MATEMÁTICA PARA INTELIGÊNCIA ARTIFICIAL

UNIDADE IV

EXPLORANDO DADOS E UMA APLICAÇÃO ESTATÍSTICA

Elaboração

Erika L P Borges

Juciara do Nascimento César

Produção

Equipe Técnica de Avaliação, Revisão Linguística e Editoração

SUMÁRIO

EXPLORANDO DADOS E UMA APLICAÇÃO ESTATÍSTICA

UNIDADE IV

CAPÍTULO 1

MEDIDAS DE TENDÊNCIA CENTRAL

1.1. Medidas de Tendência Central

Um dos aspectos mais importantes no estudo da distribuição é a posição de um valor central, isto é, um valor representativo sobre o qual as observações estão distribuídas.

Qualquer medida numérica com este objetivo é chamada de medida de tendência central ou de locação. As três mais importantes são a *média aritmética*, *mediana* e *moda*.

1.1.1. Média aritmética

1.1.2. Média aritmética para dados não agrupados

A média aritmética de um conjunto de n elementos é a soma dos n elementos dividida por n.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

onde:

x_i = valor genérico da observação

n = número de observações

Exemplo: Calcular a média aritmética dos seguintes valores: 37, 35, 640, 52, 60 e 40.

$$\overline{x} = \frac{\sum x_i}{n} = \frac{37+35+640+52+60+40}{6} = 144$$

Solução:

1.1.3. Média aritmética para dados agrupados

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{n}$$

onde

x_i = valores agrupados, ou pontos médios de classe, x

$\sum_{i=1}^{n} f_i = n ==$ número total de observações

Exemplo: Determinar a média da distribuição:

Tabela 1. Renda familiar.

Renda Familiar (em salários mínimos)	Nº de Famílias	x_i	$x_i . f_i$
2 |- 4	5	3	15
4 |- 6	10	5	50
6 |- 8	14	7	98
8 |- 10	8	9	72
10 |- 12	3	11	33
Total	40		268

Fonte: autor.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{n} = \frac{268}{40} = 6,7$$

Solução:

Como a renda familiar foi dada em salários mínimos, podemos afirmar que a renda média desse grupo de 40 famílias é de 6,7 salários mínimos.

1.1.4. Mediana

A mediana de um conjunto de n elementos é o valor que ocupa a posição central, quando os dados são ordenados do menor para o maior.

1.1.5. Mediana para dados não agrupados

Se "n" é **ímpar**, existe um único valor que ocupa a posição do meio e este valor é a mediana.

Se "n" é **par**, há dois valores que ocupam a posição central, e a mediana é definida como sendo a média entre esses valores.

A mediana é o valor que divide os dados em duas metades, ou seja, 50% dos dados estão abaixo da mediana e 50% estão acima.

Exemplo: Encontre a mediana dos pesos de recém-nascidos, dados abaixo:

6.4; 9.2; 8.1; 7.8; 10.5

Solução: Primeiro, ordena-se em ordem crescente:

6.4 7.8 8.1 9.2 10.5.

Como n = 5, então a posição do meio é a 3ª posição. Assim, a mediana é o valor que ocupa esta posição, ou seja, 8.1.

Se houvesse seis observações, a mediana seria a média aritmética entre a 4ª e a 3ª posições.

Para "n" ímpar

$$Md = x_{\left(\frac{n+1}{2}\right)^0}$$

Para "n" par

$$Md = \frac{\left[x_{\left(\frac{n}{2}\right)^0} + x_{\left(\frac{n}{2}\right)^0+1}\right]}{2}$$

1.1.6. Mediana para valores tabulados em classe

$$Md = l + h \cdot \left(\frac{E_{Md} - F_{ant}}{f_{Md}}\right)$$

Para os dados agrupados, a mediana pode ser obtida pela fórmula:

l = limite inferior da classe mediana.

h = amplitude do intervalo de classe.

f_{Md} = frequência simples da classe mediana.

F_{ant} = frequência acumulada absoluta da classe anterior à classe mediana.

E_{Md} = elemento mediano.

Exemplo: Determine a mediana para os dados da tabela abaixo.

Tabela 2. Dados fictícios.

Classes	f_i	$F_{ac\downarrow}$
35 |- 45	5	5
45 |-55	12	17
55 |-65	18	35
65 |-75	14	49
75 |-85	6	55
85 |- 95	3	58
Total	58	

Fonte: autor.

Solução: Primeiro, determine o $E_{Md} = n/2$.

Como n = 58, tem-se que $E_{Md} = 58/2 = 29^{o}$ (vigésimo nono elemento).

Depois, identifique a classe mediana pela $F_{ac}\downarrow$.

Neste exemplo, a classe mediana é a 3ª.

$$Md = l + h \cdot \left(\frac{E_{Md} - F_{ant}}{f_{Md}} \right) = 55 + 10 \cdot \left(\frac{29 - 17}{18} \right) = 61,67$$

Em seguida, aplique a fórmula:

1.1.7. Moda

É o valor de maior frequência.

Para distribuições simples (sem agrupamento em classes), a identificação da Moda é facilitada pela simples observação do elemento que apresenta maior frequência.

Assim, para a distribuição:

x_i	243	245	248	251	307
f_i	7	17	23	20	8

A moda será 248. Indica-se Mo = 248.

Exemplo: Calcule a moda do seguinte conjunto de valores:

$$X = \{4, 5, 5, 6, 6, 6, 7, 7, 8, 8\}.$$

Solução: Moda de X: Mo = 6.

Exemplo: Calcule a moda do seguinte conjunto de valores:

$$Y = \{4, 4, 5, 5, 6, 6\}.$$

Solução: Moda de Y: não existe é Conjunto *amodal*.

Exemplo: Calcule a moda do seguinte conjunto de valores:

$$Z = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6\}.$$

Solução: Moda de Z: Mo_1=2 e Mo_2=5; este é um conjunto *bimodal*.

Exemplo: Calcule a moda do seguinte conjunto de valores:

$$W = \{1, 2, 3, 4, 5\}.$$

Solução: Moda de W: não existe é Conjunto *amodal*.

1.1.8. Moda para valores tabulados agrupados

Existem três métodos para a determinação da moda de valores tabulados.

» **Moda bruta**

Consiste em tomar o ponto médio da classe modal. A classe modal é a classe de maior frequência.

» **Método de Czuber**

O Método de Czuber leva em consideração não apenas as frequências das classes adjacentes, mas também a frequência da classe modal.

É obtido através da seguinte expressão:

$$Mo = l + h \cdot \left(\frac{f_{mo} - f_{ant}}{2 \cdot f_{mo} - \left(f_{ant} + f_{post} \right)} \right)$$

l = limite inferior da classe modal.

h = amplitude do intervalo de classe.

f_{ant} = frequência simples da classe adjacente anterior à classe modal.

f_{post} = frequência simples da classe posterior à classe modal.

f_{mo} = frequência simples da classe modal.

» **Método de King**

O método de King baseia-se na influência das frequências das classes adjacentes sobre a classe modal.

Ë calculado pela seguinte expressão:

$$Mo = l + h \cdot \left(\frac{f_{post}}{f_{ant} + f_{post}} \right)$$

l = limite inferior da classe modal.

h = amplitude do intervalo de classe.

f_{ant} = frequência simples da classe adjacente anterior à classe modal.

f_{post} = frequência simples da classe posterior à classe modal.

Exercícios:

1. Dado o seguinte conjunto de dados {7, 8, 6, 10, 5, 9, 4, 12, 7, 8}, encontre a média, a mediana e a moda do conjunto.
2. O número de acidentes ocorridos durante um dado mês em 13 departamentos de manufaturas em um estabelecimento industrial foi: 2, 0, 0, 3, 3, 12, 1, 0, 8, 1, 0, 5, 1. Calcular:
   a. a) média;
   b. b) mediana;
   c. c) a moda para o número de acidentes por departamento.
3. Determine a média, a mediana e a moda dos dados da tabela abaixo.

Rendimento em milhas por galão de automóveis em 25 viagens realizadas por veículos de propriedade de uma companhia.

Tabela 3. Viagens.

Milhas por galão	Número de viagens
15,9	3
17,9	5
19,9	10
21,9	4
23,9	2
25,9	1
Total	**25**

Fonte: Próprio autor.

4) As mortes em acidentes de trânsito são devastadoras para as famílias envolvidas e, em geral, envolvem processos na justiça e pagamentos de altos seguros. Abaixo estão apresentadas uma distribuição de frequência dos motoristas com habilitação, por idade.

