

# R para Ciência de Dados

## Intervalos de Confiança e Teste de Hipóteses

Instituto de Matemática e Estatística  
Universidade Federal da Bahia

Professor: Gilberto Sassi

# Preparando o ambiente

## Durante o curso

- Usaremos nas aulas: [posit.cloud](https://posit.cloud).
  - Recomendamos instalar e usar R com versão pelo menos 4.1: [cran.r-project.org](https://cran.r-project.org).
  - usaremos o *framework* [tidyverse](https://www.tidyverse.org):
    - Instalação: `install.packages("tidyverse")`
- 

## Na sua casa

- **IDE** recomendadas: [RStudio](https://rstudio.com) e [VSCode](https://code.visualstudio.com).
  - Caso você queira usar o [VSCode](https://code.visualstudio.com), instale a extensão da linguagem R: [REditorSupport](https://marketplace.visualstudio.com/items?itemName=RobertoHernandez.REditorSupport).
- Outras linguagens interessantes: [python](https://python.org) e [julia](https://julialang.org).
  - [python](https://python.org): linguagem interpretada de propósito geral, contemporânea do R, simples e fácil de aprender.
  - [julia](https://julialang.org): linguagem interpretada para análise de dados, lançada em 2012, promete simplicidade e velocidade.

Revisão

# Revisão de Estatística Descritiva no R

## Gráficos e Tabelas

### Alguns conceitos básicos

- **População:** todos os elementos ou indivíduos alvo do estudo.
- **Amostra:** parte da população.
- **Parâmetro:** característica numérica da população. Usamos letras gregas para denotar parâmetros populacionais.
- **Estatística:** função ou *cálculo* da amostra
- **Estimativa:** característica numérica da amostra, obtida da estatística computada na amostra. Em geral, usamos uma estimativa para estimar o parâmetro populacional.
- **Variável:** *característica mensurável comum a todos os elementos da população.*

## Exemplo

- **População:** todos os eleitores nas eleições gerais de 2023.
- **Amostra:** 3.500 pessoas abordadas pelo datafolha.
- **Variável:** candidato a presidente de cada pessoa.
- **Parâmetro:** porcentagem de pessoas que escolhem Lula como presidente entre todos os eleitores.
- **Estatística:** porcentagem de pessoas que escolhem o lula
- **Estimativa:** porcentagem de pessoas que escolhem Lula como presidente entre todos os eleitores da amostra de 3.500 pessoas entrevistadas pelo datafolha.

# Revisão de Classificação de variáveis

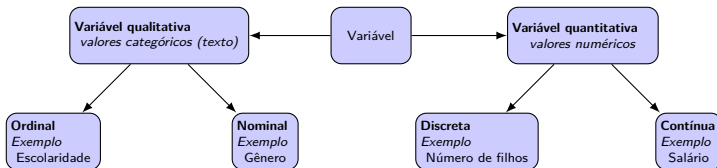


Figura 1: Classificação de variáveis.

# Revisão de tabela de distribuição de frequências

## Variável quantitativa discreta

A primeira coisa que fazemos é contar!

$X$	frequência	frequência relativa	porcentagem
$B_1$	$n_1$	$f_1$	$100 \cdot f_1 \%$
$B_2$	$n_2$	$f_2$	$100 \cdot f_2 \%$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$B_k$	$n_k$	$f_k$	$100 \cdot f_k \%$
Total	$n$	1	100%

Em que  $n$  é o tamanho da amostra.

# Revisão de tabela de distribuição de frequências

## Variável quantitativa discreta

```
dados_iris <- read_xlsx("dados/brutos/iris.xlsx")
dados_iris <- clean_names(dados_iris)

tab <- tabyl(dados_iris, especies) |>
  adorn_totals() |>
  adorn_pct_formatting(digits = 2) |>
  rename(
    "Espécies" = especies,
    "Frequência" = n,
    "Porcentagem" = percent
  )
tab
```



#	Espécies	Frequência	Porcentagem
#	setosa	50	33.33%
#	versicolor	50	33.33%
#	virginica	50	33.33%
#	Total	150	100.00%

# Revisão de tabela de distribuição de frequências

## Variável quantitativa contínua

Para variáveis quantitativas discretas com muitos valores distintos, e para variáveis quantitativas contínuas.

$X$	frequência	frequência relativa	porcentagem
$[l_0, l_1)$	$n_1$	$f_1$	$100 \cdot f_1 \%$
$[l_1, l_2)$	$n_2$	$f_2$	$100 \cdot f_2 \%$
$[l_2, l_3)$	$n_3$	$f_3$	$100 \cdot f_3 \%$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[l_{k-1}, l_k]$	$n_k$	$f_k$	$100 \cdot f_k \%$
Total	$n$	1	100%

Em que  $n$  é o tamanho da amostra.

# Revisão de tabela de distribuição de frequências

## Variável quantitativa contínua

```
dados_iris <- read_xlsx("dados/brutos/iris.xlsx")
dados_iris <- clean_names(dados_iris)

k <- floor(1 + log2(nrow(dados_iris)))
dados_iris <- dados_iris |>
  mutate(comprimento_sepala_int = cut(
    comprimento_sepala,
    breaks = k,
    include.lowest = TRUE,
    right = FALSE
  ))

tab <- tabyl(dados_iris, comprimento_sepala_int) |>
  adorn_totals() |>
  adorn_pct_formatting(digits = 2) |>
  rename(
    "Comprimento de Sépala" = comprimento_sepala_int,
    "Frequência" = n,
    "Porcentagem" = percent
  )
tab
```

#	Comprimento de Sépala	Frequência	Porcentagem
#	[4.3,4.75)	11	7.33%
#	[4.75,5.2)	30	20.00%
#	[5.2,5.65)	24	16.00%
#	[5.65,6.1)	24	16.00%
#	[6.1,6.55)	31	20.67%
#	[6.55,7)	17	11.33%
#	[7,7.45)	7	4.67%
#	[7.45,7.9]	6	4.00%
#	Total	150	100.00%

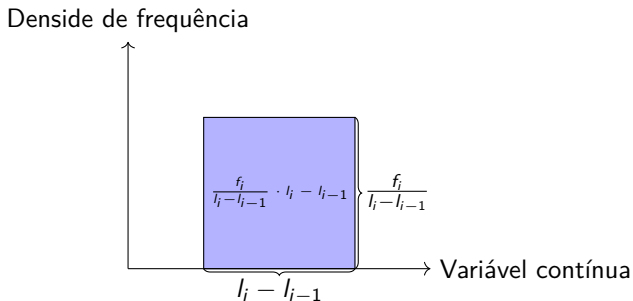
## Revisão de histograma

Para variáveis quantitativas contínuas, geralmente não construímos gráficos de barras, e sim uma figura geométrica chamada de *histograma*.

- O histograma é um gráfico de barras contíguas em que a área de cada barra é igual à frequência relativa.
- Cada faixa de valor  $[l_{i-1}, l_i)$ ,  $i = 1, \dots, n$ , será representada por um barra com área  $f_i$ ,  $i = 1, \dots, n$ .
- Como cada barra terá área igual a  $f_i$  e base  $l_i - l_{i-1}$ , e a altura de cada barra será  $\frac{f_i}{l_i - l_{i-1}}$ .
- $\frac{f_i}{l_i - l_{i-1}}$  é denominada de densidade de frequência.
- Podemos usar os seguintes parâmetros (**obrigatório o uso de apenas um deles**):
  - bins: número de intervalos no histograma (usando, por exemplo, a regra de Sturges)
  - binwidth: tamanho (ou largura) dos intervalos
  - breaks: os limites de cada intervalo

## Revisão de Histograma

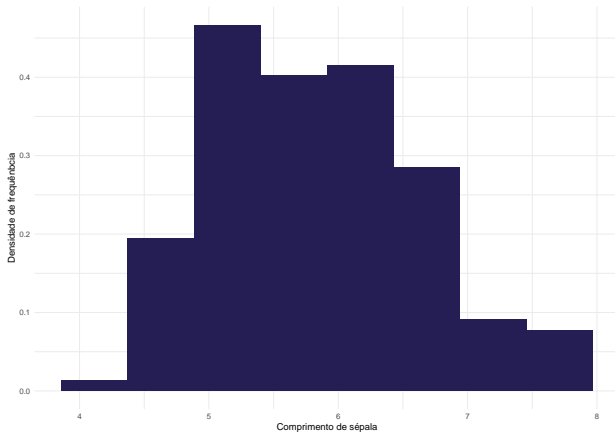
Figura 2: Representação de uma única barra de um histograma.



## Revisão de Histograma

```
dados_iris <- read_xlsx("dados/brutos/iris.xlsx")
dados_iris <- clean_names(dados_iris)

ggplot(dados_iris) +
  geom_histogram(
    aes(comprimento_sepala, after_stat(density)),
    bins = k,
    fill = "#251e54"
  ) +
  labs(
    x = "Comprimento de sépala",
    y = "Densidade de frequênbcia"
  ) +
  theme_minimal()
```





## Medidas de resumo

```
tab <- group_by(dados_iris, especie) |>
  summarise(
    media = mean(comprimento_sepala),
    dp = sd(comprimento_sepala),
    cv = dp / media,
    q1 = quantile(comprimento_sepala, probs = 1 / 4),
    q2 = quantile(comprimento_sepala, probs = 2 / 4),
    q3 = quantile(comprimento_sepala, probs = 3 / 4)
  )
tab
```

```
# # A tibble: 3 x 7
#   especie    media    dp    cv    q1    q2    q3
#   <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
# 1 setosa     5.01 0.352 0.0704 4.8    5     5.2
# 2 versicolor 5.94 0.516 0.0870 5.6    5.9   6.3
# 3 virginica  6.59 0.636 0.0965 6.22   6.5   6.9
```

# Inferência estatística

## O que faremos nesse curso?

- **Estimação pontual:** Aproximar um parâmetro.  
*Exemplo:* Estimar o teor alcoólico de uma bebida.
- **Intervalo de confiança:** Encontrar uma estimativa intervalar para um parâmetro.  
*Exemplo:* Encontrar números  $a$  e  $b$  tal que o teor alcoólico verdadeiro está entre  $a$  e  $b$  com uma probabilidade estabelecida pelo pesquisador.
- **Teste de hipóteses:** Decidir entre duas hipóteses  $H_0$  e  $H_1$  : negação de  $H_0$ .  
*Exemplo:* Decidir entre duas hipóteses:

$H_0$  : A nota média em matemática no ENEM 2021 é maior que 600,

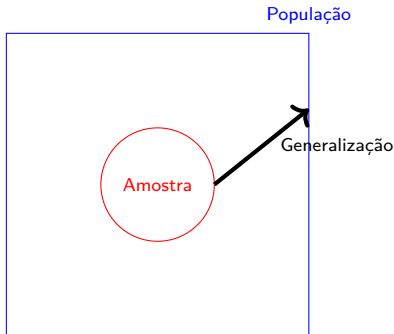
$H_1$  : A nota média em matemática no ENEM 2021 é menor ou igual 600.

