

R para Ciência de Dados

Intervalos de Confiança e Testes de Hipóteses

Profa Carolina Paraíba e Prof Gilberto Sassi

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade Federal da Bahia

Abril de 2026

Curso R para Ciência de Dados: Intervalos de Confiança e Testes de Hipóteses

- Introdução
- Revisão: Estatística Descritiva
- Inferência Estatística
- Revisão: Probabilidade
- Intervalos de Confiança
- Testes de Hipóteses

Durante o curso

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

Na sua casa

- **IDE** recomendadas: *RStudio* e *VSCode*.
 - Caso você queira usar o *VSCode*, instale a extensão da linguagem R: `REditorSupport`.
- Outras linguagens interessantes: `python` e `julia`.
 - `python`: linguagem interpretada de propósito geral, contemporânea do R, simples e fácil de aprender.
 - `julia`: linguagem interpretada para análise de dados, lançada em 2012, promete simplicidade e velocidade.

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.

Classificação de variáveis

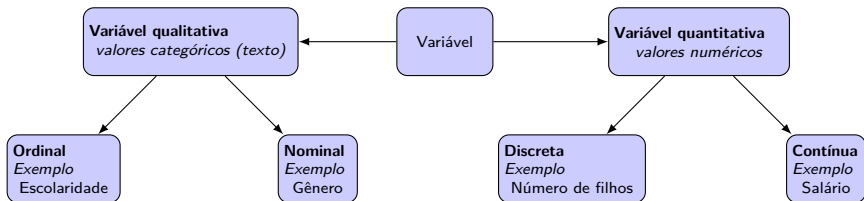


Figura 1: Classificação de variáveis.

Tabela de distribuição de frequências

Variável quantitativa discreta

A primeira coisa que fazemos é contar!

X	frequência	frequência relativa	porcentagem
X_1	n_1	f_1	$100 \cdot f_1\%$
X_2	n_2	f_2	$100 \cdot f_2\%$
\vdots	\vdots	\vdots	\vdots
X_k	n_k	f_k	$100 \cdot f_k\%$
Total	n	1	100%

Observação

n é o tamanho da amostra.

Tabela de distribuição de frequências

Variável quantitativa discreta

```
dados <- read_xlsx(  
  "../dados/brutos/mulheres_20242.xlsx")  
  
tab <- tabyl(dados, filhos) |>  
  adorn_totals() |>  
  adorn_pct_formatting(digits = 2) |>  
  rename(  
    "Nro de filhos" = filhos,  
    "Frequência" = n,  
    "Porcentagem" = percent)
```

tab

##	Nro de filhos	Frequência	Porcentagem
##	0	16	26.67%
##	1	7	11.67%
##	2	10	16.67%
##	3	6	10.00%
##	4	12	20.00%
##	5	9	15.00%
##	Total	60	100.00%

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%

Observação

n é o tamanho da amostra.

Tabela de distribuição de frequências

Variável quantitativa contínua

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

k <- floor(1 + log2(nrow(dados)))

dados <- dados |>
  mutate(comprimento_sepala_int = cut(
    comprimento_sepala,
    breaks = k,
    include.lowest = TRUE,
    right = FALSE))
```

