TÓPICOS EM CIÊNCIA DE DADOS PARA O ESPORTE

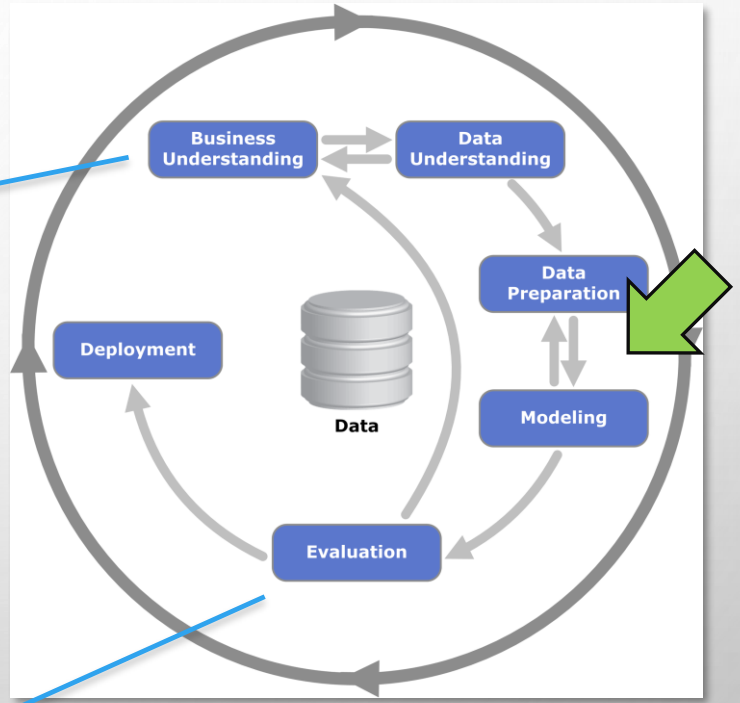
ANÁLISE EXPLORATÓRIA

DIEGO RODRIGUES DSC

INFNET

CRONOGRAMA

NÚMERO	ÁREA	AULA	TRABALHOS
1	Intro	Introdução a Disciplina e Organização do Ambiente	
2	Dados	Coleta de Dados e Sensoriamento	
3	Estatística	Variáveis Aleatórias	Grupos
4		Análise Exploratória	
5		Estatísticas para Ranqueamento	
6		Ranqueamento Estatístico : ELO	
7		Ranqueamento Estatístico : Glicko	
8		Ranqueamento Estatístico : TrueSkill	
9		Ranqueamento Estatístico : XELO	Base de Dados
10	ML	Modelos de Aprendizado de Máquina	
11		Machine Learning: Classificação	
12		Machine Learning: Regressão	
13		Machine Learning: Agrupamento	Pesquisa
14		Machine Learning: Visão Computacional	
15	Esportes	Aplicações & Artigos: Esportes Independentes	Modelo
16		Aplicações & Artigos: Esportes de Objeto	
17		Aplicações & Artigos: Esportes de Combate	
18		Aplicações & Artigos : Betting	
19	Workshop	Workshop	
20		Apresentações de Trabalhos I	Apresentação
21		Apresentações de Trabalhos II	



CRISP-DM

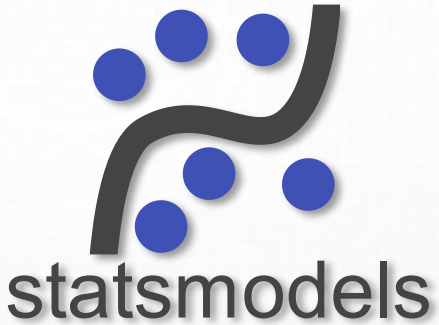
AGENDA

- PARTE 1 : TEORIA
- ANÁLISE EXPLORATÓRIA UNIVARIADA
  - VARIÁVEL BINÁRIA
  - VARIÁVEL CATEGÓRICA
  - VARIÁVEL DISCRETA
  - VARIÁVEL CONTÍNUA
  - VARIÁVEL CONTÍNUA PARAMETRIZADA
- PARTE 2 : PRÁTICA
  - PANDAS + MATPLOTLIB → VARIÁVEIS ALEATÓRIAS NO ESPORTE

