FIAP

Statistics for Machine Learning

Aula 22: Teste F

Prof. Jones Egydio

profjones.egydio@fiap.com.br

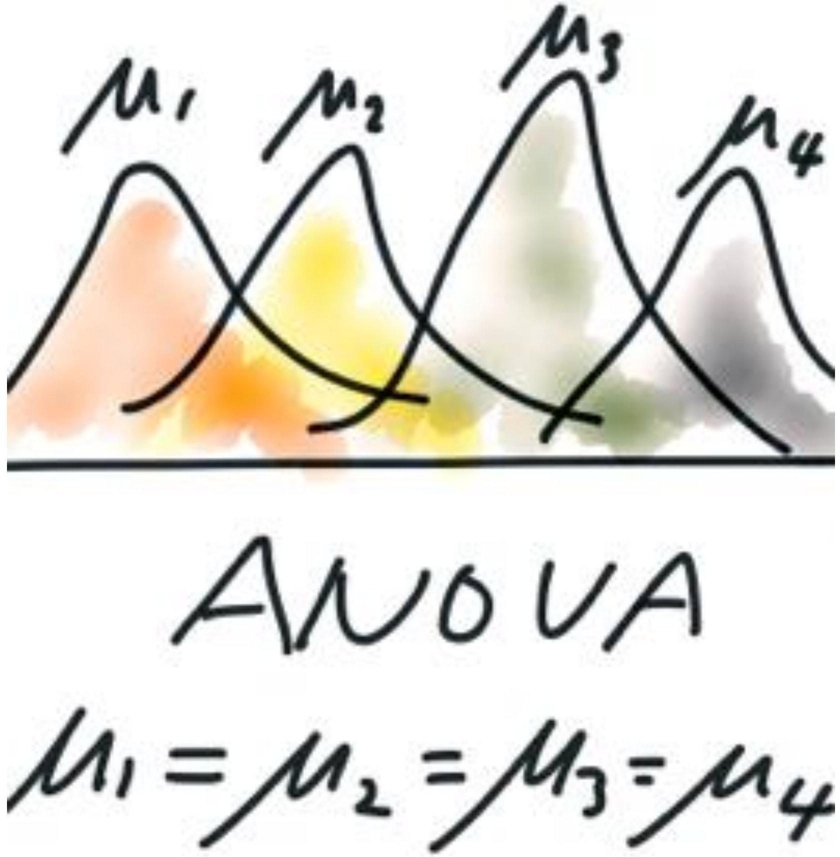
2024

Objetivos

- Introduzir os conceitos de testes F;
- Formas de representação;
- Exemplos e exercícios;
- Conclusão;
- Perguntas.

