



TÓPICOS EM CIÊNCIA DE
DADOS PARA O ESPORTE

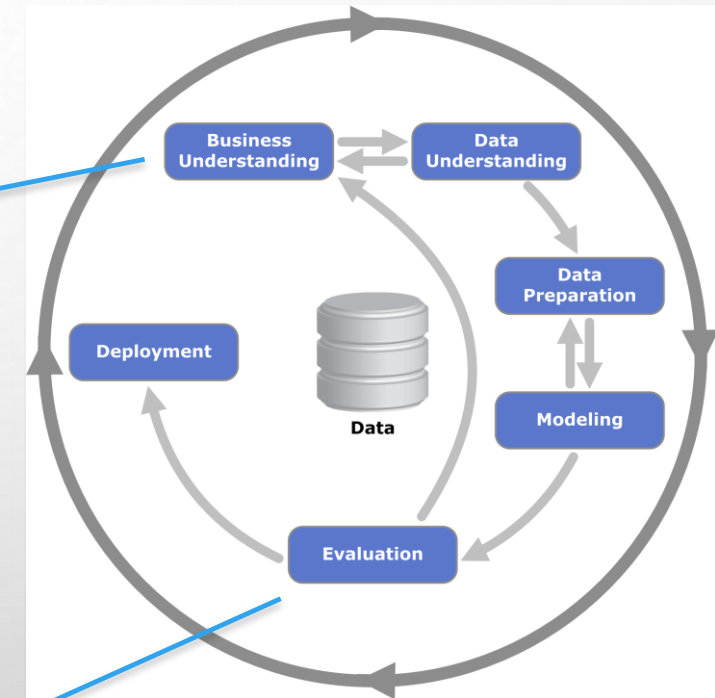
VARIÁVEIS ALEATÓRIAS

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	Workshop	Apresentações de Trabalhos I	Apresentação
21		Apresentações de Trabalhos II	

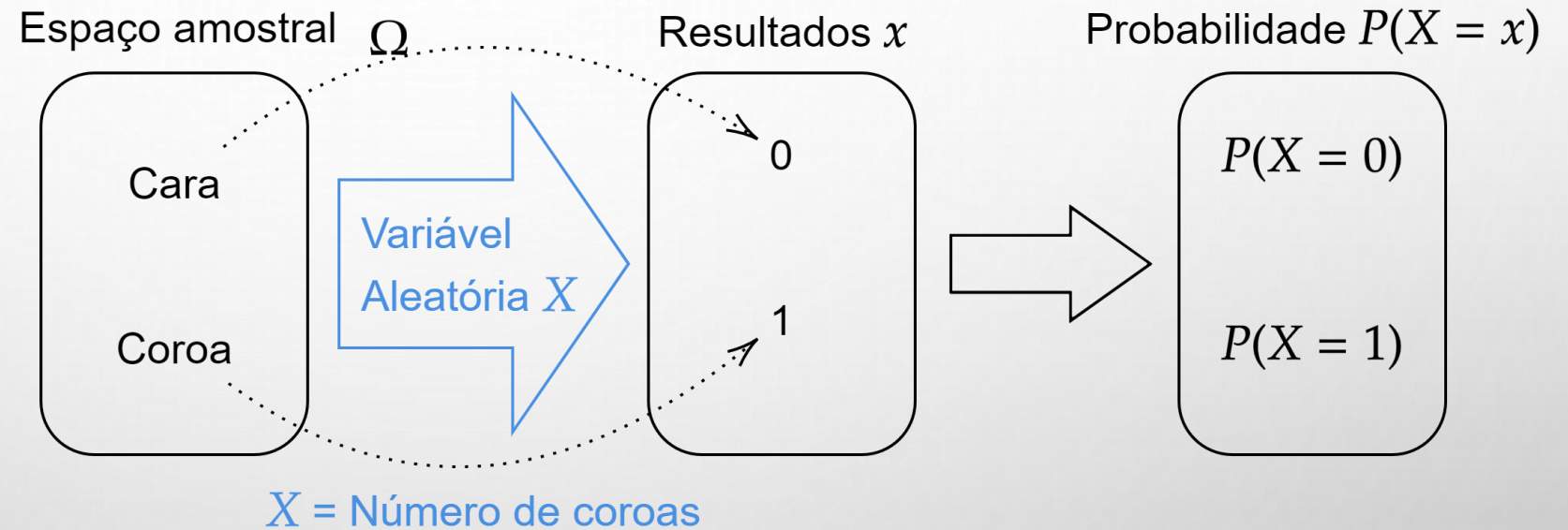


CRISP-DM

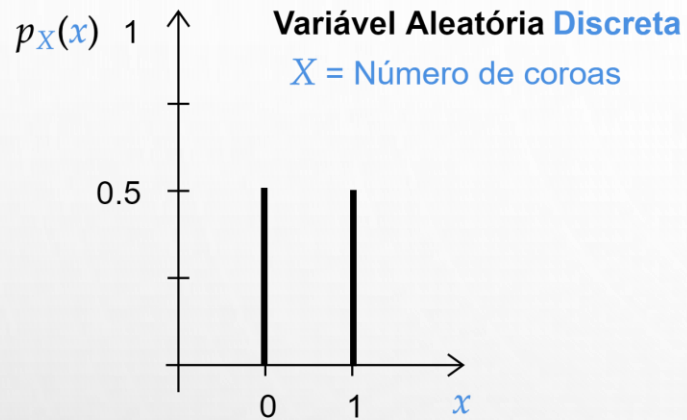
AGENDA

- PARTE 1 : TEORIA
 - VARIÁVEL ALEATÓRIA
 - DISTRIBUIÇÃO DE PROBABILIDADE
 - FUNÇÃO DENSIDADE DE PROBABILIDADE
 - ESTATÍSTICAS DE UMA DISTRIBUIÇÃO
 - DISTRIBUIÇÃO DE BERNOULLI E ESTIMATIVA DE PARÂMETROS
 - DISTRIBUIÇÃO UNIFORME DISCRETA
 - DISTRIBUIÇÃO CATEGÓRICA
 - DISTRIBUIÇÃO BINOMIAL
 - TEOREMA CENTRAL DO LIMITE
 - DISTRIBUIÇÃO NORMAL (GAUSSIANA)
- PARTE 2 : PRÁTICA
 - PANDAS + MATPLOTLIB → VARIÁVEIS ALEATÓRIAS NO ESPORTE

