FIAP

MBA

MBA EM DATA SCIENCE & AI

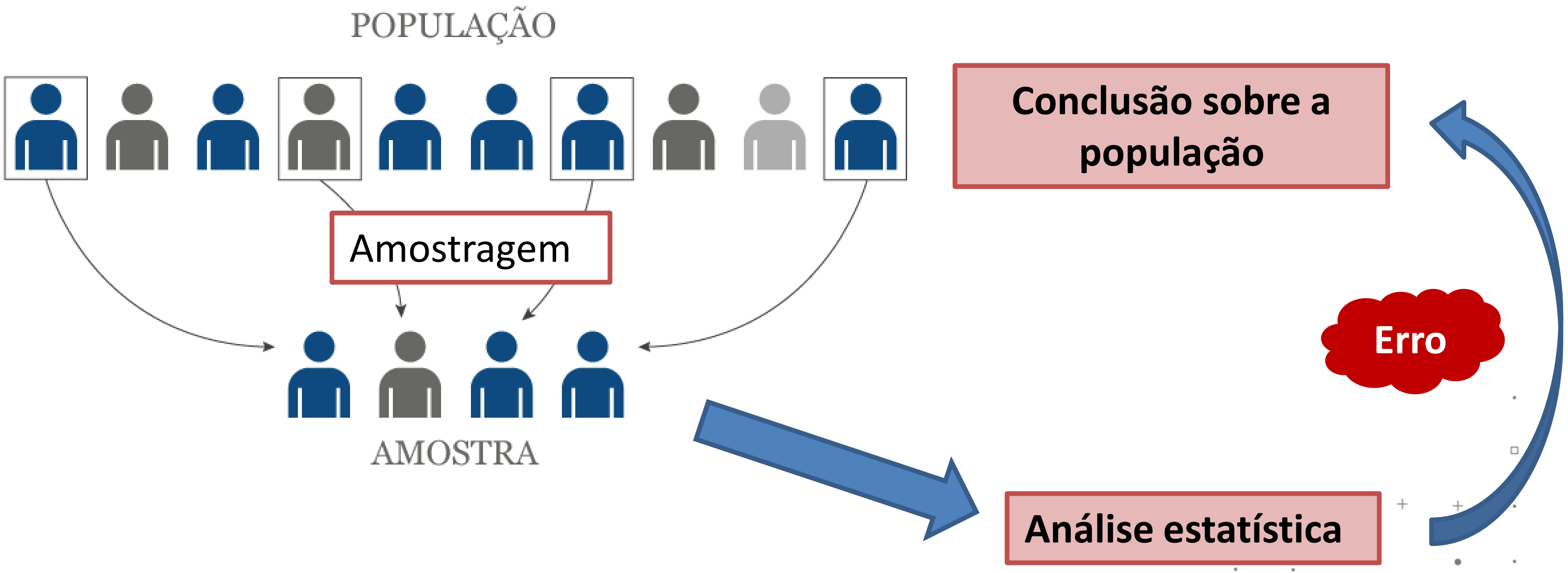
APPLIED STATISTICS

AULA 4

Amostragem

Testes de Hipótese

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.

