Universidade Estadual de Santa Cruz

Ciência da Computação – Probabilidade e Estatística 2019.2

Alunos: Ariel Narciso e Breno Moreira

Adaptação: Fabiano Santos
Atualizado: Ana Cristina, Daniel Lago e Maria Gabriella

DISPERSÃO

Índice

- Amplitude total;
- Amplitude Interquartil;
- Amplitude Semi-Interquartil;
- Desvios:
  - Desvio médio;
  - Desvio quadrático médio;
  - Desvio padrão;
  - Desvio padrão relativo;
- Variância;
- Coeficiente de variação;
- Erro padrão da média.

Dispersão

- Refere-se ao efeito que se produz quando vários elementos se separam de sua origem ou do seu núcleo e se expandem no espaço ou no tempo;
- Dispersão da população;
- Sinônimo: Heterogeneidade;
- Antônimo: Concentração / homogeneidade;
- Na área da matemática este termo também é utilizado e através dele se determina o distanciamento de quaisquer dados sobre a média destes mesmos dados;

Medidas de Dispersão

- O objetivo das medidas de dispersão é medir quão próximos uns dos outros estão os valores de um grupo (e algumas mensuram a dispersão dos dados em torno de uma medida de posição).

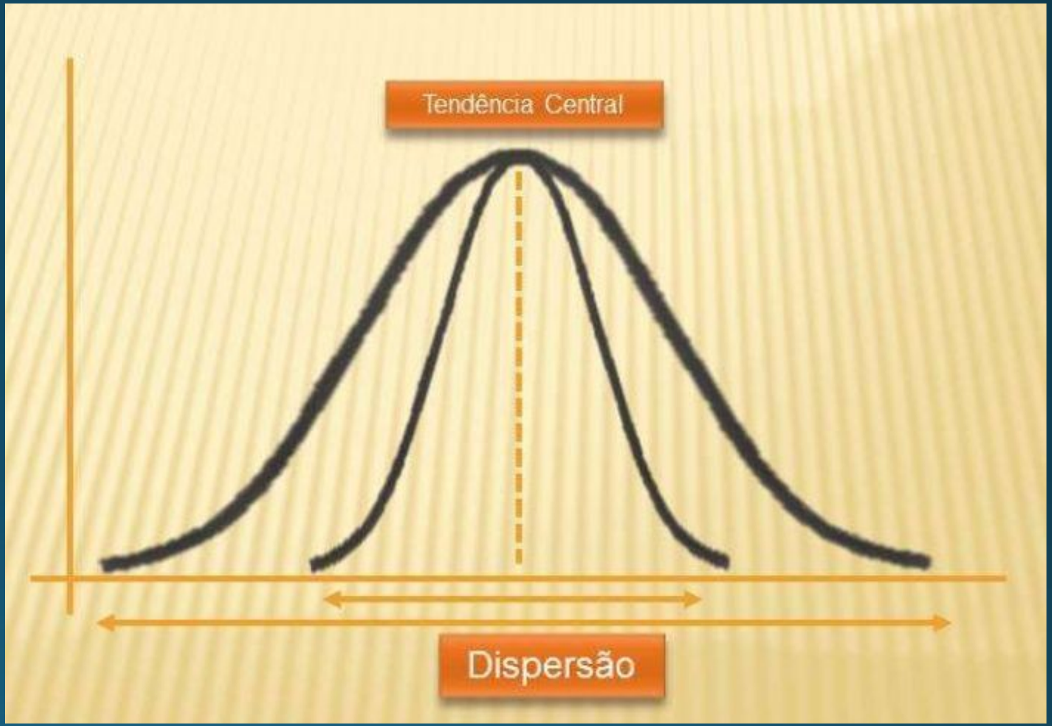


Figura 1: Dispersão;
Fonte: http://oficinadamente.com/medidas-de-dispersao/

Medidas de Dispersão

- É muito incomum ter uma situação em que o salário médio mensal é 1000 euros e todos ganham 1000 euros, ou ter o mesmo salário médio mas em que metade das pessoas ganha o (zero) e outra metade ganha 2000 euros;
- Para caracterizar um conjunto de dados é importante não só a média, mas também a dispersão dos valores em torno da média (os valores estão todos próximos da média ou, pelo contrário, há valores muito afastados da média?).

Medidas de Dispersão

- Uma medida de dispersão estatística é um número real não-negativo que é zero se todos os dados são os mesmos e aumentam conforme os dados vão ficando mais diversos;

- Permitem distinguir séries de dados em relação à homogeneidade:

 Séries homogêneas: menor valor da medida;

 Séries heterogêneas: maior valor da medida.

Medidas de Dispersão

- Exemplo:

 Considere o exemplo de duas linha de produção de uma peça.

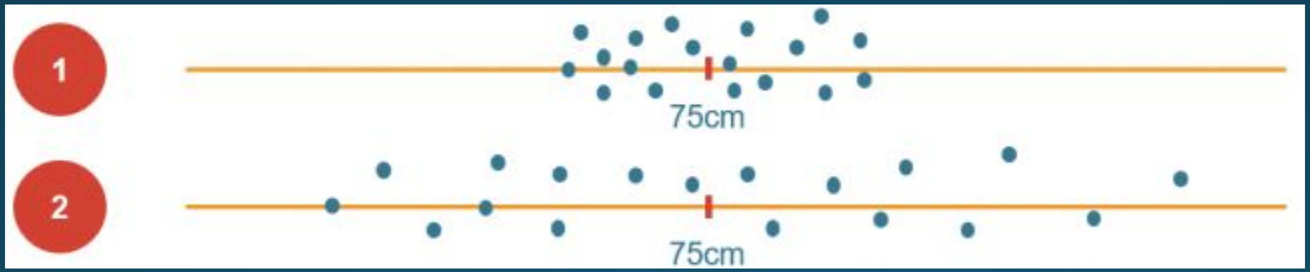


Figura 2: Duas linhas de produção;
http://www.portalaction.com.br/estatistica-basica/22-medidas-de-dispersao

 As peças produzidas pela primeira linha de produção são mais precisas que a segunda.