Tabela 4. Motoristas com habilitação.

Classe de idade	Motoristas com habilitação
10|--20	5
20|--30	17
30|--40	21
40|--50	27
50|--60	12
60|--70	8
70|--80	6
80|--90	4
Total	**100**

Fonte: Próprio autor.

Determine os valores:

a. média;

b. mediana;

c. moda.

CAPÍTULO 2

MEDIDAS DE DISPERSÃO OU DE VARIABILIDADE

2.1. Medidas de Dispersão ou de Variabilidade

O termo dispersão indica o grau de afastamento de um conjunto de números em relação a sua média, pois, ainda que consideremos a média como um número que tem a faculdade de representar uma série de valores, ela não pode por si mesma destacar o grau de homogeneidade ou heterogeneidade que existe entre os valores que compõem o conjunto.

O nosso objetivo é construir medidas que avaliem a representatividade da média. Para isso, usaremos as medidas de dispersão. As principais são *amplitude total, variância e desvio-padrão*.

2.2. Amplitude Total

Esta medida nos dá uma ideia do campo de variação dos elementos da distribuição. É a diferença entre o maior e o menor valor da sequência.

$$At = X_{(max)} - X_{(min)}$$

Exemplo: Considere os seguintes dados brutos:

X: 11, 12, 9, 10, 10, 15

Solução:

$$At = 15 - 9 = 6.$$

A amplitude total é uma medida que tem pouca sensibilidade estatística, uma vez que ela leva em consideração apenas os valores extremos da série.

Exercícios:

1. Calcule a amplitude total dos conjuntos de dados:

   a) 1, 3, 5, 9

   b) 20, 14, 15, 19, 21, 22, 20

   c) 17,9; 22,5; 13,3; 16,8; 15,4; 14,2

   d) -10, -6, 2, 3, 7, 9, 10

2. Considerando as séries do item b e c do exercício anterior, qual delas apresenta maior dispersão?

Para dados dispostos em tabelas de frequência, com valores agrupados em intervalo de classe, a amplitude total é definida da seguinte forma:

At = (limite superior da última classe – limite inferior da primeira classe)

A amplitude total é útil em casos como a medida de temperatura de localidade, em que se pode estabelecer a amplitude da temperatura em um dia, semana, mês ou ano.

Para a programação de uma viagem, é importante conhecer a amplitude da região durante os últimos dias, semanas, meses ou anos.

2.3. Variância e Desvio Padrão

2.3.1. Variância

Os dados distribuem-se em torno da média. Então, o grau de dispersão de um conjunto de dados pode ser medido pelos desvios dos valores observados em relação à média.

Entende-se por desvio em relação à média a diferença entre o valor observado e a média do conjunto de dados.

Por exemplo, se a média de estatura dos jogadores de um time de futebol for $\bar{x}$ =1,72 m, o jogador que tiver estatura x_i = 1,82 m terá um desvio em relação à média de:

$$x_i - \bar{x} = 1{,}82 - 1{,}72 = 0{,}10 \text{ m}$$

Os desvios em relação à média medem a dispersão. É preciso considerar, no entanto, que cada dado tem um desvio em relação à média.

Então, para julgar o grau de dispersão de todo o conjunto de dados com base nos desvios, seria preciso observar todos os desvios.

Não se pode usar a soma dos desvios, pois ela é necessariamente igual a zero, como se observa no exemplo abaixo:

Tabela 5. Desvios.

Dados (x_i)	Desvios ($x_i - \bar{x}$)
2	2 - 7 = -5
6	6 – 7 = -1
8	8 – 7 = 1
9	9 – 7 = 2
10	10 – 7 = 3
$\bar{x} = 7$	$\Sigma (x_i - \bar{x}) = 0$

Fonte: Próprio autor.

Este problema de achar um valor único para representar os desvios fica resolvido se, em lugar da soma dos desvios, for usada a *soma dos quadrados dos desvios*.

Como os quadrados de números negativos são positivos, toda soma de quadrados é positiva ou, no mínimo, nula.

O procedimento para obter a soma de quadrados dos desvios é mostrado na tabela abaixo.

Tabela 6. Desvios ao quadrado.

dados (x_i)	desvios ($x_i - \bar{x}$)	Quadrados dos desvios $(x_i - \bar{x})^2$
2	-5	25
6	-1	1
8	1	1
9	2	4
10	3	9
$\bar{x} = 7$	$\Sigma (x_i - \bar{x}) = 0$	$\Sigma (x_i - \bar{x})^2 = 40$

Fonte: Próprio autor.

Para medir a dispersão dos dados em torno da média, usa-se a variância, que pode ser definida como a soma dos quadrados dos desvios

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

dividida pelo número de dados, isto é, por n.

Notação: σ^2 – variância populacional

Quando se trabalha com amostras, é mais correto definir a variância como a soma dos quadrados dos desvios, dividida pelo número de graus de liberdade da amostra, que é n-1.

Então, a variância de uma amostra, que é indicada por $\mathbf{s^2}$, é dada pela fórmula:

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

Notação: $\mathbf{s^2}$ – variância amostral.

Para mostrar como se calcula a variância, são dados os valores 1, 2, 3, 4 e 5.

Os cálculos intermediários para a obtenção da variância estão apresentados abaixo.

Tabela 7. Quadrado dos desvios.

Dados (x_i)	Desvios ($x_i - \bar{x}$)	Quadrados dos desvios $(x_i - \bar{x})^2$
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4
$\bar{x} = 3$	$\Sigma (x_i - \bar{x}) = 0$	$\Sigma (x_i - \bar{x})^2 = 10$

Fonte: Próprio autor.

Aplicando a fórmula, tem-se:

$$S^2 = \frac{10}{4} = 2{,}5$$

É importante deixar claro que variância mede dispersão.

Para entender esta afirmativa, imagine que quatro alunos fizeram quatro provas e foram todos aprovados com média igual a 7.

As notas de quatro alunos estão apresentadas na tabela abaixo:

Tabela 8. Notas de quatro alunos.

Alunos	Notas				Média
Antônio	7	7	7	7	7
Carlo	6	8	6	8	7
João	4	6	8	10	7
Pedro	0	8	10	10	7

Fonte: Próprio autor.

Observe as notas apresentadas na tabela acima. É fácil ver que:

a. as notas de Antônio não têm dispersão;

b. as notas de Carlos têm dispersão pequena;

c. as notas de João têm dispersão maior do que as de Carlos;

d. as notas de Pedro têm a maior dispersão pequena.

Essas observações são confirmadas pelas variâncias dessas notas, apresentadas na tabela abaixo.

Tabela 9. Média e variância das notas.

Alunos	Média	Variância
Antônio	7	0
Carlo	7	1,33
João	7	6,67
Pedro	7	22,67

Fonte: Próprio autor.

2.3.2. Variância para dados agrupados

Se os dados estão apresentados em uma tabela de distribuição de frequência, para obter a variância aplica-se a fórmula:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot f_i}{n - 1}$$

Onde x_i são os valores dos dados, f_i são as frequências e $n = \Sigma f_i$, ou seja, o número total de observações.

Os dados apresentados na Tabela 10 serão utilizados para mostrar como se calcula a variância. Os cálculos intermediários estão na Tabela 11.

Tabela 10. Distribuição de 40 alunos segundo as notas.

Notas	Número de alunos
0	15
1	10
2	5
3	5
4	1
5	1
6	0
7	3

Fonte: Próprio autor.

Tabela 11. Cálculos intermediários para obtenção da variância.

Notas	Número de alunos	$x_i f$	$(x_i - \overline{x})^2 f_i$
0	15	0	$(0 - 1{,}62)^2 \cdot 15 = 39{,}3$
1	10	10	$(1 - 1{,}62)^2 \cdot 10 = 3{,}8$
2	5	10	$(2 - 1{,}62)^2 \cdot 5 = 0{,}7$
3	5	15	$(3 - 1{,}62)^2 \cdot 5 = 9{,}5$
4	1	4	$(4 - 1{,}62)^2 \cdot 1 = 5{,}6$
5	1	5	$(5 - 1{,}62)^2 \cdot 1 = 11{,}4$
6	0	0	$(6 - 1{,}62)^2 \cdot 0 = 0$
7	3	21	$(7 - 1{,}62)^2 \cdot 3 = 86{,}8$
	40	**65**	**157,1**

Fonte: Próprio autor.

