FIAP

MBA

MBA EM DATA SCIENCE & AI

STATISTICS WITH R

AULA 4

Amostragem

Testes de Hipótese

Um 'cadinho' de R ...

Se DF é um data frame, então:

DF[**x**, **y**] :acessa a linha **x** e coluna **y**
DF[**x**,] :acessa a linha **x** e todas as colunas
DF[, **y**] :acessa todas as linhas e coluna **y**

- `mtcars[1,10]`
- `mtcars[2,   ]`
- `mtcars[ , 3]`
- `mtcars[1:4, 2]`
- `mtcars[ ,c(2, 10, 1)]`

Reproduzindo experimentos

- Trabalhar com números aleatórios naturalmente faz com que se obtenha resultados igualmente aleatórios

- Quando se deseja que os resultados sejam reproduzidos (apesar da aleatoriedade), é preciso “plantar a semente” da aleatoriedade.

- para isso, se usa a instrução
  - `set.seed(seed)`

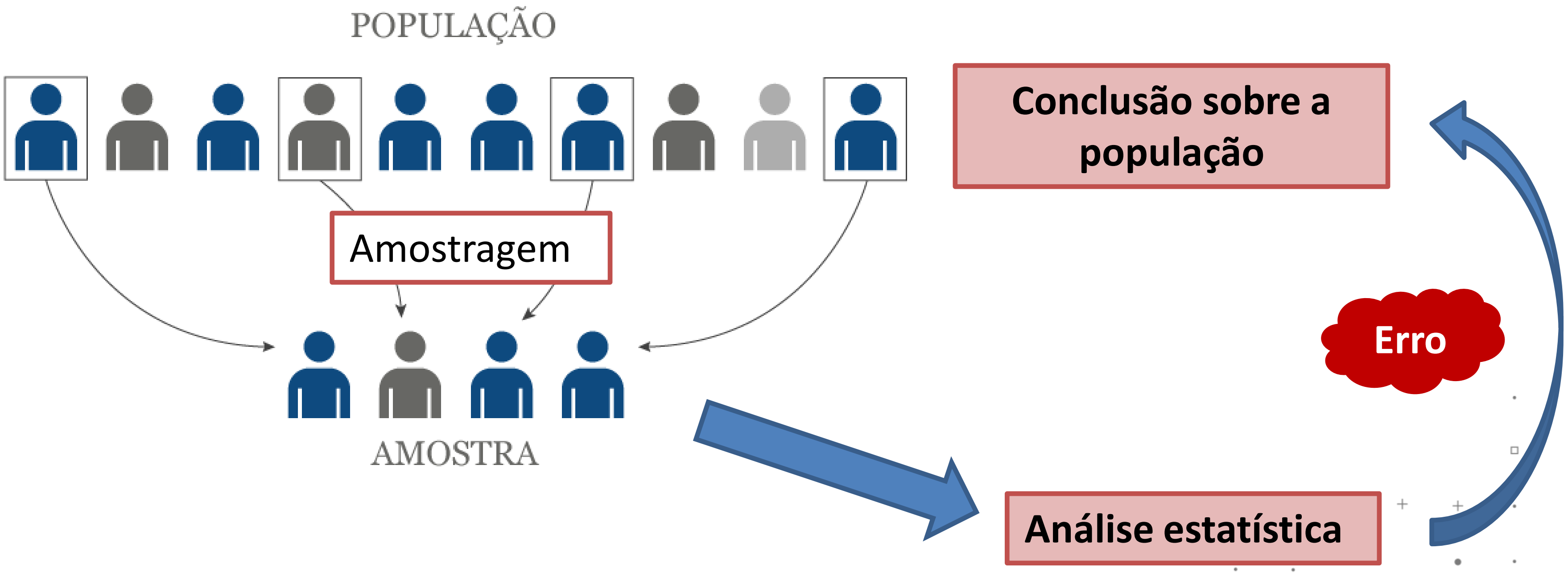
```
• set.seed(1)
• rnorm(5)
• rnorm(5)
• set.seed(1)
• rnorm(5)
```

Amostras a partir de um domínio

- comando `sample`

```
• set.seed(1)
• amostra = c( "T", "R", "I", "A", "N", "G", "U",
  "L", "O", "S")
• sample(x = amostra, replace = FALSE)
• sample(x = amostra, replace = TRUE)
• sample(x = amostra, size = 5)
• sample(x = amostra, size = 10, replace = TRUE,
  prob = c(1, 1, 5, 1, 1, 1, 1, 1, 1, 5))
```

Amostragem



POPULAÇÃO
Amostragem
AMOSTRA
Conclusão sobre a população
Erro
Análise estatística

Amostragem

- Pesquisa eleitoral
- Pesquisa com clientes
- Controle de qualidade de produtos
- Desenvolvimento de modelos estatísticos
  - Amostra de desenvolvimento (Treino)
  - Amostra de validação (Teste/OOS)

Amostragem

O que é necessário garantir?

- Que a amostra seja representativa da população A amostra deve possuir as mesmas características básicas da população, no que diz respeito às variáveis que desejamos pesquisar.

