Estatística e Probabilidade

Medidas de Dispersão

Desenvolvimento do material

Gregório Dalle Vedove Nosaki

1ª Edição

Sumário

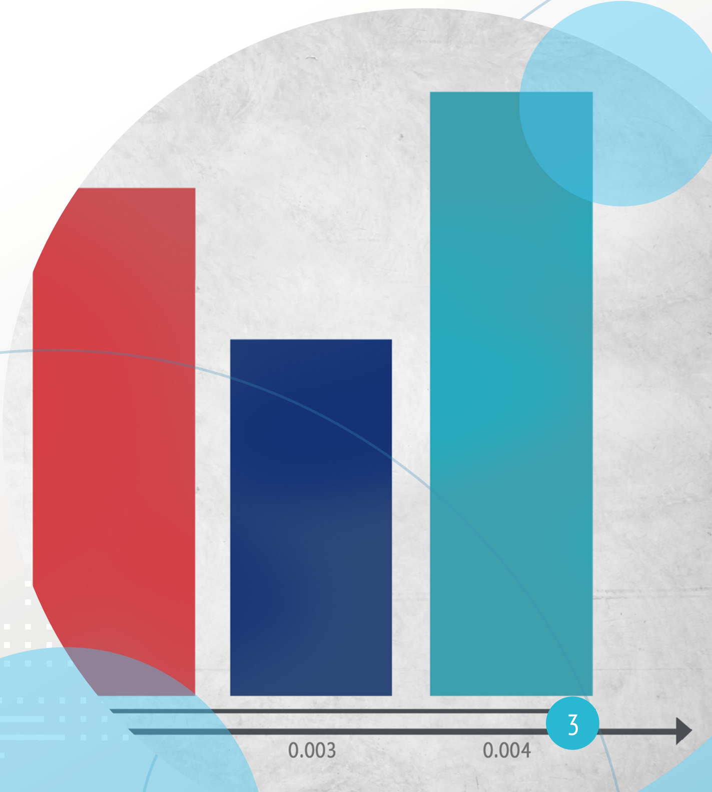
Medidas de Dispersão

Para início de conversa...

Já trabalhamos com as medidas de posição de tendência central que, de maneira geral, tentavam descrever o comportamento dos valores de uma variável em uma amostra. Vimos que nem sempre essa informação é suficiente para dar uma ideia de como os valores estão distribuídos. Neste capítulo, vamos trabalhar com as medidas de dispersão. Essas medidas, juntamente com as medidas de posição, fornecem uma visão mais completa para uma análise rápida de um conjunto de dados. Começaremos apresentando os conceitos de amplitude, variância e desvio padrão. Esses dois últimos são os mais utilizados em estudos estatísticos de banco de dados. Veremos como as unidades de medidas estão relacionadas para essas duas medidas de dispersão. Trabalharemos ainda com o coeficiente de variação e o desvio médio de uma coleção de valores e como interpretá-los dentro da nossa amostra. Todas as definições serão exemplificadas com exemplos numéricos e também com situações-problemas comentadas, para que fiquem mais claros todos os conceitos que serão introduzidos neste capítulo.

Objetivos

- Calcular e interpretar as medidas de dispersão, amplitude total, variância, desvio padrão e coeficiente de variação.
- Entender as propriedades da média e o desvio padrão.



1. Variância e Desvio Padrão

Uma medida de tendência central pode não fornecer informações suficientes sobre o nosso conjunto de dados. Veja, por exemplo, as seguintes listas de observáveis.

A = 6   6   6   6   6   6

B = 1   2   3   3   12   15

C = 1   3   3   7   10   12

A média aritmética das três coleções de dados A, B e C é igual a 6, mas apenas essa medida nos dá pouquíssima informação sobre o conjunto dos dados que estamos trabalhando. Até mesmo a moda e a mediana não fornecem informações suficientes: a moda e a mediana do conjunto A é igual a 6, a mediana e a moda do conjunto B é 3 e a moda do conjunto C também é 3 e a mediana é 6. Complementando as medidas de tendência central, vamos apresentar as medidas de dispersão.

Definição: chamamos de **amplitude** de um conjunto de dados a diferença entre o maior e o menor valor dos dados.