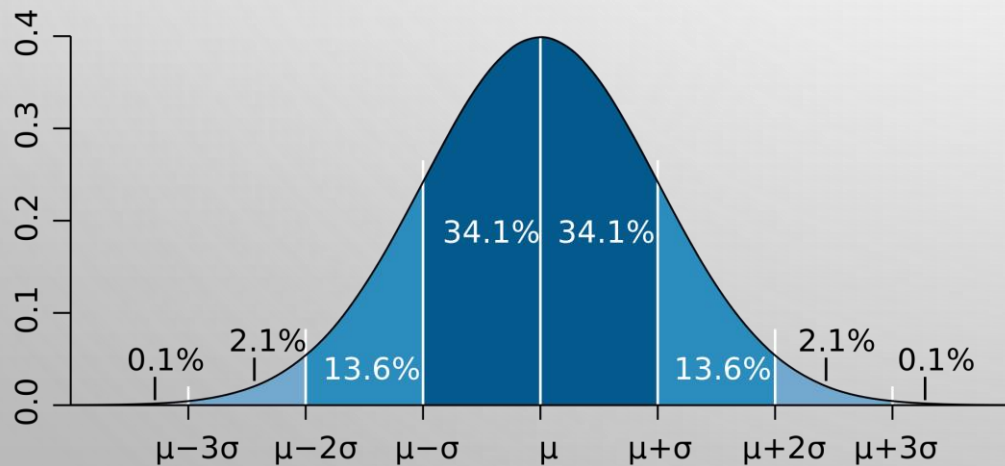
VARIÁVEL ALEATÓRIA



DISTRIBUIÇÃO DE PROBABILIDADE



Zibetti [1]

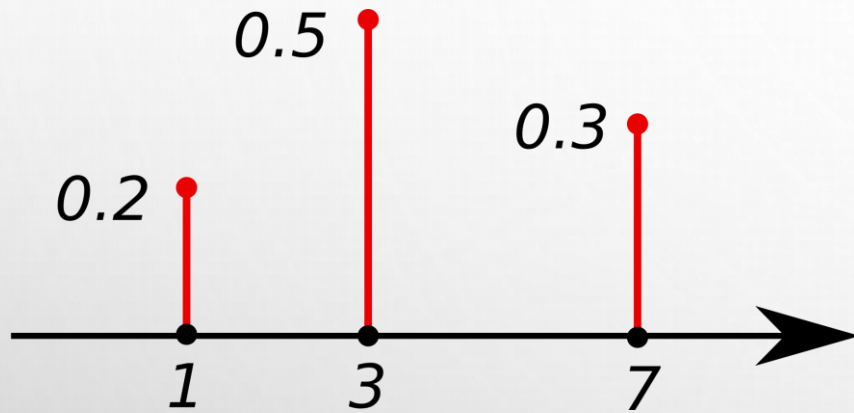


Wikipedia [2]



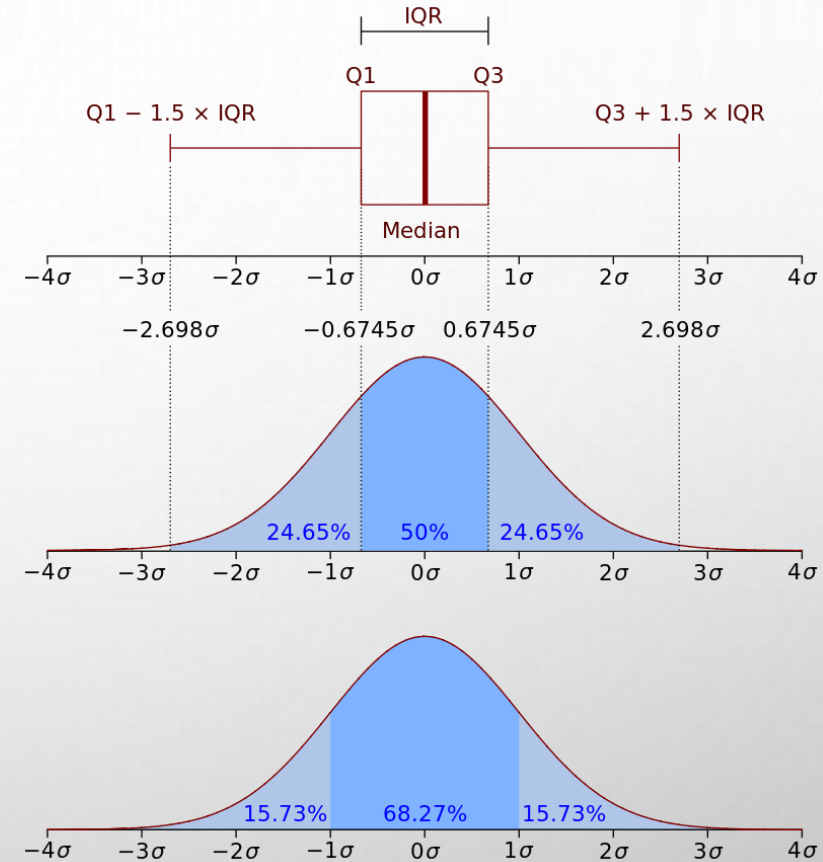
Distribuição teórica, paramétrica, e a realização experimental, não paramétrica.

FUNÇÃO DENSIDADE DE PROBABILIDADE



Função Massa de Probabilidade (Discreta)