$\overline{x} = 1{,}62$

Aplicando a fórmula:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot f_i}{n-1} = \frac{157{,}1}{39} = 4{,}02$$

2.3.3. Desvio-padrão

Como medida de dispersão, a variância apresenta a desvantagem de ter unidade de medida igual ao quadrado da unidade de medida dos dados. Para entender essa ideia, imagine que um aluno fez três provas, com 60 questões cada uma.

Na primeira prova acertou 40 questões, na segunda acertou 45 e na terceira acertou 50. A média de acertos, nas três provas, é:

$$\overline{x} = \frac{40 + 45 + 50}{3} = 45$$

ou seja, o aluno acertou, em média, 45 questões.

A variância é:

$$s^2 = \frac{(40-45)^2 + (45-45)^2 + (50-45)^2}{3-1} = \frac{50}{2} = 25$$

ou seja, 25 questões (porque os desvios são elevados ao quadrado).

Evidentemente, "questões ao quadrado" não têm qualquer sentido prático.

Então, é preciso definir uma medida de dispersão que seja a raiz quadrada da variância.

Por definição, desvio-padrão é a raiz quadrada, com sinal positivo, da variância. O desvio-padrão é representado por s.

Para o exemplo apresentado, tem-se que:

$$s = \sqrt{25} = 5$$

Ou seja, o desvio-padrão é 5 questões.

O desvio-padrão, como a variância, mede a dispersão dos dados, mas tem a vantagem de usar a mesma unidade de medida dos dedos.

Exemplo: Dados das idades de 50 funcionários da empresa XPTO. Vamos determinar a variância, o desvio-padrão.

Solução:

Tabela 12. Cálculo da variância.

Classe de idade	fi	xi	xi*fi	$(x_i - \bar{x})^2 \cdot f_i$
18|--25	6	21,5	129	$(21,5 - 38,44)^2 * 6 = 1721,78$
25|--32	10	28,5	285	$(28,5 - 38,44)^2 * 10 = 988,03$
32|--39	13	35,5	461,5	$(35,5 - 38,44)^2 * 13 = 112,36$
39|--46	8	42,5	340	$(42,5 - 38,44)^2 * 8 = 131,86$
46|--53	6	49,5	297	$(49,5 - 38,44)^2 * 6 = 733,94$
53|--60	5	56,5	282,5	$(56,5 - 38,44)^2 * 5 = 1630,81$
60|--67	2	63,5	127	$(63,5 - 38,44)^2 * 2 = 1256,00$
Total	**50**		**1922**	**6.574,78**

Fonte: Próprio autor.

Logo, a média amostral será:

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{1922}{50} = 38,44\, anos.$$

Quanto à variância, tem-se que:

$$S^2 = \frac{\sum (x_i - \bar{x})^2 \cdot f_i}{n-1} = \frac{6.574,78}{50-1} = \frac{6.574,78}{49} = 134,18.$$

O desvio-padrão é:

$$S = \sqrt{S^2} = \sqrt{134,18} = 11,58 \text{ anos.}$$

O desvio-padrão goza de algumas propriedades, dentre as quais destacamos:

1ª) Somando-se (ou subtraindo-se) uma constante a todos os valores de uma variável, o desvio-padrão não se altera:

$$y_i = x_i \pm c \Rightarrow s_y = s_x$$

2ª) Multiplicando-se (ou dividindo-se) todos os valores de uma variável por uma constante (diferente de zero), o desvio-padrão fica multiplicado por uma constante:

$$y_i = x_i \cdot c \Rightarrow s_y = c \cdot s_x$$

Exemplo:

$$\bar{x} \pm S = 38,44 \pm 11,58 = (26,86; 50,02)$$

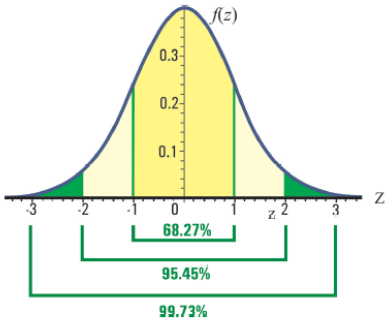
Conclui-se que entre 27 e 50 anos temos 66% das observações. Isto é: o intervalo compreendido entre a média menos um desvio-padrão e a média mais um desvio-padrão contém, nesse exemplo, 66% das 50 idades.

Exemplo:

$$\bar{x} \pm 2S = 38,44 \pm 2 \cdot (11,58) = (15,28; 61,60)$$

Resulta que entre 16 e 62 anos temos 98% das observações. Isto é: o intervalo compreendido entre a média menos duas vezes desvio-padrão e a média mais duas vezes desvio-padrão contém, nesse exemplo, 98% das 50 idades.

Figura 27. Probabilidades quando uma distribuição é normal



Fonte: Crespo, 2004.

2.4. Coeficiente de variação

Uma medida de dispersão relativa resulta da comparação entre uma medida de variabilidade absoluta e uma medida de tendência central.

Coeficiente de variação de Pearson: é a relação entre o desvio-padrão e a média aritmética, isto é:

$$CV = \frac{s}{\overline{x}} \times 100$$

Exemplo:

Pense em uma sala de aula com 50 alunos possuidores de uma estatura média de 1,70 m, com desvio-padrão igual a 0,063m e um peso médio de 53kg com desvio-padrão igual a 6,0kg. Qual a maior variabilidade relativa, a dos pesos ou a das alturas?

Para a variável altura:

$$\overline{x} = 1,70m$$
$$s = 0,063m$$

Para a variável peso:

$$\overline{x} = 53kg$$
$$s = 6,0kg$$

O coeficiente de variação para a altura é:

$$CV = \frac{0,063}{1,70} \cdot 100 = 3,7\%$$

e para o peso é:

$$CV = \frac{6,0}{53} \cdot 100 = 11,3\%$$

Diz-se que a distribuição possui pequena variabilidade (dispersão) quando o coeficiente der até 10%; média dispersão quando estiver acima de 10% até 20%; e grande dispersão quando superior a 20%.

Exercícios:

1. Dada a amostra: 2, 3, 4, 5, 7, 10, 12.
   a. Qual é a amplitude total?
   b. Calcule a variância.
   c. Determine o desvio-padrão.
2. Calcule a variância amostral e populacional:

Tabela 13. Dados de uma população.

Classes	Fi
2|--4	3
4|--6	5
6|--8	8
8|--10	6
10|--12	3

Fonte: Próprio autor.

3. A seguir, temos a distribuição de frequência dos pesos de uma amostra de 45 alunos:

Tabela 14. Peso de 45 alunos.

Peso em Kg	F_i
40|--45	4
45|--50	10
50|--55	15
55|--60	8
60|--65	5
65|--70	3

Fonte: Próprio autor.

a. Determine o peso médio.

b. Determine a variância.

c. Qual é o valor do coeficiente de variação?

4. Cronometrando o tempo para várias provas de uma incana automobilística, encontramos:

Equipe1:	40 provas Tempo médio: 40 segundos Variância: 400 segundos ao quadrado
Equipe2:	Tempo: 20 40 50 80 Nº de provas: 10 15 30 5

a. Qual o coeficiente de variação relativo à equipe 1?

b. Qual a média da equipe 2?

c. Qual o desvio-padrão relativo à equipe 2?

d. Qual a média aritmética referente às duas equipes consideradas em conjunto?

e. Qual a equipe que apresentou resultados mais homogêneos? Justifique.

f. Qual a equipe que apresentou menor dispersão relativa?

CAPÍTULO 3

CORRELAÇÃO E REGRESSÃO LINEAR SIMPLES

3.1. Correlação

O termo “correlação” significa até que ponto duas variáveis estão relacionadas entre si. Essa correlação ou associação pode ser verificada através do diagrama de dispersão e medida através do coeficiente de correlação.

3.1.1. Diagrama de dispersão

Dispositivo gráfico utilizado para verificar o grau de associação entre duas variáveis quantitativas. Basta traçar um sistema de eixos cartesianos, onde fazemos corresponder um ponto para cada par de valores (x, y).

