Lição 9: Testes de Proporção e Qui-Quadrado

A lição anterior, sobre conceitos fundamentais da Estatística Inferencial, é a base de todas as próximas lições, que apresentam diferentes testes estatísticos que podem ser aplicados aos dados. Esta e as próximas duas lições tratam de análises *univariadas*, que verificam se há correlação entre duas variáveis – em geral, uma variável *dependente* e uma variável *independente*.

Rode as linhas de comando a seguir para deixar disponível o dataframe `ds` nesta sessão. Defina como diretório de trabalho aquele que, em seu computador, contém o arquivo LabovDS.csv.

```
# Definir diretório de trabalho

#setwd()

# Importar planilha de dados

ds <- read_csv("LabovDS.csv",
               col_types = cols(.default = col_factor(),
                                r = col_factor(levels = c("r0", "r1", "d
"))
                                )
               ) %>%
  filter(r != "d") %>%
  droplevels()
```

Antes de tudo, carregue o pacote tidyverse, que vamos usar nesta lição (ó, surpresa!).

```
library(tidyverse)
```

A escolha do teste que pode ser aplicado depende fundamentalmente da natureza das variáveis. Que tipo de variável é o apagamento (r0) vs. a realização (r1) de /r/ pós-vocálico no inglês, como é o caso da variável estudada por Labov em seu famoso estudo nas lojas de departamento?

- nominal
- numérica

- ordinal

Para variáveis nominais, como é o caso da variável /r/ pós-vocálico, podemos aplicar *testes de proporção* e *testes de qui-quadrado* para avaliar se há diferenças significativas entre proporções das variantes.

Nesta lição, vamos trabalhar com o arquivo de dados LabovDS.csv, que havíamos visto na Lição 4. Cheque a estrutura do dataframe `ds`.

```
str(ds)
```

```
## spec_tbl_df [730 × 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ r       : Factor w/ 2 levels “r0”,”r1”: 2 2 2 2 2 2 2 2 2 2 ...
##  $ store   : Factor w/ 3 levels “Saks”,”Macys”,..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ emphasis: Factor w/ 2 levels “casual”,”emphatic”: 1 1 1 1 1 1 1
1 1 1 ...
##  $ word    : Factor w/ 2 levels “fouRth”,”flooR”: 1 1 1 1 1 1 1 1 1
1 ...
##  - attr(*, “spec”)=
##   .. cols(
##   ..   .default = col_factor(),
##   ..   r = col_factor(levels = c(“r0”, “r1”, “d”), ordered = FALSE,
include_na = FALSE),
##   ..   store = col_factor(levels = NULL, ordered = FALSE, include_n
a = FALSE),
##   ..   emphasis = col_factor(levels = NULL, ordered = FALSE, includ
e_na = FALSE),
##   ..   word = col_factor(levels = NULL, ordered = FALSE, include_na
= FALSE)
##   .. )
##  - attr(*, “problems”)=<externalptr>
```

Neste dataframe, a variável r contém apenas r0 e r1, pois os dados d (= duvidosos) já foram excluídos.

Como vimos nas Lições 4 e 5, a análise estatística começa com a estatística descritiva, com tabelas e gráficos. Façamos então cálculo de frequências e proporções da distribuição da variável r. No *script*, vamos criar um dataframe chamado `prop.r`, com os dados de `ds`, a frequência de r, e uma nova coluna chamada prop por meio da aplicação da função `prop.table()` aos dados.

```
prop.r <- ds %>%
  count(r) %>%
  mutate(prop = prop.table(n)) %>%
  print()
```

```
## # A tibble: 2 × 3
##   r         n  prop
##   <fct> <int> <dbl>
## 1 r0      499 0.684
## 2 r1      231 0.316
```

Como já havíamos visto na Lição 4, há 499 dados de r0 e 231 dados de r1, que correspondem a 68% e 32%. Vamos agora plotar um gráfico de barras exploratório para visualizar essas proporções (Figura 9.1).

```
ggplot(prop.r, aes(x = r, y = prop, fill = r)) +
  geom_bar(stat = "identity")
```

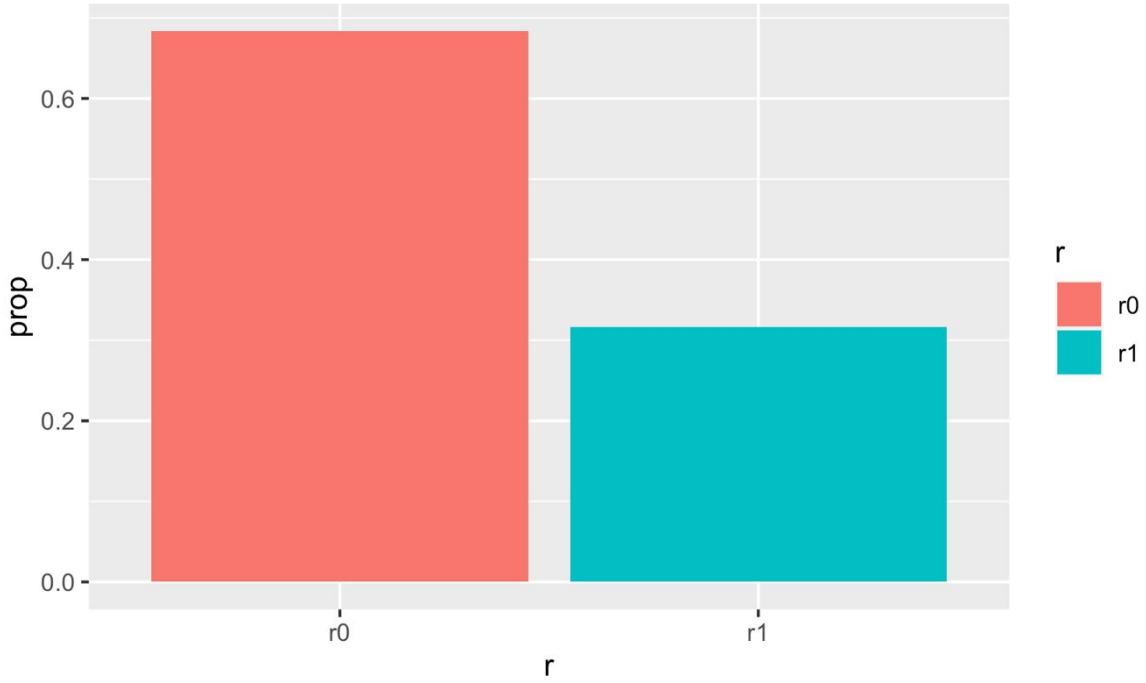

Figura 9.1: Distribuição das variantes de /r/ pós-vocálico nos dados de Labov (1972). Fonte: própria.

