

4

UNIDADE

Análise Exploratória de Dados II



Nesta Unidade você aprenderá mais uma maneira de descrever e analisar um conjunto de dados referente a uma variável quantitativa (discreta ou contínua): através das medidas de síntese. Serão apresentadas as medidas de posição e de dispersão que permitem sintetizar o comportamento da variável individualmente ou em função dos valores de outra variável.

Medidas de Posição ou de Tendência Central

Caro estudante!

Na Unidade 3 estudamos como fazer a descrição tabular e gráfica das variáveis, seja isoladamente ou relacionadas a outras, e interpretar os resultados obtidos. Além daquelas técnicas, nos casos em que a variável sob análise for **quantitativa discreta** ou **quantitativa contínua**, há uma terceira forma de descrição: as **medidas de síntese**, ou estatísticas. A sua utilização pode ser feita de forma a complementar as técnicas vistas na Unidade 3, ou como alternativa a elas.

As medidas de síntese subdividem-se em **medidas de posição (ou de tendência central)** e **medidas de dispersão**. Vamos estudar as medidas de posição: média, mediana, moda e quartis; e as medidas de dispersão: intervalo, variância, desvio padrão e coeficiente de variação percentual. Cada uma delas pode ser muito útil para caracterizar um conjunto de dados referente a uma variável quantitativa.

Tenha sempre em mente que é indispensável que o administrador conheça as medidas de síntese para que possa realizar Análise Exploratória de Dados através delas. Vamos ver que são ferramentas que geram resultados objetivos, o que torna mais racional o processo de tomada de decisão.

As Medidas de Posição procuram caracterizar a tendência central do conjunto, um valor numérico que “represente” o conjunto. Esse valor pode ser calculado levando em conta todos os valores do conjunto ou apenas alguns valores ordenados. As medidas mais importantes são média, mediana, moda e quartis.

Média (\bar{x})

A Média aqui citada é a **média aritmética simples**, a soma dos valores observados dividida pelo número desses valores. Seja um conjunto de **n** valores de uma variável quantitativa X, a média do conjunto será:

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

Onde x_i é um valor qualquer do conjunto, $\sum_{i=1}^n x_i$ é a soma dos valores do conjunto, e n é o tamanho do conjunto.

Vamos ver um exemplo que irá nos acompanhar por algum tempo. O Quadro 10 se refere às notas finais de três turmas de estudantes.

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Quadro 10: Notas finais das turmas A, B, e C.

Fonte: elaborado pelo autor.

Como o objetivo é calcular a média de cada turma, ao somar os valores teremos o mesmo resultado: 48. Como cada turma tem 8 alunos, as três turmas terão a mesma média: 6.

No exemplo que acabamos de ver as três turmas têm a mesma média (6), então se apenas essa medida fosse utilizada para caracterizá-las poderíamos ter a impressão de que as três turmas têm desempenhos idênticos. Será? Observe atentamente o Quadro 10.

Veja que na primeira turma temos realmente os dados distribuídos regularmente em torno da média, com a mesma variação tanto

No Microsoft Excel e no Br.Office Calc a média aritmética simples é implementada através da função MÉDIA().

Média aritmética simples – medida de posição que é o resultado da divisão da soma de todos os elementos do conjunto divididos pela quantidade de elementos do conjunto. Conceitualmente, é o centro de massa do conjunto de dados. Fonte: Barbetta (2007).

abaixo quanto acima. Já na segunda vemos uma distorção maior, embora, a maioria das notas seja alta algumas notas baixas “puxam” a média para um valor menor. E no terceiro grupo há apenas uma nota baixa, mas seu valor é tal que realmente consegue diminuir a média do conjunto.

Um dos problemas da utilização da média é que, por levar em conta todos os valores do conjunto, ela pode ser distorcida por **valores discrepantes** (“*outliers*”) que nele existam. É importante então interpretar corretamente o valor da média.

O valor da média pode ser visto como o centro de massa de cada conjunto de dados, ou seja, o ponto de equilíbrio do conjunto: “se os valores do conjunto fossem pesos sobre uma tábua, a média é a posição em que um suporte equilibra essa tábua”.

Vamos ver como os valores do exemplo distribuem-se em um diagrama apropriado (Figura 27):

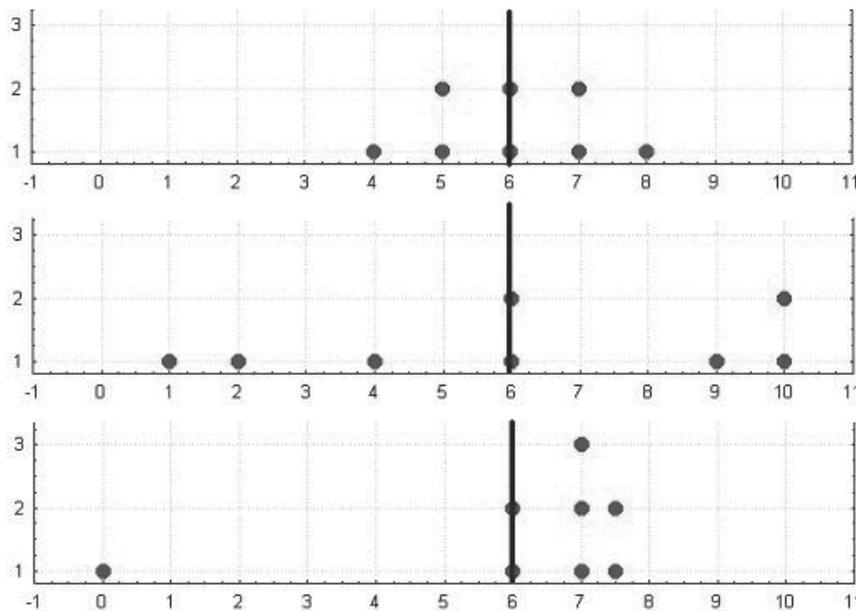


Figura 27: Interpretação do valor da média.

