



MBA EM DATA SCIENCE & AI

APPLIED STATISTICS





Leandro Augusto Ferreira

Cientista de Dados / Professor/ Pesquisador profleandro.ferreira@fiap.com.br

Formação Acadêmica

Bach. e Lic. em Matemática – ICMC/USP

Bach. em Estatística - FMU

Mestre em Sistemas Complexos – EACH/USP

Doutor em Estatística - IME/USP

Experiência Profissional

Estatístico – FM/USP

Cientista de Dados - ShareCare Brasil

Cientista de Dados Sr.- TransUnion

Espec. 2 de Modelagem de risco de Crédito – Banco Safra

Gerente de Data Science & Analytics – Sulamérica Seguros *

_ · · •



AULA 1 Conceitos Básicos de Estatística Estatística descritiva



Avaliação da disciplina

Avaliação	Peso
Listas de Exercícios	0.5
Projeto Integrado	0.5



Conceito de Estatística

Statistics ≠ **Statistic**

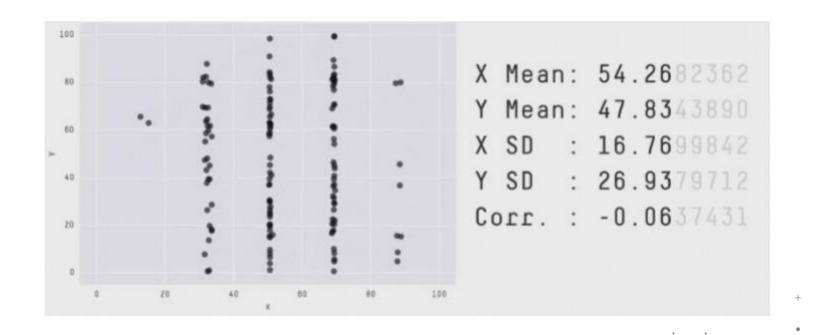


O que é um estatístico

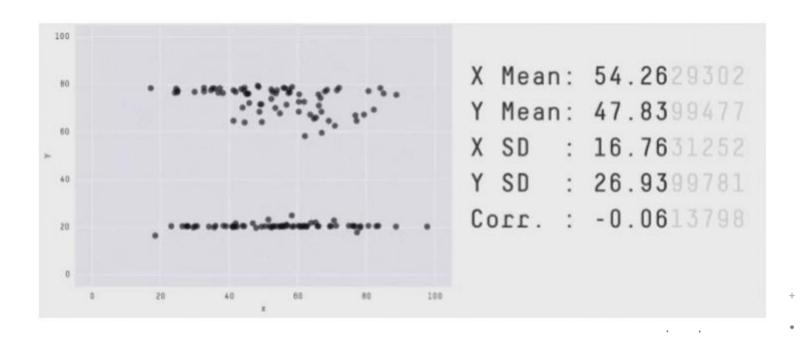
O estatístico é um "mágico" capaz de transformar cientificamente dados em informação. No entanto, ao contrário de um verdadeiro **guru**, ele atribui incertezas as suas tomadas de decisão.