Em todos estes casos, precisamos usar probabilidade.

## Por que precisamos de probabilidade

- Queremos fazer afirmações válidas para toda população.
- *inferência estatística*: generalização da **amostra** para toda **população** precisa de probabilidade.

Figura 3: Ilustração da estatística inferencial.



# Probabilidade

# Probabilidade

## **Fenômeno Aleatório**

Procedimento ou evento cujo resultado não é possível antecipar de forma determinística. Por exemplo:

- Teremos uma guerra total na Venezuela envolvendo o Brasil, Colômbia e Estados Unidos da América?
- Qual será o resultado do lançamento de um dado “justo”?

# Probabilidade

## Notação e nomes

- **Espaço amostral:** O conjunto de todos os resultados de um fenômeno aleatório.  
**Notação:**  $\Omega$
- **Evento:** Subconjunto de um espaço amostral.  
**Notação:**  $A, B, C, \dots$
- **Ponto amostral:** Um resultado possível de um fenômeno aleatório.  
**Notação:**  $\omega$ .
- **Probabilidade:** A plausibilidade de um ponto amostral  $\omega$  de  $A$  ser o resultado do fenômenos aleatório.  
**Notação:**  $P(A)$ .
- **Variável aleatória:** Função com domínio em um espaço amostra e contra-domínio no conjunto dos números reais  $X : \Omega \rightarrow \mathbb{R}$ .

## Classificação de variáveis aleatórias

- Dizemos que  $X$  é uma variável aleatória discreta, se os valores possíveis desta variável são números inteiros, geralmente resultado de contagem;
- Dizemos que  $X$  é uma variável aleatória contínua, se os valores possíveis desta variável pode ser qualquer número (incluindo aqueles por parte decimal);
- O conjunto dos valores possíveis de  $X$  representamos por  $\chi$ .



# Variável aleatória discreta

**Função de probabilidade (FP):**

$$f(x) = P(X = x)$$

**Interpretação:**  $f(x)$  pode ser interpretada como a frequência relativa em toda população de  $x$ .

---

**Para amostra**

$X$	frequência relativa
$x_1$	$f_1$
$x_2$	$f_2$
$x_3$	$f_3$
$\vdots$	$\vdots$
$x_k$	$f_k$

**Para população**

$X$	função de probabilidade
$x_1$	$f(x_1)$
$x_2$	$f(x_2)$
$x_3$	$f(x_3)$
$\vdots$	$\vdots$
$x_k$	$f(x_k)$

# Medidas de resumo para variável aleatória discreta

## Para amostra

$X$  uma variável quantitativa discreta

- **Média:**

$$\bar{X} = x_1 \cdot f_1 + \cdots + x_k \cdot f_k$$

- **Variância:**

$$\begin{aligned} \text{Var}(X) = \\ (x_1 - \bar{x})^2 \cdot f_1 + \cdots + (x_k - \bar{x})^2 \cdot f_k \end{aligned}$$

- **Desvio padrão:**

$$dp(X) = \sqrt{\text{Var}(X)}$$

- **Mediana:**

$Md$  tal que:

- $f_1 + \cdots + f_{Md} \geq 0,5$
- $f_{Md} + \cdots + f_k \leq 0,5$

## Para população

$X$  para uma variável aleatória discreta

- **Média:**

$$\mu = x_1 \cdot f(x_1) + \cdots + x_k \cdot f(x_k)$$

- **Variância:**

$$\begin{aligned} \sigma^2 = (x_1 - \mu)^2 \cdot f(x_1) + \cdots + \\ (x_k - \mu)^2 \cdot f(x_k) \end{aligned}$$

- **Desvio padrão:**

$$\sigma = \sqrt{\text{Var}(X)}$$

- **Mediana:**

$Md$  tal que:

- $f(x_1) + \cdots + f(Md) \geq 0,5$
- $f(Md) + \cdots + f(x_k) \leq 0,5$

## Distribuição Bernoulli

# Distribuição Bernoulli

Cada elemento da população pode ter **sucesso** ou **fracasso**.

**Sucesso:** caso de interesse ou mais importante.

---

Sucesso
Município tem secretaria cultura
Pessoa infectada
Pessoa alta
Bahia ganha o jogo

---

---

Fracasso
Município <b>não</b> tem secretaria cultura
Pessoa sadia
Pessoa baixa
Bahia <b>não</b> ganhou jogo

---

Precisamos descobrir a proporção (ou porcentagem) de Sucesso.

**Notação:**  $p$  é a prporção (ou porcentagem) de Sucesso.

# Parâmetros da distribuição Bernoulli

Usamos letras gregas para representar parâmetros:

- **Média populacional:**  $\mu$
  - **Variância populacional:**  $\sigma^2$
  - **Desvio padrão populacional:**  $\sigma$
- 

## Distribuição Bernoulli

- Média (populacional):  $\mu = p$
- Variância (populacional):  $\sigma^2 = p \cdot (1 - p)$
- Desvio padrão (populacional):  $\sigma = \sqrt{p \cdot (1 - p)}$

# Estimação pontual

## Distribuição Bernoulli

- 1 Definimos o sucesso.
- 2 Encontramos a estimativa de  $p$ .

---

**Variável aleatória:** transmissão (do conjunto de dados `mtcarros.xlsx`).

- 0: Carro com transmissão automática
- 1: Carro com transmissão manual (**Sucesso**)

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarros.csv")
```

```
tab <- dados_mtcarrros |>  
  summarise(prop_sucesso = mean(transmissao))  
tab
```

```
# # A tibble: 1 x 1  
#   prop_sucesso  
#           <dbl>  
# 1           0.406
```

**Variável aleatória:** Cidade realizou Conferência Municipal de Cultura? (coluna Mcul14 em munic\_amostra.xlsx).

- **Sucesso:** Sim (realizou a Conferência Municipal de Cultura).
- **Fracasso:** Não (**não** realizou a Conferência Municipal de Cultura).

Vamos criar uma nova coluna com 1 e 0.

```
munic_cultura <- read_xlsx("dados/brutos/munic_amostra.xlsx")
munic_cultura <- munic_cultura |>
  mutate(in_mcul14 = Mcul14 == "Sim")
tab <- munic_cultura |>
  summarise(prop_mcul14 = mean(in_mcul14))
tab
```

```
# # A tibble: 1 x 1
#   prop_mcul14
#         <dbl>
# 1         0.152
```

## Exercício

### Distribuição Bernoulli

Responda as seguintes perguntas:

- Qual a proporção de cidades que executaram a LAB (Mcu142 de munic\_amostra)?
- Qual a proporção de treineiros no ENEM na edição 2023 (in\_treineiro)? (Cada pessoa tem sua cidade).
- Qual a proporção de candidatas/os **sem acesso** à internet entre as/os candidatos no ENEM na edição de 2023 (q025)? (Cada pessoa tem sua cidade).
- Qual a proporção de candidatas/os que escolheram fazer a prova de inglês no ENEM na edição de 2023 (tp\_lingua)? (Cada pessoa tem sua cidade).



## Distribuição Binomial

# Distribuição Binomial

- Temos  $n$  casos
  - Cada caso pode ser **sucesso** ou **fracasso**
- 

## Parâmetros:

- Proporção de sucesso:  $p$
  - Número de casos:  $n$
  - Média:  $\mu = n \cdot p$
  - Variância:  $\sigma^2 = n \cdot p \cdot (1 - p)$
  - Desvio padrão:  $\sigma = \sqrt{n \cdot p \cdot (1 - p)}$
- 

Precisamos estimar  $p$ .

Geralmente conhecemos previamente  $n$ .

Soma de *Bernoulli* produz *Binomial*.

# Estimação pontual

## Distribuição Binomial

**Variável aleatória:** Número de semente germinado (coluna germinado de estudos\_sementes.xlsx).

```
sementes <- read_xlsx("dados/brutos/estudo_sementes.xlsx")
tab <- sementes |>
  summarise(prop = sum(germinado) / sum(numero_sementes_plantadas))
tab
```

