Estatística descritiva usando R bem-vinde ao `tidyverse`

Teste de hipóteses usando R

Profa Carolina e Prof Gilberto

Parte 5

Inferência estatística

Estimação pontual: aproximação de parâmetro.
Exemplo: Estimar a nota média em matemática dos candidatos do ENEM na cidade de Salvador.

Estimação intervalar: estimativa intervalar para o parâmetro.
Exemplo: Encontrar a e b tal que a nota média de matemática esteja entre a e b com alguma *confiança*.

Teste de hipóteses: decisão entre duas hipóteses complementares.
Exemplo: Decidir entre duas hipóteses

$$H_0 : \text{a média em matemática no enem em salvador é no máximo } 600$$
$$H_1 : \text{a média em matemática no enem em salvador é maior que } 600$$

Introdução
Definições iniciais

Introdução

Objetivo

- Decidir entre H_0 e H_1 usando apenas a amostra
- H_0 é negação de H_1 e H_1 é negação de H_0 (complementares)
- H_0 é chamada de hipótese nula
- H_1 é chamada de hipótese alternativa

Erros que podemos cometer

- **Erro tipo I** ou **falso positivo**: rejeitar H_0 quando H_0 é verdadeira
- **Erro tipo II** ou **falso negativo**: não rejeitar H_0 quando H_1 é verdadeira

Introdução

		Situação na população: H_0	Situação na população: H_1
Decisão	H_0 (negação de H_1)	Sem erro (verdadeiro positivo)	**Erro tipo II** (*falso negativo*)
Decisão	H_1 (negação de H_0)	**Erro tipo I** (*falso positivo*)	Sem erro (verdadeiro negativo)

Sobre H_0 e H_1

Uso mais comum

- Verificar se o parâmetro mudou de valor em um novo cenário
- Validar uma hipótese científica (modelo ou teoria)
- Checar especificações (do mercado e/ou regulador)

Roteiro para especificar H_0 e H_1

- H_0: valor padrão ou comum (do mercado e/ou do regulador)
- H_1: sua hipótese de pesquisa ou pergunta
- **Dica prática:** igual matemática sempre fica em H_0

Nomenclaturas

Testes mais usados (e com nomes especiais).

- **Teste bilateral:** $H_0 : \theta = \theta_0$ contra $H_1 : \theta \neq \theta_0$
- **Teste unilateral superior:** $H_0 : \theta \leq \theta_0$ contra $H_1 : \theta > \theta_0$
- **Teste unilateral inferior:** $H_0 : \theta \geq \theta_0$ contra $H_1 : \theta < \theta_0$

Geralmente θ é: média da população (μ), desvio padrão da população (σ) e proporção de sucesso na população (p).

Interpretação e uso

- Por convenção, o erro mais grave é **falso positivo** ou **erro tipo I**
- Estabelecemos H_0 e H_1 para controlar o **falso positivo**

Exemplo

Em um julgamento podemos cometer duas hipóteses:

- o réu é culpado
- o réu é inocente

e podemos cometer dois erros:

- o réu é inocente mas o Juiz *decide* que o réu é culpado **(erro mais grave)**
- o réu é culpado mas o Juiz *decide* que o réu é inocente **(erro menos grave)**

Interpretação e uso

O **falso positivo** é $\underbrace{\text{o réu é inocente}}_{H_0}$ mas o Juiz *decide* que $\underbrace{\text{o réu é culpado}}_{H_1}$.

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

Interpretação e uso

Muito importante:

- **Apenas decidimos por o réu é culpado se tivermos evidência (prova).**
  Na ausência de evidências (provas), melhor continuar acreditando que o réu é inocente.
- **Sempre começamos acreditando na inocência do réu, apenas passamos a acreditar na culpa do réu com evidência(prova).**

- **Apenas decidimos por H_1 se tivermos evidência.**
  Na ausência de evidências, melhor continuar acreditando que H_0 é verdade.
- **Sempre começamos acreditando em H_0, apenas passamos a acreditar em H_1 com evidência.**

Para destacar isso, em Estatística falamos:

- **Decisão por H_0:** não rejeitamos H_0
- **Decisão por H_1:** rejeitamos H_0

Digressão

Sem provas robustas e convincentes, o juiz não rejeita a inocência do réu, ou seja, o juiz não rejeita H_0.

