Estatística Descritiva

Rodrigo R. Pescim

Universidade Estadual de Londrina

Departamento de Estatística - CCE/UEL

- A **ciência estatística** tem por interesse a coleta, a apresentação, a análise e a utilização dos dados para tomar decisões, resolver problemas e planejar processos

- Os **métodos estatísticos** são utilizados para compreender o conceito de variabilidade

- **Variabilidade** $\rightarrow$ "Sucessivas observações de um fenômeno que não produz exatamente o mesmo resultado"

Exemplo 1

- Algumas estruturas nas áreas de construção e engenharia civil são expostas a “forças naturais”.

- Essas estruturas podem sofrer, ao longo do tempo, **degradação** (deterioração, fadiga, deformação, etc.) e também sofrer do efeito de **fatores externos** (corrosão, sobrecarga ou riscos ambientais).

- Assim, **as respostas associadas a esses fatores** nas estruturas não devem ser consideradas constante, mas sim, como **uma variação de um fenômeno aleatório**.

Exemplo 2

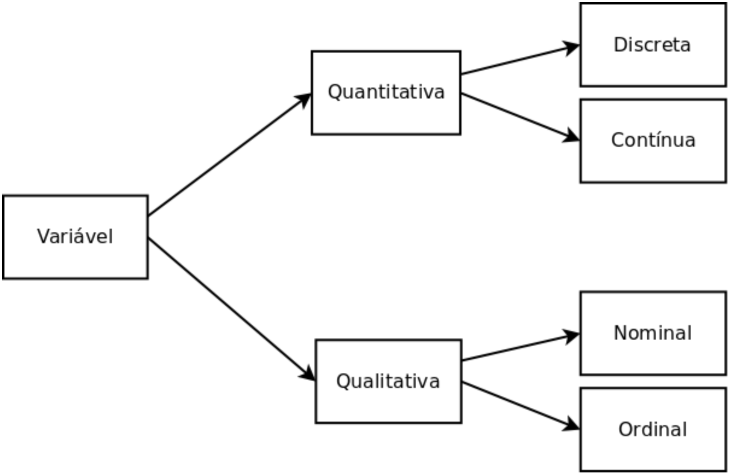
- Um artigo do periódico *Quality Engineering* apresenta dados de viscosidade (Pa.s) de um processo químico em batelada. Uma amostra desses dados é apresentada a seguir.

1,570	1,720	1,900	2,400	2,522	2,700
2,720	2,750	2,800	3,125	3,200	3,250
3,400	3,450	3,600	3,720	4,100	4,600

- Como retirar as primeiras informações sobre a viscosidade de um fluido num processo químico ?

Estatística Descritiva

- Tem por objetivo resumir e apresentar os dados sob a forma de **tabelas e gráficos**, além de estudar os parâmetros e suas estimativas.
- **Organização**: Como "tratar" os dados a fim de extrair informações a respeito de uma ou mais características de interesse ?
- **Variáveis**: Uma variável, neste contexto, é uma medida numérica ou classificação obtida a partir de uma característica de interesse de uma população em estudo.
- Exemplo:
  - Variável $X \rightarrow$ refere-se a viscosidade de um processo químico
  - x_i, $i = 1, \ldots, 18$ é o valor observado da viscosidade em cada unidade experimental
  - $x_1 = 1,570$; $x_2 = 1,720$; $x_3 = 1,900$; $\ldots$


Variável
Quantitativa
Discreta
Contínua
Qualitativa
Nominal
Ordinal

Tipos de Variáveis

- Os tipos de variáveis mais comumente utilizadas na descrição dos dados são: Quantitativas (Discretas ou Contínuas) e Qualitativas (nominais ou ordinais)
- Exemplos de variáveis quantitativas discretas:
  - Número de peças defeituosas;
  - Número de bits transmitidos que foram recebidos com erros;
  - Número de lâmpadas queimadas
- Exemplos de variáveis quantitativas contínuas:
  - Intensidade de corrente elétrica;
  - Tempo de falha de uma estrutura metálica;
  - Medidas de voltagem, comprimento, pressão, temperatura, etc.

- Exemplos de variáveis qualitativas nominais (ou categóricas):
  - Sexo;
  - Nacionalidade;
  - Área de atividade

- Exemplos de variáveis qualitativas ordinais:
  - Classes sociais;
  - Nível de instrução;
  - Categoria de clientes (ouro, prata e bronze)

Exercício 1

- Um levantamento foi realizado com 10 funcionários de uma construtora e os resultados obtidos são informados na tabela abaixo.

