

# Estatística e Data Science

July 14, 2021

## 1 Trilha de Estatística para Data Science

- Autor: Ronisson Lucas Calmon da Conceição
- [Github](#)
- [Linkedin](#)

**O que é Data Science?** Para além da programação e das demais ferramentas, Ciência de Dados é um campo interdisciplinar e desta forma envolve a aplicação de outras disciplinas como Matemática e Estatística, com ênfase na análise de dados para possíveis tomadas de decisão. Assim, o objetivo desta trilha é demonstrar os principais conceitos estatísticos a serem utilizados em Ciência de Dados. Não pretende ser um resumo exaustivo ou definitivo, visto a complexidade destas áreas.

Fonte: <https://abracd.org/o-que-e-data-science-a-famosa-ciencia-de-dados/>

### Pré-requisitos matemáticos

- a. Nivelamento matemático
  - Teoria dos Conjuntos
  - Funções
- b. Cálculo Integral e Diferencial
  - Limites
  - Derivadas
  - Derivadas parciais
  - Vetor gradiente
  - Otimização
  - Integral

Onde estudar?

Cálculo I

- [Curso UNICAMP](#)
- [Curso USP](#)
- [Professor Paulo Pereira](#)
- [Professor Ferreto](#)
- [Univesp](#)
- [Professor Aquino](#)

Cálculo II

- [Aulas USP](#)
- [Univesp](#)

Cálculo III

- [Aulas USP](#)
- [Univesp](#)

Cálculo IV:

- [Aulas USP](#)

Por tópicos:

Limites

- [Me Salva!](#)
- [Grings](#)

Derivadas

- [Paulo Pereira](#)
- [Me Salva!](#)
- [Grings](#)

Derivadas parciais

- [A matemática das redes neurais](#)

Matemática para Machine Learning

- [Fundamentos de matemática \(Didática Tech\)](#)
- [Matemática para Machine Learning \(Didática Tech\)](#)

c. Álgebra Linear

- Matrizes
  - Tipos de matrizes
  - Operações matriciais
  - Matriz inversa, transposta e adjunta
- Determinantes
- Sistemas Lineares
- Espaços vetoriais e subespaços
- Base e dimensão
- Produto interno, ortogonalidade e projeções
- Transformação Linear
- Autovalor e autovetor
- Diagonalização

Onde estudar?

Playlists completas no YouTube:

- [Professor Alê](#)
- [Escola Politécnica da USP](#)
- [Canal Matemática Universitária](#)
- [Resumão de Álgebra Linear](#)

- [Linear Algebra - Full College Course](#)
- [Univesp](#)

Revisão de Estatística e Álgebra linear:

- Apêndices do Gujarati e Porter (2011)[Econometria Básica]
- Apêndices do Wooldridge (2006)[Introdução à Econometria]

## 1.1 Trilha de Estatística

### 1. Estatística descritiva univariada 1.1 Medidas de posição central

- Média
- Moda
- Mediana

### 1.2 Medidas de dispersão

- Amplitude
- Desvio-médio
- Variância
- Desvio-padrão
- Erro-padrão
- Coeficiente de variação

### 1.3 Medidas de assimetria e curtose

## 2. Estatística Descritiva bivariada 2.1 Medidas de associação entre variáveis qualitativas

### 2.2 Medidas de associação entre variáveis quantitativas

- Diagrama de dispersão
- Covariância
- Correlação

## 2. Probabilidade

- Experimento Aleatório
- Espaço amostral
- Eventos
- Operações com eventos
- Eventos mutuamente excludentes
- Eventos independentes
- Probabilidade condicional
- Teorema de Bayes

## 3. Variáveis Aleatórias e Distribuições de probabilidade

- Esperança Matemática (valor esperado ou expectância)
  - Expectância de uma variável aleatória (v.a.) discreta
  - Expectância de uma v.a. contínua
  - Propriedades do valor esperado
  - Esperança Condicional

- Lei das Expectativas Iteradas
- Variância
  - Propriedades da variância
- Covariância
  - Propriedades da covariância
- Coeficiente de correlação
  - Propriedades da correlação
- Desigualdade de Tchebycheff
- Momentos de uma distribuição
  - Função Geratriz de Momentos
  - Propriedades da Função Geratriz de Momentos
- A seguir as principais distribuições de probabilidade contínuas e discretas.