SETUP INICIAL DO AMBIENTE PYTHON

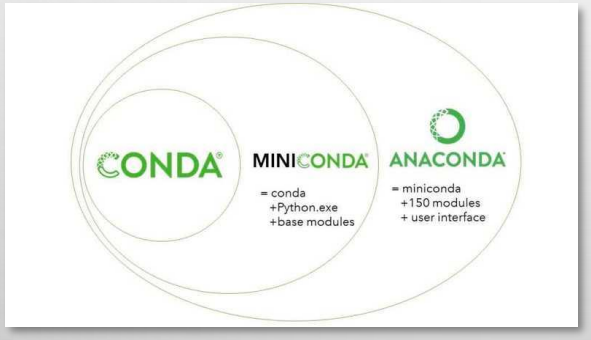
4. Variáveis Aleatórias

5. Visualização



5. Estimação e Inferência

1. Editor de Código



2. Gestor de Ambiente

3. Ambiente Python do Projeto

3. Notebook Dinâmico

VARIÁVEL BINÁRIA

$$\text{odds} = \frac{p}{1-p}$$

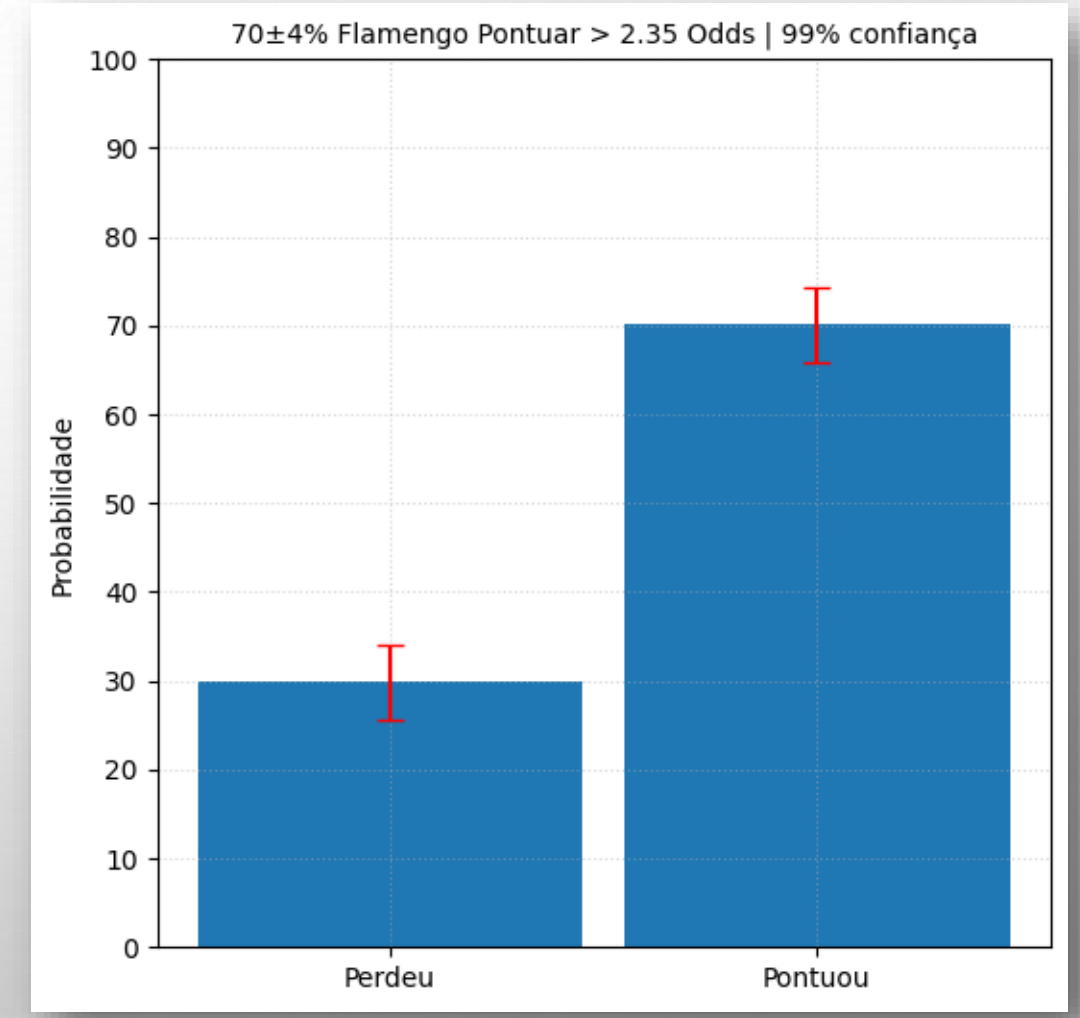
Parameters	$0 \le p \le 1$ $q = 1 - p$
Support	$k \in \{0, 1\}$
PMF	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \le k < 1 \\ 1 & \text{if } k \ge 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Mode	$\begin{cases} 0 & \text{if } p < 1/2 \\ 0, 1 & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Variance	$p(1-p) = pq$
MAD	$\frac{1}{2}$
Skewness	$\frac{q-p}{\sqrt{pq}}$
Ex. kurtosis	$\frac{1-6pq}{pq}$
Entropy	$-q \ln q - p \ln p$
MGF	$q + pe^t$
CF	$q + pe^{it}$
PGF	$q + pz$
Fisher information	$\frac{1}{pq}$