Mas o réu pode ser culpado, você apenas não conseguiu provas e a decisão por H_0 é "mais fraca".

Em um teste de hipóteses, o pesquisador exerce o papel de Juiz, ou seja, os dados precisam fornecer provas robustas e convicentes para rejeitar H_0 e se os dados não oferecem provas robustas e convincentes não rejeitamos H_0.

Controlando os erros

Probabilidade dos erros:

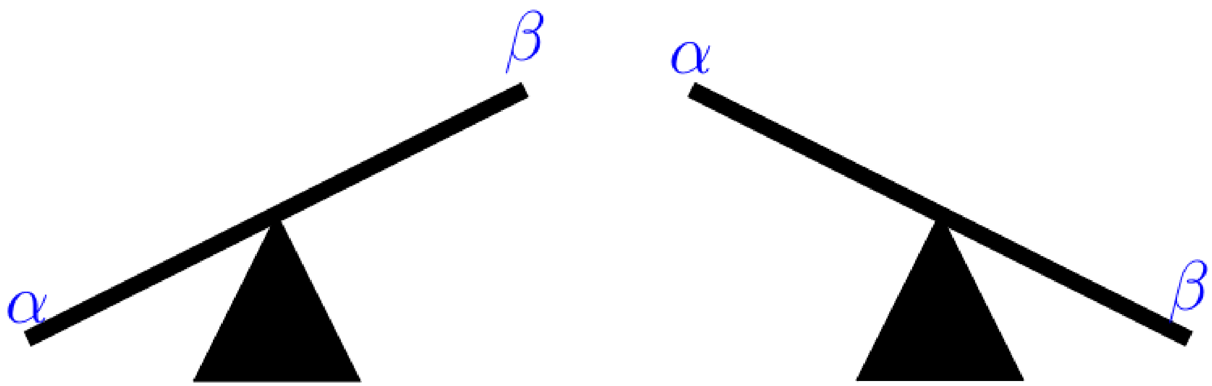
- $\alpha = P(\text{falso positivo}) = P(\text{erro tipo I})$
- $\beta = P(\text{falso negativo}) = P(\text{erro tipo II})$

Tomar uma decisão que minimize simultaneamente α e β.

Problema

Impossível decidir minizando simultaneamente α e β.

O que queremos



Controlando os erros

- α: Nível de significância, erro α ou tamanho do teste. Geralmente usamos $\alpha = 5\%$.
- β: erro β.
- $1 - \beta$: poder do teste de hipóteses.

Tamanho dos erros, nível de significância e poder do teste

Quanto maior a amostra, menor o nível de significânccia e maior o poder do teste.

- $X \sim N(\mu; 1,25^2)$
- $H_0 : \mu = 5$ e $H_1 : \mu \neq 5$
- Regra de decisão:
  - se $4,80 \leq \bar{x} \leq 5,20$, então não rejeitamos H_0
  - se $\bar{x} < 4,80$ ou $\bar{x} > 5,20$, então rejeitamos H_0

Nível de significância e poder do teste ao aumentarmos o tamanho da amostra.			
Tamanho da amostra	**Falso positivo**	**Falso negativo** (μ=4,6)	**Falso negativo** (μ=5,3)
25	0,42	0,20	0,32
50	0,26	0,13	0,28
75	0,17	0,08	0,24
100	0,11	0,05	0,21
250	0,01	0,01	0,10
500	0,00	0,00	0,04
750	0,00	0,00	0,01
1000	0,00	0,00	0,01

