Determinação de medidas de posição a partir de dados agrupados

Rinaldo Artes

Em algumas situações, o acesso aos microdados de uma pesquisa é restrito ou tecnicamente difícil. Em seu lugar, são divulgados ao público informações em forma de tabelas (distribuições de frequências – dados agrupados). Discutiremos como obter estimativas de medidas de posição a partir de dados agrupados. Admitimos que o leitor já conheça as medidas descritivas utilizadas neste texto, deste modo, o foco estará em aspectos operacionais ligados à obtenção dos coeficientes.

Considere, como ponto de partida, o seguinte exemplo:

Exemplo: Uma instituição financeira (IF) utiliza um indicador do grau de endividamento para avaliar a situação de suas carteiras de empréstimo. O indicador é calculado para cada participante da carteira e, grosseiramente, quanto maior o valor do indicador, pior é a situação do cliente. A IF está interessada em descrever a situação de duas carteiras distintas:

Carteira Alfa – com empréstimos concedidos em 2012 e

Carteira Beta – com empréstimos concedidos em 2013.

Ela conta com uma amostra de 500 clientes de cada carteira. As Tabelas 1 e 2 resumem os resultados desta pesquisa.

Notação:

k: número de classes da tabela;

n_i: frequência absoluta (contagem) da classe i;

f_i: frequência relativa (proporção) da classe i;

F_i: frequência (relativa) acumulada da classe i e

$\delta_i = \frac{n_i}{\Delta_i}$: densidade de frequência da classe i, sendo Δ_i a amplitude (largura) da classe;

Tabela 1: Distribuição de frequências e densidades de frequência do grau de endividamento de clientes da carteira Alfa

Grau de Endividamento	n_i	$f_i \times 100$	$F_i \times 100$	δ_i
0 |-- 5	61	12,2	12,2	0,02440
5 |-- 10	107	21,4	33,6	0,04280
10 |-- 15	97	19,4	53,0	0,03880
15 |-- 20	77	15,4	68,4	0,03080
20 |-- 30	77	15,4	83,8	0,01540
30 |-- 50	63	12,6	96,4	0,00630
50 |-- 75	18	3,6	100	0,00144
Total	500	1,000		

Tabela 2: Distribuição de frequências e densidades de frequência do grau de endividamento de clientes da carteira Beta

Grau de Endividamento	n_i	f_i	F_i	δ_i
0 |-- 5	55	0,110	0,110	0,02200
5 |-- 10	83	0,166	0,276	0,03320
10 |-- 15	79	0,158	0,434	0,03160
15 |-- 20	77	0,154	0,588	0,03080
20 |-- 30	75	0,15	0,738	0,01500
30 |-- 50	100	0,2	0,938	0,01000
50 |-- 100	31	0,062	1	0,00124
Total	500	1,000		

Deseja-se comparar as duas carteiras em termos de medidas de posição e de variabilidade.

1. Obtenção de medidas de posição a partir de dados agrupados

Serão discutidas formas de obtenção da média aritmética ($\bar{x}$), da mediana (m_d) e da moda (m_o) dessas distribuições.

1.1. Média aritmética ($\bar{x}$)

Seja $x_1, \ldots, x_n$ uma amostra de n observações de uma variável de interesse, X. A média aritmética é obtida a partir de

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}. \qquad (1)$$

No caso dos dados da Tabela 1, por exemplo, não temos mais as informações de todas as 500 observações separadamente. Sabemos apenas que há 61 observações menores que 5, 107 entre 5 e 10 e assim por diante. Isso, em princípio, impede o uso da expressão (1) para determinar a média aritmética. A obtenção desse indicador exige que se faça uma suposição sobre o comportamento dos dados apresentados na tabela. Admita que seja possível supor que, em cada classe, os dados estejam uniformemente distribuídos. Sendo isso verdadeiro, é possível afirmar que a soma de todas as observações pertencentes à classe i será igual ao ponto médio da classe (c_i) vezes a quantidade de observações da classe (n_i), conforme descrito na equação (2).

$$\sum_{j \in Classe\ i} x_j = c_i\, n_i \tag{2}$$

Sob a suposição de uniformidade em cada classe,

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{\sum_{i=1}^{k} c_i\ n_i}{n}. \tag{3}$$

Levando-se em conta que $f_i = {}^{n_i}/_{n}$, temos

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{\sum_{i=1}^{k} c_i\ n_i}{n} = \sum_{i=1}^{k} c_i \frac{n_i}{n} = \sum_{i=1}^{k} c_i\, f_i. \tag{4}$$

Na prática, será difícil definir classes com as observações distribuídas de modo perfeitamente uniforme, assim, os resultados apresentados são aproximados. A aproximação será melhor à medida que as classes forem definidas de modo que a distribuição dos pontos em cada uma se aproxime da suposição de uniformidade.

A Tabela 3 resume os cálculos para a obtenção da média para a Carteira Alfa utilizando os resultados (3) e (4).

