Parabola Volume 62 Issue 2 (2026), Article 15

Pearson's median approximation and its limitations

Larsson A. Laldingliana[1]

1 Introduction

In statistics, one often aims to find the "centre" of a dataset. For this purpose, three common measures are typically used: the mean, median and mode.

Given the observations $x_1, x_2, \dots, x_n$, the *mean* is defined as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}\,.$$

The *median* is the middle value when the observations are arranged in increasing order, while the *mode* is the value that occurs most frequently.

For perfectly symmetric distributions, these three quantities coincide. However, many real datasets are skewed, so the distribution extends further in one direction than the other.

In the late 19th century, the statistician Karl Pearson observed [1] that, in many moderately skewed unimodal distributions, the mean, median and mode satisfy an approximate relationship. Rearranged, *Pearson's Formula* is the approximation

$$\mathrm{median} \approx \frac{\mathrm{mode} + 2\,\mathrm{mean}}{3}\,.$$

This formula provides a convenient way to estimate the median when the mean and mode are known and direct computation of the median may be time-consuming for large data sets.

The purpose of this article is to examine the reliability of this approximation. We will see that, in symmetric datasets, the formula gives the exact median, while in the presence of sufficiently large outliers, the approximation can become arbitrarily inaccurate.

2 The approximation error $\Delta(X)$

Let, $X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}$ be a finite dataset where $\mathrm{med}(X)$ is the true median, $\mathrm{mo}(X)$ is the mode and $\bar{X}$ is the mean. Pearson's *empirical median* is defined as

$$\mathrm{med}_{\mathrm{emp}}(X) = \frac{\mathrm{mo}(X) + 2\bar{x}}{3}$$

[1] Larsson is a Class 10th Student at Kendriya Vidyalaya Sangathan, Aizawl

and Pearson's Formula is then

$$\operatorname{med}(X) \approx \operatorname{med}_{\mathrm{emp}}(X)\,.$$

To measure how accurate this approximation is, we define the error

$$\Delta(X) = |\operatorname{med}(X) - \operatorname{med}_{\mathrm{emp}}(X)|\,.$$

Here, $\Delta(X)$ represents the absolute difference between the true median and the value predicted by Pearson's empirical formula. The smaller $\Delta(X)$ is, the better the approximation. However, the larger $\Delta(X)$ is, the more inaccurate the approximation becomes.

3 The accuracy

Now let us test the accuracy of Pearson's Formula. We first consider symmetric unimodal datasets X. A dataset is *symmetric about* c if, for every value $c + d$ in the dataset, there is a corresponding value $c - d$. The dataset is *unimodal* if it has a unique mode, that is, only one value occurring with the highest frequency.

Proposition 1. *Let X be a symmetric unimodal dataset centred at c. Then*

$$\bar{x} = \operatorname{med}(X) = \operatorname{mo}(X) = c\,.$$

Consequently, $\Delta(X) = 0$, and Pearson's Formula is strictly accurate:

$$\operatorname{med}_{\mathrm{emp}}(X) = \frac{\operatorname{mo}(X) + 2\bar{x}}{3}\,.$$

Proof. The data set is symmetric about c, so for each value $c+d$ there is a corresponding value $c-d$. The mean of thee two values is then $\frac{1}{2}\big((c+d)+(c-d)\big) = c$. Summing over all values, the mean of the data set is c. By symmetry, exactly half of the observations lie on each side of c, so the median is also c. Since the dataset is unimodal with centre c, the mode must also be c. Putting these values into Pearson's empirical formula yields

$$\operatorname{med}_{\mathrm{emp}}(X) = \frac{c + 2c}{3} = c\,.$$

Hence, the empirical median coincides with the true median, and $\Delta(X) = 0\,.$ □

This proposition states that Pearson's Formula is exact for symmetric unimodal datasets. More generally, it explains why Pearson's Formula performs well for distributions that are close to symmetric. This is not true for less symmetric distributions, as the following proposition shows.

Proposition 2. *Let $X_N = \{a, a, \ldots, a, N\}$ be a dataset in which the value a appears $k \geq 2$ times. Then the approximation error $\Delta(X_N) = |\operatorname{med}(X_N) - \operatorname{med}_{\mathrm{emp}}(X_N)|$ satisfies*

$$\lim_{N\to\infty} \Delta(X_N) = \infty\,.$$

Proof. Since the value a appears at least twice, the middle position of the ordered dataset is occupied by a, and s is the most-frequently occuring value. Hence,

$$\mathrm{med}(X_N) = \mathrm{mo}(X_N) = a\,.$$

The mean of the dataset is

$$\bar{x} = \frac{ka+N}{k+1}\,,$$

so $\Delta(X_N) = |\mathrm{med}(X_N) - \mathrm{med}_{\mathrm{emp}}(X_N)|$ equals

$$\left|a - \frac{\mathrm{mo}(X)+2\bar{x}}{3}\right| = \left|a - \frac{a+2\frac{ka+N}{k+1}}{3}\right| = \frac{2}{3}\left|\Big(1-\frac{ka}{k+1}\Big)a - \frac{N}{k+1}\right|.$$

