A Simple Class of Measures of Skewness

ALTINAY, Galip

Bandirma Onyedi Eylul University, Turkey.

3 July 2016

A Simple Class of Measures of Skewness

Galip ALTINAY

Bandirma Onyedi Eylul University
Department of Econometrics
10200 Bandirma, Balikesir, Turkey

e-mail address: galtinay@yahoo.com

Abstract: In this paper, a simple class of measures for detecting skewness in samples is introduced. The new class of measures is based on a new definition of skewness that takes midrange into consideration. The proposed coefficients of skewness can be computed easily with only three of the summary statistics, i.e., the minimum value, the maximum value and the median (or the mode, or the mean). Another advantage of the new statistics is that they are bounded by -1 and +1, hence, the coefficients of skewness can be interpreted easily. The powers of the proposed statistics to detect skewness are investigated by a limited Monte Carlo simulation in order to have an idea. The preliminary results indicate that the performances of the new statistics look generally good in a limited simulation. However, a more comprehensive investigation is needed.

Keywords: Symmetry, Measure of Skewness, Monte Carlo Study, Midrange, Critical Values.

1. Introduction

Skewness is usually described with reference to symmetry. On the other hand, symmetry is not usually defined clearly, and it is assumed that everyone understands it. There may be many definitions of symmetry depending on the areas where it is used. As Murphy (1982) explains, any statement about symmetry of a structure must be made with reference to some principle of symmetry, a point, a line, an axis. In statistical distributions, the significant point or axis is taken as the center of a distribution. Thus, for unimodal case, the mass is concentrated around the center evenly in a symmetrical distribution. As explained in many statistics textbooks or elsewhere, in a symmetrical distribution, the three popular measures of center (or central tendency), namely, the mean, median and mode coincide at the center. This equality can be considered as the most important characteristic of a unimodal symmetric distribution. Thus a deviation from the symmetry condition is called asymmetry, or simply skewness to Arnold and Groeneveld (1992). In a positively skewed distribution, the ordering of the measures of central tendency generally occurs as mode < median < mean, and the reverse ordering in negatively skewed distributions. The mean-median-mode inequality has been investigated by Groeneveld and Meeden (1977), Runnenburg (1978), MacGillivray (1981), van Zwet (1979), Abdous and Theodorescu (1998), Abadir (2005), and von Hippel (2005), among others, for both continuous and discrete distributions. It is shown in these studies that, although there are some exceptions, the mean-median-mode inequality generally holds in unimodal continuous distributions. However, there are many counter-examples for the mean-median-mode ordering in discrete distributions. Despite the fact that the mean-median-mode inequality is not universal, many measures of skewness are based on this inequality, to be more precise, on the difference between the location parameters in asymmetrical distributions.

As Arnold and Groeneveld (1995) explains, several measures of skewness had been proposed by 1920. Let denote the mean μ, the median m, the mode M, σ standard deviation, Q_1 and Q_3 for the first and the third quartiles, respectively. The measures are, Pearson's coefficient of skewness: $SK_P = \frac{\mu - M}{\sigma}$, Pearson's second coefficient of skewness (see, Doane and Seward, 2011): $SK_{P2} = \frac{3(\mu - m)}{\sigma}$, Yule's coefficient of skewness: $SK_Y = \frac{\mu - m}{\sigma}$, the standardized third central moment: $\gamma_1 = \frac{\mu_3}{\sigma^3}$, and Bowley's coefficient of skewness: $SK_B = \frac{Q_3 + Q_1 - 2m}{Q_3 - Q_1}$. Although several other measures, generally extensions of the above coefficients, have been introduced later on, the early measures are still used today, especially γ_1 (or its variants) is widely used in many statistical software. The first three of the measures of skewness are apparently based on the mean-median-mode inequality, generally encountered in asymmetrical distributions. In cases where the inequality does not hold, the skewness coefficients may give contradictory results. This study attempts to define skewness from a new perspective by taking midrange, a neglected measure of central tendency, into consideration. Based on this definition of skewness, a new class of statistics to measure sample skewness is introduced.

In next section, a new definition of skewness, hence a new method for measuring skewness is developed. In section 3, the properties of the new statistics are explained and the critical values for the new statistics using Monte Carlo simulation are obtained. In section 4, the powers of the proposed statistics are compared to the conventional measures of skewness. An empirical example using the General Social Surveys data is given in section 5, and section 6 concludes.

