Material Teórico - Módulo de ESTATÍSTICA

Tópicos Extras

Primeiro Ano do Ensino Médio

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

Esta aula discute dois tópicos adicionais, relacionados a medidas de dispersão. Primeiramente, utilizaremos as desigualdades estudadas na aula anterior para relacionar algumas das medidas de dispersão estudadas. Depois, veremos uma representação gráfica (o *gráfico boxplot*) para os quartis e medianas de um conjunto de dados.

1 A média quadrática e as medidas de dispersão

Começamos tratando de maneira mais aprofundada algumas propriedades das medidas de dispersão que definimos na terceira aula. Mais especificamente, demonstraremos algumas relações de desigualdades que revelam a existência de uma ordenação bem definida entre as diferentes medidas de dispersão. Tais são dadas por

$$dm \leq \sigma \leq A, \tag{1}$$

em que dm é o desvio médio, σ é o desvio padrão e A é a amplitude de um conjunto de dados **positivos**. É importante destacar que tais desigualdades são válidas apenas quando o conjunto de dados é positivo.

Antes de prosseguirmos, considere o seguinte exemplo em que verificamos as desigualdades (1) diretamente.

Exemplo 1. *Considere os dados a seguir, sobre as alturas (em metros) dos* 16 *alunos de uma determinada turma*[1]

1.65 — 1.61 — 1.46 — 1.66 — 1.49 — 1.48 — 1.70 — 1.54 — 1.55 — 1.65 — 1.69 — 1.65 — 1.66 — 1.54 — 1.68 — 1.68

Cálculos simples fornecem

$$dm = 0.071, \quad \sigma = 0.082 \quad e \quad A = 0.24,$$

valores que refletem as desigualdades (1).

Veremos, agora, como justificar (1). Para tanto, considere um conjunto de dados positivos $x_1, x_2, \ldots, x_n$, e seja $a_i = |x_i - \overline{x}|$, para $1 \leq i \leq n$. Note que a média quadrática dos termos a_i é igual ao desvio-padrão de $x_1, \ldots, x_n$, enquanto a média aritmética dos a_i é igual ao desvio médio de $x_1, \ldots, x_n$. Em símbolos,

$$MQ(a_i) = \sigma, \quad \text{e} \quad MA(a_i) = dm.$$

Dessa forma, a desigualdade entre as médias quadrática e aritmética, vista na aula passada, garante que $dm \leq \sigma$.

Para demonstrar que $\sigma \leq A$, suponha (sem perda de generalidade) que os dados estão ordenados, ou seja, que $x_1 \leq x_2 \leq \ldots \leq x_n$. Dessa forma, x_n e x_1 são, respectivamente, o máximo e o mínimo dos dados observados. Assim, por definição, temos:

$$x_1 \leq x_1 \leq x_n$$

$$x_1 \leq x_2 \leq x_n$$

$$\vdots$$

$$x_1 \leq x_n \leq x_n$$

Somando membro a membro tais desigualdades, e dividindo por n em seguida, podemos verificar que

$$x_1 \leq \overline{x} \leq x_n.$$

Assim, a média de um conjunto de observações deve estar entre o o máximo e o mínimo dos dados observados.

Agora, veja que a amplitude A é dada por

$$A = x_n - x_1 = \max\{|x_i - x_j|;\ 1 \leq i, j \leq n\}.$$

Uma vez que a média quadrática de um conjunto finito de números não negativos é menor ou igual que o maior deles, obtemos

$$\sigma(x_i) = MQ(a_i) \leq M,$$

onde $M = \max\{a_1, \ldots, a_n\}$.

Por outro lado, como $A = x_n - x_1$ e $x_1 \leq \overline{x} \leq x_n$, devemos ter $M \leq A$. Portanto,

$$\sigma \leq A.$$

Assim, concluímos a demonstração das desigualdades (1).

Finalizamos esta aula apresentando um resultado que fornece uma cota superior para σ melhor do que A.

Teorema 2. *Se $x_1, ..., x_n$ são dados positivos, então:*

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

Prova. Sejam $a_i = |x_i - \overline{x}|$ os desvios médios das observações. Pelo primeiro corolário do Teorema 3 da aula anterior, temos que

$$\sigma^2 < dm^2 + \frac{1}{2}\Delta^2, \tag{2}$$

onde $\Delta = \max\{|a_i - a_j|;\ 1 \leq i, j \leq n\}$.