- PROBABILIDADE DE **OCORRÊNCIA DE UM EVENTO**
- DISTRIBUIÇÃO DE **BERNOULLI**
- PARÂMETRO **P → PROPORÇÃO**
- VISUALIZAÇÃO **GRÁFICO BARRA**
- **TESTE DE PROPORÇÃO**

- *If* the average, $\hat{p}$, is not near 1 or 0, and sample size n is sufficiently large (i.e. $n\hat{p} > 5$ and $n(1-\hat{p}) > 5$, the confidence interval can be estimated by a normal distribution and the confidence interval constructed thus:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- *If* $\hat{p} = 0$ and $n > 30$, the 95% confidence interval is approximately $[0, \frac{3}{n}]$ (Javanovic and Levy, 1997); the opposite holds for $\hat{p} = 1$. The reference also discusses using using $n + 1$ and $n + b$ (the later to incorporate prior information).



VARIÁVEL CATEGÓRICA

Categorical

Parameters	$k > 0$ number of categories (integer) $p_1, \ldots, p_k$ event probabilities $(p_i \geq 0,\ \Sigma p_i = 1)$
Support	$x \in \{1, \ldots, k\}$
PMF	(1) $p(x = i) = p_i$ (2) $p(x) = p_1^{[x=1]} \cdots p_k^{[x=k]}$ (3) $p(x) = [x = 1] \cdot p_1 + \cdots + [x = k] \cdot p_k$ where $[x = i]$ is the Iverson bracket
Mode	i such that $p_i = \max(p_1, \ldots, p_k)$

PROBABILIDADE DE **OCORRÊNCIA DE UM SUBCONJUNTO DE CATEGORIAS**

DISTRIBUIÇÃO **CATEGÓRICA**

PARÂMETROS $\mathbf{P_c}$ → **PROPORÇÃO DA CATEGORIA C**

VISUALIZAÇÃO **GRÁFICO BARRA**

TESTE DE PROPORÇÃO

Sum of normally distributed random variables

2 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

In probability theory, calculation of the **sum of normally distributed random variables** is an instance of the arithmetic of random variables.

This is not to be confused with the **sum of normal distributions** which forms a mixture distribution.

