

Universidade Federal do Rio Grande do Norte
Instituto Metr pole Digital
IMD0033 - Probabilidade

Apresenta  o da disciplina

Prof. Dr. Tetsu Sakamoto
Instituto Metr pole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

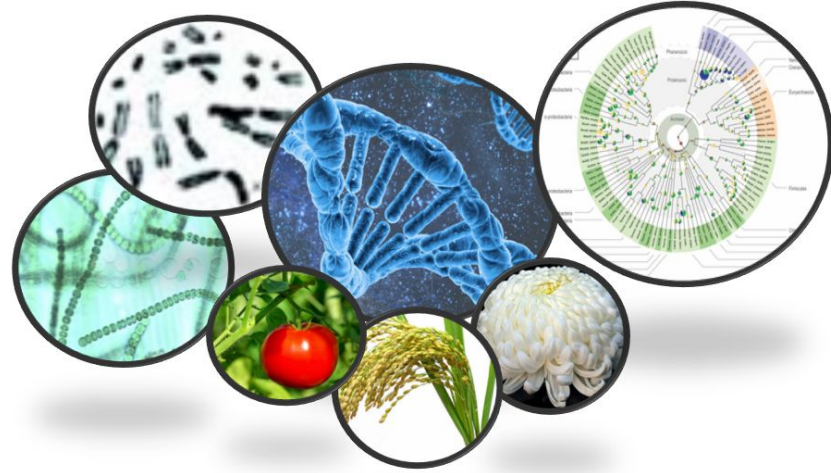


Sobre mim



Prof. Tetsu Sakamoto

- Biólogo/Bioinformata
- tetsu@imd.ufrn.br
- Sala A224
- Horários de atendimento: 24T12



Probabilidade (e Estatística)

IMD0033

O que é e por
quê?

Probabilidade (e Estatística)

