

# Introdução à estatística

Wasim A. Prates-Syed

Farmacêutico (FCFRP-USP), doutorando em biotecnologia (ICB-USP), divulgador científico pela UPVacina (IEARP-USP) e Project Halo (ONU).

**Email:** wasim.syed@usp.br

**Instagram:** @wasimvacinas

Dennyson L. M. Fonseca

Biomédico e doutorando no Programa Interunidades em Bioinformática USP, com bolsa FAPESP. Pesquisador convidado no Berlin institute of healthy - BIH, Universitätsmedizin Charité, Alemanha.

**Email:** dennyson@usp.br

# **Fundamentos da Análise de dados**

# **Introdução à estatística**

A mensuração do universo **nunca** é completamente **exata**.

Sempre há **variações** que podem surgir de diferentes **fontes**, como a **aleatoriedade** inerente aos processos naturais, as **limitações** intrínsecas dos instrumentos, ou até mesmo **influências externas e sistemáticas**.

A **estatística** surge como uma **ferramenta** essencial para **lidar com essas incertezas e identificar padrões**.

Para isso são necessárias a **descrição** e **análise** dos dados.

Além disso, é possível realizar **inferências** sobre a realidade com base em amostras observadas.

## Medidas de tendência central

$$\text{Média} = \frac{\text{Soma}}{n} = 1+2+3+4+5 = 15/5 = 3$$

$$\text{Mediana} = 1 \ 2 \ \mathbf{3} \ 4 \ 5 = 3$$

$$\text{Mediana} = 1 \ 2 \ \mathbf{3} \ \mathbf{4} \ 5 \ 6 = 3.5$$

## Medidas de dispersão

Quanto os dados variam?

**Variância** =  
**Dispersão média**  
**(ao quadrado)**

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

**Desvio padrão** = Raíz quadrada da variância  
**Mesma escala dos dados**

Erro Padrão =

$$\frac{\sigma}{\sqrt{n}}$$

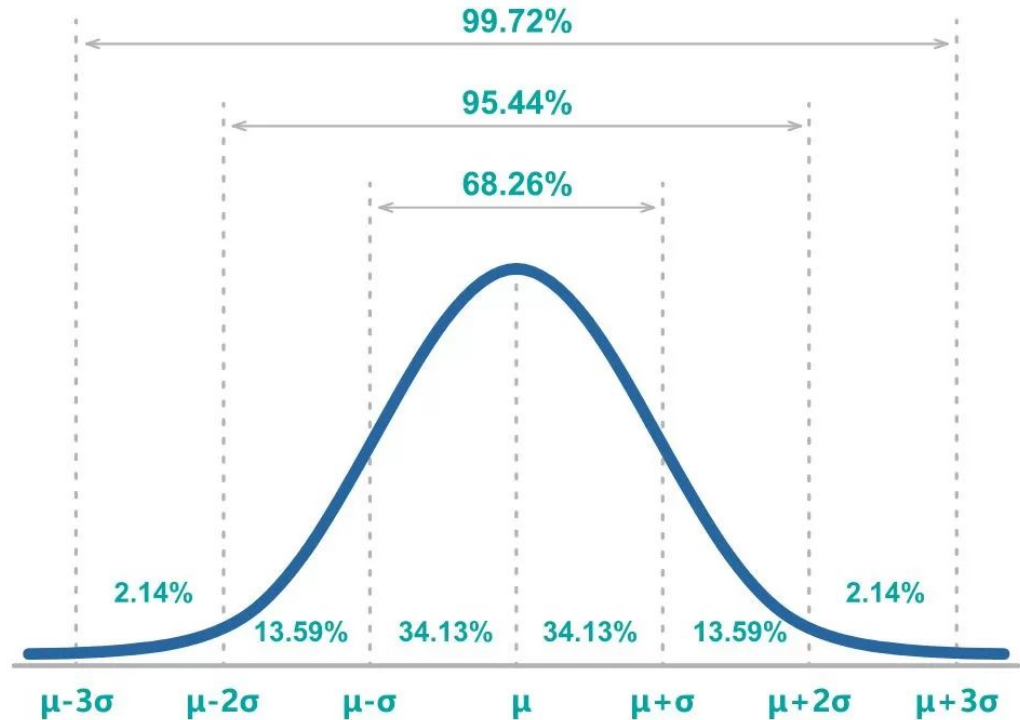
# Normalização

$$z = \frac{\overset{\text{Valor}}{x} - \overset{\text{Média}}{\mu}}{\underset{\text{Desvio padrão}}{\sigma}}$$