### 3.1 Distribuições discretas

- Distribuição de Bernoulli
- Distribuição binomial
- Distribuição binomial negativa
- Distribuição geométrica
- Distribuição hipergeométrica
- Distribuição de Poisson
- Distribuição de Pascal
- Distribuição Multinomial

### 3.2 Distribuições contínuas

- Distribuição uniforme
- Distribuição normal (gaussian distribution)
- Distribuição Log-normal
- Distribuição exponencial
- Distribuição Gama
- Distribuição qui-quadrado
- Distribuição  $t$  de Student
- Distribuição  $F$  de Snedecor

### 3.3 Distribuições de probabilidade conjunta

- Distribuição conjunta de variáveis discretas
- Distribuição conjunta de variáveis contínuas

## 4. Inferência Estatística e Principais Teoremas de Probabilidade 4.1 Principais teoremas de probabilidade

- Teorema de Tchebycheff
- Lei dos grandes números
- Teorema Central do Limite

### 4.2 Estimadores

- Estimadores não viesados
- Estimadores eficientes

- Estimadores consistentes
- Melhor estimador linear não viesado
- Propriedades ideais dos estimadores em pequenas e grandes amostras
- Estimação por máxima verossimilhança

#### 4.3 Intervalo de confiança

- Intervalo de confiança para  $\mu$  quando o valor de  $\sigma$  é conhecido
- Intervalo de confiança para  $\mu$  quando o valor de  $\sigma$  é desconhecido
- Intervalo de confiança para uma proporção

#### 4.4 Testes de hipóteses

- Tipos de erro
- Testes paramétricos
- Testes não paramétricos

### 5. Regressão Linear 5.1 Regressão Linear Simples

- Hipóteses e propriedades
- Derivação das estimativas de Mínimo Quadrados Ordinários
- Formas funcionais
- Valores esperados e variâncias dos estimadores de MQO
- Regressão através da origem
- Descrição dos resultados

#### 5.2 Regressão Linear Múltipla

- Hipóteses e propriedades
- Variáveis dummy
- Seleção de modelos
  - $R^2$  ajustado
  - Critérios de informação
  - Teste  $F$
- Modelo de regressão múltipla em notação matricial
- Derivação das estimativas de MQO com notação matricial
- Interpretação das estimativas
- Teorema de Gauss-Markov
- Distribuição amostral dos estimadores de MQO
- Teste  $t$
- Intervalos de confiança
- MQO Assintótico
- Modelo de equações simultâneas
- Modelos não lineares

#### 5.3 Violando Hipóteses Básicas

- Multicolinearidade
- Autocorrelação
- Heterocedasticidade
- Simultaneidade

## 6. Séries Temporais

- Processos estocásticos
- Autocovariância e autocorrelação
- Estacionaridade
- Ergodicidade
- Ruído branco
- Passeio aleatório
- Passeio aleatório com drift
- Médias móveis (MA)
- Processos autorregressivos (AR)
- Processo autorregressivo de médias móveis (ARMA)
- Processo autorregressivo integrado de médias móveis (ARIMA)
- Testes de estacionariedade
- Metodologia de Box-Jenkins
- Função de autocorrelação
- Função de autocorrelação parcial
- Diagnóstico de resíduos
- Processos não estacionários
- Modelos multivariados
- Cointegração

## 7. Econometria (Opcional)

- Variáveis instrumentais
- Modelo de regressão generalizado (GLS)
- Modelo de regressão generalizado factível (FGLS): estimação em dois estágios
- Método dos momentos generalizados (GMM)
- Estimação não-paramétrica e semiparamétrica
- Modelos não lineares e transformação de Box-Cox e Mínimos Quadrados Não-Lineares
- Máxima verossimilhança
- Modelos de escolha discreta
- Modelos de equações simultâneas
- Dados em Painel
- Modelos com variáveis dependentes defasada
- Regressão quantílica
- Regressão descontínua
- Econometria espacial
- Econometria financeira

## Principais Bibliotecas do Python para Data Science

- Pandas
- NumPy
- Statsmodels
- Scikit-learn
- Keras
- PyTorch
- TensorFlow

- XGBoost
- SciPy
- SymPy
- PyCaret
- Scrappy
- BeautifulSoup
- Linear Models

#### Visualização de dados

- Matplotlib
- Seaborn
- Bokeh
- Plotly

Para mais informações:

- [Top 10 Python Libraries For Data Science for 2021](#)
- [Best Python Libraries For Data Science In 2021](#) )
- [Top 13 Python Libraries Every Data science Aspirant Must know! \(and their Resources\)](#)

#### Sugestões de Livros Matemática

- BOLDRINI, J. et al. Álgebra Linear. São Paulo: Harbra, 1986.
- ÁVILA, G. Cálculo. Vols. I, II e III. Rio de Janeiro: Livros Técnicos e Científicos S.A., 1987.
- LIMA, E. L. Álgebra Linear. Coleção Matemática Universitária. Rio de Janeiro: IMPA, 1996.
- GUIDORIZZI, H.L. Um Curso de Cálculo. Vols. 1 a 4. 2ª ed. Rio de Janeiro: Forense-Universitária.

