

Análise Estatística de Dados

CAPÍTULO 1. INTRODUÇÃO A DISCIPLINA

PROF. MÁIRON CHAVES

Análise Estatística de Dados

AULA 1.1. INTRODUÇÃO A DISCIPLINA

PROF. MÁIRON CHAVES

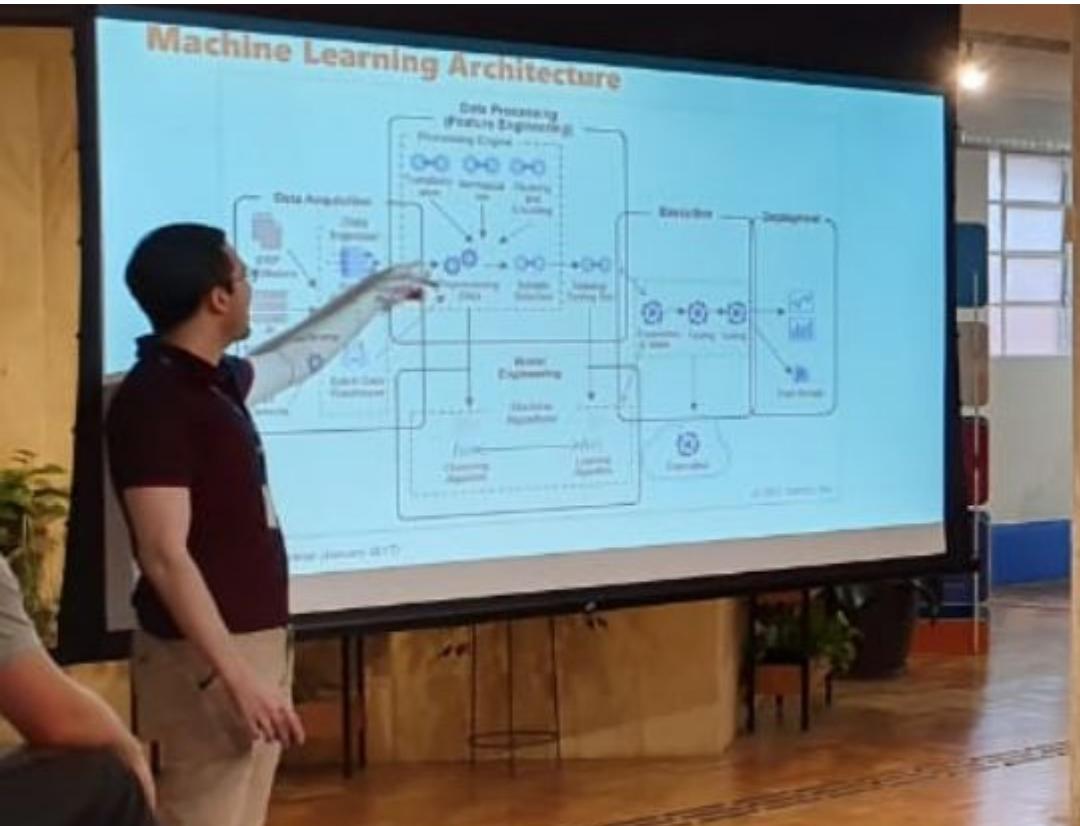
Nesta aula



- Objetivos do Curso.
- Introdução a Disciplina.
- O que é Estatística?
- O que é Ciência de Dados?
- Ciência de Dados vs. Estatística.

Objetivos do curso

IGTI



Sobre o professor:
Máiron Chaves

Formações Acadêmicas

- Mestrado em Modelagem Matemática e Computacional – CEFET (disciplinas isoladas).
- Especialista em Estatística, pela UFMG.
- Especialista em Inteligência de Mercado, pela UNA.

Experiência Profissional

- Atua desde 2010 com análise de dados e inteligência de negócio.
- Atua desde 2015 com modelagem estatística e aprendizado de máquina.

Contatos

maironchaves@hotmail.com

<https://www.linkedin.com/in/maironchaves/>

Objetivos do curso



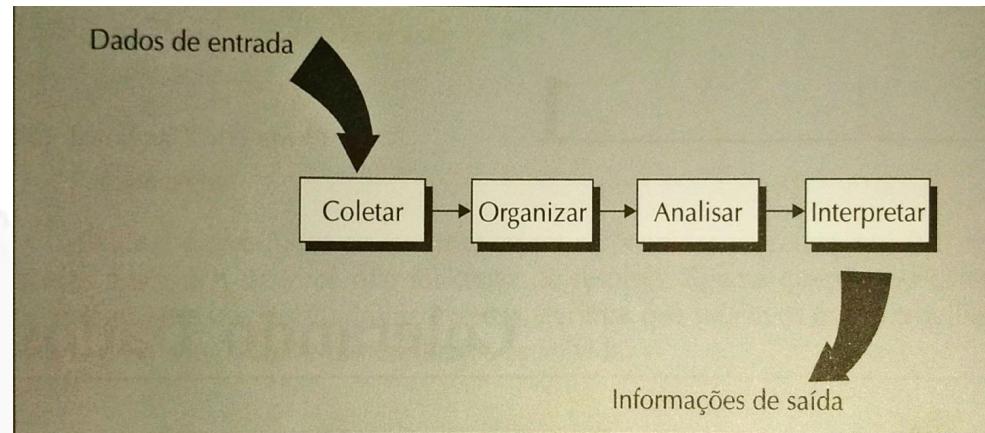
Espera-se que o aluno consiga, ao final da disciplina:

1. Saber o que é a Estatística e sua importância para o profissional que lida com dados.
2. Realizar análise exploratória em bases de dados.
3. Entendimento sobre teoria das probabilidades e como aplicar no mundo real.
4. Propor hipóteses e buscar por evidências nos dados.
5. Construir modelos preditivos através de Regressão Linear.

O que é a Estatística?

IGTI

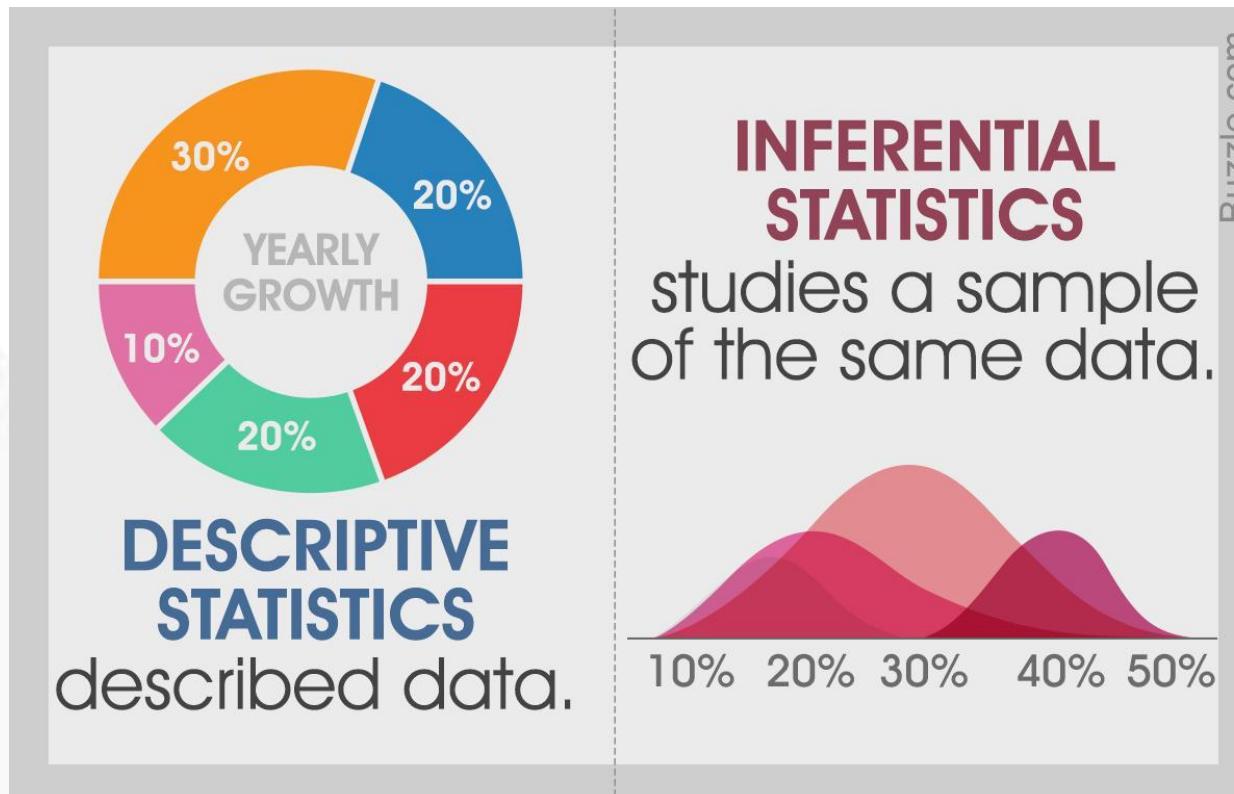
A Estatística é um ramo das ciências exatas que visa obter conclusões a partir de dados, envolvendo técnicas para coletar, organizar, descrever, analisar e interpretar dados. (SMAILES & MCGRANE, 2012)



O que é a Estatística?

IGTI

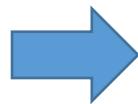
Estatística Descritiva vs. Estatística Inferencial



O que é a Estatística?

Dos laboratórios para as empresas

iGTD



O que é a Estatística?

IGTI

Dos laboratórios para as empresas



Manuela

H_0 : Quando o produto esta na posição B, seu faturamento é igual ao seu faturamento quando está na posição A

H_1 : Quando o produto esta na posição B, seu faturamento é maior do que quando está na posição A



Letícia



O que é a Ciência de Dados?

“A ciência de dados é um conjunto de métodos que cercam a extração do conhecimento a partir dos dados” - Fawcett e Provost (2016) em *Data Science para Negócios*.

É uma área multidisciplinar

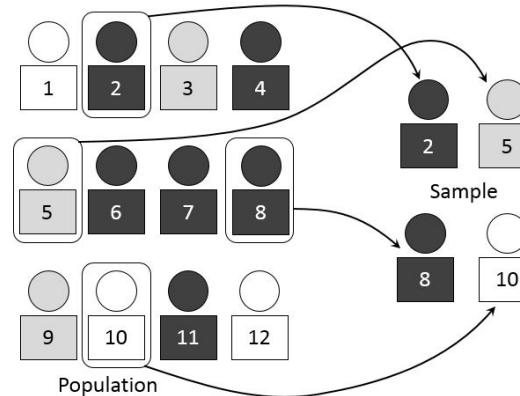


Ciência de Dados vs. Estatística

IGTI

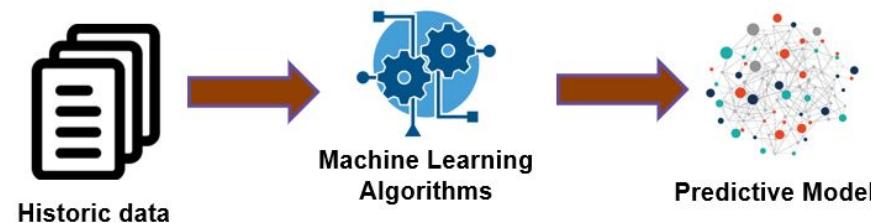
Estatística:

Foco em tirar conclusões de uma população, a partir de um amostra.



Ciência de Dados:

Foco em construir modelos preditivos.



Conclusão



- ✓ O que é Estatística?
- ✓ O que é Ciência de Dados?
- ✓ Ciência de Dados vs. Estatística.



Próxima aula



- Medidas de Centralidade.
- Medidas de Dispersão.

Análise Estatística de Dados

AULA 1.2. MEDIDAS DE CENTRALIDADE E DISPERSÃO

PROF. MÁIRON CHAVES

Nesta aula



- Medidas de Centralidade (média aritmética e mediana).
- Medidas de Dispersão (variância, desvio padrão, coeficiente de variação, amplitude e quartis).

Medidas de Centralidade e Dispersão

IGTI

São medidas quantitativas para resumir e compreender o comportamento de uma variável.

Contextualização

Preco_Cafe
4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79

n = 30

Medidas de Centralidade e Dispersão

Média Aritmética

É a soma dos valores dividido pela quantidade de amostras. Ou seja, o resultado dessa divisão equivale a um valor médio entre todos os valores.

Preco_Cafe
4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79

n = 30

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\bar{x} = \frac{132,79}{30}$$

$$\bar{x} = 4,42$$

A interpretação fica: O hipermercado geralmente pratica o preço de 4,42 reais para o café.



Medidas de Centralidade e Dispersão

Desvio Padrão

É uma medida de dispersão que indica o quanto longe os valores de uma variável estão do valor médio.



Preco_Cafe	Média	(Preço - Média)^2
4,77	4,4263	0,1181
4,67	4,4263	0,0594
4,75	4,4263	0,1048
4,74	4,4263	0,0984
4,63	4,4263	0,0415
4,56	4,4263	0,0179
4,59	4,4263	0,0268
4,75	4,4263	0,1048
4,75	4,4263	0,1048
4,49	4,4263	0,0041
4,41	4,4263	0,0003
4,32	4,4263	0,0113
4,68	4,4263	0,0643
4,66	4,4263	0,0546
4,42	4,4263	0,0000
4,71	4,4263	0,0805
4,66	4,4263	0,0546
4,46	4,4263	0,0011
4,36	4,4263	0,0044
4,47	4,4263	0,0019
4,43	4,4263	0,0000
4,4	4,4263	0,0007
4,61	4,4263	0,0337
4,09	4,4263	0,1131
3,73	4,4263	0,4849
3,89	4,4263	0,2877
4,35	4,4263	0,0058
3,84	4,4263	0,3438
3,81	4,4263	0,3799
3,79	4,4263	0,4049

n = 30

$\Sigma = 3,00$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$s = \sqrt{\frac{3,00}{30 - 1}}$$

$$s = \sqrt{0,10}$$

$$\underline{s = 0,31}$$

A interpretação fica: Os preços praticados para o café variam em média 0,31 centavos em torno do seu preço médio.

Medidas de Centralidade e Dispersão

Coeficiente de Variação

É uma medida padronizada de dispersão, é a razão entre o desvio padrão e a média.

Preco_Cafe
4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79

$$\text{Coeficiente de Variação: } \frac{s}{\bar{x}} * 100$$

$$\text{Coeficiente de Variação: } \frac{0,31}{4,42} * 100$$

$$\text{Coeficiente de Variação: } 7,01 \%$$



A interpretação fica: Os preços praticados para o café variam em média 7,01% em torno do preço médio.

Medidas de Centralidade e Dispersão

Quartis

São valores que dividem a variável em quatro partes iguais, e assim cada parte representa 25% da variável.

	Preco_Cafe
1	3,73
2	3,79
3	3,81
4	3,84
5	3,89
6	4,09
7	4,32
8	4,35
9	4,36
10	4,40
11	4,41
12	4,42
13	4,43
14	4,46
15	4,47
16	4,49
17	4,56
18	4,59
19	4,61
20	4,63
21	4,66
22	4,66
23	4,67
24	4,68
25	4,71
26	4,74
27	4,75
28	4,75
29	4,75
30	4,77

Medidas de Centralidade e Dispersão



Quartis

A mediana é robusta a outliers!

Preco_Cafe

4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79

Preco_Cafe

4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79
381,00

Média 4,43
Mediana 4,48

Média 17,00
Mediana 4,53

Medidas de Centralidade e Dispersão



Amplitude

É o intervalo entre o valor máximo e o valor mínimo.

Preco_Cafe
4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79

Amplitude: $\max(x)$ até $\min(x)$

Amplitude: 3,73 até 4,77

A interpretação fica: Os preços praticados para o café variam entre 3,73 e 4,77

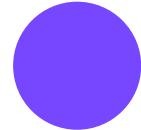
Conclusão



- ✓ Média.
- ✓ Desvio Padrão.
- ✓ Variância.
- ✓ Coeficiente de Variação.
- ✓ Quartis (Q1, Q2 e Q3).
- ✓ Amplitude.



Próxima aula



- Análise de Dados Através de Gráficos.

Análise Estatística de Dados

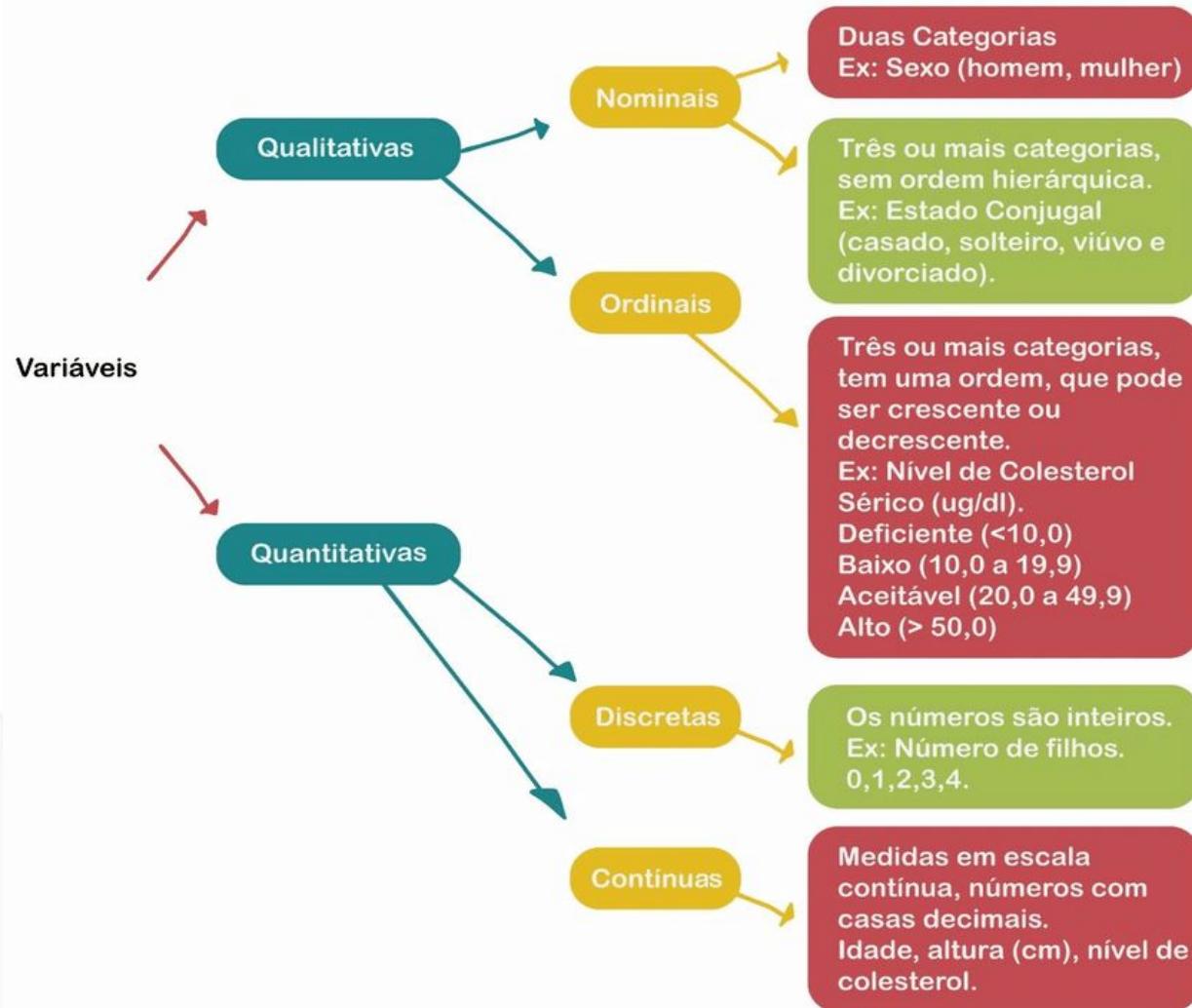
AULA 1.3. ANÁLISE DE DADOS ATRAVÉS DE GRÁFICOS

PROF. MÁIRON CHAVES

Nesta aula

- Tipos de variáveis.
- Histograma.
- Boxplot.
- Gráfico de Linhas.
- Gráfico de Dispersão.
- Gráfico de Setores.

Tipos de variáveis



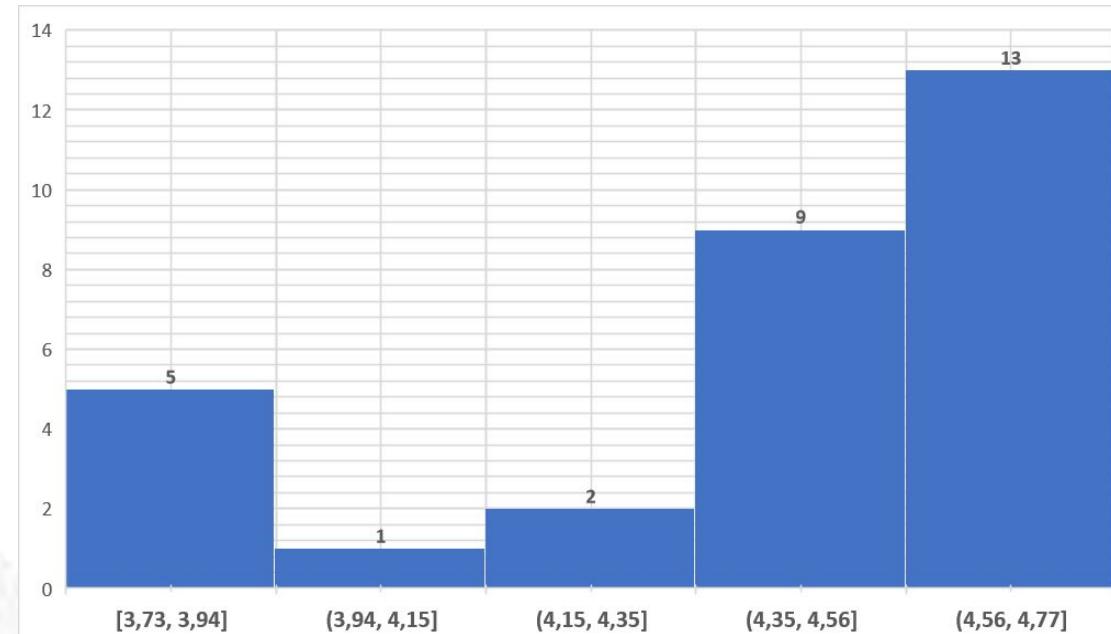
Análise de dados através de gráficos

Histograma

É uma representação gráfica em barras, dividida em classes. A altura de cada barra representa a frequência com que o valor da classe ocorre.

Preco_Cafe
4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79

Distribuição dos Preços do Café



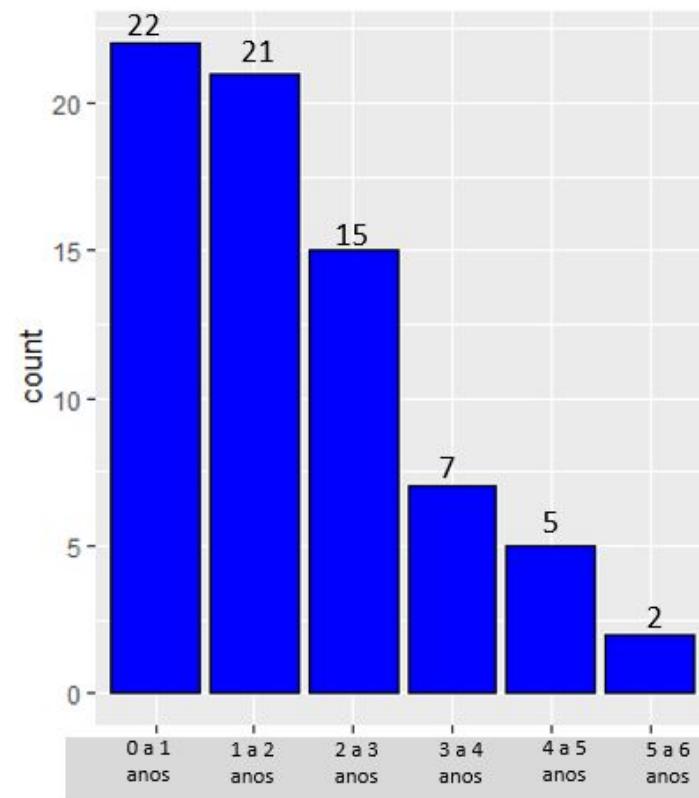
Análise de dados através de gráficos



Histograma (exemplo dois)

É uma representação gráfica em barras, dividida em classes. A altura de cada barra representa a frequência com que o valor da classe ocorre.

Distribuição do tempo que os clientes permanecem com contrato ativo

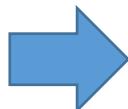


Análise de dados através de gráficos

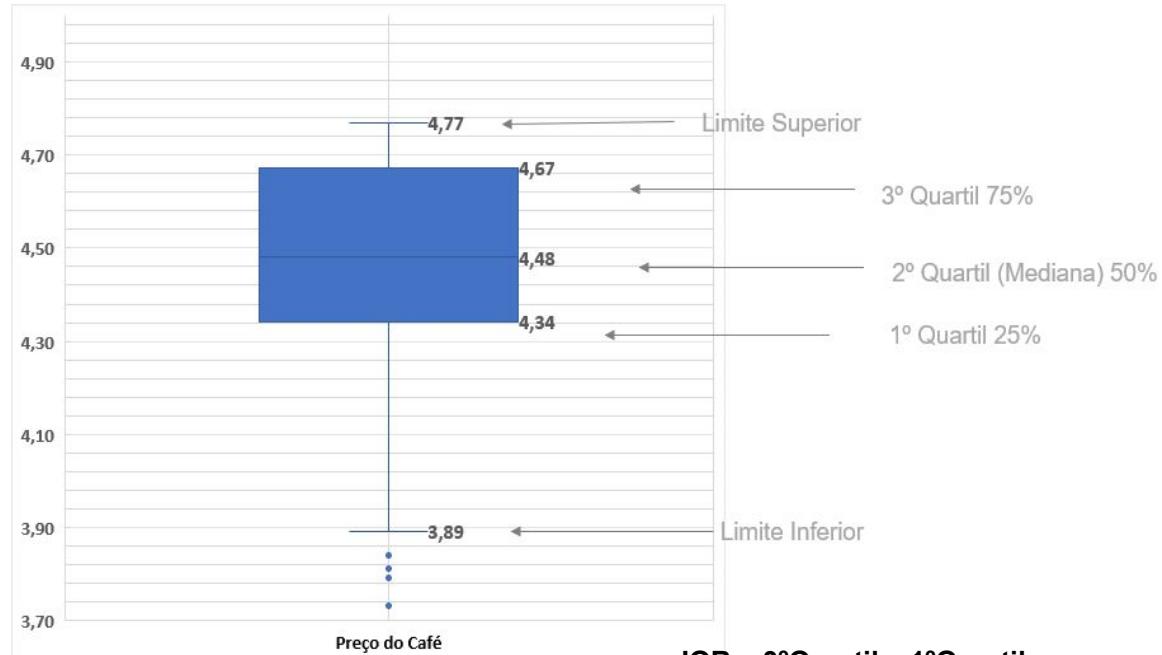
Boxplot

Visualização de uma variável baseada em seus quartis.

Preço_Café
4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79



Distribuição dos Preços do Café



$$\text{Limite Inferior} = \text{1ºQuartil} - (1.5 * \text{IQR})$$

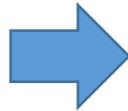
$$\text{Limite Superior} = \text{3ºQuartil} + (1.5 * \text{IQR})$$

Análise de dados através de gráficos

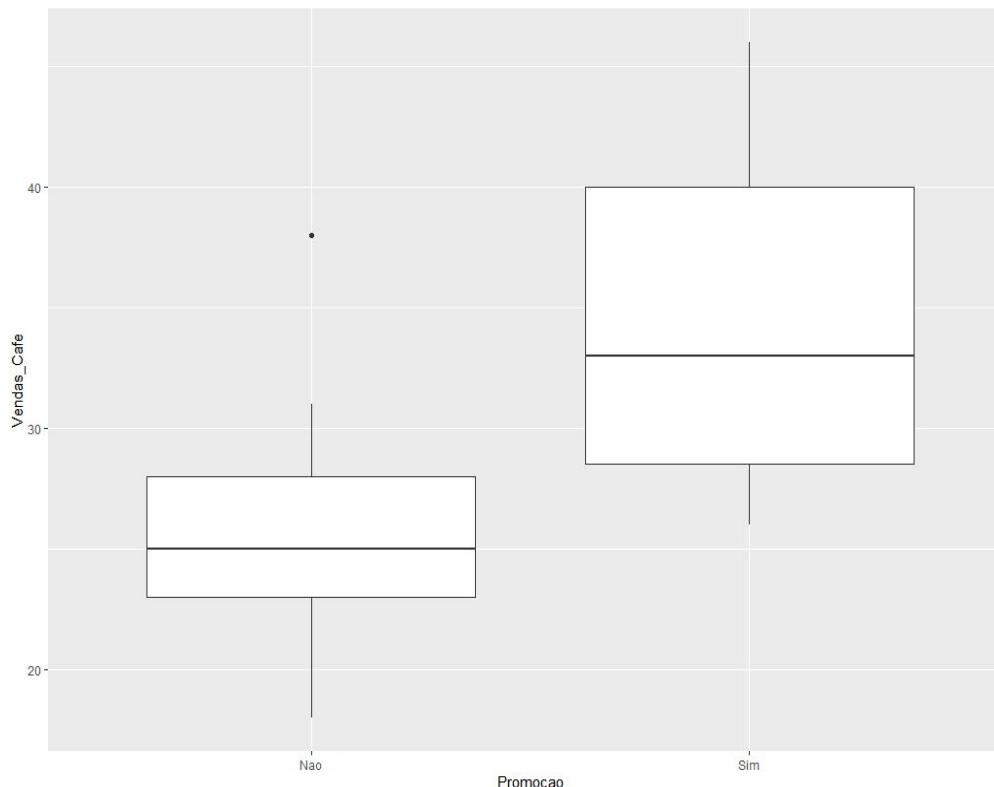
Boxplot – Exemplo dois

Relação entre uma variável qualitativa e uma

Promocao	Preco_Cafe	Vendas_Cafe
Nao	4,77	18
Nao	4,67	20
Nao	4,75	23
Nao	4,74	23
Nao	4,63	23
Nao	4,56	23
Nao	4,59	24
Nao	4,75	25
Sim	4,75	26
Nao	4,49	26
Sim	4,41	26
Nao	4,32	26
Nao	4,68	27
Sim	4,66	28
Sim	4,42	28
Nao	4,71	29
Sim	4,66	29
Sim	4,46	30
Sim	4,36	30
Nao	4,47	31
Nao	4,43	31
Sim	4,4	33
Sim	4,61	34
Sim	4,09	35
Nao	3,73	38
Sim	3,89	39
Sim	4,35	41
Sim	3,84	44
Sim	3,81	44
Sim	3,79	46



Vendas Durante a Promoção VS Vendas fora da Promoção

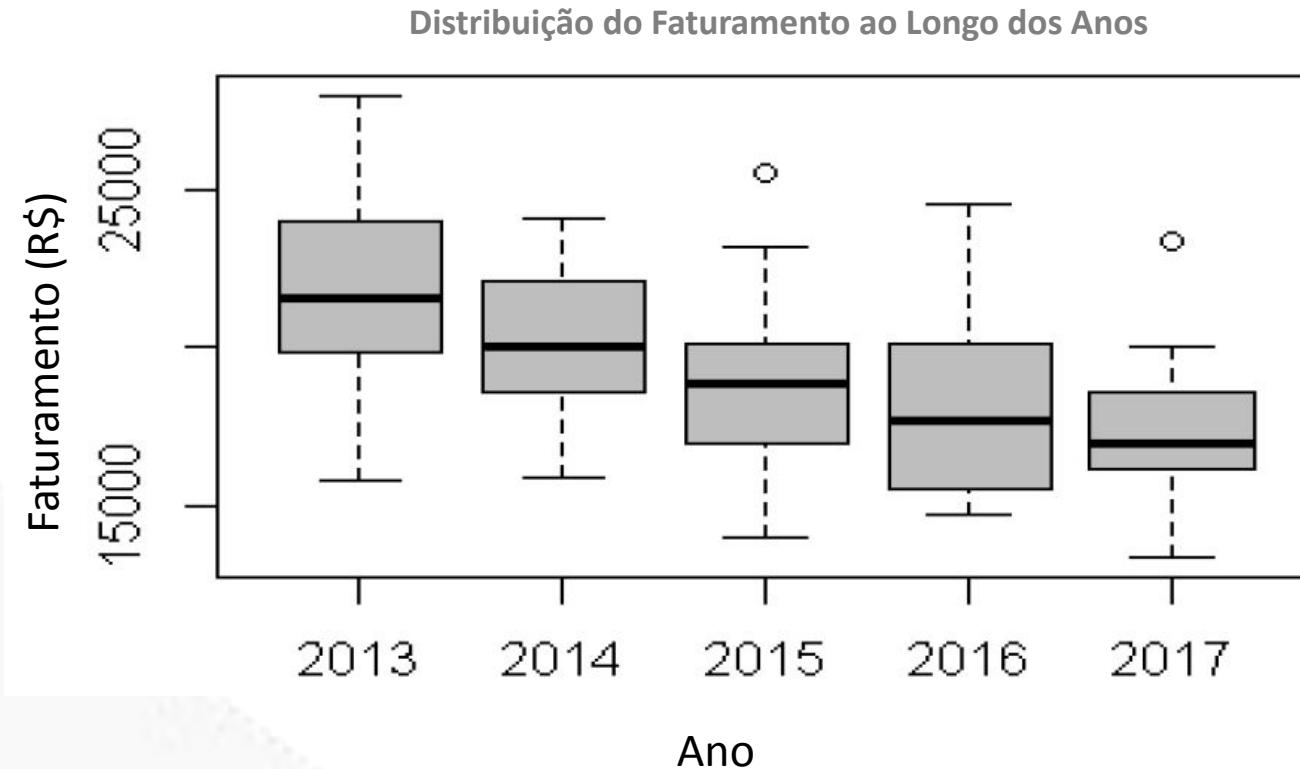


Análise de dados através de gráficos

IGTI

Boxplot – Exemplo três

Visualização de distribuição de uma série temporal ao longo do tempo.

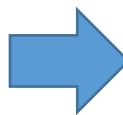


Análise de dados através de gráficos

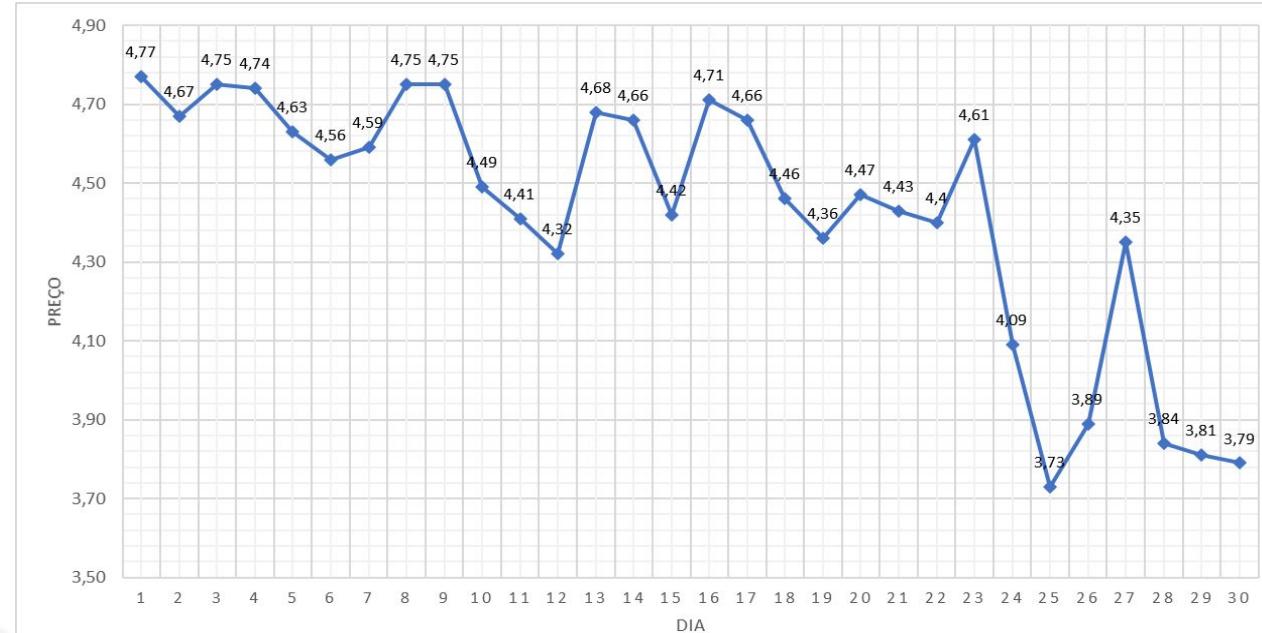
Gráfico de Linhas

Mostra a evolução de uma variável ao longo do tempo.

Preco_Cafe
4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79



Evolução dos Preços Diários do Café

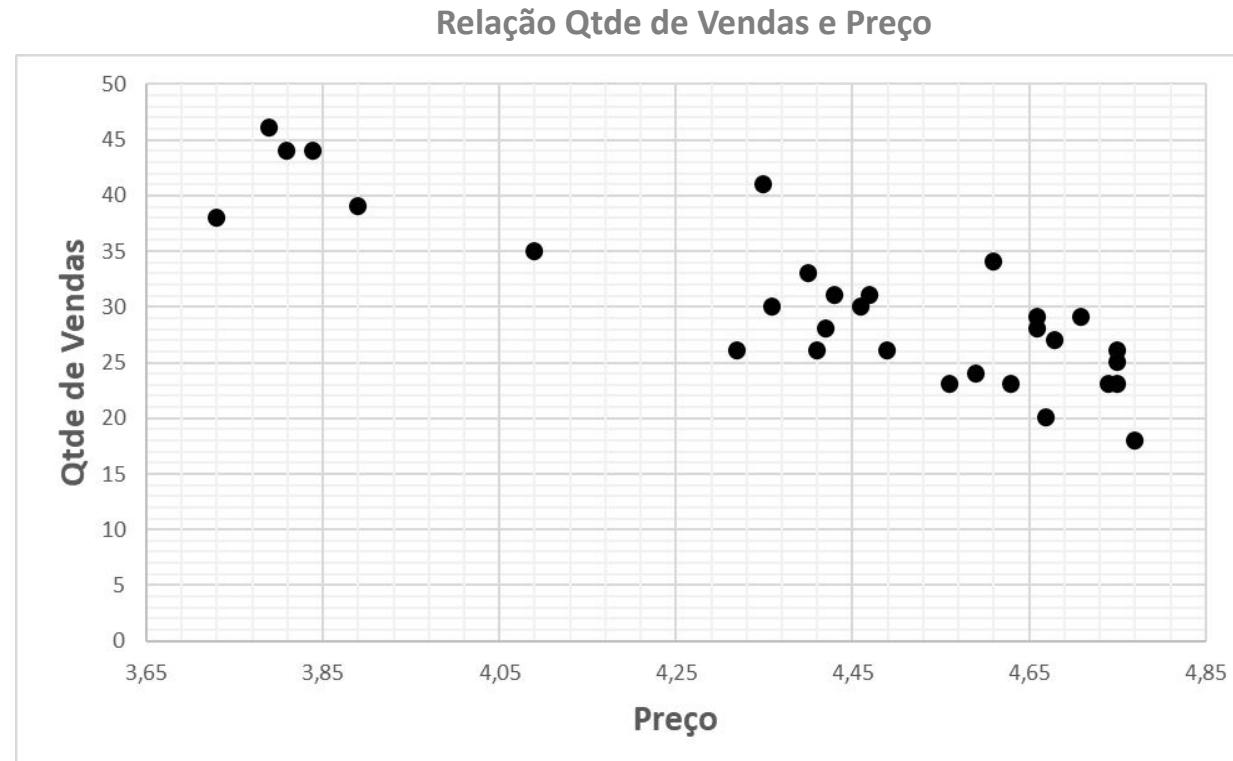
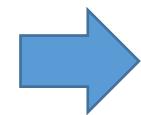


Análise de dados através de gráficos

Gráfico de Dispersão (scatterplot)

Utilizado para analisar a relação entre duas variáveis quantitativas.