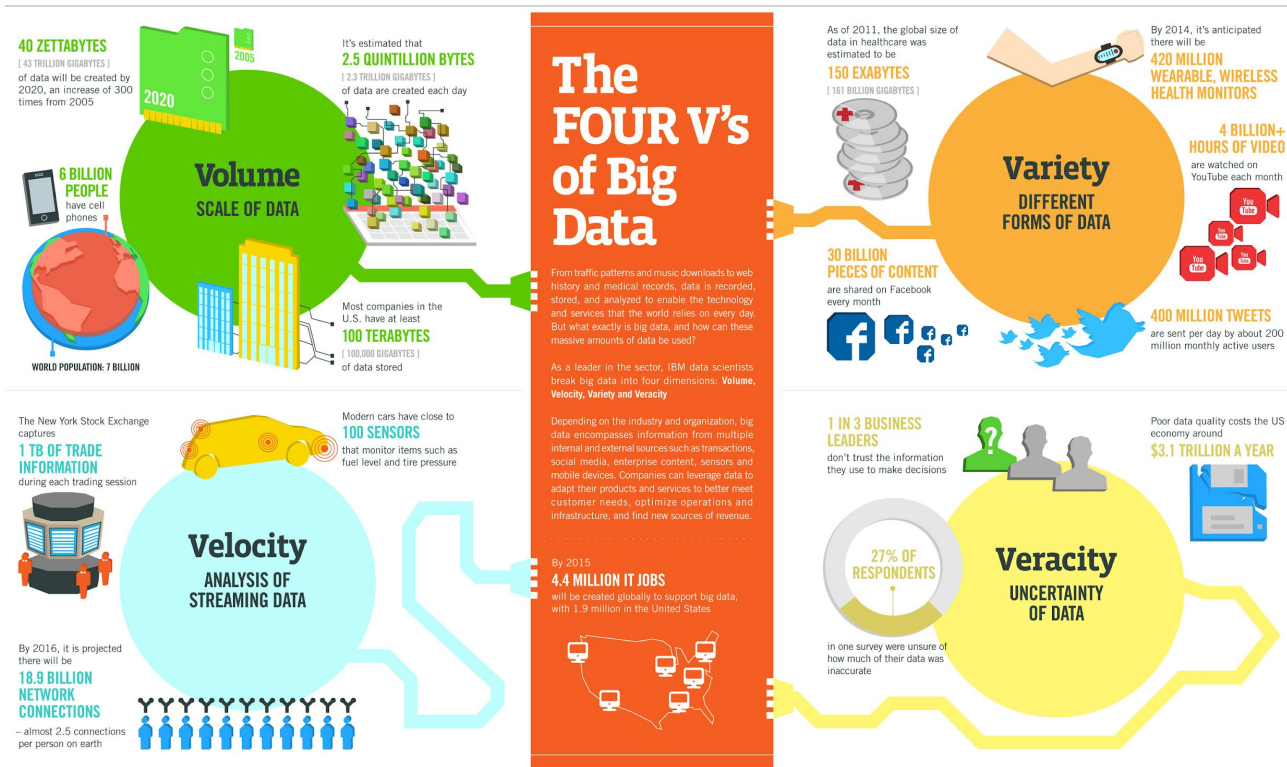
IMD0033

O que é e por quê?

Estatística - uma área da ciência que procura responder questões baseando-se em dados.

- Desenhar o experimento;
 - Coletar dados de forma apropriada;
 - Analisar os dados e checar as hipóteses;
 - Extrair conclusões confiáveis;
-

Big Data



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTec, QAS

IBM

<https://www.ibmbigdatahub.com/infographic/four-vs-big-data>

Big Data



<https://thenextweb.com/contributors/2017/07/06/will-big-data-change-use-social-media/>

Um TI “moderno”

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



Objetivos

IMD0033 - Probabilidade

Abordar noções básicas de probabilidade (e de estatística).

Estrutura da disciplina

IMD0033 - Probabilidade

Carga horária: 60 horas (72 aulas)

Data: 22/07/2019 a 20/11/2019

Horário: 24T34

Local: A101

Avaliações: 1 trabalho e 2 avaliações

de faltas permitidas: 18 aulas (9 dias)

Cronograma e temas



Introdução ao Python

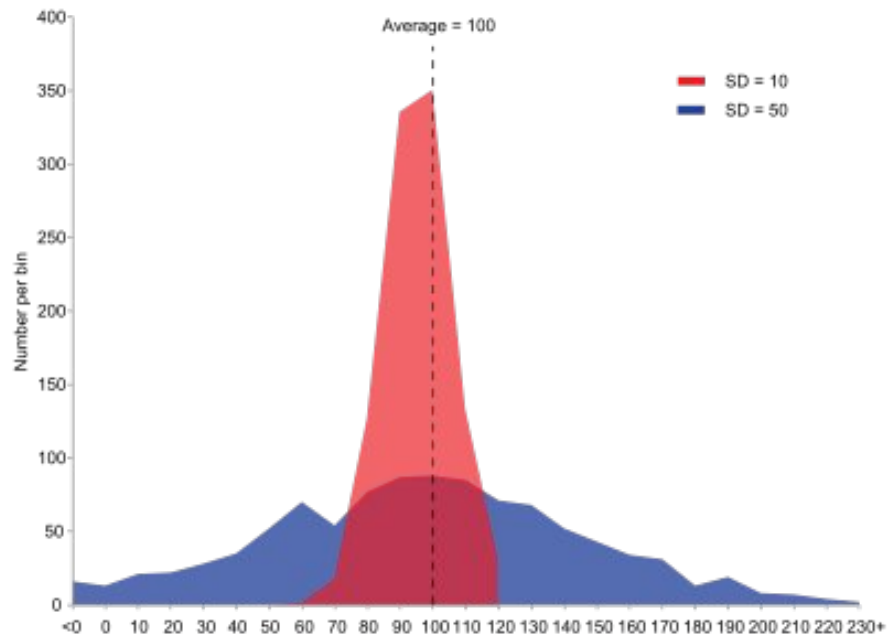
Organização dos dados

Visualização dos dados

$$\text{Mean}(\text{population}) = \mu = \frac{\sum_{i=1}^k f_i x_i}{n}$$

$$\text{StandardDeviation}(\text{population}) = \sigma = \sqrt{\frac{\sum_{i=1}^k f_i (x_i - \mu)^2}{n}}$$

