A Generalized Z-score for Both Symmetric and Asymmetric Distribution

Mian Arif Shams Adnan
Bowling Green State University, Bowling Green, Ohio, 43402

Abstract
A new form of Z score/test statistic (along with its mathematical form) has been suggested for both symmetric an asymmetric distribution considering simultaneously the appropriate measures of location and dispersion of a distribution. It is relatively less affected by outlier(s).

Key Words: Mean deviation from median.

1. Introduction

The shape and position of a normal distribution curve depend on two parameters, the mean and the standard deviation. Since each normally distributed variable has its own mean and standard deviation, the shape and location of these curves will vary. So, a table of areas under the curve for each variable is required which is known as z table. To simplify this situation, statisticians use standard normal distribution. The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. The standard normal distribution shows the values under the curve indicate the proportion of area in each section.

The formula of a standard normal distribution is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

For example, the area between the mean and 1 standard deviation above or below the mean is about 0.3413, or 34.13%. That's, 68% within 1 sigma, 95% within 2 sigma and 99.7% within 3 sigma. The formula for the standard normal distribution is all normally distributed variables can be transformed into the standard normally distributed variable by using the formula for the standard score:

$$\mathrm{z} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

$$\mathrm{z} = \frac{\mathrm{x}-\mu}{\sigma}.$$

A *Z*-test is a statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples.

2. Types of Z Test

Z test statistic is used for various types of tests like one sample mean, one sample proportion, two samples' equality of means, two samples' equality of proportions, etc.

3. Proposed Z Test

One way of overcoming the problem of outlier(s) and/or extreme value(s) in z test as well as setting the limit lines is measuring the extent of asymmetry of a distribution and to detect and eliminate outlier(s)/extreme values in a symmetric/asymmetric data set is z based limits should be less affected by the outlier and/or extreme value for its proper location and one dimensional dispersion measure.

Adnan *et al* (2007) suggested a new Non-Robust measure for determining skewness along with its mathematical form to measure the extent of asymmetry that is also less affected by outliers considering simultaneously the shape, appropriate measure of location and appropriate measure of dispersion of a distribution. They also suggested a new method to detect outliers for asymmetric distribution which has a step wise detection criterion considering simultaneously the appropriate measure of location and appropriate measure of dispersion of the asymmetric distribution. They found appropriate ways to detect asymmetry and to detect outlier using mean deviation measured from median as in this days of computer, computation of absolute mean deviation from median, and determination of median even for a very large data set is not difficult. Forhad and Adnan's Skewness coefficient is

$$SK_{FA} = \frac{\sum_{i=1}^{n}(x_i - sample\ median)}{\sum_{i=1}^{n}|x_i - sample\ median|}$$

They have proposed that any data point is said to be outlier if it is outside 3md (or more i.e. 4md, 5md and so on depending on the amount of Type one error the experimenter ready to accept) from median. If 6md is taken apart from median it covers more than 97% of the total area (according to Chebyshev Inequality). Based on this cutoff point EUPP (Exceptional Until Proven Pure) method is proposed for an asymmetric distribution where EUPP method assumes a data point to be an outlier if it is outside 3 md (or more) from median unlike the traditional three standard deviation from mean (the central value for normal or symmetric (distribution). EUPP method first drops the most extreme data points (one from upper tail and one from lower tail) before computing the md (mean deviation about median) of the sample. If those data points are still within three md (or more) from the median, then they are retained, if not, they are dropped and the same process is repeated for the second most extreme data points. This procedure is consequently iterated until the extreme data points can no longer be expelled using this heuristic.

A new way of constructing the control charts based on new control limits along with its mathematical form have been suggested by Adnan (2018) which is also less affected by outliers considering simultaneously the shape (symmetric or asymmetric), median as an appropriate measure of location and mean deviation measured from median as an appropriate measure of dispersion of a distribution. The mathematical form of the proposed control limits were

