Enap Escola Nacional de Administração Pública

Programa Avaliação Socioeconômica de Projetos

Unidade 1
Estatística Descritiva

Brasília - 2015

Conteúdo para impressão

Fundação Escola Nacional de Administração Pública

Presidente
Gleisson Rubin

Diretor de Desenvolvimento Gerencial
Paulo Marques

Coordenadora-Geral de Educação a Distância
Natália Teles da Mota Teixeira

Conteudista

Alexandre da Costa Pereira (2013)

Diagramação realizada no âmbito do acordo de Cooperação Técnica FUB/CDT/Laboratório Latitude e Enap.

SUMÁRIO

Unidade 1
Estatística Descritiva

Unidade 1 - ESTATÍSTICA DESCRITIVA

Objetivos da Unidade

Esta unidade tem por objetivo fazer com que você tenha condições de descrever e apresentar os resultados de um conjunto de observações de forma clara, objetiva e passando o máximo de informações possíveis.

Para tal objetivo, serão abordadas questões relacionadas a dados estatísticos, distribuições de frequências, representações gráficas, medidas de posição e dispersão.

Após concluir esta unidade, espera-se que você seja capaz de:

- Reconhecer a importância dos métodos estatísticos para o estudo de variáveis.
- Compreender os conceitos fundamentais da estatística descritiva.
- Analisar distribuições de freqüências para dados estatísticos e suas formas de representação.
- Distinguir e saber aplicar as diversas medidas de posição. Calcular a média, mediana e moda para uma amostra.
- Distinguir e saber aplicar as diversas medidas de dispersão.
- Calcular a variância, o desvio-padrão e coeficiente de variação para uma amostra.

1.1. DADOS ESTATÍSTICOS

A Estatística pode ser definida como sendo "parte da matemática aplicada que fornece métodos para a coleta, organização, descrição, análise e interpretação de dados e para a utilização dos mesmos na tomada de decisões".

A estatística, portanto, é uma ciência que se dedica à coleta, análise e interpretação de dados, preocupando-se com os métodos de coleta, organização, síntese, apresentação e interpretação dos dados, assim como em tirar conclusões sobre as características das fontes donde estes foram retirados para melhor compreender as situações analisadas.

Como podemos apreciar no portal do Instituto Brasileiro de Geografia e Estatística – IBGE (<http://www.ibge.gov.br/brasil_em_sintese/>), são diversos os dados considerados como de interesse para utilização nos processos de planejamento e gestão de projetos e políticas públicas, apresentando-se de forma destacada no site os dados referentes às temáticas relacionadas à população, educação, trabalho, habitação, agropecuária, indústria, comércio, serviços e contas nacionais, bem como as séries estatísticas à disposição no sítio <http://seriesestatisticas.ibge.gov.br/apresentacao.aspx>.

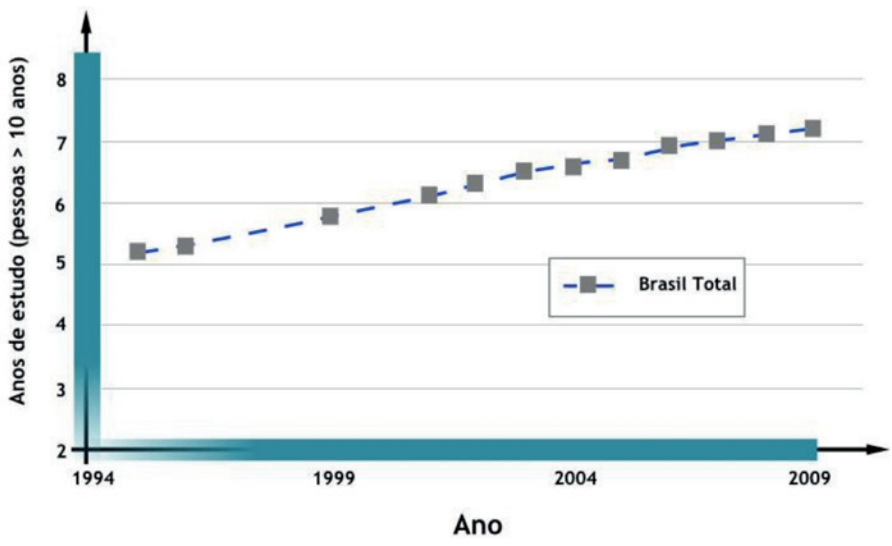
O banco de dados “Séries Estatísticas & Séries Históricas”, à disposição no sítio supracitado, tem por objetivo disseminar, para um público diversificado (instituições governamentais, setor privado, área acadêmica, estudantes, ONGs), informações provenientes de dados oficiais oriundos de pesquisas do IBGE, em sua maior parte, e de outras fontes governamentais.

Ordenadas segundo um intervalo de tempo, essas informações constituem Séries Estatísticas Históricas, sendo, em muitos casos, séries longas (períodos superiores a 20 anos).

Os temas contemplados pelas Séries Estatísticas & Séries Históricas reúnem um acervo de informações sobre a realidade brasileira em suas dimensões social (educação, habitação, trabalho, saúde, organização familiar), demográfica (características da população, dinâmica demográfica e indicadores demográficos) e econômica (sistema de contas nacionais; índices de preços, de produção, de comércio e agropecuária) que, sem excluir os dados de anos recentes, cobrem períodos tão longos quanto possível, importantes enquanto dados estatísticos de interesse em atividades de avaliação socioeconômica de projetos.

Na Figura 01 apresentada abaixo, podemos ver um dos resultados dos levantamentos de séries de dados estatísticos relacionados com a demografia brasileira, referente à série para a evolução dos anos de escolaridade para a parcela da população com idade superior a 10 anos.

Figura 01- Evolução da escolaridade da parcela da população brasileira com idade > 10 anos.