Preco_Cafe	Vendas_Cafe
4,77	18
4,67	20
4,75	23
4,74	23
4,63	23
4,56	23
4,59	24
4,75	25
4,75	26
4,49	26
4,41	26
4,32	26
4,68	27
4,66	28
4,42	28
4,71	29
4,66	29
4,46	30
4,36	30
4,47	31
4,43	31
4,4	33
4,61	34
4,09	35
3,73	38
3,89	39
4,35	41
3,84	44
3,81	44
3,79	46

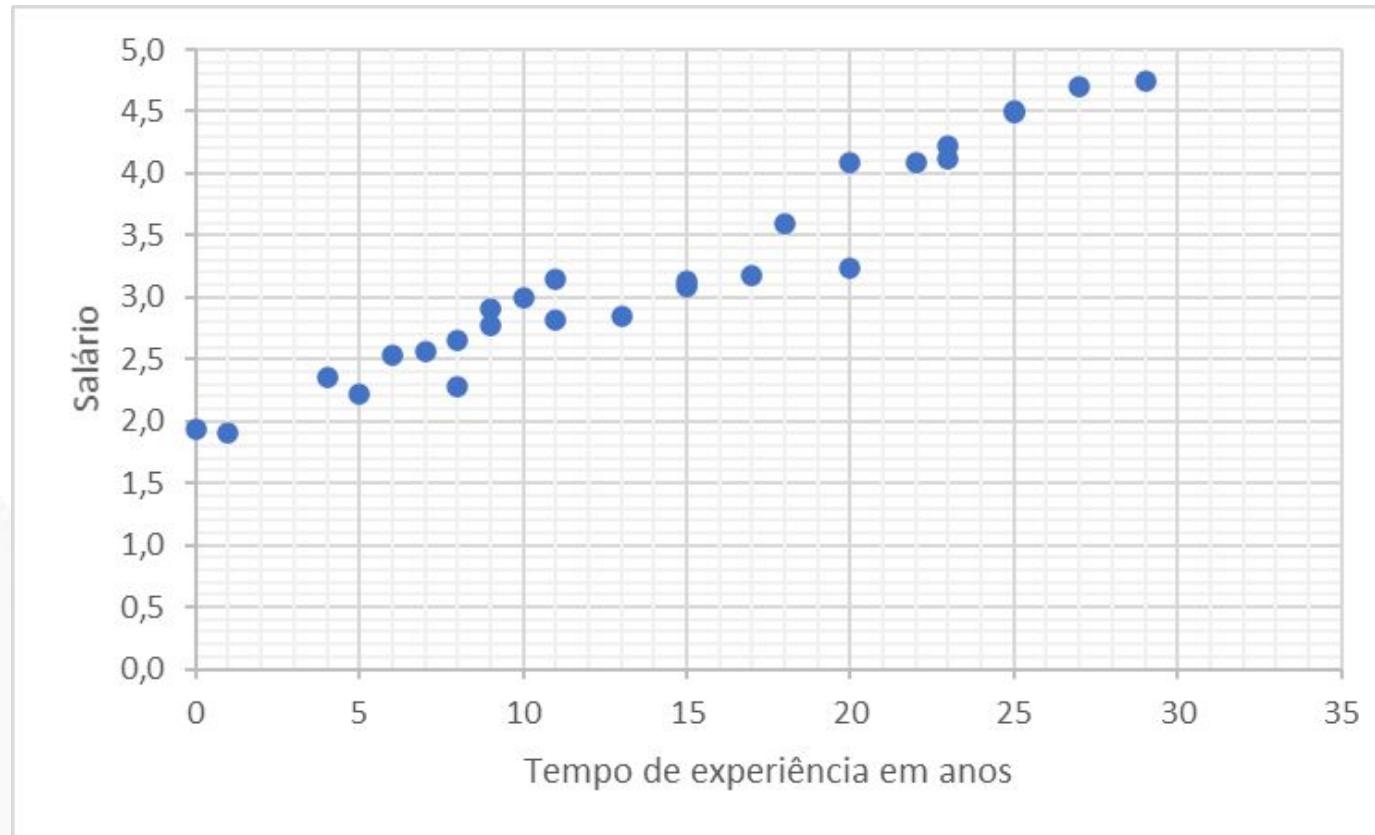


Análise de dados através de gráficos

Gráfico de Dispersão – Exemplo dois

Utilizado para analisar a relação entre duas variáveis quantitativas.

Relação Salário e Tempo de Experiência

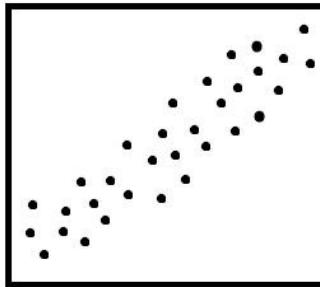


Análise de dados através de gráficos

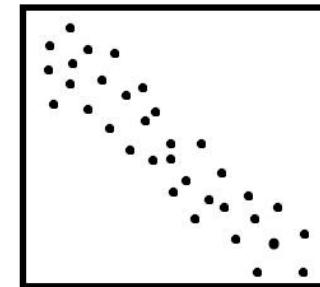


Gráfico de Dispersão – Exemplos hipotéticos

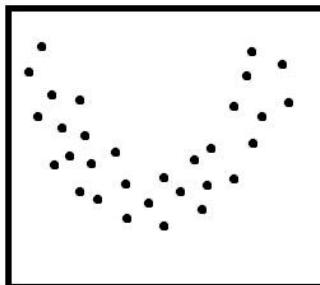
Utilizado para analisar a relação entre duas variáveis quantitativas.



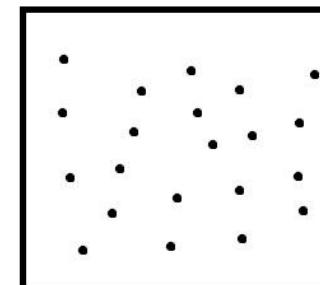
positive linear association



negative linear association



nonlinear association



no association

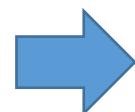
Fonte: <https://www.tes.com/teaching-resource/scatter-plots-19eCRQ/unit-2-data-and-statistics-review-help>

Análise de dados através de gráficos

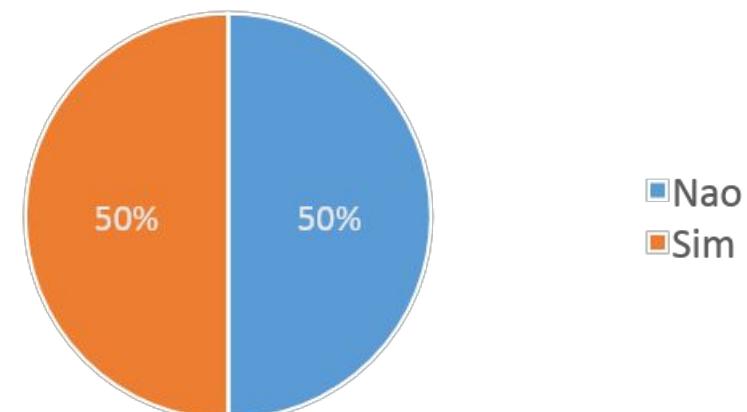
Gráfico de Setores

Exibe a proporção de cada nível categórico de uma variável.

Promocao
Nao
Sim
Nao
Sim
Nao
Nao
Sim
Sim
Nao
Sim
Sim
Sim
Nao
Nao
Sim
Sim
Sim
Sim
Nao
Sim



Percentual de Vendas em Promoção



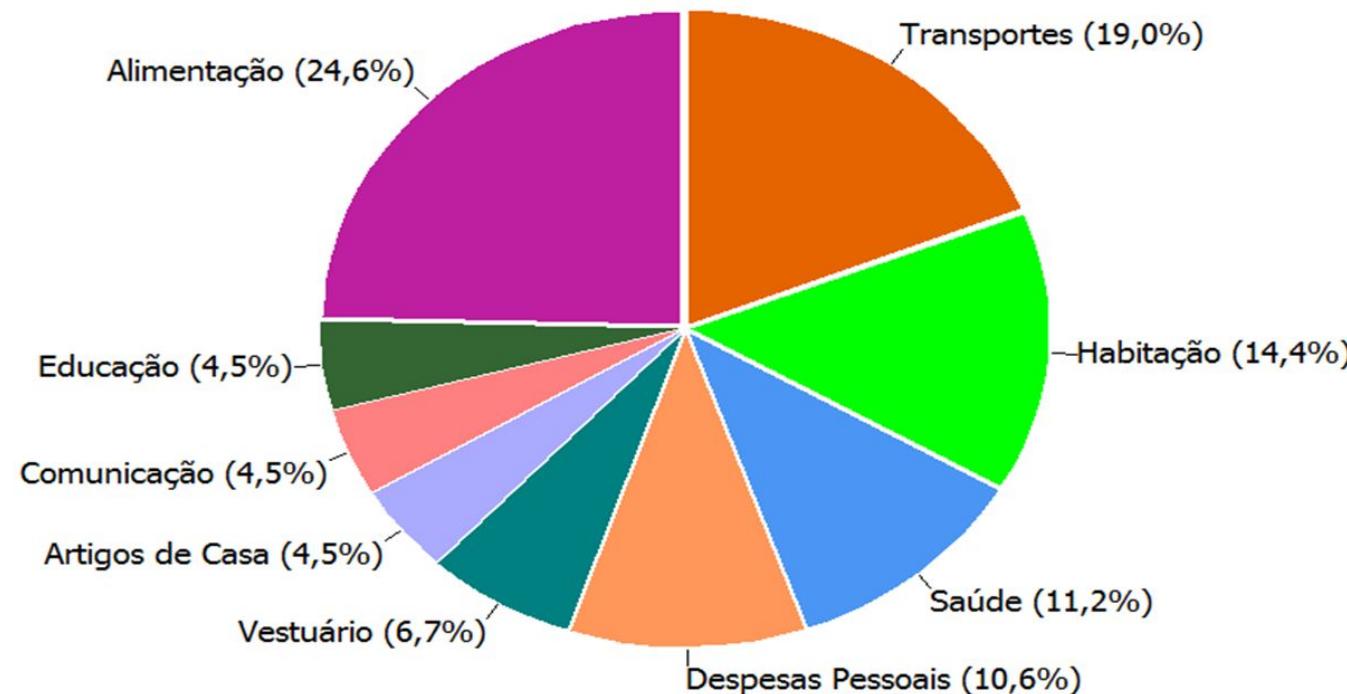
Análise de dados através de gráficos

Gráfico de Setores – CUIDADOS

Exibe a proporção de cada nível categórico de uma variável.

iGTD

Peso dos itens no IPCA (em jan/14)



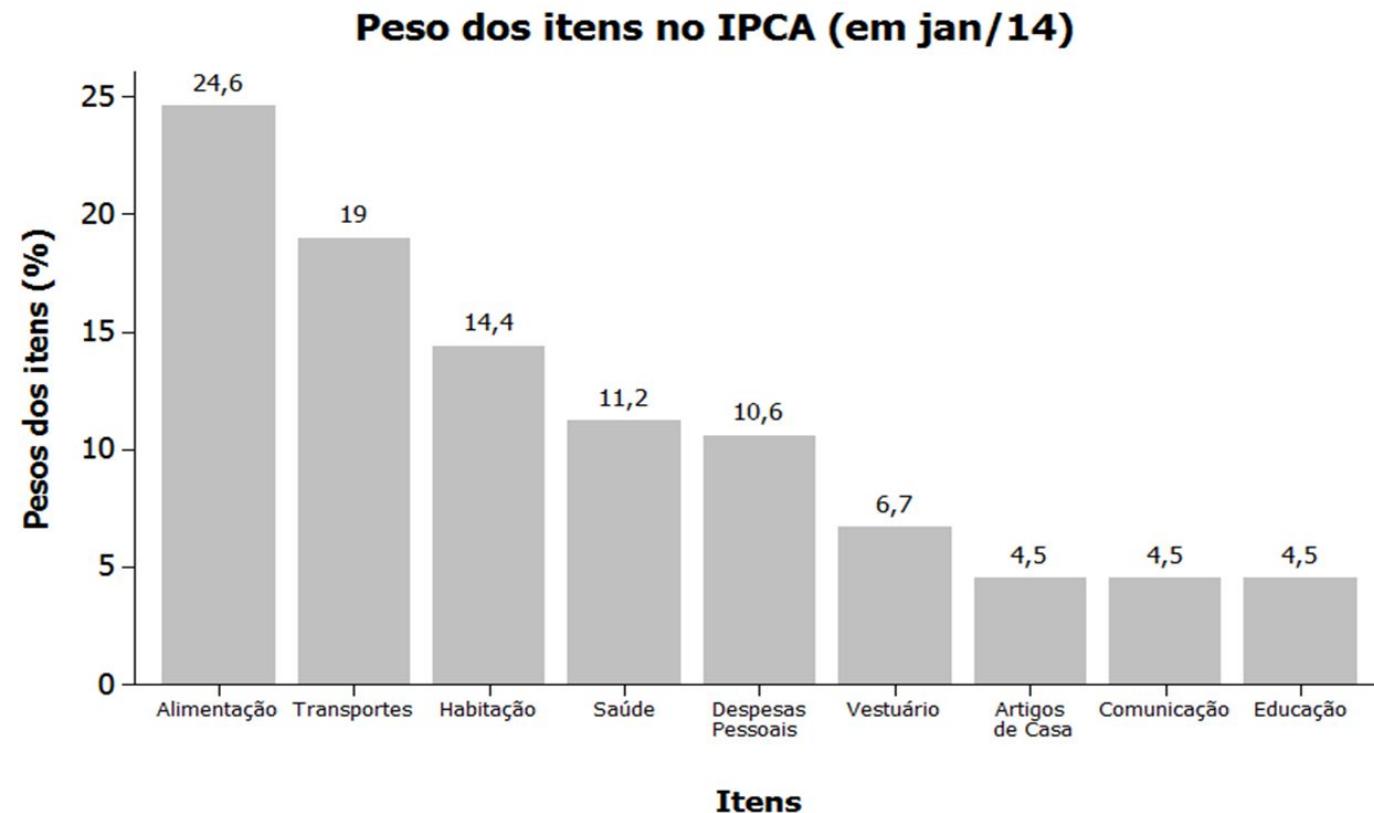
Fonte: <http://www.atireiopaunografico.com.br/2014/03/grafico-de-pizza-usar-ou-nao.html>

Análise de dados através de gráficos

IGTI

Gráfico de Setores - CUIDADOS

Exibe a proporção de cada nível categórico de uma variável



Fonte: <http://www.atireiopaunografico.com.br/2014/03/grafico-de-pizza-usar-ou-nao.html>

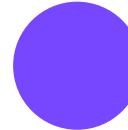
Conclusão



- ✓ Tipos de variáveis.
- ✓ Histograma.
- ✓ Boxplot.
- ✓ Gráfico de Linhas.
- ✓ Gráfico de Dispersão.
- ✓ Gráfico de Setores.



Próxiam aula



- A ferramenta R.

Análise Estatística de Dados

AULA 1.4. A FERRAMENTA R

PROF. MÁIRON CHAVES

Nesta aula

- Ferramentas e Linguagens para trabalhar com Estatística.
- O que é a Linguagem R?
- RStudio.
- Principais Tipos de Objetos no R.
- Bibliotecas relevantes no R.

Ferramentas e linguagens para trabalhar com Estatística

IGTI

Gratuitas



Pagas



Linguagem R

IGTI

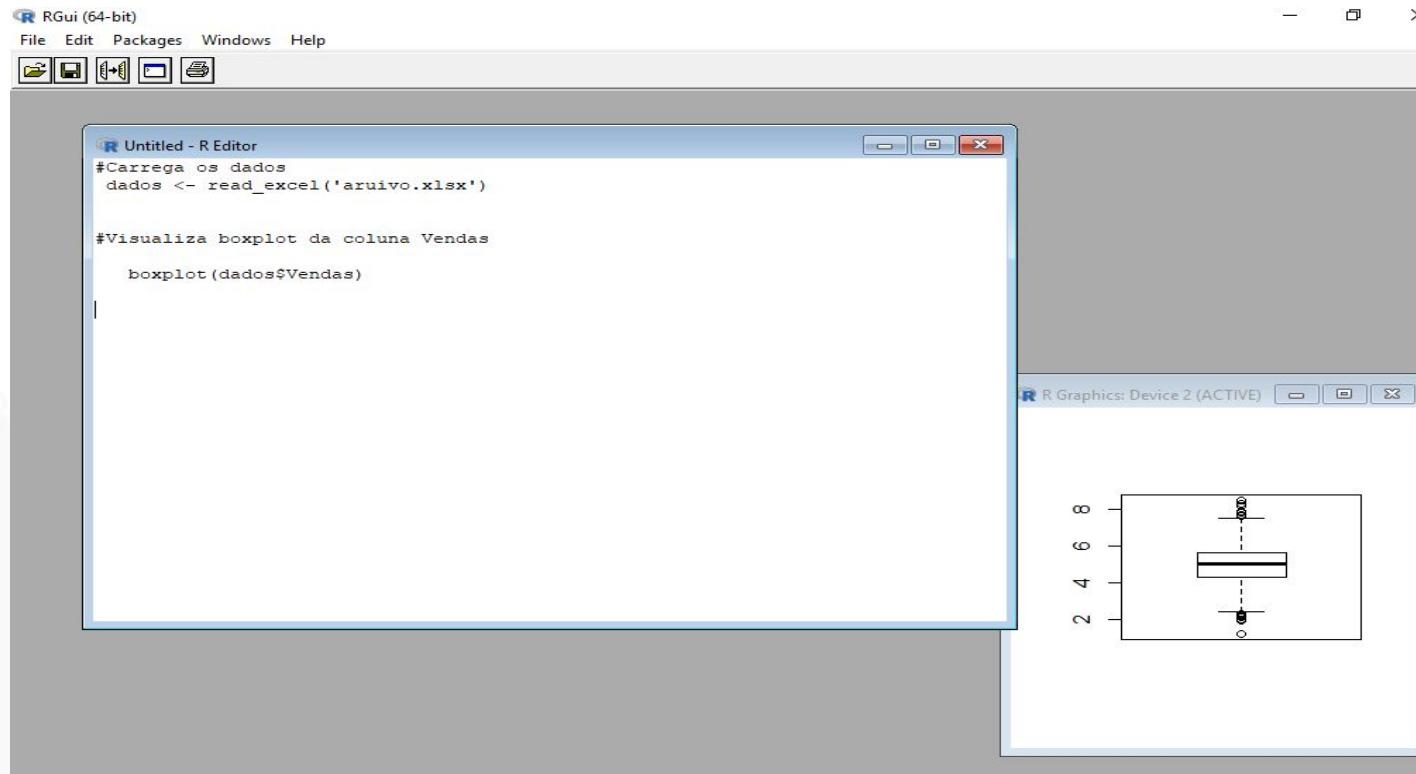
- R é uma linguagem e também um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos.
- Foi criada originalmente por Ross Ihaka e por Robert Gentleman, no departamento de Estatística da Universidade de Auckland, Nova Zelândia.
- É utilizado por profissionais em diversas áreas, como ciências sociais, saúde, psicologia, computação, dentre outras.
- É gratuito.



Linguagem R

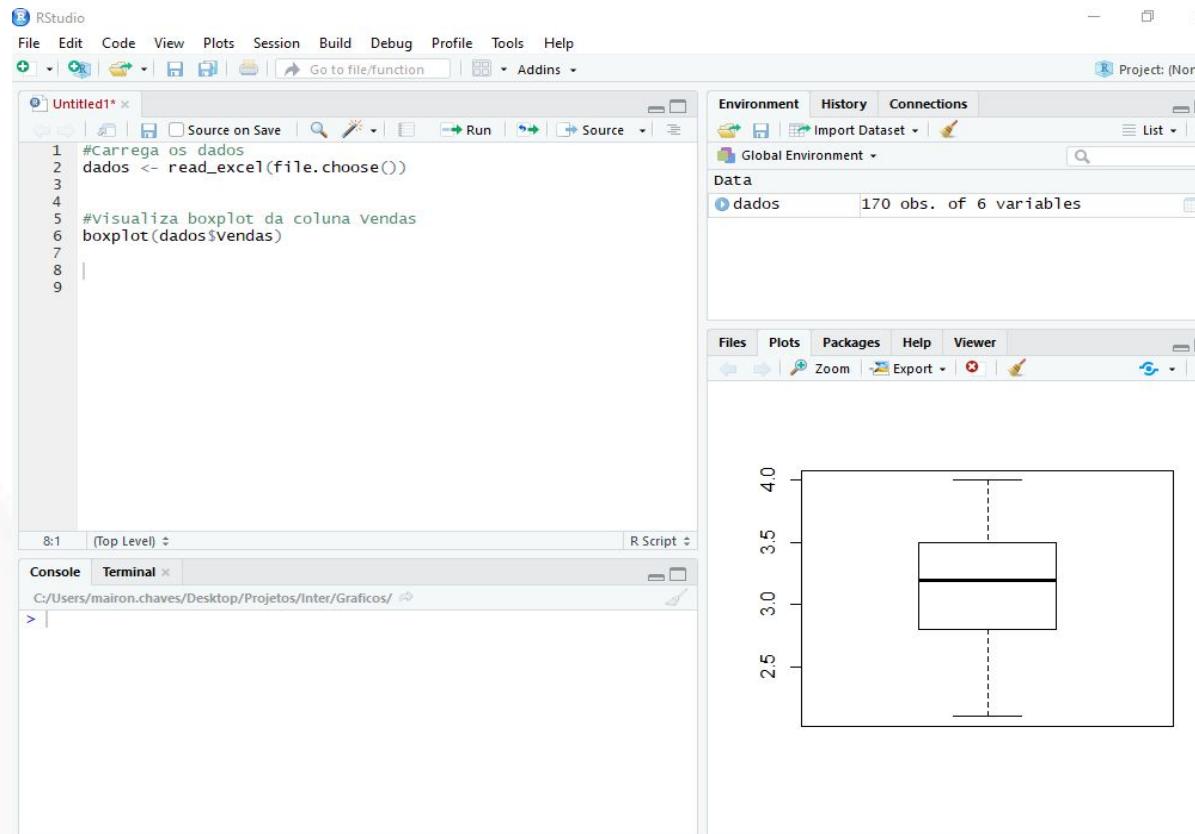
Interface do R

IGTI



Linguagem R

Interface do RStudio



Linguagem R

Principais tipos de Objetos

Vetor

Preço
500,00
340,00
177,00
308,00

Matriz

Preço	Lucro
500,00	123,12
340,00	201,00
177,00	78,00
308,00	234,20

Data Frame

Preço	Lucro	Categoria do Cliente	Data da Compra
500,00	123,12	A	01/05/2019
340,00	201,00	B	01/05/2019
177,00	78,00	A	02/05/2019
308,00	234,20	C	03/05/2019

Lista

Linguagem R

Principais Tipos de Objetos

Numeric - Inteiro ou Decimal (int ou float).

Character – Texto (string).

Date Time – Data, hora.

Factor – Atribui codificação inteira ao dado.

Character

Estado
MG
SP
RJ
GO
MG
SP
MG
SP
GO
GO
MG
SP

Factor

Estado_MG	Estado_SP	Estado_RJ
1	0	0
0	1	0
0	0	1
0	0	0
1	0	0
0	1	0
1	0	0
0	1	0
0	0	0
0	0	0
1	0	0
0	1	0



Linguagem R

Bibliotecas Relevantes no R – Visualização de Dados

IGTI



<https://www.r-graph-gallery.com/ggplot2-package.html>



plotly

<https://plotly-r.com/animating-views.html>

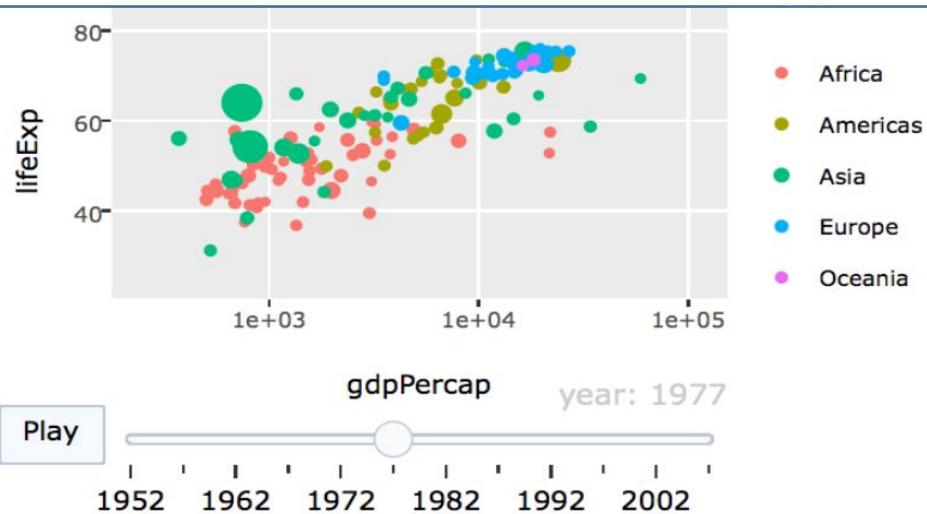


FIGURA 14.1: Animação da evolução da relação entre o PIB per capita e a expectativa de vida em vários países.

Linguagem R

Bibliotecas Relevantes no R – Machine Learning



Caret – Mais de 180 algoritmos de Machine Learning.

The screenshot shows a blog post titled "O `caret` pacote" by Max Kuhn, published on 27/03/2019. The post is an introduction to the `caret` package, which provides a unified interface for various machine learning algorithms. The sidebar on the left lists other posts related to data visualization, pre-processing, division of data, training, and adjusting models. The main content discusses the package's purpose, its features (like data division, preprocessing, and model selection), and its integration with other R packages. It also mentions the CRAN repository and GitHub.

O `caret` pacote

Max Kuhn
27/03/2019

1 Introdução

O `caret` pacote (curto para **C**lassificação **U**m **R**egressão **T**chover) é um conjunto de funções que a tentativa para simplificar o processo para a criação de modelos preditivos. O pacote contém ferramentas para:

- divisão de dados
- pré-processando
- seleção de recursos
- ajuste de modelo usando reamostragem
- estimativa de importância variável

bem como outras funcionalidades.

Existem muitas funções de modelagem diferentes em R. Algumas têm sintaxe diferente para treinamento e / ou previsão de modelo. O pacote começou como uma forma de fornecer uma interface uniforme às próprias funções, bem como uma forma de padronizar tarefas comuns (como ajuste de parâmetros e importância variável).

A versão de lançamento atual pode ser encontrada no [CRAN](#) e o projeto está hospedado [no github](#).

Alguns recursos:

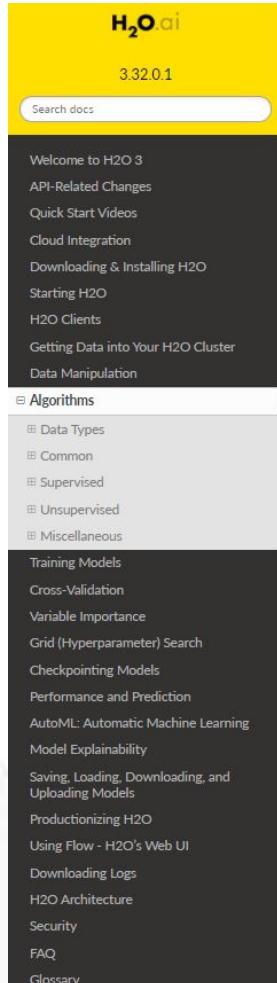
- O livro *Applied Predictive Modeling* apresenta [ácento circunflexo](#) e mais de 40 outros pacotes R. Está à venda na [Amazon](#) ou no [site da editora](#). Também existe um [site complementar](#).

Fonte: <https://topepo.github.io/caret/>

Linguagem R

Bibliotecas Relevantes no R – Machine Learning

H2O



• Supported Data Types

Common

- Quantiles
- Early Stopping

Supervised

In supervised learning, the dataset is labeled with the answer that algorithm should come up with. Supervised learning takes input variables (x) along with an output variable (y). The output variable represents the column that you want to predict on. The algorithm then uses these variables to learn and approximate the mapping function from the input to the output. Supervised learning algorithms support classification and regression problems.

H2O supports the following supervised algorithms:

- AutoML: Automatic Machine Learning
- Cox Proportional Hazards (CoxPH)
- Deep Learning (Neural Networks)
- Distributed Random Forest (DRF)
- Generalized Linear Model (GLM)
- Generalized Additive Models (GAM)
- Gradient Boosting Machine (GBM)
- Naïve Bayes Classifier
- RuleFit
- Stacked Ensembles
- Support Vector Machine (SVM)
- XGBoost

Unsupervised

In unsupervised learning, the model is provided with a dataset that isn't labeled - i.e., without an explicit outcome that the algorithm should return. In this case, the algorithm attempts to find patterns and structure in the data by extracting useful features. The model organizes the data in different ways, depending on the algorithm (clustering, anomaly detection, autoencoders, etc).

H2O supports the following unsupervised algorithms:

- Aggregator
- Generalized Low Rank Models (GLRM)
- Isolation Forest
- K-Means Clustering
- Principal Component Analysis (PCA)

Fonte: <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science.html>

IGTI

Linguagem R

Bibliotecas Relevantes no R – Manipulação de Dados

IGTI



dplyr

<https://blog.rstudio.com/2015/01/09/dplyr-0-4-0/>



Data.table

<https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html>

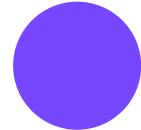
Conclusão



- ✓ Ferramentas para trabalhar com Estatística.
- ✓ R e RStudio.
- ✓ Principais objetos do R.
- ✓ Bibliotecas relevantes no R.



Próxima aula



- Estatística Computacional – Análise Exploratória de Dados com o R.

Análise Estatística de Dados

AULA 1.5. ESTATÍSTICA COMPUTACIONAL – ANÁLISE EXPLORATÓRIA DE DADOS COM O R

PROF. MÁIRON CHAVES

Nesta aula



- Análise Exploratória de Dados com o R.

Análise Exploratória de Dados com o R



Estatística Computacional – Análise Exploratória de Dados com o R

```
##### Análise exploratória de dados #####
```

```
## AED - Capítulo 01 - Prof. Máiron Chaves ##
```

```
#Copie este código, cole no seu R e execute para ver os resultados
```

```
rm(list=ls(all=TRUE)) #Remove objetos da memória do R
```

```
#Cria o data frame contendo o histórico de vendas do café
```

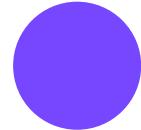
```
dados <- data.frame(Vendas_Café = c(18, 20, 23, 23, 23, 23, 24, 25, 26, 26, 26,  
27, 28, 28,
```

```
29, 29, 30, 30, 31, 31, 33, 34, 35, 38, 39, 41, 44, 44, 46),
```

Conclusão



- ✓ Medidas de Centralidade e Dispersão.
- ✓ Análise gráfica.
- ✓ Customizando gráficos.



Próxima aula



- Leis de Probabilidade e Diretrizes para sua Aplicação.

Análise Estatística de Dados

CAPÍTULO 2. LEIS DE PROBABILIDADE E DIRETRIZES PARA SUA APLICAÇÃO

PROF. MÁIRON CHAVES

Análise Estatística de Dados

AULA 2.1. LEIS DE PROBABILIDADE E DIRETRIZES PARA SUA APLICAÇÃO

PROF. MÁIRON CHAVES

Nesta aula

- Eventos aleatórios.
- Probabilidade Frequentista.
- Regra Aditiva.
- Regra Multiplicativa.

Eventos aleatórios



IGTI

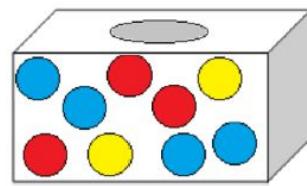
Um experimento aleatório é aquele cujo resultado é incerto, embora se saiba quais são os resultados possíveis.



$$P(\text{face}) = 1/2, \text{ face} = \text{cara ou coroa}.$$



$$P(\text{face}) = 1/6, \text{ face} = 1, 2, 3, 4, 5 \text{ ou } 6.$$



$$P(\text{cor}) = \begin{cases} 2/9, & \text{cor} = \text{amarela}; \\ 4/9, & \text{cor} = \text{azul}; \\ 3/9, & \text{cor} = \text{vermelha}. \end{cases}$$

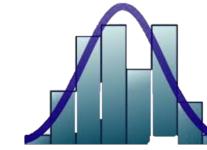
De tal forma que:

$$P(\text{Sucesso}) + P(\text{Fracasso}) = P(1) + P(0) = 1$$

Ou seja:

$$P(\text{Fracasso}) = 1 - P(\text{Sucesso})$$

Probabilidade Frequentista



Sendo A um evento aleatório qualquer, podemos encontrar a probabilidade de A, utilizando a probabilidade frequentista.

$$P(A) = \frac{\text{Número de Vezes que o evento } A \text{ ocorreu}}{\text{Número total de observações}}$$

$$P(\text{Sucesso}) + P(\text{Fracasso}) = P(1) + P(0) = 1$$

De tal forma que:

$$0 \leq P(A) \leq 1$$

Regra Aditiva

Operador **OU** – Probabilidade de um ou outro evento aleatório ocorrer.

- Uma sobremesa possui cobertura de menta.
- Duas sobremesas possuem cobertura de chocolate.
- Três sobremesas possuem cobertura de morango.
- Uma sobremesa possui cobertura de chocolate e cobertura de morango.
- Três sobremesas possuem cobertura de baunilha.



Regra Aditiva

Operador **OU** – Probabilidade de um ou outro evento aleatório ocorrer.



Qual a probabilidade do cliente receber uma sobremesa com cobertura de menta ou uma sobremesa com cobertura de chocolate?

- Uma sobremesa possui cobertura de menta.
- Duas sobremesas possuem cobertura de chocolate.
- Três sobremesas possuem cobertura de morango.
- Uma sobremesa possui cobertura de chocolate e cobertura de morango.
- Três sobremesas possuem cobertura de baunilha.

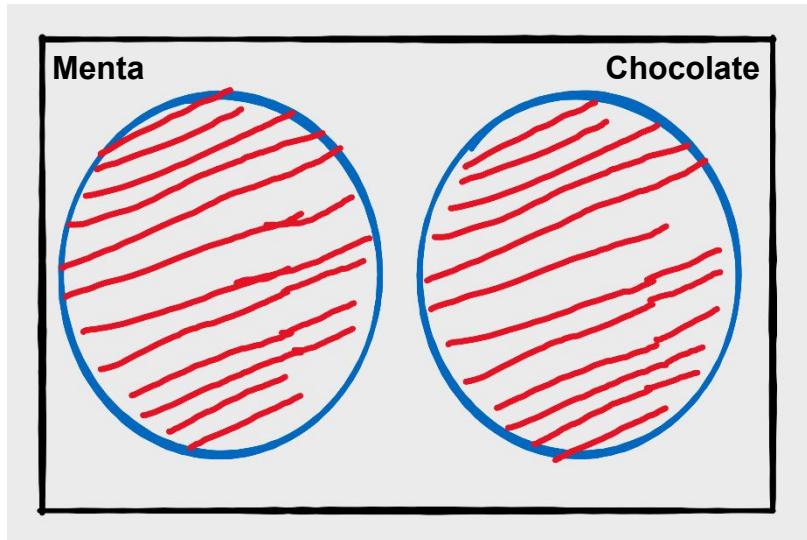
Uma possui cobertura de menta, portanto a probabilidade do cliente receber aleatoriamente uma sobremesa com cobertura de menta, é $\frac{1}{10}$ (ou 10%).

Três sobremesas possuem cobertura de chocolate, portanto, a probabilidade do cliente receber aleatoriamente uma sobremesa com cobertura de chocolate é de $\frac{3}{10}$ (ou 30%).

Regra Aditiva

Operador **OU** – Probabilidade de um ou outro evento aleatório ocorrer.

Eventos mutuamente exclusivos



$$P(Menta \cap Chocolate) = \emptyset$$

Dois eventos são eventos mutuamente exclusivos se eles não podem ocorrer ao mesmo tempo.

Um exemplo disso é o lançamento de uma moeda, o qual pode resultar em cara ou coroa, mas não ambos.

Regra Aditiva

Operador **OU** – Probabilidade de um ou outro evento aleatório ocorrer.

$$P(\text{Menta}) \text{ ou } P(\text{Chocolate}) =$$

$$P(\text{Menta}) + P(\text{Chocolate}) =$$

$$\frac{1}{10} + \frac{3}{10} = \frac{4}{10} \text{ ou } 40\%$$

Regra Aditiva

Operador **OU** – Probabilidade de um ou outro evento aleatório ocorrer (evento não mutuamente exclusivo).

Qual a probabilidade do cliente receber, aleatoriamente, uma sobremesa com cobertura de chocolate ou uma sobremesa com cobertura de morango? Veja que existe uma sobremesa que vai com as **duas coberturas**.

- Uma sobremesa possui cobertura de menta.
- Duas sobremesas possuem cobertura de chocolate.
- Três sobremesas possuem cobertura de morango.
- Uma sobremesa possui cobertura de chocolate e cobertura de morango.
- Três sobremesas possuem cobertura de baunilha.

Duas sobremesas que possuem cobertura de chocolate (apenas chocolate), portanto, a probabilidade do cliente receber aleatoriamente uma sobremesa com cobertura de chocolate é de $\frac{2}{10}$ (ou 20%).

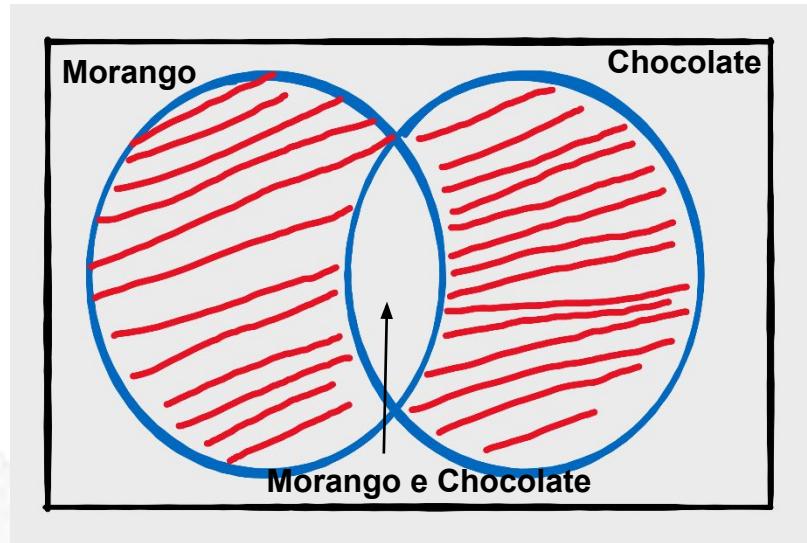
Três possuem cobertura de morango(apenas morango), portanto a probabilidade do cliente receber aleatoriamente uma sobremesa com cobertura de morango é $\frac{3}{10}$ (ou 30%).

Uma sobremesa que possui cobertura de chocolate e de morango ao mesmo tempo, portanto, a probabilidade do cliente receber aleatoriamente uma sobremesa com cobertura de chocolate e morango é de $\frac{1}{10}$ (ou 10%).

Regra Aditiva

Operador **OU** – Probabilidade de um ou outro evento aleatório ocorrer.

Eventos não mutuamente exclusivos



Dois eventos são não mutuamente exclusivos, se eles podem ocorrer ao mesmo tempo.

Regra Aditiva

Operador **OU** – Probabilidade de um ou outro evento aleatório ocorrer.

$$P(\text{Morango}) \text{ ou } P(\text{Chocolate}) =$$

$$P(\text{Morango}) + P(\text{Chocolate}) - P(\text{Morango e Chocolate})$$

$$P((\text{Morango}) \text{ ou } (\text{Chocolate})) = \frac{3}{10} + \frac{2}{10} - \frac{1}{10} = \frac{4}{10} \text{ ou } 40\%$$

Regra Multiplicativa

Operador **E** – Probabilidade de um e outro evento aleatório ocorrer.

- Expresso.
- Cappuccino.



Regra Multiplicativa

Operador **E** – Probabilidade de um **e** outro evento aleatório ocorrer.

Dois eventos são INDEPENDENTES quando um não altera a probabilidade de outro ocorrer.

- Uma sobremesa possui cobertura de menta.
- Duas sobremesas possuem cobertura de chocolate.
- Três sobremesas possuem cobertura de morango.
- Uma sobremesa possui cobertura de chocolate e cobertura de morango.
- Três sobremesas possuem cobertura de baunilha.



- Espresso
- Cappuccino



Qual a probabilidade do cliente receber aleatoriamente uma sobremesa com cobertura de menta e um café expresso?

$$P(\text{Menta}) \text{ e } P(\text{Expresso}) =$$

$$P(\text{Menta}) * P(\text{Expresso}) =$$

$$\frac{1}{2} * \frac{1}{10} = \frac{1}{20} (\text{ou } 5\%)$$

Regra Multiplicativa



Operador **E** – Probabilidade de um **e** outro evento aleatório ocorrer.

Dois eventos são DEPENDENTES quando um altera a probabilidade de outro ocorrer.

Dado que o cliente já recebeu uma sobremesa com cobertura de menta, qual a probabilidade da próxima sobremesa sorteada ser de cobertura de baunilha?

- Uma sobremesa possui cobertura de menta.
- Duas sobremesas possuem cobertura de chocolate.
- Três sobremesas possuem cobertura de morango.
- Uma sobremesa possui cobertura de chocolate e cobertura de morango.
- Três sobremesas possuem cobertura de baunilha

$$P(\text{menta}) = \frac{1}{10}$$

$$P(\text{baunilha} | \text{menta}) = \frac{3}{9}$$

$$P(\text{baunilha} | \text{menta}) = P(\text{menta}) * P(\text{baunilha} | \text{menta}) = \frac{1}{10} * \frac{3}{9} = \frac{3}{90} (0,03 \text{ ou } 3\%)$$

Regra Multiplicativa

IGTI

Algoritmo Naive Bayes – Machine Learning

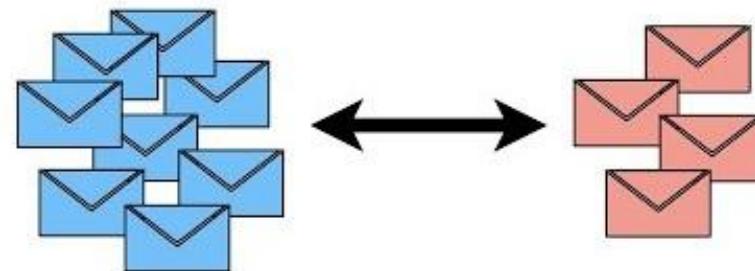
Naïve Bayes Classifier

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



Thomas Bayes
1702 - 1761

Naive Bayes....



...Clearly Explained!!!

Fonte: <https://www.youtube.com/watch?v=O2L2Uv9pdDA>

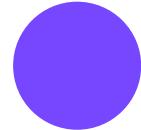
Connclusão



- ✓ Probabilidade Frequentista.
- ✓ Eventos Aleatórios.
- ✓ Regras Aditiva (Eventos mutuamente exclusivos).
- ✓ Regras Aditiva (Eventos não mutuamente exclusivos).
- ✓ Regras Multiplicativa (Eventos independentes).
- ✓ Regras Multiplicativa (Eventos dependentes e probabilidade condicional).



Próxima aula



- Variáveis Aleatórias Discretas.
- Distribuições de Probabilidades Discretas.

Análise Estatística de Dados

AULA 2.2. VARIÁVEIS ALEATÓRIAS E DISTRIBUIÇÕES DE PROBABILIDADES DISCRETAS

PROF. MÁIRON CHAVES

Nesta aula



- Variáveis Aleatórias Discretas.
- Distribuições Discretas.

