ESTATÍSTICA APLICADA I

VOLUME 1

ANA SCARDINO

ANA SCARDINO

ESTATÍSTICA APLICADA I

1ª edição

São Paulo

Edição do autor

2020

Sumário

Capítulo 1

1. INTRODUÇÃO À ESTATÍSTICA DESCRITIVA.

Foi no século XVIII que Godofredo Achenwall batizou a ciência que estuda como chegar a conclusões do todo partindo da observação de partes desse todo com o nome de Estatística, cuja origem lembra Estado.
Vamos inicialmente proceder à algumas definições:

VARIÁVEIS: Conjunto de resultados possíveis de um fenômeno.

TIPOS DE VARIÁVEIS:

- Qualitativas - seus valores são expressos por atributos.

Podem ser:

- Ordinais – As categorias têm uma ordem natural. Exemplos: faixa de renda salarial, escolaridade.
- Nominais – As categorias não se apresentam com uma ordem. Exemplos: sexo, cor da pele.

- Quantitativas - seus valores são expressos em números.

Exemplos: salário, idade dos alunos etc.
Podem ser:

- Contínuas: podem assumir qualquer valor num intervalo contínuo. Exemplo: peso dos alunos, volume.
- Discretas: podem assumir valores inteiros. Exemplo: número de alunos de uma sala.

Uma pergunta sobre renda pode ser apresentada como dado numérico (valor da renda), dado ordinal (faixa de renda) ou dado nominal (rico ou pobre).

Os dados quantitativos permitem uma avaliação estatística mais rica e podem ser sempre transformados em dados qualitativos e vice-versa.

ATIVIDADE

1.1. Classifique as variáveis:

a) Estado civil
a) Salários
b) Lucro de uma empresa
c) Religião
d) Altura
e) Número de filhos.

Capítulo 2

2. TÉCNICAS DE AMOSTRAGEM

POPULAÇÃO
Conjunto de entes portadores de, pelo menos, uma característica comum. Ex. estudantes.

AMOSTRA
Subconjunto representativo e finito de uma população. Ex. retirada de amostras para controle de qualidade de uma indústria.

AMOSTRAGEM
Existem técnicas de amostragem que garantem o acaso na escolha:
Amostragem casual: pode-se numerar a população de 1 a n e sortear k números dessa sequência ou usar uma tabela de números aleatórios.

Amostragem proporcional estratificada: Quando a população se divide em subpopulações. Quando de estrato em estrato se verifica um comportamento heterogêneo. Ex. Uma classe com 90 alunos, 62 homens e 28 mulheres. Queremos uma amostra de 10% da população.

Determinação de uma amostra proporcional estratificada.

Sexo	População	10%	Amostra
M	62	6,2	6
F	28	2,8	3
Total	90	9,0	9

Amostragem sistemática: seleção dos elementos pode ser feita por um sistema determinado pelo pesquisador. Ex. Numa produção diária de um produto, retirar um a cada 10 itens, para controle de qualidade.

Capítulo 3

3. TRATAMENTO DE DADOS

O exemplo abaixo nos mostra como obter algumas medidas dentro da estatística descritiva. Para isso considerou-se a idade, peso e estatura dos alunos em uma sala de aula.

Coleta dos dados da idade, peso e estatura dos alunos de uma sala de aula.

nº	Idade (anos)	Peso (kg)	Estatura (m)
1	33	64	1,64
2	37	55	1,61
3	40	55	**1,61**
4	33	60	1,61
5	33	48	1,6
6	36	68	1,68
7	38	60	1,5
8	35	63	1,6
9	40	75	1,67
10	36	62	1,6
11	32	67	1,7
12	54	66	1,58
13	31	47	1,61
14	29	52	1,59
15	28	70	1,69
16	46	75	1,69
17	38	95	1,6
18	32	100	1,9
19	40	62	1,6
20	45	63	1,67
21	48	67	1,53
22	36	75	1,76

A tabela assim obtida chama-se tabela primitiva, pois os elementos não foram numericamente organizados.

Partindo dos dados de idade acima, vamos ordená-los em ordem crescente.

Idade (anos)
28
29
31
32
32
33
33
33
35
36
36
36
37
38
38
40
40
40
45
46
48
54

3.1. ROL

A tabela obtida após a ordenação chama-se **ROL**. Podemos saber agora, com mais facilidade qual a idade do aluno mais novo, do mais velho, a amplitude de variação e onde há maior concentração de idades.

Se os valores ordenados forem dispostos em uma coluna e colocarmos, ao lado de cada valor, o número de vezes que aparece repetido verifica-se uma maneira mais fácil de observar os dados.

Idade(anos)	Frequência (f_i)
28	1
29	1
31	1
32	2
33	3
35	1
36	3
37	1
38	2
40	3
45	1
46	1
48	1
54	1

3.2. DISTRIBUIÇÃO DE FREQUÊNCIA

O número de alunos que fica relacionado a uma determinada idade chama-se frequência. A tabela obtida dessa maneira recebe o nome de **DISTRIBUIÇÃO DE FREQUÊNCIA.**

A tabela obtida assim ainda é inconveniente, pois se torna muito grande. Uma maneira melhor de apresentá-la será agrupando os valores em intervalos.

3.2.1. NÚMERO DE CLASSES

A escolha do número de classes deve ficar entre 5 e 15 classes. Menos de 5 classes pode ocultar detalhes importantes dos dados, e mais de 15 torna a apresentação demasiadamente detalhada. Uma regra prática consiste em tirar a raiz quadrada do número de observações e ajustá-la aos limites 5 a 15. Por exemplo, para 40 observações, teríamos $\sqrt{40} = 6{,}32$ que arredondamos para 6 ou 7. Para 900 observações, $\sqrt{900} = 30$, que será ajustado para 15.

Chamando de i = número de classes e
N =número total de dados

$$\boxed{i = \sqrt{N}}$$

Exemplo:

Para n=100, $i = \sqrt{100} = 10$

Para n=1000 teremos i = $\sqrt{1000}$ = 31,6, que será ajustado para 15.

Outra regra utilizada é a regra de Sturges, que nos dá o número de classes em função do número de dados.

$$\boxed{i=1 + 3{,}3\log N}$$

Exemplos:

a) Para n= 1000 → i = 1 + 3,3log1000 = 10,9
Aplicando a regra de Sturges para n = 1000, temos número de classes igual a 11.

b) Para n = 22 → i = 1 + 3,3 log22 = 5,43 = 5 (com a regra prática temos, $i = \sqrt{22} = 4{,}7 = 5$)

c) Para n = 40 → i = 1 + 3,3log40 = 6,29 = 6

3.2.2. INTERVALOS DE CLASSES

Qual a amplitude do intervalo de cada classe?

Chamando A =amplitude total = l_s - l_i
i =número de classes E
h = intervalo de classe, tem-se

$$\boxed{h= A/i}$$

Tem-se para o exemplo do rol anterior:
Número de classes i = 5
Amplitude total A = 54-28 =26
h = 26:5 = 5,2, que é o intervalo de classe.
No caso podemos optar em ter h=6 e i =5

3.2.3. TABELA DE DISTRIBUIÇÃO DE FREQUÊNCIA POR CLASSES

Distribuição de FREQUÊNCIA(D.F.): É o arranjo ordenado dos valores com suas respectivas FREQUÊNCIA

Limites de classes:
Limite Inferior de classe (l_I) menor número da classe
Limite Superior de classe (L_S)- maior número da classe

FREQUÊNCIA Absoluta (Fi): É o número de vezes que o dado aparece no rol.

Pontos médios das classes(xi): é a média aritmética entre o limite superior e o limite inferior: xi =(li+ls)/2

FREQUÊNCIA relativa(fr): Obtém-se dividindo a FREQUÊNCIA absoluta de cada classe por n.
$fr = F_i/n$

FREQUÊNCIA acumulada(fac): é a soma das FREQUÊNCIA absolutas, linha por linha.

FREQUÊNCIA relativa acumulada (facr): é a soma linha a linha das FREQUÊNCIAS relativas. Obtém-se também se dividindo a frequência acumulada pela FREQUÊNCIA total.

Exemplo:

i	*Idade*	X_i	*Fi*	*Fr %*	*fac*	*Facr%*
1	28⊢34	31	8	36	8	36
2	34⊢40	37	7	32	15	68
3	40⊢46	43	4	18	19	86
4	46⊢52	49	2	9	21	96
5	52⊢58	51	1	5	22	100
Total			22	1.0		

ATIVIDADE

3.1. Dado o rol de 50 notas, agrupar os elementos em uma distribuição de frequência:
33-35-35-39-41-41-42-45-47-48-50-52-53-54-55-55-57-59-60-60
61-64-65-65-65-66-66-66-67-68-69-71-73-73-74-74-76-77-77-78-80-81-84-85-85-88-89-91-94-97.

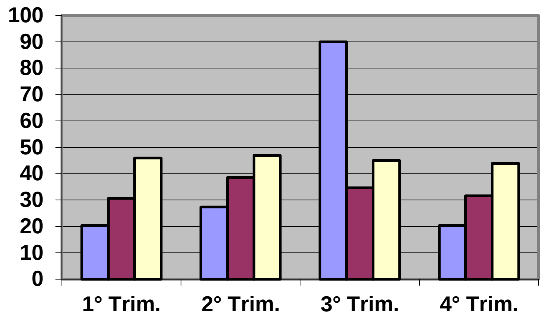
3.2. Responda:

a) Quantas notas estão no intervalo de 53 a 63?
a) Quantos por cento das notas estão acima de 73?
b) Quantas notas estão abaixo de 63?
c) Quantos por cento das notas estão no intervalo de 63 a 83 (excluindo o 83)?