Carlos A. Bragança Pereira

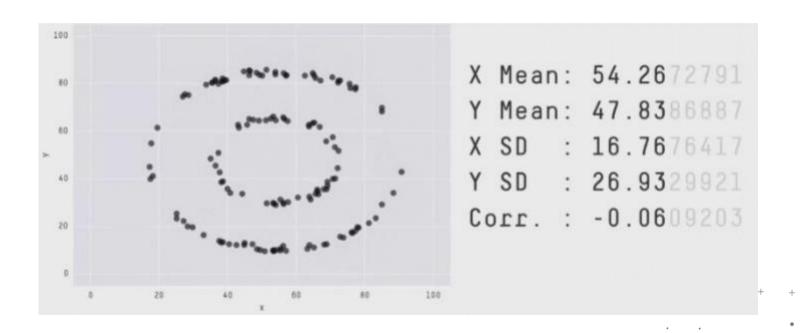




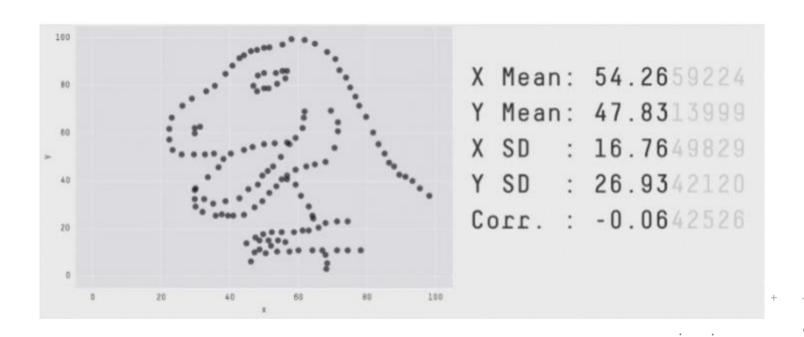












Dado

Registro do atributo de um

- ser vivo
- objeto
- fenômeno.



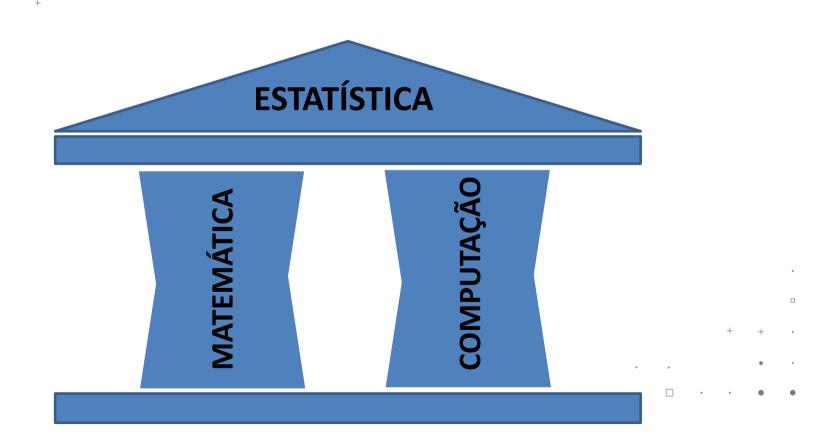
Informação

Resultado da

- manipulação
- processamento
- organização

de um dado

Bases da estatística

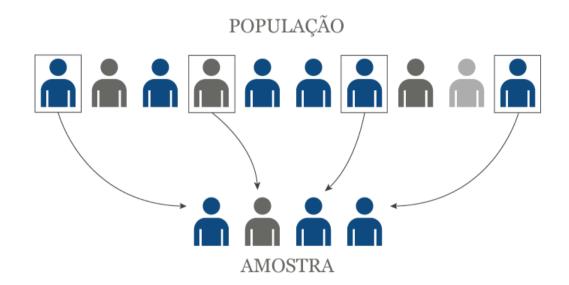


Quais dados uma pessoa poderia ter?

- Nome
- Idade
- Profissão
- Sexo
- Estado Civil
- Salário
- Tem alguma doença
- Usa óculos
- Quanto gasta com cartão de crédito
- Religião
- Onde mora
- Onde nasceu ...



E como é que a gente faz?



