
# Elementos de Estatística Aplicada. Métodos Computacionais em Física Médica.

Departamento de Física e Astronomia

FCUP - 2022

**Pedro Teles**

# 1 Introdução.

Na primeira parte da Unidade Curricular de Métodos Numéricos em Física Médica, vamos introduzir uma série de conceitos e métodos estatísticos para serem utilizados como ferramentas de aplicação para os Métodos Numéricos, tão essenciais em Física Médica, sobretudo no que toca às técnicas de Monte Carlo.

Pretende este documento ser um apoio ao estudante nesta primeira parte da disciplina.

Iremos assim definir de forma explícita, ainda que introdutória, alguns dos conceitos mencionados na sala de aula.

Relembra-se ao estudante que o foco desta componente da UC não é tanto na rigidez das suas definições matemáticas, mas numa introdução abrangente dos tópicos, assim como na sua importância enquanto poderosas ferramentas matemáticas. Focaremos também na sua codificação utilizando linguagens computacionais, como o Python.

# 2 Natureza das Variáveis Estatísticas.

A Estatística e os seus métodos são hoje fundamentais em quase todas as áreas científicas, mas também nas áreas sociais e nas finanças. Com o aumento progressivo da capacidade computacional, é hoje possível, a partir de códigos computacionais, utilizar estas ferramentas para o tratamento de grandes quantidades de dados, de forma rápida e eficaz.

Podemos começar então por definir o domínio de aplicação da Estatística. A Estatística utiliza métodos científicos para catalogar, organizar, descrever, sumariar e analizar conjuntos de dados, e pode ser dividida em duas grandes áreas – A **Estatística Descritiva ou Dedutiva**, que trata da descrição de conjuntos de dados, e a **Estatística Inferencial ou Indutiva**, que procura retirar conclusões, ou prever comportamentos futuros de uma determinada amostragem de dados.

Os dados utilizados em Estatística são chamados **dados estatísticos**. Estes dados são obtidos a partir de uma determinada **população** que pode ser finita ou infinita. Mesmo quando a população é finita, normalmente não é possível tratar a totalidade da população, retirando-se desta um conjunto de observações, ou dados. A este conjunto de dados, chama-se **amostra**. O número total de dados numa determinada amostra é normalmente chamado de $N$.

Aos dados estatísticos *(observações)*, correspondem **varíáveis estatísticas** *(observáveis)*. Estas variáveis estão definidas num determinado domínio, e podem ser de diferentes tipos:

1. **variáveis numéricas:** Varíaveis que descrevem dados de natureza numérica. Estas variáveis têm sempre unidades (físicas, ou alternativamente "pessoas"). Às variáveis numéricas podem aplicar-se todos as operações matemáticas possíveis (soma, subtração, divisão, exponenciação, etc.). Exemplos destas variáveis são: altura, peso, índice de massa corporal, número de pessoas com diabetes, idade, ...). Dependendo do seu domínio de aplicação, estas variáveis podem ainda ser:

    a. **discretas:** Variáveis definidas ou mapeáveis ao conjunto dos números inteiros (seja $x$ uma variável numérica discreta, então $x \in \mathbb{Z}$). Exemplos: números de pessoas, números de eventos, interações, colisões, etc.)

    b. **contínuas:** Variáveis definidas ou no conjunto dos números reais ou complexos (seja $x$ uma variável numérica contínua, então $x \in \mathbb{R}$ ou $x \in \mathbb{C}$). Exemplos: idade, altura, dose efetiva, dose absorvida, massa, energia, etc.)