Variáveis Aleatórias

Variável Aleatória Discreta – Assume um número infinito contável de valores.



Exemplos:

- Quantidade de ligações por dia que um call center recebe.
- Quantidade de clientes por hora que entram em uma loja.



Variáveis Aleatórias

Distribuições Discretas

Experimento Bernoulli - Número de sucessos em n tentativas.

$$p(\text{sucesso}) = p(1) = p$$

$$p(\text{fracasso}) = p(0) = 1 - p$$

Onde:

p é a probabilidade do sucesso ocorrer

Exemplos:

- Jogada de uma moeda observando se deu cara ou coroa.
- Jogada de um dado e observação da face de cima.



Variáveis Aleatórias

Distribuições Discretas

Experimento Bernoulli - Número de sucessos em n tentativas.

$$p(\text{sucesso}) = p(1) = p$$

$$p(\text{fracasso}) = p(0) = 1 - p$$

Onde:

p é a probabilidade do sucesso ocorrer.

Exemplos:

- Jogada de uma moeda observando se deu cara ou coroa.
- Jogada de um dado e observação da face de cima.



Variáveis Aleatórias

Distribuições Discretas

Análise Combinatória:

Se n e x são inteiros positivos ($n \geq x$), então combinações de n elementos x a x , é:



Variáveis Aleatórias

Distribuições Discretas

Distribuição Geométrica – Número de tentativas até o primeiro sucesso.

$$f(x) = (1 - p)^{x-1} \cdot p$$

Onde:

x é o número de tentativas.

p é a probabilidade de sucesso.

Exemplo: Sabendo que a probabilidade de sucesso é 50%(p),
qual a probabilidade da primeira venda ocorrer quando o quinto(x) cliente entrar na loja?

→ $f(x) = 0,0156$

Variáveis Aleatórias

Distribuições Discretas

Distribuição Binomial Negativa – É uma generalização da distribuição geométrica.

Número de tentativas até que uma quantidade de sucessos ocorra.

$$f(x) = \binom{x-1}{r-1} (1-p)^{x-r} \cdot p^r$$

Onde:

r é a quantidade de sucessos.

x é o número de tentativas.

p é a probabilidade de sucesso.

Exemplo: Sabendo que a probabilidade de sucesso é 50%(p),

qual a probabilidade de ter que entrar 8(x) clientes até que a segunda(r) venda ocorra?

$$f(x) = \binom{8-1}{2-1} (1 - 0,5)^{8-2} \cdot 0,5^2 \quad \rightarrow$$

Variáveis Aleatórias

Distribuições Discretas



Distribuição de Poisson – Expressa a probabilidade de um evento ou uma série de eventos ocorrerem em um determinado período de tempo.

Onde:

$e = 2,71$.

$x!$ é o fatorial de número de vezes que o evento ocorre.

λ é o número de ocorrências que de um evento em um intervalo de tempo.

Exemplo a: Uma loja recebe em média, 6 (λ) clientes por minuto. Qual a probabilidade de que 5(x) clientes entrem em um minuto?



Variáveis Aleatórias

Distribuições Discretas

Exemplo b: Uma loja recebe em média, 6 (λ) clientes por minuto. Qual a probabilidade de que até 2(x) clientes entrem em um minuto?

$$P(x \leq 2) \rightarrow P(X=0) + P(X=1) + P(X=2)$$

$$0,002479 + 0,014873 + 0,044618 \rightarrow 0,0619$$

Exemplo c: Uma loja recebe em média, 6 (λ) clientes por minuto. Qual a probabilidade de que 6(x) clientes ou mais em um minuto?

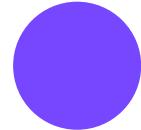
$$P(x \geq 3) \rightarrow 1 - P(X < 3) \rightarrow 1 - [P(X=0) + P(X=1) + P(X=2)] \rightarrow 1 - 0,0619 \rightarrow 0,9381$$

Conclusão



- ✓ Variável Aleatória Discreta.
- ✓ Experimento Bernoulli.
- ✓ Distribuição Binomial.
- ✓ Distribuição Geométrica.
- ✓ Distribuição Binomial Negativa.
- ✓ Distribuição de Poisson.

Próxima aula



- Variáveis Aleatórias Contínuas.
- Distribuições de Probabilidades Contínuas.

Análise Estatística de Dados

AULA 2.3. VARIÁVEIS ALEATÓRIAS E DISTRIBUIÇÕES DE PROBABILIDADES CONTÍNUAS

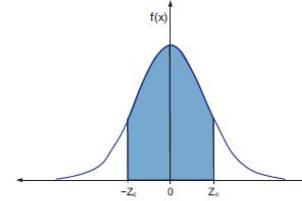
PROF. MÁIRON CHAVES

Nesta aula



- Variáveis aleatórias contínuas.
- Distribuições contínuas.

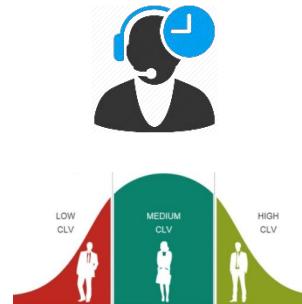
Variáveis Aleatórias



Variável Aleatória Contínua – Assume um número infinito incontável de valores.

Exemplos:

- **Tempo das ligações que um call center recebe por dia.**
- **Valor que os clientes compram em uma loja em reais R\$.**



Variáveis Aleatórias

Distribuições Contínuas

Distribuição Normal – É uma das distribuições mais importantes.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Onde:

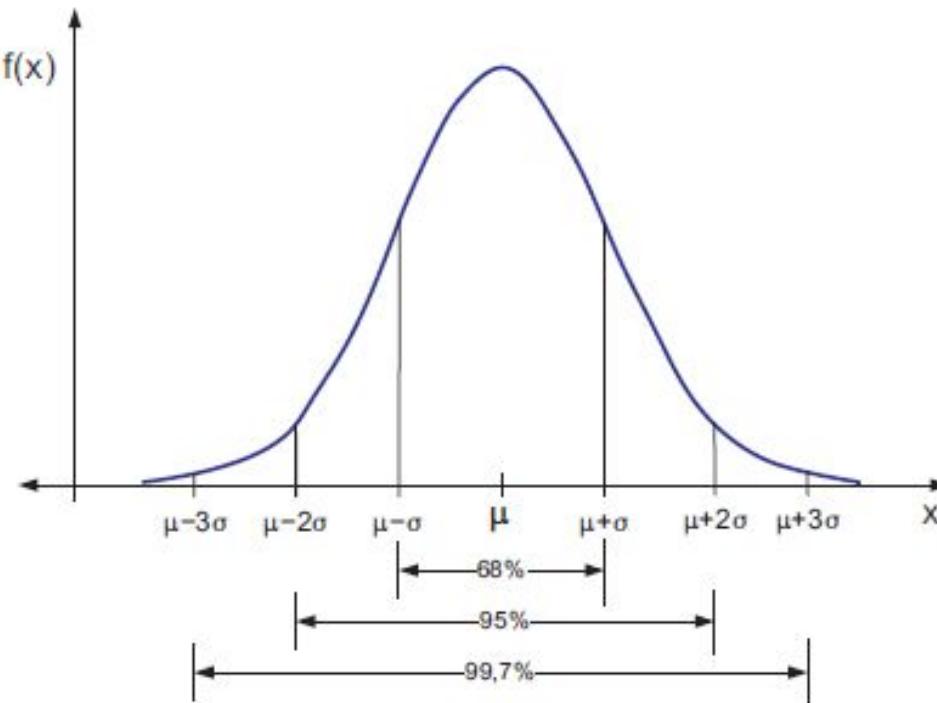
x é o valor da variável aleatória.

μ é a média.

σ é o desvio padrão.

$\pi = 3,14$.

$e = 2,71..$



Variáveis Aleatórias

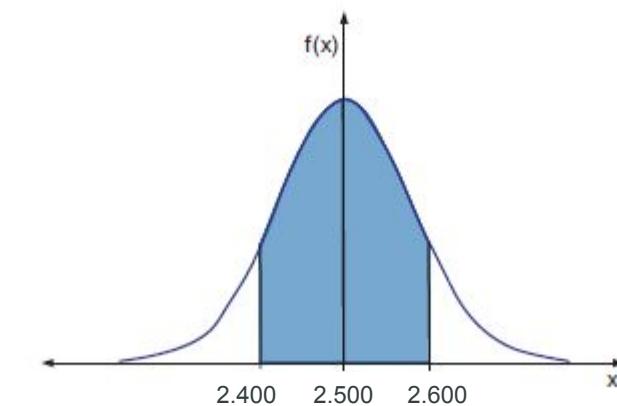
Distribuições Contínuas

Distribuição Normal

Exemplo: suponha que a distribuição dos salários dos funcionários sigam uma distribuição normal com média $\mu=2.500$ e desvio padrão $\sigma= 170$.

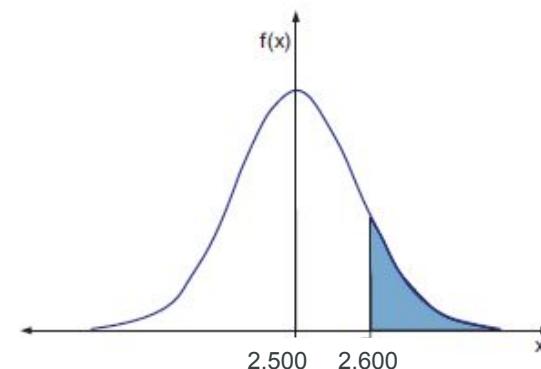
Dizemos que a variável $\text{salário} \sim N(\mu=2500, \sigma=170)$

Qual a probabilidade de um funcionário ter salário entre 2.400 e 2.600?



$$P(2.400 \leq x \geq 2.600) = 0,4436$$

Qual a probabilidade de um funcionário ter salário acima de 2.600?



$$P(x \geq 2.600) = 0,2781$$

Variáveis Aleatórias

Distribuições Contínuas

Distribuição Normal Padrão (distribuição z) – É um caso especial da distribuição normal.

Se uma variável aleatória segue uma distribuição normal, uma transformação (chamada de padronização) é aplicada de modo que essa variável tenha média zero e desvio padrão unitário. $\rightarrow X \sim Z(\mu=0, \sigma=1)$

Ainda no exemplo anterior da variável $\text{salário} \sim N(\mu=2500, \sigma=170)$, como ficaria o salário padronizado de um indivíduo que recebe 2.600?

$$z = \frac{(x - \mu)}{\sigma} \quad \rightarrow \quad z = \frac{(2600 - 2500)}{170} \quad \rightarrow \quad z = 0,5882$$

Variáveis Aleatórias

IGTI

Distribuições Contínuas

Distribuição Normal Padrão (distribuição z) – É um caso especial da distribuição normal.

$$z = 0,5882$$



:
=DIST.NORM.N(2600;2500;170;VERDADEIRO)
Ou
=DIST.NORM.P.N(0,5882;VERDADEIRO)



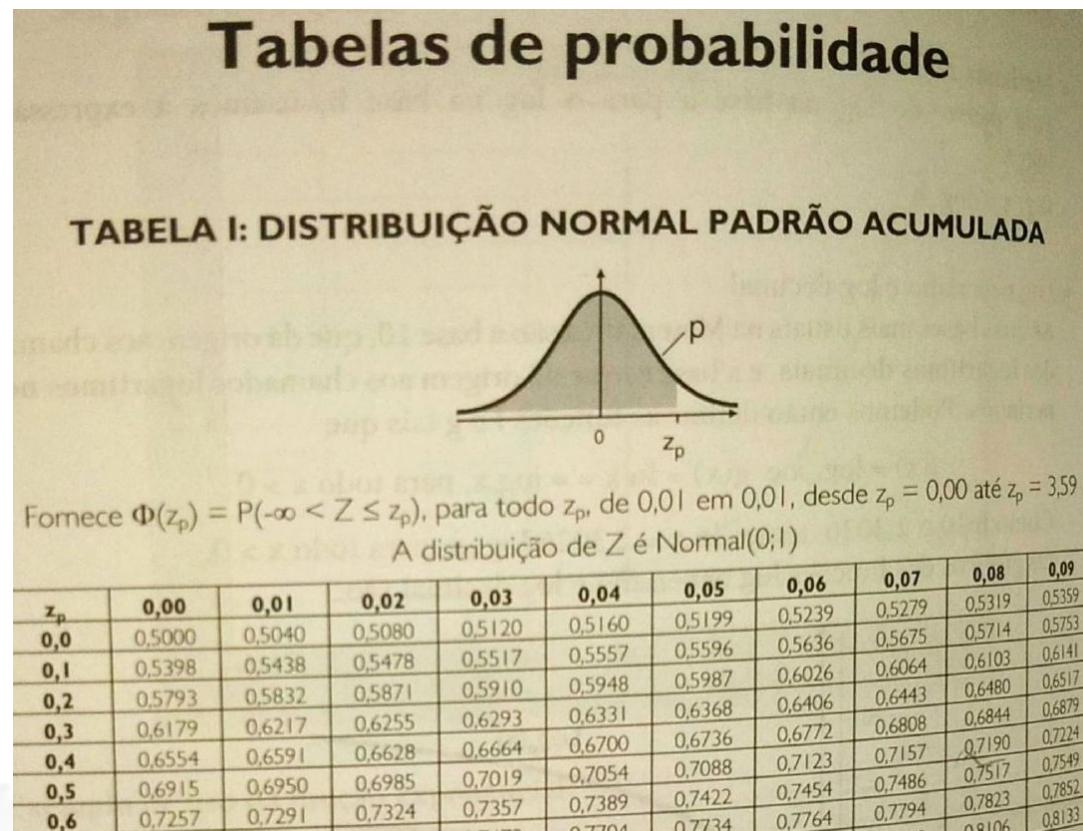
:
pnorm(2600 , 2500 , 170)
Ou
pnorm(0.52882)

Calculando a probabilidade complementar no R:

$$P(\text{Salário} < 2600) = \text{pnorm}(2600 , 2500 , 170) = 0,7218$$

Ou seja:

$$P(\text{Salário} \geq 2600) = 1 - 0,7218 = 0,2781$$



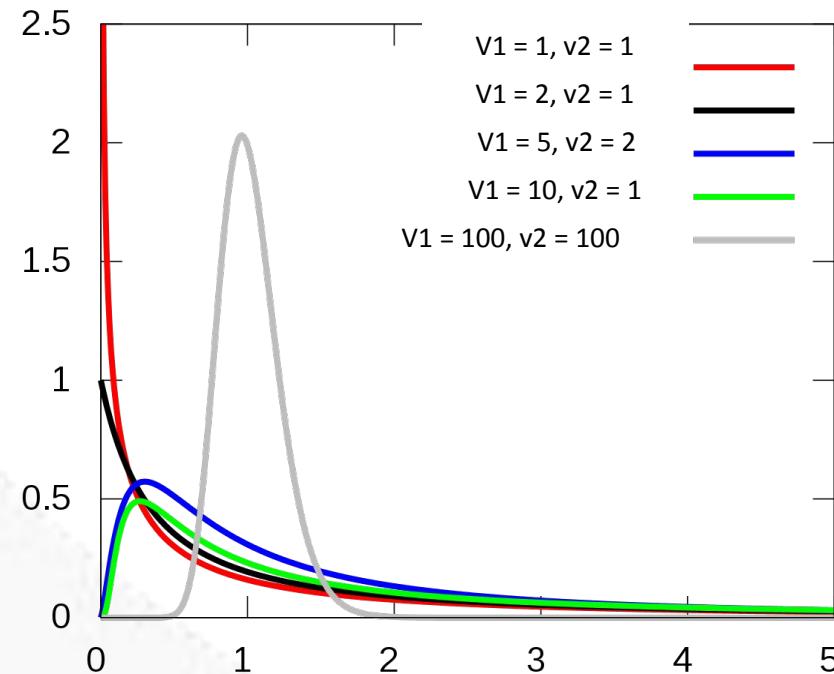
Variáveis Aleatórias

Distribuições Contínuas

Distribuição F de Fisher-snedecor

É uma distribuição positivamente assimétrica, não admite valores negativos. Geralmente utilizada para testar variâncias e depende de dois parâmetros chamados de graus de liberdade.

Se uma variável aleatória X segue uma distribuição F com v_1 e v_2 graus liberdades, então dizemos que: $\rightarrow X \sim F(v_1, v_2)$



Variáveis Aleatórias

Distribuições Contínuas

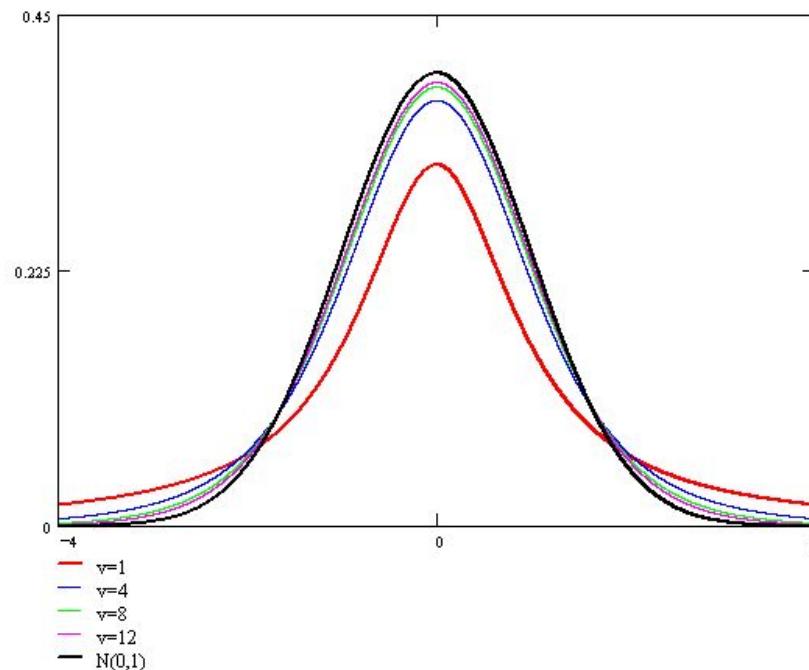
Distribuição t de Student

É simétrica e semelhante à curva normal padrão. Depende de um único parâmetro, que também é um grau de liberdade. É utilizada para testar médias.



Se uma variável aleatória X segue uma distribuição t com v grau de liberdade, então

dizemos que: $X \sim t(v)$



Variáveis Aleatórias

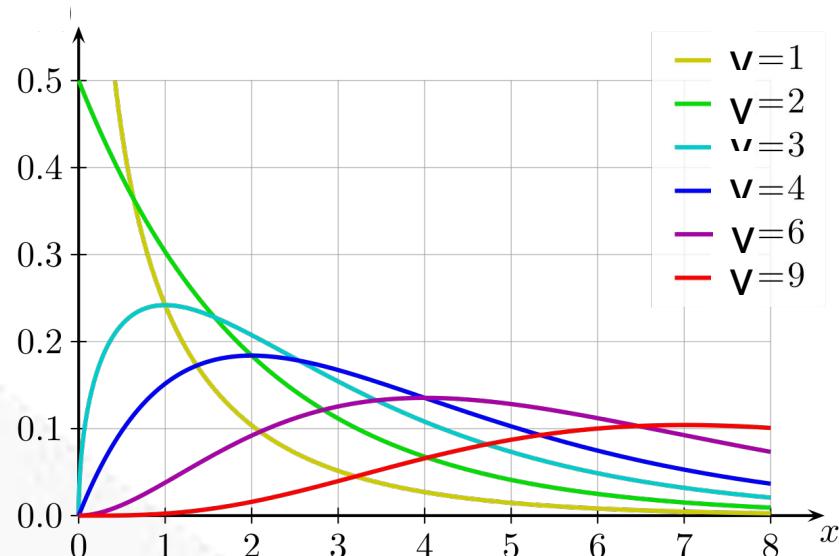
Distribuições Contínuas

Distribuição Qui-Quadrado

O quadrado de uma v.a. com distribuição normal padrão é uma Qui-Quadrado. Frequentemente utilizado para testar associação entre variáveis categóricas.

Dizemos que v.a. contínua, com valores positivos, tem uma distribuição qui-quadrado com v graus de liberdade.

$$\Rightarrow X \sim \chi^2(v)$$



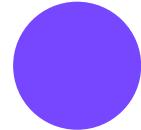
Conclusão



- ✓ Variável Aleatória Contínua.
- ✓ Distribuição Normal.
- ✓ Distribuição Normal Padrão.
- ✓ Distribuição F.
- ✓ Distribuição t de Student.
- ✓ Distribuição Qui-Quadrado.



Próxima aula



- Estatística Computacional – Probabilidades com R.

Análise Estatística de Dados

AULA 2.4. ESTATÍSTICA COMPUTACIONAL – PROBABILIDADES COM O R

PROF. MÁIRON CHAVES

Nesta aula



- Probabilidades com R.

Probabilidades com o R



Estatística Computacional – Probabilidades com o R

```
#####
```

```
##### Distribuições de Probabilidades #####
```

```
## AED - Capítulo 02 - Prof. Máiron Chaves ####
```

```
#####
```

```
#Copie este código, cole no seu R e execute para ver os resultados
```

```
#####
```

```
#### DISTRIBUIÇÃO BINOMIAL ####
```

```
#####
```

```
# Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade  
de sucesso é 50%.
```

```
# Ao passar 10 clientes em nossa loja, qual a probabilidade de realizarmos 2 vendas?
```

```
#Ou seja, queremos encontrar a probabilidade de dois sucessos, em dez tentativas.
```

```
Cuja probabilidade de sucesso
```

Conclusão



- ✓ Distribuições Discretas e Contínuas com o R.



Próxima aula



- Intervalos de Confiança e Teorema Central do Limite.

Análise Estatística de Dados

CAPÍTULO 3. INTERVALOS DE CONFIANÇA

PROF. MÁIRON CHAVES

Análise Estatística de Dados

AULA 3.1. INTERVALOS DE CONFIANÇA

PROF. MÁIRON CHAVES

Nesta aula



- Intervalo de confiança.
- Teorema Central do Limite.

Intervalo de Confiança

Estimativa Pontual vs. Estimativa Intervalar

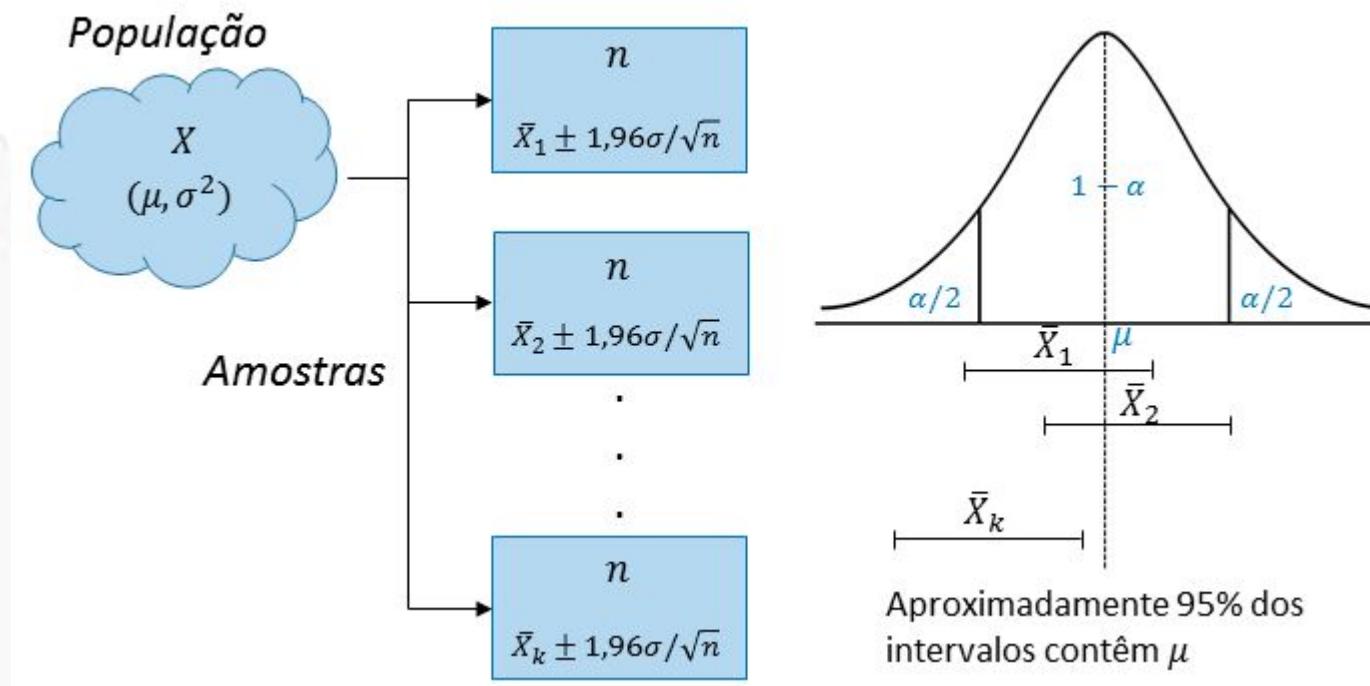


Intervalo de Confiança

Estimativa Pontual vs. Estimativa Intervalar

Temos 95% de “chance” do intervalo conter o verdadeiro valor da média populacional. Em outras palavras, se produzirmos diversos intervalos de confiança provenientes de diferentes amostras independentes de mesmo tamanho, podemos esperar que aproximadamente 95% destes intervalos devem conter o verdadeiro valor da média populacional.

iGTi



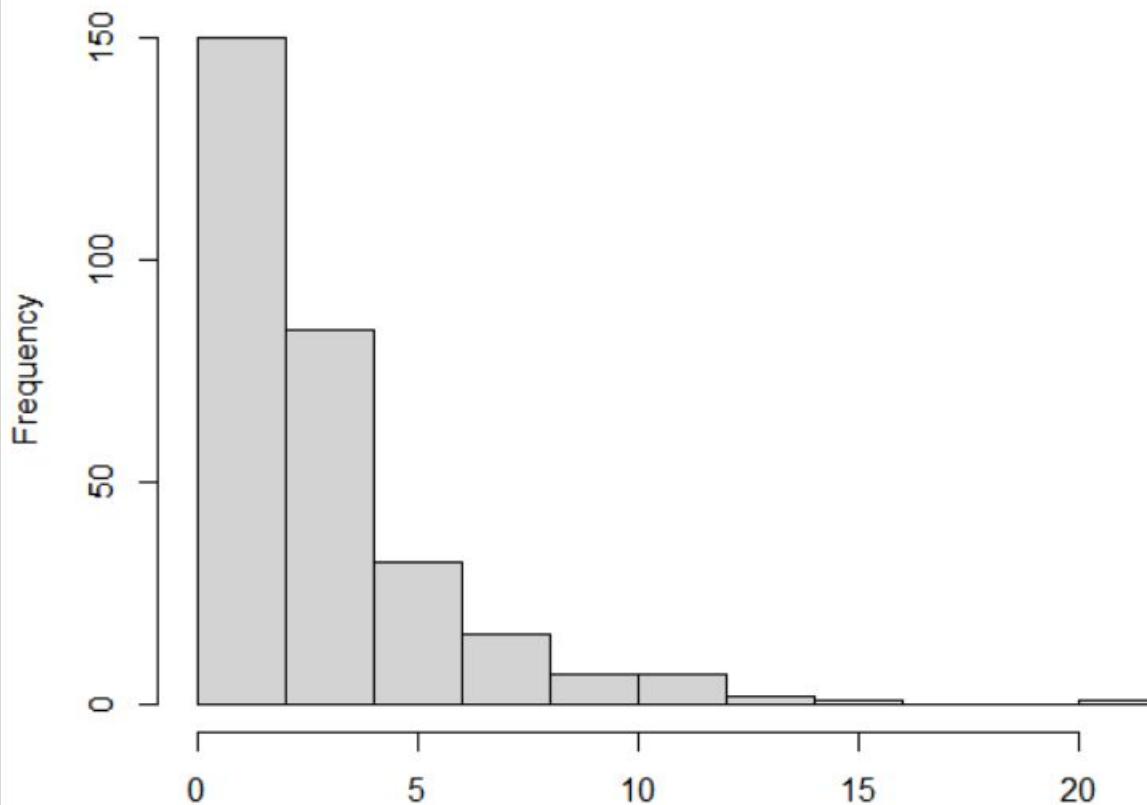
Teorema Central do Limite

O Teorema Central do Limite (TCL) afirma que independentemente de qual seja a distribuição original de X , a distribuição de \bar{x}_n se aproxima da distribuição Normal a medida que n cresce.

IGTI

V1
1 3.74485197
2 1.00887491
3 0.85788444
4 0.23472044
5 0.01121126
6 2.60001475
7 8.35325273
8 1.08187553
9 1.58768780
10 4.68492488
11 1.18704253
12 1.74471697
13 1.50570445
14 6.55945360
15 1.56107093
16 3.30260565
...
297 4.371387290
298 1.268781794
299 1.848694822
300 0.560306171

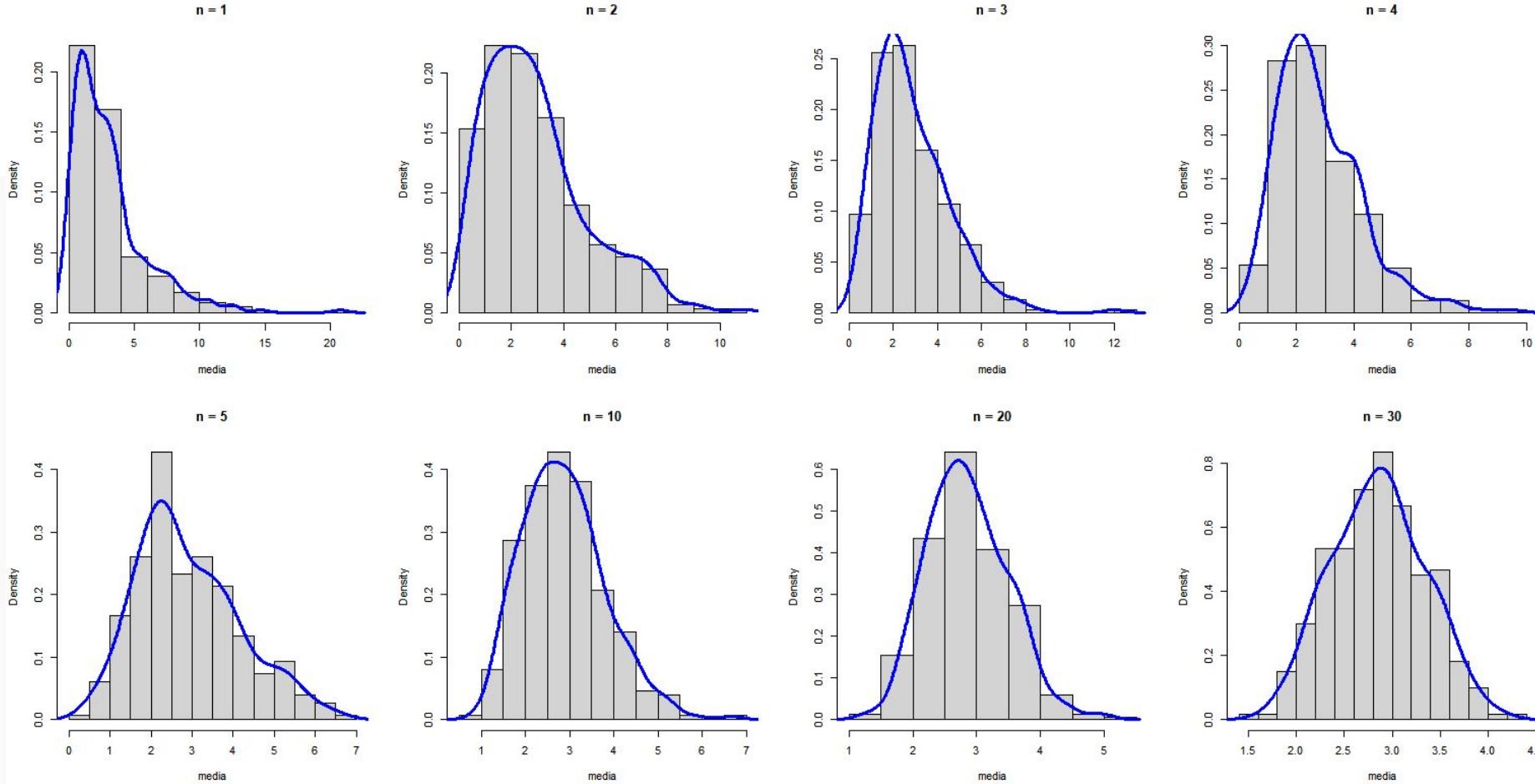
Histograma com 300 observações geradas aleatoriamente seguindo uma distribuição Exponencial



```
va <- rexp(300, rate = 1/3 )  
hist(va)
```

Teorema Central do Limite

O Teorema Central do Limite (TCL) afirma que independentemente de qual seja a distribuição original de X , a distribuição de \bar{x}_n se aproxima da distribuição Normal a medida que n cresce.



Teorema Central do Limite

Simulando o TCL no R.



R:

```
va <- rexp(300 , rate = 1/3 )
hist(va)

n <-1
media <- c()

for (i in 1:300 {

  subamostra <- sample(va, size = n,replace = F)

  media[i] <- mean(subamostra)

}

hist(media, main = paste('n = ',n),prob = T)
lines(density(media),lwd = 3,col = 'blue')
```

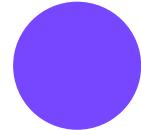
Conclusão



- ✓ Estimativa Pontual vs. Estimativa Intervalar.
- ✓ Teorema Central do Limite (TCL).



Próxima aula



- Intervalo de Confiança para Média.

Análise Estatística de Dados

AULA 3.2. INTERVALO DE CONFIANÇA PARA MÉDIA

PROF. MÁIRON CHAVES

Nesta aula



- ❑ Intervalo de Confiança para Média.

Intervalo de Confiança para Média

Vendas:

Intervalo de Confiança de 95% para essa média, utilizando a distribuição Z (normal padrão):



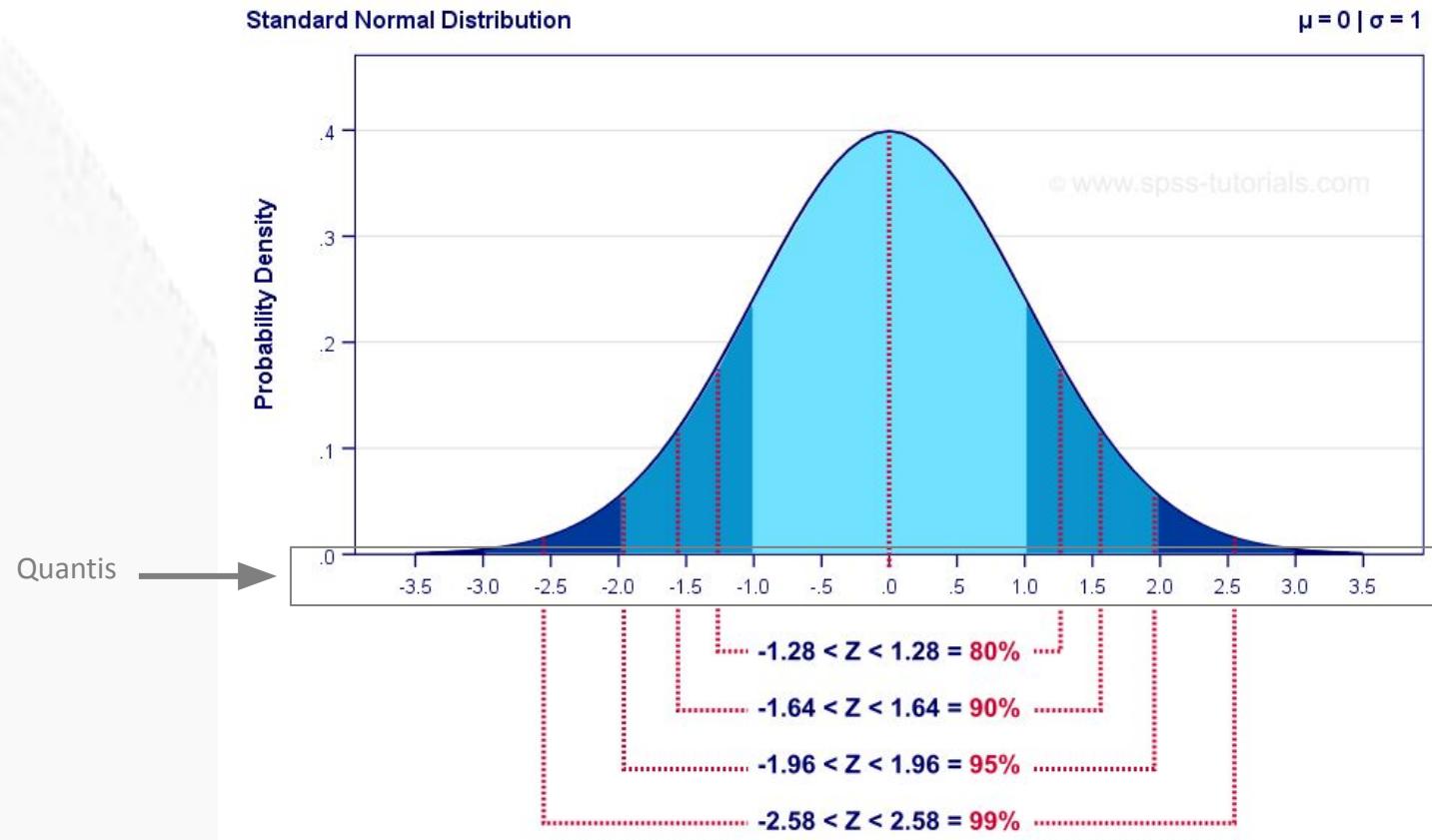
*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Intervalo de Confiança para Média

Quantis são pontos estabelecidos em intervalos regulares, a partir da função distribuição de probabilidade de uma variável aleatória.

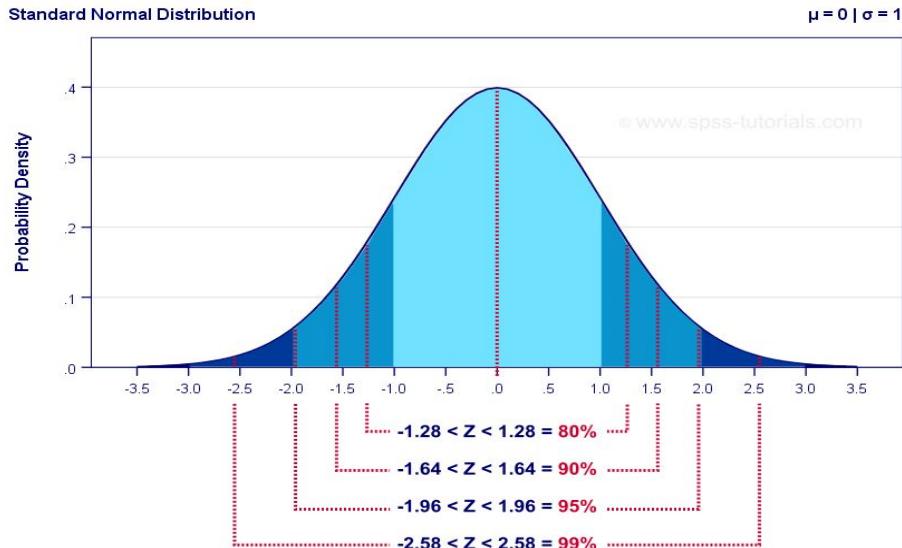
Os quantis dividem os dados ordenados em subconjuntos de dados de dimensão essencialmente igual.



*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Intervalo de Confiança para Média



R:

```
ic <- 0.99  
alfa <- 1-ic  
area <- 1-(alfa/2)  
qnorm(area) #0.995  
>2.5758
```

$$IC_{99\%} = 30 - 2,58 * \frac{7,31}{\sqrt{30}} < \mu > 30 + 2,58 * \frac{7,31}{\sqrt{30}}$$

$$IC_{99\%} = 30 - 3,44 < \mu > 30 + 3,44$$

$$IC_{99\%} = 26,55 < \mu > 33,44$$

R:

```
ic <- 0.80  
alfa <- 1 - ic  
area <- 1-(alfa/2)  
qnorm(area) #0.9  
>1,2815
```

$$IC_{80\%} = 30 - 1,28 * \frac{7,31}{\sqrt{30}} < \mu > 30 + 1,28 * \frac{7,31}{\sqrt{30}}$$

$$IC_{80\%} = 30 - 1,70 < \mu > 30 + 1,70$$

$$IC_{80\%} = 28,29 < \mu > 31,70$$

*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Intervalo de Confiança para Média



R:

```
ic <- 0.95  
alfa <- 1-ic  
area <- 1-(alfa/2)  
qt(area, df = n-1) #0.975  
>2,0452
```

Vendas:

*Esta área precisa ser preservada em todos os slides. É onde você aparecerá na transmissão.

