

# Aula 01 • 15.05

- Assistir video que de encontro
- Solicitar excel se nível

Diogenes

Matemático de formação

Econômica

Dirектор Global das Algaristas

- |             |                           |
|-------------|---------------------------|
| 1a semana - | estatística descritiva    |
| 2a semana - | regressão linear          |
| 3a semana - | teoria das probabilidades |
| 4a semana - | exercícios                |

## Cluster computing

Hadoop - código aberto, rápida aplicação

↳ Estrutura "map reduce" que  
o google utilizava

↳ Possui um sucessor chamado Spark

↳ Técnica p/  
resolver maior volume  
de dados

Data Mining ≈ Big Data ≈

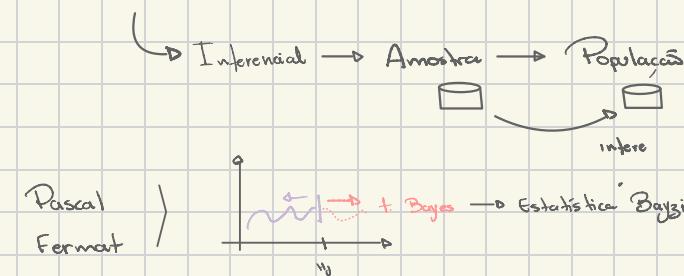
Predictive Analytics ≈

Data Science

Parecido,  
mas diferente

Estatística - métricas em cima de números

Clássica - não tem olhar p/ ciência de dados



Estatística Descritiva: Descrever um conjunto de dados

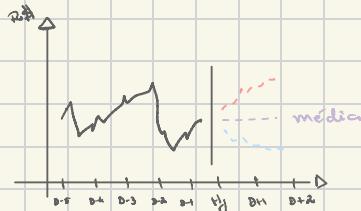
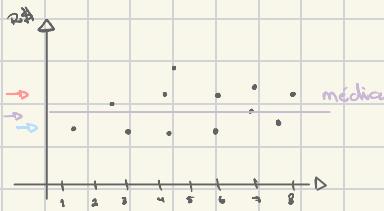
### 01. Média Aritmética

$\bar{x}$  → Literatura

$$\bar{x} = (x_1, x_2, \dots, x_n)$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Bayes: Média mais provável que vai acontecer



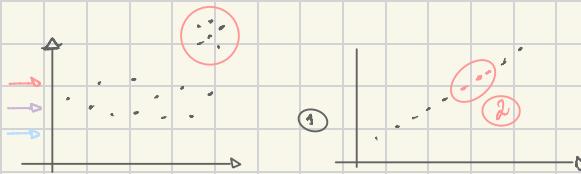
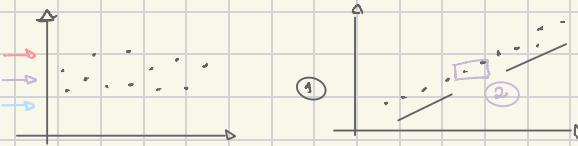
## Ord. Medianas

$$x = (1, 5, 7, 0)$$

① Ord. 0, 1, 5, 7

② ímpar  $\rightarrow$  elemento central

par  $\rightarrow \bar{x}$  dos 2 elementos centrais



Variância média da distância dos pontos médios

Desvio Padrão  $\sqrt{\text{var}(x)}$   $\rightarrow \sigma$

Score Z

$$\frac{\bar{x} - x_i}{\sigma}$$

$Z > 3$  é outlier

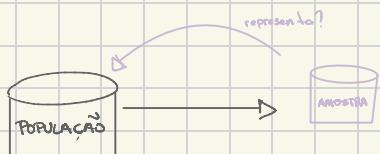
# Aula 02 • 17.05

## Bibliografia

- BAILEY A., BRUCK, P. ESTATÍSTICA PARA DATA SCIENCE, 1ª edição.
- LIMA, Sérgio, Arnaud, Bernardo. ESTATÍSTICA TEÓRICA E APLICAÇÕES com uso de R para Ciências Sociais e Humanas, 1ª edição.
- FARAWI, I. F. L. et al. Análise de Dados - Modelagem Matemática para Tomada de Decisões em Big Data. São Paulo: EDUSP, 2018.
- KERSEBAUM, K. Data Mining e Big Data. São Paulo: EDUSP, 2018.
- DALE, Peter, HANNAH, SICK VERNON. CONCEPTS AND TECHNIQUES, 2ª edição. Morgan Kaufmann.