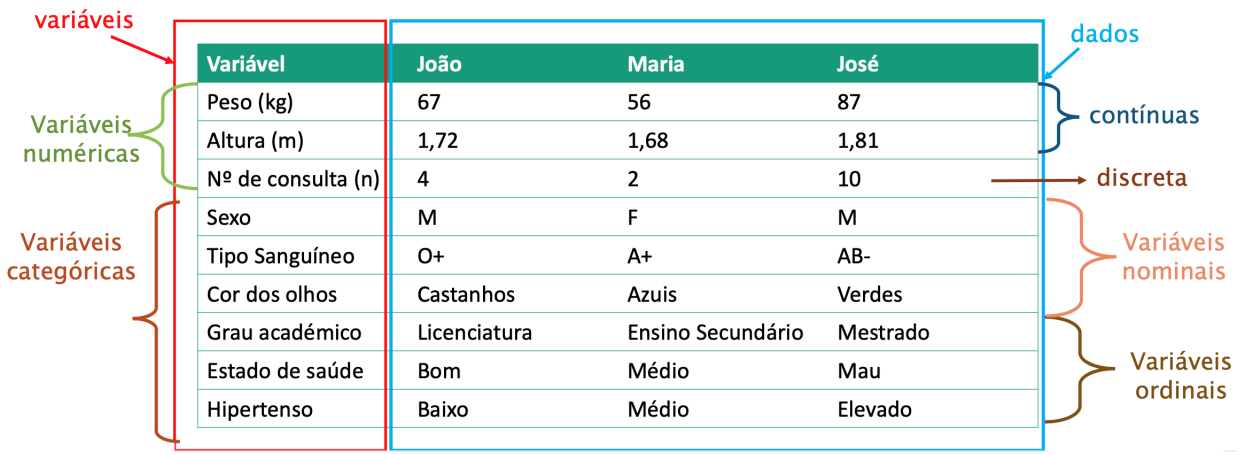
    *NB: É por vezes conveniente, sobretudo em estatística descritiva, considerar variáveis discretas como variáveis contínuas, no caso em que estas estão definidas num domínio com um grande número de possibilidades. Por outro lado, é por vezes conveniente tratar variáveis contínuas como discretas (por exemplo: altura, ou idade).*

2. **variáveis categóricas:** Variáveis definidas num domínio de valores diferentes entre si, de acordo com algum tipo de propriedade qualitativa. (Exemplos: cor dos olhos, tipo sanguíneo, Género, etc.). Apesar de poderem muitas vezes ser mapeáveis num conjunto de números discretos, estas variáveis são não-numéricas (não se pode realizar com elas todo o tipo de operação matemática), e não têm unidades (físicas, ou "números de pessoas). As variáveis categóricas podem ainda ser divididas em:

   a. **nominais:** Variáveis em que não existe qualquer tipo de ordem específica em que podem ser colocadas. (exemplo: Género, cor da pele, tipo sanguíneo, nacionalidade, Nome, etc.)

   b. **ordinais:** Variáveis que podem ser ordenadas de acordo com algum critério específico (exemplo: Notas de uma disciplina, graus de escolaridade, etc.). Apesar destas variáveis poderem ser definidas, por exemplo, com recurso aos números naturais *não são variáveis numéricas*. Por exemplo, não faz sentido qualquer tipo de operação matemática entre elas.

Na figura em baixo dá-se um exemplo de uma amostra hipotética de 3 pacientes de uma determinada clínica, dos quais foram guardados uma série de dados estatísticos, e em que os tipos de variáveis estatísticas estão identificados.




| Variável | João | Maria | José |
|---|---|---|---|
| Peso (kg) | 67 | 56 | 87 |
| Altura (m) | 1,72 | 1,68 | 1,81 |
| Nº de consulta (n) | 4 | 2 | 10 |
| Sexo | M | F | M |
| Tipo Sanguíneo | O+ | A+ | AB- |
| Cor dos olhos | Castanhos | Azuis | Verdes |
| Grau académico | Licenciatura | Ensino Secundário | Mestrado |
| Estado de saúde | Bom | Médio | Mau |
| Hipertenso | Baixo | Médio | Elevado |

**Figura 1:** Exemplos de tipos de variáveis estatísticas.

# 3 Introdução à Estatística Descritiva.

*"Lies, damn lies, and statistics- Mark Twain*

A Estatística Descritiva lida, como o próprio nome indica, com a descrição de amostras de dados estatísticos. Para isso utiliza ferramentas que permitem veicular, de forma eficaz, o máximo de informação sobre uma determinada amostra estatística.