3.2.4. REPRESENTAÇÃO GRÁFICA DE UMA DISTRIBUIÇÃO

HISTOGRAMA

É a representação gráfica da distribuição de frequência, feita através de retângulos justapostos, cujas bases se localizam sobre o eixo horizontal, de tal modo que seus pontos médios coincidam com os pontos médios dos intervalos de classes.



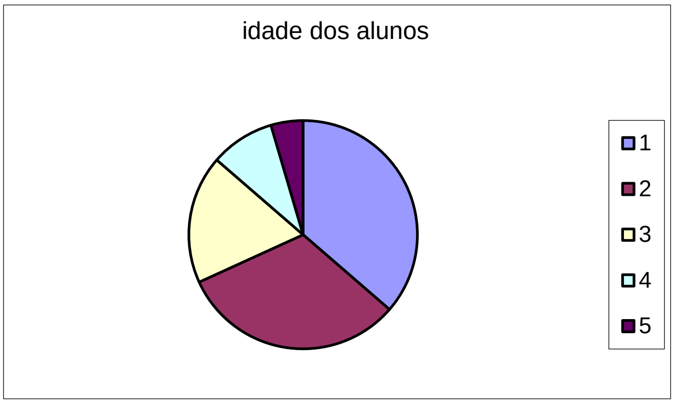
POLÍGONO DE FREQUÊNCIA

É um gráfico em linha, feito pelas ligações dos pontos médios, formando assim um polígono.

GRÁFICO EM SETORES

Este gráfico é construído com base em um círculo, e é empregado sempre que desejamos ressaltar a participação do dado no total.

O total é representado pelo círculo, que fica dividido em tantos setores quantas são as partes. Os setores são proporcionais aos dados da série. Obtemos cada setor por meio de uma regra de três simples, onde o total corresponde a 360°.



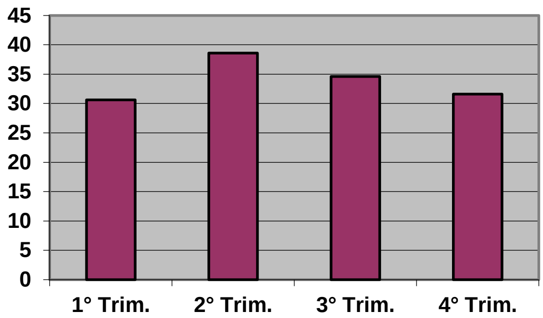
$\frac{22}{8} = \frac{360°}{x}$ → 22x = 360.8 → x = 360.8/22 → x = 131°

$\frac{22}{7} = \frac{360°}{x} \rightarrow 22x = 360.7 \rightarrow x = 115°$

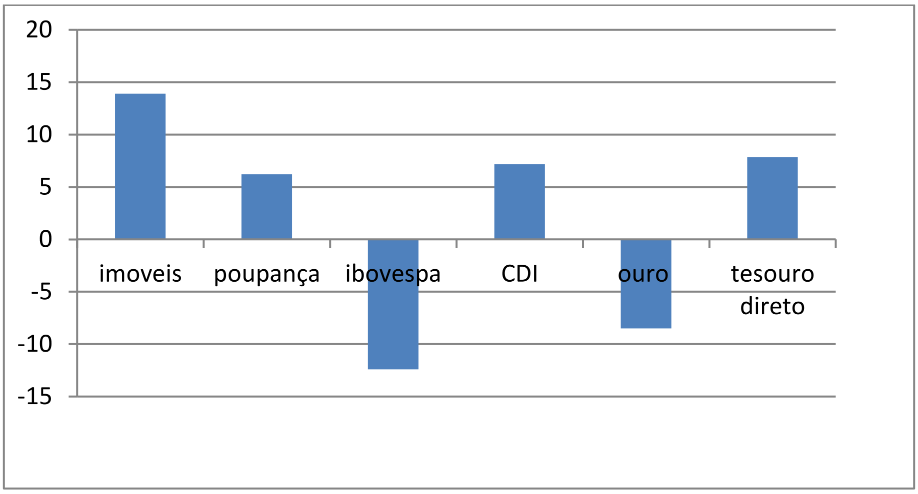
$$\frac{22}{4} = \frac{360}{x} \rightarrow 22x = 360.4 \rightarrow x = \frac{360.4}{22} = 65°$$

GRÁFICO EM COLUNAS

Constroem-se barras retangulares, separadas. Normalmente utilizado para representar variáveis qualitativas.



O gráfico abaixo mostra de forma comparativa a rentabilidade de alguns investimentos entre os meses de agosto 2012 e agosto 2013.



Fonte: *Jornal meu Imóvel*-Verde Brasil Editorial- São Paulo (Jornal de propaganda de imóveis) .

ATIVIDADE

3.3. Construir um gráfico, utilizando os dados da tabela abaixo:

a) Evolução do PIB brasileiro, ano a ano em percentual.
Governo Lula de 2003 a 2010, governo Dilma de 2011 a 2015 e governo Temer de 2016 a 2017.

ano	evolução (%)
2003	1,1
2004	5,8
2005	3,2
2006	4,0
2007	6,1
2008	5,1
2009	-0,1
2010	7,5
2011	4,0
2012	1,9
2013	3,0
2014	0,5
2015	-3,5
2016	-3,5
2017	1

Fonte: Banco Mundial

3.4. Seja a distribuição percentual da população ocupada por anos de estudos, em 2003 e 2012.

Distribuição	2003	2012
11 anos ou mais de		
estudo	53,50%	68,70%
8 a 10 anos de estudo	19,80%	16,10%
sem instrução ou com		
menos de 8 anos de		
estudo	26,80%	15,30%

Fonte: IBGE

Construa gráficos em setor para expressar a evolução de 2003 a 2012.

3.5. Complete a distribuição de frequências representando a amostra dos salários de 25 funcionários selecionados em uma empresa.

Classe	**Salários R$**	**F_i**
1	1.000,00 ⊢ 1.200,00	2
2	1.200,00 ⊢ 1.400,00	6
3	1.400,00 ⊢ 1.600,00	10
4	1.600,00 ⊢ 1.800,00	5
5	1.800,00 ⊢ 2.000,00	2

3.6. Construa o histograma, o polígono de frequência absoluta e o polígono de frequência acumulada para a distribuição acima.

3.7. Construa um gráfico em barras para expressar a dívida pública do Brasil em % do PIB, conforme tabela abaixo:

Ano	%
2013	51,5
2014	56,3
2015	65,5
2016	69,9

Fonte: www1.folha.uol.com.br de 01.01.2018

Capítulo 4

4. MEDIDAS DE TENDÊNCIA CENTRAL

O estudo que fizemos sobre distribuições de frequência, permite-nos localizar, por exemplo, a maior ou menor concentração de valores, mas se quisermos ressaltar as tendências características de cada distribuição, precisaremos estudar outras medidas, como exemplo as medidas de posição.

As medidas de posição mais importantes são as medidas de tendência central, cujo nome vem do fato de os dados observados tenderem em geral, a se agrupar em torno dos valores centrais.

4.1. MÉDIA ARITMÉTICA

4.1.1 MÉDIA ARITMÉTICA SIMPLES

É o quociente da divisão da soma dos valores da variável pelo número deles:

$$\boxed{\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}}$$

Exemplo:

a) Sabendo-se que o número de aulas semanais dadas por 7 professores foi de 10; 14; 13; 15; 16; 18 e 12 horas/aula, o número médio de aulas dadas será

$x_1 = 10$; $x_2 = 14$; $x_3 = 13$.......

$$\overline{X} = \frac{10+14+13+15+16+18+12}{7} = 14$$

b) Notas: 6; 8; 7; 10; 9; 6; 6; 8; 9; 7

$$\bar{X} = \frac{6+8+7+10+9+6+6+8+9+7}{10} = 7{,}6$$

<u>Desvio em relação à média:</u>

É a diferença entre cada elemento de um conjunto de valores e a média aritmética.

$D_i = x_i - \overline{X}$

Para o exemplo dado, temos:

$d_1 = x_1 - \overline{X} = 10-14 = -4$

$d_2 = x_2 - \overline{X} = 14-14 = 0$

PROPRIEDADES DA MÉDIA

1. $\sum d_i = 0$
   Nos exemplos anteriores, temos:

a) -4+0-1+1+2+4-2=0

b) 6; 8; 7; 10; 9; 6; 6; 8; 9; 7 = 6; 6; 6; 7; 7; 8; 8; 9; 9; 10
   $D_1 = d_2 = d_3 =$ 6 -7,6 = -1,6
   $D_4 = d_5$ = 7-7,6 = -0,6
   $D_6 = d_7$ = 8 -7,6 = 0,4
   $D_8 = d_9$ = 9 – 7,6 = 1,4
   D_{10} = 10 - 7,6 = 2,4

$$\sum d_i = -1{,}6 - 1{,}6 - 1{,}6 - 0{,}6 - 0{,}6 + 0{,}4 + 0{,}4 + 1{,}4 + 1{,}4 + 2{,}4 = 0$$

2. Somando-se ou subtraindo-se uma constante de todos os valores de uma variável, a média do conjunto fica aumentada ou diminuída dessa constante.
   Some 2 a cada um dos valores da variável do exemplo dado:
   10; 14; 13; 15; 16; 18 e 12 horas/aula
   10 + 2 = 12
   14 + 2 = 16
   13 + 2 = 15
   15 + 2 = 17
   16 + 2 = 18
   18 + 2 = 20
   12 + 2 = 14

   A média será $\overline{X}$ = 16 = 14 + 2

3. Multiplicando-se ou dividindo-se todos os valores de uma variável por uma constante, a média do conjunto fica multiplicada (ou dividida) por essa constante.
   Multiplique por 3 cada elemento da série dada. A média será

$\overline{X}$ = X.3 = 14.3 = 42

4.1.2 MÉDIA ARITMÉTICA PONDERADA

Se os dados estão apresentados na forma de uma variável discreta, utilizaremos a média aritmética ponderada:

$$\boxed{\overline{\boldsymbol{X_i}} = \frac{\sum \boldsymbol{x_i}.\boldsymbol{f_i}}{\sum \boldsymbol{f_i}}}$$

ARREDONDAMENTOS

Se o último algarismo for:

- >5, então arredondamos para mais, ou seja, o algarismo anterior terá acréscimo de uma unidade.