Fonte: adaptada pelo autor de Microsoft® (2003) e Statsoft® (2004).

A média dos três conjuntos é a mesma, mas observe as diferentes disposições dos dados. O primeiro grupo apresenta os dados distribuídos de forma **simétrica** em torno da média. No segundo grupo a distribuição já é mais irregular, com valores mais “distantes” na parte de baixo, e o terceiro grupo é claramente **assimétrica** em relação à média (que foi distorcida pelo valor discrepante 0). Portanto muito cuidado ao caracterizar um conjunto apenas por sua média.

Valores discrepantes (outliers) – valores de uma variável quantitativa que se distanciam muito (para cima ou para baixo) da maioria das observações. Por exemplo, a renda de Bill Gates é um valor discrepante da variável renda de pessoas morando nos EUA. Fonte: adaptado pelo autor de Bussab e Morettin (2003).

Assimétrica – uma distribuição dos valores de uma variável quantitativa é dita assimétrica caso a média e a mediana sejam diferentes, indicando que os valores do conjunto se estendem mais, apresentam maior variabilidade, em uma direção do que na outra. Fonte: Barbetta (2007).

Essa era a grande crítica que era feita nas décadas de 60 e 70 sobre as medições de nível de desenvolvimento. Era comum medir o nível de desenvolvimento de um país por sua renda per capita (PIB/número de habitantes), uma média, que não revelava, porém, a concentração de renda do país, levando a conclusões errôneas sobre a qualidade de vida em muitos países.

Outro aspecto importante a ressaltar é que a média pode ser um valor que a variável não pode assumir. Isso é especialmente verdade para variáveis quantitativas discretas, resultantes de contagem, como número de filhos, quando a média pode assumir um valor “quebrado”, 4,3 filhos, por exemplo.

Rompemos com o mito de que “média é o valor mais provável do conjunto”, erro que é cometido quase que diariamente pela mídia, em vários países.

É extremamente comum calcular médias de variáveis quantitativas a partir de distribuições de frequências representadas em tabelas: simplesmente multiplica-se cada valor (ou o ponto médio da classe) pela frequência associada, somam-se os resultados e divide-se o somatório pelo número de observações do conjunto. Na realidade trata-se de uma média ponderada pelas frequências de ocorrência de cada valor da variável.

$$\bar{x} = \frac{\sum_{i=1}^k (x_i \times f_i)}{n}$$

Onde k é o número de valores da variável discreta, ou o número de classes da variável agrupada, x_i é um valor qualquer da variável discreta, ou o ponto médio de uma classe qualquer, f_i é a frequência de um valor qualquer da variável discreta ou de uma classe qualquer, e n é o número total de elementos do conjunto.

Neste segundo exemplo vamos calcular a média do número de pessoas usualmente transportadas no veículo, através da distribuição de frequências obtida no terceiro exemplo exposto na Unidade 3 (Quadro 10).

VALORES	FREQUÊNCIA	PERCENTUAL
1	19	7,60%
2	29	11,60%
3	43	17,20%
4	42	16,80%
5	57	22,80%
6	60	24,00%
Total	250	100%

Quadro 11: Número de pessoas usualmente transportadas no veículo.

Fonte: elaborado pelo autor.

Precisamos multiplicar a coluna de valores x_i pela da frequência f_i , somar os resultados, e dividi-los por 250, que é o número de elementos do conjunto (n). Observe que a variável discreta pode assumir seis valores diferentes, logo $k = 6$. No Quadro 12 podemos observar o resultado:

VALORES X_i	FREQUÊNCIA f_i	$f_i \times x_i$
1	19	19
2	29	58
3	43	129
4	42	168
5	57	285
6	60	360
Total	250	1019

Quadro 12: Número de pessoas usualmente transportadas no veículo.

Fonte: elaborado pelo autor.

Agora podemos calcular a média:

$$\bar{x} = \frac{\sum_{i=1}^k (x_i \times f_i)}{n} = \frac{\sum_{i=1}^6 (x_i \times f_i)}{250} = \frac{1019}{250} = 4,076 \text{ pessoas usualmente transportadas}$$

no veículo.

Veja novamente a Figura 21 da Unidade 3, e observe como o va-

lor da média permite equilibrar os pesos, as frequências, dos vários valores da variável.

No Exemplo 2 o resultado da média é um valor (4,076) que a variável número de pessoas usualmente transportadas não pode assumir. Mas, se trata do centro de massa do conjunto

Se quisermos calcular média aritmética simples a partir de uma distribuição de frequências para dados agrupados devemos tomar cuidado. Os pontos médios das classes serão usados no lugar dos x_i da expressão da média vista acima. Eles podem ou não ser bons representantes das classes (geralmente serão melhores representantes quanto maiores forem as frequências das classes), pois perdemos a informação sobre o conjunto original de dados ao agrupá-lo em classes. Sendo assim, as medidas calculadas a partir de uma distribuição de frequências para dados agrupados, não apenas a média aritmética simples, mas todas as outras, tornam-se meras estimativas dos valores reais.

Importante! Não calcule nenhuma medida estatística com base em uma distribuição de frequência para dados agrupados se você tiver acesso aos dados originais.

Além da média aritmética simples, outra medida de posição bastante usada é a mediana, que veremos a seguir.

Mediana (M_d)

A **mediana** é o ponto que divide o conjunto em duas partes iguais: 50% dos dados têm valor menor do que a mediana e os outros 50% têm valor maior do que a mediana.

Pouco afetada por eventuais **valores discrepantes** existentes no conjunto (que costumam distorcer substancialmente o valor da média).

