



Estatística descritiva usando R bem-vind@ ao tidyverse

Curso livre de R

Profa Carolina e Prof Gilberto
Parte 2

parte 2

Na aula de hoje

Continuando com o conteúdo do dia 06/11/2021, hoje vamos aprender:

- Estatística descritiva
 - tabela de distribuição de frequências
 - gráficos
- Probabilidade (bem rapidamente)
- Inferência para uma população (normal ou Bernoulli):
 - Intervalo de confiança para média, desvio padrão e proporção
 - Teste de hipóteses para média, desvio padrão e proporção

Na aula de hoje, usaremos a IDE **RStudio**.



Pacotes da aula de hoje

- `glue`: facilita a manipulação de *string* (caracteres)
- `readxl`: permite a leitura de arquivos `.xlsx`
- `writexl`: permite salvar um `data.frame` (`tibble`) como um arquivo excel
- `statBasics`: pacote criado pela equipe técnica para construir intervalos de confiança e teste de hipóteses
- `ggthemes`: pacote com diversos temas para gráficos no pacote `ggplot2`
- `xtable`: pacote para salvar tabelas html e latex
- `gt`: permite construir e salvar tabelas customizadas e formatadas
- `tidyverse`: framework para simplificar de forma moderna a análise de dados

Para acompanhar a aula de hoje, instale e carregue todos estes pacotes.



estadística descriptiva

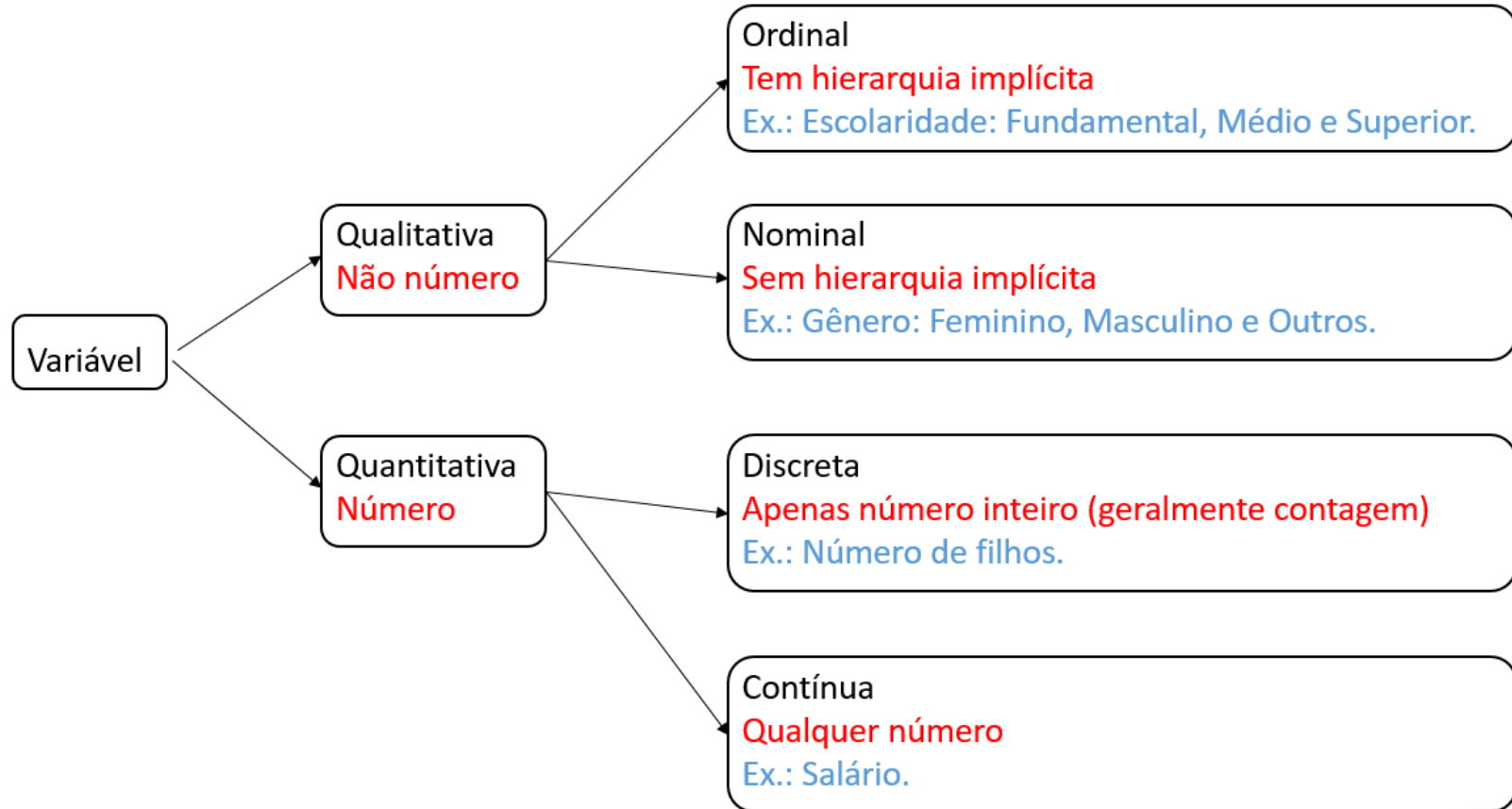
Conceitos básicos

Vamos começar com alguns conceitos básicos, que usaremos durante a aula de hoje.

- **População:** Todos os elementos (ou indivíduos) alvos do seu estudo.
- **Amostra:** parte da população.
- **Parâmetro:** característica da população.
- **Amostra:** característica da amostra. (Usamos estimativa para aproximar a população).
- **Variável aleatória:** *característica de um elemento da população:*
 - usamos uma letra maiúscula para denotar ou representar uma variável aleatória;
 - usamos uma letra minúscula para representar coletado ou observado da variável aleatória.



