

# Fundamentos de Programação e Estatística

Prof. Dr. Henrique Ewbank

[henrique.vieira@facens.br](mailto:henrique.vieira@facens.br)

# Apresentação

Professor da Faculdade de Engenharia de Sorocaba – FACENS

Pós-doutor em Ciências Ambientais – UNESP Sorocaba

Doutor em Ciências em Administração – UFRJ

Linhas de Pesquisa:

- Sustentabilidade na cadeia de suprimentos
- Incerteza na cadeia logística
- Logística urbana

[Currículo Lattes](#)



# Apresentações



Sobre



Linguagem de programação

Software gratuito

Criado por Guido van Rossum, em 1991, inspirado na linguagem ABC

Python Software Foundation (PSF)

[www.python.org](http://www.python.org)

*“Promover, proteger e avançar a linguagem de programação Python, e dar suporte e ajudar o crescimento de uma comunidade diversificada e internacional de programadores Python”*



## Pontos Fortes

- Grande comunidade que contribui com a elaboração de pacotes (bibliotecas)
- Linguagem de alto nível
- Múltiplos paradigmas de programação
  - Imperativa (Fortran, BASIC, C)
  - Funcional (R, Javascript – parcial)
  - Procedural (Pascal, C)
  - Orientada a objetos (C++, Python, PHP)
- Fácil aprendizado

## Pontos Fracos

- Seus usuários têm dificuldades em usar outras linguagens
- Linguagem interpretada
- Não funciona bem com aplicações para dispositivos móveis

# Python Package Index – Pacotes disponíveis

*pypi.org*

Em 15 de março de 2019:

203,578 projects

1,531,340 releases

2,277,578 files

385,054 users

# Outcomes da Disciplina





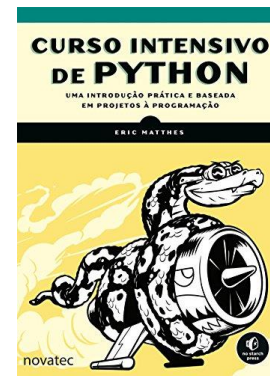
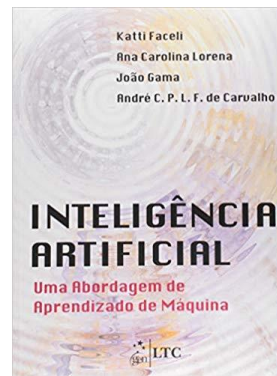
# Ementa

- Estatística Descritiva
- Probabilidades
- Testes de Hipótese
- ANOVA
- Bonferroni
- Regressão Linear
- K Vizinhos Mais Próximos

# Referências

## Livro-texto

- FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André C. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.
- MATTHES, Eric. Curso intensivo de Python: uma introdução prática e baseada em projetos à programação. São Paulo, SP: Novatec, 2016.



# Algumas versões



## **IDLE Python**

[www.python.org](http://www.python.org)



## **Jupyter Notebook**

Ambiente amigável. Roda no browser.

[www.jupyter.org](http://www.jupyter.org)



## **Spyder**

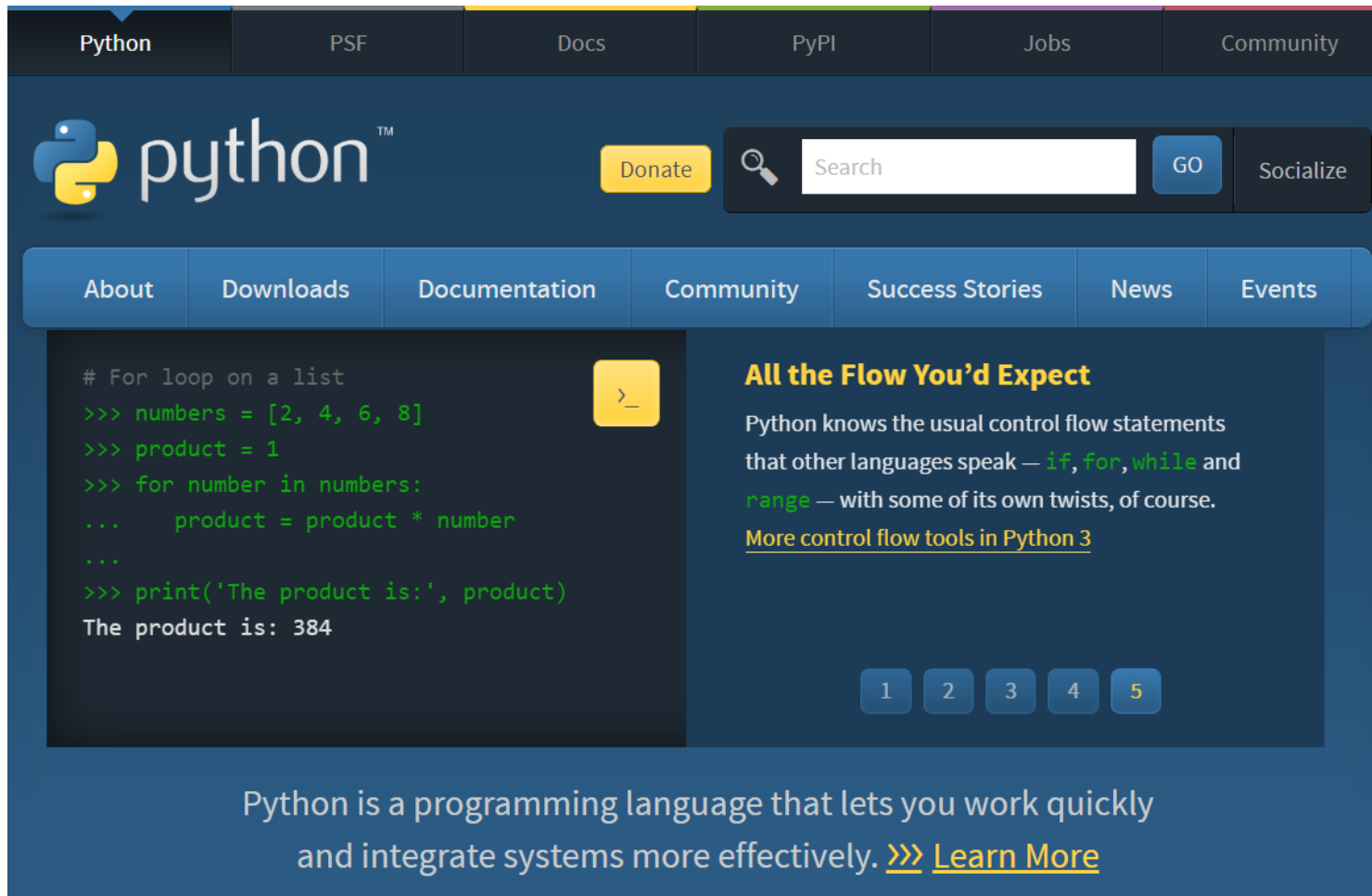
Ambiente amigável, similar ao Matlab

[www.spyder-ide.org](http://www.spyder-ide.org)



**ANACONDA®**

# Python Software Foundation – www.python.org



The screenshot shows the Python Software Foundation website. At the top, there is a navigation bar with links for Python, PSF, Docs, PyPI, Jobs, and Community. Below this is a dark blue header featuring the Python logo, a 'Donate' button, a search bar with a 'GO' button, and a 'Socialize' button. A secondary navigation bar contains links for About, Downloads, Documentation, Community, Success Stories, News, and Events. The main content area is split into two columns. The left column displays a code snippet for a for loop that calculates the product of a list of numbers, with a yellow prompt character button next to it. The right column has the heading 'All the Flow You'd Expect', followed by text about Python's control flow statements and a link to 'More control flow tools in Python 3'. At the bottom of the right column are five numbered buttons (1-5). A footer at the very bottom states: 'Python is a programming language that lets you work quickly and integrate systems more effectively. >>> [Learn More](#)'.

Python

PSF

Docs

PyPI

Jobs

Community

python™

Donate

Search

GO

Socialize

About

Downloads

Documentation

Community

Success Stories

News

Events

```
# For loop on a list
>>> numbers = [2, 4, 6, 8]
>>> product = 1
>>> for number in numbers:
...     product = product * number
...
>>> print('The product is:', product)
The product is: 384
```

>\_

### All the Flow You'd Expect

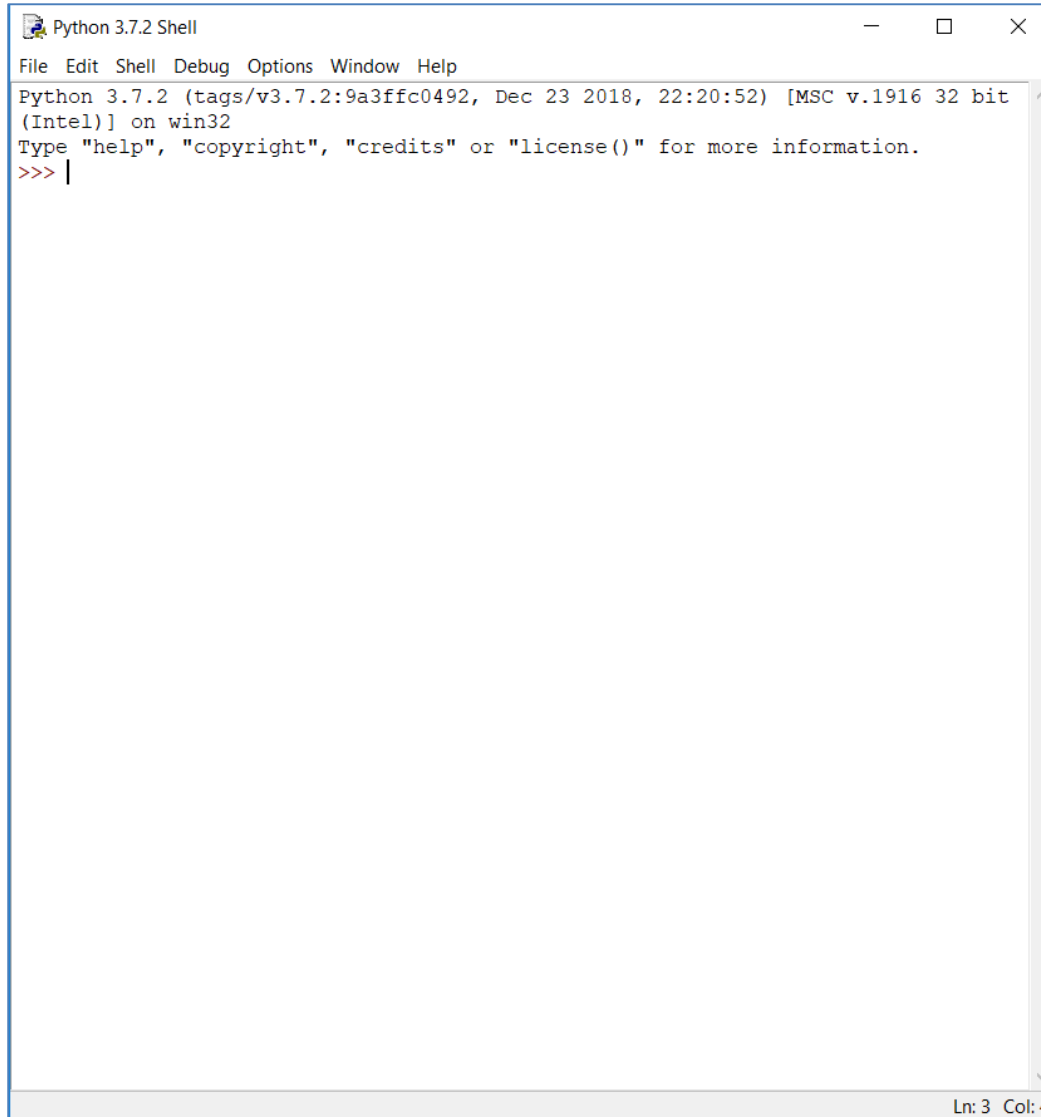
Python knows the usual control flow statements that other languages speak — `if`, `for`, `while` and `range` — with some of its own twists, of course.

