Estatística Descritiva com R

Curso livre de R

Profa Carolina e Prof Gilberto

Parte 2

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.

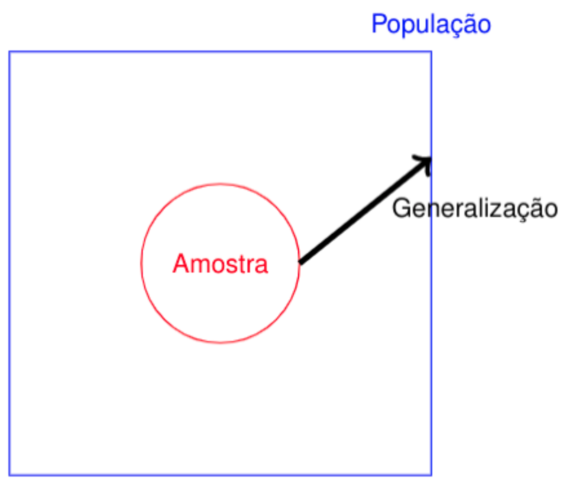


Ilustração da inferência estatística.

O que podemos fazer com Inferência Estatística?

- **Estimação pontual:**
  - utilizar os dados observados (amostra) para encontrar o melhor palpite sobre o parâmetro (populacional). Usamos uma estimativa para ``aproximar" o parâmetro.
- **Exemplo:**
  - com base em uma amostra da população de Salvador com 25 anos ou mais (conjunto de dados observados), qual seria nosso melhor chute para a média salarial (em R$) dessa população?

O que podemos fazer com Inferência Estatística?

- **Intervalo de confiança:**
  - utilizar os dados observados (amostra) para encontrar um intervalo numérico (a, b), tal que o parâmetro populacional de interesse esteja contido nesse intervalo com algum *nível de confiança* pré-fixado.
- **Exemplo:**
  - com base em uma amostra da população de Salvador com 25 anos ou mais (conjunto de dados observados), qual seria o intervalo numérico que contém o valor da média salarial (em R$) dessa população *95% de confiança*?

O que podemos fazer com Inferência Estatística?

- **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:

$$\begin{cases} H_0 : \text{a média salarial é menor ou igual a R\$ 1000,00} \\ H_1 : \text{a média salarial é maior do que R\$ 1000,00} \end{cases}$$

Intervalos de confiança

Intervalo de confiança para a média populaional

- Usamos quando a variável de interesse é quantitativa.

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

Intervalo de confiança para a proporção populacional

- Usamos quando a variável de interesse assume uma de duas opções (`sucesso` e `fracasso`).

- Seja p aproporção de sucesso na população. Queremos encontrar a e b tal que $a < p < b$ com coeficiente de confiança γ.

Intervalo de confiança para a média populacional

Considere a variável `salario` do conjunto de dados `empresa.xlsx`.

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 $b < \mu$).