A maior parte das vezes, a amostra estatística não está tratada. A este conjunto de dados é comum chamar-se **dados brutos ou em bruto**. Em seguida vamos indicar algumas formas de tratar estes dados em bruto de forma a veicular com eles informação.

Uma das formas mais diretas e naturais de descrever uma amostra de dados estatísticos é através da utilização de gráficos e tabelas.

## 3.1 Tabelas de frequências.

É comum, em todos os tipos de variáveis estatísticas, encontrar repetições dos mesmos valores numa determinada amostra. Às somas dessas repetições dá-se o nome de **frequências**.

Alternativamente, pode ser conveniente agrupar os dados da amostra em **intervalos ou classes de frequências**.

A uma tabela que contenha esta informação dá-se o nome de **tabela de frequências** ou **distribuição de frequências**. É possível ainda normalizar estas frequências ao número total de observações $N$. Nesta situação obtêm-se as **frequências relativas**.

Finalmente, é possível ir somando os valores das frequências para cada valor ou intervalo. Nestes casos, tratam-se de **frequências acumuladas**. Quando estas estão também normalizadas, chamam-se **frequências acumuladas relativas**.

| Peso (kg) | Frequências | Relativas (%) | Acumuladas | Acumuladas relativas (%) |
|---|---|---|---|---|
| 0-10 | 9 | 45.0 | 9 | 45.0 |
| 10-20 | 6 | 30.0 | 15 | 75.0 |
| 20-30 | 2 | 10.0 | 17 | 85.0 |
| 30-40 | 1 | 5.0 | 18 | 90.0 |
| 40-50 | 0 | 0.0 | 18 | 90.0 |
| 50-60 | 1 | 5.0 | 19 | 95.0 |
| 60-70 | 0 | 0.0 | 19 | 95.0 |
| 70-80 | 1 | 5.0 | 20 | 100.0 |
| 80-90 | 0 | 0.0 | 20 | 100.0 |
| 90-100 | 0 | 0.0 | 20 | 100.0 |

**Tabela 1:** Exemplo de uma tabela de frequências.

É possível visualizar, na tabela 3.1 uma tabela de frequências de intervalos de pesos de uma amostra de dados de crianças submetidas a cintigrafias renais num centro de medicina nuclear.

## 3.2 Gráficos de frequências.

Estas frequências podem ser também visualizadas em gráficos. É comum utilizar **gráficos de barras**, **gráficos circulares** ou outros para visualização. Na figura 2 apresenta-se um gráfico de barras e um gráfico circular *(pie chart)* das frequências dos intervalos de pesos da tabela 3.1. Para o pie chart, os valores de 0 não são mostrados.