```
# # A tibble: 1 x 1
#   prop
#   <dbl>
# 1 0.700
```

## Exercício

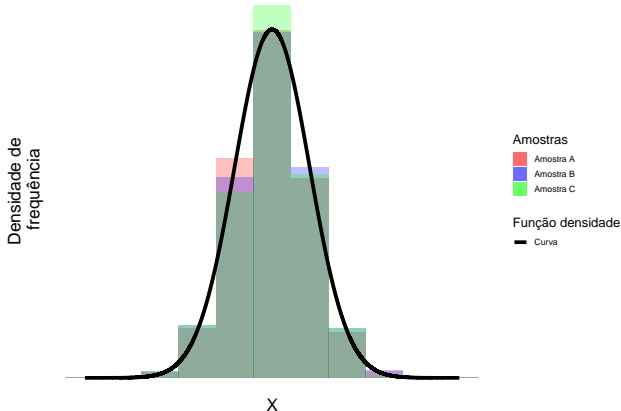
### Distribuição Binomial

Qual a proporção de email com alguma tipo de resposta em 50 campanhas de *mailing* (conjunto de dados `campanha_mailng.xlsx`)?

# Variável aleatória contínua

## Motivação

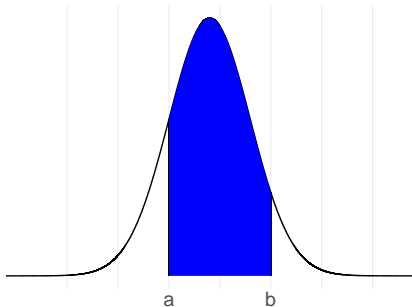
- Para cada amostra, temos um histograma;
- Queremos encontrar uma curva que aproxime bem todos os histogramas possíveis;



Chamamos a curva preta de **função densidade**.

## Propriedades de variável aleatória contínua

- Proporção de elementos da população com variável aleatória  $X$  entre  $a$  e  $b$ :  $P(a < X < b)$ .
- $P(a < X < b)$ : área sob a curva (região azul).



## Distribuição normal

# Distribuição normal

## Quando usar?

- Valores da variável aleatória concentrados em torno da média;
- Valores da variável aleatória afastados da média são pouco prováveis;
- Função densidade de probabilidade em curva em formato de sino;
- Simetria em torno da média.

Checamos isso com histograma.

---

## Parâmetros

- Média:  $\mu$
- Variância:  $\sigma^2$
- Desvio padrão:  $\sigma$



## Exemplos de aplicação

### Distribuição normal

- Altura:
  - As médias no Brasil tem em média 170cm
  - Algumas pessoas são menores que 170cm
  - Algumas pessoas são maiores que 170cm
  - poucas ficam muito longe de 170cm
- Uso de caixa eletrônico:
  - Em média, as pessoas demoram 2 minutos no caixa eletrônico
  - Algumas pessoas são mais lentas
  - Algumas pessoas são mais rápidas
  - poucas pessoas ficam longe de 2 minutos

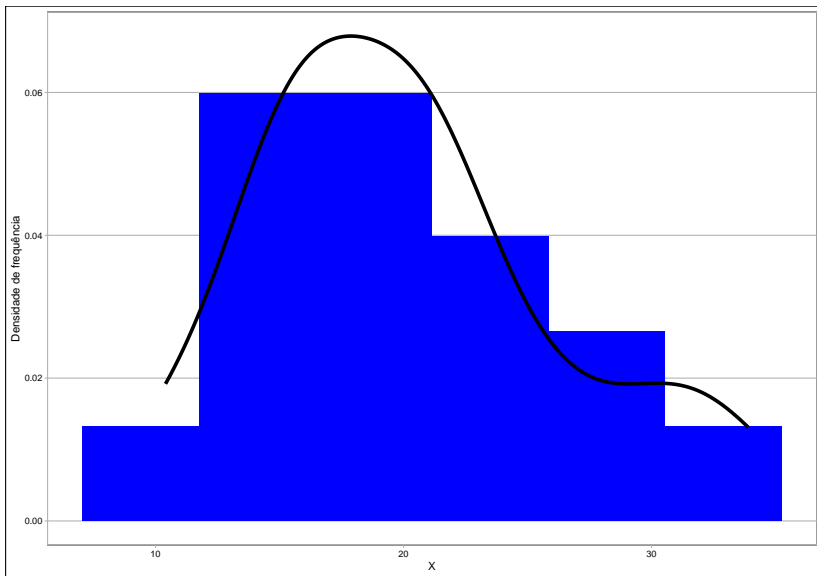
## Exemplos

### Distribuição normal

**Variável aleatória:** milhas por galão (milhas\_por\_galao em mtcarros.xlsx).

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarros.csv")
k <- ceiling(1 + log2(nrow(dados_mtcarrros)))

ggplot(dados_mtcarrros, aes(x = milhas_por_galao,
                           y = after_stat(density))) +
  geom_histogram(bins = k, fill = "blue") +
  geom_density(color = "black", linewidth = 1.5) +
  labs(x = "X", y = "Densidade de frequência") +
  theme_calc()
```



## Estimativa pontual

### Distribuição Normal

```
tab <- dados_mtcarrros |>
  summarise(media = mean(milhas_por_galao),
            dp = sd(milhas_por_galao))
tab
```

```
# # A tibble: 1 x 2
#   media    dp
#   <dbl> <dbl>
# 1  20.1  6.03
```

## Exercício

### Distribuição normal

- Cheque se as seguintes variáveis aleatórias têm distribuição normal:
  - `nu_nota_mt` do conjunto de dados do ENEM/2023 (cada pessoa tem o seu conjunto de dados);
  - `nu_nota_lc` do conjunto de dados do ENEM/2023 (cada pessoa tem o seu conjunto de dados);
  - `nu_nota_ch` do conjunto de dados do ENEM/2023 (cada pessoa tem o seu conjunto de dados);
  - `nu_nota_cn` do conjunto de dados do ENEM/2023 (cada pessoa tem o seu conjunto de dados).
- Para cada uma das variáveis acima, calcule a média e o desvio padrão.

## Intervalo de Confiança uma população

# Intervalo de Confiança

**Objetivo:** Para parâmetro  $\mu$  ( $\sigma$  e  $p$ ), encontrar  $L$  e  $U$  tal que  $L < \mu < U$  com alguma probabilidade associada  $\gamma$ .

Chamamos  $\gamma$  de **coeficiente de confiança**.

Vamos usar o pacote `statBasics`.

# Interpretação Intervalo de Confiança

O parâmetro  $\mu$  ( $\sigma$  e  $p$ ) **pode** ou **não pode** estar entre **L** e **U** do intervalo de confiança com coeficiente de confiança  $\gamma$ .

```
dados_estudos <- read_xlsx("dados/brutos/motivacao_intervalo_confianca.xlsx", sheet = 1)
dados_pop <- read_xlsx("dados/brutos/motivacao_intervalo_confianca.xlsx", sheet = 2)
media_pop <- mean(dados_pop$variavel)

tab <- dados_estudos |>
  group_by(estudo) |>
  summarise(lower_ci = ci_1pop_norm(variavel)$lower_ci,
            upper_ci = ci_1pop_norm(variavel)$upper_ci,
            media_pop = media_pop)

tab
```

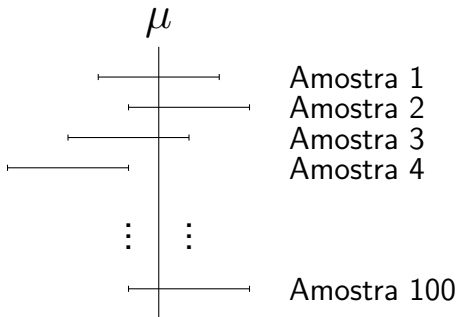


```
# # A tibble: 6 x 4
#   estudo lower_ci upper_ci media_pop
#   <chr>      <dbl>    <dbl>    <dbl>
# 1 amostra1    5.40      6.30      6.37
# 2 amostra2    4.53      6.17      6.37
# 3 amostra3    5.76      6.19      6.37
# 4 amostra4    5.74      8.29      6.37
# 5 amostra5    5.63      7.99      6.37
# 6 amostra6    4.12      8.70      6.37
```

# Interpretação Intervalo de Confiança

$\gamma\%$  das amostras vão gerar intervalos de confiança que contém o parâmetro.

Figura 4: Interpretação de intervalo de confiança.



Geralmente  $\gamma$  é 99%, 95% ou 90%.

## Intervalo de Confiança Distribuição Bernoulli

**Primeira forma:** Vetor de 1 e 0.

- **Variável aleatória:** Carro tem transmissão manual? (variável `transmissao` de `mtcarros.csv`).
- **Sucesso:** 1 (transmissão manual)
- **Fracasso:** 0 (transmissão automática)

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarros.csv")
ic_transmissao <- ci_1pop_bern(dados_mtcarrros$transmissao)
ic_transmissao
```

```
# # A tibble: 1 x 3
#   lower_ci upper_ci conf_level
#   <dbl>     <dbl>     <dbl>
# 1    0.233    0.579         0.95
```

---

A proporção de carros com transmissão manual está entre 0,233 e 0,5795 com coeficiente de confiança 95%.

## Intervalo de confiança Distribuição Bernoulli

**Segunda forma:** Número de tentativas e número de sucessos.

- **Variável aleatória:** Carro tem transmissão manual? (variável transmissão de `mtcarros.csv`).
- **Sucesso:** 1 (transmissão manual)
- **Fracasso:** 0 (transmissão automática)

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarros.csv")
n_tentativas <- nrow(dados_mtcarrros)
n_sucessos <- sum(dados_mtcarrros$transmissao)
ic_transmissao <- ci_1pop_bern(n_sucessos, n_tentativas,
                               conf_level = 0.99)

ic_transmissao
```

```
# # A tibble: 1 x 3
#   lower_ci upper_ci conf_level
#   <dbl>     <dbl>     <dbl>
# 1    0.179    0.634         0.99
```

A proporção de carros com transmissão automática está entre 0,1786 e 0,6339 com coeficiente de confiança 99%.

## Intervalo de Confiança Distribuição Bernoulli

**Pesquisa de Intenção de voto:** Eleição 2023.

- **Número de entrevistados:** 8308
- **Número de eleitores de Lula:** 4403
- **Coeficiente de Confiança:** 99%

```
eleicao_lula_22 <- ci_1pop_bern(4403, 8308, conf_level = 0.99)
eleicao_lula_22
```

```
# # A tibble: 1 x 3
#   lower_ci upper_ci conf_level
#   <dbl>    <dbl>    <dbl>
# 1    0.516    0.544    0.99
```

---

Lula teria uma proporção entre 0,5158 e 0,5441 de votos com coeficiente de 99%.

## Exercício

### Intervalo de Confiança

### Distribuição Bernoulli

Construa os seguintes intervalos de confiança:

- Proporção de candidatas/os que escolheram fazer a prova de espanhol no ENEM/2023 (tp\_lingua) com coeficiente de confiança 99%;
- Proporção de candidatas/os que **não** tem acesso a internet em casa no ENEM/2023 (q025) com coeficiente de confiança 95%;
- Proporção de treineiras/os no ENEM/2023 (in\_treineiro) com coeficiente de confiança 92,5%;
- Proporção de cidades que executaram a LAB (Mcul42 em munic\_amostra) com coeficiente de confiança 97,5%;
- Proporção de cidades que realizaram Conferência Municipais de Cultura (Mcul14 em munic\_amostra) com coeficiente de confiança 90%.