Amplitude Total

- A amplitude total dos dados de uma amostra é a diferença entre o maior e o menor valor do conjunto de dados;
- Notação: "AT" para o parâmetro e "at" para a estimativa.

$$AT \ ou \ at = y_{max} - y_{min}$$

- Exemplo básico:
- Para o conjunto {2, 3, 4, 6, 6, 7, 7, 9, 9, 10, 12}, a amplitude total é:

AT = 12 - 2= 10

Amplitude Total

Exemplo 1:

- O setor de controle de qualidade de uma empresa seleciona ao acaso peças de um lote. Quando a amplitude das medidas dos diâmetros das peças ultrapassa 0,8 cm o lote é rejeitado.
- Considerando que em um lote foram encontrados os seguintes valores:

2,1 cm; 2,0 cm; 2,2 cm; 2,9 cm; 2,4 cm,

- esse lote foi aprovado ou rejeitado?
- Solução:
- Para calcular a amplitude, basta identificar o menor e o maior valor, que neste caso são 2,0 cm e 2,9 cm. Calculando a amplitude, temos:

$$A = 2{,}9 - 2 = 0{,}9 \text{ cm}$$

- Nesta situação o lote foi rejeitado, pois a amplitude ultrapassou o valor limite.

Amplitude Total

- Exemplo 2 - Histograma:

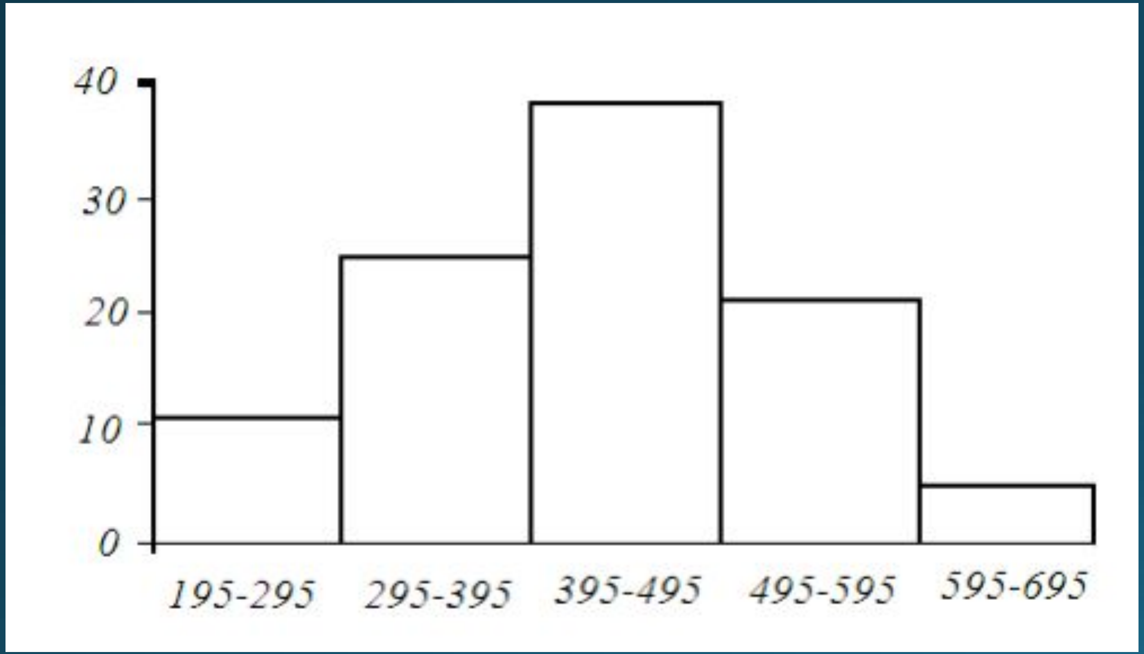


Figura 3: Histograma de amplitude total
http://sisne.org/Disciplinas/Grad/ProbEstat1/aula%206.pd
f

- A amplitude total dos dados é AT = 645 - 245 = 400.
- Note que esta amplitude foi calculada como a diferença entre os pontos médios da última e da primeira classe.

Amplitude Total

- Apesar de sua simplicidade, a amplitude total não dá ideia de como os dados estão agrupados entre os extremos.

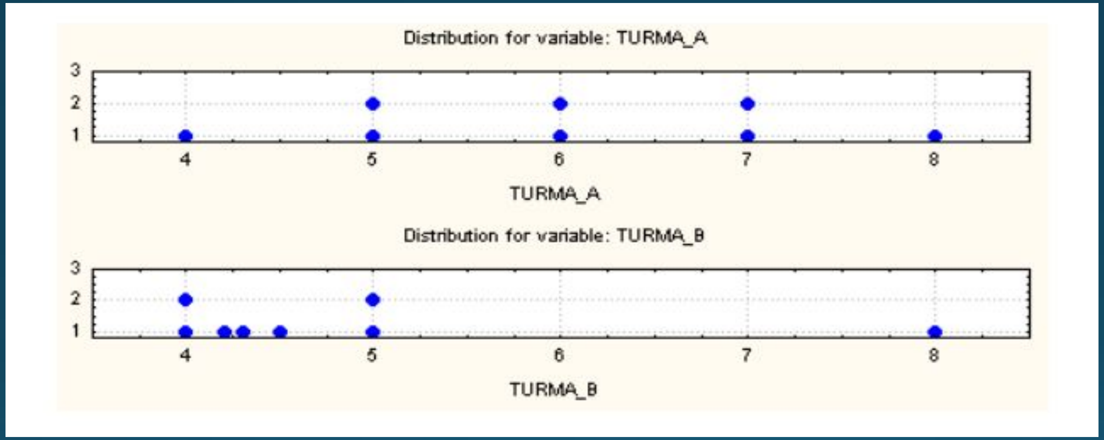


Figura 4: Distribuição de turmas;
http://www.inf.ufsc.br/~marcelo.menezes.reis/AED04.pdf

- No caso acima ambos os grupos têm o mesmo intervalo (4,8), mas no primeiro grupo os dados estão bem dispersos, enquanto no segundo estão próximos do valor mínimo.

Amplitude Total

- Vantagem:

✔ Cálculo rápido e fácil.

- Desvantagens:

✔ Baseada apenas nos valores extremos;

✔ Sensível aos valores anômalos;

Figura 5: Reprodução da Estatística
https://www.estudopratico.com.br/estatistica/

Amplitude Interquartil

- A amplitude interquartil, também chamada de desvio de quartílico, é uma medida de dispersão dos 50% centrais de um conjunto de dados;
- É a distância entre o primeiro quartil e o terceiro quartil, ou seja, Q 3 – Q 1.
- Podemos concluir que 50% dos elementos do meio da amostra, estão contidos num intervalo com aquela amplitude.
- Esta medida é não negativa e será tanto maior quanto maior for a variabilidade nos dados.

