1	2	3	4	5	6
O_i	141	48	9	2	0	0	0
p_{i0}	0.73509189	0.23213428	0.03054398	0.00214344	0.00008461	0.00000178	0.00000002
$E_i = np_{i0}$	147.018378	46.426856	6.108797	0.428688	0.016922	0.000356	0.000003

Note-se que o critério de Cochran não é agora verificado. Efectivamente as frequências esperadas para $x_i = 3, 4, 5, 6$ são todas inferiores a 1. Observe-se que ao executar o comando no R para realizar o teste ele envia um Aviso, dizendo que a aproximação usada pode estar incorrecta.

Vamos então agrupar os 5 últimos valores, i.e. ficamos com a tabela

x_i	0	1	≥ 2
O_i	141	48	11
p_{i0}	0.73509189	0.23213428	0.03277383
$E_i = np_{i0}$	147.018378	46.426856	6.554766

Ver o Anexo, onde se tem $X^2_{cal} = X - squared = 3.3143$ e $p - value = 0.1907$, por conseguinte não se rejeita a hipótese de "o n º de latas impróprias‘" seguir uma lei Binomial(6,0.05).

Ex5. Dados: $n = 100$; $O_i = (15, 25, 30, 20, 10)$ e o modelo a testar é que X, "número de utilizadores da caixa multibanco em períodos de meia hora", segue uma lei Poisson(2)

Comecemos então por formular as seguintes hipóteses

$\mathbf{H}_0$ **:** $X \frown Poisson(2)$ *vs* $\mathbf{H}_{1:}$ $X \not\frown Poisson(2)$

$\alpha = 0.05$

Teriamos então a seguinte tabela, onde $p_{i0} = P[X = x_i | H_0]$:

x_i	0	1	2	3	4
O_i	15	25	30	20	10
p_{i0}	0.13533528	0.27067057	0.27067057	0.18044704	0.09022352
$E_i = np_{i0}$	13.533528	27.067057	27.067057	18.044704	9.022352

Mas nesta tabela não temos contruida uma lei de probabilidade, note-se que

$\sum p_{i0} = 0.947347$ e, claro, $\sum E_i = 94.7347$ e não 100 como deveria ser.

Como a v.a. com distribuição de Poisson pode tomar uma infinidade de valores temos que considerar mais uma classe "$x_i \geq 5$", embora nesta experiência não tenha sido observado nenhum valor igual ou superior a 5.

A tabela deve ser então:

x_i	0	1	2	3	4	≥ 5
O_i	15	25	30	20	10	0
p_{i0}	0.13533528	0.27067057	0.27067057	0.18044704	0.09022352	0.05265302
$E_i = np_{i0}$	13.533528	27.067057	27.067057	18.044704	9.022352	5.265302

Agora as frequências esperados satisfazem o critério de Cochran pelo que, sob a hipótese nula, pode admitir-se que a estatística de Pearson, $X^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$ tem, assintoticamente, distribuição $\chi^2_{(5)}$.

Os resultados em Anexo permitem concluir que, com um $p.value = 0.2856$, não se deve rejeitar a hipótese nula, pelo que a distribuição de Poisson(2) pode ser admitida para modelar "o número de utilizadores da caixa multibanco em períodos de meia hora"

Ex11. A resolução apresenta-se em R. Encontra-se em Anexo, acompanhada de comentários.