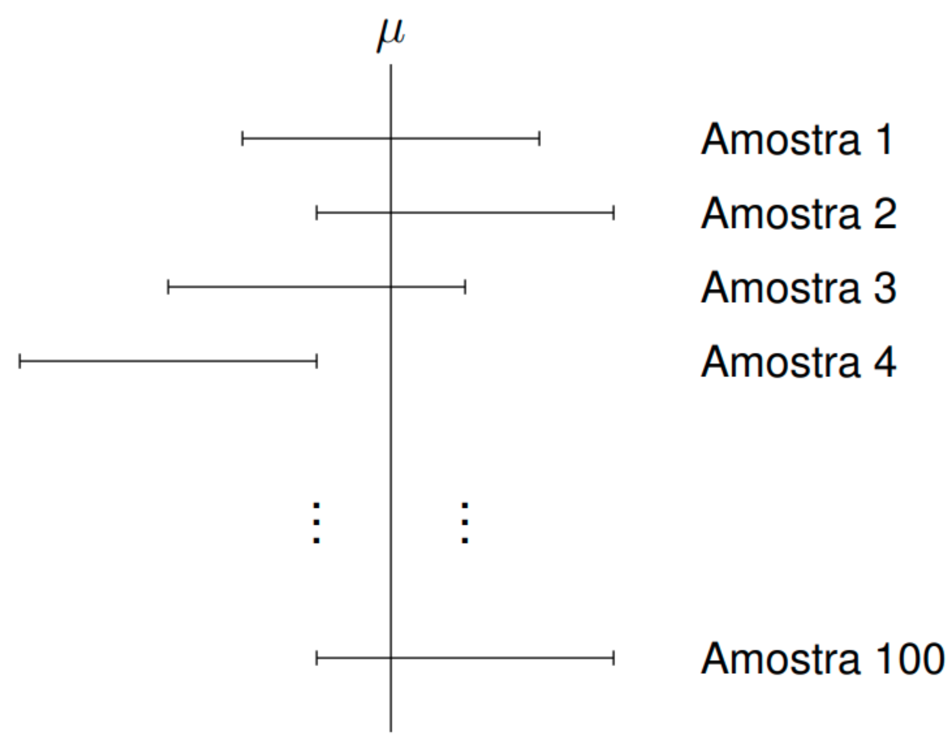
No conjunto de dados `amostras.xlsx`, temos seis amostras de uma população com média 25, e vamos calcular o intervalo de confiança para cada amostra.

```
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 verdadeiro valor (desconhecido) do parâmetro populacional.



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

Intervalo de confiança para a proporção populacional

Considere a variável `procedencia` do conjunto de dados `empresa.xlsx`.

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 miniza, simultaneamente, α e β (ou minimiza α e maximiza $1 - \beta$).

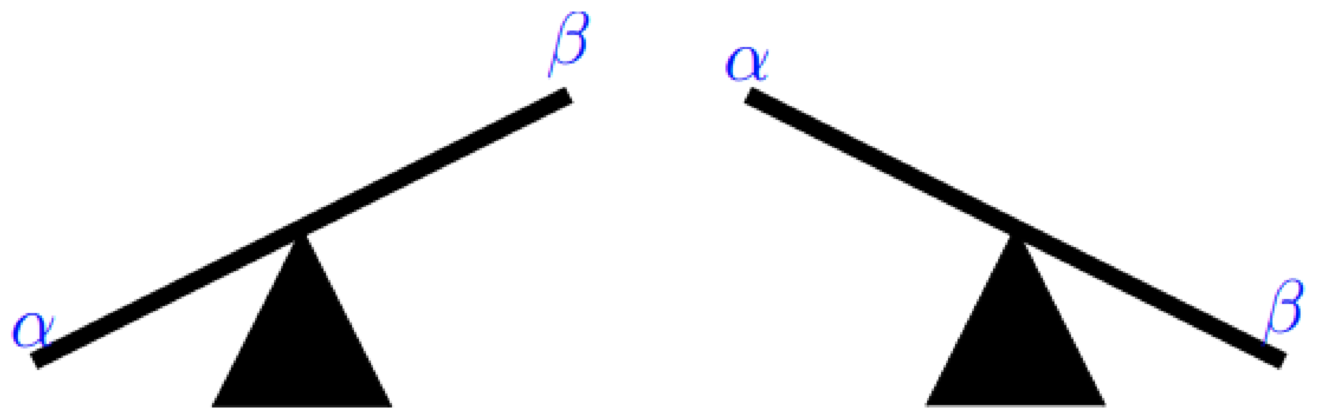


Ilustração dos erros tipos I e II. Impossível minimizar, simultaneamente, α e β.

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.

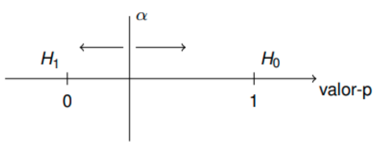
Determinando as hipóteses nulas e alternativas:

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.

