Medidas de Posição ou Tendência Central

As medidas de posição ou medidas de tendência central indicam um valor que melhor representa todo o conjunto de dados, ou seja, dão a tendência da concentração dos valores observados. As principais medidas de posição são: a média, a mediana e a moda.

Média

Média da população (μ, lê-se mi): a média de população finita é encontrada somando-se todos os valores da população e dividindo-se pelo tamanho N da mesma.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Exemplo: *Calcule o salário médio dos seis empregados de uma pequena empresa:*

R$ 860,00 – R$ 750,00 – R$ 980,00 – R$ 1.200,00 – R$ - R$ 790,00 – R$ 950,00

Solução: *Como a empresa possui apenas seis empregados podemos calcular a média dos salários considerando todos os empregados. Neste caso, estamos trabalhando com toda a população.*

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{860 + 750 + 980 + 1200 + 790 + 950}{6} = \frac{5530}{6} = 921{,}67$$

O salário médio nessa empresa é de μ = R$ 921,67

A forma de calcular a média é a mesma tanto para uma população como para uma amostra, mas usamos uma notação diferente, para indicar que estamos trabalhando com uma amostra. O número de elementos em uma amostra é denotado por **n** e a média da amostra por $\bar{x}$.

Média da amostra ($\bar{x}$, lê-se x-barra): a média de um conjunto de dados é a medida de tendência central encontrada pela soma de todos os valores, e esta soma é dividida pelo número total de valores. A média é considerada o ponto de equilíbrio no conjunto de dados. Se

as observações em uma amostra de tamanho n são $x_{1+} x_{2+} \ldots + x_n$, então, a média amostral é calculada pela seguinte expressão:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \quad \text{que pode ser representada por} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

onde $\mathbf{x_i}$ é o valor da observação i, **n** o número de observações e **Σ** a letra sigma maiúscula do alfabeto grego que, na fórmula, indica o símbolo de somatório.

Exemplo*: Para verificar se o fabricante está respeitando o peso indicado nas embalagens de feijão, selecionou-se uma amostra de cinco pacotes. Foram obtidos os seguintes pesos em cada pacote da amostra: 495 g – 508 g – 492 g – 500 g – 496 g – 489 g. Calcule a média de peso das embalagens de feijão.*

Solução:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{495 + 507 + 492 + 500 + 496}{5} = \frac{2490}{5} = 498$$

Como ressaltado acima podemos observar que a forma de obter o valor da média da amostra é a mesma forma de obter a média populacional, diferindo apenas nas nomenclaturas. Para o "estimador" da média da amostra usamos $\bar{x}$, para o parâmetro da média da população usamos μ; para o número de elementos da amostra usamos "n", e para o número finito de elementos de uma população usamos "N". Tudo que é calculado, a partir de uma amostra, é chamado de estimador, e o que é calculado baseado em toda a população é chamado de parâmetro.

A importância na distinção das nomenclaturas das estatísticas calculadas com base em dados amostrais e populacionais se dá pelo fato de que a estimativa amostral é variável, pois depende da amostra coletada. O parâmetro populacional é constante, pelo menos, até que a população mude.

Nem sempre a média é uma boa medida de posição dos dados, pois ela é influenciada por valores extremos. Observando a tabela 1, podemos verificar que a turma A apresenta uma distribuição simétrica, ou seja, os valores estão distribuídos de forma homogênea em torno do centro do conjunto de dados, nesse caso, a média é uma boa medida de tendência central. A turma B apresenta um valor extremo que desvia a média mais para a esquerda do conjunto de dados.

Tabela 1 - Na tabela abaixo, são apresentadas as notas de 9 alunos de duas turmas.

Turma	**Notas dos alunos**									$\overline{x}$
A	7	7,5	7,5	8	8	8	8,5	8,5	9	8
B	0	7	7,5	7,8	8	8	8,2	8,5	9	7,1

Neste caso, a mediana é mais indicada como medida de tendência central, pois ela reflete melhor a tendência dos dados. Veja a tabela 2

Tabela 2 - Comparação entre média e mediana.

Turma	**Notas dos alunos**									$\overline{x}$	**Me**
A	7	7,5	7,5	8	8	8	8,5	8,5	9	8	8
B	0	7	7,5	7,8	8	8	8,2	8,5	9	7,1	8

Essa é uma das diferenças marcantes entre a mediana e a média. Vejamos agora como obter a mediana de um conjunto de dados.

Mediana

Mediana (Me): é o valor cuja posição separa o conjunto de dados em duas partes iguais, metade do número de elementos está acima do valor mediano e a outra metade abaixo do valor mediano. Para obter o valor mediano de uma distribuição de dados, primeiro ordene os valores. Isso poderá ser feito tanto em ordem crescente quanto em ordem decrescente. Depois, determine a posição que o valor mediano ocupa pela seguinte expressão:

$$\text{Pos Me} = \frac{n+1}{2}$$

Esta fórmula não fornece o valor mediano, mas sim sua localização no conjunto de dados. A forma de determinar o valor mediano depende se o número de observações que compõe o conjunto de dados é par ou ímpar.