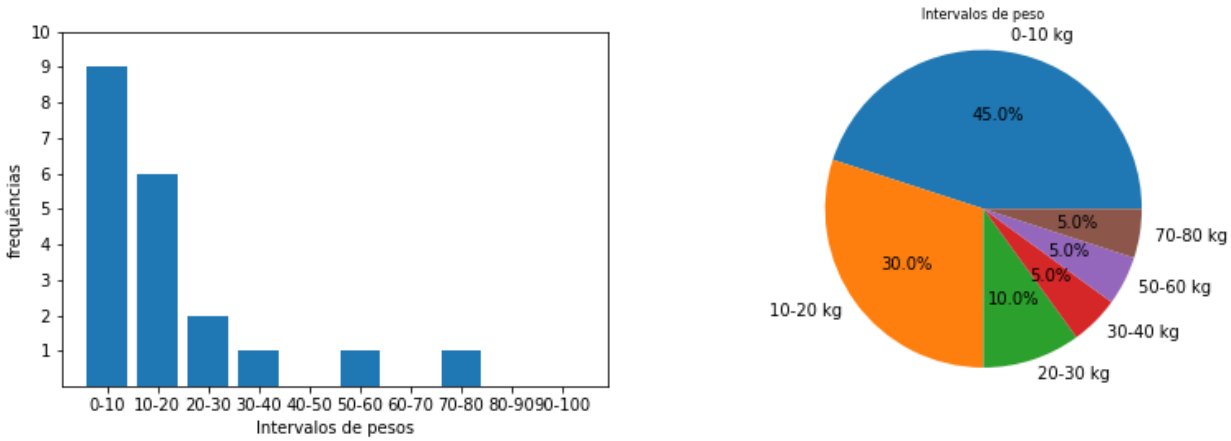



**Figura 2:** Gráfico de barras e *pie chart* das frequências da tabela 3.1

O mesmo pode ser feito para gráficos de frequências relativas, frequências acumuladas etc.

## 3.3 Histogramas.

No caso de variáveis numéricas contínuas, é possível ainda a sua visualização com recurso a um **histograma**. Os histogramas diferem de gráficos de barras, já que podem ser utilizados apenas com variáveis contínuas. Apesar disso a sua forma de construção é equivalente.

- Agrupam-se os valores da amostra em intervalos *(bins)* devidamente espaçados; a largura dos intervalos deve ser escolhida de forma conveniente, pois pode ter uma influência grande na forma de visualização.
- Faz-se um gráfico em que no eixo dos xx são colocados os intervalos e no dos yy as frequências.
- são desenhadas "barras"entre cada intervalo, com uma altura correspondente ao valor da frequência.
- Não devem existir espaços entre os diferentes intervalos, para indicar *continuidade*.

- Se no histograma as frequências estiverem normalizadas, então temos um histograma de frequências relativas.

Os intervalos de um histograma não são necessariamente todos iguais. Neste caso, deve-se desenhar a barra em cada intervalo, não de forma proporcional à frequência, mas normalizado à largura do intervalo. Neste caso, a barra indica não frequência, mas *"densidade de frequências"*(não confundir com frequência relativa). Nestes casos, o valor da frequência é obtido multiplicando a largura do intervalo pela sua altura.

Os histogramas também podem ser cumulativos. Na figura 3, dá-se um exemplo de um histograma cumulativo, para os pesos das crianças referidos nos gráficos e tabelas anteriores.