Tabela 3: Determinação da média aritmética para os dados da Carteira Alfa.

Endividamento	c_i	$\bar{x} = \frac{\sum_{i=1}^{k} c_i \ n_i}{n}$		$\bar{x} = \sum_{i=1}^{k} c_i \, f_i$	
		n_i	$c_i \, n_i$	f_i	$c_i \, f_i$
0 |-- 5	2,5	61	152,5	0,122	0,305
5 |-- 10	7,5	107	802,5	0,214	1,605
10 |-- 15	12,5	97	1212,5	0,194	2,425
15 |-- 20	17,5	77	1347,5	0,154	2,695
20 |-- 30	25,0	77	1925,0	0,154	3,85
30 |-- 50	40,0	63	2520,0	0,126	5,04
50 |-- 75	62,5	18	1125,0	0,036	2,25
Total		**500**	**9085**	**Média=**	**18,17**
		Média=	**9085/500=18,17**		

1.2. Mediana (m_d)

Por definição, a mediana é o ponto que divide a amostra ordenada em duas partes com o mesmo número de observações. A frequência acumulada nos dá a localização desta estatística: a mediana será o valor que deixa 50% das observações abaixo dela. Tomando por base a Tabela 1, o grau de endividamento mediano dos clientes da Carteira Alfa está entre 10 e 15. Notem que até 10, temos 33,6% das observações (coluna F_i) e que até 15, temos 53,0%, logo o valor que supera 50% da amostra ordenada, deve estar entre 10 e 15.

Discutiremos duas abordagens para o cálculo da mediana. A primeira passa pela construção de uma regra de três, seguindo o seguinte raciocínio:

a) A mediana está entre 10 e 15.
b) Há 19,4% de observações nesta classe.
c) Até 10, temos 33,6% dos dados; faltam 50-33,6=16,4% para se atingir a mediana.

A partir disso, construímos a seguinte regra de três:

$$\begin{matrix} 15-10 & --- & 19{,}4\% \\ m_d-10 & --- & 16{,}4\% \end{matrix}.$$

De onde vem que

$$(m_d - 10) \times 19{,}4\% = (15 - 10) \times 16{,}4\%, \text{ portanto,}$$

$$m_d = 10 + (15 - 10) \times \frac{16{,}4\%}{19{,}4\%} = 14{,}23.$$

Este valor também pode ser obtido a partir da densidade de frequência. Lembre que num histograma (neste texto, definimos como histograma um gráfico de barras construído a partir de δ_i), a área de cada barra coincide com a frequência relativa. A Figura 1 corresponde ao histograma do Grau de Endividamento dos clientes da Carteira Alfa, obtido a partir da Tabela 1.

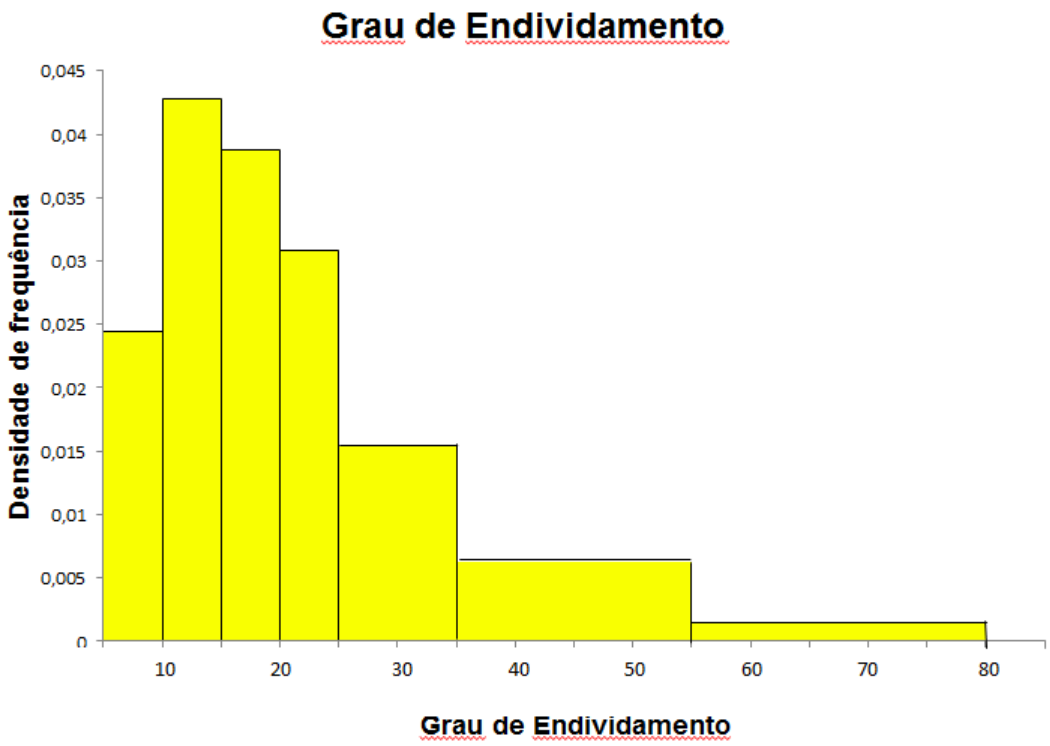

Figura 1: Histograma do Grau de Endividamento dos clientes da Carteira Alfa.

