Journal of Contemporary Educational Research

Research Article

BIO-BYWORD SCIENTIFIC PUBLISHING PTY LTD

# Discussion about the Mean, Median, Mode and their Validity, and the Representative Number

Weiwen Cao

Yinchuan Elite School, Yinchuan 750002, Ningxia Hui Autonomous Region, China

**Abstract:** The mean, median and mode are statistical means often used in statistics to describe the concentrative trend of a group of data. However, the three kinds of data are not only different in definition, but also different in application. For the same group of data, there is no definite and strict standard for which one should be used. What's more, different data may even lead to diametrically opposite conclusions. Even if we finally choose one, what extent the data can describe the concentrative trend of the group of data? Is there any statistical method that can describe the concentrative trend of a group of data more objectively and fairly? It is not only a very interesting issue, but also an issue including great research value.

**Keywords:** Mean; Median; Mode; Validity; Representative Number

***Publication date:*** March, 2021
***Publication online:*** 31 March, 2021
****Corresponding author:*** Weiwen Cao, caoweiwen 2008@126.com

## 1 Definition and properties of mean, median and mode

Mean: In general, for $n$ numbers: $x_1$, $x_2$,$\cdots$ $x_n$, we call $\frac{1}{n}(x_1+x_2+\cdots+x_n)$. that the mean of the $n$ data. The mean can make full use of the information provided by all data, so it is more commonly used in real life, but it is easily affected by extreme data[1].

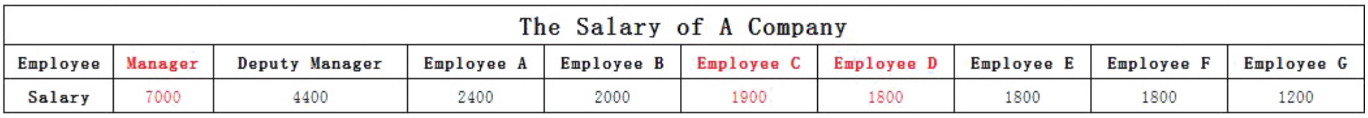
Median: Generally, if the data are arranged in order of size, and the data in the middle (or the mean of the two data in the middle) is called the median of the group of data. The advantage of median is that it is easy to calculate and less affected by extreme data, but it can't make full use of all data information.

Mode: The data that appears most frequently in a group of data is called mode. However, when the repeat times of each data is approximately equal, the mode often has no special significance[2].

## 2 Generation of the representative number

Although the definition and nature of mean, median and mode are very clear, for a group of large or small, long or short, uneven or neat data, which one should be used, or is there any statistical method to describe the concentrative trend of a group of data more objectively and fairly?

However, we don't have certain definite and strict standards to judge which data should be used. Even if we choose one of them through inexplicable, uncertain sensibility or rationality, what extent the kind of data can effectively describe the group of data, namely, what is its validity? Now let's study a common case in our life, so as to understand and think about the above series of problems more deeply.



The Salary of A Company

| Employee | Manager | Deputy Manager | Employee A | Employee B | Employee C | Employee D | Employee E | Employee F | Employee G |
|---|---|---|---|---|---|---|---|---|---|
| Salary | 7000 | 4400 | 2400 | 2000 | 1900 | 1800 | 1800 | 1800 | 1200 |

For the group of data, manager, employee C and employee D respectively describe the salary of company from three different perspectives: mean, median and mode. No matter for a college student who has just graduated, or for a middle-aged person with greater pressure, it is better to earn more 100 yuan per month than less 100 yuan. So which kind of salary the job seeker should use as a reference standard? How much is the final salary? If the job seeker is not faced with three kinds of data, but with three employees: manager, employee C and employee D, the choice will become more difficult after some conversation,