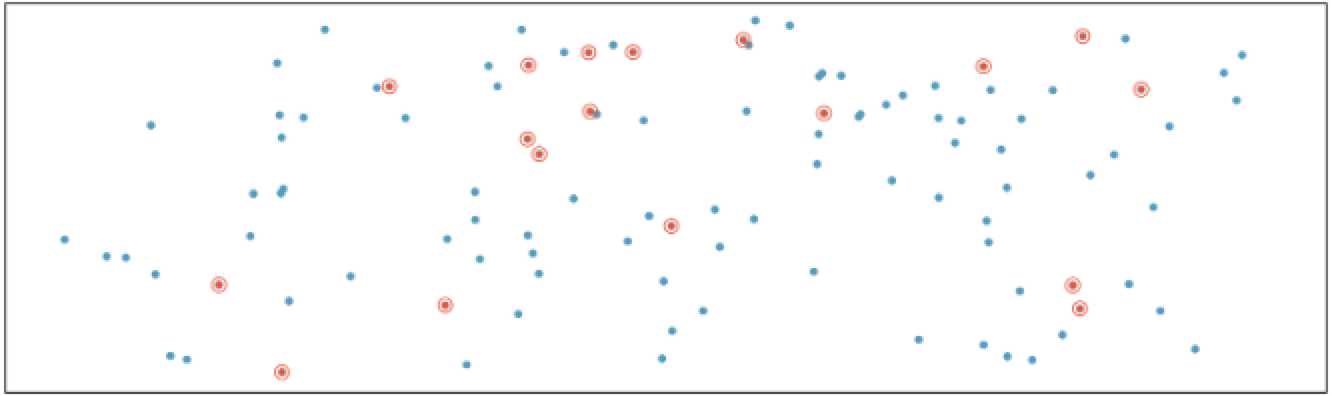
Tipos de amostragem

- PROBABILÍSTICA
  - ALEATÓRIA SIMPLES
  - SISTEMÁTICA
  - ESTRATIFICADA
  - CONGLOMERADO
- NÃO PROBABILÍSTICA (INTENCIONAL)
  - COTAS
  - PROCURA
  - ...

Tipos de amostragem

- PROBABILÍSTICA
  - **ALEATÓRIA SIMPLES**
  - **SISTEMÁTICA**
  - **ESTRATIFICADA**
  - **CONGLOMERADO**
- NÃO PROBABILÍSTICA (INTENCIONAL)
  - COTAS
  - PROCURA
  - ...

Aleatória simples



Sorteio de forma aleatória.

Aleatória simples

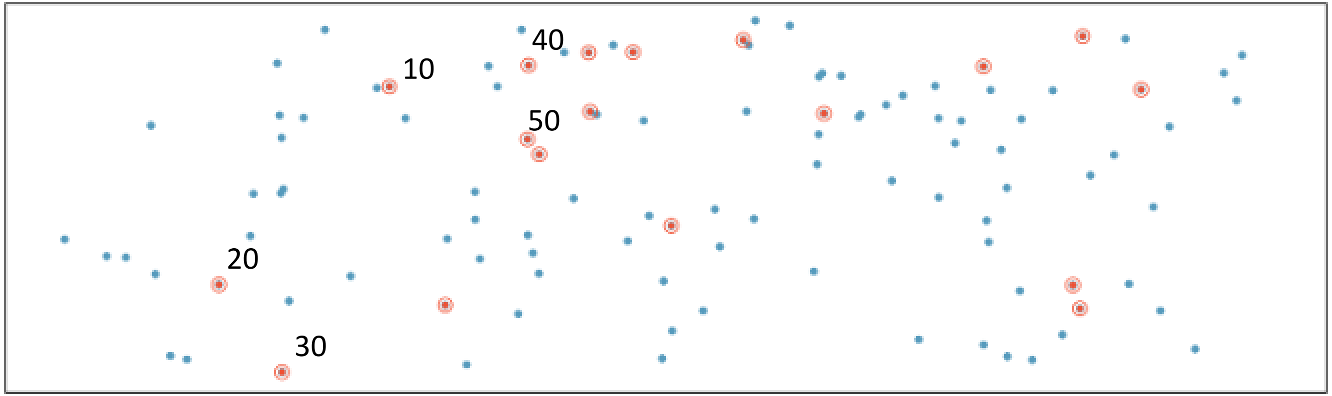
Comando `sample.int`

```
sample.int(n = 5, size = 2 )
sample.int(n = 5, size = 10, replace = TRUE )
```

Amostrando...

```
n <- nrow(imdb)
index <- sample.int(n, 100)
amostraAleat <- imdb[index,]
```

Sistemática



Sorteio baseado em uma estratégia. Ex: Selecionar a cada 10.