Faça a tabela de distribuição de dados da variável r pelas funções da instalação base do R – `with()` e `table()`. Guarde-a em um objeto chamado `tab.r`.

```
tab.r <- with(ds, table(r))
```

Inspecione o objeto `tab.r`.

```
tab.r
```

```
## r
##  r0  r1
## 499 231
```

Até aqui, refizemos alguns dos passos da Lição 4, certo? A partir das tabelas e dos gráficos – o primeiro passo de qualquer boa análise quantitativa! –, o pesquisador deve começar a avaliar quais diferenças podem ter ocorrido por acaso e quais têm menor chance de terem ocorrido aleatoriamente. O teste de proporções contrasta uma distribuição observada com uma distribuição esperada.

Em sua forma mais simples, o teste de proporções indica se a diferença entre as proporções das variantes de uma variável é significativa. No caso em questão, podemos nos perguntar se a diferença entre as proporções das variantes de /r/ (aproximadamente 68% e 32%) é significativa. A função para responder a essa questão é `prop.test()`, que é aplicada sobre uma tabela de *frequências*. Aplique-a então à tabela `tab.r` e guarde o resultado em um objeto chamado `teste.prop`.

```
teste.prop <- prop.test(tab.r)
```

Veja o resultado do teste de proporções.

```
teste.prop

##
##  1-sample proportions test with continuity correction
##
## data:  tab.r, null probability 0.5
## X-squared = 97.656, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.6482366 0.7169260
## sample estimates:
##         p
## 0.6835616
```

Vejamos o que o R nos informa. A primeira linha nos diz o teste estatístico que foi realizado: um teste de proporções com uma amostra (a distribuição de uma variável). Em seguida, o R informa os dados utilizados – `tab.r` – e a proporção contra a qual contrastou a proporção observada – 0.5. Ou seja, o R comparou a distribuição de proporções de aproximadamente 68%-32% com uma proporção de 50%-50%. Este caso é semelhante ao da moeda, que vimos na lição passada! Se a hipótese nula prevê que a distribuição deveria ter sido meio a meio, qual é a probabilidade de se ter observado uma distribuição de 68% pra 32%?

Na terceira linha, o R informa que tal probabilidade é extremamente pequena: 2.2e-16. (Veremos daqui a pouco o que são os valores X-squared (= qui-quadrado) e df (= graus de liberdade).) A quarta linha é a afirmação que expressa a hipótese alternativa, H1: "a verdadeira p (= proporção) não é igual a 0.5". Comparando-se então o valor-*p* gerado (2.2e-16) e a hipótese alternativa, o que deve fazer o pesquisador?

- Rejeitar a hipótese nula e acatar a hipótese alternativa
- Rejeitar a hipótese alternativa e acatar a hipótese nula
- Refazer o teste mais algumas vezes, para verificar se o resultado é o mesmo

Continuemos com a leitura dos resultados. Após a hipótese alternativa, o R indica quais são os valores dentro do intervalo de confiança de 95%. Esses valores são 64,8% e 71,7%, que se referem ao primeiro nível da variável `r`. Entre r0 e r1, qual é o primeiro nível? Dica: olhe o gráfico de barras!

- r0
- r1

O R contrastou a proporção observada do primeiro nível da variável, tomada como valor de referência – r0 = 68% – com a proporção esperada – r0 = 50%. Como 50% não está contido no intervalo de confiança, entre 64,8% e 71,7%, a estimativa de probabilidade de que o verdadeiro parâmetro da distribuição é 50% está abaixo de 5%, pois está efetivamente fora dos 95% do nível de confiança.

Se esse raciocínio pareceu complicado, lembre-se do exemplo da moeda: se se espera que haja 50% de caras e 50% de coroas sob a hipótese nula, ter observado 68% de uma das faces é pouco provável, a depender do número de tentativas. A diferença aqui é que estamos falando de apagamento e de realização de /r/ em vez de cara ou coroa.

O R usou o intervalo de confiança de 95% como *default*, pois não especificamos nenhum outro valor. Mas, como visto na lição anterior, o pesquisador pode determinar outro nível α para seu teste (diferente do valor *default* 5%) e, consequentemente, mudar o nível de confiança. Veja na ajuda da função `prop.test()` qual argumento precisaríamos ter especificado para operar com um IC de 99%.

```
?prop.test
```

N.B.: Resultado aqui omitido.

O argumento é conf.level, cujo valor *default* é 0.95. Então se quiséssemos realizar um teste com 99% de nível de confiança, bastaria especificar conf.level = 0.99. Por fim, o R apresenta a estimativa do primeiro nível da variável que, como já visto, é 68% de r0.

A ajuda de `prop.test()` também informa outros argumentos possíveis da função. No exemplo acima, contrastamos a proporção 68%-32% com 50%-50%, mas também poderíamos ter contrastado com outra proporção. Imagine que um estudo prévio sobre a variável r no inglês de Nova Iorque houvesse notado uma proporção de 70% de apagamento, e Labov quisesse saber se a proporção de 68% de apagamento diferia significativamente. Neste caso, seria necessário especificar p = 0.7 na função `prop.test()`. Faça isso agora com a tabela `tab.r`. Não se preocupe em guardar o resultado em um objeto, pois queremos visualizá-lo imediatamente.

```
prop.test(tab.r, p = 0.7)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  tab.r, null probability 0.7
## X-squared = 0.86269, df = 1, p-value = 0.353
## alternative hypothesis: true p is not equal to 0.7
## 95 percent confidence interval:
##  0.6482366 0.7169260
## sample estimates:
##         p
## 0.6835616
```

O R agora apresenta os resultados em relação a uma probabilidade esperada de 70%. Qual é o intervalo de confiança neste caso?

- de 0,35 a 0,86
- de 0,64 a 0,71
- de 0,68 a 0,86

Você deve ter notado que se trata exatamente do mesmo intervalo de confiança para o teste com probabilidade esperada de 50%. O que mudou agora, no entanto, é que a probabilidade 0,7 está dentro do intervalo de confiança, entre 0,64 e 0,71, de modo que

o valor-*p* agora está acima do nível α de 5%. Qual é a probabilidade de se ter observado 68% de r0 sob a hipótese nula de que o verdadeiro parâmetro é 70%?

- 0,35
- 0,68
- 0,7
- 0,86
- 1

No exemplo da moeda, também vimos dois cenários: um em que você avaliava a hipótese de eu estar roubando e outro em que um juiz imparcial avaliava se um dos jogadores estava roubando. Você sabia não estar roubando, de modo que podia estabelecer uma hipótese unidirecional, enquanto o juiz deveria estabelecer uma hipótese bidirecional. Na função `prop.test()`, o *default* é a realização de um teste bidirecional, que simplesmente avalia se as proporções diferem, independentemente de o valor de referência estar acima ou abaixo do valor esperado. Contudo, também se pode estabelecer um teste unidirecional com o argumento alternative. Dê uma olhada na ajuda da função para ver como especificar esse argumento.