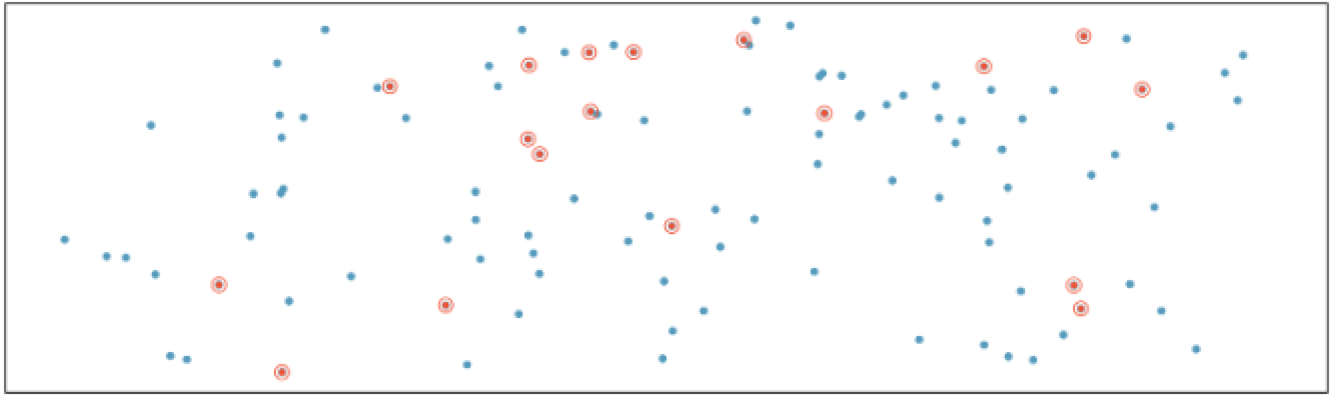
Tipos de amostragem

- 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

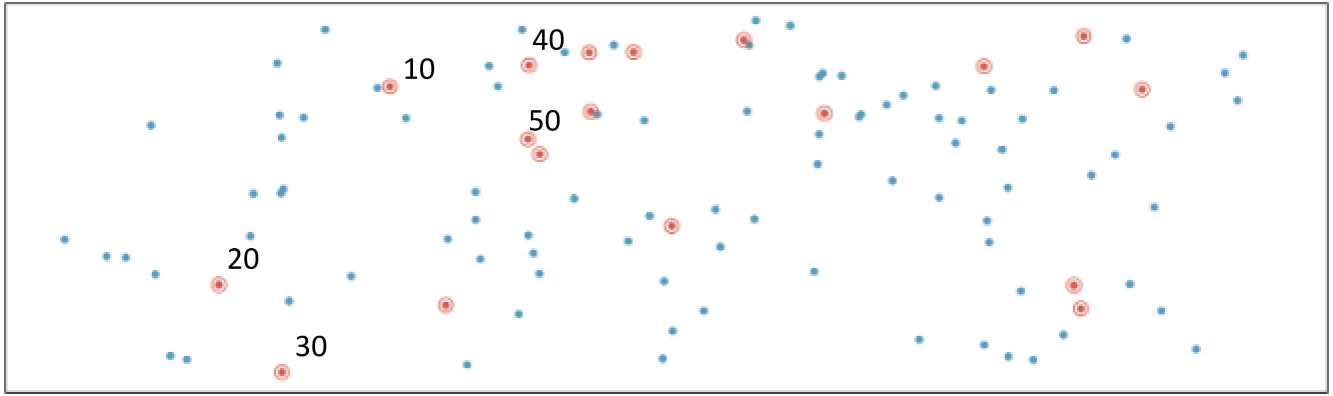


Sorteio de forma aleatória.

Aleatória simples

- Devemos utilizar essa abordagem quando **não** temos que garantir representatividade de nenhum grupo em específico.