Ainda trabalhando com os conjuntos A, B e C apresentados anteriormente, a amplitude do conjunto A é igual a 0, a do conjunto B é igual a 14 e a do conjunto C é igual a 11. Baseado na amplitude já podemos ter uma noção melhor de como os valores do nosso conjunto estão distribuídos. Apresentaremos agora de maneira intuitiva as duas principais medidas de dispersão: a variância e o desvio padrão.

Considere o conjunto de dados B dado anteriormente, ou seja, estamos considerando os valores

1   2   3   3   12   15

Para determinarmos o grau de variação que os valores na nossa coleção têm em relação à média, uma primeira abordagem seria subtrair a média de cada um desses valores. Dessa forma, obtemos

1 - 6 = - 5

2 - 6 = - 4

3 - 6 = - 3

3 - 6 = - 3

12 - 6 = 6

15 - 6 = 9

Esses valores podem ser interpretados como as “distâncias” entre os valores reais da amostra e o valor da média aritmética da coleção. Tais valores serão chamados de **desvios**. Note que, ao realizarmos tal tipo de análise, a soma dos desvios é igual a zero

(-5) + (-4) + (-3) + (-3) + 6 + 9 = 0

Essa propriedade é sempre válida para qualquer que seja a coleção de observáveis. Realize as verificações para os conjuntos A e C apresentados anteriormente.

Para que possamos ter uma visão mais objetiva e sem sinais negativos nos valores que obtivemos anteriormente, elevamos ao quadrado todos os valores dos desvios obtidos anteriormente. Dessa forma

$$(1 - 6)^2 = (-5)^2 = 25$$

$$(2 - 6)^2 = (-4)^2 = 16$$

$$(3 - 6)^2 = (-3)^2 = 9$$

$$(3 - 6)^2 = (-3)^2 = 9$$

$$(12 - 6)^2 = (6)^2 = 36$$

$$(15 - 6)^2 = (9)^2 = 81$$

A média da soma dos quadrados dos desvios é o que chamamos de variância. No caso do conjunto B, a variância é igual a

$$\frac{25+16+9+9+36+81}{6} = \frac{176}{6} \approx 29,33 \ .$$

Utilizaremos o sinal “ $\approx$ ” para denotar um arredondamento realizado entre os valores, ou seja, “ $\approx$ ” indica que um valor é aproximadamente igual a outro.

Vamos definir de maneira generalizada como obter a variância para uma coleção qualquer de valores. Para as próximas definições considere um conjunto de dados X com n elementos dada como sendo $X = \{x_1, x_2, ..., x_n\}$. Denotaremos por $\overline{x}$ a média aritmética da coleção X.

Definição: dado um conjunto de dados $X = \{x_1, x_2, ..., x_n\}$, a **variância** do conjunto X é dada por

$$var(X) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \ .$$

O desvio padrão é igual à raiz quadrada de valor positivo da variância. No caso do conjunto B, vale que o desvio padrão é igual a

$$\sqrt{29,33} \approx 5,4157.$$

Definição: dado um conjunto de dados $X = \{x_1, x_2, ..., x_n\}$, o **desvio padrão** do conjunto X é dado por

$$dp(X) = \sqrt{var(X)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}.$$

A unidade de medida do desvio padrão é igual à unidade de medida das variáveis, enquanto que a unidade de medida da variância é igual à unidade de medida das variáveis ao quadrado. Vejamos um exemplo dos cálculos dessas medidas de dispersão em um exemplo.

Exemplo 1: para a realização de uma mesma tarefa, o tempo de realização de seis funcionários diferentes são apresentados na tabela a seguir.

Funcionário	Tempo de realização (min.)
A	52
B	45
C	50
D	54
E	52
F	42

A primeira medida que devemos calcular é a média aritmética entre os valores apresentados. Seja x_A o tempo do funcionário A, x_B o tempo do funcionário B e assim por diante. Denotando a média do tempo de realização desta tarefa por $\overline{x}$, temos que

$$\overline{x} = \frac{x_A + x + B + x_C + x_D + x_E + x_F}{6} = \frac{52 + 45 + 50 + 54 + 52 + 42}{6} = 50.$$

A partir dessa medida de posição, podemos determinar os desvios de cada um dos funcionários e proceder com o cálculo da variância e do desvio padrão. Observe a tabela abaixo com os valores correspondentes aos desvios de cada funcionário e o quadrado dos desvios.

