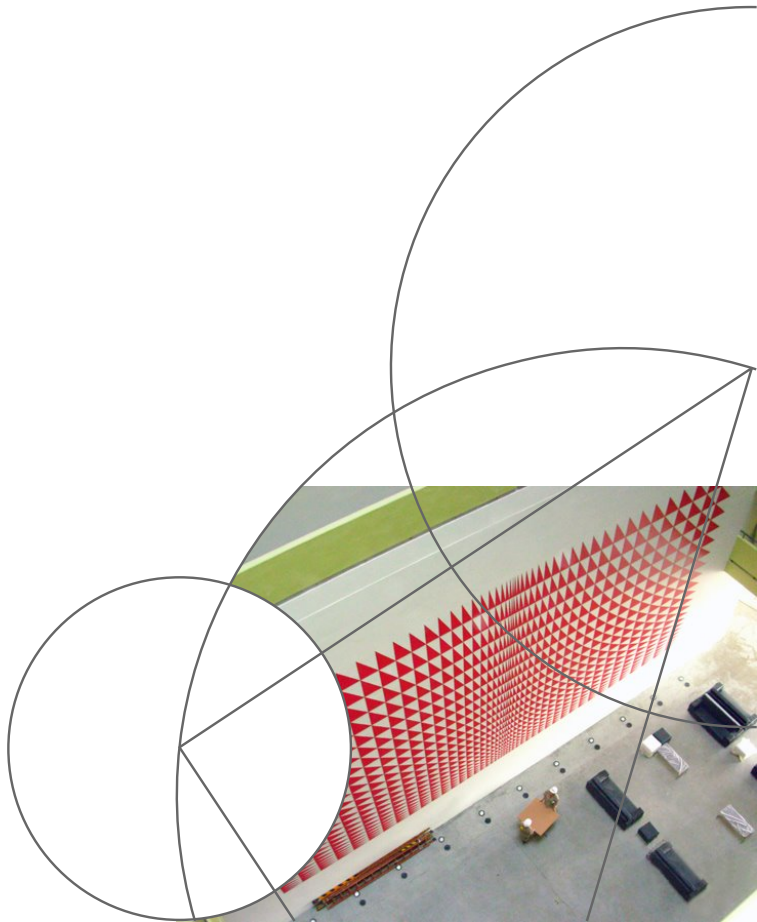
UFABC

Estatística Descritiva

Centro de Matemática, Computação e Cognição



Estatística

- A **Estatística** é um ramo da matemática dedicado ao estudo de técnicas e métodos para resumir, analisar e interpretar dados de observações e realizar inferências a partir desses dados.

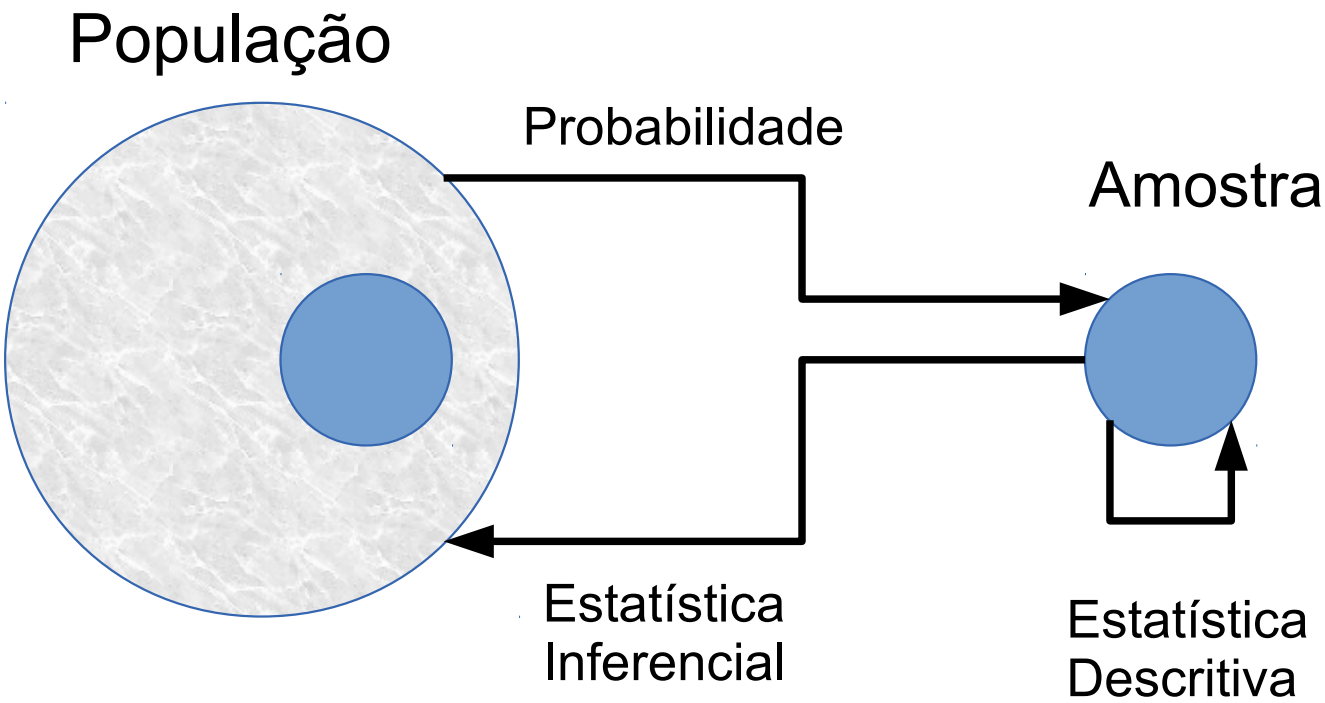
- A **Estatística Descritiva** é a parte da estatística que se dedica ao estudo dos procedimentos utilizados para resumir, organizar e fazer sentido de um conjunto de pontos ou as observações.
- A **Estatística Inferencial** é a parte da estatística que se dedica ao estudo dos procedimentos utilizados que permitem inferir ou generalizar as observações feitas com amostras de uma população maior de que foram selecionados.

