AGA 0505 - Análise de Dados em Astronomia

0. Informações Gerais

1. Introdução: dados, probabilidades e estatística

Laerte Sodré Jr.

1o. semestre, 2025

Informações gerais

- webpage do curso:
  http://www.astro.iag.usp.br/∼laerte/aga0505_25.html

  professor: Laerte Sodré Jr. (laerte.sodre@iag.usp.br)
  monitor: Nathan Vieira de Oliveira (nathanoliveira044@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 ANTES da aula seguinte (ou subtração de 0.1 na nota por dia de atraso)
    o máximo atraso permitido será de 2 semanas
  - prova final: 24/6
  - nota final: 60% listas de exercícios (removendo a nota de exercício mais baixa);
    30% prova; 10% participação nas aulas
- os exercícios são *individuais*!
  - você pode trabalhar com outros alunos, mas deverá apresentar seus próprios materiais e código, mostrando a lógica por trás do trabalho.
  - você pode usar o chatGPT ou outra ferramenta, mas não apresente materiais gerados por IA como se fossem seus e verifique sempre se eles estão certos!
  - você é responsável por todo trabalho que apresentar como sendo seu!

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

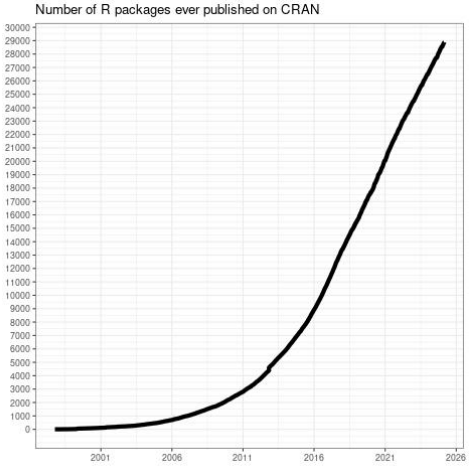
por que R?

- 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"

Feb 2025	Feb 2024	Change	Programming Language	Ratings	Change
1	1		Python	23.88%	+8.72%
2	3	↑	C++	11.37%	+0.84%
3	4	↑	Java	10.66%	+1.79%
4	2	↓	C	9.84%	-1.14%
5	5		C#	4.12%	-3.41%
6	6		JavaScript	3.78%	+0.61%
7	7		SQL	2.87%	+1.04%
8	8		Go	2.26%	+0.53%
9	12	↑	Delphi/Object Pascal	2.18%	+0.78%
10	9	↓	Visual Basic	2.04%	+0.52%
11	11		Fortran	1.75%	+0.35%
12	15	↑	Scratch	1.54%	+0.36%
13	18	⇈	Rust	1.47%	+0.42%
14	10	↓	PHP	1.14%	-0.37%
15	21	⇈	R	1.06%	+0.07%
16	13	↓	MATLAB	0.98%	-0.28%
17	14	↓	Assembly language	0.95%	-0.24%
18	19	↑	COBOL	0.82%	-0.18%
19	20	↑	Ruby	0.82%	-0.17%
20	24	⇈	Prolog	0.80%	+0.03%

https://www.tiobe.com/tiobe-index/

por que R?



introdução ao R:

- https://www.r-project.org/
- http://swirlstats.com/
- https://rstudio-education.github.io/hopr/
- R para Ciência de Dados (2ª edição),
  Wickham, Çetinkaya-Rundell & Grolemund
  https://pt.r4ds.hadley.nz/
- https://www.datacamp.com/courses/free-introduction-to-r
- informações sobre bibliotecas de vários tipos:
  https://cran.r-project.org/web/views/
- ache uma biblioteca ou função: https://rseek.org/
- novidades: https://www.r-bloggers.com/

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, 2nd. edition, https://www.manning.com/books/deep-learning-with-r
- 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:

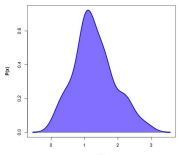
- Introdução
  1. dados e ciência
  2. o que é estatística?
  3. probabilidades
  4. análise de dados exploratória
- R:
  - introdução ao R
  - estatística descritiva