Funcionário	Tempo de realização (min.)	$x - \overline{x}$	$(x - \overline{x})^2$
A	52	2	4
B	45	-5	25
C	50	0	0
D	54	4	16
E	52	2	4
F	42	-8	64

A variância dos valores apresentados é igual à média dos valores da quarta coluna, ou seja,

$$\frac{4+25+0+16+4+64}{6} = \frac{113}{6} \approx 18,83\, min^2$$

e, portanto, o desvio padrão é igual a

$$\sqrt{18,83} \approx 4,34\, min.$$

Note que a unidade de medida da variância é min² (minutos ao quadrado) e a unidade de medida do desvio padrão é min (minutos). Utilizamos geralmente o desvio padrão para estudar a dispersão de uma amostra pois ele é dado na mesma unidade de medida que as variáveis analisadas tornando essa comparação mais fácil e intuitiva.

É comum adicionar colunas como feito no exemplo anterior para representar os desvios e o quadrado dos desvios. Dessa forma, fica mais claro e evita erros de cálculo quando estamos trabalhando com uma amostra com muitos valores.

Importante

Em algumas bibliografias, a variância e o desvio padrão são representados por s^2 e s, respectivamente. Outra notação que é bastante comum é σ^2 para a variância e σ para o desvio padrão. Na notação apresentada anteriormente
($var(X)$ e $dp(X)$) fica claro sobre qual conjunto de valores estamos nos referindo, mas caso não haja confusão, as notação apresentadas aqui também podem ser utilizadas.

2. Coeficiente de variação

O coeficiente de variação de uma amostra é calculado baseado no desvio padrão dessa amostra com relação à média. Apresentaremos aqui uma definição na qual o coeficiente de variação é dado em porcentagem, para que fique mais simples a análise dos seus resultados. Considere um conjunto de dados $X = \{x_1, x_2, ..., x_n\}$ sendo $\overline{x}$ a média aritmética entre os valores e $dp(X)$ o desvio padrão de X.

Definição: dado uma coleção $X = \{x_1, x_2, ..., x_n\}$, definimos o **coeficiente de variação** como sendo

$$CV(X) = \frac{dp(X)}{\overline{x}} \cdot 100.$$

Interpretaremos o coeficiente de variação segundo as faixas:

- até 15% diremos que há uma baixa dispersão, ou seja, que os dados são mais homogêneos;
- entre 15% e 30% diremos que a dispersão é moderada;
- acima de 30% diremos que há uma alta dispersão, ou seja, os dados são mais heterogêneos.

Vamos calcular o coeficiente de dispersão para o Exemplo 1. Temos que o desvio padrão é igual a 4,34 e a média é igual a 50, portanto o coeficiente de variação é igual a

$$\frac{4,34}{50} \cdot 100 = 8,68\%.$$

Neste exemplo, os dados têm uma baixa dispersão, o que significa que eles tendem a ficar próximos ao valor da média.

No caso do conjunto de valores

1 2 3 3 12 15

temos que a média é igual a 6 e o desvio padrão é de aproximadamente 5,4157 e, portanto, o coeficiente de variação é igual a

$$\frac{5,4157}{6} \cdot 100 \approx 90,26\%.$$

O conjunto de dados tem uma alta dispersão de valores em comparação com a média, que, neste caso, é 6.

Importante

O coeficiente de variação será sempre apresentado em porcentagem, independentemente da unidade da variável que estaremos trabalhando.

Antes de apresentarmos mais exemplos com o cálculo dessa medida de dispersão, iremos definir a última medida de dispersão a ser abordada neste capítulo.

3. Desvio Médio

A última medida de dispersão que trabalharemos é o desvio médio. Considere novamente um conjunto de dados $X = \{x_1, x_2, ..., x_n\}$ sendo $\overline{x}$ a média aritmética entre os valores.

Definição: dado uma coleção $X = \{x_1, x_2, ..., x_n\}$, definimos o **desvio médio** como sendo

$$DM(X) = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}.$$

No Exemplo 1, temos que o desvio médio é igual a

$$\frac{|52-50|+|45-50|+|50-50|+|54-50|+|52-50|+|42-50|}{6}=3,5\ .$$

No exemplo do conjunto B = 1 2 3 3 12 15, o desvio médio é igual a

$$\frac{|1-6|+|2-6|+|3-6|+|3-6|+|12-6|+|15-6|}{6}=5\ .$$

Importante

O desvio padrão e o desvio médio são medidas de dispersão distintas, mas podem ser interpretadas da mesma forma, ou seja, quanto maior o valor desses desvios, mais variada será nossa base de dados considerada, pois existe uma maior variação entre os valores com relação à média aritmética.