## Intervalo de Confiança Distribuição Binomial

- **Variável aleatória:** Número de semente germinado (coluna germinado de estudos\_sementes.xlsx);
- **Coefficiente de confiança:** 92,5%.

```
sementes <- read_xlsx("dados/brutos/estudo_sementes.xlsx")
n_tentativas <- sum(sementes$numero_sementes_plantadas)
n_sucessos <- sum(sementes$germinado)
ic_germinado <- ci_1pop_bern(n_sucessos, n_tentativas, 0.925)
ic_germinado
```

```
# # A tibble: 1 x 3
#   lower_ci upper_ci conf_level
#   <dbl>     <dbl>     <dbl>
# 1    0.690    0.710    0.925
```

---

A proporção de sementes geminadas está entre 0,6895 e 0,7099 com coeficiente de confiança 92,5%.



## Exercício

### Intervalo de Confiança

### Distribuição Binomial

Construa um intervalo de confiança para a proporção de resposta positiva em 50 campanhas de *mailing* com coeficiente de confiança 94% (conjunto de dados `campanha_mailng.xlsx`).

## Intervalo de Confiança para média

- Variável aleatória tem distribuição normal;
- Intervalo de confiança para média.

Função `ci_1pop_norm` do pacote `statBasics`.

## Exemplo

### Intervalo de Confiança para média

- **Variável aleatória:** milhas por galão (milhas\_por\_galao em mtcarros.csv).
- **Coeficiente de confiança:** 99%.

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarros.csv")
ic <- ci_1pop_norm(dados_mtcarrros$milhas_por_galao,
                   conf_level = 0.99)
ic
```

```
# # A tibble: 1 x 3
#   lower_ci upper_ci conf_level
#   <dbl>     <dbl>     <dbl>
# 1    17.2     23.0       0.99
```

---

Os carros fazem, em média, entre 17,17 e 23,01 milhas por galão com coeficiente de confiança 99%.

## Exercício

### Intervalo de Confiança para média

Nos exercícios abaixo, cada pessoa tem sua própria cidade.

- Encontre o Intervalo de Confiança para a média de matemática (`nu_nota_mt`) na edição ENEM/2023 com coeficiente de confiança 99% e escreva a frase com o resultado.
- Encontre o Intervalo de Confiança para a média de linguagens e código (`nu_nota_lc`) na edição ENEM/2023 com coeficiente de confiança 95% e escreva a frase com o resultado.
- Encontre o Intervalo de Confiança para a média de ciências humanas (`nu_nota_ch`) na edição ENEM/2023 com coeficiente de confiança 90% e escreva a frase com o resultado.
- Encontre o Intervalo de Confiança para a média de ciências naturais (`nu_nota_cn`) na edição ENEM/2023 com coeficiente de confiança 97,5% e escreva a frase com o resultado.

## Intervalo de Confiança para variância

- Variável aleatória tem distribuição normal;
- Intervalo de confiança para a variância.

Função `ci_1pop_norm` do pacote `statBasics`, com parâmetro `parameter='variance'`.

## Exemplo

### Intervalo de Confiança para variância

- **Variável aleatória:** milhas por galão (milhas\_por\_galao em mtcarros.csv).
- **Coeficiente de confiança:** 99%.

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarros.csv")
ic <- ci_1pop_norm(dados_mtcarrros$milhas_por_galao,
                   conf_level = 0.99, parameter = "variance")
ic
```

```
# # A tibble: 1 x 3
#   lower_ci upper_ci conf_level
#   <dbl>     <dbl>     <dbl>
# 1    20.5     77.9       0.99
```

---

Os carros fazem, em média, entre 20,47 e 77,89 milhas por galão com coeficiente de confiança 99%.

## Exercício

### Intervalo de Confiança para variância

Nos exercícios abaixo, cada pessoa tem sua própria cidade.

- Encontre o Intervalo de Confiança para o desvio padrão de matemática (`nu_nota_mt`) na edição ENEM/2023 com coeficiente de confiança 99% e escreva a frase com o resultado.
- Encontre o Intervalo de Confiança para o desvio padrão de linguagens e código (`nu_nota_lc`) na edição ENEM/2023 com coeficiente de confiança 95% e escreva a frase com o resultado.
- Encontre o Intervalo de Confiança para o desvio padrão de ciências humanas (`nu_nota_ch`) na edição ENEM/2023 com coeficiente de confiança 90% e escreva a frase com o resultado.
- Encontre o Intervalo de Confiança para o desvio padrão de ciências naturais (`nu_nota_cn`) na edição ENEM/2023 com coeficiente de confiança 97,5% e escreva a frase com o resultado.

## Teste de Hipóteses uma população



# Teste de hipóteses

## Objetivo:

Decidir entre  $H_0$  (hipótese nula) e  $H_1$  (hipótese alternativa) usando as evidências da amostra.

---

- $H_0$  é a negação de  $H_1$
- $H_1$  é a negação de  $H_0$
- $H_1$  é *aquilo que desejamos provar que é verdade*
  - $H_1$  é afirmação *extraordinária* que precisa de evidências para acreditarmos
- $H_0$  é o *padrão, valor padrão de mercado ou valor padrão do regulador* (ex. ANVISA)
  - $H_0$  é a afirmação *ordinária* que assumimos como verdade quando não temos evidência para acreditar em  $H_1$

- Decisão através de *evidência* na amostra:
    - Decisão *embasada* com *evidência*  $\implies$  hipótese alternativa  $H_1$
    - Decisão *sem evidência* ou *na dúvida*  $\implies$  hipótese nula  $H_0$
- 

Como temos uma **tendência de continuar em  $H_0$**  na ausência de *evidências*, escrevemos:

- Decisão por  $H_0$ : **Não rejeitamos  $H_0$** ;
- Decisão por  $H_1$ : **Não rejeitamos  $H_1$** .

# Teste de hipóteses

Podemos cometer dois erros ao decidir:

- **Erro tipo I** ou **Falso positivo**: Decisão por  $H_1$ , mas  $H_0$  é a verdade.  
Erro GRAVÍSSIMO!
- **Erro tipo II** ou **Falso negativo**: Decisão por  $H_0$ , mas  $H_1$  é a verdade
- **Nível de significância**:  $\alpha = P(\text{Falso positivo})$
- **Poder do teste**:  $1 - \beta = P(\text{Verdadeiro positivo})$

---

Escândalo dos *falso positivo* na Colômbia.

		Situação na população	
		$H_0$	$H_1$ (Negação de $H_0$ )
Decisão	$H_0$	Sem erro (verdadeiro negativo)	Falso negativo (Erro tipo II)
	$H_1$ (Negação de $H_0$ )	Falso positivo (Erro tipo I)	Sem erro (Verdadeiro positivo)

## Teste de hipóteses

**Objetivo:** Como  $H_1$  (positivo) é a hipótese mais importante, então queremos decidir entre  $H_0$  e  $H_1$  garantindo que:

- o *nível de significância* seja pequeno (geralmente 5%)
  - o *poder do teste* seja máximo possível
- 

- Sem evidência, continuamos acreditando em  $H_0$ .
- Com evidência, desistimos de  $H_0$  e passamos a acreditar em  $H_1$ .

Neste contexto, usamos o verbo **rejeitar** em estatística:

- Sem evidência, **não rejeitamos**  $H_0$ .
- Com evidência, **rejeitamos**  $H_0$ .

## Teste de hipóteses

### Exemplo

Em um julgamento, temos as seguintes hipóteses:

- $H_0$ : o réu é inocente
  - $H_1$ : o réu é culpado
- 

Em um julgamento, o sistema de justiça pode cometer dois erros:

- **falso positivo:** uma pessoa inocente é condenada
  - **falso negativo:** um pessoa culpada é inocentada
- 

Em um julgamento, o sistema de justiça usa a seguinte regra de decisão:

- **réu é culpado:** apenas se tiver *evidências* fortes e concretas
- **réu é inocente:** na dúvida ou na ausência de *evidências*

# Teste de Hipóteses

## Como decidir?

Hipóteses nula e alternativa geralmente são *declarações matemáticas* envolvendo parâmetros.

**Ideia:** Calculamos uma distância entre a estimativa e o valor do parâmetro quando a hipótese nula é verdade.

- Se essa distância for pequena, decidimos por  $H_0$
- Se essa distância for grande, decidimos por  $H_1$

Chamamos esta distância de **estatística de teste**.

---

Existem duas formas de determinar o que é *pequeno* ou *grande* (**extrema**):

- 1 **Procedimento Geral de Testes de Hipóteses** ou **Procedimento de Neymann-Pearson**
- 2 **valor-p** (*p-value* em inglês)

## Teste de Hipóteses

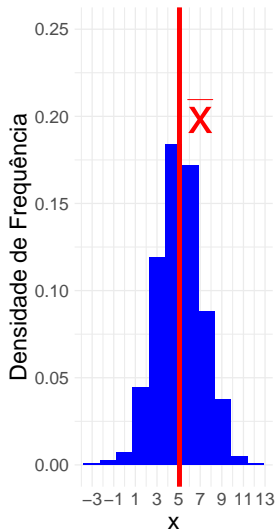
### Como decidir?

**População:**  $N(\mu, 4)$ .

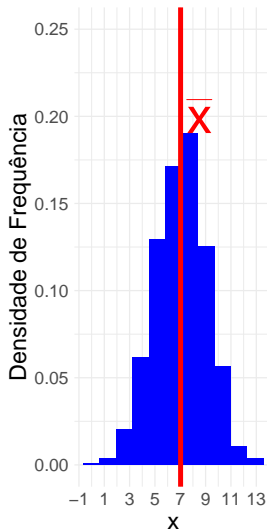
- Hipóteses:  $H_0 : \mu = 5$  contra  $H_1 : \mu \neq 5$ .
- Regra de Decisão: se  $4,80 \leq \bar{x} \leq 5,20$ , então não rejeitamos  $H_0$ .



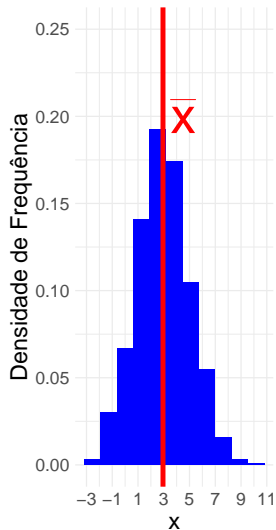
Não rejeitamos  $H_0$



Rejeitamos  $H_0$



Rejeitamos  $H_0$



# Teste de Hipóteses

Erros decaem quando o tamanho amostral aumenta.

