Research Article

A New coefficient of Skewness for grouped data

Abstract

The primary objective of this paper is to introduce a new measure for detecting skewness for grouped data, which is simpler than the current measures in its application. The new proposed coefficient of skewness based on the cumulative frequency data and hence uses more information from the tails of the distribution and thus will be more appropriate to detect asymmetry in the data. Another advantage of the new statistic is that it is bounded by -1 and +1; hence, the coefficients of skewness can be interpreted easily. Simulation study is employed to assess the performance of the proposed coefficient of skewness with three of the classical measure of skewness appeared in the literature using the mean square error (MSE) and mean absolute error (MAE). The simulation study strongly supports the use of the proposed measure for comparing the degrees of skewness of different frequency distributions.

Keywords: coefficient of skewness, symmetry, mean absolute error, mean square error

Mahmoud.A.Eltehiwy,[1] Abu-Bakr A. Abdul-Motaal[2]

[1]Faculty of Politics and Economics, Department of Statistics, Beni Suef University, Egypt
[2]Faculty of Commerce, Department of quantitative methods, South Valley University, Egypt

Correspondence: Mahmoud.A.Eltehiwy, Faculty of Politics and Economics, Department of Statistics, Beni Suef University, Egypt, Email Mahmoud.Mohamed111@com.svu.edu.eg

Received: February 17, 2020 | **Published:** April 06, 2020

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[1] 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.[2] 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,[3] Runnenburg,[4] MacGillivray,[5] van Zwet,[6] Abdous and Theodorescu,[7] Abadir,[8] and von Hippel,[9] 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[10] explains, several measures of skewness had been proposed by 1920. Let denote the mean μ, the median m, the mode M, σ standard deviation, Q1 and Q3 for the first and the third quartiles, respectively. The measures are as follows:

1-Pearson's coefficient of skewness (K_1):

$$K_1 = \frac{\mu - M}{\sigma},$$

The numerical value given by this coefficient usually varies between ± 3. In fact, the mode is the least reliable measure of average as it is so much affected by grouping errors. Therefore, this coefficient is unreliable.

2- Pearson's second Coefficient: (K_2)

$$K_2 = \frac{3(\mu - m)}{\sigma}$$

This coefficient has also the value limits between ± 3, and is used when mode cannot be properly defined. In fact, both of the Karl Pearson's coefficients give too much importance to the extreme values.

3- Bowly's Coefficient (B)

$$B = \frac{Q_1 + Q_3 - 2 * Median}{Q_1 - Q_3}$$

where Q_1 and Q_3 are respectively the first and third quartiles. This measure has the defects that it fails to take into consideration the magnitude of the extreme values, and really speaking it measures the skewness of the middle half and not of the whole distribution. As such this measure is rarely used for determing asymmetry. The numerical value of this coefficient varies between ± 1.

4- The standardized third central moment:

$$\gamma_1 = \frac{\mu^3}{\sigma^3}$$

The measure of skewness based on the moments is very good measure. However, it is unpopular because of the difficulty of its calculation and also because of the fact that this measure gives too much importance to the extreme values.

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 two 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.

In the light of the above argument, it is proposed in this paper to develop a new measure of skewness which takes into account the entire set of values, a neglected measure of central tendency and is simpler than the current measures in its application.

In next section, a new method for measuring skewness is developed. In sections 3 and 4, Simulated frequency distributions under different conditions of symmetry and asymmetry provided an opportunity to compare the performance of the proposed coefficient of skewness with that of the Pearson's and Bowly's coefficients using Monte Carlo simulation. An empirical example using the General Social Surveys data is given in section 5, and section 6 concludes.

Proposed measure of skewness

For a symmetrical frequency distribution, let :

C the number of classes,

f_i the frequency of the ith class, $i = 1, 2, 3 \ldots, C$

F_i the cumulative frequency of the ith class as obtained by summing its frequency with the frequencies of all classes below it.

The proposed measure of skewness is defined in terms of F where

$$F = \sum_{i=1}^{C} F_i ,$$

and is based on the assumption that the frequency distribution has equal classes among which no classes have a frequency of zero.

Some properties of F

Some properties of F are now discussed to be used for defining the proposed measure of skewness which will be denoted by (A).

Theorem (1)

For a frequency distribution of equal classes and $f_i \neq 0$; $i = 1, 2, \ldots, C$, the lowest and highest values of F are respectively given by:

$$F_L = \frac{C(C-1)}{2} + f \quad \text{and} \quad F_H = \frac{C(C-1)}{2} + C(f - C + 1) \quad \text{Where}$$

C is the number of classes, $f = \sum_{i=1}^{C} f_i$; f_i is as defined above.