```
tab <- taby1(dados, 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%
```

Histograma

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

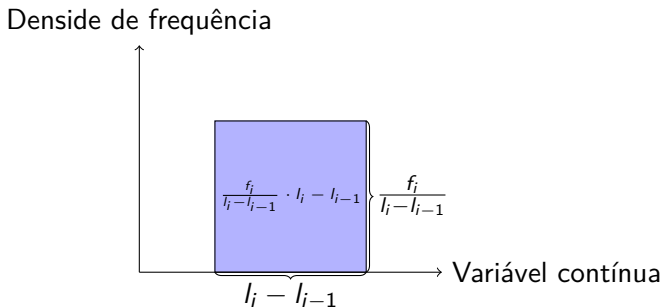
- O histograma é um gráfico de barras contíguas em que a área de cada barra é igual à frequência relativa.
- Cada faixa de valor $[l_{i-1}, l_i)$, $i = 1, \dots, n$, será representada por um barra com área f_i , $i = 1, \dots, n$.
- Como cada barra terá área igual a f_i e base $l_i - l_{i-1}$, e a altura 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.

Histograma

- 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.

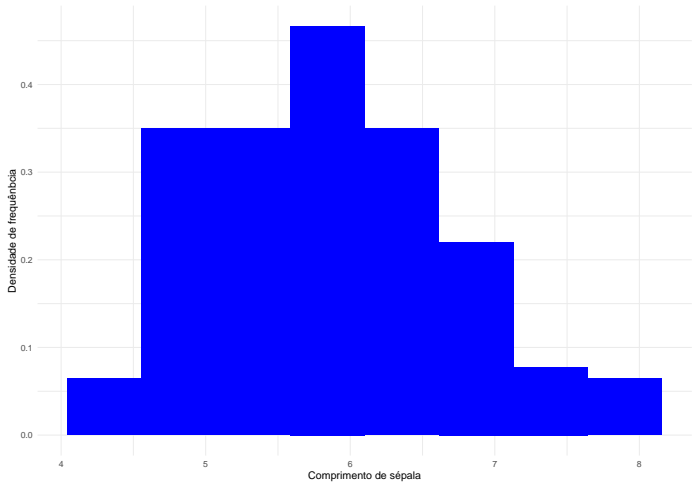
Histograma

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



Histograma

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



Medidas Resumo

```
tab <- group_by(dados, 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
##   species      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
```

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 **confiança** estabelecida pelo pesquisador.

O que faremos nesse curso?

- **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.

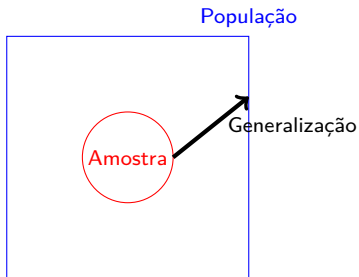
Inferência Estatística

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.



Fenômeno Aleatório

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

- Qual será o resultado do lançamento de um dado “justo”?

Espaço amostral

O conjunto de todos os resultados de um fenômeno aleatório.

Notação: Ω .

Evento

Subconjunto de um espaço amostral. Notação: A, B, C, \dots

Ponto amostral

Um resultado possível de um fenômeno aleatório. Notação: ω .

Probabilidade

A plausibilidade de um ponto amostral ω de A ser o resultado do fenômeno aleatório. Notação: $P(A)$.

Variável aleatória

Função com domínio em um espaço amostral 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 χ .

Variável Aleatória Discreta

Função de probabilidade (FP):

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

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

Na amostra

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

Na população

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

Variável Aleatória Discreta

Medidas resumo para variável aleatória discreta

Na amostra

X uma variável quantitativa discreta

- **Média:**

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

- **Variância:**

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

- **Desvio padrão:**

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

- **Mediana:**

Md tal que:

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

Na população

X para uma variável aleatória discreta

- **Média:**

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

- **Variância:**

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

- **Desvio padrão:**

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

- **Mediana:**

Md tal que:

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

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 não tem secretaria cultura
Pessoa sadia
Pessoa baixa
Bahia não ganha o jogo

Precisamos descobrir a proporção (ou porcentagem) de Sucesso.

Notação: p é a proporção (ou porcentagem) de Sucesso.

Parâmetros da distribuição Bernoulli

Usamos letras gregas para representar parâmetros:

- **Média populacional:** μ .
 - **Variância populacional:** σ^2 .
 - **Desvio padrão populacional:** σ .
-

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

Disribuição 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**).

Estimação pontual

Disribuição Bernoulli

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

```
tab <- dados |>
```

```
  summarise(prop_sucesso = mean(transmissao))
```

```
tab
```

```
## # A tibble: 1 x 1
```

```
##   prop_sucesso
```

```
##           <dbl>
```

```
## 1           0.406
```

Estimação pontual

Disribuição Bernoulli

Variável aleatória: Mulherer já teve gravidez? (coluna gravidez em `mulheres_20242.xlsx`).

- **Sucesso:** Sim (**sim** a mulher já esteve grávida).
- **Fracasso:** Não (**nao** a mulhere nunca esteve grávida).

Vamos criar uma nova coluna com 1 e 0.