Exemplo:

Verificar, através do diagrama de dispersão, a correlação ou associação existente entre as variáveis abaixo:

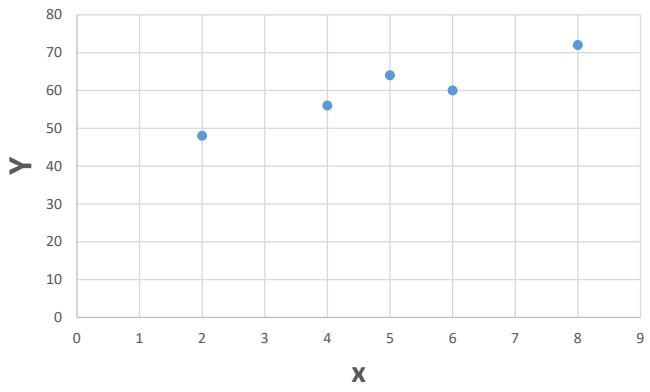
Número de anos de atendimento ao público (x) e número de clientes (y) de 5 agentes de seguro:

Tabela 15. Dados de um agente de seguro.

Agentes	A	B	C	D	E
x	2	4	5	6	8
y	48	56	64	60	72

Fonte: Próprio autor.

Gráfico 1. Dispersão para os dados acima.



Fonte: Próprio autor.

Solução: Quando a "nuvem" de pontos se apresenta como no diagrama abaixo, dizemos que existe uma **Correlação Linear Direta** entre as variáveis x e y. Além disso, quanto mais próxima estiver de uma reta, mais **forte** será essa correlação.

Exemplo:

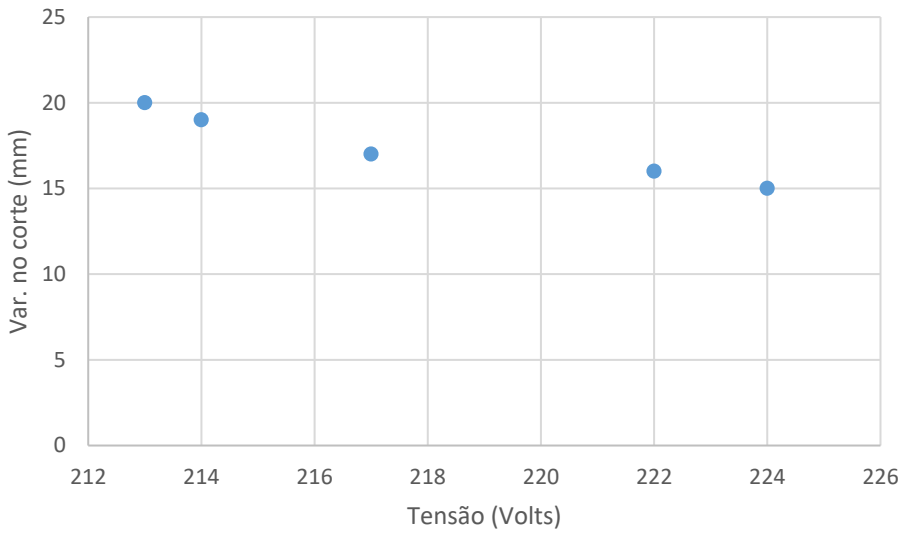
Desconfia-se que a tensão da rede elétrica (volts) esteja influenciando na máquina que corta o material que fabrica gavetas para geladeiras, provocando uma certa variação no corte (mm). Para verificar se existe realmente ligação entre as duas variáveis acima, foram considerados os seguintes valores:

Tabela 16. Tensão da rede elétrica.

Tensão (volts)	222	214	224	217	213
Var. no corte (mm)	16	19	15	17	20

Fonte: Próprio autor.

Gráfico 2. Dispersão para os dados da rede elétrica.



Fonte: Próprio autor.

Solução: Quando a "nuvem" de pontos se apresenta como no diagrama abaixo, dizemos que existe uma **Correlação Linear Inversa** entre as duas variáveis x e y.

Exemplo:

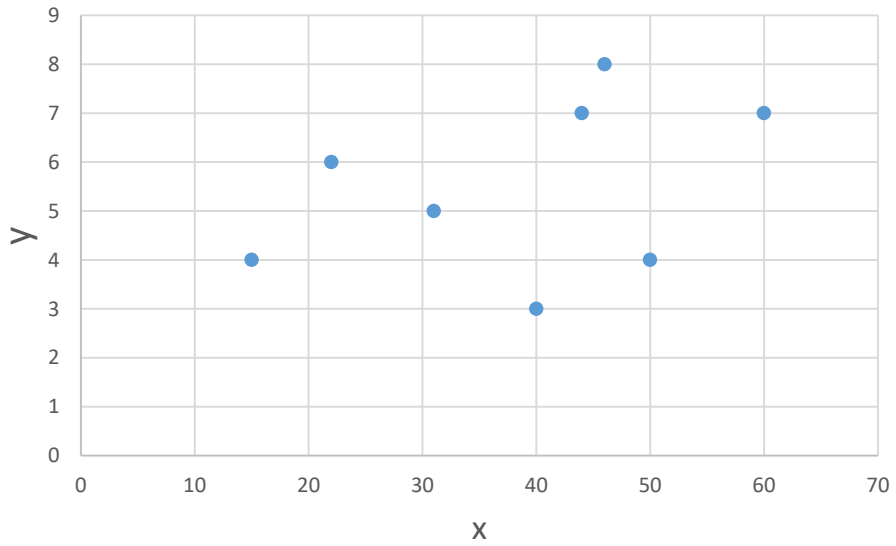
Resultado de um teste de inglês (x) e tempo gasto em dias (y) por um aluno de Administração para aprender a operar uma determinada máquina:

Tabela 17. Teste de inglês e tempo.

x	15	22	31	40	44	46	50	60
y	4	6	5	3	7	8	4	7

Fonte: Próprio autor.

Gráfico 3. Dispersão para teste de inglês e tempo.



Fonte: Próprio autor.

Solução: Quando a "nuvem" de pontos se apresenta como no diagrama abaixo, dizemos que **não existe correlação entre as duas variáveis x e y** ou, se existir, é irrelevante.

3.1.2. Coeficiente de Correlação (Pearson) – $r_{(x, y)}$

Medida que quantifica e julga o quanto a "nuvem" de pontos do diagrama de dispersão se aproxima de uma reta.

Quanto mais se aproxima de ± 1 (não podendo ultrapassar), mais se aproxima de uma reta, ou seja, maior é a correlação ou dependência entre as variáveis.

A correlação é dita perfeita e direta ou positiva quando r (x, y) =1; é perfeita e inversa ou negativa quando r (x, y) = -1.

Além disso, quanto mais o valor se aproxima de ZERO, menor é a correlação, tornando-se inexistente quando r (x, y) = 0, o que significa que as variáveis são independentes.

Expressão:

$$r = \frac{n\sum(X \cdot Y) - \left(\sum X\right) \cdot \left(\sum Y\right)}{\sqrt{n \cdot \sum X^2 - \left(\sum X\right)^2} \cdot \sqrt{n \cdot \sum Y^2 - \left(\sum Y\right)^2}}$$

Para obter os somatórios da expressão acima, procede-se da seguinte maneira.

Σ(XY): fazem-se os produtos x.y, referentes a cada par de observações e depois, efetua-se a soma;

ΣX: somam-se os valores da variável X;

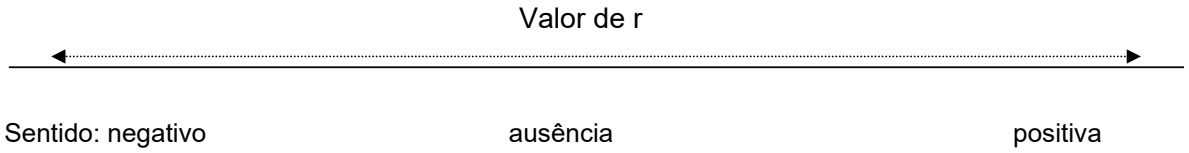
ΣY: somam-se os valores da variável Y;

ΣX²: eleva-se ao quadrado cada valor de X e, depois, efetua-se a soma;

ΣY²: eleva-se ao quadrado cada valor de Y e, depois, efetua-se a soma.

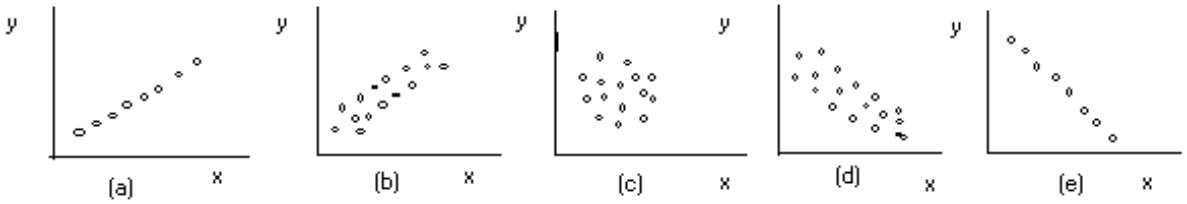
É importante saber que o coeficiente de correlação assume valores entre –1 e 1 ($-1 \leq r \leq 1$)

Figura 28. Sentido da correlação em função do valor de r.