Vejamos mais alguns exemplos do cálculo das medidas de dispersão que foram apresentadas neste capítulo.

Exemplo 2: considere as médias finais de 10 alunos de uma sala nas disciplinas de Matemática, Física e Química apresentadas na tabela a seguir.

Nome	Matemática	Física	Química
Márcio	8,5	9	8
Roberta	7,5	8,5	9
Felipe	6	7	7
Guilherme	5,5	6	9
Gustavo	9	8,5	8
Paula	9,5	7	8,5
Sabrina	8	8	8
Mônica	8,5	7,5	6
Fernando	7	8	7,5
Tiago	6,5	9	5,5

Vamos calcular a média aritmética de cada uma das disciplinas e depois as medidas de dispersão apresentadas neste capítulo. Denotaremos as médias de Matemática, Física e Química como x_M, x_F e x_Q, respectivamente.

Efetuando os cálculos, obtemos que

$$x_M=\frac{8,5+7,5+6+5,5+9+9,5+8+8,5+7+6,5}{10}=\frac{76}{10}=7,6$$

$$x_F=\frac{9+8,5+7+6+8,5+7+8+7,5+8+9}{10}=\frac{78,5}{10}=7,85$$

$$x_Q=\frac{8+9+7+9+8+8,5+8+6+7,5+5,5}{10}=\frac{76,5}{10}=7,65$$

Efetuamos agora detalhadamente o cálculo das medidas de dispersão para a disciplina de Matemática. Vamos construir uma tabela com o nome dos alunos, suas médias finais em Matemática e adicionaremos duas colunas: uma com a diferença entre a média do aluno e a média x_M e outra coluna como essa diferença elevada ao quadrado. Dessa forma, obtemos a seguinte tabela:

Nome	Matemática	$x - x_M$	$(x - x_M)^2$
Márcio	8,5	0,9	0,81
Roberta	7,5	-0,1	0,01
Felipe	6	-1,6	2,56
Guilherme	5,5	-2,1	4,41
Gustavo	9	1,4	1,96
Paula	9,5	1,9	3,61
Sabrina	8	0,4	0,16
Mônica	8,5	0,9	0,81
Fernando	7	-0,6	0,36
Tiago	6,5	-1,1	1,21

Considerando o conjunto de notas finais em Matemática como M, a variância desse conjunto é igual a

$$var(M) = \frac{0,81 + 0,01 + 2,56 + 4,41 + 1,96 + 3,61 + 0,16 + 0,81 + 0,36 + 1,21}{10} = 1,59$$

e o desvio padrão é

$$dp(M) = \sqrt{var(M)} = \sqrt{1,59} \approx 1,261 \, .$$

Podemos calcular ainda o coeficiente de variação e o desvio médio para as notas de Matemática. O coeficiente de variação é

$$CV(M) = \frac{dp(M)}{x_M} \cdot 100 = \frac{1,261}{7,6} \cdot 100 \approx 16,59\%$$

e o desvio médio é

$$DM(M) = \frac{|0,9| + |-0,1| + |-1,6| + |-2,1| + |1,4| + |1,9| + |0,4| + |0,9| + |-0,6| + |-1,1|}{10} = 1,1 \, .$$

Deixaremos como exercício o cálculo das medidas de dispersão para as notas finais das disciplinas de Física e Química. Para que você possa comparar seus resultados finais, apresentaremos os valores finais desses cálculos. Denotaremos o conjunto de notas finais de Física por F e o conjunto de notas finais de Química por Q.

$$var(F) = 0,8525 \quad var(Q) = 1,2525$$

$$dp(F) \approx 0,9233 \quad dp(Q) \approx 1,1192$$

$$CV(F) = 11,76\% \quad CV(Q) = 14,63\%$$

$$DM(F) = 0,78 \quad DM(Q) = 0,92$$

Exemplo 3: considerando ainda a tabela de notas finais apresentada no Exemplo 2, vamos determinar qual aluno tem a maior e a menor variação entre as notas das três disciplinas apresentadas. Para realizar essa análise, vamos tomar como exemplo as notas do aluno Márcio.