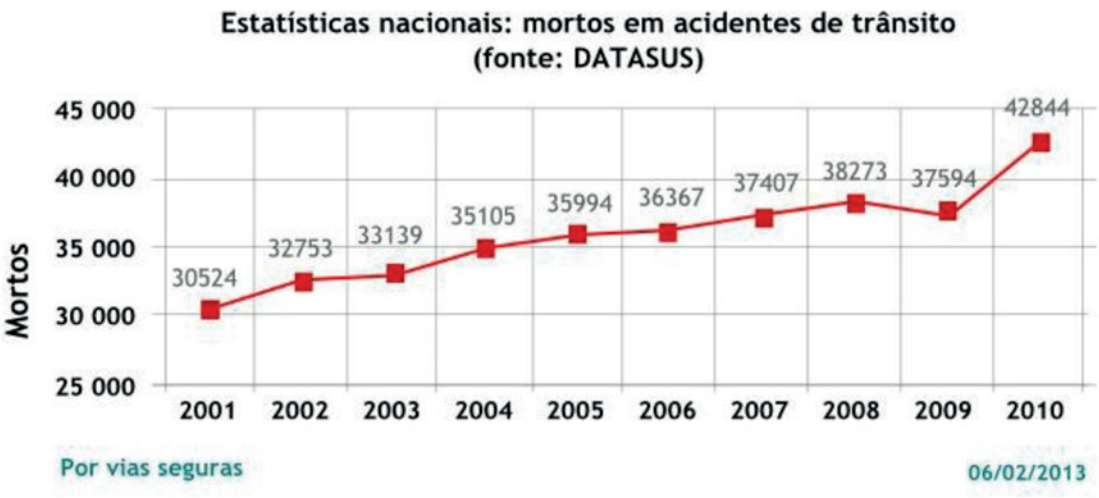


Fonte: <http://seriesestatisticas.ibge.gov.br/series.aspx?no=4&op=0&vcodigo=ECE370&t=media-anos- estudo-pessoas-10-anos>.

Outro exemplo interessante de fonte de dados para séries históricas consiste na compilação de dados realizada por órgãos da Administração Direta, como é o caso do registro feito pelo Ministério da Saúde para o número de óbitos e de internações de vítimas de acidentes do trânsito, cujos dados mais recentes indicam 42.800 óbitos em 2010, 174.000 feridos internados em 2011.

A Figura 02 abaixo mostra a evolução do número de óbitos de 2001 a 2010, com crescimento de 40% no período.

Figura 02- Série de dados de óbitos em acidentes de trânsito no período 2001-2010.



Fonte: <http://www.vias-seguras.com/layout/set/print/os_acidentes/estatisticas/estatisticas_nacionais>.

Para que gráficos como os apresentados nas Figuras 03 e 04 possam ser produzidos, faz-se necessária a mobilização de um esforço importante, com mobilização de recursos humanos, materiais e financeiros de monta para a coleta, tratamento e análise de dados, conforme metodologia adequada com o estudo das distribuições de frequências e representações gráficas, medidas de posição e dispersão para que possam ser adequadamente utilizados para atender, por exemplo, às necessidades do Estado na formulação de políticas públicas, fornecendo dados estatísticos demográficos e econômicos.

Passaremos, então, à apresentação dos fundamentos da estatística descritiva, essenciais para a compreensão do processo estatístico utilizado para a análise de dados e o adequado tratamento dos mesmos.

Como primeiro tópico a ser trabalhado, teremos o estudo das distribuições de frequências para os dados analisados para a adequada compreensão de determinada variável à luz da Estatística.

1.2. DISTRIBUIÇÕES DE FREQUÊNCIAS

Para o estudo do tema relacionado à distribuição de frequências para uma determinada variável, serão apresentados os conceitos referentes à forma de organização dos dados com a adequada determinação do número de classes, sua amplitude, obtenção de suas frequências absoluta e relativa, bem como as formas de representação gráfica e de análise para as variáveis de estudo.

Tipos de Variáveis e a Determinação do Número de Classes (K)

É importante que a distribuição conte com um número adequado de classes. Se o número de classes for excessivamente pequeno, acarretará perda de detalhe e pouca informação se poderá extrair da tabela. Por outro lado, se for utilizado um número excessivo de classes, haverá alguma classe com frequência nula ou muito pequena, não atingindo o objetivo de classificação que é tornar o conjunto de dados supervisionáveis.

Qualquer conjunto de dados, tais como a precipitação pluviométrica em um determinado mês, o PIB de municípios em um Estado, a ocorrência de acidentes em uma rodovia etc., contém informação sobre algum grupo de indivíduos. As possíveis diferenças entre indivíduos determinam a variação que está sempre presente na análise de dados.

Uma característica que pode assumir diferentes valores de indivíduo para indivíduo é denominada variável, pois de outra forma, seria denominada constante. As variáveis podem ser classificadas em qualitativas e quantitativas. Os dados qualitativos são divididos em nominais e ordinais, enquanto os dados quantitativos são divididos em discretas e contínuas.

Portanto, ao conjunto de distintos valores numéricos que adota um caráter quantitativo, denomina-se "variável estatística", que pode apresentar-se conforme dois tipos:

Variáveis qualitativas ou categóricas: que não se podem medir numericamente (ex: nacionalidade, sexo, escolaridade etc).
Variáveis quantitativas: Possuem valor numérico (ex: idade, preço de um produto, renda etc).

As variáveis também podem se classificar em:

- Variáveis unidimensionais: somente transmitem informação sobre uma característica (ex: idade de alunos de uma turma).
- Variáveis bidimensionais: guardam informação sobre duas características da população (ex: idade e renda de trabalhadores de uma cidade).
- Variáveis pluridimensionais: detém informação sobre três ou mais características (ex: idade, altura e peso de alunos de uma turma).

Por sua parte, as variáveis quantitativas podem classificar-se em discretas e contínuas:

- Discretas: só podem tomar valores inteiros (ex: 1, 2, 8, -4, etc.), como é o caso do número de espécies de orquídeas em um determinado bioma.
- Contínuas: Podem tomar qualquer valor real dentro de um intervalo. Por exemplo, a velocidade de um veículo em uma rodovia, que pode ser 70,4 km/h, 94,57 km/h...