Proof

The lowest value of F, F_L , is achieved when each of the first (C-1) classes has a frequency of one and the last class has a frequency of (f-C+1), i.e., when:

$$f_i = 1 \text{ for } i = 1, 2, \ldots, (C-1) ,$$

and $f_C = f - C + 1$

That is ,

$$F_L = \sum_{i=1}^{C} F_i = \sum_{i=1}^{C-1} F_i + F_C = \sum_{i=1}^{C-1} i + F_C$$

Since $F_C = f$,

$$\text{Then} \quad F_L = \frac{C(C-1)}{2} + f \quad (1)$$

The highest value of F, F_H , is achieved when:

$$f_1 = f - C + 1 ,$$

and $f_i = 1$ for $i = 2, 3, \ldots, C$

That is,

$$F_H = \sum_{i=1}^{C} F_i = \sum_{i=1}^{C-1} i + \sum_{i=1}^{C} (f - C + 1) ,$$

$$\text{Then} \quad F_H = \frac{C(C-1)}{2} + C(f - C + 1) \quad (2)$$

Theorem (2)

For a symmetrical frequency distribution, the value of F is always equal to $f(C+1)/2$.

Proof

For any frequency distribution, F can be expressed as follows:

$$F = \sum_{i=1}^{C} i f_{c-i+1} \quad (3)$$

Since $f_i = f_{C-i+1}$ for a symmetrical distribution, then, for a symmetrical distribution, F can also take the formula:

$$F = \sum_{i=1}^{C} i f_i \quad (4)$$

Summing (4) with (3) will give

$$2F = \sum_{i=1}^{C} i (f_i + f_{C-i+1})$$

$$= (f_1 + f_C) + 2(f_2 + f_{C-1}) + 3(f_3 + f_{C-2}) + \ldots + C(f_C + f_1)$$

$$= (C+1) f$$

That is

$$F = f(C+1)/2 \qquad (5)$$

Corollary

In terms of F_L and F_H, the value of F for a symmetrical distribution is given by:

$$F = \frac{F_L + F_H}{2} \qquad (6)$$

Proof

Using theorems [1] and [2], (6) comes as a consequence of (1), (2), (5)

Now using formulas (1), (2), (5) and (6), the proposed coefficient of skewness (A) is defined so that it has the value limits between ± 1. That is,

$$A = \frac{2F_{ob} - f(C+1)}{(f-C)(C-1)}, \qquad (7)$$

where F_{ob} is the observed F value for the frequency distribution; $F = \sum_{i=1}^{C} i f_{c-i+1}$, f and C are as defined above.

The notion on which the proposed coefficient of skewness is based is that the larger value of f_{ob} the more likely the bulk of items are of low values and hence the stronger is the evidence that the frequency distribution is positively skewed, and vice versa. In addition the more closer to f(C+1)/2 is the value of F_{ob} the more likely the frequency distribution will be symmetrical.

Simulation study

In this study, only three well known measures of skewness are considered for purposes of comparison with the proposed measure. These three measures are based on either the Karl Pearson's method of measuring skewness or on Bowly's method. Simulated frequency distributions under different conditions of symmetry and asymmetry provided an opportunity to compare the performance of the proposed coefficient of skewness with that of the Pearson's and Bowly's coefficients.

The simulations

1. For symmetrical distributions: 1000 samples, each of size 500 observations, were generated from the normal distribution using the R program . The 500 observations for each sample were then grouped into a frequency distribution of equal classes; each of width 2. Then, using the Kolmogorov-Smirnow test, it was determined that each frequency distribution included in the analysis was consistent with the normal distribution. It was also determined that none of the frequency distributions considered has any class of a frequency zero.

2. For skewed distributions: 5000 samples, each of size 200 observations, were generated from a chi-square distribution of 10 degrees of freedom using the statistical program R. The 200 observations for each sample were then grouped into a frequency distribution of equall classes: each of width 2.

The following frequency distributions were discarded from the analysis and replaced by the convenient ones:

i. Frequency distributions found inconsistent with a chi-square distribution of 10 degrees of freedom.

ii. Frequency distributions with any class of a frequency zero.

iii. Bimodal frequency distributions.

This total operation was then repeated with a chi-square distribution of 15 degrees of freedom.The number of classes was set at 11 and 13 for the chi-square distributions with 10 and15 degrees of freedom respectively.

Methodology

Methods used: the Coefficients of skewness considered in the analysis were:

The Pearson's two measures of skewness (K_1 and K_2).

The Bowly's coefficient (B)

The proposed coefficient (A).