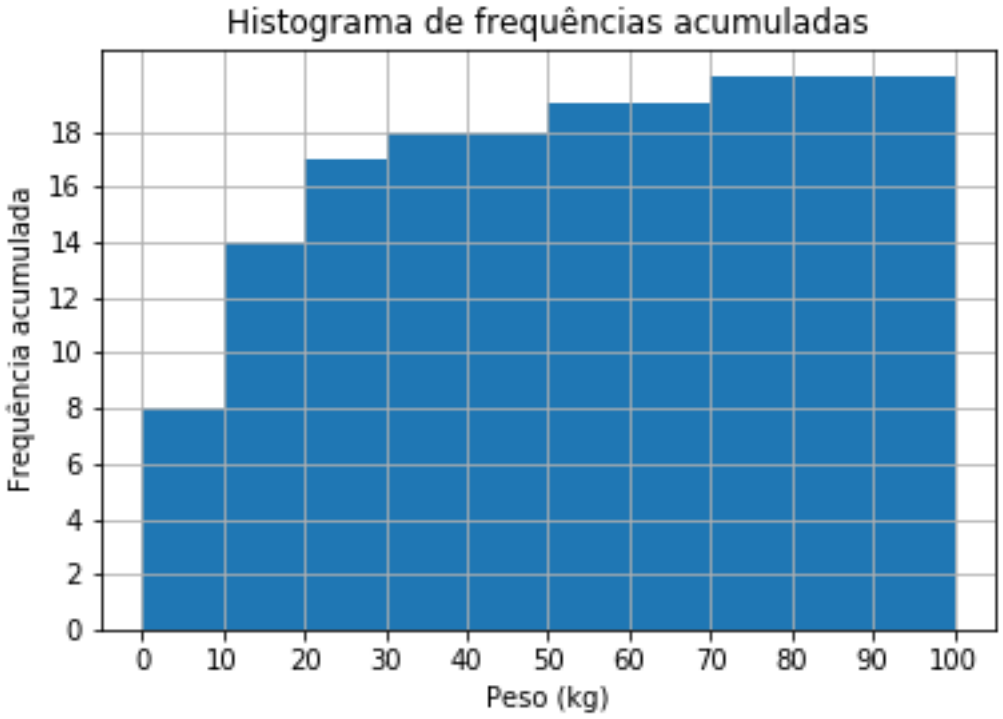



**Figura 3:** Histograma de frequências acumuladas ou histograma cumulativo para a mesma variável da tabela 3.1

# 4 Valores-sumário.

*"I couldn't claim that I was smarter than sixty-five other guys – but the average of sixty-five other guys, certainly!" - Richard P. Feynman*

Os dados estatísticos de uma determinada amostra são frequentemente numerosos, sofrem grandes variações, e são difíceis de interpretar *em bruto*; Por vezes, também não basta descrevê-los apenas com tabelas e gráficos de frequências.

Em estatística descritiva, podemos definir valores a partir dos dados em bruto ou das frequências que procuram descrever de forma rápida a amostra, fornecendo o máximo de informação possível. Estes valores são chamados de **valores-sumário**, ou valores de sumário.

## 4.1 Medidas de tendência central.

Muitos conjuntos de dados estatísticos tendem a aglomerar-se em torno de um valor, muitas vezes um valor central. Os três principais indicadores de tendência central em estatística descritiva são:

- **a moda** – Seja $X$ uma variável estatística aleatória *(iremos ver como definir posteriormente uma variável aleatória)* númerica discreta ou categórica (nominal ou ordinal) definida num determinado intervalo de valores, que definem uma **distribuição** de tamanho $N$, $X \in \{X_1, X_2, ..., X_N\}$. Se $f_i$ representar a frequência *(número de ocorrências – ver secção anterior)* de cada variável $X_i$ no conjunto, a moda corresponde ao valor máximo de $f_i$. Se a distribuição tiver apenas um máximo, diz-se *unimodal*. Se tiver mais do que um máximo, diz-se que *multimodal* (bimodal – 2 modas; trimodal – 3 modas, etc.). A moda é o único valor-sumário de tendência central possível de ser utilizado nas variáveis categóricas nominais. Não é possível definir moda para uma variável numérica contínua, pois cada valor toma uma única frequência. Como é possível dividir variáveis numéricas contínuas em intervalos, fala-se muitas vezes em **intervalo modal**, que corresponde ao intervalo com máxima frequência. Alguns autores definem a moda como sendo o valor intermédio do intervalo.

- **a mediana** – Seja $P(X_1, X_2, ..., X_N)$ uma distribuição estatística *ordenada* de tamanho $N$, ou seja $X_1 < X_2 < X_3 < ... < X_N$, em que $X_i$ é uma variável numérica ou categórica ordinal.

  - Se $N$ é ímpar, a mediana é o valor que divide a distribuição em duas partes iguais, uma superior e uma inferior, ou seja $X_{(N+1)/2}$.

  - Se $N$ é par, não existe um só valor mediano, define-se então a mediana como sendo o valor médio entre os dois valores centrais, ou seja, $\frac{X_{N/2} + X_{N/2+1}}{2}$.

  A mediana é um valor que não é afetado por grandes variações nos valores não próximos da média, o que a torna um bom medidor de **obliquidade ou assimetria** de uma distribuição, em comparação com a média.