Amplitude Interquartil

- Exemplo:

- $Q_1 = (21+31)/2 = 26$
- $Q_3 = (119+119)/2 = 119$
- A.I. $= (Q_3 - Q_1) = 119 - 26 = 93$

i	x[i]	
1	7	
2	7	
3	21	Q_1
4	31	
5	47	
6	75	
7	87	Q_2
8	115	
9	116	
10	119	Q_3
11	119	
12	155	
13	177	

Figura 6: Tabela de dados Spiegel, Murray R. (2006). Estatística. São Paulo: Pearson. p. 115. 643 páginas

Amplitude Semi-Interquartil

- A amplitude semi-interquartil, também chamado de desvio quartil, se define como a metade da distância entre o primeiro e terceiro quartis, ou sendo:
- É a distância entre o primeiro quartil e o terceiro quartil, dividido por dois, ou seja, $(Q3 – Q1)/2$.
- O desvio quartil mede a amplitude ocupada pela metade central dos dados, ou seja, mede a concentração dos valores em torno da mediana. Tal concentração é inversamente proporcional ao tamanho do desvio quartil.

Amplitude Semi-Interquartil

- Exemplo:

i	x[i]
1	7
2	7
3	21
4	31
5	47
6	75
7	87
8	115
9	116
10	119
11	119
12	155
13	177

Q_1 Q_2 Q_3

Figura 6: Tabela de dados
Spiegel, Murray R. (2006). Estatística. São Paulo: Pearson. p. 115. 643 páginas

- Q 1 = (21+31)/2 = 26
- Q 3 = (119+119)/2 = 119
- D.Q = (Q3 - Q1)/2 = (119 – 26)/2 = (93)/2 = 46,5

Desvio

- Diferença entre o valor e a média.
- Notação: “D” para o parâmetro e “d” para a estimativa.
- Seja **Y** = (y_1, y_2,..., y_n) e $\bar{\mathbf{Y}}$ a média aritmética de Y, o desvio **i** é dado por:

$$D_i \; ou \; d_i = (y_i - \bar{Y})$$

Desvio

$Y_1 = \{3, 4, 5, 6, 7\}$
$\bar{Y}_1 = 5$
$d_4 = 6 - 5 = 1$

$Y_2 = \{2.5, 3.7, 6.3, 8.8, 9.5\}$
$\bar{Y}_2 = 6.16$
$d_2 = 3.7 – 6.16 = -2.46$

- Observação: $\sum Di = \sum di = 0$ (zero)

$Y1 = \{3, 4, 5, 6, 7\}$
$\sum Di = \sum di = d_1 + d_2 + d_3 + d_4 + d_5$
$\sum Di = \sum di = (-2) + (-1) + (0) + (1) + (2) = -3 + 3 = 0$

Desvio Médio

- Média dos desvios absolutos em relação à média aritmética.
- Notação: (DM) para o parâmetro e (dm) para a estimativa.
- Quantifica a dispersão dos dados.
- Permite distinguir séries de dados em relação à homogeneidade:

✔ Séries homogêneas: menor valor de desvio médio;

✔ Séries heterogêneas: maior valor de desvio médio.

$$DM = \frac{\sum |D_i|}{N} \qquad dm = \frac{\sum |d_i|}{n}$$

Desvio Médio

- Exemplo 1: Qual o desvio médio da série (3,4,5,6,7) ?

- Passo 1: Média = (3+4+5+6+7)/5 = 25/5 = 5
- Passo 2: Cálculo dos desvios em módulo:

$$D_i = (y_i - \bar{Y})$$

$$|D_1| = |3 - 5| = |-2| = 2$$
$$|D_2| = |4 - 5| = |-1| = 1$$
$$|D_3| = |5 - 5| = |0| = 0$$
$$|D_4| = |6 - 5| = |1| = 1$$
$$|D_5| = |7 - 5| = |2| = 2$$

Desvio Médio

- Exemplo 1: Qual o desvio médio da série (3,4,5,6,7) ?

- Passo 3: Soma dos desvios em módulo:
  - ✔ $\sum |Di| = 2 + 1 + 0 + 1 + 2 = 6$

- Passo 4: Dividir o resultado do passo anterior, ou seja, do somatório dos módulos, por N, quantidade de elementos da série:
  - ✔ $DM = \sum |Di| / N = 6 / 5 = 1{,}2$

Desvio Médio

Exemplo 2:

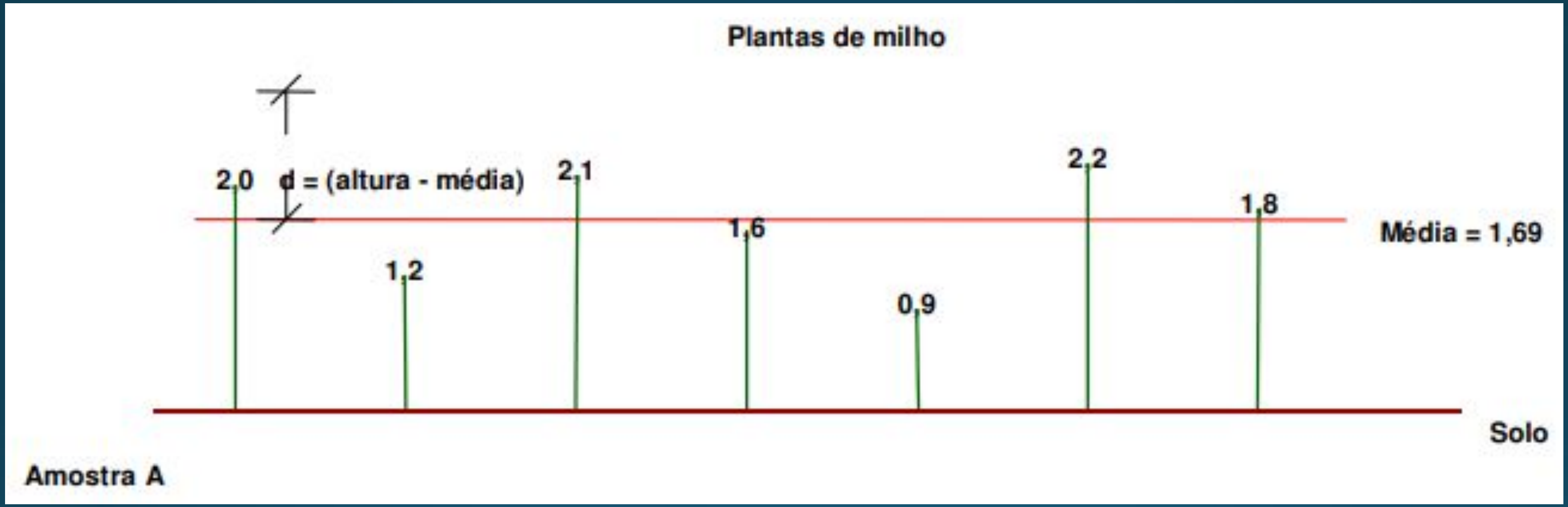


Figura 8 – Ilustração de uma amostra de plantas de milho
http://nbcgib.uesc.br/lec/download/faria/cet756/apresentacoes/medidas_estatisticas_tcpd.pdf

$$dm = \frac{|2{,}0 - 1{,}69| + \cdots + |1{,}8 - 1{,}69|}{7} = 0{,}39m$$

Desvio Quadrático Médio

- Média dos quadrados dos desvios em relação à média aritmética.
- Notação: (DQM) para o parâmetro e (dqm) para a estimativa.

$$DQM = \frac{\sum (D_i)^2}{N} \qquad dqm = \frac{\sum (d_i)^2}{n}$$

Desvio Quadrático Médio

- Exemplo 1: Qual o desvio quadrático médio da série (3,4,5,6,7) ?

- Passo 1: Média = (3+4+5+6+7)/5 = 5
- Passo 2: Cálculo dos desvios ao quadrado:

$$D1^2 = (3 - 5)^2 = (-2)^2 = 4$$
$$D2^2 = (4 - 5)^2 = (-1)^2 = 1$$
$$D3^2 = (5 - 5)^2 = (0)^2 = 0$$
$$D4^2 = (6 - 5)^2 = (1)^2 = 1$$
$$D5^2 = (7 - 5)^2 = (2)^2 = 4$$

Desvio Quadrático Médio

- Exemplo 1: Qual o desvio quadrático médio da série (3,4,5,6,7) ?

- Passo 3: Soma dos quadrados dos desvios:

  ✔ $\sum (Di)^2 = 4 + 1 + 0 + 1 + 4 = 10$

- Passo 4: Dividir o resultado do passo anterior, ou seja, do somatório dos módulos, por N, quantidade de elementos da série:

  ✔ $DQM = \sum (Di)^2 / N = 10 / 5 = 2$

Desvio Quadrático Médio

Exemplo 2:

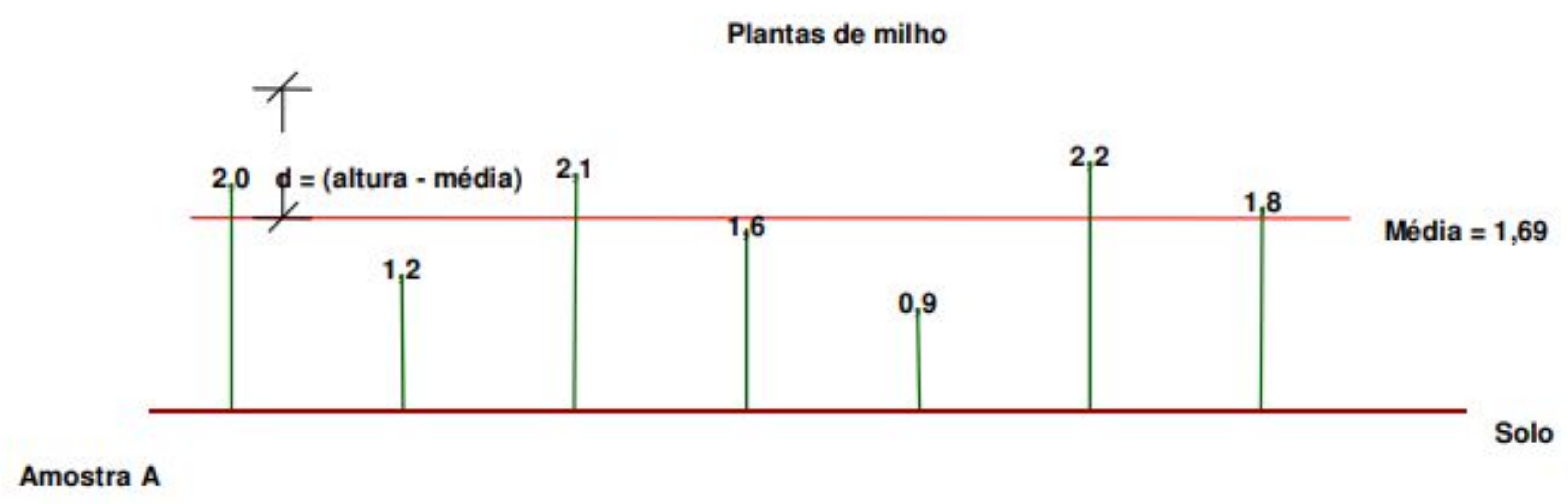


Figura 9 – Ilustração de uma amostra de plantas de milho
http://nbcgib.uesc.br/lec/download/faria/cet756/apresentacoes/medidas_estatisticas_tcpd.pdf

$$dqm = \frac{(2{,}0 - 1{,}69)^2 + \cdots + (1{,}8 - 1{,}69)^2}{7} = 0{,}19m^2$$

Variância

- Razão entre o somatório dos quadrados dos desvios pelo número de elementos da série, N para população e n-1 para amostra.
- Notação: (σ^2) para o parâmetro e (s^2) para a estimativa (sigma ao quadrado).