Exemplo: 5,8 = 6
7,6 = 8
4,666666 = 4,67
3,8999 = 3,90

- <5, então mantem-se o algarismo anterior.

Exemplos: 5,4 = 5
7,1 = 7
5,339 = 5,3
5,33333 = 5,33

- =5, verifica-se o algarismo anterior: se for par, então mantem-se o algarismo. Se for ímpar, então aumenta-se de uma unidade o algarismo anterior

Exemplos:
5,5 = 6
8,5 = 8
5,45 = 5,4
6,55 = 6,6
6,445 = 6,4
7,475 = 7,48

Exemplos:

1. Determinar a média da distribuição de notas obtidas pelas lojas on-line, sendo consultados 10 clientes:

X_i = notas	f_i = frequência
2	1
5	4
6	3
8	2

2;5;5;5;5;6;6;6;8;8 → $\bar{x} = \frac{\sum_1^{10} x_i}{10} = \frac{2+5+5+5+5+6+6+6+8+8}{10} = 5{,}6$

Devemos acrescentar mais uma coluna: $x_i f_i$

x_i	f_i	$X_i f_i$
2	1	02
5	4	20
6	3	18
8	2	16
	$\sum f_i = 10$	$\sum x_i f_i = 56$

$\overline{X}$ = 56/10 = 5,6

Se os dados estão apresentados na forma de uma variável contínua, utilizaremos a média aritmética ponderada, considerando as FREQUÊNCIA simples das classes como sendo as ponderações dos pontos médios destas classes (x_i).

2. Determinar a média da distribuição de falhas ocorridas em um sistema durante 14 dias:

Classe	Nº falhas	f_i	Ponto médio	$x_i.fi$
1	0 ⊢ 2	3	1	3
2	2 ⊢ 4	4	3	12
3	4 ⊢ 6	6	5	30
4	6 ⊢ 8	1	7	7
total		14		52

$\overline{Xi} = \frac{52}{14} = 3{,}71$

Exemplo de notas com ponderação:

A_1 →peso =1 → nota = 5

A_2→ peso =2 → nota = 5

A_3 → peso= 3→ nota = 7

$$\bar{x} = \frac{5x1 + 5x2 + 7x3}{6} = 6$$

Se fosse sem ponderação: $\bar{x} = \frac{7+5+5}{3} = 5{,}666666\ldots = 5{,}67$

ATIVIDADE

4.1. Em uma escola o número de funcionários em cada função é: 4 serventes, 15 professores sub, 20 professores P1, 50 professores P2 e 3 coordenadores pedagógicos. Os salários para cada função são, respectivamente, R$200,00; R$300,00; R$500,00; R$1.000,00 e R$3.000,00. Qual a média de salários pagos por essa escola?

4.2. O salário de 40 funcionários de uma empresa está distribuído segundo o quadro abaixo. Calcule o salário médio destes funcionários

Classe	Salários $	Nº de funcionários f_i		
1	400 ⊢ 500	12		
2	500 ⊢ 600	15		
3	600 ⊢ 700	8		
4	700 ⊢ 800	3		
5	800 ⊢900	1		
6	900 ⊢1000	1		

4.3. Uma empresa de TI observou em seus registros recentes, o tempo de mão de obra gasto na revisão de um programa:

Classe	Tempo de mão de obra(horas)	Nº de programas f_i		
1	0 ⊢4	1		
2	4 ⊢8	5		
3	8 ⊢12	10		
4	12⊢16	12		
5	16⊢20	4		

a) Determine o número de horas necessárias para a revisão de cada programa
b) Qual deve ser o tempo total para a revisão de dez programas?
c) Se a empresa dispõe no momento de três funcionários trabalhando 8 horas por dia nestas revisões conseguirá provavelmente revisar estes dez programas em quatro dias?

4.4. Considere uma pesquisa realizada com famílias de quatro filhos, tomando para variável o nº de filhos do sexo masculino. Determine a média de filhos homens por família.

Nº de meninos	f_i	
0	1	
1	4	
2	8	
3	1	
4	4	
	∑=18	

4.2 MÉDIA GEOMÉTRICA

4.2.1. Média Geométrica Simples

$$\mathbf{Mg} = \sqrt[N]{x_{1}.\,x_{2}.x_{3}\;\ldots.\,x_{N}}$$

Exemplo:

4.2.1 Calcule a média aritmética e a média geométrica simples entre:
3,5,7 e 9.
Ma = 6
Mg= 5,54

Média Geométrica Simples para número de dados grande ou valores muito grandes

Quando o número de observações for muito grande, é aconselhável o emprego de logaritmos (decimal ou neperiano)

Temos que $\mathbf{Mg} = \sqrt[N]{x_{1}.\,x_{2}.\,x_{3}\;\ldots.\,x_{N}}$

Lembrando de logaritmos: x = 100 → log100 = 2 ↔ $10^2 = 100$

$\log 10^2 = 2.\log 10 = 2.1 = 2$

log(10.10) = log10 + log10 = 1 + 1 = 2

Logaritmando, então os dois lados da equação, temos:

$$\log Mg = \log \sqrt[N]{\boldsymbol{x_1.x_2.x_3 \ldots .x_N}} = \boldsymbol{\log(x_1.x_2 \ldots x_N)^{\frac{1}{N}}}$$

$$\log Mg = \frac{1}{N}\log(\boldsymbol{x_1.x_2 \ldots x_N})$$

$$\log Mg = \frac{1}{N}(\log \boldsymbol{x_1} + \boldsymbol{\log x_2} \ldots + \boldsymbol{\log x_N})$$

$$\log Mg = \frac{1}{N}\sum \log x_i$$

$$\boxed{Mg = \text{antilog}\left[\frac{\sum \log x_i}{N}\right]}$$

Comparando com a média aritmética: $\bar{x} = \frac{\sum x_i}{N}$, obtém-se o antilog da média aritmética dos logaritmos dos valores.

Exemplo:

4.2.2 Calcular Mg de 5,7,15,20 e 23

$$Mg = \sqrt[5]{5.7.15.20.23} = \sqrt[5]{241.500} = 11{,}93$$

x_i	$\log x_i$
5	0,6990
7	0,8451
15	1,1761
20	1,3010
23	1,3617
total	5,3829

Mg = antilog$\left[\frac{\sum \log x_i}{N}\right]$ = antilog($\frac{5{,}3829}{5}$) = antilog(1,07658) = 11,93

Exercícios

1. *Calcular a média geométrica dos salários:*

x_i	$logx_i$
1500	*3,1761*
1700	*3,2304*
2100	*3,3222*
2300	*3,3617*
2500	*3,3979*
3500	*3,5441*
6700	*3,8261*
10.350	*4,0149*
15.000	*4,1761*
Total	*32,0495*

Mg = $\sqrt[9]{1{,}12090.\,10^{32}} = 3.639{,}71$

Ou Mg = antilog$\left[\frac{\sum logx_i}{N}\right]$ = antilog($\frac{32{,}0495}{9}$) = antilog(3,5611) = 3.639,61

Obs. A diferença entre os valores ocorre devido aos arredondamentos realizados na segunda coluna.

No exemplo abaixo, não será possível realizar a conta através da fórmula

Mg = $\sqrt[N]{\boldsymbol{x_1.\,.x_2.x_3 \ldots .x_N}}$

Teremos que utilizar a fórmula a outra fórmula.

2. *Calcular a média geométrica dos salários relacionados abaixo:*

x_i	$logx_i$
1500	*3,1761*
1700	*3,2304*
2100	*3,3222*
2300	*3,3617*
2500	*3,3979*
3500	*3,5441*
6700	*3,8261*
10.350	*4,0149*
15.000	*4,1761*
28.000	*4,4472*
30.000	*4,4771*
33.900	*4,5302*
45.000	*4,6532*
57.000	*4,7559*
Total	*54,9131*

Mg = antilog$\left[\frac{\sum logx_i}{N}\right]$ = antilog($\frac{54{,}9132}{14}$) = antilog(3,9224) = 8.363,14

4.2.2 Média Geométrica Ponderada

Exemplo:

x_i	f_i
2	3
8	2
10	1

$$\boldsymbol{Mg = \sqrt[6]{2^3.8^2.10} = \sqrt[6]{5120} = 4{,}15}$$

$$Mg = \sqrt[f_1+f_2+..f_N]{x_1^{f_1}.x_2^{f_2}\dots x_N^{f_N}}$$

$$\boldsymbol{Mg = \sqrt[\Sigma f_i]{\prod x_i^{f_i}}}$$

Ou

$$E \log Mg = \log\left(\prod (x_i)^{fi}\right)^{1/\Sigma fi}$$

$$\log Mg = \frac{1}{\Sigma f_i}\log\left(\prod (x_i)^{fi}\right)$$

$$\log Mg = \frac{1}{N}\sum f_i \log(x_i)$$

$$\boxed{Mg = \text{antilog}\left[\frac{\Sigma f_i.\,log x_i}{N}\right]}$$

ATIVIDADE

4.5 Qual a média geométrica de x= {3,6,12,24,48}

Resp: 12

4.6 Dê a Mg de 2,3,4,5 e 6 ponderados por 6,5,4,3,2

Resp: 3,26

4.7 Uma experiência com 25 indivíduos consistiu em determinar a *concentração mínima inibitória(CMI)*de determinado antibiótico para um certo tipo de bactéria:

CMI(μg/ml)	**Número de indivíduos**
1,00000	1
0,50000	2
0,25000	6
0,12500	4
0,06250	3
0,03125	9
Total	25

Fonte: Arango, H.G.,Bioestatística, teórica e computacional, Guanabara-Koogan, 2001

Obs: Note que a distribuição de frequências da CMI é concentrada do lado esquerdo.

