escola britânica de artes criativas & tecnologia

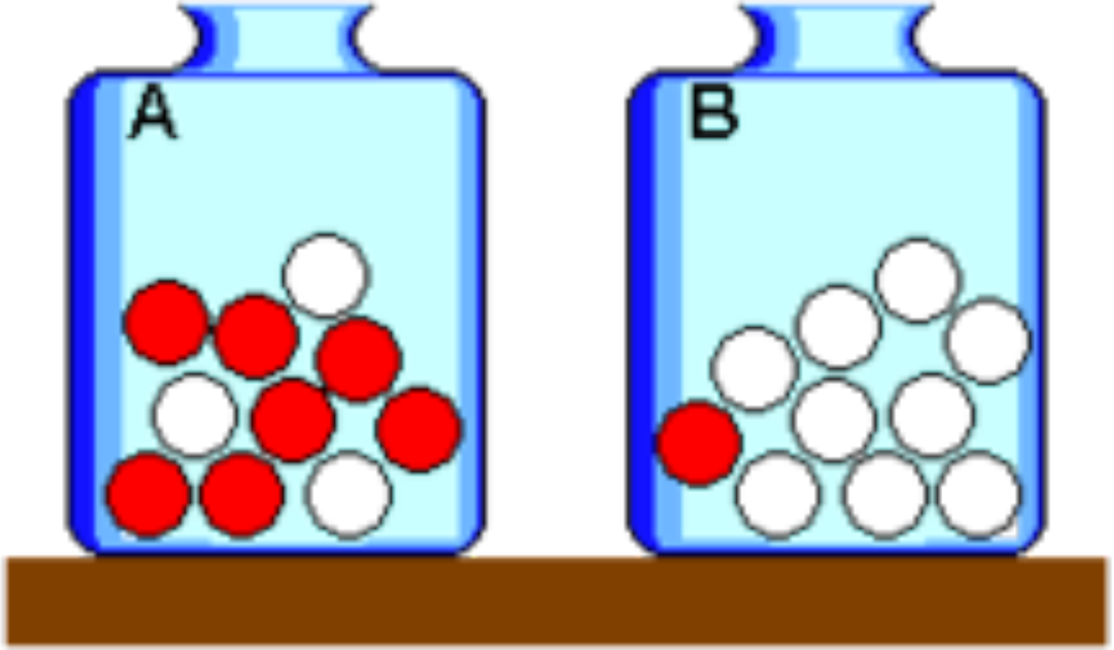
Profissão: Cientista de Dados

Probabilidade

Introdução:

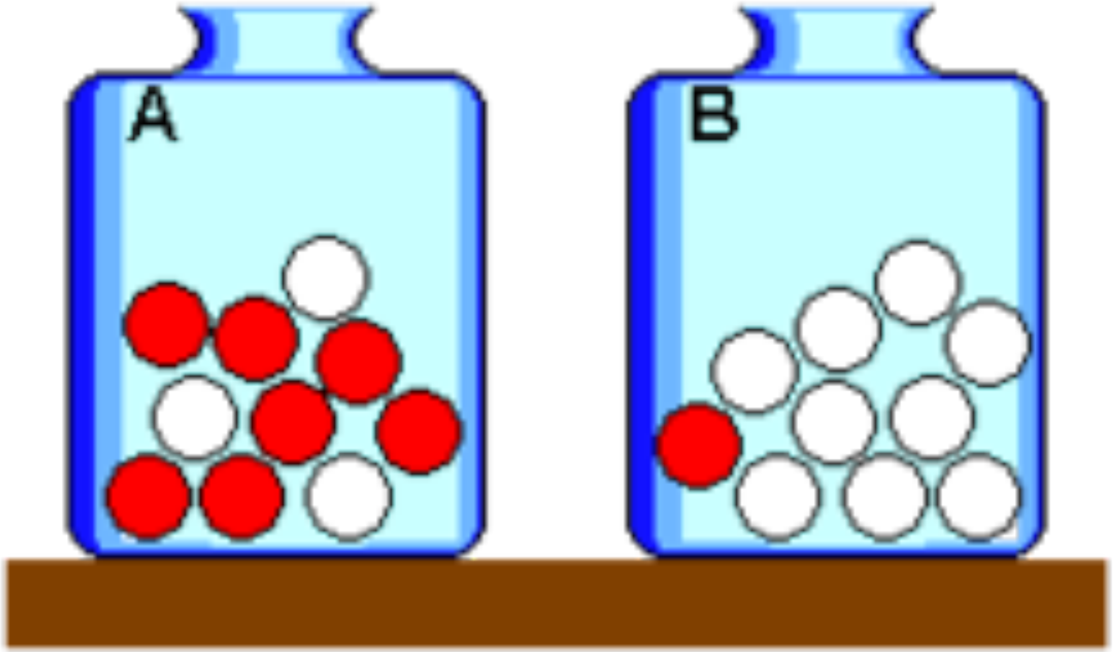
Variáveis Aleatórias e Amostragem

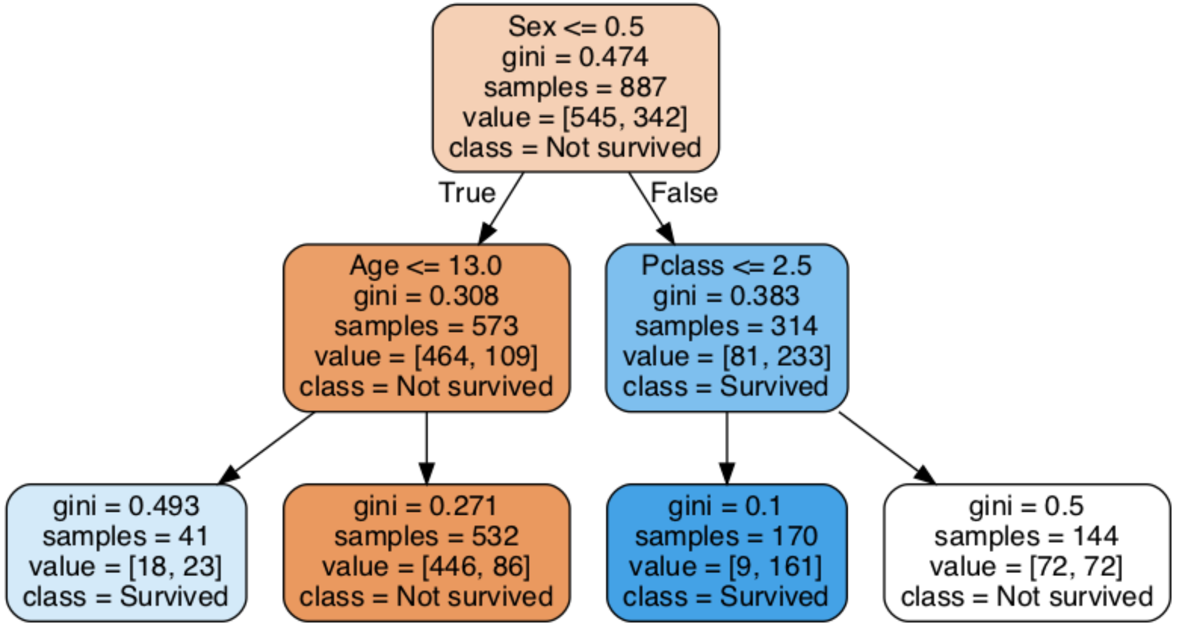
Probabilidade



https://upload.wikimedia.org/wikipedia/commons/5/5b/Stochastik_Bayestheorem_Urnenversuch.png

Probabilidade

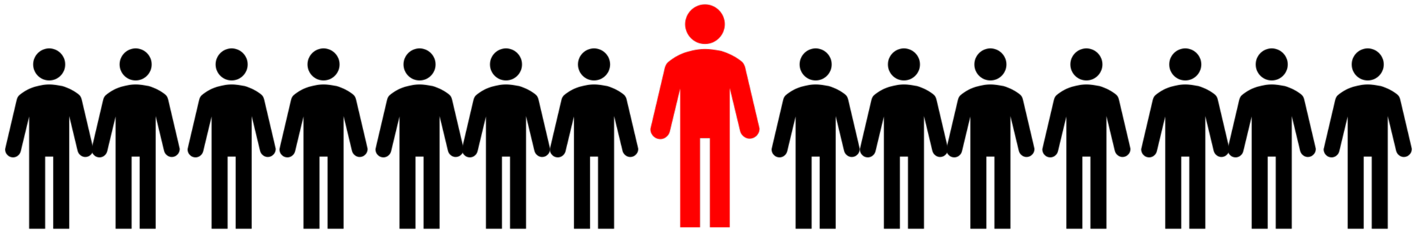


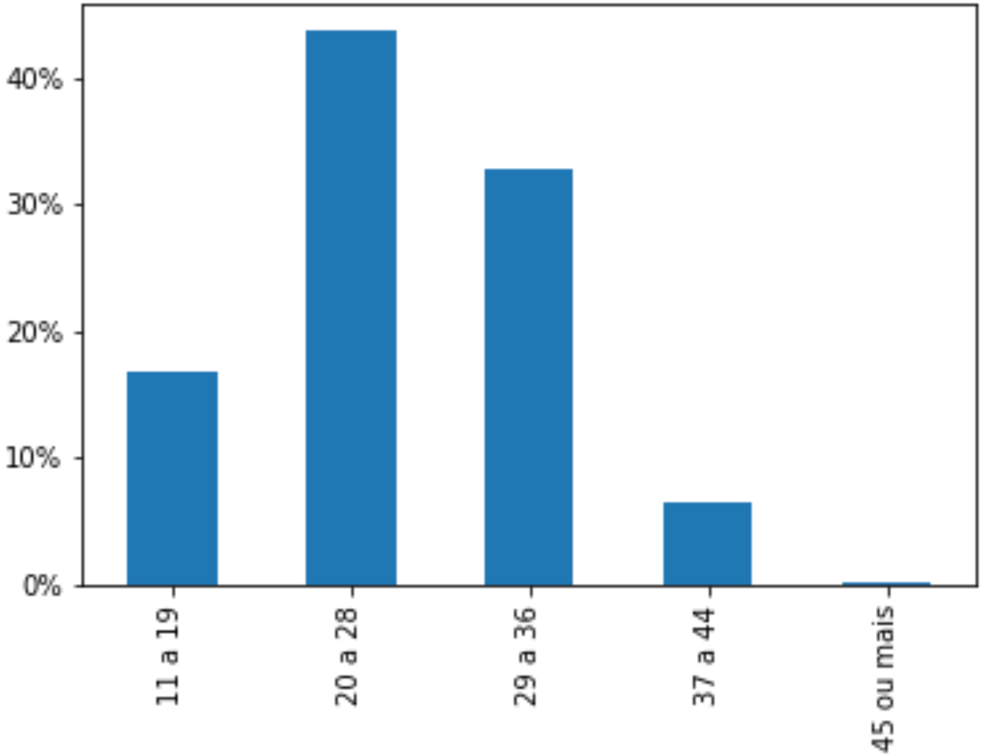


https://upload.wikimedia.org/wikipedia/commons/5/5b/Stochastik_Bayestheorem_Urnenversuch.png