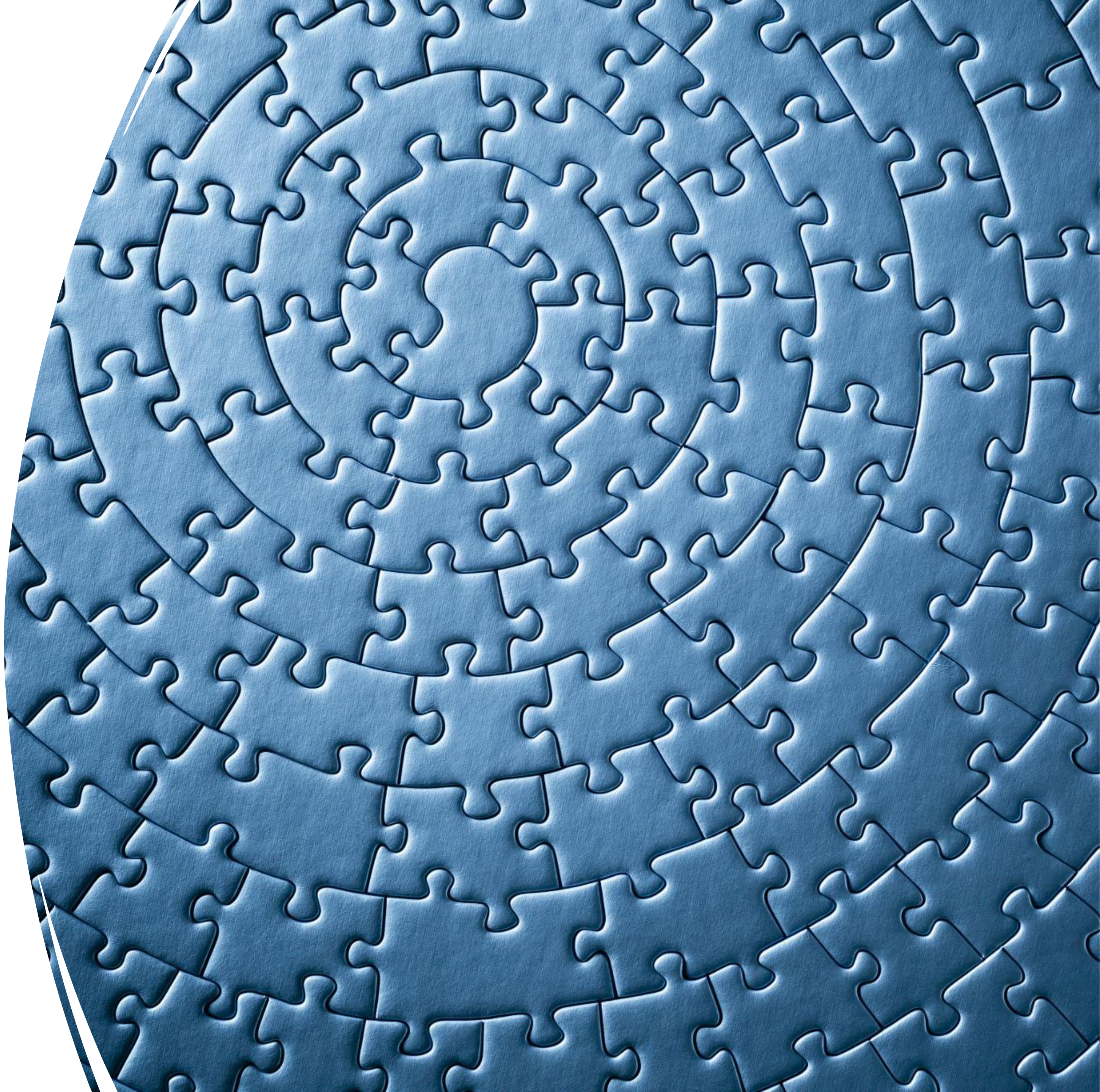
Conceitos iniciais

- Teste F ANOVA



Conceitos

- Revisão Procedimento geral
- Teste F e problemas
- Teste F aplicado



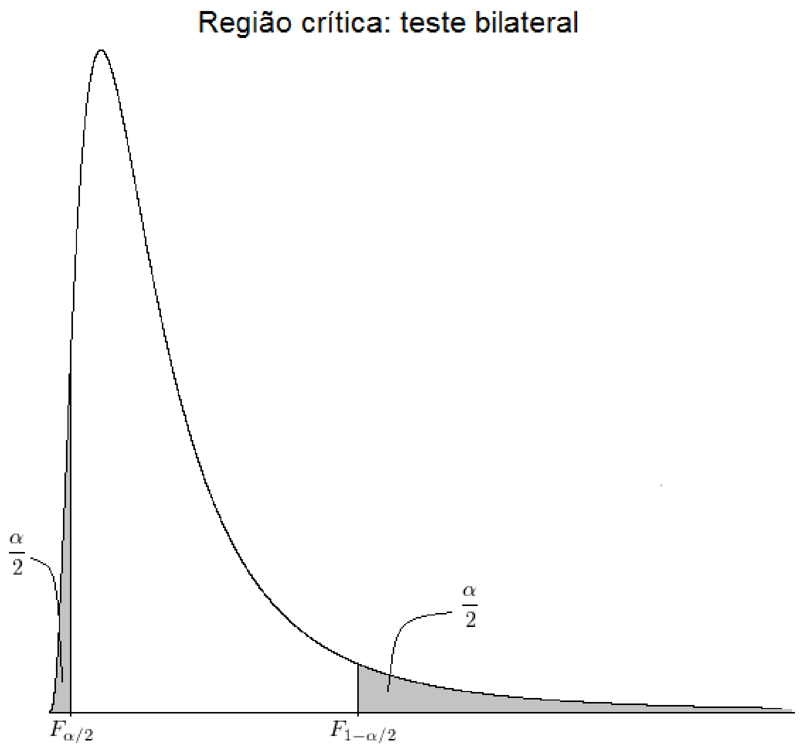
Procedimento Geral

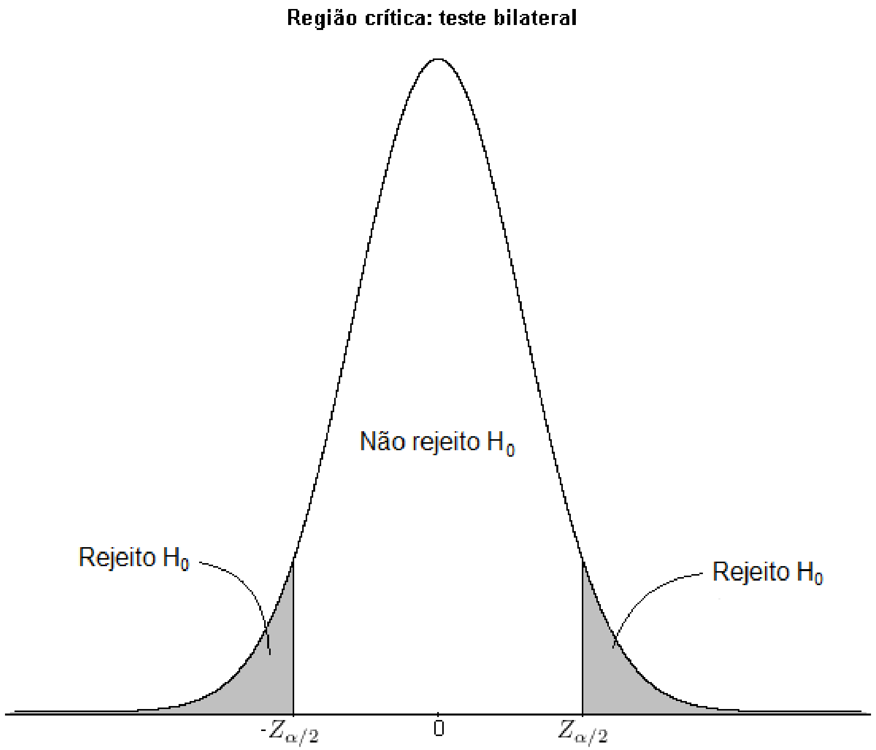
- Etapas:

- Fixar a hipótese nula (H_0) e a hipótese alternativa (H_1).
- Definir qual estimador será utilizado (média, variância).
- Fixar o nível de confiança (90%, 95%, 99%).
- Utilizar os dados da amostra obtida para calcular o valor da estatística teste (t, F, Qui-Quadrado).
- Verificar se o valor da estatística teste esta contida ou não na região de rejeição. Caso esteja, rejeita H_0, contrário não rejeita H_0.

Procedimento geral

Exemplos de regiões de rejeição.





Quais situações aplicamos?

- Geralmente aplicamos quando temos algumas dúvidas sobre um objetivo e necessitamos realizar uma pesquisa.

- Exemplos:

- Comparar as variabilidade de renda de duas cidades.
- Comparar o risco de ações do mercado financeiro
- Comparar as médias de vários perfis de clientes.

Teste F

- Será apresentado um dos primeiros testes de hipóteses que podem ser aplicados.
- O teste F pode ser aplicado quando:
- **Deseja-se comparar as variâncias de 2 populações**
- **Deseja-se comparar as médias de mais de 2 grupos.**

Teste *F*

- Teste F para comparar duas variâncias:

- Problema exemplo:

- Para verificar o grau de satisfação dos funcionários de uma área da empresa. Estimou-se o desvio-padrão de 2 filiais presentes em duas cidades semelhantes. Foram obtidos salários de 10 funcionários da cidade A e 15 da cidade B, sendo o desvio-padrão de A R$ 1.000,00 e o desvio-padrão de B R$ 1.600,00.

Teste *F* - resolução

- Neste teste F precisamos dos tamanhos das amostras, dos desvios-padrões e do nível de confiança.

- Vemos o seguinte:

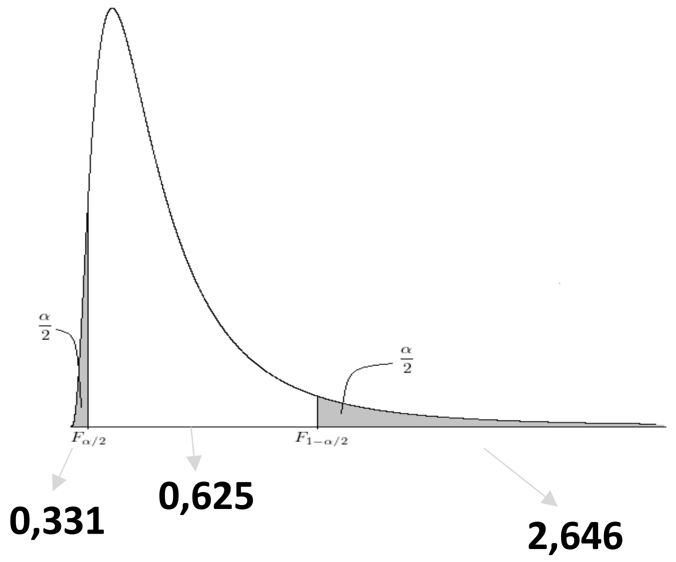
- Desvio-padrão de A: **1000**
- Desvio-padrão de B: **1600**
- Amostra de A: **10**
- Amostra de B: **15**
- Nível de confiança: **95%**

Teste F - resolução

- Fixar a hipótese nula (H_0) e a hipótese alternativa (H_1)

- $\boldsymbol{H_0: \sigma_A^2 = \sigma_B^2}$ **ou** $\boldsymbol{H_0: \sigma_A^2 / \sigma_B^2 = 1}$
- $\boldsymbol{H_a: \sigma_A^2 \neq \sigma_B^2}$ **ou** $\boldsymbol{H_a: \sigma_A^2 / \sigma_B^2 \neq 1}$

- Definir o estimador: **dividir as variâncias**
- Fixar o nível de confiança: **95%**
- Dividir as variâncias.
- $F = \frac{1000}{1600} = 0{,}625$
- $F_{95\%,9,14} \cong 2{,}646$, $t_{5\%,9,14} \cong 0{,}331$



O valor de **0,625 NÃO** está dentro da região de rejeição

- Rejeitar ou não rejeitar H_0?
- O valor de **0,625 NÃO** está dentro da região de rejeição. Deste modo, **não rejeitamos a hipótese nula**, **a variância entre os salários são estatisticamente iguais**.

Teste *F*

- Teste *F* para comparar médias de mais de 2 grupos:

- Neste caso, temos o problema de testar as diferenças de médias quando existem mais de 2 grupos a serem comparados.

- As hipóteses são construídas conforme o seguinte:

- $\boldsymbol{H_0: \mu_1 = \mu_2 = \cdots = \mu_k}$
- $\boldsymbol{H_a: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_k}$

- Para resolver este problema precisaremos falar um pouco de teoria.

Teste *F* – *ANOVA*

- Para testar este tipo de hipótese precisamos aplicar uma técnica denominada Análise de Variância (**ANOVA – *Analysis of Variance***).
- Como temos vários grupos precisaremos estimar a variabilidade **DENTRO** dos grupos e **ENTRE** os grupos. Aplicamos o procedimento:

- **SQT = SQF + SQE**

- **Soma dos Quadrados Totais** = **Soma dos Quadrados dos Fatores** + **Soma dos Quadrados dos Erros**.