“A mediana de um conjunto de valores é o valor que ocupa a posição $(n + 1)/2$, quando os dados estão **ordenados** crescente ou decrescentemente. Se $(n + 1)/2$ for fracionário toma-se como mediana a média dos dois valores que estão nas posições imediatamente

abaixo e acima de $(n + 1)/2$ ". Onde n é o número de elementos do conjunto.

Neste terceiro exemplo vamos calcular a mediana para as notas das três turmas do Exemplo 1.

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Quadro 13: Notas finais das turmas A, B, e C.

Fonte: elaborado pelo autor.

Posição mediana = $(n + 1)/2 = (8+1)/2 = 4,5^a$ significa que o valor da mediana será calculado através da média entre os valores que estiverem na 4ª e na 5ª posição do conjunto.

Turma A: $Md = (6 + 6)/ 2 = 6.$

Turma B: $Md = (6 + 6)/ 2 = 6.$

Turma C: $Md = (7 + 7)/ 2 = 7.$

Observe que a mediana da Turma C é diferente, mais alta, refletindo melhor o conjunto de dados, uma vez que há apenas uma nota baixa. Perceba também que apenas os dois valores centrais foram considerados para obter a mediana, deixando o resultado “imune” aos valores discrepantes.

No exemplo 4, vamos Calcular a mediana para o grupo a seguir:

10 11 12 13 15 16 16 35 60.

Posição mediana = $(n + 1)/2 = (9+1)/2 = 5^a$ como o conjunto tem um número ímpar de valores, o valor da mediana será igual ao valor que estiver na quinta posição.

Mediana = 15

Média = 20,89

Observe que neste caso, média e mediana são diferentes, pois a média foi distorcida pelos valores mais altos 35 e 60, que constituem uma minoria. Neste caso a medida de posição que melhor representaria o conjunto seria a mediana. Se a média é diferente da mediana a distribuição da variável quantitativa no conjunto de dados é dita **assimétrica**.

Por esse motivo os dados precisam estar ordenados crescentemente.

No Microsoft Excel® e no Br.Office Calc ® a mediana é implementada através da função MED(), tal como explicado no texto “Como realizar análise exploratória de dados no Microsoft Excel®”.

Tal como a média, a mediana pode ser calculada a partir de uma tabela de frequências, com as mesmas ressalvas feitas para aquela medida. Os programas estatísticos, e muitas planilhas eletrônicas dispõem de funções que calculam a mediana.

Moda (Mo)

A **moda** é o valor da variável que ocorre com maior frequência no conjunto. Pode então ser considerado o mais provável.

É a medida de posição de obtenção mais simples, e também pode ser usada para variáveis qualitativas, pois apenas registra qual é o valor mais freqüente, podendo este valor ser tanto um número quanto uma categoria de uma variável nominal ou ordinal.

Um conjunto pode ter apenas uma Moda, várias Modas ou nenhuma Moda. Esse último caso geralmente ocorre com variáveis quantitativas contínuas.

A proposta no exemplo 5 é encontrar a moda das notas das três turmas do Exemplo 1 (Quadro 14).

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Quadro 14: Notas finais das turmas A, B, e C.

Fonte: elaborado pelo autor.

A turma A tem 3 modas: os valores 5, 6 e 7 ocorrem duas vezes cada. A turma B tem duas modas: os valores 6 e 10 ocorrem duas vezes cada. A turma C tem uma moda apenas: o valor 7 ocorre 3 vezes.

Quartis

Para alguns autores os **quartis** não são medidas de posição, são separatrizes. Porém, como sua forma de cálculo é semelhante a da mediana, resolvemos incluí-los no tópico de Medidas de Posição. Os quartis são medidas que dividem o conjunto em quatro partes iguais.

O primeiro quartil ou **quartil inferior (Qi)** é o valor do conjunto que delimita os 25% menores valores: 25% dos valores são menores do que **Qi** e 75% são maiores do que **Qi**.

O segundo quartil ou **quartil do meio** é a própria mediana (**Md**), que separa os 50% menores dos 50% maiores valores.

O terceiro quartil ou **quartil superior (Qs)** é o valor que delimita os 25% maiores valores: 75% dos valores são menores do que **Qs** e 25% são maiores do que **Qs**.

Como são medidas baseadas na ordenação dos dados é necessário, primeiramente, calcular as posições dos quartis.

$$\text{Posição do quartil inferior} = (n + 1)/4.$$

$$\text{Posição do quartil superior} = [3 \times (n + 1)]/4.$$

Onde **n** é o número total de elementos do conjunto.

Após calcular a posição, encontrar o elemento do conjunto que nela está localizado. O conjunto de dados precisa estar ordenado! Se o valor da posição for fracionário deve-se fazer a média entre os dois valores que estão nas posições imediatamente anterior, e imediatamente posterior à posição calculada. Se os dados estiverem dispostos em uma distribuição de frequências, utilizar o mesmo procedimento observando as frequências associadas a cada valor (variável discreta) ou ponto médio de classe.

No exemplo 6 iremos encontrar os quartis para a renda no conjunto de dados apresentados no Quadro 15:

No Microsoft Excel® e no Br.Office Calc® os quartis são implementados através da função QUARTIL(;1) para quartil inferior, e QUARTIL(;3) para quartil superior.

VALORES									
4,695	5,750	7,575	12,960	13,805	14,000	15,820	18,275	18,985	18,985
19,595	19,720	20,600	22,855	22,990	23,685	24,400	24,400	24,685	24,980
24,980	26,775	27,085	27,240	28,340	31,480	40,050	43,150	47,075	

Quadro 15: Renda em salários mínimos.

Fonte: elaborado pelo autor.