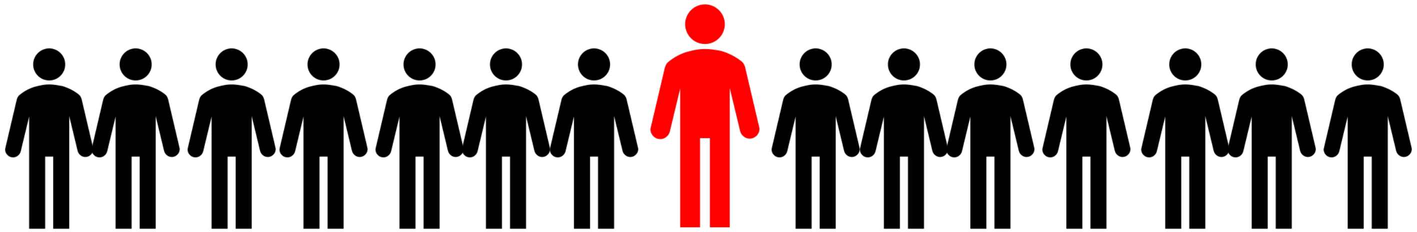
Amostragem tipo *cross section*

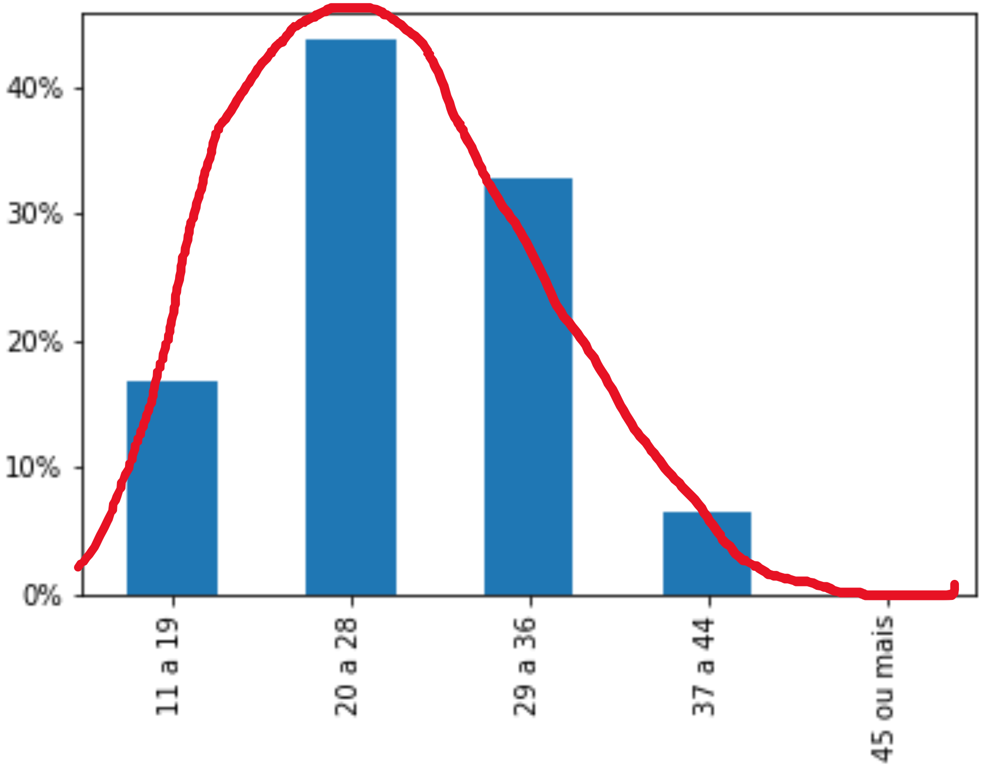
Probabilidade





Probabilidade





Amostragem de Processo Estocástico

Modelos discretos

Modelos discretos - Bernoulli

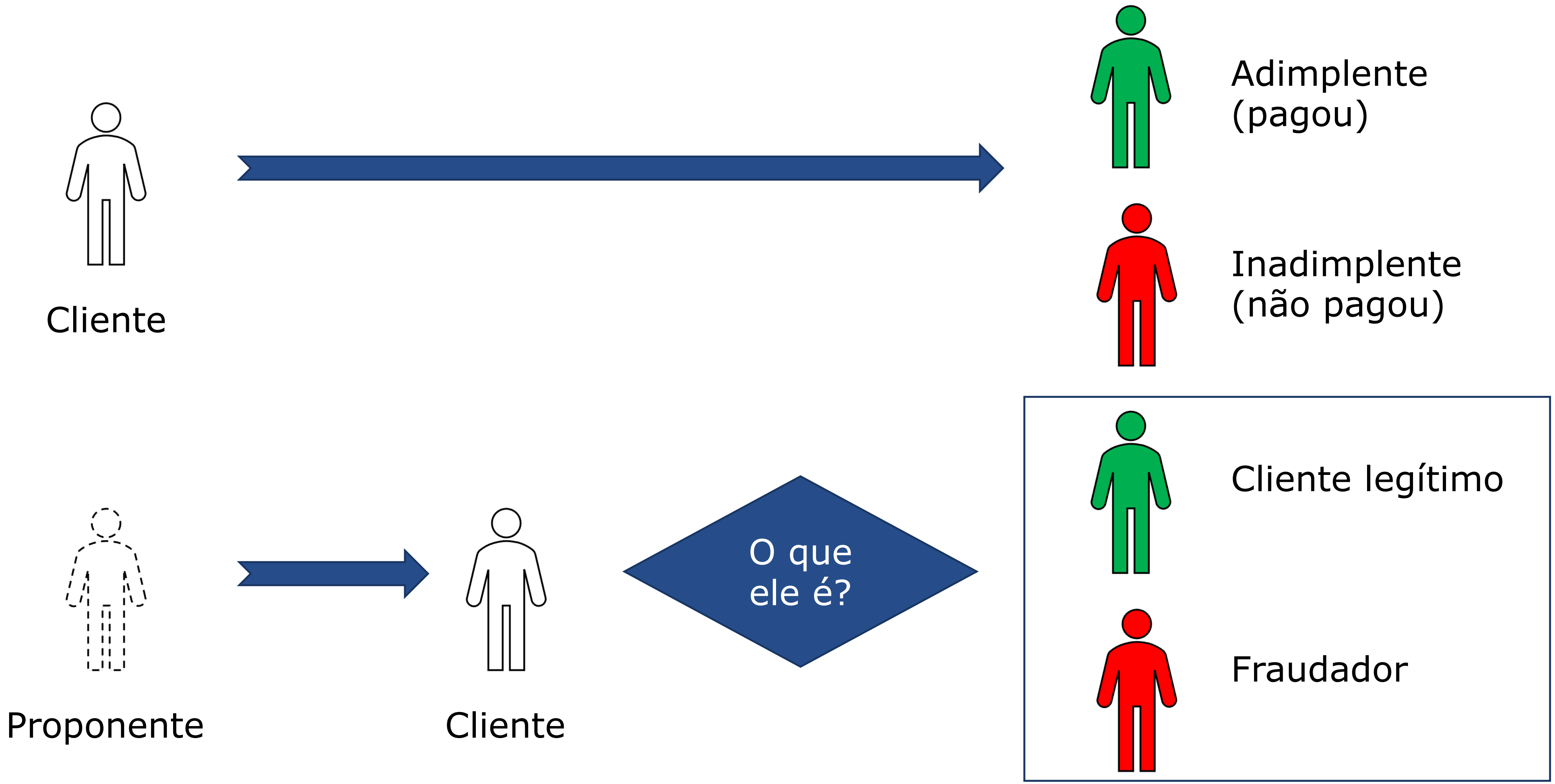
$$X = \begin{cases} 1 \; se \; o \; resultado \; do \; ensaio \text{ é } “sucesso” \\ 0 \; se \; o \; resultado \; do \; ensaio \text{ é } “fracasso” \end{cases}$$

Definimos: “cara” = “sucesso”, ”coroa” = “fracasso”

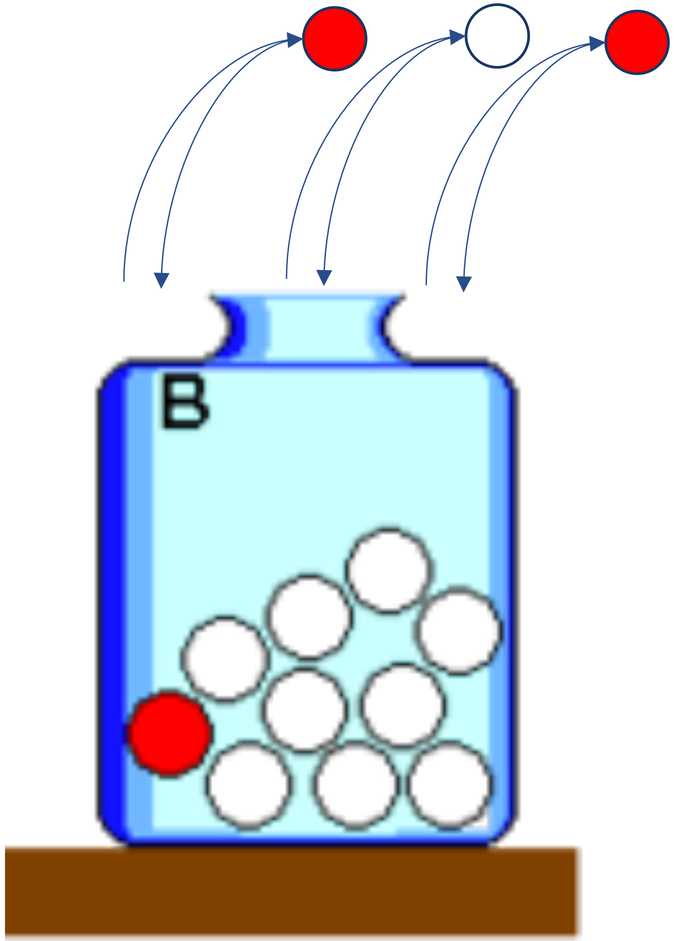
$$P(X = 1) = 0{,}5 \qquad P(X = 1) = p$$