2. An Alternative Definition of Skewness

What we expect from a symmetrical distribution when we inspect visually a graphical display of data is that one half of the distribution is a mirror image of the other half with respect to the center. Hence, the center of any distribution plays an important role in deciding whether it is symmetrical or not. On the other hand, in measuring skewness, which of the measure of central tendency, (i.e., the mean, mode or median) should be considered as the '*true center*' of a distribution is a critical issue that has not been settled in the literature (see Groeneveld and Meeden, 1984; Arnold and Groeneveld, 1995; Cabilio and Masaro, 1996; Tajuddin, 1996; Das et al., 2009). We can make a statement that, in a symmetrical distribution, all the measures of central tendency located on the center, whereas they depart from the center in case of asymmetry. Thus, to measure any skewness we can just measure how far they are departed from the center. Of course, the question here is, if the mean, mode or median is not the center, what is the center? Here, it is assumed that the best candidate for the center is midrange, a neglected measure of central tendency. The midrange of a dataset is just halfway of its range, i.e., the arithmetic mean of the minimum (X_{min}) and the maximum (X_{max}) values (i.e., midrange = $\frac{X_{min}+X_{max}}{2}$). By considering the midrange as the center, we can define absolute skewness, depicted in Figure 1, as (midrange - θ), where θ is either the mean, median or mode (if exists) of a sample. Therefore, relative skewness (or coefficient of skewness) with respect to the mean, mode and median can be written as follows:

$$SK_{G1} = \frac{\text{midrange} - \text{mode}}{\text{range}/2} = \frac{X_{min} + X_{max} - (2 \times \text{mode})}{X_{max} - X_{min}} \tag{1}$$

$$SK_{G2} = \frac{\text{midrange} - \text{median}}{\text{range}/2} = \frac{X_{min} + X_{max} - (2 \times \text{median})}{X_{max} - X_{min}} \tag{2}$$

$$SK_{G3} = \frac{\text{midrange} - \text{mean}}{\text{range}/2} = \frac{X_{min} + X_{max} - (2 \times \text{mean})}{X_{max} - X_{min}} \tag{3}$$

Division of the absolute skewness by (range/2) guarantees that the coefficients of skewness are bounded by [-1, 1]. Thus, any value of the coefficients, other than zero, indicates skewness as percentage departure from the center, whereas a coefficient value of zero denotes a symmetrical distribution. Although the ordering of the mean, mode and median in Figure 1 (b) and (c) are shown in conventional way, the statistics proposed do not depend on the ordering.