Há 29 elementos no conjunto, que já estão ordenados crescentemente. Podemos calcular as posições dos quartis.

$$\text{Posição do quartil inferior} = (n + 1)/4 = (29 + 1)/4 = 7,5^{\text{a}}.$$

$$\text{Posição do quartil superior} = [3 \times (n + 1)]/4 = [3 \times (29 + 1)]/4 = 22,5^{\text{a}}.$$

Para encontrar o quartil inferior precisamos calcular a média dos valores que estão na 7^a e 8^a posição do conjunto: no caso, 15,820 e 18,275, resultando:

$$Q_i = (15,820 + 18,275)/2 = 17,0475.$$

Imagine que fosse um grande conjunto de dados, referente a salários de uma população: apenas 25% dos pesquisados teriam renda **abaixo** de 17,0475 salários mínimos (ou R\$ 6478,05 pelo salário mínimo de maio de 2007). Com base nisso poderíamos ter uma ideia do nível de renda daquela população.

Para encontrar o quartil superior precisamos calcular a média dos valores que estão na 22^a e 23^a posição do conjunto: no caso, 15,820 e 18,275, resultando:

$$Q_s = (26,775 + 27,085)/2 = 26,93.$$

Novamente, imagine que fosse um grande conjunto de dados, referente a salários de uma população: apenas 25% dos pesquisados teriam renda acima de 26,93 salários mínimos (ou R\$ 10.233,40 pelo salário mínimo de maio de 2007).

Com todas as medidas de posição citadas, já é possível obter um retrato razoável do comportamento da variável. Mas as medidas de posição são insuficientes para caracterizar adequadamente um conjunto de dados. É preciso calcular também medidas de dispersão.

Medidas de dispersão ou de variabilidade

O objetivo das medidas de dispersão é medir quão próximos uns dos outros estão os valores de um grupo (e algumas mensuram a dispersão dos dados em torno de uma medida de posição). Com isso é obtido um valor numérico que sintetiza a variabilidade.

Vamos estudar o intervalo, a variância, o desvio padrão e o coeficiente de variação percentual.

Intervalo

O intervalo é a medida mais simples de dispersão. Consiste em identificar os valores extremos do conjunto (mínimo e máximo), podendo ser expresso:

- pela diferença entre o valor máximo e o mínimo; e
- pela simples identificação dos valores.

O intervalo é muito útil para nos dar uma ideia da variabilidade geral do conjunto de dados. Alguém que calculasse o intervalo da variável renda mensal familiar no Brasil provavelmente ficaria abismado pela gigantesca diferença entre o valor mais baixo e o mais alto. Se essa mesma pessoa fizesse o mesmo cálculo na Noruega a diferença não seria tão grande.

No exemplo 7 vamos obter o Intervalo para os conjuntos de notas das duas turmas apresentadas no Quadro 16:

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	4 4 4,2 4,3 4,5 5 5 8

Quadro 16: Notas das turmas A e B.

Fonte: elaborado pelo autor.

Medidas de dispersão – medidas numéricas que visam avaliar a variabilidade do conjunto de dados, sintetizando-a em um número. Fonte: elaborado pelo autor.

O intervalo será o mesmo para ambas as turmas: [4,8] ou 4.

Observe que no Exemplo 7 as duas turmas apresentam o mesmo intervalo (4). Mas observando os dados percebemos facilmente que a dispersão dos dados tem comportamento diferente nas duas turmas, e essa é principal desvantagem do uso do intervalo como medida de dispersão.

Se colocarmos os dados do Exemplo 7 em um diagrama apropriado (Figura 28):

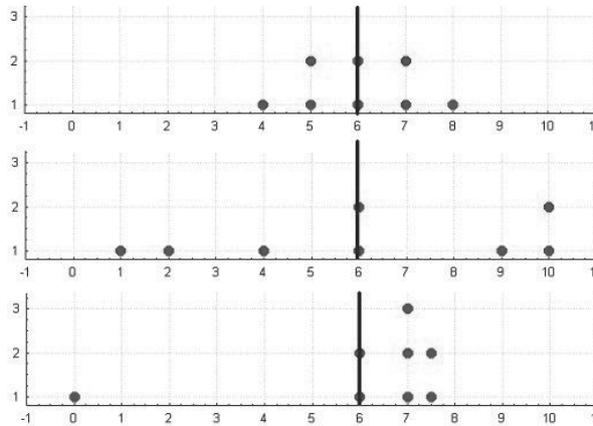


Figura 28: Desvantagem do uso do intervalo como medida de dispersão.

Fonte: adaptada pelo autor de Statsoft® e Microsoft®.

No Microsoft Excel® e no Br.Office Calc® podemos obter o Intervalo através das funções MÁXIMO () e MÍNIMO ().

Observamos claramente que os dados da turma A apresentam uma dispersão bem mais uniforme do que os da turma B, embora ambos os conjuntos tenham o mesmo intervalo. O intervalo não permite ter ideia de como os dados estão distribuídos entre os extremos (não permite identificar que o valor 8 na turma B é um valor discrepante).

Torna-se necessário obter outras medidas de dispersão, capazes de levar em conta a variabilidade entre os extremos do conjunto, o que nos leva a estudar variância e desvio padrão.

Variância (s²)

A variância é uma das medidas de dispersão mais importantes. É a média aritmética dos quadrados dos desvios de cada valor em relação à média: proporciona uma mensuração da dispersão dos dados em torno da média.

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

Onde x_i é um valor qualquer do conjunto, \bar{x} é a média do conjunto e n é o número de elementos do conjunto. Se os dados referem-se a uma POPULAÇÃO usa-se n no denominador da expressão.

Você sabe por que é preciso elevar os desvios ao quadrado para avaliar a dispersão? Não podemos apenas somar os desvios dos valores em relação à média do conjunto? Deixamos como exercício para você os cálculos dos desvios (diferença entre cada valor e a média) para as notas das três turmas descritas no quadro 10, do Exemplo 1. Após calcular os desvios, some-os e veja os resultados. Lembre-se de que a média é o centro de massa do conjunto.