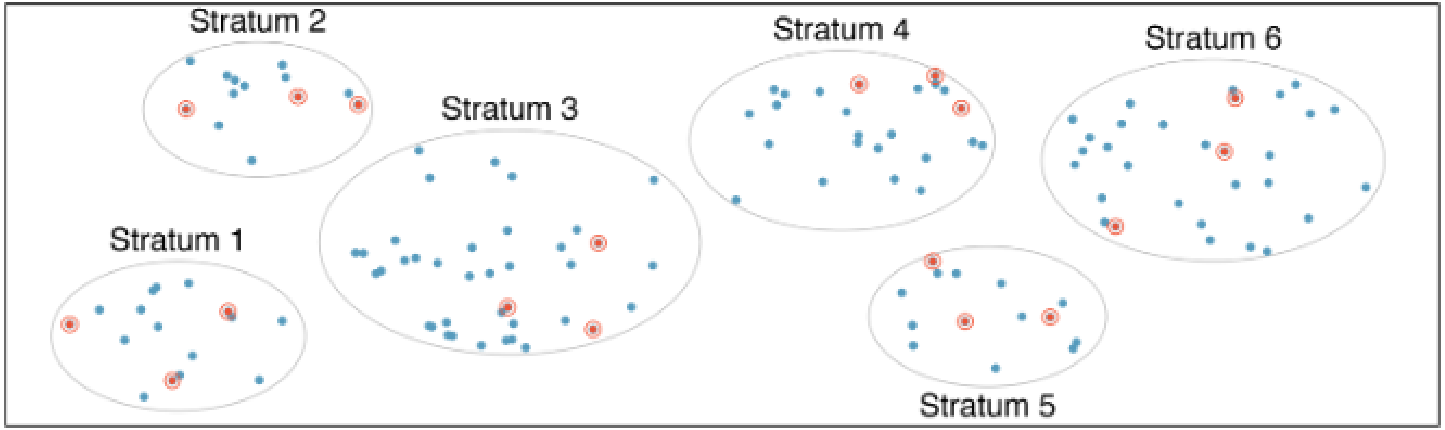
Sistemática

```
imdbindex <- imdb %>% mutate(id = 1:nrow(imdb)) ##opcional
##Amostrando a cada 10 filmes
index <- seq(10, nrow(imdbindex), by = 10)
amostraSist <- imdbindex[index,]
```

Ou usando dplyr

```
imdbindex <- imdb %>% mutate(id = 1:nrow(imdb)) ##necessária
##Amostrando a cada 10
filmesindex <- seq(10, nrow(imdbindex), by = 10)
amostraSist <- imdbindex %>% dplyr::filter(id %in% index)
```

Estratificada



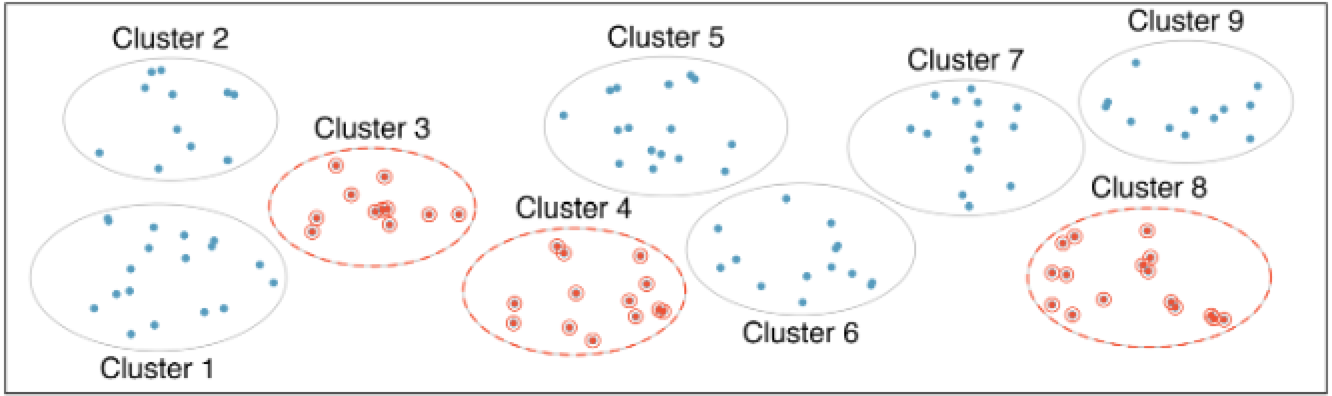
Sorteio de indivíduos dentro dos estratos

Estratificada

```
imdbcor <- imdb %>% filter(cor == "Color")
imdbbw <- imdb %>% filter(cor == "Black and White")

##Amostrando 100 filmes de cada tipo
idcor <- sample.int(nrow(imdbcor), 100)
idbw  <- sample.int(nrow(imdbbw), 100)
amostracor <- imdbcor[idcor,]
amostrabw  <- imdbbw[idbw,]
amostra <- bind_rows(amostracor, amostrabw)
```

Conglomerados



Sorteio de clusters e não dos indivíduos.