A Figura 2 representa determinação da mediana a partir de um histograma. Estão representadas a barra entre as observações 10 e 15 e a área entre 10 e a mediana (retângulo preenchido). Como num histograma a área sobre um intervalo corresponde à proporção de observações encontradas no intervalo, temos

$$16{,}4\% \;=\; 0{,}164 \;= (m_d - 10) \times \delta_i = (m_d - 10) \times 0{,}0388$$

Assim,

$$m_d = 10 + \frac{0{,}164}{0{,}0388} = 14{,}23.$$

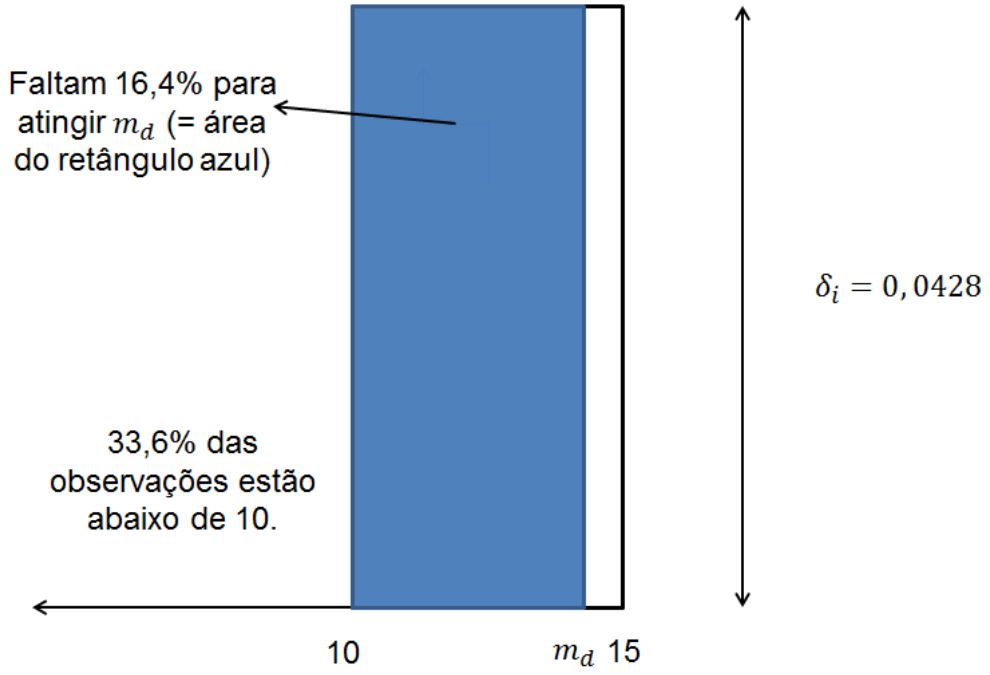

Figura 2: Representação gráfica do cálculo da mediana a partir de um histograma.

1.3. Moda (m_o)

Uma maneira adequada de se obter a moda para um variável contínua é defini-la como o valor com maior densidade, ou seja, o ponto mais alto de um histograma.

No caso dos dados da Carteira Alfa, a classe modal (classe que contém a moda) está entre 5 e 10 (maior densidade da Tabela 1). Uma maneira simples de obter um valor para a moda é considerar o ponto médio da classe mediana. Neste caso, $m_o = 7{,}5$.

Ao se observar a Figura 1, no entanto, repare que a barra que sucede a classe de 5 a 10 é maior do que a barra que antecede esta classe. Seria de se esperar que a moda estivesse mais próxima da classe vizinha com barra mais

alta do que da com barra mais baixa. Uma maneira de fazer isso é considerar a moda como

$$m_o = \frac{\alpha\ \delta_a + \beta\ \delta_p}{\alpha + \beta},$$

na qual, α é o limite inferior da classe modal, β é o limite superior da classe modal, δ_a é a densidade de frequência da classe anterior à modal (igual a zero se a classe modal for a primeira classe) e δ_p é a densidade de frequência da classe posterior à modal (igual zero se a classe modal for a última classe). Para os dados da Carteira Alfa, temos por este método:

$$m_o = \frac{\alpha\ \delta_a + \beta\ \delta_p}{\alpha + \beta} = \frac{5\ \times 0{,}0244 + 10\ \times 0{,}0388}{0{,}0244 + 0{,}0388} = 8{,}07.$$

Repare que os valores obtidos pelos dois métodos não coincidem.