Valor normalizado

## Exemplo:

$$z = (6 - 3)/2 = 1.5$$

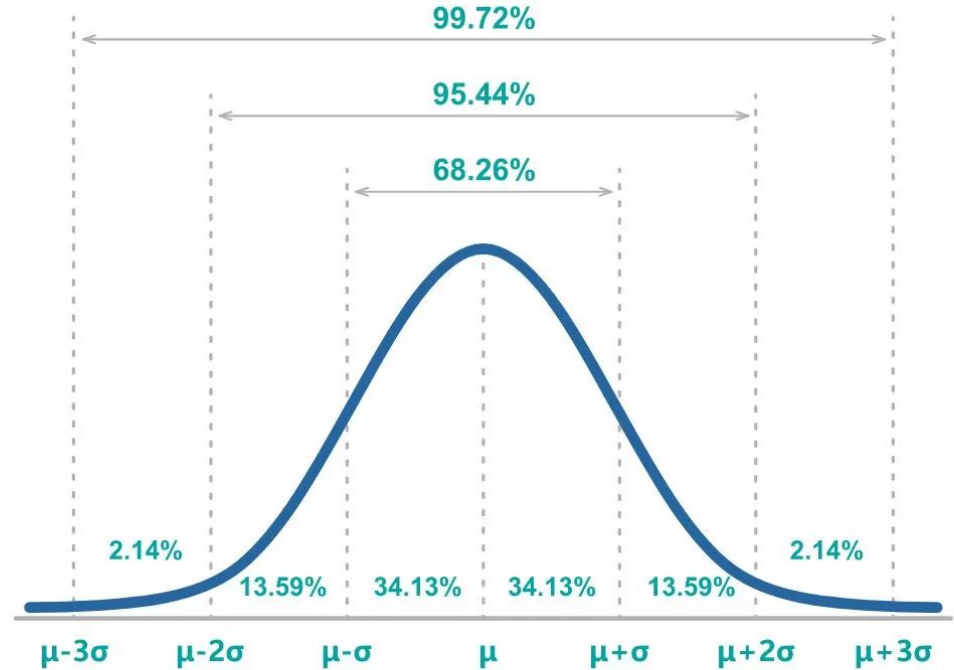
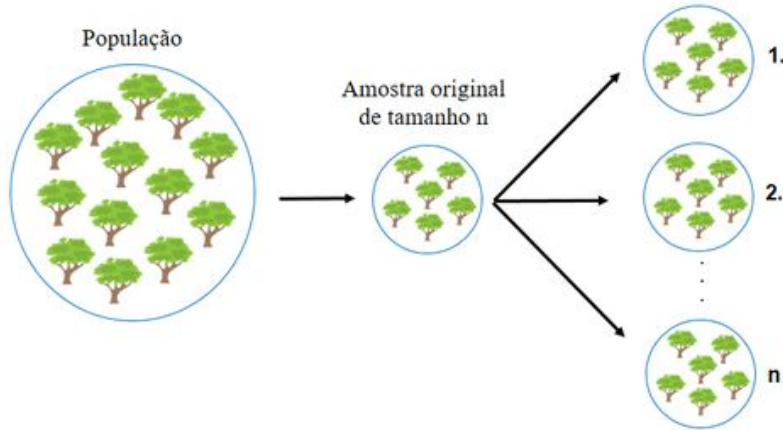




# Intervalo de confiança

O que aconteceria se fizéssemos a **amostragem** 100 vezes, onde "cairiam" os dados?

**Interpreta-se:** Usando um intervalo de 95%, 95 de 100 vezes que fizéssemos reamostragem, os valores cairiam nesse intervalo.

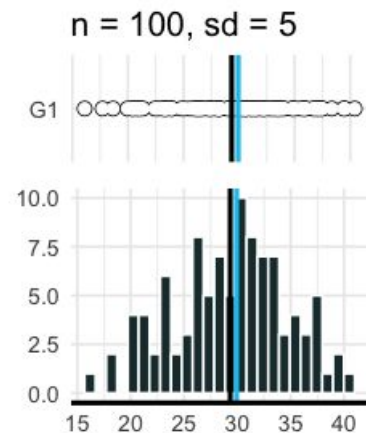
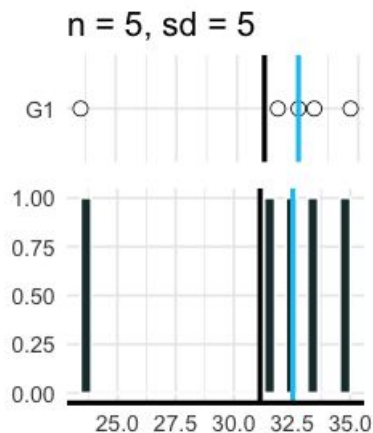
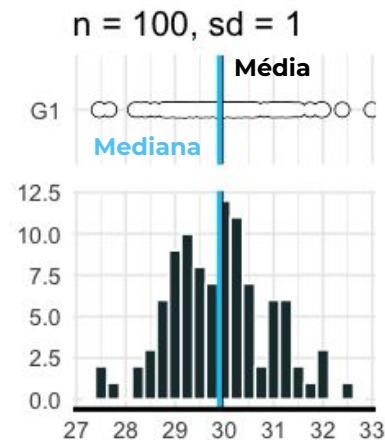
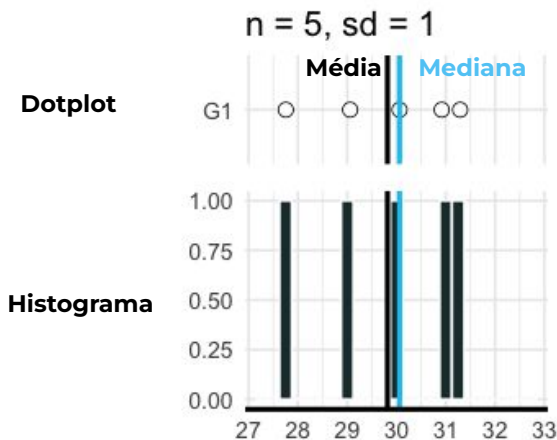