Sistemática

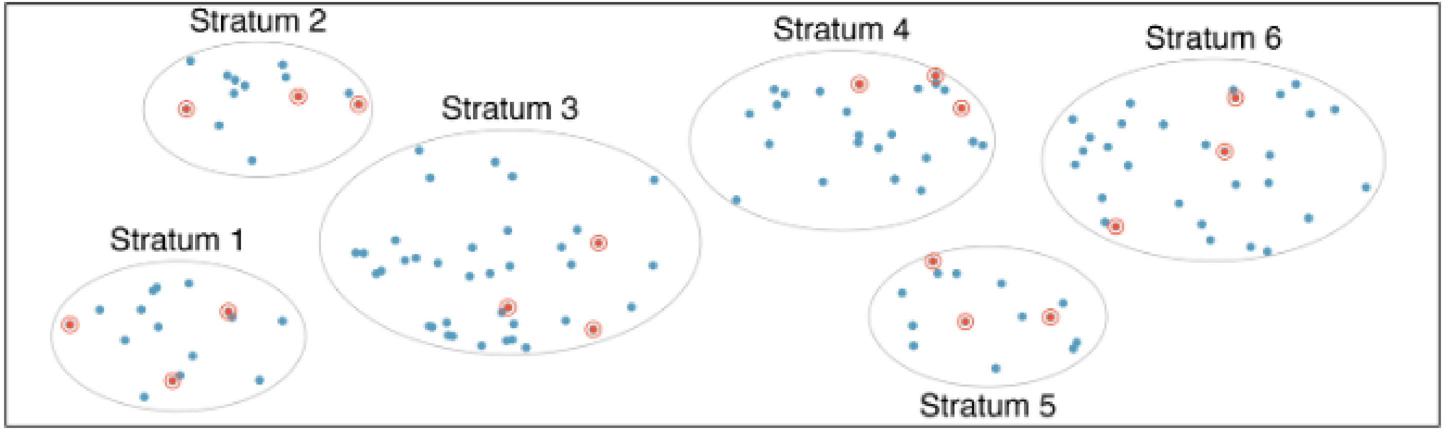


Sorteio baseado em uma estratégia. Ex: Selecionar a cada 10.

Sistemática

- Técnica bastante utilizada em controle de qualidade de processos industriais, onde não há uma especificação de qual elemento será coletado.

Estratificada

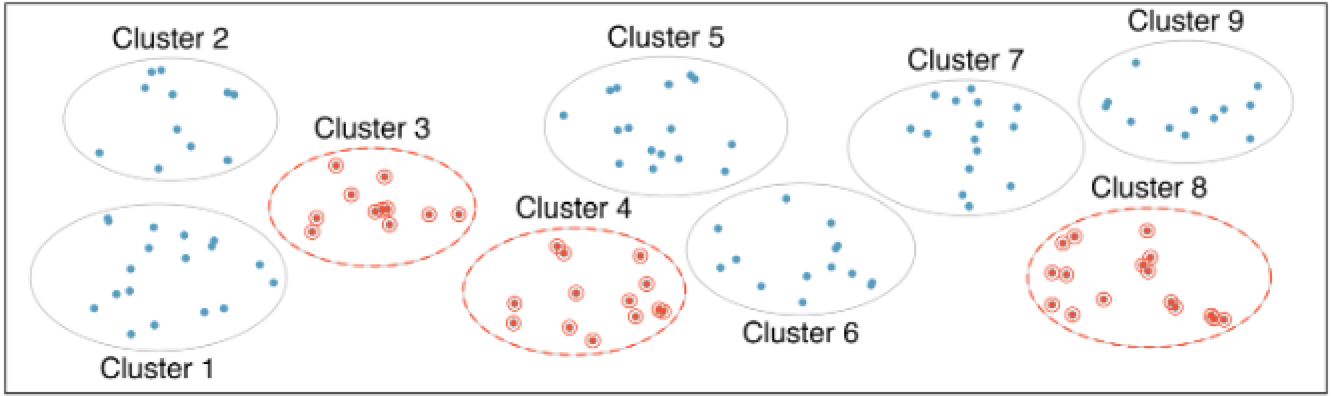


Sorteio de indivíduos dentro dos estratos

Estratificada

- Queremos garantir que cada estrato tenha uma quantidade de representantes pré-definida.
- **Ex:** Em problemas de Fraude onde a ocorrência é muito baixa. Gostaríamos de garantir uma proporção maior de ocorrência.

Conglomerados



Sorteio de clusters e não dos indivíduos.

Exercício (Claims.csv)

O arquivo **claims.csv** contém uma amostra aleatória de 996 apólices de seguros de veículos referente ao período 2004-2005. As variáveis do arquivo estão na seguinte ordem : (i) **valorv** (valor do veículo em 10000 dolares australianos), (ii) **expos** (exposição do veículo), (iii) **nsinistros** (número de sinistros no período), (iv) **csinistros** (custo total dos sinistros em dolares australianos), (v) **tipov** (tipo do veículo em 11 categorias), (vi) **idadev** (idade do veículo em 4 categorias), (vii) **sexoc** (sexo do condutor principal), (viii) **areac** (área de residência do condutor principal) e (ix) **idadec** (idade do condutor principal em 6 categorias).