Quando se estuda o comportamento de uma variável, é necessário que se proceda à distinção dos seguintes conceitos:

- Indivíduo: qualquer elemento que porte informação sobre o fenômeno que se estuda. Assim, se estudamos à altura dos alunos em uma turma, cada aluno é um indivíduo; se o objeto de estudo é o preço de uma habitação, cada unidade de habitação é um indivíduo.
- População: conjunto de todos os indivíduos (pessoas, objetos, animais etc) que portem informação sobre o fenômeno que se estuda. Por exemplo, se estudamos o preço de habitações em uma cidade, a população será o total de habitações em dita cidade.
- Amostra: subconjunto que, selecionando uma população, por exemplo no caso do estudo do preço de habitações em uma cidade, o normal seria não obter informações sobre todas as moradias da cidade, pois seria um trabalho muito complexo, mas deveria selecionar um subgrupo (amostra) que se entenda suficientemente representativo.

Convém destacar que quando coletamos dados para uma pesquisa, estas observações são chamadas de dados brutos. Um exemplo de dados brutos corresponde à precipitação pluviométrica mensal, medida em mm, consistindo em dados obtidos conforme metodologia utilizada na hidrologia, na forma que foram coletados em uma determinada estação climatológica, sendo, por este motivo, denominados de dados brutos[1].

Geralmente, este tipo de dado não possui as informações que são necessárias pelo observador para a utilização direta em um projeto, ou conclusões que o auxiliem na tomada de decisão, que somente poderá ser viabilizada após a adequada organização dos dados, visando potencializar a sua capacidade de fornecer informações úteis ao analista e minimizar o erro na tomada de decisão na avaliação de um dado problema a ser enfrentado.

A simples organização dos dados em um rol[2] aumenta muito a capacidade de informação destes. Será possibilitada a verificação da amplitude total[3] de variação para os dados observados e também será possível a organização de rol crescente dos dados, com a identificação dos dados mais frequentes na amostra.

Para a organização de um conjunto de dados, recomenda-se a elaboração de tabela de distribuição de frequências, onde são apresentadas as frequências de cada uma das classes[4], contando o número de observações contidas em cada uma delas, obtendo-se a frequência de classe. Denomina-se distribuição de frequências a disposição tabular dos dados agrupados em classes, associados às suas frequências correspondentes.

Para o caso, por exemplo, da precipitação pluviométrica mensal em uma determinada região, poder-se-ia incluir, em uma única classe, os meses em que a precipitação pluviométrica, para um certo período estivesse compreendida no intervalo de 100 e 200mm.

Considerando os dados de chuva referentes ao registro de observações em estação climatológica, tomado como exemplo, será possível identificar conceitos presentes em uma distribuição de frequências. Para a elaboração da distribuição de frequências, é necessário que, primeiramente, se determine o número de classes (k) em que os dados serão agrupados.

Por questões de ordem prática e estética, sugere-se utilizar de 5 a 20 classes, e uma recomendação de ordem prática que poderá ser seguida é a definição do número de classes (k) em função do número de observações (n), segundo a expressão $k = \sqrt{n}$.

Quadro 01: Precipitação pluviométrica média mensal para um período de 4 anos.

Mês/Ano	P (mm)	Mês/Ano	P (mm)	Mês/Ano	P (mm)	Mês/Ano	P (mm)
jan/08	70,9	jan/09	103,3	jan/10	25,5	jan/11	139,8
fev/08	0,9	fev/09	154,2	fev/10	21,7	fev/11	75,9
mar/08	154,4	mar/09	131,8	mar/10	15,3	mar/11	89,5
abr/08	219,6	abr/09	145,6	abr/10	94,3	abr/11	169,1
mai/08	111,3	mai/09	160,6	mai/10	69,7	mai/11	168,9
jun/08	294,5	jun/09	190,6	jun/10	62,7	jun/11	194,1
jul/08	146,1	jul/09	139,5	jul/10	22,8	jul/11	122,8
ago/08	204,8	ago/09	146,7	ago/10	9,9	ago/11	33,8
set/08	25,0	set/09	26,5	set/10	0,0	set/11	2,3
out/08	23,6	set/09	0,2	set/10	0,3	set/11	5,3

1. Dados na forma em que foram coletados, sem nenhum tratamento
2. É a mais simples organização numérica. É a ordenação dos dados em ordem crescente ou decrescente
3. Corresponde à diferença entre o maior e o menor valor observado em um conjunto de dados. Notaremos por A
4. Intervalos nos quais os valores da variável analisada são agrupados.

nov/08	3,5	nov/09	2,0	nov/10	34,7	nov/11	2,7
dez/08	1,8	dez/09	2,6	dez/10	2,6	dez/11	2,6
Obs. P correspondendo a precipitações totais mensais, em mm.							

Para n = 48 observações, teremos, então, o número de classes definido por $k = \sqrt{48} = 6{,}9$, o que implicaria na definição de 7 classes.

Convém destacar que a definição do número de classes, poderá, inclusive, ser realizada conforme critérios fixados pelo próprio analista, respeitando-se o princípio de evitar classes excessivamente ou fracamente populosas.

Em seguida à determinação do número de classes (k) para o agrupamento dos dados, se procederá à determinação da amplitude[5] do intervalo de classe (c), sendo necessária[6], antes, a determinação da amplitude total dos dados (A), que corresponde à diferença entre o maior valor observado e o menor valor observado. Para o cálculo da amplitude total dos dados (A) para o exemplo considerado, que corresponde à diferença entre o maior valor observado e o menor valor observado, teríamos:

$$\boxed{A = 294{,}5 - 0 = 294{,}5 \text{ mm}}$$

Para a amplitude do intervalo de classe (c), considerando o valor da amplitude total (A) calculado, teremos:

$$\boxed{C = 294{,}5 \,/\, (7 - 1) = 49{,}1 \text{ mm}}$$

O próximo passo, após conhecida a amplitude de classes, será a determinação dos intervalos de classe, mediante a definição dos limites inferior (LI) e superior (LS) das classes, escolhidos de modo que o menor valor observado esteja localizado no ponto médio (PM) da primeira classe, ou seja:

$$\boxed{PM = (LI + LS) \,/\, 2}$$

Para o valor do limite inferior para um intervalo, teríamos:

$$\boxed{\text{Limite Inferior} = \text{Menor Valor} - c \,/\, 2}$$

Como exemplo de cálculo, o limite Inferior da 1ª Classe será dado por:

$$\boxed{LI = 0 - 49{,}1/2 = -\ 24{,}5 \text{ mm}}$$

5. Intervalos abertos – os limites da classe (inferior e superior) não pertencem a ela. Intervalos fechados – os limites de classe (superior e inferior) pertencem à classe em questão. Intervalos mistos – um dos limites pertence à classe, e o outro, não.
6. Existem outros procedimentos para determinação da amplitude do intervalo de classe que podem ser encontrados na literatura.

