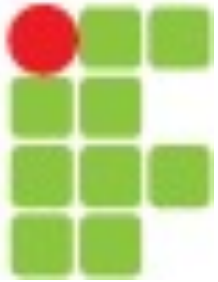


Instituto Federal de Santa Catarina
Campus Florianópolis

Introduçao a Estatística

Prof. Glauco Cardozo

glauco.cardozo@ifsc.edu.br

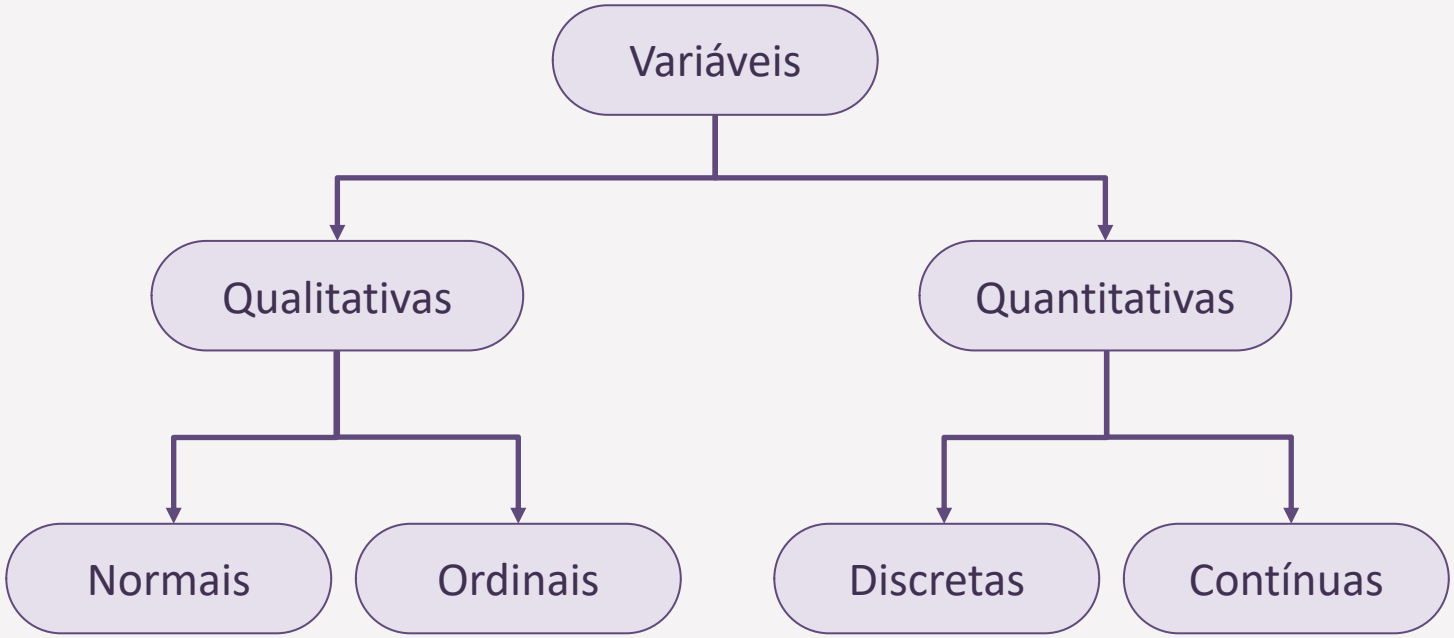
Estatística

Estatística é a ciência que utiliza as teorias probabilísticas para explicar a frequência da ocorrência de eventos, tanto em estudos observacionais quanto em experimentos para modelar a aleatoriedade e a incerteza de forma a estimar ou possibilitar a previsão de fenômenos futuros, conforme o caso. Wikipédia

A **estatística descritiva** é um ramo da estatística que aplica técnicas para descrever e sumarizar um conjunto de dados. Para isso são coletadas amostras aleatórias e representativas da variabilidade dessa população.

Estatística Descritiva

Cada dado dessa amostra pode ser de diferentes tipos:



O tipo de tratamento e de gráfico escolhido dependem do tipo dessa variável.

Estatística Descritiva

Variáveis Qualitativas: ordinais ou nominais

As variáveis qualitativas **ordinais** expressam uma relação de posicionamento e ordem. Alguns exemplos são: escolaridade, estágio de doença, classe social, etc.

As variáveis qualitativas **nominais** são as que não expressam nenhuma ordem. Como exemplos temos: sexo, profissão, religião, etc.

Estatística Descritiva

Variáveis Qualitativas: ordinais ou nominais

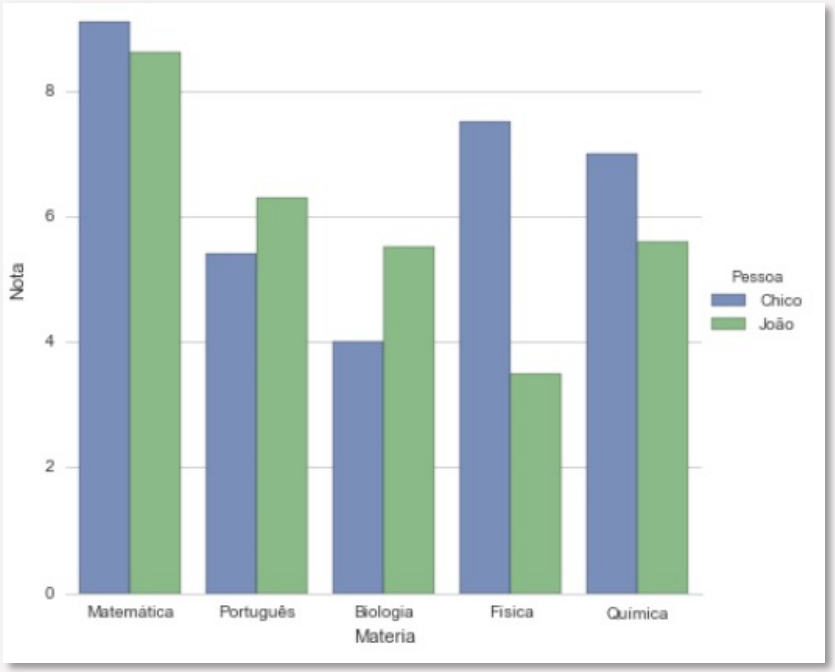
Mesmo quando as variáveis qualitativas são transformadas em números elas continuam representando categorias, logo, elas continuam sendo categóricas!