Exercício (Claims.csv)

Utilizar a base 'claims.csv' e faça amostragens:

- Aleatória simples (200)
- Estratificada (100 pelo segmento sexo)

Compare a variável cmsinistros = csinistros/nsinistros por tipo de amostragem usando boxplot.

```
df = pd.read_csv('claims.csv',  delimiter=';', decimal=',')
```

Teste de Hipótese

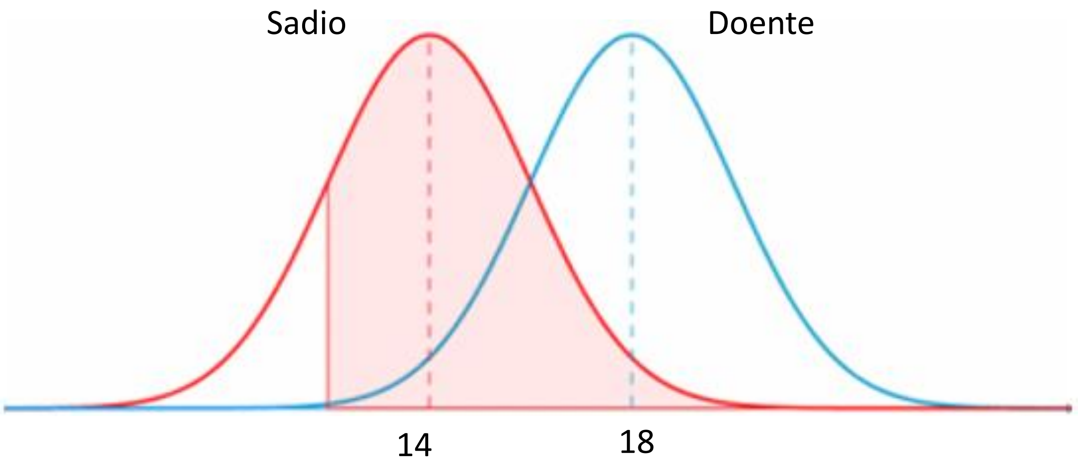
S -> P

Se fizer Sol, então vou à Praia

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.

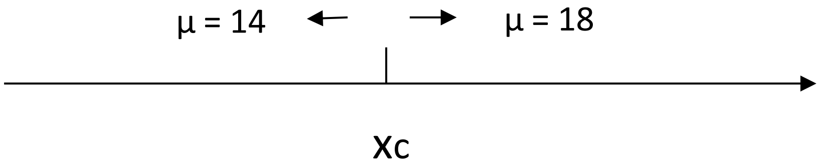
Introdução: Teste de Hipótese



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

- Sobre a eficácia do tratamento podemos formular as seguintes hipóteses

H0: O tratamento não é eficaz

H1: O tratamento é eficaz

$H_0: \mu = 18$

$H_1: \mu < 18$

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

FIAP MBA+

Situação

Decisão		H0 Verdadeira	H0 Falsa
	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

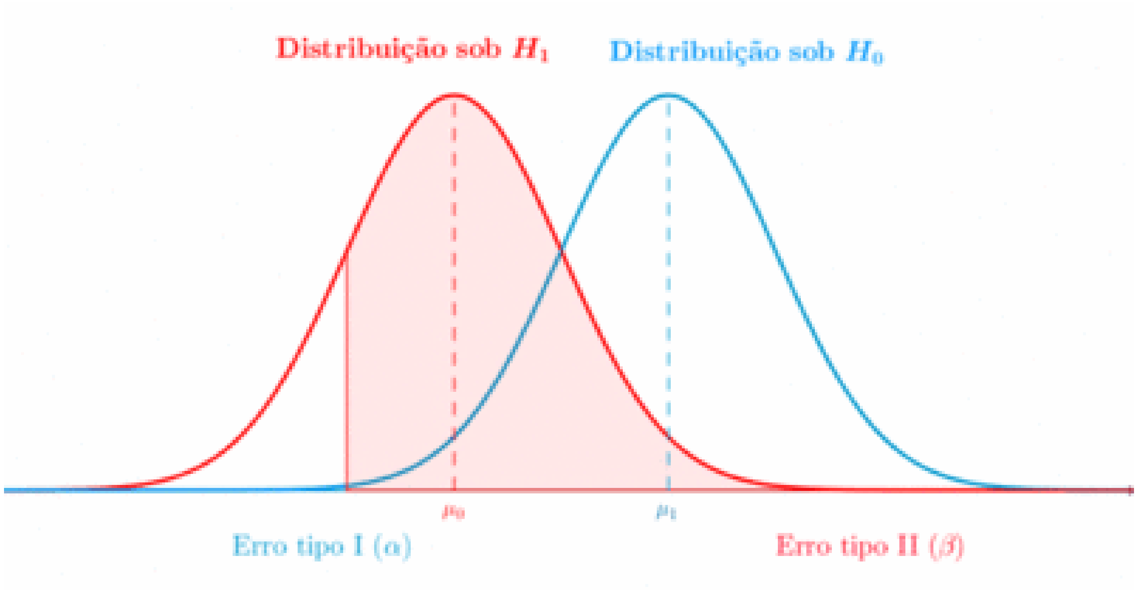
- Sobre a eficácia do tratamento podemos formular as seguintes hipóteses

