Material Teórico - Módulo de ESTATÍSTICA

Medidas de Dispersão

Primeiro Ano do Ensino Médio

Autor: Prof. Francisco Bruno Holanda
Revisor: Prof. Antonio Caminha Muniz Neto

1 Introdução

Na aula anterior, vimos como as medidas de posição (também conhecidas como *medidas de tendência central*) nos fornecem informações que resumem o conjunto dos dados em valores únicos. As principais medidas apresentadas foram a média, a mediana e a moda. Apesar do fato de que tais medidas propiciam uma visão geral dos dados, fazer uma análise estatística baseando-se apenas em tais medidas pode gerar interpretações enganosas. A fim de ilustrar esse fato, tome como exemplo os dois grupos de dados a seguir:

Primeiro grupo:

$$-1, 0, 0, 0, 1.$$

Segundo grupo:

$$-16, -9, -7, 0, 0, 0, 7, 9, 16.$$

Apesar de ambos possuírem as mesmas média, mediana e moda (todos tais valores são iguais a zero), os dois conjuntos de dados são bem distintos. Enquanto o primeiro possui seus valores mais próximos da média, o segunto tem valores extremos mais distantes. Dessa forma, se quisermos analisar mais profundamente um conjuto de dados, faz-se necessário considerarmos outras medidas que não apenas as de tendência central. Nesse sentido, uma **medida de dispersão** para uma variável quantitativa é um indicador do *grau de espalhamento* dos valores da amostra em torno de uma medida de centralidade. Nesta aula, estudaremos algumas de tais medidas.

2 Amplitude e distância interquartil

A forma mais simples de se medir a variabilidade de um conjunto de dados é calculando sua **amplitude**, que é definida como a diferença entre o maior e o menor valores observados. Por exemplo, considere a série de dados a seguir:

$$5, 9, 16, 4, 9, 5, 2, 6, 1, 7.$$

Como a maior observação é 16 e a menor 1, a amplitude é

$$A = 16 - 1 = 15.$$

Apesar da amplitude ser uma medida bastante simples do ponto de vista teórico, na prática ela não é considerada uma boa medida de dispersão, principalmente por dois motivos. O primeiro deles é devido ao custo operacional de se encontrar os valores máximo e mínimo das observações quando temos um grande número de dados (na ordem das dezenas de milhares, em aplicações típicas). O segundo é em razão dessa medida ser muito sensível a valores distoantes (*outliers*). De fato, considere os conjuntos de dados a seguir:

$$G_1 : 1, 3, 5, 2, 3, 4, 5, 1, 2, 3, 4, 3, 3, 2, 1;$$

$$G_2 : 1, 3, 5, 2, 3, 4, 5, 1, 2, 3, 4, 3, 3, 2, 1, 55.$$

Veja que, apesar dos dois conjuntos serem semelhantes, a amplitude do primeiro é $A_1 = 5 - 1 = 4$, enquanto a do segundo é $A_2 = 55 - 1 = 54$. A grande diferença constatada nesse exemplo é devida à presença de um valor distoante (55) no segundo grupo. Uma forma de dimunir o efeito de valores distoantes sobre a distorção é considerar a **distância interquartil**, conforme descrita a seguir.

Na aula anterior, vimos que a mediana é uma medida de tendência central na qual metade dos dados é menor que ela, ao passo que a outra metade é maior que ela. Analogamente, os três **quartis** do conjunto de dados são os valores reais que o dividem em quatro grupos, cada um deles contendo 1/4 do tamanho total da amostra. Mais precisamente:

- o primeiro quartil Q_1 tem 1/4 dos dados abaixo dele e 3/4 dos dados acima dele. Ou seja, é a mediana da primeira metade dos dados;
- o segundo quartil Q_2 é a mediana;
- o terceiro quartil Q_3 tem 3/4 dos dados abaixo dele e 1/4 dos dados acima dele. Ou seja, é a mediana da segunda metade dos dados.

A **distância interquartil** é definida como $DIQ = Q3 - Q1$.

Exercício 1. *Calcule as distâncias interquartis dos conjuntos G_1 e G_2 apresentados anteriormente nesta seção.*

Solução. Primeiramente, reescrevamos os conjuntos de dados G_1 e G_2 na ordem crescente:

$$G_1 : 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 5;$$

$$G_2 : 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 5, 55.$$

O primeiro grupo tem 15 elementos de forma que a mediana é o elemento central, que é igual a 3. Seu primeiro quartil é a mediana dos $\frac{15-1}{2} = 7$ primeiros termos, de forma que é novamente seu elemento central $Q_1 = 2$. O terceiro quartil é a mediana dos sete últimos termos, de forma que $Q_3 = 4$. Portanto, a distância interquartil para G_1 é $DIQ = 4 - 2 = 2$.