The four coefficients of skewness were developed for each frequency distribution of the three families of generated samples described above. Sine the Pearson's coefficients have the value limits between ± 3 whereas the Bowly's coefficient has the value limits between ± 1. Therefore, to make a meaningful comparison, the measures 3A and 3B were considered instead of A and B respectively

Criteria used for evalution of various measures of skewness

For the measures of skewness considered , the mean-square error (MSE) and mean absolute error (MAE) were used for evaluating performances in the following manner:

I. The Mean-Square Error (MSE):

Let $\hat{K}_1$, $\hat{K}_2$, $\hat{B}$ and $\hat{A}$ respectively denote the coefficients K_1, K_2, B and A when used as estimators for their correspondent true values K_{1t}, K_{2t}, B_t and A_t.

a. for symmetrical distributions, the MSE for each coefficient was obtained. Since the true value for any coefficient of skewness for a symmetrical distribution is equal to zero, that is:

$$K_{1t} = K_{2t} = B_t = A_t = 0$$

Therefore, the mean-square error (MSE) for the Pearson's coefficient (K_1), for example, is given by:

$$MSE(\hat{K}) = E(\hat{K}_1)^2,$$

and so for other measures of skewness.

b. for the two skewed distributions, the MSE was obtained for each coefficient when used for estimating its corresponding true value. In this case, the MSE for the Pearson's coefficient (K_2), for example, is given by:

$$MSE(\hat{K}_2) = Var(\hat{K}_2) + E(K_{2t} - E(K_{2t}))^2,$$

and so for other measures of skewness.

To determine the MSE and MAE for different coefficients cosidering the two chi-square distributions, it is required to determine the true values of these coefficients for both distributions which requires, in turn , some measures to be computed (Table 2 shows the values of these measures). It should be pointed out that these measures were determined as follows:

The mean (μ) and standard diviation (σ); it is known that they are K and $\sqrt{2K}$ respectively for a chi-square distribution of K degrees of freedom.

The median (Q_2), first Quartile (Q_1) and third Quartile (Q_3): were determined with the required precision using the statistical package MathCad

The mode: it can be proven that a chi-square distribution of K degrees of freedom has a unique maximum at $X = K - 2$, that is, the mode of such a distribution is equal to K-2.

The true values of the Pearson's and Bowly's coefficients were then obtained (Table 3). In case of the proposed coefficient, the expected frequencies for the theoretical chi-square distributions with 200 observations and equal classes of width 2 were determined first considering 10 and 15 degrees of freedom and using the statistical program R. Then, the true values of the proposed coefficient were obtained for each chi-square distribution considered (Table 3).

Since the true values of various coefficients are different from each other, it might be convenient to obtain the relative mean-square error (RMSE) for each coefficient by dividing its mean-square error by its true value (e.g. $RMSE(\hat{K}_1) = MSE(\hat{K})/K_{1t}$)

It may be emphasized that the old and proposed coefficients of skewness are based on entirely different principles and hence the results obtained will be different. Therefore, each coefficient ,by itself, is of little use and it is useful when we try to decide which of distributions shows the greater degree of skewness. This was the main reason for considering two different skewed distributions. The proposed and old coefficients of skewness were used to compare the degrees of skewness of the two chi-square distributions as shown in the following point.

c. To compare the degrees of skewness of the two chi-square distributions, the 50 frequency distributions constructed for each distribution were ranked from 1 to 50 according to order of execution on computer. Let $S(k), k = 10, 15$ denote the degree of skewness of a chi-square distribution with k degrees of freedom as obtained by a measure S. then, the amount $\dfrac{S(10)}{S(15)} = R(S)$ say, was obtained for each pair of frequency distribution of the same rank. Thus, 50 values for R were obtained for each coefficient of skewness considered. The MSE of R was then determines for each coefficient when used for estimating the true value of its corresppondent R. that is , the mean-square error of R for the proposed coefficient, for example, will take the formula:

$$MSE\{R(3\hat{A})\} = Var\{R(3\hat{A})\} + E\{R(3A_t) - E[R(3\hat{A})]\}^2 ,$$

and so for other measures of skewness.

(II). The mean Absolute Error (MAE):

The total procedures described in (a), (b) and (c) were then repeated with the mean absolute error as obtained from the risk function which takes the formula:

$$R_\tau(\theta) = E|T - \tau(\theta)|$$

where T is an estimator of $\tau(\theta)$. For example the mean absolute error of $\hat{K}_1$ is given by :

$$MAE(\hat{K}_1) = E|\hat{K}_1 - K_{1t}|$$

Results

Symmetrical distributions

For symmetrical distribution (normal distribution), Table 1 shows the results for the mean-square error (MSE) and mean absolute error (MAE).