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

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

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) $= P(\bar{X} < x_c \mid \mu = 18) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{x_c - 18}{6/\sqrt{30}}\right) = P(Z < z_c)$

$Z \sim N(0,1)$

Tipos de erro: Exemplo eficácia do tratamento

$z_c = \frac{x_c - 18}{6/\sqrt{30}}$, 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, $z_c = -1{,}64$.

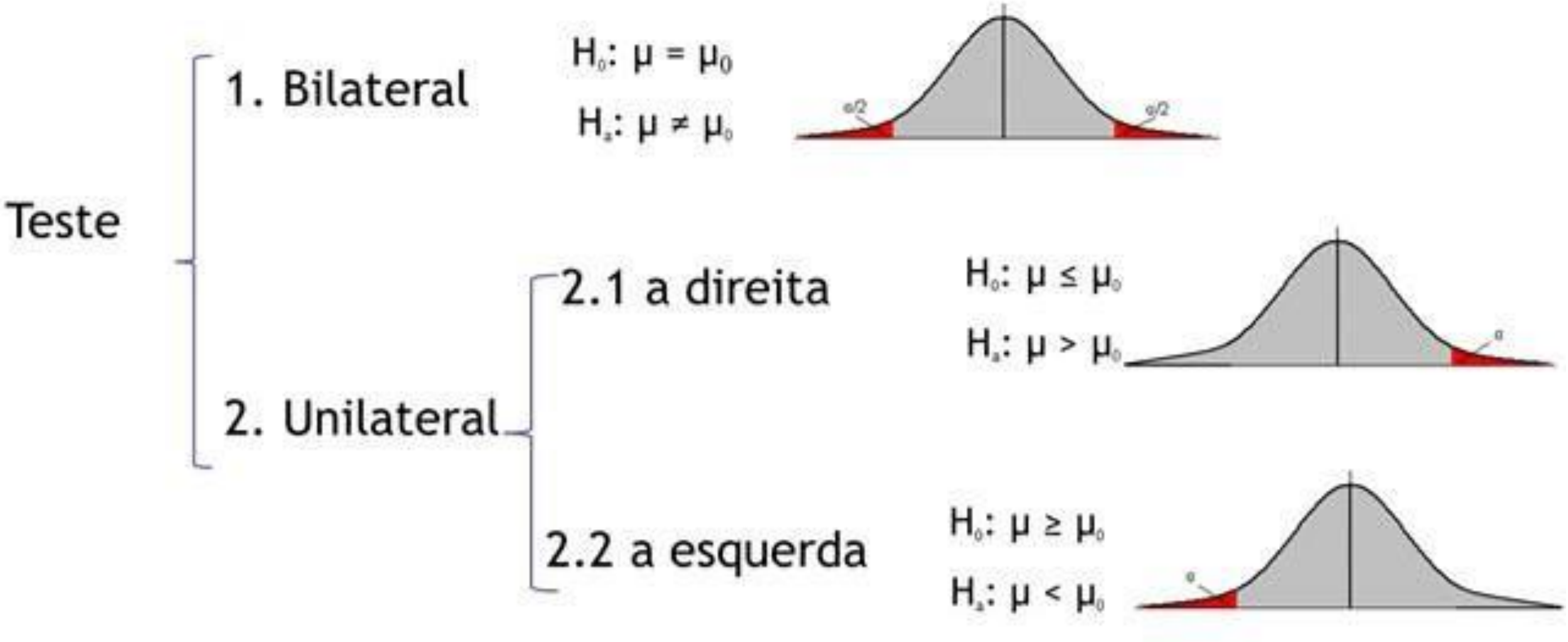
Portando $x_c = 16{,}20$.

