

AGA 0505 - Análise de Dados em Astronomia 2023

0. Informações Gerais

1. Introdução: dados, probabilidades e estatística

Laerte Sodré Jr.

1o. semestre, 2023

Informações gerais

- webpage do curso:
<https://edisciplinas.usp.br/course/view.php?id=106918>
http://www.astro.iag.usp.br/~laerte/aga0505_23.html
professor: Laerte Sodré Jr. (laerte.sodre@iag.usp.br)
monitora: Melissa de Andrade Nunes (melissa.nunes@usp.br)
- objetivos:
 - ensinar noções práticas de estatística aplicada à análise de dados
 - análises frequentistas e bayesianas
 - aplicações em R (ver <https://www.r-project.org/>)
 - curso prático: entender e saber aplicar os conceitos
- o que vocês vão aprender:
 - analisar dados
 - modelar dados

Informações gerais

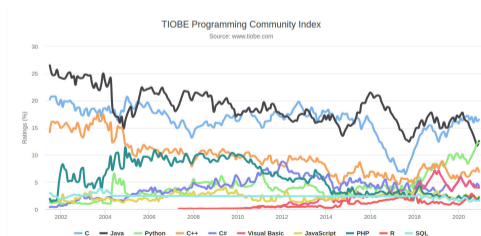
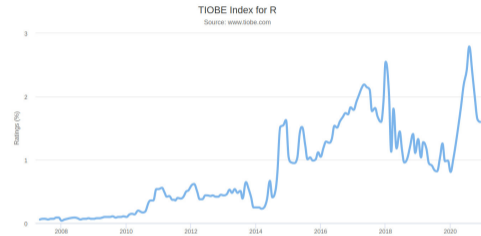
- aulas: 3as. Feiras, 14-16h, presenciais
- avaliação:
 - listas – entregar na aula seguinte (ou subtração de 0.1 na nota por dia de atraso)
 - trabalho de conclusão da disciplina (TCD): entrega 28/6
- nota final: 30% TCD; 70% listas de exercícios (removendo a nota de exercício mais baixa)

programa

1. Introdução: Probabilidades e Estatística
2. Probabilidades
3. Distribuições de Probabilidades
4. Simulações
5. Testes de Hipótese
6. O Método da Máxima Verossimilhança
7. Inferência Bayesiana
8. Comparação de Modelos
9. Aprendizado de Máquina: Princípios Gerais
10. Aprendizado de Máquina: Regressão e Classificação
11. Aprendizado de Máquina: Deep Learning
12. Séries Temporais

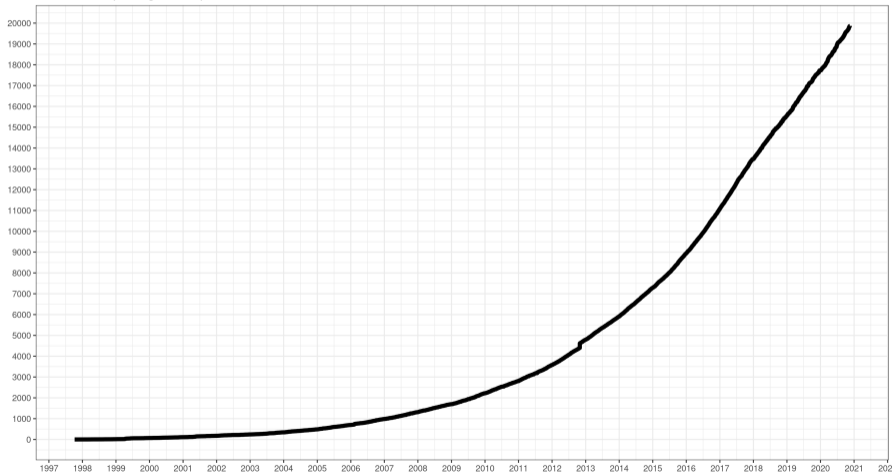
por que R?