```
dados <- read_xlsx(  
  "../dados/brutos/mulheres_20242.xlsx")  
  
dados <- dados |>  
  mutate(ind_gravidez = ifelse(gravidez == "sim", 1, 0))  
  
tab <- dados |>  
  summarise(prop_gravidez = mean(ind_gravidez))  
tab
```

```
## # A tibble: 1 x 1  
##   prop_gravidez  
##           <dbl>  
## 1           0.733
```

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 mulheres que com diagnóstico de endometriose (coluna endometriose de mulheres_20242.xlsx).

```
dados <- read_excel(
  "../dados/brutos/mulheres_20242.xlsx")

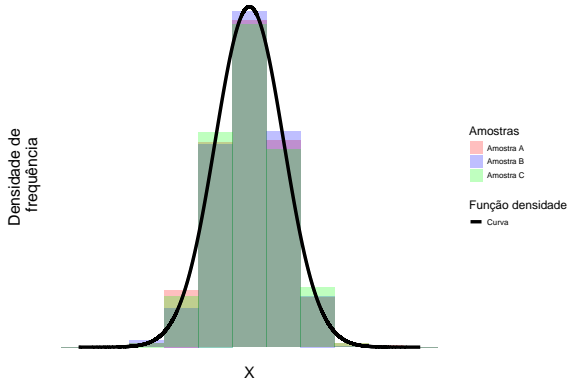
tab <- dados |>
  summarise(
    prop = sum(endometriose) / nrow(dados))
tab
```

```
## # A tibble: 1 x 1
##   prop
##   <dbl>
## 1 0.167
```

Variável Aleatória Contínua

Motivação

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

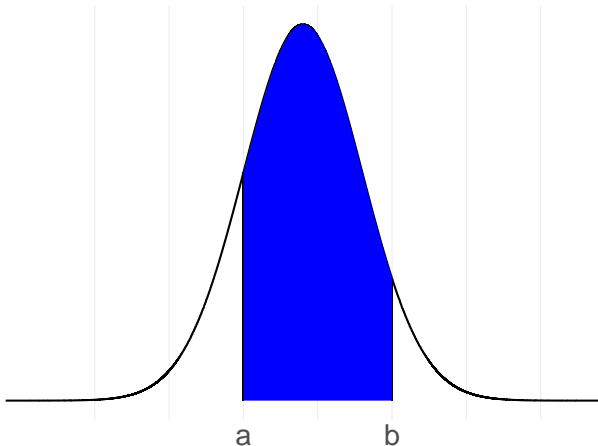


Chamamos a curva preta de **função densidade**.

Variável Aleatória Contínua

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).



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: μ .
- Variância: σ^2 .
- Desvio padrão: σ .

Exemplos de aplicação

Distribuição normal

- Altura:
 - As pessoas 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`).

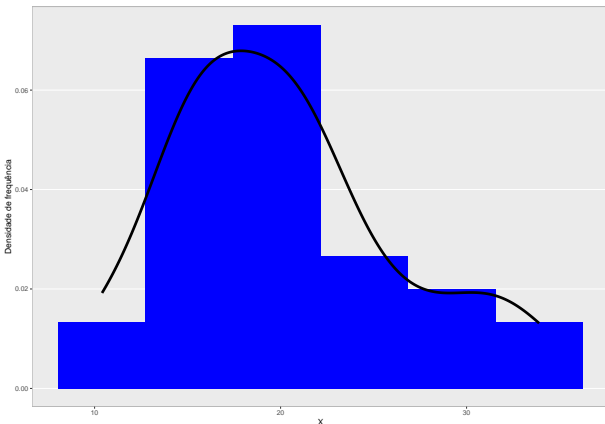
```
dados <- read_xlsx(  
  "../dados/brutos/mtcarros.xlsx")  
  
k <- ceiling(1 + log2(nrow(dados)))
```

```
ggplot(dados, 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()
```

Exemplos

Distribuição normal

Variável aleatória: milhas por galão (`milhas_por_galao` em `mtcarros.xlsx`).



Estimativa pontual

Distribuição Normal

```
tab <- dados |>  
  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
```

Intervalo de Confiança

Objetivo: Para parâmetro μ (σ e p), encontrar L e U tal que $L < \mu < U$ com alguma probabilidade associada γ .

Chamamos γ de *coeficiente de confiança*.

Vamos usar o pacote `statBasics`.

Interpretação de Intervalo de Confiança

O parâmetro μ (σ e p) **pode** ou **não pode** estar entre L e U do intervalo de confiança com coeficiente de confiança γ .