- Número ímpar de elementos: o valor mediano é a observação que ocupa a posição (n+1)/2 desse conjunto de dados.

Exemplo: *Vejamos como calcular a mediana das notas da Turma A da tabela anterior.*

Notas da turma A	7	7,5	7,5	8	8	8	8,5	8,5	9

Solução: *Como os dados já estão ordenados podemos partir para determinar a posição do valor mediano. O número de observações é ímpar (n=9), logo o valor mediano é a observação central desse conjunto de dados. Pela fórmula da posição da mediana, tem-se:*

$$\text{Pos Me} = \frac{n+1}{2} \quad \rightarrow \quad \text{Pos Me} = \frac{9+1}{2} = 5^{\text{ª}}\ \textit{posição} \quad \rightarrow \quad \textit{Me = 8}$$

O valor que se encontra na 5ª posição é o oito. O número oito possui quatro observações à sua esquerda e quatro observações à sua direita, ou seja, 50% dos valores do conjunto de dados são inferiores a oito e 50% dos valores são superiores a oito.

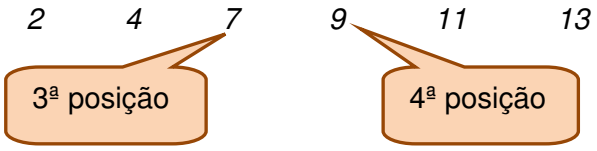
- Número par de elementos: quando o número de observações no conjunto de dados é par, a posição (n+1)/2 não será um número inteiro. A mediana será dada pela média aritmética das duas observações centrais dos dados ordenados.

Exemplo: *Calcule a mediana do seguinte conjunto de dados: 2, 4, 7, 9, 11, 13.*

Solução: *como o número de observações n=6 é par, não existe um valor central. Pela fórmula da posição da mediana, tem-se:*

$$\text{Pos Me} = \frac{6+1}{2} = \ 3{,}5^{\text{ª}}\ \textit{posição.}$$

O valor mediano está entre a 3ª e a 4ª posição. Nesses casos, o valor mediano não será um dos valores da distribuição e sim a média aritmética dos valores que se encontram nessas duas posições. A terceira posição é ocupada pelo valor sete e a quarta posição é ocupada pelo valor nove.



$$\text{Me} = \frac{7+9}{2} = 8$$

A mediana é o valor oito (Me = 8). Este valor possui três observações à sua esquerda e três observações à sua direita, ou seja, 50% dos valores do conjunto de dados são inferiores a oito e 50% dos valores são superiores a oito.

A vantagem da mediana é que ela não é influenciada por valores extremos, pois ela depende da posição e não dos valores das observações no conjunto de dados.

Moda

Moda (Mo): é definida como sendo aquele valor ou aqueles valores que ocorrem com maior freqüência.

Exemplo: Determine a moda de cada conjunto de dados abaixo:

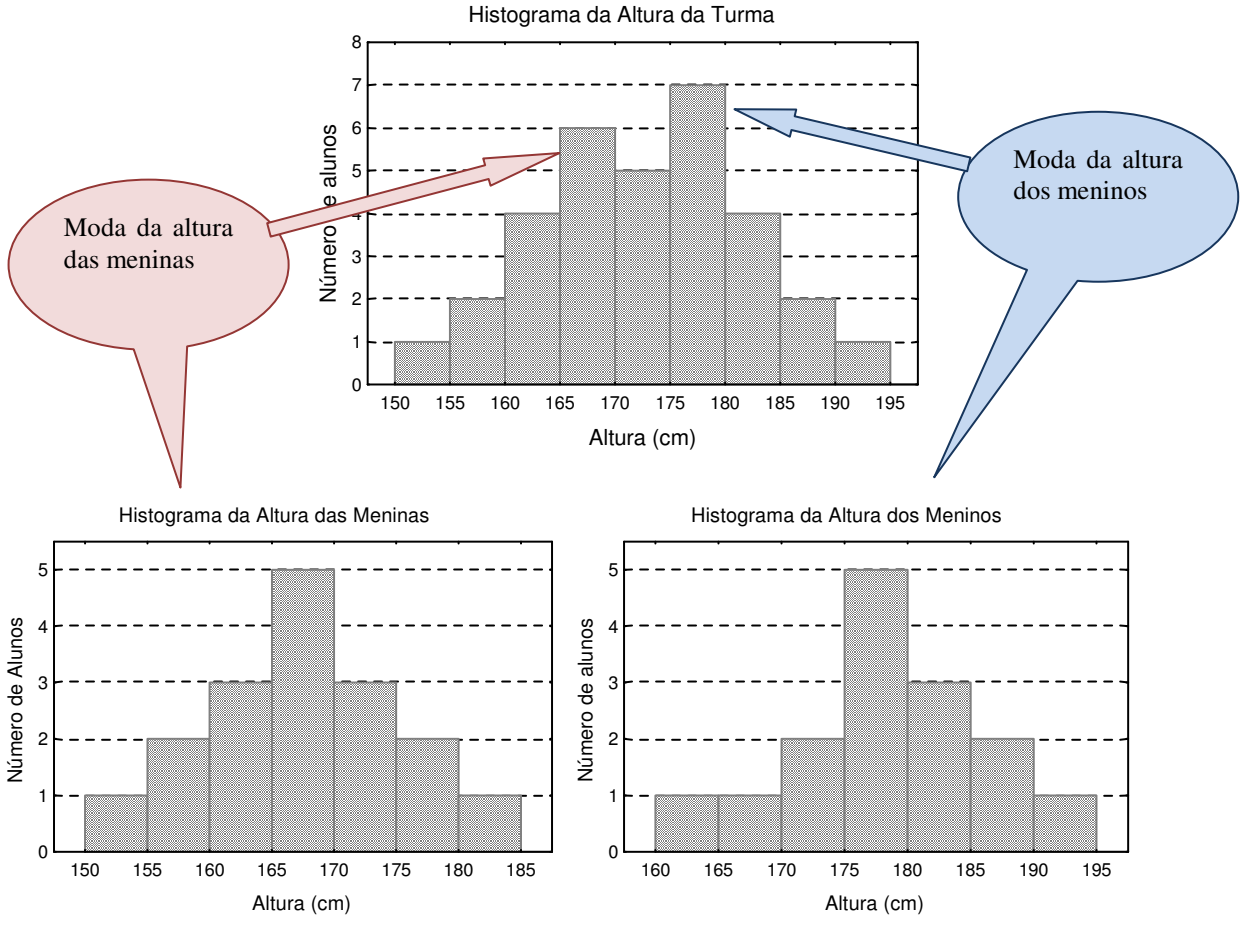
a-) 26, 39, 39, 41, 41, 41, 43, 47 ➔ Mo = 41, e a distribuição é unimodal, pois possui apenas uma moda.