O segundo grupo tem 16 elementos. Logo, a mediana é a média aritmética dos dois elementos centrais, que nesse caso vale 3. O primeiro quartil é a mediana do conjunto formado pelos oito primeiros termos, de sorte que $Q_1 = \frac{2+2}{2} = 2$. O terceiro quartil é a mediana do conjunto de dados formado pelos oito últimos termos, logo, $Q_3 = \frac{4+4}{2} = 4$. Portanto, a distância interquartil para G_2 vale $DIQ = 4 - 2 = 2$. □

No exercício anterior, a distância interquartil é a mesma nos dois grupos, apesar da presença de um valor distoante no segundo conjunto de dados. Este é um exemplo de como a *DIQ* é menos sensível à *outliers* do que a amplitude, e pode-se mostrar que tal fato é típico.

3 Desvio-médio

Outra opção de medida de dispersão é a média aritmética dos valores absolutos dos desvios das várias observações em relação à média dos dados. Em símbolos, se x_1, x_2, ..., x_n são os dados observados e $\overline{x}$ é sua média, o **desvio médio** (dm) da amostra é definido por:

$$dm = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|.$$

Vejamos um exemplo.

Exemplo 2. *Considere um conjunto de dados que representam as idades nas quais um grupo de pacientes começou a apresentar problemas de saúde:*

$$65, 72, 70, 72, 60, 67, 69, 68.$$

Para calcularmos o desvio médio, devemos primeiramente calcular a média $\overline{x}$ do conjunto de dados:

$$\overline{x} = \frac{65+72+70+72+60+67+69+68}{8} = 67,875.$$

Agora, vamos subtrair $\overline{x}$ de cada valor da amostra, calcular o valor absoluto de cada uma dessas diferenças e somá-los. Para tanto, considere a tabela a seguir:

	x_i	$x_i - \overline{x}$	$\lvert x_i - \overline{x}\rvert$
	65	-2,875	2,875
	72	4,125	4,125
	70	2,125	2,125
	72	4,125	4,125
	60	-7,875	7,875
	67	-0,875	0,875
	69	1,125	1,125
	68	0,125	0,125
Soma	543		23,25
Média	67,875		2,90625

Então, o desvio médio é dado por:

$$dm = 2,90625$$

Apesar de ser uma medida fácil de ser computada, já que não envolve a ordenação dos valores, o desvio-médio não é muito usado na prática, uma vez que não é tão fácil de se estabelecer as propriedades matemáticas de uma medida definida com o uso da função valor absoluto. Para resolvermos esse problema e, ainda assim, mantermo-nos próximos a uma noção de *desvio da média*, definiremos a seguir as medidas de dispersão mais populares, quais sejam, a **variância** e o **desvio-padrão**.

4 Variância e desvio-padrão

A **variância** σ^2 de uma população $x_1, \ldots, x_N$ de N dados é a medida de dispersão definida como a média aritmética do quadrado dos desvios dos elementos da população em relação à média $\overline{x}$ da mesma. De outra forma, a variância é dada por:

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

Exemplo 3. *Considere o conjunto de dados x_i, com $1 \leq i \leq 10$, colecionados na primeira coluna da tabela a seguir:*

	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
	3	-2,3	5,29
	0	-5,3	28,09
	1	-4,3	18,49
	9	3,7	13,69
	3	-2,3	5,29
	12	6,7	44,89
	1	-4,3	18,49
	6	0,7	0,49
	4	-1,3	1,69
	14	8,7	75,69
Soma	53		212,1
Média	5,3		21,21

Para calcularmos a variância do conjunto, o primeiro passo é calcular a média do mesmo, dada por $\overline{x} = 5,3$. Em seguida escrevemos uma segunda coluna, na qual listamos os vários desvios da média *$x_i - \overline{x}$. Na terceira coluna, calculamos os quadrados dos desvios, i.e., os números $(x_i - \overline{x})^2$. Por fim, obtemos a variâcia como a média aritmética desses valores, de sorte que*

$$\sigma^2 = 21,21$$

Note que, se as observações forem dadas em uma certa unidade de medida, a média virá expressa nessa mesma unidade. Por outro lado, variância estará expressa em termos dessa unidade ao quadrado. Por exemplo, observações medidas em metros (m) terão variância medida em metros quadrados (m^2).

