TÓPICOS EM CIÊNCIA DE DADOS PARA O ESPORTE



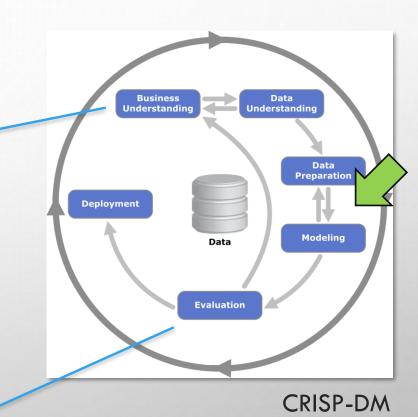
ANÁLISE EXPLORATÓRIA

DIEGO RODRIGUES DSC

INFNET

CRONOGRAMA

DIA	NÚMERO	ÁREA	AULA	TRABALHOS
30/1/2024	1	Intro	Introdução a Disciplina e Organização do Ambiente	
1/2/2024	2	Dados	Coleta de Dados e Sensoriamento	
6/2/2024	3	Estatística	Variáveis Aleatórias	Grupos
8/2/2024	4		Análise Exploratória	
15/2/2024	5		Estatísticas para Ranqueamento	
20/02/2024	6		Ranqueamento Estatístico : ELO	
22/02/2024	7		Ranqueamento Estatístico : Glicko	
27/2/2024	8		Ranqueamento Estatístico : TrueSkill	
29/2/2024	9		Ranqueamento Estatístico : XELO	Base de Dados
5/3/2024	10	ML	Modelos de Aprendizado de Máquina	
7/3/2024	11		Machine Learning: Classificação	
12/3/2024	12		Machine Learning: Regressão	
14/3/2024	13		Machine Learning: Agrupamento	Pesquisa
19/3/2024	14		Machine Learning: Visão Computacional	
21/3/2024	15	Esportes	Aplicações & Artigos: Esportes Independentes	Modelo
26/3/2024	16		Aplicações & Artigos: Esportes de Objeto	
28/3/2024	17		Aplicações & Artigos: Esportes de Combate	
2/4/2024	18		Aplicações & Artigos : Betting	
4/4/2024	19	Workshop	Workshop	
9/4/2024	20		Apresentações de Trabalhos I	Apresentação
11/4/2024	21		Apresentações de Trabalhos II	



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

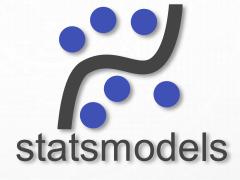
SETUP INICIAL DO AMBIENTE PYTHON



4. Variáveis Aleatórias



5. Visualização



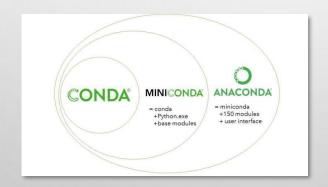
5. Estimação e Inferência







1. Editor de Código



2. Gestor de Ambiente



3. Ambiente
Python do Projeto



3. Notebook Dinâmico

VARIÁVEL BINÁRIA

Parameters	$0 \leq p \leq 1$
	q = 1 - p
Support	$k \in \{0,1\}$
PMF	$\left\{egin{array}{ll} q=1-p & ext{if } k=0 \ p & ext{if } k=1 \end{array} ight.$
CDF	$\left\{egin{array}{ll} 0 & ext{if } k < 0 \ 1-p & ext{if } 0 \leq k < 1 \ 1 & ext{if } k \geq 1 \end{array} ight.$
Mean	p
Median	$\begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$ $\begin{cases} 0 & \text{if } p < 1/2 \\ 0,1 & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
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	$\frac{1}{pq}$

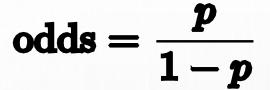
•	PROBABILIDADE DE OCORRÊNCIA
	DE UM EVENTO

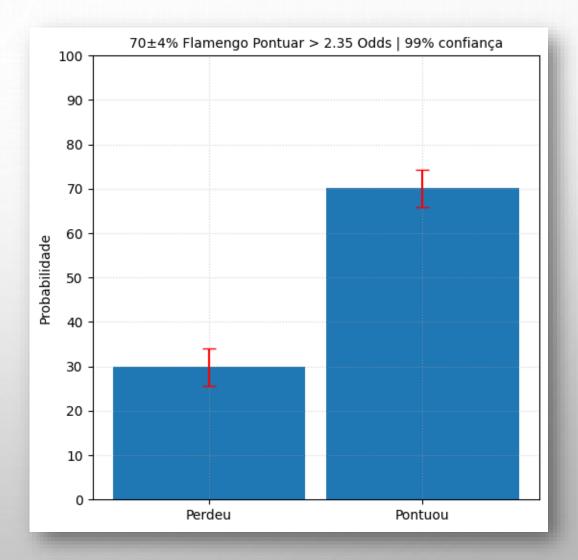
- DISTRIBUIÇÃO DE **BERNOULLI**
- PARÂMETRO P → PROPORÇÃO
- VISUALIZAÇÃO GRÁFICO BARRA
- TESTE DE PROPORÇÃO