População

- Uma **população** é o conjunto de todos os indivíduos, itens, ou os dados de interesse.
- Uma característica (geralmente numérico), que descreve uma população é referido como um **parâmetro da população**.

Amostra

- Uma **amostra** é definido como um conjunto de indivíduos selecionados, itens, ou os dados extraídos um população de interesse.
- Uma característica (geralmente numérico) que descreve um exemplo é referido como um **estatística da amostra.**


População
Probabilidade
Amostra
Estatística
Inferencial
Estatística
Descritiva

Tipos de Amostragem

Amostragem Probabilística: é uma amostragem :

- envolve a seleção aleatória em algum ponto da amostragem;
- em que cada elemento da população tem uma probabilidade maior do que zero de ser selecionada para a amostra;
- a probabilidade de cada elemento ser escolhido para a amostra pode ser determinada com precisão.

A combinação destas características permite produzir estimativas imparciais de totais populacionais.

Amostragem probabilística inclui:

- amostragem aleatória simples
- amostragem estratificada, com probabilidade proporcional ao estrato;

Amostragem não probabilística: é qualquer método de amostragem, onde alguns elementos da população não tem chance de seleção , ou onde a probabilidade de seleção não pode ser determinada com precisão.

- Ela envolve a seleção de elementos com base em suposições sobre a população de interesse, que forma os critérios para a seleção.
- Como a seleção de elementos é não aleatória, a amostragem não-probabilística não permite a estimativa do erro de amostragem.
- Estas condições dão origem a um viés da exclusão, colocando limites a quantidade e a qualidade de informação que uma amostra pode fornecer sobre a população.

Tipos de Dados

- **Dados Qualitativos:** atributos e rótulos não numéricos.
- **Dados Quantitativos:** consistem de medidas numéricas
  - **Discretos**
  - **Contínuos**

Tipos de Escalas

- **Escala nominal:** Variáveis expressas na escala nominal não possuem uma ordem natural. **Exemplos**: Matrículas de automóveis, códigos postais, estado civil, sexo, cor dos olhos, código de artigo, código de barras.
- **Escala ordinal:** A variável utilizada para medir uma determinada característica identifica que é pertencente a uma classe e pressupõe que as diferentes classes podem ser ordenadas. **Exemplos:** Escala social, escalas usadas na medida de opiniões.

- **Escala métrica:** além de ser possível ordenar os indivíduos, é também feita uma quantificação das diferenças entre eles. As escalas métricas dividem-se em dois subtipos:
  - **Escala intervalar:** é possível quantificar as distâncias entre as medições mas onde não há um ponto nulo natural e uma unidade natural. Exemplo clássico são as escalas de temperatura Celsius e Fahrenheit, onde não se pode assumir um ponto 0 (ponto de nulidade) ou dizer que a temperatura X é o dobro da temperatura Y.
  - **Escala racional** A escala onde não só é possível quantificar as diferenças entre as medições como como existe um ponto nulo natural (absoluto). Isto permite o quociente de duas medições, independentemente da unidade de medida. Exemplos de escalas de razão são a idade, salário, preço, volume de vendas, distâncias, escala Kelvin de temperatura.

Tipo de Escala	**Operações Matemáticas Permitidas**	**Exemplos**	**Medida de tendência Central**
Nominal	$=, \neq$	Gênero, Nacionalidade	Moda
Ordinal	$=, \neq <, >$	Opinião ("concordo completamente" ... "discordo completamente")	Mediana
Intervalar	$=, \neq <, >, +, -$	Data	Média Aritmética
Racional	$=, \neq <, >$ $, +, -, \times \div$	Idade	Média Aritmética ou Geométrica

Medidas

- **medidas de posição ou de tendência central**: são medidas que descrevem um "centro" em torno do qual as medições dos dados estão distribuídas.
- **medidas de dispersão ou variabilidade**: são medidas que descrevem o espalhamento dos dados ou quanto as medições se afastam do centro.

Média Aritmética

Consideremos uma coleção formada por n números: $x_1, x_2, \ldots, x_n$, a média aritmética, denotada por $\overline{x}$ é definida como:

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

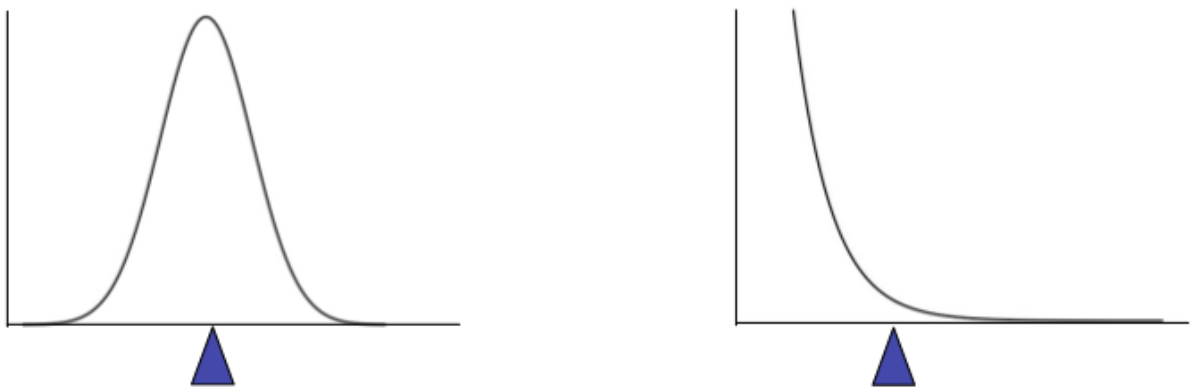


Figura: Média Aritmética

Média Aritmética Ponderada

Consideremos uma coleção formada por n números: $x_1, x_2, \ldots, x_n$, de forma que cada um esteja sujeito a um peso, respectivamente, $p_1, p_2, \ldots, p_n$.
A **média aritmética** ponderada desses n números é a soma dos produtos de cada um multiplicados por seus respectivos pesos, dividida pela soma dos pesos, isto é:

$$\overline{p} = \frac{x_1 p_1 + x_2 p_2 + \ldots. + x_n p_n}{p_1 + p_2 + \ldots. + p_n}$$

Outras Médias

- Média Geométrica

$$G = \left(\prod_{i=1}^{n} a_i \right)^{1/n} = (a_1 \cdot a_2 \cdots a_n)^{1/n} = \sqrt[n]{a_1 \cdot a_2 \cdots a_n}$$

- Média Harmônica

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}, \qquad \text{com } x_i > 0$$

Desigualdades:

Média Aritmética $\geqslant$ Média Geométrica $\geqslant$ Média Harmõnica

Moda

- A moda é o valor que detém o maior número de observações, ou seja, o valor ou valores mais frequentes,
- A moda não é necessariamente única, ao contrário da média ou da mediana.
- É especialmente útil quando os valores ou observações não são numéricos, uma vez que a média e a mediana podem não ser bem definidas.

