MEDIDAS DE POSIÇÃO E DISPERSÃO

Flávia Landim – flavia@im.ufrj.br (IM/UFRJ)
J. Ezequiel Soto – cheque@impa.br (IMPA – doutorando em Computação Gráfica)
Nei Rocha – rocha@im.ufrj.br (IM/UFRJ)
Vanessa Matos Leal – vanesamatosleall@gmail.com (SME Angra dos reis e Mesquita – RJ)
Alexandre S. Silva – alexandre.silva@uniriotec.br (UNIRIO)
Projeto Livro Aberto de Matemática – www.umlivroaberto.com

1) INTRODUÇÃO

O quê? Medidas de posição: média, mediana, moda e quartis. Medidas de dispersão: amplitude amostral, distância entre quartis, desvio médio absoluto, variância, desvio padrão e coeficiente de variação. Construção do boxplot (gráfico caixa).

Por quê? As medidas resumo (posição e dispersão) correspondem a uma síntese do conjunto de dados observados e ao passo preliminar para fazer uma inferência estatística, ou seja, a partir das informações obtidas na amostra, expandir nossas conclusões para a população. Como as distribuições podem apresentar formas variadas é importante conhecer diferentes tipos de medidas resumo, para usar medidas apropriadas em cada caso.

Dado um conjunto de dados quantitativos, buscaremos responder as seguintes questões.

(a) É possível encontrar valor(es) para resumir as observações? Qual(is) seria(m) este(s) valor(es)? Como encontrá-lo(s)?

(b) Como medir se os dados estão "próximos" ou "afastados" uns dos outros?

(c) Como você classifica a forma do gráfico construído para representar os dados?

(d) Existe algum valor muito diferente dos demais? Como identificá-lo?

2) MEDIDAS DE POSIÇÃO

Medidas de posição, como o próprio termo indica, visam a resumir um conjunto de dados em geral numa única medida em algum lugar geométrico entre os extremos observados do conjunto (mínimo e máximo). Na figura 1 apresentam-se as marcações da média e da mediana de uma distribuição de notas representada por um histograma.

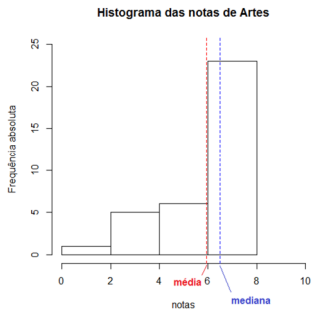

Figura 1: Histograma de uma distribuição de notas com destaque para a média e a mediana

Só é possível obter medidas como a média e a mediana, se nossas observações são de natureza quantitativa, pois, como já vimos, as variáveis qualitativas estão no domínio da frequência apenas, ou seja, só podemos contar quantas observações ocorrem em cada categoria da variável qualitativa, mas não podemos operar matematicamente com as categorias em si.

As principais medidas de posição usadas na Estatística são a média, a mediana, a moda e os quartis da distribuição. Outras medidas de posição existem, mas não são tão usuais. Para definir várias medidas a serem estudadas neste capítulo vamos adotar a notação descrita no exemplo 1.

Exemplo 1: Idade de pessoas que tomaram a vacina da febre amarela

Suponha que na primeira segunda-feira do mês de março de 2018, um Posto de Saúde tenha registrado as idades (em anos completos) das seis primeiras pessoas que chegaram para tomar a vacina da febre amarela e, os registros, obtidos foram 55, 22, 30, 14, 25 e 40, nessa ordem. O número de observações, denotado por *n*, é 6 e observações são dadas por $x_1 = 55, x_2 = 22, x_3 = 30, x_4 = 14, x_5 = 25 e x_6 = 40$.

De um modo geral, sejam $x_1, x_2, \cdots, x_n$ os n valores observados de uma variável quantitativa tal que x_1 é o primeiro valor observado; x_2 é o segundo valor observado; e, assim por diante, tal que x_n é o último valor observado. Os valores observados não ocorrem necessariamente de forma ordenada do menor para o maior. Neste exemplo, x_1=55, x_2=22 e $x_3 = 30$ de modo que $x_1 > x_2$ e $x_2 < x_3$. Para definir a mediana, será útil usar uma notação para representar os dados ordenados.

Defina $x_{(1)}$ o menor valor do conjunto $\{x_1, x_2, x_3, \cdots, x_n\}$; $x_{(2)}$, o segundo menor valor do conjunto $\{x_1, x_2, x_3, \cdots, x_n\}$;e assim sucessivamente até $x_{(n)}$, o maior valor do conjunto $\{x_1, x_2, x_3, \cdots, x_n\}$. Desse modo, $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)}$ são os valores ordenados do conjunto $\{x_1, x_2, x_3, \cdots, x_n\}$.