Convém destacar que o ponto médio do intervalo para a primeira classe corresponderá ao valor 0mm e, tendo em vista que não existe valor negativo para precipitação pluviométrica (desconsiderando-se a evapotranspiração), teremos o Quadro 02 abaixo com as classes da nossa distribuição, basta que somemos a amplitude do intervalo de classe a cada limite inferior.

Quadro 02: Definição dos limites inferior e superior de cada uma das classes determinadas para o universo de dados analisado.

Classe	LI	LS
1a	-24,5	24,6
2a	24,6	73,7
3a	73,7	122,8
4a	122,8	171,9
5a	171,9	221
6a	221	270,1
7a	270,1	319,2

Poderemos, então, elaborar um quadro de frequências[7] absolutas e relativas, conforme indicado no quadro apresentado a seguir:

Quadro 03: Frequências absoluta (f_a) e relativa (f_r) para cada uma das classes.

Classe	LI	LS	f_a	f_r
1a	-24,5	24,6	18	0,375
2a	24,6	73,7	8	0,167
3a	73,7	122,8	6	0,125
4a	122,8	171,9	11	0,229
5a	171,9	221	4	0,083
6a	221	270,1	0	0,000
7a	270,1	319,2	1	0,021
Total			48	1,000

Em relação à interpretação das informações contidas no Quadro 03, pode-se observar que os valores para precipitação ocorrida nos 48 meses avaliados estão concentrados na primeira, segunda e quarta classes, decrescendo em direção às classes do fim da tabela.

A apresentação dos dados na forma de distribuição de freqUências facilita bastante o cálculo manual de várias medidas estatísticas de interesse, bem como a sua apresentação gráfica, consistindo em ferramenta à disposição do analista.

Caso o interesse do analista, além da determinação das frequências absolutas e relativas, se dirija à determinação da quantidade de observações que existe acima ou abaixo de um

7. A frequência absoluta (f) corresponde ao número de observações que temos em uma determinada classe ou em um determinado atributo de uma variável qualitativa, e a frequência relativa (fr) corresponde à proporção do número de observações em uma determinada classe em relação ao total de observações que temos. Esta freqência pode ser expressa em termos porcentuais. Para isto, basta multiplicar a frequência relativa obtida por 100.

determinado ponto na distribuição, teríamos o recurso da utilização da frequência acumulada[8]. A frequência acumulada, apresentada no Quadro 04, pode ser obtida da seguinte forma: abaixo do limite superior da primeira classe temos 18 registros que estão presentes nesta classe, como pode ser visto na distribuição de frequências absoluta.

Quando consideramos a segunda classe, a frequência acumulada corresponde ao número de pessoas que temos abaixo do limite superior desta classe, ou seja, os 8 registros da segunda classe mais os 18 da primeira classe totalizam 26 registros mensais de pluviometria abaixo de 122,8mm, correspondendo a 54,2% do total dos registros contidos na amostra. Para as outras classes, o raciocínio é semelhante.

Quadro 04: Frequências acumuladas para cada uma das classes.

Classe	LI	LS	f acum.	f acum.r
1a	-24,5	24,6	0	0,000
2a	24,6	73,7	18	0,375
3a	73,7	122,8	26	0,542
4a	122,8	171,9	32	0,667
5a	171,9	221	43	0,896
6a	221	270,1	47	0,979
7a	270,1	319,2	47	0,979
Total	Total	Total	48	1,000

Um exemplo típico de aplicação das distribuições de frequências acumuladas corresponde à identificação de uma determinada frequência abaixo ou acima de um determinado valor que não corresponde ao limite superior ou inferior de uma classe qualquer. Podemos, então, querer verificar qual a porcentagem de registros mensais de chuva com intensidade superior a 122,8mm. Para isto, basta consultar diretamente a frequência acumulada acima deste valor (100% – 54,2% = 45,8%), pois o valor 122,8mm/mês corresponde a um dos limites de classe apresentados nesta tabela. Para a determinação de frequências acumuladas correspondentes a valores não coincidentes a limites superiores ou inferiores de classes, o procedimento é o mesmo, bastando realizar a interpolação para a classe correspondente para encontrar a frequência acumulada desejada.

Por exemplo, para a determinação da frequência acumulada para registros mensais de pluviometria com intensidade superior a 100mm nos 48 meses que constituem o intervalo de tempo, referente à amostra estudada, teríamos:

73,7 mm --- 62,5%
100,0 mm --- X
122,8 mm --- 45,8%

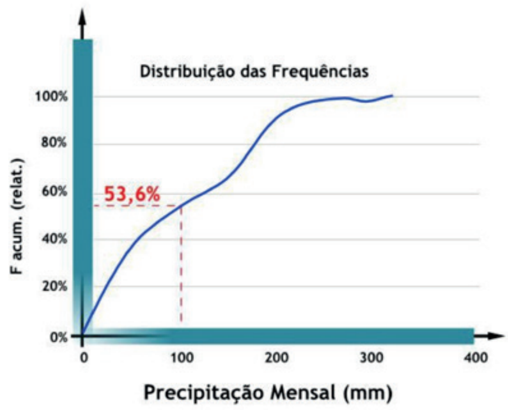
Logo:

$$\boxed{X} = \begin{array}{l}(122,8 - 100).(62,5 - 45,8)/(122,8 - 73,7) + 45,8 \\ = 16,7.22,8/49,1+45,8 = 53,6\%\end{array}$$

8. Freqüência acumulada é o total acumulado (soma) de todas as classes anteriores até a classe atual.

Ou seja, teríamos a frequência de 46,4% de registros mensais de precipitação pluviométrica superior a 100mm no período de 48 meses em questão. Na Figura 05 apresentada abaixo, teremos a representação gráfica da distribuição de frequências na forma relativa (percentuais) para o caso utilizado no exemplo, sendo destacado o valor correspondente à frequência de registros de precipitação pluviométrica com intensidade superior a 100mm/mês.

