SPLIT SAMPLE SKEWNESS

Iftikhar Hussain Adil
School of Social Sciences and Humanities
National University of Sciences and Technology, Islamabad, Pakistan
Email: (a) iftikhar.adil@s3h.nust.edu.pk

Abdul Wahid
National University of Modern Languages, (NUML) Islamabad, Pakistan
abwahid@numl.edu.pk

Abstract:

Shape of the distribution is an important feature of the data mining and it remained the important part for descriptive analysis, risk analysis and decision making. It is also important in finance and investment data analysis. For this purpose numbers of techniques have been formulated by statisticians to compute skewness. Classical skewness is based on mean which is affected by outliers so mean based measures are inefficient especially in small or medium datasets. Quartile and Octile based skewness are also existent to handle the issue of outliers. More recently another measure of skewness has emerged with the name of medcouple. This paper is another attempt to devise new measure of skewness that has been named as split sample skewness due to the fact that sample is divided into two subgroups from median. Simulation study has been done to prove the superiority of new technique over the existing robust measures of skewness.

Substantial literature can be seen on skewness measures as a wide variety of skewness measures have been introduced in literature. These techniques are either parametric or non-parametric; former being the classical measure and Pearson skewness while the latter being the quartile, octile and the recently introduced measure of skewness i.e. medcouple. This study is an attempt to introduce a novel non parametric measure of skewness. The performance of the proposed measure is compared by the existing parametric and non parametric techniques. Monte Carlo simulation study is carried out to prove the superiority of the new measure.

Asymmetry in the probability distribution of the random variable is known to be the skewness of that random variable. Using the conventional third moment measure, the value of skewness might be either positive or negative while in some cases it may be undefined. If the distribution is negatively skewed, it implies that tail on the left side of the probability density function is longer than the right hand side tail of the distribution. It also shows that larger amount of the values including median lie to the right of the mean. Alternatively, positively skewed distribution indicates that the tail on the right side is longer than the left side and the bulk of the values lie to the right of the mean. If the value of the skewness is exactly zero, this suggests symmetry of the distribution. However, it must be noted that the third moment is a crude measure of symmetry and in fact highly asymmetric distributions may have zero third moment. In addition, the third moment is extremely sensitive to outliers, which makes it unreliable in many practical situations. For example, presence of even one outlier may change the sign of the skewness. It is therefore useful to develop alternative measures of skewness which are insensitive to outliers and more direct measures of symmetry. For example, if exactly symmetric data are like{ -5,-4,-3,-2,-1, 0, 1, 2, 3, 4, 5} then its classical skewness is exactly zero but by

replacing just last observation by 50, the classical skewness approaches to 2.66 whereas the other nonparametric measures perform much better in presence of this outlier.

Classical skewness is moment based skewness and can be calculated by the below given formula

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{3/2}} = \frac{\mu_3}{\sigma^3} = \frac{k_3}{k_2^{3/2}}$$

Where E is the expectation operator, $k_2$and k_3 are second and third commulants respectively. This formula can also be transformed into non central moments just by expanding the above formula as

$$\gamma_1 = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{3/2}} = \frac{E[X^3] - 3\mu E[X^2] + 2\mu^3}{\sigma^3} = \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

Another measure of skewness is Pearson (1895) skewness which is based on median and mode as

$$Sk = \frac{Mean - Mode}{Standard\ Deviation}$$

As there is problem with finding mode because some distributions have more than one mode and other may have no mode. So the formula for measuring skewness may be modified as

$$Sk = \frac{3(Mean - Median)}{Standatd\ Deviation}$$

Similarly quartile and octile skewness are introduced by Bowley (1920) and Hinkley (1976) for the better measure of skewness. In quartile skewness data is divided into four parts and in octile it is dividing into eight parts. Their formulae are given as under

$$Quartile\ Skewness = \frac{Q_1 + Q_3 - 2Median}{Q_3 - Q_1}$$

Where Q_1, Median and Q_3 are first second and third quartile respectively

$$Octile\ Skewness = \frac{Q_{0.875} + Q_{0.125} - 2 * Q_{0.50}}{Q_{0.875} - Q_{0.125}}$$

Where $Q_{0.875}$, $Q_{0.125}$ and $Q_{0.50}$ represent 87.5th, 12.5th and 50th (median) percentiles of data respectively. The value of octile skewness varies from -1 to +1 with zero being the point of reference i.e. zero implies symmetry.

Medcouple was introduced by Brys, Hubert Struyf, in 2004. Authors proved it better from the existing techniques by using simulation and formula is given as

$$h(x_i, x_j) = \frac{(x_j - med_k) - (med_k - x_i)}{(x_j - x_i)}$$

Where med_k is the median of X_n, and i and j have to satisfy $x_i \leq med_k \leq x_j$, and $x_i \neq x_j$. This study introduces another measure of skewness names as split sample skewness. The title given to it is based on the reason we split data into two parts from the median and treat each part as independent data set. The new measure introduced is

$$Split\ Sample\ Skewness = Ln\ (IQR_R / IQR_L)$$

Where IQR_L =37.5th -12.5th percentiles and IQR_R =87.5th -62.5th . The reason of keeping IQR_R in the numerator is that if IQR_R is more than IQR_L then ratio will be more than 1 and natural log will be a positive number indicating positively skewed and vice versa. This technique is very easy to measure and has been proved better than all the existing techniques using simulations. It is also best in estimating the risk on either side of the distribution.