Aula 03 – Introdução à Estatística Descritiva – Medidas Descritivas

Luciana Rocha Pedro

GCC 1518 – Estatística e Probabilidade – CEFET Maracanã

USP – Projeto de Ensino

Aprender Fazendo Estatística

Medidas Descritivas

Uma outra maneira de resumirmos os dados de uma variável quantitativa, além de tabelas e gráficos, é apresentá-los na forma de valores numéricos, denominados **medidas descritivas**.

Estas medidas, se calculadas a partir de dados populacionais, são denominadas **parâmetros** e, se calculadas a partir de dados amostrais, são denominadas **estimadores** ou **estatísticas**.

As medidas descritivas auxiliam a análise do comportamento dos dados. Tais dados são provenientes de uma população ou de uma amostra, o que exige uma notação específica para cada caso.

Medidas Descritivas

Classificamos as medidas descritivas como: **medidas de posição** (tendência central e separatrizes), **medidas de dispersão**, **medidas de assimetria** e **medidas de curtose**.

Quadro 01: Notações de algumas estatísticas.

Medidas	Parâmetros	Estimadores
Número de elementos	N	n
Média	μ	$\overline{X}$
Variância	σ^2	S^2
Desvio padrão	σ	S

Medidas de Tendência Central

As **medidas de tendência central** são assim denominadas por indicarem um ponto em torno do qual se concentram os dados. Este ponto tende a ser o **centro** da distribuição dos dados.

A seguir, são definidas as principais medidas de tendência central:

- **média**,
- **mediana**,
- **moda**.

Média Aritmética

A **média aritmética** ($\bar{x}$) é a soma de todos os valores observados da variável dividida pelo número total de observações.

Sob uma visão geométrica, a média de uma distribuição é o centro de gravidade e representa o ponto de equilíbrio de um conjunto de dados.

É a medida de tendência central mais utilizada para representar um conjunto de dados.

Média Aritmética

Seja $X = (x_1, \ldots, x_n)$ um conjunto de dados. A média é dada por:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \text{ou} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

para dados populacionais ou amostrais, respectivamente.

Média Aritmética

Caso os dados estejam apresentados segundo uma distribuição de freqüência, temos:

$$\mu = \frac{\sum_{i=1}^{k} x_i \cdot F_i}{N} \qquad \text{ou} \qquad \bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot F_i}{n}.$$

Observe que, no caso de dados agrupados, a média é obtida a partir de uma ponderação, em que os pesos são as freqüências absolutas de cada classe e x_i é o ponto médio da classe i.

Propriedades da Média Aritmética

Citamos, a seguir, algumas propriedades da média aritmética:

1. A média é um valor calculado facilmente e depende de todas as observações.

2. A média é única em um conjunto de dados e nem sempre tem existência real, ou seja, nem sempre é igual a um determinado valor observado.

3. A média é afetada por valores extremos observados.

Propriedades da Média Aritmética

4. Por depender de todos os valores observados, qualquer modificação nos dados fará com que a média tenha a mesma alteração. Isto quer dizer que somando-se, subtraindo-se, multiplicando-se ou dividindo-se uma constante a cada valor observado, a média ficará acrescida, diminuída, multiplicada ou dividida por esse valor.

5. A soma da diferença de cada valor observado em relação à média é zero, ou seja, a soma dos desvios é zero

$$\sum(x_i - \bar{x}) = 0.$$

Observações

A propriedade 5 é de extrema importância para a definição da **variância**, uma medida de dispersão a ser definida posteriormente.

Destacamos, ainda, que a propriedade 3, quando temos um conjunto de dados discrepantes, faz da média uma medida não apropriada para representar os dados.

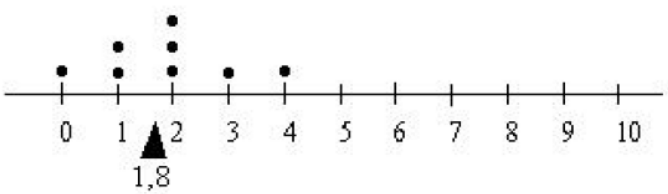
Neste caso, não existe uma regra prática para a escolha de uma outra medida. O ideal é, a partir da experiência do pesquisador, decidir pela **moda** ou **mediana**.

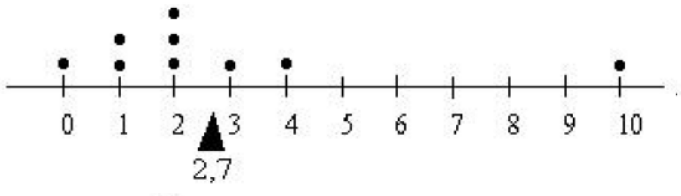
Exemplo

Para ilustrar, considere o número de filhos, por família, para um grupo de 8 famílias: $0, 1, 1, 2, 2, 2, 3, 4$. Neste caso, a média é $x = 1,875$ filhos por família.

Entretanto, incluindo ao grupo uma nova família com 10 filhos, a média passa a ser $x = 2,788$, o que eleva em $48,16\%$ o número médio de filhos por família.

Assim, ao observar a média, podemos pensar que a maior parte das famílias deste grupo tem três filhos quando, na verdade, apenas uma tem três filhos.





Exemplo

Considerando a idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, a idade média é

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{20 + 26 + 18 + \ldots + 21 + 22}{22} = \frac{518}{22} = 23,5 \text{ anos.}$$

Portanto, a idade média dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá é 23,5 anos.

Exemplo

No entanto, ao considerar os dados agrupados como na Tabela 10, a média é:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot F_i}{n} = \frac{20 \cdot 11 + 24 \cdot 6 + \ldots + 36 \cdot 2}{22} = \frac{524}{22} = 23,8 \text{ anos.}$$

Notamos que esta diferença ocorre devido ao fato de utilizarmos os dados sem o conhecimento de seus valores individuais.

Neste caso, tornou-se necessário representá-los pelos pontos médios de suas respectivas classes, resultando numa certa perda de informação.

Tabela 10

Tabela 10 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Idade	x_i	F_i	f_i %	F_{a_i}	f_{a_i} %
18 |---22	20	11	50,00	11	50,00
22 |---26	24	6	27,27	17	77,27
26 |---30	28	2	9,09	19	86,36
30 |---34	32	1	4,55	20	90,91
34 |---38	36	2	9,09	22	100,00
Total	-	22	100,00	-	-

Fonte: Tabela 01.