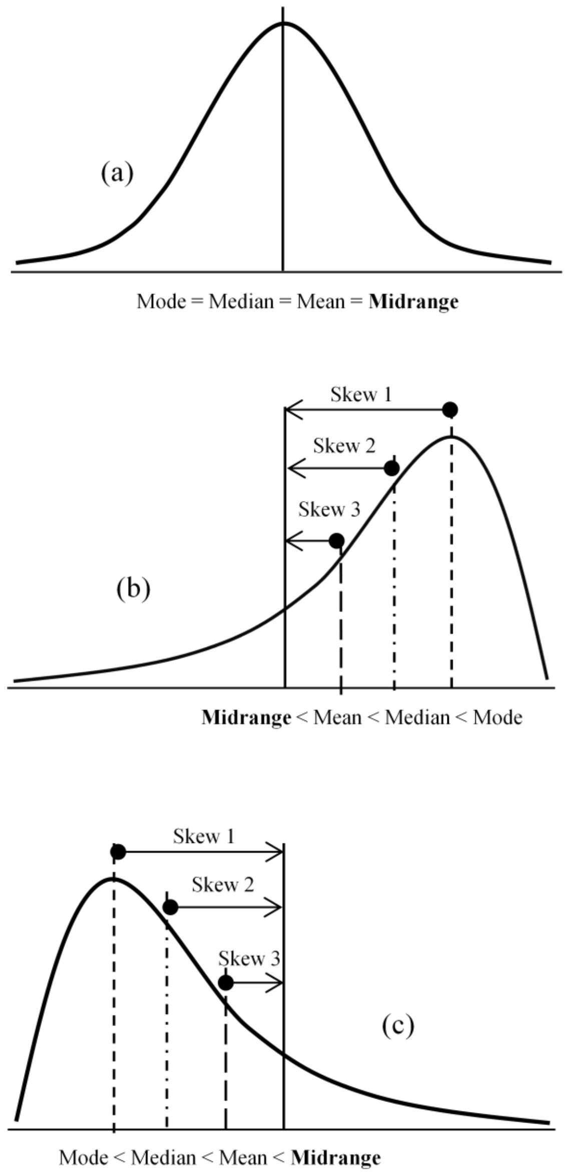

Figure 1. (a) Symmetric distribution, (b) Skewness to the left, (c) Skewness to the right.

3. Properties of the Proposed Statistics

Let γ refer to any of the three skewness measures proposed, and X a sample from either a continuous or discrete distribution. The following properties can be written:

1) $\gamma(aX + b) = \gamma(X)$, for any $a > 0$ and $b \in \mathbb{R}$.

2) $\gamma(-X) = -\gamma(X)$.

3) If X has a symmetrical distribution, then $\gamma(X) = 0$.

4) $\gamma(X) \in [-1, 1]$.

In other words, the coefficient of skewness is not affected by a scale or location change. If the values of a dataset are inverted, the coefficient is also inverted. It is obvious that for a symmetrical distribution, all the measures of central tendency coincide (see Figure 1(a)), thus the coefficient of skewness will be zero. All of the measures proposed are bounded by [-1, 1], i.e., 1 represents extreme right skewness, while -1 represents extreme left skewness. The last property, which most of the other measures of skewness do not possess, is particularly useful for interpreting coefficient of skewness. The other advantage of the proposed measures is that they are very easy to implement, they can be computed with knowledge of a few summary statistics. On the other hand, all of the three statistics are very non-robust since they take only the extreme values into consideration.

By using Monte Carlo simulation based on 20,000 samples drawn from N(0,1), preliminary critical values for the sample skewness statistics, SK_{G2} and SK_{G3}, are obtained. For each sample size from 5 to 150, SK_{G2} and SK_{G3} and their percentiles were calculated. Table 1 shows only the upper 1%, 5% and 10% critical values obtained from the simulation. The lower percentiles are the same except for sign. Note that, the critical values in Table 1 can be used to determine whether the sample is drawn from a normal population or not. Thus, they are not for a test for symmetry in general. As can be seen from Table 1, the critical values of SK_{G2} are higher (in absolute terms) than those of SK_{G3} as expected, since median is farther from the center than the mean is in these samples, taken from normal population. However, they converge as sample size gets larger. The critical values of SK_{G1} are not provided since mode does not exist in this simulation setting.

Table 1. Critical Values for SK_{G2} and SK_{G3} (based on 20,000 replications)

	SK_{G2}				SK_{G3}		
	Upper Percentiles				Upper Percentiles		
n	10%	5%	1%	n	10%	5%	1%
5	0.5743	0.6973	0.8586	5	0.2674	0.3352	0.4374
10	0.3683	0.4640	0.6233	10	0.2360	0.3004	0.4073
15	0.3156	0.3988	0.5460	15	0.2158	0.2732	0.3750
20	0.2720	0.3488	0.4763	20	0.2025	0.2587	0.3578
25	0.2530	0.3179	0.4373	25	0.1914	0.2430	0.3387
30	0.2310	0.2941	0.4067	30	0.1837	0.2337	0.3268
40	0.2092	0.2681	0.3715	40	0.1719	0.2200	0.3067
50	0.1927	0.2467	0.3445	50	0.1641	0.2092	0.2943
60	0.1824	0.2334	0.3278	60	0.1576	0.2026	0.2837
70	0.1726	0.2209	0.3117	70	0.1523	0.1950	0.2746
80	0.1683	0.2146	0.3025	80	0.1494	0.1904	0.2695
90	0.1623	0.2082	0.2886	90	0.1459	0.1863	0.2619
100	0.1575	0.2014	0.2834	100	0.1422	0.1828	0.2568
150	0.1390	0.1793	0.2541	150	0.1294	0.1672	0.2373

Note: The lower percentiles are the same except for sign.