Classificação de variáveis aleatórias



Classificação de variáveis aleatórias.

tabelas

Tabelas da distribuição de frequências variável qualitativa

A primeira coisa que podemos fazer é contar!

Seja X uma variável qualitativa com valores possíveis B_1, \dots, B_k .

Tabela de distribuição de frequências.

| 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 | $100 \cdot f_k \%$ |

Tabelas da distribuição de frequências variável qualitativa

Vamos usar a variável escolaridade o conjunto de dados empresa.xlsx.

```
df_empresa <- read_xlsx("../data/raw/empresa.xlsx")
tab <- df_empresa |>
  group_by(escolaridade) |>
  summarise(frequencia = n()) |>
  mutate(fr = frequencia / sum(frequencia), p = 100 * fr)
tab
```

```
## # A tibble: 3 × 4
##   escolaridade      frequencia    fr      p
##   <chr>          <int> <dbl> <dbl>
## 1 ensino fundamental      12 0.333  33.3
## 2 ensino médio           18 0.5    50
## 3 superior                6 0.167  16.7
```



Como podemos salvar uma tabela

Exportar como latex

```
xtable(tab) |>  
  print.xtable(digits = 2, include.rownames = F,  
               booktabs = T, format.args = list(decimal.mark = ","))
```

```
## % latex table generated in R 4.1.2 by xtable 1.8-4 package  
## % Fri Nov 19 17:25:43 2021  
## \begin{table}[ht]  
## \centering  
## \begin{tabular}{lrrr}  
## \toprule  
## escolaridade & frequencia & fr & p \\  
## \midrule  
## ensino fundamental & 12 & 0,33 & 33,33 \\  
## ensino médio & 18 & 0,50 & 50,00 \\  
## superior & 6 & 0,17 & 16,67 \\  
## \bottomrule  
## \end{tabular}  
## \end{table}
```



Como podemos salvar uma tabela

Exportar como html

```
xtable(tab) |>  
  print.xtable(digits = 2, include.rownames = F,  
               booktabs = T, format.args = list(decimal.mark = ","),  
               type = "html")
```

```
## <!-- html table generated in R 4.1.2 by xtable 1.8-4 package -->  
## <!-- Fri Nov 19 17:25:43 2021 -->  
## <table border=1>  
## <tr> <th> escolaridade </th> <th> frequencia </th> <th> fr </th> <th> p </th> </tr>  
## <tr> <td> ensino fundamental </td> <td align="right"> 12 </td> <td align="right"> 0,33 </td> <td align="right"> 0,33 </td> </tr>  
## <tr> <td> ensino médio </td> <td align="right"> 18 </td> <td align="right"> 0,50 </td> <td align="right"> 0,50 </td> </tr>  
## <tr> <td> superior </td> <td align="right"> 6 </td> <td align="right"> 0,17 </td> <td align="right"> 0,17 </td> </tr>  
## </table>
```

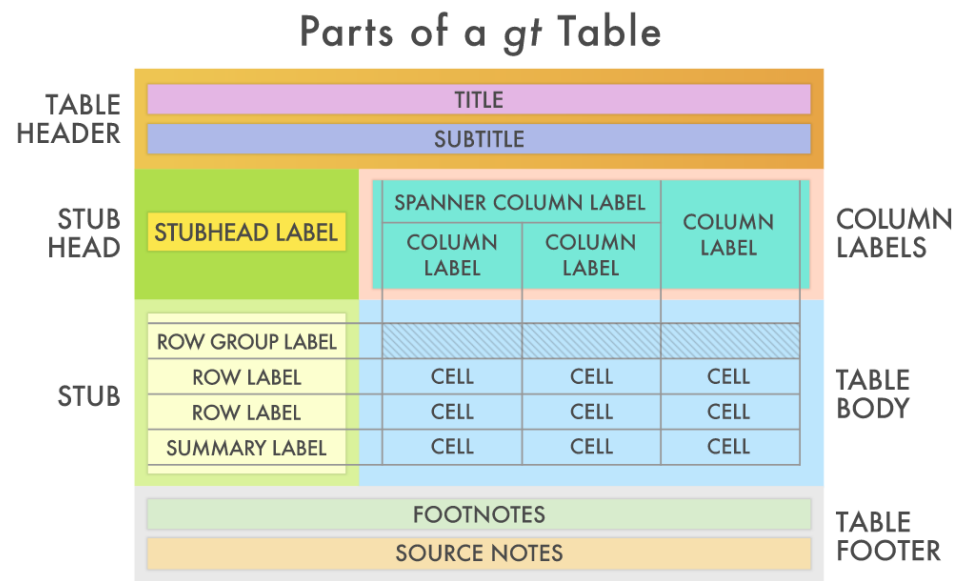


Customizando tabelas

Vamos usar o pacote `gt` para customizar a apresentação de uma tabela.

A ideia do pacote `gt` é melhorar apresentação por camadas.

```
knitr::include_graphics("../figures/gt-table.png")
```



Customizando tabelas usando o pacote `gt`.

Customizando tabelas

Incluindo cabeçalho e sub-cabeçalho

```
gt_tab <- gt(tab) |>
  tab_header(
    title = md("**Escolaridade dos funcionário:** _Empresa tal_ "),
    subtitle = md("**Criado por:** _Gilberto Sassi_")
  )
gtsave(gt_tab, filename = "../output/gt_tab.html")
gtsave(gt_tab, filename = "../output/gt_tab.tex")
gtsave(gt_tab, filename = "../output/gt_tab.rtf")
```



Customizando tabelas

Incluindo fonte

```
gt_tab <- gt_tab |>
  tab_source_note(
    source_note = md("Exemplo didático.")
  ) |>
  tab_source_note(
    source_note = md("BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica.**")
  )
gt_tab
```



Escolaridade dos funcionários: *Empresa tal*

Criado por: *Gilberto Sassi*

| escolaridade | frequencia | fr | p |
|--------------------|------------|-----------|----------|
| ensino fundamental | 12 | 0.3333333 | 33.33333 |
| ensino médio | 18 | 0.5000000 | 50.00000 |
| superior | 6 | 0.1666667 | 16.66667 |

Exemplo didático.

BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica.**



Customizando tabelas

Agrupando linhas

- Usamos a função `tab_row_group` para agrupar linhas.

```
gt_tab <- gt_tab |>
  tab_row_group(
    label = "Ensino básico",
    rows = 1:2
  ) |>
  tab_row_group(
    label = "Ensino superior",
    row = 3
  )
gt_tab
```



Escolaridade dos funcionários: *Empresa tal*

Criado por: *Gilberto Sassi*

| escolaridade | frequencia | fr | p |
|--------------|------------|----|---|
|--------------|------------|----|---|

Ensino superior

| | | | |
|----------|---|-----------|----------|
| superior | 6 | 0.1666667 | 16.66667 |
|----------|---|-----------|----------|

Ensino básico

| | | | |
|--------------------|----|-----------|----------|
| ensino fundamental | 12 | 0.3333333 | 33.33333 |
|--------------------|----|-----------|----------|

| | | | |
|--------------|----|-----------|----------|
| ensino médio | 18 | 0.5000000 | 50.00000 |
|--------------|----|-----------|----------|

Exemplo didático.

BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica.**



Customizando tabelas

Agrupando colunas

- Usamos a função `tab_spanner` para agrupar linhas.

```
gt_tab <- gt_tab |>
  tab_spanner(
    label = md("_Variável aleatória_"),
    columns = "escolaridade"
  ) |>
  tab_spanner(
    label = md("***Informações numéricas***"),
    columns = c(frequencia, fr, p)
  )
gt_tab
```



Escolaridade dos funcionários: *Empresa tal*

Criado por: *Gilberto Sassi*

| <i>Variável aleatória</i> | Informações numéricas | | |
|---------------------------|-----------------------|-----------|----------|
| escolaridade | frequencia | fr | p |
| Ensino superior | | | |
| superior | 6 | 0.1666667 | 16.66667 |
| Ensino básico | | | |
| ensino fundamental | 12 | 0.3333333 | 33.33333 |
| ensino médio | 18 | 0.5000000 | 50.00000 |

Exemplo didático.

BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica.**



Customizando tabelas

Modificando os rótulos da coluna

Podemos:

- Mudar a ordem das colunas usando a função `cols_move_to_start`
- Mudar o nome das colunas usando a função `cols_label`

```
gt_tab <- gt_tab |>
  cols_move_to_start(
    columns = c(escolaridade, frequencia)
  ) |>
  cols_label(
    escolaridade = md("**Grau de Escolaridade**"),
    frequencia = md("**Frequência**"),
    fr = md("**Frequência relativa**"),
    p = md("**Porcentagem**")
  )
gt_tab
```



Escolaridade dos funcionários: *Empresa tal*

Criado por: *Gilberto Sassi*

| <i>Variável aleatória</i> | Informações numéricas | | |
|---|------------------------------|----------------------------|--------------------|
| Grau de Escolaridade | Frequência | Frequência relativa | Porcentagem |
| Ensino superior | | | |
| superior | 6 | 0.1666667 | 16.66667 |
| Ensino básico | | | |
| ensino fundamental | 12 | 0.3333333 | 33.33333 |
| ensino médio | 18 | 0.5000000 | 50.00000 |
| Exemplo didático. | | | |
| BUSSAB, Wilton de O.; MORETTIN, Pedro A. Estatística básica. | | | |



Customizando tabelas

Modificando o formato numérico dos valores

```
gt_tab <- gt_tab |>
  fmt_number(
    columns = c(fr),
    decimals = 2,
    dec_mark = ",",
    big_mark = "."
  ) |>
  fmt_number(
    columns = p,
    decimals = 2,
    dec_mark = ",",
    big_mark = ".",
    pattern = "{x}%"
  )
gt_tab
```



Escolaridade dos funcionários: *Empresa tal*

Criado por: *Gilberto Sassi*

| <i>Variável aleatória</i> | Informações numéricas | | |
|-----------------------------|------------------------------|----------------------------|--------------------|
| Grau de Escolaridade | Frequência | Frequência relativa | Porcentagem |
| Ensino superior | | | |
| superior | 6 | 0,17 | 16,67% |
| Ensino básico | | | |
| ensino fundamental | 12 | 0,33 | 33,33% |
| ensino médio | 18 | 0,50 | 50,00% |

Exemplo didático.

BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica.**



gráficos

Gráficos no R

- Pacote: `ggplot2`
- Permite gráficos personalizados com uma sintaxe simples e rápida, e iterativa *por camadas*
- Começamos com um camada com os dados `ggplot(dados)`, e vamos adicionando as camadas de anotações, e sumários estatísticos
- Usa a *gramática de gráficos* proposta por Leland Wilkinson: [Grammar of Graphics](#)
- Ideia desta gramática: delinear os atributos estéticos das figuras geométricas (incluindo transformações nos dados e mudança no sistema de coordenadas)
- Para mais detalhes, você pode consultar [ggplot2: elegant graphics for data analysis](#) e [documentação do ggplot2](#)



Gráficos no R

Estrutura básica de ggplot2

```
ggplot(data = <data possible tibble>) +  
  <Geom functions>(mapping = aes(<MAPPINGS>)) +  
  <outras camadas>
```

Você pode usar diversos temas e extensões que a comunidade cria e criou para melhorar a aparência e facilitar a construção de ggplot2.