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: $\mu = 18$

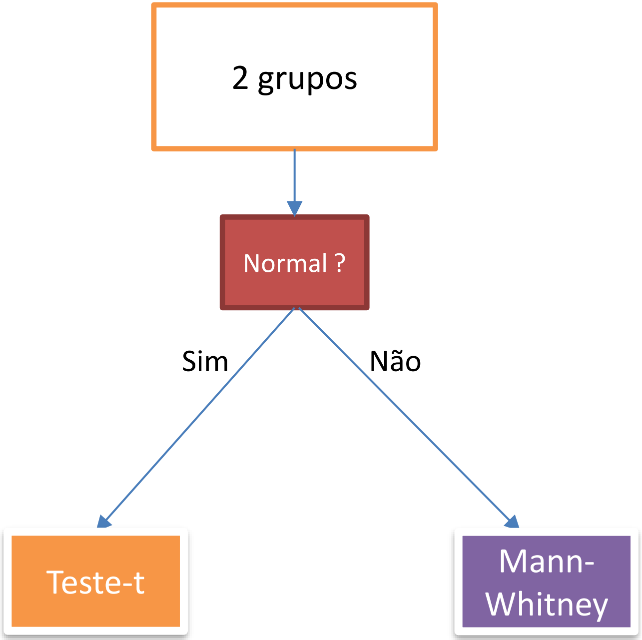
H1: $\mu < 18$

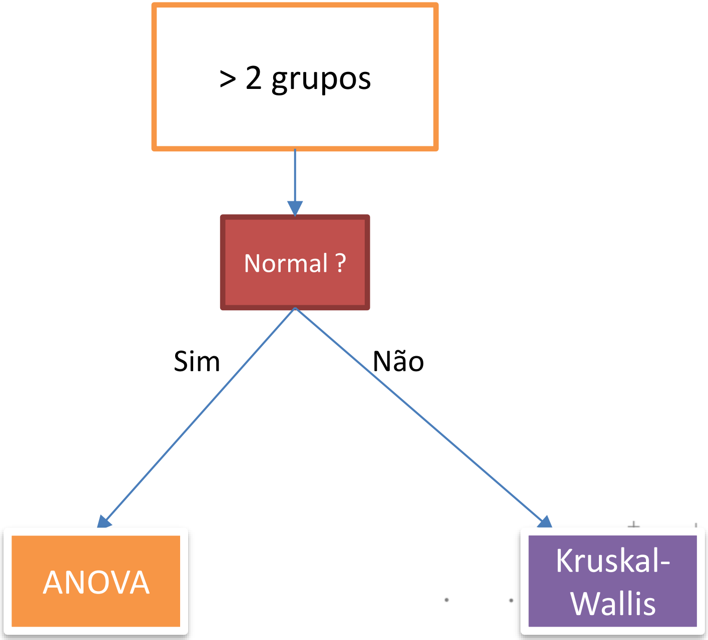
Supondo média amostral igual a 16 e $\alpha = 5\%$.

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

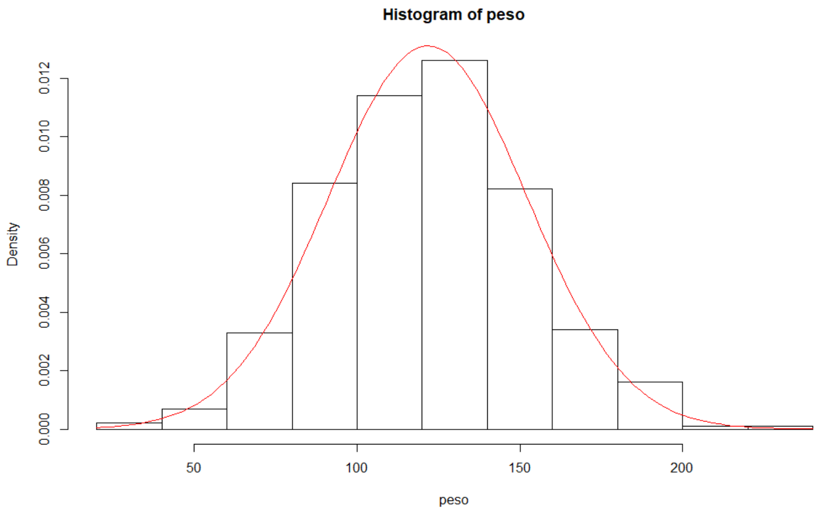
Rejeitamos H0, e concluímos que o tratamento é eficaz ao nível de 5% de significância.