$$\text{Variance}(\text{population}) = \sigma^2 = \frac{\sum_{i=1}^k f_i (x_i - \mu)^2}{n}$$



Estatística descritiva

Medidas de Tendência Central

Medidas de Dispersão



Probabilidade

Teoria de conjunto

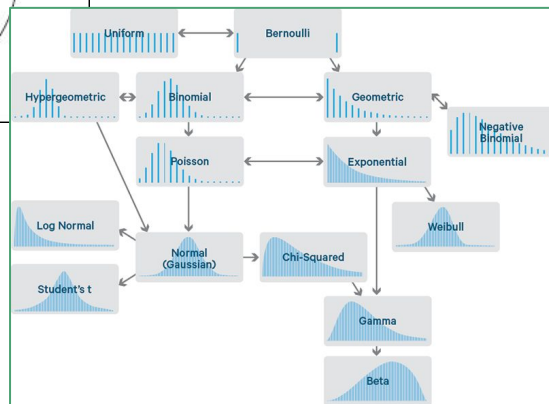
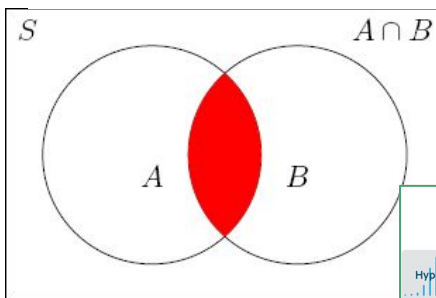
Métodos de contagem

Probabilidade

Probabilidade condicional

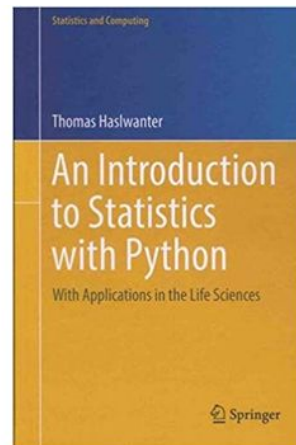
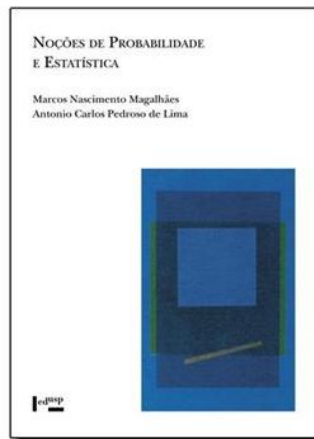
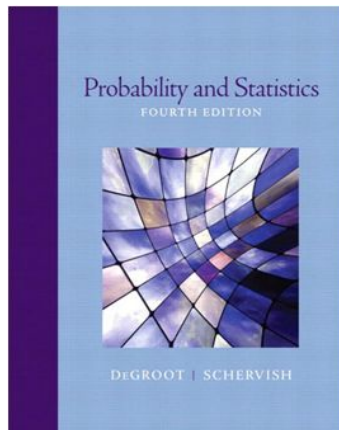
Inferência Bayesiana

Distribuições discretas e contínuas



Referências bibliográficas

- DeGroot, MH e Schervish, MJ, **Probability and Statistics**, 4a Ed, 2012;
- Magalhães, MN e de Lima ACP, **Noções de Probabilidade e Estatística**, 7a Ed, 2013;
- Hashmanter, T, **An Introduction to Statistics with Python**, 2018.