• 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:

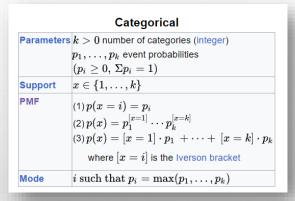
$$\hat{p}\pm z_{1-lpha/2}\sqrt{rac{\hat{p}(1-\hat{p})}{n}}$$

• 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



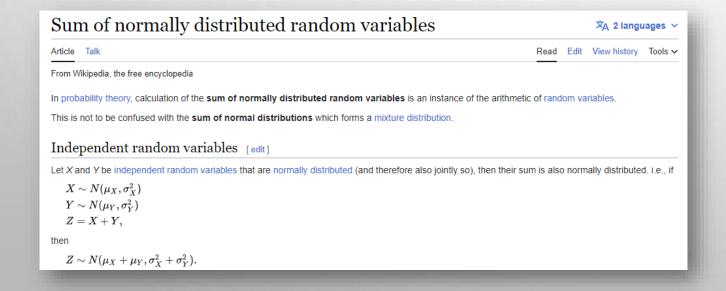
PROBABILIDADE DE **OCORRÊNCIA DE UM**SUBCONJUNTO DE CATEGORIAS

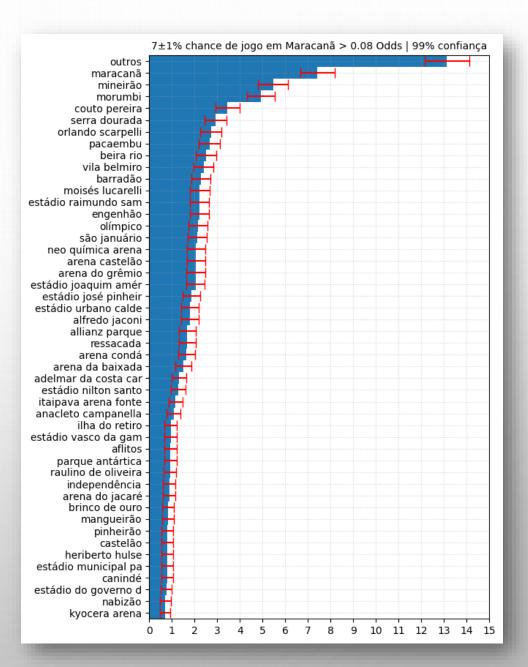
DISTRIBUIÇÃO CATEGÓRICA

PARÂMETROS P_C → PROPORÇÃO DA CATEGORIA C

VISUALIZAÇÃO GRÁFICO BARRA

TESTE DE PROPORÇÃO





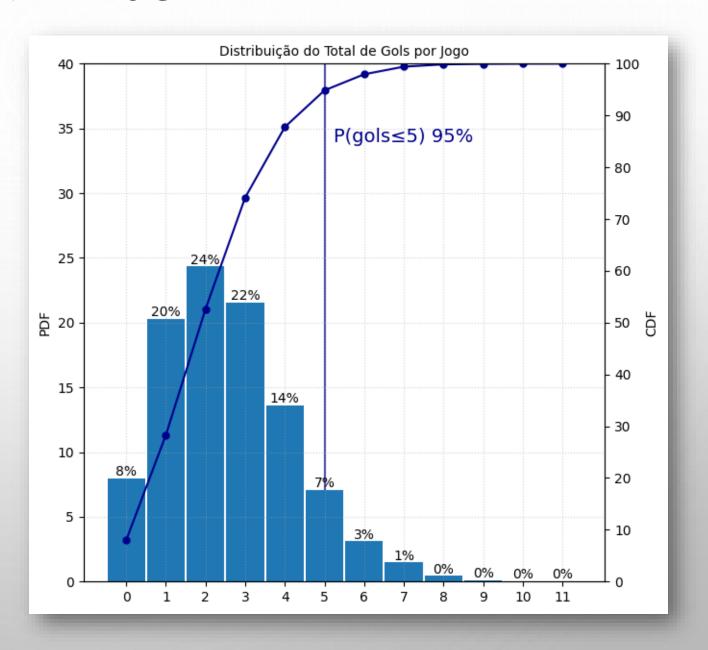
VARIÁVEL DISCRETA

PROBABILIDADE DE UM

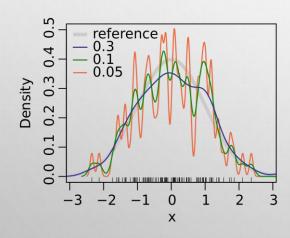
SUBCONJUNTO DE EVENTOS OU

INTERVALO

DISTRIBUIÇÃO DE MASSA P_X → PROPORÇÃO DO VALOR X
VISUALIZAÇÃO GRÁFICO BARRA E GRÁFICO LINHA
TESTE DE MÉDIAS

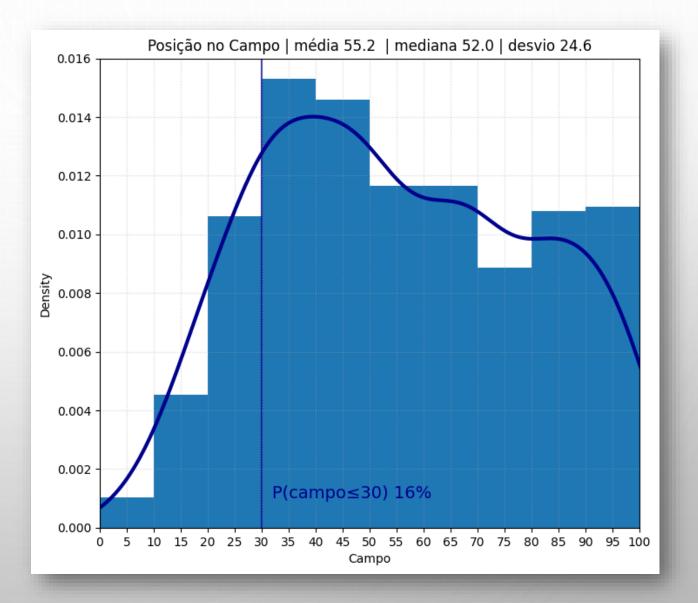


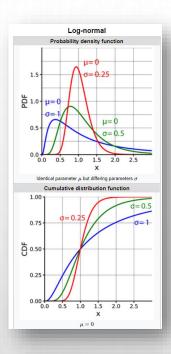
PROBABILIDADE DE UM INTERVALO DISTRIBUIÇÃO CONTÍNUA EM X VISUALIZAÇÃO HISTOGRAMA / DENSIDADE TESTE DE MÉDIAS

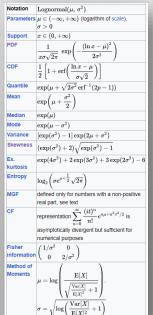