## Teorema do limite central

Quanto **maior a amostra**, mais a **distribuição** se aproxima de uma **distribuição normal**.

O valor da **média** amostral se aproxima da média da **população**.

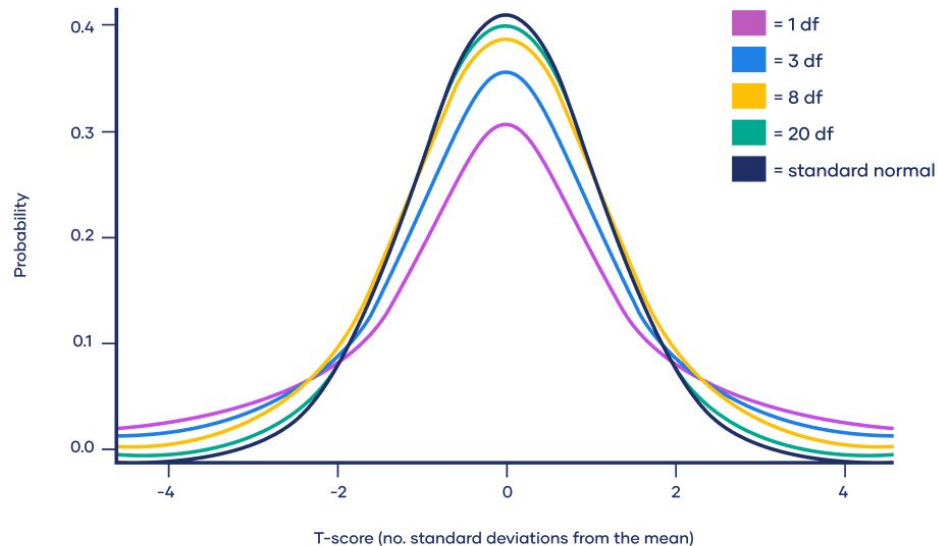
E a **dispersão** dos dados não impactam significativamente a média e a mediana.



## Distribuições

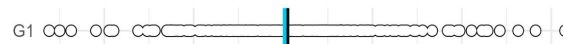
**Normal:** é simétrica em torno da média e tem a forma de um sino.

