Preparado para: AT autoridade tributária e aduaneira

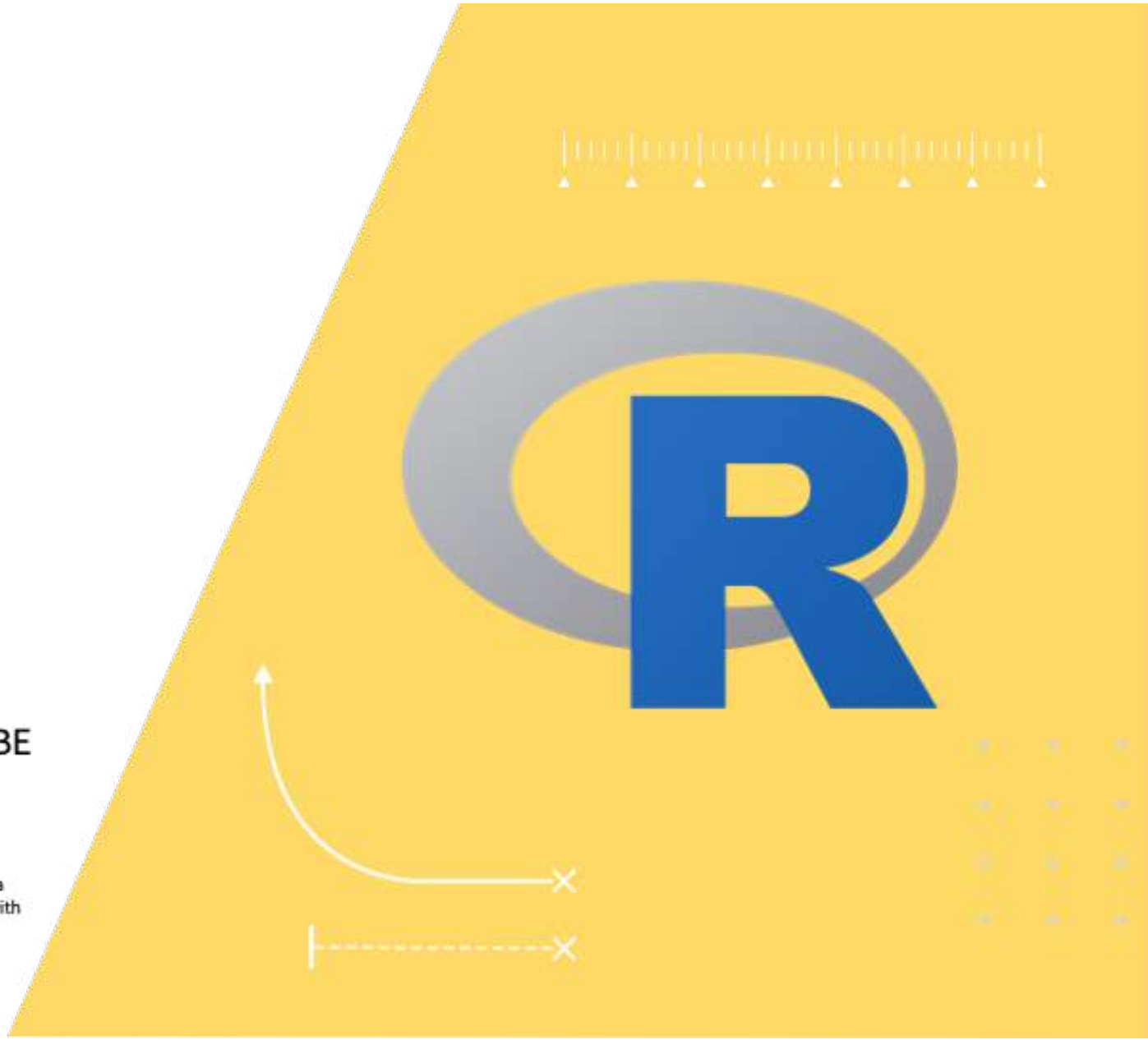
REFORM/SC2022/126
DELIVERABLE 4
MÓDULO 3
ESTATÍSTICA BÁSICA EM R

DESIGNING A NEW VALUATION MODEL
FOR RURAL PROPERTIES IN PORTUGAL

Parte II (cont.)