$$P(X = 0) = 0{,}5 \qquad P(X = 0) = 1 - p$$

Modelos discretos - Bernoulli

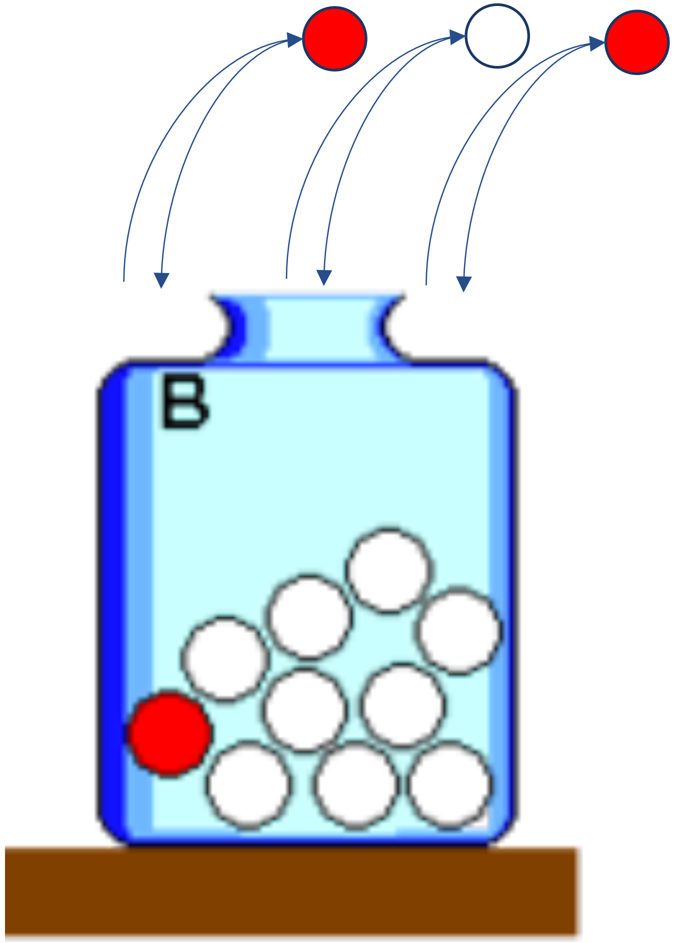


Modelos discretos - Binomial

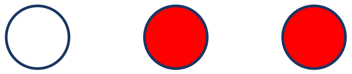


Se retirarmos N bolas, com reposição, qual a probabilidade de retirarmos x bolas brancas?

Modelos discretos - Binomial



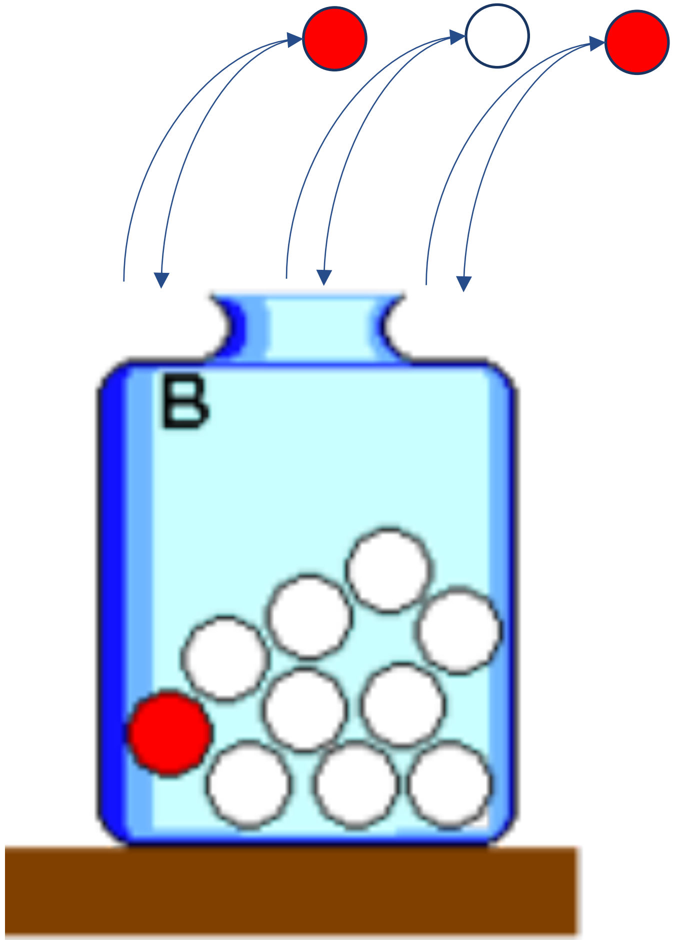
Qual a probabilidade de retirarmos 1 bola branca?



$$p * (1 - p) * (1 - p)$$

$$= p(1 - p)^2$$

Modelos discretos - Binomial



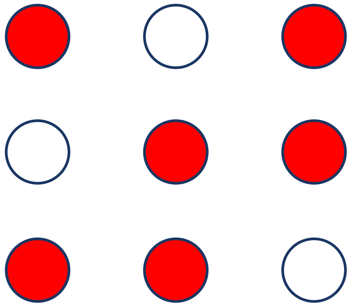
Qual a probabilidade de retirarmos 1 bola branca?

OBS: Considerando a ordem, podemos fazer isso de 3 formas diferentes:

Forma 1:

Forma 2:

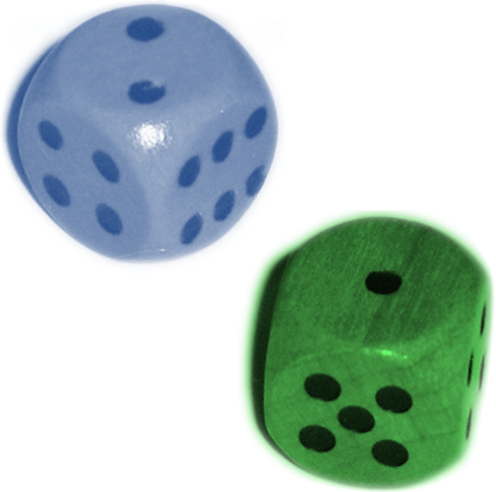
Forma 3:



$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Distribuição condicional e independência

Conjunta



X_1: Resultado do dado 1

X_2: Resultado do dado 2

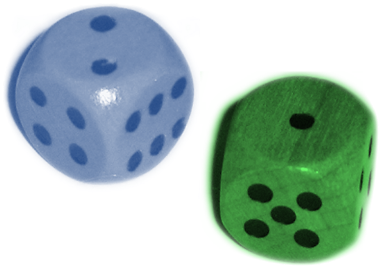
Y: Soma

$$\boxed{1} + \boxed{1} = \boxed{2}$$

Conjunta

1 + 1 = 2	2 + 1 = 3	3 + 1 = 4
1 + 2 = 3	2 + 2 = 4	3 + 2 = 5
1 + 3 = 4	2 + 3 = 5	3 + 3 = 6
1 + 4 = 5	2 + 4 = 6	3 + 4 = 7
1 + 5 = 6	2 + 5 = 7	3 + 5 = 8
1 + 6 = 7	2 + 6 = 8	3 + 6 = 9

4 + 1 = 5	5 + 1 = 6	6 + 1 = 7
4 + 2 = 6	5 + 2 = 7	6 + 2 = 8
4 + 3 = 7	5 + 3 = 8	6 + 3 = 9
4 + 4 = 8	5 + 4 = 9	6 + 4 = 10
4 + 5 = 9	5 + 5 = 10	6 + 5 = 11
4 + 6 = 10	5 + 6 = 11	6 + 6 = 12