Nos dados das lojas de departamento, se o pesquisador tem evidências de que a proporção de apagamento de /r/ está diminuindo em Nova Iorque, ele pode estabelecer uma hipótese unidirecional, em que a proporção esperada é *menor* do que 0.7. Aplique este teste sobre os mesmos dados acima, com a adição de alternative = “less”.

```
prop.test(tab.r, p = 0.70, alternative = “less”)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  tab.r, null probability 0.7
## X-squared = 0.86269, df = 1, p-value = 0.1765
## alternative hypothesis: true p is less than 0.7
## 95 percent confidence interval:
##  0.0000000 0.7118195
## sample estimates:
##         p
## 0.6835616
```

Veja que, neste caso, mudam as estimativas do intervalo de confiança, a hipótese alternativa e as medidas estatísticas de qui-quadrado, graus de liberdade e valor-*p*. Por curiosidade, faça o teste agora com alternative = "greater".

```
prop.test(tab.r, p = 0.70, alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  tab.r, null probability 0.7
## X-squared = 0.86269, df = 1, p-value = 0.8235
## alternative hypothesis: true p is greater than 0.7
## 95 percent confidence interval:
##  0.6539154 1.0000000
## sample estimates:
##         p
## 0.6835616
```

Este último também foi um teste unidirecional, mas que testou a hipótese de que a proporção verdadeira é *maior* do que 70%. Compare os valores do intervalo de confiança, a enunciação da hipótese alternativa e as medidas de qui-quadrado, graus de liberdade e valor-*p* em relação ao teste unidirecional anterior. Diante da hipótese alternativa de que a proporção fosse 70% ou mais, a proporção de 68% torna a hipótese nula (= proporção não é maior do que 70%) mais provável, com probabilidade igual a 82%.

Até agora, só comparamos a proporção da VD com outras proporções esperadas. Mas é provável que, em seus dados, você queira comparar proporções entre dois grupos. A partir do dataframe `ds`, crie um novo dataframe chamado `prop.word`, computando as frequências de word e r (nessa ordem), agrupando os dados pela variável word, e computando as proporções de word numa variável chamada prop com a função `prop.table()`.

```
prop.word <- ds %>%
  count(word, r) %>%
  group_by(word) %>%
  mutate(prop = prop.table(n)) %>%
  print()
```

```
## # A tibble: 4 × 4
## # Groups:   word [2]
##   word   r         n  prop
##   <fct>  <fct> <int> <dbl>
```

```
## 1 fouRth r0      295 0.770
## 2 fouRth r1       88 0.230
## 3 flooR  r0      204 0.588
## 4 flooR  r1      143 0.412
```

Visualize essas proporções com um gráfico de barras (Figura 9.2). Coloque na função `ggplot()` o nome do dataframe relevante, e as variáveis x, y e fill.

```
ggplot(prop.word, aes(x = word, y = prop, fill = r)) +
  geom_bar(stat = "identity")
```

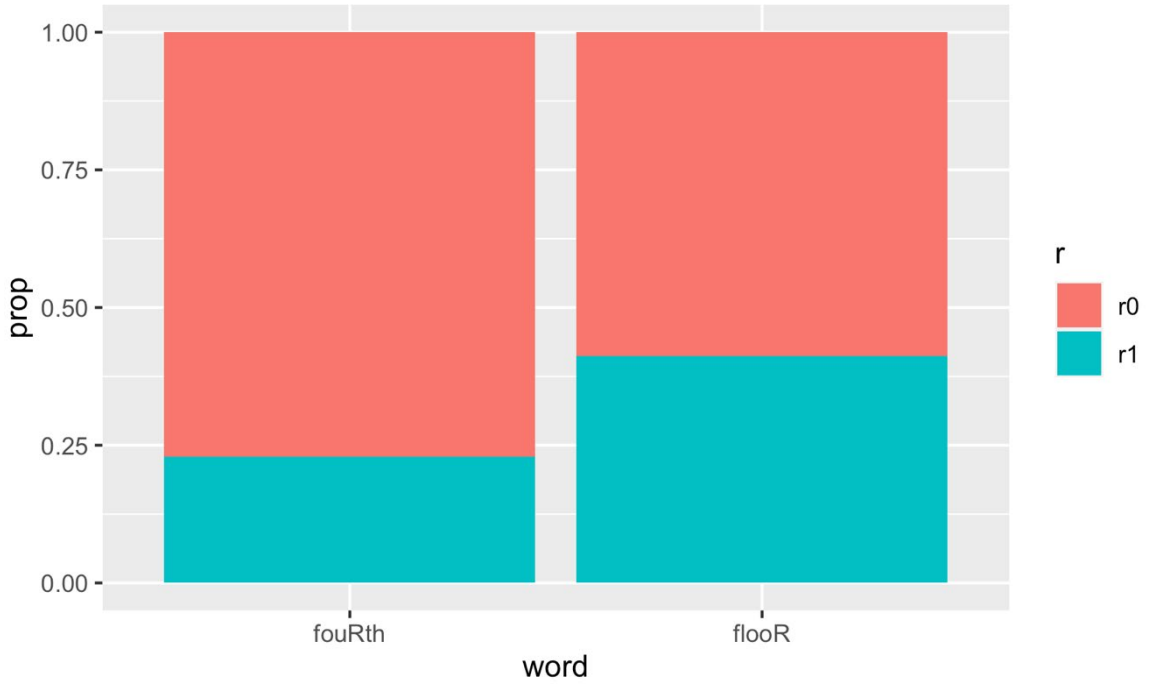

Figura 9.2: Distribuição das variantes de /r/ pós-vocálico por palavra, nos dados de Labov (1972). Fonte: própria.

Faça também uma tabela de frequências da variável word pela variável dependente r com as funções da instalação base do R, e guarde o resultado em um objeto chamado `tab.word`.

```
tab.word <- with(ds, table(word, r))
```

Inspecione agora a tabela `tab.word`.

```
tab.word

##         r
## word      r0  r1
##   fouRth 295  88
##   flooR  204 143
```

Novamente, refizemos os passos de lições anteriores, com a criação de tabelas e gráficos. Isso porque os testes estatísticos relevantes são sugeridos justamente por essas primeiras inspeções dos dados. Na figura, vemos que as barras de proporção de apagamento de /r/ nas palavras floor e fourth parecem ser diferentes. Será que as proporções são as mesmas ou diferem significativamente?

