

FIAP

MBA

MBA em DATA SCIENCE

STATISTICS FUNDAMENTALS



Dra. Regina Tomie Ivata Bernal

Cientista de Dados na área da Saúde

Formação Acadêmica:

Estatístico - UFSCar

Mestre em Saúde Pública - FSP/USP

Doutor em Ciências - Epidemiologia - FSP/USP

profregina.bernal@fiap.com.br
reginabernal@terra.com.br

Atividades Profissionais:

Professora de pós-graduação na FIAP

Consultora externa da SVS/MS

Cientista de Dados em Saúde

Aula 1

Estatística Descritiva

Avaliação da disciplina

Avaliação	Peso
Listas de exercícios	0.5
Exercício prático em grupo	0.5

Objetivos da Disciplina

- Disseminar a cultura estatística quanto ao uso das técnicas descritivas, técnicas de associação e correlação tendo em vista a modelagem para previsão.
- Desenvolver conceitos de preparação de dados para fins estatísticos e informações para a geração de competitividade organizacional.
- Preparar o aluno para implementar as técnicas estatísticas.

Referências Bibliográficas

- BUSSAB, W.O.; MORETTIN, P. A., Estatística Básica, 5a. ed., São Paulo: Saraiva, 2006.
- LARSON, R.; FARBER, B. Estatística Aplicada. Tradução técnica Cyro Patarra, 2^a ed., São Paulo, 2007.
- MARTINS, G.A; DENIS, G.A.M. Princípios da Estatística. 4^a ed. São Paulo: Atlas, 1990.
- BERRY, M.J.A.; LINOFF, G. Data Mining Techniques: for marketing, sales, and customer. Wiley Computer Publishing, 1997.



Referências Bibliográficas

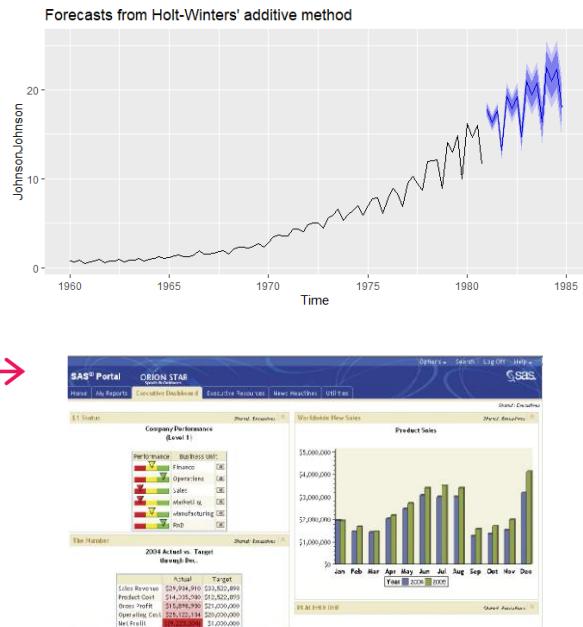
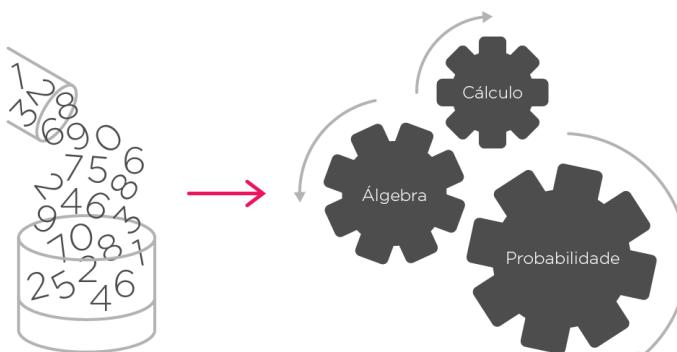
- MOORE, S.D.; MCCABE, G.P.; DUCKWORTH, W.M.; SCLOVE, S.S.
Estatística Empresarial como usar dados para tomar decisões.
Tradução Luis
Antonio Forjado. Rio de Janeiro: LTC, 2006.
- MORETIM, P.A.; TOLOI, C.M.C. Análise de Séries Temporais, 2^a ed., São Paulo: Edgard Blücher, 2006.
- SILVA, NN. Amostragem Probabilística. 2^a ed., São Paulo: Editora da Universidade de São Paulo, 2001.
- SOARES, J.; FARIA, A. A.; CESAR, C. C., Introdução a Estatística, LTC, 2002.



DATA ANALYTICS

DATA ANALYTICS

Métodos



UNIVERSO DE FORNECEDORES

MicroStrategy®

IBM

SAP®

sas

Microsoft

alteryx

+able

Qlik

TIBCO®
The Power of Now®

Information
Builders

julia

ORACLE

OPEN TEXT
The Content Experts®

R

R Studio



python

O QUE
ACONTECEU?

DESCRITIVO

QUANTOS CANCELAMENTOS?

QUANTOS CLIENTES NOVOS, QUANTOS ANTIGOS?

QUAL REGIÃO?

QUE TIPO DE CLIENTE?

POR QUE ISTO
ACONTEceu?

DIAGNÓSTICO

QUAL A RELAÇÃO ENTRE CANCELAMENTO
VOLUNTÁRIO POR TEMPO DE CONTA E TIPO DE
CLIENTE?

O QUE
ACONTECERÁ?

PREDITIVO

QUAL A PROBABILIDADE DE UM CLIENTE
CANCELAR O SERVIÇO EM UMA CERTA
REGIÃO NOS PRÓXIMOS 3 MESES?

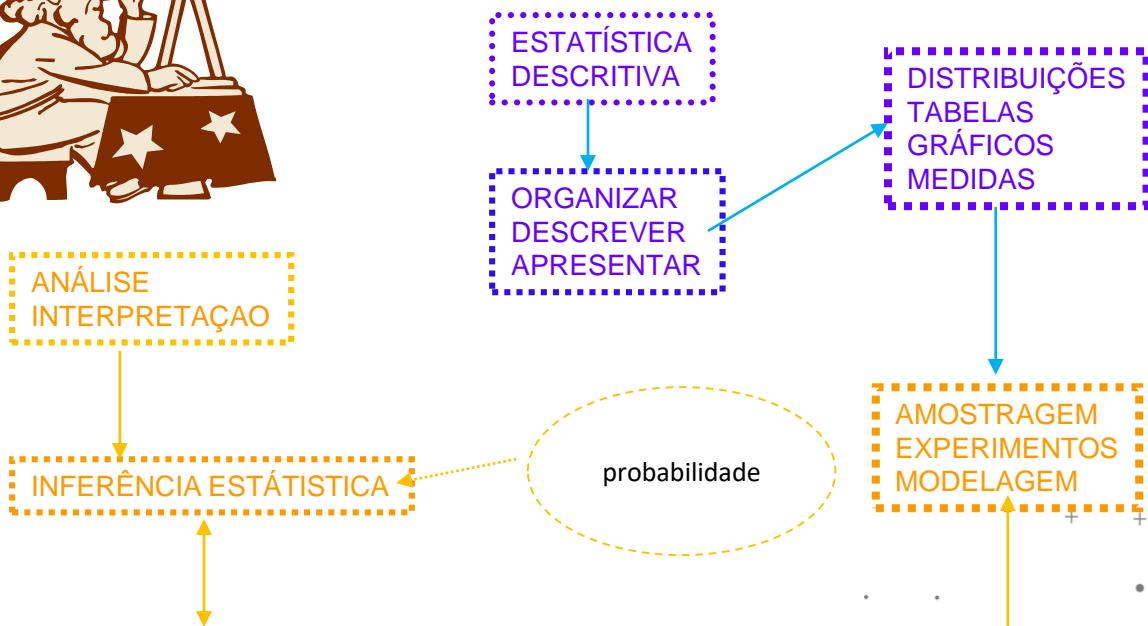
O QUE
POSSO FAZER?

PRESCRITIVO

LISTA DE AÇÕES PARA RETER OS CLIENTES,
SEGUNDO SEU VALOR?



ESTATÍSTICA



Estatística

É a ciência que trata dados numéricos provenientes de mensuração em grupos de indivíduos.

Trata da organização, descrição, apresentação análise e interpretação de dados resultantes da observação de fenômenos coletivos. Produz métodos para inferência estatística.

✓ Propriedades

Estuda as variações:

- entre indivíduos;
- em um mesmo indivíduo.





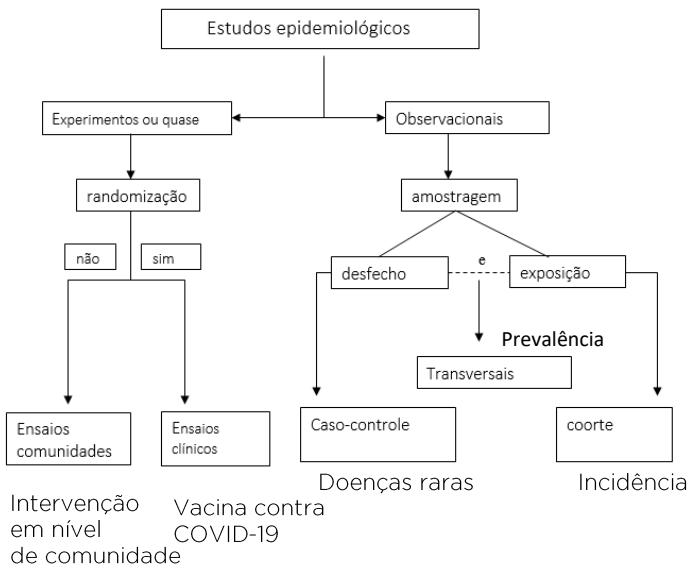
ESTATÍSTICA

ÁREA	ESTATÍSTICA
Saúde	Bioestatística
Economia	Econometria
População	Demografia
Jurídica	Jurimetria
Biologia	Biometria
Contabilidade	Contabilometria



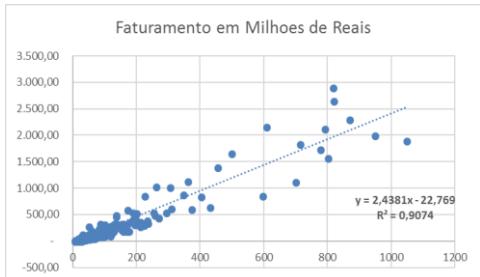
BIOESTATÍSTICA

“Bioestatística é a Estatística aplicada às ciências médica e biológica.”