Exemplos

- A série {1, 3, 5, 5, 6, 6} apresenta duas modas (BIMODAL): 5 e 6
- A série {1, 3, 2, 5, 8, 7, 9} não apresenta moda (AMODAL).

Mediana

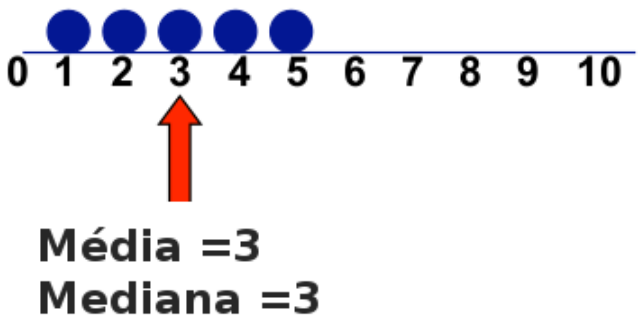
- **Mediana** é uma medida de tendência central. A mediana de um conjunto de dados ordenados caracteriza-se por separar a metade inferior da amostra, população ou distribuição de probabilidade, da metade superior.
- Concretamente, 1/2 dos dados terá valores inferiores ou iguais à mediana e 1/2 dos dados terá valores superiores ou iguais à mediana.

No caso de **dados ordenados** (crescentes ou decrescentes) de amostras de tamanho n,

- se n for ímpar, a mediana será o elemento central $\frac{(n+1)}{2}$.
- se n for par, a mediana será o resultado da média aritmética entre os elementos $\frac{n}{2}$ e $\frac{n}{2}+1$.

Qual é melhor?

- A média é melhor para distribuições simétricas sem valores extremos.
- A mediana é útil para distribuições distorcidas ou de dados com discrepantes.



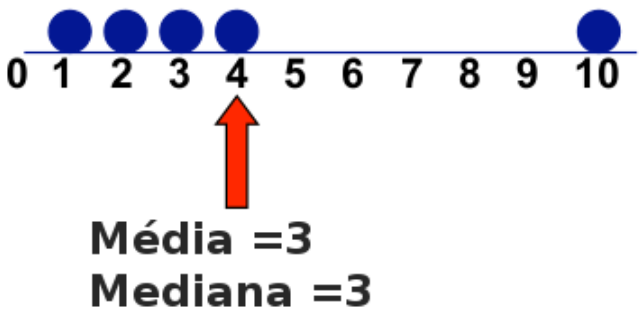


Figura:

Amplitude

A **amplitude** é uma medida de dispersão que pode ser definida como a diferença entre o maior valor e o menor valor menor de um conjunto de observações.
É a medida de dispersão mais simples. Será denotada por A e é calculada da seguinte maneira:

$$A = X_{max} - X_{min}$$

A amplitude é uma medida imperfeita de variação, pois:

- seu cálculo utiliza apenas as valores extremos, não avaliando os valores intermediários;
- seu valor tende a crescer com o aumento do número de observações.

Variância

A **variância da população** y_i onde $i = 1, 2, ..., n$ é dada por :

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

onde μ é a média da população.

Um método comum de estimar a variância da população é através de amostras. Quando estimando a variância da população usando uma amostra $\{x_i$, com $i = 1, 2, \ldots, n\}$, a fórmula seguinte é um estimador não enviesado, denominado **variância da amostra**:

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

onde $\overline{x}$ é a média da amostra.

Desvio Padrão

O desvio padrão amostral é calculado como:

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

Independentemente da forma como os dados são distribuídos, uma certa porcentagem de valores deve estar dentro k desvios padrão a partir da média:

ao menos	dentro de
75%	$\mu \pm 2\sigma$
89%	$\mu \pm 3\sigma$

Para muitos conjuntos de dados, especialmente se o seu histograma é em forma de sino:

ao menos	dentro de
68%	$\mu \pm \sigma$
95%	$\mu \pm 2\sigma$

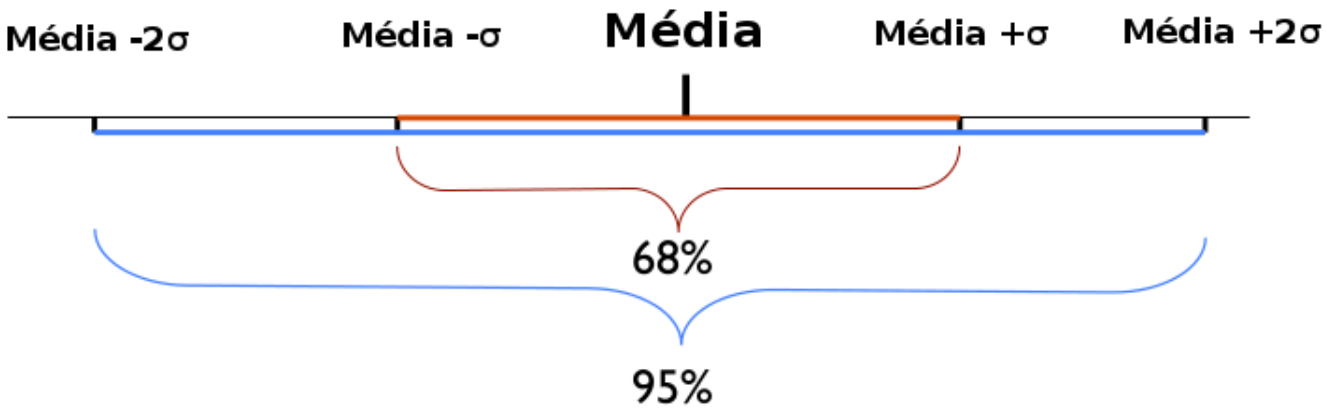


Figura:

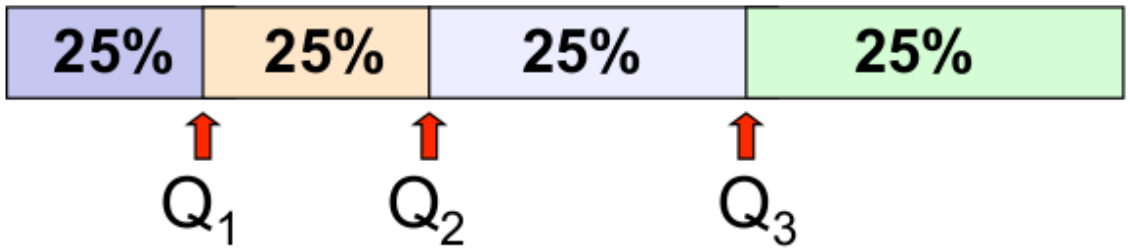
Coeficiente de Variação

O **coeficiente de variação** é uma medida de dispersão empregada para estimar a precisão de experimentos e representa o desvio-padrão expresso como porcentagem da média.

$$c_v = \frac{\sigma}{\mu}$$

Sua principal qualidade é a capacidade de comparação entre diferentes distribuições.

Quartis



- O primeiro quartil, Q_1, é o valor para o qual 25% da observações são menores e 75% são maiores
- O segundo quartil, Q_2, é a mediana (50% dos valores são menores de 50% dos valores são maiores)
- O terceiro quartil, Q_3, é o valor para o qual 75% da observações são menores e 25% são maiores

O intervalo interquartil, denotado Iq ou IQR, é uma medida de de dispersão é definido como:

$$IQR = Q_3 - Q_1$$

Outliers

Um **outlier**, é uma observação que parece desviar-se acentuadamente do outros membros da amostra em que ela ocorre. Não existe uma definição rígida matemática do que constitui um outlier.
Determinar se uma observação é um outlier é fundamentalmente um exercício subjetivo.

Uma possível caracterização matemática de outlier pode ser dada usando intervalos interquartis: outliers são os pontos que são menores que $Q_1 - 1,5 \cdot IQR$ ou maiores que $Q_3 + 1,5 \cdot IQR$.

Box-Plot

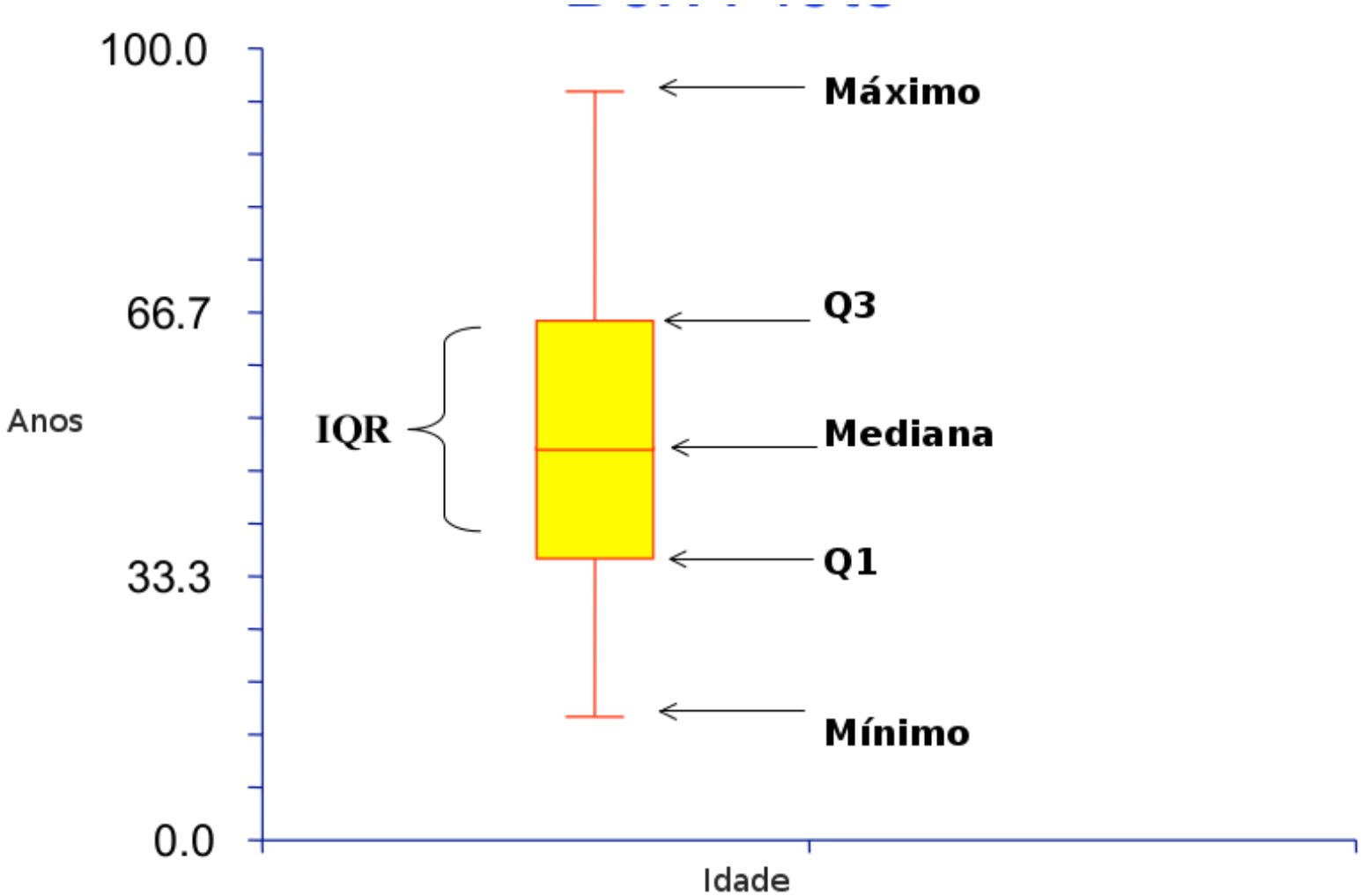


Figura: Box-Plot

- Box-Plot é uma ferramenta gráfica para apresentar as diferenças entre grupos de dados sem fazer qualquer suposição sobre a distribuição estatística subjacente.
- Os espaçamentos entre as diferentes partes da caixa de ajuda indicar o grau de dispersão e assimetria nos dados, e identificar outliers.