- **a média** – Seja $P(X_1, X_2, ..., X_N)$ uma distribuição estatística de variáveis numéricas discretas de tamanho $N$. A **média aritmética** ($\overline{X}$) desta distribuição é definida como:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}. \tag{1}$$

  A média aritmética pode também ser definida em termos das frequências $f$ de cada valor:

$$\overline{X} = \frac{\sum_{k=1}^{n} f_k \cdot X_k}{N}, \tag{2}$$

  em que $k$ representa a ocorrência de cada valor $X_k$, e $n$ o número total de ocorrências. Finalmente, isto pode ser generalizado para uma variável contínua $x$ distribuída como $f(x)$ no conjunto de números reais $\mathbb{R}$:

$$\overline{X} = \int_{-\infty}^{+\infty} x f(x) dx. \tag{3}$$

  Para além da média aritmética, em especial para variáveis numéricas com variações pronunciadas entre si (p.ex. valores distribuídos exponencialmente), é comum definir também a **média geométrica** como sendo:

$$\overline{X} = \sqrt[N]{X_1 \cdot X_2 \cdot X_3 \cdots X_N} = \sqrt[N]{\prod_{i=1}^{N} X_i}. \tag{4}$$

  A média não é geralmente aplicável a variáveis categóricas, no entanto, em alguns casos, pode fazer sentido (ex.: notas de 0 a 20, etc.).

  Finalmente, é comum designar por $\mu$ a **média de uma população**, e como $\overline{X}$ a **média de uma amostra**.

Na sala de aula vimos um exemplo de como programar o cálculo de uma média aritmética no Jupyter notebook a partir de um ficheiro inicial gravado separado por vírgulas ('csf' – *comma separated file*). Na listagem abaixo podemos ver um exemplo geral de código:

```
import math
import numpy as np
import pandas as pd

dados = pd.read_csv("O_meu_ficheiro_csv.csv")
dados = dados["coluna_contendo_a_variavel"].value_counts().reset_index(name=
    'Count').rename(columns={'Col1':'Col_value'})
dados.columns=['variavel','contagem']
N=sum(dados["contagem"]) # usamos a frequencia total
soma=sum(dados["variavel"]*dados["contagem"]) #somamos ponderando com
    frequencias
media=soma/N
print("A media da variavel e % s." % media)
```

**Listing 1:** Como calcular uma média aritmética de um "dataframe"importado de um ficheiro .csv usando o Pyhton

Para o cálculo de medianas:

```
import math
import numpy as np
import pandas as pd

dados = pd.read_csv("O_meu_ficheiro_csv.csv")
N=len(dados) #vamos precisar
var=dados["coluna_contendo_a_variavel"].sort_values()  # precisamos de
    ordenar os

if N % 2 == 0:  # casos de distribuicao par
    mediana1=var[N//2]
    mediana2=var[N//2 +1] # buscar os dois valores centrais
    mediana= (mediana1+mediana2)/2
else:
    mediana=var[(N+1)//2]
print("A mediana da variavel e % s." % mediana)
```

**Listing 2:** Como calcular a mediana de um "dataframe"importado de um ficheiro .csv usando o Pyhton

E finalmente para o cálculo de modas:

```
import math
import numpy as np
import pandas as pd
from collections import Counter

dados = pd.read_csv("O_meu_ficheiro_csv.csv")
var = dados["coluna_contendo_a_variavel"]
N=len(dados) # continuamos a precisar
```

```
contados=Counter(var) # conta os valores
v=list(contados.values())
k=list(contados.keys())

moda2 = k[v.index(max(v))]

if max(v) == N:
    moda = "Este conjunto de dados nao tem moda"
else:
    moda = moda2
print("A(s) moda(s) da variavel e(sao)% s." % moda)
```

**Listing 3:** Como calcular a moda de um "dataframe"importado de um ficheiro .csv usando o Pyhton

## 4.2 Valores de localização não-central.

Para além das medidas de tendência central, é possível e muitas vezes necessário definir indicadores de localização não-centrais. Estes valores chamam-se **valores de localização não central**, ou **quantis**.