Variância

- Dados não agrupados

Variância Populacional

$$\sigma^2 = \frac{\sum {D_i}^2}{N}$$

- $\bar{Y}$ é conhecido
- Caso raro

Variância Amostral

$$s^2 = \frac{\sum {d_i}^2}{n-1}$$

- $\bar{Y}$ não é conhecido
- Caso comum

Variância Populacional

Média aritmética da população = $\bar{Y}$ = (0,20+0,21+0,22+0,20+0,19)/5 = 0,204

Yi	Desvio (Yi - $\bar{Y}$)	Desvio ao quadrado $(Yi - \bar{Y})^2$
0,20	0,20 – 0,204 = -0,004	0,000016
0,21	0,21 – 0,204 = 0,006	0,000036
0,22	0,22 – 0,204 = 0,016	0,000256
0,20	0,20 – 0,204 = -0,004	0,000016
0,19	0,19 – 0,204 = -0,014	0,000196
Total		0,00052

σ^2 = 0,00052/5 = 0,000104 (variância da população que é a dispersão da aplicação com relação a sua média simples)

Variância Amostral

Refere-se a parcela de dados retirados de um grande universo da qual desejamos obter informações e/ou conhecimento.

Exemplo: Suponhamos uma amostra aleatória de 5 elementos que são = (20, 18, 15, 0, 25). Imaginemos agora que seja a amostra de uma população, então temos:

Média amostral = $\bar{Y}$ = (20 + 18 + 15 + 0 + 25) / 5 = 15,6

yi	Desvio (yi - $\bar{Y}$)	Desvio ao quadrado (yi - $\bar{Y}$)^2
20	20 – 15,6 = 4,4	19,36
18	18 – 15,6 = 2,4	5,76
15	15 – 15,6 = - 0,6	0,36
0	0 – 15,6 = - 15,6	243,36
25	25 – 15,6 = 9,4	88,36
Total		357,20

s^2 = 357,20/ (5-1) = 357,20/4 = 89,3

Variância

- Ilustração da tendenciosidade da estimativa de s^2 se o somatório dos desvios em relação à média for dividido por n, ao invés de n-1.

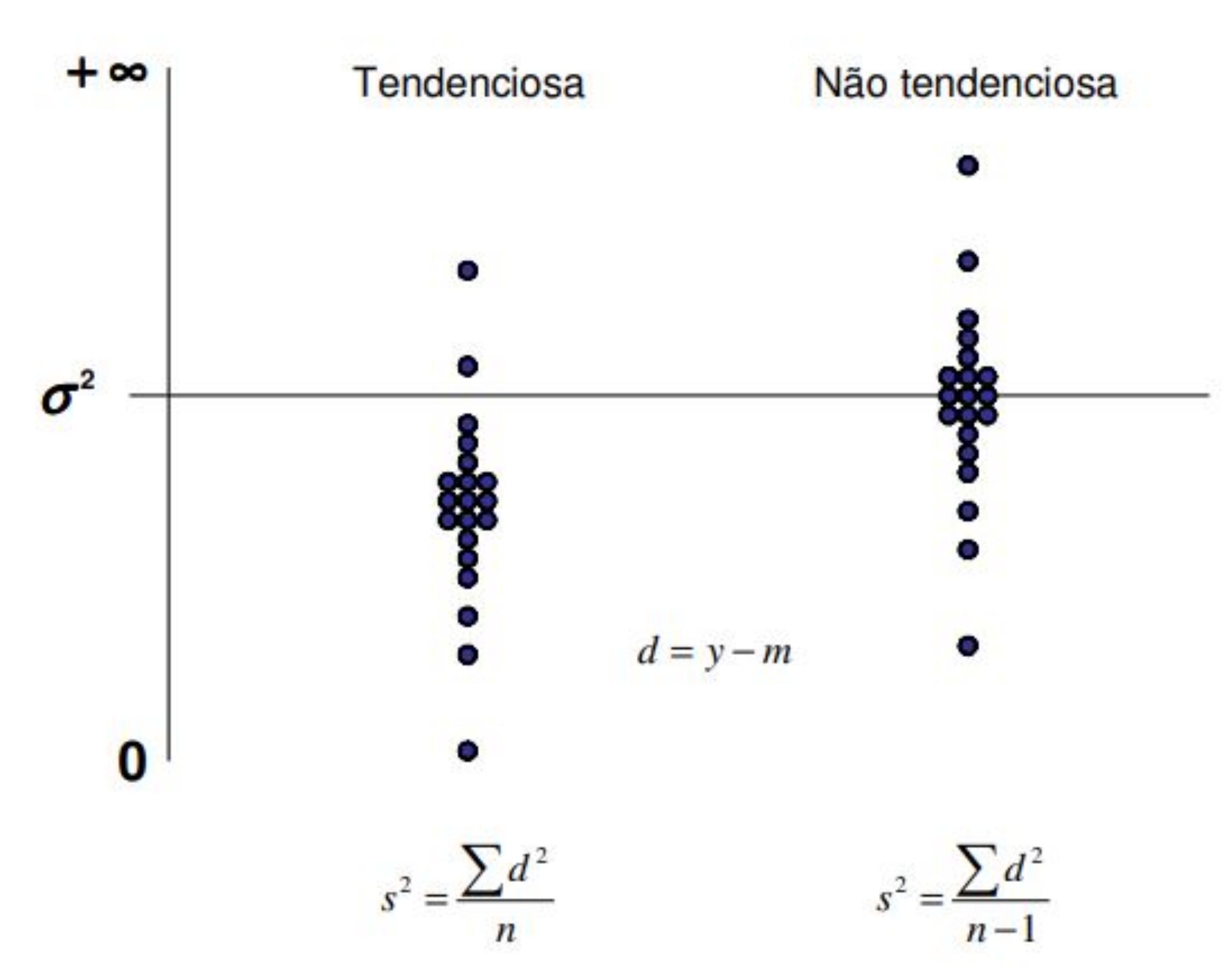


Figura 10: Estimativa das variâncias

Variância

- Quanto maior for a variância, mais distantes da média estarão os valores;
- Quanto menor for a variância, mais próximos os valores estarão da média;
- Pode não ser suficiente, pois é muito influenciada por valores que estão muito distantes da média;
- Como é calculada “ao quadrado”, causa uma certa dificuldade na sua interpretação.