Fonte: Vieira, Sonia. Introdução à Bioestatística. Rio de Janeiro: Elsevier, 1980

ECONOMETRIA



A Econometria consiste em uma série de ferramentas estatísticas que visam obter relações relevantes entre as variáveis econômicas a partir da aplicação de modelos matemáticos.

Fonte: <http://www.suno.com.br/artigos/econometria/>

Exemplos de modelos econométricos

Um modelo econométrico para previsão de impostos no Brasil.

<https://doi.org/10.1590/S1413-80502013000200006>

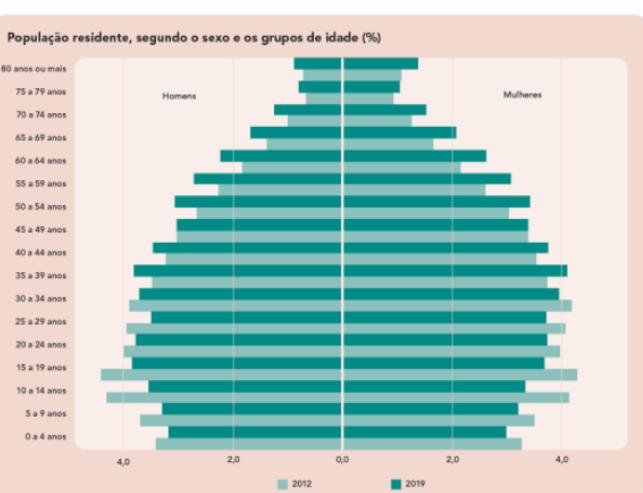
Análise Econométrica do Comportamento dos Preços do Minério de Ferro no Mercado Mundial

http://www.repositorio.ufop.br/bitstream/123456789/3608/1/DISSERTAÇÃO_AnáliseEconométricaComportamento.pdf

DEMOGRAFIA

“A Demografia é uma área do conhecimento que estuda a dinâmica das populações sejam elas humanas ou não.”

Fonte: <http://www.todamateria.com.br/demografia/#:-text=A%20demografia%20baseia-se%20em,dos%20seres%20vivos%20pelo%20planeta>.



Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Nacional por Amostra de Domicílios Contínua 2012/2019.



BIOMETRIA

“É um ramo da ciência que estuda a mensuração dos seres vivos. É a parte da Estatística que investiga atributos biológicos quantitativos, pertinentes a uma população de seres vivos.”

<http://www.youtube.com/watch?v=DiUejcMFsjY> Fonte: <http://ufpa.br/biome/bioconba.htm>

GARCIA, I. A. A segurança na identificação: a biometria da íris e da retina. 2009. 129 f. Dissertação (Mestrado em Direito Penal) – Faculdade de Direito, Universidade de São Paulo, São Paulo, 2009.

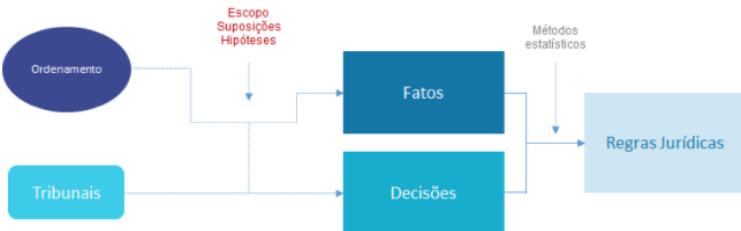
MORAES, A. F. Método para avaliação da tecnologia biométrica na segurança de aeroportos. 2006. 203 f. Dissertação (Mestrado em Engenharia Elétrica) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2006.

NAKASHIRO, M. M. Biometria no Brasil e o registro de identidade civil: novos rumos para identificação. 2011. 126 f. Tese (Doutorado em Sociologia) – Departamento de Pós-Graduação em Sociologia, Universidade do Estado de São Paulo, São Paulo, 2011.

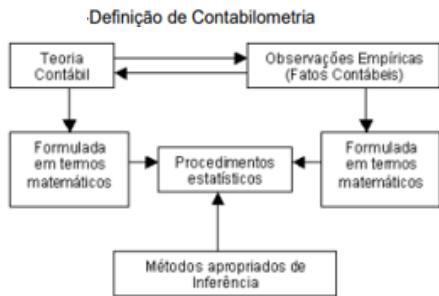
JURIMETRIA

“Os avanços da computação possibilitaram uma nova forma de encarar as normas e a sua aplicação que baseia-se em dados e, consequentemente, em estatísticas. Por isso, ela pode ser genericamente definida como “a estatística aplicada do Direito”.

Fonte: <http://abj.org.br/conteudo/jurimetria/>



CONTABILOMETRIA



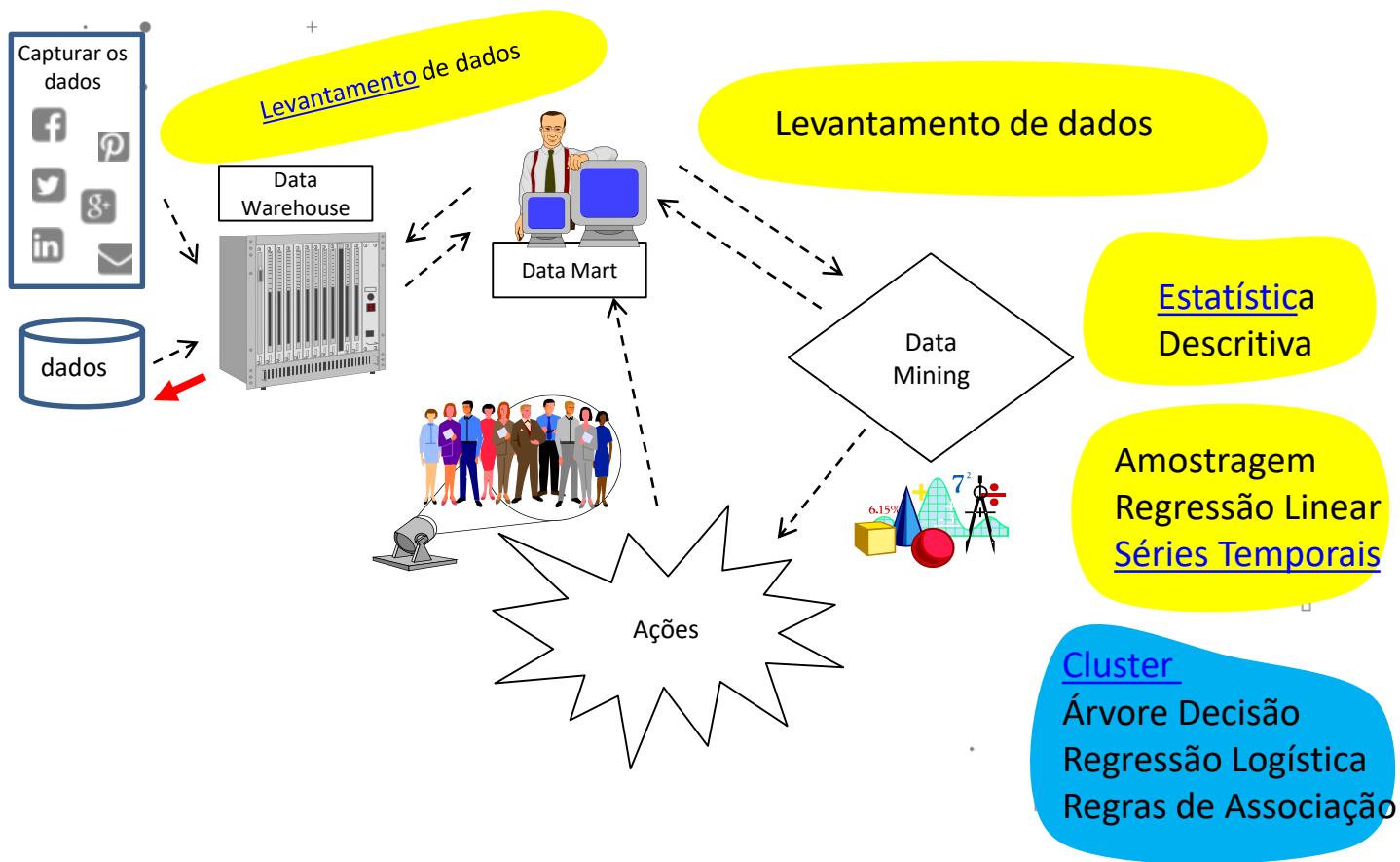
Fonte: (MARION; SILVA, 1986)

“A Contabilometria tem como característica fundamental a sua capacidade preditiva, ou seja, através da Contabilometria é possível projetar modelos de decisão eficazes que possam antecipar, prever ou estimar, de alguma forma o que irá ocorrer no futuro, provocando assim, uma melhor utilização que se pode fazer dos dados contábeis, como instrumento informativo projetado para o futuro, tornando a Contabilidade uma disciplina mais forte e mais útil.”