Para mantermos as mesmas unidades de medida no cálculo de medidas de dispersão, ua estratégia interessante é calcularmos o **desvio padrão** σ da população, que por definição é igual à raiz quadrada da variância da mesma. Dessa forma, o desvio padrão é dado por:

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

Assim, no exemplo 3, o desvio padrão é

$$\sigma = \sqrt{21,21} \cong 4,605.$$

Suponha, agora, que temos um conjunto de dados agrupados em uma tabela de frequência, de forma que as amostras *distintas* do conjunto valham $x_1, \ldots, x_n$, com x_i ocorrendo com frequência absoluta fa_i e frequência relativa fr_i. Então, nosso total de observações é

$$N = fa_1 + fa_2 + \cdots + fa_n,$$

de forma que

$$fr_i = \frac{fa_i}{N}.$$

A média dos dados é calculada por

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{n} fa_i x_i = \sum_{i=1}^{n} fr_i x_i \qquad (1)$$

e a variância por

$$\sigma^2 = \sum_{i=1}^{n} \frac{fa_i(x_i - \overline{x})^2}{N} = \sum_{i=1}^{n} fr_i(x_i - \overline{x})^2. \qquad (2)$$

Para calcular o desvio-padrão basta, obviamente, tirar a raiz quadrada da fórmula acima.

Exemplo 4 (FGV 2002)**.** *Numa pequena ilha, há 100 pessoas que trabalham na única empresa ali existente. Seus salários (em moeda local) têm a seguinte distribuição de frequências:*

Salários	Frequência
\$ 50	30
\$ 100	60
\$ 150	10

a) Qual a média dos salários das 100 pessoas?

b) Qual a variância dos salários? Qual o desvio padrão dos salários?

Solução. Veja que as frequências relativas para os salários de \$50, \$100 e \$150 são, respectivamente, iguais a 30%, 60% e 10%. Assim, utilizando (1), obtemos a média dos salários:

$$\overline{x} = 50 \cdot \frac{30}{100} + 100 \cdot \frac{60}{100} + 150 \cdot \frac{10}{100} = 90$$

Após termos calculado a média, podemos calcular a variância com o auxílio de (2):

$$\begin{aligned}\sigma^2 &= \frac{30}{100}(50-90)^2 + \frac{60}{100}(100-90)^2 + \frac{10}{100}(150-90)^2 \\ &= \frac{30}{100}\cdot 1600 + \frac{60}{100}\cdot 100 + \frac{10}{100}\cdot 3600 \\ &= 480 + 60 + 360 = 900.\end{aligned}$$

Portanto, o desvio-padrão será

$$\sigma = \sqrt{900} = 30.$$

□

Na segunda parte dessa aula, deduziremos algumas comparações entre as medidas de dispersão estudadas aqui, para um conjunto positivo de dados.

O primeiro resultado que desejamos obter diz que o desvio-padrão σ é sempre maior ou igual que o desvio médio dm e menor ou igual que a amplitude, ocorrendo a igualdade se, e só se, todos os dados de nossa amostra coincidirem. Em símbolos,

$$dm \leq \sigma \leq A.$$

O segundo resultado diz que, se o desvio médio não for muito grande, então o desvio padrão não excede $\sqrt{dm^2 + \frac{A^2}{2}}$. Em símbolos,

$$dm \leq \frac{A}{\sqrt{2}} \Rightarrow \sigma < \sqrt{dm^2 + \frac{A^2}{2}}.$$

5 Sugestões ao professor

Separe três encontros de 50 minutos cada para desenvolver o conteúdo desta aula. No primeiro encontro, apresente as definições e exemplos sobre amplitude e distância interquartil. No segundo, apresente as definições de desvio-médio, variância e desvio-padrão. Por se tratarem das medidas de dispersão mais populares, resolva o máximo possível de exercícios sobre essa seção (recorrendo à bibliografia sugerida, se preciso) em uma terceiro encontro.

Referências

[1] Pedro A. Morettin e Wilton de O. Bussab. *Estatística Básica*. Saraiva, 2010.

[2] João Ismael Pinheiro et al. *Estatística Básica: a Arte de Trabalhar com Dados*. Campus, 2009.