- Lista com extensões do ggplot: [extensões do ggplots](#)

Indicação de extensões:

- Temas adicionais para o pacote ggplot2: [ggthemes](#)
- Gráfico de matriz de correlação: [ggcorrplot](#)
- Gráfico quantil-quantil: [qqplotr](#)



Gráficos no R

Gráfico de Barras no ggplot2

- **função:** `geom_bar()`. Para porcentagem: `geom_bar(x = <variável no eixo x>, y = ..prop.. * 100)`.
- Argumentos adicionais:
 - **fill:** mudar a cor do preenchimento das figuras geométricas
 - **color:** mudar a cor da figura geométrica

Rótulos dos eixos

- **Mudar os rótulos:** `labs(x = <rótulo do eixo x>, y = <rótulo do eixo y>)`
- Trocar o eixo-x pelo eixo-y: `coord_flip()`



```
library(ggthemes)
ggplot(df_empresa) +
  geom_bar(mapping = aes(x = escolaridade, y = ..prop.. * 100, group = 1),
    fill = "blue", color = "red") +
  labs(x = "Espécies", y = "Porcentagem") +
  theme_gdocs()
```

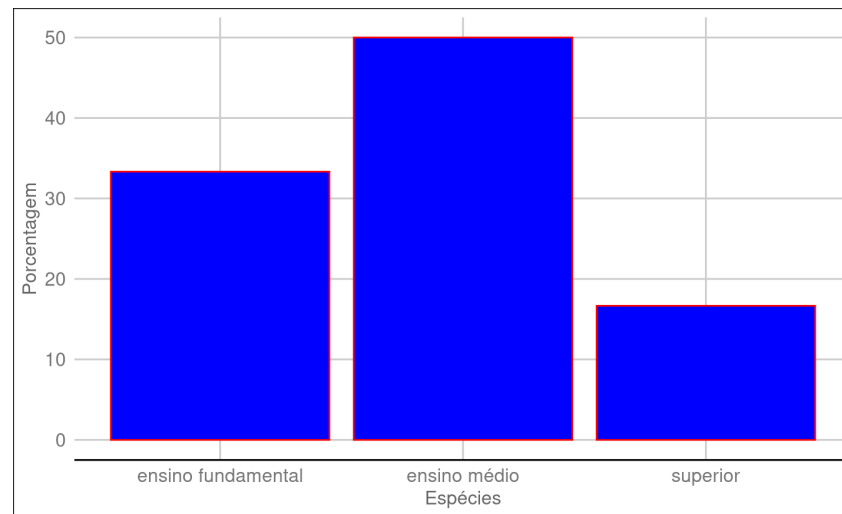


Gráfico de barras para escolaridade.

Tabela de distribuição de frequência – 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 | $100 \cdot f_2 \%$ |
| x_3 | n_3 | f_3 | $100 \cdot f_3 \%$ |
| \vdots | \vdots | \vdots | \vdots |
| x_k | n_k | f_k | $100 \cdot f_k \%$ |
| Total | n | 1 | 100% |

Em que n é o tamanho da amostra.



Tabela de distribuição de frequência – Variável quantitativa discreta

A primeira coisa que podemos fazer é construir a tabela de distribuição de frequência.

```
tab <- df_empresa |>
  group_by(n_filhos = as.character(numero_filhos)) |>
  summarise(frequencia = n()) |>
  mutate(fr = frequencia / sum(frequencia),
         percentagem = 100 * fr)
tab <- tab |>
  add_case(n_filhos = "Total",
          frequencia = sum(tab$frequencia),
          fr = sum(tab$fr),
          percentagem = sum(tab$percentagem))
tab
```



| n_filhos | frequencia | fr | porcentagem |
|----------|------------|------|-------------|
| 0 | 20 | 0,56 | 55,56 |
| 1 | 5 | 0,14 | 13,89 |
| 2 | 7 | 0,19 | 19,44 |
| 3 | 3 | 0,08 | 8,33 |
| 5 | 1 | 0,03 | 2,78 |
| Total | 36 | 1,00 | 100,00 |



Gráfico de barras no R

```
ggplot(df_empresa) +  
  geom_bar(aes(x = numero_filhos, y = ..prop.., group = 1)) +  
  labs(x = "Número de filhos", y = "Frequência relativa") +  
  theme_calc()
```

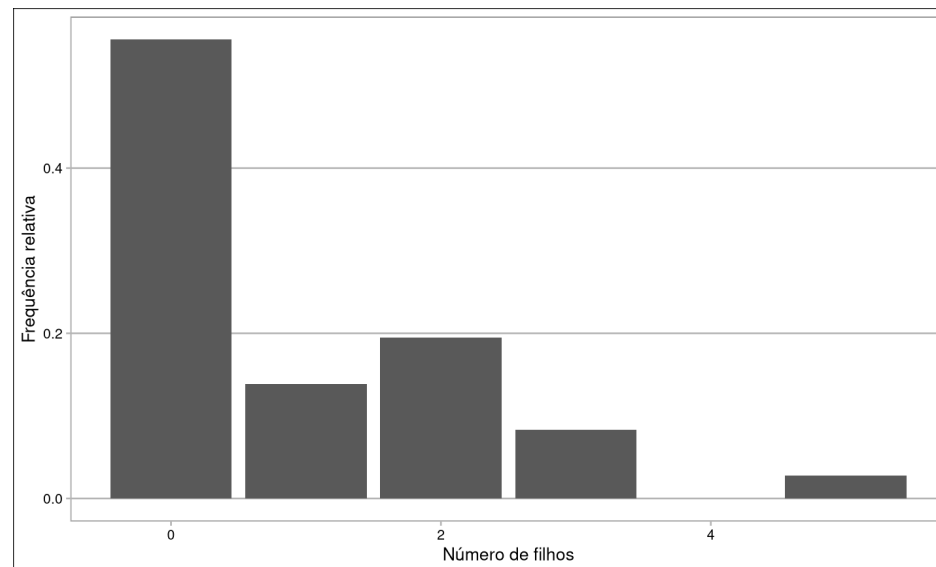


Tabela de distribuição de frequência – Variável quantitativa contínua

- X: variável quantitativa contínua

Tabela de Distribuição de Frequências para a variável quantitativa contínua.