Exercício 06

Calcule a média aritmética para a variável altura dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM:

1. utilizando os dados brutos;

2. utilizando a distribuição de freqüência (dados agrupados).

Moda

A **moda** (M_o) é o valor que apresenta a maior freqüência da variável entre os valores observados.

Para o caso de valores individuais, a moda pode ser determinada imediatamente observando-se o rol ou a freqüência absoluta dos dados.

Por outro lado, em se tratando de uma distribuição de freqüência de valores agrupados em classes, primeiramente é necessário identificar a classe modal, aquela que apresenta a maior freqüência e, a seguir, a moda pode ser calculada.

Moda

A moda é calculada aplicando-se a fórmula:

$$M_o = l_i + \frac{h \cdot (F_i - F_{i-1})}{(F_i - F_{i-1}) + (F_i - F_{i+1})},$$

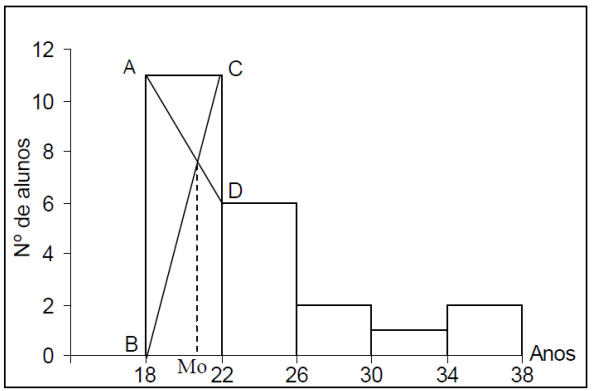
em que

- i é a ordem da classe modal;
- l_i é o limite inferior da classe modal;
- h é a amplitude da classe modal;
- F_i é a freqüência absoluta da classe modal;
- F_{i-1} é a freqüência absoluta da classe anterior à classe modal;
- F_{i+1} é a freqüência absoluta da classe posterior à classe modal.

Moda – Representação Gráfica

Graficamente, utilizando um conjunto de dados hipotéticos, identificamos a classe modal como aquela que apresenta o retângulo de maior altura (freqüência).

A intersecção das retas que unem os pontos AD e os pontos BC determina o ponto P que, projetado perpendicularmente no eixo da variável, corresponderá ao valor da moda M_o.



Observações

É relevante salientar que um conjunto de dados pode apresentar todos os seus elementos com a mesma freqüência absoluta e, neste caso, não existirá um valor modal, o que significa que a distribuição será classificada como **amodal**.

Pode ocorrer, também, casos em que a seqüência de observações apresente vários elementos com freqüência iguais, implicando numa distribuição **plurimodal**.

O uso da moda é mais indicado quando se deseja obter, rapidamente, uma medida de tendência central. Um outro aspecto que favorece a utilização da moda é que seu valor não é afetado pelos valores extremos do conjunto de dados analisado.

Exemplo

A moda da idade dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, determinada pontualmente, é $M_o = 20$ anos. Isto significa que a idade mais freqüente entre estes alunos é de 20 anos.

Ao considerar a distribuição apresentada na Tabela 10, a moda é

$$\begin{aligned} M_o &= l_i + \frac{h \cdot (F_i - F_{i-1})}{(F_i - F_{i-1}) + (F_i - F_{i+1})} \\ &= 18 + \frac{4 \cdot (11 - 0)}{(11 - 0) + (11 - 6)} = 18 + \frac{44}{16} = 18 + 2,75 = 20,75 \text{ anos.} \end{aligned}$$

A interpretação é análoga à determinada pontualmente.

Tabela 10

Tabela 10 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Idade	x_i	F_i	f_i %	F_{a_i}	f_{a_i} %
18 |---22	20	11	50,00	11	50,00
22 |---26	24	6	27,27	17	77,27
26 |---30	28	2	9,09	19	86,36
30 |---34	32	1	4,55	20	90,91
34 |---38	36	2	9,09	22	100,00
Total	-	22	100,00	-	-

Fonte: Tabela 01.

Exercício 07

Calcule a moda para a variável altura dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM.

1. utilizando os dados brutos;

2. utilizando a distribuição de freqüência (dados agrupados).

Mediana

A **mediana** (M_d) é o valor que ocupa a posição central da série de observações de uma variável, em rol, dividindo o conjunto em duas partes iguais, ou seja, a quantidade de valores inferiores à mediana é igual à quantidade de valores superiores.

Retomando o exemplo do número de filhos por famílias, verificamos que, para o caso de oito famílias, $n = 8$, a mediana é determinada como a seguir:

X	x_1	x_2	x_3	x_4		x_5	x_6	x_7	x_8
Valor observado	0	1	1	2	$\frac{x_4+x_5}{2}$	2	2	3	4
	← 4 observações →				Md=2	← 4 observações →			

Mediana

Quando acrescentamos ao grupo uma outra família com 10 filhos, o tamanho da amostra passa a ser n=9. Neste caso, a mediana é:

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Valor observado	0	1	1	1	2	2	3	4	10

$\longleftarrow$ 4 observações $\longrightarrow$ Md=2 $\longleftarrow$ 4 observações $\longrightarrow$

Observe que, nos dois casos, por coincidência, a mediana manteve-se a mesma, $M_d = 2$, significando que 50% das famílias possuem menos de 2 filhos e 50% possuem mais de 2 filhos.

Mostramos, assim, que a mediana não é influenciada por valores extremos.

Mediana

Este procedimento pode ser inadequado quando o conjunto de dados for composto por muitos elementos. Os passos a seguir indicam uma forma para o cálculo da mediana, independentemente do tamanho da amostra.

- Ordenar as observações em ordem crescente ou decrescente (rol).
- Calcular a posição, p, que a mediana ocupa no conjunto de dados:

$$p = 0,5 \cdot (n + 1)$$

- Obter a mediana pela equação

$$M_d = x_{I_p} + F_p \cdot (x_{I_{p+1}} - x_{I_p})$$

em que I_p é a parte inteira de p e F_p a parte fracionária (ou decimal).

