Mediana de dados não agrupados: a questão de ser pelo menos 50%

Adelaide Freitas
João Pedro Cruz
Nélia Silva

Com o presente trabalho temos como objetivo ilustrar exemplos que abrangem as diferentes situações que podem ocorrer no cálculo da mediana de uma coleção de dados. Pretendemos mostrar, com exemplos, as possíveis interpretações válidas de mediana que, de forma similar, podem ser extensíveis aos outros quartis. Com esses exemplos pretendemos discutir o eventual confronto entre a noção intuitiva de mediana e sua interpretação. Deste modo, esperamos contribuir para uma melhor compreensão de possíveis deduções, consequentes da interpretação de mediana, que por vezes causam alguma estranheza entre os estudantes, nomeadamente quando, por exemplo, existem observações repetidas e iguais à mediana numa coleção de dados.

CONCEITO EMPÍRICO DE MEDIANA: CÁLCULO E INTERPRETAÇÃO

Associado a uma coleção de dados, Murteira et al. (2010) refere que a mediana é, "de forma aproximada, o valor da coleção que tem 50% de observações inferiores e 50% de observações superiores" (p. 26), acrescentando que "em termos rigorosos, a mediana pode ser definida" do seguinte modo em termos de

$$med = \begin{cases} x_{((n+1)/2)} & , se\ n\ ímpar \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & , se\ n\ par \end{cases} , \quad (1)$$

onde n é o número total de observações e $x_{(i)}$ representa a i-ésima estatística ordinal (i.e., a i-ésima observação na amostra ordenada). De forma similar, Pestana & Velosa (2008) referem o conceito empírico de mediana, indicando que "a mediana é um conceito que envolve a ideia de ordenação: informalmente (…), é o valor que separa os 50% inferiores dos 50% superiores." (p. 86), e assinala também a fórmula de cálculo (1) para o caso de uma coleção de dados.

Em Guimarães & Cabral (1997) e em Mood et al. (1974), a definição de mediana empírica é dada pela fórmula (1). Em Hall et al. (2011) é dada a noção de mediana amostral usando a definição (1) e interpretando-a como sendo o "elemento que se encontra ao centro" na amostra depois de ordenada (pp. 7-8).

Na verdade, o uso da fórmula (1) é consensual. Ela é encontrada na generalidade dos atuais manuais escolares, dos 7.º e 8.º anos de escolaridade, em concordância com os programas curriculares de matemática (MEC, 2013), e dos livros de estatística que, ao nível do ensino superior, abordam a parte da estatística descritiva.

Se no cálculo do valor da mediana de uma coleção de dados (não agrupados), no contexto da estatística descritiva, é prática comum seguir a fórmula (1), não podemos dizer que encontramos uma uniformização de escrita quando se pretende descrever, por extenso, o significado de mediana. Em Amaro et al. (2009), para além de indicar a fórmula (1) no cálculo da mediana de "dados exaustivos", é esclarecido que se deve ter cuidado com a interpretação daquela medida consoante a natureza dos dados. Nomeadamente, refere que, para dados quantitativos contínuos, "50% das observações são inferiores – ou superiores à mediana" enquanto que para dados quantitativos discretos ou qualitativos ordinais, o significado de mediana deve ser adaptado e dizer "pelo menos 50% das observações são inferiores à mediana" (p. 54). Esta adaptação surge do facto de, na coleção de dados, poderem ocorrer observações repetidas iguais ao valor da mediana. Também Pestana e Gageiro (2008) define a mediana de acordo com a fórmula (1) mas distingue duas situações ao interpretá-la consoante existam observações diferentes ou iguais. Concretamente, aqueles autores referem que sendo "as observações diferentes entre si, a mediana define-se como o valor da sucessão [da amostra ordenada] que tem tantas observações inferiores ou superiores a ela. Caso existam observações repetidas, a mediana define-se como o valor da sucessão que tem tantas observações à sua esquerda como à sua direita" (p. 72). Mais ainda, chamam a atenção para uma particularidade da fórmula (1), referindo que "se n for par, a mediana é indeterminada, podendo ser qualquer valor entre" $x_{(\frac{n}{2})}$ e $x_{(\frac{n}{2}+1)}$, acrescentando que, nesse caso, "convenciona-se tomar para mediana a média aritmética simples dos dois valores centrais" (p. 72). Consultando o programa das Metas

Curriculares (MEC, 2013) verificámos que um dos descritores, no domínio Organização e Tratamento de Dados para o 7.º ano, solicita a necessidade do estudante "reconhecer, considerado um conjunto de dados numéricos, que pelo menos metade dos dados têm valores não superiores à mediana" (p. 61).