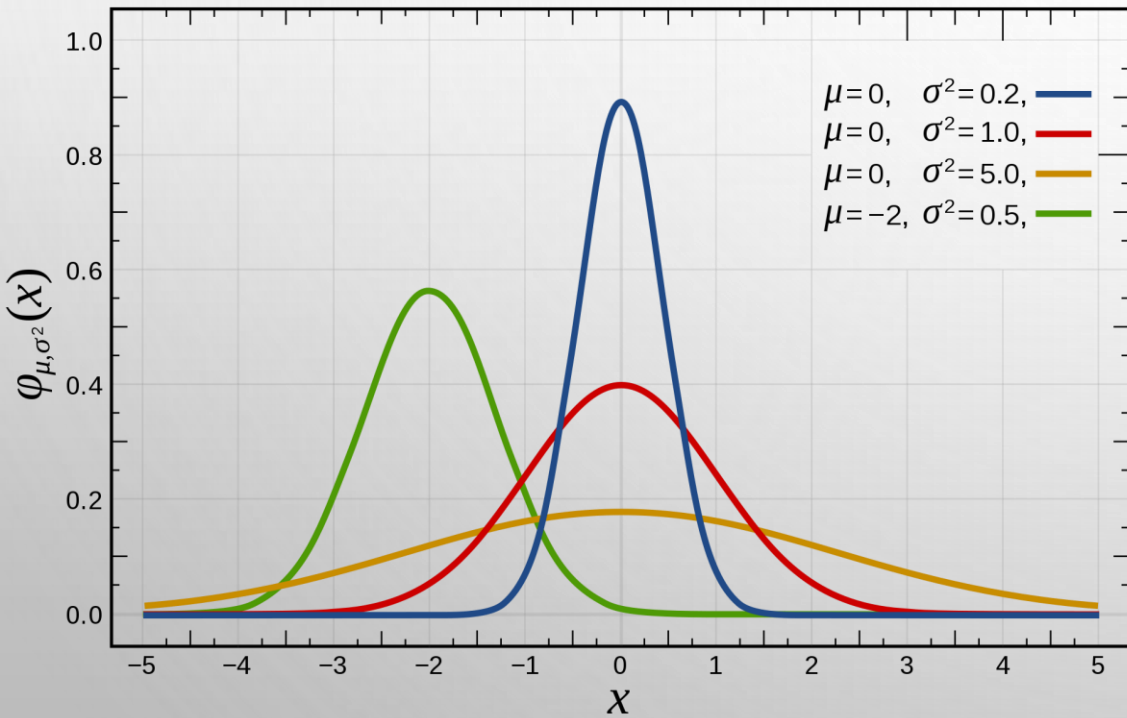
[2]



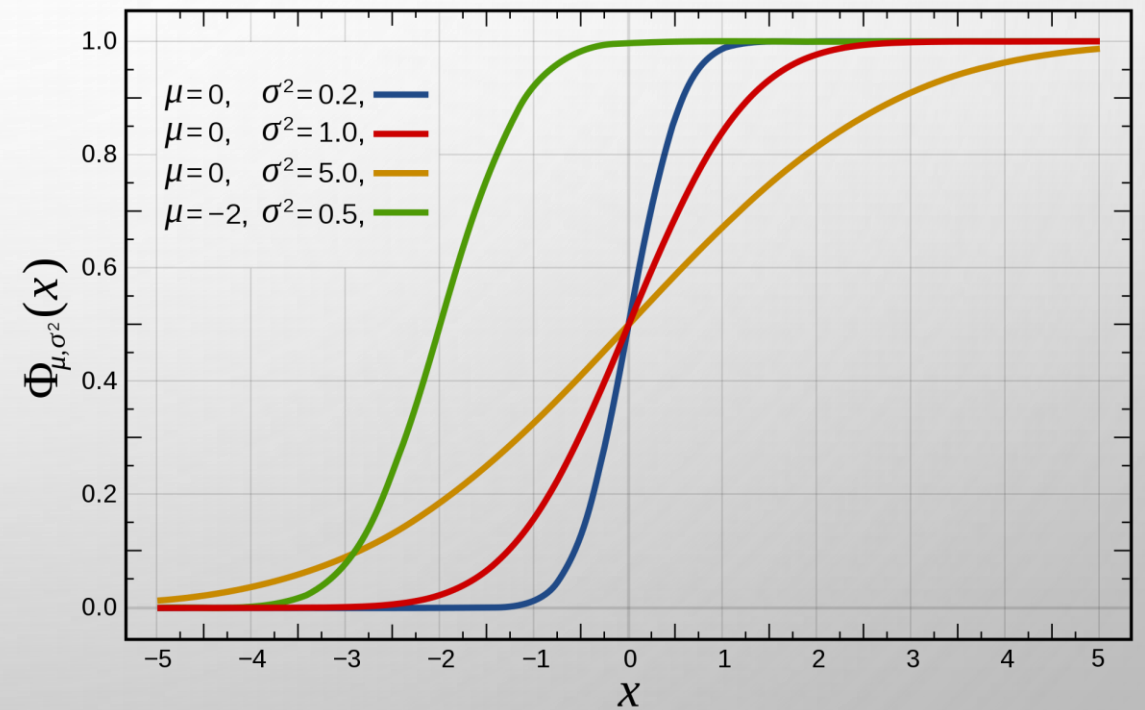
Função Densidade de Probabilidade (Contínua)

[2]

FUNÇÃO DISTRIBUIÇÃO DE PROBABILIDADE ACUMULADA

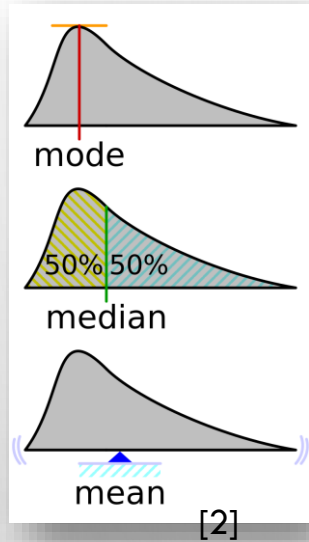


$\int \varphi$
 \Rightarrow



$$F(x) = P(X \leq x)$$

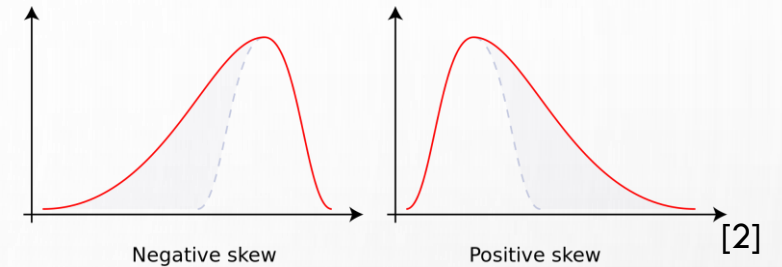
ESTATÍSTICAS DE UMA DISTRIBUIÇÃO



$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

MÉDIA – O MOMENTO CENTRAL

$$\begin{aligned} \tilde{\mu}_3 &= E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \\ &= \frac{E[X^3] - 3\mu E[X^2] + 3\mu^2 E[X] - \mu^3}{\sigma^3} \\ &= \frac{E[X^3] - 3\mu(E[X^2] - \mu E[X]) - \mu^3}{\sigma^3} \\ &= \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}. \end{aligned}$$