| x | Frequência | Frequência relativa | Porcentagem |
|------------------|------------|---------------------------------------|-----------------------|
| $[l_0, l_1)$ | n_1 | $f_1 = \frac{n_1}{n_1 + \dots + n_k}$ | $p_1 = f_1 \cdot 100$ |
| $[l_1, l_2)$ | n_2 | $f_2 = \frac{n_2}{n_1 + \dots + n_k}$ | $p_2 = f_2 \cdot 100$ |
| \vdots | \vdots | \vdots | \vdots |
| $[l_{k-1}, l_k]$ | n_k | $f_k = \frac{n_k}{n_1 + \dots + n_k}$ | $p_k = f_k \cdot 100$ |

Em que $\min = l_0 \leq l_1 \leq \dots \leq l_{k-1} \leq l_k = \max$ (\min é o menor valor do suporte da variável X e \max é o maior valor do suporte da variável X), n_i é número de valores de X entre l_{i-1} e l_i , e l_0, l_1, \dots, l_k quebram o suporte da variável X (*breakpoints*).

l_0, l_1, \dots, l_k são escolhidos de acordo com a teoria por trás da análise de dados (ou pelo regulador). Se você está em uma nova área, use l_0, l_1, \dots, l_k igualmente espaçados, e use a [regra de Sturges](#) para determinar o valor de k : $k = 1 + \log_2(n)$ onde n é tamanho da amostra. Se $1 + \log_2(n)$ não é um número inteiro, usamos $k = \lceil 1 + \log_2(n) \rceil$.



Tabela de distribuição de frequência – Variável quantitativa contínua

```
df_iris <- read_xlsx("../data/raw/iris.xlsx")

k <- round(1 + log2(nrow(df_iris)))
tab <- df_iris |>
  group_by(sepal_length_intervalo = cut(Sepal.Length, breaks = k,
                                       include.lowest = T, right = F)) |>
  summarise(freq = n()) |>
  mutate(fr = freq / sum(freq), p = fr * 100)
tab <- tab |>
  add_case(
    sepal_length_intervalo = "Total",
    freq = sum(tab$freq),
    fr = sum(tab$fr),
    p = sum(tab$p)
  )
tab
```



| sepal_length_intervalo | freq | fr | p |
|------------------------|------|------|--------|
| [4.3,4.75) | 11 | 0,07 | 7,33 |
| [4.75,5.2) | 30 | 0,20 | 20,00 |
| [5.2,5.65) | 24 | 0,16 | 16,00 |
| [5.65,6.1) | 24 | 0,16 | 16,00 |
| [6.1,6.55) | 31 | 0,21 | 20,67 |
| [6.55,7) | 17 | 0,11 | 11,33 |
| [7,7.45) | 7 | 0,05 | 4,67 |
| [7.45,7.9] | 6 | 0,04 | 4,00 |
| Total | 150 | 1,00 | 100,00 |



Histograma

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

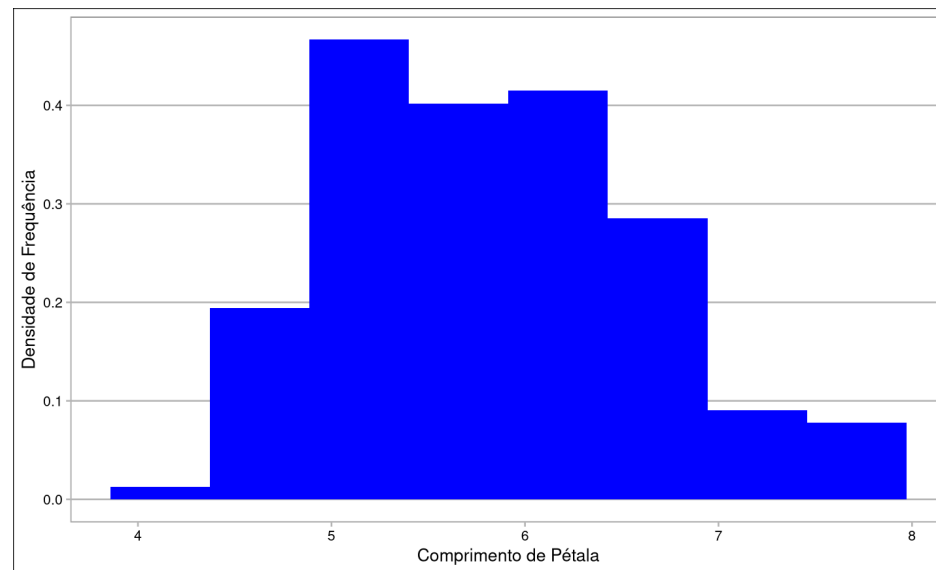
- O histograma é um gráfico de barras contíguas em que a área de cada barra é igual à frequência relativa.
- Cada faixa de valor $[l_{i-1}, l_i)$, $i = 1, \dots, n$, será representada por uma barra com área f_i , $i = 1, \dots, n$.
- Como cada barra terá área igual a f_i e base $l_i - l_{i-1}$, e a altura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

```
df_iris <- read_xlsx("../data/raw/iris.xlsx")
k <- round(1 + log2(nrow(df_iris)))

ggplot(df_iris) +
  geom_histogram(aes(x = Sepal.Length, y = ..density..),
                 bins = k, fill = "blue") +
  theme_calc() +
  labs(x = "Comprimento de Pétala", y = "Densidade de Frequência")
```



Medidas Resumo (variável quantitativa)

A ideia é encontrar um ou alguns valores que sintetizem todos os valores.

Medidas de posição (tendência central)

A ideia é encontrar um valor que representa *bem* todos os valores.

- **Média:** $\bar{x} = \frac{x_1 + \dots + x_n}{n}$
- **Mediana:** valor que divide a sequência ordenada de valores em duas partes iguais.

Medidas de dispersão

A ideia é medir a homogeneidade dos valores.