ID	Est. Civil	Instrução	Filhos	Salário	Idade
1	1	1	0	4,00	26
2	2	1	1	4,56	32
3	2	1	2	5,25	36
4	1	2	0	5,73	20
5	1	1	0	6,26	40
6	2	1	0	6,66	28
7	1	1	0	6,86	41
8	1	1	0	7,39	43
9	2	2	1	7,59	34
10	1	2	0	7,44	23

- Classifique as variáveis em quantitativas (discretas ou contínuas) ou qualitativas (nominais ou ordinais).

Tabela de Distribuições de Frequências

- Uma **distribuição de frequências** relaciona **categorias ou classes de valores** juntamente com as **frequências do número de valores** que se enquadram em cada categoria ou classe.

- Nas tabelas de distribuição de frequências - é usual fornecer a proporção (frequência relativa) das unidades que são de cada uma das variáveis

- **Para obter a frequência absoluta de uma dada variável**
  - Frequência absoluta (f_i)
  - Frequência absoluta acumulada (F_i)

- **Para obter a frequência relativa de uma dada variável**
  - Frequência relativa (f_i')
  - Frequência relativa (F_i')

- Como organizar os dados da variável Filhos em uma tabela de distribuição de frequências ?

Tabela 1: Número de filhos de uma certa empresa

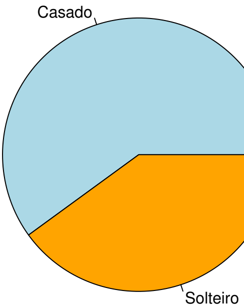
Nº de Filhos	f_i	f_i'	F_i	F_i'
0	7	0,7	7	0,7
1	2	0,2	9	0,9
2	1	0,1	10	1
Total	10	1	-	-

Representação Gráfica

- **Gráfico de Barras:** É mais apropriado para variáveis qualitativas ordinais e para variáveis quantitativas discretas.
  - Ex: Número de filhos (variável quantitativa discreta); Grau de instrução (variável qualitativa ordinal)

- **Gráfico de Setores:** É mais apropriado para variáveis qualitativas nominais, pois tem por objetivo a visualização do quanto a informação de cada parte está representada no todo.
  - Ex: Estado civil (variável qualitativa nominal)


Frequência Absoluta
0
1
2
3
4
5
6
7
0
1
2
Número de Filhos


Casado
Solteiro

Distribuições de Frequências e sua representação Gráfica para Variáveis Quantitativas Contínuas

- As variáveis quantitativas contínuas diferem um pouco das discretas na sua forma de representação gráfica
- Observa-se que essas variáveis, por definição, têm seus valores definidos no conjunto dos números reais
- Assim, não tem sentido falar em frequência de repetição de um determinado valor, pois os valores raramente se repetem (exemplo 2)

Método da raiz quadrada

1. Determinar o número de classes (ou intervalos) (k)

$$k = \sqrt{n}$$

2. Determinar a amplitude de cada intervalo (c)

$$c = \frac{A}{k-1},$$

em que $A = x_{max} - x_{min}$

3. Determinar a 1ª classe (ou intervalo) : $[LI; LS)$ ou $LI \vdash LS$

$$LI = x_{min} - \frac{c}{2} \qquad \text{e} \qquad LS = x_{min} + \frac{c}{2}$$

Classe	Notação	Denominação	Resultado
$[a, b)$	$a \vdash b$	Fechado em a, aberto em b	Inclui a, não inclui b
$(a, b]$	$a \dashv b$	Aberto em a, fechado em b	Não inclui a, inclui b