[More control flow tools in Python 3](#)

1 2 3 4 5

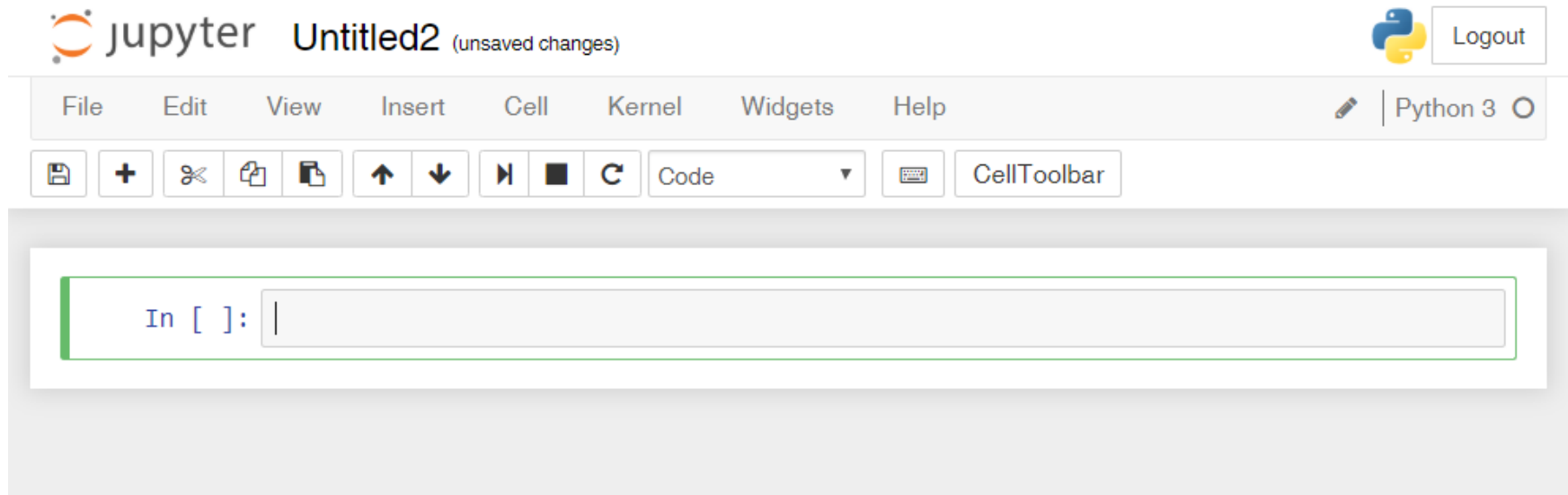
Python is a programming language that lets you work quickly and integrate systems more effectively. >>> [Learn More](#)

# IDLE - Python

A screenshot of the Python 3.7.2 Shell window. The window has a title bar that says "Python 3.7.2 Shell" and standard window controls (minimize, maximize, close). Below the title bar is a menu bar with the following items: File, Edit, Shell, Debug, Options, Window, and Help. The main text area contains the following text: "Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 22:20:52) [MSC v.1916 32 bit (Intel)] on win32", followed by "Type 'help', 'copyright', 'credits' or 'license()' for more information.", and then the prompt ">>>" followed by a vertical cursor. The status bar at the bottom right of the window shows "Ln: 3 Col: 4".

```
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 22:20:52) [MSC v.1916 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> |
```

# Jupyter Notebook



# Operadores do Python

Operador	Descrição
=	atribui um nome a um objeto. Ex: x=12
+, -, *, /	adição, subtração, multiplicação e divisão
//	fornece o maior valor inteiro menor ou igual ao valor da divisão Exemplo: 17//5 = 3
%	fornece o resto da divisão, como por exemplo, 17%5 = 2
x**b	x elevado a b, como por exemplo, 2**3 =8
=, <, >, <=, >=, !=	Utilizado em programação lógica: igual, menor, maior, menor ou igual, maior ou igual e diferente, respectivamente
and	Utilizado em programação lógica, significa “e” (adiciona uma condição)
or	Utilizado em programação lógica, significa “ou” (adiciona outra possibilidade)
not, True, False	Variáveis lógicas

# Operadores do Python

Operador	Descrição
"	delimita um caractere
'	utilizada dentro da área de abrangência do operador anterior, ou seja, dentro de um caractere
( )	delimita os argumentos de uma função
{ }	indica o início e o fim de uma função
[ ]	seleciona parte de um objeto
#	adiciona algum comentário na janela de comandos (console)
?	obtem ajuda ou informações sobre alguma função, como por exemplo, <i>?nome.da.funcao</i> irá abrir uma nova janela com informações a respeito da função desejada



# Pesquisa

## Github

- Site onde usuários postam e tiram suas dúvidas  
<https://github.com/>

## Stack Overflow

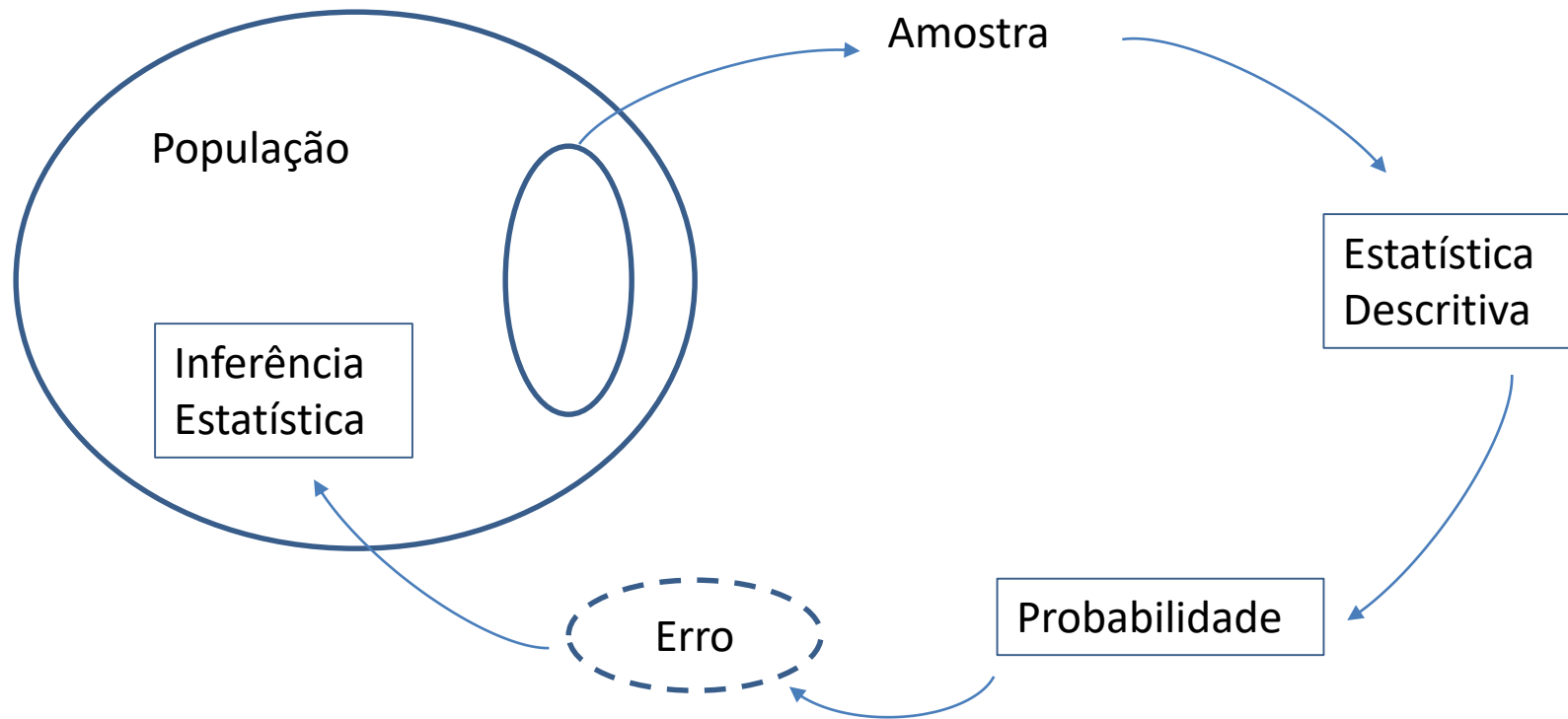
- Site onde usuários postam e tiram suas dúvidas  
<http://stackoverflow.com/>

## Google

- Observar os primeiros links ([www.google.com](http://www.google.com))

[illegible]

# Conceitos Iniciais



# Estatística Descritiva

Através da Estatística Descritiva, entendemos melhor um conjunto de dados através de suas características. As 3 principais características são:

- Um valor representativo do conjunto de dados. Ex: média
- Uma medida de dispersão ou variação
- A natureza ou forma da distribuição dos dados. Ex: sino, uniforme, assimétrica

# Estatística Descritiva

Através da Estatística Descritiva, entendemos melhor um conjunto de dados através de suas características. As 3 principais características são:

- **Um valor representativo do conjunto de dados. Ex: média**
- Uma medida de dispersão ou variação
- A natureza ou forma da distribuição dos dados. Ex: sino, uniforme, assimétrica

## Média Aritmética ( $\bar{x}$ )

É o quociente da soma dos valores da variável dividido pelo número de elementos

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \quad \text{ou} \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Ex:  $x = \{1, 5, 6, 8\}$        $\bar{x} = \frac{1+5+6+8}{4} = \frac{20}{4} = 5$

## Média Aritmética Ponderada

No cálculo da média ponderada podem ser atribuídos pesos diferentes para cada variável

i	$x_i$
1	17
2	18
3	19
4	20
5	21
6	22
7	23

$$\bar{x} = \frac{\sum_i x_i f_i}{\sum_i f_i}$$

$$\bar{x} = \frac{784}{40} = 19,60$$

## Exercício

Escreva uma função para a média ponderada, que receba uma lista com os valores e outra lista com as frequências

```
L=list(range(17,24))  
f = [1,11,8,7,10,2,1]  
print(L)  
print(f)
```

```
import numpy as np  
weighted_avg = np.average(L, weights=f)  
print(weighted_avg)
```



## Moda ( $m_o$ )

- Valor que ocorre com maior frequência em um conjunto de dados.
- Pode haver mais de 1 valor

Exemplo:  $peso = \{52; 60; 71; 75; 75; 90\}$

Mod = 75

## Exercício - Moda

Escreva um código que retorne a moda de  
{52; 60; 71; 75; 75; 90}

```
import pandas as pd
L = [52, 60, 71, 75, 75, 90]
Lpd = pd.Series(L)
Lpd.value_counts()

x = Lpd.value_counts()
x[x==max(x)]
```

## Mediana ( $\tilde{x}$ )

Valor localiza-se no centro exato da série ordenada

1. Liste em ordem crescente os valores (Rol)
2. Encontre a posição  $L = \frac{n+1}{2}$ 
  - Se  $n$  é ímpar, mediana é o valor da posição.  $\tilde{x} = x_L$
  - Se  $n$  é par, é a média aritmética entre os 2 números em torno da posição  $L$ .  $\tilde{x} = \frac{x_{n/2} + x_{n/2+1}}{2}$

Exemplo:  $altura = \{1,60; 1,62; 1,74; 1,78; 1,79; 1,79; 1,81; 1,90\}$

$$\tilde{x} = \frac{1,78 + 1,79}{2} = 1,785$$

## Medidas Separatrizes

São números reais que dividem a sequência ordenada de dados em partes que contêm a mesma quantidade de elementos da série.