Nome	Matemática	Física	Química
Márcio	8,5	9	8

Primeiramente, devemos determinar a média entre as três disciplinas para Márcio. Obtemos que a média de Márcio é

$$\overline{x} = \frac{8,5+0+8}{3} = \frac{22,5}{3} = 8,5.$$

A partir daí, podemos calcular a variância e o desvio padrão de Márcio. Seja $var(\text{Márcio})$ a variação das notas de Márcio e $dp(\text{Márcio})$ o desvio padrão das notas de Márcio. Obtemos que

$$var(\text{Márcio}) = \frac{(8,5-8,5)^2+(9-8,5)^2+(8-8,5)^2}{3} = \frac{0,5}{3} \approx 0,16$$

e

$$dp(\text{Márcio}) = \sqrt{0,16} = 0,4.$$

Podemos também calcular o coeficiente de variação das notas de Márcio ($CV(\text{Márcio})$) e o desvio médio ($DM(\text{Márcio})$) como apresentado:

$$CV(\text{Márcio}) = \frac{dp(\text{Márcio})}{\overline{x}} \cdot 100 = \frac{0,4}{8,5} \cdot 100 = 4,7\%$$

e

$$DM(\text{Márcio}) = \frac{|8,5-8,5|+|9-8,5|+|8-8,5|}{3} = \frac{1}{3}.$$

Repetindo os cálculos para todos os alunos, obtemos os resultados aproximados apresentados na tabela a seguir.

Nome	Variação	Desvio Padrão	Coeficiente de Variação	Desvio Médio
Márcio	0,16	0,4	4,7%	0,33
Roberta	0,39	0,624	7,48%	0,55
Felipe	0,22	0,47	7,07%	0,44
Guilherme	2,39	1,546	22,61%	1,44

Gustavo	0,16	0,4	4,7%	0,33
Paula	1,05	1,025	12,3%	0,88
Sabrina	0	0	0%	0
Mônica	1,05	1,025	14%	0,88
Fernando	0,16	0,4	5,44%	0,33
Tiago	2,16	1,47	21%	1,33

Analisando os valores da tabela, podemos concluir que a aluna Sabrina tem a menor variação entre as notas das três disciplinas, e o aluno Guilherme é o que apresenta maior variação entre as notas das disciplinas consideradas.

As medidas de dispersão nos fornecem informações importantes sobre a distribuição dos valores da nossa amostra. Com base nessas medidas e nas medidas de posição introduzidas no capítulo anterior, podemos ter uma visão mais completa dos nossos dados. Trabalhamos inicialmente com os conceitos de amplitude de uma amostra, variância e desvio padrão. O desvio padrão e a variância têm fórmulas gerais muito similares e são as principais medidas de dispersão consideradas em estudos estatísticos. Além dessas medidas, introduzimos os conceitos de coeficiente de variação e desvio médio, que também ajudam a estabelecer uma relação entre a distribuição dos valores das observáveis. Apresentamos exemplos comentados e deixamos alguns exercícios com o cálculo dessas medidas, para fixar suas definições e realizar o cálculo de cada uma delas.

Referências

FONSECA, J. S.; MARTINS, G. A. **Curso de estatística**. 6. ed. São Paulo: Atlas, 2012.

MOORE, D.; NOTZ, W.; FLIGNER, M. **A estatística básica e sua prática**. 7. ed. Rio de Janeiro: LTC, 2017.

MORETTIN, P.; BUSSAB, W. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017.

RIGONATTO, M. Coeficiente de variação. Cálculo de coeficiente de variação. **Mundo Educação** [*on-line*]. Disponível em: https://mundoeducacao.uol.com.br/matematica/coeficiente-variacao.htm. Acesso em: 28 jun. 2021.

SPIEGEL, M.; SCHILLIER, J.; SRINIVASAN, A. **Probabilidade e estatística**. 3. ed. Porto Alegre: Bookman, 2013.

TRIOLA, M. **Introdução à estatística**. 12. ed. Rio de Janeiro: LTC, 2017.

VIEIRA, S. **Fundamentos de estatística**, 6. ed. São Paulo: Atlas, 2019.