

## Teste qui-quadrado de bondade de ajuste – distribuições Poisson e normal

### 1. Poisson

Neste exemplo utilizamos os dados do exercício 4.21, p. 153 em Gibbons and Chakraborti (2003, *Nonparametric Statistical Inference*, 4th ed., Marcel Dekker: New York). Os dados consistem do número de erros de digitação em conjuntos de 1000 palavras. Argumenta-se que o número de erros tem distribuição de Poisson com uma taxa média de 4 erros/1000 palavras digitadas. A quantidade de erros e sua frequência em 100 amostras de 1000 palavras são fornecidas abaixo.

```
erros <- 0:5
f <- c(10, 16, 20, 28, 12, 14)
```

```
cat("\n n =", n <- sum(f))
```

```
n = 100
```

Iniciamos testando a hipótese nula simples  $H_0$  de que a média da distribuição é  $\lambda = \lambda_0 = 4$ . Com este valor calculamos as probabilidades,

```
lambda0 <- 4
prob0 <- dpois(erros, lambda = lambda0)
```

que totalizam 0,785 ( $\text{sum}(\text{prob0})$ ). Assim, acrescentamos uma frequência nula ao vetor de frequências (correspondente a 6 ou mais erros) e completamos o vetor de probabilidades.

```
f0 <- c(f, 0)
prob0 <- c(prob0, 1 - sum(prob0))
```

As frequências esperadas sob  $H_0$  são calculadas por `print(n * prob0, digits = 3)` resultando em

```
3.99 15.97 31.94 42.59 42.59 34.07 46.84
```

notando que a primeira é menor do que 5. O teste qui-quadrado segue abaixo.

```
chisq.test(f0, p = prob0)
```

```
Chi-squared test for given probabilities
```

```
data: f0
X-squared = 76.8809, df = 6, p-value = 1.573e-14
```

```
Warning message:
```

```
In chisq.test(f, p = prob0) : Chi-squared approximation may be incorrect
```

O número de g.l. está correto porque  $H_0$  é simples. O resultado do teste indica a rejeição de  $H_0$ . A mensagem na última linha é exibida porque uma das frequências esperadas é menor do que 5. Vale a pena realizar o teste com o valor- $p$  simulado. A conclusão baseada no valor- $p$  simulado a partir de  $B = 5000$  réplicas é a rejeição de  $H_0$ , notando que em todas elas o valor da estatística  $Q$  de Pearson foi pelo menos igual ao valor de  $Q$  na amostra (por quê?).

```
chisq.test(f0, p = prob0, simulate.p.value = TRUE, B = 5000)
```

```
Chi-squared test for given probabilities with
simulated p-value (based on 5000 replicates)
```

```
data: f0
X-squared = 76.8809, df = NA, p-value = 2e-04
```

Em seguida realizamos o teste mudando a hipótese nula, sem especificar o valor de  $\lambda$ . Devemos estimar  $\lambda$ .

```
lambdac <- sum(erros * f) / n
cat("\n EMV de lambda =", lambdac)
```

```
EMV de lambda = 2.58
```

Nota 1. Verifique o resultado do comando `fitdistr(rep(erros, times = f), "poisson")`.

Tendo a EMV de  $\lambda$  obtemos as probabilidades sob  $H_0$  e as frequências esperadas estimadas,

```
prob0 <- dpois(erros, lambda = lambdac)
prob0 <- c(prob0, 1 - sum(prob0))
print(n * prob0, digits = 3)
```

```
7.58 19.55 25.22 21.69 13.99 7.22 4.76
```

cujos valores devem ser comparados com  $f_0$  (10 16 20 28 12 14 0). Uma das estimativas das frequências esperadas é menor do que 5. Realizamos o teste.

```
(testeh0 <- chisq.test(f0, p = prob0))
```

```
Chi-squared test for given probabilities
```

```
data: f0
X-squared = 15.7481, df = 6, p-value = 0.01517
```

```
Warning message:
```

```
In chisq.test(f0, p = prob0) : Chi-squared approximation may be incorrect
```

O valor- $p$  obtido da distribuição  $\chi^2$  deve levar em conta que  $H_0$  é composta.

```
pchisq(testeh0$statistic, testeh0$parameter - 1, lower.tail = FALSE)
```

```
0.007601323
```

Assim, os dados apontam que a distribuição não é Poisson. Para um nível de significância de 5%, a conclusão é a mesma se observarmos o valor- $p$  simulado.

```
chisq.test(f0, p = prob0, simulate.p.value = TRUE, B = 10000)
```

```
Chi-squared test for given probabilities with
simulated p-value (based on 10000 replicates)
```

```
data: f0
```

```
X-squared = 15.7481, df = NA, p-value = 0.0147
```

Nota 2. Refaça o exemplo com a estatística  $-2 \log \Lambda$  no lugar da estatística  $Q$  de Pearson.

## 2. Normal

Além de testes de normalidade como os de Anderson-Darling e Cramér-von Mises, no pacote `nortest` também temos a função `pearson.test`. Um dos argumentos da função é o número de classes (`n.classes`), com valor *default*

$$k = \lceil 2n^{2/5} \rceil.$$

São consideradas classes equiprováveis. Por *default* (`adjust = TRUE`), os graus de liberdade são  $k - 3$ , levando que conta que  $H_0$  é composta.

```
# Dados. Distância percorrida por carros até parar (em pés)
```

```
# Conjunto de dados "cars"
```

```
dados <- cars$dist
```

```
dados <- dados * 12 * 2.54 / 100 # em m
```

```
cat("\n n =", length(dados))
```

```
n = 50
```

```
summary(dados)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.6096  7.9250 10.9700 13.1000 17.0700 36.5800
```

```
(testep <- pearson.test(dados))  
  
      Pearson chi-square normality test  
data:  dados  
      P = 11.2, p-value = 0.1301
```

A um nível de significância de 5%, a hipótese de normalidade não é rejeitada ( $p = 0,1301$ ).

Nota 3. Apresente as estimativas de máxima verossimilhança dos parâmetros (vide a função `fitdistr`).

Nota 4. O número de classes é  $k = 10$  (`testep$n.classes`). Realize o teste com diferentes números de classes para ter uma ideia da sensibilidade do resultado quanto a esta escolha.

Nota 5. Compare os resultados deste teste com os de outros testes de normalidade.