Anexo – comandos e sua execução em R, para apoio à resolução

```
###############################
##ex1
####
###############################
###dados

> pi0<-c(1/4,1/4,1/2)
> Oi<-c(36,32,73)

#teste
> chisq.test(Oi,p=pi0)

        Chi-squared test for given probabilities

data:  Oi
X-squared = 0.4043, df = 2, p-value = 0.817

###############################
##ex2
####
###############################
```

```
###dados

> Oi<-c(45, 50, 65, 52, 88)
> (pi0<-rep(0.2,5))
> (teste2<-chisq.test(Oi,p=pi0))

> teste2$expected  #Mostra as frequências esperadas

> qchisq(0.05,4,lower.tail=FALSE) ## dá o valor critico


###############################
##ex3
####
###############################
###dados
>
> Oi<-c(25, 27, 31, 34, 32, 29, 22)
> sum(Oi)
[1] 200
> (pi0<-rep(1/7,7))
[1] 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571
> (teste3<-chisq.test(Oi,p=pi0))

        Chi-squared test for given probabilities

data:  Oi
X-squared = 3.7, df = 6, p-value = 0.7172

>
> (Ei<-200*pi0)  #valores esperados sob H0
[1] 28.57143 28.57143 28.57143 28.57143 28.57143 28.57143 28.57143
>
> qchisq(0.05,6,lower.tail=FALSE) ## valor critico
[1] 12.59159
>


###############################
##ex4
####
###############################
###dados
```

```
> Oi<-c(141, 48, 9, 2, 0,0,0)
> sum(Oi)
[1] 200
> (x<-seq(0,6))
[1] 0 1 2 3 4 5 6
> (pi0<-c(dbinom(x,6,0.05)))
[1] 7.350919e-01 2.321343e-01 3.054398e-02 2.143438e-03 8.460937e-05
[6] 1.781250e-06 1.562500e-08
>
## por curiosidade consulte as tabelas  da binomial e veja os valores
##das probabilidades
>
> (teste4<-chisq.test(Oi,p=pi0))

        Chi-squared test for given probabilities

data:  Oi
X-squared = 7.4448, df = 6, p-value = 0.2817


 ## Note que o R envia o seguinte  Aviso:
 "In chisq.test(Oi, p = pi0) : Chi-squared approximation may be incorrect"

 ##Na verdade o critério de Cochran não se verifica, veja os valores esperados
>
> (Ei<-round(200*pi0,4))
[1] 147.0184  46.4269   6.1088   0.4287   0.0169   0.0004   0.0000
>
## Portanto há frequências esperadas inferiores a 1 e só  43% são
superiores a 5. Deve então juntar-se classes, vamos então considerar
 juntar classes   (0,1, ">=2")

> Oi_1<-c(141,48,11)

## vamosm usar o R para calcular as probabilidades para cada valor de x, sob H0
> (pi0_1<-c(dbinom(0,6,0.05),dbinom(1,6,0.05),1-pbinom(1,6,0.05)))
[1] 0.73509189 0.23213428 0.03277383
> sum(pi0_1)
[1] 1
>
> (teste4_1<-chisq.test(Oi_1, p=pi0_1))
```

```
        Chi-squared test for given probabilities

data:  Oi_1
X-squared = 3.3143, df = 2, p-value = 0.1907

>
> (valor_critico<-qchisq(0.05,2,lower.tail=FALSE))
[1] 5.991465
>

###############################
##ex5
####
###############################
###dados


> xi<-seq(0,4)
> Oi<-c(15,25,30,20,10)
> sum(Oi)
[1] 100
> (pi0<-dpois(xi,2))
[1] 0.13533528 0.27067057 0.27067057 0.18044704 0.09022352
>
> chisq.test(Oi,p=pi0)
Error in chisq.test(Oi, p = pi0) : probabilities must sum to 1.

# Verifique que o R avisa que não tem uma lei de probabilidade,
uma vez que as probabilidades não somam 1


> sum(pi0)
[1] 0.947347
# verifique-se que esta soma não dá 1, de facto a lei de Poisson
# pode tomar uma infinidade de valores portanto, embora não se
# tenha observado valores superiores a 4, tem que ser considerado
# no modelo deve então considerar-se a classe >=5, para a qual
# foram recolhidas 0 observações

>
> (Ei<-sum(Oi)*pi0)
[1] 13.533528 27.067057 27.067057 18.044704  9.022352
```

```
> sum(Ei)
[1] 94.7347
# Como é óbvio também a soma dos  valores esperados não deu 100

> ##  consideremos então a possibilidade de x>=5
> (Oi_1<-c(Oi,0))
[1] 15 25 30 20 10  0
> (pi0_1<-c(pi0, 1-ppois(4,2)))
[1] 0.13533528 0.27067057 0.27067057 0.18044704 0.09022352 0.05265302
>
> sum(pi0_1)
[1] 1
>
> (teste5<-chisq.test(Oi_1,p=pi0_1))

        Chi-squared test for given probabilities

data:  Oi_1
X-squared = 6.2177, df = 5, p-value = 0.2856

>
> (Ei_1<-sum(Oi_1)*pi0_1)
[1] 13.533528 27.067057 27.067057 18.044704  9.022352  5.265302
> sum(Ei_1)
[1] 100

###############################
##ex11
####
###############################
###dados

> sort(y)
 [1]  0.01  0.04  0.08  0.10  0.11  0.13  0.13  0.15  0.15  0.23  0.23  0.24
[13]  0.26  0.29  0.37  0.41  0.42  0.49  0.51  0.53  0.57  0.65  0.67  0.67
[25]  0.70  0.75  0.76  0.77  0.82  0.88  0.90  0.98  1.01  1.04  1.06  1.10
[37]  1.28  1.32  1.35  1.41  1.42  1.43  1.55  1.58  1.63  1.68  1.72  1.76
[49]  1.78  1.80  2.01  2.03  2.04  2.05  2.06  2.07  2.44  2.49  2.58  2.62
[61]  2.71  2.77  2.83  2.86  2.90  3.08  3.21  3.39  3.53  3.57  3.90  4.29
[73]  4.39  4.57  4.86  5.50  6.49  9.48 11.66 14.12

#estamos agora perante uma v.a. contínua
# como é sugerido considerar  as classes
```

```
##   [0,2]     ]2,4]   ]4,6]    ]6,8]    ]8,10]   ]10,12]   ]12,16[


# mais uma vez a v.a. exponencial y pode ir até infinito,
então a probabilidade a calcular na última classe será P[X>=12]


Oi<-c(55, 16,  5,  1,  1,  1,  1)  #freq. observadas

#é necessário calcular as probabilidades, sob H0, em cada classe

> pi0<-c(pexp(2,1/3),pexp(4,1/3)-pexp(2,1/3),pexp(6,1/3)-pexp(4,1/3),
+ pexp(8,1/3)-pexp(6,1/3),pexp(10,1/3)-pexp(8,1/3),pexp(12,1/3)-pexp(10,1/3),
+ 1-pexp(12,1/3))
> pi0
[1] 0.48658288 0.24981998 0.12826185 0.06585183 0.03380946
0.01735835 0.01831564
> sum(pi0)
[1] 1
> (Ei<-80*pi0)
[1] 38.926630 19.985598 10.260948  5.268147  2.704757  1.388668  1.465251

# De novo não se verifica o critério de Cochran,
vamos juntar as 3 últimas classes

> (pi0.1<-c(0.48658288, 0.24981998, 0.12826185, 0.06585183,
+ 0.03380946+ 0.01735835+ 0.01831564))
[1] 0.48658288 0.24981998 0.12826185 0.06585183 0.06948345

> (Ei.1<-80*pi0.1)
[1] 38.926630 19.985598 10.260948  5.268146  5.558676

> Oi.1<-c(55, 16,  5, 1,3)

> chisq.test(Oi.1,p=pi0.1)

        Chi-squared test for given probabilities

data:  Oi.1
X-squared = 14.7649, df = 4, p-value = 0.005215

## Resposta: com o p-value=0.005215 rejeitamos H0;
portanto não é de admitir que a distribuição seja  Exponencial(3)
```