As idades das seis primeiras pessoas que chegaram para tomar a vacina no Posto de Saúde foram 55, 22, 30, 14, 25, 40 tal que $x_1 = 55, x_2 = 22, x_3 = 30, x_4 = 14, x_5 = 25 e x_6 = 40$. Já os valores ordenados são $x_{(1)} = 14$, $x_{(2)} = 22$, $x_{(3)}$=25, $x_{(4)}$=30, $x_{(5)} = 40$ e $x_{(6)} = 55$.

A letra maiúscula sigma (Σ) é usada para denotar somatório, simplificando algumas fórmulas. Por exemplo,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n \quad \text{e} \quad \sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2$$

Observe que no exemplo das idades das seis primeiras pessoas que chegaram para tomar a vacina no Posto de Saúde, $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 186$ e

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2 = 6.830$$

2.1) MÉDIA

A definição de média de um conjunto de dados quantitativos já é conhecida desde o Ensino Fundamental e, consiste na soma dos valores do conjunto dividida pelo número de observações. No exemplo 1, das idades das seis primeiras pessoas que chegaram para tomar a vacina no Posto de Saúde, a soma das idades é 186 tal que a média será dada por $\frac{186}{6} = 31$ anos.

De modo mais geral, considere um conjunto contendo *n* valores de uma variável quantitativa representado por $\{x_1, x_2, x_3, \cdots, x_n\}$,. Então a média deste conjunto, denotada por $\acute{x}$, é definida por

$$\acute{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Observe que a média pode substituir todas as observações sem alterar a soma dos valores, isto é, $x_1 + x_2 + x_3 + \cdots + x_n = \acute{x} + \acute{x} + \acute{x} + \cdots + \acute{x} = n \cdot \acute{x}$, fornecendo a expressão que define a média .

Esta é justamente a ideia por trás da definição de qualquer média: uma medida que de alguma forma representa o conjunto de dados, segundo uma formulação, e se situa entre os extremos das observações. É claro que, em geral, haverá valores diferentes no conjunto e, neste caso, a média será um valor pertencente ao intervalo de variação dos valores neste conjunto e não necessariamente, um valor que tenha sido observado. No exemplo das idades das seis primeiras pessoas que chegaram para tomar a vacina no Posto de Saúde a média é 31 anos, porém não se observou uma idade igual a 31 anos.

2.2) MÉDIA PARA DADOS AGRUPADOS

Quando os dados disponíveis estão agrupados em intervalos de classe, não é possível calcular a soma total exata dos dados. Neste caso, usamos uma aproximação para o cálculo da média como mostra o exemplo a seguir.

Exemplo 2: Notas de artes

Suponha que um coordenador tenha tido acesso apenas à tabela a seguir, das notas de Artes obtidas por 35 alunos de uma turma. Como este coordenador poderia calcular a média da turma?

Tabela 1: Notas de Artes agrupadas

Intervalo	Frequência absoluta	Ponto médio do intervalo
[0 ; 2[	1	1
[2 ; 4[	5	3
[4; 6[	6	5
[6 ; 8[	23	7
Total	25	--------

Apenas sabemos que, por exemplo, entre 2 e 4 existem cinco notas, mas não conhecemos o valor de cada uma destas cinco notas. Portanto, a soma exata destas cinco notas não é conhecida. A estratégia é tomar o ponto médio desta classe $\left(\frac{2+4}{2}\right) = 3$ como a nota representativa das cinco observações, pois espera-se que os erros cometidos para mais e para menos sejam compensados na classe. Desse modo estimamos a soma das notas neste intervalo como 3+3+3+3+3=5.3=15. Esse procedimento é adotado para todas as classes para obter uma estimativa da soma total dos dados, dada por $1 \cdot 1 + 5 \cdot 3 + 6 \cdot 5 + 23 \cdot 7 = 207$.

Assim, a média correspondente a este agrupamento, a ser considerada pelo coordenador, é estimada por $\acute{x} = \frac{207}{35} \cong 5{,}91$.

A soma exata das 35 notas é 207,5 de modo que o agrupamento das notas resultou numa soma muito próxima da soma exata. Por esta razão dizemos que o agrupamento não incorreu em grande perda de informação para efeito de calcular a soma dos dados: em vez de usar as 35 notas, foi possível com quatro intervalos de classe avaliar de forma precisa a soma original dos dados. Consequentemente, a média estimada por este agrupamento (5,91) não se diferencia muito da média considerando os dados brutos (dados não agrupados) (5,93).