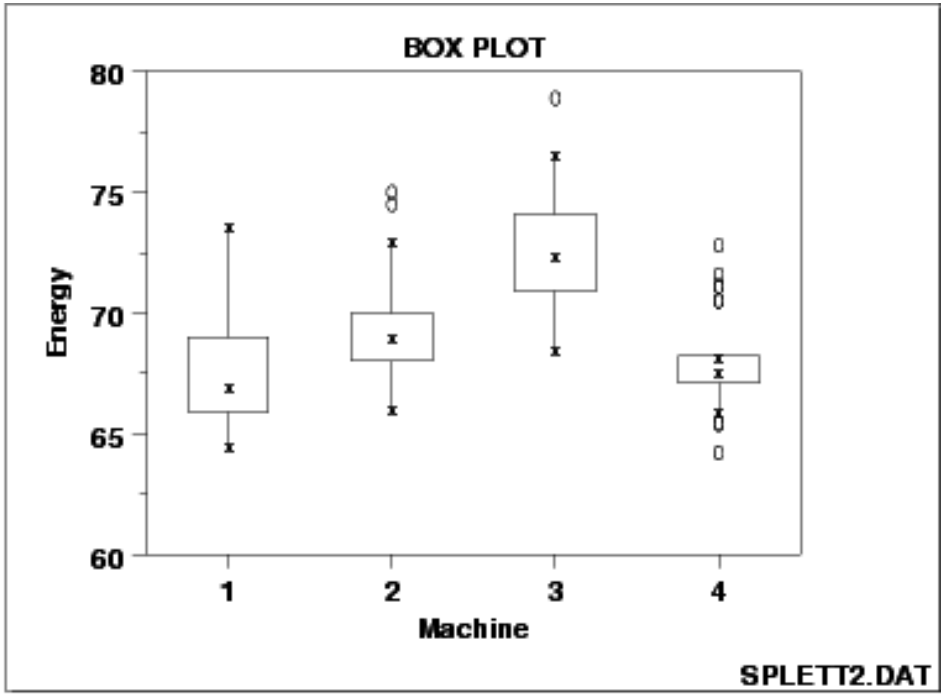
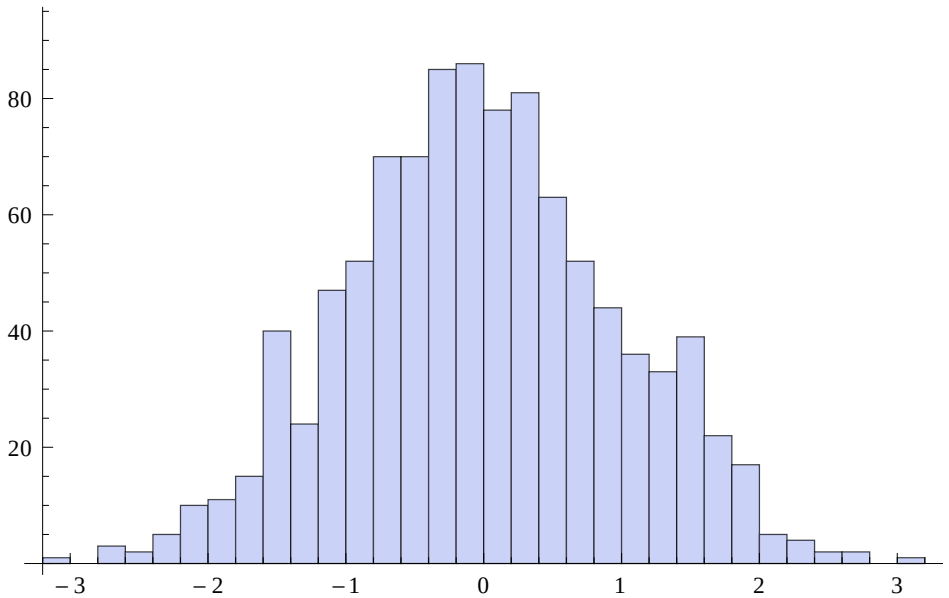


Figura: Este gráfico de caixa, comparando quatro máquinas para a produção de energia, mostra que a máquina tem um efeito significativo sobre a energia no que diz respeito tanto à localização e variação.

Histograma

Um **Histograma** ou **Distribuição de Frequências** é uma representação gráfica na qual um conjunto de dados é agrupado em classes uniformes, representado por um retângulo cuja base horizontal são os intervalos das classes e a altura vertical representa a frequência (absoluta ou percentual) com que os os valores desta classe estão presente no conjunto de dados.