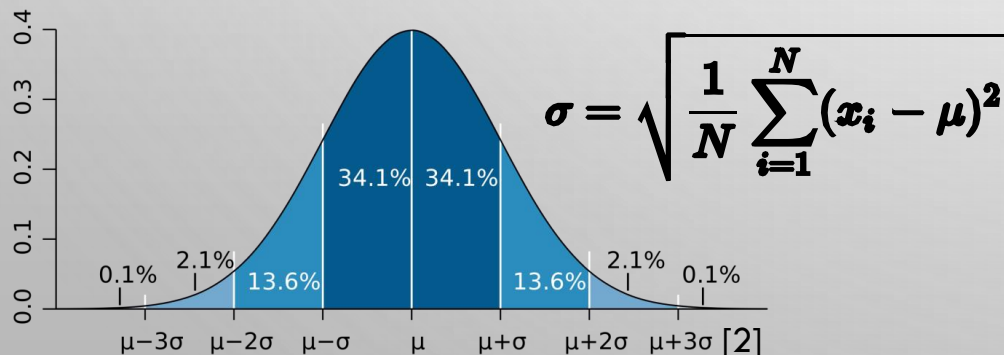


ASSIMETRIA - DESEQUILIBRIO

CURTOSE - HOMOGENEIDADE

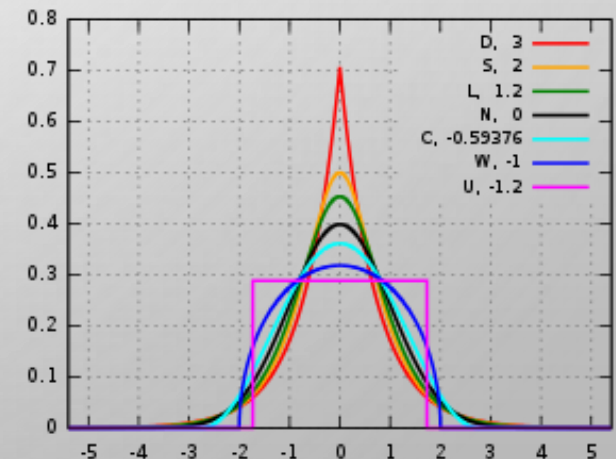
[2]