Exemplo

Considere o rol da idade dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM:

18, 18, 19, 20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 25, 25, 25, 26, 29, 30, 35, 37

A posição p da mediana é

$$p = 0,5 \cdot (22 + 1) = 11,5.$$

Assim,

$$M_d = x_{11} + 0,5 \cdot (x_{12} - x_{11}) = 21 + 0,5 \cdot (22 - 21) = 21,5 \text{ anos.}$$

Logo, 50% dos alunos têm idade inferior a 21,5 anos.

Mediana

Para os dados em distribuição de freqüências em classes, temos:

$$M_d = l_i + \frac{h \cdot (p - F_{a_{i-1}})}{F_i}$$

em que:

- $p = \frac{n}{2}$ indica a posição central da série;
- i é a ordem da classe que contém o menor valor de F_{a_i} tal que $F_{a_i} \geq p$;
- $F_{a_{i-1}}$ é a freqüência acumulada da classe anterior à da mediana;
- F_i é a freqüência absoluta da classe i.

Exemplo

Ao considerar a distribuição apresentada na Tabela 10, a mediana é

$$p = \frac{22}{2} = 11 \Rightarrow F_{a_i} \geq 11 \Rightarrow i = 1$$

$$\begin{aligned} M_d &= l_i + \frac{h \cdot (p - F_{a_{i-1}})}{F_i} \\ &= 18 + \frac{4 \cdot (11 - 0)}{11} = 18 + \frac{44}{11} = 18 + 4 = 22 \text{ anos.} \end{aligned}$$

A idade mediana é 22 anos, ou seja, 50% dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM têm idade inferior ou igual a 22 anos.

Tabela 10

Tabela 10 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Idade	x_i	F_i	f_i %	F_{a_i}	f_{a_i} %
18 |---22	20	11	50,00	11	50,00
22 |---26	24	6	27,27	17	77,27
26 |---30	28	2	9,09	19	86,36
30 |---34	32	1	4,55	20	90,91
34 |---38	36	2	9,09	22	100,00
Total	-	22	100,00	-	-

Fonte: Tabela 01.

Mediana – Representação Gráfica

Para ilustrarmos graficamente o cálculo da mediana, considere novamente um conjunto de pesos fictícios.

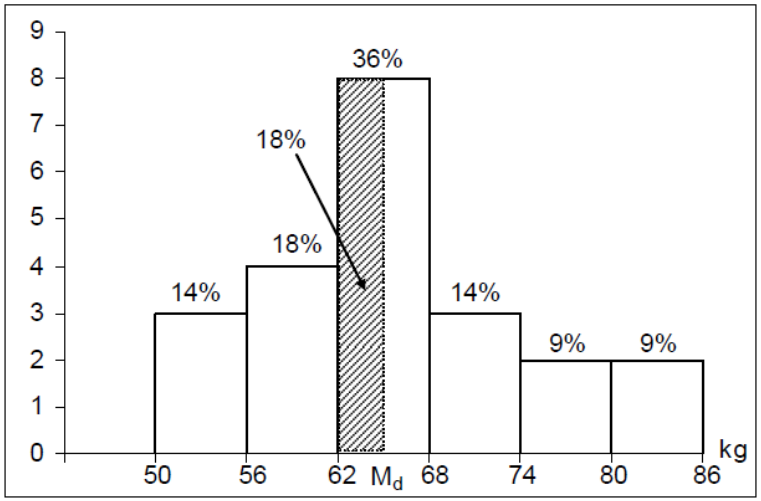
Devemos localizar, no eixo da variável, o ponto que divide o histograma ao meio. Isto é feito somando-se as áreas (freqüências relativas) até que obtenhamos 50%.

No histograma a seguir, a classe que contém a mediana é a classe de 62 a 68 kg, com freqüência relativa igual a 36%.

Mediana – Representação Gráfica

Podemos observar, então, que faltam 18%, $50\% - (14\% + 18\%)$ para completar 50% da distribuição.

Portanto, o limite superior da base do retângulo hachurado é a mediana da distribuição.



Mediana – Representação Gráfica

Aplicando a proporcionalidade entre a área e a base do retângulo teremos a mediana:

$$\frac{68-62}{36\%} = \frac{M_d - 62\%}{18\%}.$$

Portanto, a mediana é igual a 65 kg.

Exercício 08

Calcule a mediana para a variável altura dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM.

1. utilizando os dados brutos;

2. utilizando a distribuição de freqüência (dados agrupados).

Medidas Separatrizes

As **medidas separatrizes** são valores que ocupam posições no conjunto de dados, em rol, dividindo-o em partes iguais e podem ser:

- **Quartil**: Os quartis dividem o conjunto de dados em quatro partes iguais.
- **Decil**: Os decis dividem o conjunto de dados em dez partes iguais.
- **Percentil**: Os percentis dividem o conjunto de dados em cem partes iguais.

Quartil

Quartil: Os quartis dividem o conjunto de dados em quatro partes iguais.

Quadro 02: Descrição dos quartis (dados amostrais).

Estatística	Notação	Interpretação	Posição
1º quartil	Q_1	25% dos dados são valores menores ou iguais ao valor do primeiro quartil.	p=0,25(n+1)
2º quartil	$Q_2 = M_d$	50% dos dados são valores menores ou iguais ao valor do segundo quartil.	p=0,50(n+1)
3º quartil	Q_3	75% dos dados são valores menores ou iguais ao valor do terceiro quartil.	p=0,75(n+1)

Decil

Decil: Os decis dividem o conjunto de dados em dez partes iguais.

Quadro 03: Descrição dos decis (dados amostrais).

Estatística	Notação	Interpretação	Posição
1º decil	D_1	10% dos dados são valores menores ou iguais ao valor do primeiro decil.	p=0,10(n+1)
2º decil	D_2	20% dos dados são valores menores ou iguais ao valor do segundo decil.	p=0,20(n+1)
3º decil	D_3	30% dos dados são valores menores ou iguais ao valor do terceiro decil.	p=0,30(n+1)
4º decil	D_4	40% dos dados são valores menores ou iguais ao valor do primeiro decil.	p=0,40(n+1)
5º decil	$D_5=Q_2=Md$	50% dos dados são valores menores ou iguais ao valor do segundo decil.	p=0,50(n+1)
6º decil	D_6	60% dos dados são valores menores ou iguais ao valor do terceiro decil.	p=0,60(n+1)
7º decil	D_7	70% dos dados são valores menores ou iguais ao valor do primeiro decil.	p=0,70(n+1)
8º decil	D_8	80% dos dados são valores menores ou iguais ao valor do segundo decil.	p=0,80(n+1)
9º decil	D_9	90% dos dados são valores menores ou iguais ao valor do terceiro decil.	p=0,90(n+1)

Percentil

Percentil: Os percentis dividem o conjunto de dados em cem partes iguais.