Formador: Luís Teles Morais | Nova SBE

Lisboa, 22 junho 2023



This project is carried out with funding by the European Union via the Structural Reform Support Programme and in cooperation with the Directorate General for Structural Reform Support of the European Commission

AARC NOVA LOBO VASQUES esri INNERLANDS

Programa

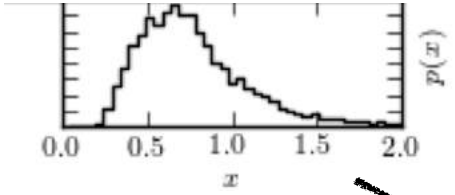
Módulos	Duração
Módulo 1 – Introdução ao R: - O que é o R? - Como instalar e configurar o R. - Sintaxe básica e comandos. - Tipos de dados, objetos e classes.	**4 Horas**
Módulo 2 – Gestão e tratamento de dados em R: - Carregar dados no R. - Perceber as estruturas de dados e *subsetting*. - Limpeza de dados: *missing values, outliers* e transformações - Juntar bases de dados	**8 Horas**
Módulo 3 – Estatística básica em R: - Estatísticas descritivas: medidas de dispersão central e variação. - Distribuições probabilísticas: variáveis discretas e contínuas. - Testes de hipóteses.	**8 Horas**

Módulos	Duração
Módulo 4 – Regressão Linear: - O modelo classico linear. - Estimação de parametros segundo o MMQ. - Testes de hipóteses: significância estatística e ajuste do modelo. - Modelo de regressão múltipla. - Testar as premissas: multicolinearidade, heteroscedasticidade e normalidade dos resíduos. - Critérios de seleção dos modelos.	**12 Horas**
Módulo 5 – O modelo: - Estrutura do modelo e premissas – Perceber o modelo (4 Hours). - Uso e tratamento dos dados (4 Hours). - Descrição do modelo (4 Hours). - Aplicação do modelo a cada piloto (12 Hours). - Aplicação autónoma do modelo a uma região (8 Hours).	**32 Horas**

Estatística básica em R (III)