Intervalo de Confiança de 95% para essa média, utilizando a distribuição t de Student:

Quantil da Distribuição t com n-1 graus de liberdade

$$IC_{95\%} = \bar{x} - t * \frac{s}{\sqrt{n}} < \mu > \bar{x} + t * \frac{s}{\sqrt{n}}$$

*Essa mensagem será retirada pela equipe de revisão.

$$IC_{95\%} = 30 - 2,0452 * \frac{7,31}{\sqrt{29}} < \mu > 2,0452 * \frac{7,31}{\sqrt{29}}$$

$$IC_{95\%} = 30 - 2,72 < \mu > 30 + 2,72$$

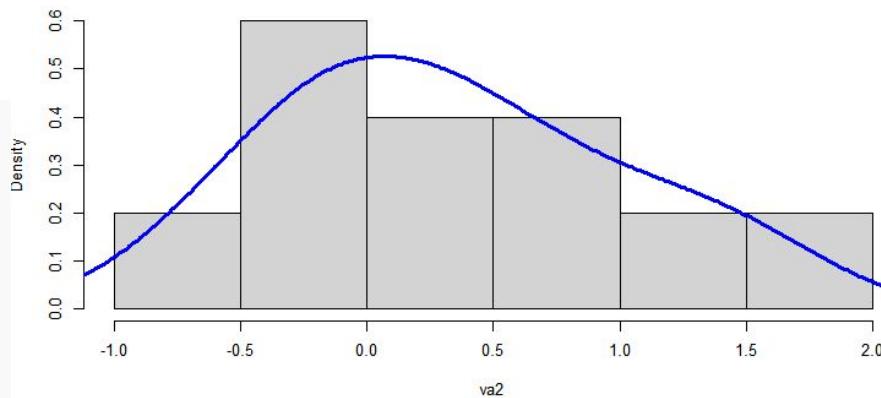
$$IC_{95\%} = 27,27 < \mu > 32,72$$

Intervalo de Confiança para Média

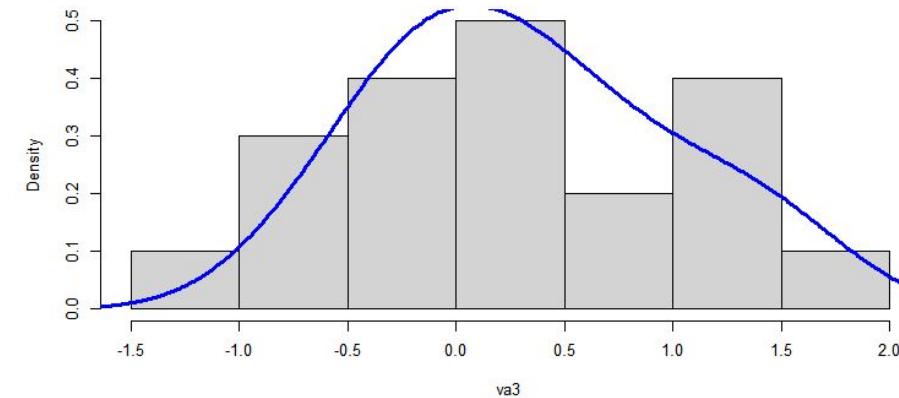
Distribuição t de Student com diferentes graus de liberdade



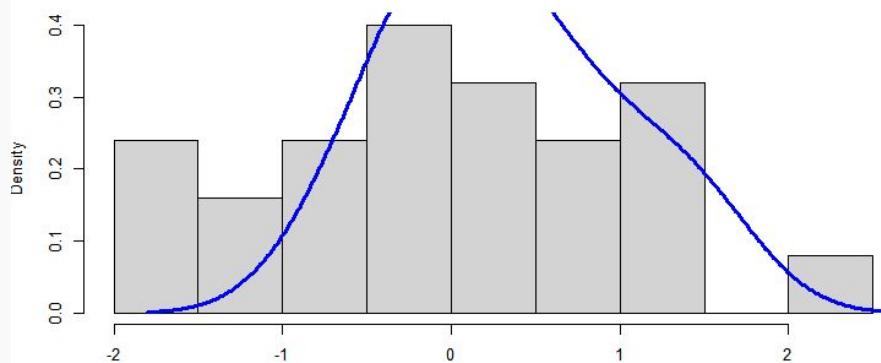
Distribuição t de Student com 09 graus de liberdade



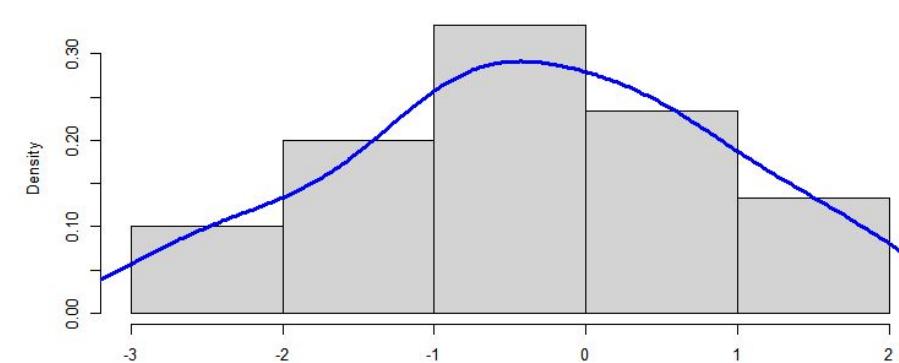
Distribuição t de Student com 19 graus de liberdade



Distribuição t de Student com 24 graus de liberdade



Distribuição t de Student com 29 graus de liberdade



*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Conclusão



- ✓ Intervalo de Confiança para Média, utilizando a distribuição Z (quando temos o desvio padrão populacional).
- ✓ Intervalo de Confiança para Média, utilizando a distribuição t de Student (quando não temos o desvio padrão populacional).



Próxima aula



- Intervalo de Confiança para Proporção.

Análise Estatística de Dados

AULA 3.3. INTERVALO DE CONFIANÇA PARA PROPORÇÃO

PROF. MÁIRON CHAVES

Nesta aula



- ❑ Intervalo de Confiança para Proporção.

Intervalo de Confiança para Proporção

Devoluções de um produto:

Proporção de Devoluções = 27,6%

Intervalo de Confiança de 95% para essa proporção, utilizando a distribuição Z (normal padrão):

*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Onde:

\hat{p} é a proporção amostral,

n é o tamanho da amostra,

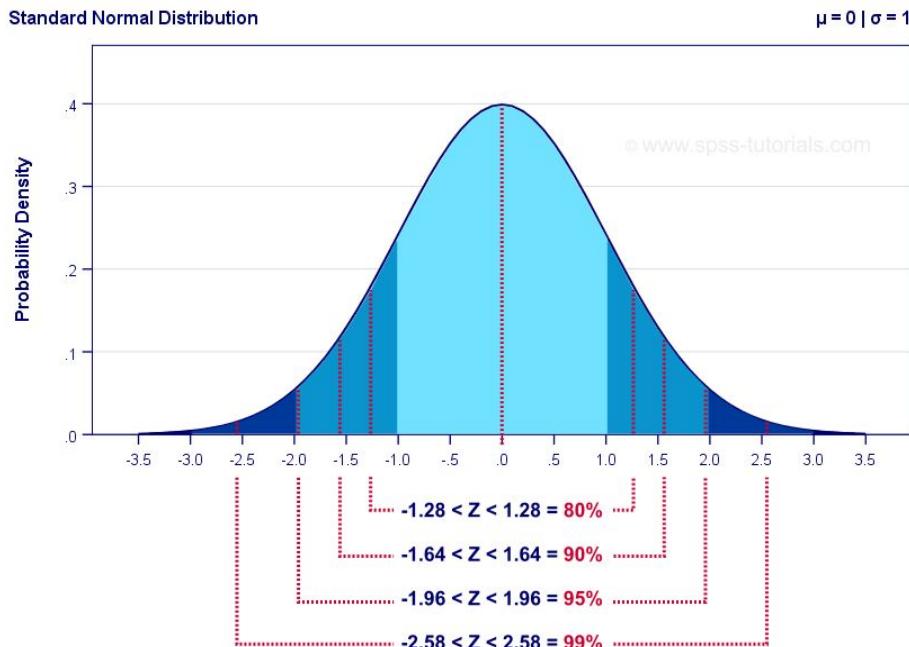
z é quantil da distribuição Z para nível $1-\alpha$ de confiança.



Intervalo de Confiança para Proporção

R:

```
ic <- 0.99  
alfa <- 1-ic  
area <- 1-(alfa/2)  
qnorm(area) #0.995  
>2.5758
```



$$IC_{99\%} = \hat{p} - z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p > \hat{p} + z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$IC_{99\%} = 0,276 - 2,58 * \sqrt{\frac{0,276(1-0,276)}{500}} < p > 0,276 + 2,58 * \sqrt{\frac{0,276(1-0,276)}{500}}$$

$$IC_{99\%} = 0,276 - 0,0515 < p > 0,276 + 0,0515$$

$$IC_{99\%} = 0,2245 < p > 0,3275$$

$$IC_{80\%} = \hat{p} - z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p > \hat{p} + z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$IC_{80\%} = 0,276 - 1,28 * \sqrt{\frac{0,276(1-0,276)}{500}} < p > 0,276 + 1,28 * \sqrt{\frac{0,276(1-0,276)}{500}}$$

$$IC_{80\%} = 0,276 - 0,0255 < p > 0,276 + 0,0255$$

$$IC_{80\%} = 0,2504 < p > 0,3015$$



*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

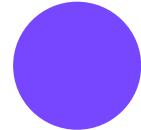
Conclusão



- ✓ Intervalo de Confiança para Proporção.



Próxima aula



- ❑ Intervalo de Confiança via método Bootstrap.

Análise Estatística de Dados

AULA 3.4. INTERVALO DE CONFIANÇA POR BOOTSTRAP

PROF. MÁIRON CHAVES

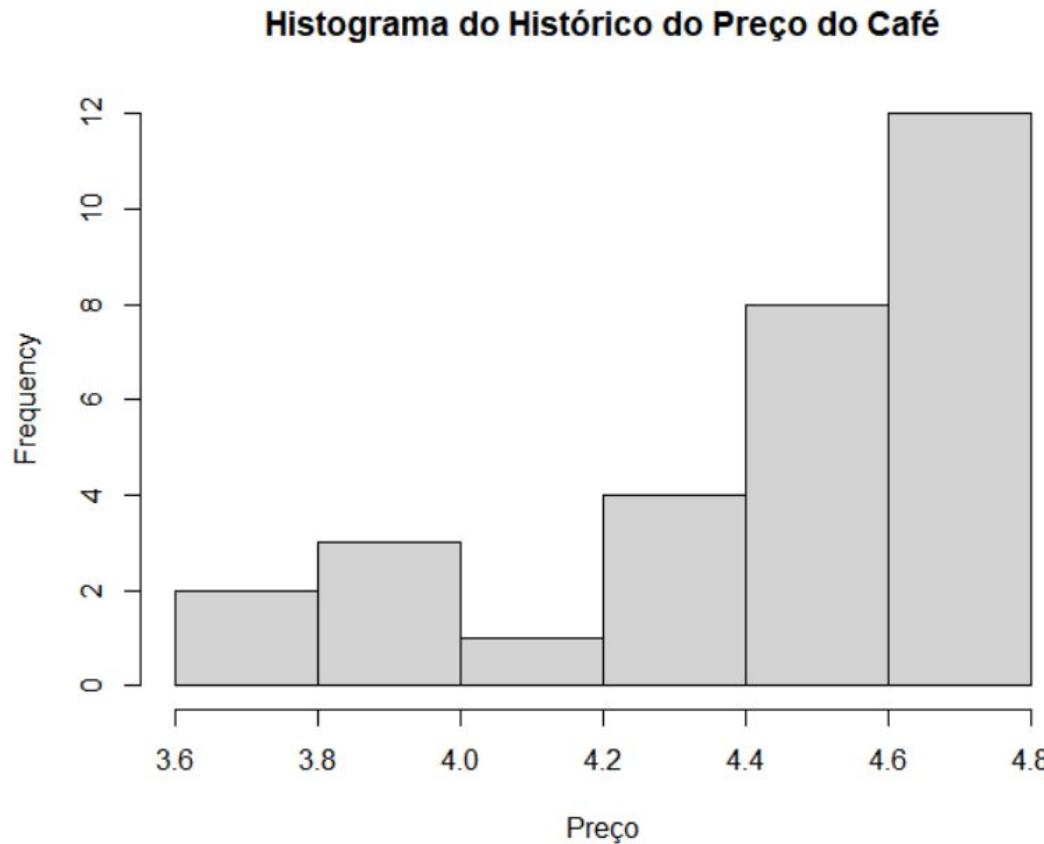
Nesta aula



- ❑ Intervalo de Confiança por Bootstrap.

Intervalo de Confiança por Bootstrap

ID	Amostra	Preco_Cafe
1		4,77
2		4,67
3		4,75
4		4,74
5		4,63
6		4,56
7		4,59
8		4,75
9		4,75
10		4,49
11		4,41
12		4,32
13		4,68
14		4,66
15		4,42
16		4,71
17		4,66
18		4,46
19		4,36
20		4,47
21		4,43
22		4,4
23		4,61
24		4,09
25		3,73
26		3,89
27		4,35
28		3,84
29		3,81
30		3,79



*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Intervalo de Confiança por Bootstrap

ID	Amostra
	Preco_Cafe
1	4,77
2	4,67
3	4,75
4	4,74
5	4,63
6	4,56
7	4,59
8	4,75
9	4,75
10	4,49
11	4,41
12	4,32
13	4,68
14	4,66
15	4,42
16	4,71
17	4,66
18	4,46
19	4,36
20	4,47
21	4,43
22	4,4
23	4,61
24	4,09
25	3,73
26	3,89
27	4,35
28	3,84
29	3,81
30	3,79

Algoritmo do Bootstrap:

1. Extrair uma subamostra aleatória de tamanho n, com reposição.
2. Calcular a média da subamostra e registrar.
3. Repetir o passo um 1 e 2 R vezes.



*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Intervalo de Confiança por Bootstrap



Amostra	ID	Preco_Cafe
1	5	4,63
2	1	4,77
3	5	4,63
4	8	4,75
5	5	4,63
6	9	4,75
7	5	4,63
8	1	4,77
9	3	4,75
10	7	4,59
11	10	4,49
12	5	4,63
13	10	4,49
14	6	4,56
15	1	4,77
16	8	4,75
17	3	4,75
18	8	4,75
19	10	4,49
20	3	4,75
21	9	4,75
22	2	4,67
23	4	4,74
24	2	4,67
25	10	4,49
26	5	4,63
27	9	4,75
28	6	4,56
29	4	4,74
30	3	4,75

ID Subamostra 1

5	4,63
1	4,77
5	4,63
8	4,75
5	4,63
9	4,75
9	4,75
5	4,63
1	4,77
3	4,75

Média 4,71

ID Subamostra 2

7	4,59
10	4,49
5	4,63
8	4,75
5	4,63
10	4,49
6	4,56
6	4,56
1	4,77
3	4,75

Média 4,63

ID Subamostra 3

10	4,49
1	4,77
3	4,75
6	4,56
6	4,56
7	4,59
4	4,74
4	4,74
3	4,75
5	4,63

Média 4,67

ID Subamostra 4

1	4,77
4	4,74
2	4,67
5	4,63
9	4,75
6	4,56
6	4,56
9	4,75
2	4,67
8	4,75

Média 4,69

ID Subamostra 5

10	4,49
8	4,75
7	4,59
6	4,56
5	4,63
9	4,75
7	4,59
5	4,63
1	4,77
2	4,67

Média 4,64

ID Subamostra 6

3	4,75
9	4,75
6	4,56
9	4,75
5	4,63
1	4,77
1	4,77
2	4,67
5	4,63
6	4,56

Média 4,70

ID Subamostra 7

10	4,49
5	4,63
8	4,75
3	4,75
8	4,75
1	4,77
8	4,75
3	4,75
10	4,49
1	4,77

Média 4,70

ID Subamostra 8

5	4,63
8	4,75
9	4,75
3	4,75
9	4,75
6	4,56
2	4,67
10	4,49
1	4,77
6	4,56

Média 4,67

ID Subamostra 9

9	4,75
2	4,67
8	4,75
3	4,75
5	4,63
9	4,75
6	4,56
1	4,77
6	4,56
10	4,49

Média 4,65

ID Subamostra 10

9	4,75
2	4,67
8	4,75
5	4,63
9	4,75
6	4,56
1	4,77
3	4,75
7	4,59
10	4,49

Média 4,68

ID Subamostra 11

4	4,74
1	4,77
5	4,63
6	4,56
3	4,75
8	4,75
1	4,77
2	4,67
5	4,63
6	4,56

Média 4,70

ID Subamostra 12

3	4,75
6	4,56
10	4,49
9	4,75
3	4,75
8	4,75
6	4,56
2	4,67
5	4,63
1	4,77

Média 4,66

*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Intervalo de Confiança por Bootstrap



ID	Amostra
	Preco_Cafe
1	4,77
2	4,67
3	4,75
4	4,74
5	4,63
6	4,56
7	4,59
8	4,75
9	4,75
10	4,49
11	4,41
12	4,32
13	4,68
14	4,66
15	4,42
16	4,71
17	4,66
18	4,46
19	4,36
20	4,47
21	4,43
22	4,4
23	4,61
24	4,09
25	3,73
26	3,89
27	4,35
28	3,84
29	3,81
30	3,79

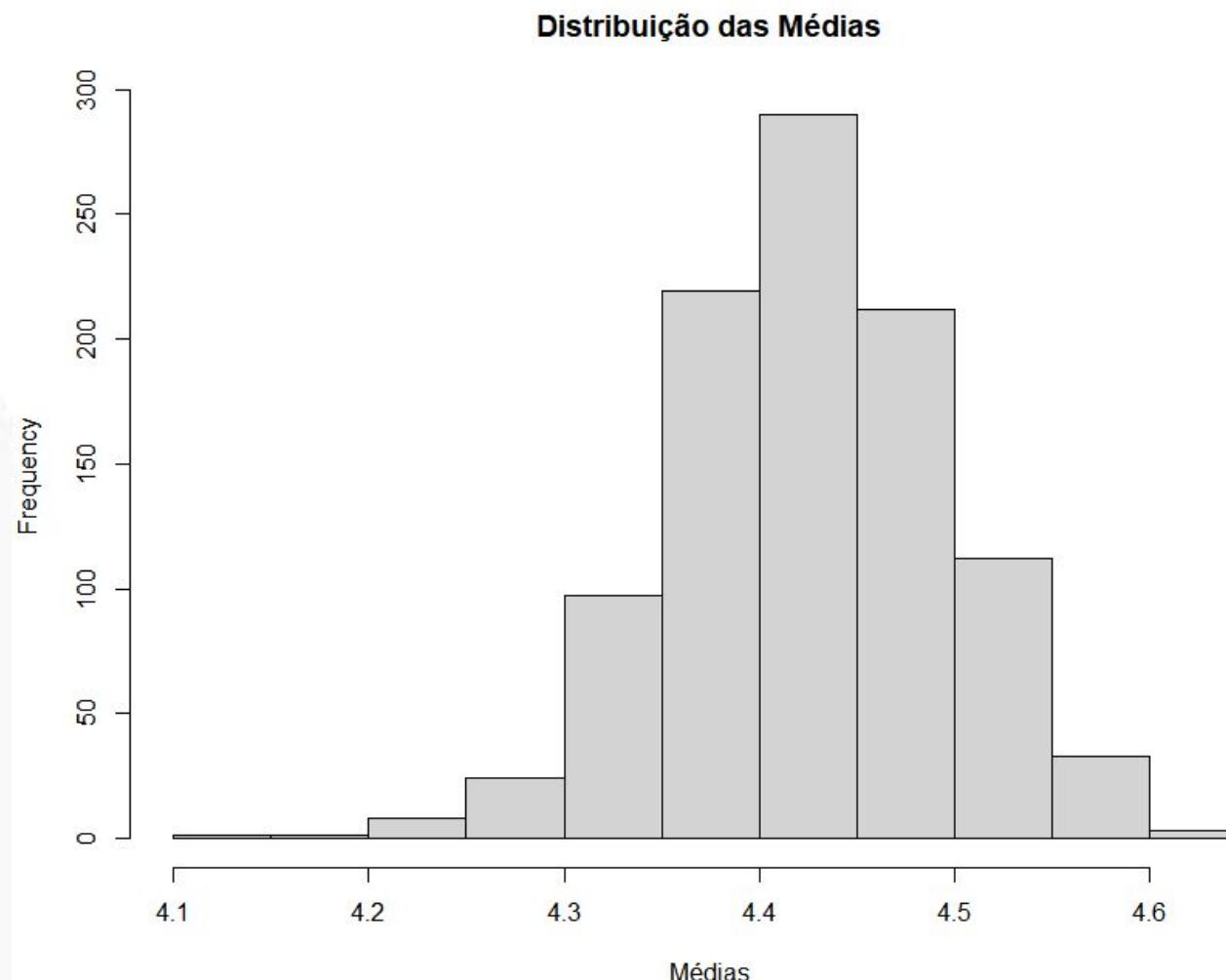


```
for (i in 1:1000) {  
  reamostra <- sample(dados$Preco_Cafe, size = 20, replace = T)  
  medias[i] <- mean(reamostra)  
}
```

*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Intervalo de Confiança por Bootstrap



*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Intervalo de Confiança por Bootstrap



n=100	4.1490
	4.1825
	4.2305
	4.2375
	4.2395
	4.2450
	4.2450
	4.2455
	4.2455
	4.2465
	..
	4.5840
	4.5895
	4.5995
	4.6000
	4.6060
	4.6115
	4.6220



```
> quantile( medias, probs = c(0.025,0.975))  
 2.5% 97.5%  
4.286000 4.560512
```



```
> quantile( medias, probs = c(0.05,0.95))  
 5% 95%  
4.318975 4.541525
```

*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

Conclusão



- ✓ Intervalo de Confiança utilizando o método Bootstrap.



Próxima aula



- Estatística Computacional – Intervalo de Confiança
com o R.

Análise Estatística de Dados

AULA 3.5. ESTATÍSTICA COMPUTACIONAL – INTERVALOS DE CONFIANÇA COM R

PROF. MÁIRON CHAVES

Nesta aula



- Estatística Computacional – Intervalos de Confiança com R.

Estatística Computacional – Intervalos de Confiança com R

Estatística Computacional – Intervalos de Confiança com o R

```
#####
```

```
##### Intervalo de Confiança #####
```

```
## AED - Capítulo 03 - Prof. Máiron Chaves ##
```

```
#####
```

```
#Copie este código, cole no seu R e execute para ver os resultados
```

```
##### Intervalo de confiança para média amostral pela distribuição Normal
```

```
Padrão #####
```

```
# Obter o intervalo de confiança para uma variável cuja média = 30, desvio padrão =  
7,31 e n = 30
```

```
#Temos que definir o nível de confiança do nosso intervalo.
```

```
#Podemos obter o valor do quantil para o nível de confiança desejado com a função  
qnorm()
```

```
#O quantil na distribuição normal padrão para 95% de confiança
```

```
qnorm(0.95)
```



*Esta área
precisa ser
preservada em
todos os slides.
É onde você
aparecerá na
transmissão.

*Essa mensagem
será retirada pela
equipe de revisão.

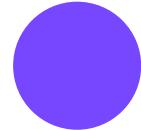
Conclusão



- ✓ Intervalos de Confiança no R.



Próxima aula



- Teste de Hipótese.

Análise Estatística de Dados

CAPÍTULO 4. TESTE DE HIPÓTESE

PROF. MÁIRON CHAVES

Análise Estatística de Dados

AULA 4.1. TESTE DE HIPÓTESE

PROF. MÁIRON CHAVES

Nesta aula

- O que é um teste de hipótese?
- Passos para condução de um teste de hipótese.
- Testes Unilaterais.

Teste de Hipótese

Conceito



É um procedimento estatístico em que, por meio da teoria das probabilidades, auxilia na tomada de decisão no sentido de rejeitar ou não hipóteses em um experimento científico.

Exemplos de hipóteses:

- Quando colocamos o produto na posição A da gôndola, vende significativamente mais do que quando ele está na posição B?
- A equipe de vendas após receber um treinamento, aumentou de forma significativa sua performance?
- Quando o cliente compra o produto A, ele também compra o produto B?
- A tendência de vendas ao longo dos meses parece estar aumentando, mas esse aumento é significativo ou são pequenas variações devido ao acaso?
- Clientes que entram em nossa loja acompanhados de crianças, compram mais do que aqueles que entram acompanhados de adultos?

Teste de Hipótese

Seis passos para condução de um teste de hipóteses

IGTI

Contexto



Sabemos que quando o produto está na posição A da gôndola, vende em média $\mu=R\$140$ com $\sigma=R\$27$.



Estou céтика. Vou conduzir um experimento. Deixarei o produto na posição A por $n = 15$ dias, e coletarei a média amostral.

Teste de Hipótese

Seis passos para condução de um teste de hipóteses

Contexto



Após os 15 dias, os resultados observados foram:

$$\bar{x} = 134 \quad \sigma = 27 \quad n = 15$$

Baseado nos resultados dessa amostra, você rejeita ou não rejeita a hipótese afirmada no início do enunciado, de que a venda média quando o produto está na posição A é R\$140?



Teste de Hipótese

Seis passos para condução de um teste de hipóteses

Passo 01 | Definir a hipótese nula (H_0) e a Hipótese Verdadeira (H_1)

Hipótese Nula (H_0) - É assumida como verdadeira na construção do teste.

Hipótese Alternativa (H_1) – É considerada verdadeira quando não há evidências para sustentar a hipótese nula.

$$H_0 : \mu = 140$$

$$H_1 : \mu \neq 140$$

Nosso H_0 (tido como verdade até então), é de que a média (populacional) de vendas quando o produto está na posição A da gôndola é de R\$140. Através da amostra que coletamos, iremos rejeitar ou não rejeitar essa hipótese probabilisticamente.

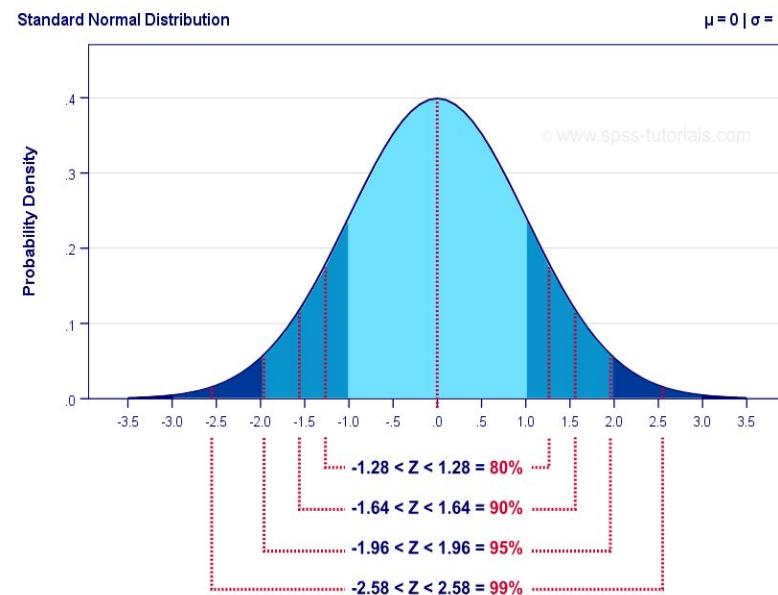
Teste de Hipótese

Seis passos para condução de um teste de hipóteses

Passo 02 | Definir o nível de Confiança e Significância

Nível de confiança: 90%, 95% ou 99%.

Nível de significância α : ao assumir o nível de confiança de 95%, a probabilidade de rejeitarmos a Hipótese Nula sendo que ela é verdadeira é de 5%.



Erro Tipo I: Rejeitar a Hipótese Nula quando ela é verdadeira

Teste de Hipótese

Seis passos para condução de um teste de hipóteses

Passo 03 | Calcular a Estatística de Teste

Uma estatística de teste é um valor calculado a partir de uma amostra de dados.

O seu valor é usado para decidir se podemos ou não rejeitar a hipótese nula.

A natureza do experimento irá direcionar qual distribuição de probabilidade usar.

Neste caso, estamos testando média e temos o desvio padrão da população.

Portanto, adotaremos a distribuição normal padrão para nosso experimento.

$$Z_{calculado} = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

Teste de Hipótese

Seis passos para condução de um teste de hipóteses



Passo 03 | Calcular a Estatística de Teste



Sabemos que quando o produto está na posição A da gôndola, vende em média $\mu=\text{R\$}140$ com $\sigma=\text{R\$}27$.

Após os 15 dias, os resultados observados foram:

$\bar{x} =$

$n =$

$$Z_{calculado} = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

$$Z_{calculado} = \frac{(134 - 140)}{\frac{27}{\sqrt{15}}}$$

$$Z_{calculado} = -0,8606$$

Teste de Hipótese

Seis passos para condução de um teste de hipóteses

Passo 04 | Delimitar a Região Crítica

Baseado no nível de significância α que adotamos para o teste, iremos delimitar a região crítica. Iremos rejeitar H_0 se nossa estatística de teste se encontrar na região crítica.



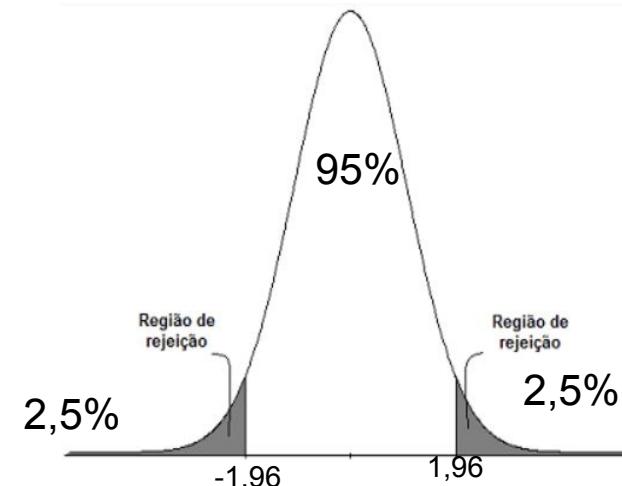
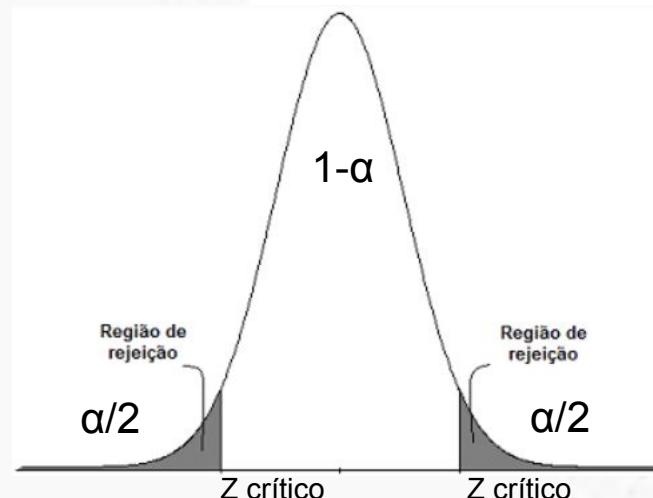
Teste de Hipótese

Seis passos para condução de um teste de hipóteses

Passo 04 | Delimitar a Região Crítica

Baseado no nível de significância α que adotamos para o teste, iremos delimitar a região crítica. Iremos rejeitar H_0 se nossa estatística de teste se encontrar na região crítica.

Região crítica bilateral na curva normal padrão para $\alpha = 5\%$.



Teste de Hipótese

Seis passos para condução de um teste de hipóteses

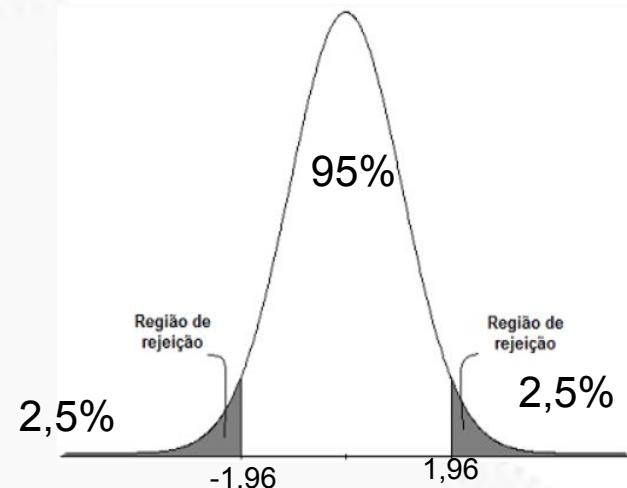
Passo 04 | Delimitar a Região Crítica

Baseado no nível de significância α que adotamos para o teste, iremos delimitar a região crítica. Iremos rejeitar H_0 se nossa estatística de teste se encontrar na região crítica.

iGTI

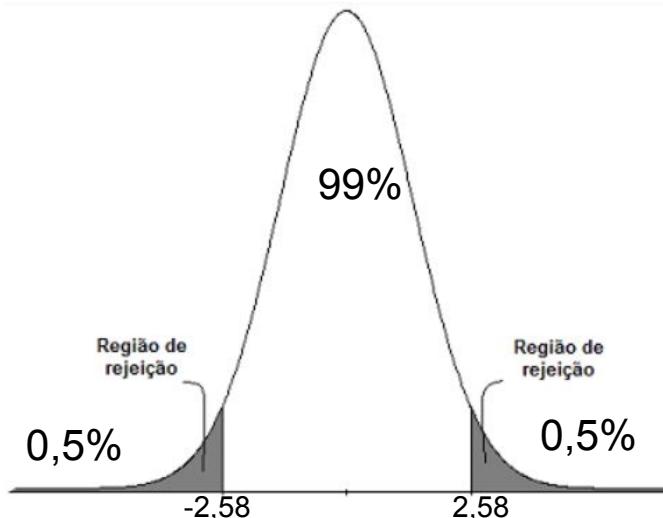
```
R:  
ic <- 0.95  
alfa <- 1-ic  
area <- 1-(alfa/2)  
qnorm(area) #0.975  
>1.9599
```

Região crítica bilateral na curva normal padrão para $\alpha = 5\%$



```
R:  
ic <- 0.99  
alfa <- 1-ic  
area <- 1-(alfa/2)  
qnorm(area) #0.995  
>2.5758
```

Região crítica bilateral na curva normal padrão para $\alpha = 1\%$



Teste de Hipótese

Seis passos para condução de um teste de hipóteses

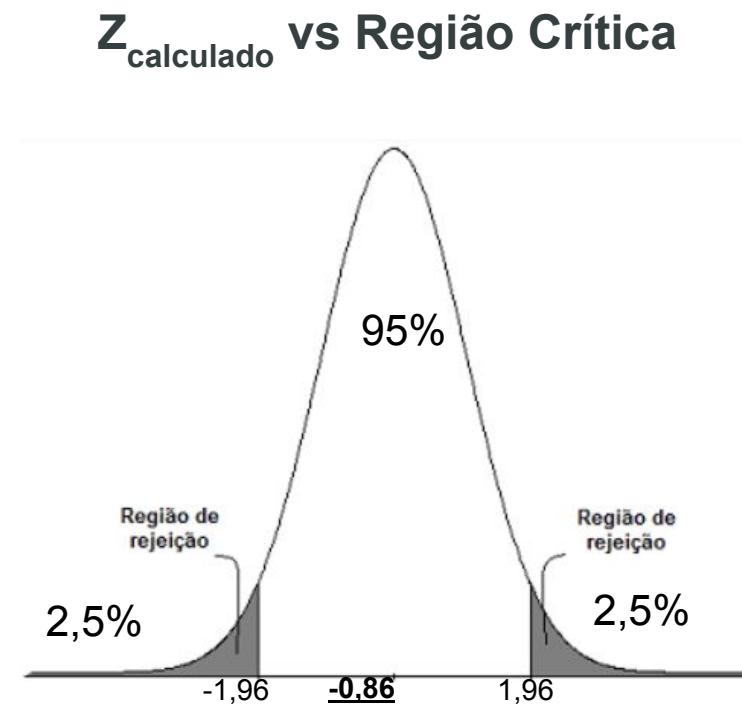
Passo 04 | Delimitar a Região Crítica

Baseado no nível de significância α que adotamos para o teste, iremos delimitar a região crítica. Iremos rejeitar H_0 se nossa estatística de teste se encontrar na região crítica.

$$Z_{calculado} = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

$$Z_{calculado} = \frac{(134 - 140)}{\frac{27}{\sqrt{15}}}$$

$$Z_{calculado} = -0,8606$$



Teste de Hipótese

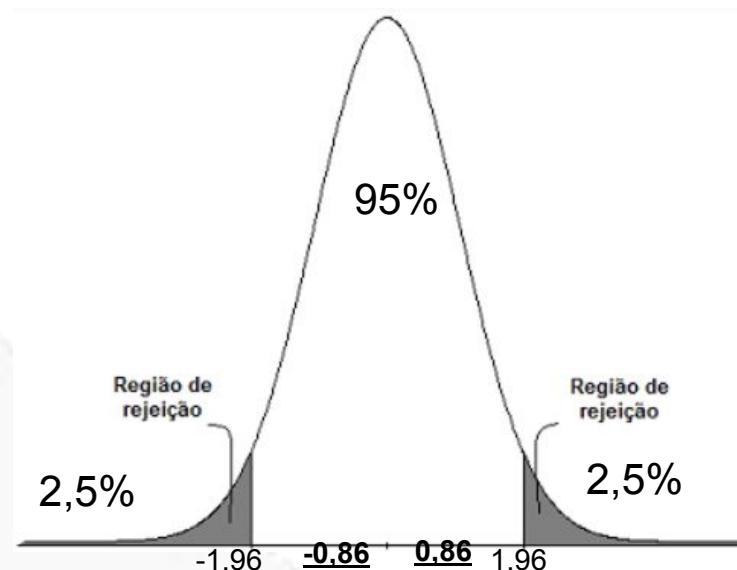
Seis passos para condução de um teste de hipóteses

Passo 05 | Obter o Valor p

O valor p é a probabilidade de rejeitar a Hipótese Nula quando ela é verdadeira (erro tipo I).

Ou seja, supondo que o nível de confiança adotado seja 95%, iremos rejeitar H₀ caso o valor p seja abaixo de 5% (nível de significância).

Resumindo: rejeitaremos a Hipótese Nula se o valor p for menor do que o nível de significância.



R:

```
>pnorm(0.86)-pnorm(-0.86)  
0.610211
```

Teste de Hipótese

Seis passos para condução de um teste de hipóteses

Passo 05 | Obter o Valor p

Supondo que o nível de confiança adotado seja 95%, iremos rejeitar H₀ caso o valor p seja abaixo de 5% (nível de significância).

A interpretação final seria:

Caso $p \leq 0.05$:

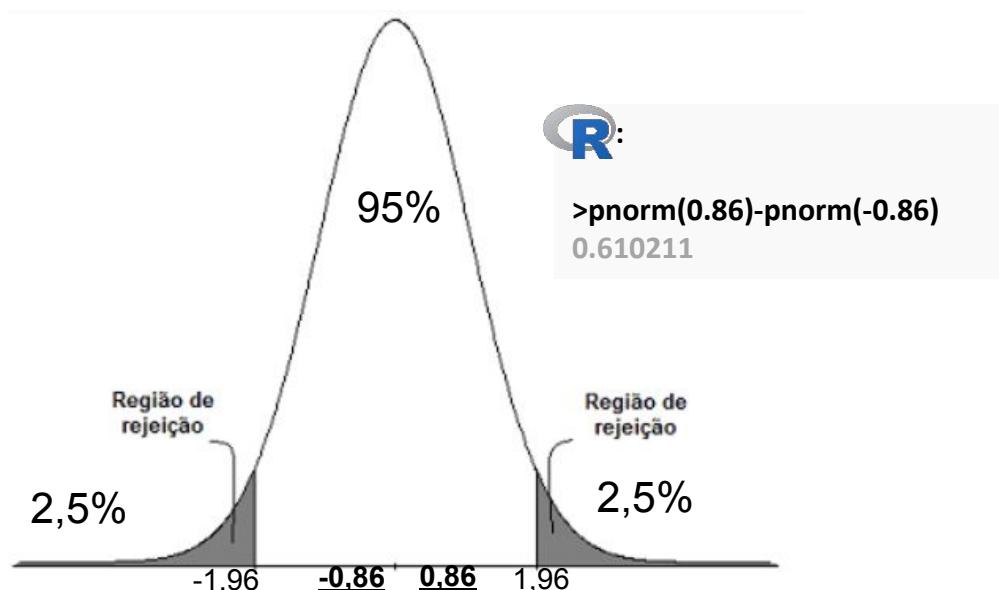
Com 95% de confiança, **há evidências para rejeitar a hipótese nula.**

Ou seja, a média de vendas quando o produto está na posição A é estatisticamente diferente de R\$140,00.

Caso $p > 0.05$:

Com 95% de confiança, **não há evidências para rejeitar a hipótese nula.**

Ou seja, a média de vendas quando o produto está na posição A, estatisticamente não é diferente de R\$140,00.



Teste de Hipótese

Seis passos para condução de um teste de hipóteses

Passo 06 | Rejeitar ou Não Rejeitar H_0

Já vimos que o Zcalculado está fora da região crítica e consequentemente o valor p é maior do que o nível de significância fixado.

Portanto, a resposta formal para concluir o teste fica:

Com 95% de confiança, não há evidências para rejeitar H_0 . Ou seja, a média de vendas quando o produto está na posição A da gôndola é estatisticamente igual a R\$140.

Teste de Hipótese

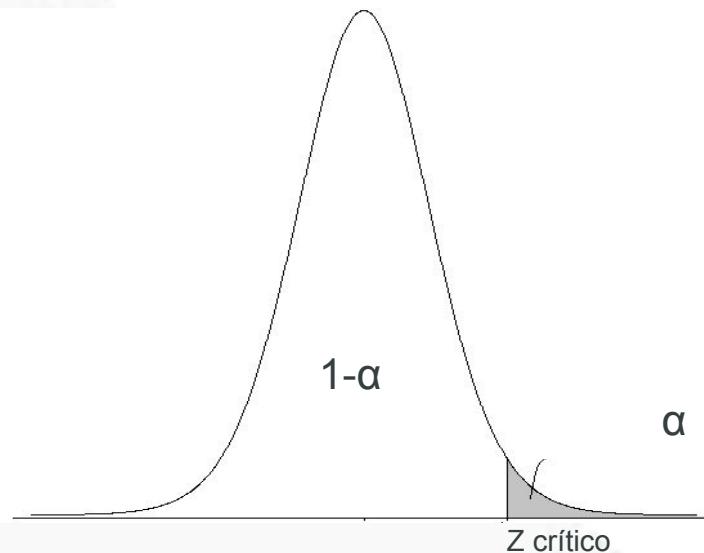
Teste Unilaterais

Um teste unilateral a direita fica:

$$H_0 : \mu = 140$$

$$H_1 : \mu > 140$$

Região crítica para um teste unilateral a direita.