**População:**  $N(\mu, 4)$ .

- Hipóteses:  $H_0 : \mu = 5$  contra  $H_1 : \mu \neq 5$ .
- Regra de Decisão: se  $4,80 \leq \bar{x} \leq 5,20$ , então não rejeitamos  $H_0$ .

Tabela 7: Porcentagens de falso positivo e falso negativo diminuem quando o tamanho da amostra aumenta.

Tamanho amostral	$\alpha$	$\beta(\mu = 4, 4)$	$\beta(\mu = 5, 5)$
n = 25	0,6170751	0,1359051	0,1865682
n = 50	0,4795001	0,0763107	0,1377580
n = 75	0,3864762	0,0413663	0,0957471
n = 100	0,3173105	0,0227185	0,0665746
n = 250	0,1138463	0,0007827	0,0088530
n = 500	0,0253473	0,0000039	0,0003981
n = 750	0,0061699	0,0000000	0,0000200
n = 1000	0,0015654	0,0000000	0,0000011

# Procedimento de Neymann-Pearson

## Etapas:

- 1 Estabeleça  $H_0$  e  $H_1$
- 2 Estaleça o (máximo) nível de significância
- 3 Encontre a *região crítica* (conjunto onde a *estatística de teste* é grande)
- 4 Verifique se a *estatística de teste* está na *região crítica*

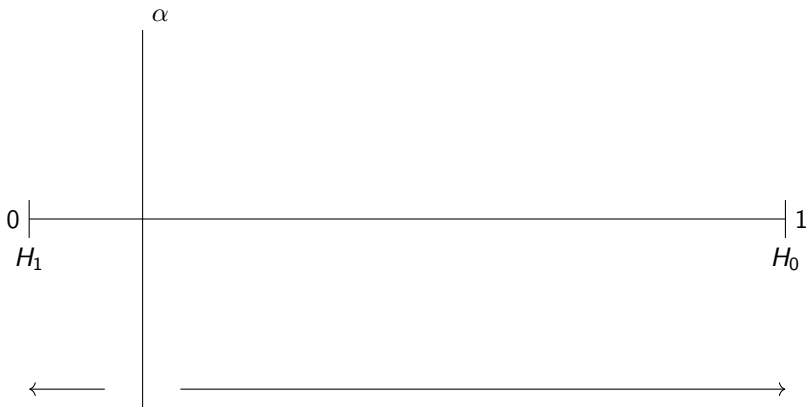
A região crítica é construída usando o nível de significância.

Para detalhes, consulte [Montgomery & Runger. Applied statistics and probability for engineers.](#)

## valor-p ou nível crítico p-value

- Valor-p é uma *medida de evidência* contra o hipótese nula.
- Valor-p **NÃO É A PROBABILIDADE DO FALSO POSITIVO.**
- Para cada amostra, temos um **valor-p diferente**.
- **Formalmente:** probabilidade de coletar uma outra amostra (de mesmo tamanho) com *estatística de teste* **mais extrema** do que a amostra que eu tenho se a hipótese nula é verdadeira.
- Rejeitamos  $H_0$  se o valor-p for menor que o nível de significância.

Rejeitamos o valor-p menor que nível de significância:  $p < \alpha$ .



## valor-p ou nível crítico p-value

Para cada amostra, temos um valor-p diferente.

O valor-p ( $p$ ) **pode ser pequeno** ou **pode ser grande**.

---

**Suposições:** Distribuição normal e  $H_0 : \mu = 20$  é verdade. Vamos usar  $\alpha = 5\%$ .

```
dados <- read_xlsx("dados/brutos/motivacao_valor_p.xlsx")

tab <- group_by(dados, amostragem) |>
  summarise(valor_p = ht_1pop_mean(amostras, mu = 20)$p_value)
print(tab)
```

```
# # A tibble: 6 x 2
#   amostragem valor_p
#   <chr>         <dbl>
# 1 amostra 1    0.0289
# 2 amostra 2    0.000306
# 3 amostra 3    0.0381
# 4 amostra 4    0.226
# 5 amostra 5    0.184
# 6 amostra 6    0.441
```

## valor-p ou nível crítico p-value

Se  $H_0$  é verdade, aproximadamente  $\alpha\%$  das amostras produzem o falso positivo quando usamos o valor-p.

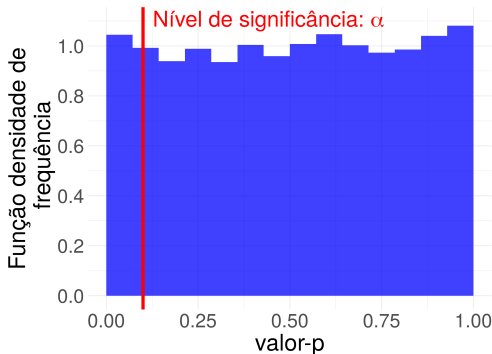


Figura 5: Histograma de valor-p de 1000 amostras quando  $H_0$  é verdade.



## valor-p ou nível crítico p-value

### Etapas:

- 1 Estabeleça  $H_0$  e  $H_1$
- 2 Estaleça o (máximo) nível de significância
- 3 Calcule o *valor-p*
- 4 Verifique se o *valor-p* é menor que *nível de significância*

Para detalhes, consulte [Montgomery & Runger. Applied statistics and probability for engineers.](#)

## Teste t para média

- Hipóteses sobre a média da população ( $\mu$ );
- Variável aleatória tem distribuição normal.

No pacote `statBasics`: `ht_1pop_mean`.

---

Testes de hipóteses deste curso:

- Teste unilateral à esquerda:  $H_1 : \mu < \mu_0$ 
  - `alternative = 'less'`
- Teste unilateral à direita:  $H_1 : \mu > \mu_0$ 
  - `alternative = 'greater'`
- Teste bilateral:  $H_1 : \mu \neq \mu_0$ 
  - `alternative = 'two.sided'` - valor padrão

Especificamos  $\mu_0$  como parâmetro `mu`.

## Teste t para média

Temos evidência para afirmar que os carros americanos conseguem fazer no máximo 15 milhas por galão, em média, ao nível de significância 1%?

- $H_0$  (negação de  $H_1$ ):  $\mu \geq 15$
- $H_1$  (o que queremos provar):  $\mu < 15$

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarrros.csv")
ht_milhas_galao <- ht_1pop_mean(
  dados_mtcarrros$milhas_por_galao,
  mu = 15,
  alternative = "less",
  sig_level = 0.01
)
ht_milhas_galao
```

```
# # A tibble: 1 x 7
#   statistic p_value critical_value critical_region alternative    mu sig_level
#   <dbl>    <dbl>         <dbl> <chr>          <chr>         <dbl>    <dbl>
# 1     4.78     1.00        -2.45 (-Inf, -2)    less          15     0.01
```

---

Não tem evidência para afirmar que os carros americanos fazer no máximo 15 milhas por galão, em média, ao nível de significância 1%.

## Exercício

### Teste t para média

Responda estas perguntas ao nível de significância 1%:

- As/os candidatas/os do ENEM/2023 tiraram nota em matemática (`nu_nota_mt`) maior que 650, em média?
- As/os candidatas/os do ENEM/2023 tiraram nota em ciências naturais (`nu_nota_cn`) menor que 400, em média?
- As/os candidatas/os do ENEM/2023 tiraram nota em ciências humanas (`nu_nota_ch`) diferente de 500, em média?
- As/os candidatas/os do ENEM/2023 tiraram nota em línguas e códigos (`nu_nota_lc`) maior que 900, em média?

Lembre que cada pessoa tem sua própria cidade.

## Teste z para proporção

- Hipóteses sobre a proporção de sucessos ( $p$ );
  - Variável aleatória tem distribuição Bernoulli ou distribuição binomial.
- 

Testes de hipóteses deste curso:

- Teste unilateral à esquerda:  $H_1 : p < p_0$ 
  - `alternative = 'less'`
- Teste unilateral à direita:  $H_1 : p > p_0$ 
  - `alternative = 'greater'`
- Teste bilateral:  $H_1 : p \neq p_0$ 
  - `alternative = 'two.sided'` - valor padrão

Especificamos  $p_0$  como parâmetro `proportion`.

## Teste z para proporção Distribuição Bernoulli

Temos evidência para afirmar que a proporção de carros americanos com transmissão manual é maior que 25% ao nível de significância 1%?

- $H_0$  (negação de  $H_1$ ):  $p \leq 0,25$
- $H_1$  (o que desejamos provar):  $p > 0,25$

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarrros.csv")
teste_transmissao <- ht_1pop_prop(
  dados_mtcarrros$transmissao,
  proportion = 0.25,
  alternative = 'greater',
  sig_level = 0.01
)
teste_transmissao
```

```
# # A tibble: 1 x 7
#   statistic p_value critical_value critical_region alternative proportion
#   <dbl>    <dbl>          <dbl> <chr>          <chr>          <dbl>
# 1      2.04 0.0206          2.33 (2.326, Inf)    greater          0.25
# # i 1 more variable: sig_level <dbl>
```

---

Ao nível de significância 1%, não temos evidência para afirmar que a proporção de carros americanos com transmissão automática é maior 25%.



## Exercício

### Teste z para proporção

### Distribuição Bernoulli

Responde as seguintes perguntas ao nível de significância 5%:

- A maioria de cidades executou a LAB (`Mcul42` de `munic_amostra`) em 2021?
- Temos evidência para afirmar que um terço das/os candidatas/os do ENEM/2023 eram treineiros (`in_treineiro`)? (cada pessoa tem sua cidade)
- Mais de três quartos das/os candidatas/os do ENEM/2023 **tem** acesso à internet (`q_025`)?
- Mais de um quarto das candidatas/os do ENEM/2023 escolhem fazer a prova em Espanhol (`tp_lingua`)?

