UNIDADE 3

Medidas de Posição e Dispersão

Objetivos específicos de aprendizagem

Ao finalizar esta Unidade, você deverá ser capaz de:

- Calcular e interpretar as medidas de posição média, moda, mediana;
- Entender como as medidas de posição influenciam na forma da distribuição dos dados;
- 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; e
- Calcular e interpretar resultados de medidas separatrizes.

MEDIDAS DE POSIÇÃO

Caro estudante,

A partir de agora, iremos conhecer uma nova forma de caracterizar um conjunto de observações. Para isso, você irá aprender novos conceitos de medidas de posição e de dispersão.

Para o entendimento dessas medidas de posição e de dispersão, serão utilizadas as duas situações apresentadas a seguir. Sempre que mencionarmos as situações, você deve vir até esta página para entender como estão sendo realizados os cálculos.

Preparado para mais esse desafio?

Então, vamos lá!

Vamos iniciar nossa discussão pelas duas situações que utilizaremos como base.

- Para facilitar um projeto de aplicação da rede de esgoto de certa região de uma cidade, os engenheiros da Prefeitura Municipal tomaram uma amostra de 52 ruas, (tamanho total da amostra ou a soma de todas as frequências absolutas) contando o número de casas por rua. Os dados referentes a uma pesquisa de mercado foram agrupados como segue na Tabela 11:

Tabela 11: Distribuição de frequências do número de casas por rua de certa região de uma cidade

Número de casas por rua	Frequência absoluta
0 |— 2	5
2 |— 4	7
4 |— 8	11
8 |— 12	16
12 |— 16	8
16 |— 20	5

Fonte: Elaborada pelo autor

- Taxa de efetivação da cobrança de um determinado tributo que se apresentava atrasado em uma prefeitura após uma campanha realizada para que ele fosse saldado. Esses resultados são diários, conforme mostra a Tabela 12.

Tabela 12: Taxa de efetivação da cobrança

44	46	51	54	54	55	56	56	56
58	59	60	61	61	61	62	63	63

Fonte: Elaborada pelo autor

É importante destacarmos ainda que as medidas de posição ou de tendência central constituem uma forma mais sintética de apresentar os resultados contidos nos dados observados, pois representam um valor central, em torno do qual os dados se concentram. As medidas de tendência central mais empregadas são a média, a mediana e a moda. A seguir, veremos cada uma delas.

Para você fazer cálculos de medidas de posição e de dispersão, utilize o programa estatístico Bioestat 5.0 e, também, planilhas eletrônicas visitando o *site*: <http://www.juliobattisti.com.br/tutoriais/celsonunes/openoffice007.asp>. Acesso em: 19 nov. 2010.

Média

Das três medidas de posição mencionadas, a **média aritmética** é a mais usada por ser a mais comum e compreensível delas e pela relativa simplicidade do seu cálculo, além de prestar-se bem ao tratamento algébrico.

É importante termos claro que a **média aritmética** ou simplesmente média de um conjunto de n observações, x_1, x_2, ..., x_n, é definida por:

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

Onde o somatório (Σ) corresponde à soma de todos os valores obtidos. Por exemplo, considerando o caso da taxa de efetivação (%) da cobrança de um determinado tributo que está atrasado em uma prefeitura (ver Tabela 12), se você somar todos os valores do número das taxas e dividi-los pelo total de dias avaliados, você terá, então, a **média aritmética** ($\bar{x}$), a taxa de efetivações de cobrança por dia. Logo, o valor obtido será: $\bar{x}$ = 56,67%.

Como podemos, então, fazer a interpretação da média?

Podemos interpretar o resultado da média como sendo o número de efetivações diárias que é de 56,67%, podendo ocorrer taxas maiores, menores ou até iguais ao valor médio encontrado.

Portanto, de uma forma mais geral, podemos interpretar a média como sendo um valor típico do conjunto de dados que pode assumir um valor que não pertence ao conjunto de dados, pois como nos dados utilizados para cálculo (exemplo anterior) não existe um taxa de efetivação diária de 56,67%.

Todavia, se os dados estiverem agrupados na forma de uma distribuição de frequência em classes, lança-se mão da **Hipótese Tabular Básica*** para o cálculo da média.

Então, você irá calcular a média por meio da seguinte expressão:

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

***Hipótese Tabular Básica** – todas as observações contidas em uma classe são consideradas iguais ao ponto médio da classe. Fonte: Elaborado pelo autor.

Onde:

x_i é o ponto médio da classe i;
fa_i representa frequência absoluta da classe i; e
fr_i é a frequência relativa da classe i.

Considerando a situação do número de casas na rua (Tabela 11), a média será dada por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i fa_i}{\sum_{i=1}^{n} fa_i} = \frac{(1\times.5)+(3\times7)+...+(18\times5)}{5+7+...+5} = 8{,}73 \ casas$$

O valor de 1, apresentado na expressão, corresponde ao ponto médio da primeira classe, que foi obtido pela soma dos limites superior e inferior (0 + 2) divididos por dois, ou seja, a média aritmética. Os pontos médios das outras classes são obtidos de forma similar.