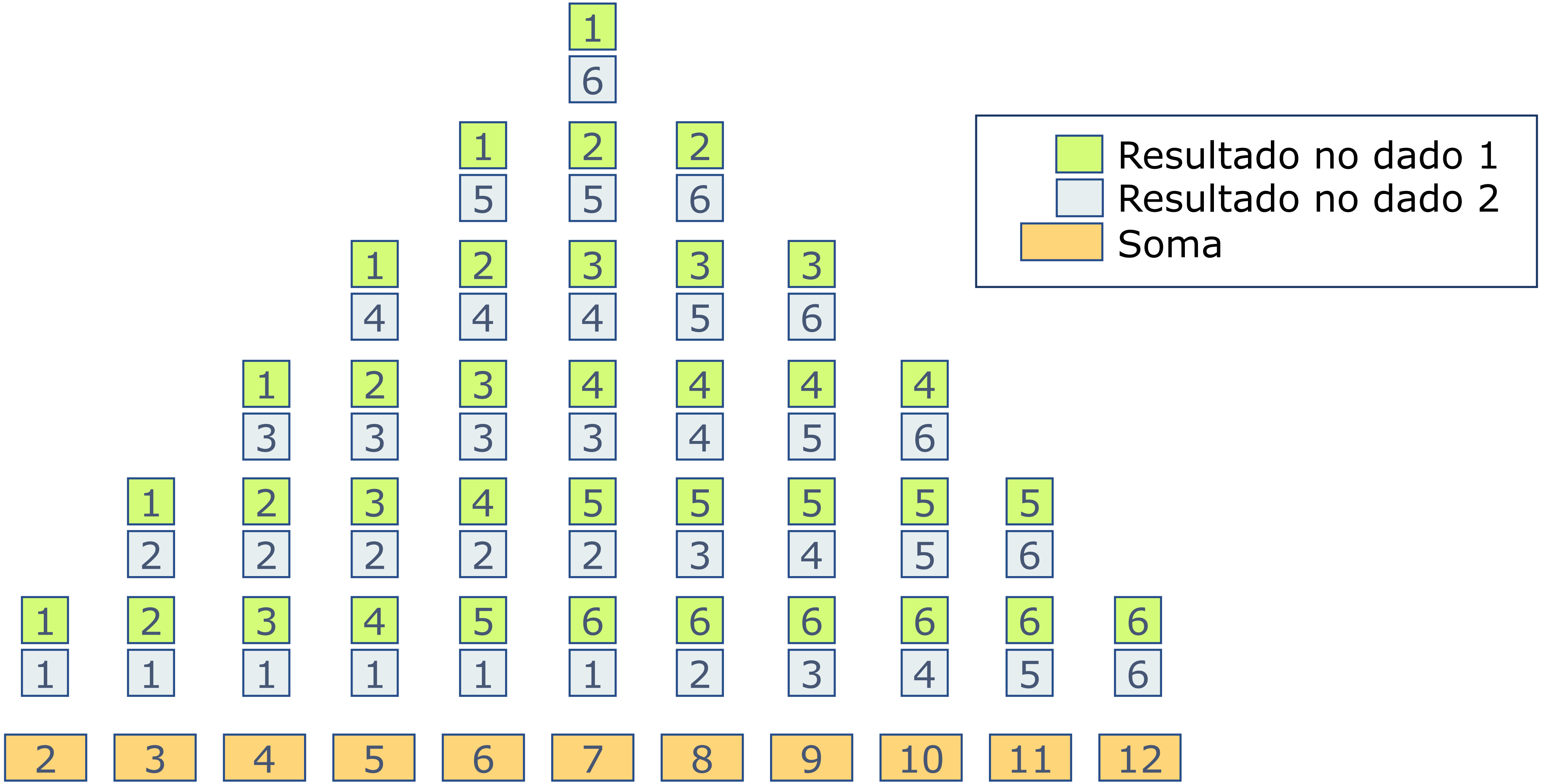
dado 1
dado 2
Soma

$$P(X_1 = i, X_2 = j)$$

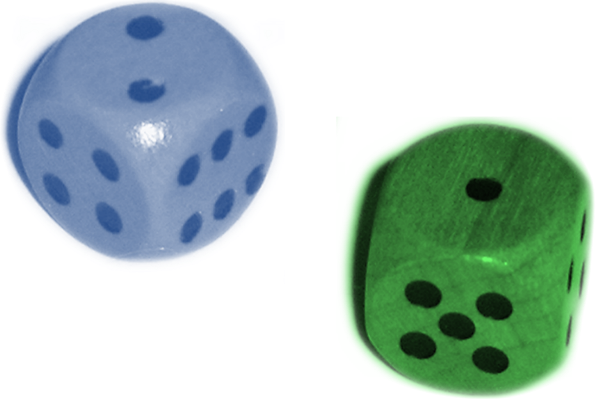
$$= P(X_1 = i).P(X_2 = j)$$

$$= \frac{1}{6}.\frac{1}{6} = \frac{1}{36}$$

Conjunta

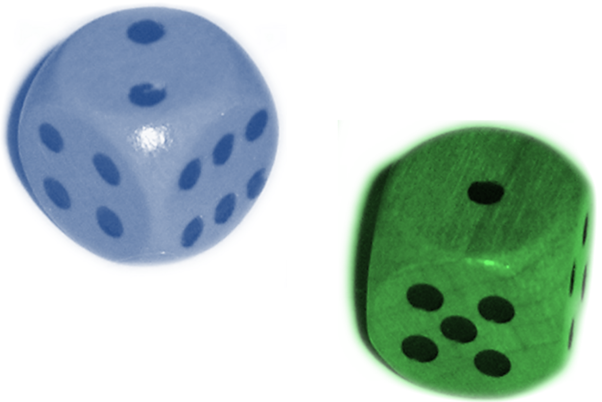

Resultado no dado 1
Resultado no dado 2
Soma
2
3
4
5
6
7
8
9
10
11
12

Independência



$$P(X_1 = i, | X_2 = j) = P(X_1 = i)$$

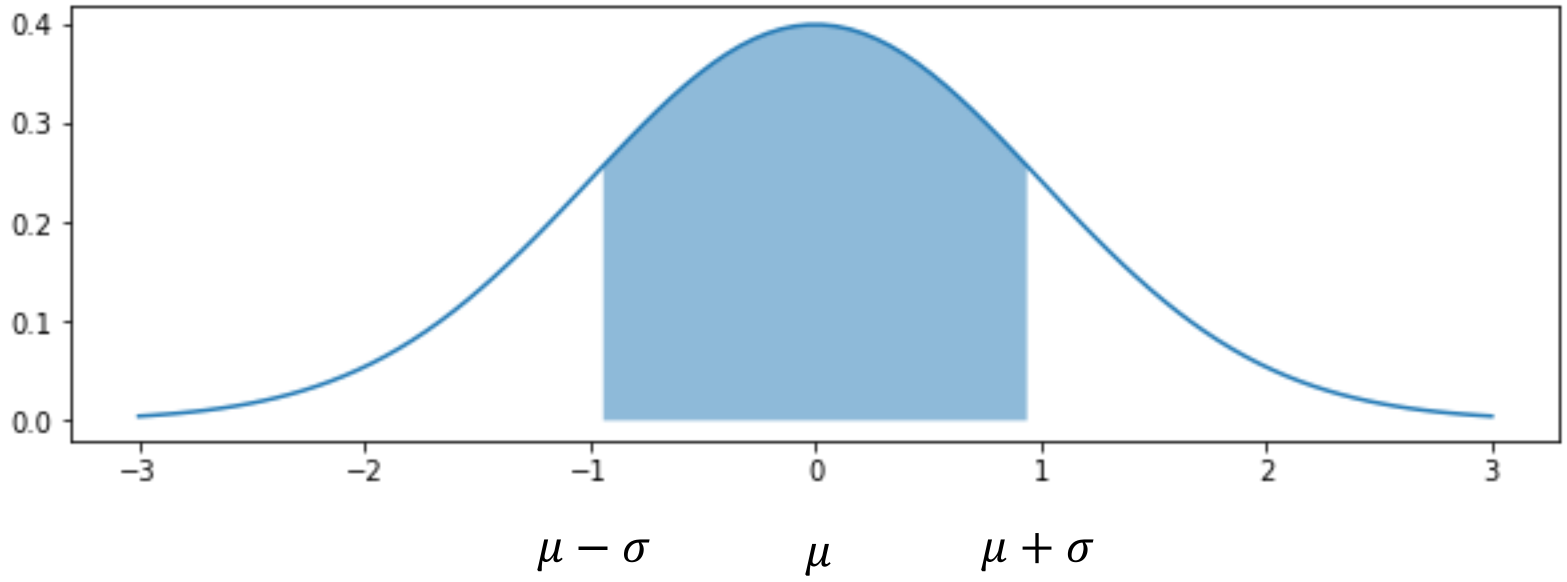
Condicional



$$P(Y = i|X_1 = 6) = \begin{cases} \frac{1}{6}\ se\ i = 7 \\ \frac{1}{6}\ se\ i = 8 \\ \frac{1}{6}\ se\ i = 9 \\ \frac{1}{6}\ se\ i = 10 \\ \frac{1}{6}\ se\ i = 11 \\ \frac{1}{6}\ se\ i = 12 \\ 0\ caso\ contrário \end{cases}$$