- **Variância:** $s^2 = \frac{(x_1 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{n - 1}$;
- **Desvio padrão:** $s = \sqrt{s^2}$ (mesma unidade dos dados);
- **coeficiente de variação** $cv = \frac{s}{\bar{x}} \cdot 100\%$ (adimensional, ou seja, “sem unidade”)



Medidas de Resumo: exemplo

Podemos usar a função `summarise` do pacote `dplyr` (incluso no pacote `tidyverse`).

```
tab <- df_empresa |>
  summarise(media = mean(salario), dp = sd(salario), mediana = median(salario),
            cv = dp / media) |>
  gt() |>
  tab_header(title = "Medidas de resumo para salário.",
             subtitle = "Média, mediana, desvio padrão e coeficiente de variação") |>
  cols_label(
    media = md("***Média salarial***"),
    dp = md("***Desvio padrão do salário***"),
    mediana = md("***Salário mediano***"),
    cv = md("***Coeficiente de variação***")) |>
  fmt_number(
    columns = everything(),
    decimals = 2,
    dec_mark = ",",
    sep_mark = "."
  )
tab
```



Medidas de resumo para salário.

Média, mediana, desvio padrão e coeficiente de variação em salários mínimos.

| Média salarial | Desvio padrão do salário | Salário mediano | Coeficiente de variação |
|----------------|--------------------------|-----------------|-------------------------|
| 11,12 | 4,59 | 10,16 | 0,41 |

Medidas de Resumo: exemplo

Podemos usar a função `group_by` para calcular medidas de resumo por categorias de uma variável qualitativa.

```
df_empresa |>
  group_by(escolaridade) |>
  summarise(media = mean(salario), dp = sd(salario), md = median(salario), cv = dp / media) |>
  gt() |>
  tab_header(
    title = "Medidas de resumo por escolaridade.",
    subtitle = "Média, desvio padrão, mediana e coeficiente de variação em salários mínimos."
  ) |>
  cols_label(
    escolaridade = "Escolaridade",
    media = "Média salarial",
    dp = "Desvio padrão de salário",
    md = "Salário mediano",
    cv = "Coeficiente de variação"
  ) |>
  fmt_number(
    columns = c(media, dp, md, cv),
    decimals = 2,
    dec_mark = ",",
    sep_mark = "."
  )
```



Medidas de resumo por escolaridade.

Média, desvio padrão, mediana e coeficiente de variação em salários mínimos.

| Escolaridade | Média salarial | Desvio padrão de salário | Salário mediano | Coeficiente de variação |
|--------------------|----------------|--------------------------|-----------------|-------------------------|
| ensino fundamental | 7,84 | 2,96 | 7,12 | 0,38 |
| ensino médio | 11,53 | 3,72 | 10,91 | 0,32 |
| superior | 16,48 | 4,50 | 16,74 | 0,27 |

Inferência Estatística

O processo da inferência estatística

- Usando as técnicas de Estatística Descritiva, podemos fazer afirmações válidas para uma amostra.
- Já em Inferência Estatística, queremos fazer afirmações válidas para toda a população. Isto é, queremos fazer generalizações para a população a partir da amostra, conforme ilustrado na Figura abaixo.

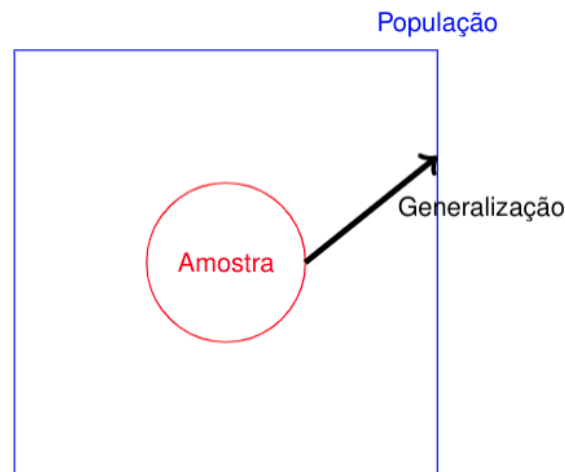


Ilustração da inferência estatística.

O que queremos?

- **Estimação pontual:** Aproximar um parâmetro usando a estimativa. Usamos estimativa para aproximar o parâmetro.

Exemplo: Média salarial dos funcionários.

- **Intervalo de confiança:** Encontrar a e b tal que $a < \text{parâmetro} < b$ com alguma *confiança* fixada pelo pesquisador.

Exemplo: Encontrar $a < \text{parâmetro} < b$ com alguma *confiança* pré-estabelecida.

- **Teste de hipóteses:** Decidir entre duas hipóteses científicas H_0 e H_1 , onde H_1 é negação de H_0 .

Exemplo: Queremos decidir entre

$$\begin{cases} H_0 : \text{a média salarial é no máximo 3 salários mínimos} \\ H_1 : \text{a média salarial é maior que 3 salários mínimos} \end{cases}$$



Intervalo de confiança

Intervalo de confiança para proporção

Usamos quando temos apenas duas opções (sucesso e fracasso). Seja p proporção de sucesso na população, e queremos encontrar a e b tal que $a < p < b$ com coeficiente de confiança γ .

Intervalo de confiança para média

Seja μ a média na população, e queremos encontrar a e b tal que $a < \mu < b$ com coeficiente de confiança γ .



Intervalo de confiança para a média

Considere a variável `salario` do conjunto de dados `empresa.xlsx`, e suponha que desejamos construir um intervalo de confiança para a média salarial com coeficiente de confiança $\gamma = 98\%$.

```
dados <- read_xlsx("../data/raw/empresa.xlsx")
ci_general(dados$salario, conf_level = 0.98)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci conf_level
##   <dbl>    <dbl>    <dbl>
## 1     9.26    13.0     0.98
```



Interpretação do intervalo de confiança

Para cada amostra (ou estudo), o intervalo de confiança pode estar correto ($a < \mu < b$) ou pode estar incorreto ($\mu < a$ ou $b < \mu$).

No conjunto de dados `amostras.xlsx`, temos seis amostras de uma população com média 25, e vamos calcular o intervalo de confiança para cada amostra.

