EXPLICANDO PORQUE A MEDIANA PERTENCE À CLASSE MEDIANA EM UM HISTOGRAMA

R. C. SANTOS[1]

Resumo

Costuma-se dizer que a mediana é o valor que "divide" o conjunto em dois subconjuntos de mesmo tamanho, ou de mesma ordem. Essa interpretação nem sempre é esclarecida nos livros. Vejamos como deve ser entendida essa divisão em dois subconjuntos para cada caso, n par e n ímpar. Vamos explorar também o conceito de mediana a partir de dados agrupados em classes, isto é, a partir de um histograma. Além disso, será demonstrado que a Mediana de fato pertence à Classe Mediana em um Histograma.

1 Definição de Mediana a partir de um conjunto numérico finito

De acordo com [LEONARDECZ – 2004], a mediana de um conjunto numérico finito e ordenado $\{y_1 \leq y_2 \leq y_3 \leq \cdots \leq y_n\}$ é, por definição:

- O elemento de ordem $\frac{n+1}{2}$, se n for ímpar, ou seja, $y_{(n+1)/2}$;
- A média aritmética dos elementos de ordem n/2 e (n/2) + 1, se n for par, ou seja, $\frac{y_{n/2}+y_{(n/2)+1}}{2}$.

Sendo a Mediana o valor que divide um conjunto ordenado em dois subconjuntos de mesma ordem, considere os três exemplos: o conjunto A = {2, 2, 3}, cuja ordem n = 3, é ímpar, que possui mediana $y_2 = 2$; o conjunto B = {1, 2, 2, 5}, com n = 4 par, que possui mediana (2 + 2) / 2 = 2, e o conjunto C = {1, 3, 6, 7}, n = 4 também par, que possui mediana (3 + 6) / 2 = 4,5.

Em A, a mediana é 2, e a quantidade de elementos menores ou iguais a ela é a mesma quantidade de elementos que são maiores ou iguais, por isso a mediana "divide" A em dois subconjuntos. Mas, a intersecção destes conjuntos não é vazia, e isso acontece toda vez que n é ímpar.

No caso em que n é par, a mediana pode coincidir com algum elemento do conjunto (como em B e em C). Observe que $\frac{y_{n/2}+y_{(n/2)+1}}{2} = y_{n/2}$ ou $y_{(n+1)/2}$, se e somente se, $y_{n/2} = y_{(n+1)/2}$.

Em B, há três valores menores ou iguais do que a mediana e três valores maiores ou iguais a ela, os dois subconjuntos têm um elemento em comum: a própria mediana. Além disso, há apenas um valor maior do que a mediana, o que caracteriza a *assimetria* do conjunto.

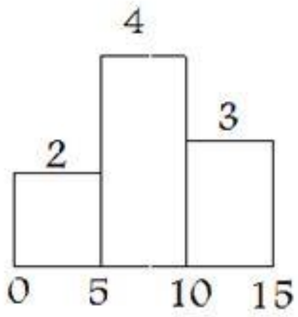
Em C, há dois valores menores do que a mediana 4,5 e dois valores maiores do que 4,5, os dois subconjuntos são disjuntos.

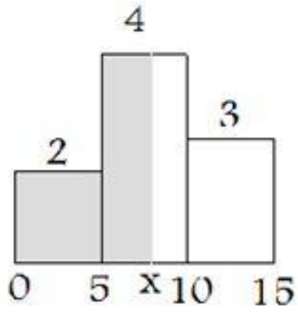
[1] Universidade de Brasília (Planaltina), e-mail: professorrogeriocesar@gmail.com

Em resumo, se n for ímpar, os subconjuntos não são disjuntos, mas quando n é par, os subconjuntos podem ser disjuntos ou não. Em todo caso, os dois subconjuntos têm mesma ordem ou tamanho.