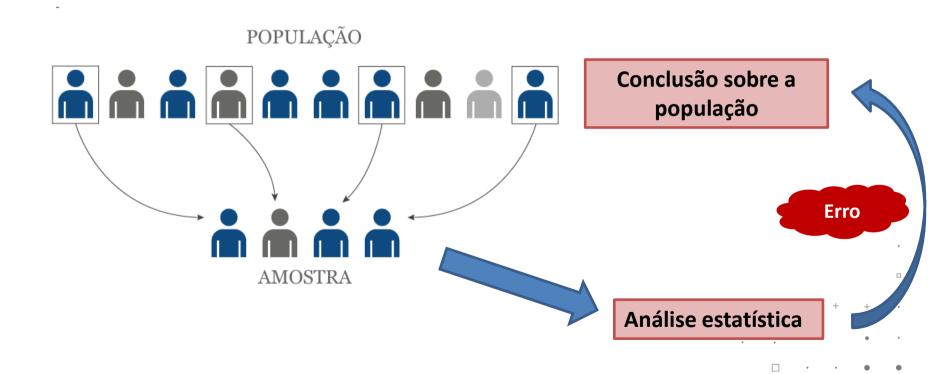


E como é que a gente faz?

POPULAÇÃO **AMOSTRA** Análise estatística

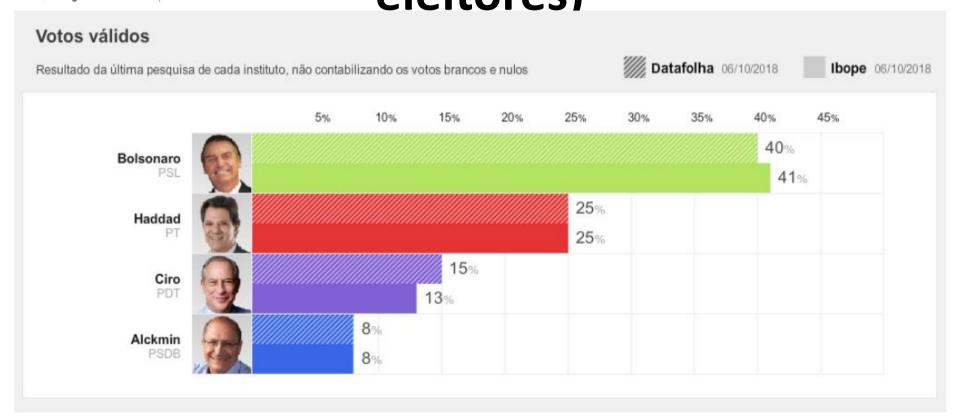


E como é que a gente faz?

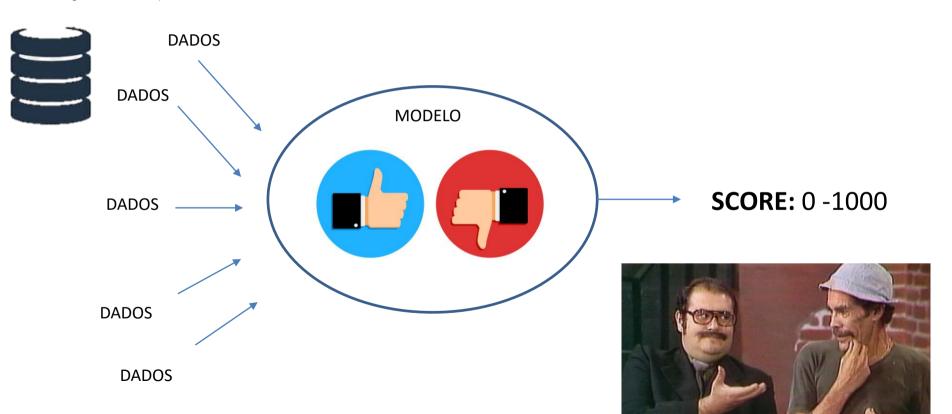


FIND WBA+

Pesquisa eleitoral (147,3 M de eleitores)

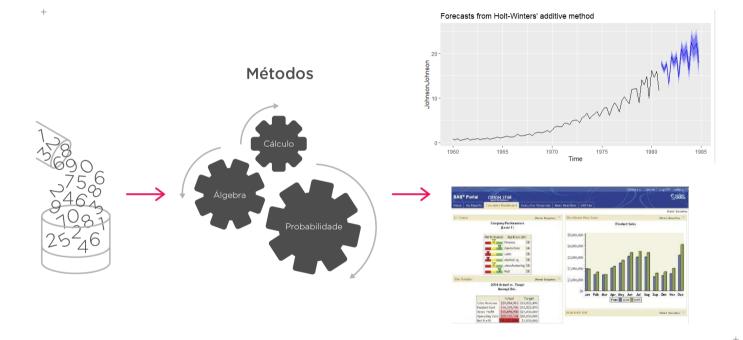


Modelagem estatística - Crédito





DATA **ANALYTICS**



. . .