Para fazer a comparação de proporções entre dois grupos – aqui, entre os dois itens lexicais –, usamos a função `chisq.test()`. Esta função toma como argumento uma tabela de *frequências* (assim como `prop.test()`). Aplique então `chisq.test()` à tabela `tab.word` e guarde o resultado em um objeto chamado `x2.word`.

```
x2.word <- chisq.test(tab.word)
```

Veja agora o resultado do teste de qui-quadrado.

```
x2.word

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab.word
## X-squared = 27.147, df = 1, p-value = 1.886e-07
```

Voilà! Este é um teste de qui-quadrado. Sua aplicação é extremamente fácil, não? Faz-se uma tabela de frequências e aplica-se o teste sobre ela. Na prática, é isso que você vai fazer com suas variáveis nominais: tabela, gráfico, teste.

Mas importa saber não apenas como aplicá-lo no R, mas também *quando* aplicá-lo e *como ler os resultados*. Você sabe agora quando aplicá-lo: quando se tem uma VD nominal e se quer comparar proporções entre grupos (os níveis de uma VI também nominal). O resultado do teste de qui-quadrado é bastante simples e direto: o R informa o teste realizado (teste de qui-quadrado de Pearson), o conjunto de dados (`tab.word`), e valores de qui-quadrado, graus de liberdade e valor-*p*.

Você já sabe interpretar o valor-*p*: a probabilidade de se ter observado tal distribuição dos dados em caso de a hipótese nula ser verdadeira. Aqui, o teste de qui-quadrado indica que as proporções entre os grupos (59% de r0 em "floor" e 77% em "fourth") são significativamente diferentes, pelo valor-*p* abaixo de 5%. Mas o que querem dizer as medidas de qui-quadrado e de graus de liberdade?

Apesar de o resultado do teste de qui-quadrado ser bastante simples e direto, o R na verdade computou outros valores com o teste. Verifique a estrutura do objeto `x2.word` por meio da função `str()`.

```
str(x2.word)
```

```
## List of 9
##  $ statistic: Named num 27.1
##   ..- attr(*, “names”)= chr “X-squared”
##  $ parameter: Named int 1
##   ..- attr(*, “names”)= chr “df”
##  $ p.value  : num 1.89e-07
##  $ method   : chr “Pearson's Chi-squared test with Yates' continuit
y correction”
##  $ data.name: chr “tab.word”
##  $ observed : 'table' int [1:2, 1:2] 295 204 88 143
##   ..- attr(*, “dimnames”)=List of 2
##   .. ..$ word: chr [1:2] “fouRth” “flooR”
##   .. ..$ r   : chr [1:2] “r0” “r1”
##  $ expected : num [1:2, 1:2] 262 237 121 110
##   ..- attr(*, “dimnames”)=List of 2
##   .. ..$ word: chr [1:2] “fouRth” “flooR”
##   .. ..$ r   : chr [1:2] “r0” “r1”
##  $ residuals: 'table' num [1:2, 1:2] 2.05 -2.16 -3.02 3.17
##   ..- attr(*, “dimnames”)=List of 2
##   .. ..$ word: chr [1:2] “fouRth” “flooR”
##   .. ..$ r   : chr [1:2] “r0” “r1”
##  $ stdres   : 'table' num [1:2, 1:2] 5.29 -5.29 -5.29 5.29
##   ..- attr(*, “dimnames”)=List of 2
##   .. ..$ word: chr [1:2] “fouRth” “flooR”
##   .. ..$ r   : chr [1:2] “r0” “r1”
##  - attr(*, “class”)= chr “htest”
```

O objeto `x2.word` é uma lista que contém outros valores, e que podem ser acessados por meio do operador `$`. Já usamos esse operador para acessar colunas de um dataframe. Aqui, vamos usá-lo para acessar valores da lista. O que veremos na sequência é como o R computou os valores de qui-quadrado e de graus de liberdade, para mais bem compreender essas medidas estatísticas. A rigor, você não precisa saber disso para saber aplicar o teste – que já foi feito acima. Mas é importante saber um pouco do raciocínio por trás desse teste para saber como foram computados esses valores.

Um dos itens da lista são os valores observados, que são acessados por `x2.word$observed`. Guarde os valores observados em um objeto chamado `O` (“o” maiúsculo).

```
O <- x2.word$observed
```

Faça o mesmo com os valores esperados (x2.word$expected), guardando-os num objeto chamado `E`.

```
E <- x2.word$expected
```

Inspecione agora o objeto `O`, criado acima.

```
O
```

```
##         r
## word      r0  r1
##   fouRth 295  88
##   flooR  204 143
```

Ora, nada mais é do que a tabela de frequências, `tab.word`, que você já havia criado, não? Inspecione agora o objeto `E`.

```
E
```

```
##         r
## word           r0       r1
##   fouRth 261.8041 121.1959
##   flooR  237.1959 109.8041
```

Esta tabela é nova. O que são esses números? Para saber como computá-los, adicione as margens na tabela de valores observados `O`, por meio da função `addmargins()`.

```
addmargins(O)
```

```
##         r
## word      r0  r1 Sum
##   fouRth 295  88 383
##   flooR  204 143 347
##   Sum    499 231 730
```

Com `addmargins()`, temos os totais de linhas e colunas. Os valores esperados são obtidos por meio da conta: (T-Linha * T-Coluna) / T-Geral. Rode o comando (499 * 383) / 730 – ou seja, o total da primeira coluna vezes o total da primeira linha, dividido pelo total geral de dados – para ver o resultado.

```
(499 * 383) / 730
```

```
## [1] 261.8041
```

O valor corresponde exatamente àquele na primeira linha e primeira coluna da tabela E. Compute agora o valor esperado para a segunda linha da primeira coluna.

```
(499 * 347) / 730
```

```
## [1] 237.1959
```

Calcule o valor esperado para a primeira linha da segunda coluna.

```
(231 * 383) / 730
## [1] 121.1959
```

E calcule o valor esperado para a segunda linha da segunda coluna.

```
(231 * 347) / 730
## [1] 109.8041
```

Ok, agora você sabe computar os valores esperados a partir dos valores observados. O que são esses valores? Faça a divisão do valor esperado de r0 para a palavra fourth (E[1, 1]) pelo total de dados de fourth (383), linha de comando que já está no *script*.

```
E[1, 1] / 383
## [1] 0.6835616
```

Faça também a divisão do valor esperado de r0 para a palavra floor (E[2, 1]) pelo total de dados de floor (347).

```
E[2, 1] / 347
## [1] 0.6835616
```

Ambos dão o valor de 0,683561. Você se lembra onde já viu esse número?