Figura 03- Representação gráfica da distribuição de frequências.



Interessante ressaltar que quando se trabalha com variáveis qualitativas, os atributos são as variações nominativas da variável, bem como a construção do quadro, que consiste em contar as ocorrências de cada atributo, e o resultado da contagem definirá a frequência absoluta do atributo em questão. Para exemplificar o caso do trabalho com variáveis qualitativas, teríamos o caso de pesquisa, na qual se procurou avaliar o número de habitações assistidas por esgotamento sanitário em uma determinada população, apresentando-se o Quadro 05 a seguir apresentado:

Quadro 05: Exemplo para frequências absoluta e relativa para variáveis qualitativas.

Esgotamento Sanitário	f_a	f_r
Atendimento	20	0,25
Não Atendimento	60	0,75
Total	80	1,0

Em relação às formas de representação gráfica de tabelas de frequência, é importante destacar que, dependendo do tipo de variável, há um gráfico mais adequado a ser utilizado para a representação. Os diferentes tipos de gráfico (histogramas, polígonos de frequência, ogivas, gráficos de setores, pictogramas e outros) permitem uma melhor visualização de resultados.

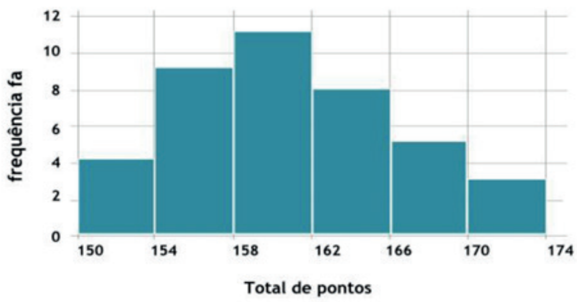
SAIBA MAIS

Estes gráficos podem ser obtidos utilizando planilhas eletrônicas como, por exemplo, o Excel®.

Os histogramas[9] são gráficos utilizados para representar tabelas intervalares. Já o polígono de frequência[10], você pode obter pela simples união dos pontos médios dos topos dos retângulos de um histograma, conforme mostra a Figura 04, que apresenta o histograma e os polígonos de frequência absoluta e acumulada para a distribuição dos dados referentes às classes correspondentes contidas no Quadro 06, apresentado a seguir:

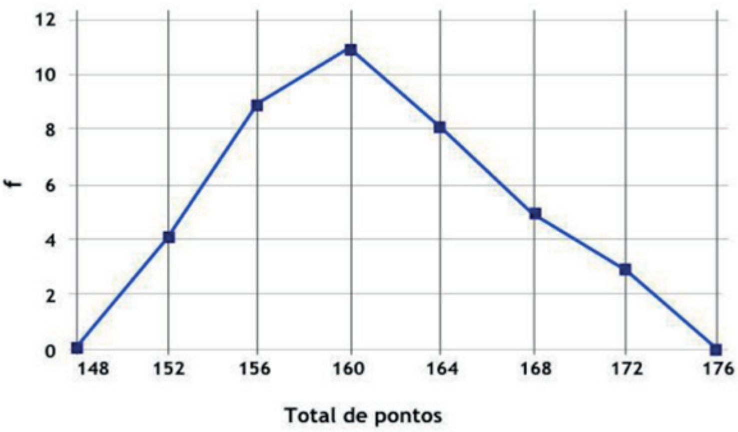
Quadro 06: Dados e frequências absoluta (fa) e acumulada (facum.) para 6 classes.

i	Classes	xi	fai	facumi
1	150 |- 154	152	4	4
2	154 |- 158	156	9	13
3	158 |- 162	160	11	24
4	162 |- 166	164	8	32
5	166 |- 170	168	5	37
6	170 |- 174	172	3	40
Total			40	



a) Histograma

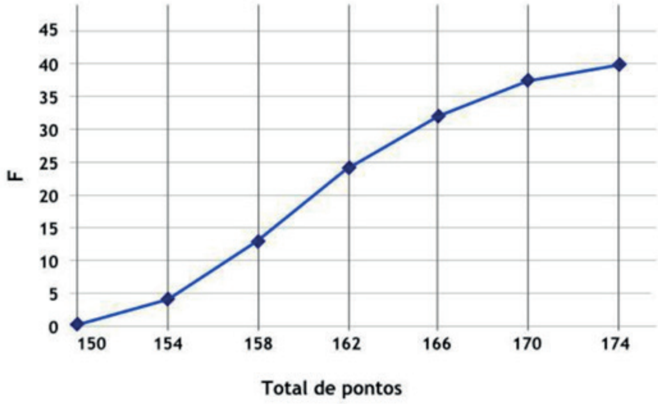
Figura 4a - Exemplos de Histograma



9. Histogramas: são constituídos por um conjunto de retângulos, com as bases assentadas sobre um eixo horizontal, tendo o centro da mesma no ponto médio da classe que representa, e cuja altura é proporcional à frequência da classe.
10. Polígono de frequências é um gráfico de análise no qual as frequências das classes são localizadas sobre perpendiculares levantadas nos pontos médios das classes. Completa-se o polígono unindo as extremidades da linha que ligam os pontos representativos das frequências de classe aos pontos médios das classes imediatamente anterior e posterior às classes extremas, que têm frequência nula.

b) Polígono de frequência absoluta

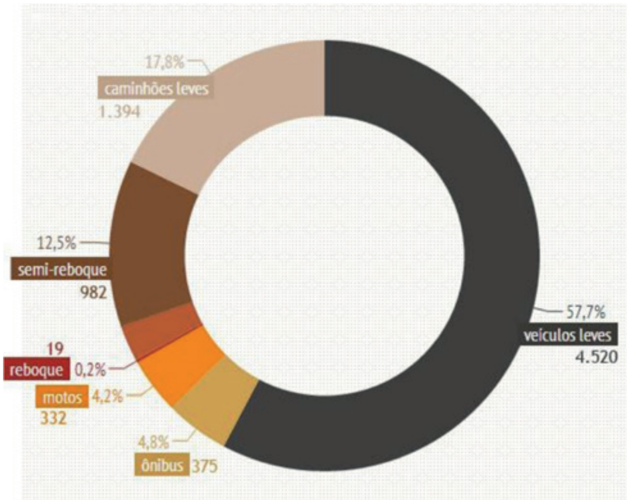
Figura 4b - Exemplos de Polígono de frequência absoluta



Gráficos para a representação de polígono de frequências acumuladas são chamados de ogivas e correspondem a um gráfico onde estas frequências são localizadas sobre perpendiculares, levantadas nos limites inferiores ou superiores das classes.