```
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_lpop_norm(variavel)$lower_ci,  
            upper_ci = ci_lpop_norm(variavel)$upper_ci,  
            media_pop = media_pop)
```

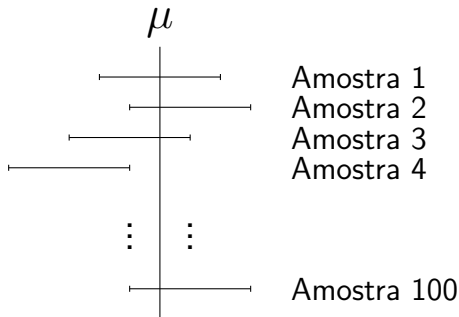
```
tab
```

```
## # 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
```

Intervalo de Confiança

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

Figura 4: Interpretação de intervalo de confiança.



Geralmente γ é 99%, 95% ou 90%.

Intervalo de Confiança Distribuição Bernoulli

Primeira forma: Vetor de 1 e 0.

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

```
dados <- read_xlsx("../dados/brutos/mtcarros.xlsx")
ic_transmissao <- ci_lpop_bern(dados$transmissao,
                              conf_level = 0.95)
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 `transmissao` de `mtcarros.xlsx`).
- **Sucesso:** 1 (transmissão manual).
- **Fracasso:** 0 (transmissão automática).

```
dados <- read_xlsx("../dados/brutos/mtcarros.xlsx")
n_tentativas <- nrow(dados)
n_sucessos <- sum(dados$transmissao)
ic_transmissao <- ci_lpop_bern(n_sucessos, n_tentativas,
                              conf_level = 0.95)

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 automática está entre 0,233 e 0,5795 com coeficiente de confiança 95%.

Intervalo de Confiança

Distribuição Binomial

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%.

Intervalo de Confiança

Distribuição Binomial

- **Variável aleatória:** Número de mulheres que já engravidaram (coluna gravidez de mulheres_20242.xlsx);
- **Coefficiente de confiança:** 92,5%.

```
dados <- read_xlsx("../dados/brutos/mulheres_20242.xlsx")
dados <- dados |>
  mutate(ind_gravidez = ifelse(gravidez == "sim", 1, 0))
n_tentativas <- nrow(dados)
n_sucessos <- sum(dados$ind_gravidez)
ic <- ci_lpop_bern(n_sucessos, n_tentativas, 0.925)
ic
```

```
## # A tibble: 1 x 3
##   lower_ci upper_ci conf_level
##   <dbl>    <dbl>    <dbl>
## 1     0.618     0.848     0.925
```

A proporção de mulheres que já engravidaram está entre 0,6184 e 0,8483 com coeficiente de confiança 92,5%.

Intervalo de Confiança para média

- Variável aleatória tem distribuição normal.
- Variância populacional desconhecida.
- Intervalo de confiança para média.

Função `ci_1pop_norm` do pacote `statBasics`.

Intervalo de Confiança para média

- **Variável aleatória:** milhas por galão (`milhas_por_galao` em `mtcarros.xlsx`).
- **Coefficiente de confiança:** 99%.

```
dados <- read_xlsx("../dados/brutos/mtcarros.xlsx")  
  
ic <- ci_lpop_norm(dados$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%.

Intervalo de Confiança para variância

- Variável aleatória tem distribuição normal.
- Média populacional é desconhecida.
- Intervalo de confiança para a variância.

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

Intervalo de Confiança para variância

- **Variável aleatória:** milhas por galão (`milhas_por_galao` em `mtcarros.xlsx`).
- **Coeficiente de confiança:** 99%.

```
dados <- read_xlsx("../dados/brutos/mtcarros.xlsx")  
  
ic <- ci_lpop_norm(dados$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
```

A variabilidade do consumo de milhas por galão está entre 20,47 e 77,89, com coeficiente de confiança 99%.