Agora, relacione as letras de cada diagrama às letras abaixo dos seguintes gráficos:

Figura 29. Configuração dos dados de dispersão.



Fonte: Próprio autor.

Quadro 4. Classificação do desempenho dos dados.

a) correlação linear perfeita positiva $r \approx 1$	b) correlação linear moderada positiva $0 \leq r \leq 1; r \approx 0,5.$	c) ausência de correlação linear $r = 0$	d) correlação linear moderada negativa $-1 \leq r \leq 0; r \approx -0,5.$	e) correlação linear perfeita negativa $r \approx -1$

Fonte: Próprio autor.

Exercício: Calcule o coeficiente de correlação das variáveis do exemplo anterior e analise.

Solução:

N=5.

Σ(XY): 1576.

ΣX: 25.

ΣY: 300.

ΣX²: 145.

ΣY²: 18320.

$$r=\frac{5*1576-25*300}{\sqrt{5*145-625}*\sqrt{5*18320-900}}$$

$$r=\frac{7880-7500}{\sqrt{100}*\sqrt{1600}}$$

$$r=\frac{380}{400}=0,95$$

Por este resultado, e observando o quadro 4, entende-se que a correlação é linear e positiva.

3.1.3. Coeficiente de Determinação – $CD_{(x, y)}$

Diz até que ponto a variação de y é explicada pela variação de x. É obtido utilizando-se a seguinte expressão:

$$CD_{(x, y)} = r^2$$

Exemplo: Calcule o coeficiente de determinação das variáveis do exemplo anterior e analise.

Solução: o CD = $(0,95)^2$ = 0,9025.

Isso nos diz que a variável y é explicada pela x em 90,25%.

3.1.4. Regressão

Se existir correlação linear entre X e Y, podemos dizer que $Y=a + b \cdot X$,

Onde:

X - variável independente.

Y - variável dependente.

O objetivo é estimarmos os valores de a e b, através dos valores X e Y, para obtermos a *Reta de Regressão*.

3.1.5. Estimativas dos parâmetros "a" e "b"

Descreve, através de um modelo matemático, a relação existente entre duas variáveis quantitativas, a partir de n observações dessas variáveis.

Seu principal objetivo é estimar valores com base em dados passados, através de um ajustamento pelo método dos mínimos quadrados.

O ajustamento é feito através de uma Reta, de uma Parábola ou de uma Exponencial, conforme se dispuserem os dados no diagrama de dispersão.

$$b = \frac{n \cdot \sum(X \cdot Y) - \left(\sum X\right) \cdot \left(\sum Y\right)}{n \cdot \sum X^2 - \left(\sum X\right)^2}$$

$$a = \frac{\sum(Y) - b \cdot \sum X}{n}$$

$$SXY = n \cdot \sum(X \cdot Y) - \left(\sum X\right) \cdot \left(\sum Y\right)$$

$$SXX = n \cdot \sum X^2 - \left(\sum X\right)^2$$

$$SYY = n \cdot \sum Y^2 - \left(\sum Y\right)^2$$

Onde:

n: número de pares (x,y) observados (tamanho da amostra);

Σ(XY): fazem-se os produtos x.y, referentes a cada par de observações e, depois, efetua-se a soma;

ΣX: somam-se os valores da variável X;

ΣY: somam-se os valores da variável Y;

ΣX^2: eleva-se ao quadrado cada valor de X e, depois, efetua-se a soma;

ΣY^2: eleva-se ao quadrado cada valor de Y e, depois, efetua-se a soma.

Exercício:

Calcule os parâmetros, obtenha a equação de regressão das variáveis do exemplo anterior e depois faça estimativas.

Resolução para o primeiro conjunto de dados:

N = 5.

Σ(XY): 1576.

ΣX: 25.

ΣY: 300.

ΣX²: 145.

ΣY²: 18320.

$$b = \frac{5*1576 - 25*300}{5*145 - 625}$$

$$b = \frac{380}{100} = 3,8$$

$$a = \frac{300 - 3,8*25}{5}$$

$$a = \frac{300 - 3,8*25}{5} = 42,5$$

Reta de regressão ou modelo

$$y = 42,5 + 3,5*X$$

Exercício:

1. A tabela abaixo mostra o consumo de combustível e o número de quilômetros rodados de um carro.

Tabela 18. Dados de consumo por quilômetros.

Y	Consumo (litros)	1	2	3	4	5	6
X	Km rodados	6	13	18	25	33	40

Fonte: Próprio autor.

a. Faça o diagrama de dispersão.

b. Calcule o coeficiente de correlação. Caso exista uma correlação significante, encontre a reta de regressão y = a + b·x e estime o consumo (em litros) para 100 quilômetros rodados.

c. Se o carro só dispõe de 8 litros de combustível, quantos quilômetros podem ser rodados?

d. Calcule o coeficiente de determinação.

2. A tabela abaixo apresenta dados referentes ao número de horas de estudo fora de sala de aula para determinados alunos, no período de uma semana, bem como as notas obtidas na avaliação.

Tabela 19. Horas de estudo por notas.

Horas de estudo (x)	3 4 5 6 7 8
Notas obtidas (y)	6 7 6 7 8 9

Fonte: Próprio autor.

a. Faça o diagrama de dispersão.

b. Calcule o coeficiente de correlação linear.

c. Encontre a reta de regressão ($\hat{y} = a + bx$).

d. Estime a nota para um aluno que venha a estudar **3,5** horas no período.

e. Calcule o coeficiente de determinação.

CAPÍTULO 4

NORMALIZAÇÃO DOS DADOS

4.1. Normalização dos dados

A normalização de dados é um processo que usa transformações e algumas estatísticas para modificar um conjunto de dados a fim de que possuam alguma característica desejável na hora de processá-los. Existem diversos tipos.

Qual vamos usar depende do objetivo final e de como os dados se encontram. Por exemplo, um conjunto de transformações que é eficiente para a análise de componentes principais pode não ser para uma regressão.

Exemplo:

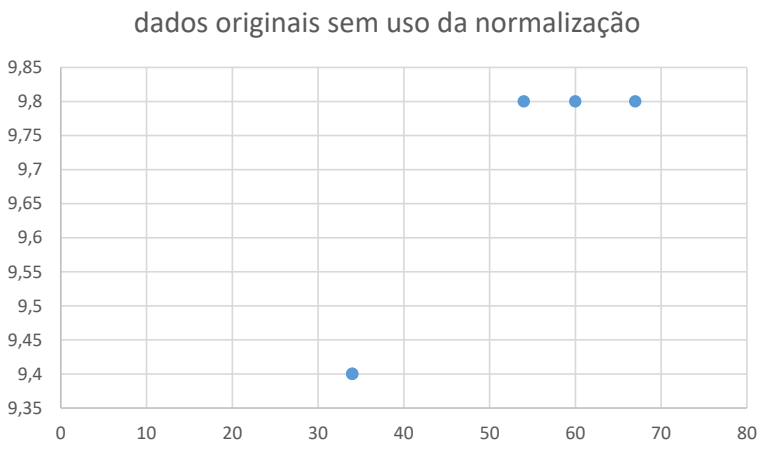
Para fazer diferentes tipos de normalizações, selecionamos uma amostra de cinco observações de dois ingredientes que compõem o vinho, total de dióxido de enxofre e teor alcoólico (ANEXO I).

Tabela 20. Medida de dois componentes em vinhos.

total.sulfur.dioxide	álcool
34	9,4
67	9,8
54	9,8
60	9,8
34	9,4

Fonte: anexo I.

Gráfico 4. Dispersão de dados do componente químicos da Tabela 20.



Fonte: Próprio autor.