É um facto curioso que, a nível da interpretação da mediana, urge a necessidade de algum esclarecimento adicional perante tanta diversidade! Esta multiplicidade de formas, não equivalentes, de interpretar a mediana empírica demonstra a necessidade do cuidado a ter quando pretendemos verbalmente descrevê-la no contexto do problema associado aos dados.

Dado o carácter finito inerente a uma coleção de dados numéricos, sejam eles provenientes de uma variável de natureza contínua ou discreta, podemos socorrer-nos do conceito probabilístico de mediana para obter uma condição universal para a interpretação de mediana de uma coleção finita de dados. Fórmulas diferentes (mas equivalentes) podem ser encontradas para definir a mediana de uma variável aleatória X. Por exemplo, podemos definir mediana de X como qualquer número m tal

$$P(X \leq m) \geq \frac{1}{2} \quad \text{e , simultaneamente,} \quad P(X \geq m) \geq \frac{1}{2} \qquad (2)$$

Assim, transpondo as condições (2) para o contexto empírico, transformando as probabilidades em frequências relativas, diremos que *m é mediana de uma coleção de dados se e somente se pelo menos 50% dos dados são inferiores ou iguais a m e pelo menos 50% são superiores ou iguais a m*.

Podemos, todavia, simplificar este texto na interpretação de mediana definida pela conjunção de duas condições? Será que poderemos referir que *exatamente, aproximadamente* ou *pelo menos* 50% das observações ordenadas são inferiores ou superiores à mediana? Será suficiente referir apenas uma parte da distribuição dos dados (inferiores ou superiores à mediana), ou devem ser referidas ambas? Será que deve ser incluído o valor da mediana nessa parte uni ou bilateral de referência?

EXEMPLOS

O cálculo da mediana, de acordo com a fórmula (1), depende se o número total de dados na coleção é par ou é ímpar. Mais ainda, os dois ramos da fórmula (1) podem ser analisados segundo três situações distintas. Concretamente, a mediana de uma coleção de dados pode corresponder a: um valor não observado na coleção, um valor observado com frequência absoluta unitária na coleção de dados, ou um valor observado que se repete na coleção, ou seja, com frequência absoluta superior a 1. Assim, das três situações, tendo em conta a paridade de n e a frequência da mediana, podem resultar quatro possíveis casos distintos, assinaladas na Tabela 1, no cálculo da mediana de uma coleção finita de dados numéricos. No caso de n ser par, o valor da mediana não poderá ser uma observação única na coleção de dados e, no caso de n ser ímpar, a mediana será sempre um valor observado na coleção de dados.

Tabela 1. Situações possíveis no cálculo da mediana empírica

Frequência da mediana numa coleção de dados	Tamanho da coleção	
	é par	é ímpar
0	Caso 1	Impossível
1	Impossível	Caso 2
> 1	Caso 3	Caso 4

Pretendemos discutir a interpretação do valor da mediana face a cada um dos quatro casos possíveis especificados na tabela 1. Se m é o valor da mediana de uma coleção de dados, então que condições podemos mencionar que conduzem a uma interpretação correta da mediana? Com base nas diferentes formas que encontramos, em livros e manuais, de comentar o valor da mediana de um conjunto (finito) de observações, estabelecemos uma lista de 12 condições (simples), tão exaustiva quanto possível, associadas à interpretação de mediana:

i. Exatamente 50% dos valores da coleção de dados são inferiores a a;
ii. Exatamente 50% dos valores da coleção de dados são superiores a a;
iii. Exatamente 50% dos valores da coleção de dados são inferiores ou iguais a a;
iv. Exatamente 50% dos valores da coleção de dados são superiores ou iguais a a;
v. Aproximadamente 50% dos valores da coleção de dados são inferiores a a;
vi. Aproximadamente 50% dos valores da coleção de dados são superiores a a;
vii. Aproximadamente 50% dos valores da coleção de dados são inferiores ou iguais a a;
viii. Aproximadamente 50% dos valores da coleção de dados são superiores ou iguais a a;
ix. Pelo menos 50% dos valores da coleção de dados são inferiores a a;
x. Pelo menos 50% dos valores da coleção de dados são superiores a a;
xi. Pelo menos 50% dos valores da coleção de dados são inferiores ou iguais a a;
xii. Pelo menos 50% dos valores da coleção de dados são superiores ou iguais a a.

Sem perda de generalidade, tomámos quatro exemplos concretos, um para cada um dos quatro casos apontados na tabela 1. Na tabela 2 encontram-se os exemplos considerados e a indicação

do valor lógico de cada uma das 12 condições acimas listadas (I,…, XII) em cada exemplo. Exemplo 1 (Caso 1: o valor atribuído à mediana corresponde a um valor não observado na coleção de dados), Exemplo 2 (Caso 2: o valor atribuído à mediana corresponde a um valor com frequência unitária na coleção de dados), Exemplo 3 (Caso 3: o valor atribuído à mediana corresponde a um valor que se repete na coleção de dados de dimensão par) e Exemplo 4 (Caso 4: o valor atribuído à mediana corresponde a um valor que se repete na coleção de dados de dimensão ímpar). Em cada exemplo apresenta-se uma coleção de observações, uma tabela com as frequências relativas ($f(x_i)$) e as frequências relativas acumuladas ($F(x_i) = \sum_{x \leq x_i} f(x)$) associadas a cada observação distinta (x_i).

Na nossa opinião, a aparente falta de rigor nas várias definições de mediana encontradas nos livros, correspondente ao procedimento prático de procurar a mediana pelo valor posicionado ao meio na amostra ordenada, referindo-se a haver *50% de observações inferiores* ou *50% de observações superiores* à mediana, resulta do recurso aos vocábulos "inferiores a" ou "superiores a". Estes vocábulos levam intuitivamente à representação das observações na reta ordenada concentrando, consequentemente, observações repetidas num mesmo ponto, e diluindo-se a noção de volume de dados que está efetivamente associada à definição de mediana (50% das observações estão à direita da mediana e 50% estão à esquerda da mediana na coleção ordenada dos dados). Ainda da tabela 2 verifica-se que apenas uma das condições, XI ou XII, não é suficiente para identificar corretamente o valor da mediana. Efetivamente, dizer *pelo menos 50% dos valores da coleção de dados são superiores ou iguais a 3* (condição XII) não implica que 3 seja a mediana da coleção de dados. Vários contraexemplos de coleções de dados podem ser considerados para ilustrar tal situação. Por exemplo, para a coleção: 2, 3, 4, 4, 4, 4, 4, 4, tem-se que mais de 50% dos dados são não inferiores a 3 e, no entanto, a mediana não é 3.

Notemos, também, que o exemplo 1 da tabela 2 ilustra a situação da mediana ser indeterminada como mencionado atrás. Na realidade, no caso da frequência da mediana ser nula (caso 1), e denotando por x_i^- o valor observado na coleção de dados tal que $F(x_i^-) = 0.5$ e por x_i^+ o menor valor tal que $F(x_i^+) > 0.5$, verifica-se que qualquer valor entre x_i^- e x_i^+ pode ser tomado como mediana da coleção; a fórmula (1) convenciona tomar o ponto médio do intervalo $[\, x_i^-, x_i^+ \,]$. No exemplo 1, tem-se $x_i^- = 2$, $x_i^+ = 4$.