4. Power Comparisons

In order to have an idea about the performances of the proposed statistics, the powers of the statistics are compared to the conventional measures of skewness by using a Monte Carlo simulation. Recently, Tabor (2010) tested the power[1] of eleven different statistics, including the coefficients of skewness given in introduction section, for detecting skewness in samples of size 10 taken from strongly skewed (χ^2 with d.f.= 1), moderately skewed (χ^2 with d.f.= 5) and slightly skewed (χ^2 with d.f.= 40) populations. The same procedure is used in this study in order to make the power of the proposed statistics comparable to those in Tabor (2010). In addition to sample size of 10, to find out the power of the statistics in larger samples, the same procedure based on 10,000 replications is applied to samples with sizes of 30 and 60, and the results are

[1] Probability of rejecting the null hypothesis of normally distributed population against the alternative hypothesis of positively skewed population, when the alternative is presumably true.

presented in Table 2. Only three of the eleven statistics (γ_1, SK_{P2}, SK_B), the most popular ones, are included in the Monte Carlo study. Note that the powers of γ_1, SK_{P2}, SK_B found here are very similar to those in Tabor (2010).

Table 2. Power Comparisons of the Proposed Statistics with the Conventional Statistics

Sample Size	Statistic	Extremely Skewed	Moderately Skewed	Slightly Skewed
		Power at 5 % significance level		
n = 10	SK_{G2}	0.8415	0.3121	0.1090
	SK_{G3}	0.8199	0.3210	0.1176
	γ_1	0.6828	0.2873	0.1145
	SK_{P2}	0.6254	0.1858	0.0838
	SK_B	0.2644	0.0897	0.0626
n = 30	SK_{G2}	0.9998	0.7816	0.2068
	SK_{G3}	0.9998	0.7365	0.1984
	γ_1	0.9943	0.7033	0.2095
	SK_{P2}	0.9654	0.4354	0.1252
	SK_B	0.5331	0.1468	0.0739
n = 60	SK_{G2}	1.0000	0.9690	0.3213
	SK_{G3}	1.0000	0.9467	0.2952
	γ_1	1.0000	0.9494	0.3453
	SK_{P2}	0.9989	0.7118	0.1851
	SK_B	0.7776	0.2239	0.0841

As preliminary results, the performances of SK_{G2} and SK_{G3} seem to be very promising. In extremely skewed and moderately skewed distributions, the powers of SK_{G2} and SK_{G3} are the best; especially in small samples they detect skewness much better. In slightly skewed distributions, the performance of γ_1 is either similar to or slightly better than those of SK_{G2} and SK_{G3}. The power of Pearson's second coefficient of skewness (SK_{P2}), equivalent to Yule's coefficient of skewness (SK_Y) in power, ranks fourth, while Bowley's coefficient of skewness (SK_B) has the worst performance in all cases. In overall assessment, SKG2 may be preferable, not only because of its higher power in general, but also because of its ease of computing.

5. An Empirical Example

So far we have had some idea about the performance of the proposed statistics in continuous data. To find out the performances of the proposed statistics in discrete data, especially in real world data, we consider the General Social Surveys (1972-2010) data, as they were used in von Hippel (2005) and in Garcia et al. (2015). The data given in Table 3. correspond to a survey of respondents who are asked how many people older than 17 live in their household in the USA in 2002.

Table 3. Number of Adult Household Members in the U.S. in 2002 (n = 2,765)

# of Members	1	2	3	4	5
Frequency	1045	1365	259	75	21

The summary statistics of the data in Table 3 are as follows.

mean	median	mode	s.d.	min	max	midrange	range	Q_1	Q_3
1.7928	2	2	0.7783	1	5	3	4	1	2

Although the frequencies suggest a likely skewness to the right, the mean is lower than the median and the mode. This is one of the counter-examples for the mean-median-mode inequality in discrete data. The coefficients of skewness corresponding to the data in Table 3 are as follows.