A razão dessa distinção será explicada mais adiante na Unidade 7. Podemos adiantar que a utilização de $n - 1$ no denominador é indispensável para que a variância da variável na amostra possa ser um bom estimador da variância da variável na população.

A unidade da variância é o quadrado da unidade dos dados e, portanto, o quadrado da unidade da média, causando dificuldades para avaliar a dispersão: se por exemplo temos a variável peso com média de 75 kg em um conjunto e ao calcular a variância obtemos 12 kg² a avaliação da dispersão torna-se difícil. Não obstante, a variância e a média são as medidas geralmente usadas para caracterizar as distribuições probabilísticas (que serão vistas adiante, na Unidade 6).

O que podemos afirmar, porém, é que quanto maior a variância, mais dispersos os dados estão em torno da média (maior a dispersão do conjunto).

Para fins de Análise Exploratória de Dados, caracterizar a dispersão através da variância não é muito adequado. Costumamos usar a raiz quadrada positiva da variância, o desvio padrão. Vamos ver mais sobre isso? Continuemos, então, a estudar!

No Microsoft Excel® e no Br.Office Calc® a variância populacional é obtida através da função VARP(), e a variância amostral através da função VAR().

Desvio Padrão (s)

É a raiz quadrada positiva da variância, apresentando a mesma unidade dos dados e da média, permitindo avaliar melhor a dispersão.

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

As mesmas observações sobre população e amostra feitas para a variância são válidas para o desvio padrão. É prática comum ao resumir através de várias medidas de síntese um conjunto de dados referente a uma variável quantitativa, apresentar apenas a média e o desvio padrão desse conjunto, para que seja possível ter uma ideia do valor típico e da distribuição dos dados em torno dele.

Deixamos como exercício para você elevar os desvios obtidos com os dados das turmas, expressos no Quadro 10, Exemplo 1, ao quadrado, somá-los e dividi-los por 7 (suponha que se trata de uma amostra). Assim, você obterá os desvios padrões das notas das turmas.

O desvio padrão pode assumir valores menores do que a média, da mesma ordem de grandeza da média, ou até mesmo maiores do que a média. Obviamente se todos os valores forem iguais, não haverá variabilidade e o desvio padrão será igual a zero.

A fórmula acima costuma levar a consideráveis erros de arredondamento, basicamente porque exige o cálculo prévio da média. Se o valor desta for uma dízima um arredondamento terá que ser feito, causando um pequeno erro, e esse erro será propagado pelas várias operações de subtração (de cada valor em relação à média) e potenciação (elevação ao quadrado da diferença entre cada valor e a média). Assim a fórmula é modificada para reduzir o erro de arredondamento apenas ao resultado final:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i^2) - \frac{\left[\sum_{i=1}^n x_i\right]^2}{n}}{n-1}} \quad (\text{amostra}).$$

Primeiramente cada valor (x_i) do conjunto é elevado ao quadrado e somam-se todos os resultados obtendo $\sum_{i=1}^n x_i^2$. Somam-se tam-

bém todos os valores do conjunto para obter $\sum_{i=1}^n x_i$, somatório este que será elevado ao quadrado. Os somatórios e o valor de **n** (número de elementos no conjunto) são substituídos na fórmula para obter os resultados.

É desta forma que os programas computacionais calculam o desvio padrão.

Tal como no caso da média pode haver interesse em calcular o desvio padrão de variáveis quantitativas a partir de distribuições de frequências representadas em tabelas. Tal como no caso da média os valores da variável (ou os pontos médios das classes), e os quadrados desses valores, serão multiplicados por suas respectivas frequências:

$$s = \sqrt{\frac{\sum_{i=1}^k (x_i^2 \times f_i) - \frac{\left(\sum_{i=1}^k x_i \times f_i\right)^2}{n}}{n-1}} \quad (\text{amostra}).$$

Onde x_i é o valor da variável ou ponto médio da classe, f_i a frequência associada, k é o número de valores da variável discreta (ou o número de classes da variável agrupada), e n é o número de elementos do conjunto.

No Microsoft Excel® e no Br.Office Calc® podemos obter o desvio padrão populacional através da função DESVPADP() e amostral através da função DESVPAD().

Veremos neste oitavo exemplo como calcular o desvio padrão da renda para os dados do Exemplo 6.

VALORES									
4,695	5,750	7,575	12,960	13,805	14,000	15,820	18,275	18,985	18,985
19,595	19,720	20,600	22,855	22,990	23,685	24,400	24,400	24,685	24,980
24,980	26,775	27,085	27,240	28,340	31,480	40,050	43,150	47,075	

Quadro 17: Renda em salários mínimos.

Fonte: elaborado pelo autor.

Há 29 elementos no conjunto, $n = 29$.

Somando os valores vamos obter: $\sum_{i=1}^n x_i = \sum_{i=1}^{29} x_i = 654,935$

Elevando cada valor ao quadrado e somando-os vamos obter:

$$\sum_{i=1}^n x_i^2 = \sum_{i=1}^{29} x_i^2 = 17497,919125$$

Agora basta substituir os somatórios na expressão e calcular o desvio padrão, supondo que se trata de uma amostra:

$$s = \sqrt{\frac{\sum_{i=1}^{29} (x_i^2) - \left[\frac{\left(\sum_{i=1}^{29} x_i \right)^2}{29} \right]}{29-1}} = \sqrt{\frac{17497,919125 - \left[\frac{(654,935)^2}{29} \right]}{29-1}} =$$

$$s = \sqrt{\frac{17497,919125 - 14791,02946}{28}}$$

$s \cong 9,83$ salários mínimos.