## Teste z para proporção

### Distribuição Binomial

- **Variável aleatória:** Número de semente germinado (coluna germinado de estudos\_sementes.xlsx)
- **Nível de significância:** 2,5%
- A maioria das sementes germinaram?
  - $H_0 : p \leq 0,5$
  - $H_1 : p > 0,5$  (o que desejamos provar)

```
sementes <- read_xlsx("dados/brutos/estudo_sementes.xlsx")
n_tentativas <- sum(sementes$numero_sementes_plantadas)
n_sucessos <- sum(sementes$germinado)
teste_germinado <- ht_1pop_prop(
  n_sucessos, n_tentativas, proportion = 0.5,
  alternative = "greater", sig_level = 0.025
)
teste_germinado
```

```
# # A tibble: 1 x 7
#   statistic p_value critical_value critical_region alternative proportion
#   <dbl>    <dbl>         <dbl> <chr>          <chr>          <dbl>
# 1      34.8      0           1.96 (1.960, Inf) greater          0.5
# # i 1 more variable: sig_level <dbl>
```

---

Ao nível de significância 2,5%, a maior parte das sementes germinaram.

# Teste qui-quadrado para variância

## Distribuição normal

- Hipóteses sobre a variância da população ( $\sigma$ );
- Variável aleatória tem distribuição normal.

No pacote `statBasics`: `ht_1pop_var`.

---

Testes de hipóteses deste curso:

- Teste unilateral à esquerda:  $H_1 : \sigma^2 < \sigma_0^2$ 
  - `alternative = 'less'`
- Teste unilateral à direita:  $H_1 : \sigma^2 > \sigma_0^2$ 
  - `alternative = 'greater'`
- Teste bilateral:  $H_1 : \sigma^2 \neq \sigma_0^2$ 
  - `alternative = 'two.sided'` - valor padrão

Especificamos  $\sigma_0^2$  como parâmetro `sigma`.

## Teste qui-quadrado para variância

### Distribuição normal

Temos evidência para afirmar o desvio padrão da distância percorrida por galão nos carros americanos é menor 2 (milhas por galão), ao nível de significância 5%?

- $H_0$  (negação de  $H_1$ ):  $\sigma^2 \geq 2^2$
- $H_1$  (o que queremos provar):  $\sigma^2 < 2^2$

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarrros.csv")
ht_milhas_galao <- ht_1pop_var(
  dados_mtcarrros$milhas_por_galao,
  alternative = "less",
  sigma = 2,
  sig_level = 0.05
)
ht_milhas_galao
```

```
# # A tibble: 1 x 7
#   statistic p_value critical_value critical_region alternative sigma sig_level
#   <dbl>     <dbl>         <dbl> <chr>          <chr>         <dbl>     <dbl>
# 1      282.         1          19.3 (0, 19)      less          2         0.05
```

---

Ao nível de significância 5%, não temos evidência para afirmar o desvio padrão da distância percorrida por galão nos carros americanos é menor 2 (milhas por galão).

## Exercício

### Teste qui-quadrado para variância

### Distribuição normal

Responda as seguintes perguntas, ao nível de significância 1%:

- O desvio padrão de matemática no ENEM/2023 ( $\text{nu\_nota\_mt}$ ) é diferente de 100?
- O desvio padrão de linguagens e código no ENEM/2023 ( $\text{nu\_nota\_lc}$ ) é menor que 100?
- O desvio padrão de ciências naturais no ENEM/2023 ( $\text{nu\_nota\_cn}$ ) é maior que 100?
- O desvio padrão de ciências humanas no ENEM/2023 ( $\text{nu\_nota\_lc}$ ) é maior que 150?

---

Cada pessoa tem sua própria cidade.

Intervalo de Confiança  
Teste de Hipóteses  
duas populações



# Experimento Comparativo

## Experimento completamente aleatório

Medimos uma mesma variável em duas populações independentes.

- ① População 1
- ② População 2
- ③ As duas populações são independentes

Se decidirmos por  $H_1$ , temos uma relação de *causa-e-efeito* .

---

## Estudo observacional

- ① Acompanhamos cada elemento da amostra *antes* e *depois* de uma *intervenção*.
- ② As duas populações não são independentes.
- ③ Teste t pareado.

## Comparação de variâncias

Antes de comparar  $\mu_1$  e  $\mu_2$ , precisamos verificar se  $\sigma_1 = \sigma_2$

---

- População 1:  $N(\mu_1, \sigma_1^2)$
- População 2:  $N(\mu_2, \sigma_2^2)$
- Teste de Hipóteses envolvendo  $\sigma_1$  e  $\sigma_2$

No pacote statBasics: `ht_2pop_var`.

---

Testes de hipóteses deste curso:

- Teste bilateral:  $H_1 : \sigma_1 \neq \sigma_2$ 
  - `alternative = 'two.sided'`- valor padrão
- Teste unilateral à esquerda:  $H_1 : \sigma_1 < \sigma_2$ 
  - `alternative = 'less'` (*Atenção para ordem das populações*)
- Teste unilateral à direita:  $H_1 : \sigma_1 > \sigma_2$ 
  - `alternative = 'greater'` (*Atenção para ordem das populações*)

Especificamos `ratio` fornecendo  $\frac{\sigma_1}{\sigma_2}$ . Valor padrão: `ratio = 1` (neste caso, estamos testando a igualdade).

## Comparação de variâncias

Ao nível de significância 1%, existe diferença entre os desvios padrões da distância percorrida em milhas por um galão entre carros com transmissão manual e automática.

- **Variável aleatória:** Milhas por galão
- **População 1:** carros com transmissão manual (`transmissao == 1`)
- **População 2:** carros com transmissão manual (`transmissao == 0`)

---

```
dados_mtcarrros <- read_csv2("dados/brutos/mtcarrros.csv")
carros_manuais <- dados_mtcarrros |> filter(transmissao == 1)
carros_auto <- dados_mtcarrros |> filter(transmissao == 0)
comparacao_var <- ht_2pop_var(
  carros_manuais$milhas_por_galao,
  carros_auto$milhas_por_galao,
  ratio = 1,
  sig_level = 0.01
)
comparacao_var
```

```
# # A tibble: 2 x 7
#   statistic p_value critical_vale critical_region alternative ratio sig_level
#   <dbl>    <dbl>         <dbl> <chr>         <chr>         <dbl>    <dbl>
# 1      2.59 0.0669         0.218 (0,0.218)U(3.860,~ two.sided      1      0.01
# 2      2.59 0.0669         3.86  (0,0.218)U(3.860,~ two.sided      1      0.01
```

---

Não temos evidência para assumir que as variâncias são diferentes ao nível de significância 1%, e assumimos que as variâncias são iguais.

## Exercício

### Comparação de variâncias

Responde as seguintes questões ao nível de significância 2,5%:

- Os desvios padrões das notas de matemáticas do ENEM/2023 (`nu_nota_mt`) entre pessoas brancas (`branca`) e pessoas negras (`parda` e `preta`) são iguais?
- Os desvios padrões das notas de português do ENEM/2023 (`nu_nota_lc`) entre pessoas brancas (`branca`) e pessoas negras (`parda` e `preta`) são iguais?

---

Cada pessoa tem sua própria cidade.

# Teste t para duas populações

## Variâncias iguais

Primeiro precisamos checar se os desvios padrões são iguais para duas populações.

- **Variável aleatória:** milhas percorridas por galão (milhas\_por\_galao)
- **População 1:** carros com transmissão manual (transmissao == 1)
- **População 2:** carros com transmissão automática (transmissao == 0)

---

```
dados_mtcarros <- read_csv2("dados/brutos/mtcarros.csv")
carros_manuais <- dados_mtcarros |> filter(transmissao == 1)
carros_auto <- dados_mtcarros |> filter(transmissao == 0)
comparacao_var <- ht_2pop_var(
  carros_manuais$milhas_por_galao,
  carros_auto$milhas_por_galao
)
comparacao_var
```

```
# # A tibble: 2 x 7
#   statistic p_value critical_vale critical_region alternative ratio sig_level
#   <dbl>    <dbl>         <dbl> <chr>         <chr>         <dbl>    <dbl>
# 1      2.59  0.0669         0.322 (0,0.322)U(2.769,~ two.sided      1      0.05
# 2      2.59  0.0669         2.77  (0,0.322)U(2.769,~ two.sided      1      0.05
```

---

Ao nível de significância 5%, continuamos acreditando que os desvios padrões das duas populações são iguais.

## Teste t para duas populações Variâncias iguais

Quando sabemos que as variâncias populacionais são iguais.

---

- **População 1:**  $N(\mu_1, \sigma)$
- **População 2:**  $N(\mu_2, \sigma)$
- Teste de Hipóteses envolvendo  $\mu_1$  e  $\mu_2$

No pacote `statBasics`: `ht_2pop_mean` com argumento `var_equal = T`.



Testes de hipóteses deste curso:

- Teste bilateral:  $H_1 : \mu_1 - \mu_2 = \Delta_0$ 
  - `alternative = 'two.sided'`- valor padrão
- Teste unilateral à esquerda:  $H_1 : \mu_1 - \mu_2 < \Delta_0$ 
  - `alternative = 'less'` (*Atenção para ordem das populações*)
- Teste unilateral à direita:  $H_1 : \mu_1 - \mu_2 > \Delta_0$ 
  - `alternative = 'greater'` (*Atenção para ordem das populações*)

Especificamos delta fornecendo  $\Delta_0 = \mu_1 - \mu_2$ . Valor padrão: `delta = 0` (neste caso, estamos testando a igualdade).

## Teste t para duas populações Variâncias iguais

Ao nível de significância 1%, carros com transmissão automática andam mais com galão de gasolina que carros com transmissão manual?