**t de Student:** semelhante à Normal, mas com caudas mais largas.

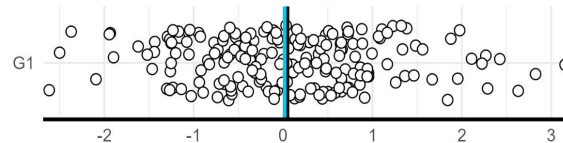


# Gráficos comuns

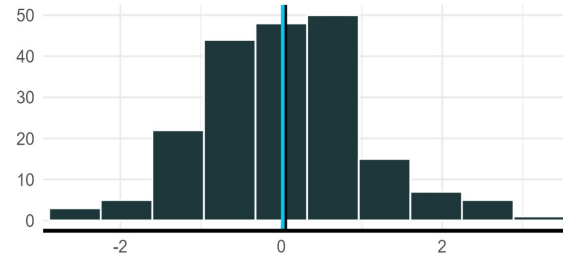
Dotplot



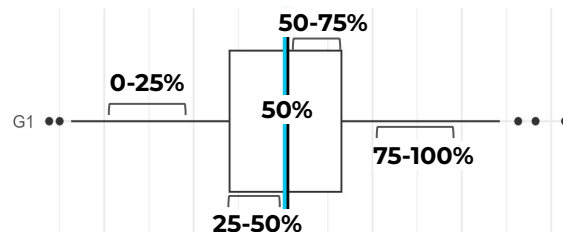
Jittered Dotplot



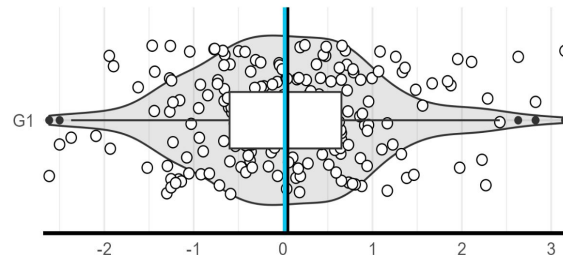
Histograma



Boxplot

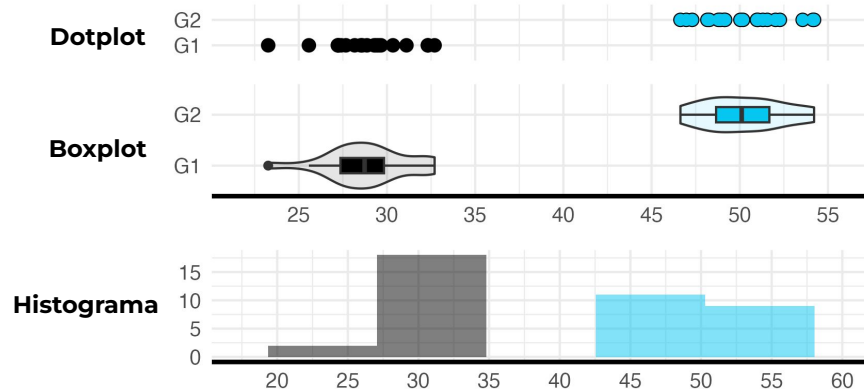


Violin plot +  
Boxplot +  
Jittered dotplot



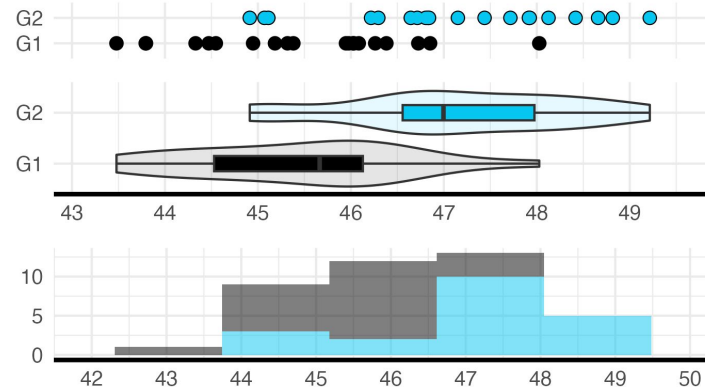
# Comparando grupos

## Muita separação



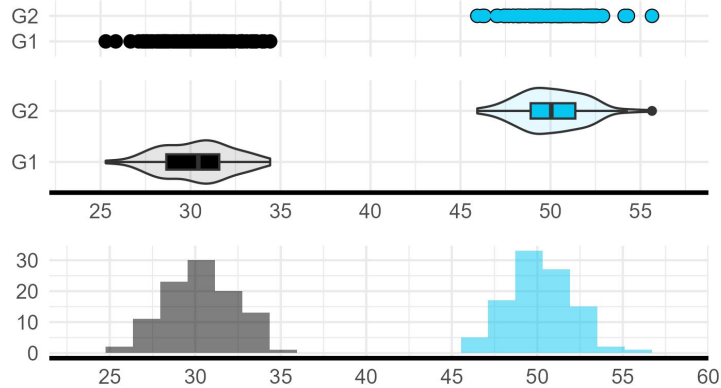
$sd = 1, n = 20$   
Distribuição t

## Pouca separação



$sd = 1, n = 20$   
Distribuição t

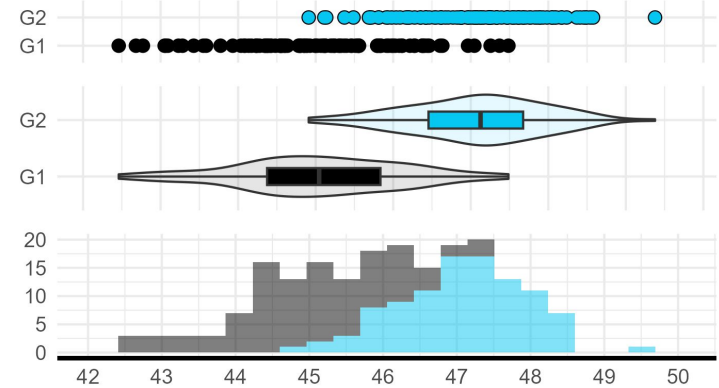
## Muita separação



$sd = 1, n = 100$

Distribuição normal  
valor-p

## Pouca separação



$sd = 1, n = 100$

Distribuição normal

Diferentes

# Estatísticas

## Descritiva

**Resumir, organizar** e apresentar os dados.

Média, mediana, moda, variância, desvio padrão, teste de normalidade

## Inferencial

Tirar **conclusões** sobre uma **população** com base em uma **amostra**.