Determine a média geométrica ponderada.

O uso da média aritmética simples, neste caso, não fornece um resultado adequado. O ideal é calcular a média geométrica simples. A CMI normal para este tipo de bactéria é de *0,10 μg/ml.*
Resp: 0,100

4.3. MEDIANA

É o número que se encontra no centro de uma série de números, estando estes dispostos segundo uma ordem. Separa em dois subconjuntos de mesmo número de elementos.

Exemplo:

Dada a série: 5,13,10,2,18,15,6,16,9

Primeiro é necessário ordenar:
2,5,6,9,10,13,15,16,18

O valor central que apresenta o mesmo nº de elementos à direita e à esquerda é o 10.
Md = 10

Se a série tiver um nº par de termos, a mediana será qualquer dos números compreendidos entre os dois valores centrais. Utiliza-se o ponto médio nesse caso.

Exemplos:
a) 2,6,7,10,12,13,18,21
Md = 11

b) 5; 7; 10; 13; 15
Média = 10 e M_d = 10

c) 5; 7; 10; 13; 65
Média = 20 e M_d =10

Observações:
1- A mediana depende da posição, e não do valor dos elementos.
2- O valor da mediana pode ou não coincidir com um elemento da série.
3- A mediana e a média não têm, necessariamente o mesmo valor.
4- A mediana pode ser chamada por valor mediano

4.3.1 MEDIANA DE DADOS AGRUPADOS

Posição da Mediana = $\sum f_i/2$

- *<u>Sem intervalos de classe</u>*

Determinar a mediana da distribuição de notas a seguir:

Notas (X_i)	f_i	f_{ac}
2	1	1
5	4	5
6	3	8
8	2	10
	N = 10	

2,5,5,5,5,6,6,6,8,8

A mediana será o valor que ocupa a posição: 10/2 → Entre 5ª e 6ª. Se construirmos a coluna das f_{ac} verificamos que esse elemento deve estar na 2ª e 3ª classe, ou seja, a média entre as notas 5 e 6 : Md = (5 + 6)/2= 5,5

Exemplo:

Calcule a mediana:

x_i	f_i	f_{ac}
12	1	1
14	2	3
15	1	4
16	2	6
17	1	7
20	1	8
	∑=8	

$\sum f_i/2 = 8/2 = 4$. Mas fac3 = 4, então $Md = \frac{x_i + x_{i+1}}{2}$

Md = (15+16)/2 = 15,5

ATIVIDADE

4.3.1. Determine a mediana:

x_i	f_i	f_{ac}
2	3	3
4	7	10
6	12	22
8	8	30
10	4	34

Posição da mediana: N/2 = 34/2 = 17→ Posição da mediana estará entre 17ª e a 18ª
Md = (6+6)/2 = 6

- *Com intervalos de classes*

4.3.2. Considere a distribuição de FREQUÊNCIA dos salários de funcionários selecionados em uma empresa.

Classe	**Salários R$**	**F_i**	**Fac**	
1	1.000,00 ⊢ 1.200,00	2	2	
2	1.200,00 ⊢ 1.400,00	6	8	
3	1.400,00 ⊢ 1.600,00	10	18	
4	1.600,00 ⊢ 1.800,00	5	23	
5	1.800,00 ⊢ 2.000,00	2	25	
total		25		

A mediana é o valor de separa o nº de elementos da série em dois grupos, contendo cada um deles 50% dos elementos. Portanto, a posição da mediana na série é N/2. No ex. 25/2 = 12,5ª posição.

O valor decimal 12,5 indica que a mediana é um valor posicionado entre o décimo segundo e o décimo terceiro elemento da série.

Verificando na coluna da FREQUÊNCIA acumulada para identificar em qual classe estão situados esses elementos: 3ª classe.

A mediana deve ser um valor entre R$1.400,00 e R$1.600,00. Esse intervalo contém dez elementos. Supondo que eles estão uniformemente distribuídos neste intervalo, então dividiremos este intervalo de modo proporcional à posição da mediana na série.

POSIÇÃO	8ª	12,5ª	18ª
VALOR	1400	Md	1600

Ou seja $\frac{18-8}{1600-1400} = \frac{12{,}5-8}{Md-1400}$

(Md –1400). 10 = (12,5 – 8).200

$$Md = 1400 + \frac{12{,}5 - 8}{10} \cdot 200 = 1.490{,}00$$

1400 é o l_i da classe mediana

12,5 é posição central dos elementos da série, N/2

8 é a FREQUÊNCIA acumulada da classe anterior à classe mediana

10 é a FREQUÊNCIA simples da classe mediana

200 é a amplitude do intervalo de classe

$$\boxed{Md = li_{(md)} + \frac{N/2 - Fac_{ant.}}{f_{md}} \cdot h}$$

ATIVIDADE

4.6. Calcule a mediana da sequência:

2,5,8,10,12,15,8,5,12

4.7. Calcule a mediana da distribuição:

x_i	f_i	
2	5	
4	20	
5	32	
6	40	
8	2	

4.8. Salário de 25 funcionários:

classe	Salários –R$	Nº de func.	
1	1000⊢1400	4	
2	1400⊢1800	10	
3	1800⊢2200	16	
4	2200⊢2600	6	
5	2600⊢3000	2	

4.9. A coordenação de vendas de uma empresa, resolveu premiar com uma viagem a metade dos vendedores que contribuírem com mais "vendas". Para isto, fez um levantamento mensal, obtendo a tabela: A partir de qual volume de vendas (em real) a turma será premiada?

Vendas- R$	Nº de turmas		
0⊢100,00	1		
100,00⊢200,00	12		
200,00⊢300,00	27		
300,00⊢400,00	31		
400,00⊢500,00	10		

LISTA DE EXERCÍCIOS

4.1. Dada a tabela abaixo, determine a mediana e a média aritmética simples

notas	f_i	F_{ac}	$x_i f_i$
4	**1**		
5	**2**		
6	**8**		
7	**8**		
8	**4**		
9	**3**		
10	**1**		
total			

4.2. Dada a sequência, determine a mediana e a média aritmética
123; 124; 131; 144; 156; 207

4.3. A tabela abaixo mostra a distribuição do peso de crianças em uma escola:

classes	Pesos(kg)	f_i	x_i	$x_i f_i$	f_{ac}
1	$20 \vdash 30$	15			
2	$30 \vdash 40$	23			
3	$40 \vdash 50$	5			
4	$50 \vdash 60$	2			
5	$60 \vdash 70$	1			
total					

Determinas a média aritmética e a mediana.

4.4. MODA

É o valor que ocorre com maior frequência em uma série de valores.
Moda = valor que mais se repete

Exemplos:

4.4.1. Seja a série 3,4,5,5,5,6,6,6,6,6,7,8,9,9
A moda Mo = 6

Se não ocorrer valores que se repetem a série é amodal
Ex. 3,5,7,8,9

Se houver dois ou mais valores de concentração, dizemos que a série tem dois ou mais valores modais.
Ex.2,3,4,4,4,5,6,7,7,7,8,9 – série bimodal → Mo = 4;7

- Variável discreta: É só identificar o elemento de MAIOR FREQUÊNCIA na série:

4.4.2. Notas obtidas em uma turma:

Notas (x_i)	f_i
0	2
2	5
5	8
7	3
9	1

A maior FREQUÊNCIA é 8 e corresponde ao elemento 5. Mo = 5

- Variável contínua:

Para determinar a moda de uma variável contínua, podemos optar por vários processos:

4.4.1. MODA de PEARSON

Mo = 3 Md – 2 $\overline{X}$

Exemplo:

4.4.3. Salários em uma empresa

Classe	**Salários R$**	**F_i**	**X_i**	**X_i f_i**	**fac**	
1	1.000,00 ⊢ 1.200,00	2	1100	2200	2	
2	1.200,00 ⊢ 1.400,00	6	1300	7800	8	
3	1.400,00 ⊢ 1.600,00	10	1500	15000	18	
4	1.600,00 ⊢ 1.800,00	5	1700	8500	23	
5	1.800,00 ⊢ 2.000,00	2	1900	3800	25	
total		25		37300		

Cálculo da média e da mediana

$\overline{X}$ = 37300/25= R$ 1.492,00

Md= R$ 1.490,00

Mo = 3x 1490 – 2 x 1492 = R$1.486,00

4.4.2. MODA de KING

King propôs outra fórmula para cálculo da moda, onde se considera tanto a frequência da classe anterior como da posterior:

$$Mo = l_{Mo} + \frac{f_{post}}{f_{ant} + f_{post}} . h$$

Onde:

$$l_{mo} = limite\ inferior\ da\ classe\ modal$$
$$f_{post} = frequência\ simples\ da\ classe\ posterior\ à\ classe\ modal$$
$$f_{ant} = frequência\ simples\ anterior\ à\ classe\ modal$$
$$h = amplitude\ do\ intervalo\ de\ classe.$$