Dados de viscosidade: Método da raiz quadrada

Observa-se que $n = 18$, então

Dados de viscosidade: Método da raiz quadrada

Observa-se que $n = 18$, então

$$k = \sqrt{n} \; = \; \sqrt{18} \; \approx \; 4,242 \; = \; 4.$$

Dados de viscosidade: Método da raiz quadrada

Observa-se que $n = 18$, então

$$k = \sqrt{n} \ = \ \sqrt{18} \ \approx \ 4,242 \ = \ 4.$$

Assim, temos $k = 4$ **intervalos (ou classes) de frequências.**

Dados de viscosidade: Método da raiz quadrada

Observa-se que $n = 18$, então

$$k = \sqrt{n} \; = \; \sqrt{18} \; \approx \; 4,242 \; = \; 4.$$

Assim, temos $k = 4$ **intervalos (ou classes) de frequências.**

Como $A = x_{\text{máx}} - x_{\text{min}}$ então

Dados de viscosidade: Método da raiz quadrada

Observa-se que $n = 18$, então

$$k = \sqrt{n} \;=\; \sqrt{18} \;\approx\; 4,242 \;=\; 4.$$

Assim, temos $k = 4$ **intervalos (ou classes) de frequências.**

Como $A = x_{\text{máx}} - x_{\text{min}}$ então

$$c = \frac{A}{k-1} = \frac{4,6 - 1,570}{3} = 1,01$$

Portanto, **a amplitude de cada intervalo é** $c = 1,01$.

Dados de viscosidade: Método da raiz quadrada

Para determinar a 1ª classe (ou intervalo) : $[LI; LS)$, temos

$$LI = x_{min} - \frac{c}{2} \qquad \text{e} \qquad LS = x_{min} + \frac{c}{2}$$

Então,

$$LI = 1,570 - \frac{1,01}{2} = 1,065 \qquad \text{e} \qquad LS = 1,570 + \frac{c}{2} = 2,075$$

Portanto, 1ª classe (ou intervalo) é representado por

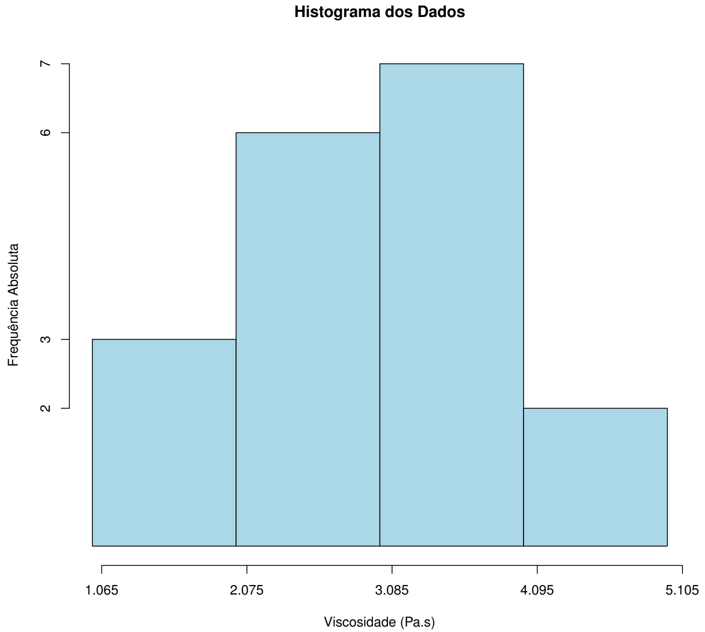
$$[1,065; 2,075) \quad \text{ou} \quad 1,065 \vdash 2,075$$

Dados de viscosidade: Tabela de frequências para variável quantitativa contínua

Viscosidade	f_i	f'_i	F_i	F'_i
1,065 $\vdash$ 2,075	3	0,167	3	0,167
2,075 $\vdash$ 3,085	6	0,333	9	0,5
3,085 $\vdash$ 4,095	7	0,389	16	0,889
4,095 $\vdash$ 5,105	2	0,111	18	1
Total	18	1	-	-

Histograma

- Muito utilizado para avaliar o comportamento de variáveis quantitativas, principalmente as contínuas
- Quando os dados estão agrupados em classes - o histograma apresenta as frequências das classes em colunas
- As frequências representadas podem ser simples ou relativas
- As colunas possuem bases com mesma largura
- Não existe espaço entre as classes


Histograma dos Dados
Frequência Absoluta
7
6
3
2
1.065
2.075
3.085
4.095
5.105
Viscosidade (Pa.s)

Medidas de Posição

- Às vezes, tem-se o interesse em resumir um conjunto de dados, apresentando um ou alguns valores que representem todo conjunto de dados.

- Para determinar esses valores, utiliza-se as **medidas de posição central**: Média, Mediana e Moda.