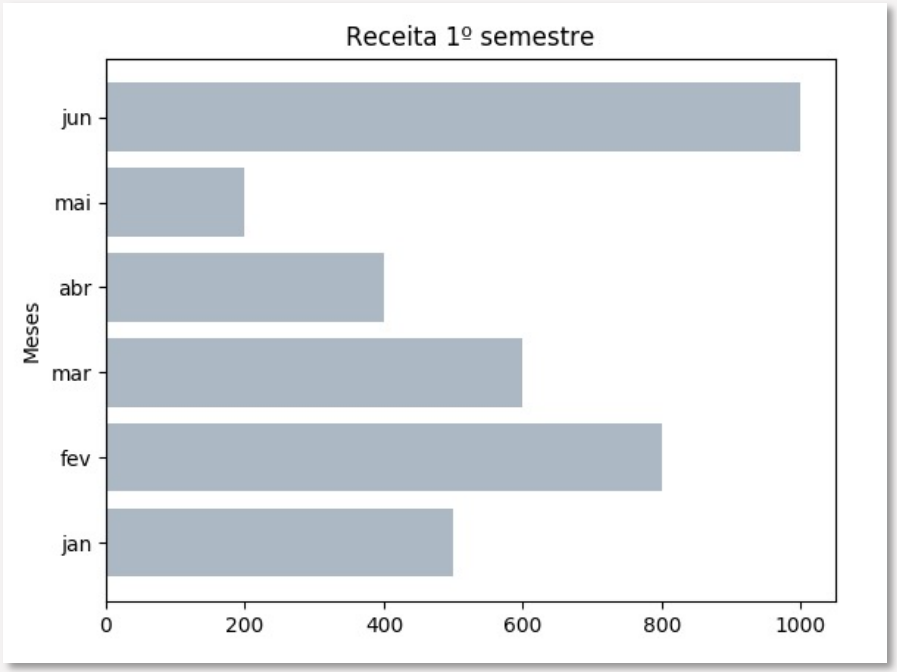
São melhor apresentadas por:

- **Gráficos de barra verticais ou horizontais**: representam sequências simples de valores e a frêquencia absoluta ou relativa destes;
- **Gráficos de torta/pizza**: fazem divisão por setores e proporções.

Estatística Descritiva

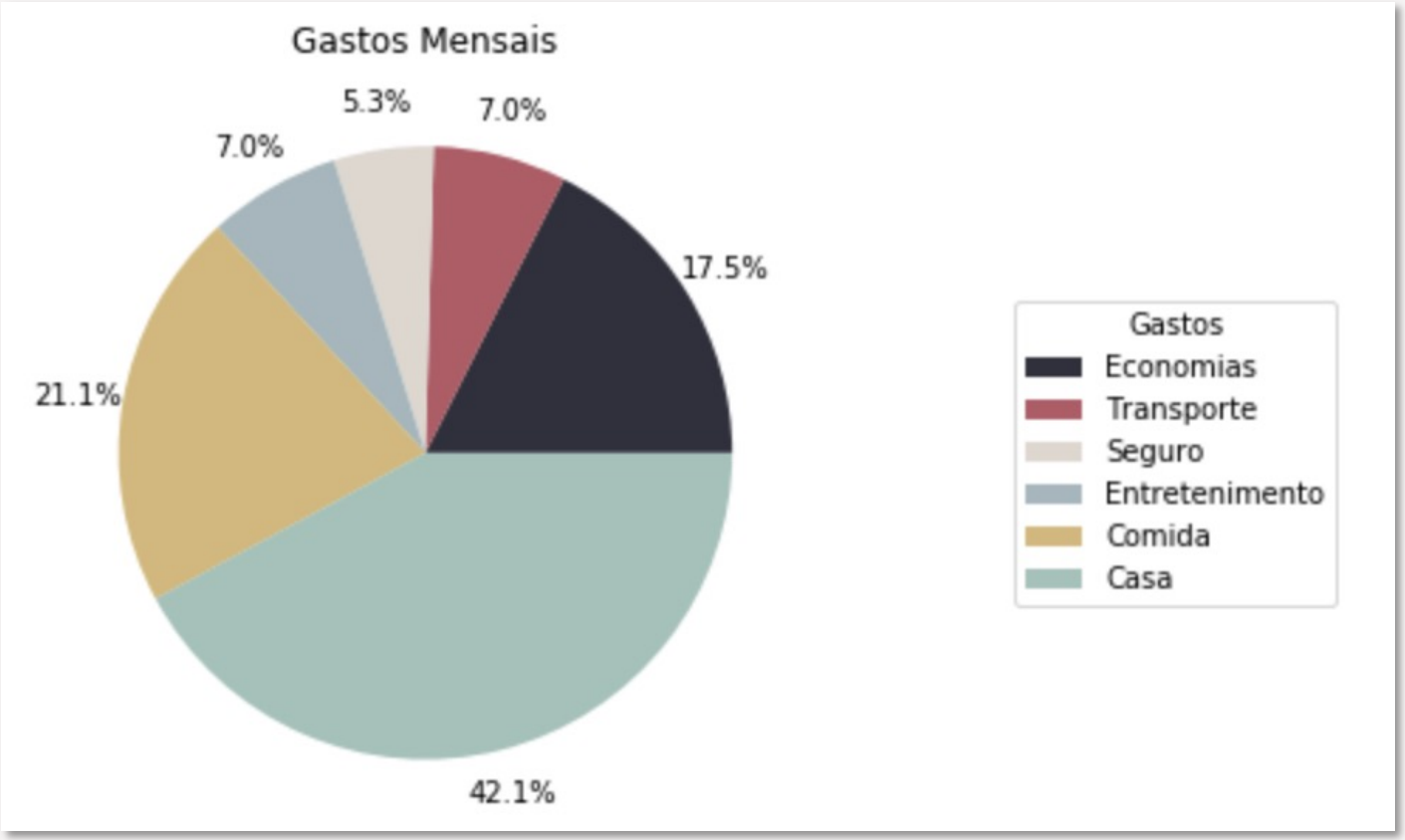
- **Gráficos de barra verticais ou horizontais**