Fonte:

<http://congressousp.fipecafi.org/anais/artigos32006/255.pdf>

DATA **ANALYTICS**



Levantamento de Dados

Data Cleaning:

- Padronização
- Transformação de Dados
- Adoção de De-Para de Atributos

Atributo Descrição

		De	Para
Sexo:	Masculino	2	0
	Feminino	4	1
Idade:	Criação de Faixa Etária		
	0-10		1
	11-18		2
	19-25		3
	26-30		4
	31-35		5
	36-40		6
	41-45		7
	46	8	
	sem informação		0

Levantamiento de Dados

Atributo	Descrição	De	Para
Internet: Acesso últimos 3 meses		1	0

Anos

de Estudo: Criação da Faixa Grau de Instrução

0-4	1
5-8	2
9-11	3
12	4

Renda: Criacão de Faixa Salarial baseando em salários mínimos (Valor atual R\$380,00)

0-380	1
381-760	2
761-1900	3
1901-3800	4
3801	5
sem informação	0

Área: Agrupamento da área de residência em 1-Urbana e 2-Rural

|1-3| 1
|4-8| 2

Estatística Descritiva

SAS® Portal ORION STAR

Home My Reports Executive Dashboard Executive Resources News Headlines Utilities Options Search Log Off Help

Shared: Dashboard

EE Status Shared: Dashboard

Company Performance (Level 1)

Performance BUSINESS UNIT:

- Finance
- Operations
- Sales
- Marketing
- Manufacturing
- R&D

The Number Shared: Dashboard

2004 Actual vs. Target through Dec.

	Actual	Target
Sales Revenue	\$29,934,910	\$23,522,898
Product Cost	\$14,235,930	\$12,522,893
Gross Profit	\$15,698,980	\$21,000,000
Operating Cost	\$25,122,134	\$20,000,000
Net Profit	(\$9,223,154)	\$1,000,000

Worldwide New Sales Shared: Dashboard

Predict Sales

Month	2004 Actual	2005 Predicted
Jan	\$28,000,000	\$28,000,000
Feb	\$25,000,000	\$26,000,000
Mar	\$26,000,000	\$24,000,000
Apr	\$28,000,000	\$30,000,000
May	\$29,000,000	\$32,000,000
Jun	\$31,000,000	\$34,000,000
Jul	\$32,000,000	\$35,000,000
Aug	\$31,000,000	\$33,000,000
Sep	\$15,000,000	\$18,000,000
Oct	\$18,000,000	\$21,000,000
Nov	\$19,000,000	\$23,000,000
Dec	\$24,000,000	\$45,000,000

Year 2004 2005

Shared: Dashboard

EE Status Shared: Dashboard

EE Status Shared: Dashboard

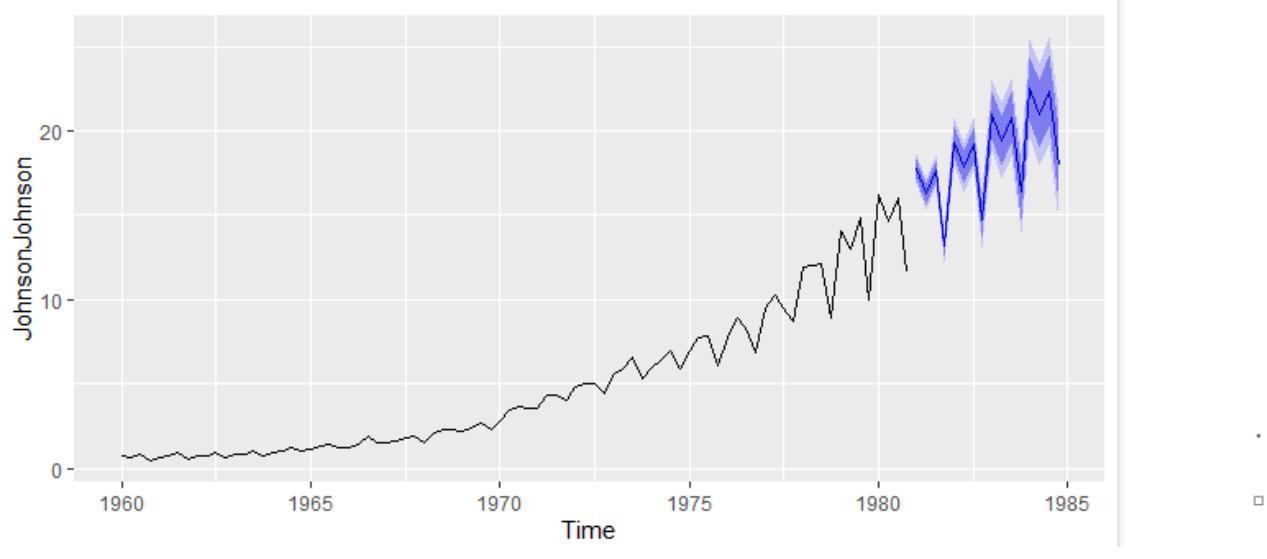
Fonte: <http://www.sas.com/technologies/bi/entbiserver/index.html#screenshot>