Quadro 04: Descrição de alguns percentis (dados amostrais).

Estatística	Notação	Interpretação	Posição
5º Percentil	P_5	5% dos dados são valores menores ou iguais ao valor do primeiro percentil.	p=0,05(n+1)
10º Percentil	P_{10}	10% dos dados são valores menores ou iguais ao valor do décimo percentil.	p=0,10(n+1)
25º Percentil	$P_{25}=Q_1$	25% dos dados são valores menores ou iguais ao valor do percentil cinquenta.	p=0,25(n+1)
50º Percentil	$P_{50}= D_5 =Q_2 = Md$	50% dos dados são valores menores ou iguais ao valor do primeiro percentil.	p=0,50(n+1)
75º Percentil	$P_{75}=Q_3$	75% dos dados são valores menores ou iguais ao valor do primeiro percentil. (Q_3)	p=0,75(n+1)
90º Percentil	P_{90}	90% dos dados são valores menores ou iguais ao valor do percentil noventa.	p=0,90(n+1)
95º Percentil	P_{95}	95% dos dados são valores menores ou iguais ao valor do percentil noventa e cinco.	p=0,95(n+1)

Cálculo das Medidas Separatrizes

Para os dados em rol, o cálculo das medidas separatrizes é o mesmo que o da mediana, ou seja:

$$S_k = x_{I_p} + F_p \cdot (x_{I_{p+1}} - x_{I_p})$$

em que I_p é a parte inteira de p e F_p a parte fracionária (ou decimal).

Cálculo das Medidas Separatrizes

Para os dados em distribuição de freqüências em classes, o cálculo das medidas separatrizes é o mesmo que o da mediana, ou seja:

$$S_K = l_i + \frac{h \cdot (p - F_{a_{i-1}})}{F_i}$$

em que

- $p = \frac{n}{4} \cdot k$, com $k = 1, 2, 3$ para determinação dos quartis;
- $p = \frac{n}{10} \cdot k$, com $k = 1, 2, \ldots, 9$ para o cálculo dos decis;
- $p = \frac{n}{100} \cdot k$, com $k = 1, 2, \ldots, 99$ para os percentis;
- i é a ordem da classe que contém o menor valor de F_{a_i} tal que $F_{a_i} \geq p$;
- $F_{a_{i-1}}$ é a freqüência acumulada da classe anterior à da separatriz.

Exemplo

Considere, novamente, o rol da idade dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM:

18, 18, 19, 20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 25, 25, 25, 26, 29, 30, 35, 37

Calcule:

- o terceiro quartil,
- o quadragésimo percentil,

Exemplo

Terceiro quartil: $p = 0,75 \cdot (22 + 1) = 17,25$ e

$$Q_3 = S_3 = x_{17} + 0,25 \cdot (x_{18} - x_{17}) = 25 + 0,25 \cdot (26 - 25) = 25,25 \text{ anos.}$$

Assim, podemos afirmar que 75% dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM têm idade inferior ou igual a 25,25 anos.

Exemplo

Quadragésimo percentil: $p = 0,40 \cdot (22 + 1) = 9,2$ e

$$P_{40} = S_{40} = x_9 + 0,20 \cdot (x_{10} - x_9) = 20 + 0,20 \cdot (21 - 20) = 20,2 \text{ anos.}$$

Logo, 40% dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM têm idade inferior ou igual a 20,2 anos.

Exemplo

Considerando, agora, a Tabela 10, calcule:

- o primeiro quartil,
- o terceiro quartil,
- o sétimo decil,
- o nonagésimo percentil.

Tabela 10 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

Idade	x_i	F_i	f_i %	F_{a_i}	f_{a_i} %
18 |---22	20	11	50,00	11	50,00
22 |---26	24	6	27,27	17	77,27
26 |---30	28	2	9,09	19	86,36
30 |---34	32	1	4,55	20	90,91
34 |---38	36	2	9,09	22	100,00
Total	-	22	100,00	-	-

Fonte: Tabela 01.

Exemplo

Primeiro quartil: $p = \frac{n}{4} \cdot k = \frac{22}{4} \cdot 1 = 5,5 \Rightarrow F_{a_i} \geq 5,5 \Rightarrow i = 1$

$$Q_1 = l_1 + \frac{h \cdot (p - F_{a_{1-1}})}{F_1} = 18 + \frac{4 \cdot (5,5 - 0)}{11} = 20 \text{ anos.}$$

Terceiro quartil: $p = \frac{n}{4} \cdot k = \frac{22}{4} \cdot 3 = 16,5 \Rightarrow F_{a_i} \geq 16,5 \Rightarrow i = 2$

$$Q_3 = l_2 + \frac{h \cdot (p - F_{a_{2-1}})}{F_2} = 22 + \frac{4 \cdot (16,5 - 11)}{6} = 25,67 \text{ anos.}$$

Exemplo

Sétimo decil: $p = \frac{n}{10} \cdot k = \frac{22}{10} \cdot 7 = 15,4 \Rightarrow F_{a_i} \geq 15,4 \Rightarrow i = 2$

$$D_7 = l_2 + \frac{h \cdot (p - F_{a_1})}{F_2} = 22 + \frac{4 \cdot (15,4 - 11)}{6} = 24,93 \text{ anos.}$$

Nonagésimo percentil: $p = \frac{n}{100} \cdot k = \frac{22}{100} \cdot 90 = 19,8 \Rightarrow F_{a_i} \geq 19,8 \Rightarrow i = 4$

$$P_{90} = l_4 + \frac{h \cdot (p - F_{a_3})}{F_4} = 30 + \frac{4 \cdot (19,8 - 19)}{1} = 33,2 \text{ anos.}$$

Exemplo

Concluímos, que

- 25% dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM têm idade inferior ou igual a 20 anos,
- 75% tem idade inferior a 25,67 anos,
- 70% tem idade inferior a 24,93 anos,
- 90% tem idade inferior a 33,2 anos.

Medidas de Dispersão

Fenômenos que envolvem análises estatísticas caracterizam-se por suas semelhanças e variabilidades.

