Medidas de dispersão: os valores estão próximos entre si ou variam muito?

Measures of dispersion: are all values close to each other or do they vary a lot?

JOÃO LUIZ DORNELLES BASTOS[1]
RODRIGO PEREIRA DUQUIA[2]

DESCRITORES: MEDIDAS EM EPIDEMIOLOGIA; EPIDEMIOLOGIA E BIOESTATÍSTICA; ANÁLISE DE DADOS; FATORES EPIDEMIOLÓGICOS.

KEY WORDS: *EPIDEMIOLOGIC MEASUREMENTS; EPIDEMIOLOGY AND BIESTATISTICS; DATA ANALYSIS; EPIDEMIOLOGIC FACTORS.*

As informações e os conhecimentos adquiridos com a leitura da segunda *Nota de Epidemiologia e Bioestatística*, cujo título é "Medidas de tendência central: onde a maior parte dos indivíduos se encontra?" impõe algumas novas necessidades. Para além de expressar através de um único valor em torno do qual tende a se concentrar um conjunto de dados numéricos, importa saber como estas observações estão distribuídas em nossa população de estudo – são elas bastante próximas entre si ou variam muito?[1]

Isto ocorre porque duas distribuições podem apresentar médias aritméticas idênticas e, ao mesmo tempo, possuir valores que se distribuem de maneiras completamente diferentes em relação a ela. Para ilustrar, considere a Tabela 1, que apresenta a situação hipotética de duas distribuições das pontuações obtidas por um grupo de alunos de ensino médio nas disciplinas de Biologia e Matemática. As médias nas duas disciplinas são iguais e equivalem a 5 (cinco). No entanto, ao examinar a tabela mencionada e as Figuras 1 e 2, percebe-se que as distribuições são diferentes entre si. Enquanto na disciplina de Biologia a maior parte dos indivíduos tendeu a uma nota próxima de 5, em Matemática houve maior dispersão das pontuações, isto é, as notas variaram mais entre os alunos.

A partir desta constatação, coloca-se a seguinte pergunta: Existe alguma medida capaz de expressar a forma como as observações se distribuem em um conjunto de dados? A resposta a este questionamento remete o(a) leitor(a) às chamadas medidas de dispersão, que nada mais são do que medidas que indicam como as observações estão dispostas em uma dada distribuição (se estão dispersas ou próximas entre si na amostra estudada).

[1] Odontólogo. Mestre em Epidemiologia pela Universidade Federal de Pelotas.
[2] Dermatologista do Hospital Santa Casa de Porto Alegre. Mestre em Epidemiologia pela Universidade Federal de Pelotas.

TABELA 1 – Pontuações obtidas por alunos do ensino médio conforme as disciplinas cursadas (dados hipotéticos).

Aluno	Biologia	Matemática
Ana	5	5
Carla	6	3
César	5	8
João Paulo	4	5
José Nilton	5	2
Luiz Roberto	5	5
Marcelo	5	10
Maria	6	10
Mariana	4	5
Pâmela	7	3
Pedro	3	2
Roberta	5	2
Média aritmética ($\overline{X}$)	**5**	**5**

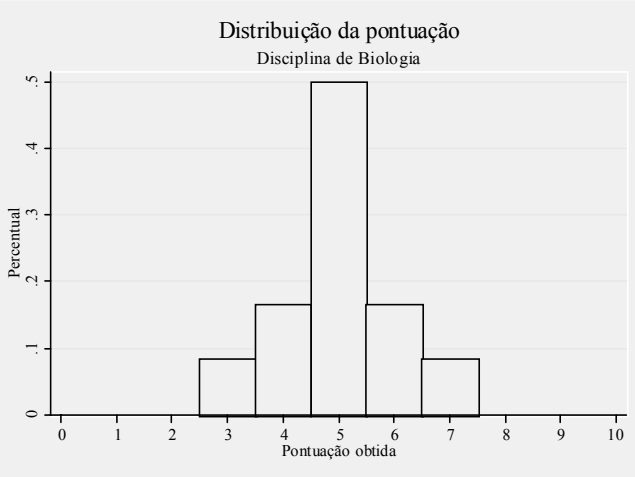

Figura 1 – Distribuição da pontuação de alunos do ensino médio na disciplina de Biologia (dados hipotéticos).