Conglomerados

```
#Criando Artificialmente 50 clusters (conglomerados)
imdbCluster <- imdb %>% mutate(cluster = sample.int(50,
nrow(imdb), replace = TRUE))

idcluster <- imdbCluster$cluster %>% unique()

#Amostrando 10 clusters
amostraid <- sample(x = idcluster, size = 10, replace =
FALSE)

amostraCluster <- imdbCluster %>% dplyr::filter(cluster
%in% amostraid)
```

Amostra de treino e teste

- Na construção de modelos é comum criarmos bases de treino e teste
- Vamos criar uma base de treino com 70% dos dados

```
• n <- nrow(mtcars)
• index <- sample.int(n, 0.7*n)
• treino <- mtcars[index,]
• teste <- mtcars[-index,]
```

Amostra de treino e teste

- Uma boa opção é usar o pacote Dplyr

```
treino <- imdbindex %>% sample_frac(0.7)
teste  <- dplyr::anti_join(imdbindex, treino, by = 'id')
```

Exercícios

Utilizar a base 'Customer Data for a Clothing Company' (veja o html para ver a descrição das variáveis)

- Faça amostragens:
  - Aleatória simples (400)
  - Estratificada (100 por segmento)

Use a semente 1234 e compare a variável store_exp por tipo de amostragem usando boxplot.

Definição: Probabilidade Condicional

Sejam A e B dois eventos pertencentes a um espaço amostral Ω.

Dizemos que a probabilidade de acontecer o evento A dado que aconteceu o evento B é definido por

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

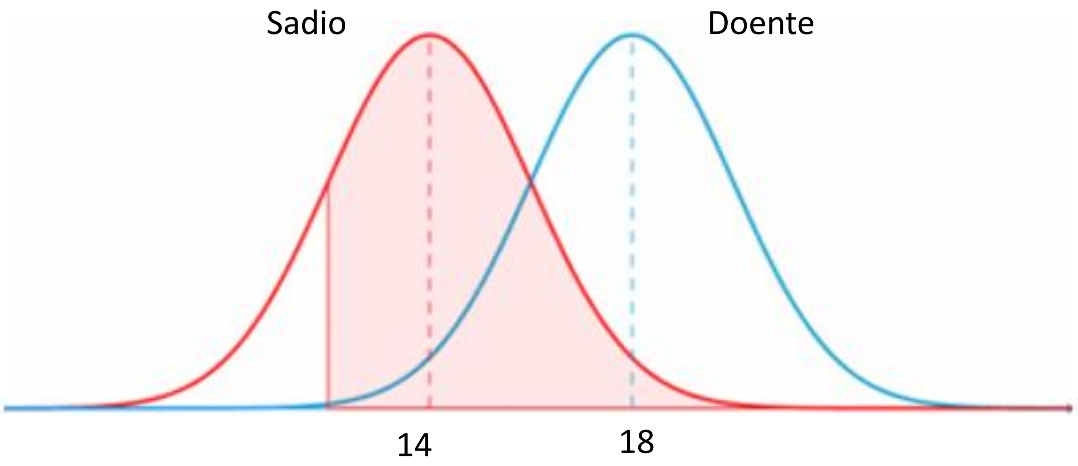
Exemplos

- P(Levar um guarda-chuva | Choveu)
- P(Sair mais cedo | Perdeu o ônibus)
- P(Chutar um pênalti no lado esquerdo | Goleiro pegou no lado direito)
- P(Estar doente | Exame deu positivo)
- P(Estar sadio | Exame deu negativo)

Introdução: Teste de Hipótese

- Suponha que, entre pessoas sadias, a concentração de certa substância no sangue se comporta segundo um modelo Normal com média **14 um/ml** e desvio padrão de **6 um/ml**. Pessoas sofrendo de uma doença específica têm a concentração média as substância alterada para **18 um/ml**. Admitindo que o modelo com desvio padrão continua representando de forma adequada a concentração da substância em pessoas com a doença, vejamos a ilustração.

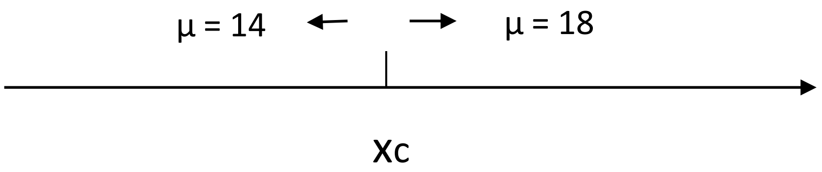
Introdução: Teste de Hipótese



Note que as curvas se cruzam fazendo com que pessoas sadias possam ter níveis tão alto de concentração quanto aqueles dito doentes.

Introdução: Teste de Hipótese

- Suponha que desejamos saber sobre a eficácia de um tratamento e para tanto coletamos uma amostra de tamanho 30.
- O objetivo é encontrar um valor crítico xc que nos permita decidir se acima dele o tratamento não foi eficaz ou abaixo dele o tratamento foi eficaz.



Introdução Teste de Hipótese

FIAP MBA+

- Sobre a eficácia do tratamento podemos formular as seguintes hipóteses

H0: O tratamento não é eficaz

H1: O tratamento é eficaz

$H0: \mu = 18$

$H1: \mu < 18$

Testes de hipóteses

Teste de hipóteses, teste estatístico ou teste de significância é um procedimento estatístico que permite tomar uma decisão (rejeitar ou não) a hipótese nula **H0** entre duas ou mais hipóteses (hipótese nula **H0**) ou (hipótese alternativa **H1)**, utilizando os dados observados de um determinado experimento.