$$CL = median \mp 6 * E(x - \widehat{median}).$$

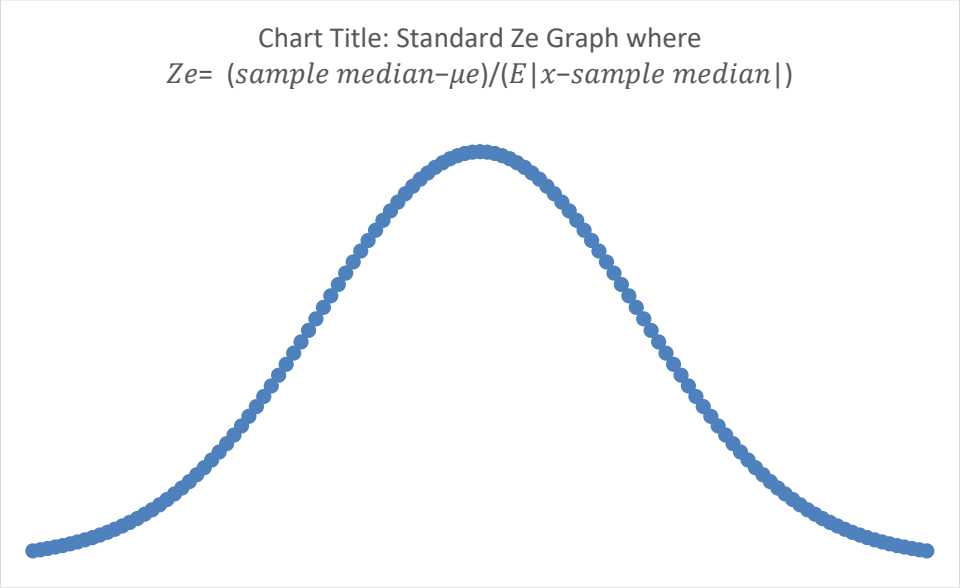
The traditional Z statistic can be viewed as below

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x}-\mu}{\frac{\frac{5}{4}md}{\sqrt{n}}} = \frac{\bar{x}-\mu}{\frac{\frac{5}{4}E|x-\mu|}{\sqrt{n}}}.$$

The proposed Z could be of the form:

$$Z_e = \frac{sample\ median-\mu_e}{E|x-sample\ median|}.$$

As we know that for a skewed distribution (or even symmetric distribution) median is the middle most or central value, so median (m) may be considered as the measures of central tendency instead of mean and scatteredness of the data set of previous type of asymmetric (or



symmetric) distribution may be measured by using mean deviation from median (md) as it is least than any other measures of dispersion. If we take 6 md (mean deviation measured from median) from median, it covers 97% of the total area. Both median and mean deviation from median are less affected by outlier(s) and/or extreme value(s).

Besides, the range of mean deviation measured from mean is greater than that from median and range of standard deviation is greater than that of mean deviation measured from mean [Adnan *et al* (2013)]. The ranges of them are as below respectively

$$\frac{d}{\sqrt{2n}} \leq sd \leq d,$$

$$\frac{0.8d}{\sqrt{2n}} \leq md_{mean} \leq 0.8d,$$

$$\frac{d}{n} \leq md_{median} \leq \frac{d}{2}.$$

4. A Real Life Example

At the end of a semester the university authority asked to apply statistical method to analyze the student evaluation data with an aim to improve the evaluation system of the university as some unexpected evaluation reports regarding some very reputed faculty were received [4]. It was noticed that two instructors have same mean point with exactly same standard deviation (one experienced and other with no previous experience) so the experienced teacher raised a genuine question about the evaluation procedure that need to be investigated. [With this aim in view the relevant data were from the students and also from the course teachers. The whole process was done in such a way that neither the students nor the instructors of the courses could understand anything, in order to avoid possible biasness.] It was observed that there was a negative relationship between the number of lecturers not attended by the students and scores of the students in the examination. Surprisingly enough it was noticed that the evaluation by the irregular students were very poor. It may happen that the irregular student could not follow class lectures properly and for this reason they evaluated instructors teaching procedure poorly which seems to be a clue to resolve the problem. The devils of these problems were the outliers. So before calculating the central value (mean) of a data set the data set should be checked with some suitable statistical tools whether it contains any outlier data or not. If there exist any outliers those must be eliminated first then mean or central value of the data set with its measures of dispersion may be calculated. It was also noticed that the two data set seems to have different shape structures. One seems to be almost symmetric and another seems to be negatively skewed. Therefore, the assumption of symmetric parent distribution may not always be tenable in some real life situations.