Média Amostral

- Sejam $X_1, \ldots, X_n$, n valores amostrais. A média amostral ($\bar{X}$) é definida por

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + \ldots + X_n}{n}$$

- Exemplo: Notas das provas dos alunos da UEL: 6,4,5,4,6

$$\bar{X} = \frac{6+4+5+4+6}{5} = 5,0.$$

- A média é considerada a **melhor medida de centro**, pois possui boas propriedades matemáticas
- Entretanto, ela é **falha** quando alguns dos valores estão muito **afastados da maioria dos dados.** (Valores extremos, valores aberrantes, outliers, etc).

Exemplo

Imagine o seguinte banco de dados: 6, 4, 5, 4, 50

Qual é o valor da média amostral ?

$$\bar{X} = \frac{6+4+5+4+50}{5} = 13,8.$$

Observa-se que a média amostral foi afetada por um valor aproximadamente 10 vezes maior do que a maioria dos dados na amostra. Algumas perguntas podem surgir nesse momento, como:

- O Valor $x = 50$ é um valor extremo (outlier)?
- Como podemos identificar um valor extremo ?
- É razoável pensar, neste caso, que a média não seja a melhor medida para representar a informação central dos dados. Neste sentido, o que podemos fazer para contornar esse problema?

Mediana e Moda

- A Mediana (Md) divide um conjunto ordenado (ordem crescente) de valores em duas partes iguais. É definida por

$$Md = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se} \quad \text{n for ímpar} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se} \quad \text{n for par} \end{cases}$$

- A Moda (Mo) é definida como a realização mais frequente do conjunto de dados observados.

Exemplo: Notas dos alunos da UEL

Imagine o seguinte banco de dados: 6, 4, 5, 4, 6

Qual é o valor da mediana ?

Exemplo: Notas dos alunos da UEL

Imagine o seguinte banco de dados: 6, 4, 5, 4, 6

Qual é o valor da mediana ?

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{6}^{x_5}$

Imagine o seguinte banco de dados: 6, 4, 5, 4, 6

Qual é o valor da mediana ?

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{6}^{x_5}$

Como $n = 5$, então a mediana é dada por

$$\text{Md} = \text{x}_{\left(\frac{n+1}{2}\right)} = \text{x}_{\left(\frac{5+1}{2}\right)} = x_3 = 5,0.$$

Exemplo: Notas dos alunos da UEL

Imagine o seguinte banco de dados: 6, 4, 5, 4, 6

Qual é o valor da mediana ?

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{6}^{x_5}$

Como $n = 5$, então a mediana é dada por

$$\text{Md} = \text{x}_{\left(\frac{n+1}{2}\right)} = \text{x}_{\left(\frac{5+1}{2}\right)} = x_3 = 5,0.$$

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \color{red}{\overbrace{5}^{x_3}}, \overbrace{6}^{x_4}, \overbrace{6}^{x_5}$

Exemplo: Notas dos alunos da UEL

Imagine o seguinte banco de dados: 6, 4, 5, 4, 6

Qual é o valor da mediana ?

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{6}^{x_5}$

Como $n = 5$, então a mediana é dada por

$$\text{Md} = \text{x}_{\left(\frac{n+1}{2}\right)} = \text{x}_{\left(\frac{5+1}{2}\right)} = x_3 = 5,0.$$

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{\color{red}{5}}^{\color{red}{x_3}}, \overbrace{6}^{x_4}, \overbrace{6}^{x_5}$

Observa-se que **os valores da média amostral e da mediana são iguais**.
Isso foi coincidência ou sempre teremos $\bar{X} = \text{Md}$?

Exemplo

Imagine o seguinte banco de dados: 6, 4, 5, 4, 50

Qual é o valor da mediana ?

Exemplo

Imagine o seguinte banco de dados: 6, 4, 5, 4, 50

Qual é o valor da mediana ?

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{50}^{x_5}$

Exemplo

Imagine o seguinte banco de dados: 6, 4, 5, 4, 50

Qual é o valor da mediana ?

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{50}^{x_5}$

Como $n = 5$, então a mediana é dada por

$$\text{Md} = \text{x}_{\left(\frac{n+1}{2}\right)} = \text{x}_{\left(\frac{5+1}{2}\right)} = x_3 = 5,0.$$

Exemplo

Imagine o seguinte banco de dados: 6, 4, 5, 4, 50

Qual é o valor da mediana ?