H0: Algo que se queira refutar

H1: Algo que se queira evidenciar

Tipos de erro

Situação

Decisão

	H0 Verdadeira	H0 Falsa
Rejeitar H0	Erro tipo I	Acerto
Não rejeitar H0	Acerto	Erro tipo II

Tipos de erro

H0: Não estar grávida(o)

H1: Estar grávida(o)

Tipos de erro

Situação

		H0 Verdadeira	H0 Falsa
Decisão	Rejeitar H0	Erro tipo I	Acerto
	Não rejeitar H0	Acerto	Erro tipo II

α= P(erro tipo I) = P(rejeitar H0 | H0 Verdadeira)

β= P(erro tipo II) = P(não rejeitar H0 | H0 Falsa)

Tipos de erro: Exemplo eficácia do tratamento

- Sobre a eficácia do tratamento podemos formular as seguintes hipóteses

H0: O tratamento não é eficaz

H1: O tratamento é eficaz

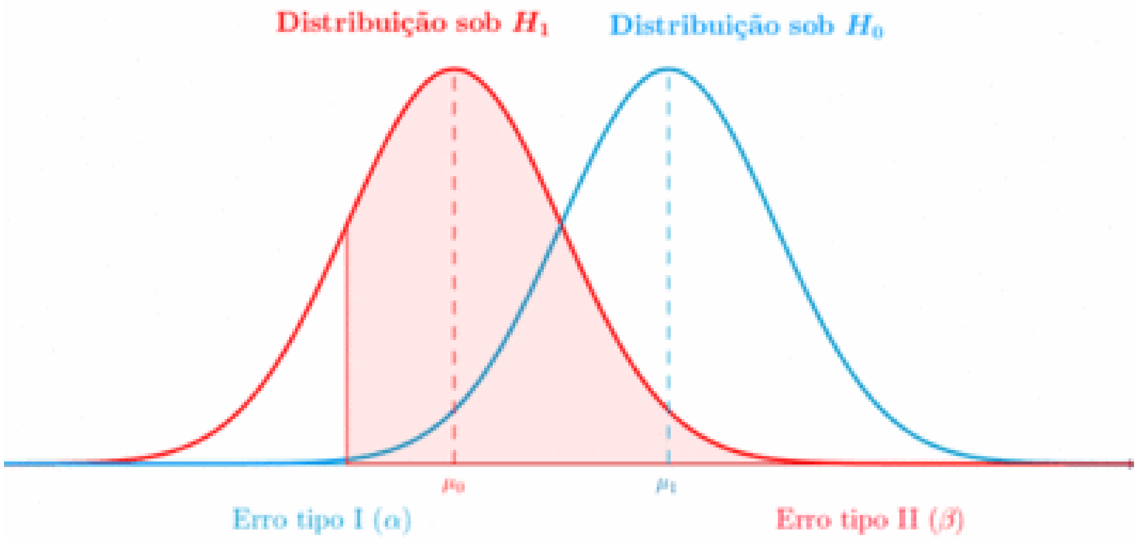
$H0: \mu = 18$

$H1: \mu < 18$

α= P(erro tipo I) = P(rejeitar H0 | H0 Verdadeira) =P(concluir que o tratamento é eficaz quando na verdade ele não é)

β= P(erro tipo II) = P(não rejeitar H0 | H0 Falsa) = P(concluir que o tratamento não é eficaz quando na verdade ele é)

Controle dos tipos de erro



Tipos de erro: Exemplo eficácia do tratamento

- Sobre a eficácia do tratamento podemos formular as seguintes hipóteses

$$\begin{cases} H0: \text{O tratamento não é eficaz} \\ H1: \text{O tratamento é eficaz} \end{cases} \qquad \begin{cases} H0: \mu = 18 \\ H1: \mu < 18 \end{cases}$$

α= P(erro tipo I) = P(rejeitar H0 | H0 Verdadeira) $= P(\bar{X} < x_c \mid \mu = 18) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{x_c - 18}{6/\sqrt{30}}\right) = P(Z < z_c)$

$Z \sim N(0,1)$

Tipos de erro: Exemplo eficácia do tratamento

$$z_c = \frac{x_c - 18}{6/\sqrt{30}}, \text{ então } x_c = 18 + z_c . 6/\sqrt{30}$$

Usando α = 5% = 0,05, então 0,05 = P(Z < zc), ou seja, $z_c = -1{,}64$.

Portando $x_c = 16{,}20$.

RC = { x < 16,20}

Rejeita H0 se x < 16,20

Tipos de erro: Exemplo eficácia do tratamento

$$z_c = \frac{x_c - 18}{6/\sqrt{30}}, \text{ então } x_c = 18 + z_c . {}^{6}/_{\sqrt{30}}$$

Usando α = 5% = 0,05, então 0,05 = P(Z < zc), ou seja, $z_c = -1{,}64$.