Testes de hipóteses

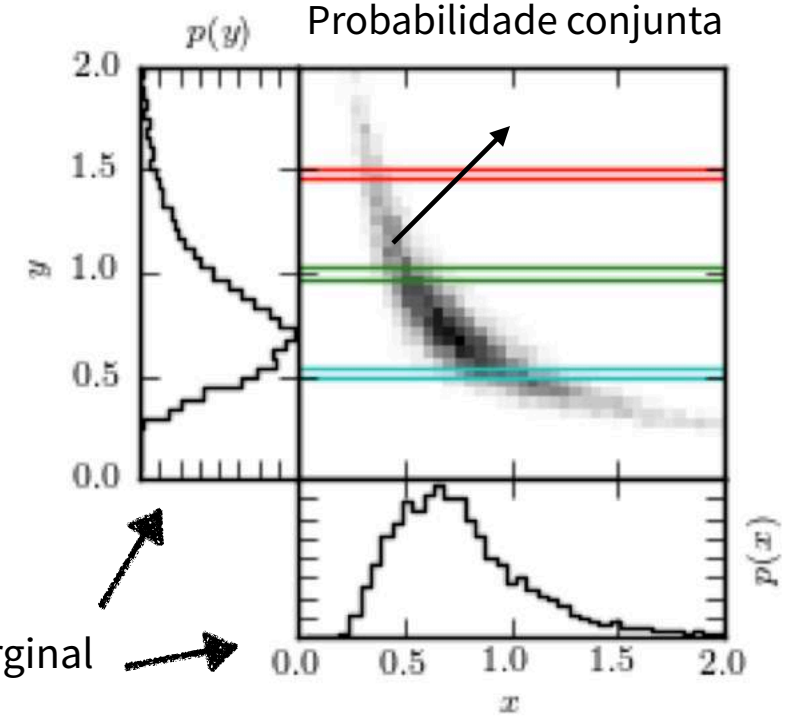
Correlação vs. causalidade



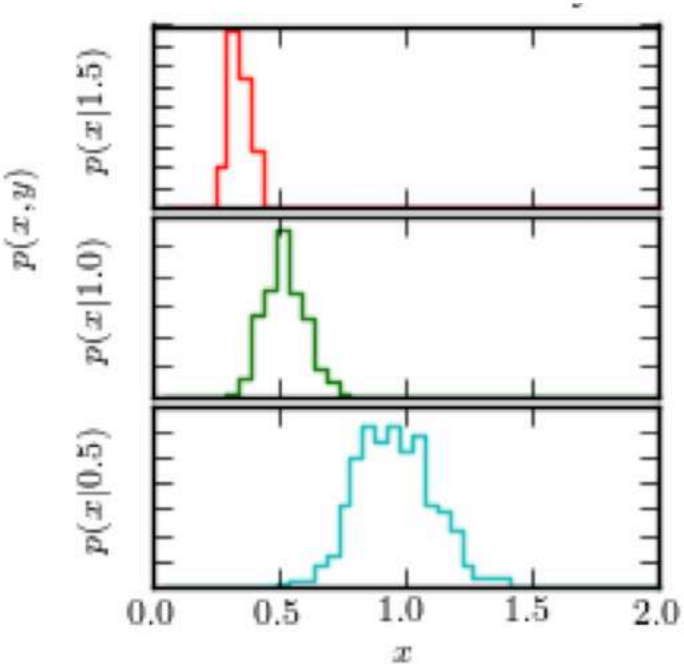
Distribuição univariada

Distribuição bivariada

Probabilidade conjunta

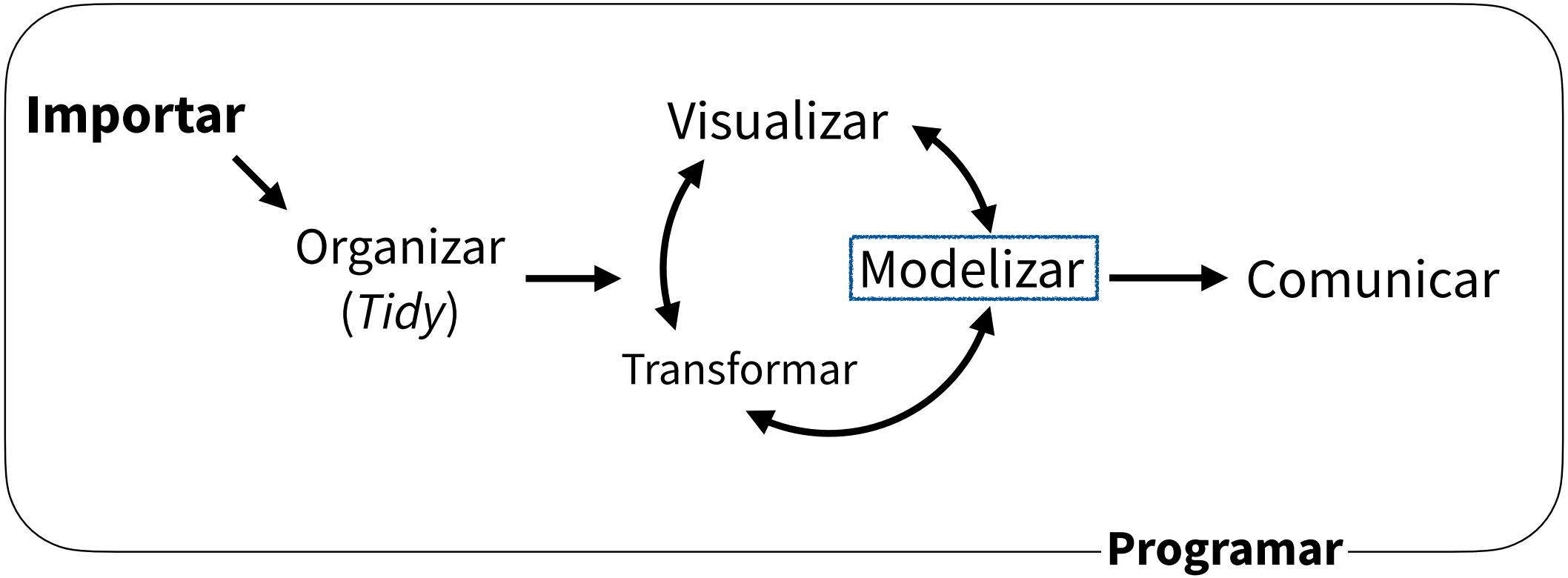


Probabilidade condicional



Probabilidade(s) marginal

Ciência de dados



Alguns conceitos

Visualizar

Estatística **descritiva**

- Analisar um conjunto de dados, reduzindo-o a medidas sumárias simples

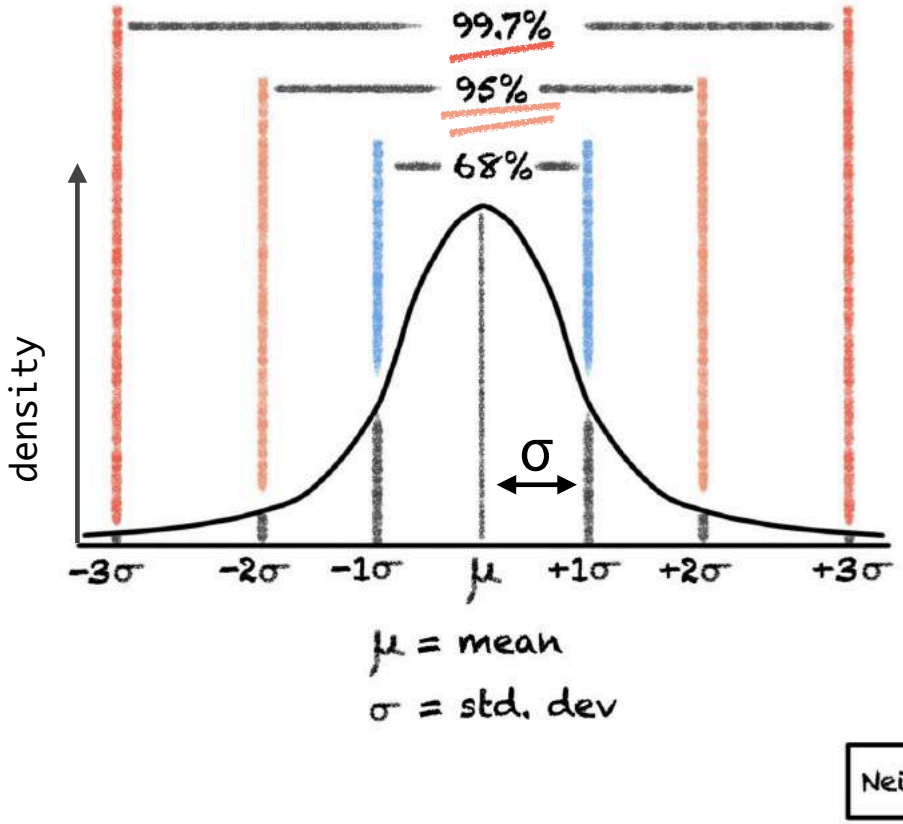
Modelizar

Inferência estatística