Estatística Descritiva

- **Gráficos de torta/pizza**



Estatística Descritiva

Variáveis Quantitativas: discretas ou contínuas

As variáveis **discretas** são resultantes de um processo de contagem e, logo, são representadas pelos números naturais. Alguns exemplos são: número de filhos, número de dias sem chuva, número de acessos a uma plataforma, etc.

Estatística Descritiva

Variáveis Quantitativas: discretas ou contínuas

As variáveis **contínuas** são resultantes de um processo de medição; logo, representadas pelos números reais. Como exemplos temos: altura, peso, salário, vazão de um rio, etc.

Estatística Descritiva

Variáveis Quantitativas: discretas ou contínuas

Esses valores amostrados podem ser ordenados e apresentados em **tabelas de frequências**.

- **Frequência simples (f):** contagem dos elementos, frequência com que determinado elemento ocorre na amostra;
- **Frequência simples acumulada (fac):** mostra quantos dados apresentam valores menores ou iguais ao elemento analisado;

Estatística Descritiva

Variáveis Quantitativas: discretas ou contínuas

Esses valores amostrados podem ser ordenados e apresentados em **tabelas de frequências**.

- **Frequência relativa (fr):** a porcentagem referente à frequência simples, estima a probabilidade de ocorrência do elemento;
- **Frequência relativa acumulada (frac):** a porcentagem de valores que são menores ou iguais ao elemento analisado.

Estatística Descritiva

Variáveis Quantitativas: discretas ou contínuas

Se existe uma quantidade grande de dados, eles devem ser agrupados por **intervalos de classes** de igual largura; assim, a tabela de frequência apresentará o número de dados existentes no intervalo da classe. É possível estimar o número de classes pela raiz quadrada do número de observações (n).

Estatística Descritiva

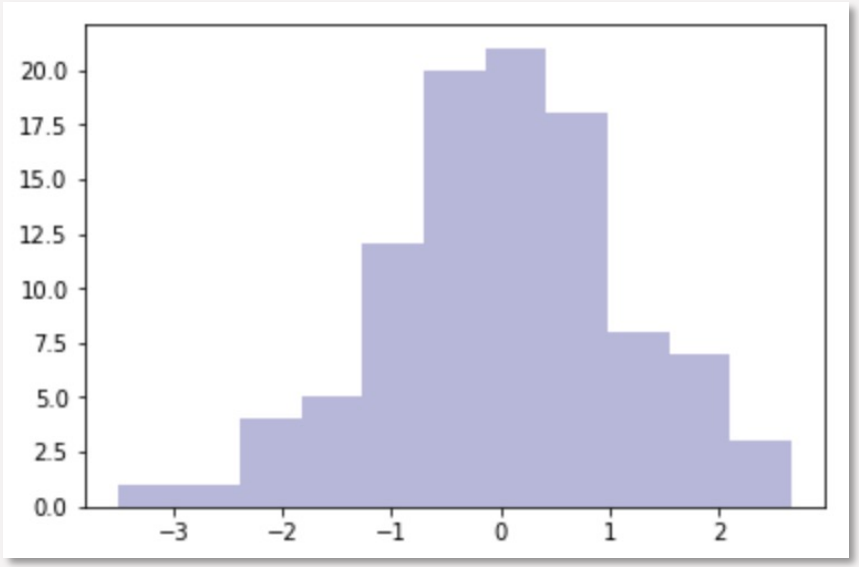
Variáveis Quantitativas: discretas ou contínuas