Inferência Estatística

• • • + • □

• • +

Forecasts from Holt-Winters' additive method



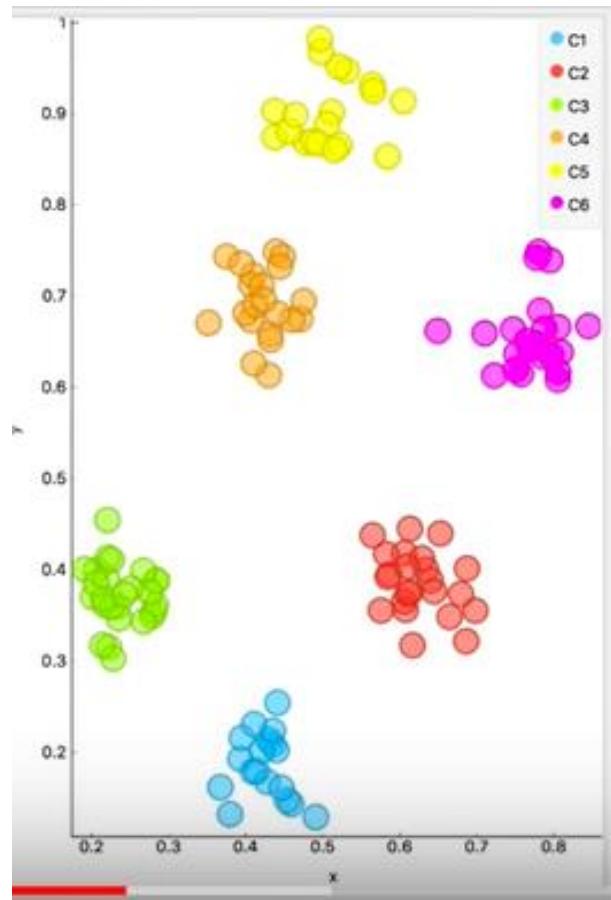
+ + +



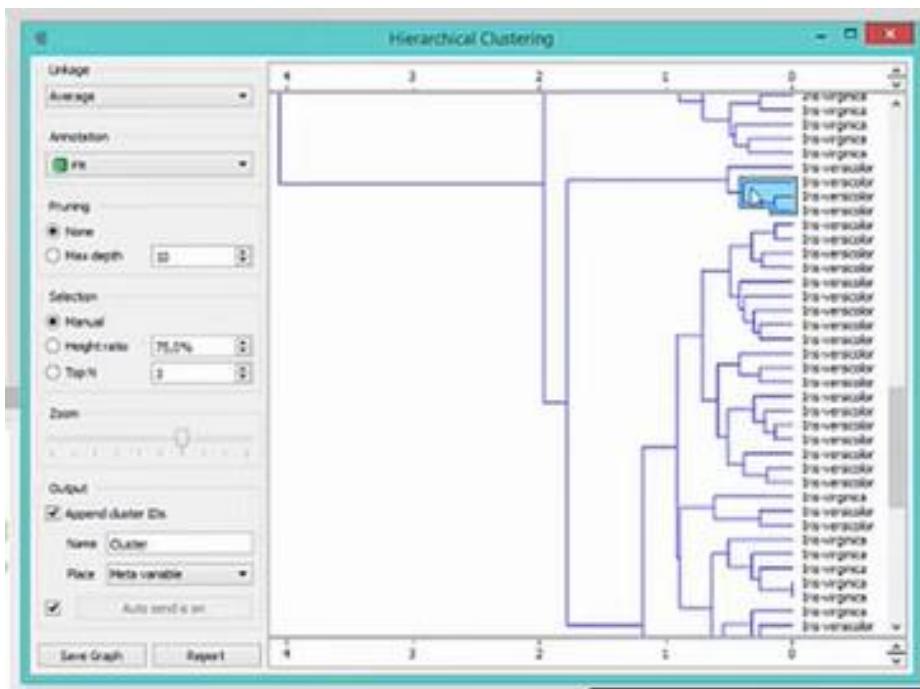
•

□ • • • •

Exemplo de Cluster Não Hierárquico

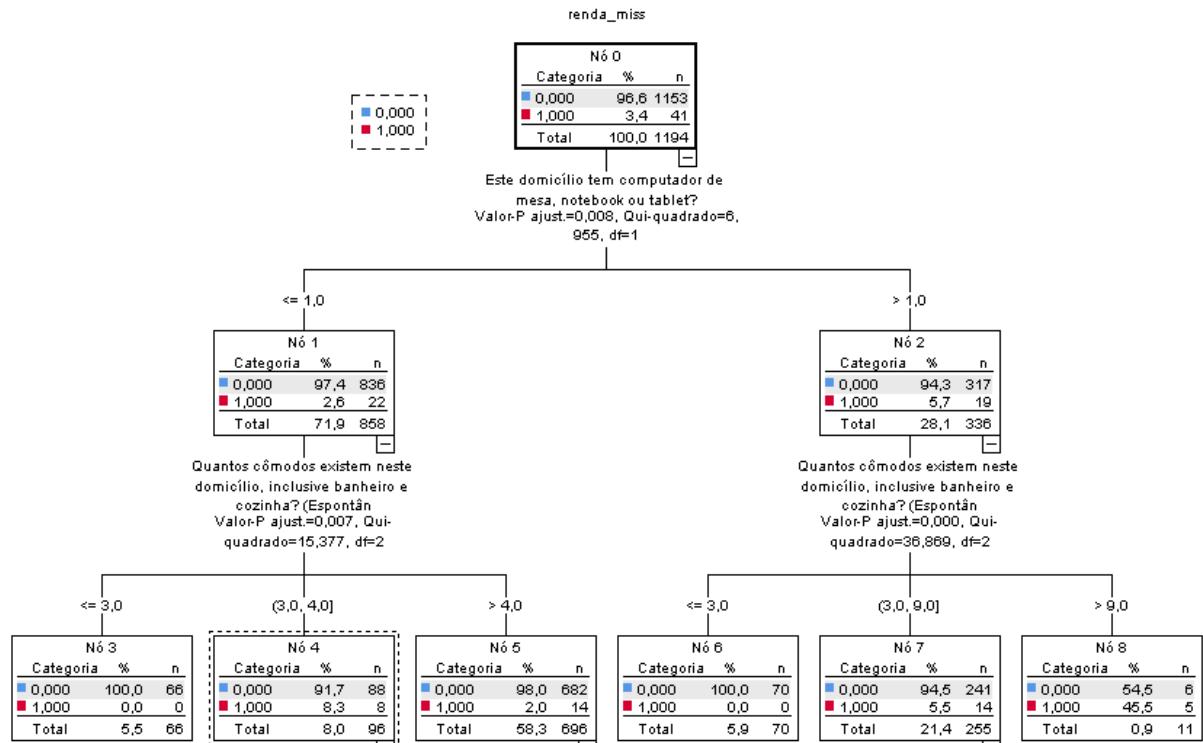


Exemplo de Cluster Hierárquico

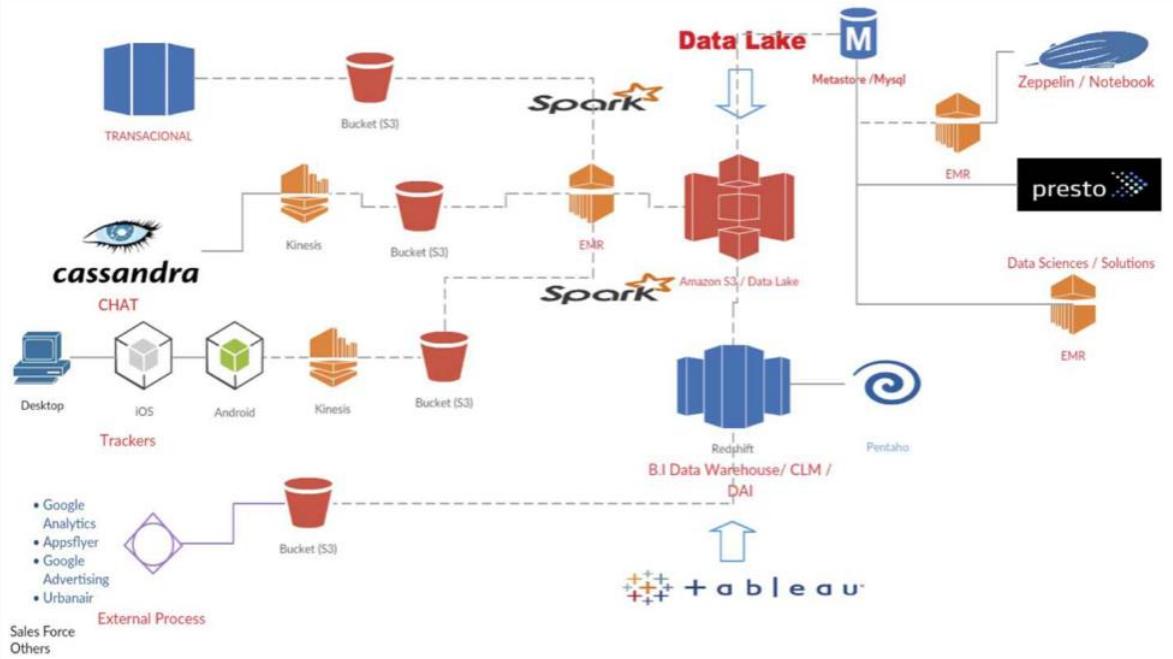


Técnicas de Discriminação

Exemplo de Árvore de Decisão



Plataforma de Dados

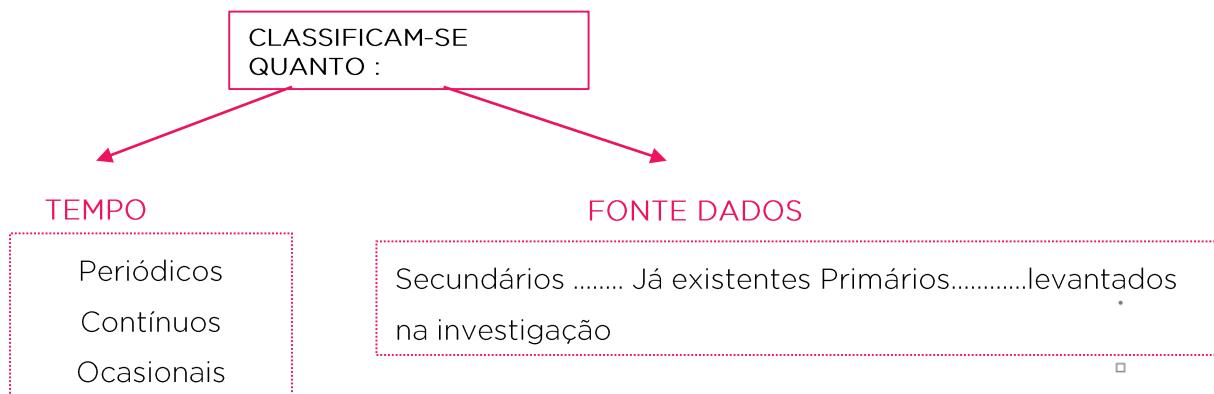


Fonte: <https://www.infoq.com/br/presentations/data-science-na-olx>

DADOS

LEVANTAMENTO **DE DADOS**

“É A OPERAÇÃO DE COLETA PARA DESCRIÇÃO E/OU ANÁLISE DAS CARACTERÍSTICAS DE UMA POPULAÇÃO”



LEVANTAMENTO DE DADOS

Exemplos de dados secundários do IBGE :

- Pesquisa Mensal de Emprego.
- Pesquisa Industrial Mensal de Empregos e Salários.
- Pesquisa Mensal de Comércio.
- Pesquisa Nacional de Saúde.

- Censo Demográfico.
- Pesquisa de Orçamentos Familiares (POF).
- Pesquisa Nacional por Amostra de Domicílios (PNAD).
- Contagem Populacional.

Link: <https://www.ibge.gov.br/>

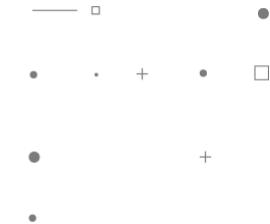
LEVANTAMENTO DE DADOS

- Exemplos de dados secundários da Agência Nacional de Saúde suplementar (ANS)

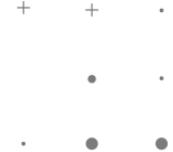
<http://www.ans.gov.br/anstabnet/>

- Pesquisa Mensal de Comércio
- Susep

<http://www2.susep.gov.br/menuestatistica/Autoseg/menu1.aspx>

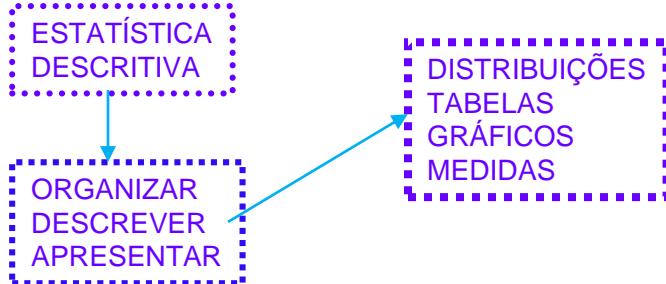


ESTATÍSTICA **DESCRITIVA**





ESTATÍSTICA



Estatística Descritiva

Tem por objetivo organizar, descrever e apresentar os dados, de uma determinada população, em tabelas, gráficos e medidas de resumo.

População



População

Elementos ($N=8$)

Variáveis (atividade física, sexo, idade, filhos ...)



Quais as ocorrências possíveis para atividade física?

Como você representaria essas ocorrências?