CONCLUSÃO

A linguagem associada à interpretação de mediana pode ser mais ou menos simplista. Podemos substituir, de forma equivalente, a expressão "inferior ou igual" por "não superior" (similar para "superior ou igual"). Contudo, pretendendo interpretar a mediana de um conjunto de observações, de uma forma completa, teremos que mencionar a conjunção de duas condições e salvaguardar a possibilidade de repetições de observações, ou seja, *m* é mediana de uma colecção de dados se e somente se pelo menos 50% dos dados são inferiores ou iguais a *m* e pelo menos 50% são superiores ou iguais a *m*

Se pretendemos tirar alguma ilação, no contexto do problema dado, do facto do valor *m* ser a mediana de uma coleção de dados, basta usar qualquer condição subsequente resultante daquela conjunção, como é proposto nas (atuais) Metas Curriculares: se *m* é mediana de uma coleção de dados, então pelo menos 50% dos dados são inferiores ou iguais a *m.* Mas tenhamos consciência de que, omitindo uma ou parte das duas condições XI e XII na interpretação da mediana de uma coleção de dados, estaremos a mostrar apenas uma perspetiva do contexto dos dados!

Agradecimentos. Trabalho subsidiado por fundos portugueses através do CIDMA (Centro de Investigação e Desenvolvimento em Matemática e Aplicações) da Universidade de Aveiro e FCT (Fundação para a Ciência e a Tecnologia), dentro do projeto UID/MAT/04106/2013.

Referências Bibliográficas

Amaro, A., Silvestre, C. Fernandes, L. (2009) *Estatística Descritiva. O segredo dos dados.* Lulu.com. Lisboa.

Guimarães, R.C., Cabral, J.A.S. (1997) *Estatística*. McGraw Hill. Portugal.

Hall, A., Neves, C. Pereira, A. (2011) *Grande maratona de Estatística no SPSS*, Escolar Editora. Lisboa.

Ministério da Educação e Ciência (2013). Programa e Metas Curriculares de Matemática do ensino básico. Lisboa: Ministério da Educação, DGIDC.

Mood, A.M., Graybill, F.A., Boes,D.C. (1974) *Introduction to the theory of Statistics*. 3rd Edition. McGraw-Hill. Lisbon.

Murteira, B., Antunes, M. (2012) Probabilidades e Estatística, Vol. I, Escolar Editora. Lisboa.

Murteira, B., Ribeiro, C.S., Andrade e Silva, J., Pimenta, C. (2010) *Introdução à Estatística*. Escolar Editora. Lisboa.

Pestana, D. D., Velosa, S. (2008) *Introdução à Probabilidade e à Estatística*. 3.ª Edição, Vol I, Fundação Calouste Gulbenkian. Lisboa.

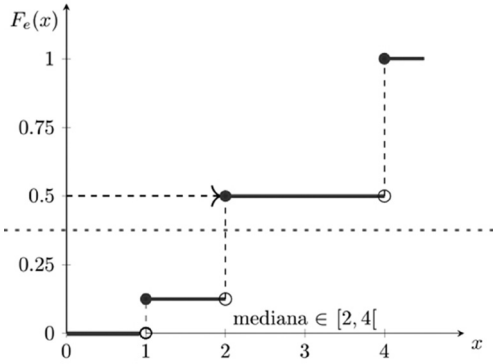
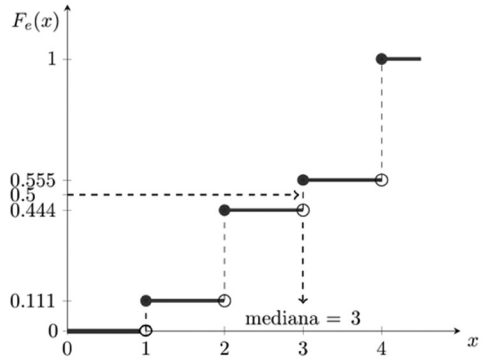
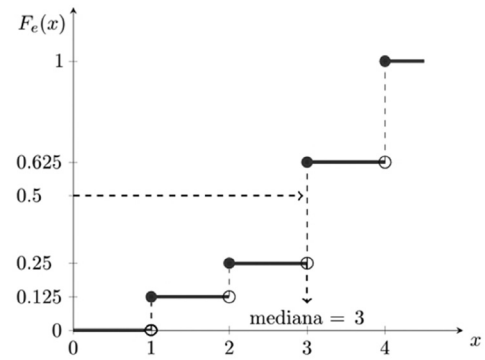
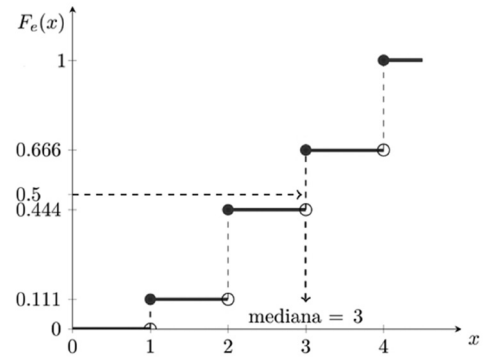
Pestana, M. H., Gageiro, J. N. (2008). Análise de Dados para Ciências Sociais – a complementaridade do SPSS, Edições Silabo. Lisboa.