Exemplo:

4.4.4. Calcule a moda de King para a distribuição de salários a seguir:

Classe	Salários (R$)	Número funcionários
1	1000 ⊢ 1400	10
2	1400 ⊢ 1800	16
3	1800 ⊢ 2200	08
4	2200 ⊢ 2600	06
5	2600 ⊢ 3000	04
Total		44

$$M_o = 1400 + \frac{8}{10 + 8} . 400 = 1577,78$$

4.4.5. Determinar a moda de King para a tabela do exemplo 4.4.3

Resp: R$1.490,91

4.4.3. MODA DE CZUBER

A fórmula de Czuber é a mais precisa das três, a de King é a mais simples e a de Pearson é a mais fácil, desde que se tenha a média e a mediana. Geralmente, é a mais utilizada em concursos públicos.

$$M_o = l_{infMo} + \frac{f_{Mo} - f_{ant}}{2f_{Mo} - (f_{ant} + f_{post})} . h$$

Onde

$$l_{mo} = limite\ inferior\ da\ classe\ modal$$
$$f_{Mo} = frequência\ da\ classe\ modal$$
$$f_{post} = frequência\ simples\ da\ classe\ posterior\ à\ classe\ modal$$
$$f_{ant} = frequência\ simples\ anterior\ à\ classe\ modal$$
$$h = amplitude\ do\ intervalo\ de\ classe.$$

Exemplo:

4.4.6. Determinar a moda de Czuber da tabela do exemplo 4.4.4

$$M_o = 1400 + \frac{16 - 10}{2x16 - (10 + 8)} . 400 = 1.571,43$$

4.4.7. Determinar a moda de Czuber da tabela do exemplo 4.4.3

Resp: R$1.488,89

ATIVIDADE

4.10. Determine a moda:

Número de faltas dos alunos em um bimestre	f_i	
0	1	
1	4	
2	8	
3	1	
4	4	
	∑=18	

4.11. Determine a moda de Pearson da distribuição abaixo, com variável contínua:

Classe	**Salários R$**	**F_i**
1	1.000,00 ⊢ 3.000,00	20
2	3.000,00 ⊢ 5.000,00	60
3	5.000,00 ⊢ 7.000,00	100
4	7.000,00 ⊢ 9.000,00	50
5	9.000,00 ⊢ 11.000,00	20
Total		

4.12. (ESAF Oficial de Justiça Avaliador TJ CE/02) A tabela abaixo apresenta a distribuição de frequências do atributo salário mensal medido em quantidade de salários-mínimos para uma amostra de 200 funcionários da empresa X. Note que a coluna classes refere-se a classes salariais em quantidades de salários mínimos e que a coluna P refere-se ao percentual da frequência acumulada relativo ao total da amostra. Não existem observações coincidentes com os extremos da classe.

classes	P
4 – 8	20
8 – 12	60
12 – 16	80
16 – 20	98
20 - 24	100

a) Qual o salário médio amostral calculado a partir de dados agrupados?
b) Qual o salário modal no conceito de Czuber?
c) Qual o salário modal no conceito de Pearson?
d) Qual o salário modal no conceito de King?

Adaptado do livro Estatística para concursos de Adriano Leal Bruni, ed.Atlas, 2008.

Capítulo 5

5. MEDIDAS SEPARATRIZES

5.1 QUARTIS:

Dividem um conjunto de dados em 4 partes iguais

0 → 25% → 50% → 75% → 100%

O quartil 2 coincide com a mediana
Determinação:
1º) Calcula-se a posição do quartil pela fórmula in/4
2º) Identifica-se a classe Q1 pela fac
3º) Aplica-se a fórmula:

$$Q_i = l_{Qi} + \frac{\left(\frac{in}{4} - fac_{ant}\right).h}{fQ_i}$$

Exemplos:

Xi (idade)	fi	Fac
7⊢17	6	6
17⊢27	15	21
27⊢37	20	41
37⊢47	10	51
47⊢57	5	56
total	56	

Posição do quartil 1: iN/4 = 1.56/4 = 14
$Q_1 = 17 + \frac{14-6}{15}.10 = 22{,}33$

Posição do quartil 3: 3x56/4 = 42

$$Q_3 = 37 + \frac{42-41}{10}.10 = 38$$

5.2. DECIS

São os valores que dividem a série em 10 partes iguais

0% → 10% → 20% → 30% → 40% → 50% → 60% → 70% → 80% → 90% → 100%

$$\mathrm{D_i} = \mathrm{l_{di}} + \frac{\left(\frac{\mathrm{in}}{10} - \mathrm{fac_{ant}}\right).\mathrm{h}}{\mathrm{f_{Di}}}$$

Exemplo:

xi	Fi	Fac
4⊢9	8	**8**
9⊢14	12	**20**
14⊢19	17	**37**
19⊢24	3	**40**
total	40	

Posição do decil 4: in/10 = 4.40/10 = 16

$$D_4 = 9 + \frac{16 - 8}{12} . 5 = 12{,}33$$

5.3. PERCENTIS

Dividem a série em 100 partes iguais

$$\text{Pi} = l_{pi} + \frac{\left(\frac{in}{100} - fac_{ant}\right).h}{f_{pi}}$$

Exemplo:

Considerando a tabela anterior,
Posição do percentil 72: in/100 = 72.40/100 =28,8

$$P_{72} = 14 + \frac{28{,}8 - 20}{17} . 5 = 16{,}59$$

ATIVIDADE

5.1. Calcule a moda, quartil 1 e percentil 75 das séries abaixo:
a)2,3,5,4,2,5,7
b)

x_i	f_i
2	1
3	7
4	2
5	2

5.2. Calcule a média aritmética, a moda de Pearson, quartil 1, decil 8 e percentil 70 para a distribuição que representa a nota de 60 alunos em uma prova:

Classe	Notas	Nº de alunos
1	0 ⊢2	5
2	2 ⊢4	20
3	4 ⊢6	12
4	6 ⊢8	20
5	8⊢10	3

5.3. Calcule a média aritmética simples, a mediana, a moda de King, o decil 3 e o percentil 10 da distribuição que representa as alturas de 80 alunos de uma classe:

Classe	Altura	N º de alunos
1	150 ⊢160	2
2	160 ⊢170	15
3	170 ⊢180	18
4	180 ⊢190	28
5	190 ⊢200	16
6	200 ⊢210	1

Capítulo 6

6. MEDIDAS DE DISPERSÃO OU DE VARIABILIDADE

Podemos sintetizar um conjunto de valores em poucos valores representativos: a média, a mediana e a moda. No entanto, mesmo essas medidas de tendência central não são suficientes para ter-se uma ideia retrospectiva de como os dados se apresentavam na tabela.

Exemplo:

O preço médio de uma mercadoria ser igual a R$70,00 em várias cidades pode significar uma oscilação de R$5,00 à R$160,00 ou então de R$68,00 à R$72,00. Vamos analisar os seguintes conjuntos de valores:

A: 70, 70, 70, 70, 70
B: 68, 69, 70, 71, 72
C: 5, 15, 50, 120, 160

Calculando a média aritmética de cada um desses conjuntos, obtemos:

X_A = 350/5 = 70
X_B = 350/5 = 70
X_C = 350/5 = 70

Vejamos a amplitude total de cada série:

A_a = 70 - 70 = 0
A_b = 72 - 68 = 4
A_c = 160 – 5 = 155

A amplitude total é um valor instável, por se deixar influenciar por valores extremos, que são na sua maioria, devido ao acaso
Para evitar essa falha define-se as medidas: variância e desvio padrão
Vejamos inicialmente o conceito de desvio médio simples:

6.1. DESVIO MÉDIO SIMPLES (D.M.S.)

Rol: DMS = $\frac{\sum|x_i - \bar{x}|}{N} = \sum\frac{|x_i - \bar{x}|}{N}$

A: 70, 70, 70, 70, 70
Desvios:
70-70 = 0
70-70 = 0
70-70 = 0
70-70 = 0
70-70 = 0
Média dos desvios = 0

B: 68, 69, 70, 71, 72
Desvios em relação à média:
D_1 = 68-70 = -2
D_2 = 69-70 = -1
D_3 = 70-70 = 0
D_4 = 71-70 = 1
D_5 = 72-70 = 2

C: 5, 15, 50, 120, 160
D_1 = 5-70 = -65
D2 = 15-70 = -55
D_3 = 50-70 = -20
D_4 = 120-70 = 50
D_5 = 160-70 = 90

Em módulo:

D_1 = 68-70 = -2 → $|68-70| = 2$
D_2 = 69-70 = -1 → $|69-70| = 1$
D_3 = 70-70 = 0 → $|70-70| = 0$
D_4 = 71-70 = 1 → $|71-70| = 1$
D_5 = 72-70 = 2 → $|72-70| = 2$

Média dos desvios: $\frac{2+1+0+1+2}{5} = \frac{6}{5} = 1{,}2$

D_1 = 5-70 = -65 → $|5-70| = 65$
D2 = 15-70 = -55 → $|15-70| = 55$
D_3 = 50-70 = -20 → $|50-70|$ = 20
D_4 = 120-70 = 50 → $|120-70|$ = 50
D_5 = 160-70 = 90 → $|160-70| = 90$
Média dos desvios = 56

Ex.2,8,5,6
$\overline{X}$ = (2+8+5+6)/4 = 5,25

$|x_1 - \overline{X}|$ = 3,25
$|x_2 - \overline{X}|$ = 2,75
$|x_3 - \overline{X}|$ = 0,25
$|x_4 - \overline{X}|$ = 0,75