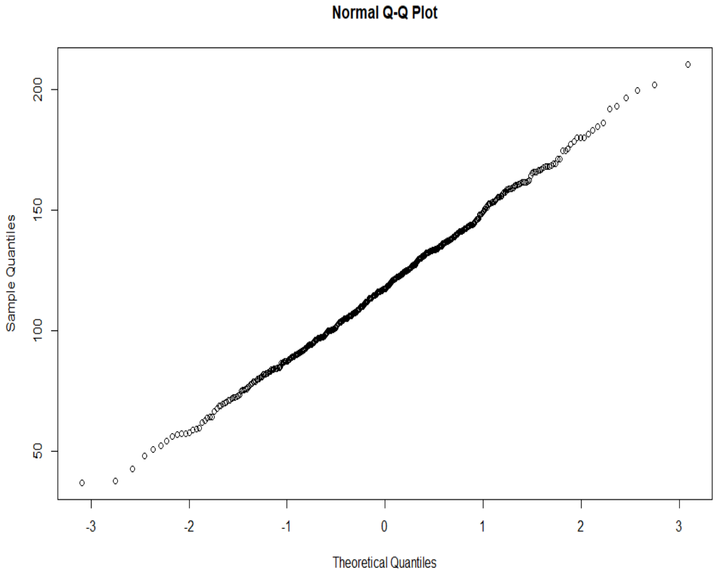
Comparação de Grupos





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.

Comparação 3 ou mais grupos

H0: Os grupos são iguais

H1: Pelo menos um grupo é diferente

H0: $m1 = m2 = m3 = \dots = mn$

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: $r \neq 0$

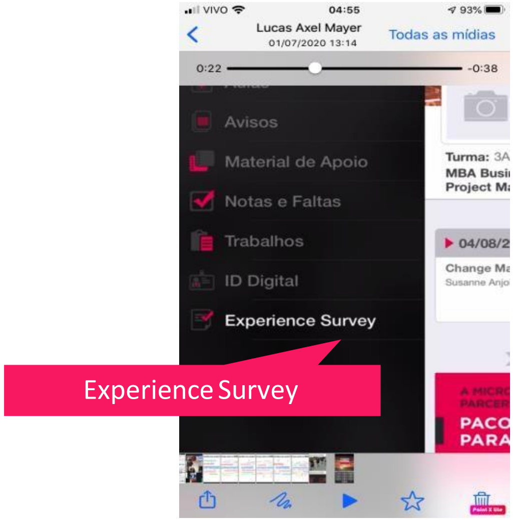
$p < \alpha$: Rejeita a Hipótese Nula, ou seja, há correlação ao nível de significância α.

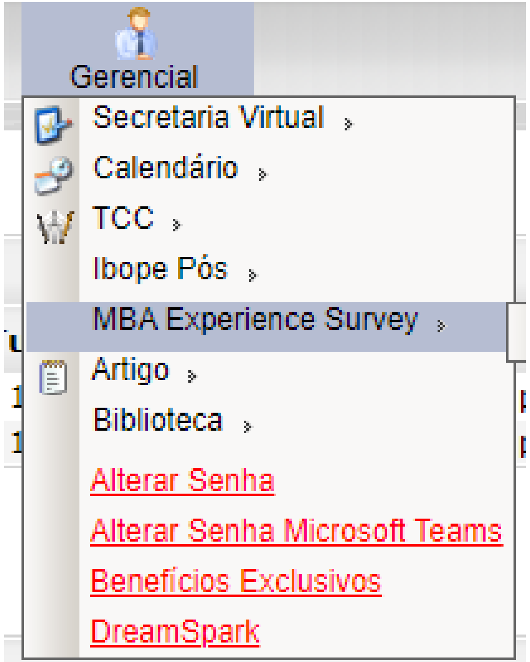
$p >= \alpha$: Não Rejeita a Hipótese Nula, ou seja, não há correlação ao nível de significância α.

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 MBA+

FIAP