- Calcular ou estimar algo que não podemos observar diretamente, a partir dos dados existentes

Distribuição normal

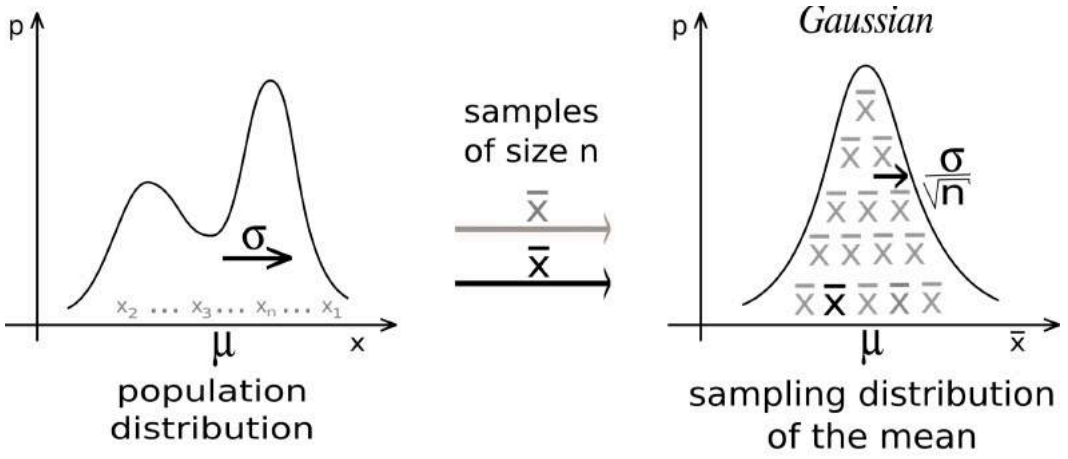
- Muitas variáveis na realidade são aproximadamente “normais”



- Também curva de Gauss
- ~N(μ,σ):
  - média μ e desv. pad. σ
  - ditam posição e dispersão

Lei dos grandes números

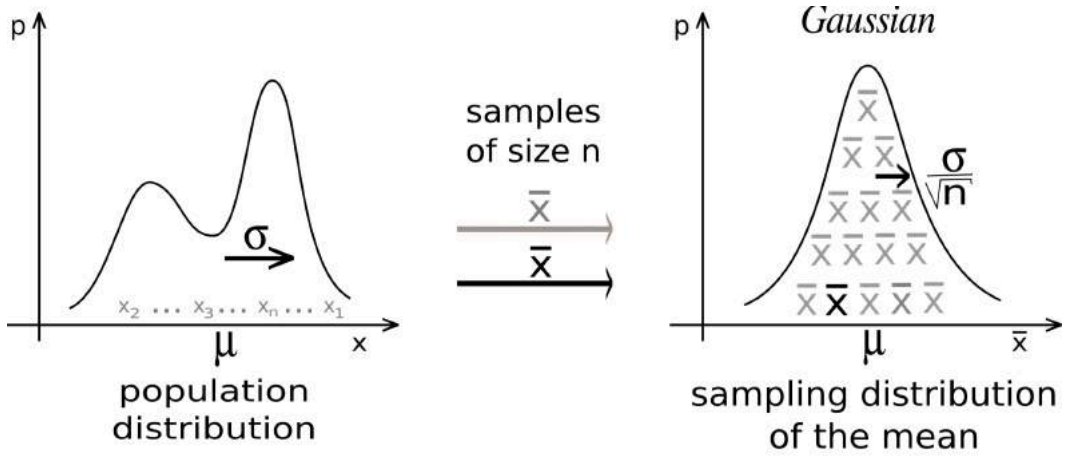
- **Sim!** Lei dos grandes números: a média da amostra acerta, em média, na média da população, seja qual for a distribuição da variável
- Formalmente: $$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{X_i}{n} = \overline{X}$$