Teste *F – ANOVA* (Fórmulas)

- Soma dos Quadrados Totais:
- $SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}\right)^2$

- Soma dos Quadrados dos fatores:
- $SQF = \sum_{i=1}^{k} n_i \, (\bar{y}_i - \bar{y})^2$

- Soma dos Quadrados dos resíduos (erros):
- $SQE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$

Teste *F – ANOVA* (Tabela)

- A análise é resumida nesta tabela:

Fontes de variação	Soma dos quadrados	Graus de liberdade	Médias	F
Var. entre os grupos	SQF	k-1	QMF=SQF/k-1	QMF/QME
Var. dentro dos grupos	SQE	n-k	QME=SQE/n-k	
Total	SQT	n-1		

Teste *F*

- Teste F para comparar média de mais de 2 grupos:
- Problema exemplo:
- Uma pesquisa busca comparar a perda de peso em 3 tipos de dieta. Vamos analisar a tabela abaixo e testar a hipótese de igualdade de médias.

Métricas	Baixa Caloria	Baixa Gordura	Baixo Carboidrato	Grupo de controle
n	5	5	5	5
Média	6,6	3	3,4	1,2

Teste *F* - resolução

- Fixar a hipótese nula (H_0) e a hipótese alternativa (H_1)

- $\boldsymbol{H_0: \mu_{caloria} = \mu_{gordura} = \mu_{carb} = \mu_{controle}}$
- $\boldsymbol{H_a: \mu_{caloria} \neq \mu_{gordura} \neq \mu_{carb} \neq \mu_{controle}}$

- Definir o estimador: **ANOVA**
- Fixar o nível de confiança: **95%**

- Vamos criar a tabela e continuar.

Teste *F* – *ANOVA* (Tabela)

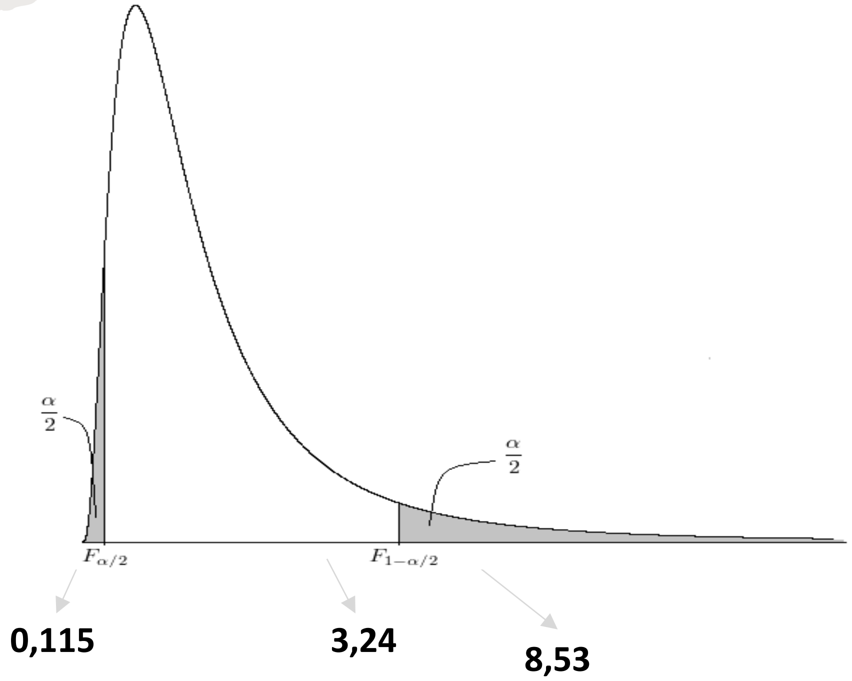
- Neste exemplo vamos pular as etapas de calcular SQF, SQE e SQT. Iremos obter estes valores no *python* mais adiante.

- Vamos guardar o valor de **8,53** e verificar a hipótese no próximo slide.

Fontes de variação	Soma dos quadrados	Graus de liberdade	Médias	F
Var. entre os grupos	75,8	4-1	75,8/3=25,3	25,3/3=8,53
Var. dentro dos grupos	47,4	20-4	47,4/16=3	
Total	123,2	20-1		

Teste *F* - resolução

- Valor de F: **8,53**
- F crítico: **F(5%, 3, 16) = 0,115**
- F crítico: **F(95%, 3, 16) = 3,24**



O valor de **8,43** está dentro da região de rejeição

- Rejeitar ou não rejeitar H_0?
- O valor de **8,53** está dentro da região de rejeição. Deste modo, **rejeitamos a hipótese nula**, **as médias de perda de peso entre os diferentes tratamentos são diferentes**.

Problema

Existe diferença na satisfação dos clientes quanto aos perfis PJ, PF e digital?

- O *Net Promoter Score (NPS)* é um índice utilizado por empresas para avaliar a satisfação dos clientes.
- Os consumidores avaliam a satisfação com o serviço prestado pela empresa atribuindo um nota de zero a dez.
- Esta avaliação fornece um diagnóstico simplificado pelos clientes em relação a empresa.

Teste *F*

- Alguns dados

Métricas	PJ	PF	Digital
n	50	150	100
Média	8	7,5	7

Perguntas

- Qual dos 2 testes F você utilizaria para comparar a média dos NPS entre os perfis de clientes?

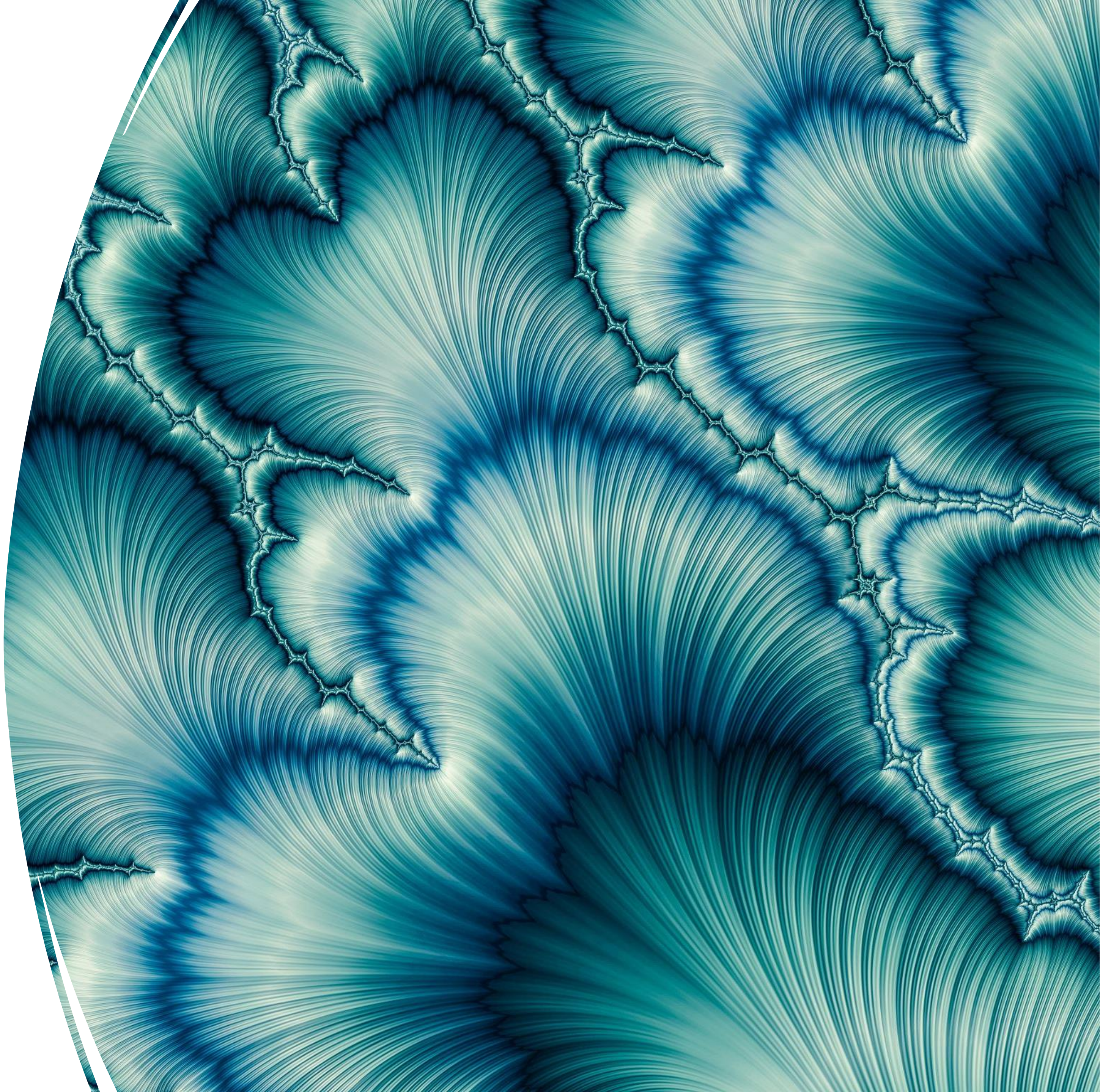
- Como seria a construção do teste de hipótese para este problema?