Figura 2 – Distribuição da pontuação de alunos do ensino médio na disciplina de Matemática (dados hipotéticos).

A variabilidade (ou dispersão) de um conjunto de dados pode ser quantificada através da amplitude de variação, da variância, do desvio-padrão e do coeficiente de variação, entre outras.[1-4] Nas seções que seguem, são apresentadas as fórmulas e exemplos do cálculo de cada uma das quatro medidas, bem como suas vantagens e desvantagens para utilização na análise de dados e leitura crítica de trabalhos científicos.

1 AMPLITUDE DE VARIAÇÃO

A amplitude de variação pode ser obtida facilmente através da diferença entre o maior e o menor valor de uma distribuição de dados.[3] Aproveitando o exemplo das notas nas disciplinas de Biologia e Matemática, a amplitude de variação em cada um dos casos foi de 4 (7 [maior pontuação] – 3 [menor pontuação] = 4) e de 8 (10 [maior pontuação] – 2 [menor pontuação] = 8), respectivamente. A maior amplitude de variação nas notas de Matemática está de acordo com o que foi observado na Figura 2, ou seja, em uma distribuição com maior dispersão dos dados, a amplitude de variação tende a ser maior. Entretanto, cabe salientar que o cálculo da medida em questão não leva em consideração os valores intermediários da distribuição, de forma que estes não influenciam seu resultado final.[2] Esta poderia ser considerada uma desvantagem, uma vez que as medidas de dispersão deveriam levar em conta todas as observações e não somente os limites do conjunto de dados, isto é, seus valores máximo e mínimo.

2 VARIÂNCIA (s^2)

Ao contrário da medida supracitada, a variância consiste em uma medida de dispersão que leva em conta todos os valores de uma distribuição para seu cálculo.[2] Ela é estimada a partir do somatório do quadrado da distância de cada valor em relação à média, dividido pelo total de observações menos um, tal como na fórmula:[2]

$$s^2 = \frac{\Sigma\,(x - \overline{X})^2}{(n - 1)},$$

onde s^2 corresponde à variância, Σ ao somatório, x aos valores observados, $\overline{X}$ à média da distribuição e n ao tamanho da amostra estudada.

A aplicação desta fórmula pode ser ilustrada com as pontuações obtidas nas disciplinas des-

tacadas na Tabela 1. Em Biologia, a variância poderia ser calculada pela fórmula:

$$s^2 = \frac{(5-5)^2+(6-5)^2+(5-5)^2+(4-5)^2+(5-5)^2+(5-5)^2+(5-5)^2+(6-5)^2+(4-5)^2+(7-5)^2+(3-5)^2+(5-5)^2}{(12-1)} = 1{,}09$$

Realizando o mesmo cálculo para Matemática, chega-se ao valor de variância de 8,54. A maior variância na distribuição das notas desta última é conseqüente à maior dispersão dos dados nesta disciplina, quando comparada com a Biologia. Perceba que esta maior dispersão nas notas de Matemática já havia sido acusada na Figura 2, o que significa que a dispersão de um conjunto de dados também pode ser verificada visualmente, através de gráficos do tipo histograma, por exemplo.

Uma desvantagem considerável desta medida de variabilidade reside no fato de que seu resultado é oferecido na unidade de medida dos dados elevada ao quadrado.[2] Exemplificando, a variância da altura em metros de indivíduos incluídos em um estudo será expressa em metros quadrados. Isto confere maior complexidade de interpretação à medida e, como forma de contornar o problema, calcula-se sua raiz quadrada. A raiz quadrada da variância é denominada desvio-padrão, que receberá maior atenção na seção abaixo.

3 DESVIO-PADRÃO (s)

O desvio-padrão é amplamente utilizado na literatura científica como medida de dispersão dos dados. Ele estima o quanto, em média, cada valor se distancia da própria média aritmética de uma distribuição com a vantagem de preservar a unidade de mensuração original das observações, algo que não ocorre com a variância. Para calculá-lo, basta extrair a raiz quadrada da fórmula da variância:[2,3]

$$s = \sqrt{\frac{\Sigma\,(x - \bar{X})^2}{(n-1)}},$$

onde s equivale ao desvio-padrão, Σ ao somatório, x aos valores observados, $\bar{X}$ à média da distribuição e n ao tamanho da amostra estudada.