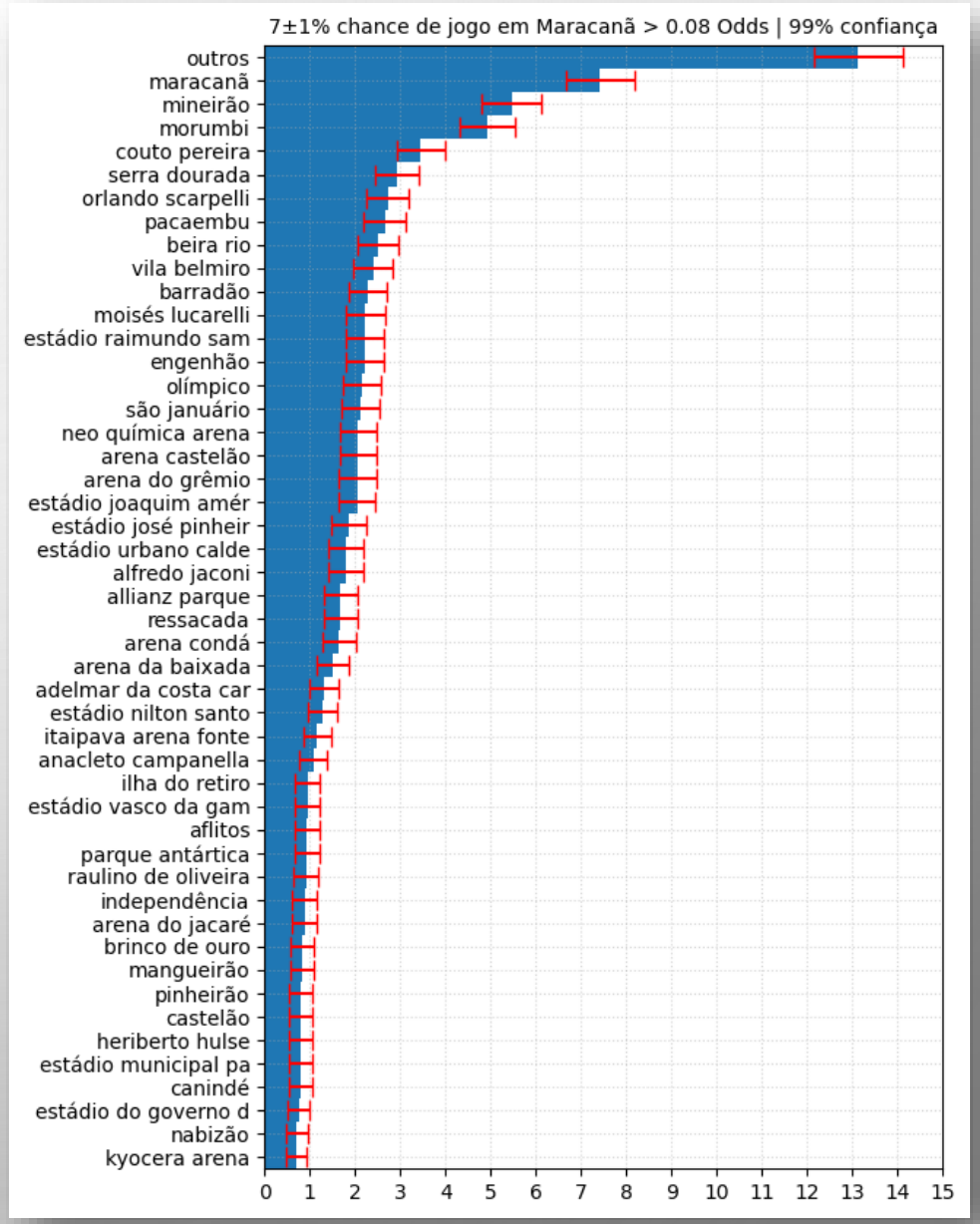
Independent random variables [edit]

Let X and Y be independent random variables that are normally distributed (and therefore also jointly so), then their sum is also normally distributed. i.e., if