O DMS é a média aritmética simples destes valores:

DMS = $\frac{3{,}15+2{,}75+0{,}25+0{,}75}{4}$ = 1,75

Em média, cada elemento da sequência está afastado do valor 5,25 por 1,75 unidades

a) Variável discreta:

x_i	f_i	X_if_i	$\vert x_i - \overline{X} \vert . f_i$
1	2	2	2x2 = 4
3	5	15	0x5 = 0
4	2	8	1x2 = 2
5	1	5	2x1 = 2
	10	30	8

A média da série é 30/10 =3

O DMS é:

DMS = $\frac{\sum|x_i - \overline{x}|.f_i}{\sum f_i}$ = 8/10 = 0,8

Em média cada elemento da série está afastado do valor 3 por 0,8

b) Variável contínua:

x	f_i	x_i	x_if_i	$\vert x_i - X \vert . f_i$
2⊢ 4	5	3	15	$\vert 3 - 5{,}1\vert .5 = 10{,}50$
4⊢ 6	10	5	50	$\vert 5 - 5{,}1\vert .10 = 1{,}00$
6⊢ 8	4	7	28	$\vert 7 - 5{,}1\vert .4 = 7{,}60$
8⊢ 10	1	9	9	$\vert 9 - 5{,}1\vert .1 = 3{,}90$
	20		102	Soma = 23

$$\boxed{DMS = \frac{\sum|x_i - X|.f_i}{\sum f_i}}$$

DMS = 23/20 = 1,15

Em média cada elemento da série está afastado de 1,15 da média 5,1

ATIVIDADE

6.1. Calcule o DMS das séries:

a)3,8,12,3,9,7

b)

x_i	f_i
2	3
4	8
5	10
6	6
8	2
10	1

c)

Classe	Notas	Nº de alunos
1	0 ⊢2	5
2	2 ⊢4	20
3	4 ⊢6	12
4	6 ⊢8	20
5	8⊢10	3

d)

Nº salários-mínimos	Nº de funcionários	
0 ⊢ 5	105	
5 ⊢ 10	92	
10 ⊢ 15	54	
15 ⊢ 20	32	
20 ⊢ 25	6	
25 ⊢ 30	1	

e)

idades	Nº alunos
4 ⊢ 8	2
8 ⊢ 12	6
12 ⊢ 16	8
16 ⊢ 20	9
20 ⊢24	10

f)

Idade (anos)	Nº de alunos			
17	3			
18	18			
19	17			
20	8			
21	4			

6.2. VARIÂNCIA

Uma forma de fazer com que as diferenças $x_i - \overline{X}$ sejam sempre positivas é considerar o quadrado dessas diferenças:

$(x_i - \overline{X})^2$

Se substituirmos esse quadrado na fórmula do DMS obtemos a fórmula da variância σ^2.

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{\sum f_i}$$

Exemplo:

a) Rol:
4,5,5,8
X =(4+5+8+5)/4=5,5
$(x_1\text{-}X)^2=(4\text{-}5{,}5)^2= 2{,}25$
$(x_2\text{-}X)^2= (5\text{-}5{,}5)^2=0{,}25$
$(x_3\text{-}X)^2= (8\text{-}5{,}5)^2=6{,}25$
$(x_4\text{-}X)^2 =(5\text{-}5{,}5)^2=0{,}25$

$\sum(x_i\text{-}X)^{2=}9$

σ^2= 9/4 = 2,25

6.3. DESVIO PADRÃO (σ)

É a raiz quadrada positiva da variância.

Sendo a variância calculada a partir dos quadrados dos desvios, ela é um número em unidade quadrada em relação à variável em questão, o que é inconveniente. Por isso, imaginou-se uma nova medida que tem utilidade e interpretação práticas, denominada desvio padrão.

$$\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2}{\sum f_i}}$$

Exemplos:

Cálculo do desvio padrão do exemplo anterior:
σ(x) = $\sqrt{\sigma^2} = \sqrt{2{,}25}$ = 1,5 unidades

a) Tabela com Variável discreta

x_i	f_i	x_if_i	$(x_i\text{-}X)^2.fi$
2	3	6	$(2\text{-}3{,}65)^2.3$ = 8,1675
3	5	15	$(3\text{-}3{,}65)^2.5$ = 2,1125
4	8	32	$(4\text{-}3{,}65)^2.8$ = 0,9800
5	4	20	$(5\text{-}3{,}65)^2.4$ = 7,2900
	20	73	18,55

Variância
Média = 73/20 = 3,65

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{x})^2 . f_i}{\sum f_i}$$

σ^2= 18,55/20 = 0,9275 → σ = √0,9275 = 0,963

b) Variável contínua:

Int.classe	f_i	x_i	$x_i f_i$	$(x_i\text{-}X)^2.f_i$
2 ⊢4	5	3	15	$(3\text{-}5{,}1)^2\ .5 = 22{,}05$
4 ⊢6	10	5	50	$(5\text{-}5{,}1)^2.10 = 0{,}1$
6 ⊢ 8	4	7	28	$(7\text{-}5{,}1)^2.4 = 14{,}44$
8 ⊢10	1	9	9	$(9\text{-}5{,}1)^2.1 = 15{,}21$
	$\sum f_i$=20		$\sum x_i f_i$=102	$\sum(x_i\text{-}x)^2.f_i = 51{,}8$

Média = 102/20 = 5,1

$$\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2 . f_i}{\sum f_i}}$$

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2 . f_i}{\sum f_i} = 51{,}8/20 \ = \ 2{,}59 \rightarrow \sigma \ = 1{,}61$$

Observações

No cálculo da variância quando elevamos ao quadrado a diferença (x_i- $\bar{X}$), a unidade de medida da série fica também elevada ao quadrado.
Ex. Dados expressos em metros, variância expressa em m^2. Em alguns casos a unidade de medida da variância nem faz sentido. Ex. litros^2.
Para suprir essa deficiência é que se define o desvio padrão, que tem sempre a mesma unidade de medida do dado.
Para o caso de dados amostrais, a fórmula para o desvio padrão será:

$$\boxed{\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2 . f_i}{N - 1}}}$$

Como em geral, os dados coletados pertencem a uma amostra extraída de uma população, podemos imaginar as amostras tornando-se cada vez mais amplas e a amplitude das classes ficando cada vez menor, o que nos permite concluir que a linha poligonal (contorno do polígono de frequência) tende a se transformar numa curva - a curva de frequência- mostrando, de modo mais evidente, a verdadeira natureza da distribuição da população. Podemos então dizer que enquanto o polígono de frequência nos dá a imagem real, a curva nos dá a imagem tendencial.
Assim, após o traçado de um polígono, é desejável, muitas vezes, que se faça um polimento no polígono, de modo a mostrar o que seria tal polígono com um nº maior de dados.

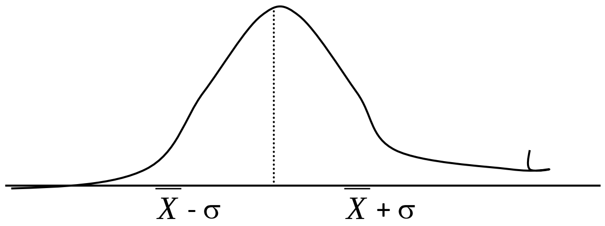
O polimento é obtido eliminando os vértices da linha poligonal. Utiliza-se a seguinte fórmula para obtenção dos pontos da curva:

$$F_i = \frac{f_{i-1} + 2\, f_i \ + f_{i+1}}{4}$$

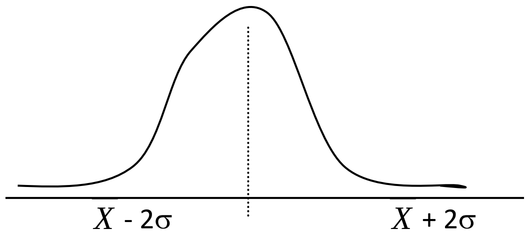
O desvio padrão é a medida de dispersão mais importante. È importante saber relacionar o valor obtido do desvio com os dados da série.

Quando uma curva de frequência é perfeitamente simétrica, como a curva abaixo, podemos afirmar que:

O intervalo [x-σ,x+σ] contém aproximadamente 68% dos valores da série.



O intervalo [x-2σ,x+2σ] contém aproximadamente 95% dos valores da série



O intervalo [x-3σ,x+3σ] contém aproximadamente 99% dos dados

Quando a distribuição não é simétrica, esses percentuais apresentam pequenas variações para mais ou para menos.