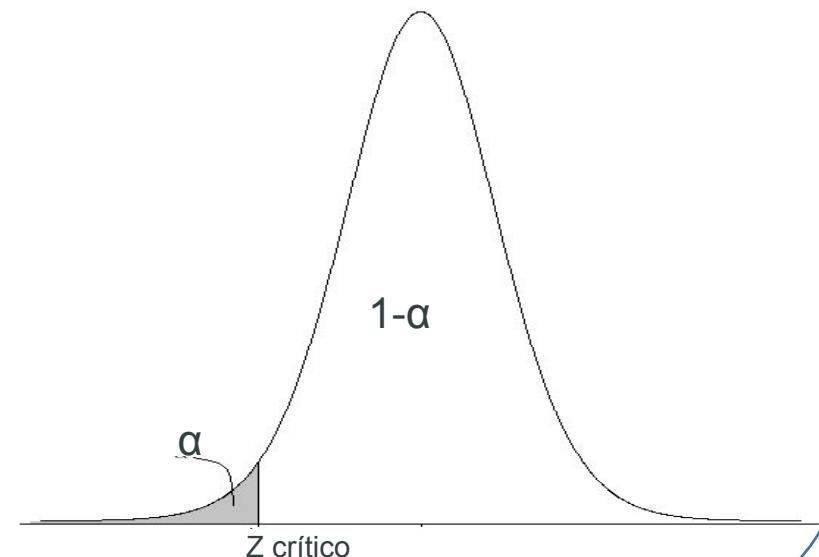


Um teste unilateral a esquerda fica:

$$H_0 : \mu = 140$$

$$H_1 : \mu < 140$$

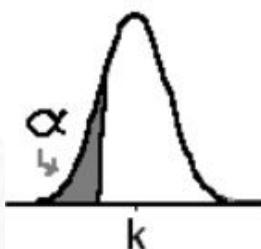
Região crítica para um teste unilateral a esquerda.



Teste de Hipótese

Níveis de Confiança e Valores críticos para distribuição Z

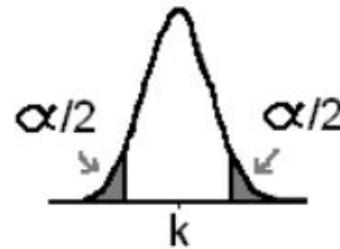
R:
ic <- 0.95
qnorm(1-ic)
>-1.6448



$$\begin{aligned}H_0: \mu &= k \\H_1: \mu &< k\end{aligned}$$

α	z critical
0.10	-1.28
0.05	-1.65
0.01	-2.33

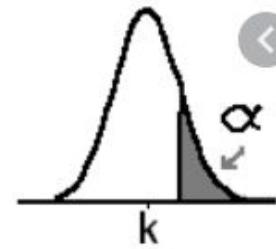
R:
c <- 0.95
alfa <- 1-ic
area <- 1-(alfa/2)
qnorm(area) #0.975
>1.9599



$$\begin{aligned}H_0: \mu &= k \\H_1: \mu &\neq k\end{aligned}$$

α	z critical
0.10	± 1.65
0.05	± 1.96
0.01	± 2.58

R:
ic <- 0.95
qnorm(ic)
>1.6448



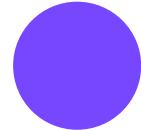
$$\begin{aligned}H_0: \mu &= k \\H_1: \mu &> k\end{aligned}$$

α	z critical
0.10	1.28
0.05	1.65
0.01	2.33

Conclusão



- ✓ O que é um Teste de Hipótese?
- ✓ Seis passos para condução de um Teste de Hipótese.
- ✓ Testes Unilaterais.



Próxima aula



- Avaliando a normalidade de uma variável aleatória.

Análise Estatística de Dados

AULA 4.2. AVALIANDO A NORMALIDADE DE UMA VARIÁVEL ALEATÓRIA

PROF. MÁIRON CHAVES

Nesta aula



- Histograma.
- QQ-Plot.
- Teste Shapiro-Wilk.

Avaliando a normalidade de uma Variável Aleatória



Seja **va1** e **va2** duas variáveis aleatórias, iremos avaliar se seguem uma distribuição normal.

Gerando as variáveis aleatórias:

:

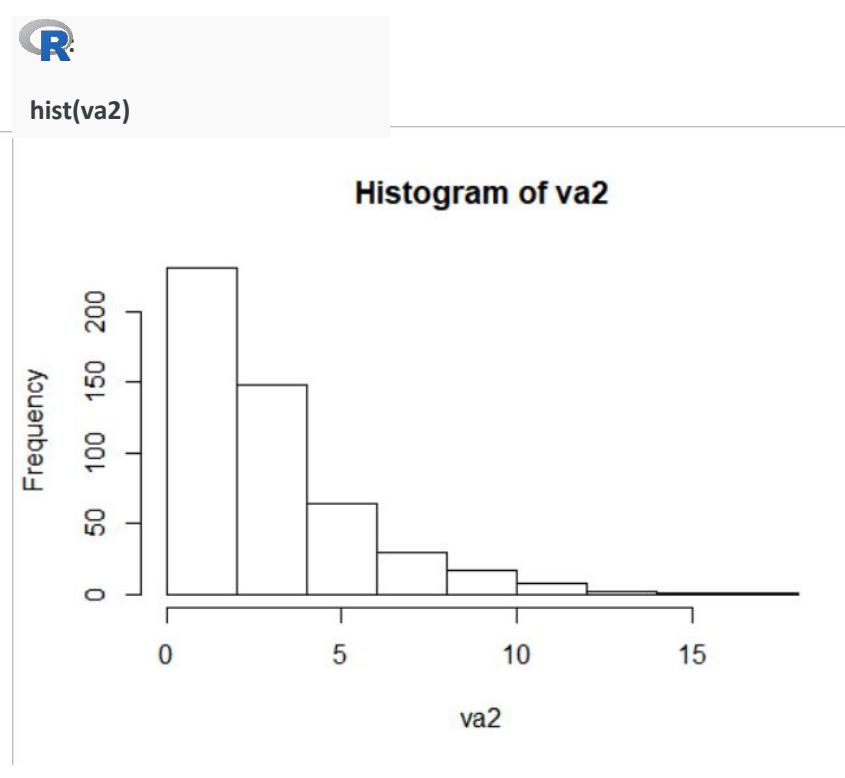
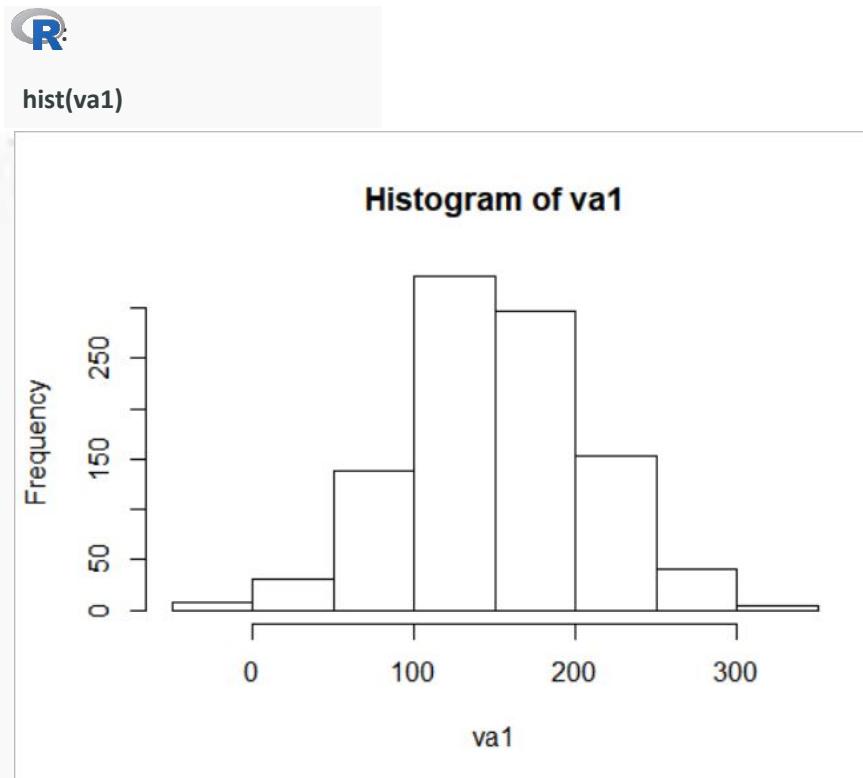
```
va1 <- rnorm(1000,mean = 150, sd = 60)  
va2 <- rchisq(500, df = 3)
```

Onde:

$va1 \sim N(\mu = 150, \sigma = 60)$
 $va2 \sim \chi^2(df = 3)$

Avaliando a normalidade de uma Variável Aleatória

Histograma | Se o histograma da variável aleatória for simétrico, é um indício de normalidade.

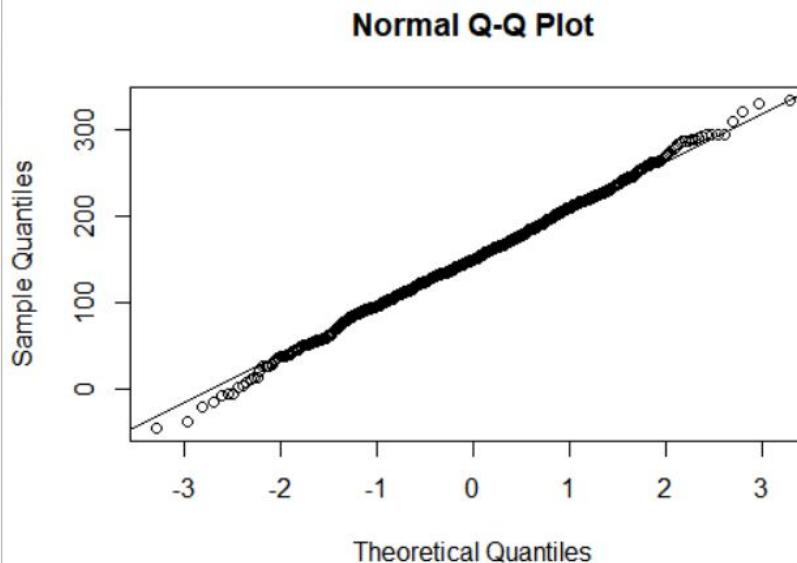


Avaliando a normalidade de uma Variável Aleatória

QQ Plot (Quantile-Quantile Plots) | Os pontos devem acompanhar a reta

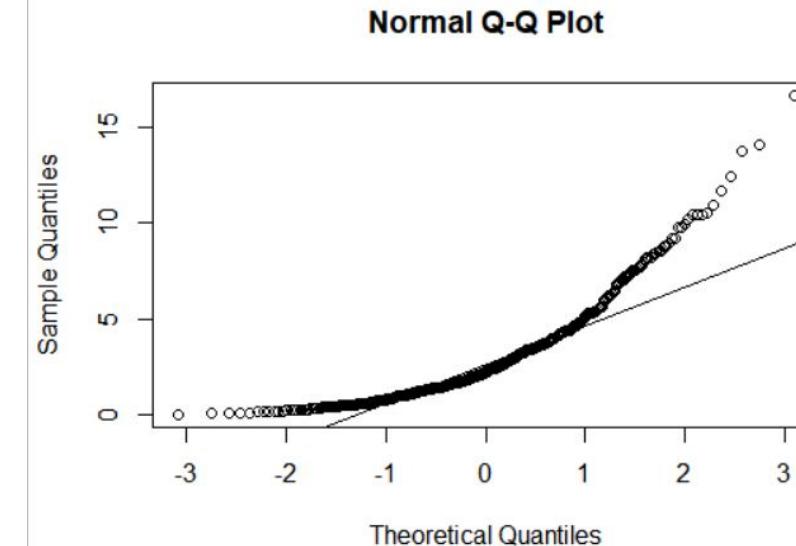
R:

```
qqnorm(va1)  
qqline(va1)
```



R:

```
qqnorm(va2)  
qqline(va2)
```



Avaliando a normalidade de uma Variável Aleatória



Teste Shapiro-Wilk | É um teste hipótese, onde:

H0: os dados seguem uma distribuição normal vs **H1**: os dados não seguem uma distribuição normal.

R:

```
shapiro.test(va1)
```

R:

```
shapiro.test(va2)
```

```
Shapiro-Wilk normality test  
data: val  
W = 0.99808, p-value = 0.317
```

```
Shapiro-Wilk normality test  
data: va2  
W = 0.8527, p-value <0.0000000000000002
```

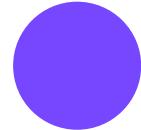
Conclusão



- ✓ Histograma para avaliar normalidade.
- ✓ QQ-Plot.
- ✓ Teste Shapiro-Wilk.



Próxima aula



- Teste t para diferença de médias em duas amostras independentes.

Análise Estatística de Dados

AULA 4.3. TESTE T PARA DIFERENÇA DE MÉDIAS EM DUAS AMOSTRAS INDEPENDENTES

PROF. MÁIRON CHAVES

Nesta aula



- Teste t para amostras independentes.

Teste t para diferença de médias em duas amostras independentes



Contexto



H₀: Quando o produto esta na posição B, seu faturamento é igual ao seu faturamento quando está na posição A.



H₁: Quando o produto esta na posição B, seu faturamento é diferente de quando está na posição A.



Teste t para diferença de médias em duas amostras independentes



Formalizando as hipóteses

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Ou, traduzindo pro nosso contexto, fica:

$$H_0: \mu_{\text{Posição A}} = \mu_{\text{Posição B}}$$

$$H_1: \mu_{\text{Posição A}} \neq \mu_{\text{Posição B}}$$

Iremos assumir 95% de confiança para conduzir o teste.

Consequentemente, nosso nível de significância será $\alpha = 5\%$.

Teste t para diferença de médias em duas amostras independentes



Contexto

Posição A

Vendas observadas na posição A:

$$s_1 = 17$$

Posição B

Vendas observadas na posição B:

$$s_2 = 19,2$$

Teste t para diferença de médias em duas amostras independentes

Calculando a estatística de teste (t calculado)

Posição A

Vendas observadas na posição A:

$$s_1 = 17$$

Posição B

Vendas observadas na posição B:

$$s_2 = 19,2$$

$$t_{calculado} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Teste t para diferença de médias em duas amostras independentes

Calculando a estatística de teste | t calculado



Posição A

Vendas observadas na posição A:

$$s_1 = 17$$

Posição B

Vendas observadas na posição B:

$$s_2 = 19,2$$

$$t_{calculado} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t_{calculado} = \frac{(150,1 - 182,1)}{\sqrt{\frac{17^2}{25} + \frac{19,2^2}{30}}}$$

$$t_{calculado} = -6,5527$$

Teste t para diferença de médias em duas amostras independentes

Calculando a estatística de teste | Graus de Liberdade



Posição A

Vendas observadas na posição A:

$$s_1 = 17$$

Posição B

Vendas observadas na posição B:

$$s_2 = 19,2$$

$$\text{Graus de liberdade} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1 - 1}\right)^2 + \left(\frac{s_2^2}{n_2 - 1}\right)^2}$$

$$\text{Graus de liberdade} = \frac{\left(\frac{17^2}{25} + \frac{19,2^2}{30}\right)^2}{\left(\frac{17^2}{25 - 1}\right)^2 + \left(\frac{19,2^2}{30 - 1}\right)^2}$$

$$\text{Graus de liberdade} = 52,7831$$

Teste t para diferença de médias em duas amostras independentes

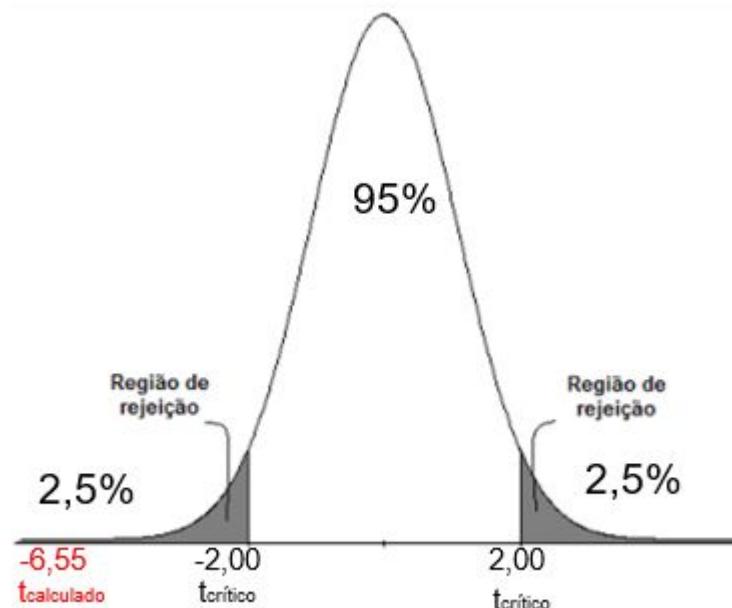
Delimitando a região crítica | Obtendo o t crítico

Nossa estatística de teste $t_{calculado} = -6,5527$ segue uma distribuição t de Student com 52,7831 graus de liberdade.

R:
>qt(p=0.975, df=52.7831)
2.005938

R:
>2*pt(q = -6.5527, df = 52.7831)
0.00000002403875

Região crítica para uma t de Student, com 52,78 graus de liberdade e $\alpha = 5\%$.



Valor p < α então:

A resposta formal fica: com 95% de confiança, há evidência para rejeitar a hipótese nula, ou seja, as vendas do produto na posição A são estatisticamente diferentes das vendas do produto na posição B.

Teste t para diferença de médias em duas amostras independentes

Obtendo o t crítico pela tabela da distribuição t de Student

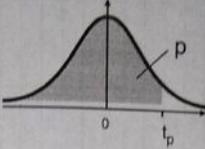


R:

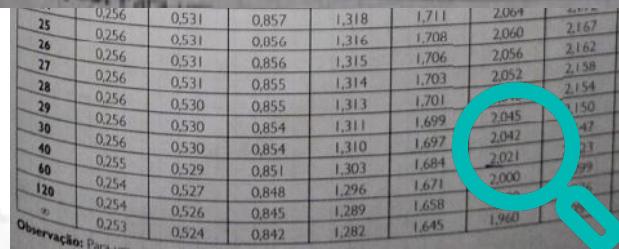
```
>qt(p=0.975, df=52.7831)  
2.005938
```

R:

```
>2*pt(q = -6.5527, df = 52.7831)  
0.00000002403875
```

		TABELAS DE PROBABILIDADE								
		TABELA II: DISTRIBUIÇÃO T DE STUDENT								
										
v	P	0,6	0,7	0,8	0,9	0,95	0,975	0,98	0,99	0,995
15		0,256		0,531		0,856		1,316		1,708
26		0,256		0,531		0,856		1,315		1,706
27		0,256		0,531		0,855		1,314		1,703
28		0,256		0,530		0,855		1,313		1,701
29		0,256		0,530		0,855		1,311		1,699
30		0,256		0,530		0,854		1,310		1,697
40		0,256		0,530		0,854		1,303		1,684
60		0,255		0,529		0,851		1,296		1,671
120		0,254		0,527		0,848		1,289		1,658
90		0,254		0,526		0,845		1,282		1,645
	Observação: Para v > 30, t_p = t_0,975	0,253		0,524		0,842		1,282		1,645

v	0,256	0,531	0,857	1,318	1,711	2,064	2,167	2,485	2,787
25	0,256	0,531	0,856	1,316	1,708	2,060	2,162	2,479	2,779
26	0,256	0,531	0,856	1,315	1,706	2,056	2,158	2,473	2,771
27	0,256	0,531	0,855	1,314	1,703	2,052	2,154	2,467	2,763
28	0,256	0,530	0,855	1,313	1,701	2,048	2,150	2,462	2,756
29	0,256	0,530	0,855	1,311	1,699	2,045	2,147	2,457	2,750
30	0,256	0,530	0,854	1,311	1,697	2,042	2,144	2,454	2,744
40	0,256	0,530	0,854	1,310	1,697	2,042	2,141	2,452	2,740
60	0,255	0,529	0,851	1,303	1,684	2,021	2,129	2,423	2,660
120	0,254	0,527	0,848	1,296	1,671	2,000	2,116	2,390	2,617
90	0,254	0,526	0,845	1,289	1,658	1,960	2,106	2,358	2,576



Conclusão

- ✓ Conduzir um teste t de Student para comparar duas médias (amostras independentes).
- ✓ Formalizar as hipóteses.
- ✓ Definir o nível de confiança e o nível de significância α .
- ✓ Obter o t calculado.
- ✓ Obter os graus de liberdade.
- ✓ Delimitar a região crítica.
- ✓ Obter o valor p.
- ✓ Formalizar a conclusão do teste.



Próxima aula



- Teste t para diferença de médias em duas amostras dependentes.

Análise Estatística de Dados

AULA 4.4. TESTE T PARA DIFERENÇA DE MÉDIAS EM DUAS AMOSTRAS DEPENDENTES

PROF. MÁIRON CHAVES

Nesta aula



- Teste t para amostras dependentes (teste t pareado).

Teste t para diferença de médias em duas amostras dependentes

O teste t pareado é útil para analisar o mesmo conjunto de indivíduos que foram medidos sob condições diferentes.

Geralmente utilizado para verificar a eficácia de um tratamento.

Exemplos:



Submeter um grupo de n indivíduos a um conjunto de atividades físicas (tratamento) e comparar sua capacidade respiratória antes e após o tratamento.



Submeter um grupo de n indivíduos portadores de pressão alta ao medicamento A, e comparar os resultados antes e após o tratamento.

Teste t para diferença de médias em duas amostras dependentes

Contexto



Submeter um grupo de n indivíduos a uma dieta e comparar seus respectivos pesos, antes e após o tratamento.

Teste t para diferença de médias em duas amostras dependentes

Formalizando as hipóteses



$$H_0: \mu_2 = \mu_1$$

$$H_1: \mu_2 < \mu_1$$

Ou traduzindo pro nosso contexto, fica:

$$H_0: \mu_{\text{Após a Dieta}} = \mu_{\text{Antes da Dieta}}$$

$$H_1: \mu_{\text{Após a Dieta}} < \mu_{\text{Antes da Dieta}}$$

Iremos assumir 90% de confiança para conduzir o teste.

Consequentemente, nosso nível de significância será $\alpha = 10\%$.

Teste t para diferença de médias em duas amostras independentes

Contexto

Antes da dieta

Resultados observados antes da dieta:

$$s_{antes} = 18$$

Após a dieta

Resultados observados após a dieta:

$$s_{antes} = 28$$

Teste t para diferença de médias em duas amostras dependentes

Calculando a estatística de teste (t calculado)

Antes da dieta

Resultados observados antes da dieta:

$$s_{antes} = 18$$

Após a dieta

Resultados observados após a dieta:

$$s_{depois} = 28$$

	antes_da_dieta	depois_da_dieta	diferença
1	112.3106	122.0665	9.75584270
2	122.0929	114.1543	-7.93861938
3	118.7420	116.6106	-2.13137439
4	132.4961	114.7139	-17.78221867
5	121.8623	121.8264	-0.03592273
6	124.9999	112.4697	-12.53011447
7	111.2142	121.9515	10.73735225
8	130.5020	119.2740	-11.22801973
9	108.1845	129.2115	21.02697536
10	114.3790	118.9030	4.52395584
11	121.4324	119.1146	-2.31781703
12	121.5339	114.9019	-6.63209615
13	116.8041	118.2447	1.44068268
14	130.8121	123.2356	-7.57648395
15	121.9639	127.1331	5.16919244
16	119.5322	126.6805	7.14830221
17	127.8110	119.1905	-8.62045583
18	107.2163	110.7567	3.53976178
19	139.5826	124.6663	-14.91628210
20	131.1034	116.0147	-15.08864708

$$\bar{x} = -2,1727$$
$$s = 10,1803$$

Teste t para diferença de médias em duas amostras dependentes



Calculando a estatística de teste (t calculado)

Nossa estatística de teste seguirá uma distribuição t de Student com $n - 1$ graus de liberdade:

Resultados observados na diferença

$$s = 10,1803$$

$$t_{calculado} = \frac{\bar{x}_{diferença}}{\frac{s_{diferença}}{\sqrt{n}}}$$

$$t_{calculado} = \frac{-2,1728}{\frac{10,1803}{\sqrt{20}}}$$

$$t_{calculado} = -0,9544$$

Teste t para diferença de médias em duas amostras dependentes

Calculando a estatística de teste | Graus de Liberdade



Nossa estatística de teste seguirá uma distribuição t de Student com $n - 1$ graus de liberdade.

Resultados observados na diferença:

$$s = 10,1803$$

$$\text{Graus de Liberdade} = n - 1$$

$$\text{Graus de Liberdade} = 20 - 1$$

$$\text{Graus de Liberdade} = 19$$

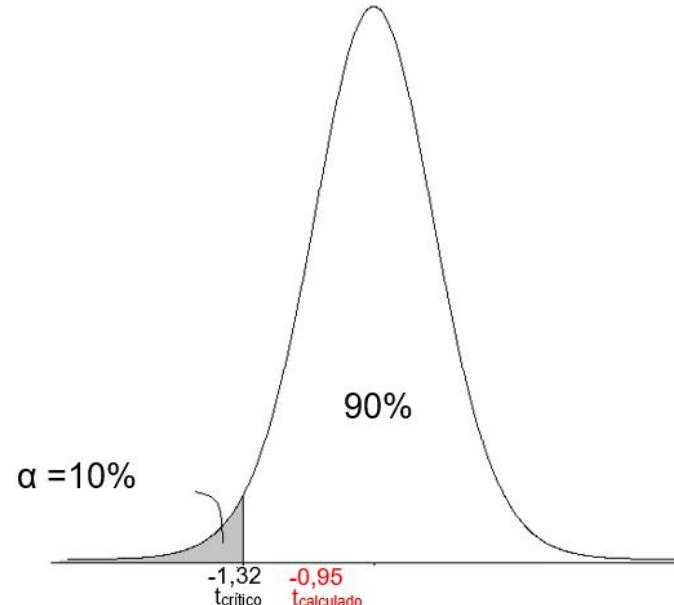
Teste t para diferença de médias em duas amostras dependentes

Delimitando a região crítica | Obtendo o t crítico

Nossa estatística de teste $t_{calculado} = -0.9544$ segue uma distribuição t de Student com 19 graus de liberdade.

R:
`>qt(p=0.90, df=19)
1.3277`

Região crítica unilateral a esquerda para uma distribuição t de Student, com 19 graus de liberdade ao nível de confiança de 90%.



Valor p > α, então:

A resposta formal fica: com 90% de confiança, não há evidências para rejeitar a hipótese nula, ou seja, o peso médio dos indivíduos após o tratamento com a dieta não é estatisticamente menor que o peso médio antes da dieta.

Teste t para diferença de médias em duas amostras dependentes

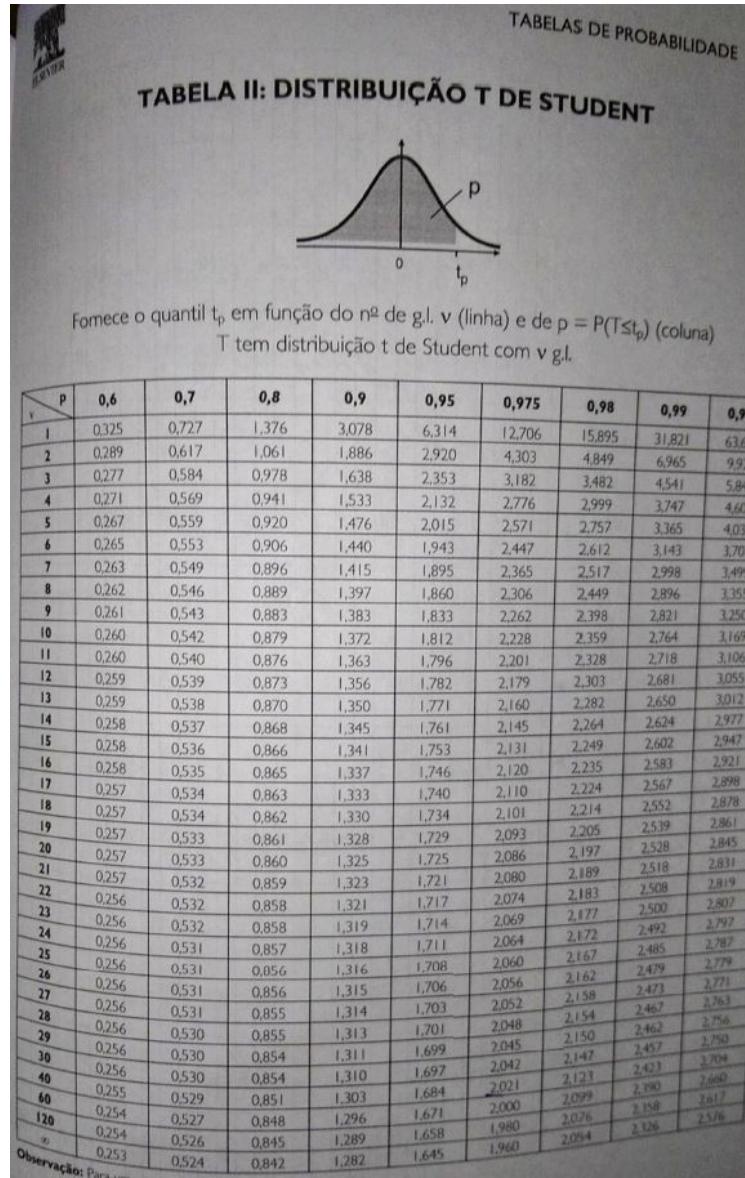
Obtendo o t crítico pela tabela da distribuição t de Student



```
>qt(p=0.90, df=19)  
1.3277
```



```
>pt(q = -0.9544, df=19)  
0.1759
```

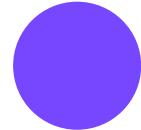


Conclusão

- ✓ Conduzir um teste t de Student para comparar duas médias (amostras dependentes).
- ✓ Formalizar as hipóteses.
- ✓ Definir o nível de confiança e o nível de significância α .
- ✓ Obter o t calculado.
- ✓ Obter os graus de liberdade.
- ✓ Delimitar a região crítica.
- ✓ Obter o valor p.
- ✓ Formalizar a conclusão do teste.



Próxima aula



- Teste Qui-Quadrado para independência entre variáveis categóricas.

Análise Estatística de Dados

AULA 4.5. TESTE QUI-QUADRADO PARA INDEPENDÊNCIA ENTRE VARIÁVEIS CATEGÓRICAS

PROF. MÁIRON CHAVES

Nesta aula



- Teste Qui-Quadrado para independência entre variáveis categóricas.

Teste Qui-Quadrado para independência entre variáveis categóricas

É um teste para mensurar se duas variáveis categóricas(qualitativas) são independentes.

Exemplos:



O fato do cliente comprar o Produto A está associado ao fato dele comprar o Produto B?



O fato do indivíduo ter curso superior está associado ao fato dele conseguir cargo de gerência.

Teste Qui-Quadrado para independência entre variáveis categóricas

Contexto



Desejamos investigar se um produto vende mais quando o cliente adulto está acompanhado de uma criança.

Achamos que o cliente compra independente de estar ou não com criança.

Teste Qui-Quadrado para independência entre variáveis categóricas

Contexto



Desejamos investigar se um produto vende mais quando o cliente adulto está acompanhado de uma criança.

Achamos que o cliente compra independente de estar ou não com criança.

Para isso, observamos 50 clientes, com criança e sem criança, que compraram e não compraram o produto.

Características observadas dos clientes

Cliente	Comprou
1 Adulato_com_Crianca	Não_Comprou
2 Adulato_com_Crianca	Não_Comprou
3 Adulato_com_Crianca	Não_Comprou
4 Adulato	Não_Comprou
5 Adulato	Não_Comprou
6 Adulato	Não_Comprou
7 Adulato_com_Crianca	Comprou
8 Adulato_com_Crianca	Comprou
9 Adulato_com_Crianca	Comprou
10 Adulato_com_Crianca	Comprou
11 Adulato_com_Crianca	Comprou
12 Adulato_com_Crianca	Comprou
13 Adulato_com_Crianca	Comprou
14 Adulato_com_Crianca	Comprou
15 Adulato_com_Crianca	Comprou
16 Adulato_com_Crianca	Comprou
17 Adulato_com_Crianca	Comprou
18 Adulato_com_Crianca	Comprou
...	...
49 Adulato	Comprou
50 Adulato	Comprou

Teste Qui-Quadrado para independência entre variáveis categóricas

Formalizando as hipóteses

H_0 : Não existe associação significativa entre as variáveis.

H_1 : Existe associação significativa entre as variáveis.

Traduzindo pro contexto proposto fica:

H_0 : o fato do cliente estar ou não com criança, não tem relação com o fato de comprar ou não comprar.

H_1 : o fato do cliente estar ou não com criança, tem relação com fato de comprar ou não comprar.

Iremos assumir 95% de confiança para conduzir o teste.

Consequentemente nosso nível de significância será $\alpha = 5\%$.

Teste Qui-Quadrado para independência entre variáveis categóricas



Contexto

Resultados observados após a coleta de dados:

Tabela de contingência 2x2

	Comprou	Não_Comprou	Total (Colunas)
Adulto	6	14	20
Adulto_com_Crianca	23	7	30
Total (linhas)	29	21	50

Teste Qui-Quadrado para independência entre variáveis categóricas



Calculando a estatística de teste | $\chi^2_{calculado}$

Calcular a frequência esperada para cada casela, pois se a distância entre a frequência observada e a esperada for “grande” o suficiente, teremos evidências para rejeitar H_0

Frequência observada:

	Comprou	Não_Comprou	Total (Colunas)
Adulto	6	14	20
Adulto_com_Crianca	23	7	30
Total (linhas)	29	21	50

$$E_{ij} = \frac{n_i * n_j}{n}$$

Onde:

i representa as linhas.

j representa as colunas.

n a quantidade de observações.

Frequência esperada:

	Comprou	Não_Comprou	Total (Colunas)
Adulto	11,6	8,4	20
Adulto_com_Crianca	17,4	12,6	30
Total (linhas)	29	21	50

Teste Qui-Quadrado para independência entre variáveis categóricas

Calculando a estatística de teste | $\chi^2_{calculado}$

Calcular a frequência esperada para cada casela, pois se a distância entre a frequência observada e a esperada for “grande” o suficiente, teremos evidências para rejeitar H_0

Frequência observada:

	Comprou	Não_Comprou	Total (Colunas)
Adulto	6	14	20
Adulto_com_Crianca	23	7	30
Total (linhas)	29	21	50

Frequência esperada:

	Comprou	Não_Comprou	Total (Colunas)
Adulto	11,6	8,4	20
Adulto_com_Crianca	17,4	12,6	30
Total (linhas)	29	21	50

$$E_{ij} = \frac{n_i * n_j}{n}$$

$$E_{11} = (29*20)/50 = 11,6$$

$$E_{12} = (21*20)/50 = 8,4$$

$$E_{21} = (29*30)/50 = 17,4$$

$$E_{22} = (21*30)/50 = 12,6$$

Teste Qui-Quadrado para independência entre variáveis categóricas



Calculando a estatística de teste | $\chi^2_{calculado}$

Calcular a frequência esperada para cada casela, pois se a distância entre a frequência observada e a esperada for “grande” o suficiente, teremos evidências para rejeitar H0.

Frequência observada:

	Comprou	Não_Comprou	Total (Colunas)
Adulto	6	14	20
Adulto_com_Crianca	23	7	30
Total (linhas)	29	21	50

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Onde:

r representa as linhas da tabela de contingência.

c representa as colunas da tabela de contingência.

$$\chi^2 = 10,7279$$

Frequência esperada:

	Comprou	Não_Comprou	Total (Colunas)
Adulto	11,6	8,4	20
Adulto_com_Crianca	17,4	12,6	30
Total (linhas)	29	21	50

Teste Qui-Quadrado para independência entre variáveis categóricas



Calculando a estatística de teste | $\chi^2_{calculado}$ | Graus de Liberdade

Nossa estatística de teste qui-quadrado terá (linhas-1)*(colunas-1) graus de liberdade.

Frequência observada:

	Comprou	Não_Comprou	Total (Colunas)
Adulto	6	14	20
Adulto_com_Crianca	23	7	30
Total (linhas)	29	21	50

$$\chi^2 = 10,7279 \text{ com um } 1 \text{ grau de liberdade}$$

Teste Qui-Quadrado para independência entre variáveis categóricas

Delimitando a região crítica | Obtendo o $\chi^2_{crítico}$

Região crítica de 95% de confiança para uma distribuição qui-quadrado com 1 grau de liberdade.



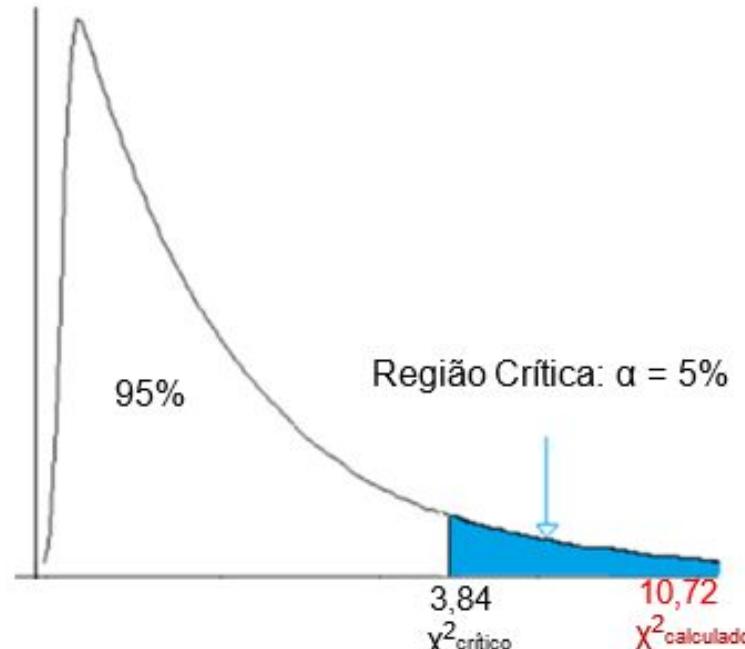
```
>qchisq(p=0.95,df = 1)  
3.8414
```



```
>1-pchisq(q=10.72,df = 1)  
0.0010
```

Valor p < α , então:

A resposta formal fica: com 95% de confiança, temos evidências para rejeitar a hipótese nula. Ou seja, o fato do cliente estar ou não com criança, tem relação com o fato do cliente comprar ou não comprar.



Teste Qui-Quadrado para independência entre variáveis categóricas

IGTI

Delimitando a região crítica | Obtendo o $\chi^2_{crítico}$

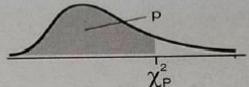
R :

```
>qchisq(p=0.95,df = 1)  
3.8414
```

R :

```
>1-pchisq(q=10.72,df = 1)  
0.0010
```

TABELA IV: DISTRIBUIÇÃO QUI-QUADRADO



Fornecendo o quantil χ^2_p em função do nº de gl. v (linha) e de p = P($\chi^2 \leq \chi^2_p$) (coluna). χ^2 tem distribuição qui-quadrado com v gl.

v \ P	0,005	0,010	0,025	0,050	0,100	0,250	0,500	0,750	0,900	0,950	0,975	0,990	0,995
1	0,000	0,000	0,001	0,004	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338	22,718	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337	28,241	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,435	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	24,478	29,336	34,800	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	29,051	33,660	39,335	45,616	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335	56,334	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335	66,981	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	61,698	69,334	77,577	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	71,145	79,334	88,130	96,578	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	73,291	80,625	89,334	98,650	107,565	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	82,358	89,133	99,334	109,141	118,498	124,342	129,561	135,807	140,169

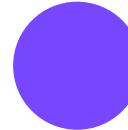
Conclusão



- ✓ Conduzir um teste qui-quadrado para independência entre variáveis categóricas.
- ✓ Formalizar as hipóteses.
- ✓ Definir o nível de confiança e o nível de significância α .
- ✓ Obter o qui-quadrado calculado.
- ✓ Obter os graus de liberdade.
- ✓ Delimitar a região crítica.
- ✓ Obter o valor p.
- ✓ Formalizar a conclusão do teste.



Próxima aula



- Teste F para análise de variância (ANOVA).

Análise Estatística de Dados

AULA 4.6. TESTE F PARA ANOVA

PROF. MÁIRON CHAVES

Nesta aula



- Teste F para ANOVA.

Teste F para ANOVA

A ANOVA (Analysis of Variance) utiliza um teste F para identificar se há variabilidade significativa ao realizar as comparações das médias das n populações. Nos permite comparar mais de duas médias.



Exemplos:



A pontuação média no ENEM entre indivíduos que estudaram em Escolas Estaduais, Escolas Municipais e Escolas Particulares é estatisticamente diferente?



Tomando n indivíduos com pressão alta, existe diferença na pressão média entre os indivíduos que tomaram o Medicamento A, o Medicamento B e o Medicamento Placebo?

Teste F para ANOVA

Contexto



Estamos pesquisando o consumo de uma determinada bebida entre três populações (públicos) em um restaurante: solteiros, casados e divorciados.

O consumo médio com esta bebida oriundo desses três públicos pode ser considerado igual? Ou pelo menos um deles pode ser considerado estatisticamente diferente dos demais?

Teste F para ANOVA

Contexto

Para isso, observamos os gastos com a bebida oriundos de $n_1=17$ solteiros , $n_2=98$ casados e $n_3=15$ divorciados.