- Teste esta hipótese

Resolução

- Qual dos 2 testes F você utilizaria para comparar a média dos NPS entre os perfis de clientes?

- **Neste caso iremos utilizar o teste ANOVA.**



Resolução

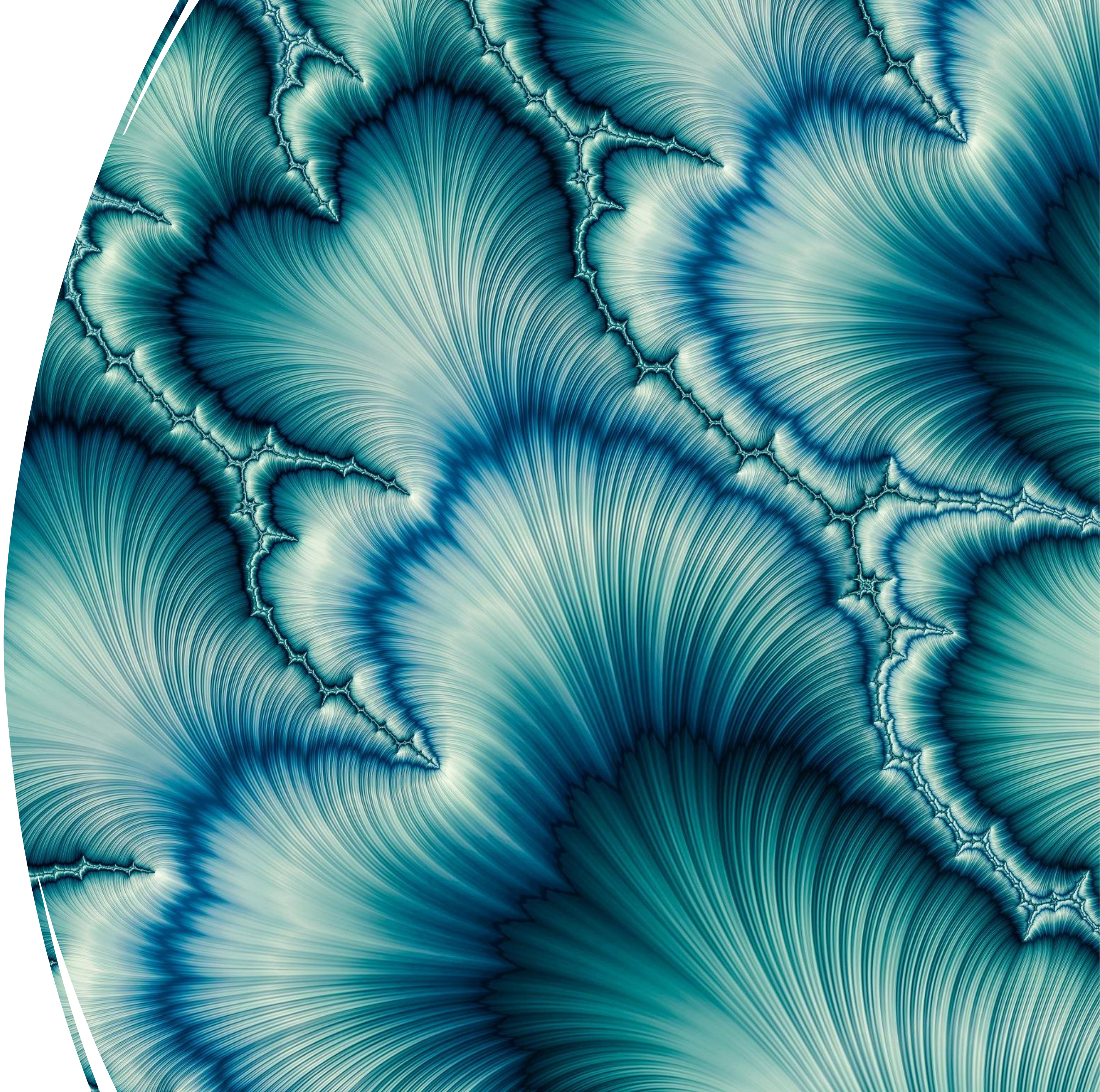
- Como seria a construção do teste de hipótese para este problema?

- Fixar a hipótese nula (H_0) e a hipótese alternativa (H_1)

- $\boldsymbol{H_0: \mu_{PJ} = \mu_{PF} = \mu_{Digital}}$
- $\boldsymbol{H_a: \mu_{PJ} \neq \mu_{PF} \neq \mu_{Digital}}$

- Definir o estimador: **ANOVA**
- Fixar o nível de confiança: **95%**

- Vamos criar a tabela e continuar.



Teste *F* – *ANOVA* (Tabela)

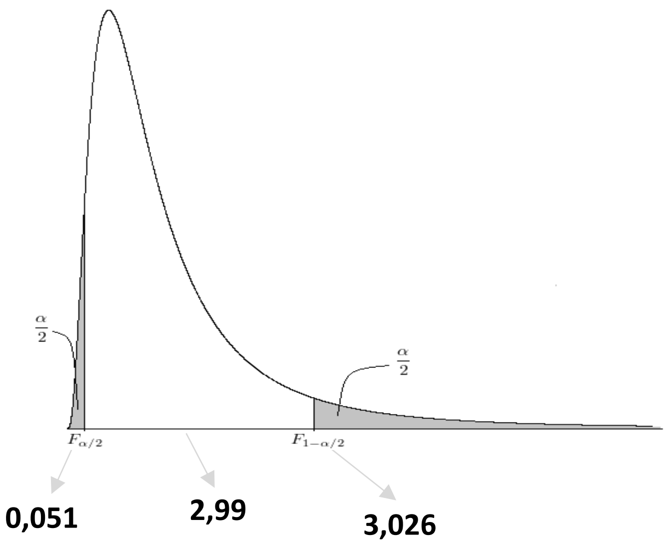
- Este é um exemplo fictício:

FONTES DE VARIAÇÃO	SOMA DOS QUADRADOS	GRAUS DE LIBERDADE	MÉDIAS	F
Var. entre os grupos	5,7	3-1	5,7/2=2,85	2,85/0,95=2,99
Var. dentro dos grupos	283,4	300-3	283,4/297=0,95	
Total	390	300-1		

- Vamos guardar o valor de **2,99** e verificar a hipótese no próximo slide.