stes valores determinam-se exactamente como a mediana, isto é, para uma determinada distribuição estatística ordenada $P(X_1, X_2, ..., X_N)$ com $N$ valores, o valor do 'n'-til (em que 'n' representa o número de intervalos do quantil desejado mais um: "quartil"(3 intervalos), "decil"(9 intervalos), "percentil"(99 intervalos), etc.). *(Nota: Para amostras com valores de N menores que a sivisão dos 'n'-tis, iremos utilizar o método de arredondamento à ordem, ou seja utiliza-se a expressão e arredonda-se ao valor inteiro correspondente ao valor da ordem do valor na amostra)*:

- Se $N$ é ímpar – $X_{(N+1)/n}$ é o 1º 'n'-til, $X_{2\cdot(N+1)/n}$ é o 2º, $X_{k\cdot(N+1)/n}$ é o kº, etc (com $k \leq (n-1)$).
- Quando o tamanho da amostra é par, faz-se a média dos dois dados adjacentes $X_{k\cdot N/n}$ e $X_{k\cdot((N/n)+1)}$ para o kº 'n'-til. (com $k \leq (n-1)$).

Alguns exemplos em baixo:

- 1º, 2º, e 3º Quartis: valores que separam 25, 50, ou 75% dos valores inferiores, dos 75, 50, ou 25% dos valores superiores. O 2º quartil corresponde à mediana.
- Decis: 9 valores que separam 10, 20, 30, 40, 50, 60, 70, 80, 90% dos valores inferiores, dos 90, 80, 70, 60, 50, 40, 30, 20, 10% dos valores superiores. O 5º decil corresponde à mediana.
- Percentis: o mesmo que acima mas intervalado em percentagens.

## 4.3 Medidas de de alcance ou dispersão.

Numa distribuição estatística, para além de determinar valores de tendência central ou não-central, é importante saber a que *distância* de um determinado valor (usualmente de um valor de tendência central, e em particular, da média) se encontram os valores da distribuição. A estas medidas de distância, chamam-se **medidas de alcance ou de dispersão.**

- **a variância (população)** – Seja $P(X_1, X_2, ..., X_N)$ uma população estatística com uma média $\mu$, A variância desta população pode ser definida como:

$$V = \overline{X^2} - \mu^2, \tag{5}$$

  ou seja, a variância corresponde à diferença entre a média dos valores da população ao quadrado:

$$\overline{X^2} = \frac{\sum_{i=1}^{N} X_i^2}{N}, \tag{6}$$

  pela média ao quadrado da distribuição.

- **o desvio-padrão (população)** – O desvio-padrão da população é simplesmente a raiz quadrada da sua variância.

$$\sigma = \sqrt{V} = \sqrt{\overline{X^2} - \mu^2} \tag{7}$$

- **a variância (amostra)** – Para determinar a variância de uma amostra, deve-se utilizar a correção de Bessel, que permite retirar à variância o viés quem vem da amostragem. Seja $\overline{X}$ a média da amostra:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \tag{8}$$

- **o desvio-padrão (amostra)** – O desvio-padrão da amostra é simplesmente a raiz quadrada da sua variância.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \tag{9}$$

Estes dois valores são essenciais em estatística descritiva e inferencial, como iremos ver mais adiante.

## 4.4 Resumo.

A estatística descritiva utiliza valores-sumário que permitem transmitir o máximo de informação sobre um determinado conjunto estatístico.

- Estes valores podem ser valores de tendência central (assumindo que os valores tendem a agregar- se em torno de um ponto central) – a moda, a mediana, e a média.
- podem também ser valores de localização, os quantis, que informam sobre uma determinada posição num conjunto estatístico – os quartis, os decis, e os percentis (por exemplo).
- finalmente, podem também ser valores de dispersão, que medem a "distância"a que os valores da amostra estão dispersos em torno de um determinado valor central, em particular, da média.

Estes valores devem ser utilizados de forma científica, de maneira a "cumprirem o seu dever", i.e. o de fornecerem informação relevante sobre um conjunto estatístico. Devem ser evitados todos os tipos de viés e erros.