O DESVIO PADRÃO – DISPERSÃO DOS DADOS

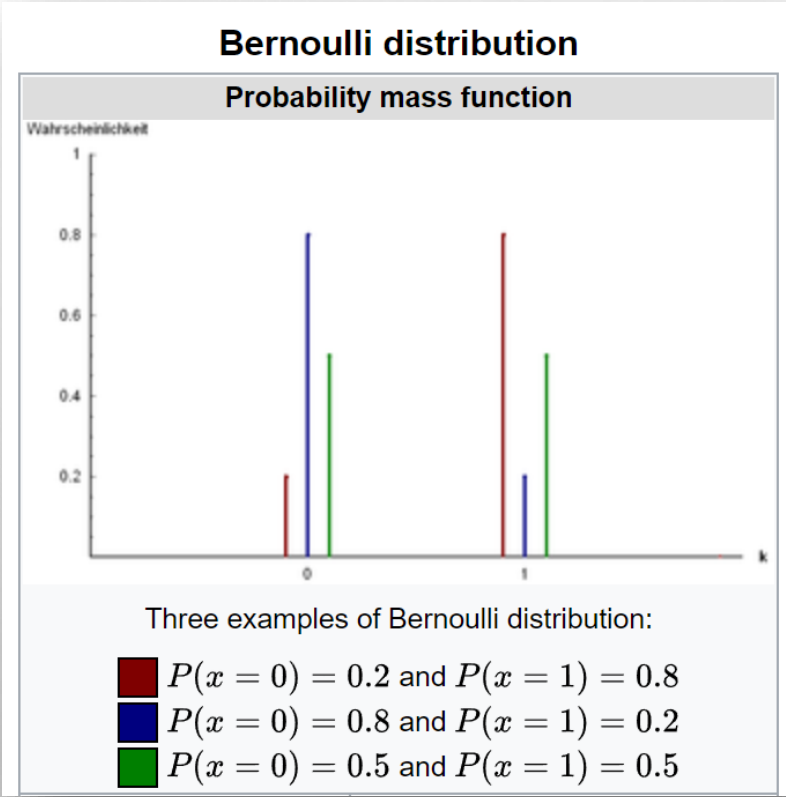


$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right]^2} - 3$$



DISTRIBUIÇÃO DE BERNOULLI E ESTIMATIVA DE PARÂMETROS

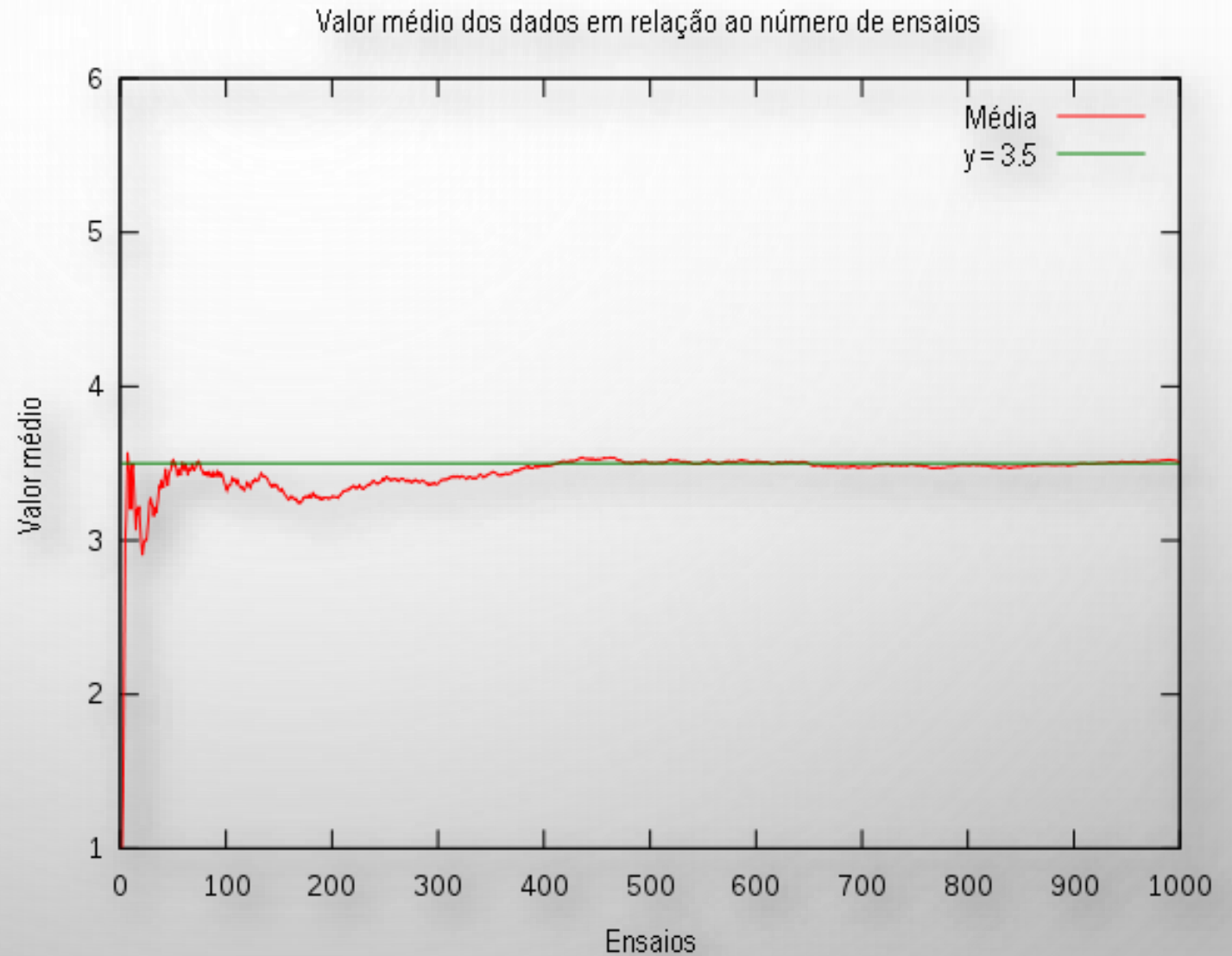


[2]

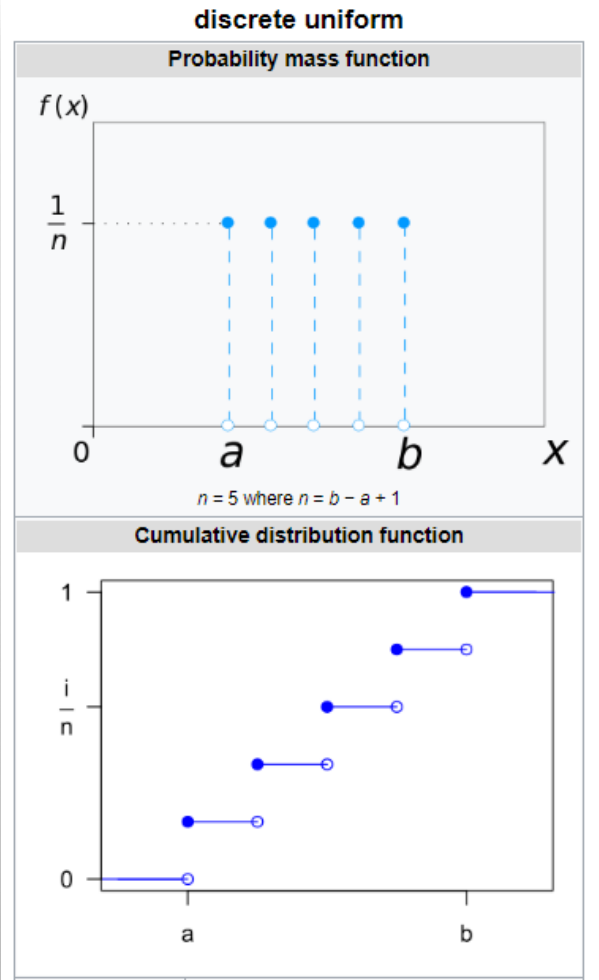
Parameters	$0 \leq p \leq 1$ $q = 1 - p$
Support	$k \in \{0, 1\}$
PMF	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if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

[2]

LEI DOS GRANDES NÚMEROS



DISTRIBUIÇÃO DE MASSA UNIFORME

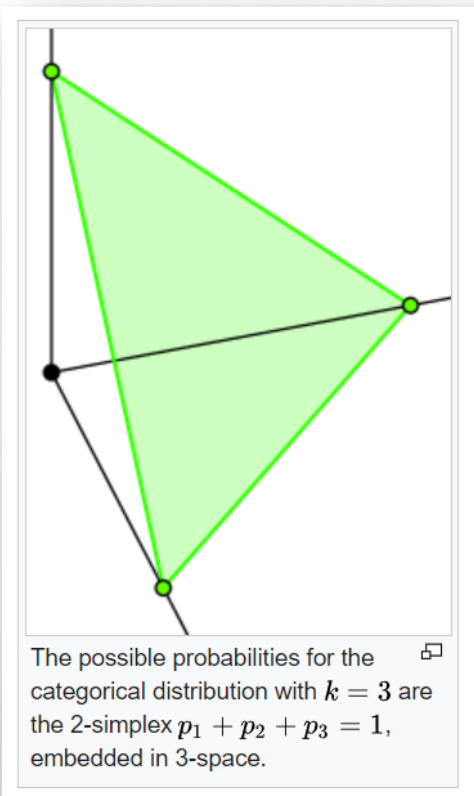
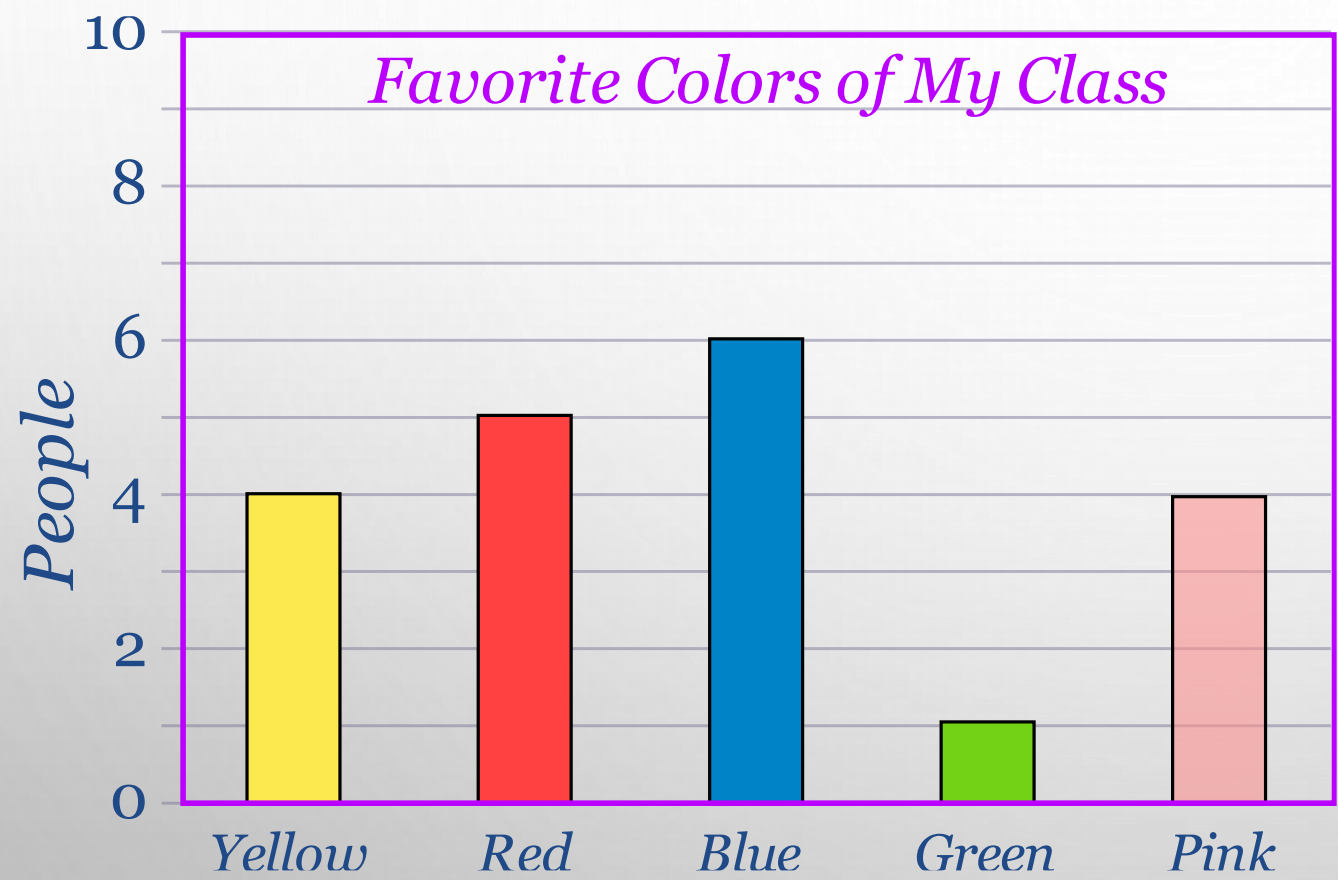