Variância

- Dados agrupados

Variância Populacional

$$\sigma^2 = \frac{\sum {D_i}^2 \, x \, f_i}{\sum f_i}$$

- Ȳ é conhecido
- Caso raro

Variância Amostral

$$s^2 = \frac{\sum {d_i}^2 \, x \, f_i}{\sum f_i - 1}$$

- Ȳ não é conhecido
- Caso comum

Variância

Exemplo

Dados agrupados sem intervalo de classe

Altura (m)	fi	yi
1,8	1	1,8
1,7	4	6,8
1,5	2	3
	7	11,6

Tabela 1 – Alturas de plantas de milho com suas respectivas frequências

Média:

$$\frac{\sum y_i}{\sum f_i} = 1{,}6571$$

$$s_B^2 = \frac{(1{,}8 - 1{,}66)^2 . 1 + (1{,}7 - 1{,}66)^2 . 4 + (1{,}5 - 166)^2 . 2}{7 - 1} = 0{,}01m^2$$

Variância

Exemplo

Dados agrupados com intervalo de classe

Altura (m)	f_i	y_i	$f_i y_i$
155 |-- 165	3	160	480
165 |-- 175	8	170	1360
175 |-- 185	5	180	900
185 |-- 195	9	190	1710
	25		4450

Média: $\frac{\sum y_i f_i}{\sum f_i} = 178$

y_i = ponto médio

Tabela 2 – Alturas de plantas de milho agrupadas em intervalos de classes

$$s^2 = \frac{(160 - 178)^2 . 3 + \cdots + (190 - 178)^2 . 9}{25 - 1} = 116{,}66 m^2$$

Desvio Padrão

- Raiz quadrada da variância
- Quantifica a dispersão dos dados.
- Junto a variância, o desvio padrão é uma das medidas mais usadas para quantificar a dispersão dos dados em torno da média.
- Notação: (σ) para o parâmetro e (s) para a estimativa (sigma).

$$\sigma = \sqrt{\sigma^2} \qquad s = \sqrt{s^2}$$

Desvio Padrão

- É sempre positivo;
- Só é nulo se não houver variabilidade;
- Tem a mesma unidade dos dados originais;
- Um desvio padrão grande significa que os valores estão mais dispersos em relação a média;
- Um desvio padrão pequeno indica que os valores estão concentrados próximos da média.

Desvio Padrão

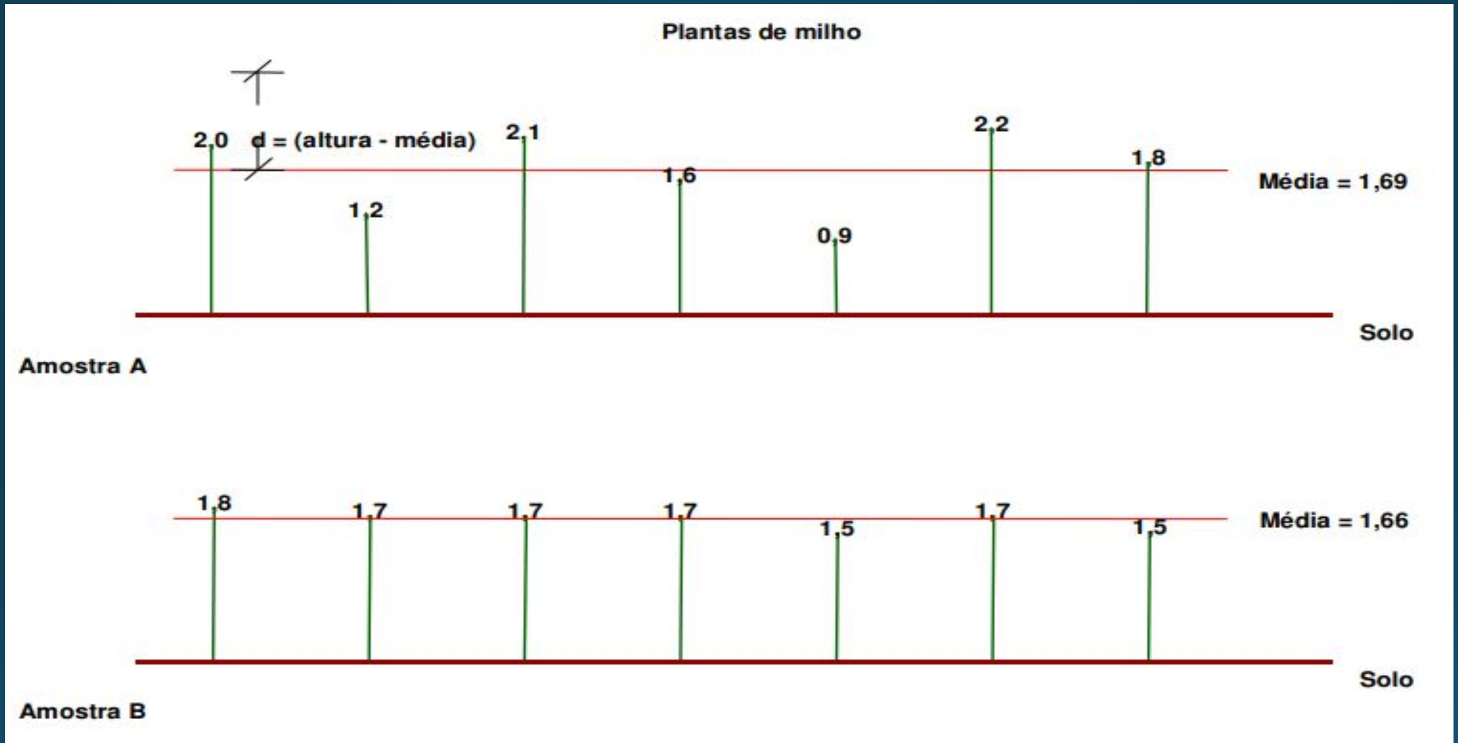


Figura 11 – Ilustração de amostras de plantas de milho
http://nbcgib.uesc.br/lec/download/faria/cet756/apresentacoes/medidas_estatisticas_tcpd.pdf

$$S_A^2 = \frac{(2{,}0 - 1{,}69)^2 + \cdots + (1{,}8 - 1{,}69)^2}{7 - 1} = 0{,}23m^2$$