4.1.1. Reescalonamento

O reescalonamento também é chamado de escalonamento min-máx e consiste em modelar o conjunto de dados para que eles fiquem contidos no intervalo $[0,1]$ ou $[-1,1]$ dependendo da natureza dos elementos do conjunto original. A transformação é:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Em que x' são os novos elementos do conjunto de dados transformados. Essa transformação leva:

Exemplo:

$$x' = \begin{Bmatrix} 1 \\ 4 \\ 8 \end{Bmatrix} \Rightarrow \begin{Bmatrix} \frac{1-1}{8-1} \\ \frac{4-1}{8-1} \\ \frac{8-1}{8-1} \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0,42 \\ 1 \end{Bmatrix}$$

O novo vetor:

$$x' = \begin{bmatrix} 0 & 0,42 & 1 \end{bmatrix}$$

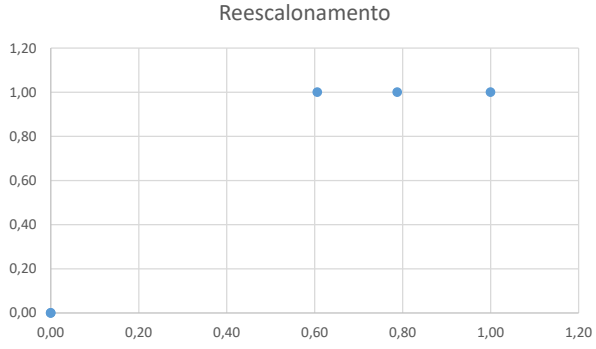
Exemplo dos ingredientes do vinho:

Tabela 21. componentes do vinho.

total.sulfur.dioxide	alcool
0,00	0
1,00	1
0,61	1
0,79	1
0,00	0

Fonte: autor

Gráfico 5. Dispersão dos componentes químicos da Tabela 21.



Fonte: autor.

Exemplo:

$$\begin{pmatrix} 0.22192759 \\ 0.71370026 \\ 0.10149519 \\ 0.93973468 \\ 0.57237254 \\ 0.30804023 \\ 0.17388386 \\ 0.68692274 \\ 0.69140697 \\ 0.87739222 \\ 0.8134652 \\ 0.44172751 \\ 0.23980563 \\ 0.62331624 \\ 0.44377312 \end{pmatrix} \rightarrow \begin{pmatrix} 0.14367302 \\ 0.73034625 \\ 0. \\ 1. \\ 0.56174561 \\ 0.24640338 \\ 0.08635798 \\ 0.6984013 \\ 0.70375088 \\ 0.92562691 \\ 0.84936348 \\ 0.40588916 \\ 0.1650011 \\ 0.62252024 \\ 0.40832952 \end{pmatrix}$$

4.1.2. Normalização pela média

A normalização pela média consiste em subtrair de cada elemento do conjunto a média.

Esse procedimento faz com que a média do novo conjunto seja nula. Com a divisão pela amplitude, os elementos do novo conjunto estarão contidos entre $[-1,1]$.

Podemos calcular os elementos do novo conjunto fazendo:

$$x' = \frac{x - \bar{x}}{\max(x) - \min(x)}$$

Em que x' são os elementos do novo conjunto de dados transformados.

Essa transformação leva:

$$\bar{x} = \frac{1+4+8}{3} = 4,33$$

$$x' = \begin{Bmatrix} 1 \\ 4 \\ 8 \end{Bmatrix} \Rightarrow \begin{Bmatrix} \frac{1-4,33}{8-1} \\ \frac{4-4,33}{8-1} \\ \frac{8-4,33}{8-1} \end{Bmatrix} = \begin{Bmatrix} -0,47 \\ -0,04 \\ 0,52 \end{Bmatrix}$$

$$x' = \begin{bmatrix} -0,47 & -0,04 & 0,52 \end{bmatrix}$$

Outro exemplo:

$$\begin{pmatrix} 0.22192759 \\ 0.71370026 \\ 0.10149519 \\ 0.93973468 \\ 0.57237254 \\ 0.30804023 \\ 0.17388386 \\ 0.68692274 \\ 0.69140697 \\ 0.87739222 \\ 0.8134652 \\ 0.44172751 \\ 0.23980563 \\ 0.62331624 \\ 0.44377312 \end{pmatrix} \rightarrow \begin{pmatrix} -0.35948757 \\ 0.22718566 \\ -0.50316059 \\ 0.49683941 \\ 0.05858502 \\ -0.25675721 \\ -0.41680261 \\ 0.19524071 \\ 0.20059029 \\ 0.42246632 \\ 0.34620289 \\ -0.09727143 \\ -0.33815949 \\ 0.11935965 \\ -0.09483107 \end{pmatrix}$$

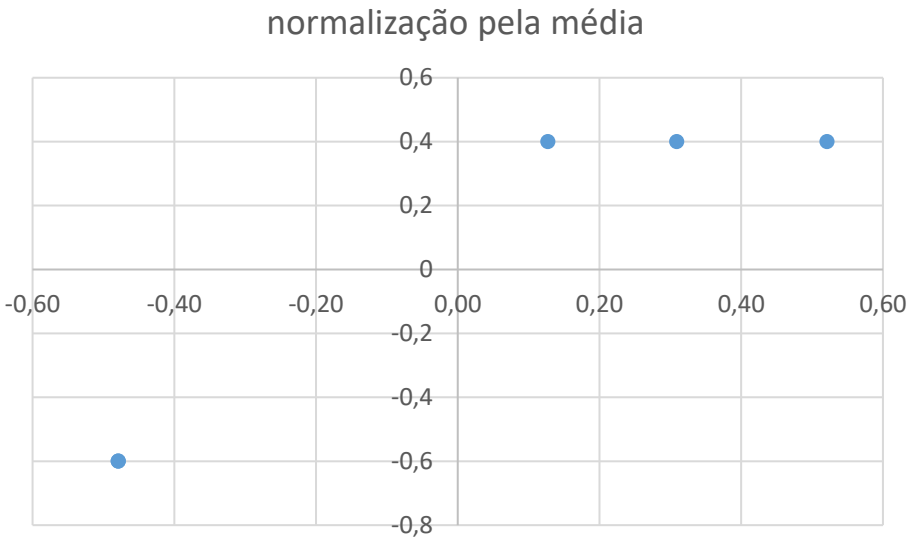
Exemplo:

Tabela 22. Componentes do vinho.

total.sulfur.dioxide	álcool
-0,48	-0,6
0,52	0,4
0,13	0,4
0,31	0,4
-0,48	-0,6

Fonte: autor.

Gráfico 6. Dispersão dos componentes químicos da Tabela 22.



Fonte: autor.

4.1.3. Padronização dos dados

Na padronização de dados, transformamos os dados a fim de fazer com que a média do novo conjunto de dados seja nula e o novo desvio seja um.

Os novos elementos são dados por:

$$x' = \frac{x - \overline{x}}{\sigma_x}$$

Os elementos se tornam:

Média:

$$\overline{x} = \frac{1+4+8}{3} = 4,33$$

$$\overline{x} = \frac{(1-4,33)^2 + (4-4,33)^2 + (8-4,33)^2}{3} =$$

Variância:

$$\delta^2 = \frac{11,08 + 0,108 + 13,46}{3} = 8,22$$

Desvio-padrão:

$$\delta = \sqrt{8,22} = 2,86$$

$$x' = \begin{Bmatrix} 1 \\ 4 \\ 8 \end{Bmatrix} \Rightarrow \begin{Bmatrix} \frac{1-4,33}{2,86} \\ \frac{4-4,33}{2,86} \\ \frac{8-4,33}{2,86} \end{Bmatrix} = \begin{Bmatrix} -1,164 \\ -0,115 \\ 1,28 \end{Bmatrix}$$

$$x' = \begin{bmatrix} -1,164 & -0,115 & 1,28 \end{bmatrix}$$

$$\begin{pmatrix} 0.22192759 \\ 0.71370026 \\ 0.10149519 \\ 0.93973468 \\ 0.57237254 \\ 0.30804023 \\ 0.17388386 \\ 0.68692274 \\ 0.69140697 \\ 0.87739222 \\ 0.8134652 \\ 0.44172751 \\ 0.23980563 \\ 0.62331624 \\ 0.44377312 \end{pmatrix} \rightarrow \begin{pmatrix} -1.15715614 \\ 0.73128895 \\ -1.61962587 \\ 1.5992786 \\ 0.18857957 \\ -0.82647694 \\ -1.34164778 \\ 0.6284612 \\ 0.64568099 \\ 1.35987873 \\ 1.11439403 \\ -0.31310745 \\ -1.08850308 \\ 0.38420732 \\ -0.30525215 \end{pmatrix}$$