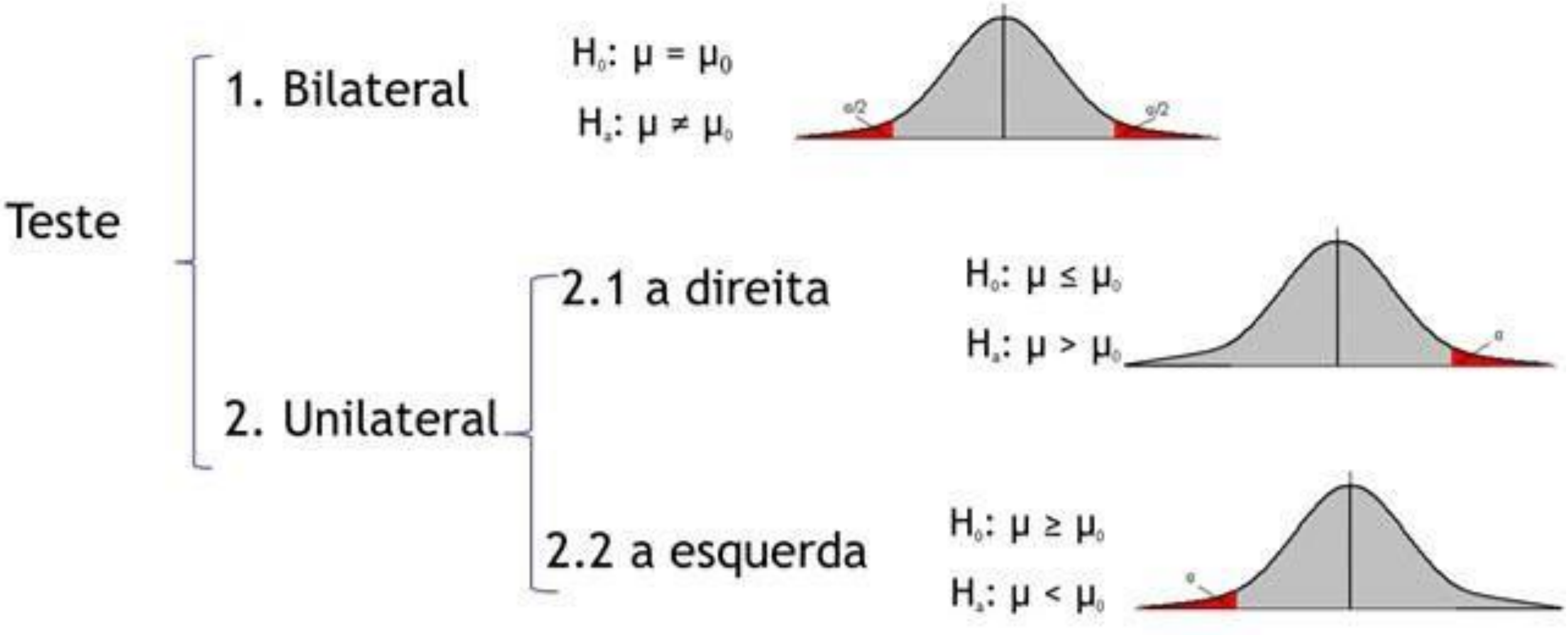
Portando $x_c = 16{,}20$.

RC = { x < 16,20}

Rejeita H0 se x < 16,20

Ou seja, o tratamento é eficaz.

Testes Unilaterais e bilaterais



P-valor: Nível descritivo

Probabilidade de se obter estimativas mais desfavoráveis ou extremas (à luz da hipótese alternativa) do que a que está sendo fornecida pela amostra.

Em outras palavras

Probabilidade do valor obtido da estimativa pela amostra ter sido ao acaso.

P-valor= P(X < média(observada) | H0 Verdadeira)