As **medidas de dispersão** auxiliam as **medidas de tendência central** a descrever o conjunto de dados adequadamente.

As medidas de dispersão indicam se os dados estão, ou não, próximos uns dos outros.

Medidas de Dispersão

Desta forma, não há sentido em calcularmos a média de um conjunto em que não haja variação dos seus elementos. Existe ausência de dispersão e a medida de dispersão é igual a zero.

Por outro lado, aumentando a dispersão, o valor da média varia e, se a variação for muito grande, a média não será uma medida de tendência central representativa.

É necessário, portanto, ao menos **uma medida de tendência central** e **uma medida de dispersão** para descrever um conjunto de dados.

Medidas de Dispersão

As quatro medidas de dispersão que serão definidas a seguir são:

- amplitude total,
- amplitude interquartílica,
- desvio padrão,
- variância.

Com exceção da primeira, todas têm como ponto de referência a **média**.

Amplitude Total

A **amplitude total** de um conjunto de dados é a diferença entre o maior e o menor valor observado.

$$AT = x_{max} - x_{min}$$

Esta medida de dispersão não leva em consideração os valores intermediários, perdendo a informação de como os dados estão distribuídos e/ou concentrados.

Exemplo

A amplitude total da idade dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM é

$$AT = 37 - 18 = 19 \text{ anos},$$

isto é, as idades dos alunos diferem em 19 anos.

Amplitude Interquartílica

A **amplitude interquartílica** é a diferença entre o terceiro e o primeiro quartil.

Esta medida é mais estável que a amplitude total por não considerar os valores mais extremos.

Esta medida abrange 50% dos dados e é útil para detectar valores discrepantes.

$$d_q = Q_3 - Q_1.$$

Amplitude Semi-Interquartílica

Por outro lado, a **amplitude semi-interquartílica** é definida como a metade da diferença entre os quartis:

$$dq_m = \frac{Q_3 - Q_1}{2}.$$

Exemplo

A amplitude interquartílica da idade dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM, considerando a Tabela 10, é:

$$dq = 25,67 - 20 = 5,67 \text{ anos.}$$

A amplitude entre o terceiro e primeiro quartil, que envolve 50% (centrais) dos alunos, é de 5,67 anos.

Exemplo

Calculando a amplitude semi-interquartílica da idade dos alunos que cursam a disciplina Inferência Estatística do curso de Estatística da UEM, temos:

$$dq_m = 2,84 \text{ anos.}$$

Observamos que a distância entre a mediana e o primeiro quartil, $(22 - 20)$, é 2. Como $2 < 2,84$, isto indica que há uma concentração de dados à esquerda da mediana e que os dados localizados à direita da mediana são mais dispersos.

Desvio

A diferença entre cada valor observado e a média é denominado **desvio** e é dado por $(x_i - \mu)$ se o conjunto de dados é populacional e por $(x_i - \bar{x})$ se os dados são amostrais.

Ao somar todos os desvios, ou seja, ao somar todas as diferenças de cada valor observado em relação a média, o resultado é igual a zero (propriedade 5 da média).

Isto significa que esta medida não mede a variabilidade dos dados.

Desvio Médio

Para resolvermos este problema, podemos considerar as diferenças em módulo. A média destas diferenças, em módulo, é denominada **desvio médio**:

$$d_m = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} \qquad \text{ou} \qquad d_m = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n},$$

para dados populacionais ou amostrais, respectivamente. Caso os dados estejam apresentados segundo uma distribuição de freqüência, temos:

$$d_m = \frac{\sum_{i=1}^{N} |x_i - \mu| \cdot F_i}{N} \qquad \text{ou} \qquad d_m = \frac{\sum_{i=1}^{n} |x_i - \bar{x}| \cdot F_i}{n},$$

Variância

Não há nada conceitualmente errado em considerarmos o desvio médio, mas esta medida não tem certas propriedades importantes e não é muito utilizada.

O mais comum é considerarmos o quadrado dos desvios em relação à média e, então, calcularmos a medida. Obtemos, assim, a **variância**, que é definida por:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad \text{ou} \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1},$$

se os dados são populacionais ou amostrais, respectivamente.

Variância

Caso os dados estejam apresentados segundo uma distribuição de freqüência, temos:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2 \cdot F_i}{N} \qquad \text{ou} \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot F_i}{n-1},$$

Desvio Padrão

Entretanto, ao calcularmos a variância, observamos que o resultado será dado em unidades quadráticas, o que dificulta a sua interpretação.

O problema é resolvido extraindo-se a raiz quadrada da variância, definindo-se, assim, o **desvio padrão**:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \qquad \text{ou} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}},$$

se os dados são populacionais ou amostrais.

Desvio Padrão

Se estiverem em distribuição de freqüências, o desvio-padrão é definido como

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2 \cdot F_i}{N}} \qquad \text{ou} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot F_i}{n-1}}.$$

Geralmente, o desvio padrão é maior ou igual ao desvio médio, pois o cálculo do desvio-padrão considera cada desvio elevado ao quadrado, aumentando desproporcionalmente o peso dos valores extremos.

Exemplo

É importante destacar que, se duas populações apresentam a mesma média, mas os desvios padrão não são iguais, isto significa que as populações não têm o mesmo comportamento.

Considere três alunos cujas notas em uma disciplina estão apresentadas na Tabela 13.

Observamos que as médias das notas dos três alunos são iguais. Porém, seus desvios em torno da média são diferentes.

Isto quer dizer que seus desempenhos são diferentes.

Exemplo

O aluno A é constante em seu desempenho, o aluno B vai progredindo aos poucos e o aluno C diminui abruptamente seu desempenho. Em outras palavras, apesar dos três alunos terem o mesmo desempenho médio, a variabilidade difere.

Tabela 13. Notas, desvios e média dos alunos em uma disciplina.