Teste de hipótese (**pvalor**), **Intervalo de confiança**, **ANOVA**, **Correlação**, testes **paramétricos** e não **paramétricos**, Redução de dimensionalidade (PCA)



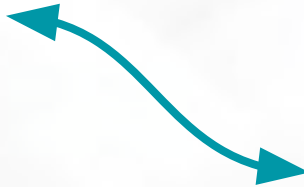
**Ciência de dados**

**x**

**Análise de dados**

## **Ciência de dados**

- Estatística inferencial
- Machine learning
- Algoritmos
- Testes estatísticos mais específicos
- Modelagem
- Foca no futuro e em novos dados
- Explicabilidade dos dados



## **Análise de dados**

- Subárea da ciência de dados
- Processamento
- Exploração
- Análises estatísticas gerais
- Foca no passado e presente

# Data analysis

Data acquisition → Data processing and cleaning → Descriptive statistics → Comparisons

---

Data visualization

Para visualizar e analisar dados, é preciso entender os **tipos de variáveis** que podem ser **plotadas**.

# Distribuição de frequências

## Variável **numérica**

Gráfico de  
**histograma**

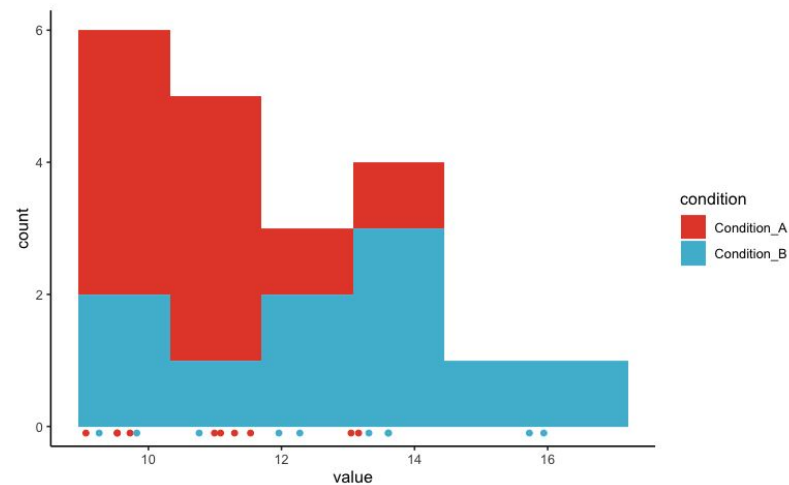
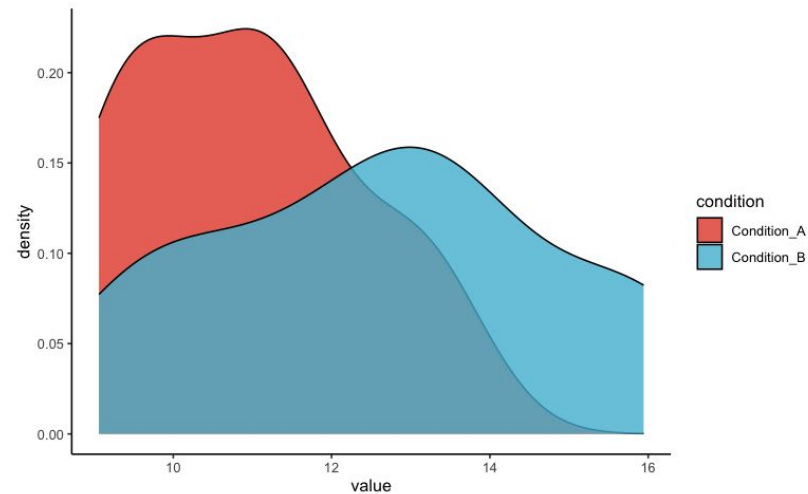