b-) 12, 17, 17, 23, 25, 25, 32 ➔ esta distribuição apresenta duas modas, $Mo_1 = 17$ e $Mo_2 = 25$, sendo denominada de distribuição bimodal.

c-) 22, 25, 27, 29, 33, 35, 42 ➔ este conjunto de dados não possui moda, pois todos os valores ocorrem o mesmo número de vezes. Nesse caso, dizemos que o conjunto de dados apresenta uma distribuição amodal.

Quando o conjunto de dados apresentar três modas, denomina-se trimodal, e quatro ou mais, multimodal.

Nos gráficos da figura abaixo, é apresentado o Histograma da variável altura de uma turma de alunos. No primeiro gráfico, observamos dois picos, ou seja o conjunto de dados apresenta altas freqüências em dois intervalos de dados. Separando a turma de alunos pelo sexo: masculino e feminino; podemos observar que para o sexo feminino a moda encontra-se entre 165 cm e 170 cm, já para os meninos a moda está entre 175 e 180 cm. Portanto, quando um histograma apresentar dois ou mais picos, significa que os dados apresentam modas diferentes dependendo da forma que foram agrupados. Também podemos concluir que quando a distribuição apresentar mais de uma moda, o histograma terá mais de um pico.



Medidas de Dispersão

As medidas de dispersão são medidas estatísticas que caracterizam o quanto um conjunto de dados está disperso em torno de sua tendência central.

Não há razão alguma para se calcular a média de um conjunto de dados, onde não haja variação desses elementos (Exemplo: 6 6 6 6 , a média = 6). No entanto, se a variabilidade dos dados for muito grande, sua média terá um grau de confiabilidade tão pequeno que será inútil calculá-la. Portanto, não é possível analisar um conjunto de dados apenas através de uma medida de tendência central, também é necessário analisar de que forma os valores observados estão espalhados em torno de seu centro. Além disso, dois conjuntos de dados podem possuir a mesma média e, no entanto, os valores podem estar distribuídos de forma diferente. Por exemplo, considere os resultados das notas de oito alunos de duas turmas:

Exemplo 1

Tabela 3 – Notas de oito alunos de duas turmas

Turma A	0	2	4	5	5	6	8	10
Turma B	4	4,5	5	5	5	5	5,5	6

Embora as duas turmas de alunos possuam a mesma média, 5, diferem bastante na variabilidade das notas. Enquanto a turma A apresenta notas mais dispersas, na turma B, observam-se pequenas variações nas notas obtidas pelos alunos. Dessa forma, para descrever adequadamente um conjunto de dados, além de uma medida que descreva sua tendência central, é necessário uma medida que descreva sua dispersão.

Para avaliar o grau de dispersão ou variabilidade dos valores de um conjunto de dados, usaremos dois tipos de medidas de dispersão: **absoluta** (desvio médio, variância e desvio padrão) e **relativa** (coeficiente de variação de Pearson).