- E.x.: valor médio do dado -> converge sempre para 3.5, se o atirar vezes suficientes

Teorema do limite central

- **E mais:** a distribuição das estimativas converge para uma distribuição normal, independentemente da distribuição da variável inicial
- Formalmente: $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{a} \mathcal{N}\left(0, \sigma^2\right)$



- Permite a inferência: sei a probabilidade associada ao valor obtido como estimativa, mesmo sem conhecer a forma da distribuição da variável

Teste de hipóteses

Exemplo médico

- **Facto para efeitos desta aula:** está comprovado que a probabilidade de surgirem complicações graves é, para qualquer hospital em Portugal, **10%**.

> - *Chefe do Serv. de Cirurgia do Hospital da Guarda:*
> Graças às práticas que implementámos, apenas tivemos 30 casos complicados no ano de 2022.

- A sua tese: isto comprova que as melhorias que implementou conduziram a uma redução da probabilidade de existirem complicações graves no Hospital da Guarda, face ao que acontece no resto do país.

Teste de hipóteses

Exemplo médico

- **Facto para efeitos desta aula:** está comprovado que a probabilidade de surgirem complicações graves é, para qualquer hospital em Portugal, **10%**.

> - *Chefe do Serv. de Cirurgia do Hospital da Guarda:*
> Graças às práticas que implementámos, apenas tivemos 30 casos complicados no ano de 2022.

- A sua tese: isto comprova que as melhorias que implementou conduziram a uma redução da probabilidade de existirem complicações graves no Hospital da Guarda, face ao que acontece no resto do país. **Será?**

Dados: Exercício

- O n.º total de cirurgias no Hospital da Guarda em 2022 foi de 500.
- Crie de raíz, através de código, uma tabela com um registo para cada cirurgia e uma variável binária que assinala se existiram complicações. Dica:
  - Primeiro, crie uma tabela em que a variável indica que não houve complicações em todos os registos
  - Segundo, altere o valor apenas das primeiras 30 linhas

Dados: Exercício

- O n.º total de cirurgias no Hospital da Guarda em 2022 foi de 500.
- Crie de raíz, através de código, uma tabela com um registo para cada cirurgia e uma variável binária que assinala se existiram complicações. Dica:
  - Primeiro, crie uma tabela em que a variável indica que não houve complicações em todos os registos
  - Segundo, altere o valor apenas das primeiras 30 linhas

```
cirurgias_Guarda <- tibble(
  no_cirurgia = seq(0, 500),
  I_Complicacoes = FALSE
)
# Variável binária => utilizar valores do tipo logical
cirurgias_Guarda$I_Complicacoes[0:30] = TRUE
```

Dados: Exercício 2

- Como é que poderia realizar o mesmo procedimento acima (alterar só aqueles valores para TRUE), mas utilizando uma função do tidyverse?

Dados: Exercício 2

- Como é que poderia realizar o mesmo procedimento acima (alterar só aqueles valores para TRUE), mas utilizando uma função do tidyverse?

```
cirurgias_Guarda <- cirurgias_Guarda %>%
  mutate(resultado = if_else(no_cirurgia <= 30, TRUE, FALSE))

# Bonus: sem utilizar a variável n.º da cirurgia
cirurgias_Guarda %>% mutate(resultado = if_else(n() <= 30, TRUE, FALSE))
```

```
## # A tibble: 501 × 3
##   no_cirurgia I_Complicacoes resultado
##         <int> <lgl>          <lgl>
## 1           0 TRUE           FALSE
## 2           1 TRUE           FALSE
## 3           2 TRUE           FALSE
## 4           3 TRUE           FALSE
## 5           4 TRUE           FALSE
## 6           5 TRUE           FALSE
## # ℹ 495 more rows
```

Parâmetro e medida ou estimativa

p : probabilidade de existirem complicações graves em cirurgias em Portugal $= 10\%$.

$\hat{p}$: percentagem de complicações na Guarda em 2022 (na amostra) = $\frac{30}{500} = 6\%$

Parâmetro e medida ou estimativa

p : probabilidade de existirem complicações graves em cirurgias em Portugal $= 10\%$.

$\hat{p}$: percentagem de complicações na Guarda em 2022 (na amostra) = $\frac{30}{500} = 6\%$

Correlação e causalidade

> É possível confirmar a afirmação do diretor com estes dados?

Parâmetro e medida ou estimativa

p : probabilidade de existirem complicações graves em cirurgias em Portugal $= 10\%$.