UNIVERSO DE FORNECEDORES





























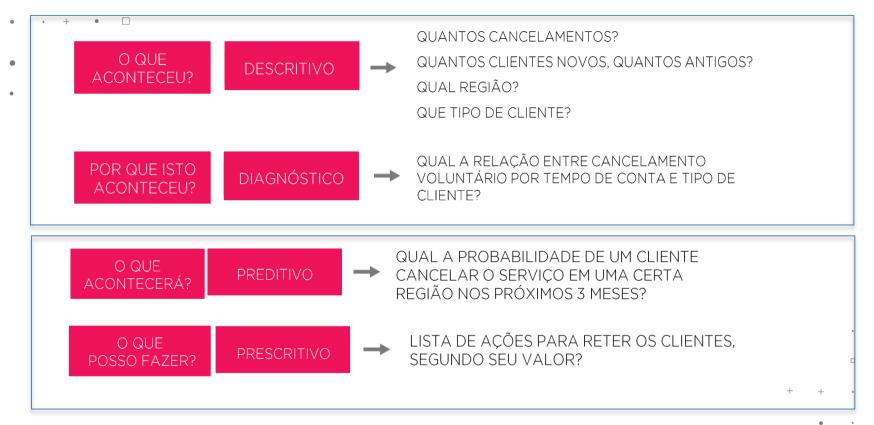






COMO TIRAR INFORMAÇÕES DELES





ESTATÍSTICA ESTATÍSTICA DESCRITIVA DISTRIBUIÇÕES **TABELAS GRÁFICOS MEDIDAS** ORGANIZAR **DESCREVER APRESENTAR ANÁLISE** INTERPRETAÇÃO **AMOSTRAGEM EXPERIMENTOS** probabilidade INFERÊNCIA ESTÁTISTIC 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.





Como medir a incerteza?





ESTATÍSTICA

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



BIOESTATÍSTICA

Estudos epidemiológicos Observacionais Experimentos ou quase randomização amostragem não sim exposição desfecho Transversais Caso-controle Ensaios Ensaios coorte comunidades clínicos

"Bioestatística é a Estatística aplicada às ciências médica e biológica."

Fonte: Vieira, Sonia. Introdução a 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 Domicilios Continua 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."

/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.

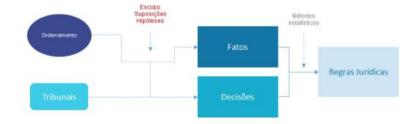
NAKASHIRO, M. M. Biometria no Brasil e o registro de identidade civil: novos rumos para identificação. 2011. 126 f. Tese (Doutorado em Sociologia) - Departamento de Pós-Graduação em Sociologia, Universidade do Estado de São Paulo, São Paulo, 2011.

• • + • □

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

.

•



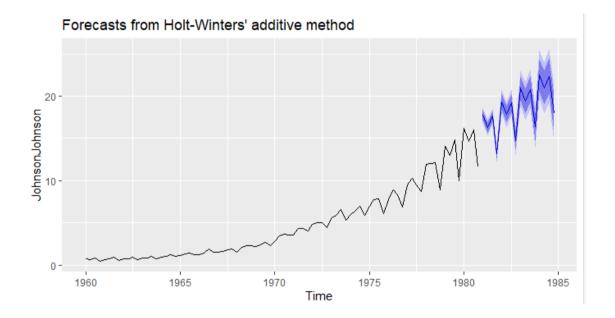
Estatística Descritiva



Inferência Estatística

. • +

.

