Complexities on Median Calculation

M A Jalil* and Jamil H Karami

Department of Statistics, University of Dhaka, Bangladesh

Submission: February 8, 2018; **Published:** May 05, 2018

*Corresponding author: M A Jalil, Department of Statistics, University of Dhaka, Dhaka-1000; Bangladesh, Tel: (+88)01974799788;
Fax: 880-2-9667222; email: majalil@du.ac.bd

Abstract

The complexity of median calculation is explored with some examples in this article. We have pointed out how to identify ungrouped and grouped data, and thereby the appropriate method of computing median and other percentiles.

Keywords: Ungrouped data; Grouped data; Median; Percentile; Location parameter; Ungrouped; Frequency; Evaluating median; Sample data; Values; Distribution; Continuous scale; Repeated values; Domain; Horizontal line; Equal halves; Lower boundary; Cumulative frequency; Statistical packages; Economic status; Central tendency

Introduction

Location is an important measure in describing a sample data for a variable. For a discrete or ordinal data we can calculate percentiles for location parameter. Among these location parameters, the median is the most important one in describing a variable, as it informs the central position of the distribution underlying variable. In this article, we show the importance of identifying whether a data set is ungrouped or grouped. Although some articles have questioned about the usefulness of grouped data [1], we think the existence of grouped (frequency) data is very frequent in real-world examples compared to ungrouped data. In this research, we have suggested an appropriate method of computing median and other percentiles. The concept has been illustrated with several examples.

Methods

Several methods can be used for evaluating median from sample data [2]. It is crucial to decide which of them would be appropriate for the problem at hand. This difficulty can be resolved if we simply think of the nature of the data for a particular problem. By the `nature' we mean whether a data set is grouped or ungrouped. If we have raw data (observed) of a continuous variable with no repeated observations (i.e., single frequency for each case), the data is called ungrouped data. Suppose, three values of grades obtained by a student are 2.8, 2.9 and 3.3; it is an example of ungrouped data. Therefore, we can compute median using the formula for ungrouped data. However, in real world examples, the ungrouped data is not frequently encountered. It is hardly possible to get scope of work with this type of data. On the other hand, if there are any repeated observations in the data set, it is called a grouped data. If the three values of grades were 2.8, 2.8 and 3.3, it can be considered as grouped data. In fact, practically observed

data obtained from any field of enquiry or experiment for a continuous variable is mostly grouped data, frequency of one or more observations are present.

Median is the middlemost location of a distribution or of a set of values for a variable with respect to number of observations. To find the median, we locate the center point keeping 50% observations left to the point and 50% observations to the right of a distribution. The value corresponding to this center point is the median. Percentiles are measures of location and median is the 50th percentile. Median lies on a continuous scale corresponding the domain of the variable under consideration. It is interesting to note that whatever be the type of a variable, median is always continuous. For an observed set of values we can scale a horizontal line with the values and locate the point where it divides into two equal halves with respect to number of observations. The value on the scale of the located point is the median.

Results and Discussion

For an ungrouped raw data the median is calculated as: $x_1, x_2, \ldots, x_n$ (ordered). If number of observations (n) is odd, the sample median is the $\frac{n+1}{2}^{th}$ observation. But if there are repeated values in the central area of the distribution (of underlying variable), then $\frac{n+1}{2}^{th}$ observation may not be the median. We illustrate this with an example [3].

Example 1

Let us consider some observations: 10, 12, 13, 14, 15, 16, 17, 17, 17, 17, 17, 17, 17, 18, 19, 19, 20. Clearly, $\frac{n+1}{2} = \frac{17+1}{2} = 9^{th}$ observation = 17, which many people will take as median. But, if we notice carefully it can be seen that 7th, 8th, 10th, 11th, 12th and 13th observations are also 17. Therefore, the value actually does not

divide the whole distribution into two equal halves. Still then, how can we accept it as median?