Se calcularmos a média, obteremos 22,584 salários mínimos. Observe que o desvio padrão é menor do que a média, não chega à metade. Com base nisso poderíamos avaliar a variabilidade do conjunto.

Quanto menor o desvio padrão, mais os dados estão concentrados em torno da média. Pensando nisso, alguém teve a ideia de criar uma medida de dispersão que relacionasse média e desvio padrão, o coeficiente de variação percentual, que veremos a seguir.

Coeficiente de Variação Percentual (c.v.%)

O coeficiente de variação percentual é uma medida de dispersão relativa, pois permite comparar a dispersão de diferentes distribuições (com diferentes médias e desvios padrões).

$$\text{c.v. \%} = \frac{s}{\bar{x}} \times 100\%$$

Onde s é o desvio padrão da variável no conjunto de dados, e \bar{x} é a média da variável no mesmo conjunto.

Quanto menor o coeficiente de variação percentual, mais os dados estão concentrados em torno da média, pois o desvio padrão é pequeno em relação à média.

Neste exemplo, vamos calcular o coeficiente de variação percentual para as notas das turmas do Exemplo 1, e indicar qual das três apresenta as notas mais homogêneas.

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Quadro 18: Notas finais das turmas A, B, e C.

Fonte: elaborado pelo autor.

Coeficiente de variação percentual – resultado da divisão do desvio padrão pela média do conjunto, multiplicado por 100, permite avaliar o quanto o desvio padrão representa da média. Fonte: Barbetta, Reis e Bornia (2008); Anderson, Sweeney e Williams (2007).

Para a turma A: $\bar{x} = 6$ $s = 1,31$ $c.v.\% = (1,31/6) \times 100 = 21,82\%$.

Para a turma B: $\bar{x} = 6$ $s = 3,51$ $c.v.\% = (3,51/6) \times 100 = 58,42\%$.

Para a turma C: $\bar{x} = 6$ $s = 2,49$ $c.v.\% = (2,49/6) \times 100 = 41,55\%$.

A turma mais homogênea é a A, pois apresenta o menor coeficiente de variação das três. Isso era esperado, uma vez que as notas da turma A estão distribuídas mais regularmente do que as das outras.

No caso apresentado anteriormente a comparação ficou ainda mais simples, pois as médias dos grupos eram iguais, bastaria avaliar apenas os desvios padrões dos grupos, mas para comparar a dispersão de distribuições com médias diferentes é imprescindível a utilização do coeficiente de variação percentual.

Você deve se perguntar, mas porque é tão importante calcular a média e o desvio padrão dos valores de uma variável registrados em um conjunto de dados? Argumentam que talvez a mediana seja uma melhor medida de posição, e que os quartis permitem ter uma boa ideia da dispersão. Contudo há um teorema que permite, a partir da média e do desvio padrão, obter estimativas dos extremos do conjunto, especialmente quando se trata de uma amostra: é o teorema de Chebyshev, também chamado de Desigualdade de Chebyshev.

há um pequeno texto sobre o Teorema de Chebyshev no ambiente virtual.

Cálculo de medidas de síntese de uma variável em função dos valores de outra

Na Unidade 3, estudamos como analisar em conjunto uma variável quantitativa e outra qualitativa. Naquela ocasião mostramos como os dados da variável quantitativa poderiam ser avaliados em função dos valores da variável qualitativa, uma vez que esta costuma ter menos opções, possibilitando resumir mais o conjunto.

Recomendamos que você olhe novamente o oitavo exemplo da Unidade 3 verá que construímos distribuições de frequências agrupadas em classes, para a variável renda (quantitativa), em função dos valores da variável modelo (qualitativa). Poderíamos fazer o mesmo com as medidas de síntese! Vamos ver o exemplo a seguir.

Para a mesma situação dos Exemplos 1 e 7 da Unidade 3. Gostaríamos de avaliar, neste exemplo, se existe algum relacionamento entre a renda do consumidor e o modelo adquirido. Esperamos que exista tal relacionamento, pois os modelos Chiconaultla e DeltaForce3 são os mais baratos, e o sofisticado LuxuriousCar é o mais caro de todos.

Através do Microsoft Excel® e do Br.Office Calc®, podemos calcular várias medidas de síntese da variável renda, em função dos modelos de veículos. O Excel® permite obter as seguintes medidas em função dos valores de outra variável: média, desvio padrão (amostral e populacional), variância (amostral e populacional), mínimo e máximo (infelizmente não permite cálculo de mediana ou quartis). O Calc permite obter as mesmas medidas, mas para cada uma delas é necessário acionar o assistente de dados, enquanto no Excel é possível agrupá-las em uma única tabela. Ao realizar este procedimento, usando os dados do arquivo AmostraToyord.xls vamos obter (Quadro 19):

MODELO	MEDIDA	VALOR
Chiconaultla	Frequência	81
	Mínimo	1,795
	Máximo	40,160
	Média	12,704
	Desvio padrão (amostral)	6,038
DeltaForce3	Frequência	56
	Mínimo	10,820
	Máximo	48,220
	Média	22,063
	Desvio padrão (amostral)	6,956

Quadro 19: Medidas de síntese de Renda por Modelo.
Fonte: elaborado pelo autor.

MODELO	MEDIDA	VALOR
LuxuriousCar	Frequência	29
	Mínimo	29,800
	Máximo	86,015
	Média	50,932
	Desvio padrão (amostral)	14,922
SpaceShuttle	Frequência	42
	Mínimo	18,865
	Máximo	47,300
	Média	33,050
	Desvio padrão (amostral)	7,620
Valentiniana	Frequência	41
	Mínimo	13,055
	Máximo	65,390
	Média	27,353
	Desvio padrão (amostral)	8,383
Frequência Total		249
Mínimo Total		1,795
Máximo Total		86,015
Média Total		25,105
Desvio padrão (amostral) Total		14,505

Quadro 19: Medidas de síntese de Renda por Modelo.