P-valor: Exemplo eficácia do tratamento

- Sobre a eficácia do tratamento podemos formular as seguintes hipóteses

H0: O tratamento não é eficaz

H1: O tratamento é eficaz

H0: $\mu = 18$

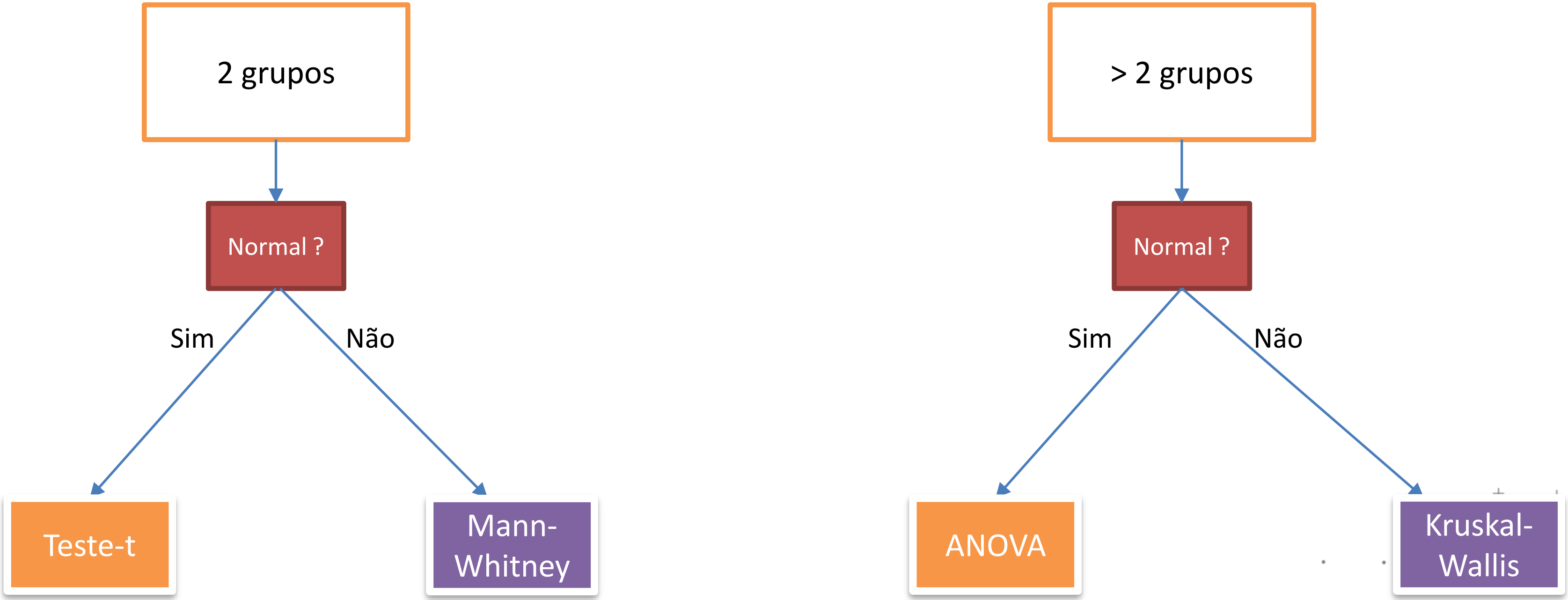
H1: $\mu < 18$

Supondo média amostral igual a 16 e $\alpha = 5\%$.

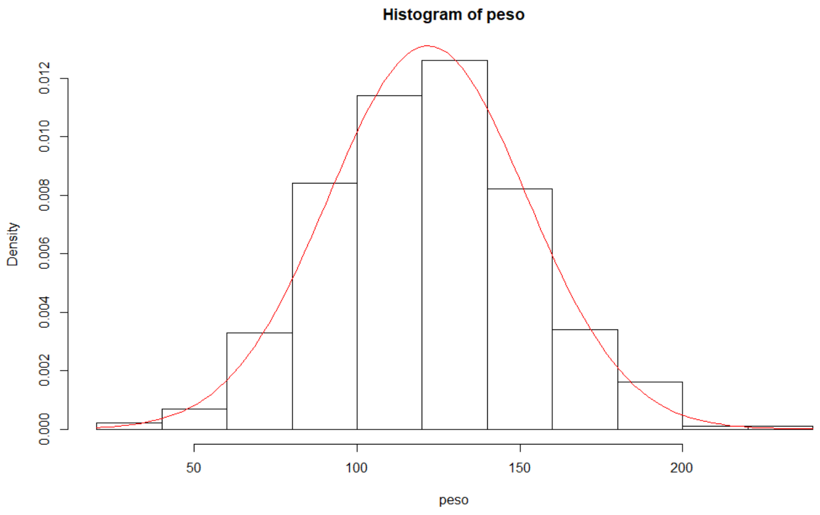
P-valor= $P(\bar{X} < 16 | \mu = 18) = P(Z < -1{,}826) = 0{,}033$

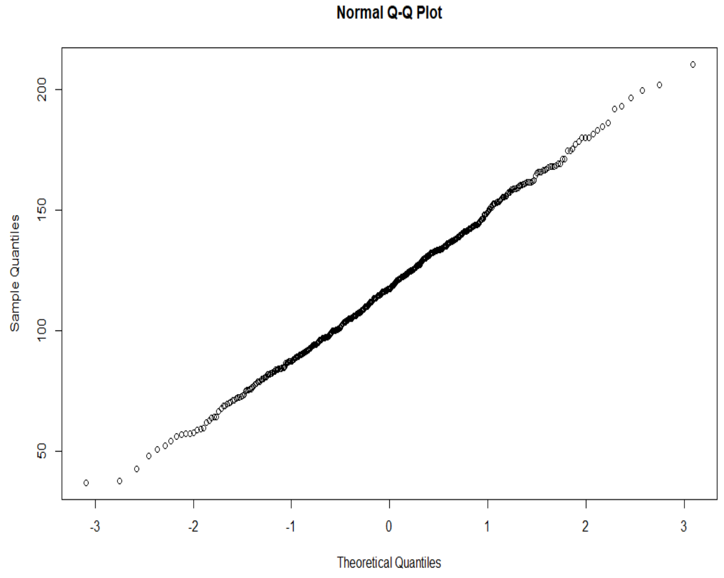
Rejeitamos H0, e concluímos que o tratamento é eficaz ao nível de 5% de significância.