Desta forma, a mediana que divide a sequência ordenada em dois grupos, cada um deles contendo 50% dos valores da sequência, é também uma medida separatriz.

Além da mediana, as outras medidas separatrizes que destacaremos são:

- Quartis;
- Quintis;
- Decis; e
- Percentis.

# Medidas Separatrizes

## Quartis:

- Se dividirmos a série ordenada em quatro partes, cada uma ficará com 25% de seus elementos. Os elementos que separam estes grupos são chamados quartis.
- Assim, o primeiro quartil, que indicaremos por  $Q_1$ , separa a sequência ordenada, deixando 25% de seus elementos, à esquerda e 75% de seus elementos à direita.
- O segundo quartil, que indicaremos por  $Q_2$ , separa a sequência ordenada, deixando 50% de seus elementos à esquerda e 50% de seus elementos à direita. (note que  $Q_2$  é a mediana).
- O terceiro quartil, que indicaremos por  $Q_3$ , separa a sequência ordenada deixando à esquerda 75% de seus elementos e 25% de seus elementos à direita.

# Medidas Separatrizes

## Quartil:

- Se dividirmos a sequência ordenada em cinco partes, cada um ficará com 20% de seus elementos. Os elementos que separam estes grupos são chamados quintis.
- Assim, o primeiro quartil, que indicaremos por  $K_1$ , separa a sequência ordenada, deixando 20% de seus elementos à esquerda(ou abaixo) e 80% de seus elementos à direita(ou acima)

# Medidas Separatrizes

Decil:

- Se dividirmos a sequência ordenada em dez partes, cada uma ficará com 10% de seus valores. Os elementos que separam estes grupos são chamados decis.
- Assim, o primeiro decil, que indicaremos por  $D_1$  separa a sequência ordenada, deixando à sua esquerda 10% de seus elementos e 90% de seus elementos à direita.
- De modo análogo são definidos os outros decis.

# Medidas Separatrizes

## Percentil:

- Se dividirmos a sequência em 100 partes, cada uma ficará com 1% de seus elementos. Os elementos que separam estes grupos são chamados centis ou percentis.
- Assim, o primeiro percentil, que indicaremos por  $P_1$ , separa a sequência ordenada deixando à sua esquerda 1% de seus valores e 99% de seus valores à direita. De modo análogo são definidos os outros percentis.



# Medidas Separatrizes

Se observarmos que os quartis e decis são múltiplos dos percentis, então basta estabelecer a fórmula de cálculo de percentis. Todas as outras medidas podem ser identificadas como percentis. *Desta forma:*

$$Q_1 = P_{25}$$

$$Q_2 = P_{50}$$

$$Q_3 = P_{75}$$

Observe que:

$$D_5 = Q_2 = P_{50} = \tilde{x} \text{ (Mediana)}$$

$$D_1 = P_{10}$$

$$D_2 = K_1 = P_{20}$$

$$D_3 = P_{30}$$

$$D_4 = K_2 = P_{40}$$

$$D_5 = Q_2 = P_{50}$$

$$D_6 = K_3 = P_{60}$$

$$D_7 = P_{70}$$

$$D_8 = K_4 = P_{80}$$

$$D_9 = P_{90}$$

## Separatrizes

`np.percentile(lista, percentis)`

```
a = np.array([1,2,3,4,5,6,7,8,9,10])  
np.percentile(a, (25,50,75))
```

## Exercício

Calcule:

- Média aritmética
- Moda
- Mediana
- Os 3 Quartis
- 2º Decil
- 9º Decil
- 59º Centil
- 4º Centil

Classe	Limites	Frequência	Frequência Acumulada
1	20 ┤ 25	3	3
2	25 ┤ 30	4	7
3	30 ┤ 35	6	13
4	35 ┤ 40	14	27
5	40 ┤ 45	20	47
6	45 ┤ 50	12	59
7	50 ┤ 55	8	67
8	55 ┤ 60	4	71
9	60 ┤ 65	1	72
		N=72	

```
L = [22.5]*3 + [27.5]*4 + [32.5]*6 + [37.5]*14 + [42.5]*20 +  
[47.5]*12 + [52.5]*8 + [57.5]*4 + [67.5]*1
```

```
npl = np.array(L)
```

```
print(npl.mean())
```

```
print(pd.value_counts())
```

```
np.percentile(npl, (25, 50, 75, 20, 90, 59, 4))
```

# Estatística Descritiva

Através da Estatística Descritiva, entendemos melhor um conjunto de dados através de suas características. As 3 principais características são:

- Um valor representativo do conjunto de dados. Ex: média
- **Uma medida de dispersão ou variação**
- A natureza ou forma da distribuição dos dados. Ex: sino, uniforme, assimétrica

## Características de variação de um conjunto de dados

- AMPLITUDE: diferença entre o maior e o menor valor
- DESVIO (ou ERRO): diferença entre cada valor e a media

$$x - \bar{x}$$

- DESVIO MÉDIO ABSOLUTO: Média dos desvios em termos absolutos

$$DMA = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

## Variância

- Descreve a variação dos dados.
- Ameniza problemas computacionais associados à extração de módulos.
- Dificuldade é a interpretação da sua dimensão, que é elevada ao quadrado

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

População

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Amostra

# Desvio Padrão

- É a raiz quadrada da variância.
- Quanto mais dispersos os dados, maior o desvio padrão

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

População

DESVPAD.P  
STDEV.P

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Amostra

DESVPAD.A  
STDEV.A

# Coeficiente de Variação

O desvio padrão por si só não nos diz muita coisa.

Assim, um desvio padrão de duas unidades pode ser considerado pequeno para uma série de valores cujo valor médio é 200; no entanto, se a média for igual a 20, o mesmo não pode ser dito.

Além disso, o fato de o desvio padrão ser expresso na mesma unidade dos dados limita o seu emprego quando desejamos **comparar duas ou mais séries de valores, relativamente** à sua dispersão ou variabilidade, quando expressas em unidades diferentes.

Para contornar essas dificuldades e limitações, podemos caracterizar a dispersão ou variabilidade dos dados em termos relativos a seu valor médio, medida essa denominada **coeficiente de variação** (cv):

$$\text{Coeficiente de Variação} = \frac{\text{Desvio Padrão}}{\text{Média}} \qquad cv = \frac{s}{\bar{x}}$$

qual demanda possui maior variabilidade:  
toneladas de carvão ou caixas de leite?



## Exercício:

$altura = \{1,72; 1,60; 1,74; 1,88; 1,82; 1,75; 1,82; 1,75; 1,75; 1,73\}$

Aluno #	Altura (x)
1	1,72
2	1,60
3	1,74
4	1,88
5	1,82
6	1,75
7	1,82
8	1,75
9	1,75
10	1,73
$\Sigma$	17,56
$\bar{x}$	1,756

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \\&= \frac{0,0502}{9} \\&= 0,0056m^2\end{aligned}$$

$$\begin{aligned}s &= \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \\&= \sqrt{0,0056} \\&= 0,0747m\end{aligned}$$

$$cv = \frac{s}{\bar{x}} = \frac{0,0747}{1,756} = 0,04$$

## Exercício:

$altura = \{1,72; 1,60; 1,74; 1,88; 1,82; 1,75; 1,82; 1,75; 1,75; 1,73\}$

Aluno #	Altura (x)
1	1,72
2	1,60
3	1,74
4	1,88
5	1,82
6	1,75
7	1,82
8	1,75
9	1,75
10	1,73
$\Sigma$	17,56
$\bar{x}$	1,756

```
import numpy as np
```

```
#Desvio padrão populacional
```

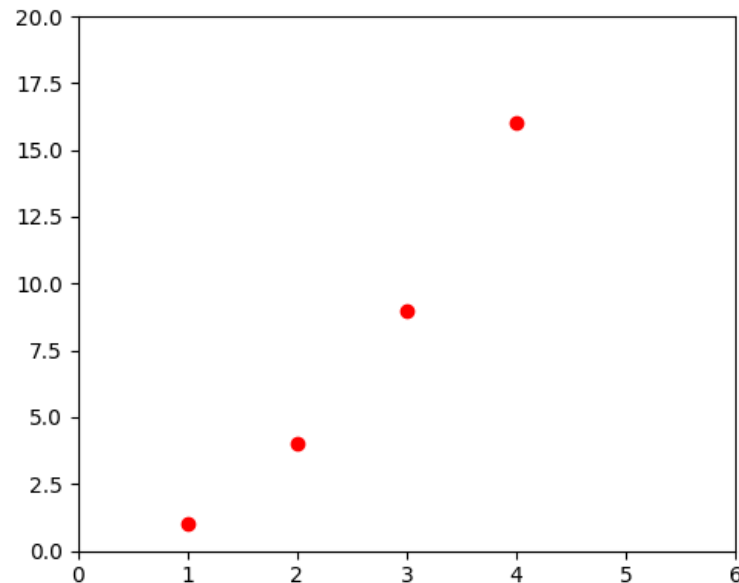
```
np.std(L)
```

```
#Desvio padrão amostral
```

```
np.std(L, ddof=1)
```

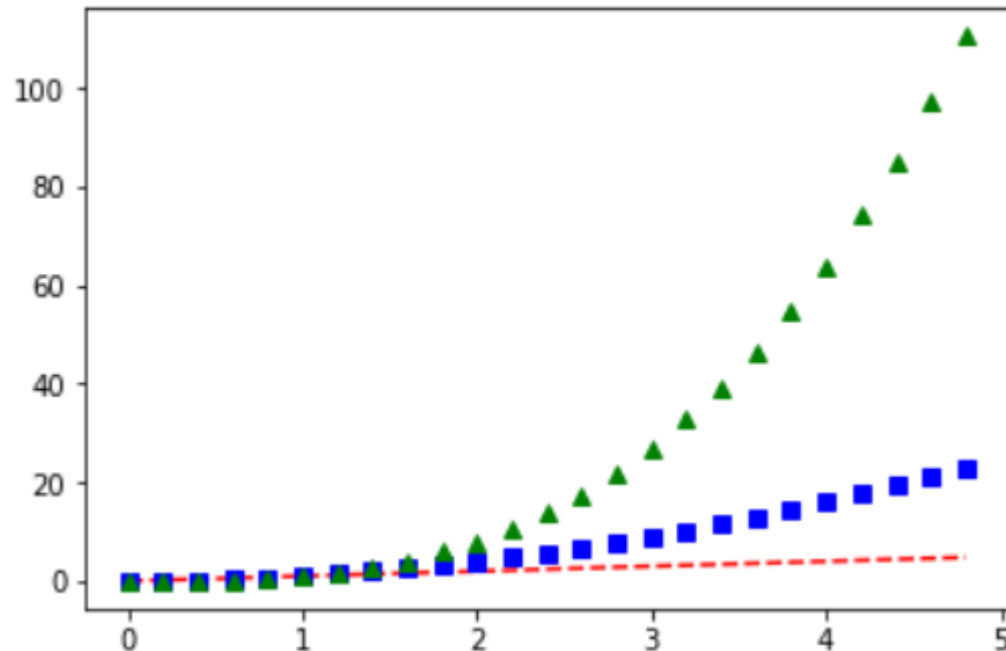
# Gráficos de Dispersão

```
import matplotlib.pyplot as plt
plt.plot([1,2,3,4], [1,4,9,16], 'ro') #red dot
plt.plot([1,2,3,4], [1,4,9,16], 'b-') #blue line
plt.axis([0, 6, 0, 20]) #Define amplitude dos eixos X e Y
plt.show()
```



