Aula 4 – Medidas de dispersão

Nesta aula, você estudará as medidas de dispersão de uma distribuição de dados e aprenderá os seguintes conceitos:

- amplitude
- desvios em torno da média
- desvio médio absoluto
- variância
- desvio padrão

Amplitude

Considere os conjuntos de dados apresentados por um *diagrama de pontos* na **Figura 4.1**. Nesse gráfico, as "pilhas" de pontos representam as freqüências de cada valor. Podemos ver facilmente que ambos os conjuntos têm a mesma média (o centro de gravidade ou ponto de equilíbrio é o mesmo), a mesma mediana e a mesma moda. No entanto, esses dois conjuntos têm características diferentes e ao sintetizá-los apenas por alguma medida de posição, essa característica se perderá. Tal característica é a *dispersão* dos dados: no primeiro conjunto, os dados estão mais concentrados em torno da média do que no segundo conjunto.

DISPERSÃO

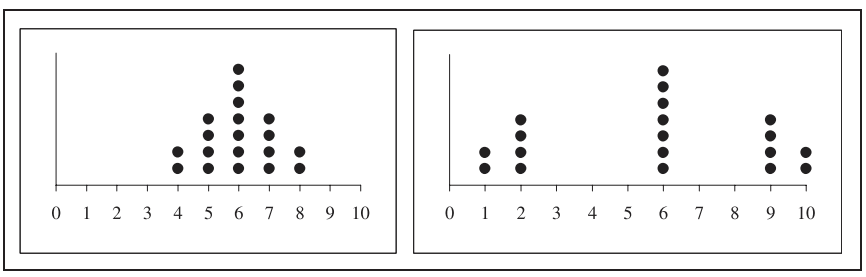

Figura 4.1: Conjuntos de dados com medidas de posição iguais e dispersão diferente.

Como podemos "medir" essa dispersão? Uma primeira idéia é considerar a *amplitude* dos dados, que é, como já visto, a diferença entre o maior e o menor valor.

Definição
A **amplitude** de um conjunto de dados é a distância entre o maior valor e o menor valor.

$$\Delta_{total} = V_{\max} - V_{\min}. \tag{4.1}$$

A amplitude tem a mesma unidade dos dados, mas ela tem algumas limitações, conforme ilustrado na **Figura 4.2**. Aí os dois conjuntos têm a mesma média, a mesma mediana e a mesma amplitude, mas essas medidas não conseguem caracterizar o fato de a distribuição dos valores entre o mínimo e o máximo ser diferente nos dois conjuntos. A limitação da amplitude também fica patente pelo fato de ela se basear em apenas duas observações, independentemente do número total de observações.

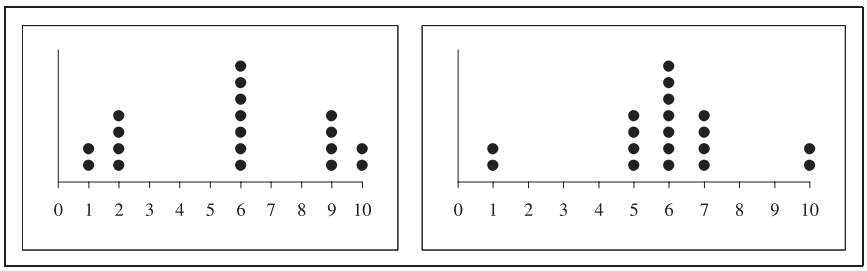

Figura 4.2: Conjuntos de dados com medidas de posição e amplitude iguais.