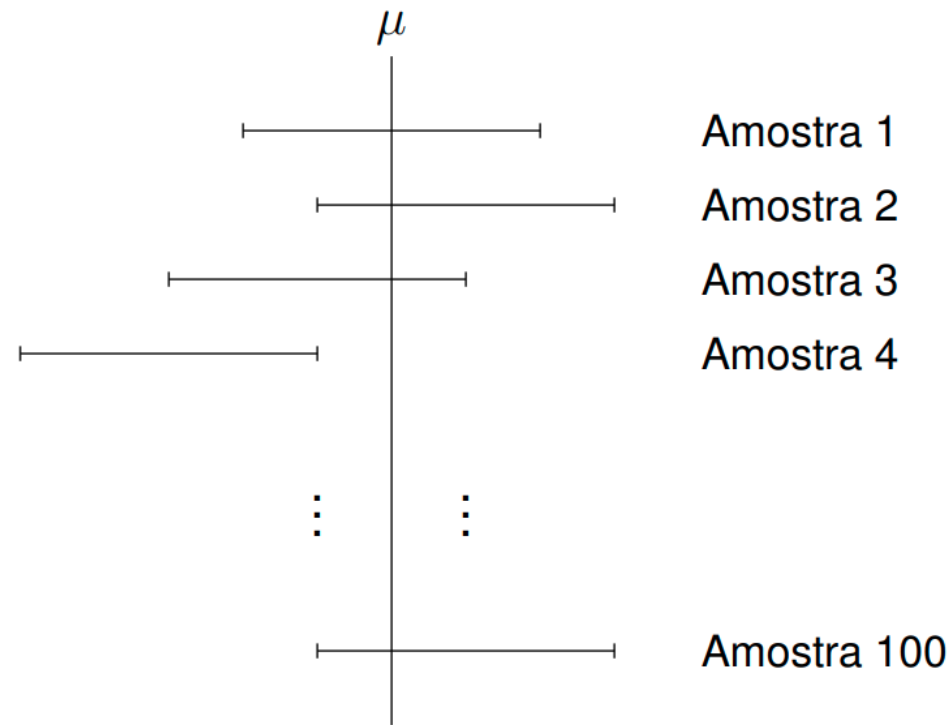
```
dados <- read_xlsx("../data/raw/amostras.xlsx")
intervalos <- dados |>
  group_by(amostra) |>
  summarise(li = ci_general(valores)$lower_ci, ls = ci_general(valores)$upper_ci)
gt(intervalos) |>
  fmt_number(
    columns = c(li, ls),
    decimals = 2,
    dec_mark = ",",
    sep_mark = "."
  ) |>
  cols_label(
    amostra = md("**amostras**"),
    li = md("**Limite inferior**"),
    ls = md("**Limite superior**")
  )
```



| amostras | Limite inferior | Limite superior |
|-----------|-----------------|-----------------|
| amostra_1 | 24,33 | 26,00 |
| amostra_2 | 24,24 | 26,01 |
| amostra_3 | 24,33 | 25,75 |
| amostra_4 | 23,02 | 24,51 |
| amostra_5 | 25,13 | 25,94 |
| amostra_6 | 24,16 | 24,89 |

Interpretação do intervalo de confiança

Importante: $\gamma\%$ dos intervalos de confiança estão corretos e contêm o parâmetro da população.



Interpretação do intervalo de confiança: $\gamma\%$ dos intervalos estão corretos.

Intervalo de confiança para proporção

Considere a variável `procedencia` do conjunto de dados `empresa.xlsx`, e suponha que desejamos construir um intervalo de confiança para a proporção de pessoas que vieram da capital com coeficiente de confiança $\gamma = 99\%$.

Nesse caso, temos

- sucesso: funcionário nasceu na capital;
- fracasso: funcionário não nasceu na capital.

```
dados <- read_xlsx("../data/raw/empresa.xlsx")
ci_bern(dados$procedencia == 'capital', conf_level = 0.99)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci conf_level
##   <dbl>     <dbl>     <dbl>
## 1  0.0909    0.520      0.99
```



Teste de hipóteses

Objetivo: decidir entre duas hipóteses científicas H_0 e H_1 , onde H_0 é chamada de hipótese nula e H_1 é chamada de hipótese alternativa.

Como estabelecer H_0 e H_1

- Valor padrão (*benchmark* do mercado ou *benchmark* do regulador) ou especificação do cliente vai sempre no H_0 .
- Hipótese científica ou pergunta vai sempre no H_1 .

Ao decidirmos, podemos errar de duas formas:

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

Tipos de erros que um analista pode cometer ao decidir usando as informações (*evidências estatísticas*) de uma amostra.

Teste de hipóteses

Usamos probabilidade para controlar os *falsos positivos* ou *falsos negativos*:

- $\alpha = P(\text{falso positivo}) = P(\text{Erro tipo I})$ - nível de significância.
- $\beta = P(\text{falso negativo}) = P(\text{Erro tipo II})$.
- $1 - \beta = P(\text{verdadeiro negativo})$ - poder do teste.

Impossível estabelecer uma decisão que minimize, simultaneamente, α e β (ou minimiza α e maximiza $1 - \beta$).