```
comparacao_medias <- ht_2pop_mean(  
  carros_auto$milhas_por_galao,  
  carros_manuais$milhas_por_galao,  
  alternative = "greater",  
  delta = 0,  
  sig_level = 0.01  
)  
comparacao_medias
```

```
# # A tibble: 1 x 7
#   statistic p_value critical_value critical_region delta alternative sig_level
#   <dbl>    <dbl>         <dbl> <chr>          <dbl> <chr>          <dbl>
# 1    -3.77    0.999           2.55 (2.548, Inf)      0 greater          0.01
```

---

Ao nível de significância 1%, não tem evidência para afirmar que carros automáticos são mais eficientes.

## Teste t de Welch

### Variâncias diferentes

Primeiro precisamos checar se os desvios padrões são iguais para duas populações.

- **Variável aleatória:** comprimento de pétala
- **População 1:** espécie setosa (`especies == 'setosa'`)
- **População 2:** espécie versicolor (`especies == 'versicolor'`)

---

```
dados_iris <- read_xlsx("dados/brutos/iris.xlsx")
iris_setosa <- dados_iris |> filter(especies == "setosa")
iris_versicolor <- dados_iris |> filter(especies == "versicolor")
comparacao_var <- ht_2pop_var(
  iris_setosa$comprimento_petala,
  iris_versicolor$comprimento_petala
)
comparacao_var
```

```
# # A tibble: 2 x 7
#   statistic p_value critical_vale critical_region alternative ratio sig_level
#   <dbl>     <dbl>         <dbl> <chr>         <chr>         <dbl>     <dbl>
# 1    0.137 1.03e-10         0.567 (0,0.567)U(1.762~ two.sided      1      0.05
# 2    0.137 1.03e-10         1.76  (0,0.567)U(1.762~ two.sided      1      0.05
```

---

Ao nível de significância 5%, as variâncias dos comprimentos de pétalas para as duas espécies são diferentes.

## Teste t de Welch

### Variâncias diferentes

Quando sabemos que as variâncias populacionais são diferentes

---

- **População 1:**  $N(\mu_1, \sigma)$
- **População 2:**  $N(\mu_2, \sigma)$
- Teste de Hipóteses envolvendo  $\mu_1$  e  $\mu_2$

No pacote `statBasics`: `ht_2pop_mean` com argumento `var_equal = F` (valor padrão).

Testes de hipóteses deste curso:

- Teste bilateral:  $H_1 : \mu_1 - \mu_2 = \Delta_0$ 
  - `alternative = 'two.sided'`- valor padrão
- Teste unilateral à esquerda:  $H_1 : \mu_1 - \mu_2 < \Delta_0$ 
  - `alternative = 'less'` (*Atenção para ordem das populações*)
- Teste unilateral à direita:  $H_1 : \mu_1 - \mu_2 > \Delta_0$ 
  - `alternative = 'greater'` (*Atenção para ordem das populações*)

Especificamos delta fornecendo  $\Delta_0 = \mu_1 - \mu_2$ . Valor padrão: `delta = 0` (neste caso, estamos testando a igualdade).

## Teste t de Welch

### Variâncias diferentes

Existe diferença entre os comprimentos médios de pétalas das espécies setosa e versicolor ao nível de significância 5%?

```
comparacao_medias_iris <- ht_2pop_mean(  
  iris_setosa$comprimento_petala,  
  iris_versicolor$comprimento_petala,  
  delta = 0,  
  var_equal = T,  
  alternative = "two.sided",  
  sig_level = 0.05  
)  
comparacao_medias_iris
```



```
# # A tibble: 1 x 7
#   statistic p_value critical_value critical_region delta alternative sig_level
#   <dbl>    <dbl>         <dbl> <chr>          <dbl> <chr>          <dbl>
# 1    -39.5      0           1.98 (-Inf,-1.984)U(1~    0 two.sided      0.05
```

---

Ao nível de significância 5%, os comprimentos médios de pétalas para as espécies setosa e versicolor são diferentes.

## Exercício

### Comparação de médias

Responda as seguintes perguntas ao nível de significância 1%:

- 1 Existe diferença entre a nota média de matemática (`nu_nota_mt`) entre pessoas negras e brancas?
- 2 A nota em português (`nu_nota_lc`) entre as pessoas brancas é maior que entre as pessoas negras?
- 3 A nota em ciências naturais (`nu_nota_cn`) entre as pessoas brancas é maior que entre as pessoas negras?
- 4 A nota em ciências humanas (`nu_nota_ch`) entre as pessoas brancas é maior que entre as pessoas negras?

## Teste z para proporção

- **População 1:** Bernoulli( $p_1$ )
- **População 2:** Bernoulli( $p_2$ )
- Teste de Hipóteses envolvendo  $p_1$  e  $p_2$

No pacote statBasics: `ht_2pop_prop`.

---

Testes de hipóteses deste curso:

- Teste bilateral:  $H_1 : p_1 - p_2 = \Delta_0$ 
  - `alternative = 'two.sided'` - valor padrão
- Teste unilateral à esquerda:  $H_1 : p_1 - p_2 < \Delta_0$ 
  - `alternative = 'less'` (*Atenção para ordem das populações*)
- Teste unilateral à direita:  $H_1 : p_1 - p_2 > \Delta_0$ 
  - `alternative = 'greater'` (*Atenção para ordem das populações*)

Especificamos delta fornecendo  $\Delta_0 = p_1 - p_2$ . Valor padrão: `delta = 0` (neste caso, estamos testando a igualdade).

Duas formas de realizar este Teste de Hipóteses:

- **Primeira forma:** usando dois vetores de 1 e 0
- **Segunda forma:** usando número de sucessos e tamanhos das amostras das duas populações

## Teste z para proporção

No conjunto de crédito.xlsx, a proporção de estudantes é igual entre pessoas brancas e negras no contexto de solicitação de crédito ao nível de significância 1%?

---

### Primeira forma.

```
df_credito <- read_xlsx("dados/brutos/credito.xlsx")
df_credito_branca <- df_credito |> filter(raca == "Branca")
df_credito_negra <- df_credito |> filter(raca == "Negra")
comparacao_prop <- ht_2pop_prop(
  df_credito_branca$estudante == "Sim",
  df_credito_negra$estudante == "Sim",
  alternative = "two.sided",
  sig_level = 0.01
)
comparacao_prop
```

```
# # A tibble: 1 x 7
#   statistic p_value critical_value critical_region delta alternative sig_level
#   <dbl>    <dbl>         <dbl> <chr>          <dbl> <chr>          <dbl>
# 1    -0.441    0.659           2.58 (-Inf,-2.576)U(2~    0 two.sided    0.01
```

---

Ao nível de significância 1%, não temos evidência para afirmar que a proporção de estudantes entre pessoas brancas negras é diferente.

## Teste z para proporção

A proporção de carros com transmissão manual é maior nos carros produzidos no exterior, ao nível de significância 10%?

Tabela 8: Tabela de contingência entre duas variáveis: origem e transmissão manual.

Transmissão manual	EUA	Importado	Total
Não	26	6	32
Sim	22	39	61
Total	48	45	93

## Segunda forma.

```
comparacao_prop <- ht_2pop_prop(  
  39, 22, 45, 48,  
  alternative = "greater",  
  sig_level = 0.1  
)  
comparacao_prop
```

```
# # A tibble: 1 x 7  
#   statistic p_value critical_value critical_region delta alternative sig_level  
#   <dbl>    <dbl>         <dbl> <chr>         <dbl> <chr>         <dbl>  
# 1      4.14 0.0000172         1.28 (1.282, Inf)      0 greater         0.1
```

---

Ao nível de significância 10%, temos evidência para afirmar que a porcentagem de carros com transmissão manual é maior entre carros importados.



## Exercício

### Comparação de proporções

Responda as seguintes perguntas ao nível de significância 1%:

- 1 Existe diferença entre a porcentagem de treineiros (`in_treineiro`) entre homens e mulheres (`tp_sexo`)?
- 2 A porcentagem de treineiros (`in_treineiro`) entre pessoas brancas é maior que entre pessoas negras (`tp_cor_raca`)?
- 3 Existe diferença no acesso a internet (`q_05`) entre pessoas brancas e negras (`tp_cor_raca`)?
- 4 A porcentagem de pessoas oriundas de escolas privadas (`tp_escola`) é menor em pessoas negras em relação a pessoas brancas (`tp_cor_raca`)?

Cada pessoa tem sua cidade, e todas as variáveis estão no conjunto de dados do ENEM/2023.

## Teste t pareado

- Uma mesma observação é mensurada **antes** e **depois** de um intervenção
- Desejamos checar se a intervenção *produziu* efeito

Vamos usar a função `t.test` com o argumento `paired = TRUE`.

---

Testes de hipóteses deste curso:

- Teste bilateral:  $H_1 : \mu_{\text{antes}} \neq \mu_{\text{depois}}$ 
  - `alternative = 'two.sided'` - valor padrão
- Teste unilateral à esquerda:  $H_1 : \mu_{\text{antes}} < \mu_{\text{depois}}$ 
  - `alternative = 'less'` (*Atenção para ordem*)
- Teste unilateral à direita:  $H_1 : \mu_{\text{antes}} > \mu_{\text{depois}}$ 
  - `alternative = 'greater'` (*Atenção para ordem*)

## Teste t pareado

Existe evidência que combinação de dieta e exercício diminuiu a pressão sanguínea ao nível de significância 5%?