Quando afirmamos que $\overline{X}$ = 100 e σ = 5 significa que:

a) os valores da série estão concentrados em torno de 100
b) o intervalo [95,105] contém aproximadamente 68% dos valores da série
c) o intervalo [90,110] contém aproximadamente 95% dos valores da série
d) o intervalo [85,115] contém aproximadamente 99% dos valores da série

ATIVIDADE

6.2. Calcule a variância e o desvio padrão da população:

a) 2, 3, 7, 9, 11, 13
b) 5, 12, 4, 20, 13, 17
c) 6, 5, 10, 12, 19
d)

Idade (anos)	Nº de alunos			
17	3			
18	18			
19	17			
20	8			
21	4			

e) Estatura de 70 alunos

Estatura(cm)	FREQUÊNCIA			
150├─160	2			
160├─170	15			
170├─180	18			
180├─190	18			
190├─200	16			
200├─210	1			
	70			

f)

x_i	f_i
1	2
3	5
4	2
5	1
	10

g)

Classe	Notas	Nº de alunos
1	0 ⊢2	5
2	2 ⊢4	20
3	4 ⊢6	12
4	6 ⊢8	20
5	8⊢10	3

h) Uma série estatística simétrica apresenta: $\overline{X}$ = 10, σ^2 = 4 e σ = 2. Interprete esses valores

Capítulo 7

7. MEDIDAS DE DISPERSÃO RELATIVA

Se uma série X apresenta X=10 e desvio padrão σ=2 e uma série Y apresenta Y=100 e σ=5, do ponto de vista da dispersão absoluta, a série Y apresenta maior dispersão que a série X. No entanto, se levarmos em consideração as médias das séries, o desvio padrão de Y que é 5 em relação a 100 é um valor menos significativo que o desvio padrão de X que é 2 em relação a 10. Isto nos levar a definir as medidas de dispersão relativas:

7.1. COEFICIENTE DE VARIAÇÃO E VARIÂNCIA RELATIVA

O coeficiente de variação de uma série X é indicado por CV e definido por:

$$\boxed{CV = \frac{\sigma}{X}}$$

A variância relativa é indicada por v(x):

$$\boxed{V(x) = \sigma^2/(x)^2}$$

Assim CV (x) = 2/10=0,2 = 20%
e CV(y) = 5/100=0,05=5%

ATIVIDADE

7.1. Qual das séries abaixo apresenta maior dispersão? Dê a dispersão absoluta e relativa.

a)$X_a = 20$, $\sigma_A = 2$
$X_B = 20$, $\sigma_B = 5$

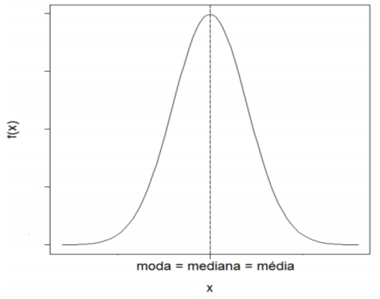
b)$X_A = 20$, $\sigma_A = 3$
$X_B = 60$, $\sigma_B = 9$

Capítulo 8

8.1. POSIÇÃO RELATIVA DA MÉDIA, MEDIANA E MODA.

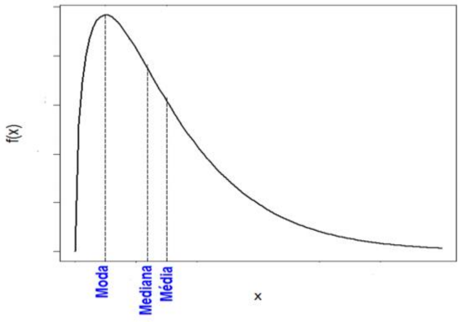
DISTRIBUIÇÃO SIMÉTRICA

As três medidas coincidem: $\overline{X}$= Md = Mo



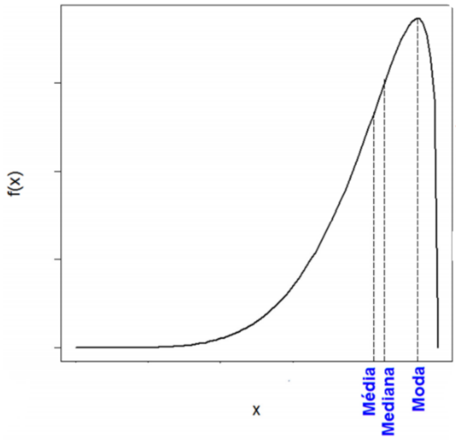
DISTRIBUIÇÃO ASSIMÉTRICA À DIREITA OU ASSIMÉTRICA POSITIVA:

Mo < Md < $\overline{X}$



DISTRIBUIÇÃO ASSIMÉTRICA À ESQUERDA OU ASSIMÉTRICA NEGATIVA:

$\overline{X}$ < Md < Mo



8.2. MEDIDAS DE ASSIMETRIA

COEFICIENTE DE PEARSON

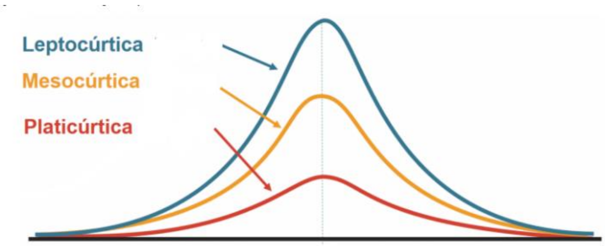
$$A_s = \frac{\bar{X} - M_o}{\sigma}$$

$A_s = 0$ distribuição é simétrica
$A_s < 0$ distribuição é assimétrica negativa
$A_s > 0$ distribuição é assimétrica positiva

8.3. MEDIDAS DE CURTOSE

Mede o grau de achatamento da curva de distribuição
- Observando a concentração dos valores de uma série em torno de sua moda, podemos observar três situações:

1-Os dados estão fortemente concentrados em torno da moda.
2-Os dados estão razoavelmente concentrados em torno da moda, o que faria a curva de frequência ser razoavelmente afilada
3-Os dados estão fracamente concentrados em torno da moda:



Para classificar uma distribuição quanto a sua curtose, podemos utilizar o coeficiente de curtose dado por:
Índice Momento de Curtose: Será dado pela seguinte fórmula:

$$\frac{m^4}{s_4}$$

Onde:

m^4 é o Momento de 4a Ordem Centrado na Média Aritmética; e S4 é o Desvio-Padrão do conjunto, elevado à quarta potência.

O numerador (m4): Quarto Momento Centrado na Média:

$$\frac{\sum(x_i - \bar{X})^4 . f_i}{N}$$

O denominador s_4: Quarta potência Desvio-Padrão ou segunda potência da variância:

$$\left(\frac{\sum(x_i - \bar{X})^2 . f_i}{N}\right)^2$$

$$C = \frac{\frac{\sum(x_i - \bar{X})^4 . f_i}{N}}{\left(\frac{\sum(x_i - \bar{X})^2 . f_i}{N}\right)^2}$$

C = 3 mesocúrtica
C > 3 leptocúrtica
C < 3 platicúrtica

O valor 3 representa o valor da curtose para uma distribuição teórica de probabilidades chamada distribuição normal padrão, que caracteriza distribuição mesocúrtica.

Exemplo:

Questão Extraída do AFRF-2002-2:

O atributo do tipo contínuo X, observado como um inteiro, numa amostra de tamanho 100 obtida de uma população de 1000 indivíduos, produziu a tabela de frequências seguinte:

	classes	frequência			
	29,5 – 39,5	4			
	39,5 – 49,5	8			
	49,5 – 59,5	14			
	59,5 – 69,5	20			
	69,5 – 79,5	26			
	79,5 – 89,5	18			
	89,5 – 99,5	10			

Para a distribuição de frequências do atributo X, sabe-se que:

$\sum(x_i - \bar{X})^2 . f_i = 24.500$ e $\sum(x_i - \bar{X})^4 . f_i$ = 14.682.500.

Nessas expressões os Xi representam os pontos médios das classes e $\overline{X}$ a média amostral. Assinale a opção correta. Considere para sua resposta a fórmula da curtose com base nos momentos centrados e suponha que o valor de curtose encontrado é populacional.

a) A distribuição do atributo X é leptocúrtica.

b) A distribuição do atributo X é platicúrtica.

c) A distribuição do atributo X é indefinida do ponto de vista da intensidade da curtose.

d) A informação dada se presta apenas ao cálculo do coeficiente de assimetria com base nos momentos centrados de X.

e) A distribuição de X é normal.

Resp: b) platicúrtica C = 2,45

Utiliza-se também para a medida de curtose o Índice Percentílico de Curtose: Encontraremos este índice usando a seguinte fórmula:

$$K = \frac{Q_3 - Q_1}{2(D_9 - D_1)}$$

Onde Q3 é o terceiro quartil;
Q1 é o primeiro quartil;
D9 é o nono decil e
D1 é o primeiro decil.

Se k < 0,263 → leptocúrtica
Se k = 0,263 → mesocúrtica
Se k > 0,263 → platicúrtica

Exemplo:

Questão extraída do concurso público AFRF-2002.1:

Em um ensaio para o estudo da distribuição de um atributo financeiro (X), foram examinados 200 itens de natureza contábil do balanço de uma empresa. Esse exercício produziu a tabela de frequência abaixo. A coluna Classes representa intervalos de valores de X em reais e a coluna P representa a frequência relativa acumulada. Não existem observações coincidentes com os extremos das classes.

classes	x	P(%)	Fi %	Fi	Fac	
1	70 - 90	5	05	10	10	
2	90 – 110	15	10	20	30	
3	110 – 130	40	25	50	80	
4	130 – 150	70	30			
5	150 – 170	85	15			
6	170 – 190	95	10			
7	190 - 210	100	05			
			100	200		

Entende-se por curtose de uma distribuição seu grau de achatamento em geral medido em relação à distribuição normal. Uma medida de curtose é dada pelo quociente k = Q / (P90-P10), onde Q é a metade da distância interquartílica e P90 e P10 representam os percentis de 90% e 10%, respectivamente. Assinale a opção que dá o valor da curtose k para a distribuição de X.

a) 0,263
b) 0,250
c) 0,300
d) 0,242
e) 0,000

quartil 1 → Posição : iN/4 = 1.200/4= 50→ 3ª classe

Q1 = $l_{inf} + \frac{\frac{iN}{4} - f_{acant}}{f_Q} . h \rightarrow Q_1 = 110 + \frac{(50-30)}{50} . 20 = 118$

Resp: d)

ATIVIDADE

8.1. Classifique quanto à assimetria a distribuição abaixo:
(amostra)

X_i	f_i			
1	2			
2	10			
3	6			
4	4			
5	2			
6	1			

8.2. Classifique quanto à assimetria e quanto à curtose utilizando o índice percentílico:

x_i	f_i			
2	2			
3	4			
4	6			
5	10			
6	6			
7	4			
8	2			

8.3. Uma empresa que monta microcomputadores verificou que o nº de defeitos encontrados por computador segue a tabela:

Nº de defeitos	Nº de computadores			
0	32			
1	28			
2	11			
3	4			
4	3			
5	1			

Pede-se:

a) O nº médio de defeitos por computador
b) A porcentagem de computadores com 2 defeitos
c) A moda
d) O D.M.S
e) A variância
f) O desvio padrão
g) O coeficiente de variação
h) A classificação quanto à assimetria
i) A classificação quanto à curtose utilizando o momento de curtose.