Gráfico de  
**densidade**



Variável **categórica**  
vs.  
Variável **numérica**

Boxplot

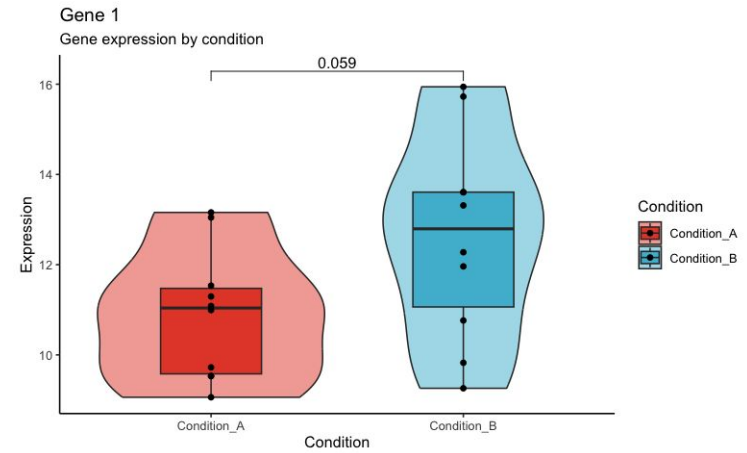
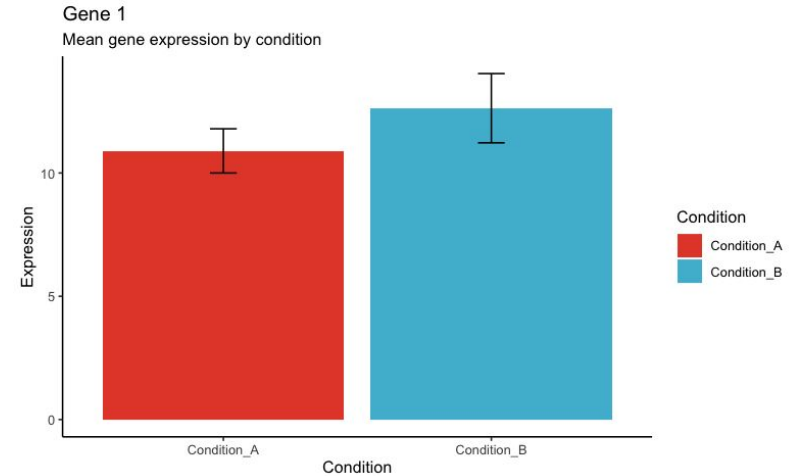


Gráfico de **barras**

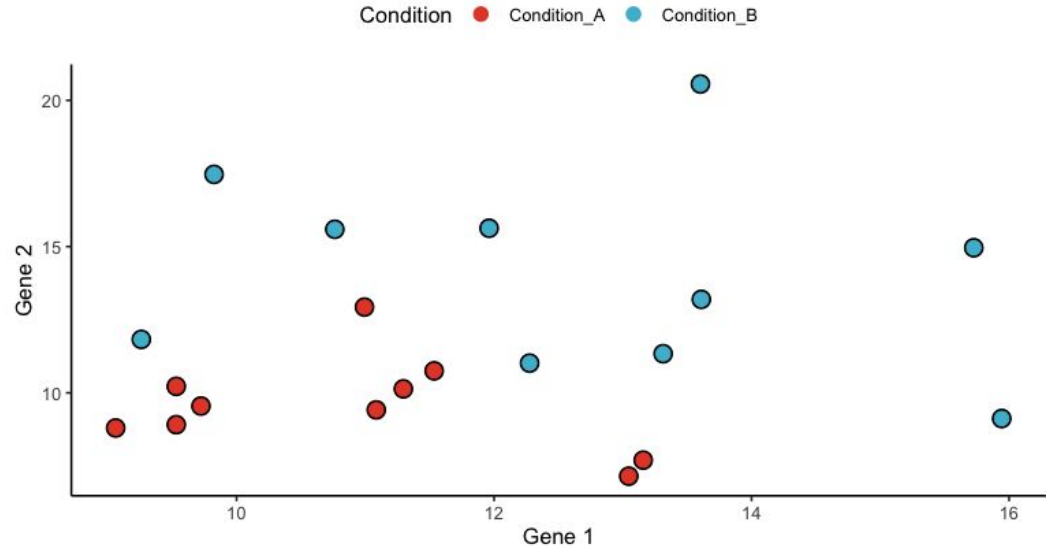


Variável **numérica**  
vs.  
Variável **numérica**

## Gráfico de **dispersão**

Gene 1 vs Gene 2

Gene expression by condition



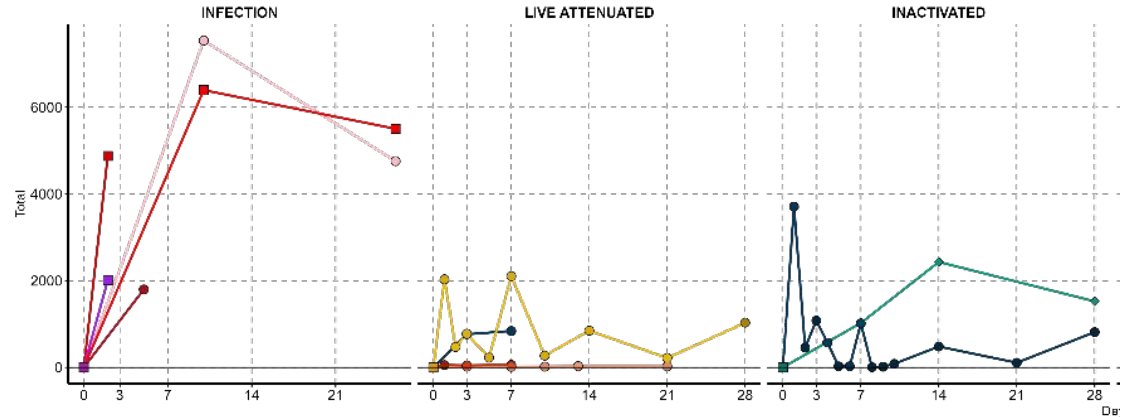
# Séries temporais

Variável **temporal**

vs.