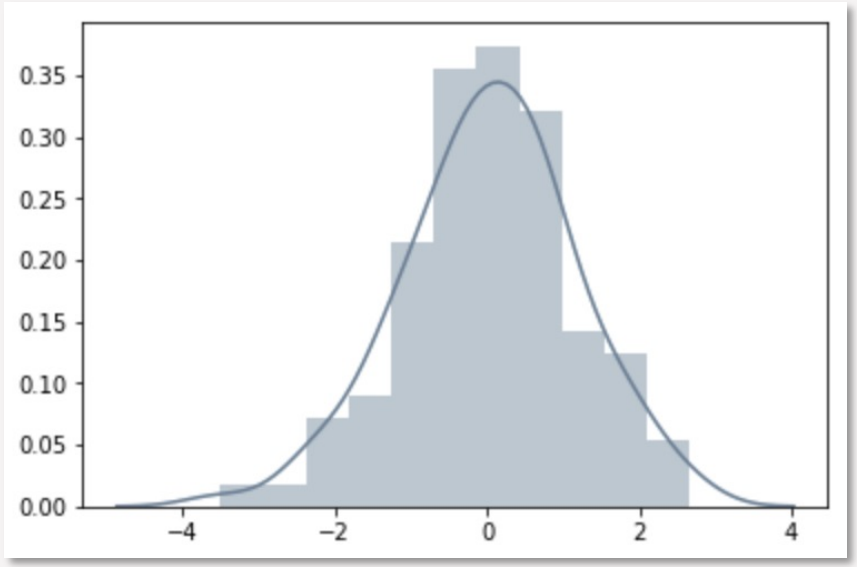
As variáveis quantitativas podem ser representadas por **gráficos** como:

- **Histogramas**: mostram a distribuição das frequências;
- **Gráficos de *box e whisker* (box plot):** mostram a assimetria da distribuição, quartis, presença de outliers e variabilidade dos dados;
- **Gráficos de dispersão**: mostram a relação entre duas variáveis;
- **Séries temporais**: mostram a alteração de uma variável no tempo.