Considere um conjunto de *n* dados agrupados em *c* intervalos de classe. Defina ɤ_i, o ponto médio do *i*-ésimo intervalo de classe. Defina n_i a frequência absoluta do *i*-ésimo intervalo de classe, $i = 1{,}2{,}3{,}\cdots, c$. Neste caso a média é calculada por

$$\acute{x} = \frac{1}{n} \cdot \sum_{i=1}^{c} n_i \cdot \text{ɤ}_i = \frac{n_1 \cdot \text{ɤ}_1 + n_2 \cdot \text{ɤ}_2 + \cdots + n_c \cdot \text{ɤ}_c}{n}$$

Denotando por $f_i = \frac{n_i}{n}$ a frequência relativa do *i*-ésimo intervalo classe, tem-se

$$\acute{x} = \sum_{i=1}^{c} f_i \cdot \text{ɤ}_i = f_1 \cdot \text{ɤ}_1 + f_2 \cdot \text{ɤ}_2 + \cdots + f_c \cdot \text{ɤ}_c$$

Exemplo 3 (A): Promoção de cartão de crédito de supermercado

Numa tarde, 10 clientes interessados em obter um cartão de crédito oferecido por uma rede de supermercados informaram a uma atendente seus salários (em salários mínimos): `, 1, 2, 3, 4, 5, 5, 6, 9 e 10. A média destes dados é, então, $\acute{x} = \frac{1+2+3+4+5+5+6+9+10}{10} = \frac{46}{10} = 4{,}6$. Esse valor de média representa bem este conjunto, pois nele existem cinco valores acima da média e cinco valores abaixo da média e, estes valores não estão muito afastados do valor da média, conforme ilustrado na figura 2.

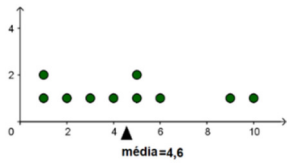

Figura 2: Diagrama de pontos dos salários com destaque para a média do conjunto

Suponha uma pequena variação do conjunto de dez salários na qual o salário igual a 10 salários mínimos foi substituído por um igual a 100 salários mínimos. Assim, os dados agora são: 1, 1, 2, 3, 4, 5, 5, 6, 9 e 100. Há apenas uma diferença entre os dois conjuntos no valor máximo: no primeiro é 10 e no segundo é 100. O que esta única diferença nos dois conjuntos acarreta na média?

Com os dados do segundo conjunto, a média é dada por $\acute{x} = \frac{136}{10} = 13{,}6$, valor maior do que a maioria dos dados observados no conjunto, a saber, apenas uma observação é bem superior a 13,6. Observe, que para representar o diagrama de pontos destes dados (figura 3), usou-se um recurso de quebra do eixo dos dados devido ao valor atípico 100, em relação aos demais valores.

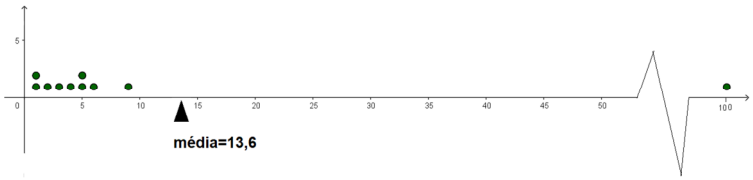

Figura 3: Diagrama de pontos do conjunto {1, 1, 2, 3, 4, 5, 5, 6, 9, 100} com destaque para a média do conjunto e quebra do eixo devido ao valor atípico

Este exemplo simples mostra que na presença de dados atipicamente altos, deve-se tomar cuidado em escolher a média como medida de posição das observações coletadas. Uma medida pouco afetada para valores atípicos, conhecida como medida robusta deve ser considerada em situações deste tipo. A mediana, que trataremos a seguir, é considerada uma medida robusta.

1.3) MEDIANA

A mediana de um conjunto de valores numéricos é definida como o valor que ocupa a posição central dos dados ordenados.

$$mediana = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, sen\acute{e}impar \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, sen\acute{e}par \end{cases}$$

Exemplo 3 (B): Promoção de cartão de crédito de supermercado

Considerando o segundo conjunto de 10 salários dos clientes do supermercado interessados em obter o cartão de crédito (em salários mínimos) tem-se que a mediana é dada por

$$mediana = \frac{x_{(5)}+x_{(6)}}{2} = \frac{4+5}{2} = 4{,}5$$

A média destes dados resultou em 13,6. Este exemplo ilustra a propriedade de que a mediana é pouco afetada na presença de valores atipicamente grandes (ou pequenos). Já a média não possui esta propriedade, sendo muito afetada na presença de valores atípicos.

1.4) MEDIANA PARA DADOS AGRUPADOS

Para obter uma aproximação da mediana quando os dados estão agrupados, deve-se primeiro determinar as frequências acumuladas (absoluta ou relativa) associadas a cada intervalo. Se as frequências forem absolutas, deve-se identificar em qual intervalo encontra-se a observação na posição central $(n+1)/2$ se n for ímpar, ou as duas posições centrais ($n/2$ e $n/2$ +1) se n for par. Depois, tome como uma aproximação para a mediana o ponto médio do intervalo de classe que compreende a(s) posição(ões) central(is).

Existem outras formas de avaliar a mediana quando os dados estão agrupados e uma delas, proposta no exercício 17 do capítulo "A Natureza da Estatística" do Livro Aberto de Matemática para o Ensino Médio, corresponde a encontrar, o valor da observação para o qual a área à sua esquerda e a área a sua direita no histograma sejam iguais.

1.5) MODA

A moda é a observação mais frequente de um conjunto de dados.