DataCamp



Plataforma de cursos online:

- Programação;
- Ciências de dados;

Nesta disciplina:

- **Introduction to Python** (Até dia 04/08/2019, 10% da 1ª avaliação)

Vocês receberão convites por email (cadastrado no SIGAA) para entrar na plataforma e completar o curso.

Aproveite esta licença para realizar outros cursos!

Perguntas?

Tem git instalado em suas máquinas?

Verificando se git está instalado

git (<https://git-scm.com/>)

Abra um terminal e dê o seguinte comando:

```
> git help
```

Os arquivos e os slides da aula estarão em **github.com/tetsufmbio/IMD0033**. Para clonar o repositório no seu computador, dê o seguinte comando:

```
> git clone https://github.com/tetsufmbio/IMD0033.git
```

Para atualizar o git, dê o seguinte comando:

```
> git pull
```



Tem Python 3 instalado em suas máquinas?

Verificando se Python 3 está instalado

Python 3 (<https://www.python.org/download/releases/3.0/>)

Abra um terminal e dê o seguinte comando:

```
> python --version
```

```
Python 3.6.8 :: Anaconda, Inc.
```

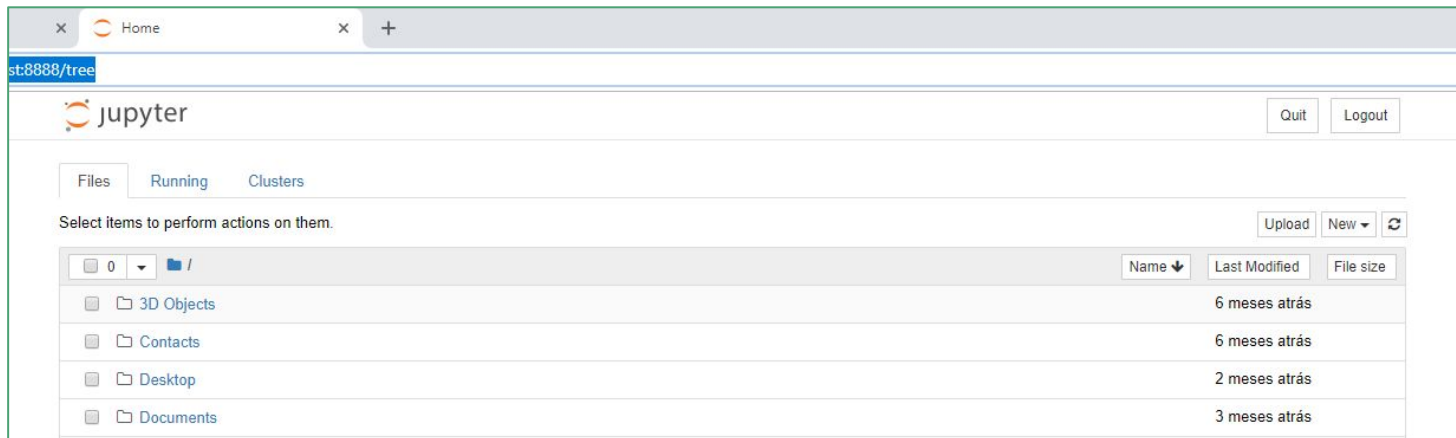
Tem Jupyter Notebook instalado em suas máquinas?

Verificando se Jupyter Notebook está instalado

Jupyter Notebook (<https://jupyter.org/>)

Abra um terminal e dê o seguinte comando:

```
> jupyter notebook
```



Não tem git,
Python3 ou
Jupyter?

Baixe o instalador do **Anaconda**
com Python 3.

(www.anaconda.com/distribution/);

Dê a permissão de execução
(`chmod +x Anaconda*`)

Execute o instalador;



Não tem git, Python3 ou Jupyter?

Instalando git via Anaconda

```
conda install -c anaconda git
```

O **Python3** e o **Jupyter Notebook**
são instalados automaticamente
pelo Anaconda.