[2]

Notation	$\mathcal{U}\{a, b\}$ or $\text{unif}\{a, b\}$
Parameters	a, b integers with $b \geq a$ $n = b - a + 1$
Support	$k \in \{a, a + 1, \dots, b - 1, b\}$
PMF	$\frac{1}{n}$
CDF	$\frac{\lfloor k \rfloor - a + 1}{n}$
Mean	$\frac{a + b}{2}$
Median	$\frac{a + b}{2}$
Mode	N/A
Variance	$\frac{n^2 - 1}{12}$
Skewness	0
Ex. kurtosis	$-\frac{6(n^2 + 1)}{5(n^2 - 1)}$
Entropy	$\ln(n)$
MGF	$\frac{e^{at} - e^{(b+1)t}}{n(1 - e^t)}$
CF	$\frac{e^{iat} - e^{i(b+1)t}}{n(1 - e^{it})}$
PGF	$\frac{z^a - z^{b+1}}{n(1 - z)}$

[2]

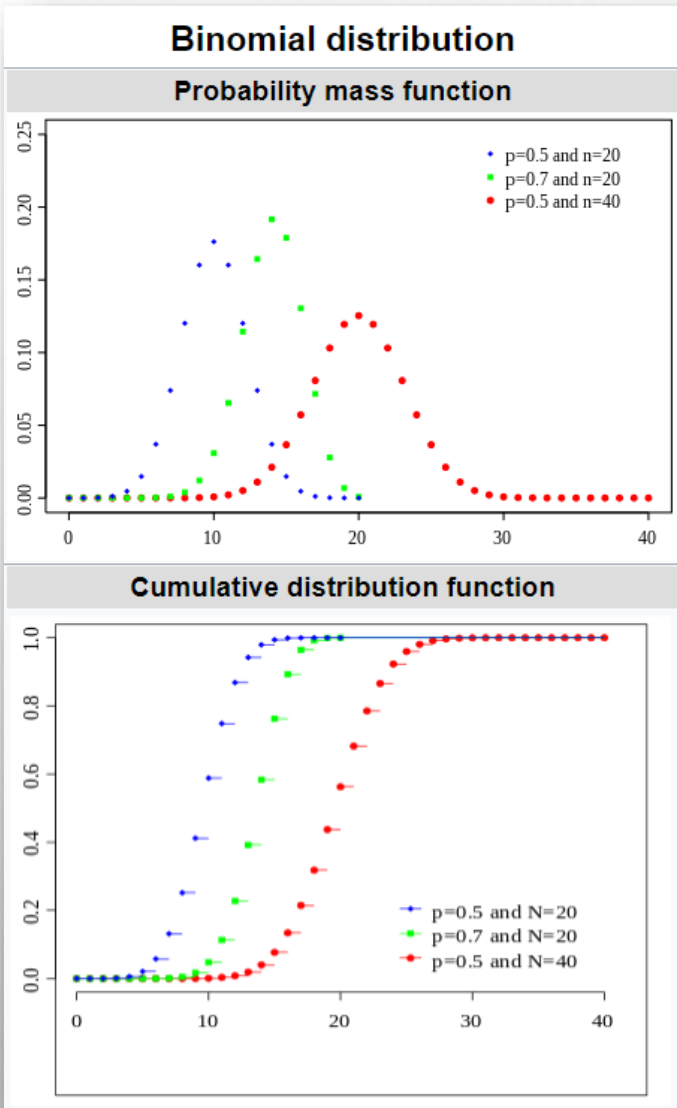
DISTRIBUIÇÃO CATEGÓRICA



[2]