$$X \sim N(\mu_X, \sigma_X^2)$$
$$Y \sim N(\mu_Y, \sigma_Y^2)$$
$$Z = X + Y,$$

then

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$



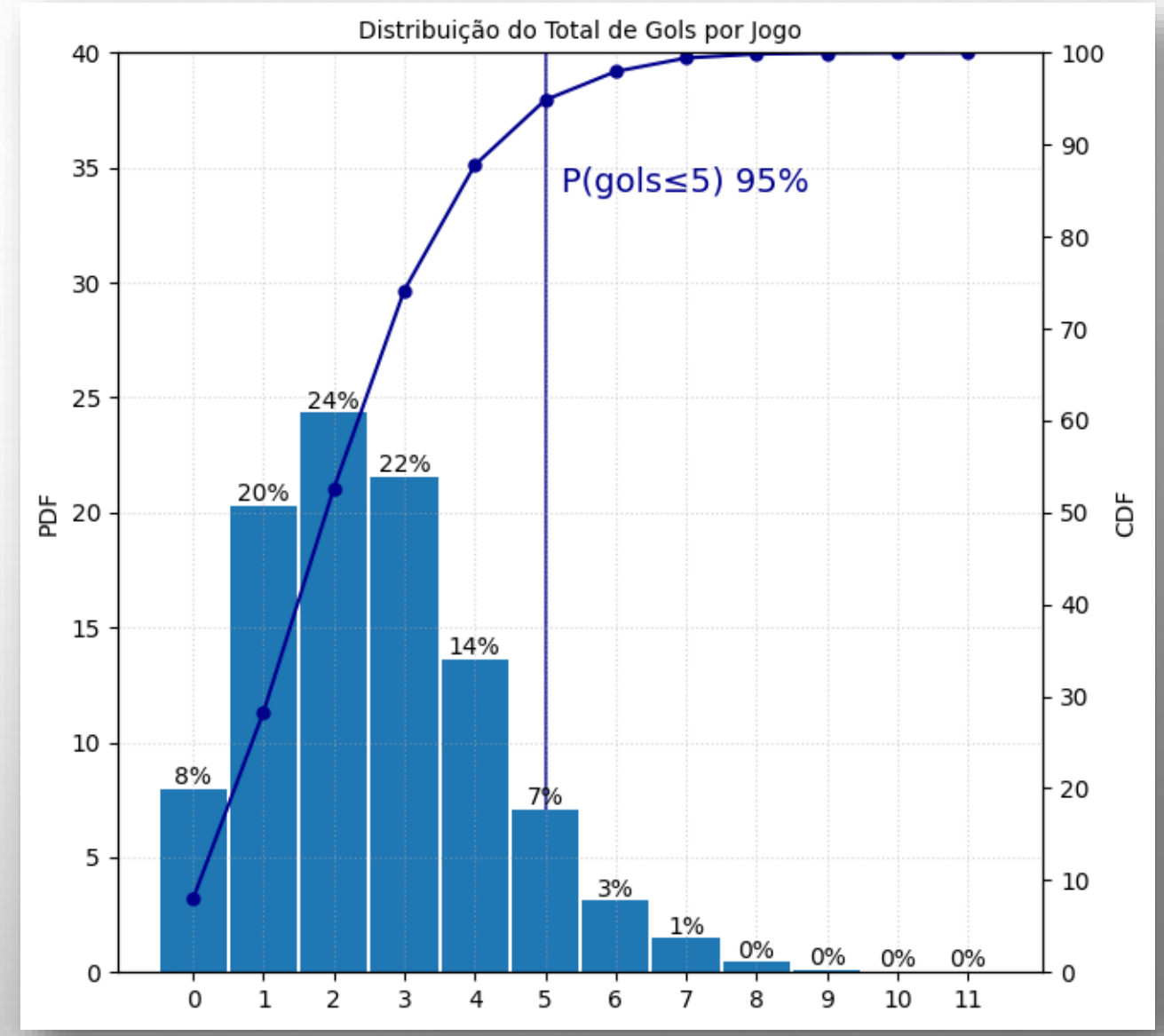
VARIÁVEL DISCRETA

PROBABILIDADE DE UM **SUBCONJUNTO DE EVENTOS OU INTERVALO**

DISTRIBUIÇÃO DE MASSA $\mathbf{P_X}$ → **PROPORÇÃO DO VALOR X**

VISUALIZAÇÃO **GRÁFICO BARRA E GRÁFICO LINHA**

TESTE DE MÉDIAS



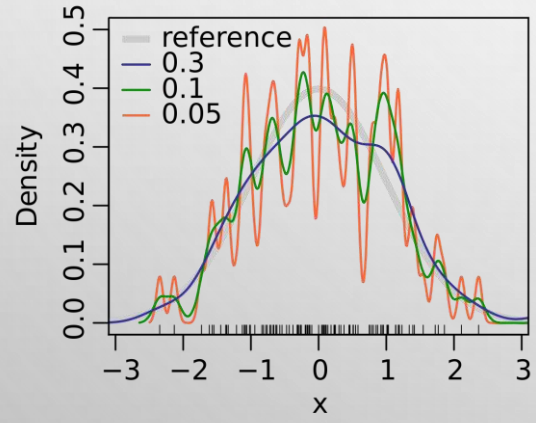
VARIÁVEL CONTÍNUA

PROBABILIDADE DE UM **INTERVALO**

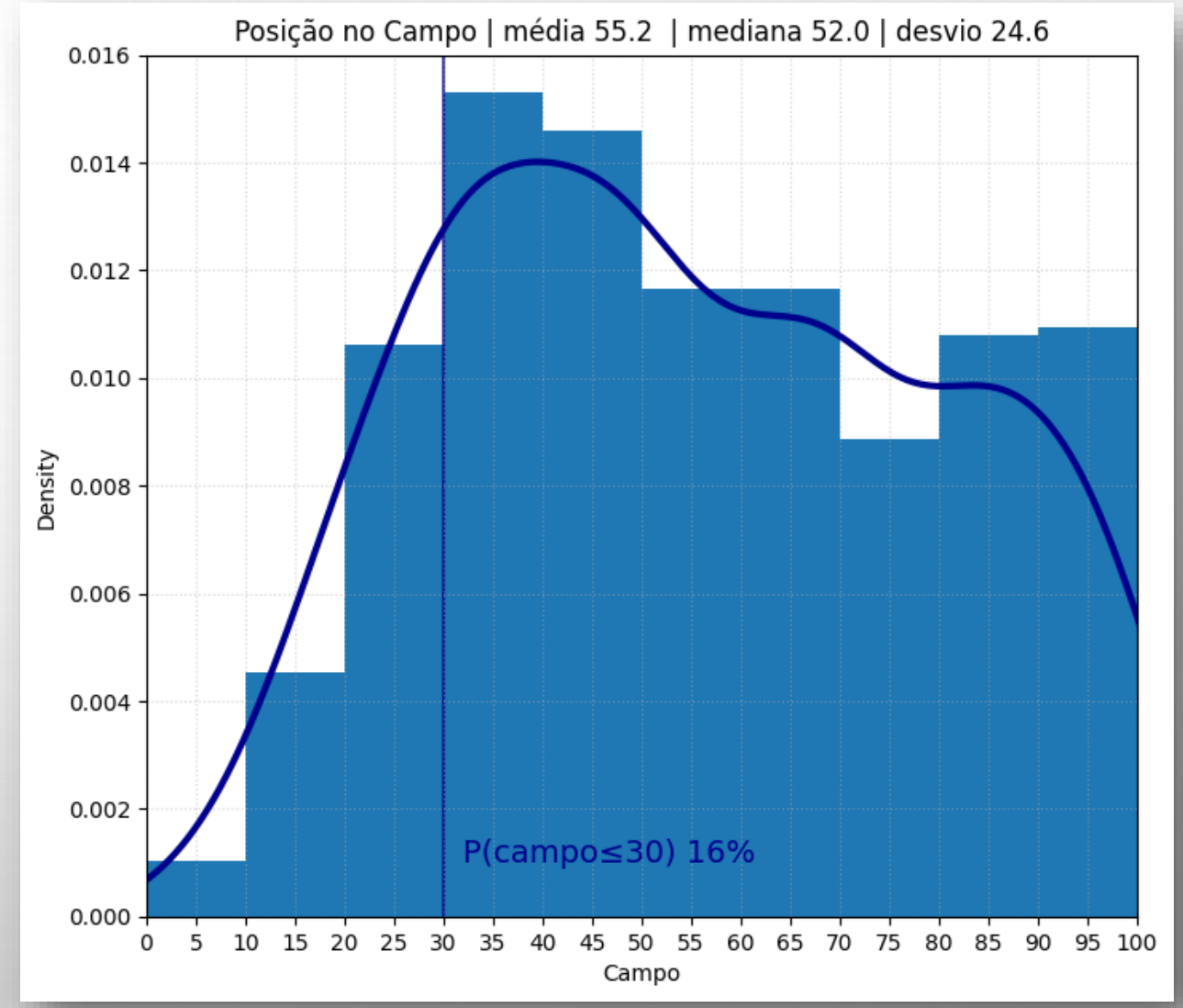
DISTRIBUIÇÃO **CONTÍNUA EM X**

VISUALIZAÇÃO **HISTOGRAMA / DENSIDADE**

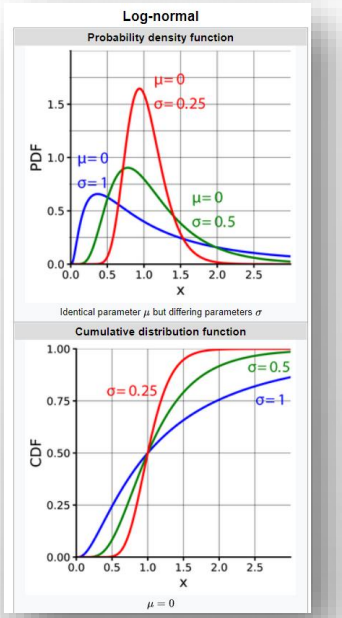
TESTE DE MÉDIAS



Kernel density estimation of 100 normally distributed random numbers using different smoothing bandwidths [Kernel Density Estimation, Wikipedia]



VARIÁVEL CONTÍNUA - PARAMETRIZAÇÃO



- ESCOLHA DE UMA **DISTRIBUIÇÃO CONTÍNUA** E **ESTIMAÇÃO DOS PARÂMETROS**