Caso não haja observação mais frequente, ou seja, todos os valores aparecem apenas uma única vez no conjunto de dados, a distribuição é dita amodal. Um conjunto é dito unimodal se houver apenas uma moda; bimodal se houver duas modas; ou multimodal se houver três ou mais modas no conjunto de dados coletados.

Exemplo 4: Classificação da distribuição quanto à presença de moda(s)

Considere os conjuntos de notas da prova de Matemática dos alunos de quatro turmas diferentes dadas pela tabela a seguir.

Tabela 2: Exemplos de conjuntos de notas e classificação quanto à presença de moda(s)

Turma	Notas	Moda	Classificação da distribuição quanto à presença de moda(s)
I	2;4;7;8;9;10	Não há	amodal
II	2;4;5;5;5;8;9	5	unimodal
III	2;4;4;5;6;8;8;9	4 e 8	bimodal
IV	1;2;2;2;3;4;4;4;5;7;8;8;8;10	2 4 e 8	multimodal

O conceito de moda é adequado para conjuntos de dados qualitativos ou quantitativos discretos, pois quando os dados são quantitativos contínuos, potencialmente todas as observações são distintas entre si tal que raramente existirá um valor mais frequente e, mesmo quando um valor se repetir, não necessariamente é por que ele corresponderá a uma moda. Neste último caso, o que fazemos é, agrupar os dados em intervalos de classe para identificar um intervalo de classe modal ou intervalos de classe modais, isto é, o(s) intervalo(s) de classe com maior frequência. Uma vez identificado(s) o(s) intervalo(s) de classe modal(ais), uma estimativa para a(s) moda(s) é dada pelo ponto médio do intervalo de classe modal correspondente.

Quando a moda será preferível à média ou à mediana como medida para representar a distribuição? Se o histograma da distribuição é aproximadamente simétrico, e há uma única moda, então as três medidas resumo (média, mediana e moda) serão valores aproximadamente iguais. Nesse caso, em geral, preferiremos

usar a média como medida de posição, pois ela possui propriedades relevantes para a inferência estatística. No entanto, se a distribuição for simétrica e bimodal, tanto a média como a mediana estarão localizadas em intervalo de baixa frequência. Veja na figura 4 uma ilustração desse caso em que as duas modas são mais adequadas para representar a distribuição.

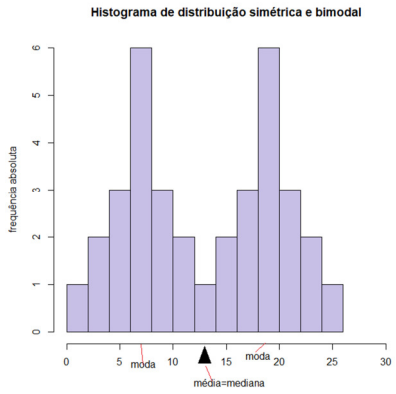

Figura 4: Histograma de uma distribuição simétrica bimodal

1.6) QUARTIS

Os quartis são os três valores que dividem a distribuição em quatro partes de frequências iguais.

Primeiro quartil (Q1): valor da distribuição para o qual a frequência relativa de valores abaixo dele é igual 25% do número de observações do conjunto de dados e, consequentemente, acima dele, é 75% do número de observações do conjunto de dados.

Segundo quartil (Q2): equivalente à mediana, é o valor da distribuição para o qual que a frequência relativa de valores abaixo dele é 50% do número de observações do conjunto de dados e, consequentemente, acima dele, é 50% do número de observações do conjunto de dados.

Terceiro quartil (Q3): valor da distribuição para o qual a frequência relativa de valores abaixo dele é igual 75% do número de observações do conjunto de dados e, consequentemente, acima dele, é 25% do número de observações do conjunto de dados.

Observe que o conhecimento dos valores extremos (mínimo e máximo) e dos quartis (Q1, Q2 e Q3) dispõe-se de um agrupamento dos dados em quatro intervalos de comprimentos (possivelmente) desiguais, a saber,

[mínimo, Q1[; [Q1 , mediana[; [mediana, Q3[e [Q3, máximo], porém todos eles com frequências relativas iguais a $\frac{1}{4} = 0{,}25$.

Um método simples para obter os demais quartis, Q1 e Q3, é considerar dois novos conjuntos de dados, o primeiro, consistindo da primeira metade dos valores ordenados e, o segundo, consistindo da segunda metade. Depois, tome como primeiro

quartil a mediana da primeira metade do conjunto e, como terceiro quartil, a mediana da segunda metade do conjunto de dados.

Uma outra possibilidade é considerar a observação na posição $(n+1)/4$ para o primeiro quartil e a observação da posição (3n+1)/4 para o terceiro quartil. Se o resultado de $(n+1)/4$ ou $(3n+1)/4$ não for um número inteiro, arredonde para o inteiro mais próximo ou, caso a parte decimal seja 0,5, tome a média das duas posições correspondentes à posição. Por exemplo, suponha $n=21$ tal que (21+1)/4=5,5. Assim, neste caso, para obter o primeiro quartil, calcule a média dos valores nas posições 5 e 6.

