
FIAP
MBA

MBA em DATA SCIENCE & ARTIFICIAL INTELLIGENCE

APPLIED STATISTICS

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

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

Atividades Profissionais:

- Professora de pós-graduação na FIAP
- Consultora externa da SVS/MS
- Cientista de Dados em Saúde

Programa

DATA	CONTEÚDO PROGRAMÁTICO
22/01	Introdução; Estatística Descritiva. Exercício prático usando no Python.
24/01	Modelos de distribuição (Probabilidade). Probabilidade condicional. Exercício prático usando no Python
27/01	Inferência Estatística: Amostragem. Exercício prático usando no Python
29/01	Inferência estatística: Teste de hipóteses paramétrico e não paramétrico. Exercício prático usando no Python.
03/02	Correlação de Pearson. Gráfico de Dispersão. Regressão Linear Simples Exercício prático usando no PythonR.
05/02	Regressão Linear Múltipla. Exercício prático usando no Python.
10/02	Regressão Logística. Exercício prático usando no Python
17/02	Análise de Cluster. Exercício prático usando no Python
21/02	Modelos de Séries Temporais. Exercício prático usando no Python

Avaliação da disciplina

Avaliação	Peso
Listas de exercícios	0.5
Projeto Integrado	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.

- Apresentar os conceitos básicos e metodologias para que seja extraído conhecimento de grandes bases de dados.

- Desenvolver conceitos de preparação de dados para fins estatísticos e informações para a geração de competitividade organizacional.

- Proporcionar o conhecimento necessário para reconhecer as seguintes técnicas Supervisionadas (Árvore de Decisão, Regressão Linear e Regressão Logística) e Não Supervisionadas como Componentes Principais e Análise de Cluster .

Referências Bibliográficas

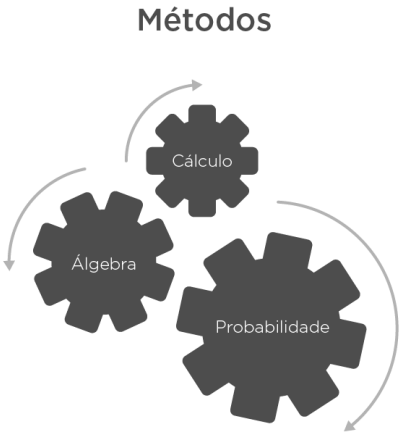
- BERRY, M.J.A.; LINOFF, G. Data Mining Techniques: for marketing, sales, and customer. Wiley Computer Publishing, 1997.
- BUSSAB, W.O.; MORETTIN, P. A., Estatística Básica, 5a. ed., São Paulo: Saraiva, 2006.
- KUHN, M. / JOHNSON, K. , Applied Predictive Modeling, 2013
- HAIR, J.F. / ANDERSON, R.E. / TATHAN, R.L. / BLACK, W.C. Análise multivariada de dados, 2009
- JAMES, G, / WITTEN, D. / HASTIE, T. / TIBSHIRANI, R. Na Introduction to Statistical Learning with Aplications in R, 2013
- LANTZ, B. Machine Learning with R. 2a. ed. Packt Publishing, 2015

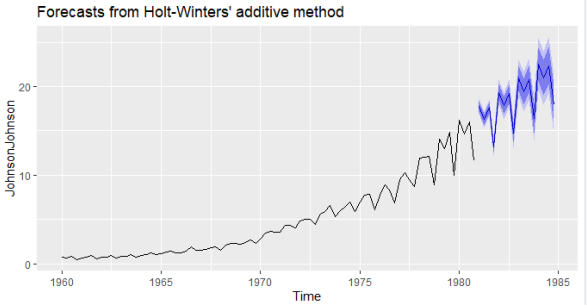
Referências Bibliográficas

- 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**, 2ª ed., São Paulo: Edgard Blücher, 2006.
- SILVA, NN. **Amostragem Probabilística**. 2ª ed., São Paulo: Editora da Universidade de São Paulo, 2001.
- SOARES, J.; FARIAS, A. A.; CESAR, C. C., **Introdução a Estatística,** LTC**,** 2002.
- TORGO, L. Data Mining with R: Learning with Case Studies. 2.a ed. Chapman and Hall/CRC, 2007

DATA ANALYTICS

DATA ANALYTICS





UNIVERSO DE FORNECEDORES

MicroStrategy

IBM

SAP

sas

Microsoft

alteryx

tableau

Qlik

TIBCO The Power of Now

Information Builders

julia

ORACLE

OPEN TEXT The Content Experts

R

R Studio

orange

python

ESTATÍSTICA

COMO TIRAR INFORMAÇÕES DELES

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 DESCRITIVA

ORGANIZAR
DESCREVER
APRESENTAR

DISTRIBUIÇÕES
TABELAS
GRÁFICOS
MEDIDAS

ANÁLISE
INTERPRETAÇAO

INFERÊNCIA ESTÁTISTICA

probabilidade

AMOSTRAGEM
EXPERIMENTOS
MODELAGEM

Estatística

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

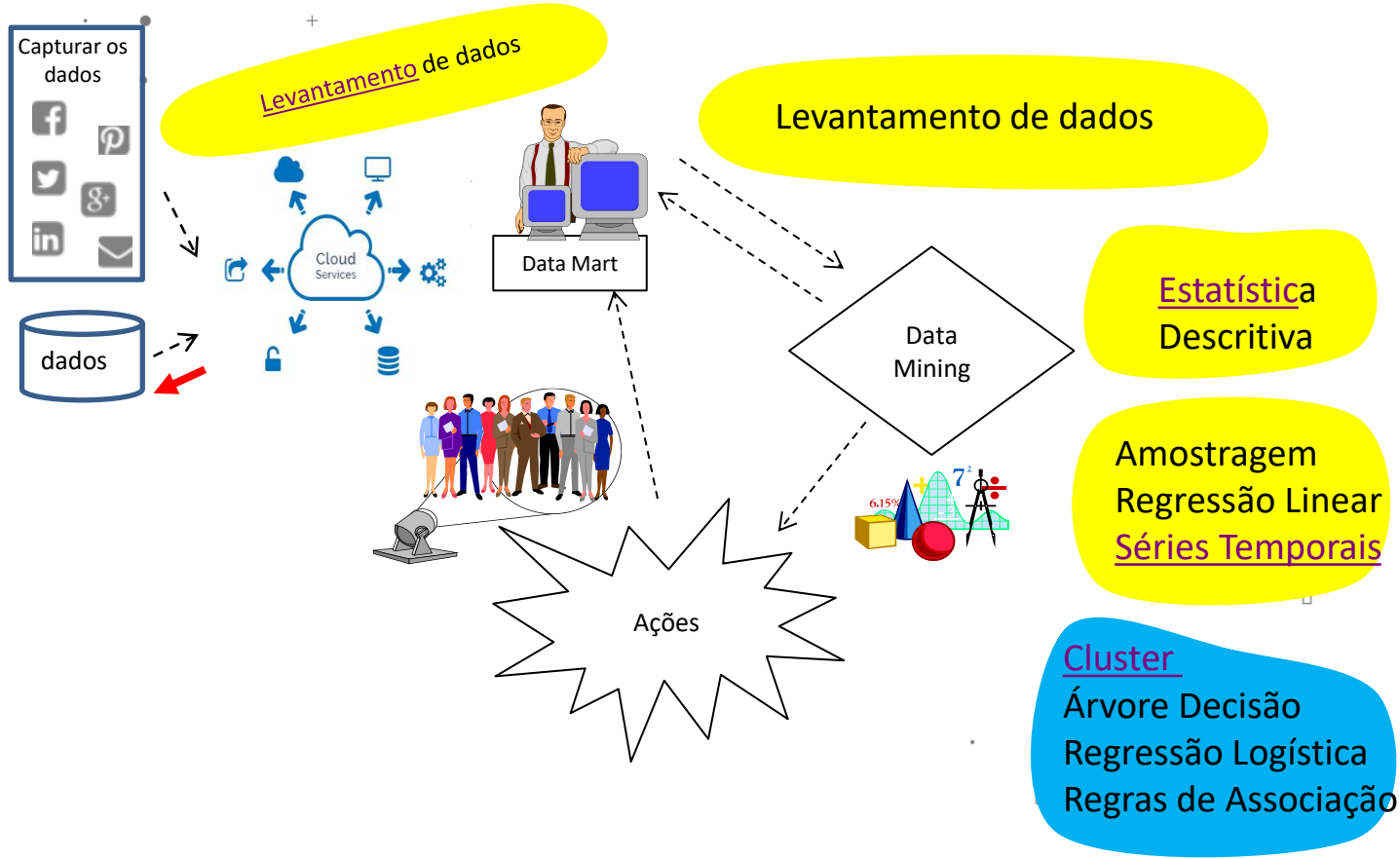
Trata da organização, descrição, apresentação análise e interpretação de dados resultantes da observação de fenômenos coletivos. Produz métodos para inferência estatística.

- ✓ Propriedades

Estuda as variações:

- ➢ entre indivíduos;
- ➢ em um mesmo indivíduo.