Agora, fazendo $u = x_i - \overline{x}$ e $v = x_j - \overline{x}$ e utilizando o terceiro corolário da desigualdade triangular (veja uma vez mais a aula anterior), obtemos

$$|a_i - a_j| = ||x_i - \overline{x}| - |x_j - \overline{x}|| \leq |x_i - x_j| \leq A,$$

[1] Observe que, na tabela a seguir, utilizamos a notação inglesa para números decimais, com pontos em vez de vírgulas. Essa é meramente uma questão de conveniência, tendo sido motivada pela rotina de programação apresentada no final desta aula.

para todos $1 \leq i, j \leq n$. Então,

$$\Delta = \max\{|a_i - a_j|;\, 1 \leq i, j \leq n\} \leq A. \qquad (3)$$

Por fim, segue de (2) e (3) que

$$\sigma^2 < dm^2 + \frac{1}{2}A^2.$$

Para finalizar, basta calcular a raiz quadrada em ambos os membros da última desigualdade acima. □

Exemplo 3. *Considere os dados a seguir, obtidos a partir da produção (em toneladas) de café em nove fazendas:*

4.28 — 6.53 — 7.46 — 4.19 — 3.70 — 2.84 — 4.67 — 4.42 — 5.51

Cálculos fáceis fornecem

$$dm = 1.103, \quad \sigma = 1.432 \quad e \quad A = 4.62,$$

de sorte que

$$1.432 < \sqrt{(1.103)^2 + \frac{(4.62)^2}{2}} = \sqrt{1.216 + 10.672} = 3.447.$$

2 Gráficos boxplot

Um **gráfico boxplot** (ou um *diagrama de caixas*) é uma representação gráfica de um conjunto de dados, a qual nos permite avaliar rapidamente a dispersão dos mesmos, destacando também os valores discrepantes presentes.

Para construir o gráfico boxplot correspondente a um certo conjunto de dados, necessitamos compilar algumas informações específicas sobre os quartis Q_1, Q_2 e Q_3 das observações. (A partir desse ponto, sugerimos ao leitor acompanhar a leitura observando o gráfico boxplot do primeiro exemplo a seguir, a fim de melhor compreender a descrição de sua construção.)

Mais precisamente, para compor o diagrama boxplot, inicialmente desenhamos uma caixa retangular maior, na qual a base é vista como um segmento da reta real, orientada da esquerda para a direita. Em seguida, desenhamos o diagrama boxplot em si, o qual consiste em um par de segmentos verticais, conectados por hastes horizontais a uma caixa retângular menor situada entre ambas, e no interior da qual um terceiro segmento vertical é traçado; eventualmente (e conforme descrito mais adiante), o diagrama conterá alguns pontos adicionais.

Os lados verticais da caixa menor têm abscissas iguais aos quartis Q_1 e Q_3 do conjunto de dados, ao passo que o segmento vertical interior a ela tem abscissa igual à mediana do conjunto de dados.

As hastes horizontais que partem da caixa menor são segmentos de reta horizontais, cujas abscissas variam:

(i) à esquerda, do quartil inferior Q_1 até a menor observação que seja maior ou igual ao *limite inferior* dos dados (definido a seguir).

(ii) à direita, do quartil superior Q_3 até a maior observação que seja menor ou igual ao *limite superior* dos dados (também definido a seguir).

Os segmentos verticais situados à esquerda e à direita da caixa menor têm abscissas respectivamente iguais àquelas da extremidade esquerda da haste mais à esquerda e da extremidade direita da haste mais à direita da caixa menor.

O **limite inferior** e o **limite superior** do conjunto de dados são calculados por:

$$\textbf{Lim. inferior} = \max\{\min(\text{dados}); Q_1 - 1,5(Q_3 - Q_1)\}.$$

$$\textbf{Lim. superior} = \min\{\max(\text{dados}); Q_3 + 1,5(Q_3 - Q_1)\}.$$

Aqui, min(dados) e max(dados) representam, respectivamente, o menor valor e maior valor encontrado nas observações.

Por fim, as observações situadas fora do intervalo formado pelas abscissas dos segmentos verticais à esquerda e à direita da caixa menor são considerados *valores discrepantes* (*outliers*), sendo denotados no boxplot por pontos (•), marcados nas abscissas adequadas.

Para melhor entender a construção de diagramas boxplot, temos os dois exemplos a seguir.

Exemplo 4. *Considere as observações a seguir, retiradas de uma amostra dos valores pagos pelos clientes de um restaurante a quilo no almoço.*

11	23	50	19
49	50	15	41
49	16	17	22
15	25	50	42
36	11	20	33
29	44	150	30
29	48	40	28
26	37	15	42
19	13	24	22
22	25	25	16

Para construir o gráfico boxplot correspondente, precisamos calcular os seguintes valores constantes da primeira coluna da tabela a seguir:

Primeiro Quartil	19
Mediana	25,5
Terceiro Quartil	41,25
$Q_1 - 1,5(Q_3 - Q_1)$	-14,375
$Q_3 + 1,5(Q_3 - Q_1)$	74,625
max(dados)	150
min(dados)	11
Limite superior	11
Limite inferior	74,625

Deixamos ao leitor a tarefa de verificar os valores constantes das sete primeiras linhas, concentrando-nos naqueles das duas últimas.