$\hat{p}$: percentagem de complicações na Guarda em 2022 (na amostra) = $\frac{30}{500} = 6\%$

Correlação e causalidade

> É possível confirmar a afirmação do diretor com estes dados?

- Não -- **correlação vs. causalidade**: a probabilidade de existirem complicações na Guarda pode ser diferente do resto do país por outros motivos que não as alterações promovidas
  - Ainda assim, podemos aferir se a taxa obtida, 6%, significa realmente que a probabilidade na Guarda é mais baixa na Guarda face ao resto do país.
  - Ou se pelo contrário, em 2022 a taxa foi mais baixa por mero acaso.

2 hipóteses

O teste de hipóteses começa por definir a hipótese nula e a alternativa:

- **Hipotese nula:** "Não há diferença entre a Guarda e o resto do país"

2 hipóteses

O teste de hipóteses começa por definir a hipótese nula e a alternativa:

- **Hipotese nula:** "Não há diferença entre a Guarda e o resto do país"
- **Hipótese alternativa:** "A taxa é de facto mais baixa na Guarda"

Um teste de hipóteses é como um julgamento no tribunal

- **Hipotese nula**, H_0: O réu é inocente
- **Hipótese alternativa**, H_A: O réu é culpado

Um teste de hipóteses é como um julgamento no tribunal

- **Hipotese nula**, H_0: O réu é inocente
- **Hipótese alternativa**, H_A: O réu é culpado
- **Apresentar o material probatório:** Recolher os *dados*

Um teste de hipóteses é como um julgamento no tribunal

- **Hipotese nula**, H_0: O réu é inocente
- **Hipótese alternativa**, H_A: O réu é culpado
- **Apresentar o material probatório:** Recolher os *dados*

Julgar com base nas provas:

- Se a hipótese nula for verdadeira:
  - (Ou seja, se a probabilidade de complicações continuar igual ao resto do país)
  - Será plausível ter sido obtido um valor tão baixo este ano?

Sim: Não se rejeita a H0 vs. **Não**: Rejeita-se a H0

Estrutura de qualquer teste de hipóteses

- Começa-se com a hipótese nula, H_0, que representa o status quo
- A hipótese alternativa, H_A, representa a pergunta de investigação, i.e. a informação nova, cuja verosimilhança queremos testar.
  - Para testar a hipótese de H_0 ser verdadeira, obtém-se o **p-value** $\rightarrow$ probabilidade de um resultado tão ou mais *extremo* que o observado, sob a hipótese nula
- Duas possibilidades:

p-value alto $\Rightarrow$ as provas não permitem rejeitar H_0

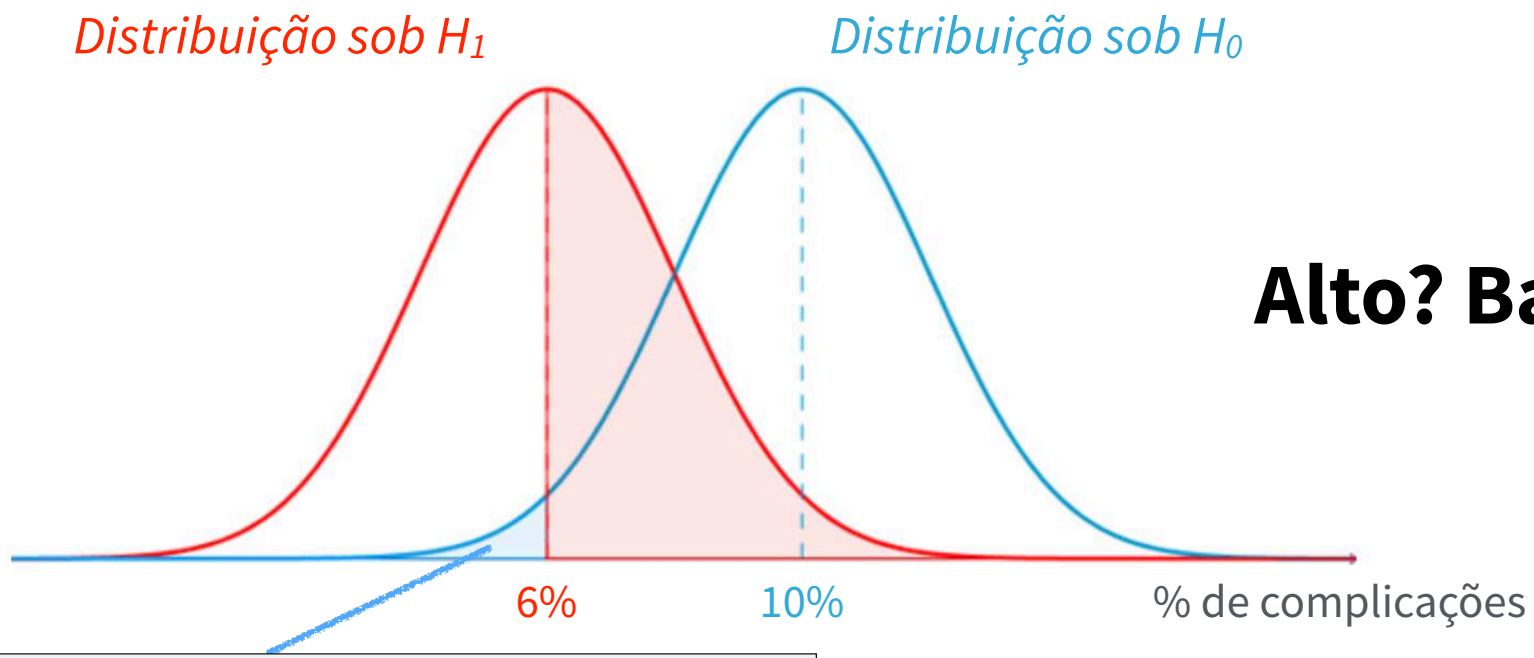
- A taxa de complicações observada (6%) parece plausível sob H_0