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

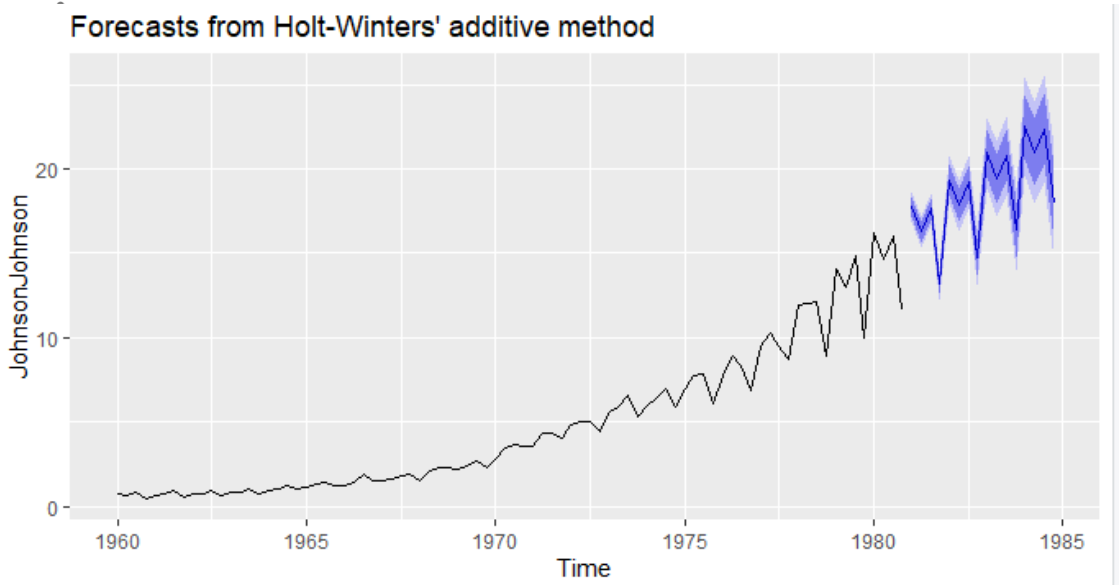
Levantamento de Dados

Atributo	Descrição		
	De	Para	
Internet:	Acesso últimos 3 meses		
	1	1	
	3	0	
Anos de Estudo:	Criação da Faixa Grau de Instrução		
	|0-4|	1	
	|5-8|	2	
	|9-11|	3	
	|12	4	
Renda:	Criação 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	