It is necessary to find the distribution pattern and probability distribution function that the data under study follow, otherwise, outliers may be falsely detected for a set of data assuming symmetric or normal parent distribution. As the observed data set (fig. 1(b)) follows negatively skewed distribution where n> 40. We search and find a negatively skewed probability distribution function that has range 0<x<u where u is any positive value greater than or equal to unity. The density function for the proposed distribution is as follows:

$$f(x) = \frac{(\gamma+1)(\gamma+2)}{u^{\gamma+2}} x^{\gamma}(u-x), \text{ where } 0 < x < u$$

and u is a positive constant with $u \geq 1$ for $\gamma \geq 1$, it is a negatively skewed distribution.

The mean and variance of this distribution are

$$E(X) = u\frac{(\gamma+1)}{(\gamma+3)} \quad \text{and} \quad V(X) = \frac{2u^2(\gamma+1)}{(\gamma+3)^2(\gamma+4)} \text{ respectively.}$$

The measures of skewness of this distribution

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(\gamma+3)^6(\gamma+4)^3}{8(\gamma+1)}\left[\frac{(\gamma+2)}{(\gamma+4)(\gamma+5)} - \frac{3(\gamma+1)(\gamma+2)}{(\gamma+3)^2(\gamma+4)} + \frac{2(\gamma+1)^2}{(\gamma+3)^3}\right].$$

The 3rd corrected moment of this distribution is

$$\mu_3 = u^3(\gamma+1)\left[\frac{(\gamma+2)}{(\gamma+4)(\gamma+5)} - \frac{3(\gamma+1)(\gamma+2)}{(\gamma+3)^2(\gamma+4)} + \frac{2(\gamma+1)^2}{(\gamma+3)^3}\right]$$

which is negative for any value of γ larger than 1. It indicates that the distribution is negatively skewed.