Table 1 MSE and MAE for coefficients of skewness (Symmetrical distributions)

Criterion.	Coefficient of Skewness			
	Pearson (K_1)	Pearson (K_2)	Bowly 3B	Proposed (3A)
MSE	0.0581	0.0075	0.0248	0.0223
MAE	0.1973	0.0699	0.1253	0.1236

It can be concluded from Table 1 that the Pearson's coefficient of skewness (K_2) gave the best results with respect to the MSE and MAE criteria. The proposed coefficient (A) came to be the second best measure in terms of both criteria. The appropriate interpretation of the worst results obtained by the Pearson's coefficient (K_1) is perhaps that the mode is so much affected by grouping errors that it becomes unreliable.

Skewed distributions

Table 2 shows the values of measures required to compute the true values of the Pearson's and Bowly's coefficients of skewness for the chi-square distributions with 10 and 15 degrees of freedom, whereas Table 3 presents the true values of these coefficients together with the true value of the proposed one.

Table 2 The true values of measures required for determining the coefficients of skewness for the Chi-square distributions

Distribution	Measures					
	μ	σ	Q_2	Q_1	Q_3	Mode
Chi-square (10 d.f.)	10	4.4721	9.342	6.737	12.549	8
Chi-square (15 d.f.)	15	5.4772	14.339	11.036	18.245	13

Table 3 The true values of various measures of skewness (the chi-square distributions)

Skewness	Coefficient of Skewness			
(Dist.)	Pearson (K_1)	Pearson (K_2)	Bowly 3B	Proposed (3A)
S(10)	0.4472	0.4414	0.3107	0.8965
S(15)	0.3652	0.362	0.2509	0.7205
R(S)	1.2245	1.2193	1.2383	1.2443

It can be concluded from Table 3 that similar results were obtained when the true values of different coefficients were used for comparing the skewness of the two chi-square distributions (the true values of S(10)/S(15) will be identical for different coefficients when rounded to the nearest tenth). It should be pointed out here that this value was found to be 1.2247 for the coefficient of skewness based on the moments which coincides with our results.

Using the true values of different coefficients (Table 3) , the mean square error (MSE), square error (RMSE), mean absolute error (MAE) and relative mean absolute error (RMAE) were obtained for each coefficient when used either for estimationg its true value or for comparing the skewness of the two chi-square distributions (Tables 4 and 5). from these Tables, the following points can be drawn:

Table 4 The MSE and MAE for different coefficients of skewness for the two chi-square distributions

criterion	Skew. (Dist.)	Coefficient of Skewness			
		Pearson (K_1)	Pearson (K_2)	Bowly (3B)	Proposed (3A)
MSE	S(10)	0.0493	0.0190	0.0361	0.0216
	S(15)	0.0391	0.0119	0.0315	0.0178
	R(S)	1.0751	0.4754	3.5957	0.0534
MAE	S(10)	0.1816	0.1085	0.1597	0.1233
	S(15)	0.1685	0.0905	0.1352	0.1117
	R(S)	0.8081	0.5006	1.2503	0.1695

Table 5 The RMSE and RMAE for different coefficients of skewness for the two Chi-square distributions

criterion	Skew. (Dist.)	Coefficient of Skewness			
		Pearson (K_1)	Pearson (K_2)	Bowly (3B)	Proposed (3A)
RMSE	S(10)	0.1102	0.0431	0.1162	0.0240
	S(15)	0.1071	0.0329	0.1256	0.0247
	R(S)	0.8780	0.3899	2.9037	0.0429
RMAE	S(10)	0.4061	0.2459	0.5139	0.1375
	S(15)	0.4614	0.2499	0.5388	0.1551
	R(S)	0.6599	0.4105	1.0097	0.1362

a. When different coefficients were used for estimating their correspondent true values:

i. In terms of the mean-square and mean absolute errors, the Pearson's coefficient (K_2) gave the best results and was followed by the proposed coefficient. However, the proposed coefficient gave more competitive results to that obtained by the K_2 coefficient that it was in the case of symmetrical distribution.

ii. The proposed coefficient gave considerably the best result with respect to the RMSE and RMAE. It was followed by the K_2 coefficient.

iii. The results for the K_1 and B coefficients were so much less satisfactory than the proposed and K_2 coefficients with respect to all criteria used.

b. when different coefficients were used for comparing the degrees of skewness of the two chi-square distributions

i. the performance of the proposed measure of skewness was superior to the other measures. This was true with respect to all criteria used (the MSE, MAE, RMSE and RMAE criteria).

ii. Again the Pearson's K_1 and Bowly's coefficients gave the worst results in terms of all criteria used.