- R está sempre na lista das 'melhores' linguagens para análise de dados
- mantida por comunidade de estatísticos
- tem muitos recursos disponíveis: livros/textos/programas
- pode produzir visualizações de alta qualidade
- “R não se aprende, se usa”



por que R?

Number of R packages ever published on CRAN



introdução ao R

- <https://www.r-project.org/>
- <http://swirlstats.com/>
- <https://edisciplinas.usp.br/course/view.php?id=3739§ion=6>
- <https://www.datacamp.com/courses/free-introduction-to-r>
- ...

bibliografia

- Numerical Recipes: The Art of Scientific Computing, Press et al, 2007
- Bayesian Data Analysis, Gelman, Carlin, Stern, Dunson, Vehtari e Rubin, 2014 (3a. edição)
- An Introduction to Statistical Learning, James, Witten, Hastie & Tibishirani, 2021 (<https://www.statlearning.com/>)
- Modern Statistical Methods for Astronomy: With R Applications, Feigelson & Babu, 2012
- Data Analysis: a Bayesian Tutorial, Sivia & Skilling, 2006
- Statistics, Data Mining, and Machine Learning in Astronomy, Ivezić, Connolly, VanderPlas & Gray, 2014 (<https://www.astroml.org/>)
- Bayesian Models for Astrophysical Data, Hilbe, de Souza & Ishida, 2017
- Bayesian Methods for the Physical Sciences, Andreon & Weaver, 2015
- Bayesian Methods in Cosmology, Trotta, arXiv:1701.01467, 2017

bibliografia (cont.)

- Deep Learning with R, Chollet & Allaire, 2018
- Deep Learning, Goodfellow, Bengio & Courville, 2016 (<https://www.deeplearningbook.org/>)
- The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys, M. Huertas-Company & F. Lanusse, arXiv:2210.01813, 2022
- The theory that would not die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy, Sharon Bertsch Mcgrayne, 2011

aula de hoje:

- Introdução

1. dados e ciência
2. o que é estatística?
3. probabilidades
4. análise de dados exploratória



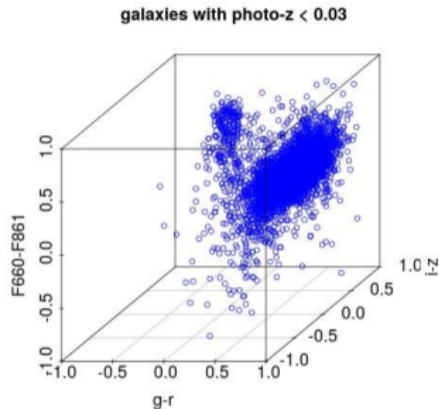
"Data! data! data!", he cried impatiently. "I can't make bricks without clay."

Arthur Conan Doyle, *The Adventure of the Copper Beeches*

- R:
 - introdução ao R
 - estatística descritiva

dados na ciência

- os **dados** são essenciais para a ciência: nossas conclusões devem ser baseadas em fatos e evidências
- os dados geralmente têm uma **natureza estatística**: devido a incertezas nas medidas ou observações, amostragem, etc
- os dados carregam **informação**: um dado é um ponto no **espaço de dados**, e este espaço tem **estrutura**
- muitas vezes o propósito da análise de dados é descobrir, analisar ou modelar essa estrutura



o que é ciência?

- ?
- interpretação da natureza através da análise de observações/experimentos e, na Física/Astronomia, usando a linguagem da matemática
- interpretação da natureza: via testes de modelos e hipóteses, através da análise de observações/experimentos/dados
- os dados têm uma natureza estatística, devido a incertezas de várias naturezas
- a análise de dados deve ser estatística!