Teste *F* - resolução

- Valor de F: **2,99**
- F crítico: **F(5%, 2, 297) = 0,051**
- F crítico: **F(95%, 2, 297) = 3,026**

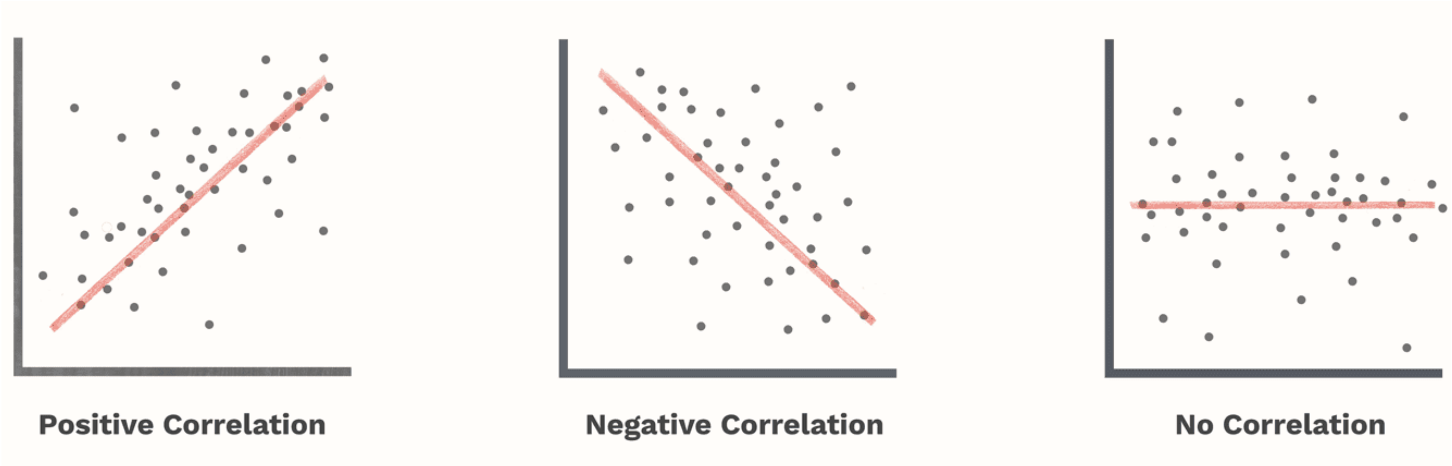


O valor de **2,99 NÃO** está dentro da região de rejeição

- Rejeitar ou não rejeitar H_0?
- O valor de **2,99 NÃO** está dentro da região de rejeição. Deste modo, **não rejeitamos a hipótese nula**, **as médias estão idênticas entre os perfis dos clientes quanto a satisfação da empresa**.

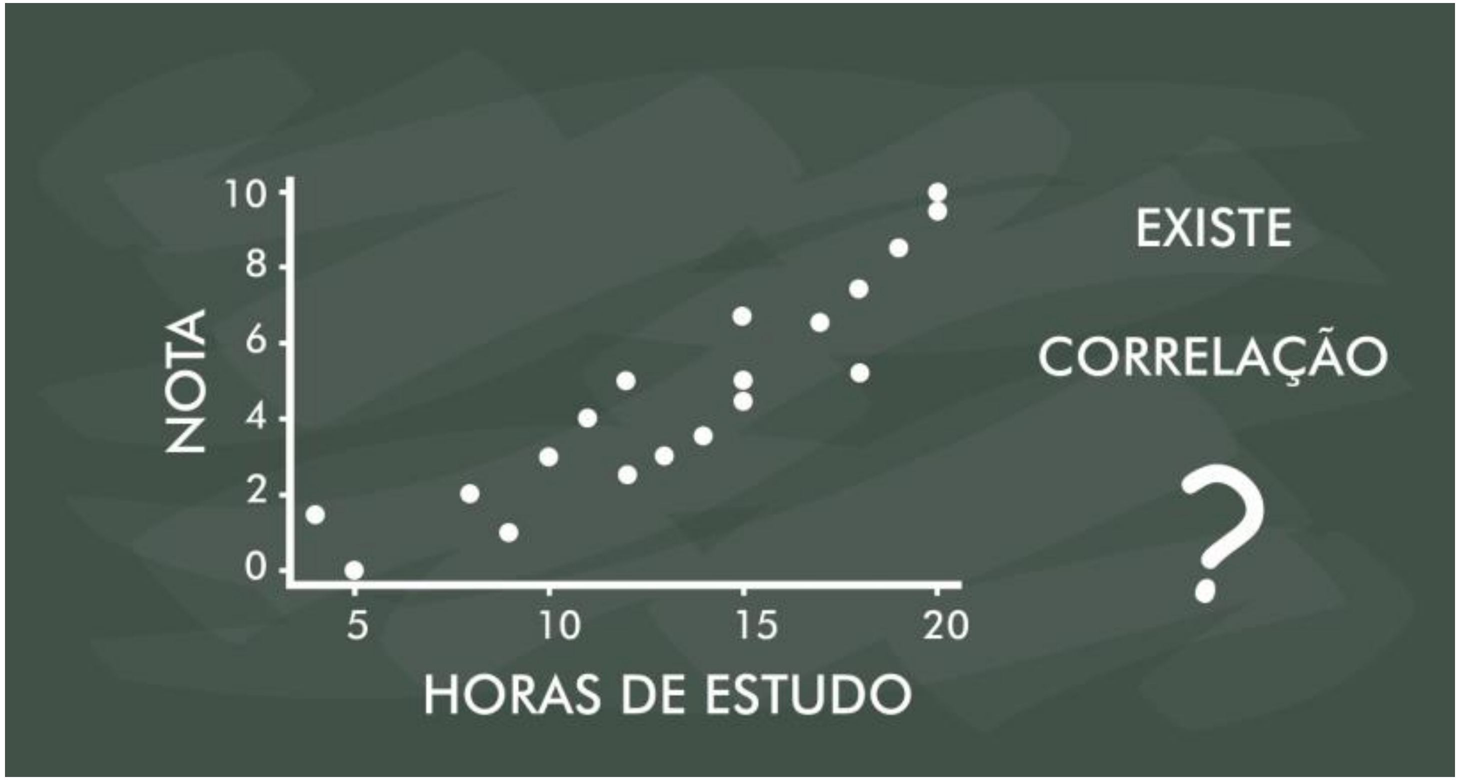
Conceitos iniciais

- Teste Correlação



Conceito

- **Problema:** Dado que temos duas variáveis X e Y e buscamos conhecer o quanto estão relacionadas, como poderíamos verificar esta relação?

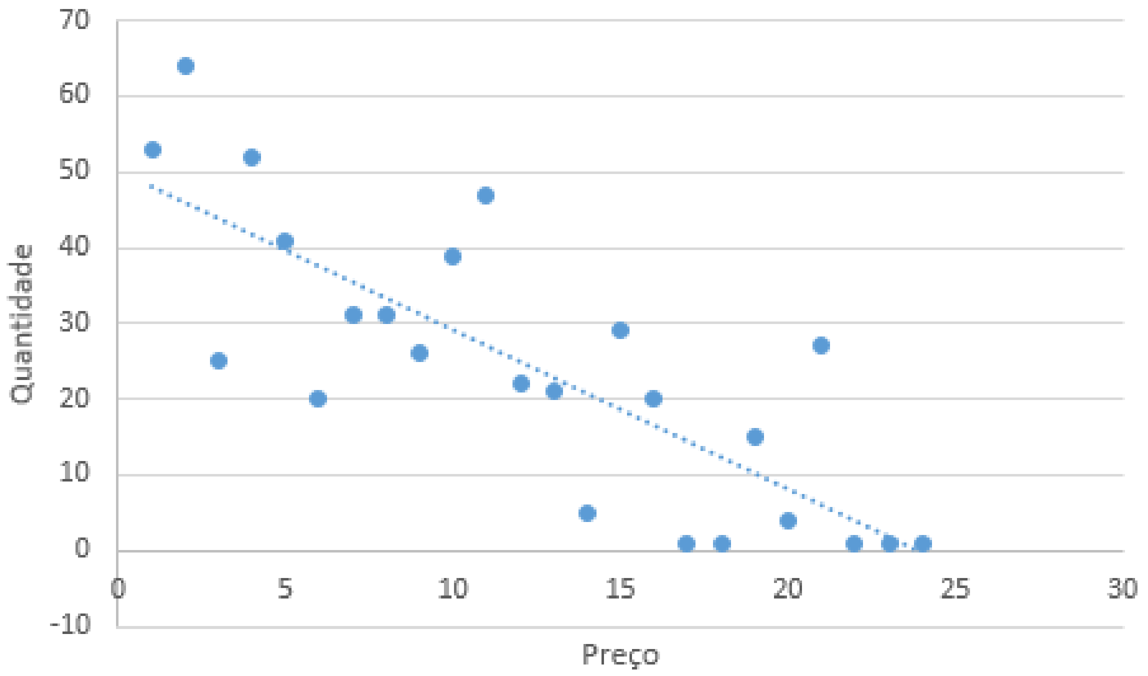


Conceito

- Neste caso, a métrica a ser utilizada é denominada “Coeficiente de correlação de Pearson”.
- A medida possui amplitude no intervalo -1 a 1, sendo:
- 0: nenhuma correlação
- -1: completa correlação negativa
- 1: completa correlação positiva

Conceito

- Como caso prático, vamos verificar a correlação entre Preço e Quantidade



Exemplo: exemplo_testf_correlação.xlsx

Conceito

- A medida de correlação é obtida pela fórmula:
- $r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$
- Aplicando a fórmula acima nos dados do slide anterior, temos uma correlação de -0,8008.
- Indica que existe uma forte correlação preço e quantidade vendida.