```
df_pressao_sanguinea <- read_xlsx("dados/brutos/pressao_sanguinea.xlsx")
teste_pressao <- t.test(
  df_pressao_sanguinea$antes_exercicio,
  df_pressao_sanguinea$depois_exercicio,
  alternative = "greater",
  paired = T
)
teste_pressao
```

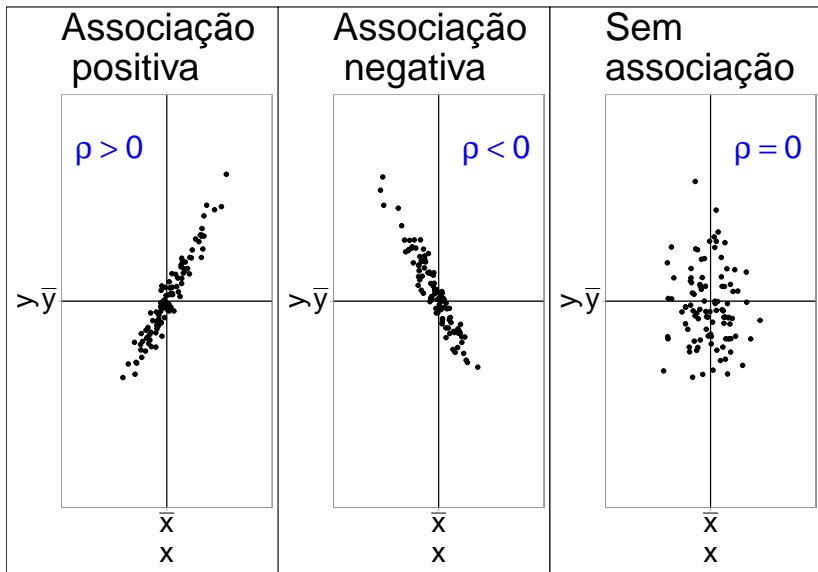
```
#
#   Paired t-test
#
# data:  df_pressao_sanguinea$antes_exercicio and df_pressao_sanguinea$depois_exercicio
# t = 25.364, df = 9, p-value = 5.537e-10
# alternative hypothesis: true mean difference is greater than 0
# 95 percent confidence interval:
#   8.627859      Inf
# sample estimates:
# mean difference
#           9.3
```

---

Ao nível de significância 5%, o exercício e a dieta produziram efeito na diminuição da pressão sanguínea.

Teste de associação para variáveis quantitativas

## Gráfico de dispersão



## Teste de associação para variáveis quantitativas

- **Variável 1:** variável quantitativa
- **Variável 2:** variável quantitativa
- Teste de hipóteses envolvendo o coeficiente de correlação linear de Pearson

Usamos a função `cor_test` do pacote `rstatix`.

---

Testes de hipóteses deste curso:

- Teste bilateral:  $H_1 : \rho \neq 0$ 
  - `alternative = 'two.sided'` - valor padrão
- Teste unilateral à esquerda:  $H_1 : \rho < 0$ 
  - `alternative = 'less'`
- Teste unilateral à direita:  $H_1 : \rho > 0$ 
  - `alternative = 'greater'`

## Teste de associação para variáveis quantitativas

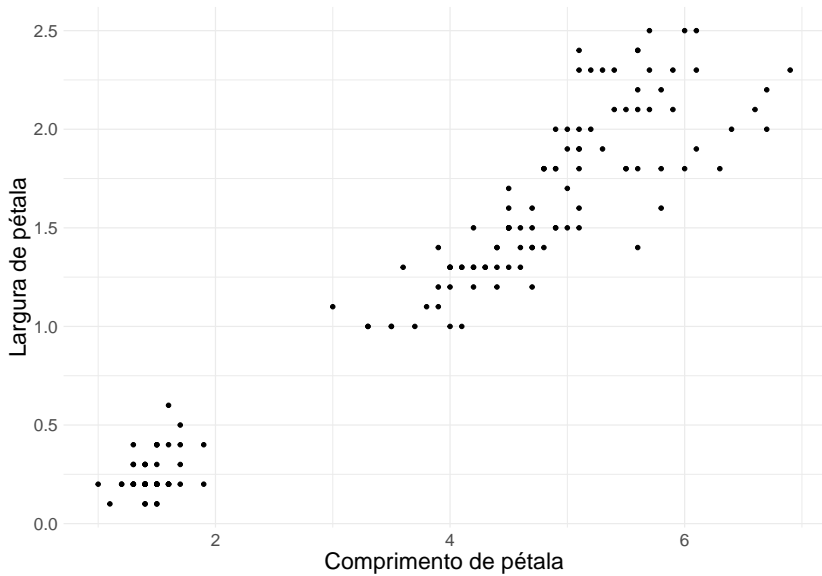
Comprimento e largura de pétalas estão positivamente associadas, ao nível de significância 1%?

---

### gráfico de dispersão

```
dados_iris <- read_xlsx("dados/brutos/iris.xlsx")
ggplot(dados_iris, aes(comprimento_petala, largura_petala)) +
  geom_point() +
  labs(x = "Comprimento de pétala", y = "Largura de pétala") +
  theme_minimal()
```





**coeficiente de correlação linear de Pearson:**

```
coef_cor <- cor(  
  dados_iris$comprimento_petala,  
  dados_iris$largura_petala  
)  
coef_cor
```

```
# [1] 0.9628654
```

---

Aparentemente, existe uma associação positiva entre o comprimento e largura de pétalas.

```
teste_cor <- dados_iris |>
  cor_test(
    comprimento_petala,
    largura_petala,
    alternative = "greater",
    conf.level = 0.99 # coeficiente de confiança
  )
teste_cor
```

```
# # A tibble: 1 x 8
#   var1          var2      cor statistic      p conf.low conf.high method
#   <chr>        <chr>  <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <chr>
# 1 comprimento_petala largura~ 0.96      43.4 2.34e-86 0.946      1 Pears~
```

---

Ao nível de significância 1%, o comprimento e largura de pétalas estão associadas.

## Exercício

### Teste de associação para variáveis quantitativas

Responda as seguintes perguntas ao nível de significância 5%:

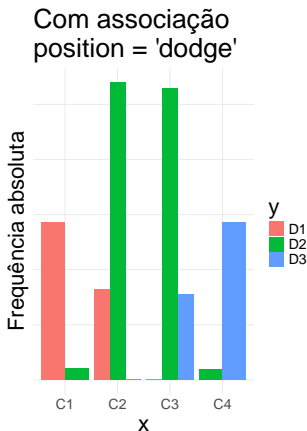
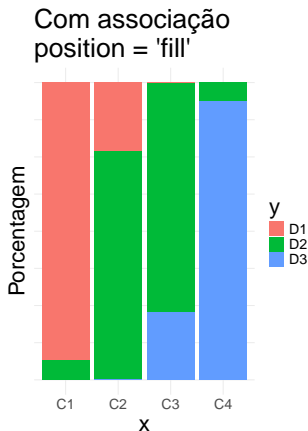
- 1 Existe associação entre a nota em ciências naturais (`nu_nota_cn`) e a nota em ciências humanas (`nu_nota_ch`) no ENEM/2023?
- 2 Existe associação positiva entre a nota em matemática (`nu_nota_mt_`) e a nota em português (`nu_nota_lc`) no ENEM/2023?
- 3 Existe associação positiva entre a nota em português (`nu_nota_lc`) e a nota em redação (`nu_nota_redacao`) no ENEM/2023?

Cada pessoa tem sua cidade.

## Teste de associação para variáveis qualitativas

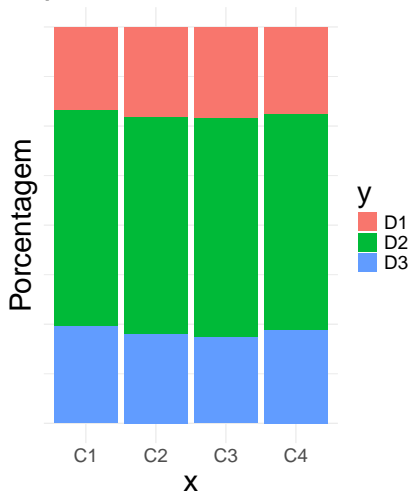
## Gráfico de barras

Quando existe associação.

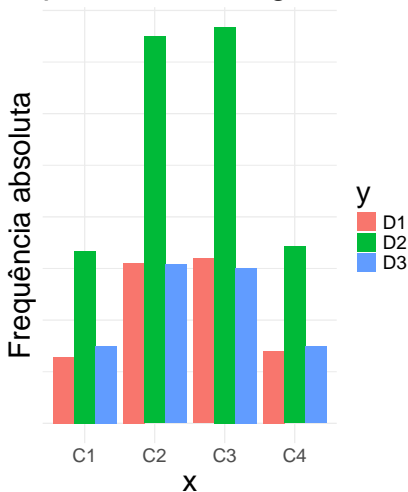


Quando não existe associação.

Sem associação  
position = 'fill'



Sem associação  
position = 'dodge'



## Teste de associação para variáveis qualitativas

- **Variável 1:** variável quantitativa
- **Variável 2:** variável quantitativa
- Queremos checar se *Variável 1* e *Variável 2* estão associadas

Usamos a função `chisq.test` do pacote `janitor`.

---

Queremos testar as seguintes hipóteses:

- $H_0$ : **não** existe associação entre as duas variáveis
- $H_1$ : existe associação entre as duas variáveis



## Teste de associação para variáveis qualitativas

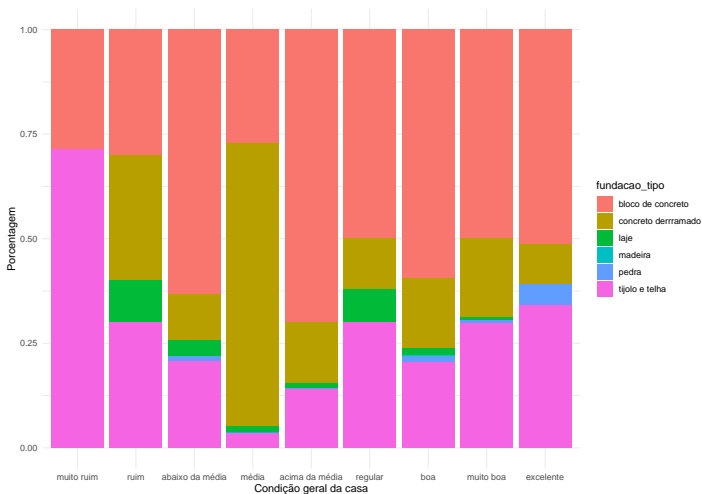
O tipo de fundação (`fundacao_tipo`) e a condição geral da casa (`geral_condicao`) estão associadas, ao nível de significância 1%? (do conjunto de dados `casas.xlsx`)

---

### Vamos construir gráficos de barras

```
casas <- read_xlsx("dados/brutos/casas.xlsx")
casas <- casas |>
  mutate(geral_condicao = fct(
    geral_condicao,
    levels = c(
      "muito ruim", "ruim", "abaixo da média",
      "média", "acima da média", "regular",
      "boa", "muito boa", "excelente"
    )
  )
)
```

```
ggplot(casas, aes(x = geral_condicao, fill = fundacao_tipo)) +  
  geom_bar(position = "fill") +  
  labs(x = "Condição geral da casa", y = "Porcentagem") +  
  theme_minimal()
```



**Vamos calcular o coeficient V de Cramer.**

```
coef_cramer <- CramerV(  
  casas$geral_condicao,  
  casas$fundacao_tipo,  
  correct = T,  
  conf.level = 0.95  
)  
coef_cramer
```

```
#   Cramer V      lwr.ci      upr.ci  
# 0.2535774 0.2371296 0.2698504
```

---

```
teste_associacao <- casas |>  
  tabyl(geral_condicao, fundacao_tipo) |>  
  chisq.test()  
teste_associacao
```

```
#  
#   Pearson's Chi-squared test  
#  
#   data:  tabyl(casas, geral_condicao, fundacao_tipo)  
#   X-squared = 980.42, df = 40, p-value < 2.2e-16
```

---

Ao nível de significância 1%, a condição geral da casa e o tipo de fundação estão associadas.