Estatística Descritiva

- **Histograma**

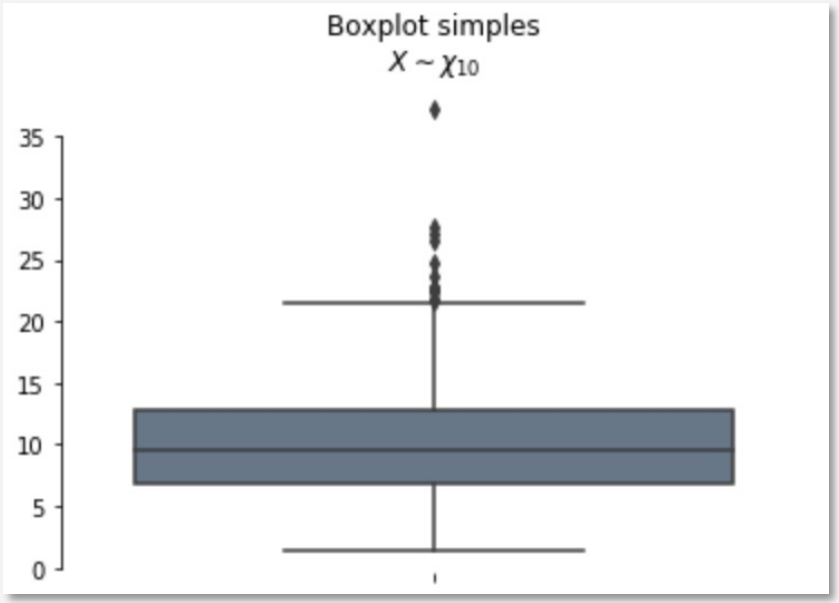


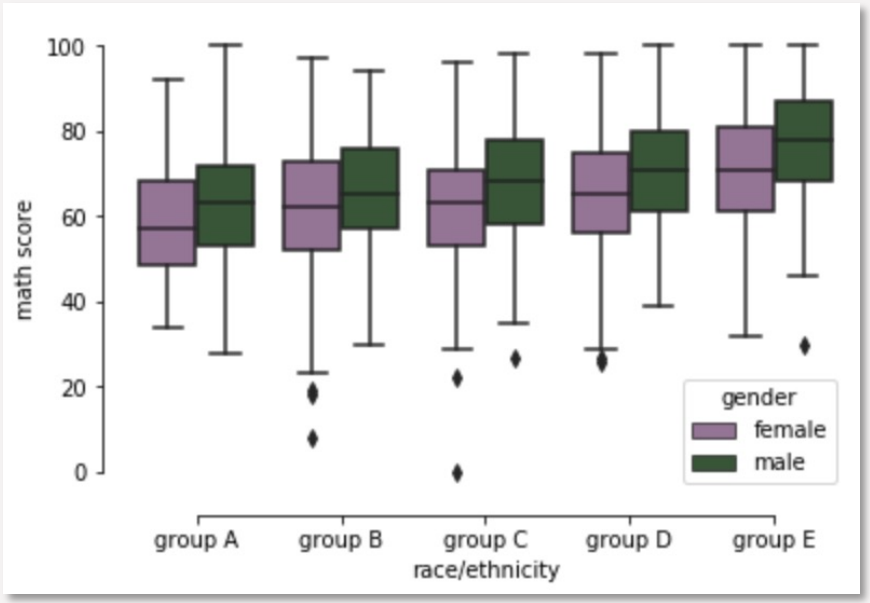


Usados com dados univariados

Estatística Descritiva

- **Box plot**

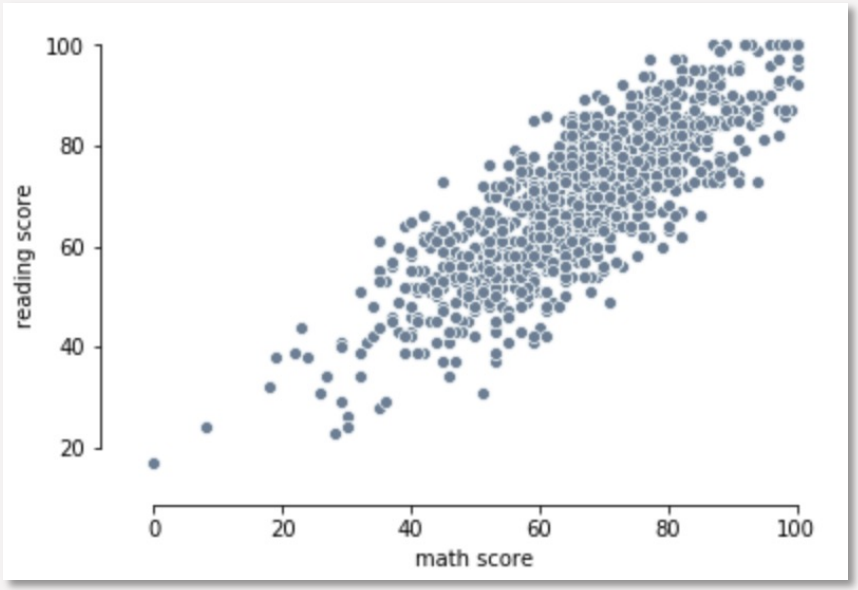


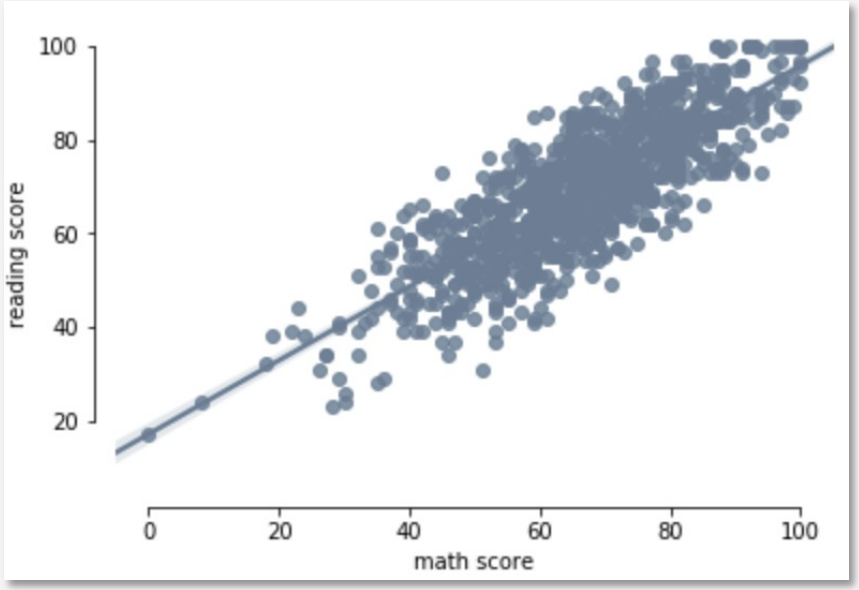


Usados com dados univariados

Estatística Descritiva

- **Dispersão**

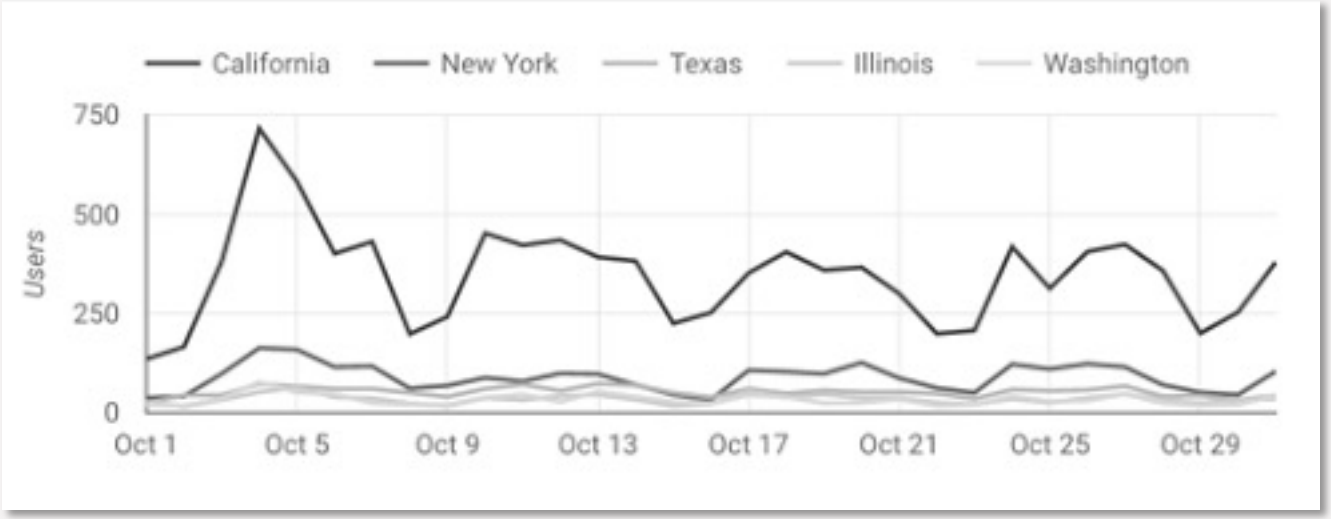




Usados com dados bivariados

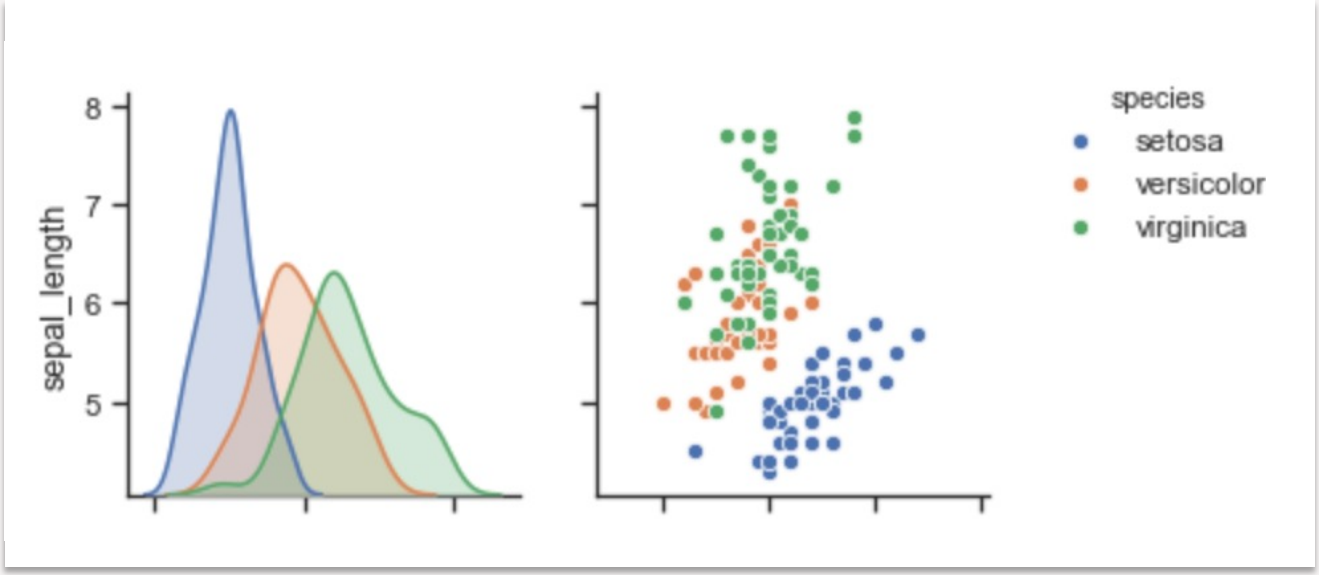
Estatística Descritiva

- **Série Temporal**



Estatística Descritiva

- **Dados Multivariados**



Estatística Descritiva

Medidas de tendência central

- **Média** (*mean*, *average*): é como se fosse o ponto de equilíbrio da distribuição e pode ser calculada por:

$$\mu = \frac{\sum x_i}{N} \qquad \bar{x} = \frac{\sum x_i}{n}$$

Média populacional x amostral

Estatística Descritiva

Medidas de tendência central

- **Média** (*mean*, *average*): é como se fosse o ponto de equilíbrio da distribuição

  e pode ser calculada por:

```
Numpy:
np.mean(array)

Pandas:
dataframe.mean() #retorna a média de cada coluna
serie.mean()
```

Estatística Descritiva

Medidas de tendência central

- **Moda**: representa o valor mais comum do conjunto de dados e é mais utilizada para dados categóricos. Se, por exemplo, dois valores apresentarem uma mesma frequência, seu conjunto de dados contém duas modas.

```
Pandas:
dataframe.mode()
serie.mode()
```

Estatística Descritiva

Medidas de tendência central

- **Mediana**: A mediana é o valor que separa a metade superior da metade inferior de uma distribuição de dados, ou o valor no centro da distribuição.

Pandas:
dataframe.median()
serie.median()