Antes de darmos continuidade, é muito importante você saber que, em relação à notação matemática, quando calculamos a média a partir dos dados de uma população, devemos utilizar a letra μ para designar a média populacional e para média amostral a notação a ser utilizada é $\bar{x}$. Na grande maioria dos casos, iremos trabalhar com amostras. A forma de cálculo é a mesma nas duas situações, mas as notações são diferentes, ou seja:

Média populacional ➔ μ
Média amostral ➔ $\bar{x}$

As médias são comumente utilizadas e apresentam propriedades específicas. As principais propriedades são:

- A soma dos **desvios*** de um conjunto de dados em relação a sua média é nula, ou seja, igual a zero. Para entender essa propriedade, tomemos como exemplo a quantidade consumida de arroz do tipo A em um refeitório de uma prefeitura: 10, 14, 13, 15, 16, 18,

*Desvios – diferenças entre cada valor e um valor padrão, que pode ser a média. Fonte: Elaborado pelo autor.

12 quilos, no qual o consumo médio diário encontrado foi de 14 quilogramas (Kg).

A soma desvios será:

$$(10 - 14) + (14 - 14) + (13 - 14) + (15 - 14) + (16 - 14) + (18 - 14) + (12 - 14) = 0$$

- A soma ou a subtração de uma constante (c) a todos os valores de uma variável, a média do conjunto fica aumentada ou diminuída dessa constante. Assim, voltando ao caso do consumo de arroz, apresentado no tópico anterior, se somarmos 2 a cada um dos valores, teremos:

$$Y = (12 + 16 + 15 + 17 + 18 + 20 + 14) / 7 = 16 \text{ kg ou}$$
$$Y = 14 + 2 = 16 \text{ kg}$$

- Na multiplicação ou na divisão de todos os valores de uma variável por uma constante (c), a média do conjunto fica multiplicada ou dividida por essa constante. Novamente pensando no caso do consumo de arroz, se multiplicarmos 3 a cada um dos valores, teremos:

$$Y = (30 + 42 + 39 + 45 + 48 + 54 + 36) / 7 = 42 \text{ kg ou}$$
$$Y = 14 \,.\, 3 = 42 \text{ kg}$$

Existem outros tipos de médias que podemos utilizar, por exemplo, média ponderada (utilizada quando existe algum fator de ponderação); media geométrica (quando os dados apresentam uma distribuição que não é simétrica); entre outras.

Às vezes, podemos, ainda, associar às observações X_1, X_2, ..., X_n determinadas ponderações, ou pesos, W_1, W_2, ..., W_n que dependem da importância atribuída a cada uma das observações, nesse caso, a média ponderada será dada por:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i W_i}{\sum_{i=1}^{n} W_i}$$

Para entender melhor, imagine um processo de avaliação de funcionários públicos que foi divido em três etapas. Nessa avaliação, suponha que um dos colaboradores apresentou as seguintes notas durante a avaliação: 1ª etapa = 90; 2ª etapa = 70; 3ª etapa = 85; e os pesos de cada etapa são: 1, 1 e 3, respectivamente. Qual o escore médio final do funcionário público?

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i\, w_i}{\sum_{i=1}^{n} w_i} = \frac{(1\times 70)+(1\times 90)+(3\times 85)}{1+1+3} = \frac{415}{5} = 83$$

Outro tipo de média corresponde à geométrica (Mg), calculada pela raiz n-ésima do produto de um conjunto de n observações, X_1, X_2, ... , X_n, associadas às frequências absolutas $f_1, f_2, \ldots, f_n$ (número de vezes que aquele valor acontece) e respectivamente dada por:

Este tipo de média você irá utilizar na disciplina *Matemática Financeira* que trabalharemos no próximo módulo.

$$Mg = \sqrt[n]{x_1^{f_1} \times x_2^{f_2} \times \ldots \times x_n^{f_n}}$$

Sendo assim, considerando o caso da taxa de efetivação para pagamento do tributo atrasado (exemplo apresentado anteriormente), teremos:

$$Mg = \sqrt[18]{44^1 \times 46^1 \times 51^1 \times 54^2 \times \ldots\ldots \times 61^3 \times 62^1 \times 63^2} = 56{,}40\%$$

Moda

Em algumas situações, você verá que é necessária a informação do número de observações que mais ocorre em um conjunto de dados. No caso da taxa de efetivação da cobrança, verificamos que a taxa que mais ocorre é 56 e 61. Assim, podemos definir a **moda (Mo) como sendo o valor em um conjunto**

de dados que ocorre com maior frequência. Um conjunto de dados pode ser em relação à moda:

- unimodal ➔ possui apenas uma moda;
- amodal ➔ não possuir moda, pois não existe nenhum valor que ocorre com maior frequência; e
- multimodal ➔ possui mais de uma moda.

Na situação comentada anteriormente, a distribuição é multimodal ou bimodal, pois apresenta duas modas, ou seja, dois valores com maior frequência, 56 e 61.