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{50}^{x_5}$

Como $n = 5$, então a mediana é dada por

$$\text{Md} = \text{x}_{\left(\frac{n+1}{2}\right)} = \text{x}_{\left(\frac{5+1}{2}\right)} = x_3 = 5,0.$$

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{50}^{x_5}$

Exemplo

Imagine o seguinte banco de dados: 6, 4, 5, 4, 50

Qual é o valor da mediana ?

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{50}^{x_5}$

Como $n = 5$, então a mediana é dada por

$$\text{Md} = \text{x}_{\left(\frac{n+1}{2}\right)} = \text{x}_{\left(\frac{5+1}{2}\right)} = x_3 = 5,0.$$

Dados ordenados: $\overbrace{4}^{x_1}, \overbrace{4}^{x_2}, \overbrace{5}^{x_3}, \overbrace{6}^{x_4}, \overbrace{50}^{x_5}$

Observa-se que os valores da média amostral e da mediana **são diferentes**. Isso significa **um indicativo** de que esse último conjunto de dados possua **pontos influentes** (pontos extremos ou outliers).

- Às vezes, tem-se o interesse em resumir um conjunto de dados, apresentando um ou alguns valores que representem todo conjunto de dados
- Exemplos:
  - Aluno A (variável X): 6,4,5,4,6
  - Aluno B (variável Y): 5,5,5,5,5
  - Aluno C (variável Z): 10,10,5,0,0
- Observa-se $\bar{X} = \bar{Y} = \bar{Z} = 5$, porém nada importa sobre suas diferentes variabilidades
- Dessa forma, deve-se construir medidas que sumarizem a variabilidade nos dados. Tais medidas são: **Variância amostral, desvio-padrão amostral e coeficiente de variação.**

Como medir ou quantificar variabilidade ?

Podemos pensar na soma dos **desvios** d_i, isto é, **na diferença de cada valor observado com a média amostral**, ou seja $d_i = x_i - \bar{x}$. Para o exemplo das notas do Aluno A, pode-se calcular os desvios

Como medir ou quantificar variabilidade ?

Podemos pensar na soma dos **desvios** d_i, isto é, **na diferença de cada valor observado com a média amostral**, ou seja $d_i = x_i - \bar{x}$. Para o exemplo das notas do Aluno A, pode-se calcular os desvios

$$d_1 = x_1 - \bar{x} = 6 - 5 = 1$$

Como medir ou quantificar variabilidade ?

Podemos pensar na soma dos **desvios** d_i, isto é, **na diferença de cada valor observado com a média amostral**, ou seja $d_i = x_i - \bar{x}$. Para o exemplo das notas do Aluno A, pode-se calcular os desvios

$$
\begin{aligned}
d_1 &= x_1 - \bar{x} = 6 - 5 = 1 \\
d_2 &= x_2 - \bar{x} = 4 - 5 = -1
\end{aligned}
$$

Podemos pensar na soma dos **desvios** d_i, isto é, **na diferença de cada valor observado com a média amostral**, ou seja $d_i = x_i - \bar{x}$. Para o exemplo das notas do Aluno A, pode-se calcular os desvios

$$
\begin{aligned}
d_1 &= x_1 - \bar{x} = 6 - 5 = 1 \\
d_2 &= x_2 - \bar{x} = 4 - 5 = -1 \\
d_3 &= x_3 - \bar{x} = 5 - 5 = 0
\end{aligned}
$$

Como medir ou quantificar variabilidade ?

Podemos pensar na soma dos **desvios** d_i, isto é, **na diferença de cada valor observado com a média amostral**, ou seja $d_i = x_i - \bar{x}$. Para o exemplo das notas do Aluno A, pode-se calcular os desvios

$$\begin{aligned}
d_1 &= x_1 - \bar{x} = 6 - 5 = 1 \\
d_2 &= x_2 - \bar{x} = 4 - 5 = -1 \\
d_3 &= x_3 - \bar{x} = 5 - 5 = 0 \\
d_4 &= x_4 - \bar{x} = 4 - 5 = -1
\end{aligned}$$

Como medir ou quantificar variabilidade ?