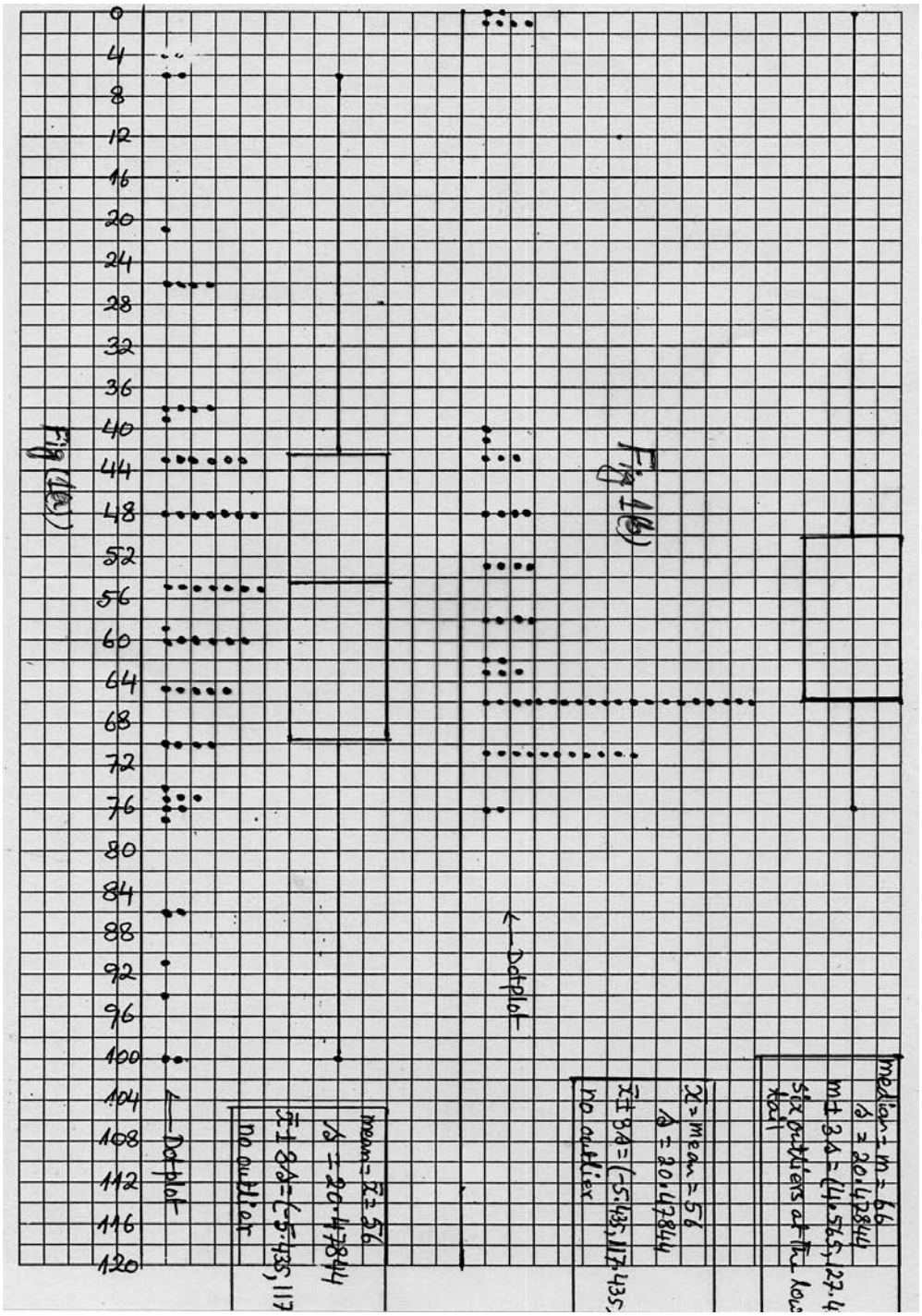


Figure 1: [1(a) dot plot of almost symmetric data set, 1(b):dot plot of asymmetric data set]

For u=1, this distribution converses to standard beta distribution is (Johnson *et. al* (1996))

$$\frac{2m^p(1-m)^q}{(p+q)B(p,q)} = 2V(X)\left\{\frac{1}{B(p+1,q+1)}m^p(1-m)^q\right\}.$$

On the other hand, the mean deviation from median for the proposed distribution is

$$\frac{2m^{\gamma+2}}{u^{\gamma+1}}\left[\frac{\gamma(u-m)+(3u-m)}{u(\gamma+3)}\right]+\left[\frac{\gamma(u-m)+(u-3m)}{\gamma+3}\right].$$

From the dotplots (Fig. 1 (a), 3 (b)) the structure of the two data set may be checked. Before eliminating outliers the mean and standard deviation of the two sets are same as 56 and 20.47844 respectively. But after eliminating the outliers [(by using same method for the two data sets) by using EUPP method] the mean and standard deviation for the first (fig. 1 (a)) set do not change but for the second data set (fig. 1(b)) mean changes to 62.03 and standard deviation to 9.41. It has been observed that their means differ significantly at a = 0.05. The mean and standard deviation do not change for first data set (fig: 1(a)) after applying the "traditional 3-standard deviation cutoff point from mean" [even also eliminating outliers by using GUPI method (as it is a symmetric data set)]. It has been observed that their means differ significantly (p-value = .0192) for the second data set. This indicates that the second set of data has higher central (mean) value than the first one. On the other hand, if we consider median as the measures of central value for the skewed distribution, the two central values become significantly apart from each other. [For asymmetric one 70% observations are larger than the mean value 56, on the other hand for approximately symmetric distribution 54% observations are below mean. So, we may consider median =m=66 as the central value for asymmetric case and mean =56 as the measure of central value for the approximately symmetric case. Hence there exist significant difference between the two central values and obviously the experience and more skilled teacher's ability and quality is much better than his junior colleague.]

Conclusion

The author is trying to develop the distribution of this new Ze, as well as a table of it are more theoretical backgrounds for the new Ze.

References

Adnan, M. A. S., Hossain, F. M. 2007. A New Approach to Determine Skewness and to detect outlier for asymmetric distribution. Journal of Applied Statistical Science. Vol 15 Issue 1, 127-134.

Hossain, F. M., Adnan, M. A. S. and Joarder, A. H. 2013. Shorter variation of standard deviation for small sample. International Journal of Mathematical Education in Science and Technology. 45 (2). DOI:10.1080/0020739X.2013.822588.

Adnan, M. A. S. 2018. Quality Control Charts not based on Sigma Limits. Presented in JRC 2018. Unpublished paper.

https://en.wikipedia.org/wiki/Z-test