Fonte: elaborado pelo autor.

Se analisarmos as medidas de renda para os cinco modelos vamos identificar alguns aspectos interessantes:

- os mínimos de Chiconaultla e DeltaForce3 são efetivamente menores do que os dos outros modelos (o mínimo de Chiconaultla é o menor do conjunto todo);
- o mínimo de LuxuriousCar é o maior de todos, e seu máximo também (sendo o valor máximo do conjunto todo);
- quanto às médias podemos observar um comportamento na seguinte ordem crescente: Chiconaultla, DeltaForce3, Valentiniana, SpaceShuttle e LuxuriousCar; e
- a média de renda dos clientes do LuxuriousCar é quase quatro vezes maior do que as dos compradores do Chiconaultla.

Portanto, o relacionamento entre renda e modelo parece realmente existir.

Agora devemos avaliar a dispersão da renda em função dos modelos. Como as médias são diferentes é recomendável calcular os coeficientes de variação percentual, mostrados no Quadro 20.

MODELO	MEDIDA	VALOR
Chiconaultla	Coeficiente de Variação Percentual	47,526%
DeltaForce3	Coeficiente de Variação Percentual	31,528%
LuxuriousCar	Coeficiente de Variação Percentual	29,298%
SpaceShuttle	Coeficiente de Variação Percentual	23,054%
Valentiniana	Coeficiente de Variação Percentual	30,646%
Coeficiente de Variação Percentual Total		57,777%

Quadro 20: Coeficientes de Variação Percentual de Renda por Modelo.

Fonte: elaborado pelo autor.

Aparentemente, a relação existente entre renda média e os modelos não se reproduz completamente no que tange à dispersão. Embora o Chiconaultla (modelo mais barato e cujos compradores tem a média mais baixa de renda) tenha o maior coeficiente de variação percentual (47,526%), o modelo mais sofisticado, LuxuriousCar, cujos compradores têm a média mais alta, não apresenta o menor coeficiente de variação percentual. O modelo cujos compradores possuem a renda mais concentrada em torno da média é o SpaceShuttle, cujo coeficiente de variação percentual vale 23,054%. Podemos concluir que, embora o Chiconaultla seja um modelo mais “simples”, teoricamente visando um público de menor renda, ele também é adquirido por compradores mais abastados. Já o SpaceShuttle tem compradores de nível mais elevado (segunda maior média de renda), com pouca variação entre eles.

Utilizando um *software* estatístico podemos calcular outras medidas além das mostradas nos Quadros anteriores. No nosso caso, usando o Statsoft Statistica 6.0®, podemos obter:

MODELO	MEDIDAS							
	\bar{X}	FREQ.	s	MIN.	MAX.	Q _i	M _D	Q _s
Deltaforce3	22,064	56	6,956	10,82	48,22	16,575	21,378	26,392
SpaceShuttle	33,05	42	7,62	18,865	47,3	26,62	33,85	39,65
Valentiniana	27,353	41	8,383	13,055	65,39	23,685	25,715	30,13
Chiconaultla	12,705	81	6,038	1,795	40,16	8,88	12,245	15,4
LuxuriousCar	50,932	29	14,922	29,800	86,015	41,89	47,525	58,92
Total	25,105	249	14,505	1,795	86,015	14,095	23,545	32,17

Quadro 21: Medidas de síntese de Renda por Modelo.

Fonte: adaptado pelo autor de Statsoft®.

Observe que as medianas, quartis inferiores e superiores comportam-se de forma semelhante às médias. A propósito, médias e medianas são próximas, o que indicaria simetria das distribuições das rendas para todos os modelos.

Saiba mais...

Sobre medidas de síntese, assimetria, diagramas em caixa e outros aspectos, procure em: BARBETTA, Pedro A. *Estatística Aplicada às Ciências Sociais*. 7. ed. Florianópolis: Ed. da UFSC, 2007, Capítulo 7.

Sobre outros tipos de médias (harmônica, geométrica), SPIEGEL, Murray R. *Estatística*. 3. ed. São Paulo: Makron Books, 1993, Capítulo 3.

Sobre outros aspectos de Análise Exploratória de Dados com medidas de síntese, teorema de Chebyshev e assimetria, ANDERSON, David R.; SWEENEY, Dennis J.; WILLIAMS, Thomas A. *Estatística Aplicada à Administração e Economia*. 2. ed. São Paulo: Thomson Learning, 2007, Capítulo 3.

Sobre Análise Exploratória de Dados utilizando o Excel, LEVINE, David M.; STEPHAN, David; KREHBIEL, Timothy C.; BERENSON, Mark L. *Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português*. 5. ed. Rio de Janeiro: LTC, 2005.

Para saber como realizar as análises descritas nesta Unidade e na Unidade 4 através do Microsoft Excel® consulte *Como realizar análise exploratória de dados no Microsoft Excel®*, disponível no Ambiente Virtual de Ensino-Aprendizagem, assim como o arquivo de dados usado nos exemplos apresentados.

Sobre como realizar as análises descritas nesta Unidade e na Unidade 4 através do Br.Office Calc® consulte *Como realizar análise exploratória de dados com o Br.Office Calc®*, disponível no Ambiente Virtual assim como o arquivo de dados usado nos exemplos apresentados.