- **MEMÓRIA** DO PROCESSO OU **COMPACTAÇÃO** DOS DADOS
- VISUALIZAÇÃO **HISTOGRAMA** / **DENSIDADE**

Notation	$\text{Lognormal}(\mu, \sigma^2)$
Parameters	$\mu \in (-\infty, +\infty)$ (logarithm of scale), $\sigma > 0$
Support	$x \in (0, +\infty)$
PDF	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\exp(\mu + \sqrt{2\sigma^2}\operatorname{erf}^{-1}(2p-1))$
Mean	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$
Median	$\exp(\mu)$
Mode	$\exp(\mu - \sigma^2)$
Variance	$[\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$
Skewness	$(\exp(\sigma^2) + 2)\sqrt{\exp(\sigma^2) - 1}$
Ex. kurtosis	$\exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 6$
Entropy	$\log_2\left(\sigma e^{\mu + \frac{1}{2}}\sqrt{2\pi}\right)$
MGF	defined only for numbers with a non-positive real part, see text
CF	representation $\sum_{n=0}^{\infty} \frac{(it)^n}{n!} e^{n\mu + n^2\sigma^2/2}$ is asymptotically divergent but sufficient for numerical purposes
Fisher information	$\begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$
Method of Moments	$\mu = \log\left(\frac{\mathrm{E}[X]}{\sqrt{\frac{\mathrm{Var}[X]}{\mathrm{E}[X]^2} + 1}}\right)$, $\sigma = \sqrt{\log\left(\frac{\mathrm{Var}[X]}{\mathrm{E}[X]^2} + 1\right)}$

Statistical inference [edit]

Estimation of parameters [edit]

For determining the maximum likelihood estimators of the log-normal distribution parameters μ and σ, we can use the same procedure as for the normal distribution. Note that

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{x_i} \varphi_{\mu,\sigma}(\ln x_i),$$

where φ is the density function of the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Therefore, the log-likelihood function is

$$\ell(\mu, \sigma \mid x_1, x_2, \ldots, x_n) = -\sum_i \ln x_i + \ell_N(\mu, \sigma \mid \ln x_1, \ln x_2, \ldots, \ln x_n).$$