Intervalo de Confiança para média

Grandes amostras

- Variável aleatória não tem distribuição normal.
- Variância populacional é desconhecida.
- Tamanho amostral é *suficientemente grande*.
- Intervalo de confiança para média.

Função `ci_1pop_general` do pacote `statBasics`.

Intervalo de Confiança para média

Grandes amostras

- **Variável aleatória:** número de filhos (filhos em mulheres_20242.xlsx).
- **Coefficiente de confiança:** 95%.

```
dados <- read_xlsx("../dados/brutos/mulheres_20242.xlsx")
```

```
ic <- ci_lpop_general(dados$filhos,  
                    conf_level = 0.95)
```

```
ic
```

```
## # A tibble: 1 x 3  
##   lower_ci upper_ci conf_level  
##   <dbl>    <dbl>    <dbl>  
## 1     1.82     2.78     0.95
```

As mulheres têm, em média, entre 1,824 e 2,776 filhos com coeficiente de confiança 95%.

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 .

Teste de Hipóteses

- Decisão através de *evidência* na amostra:
 - Decisão baseada em ***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 : **Rejeitamos H_0** .

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**: $1 - \beta = P(\text{Verdadeiro positivo})$.

		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 Teste**.

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

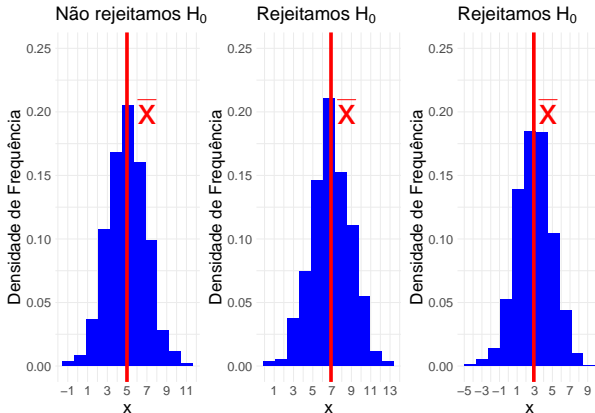
- 1 **Procedimento Geral de Testes de Hipóteses** ou **Procedimento de Neymann-Pearson**.
- 2 **valor-p** (*p-value* em inglês).

Teste de Hipóteses

Como decidir?

População: $N(\mu, 4)$.

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



Procedimento de Neymann-Pearson

Etapas:

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

A região crítica é construída usando o nível de significância.

Erros decaem quando o tamanho amostral aumenta.

As porcentagens de falso positivo e de falso negativo diminuem quando o tamanho da amostra aumenta.

valor-p ou nível crítico

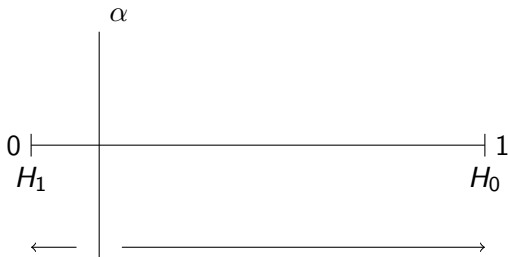
p-value

- 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.

valor-p ou nível crítico

p-value

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



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 cada amostra, temos um valor-p diferente.

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

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

Teste t para média

- Variável aleatória tem distribuição normal.
- Variância populacional é desconhecida.
- Hipóteses sobre a média da população (μ).

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 μ_0 com o 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): $\mu < 15$.

Teste t para média

```
dados <- read_xlsx("../dados/brutos/mtcarros.xlsx")
ht_milhas_galao <- ht_lpop_mean(
  dados$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 há evidência para afirmar que os carros americanos fazem no máximo 15 milhas por galão, em média, ao nível de significância 1%.

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 com o 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): $p > 0,25$.

Teste z para proporção

Distribuição Bernoulli

```
dados <- read_xlsx("../dados/brutos/mtcarros.xlsx")
teste_transmissao <- ht_1pop_prop(
  dados$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%.

Teste z para proporção

Distribuição Binomial

- **Variável aleatória:** Número de mulheres com diagnóstico de endometriose (coluna `endometriose de mulheres_20242.xlsx`)
- **Nível de significância:** 2,5%
- A maioria das mulheres têm endometriose?
 - $H_0 : p \leq 0,5$.
 - $H_1 : p > 0,5$ (o que desejamos provar).