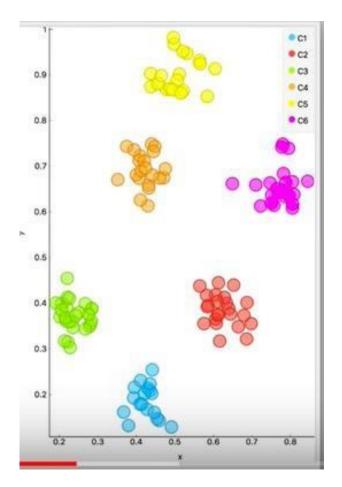






Exemplo de Cluster Não Hierárquico

+



.

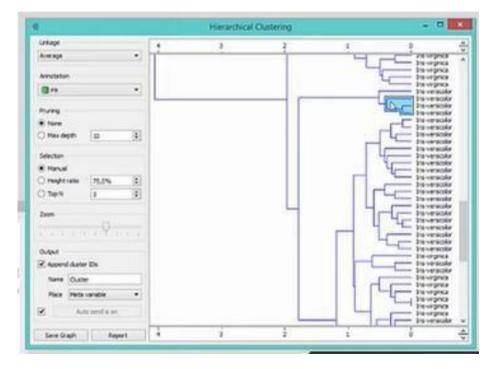
+

•

□ · · • •



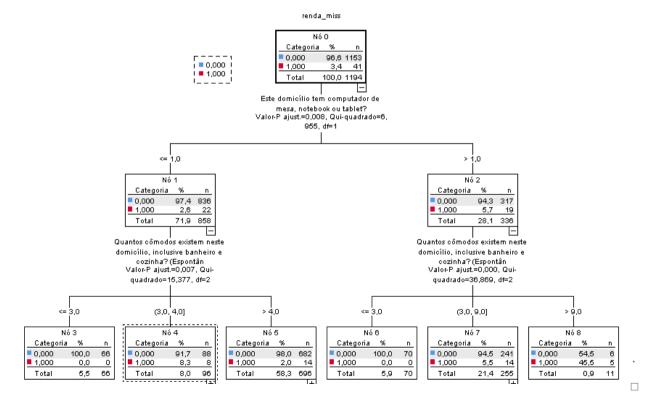
Exemplo de Cluster Hierárquico





Técnicas de Discriminação

Exemplo de Árvore de Decisão







Estatística Descritiva

Medidas de resumo



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 civ		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	О	2	О
2	1	1	0	2	O	2
3	2	2	0	4	2	2



Escala de Mensuração

Tipos de Variáveis Resultantes de contagem Resultantes de Classificação ou mensuração Qualitativas Quantitativas Ordinal Contínua Nominal Discreta Exemplo: Sexo Classe social Número de filhos + Idade



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:

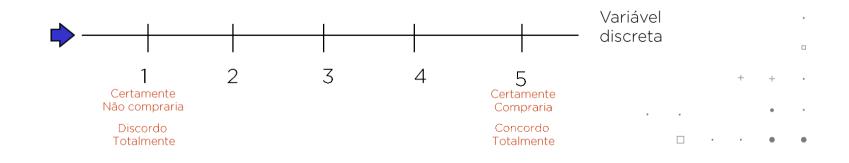
Variável	Ocorrência	Tipo (escala de mensuração)	
Estado civil	Solteiro		
	Casado	Qualitativa Nominal	
	Viúvo		
	Divorciado		
Faz atividade física	0=Não ; 1=Sim	Qualitativa Nominal	
Idade (anos)	[0-110]	Quantitativa contínua	
Anos de estudo	[0-99]	Quantitativa contínua	



Escala de Mensuração

Exemplo: Escala de questionário:



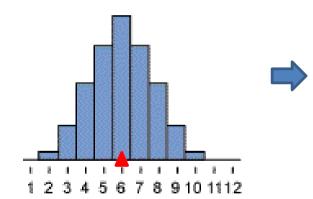




São estatísticas que resumem, em um único valor, a tendência central (média, mediana, moda), a variabilidade (variância, desvio padrão) e a forma da distribuição (simétrica ou assimétrica) da variável.