We could see how difficult it is to select a representative data describing the concentrative trend of a group of data. This is because the three kinds of data describe the concentrative trend of a group of data from three different angles, so it is too one-sided to choose any one of them. At the same time, the validity of three kinds of data for different sequences will also have some differences, sometimes the difference is even very huge. For example, the mean is obviously not suitable to describe the sequence with more extreme data, and sometimes an extreme data will seriously affect the validity of mean. For the median, although the amount of data on both sides is the same, if the data on both sides are unbalanced, the validity of median will be greatly reduced. As for mode, unless the frequency of certain data is obviously dominant, the reference value will not be too significant. To sum up, in order to describe the concentrative trend of a group of data more objectively and fairly, combining the characteristics of the three kinds of data, organically forming a more representative data——the representative number, is the best and final way to solve the above complex problems.

So how should we "combine" the three kinds of data? Since the mean, median and mode are three necessary elements to form the representative number, the greater the validity, the greater proportion of the data, and vice versa. We use $P$ to represent mean, $V_{\mathrm{P}}$ to represent mean validity; We use $N$ to represent median, $V_{\mathrm{N}}$ to represent median validity; We use $M$ to represent mode, $V_{\mathrm{M}}$ to represent mode validity. We use $E$ to represent the representative number. According to the above analysis, we define the representative number as follows:

$$E=\frac{V_P P}{V_P+V_N+V_M}+\frac{V_N N}{V_P+V_N+V_M}+\frac{V_M M}{V_P+V_N+V_M}$$

## 3 Validity of mean, median and mode

### 3.1 Mean validity

Although we have solved the "weighted" problem of mean, median and mode validity, how to determine the validity of the three kinds of data? We know that mean validity is negatively correlated with the number of extreme data and extreme degree in the sequence. We usually use range ($R$) and variance ($S^2$) or standard deviation ($S$) to describe the fluctuation of a group of data. Therefore, the larger the range, variance or standard deviation of the group of data, the lower the validity of mean, and vice versa.. However, the size of range, variance or standard deviation cannot be determined by the group of data itself——this method is generally used to compare the fluctuation between different sequences. Since the standard deviation represents the average distance between each data and the mean, and the range represents the distance between the maximum data ($a$) and the minimum data ($b$), we can define the mean validity as follows: If $a \neq b$ ,

$$V_P = 1-\frac{s}{\frac{R}{2}} = 1-\frac{2s}{R}$$

; If $a=b$, $V_{\mathrm{P}}=1$; In summary,

$$V_P = \begin{cases} 1-\frac{2s}{R}, & (a \neq b) \\ 1, & (a = b) \end{cases}$$

### 3.2 Median validity

For a group of data, the ideal median should not only ensure that the amount of data on both sides is equal, but also ensure that the distribution of data on both sides is uniform and symmetrical. Therefore, it should be exactly equal to the intermediate value between the maximum data and the minimum data, so the closer the median is to the intermediate value, the higher its validity, and vice versa. Therefore, we can define the median validity as follows: If $a \neq b$ ,

$$,V_N = 1-\frac{\left|N-\frac{a+b}{2}\right|}{\frac{R}{2}} = 1-\frac{|2N-a-b|}{R} \quad ;$$

If $a=b$, $V_N=1$; In summary,

$$V_N = \begin{cases} 1 - \dfrac{|2N - a - b|}{R}, & (a \neq b) \\ 1, & (a = b) \end{cases}$$