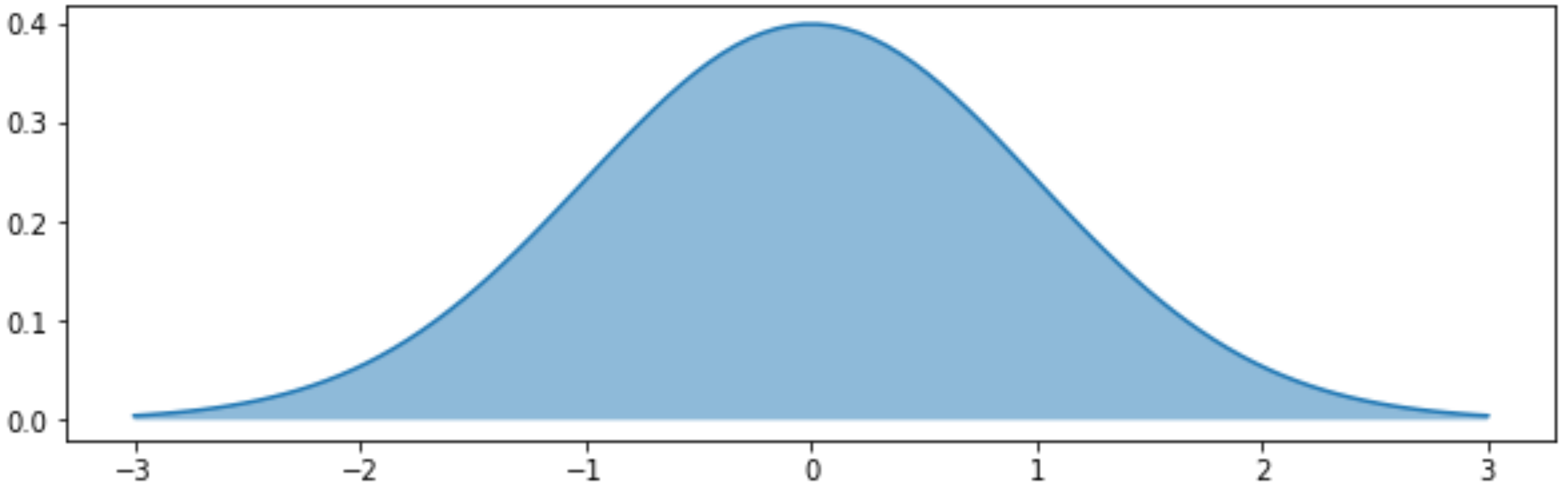
Função Densidade de Probabilidade

Função Densidade de Probabilidade



$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

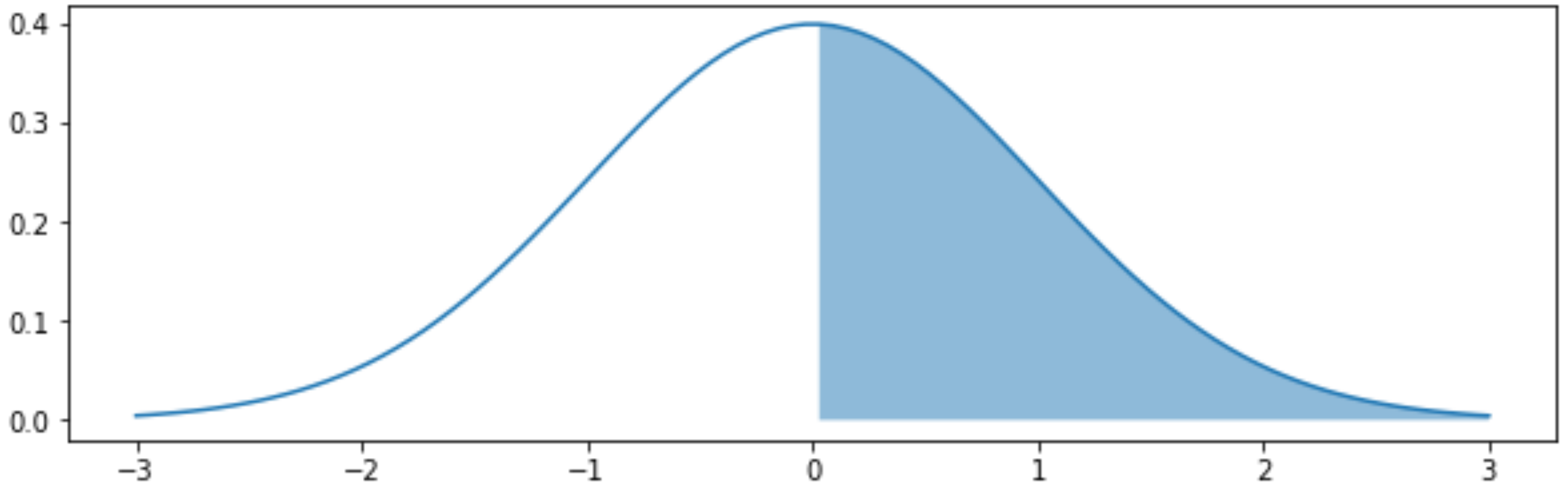
Função Densidade de Probabilidade



$\mu - \sigma$ $\qquad \mu \qquad$ $\mu + \sigma$

$$P(-\infty \leq X \leq +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$

Função Densidade de Probabilidade



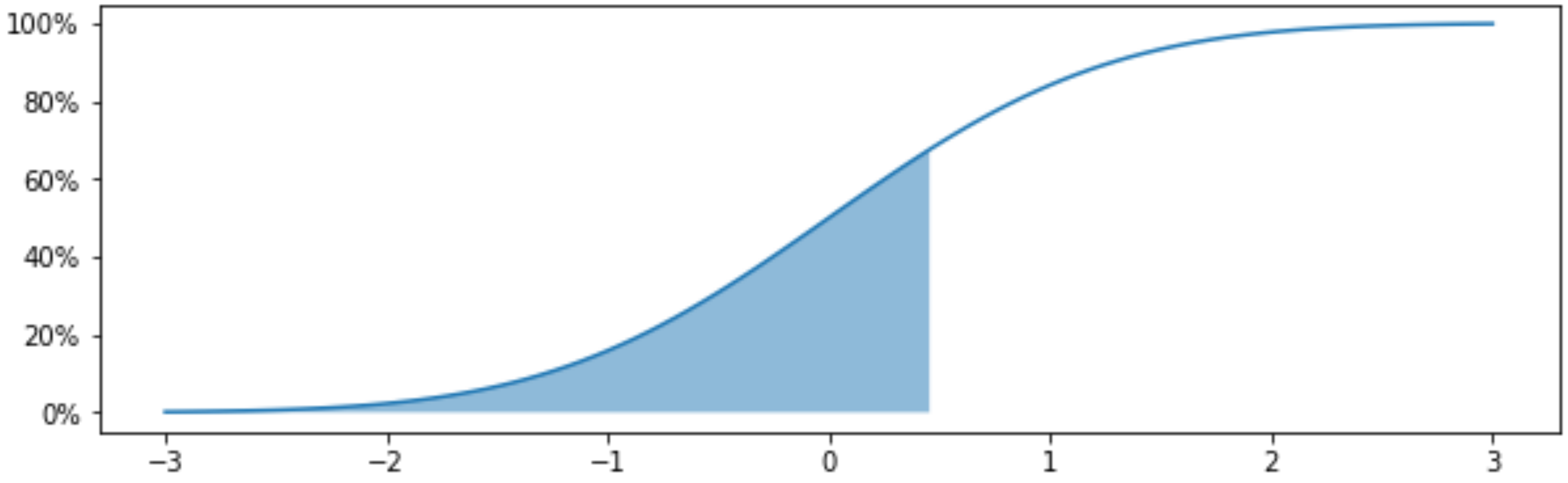
$\mu - \sigma$ μ $\mu + \sigma$

$$P(0 \leq X \leq +\infty) = \int_{0}^{+\infty} f(x)dx = 50\%$$

Quando a função é simétrica, a probabilidade para valores maiores que média é de 50%