Quando os dados não estão em intervalos de classes, basta olhar o valor que ocorre com maior frequência.

Para dados agrupados em intervalos de classes, você pode calcular a moda por meio do método de Czuber, que se baseia na influência das classes adjacente na moda deslocando-se no sentido da classe de maior frequência. A expressão que você utilizará é:

$$Mo = L_i + \frac{d_1}{d_1 + d_2} \times c$$

Onde:

L_i : limite inferior da classe modal;

d_1 : diferença entre a frequência da classe modal e a imediatamente anterior;

d_2 : diferença entre a frequência da classe modal e a imediatamente posterior; e

c : amplitude da classe modal.

No caso em que, para facilitar um projeto de aplicação da rede de esgoto de certa região de uma cidade, os engenheiros da Prefeitura Municipal tomaram uma amostra de 52 ruas, contando o número de casas (Tabela 11), teremos que a classe modal é a **quarta**, pois apresenta maior frequência (valor igual a 16). Utilizando a expressão mostrada anteriormente, teremos:

$$Mo = L_i + \frac{d_1}{d_1 + d_2} \times c = 8 + \frac{5}{5+8} \times 4 = 9,54 \ casas$$

Uma característica importante da moda é que ela não é afetada pelos valores extremos da distribuição, desde que esses valores não constituam a classe modal.

Dessa forma, a moda deve ser utilizada quando desejamos obter uma medida rápida e aproximada de posição ou quando a medida deva ser o valor mais frequente da distribuição.

Mediana

Outra medida de posição que você pode utilizar é a **mediana (Md)**, que consiste em um conjunto de valores dispostos segundo uma ordem (crescente ou decrescente). A mediana é o valor situado de tal forma no conjunto ordenado que o separa em dois subconjuntos de mesmo número de elementos, ou seja, 50% dos dados são superiores à mediana e 50% são inferiores.

O símbolo da mediana é dado por Md ou $\tilde{x}$, e a sua posição é dada por meio do da expressão:

E (elemento central) = (n+1) / 2

Considerando um conjunto de dados com número ímpar de elementos (1, 2, 5, 9, 10, 12, 13), a posição da mediana será dada por (7 + 1)/2 = 4ª posição. Portanto, a partir dos dados ordenados, o número que se encontra na 4ª posição é o 9 e, assim, a mediana será igual a 9 (temos três valores abaixo e três valores acima, ou 50% acima da mediana e 50% abaixo).

E, caso o número de elementos do conjunto de dados seja par, por exemplo, (1, 2, 6, 8, 9, 12, 11, 13) a posição da mediana será:

E = (8 + 1)/2 = 4,5ª posição

Como a posição 4,5 está entre a 4ª e a 5ª posição, calculamos a média entre os valores que ocupam essas posições.

O valor encontrado de 8,5, (vem de (8 + 9) / 2), corresponde à mediana.

Quando os dados estão agrupados na mediana, devemos encontrar a classe mediana. Se os dados estão agrupados em intervalos de classe, como no caso do número de casa por rua, utilizaremos a seguinte expressão:

$$Md = li + \left(\frac{(n/2) - f_{antac}}{f_{med}} \right) \times c$$

Onde:

li : limite inferior da classe mediana;

n : número total de elementos;

f_{antac} : frequência acumulada anterior à classe mediana;

f_{med} : frequência absoluta da classe mediana; e

c: amplitude da classe mediana.

Portanto, resolvendo o caso em que, para facilitar um projeto de aplicação da rede de esgoto de certa região de uma cidade, os engenheiros da Prefeitura Municipal tomaram uma amostra de 52 ruas, contando o número de casas por rua; você verá que a posição da mediana será dada por:

E = (52+1)/2 = 26,5º elemento, o qual está na quarta classe (8 ⊢ 12), que corresponde à classe mediana.

$$Md = li + \left(\frac{(n/2) - f_{antac}}{f_{med}} \right) \times c = 8 + \left(\frac{(52/2) - 23}{16} \right) \times 4 = 8{,}75 \ casas$$

Em um conjunto de dados, a mediana, a moda e a média não necessariamente devem apresentar o mesmo valor. Uma informação importante é que a mediana não é influenciada pelos valores extremos. Comparando os resultados encontrados para uma amostra em relação às medidas de posição estudadas e verificando a inter-relação entre elas, você pode concluir que seus valores podem

nos dar um indicativo da natureza da distribuição dos dados, em função das regras definidas pela Figura 12:

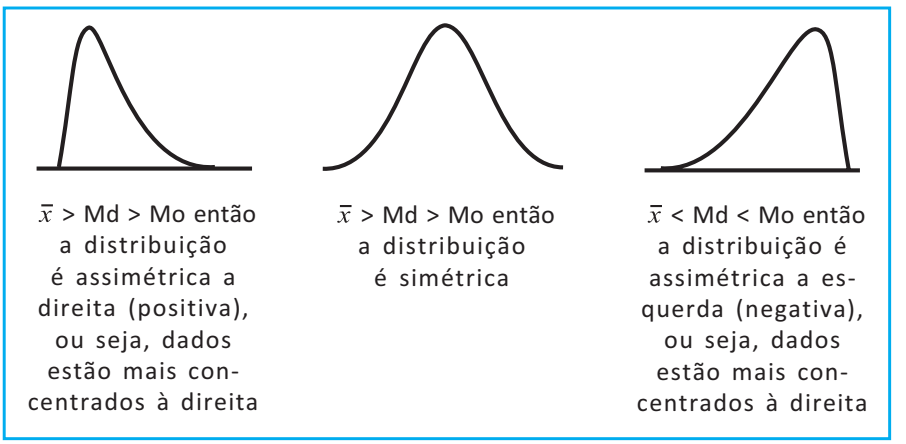

Figura 12: Natureza de distribuição de dados
Fonte: Elaborada pelo autor

Separatrizes

A principal característica das medidas separatrizes consiste na separação da série de dados ordenados em partes iguais que apresentam o mesmo número de valores. As principais são os quartis, os decis e os percentis.

Os **quartis** são valores que dividem um conjunto de dados ordenados em quatro partes iguais. São necessários, portanto, três quartis (Q_1, Q_2 e Q_3) para dividir um conjunto de dados ordenados em quatro partes iguais.

Q_1 : deixa 25% dos elementos abaixo dele.

Q_2 : deixa 50% dos elementos abaixo dele e coincide com a mediana.

Q_3 : deixa 75% dos elementos abaixo dele.

A Figura 13 mostra bem a divisão dos quartis, observe.

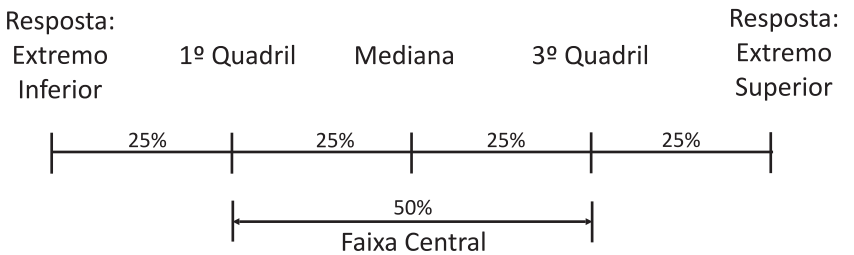

Figura 13: Representação dos quartis
Fonte: Elaborada pelo autor

Se considerarmos a situação da taxa de efetivação da cobrança de um determinado tributo, que estava atrasado em uma prefeitura, após uma campanha realizada para que ele fosse saldado, teremos, de forma semelhante à Figura 13, a Figura 14:

$$\underset{44}{Mínimo} \quad \underset{54}{Q_1} \quad \underset{57}{Q_2} \quad \underset{61}{Q_3} \quad \underset{63}{Máximo}$$

Figura 14: Quartis da taxa de efetivação da cobrança de um determinado tributo
Fonte: Elaborada pelo autor

Sendo assim, temos o cálculo da posição do elemento quartil dado por:

$$EQi = in/4 \quad (i = 1, 2, 3)$$

A regra para obtenção dos valores dos quartis, a partir da posição encontrada, será dada por:

- quando n é impar, o arredondamento deve ser para cima da posição encontrada; e
- quando n é par, devemos fazer a média do valor encontrado e do subsequente.

Para melhor entendimento, elaboramos um exemplo para realizarmos juntos. Para tanto, considere a seguinte sequência de números para cálculo dos quartis: (5, 2, 6, 9, 10, 13, 15).

Agora, precisamos ordenar o conjunto de dados e, então, temos: (2, 5, 6, 9, 10, 13, 15).

Sendo assim, obtemos a posição e, olhando no conjunto ordenado de dados, encontramos os valores dos quartis, conforme você pode observar a seguir.

$$EQ1 = 1.7/4 = 1{,}75 \cong 2^a \text{ posição} \Rightarrow Q1 = 5$$
$$EQ2 = 2.7/4 = 3{,}5 \cong 4^a \text{ posição} \Rightarrow Q2 = 9$$
$$EQ3 = 3.7/4 = 5{,}25 \cong 6^a \text{ posição} \Rightarrow Q3 = 13$$

Agora, vamos a outro exemplo, para tanto, considere um conjunto de dados com uma quantidade par de observações, a saber: (1, 1, 2, 3, 5, 5, 6, 7, 9, 9, 10, 13) $\Rightarrow$ já ordenados. Então, temos:

$$EQ1 = 1.12/4 = 3^a \text{ posição} \Rightarrow Q1 = (2 + 3) / 2 = 2{,}5$$
$$EQ2 = 2.12/4 = 6^a \text{ posição} \Rightarrow Q2 = (5 + 6) / 2 = 5{,}5$$
$$EQ3 = 3.12/4 = 9^a \text{ posição} \Rightarrow Q3 = (9 + 9) / 2 = 9$$

Os **decis** são valores que dividem um conjunto de dados ordenados em dez partes iguais.