1.7) BOXPLOT SEM EXIBIÇÃO DE VALORES DISCREPANTES

O boxplot é um esquema para representar dados quantitativos, alternativo ao histograma. Para construir o boxplot é necessário conhecer o esquema dos cinco números, a saber, mínimo, Q1, Q2, Q3 e máximo. O boxplot é composto por um retângulo cortado por um segmento e das bases desses retângulos partem segmentos paralelos ao eixo dos valores da variável.

As bases do retângulo correspondem aos quartis Q1 e Q3 e, o segmento que corta o retângulo corresponde ao segundo quartil (Q2 ou mediana). Da base correspondente ao primeiro quartil (Q1) traça-se um segmento até o valor mínimo observado e, da base correspondente ao terceiro quartil (Q3), traça-se um segmento até o valor máximo observado. Essas são as regras da construção do boxplot simplificado, sem a marcação de valores atípicos (discrepantes). Veremos adiante como marcar valores discrepantes no boxplot.

Por exemplo, o esquema dos cinco números das notas obtidas pelos alunos é {min=0,8; Q1=5,4; Q2=6,8; Q3=7,4; max=8,0} e o boxplot simplificado correspondente ilustrado na figura 5.

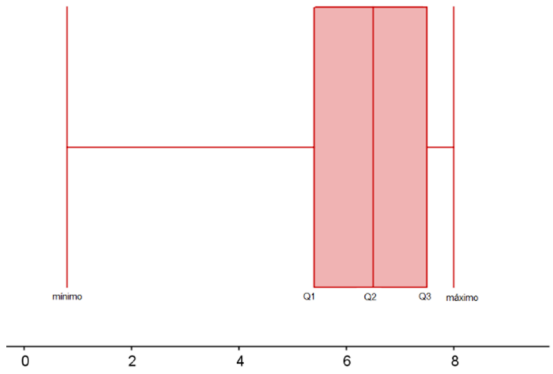

Figura 5: Boxplot sem a exibição de valores discrepantes das notas de Artes

2. MEDIDAS DE DISPERSÃO

Enquanto as medidas de posição procuram resumir o conjunto de dados em alguns valores situados entre dados coletados, as medidas de dispersão buscam avaliar quão dispersos são os dados coletados.

Há uma piada que conta que o Estatístico é o profissional que diz que uma pessoa, ao se sentar numa cadeira com duas placas de metal, uma aquecida a 100°C e outra resfriada a -40 °C, estará em média confortável, pois temperatura média é de 30°C. Na verdade, um Estatístico jamais diria isso, pois ele não toma decisões apenas por uma medida de posição, mas leva em conta também a dispersão dos dados em torno de uma medida de posição. Uma cadeira com duas placas de metal, uma aquecida a 35°C e outra a 25°C, também tem temperatura média de 30°C, mas há menos dispersão da temperatura nessa cadeira que na outra. Assim, embora quantitativamente iguais, os dois valores de 30°C não são qualitativamente equivalentes. Há, portanto, que se avaliar a dispersão dos dados coletados, a fim de poder obter conclusões adequadas.

2.1) AMPLITUDE AMOSTRAL (*R*)

Entre as medidas de dispersão mais simples, define-se a amplitude amostral (*R*) como a diferença entre o maior valor e menor valor observados. Usando a notação apresentada anteriormente, dado um conjunto com n observações, tem-se

$$R = máximo - mínimo = x_{(n)} - x_{(1)}.$$

Uma desvantagem desta medida é que ela considera apenas os dois extremos do conjunto. Ainda é possível que dois conjuntos, tendo mesmas média, moda e mediana, apresentem a mesma amplitude e, no entanto, eles tenham comportamentos diferentes.

2.2) DISTÂNCIA ENTRE QUARTIS (DQ)

Uma medida de dispersão, um pouco mais refinada do que a amplitude amostral, é dada pela distância entre quartis (DQ), definida pela diferença entre os terceiro e primeiro quartis, DQ=Q3-Q1. A distância entre quartis também apresenta a desvantagem de somente levar em conta o primeiro e terceiro quartis, não considerando todas as observações do conjunto.

A distância entre quartis é a medida de dispersão considerada na construção do boxplot para a classificação de valores discrepantes. Defina

cerca inferior=Q1-1,5.DQ e **cerca superior**=Q3+1,5.DQ

Qualquer valor observado menor do que cerca inferior ou maior do que a cerca superior é considerado um valor discrepante. Tais valores, se existirem, são assinalados de forma destacada no boxplot com um símbolo. Para fechar a construção do boxplot com a exibição de valores discrepantes, os segmentos que partem dos primeiro e terceiro quartil são prolongados até o menor valor observado não discrepante e até o maior valor observado não discrepante. Veja, na figura 6, o boxplot construído no final da seção 2, exibindo valores discrepantes.