Função Distribuição Acumulada de Probabilidade

Distribuição acumulada

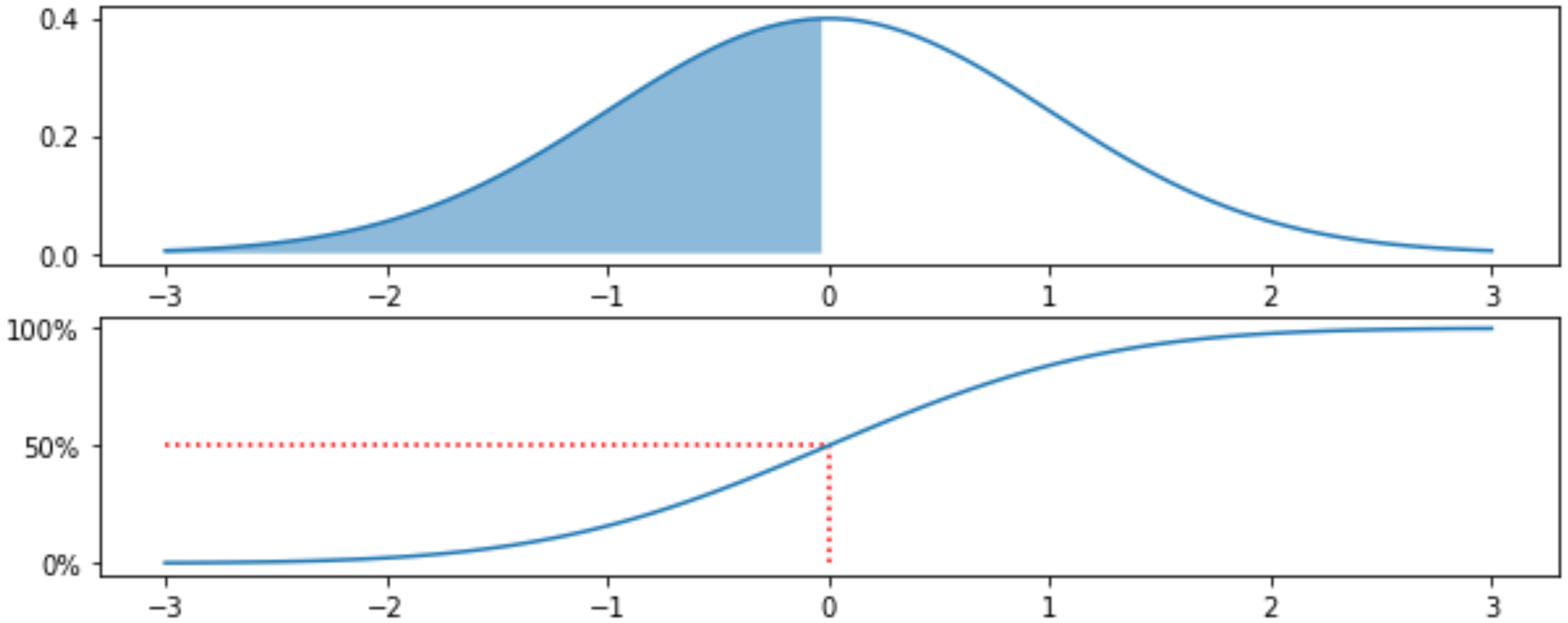


$\mu - \sigma$ μ $\mu + \sigma$

$$F(x) = P(X < x)$$

$$F(+\infty) = 1$$

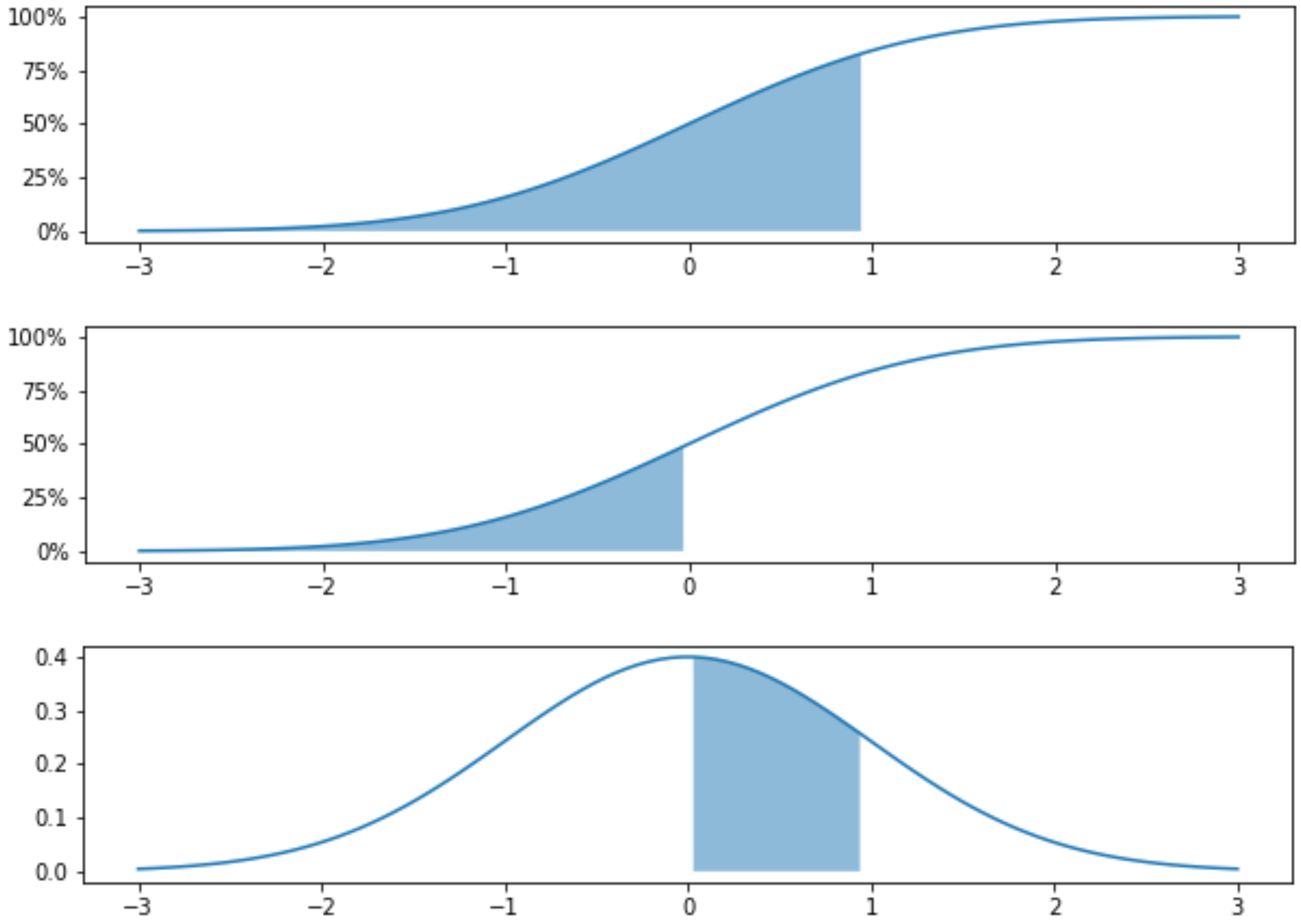
Distribuição acumulada



$$F(x) = \int_{-inf}^{x} f(x)dx = P(X < x)$$

$$F(+\infty) = 1$$

Função Densidade de Probabilidade



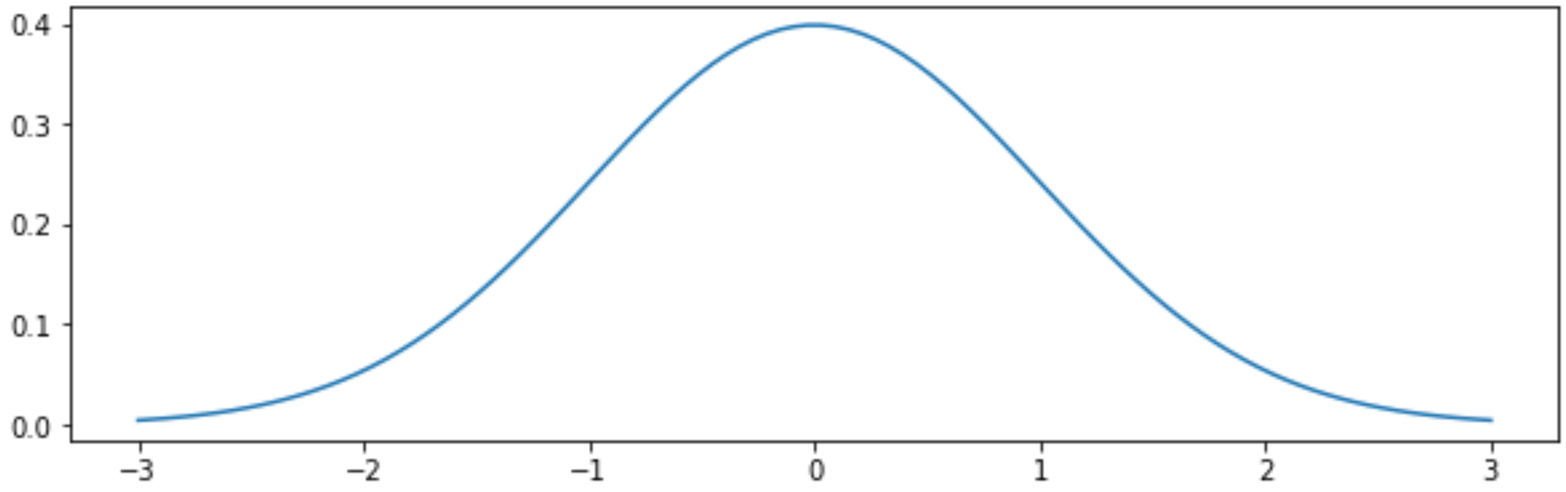
$F(1)$

$F(0)$

$P(0 < X < 1) = F(1) - F(0)$

Distribuição Normal

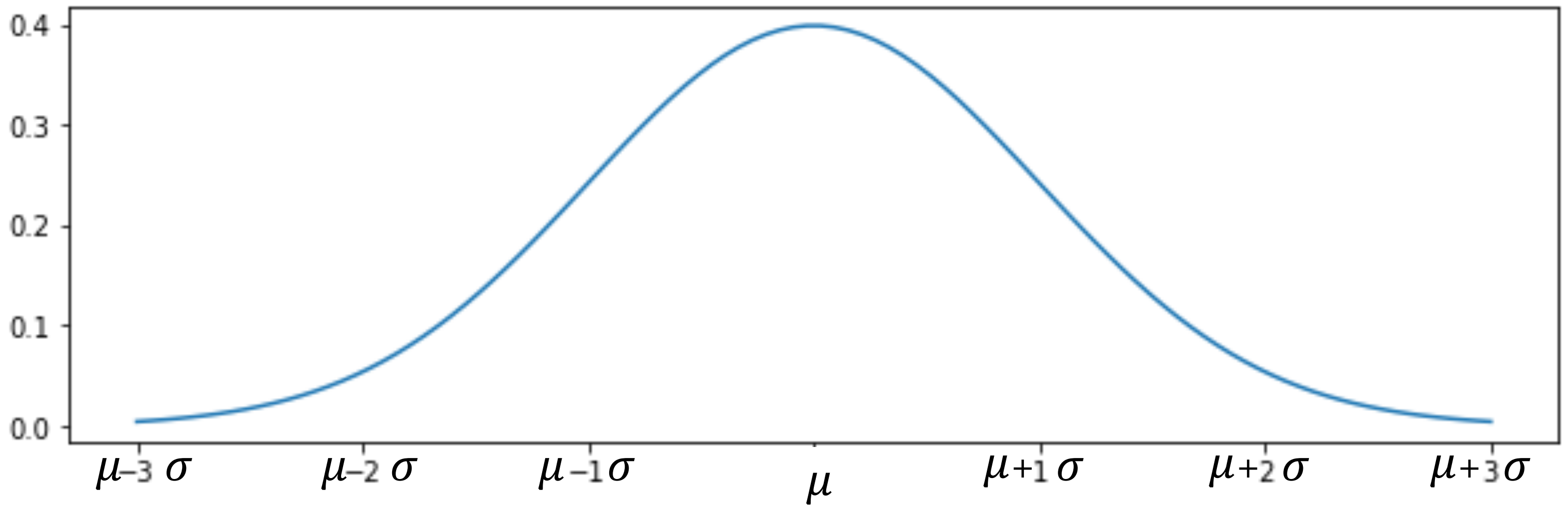
Distribuição normal



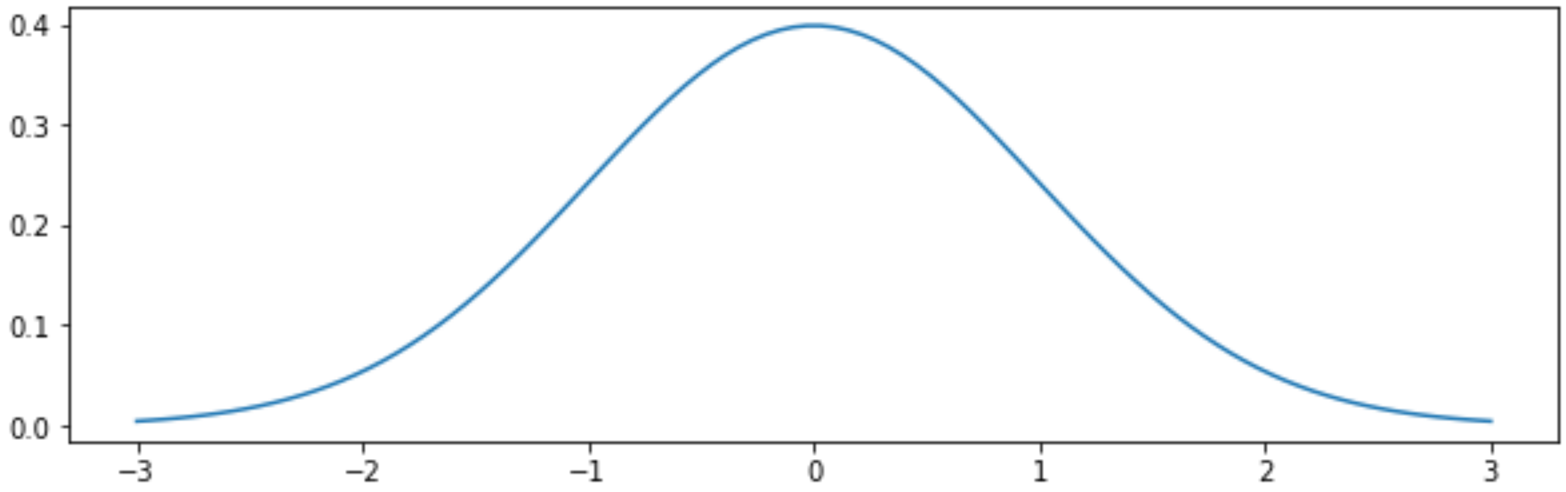
$\mu - \sigma$ μ $\mu + \sigma$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Distribuição normal



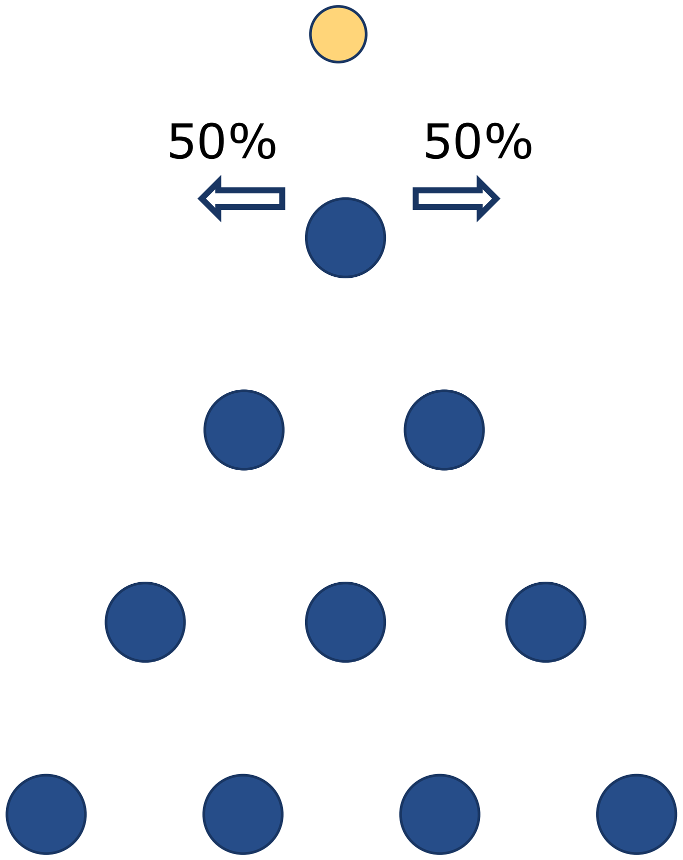
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



$$x^* = \frac{x - \mu}{\sigma}$$

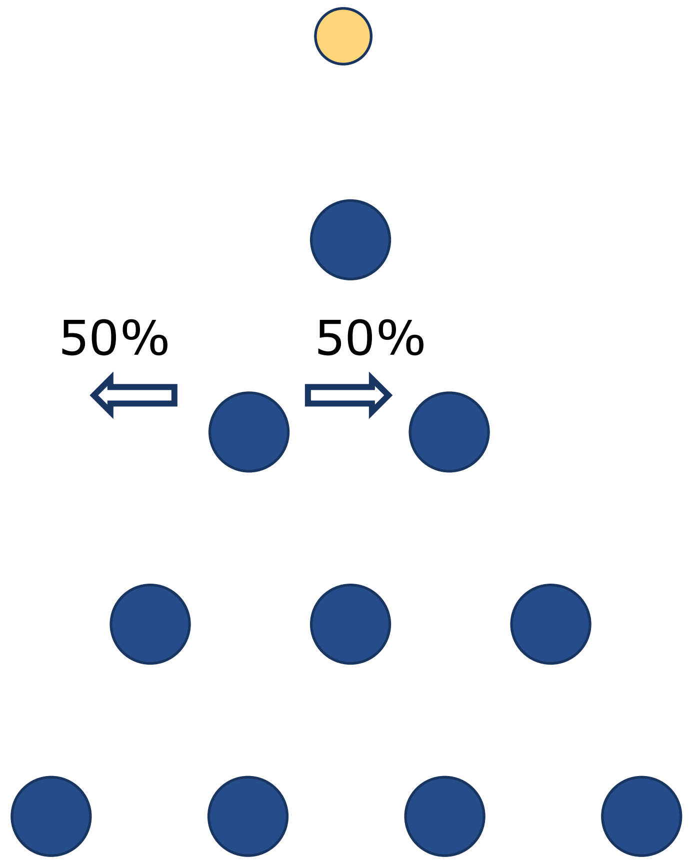
$$f(x^*) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x)^2}$$

Por que a Normal é tão importante?