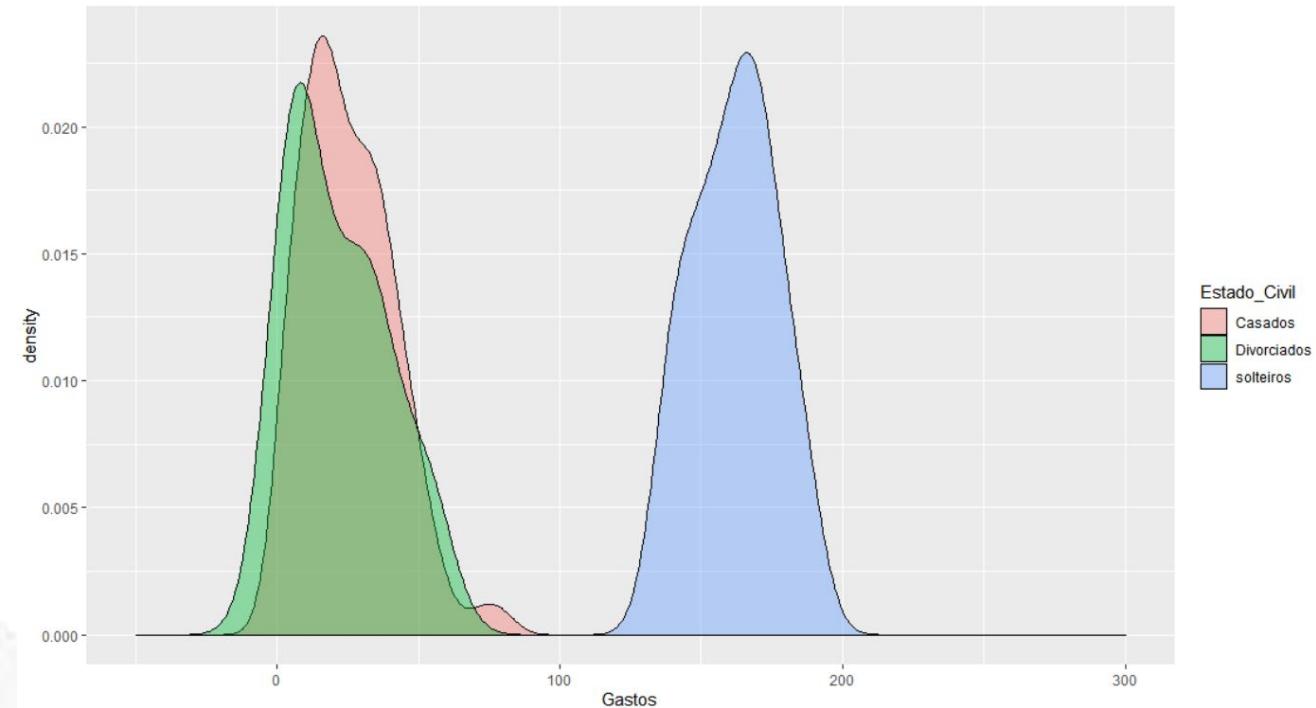


Solteiros:

Distribuição com consumo para cada uma das três populações na amostra coletada

Casados:

Divorciados:



Teste F para ANOVA

Formalizando as hipóteses



H_0 : as médias são iguais ($\mu_1 = \mu_2 = \mu_3 = \dots = \mu_n$).

H_1 : pelo menos uma das médias são diferentes ($\mu_i \neq \mu_j$ para pelo menos um par de médias (i,j)).

Traduzindo pro contexto proposto fica:

H_0 : não há diferença significativa entre o consumo médio em nenhum dos públicos.

H_1 : em pelo menos um dos públicos o consumo médio é significativamente diferente.

Iremos assumir 95% de confiança para conduzir o teste.

Consequentemente, nosso nível de significância será $\alpha = 5\%$.

Teste F para ANOVA

Calculando a estatística de teste | $F_{calculado}$

O F calculado é uma razão entre dois valores.

$$F_{calculado} = \frac{\frac{SS_{entre}}{m - 1}}{\frac{SS_{dentro}}{n - m}}$$

Onde:

m é a quantidade de populações que estão sendo testadas,

n é a quantidade de observações disponíveis,

SSentre é a soma dos quadrados das diferenças entre as médias de cada população em relação média global; e **SSdentro** é a soma dos quadrados das diferenças das observações dentro das populações em relação a média daquela população.

Teste F para ANOVA

Calculando a estatística de teste | $F_{calculado}$

O F calculado é uma razão entre dois valores.

$$SS_{entre} = \sum_{i=1}^m n_i (\bar{Y}_i - \bar{Y})^2$$

Onde:

m é a quantidade de populações.
 n_i é a quantidade de observações da i-ésima população.
 \bar{Y}_i é a média da i-ésima população.
 \bar{Y} é a média global da variável estudada.

$$F_{calculado} = \frac{\frac{SS_{entre}}{m-1}}{\frac{SS_{dentro}}{n-m}}$$

$$SS_{dentro} = \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

Onde:

m é a quantidade de populações.
 n é quantidade de observações dentro da população i.
 Y_{ij} é o valor da j-ésima observação dentro da i-ésima População.
 \bar{Y}_i é a média da i-ésima população.

Onde:

m é a quantidade de populações que estão sendo testadas.

n é a quantidade de observações disponíveis.

SSentre é a soma dos quadrados das diferenças entre as médias de cada população em relação média global.

SSdentro é a soma dos quadrados das diferenças das observações dentro das populações em relação a média daquela população.

Teste F para ANOVA

Calculando a estatística de teste | $F_{calculado}$ | Graus de Liberdade

O F calculado é uma razão entre dois valores.

$$F_{calulado} = \frac{\frac{SS_{entre}}{m - 1}}{\frac{SS_{dentre}}{n - m}}$$

Graus de liberdade do denominador:
Quantidade de observações – Quantidade de Populações.

Onde:

m é a quantidade de populações que estão sendo testadas,

n é a quantidade de observações disponíveis,

SSentre é a soma dos quadrados das diferenças entre as médias de cada população em relação à média global.

SSdentro é a soma dos quadrados das diferenças das observações dentro das populações em relação a média daquela população.

Teste F para ANOVA

Calculando a estatística de teste | $F_{calculado}$

O F calculado é uma razão entre dois valores.

$$F_{calculado} = \frac{\frac{SS_{entre}}{m - 1}}{\frac{SS_{dentro}}{n - m}}$$

$$SS_{entre} = \sum_{i=1}^m n_i (\bar{Y}_i - \bar{Y})^2$$

$$SS_{dentro} = \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

$$F_{calculado} = \frac{\frac{276.693}{3 - 1}}{\frac{32.665}{130 - 3}}$$

$$F_{calculado} = \frac{138.319,50}{257,2}$$

$$F_{calculado} = 538$$

Teste F para ANOVA

Calculando a estatística de teste | $F_{calculado}$

O F calculado é uma razão entre dois valores.

$$F_{calculado} = \frac{\frac{SS_{entre}}{m - 1}}{\frac{SS_{dentro}}{n - m}}$$

$$F_{calculado} = \frac{\frac{276.693}{3 - 1}}{\frac{32.665}{130 - 3}}$$

Tabela ANOVA

Fonte de Variação	Soma dos Quadrados (SS)	Graus de Liberdade	Quadrados médios (MS)	F
Entre populações	SS _{entre}	m-1	$MS_{entre} = \frac{SS_{entre}}{m - 1}$	$\frac{MS_{entre}}{MS_{dentro}}$
Dentro das populações (erro)	SS _{dentro}	n-m	$MS_{dentro} = \frac{SS_{dentro}}{n - m}$	

Tabela ANOVA para os gastos com bebidas nas populações avaliadas

$$F_{calculado} = \frac{138.319,50}{257,2}$$

$$F_{calculado} = 538$$

Fonte de Variação	Soma dos Quadrados (SS)	Graus de Liberdade	Quadrados médios (MS)	F
Entre populações	276.639	2	138.319	
Dentro das populações (erro)	32.665	127	257	538

Teste F para ANOVA

Delimitando a região crítica | Obtendo o $F_{crítico}$

IGTI

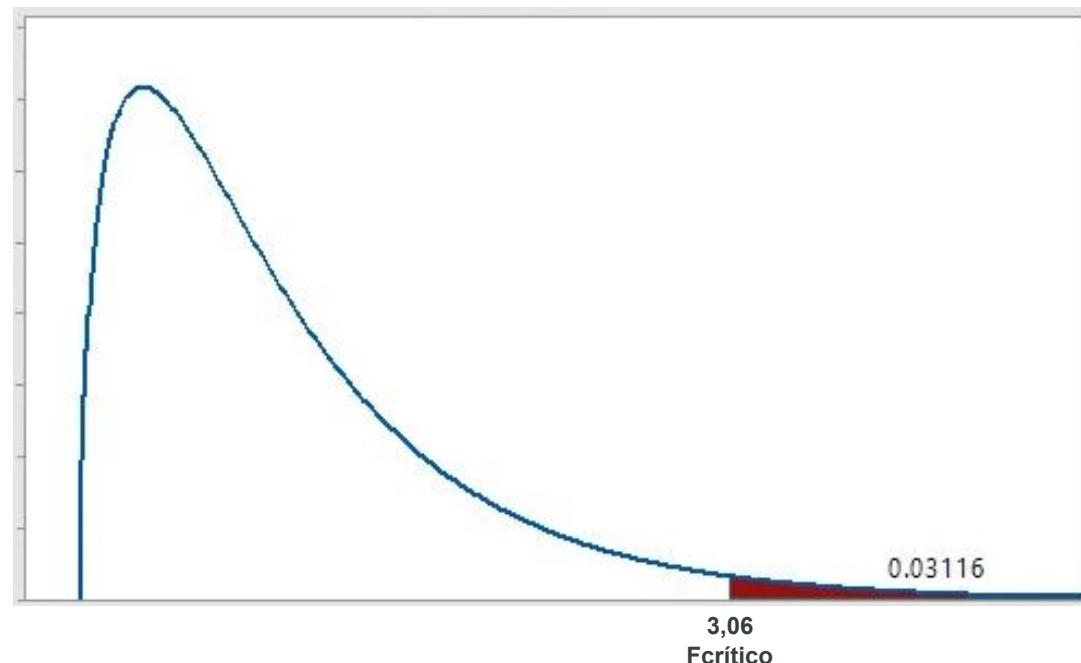
Região crítica de 95% de confiança para uma distribuição F com 2 (numerador) e 127
(denominador) graus de liberdade

R:

```
>qf(p = 0.95,df1=2,df2=127)  
3.0675
```

R:

```
>1-pf(q=538,df1=2,df2=127)  
0.0000
```



Valor $p < \alpha$, então:

A resposta formal fica: com 95% de confiança, há evidências para rejeitar a hipótese nula. Ou seja, pelo menos uma das médias são estatisticamente diferentes.

Teste F para ANOVA

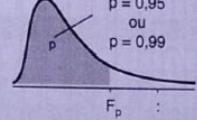
Delimitando a região crítica | Obtendo o $F_{crítico}$

R :

```
>qf(p = 0.95,df1=2,df2=127)  
3.0675
```

R :

```
>1-pf(q=538,df1=2,df2=127)  
0.0000
```

		TABELA III: DISTRIBUIÇÃO F DE FISHER-SNEDECOR													
															
		Fornece os quantis $F_{0.95}$ (em cima) e $F_{0.99}$ (embaixo) em função do nº de g.l. numerador v_1 (coluna) e do nº de g.l. denominador v_2 (linha)													
$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9	10	20	40	60	120	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	248,01	251,14	252,20	253,25	254,31
	4052,18	4999,50	5403,35	5624,58	5763,65	5858,99	5928,36	5981,07	6022,47	6055,85	6208,73	6286,78	6313,03	6339,39	6365,76
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,45	19,47	19,48	19,49	19,50
	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,45	99,47	99,48	99,49	99,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,66	8,59	8,57	8,55	8,53
	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	26,69	26,41	26,32	26,22	26,13
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,80	5,72	5,69	5,66	5,63
	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,02	13,75	13,65	13,56	13,46
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,56	4,46	4,43	4,40	4,37
	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,55	9,29	9,20	9,11	9,02
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,87	3,77	3,74	3,70	3,67
	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,40	7,14	7,06	6,97	6,88
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,44	3,34	3,30	3,27	3,23
	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,16	5,91	5,82	5,74	5,65
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,15	3,04	3,01	2,97	2,93
	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,36	5,12	5,03	4,95	4,86
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	2,94	2,83	2,79	2,75	2,71
	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	4,81	4,57	4,48	4,40	4,31
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,77	2,66	2,62	2,58	2,54
	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,41	4,17	4,08	4,00	3,91
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,12	1,99	1,95	1,90	1,84
	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	2,94	2,69	2,61	2,52	2,42
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,84	1,69	1,64	1,58	1,51
	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,37	2,11	2,02	1,92	1,81
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,75	1,59	1,53	1,47	1,39
	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,20	1,94	1,84	1,73	1,60
120	3,92	4,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,66	1,50	1,43	1,35	1,25
	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,03	1,76	1,66	1,53	1,38
∞	3,84	3,00	2,61	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,57	1,39	1,32	1,22	1,02
	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	1,88	1,59	1,47	1,33	1,03

Observação: O quantil F_p correspondente a v_1 g.l. no numerador e v_2 g.l. no denominador coincide com o inverso do quantil F_{1-p} correspondente a v_2 g.l. no numerador e v_1 g.l. no denominador.

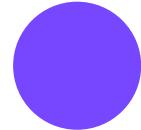
Conclusão



- ✓ Conduzir um teste F para ANOVA.
- ✓ Formalizar as hipóteses.
- ✓ Definir o nível de confiança e o nível de significância α .
- ✓ Obter o qui-quadrado calculado.
- ✓ Obter os graus de liberdade.
- ✓ Delimitar a região crítica.
- ✓ Obter o valor p.
- ✓ Formalizar a conclusão do teste.



Próxima aula



- Estatística Computacional – Teste de Hipótese com o R.

Análise Estatística de Dados

AULA 4.7. ESTATÍSTICA COMPUTACIONAL – TESTE DE HIPÓTESE COM O R

PROF. MÁIRON CHAVES

Nesta aula



- Estatística Computacional – Teste de Hipótese com o R.

Estatística Computacional – Teste de Hipótese com o R



Estatística Computacional – Teste de Hipótese com o R

```
#####
#####      Teste de Hipótese      #####
##   AED - Capítulo 04 - Prof. Máiron Chaves #####
#####

#Copie este código, cole no seu R e execute para ver os resultados

rm(list = ls()) #Limpa memória do R

#####
##### Avaliando a normalidade de uma variável aleatória #####
#####

set.seed(10)

#Gera v.a. que segue distribuição normal com n = 70, média = 40 e desvio padrão =
8

va_normal <- rnorm(n = 70, mean = 25, sd = 8)

#Gera v.a. que segue uma distribuição F (não normal) com n = 15, 2 graus de
liberdade no numerador e 10 graus de liberdade no denominador

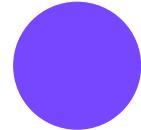
va_nao_normal <- rf(n = 15, df1 = 2, df2 = 10)
```

Conclusão

- ✓ Gerar variáveis aleatórias.
- ✓ Obter valores críticos (quantis) para as distribuições de probabilidades.
- ✓ Obter o valor p.
- ✓ Avaliar a normalidade de uma variável aleatória.
- ✓ Teste t para amostras independentes.
- ✓ Teste t pareado (amostras dependentes).
- ✓ Teste qui-quadrado para independência entre variáveis categóricas.
- ✓ ANOVA.



Próxima aula



- Regressão Linear e Correlação.

Análise Estatística de Dados

CAPÍTULO 5. REGRESSÃO LINEAR E CORRELAÇÃO

PROF. MÁIRON CHAVES

Análise Estatística de Dados

AULA 5.1. REGRESSÃO LINEAR E CORRELAÇÃO

PROF. MÁIRON CHAVES

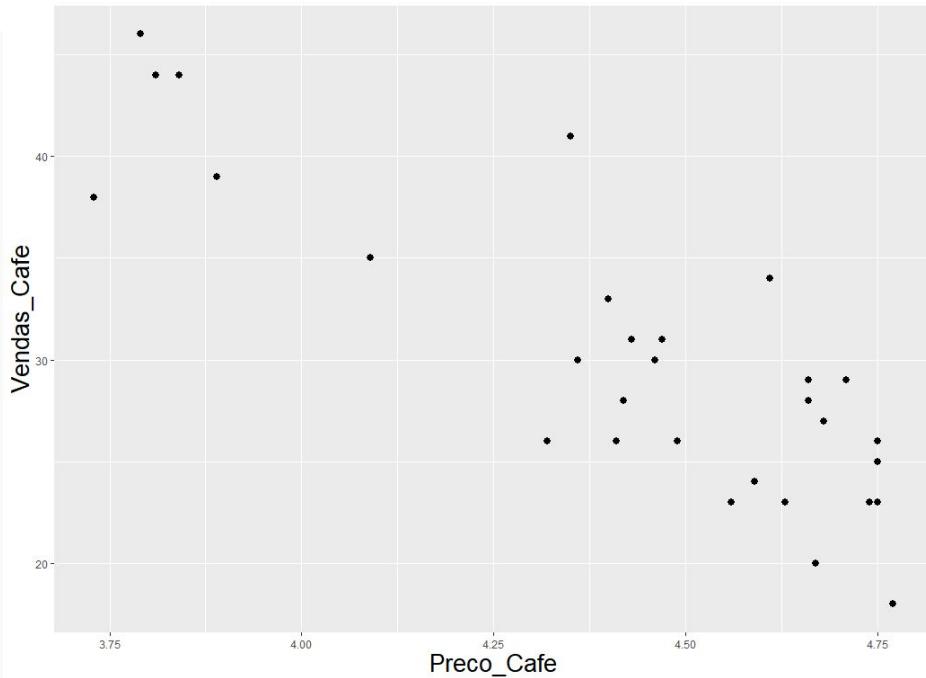
Nesta aula

- Relações Lineares.
- Coeficiente de Correlação Linear.
- Método dos Mínimos Quadrados Ordinários.

Regressão Linear e Correlação

O que é a Regressão Linear

A Regressão Linear permite gerar um modelo matemático através de uma reta que explique a relação linear entre variáveis.



- Como a variável **X** explica a variável **Y** ?
- Posso prever os valores da variável **Y** usando os valores da variável **X** ?

Y: variável resposta (dependente)

X: variável preditora(explicativa, independente)

Regressão Linear e Correlação

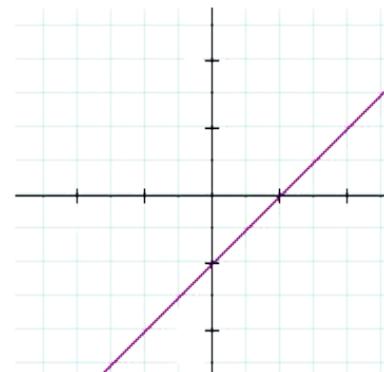
Função de primeiro grau

A Regressão Linear permite gerar um modelo matemático através de uma reta que explique a relação linear entre variáveis.

$$\hat{y} = \beta_0 + (\beta_1 * x_1) + \varepsilon$$

β_0 = Intercepto, onde a reta toca no eixo Y.

β_1 = Coeficiente angular, é o quanto Y varia no aumento unitário em X.



Condição para usar o modelo de regressão linear:

- A variável resposta (Y) deve ser do tipo contínua.
- A relação entre Y e a variável explicativa X deve ser linear.
- Os erros do modelo (ε) devem ser independentes e seguir a distribuição Normal com média igual a zero e variância constante ao longo da reta.

$$\varepsilon \sim N(\mu=0, \sigma=1)$$

Regressão Linear e Correlação

Coeficiente de Correlação Linear de Pearson

Mensura a força da correlação entre um par de variáveis.

X	y
Preco_Cafe	Vendas_Cafe
4.77	18
4.67	20
4.75	23
4.74	23
4.63	23
4.56	23
4.59	24
4.75	25
4.75	26
4.49	26
4.41	26
4.32	26
4.68	27
4.66	28
4.42	28
4.71	29
4.66	29
4.46	30

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]}}$$

$$\rho = -0,84$$

Valor do Coeficiente	Interpretação
Entre 0,10 e 0,29	Correlação positiva fraca
Entre 0,30 a 0,49	Correlação positiva moderada
Entre 0,5 e 1	Correlação positiva alta
Entre -0,10 e -0,29	Correlação negativa fraca
Entre -0,30 a -0,49	Correlação negativa moderada
Entre -0,5 e -1	Correlação negativa alta

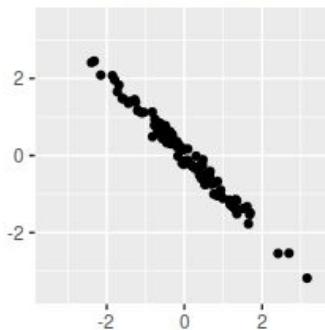
Fonte: (COHEN, *Multiple Commitment in the workplace: an integrative approach*, 2003)

Regressão Linear e Correlação

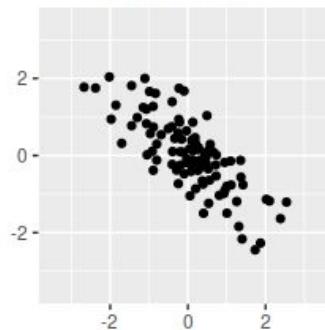
Coeficiente de Correlação Linear de Pearson

Exemplos hipotéticos de correlação e o respectivo gráfico de dispersão

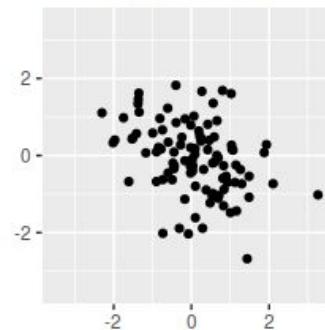
Correlation = -0.99



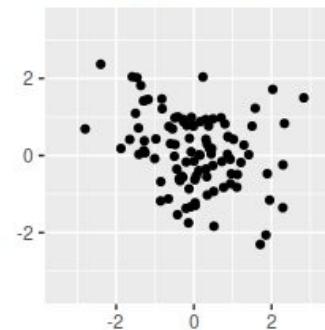
Correlation = -0.75



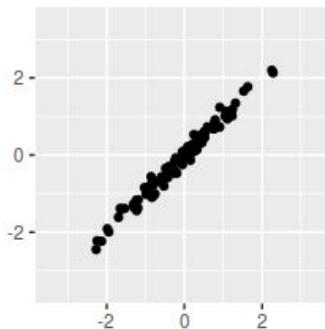
Correlation = -0.50



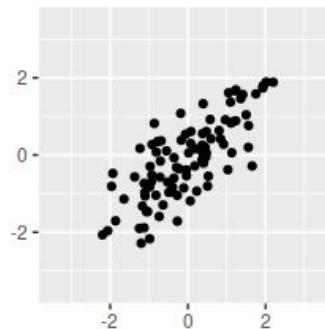
Correlation = -0.25



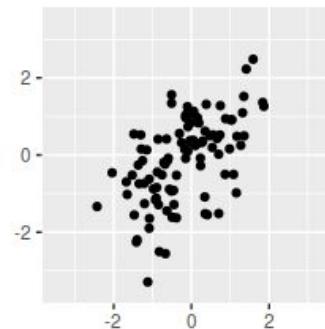
Correlation = 0.99



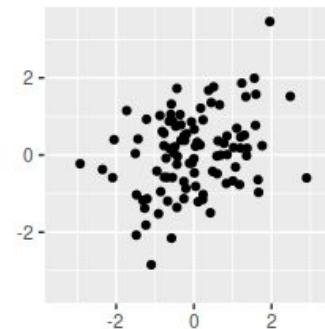
Correlation = 0.75



Correlation = 0.50



Correlation = 0.25

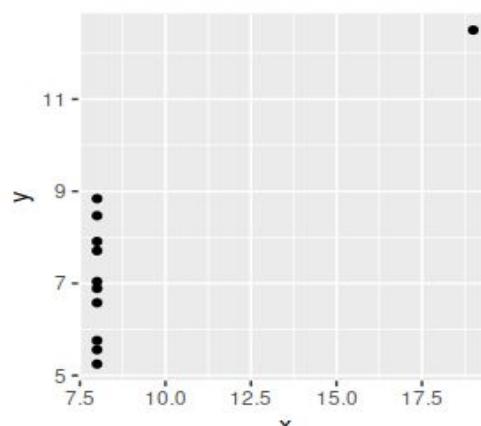
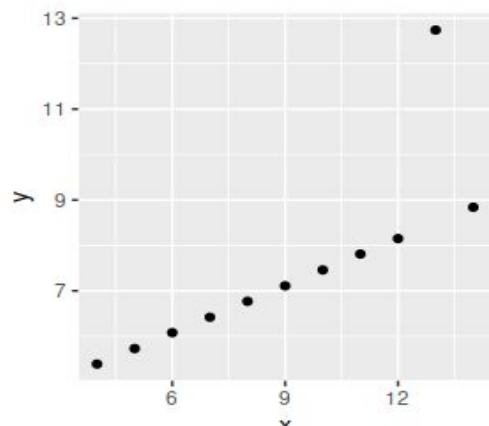
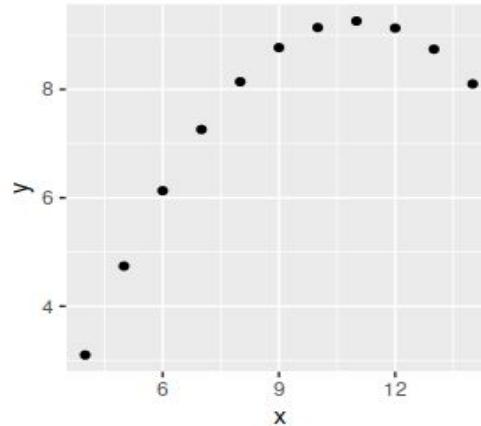
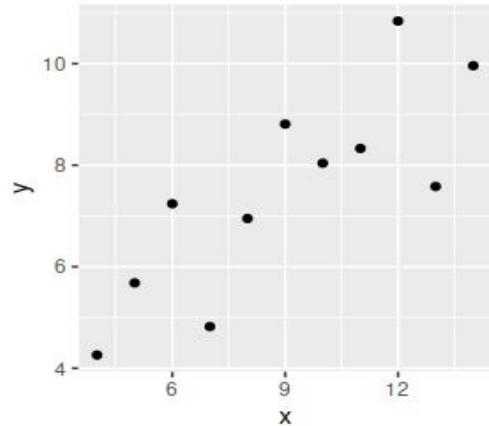


Regressão Linear e Correlação

IGTI

Coeficiente de Correlação Linear de Pearson

Todos os gráficos resultam em um coeficiente de correlação de 0,82

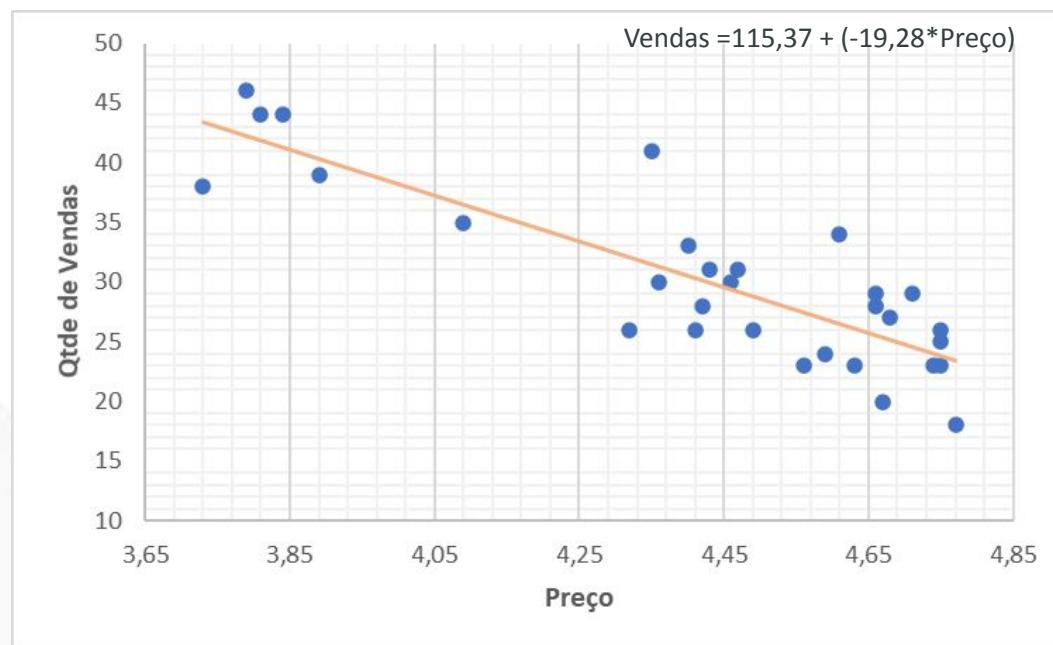


Fonte: <https://otexts.com/fpp2/scatterplots.html>

Regressão Linear e Correlação

Método dos Mínimos Quadrados Ordinário (MQO)

Reta de regressão linear entre o preço do café e suas vendas



Regressão Linear e Correlação

Método dos Mínimos Quadrados Ordinário (MQO)

Exemplo 2:

Imagine que um pesquisador registrou as medidas de 31 árvores. Para cada árvore ele coletou a circunferência (em polegadas), altura (em pés) e o volume (em pés cúbicos).

	circunferencia	altura	volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
29	18.0	80	51.5
30	18.0	80	51.0
31	20.6	87	77.0

Será que existe correlação entre a circunferência e o volume? Será que é possível que predizer o volume baseado na circunferência?

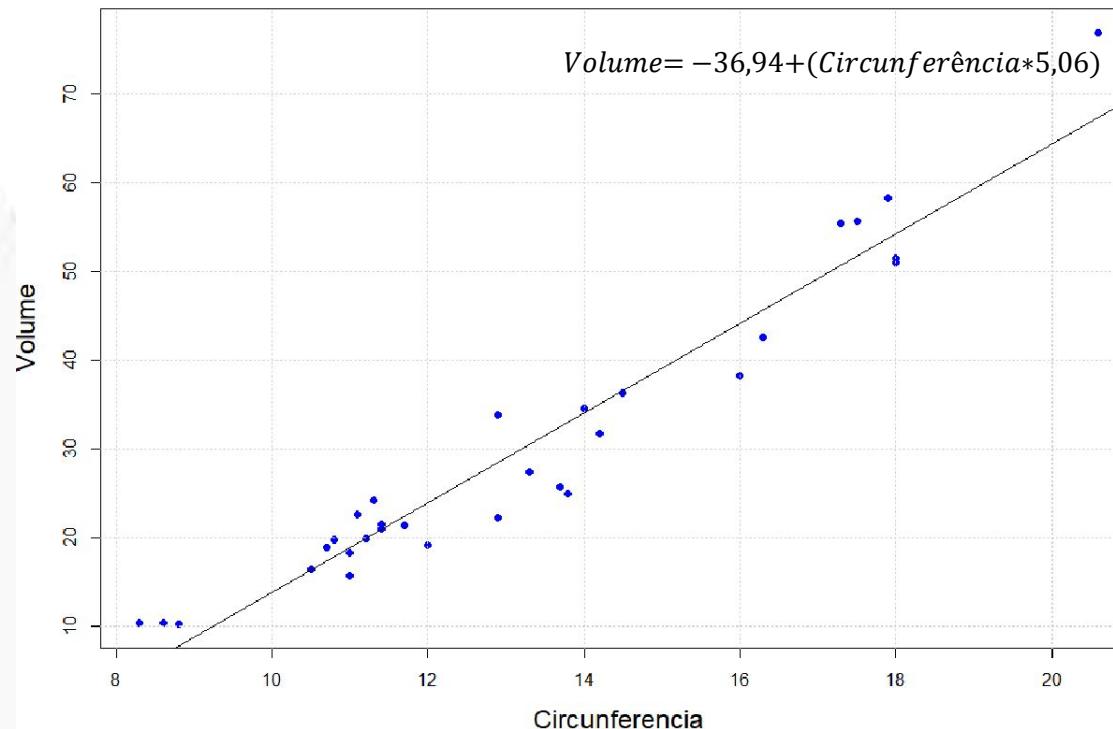
Regressão Linear e Correlação

Método dos Mínimos Quadrados Ordinário (MQO)

Exemplo 2:

Será que existe correlação entre a circunferência e o volume? Será que é possível predizer o volume baseado na circunferência?

Relação entre volume e circunferência $\rho = 0,96$



Regressão Linear e Correlação



Método dos Mínimos Quadrados Ordinário (MQO)

Obtendo os coeficientes da equação de regressão

O método tem como função objetivo minimizar a soma dos erros ao quadrado.

$$SQE = \min \sum_{i=1}^n (\varepsilon_i)^2$$

$$\hat{y} = \beta_0 + (\beta_1 * x_1) + \varepsilon$$

Y	\hat{Y}	ε	$(\varepsilon)^2$
16,68	21,71	-5,03	25,28
11,5	8,01	3,49	12,16
12,03	9,74	2,29	5,25
14,88	7,61	7,27	52,79
13,75	11,85	1,90	3,60
18,11	16,06	2,05	4,21
8	4,81	3,19	10,15
17,83	14,33	3,50	12,23
79,24	69,48	9,76	95,28
21,5	16,78	4,72	22,26
40,33	35,75	4,58	20,96
21	19,25	1,75	3,06
13,5	10,13	3,37	11,35
19,75	16,34	3,41	11,62
24	20,99	3,01	9,07
29	27,32	1,68	2,82
15,35	12,57	2,78	7,72
19	13,21	5,79	33,52
9,5	5,37	4,13	17,09
35,1	38,55	-3,45	11,88
17,9	18,17	-0,27	0,07
52,32	53,67	-1,35	1,81
18,75	21,02	-2,27	5,14
19,83	22,06	-2,23	4,98
10,75	8,62	2,13	4,53
		Σ	388,83

$$\hat{y} = ? + (? * x_1) + \varepsilon$$

Regressão Linear e Correlação

Método dos Mínimos Quadrados Ordinário (MQO)

Obtendo os coeficientes da equação de regressão

O método possui solução analítica: $\hat{\beta} = (X^T X)^{-1} X^T Y$

X =

	Preço Café
1	4,77
1	4,67
1	4,75
1	4,74
1	4,63
1	4,56
1	4,59
1	4,75
1	4,75
1	4,49
1	4,41
1	4,32
1	4,68
1	4,66
1	4,42
1	4,71
1	4,66
1	4,46
1	4,36
1	4,47
1	4,43
1	4,4
1	4,61
1	4,09
1	3,73
1	3,89
1	4,35
1	3,84
1	3,81
1	3,79

Y =

	Vendas
18	
20	
23	
23	
23	
24	
25	
26	
26	
26	
27	
28	
28	
29	
29	
30	
30	
31	
31	
33	
34	
35	
38	
39	
41	
44	
44	
46	

$$(X^T X) \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} X^T Y \end{bmatrix}$$

30	132,79
132,79	590,7807

900
3925,69

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} \cdot \begin{bmatrix} X^T Y \end{bmatrix}$$

6,546997	-1,47157
-1,47157	0,332458

900
3925,69

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 115,3658 \\ -19,2859 \end{bmatrix}$$

R:
lm(Vendas_Cafe ~ Preco_Cafe, data = dados)

Coefficients:
(Intercept) Preco_Cafe
115.37 -19.29

$$\hat{y} = \beta_0 + (\beta_1 * x_1) + \varepsilon$$

$$\hat{y} = ? + (? * x_1) + \varepsilon$$

$$\hat{y} = 115,37 + (-19,29 * x_1)$$



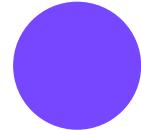
Conclusão



- ✓ Função de primeiro grau.
- ✓ Relações lineares.
- ✓ Interpretando a correlação pelo gráfico de dispersão.
- ✓ Método dos Mínimos Quadrados Ordinários.



Próxima aula



- Regressão Linear Múltipla.

Análise Estatística de Dados

AULA 5.2. REGRESSÃO LINEAR MÚLTIPLA

PROF. MÁIRON CHAVES

Nesta aula



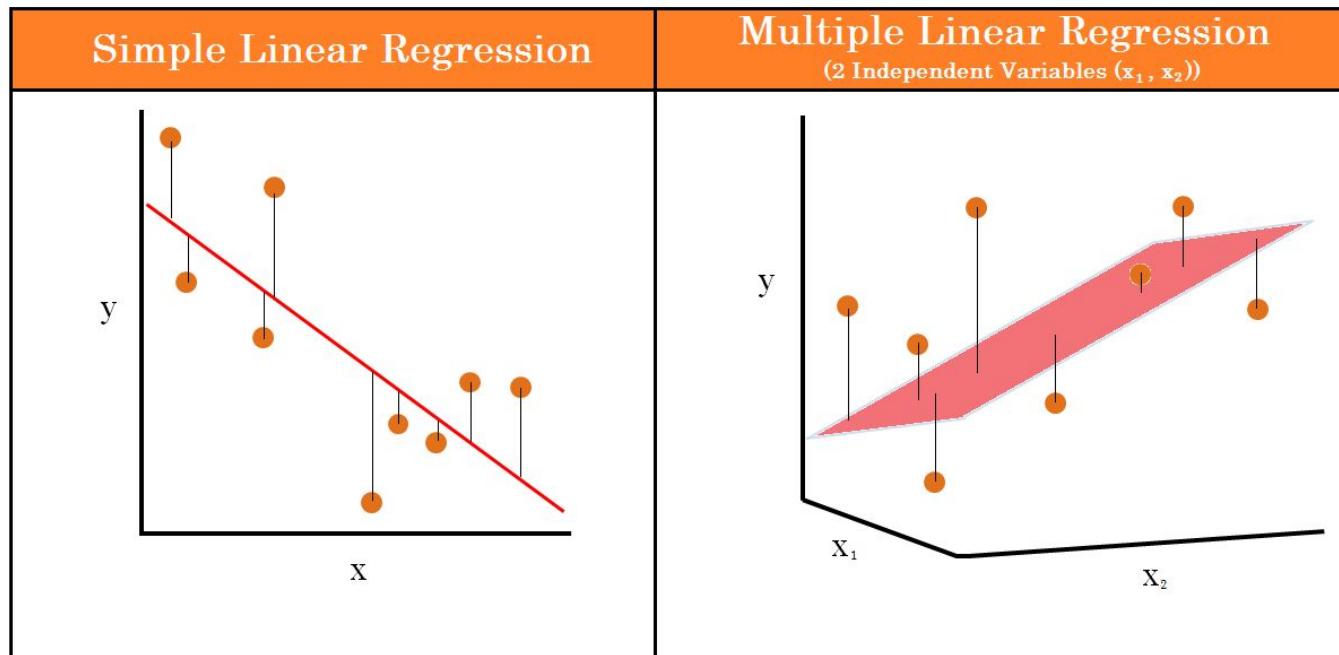
- ❑ Regressão Linear Múltipla.

Regressão Linear Múltipla

Dois ou mais preditores

A regressão linear múltipla pode comportar p variáveis preditoras ao invés de somente uma, como na regressão linear simples.

$$\hat{Y} = \beta_0 + (\beta_1 * X_1) + (\beta_2 * X_2) + \dots + (\beta_p * X_p) + \varepsilon$$



Regressão Linear Múltipla

Dois ou mais preditores

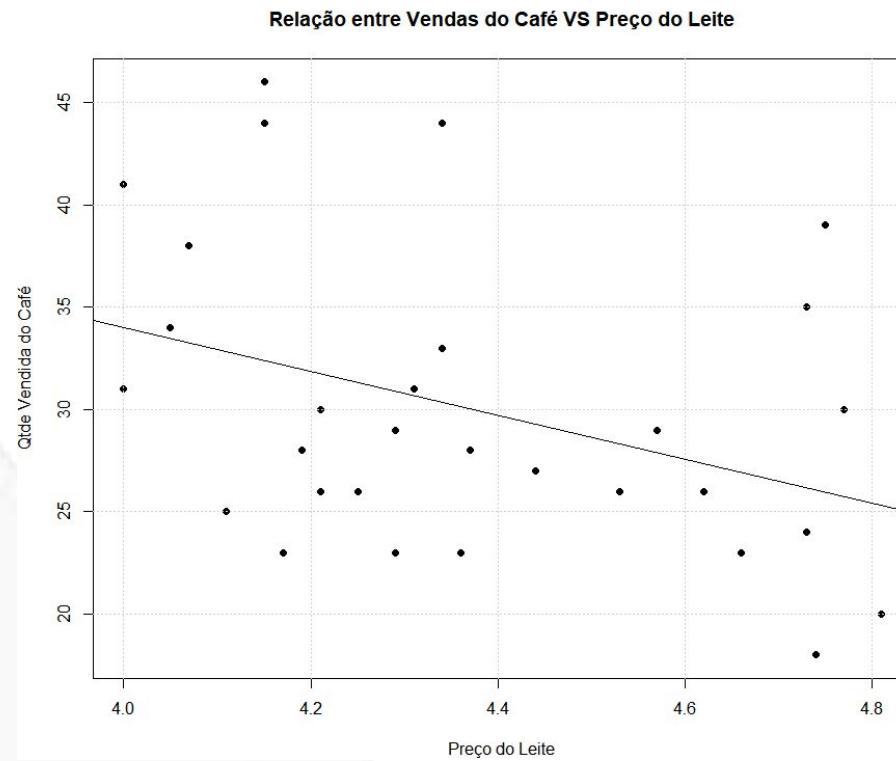
A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.



Regressão Linear Múltipla

Dois ou mais preditores

A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.