Retomando o exemplo das disciplinas de Biologia e Matemática, o desvio padrão em ambas distribuições de notas seria 1,04 ($\sqrt{1{,}09}$ = 1,04, onde 1,09 equivale à variância calculada no item 2) e 2,92 ($\sqrt{8{,}54}$ = 2,92, onde 8,54 corresponde à variância calculada no item 2), respectivamente. À primeira vista, utilizar o desvio padrão como medida de dispersão não ofereceria qualquer vantagem em relação ao uso da variância, a não ser pelo fato de conservar a unidade original de medida das observações.

A maior vantagem desta medida de dispersão é que, em distribuições Normais ou Gaussianas, 68% das observações encontram-se distanciadas em até um desvio-padrão em relação à média, para mais e para menos.[1] De modo análogo, 95% e 100% das observações de uma distribuição Gaussiana encontram-se entre mais e menos dois e mais e menos três desvios-padrão da média.[1] A Figura 4 mostra que 68%, 95% e 100%* dos valores estão contidos entre um, dois e três desvios-padrão da média aritmética em distribuições Normais. Esta informação é importante quando do cálculo de intervalos de confiança e do estabelecimento de inferências, assuntos a serem tratados em notas futuras.

Além disso, conhecendo-se o valor do desvio-padrão e da média aritmética de uma distribuição é possível saber se esta tende a uma forma simétrica, também dita Normal, ou assimétrica. Nos casos em que a distribuição dos dados for assimétrica, o desvio padrão será maior do que a metade da média aritmética (em distribuições assimétricas $s > \bar{X}/2$, onde s é o desvio padrão e $\bar{X}$ a média aritmética).[1] É importante levar em consideração este fato, pois boa parte dos testes utilizados nas análises estatísticas tem como pressuposto que a distribuição dos dados seja, pelo menos, próxima à Normal. O teste t de *Student*, por exemplo, largamente utilizado na comparação de médias entre dois grupos, tem como um de seus requisitos (pressupostos) que a distribuição da variável em questão seja Normal. Na ausência de informações gráficas sobre como se distribui uma variável, mas tendo-se à disposição

* A área sob a curva Normal compreendida entre um, dois e três desvios-padrão para mais e para menos da média aritmética é de 68,3%, 95,4% e 99,7%, respectivamente. Por motivos de simplificação, estamos arredondando estes valores ao longo do presente texto.

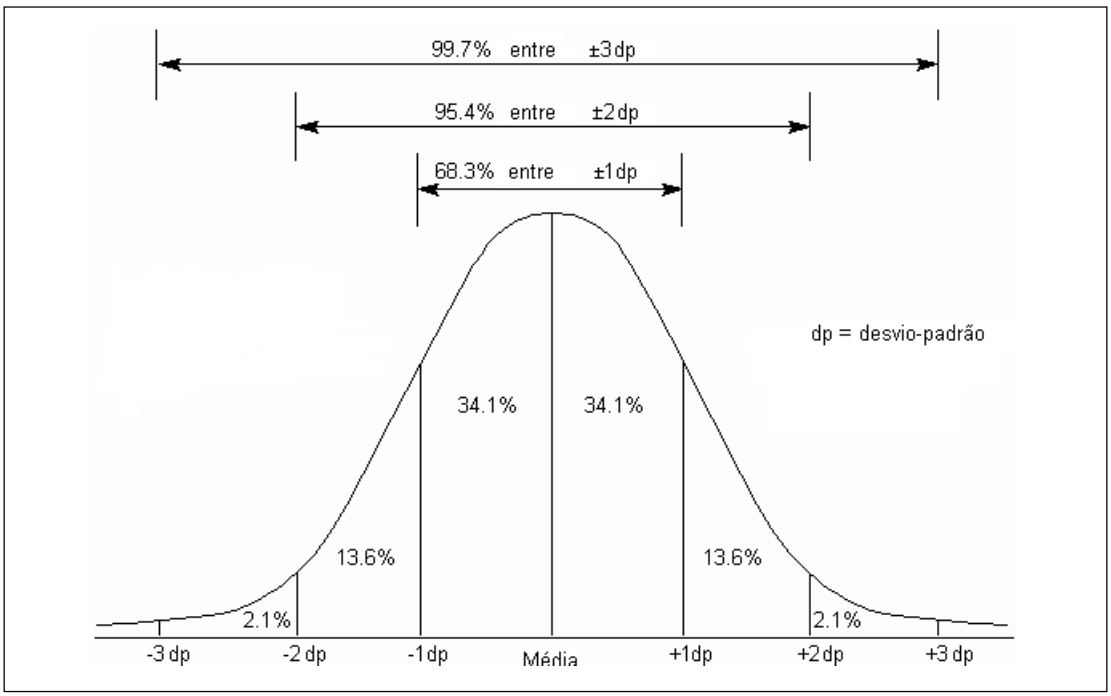

Figura 3 – Área sob a curva da distribuição Normal (Gaussiana) e sua relação com os desvios padrão.