Sim, é a proporção geral de r0 na amostra! Os valores esperados são justamente aqueles que se esperariam caso não houvesse diferença entre as proporções (no exemplo, entre as proporções de r0 nas palavras fourth e floor) – que seriam, então, iguais à proporção geral de r0 na amostra (68%).

O teste de qui-quadrado compara valores observados com valores esperados de acordo com a hipótese nula. Foi o mesmo que fizemos no teste de proporções, certo? Aliás, essencialmente, *todo* teste estatístico faz isso: compara valores observados com valores esperados.

O valor de qui-quadrado é calculado a partir dos valores observados e esperados, por meio da fórmula na sequência. Não se assuste com a notação matemática, pois essa fórmula não é nada complicada. Ela define o valor de qui-quadrado como a soma das

diferenças entre valores observados e esperados elevadas ao quadrado, divididas pelos valores esperados.

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

No R, qual fórmula a seguir expressa a fórmula do qui-quadrado?

- `(O - E)^2 / E`
- `sum(O ^ 2 - E ^ 2 / E)`
- `sum((O - E) ^ 2 / E)`
- `sum(O - E)^2 / E`

Acima, havíamos definido os valores observados e esperados nos objetos `O` e `E` respectivamente. Portanto, podemos aplicar a fórmula `sum((O - E) ^ 2 / E)` para calcular o valor de qui-quadrado. Faça isso agora.

```
sum((O - E) ^ 2 / E)
```

```
## [1] 27.98314
```

O R fornece o valor de qui-quadrado = 27,98314. Como esse valor se compara com aquele do teste feito acima? Inspecione novamente o objeto `x2.word`.

```
x2.word
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab.word
## X-squared = 27.147, df = 1, p-value = 1.886e-07
```

No teste, o valor de qui-quadrado foi de 27,147, e no nosso teste foi 27,983. Parecido, mas não igual. Isso porque o teste de qui-quadrado no R tem como valor *default* uma correção feita para tabelas 2 x 2 (2 linhas e 2 colunas), como é o caso do nosso exemplo. Esse argumento pode ser definido como correct = F para que o R não faça a correção. Aplique a função `chisq.test()` sem correção da medida de qui-quadrado à tabela `tab.word`. Não se preocupe em guardar o resultado.

```
chisq.test(tab.word, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  tab.word
## X-squared = 27.983, df = 1, p-value = 1.224e-07
```

O mesmíssimo valor que computamos! Agora você sabe que o valor de qui-quadrado é uma medida da diferença entre valores observados e esperados em uma distribuição. Quanto mais próximo o qui-quadrado estiver de zero, mais os valores observados se aproximam dos valores esperados – e, portanto, maior a chance de se ter observado tal distribuição em caso de a hipótese nula ser verdadeira (i.e. maiores valores de *p*!). No entanto, a interpretação do valor de qui-quadrado depende dos graus de liberdade, pois tabelas maiores tendem a gerar valores de qui-quadrado maiores.

O cálculo dos graus de liberdade é bastante simples. A fórmula é (n-linhas - 1) * (n-colunas - 1). Faça essa conta para nossa tabela de frequências `tab.word`, que tem 2 linhas e 2 colunas.

```
(2 - 1) * (2 - 1)
```

```
## [1] 1
```

Para tabelas 2 x 2, os graus de liberdade são sempre 1. Você pode entender o valor de graus de liberdade como o número de células de que você precisa, junto com os valores totais de linhas e colunas, para conseguir deduzir os demais valores. Isso está ilustrado na Tabela 9.1. Com apenas uma das células e os totais, você conseguiria definir quais foram as demais frequências.

Tabela 9.1: Graus de liberdade em tabelas 2 x 2.

	r0	r1	soma
flooR	?	?	347
fouRth	295	?	383
soma	499	231	730

Fonte: própria.

Com os valores de qui-quadrado e graus de liberdade, o pesquisador pode consultar uma tabela de distribuição de qui-quadrado para determinar o valor-*p*, a

probabilidade de se ter observado tal distribuição em caso de a hipótese nula ser verdadeira.

A Tabela 9.2 é uma tabela de qui-quadrado. As linhas apresentam os valores de qui-quadrado de 1 a 10 graus de liberdade; as colunas indicam as probabilidades associadas a cada valor de qui-quadrado.

Tabela 9.2: Tabela de distribuição de qui-quadrado.

	Probabilidade										
df	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,01	0,001
1	0,004	0,02	0,06	0,15	0,46	1,07	1,64	2,71	3,84	6,64	10,83
2	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	9,21	13,82
3	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,82	11,34	16,27
4	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	13,28	18,47
5	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	15,09	20,52
6	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	16,81	22,46
7	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	18,48	24,32
8	2,73	3,49	4.59	5,53	7,34	9,52	11,03	13,36	15,51	20,09	26,12
9	3,32	4,17	5.38	6,39	8,34	10,66	12,24	14,68	16,92	21,67	27,88
10	3,94	4,86	6,18	7,27	9,34	11,78	15,99	15,99	18,31	23,21	29,59

Caso o R não houvesse calculado a significância pra nós, consultaríamos a primeira linha da tabela até encontrar o valor $\chi^2 = 27{,}983$. Da esquerda pra direita, vemos que os valores de qui-quadrado aumentam. O valor $\chi^2 = 27{,}983$, portanto, está à direita da última coluna. Vemos também que os valores de probabilidade, da esquerda pra direita, diminuem. Dessa forma, o valor-*p* para $\chi^2 = 27{,}983$, com 1 grau de liberdade, é menor do que 0,001. No teste de qui-quadrado, o R nos forneceu esse valor com maior precisão: $p = 1.224e\text{-}07$.

Vejamos agora outro exemplo, com uma tabela maior do que 2 x 2. No caso acima, em que testamos a diferença entre proporções para os itens floor e fourth, sabemos que, se há diferença, só pode ser entre esses itens. Mas e se estivermos tratando de uma variável com três ou mais níveis? Numa variável com os fatores A, B e C, eventuais diferenças verificadas podem estar entre A-B, A-C, B-C ou entre todos eles.