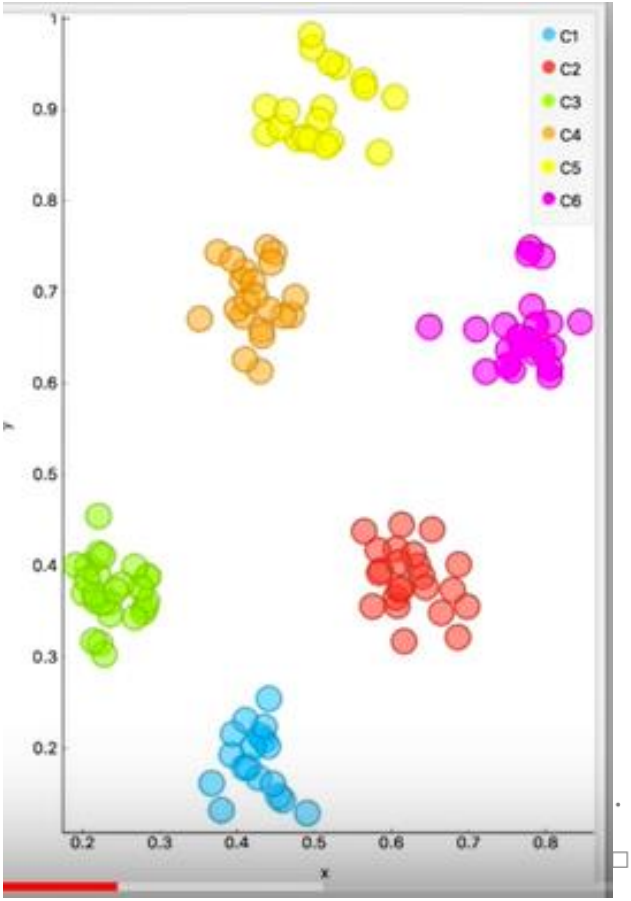
Estatística Descritiva

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

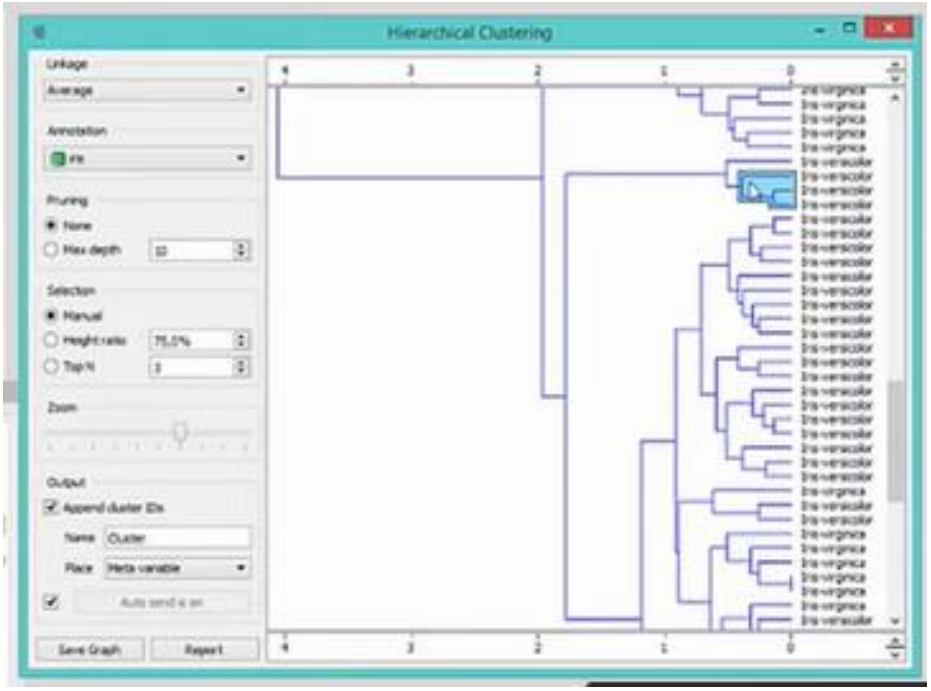
Inferência Estatística



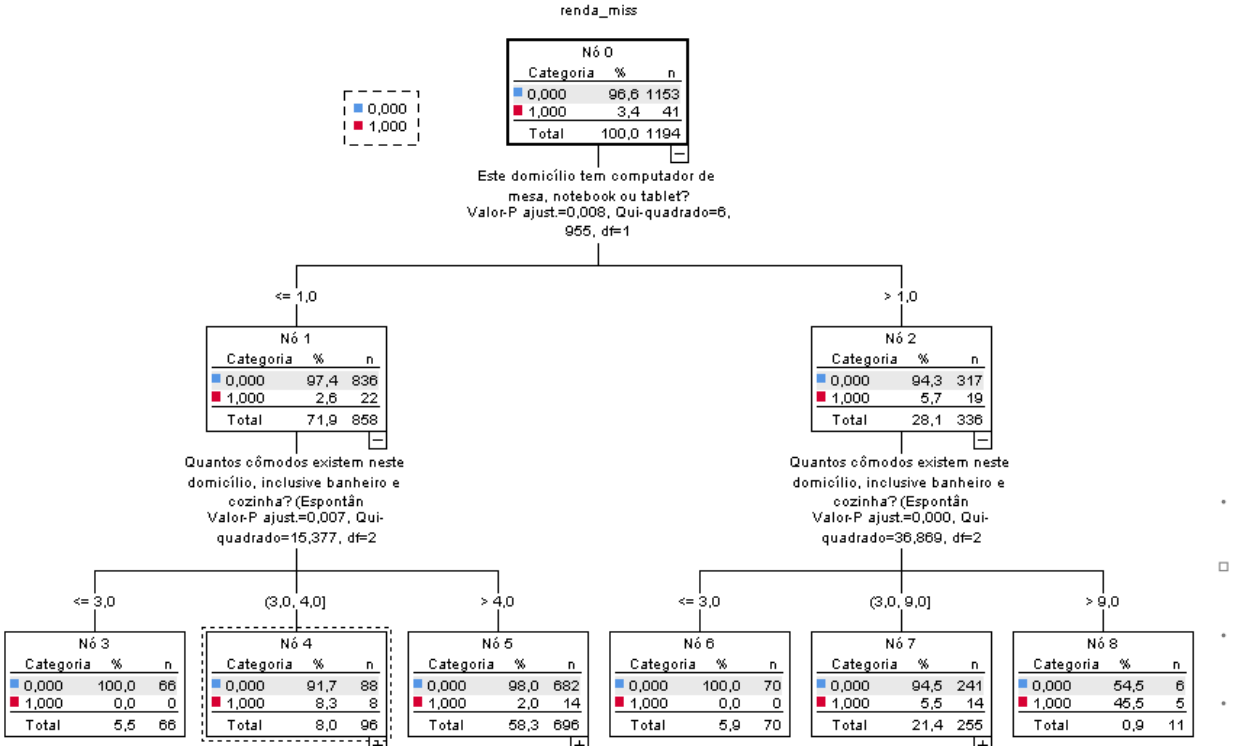
Exemplo de Cluster Não Hierárquico



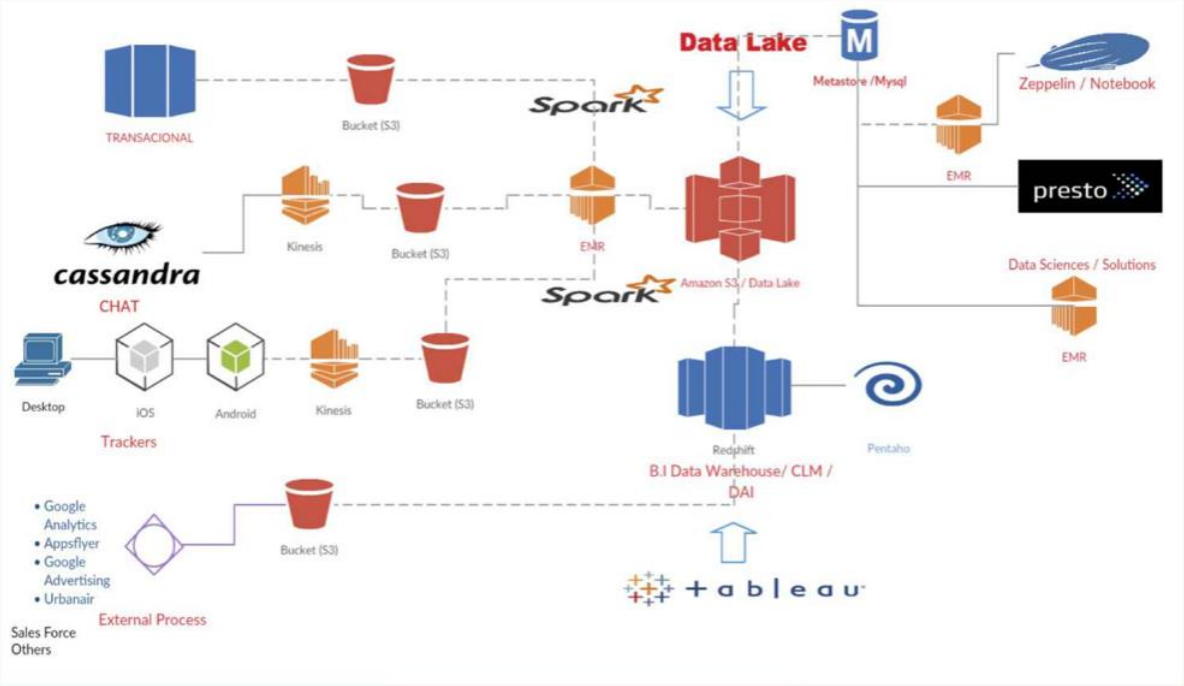
Exemplo de Cluster Hierárquico



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"

CLASSIFICAM-SE QUANTO :

TEMPO

Periódicos

Contínuos

Ocasionais

FONTE DADOS

Secundários ..... Já existentes Primários.....levantados na investigação

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

LEVANTAMENTO DE DADOS

Segmento	Dados	Fonte
Seguradora de veículos	Susep Frota de veículos	http://www2.susep.gov.br/menuestatistica/Autoseg/menu1.aspx https://www.gov.br/transportes/pt-br/assuntos/transito/conteudo-Senatran/frota-de-veiculos-2023
Operadora de Saúde	Agência Nacional de Saúde suplementar (ANS)	http://www.ans.gov.br/anstabnet/
População		https://www.ibge.gov.br

ESTATÍSTICA DESCRITIVA

ESTATÍSTICA

ESTATÍSTICA DESCRITIVA

ORGANIZAR
DESCREVER
APRESENTAR

DISTRIBUIÇÕES
TABELAS
GRÁFICOS
MEDIDAS

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

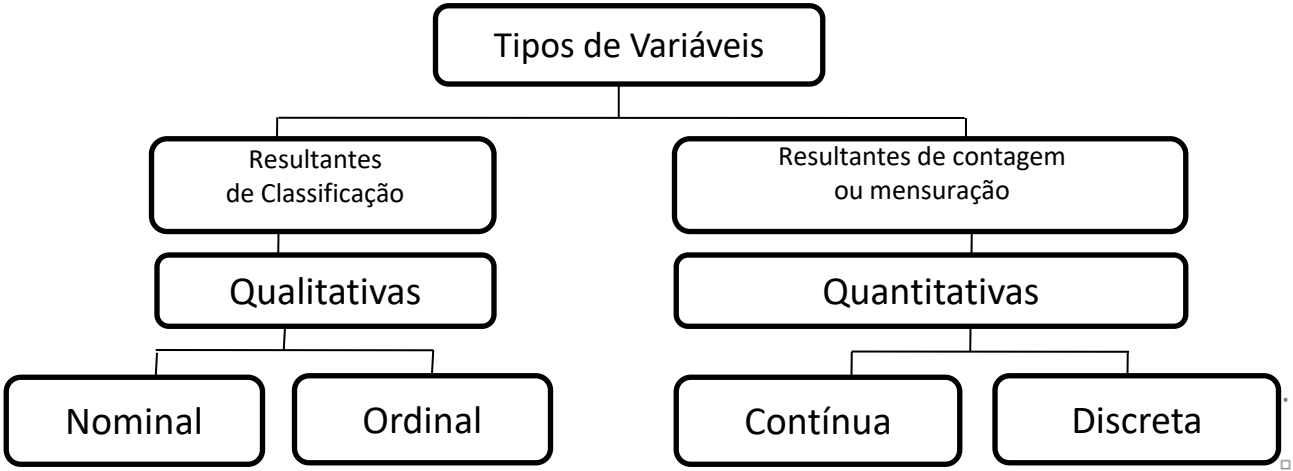
Apresentação dos dados

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

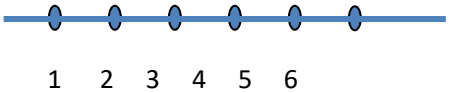


Escala de Mensuração

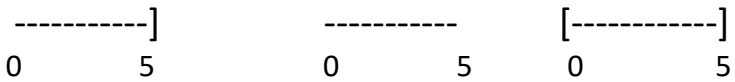
Variável qualitativa nominal: não existe nenhuma ordenação nos possíveis resultados. CATEGORIAS

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

Exercitando!!!!!

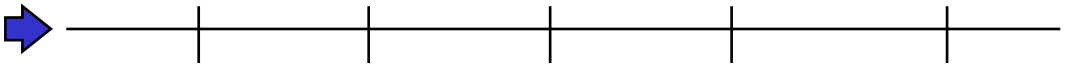
Base
Cadastro

Escala de Mensuração

Exemplo: Escala de questionário:

péssimo ( ) regular ( ) bom ( ) ótimo ( ) excelente ( )

Variável Qualitativa ordinal



1 2 3 4 5

1 Certamente Não compraria

Discordo Totalmente

5 Certamente Compraria

Concordo Totalmente

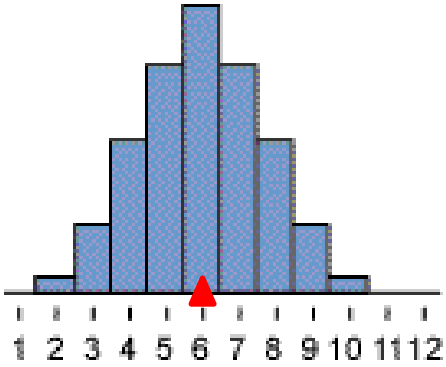
Variável discreta

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.

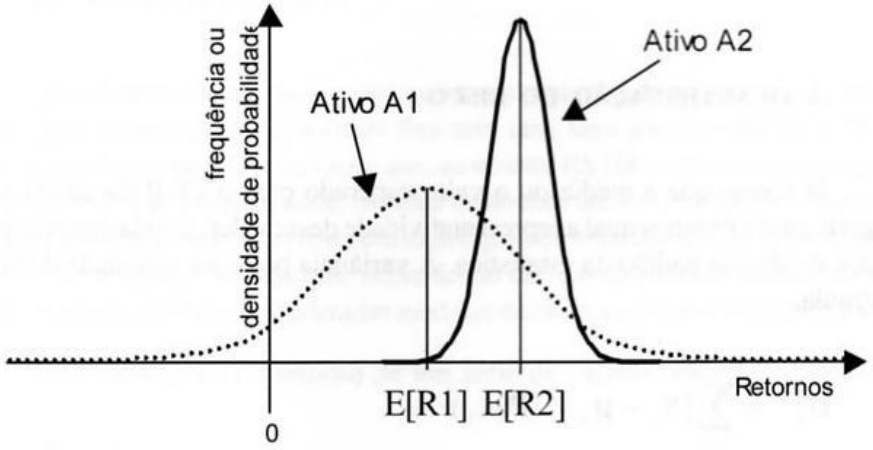
Medidas de dispersão:

- 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

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

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

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

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

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):

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

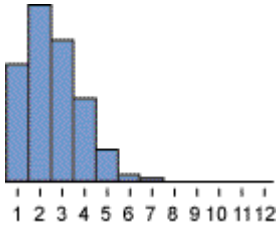
- Média Quadrática:

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

Medidas Resumo

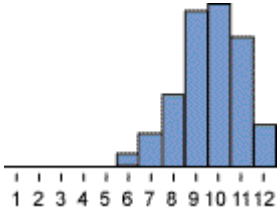
Decisão pela média ?????

