7. 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. A **média aritmética** é uma das principais medidas de posição, cuja aplicação é a mais usada. A **mediana** é o valor central de um rol, ou seja, é o valor que fica no meio da seqüência, quando os dados são arranjados na ordem crescente. A **moda** é definida como sendo aquele valor ou aqueles valores que ocorrem com maior freqüência. Evidentemente, um conjunto de valores pode não apresentar moda, sendo, então, denominado amodal.

7.1 Média

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**: Calcule a média dos seguintes dados: 2, 4, 6, 7, 11*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+4+6+7+11}{5} = 6$$

A forma de calcular a média é a mesma tanto para uma amostra como para uma população finita, mas usamos uma notação diferente, para indicar que estamos trabalhando

com uma população. O número de elementos em uma população é denotado por N e a média da população por μ.

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.

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

A média é influenciada por valores extremos. No exemplo abaixo, podemos verificar que a mediana depende da posição, e não dos valores dos elementos na série ordenada. Essa é uma das diferenças marcantes entre a mediana e a média. Além disso, a média nem sempre representa a tendência central dos dados, como na série 2.

Exemplo:

Série 1: 5, 7, 10, 13, 15 - $\bar{x} = 10$ e Me = 10

Série 2: 5, 7, 10, 13, 65 - $\bar{x} = 20$ e Me = 10

Vejamos outro exemplo, na tabela 7.1, 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. Neste caso, a mediana é mais indicada como medida de tendência central, pois ela reflete melhor a tendência dos dados.

Tabela 7.1 - Na tabela abaixo, são apresentadas as notas de 9 alunos de três turmas.

Turma	Notas dos alunos									$\bar{x}$	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

7.2 .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: *Calcule a mediana do seguinte conjunto de dados: 2, 4, 6, 7, 11*

Solução: o número de observações n=5 é ímpar. O valor mediano é a observação central desse conjunto de dados. Pela fórmula da posição da mediana, tem-se $\text{Pos Me} = \frac{5+1}{2} = 3ª$ posição. A mediana é o valor seis que se encontra na terceira posição no conjunto de dados. O número seis possui duas observações à sua esquerda e duas observações à sua direita, ou seja, 50% dos valores do conjunto de dados são inferiores a seis e 50% dos valores são superiores a seis.

- 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.

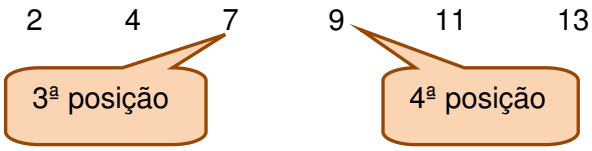
Exemplo: *Calcule a mediana do seguinte conjunto de dados: 2, 4, 7, 9, 11, 13.*

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

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

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.



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

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.

7.3 Moda

Moda (Mo): é o valor que ocorre com maior freqüência em um conjunto de dados.

Exemplo: Determine a moda de cada um dos conjuntos de dados:

a-) 2, 5, 7, 9, 13, 15, 22 ➔ 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.

b-) 16, 19, 19, 21, 21, 21, 23, 27 ➔ Mo = 21, e a distribuição é unimodal, pois possui apenas uma moda.

c-) 2, 7, 7, 13, 15, 15, 22 ➔ esta distribuição apresenta duas modas, $Mo_1 = 7$ e $Mo_2 = 15$, sendo denominada de distribuição bimodal.

Quando a distribuição apresentar mais de uma moda, o histograma terá mais de um pico. Quando o conjunto de dados apresentar três modas, denomina-se trimodal, e quatro ou mais, multimodal.

8. 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: 5 5 5 5 $\bar{x} = 5$) No entanto, se a variabilidade dos dados for muito grande, sua média terá um grau de confiabilidade tão pequeno que será inútil calculá-la, como discutido no capítulo anterior, na série 2 do exemplo. 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 8.1 – 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.

Exemplo 2

- Em duas cidades A e B, foram coletados dados sobre temperatura diária, durante um mês, verificando-se que a média das duas cidades foi de $\bar{x} = 20°C$. Em qual das duas cidades é mais agradável de se viver, em termos de clima? Apenas conhecendo a média da temperatura do ar, fica difícil de responder. Nesse caso, as medidas de dispersão podem nos auxiliar a verificar qual cidade possui menor variação em termos de temperatura. E, assim, apontar qual cidade possui o clima mais agradável.

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

8.1. 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.

8.2. 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. Portanto, o desvio médio é definido como a média aritmética dos desvios em módulo.

Desvio médio para uma amostra $$DM = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Desvio médio para uma população $$DM = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

Podemos observar que o equacionamento é o mesmo, tanto para a amostra quanto para a população. O que difere nas equações são as nomenclaturas. Para o "estimador" da média da amostra usamos $\bar{x}$, para o parâmetro, a 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 estimativa, 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 ela é variável,

pois depende da amostra coletada. O parâmetro populacional é constante, até que a população mude.

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

Solução: Podemos utilizar uma tabela para realizar os cálculos.

Tabela 8.2 - Notas dos alunos da turma A.

Aluno	Nota	$(x_i - \bar{x})$	$\lvert x_i - \bar{x} \rvert$
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
Total	**40**	**0**	**18**

Substituindo na fórmula, temos: $DM = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} = \frac{18}{8} = 2{,}25$

Também podemos resolver direto pela fórmula:

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

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

O desvio médio já não é tão usado, pois ele utiliza a função módulo que nem sempre é útil em análises estatísticas mais avançadas. 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 o conjunto de dados. Além disso, auxilia na compreensão de outras medidas de dispersão como a variância e o desvio padrão.

8.3. 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. A variância é definida como a média aritmética dos quadrados dos desvios.

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

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

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 uma estimativa.

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

Solução: Pela tabela,

Tabela 8.3 - 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**

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

Pela fórmula:

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

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

Em pequenas amostras, os elementos de um conjunto de dados tendem a ficar mais perto da sua média amostral do que da média populacional. 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 de pequenas amostras (n < 30), utiliza-se (n-1) no denominador. Para valores grandes de n (n ≥ 30), não há grande

diferença entre os resultados proporcionados pela utilização de qualquer dos dois divisores, n ou n – 1.

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

Se desenvolvermos o numerador da expressão sob o radical, chegaremos a uma fórmula mais prática da variância:

Variância para amostras pequenas $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$

Variância para amostras grandes $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ ou $s^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$

8.4. 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”. Pode-se retomar a unidade original dos dados, extraindo-se a raiz quadrada da variância, denominada de desvio padrão.

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

Desvio padrão para amostras grandes (n≥30) $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$

Desvio padrão para amostras pequenas (n<30) $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

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:

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

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

8.5. 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

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

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 parâmetros:

$CV \geq 30\%$ ➔ 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.

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

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

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

Desvio médio amostral

$$DM = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Variância para amostras grandes

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

Variância para amostras pequenas

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

Desvio padrão para amostras grandes (n≥30)

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

Desvio padrão para amostras pequenas (n<30)

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

Coeficiente de Variação Amostral

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