# Gráficos de Dispersão

```
import numpy as np  
t = np.arange(0., 5., 0.2)  
plt.plot(t, t, 'r--', t, t**2, 'bs', t, t**3, 'g^')  
plt.show()
```



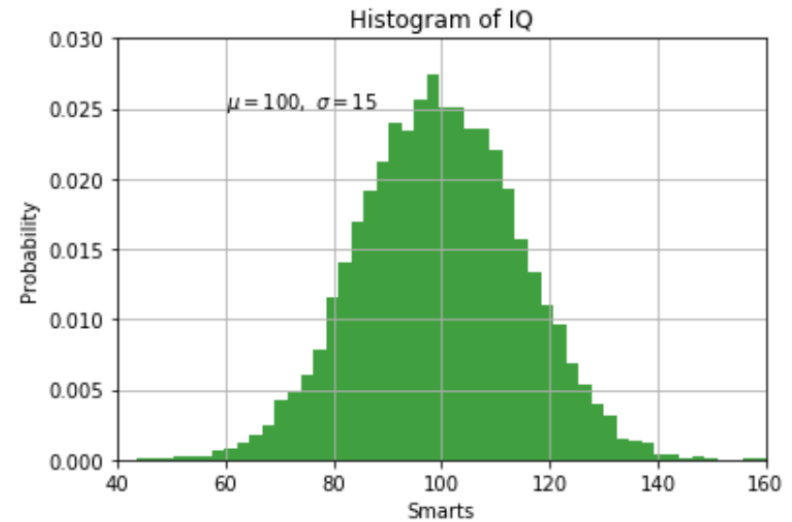
# Histograma

```
np.random.seed(19680801)

mu, sigma = 100, 15
x = mu + sigma * np.random.randn(10000)

# Histograma - cumulative=True
n, bins, patches = plt.hist(x, 50, normed=1, facecolor='g', alpha=0.75)

plt.xlabel('Smarts')
plt.ylabel('Probability')
plt.title('Histogram of IQ')
plt.text(60, .025, r'$\mu=100, \sigma=15$')
plt.axis([40, 160, 0, 0.03])
plt.grid(True)
plt.show()
```

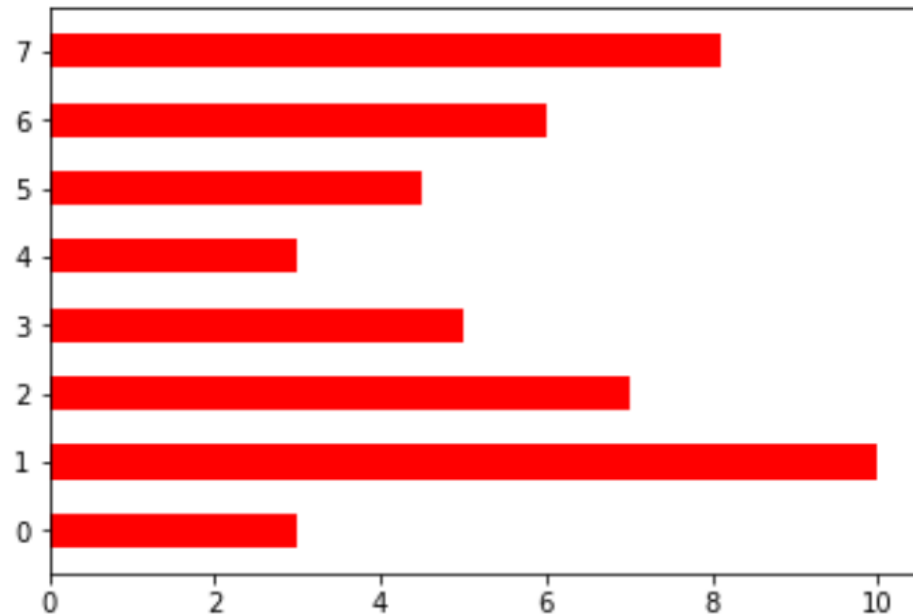


## Gráfico de Barras

```
y = [3, 10, 7, 5, 3, 4.5, 6, 8.1]
N = len(y)
x = range(N)
width = 0.5
plt.bar(x, y, width, color="red")
plt.show()
```

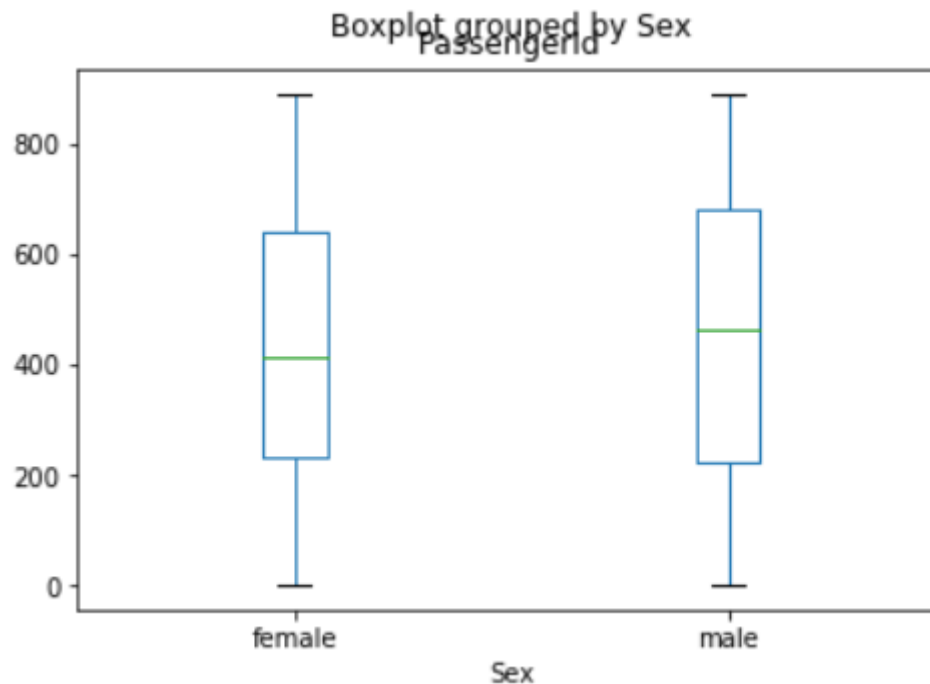
### Barras horizontais:

```
plt.barh()
```



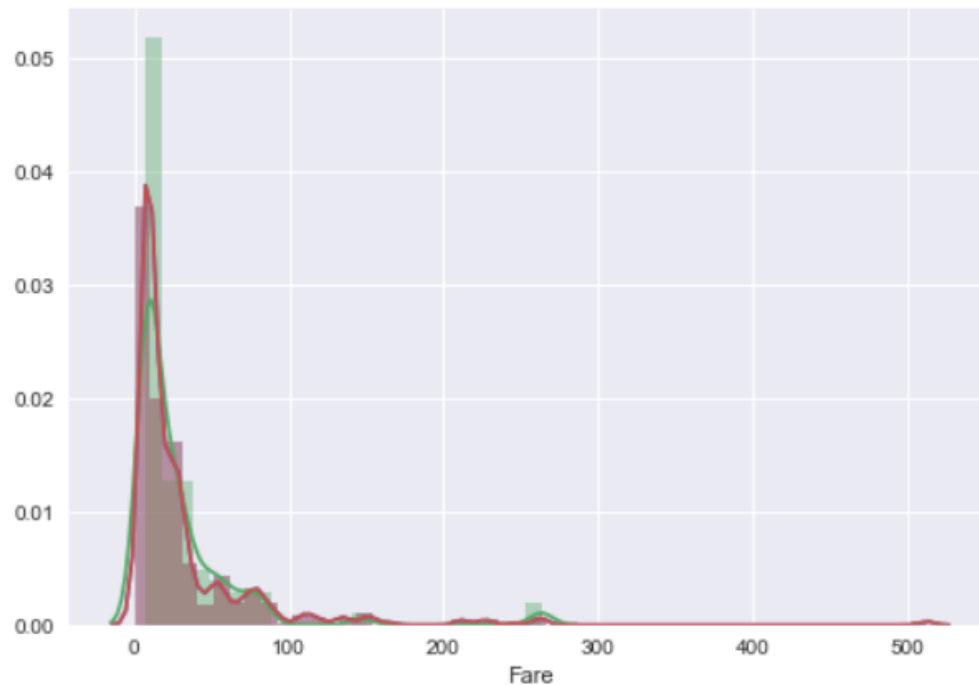
# Boxplot

```
import pandas as pd
train_df = pd.read_csv('train.csv')
train_df['PassengerId']
train_df.boxplot(column="PassengerId", by="Sex", grid=False)
```



## Densidade

```
import seaborn as sns
sns.distplot(train_df['Fare'])
sns.distplot(train_df['Fare'].head(100))
sns.distplot(train_df['Fare'].head(1000))
```





# Probabilidades

Probabilidade é a medida das incertezas relacionadas a um evento, ou, é a chance de ocorrência de um evento.

## Exemplos:

- Probabilidade de jogar uma moeda e sair cara
- Chance de ocorrer uma tempestade
- Probabilidade de ganhar na Mega Sena

$$p = \frac{\# \text{ sucesso}}{\# \text{ possibilidades}}$$



Todos os eventos  
possuem a mesma  
chance de ocorrer

## Exemplos

1. Qual a probabilidade de obter um 10 em um baralho de 52 cartas?

Há 4 cartas 10 dentre as 52 cartas:  $\frac{4}{52} = \frac{1}{13}$

2. Qual a probabilidade de tirar um 2 ou um 3 em uma jogada de um dado honesto?

Sucessos: {2;3}

Possibilidades: {1;2;3;4;5;6}

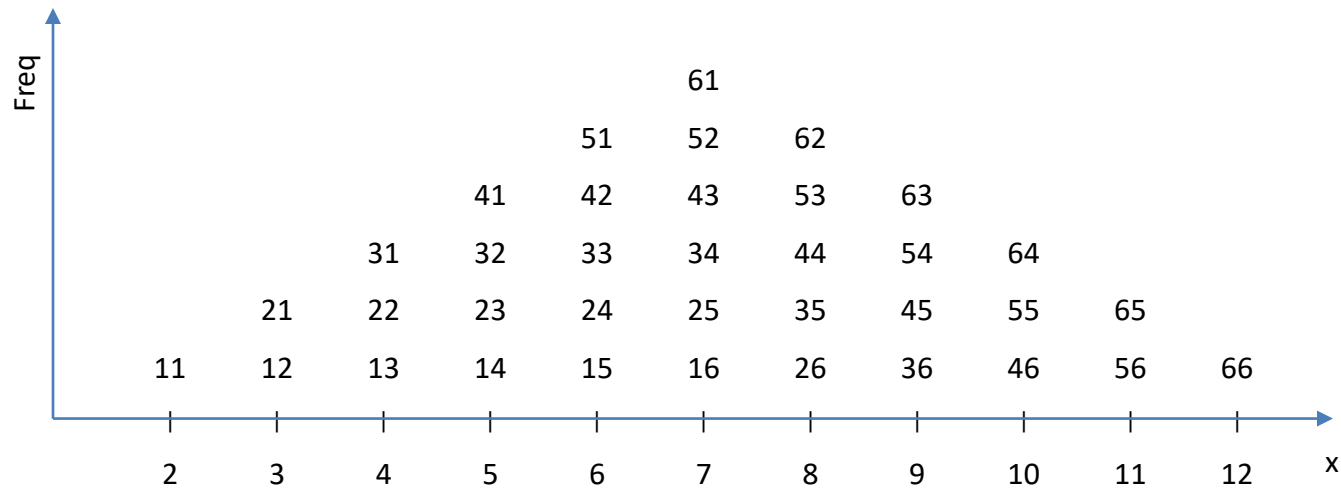
$$\frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

## Exercícios

3. Ao jogarmos 2 moedas, qual a probabilidade de tirarmos:

- a) Zero caras
- b) Uma cara
- c) Duas caras

4. Qual a probabilidade de obtermos 8 jogando duas vezes um dado?



➡ Qual a probabilidade de chover amanhã?