2 Dados agrupados em classes, através de um Histograma

Ao ministrar uma aula de estatística, pedi para meus alunos calcularem a mediana a partir do histograma ilustrado abaixo:





Como os dados estão organizados em classes através de um histograma, a mediana é o valor x que deixa a soma das áreas dos retângulos hachurados, à esquerda do histograma, igual à soma das áreas dos retângulos não hachurados à direita.

Então, os alunos deveriam fazer a seguinte conta: a soma das áreas dos retângulos hachurados, na figura da direita, é igual à soma das áreas dos demais retângulos. Observe que o valor x deve pertencer à *classe* mediana, cujo conceito será falado adiante. A classe mediana nesse exemplo é a classe de intervalo]5, 10]. No entanto, um dos alunos encontrou um valor x que não pertencia à classe mediana, ou seja, estava entre 0 e 5, e não entre 5 e 10 como deveria. Daí pode surgir a seguinte questão: o valor x que divide a área do histograma sempre pertencerá à classe mediana?

Considere o histograma abaixo, onde o vértice inferior esquerdo do primeiro retângulo da esquerda foi colocado na origem do semi-eixo x positivo. Para tanto, considere que:

- $f_1, f_2, f_3, ..., f_n$ são as frequências de cada classe, que coincidem com as alturas dos retângulos;
- c é o comprimento de cada classe, que coincide com o tamanho da base dos retângulos;
- os valores $0, 2c, 3c, ..., kc, ..., nc$ são as abscissas correspondentes aos vértices dos retângulos, localizados no semi-eixo x positivo;

os valores observados coincidentes com o extremo esquerdo de cada intervalo ou classe pertencem à classe da esquerda, e os correspondentes ao extremo direito pertencem à própria classe correspondente, isto é, se há uma observação $y = ic$ para algum i, então y pertence à classe da esquerda, isto é, a classe correspondente ao intervalo semi-aberto $]\ (i-1)c, ic]$ para $i = 1, \ldots, n$.

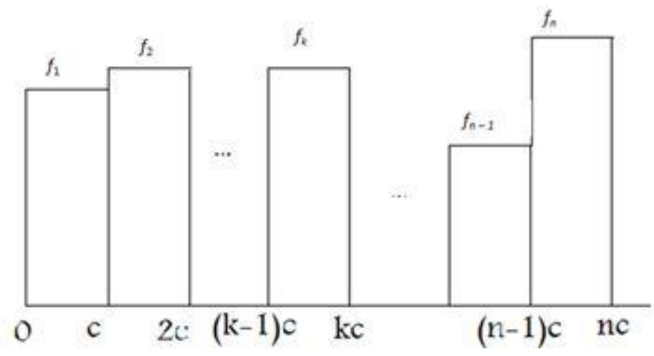


Fig. 1 Histograma

Nota: Em cada intervalo semi-aberto $](i-1)c, ic]$ há f_i valores, para $i = 1, \ldots, n$.

3 Definição de Classe Mediana

De acordo com a definição, a classe mediana é a primeira das classes, da esquerda para a direita, ou da direita para a esquerda, cuja frequência acumulada é maior ou igual do que a metade da soma S das frequências. Isto é, se $S = \sum_{i=1}^{n} f_i$, então a classe mediana, que na figura 1 será por suposição o intervalo semi-aberto $](k-1)c, kc]$, é tal que $\sum_{i=1}^{k-1} f_i < S/2$ e $\sum_{i=1}^{k} f_i \geq S/2$.

4 Definição de Mediana para Distribuição de Frequência

Dado o histograma, a mediana é o valor x tal que a soma das áreas dos retângulos hachurados na figura abaixo (os da esquerda) é igual à soma das dos retângulos à direita. A nossa pergunta é: existe um tal valor x? E se existe, ele se localiza obrigatoriamente na classe mediana, ou seja, $(k-1)c < x \leq kc$?

O valor x que divide a área do histograma em dois subconjuntos de igual frequência pertence à classe mediana?