Aluno	Notas	Soma	Média μ	$d=x_i-\mu$	$\lvert x_i-\mu\rvert$	$(x_i-\mu)^2$	$\sqrt{\sum(x_i-\mu)^2}$
A	8	40	8	0	0	0	$\sqrt{0}=0$
	8			0	0	0	
	8			0	0	0	
	8			0	0	0	
	8			0	0	0	
Total				0	0	0	
B	6	40	8	-2	2	4	$\sqrt{16}=4$
	6			-2	2	4	
	8			0	0	0	
	10			2	2	4	
	10			2	2	4	
Total				0	8	16	
C	10	40	8	2	2	4	$\sqrt{30}=5{,}48$
	10			2	2	4	
	10			2	2	4	
	5			-3	3	9	
	5			-3	3	9	
Total				0	12	30	

Exemplo

Retomando a idade dos alunos apresentada na Tabela 10, temos:

$$\text{Desvio médio: } d_m = \frac{|20 - 23,8| \cdot 11 + \ldots + |36 - 23,8| \cdot 2}{22} = 3,82 \text{ anos.}$$

$$\text{Variância: } \sigma^2 = \frac{(20 - 23,8)^2 \cdot 11 + \ldots + (36 - 23,8)^2 \cdot 2}{22 - 1} = 26,63 \text{ anos.}$$

$$\text{Desvio padrão: } \sigma = \sqrt{26,63} = 5,16 \text{ anos.}$$

Coeficiente de Variação

O **coeficiente de variação** é uma medida de dispersão relativa definida como a razão entre o desvio padrão e a média:

$$CV = \frac{\sigma}{\mu} \cdot 100 \qquad \text{ou} \qquad CV = \frac{s}{\bar{x}} \cdot 100,$$

se os dados são populacionais ou amostrais.

Coeficiente de Variação

A partir do coeficiente de variação podemos avaliar a homogeneidade do conjunto de dados e, conseqüentemente, se a média é uma boa medida para representar estes dados.

O coeficiente de variação também pode ser utilizado para comparar conjuntos com unidades de medidas distintas.

Uma desvantagem do coeficiente de variação é que ele deixa de ser útil quando a média está próxima de zero. Uma média muito próxima de zero pode inflacionar o coeficiente de variação.

Coeficiente de Variação

Um coeficiente de variação superior a 50% sugere alta dispersão, o que indica heterogeneidade dos dados. Quanto maior for este valor, menos representativa será a média.

Neste caso, optamos pela mediana ou moda, não existindo uma regra prática para a escolha de uma destas medidas. O pesquisador, com sua experiência, é que deverá decidir por uma ou outra.

Por outro lado, quanto mais próximo de zero, mais homogêneo é o conjunto de dados e mais representativa será a sua média.

Exemplo

Para as idades apresentadas na Tabela 10, temos:

$$CV = \frac{5,16}{23,8} \cdot 100 = 21,68\%.$$

Como $CV < 50\%$, podemos afirmar que a média é uma medida descritiva representativa para a variável idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá do ano de 2002.

Exercício 09

Calcule as medidas de dispersão para a variável altura da Tabela 10.

Medidas de Assimetria

A **medida de assimetria** é um indicador da forma da distribuição dos dados.

Ao construir uma distribuição de freqüências e/ou um histograma, estamos buscando, também, identificar visualmente a forma da distribuição dos dados, fornecida pelo coeficiente de assimetria de Pearson (A_s), definido como:

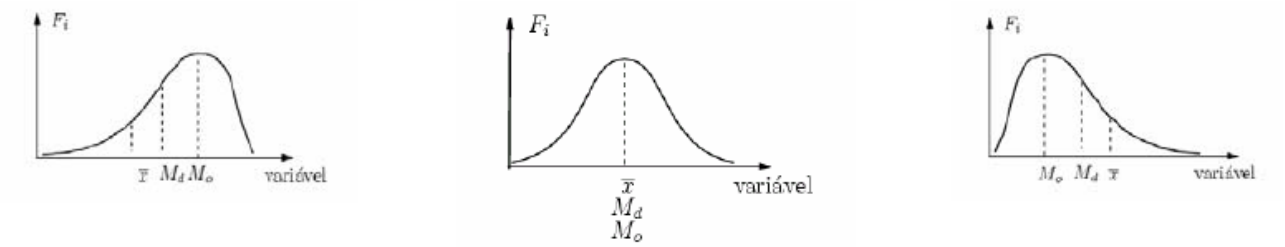
$$A_s = \frac{\mu - M_o}{\sigma} \qquad \text{ou} \qquad A_s = \frac{\bar{x} - M_o}{s},$$

para dados populacionais e amostrais, respectivamente.

Classificação das Distribuições

Uma distribuição é classificada como:

- **simétrica**: se média = mediana = moda ou $A_s = 0$;
- **assimétrica negativa**: se média $\leq$ mediana $\leq$ moda ou $A_s < 0$. O lado mais longo do polígono de freqüência (cauda da distribuição) está à esquerda do centro.
- **assimétrica positiva**: se moda $\leq$ mediana $\leq$ média ou $A_s > 0$. O lado mais longo do polígono de freqüência está à direita do centro.



Assimétrica negativa | Simétrica | Assimétrica positiva

Figura 15 - Classificação quanto à forma da distribuição

Exemplo

A distribuição das idades apresentadas na Tabela 10 é classificada como assimétrica positiva, pois:

$$A_s = \frac{\mu - M_o}{\sigma} = \frac{23,8 - 20,75}{5,16} = 0,59.$$

Medidas de Curtose

A **medida de curtose** representa o grau de achatamento de uma distribuição e é um indicador da forma desta distribuição. É definida como:

$$K = \frac{(Q_3 - Q_1)}{2 \cdot (P_{90} - P_{10})}.$$

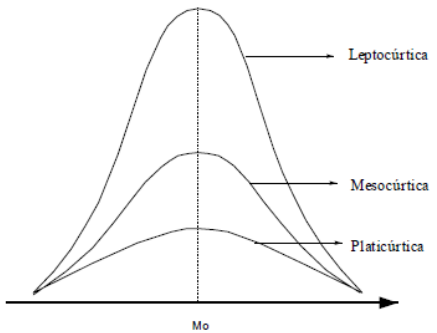
A curtose é mais uma medida com a finalidade de complementar a caracterização da dispersão em uma distribuição.

Esta medida quantifica a concentração ou dispersão dos valores de um conjunto de dados em relação às medidas de tendência central em uma distribuição de freqüências.

Classificação das Distribuições

Uma distribuição é classificada quanto ao grau de achatamento como:

- **Leptocúrtica**: quando a distribuição apresenta uma curva de freqüência bastante fechada, com os dados fortemente concentrados em torno de seu centro, $K < 0,263$.
- **Mesocúrtica**: quando os dados estão razoavelmente concentrados em torno de seu centro, $K = 0,263$.
- **Platicúrtica**: quando a distribuição apresenta uma curva de freqüência mais aberta, com os dados fracamente concentrados em torno de seu centro, $K > 0,263$.