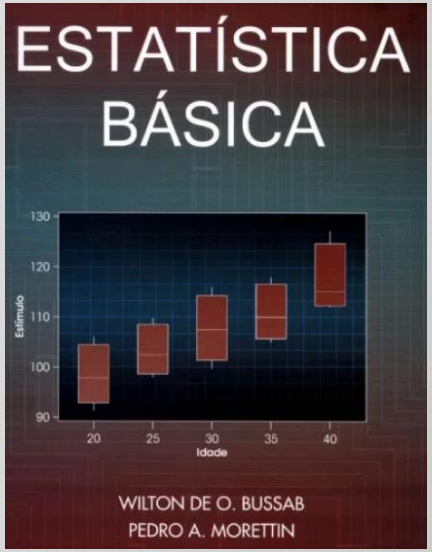
Conceito

- Tabela de referência da correlação:

Valores (em módulo)	Interpretação
0 – 0,3	Correlação **fraca**
0,3 – 0,7	Correlação **moderada**
0,7 - 1	Correlação **forte**

Onde estudar mais!!

- Leitura



- Aplicações teste F: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ANOVA/BS704_HypothesisTesting-Anova_print.html

- Vídeos

- Teste de ANOVA: https://pt.khanacademy.org/math/statistics-probability/analysis-of-variance-anova-library/analysis-of-variance-anova/v/anova-3-hypothesis-test-with-f-statistic

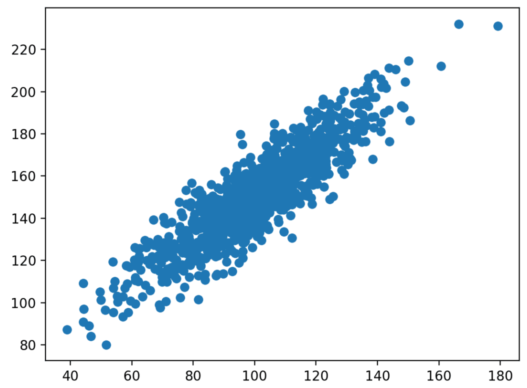
- Correlação:
- https://pt.khanacademy.org/math/statistics-probability/designing-studies/sampling-and-surveys/v/correlation-and-causality

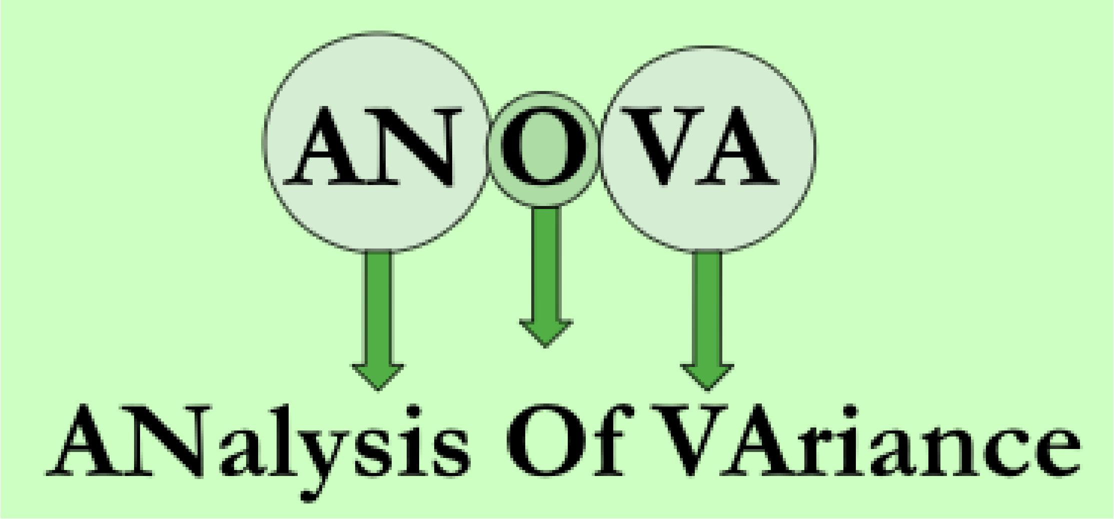
- Leitura

- https://link.springer.com/article/10.1057/jt.2009.5

Conceitos iniciais

- Teste F ANOVA (Prática no *Python*)
- Análise de correlação (Prática no *Python*)





Conceitos

- Teste F para comparar duas variâncias
- Teste F para comparar mais de 2 médias
- Análise de correlação

Procedimento Geral

- Etapas:

- Fixar a hipótese nula (H_0) e a hipótese alternativa (H_1).
- Definir qual estimador será utilizado (média, variância).
- Fixar o nível de confiança (90%, 95%, 99%).
- Utilizar os dados da amostra obtida para calcular o valor da estatística teste (t, F, Qui-Quadrado).
- Verificar se o valor da estatística teste esta contida ou não na região de rejeição. Caso esteja, rejeita H_0, contrário não rejeita H_0.

Teste *F e Correlação*

- No *python*, o teste *F* para comparar duas variâncias é realizada pela função *"f"* do módulo scipy.stats.
- No caso do teste F para comparação de médias podemos utilizar a função *"f_oneway"* do mesmo módulo.

```
from scipy.stats import f, f_oneway
```

- Para obter a correlação, uma das opções é utilizar a função *"corr"*, uma função nativa do *pandas*.

Teste F

- Será apresentado um dos primeiros testes de hipóteses que podem ser aplicados.
- O teste F pode ser aplicado quando:
- **Deseja-se comparar as variâncias de 2 populações**
- **Deseja-se comparar as médias de mais de 2 grupos.**

Teste *F*

- Teste F para comparar duas variâncias:

- Problema exemplo:

- Para verificar o grau de satisfação dos funcionários de uma área da empresa. Estimou-se o desvio-padrão de 2 filiais presentes em duas cidades semelhantes. Foram obtidos salários de 10 funcionários da cidade A e 15 da cidade B, sendo o desvio-padrão de A R$ 1.118,00 e o desvio-padrão de B R$ 1.342,00.

Teste *F* - resolução

- Fixar a hipótese nula (H_0) e a hipótese alternativa (H_1)

- $\boldsymbol{H_0: \sigma_A^2 = \sigma_B^2}$ **ou** $\boldsymbol{H_0: {\sigma_A^2}/{\sigma_B^2} = 1}$
- $\boldsymbol{H_a: \sigma_A^2 \neq \sigma_B^2}$ **ou** $\boldsymbol{H_a: {\sigma_A^2}/{\sigma_B^2} \neq 1}$

- Definir o estimador: **dividir as variâncias**
- Fixar o nível de confiança: **95%**

Teste *F*

- Carregar os dados do problema

```
dados_salarios = pd.read_csv('dados_salarios.csv')
```

- Estimar as estatísticas descritivas:

```
dados_salarios.groupby('cidade') \
             .agg(media_salarios = pd.NamedAgg('salarios', 'mean'),
                  dp_salarios = pd.NamedAgg('salarios', 'std'),
                  n = pd.NamedAgg('salarios', 'size')) \
             .reset_index()
```

Média, desvio-padrão e tamanho das amostras

cidade	media_salarios	dp_salarios	n
A	2964.052109	1117.505582	10
B	2432.859069	1342.126772	15

Vamos comparar estes desvios padrões de salários

Teste *F* - resolução

- Aplicamos o teste F no *python*:
- O teste é realizado em 3 etapas:
- Na primeira temos q obter a estatística F, obter os graus de liberdade e depois o valor *p*.
- Etapa 1: Obter o F:

```
salarios_cidade_a = dados_salarios[dados_salarios['cidade'] == 'A']['salarios']
salarios_cidade_b = dados_salarios[dados_salarios['cidade'] == 'B']['salarios']
```

```
f_valor = np.var(salarios_cidade_a, ddof=1) / np.var(salarios_cidade_b, ddof=1)
```