Apresentação dos dados

Arquivo

estrutura matricial : linhas e colunas

ordem	Sexo	Atividade física	Estado civil	Grupo
1	F	Sim	Solteira	1
2	M	Sim	solteiro	1
3	F	Não	Casada	2
4	M	Não	Casado	2
5	F	Não	Casada	3
6	M	Não	Casado	3
7	F	Não	Solteira	3
8	M	Não	Solteiro	3

Apresentação dos dados

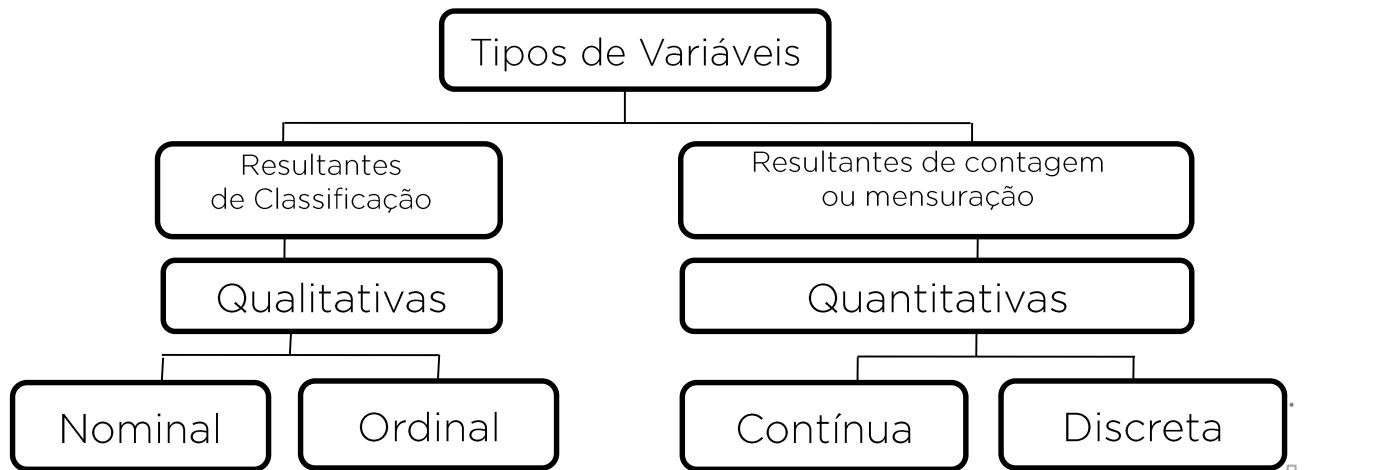
Arquivo

estrutura matricial : linhas e colunas

Grupo	Masculino	Feminino	Atividade física_Sim	Atividade física_Nao	Solteira	Casada
1	1	1	2	0	2	0
2	1	1	0	2	0	2
3	2	2	0	4	2	2



Escala de Mensuração



Exemplo: Sexo Classe social

Idade

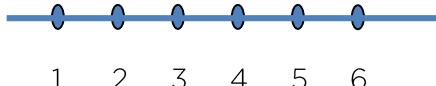
Número de filhos

Escala de Mensuração

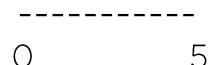
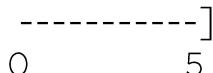
Variável qualitativa nominal: não existe nenhuma ordenação nos possíveis resultados. CATEGORIAS

Variável qualitativa ordinal: os possíveis resultados são ordenados. POSTOS

Variável quantitativa discreta: resultam de operação de contagem



Variável quantitativa contínua: possíveis resultados (valores) formam um intervalo de números reais



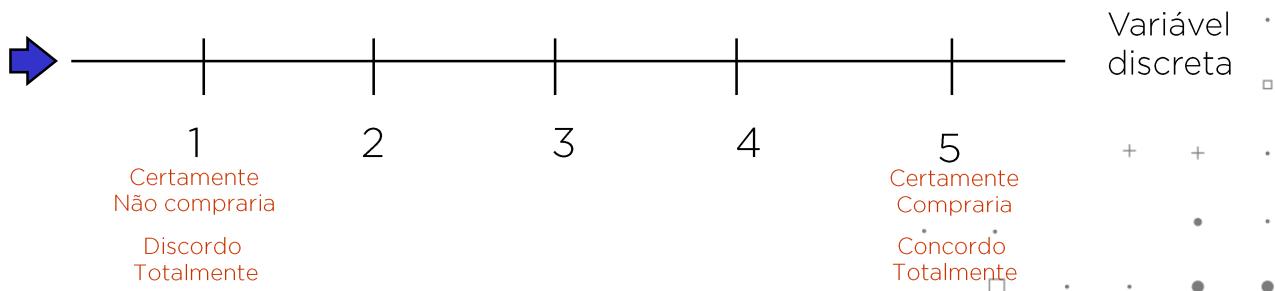
Aplicando conhecimento

Classifique cada variável de acordo com seu tipo:

Variável	Ocorrência	Tipo (escala de mensuração)
Estado civil	Solteiro	Qualitativa Nominal
	Casado	
	Viúvo	
	Divorciado	
Faz atividade física	0=Não ; 1=Sim	Qualitativa Nominal
Idade (anos)	[0 – 110]	Quantitativa contínua
Anos de estudo	[0 – 99]	Quantitativa contínua

Escala de Mensuração

Exemplo: Escala de questionário:

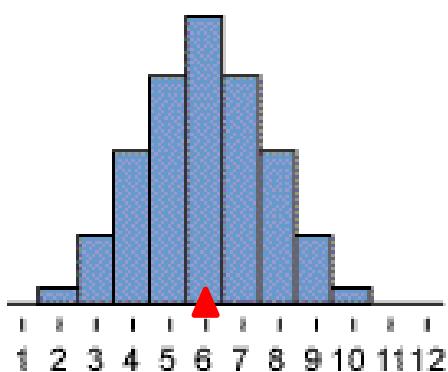


Medidas Resumo

São estatísticas que resumem, em um único valor, a tendência central (média, mediana, moda), a variabilidade (variância, desvio padrão) e a forma da distribuição (simétrica ou assimétrica) da variável.

Medidas Resumo

Distribuição simétrica



Distribuição do tempo de uso de internet (horas)

Medidas de tendência central:

- Média
- Mediana
- Moda

Indicam o centro da distribuição de frequências ou a região de maior concentração de frequência na distribuição.

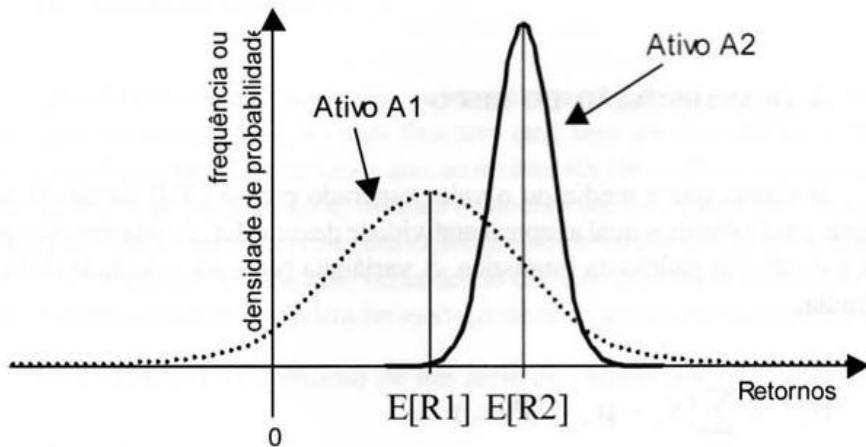
Medidas de dispersão:

- Variância
- Desvio padrão

Indicam o grau de homogeneidade dos valores, até que ponto eles se encontram concentrados ou dispersos da média.

Medidas Resumo

Decisão pela média



Qual ativo você escolheria para investir? Justifique sua escolha.

Medidas Resumo

Exemplo 2

Durante uma verificação de qualidade no conteúdo de seis recipientes de café instantâneo, foram obtidas as seguintes notas:

6,03 5,59 6,40 6,00 5,99 6,02

Qual a média e a mediana encontrada?

Média aritmética: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ $\rightarrow \bar{x} = \frac{6,03 + 5,59 + 6,40 + 6,00 + 5,99 + 6,02}{6} \rightarrow \bar{x} = 6,00$

Mediana: 5,59 5,99 6,00 6,02 6,03 6,40

$$\text{mediana} = \frac{6,00 + 6,02}{2} = 6,01$$

Medidas Resumo

Exemplo 1

Durante uma verificação de qualidade no conteúdo de seis recipientes de café instantâneo, foram obtidas as seguintes notas:

6,03 5,59 6,40 6,00 5,99 6,02

Qual a média e a mediana encontrada?

$$\bar{x} = 6,00 \quad \text{mediana} = 6,01$$

Suponha que o terceiro valor tenha sido incorretamente medido e que na verdade seja de 6,04. Determine novamente a nota média e mediana.

Média aritmética:

$$\bar{x} = \frac{6,03 + 5,59 + 6,04 + 6,00 + 5,99 + 6,02}{6} = 5,95$$

Mediana:

5,59 5,99 6,00 6,02 6,03 6,04



$$\text{mediana} = \frac{6,00 + 6,02}{2} = 6,01 \square$$

Medidas Resumo

Comparação entre Média, Mediana e Moda

	VANTAGENS	LIMITAÇÕES	TIPO DE VARIÁVEIS
MÉDIA	Reflete todos os valores da amostra	É influenciada por valores extremos	Contínua e discreta
MEDIANA	Menos sensível a valores extremos que a média	Mais difícil de ser determinada para grande quantidade de dados	Contínua e discreta
MODA	Representa um valor típico	Não tem função em certos conjuntos de dados	Contínua, discreta, nominal e ordinal

Medidas Resumo

MEDIDAS DE POSIÇÃO - MÉDIA

- Média Aritmética Simples:

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

- Média Aritmética Ponderada:

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

- Média Geométrica (evolução):

$$Mg = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

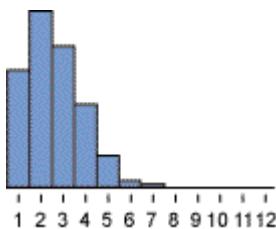
- Média Quadrática:

$$\bar{x}^2 = \frac{\sum_{i=1}^n x_i^2}{n}$$

Medidas Resumo

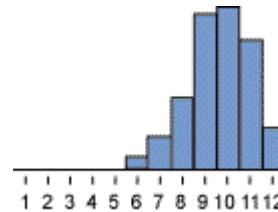
Decisão pela média ?????