Desvio médio absoluto

Uma maneira de se medir a dispersão dos dados é considerar os tamanhos dos *desvios* $x_i - \overline{x}$ de cada observação em relação à média. Note nas figuras acima que, quanto mais disperso o conjunto de dados, maiores esses desvios tendem a ser. Para obter uma medida-resumo, isto é, um único número, poderíamos somar esses desvios, ou seja, considerar a seguinte medida:

$$D = \sum_{i=1}^{n} (x_i - \overline{x}). \tag{4.2}$$

Vamos desenvolver tal fórmula, usando as propriedades de somatório e a definição da média amostral.

$$\begin{aligned} D &= \sum_{i=1}^{n} (x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = \sum_{i=1}^{n} x_i - n\overline{x} = \\ &= \sum_{i=1}^{n} x_i - n \times \frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0. \end{aligned}$$

Ou seja: essa medida, que representa a soma dos desvios em relação à média, é sempre nula, não importa o conjunto de dados! Logo, ela não serve para diferenciar quaisquer conjuntos!

Vamos dar uma explicação intuitiva para esse fato, que nos permitirá obter correções para tal fórmula. Ao considerarmos as diferenças entre cada valor e o valor médio, obtemos valores negativos e positivos, pois, pela definição de média, sempre existem valores menores e maiores que a média; esses valores positivos e negativos, ao serem somados, se anulam.

Bom, se o problema está no fato de termos valores positivos e negativos, por que não trabalhar com o valor absoluto das diferenças? De fato, esse procedimento nos leva à definição de *desvio médio absoluto.*

Definição
O **desvio médio absoluto** de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por

$$DMA = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}| \tag{4.3}$$

onde as barras verticais representam o valor absoluto ou módulo.

Note que nesta definição estamos trabalhando com o desvio médio, isto é, tomamos a média dos desvios absolutos. Isso evita interpretações equivocadas, pois, se trabalhássemos apenas com a soma dos desvios absolutos, um conjunto com um número maior de observações tenderia a apresentar um resultado maior para a soma devido apenas ao fato de ter mais observações. Esta situação é ilustrada com os seguintes conjuntos de dados:

- Conjunto 1: $\{1, 3, 5\}$
- Conjunto 2: $\left\{1, \frac{5}{3}, 3, \frac{13}{3}, 5\right\}$

Para os dois conjuntos, $\overline{x} = 3$ e para o conjunto 1

$$\sum_{i=1}^{3} |x_i - \overline{x}| = |1-3| + |3-3| + |5-3| = 4$$

e para o conjunto 2

$$\sum_{i=1}^{5} |x_i - \overline{x}| = |1-3| + \left|\frac{5}{3} - 3\right| + |3-3| + \left|\frac{13}{3} - 3\right| + |5-3| = \frac{20}{3} = 6,667.$$

Então, o somatório para o segundo conjunto é maior, mas o desvio absoluto médio é o mesmo para ambos; de fato, para o primeiro conjunto temos

$$DMA = \frac{4}{3}$$

e para o segundo conjunto

$$DMA = \frac{\frac{20}{3}}{5} = \frac{4}{3}$$

Ao dividirmos o somatório pelo número de observações, compensamos o fato de o segundo conjunto ter mais observações que o primeiro.

O desvio médio absoluto tem a mesma unidade dos dados.

Atividade 4.1

Para o conjunto de dados $2, 4, 7, 8, 9, 6, 5, 8$, calcule os desvios em torno da média e verifique que eles somam zero. Em seguida, calcule o desvio médio absoluto.

Variância e desvio padrão

Considerar o valor absoluto das diferenças $(x_i - \overline{x})$ é uma das maneiras de se contornar o fato de que $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. No entanto, a função módulo tem a desvantagem de ser não diferenciável no ponto zero. Outra possibilidade de correção, com propriedades matemáticas e estatísticas mais adequadas, é considerar o quadrado das diferenças. Isso nos leva à definição de *variância*.

Definição

A **variância**[1] de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definida por

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 . \tag{4.4}$$

Note que esta definição de variância nos diz que a variância é *a média dos desvios quadráticos*.