$s = \# \text{ sucesso} = 1$

$n = \# \text{ possibilidades} = 2$

$p = \frac{1}{2} ?$

**FALSO!**

Não se pode fazer esta afirmação, pois os eventos não possuem a mesma chance de ocorrer

Nestes casos, a interpretação frequencial deve ser usada para determinar a possibilidade de ocorrência de um evento – PROBABILIDADE

$$P(A) = \frac{\# \text{ocorrências de } A}{\# \text{repetições do experimento}}$$

## Exercício

1. Sabendo que 567 voos dentre 700 voos da Brasil Airlines chegam no horário, qual a probabilidade de que um voo daquela companhia chegue no horário?

$$P = \frac{567}{700} = 0,81$$

2. No ultimo ano, 175 máquinas dentre 625 vendidas precisaram de algum reparo. (a) Qual a probabilidade de que uma máquina não precise de manutenção? (b) E qual a probabilidade de que uma máquina precise de manutenção?

(a) 
$$P(A) = \frac{625 - 175}{625} = \frac{450}{625} = 0,72$$

(b) 
$$P(\bar{A}) = 1 - P(A) = 1 - 0,72 = 0,28$$

## Comentários

Observa-se que a conclusão de probabilidade de eventos futuros está toda baseada em experimentos passados. Portanto surge a pergunta:

- Que garantias temos sobre a estimativa feita?
- Mais adiante será apresentado um método que estima a precisão do resultado
- Por enquanto, é importante saber a LEI DOS GRANDES NÚMEROS

## Lei dos Grandes Números

*“Quanto maior for a repetição do experimento, maior a aproximação da probabilidade efetiva de acontecimento de um determinado evento através da frequência relativa”*

## Comentários

- Alguns experimentos, mesmo que tenham os resultados todos com a mesma chance de ocorrer, são muito complexos para serem resolvidos através de abordagem clássica  $\left(\frac{s}{n}\right)$
- Utiliza-se então a frequência relativa. Ex: Probabilidade de ganhar no jogo de paciência
- Nestes casos há métodos de simulação para gerar experimentos a partir de poucos resultados



## Amostras aleatórias

Para gerar experimentos, os eventos devem ser escolhidos de tal maneira que toda amostra de “n” elementos da população tenha a mesma chance de ser escolhido, sendo um conjunto de dados imparcial, representativo e não tendencioso.

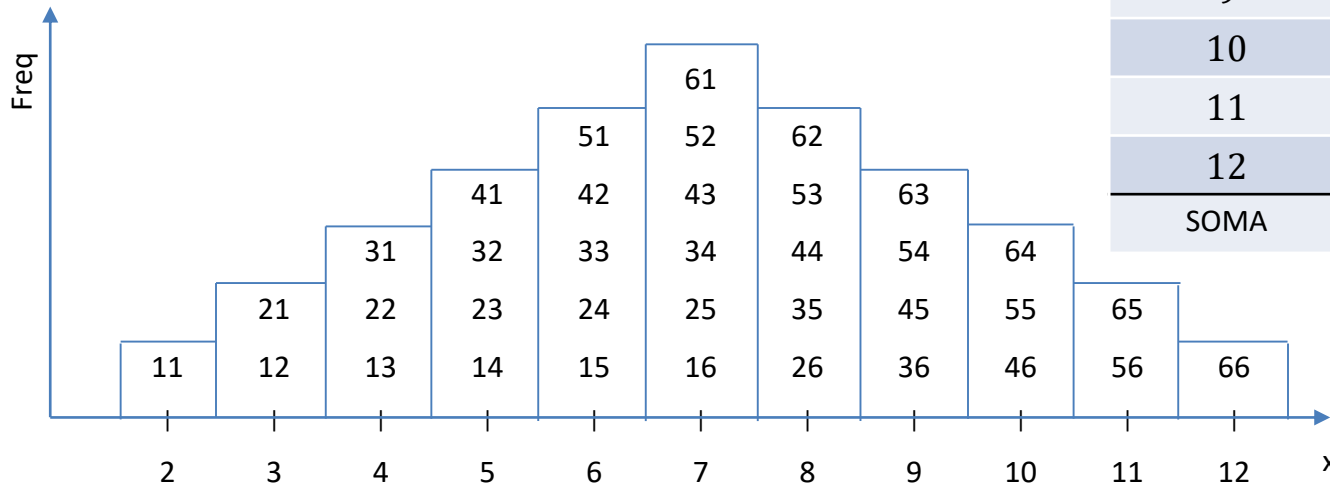
# Variáveis aleatórias

- Corresponde ao resultado de um experimento.
- Geralmente representada por um “X”
- Exemplos:
  - Números de alunos que aparecem às aulas de estatística;
  - Resultado de uma jogada de um dado
- Variável aleatória Discreta: num intervalo determinado, admite um número finito de valores
- Variável aleatória Contínua: pode tomar um número infinito de valores e pode ser associada a uma mensuração em uma escala contínua

# Exemplo

Resultado do lançamento de dois dados honestos simultaneamente

Distribuição de Probabilidades



Soma das Faces ( $X_i$ )	# de Possibilidades	Probabilidade $P(X)$
2	1	2,87%
3	2	5,56%
4	3	8,33%
5	4	11,11%
6	5	13,89%
7	6	16,67%
8	5	13,89%
9	4	11,11%
10	3	8,33%
11	2	5,56%
12	1	2,87%
SOMA	36	100%

## Exemplo (CONT.)

Qual a probabilidade da soma de dois dados ser igual a 5?

$$P(X=5) = 11,11\%$$

Qual a probabilidade da soma de dois dados ser menor do que 5?

$$\begin{aligned} P(X<5) &= P(X=2) + P(X=3) + P(X=4) \\ &= 2,87 + 5,56 + 8,33 \\ &= 16,76\% \end{aligned}$$

Qual a probabilidade da soma de dois dados ser maior do que 10?

$$\begin{aligned} P(X>10) &= P(X=11) + P(X=12) \\ &= 5,56 + 2,87 \\ &= 8,43\% \end{aligned}$$

Qual a probabilidade da soma de dois dados ser maior do que 4 e menor do que 7?

$$\begin{aligned} P(4<X<7) &= P(X=5) + P(X=6) \\ &= 11,11 + 13,89 \\ &= 25,00\% \end{aligned}$$

## Exercício

Dada a distribuição de probabilidades a seguir, calcule:

a)  $P(X \leq 3)$

b)  $P(X \text{ ímpar})$

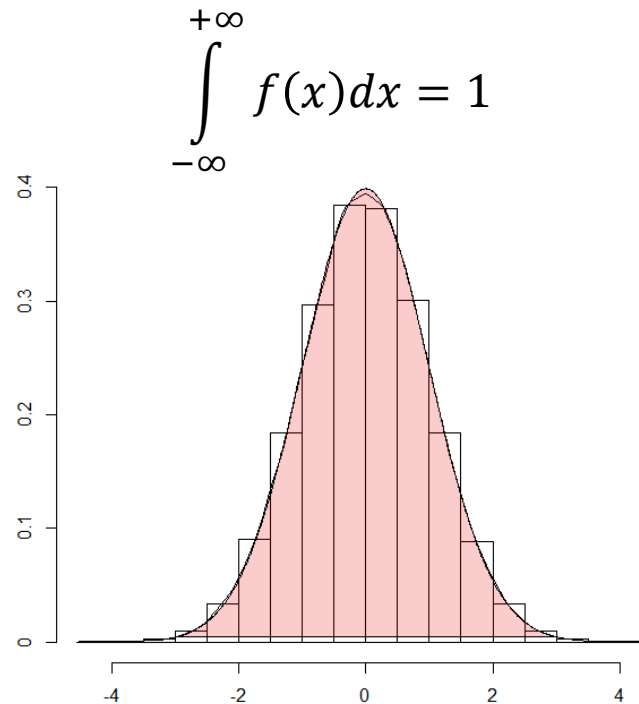
c)  $P(X < 2 \text{ ou } X > 4)$

X	P(X)
1	0,1
2	0,2
3	0,4
4	0,2
5	0,1

# Função Densidade de Probabilidade

A probabilidade é igual à área sob a curva

Logo, se  $f(x)$  descreve a curva, chamada **função densidade de probabilidade**, a área sob a curva:



# Distribuição Normal

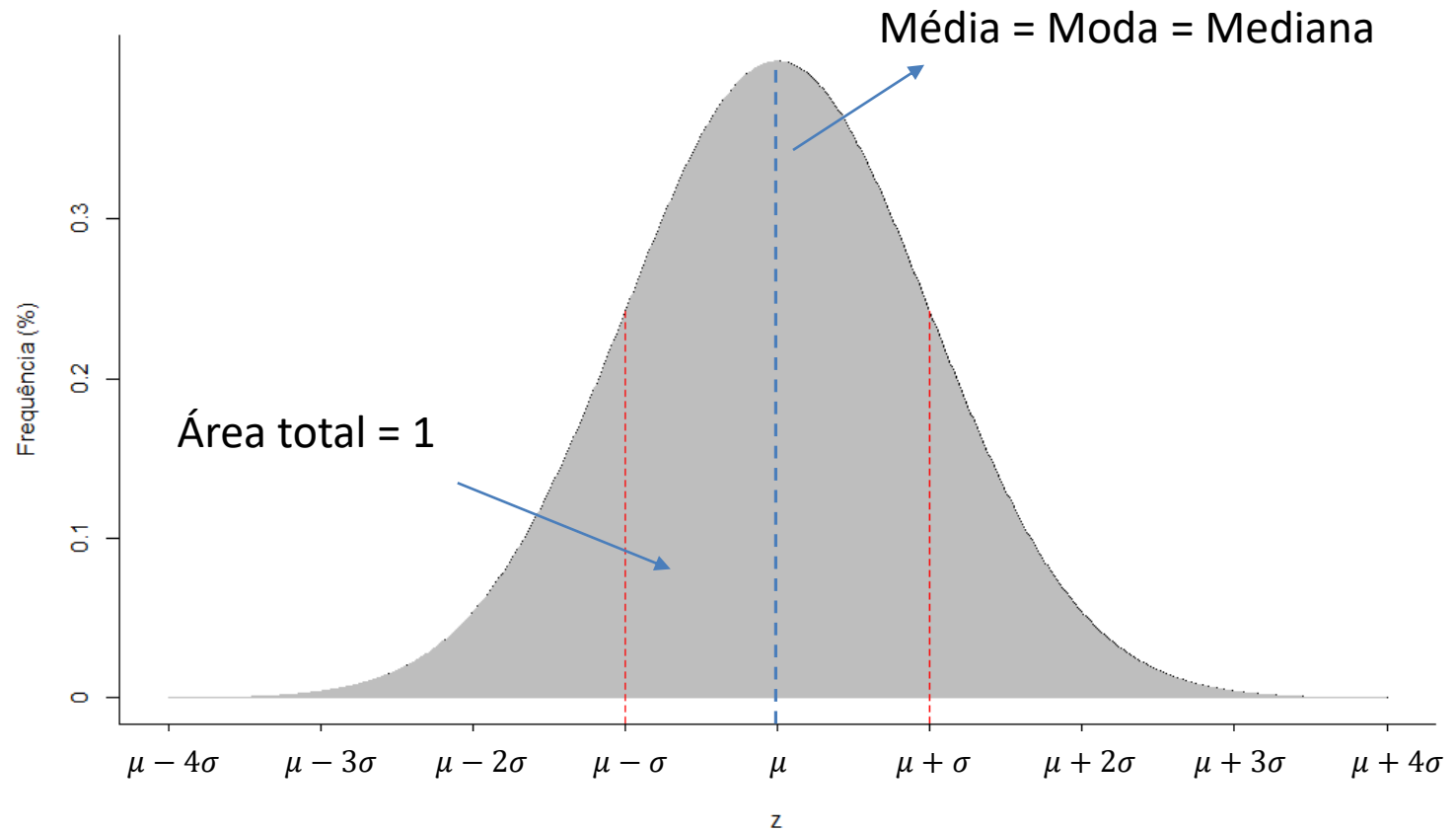
# Definição

Um distribuição normal é uma distribuição de probabilidade contínua para uma variável aleatória  $X$  (Larson e Farber, 2010)