```r
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 a média populacional

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

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

```r
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 a proporção populacional

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

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

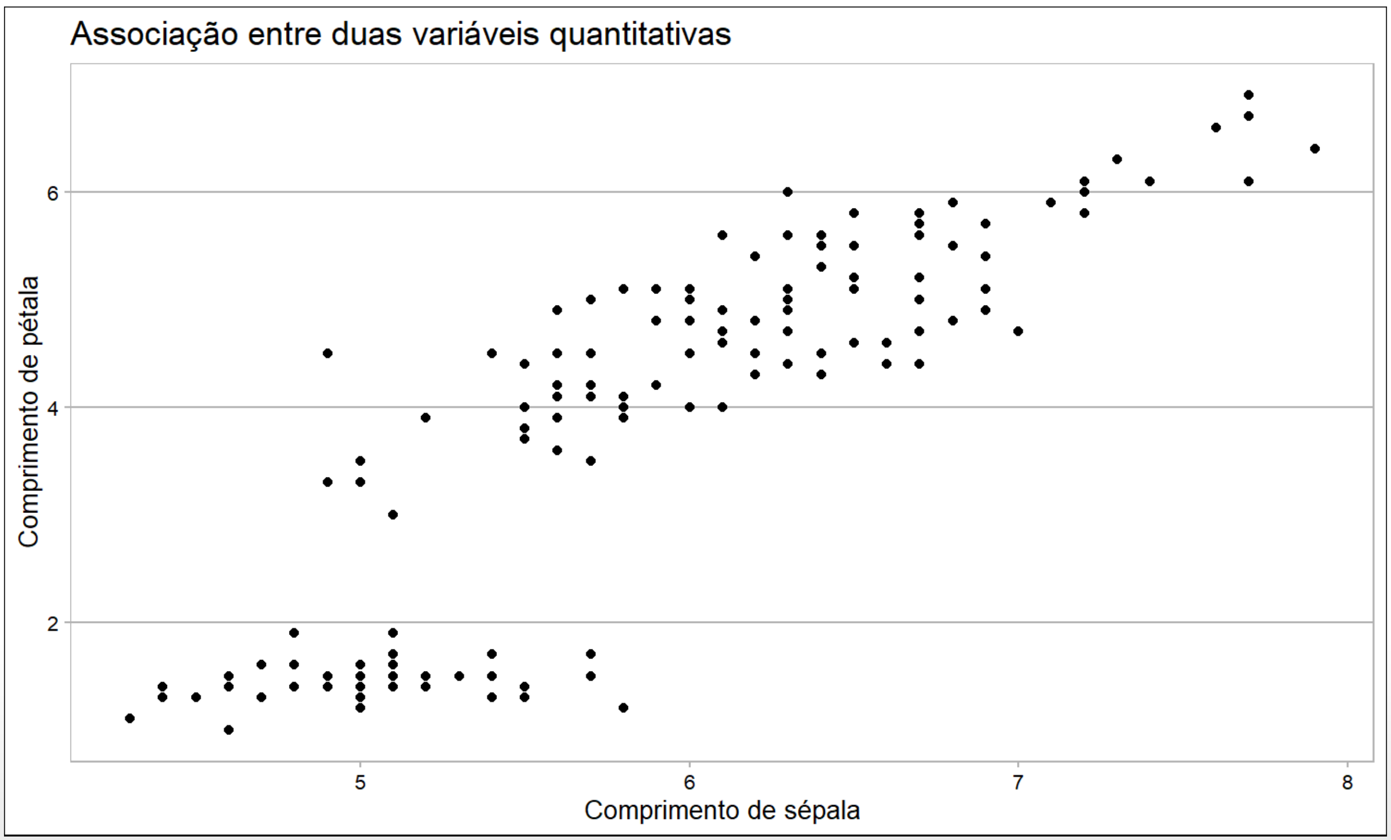
```r
dados <- read_xlsx("../data/raw/empresa.xlsx")
num_sucessos <- sum(sum(dados$procedencia == 'capital'))
tamanho_amostra <- nrow(dados)
prop.test(num_sucessos, tamanho_amostra, p = 0.5, alternative = "greater")
```