Exemplo

Em relação ao grau de achatamento, a distribuição das idades apresentadas na Tabela 10 é classificada como leptocúrtica, pois:

$$K = \frac{(25,67 - 20)}{2 \cdot (33,2 - 18,8)} = 0,1969.$$

Variável Idade

Centrais	Separatrizes	Dispersão	Assimetria	Curtose
$\bar{x} = 23.8$	$Q_1 = 20$	$AT = 20$	$A_s = 0.89$	$K = 0.1969$
$M_o = 20.75$	$Q_3 = 25.67$	$dq = 5,67$		
$M_d = 22$	$P_{10} = 18,8$	$D_m = 23,82$		
	$P_{90} = 33,2$	$\sigma^2 = 26,63$		
		$\sigma = 5,16$		
		$CV = 21,68\%$		

Resumo Descritivo da Variável Idade – Tabela 10

Exercício 10

Determine e interprete as medidas de assimetria e curtose para a variável altura da Tabela 10.

Box Plot

O gráfico **Box Plot** (ou desenho esquemático) é uma análise gráfica que utiliza cinco medidas estatísticas:

- valor mínimo,
- valor máximo,
- mediana,
- primeiro quartil,
- terceiro quartil.

Box Plot

Este conjunto de medidas oferece a idéia da **posição**, **dispersão**, **assimetria**, **caudas** e **dados discrepantes**.

A posição central é dada pela mediana e a dispersão pelo desvio interquartílico, $dq = Q_3 - Q_2$.

As posições relativas de Q_1, Q_2 e Q_3 dão uma noção da assimetria da distribuição.

Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores atípicos.

Outlier

Um **outlier**, ou valor atípico, é um valor que se localiza distante de quase todos os outros pontos da distribuição.

A distância a partir da qual consideramos um valor como atípico é aquela que supera $1,5 \cdot dq$.

De maneira geral, são considerados outliers todos os valores inferiores a $L_i = Q_1 - 1,5 \cdot dq$ ou os superiores a $L_s = Q_3 + 1,5 \cdot dq$.

Exemplo

A construção do gráfico Box Plot pode ser exemplificada considerando a variável idade da Tabela 10.

Sua elaboração segue os seguintes passos:

1. Ordenar os dados em seqüência crescente.

18	18	19	20	20	20	20	20	20	21	21
22	23	24	25	25	25	26	29	30	35	37

2. Determinar as cinco medidas: mediana, primeiro quartil, terceiro quartil, limite inferior, limite superior.

Exemplo

Mediana

$$i = 0,5 \cdot (22 + 1) = 11,50.$$

$$\begin{aligned} M_d &= x_{11,50} \\ &= x_{11} + 0,50 \cdot (x_{12} - x_{11}) \\ &= 21 + 0,50 \cdot (22 - 21) \\ &= 21,50 \end{aligned}$$

Exemplo

Primeiro quartil

$$i = 0,25 \cdot (22 + 1) = 5,75.$$

$$\begin{aligned} Q_1 &= x_{5,75} \\ &= x_5 + 0,75 \cdot (x_6 - x_5) \\ &= 20 + 0,75 \cdot (20 - 20) \\ &= 20 \end{aligned}$$

Exemplo

Terceiro quartil

$$i = 0,75 \cdot (22 + 1) = 17,25.$$

$$\begin{aligned} Q_3 &= x_{17,25} \\ &= x_{17} + 0,25 \cdot (x_{18} - x_{17}) \\ &= 25 + 0,25 \cdot (26 - 25) \\ &= 25,75 \end{aligned}$$

Exemplo

Desvio interquartílico:

$$dq = Q_3 - Q_1 = 25,75 - 20,00 = 5,75$$

Limite inferior:

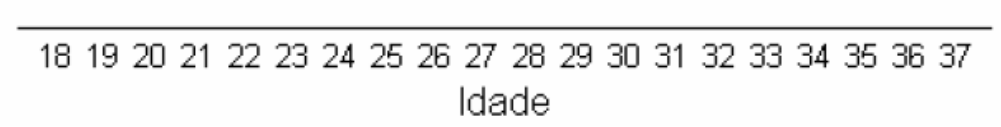
$$\begin{aligned} L_i &= Q_1 - 1,5 \cdot dq \\ &= 20 - 1,5 \cdot 5,75 = 11,375 \end{aligned}$$

Limite superior:

$$\begin{aligned} L_s &= Q_3 + 1,5 \cdot dq \\ &= 25,75 + 1,5 \cdot 5,75 = 34,375 \end{aligned}$$

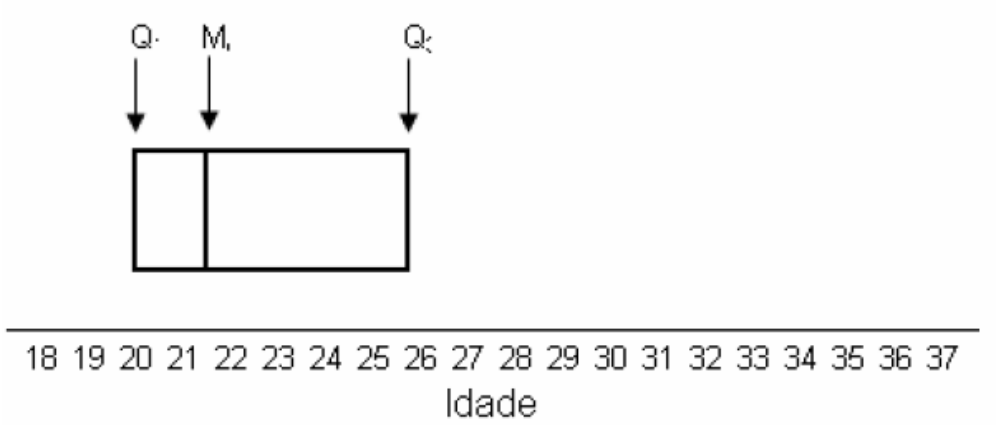
Exemplo

Construir uma escala com valores que incluam os valores máximo e mínimo dos dados.