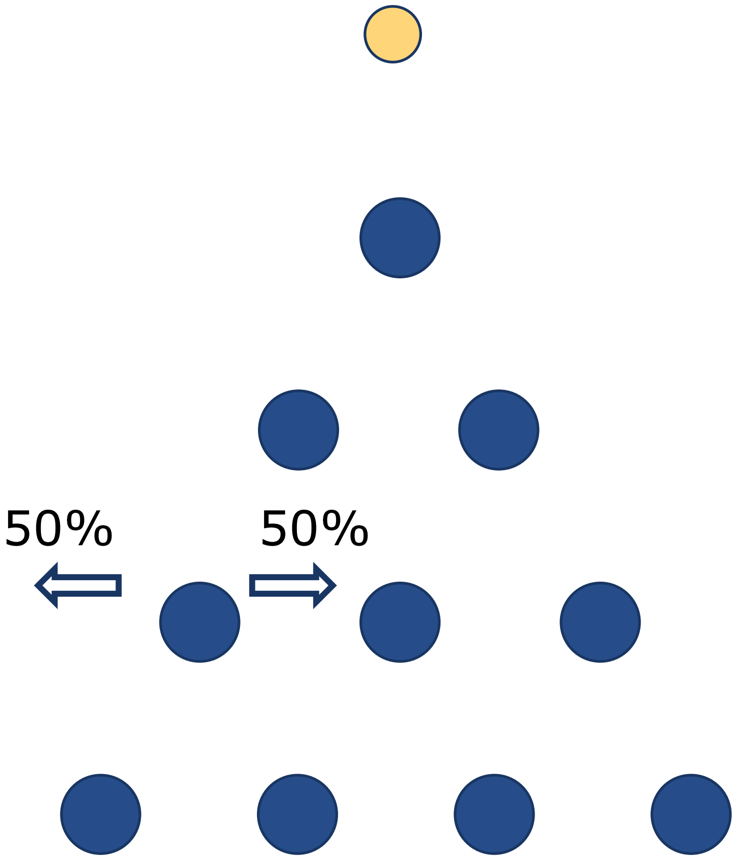
Galton Board ou Tabuleiro de Galton:

Por que a Normal é tão importante?



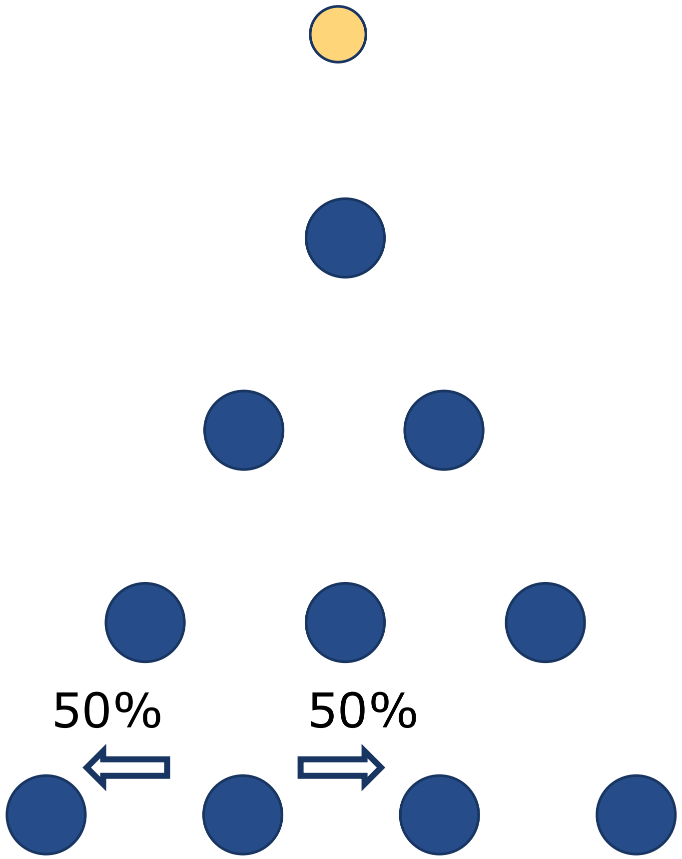
Galton Board ou Tabuleiro de Galton:

Por que a Normal é tão importante?



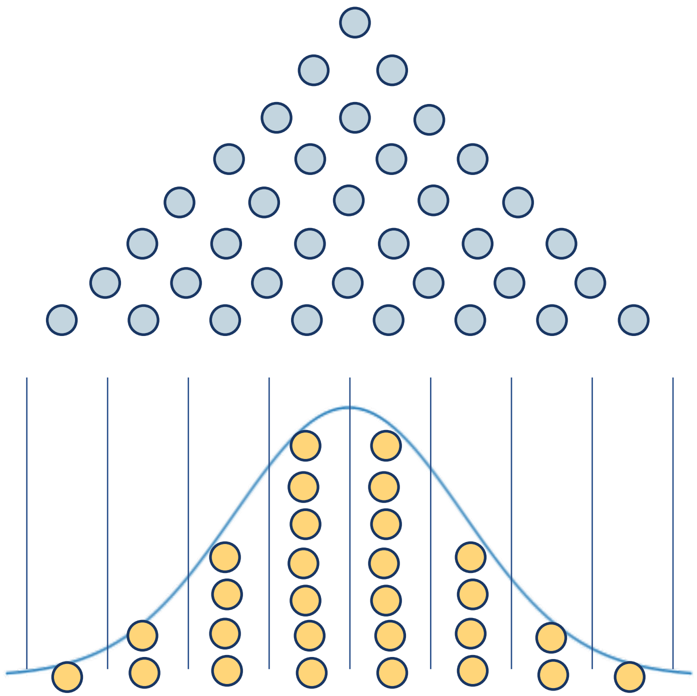
Galton Board ou Tabuleiro de Galton:

Por que a Normal é tão importante?



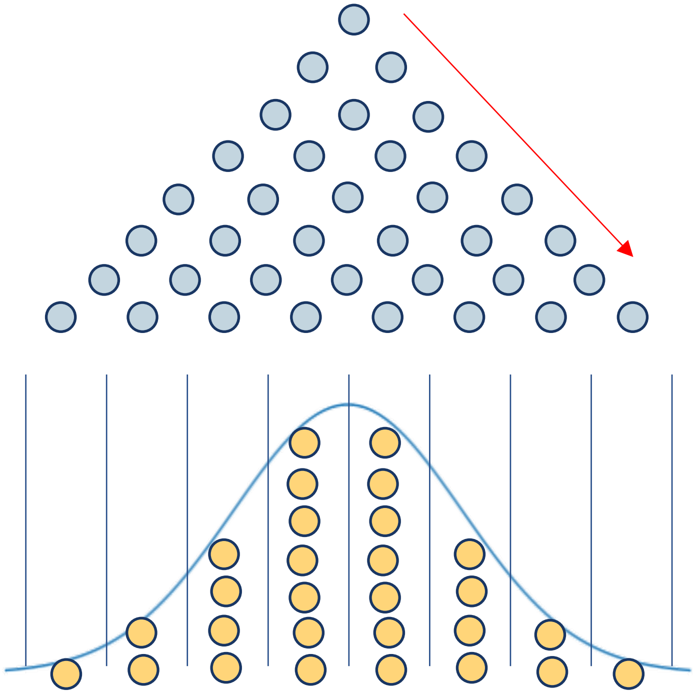
Galton Board ou Tabuleiro de Galton:

Por que a Normal é tão importante?



Galton Board ou Tabuleiro de Galton:

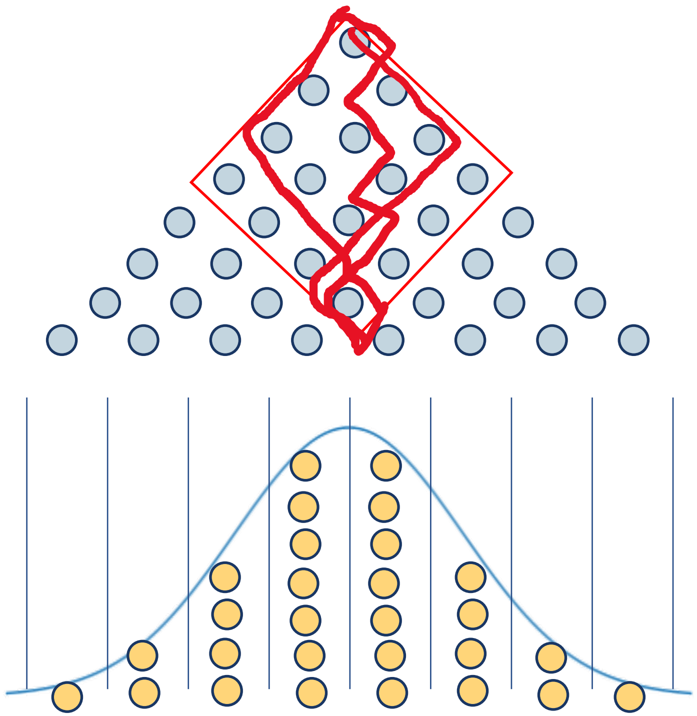
Por que a Normal é tão importante?



Somente um caminho
leva aos extremos

Galton Board ou Tabuleiro de Galton:

Por que a Normal é tão importante?