Teste z para proporção

Distribuição Binomial

```
dados <- read_xlsx("../dados/brutos/mulheres_20242.xlsx")
n_tentativas <- nrow(dados)
n_sucessos <- sum(dados$endometriose)
teste_endometriose <- ht_lpop_prop(
  n_sucessos, n_tentativas, proportion = 0.5,
  alternative = "greater", sig_level = 0.025)
```

```
teste_endometriose
```

```
## # A tibble: 1 x 7
##   statistic p_value critical_value critical_region alternative proportion
##   <dbl> <dbl> <dbl> <chr> <chr> <dbl>
## 1 -5.16 1.00 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 mulheres têm endometriose.

Teste z para média

Grandes amostras

- Hipóteses sobre a média da população (μ).
- Variável aleatória não tem distribuição normal.
- Tamanho amostral é *suficientemente grande*.

No pacote `statBasics`: `ht_1pop_mean` (com adaptação).

Teste Qui-Quadrado para variância

Distribuição normal

- Hipóteses sobre a variância da população (σ).
- Variável aleatória tem distribuição normal.
- Média populacional é desconhecida.

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 σ_0^2 com o 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 <- read_xlsx("../dados/brutos/mtcarros.xlsx")
ht_milhas_galao <- ht_lpop_var(
  dados$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).

Intervalos de Confiança e Testes de Hipóteses para duas populações

Experimentos Comparativos

Experimento completamente aleatório:

Medimos uma mesma variável em duas populações independentes.

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

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.

Intervalos de Confiança e Testes de Hipóteses para duas populações

Comparação de variâncias

Antes de comparar μ_1 e μ_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 σ_1 e σ_2 .

No pacote `statBasics`: `ht_2pop_var`.

Intervalos de Confiança e Testes de Hipóteses para duas populações

Comparação de variâncias

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).

Intervalos de Confiança e Testes de Hipóteses para duas populações

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).

Intervalos de Confiança e Testes de Hipóteses para duas populações

Comparação de variâncias

```
dados <- read_xlsx("../dados/brutos/mtcarros.xlsx")
carros_manuais <- dados |> filter(transmissao == 1)
carros_auto <- dados |> 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.

Teste t para duas populações Variâncias iguais

Primeiro precisamos verificar 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`).

Teste t para duas populações Variâncias iguais

```
dados <- read_xlsx("../dados/brutos/mtcarros.xlsx")
carros_manuais <- dados |> filter(transmissao == 1)
carros_auto <- dados |> 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
##   <dbl>     <dbl>         <dbl> <chr>
## 1     2.59  0.0669           0.322 (0,0.322)U(2.769,~
## 2     2.59  0.0669           2.77  (0,0.322)U(2.769,~
```

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 μ_1 e μ_2 .

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

Teste t para duas populações Variâncias iguais

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 verificar 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'`).

Teste t de Welch

Variâncias diferentes

```
dados <- read_xlsx("../dados/brutos/iris.xlsx")
iris_setosa <- dados |> filter(especies == "setosa")
iris_versicolor <- dados |> filter(especies == "versicol")
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 μ_1 e μ_2 .

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

Teste t de Welch

Variâncias diferentes

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.

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`.

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

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).

Teste z para proporção

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

```
dados <- read_xlsx("../dados/brutos/credito.xlsx")
dados_branca <- dados |> filter(raca == "Branca")
dados_negra <- dados |> filter(raca == "Negra")
comparacao_prop <- ht_2pop_prop(
  dados_branca$estudante == "Sim",
  dados_negra$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 e negras é diferente.

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%?

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

```
teste_pressao
```

```
##  
## Paired t-test  
##  
## data: dados$antes_exercicio and dados$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 qualitativas

- **Variável 1:** variável qualitativa.
- **Variável 2:** variável qualitativa.
- 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.

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`)

Teste de associação para variáveis qualitativas

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

```

Teste de associação para variáveis qualitativas

Vamos calcular o coeficient V de Cramer.

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

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

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

```
teste_associacao
```

```
##  
## Pearson's Chi-squared test  
##  
## data:  tabyl(dados, 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.