Amplitude total

Para uma rápida medida da variabilidade, podemos calcular a amplitude total (AT), que é a diferença entre o mais alto e o mais baixo valor em uma distribuição.

$$AT = V_{max} - V_{min}$$

A amplitude total considera apenas o valor máximo e o valor mínimo, ignorando todos os outros valores no conjunto de dados. Além disso, esses valores podem ser valores extremos

ou atípicos. Podemos aperfeiçoar nossa descrição da dispersão, através de outras medidas como o desvio médio.

Exemplo: *Calcule a média e a amplitude total para os dados abaixo:*

Conjunto de dados A ➔ 1 – 2 - 3 – 3 – 3 – 4 – 5 ➔ $\mu = 3$ e AT = 4

Conjunto de dados B ➔ 1 – 2- 3 – 4 – 5 ➔ $\mu = 3$ e AT = 4

Os dois conjuntos de dados possuem mesma média e AT, no entanto são diferentes. O conjunto de dados A apresenta maior número de dados em torno da média do que o conjunto de dados B. Com isso concluímos que a AT não é uma boa medida de dispersão, pois apenas avalia os valores que estão nos extremos.

Desvio Médio

Para levar em consideração todos os valores da distribuição, além dos extremos, subtrai-se a média aritmética de cada elemento do conjunto de dados e somam-se as diferenças, calculando, dessa forma, o desvio de cada elemento a média. Como essa soma é sempre igual a zero, pois alguns valores são negativos e outros positivos, considera-se apenas o módulo das diferenças.

Exemplo: Calcule o desvio médio para as notas dos alunos da Turma A da tabela abaixo.

Solução: *Primeiro calcula-se a média do conjunto de dados:*

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{0+2+4+5+5+6+8+10}{8} = \frac{40}{8} = 5$$

A seguir calcula-se o desvio médio. Podemos utilizar a própria tabela para desenvolver os cálculos e substituir o valor do somatório direto na fórmula ou podemos desenvolver os cálculos na própria fórmula.

Usando a tabela para desenvolver os cálculos:

Tabela 4 - Notas dos alunos da turma A.

Aluno	**Nota**	$(x_i - \mu)$	$\lvert x_i - \mu \rvert$
1	0	-5	5
2	2	-3	3
3	4	-1	1
4	5	0	0
5	5	0	0
6	6	+1	1
7	8	+3	3
8	10	+5	5
Σ	**40**	**0**	**18**

Como o desvio médio é definido como a média aritmética dos desvios em módulo, basta dividir o valor obtido na tabela do $\sum|x_i - \mu|$ pelo número de elementos do conjunto de dados.

$$DM = \frac{18}{8} = 2{,}25$$

Portanto, o desvio médio é definido pela seguinte expressão:

$$DM = \frac{\sum_{i=1}^{N}|x_i - \mu|}{N}$$

Desenvolvendo os cálculos na fórmula :

$$DM = \frac{\sum_{i=1}^{n}|x_i - \mu|}{N} = \frac{|0-5|+|2-5|+|4-5|+|5-5|+|5-5|+|6-5|+|8-5|+|10-5|}{8} =$$

$$DM = \frac{5+3+1+0+0+1+3+5}{8} = \frac{18}{8} = 2{,}25$$

Apesar de não ser muito utilizado na inferência estatística, o desvio médio é considerado uma boa medida de dispersão, quando o objetivo é apenas descrever um conjunto de dados. Entretanto, o uso de valores absolutos cria dificuldades algébricas nos métodos de inferência estatística. Assim, em vez de usarmos os valores absolutos, obtemos uma melhor medida de variação fazendo os desvios ao quadrado, obtendo-se a variância.

Variância

Para o cálculo da **variância**, ao invés de considerar o módulo da diferença, eleva-se esta diferença ao quadrado, eliminando-se, assim, o problema do sinal negativo sem fazer uso da função módulo. A variância é definida como a média aritmética dos quadrados dos desvios.

Variância populacional $\qquad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

Exemplo: Calcule a variância para as notas dos alunos da Turma A

Solução: *Utilizando a tabela para auxiliar no cálculo:*

Tabela 5 - Notas dos alunos da Turma A.

Turma A	**Nota**	$(x_i - \mu)$	$(x_i - \mu)^2$
1	0	-5	25
2	2	-3	9
3	4	-1	1
4	5	0	0
5	5	0	0
6	6	+1	1
7	8	+3	9
8	10	+5	25
Total	**40**	**0**	**70**

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{70}{8} = 8{,}75$$

Desenvolvendo os cálculos na fórmula:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{(0-5)^2 + (2-5)^2 + (4-5)^2 + (5-5)^2 + (5-5)^2 + (6-5)^2 + (8-5)^2 + (10-5)^2}{8} =$$

$$\sigma^2 = \frac{25+9+1+0+0+1+9+25}{8} = \frac{70}{8} = 8{,}75$$

Usamos o símbolo σ^2 para representar a variância calculada com base em dados em todos os elementos da população, portanto, a variância populacional é um parâmetro. Quando usamos uma amostra para calcular a variância, o símbolo usado é s^2, a variância amostral é um estimador da variância populacional.