8.4. Um teste de 50 questões foi aplicado em uma turma, tendo-se obtido os seguintes resultados:

Classe	acertos	fi		
1	7⊢12	2		
2	12⊢17	5		
3	17⊢22	13		
4	22⊢27	10		
5	27⊢32	9		
6	32⊢37	6		
7	37⊢42	5		

a) O valor médio dos acertos
b) O polígono de frequência
c) O valor mediano dos acertos
d) A moda
e) O D.M.S
f) A variância
g) O desvio padrão
h) O coeficiente de variação
i) Classificar quanto à assimetria
j) Classificar quanto à curtose

8.5. Os dados abaixo se referem às notas de Estatística do 1º Bi 3º:

8,0 4,5 6,0 9,5 5,0
4,5 6,0 9,5 7,5 9,0
5,5 7,0 5,0 6,0 6,0
6,0 3,5 8,0 2,5 9,0
5,5 8,5 6,0 6,5 6,0
6,0 7,5 8,5 6,0 10,0
9,5 4,0 2,0 9,0 3,5
8,5 6,0 8,5 7,0 4,5
5,0 8,5 9,0 9,0 6,5
10,0

a) Construir o rol e calcular a média aritmética.
b) Construir a distribuição de frequência com h=2, começando de zero
c) Calcular a média aritmética da distribuição de frequência
d) Localizar no rol a mediana e a moda.
e) Calcular a mediana e a moda
f) Calcular a variância e o desvio padrão
g) calcular o coeficiente de variação
h) verificar se a percentagem de dados nos intervalos considerados no item acima coincide com os valores obtidos do rol
i) Qual a percentagem de alunos que tiraram nota igual ou maior do que 7,0
j) Qual a percentagem de alunos que tiraram nota inferior à 4,0

8.6. Qual das séries abaixo apresenta maior dispersão relativa?
a) $X_a = 10$, $\sigma_a = 1$
ou $X_b = 100$, $\sigma_b = 20$

Capítulo 9

9. REGRESSÃO E CORRELAÇÃO.

Quando consideramos observações de duas ou mais variáveis, a relação quantitativa entre elas é chamada correlação. Uma vez caracterizada a relação entre as variáveis, é possível descrevê-la através de uma função matemática. O instrumento utilizado para a determinação dos parâmetros dessa função e a regressão.
Exemplo de relação: pressão e temperatura; uso do cigarro e incidência de câncer; número de evasão e número de matrículas, notas e disciplinas etc.

CORRELAÇÃO

Exemplo: Amostra aleatória das notas em Estatística e Teoria da administração.

Tabela 9.1.

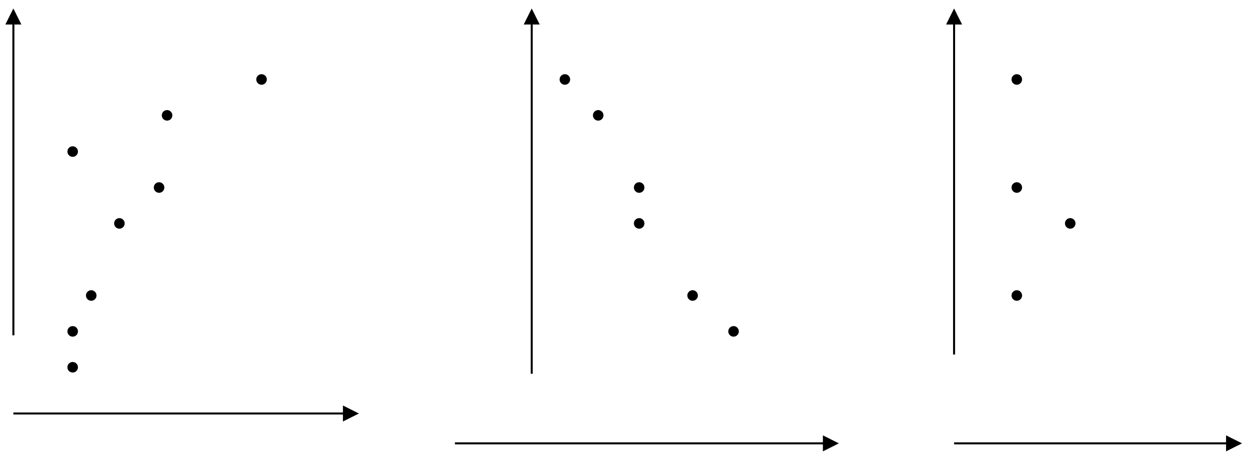
Estatística(x)	Teoria da administração(y)			
5,0	6,0			
8,0	9,0			
7,0	8,0			
10,0	10,0			
6,0	5,0			
7,0	7,0			
9,0	8,0			
3,0	4,0			
8,0	6,0			
2,0	2,0			

Represente, em um sistema cartesiano os pares ordenados (x,y)

Esses pontos obtidos formam o chamado diagrama de dispersão

9.1. CORRELAÇÃO LINEAR

Os pontos obtidos nesse diagrama de dispersão podem ser aproximados a uma reta. Quando isto ocorre, dizemos que houve uma correlação linear. Caso os pontos apresentarem-se muito dispersos, dizemos que não há correlação, evidenciando que não há relação alguma entre as variáveis.



positiva negativa nula

COEFICIENTE DE CORRELAÇÃO LINEAR:

É o instrumento empregado para a medida da correlação linear.

$$R = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{[n\Sigma x_i^2 - (\Sigma x_i)^2][n\Sigma y_i^2 - (\Sigma y_i)^2]}}$$

Coeficiente de Pearson

Se
r = +1, há uma correlação perfeita e positiva entre as variáveis
r = 0 não há correlação, ou a relação não é linear
r = -1, há uma correlação perfeita e negativa entre as variáveis

Exemplos

1- Determinar o coeficiente de Pearson para o exemplo dado anteriormente.

2- Verificar se há uma relação entre o dinheiro gasto com propaganda e as vendas em uma empresa.

Tabela 9.2

Gastos com propaganda (1000s de R$) x	Vendas da empresa (1000s de R$) y	xy	X^2	Y^2
2,4	225	540	5,76	50.625
1,6	184	294,4	2,56	33.856
2,0	220	440,0	4,00	48.400
2,6	240	624,0	6,76	57.600
1,4	180	252,0	1,96	32.400
1,6	184	294,4	2,56	33.856
2,0	186	372,0	4,00	34.596
2,2	215	473,0	4,84	46.225
$\sum x = 15,8$	$\sum y = 1.634$	$\sum xy = 3.289,80$	$\sum x^2 = 32,44$	$\sum y^2 = 337.558$

$$r = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{[n\Sigma x_i^2 - (\Sigma x_i)^2][n\Sigma y_i^2 - (\Sigma y_i)^2]}}$$

$$r = \frac{8.(3.289,80)-(15,8)(1.634)}{\sqrt{8.(32.44)-15,8^2}\sqrt{8.(337.558-1.634^2}} = \frac{501,20}{\sqrt{9,88}.\sqrt{30.508}} = 0,9129$$

Sugere correlação positiva forte. Conforme aumenta o gasto com propaganda, as vendas também aumentam.

Fazer o diagrama de dispersão.

9.1. REGRESSÃO LINEAR

Sempre que desejamos estudar determinada variável em função de outra, fazemos uma análise de regressão.

O objetivo, é descrever através de um modelo matemático a relação entre duas variáveis.

A variável sobre a qual desejamos fazer uma estimativa recebe o nome de variável dependente e a outra recebe o nome de independente.

X – independente

Y – dependente

Ajustando os dados à uma reta: Y= aX + b

Onde $a = \dfrac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$

$a = \frac{8.(3.289{,}8)-(15{,}8).(1634)}{8(32{,}44)-15{,}8^2} = \frac{501{,}2}{9{,}88} = 50{,}72874$

b = $\bar{y}$ - a $\bar{x}$, x é a média dos valores x_i , e y é a média dos valores y_i

$$b = \frac{1.634}{8} - 50{,}72874.\frac{15{,}8}{8} = 104{,}0607$$
$$y = 50{,}73x + 104{,}06$$

BIBLIOGRAFIA

BUSSAB, W. O; MORETTIN, P. A. Estatística básica. 6ª Ed. São Paulo: Saraiva. 2010.

LARSON, Ron; FARBER, Betsy. Estatística aplicada. São Paulo: Prentice Hall Brasil, 2010.

CRESPO, Antônio Arnot. Estatística Fácil. Saraiva, 2009.

TRIOLA, Mario F. Introdução à estatística: Atualização da tecnologia. Rio de Janeiro: LTC, 2013.

NEUFELD, John L. Estatística Aplicada à Administração usando Excell. São Paulo: Prentice Hall, 2003.

LEVINE, D. M; STEPHAN, D. F; KREHBIEL, T. C; BERENSON, M. L. Estatística: Teoria e aplicações. Rio de Janeiro: LTC, 2008.