O resumo desta Unidade está demonstrado na Figura 29:

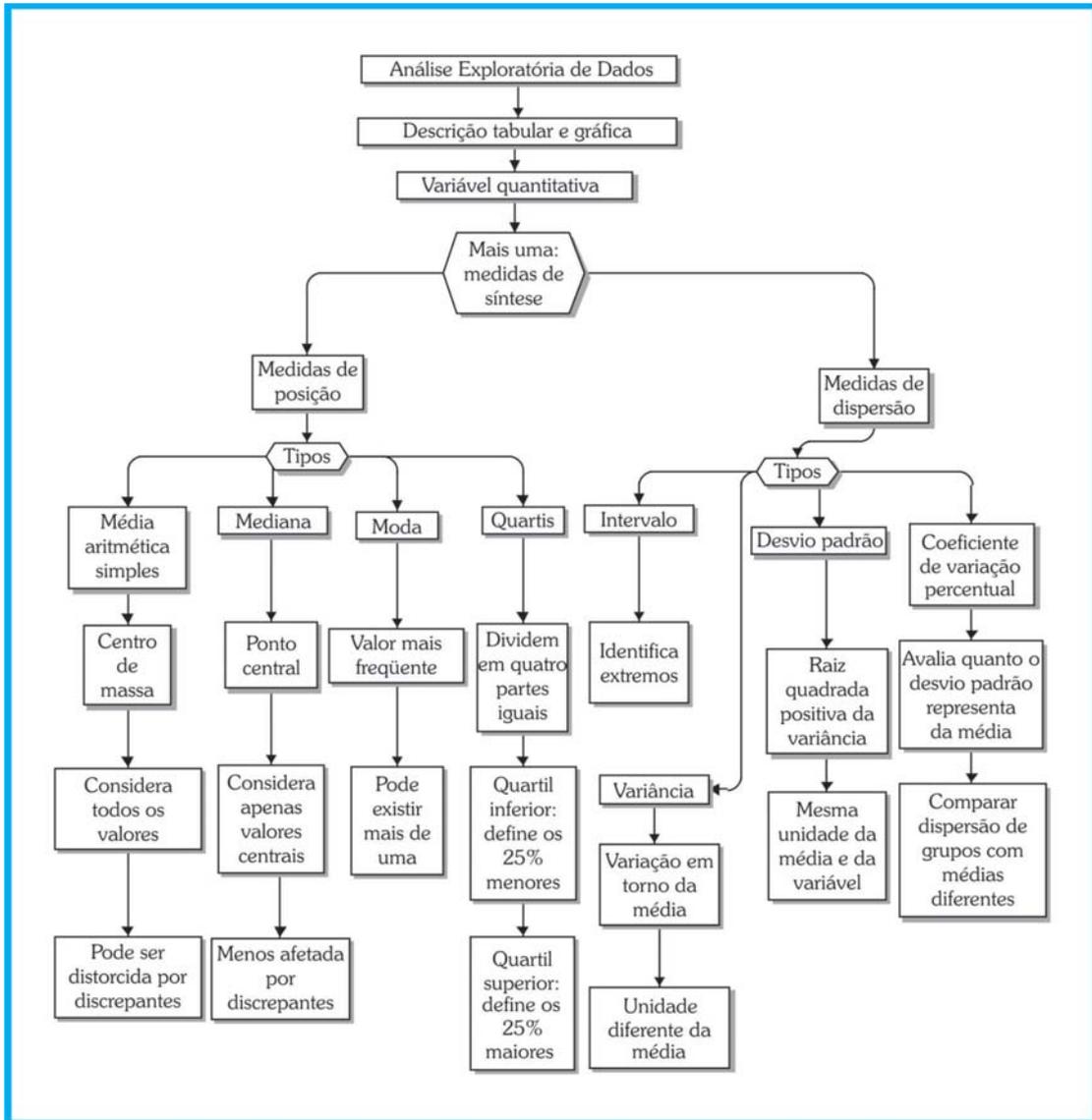


Figura 29: Resumo da Unidade 4.

Fonte: elaborado pelo autor.

Com este tópico finalizamos Análise Exploratória de Dados. É extremamente importante que você faça todos os exercícios, entre em contato com a tutoria para tirar dúvidas, pois não há outra forma de aprender a não ser praticando. Na Unidade 5, veremos os conceitos de Probabilidade, que são indispensáveis para compreender o processo de inferência (generalização) estatística. Vamos em frente e ótimos estudos!!!



Atividades de aprendizagem

As atividades devem ser feitas usando o Microsoft Excel® ou o Br.Office Calc®, através do arquivo AmostraToyord.xls que está no Ambiente Virtual de Ensino-Aprendizagem.

- 1) A variável anos de remodelação dos veículos (na percepção do cliente) está representada na distribuição de frequências expressa no quadro a seguir:

ANOS DE REMODELAÇÃO	FREQUÊNCIAS
0	2
1	57
2	123
3	59
4	9
Total	250

Fonte: elaborado pelo autor.

- a) Calcule a média, mediana, moda e quartis da variável anos de remodelação.
- b) A direção da Toyord acredita que se uma parcela considerável dos clientes perceber que seus modelos são atualizados (foram remodelados há no máximo 2 anos) o

design e o marketing dos veículos estão coerentes. Com base nos resultados da letra a, os dados mostram isso? Justifique.

c) Calcule o intervalo, desvio padrão e coeficiente de variação percentual da variável anos de remodelação.

d) Com base nos resultados dos itens a e c, você considera que os dados estão fortemente concentrados em torno da média? Justifique.

2) Na questão 5 das atividades de aprendizagem da Unidade 3 foi dito:

“os executivos da Toyord creem que seus clientes mais abastados são mais críticos, tendem a ser mais insatisfeitos com seus veículos”. Naquela questão foi construída uma distribuição de frequências conjunta, relacionando a renda agrupada em classes com a opinião geral dos clientes sobre seus veículos, para verificar se os executivos estavam certos. Agora, analise a renda dos clientes (variável quantitativa) em função da opinião geral dos clientes (através do Microsoft Excel® ou do Br.Office Calc®), calculando medidas de síntese de renda em função das opiniões.

a) Com base nos resultados os executivos estão certos? Justifique.

b) Compare com as conclusões que você obteve na questão 9 da Unidade 3.