O cálculo de cada decil será obtido de forma semelhante aos quartis, sendo diferente apenas a expressão de sua obtenção, que será dada por:

Posição do elemento decil ➔ $EDi = in/10 \quad (i = 1, 2, \ldots, 9)$

Os **percentis** são valores que dividem um conjunto de dados ordenados em 100 partes iguais.

A posição de cada percentil será dada pela expressão a seguir que é semelhante aos quartis e aos decis:

Posição do elemento percentil ➔ $EPi = in/100 \quad (i = 1, 2, \ldots, 99)$

Essas medidas separatrizes são importantes quando queremos dividir um conjunto de dados em parte iguais; por exemplo, em quatro partes; e, assim, você terá os quartis. Essa separação permite uma formação de grupos que podem apresentar um mesmo padrão, quando, então, poderemos identificar perfis importantes para serem utilizados em diversas áreas da Administração.

Medidas de Dispersão

Como vimos anteriormente, é possível sintetizar um conjunto de observações em alguns valores representativos, como média, mediana, moda e separatrizes. Em várias situações, é necessário visualizar como os dados estão dispersos.

Tomando como exemplo algumas funções da área de Administração Pública que apresentem salários médios iguais, podemos concluir que sua contribuição social (% do salário) será a mesma?

A resposta é sim somente com base no salário médio; mas estaríamos chegando a uma conclusão errada, pois a variação em termos de faixas salariais pode ser diferente, apesar de apresentarem a mesma média.

Suponhamos três cidades: A, B e C, que foram avaliadas durante cinco anos quanto ao número de declarantes na distribuição de patrimônio na faixa de renda mensal de 8 a 10 mil reais. Esses valores estão em milhares de pessoas.

$$A = \{120, 122, 118, 124, 121\}$$
$$B = \{121, 121, 121, 121, 121\}$$
$$C = \{116, 125, 124, 120, 120\}$$

Se nós calcularmos a média de cada cidade, teremos:

A ➔ $\bar{x}$ = 121 mil pessoas
B ➔ $\bar{x}$ = 121 mil pessoas
C ➔ $\bar{x}$ = 121mil pessoas

Note que as três cidades (A, B, C) apresentam médias iguais, apesar de elas serem bem diferentes entre si, pois enquanto na cidade B os dados são todos iguais, os das demais cidades apresentam certa variação, que é maior no conjunto C. Portanto, devemos associar medidas de posição e de dispersão para obtermos informações mais precisas de um conjunto de dados, ou seja, observar como esses dados se comportam em torno da medida de posição em questão.

Amplitude total

A amplitude total é a diferença entre o maior e o menor valor observado, como vimos na Unidade 2.

Sendo assim, retomando nossos exemplos das cidades A, B e C, temos:

$A_A = 124 - 118 = 6$ mil pessoas
$A_B = 121 - 121 = 0$ mil pessoas
$A_C = 125 - 116 = 9$ mil pessoas

Desse modo, podemos identificar que a amplitude do conjunto C é bem maior do que nos demais e o conjunto B apresenta amplitude igual a zero.

Essa medida apresenta a vantagem de ser facilmente calculada. Entretanto, o seu inconveniente é que ela é muito afetada pelos valores extremos, pois no seu cálculo não são consideradas todas as observações.

Variância

Uma boa medida de dispersão deve ter as seguintes características:

- estar baseada em todos os dados;
- ser facilmente calculada;
- ser compreensível; e
- servir bem ao tratamento algébrico.

Portanto, podemos afirmar que uma medida de dispersão deve utilizar todas as observações considerando os desvios de cada observação em relação à média (chamados erros):

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

Para obter um único número que represente a dispersão dos dados, pensamos, inicialmente, em obter a média desses desvios, mas devemos lembrar de que a soma dos desvios de um conjunto de dados em relação a sua média é nula.

Para resolver esse problema, utilizamos a soma dos quadrados dos desvios, pois, ao elevarmos cada desvio ao quadrado, eliminamos o sinal negativo que estava trazendo complicações.

Posteriormente, dividimos a soma dos quadrados dos desvios pelo número de observações para obtermos a variância populacional, chamada de σ^2, que é uma medida quantitativa da dispersão de um conjunto de dados entorno da sua média, além do fato de essa soma de quadrados de desvios ser mínima.

Sendo assim, temos a expressão para cálculo da variância populacional, conforme mostrada a seguir:

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

E não para por aí! Na maioria das vezes, trabalhamos com amostras e, nesse caso, a variância amostral (s^2) será obtida pela expressão:

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

Veja que nesse caso a soma do quadrado dos desvios é dividida por n – 1, onde n corresponde ao tamanho da amostra. Esse valor n – 1 (número de observações menos um) é denominado de **grau de liberdade***.

*Grau de liberdade – é o número de determinações independentes (dimensão da amostra) menos o número de parâmetros estatísticos a serem avaliados na população. Fonte: Elaborado pelo autor.