Observations January 1870

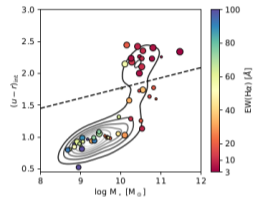
2. P. 1000' mod H. 12	○ **
30. mare'	** ○ *
2. 2000'	○ ** *
3. mare'	○ * *
3. H. 5.	* ○ *
7. mare'	* ○ **
6. mare'	** ○ *
8. mare' H. 17.	* * * ○
10. mare'	* * * ○ *
11.	* * ○ *
12. H. 4. 1000'	* ○ *
17. mare'	* ** ○ *
14. mare'	* * * ○ *

a teoria das probabilidades nos dá meios de quantificar a incerteza

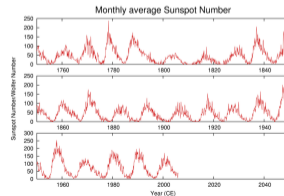
o teorema de Bayes oferece um procedimento lógico para a condução da análise estatística

tipos de dados

- dados: representação quantitativa/qualitativa da natureza
raio do Sol, tipo estelar, massas de galáxias...
- os dados podem ser de tipos diferentes:
 - números reais ou decimais (π , e , 4.1)
 - números complexos ($1+2i$)
 - lógicos (TRUE, FALSE)
 - discretos ou categóricos ('CEP1273', 'NGC 4151', '3')
- os dados podem ser de muitas formas:
 - escalares, vetores, matrizes, tensores, listas
 - tabelas, imagens, espectros, cubos de dados, séries temporais, ...



Rodríguez-Martín et al. (2022)

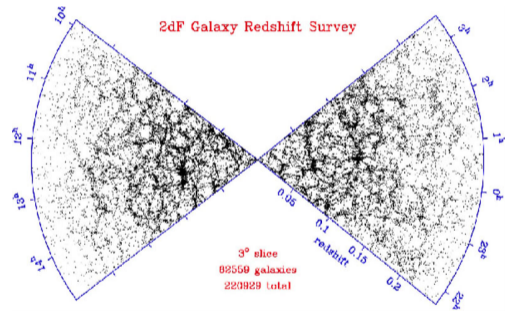


Leland McInnes (in Wikipedia)

dados em astronomia

astronomia:

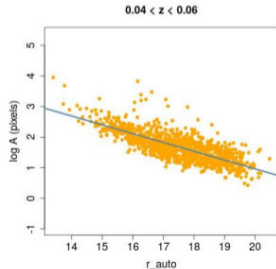
- ciência observacional e não experimental: muitas vezes o número de objetos de uma dada amostra é limitado: temos um céu! um universo!
- eventos raros/únicos: eventos não repetíveis nem sempre é possível repetir uma observação



o que é estatística?

alguns significados bem diferentes:

- **Estatística como ciência:** conjunto de métodos para análise de dados- procura-se tirar alguma conclusão a partir dos dados- estatística frequentista, estatística bayesiana



- **estatísticas:** um ou alguns números que resumem certas *propriedades dos dados*

exemplos:

- a média de um conjunto de medidas
 - o desvio padrão de um conjunto de medidas
 - a FWHM da imagem de um objeto
 - a razão sinal-ruído de um sinal
- **estatística descritiva ou exploratória:** descrição de um conjunto de dados usando *estatísticas*

estatísticas

estatísticas: números que resumem certas *propriedades dos dados*

- exemplos: média, variância, ...
- *dependem apenas dos dados!*
- esperamos que os dados possam ser descritos por uma certa *distribuição de probabilidades* e que as “estatísticas” representem propriedades desta distribuição

- em geral, quanto mais dados, melhores as estimativas
- uma boa estatística deve ser *robusta*, isto é, resistente à presença de dados espúrios (intrusos ou *outliers*)
- estatísticas podem ser justas (*unbiased*) ou viesadas (*biased*):
 - a diferença entre valores medidos e “verdadeiros” é denominada *viés* ou, em inglês, *bias*

probabilidades e estatística

inferência estatística:

- inferência: procedimento lógico para se chegar a uma conclusão baseado em evidências
- há uma disputa dentro da Estatística, tendo como base a natureza das probabilidades: bayesianos x frequentistas
- os métodos bayesianos propõem um enfoque lógico para a análise de dados baseado no teorema de Bayes
- os métodos frequentistas foram largamente dominantes durante todo o século XX
- muitos procedimentos frequentistas são muito usados em ciência ex.: método da máxima verossimilhança
- em muitos casos os resultados são semelhantes
- métodos bayesianos tendem a ser mais computacionalmente intensivos
- os procedimentos frequentistas são dominantes, mas o futuro é bayesiano (modelos bayesianos têm maior flexibilidade e poder preditivo)

natureza das probabilidades:

- frequentista: medida da frequência de eventos (em vários experimentos ou ensemble de sistemas estatisticamente equivalentes)
- bayesiana: medida da plausibilidade de uma proposição; podem ser definidas para dados, parâmetros, modelos

inferência de parâmetros:

- procedimento frequentista:
 - supõe-se que os dados são uma amostra de uma distribuição de probabilidades com parâmetros fixos
 - o ruído que afeta os dados é atribuído à amostragem dessa distribuição
 - inferência: “melhor valor dos parâmetros” + barras de erro
- procedimento bayesiano:
 - considera-se que os dados são fixos e que os parâmetros têm incertezas e são descritos por distribuições de probabilidades
 - inferência: distribuição de probabilidades dos parâmetros

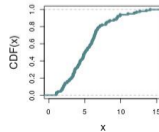
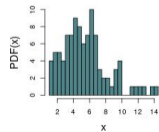
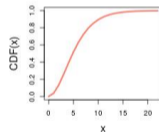
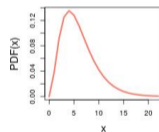
função de distribuição de probabilidades $P(x)$

- $P(x)$ pode ser uma **função discreta ou contínua**
- $P(x)$ é **normalizada**:

$$\int P(x)dx = 1 \quad \sum_{i=1}^N P(x_i) = 1$$
- se x é uma variável contínua $P(x)$ é uma *função de densidade de probabilidades*:
 $P(x)dx$: número entre 0 e 1 que mede o grau de plausibilidade ou frequência de que x esteja entre x e $x + dx$
- se x é uma variável discreta, $P(x)$ é uma *função de massa de probabilidades*

- **função de distribuição cumulativa:**

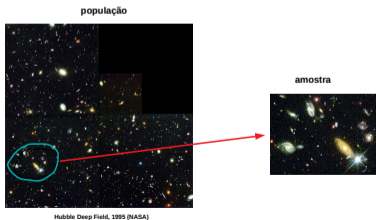
$$CDF(x) = \int_{-\infty}^x P(x')dx'$$



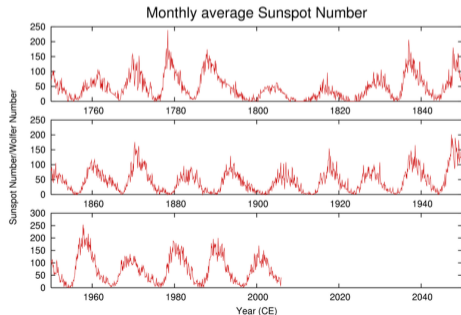
dados e probabilidades

algumas definições:

- **população**: qualquer coleção de objetos ou indivíduos que se quer analisar
ex.: as estrelas da Via Láctea
- **amostra**: um grupo extraído da população
ex.: as estrelas até 10 parsecs do Sol



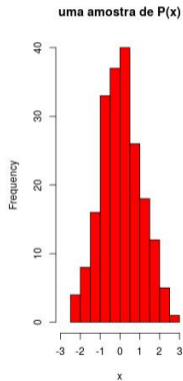
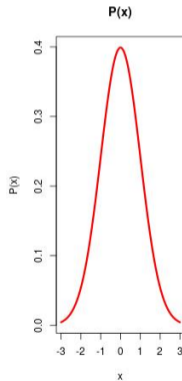
- **parâmetro**: um número que descreve alguma propriedade da população
ex.: periodicidade do ciclo solar



dados e probabilidades

dados em estatística:

- supomos que os dados são uma amostra de uma distribuição $P(x)$
- vamos considerar N medidas de uma variável x : $D = \{x_i\}$, $i = 1 \dots N$
- o conjunto de dados D pode então ser considerado uma realização/amostragem de uma variável aleatória x , com distribuição de probabilidades $P(x)$ da população



estatística descritiva ou exploratória

- objetivo: **explorar os dados para determinar propriedades da função de distribuição/massa de probabilidades da população ou da amostra**
- se $P(x)$ se refere a uma distribuição: estatísticas da população
- se $P(x)$ se refere a dados: estatísticas da amostra
- em geral se considera estatísticas de
 - posição: média, mediana, moda, percentis
 - largura: variância, desvio padrão, desvio absoluto
 - forma: skewness (assimetria), kurtosis (curtose)
- **estatísticas robustas**: são menos sensíveis à presença de *outliers*

valores esperados

- **conexão entre estatísticas e $P(x)$**
- o valor esperado de uma certa função $f(x)$ com respeito a uma distribuição de probabilidades $P(x)$ é

$$E[f] = \int_{-\infty}^{\infty} f(x)P(x)dx \qquad E[f] = \frac{1}{N} \sum_i^N f_i$$

- o valor esperado permite determinar propriedades e estatísticas para a distribuição de probabilidades $P(x)$; por exemplo, no caso de distribuições contínuas:
 - média: $\mu = \int_{-\infty}^{\infty} xP(x)dx$
 - mediana: $\int_{-\infty}^{x_{med}} P(x)dx = 1/2 = \int_{x_{med}}^{\infty} P(x)dx$
 - variância: $\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 P(x)dx$, σ : desvio quadrático médio

estatísticas que medem *posição*

- a **média** de uma população

$$\mu = E(x) = \int_{-\infty}^{\infty} xP(x)dx$$

ou de um conjunto de medidas

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- a **mediana** de uma população

$$\int_{-\infty}^{x_{med}} P(x)dx = 1/2 = \int_{x_{med}}^{\infty} P(x)dx$$

ou de um conjunto de medidas: ordene x_i do valor menor para o maior e os renumere; então

$$\bar{x}_{med} = \begin{cases} x_j & j = N/2 + 0.5, \text{ para } N \text{ ímpar} \\ (x_j + x_{j+1})/2 & j = N/2, \text{ para } N \text{ par} \end{cases}$$

a mediana é considerada uma estatística mais **robusta** que a média

estatísticas que medem *posição*

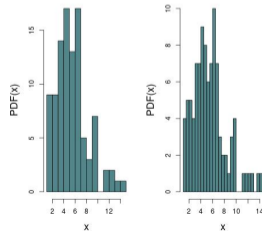
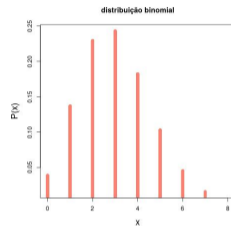
- a **moda**

x_{moda} é o valor mais provável de $P(x)$

$$\left. \frac{dP(x)}{dx} \right|_{x_{moda}} = 0$$

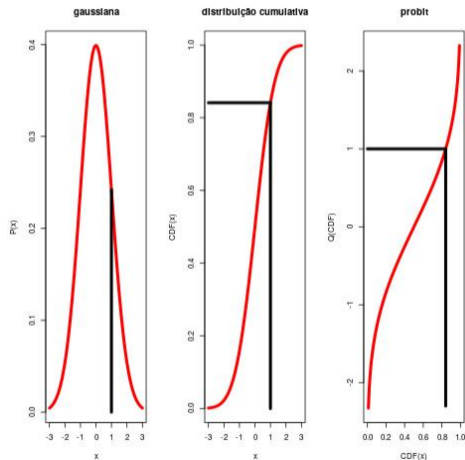
ou, no caso de um conjunto de medidas, é o valor de x_i que ocorre mais frequentemente; é a posição do pico do histograma de x_i