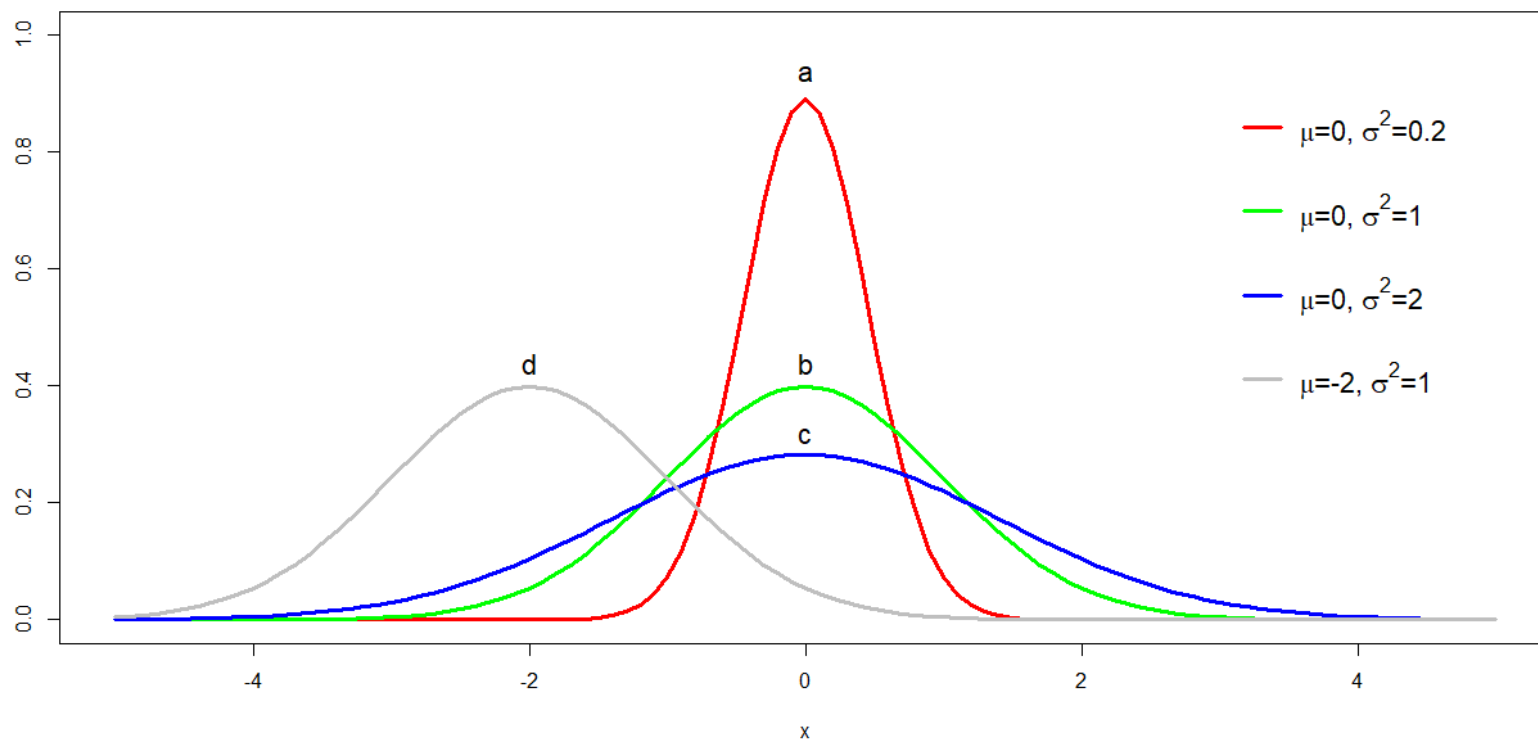


# Curva Normal



# Descrição

$N(\mu, \sigma^2)$ ,  $\mu$  = média  
 $\sigma^2$  = variância

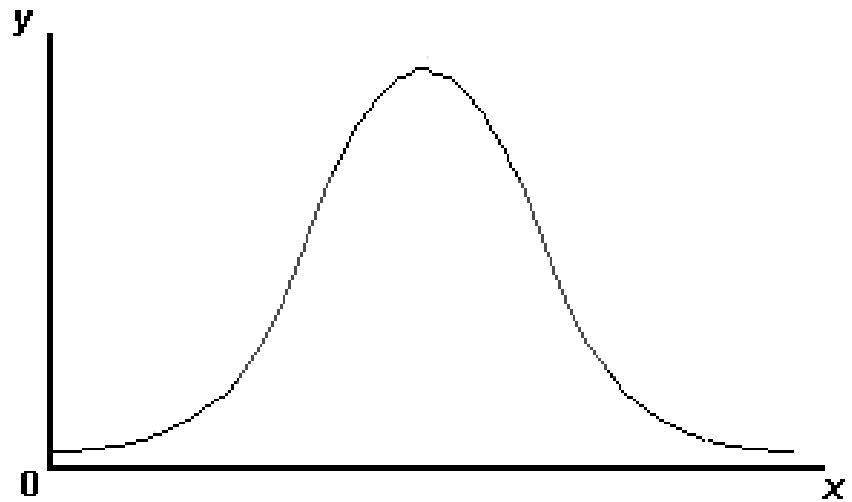


<https://hewbank.shinyapps.io/shiny/>

# Equação

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

$\mu$  = média  
 $\sigma$  = desvio padrão

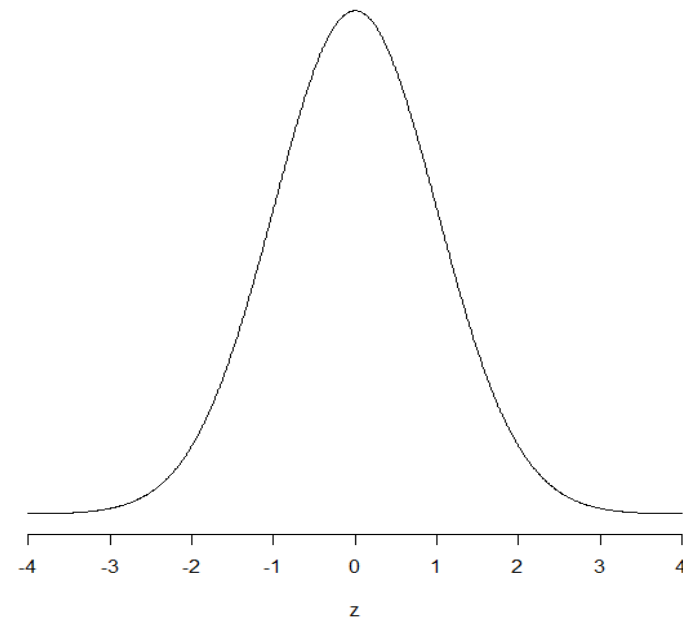


# Distribuição Normal Padrão

$N(0,1)$ ,  $\mu = \text{média}$   
 $\sigma = \text{desvio padrão}$

$$z = \frac{\text{Valor} - \text{Média}}{\text{Desvio padrão}}$$

$$z = \frac{x - \mu}{\sigma}$$



# Geração de números aleatórios com distribuição normal

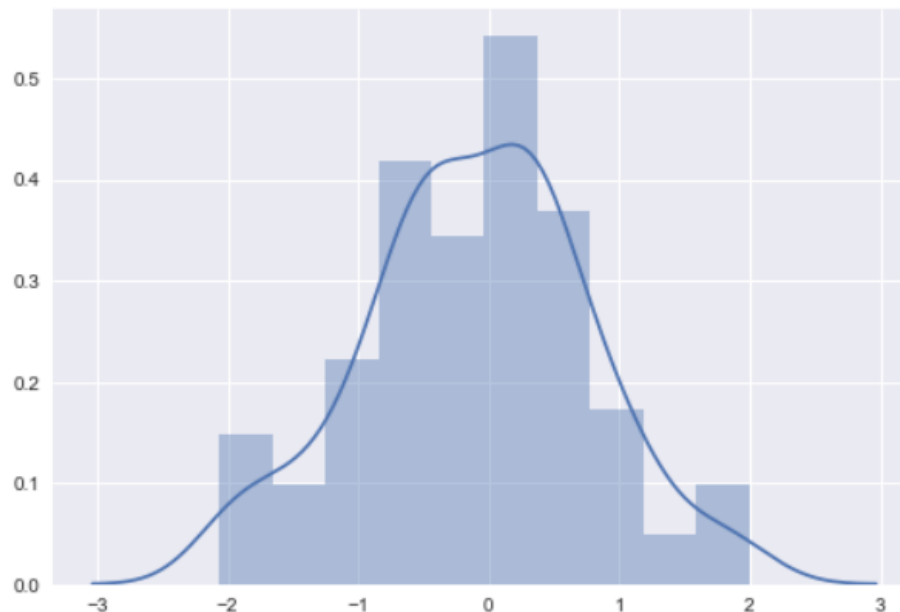
```
np.random.seed(123)
```

```
mu = 0
```

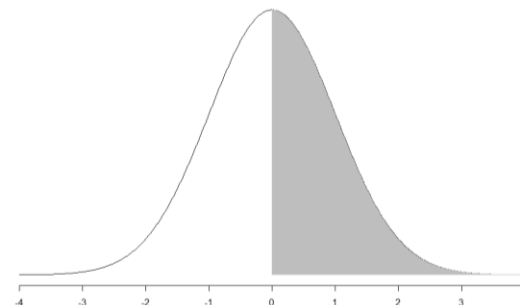
```
sigma = 1
```

```
s = np.random.normal(mu, sigma, 100)
```

```
sns.distplot(s)
```



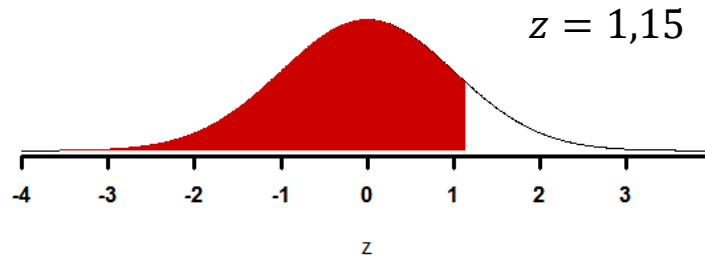
# Tabela



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964

# Cálculo de probabilidades #1

```
import scipy.stats
scipy.stats.norm(0, 1).pdf(0)
scipy.stats.norm(0, 1).cdf(0)
scipy.stats.norm(0, 1).cdf(1.15)
scipy.stats.norm(100, 12).pdf(98)
scipy.stats.norm(100, 12).cdf(98)
scipy.stats.norm(100, 12).cdf(100)
```

