

## Funções `chisq.test` e `fisher.test`

São apresentados exemplos com as funções `chisq.test` e `fisher.test` do pacote `stats` em R. Os resultados são destacados em cor azul.

### 1. `chisq.test`

Pode ser utilizada para testar a bondade do ajuste a uma distribuição multinomial. Também pode ser utilizada para testar a independência de duas variáveis a partir das contagens em uma tabela de contingências bidimensional.

Exemplo 1.1. Testar se uma distribuição multinomial com probabilidades  $(1/3, 1/3, 1/3)$  ajusta bem as contagens  $(20, 15, 25)$ .

Neste exemplo  $N = 3$  e  $n = 20 + 15 + 25 = 60$ . Como as probabilidades sob a hipótese nula são iguais a  $1/N$ , não é necessário especificar.

```
x <- c(20, 15, 25)
(ex11 <- chisq.test(x))
```

Chi-squared test for given probabilities

```
data: x
X-squared = 2.5, df = 2, p-value = 0.2865.
```

O valor da estatística de teste é  $X^2 = 2,5$  e está armazenado em `ex11$statistic`. Com dois graus de liberdade ( $N - 1 = 2$ ), obtemos  $\text{valor-p} = 0,2865$ , que pode ser calculado como

```
pchisq(ex11$statistic, df = length(x) - 1, lower.tail = FALSE)
```

ou

```
pchisq(ex11$statistic, df = ex11$parameter, lower.tail = FALSE)
```

0.2865048

Exemplo 1.2. Testar se uma distribuição multinomial com probabilidades  $(1/4, 1/8, 5/8)$  ajusta bem as contagens  $(14, 25, 81)$ .

Neste exemplo  $N = 3$  e  $n = 120$ . Como as probabilidades sob a hipótese nula são diferentes de  $1/N$ , devemos informar estas probabilidades com o argumento `p`.

```
x <- c(14, 25, 81)
prob0 <- c(1/4, 1/8, 5/8)
(ex12 <- chisq.test(x, p = prob0))
```

Chi-squared test for given probabilities

```
data: x
X-squared = 15.68, df = 2, p-value = 0.0003937
```

Exemplo 1.3. Testar se uma distribuição multinomial com probabilidades proporcionais a (2, 5, 3, 8) ajusta bem as contagens (19, 62, 31, 105).

Neste exemplo  $N = 4$  e  $n = 217$ . As probabilidades em si não foram fornecidas, mas são iguais às constantes de proporcionalidade divididas pela sua soma. Podemos especificar diretamente as constantes de proporcionalidade, bastando informar o argumento `rescale.p` como `TRUE`.

```
x <- c(19, 62, 31, 105)
prop0 <- c(2, 5, 3, 8)
chisq.test(x, p = prop0, rescale.p = TRUE)
```

Chi-squared test for given probabilities

```
data: x
X-squared = 2.6297, df = 3, p-value = 0.4523
```

Exemplo 1.4. A partir dos dados (obtidos sem que as margens fossem fixadas)

```
n <- as.table(rbind(c(762, 327, 468), c(484, 239, 477)))
dimnames(n) <- list(gender = c("M", "F"),
                    party = c("Democrat", "Independent", "Republican"))
n
```

	party		
gender	Democrat	Independent	Republican
M	762	327	468
F	484	239	477

testar a hipótese de independência entre as variáveis *gender* e *party*

Os dados compõem uma tabela  $2 \times 3$  e o tamanho da amostra é  $n = \text{sum}(n) = 2757$ . Adotando *gender* como variável explicativa, os gráficos de barras da Figura 1, obtidos com os comandos

```
tab14 <- prop.table(n, margin = 1) * 100
library(lattice)
barchart(tab14, xlab = "Percentage", ylab = "Gender", stack = FALSE,
scale = list(cex = 1.5), auto.key = list(space = "top", columns = 3))
```

sugerem que há dependência entre as variáveis. Realizando o teste com a estatística  $X^2$ ,

```
(ex14 <- chisq.test(n))
```

Pearson's Chi-squared test

```
data: n
X-squared = 30.0701, df = 2, p-value = 2.954e-07
```

encontramos evidências contra a hipótese nula de independência. Vale ressaltar que a função `chisq.test` neste exemplo serve para testar uma hipótese diferente daquelas dos exemplos anteriores.