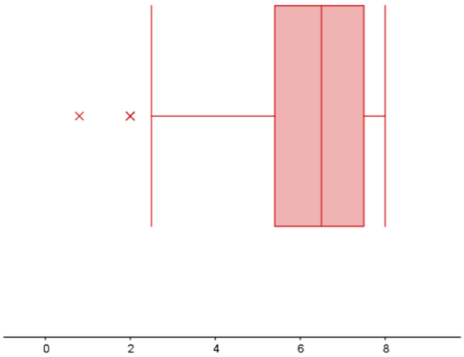

Figura 6: Boxplot das notas de Artes com exibição de valores discrepantes

Podemos observar nesse caso que há duas notas consideradas discrepantes, pois são menores do que a cerca inferior e a menor nota observada não discrepante é um pouco maior do que 2. Além disso, não há notas altas discrepantes. Para a construção do boxplot com exibição de valores discrepantes, as informações dos dados brutos é necessária para que possamos identificar tais valores.

2.3) DESVIOS DA MÉDIA

Considerando o conjunto $\{x_1, x_2, x_3, \cdots, x_n\}$ com n observações, seja $\acute{x}$ a média deste conjunto. Define-se como um desvio da média, a diferença entre uma observação e a média, a saber,

$$desviodamédia_i = x_i - \acute{x}, i = 1,2, \cdots, n$$

Poderíamos pensar em usar os desvios da média para definir uma medida de dispersão dos dados em relação à média do conjunto, no entanto, a não ser que todos os valores sejam iguais, teremos valores acima da média e valores abaixo da média de tal modo que os desvios da média poderão apresentar sinais positivos ou negativos. A média pode ser interpretada como o centro de massa (ponto de equilíbrio) dos dados e, esta propriedade pode ser descrita da seguinte forma: a soma dos desvios da média de qualquer conjunto de dados é sempre nula.

$\sum_{i=1}^{n}(x_i - \acute{x}) = (x_1 - \acute{x}) + (x_2 - \acute{x}) + \cdots + (x_n - \acute{x}) = \sum_{i=1}^{n} x_i - n \cdot \acute{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$, pois $\acute{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Portanto, a soma dos desvios da média não serve como medida de dispersão de um conjunto de dados, pois ela sempre resultará em zero.

2.4) DESVIO MÉDIO ABSOLUTO

Considerando os desvios da média em valor absoluto ($|x_i - \acute{x}|$), a soma dos desvios da média em valor absoluto ($\sum_{i=1}^{n}|x_i - \acute{x}|$)será maior ou igual a zero, com a igualdade valendo somente se todos os valores no conjunto forem iguais.

Com base na observação anterior, pode-se definir uma medida de dispersão dos dados, considerando todas as observações, chamada desvio médio absoluto (DM) que é definida como a média dos desvios da média tomados em valor absoluto.

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

2.5) VARIÂNCIA E DESVIO PADRÃO

Uma outra forma de eliminar o sinal negativo dos desvios da média é elevar ao quadrado cada um deles, tornando-os não-negativos. A variância é definida como uma média dos desvios da média elevados ao quadrado.

$$variância = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \acute{x})^2$$

Quando lidamos com grande quantidade de dados, calcular a variância usando a definição apresentada será uma tarefa entediante, pois após calcular a média de muitos dados, teremos que calcular cada desvio da média, elevá-los ao quadrado e, finalmente, somá-los. Um modo mais simples para calcular a variância é apresentado a seguir.

$$\sum_{i=1}^{n} (x_i - \acute{x})^2 = \sum_{i=1}^{n} \left(x_i^2 - 2 \cdot \acute{x} \cdot x_i + \acute{x}^2\right) = \sum_{i=1}^{n} x_i^2 - 2 \cdot \acute{x} \cdot \sum_{i=1}^{n} x_i + n \cdot \acute{x}^2$$

$$\sum_{i=1}^{n} x_i^2 - \frac{2}{n} \cdot \left(\sum_{i=1}^{n} x_i\right)^2 + \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

Assim, basta conhecer a soma simples $(\sum_i^n x_i)$ e a soma de quadrados $\left(\sum_{i=1}^n x_i^2\right)$para calcular a variância.

Quando calculamos a variância, a unidade de medida das observações é elevada ao quadrado. De modo a poder comparar diretamente essa medida de dispersão com os dados, usamos o desvio padrão que é a raiz quadrada da variância.

$$desviopadrão = \sqrt{variância}$$

Observação 1: Fórmula de aproximação para o desvio padrão

Uma expressão que estima de modo grosseiro o valor do desvio padrão s é dada por $s \approx \frac{R}{4}$, em que R é a amplitude amostral.

Uma explicação para essa aproximação pode ser dada, argumentando-se que em conjuntos de dados cuja distribuição é aproximadamente simétrica e não apresenta valores discrepantes tem-se

valor mínimo $\approx \acute{x} - 2 \cdot s$ **e valor máximo** $\approx \acute{x} + 2 \cdot s$**.**

Tomando a diferença entre os valores máximo e mínimo, obtém-se a expressão apresentada.