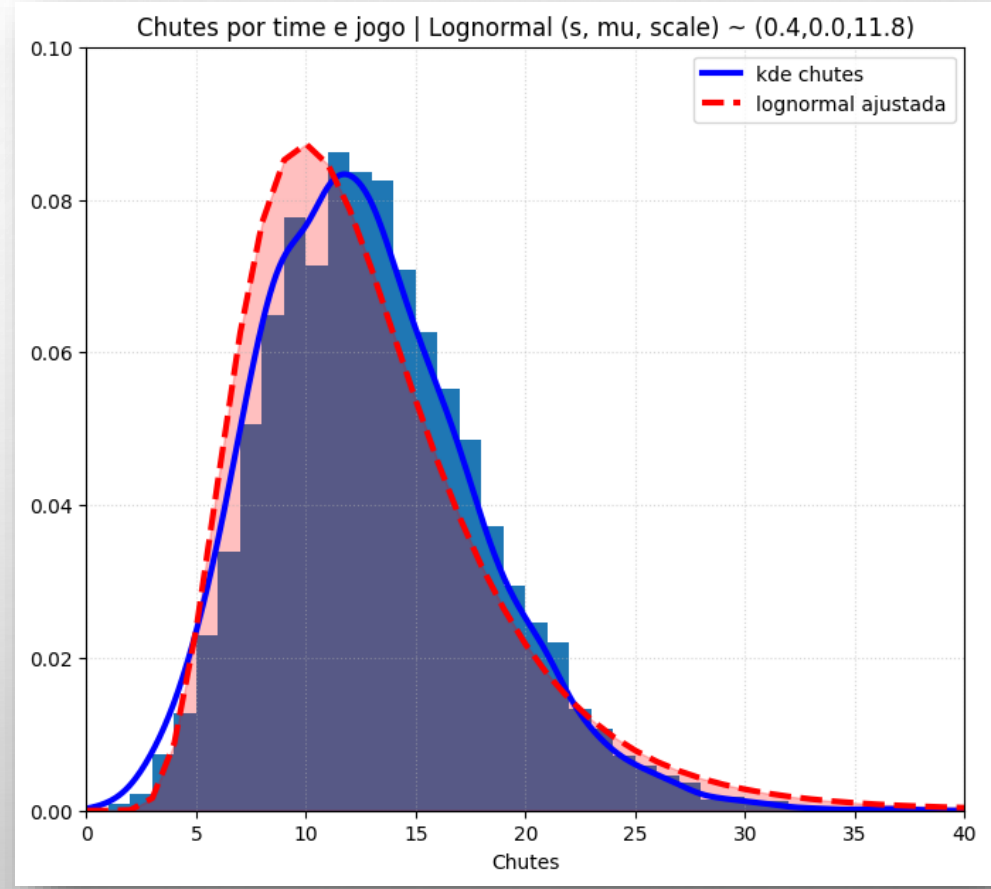
Since the first term is constant with regard to μ and σ, both logarithmic likelihood functions, ℓ and ℓ_N, reach their maximum with the same μ and σ. Hence, the maximum likelihood estimators are identical to those for a normal distribution for the observations $\ln x_1, \ln x_2, \ldots, \ln x_n$,

$$\widehat{\mu} = \frac{\sum_i \ln x_i}{n}, \qquad \widehat{\sigma}^2 = \frac{\sum_i (\ln x_i - \widehat{\mu})^2}{n}.$$

For finite n, the estimator for μ is unbiased, but the one for σ is biased. As for the normal distribution, an unbiased estimator for σ can be obtained by replacing the denominator n by $n-1$ in the equation for $\widehat{\sigma}^2$.

When the individual values $x_1, x_2, \ldots, x_n$ are not available, but the sample's mean $\bar{x}$ and standard deviation s is, then the Method of moments can be used. The corresponding parameters are determined by the following formulas, obtained from solving the equations for the expectation $\mathrm{E}[X]$ and variance $\mathrm{Var}[X]$ for μ and σ:

$$\mu = \ln\left(\frac{\bar{x}}{\sqrt{1 + \widehat{\sigma}^2/\bar{x}^2}}\right), \qquad \sigma^2 = \ln\left(1 + \widehat{\sigma}^2/\bar{x}^2\right).$$



PRÓXIMA AULA
LEITURA: ESTATÍSTICA – TESTES DE HIPÓTESE – CORRELAÇÃO DE PEARSON