Valor-P

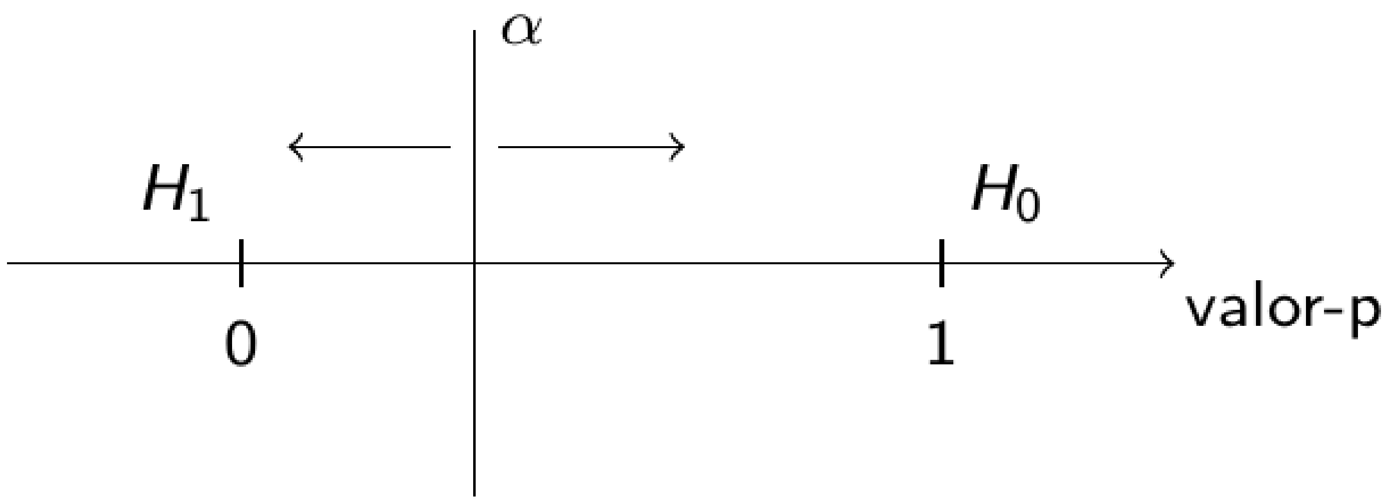
Descrição intuitiva

- Vamos chamar a **possibilidade** ou **plausibilidade** ou **indicação** da hipótese alternativa (H_1) de estatística do teste.

- O valor-p , *p-value* em inglês, é a probabilidade de coletar uma outra amostra com **estatística do teste** igual ou mais extrema do que a amostra observada 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 α.

Valor-p



Interpretação

- $X \sim N(0, 1)$
- $H_0 : \mu = 0$ contra $H_1 : \mu \neq 0$
- Algumas vezes decidimos por H_0 e outras vezes decidimos por H_1 usando o valor-p.

Interpretação

Se H_0 for verdade, em $100 \cdot (1 - \alpha)\%$ das amostras vamos decidir por H_0.
Se H_1 for verdade, em $100 \cdot (1 - \beta)\%$ das amostras vamos decidir por H_1.

Se H_0 for verdade, em $100 \cdot \alpha\%$ das amostras vamos errar o **falso positivo**.
Se H_1 for verdade, em $100 \cdot \beta\%$ das amostras vamos errar o **falso negativo**.

```r
df_amostras <- read_xlsx("data/raw/dados.xlsx", sheet = "p-valor-motivacao")

df_amostras  |>
  group_by(amostras) |>
  summarise(x1 = x[1], x2 = x[2], x3 = x[3], x4 = x[4], x5 = x[5],
            valor_p = ht_1pop_mean(x)$p_value)  |>
  mutate(Decisao = c("H1", "H1", "H0", "H0"))
```

```
## # A tibble: 4 × 8
##   amostras     x1    x2    x3    x4    x5 valor_p Decisao
##   <chr>     <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>
## 1 Amostras1 -0.98 -0.29 -0.49 -0.94 -0.59 0.00769 H1
## 2 Amostras2 -0.36 -0.69 -1.24 -0.61 -1.18 0.00866 H1
## 3 Amostras3 -2.07  1.02  0.92  0.44  0.85 0.711   H0
## 4 Amostras4 -0.74  1.11  0.21 -0.4   0.88 0.583   H0
```

Pacote `statBasics`

Testando a média da população

Nota de matemática em Salvador é maior que 600?
Decidir usando nível de significância $\alpha = 5\%$.