Observação 2: Por que o desvio padrão é preferível ao desvio médio?

Você deve estar se perguntando por que se utiliza o desvio padrão na Estatística em detrimento do desvio médio absoluto, cujo cálculo é bem mais simples. A resposta é um tanto complexa para o nível em que estamos, mas ela está associada à necessidade na Estatística de se minimizar estruturas de maneira simples. O desvio médio faz uso da função modular $f(x)=|x|$, que não possui boas propriedades matemáticas para a minimização, por possuir na sua forma uma mudança abrupta em torno de $x=0$, enquanto que a variância faz uso da função quadrática $f(x)=x^2$, representando parábolas de vértice suave e cujas propriedades analíticas são bem conhecidas. Veja a figura 7.

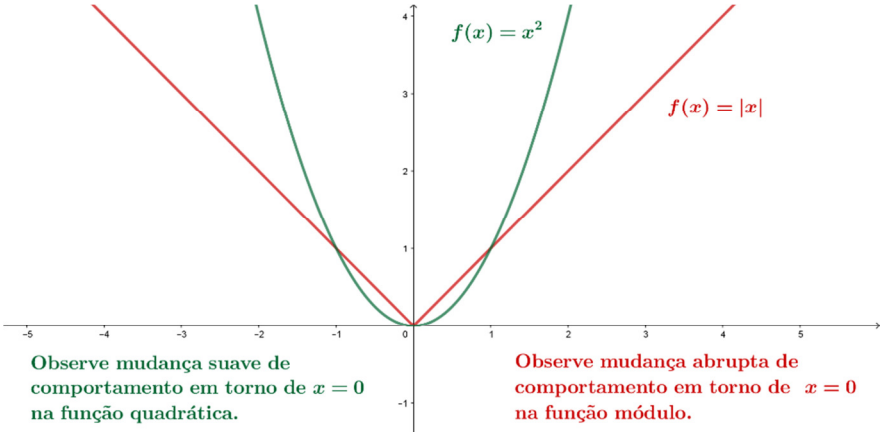

Figura 7: Funções modular e quadrática com destaque para o comportamento em torno de x=0.

Observação 3: Variância populacional e amostral, desvio padrão populacional e amostral

No capítulo *A natureza da Estatística* foram apresentados os conceitos parâmetro e estimador. Parâmetro é uma característica numérica da população, em geral desconhecida; enquanto estimador é uma função dos dados da amostra (subconjunto da população), usada para estimar o parâmetro. Em geral, usam-se letras gregas para denotar parâmetros.

Se dispomos de uma amostra da população, de fato, calculamos a média amostral ($\acute{x}$) e a variância amostral (funções dos dados da amostra) e usamos estes resultados como estimativas da média populacional (em geral denotada pela letra grega μ) e da variância populacional (em geral denotada por σ^2). Como já foi comentado anteriormente, a média amostral ($\acute{x}$) tem boas propriedades como estimador da média populacional (μ). No entanto, é possível mostrar que a variância calculada pela fórmula apresentada aqui é um estimador da variância populacional (σ^2) que tende a produzir valores menores do que o valor da variância da população. Dizemos que é um estimador viesado por essa razão.

Esse problema é resolvido, usando-se n -1 em vez de n no denominador da fórmula da variância.

Por exemplo, na planilha Excel, existem duas funções para calcular a variância, a saber, var.p(dados) para variância populacional e var.a(dados) para variância

amostral. Também existem duas funções para calcular o desvio padrão: desvpad.p(dados) para desvio padrão populacional e desvpad.a(dados) para desvio padrão amostral. No GeoGebra, a função que calcula todas as medidas resumo, retorna as seguintes informações conforme a figura a seguir.

Assim, as expressões que deverão ser usadas quando o conjunto de dados sob estudo é uma amostra da população são dadas por

variância amostral: $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \acute{x})^2$ e desvio padrão amostral: $\sqrt{s^2} = s$

Na maioria das vezes trabalhamos com amostras. Assim, neste capítulo, salvo menção em contrário, estaremos sempre calculando a variância amostral (s^2) e o desvio padrão amostral (s), mesmo que o termo "amostral" esteja omitido.

Se você estiver trabalhando com uma amostra e usar o denominador n para calcular a variância, isso implicará que você escolheu um estimador viesado, pois tende a produzir estimativas que são menores do que o verdadeiro valor da variância. Se você estiver trabalhando com amostras muito grandes, essa diferença não será importante, pois haverá pouca diferença entre dividir por n ou por n-1.

Expressões que deverão ser consideradas quando o conjunto de dados sob estudo refere-se à população com N elementos:

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

desvio padrão populacional: $\sqrt{\sigma^2} = \sigma$,

em que μ representa a média populacional.

Expressões que deverão ser consideradas quando o conjunto de dados sob estudo refere-se a uma amostra de tamanho n da população:

variância amostral: $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \acute{x})^2$

desvio padrão amostral: $\sqrt{s^2} = s$em que $\acute{x}$ é a média amostral.