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:

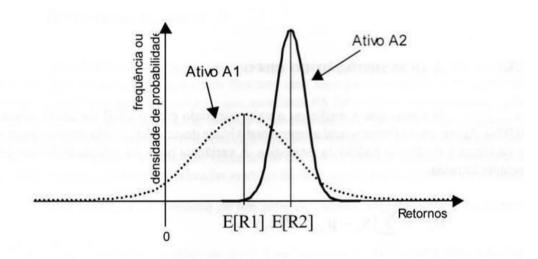
- Variância
- Desvio padrão

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.

Exemplo 1

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

Qual a média e a mediana encontrada?

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

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:
$$\bar{x} = \frac{6,03+5,59+6,04+6,00+5,99+6}{4}$$

$$mediana = \frac{6,00+6,02}{2} = 6,0$$

Exemplo 2

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

Qual a média e a mediana encontrada?



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 DE POSIÇÃO - MÉDIA

• Média Aritmética Simples:

$$\overline{\mathbf{x}} = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{\mathbf{n}}$$

• Média Aritmética Ponderada:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i \cdot F_i}{n}$$

• Média Geométrica (evolução):

$$Mg = \sqrt[n]{\cdot_1 x \cdot \cdot_2 \cdot \cdot_x} \quad n$$

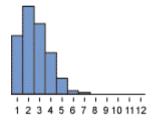
• Média Quadrática:

$$\overline{\mathbf{X}}^{2} = \frac{\sum_{i=1}^{n} \mathbf{x}_{i}^{2}}{\mathsf{n}}$$



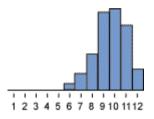
Decisão pela média ??????

Assimétrico à direita



Média > Mediana

Assimétrico à esquerda



Média < Mediana



Medidas Resumo

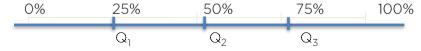
Outras Medidas de Posição

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). 0%

Quartis: dividem a série em quatro partes iguais.





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.



Medidas Resumo

Qual o desvio padrão?

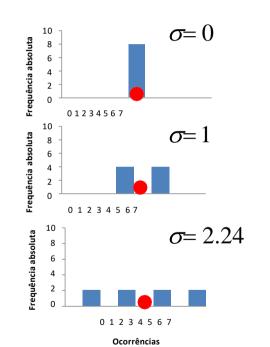
• Medidas de Dispersão

Exemplo 8:

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

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

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



Média



Medidas de Dispersão

Medidas de Dispersão: variância e desvio padrão Exemplo C

X	Média	(X-Média)	(X-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)

$$ightharpoonup$$
 variância da população $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - x)^n}{n}$

> variância da amostra

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

.

· · •



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: caracterizamse 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.



Coeficientes de Assimetria (Skewness)

```
As=0 → simétrica
```

As>0 → assimétrica positiva

As<0 → assimétrica negativa

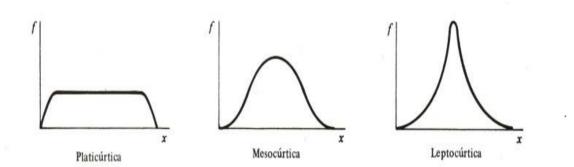
Índice de Assimetria (Pearson)

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

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

Medidas de Assimetria

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





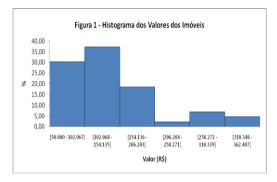
Exemplo de estatística descritiva



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			

. .