```
##
##  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
```

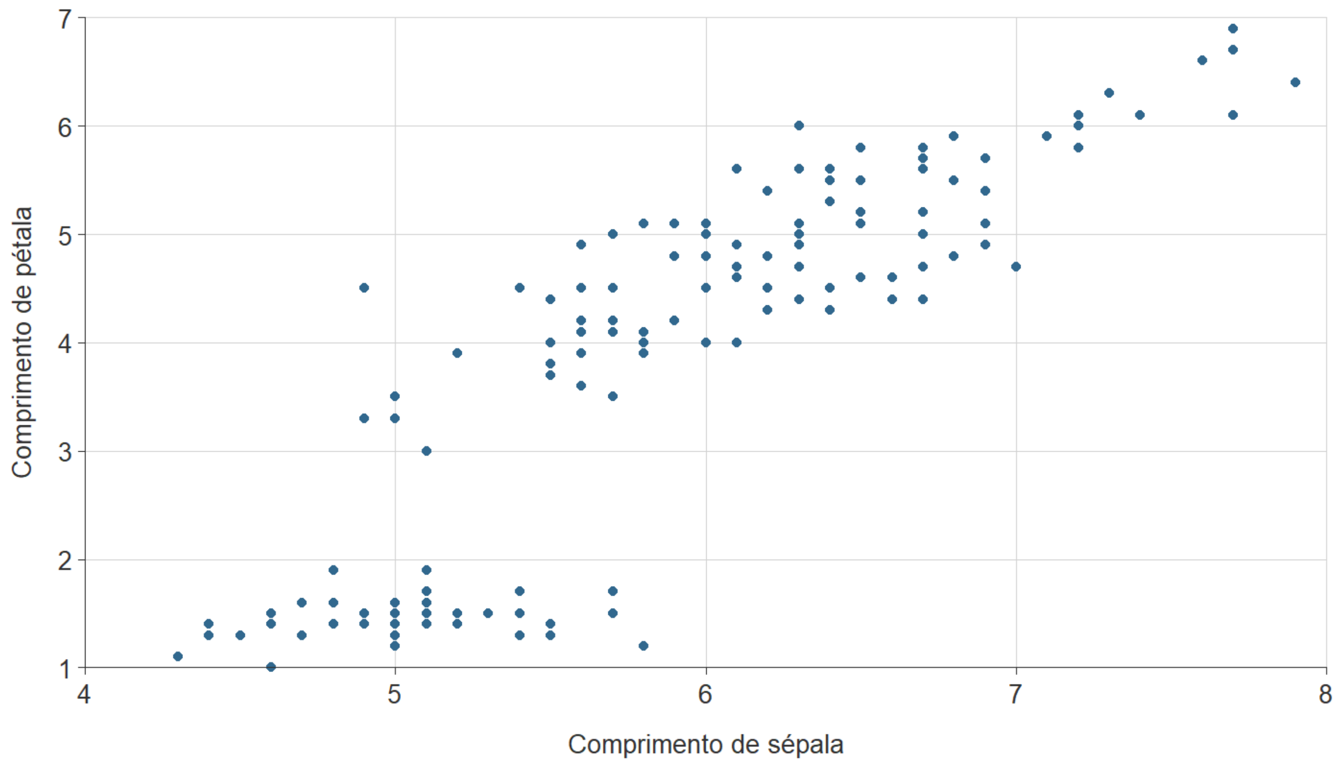
Associação entre duas variáveis quantitativas

Para duas variáveis quantitativas, estudamos a associação entre as duas variáveis usando o gráfico de dispersão. Além disso, podemos calcular o coeficiente de correlação linear de Pearson.

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

ggplot(data = df_iris) +
  geom_point(aes(x = Sepal.Length, y = Petal.Length)) +
  labs(x = "Comprimento de sépala", y = "Comprimento de pétala",
       title = "Associação entre duas variáveis quantitativas") +
  theme_calc()
```



```r
library(simplevis)

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

gg_point(df_iris,
         x_var = Sepal.Length, y_var = Petal.Length,
         x_title = "Comprimento de sépala", y_title = "Comprimento de pétala")
```



Associação entre duas variáveis quantitativas

Também podemos calcular o coeficiente de correlação linear de Pearson. Lembre que se X e Y são duas variáveis quantitativas com valores

Amostra de duas variáveis quantitativas X e Y.

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_1	$\cdots$	y_n

Então, o coeficiente de correlação linear é dado por

$$r = \left(\frac{(x_1 - \overline{x})}{s_x} \cdot \frac{(y_1 - \overline{y})}{s_y} \right) + \cdots + \left(\frac{(x_n - \overline{x})}{s_x} \cdot \frac{(y_n - \overline{y})}{s_y} \right).$$

```
cor(df_iris$Sepal.Length, df_iris$Petal.Length)
```

```
## [1] 0.8717538
```