O valor mínimo das observações é igual a 11*, i.e.,* $\min(dados) = 11$*; como* $Q_1 - 1,5(Q_3 - Q_1) = -14,375$*, temos então que*

$$\text{limite inferior} = \max\{11; -14,375\} = 11.$$

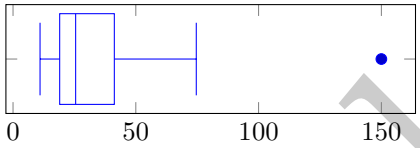
Como não nenhuma observação abaixo do limite inferior, não há outliers à esquerda da caixa menor.

Da mesma forma, o valor máximo das observações é $\max(dados) = 150$ *e* $Q_3 + 1,5(Q_3 - Q_1) = 74,625$*. Portanto,*

$$\text{limite superior} = \min\{150; 74,675\} = 74,625.$$

Nesse caso, há uma observação acima do limite inferior (o valor máximo 150*), de forma que este outlier será identificado, no boxplot, por um ponto (•).*

Chegamos assim, ao seguinte diagrama boxplot:



Além de fornecer informações importantes sobre um conjunto de dados, o diagrama boxplot também pode ser utilizado para comparar graficamente, relativamente à média, à dispersão e à distribuição, mais de um conjunto de medidas tomadas de uma mesma população (dados *multivariados*).

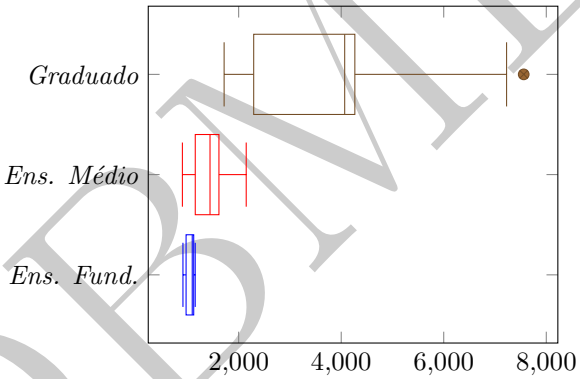
Isso pode ser conseguido desenhando-se os gráficos boxplot relativos a cada medida paralelamente, todos dentro de uma mesma caixa maior.

Faremos isso a seguir, o qual deixará claro como gráficos boxplot simplificam a análise de dados e a tomada de decisões de dados multivariados.

Exemplo 5. *Em um estudo feito sobre os salários de um grupo de* 30 *pessoas, separadas em três grupos de acordo com o nível de escolaridade, foram levantados os dados coletados na tabela a seguir:*

Calculando-se os valores necessários para a construção dos Boxplots e desenhando-os em paralelo, obtemos o diagrama a seguir (faça as verificações necessárias):

Ens. Fund.	Ens. Médio	Graduado
1152	902	2480
1152	1447	7560
926	1436	4274
961	2147	2231
1117	1669	4009
1106	1372	4328
1126	981	4128
913	1454	1716
1082	1776	4244
1014	1081	1905



Note que, apesar de encontrarmos pessoas que possuem o ensino médio completo ganhando menos do que pessoas com ensino fundamental completo, a mediana dos salários aumenta de acordo com o nível de escolaridade. Também, observa-se prontamente que a dispersão dos valores dos salários pagos aumenta com o número de anos de estudo, incrementando assim as chances de alguém que possui nível superior ter salários mais elevados.

3 Sugestões ao professor

O objetivo principal deste material é complementar a formação do professor. Por outro lado, se o professor estiver com a matéria adiantada e tiver alunos que venham apresentando bons resultados ao longo do módulo, pode ser interessante apresentar o material colecionado nesta aula. Caso contrário, uma sugestão é preparar uma aula de revisão com os conteúdos apresentados até a aula anterior.

Uma atividade interessante relacionada a gráficos boxplot é a seguinte: separe a turma em grupos de quatro ou cinco alunos e peça-os que coletem informações sobre as distâncias que os alunos da escola percorrem todos os dias para chegar em suas casas após o dia de estudo, separando as observações de acordo com o meio de transporte que cada um usa: a pé, bicicleta, transporte público ou transporte particular. Então, peça que utilizem essas informações para construir gráficos boxplot em paralelo e analisem o resultado em conjunto.

Referências

[1] João Ismael Pinheiro et al. *Estatística Básica: a arte de trabalhar com dados*. Campus, 2009.

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