Suponhamos que os valores x_i representem os pesos, em quilogramas, de um conjunto de pessoas. Então, o valor médio $\overline{x}$ representa o peso médio

dessas pessoas e sua unidade também é quilogramas, o mesmo acontecendo com as diferenças $(x_i - \overline{x})$. Ao elevarmos essas diferenças ao quadrado, passamos a ter a variância medida em quilogramas ao quadrado, uma unidade que não tem interpretação física. Uma forma de se obter uma medida de dispersão com a mesma unidade dos dados consiste em tomar a raiz quadrada da variância.

Definição
O **desvio padrão** de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por

$$\sigma = \sqrt{\text{Variância}} = \sqrt{\sigma^2} \tag{4.5}$$

A título de ilustração, vamos considerar novamente os dados analisados na aula anterior, referentes à idade dos funcionários do Departamento de Recursos Humanos. Essas idades são:

24 25 26 26 29 29 31 35 36 37 38 42 45 51 53

e sua média é $\frac{527}{15} = 35,1\overline{3}$. Assim, a variância, em anos^2 é

$$\begin{aligned}\sigma^2 &= \frac{1}{15}\left[\begin{array}{c}(24-35,13)^2+(25-35,13)^2+2\times(26-35,13)^2+2\times(29-35,13)^2+\\(31-35,13)^2+(35-35,13)^2+(36-35,13)^2+(37-35,13)^2+(38-35,13)^2+\\(42-35,13)^2+(42-35,13)^2+(45-35,13)^2+(51-35,13)^2+(53-35,13)^2\end{array}\right] = \\ &= \frac{1213,73}{15} = 80,92\end{aligned}$$

e o desvio padrão, em anos, é

$$\sigma = \sqrt{80,92} = 8,995$$

Atividade 4.2

Para o conjunto de dados da Atividade $1 - \{2, 4, 7, 8, 9, 6, 5, 8\}$ – calcule a variância e o desvio padrão.

Fórmula alternativa para o cálculo da variância

Consideremos a Equação (4.4) que define a variância. Desenvolvendo o quadrado e usando as propriedades de somatório, obtemos:

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}2\overline{x}x_i + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) + \frac{1}{n}n\overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}^2 + \overline{x}^2\end{aligned}$$

ou seja

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2 \tag{4.6}$$

Essa forma de escrever a variância facilita quando os cálculos têm que ser feitos à mão ou em calculadoras menos sofisticadas, pois o número de cálculos envolvidos é menor. Note que ela nos diz que a variância é a *média dos quadrados menos o quadrado da média*.

Vamos calcular a variância das idades dos funcionários de RH usando essa fórmula:

$$\begin{aligned}\sigma^2 &= \frac{1}{15}\left[\begin{array}{c}24^2 + 25^2 + 25^2 + 2\times 26^2 + 2\times 29^2 + 31^2 + 35^2 + 36^2 + \\ 37^2 + 38^2 + 39^2 + 42^2 + 45^2 + 51^2 + 53^2\end{array}\right] \\ &\quad - \left(\frac{527}{15}\right)^2 = \\ &= \frac{19729\times 15 - 527^2}{15^2} = \frac{295935 - 277729}{225} = \frac{18206}{225} = 80,916\end{aligned}$$

Na comparação dos resultados obtidos pelas duas fórmulas, pode haver alguma diferença por causa dos arredondamentos, uma vez que a média é uma dízima.

Atividade 4.3

Na Atividade 2 você calculou a variância do conjunto de dados $\{2, 4, 7, 8, 9, 6, 5, 8\}$ como a média dos desvios quadráticos. Calcule a variância novamente utilizando a fórmula alternativa dada na Equação (4.6).

Exemplo 4.1

Na aula anterior, analisamos os dados referentes ao número de dependentes dos funcionários do Departamento de Recursos Humanos, apresentados novamente na tabela a seguir.