Variância amostral $\quad s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$

Se fosse usado n no denominador da fórmula da variância amostral, estaríamos estimando uma medida de variabilidade menor do que a variância da população. Portanto, no cálculo da variância amostral, utiliza-se (n-1) no denominador.

Exemplo*: Para verificar se o fabricante está respeitando o peso indicado nas embalagens de feijão, selecionou-se uma amostra de cinco pacotes. Foram obtidos os seguintes pesos em cada pacote da amostra: 495 g – 508 g – 492 g – 500 g – 496 g – 489 g*

Calcule a variância do peso das embalagens de feijão.

Solução: *No exemplo da média de uma amostra já calculamos a média deste conjnto de dados que é* $\bar{x} = 498\ gramas$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(495-498)^2 + (508-498)^2 + (492-498)^2 + (500-498)^2 + (496-498)^2 + (489-498)^2}{6-1} =$$

$$s^2 = 44{,}6667\,g^2$$

Desvio padrão

A desvantagem da variância consiste no fato de suas unidades normalmente não terem sentido. A variância para as notas dos alunos, por exemplo, é medida em "notas ao quadrado". No exemplo acima a variância do peso dos pacotes de feijão estão em "gramas ao quadrado". Pode-se retomar a unidade original dos dados, extraindo-se a raiz quadrada da variância, denominada de desvio padrão. O Desvio Padrão é definido como a raiz quadrada da média aritmética dos quadrados dos desvios, ou seja, a raiz quadrada da variância.

Desvio padrão populacional $\quad \sigma = \sqrt{\dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$

Desvio padrão amostral $\quad s = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

O desvio padrão calculado usando todos os elementos da população é simbolizado por σ, o desvio padrão populacional é um parâmetro. Se o desvio padrão é calculado a partir de uma amostra, este é representado pelo símbolo s, chamado desvio padrão amostral e é considerado uma estimativa.

Exemplo 1: Calcule o desvio padrão para as notas dos alunos da Turma A.

Solução: Como no exemplo anterior a variância já foi calculada, basta extrair a raiz quadrada da variância:

$$\sigma = \sqrt{\sigma^2} = \sqrt{8{,}75} = 2{,}958$$

Exemplo 2: Calcule o desvio padrão para os seguintes dados amostrais:

25 – 26 – 33 – 21 – 30

Solução: Como trata-se de uma amostra pequena usaremos a seguinte fórmula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(25-27)^2 + (26-27)^2 + (33-27)^2 + (21-27)^2 + (30-27)^2}{5-1}} =$$

$$= \sqrt{\frac{4+1+36+36+9}{4}} = \sqrt{21{,}5} = 4{,}64 \;\rightarrow\; \mathbf{s = 4{,}64}$$

Coeficiente de variação de Pearson

O **coeficiente de variação de Pearson** (CV) é uma medida de dispersão relativa que mede a dispersão dos dados em relação à média aritmética. É calculado, dividindo-se o desvio padrão pela média, multiplicando-se por 100, para expressar o resultado em porcentagem, em vez de se utilizar a unidade de medida da variável em análise.

População ➔ $CV = \frac{\sigma}{\mu}.100$ Amostra ➔ $CV = \frac{s}{\bar{x}}.100$

Exemplo 1: *Calcule o coeficiente de variação de Pearson das notas dos alunos da turma A e B do exemplo anterior.*

Turma A	0	2	4	5	5	6	8	10
Turma B	4	4,5	5	5	5	5	5,5	6

$CV_A = \frac{2{,}96}{5}.100 = 59{,}2\%$ $CV_B = \frac{0{,}56}{5}.100 = 11{,}2\%$

A turma B apresenta menor dispersão relativa do que a turma A, o que indica que o desempenho dos alunos da turma B foi mais homogêneo.

A dispersão relativa também permite comparar duas ou mais distribuições, mesmo que essas se refiram a diferentes fenômenos e sejam expressas em unidades de medida diferentes.

Exemplo 2: *Na tabela abaixo, são apresentados os valores do desvio padrão e da média da altura e peso de um grupo de pessoas.*

	Média	Desvio padrão
Altura	174 cm	7 cm
Peso	78 kg	12 kg

Embora a diferença nas unidades de medida torne impossível comparar o desvio padrão de 7 cm com o desvio padrão de 12 kg, podemos comparar os coeficientes de variação, que não têm unidades de medida. A variável altura apresenta CV = 4% e a variável peso CV = 15,4%. Portanto, a variável peso apresenta maior dispersão relativa do que a variável altura.