Assimétrico à direita



Média > Mediana

Assimétrico à esquerda



Média < Mediana

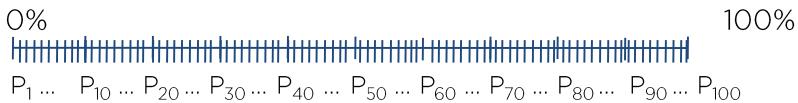
Medidas Resumo

- Outras Medidas de Posição

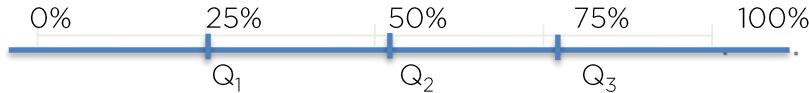
Decis: dividem um conjunto de dados em dez partes iguais.



Percentis (P): dividem a série em cem partes, de modo que $p\%$ ficam abaixo dele (P).



Quartis: dividem a série em quatro partes iguais.

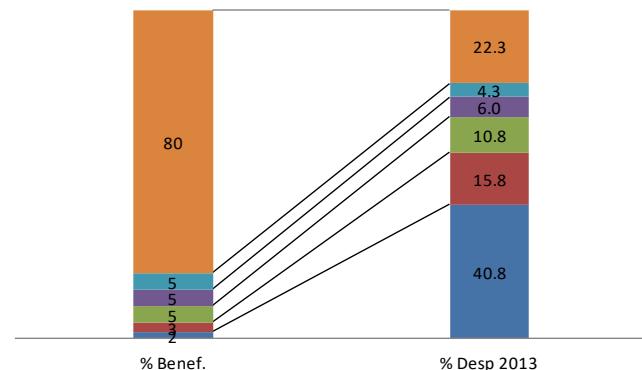


Medidas Resumo

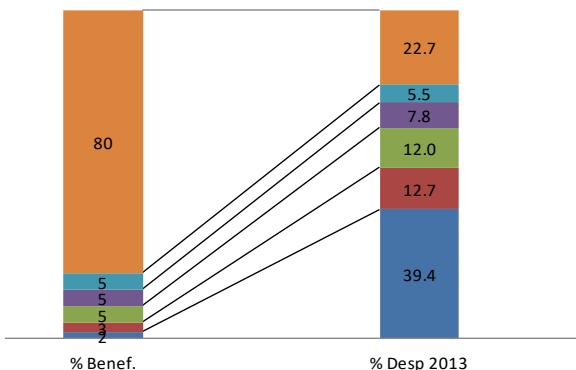
Exemplo: Despesas

Gráfico de Pareto de despesas

Casos ativos



Controle



Medidas Resumo

Economia nacional

São Paulo, Rio e Brasília respondem por 21% do PIB brasileiro

*Andrea Bruxellas
Direto do Rio de Janeiro
Especial para o Terra*

Os municípios de São Paulo, Rio de Janeiro e Brasília respondiam por 21% do Produto Interno Bruto brasileiro em 2007. Segundo dados divulgados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) nesta quarta-feira, a capital paulista responde pela maior fatia do PIB brasileiro, gerando 12% de toda riqueza produzida no País, seguida do Rio de Janeiro (5,2%), Brasília (3,8%), Belo Horizonte (1,4%) e Curitiba (1,4%).

"Com os dados de 2007 a gente pode notar uma estabilidade na série. Ou seja, na série inteira a gente vê que a renda ainda está muito concentrada em alguns municípios e isso é bastante estável. Nas cinco principais cidades a gente tem um quarto do PIB. Tirando essas cidades, a economia está concentrada em 50 cidades que geram 50% da riqueza do País", disse a coordenadora do IBGE Sheila Cristina Zani.

Já os menores PIB do Brasil foram verificados em Santo Antônio dos Milagres (PI), São Miguel da Baixa Grande (PI), Areia de Barúnas (PB), São Luís do Piauí (PI) e Olho D'Água do Piauí (PI). Segundo o IBGE, a soma dos PIB destes cinco municípios representava 0,001% da riqueza produzida em todo País em 2007.

Medidas Resumo

- Medidas de Dispersão

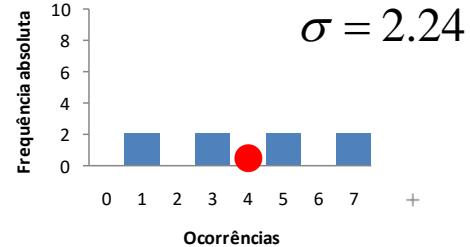
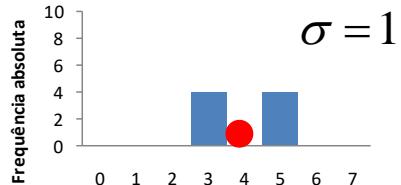
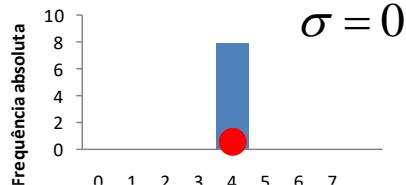
Exemplo 8:

A: 4, 4, 4, 4, 4, 4, 4 ,4

B: 3, 3, 3, 3, 5, 5, 5, 5

C: 1, 1, 3, 3, 5, 5, 7, 7

Qual o desvio padrão?



● Média

Medidas de Dispersão

Medidas de Dispersão: variância e desvio padrão

Exemplo C

X	Média	(X-Média)	$(X-\text{Média})^2$
1	4	-3	9
1	4	-3	9
3	4	-1	1
3	4	-1	1
5	4	1	1
5	4	1	1
7	4	3	9
7	4	3	9
Soma	-	0	40

Variância:

$$\sigma^2 = \frac{40}{8} = 5$$

Desvio padrão:

$$\sigma = \sqrt{\sigma^2} = \sqrt{5} = 2.24$$

Medidas de Dispersão

O quanto os pontos (dados) estão distantes da média (ponto central)

$$\text{➤ variância da população } \sigma^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}$$

$$\text{➤ variância da amostra } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Medidas Resumo

Medidas de Assimetria

As medidas de assimetria referem-se à forma da curva que representa a distribuição de frequência. A assimetria é o afastamento da simetria de uma frequência.

- Curvas de frequência simétrica ou em forma de sino: caracterizam-se pelo fato das observações equidistantes do ponto central terem a mesma frequência (curva normal)
- Curvas de frequência moderadamente assimétricas ou desviadas: a cauda de um lado da ordenada máxima é mais longa do que do outro. Se o ramo mais alongado fica à direita, a curva é dita de assimetria positiva, enquanto que, se ocorre o inverso, diz-se que a curva é de assimetria negativa.

Medidas Resumo

- Coeficientes de Assimetria (Skewness)

$$\Rightarrow As = \frac{m^3}{\sigma^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{[\frac{1}{n} (x_i - \bar{x})^2]^{\frac{3}{2}}}$$

As=0 → simétrica

As>0 → assimétrica positiva

As<0 → assimétrica negativa

Índice de Assimetria (Pearson)

$$\Rightarrow A = \frac{\text{média} - \text{moda}}{\text{desviopadão}}$$

|A|<0,15 → simétrica

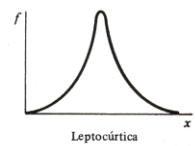
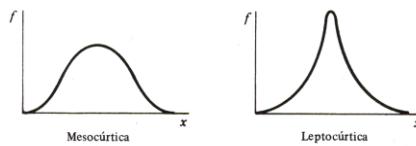
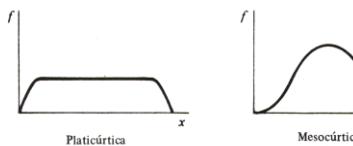
0,15<|A|<1 → assimetria moderada

|A|>1 → assimetria forte

Medidas Resumo

Medidas de Assimetria

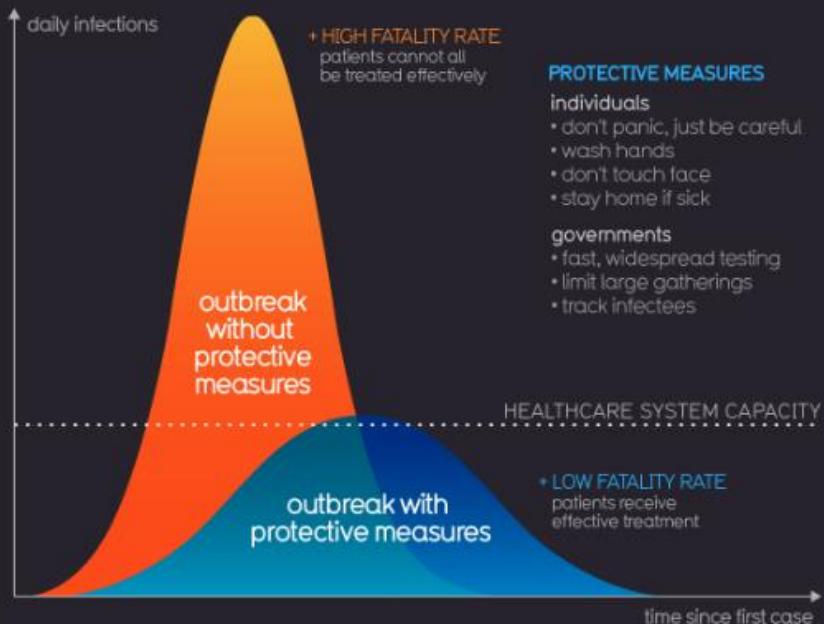
- Curtose: grau de achatamento em relação a uma curva Normal
 - Leptocúrtica (afilado) ➔ $K > 3$
 - Mesocúrtica ➔ $K = 3$
 - Platicúrtica (achatado) ➔ $K < 3$