Nome	No.de dependentes	Nome	No.de dependentes
João da Silva	3	Patrícia Silva	2
Pedro Fernandes	1	Regina Lima	2
Maria Freitas	0	Alfredo Souza	3
Paula Gonçalves	0	Margarete Cunha	0
Ana Freitas	1	Pedro Barbosa	2
Luiz Costa	3	Ricardo Alves	0
André Souza	4	Márcio Rezende	1
Ana Carolina Chaves	0		

Como o menor valor é 0 e o maior valor é 4, temos que a amplitude dos dados é de 4 dependentes. A média calculada para esses dados foi $\overline{x} = \frac{22}{15} = 1,467$. Vamos calcular a soma dos desvios em torno da média, usando o fato de que temos observações repetidas.

$$\begin{aligned}\sum(x_i - \overline{x}) &= 5 \times \left(0 - \frac{22}{15}\right) + 3 \times \left(1 - \frac{22}{15}\right) + 3 \times \left(2 - \frac{22}{15}\right) + 3 \times \left(3 - \frac{22}{15}\right) + \left(4 - \frac{22}{15}\right) = \\ &= -\frac{110}{15} - \frac{21}{15} + \frac{24}{15} + \frac{69}{15} + \frac{38}{15} = -\frac{131}{15} + \frac{131}{15} = 0\end{aligned}$$

Caso trabalhássemos com o valor aproximado $1,467$, o resultado aproximado seria $-0,005$.

O desvio médio absoluto é

$$\begin{aligned}DMA &= \frac{1}{n}\sum|x_i - \overline{x}| = \\ &= \frac{1}{15} \times \left[5 \times \left|0 - \frac{22}{15}\right| + 3 \times \left|1 - \frac{22}{15}\right| + 3 \times \left|2 - \frac{22}{15}\right| + 3 \times \left|3 - \frac{22}{15}\right| + \left|4 - \frac{22}{15}\right|\right] = \\ &= \frac{1}{15} \times \left[\frac{110}{15} + \frac{21}{15} + \frac{24}{15} + \frac{69}{15} + \frac{38}{15}\right] = \frac{1}{15} \times \left[\frac{131}{15} + \frac{131}{15}\right] = \frac{262}{225} = 1,1644\end{aligned}$$

A variância é

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum(x_i - \overline{x})^2 \\ &= \frac{1}{15} \times \left[5 \times \left(0 - \frac{22}{15}\right)^2 + 3 \times \left(1 - \frac{22}{15}\right)^2 + 3 \times \left(2 - \frac{22}{15}\right)^2 + 3 \times \left(3 - \frac{22}{15}\right)^2 + \left(4 - \frac{22}{15}\right)^2\right] = \\ &= \frac{1}{15} \times \left[\frac{2420}{225} + \frac{147}{225} + \frac{192}{225} + \frac{1587}{225} + \frac{1444}{225}\right] = \frac{5790}{15 \times 225} = 1,715556\end{aligned}$$

e

$$\sigma = \sqrt{\frac{5790}{15 \times 225}} = 1,3098$$

Vamos agora calcular a variância usando a fórmula alternativa:

$$\begin{aligned}\sigma^2 &= \frac{1}{15}\times\left(5\times 0^2+3\times 1^2+3\times 2^2+3\times 3^2+4^2\right)-\left(\frac{22}{15}\right)^2= \\ &= \frac{3+12+27+16}{15}-\frac{484}{225}=\frac{58}{15}-\frac{484}{225}=\frac{58\times 15-484}{225}= \\ &= \frac{386}{225}=1,715556\end{aligned}$$

Note que com essa fórmula os cálculos ficam bem mais simples, uma vez que temos que fazer menos conta!

Propriedades das medidas de dispersão

Como visto para as medidas de posição, vamos ver as principais propriedades das medidas de dispersão.