Adelaide Freitas, João Pedro Cruz, Nélia Silva

Departamento de Matemática e CIDMA

Universidade de Aveiro

Tabela 2. Coleções de dados, não agrupados, em 4 situações distintas no cálculo da mediana. Por aplicação da fórmula (1), a mediana é igual a 3 nos quatros exemplos.

<table>
<tr><th rowspan="2">Situação</th><th rowspan="2">Coleção de dados</th><th rowspan="2">Gráfico da função $F(x_i)$</th><th colspan="2">Condições</th></tr>
<tr><th>verdadeiras</th><th>falsas</th></tr>
<tr><td>Caso 1
(n par)</td><td>Exemplo 1:
1, 2, 2, 2, 4, 4, 4, 4
<table><tr><th>x_i</th><th>$f(x_i)$</th><th>$F(x_i)$</th></tr><tr><td>1</td><td>0.125</td><td>0.125</td></tr><tr><td>2</td><td>0.375</td><td>0.500</td></tr><tr><td>4</td><td>0.500</td><td>1.000</td></tr></table></td><td> 
</td><td>I, II, III,
IV, V, VI,
VII, VIII, IX,
X, XI, XII</td><td>----</td></tr>
<tr><td>Caso 2
(n ímpar)</td><td>Exemplo 2:
1, 2, 2, 2, 3, 4, 4, 4, 4
<table><tr><th>x_i</th><th>$f(x_i)$</th><th>$F(x_i)$</th></tr><tr><td>1</td><td>0.111</td><td>0.111</td></tr><tr><td>2</td><td>0.333</td><td>0.444</td></tr><tr><td>3</td><td>0.111</td><td>0.555</td></tr><tr><td>4</td><td>0.444</td><td>1.00</td></tr></table></td><td> 
</td><td>V, VI,
VII, VIII,
XI, XII</td><td>I, II, III,
IV,
IX,
X</td></tr>
<tr><td>Caso 3
(n par)</td><td>Exemplo 3:
1, 2, 3, 3, 3, 4, 4, 4
<table><tr><th>x_i</th><th>$f(x_i)$</th><th>$F(x_i)$</th></tr><tr><td>1</td><td>0.125</td><td>0.125</td></tr><tr><td>2</td><td>0.125</td><td>0.250</td></tr><tr><td>3</td><td>0.375</td><td>0.625</td></tr><tr><td>4</td><td>0.375</td><td>1.000</td></tr></table></td><td> 
</td><td>VII,
XI, XII</td><td>I, II, III,
IV, V, VI,
VIII, IX,
X</td></tr>
<tr><td>Caso 3
(n ímpar)</td><td>Exemplo 4:
1, 2, 2, 2, 3, 3, 4, 4, 4
<table><tr><th>x_i</th><th>$f(x_i)$</th><th>$F(x_i)$</th></tr><tr><td>1</td><td>0.111</td><td>0.111</td></tr><tr><td>2</td><td>0.333</td><td>0.444</td></tr><tr><td>3</td><td>0.222</td><td>0.666</td></tr><tr><td>4</td><td>0.333</td><td>1.00</td></tr></table></td><td> 
</td><td>V, VI,
VIII,
XI, XII</td><td>I, II, III,
IV,
VII, IX,
X</td></tr>
</table>