Assimétrico à direita



Média > Mediana

Assimétrico à esquerda

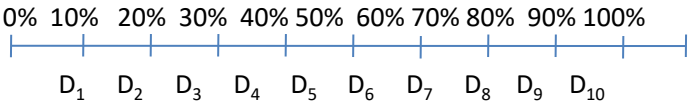


Média < Mediana

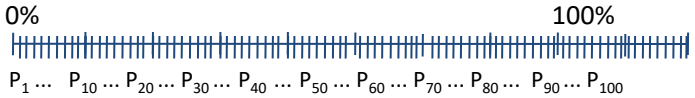
Medidas Resumo

- Outras Medidas de Posição

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



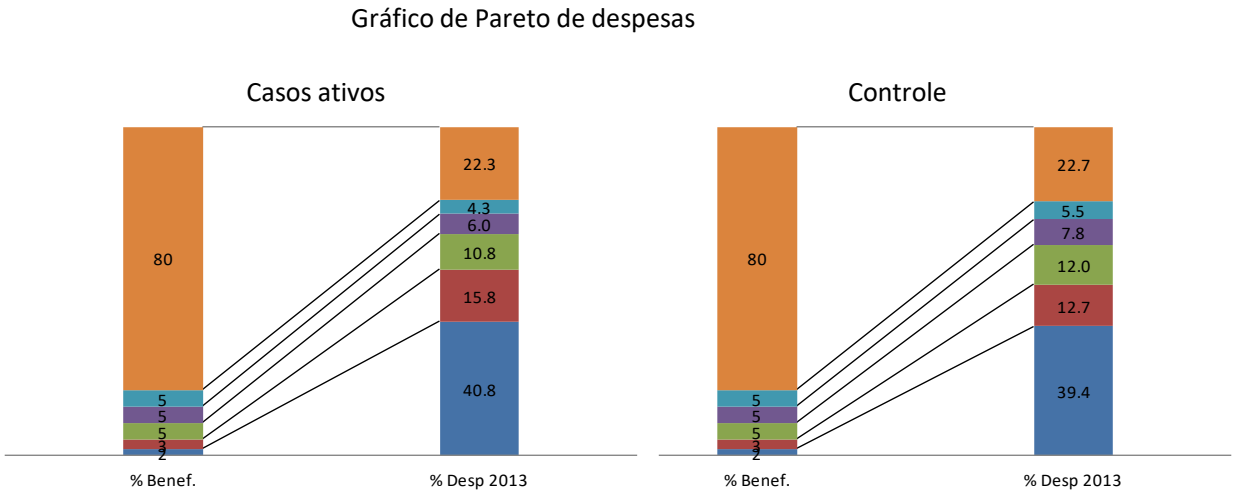
Percentis (P): dividem a série em cem partes, de modo que p% ficam abaixo dele (P).



Quartis: dividem a série em quatro partes iguais.

Medidas Resumo

Exemplo: Despesas



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 cidade a gente tem um quarto do PIB. Tirando essas cidades, a economia esta 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.

Exercitando!!!!!

Segmentação
ABC

Medidas Resumo

- Medidas de Dispersão

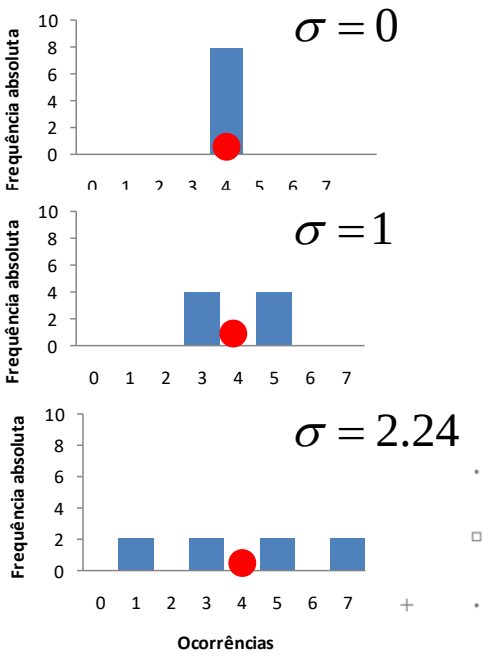
Exemplo 8:

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?



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)

- **variância da população**

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

- **variância da amostra**

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

Medidas Resumo

EXEMPLO

Controle estatístico do processo

O **Controle Estatístico de Processos** (CEP) é uma ferramenta da qualidade utilizada nos processos produtivos (e de serviços) com objetivo de fornecer informações para um diagnóstico mais eficaz na prevenção e detecção de defeitos/problemas nos processos avaliados e, consequentemente, auxilia no aumento da produtividade/resultados da empresa, evitando desperdícios de matéria-prima, insumos, produtos etc.

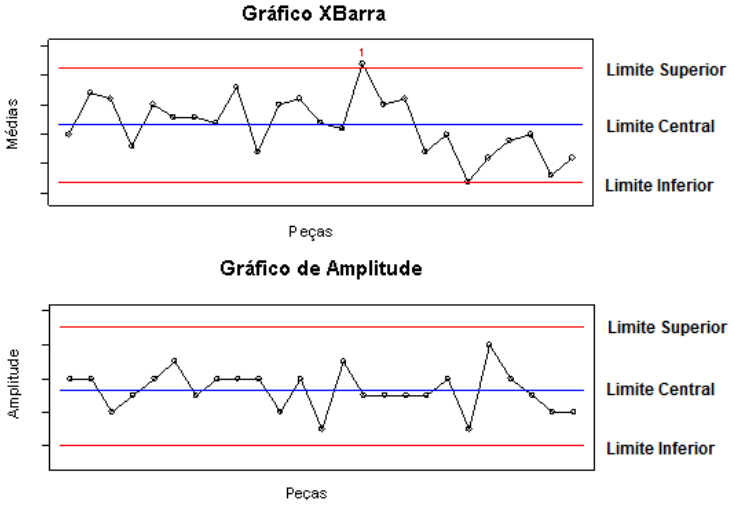
(Fonte: https://pt.wikipedia.org/wiki/Controle_estat%C3%ADstico_de_processos)

Exemplo: Fábrica de Café em Pó

Medidas Resumo

EXEMPLO

Controle estatístico do processo



O gráfico é construído a partir das medidas estatística como:
Média aritmética.
Desvio padrão.
Média das médias.
Somatórios etc.

Gráfico de controle

"Mostrar evidências de que um processo esteja operando em estado de controle estatístico e dar sinais de presença de causas especiais de variação para que medidas corretivas apropriadas sejam aplicadas".

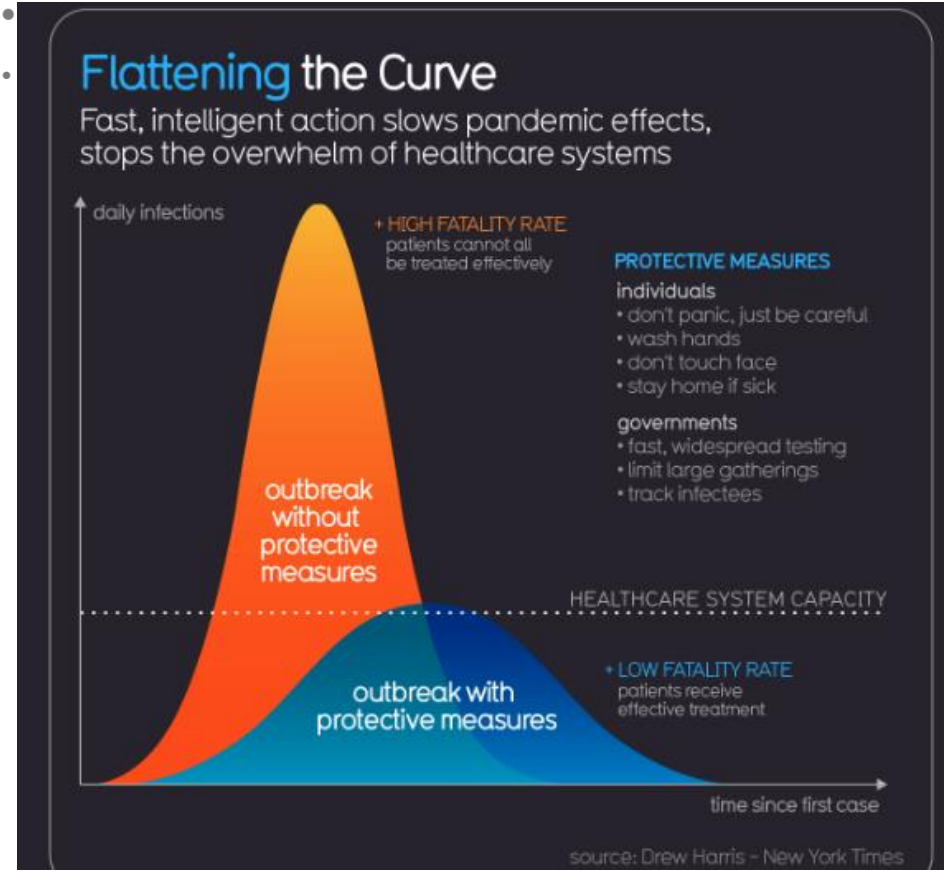
"Manter o estado de controle estatístico estendendo a função dos limites de controle como base de decisões".