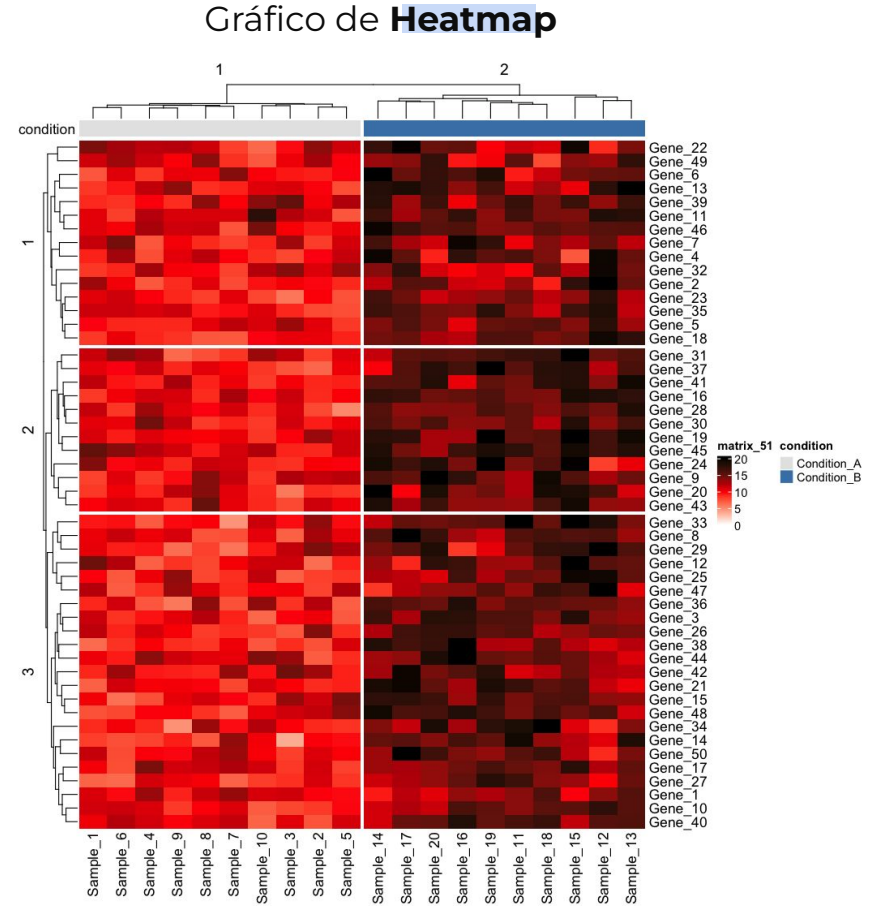
Variável **numérica**

Gráfico de **linhas**





Variável **categórica**  
vs.  
Variável **categórica**  
vs.  
Variável **numérica**





# Machine learning

[nature](#) > [nature communications](#) > [articles](#) > [article](#)

Article | Open access | Published: 09 March 2022

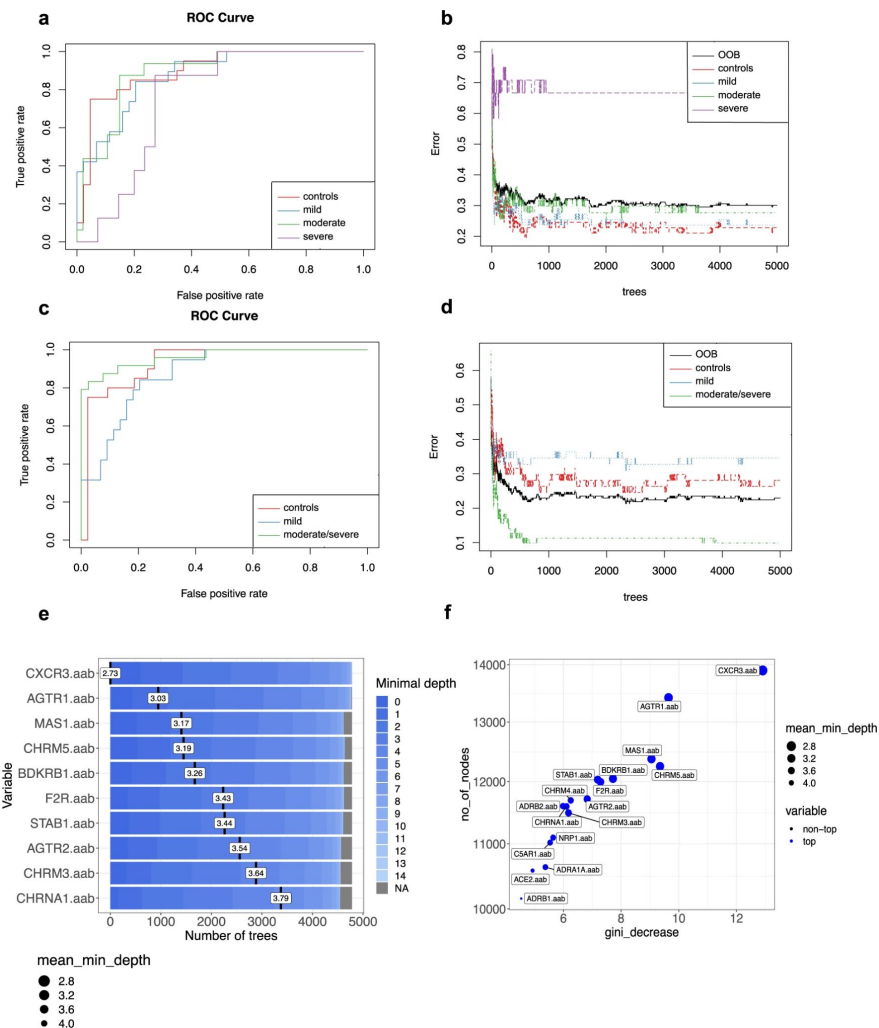
## Autoantibodies targeting GPCRs and RAS-related molecules associate with COVID-19 severity