Façamos novamente os passos da análise. Primeiro, faça um novo dataframe (`prop.store`) que computa as frequências e as proporções de store por r.

```
prop.store <- ds %>%
  count(store, r) %>%
  group_by(store) %>%
  mutate(prop = prop.table(n)) %>%
  print()

## # A tibble: 6 × 4
## # Groups:   store [3]
##   store r          n   prop
##   <fct> <fct> <int>  <dbl>
## 1 Saks  r0       93 0.522
## 2 Saks  r1       85 0.478
## 3 Macys r0      211 0.628
## 4 Macys r1      125 0.372
## 5 Klein r0      195 0.903
## 6 Klein r1       21 0.0972
```

Plote o gráfico de barras (Figura 9.3).

```
ggplot(prop.store, aes(x = store, y = prop, fill = r)) +
  geom_bar(stat = "identity")
```

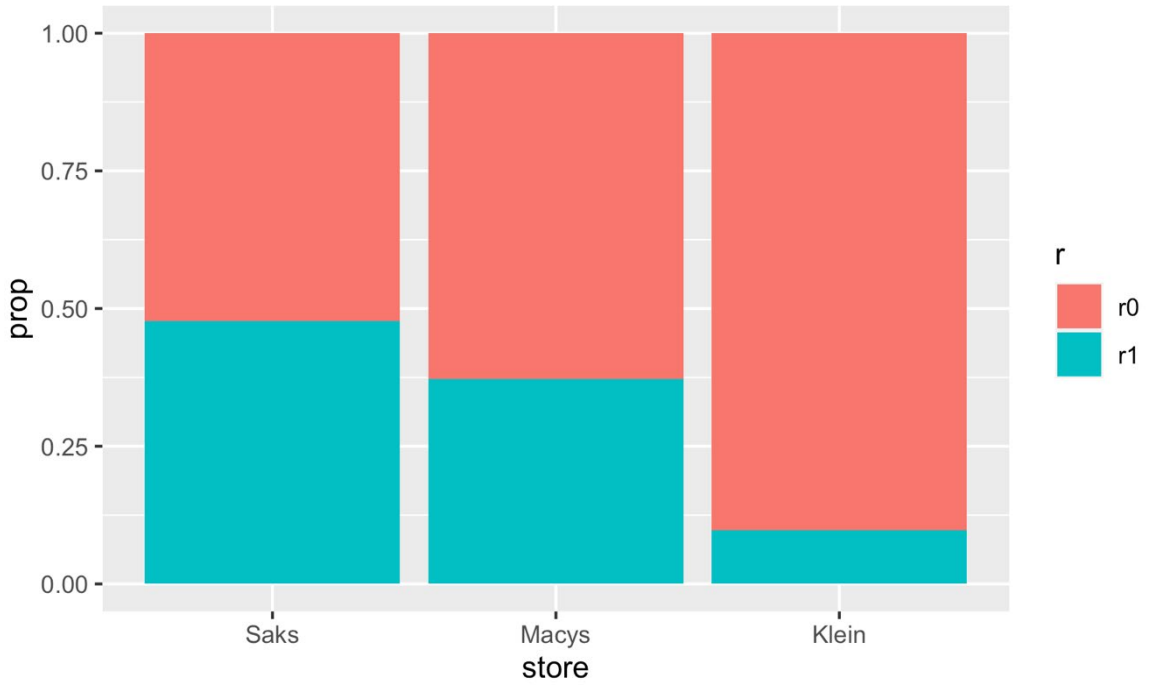

Figura 9.3: Distribuição das variantes de /r/ pós-vocálico por loja, nos dados de Labov (1972). Fonte: própria.

Crie a tabela de frequências da variável store pela VD r com as funções da instalação base do R e guarde-a num objeto chamado `tab.store`.

```
tab.store <- with(ds, table(store, r))
```

Inpecione o objeto `tab.store`.

```
tab.store
```

```
##        r
## store    r0  r1
##   Saks   93  85
##   Macys 211 125
##   Klein 195  21
```

Como já visto nas Lições 4 e 5, há mais apagamento de /r/ na S. Klein (90%), relativamente menos na Macy's (63%) e ainda menos na Saks (52%). Aplique a função de qui-quadrado à tabela `tab.store` e guarde o resultado num objeto chamado `x2.store`.

```
x2.store <- chisq.test(tab.store)
```

Veja o resultado do teste de qui-quadrado.

```
x2.store
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab.store
## X-squared = 74.142, df = 2, p-value < 2.2e-16
```

O R nos informa que o qui-quadrado é 74,1, com 2 graus de liberdade e significância menor do que 2.2e-16. Isso, contudo, é uma medida global da distribuição – lembre-se que o qui-quadrado é a *soma* das diferenças entre valores observados e esperados elevadas ao quadrado e divididas pelo valor esperado.

Vamos guardar novamente os valores observados e esperados em dois objetos. Guarde os valores observados do teste `x2.store` num objeto chamado `O`.

```
O <- x2.store$observed
```

Guarde os valores esperados do teste `x2.store` num objeto chamado `E`.

```
E <- x2.store$expected
```

Com a fórmula sum((O-E)^2/E), computamos acima o valor de qui-quadrado da tabela. Se tirarmos a função `sum()` da fórmula, teremos o valor de qui-quadrado por célula. Digite então `(O - E) ^ 2 / E` para ver o resultado.

```
(O - E) ^ 2 / E
```

```
##        r
## store          r0        r1
##   Saks   6.757375 14.597101
##   Macys  1.518742  3.280745
##   Klein 15.185220 32.802705
```

Quanto maior o valor de qui-quadrado, maior é a diferença entre o valor observado e o valor esperado. Para a distribuição dos dados de /r/ pelas lojas, vemos que a maior diferença está na S. Klein (valores acima de $\chi^2 = 15$), e a segunda maior diferença está na Saks (valores acima de $\chi^2 = 6$).

Para mais bem visualizar essa diferença, vamos fazer uma linha horizontal no gráfico de barras que indica a proporção esperada de r1, 32%. Ao comando já usado para plotar o presente gráfico, vamos adicionar a função `geom_hline()`, que plota uma linha horizontal na figura. Como argumento de `geom_hline()`, estabeleça yintercept = 0.32 (para que seja uma linha horizontal na altura 0.32).

```
ggplot(prop.store, aes(x = store, y = prop, fill = r)) +
  geom_bar(stat = "identity") +
  geom_hline(yintercept = 0.32)
```