p-value baixo $\Rightarrow$ as provas permitem rejeitar H_0

- Sob H_0, parece difícil que se observasse uma taxa tão baixa

p-value

- Suponha que sabemos que a distribuição da % de complicações é em qualquer caso normal, com um desvio padrão de 5 p.p.
- Neste caso, podemos calcular a probabilidade de obter um resultado igual ou inferior a 6%.



Alto? Baixo?

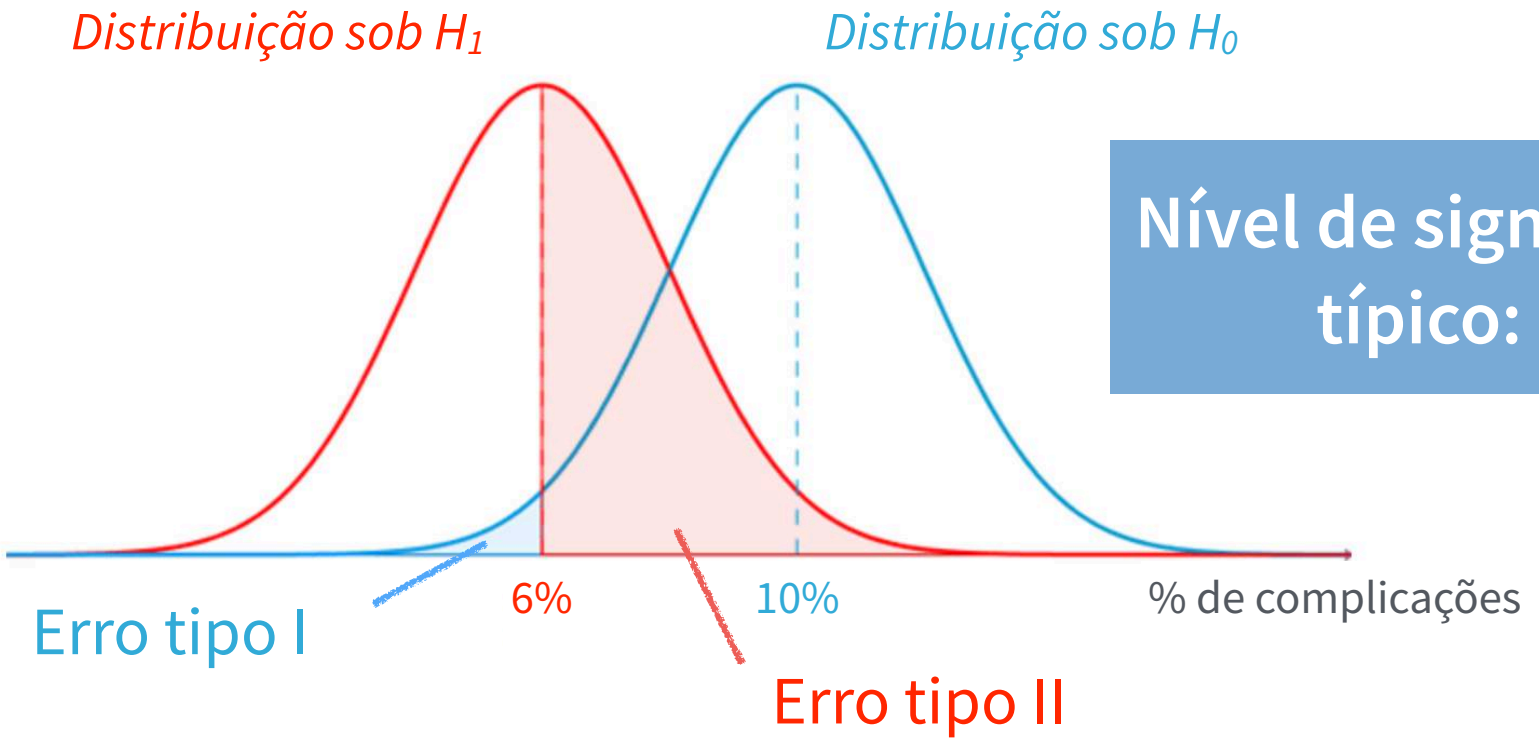
```
> pnorm(0.06, mean = 0.1, sd = 0.05)
[1] 0.2118554
```