Para uma tabela de variável qualitativa, há um tipo de gráfico adequado para apresentar os resultados correspondes ao gráfico de setores, também popularmente conhecido como gráfico tipo pizza (Figura 05), com construção simples feita mediante a proporção entre o ângulo central do setor ("fatia da pizza") e o valor para a variável representada.

Figura 05- Composição média do tráfego para os postos de contagem na BR101-NE[11].



Após o estudo das formas de determinação das distribuições de frequências e gráficos que as representam, você deverá ser capaz de organizar um conjunto de dados por meio de uma distribuição de frequências (absoluta, relativa, e acumuladas), representá-las graficamente e proceder à análise das informações contidas nos mesmos.

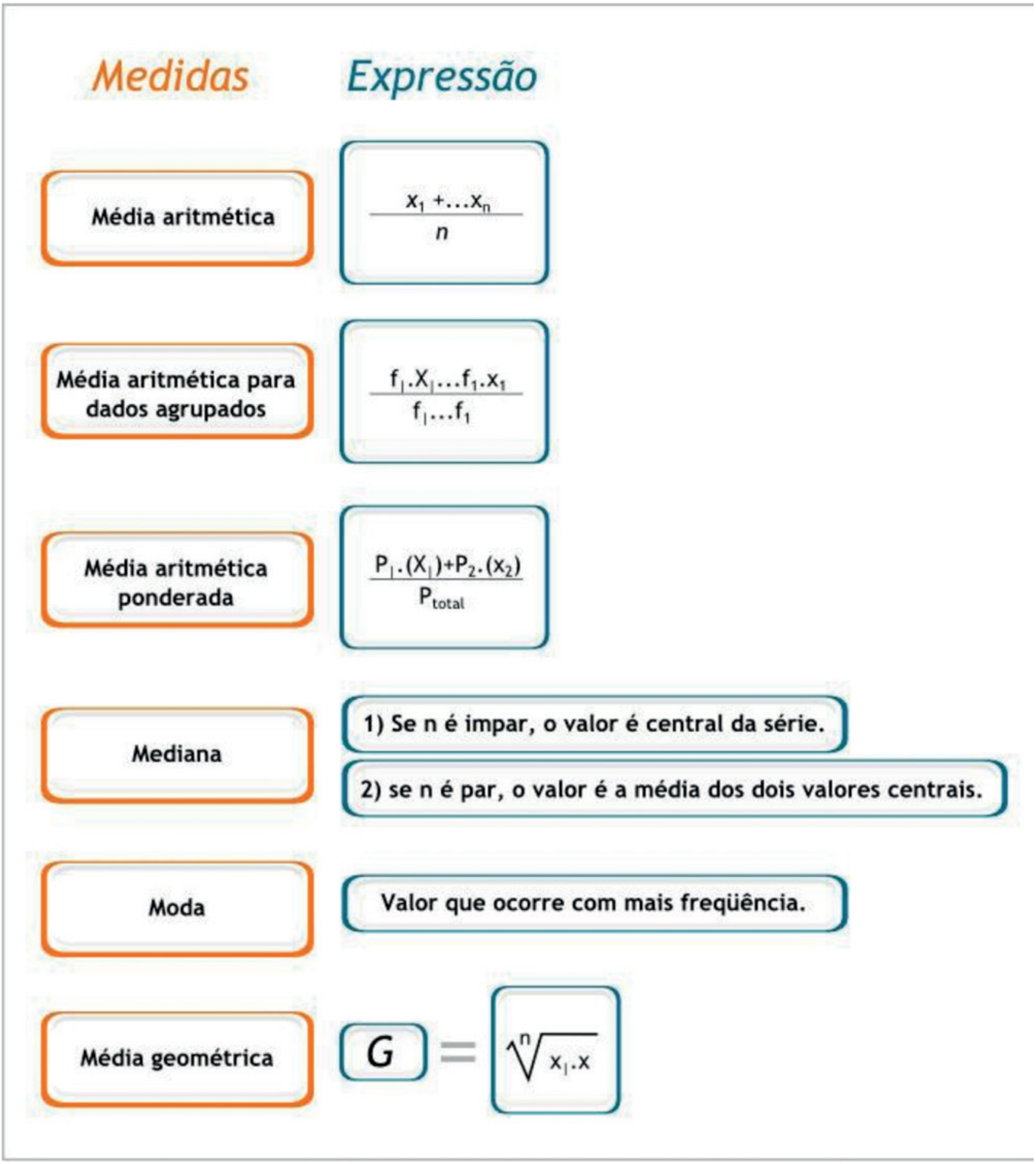
11. Pesquisa realizada pelo Exército Brasileiro (2005), mediante parceria com o Departamento Nacional de Infraestrutura de Transportes - DNIT, em 8 (oito) postos de contagem considerados no levantamento estatístico para o projeto de duplicação da BR101-NE.

1.3. MEDIDAS DE POSIÇÃO

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: média (aritmética, ponderada e geométrica), mediana e moda. Quando se estuda variabilidade, as medidas mais importantes são: amplitude, desvio padrão e variância.

No Quadro 07 mostrado abaixo, são apresentadas as expressões aritméticas para o cálculo das principais medidas de posição:

Quadro 07: Expressões para as principais medidas de tendência central.



A Utilização da Média

Sendo a Distribuição Normal uma das distribuições mais importantes e que surge com mais frequência nas aplicações (o que justifica a grande utilização da média), a média consistirá na melhor medida de localização do centro para uma série de dados. Entretanto, sendo a média uma medida bastante sensível à variabilidade dos dados, é preciso ter cuidado com a sua utilização, tendo em vista que pode propiciar uma imagem distorcida da amostra.