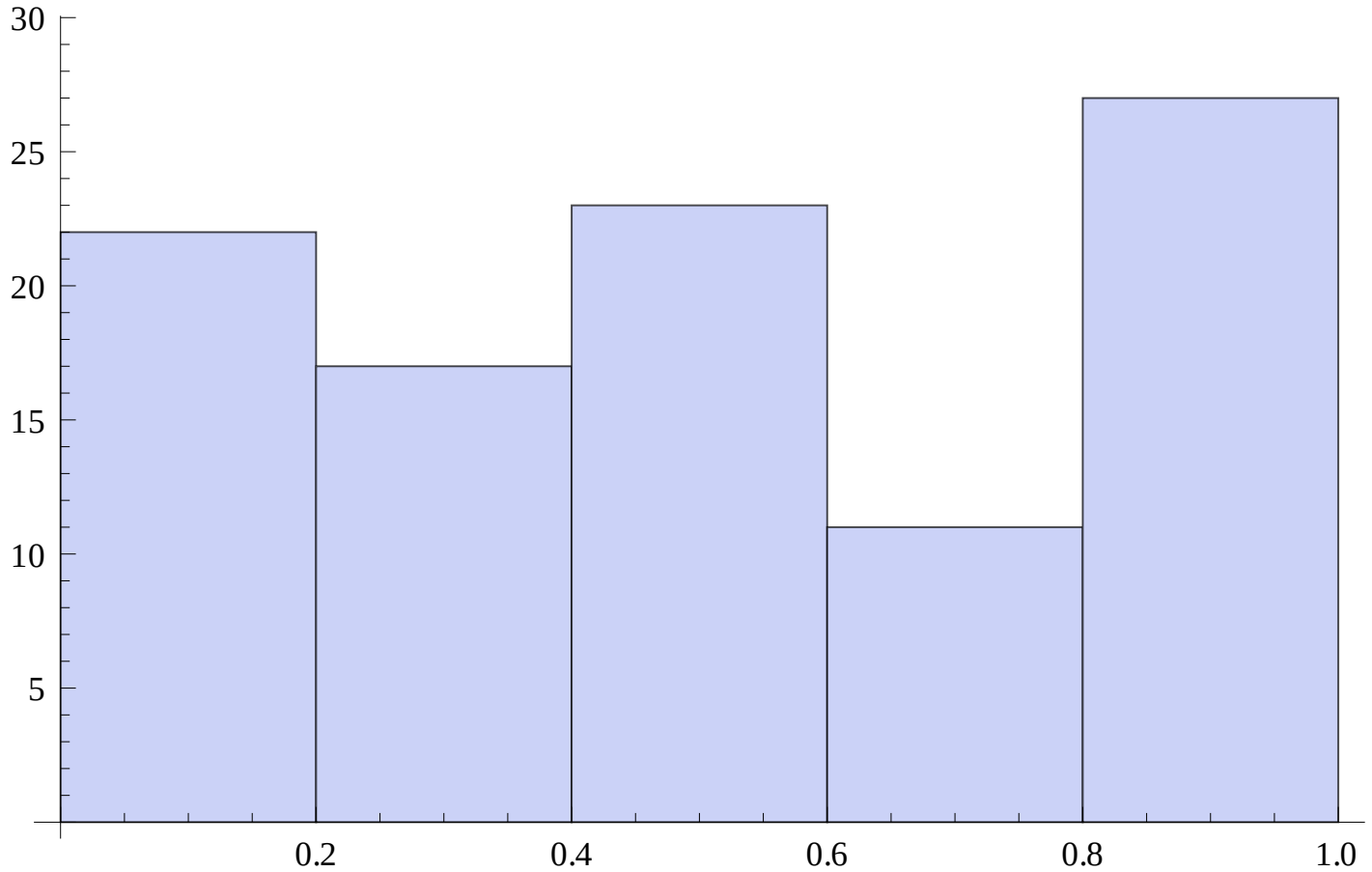


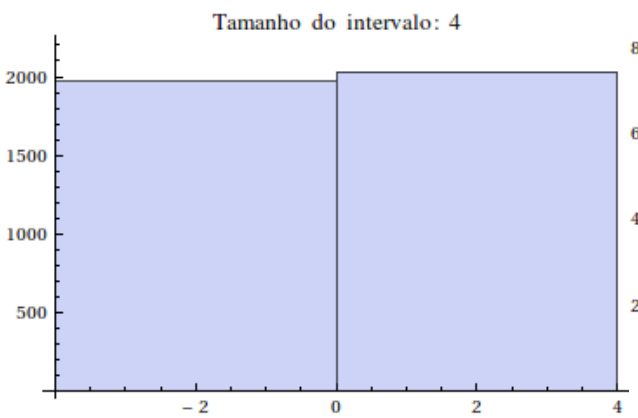
Figura: Histograma de saída de um gerador de números aleatórios.

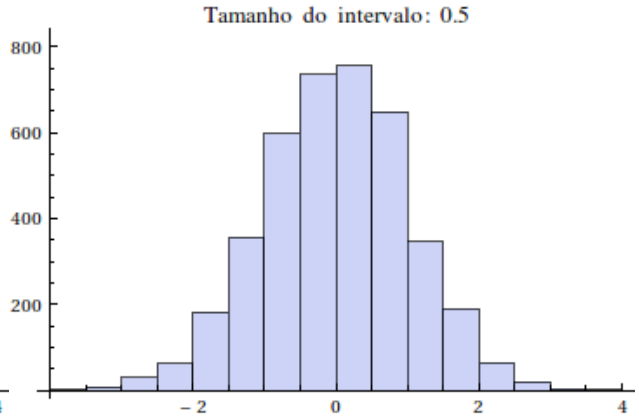
Os histogramas servem para avaliar aproximadamente a distribuição de probabilidade de uma determinada variável, descrevendo as frequências de observações que ocorrem em determinadas faixas de valores

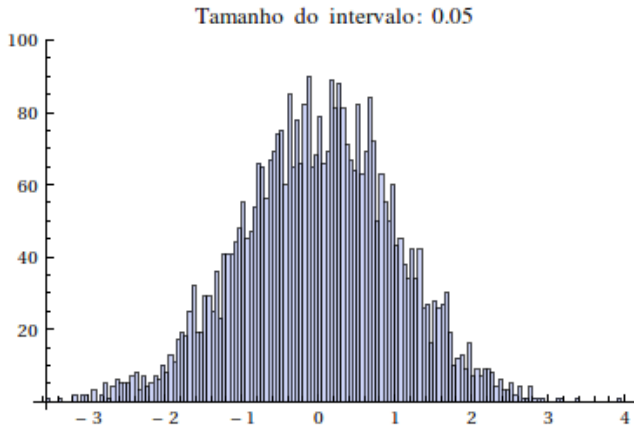
O histograma pode ser usado para verificar graficamente as seguintes propriedades:

- centro (isto é, a localização) de dados;
- dispersão (isto é, a escala) de dados;
- assimetria dos dados;
- presença de outliers;

Influência do número de classes e do tamanho do intervalo no histograma







Escolha do número de classes

Seja $k=$ número de classes num histograma. Não existe uma escolha perfeita para k

- tamanhos diferentes de classes podem revelar características diferentes dos dados.
- algumas tentativas foram feitas para determinar um número ideal de classes, mas estes métodos geralmente fazem suposições fortes sobre a forma da distribuição.