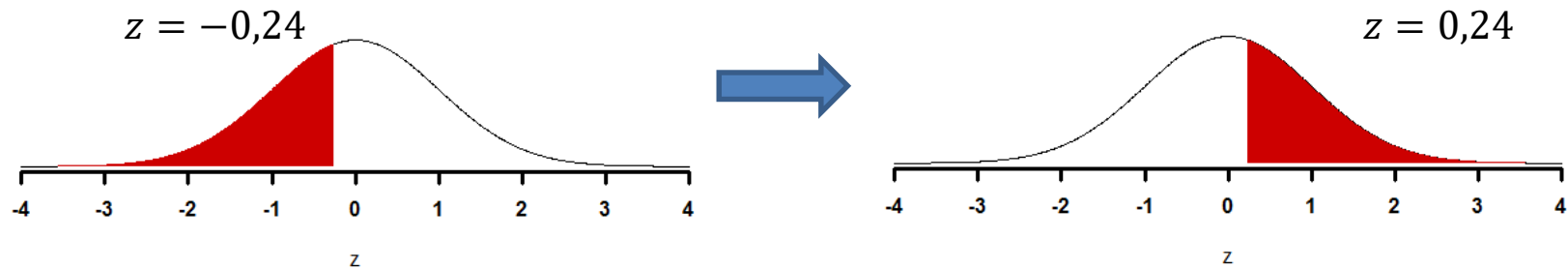


$$\begin{aligned}
 P(z < 1,15) &= 0,5000 + P(z < 1,15) \\
 &= 0,5000 + 0,3749 \\
 &= 0,8749 = 87,49\%
 \end{aligned}$$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292

`scipy.stats.norm(0, 1).cdf(-0.24)`

## Cálculo de probabilidades #2

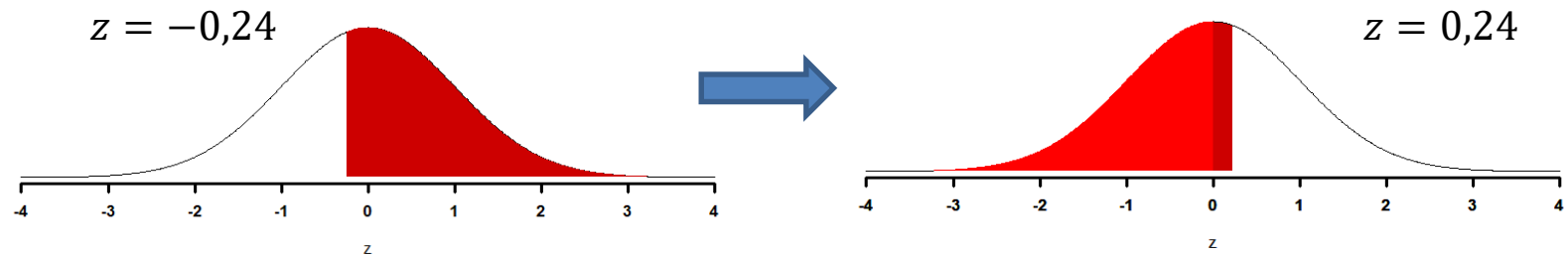


z	0,00	0,01	0,02	0,03	0,04	0,05	0,06
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026

$$\begin{aligned}P(z < -0,24) &= P(z > 0,24) \\&= 0,5000 - P(z < 0,24) \\&= 0,5000 - 0,0948 \\&= 0,4052 \\&= 40,52\%\end{aligned}$$



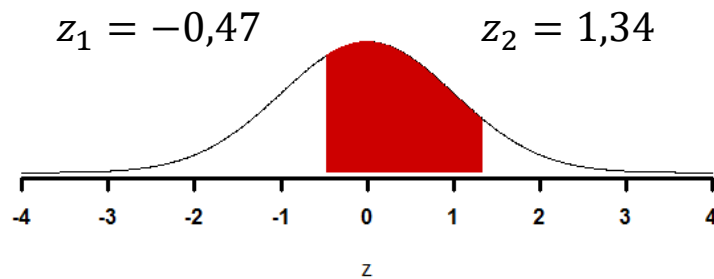
## Cálculo de probabilidades #3



$z$	0,00	0,01	0,02	0,03	0,04	0,05	0,06
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026

$$\begin{aligned}P(z > -0,24) &= 0,5 + P(z < 0,24) \\&= 0,5 + 0,0948 \\&= 0,5948 = 59,48\%\end{aligned}$$

# Cálculo de probabilidades #4



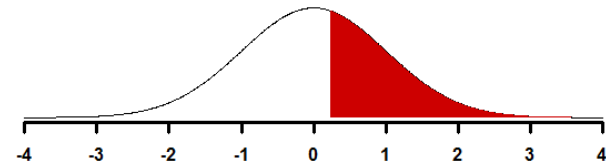
$$\begin{aligned}
 P(-0,47 < z < 1,34) &= P(z < 1,34) + P(z < 0,47) \\
 &= 0,4099 + 0,1808 \\
 &= 0,5907 \\
 &= 59,07\%
 \end{aligned}$$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441

## Probabilidades a partir de valores

Uma fábrica de chocolates comercializa barras que pesam em média 200g. Os pesos são normalmente distribuídos. Sabe-se que o desvio padrão é igual a 40g. Calcule a probabilidade de uma barra de chocolate escolhida ao acaso: (a) pesar mais que 230g; (b) pesar menos que 150g.

$$(a) \quad z = \frac{x - \mu}{\sigma} \quad \mu = 200g$$
$$\sigma = 40g$$



$$x = 230g$$

$$z = \frac{230 - 200}{40} = 0,75$$

$$P(x > 230) = P(z > 0,75)$$
$$= 0,500 - P(z < 0,75)$$
$$= 0,500 - 0,2734$$
$$= 0,2266 = 22,66\%$$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852

## Probabilidades a partir de valores

Uma fábrica de chocolates comercializa barras que pesam em média 200g. Os pesos são normalmente distribuídos. Sabe-se que o desvio padrão é igual a 40g. Calcule a probabilidade de uma barra de chocolate escolhida ao acaso: (a) pesar mais que 230g; (b) pesar menos que 150g.

$$(b) \quad z = \frac{x - \mu}{\sigma} \quad \begin{array}{l} \mu = 200g \\ \sigma = 40g \end{array}$$

$$x = 150g$$

$$z = \frac{150 - 200}{40} = -1,25$$

$$\begin{aligned} P(x < 150) &= P(z < -1,25) \\ &= 0,500 - P(z < 1,25) \\ &= 0,500 - 0,3944 \\ &= 0,1056 = 10,56\% \end{aligned}$$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177

## Exercício

O tempo de conserto de uma máquina é de  $X$  minutos. Estudos mostram que, em média, as máquinas são consertadas em 120min com variância de  $16\text{min}^2$ , com tempo de conserto normalmente distribuídos. Qual a probabilidade de, ao selecionar uma máquina aleatoriamente, seu conserto leve: (a) menos que 125 min? (b) mais que 110min? (c) menos que 115 min?

## Resumo

- ✓ Curva Normal tem formato de sino e é simétrica em relação à média
- ✓ Variável padronizada  $z = \frac{x - \mu}{\sigma}$
- ✓ Como consultar tabela da curva normal padronizada  $N(0,1)$

## Exercício 1

Faça uma análise estatística descritiva da base de dados ozone.data.txt, localizada na pasta db, da disciplina

## Exercício 2

Complete a tabela abaixo com um conjunto de 12 números entre 1 e 50.

Utilizando os dados criados por você, calcular:

- Média
- Moda
- Mediana
- Variância amostral
- Desvio padrão amostral
- Coeficiente de variação

--	--	--	--	--	--	--	--	--	--	--	--

### Regras

- Pelo menos um valor deve estar repetido.
- O conjunto de dados deve apresentar, pelo menos, 6 valores únicos.
- NÃO É PERMITIDO escolher todos os números iguais.
- **Trabalhos com os mesmos dados serão considerados inválidos.**



## Exercício 2.1

Escreva um código para ler um arquivo .txt, onde cada linha possui os 12 números escolhidos por um aluno.

Esse código deve verificar as regras estabelecidas e retornar todas as medidas de posição e variação solicitadas, com precisão de 4 casas decimais.

## Exercício 2.2

Escreva um código para gerar aleatoriamente valores para 50 alunos, com distribuição uniforme, de modo a respeitar as regras do trabalho.

# Testes de Hipóteses

Estimação e testes de hipóteses são os aspectos principais da Inferência Estatística

Um teste de hipótese decide se determinada afirmação sobre um parâmetro populacional é, ou não, apoiada pela evidência obtida de dados amostrais

Existem 2 hipóteses a serem levantadas:

- $H_0$ : hipótese nula
- $H_1$ : hipótese alternativa

Definição:  $H_0$  sempre levará sinal de igualdade

## Ponto Crucial

A diferença entre o valor alegado de um parâmetro populacional e o valor de uma estatística amostral pode ser razoavelmente atribuído à variabilidade amostral

OU

A discrepância é demasiada grande para ser vista dessa maneira

## Exemplos

Estudos indicam que a temperatura do corpo humano é  $37^{\circ}\text{C}$ .  
Pesquisadores brasileiros coletaram dados amostrais com  $\bar{x} = 36,78^{\circ}\text{C}$   
e distribuição aproximadamente normal.

Esses dados confirmam ou derrubam a afirmação inicial?

$$H_0: \mu = 37^{\circ}\text{C}$$

$$H_1: \mu \neq 37^{\circ}\text{C}$$

Uma companhia que fabrica cereais alega que o desvio padrão dos pesos de suas embalagens não é maior que 23g

$$H_0: \mu \leq 23g$$

$$H_1: \mu > 23g$$

## Tipos de Erros

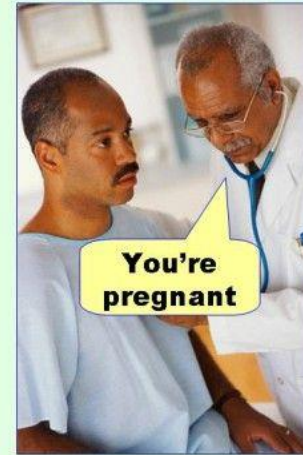
Por definição, sempre aceitamos no início que  $H_0$  é VERDADE.

A partir dessa premissa, podemos cometer dois tipos de erro:

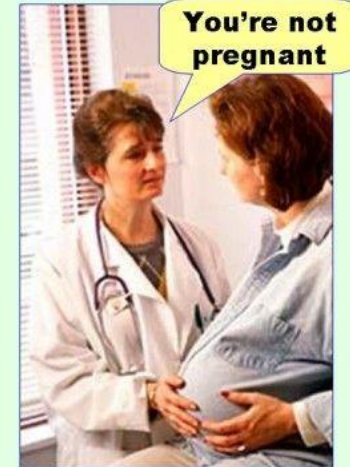
- Tipo I: rejeita  $H_0$ , quando  $H_0$  era verdadeira
- Tipo II: aceita  $H_0$ , quando  $H_0$  era falsa

# Tipos de Erros

**Type I error**  
(false positive)



**Type II error**  
(false negative)



Natureza de  $H_0$   
(Realidade)

Decisão

	VERDADEIRA	FALSA
REJEITAR	Erro Tipo I $\alpha$	✓ $1 - \beta$
ACEITAR	✓ $1 - \alpha$	Erro Tipo II $\beta$

## Nível de Significância ( $\alpha$ )

Erro máximo de se cometer o erro tipo I

$\alpha$  é fornecido pelo técnico

$\beta$  é análogo para erro tipo II, porém não é usado

## Exemplo

Uma máquina automática enche pacotes de café segundo uma distribuição normal com média  $\mu$  e desvio padrão 20g.