Propriedade 1

Todas as medidas de dispersão são não negativas!

$$\begin{aligned}&\Delta \geq 0\\ &DMA \geq 0\\ &\sigma^2 \geq 0\\ &\sigma \geq 0\end{aligned} \tag{4.7}$$

Propriedade 2

Somando-se uma mesma constante a todas as observações, as medidas de dispersão não se alteram. Essa propriedade é bastante intuitiva se notarmos que, ao somar uma constante aos dados, estamos simplesmente fazendo uma translação dos mesmos, sem alterar a dispersão.

$$y_i = x_i + k \Rightarrow \begin{cases} \Delta_y = \Delta_x \\ DMA_y = DMA_x \\ \sigma_y^2 = \sigma_x^2 \\ \sigma_y = \sigma_x \end{cases} \tag{4.8}$$

Propriedade 3

Ao multiplicarmos todos os dados por uma constante não nula temos que:

$$y_i = kx_i \Rightarrow \begin{cases} \Delta_y = |k|\ \Delta_x \\ DMA_y = |k|\ DMA_x \\ \sigma_y^2 = k^2\sigma_x^2 \\ \sigma_y = |k|\ \sigma_x \end{cases} \tag{4.9}$$

Note que é razoável que apareça o módulo da constante, já que as medidas de dispersão são não negativas.

Atividade 4.4

Se o desvio padrão das temperaturas diárias de uma determinada localidade é de $5,2°F$, qual é o desvio padrão em graus Celsius? Lembre-se que a relação entre as duas escalas é

$$C = \frac{5}{9}(F - 32)$$

Medidas de dispersão para distribuições de freqüências agrupadas

Variância

Na aula passada, vimos que, em uma tabela de freqüências agrupadas, perdemos a informação sobre os valores individuais e isso nos obriga a tomar o ponto médio de cada classe como representante da respectiva classe. Vamos ver, agora, como calcular a variância para dados agrupados. Mais uma vez, vamos considerar os dados referentes aos salários dos funcionários do Departamento de Recursos Humanos, cuja distribuição é dada na **Tabela 4.1**.

Tabela 4.1: Distribuição da renda dos funcionários do Departamento de RH

Classe de renda	Ponto médio	Freqüência Simples		Freqüência Acumulada	
		Absoluta	Relativa %	Absoluta	Relativa %
[3200,4021)	3610,5	4	26,67	4	26,67
[4021,4842)	4431,5	2	1,33	6	40,00
[4842,5663)	5252,5	2	1,33	8	53,33
[5663,6484)	6073,5	3	20,00	11	73,33
[6484,7305)	6894,5	4	26,67	15	100,00
Total		15	100,00		

Como já dito, a interpretação da tabela de freqüências nos diz que há 4 observações iguais a 3610,5; 2 observações iguais a 4431,5; 2 iguais a 5252,5; 3 iguais a 6073,5 e 4 iguais a 6894,5. Logo, para calcular a variância desses dados basta usar uma das fórmulas 4.4 ou 4.6.

Usando (4.4), a variância é calculada como:

$$\begin{aligned}
\sigma^2 &= \frac{1}{15} \times \left[\begin{array}{c} 4 \times (3610,5 - 5307,2333)^2 + 2 \times (4431,5 - 5307,2333)^2 \\ +2 \times (5252,5 - 5307,2333)^2 + 3 \times (6073,5 - 5307,2333)^2 \\ +4 \times (6894,5 - 5307,2333)^2 \end{array} \right] \\
&= \frac{4}{15} \times (3610,5 - 5307,2333)^2 + \frac{2}{15} \times (4431,5 - 5307,2333)^2 + \\
&\quad + \frac{2}{15} \times (5252,5 - 5307,2333)^2 + \frac{3}{15} \times (6073,5 - 5307,2333)^2 \\
&\quad + \frac{4}{15} \times (6894,5 - 5307,2333)^2 \\
&= 1659638,729
\end{aligned}$$

Note, na penúltima linha da equação anterior, que os desvios quadráticos de cada classe estão multiplicados pela freqüência relativa da classe. Dessa forma, chegamos à seguinte expressão para a variância de dados agrupados:

$$\sigma^2 = \sum f_i (x_i - \overline{x})^2 \tag{4.10}$$

onde x_i é o ponto médio da classe e f_i é a freqüência relativa.