## Bibliografia

- COHEN, J. A. Little Book of  $\beta$  for Multivariate Analysis. New Jersey: John Wiley & Sons, 2013.
- HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. The Elements of Statistical Learning. 2nd ed. (2009). Springer. <http://statweb.stanford.edu/~tibs/ElemStatLearn>, 2009.
- JAGGARD, R. Data Mining: Practical Machine Learning Tools and Techniques with Applications in R. 3rd Edition. Chapman and Hall/CRC, 2016.
- WILCOX, R. Introduction to Robust Statistics. 2014. <http://www.rufuswilcox.com/robcourse.html>



$$\bar{p} = 3,9K$$

$$\text{mediana } (\bar{p}) = 2,4K$$

$$\delta(p) = 3,989$$

amostra  $\alpha_1$

$$\bar{p} = 2,7K$$

$$\text{mediana } (\bar{p}) = 2,1K$$

$$\delta(p) = 0,79K$$

amostra  $\alpha_2$

$$\bar{p} = 3,3K$$

$$\text{mediana } (\bar{p}) = 3,6K$$

$$\delta(p) = 0,9K$$

não  
representa

Gerar amostra aleatória

$$\bar{p} = 3,9K$$

$$\text{mediana } (\bar{p}) = 4,1K$$

$$\delta(p) = 3,9K$$

\* amostra

temporal (ordem importancia) → janelas temporais

atemporal → aleatória

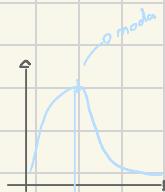
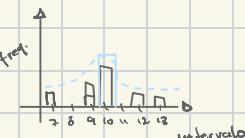
teorema do limite central

calculadora limite amostra

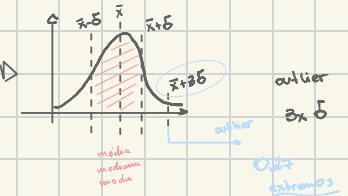
↳ term non return

## Gráficos

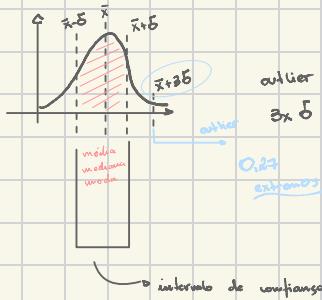
### Histograma



=>



## Estatística



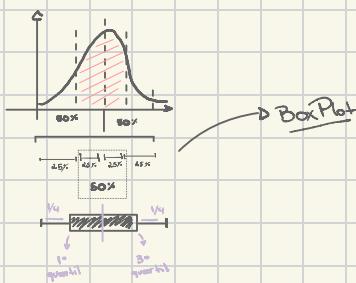
Experimento : 100x

95% da "certo" = 95% confiança

5% p-valor = 1 - nível confiança estatístico

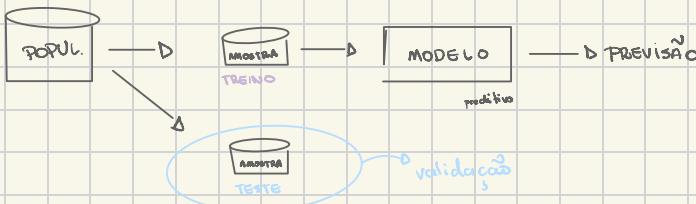
$$NC = 95\%$$

$$p\text{-valor} = 5\% = 0,05$$



Siers → abordagem determinística  
→ não exatamente preditivo

### Criação modelo preditivo



PIBIU - Itau publica uma forma de prever o PIB mensalmente

↳ índice calculado

↳ normalmente começa pelo 100

análise

Modelo preditivo: Previsão do PIB ( $y$ )