- $H_1 : \mu > 600$ contra $H_0 : \mu \leq 600$

```
df_enem <- read_xlsx("data/raw/amostra_enem_salvador.xlsx")
df_enem <- clean_names(df_enem)

ht_1pop_mean(df_enem$nu_nota_mt, mu = 600,
             alternative = "greater", sig_level = 0.05)
```

```
## # A tibble: 1 × 7
##   statistic p_value critical_value critical_region alternative    mu sig_level
##       <dbl>   <dbl>          <dbl> <chr>           <chr>       <dbl>     <dbl>
## 1     -40.8       1           1.65 (1.645, Inf)    greater       600      0.05
```

Ao nível de significâncccia 5%, não rejeitamos H_0.
Ao nível de significância 5%, não temos *evidência* de que a nota de matemática é maior que 600.

Testando a proporção de sucesso

As pessoas são negras são maioria no ENEM na cidade de Salvador?
Decidir usando nível de significânccia $\alpha = 1\%$.

- $H_0 : p \leq 0,5$ contra $H_1 : p < 0,5$

```r
df_enem <- read_xlsx("data/raw/amostra_enem_salvador.xlsx")
df_enem <- clean_names(df_enem)

df_enem <- df_enem |> mutate(raca = dplyr::recode(tp_cor_raca,
  "Preta" = 1, "Parda" = 1, "Amarela" = 0, "Branca" = 0, "Indígena" = 0,
  "Não declarado" = 0
))

ht_1pop_prop(df_enem$raca, proportion = 0.5, alternative = "greater", sig_level = 0.01)
```

```
## # A tibble: 1 × 7
##   statistic p_value critical_value critical_region alternative proportion sig_level
##       <dbl>   <dbl>          <dbl> <chr>           <chr>            <dbl>     <dbl>
## 1      33.7       0           2.33 (2.326, Inf)    greater            0.5      0.01
```

Ao nível de significância 1%, temos evidência para a proporção de pessoas negras na prova do ENEM.

Testando a proporção de sucesso

Lula ganhará no primeiro turno na eleição de 2022?
Usaremos dados da pesquisa realizada pelo PoderData: detalhes da pesquisa.
3500 responderam a pesquisa e 1505 afirmaram que votariam em Lula.

- $H_1 : p > 0,5$ e $H_0 : p \leq 0,5$
  $\alpha = 5\%$

```
ht_1pop_prop(1505, 3500, proportion = 0.5, alternative = "greater", sig_level = 0.05)
```

```
## # A tibble: 1 × 7
##   statistic p_value critical_value critical_region alternative proportion sig_level
##       <dbl>   <dbl>          <dbl> <chr>           <chr>            <dbl>     <dbl>
## 1     -8.28       1           1.64 (1.645, Inf)    greater            0.5      0.05
```

Testando o desvio padrão

O desvio da nota de matemática no ENEM dos brancos é menor que 100?

- $H_1 : \sigma < 100$ e $H_0 : \sigma \geq 100$
  $\alpha = 5\%$

```
df_enem <- read_xlsx("data/raw/amostra_enem_salvador.xlsx")
df_enem <- clean_names(df_enem) |>
  filter(tp_cor_raca == "Branca")

ht_1pop_var(df_enem$nu_nota_mt, sigma = 100, alternative = "less")
```

```
## # A tibble: 1 × 7
##   statistic p_value critical_value critical_region alternative sigma sig_level
##       <dbl>   <dbl>          <dbl> <chr>           <chr>       <dbl>     <dbl>
## 1     1589.       1          1046. (0, 1046)       less          100      0.05
```

Ao nível de significância 5%, não podemos afirmar que o desvio padrão da nota no ENEM em matemática dos brancos é menor que 100