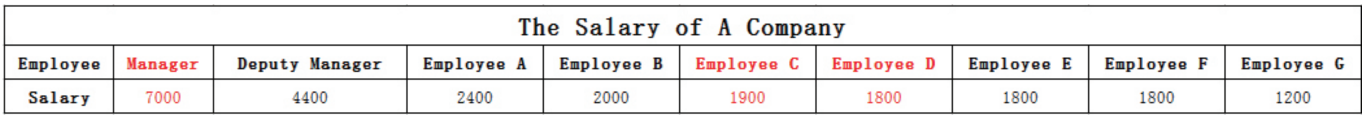
### 3.3 Mode validity

For mode, the more frequent it appears, the better it can reflect the trend of data concentration, and the higher its validity, and vice versa. Therefore, we can define mode validity as follows: $V_M = \frac{t}{n}$. But for mode validity, there are two problems as follows: (1) There may be multiple modes at the same time, so which data should be selected as the mode? In order to better describe the concentrative trend of a group of data, we should choose the data that is closest to the mean or median with higher validity as the mode; (2) No matter the mean or median is chosen as the standard of mode selection, sometimes there are two modes with the same distance from this standard, which one should we choose? In order to highlight the imbalance and disproportion of data distribution, we choose the smaller data as the mode.

## 4 Application case of the representative number

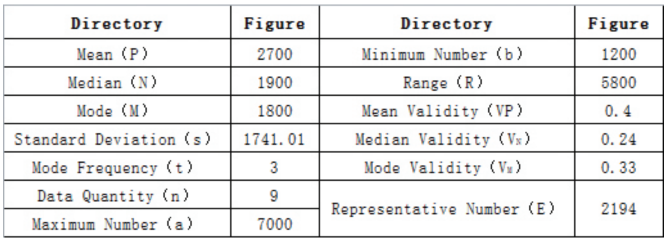
### 4.1 Case 1:

After determining the calculation method of mean, median and mode validity, we can also determine the calculation method of the representative number. Let's review the case mentioned above and feel the charm of the representative number together.



| The Salary of A Company | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Employee | Manager | Deputy Manager | Employee A | Employee B | Employee C | Employee D | Employee E | Employee F | Employee G |
| Salary | 7000 | 4400 | 2400 | 2000 | 1900 | 1800 | 1800 | 1800 | 1200 |

According to the above table, we can get the following table:



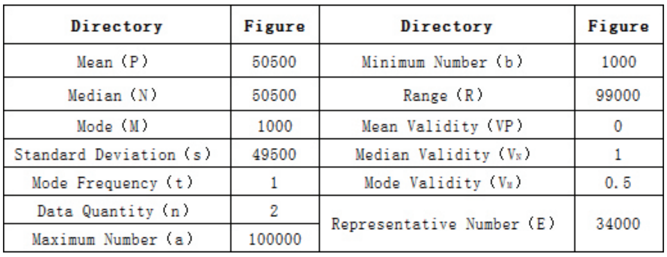
| Directory | Figure | Directory | Figure |
|---|---|---|---|
| Mean (P) | 2700 | Minimum Number (b) | 1200 |
| Median (N) | 1900 | Range (R) | 5800 |
| Mode (M) | 1800 | Mean Validity (VP) | 0.4 |
| Standard Deviation (s) | 1741.01 | Median Validity (V_N) | 0.24 |
| Mode Frequency (t) | 3 | Mode Validity (V_M) | 0.33 |
| Data Quantity (n) | 9 | Representative Number (E) | 2194 |
| Maximum Number (a) | 7000 | | |

It can be seen from the table below that the salary standard that the job seeker should refer to is  yuan, rather than the average salary  yuan that is most commonly used in life.

4.2 Case 2:

Master and apprentice jointly run a shop. The salary of master is  yuan, but the salary of apprentice is yuan. What is the salary of the shop?

According to the meaning of title, we can get the following table:

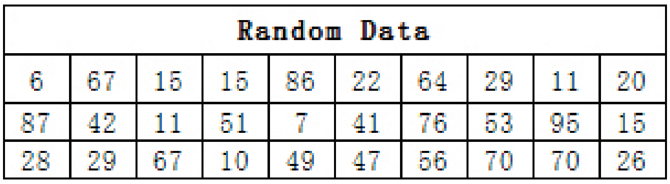


| Directory | Figure | Directory | Figure |
|---|---|---|---|
| Mean (P) | 50500 | Minimum Number (b) | 1000 |
| Median (N) | 50500 | Range (R) | 99000 |
| Mode (M) | 1000 | Mean Validity (VP) | 0 |
| Standard Deviation (s) | 49500 | Median Validity (V_N) | 1 |
| Mode Frequency (t) | 1 | Mode Validity (V_M) | 0.5 |
| Data Quantity (n) | 2 | Representative Number (E) | 34000 |
| Maximum Number (a) | 100000 | | |