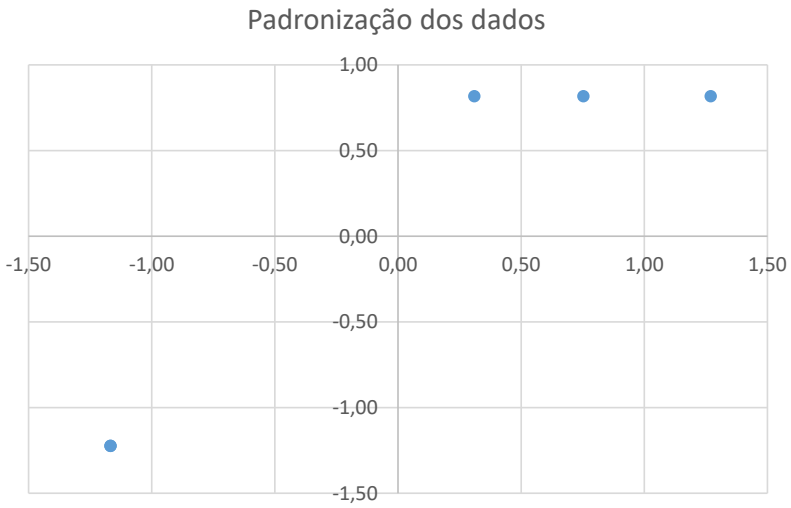
Exemplo dos ingredientes do vinho:

Tabela 23. Componentes químicos do vinho.

total.sulfur.dioxide	alcool
-1,17	-1,22
1,27	0,82
0,31	0,82
0,75	0,82
-1,17	-1,22

Fonte: autor.

Gráfico 7. Dispersão dos componentes químicos da Tabela 23.



Fonte: autor.

4.1.4. Reescalonamento para norma 1

Essa transformação se aplica a vetores e consiste em fazer com que o novo vetor tenha a mesma direção e sentido, porém módulo unitário. Os novos vetores são dados por:

$$\vec{x}' = \frac{\vec{x}}{|x|}$$

$$\begin{pmatrix} 0.22192759 & 0.9662181 \\ 0.71370026 & 0.30436127 \\ 0.10149519 & 0.41738617 \\ 0.93973468 & 0.3327575 \\ 0.57237254 & 0.04126955 \\ 0.30804023 & 0.93012659 \\ 0.17388386 & 0.04061124 \\ 0.68692274 & 0.85676488 \\ 0.69140697 & 0.49738147 \\ 0.87739222 & 0.64377773 \\ 0.8134652 & 0.79743178 \\ 0.44172751 & 0.12928483 \\ 0.23980563 & 0.15481691 \\ 0.62331624 & 0.58470379 \\ 0.44377312 & 0.0524205 \end{pmatrix} \rightarrow \begin{pmatrix} 0.223857 & 0.97462 \\ 0.919848 & 0.39227 \\ 0.919848 & 0.39227 \\ 0.942647 & 0.33378 \\ 0.997410 & 0.07191 \\ 0.314388 & 0.94929 \\ 0.973793 & 0.22743 \\ 0.973793 & 0.22743 \\ 0.811774 & 0.58397 \\ 0.806248 & 0.59157 \\ 0.714109 & 0.70003 \\ 0.959738 & 0.28089 \\ 0.840131 & 0.54238 \\ 0.729335 & 0.68415 \\ 0.993095 & 0.11730 \end{pmatrix}$$

$$\vec{U} = \begin{Bmatrix} 1 \\ 4 \\ 8 \end{Bmatrix} \Rightarrow \begin{Bmatrix} \dfrac{1}{\sqrt{1^2+4^2+8^2}} \\ \dfrac{4}{\sqrt{1^2+4^2+8^2}} \\ \dfrac{8}{\sqrt{1^2+4^2+8^2}} \end{Bmatrix} = \begin{bmatrix} 0,111 \\ 0,444 \\ 0,888 \end{bmatrix}$$

$$x' = \begin{bmatrix} 0,111 & 0,444 & 0,888 \end{bmatrix}$$

CAPÍTULO 5

UMA APLICAÇÃO DE ANÁLISE DE COMPONENTES PRINCIPAIS

5.1. Introdução a PCA

Uma Análise de Componentes Principais (PCA) está relacionada com a explicação da estrutura de variância e covariância, por meio de poucas combinações lineares das variáveis originais.

Os objetivos gerais da Análise de Componentes Principais (PCA) são redução dos dados e sua interpretação.

Embora P componentes sejam necessárias para reproduzir a variabilidade total do sistema, frequentemente muito desta variabilidade pode ser explicado por um número pequeno K de componentes principais; se isto ocorre, existe tanta informação nas K componentes principais quanto existe nas P variáveis originais.

As K componentes principais podem, então, substituir as P variáveis originais. Com isso, o conjunto de dados originais que consiste de n medidas sobre P variáveis é reduzido a um conjunto com n medidas sobre K componentes principais.

Uma Análise de Componentes Principais (PCA) frequentemente revela relações que não eram previamente suspeitadas e, assim, permite interpretações que não seriam obtidas.

A Análise de Componentes Principais (PCA) é, em geral, uma etapa para se chegar a um objetivo, visto que frequentemente serve como passo intermediário em grandes investigações.

Por exemplo, Análise de Componentes Principais (PCA) pode ser utilizada como entrada para uma regressão múltipla ou, mais comumente, para uma análise de agrupamentos.

5.2. Aplicação de Análise de Componentes Principais (PCA)

Seja o vetor aleatório $X = (X_1, X_2 \text{ e } X_3)$ tal que:

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Os autovalores ordenados são:

$$\lambda = (5{,}83;\ 2;\ 0{,}17)$$

E a matriz de autovetores correspondentes é:

$$\begin{pmatrix} 0,38 & 0 & 0,92 \\ -0,92 & 0 & 0,38 \\ 0 & 1 & 0 \end{pmatrix}.$$

Portanto, as três componentes principais são:

$$Y_1 = 0,38X_1 - 0,92X_2$$

$$Y_2 = X_3$$

$$Y_2 = 0,92X_1 + 0,38X_2$$

E as proporções da variabilidade total explicadas por cada componente são dadas por:

$$\frac{\lambda_1}{\sum_{i=1}^{3}\lambda_i} = 0,73, \quad \frac{\lambda_2}{\sum_{i=1}^{3}\lambda_i} = 0,25, \quad \frac{\lambda_3}{\sum_{i=1}^{3}\lambda_i} = 0,02.$$

Note que Y_1 e Y_2 explicam 98% da variabilidade total. Portanto, (X_1, X_2 e X_3) podem ser substituídas pelas duas primeiras componentes principais com pouca perda de informação.

Os coeficientes de correlação ficam:

$$\rho(Y_1, X_1) = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{0,38\sqrt{5,83}}{\sqrt{1}} = 0,92,$$

$$\rho(Y_1, X_2) = \frac{e_{21}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-0,92\sqrt{5,83}}{\sqrt{5}} = -0,99,$$

$$\rho(Y_1, X_3) = 0,$$

$$\rho(Y_2, X_1) = 0,$$

$$\rho(Y_2, X_2) = 0,$$

$$\rho(Y_2, X_3) = \frac{e_{32}\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = \frac{1\sqrt{2}}{\sqrt{2}} = 1,$$

$$\rho(Y_3, X_1) = \frac{e_{13}\sqrt{\lambda_3}}{\sqrt{\sigma_{33}}} = \frac{0,92\sqrt{0,17}}{\sqrt{2}} = 0,26$$

$$\rho(Y_3, X_2) = \frac{e_{23}\sqrt{\lambda_3}}{\sqrt{\sigma_{33}}} = \frac{0,38\sqrt{0,17}}{\sqrt{2}} = 0,11$$

$$\rho(Y_3, X_3) = \frac{e_{33}\sqrt{\lambda_3}}{\sqrt{\sigma_{33}}} = 0.$$

As componentes podem ser obtidas da padronização das variáveis:

$$Z_1 = \frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}$$

$$Z_2 = \frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}$$

$$\vdots$$

$$Z_p = \frac{x_p - \mu_p}{\sqrt{\sigma_{pp}}}$$

Em notação matricial:

$$Z = \left(V^{1/2}\right)^{-1}\left(X - \mu\right)$$

E(**Z**) = 0

$$Cov\left(Z\right) = \left(V^{1/2}\right)^{-1}\Sigma\left(V^{1/2}\right)^{-1} =$$

As componentes principais de **Z** podem serem obtidas dos autovetores da matriz de correlação $\boldsymbol{\rho}$ de **X**. Todos os nossos resultados anteriores se aplicam, com algumas simplificações, desde que a variância de cada Z_i seja unidade. Continuaremos a usar a notação Y_i para nos referir à i-ésima componente principal, e $\left(\lambda_i, e_i\right)$ para o par eigenvalue-eigenvector. No entanto, as quantidades derivadas de **Σ** em geral, não são as mesmas derivadas de $\boldsymbol{\rho}$.