[Otavio Cabral-Marques](#) , [Gilad Halpert](#), [Lena F. Schimke](#), [Yuri Ostrinski](#), [Aristo Vojdani](#), [Gabriela Crispim Baiocchi](#), [Paula Paccielli Freire](#), [Igor Salerno Filgueiras](#), [Israel Zyskind](#), [Miriam T. Lattin](#), [Florian Tran](#), [Stefan Schreiber](#), [Alexandre H. C. Marques](#), [Desirée Rodrigues Praça](#), [Dennyson Leandro M. Fonseca](#), [Jens Y. Humrich](#), [Antje Müller](#), [Lasse M. Gill](#), [Hanna Großhoff](#), [Anja Schumann](#), [Alexander Hackel](#), [Juliane Junker](#), [Carlotta Meyer](#), [Hans D. Ochs](#), ... [Yehuda Shoenfeld](#) 

[+ Show authors](#)

*Nature Communications* **13**, Article number: 1220 (2022) | [Cite this article](#)

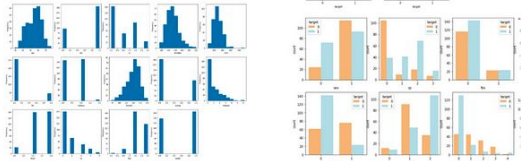
13k Accesses | 72 Citations | 84 Altmetric Metrics



# Machine Learning Algorithms - Classification

## Exploratory Data Analysis (EDA)

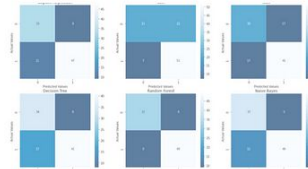
- 1) Histogram: `df.plot(kind = 'hist')`
- 2) Box Plot: `sns.boxplot()`
- 3) Grouped Bar Chart: `sns.countplot()`



## Model Evaluation

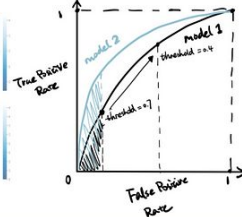
### Confusion Matrix

`confusion_matrix(y_test, y_pred)`

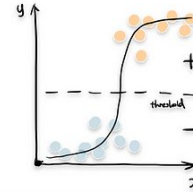


### ROC & AUC

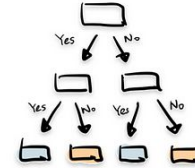
`metrics.auc(fpr, tpr)`



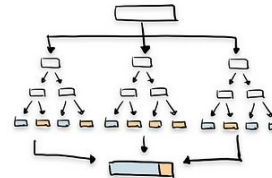
## Logistic Regression



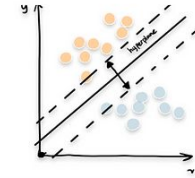
## Decision Tree



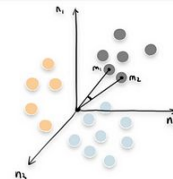
## Random Forest



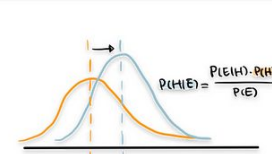
## Support Vector Machine



## K Nearest Neighbour



## Naive Bayes





Data  
resampling



Feature  
engineering



Model  
fitting



Model  
tuning



Model  
evaluation



[https://rpubs.com/chenx/tidymodels\\_tutorial](https://rpubs.com/chenx/tidymodels_tutorial)