A máquina foi regulada para  $\mu=500\text{g}$ .

De meia em meia hora foi tirada uma amostra de 16 pacotes para verificar se o empacotamento está sob controle, isto é, se  $\mu=500\text{g}$ .

Se uma dessas amostras apresentasse  $\bar{x} = 492\text{g}$ , você pararia ou não o empacotamento para verificar se o ajuste da máquina está correto?

Considere nível de significância igual a 0,01

Informado no problema:

$$\mu = 500\text{g}$$

$$\bar{x} = 492\text{g}$$

$$\sigma = 20\text{g}$$

$$\alpha = 0,01$$



## Exemplo – Solução

1º Passo: Listar hipóteses

$$H_0: \mu = 500g$$

$$H_1: \mu \neq 500g$$

2º Passo: Escolha da distribuição

Se  $\sigma$  é conhecido, então distribuição Normal

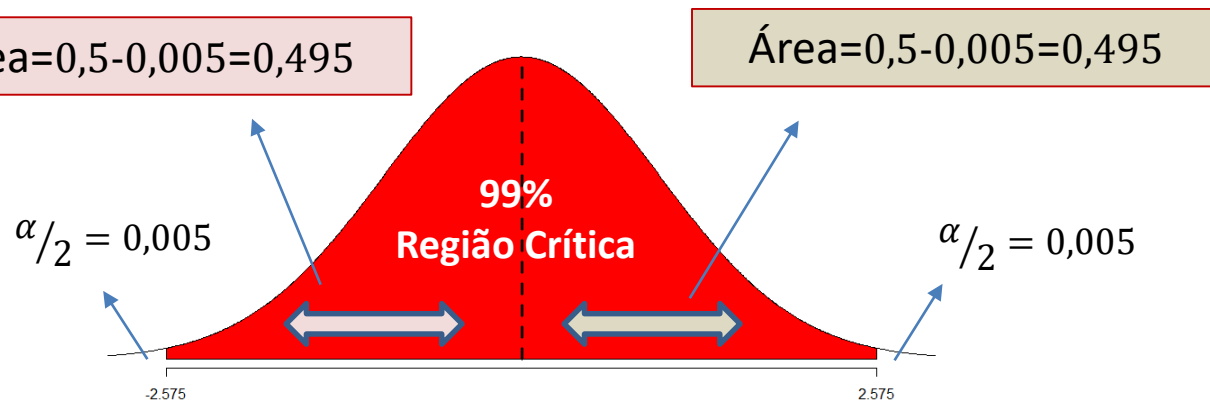
Caso contrário, distribuição t-Student

## Exemplo – Solução

3º Passo: Encontrar z de teste ( $z_t$ )

$$z_t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{492 - 500}{20 / \sqrt{16}} = -\frac{8}{5} = -1,6$$

4º Passo: Região crítica (depende de  $\alpha$ )

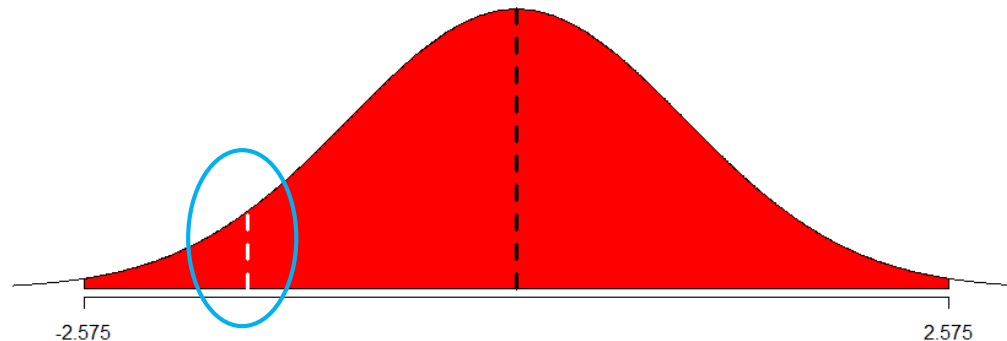


## Exemplo - Solução

### 5º Passo: Decisão do teste de hipótese

Se o valor da estatística do teste cair dentro da região crítica, aceita-se  $H_0$

Caso contrário, rejeitamos  $H_0$



Neste caso, o  $z_{teste} = -1,6$  cai dentro da região crítica

Logo, **aceitamos  $H_0$** , de que a média amostral é estatisticamente igual à média populacional

Dessa maneira, não devemos parar o empacotamento para verificar o ajuste das máquinas, pois ele está correto

## Exemplo 2

Um determinado fabricante de rações para aves produz um tipo especial de mistura. Ele alega que uma embalagem de 20kg conterá, em média, 4,5kg de um determinado composto. Esse composto não pode ser superior a 4,5kg.

Uma pesquisa realizada por um cliente, utilizando-se de uma amostra de 190 embalagens de 20kg, revelou uma quantidade média de 4,8kg com um desvio padrão de 0,3kg. Considerando um nível de confiança de 96%, é possível aceitar a alegação do fabricante?

## Exemplo 2 – Solução

1º Passo: Listar hipóteses

$$H_0: \mu \leq 4,5\text{Kg}$$

$$H_1: \mu < 4,5\text{Kg}$$

2º Passo: Escolha da distribuição

Se  $n > 30$ , então distribuição Normal

Caso contrário, distribuição t-Student

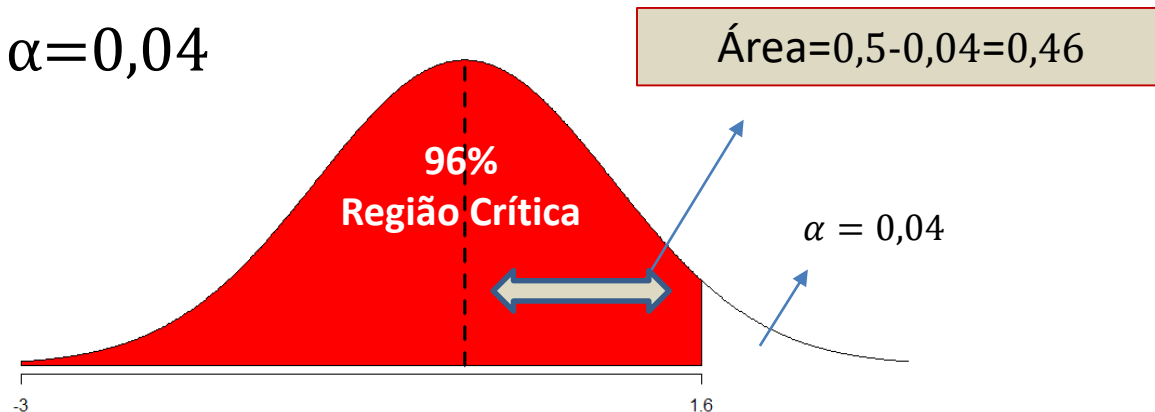
## Exemplo 2 – Solução

3º Passo: Encontrar z de teste ( $z_t$ )

$$z_t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{4,8 - 4,5}{0,3 / \sqrt{190}} = \frac{0,3}{0,3} \sqrt{190} = 13,78$$

4º Passo: Região crítica (depende de  $\alpha$ )

Neste caso,  $\alpha = 0,04$

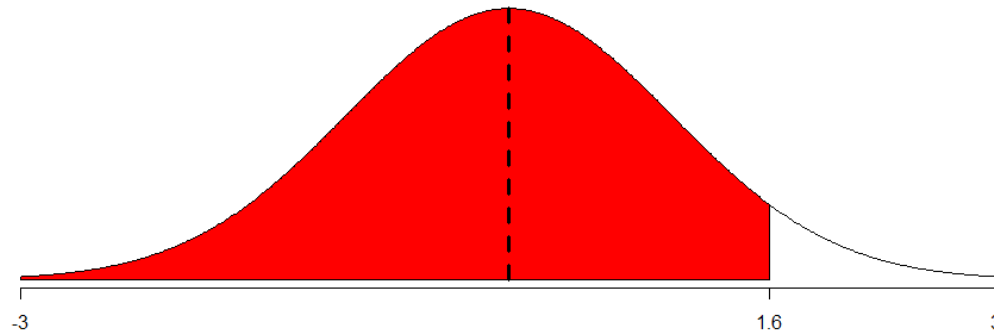


## Exemplo 2 - Solução

### 5º Passo: Decisão do teste de hipótese

Se o valor da estatística do teste cair dentro da região crítica, aceita-se  $H_0$

Caso contrário, rejeitamos  $H_0$



Neste caso,  $z_{\text{teste}} = 13,78$  cai fora da região crítica

Logo, **rejeitamos  $H_0$** , de que a média amostral é estatisticamente menor ou igual à média populacional

Dessa maneira, não aceitamos a alegação do fabricante

## Exercício

Uma empresa de pesquisas econômicas alega, que em uma determinada região da cidade, a renda familiar média anual seria, no mínimo, igual a \$100.000. Porém, uma amostra formada por 60 famílias da região apresentou renda média anual igual a \$93.300 com desvio padrão igual a \$13.400. Assumindo um nível de significância igual a 5%, seria possível concordar com a alegação formulada?

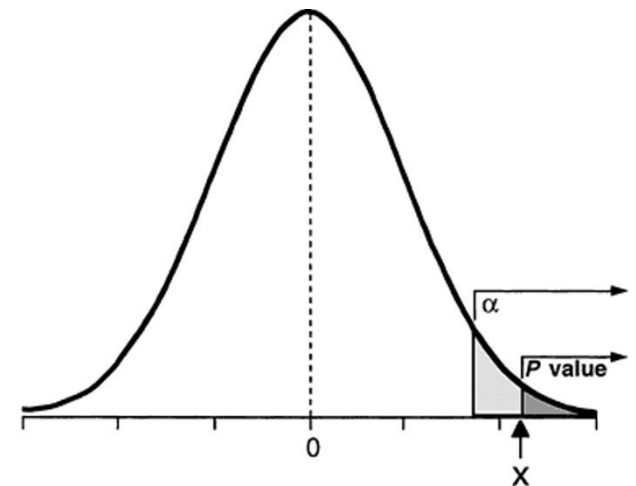


# Teste de normalidade

No código abaixo simulam-se dados com distribuição log-normal e sua normalidade é testada. Espera-se que o teste falhe:

```
from scipy.stats import shapiro  
x = np.random.randn(100)  
stats, p = shapiro(x)
```

(0.98558509349823, 0.34968388080596924)



A hipótese nula é de que os dados são normalmente distribuídos

Se o *valor-p* é menor do que o nível de significância ( $\alpha$ ), então a hipótese nula é rejeitada

Logo, os dados NÃO SÃO REJEITADOS → são normalmente distribuídos

# Comparação da Média de um Conjunto de Dados com uma Constante

```
from scipy.stats import norm
from scipy import stats

x1 = norm.rvs(loc=10000, scale=1000, size=50)
x2 = norm.rvs(loc=1000, scale=100, size=50)
x3 = norm.rvs(loc=1000, scale=100, size=50)

print(x1.mean())
print(x2.mean())
print(x3.mean())

stats.ttest_1samp(x1, 0)
Ttest_1sampResult(statistic=70.689068547402044,
pvalue=5.551055102945442e-51)
```

# Comparação da Média de Dois Conjuntos de Dados

## Comparando x1 com x2

```
stats.ttest_ind(x1,x2)
```

```
Ttest_indResult(statistic=72.022883333939518,  
pvalue=1.1191380960232934e-86)
```

## Comparando x2 com x3

```
stats.ttest_ind(x2,x3)
```

```
Ttest_indResult(statistic=0.10080661932376216,  
pvalue=0.91990992237400093)
```