os valores de média e desvio-padrão, pode-se julgar adequado ou não o uso de um teste estatístico em uma publicação científica.[5] Se, em uma publicação, tiver sido adotado o teste t para comparar a média de algum atributo entre dois grupos e o desvio-padrão for maior do que metade da média aritmética, pode-se considerar inadequado seu uso e colocar sob suspeita o resultado apresentado.

4 COEFICIENTE DE VARIAÇÃO (c_v)

O coeficiente de variação, por sua vez, refere-se à divisão entre o desvio padrão e a média de uma distribuição:[2,3]

$$c_v = \frac{s}{\overline{X}},$$

onde c_v é o coeficiente de variação, s é o desvio padrão e $\overline{X}$ a média aritmética.

Coeficientes de variação menores do que 0,2 sugerem pouca dispersão nos dados, enquanto coeficientes maiores que 1 indicam dispersão bastante elevada.[3] Especificamente, coeficientes maiores que 0,5 também sugerem que a distribuição analisada tende a uma forma assimétrica ou não-Normal.

Esta medida consiste em uma forma simples de avaliar a dispersão de uma variável, uma vez que não possui unidade de medida. Assim, é possível comparar a dispersão entre duas variáveis, mesmo que tenham sido mensuradas em escalas de medida diferentes e possuam médias diferentes. Por exemplo, através do coeficiente de variação pode-se comparar, diretamente e sem o recurso de transformações, a variabilidade existente em uma distribuição de alturas medidas em metros com outra de alturas medidas em milímetros. Mesmo com estas vantagens, o coeficiente de variação é pouco utilizado e cede lugar na maioria das vezes ao desvio-padrão e à variância nas análises estatísticas e nas publicações científicas.

5 CONSIDERAÇÕES FINAIS

Os conhecimentos introduzidos no presente artigo, somados àqueles da nota anterior, fornecem informações básicas e necessárias para se conhecer as principais características de uma distribuição, tais como sua forma e dispersão. Quando a distribuição dos dados não se aproxima de uma forma Normal, muitos testes estatísticos são contra-indicados e o uso deles pode produzir resultados inválidos. Nestes casos, pode-se transformar os dados (calculando-se o logaritmo dos valores, por exemplo) para que a distribuição assuma uma forma mais próxima da Normal ou

utilizar métodos estatísticos que não tenham como pressuposto que a distribuição seja simétrica.

Além destas aplicações práticas, os conceitos de distribuição Normal e de desvio-padrão estão intimamente relacionados com o cálculo de intervalos de confiança e com o estabelecimento de inferências. Estes intervalos são estimativas de precisão de um determinado valor e receberão destaque em notas futuras.

REFERÊNCIAS

1. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1997.
2. Kirkwood BR, Sterne JAC. Essential medical statistics. Oxford: Blackwell Science; 2003.
3. Peres KG. Apresentação de dados epidemiológicos. In: Antunes JLF, Peres MA, editores. Fundamentos de odontologia: epidemiologia da saúde bucal. Rio de Janeiro: Guanabara Koogan; 2006. p.409-21.
4. Peres MA, Antunes JLF, Frazão P. Cárie dentária. In: Antunes JLF, Peres MA, editores. Fundamentos de odontologia: epidemiologia da saúde bucal. Rio de Janeiro: Guanabara Koogan; 2006. p.49-67.
5. Altman DG, Bland JM. Statistics notes: detecting skewness from summary information. BMJ. 1996; 313:1200.

Endereço para correspondência:
JOÃO LUIZ DORNELLES BASTOS
Avenida do Antão, 353 - Morro da Cruz
CEP 88025-150, Florianópolis, SC, Brasil
Fone: (0xx48) 3028-1345
E-mail: joao@pilotis.com.br