- Etapa 2: Graus de liberdade

```
gl_a = len(salarios_cidade_a) - 1
gl_b = len(salarios_cidade_b) - 1
```

Teste *F* - resolução

- Etapa 3: Obter valor *p*:

```
def f_p_value(f_statistic, df_n, df_d, test_type):

  '''test_type: greater, less, two.sided'''

  if test_type == 'greater':
    return 1 - f.cdf(f_valor, df_n, df_d)
  elif test_type == 'less':
    return f.cdf(f_valor, df_n, df_d)
  elif test_type == 'two.sided':
    p1 = f.cdf(f_valor, df_n, df_d)
    p2 = 1 - f.cdf(f_valor, df_n, df_d)
    return np.min([p1, p2])*2
  else:
    raise TypeError("'test_type' only accept options: 'greater', 'less' or 'two.sided'")
```

```
p = f_p_value(f_valor, gl_a, gl_b, 'two.sided')

f_valor, p

(0.6932858292724237, 0.5895222755367536)
```

p-value > 0,05.
Não rejeita H0

- Rejeitar ou não rejeitar H_0?
- O valor de **0,693 NÃO** está dentro da região de rejeição. Deste modo, **não rejeitamos a hipótese nula**, **a variância entre os salários são estatisticamente iguais**.

Teste *F*

- Teste *F* para comparar médias de mais de 2 grupos:

- Neste caso, temos o problema de testar as diferenças de médias quando existem mais de 2 grupos a serem comparados.

- As hipóteses são construídas conforme o seguinte:

- $\boldsymbol{H_0: \mu_1 = \mu_2 = \cdots = \mu_k}$
- $\boldsymbol{H_a: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_k}$

- Para resolver este problema precisaremos falar um pouco de teoria.

Teste *F*

- Teste F para comparar média de mais de 2 grupos:

- Problema exemplo:

- Uma pesquisa busca comparar a perda de peso em 3 tipos de dieta. Vamos analisar os dados e testar a hipótese de igualdade de médias.

Teste *F* - resolução

- Fixar a hipótese nula (H_0) e a hipótese alternativa (H_1)

- $\boldsymbol{H_0: \mu_{caloria} = \mu_{gordura} = \mu_{carb} = \mu_{controle}}$
- $\boldsymbol{H_a: \mu_{caloria} \neq \mu_{gordura} \neq \mu_{carb} \neq \mu_{controle}}$

- Definir o estimador: **ANOVA**
- Fixar o nível de confiança: **95%**

- Vamos carregar os dados e analisar.

Teste *F* - resolução

- Dados:

```
dados_dietas.groupby('dieta') \
             .agg(media_perdapeso = pd.NamedAgg('perda_peso_kg', 'mean'),
                  dp_perdapeso = pd.NamedAgg('perda_peso_kg', 'std'),
                  n = pd.NamedAgg('perda_peso_kg', 'size')) \
             .reset_index()
```

dieta	media_perdapeso	dp_perdapeso	n
baixa_caloria	6.6	2.302173	5
baixa_gordura	3.0	1.581139	5
baixo_carboidrato	3.4	1.140175	5
grupo_controle	1.2	1.643168	5

Média, desvio-padrão e tamanho das amostras

Vamos testar estas médias

Teste *F* - resolução

- Aplicamos o teste no *python*

```
dados_baixa_cal = dados_dietas[dados_dietas['dieta'] == 'baixa_caloria']['perda_peso_kg']
dados_baixa_gor = dados_dietas[dados_dietas['dieta'] == 'baixa_gordura']['perda_peso_kg']
dados_baixo_cab = dados_dietas[dados_dietas['dieta'] == 'baixo_carboidrato']['perda_peso_kg']
dados_controle = dados_dietas[dados_dietas['dieta'] == 'grupo_controle']['perda_peso_kg']
```

```
f_oneway(dados_baixa_cal, dados_baixa_gor, dados_baixo_cab, dados_controle)
```

```
F_onewayResult(statistic=8.559322033898304, pvalue=0.0012777417892066623)
```

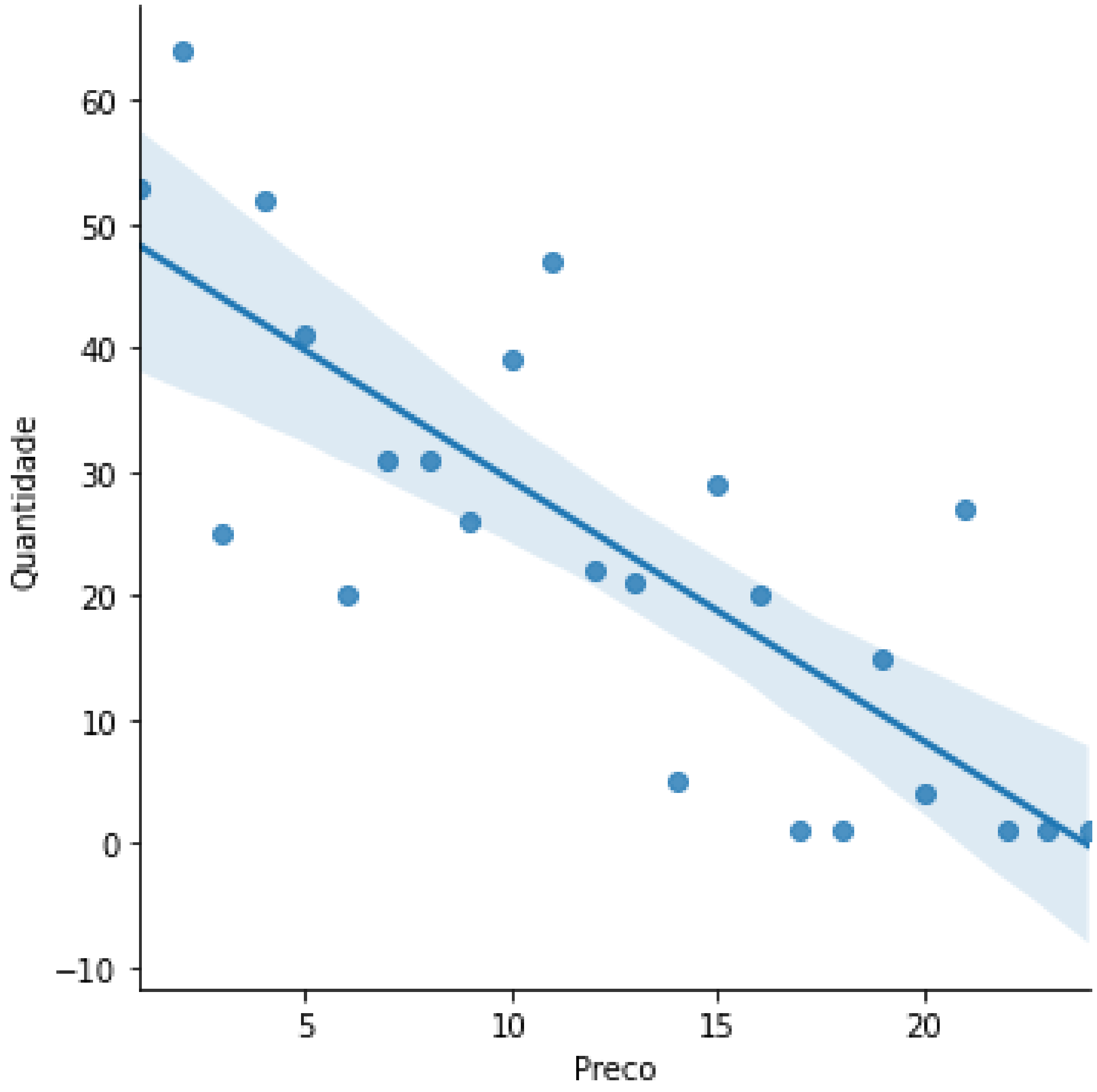
p-value < 0,05.
Rejeita H0.

- Analisamos os resultados:

- Valor de F: **8,559**
- Rejeitar ou não rejeitar H_0?
- O valor de **8,559** está dentro da região de rejeição. Deste modo, **rejeitamos a hipótese nula**, **as médias de perda de peso entre os diferentes tratamentos são diferentes**.