Estatística Descritiva

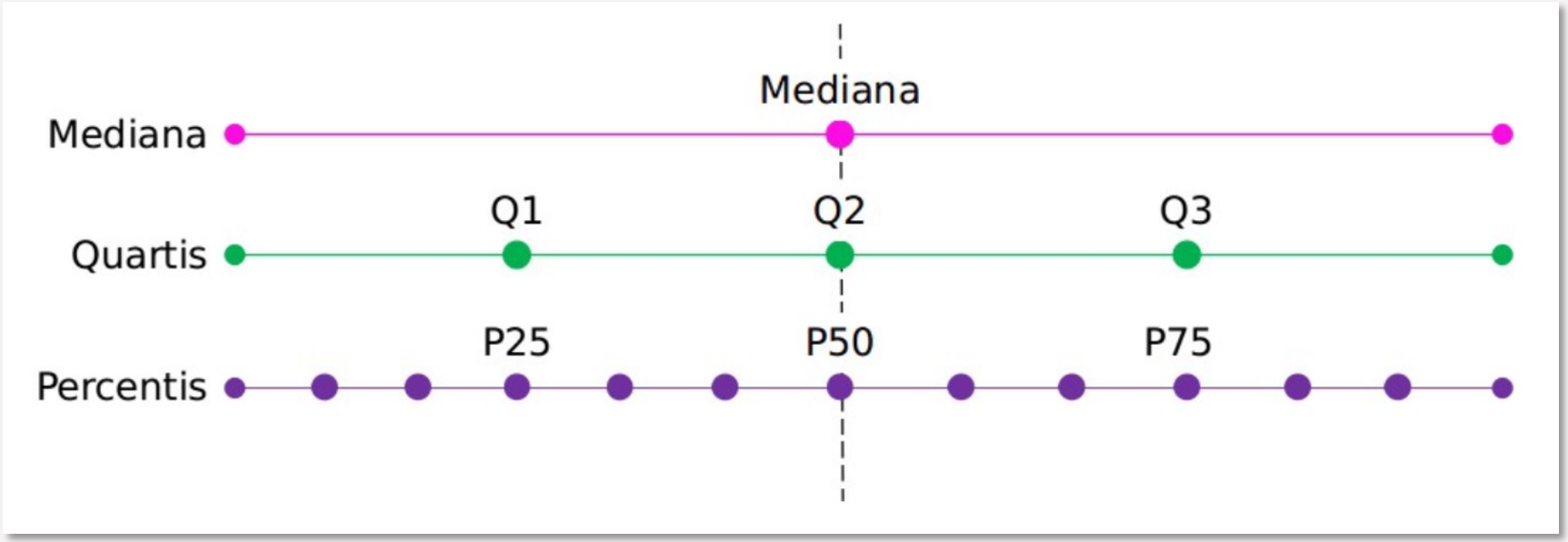
Medidas separatrizes

- **Percentis**: dividem o conjunto de dados em 100 partes iguais, ou seja, em pedaços de tamanhos iguais que contêm 1% dos dados.
- **Quartis**: dividem o conjunto de dados em 4 partes, ou seja, em pedaços de tamanhos iguais que contém 25% dos dados.
- **Mediana**: dividem o conjunto de dados em 2 partes. Acima da mediana estão 50% dos dados e abaixo dela também.

Estatística Descritiva

Medidas separatrizes

A mediana é o segundo quartil (Q2) e também o percentil 50 (P50):



Estatística Descritiva

Medidas separatrizes

A mediana é o segundo quartil (Q2) e também o percentil 50 (P50):

Pandas:
dataframe.**quantile**(q=quartil_desejado)
serie.**quantile**(q=quartil_desejado)

Estatística Descritiva

Amplitude Interquartil (*InterQuartile Range, IQR*): descreve a dispersão dos 50% dados centrais.

A Amplitude Interquartil, ou AIQ, pode ser calculada pela fórmula:

AIQ = Q3 - Q1

sendo Q3 o terceiro quartil (75%) e Q1 o primeiro quartil (25%).

Quanto maior o valor encontrado para o AIQ, mais dispersos estão os dados.

Estatística Descritiva

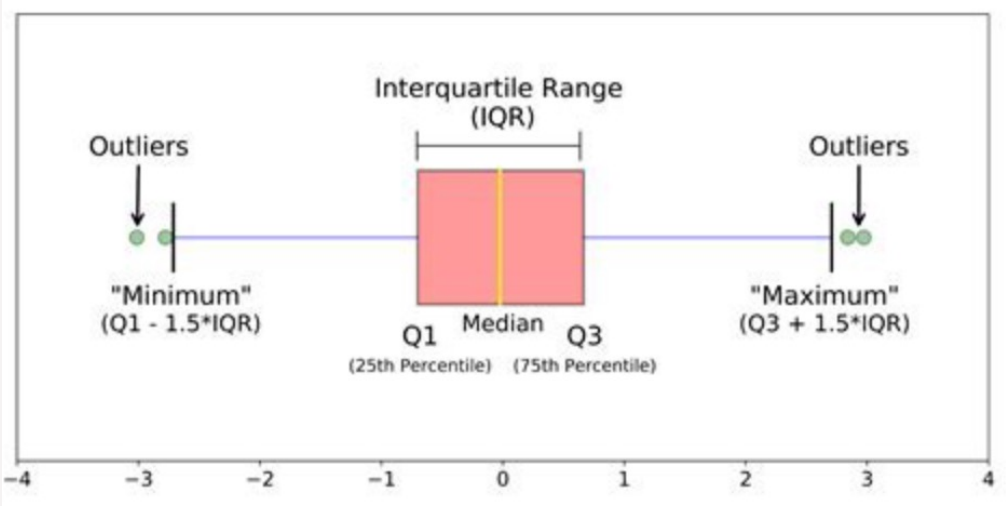
Outliers

Os **outliers** são dados que se diferenciam drasticamente de todos os outros. Em outras palavras, um **outlier** é um valor que foge da normalidade e que pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise.

Estatística Descritiva

Medidas separatrizes

O limite inferior é calculado por Q1–1,5(IQR) e o superior por Q3+1,5(IQR). Todo valor acima do superior e abaixo do inferior é considerado ourlier. IQR, ou AIQ, é a Amplitude Interquartil, AIQ=Q3-Q1.



Estatística Descritiva

Medidas de dispersão

- **Amplitude:** é do que a diferença entre o maior e o menor valor de um conjunto de dados. Para fazer este cálculo no Pandas, usaremos as funções max() e min(), que obviamente, retornam o valor máximo e mínimo de um conjunto de dados, e depois subtrairemos um do outro:

```
Pandas:
dataframe.max() - dataframe.min()
serie.max() - serie.min()
```

Estatística Descritiva

Medidas de dispersão

- **Variância** (*variance*): é mais utilizada de forma comparativa, já que não é muito intuitiva por não estar na mesma unidade dos dados.

$$\sigma^2 = \frac{\sum(xi - \mu)^2}{N} \quad \Big| \quad s^2 = \frac{\sum(xi - \bar{x})^2}{n - 1}$$

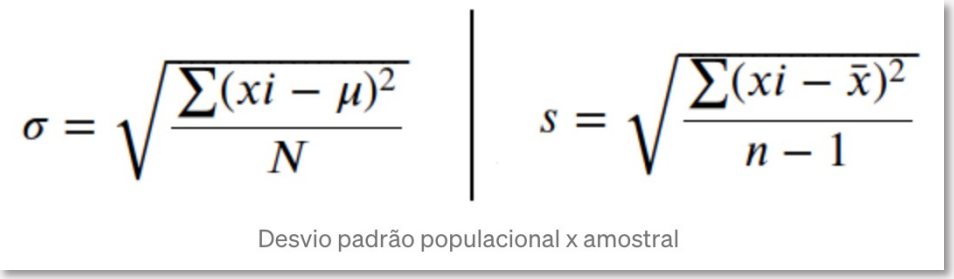
Variância populacional x amostral

Pandas:
dataframe.**var()**
serie.**var()**

Estatística Descritiva

Medidas de dispersão

- **Desvio padrão** (*standard deviation*): mais utilizado por estar na unidade dos dados. É a raiz quadrada da variância e indica quanto os dados estão afastados da média.