"Apresentar informações para que sejam tomadas ações gerenciais de melhoria dos processos".

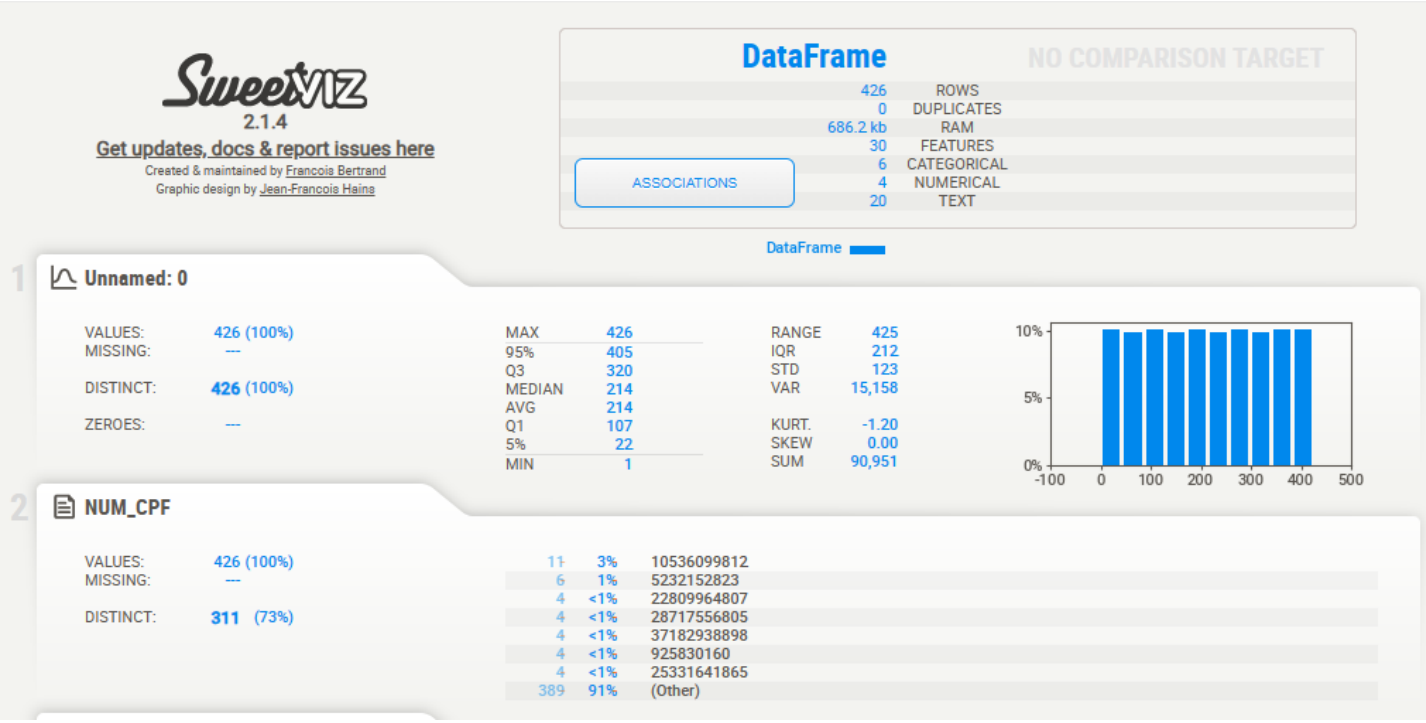
Fonte:http://www.portalaction.com.br/controle-estatistico-do-processo/graficos-ou-cartas-de-controle

MEDIDAS DE ASSIMETRIA

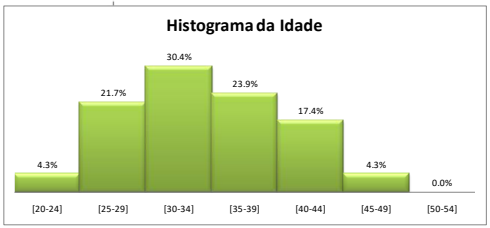
Medidas Resumo



- Exemplo de estatística descritiva da biblioteca SweetViz.

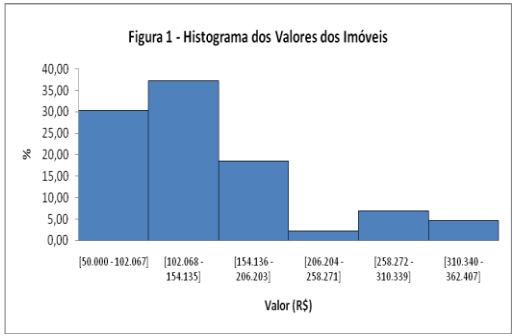


Exemplo de estatística descritiva



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 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 de Assimetria

Coeficientes de Assimetria (Skewness)

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

As=0 ➔simétrica
As>0 ➔assimétrica positiva
As<0 ➔ assimétrica negativa

Índice de Assimetria (Pearson)

➔ $A = \frac{\text{média} - \text{moda}}{\text{desviopadrão}}$

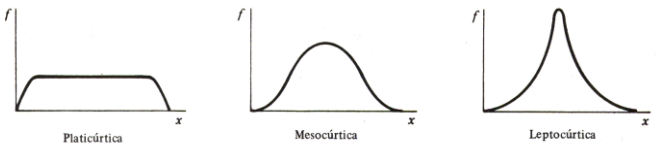
|A|<0,15➔simétrica
0,15<|A|<1➔ assimetria moderada
|A|>1 ➔ assimetria forte

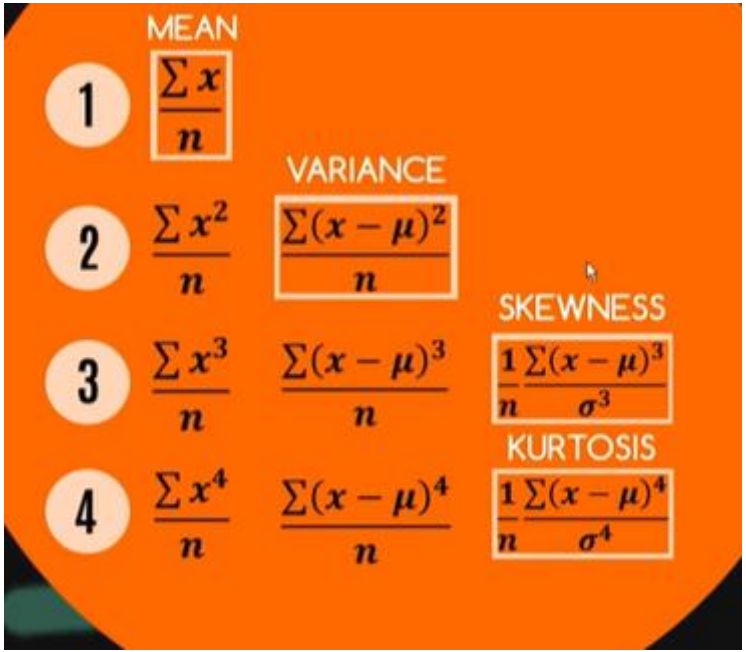
Medidas de Assimetria

Curtose (Kurtosis)

$$As = \frac{m^4}{\sigma^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{[\frac{1}{n}(x_i - \bar{x})^2\}^{\frac{4}{2}}}$$

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




MEAN
VARIANCE
SKEWNESS
KURTOSIS
1
2
3
4
$\frac{\sum x}{n}$
$\frac{\sum x^2}{n}$
$\frac{\sum x^3}{n}$
$\frac{\sum x^4}{n}$
$\frac{\sum (x-\mu)^2}{n}$
$\frac{\sum (x-\mu)^3}{n}$
$\frac{\sum (x-\mu)^4}{n}$
$\frac{1}{n}\frac{\sum (x-\mu)^3}{\sigma^3}$
$\frac{1}{n}\frac{\sum (x-\mu)^4}{\sigma^4}$

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?

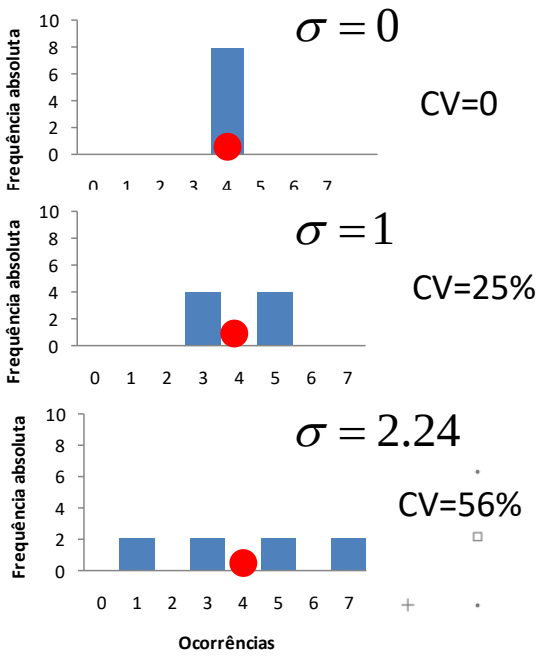
- Medidas de Dispersão

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



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.