Desvio Padrão

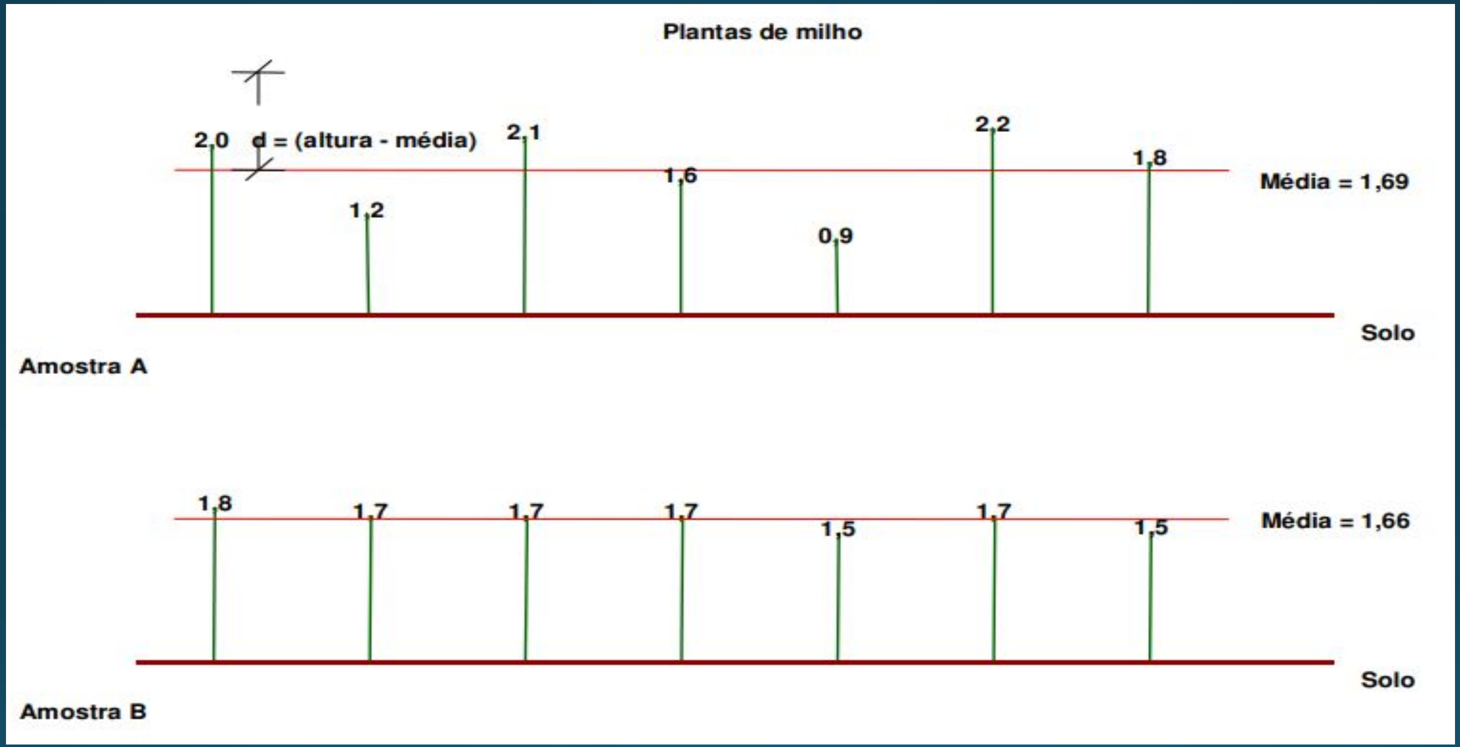
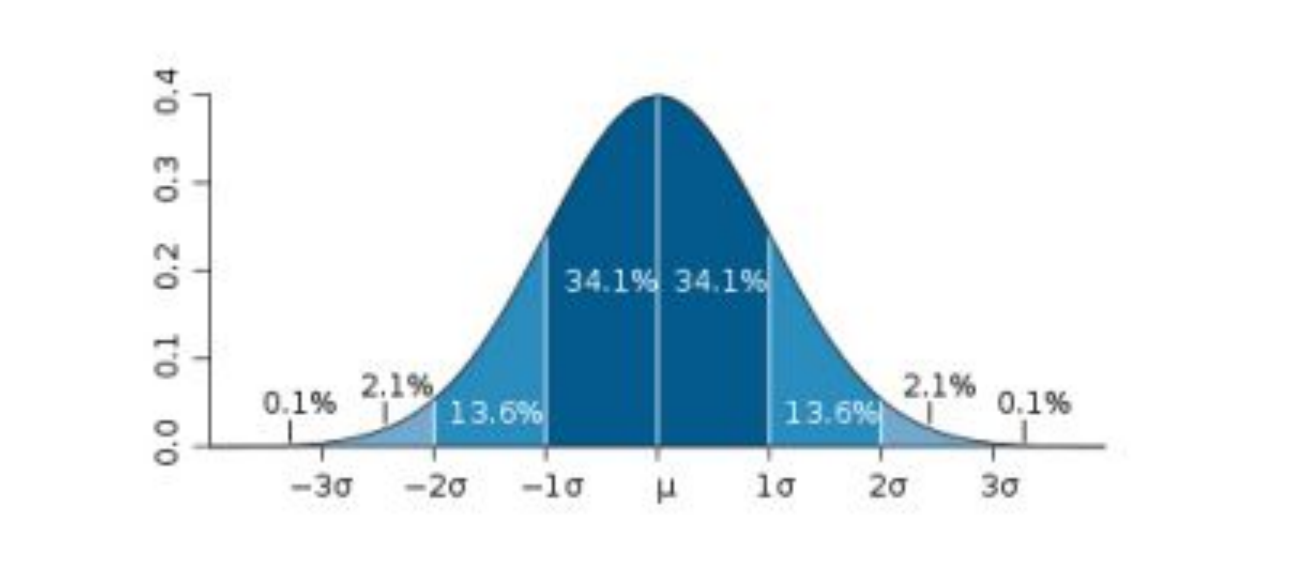


Figura 12 – Ilustração de amostras de plantas de milho
http://nbcgib.uesc.br/lec/download/faria/cet756/apresentacoes/medidas_estatisticas_tcpd.pdf

$$s_A = \sqrt{s_A^2} = \sqrt{0{,}23m^2} = 0{,}48m$$

Desvio Padrão



68,26% das ocorrências se concentrarão na área do gráfico demarcada por um desvio padrão à direita e um desvio padrão à esquerda da linha média;

95,44% das ocorrências estão a dois desvios padrão, para a direita e a esquerda da média e, finalmente;

99,72% das ocorrências ocorrem a três desvios padrão ao redor da média aritmética.