Há, no entanto, diversas orientações úteis:

- Fórmula de Sturges:

$$k = \lceil 1 + 3,322 \log_{10}(n)) \rceil$$

- Escolha da raiz quadrada (utilizada pelo Excel, por exemplo):

$$k = \lceil \sqrt{n} \rceil$$

sendo $\lceil x \rceil$ a função teto que dado um número real x retorna o menor número inteiro maior ou igual a x

Para dados quantitativos contínuos, dividimos a faixa de variação dos dados em intervalos de classes.
Para k classes dividimos o intervalo entre X_{min} e X_{max} usando $k+1$ pontos l_i com $i = 0, \ldots, k$
O intervalo ou classe pode ser representado das seguintes maneiras:

- $l_i \vdash l_{i+1}$, onde o limite inferior da classe é incluído na contagem da frequência absoluta, mas o superior não;
- $l_i \dashv l_{i+1}$, onde o limite superior da classe é incluído na contagem, mas o inferior não.

Embora não seja necessário, os intervalos são frequentemente construídos de modo que todos tenham larguras iguais, o que facilita as comparações entre as classes. Para k classes: Nesse caso:

- $\Delta = \frac{X_{max} - X_{min}}{k}$
- $l_i = X_{min} + i \cdot \Delta$ para $i = 0, \ldots, k$

Exemplo:

Altura: 1,57 1,60 1,61 1,72 1,72 1,75 1,76 1,77 1,77 1,80 1,80 1,81 1,83 1,87 1,88 1,91 1,91 1,95 1,97 1,99 1,99 2,02

Nesse caso:

- $n = 22$
- Por Sturges $k = 6$
- $\Delta = \frac{Xmin - X_{min}}{6} = 0,08$

Classe	Frequência
1,57 ⊢1,65	3
1,65 ⊢1,73	2
1,73⊢ 1,81	6
1,81⊢ 1,89	4
1,89⊢ 1,97	4
1,97 ⊢2,05	3

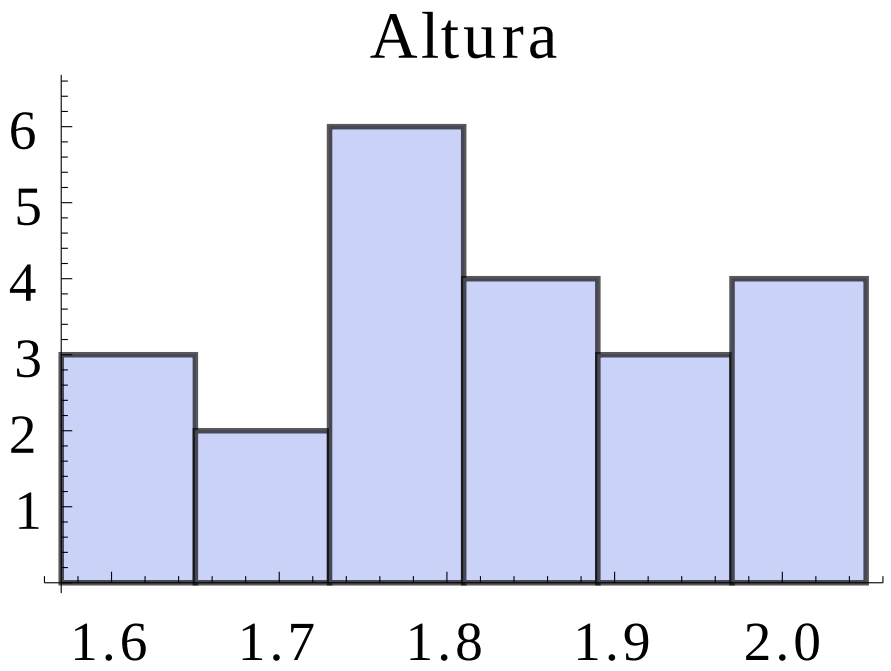


Figura: Histograma de altura

Correlação e Dependência

Em estatística

- **dependência:** refere-se a qualquer relação estatística entre duas variáveis aleatórias ou dois conjuntos de dados. Formalmente, dependência refere-se a qualquer situação em que as variáveis aleatórias não forem probabilisticamente independentes
- **correlação:** refere-se a uma ampla classe de relações estatísticas envolvendo dependência. Tecnicamente, refere-se a qualquer um dos vários tipos relação entre os valores médios.

Coeficiente de Correlação de Pearson

O mais comum dos coeficientes de correlação destes é o coeficiente de correlação de Pearson, que é sensível apenas a uma relação linear entre duas variáveis (que pode existir mesmo quando uma variável é função não linear da outra).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2 \sum_{i=1}^{n}(y_i-\overline{y})^2}},$$

ou ainda

$$r_{xy} = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

A correlação de Pearson é

- +1 , no caso de uma relação de linear positiva perfeita (crescente) (correlação)
- -1 , no caso de uma relação linear negativa perfeita (decrescente) (anti-correlação),
- algum valor entre -1 e 1 em todos os outros casos, indicando o grau de dependência linear entre as variáveis.
- Quando o coeficiente se aproxima de zero, há um relacionamento fraco ou as variáveis não estão correlacionadas.
- Quanto mais próximo o coeficiente é de -1 ou 1 , mais forte é a correlação entre as variáveis.
- Se as variáveis são independentes, o coeficiente de correlação de Pearson é 0, mas o inverso não é verdadeiro, porque o coeficiente de correlação detecta apenas as dependências lineares entre duas variáveis.