Outras Medidas de Dispersão

- Coeficiente de Variação
- > Amplitude
- Amplitude Inter-Quartílica



Outras Medidas de Dispersão

Coeficiente de variação (CV)

É o quociente entre o desvio padrão e a média.

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

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



Medidas Resumo

Qual o coeficiente de variação?

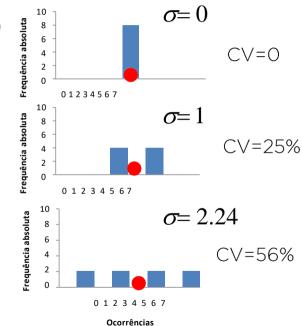
Medidas de Dispersão

Exemplo 8:

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

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

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







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.



Outras Medidas de Dispersão

Amplitude Inter-quartílica

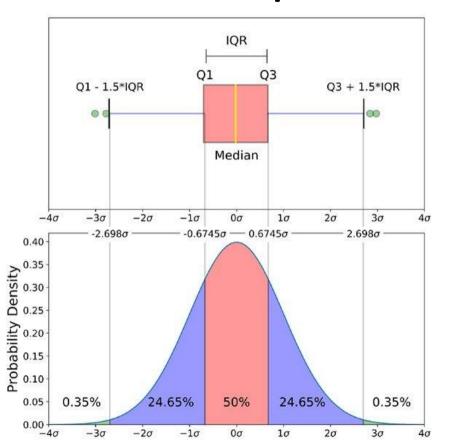
É a diferença entre o terceiro e o primeiro quartil (Q3-Q1).

Usada em análise exploratória de dados - gráficos Box Plot.



Box plot

•



Frequência Absoluta e Relativa

Considere ao variável grau de Instrução dos dados da tabela (Variável qualitativa)

Grau de instrução	Contagem	f_{i}	\int_{r_i}	f_{r_i} %
Fundamental		12	0,3333	33,3%
Médio		18	0,5000	50 %
Superior	MI	6	0,1667	16.7%
total		n=36	1,0000	100%

 f_i :Frequência absoluta da categoria i (número de indivíduos que pertencem à categoria i

$$f_{r_i} = \frac{f_i}{n}$$
 : Frequência relativa da categoria i

 $f_{r_i}\%=f_{r_i}st 100\%$: Frequência relativa percentual da categoria i



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	Cumulative
			Frequency	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		



Exemplo

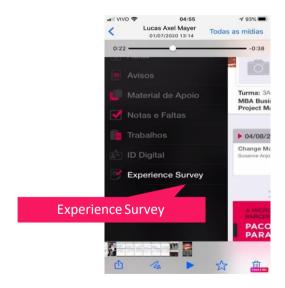
- 1) Leia a base salario.csv e selecione as variáveis salario e anosexperiencia.
- 2) Calcule todas as medidas de resumo essas variáveis.
- 3) Calcule a distribuição absoluta e relativa para as duas variáveis
- 4) Repita o cálculo das medidas de resumo cruzando com as variáveis posição e sexo (separadamente)
- 5) Repita o cálculo das medidas de resumo cruzando com as variáveis **posição** e **sexo (conjuntamente)**
 - Posição em do menor para o maior significa o cargo na Companhia
 - Salário em mil/ano
 - Sexo: 1 é Masculino 0 é Feminino

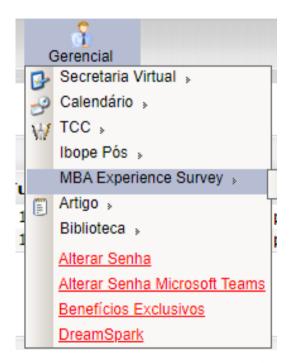


O que você achou da aula de hoje?

Pelo aplicativo da FIAP

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







OBRIGADO



profleandro.ferreira@fiap.com.br



Copyright © 2019 | Professor (a) Nome do Professor

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor.