Kernel density estimation of 100 <u>normally</u> <u>distributed random numbers</u> using different smoothing bandwidths [Kernel Density Estimation, Wikipedia]

VARIÁVEL CONTÍNUA







VARIÁVEL CONTÍNUA - PARAMETRIZAÇÃO

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

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

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

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

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

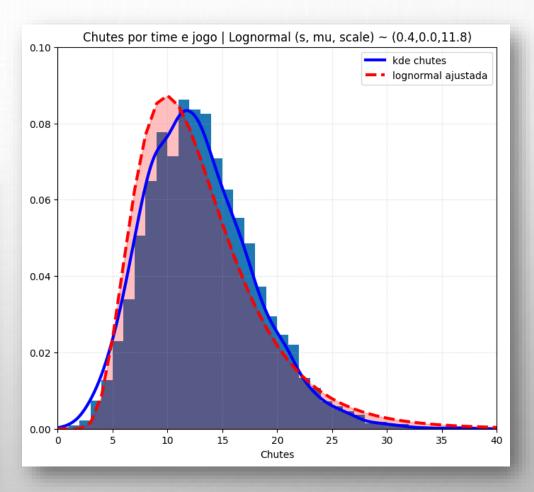
Since the first term is constant with regard to μ and σ , both logarithmic likelihood functions, ℓ and ℓ_N , reach their maximum with the same μ and σ . Hence, the maximum likelihood estimators are identical to those for a normal distribution for the observations $\ln x_1, \ln x_2, \ldots, \ln x_n$),

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

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 :

$$\mu = \ln\!\left(rac{ar{x}}{\sqrt{1+\widehat{\sigma}^2/ar{x}^2}}
ight), \qquad \sigma^2 = \ln\!\left(1+\widehat{\sigma}^2/ar{x}^2
ight)$$





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