Então, o grau de liberdade é um estimador do número de categorias independentes em um teste particular ou experiência estatística. Assim, no caso das cidades teremos:

$$s_A^2 = \frac{(120-121)^2 + (122-121)^2 + \ldots\ldots + (121-121)^2}{4} = 5\ mil\ pessoas^2$$

$$s_B^2 = \frac{(121-121)^2 + (121-121)^2 + \ldots\ldots + (121-121)^2}{4} = 0\ mil\ pessoas^2$$

$$s_C^2 = \frac{(116-121)^2 + (125-121)^2 + \ldots\ldots + (120-121)^2}{4} = 13\ mil\ pessoas^2$$

Para que você entenda melhor, veja a seguir algumas das principais propriedades da variância:

- A variância de uma constante k é nula.

  $V(k) = 0$, k = constante.

- Ao somar ou ao subtrair uma constante k a todos os dados, a variância não se altera.

  $x' = x \pm k$

  $V(x') = V(x)$

- Multiplicando todos os dados por uma constante k, a variância é multiplicada por k^2.

  $x' = x.\ k$

  $V(x') = k^2.V(x)$

Desvio Padrão

Um inconveniente da variância é que ela é expressa em unidades ao quadrado, ou seja, caso esteja trabalhando com milhares de reais, o resultado será expresso em milhares de reais^2, o que causa algumas dificuldades de interpretação.

Para resolver esse problema, podemos nos utilizar do desvio padrão que é definido como a raiz quadrada positiva da variância, sendo expresso na mesma unidade em que os dados foram coletados.

$$\sigma = \sqrt{\sigma^2} \qquad \text{(desvio padrão populacional)}$$

$$s = \sqrt{s^2} \qquad \text{(desvio padrão amostral)}$$

Para o exemplo em questão, temos:

$$s_A^2 = \sqrt{\frac{(120-121)^2 + (122-121)^2 + \ldots\ldots + (121-121)^2}{4}} = 2{,}24 \ \textit{mil pessoas}$$

$$s_B^2 = \sqrt{\frac{(121-121)^2 + (121+21)^2 + \ldots\ldots + (121-121)^2}{4}} = 0 \ \textit{mil pessoas}$$

$$s_C^2 = \sqrt{\frac{(116-121)^2 + (125-121)^2 + \ldots\ldots + (120-121)^2}{4}} = 3{,}60 \ \textit{mil pessoas}$$

Interpretando, temos que: o desvio padrão de 3,60 mil pessoas nos indica a variação dos dados em torno da média, que é de 121 mil pessoas. Quanto menor for o desvio padrão, menor será a variabilidade, ou a variação.

No caso de dados agrupados em classes, a expressão utilizada para cálculo do desvio padrão será:

$$s^2 = \sqrt{\frac{SQD}{n-1}} = \sqrt{\frac{1}{n-1}\sum_{i-1}^{n}(x_i - \overline{x})^2 . f_{ai}}$$

Para entender melhor, vamos imaginar uma situação em que, para facilitar um projeto de aplicação da rede de esgoto de certa região de uma cidade, os engenheiros da Prefeitura Municipal

tomaram uma amostra de 52 ruas (Tabela 11), contando o número de casas por rua, na qual os dados estão agrupados em classes, iremos calcular o desvio padrão da seguinte maneira:

$$s^2 = \sqrt{\frac{1}{52-1}\left((1-8{,}73)^2.5 + (3-8{,}73)^2.7 + ... + (18-8{,}73)^2.5\right)} = 5{,}075\ casas$$

Com base nessa resolução, os números 1, 3 e 18 correspondem aos pontos médios das classes primeira, segunda e última, respectivamente. Já os valores de 5 e 7 correspondem às frequências absolutas das classes. E o número 52 corresponde ao tamanho da amostra.

Existem algumas propriedades que precisamos saber sobre desvio padrão. São elas:

- Ao somar ou ao subtrair uma constante k a todos os dados, o desvio padrão não se altera.

  $x' = x \pm k$

  $\sigma(x') = \sigma(x)$

- Multiplicando todos os dados por uma constante k, o desvio padrão fica multiplicado por k

  $x' = x.k$

  $\sigma(x') = k.\ \sigma(x)$

Coeficiente de Variação

A variância e o desvio padrão são medidas de dispersão absolutas, desse modo, apenas podem ser utilizados para comparar a variabilidade de dois ou mais conjuntos de dados quando estes apresentarem:

- mesma média;
- mesmo número de observações; e
- estiverem expressos nas mesmas unidades.

Então, para você comparar qualquer conjunto de dados quanto à sua variabilidade quando, pelo menos, uma dessas condições não é satisfeita, é necessário lançar mão de uma medida de dispersão relativa como o **coeficiente de variação** (CV), que expressa a variabilidade dos dados em relação a sua média de forma percentual. Sua expressão será dada por:

$$CV = \frac{S}{\overline{x}} \cdot 100$$

Para melhor entendimento, vamos elaborar um exemplo para você.