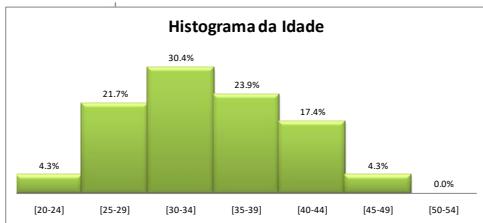
Medidas Resumo

Flattening the Curve

Fast, intelligent action slows pandemic effects, stops the overwhelm of healthcare systems

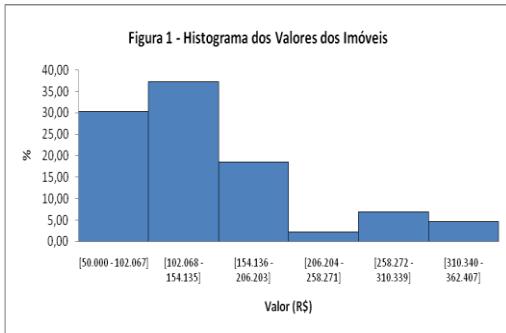


Exemplo de estatística descritiva



idade	
Média	34.6
Erro padrão	1.1
Mediana	34.5
Modo	26
Desvio padrão	6.74
Variância da amostra	45.39
Curtose	-0.54
Assimetria	-0.07
Intervalo	28
Mínimo	20
Máximo	48
Soma	1245
Contagem	36

Exemplo de estatística descritiva



Fonte: Estudo de Caso no Centro de Florianópolis

	Valor (R\$)
Média	144618.3
Erro padrão	10992.8
Mediana	120000.0
Modo	110000.0
Desvio padrão	72084.7
Variância da amostra	5196201097.5
Curtose	1.4
Assimetria	1.4
Intervalo	312400.0
Mínimo	50000.0
Máximo	362400.0
Soma	6218585.0
Contagem	43

Medidas Resumo

Outras Medidas de Dispersão

- Coeficiente de Variação
- Amplitude
- Amplitude Inter-Quartílica

Medidas Resumo

Outras Medidas de Dispersão

Coeficiente de variação

(CV)

É o quociente entre o desvio padrão e a média.

$$CV = \frac{\sigma}{\bar{X}}$$

Vantagem: caracterizar a dispersão dos dados em termos relativos a seu valor médio.

Medidas Resumo

Qual o coeficiente de variação?

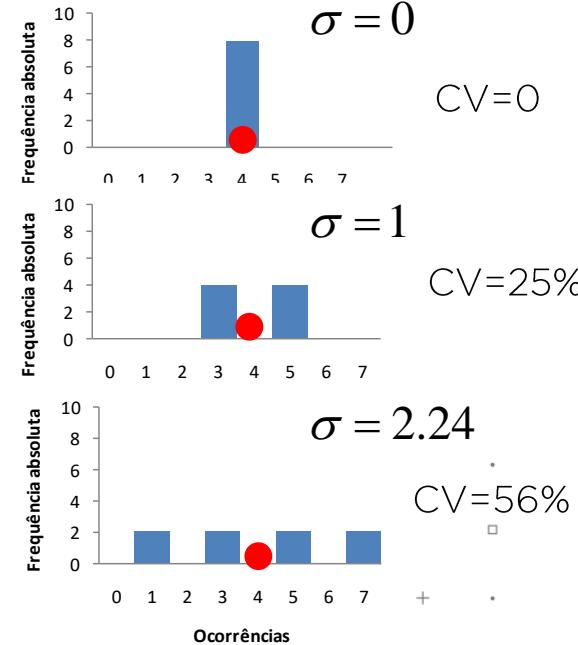
• Medidas de Dispersão

Exemplo 8:

A: 4, 4, 4, 4, 4, 4, 4 ,4

B: 3, 3, 3, 3, 5, 5, 5, 5

C: 1, 1, 3, 3, 5, 5, 7, 7



● Média

Medidas Resumo

Outras Medidas de Dispersão Amplitude

É definida como a diferença entre o maior e o menor valor de um conjunto de dados.
Fortemente relacionado com a dispersão dos dados.

A amplitude pode levar a erros de avaliação, pois não representa o conjunto dos dados.
Muitas vezes reflete muito mal a dispersão dos mesmos.

Medidas Resumo

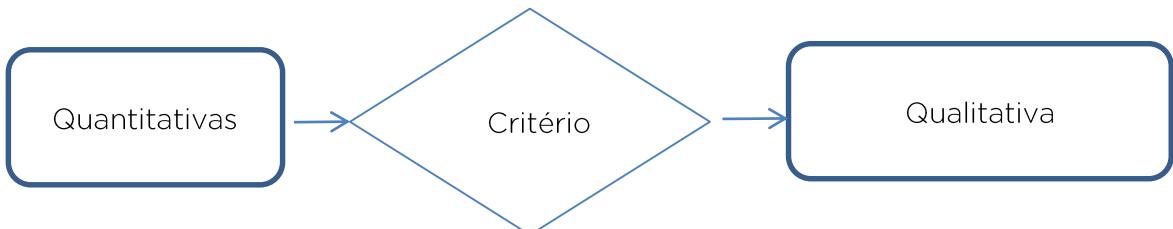
- Outras Medidas de Dispersão

Amplitude Inter-quartílica

É a diferença entre o terceiro e o primeiro quartil (Q3-Q1).

Usada em análise exploratória de dados – gráficos Box Plot.

Transformando variáveis quantitativas em qualitativas

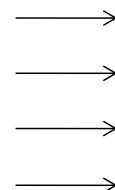


Exemplo:

Anos de
estudo



Critério
0
[1 - 9]
[10 – 12]
≥ 13



Grau instrução
Analfabeto
Fundamental
Médio
Superior

Transformando variáveis quantitativas em qualitativas

Exemplo: Quantas classes serão necessárias para representar a despesa anual?

Fórmula de Sturges

$$\rightarrow K = 1 + 3,3 * \log n$$

Medidas resumo da despesa anual

Mean	Std Dev	Minimum	Maximum	Mode	Range	Sum	N
265,22	537,55	0	4491,19	0	4491,19	16118247,5	60773

$$K = 1 + 3,3 * \log (60773) = 16,78 \sim 17$$

$$\text{Intervalo} = \frac{(Máximo - Mínimo)}{K} = \frac{4491,19}{17} = 264,18 \cong 265$$

K = número de classes

Despesa	N	%	%ac
[0 - 265)	26740	44,0	44,0
[265 - 530)	10939	18,0	62,0
[530 - 795)	4862	8,0	70,0
[795 - 1060)	4254	7,0	77,0
[1060 - 1325)	3646	6,0	83,0
[1325 - 1590)	3039	5,0	88,0
[1590 - 1855)	2431	4,0	92,0
[1855 - 2120)	1823	3,0	95,0
[2120 - 2385)	1215	2,0	97,0
[2385 - 2650)	608	1,0	98,0
[2650 - 2915)	243	0,4	98,4
[2915 - 3180)	243	0,4	98,8
[3180 - 3445)	182	0,3	99,1
[3445 - 3710)	182	0,3	99,4
[3710 - 3975)	122	0,2	99,6
[3975 - 4240)	122	0,2	99,8
[4240 - 4505)	122	0,2	100,0
Total	60773	100,0	

Distribuição de Frequência

O número de vezes que ocorreram valores em cada classe ou valores chama-se frequência absoluta. O conjunto das ocorrências, com correspondentes frequências absolutas (FA) e relativas (FR), define a distribuição de frequências da variável. Conhecer o comportamento da variável.

Distribuição etária dos trabalhadores da Empresa XXX, 01/05/2019

Faixa etária	Frequency	Percent	Cumulative Frequency	Cumulative Percent
00 - 17	19052	33,8	19052	33,8
18 - 29	16143	28,6	35195	62,4
30 - 39	13710	24,3	48905	86,7
40 - 49	5773	10,2	54678	96,9
50 - 59	1559	2,8	56237	99,7
60 - 69	174	0,3	56411	100,0
Acima 69	13	0,0	56424	100,0
Total	56424	100,0	.	.