Podemos pensar na soma dos **desvios** d_i, isto é, **na diferença de cada valor observado com a média amostral**, ou seja $d_i = x_i - \bar{x}$. Para o exemplo das notas do Aluno A, pode-se calcular os desvios

$$
\begin{aligned}
d_1 &= x_1 - \bar{x} = 6 - 5 = 1 \\
d_2 &= x_2 - \bar{x} = 4 - 5 = -1 \\
d_3 &= x_3 - \bar{x} = 5 - 5 = 0 \\
d_4 &= x_4 - \bar{x} = 4 - 5 = -1 \\
d_5 &= x_5 - \bar{x} = 6 - 5 = 1
\end{aligned}
$$

Como medir ou quantificar variabilidade ?

Observa-se que a soma dos d_i's **é igual a zero**, para n observações !!!

Como medir ou quantificar variabilidade ?

Observa-se que a soma dos d_i's **é igual a zero**, para n observações !!!

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Como medir ou quantificar variabilidade ?

Observa-se que a soma dos d_i's **é igual a zero**, para n observações !!!

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Uma alternativa, natural, é utilizar o **quadrado dos desvios** e assim computar a variabilidade dos dados por meio da **soma dos quadrados dos desvios**. Temos então,

Como medir ou quantificar variabilidade ?

Observa-se que a soma dos d_i's **é igual a zero**, para n observações !!!

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Uma alternativa, natural, é utilizar o **quadrado dos desvios** e assim computar a variabilidade dos dados por meio da **soma dos quadrados dos desvios**. Temos então,

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

Como medir ou quantificar variabilidade ?

Observa-se que a soma dos d_i's **é igual a zero**, para n observações !!!

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Uma alternativa, natural, é utilizar o **quadrado dos desvios** e assim computar a variabilidade dos dados por meio da **soma dos quadrados dos desvios**. Temos então,

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

A utilização desse total **pode causar dificuldades** quando compara-se conjuntos de dados com um número de observações diferentes. Assim, **é mais conveniente exprimir essa quantidade como uma média**, isto é dividindo por $n - 1$.

Variância e desvio-padrão amostrais

- A variância amostral (s^2) fornece a dispersão dos dados em torno do valor central (média)

- Sejam $X_1, \ldots, X_n$, n valores amostrais com média $\bar{X}$. É definida por

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

- Observa-se que essa quantidade possui uma **medida de dimensão ao quadrado da dimensão dos dados** e assim, pode causar **problemas de interpretação.**

- O **desvio-padrão (s)** é definido como a raiz quadrada da variância

- O desvio-padrão indica, em média, qual será o erro (desvio) cometido ao tentar substituir cada observação pela média

Coeficiente de Variação (CV)

- O CV é uma medida de dispersão relativa utilizada para comparar variáveis.

$$CV = \frac{s}{\bar{X}}$$

- Observa-se que quanto menor for o valor do CV, mais **homogêneo** é um conjunto de dados

- Considere o conjunto de dados abaixo, com 80 observações, sobre a resistência à compressão de uma liga de Alumínio-Lítio.
- (i) Encontre a tabela de frequências e construa os seguintes gráficos: Histograma e polígono de frequências.
- (ii) Determine a média, variância, moda, mediana e o coeficiente de variação dos dados.

76 ; 120 ; 135 ; 149 ; 157 ; 163 ; 171 ; 178 ; 190 ; 207
87 ; 121 ; 141 ; 150 ; 158 ; 163 ; 171 ; 180 ; 193 ; 208
97 ; 123 ; 142 ; 150 ; 158 ; 165 ; 172 ; 180 ; 194 ; 218
101 ; 131 ; 143 ; 151 ; 158 ; 167 ; 174 ; 181 ; 196 ; 221
105 ; 133 ; 145 ; 153 ; 158 ; 167 ; 174 ; 181 ; 199 ; 228
110 ; 133 ; 146 ; 154 ; 160 ; 168 ; 175 ; 183 ; 199 ; 229
115 ; 134 ; 148 ; 154 ; 160 ; 169 ; 176 ; 184 ; 200 ; 237
118 ; 135 ; 149 ; 156 ; 160 ; 170 ; 176 ; 186 ; 201 ; 245