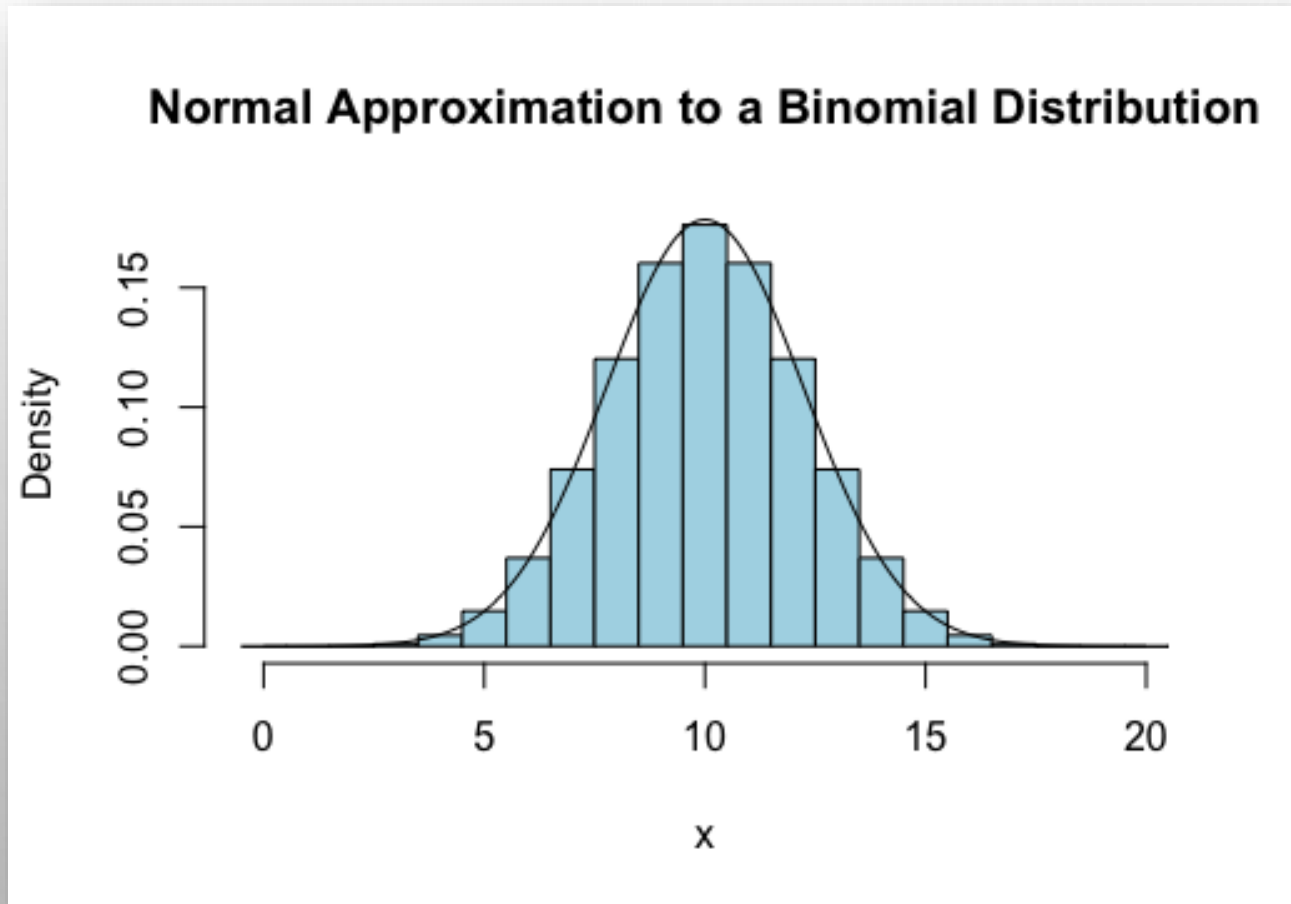
Categorical	
Parameters	$k > 0$ number of categories (<i>integer</i>) p_1, \dots, p_k event probabilities ($p_i \geq 0, \Sigma p_i = 1$)
Support	$x \in \{1, \dots, k\}$
PMF	(1) $p(x = i) = p_i$ (2) $p(x) = p_1^{[x=1]} \dots p_k^{[x=k]}$ (3) $p(x) = [x = 1] \cdot p_1 + \dots + [x = k] \cdot p_k$ where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

DISTRIBUIÇÃO BINOMIAL



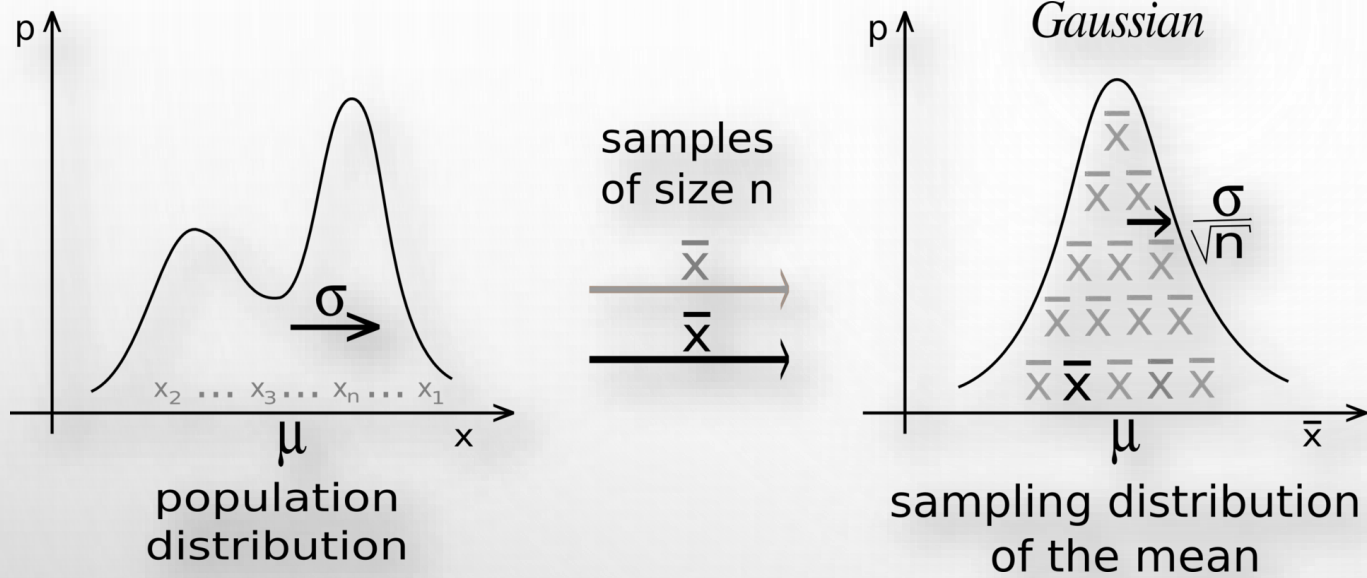
Notation	$B(n, p)$
Parameters	$n \in \{0, 1, 2, \dots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial $q = 1 - p$
Support	$k \in \{0, 1, \dots, n\}$ – number of successes
PMF	$\binom{n}{k} p^k q^{n-k}$
CDF	$I_q(n - k, 1 + k)$ (the regularized incomplete beta function)
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n + 1)p \rfloor$ or $\lceil (n + 1)p \rceil - 1$
Variance	npq
Skewness	$\frac{q - p}{\sqrt{npq}}$
Ex. kurtosis	$\frac{1 - 6pq}{npq}$
Entropy	$\frac{1}{2} \log_2(2\pi enpq) + O\left(\frac{1}{n}\right)$ in shannons . For nats , use the natural log in the log .
MGF	$(q + pe^t)^n$
CF	$(q + pe^{it})^n$
PGF	$G(z) = [q + pz]^n$
Fisher information	$g_n(p) = \frac{n}{pq}$ (for fixed n)