$$\sigma = \sqrt{\frac{\sum(xi - \mu)^2}{N}} \qquad s = \sqrt{\frac{\sum(xi - \bar{x})^2}{n - 1}}$$

Desvio padrão populacional x amostral

Pandas:
dataframe.**std()**
serie. **std()**

Estatística Descritiva

Medidas de dispersão

- **Desvio padrão**

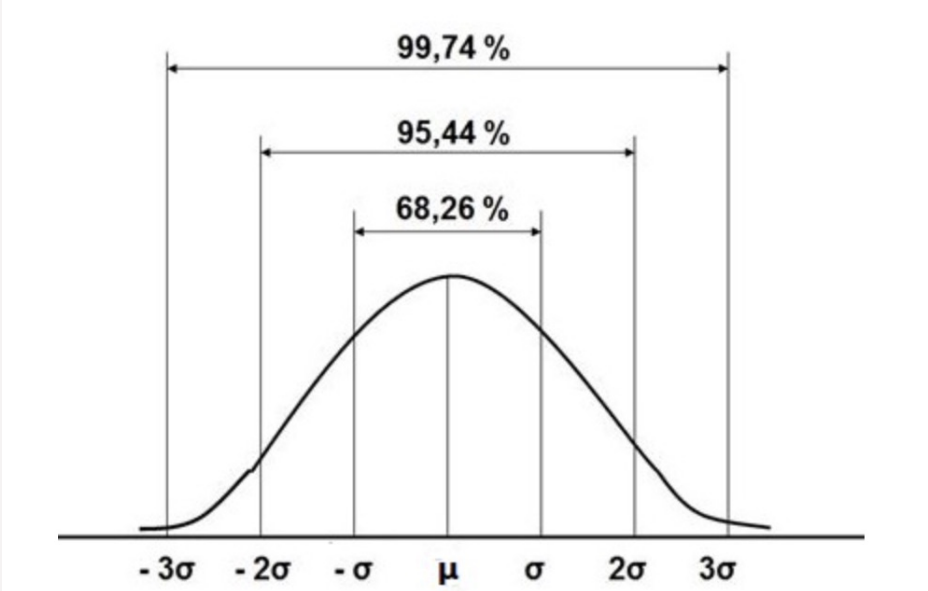
Os valores de desvio padrão também estão no conjunto dos números *Reais Positivos*, ou seja, se você encontrar um valor negativo, seus cálculos necessitam de revisão.

Considerando uma distribuição normal, 68% dos valores estão a 1 desvio padrão de distância da média:

Estatística Descritiva

Medidas de dispersão

- **Desvio padrão**



Porcentagem dos dados nos intervalos

Estatística Descritiva

Medidas de dispersão

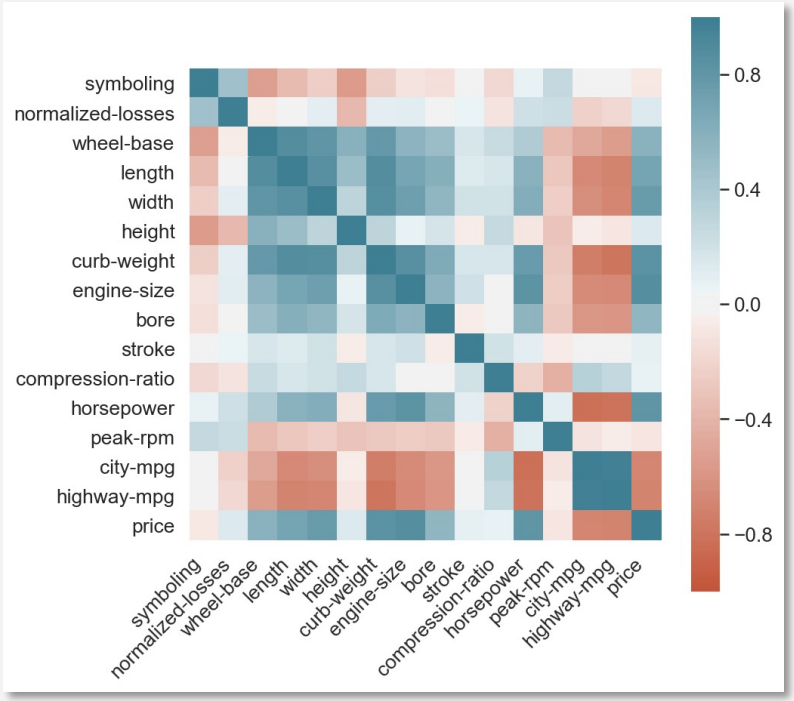
- **Correlação:** é uma medida que indica o quanto duas variáveis estão relacionadas. Seu valor fica sempre entre -1, que indica uma anti-correlação perfeita, e 1, que indica uma correlação perfeita.

Pandas:
dataframe.**corr()**
serie. **corr()**

Estatística Descritiva

Medidas de dispersão

- **Correlação**



Estatística Descritiva

Medidas de dispersão

- **Correlação**