We have mentioned in the methods section that this data set is actually a grouped data. It would be appropriate to use the formula prescribed for the grouped data. Remember the formula for computing th percentile:

$$p_k = L_k + \frac{\frac{nk}{100} - F_{-k}}{f_k} * C_w,$$

where, L_k is lower boundary of median class, F_{-k} is cumulative frequency of pre k^{th} percentile class, f_k is frequency of kth percentile class, and C_w is k^{th} percentile class width. Here, L_k is used for a grouped data. The data in this example is represented in Table 1. From this table, the median class is to be considered $16.5 - 17.5$.

Table 1: Frequency Table of the data.

Values	Frequency	Cumulative Frequency
10	1	1
12	1	2
13	1	3
14	1	4
15	1	5
16	1	6
17	7	13
18	1	14
19	2	16
20	1	17

Using the formula mentioned above, we have $p_{50} = 16.7$. However, for this same problem, p_{50} will be determined as 17 in many books and statistical packages as discussed earlier. The reason behind this is that they consider this type of data set as ungrouped data, which is not appropriate according to our definition of grouped and ungrouped data. It is interesting to note that if someone considers the data as ungrouped, then $p_{40}, p_{41}, \cdots, p_{75}$ are all equal. That is, in this case, all percentiles from 40th to 75th are equal to 17. This contradicts the concept of percentile. On the other hand, if we consider the data set as grouped data (frequency data), we have the following percentiles:

$$p_{40} = 16.5 + \frac{\frac{17*40}{100} - 6}{13} * 1 = 16.56$$

$$p_{50} = 16.5 + \frac{\frac{17*50}{100} - 6}{13} * 1 = 16.67$$

$$p_{70} = 16.5 + \frac{\frac{17*70}{100} - 6}{13} * 1 = 16.95$$

$$p_{75} = 16.5 + \frac{\frac{17*75}{100} - 6}{13} * 1 = 17.02$$

It is remarkable that the above percentile calculations seem rather appropriate.

Example 2

Suppose we have the following data on economic status (1: high class, 2: high-middle class, 3: middle class, 4: lower-middle class and 5: lower class):

1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 5.

We show how the suggested method can be used to describe the economic status with the measures of central tendency and dispersion. Using the method, it is easy to find that the central tendency (median) is 2.51, and dispersion (quartile deviation) is 0.72.

Example 3

Suppose we are given test scores of 100 students as in Table 2. This type of examples are available in Ross [4]. Using the suggested method, we have found the following percentiles: Median = 50th percentile = 5.625; 60th percentile = 5.875; 70th percentile = 6.125; 80th percentile = 6.375. However, if these percentile were calculated otherwise (i.e., considering this data set as ungrouped), then 50th to 80th percentiles will all be equal to 6, which is not acceptable at all.

Table 2: Test scores of students.

Score	0	2	4	5	6	7	8
No. of students	5	2	7	31	40	12	3

Conclusion

It is very rare to find an 'ungrouped data' where each of the values will occur only once. We can assume an ungrouped data theoretically, but practically it is rare. In contrary, grouped data is very common. We recommend the formula for grouped data in the calculation of median and other percentiles. The formula for grouped data is appropriate with all of the examples we mentioned here and beyond.

Acknowledgement

References

1. Freund EJ, Perles MB (1987) A new Look at Quartiles of Ungrouped Data. The American Statistician 41(3): 200-203.

2. Hyndman JR, Fan Y (1996) Sample Quantiles in statistical Packages. The American Statistician 50(4): 361-365.

3. Downie NM, Heath RW (1970) Basic Statistical Methods. Harper International, USA.

4. Ross MS (2010) Introductory Statistics (3rd edn). Academic Press, USA.

<table>
<tr><td>

**Your next submission with Juniper Publishers
will reach you the below assets**

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats
  (Pdf, E-pub, Full Text, Audio)
- Unceasing customer service

Track the below URL for one-step submission
https://juniperpublishers.com/online-submission.php

</td></tr>
</table>