Tipos de erro e nível de significância

	H0 é verdadeira	**H0 é falsa**
H0 rejeitada	**Erro do tipo I**	Não há erro
H0 não rejeitada	Não há erro	**Erro do tipo II**

- **Erro do tipo I:** O "inocente" é condenado
- **Erro do tipo II:** O "culpado" é ilibado
  - Como na vida real, temos mais tolerância com o tipo II

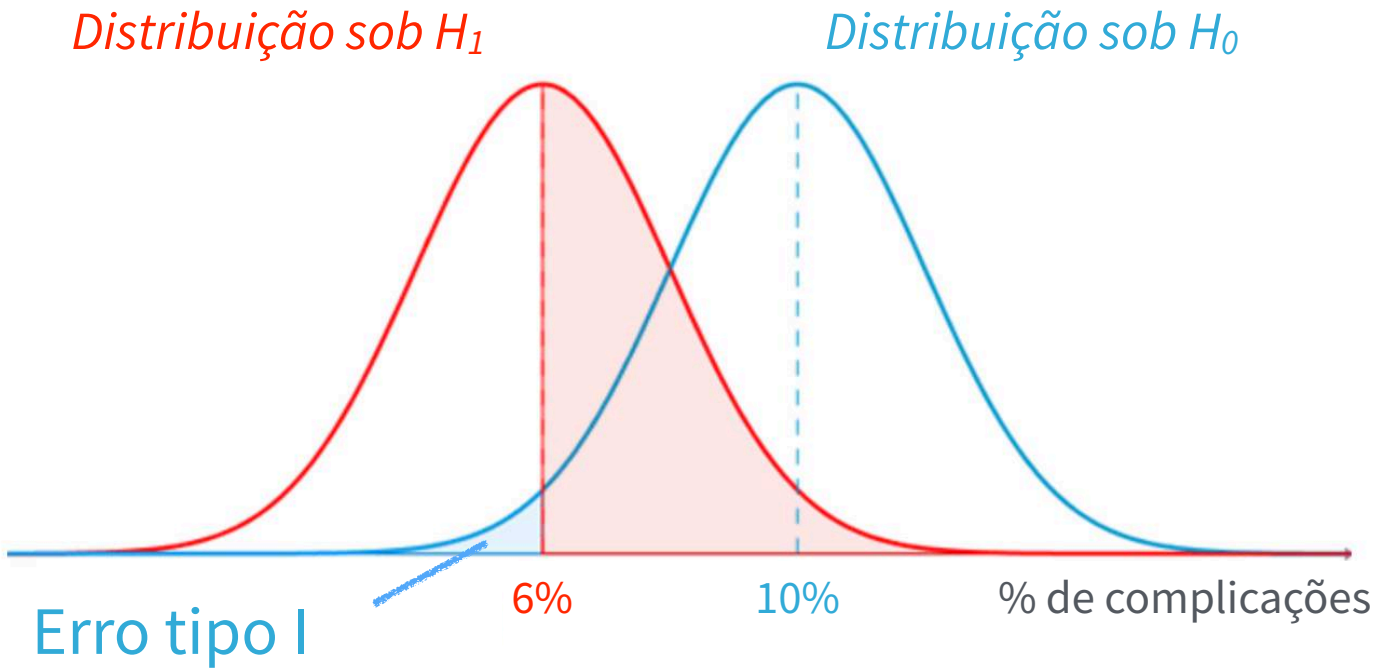
Erros e nível de significância



- **Nível de significância**: tolerância com erros do tipo I na avaliação das hipóteses.
- Preferência "subjetiva" vertida num valor concreto: a probabilidade máxima de cometer um erro do tipo I, sob a hipótese nula

Erros e nível de significância

- Neste caso, a probabilidade de cometer um erro do tipo I é 22%, na condição da hipótese nula ser verdadeira.
- **p-value = 22% > nível de significância = 5%**
  - rejeita-se H0




```
[1] 0.2118554
```

Obrigado
e bom fim-de-semana!

luis.morais@novasbe.pt