Desvio Padrão

TEOREMA DE CHEBYCHEV:

$$1 - \frac{1}{K^2}$$

- K = 2: EM QUALQUER CONJUNTO DE DADOS, **PELO MENOS** 1 - ¼ = ¾ OU 75% DOS DADOS ESTÃO DENTRO DE DOIS DESVIOS PADRÕES

- K = 3: EM QUALQUER CONJUNTO DE DADOS, **PELO MENOS** 1 - 1/9 = 8/9 OU 88,9% DOS DADOS ESTÃO DENTRO DE TRÊS DESVIOS PADRÕES

Desvio Padrão Relativo

- Razão entre o desvio padrão e a média aritmética.
- As unidades de medida das variáveis não exercem influência na magnitude das medidas de dispersão relativas.
- Notação: “DPR” para os parâmetros e “dpr” para as estimativas.

$$DPR = \frac{\sigma}{\overline{\overline{Y}}} \qquad dpr = \frac{s}{\overline{\overline{Y}}}$$

Desvio Padrão Relativo

- Exemplo: Altura e Peso de Alunos

	Média	Desvio Padrão	DPR
Altura	1,143m	0,063m	0,055
Peso	50Kg	6Kg	0,12

- Ao determinar o DPR, é possível saber de que forma o desvio padrão está para a média.

Coeficiente de Variação

- O CV é uma medida de dispersão relativa que expressa a razão entre o desvio padrão e a média (Desvio Padrão Relativo) afim de obter sua porcentagem;
- Notação: "CV" para os parâmetros e "cv" para as estimativas.
- De uma forma geral, se o CV:
  > Menor ou igual a 15% → baixa dispersão, dados homogêneos
  > Entre 15 e 30% → média dispersão
  > Maior que 30% → alta dispersão, dados heterogêneos

$$CV = \frac{\sigma}{\overline{\overline{Y}}} \times 100 \qquad cv = \frac{s}{\overline{\overline{Y}}} \times 100$$

Coeficiente de Variação

- Amostras com mesma unidade (mesma variável medida) e médias diferentes.
- Exemplo: Altura de plantas

Amostras	Média	Desvio Padrão	cv
A	58,3cm	6,9cm	11,84%
B	79,4cm	7,6cm	9,57%

$$cv(a) = \frac{s(a)}{\bar{x}(a)} * 100 = \frac{6,9}{58,3} * 100 = 11,84\%$$

$$cv(b) = \frac{s(b)}{\bar{x}(b)} * 100 = \frac{7,6}{79,4} * 100 = 9,57\%$$

A amostra A apresenta maior variabilidade dos dados

Coeficiente de Variação

- Amostras com unidades diferentes (medição de variáveis diferentes).
- Exemplo: Altura de plantas (em cm), amostra A, e produção de frutos (em gramas), amostra B.

Amostras	Média	Desvio Padrão	cv
A	52,1cm	5,8cm	11,13%
B	12,1g	1,9g	15,70%

$$cv(a) = \frac{s(a)}{\bar{x}(a)} * 100 = \frac{5,8}{52,1} * 100 = 11,13\%$$

$$cv(b) = \frac{s(b)}{\bar{x}(b)} * 100 = \frac{1,9}{12,1} * 100 = 15,70\%$$

A característica Produção de frutos, amostra B, apresenta maior variabilidade que a Altura das plantas, amostra A.

Erro padrão da média

- O erro padrão da média é uma medida de dispersão que informa a precisão com que uma média populacional é estimada pela média amostral;

$$S(\bar{x}) = \frac{s}{\sqrt{n}}$$

- Em que: $S(\bar{x})$ é o erro padrão da média, s é o desvio padrão da amostra, n é o tamanho da amostra.

Erro padrão da média

- Exemplo:

População de 150.000 plantas (N)
Altura média das plantas = μ

Amostra de 60 plantas (n):
Altura média das plantas () = 190cm
Desvio padrão (s) = 15cm

- Qual o valor de μ (média populacional)? Não sabemos.
- Estimativa é dada por $\bar{x}$ = 190cm. Qual a precisão?

Erro padrão da média

- Exemplo:

População de 150.000 plantas (N)
Altura média das plantas = μ

Amostra de 60 plantas (n):
Altura média das plantas () = 190cm
Desvio padrão (s) = 15cm

- $S(\bar{x}) = \frac{15}{\sqrt{60}} = 1{,}9\text{cm}$

A altura média das plantas foi estimada com um erro padrão de 1,9 cm, ou seja, com uma precisão de 1,9 cm

Referências

- MENEZES, Marcelo. **Estatística.** Disponível em: <http://www.inf.ufsc.br/~marcelo.menezes.reis/AED04.pdf>. Acesso em: 01 out. 2018.
- JOSÉ NETO,. **Estatística Descritiva:** Medidas de dispersão. Disponível em: <http://estatisticax.blogspot.com/2008/02/medidas-de-disperso.html>. Acesso em: 01 out. 2018.
- SILVIA. **O Diagrama de Dispersão.** Disponível em: <http://leg.ufpr.br/~silvia/CE055/node15.html>. Acesso em: 01 out. 2018.
- ROQUE, Antonio. **Medidas de Dispersão.** Disponível em: <http://sisne.org/Disciplinas/Grad/ProbEstat1/aula%206.pdf>. Acesso em: 01 out. 2018.

Referências

- PEIXOTO, Maurício. **Medidas de Dispersão.** Disponível em: <http://oficinadamente.com/medidas-de-dispersao/>. Acesso em: 01 out. 2018.
- DAVILA, Victor Hugo Lachos. **Estatística Descritiva.** Disponível em: <https://www.ime.unicamp.br/~hlachos/estdescr1.pdf>. Acesso em: 01 out. 2018.
- USP. **Medidas de Dispersão.** Disponível em: <http://www.esalq.usp.br/departamentos/lce/arquivos/aulas/2011/LCE2112/MatÃ©ria%20-%20MEDIDAS%20DE%20DISPERSÃ□O.pdf>. Acesso em: 01 out. 2018.

Referências

- SERMARINI, Renata Alcarde. **Medidas de Dispersão.** 2017. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4094191/mod_resource/content/1/4.%20Medidas%20de%20Dispersão.pdf>. Acesso em: 01 out. 2018.
- UNITINS. **Medidas de Dispersão.** Disponível em: <https://www.unitins.br/BibliotecaMidia/Files/Documento/AVA_633848868383423750estatistica___aula_7.pdf>. Acesso em: 01 out. 2018.
- WIKIPEDIA. **Amplitude interquartil.** Disponível em: <https://pt.wikipedia.org/wiki/Amplitude_interquartil>. Acesso em: 01 out. 2018.