A média possui uma particularidade bastante interessante, que consiste no seguinte: se calcularmos os desvios de todas as observações relativamente à média e somarmos esses desvios, o resultado obtido é igual a zero.

Outra característica da média que torna a sua utilização vantajosa em certas aplicações é quando o que se pretende representar é a quantidade total expressa pelos dados, e então se utiliza a média. Na realidade, ao multiplicar a média pelo número total de elementos, obtemos a quantidade pretendida.

Moda

Define-se moda como sendo o valor que surge com mais frequência se os dados são discretos ou, ainda, o intervalo de classe com maior frequência se os dados são contínuos.

Assim, da representação gráfica dos dados, obtém-se imediatamente o valor que representa a moda ou a classe modal. Esta medida é especialmente útil para reduzir a informação de um conjunto de dados qualitativos, apresentados sob a forma de nomes ou categorias para os quais não se pode calcular a média e por vezes a mediana.

Para dados agrupados com classes, teríamos o seguinte processo para a determinação do valor modal para uma determinada série de dados:

1º. Identificação da classe de maior frequência:

Para o exemplo apresentado no Quadro 06, teríamos a 3ª classe (158 |- 162).

2º passo: Cálculo da Moda:

$$\boxed{Mo} = \boxed{(l_i + l_s) / 2 = (158 + 162) / 2 = 160}$$

Sendo:

l_i: limite inferior da classe modal = 158

l_s: limite superior da classe modal = 162

Mediana

A mediana é uma medida de localização do centro da distribuição dos dados, definida do seguinte modo: Ordenados os elementos da amostra, a mediana é o valor (pertencente ou não à amostra) que a divide ao meio, isto é, 50% dos elementos da amostra são menores ou iguais à mediana e os outros 50% são maiores ou iguais à mediana.

Para sua determinação, utiliza-se a seguinte regra, depois de ordenada a amostra de **n** elementos: Se **n** é ímpar, a mediana é o elemento médio e, se **n** é par, a mediana é a semi-soma dos dois elementos médios.

Teríamos, então, para dados não agrupados, o seguinte processo:

a) Quando o número de valores observados é ímpar:

Exemplo: Considere o conjunto de dados: $X = (5, 2, 7, 10, 3, 4, 1)$

1º) Coloque os valores em ordem crescente ou decrescente:

$$X = (1, 2, 3, 4, 5, 7, 10)$$

2º) Determine a ordem ou posição (P) da Mediana por P = (n+1)/2; portanto:

$$P = (7+1) / 2 = 4^a \text{ posição}$$

O número que se encontra na 4ª posição consiste na mediana, portanto, Md = 4

b) Quando o número de valores observados é par:

Exemplo: Considere o conjunto de dados:

$$X = (4, 3, 9, 8, 7, 2, 10, 6)$$

1º) Coloque os valores em ordem crescente ou decrescente:

$$X = (2, 3, 4, 6, 7, 8, 9, 10)$$

2º) Determine a ordem ou posição (P) para n/2 e para n/2 +1

$$P = 8/2 = 4^a \text{ posição} \quad \text{e} \quad P = 8/2 + 1 = 5^a \text{ posição}$$

Os números são 6 (4ª posição) e 7 (5ª posição): 2, 3, 4, 6, 7, 8, 9, 10

Tira-se a média aritmética entre os dois números:

$$\boxed{Md = (6+7) / 2 = 6{,}5}$$

Considerações a respeito de Média e Mediana

Como medida de localização, a mediana é mais robusta do que a média, pois não é tão sensível ao valor dos dados que compõem a amostra. Dentre as observações a respeito da comparação entre média e mediana, poderíamos destacar as seguintes:

Quando a distribuição é simétrica, a média e a mediana coincidem.

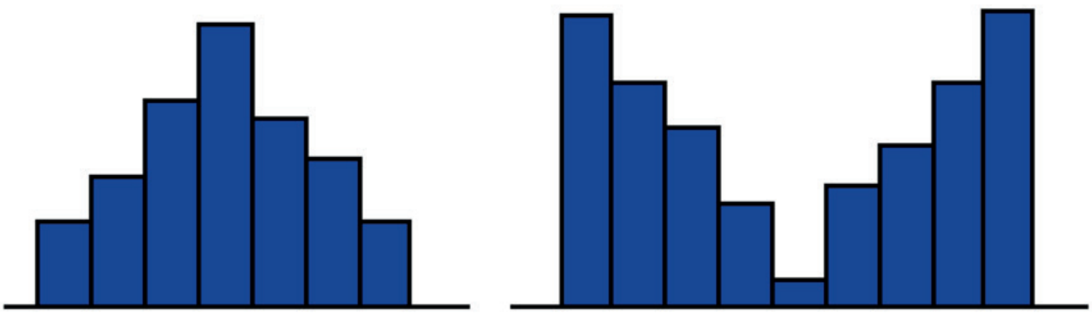
A mediana não é tão sensível, como a média, às observações que são muito maiores ou muito menores do que as restantes (*outliers*). Por outro lado, a média reflete o valor de todas as observações.

Assim, a média, ao contrário da mediana, é uma medida muito influenciada por valores “muito grandes” ou “muito pequenos”, mesmo que estes valores surjam em pequeno número na amostra. Estes valores são os responsáveis pela má utilização da média em muitas situações em que teria mais significado utilizar a mediana.

Portanto, teríamos as seguintes considerações sobre a influência da forma da distribuição dos dados:

Figura 05- Exemplos de distribuições simétricas.