#### Estatística

- MEYER, P. L. Probabilidade – Aplicações à Estatística. São Paulo: Livros Técnicos e Científicos Editora, 1983.
- TOLEDO, G.L e OVALLE, I.I. Estatística Básica. São Paulo: Atlas, 1995.
- Kroese, Dirk P., Chan, Joshua C.C. Statistical Modeling and Computation. New York: Springer, Verlag, 2014
- CASELLA, G.; BERGER, R. L. Statistical Inference, Duxbury Press, 2001.
- HOGG, McKEAN J. W.; CRAIG, J. W. Introduction to mathematical statistics. 8th ed. 2019.
- HOFFMANN, R. Estatística para Economistas. São Paulo: Pioneira Thomson Learning, 2006.
- FÁVERO, L. P.; BELFIORE, P. Manual de análise de dados. Rio de Janeiro: Elsevier, 2017.
- MORETTIN, P. A.; TOLOI, C. M. C. Análise de séries temporais. São Paulo: Egard Blucher, 2006.

#### Econometria

- SARTORIS, A. Estatística e introdução à econometria. São Paulo: Saraiva, 2003.
- HILL, C; GRIFFITHS, W & JUDGE, G. Econometria. São Paulo: Saraiva, 2000.
- MADDALA, G. Introduction to Econometrics. New York: MacMillan.
- PINDYCK, R. e Rubinfeld, D. Econometric Models and Economic Forecasts. New York: McGraw-Hill, 1997.

- WOOLDRIDGE, J.M. Introdução à Econometria: Uma Abordagem Moderna, Pioneira, 2006, Thomson Learning.
- GREENE, W. H. Econometric analysis. Upper Saddle River, N.J: Prentice Hall. 2000.
- HANSEN, B. E. Econometrics. University of Wisconsin. 2020.
- HAYASHI, F. Econometrics. Princeton: Princeton University Press, 2000.
- SHEPARD, K. Introduction to Python for Econometrics, Statistics and Data Analysis. 4th. 2020.
- WOOLDRIDGE, J. Econometric Analysis of Cross Section and Panel Data, MIT Press Books, The MIT Press, 2000.
- GUJARATI, D.M. Econometria Básica, Campus/Elsevier, 2006.
- STOCK, J.H.; WATSON, M. Econometria, Addison-Wesley, 2004.
- DAVIDSON, R.; MACKINNON, J. G. Econometric: Theory and Methods. New York: University Press, 2004.
- BUENO, R. L. S. Econometria de séries temporais. São Paulo: Cengage Learning, 2011.

### Sugestões de Livros na Amazon

- WOOLDRIDGE, J. Introdução à econometria: Uma abordagem moderna
- BUENO, R. L. S. Econometria De Séries Temporais
- STOCK, J. H.; WATSON, M. W. Econometria
- MAIA, A.G. Econometria - Conceitos e Aplicações: Conceitos e Aplicações
- HILL, R. C.; JUDGE, G. G. Econometria
- SARTORIS, A. Estatística e introdução a econometria
- SAITO, M. Econometria Prática con Python
- PEREDA, P.C.; ALVES, D. Econometria Aplicada
- MORETTIN, P. A. Econometria Financeira: um Curso em Séries Temporais Financeiras
- ROSSI, J. W.; NEVES, C. Econometria e Séries Temporais com Aplicações à Dados da Economia Brasileira
- FERREIRA, P. G. C. (org.) Análise de Séries Temporais em R: Curso Introductório
- ALMEIDA, E. Econometria espacial aplicada
- GOLGHER, A. B. Introdução à Econometria Espacial
- WHEELAN, C. J. Estatística: O que é, para que serve, como funciona
- MLODINOW, L. O andar do bêbado: Como o acaso determina nossas vidas
- BUSSAB, W. O.; MORETTIN, P. A. Estatística básica
- BRUCE, P. Estatística prática para cientistas de dados: 50 conceitos essenciais
- HUFF, D. Como Mentir com Estatística
- TRIOLA, M. F. Introdução à Estatística
- RUMSEY, D. J. Estatística para leigos
- LARSON, R. Estatística Aplicada
- CASELLA, G.; BERGER, R. L. Inferência estatística
- MAGALHÃES, M. N.; LIMA, A. C. P. Noções de Probabilidade e Estatística
- FÁVERO, L. P.; BELFIORE, P. Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata
- MCKINNEY, W. Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy e IPython
- GUJARATI, D. N.; PORTER, D. C. Econometria Básica
- SCHEIDEGGER, J. Ah, se eu soubesse (estatística)...: A estatística desmistificada