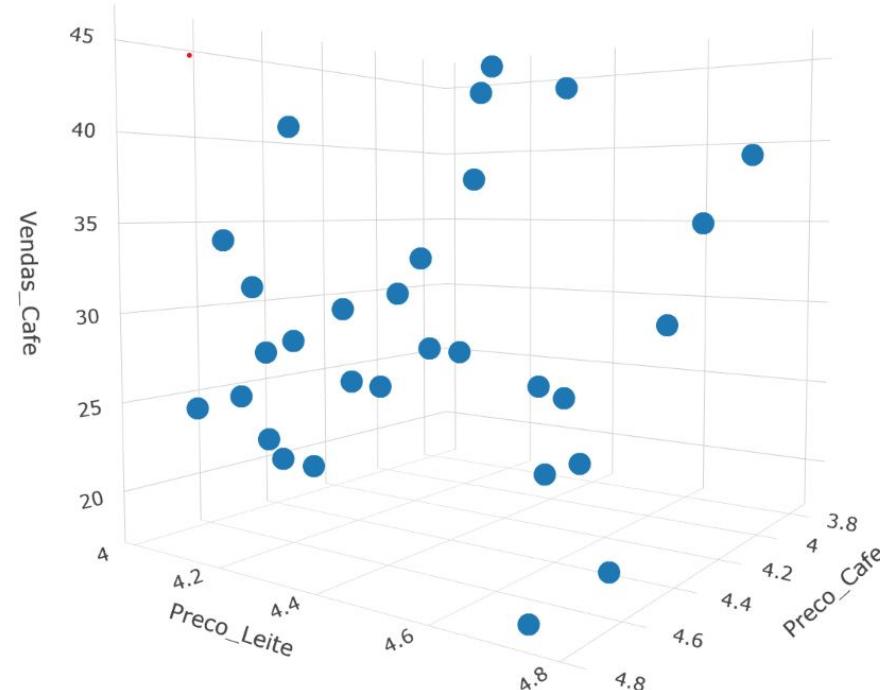


Regressão Linear Múltipla

Dois ou mais preditores

A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.

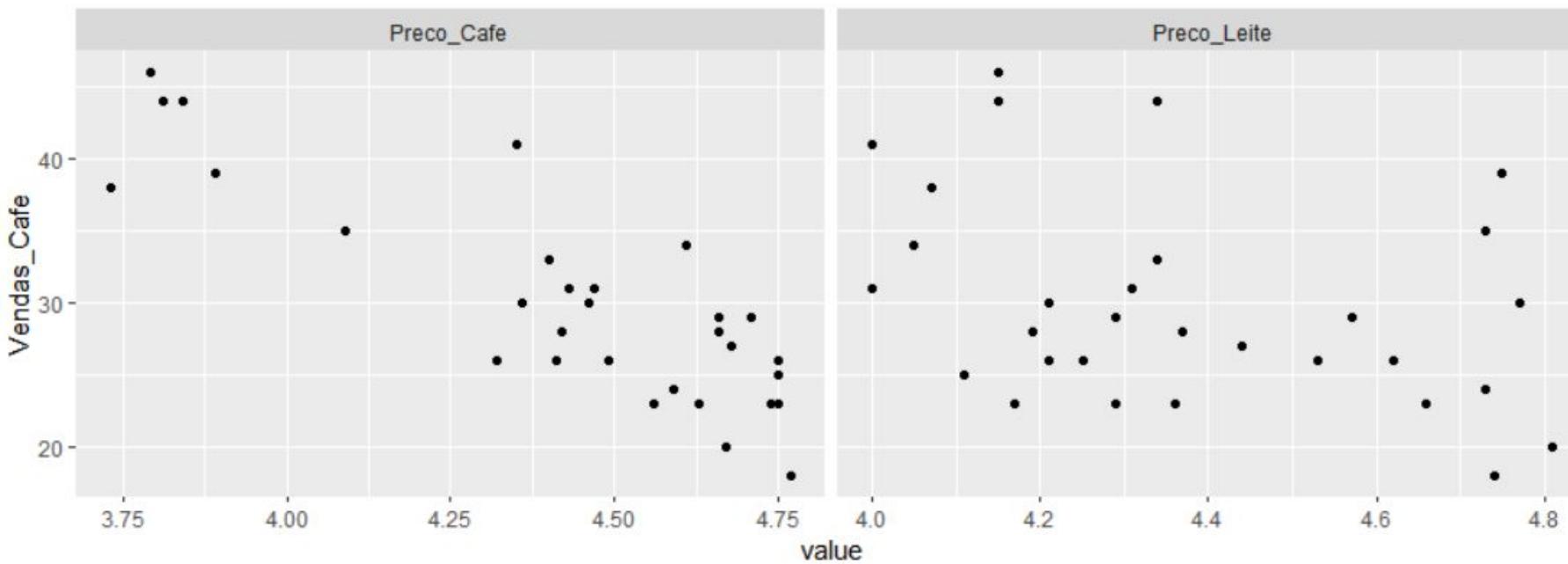
Relação entre Vendas do Café, Preço do Café e Preço do Leite



Regressão Linear Múltipla

Dois ou mais preditores

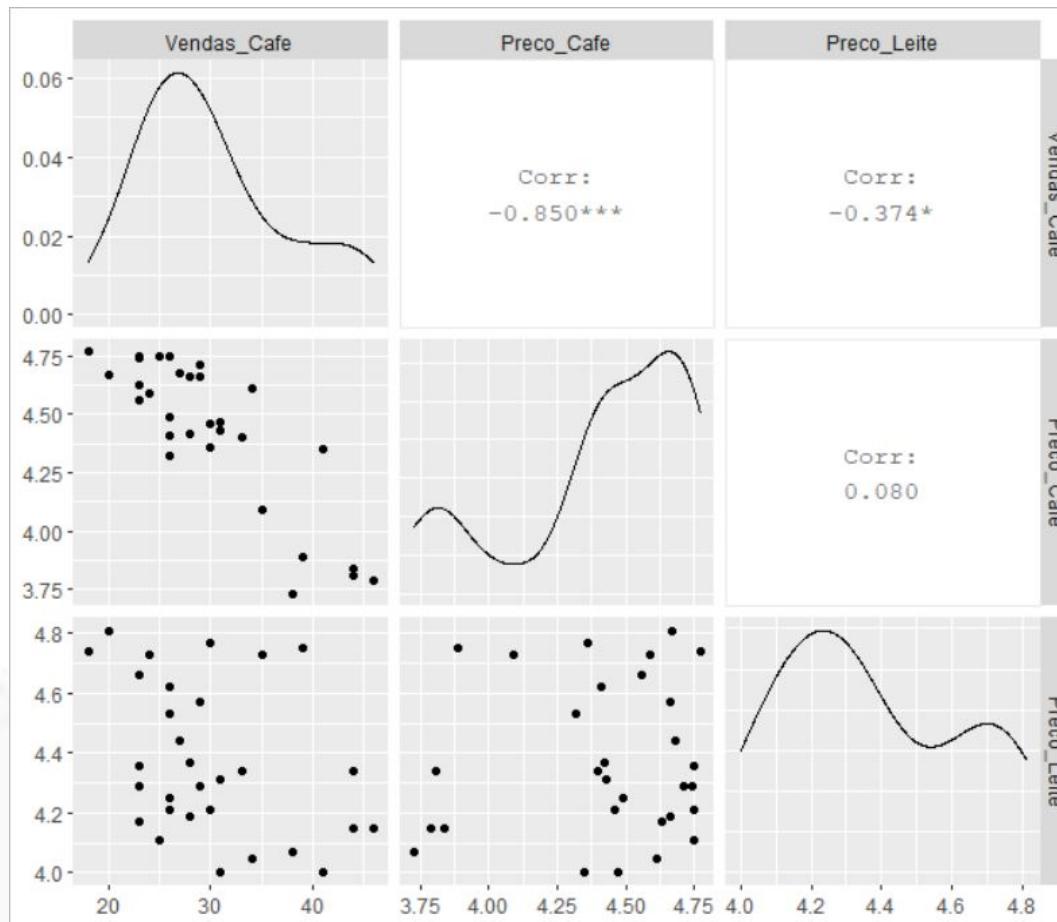
A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.



Regressão Linear Múltipla

Dois ou mais preditores

A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.



Regressão Linear Múltipla



Dois ou mais preditores

A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.

$$Vendas\ do\ Café = 151,31 + (-18,72 * Preço\ do\ Café) + (-8,78 * Preço\ do\ Leite)$$

Regressão Linear Múltipla

Dois ou mais preditores

A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.

Exemplo 2:

Imagine que um pesquisador registrou as medidas de 31 árvores. Para cada árvore ele coletou a circunferência (em polegadas), altura (em pés) e o volume (em pés cúbicos).

	circunferencia	altura	volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
29	18.0	80	51.5
30	18.0	80	51.0
31	20.6	87	77.0

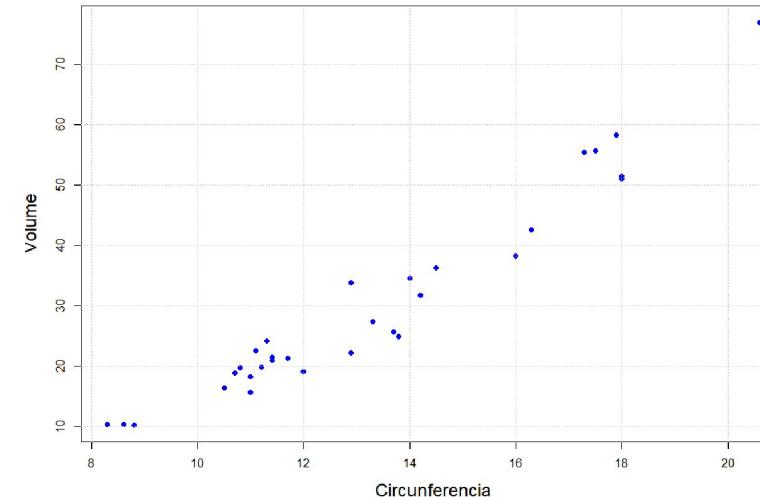
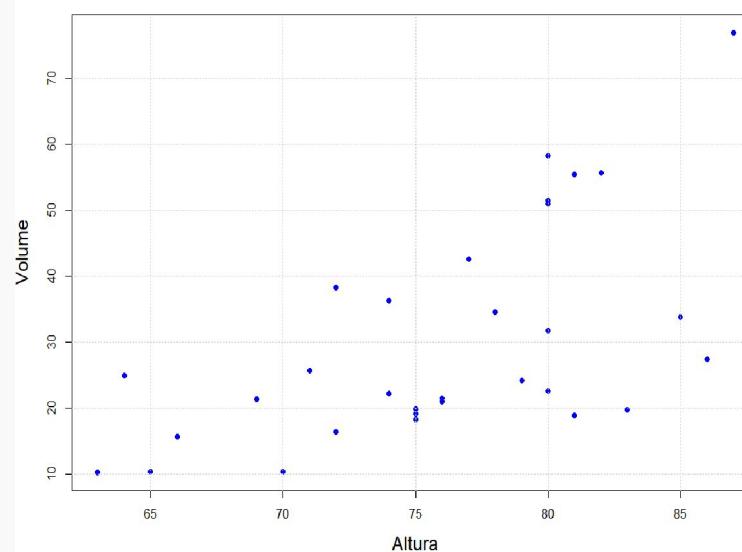
Regressão Linear Múltipla

Dois ou mais preditores

A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.

Exemplo 2:

Imagine que um pesquisador registrou as medidas de 31 árvores. Para cada árvore ele coletou a circunferência (em polegadas), altura (em pés) e o volume (em pés cúbicos).



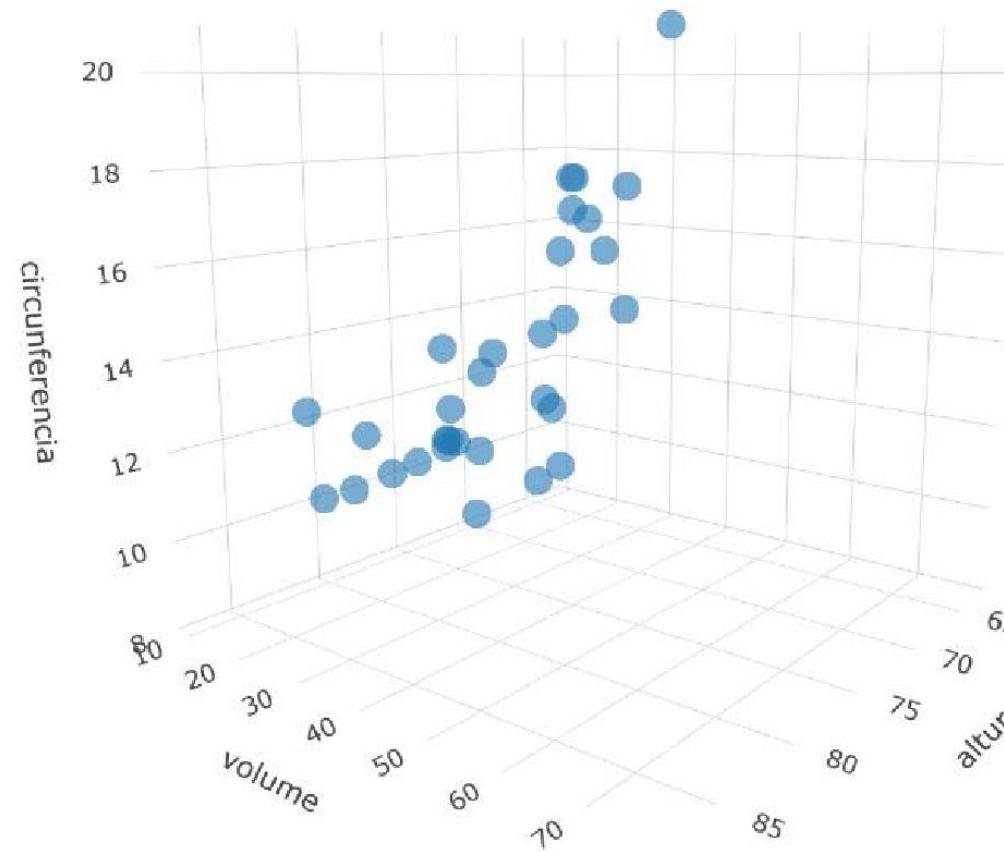
Regressão Linear Múltipla

Dois ou mais preditores

A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.

Exemplo 2:

Imagine que um pesquisador registrou as medidas de 31 árvores. Para cada árvore ele coletou a circunferência (em polegadas), altura (em pés) e o volume (em pés cúbicos).



Regressão Linear Múltipla



Dois ou mais preditores

A regressão linear múltipla pode comportar p variáveis preditoras, ao invés de somente uma como na regressão linear simples.

Exemplo 2:

Imagine que um pesquisador registrou as medidas de 31 árvores. Para cada árvore ele coletou a circunferência (em polegadas), altura (em pés) e o volume (em pés cúbicos).

$$Volume = -57,98 + (Circunferência * 4,73) + (Altura * 0,33)$$

Qual será o volume estimado para uma árvore cuja circunferência seja **11** polegadas e a altura **66** pés?

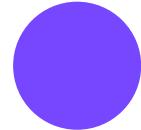
$$Volume = -57,98 + (11 * 4,73) + (66 * 0,33)$$

$$Volume = 15,83$$

Conclusão



- ✓ Regressão Linear Múltipla.
- ✓ Como apresentar visualmente uma Regressão Múltipla.



Próxima aula



- Utilizando variável categórica em um modelo de Regressão Linear.

Análise Estatística de Dados

AULA 5.3. UTILIZANDO VARIÁVEIS CATEGÓRICAS EM UM MODELO DE REGRESSÃO LINEAR

PROF. MÁIRON CHAVES

Nesta aula



- Utilizando variáveis categóricas em um modelo de Regressão Linear.

Utilizando variáveis categóricas em um modelo de Regressão Linear



Variável com dois níveis categóricos

Ainda no nosso contexto das vendas do café, vamos supor que desejamos adicionar uma nova variável indicando em quais dias o café estava em promoção.

Promocao	Preco_Cafe	Vendas_Cafe
Nao	4,77	18
Nao	4,67	20
Nao	4,75	23
Nao	4,74	23
Nao	4,63	23
Nao	4,56	23
Nao	4,59	24
Nao	4,75	25
Sim	4,75	26
Nao	4,49	26
Sim	4,41	26
Nao	4,32	26
Nao	4,68	27
Sim	4,66	28
Sim	4,42	28
Nao	4,71	29
Sim	4,66	29
Sim	4,46	30
Sim	4,36	30
Nao	4,47	31
Nao	4,43	31
Sim	4,4	33
Sim	4,61	34
Sim	4,09	35
Nao	3,73	38
Sim	3,89	39
Sim	4,35	41
Sim	3,84	44
Sim	3,81	44
Sim	3,79	46

Utilizando variáveis categóricas em um modelo de Regressão Linear

Variável com dois níveis categóricos

Ainda no nosso contexto das vendas do café, vamos supor que desejamos adicionar uma nova variável indicando em quais dias o café estava em promoção.

Promoção_Binária	Promocao	Preco_Cafe	Vendas_Cafe
0	Nao	4,77	18
0	Nao	4,67	20
0	Nao	4,75	23
0	Nao	4,74	23
0	Nao	4,63	23
0	Nao	4,56	23
0	Nao	4,59	24
0	Nao	4,75	25
1	Sim	4,75	26
0	Nao	4,49	26
1	Sim	4,41	26
0	Nao	4,32	26
0	Nao	4,68	27
1	Sim	4,66	28
1	Sim	4,42	28
0	Nao	4,71	29
1	Sim	4,66	29
1	Sim	4,46	30
1	Sim	4,36	30
0	Nao	4,47	31
0	Nao	4,43	31
1	Sim	4,4	33
1	Sim	4,61	34
1	Sim	4,09	35
0	Nao	3,73	38
1	Sim	3,89	39
1	Sim	4,35	41
1	Sim	3,84	44
1	Sim	3,81	44
1	Sim	3,79	46

Utilizando variáveis categóricas em um modelo de Regressão Linear



Variável com m níveis categóricos

Medindo o efeito do dia da semana nas vendas.

Promoção_Binária	Dia_da_Semana	Preco_Cafe	Vendas_Cafe
0	Segunda	4,77	18
0	Terça	4,67	20
0	Quarta	4,75	23
0	Quinta	4,74	23
0	Sexta	4,63	23
0	Sábado	4,56	23
0	Domingo	4,59	24
0	Segunda	4,75	25
1	Terça	4,75	26
0	Quarta	4,49	26
1	Quinta	4,41	26
0	Sexta	4,32	26
0	Sábado	4,68	27
1	Domingo	4,66	28
1	Segunda	4,42	28
0	Terça	4,71	29
1	Quarta	4,66	29
1	Quinta	4,46	30
1	Sexta	4,36	30
0	Sábado	4,47	31
0	Domingo	4,43	31
1	Segunda	4,4	33
1	Terça	4,61	34
1	Quarta	4,09	35
0	Quinta	3,73	38
1	Sexta	3,89	39
1	Sábado	4,35	41
1	Domingo	3,84	44
1	Segunda	3,81	44
1	Terça	3,79	46

Utilizando variáveis categóricas em um modelo de Regressão Linear



Variável com m níveis categóricos

Medindo o efeito do dia da semana nas vendas.

Criando m-1 variáveis binárias para o Dia da Semana

Promoção_Binária	Dia_da_Semana	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Preço_Café	Vendas_Café
0	Segunda	1	0	0	0	0	0	4,77	18
0	Terça	0	1	0	0	0	0	4,67	20
0	Quarta	0	0	1	0	0	0	4,75	23
0	Quinta	0	0	0	1	0	0	4,74	23
0	Sexta	0	0	0	0	1	0	4,63	23
0	Sábado	0	0	0	0	0	1	4,56	23
0	Domingo	0	0	0	0	0	0	4,59	24
0	Segunda	1	0	0	0	0	0	4,75	25
1	Terça	0	1	0	0	0	0	4,75	26
0	Quarta	0	0	1	0	0	0	4,49	26
1	Quinta	0	0	0	1	0	0	4,41	26
0	Sexta	0	0	0	0	1	0	4,32	26
0	Sábado	0	0	0	0	0	1	4,68	27
1	Domingo	0	0	0	0	0	0	4,66	28
1	Segunda	1	0	0	0	0	0	4,42	28
0	Terça	0	1	0	0	0	0	4,71	29
1	Quarta	0	0	1	0	0	0	4,66	29
1	Quinta	0	0	0	1	0	0	4,46	30
1	Sexta	0	0	0	0	1	0	4,36	30
0	Sábado	0	0	0	0	0	1	4,47	31
0	Domingo	0	0	0	0	0	0	4,43	31
1	Segunda	1	0	0	0	0	0	4,4	33
1	Terça	0	1	0	0	0	0	4,61	34
1	Quarta	0	0	1	0	0	0	4,09	35
0	Quinta	0	0	0	1	0	0	3,73	38
1	Sexta	0	0	0	0	1	0	3,89	39
1	Sábado	0	0	0	0	0	1	4,35	41
1	Domingo	0	0	0	0	0	0	3,84	44
1	Segunda	1	0	0	0	0	0	3,81	44
1	Terça	0	1	0	0	0	0	3,79	46

Utilizando variáveis categóricas em um modelo de Regressão Linear



Variável com m níveis categóricos

Medindo o efeito do dia da semana e das promoções nas vendas.

Dataset final

Promoção_Binária	Segunda	Terça	Quarta	Quinta	Sexta	Sabado	Preco_Cafe	Vendas_Cafe
0	1	0	0	0	0	0	4,77	18
0	0	1	0	0	0	0	4,67	20
0	0	0	1	0	0	0	4,75	23
0	0	0	0	1	0	0	4,74	23
0	0	0	0	0	1	0	4,63	23
0	0	0	0	0	0	1	4,56	23
0	0	0	0	0	0	0	4,59	24
0	1	0	0	0	0	0	4,75	25
1	0	1	0	0	0	0	4,75	26
0	0	0	1	0	0	0	4,49	26
1	0	0	0	1	0	0	4,41	26
0	0	0	0	0	1	0	4,32	26
0	0	0	0	0	0	1	4,68	27
1	0	0	0	0	0	0	4,66	28
1	1	0	0	0	0	0	4,42	28
0	0	1	0	0	0	0	4,71	29
1	0	0	1	0	0	0	4,66	29
1	0	0	0	1	0	0	4,46	30
1	0	0	0	0	1	0	4,36	30
0	0	0	0	0	0	1	4,47	31
0	0	0	0	0	0	0	4,43	31
1	1	0	0	0	0	0	4,4	33
1	0	1	0	0	0	0	4,61	34
1	0	0	1	0	0	0	4,09	35
0	0	0	0	1	0	0	3,73	38
1	0	0	0	0	1	0	3,89	39
1	0	0	0	0	0	1	4,35	41
1	0	0	0	0	0	0	3,84	44
1	1	0	0	0	0	0	3,81	44
1	0	1	0	0	0	0	3,79	46

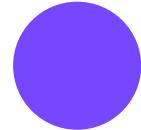
Conclusão



- ✓ Utilizando variável categórica com dois níveis.
- ✓ Utilizando variável categórica com m níveis.

Próxima aula

- Explicação vs. Predição.



Análise Estatística de Dados

AULA 5.4. EXPLICAÇÃO VS. PREDIÇÃO

PROF. MÁIRON CHAVES

Nesta aula



- ❑ Explicação vs. Predição.

Explicação vs. Predição

Inferência e Modelagem Preditiva

Um modelo de Regressão Linear pode ser utilizada para fins inferenciais, ou seja, podemos extrapolar conclusões para a população a partir da amostra. E também podemos utiliza-lo para modelagem preditiva.



Inferência

Objetivo

- Modelar a relação entre uma variável resposta e uma ou mais variáveis preditoras.

Responde a Pergunta

- Qual o impacto das variáveis preditoras na variável resposta?

Exemplo

- Em quantas unidades as vendas diminuem a cada real que aumenta no preço do café?

Predição

Objetivo

- Desenvolver um modelo para prever Y com alta precisão.

Responde a Pergunta

- Como posso prever com precisão novos pontos de dados?

Exemplo

- Qual será a quantidade vendida a um determinado preço?

Explicação vs. Predição



Inferência e Modelagem Preditiva

Explicação:

β_0

β_1

β_2

β_3

β_0 = Intercepto = É o valor que a variável resposta assume quando os preditores estão zerados.

β_1 = Preço do Café = Mantendo as demais variáveis constantes, para cada aumento unitário nessa variável, ou seja, a cada real aumentado no preço do café, as vendas caem em média 16,11 unidades.

β_2 = Promoção = Mantendo as demais variáveis constantes, quando o Café está em promoção, vende em média 4,14 unidades a mais em relação a quando não está em promoção.

β_3 = Preço do Leite = Mantendo as demais variáveis constantes para cada aumento unitário nessa variável, ou seja, a cada real aumentando no preço do Leite, as vendas do café caem em média 8,71 unidades.

O R^2 da modelagem foi de 87,25%, ou seja, o modelo estatístico criado consegue explicar 87,25% da variabilidade das vendas do Café.

Um modelo estatístico é uma representação matemática da realidade!

Explicação vs. Predição

Inferência e Modelagem Preditiva

Predição:

β_0

β_1

β_2

β_3



IGTI

Explicação vs. Predição

Inferência e Modelagem Preditiva

Predição:

β_0

β_1

β_2

β_3

Qual será a venda estimada caso coloquemos o café em promoção, ao valor de R\$2,25, e o leite acima de seu preço médio, custando R\$5,00?



IGTI

Explicação vs. Predição

Inferência e Modelagem Preditiva

Predição:

β_0

β_1

β_2

β_3



Qual será a venda estimada caso coloquemos o café em promoção, ao valor de R\$2,25, e o leite acima de seu preço médio, custando R\$5,00 ?

Qual será a venda estimada caso coloquemos o café sem promoção, ao valor de R\$4,37, e o leite em seu preço médio, R\$4,42?

iGTD

Explicação vs. Predição



Inferência e Modelagem Preditiva

Predição:

Qual será a venda estimada caso coloquemos o café em promoção, ao valor de R\$2,25, e o leite acima de seu preço médio, custando R\$5,00?

$$Vendas = 137,37 + (-16,11 * \text{Preço do Café}) + (4,14 * \text{Promoção}) + (-8,71 * \text{Preço do Leite})$$

$$Vendas = 137,37 + (-16,11 * 2,25) + (4,14 * 1) + (-8,71 * 5,00)$$

$$Vendas = 62$$

```
R :  
  
modelo <- lm(Vendas_Cafe ~ Preco_Cafe + Promocao + Preco_Leite , data = dados)  
  
novos_dados <- data.frame( Preco_Cafe = 2,25 ,Promocao = 'Sim' , Preco_Leite = 5,00)  
  
predict(object = modelo, newdata = novos_dados)
```

Explicação vs. Predição



Inferência e Modelagem Preditiva

Predição:

Qual será a venda estimada caso coloquemos o café sem promoção, ao valor de R\$4,37, e o leite em seu preço médio, R\$4,42?

$$Vendas = 137,37 + (-16,11 * \text{Preço do Café}) + (4,14 * \text{Promoção}) + (-8,71 * \text{Preço do Leite})$$

$$Vendas = 137,37 + (-16,11 * 4,37) + (4,14 * 0) + (-8,71 * 4,42)$$

$$Vendas = 29$$



```
modelo <- lm(Vendas_Cafe ~ Preco_Cafe + Promocao + Preco_Leite , data = dados)

novos_dados <- data.frame( Preco_Cafe = 4,37, Promocao = 'Nao', Preco_Leite = 4,42)

predict(object = modelo, newdata = novos_dados)
```

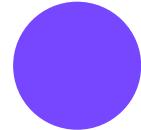
Conclusão



- ✓ Utilizando a Regressão Linear para explicação (inferência).
- ✓ Utilizando a Regressão Linear para predição (modelagem preditiva).



Próxima aula



- Diagnóstico do ajuste do Modelo de Regressão Linear.

Análise Estatística de Dados

AULA 5.5. DIAGNÓSTICO DO AJUSTE DO MODELO DE REGRESSÃO LINEAR

PROF. MÁIRON CHAVES

Nesta aula

- Teste t de Student para os coeficientes Betas.
- Teste F para ajuste geral do modelo de regressão.
- R² e R² Ajustado.
- O que é o resíduo .
- Diagnóstico de normalidade.
- Diagnóstico de homocedasticidade.

Diagnóstico do ajuste do modelo de Regressão Linear



Resumo do modelo de regressão fornecido pelo R

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	137.37	10.83	12.68	0.0000000000012	***
Preco_Cafe	-16.12	1.65	-9.79	0.0000000003258	***
PromoçãoSim	4.15	1.04	3.99	0.00048	***
Preco_Leite	-8.71	1.90	-4.58	0.00010	***

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1

Residual standard error: 2.61 on 26 degrees of freedom

Multiple R-squared: 0.886, Adjusted R-squared: 0.872

F-statistic: 67.1 on 3 and 26 DF, p-value: 0.0000000000225

Diagnóstico do ajuste do modelo de Regressão Linear



Test t de Student para os coeficientes betas

Test t:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

Ou seja

H0: a variável preditora não tem relação significativa com a variável resposta.

H1: a variável preditora tem relação significativa com a variável resposta.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	137.37	10.83	12.68	0.000000000012	***
Preco_Cafe	-16.12	1.65	-9.79	0.0000000003258	***
PromocaoSim	4.15	1.04	3.99	0.00048	***
Preco_Leite	-8.71	1.90	-4.58	0.00010	***

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’
	1				

Residual standard error: 2.61 on 26 degrees of freedom

Multiple R-squared: 0.886, Adjusted R-squared: 0.872

F-statistic: 67.1 on 3 and 26 DF, p-value: 0.0000000000225

Diagnóstico do ajuste do modelo de Regressão Linear



Test F para ajuste geral do modelo

Teste F:

H0: O modelo de regressão **não é** válido

H1: O modelo de regressão **é** válido

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	137.37	10.83	12.68	0.00000000000012	***
Preco_Cafe	-16.12	1.65	-9.79	0.0000000003258	***
PromocaoSim	4.15	1.04	3.99	0.00048	***
Preco_Leite	-8.71	1.90	-4.58	0.00010	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.61 on 26 degrees of freedom

Multiple R-squared: 0.886 Adjusted R-squared: 0.872

F-statistic: 67.1 on 3 and 26 DF, p-value: 0.0000000000225

Diagnóstico do ajuste do modelo de Regressão Linear



Avaliando o quanto da variação da resposta variável o modelo ajustado consegue explicar

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$R^2_{ajustado} = 1 - \left(\frac{n-1}{n-k} \right) (1 - R^2)$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	137.37	10.83	12.68	0.0000000000012 ***
Preco_Cafe	-16.12	1.65	-9.79	0.0000000003258 ***
PromocaoSim	4.15	1.04	3.99	0.00048 ***
Preco_Leite	-8.71	1.90	-4.58	0.00010 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.61 on 26 degrees of freedom

Multiple R-squared: 0.886, Adjusted R-squared: 0.872

F-statistic: 67.1 on 3 and 26 DF, p-value: 0.0000000000225

Para comparar modelos o R2 ajustado deve ser utilizado, pois penaliza a adição de variáveis!

Diagnóstico do ajuste do modelo de Regressão Linear



Resíduo (erro)

O resíduo é o erro de predição.

$$\varepsilon = y - \hat{y}$$

Vendas_Cafe	Predicao	Residuos
18	19.19589	-1.195890576
20	20.19798	-0.197978235
23	22.82836	0.171643614
23	23.59930	-0.599295399
23	26.41762	-3.417615962
23	23.27762	-0.277622271
24	22.18431	1.815685228
25	25.00605	-0.006051994
26	28.28236	-2.282362378
26	27.97733	-1.977332768
26	30.19121	-4.191205838
26	28.27845	-2.278445815
27	23.25978	3.740216509
28	29.90724	-1.907236212
28	32.20772	-4.207717203
29	24.08285	4.917151872
29	26.59714	2.402861112
30	32.95671	-2.956705420
30	29.69051	0.309490313

Diagnóstico do ajuste do modelo de Regressão Linear

Resíduo (erro)

O resíduo é o erro de predição.

$$\varepsilon = y - \hat{y}$$

Vendas_Cafe	Predicao	Residuos
18	19.19589	-1.195890576
20	20.19798	-0.197978235
23	22.82836	0.171643614
23	23.59930	-0.599295399
23	26.41762	-3.417615962
23	23.27762	-0.277622271
24	22.18431	1.815685228
25	25.00605	-0.006051994
26	28.28236	-2.282362378
26	27.97733	-1.977332768
26	30.19121	-4.191205838
26	28.27845	-2.278445815
27	23.25978	3.740216509
28	29.90724	-1.907236212
28	32.20772	-4.207717203
29	24.08285	4.917151872
29	26.59714	2.402861112
30	32.95671	-2.956705420
30	29.69051	0.309490313



Os resíduos devem seguir uma distribuição normal com média zero e variância constante.

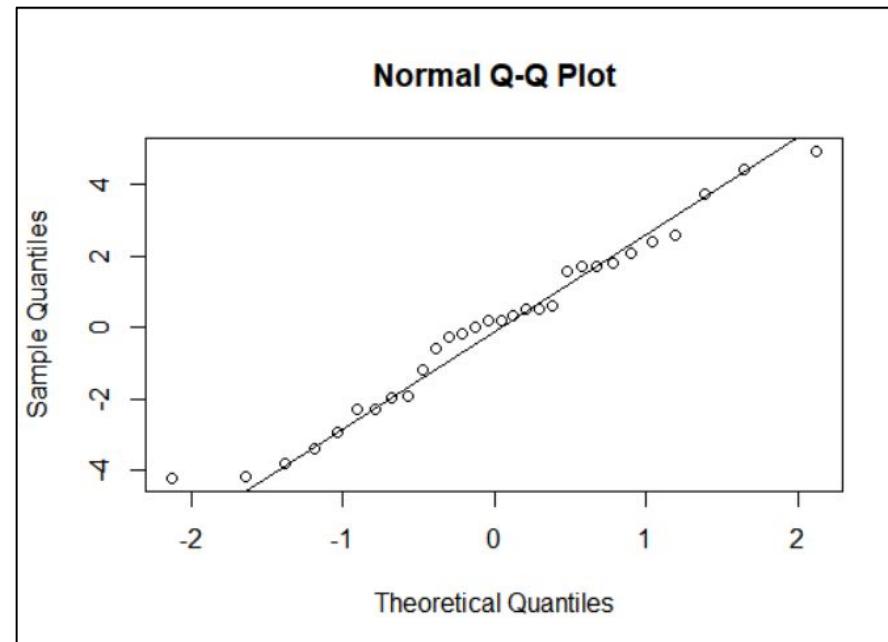
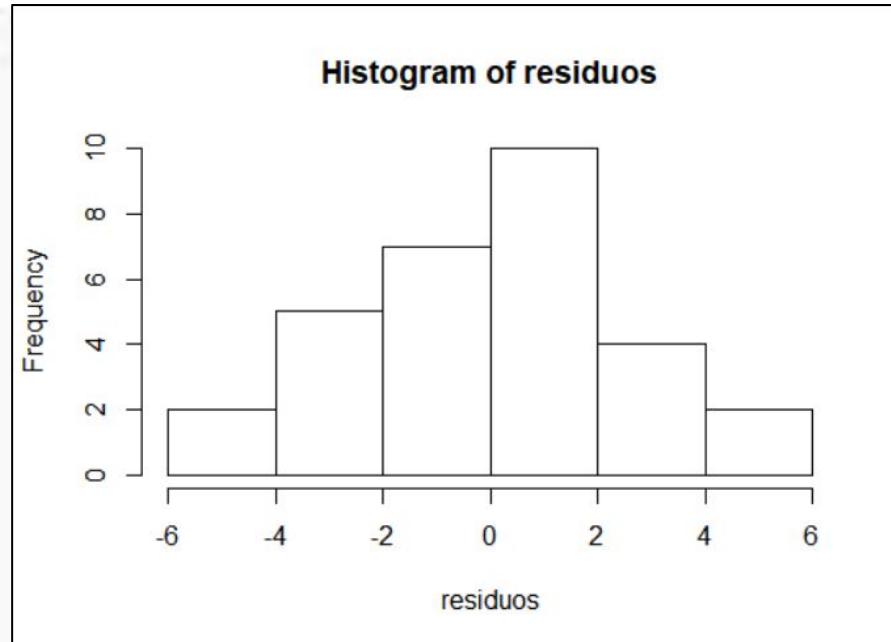
Ou seja:

$$\varepsilon \sim N(\mu=0, \sigma=1)$$

Diagnóstico do ajuste do modelo de Regressão Linear

Diagnóstico de normalidade

O resíduo é o erro de predição.

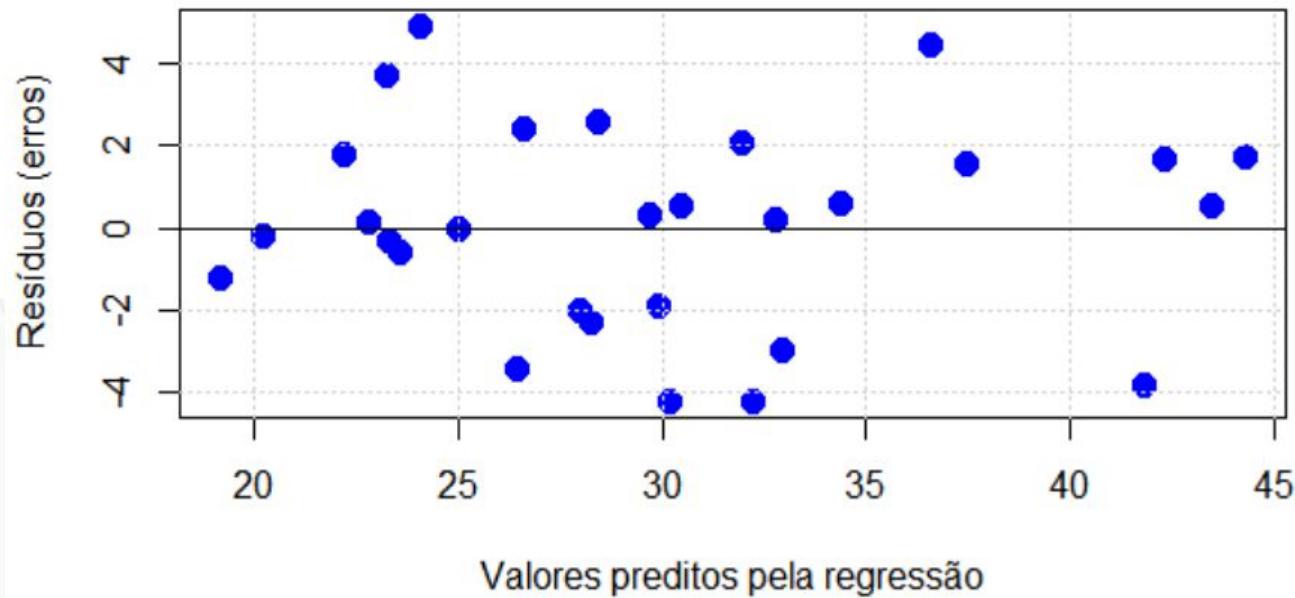


```
Shapiro-Wilk normality test  
data: resíduos W = 0.97233, p-value = 0.6048
```

Diagnóstico do ajuste do modelo de Regressão Linear

Diagnóstico de homocedasticidade

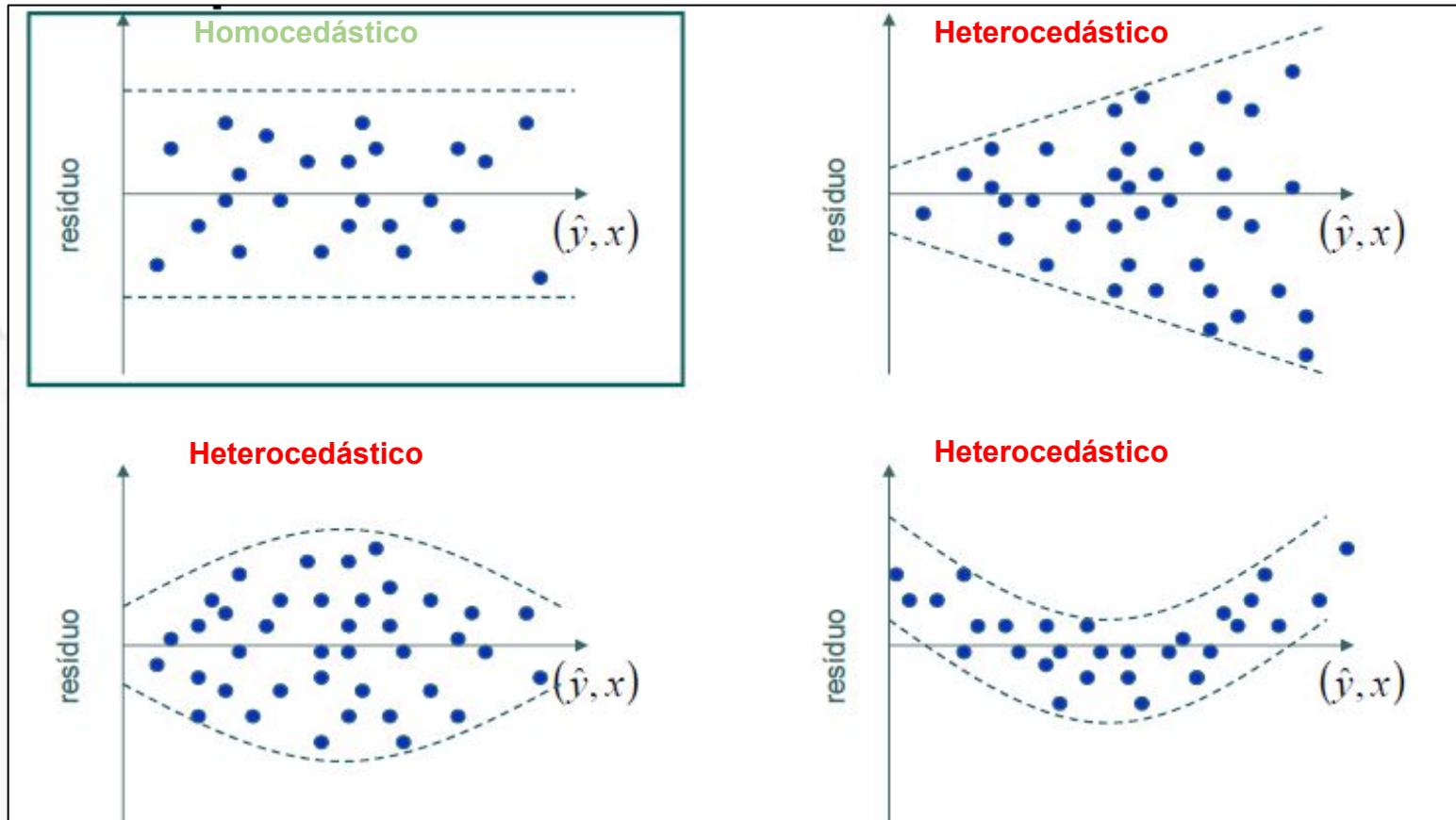
A variância dos resíduos devem ser constantes na medida que os valores da variável resposta (no eixo x) aumentam.