A i-ésima componente principal das variáveis padronizadas $\mathbf{Z}' = [Z_1, Z_2, ..., Z_p]$, com Cov(**Z**)=$\boldsymbol{\rho}$, é dada por:

$$Y_i = e_i'Z = e_i'\left(V^{\frac{1}{2}}\right)^{-1} * \left(x - \mu\right), i = 1, 2, \ldots p$$

Além disso:

$$\sum_{i=1}^{p} Var\left(Yi\right) = \sum_{i=1}^{p} Var\left(Zi\right) = p$$

$$\rho_{Y_i Z_k} = e_{ki}\sqrt{\lambda_i} \qquad i, k = 1, \ldots p$$

Neste caso:

$\left(\lambda_1, e_1\right), \left(\lambda_2, e_2\right), \ldots, \left(\lambda_p, e_p\right)$ são os pares de autovalores e autovetores para $\boldsymbol{\rho}$, com $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

Por este resultado, temos que a var. populacional total, com relação às variáveis padronizadas, é simplesmente p, ou seja, a soma dos elementos da diagonal da matriz $\boldsymbol{\rho}$.

Com isso, a proporção da variância total explicada pela k-ésima componente de **Z** é:

$$\frac{\lambda_k}{p}, k = 1, 2, \ldots, p$$

Onde $\lambda_k^{'}s$ são os autovalores de $\boldsymbol{\rho}$.

Exemplo: Oito marcas de coxinhas de galinha foram avaliadas por 5 julgadores em relação a 4 atributos: sabor (X1), aroma (X2), qualidade da massa (X3) e qualidade de recheio (X4). Cada julgador atribuiu sua nota de 1 a 5, sendo que notas maiores estão relacionadas como melhor qualidade da coxinha. Os dados estão abaixo:

Obs: Em nossos estudos, onde tiver ponto, considere uma vírgula. O ponto **é** considerado para quando estamos programando, pois o sistema **é** americano.

Quadro 5. Dados de oito marcas de coxinhas em cinco variáveis.

Marca	Sabor	Aroma	Massa	Recheio
M1	2.75	4.03	2.8	2.62
M2	3.9	4.12	3.4	3.52
M3	3.12	3.97	3.62	3.05
M4	4.58	4.86	4.34	4.82
M5	3.97	4.34	4.28	4.98
M6	3.01	3.98	2.9	2.82
M7	4.19	4.65	4.52	4.77
M8	3.82	4.12	3.62	3.71

Fonte: Mingoti, 2007.

Matriz de Variância e Covariância:

	Sabor	Aroma	Massa	Recheio
Sabor	**0.407**	0.182	0.357	0.550
Aroma	0.182	**0.110**	0.180	0.271
Massa	0.357	0.180	**0.424**	0.590
Recheio	0.550	0.271	0.590	**0.911**

Os autovalores da matriz de covariância são:

$$\widehat{\lambda}_1 = 1.737,\ \widehat{\lambda}_2 = 0.065,\ \widehat{\lambda}_3 = 0.028,\ \widehat{\lambda}_4 = 0.022$$

E os autovetores são:

$\widehat{e}_1$	$\widehat{e}_2$	$\widehat{e}_3$	$\widehat{e}_4$
0.456	-0.816	0.112	0.337
0.223	-0.215	0.269	-0.912
0.477	0.456	0.718	0.221
0.717	0.282	-0.632	-0.077

As porcentagens de variância explicadas pelas componentes são: 93,8% para $\widehat{Y}_1$: 3,5% para $\widehat{Y}_2$ 1,5% para $\widehat{Y}_3$ e 1,2% para $\widehat{Y}_4$. Juntas, as duas primeiras componentes representam 97,3% da variância total do vetor original X. As componentes principais são respectivamente dadas por:

$\widehat{Y}_1$ =0,456(sabor)+0,223(aroma)+0,477(massa)+0,717(recheio)

$\widehat{Y}_2$ =-0,816(sabor)-0,215(aroma)+0,456(massa)+0,282(recheio)

$\widehat{Y}_3$ =0,112(sabor)+0,269(aroma)+0,718(massa)-0,632(recheio)

$\widehat{Y}_2$ =0,337(sabor)-0,912(aroma)+0,221(massa)-0,077(recheio)

A primeira componente é um índice global da qualidade de coxinha de acordo com o ponto de vista dos 5 julgadores. Quanto maior o valor numérico dessa componente, melhor é a qualidade da coxinha. Todas as quatro variáveis são importantes, nesse índice, de acordo com os valores numéricos dos respectivos coeficientes na combinação linear, sendo que o de maior valor refere-se à qualidade do recheio. A segunda componente é uma comparação entre o índice de qualidade referente ao sabor e aroma com um índice referente a massa e recheio.

Tabela 24. Tabela de correlação entre componentes e as variáveis originais.

	CP1	CP2	CP3	CP4
sabor	0.6006552	**-0.20787993**	0.01873660	0.05054996
aroma	0.2945063	-0.05475124	0.04496693	-0.13670176
massa	0.6286509	0.11626198	0.11992312	0.03316016
recheio	**0.9455793**	0.07180842	-0.10563584	-0.01157992

Exercício 1

Determine a componente principal populacional Y_1 e Y_2 da matriz de covariância.

$$\Sigma = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$$

Calcule a proporção total da variância populacional explicada pela primeira componente principal.

Exercício 2

Converta a matriz $\Sigma = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$ para a matriz de correlação $\boldsymbol{\rho}$.

Determine as componentes principais Y_1 e Y_2 para $\boldsymbol{\rho}$ e calcule a proporção total da variância populacional explicada por Y_1.

Compare os componentes obtidos no exercício 1. Eles deveriam ser os mesmos?

Exercício 3

Calcule a correlação $\rho_{Y_1Z_1}$, $\rho_{Y_1Z_2}$ e $\rho_{Y_2Z_1}$.

REFERÊNCIAS

BIANCHINI, W. **Aprendendo Cálculo de Várias Variáveis.** UFRJ, 2016.

BOLDRINI, J. L.; COSTA, S. I. R.; FIGUEIREDO, V. L.; WETZLER, H. G. **Álgebra Linear**. 3. ed. Harbra, 1986.

BRYAN J. F.; MANLY, **Métodos Estatísticos Multivariados uma Introdução**. 3. ed. Porto Alegre: Bookman, 2008.

BRYAN F. J.; MANLY, ALBERTO J. A. N. **Métodos Estatísticos Multivariados**: Uma Introdução. 4. ed. Porto Alegre: Bookman, 2019.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2004.

CRESPO, A. A**. Estatística Fácil,** São Paulo: Saraiva, 2004

FONSECA, J. S.; MARTINS, G. A., **Curso de Estatística**. 6. ed. São Paulo: Atlas, 1996.

GUIDORIZZI, H. L. **Um Curso de Cálculo**. Vol. 1 e 2. LTC, 2014.

JOHONSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 3. ed. New Jersey: Prentice-Hall, 1992.

MINGOTI, S. A. **Análise de Dados Métodos de Estatística Multivariada**. Minas Gerais: Editora da UFMG, 2005. 300 p.

RORRES, C.; HOWARD, A. Álgebra Linear com Aplicações. 10. ed. Porto Alegre: Bookman, 2012.

SIMMONS, G. F. **Cálculo com Geometria Analítica**. Vol 2, McGraw-Hill, 1988.

STEINBRUCH, A.; WINTERLE, P. Álgebra Linear. 2. ed. Pearson, 2010.

TIZZIOTTI, G. C.; SANTOS, J. V. Álgebra Linear. UFU, 2012.

https://github.com/allanbreyes/udacity-data-science/blob/master/p3/data/wineQualityReds.csv

ANEXO I

Uma amostra de 5 observações com ingredientes e as medidas dos compostos de vinhos

	Unnamed: 0	fixed. acidity	volatile. acidity	citric. acid	residual. sugar	chlorides	free.sulfur. dioxide	total.sulfur. dioxide	density	pH	sulphates	alcohol	quality
0	1	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	2	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	3	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	4	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	5	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Fonte: https://github.com/allanbreyes/udacity-data-science/blob/master/p3/data/wineQualityReds.csv.