De fato, a mediana sempre pertence à classe mediana, como será demonstrado a seguir:

Proposição. *Suponha que a classe $](k-1)c, kc]$ seja a classe mediana, isto é, suponha*

$$\sum_{i=1}^{k-1} f_i < \frac{S}{2} \quad \text{(I)}$$

$$\sum_{i=1}^{k} f_i \geq \frac{S}{2}. \quad \text{(II)}$$

Então, o valor x que deixa metade da área do histograma à sua esquerda e metade da área à sua direita pertence à classe mediana. Em outras palavras a equação

$$cf_1 + cf_2 + \cdots + cf_{k-1} + (x-(k-1)c)f_k = (kc-x)f_k + cf_{k+1} + \cdots + cf_n \quad \text{(III)}$$

tem solução x pertencente ao intervalo semiaberto $](k-1)c, kc]$.

Demonstração:

Primeiro, observe que a hipótese (I) é equivalente a

$$2\sum_{i=1}^{k-1} f_i < \sum_{i=1}^{n} f_i,$$

ou seja,

$$\sum_{i=1}^{k-1} f_i < \sum_{i=k}^{n} f_i.$$

Também, a hipótese (II) é equivalente a

$$2\sum_{i=1}^{k} f_i \geq \sum_{i=1}^{n} f_i,$$

ou seja,

$$\sum_{i=1}^{k} f_i \geq \sum_{i=k+1}^{n} f_i.$$

Agora, resolvendo a equação (III) para *x*, temos:

$$cf_1 + \cdots + cf_{k-1} + x \cdot f_k - kcf_k + cf_k = kf_kc - x \cdot f_k + cf_{k+1} + \cdots + cf_n,$$

ou,

$$x = \frac{2kcf_k + c\sum_{i=k+1}^{n} f_i - c\sum_{i=1}^{k-1} f_i - cf_k}{2f_k}.$$

Esse *x* satisfaz $(k-1)c < x \leq kc$, como veremos a seguir:

A desigualdade da esquerda,

$$(k-1)c < \frac{2kcf_k + c\sum_{i=k+1}^{n} f_i - c\sum_{i=1}^{k-1} f_i - cf_k}{2f_k},$$

é equivalente a:

$$2kcf_k + c\sum_{i=k+1}^{n} f_i - c\sum_{i=1}^{k-1} f_i - cf_k > 2f_kkc - 2f_kc \Leftrightarrow$$
$$\sum_{i=k+1}^{n} f_i - \sum_{i=1}^{k-1} f_i - f_k + 2f_k > 0 \Leftrightarrow$$
$$\sum_{i=k+1}^{n} f_i - \sum_{i=1}^{k-1} f_i + f_k > 0 \Leftrightarrow$$
$$\sum_{i=k}^{n} f_i > \sum_{i=1}^{k-1} f_i,$$

que é verdadeira pela hipótese (I).

Já a desigualdade da direita,

$$\frac{2kcf_k + c\sum_{i=k+1}^{n} f_i - c\sum_{i=1}^{k-1} f_i - cf_k}{2f_k} \leq kc$$

$$2kcf_k + c\sum_{i=k+1}^{n} f_i - c\sum_{i=1}^{k-1} f_i - cf_k \leq 2fkc \Leftrightarrow$$
$$\sum_{i=k+1}^{n} f_i - \sum_{i=1}^{k-1} f_i - f_k \leq 0 \Leftrightarrow$$
$$\sum_{i=k+1}^{n} f_i \leq \sum_{i=1}^{k} f_i$$

que é verdade pela hipótese (II). □

5 Conclusão

A demonstração aqui apresentada não deixa dúvidas. A mediana sempre pertencerá à classe mediana. Não tenham receio, professores, se o valor encontrado por algum aluno estiver fora desta classe, com certeza houve algum erro de cálculo.

Referências

[LEONARDECZ – 2004] LEONARDECZ, E. **Estatística para ciências da vida.** Brasília: Universa, 2004