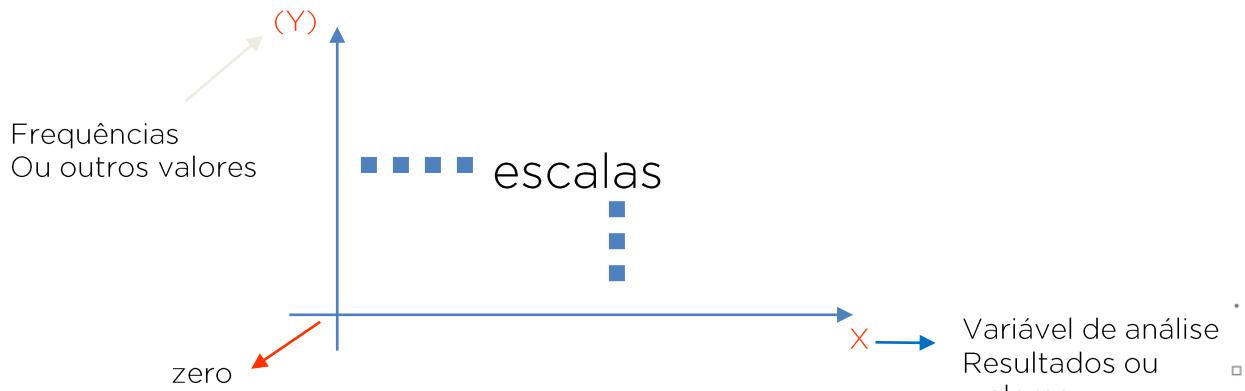
Apresentação Gráfica dos Dados

As regras básicas de elaboração de um gráfico são:

- simplicidade
- clareza
- veracidade

Apresentação Gráfica dos Dados

► EIXOS CARTESIANOS



Apresentação Gráfica dos Dados

- Variáveis qualitativas ou quantitativas discretas

Colunas: Um gráfico de colunas ilustra comparações entre itens. As categorias são organizadas na horizontal e os valores são distribuídos na vertical.

Setor ou pizza: Um gráfico de pizza mostra o tamanho proporcional de itens que constituem uma série de dados para a soma dos itens.

A frequência relativa (%) transformada em graus mediante o cálculo proporcional.

Bolhas: Um gráfico de bolhas é um tipo de gráfico dispersão (x,y). O tamanho do marcador de dados indica o valor de uma terceira variável.

Apresentação Gráfica dos Dados

- Variáveis quantitativas contínuas

Histograma: É formado por retângulos cujas áreas representam frequências dos intervalos de suas classes. Esta apresentação é indicada para séries contínuas, e portanto não há espaço entre as barras.



DETECÇÃO DE OUTLIERS



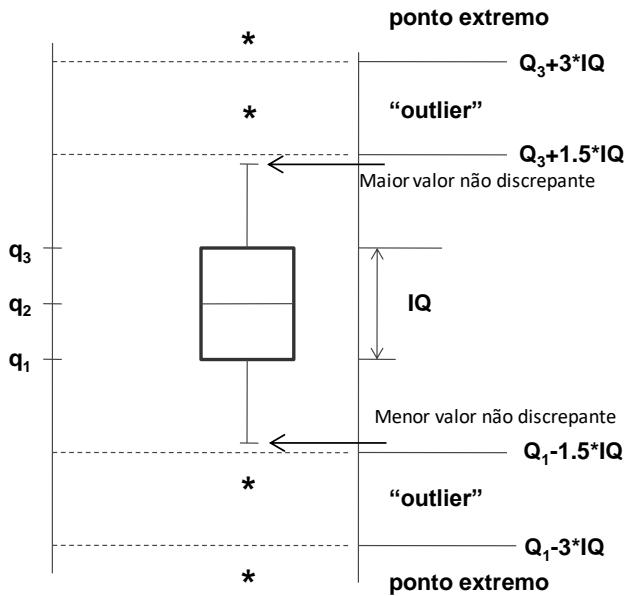
- Detecção de dados suspeitos - “Outlier”

- Dado incorreto
- População diferente
- Dado correto – Evento raro

- Detecção de dados suspeitos - “Outlier”
- +
- Representação Gráfica na Análise dos Dados

O Box Plot (desenho esquemático) informa medidas de posição, dispersão, assimetria, caudas e dados atípicos (outliers). A posição central é dada pela mediana e a dispersão pela amplitude inter-quartílica. As medidas de posição q1, q2 e q3 informam a assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores distantes e pelos valores atípicos.

- Detecção de dados suspeitos - “Outlier”
- +
- Representação Gráfica na Análise dos Dados



Legenda:

Q_1 = quartil 1

Q_2 = quartil 2 = mediana

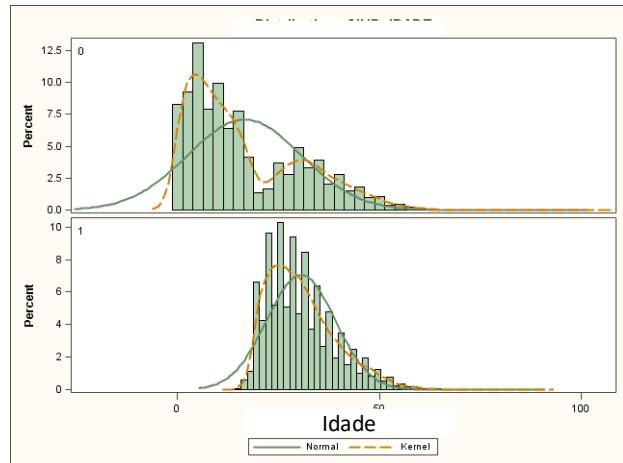
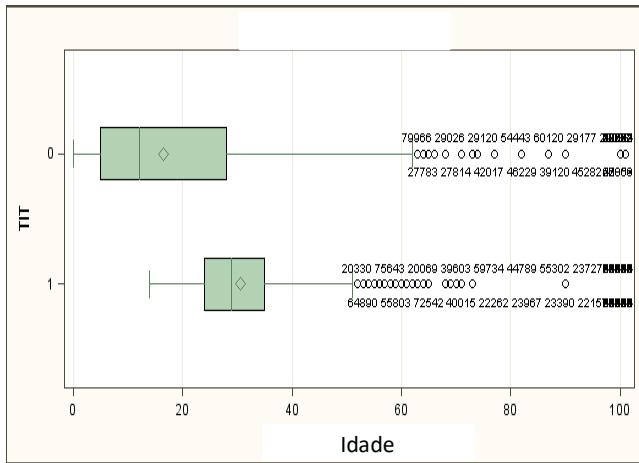
Q_3 = quartil 3

IQ = interquartil

- Detecção de dados suspeitos - “Outlier”
- +
- +
-

- Exemplo

Exemplo



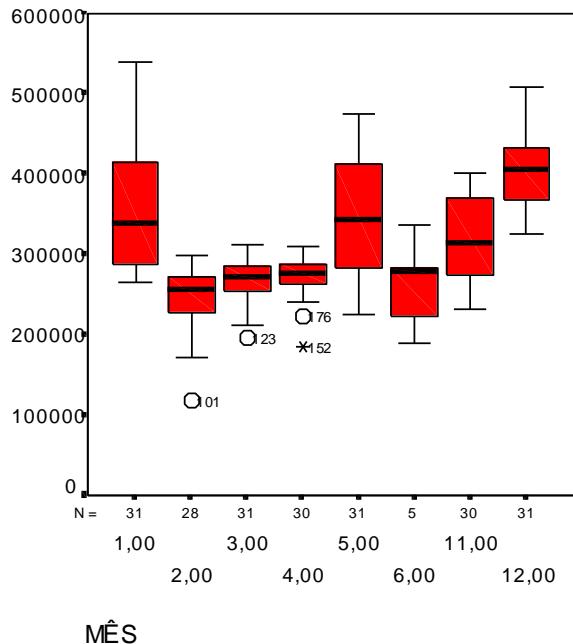
Aplicação

Detecção de dados suspeitos - “Outlier”

Exemplo

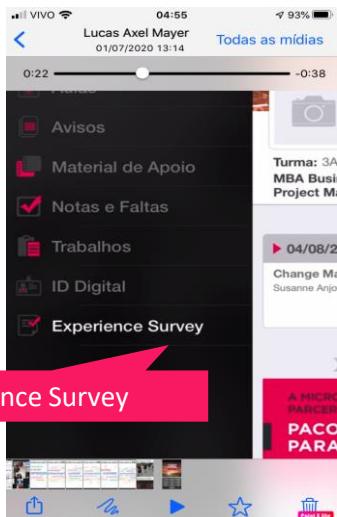
Gráfico Box-Plot

Exemplo: “Total de unidades vendidas por produto - Campanha 1 a 12 de 2016

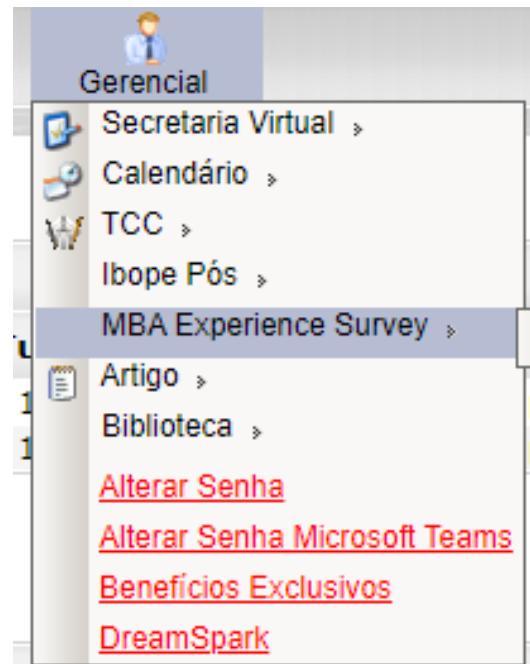


O que você achou da aula de hoje?

Pelo aplicativo da FIAP
(Entrar no FIAPP, e no menu clicar em Experience Survey)



Experience Survey



OBRIGADA



/ Regina T. I. Bernal

FIAP

Copyright © 2022 | Professora Dra. Regina Tomie Ivata Bernal
Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente
proibido sem consentimento formal, por escrito, do professor/autor.

FIAP