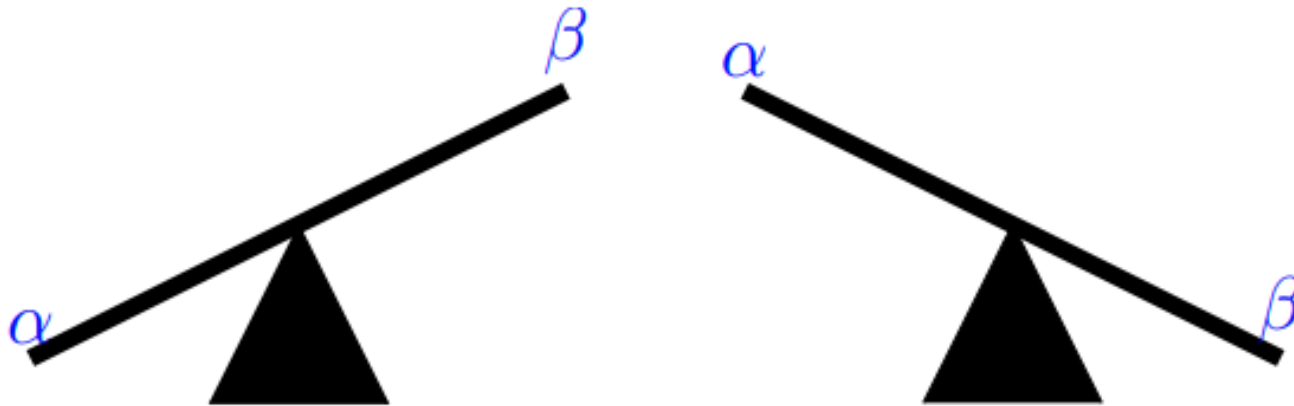


Ilustração dos erros tipos I e II. Impossível minimizar, simultaneamente, α e β .

Teste de hipóteses

Falso positivo: é o erro mais grave!

Estratégia para especificar H_0 e H_1 :

1. Determinar o erro mais grave que será o falso positivo;
2. Determino H_0 e H_1 a partir do falso positivo.

Exemplo (Ilustração do falso positivo)

Em um julgamento precisamos decidir se um *réu* é: **inocente** ou **culpado**.

Temos dois erros possíveis:

- Culpar um inocente;
- Inocentar um culpado.

Determinando as hipóteses nulas e alternativas:

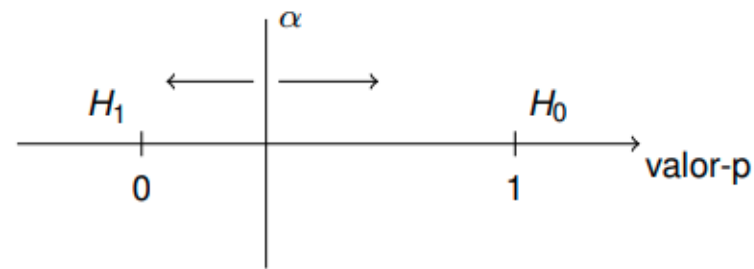
1. O erro mais grave é **culpar um inocente**;
2. **Falso positivo** é culpar um inocente;
3.
$$\begin{cases} H_0 : \text{o réu é inocente} \\ H_1 : \text{o réu é culpado} \end{cases}$$



Valor-p

Descrição intuitiva

- **estatística teste**: quantidade que indica a *evidência* contra H_0 . Quanto mais *extrema* (muito pequeno ou muito grande), mais *evidência* temos contra H_0 .
- O valor-p, ou *p-value* em inglês, é a probabilidade de coletar uma outra amostra com **estatística teste** igual ou mais extrema do que a amostra observada coletada quando H_0 é verdadeira. Lembre que o erro tipo I ou falso positivo é o mais grave.
- Rejeitamos H_0 quando o valor-p é pequeno, e usamos como valor de referência o nível de significância $\alpha\%$. Ilustramos essa ideia na Figura abaixo.



Decisão usando o valor-p.

Valor-p

Interpretação

Imagine um contexto em que H_0 é verdade. Neste contexto, o valor-p pode ser pequeno ou grande, ou seja, podemos rejeitar ou não a hipótese nula.

O importante é que para $\alpha \cdot 100\%$ das amostras rejeitaremos H_0 .

```
dados <- read_xlsx("../data/raw/amostras.xlsx")
dados |>
  group_by(amostra) |>
  summarise(valor_p = t.test(valores, mu = 25)$p.value) |>
  gt() |>
  fmt_number(
    columns = valor_p,
    decimals = 2,
    sep_mark = ".",
    dec_mark = ",",
  ) |>
  cols_label(
    amostra = md("***Amostras***"),
    valor_p = md("***Valor-p***")
  )
```



| Amostras | Valor-p |
|-----------|---------|
| amostra_1 | 0,68 |
| amostra_2 | 0,77 |
| amostra_3 | 0,91 |
| amostra_4 | 0,00 |
| amostra_5 | 0,01 |
| amostra_6 | 0,01 |



Teste de hipóteses para média

A média salarial dos funcionários é maior que 5 salários mínimos ao nível de significância 5%?

$$\begin{cases} H_0 : \text{a média salarial é no máximo 5 salários mínimos,} \\ H_1 : \text{a média salarial é maior que 5 salários mínimos.} \end{cases}$$

```
dados <- read_xlsx("../data/raw/empresa.xlsx")
t.test(dados$salario, mu = 5, alternative = "greater")
```

```
##
## One Sample t-test
##
## data: dados$salario
## t = 8.0073, df = 35, p-value = 1.006e-09
## alternative hypothesis: true mean is greater than 5
## 95 percent confidence interval:
##  9.830415      Inf
## sample estimates:
## mean of x
## 11.12222
```



Teste de hipóteses para proporção

Os funcionários com origem na capital são maioria ao nível de significância 1%?

$$\begin{cases} H_0 : \text{a percentagem de funcionários com origem na capital é no máximo 50\%,} \\ H_1 : \text{a percentagem de funcionários com origem na capital é maior que 50\%.} \end{cases}$$

```
dados <- read_xlsx("../data/raw/empresa.xlsx")
num_sucessos <- sum(sum(dados$procedencia == 'capital'))
tamanho_amostra <- nrow(dados)
prop.test(num_sucessos, tamanho_amostra, p = 0.5, alternative = "greater")
```

```
##
## 1-sample proportions test with continuity correction
##
## data:  num_sucessos out of tamanho_amostra, null probability 0.5
## X-squared = 4.6944, df = 1, p-value = 0.9849
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.1851783 1.0000000
## sample estimates:
##           p
## 0.3055556
```