Therefore, $\Delta(X_N) \to \infty$ as $N \to \infty$. □

Proposition 3. *Let $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}$ be a finite dataset and let $c \in \mathbb{R}$. Define the* translated dataset

$$X + c = \{x_1 + c, x_2 + c, \ldots, x_n + c\}.$$

Then $\Delta(X + c) = \Delta(X)$.

Proof. Adding a constant c to every observation shifts the central measures by c: $\bar{x}_{X+c} = \bar{x}_X + c$, $\mathrm{med}(X+c) = \mathrm{med}(X) + c$ and $\mathrm{mo}(X+c) = \mathrm{mo}(X) + c$. Therefore,

$$\mathrm{med}_{\mathrm{emp}}(X+c) = \frac{\mathrm{mo}(X+c)+2\bar{x}_{X+c}}{3} = \frac{\mathrm{mo}(X)+c+2(\bar{x}_X+c)}{3} = \mathrm{med}_{\mathrm{emp}}(X) + c\,,$$

and so

$$\Delta(X+c) = |\mathrm{med}(X+c) - \mathrm{med}_{\mathrm{emp}}(X+c)| = |\mathrm{med}(X) + c - (\mathrm{med}_{\mathrm{emp}}(X)+c)| = \Delta(X)\,.$$

Hence the approximation error is invariant under translation. □

Proposition 4. *Let $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}$ be a finite dataset and let $\lambda \in \mathbb{R}$. Define the* scaled dataset

$$\lambda X = \{\lambda x_1, \lambda x_2, \ldots, \lambda x_n\}.$$

Then $\Delta(\lambda X) = |\lambda|\Delta(X)$.

Proof. Multiplying every observation by λ scales the central measures by c: $\bar{x}_{\lambda X} = \lambda\bar{x}$, $\mathrm{med}(\lambda X) = \lambda \mathrm{med}(X)$ and $\mathrm{mo}(\lambda X) = \lambda \mathrm{mo}(X)$. Therefore,

$$\mathrm{med}_{\mathrm{emp}}(\lambda X) = \frac{\mathrm{mo}(\lambda X)+2\bar{x}_{\lambda X}}{3} = \frac{\lambda \mathrm{mo}(X)+2\lambda\bar{x}}{3} = \lambda\,\mathrm{med}_{\mathrm{emp}}(X)\,,$$

and so

$$\Delta(\lambda X) = |\mathrm{med}(\lambda X) - \mathrm{med}_{\mathrm{emp}}(\lambda X)| = |\lambda \mathrm{med}(X) - \lambda \mathrm{med}_{\mathrm{emp}}(X)| = |\lambda|\Delta(X)\,.$$

Hence, the approximation error scales proportionally with $|\lambda|$. □

4 How the error behaves under transformations

The previous propositions describe how the approximation error behaves under translation and scaling separately. Combining these results yields a simple law describing the effect of an affine transformation on the error.

Theorem 5. *Let* $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}$ *be a finite dataset and let* $\lambda, c \in \mathbb{R}$. *Define the* affine transformation

$$Y = \lambda X + c = \{\lambda x_1 + c, \lambda x_2 + c, \ldots, \lambda x_n + c\}.$$

Then $\Delta(Y) = |\lambda|\Delta(X)$.

Proof. By Propositions 3 and 4, $\Delta(Y) = \Delta(\lambda X + c) = \Delta(\lambda X) = |\lambda|\Delta(X)$. □

5 Conclusion

Pearson's empirical Formula offers a straightforward approximation that connects the mean, median and mode of a dataset. We looked into how accurate this approximation is by creating an error function that shows the difference between the true median and the value that Pearson's formula predicted.

We demonstrated that the approximation attains precision in symmetric contexts, whereas datasets with extreme outliers can yield errors of indeterminate magnitude. We also found that the error acts in a predictable way when affine transformations are applied. In particular, translations don't change the error, but scaling the dataset makes the error bigger by the amount of the scaling factor. The median is more stable than the mean, and it shows how Pearson's relation changes when the data is changed.

Acknowledgements

The author would like to thank Esther Vanlalrinsangi and my Mother for their encouragement and support throughout my work.

References

[1] K. Pearson, X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material, *Philosophical Transactions of the Royal Society, Series A* **186** (1895), 343–414.