DETECÇÃO DE OUTLIERS

Detecção de dados suspeitos - "Outlier"

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

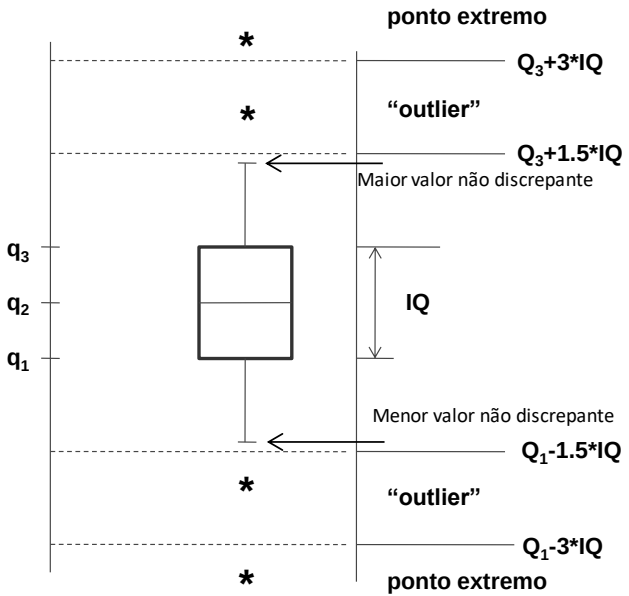
Detecção de dados suspeitos - "Outlier"

- 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



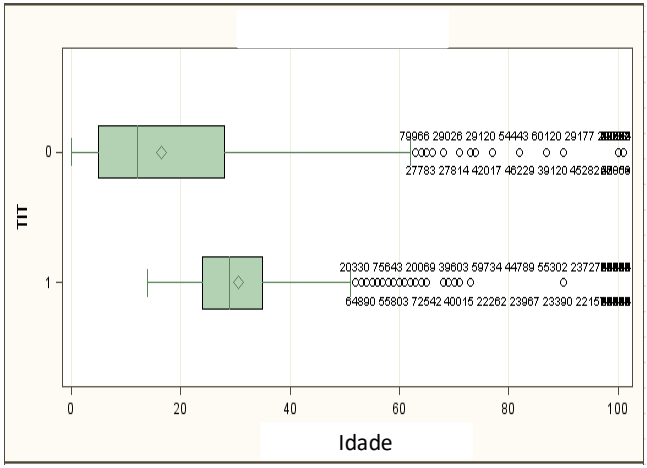
Legenda:

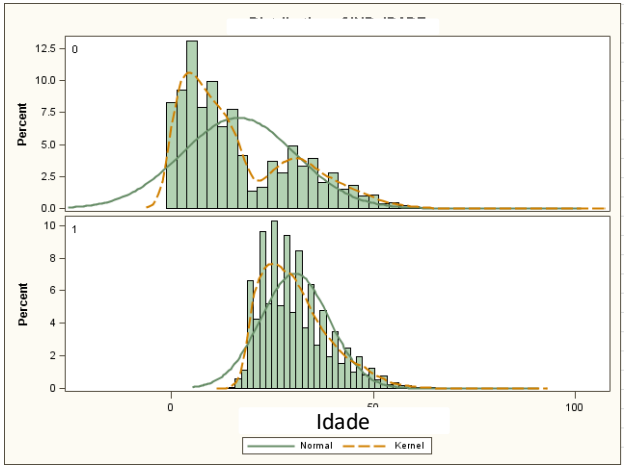
Q1 = quartil 1
Q2 = quartil 2 = mediana
Q3 = quartil 3
IQ = interquartil

Detecção de dados suspeitos - “Outlier”

Exemplo

- Exemplo





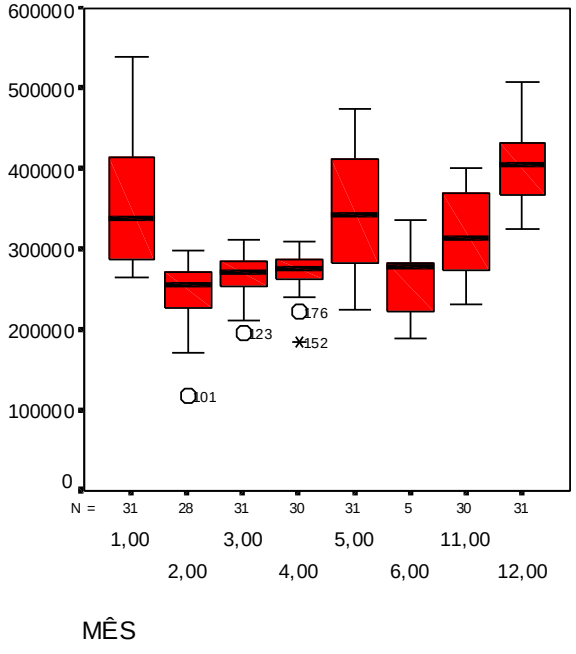
Aplicação

Detecção de dados suspeitos - “Outlier”

Exemplo

Gráfico Box-Plot

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

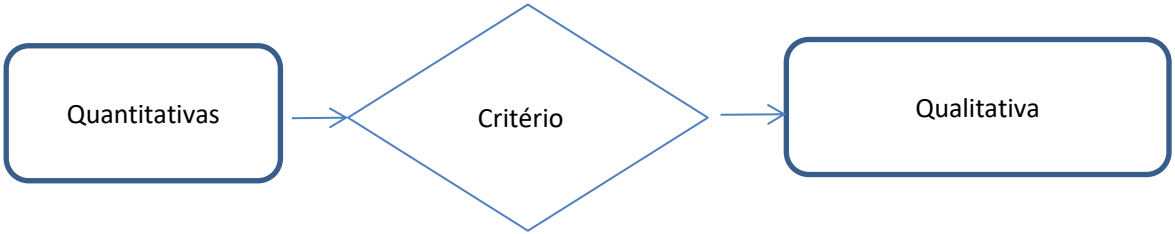


Exercitando!!!!!

Base
Cadastro

TABELAS DE FREQUÊNCIAS

Transformando variáveis quantitativas em qualitativas



Exemplo:

Anos de estudo

Critério	Grau instrução
0	Analfabeto
[1 - 9]	Fundamental
[10 – 12]	Médio
>= 13	Superior

Transformando variáveis quantitativas em qualitativas

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

Fórmula de Sturges ➡ K = 1 + 3,322 * log10(n)

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
[3170 - 3975)	122	0,2	99,6
[3975 - 4240)	122	0,2	99,8
[4240 - 4505)	122	0,2	100,0
Total	60773	100,0	

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 ~ 17

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

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	.	.

GRÁFICOS

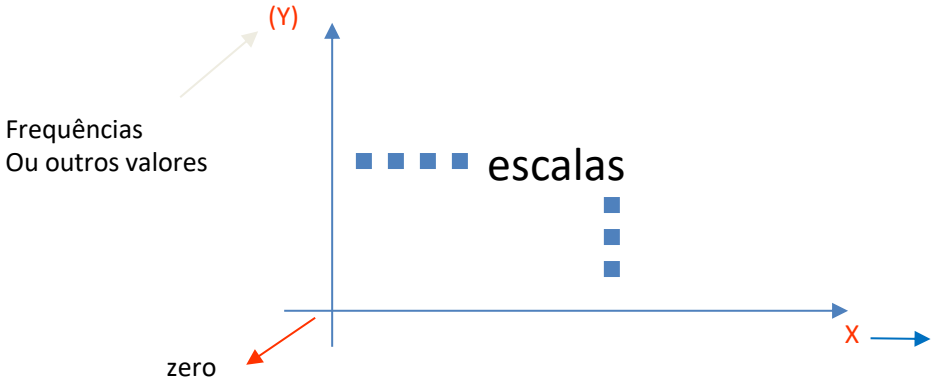
Apresentação Gráfica dos Dados

As regras básicas de elaboração de um gráfico são:

- simplicidade
- clareza
- veracidade

Apresentação Gráfica dos Dados

- EIXOS CARTESIANOS



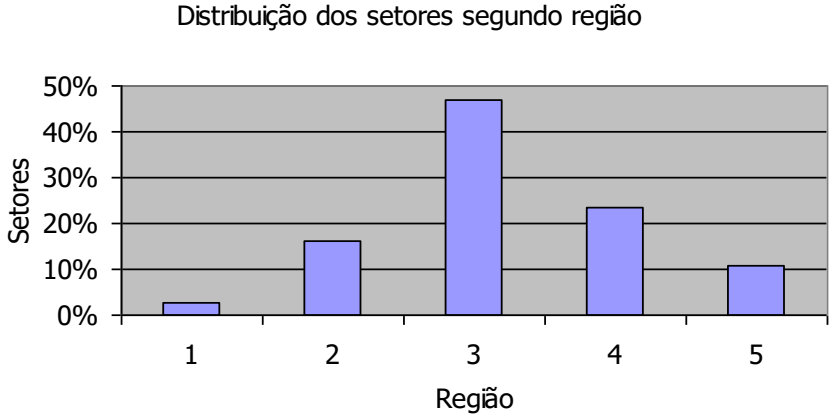
Apresentação Gráfica dos Dados

Variáveis qualitativas ou discretas

a) Colunas

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

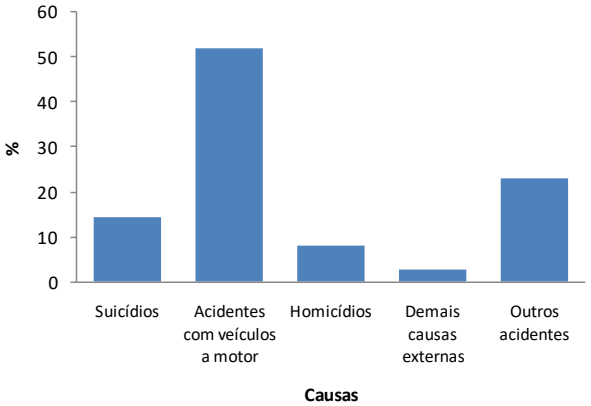
Exemplo:



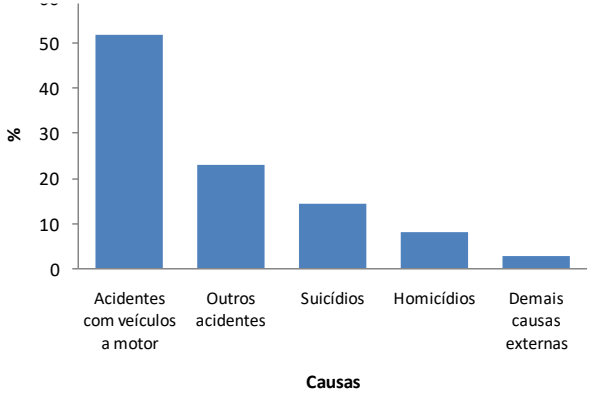
Apresentação Gráfica dos Dados

Variáveis qualitativas

Causa	%
Suicídios	14.2
Acidentes com veículos a motor	52.1
Homicídios	8.1
Demais causas externas	2.6
Outros acidentes	23
Total	100.0



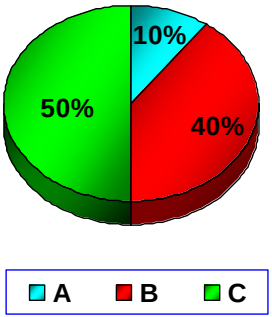
Causa	%
Acidentes com veículos a motor	52.1
Outros acidentes	23
Suicídios	14.2
Homicídios	8.1
Demais causas externas	2.6
Total	100.0



Apresentação Gráfica dos Dados

b) Setores ou pizza

Um gráfico de pizza mostra o tamanho proporcional de itens que constituem uma série de dados para a soma dos itens. A freqüência relativa (%) transformada em graus mediante o calculo proporcional.



100 — 360

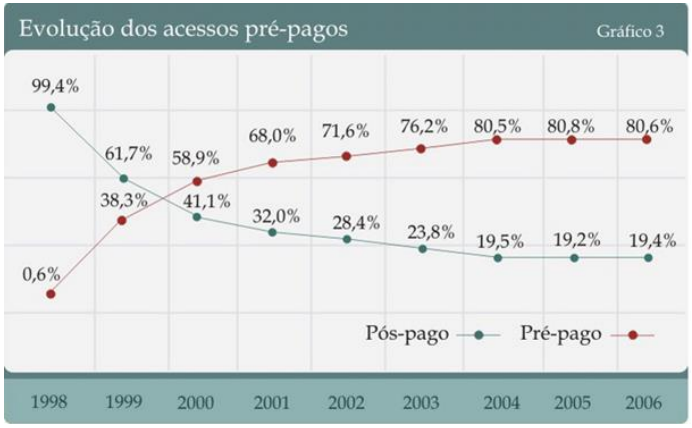
50 — X

$$X = \frac{360 \cdot 50}{100} = 180$$

Apresentação Gráfica dos Dados

c) Linha

Um gráfico de linha mostra tendências nos dados em intervalos iguais.

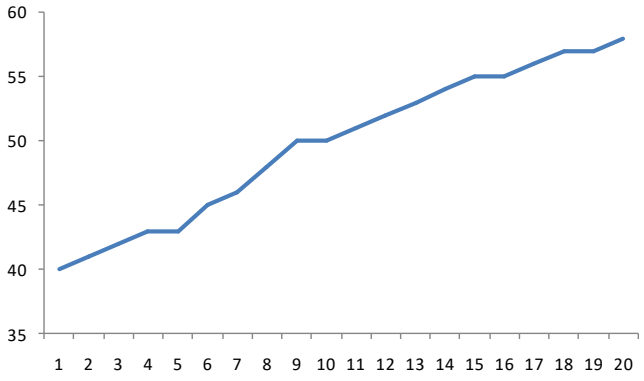


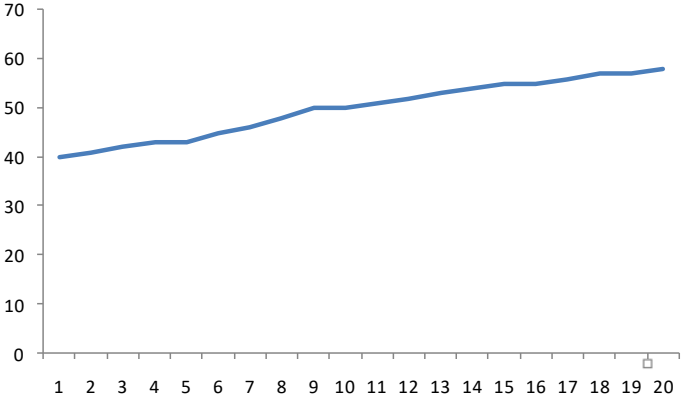
Fonte: Relatório Anual da Anatel, 2007.

O gráfico está adequado?

Apresentação Gráfica dos Dados

c) Linha





Qual gráfico está adequado?

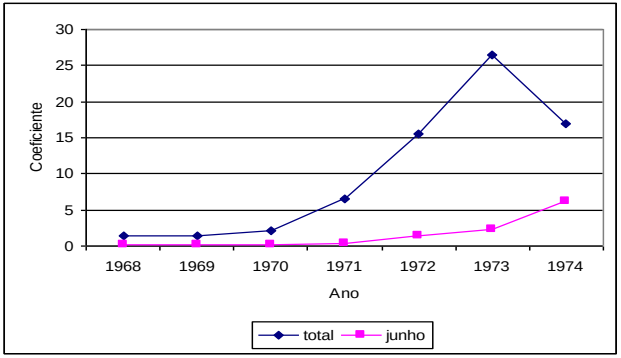
Apresentação Gráfica dos Dados

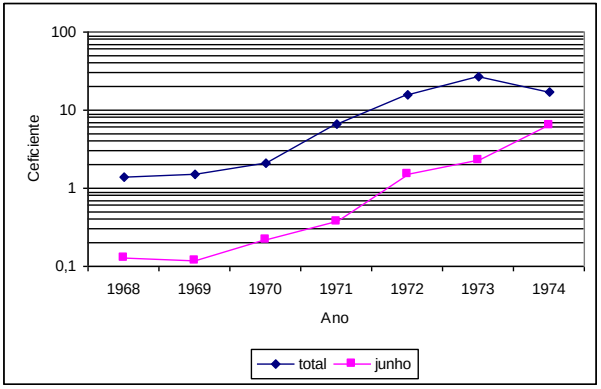
Tabela 2.4-Coeficientes de mortalidade (por 100.000 hab.) por meningite meningocócica no Município de São Paulo, no período de 1968 a 1974 observados durante todo o ano (total) e mês de junho de cada ano.

Ano	Total	Junho
1968	1,4	0,13
1969	1,5	0,12
1970	2,1	0,22
1971	6,6	0,37
1972	15,6	1,49
1973	26,5	2,24
1974	17,0	6,26

FONTE: *Rev. Saúde Públ., 10*: 1-16, 1976

Figura 1- Coeficientes de mortalidade (por 100.000 hab.) por meningite meningocócica no Município de São Paulo, no período de 1968 a 1974 observados durante todo o ano (total) e mês de junho de cada ano.





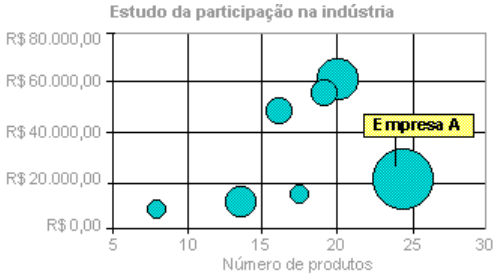
Apresentação Gráfica dos Dados

e) Bolhas

Um gráfico de bolhas é um tipo de gráfico xy (dispersão). O tamanho do marcador de dados indica o valor de uma terceira variável.