Comparação de Grupos



Testes de Normalidade





Testes de Normalidade

H0: Os dados seguem distribuição normal.

H1: Os dados não seguem distribuição normal.

Testes

- Shapiro-Wilk
- Anderson-Darling
- Kolmogorov-Smirnov

$p < \alpha$: Rejeita a Hipótese Nula, ou seja, não é normal ao nível de significância α.

$p >= \alpha$: Não rejeita a Hipótese Nula, ou seja, é normal ao nível de significância α.

Comparação 2 grupos

H0: Os grupos são iguais

H1: Grupo são diferentes

H0: $m1 = m2$

H1: $m1 \neq m2$

p < α : Rejeita a Hipótese Nula, ou seja, os grupos são diferentes ao nível de significância α.

p >= α : Não Rejeita a Hipótese Nula, ou seja, os grupos não são diferentes ao nível de significância α.

Teste-t e Mann-Whitney

```
AgeM <- sim.dat$age[sim.dat$gender == 'Male’]
AgeF <- sim.dat$age[sim.dat$gender == 'Female']
```

```
t.test(AgeM, AgeF)
```

```
wilcox.test(AgeM, AgeF)
```

Comparação 3 ou mais grupos

H0: Os grupos são iguais

H1: Pelo menos um grupo é diferente

H0: $m1 = m2 = m3 = ... = mn$

H1: $mi \neq mj$; para algum i e j

$p < \alpha$: Rejeita a Hipótese Nula, ou seja, pelo menos 1 é diferente ao nível de significância α.

$p >= \alpha$: Não Rejeita a Hipótese Nula, ou seja, grupos são iguais ao nível de significância α.

Coeficiente de correlação linear

Definição

O coeficiente de correlação linear de Pearson é expresso na seguinte forma:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y},$$

em que

$\bar{x}$ e $\bar{y}$ denotam as médias amostrais

s_X e s_Y denotam os respectivos desvios padrão amostrais

Coeficiente de correlação linear

Propriedades

O coeficiente de correlação linear de Pearson apresenta a seguinte propriedade:

$$-1 \leq r \leq 1.$$

Casos particulares

$r = 1$: correlação linear positiva e perfeita

$r = -1$: correlação linear negativa e perfeita

$r = 0$: ausência de correlação linear

Correlação

H0: Não existe correlação

H1: Existe correlação

$H0: r = 0$

$H1: r \neq 0$

$p < \alpha$: Rejeita a Hipótese Nula, ou seja, há correlação ao nível de significância α.

$p >= \alpha$: Não Rejeita a Hipótese Nula, ou seja, não há correlação ao nível de significância α.

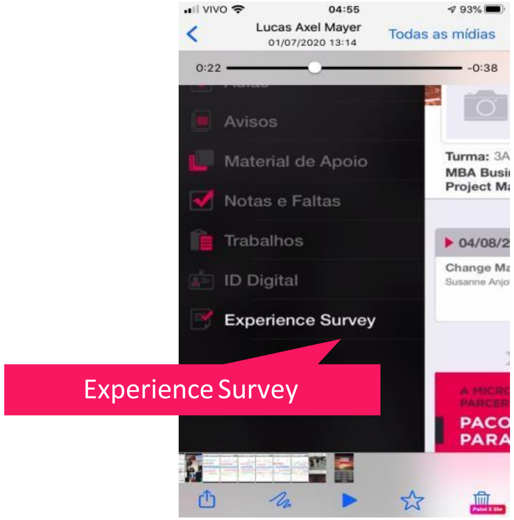
No R...

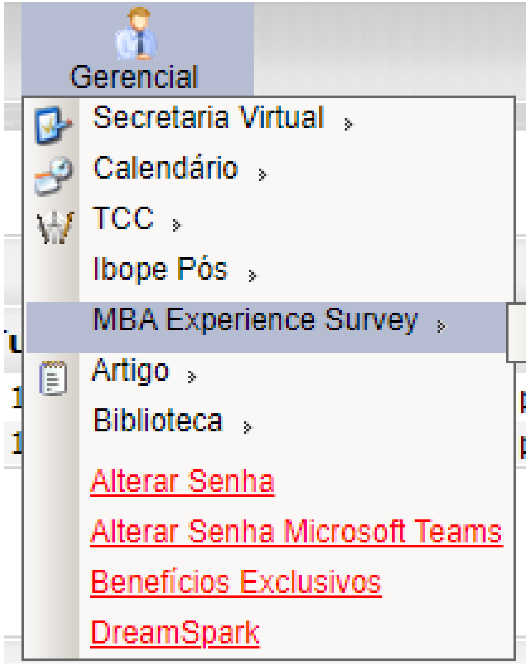
```
cor.test(sim.dat$store_exp, sim.dat$store_trans,
method = "pearson", alternative = "two.sided")
```

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





OBRIGADO

/lafphd

profleandro.ferreira@fiap.com.br

FIAP