SK_{G1}	SK_{G2}	SK_{G3}	γ_1	SK_P	SK_{P2}	SK_Y	SK_B
0.5000	0.5000	0.6036	1.1103	-0.2663	-0.7988	-0.2663	-1.0000

Since the mean-median-mode inequality does not hold in this example, four of the coefficients of skewness, the ones based on the difference between measures of central tendency (namely, SK_P,

SK_{P2}, SK_Y) and also SK_B, yield negative values indicating the dataset is skewed to the left. Especially, Bowley's coefficient of skewness (SK_B) points to extremely negative skewness. Contrary to them, the proposed coefficients of skewness (SK_{G1}, SK_{G2} and SK_{G3}) as well as γ_1 indicate that the dataset is skewed to the right. Although γ_1 indicates a positively skewed distribution, it is difficult to interpret the magnitude of 1.11, since it is not bounded. The values of SK_{G1}, SK_{G2} and SK_{G3} (around 0.5-0.6) indicate an approximately moderate (or 50% to 60%) skewness to the right.

Note that the proposed measures of skewness (SK_{G1}, SK_{G2} and SK_{G3}) generally yield similar results to the standardized third central moment (γ_1) both in continuous and discrete data.

6. Concluding Remarks

In this paper, a simple class of measures for detecting skewness in samples is introduced. The new class of measures is based on a new definition of skewness that takes midrange of a sample as the reference point. From this perspective, skewness is defined as a deviation of the mean, median and mode from the midrange or center. The powers of the proposed statistics to detect skewness in samples are investigated by a limited Monte Carlo simulation. The preliminary results indicate that the performances of the new statistics seem to be promising, since in most of the cases they have similar or better power properties. Nevertheless, a more comprehensive Monte Carlo study that uses other asymmetrical distributions is needed to make more precise power comparisons. Note that this paper does not claim that the proposed statistics are superior to the conventional measures of skewness. The advantages of the proposed coefficients of skewness are that they can be computed relatively easier, they are more intuitive, and since the coefficients are bounded they are easier to interpret. The weakness of the proposed statistics lies on their simplicity again, since the statistics proposed take only the extreme values into consideration. In case of data with outliers, the coefficients may yield misleading results. Therefore, the proposed statistics can be used to get a quick idea about skewness of sample data in which there are no outliers.

References

Abadir, K. M. (2005). The mean-median-mode inequality: counterexamples. *Econometric Theory*, 21(2): 477-482.

Abdous, B. and Theodorescu, R. (1998). Mean, median, mode IV. *Statistica Neerlandica*, 52(3): 356-359.

Arnold, B. C. and Groeneveld, R. A. (1992). Skewness and kurtosis orderings: an introduction. *Lecture Notes-Monograph Series*, 22: 17-24.

Arnold, B. C. and Groeneveld, R. A. (1995). Measuring skewness with respect to the mode. *The American Statistician*, 49(1): 34-38.

Cabilio, P. and Masaro, J. (1996). A simple test of symmetry about an unknown median. *Canadian Journal of Statistics*, 24(3): 349-361.

Das, S., Mandal, P. K. and Ghosh, D. (2009). On homogeneous skewness of unimodal distributions. *Sankhyā: The Indian Journal of Statistics Series B*, 71(2): 187-205.

Doane, D. P. and Seward, L. E. (2011). Measuring skewness: a forgotten statistic. *Journal of Statistics Education*, 19(2): 1-18.

García, V. J., Martel, M. and Vázquez-Polo, F. J. (2015). Complementary information for skewness measures. *Statistica Neerlandica*, 69(4): 442-459.

Groeneveld, R. A. and Meeden, G. (1977). The mode, median, and mean inequality. *The American Statistician*, 31(3): 120-121.

Groeneveld, R. A. and Meeden, G. (1984). Measuring skewness and kurtosis. *The Statistician*, 33(4): 391-399.

MacGillivray, H. L. (1981). The mean, median, mode inequality and skewness for a class of densities. *Australian Journal of Statistics*, 23(2): 247-250.

Murphy, E. A. (1982). Skewness and asymmetry of distributions. *Metamedicine*, 3(1): 87-99.

Runnenburg, J. T. (1978). Mean, median, mode. *Statistica Neerlandica*, 32(2): 73-79.

Tabor, J. (2010). Investigating the investigative task: Testing for skewness–An investigation of different test statistics and their power to detect skewness., *Journal of Statistics Education*, 18(2): 1-13.

Tajuddin, I. H. (1996). A simple measure of skewness. *Statistica Neerlandica*, 50(3): 362-366.

Van Zwet, W. R. (1979). Mean, median, mode II. *Statistica Neerlandica*, 33(1): 1-5.

von Hippel, P. T. (2005). Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education*, 13(2): 1-13.