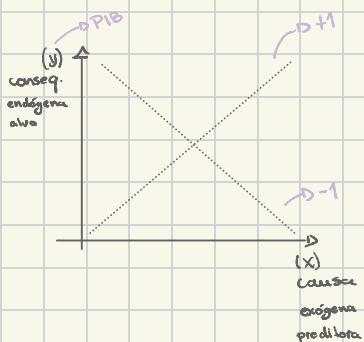
### Relação causalidade

Métrica de correlação (est. descritiva)

consequência depende da causa  
variável dependente  
varável independente  $\Rightarrow y \sim x$

Coeficiente Correlação Linear (Pearson)

↳  $-1 \leq r \leq 1$



↳  $r > 0$  - relação diretamente proporcional

↳  $r < 0$  - inversamente proporcional

↳  $|r| \rightarrow 1$  : correlação forte

↳  $|r| \rightarrow 0$  : correlação fraca

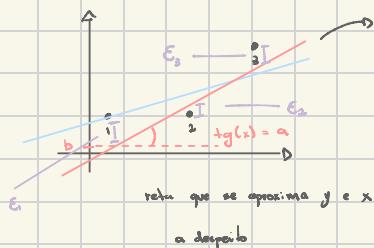
↳ ex: 

1	2	3	4
2	4	6	8
3	6	9	12
4	8	12	16

\* correlação Nô spúria = correlação não indica causalidade

## Régressão linear

Sir F. Galton → estudo sobre alturas



$\rightarrow$  aqui estamos procurando o  
 $a$  e  $b$ ,  $\circ x$  e  $y$  já sabemos

$L_b$  Coeficiente linear ou intercepto

$a$  Coeficiente angular

$$\begin{aligned} y_1 &= (ax_1 + b) + \epsilon_1 \\ y_2 &= (ax_2 + b) + \epsilon_2 \\ &\vdots \\ y_n &= (ax_n + b) + \epsilon_n \end{aligned} \quad \left\{ \begin{array}{l} y_i = ax_i + b + \epsilon_i \\ \epsilon_i = y_i - ax_i - b \end{array} \right.$$

$$\sum_{i=1}^n \epsilon_i = \sum_{i=1}^n (y_i - ax_i - b)^2$$

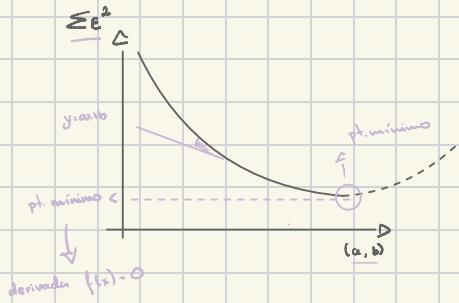
erro

reta

$$\frac{\partial}{\partial b} \frac{(\sum \epsilon_i^2)}{\partial b} = 0$$

algoritmo

gradiente descendente



tudo p/ calcular  $a$  e  $b$

$$y = ax + b$$

$$PBL = a * BPL + b$$

$$PBL = 0,51 * BPL + 66,45$$

↓

Modelo regressão linear

Parâmetros P. L.

①  $R^2 \rightarrow 1$ , melhor

$\rightarrow 0$ , pior

② Erro padrão (aprox.)

menor  $\rightarrow$  melhor

③ P-Valor  $< 0,05 \rightarrow$  menor - problema na validade estatística

$\hookrightarrow N > 0,95$

## Regressão Múltipla

modelos multivariados  $\rightarrow$  Regressões lineares Simples: 1v. x

$\rightarrow$  Regressões lineares Múltiplas: + 2v. x

$$y = \alpha_1 x_1 + \alpha_2 x_2 + b$$