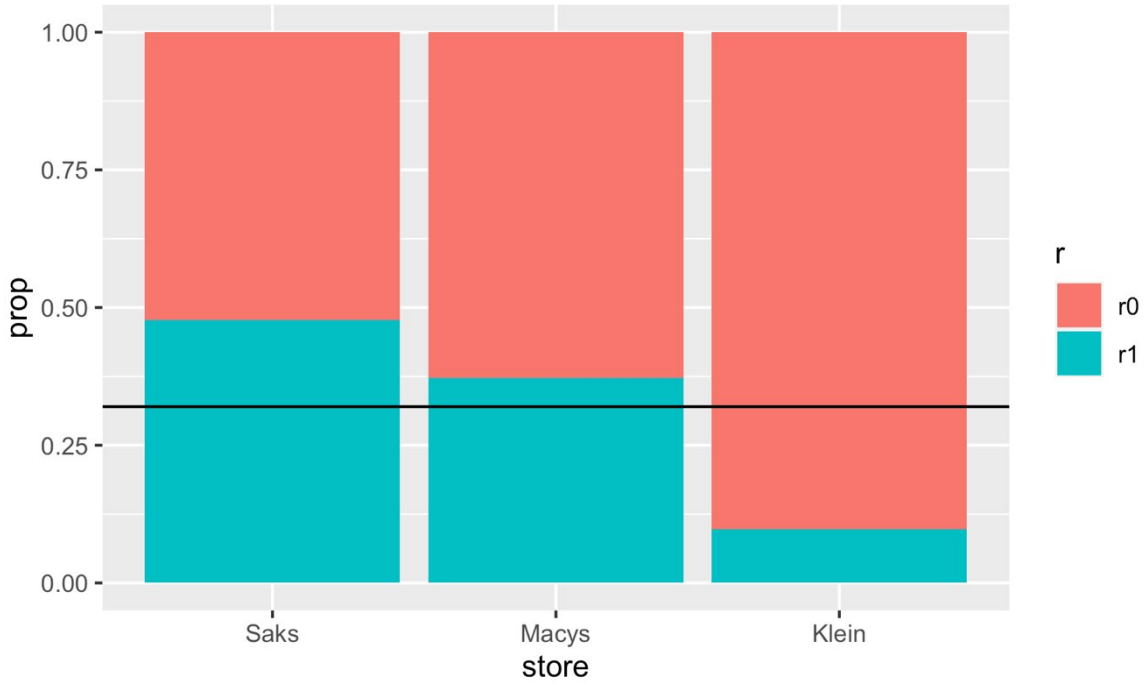

Figura 9.4: Distribuição das variantes de /r/ pós-vocálico por loja, nos dados de Labov (1972), com indicação da proporção geral de r1. Fonte: própria.

Pela , também vemos que as maiores diferenças estão nas proporções na S. Klein e na Saks, e que a proporção de r1 na Macy's se aproxima mais da proporção esperada. Os valores de qui-quadrado expressam justamente essas diferenças.

Outro modo de medir essas diferenças é pelo valor dos resíduos. Os resíduos são a diferença entre o valor observado e o valor esperado, e também foram guardados no resultado do teste. Para ver esses valores, digite `x2.store$residuals`.

```
x2.store$residuals
```

```
##        r
## store          r0        r1
##   Saks  -2.599495  3.820615
##   Macys -1.232372  1.811283
##   Klein  3.896822 -5.727365
```

Os valores de resíduos aqui são calculados pelo método de Pearson – `(O - E) / sqrt(E)`. Como esses valores não são elevados ao quadrado, mantém-se o sinal positivo ou negativo da diferença. Veja que o valor do resíduo de r1 para S. Klein é -5,727365, ou seja, o valor observado foi abaixo do esperado, enquanto os valores de resíduos de r1 para Macy's e Saks são positivos (1,811283 e 3,820615), acima da proporção esperada.

Tantos os valores de qui-quadrado por célula quanto os resíduos indicam que o maior responsável pelas diferenças são as proporções de S. Klein. Pela (e pelas estatísticas), vemos que a diferença de proporções entre Saks e Macy's é relativamente menor. Poderíamos nos perguntar se há diferença significativa entre essas duas lojas.

Acima, fizemos o teste de qui-quadrado sobre diferenças entre as três lojas por meio de `chisq.test(tab.store)`. Com que linha de comando podemos testar se há diferença apenas entre Saks e Macy's?

- `chisq.test(tab.store[1:2, ])`
- `chisq.test(tab.store[1:2])`
- `chisq.test(tab.store[1,2])`

Como `tab.store` é uma tabela, ela tem linhas e colunas, que podem ser acessadas por meio dos colchetes [ ]. Da tabela, podemos analisar apenas as duas primeiras linhas 1:2, referentes a Saks e a Macy's. Deixamos o índice de coluna vazio. Faça então o teste de qui-quadrado apenas com as proporções das lojas Saks e Macy's. Não se preocupe em guardar o resultado num objeto.

```
chisq.test(tab.store[1:2, ])
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab.store[1:2, ]
## X-squared = 4.9323, df = 1, p-value = 0.02636
```

Pelo resultado, o que o pesquisador pode concluir?

- Há diferença significativa entre as proporções de r0 de Macys e de Saks.
- Não há diferença significativa entre as proporções de r0 de Macys e de Saks.

Qual é o valor de qui-quadrado?

```
4.9323
```

```
## [1] 4.9323
```

Quantos graus de liberdade há na tabela `tab.store[1:2, ]`?

```
1
```

```
## [1] 1
```

Volte à tabela de distribuição do qui-quadrado. Em que ponto se localiza o valor χ^2 = 4,9323 para um grau de liberdade?

- entre as probabilidades 0,01 e 0,001
- entre as probabilidades 0,05 e 0,01
- entre as probabilidades 0,10 e 0,05

Faz sentido, não? O R calculou a probabilidade 0,02636, que está justamente entre 0,05 e 0,01 com um grau de liberdade.

É importante ainda mencionar que os valores de qui-quadrado e de significância são sensíveis ao tamanho da amostra. Digamos que a amostra de dados tivesse sido 20 vezes menor. Digite `tab.store/20` para ver como ficaria a distribuição dos dados de /r/ por loja neste cenário.

```
tab.store/20
```

```
##        r
## store      r0    r1
##   Saks   4.65  4.25
##   Macys 10.55  6.25
##   Klein  9.75  1.05
```

As proporções continuariam as mesmas, pois dividimos todos os valores igualmente por 20. Mas faça o teste de qui-quadrado sobre este conjunto de dados, tab.store/20.

```
chisq.test(tab.store/20)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab.store/20
## X-squared = 3.7071, df = 2, p-value = 0.1567
```

Embora as proporções sejam as mesmas, o resultado agora é a uma diferença não significativa entre as lojas. O motivo para isso é simples: com um menor número de dados, a chance de aleatoriedade é muito maior, de modo que é mais difícil rejeitar a hipótese nula.

Por fim, é importante saber como apresentar os resultados de testes de qui-quadrado. A notação convencional é $\chi^2 = 74{,}14(2)$, $p < 0{,}001$, que se lê: "Qui-quadrado igual a 74,14, com dois graus de liberdade e p menor do que 0,001." O símbolo χ deve ser representado pela letra grega chi e o quadrado é o número 2 sobrescrito. Ainda que o valor de significância fornecido pelo R seja um número muito menor – aqui, 2.2e-16 –, não é necessário reportar a significância com tanta precisão. Qualquer valor abaixo de 0,001 pode ser reportado com $p < 0{,}001$. Textualmente, o resultado pode ser assim descrito: "Um teste de qui-quadrado, com o objetivo de verificar se há diferenças entre as proporções de apagamento de /r/ nas três lojas de departamento em Nova Iorque, indica que há diferenças significativas entre as lojas ($\chi^2 = 74{,}14(2)$, $p < 0{,}001$)." O pesquisador então pode se dedicar à explicação de tal fato.