DISTRIBUIÇÃO BINOMIAL



<https://anydice.com/>

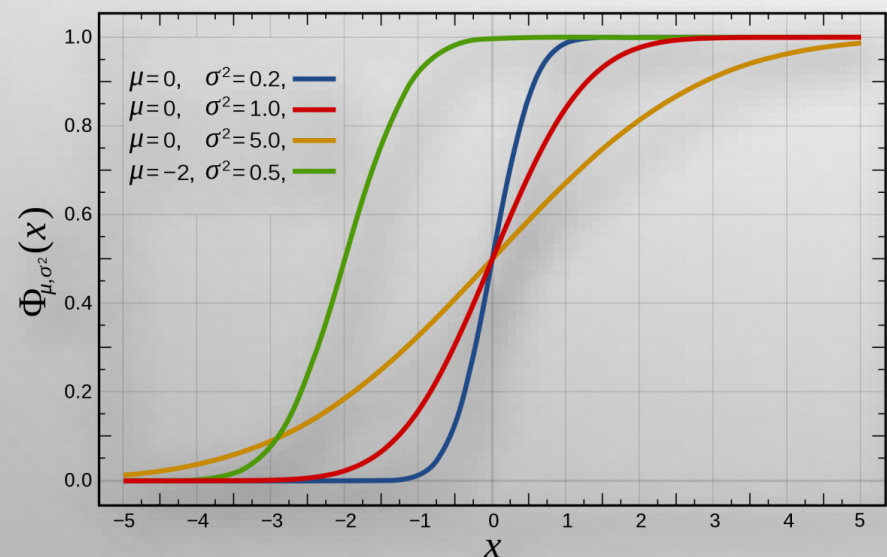
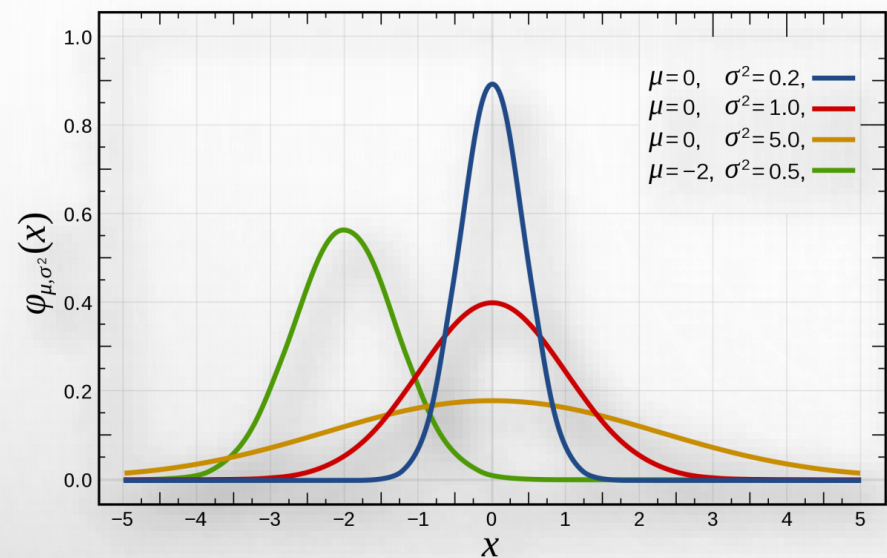
TEOREMA CENTRAL DO LIMITE



Lindeberg–Lévy CLT — Suppose $\{X_1, \dots, X_n\}$ is a sequence of **i.i.d.** random variables with $E[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2 < \infty$. Then, as n approaches infinity, the random variables $\sqrt{n}(\bar{X}_n - \mu)$ **converge in distribution** to a **normal** $\mathcal{N}(0, \sigma^2)$.^[4]

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

DISTRIBUIÇÃO NORMAL (GAUSSIANA)



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 \in \mathbb{R}_{>0}$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF	$\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \right]$
Quantile	$\mu + \sigma\sqrt{2} \operatorname{erf}^{-1}(2p - 1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
MAD	$\sigma\sqrt{2} \operatorname{erf}^{-1}(1/2)$
Skewness	0
Ex. kurtosis	0
Entropy	$\frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2}$
MGF	$\exp(\mu t + \sigma^2 t^2 / 2)$
CF	$\exp(i\mu t - \sigma^2 t^2 / 2)$
Fisher information	$\mathcal{I}(\mu, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$ $\mathcal{I}(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$
Kullback–Leibler divergence	$\frac{1}{2} \left\{ \left(\frac{\sigma_0}{\sigma_1} \right)^2 + \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2} - 1 + \ln \frac{\sigma_1^2}{\sigma_0^2} \right\}$

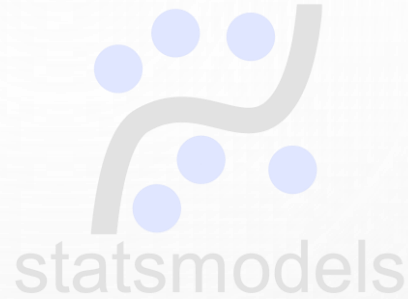
SETUP INICIAL DO AMBIENTE PYTHON



4. Variáveis
Aleatórias



5. Visualização



1. Editor de Código



2. Gestor de Ambiente



3. Ambiente
Python do Projeto



3. Notebook
Dinâmico

PRÓXIMA AULA LEITURA: ANÁLISE EXPLORATÓRIA (LIVRO OU ARTIGOS)