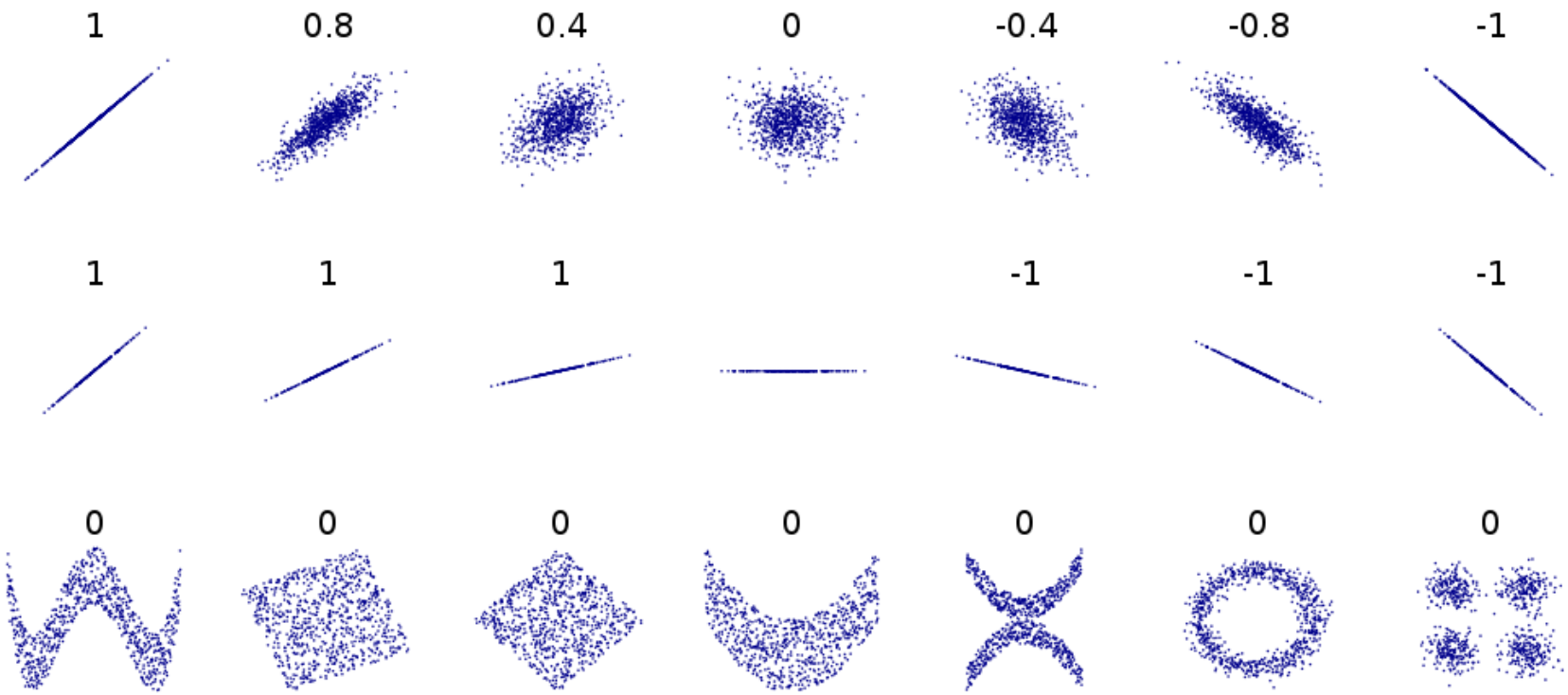
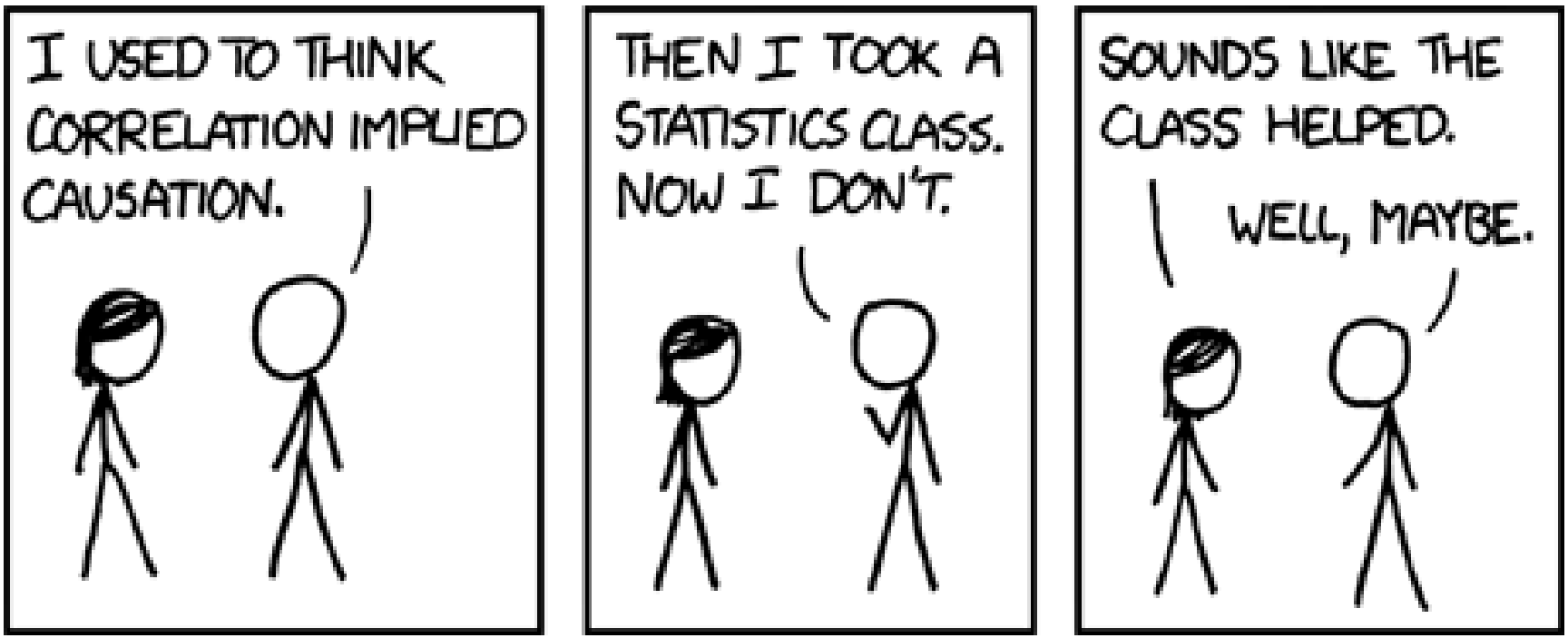


Figura: Correlação (Wikipedia)

XKCD



Fonte: xkcd.com