Observação: Para facilitar a interpretação do coeficiente de variação, usaremos os seguintes intervalos:

$CV \geq 30\%$ ➔ Alta dispersão

$15\% < CV < 30\%$ ➔ Média dispersão

$CV \leq 15\%$ ➔ Baixa dispersão

No exemplo 1, podemos verificar que a turma A apresenta alta dispersão e a turma B baixa dispersão.

$$CV_A = \frac{2{,}96}{5}.100 = 59{,}2\% \qquad CV_B = \frac{0{,}56}{5}.100 = 11{,}2\%$$

Exemplos de aplicação: *Antes de comprar uma bateria para seu celular você analisa as informações estatísticas fornecidas pelo fabricante com relação ao tempo de duração das mesmas. Explique de que forma essas informações podem auxiliar você na sua escolha.*

Bateria	A	B	C
Média	600 ciclos	650 ciclos	600 ciclos
Mediana	600 ciclos	700 ciclos	500 ciclos
Desvio Padrão	50 ciclos	150 ciclos	250 ciclos

Solução: *Na tabela acima podemos verificar a bateria A apresenta a menor dispersão e média e mediana são iguais o que indica não haver valores extremos nem para a direita nem*

para a esquerda. A bateria B apesar de apresentar média maior do que a bateria A e C apresentam maior mediana o que indica tempos de duração da bateria afastados da média para valores menores, o que explica o maior desvio padrão. Já a bateria C possui a maior dispersão de tempo de duração, sendo a menos indicada.

Quadro Resumo das Medidas de Posição

Medida	Definição	Vantagens	Desvantagens
Média	$\bar{x} = \frac{\sum x_i}{n}$	Usada em muitos métodos estatísticos	- Afetada por valores extremos
Mediana	Valor central	- Apropriada quando há valores extremos ou distribuições assimétricas. - Sempre existe	- Usada em poucos métodos estatísticos
Moda	Valor mais freqüente	- Apropriada para dados qualitativos.	- Nem sempre existe. - Pode haver mais de uma moda. - Não se presta à análise matemática

Quadro Resumo das Medidas de Dispersão

Desvio médio populacional

$$DM = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

Variância Populacional

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Desvio padrão populacional

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Coeficiente de Variação Populacional

$$CV = \frac{\sigma}{\mu}.100$$

Variância amostral

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

Desvio padrão amostral

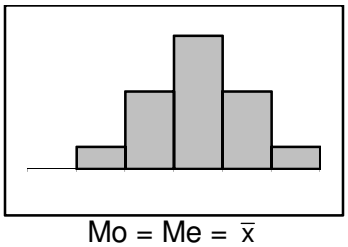
$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Coeficiente de Variação Amostral

$$CV = \frac{s}{\bar{x}}.100$$

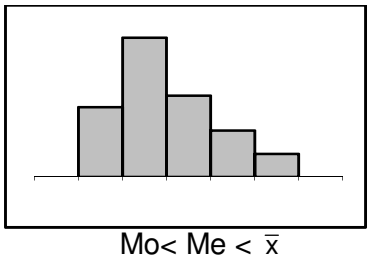
Medida de Assimetria

Além das medidas de posição e dispersão a forma da distribuição é uma importante fonte de informação sobre o comportamento dos dados. Algumas distribuições podem apresentar forma simétrica ou assimétrica.



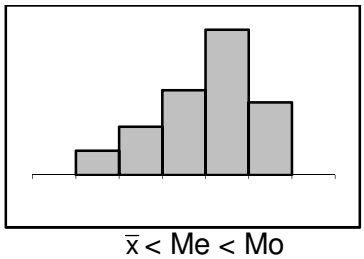
Mo = Me = $\bar{x}$
As = 0 ➔ simétrica

Figura 1(a)



Mo< Me < $\bar{x}$
As > 0 ➔ assimetria positiva

Figura 1(b)



$\bar{x}$ < Me < Mo
As < 0 ➔ assimetria negativa

Figura 1(c)

Uma distribuição é simétrica, Figura 1(a), se o gráfico apresenta o mesmo comportamento a direita e a esquerda da média. Neste caso, moda, média e mediana são iguais ou muito próximas. A distribuição apresenta uma assimetria positiva, Figura 1(b), quando sua cauda direita afasta-se mais do pico do que a cauda esquerda, o que faz com que a média seja maior do que a mediana e esta maior do que a moda. A distribuição de dados apresenta uma assimetria negativa, Figura 1(c), quando sua cauda esquerda afasta-se mais do pico do que a cauda direita, sendo a moda maior do que a mediana e esta maior do que a média.

Existe mais de uma forma de determinar o coeficiente de assimetria, aqui usaremos o **coeficiente de assimetria** pelo segundo critério de Pearson:

$$As = \frac{3.(\bar{x} - Me)}{s}$$

Tabela 6 - Interpretação do Coeficiente de Assimetria de Pearson

Assimétrica negativa	As ≤ - 1
Assimétrica negativa moderada	-1 < As < -0,15
Simétrica	-0,15 < As < +0,15
Assimétrica positiva moderada	+0,15 < As < +1
Assimétrica positiva	As ≥ 1

Exemplo: Os dados a seguir se referem a uma amostra dos salários recebidos em uma determinada empresa. Determine o coeficiente de assimetria e interprete.