- a moda é bem definida para distribuições contínuas ou discretas
- a moda é mal definida para amostras de variáveis contínuas ou discretas: o pico de um histograma depende do número de bins



a função quantil

- dada uma função de distribuição de probabilidades $P(x)$, com distribuição cumulativa $CDF(x)$, a função quantil $Q(CDF)$ dá o valor de x tal que a probabilidade de se ter um valor menor ou igual a CDF é x
- $Q(CDF)$ é a função inversa da distribuição cumulativa $CDF(x)$
- a função quantil da gaussiana é chamada *probit*

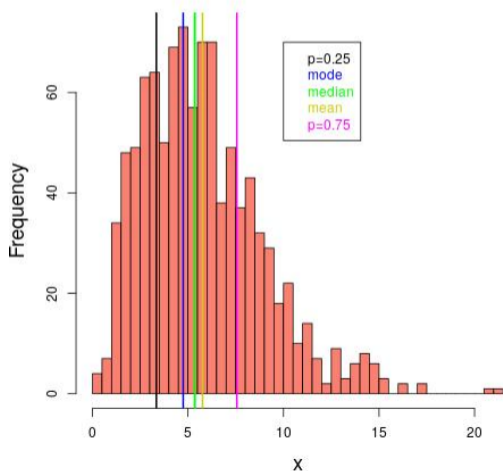


estatísticas que medem *posição*

- quartis- probabilidades que dividem uma distribuição em 4 áreas iguais:
 $q = 0.25, 0.50$ (mediana), 0.75

áreas de $P(x)$:

$0 - 0.25, 0.25 - 0.5, 0.5 - 0.75, 0.75 - 1$



estatísticas que medem *largura*

- a **variância** (desvio quadrático médio) de uma população

$$V = \int (x - \mu)^2 P(x) dx$$

ou de um conjunto de medidas (*s* de *sample*, amostra):

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

($N - 1$) e não N porque \bar{x} também é determinado dos dados

- o **desvio padrão** (*root mean square deviation*):

$$\sigma = \sqrt{V}$$

ou

$$\sigma_s = \sqrt{V_s}$$

estatísticas que medem *largura*

- o **desvio absoluto médio em relação a \bar{d}** (\bar{d} : média ou mediana de distribuições ou dados)- MAD

$$\delta = \int |x - \bar{d}|P(x)dx$$

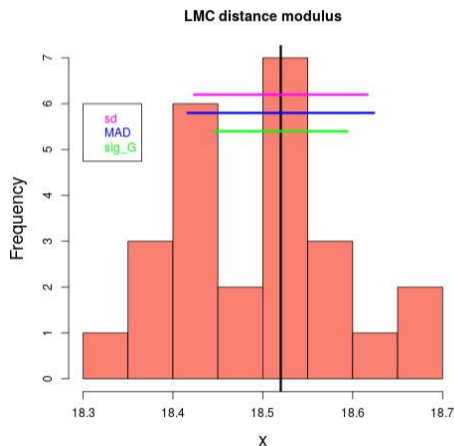
ou

$$\delta_s = \frac{1}{N} \sum_{i=1}^N |x_i - \bar{d}|$$

- pode-se obter estimativas da largura da distribuição com a distância intraquartil (com um fator para ficar equivalente ao desvio padrão de uma gaussiana):

$$\sigma_G = 0.7413(q_{75} - q_{25})$$

- estas duas estatísticas são robustas



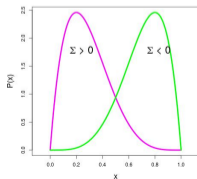
estatísticas que medem *forma*

- a **skewness** (assimetria) de uma população

$$\Sigma = \int \left(\frac{x - \mu}{\sigma} \right)^3 P(x) dx$$

ou de um conjunto de medidas:

$$\Sigma_s = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{\sigma_s} \right)^3$$



- a **kurtosis** (curtose):

$$K = \int \left(\frac{x - \mu}{\sigma} \right)^4 P(x) dx$$

ou

$$K_s = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{\sigma_s} \right)^4$$

para uma distribuição normal a curtose é igual a 3