The reason for the results given in (ii) and (iii) for (a) and given in (i) and (ii) for (b) is may be that the proposed coefficient was found to be the most stable measure of skewness as determined by the coefficient of variation (Table 6). This was true when various measures of skewness were used either for estimating their correspondent true values or for comparing the degrees of skewness of the two chi-square distributions.

Table 6 The coefficients of variation (%) for various measures of skewness for the two Chi-square distributions

Skewness (Dist.)	Coefficient of Skewness			
	Pearson (K_1)	Pearson (K_2)	Bowly 3B	Proposed (3A)
S(10)	31.65	57.18	60.12	10.65
S(15)	30.03	56.48	58.42	12.27
R(S)	49.04	78.09	103.95	17.52

Results in Table 6 indicate that the measure of skewness considered differ as to their relative stabilities. The proposed measure is the most stable measure and was followed by the Pearson's measure (K_1).

The coefficient of variation of the Pearson' measure (K_2) did not considerably differ from that of the Bowly's measure (B) and both of them were the least stable measures. In fact the differences in stability were quite marked between the proposed measure and other measures of skewness. These results may be justify the results obtained for various measures of skewness.

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[9] and in Garcia et al.[11] The data given

in Table 7 correspond to a survey of respondents who are asked how many people older than 17 live in their household in the USA in 2002.

The summary statistics of the data in Table 7 are as follows.

Table 7 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

Mean	Median	Mode	s.d.	Min	Max	Range	Q_1	Q_3
1.7928	2	2	0.7783	1	5	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 VII are as follows.

A	γ	K_1	K_2	B
0.60471	1.1103	-0.2663	-0.7988	-1.000

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, K_1, K_2 and B) yield negative values indicating the dataset is skewed to the left. Especially, Bowley's coefficient of skewness (B) points to extremely negative skewness. Contrary to them, the proposed coefficients of skewness (A) as well as γ indicate that the dataset is skewed to the right. Although γ indicates a positively skewed distribution, it is difficult to interpret the magnitude of 1.11, since it is not bounded. The values of A (0.60471) indicate an approximately moderate skewness to the right.

Conclusion

This paper shows that various measures of skewness considered could yield, as expected, different degrees of skewness for the same frequency distribution. However, it was useful to use them either for estimating their true values for the symmetrical and skewed distributions or for comparing the degrees of skewnss of the two chi-square distributions with 10 and 15 degrees of freedom. In case of symmetrical distribution, the MSE and MAE showed that the Pearson' coefficients (K_2) was the best measure for determining symmetry of the normal distribution and was followed by the proposed measure of skewness. In cases of skewed distributions, the RMES and RMAE strongly support the use of the proposed measure of skewness for comparing the degrees of skewness of the two chi-square distributions. In general, results pointed to the relative inferiority of the performance of the Bowly's and Pearson's (K_1) measures of skewness when compared with that of the proposed and Pearson's K_2 measures.

Finally, it must be stressed that each coefficient, by itself, is of little use and it becomes useful when used for comparing skewness of different frequency distributions. Therefore, the results obtained for the proposed measure of skewness are of great value bearing in mind its simplicity in application relative to the complexity of the other measure.

Acknowledgments

Conflicts of interest

None.

References

1. Murphy EA. Skewness and asymmetry of distributions. *Metamedicine.* 1982;3(1):87–99.

2. Arnold BC, Groeneveld RA. Skewness and kurtosis orderings: an introduction. *Lecture Notes-Monograph Series.* 1992;22:17–24.

3. Groeneveld RA, Meeden G. The mode median and mean inequality. *The American Statistician.* 1977; 31(3):120–121.

4. Runnenburg JT. Mean median mode. *Statistica Neerlandica.* 1978;32(2):73–79.

5. MacGillivray HL. The mean median mode inequality and skewness for a class of densities. *Australian Journal of Statistics.* 1981;23(2):247–250.

6. Van Zwet WR. Mean median mode II. *Statistica Neerlandica.* 1979;33(1):1–5.

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

8. Abadir KM. The mean-median-mode inequality: counterexamples. *Econometric Theory.* 2005;21(2):477–482.

9. Von Hippel PT. Mean and skew: Correcting a textbook rule. *Journal of Statistics Education.* 2005;13(2):1–13.

10. Arnold BC, Groeneveld RA. Measuring skewness with respect to the mode. *The American Statistician.* 1995;49(1):34–38.

11. García VJ, Martel M, Vázquez-Polo FJ. Complementary information for skewness measures. *Statistica Neerlandica.* 2015;69(4):442–459.