Usando a Equação (4.6), a variância é calculada como:

$$\begin{aligned}
\sigma^2 &= \frac{1}{15} \times \left[\begin{array}{c} 4 \times 3610,5^2 + 2 \times 4431,5^2 + 2 \times 5252,5^2 \\ +3 \times 6073,5^2 + 4 \times 6894,5^2 \end{array} \right] - 5307,2333^2 \\
&= \left[\begin{array}{c} \frac{4}{15} \times 3610,5^2 + \frac{2}{15} \times 4431,5^2 + \frac{2}{15} \times 5252,5^2 \\ +\frac{3}{15} \times 6073,5^2 + \frac{4}{15} \times 6894,5^2 \end{array} \right] - 5307,2333^2 \\
&= 1659638,729
\end{aligned}$$

Note, na penúltima linha da equação anterior, que os quadrados dos pontos médios de cada classe estão multiplicados pela freqüência relativa da classe. Dessa forma, chegamos à seguinte expressão alternativa para a variância de dados agrupados:

$$\sigma^2 = \sum f_i x_i^2 - \overline{x}^2 \tag{4.11}$$

e mais uma vez, obtemos que a variância é a *média dos quadrados menos o quadrado da média*; a diferença é que aqui a média é uma média ponderada pelas freqüências das classes.

Desvio médio absoluto

Seguindo raciocínio análogo, obtemos que o desvio médio absoluto para dados agrupados é

$$DMA = \sum f_i \,|\, x_i - \overline{x} \,|$$

que é uma média ponderada dos desvios absolutos em torno da média.

Atividade 4.5

Calcule a variância e o desvio médio absoluto para a distribuição dada na seguinte tabela, que foi analisada na Atividade 6 da aula anterior:

Classes	Freqüência
$4 \vdash 6$	10
$6 \vdash 8$	12
$8 \vdash 10$	18
$10 \vdash 12$	6
$12 \vdash 14$	4
Total	40

Resumo da Aula

Nesta aula, você estudou as principais medidas de dispersão, que medem a variabilidade dos dados. Seja $x_1, x_2, \ldots, x_n$ o nosso conjunto de dados.

- Amplitude - é a distância entre o maior e o menor valor:

$$\Delta_{total} = V_{Máx} - V_{Mín} = x_{(n)} - x_{(1)}$$

- Desvio em torno da média:

$$d_i = x_i - \overline{x}$$

para qualquer conjunto de dados, $\sum_{i=1}^{n} d_i = 0$

- Desvio médio absoluto:

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

- Variância: desvio quadrático médio

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

- Desvio padrão: raiz quadrada da variância

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

Exercícios

1. *Continuação do Exercício 3.3* Em uma pesquisa sobre atividades de lazer realizada com uma amostra de 20 alunos de um campus universitário, perguntou-se o número de horas que os alunos gastaram "navegando" na Internet na semana anterior. Os resultados obtidos foram os seguintes:

15	24	18	8	10	12	15	14	12	10
18	12	6	20	18	16	10	12	15	9

   Calcule a amplitude, o desvio médio absoluto e o desvio padrão desses dados, especificando as respectivas unidades.

2. *Continuação do Exercício 3.4* No final do ano 2005, o dono de um pequeno escritório de administração deu a seus 8 funcionários uma gratificação de 250 reais, paga junto com o salário de dezembro. Se em novembro o desvio padrão dos salários desses funcionários era de 180 reais, qual o desvio padrão dos salários em dezembro? Que propriedades você utilizou para chegar a esse resultado?

3. *Continuação do Exercício 3.5* No mês de dissídio de determinada categoria trabalhista, os funcionários de uma empresa tiveram reajuste salarial de 8,9%. Se no mês anterior ao dissídio o desvio padrão dos salários desses funcionários era de 220 reais, qual o valor do desvio padrão dos salários depois do reajuste? Que propriedades você utilizou para chegar a esse resultado?