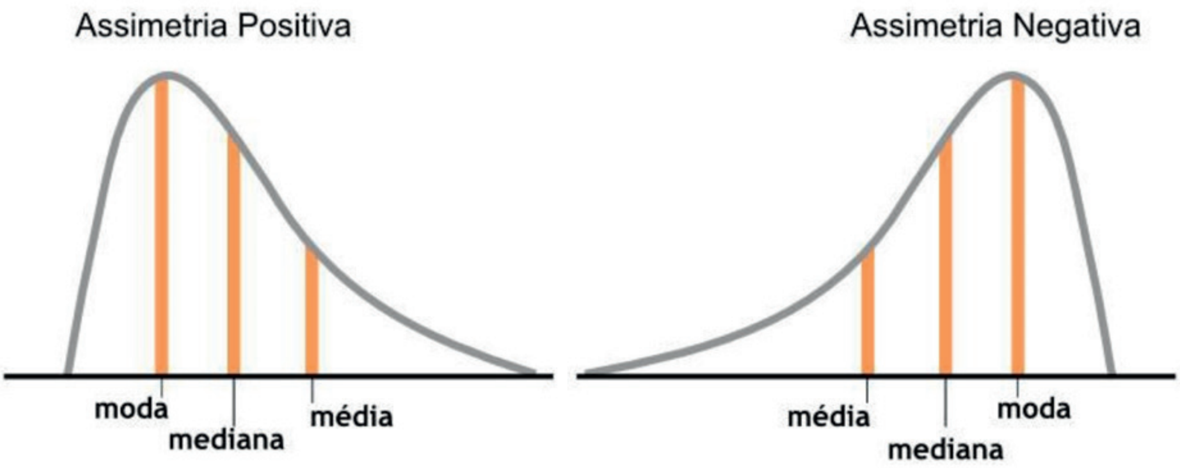
- Quando for aproximadamente simétrica, a média aproxima-se da mediana.



Quando se apresentar de forma enviesada para a direita (alguns valores grandes como *outliers*), a média tende a ser maior que a mediana.

Figura 06- Exemplos de distribuições assimétricas.

Caso a distribuição seja enviesada para a esquerda (alguns valores pequenos como *outliers*), a média tende a ser inferior à mediana.



1.4. MEDIDAS DE DISPERSÃO

No item anterior foram apresentados conceitos de algumas medidas de localização do centro de uma distribuição de dados. Veremos agora como medir a variabilidade presente num conjunto de dados. As medidas de dispersão são utilizadas para medir o grau de variabilidade (dispersão) dos valores observados em torno da média aritmética. Servem para medir a representatividade da média e proporcionam conhecer o nível de homogeneidade ou heterogeneidade dentro de cada grupo analisado.

Assim, um aspecto importante no estudo descritivo de um conjunto de dados é o da determinação da variabilidade ou dispersão desses dados, relativamente à medida de localização do centro da amostra. Supondo ser a média a medida de localização mais importante, será relativamente a ela que se define a principal medida de dispersão – a variância, apresentada a seguir.

Variância

Define-se a variância como sendo a medida que se obtém somando os quadrados dos desvios das observações da amostra, relativamente à sua média, e dividindo pelo número de observações da amostra menos um.

$$\boxed{\sum \frac{(x1 - \text{média})^2}{n - 1}}$$

Desvio-padrão

É a raiz quadrada da variância. Na fórmula original para o cálculo da variância, observa-se que é uma soma de quadrados, a unidade em que se exprime não é a mesma que a dos dados Por exemplo, se a unidade original for metro (m), o resultado será metro ao quadrado (m2).

Para retornar à unidade de medida original, extrai-se a raiz quadrada da variância, passando a chamar-se de desvio-padrão. Assim, para obter uma medida da variabilidade ou dispersão com as mesmas unidades que os dados, tomamos a raiz quadrada da variância e obtemos o desvio padrão.

$$\boxed{\sqrt{\frac{\Sigma(x1 - \text{média})^2}{n - 1}}}$$

O desvio padrão, portanto, é uma medida que só pode assumir valores não negativos e, quanto maior for, maior será a dispersão dos dados. O desvio padrão será maior, quanto mais variabilidade houver entre os dados.

Coeficiente de Variação

O coeficiente de variação (CV) consiste em uma medida relativa de dispersão, útil para a comparação em termos relativos ao grau de concentração em torno da média de séries distintas. Para uma amostra, teríamos a seguinte expressão:

$$\boxed{CV} = \boxed{\text{Desvio Padrão / Média x 100 (\%)}}$$

IMPORTANTE

O coeficiente de variação é expresso em porcentagem, avaliado para amostras segundo a seguinte referência:

Baixa dispersão: CV $\leq$ 15%

Média dispersão: 15% < CV < 30%

Grande dispersão: CV $\geq$ 30%

Distribuição Normal

A distribuição normal é a mais importante distribuição estatística, considerando a questão prática e teórica, apresentando-se em formato de sino, unimodal, simétrica em relação a sua média.

Considerando a probabilidade de ocorrência, a área sob sua curva soma 100%. Isso quer dizer que a probabilidade de uma observação assumir um valor entre dois pontos quaisquer é igual à área compreendida entre esses dois pontos.

Na figura apresentada a seguir, com as barras situadas logo abaixo do eixo das abscissas, representando os desvios-padrão, quanto mais afastado do centro da curva normal, mais área compreendida abaixo da curva haverá, ou seja, a um desvio-padrão, temos 68,26% das observações contidas, a dois desvios-padrões, possuímos 95,44% dos dados compreendidos e, finalmente, a três desvios, temos 99,73% de probabilidade de ocorrência.

Figura 07- Relação entre o desvio-padrão e a probabilidade de ocorrência de um evento.

p(x) para n desvios-padrão 68,26% => 1 desvio

95,44% => 2 desvios

99,73% => 3 desvios

O desvio-padrão, quando analisado isoladamente, não dá margem a muitas conclusões. Por exemplo, para uma distribuição cuja média é 79,7, como é o caso do exemplo com a série de dados de precipitação pluviométrica mensal, visto no início do curso, que apresentou desvio padrão de 76,9, considerado como bastante elevado, um desvio-padrão de 5mm/mês seria pequeno, mas para uma distribuição cuja média fosse 10, este desvio-padrão já não seria tão pequeno.

IMPORTANTE

Condições para se usar o desvio-padrão ou variância para comparar a variabilidade entre grupos:

- mesmo número de observações;
- mesma unidade;
- mesma média.

Além disso, se quisermos comparar duas ou mais amostras de valores expressos em unidades diferentes, não será ser possível fazer a comparação por meio do desvio-padrão, pois ele é expresso na mesma unidade dos dados.