Observação 4: Cálculo da variância para dados agrupados

Se os dados estão agrupados em c classes, as frequências relativas são dadas por $f_1, f_2, \cdots, f_c$ e os pontos médios das classes são dados por $\acute{x}_1, \acute{x}_2, \cdots, \acute{x}_n$,a variância amostral pode ser calculada por

$$s^2 \approx \sum_{i=1}^{c} f_i \cdot \acute{x}_i^2 - \acute{x}^2$$

2.6) COEFICIENTE DE VARIAÇÃO

Nem sempre uma variância pequena (e consequentemente desvio-padrão pequeno) significa pouca dispersão. Tampouco uma variância grande é sempre indicador de alta dispersão. Esses valores podem ser altos ou baixos devido à magnitude (ordem de grandeza) dos dados observados. Se medimos observações em

microscópio, por exemplo, teremos inevitavelmente valor numericamente baixo de variância, podendo, no entanto, haver alta dispersão dos dados no nível microscópico. Da mesma maneira, ao medir os produtos internos brutos brasileiros em dólares em vários anos teremos valores observados de alta magnitude, gerando variância numericamente grande, mas não necessariamente indicando alta dispersão.

O coeficiente de variação amostral é uma medida usada para calcular a variação relativa dos dados de um conjunto em torno da média: quanto maior seu valor, maior é a variação relativa em torno da média. Em geral, ele é calculado em termos percentuais.

$$CV = \frac{s}{\acute{x} \cdot 100\%}$$

em que s é o desvio padrão amostral e $\acute{x}$ é a média amostral.

Observe que o coeficiente de variação só é definido para conjuntos cuja média é diferente de zero.

3. AVALIAÇÃO DO GRAU DE ASSIMETRIADA DISTRIBUIÇÃO

O boxplot é útil para avaliar a forma da distribuição quanto ao grau de assimetria e também revela valores atípicos, se houver. O retângulo do boxplot corresponde aos 50% valores centrais da distribuição, ou seja, metade dos dados estão no intervalo delimitado pela caixa (retângulo) e, a outra metade, está nos dois intervalos delimitados fora da caixa, sendo 25% acima e 25% abaixo da caixa.

As medidas do esquema dos cinco números nos permitem avaliar o grau de assimetria da distribuição. Por exemplo, se

- mediana - Q1 ≈ Q3 - mediana
- Q1 – mínimo ≈ máximo - Q3
- mediana – mínimo ≈ máximo - mediana

podemos concluir que a distribuição é aproximadamente simétrica, porém se alguns destes pares de intervalos apresentarem comprimentos muito diferentes, isso indica que a distribuição apresenta algum tipo de assimetria.

Uma regra empírica para avaliar frequências de valores em intervalos em torno da média que pode ser útil, é obtida a partir das propriedades de um modelo teórico conhecido como densidade normal de probabilidades. Entre várias propriedades desta densidade, destaca-se que ela é simétrica e unimodal tal que média, mediana e moda são iguais. Uma ilustração da densidade normal com média μ e desvio padrão σ, também conhecida como a curva em forma de sino é apresentada na figura 8.

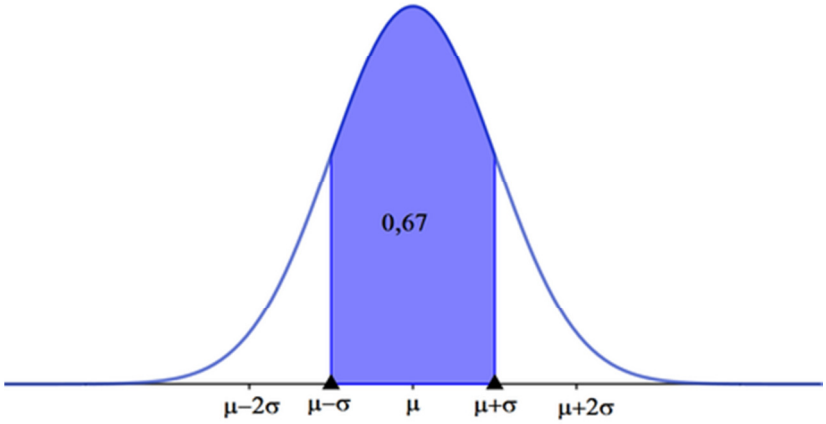

Figura 8: Densidade Normal com região colorida no intervalo entre μ-σ e μ+σ, cuja área corresponde a aproximadamente 0,67 da área total igual a 1

A regra empírica estabelece que em distribuições aproximadamente simétricas para as quais a presença de valores discrepantes é muito rara ou não existem valores discrepantes, a frequência relativa de valores no intervalo

(a) $[\acute{x} - s; \acute{x} + s]$é aproximadamente 67%,
(b) $[\acute{x} - 2 \cdot s; \acute{x} + 2 \cdot s]$ é aproximadamente 95%.