Diagnóstico do ajuste do modelo de Regressão Linear

Diagnóstico de homocedasticidade

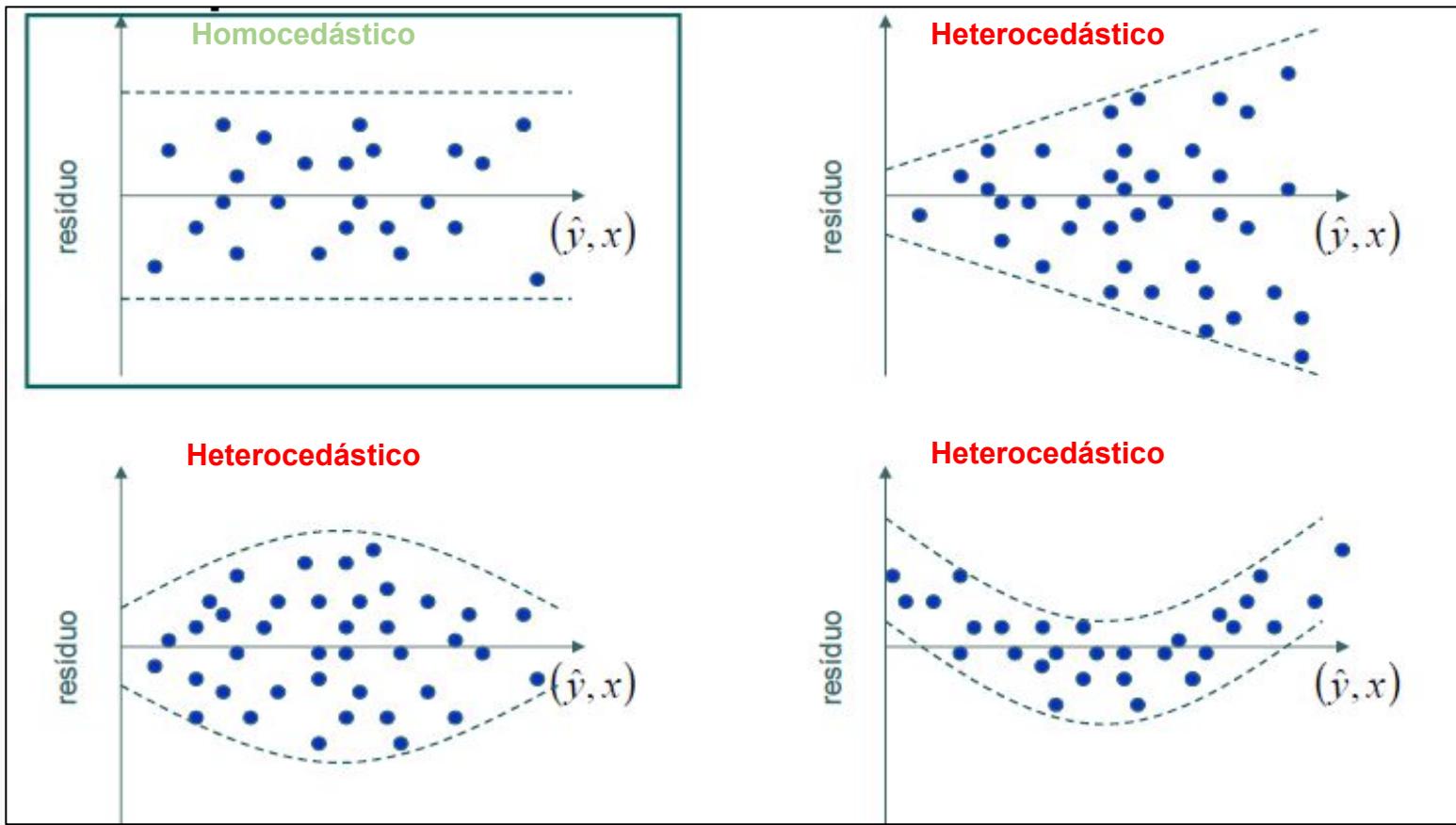
A variância dos resíduos devem ser constantes na medida em que os valores da variável resposta (no eixo x) aumentam.



Diagnóstico do ajuste do modelo de Regressão Linear

Diagnóstico de homocedasticidade

A variância dos resíduos devem ser constantes na medida que os valores da variável resposta (no eixo x) aumentam.

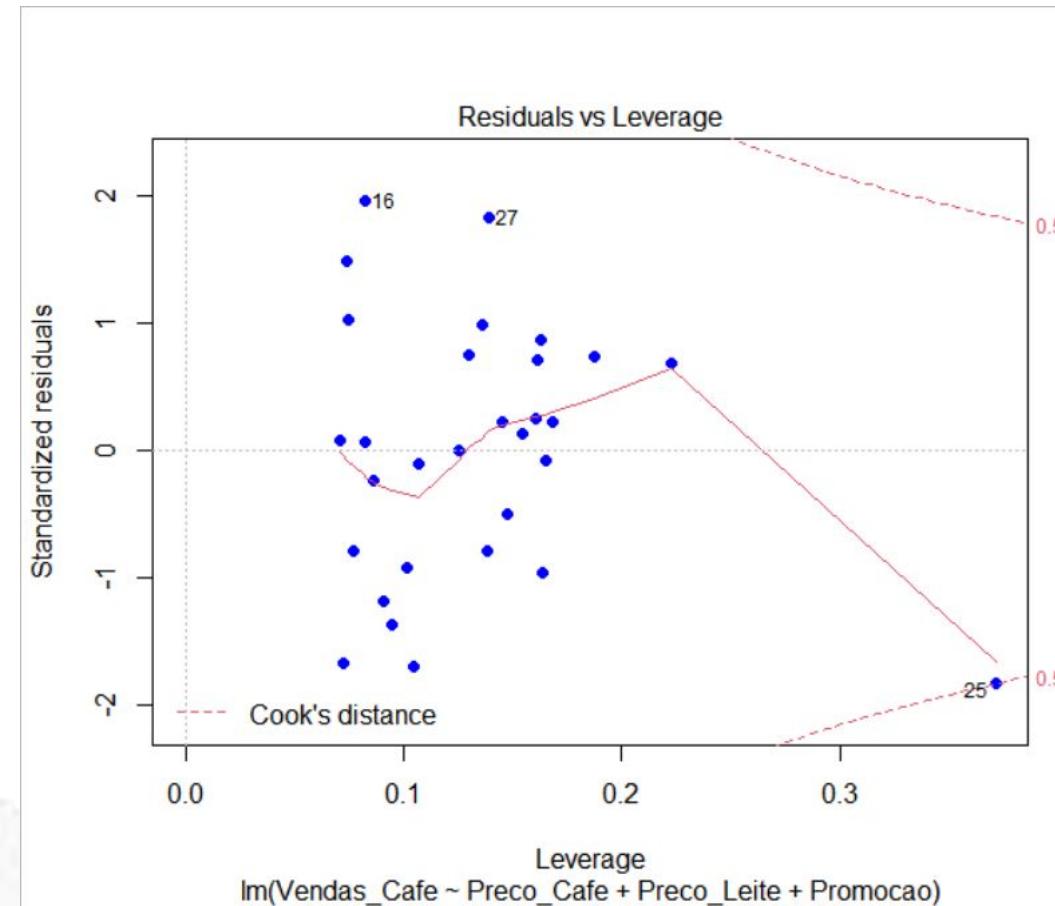


Diagnóstico do ajuste do modelo de Regressão Linear

Pontos influentes e Distância de Cook

Observações com valores muito grandes ou muito pequenos podem gerar alto resíduo e distorcer o ajuste do modelo.

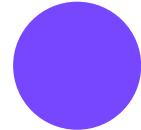
O diagnóstico dos resíduos também deve ser aplicado na ANOVA.



Conclusão



- ✓ Teste t de Student para os coeficientes betas.
- ✓ Teste F para ajuste geral do modelo de regressão.
- ✓ Diagnóstico de resíduos (normalidade e homocedasticidade).
- ✓ R² e R² Ajustado.
- ✓ Pontos influentes.



Próxima aula



- Algoritmo Stepwise para seleção automática de preditores.

Análise Estatística de Dados

AULA 5.6. ALGORITMO STEPWISE PARA SELEÇÃO DE PREDITORES

PROF. MÁIRON CHAVES

Nesta aula



- Algoritmo Stepwise.

Seleção automática de variáveis preditoras

Método Forward – Começa sem preditores no modelo, adiciona iterativamente os preditores mais contribuintes e para quando a melhoria não é mais estatisticamente significativa.

Método Backward – Começa com todos os preditores no modelo, remove iterativamente os preditores menos contributivos e para quando você tem um modelo em que todos os preditores são estatisticamente significativos.

Método Both – Começa sem preditores e, em seguida, adiciona sequencialmente os preditores mais contribuintes (como no método Forward). Depois de adicionar cada nova variável, remove todas as variáveis que não fornecem mais uma melhoria no ajuste do modelo (como no método Backward).

Métricas de qualidade de ajuste podem ser variadas, por exemplo, o AIC (*Akaike information criterion*), o BIC (*Bayesian information criterion*) e o R² ajustado.

Procedimentos automáticos não consideram o conhecimento especializado que o pesquisador possa ter sobre os dados. Por isso o modelo selecionado pode não ser o melhor sob um ponto de vista prático!

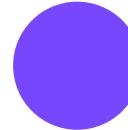
Conclusão



- ✓ Método Forward.
- ✓ Método Backward.
- ✓ Método Both.



Próxiam aula



- Regressão Linear com o R.

Análise Estatística de Dados

AULA 5.7. ESTATÍSTICA COMPUTACIONAL – REGRESSÃO LINEAR COM O R

PROF. MÁIRON CHAVES

Nesta aula



- ❑ Regressão Linear com o R.

Estatística Computacional – Regressão Linear com o R

Copie o código na apostila, cole no seu R e execute



Estatística Computacional – Regressão Linear com o R

```
#####
```

```
#####      Regressao Linear      #####
```

```
##   AED - Capitulo 05 - Prof. Máiron Chaves   ####
```

```
#####
```

#Copie este código, cole no seu R e execute para ver os resultados

```
rm(list = ls()) #Limpa memória do R
```

```
library(ggplot2) #Biblioteca pra gerar visualizacoes mais sofisticadas
```

```
library(plotly) #Biblioteca pra gerar visualizacoes mais sofisticadas
```

```
#Cria o data frame
```

Conclusão



- ✓ Gráfico de dispersão e reta de regressão.
- ✓ Gráfico de dispersão utilizando a biblioteca ggplot2.
- ✓ Gráfico 3D utilizando a biblioteca plotly.
- ✓ Obtendo o coeficiente de correlação.
- ✓ Diagnóstico de resíduos.
- ✓ Stepwise.
- ✓ Realizando predição para uma nova observação.

Fundamentos de Estatística e Aprendizado de Máquina

Capítulo 6. Regressão Logística

Prof. Máiron Chaves



Fundamentos de Estatística e Aprendizado de Máquina

Aula 6.1. Classificação Binária e Interpretação dos Coeficientes

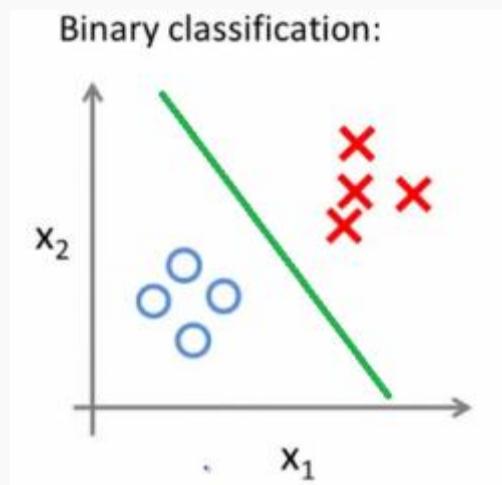
Prof. Máiron Chaves

Nesta aula



- ❑ Regressão Linear vs Regressão Logística.
- ❑ Eventos Binários.
- ❑ Interpretando os coeficientes.

Classificação Binária



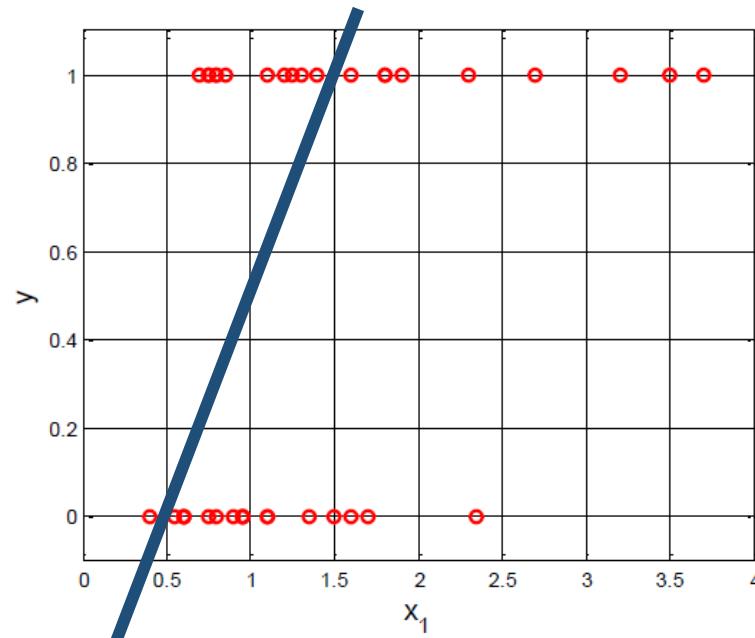
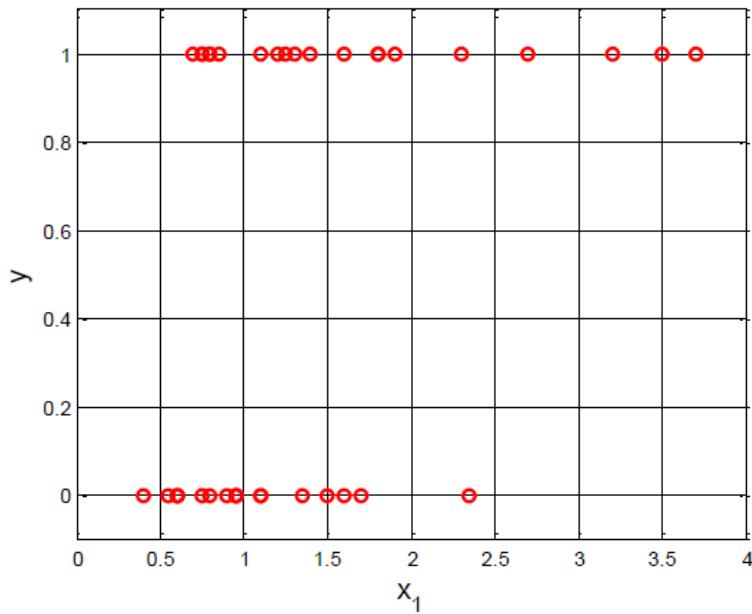
- Cliente Compra/Não Compra.
- Cliente Cancela/Não Cancela.
- Cliente Pagará/Não Pagará.
- Equipamento Estragará/Não Estragará.
- Contratar/Não Contratar um candidato.
- Tumor benigno/Tumor maligno.

Regressão Logística

IGTI

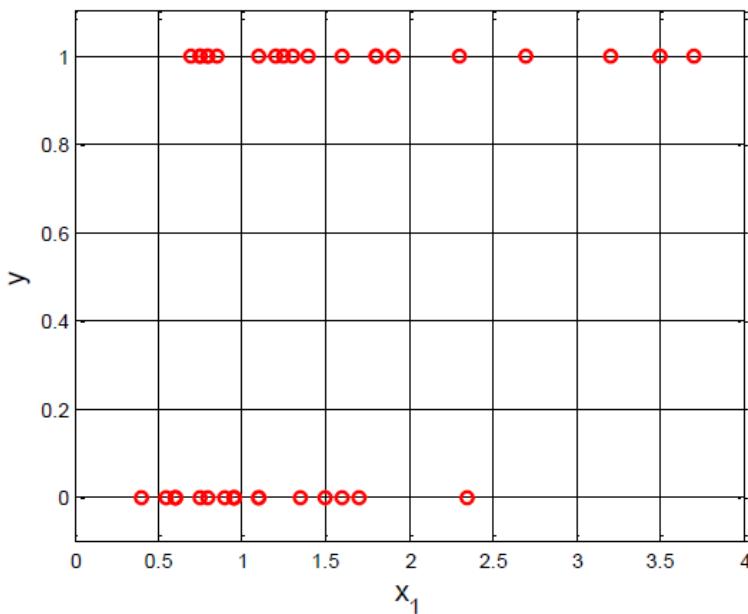
$$Y \in \{1,0\}$$

$$\hat{y} = \beta_0 + \beta_1 * X_1 + \cdots + \beta_p * X_p$$

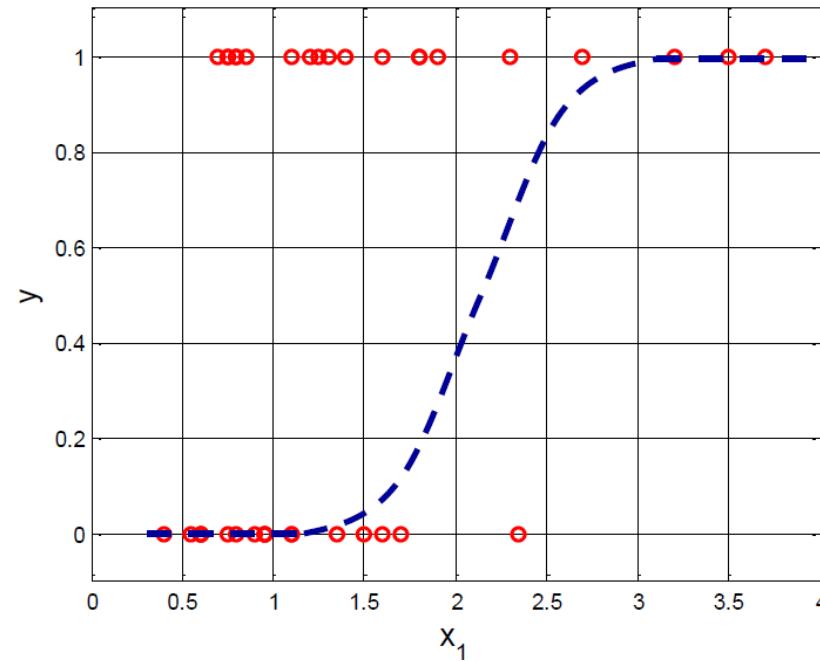


Regressão Logística

$$Y \in \{1,0\}$$



$$\hat{y} = \Pr(y = 1 | x) = \frac{e^{x^T \beta + \beta_0}}{1 + e^{x^T \beta + \beta_0}}$$

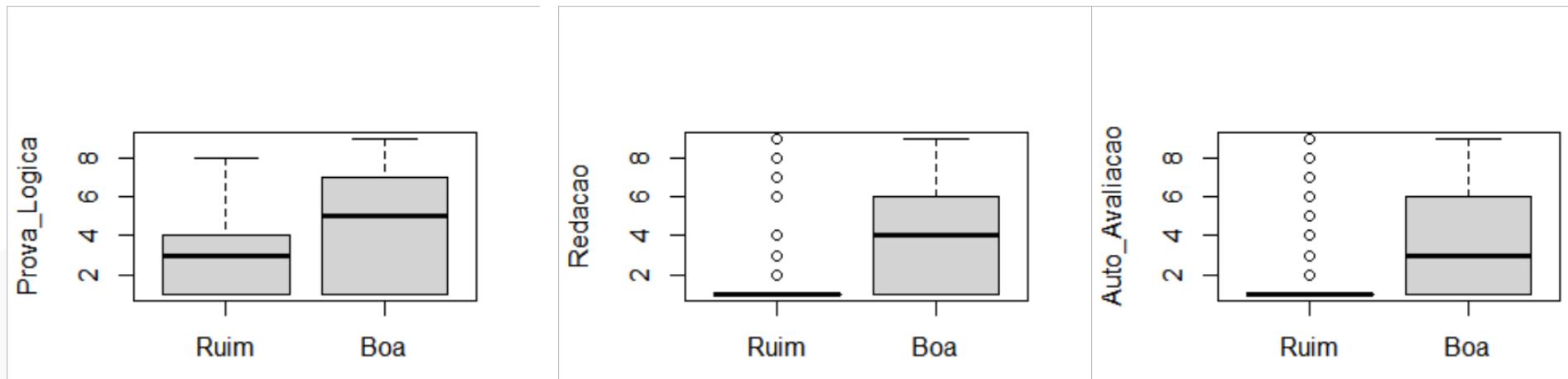


Regressão Logística



Prova_Logica	Redacao	Psicotecnico	Dinamica_Grupo	Fit_Cultural	Ingles	Avaliacao_RH	Auto_Avaliacao	Demograficos	Estado_Civil	Escolaridade	Classe
2	1	1	1	1	2	1	2	1	1 Divorciado	Superior	Ruim
2	1	1	1	1	2	1	3	1	1 Divorciado	Segundo_Grau	Ruim
5	1	1	1	1	2	1	2	1	1 Casado	Superior	Ruim
5	4	6	8	4	1	8	1	1	1 Divorciado	Superior	Boa
5	3	3	1	2	1	2	1	1	1 Solteiro	Segundo_Grau	Ruim
2	3	1	1	3	1	1	1	1	1 Solteiro	Pos_Graduacao	Ruim
3	5	7	8	8	9	7	1	7	Divorciado	Segundo_Grau	Boa
1	5	6	1	6	1	7	7	1	Casado	Segundo_Grau	Boa
1	9	8	7	6	4	7	1	3	Casado	Segundo_Grau	Boa
4	1	1	1	2	1	3	1	1	Casado	Segundo_Grau	Ruim
5	1	1	1	1	2	1	3	1	1 Casado	Pos_Graduacao	Ruim
8	1	1	1	1	3	6	3	9	1 Solteiro	Pos_Graduacao	Boa
1	1	3	1	2	1	2	1	1	1 Divorciado	Superior	Ruim
1	1	1	1	2	1	1	1	1	1 Solteiro	Pos_Graduacao	Ruim
3	4	5	2	6	8	4	1	1	1 Solteiro	Superior	Boa
4	3	3	1	2	1	3	3	1	1 Divorciado	Segundo_Grau	Ruim
3	3	2	1	3	1	3	6	1	1 Solteiro	Segundo_Grau	Ruim
2	1	1	1	1	2	1	1	1	1 Solteiro	Superior	Ruim
1	1	1	1	1	2	1	2	1	1 Solteiro	Segundo_Grau	Ruim
1	1	1	1	1	2	1	1	1	1 Casado	Segundo_Grau	Ruim
8	1	1	1	1	5	1	8	1	6 Solteiro	Segundo_Grau	Boa
8	7	4	4	5	3	5	1	1	1 Solteiro	Pos_Graduacao	Boa
1	1	1	1	1	1	2	1	1	1 Casado	Segundo_Grau	Ruim
2	1	1	1	1	2	1	1	1	1 Solteiro	Segundo_Grau	Ruim
1	8	8	4	1	1	8	1	1	1 Casado	Superior	Boa
5	1	1	1	2	2	1	2	1	1 Casado	Superior	Ruim
3	1	1	1	1	2	1	2	1	2 Divorciado	Pos_Graduacao	Ruim
3	1	1	1	1	2	1	3	1	1 Casado	Superior	Ruim

Regressão Logística



	Prova_Logica	Redacao	Auto_Avaliacao
Min.	:1.000	:1.000	:1.000
1st Qu.	:1.000	:1.000	:1.000
Median	:3.000	:1.000	:1.000
Mean	:3.531	:2.272	:2.082
3rd Qu.	:5.000	:3.000	:2.000
Max.	:9.000	:9.000	:9.000

Regressão Logística



Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.84564	0.27820	-13.823	< 2e-16	***
Prova_Logica	0.21409	0.05184	4.130	3.63e-05	***
Redacao	0.69158	0.07437	9.299	< 2e-16	***
Auto_Avaliacao	0.42409	0.06728	6.303	2.91e-10	***

$$\hat{y} = -3,8458 + (0,2143 * Prova_Logica) + (0,6914 * Redacao) + (0,4238 * Auto_Avaliacao)$$

Regressão Logística



$\beta_1 = \text{Prova_Logica} = \exp(0.2143) = 2,7182^{0,2143} = \underline{\underline{1,23}}$ – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na prova de lógica, o candidato aumenta em média 1,23 vezes as chances de pertencer à classe ‘Boa’.

$\beta_2 = \text{Redacao} = \exp(0.6914) = 2,7182^{0,6914} = \underline{\underline{1,99}}$ – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na prova de redação, o candidato aumenta em média, 1,99 vezes as chances de pertencer à classe ‘Boa’.

$\beta_3 = \text{Auto_Avaliacao} = \exp(0,4238) = 2,7182^{0,4238} = \underline{\underline{1,5278}}$ – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na auto avaliação, o candidato aumenta em 1,52 vezes as chances de pertencer à classe ‘Boa’.

Conclusão



- ✓ Classificação Binária.
- ✓ Interpretando coeficientes da Regressão Logística.

Na próxima aula



- Realizando previsões com a Regressão Logística.



Fundamentos de Estatística e Aprendizado de Máquina

Aula 6.2. Realizando Predição com a Regressão Logística

Prof. Máiron Chaves

Nesta aula



- Função logística.
- Realizando previsões.
- Seleção automática de preditores.

Regressão Logística

$$\hat{y} = -3,8458 + (0,2143 * \text{Prova_Logica}) + (0,6914 * \text{Redacao}) + (0,4238 * \text{Auto_Avaliacao})$$



$$\hat{y} = \frac{e^{-3,8458+(0,2143*\text{Prova_Logica})+(0,6914*\text{Redacao})+(0,4238*\text{Auto_Avaliacao})}}{1 + e^{-3,8458+(0,2143*\text{Prova_Logica})+(0,6914*\text{Redacao})+(0,4238*\text{Auto_Avaliacao})}}$$

Supondo que o candidato tire 3 em Prova_Logica, 5 em Redacao e 1 em Auto_Avaliacao.

A probabilidade dele(a) pertencer à classe de interesse ‘Boa’ fica:

$$\hat{y} = \frac{e^{-3,8458+(0,2143*3)+(0,6914*5)+(0,4238*1)}}{1 + e^{-3,8458+(0,2143*3)+(0,6914*5)+(0,4238*1)}}$$

$$\hat{y} = \frac{1,9697}{2,9697}$$

$$\hat{y} = 0,6632 \text{ (ou } 66,32\%)$$

Regressão Logística



Prova_Logica	Redacao	Auto_Avaliacao	Classe	probabilidade
2	1		1 Ruim	0.09097517
2	1		1 Ruim	0.09097517
5	1		1 Ruim	0.15982827
5	4		1 Boa	0.60234090
5	3		1 Ruim	0.43134889
2	3		1 Ruim	0.28523727
3	5		1 Boa	0.66343073
1	5		7 Boa	0.94240236
1	9		1 Boa	0.95332501
4	1		1 Ruim	0.13312577
5	1		1 Ruim	0.15982827

Regressão Logística



Ponto de Corte: 0,5

Prova_Logica	Redacao	Auto_Avaliacao	Classe	probabilidade	Predicao
2	1		1	Ruim	0.09097517 Ruim
2	1		1	Ruim	0.09097517 Ruim
5	1		1	Ruim	0.15982827 Ruim
5	4		1	Boa	0.60234090 Boa
5	3		1	Ruim	0.43134889 Ruim
2	3		1	Ruim	0.28523727 Ruim
3	5		1	Boa	0.66343073 Boa
1	5		7	Boa	0.94240236 Boa
1	9		1	Boa	0.95332501 Boa
4	1		1	Ruim	0.13312577 Ruim
5	1		1	Ruim	0.15982827 Ruim

Regressão Logística - Stepwise



Coefficients:

	Estimate	std. Error	z value	Pr(> z)	
(Intercept)	-7.29948	0.63721	-11.455	< 2e-16	***
Prova_Logica	0.23650	0.07342	3.221	0.00128	**
Redacao	0.20571	0.10626	1.936	0.05288	.
Psicotecnico	0.40900	0.10350	3.952	7.75e-05	***
Dinamica_Grupo	0.19739	0.09318	2.118	0.03414	*
Fit_cultural	0.28630	0.10435	2.744	0.00608	**
Inglés	-0.08452	0.03998	-2.114	0.03452	*
Avaliacao_RH	0.64018	0.09103	7.033	2.03e-12	***
Auto_Avaliacao	0.13054	0.08629	1.513	0.13031	
Demograficos	0.77714	0.15599	4.982	6.30e-07	***
Estado_civilDivorciado	0.09922	0.38401	0.258	0.79611	
Estado_civilSolteiro	-0.36401	0.40339	-0.902	0.36686	
EscolaridadesSegundo_Grau	0.07021	0.37122	0.189	0.84998	
EscolaridadesSuperior	-0.56484	0.38455	-1.469	0.14188	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Regressão Logística - Stepwise



Coefficients:

	Estimate	std. Error	z value	Pr(> z)	
(Intercept)	-7.35786	0.55520	-13.253	< 2e-16	***
Prova_Logica	0.22033	0.07196	3.062	0.002201	**
Redacao	0.22923	0.10710	2.140	0.032335	*
Psicotecnico	0.39374	0.10252	3.841	0.000123	***
Dinamica_Grupo	0.20777	0.09192	2.260	0.023793	*
Fit_cultural	0.26802	0.10217	2.623	0.008708	**
Ingles	-0.08332	0.03936	-2.117	0.034275	*
Avaliacao_RH	0.62276	0.08834	7.049	1.80e-12	***
Auto_Avaliacao	0.13426	0.08499	1.580	0.114167	
Demograficos	0.73248	0.15198	4.820	1.44e-06	***

signif. codes:	0	'***'	0.001	'**'	0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusão



- ✓ Realizando previsões com a Regressão Logística.
- ✓ Stepwise na Regressão Logística.

Na próxima aula



- Avaliando a capacidade preditiva do modelo ajustado.



Fundamentos de Estatística e Aprendizado de Máquina

Aula 6.3. Avaliando a Capacidade Preditiva

Prof. Máiron Chaves

Nesta aula



- Acurácia.
- Sensibilidade.
- Especificidade.
- Curva ROC.

Regressão Logística

IGTI

Ponto de corte: 0,5

		classe_original	
		classe_Predita	Boa Ruim
		Boa	169 23
		Ruim	72 435

$$Acurácia = \frac{169 + 435}{169 + 23 + 72 + 435} = 86,40\%$$

$$Sensitividade = \frac{Verdadeiros Positivos}{Verdadeiros Positivos + Falsos Negativos} = \frac{169}{169 + 72} = 70,12\%$$

$$Especificidade = \frac{Verdadeiros Negativos}{Verdadeiros Negativos + Falsos Positivos} = \frac{435}{435 + 23} = 94,97\%$$

Regressão Logística



Ponto de corte: 0,9

		classe_original	
		classe_Predita	Boa Ruim
classe_Original	Boa	98	9
	Ruim	143	449

$$\text{Acurácia} = \frac{98 + 449}{98 + 9 + 143 + 449} = 78,25\%$$

$$\text{Sensitividade} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} = \frac{98}{98 + 143} = 40,66\%$$

$$\text{Especificidade} = \frac{\text{Verdadeiros Negativos}}{\text{Verdadeiros Negativos} + \text{Falsos Positivos}} = \frac{449}{449 + 9} = 98,03\%$$

Regressão Logística



Ponto de corte: 0,1

		classe_Original	
		classe_Predita	Boa Ruim
classe_Original	Boa	229	291
	Ruim	12	167

$$\text{Acurácia} = \frac{229 + 167}{229 + 167 + 12 + 291} = 56,65\%$$

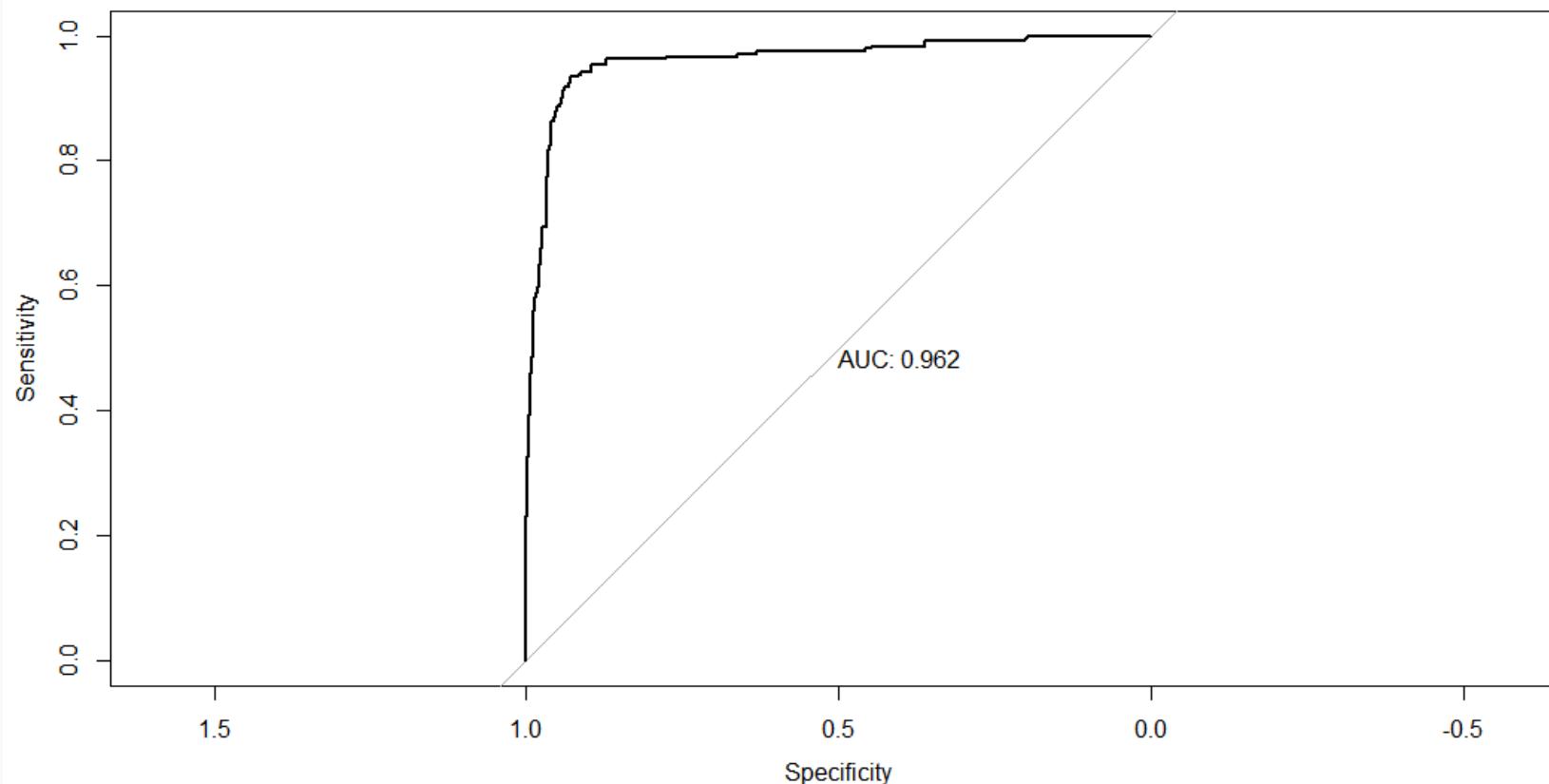
$$\text{Sensitividade} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} = \frac{229}{229 + 12} = 95,02\%$$

$$\text{Especificidade} = \frac{\text{Verdadeiros Negativos}}{\text{Verdadeiros Negativos} + \text{Falsos Positivos}} = \frac{167}{167 + 291} = 36,46\%$$

Regressão Logística

Curva ROC ‘receiver operating characteristic curve’

Em vez de verificar manualmente os pontos de corte, podemos criar uma curva ROC que irá varrer todos os cortes possíveis e traçar a sensibilidade e especificidade.



Conclusão



- ✓ Acurácia.
- ✓ Sensibilidade.
- ✓ Especificidade.
- ✓ Curva ROC.
- ✓ Área sobre a curva ROC.

Na próxima aula



- Análise de Sensibilidade e Especificidade



Fundamentos de Estatística e Aprendizado de Máquina

Aula 6.4. Análise de Sensibilidade e Especificidade

Prof. Máiron Chaves

Nesta aula

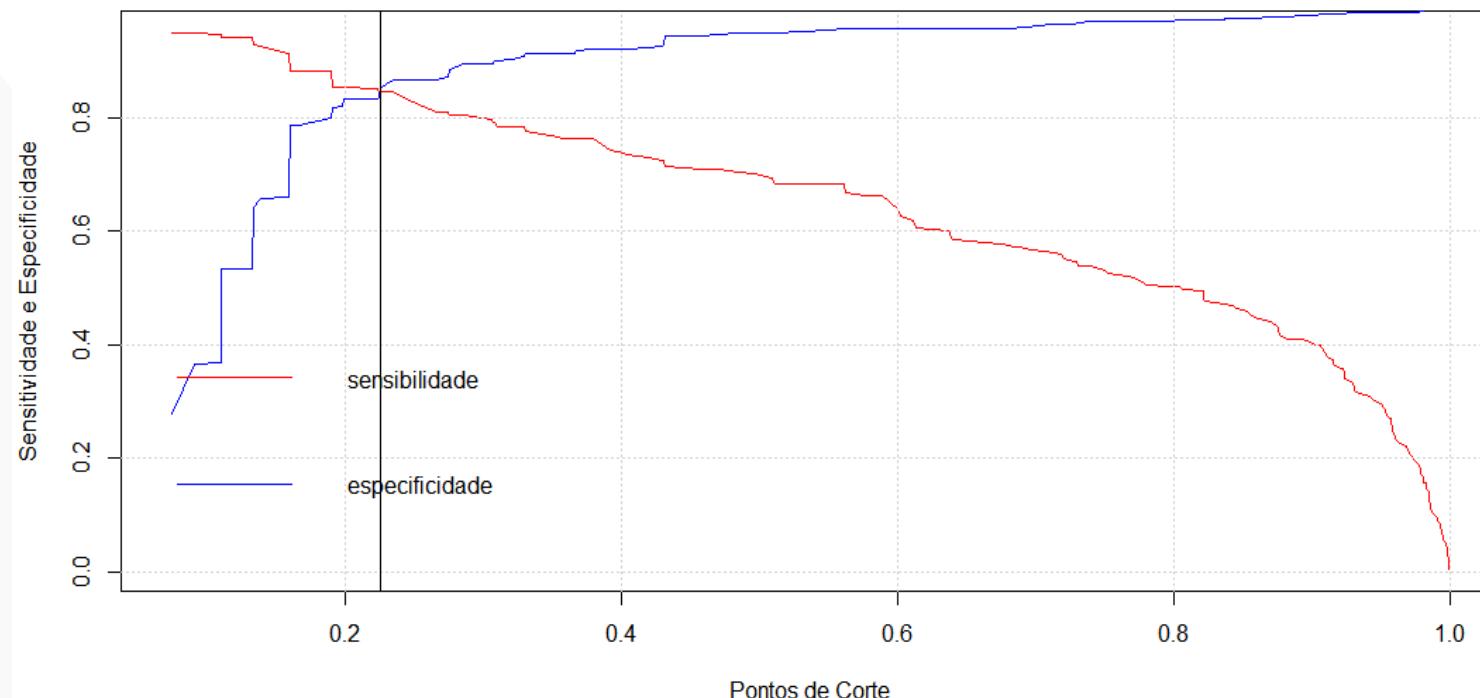


- Gráfico de Sensibilidade e Especificidade.
- Identificando o ponto de corte 'ótimo'.

Regressão Logística

IGTI

A ideia da Análise de Sensibilidade e Especificidade é simular várias matrizes de confusão, através de vários pontos de corte diferentes e identificar aquela matriz de confusão que nos dará tanto a maior Sensibilidade quanto a maior Especificidade.



Regressão Logística

Ponto de corte: 0,225



		classe_original	
		classe_Predita	Boa Ruim
		Boa	204 75
		Ruim	37 383

$$\text{Acurácia} = \frac{204 + 383}{204 + 75 + 37 + 383} = 83,97\%$$

$$\text{Sensitividade} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} = \frac{204}{204 + 37} = 84,64\%$$

$$\text{Especificidade} = \frac{\text{Verdadeiros Negativos}}{\text{Verdadeiros Negativos} + \text{Falsos Positivos}} = \frac{383}{383 + 75} = 83,62\%$$

Conclusão



- ✓ Análise de Sensibilidade e Especificidade.

Na próxima aula

- ❑ Regressão Logística com o R.





Fundamentos de Estatística e Aprendizado de Máquina

Aula 6.5. Regressão Logística com o R

Prof. Máiron Chaves

Nesta aula



- ❑ Regressão Logística com o R.

Regressão Logística com o R



INSTITUTO DE GESTÃO E TECNOLOGIA DA INFORMAÇÃO

Estatística Computacional – Regressão Logística no R

```
#####
```

```
#####      Regressao Logistica      #####
```

```
##   AED - Capitulo 06 - Prof. Máiron Chaves ##
```

```
#####
```

```
#Copie este código, cole no seu R e execute para ver os resultados
```

```
rm(list = ls()) #Limpa memória do R
```

```
#Instala e carrega biblioteca para gerar a curva ROC
```

```
install.packages('pROC') #Instala
```

```
library(pROC) #Carrega
```

```
#Monte o dataset
```

```
dados <- data.frame(Prova_Logica = c(2, 2, 5, 5, 5, 2, 3, 2, 1, 4,
```

Conclusão



- ✓ Regressão Logística com o R.