Quartis

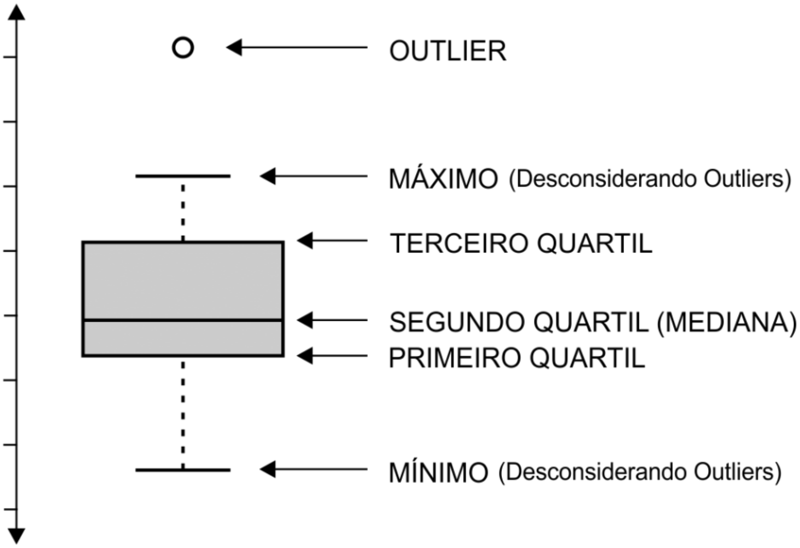
- Dê forma análoga a mediana, um conjunto de dados pode ser dividido em 4, 10, 100, etc ...
- Quartis: Divide o conjunto de dados em quatro partes iguais

$$\text{Quartis}: \begin{cases} Q_1 & \rightarrow & 1^o \text{ Quartil } (25\%) \\ Q_2 & \rightarrow & 2^o \text{ Quartil } (50\%) \\ Q_3 & \rightarrow & 3^o \text{ Quartil } (75\%) \end{cases}$$

- $Q_1 = X_{\left(\frac{n}{4}+\frac{1}{2}\right)}$ e $Q_3 = X_{\left(\frac{3n}{4}+\frac{1}{2}\right)}$

Box-plot

- É um gráfico construído com base no resumo de cinco medidas, constituído por:
  - Limite mínimo (L_1)
  - 1^o Quartil (Q_1)
  - 2^o Quartil (Q_2)
  - 3^o Quartil (Q_3)
  - Limite máximo (L_1')
- Para o cálculo do limite mínimo e máximo do boxplot, temos:
  - $L_1 = \text{máx}\{x_{min}\ ;\ Q_1 - 1,5(Q_3 - Q_1)\}$
  - $L_1' = \min\{x_{max}\ ;\ Q_3 + 1,5(Q_3 - Q_1)\}$
- Por meio do box-plot pode-se verificar: Assimetria nos dados, discutir pontos influentes (possíveis outliers), estudar a variabilidade do conjunto de dados.


OUTLIER
MÁXIMO (Desconsiderando Outliers)
TERCEIRO QUARTIL
SEGUNDO QUARTIL (MEDIANA)
PRIMEIRO QUARTIL
MÍNIMO (Desconsiderando Outliers)

Exercício 3

- (a) Utilizando os dados do exemplo 2 (viscosidade num processo químico), construa o gráfico boxplot. Comente sobre a possibilidade de pontos influentes nos dados.

- (b) Utilizando os dados do exercício 2 (resistência à compressão de uma liga de Alumínio-Lítio), construa o gráfico boxplot. Comente sobre a possibilidade de pontos influentes nos dados.

Exercício 4

- Um artigo no *Transactions of the Chemical Engineers* reportou dados sobre um experimento que investigou o efeito de variáveis de processo de oxidação em fase vapor de naftaleno. Uma amostra da conversão percentual molar de naftaleno em anidrino maléico resulta em:

$$10.20, 12.00, 13.34, 13.54, 13.59, 13.61, 14.06, 14.11, 14.14, 14.17$$
$$14.30, 14.62, 14.77, 14.78, 14.85, 15.06, 15.26, 15.55, 15.63, 15.94$$
$$16.07, 16.08, 16.12, 16.39, 16.51, 16.78, 17.09, 17.11, 17.26, 17.46$$
$$17.59, 17.60, 17.76, 17.78, 18.05, 19.12, 19.18, 19.21, 19.36, 19.38$$
$$19.56, 19.70, 20.04, 20.08, 20.35, 20.39, 20.48, 20.67, 20.90, 22.10$$

- (a) Encontre a média, variância, mediana, os quartis e o coeficiente de variação dos dados brutos.
- (b) Construa o histograma, polígono de frequências e o box-plot.
- (c) Comente sobre a variabilidade, pontos discrepantes e assimetria dos dados acima.