830 920 920 1020 1100 1150 1300 1340 2600 2950

Solução: Para determinar o coeficiente de assimetria precisamos calcular a média, mediana e o desvio padrão.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1413 \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = 740{,}65$$

$$\text{Pos Me} = \frac{n+1}{2} = \frac{10+1}{2} = 5{,}5^{\text{ª}} \text{ posição} \qquad \text{Me} = \frac{1100+1150}{2} = 1125$$

$$As = \frac{3.(\bar{x} - Me)}{s} = \frac{3.(1413-1125)}{740{,}65} = 1{,}17$$

Os salários dos empregados apresentam assimetria positiva (As=1,17). Este resultado evidência a presença de valores extremos a direita do conjunto de dados que puxam a média para cima, mas não influenciam a mediana.

Escore z ou variável padronizada

Usando a média e o desvio padrão podemos determinar a posição relativa de qualquer valor do conjunto de dados. O escore z ou variável padronizada representa o número de desvios padrões que um dado valor está afastado da média. Para obter o escore z, use a seguinte fórmula:

$$\text{Amostra} \rightarrow z = \frac{x - \bar{x}}{s} \qquad \text{População} \rightarrow z = \frac{x - \mu}{\sigma}$$

Um escore *z* pode ser negativo, positivo ou zero. Se *z* é negativo, o valor *x* correspondente está abaixo da média. Se *z* é positivo, o valor *x* correspondente está acima da média. E se *z*=0, o valor *x* correspondente é igual à média. Quanto maior o valor do escore z melhor é a posição relativa da observação no conjunto de dados.

Exemplo: Um estudante obteve nota 7,2 em Estatística e 8,0 em Álgebra. Determine em que disciplina o aluno obteve melhor posição relativa. A média da turma em Estatística foi 6,4 com desvio padrão 1,2 e a média da turma em Álgebra foi 7,6 com desvio padrão de 1,6.

Solução:

$$z_{\text{Estatística}} = \frac{x - \mu}{\sigma} = \frac{7{,}2 - 6{,}4}{1{,}2} = 0{,}67 \qquad z_{\text{Álgebra}} = \frac{x - \mu}{\sigma} = \frac{8{,}0 - 7{,}6}{1{,}6} = 0{,}25$$

Apesar da média em Estatística ser menor do que a média obtida pelo aluno na disciplina de Álgebra, ele obteve melhor posição relativa em Estatística, pois o valor do escore z desta disciplina é maior do que o escore z em Álgebra. Comparando a nota do aluno com a média e o desvio padrão da turma verifica-se que em Estatística sua nota está mais afastada da média (para cima) e com menor dispersão do que a nota obtida em Álgebra.

Coeficiente de Correlação

Até o momento estudamos apenas uma única variável. No entanto, algumas vezes o pesquisador está interessado em analisar a relação entre duas variáveis. Por exemplo, qual a relação entre o preço dos alimentos e a oferta; entre a altura das pessoas e o peso. A maneira mais simples de iniciar este estudo é através do diagrama de dispersão. Neste gráfico, cada eixo representa uma das variáveis em estudo.

Exemplo: *No gráfico abaixo é apresentado o diagrama de dispersão entre as variáveis altura e peso de um grupo de oito pessoas.*

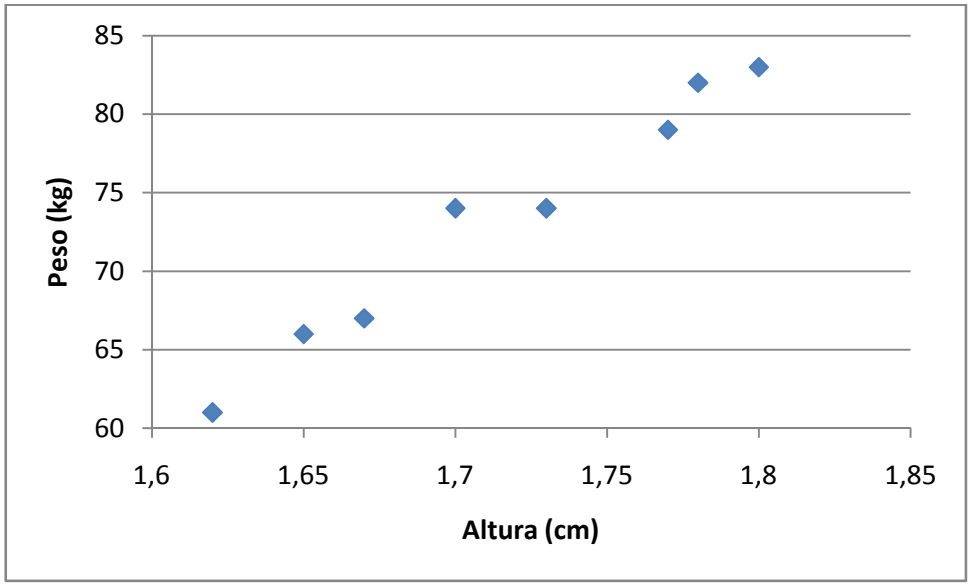
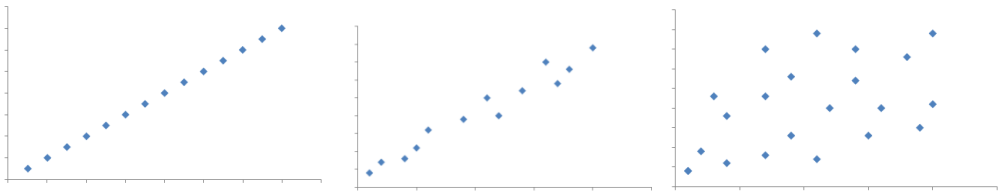