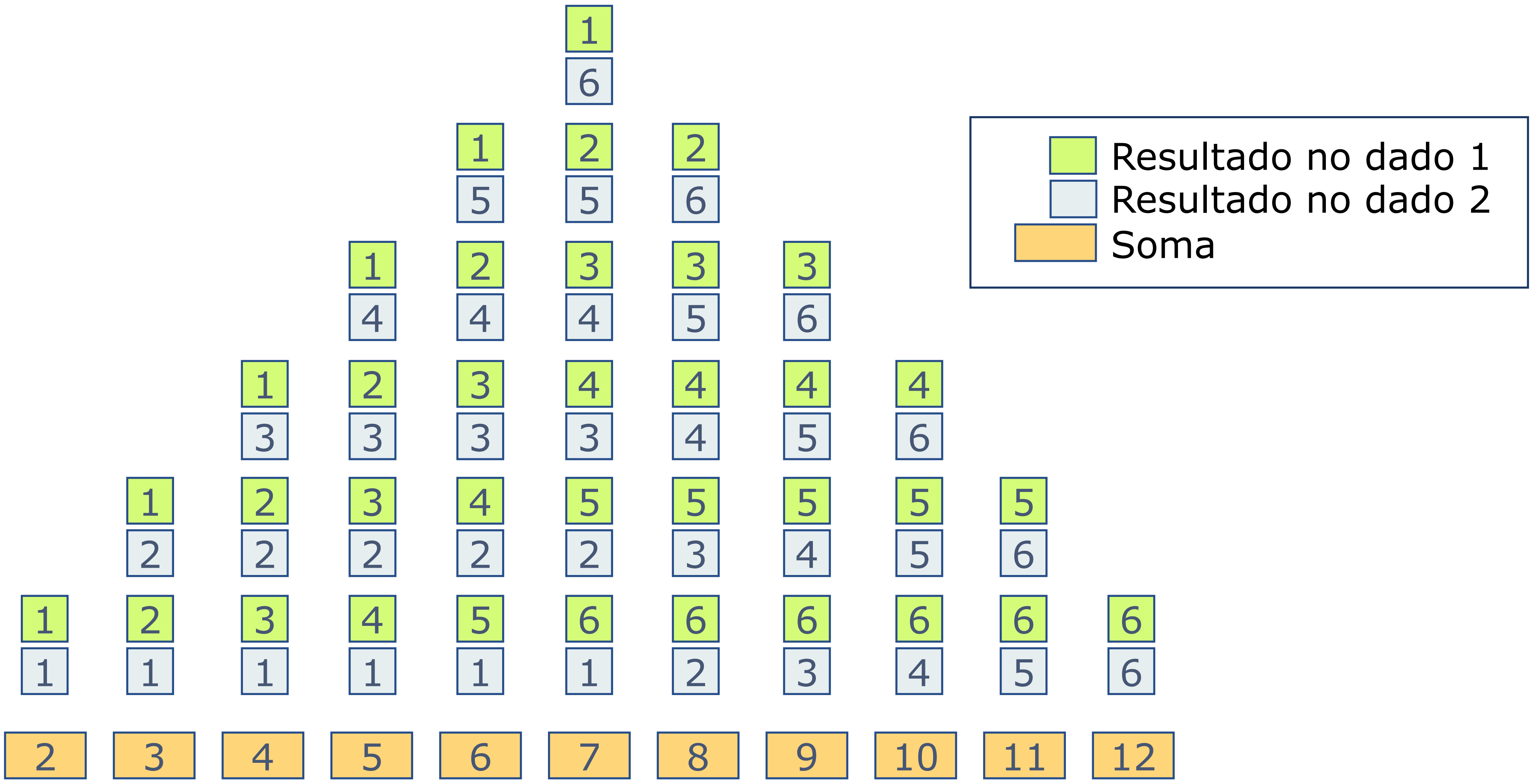
Vários caminhos levam ao meio

Galton Board ou Tabuleiro de Galton:

Por que a Normal é tão importante?

Galton Board ou Tabuleiro de Galton:

Conjunta, condicional e independência



Combinação linear

- Se $X_1 \sim N(\mu, \sigma^2)$
- $X_1^* = a + bX_1 \sim N(a + \mu,\ \ b^2\sigma^2)$ com a e b constantes.
- Ou seja:
  - A média de X_1^* é $a + \mu$
  - A variância de X_1^* é $b^2 . \sigma^2$
  - O que significa que o desvio padrão de X_1^* é $b . \sigma$ (que é $\sqrt{b^2\sigma^2}$)

- POR ISSO QUE SE $X_1 \sim N(\mu, \sigma^2)$:
  - $X_1^* = \frac{X_1 - \mu}{\sigma} \Rightarrow X_1^* \sim N(0,1)$

Combinação linear

- Se $X_1 \sim N(\mu, \sigma^2)$ e $X_2 \sim N(\mu, \sigma^2)$
- E se $X_1 \; e \; X_2$ são independentes,

$$X_1 + X_2 \sim N(2\mu, 2\sigma^2)$$

Distribuição da média

- Dos dois resultados anteriores, temos a distribuição da média:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Xi \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Ou seja:

$$\text{Desvio Padrão}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Ou ainda:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Distribuição da média

Teorema Central do Limite

Se temos uma amostra $X_1, X_2, \dots, X_n$ iid (independentes, identicamente distribuídas), então:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Xi \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Ou ainda:

$$\frac{\bar{X} - \mu}{\sigma/_{\sqrt{n}}} \approx N(0, 1)$$

Teorema do Limite Central

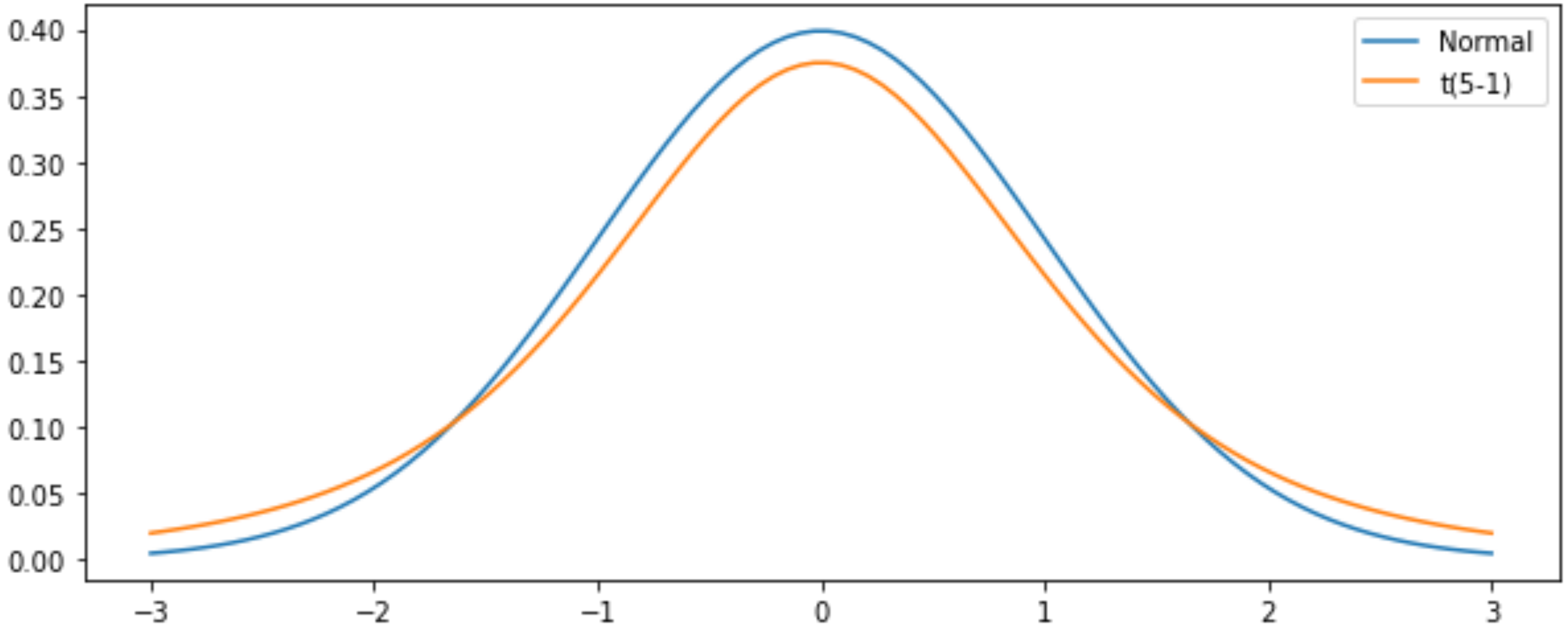
Mas se o parâmetro de variância σ^2 não é conhecido, nos resta substituí-lo pela sua estimativa S^2.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Xi \approx N\left(\mu, \frac{S^2}{n}\right)$$

Ou ainda:

$$\frac{\bar{X} - \mu}{S/_{\sqrt{n-1}}} \approx t(n-1)$$

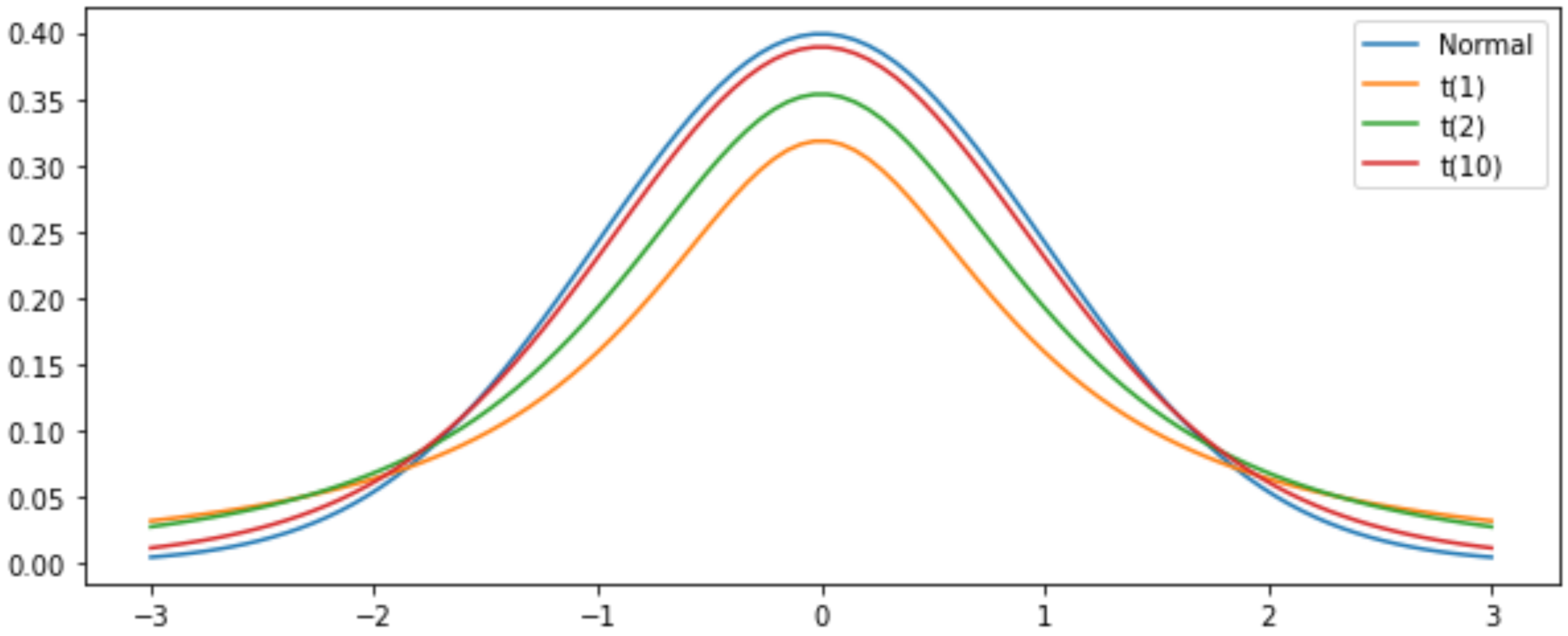
Média com Variância não conhecida



A distribuição t-student possui ‘caudas mais pesadas’.

Isso significa que, com mesma média e desvio padrão, valores mais distantes da média são mais frequentes (têm maior probabilidade)

Média com Variância não conhecida



Um fato interessante da distribuição t-student é que ela se aproxima da distribuição normal conforme o número de graus de liberdade aumenta. Na prática, a partir de 20 a aproximação pela normal já é bem razoável.