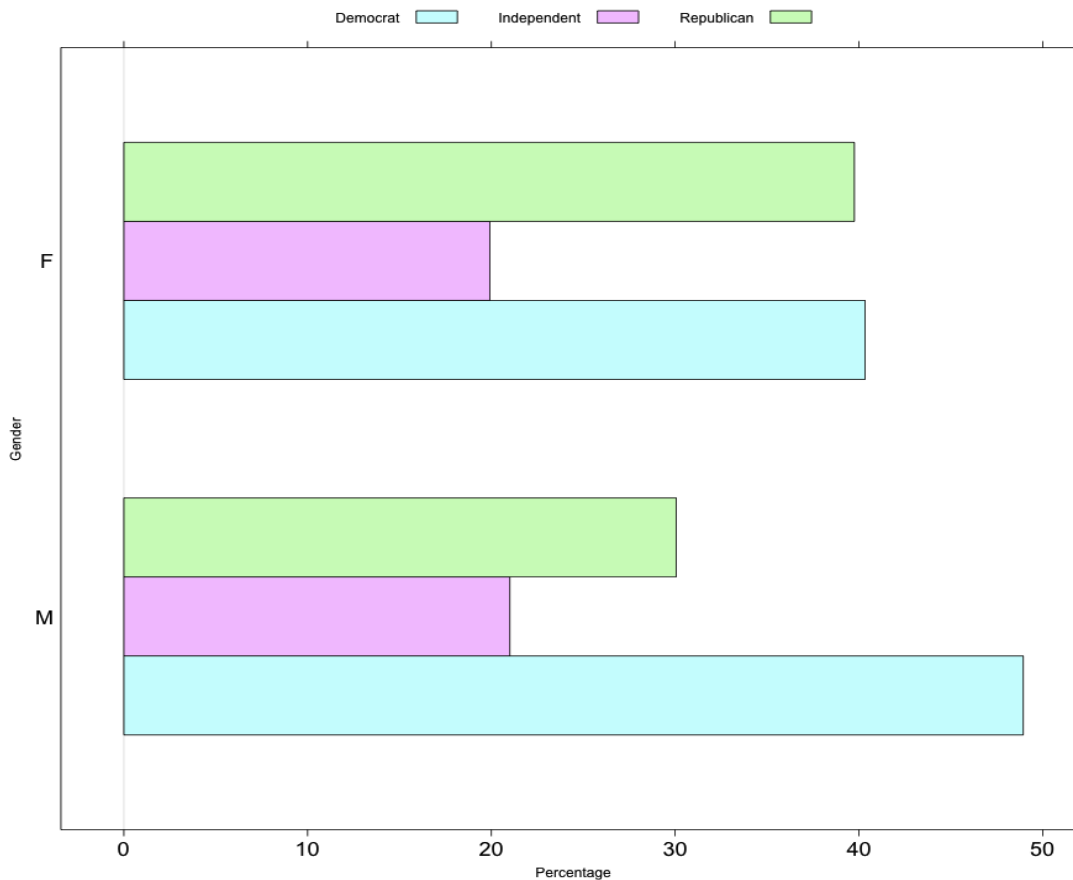


Figura 1. Gráficos de barras do exemplo 1.4.

As frequências esperadas estimadas sob  $H_0$  são

```
ex14$expected
```

```

party
gender Democrat Independent Republican
M 703.6714 319.6453 533.6834
F 542.3286 246.3547 411.3166

```

A estatística de teste  $G^2$  (baseada na razão de verossimilhanças) tem valor

```
(G2 <- 2 * sum(n * (log(n) - log(ex14$expected))))
```

```
[1] 30.01669
```

bastante próximo ao valor de  $X^2$ .

## 2. fisher.test

Pode ser mostrado que se condicionarmos nos totais marginais de uma tabela de contingências obtemos uma distribuição que não depende de parâmetros desconhecidos. Ressalte-se que (i) no modelo de Poisson até mesmo o tamanho da amostra é aleatório, (ii) no modelo multinomial apenas o tamanho da amostra é fixo e (iii) no modelo produto de multinomiais independentes os totais de uma das margens são fixados. Em uma tabela  $2 \times 2$ , sob a hipótese nula de independência entre as variáveis em (i) e (ii) e sob a hipótese nula de homogeneidade das distribuições binomiais (que equivalem a  $\theta = 1$  em uma tabela  $2 \times 2$ ), a distribuição condicional das contagens nos totais marginais é hipergeométrica como função de  $n_{11}$ . Se a hipótese alternativa for  $H_1: \theta > 1$  ( $H_1: \theta < 1$ ), quanto maior (menor)  $n_{11}$ , mais evidência contra  $H_0$ . Portanto, este resultado permite propor um teste exato para estas hipóteses, conhecido como teste exato de Fisher, implementado na função `fisher.test`.

Exemplo 2.1. Fisher's tea drinker ( $H_1: \theta > 1$ ).

```
TeaTasting <-  
matrix(c(3, 1, 1, 3),  
       nrow = 2, dimnames = list(Guess = c("Milk", "Tea"),  
                                 Truth = c("Milk", "Tea")))  
TeaTasting  
  
      Truth  
Guess Milk Tea  
Milk    3   1  
Tea     1   3  
  
fisher.test(TeaTasting, alternative = "greater")$p.value  
  
[1] 0.2428571
```

Deve ser enfatizado que em diversas situações, diferentemente do exemplo 2.1, os totais marginais não são fixados. Assim, o teste é interpretado como sendo exato condicional.

Se a tabela não é  $2 \times 2$ , são utilizadas extensões do teste. O valor-p é calculado como a soma das probabilidades das tabelas (com totais marginais fixados) que não são mais prováveis de ocorrer do que a tabela observada.

Exemplo 2.2. Job satisfaction.

```
Job <- matrix(c(1,2,1,0, 3,3,6,1, 10,10,14,9, 6,7,12,11), 4, 4,  
             dimnames = list(income=c("< 15k", "15-25k", "25-40k", "> 40k"),  
                             satisfaction=c("VeryD", "LittleD", "ModerateS", "VeryS")))  
Job
```

income	satisfaction			
	VeryD	LittleD	ModerateS	VeryS
< 15k	1	3	10	6
15-25k	2	3	10	7
25-40k	1	6	14	12
> 40k	0	1	9	11

```
fisher.test(Job)
```

```
Fisher's Exact Test for Count Data
```

```
data: Job
```

```
p-value = 0.7827 alternative hypothesis: two.sided
```

O valor-p pode ser aproximado por simulação de um certo número (argumento B) de tabelas com os totais marginais fixados.

```
fisher.test(Job, simulate.p.value = TRUE, B = 1e5)
```

```
Fisher's Exact Test for Count Data with simulated p-value (based on
1e+05 replicates)
```

```
data: Job
```

```
p-value = 0.7843
```

```
alternative hypothesis: two.sided
```