Figura 2 – Correlação entre as variáveis peso e altura

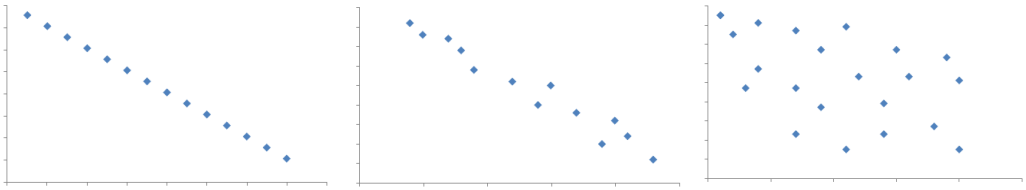
O diagrama de dispersão pode apresentar uma relação positiva entre as variáveis, ou seja, à medida que uma variável aumenta a outra variável também aumenta. Na figura 3 (a) as variáveis apresentam uma correlação perfeita, na figura 3(b) uma correlação forte e na figura 3(c) uma correlação fraca.



(a) (b) (c)

Figura 3 – correlação positiva entre as duas variáveis

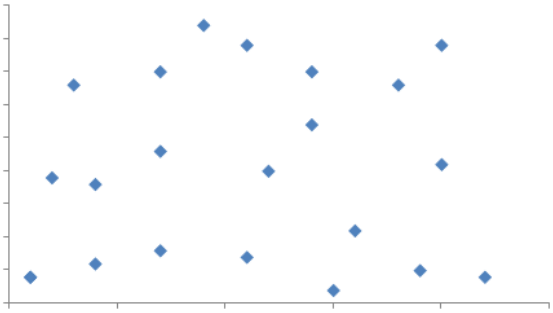
O diagrama de dispersão também pode apresentar uma relação negativa entre as variáveis, ou seja, à medida que uma variável aumenta a outra variável diminui. Na figura 4(a) as variáveis apresentam uma correlação perfeita, na figura 4(b) uma correlação forte e na figura 4(c) uma correlação fraca.



(a) (b) (c)

Figura 4 – correlação negativa entre as duas variáveis

Quando as duas variáveis não possuem relação o gráfico de dispersão apresenta uma nuvem aleatória de pontos.



O grau de associação linear entre duas variáveis pode ser medido através do coeficiente de correlação de Pearson, que é dado por:

$$r = \frac{n\sum xy - \left(\sum x\right).\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right].\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

O valor de r, que sempre pertencerá ao intervalo [-1, 1], representa uma medida de intensidade do inter-relacionamento entre duas variáveis. Se r = 1, há uma perfeita correlação positiva entre as variáveis, isto é, se os valores de uma variável aumentam (ou diminuem), em correspondência os valores da outra variável também aumentam (ou diminuem) na mesma proporção. Se, por outro lado, r = -1, há uma perfeita correlação negativa entre as variáveis, ou seja, os valores de uma variável variam em proporção inversa aos valores de outra variável. Se, entretanto, r = 0, não há correlação entre as variáveis.

Exemplo 1: *Calcule o coeficiente de correlação entre as variáveis altura e peso de oito pessoas.*

Altura (cm)	Peso (kg)
1,7	74
1,65	66
1,62	61
1,73	74
1,78	82
1,67	67
1,8	83
1,77	79

Solução: *Utilizando a tabela para auxiliar nos cálculos*

Altura (cm) - x	Peso (kg) - y	x.y	X^2	Y^2
1,7	74	125,8	2,89	5476
1,65	66	108,9	2,72	4356
1,62	61	98,82	2,62	3721
1,73	74	128,02	2,99	5476
1,78	82	145,96	3,17	6724
1,67	67	111,89	2,79	4489
1,8	83	149,4	3,24	6889
1,77	79	139,83	3,13	6241
$\sum x$ =13,72	$\sum y$ = 586	$\sum x.y$ =1.008,62	$\sum x^2$ = 23,56	$\sum y^2$ = 43372

$$r = \frac{n\sum xy - (\sum x).(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right].\left[n\sum y^2 - (\sum y)^2\right]}} = \frac{8.1008{,}62 - 13{,}72.586}{\sqrt{\left[8.23{,}56 - 13{,}72^2\right].\left[8.43372 - 586^2\right]}} = 0{,}987$$

Interpretação: *As variáveis altura e peso apresentam uma forte relação positiva, portanto, à medida que uma variável aumenta a outra também aumenta.*