Exemplo

Imagine uma situação referente ao número de documentos falsificados que aparecem em um determinado setor da prefeitura e o valor arrecadado por hora de um tipo de multa em reais. Em qual das duas variáveis ocorre maior variabilidade, ou variação?

	Documentos falsificados (nº)	Multa (reais)
Média	22	800
Desvio padrão	5	100

Utilizando o desvio padrão para comparar a variabilidade, você pode, a princípio, considerar que a multa apresenta maior variabilidade, já que tem maior desvio padrão. Entretanto, se verificarmos as condições de se utilizar o desvio padrão para comparar a variabilidade entre amostras, você vai perceber que as médias são diferentes e as unidades também são diferentes.

Calculando, então, o coeficiente de variação, teremos os valores apresentados, a seguir:

$$CV_{DOC} = \frac{S}{\bar{x}} \cdot 100 = \frac{5}{22} \cdot 100 = 22{,}7\%$$

$$CV_{MULTA} = \frac{S}{\bar{x}} \cdot 100 = \frac{100}{800} \cdot 100 = 12{,}5\%$$

Perceba, então, que estávamos concluindo erroneamente que a multa é mais variável do que o número de documentos falsificados, além de termos cometido o disparate de comparar numericamente duas variáveis expressas em unidades diferentes.

Portanto, o número de documentos falsificados apresentou maior dispersão do que a multa, já que seu coeficiente de variação foi maior, mudando assim a conclusão anterior.

Vamos ver agora outros exemplos de situações com a resolução comentada para você fixar melhor os conceitos desta Unidade.

Exemplo 1

Considere as idades dos funcionários do programa *Jovens que aprendem uma profissão* de duas prefeituras, apresentadas a seguir.

Prefeitura A: {16; 15; 18; 15; 16; 16; 17; 18; 19; 17; 16}
Prefeitura B: {15; 17; 19; 19; 17; 18; 19; 18; 18; 17; 16}

Encontre a média, moda e mediana de cada prefeitura e identifique qual das prefeituras apresenta maior variabilidade na idade de seus jovens aprendizes.

Prefeitura A

- Média: $\bar{x} = \frac{\sum x_i}{n} = \frac{16 + 15 + ... + 16}{11} = 16{,}64$
- Mediana: Md = 16, lembrando que, para encontrar a mediana, necessariamente os dados devem estar ordenados.

- Moda: Mo = 16, valor que aparece com maior frequência.

Prefeitura B

- Média: $\bar{x} = \frac{\sum x_i}{n} = \frac{15+17+...+16}{11} = 17{,}54$
- Mediana: Md = 18, lembrando que, para encontrar a mediana, necessariamente os dados devem estar ordenados.
- Moda: Mo = 17, 18 e 19 (distribuição multimodal, pois apresenta mais de duas modas).

Para sabermos quem tem maior variabilidade, temos de calcular o coeficiente de variação, pois, como o valor das médias são diferentes, não podemos usar o desvio padrão para comparar a variabilidade. Para encontrarmos o desvio padrão, precisamos primeiramente encontrar a variância usando a seguinte formula:

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

Prefeitura A

- Variância:

$$s^2 = \frac{(16-16{,}64)^2 + (15-16{,}64)^2 + ... + (16-16{,}64)^2}{11-1} = 1{,}654$$

- Desvio padrão: $s = \sqrt{1{,}654} = 1{,}2862$
- Coeficiente de variação: $CV = \frac{s}{\bar{x}}.100 = \frac{1{,}2862}{16{,}64}.100 = 7{,}7\%$

Prefeitura B

- Variância:

$$s^2 = \frac{(15-17{,}54)^2 + (17-17{,}54)^2 + ... + (16-17{,}54)^2}{11-1} = 1{,}6726$$

- Desvio padrão: $s = \sqrt{1{,}6726} = 1{,}2933$
- Coeficiente de variação: $CV = \frac{s}{\bar{x}}.100 = \frac{1{,}2933}{17{,}54}.100 = 7{,}3\%$

Sendo assim, como os coeficientes apresentam valores muito próximos, podemos concluir que a variabilidade na idade das duas prefeituras é praticamente a mesma.

Exemplo 2

Considerando os dados apresentados a seguir, que são referentes ao percentual de gastos com planejamento e com administração em cidades de diferentes portes, identifique as medidas de posição e de dispersão dos dados.

Gasto	Frequência (F_i)
5 ⊢ 15	2
15 ⊢ 25	7
25 ⊢ 35	20
35 ⊢ 45	5
45 ⊢ 55	4
55 ⊢ 65	2
Soma	40

Primeiramente, temos de encontrar os valores de x_i (ponto médio), pois ele é indispensável no cálculo da média, variância etc. Logo, temos:

X_i = 10; 20; 30; 40; 50; 60 (soma: limite inferior + limite superior dividido por 2).