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

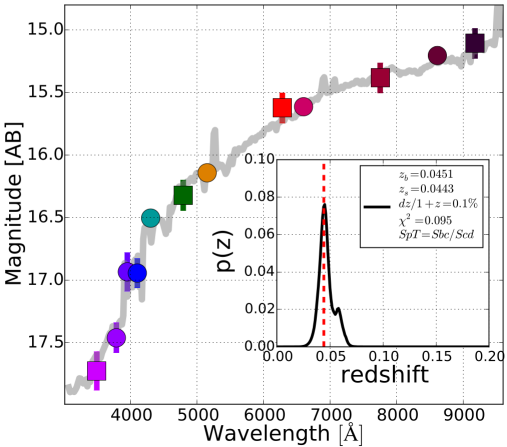
Arthur Conan Doyle, The Adventure of the Copper Beeches

dados na ciência

- **a ciência é baseada em dados**
  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
- dados podem ser descritos por **distribuições de probabilidades**



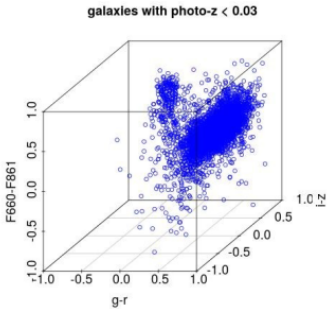
- o **teorema de Bayes** oferece um caminho natural para a análise de dados



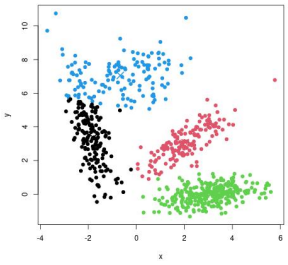
dados na ciência

- os dados carregam **informação**:
  um dado é um ponto no **espaço de dados**, e este espaço tem **estrutura**

- um dos propósitos da análise de dados é descobrir, analisar ou modelar essa estrutura

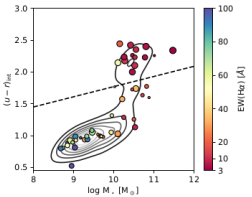


- **procedimentos comuns em análise de dados:**
  - regressão
  - classificação
  - análise não-paramétrica: análise de aglomeração, redução de dimensionalidade
  - simulações
  - ...

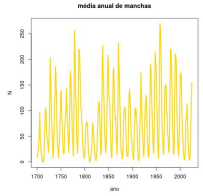


tipos de dados

- **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 (pi, 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)

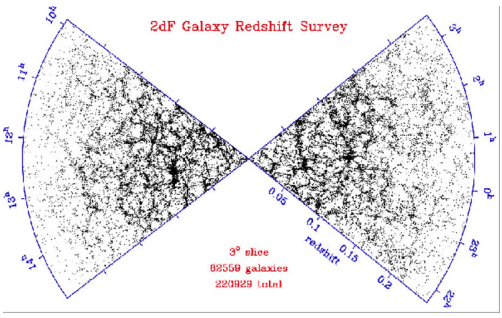


https://www.sidc.be/SILSO/datafiles

dados em astronomia

astronomia:

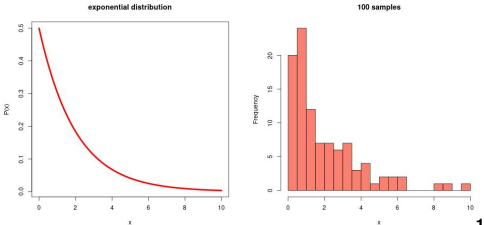
- ciência observacional e não experimental
- muitas vezes o número de objetos de uma dada amostra é limitado:
  temos um céu! um universo!
- temos eventos raros/únicos:
  eventos não repetitíveis



amostras e populações

- **variável aleatória:** variável (contínua ou discreta) cujos valores representam o resultado de um fenômeno aleatório
- assumimos que uma variável aleatória representa uma propriedade x de uma **população** e obedece a uma distribuição de probabilidades $P(x)$
- *resultados de medições ou observações podem ser tratados como variáveis aleatórias*
- resultados de medições ou observações podem ser tratados como **amostras** de $P(x)$
- amostragem de $P(x)$: o processo de obtenção de um conjunto aleatório de valores de $P(x)$
- métodos de amostragem: Monte Carlo, Monte Carlo de Cadeia de Markov, ...
- *amostragem é uma ferramenta muito importante na análise de dados!*
- dada uma amostra, podemos obter estimativas para qualquer função de x:

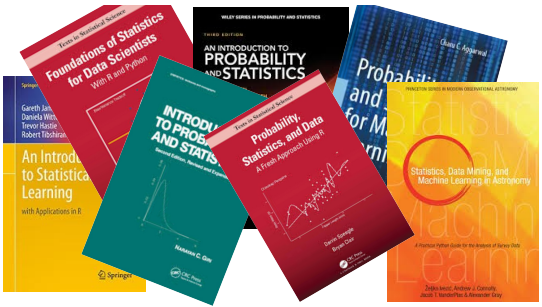
$$f(x) = \frac{1}{N}\sum_i f(x_i)$$



o que é estatística?

alguns significados bem diferentes:

- **Estatística como ciência**:
  conjunto de métodos para análise de dados



- **estatísticas**: números que resumem certas *propriedades dos dados*

  exemplos:

  - a média, o desvio padrão, os quantis de um conjunto de medidas
  - **as estatísticas dependem apenas dos dados**

- **estatística descritiva ou exploratória**:
  descrição de um conjunto de dados usando *estatísticas*
  é o primeiro passo em uma análise estatística

estatística frequentista e bayesiana

- há uma *disputa dentro da Estatística*, tendo como base a natureza das probabilidades: bayesianos x frequentistas

  natureza das probabilidades:

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

- **inferência**: procedimento lógico para se chegar a uma conclusão baseado em evidências

inferência:

- 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

estatística frequentista e bayesiana

- 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 bastante 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)

função de distribuição de probabilidades P(x)

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

$$\int P(x)dx = 1 \qquad \sum_{i=1}^{N} P(x_i) = 1$$

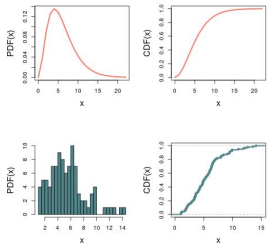
- se x é uma variável contínua, $P(x)$ é uma *função de densidade de probabilidades*:

  $P(x)dx$: 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**:

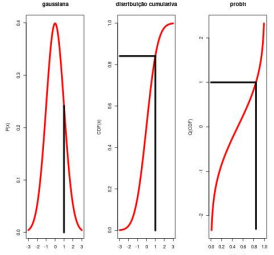
$$C(x) = \int_{-\infty}^{x} P(x')dx' \qquad (0 \leq C \leq 1)$$



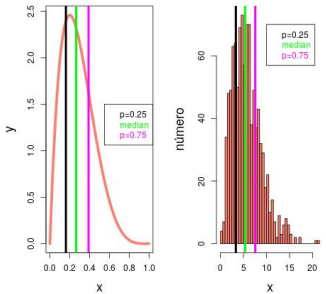
o histograma é o modo mais comum de se representar uma distribuição discreta

a função quantil

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



- **quartis**- probabilidades que dividem uma distribuição em 4 áreas iguais:
quantis $q = 0.25, 0.50$ (mediana), 0.75
áreas de $P(x)$:
$0 - 0.25, 0.25 - 0.5, 0.5 - 0.75, 0.75 - 1$



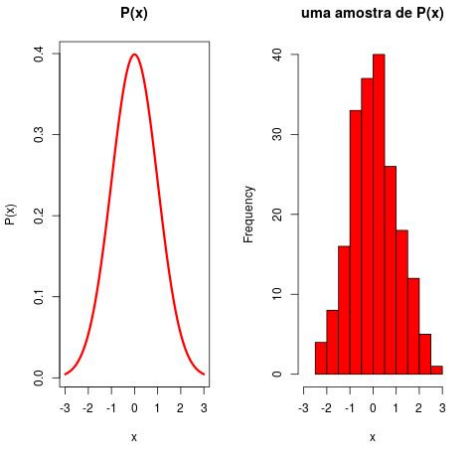
dados e probabilidades

dados em estatística:

- supomos que os dados são uma amostra de uma distribuição $P(x)$
- vamos considerar N medidas de uma variável x:

$$D = \{x_i\}, \quad i = 1...N$$

- o conjunto de dados D pode então ser considerado uma realização ou amostragem de uma variável aleatória x, cuja população tem uma distribuição de probabilidades $P(x)$



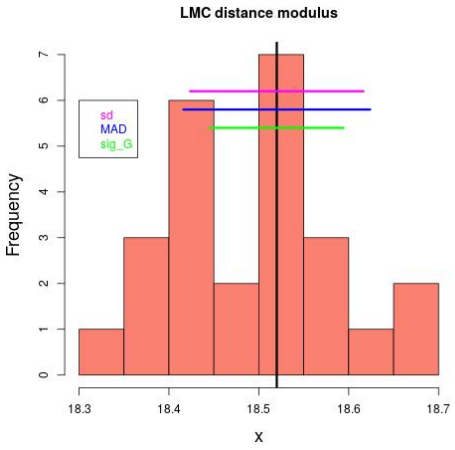
estatísticas

estatísticas: números que resumem certas *propriedades dos dados*

- exemplos: média, desvio padrão, ...
- *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*

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 consideramos estatísticas de
  - posição: média, mediana, moda
  - largura: variância, desvio padrão, desvio absoluto
  - forma: skewness (assimetria), kurtosis (curtose)



valores esperados

- **conexão entre estatísticas e $\mathbf{P(x)}$**
- 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 padrão ou desvio quadrático médio

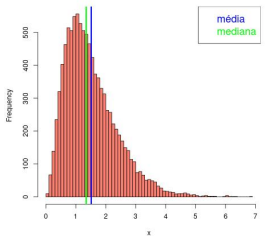
estatísticas que medem *posição*

- 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

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

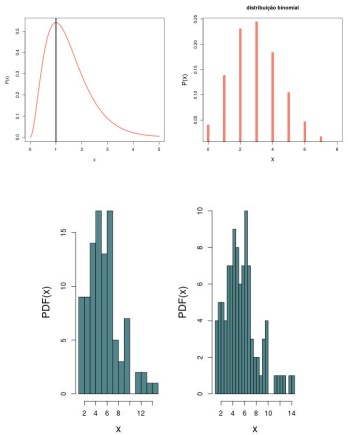
a mediana é considerada uma estatística mais **robusta** que a média

estatísticas que medem *posição*

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

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

  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



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):

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

onde $\bar{x}$ é a média das medidas

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

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

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

ou

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

estatísticas que medem *largura*

- o **desvio absoluto mediano** - MAD

$$\delta = \int |x - x_{med}| P(x) dx$$

ou

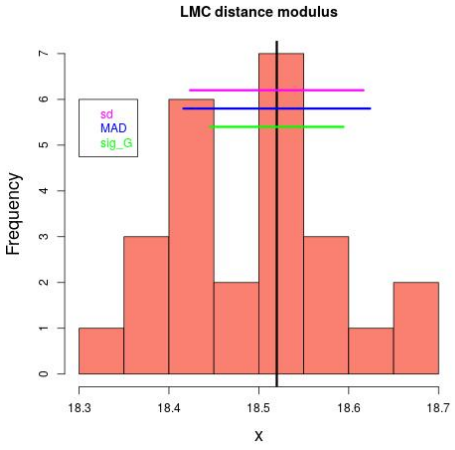
$$\delta_s = \frac{1}{N} \sum_{i=1}^{N} |x_i - x_{med}|$$

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

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

onde q_{25} e q_{75} são os quartis de probabilidade 0.25 e 0.75 da variável x

- estas duas estatísticas são robustas



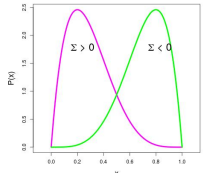
estatísticas que medem *forma*

- 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