Exemplo:

Nº de produtos	Vendas	Partic. no mercado %
14	R$ 11.200,00	13
20	R$ 60.000,00	23
18	R$ 14.400,00	5
Valores X	Valores Y	Tamanho da bolha



O gráfico nesse exemplo mostra que a Empresa A tem a maioria dos produtos e a maior fatia do mercado, mas não necessariamente as melhores vendas.

Apresentação Gráfica dos Dados

Variáveis contínuas

a) Histograma

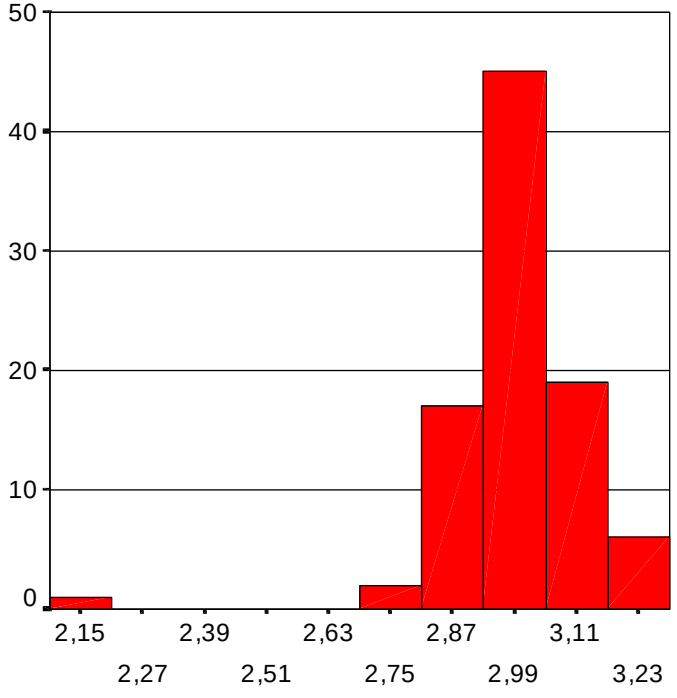
O histograma é formado por retângulos cujas áreas representam freqüências dos intervalos de suas classes. Esta apresentação é indicada para séries contínuas, e portanto não há espaço entre as barras.

Apresentação Gráfica dos Dados

Histograma

Exemplo: Preço médio (net price) do produto A (em reais)

Classes	Freqüência	Freqüência Relativa	Ponto Médio
2,09|---- 2,21	1	0,01	2,15
2,21|---- 2,33	0	0,00	2,27
2,33|---- 2,45	0	0,00	2,39
2,45|---- 2,57	0	0,00	2,51
2,57|---- 2,69	0	0,00	2,63
2,69|---- 2,81	2	0,02	2,75
2,81|---- 2,93	19	0,21	2,87
2,93|---- 3,05	45	0,50	2,99
3,05|---- 3,17	17	0,19	3,11
3,17|---- 3,29	6	0,07	3,23
Total	90	1,00	



Apresentação Gráfica dos Dados

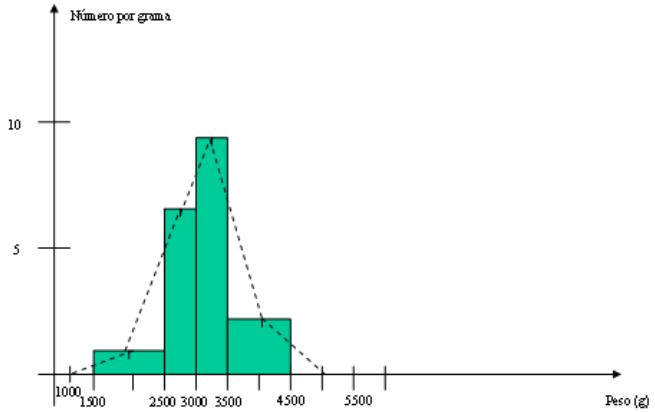
Tabela 1 – Nascidos vivos segundo peso ao nascer

Peso ao nascer (g)	Nº	Amplitude (g)	h´(frequência por gramas)
1.500 | — 2.500	1.200	1000	1,2
2.500 | — 3.000	3.600	500	7,2
3.000 | — 3.500	4.800	500	9,6
3.500 | — 4.500	2.400	1000	2,4
Total	12.000		

Tabela de frequência com amplitudes desiguais.

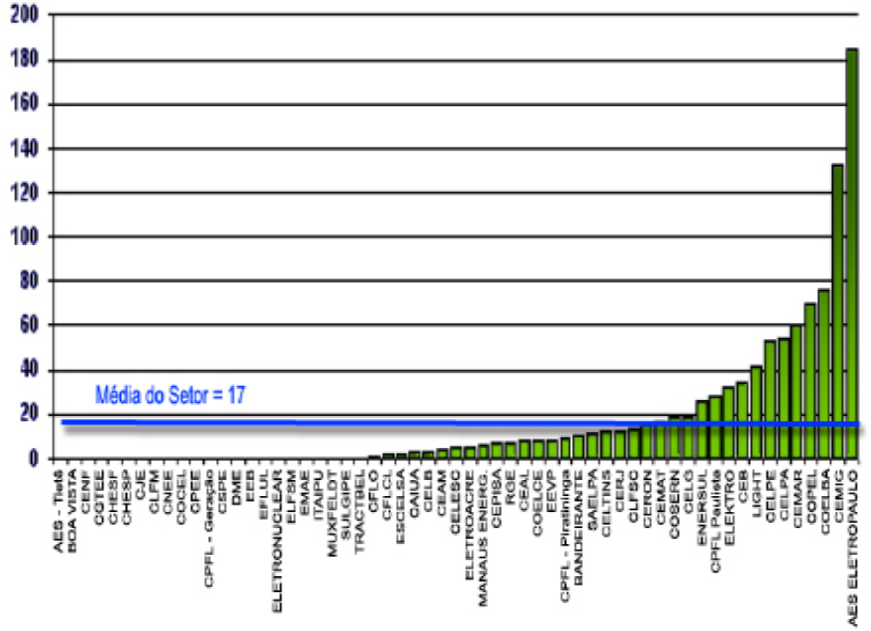
Frequência ajustada $\longrightarrow h' = \frac{N^{o}}{Amplitude}$

Figura 2.9 - Distribuição de nascidos vivos segundo peso ao nascer. Maternidade X, 1999.



FONTE: SAME, Maternidade X

Apresentação Gráfica dos Dados



O gráfico mostra que a AES Eletropaulo tem maior número de acidentes. Você concorda com esse resultado?

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 calculo 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

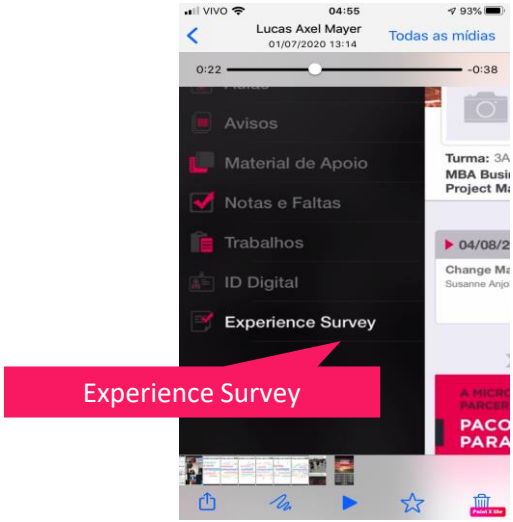
- Variáveis quantitativas contínuas

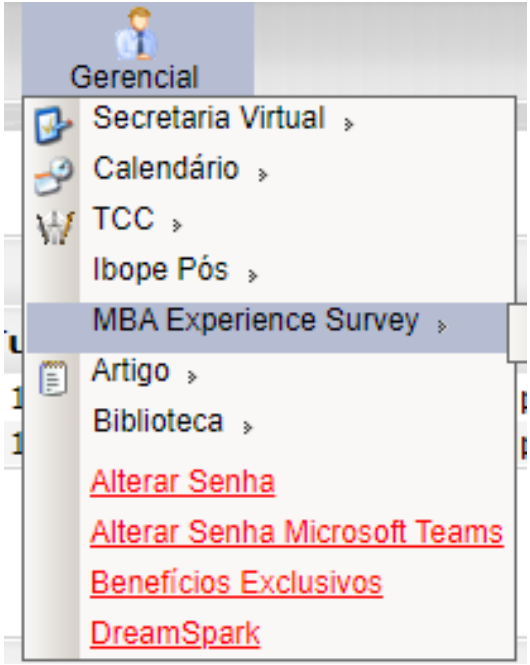
Histograma: É formado por retângulos cujas áreas representam frequências dos intervalos de suas classes. Esta apresentação é indicada para séries contínuas, e portanto não há espaço entre as barras.

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

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






NÃO VÊ QUE
ESTÁ A PISAR
A NOSSA RECOLHA
ESTATÍSTICA??

OBRIGADA

 / Regina T. I. Bernal


FIAP