- OLIVEIRA, F. E. M. Estatística e Probabilidade: com ênfase em exercícios resolvidos e propostos
- CAETANO, M. A. L. Python e Mercado Financeiro: Programação Para Estudantes, Investidores e Analistas
- PINTO, M. ESTATÍSTICA BÁSICA: para quem não é especialista
- PAULINO, C. D. et al. Estatística Bayesiana
- HAIR, J. F. et al. Análise Multivariada de Dados
- TOLEDO, G. L.; OVALLE, I. I. Estatística Básica
- GRUS, J. Data science do zero: primeiras regras com o Python
- GRUS, J. Data Science Do Zero: Noções Fundamentais com Python
- PROVOST, F.; FAWCETT, T. Data science para negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados
- KLOSTERMAN, S. Projetos de ciência de dados com Python: Abordagem de estudo de caso para a criação de projetos de ciência de dados bem-sucedidos usando Python, pandas e scikit-learn
- RAMALHO, L. Python Fluente: Programação Clara, Concisa e Eficaz
- MUELLER, J. P.; MASSARON, L. Python para data science para leigos: os primeiros passos para o sucesso
- GÉRON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems
- GÉRON, A. Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow
- MENEZES, N. N. C. Introdução à Programação com Python: Algoritmos e Lógica de Programação Para Iniciantes
- DOWNEY, A. B. Pense em Python: Pense Como um Cientista da Computação
- CHEN, D. Y. Análise de Dados com Python e Pandas
- MATTHES, E. Curso Intensivo de Python: Uma Introdução Prática e Baseada em Projetos à Programação
- NETTO, A.; MACIEL, F. Python Para Data Science: e Machine Learning Descomplicado
- RODRIGUES, P. L. V. Python Aplicado: Bolsa de Valores - Um guia para construção de análises e indicadores
- KNAFLIC, C. N. Storytelling com dados: Um guia sobre visualização de dados para profissionais de negócios

### Onde encontrar material gratuito?

- Playlist completa de Estatística básica da UFPR
  - Curso completo de estatística básica desde o básico.
- Estatística e Probabilidade (Khan Academy)
- Séries Temporais
  - Material didático
- Econometria (Estudar com Você)
- Econometria (Professor Alexandre Maia)
- Estatística e Econometria
- Estatística Básica (Professor Guru)
  - Canal conta com aulas de vários tópicos de estatística básica.
- Professor Grings
- Estatística (Univesp)
- Probabilidade e Estatística (Me Salva!)

### Onde encontrar dados públicos?

- [Kaggle](#)
- [Google Dataset Search](#)
- [UCI Machine Learning Repository](#)
- [Kdnuggets](#)
- [Dados.gov.br](#)
- [Portal da Transparência](#)

Dados financeiros, econômicos e sociais (bases genéricas):

- [Quandl](#)
- [db.nomics](#)
- [Banco Central do Brasil](#)
- [IPEADATA](#)
- [SIDRA/IBGE](#)
- [IBGE Cidades e Estados](#)
- [Atlas do Desenvolvimento Humano no Brasil](#)
- [OECD.Stat](#)
- [Portal Dados Abertos CVM](#)
- [Portal de Indicadores IBGE](#)
- [Trading economics](#)
- [Econodados](#)

Dados de comércio internacional:

- [Comex stat \(BR\)](#)
- [UN Comtrade Database \(World\)](#)
- [WTO Data \(World\)](#)
- [FAOSTAT \(World\)](#)

Dados de saúde:

- [DATASUS](#)
- [Saúde - IBGE](#)

Dados de educação:

- [Microdados INEP](#)
- [Dados abertos INEP](#)