Para saber mais

Recomendo a leitura do capítulo 8 de Dalgaard (2008), das páginas 150–177 de Gries (2019) e do capítulo 9 de Levshina (2015).

Exercícios

1. Carregue o pacote tidyverse.
2. Defina como diretório de trabalho aquele que contém a planilha DadosRT.csv.
3. Carregue os dados da planilha DadosRT.csv num objeto chamado `dados` com a função `read_csv()`. Para tanto, defina as variáveis como factor, exceto IDADE (definida como integer), e INDICE.SOCIO e FREQUENCIA (definidas como double).
4. Reorganize os níveis da variável ORIGEM.PAIS, colocando "SPcapital" como primeiro nível. Não se esqueça de guardar o resultado no mesmo vetor.
5. Cheque os níveis da variável ORIGEM.PAIS.
6. Cheque a estrutura de dados para verificar se foram carregados corretamente.
7. A planilha DadosRT.csv contém dados sobre a variação na pronúncia de /r/ em coda silábica (como em "porta" e "mulher") na fala de 118 paulistanos. Foram excluídos os dados de apagamento de /r/ e mantidos os dados das variantes retroflexa e tepe. Aplique a função `View()` sobre o dataframe e dedique um tempo para se familiarizar com o conjunto de dados.
8. No R, ao criar um objeto, é possível visualizá-lo imediatamente colocando-se parênteses () em volta de toda a expressão. Faça uma tabela de frequência da variável dependente (VD), usando as funções da instalação base do R, e guarde-a num objeto chamado `tab.RT`. Em seguida, coloque parênteses em volta de toda a linha de comando para visualizar o conteúdo de `tab.RT`.
9. Faça uma tabela de proporções chamada `prop.RT` com as proporções da variável dependente, usando a função da instalação base do R. Envolva a linha de comando com parênteses para visualizar o conteúdo de `prop.RT`.
10. Faça um teste de proporções dos dados de VD sob a hipótese nula de que a proporção de *tepes* é igual a 80%.

11. Faça uma tabela de frequências dos dados da variável SEXO.GENERO por VD, usando as funções da instalação base do R. Guarde-a num objeto chamado `tab.sexo`. Visualize o conteúdo de `tab.sexo` imediatamente.
12. Faça uma tabela de proporções por linha dos dados da variável SEXO.GENERO por VD, usando a função da instalação base do R, e guarde-a num objeto chamado `prop.sexo`. Visualize o conteúdo de `prop.sexo` imediatamente.
13. Com uso das funções do tidyverse, faça um gráfico de barras das proporções de uso de retroflexo e de tepe por parte de homens e mulheres, e que contenha uma linha horizontal que representa a proporção de tepes na comunidade como um todo. Para tanto, com auxílio do pipe e a partir do dataframe `dados`, (i) compute as frequências da VD pela VI; (ii) agrupe os dados pela variável SEXO.GENERO; (iii) compute as proporções de retroflexo e de tepe por sexo do falante, nomeando a coluna de proporções como prop; (iv) defina os parâmetros estéticos x, y e fill dentro da função `ggplot()`; (v) use a geometria de barras apenas com o argumento stat = "identity"; e (vi) adicione a função `geom_hline()`, com argumento yintercept = 0.72.
14. Faça um teste de qui-quadrado para verificar se a diferença no uso de retroflexo entre homens e mulheres é significativa. Guarde os resultados do teste num objeto chamado `x2.sexo`.
15. Visualize o resultado do teste de qui-quadrado feito acima.
16. Qual é o valor de qui-quadrado?
17. O que significa df?
    a. degrees of freedom
    b. degrees of fame
    c. degrees of force
    d. degrees of fashion
18. Qual é o valor de significância calculado para este teste?
    a. $p < 0{,}001$
    b. $0{,}001 < p < 0{,}01$

c. $0{,}05 > p > 0{,}01$

d. $p > 0{,}1$

19. A qual conclusão o pesquisador *não* pode chegar diante desse resultado?
    a. a diferença entre homens e mulheres no uso de retroflexo não é significativa
    b. homens paulistanos tendem a usar retroflexos mais frequentemente do que as mulheres paulistanas
    c. o uso de retroflexos por parte de mulheres paulistanas é significativamente mais baixo que o dos homens paulistanos
    d. homens e mulheres da cidade de São Paulo empregam o retroflexo em proporções diferentes
20. Faça uma tabela de frequências dos dados da variável ORIGEM.PAIS por VD, usando as funções da instalação base do R, e guarde-a num objeto chamado `tab.origem`. Visualize a tabela.
21. Nesta distribuição, há quantos graus de liberdade?
22. Faça uma tabela de proporções por linha dos dados da variável ORIGEM.PAIS por VD, usando a função da instalação base do R, e guarde-a num objeto chamado `prop.origem`. Visualize a tabela.
23. Usando as funções do tidyverse, faça um gráfico de barras das proporções de uso de retroflexo e de tepe, de acordo com a origem dos pais, com as mesmas especificações do gráfico plotado para a variável SEXO.GENERO.
24. Da , quais falantes mais usam retroflexo?
    a. paulistanos filhos de paulistanos
    b. paulistanos filhos de interioranos
    c. paulistanos filhos de nordestinos
    d. paulistanos filhos de estrangeiros
    e. paulistanos filhos de pais de origem mista

25. Faça um teste de qui-quadrado sobre os dados de ORIGEM.PAIS por VD, e guarde os resultados num objeto chamado `x2.origem`. Visualize os resultados.
26. A qual conclusão um pesquisador pode chegar a partir do gráfico de barras e do resultado do teste acima?
    a. a variação na pronúncia de /r/ em coda, por parte de paulistanos, correlaciona-se com a origem dos pais
    b. paulistanos cujos pais vieram do interior são os que usam o retroflexo mais frequentemente
    c. a diferença entre o uso de retroflexo por parte de filhos de interioranos e filhos de paulistanos é significativa
    d. a proporção de uso do retroflexo por parte de filhos de estrangeiros é acima da média dos paulistanos
27. Da figura, percebe-se que as barras de proporção para paulistanos cujos pais são de origem mista e de origem paulistana são bastante próximas. Faça um teste de qui-quadrado para verificar se tal diferença é significativa. Não se preocupe em guardar o resultado.
28. A diferença das proporções de retroflexos entre os dois grupos (paulistanos cujos pais são de origem mista e paulistana) é significativa? Explique sua resposta.