Exemplo

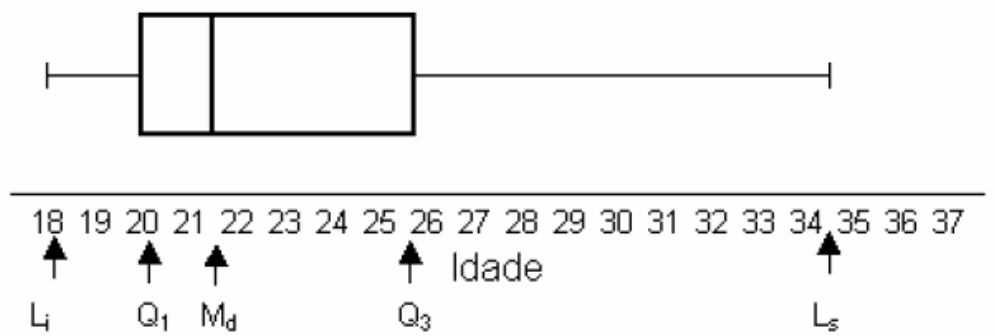
Construir uma caixa (retangular), estendendo-se de Q_1 a Q_3, e traçar uma linha na caixa no valor da mediana.



Exemplo

Traçar uma linha paralela à reta, com uma das extremidades alinhada ao limite inferior L_i e a outra no centro do lado do retângulo correspondente ao primeiro quartil.

Traçar uma outra linha paralela à reta, com uma extremidade no centro do lado do retângulo correspondente ao terceiro quartil e a outra alinhada com o limite máximo L_s.



Exemplo

Identificar os pontos discrepantes.

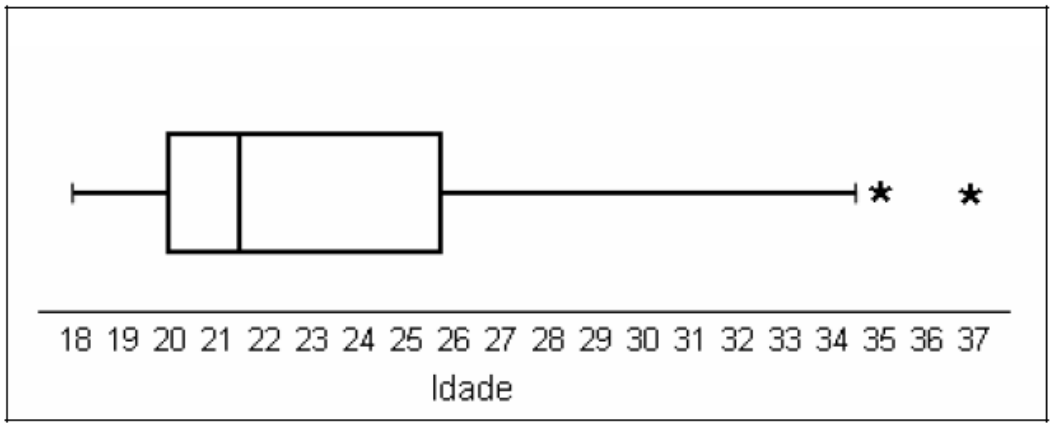


Figura 17 - Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá.

Exemplo

No conjunto de dados, não existe aluno com idade inferior a 11,375, ou seja, não há aluno com idade considerada discrepante inferiormente.

Entretanto, existem dois indivíduos cujas idades são superiores a 34,375, pontos estes considerados discrepantes neste conjunto de dados: as idades 35 e 37.

Estes pontos são identificados no Box Plot por meio de um asterisco na direção das linhas traçadas.

Exemplo

Notamos que, no intervalo interquartílico (dentro do retângulo), existem 50% dos dados, dos quais 25% estão entre a linha da mediana e a linha do primeiro quartil e os outros 25% estão entre a linha da mediana e a linha do terceiro quartil.

Cada linha da cauda mais os valores discrepantes contêm os 25% restantes da distribuição.

A distribuição das idades dos alunos apresenta assimetria positiva, ou seja, dispersam-se para os valores maiores.

Gráficos Box Plot

O gráfico Box Plot pode ser utilizado para fazer comparações entre várias distribuições.

Essa comparação é feita através de vários desenhos esquemáticos numa mesma figura.

Na Figura 18 é apresentado o gráfico para a variável idade classificada segundo o sexo do aluno.

Gráficos Box Plot

Notamos que, para o sexo feminino, não há valores discrepantes e a distribuição apresenta assimetria positiva, com idade mediana inferior à do sexo masculino.

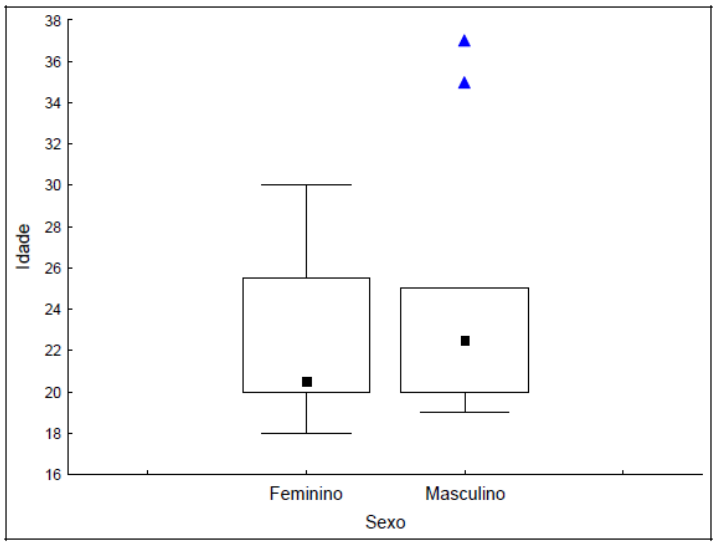


Figura 18 – Box plot da idade segundo o sexo dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá.

Exercício 11

Considere as variáveis peso, número de reprovações na disciplina Inferência Estatística e número de irmãos apresentados na Tabela 01. Determine e interprete os resultados, utilizando os dados em rol e em distribuição de frequências.

1. Média, mediana e moda.
2. Quartil 1, quartil 3, decil 4 e percentil 95.
3. Desvio médio, variância, desvio padrão e coeficiente de variação.
4. Medidas de assimetria e curtose.
5. Construir o box plot para cada uma das variáveis.