As can be seen from the table below, the reference salary in the table is  yuan, but the average salary is  yuan. It can be seen that the mean validity is very low when the sequence fluctuates greatly, so it can't effectively describe the concentrative trend of the group of data.

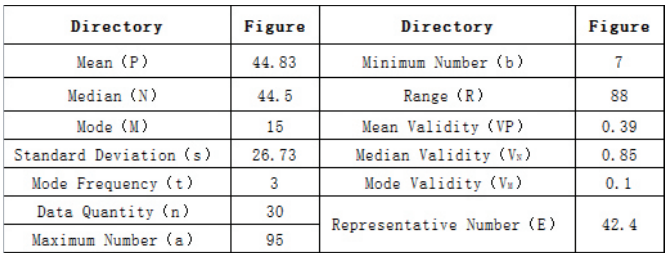
### 4.3 Case 3:

The random number generator generates  data between ~ as follows, to calculate the representative number of the group of data please:



| Random Data | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 67 | 15 | 15 | 86 | 22 | 64 | 29 | 11 | 20 |
| 87 | 42 | 11 | 51 | 7 | 41 | 76 | 53 | 95 | 15 |
| 28 | 29 | 67 | 10 | 49 | 47 | 56 | 70 | 70 | 26 |

According to the above table, we could get the following table:



| Directory | Figure | Directory | Figure |
|---|---|---|---|
| Mean (P) | 44.83 | Minimum Number (b) | 7 |
| Median (N) | 44.5 | Range (R) | 88 |
| Mode (M) | 15 | Mean Validity (VP) | 0.39 |
| Standard Deviation (s) | 26.73 | Median Validity ($V_N$) | 0.85 |
| Mode Frequency (t) | 3 | Mode Validity ($V_M$) | 0.1 |
| Data Quantity (n) | 30 | Representative Number (E) | 42.4 |
| Maximum Number (a) | 95 | | |

It can be seen from the following table that with the increase of sequence, the representative number, mean and median would gradually be similar. It is because with the increase of sequence, while the mode validity is decreasing, the fluctuation of sequence tends to be smooth, and the distribution of data tends to be uniform and symmetrical, which leads to gradual improvement of the validity of mean and median.

In fact, we could not only use the representative number to describe the concentrative trend of a group of data, but also compare the quality of different sequences through the representative number, and adjust the distribution of data according to the representative number. For example, schools, enterprises or militaries could effectively allocate staffs and works according to the representative number to avoid the performance deviation caused by uneven distribution.

## 5 Conclusion

The representative number is an organic combination of mean, median and mode. Through the way of "weighting", it ends the history of one-sided and blind description of sequences among the three kinds of data, and enables people to describe the concentrative trend of a group of data more objectively and fairly. The generation of the representative number will have a very rich research value for both academic research and practical application.

## References

[1] Compilation group of compulsory education textbook mathematics (grade 7-9) published by Beijing Normal University. Mathematics (Class 8, Volume 1)[M]. Beijing: Beijing Normal University Press, 2016.

[2] Compilation group of compulsory education textbook mathematics (grade 7-9) published by Beijing Normal University. Mathematics (Class 8, Volume 1)[M]. Beijing: Beijing Normal University Press, 2016.

[3] Compilation group of compulsory education textbook mathematics (grade 7-9) published by Beijing Normal University. Mathematics (Class 8, Volume 1)[M]. Beijing: Beijing Normal University Press, 2016.