Feita essa conta, vamos calcular a frequência acumulada, como você pode acompanhar a seguir:

F_{ac} = 2; 9; 29; 34; 38; 40

E, na sequência, com os valores do ponto médio, podemos calcular a média:

$$\bar{x} = \frac{\sum x_i \times f_i}{\sum fa} = \frac{10.2 + 20.7 + 30.20 + ... + 60.2}{40} = 32$$

Para encontrar a mediana, primeiramente temos de encontrar a classe mediana. Como n é par: $x_{n/2} = x_{40/2} = x_{20}$, a qual classe pertence o elemento de posição 20º (3ª classe)?

$$Md = Li + \left(\frac{\frac{n}{2} - f_{antac}}{f_{md}} \right) . c = 25 + \left(\frac{\frac{40}{2} - 9}{20} \right) . 10 = 30,5$$

Vamos, agora, calcular a moda e, para tanto, precisamos encontrar a classe modal, aquela com maior frequência absoluta (3ª classe).

$$Mo = LI_{mo} + \left(\frac{d_1}{d_1 + d_2} \right) . c = 25 + \left(\frac{13}{13+15} \right) . 10 = 29,6$$

Fique atento, pois as classes mediana e modal não necessariamente vão pertencer a mesma classe.

E, por fim, devemos fazer o cálculo das medidas de dispersão, como você pode acompanhar a seguir:

$$S^2 = \frac{\sum (x_i - \bar{x})^2 \times f_i}{n-1} = \frac{(10-32)^2 \times 2 + ... + (60-32)^2 \times 2}{40-1} = \frac{5240}{39} = 134,3590$$

$$S = \sqrt{S^2} = \sqrt{134,3590} = 11,5913$$

$$CV = \frac{S}{\bar{x}} \times 100 = 36,22\%$$

Observe que, com as medidas de dispersão calculadas, podemos verificar que a dispersão obtida foi média (36,22% em torno da média), ou seja, tanto para cima quanto para baixo. Se esse valor fosse bem menor, poderíamos considerar que os gastos com planejamento e com transportes seriam mais uniformes.

Exemplo 3

Considerando as séries de dados apresentadas pelos gastos com transportes em relação ao total gasto em várias prefeituras, conforme descrição a seguir, faça o seguinte: imagine que você precise efetuar uma estimativa com base nesses dados. Sobre qual série é mais fácil fazer estimativas precisas? Por quê?

Série A: {3,96; 3,17; 3,55; 3,61; 4,11; 4,57; 4,97; 5,91; 5,99; 5,74}

Série B: {1,46; 2,09; 3,04; 5,12; 7,80; 8,25; 9,95; 15,24; 17,40; 21,74}

Série A

- Média: $\bar{x} = \frac{\sum x_i}{n} = \frac{3,96 + 3,17 + ... + 5,74}{10} = 4,558$
- Variância:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(3,96 - 4,558)^2 + ... + (5,74 - 4,558)^2}{10-1} = 1,0939$$

- Desvio padrão: $S = \sqrt{S^2} = \sqrt{1,0939} = 1,0459$
- Coeficiente de variabilidade: $CV = \frac{S}{\bar{x}} \times 100 = 22,9\%$

Série B

- Média: $\bar{x} = \frac{\sum x_i}{n} = \frac{1,46 + 2,09 + ... + 21,74}{10} = 9,206$
- Variância:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(1,46 - 9,206)^2 + ... + (21,74 - 9,206)^2}{10-1} = 47,748$$

- Desvio padrão: $S = \sqrt{S^2} = \sqrt{47,748} = 6,91$
- Coeficiente de variabilidade: $CV = \frac{S}{\bar{x}} \times 100 = 75\%$

Observe que na série A é mais fácil fazermos estimativas precisas, pois ela apresenta menor dispersão.

Resumindo

Nesta Unidade, você aprendeu conceitos básicos sobre as medidas de posição e de dispersão e, desse modo, sabe, agora, caracterizar um conjunto de observações. Esses conceitos são de extrema importância para as inferências estatísticas, para os testes de hipóteses e para as informações contidas nas Unidades posteriores dessa disciplina.

Atividades de aprendizagem

Agora que você já sabe como calcular e como utilizar as principais medidas de posição e de dispersão, exercite-as fazendo as atividades, a seguir, que serão importantes na consolidação dos conhecimentos adquiridos. Em caso de dúvida, lembre-se de consultar seu tutor por meio do AVEA.

1. Considere a sequência numérica apresentada, a seguir, que mostra as idades de motociclistas e de seus caronas na época em que morreram em acidentes fatais de trânsito.

7	38	27	14	18	34	16
42	28	24	40	20	23	31
37	21	30	25	17	28	33
25	23	19	51	18	29	

   Calcule a média moda, a mediana, a variância, o desvio padrão e o coeficiente de variabilidade para os dados não agrupados.

2. Image um determinado setor de uma prefeitura que vem apresentando problemas com o afastamento de funcionários por motivos de saúde, por período muito longo. Uma amostra de dez apresentou os seguintes números de dias afastados em um semestre:

   23, 21, 10, 14, 16, 12, 39, 45, 10 e 20

   Calcule as medidas de posição e de dispersão em relação ao número de dias em que eles ficaram afastados.