Correlação

- Vamos agora verificar a correlação entre preço e quantidade:



Correlação

- Vamos agora verificar a correlação entre preço e quantidade.

- Podemos obter a correlação no *python* utilizando a função *“corr”*.

```
dos_preco_quant.cor
```

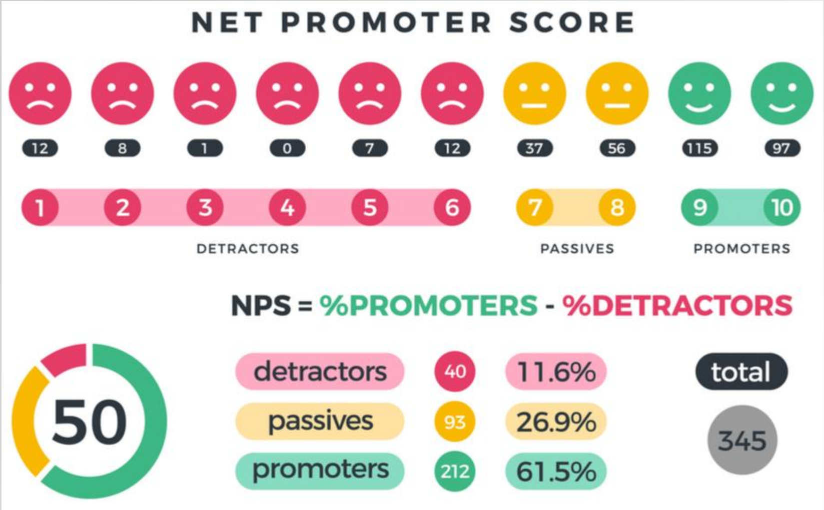
	Preco
Preco	1.000000
uantidade	-0.800847

Problema

- O *Net Promoter Score* (NPS) é um índice utilizado por empresas para avaliar a satisfação dos clientes.

- Os consumidores avaliam a satisfação com o serviço prestado pela empresa atribuindo um nota de zero a dez.

- Esta avaliação fornece um diagnóstico simplificado pelos clientes em relação a empresa.



Vamos analisar o banco de dados a seguir!!!

Teste *F*

- Carregar os dados:

```
dados_nps = pd.read_csv('nps_example.csv', sep = ';')
```

- Verificar se todas as respostas estão completas:

```
dados_nps.groupby('response_status') \
         .size() \
         .to_frame('n') \
         .reset_index()
```

response_status	n
Complete	2281
Incomplete	265
Terminated	33

Temos 2281 dados completos

Teste *F*

- Filtrar os dados (alguns dados estão faltantes):

```
dados_nps_filtrados = dados_nps[(dados_nps['response_status'] == 'Complete') & \
                                (dados_nps['nps_score'].notna())]
```

Filtramos o valor “Complete” na coluna “response_status”.

Também filtramos algum
: “nps_score”
A.

```
dados_nps[dados_nps['nps_score'].isnull()]
```

	id	response_status	how_long_listening	age	nps_score	gender
17	11706467	Incomplete	Less than 6 months	18-24	NaN	NaN
31	11706938	Incomplete	1 year to less than 3 years	25-34	NaN	NaN

Teste *F*

- Vamos obter as estatísticas descritivas:

```
dados_nps_filtrados.groupby('age') \
                   .agg(media_nps = pd.NamedAgg('nps_score', 'mean'),
                        dp_nps = pd.NamedAgg('nps_score', 'std'),
                        n = pd.NamedAgg('nps_score', 'size')) \
                   .reset_index()
```

age	media_nps	dp_nps	n
18-24	9.464539	1.116275	282
25-34	9.694828	0.957639	580
35-44	9.707612	0.979501	578
45-54	9.719039	0.928254	541
55-64	9.733871	0.923020	248
65-74	9.423077	1.361560	26
75+	8.000000	0.000000	2

Teremos que retirar esta faixa etária. Poucos dados

Teste *F*

- Filtrar a faixa “75+”:

```
dados_nps_filtrados_aj = dados_nps_filtrados[dados_nps_filtrados['age'] != '75+']
```

```
dados_nps_filtrados_aj.groupby('age') \
                    .agg(media_nps = pd.NamedAgg('nps_score', 'mean'),
                        dp_nps = pd.NamedAgg('nps_score', 'std'),
                        n = pd.NamedAgg('nps_score', 'size')) \
                    .reset_index()
```

Filtrar a faixa “75+”

age	media_nps	dp_nps	n
18-24	9.464539	1.116275	282
25-34	9.694828	0.957639	580
35-44	9.707612	0.979501	578
45-54	9.719039	0.928254	541
55-64	9.733871	0.923020	248
65-74	9.423077	1.361560	26

Vamos realizar o teste de hipótese para verificar se existe diferença entre estas médias quanto a faixa etária

Resolução

- Construção do teste de hipótese:

- Fixar a hipótese nula (H_0) e a hipótese alternativa (H_1)

- $\boldsymbol{H_0}$: $\boldsymbol{Todas\ as\ m}$é$\boldsymbol{dias\ s}$ã$\boldsymbol{o\ iguais}$
- $\boldsymbol{H_a}$: $\boldsymbol{Alguma\ das\ m}$é$\boldsymbol{dias}$ é $\boldsymbol{diferente}$

- Definir o estimador: **ANOVA**
- Fixar o nível de confiança: **95%**

- Vamos testar no Python

Teste *F* - resolução

- Aplicamos o teste no *Python*:

```
dados_18_24 = dados_nps_filtrados_aj[dados_nps_filtrados_aj['age'] == '18-24']['nps_score']
dados_25_34 = dados_nps_filtrados_aj[dados_nps_filtrados_aj['age'] == '25-34']['nps_score']
dados_35_44 = dados_nps_filtrados_aj[dados_nps_filtrados_aj['age'] == '35-44']['nps_score']
dados_45_54 = dados_nps_filtrados_aj[dados_nps_filtrados_aj['age'] == '45-54']['nps_score']
dados_55_64 = dados_nps_filtrados_aj[dados_nps_filtrados_aj['age'] == '55-64']['nps_score']
dados_65_74 = dados_nps_filtrados_aj[dados_nps_filtrados_aj['age'] == '65-74']['nps_score']
```

```
f_oneway(dados_18_24, dados_25_34, dados_35_44,
         dados_45_54, dados_55_64, dados_65_74)
```

```
F_onewayResult(statistic=3.5221660981040768, pvalue=0.0035606861304276695)
```

p-value < 0,05.
Rejeita H0.

- Valor de F: **3,522**
- Rejeitar ou não rejeitar H_0?
- O valor de **3,522 ESTÁ** dentro da região de rejeição. Deste modo, **rejeitamos a hipótese nula**, **o nível de satisfação muda conforme a idade dos respondentes**.

Problema

- A bolsa de valores é um ambiente propício para diversas análises de dados.

- Conhecer as melhores técnicas de análises neste ambiente é crucial para obter os maiores retornos sem correr altos riscos.

- Entender a correlação entre as ações do mercado financeiro é uma destas análises importantes.

Vamos analisar o banco de dados a seguir!!!

Correlação

- Carregar os dados:

```
dados_bolsa = pd.read_csv('dados_bolsa.csv', sep = ';', decimal = ',')
```

- Os *missings*:

```
dados_bolsa[dados_bolsa.isnull()]
```

	data	petr4	bbdc3	vale5	ambv4	itub4
0	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN

missings

Correlação

- Filtrar:

```
dados_bolsa_filtrados = dados_bolsa.dropna()
```

- Correlações:

```
dados_bolsa_filtrados.corr()
```

	petr4	bbdc3	vale5	ambv4	itub4
petr4	1.000000	0.539247	0.724023	0.392074	0.593834
bbdc3	0.539247	1.000000	0.592143	0.470529	0.778506
vale5	0.724023	0.592143	1.000000	0.482919	0.642838
ambv4	0.392074	0.470529	0.482919	1.000000	0.488886
itub4	0.593834	0.778506	0.642838	0.488886	1.000000

A função seleciona de forma automática somente as colunas numéricas

Obrigado!