Computational Journal of Mathematical and Statistical Sciences
3(2), 417–431
DOI:10.21608/CJMSS.2024.288527.1052
https://cjmss.journals.ekb.eg/

Research article

A Proposed Distribution-Free Test For Symmetry of Grouped Data

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

[1] Statistics department, the Politics and Economics faculty, Beni -Suef university, Egypt; mahmoudtehiwy@gmail.com
[2] Faculty of Commerce, Department of Statistics, South Valley University, Egypt; abubakrabdulmotaal@gmail.com

* **Correspondence:** mahmoudtehiwy@gmail.com

Abstract: This paper introduces a new, straightforward test for analyzing the symmetry of data represented in frequency distributions. Unlike other methods, this test doesn't require any assumptions about the underlying statistical distribution of the data. The only prerequisite is that the data has equal-sized categories (bins) and none of these categories have a frequency of zero. This makes the test both statistically robust and computationally efficient. In simpler terms, the test offers a user-friendly way to assess whether a data set leans towards one side or the other (positive or negative skew) without relying on specific statistical models. The test relies on a statistic denoted by G, which was introduced by [1]. We introduce and study the properties of G statistic and obtain its mean and variance. We explore the Asymptotic distribution of the G Statistic. Finally, a worked example and a practical example is presented to illustrate how the test is implemented.

Keywords: Asymptotic normal distribution; distribution-free test; skewness; symmetry.
Mathematics Subject Classification: 62E20, 11K16, 62F03
Received: 10 May 2024; **Revised:** 9 July 2024; **Accepted:** 11 July 2024; **Published:** 13 July 2024.

1. Introduction

Many statistical approaches assume that data follow a normal distribution. However, many such processes are resistant to breaches of normality; therefore, having data from a symmetric distribution is often sufficient for their validity. Other processes, such as non-parametric methods, use symmetric distributions instead of normal distributions. The existence or lack of symmetry is also an essential consideration when determining which parameter to estimate. According to [2] and [3], if the distribution is symmetric, the point of symmetry (μ) is the sole natural measure of position. However, if the distribution is non-symmetric, there is no longer a single fair measure of location. Thus, there are several reasons to investigate the existence or lack of symmetry, and the subject of verifying symmetry has received significant attention in the literature. Furthermore, most parametric, and distribution-free methods of testing symmetry that have appeared in the literature have been developed for ungrouped data (see [4], [5],[6],[7],[8],[9],[10],[11],[12],[13],[14],[15]).

Some authors propose non-normality robust test for skewness. Non-parametric tests for symmetry were developed by [16], based on kernel estimation techniques. Data-driven smooth tests of symmetry have been proposed by [17]. Consistent non-parametric statistical tests of symmetry hypotheses were proposed by [18], based on Cramér–von Mises statistics computed from empirical distribution and characteristic functions.

The literature on testing symmetry gives very little guidance on how to test the non-parametric symmetry of a frequency distribution. Many statistical methods rely on the assumption of a normal or at least a symmetric distribution. However, real-world data often deviates from these ideal distributions. While these methods are valuable, they are not readily applicable to data presented in frequency distributions, which are common in many fields. The main concern of this paper is to propose a novel, distribution-free test specifically designed to analyze the symmetry of grouped data (frequency distributions). This new approach offers several advantages:

1- **No assumptions about underlying distribution**: Unlike existing methods, our test doesn't require specific assumptions about the underlying statistical distribution of the data. This makes it more robust and applicable to a wider range of datasets.

2- **Computational efficiency**: The test is designed for efficiency, requiring only data with equal-sized categories (bins) and no zero frequencies. This simplifies calculations and makes the test practical for real-world applications. In simpler terms, our test provides an easy way to assess data skewness (positive or negative) without relying on complex statistical models. This makes it accessible to researchers from various disciplines.

We introduce and explore the properties of the G statistic, a key component of the test, introduced by [1]. We derive the mean, variance, and asymptotic distribution of the G statistic, which are critical for statistical inference. We illustrate the implementation of the test through a worked example and a practical example, demonstrating its real-world application.

The rest of the paper is organized as follows. We introduce and study the properties of G statistic in Sections 2 and 3, respectively. Testing Symmetry of a Frequency Distribution is discussed in Section 4. Worked example is employed in Section 5. Finally, Section 6 concludes this paper.

2. Proposed Measure of Skewness

A new measure of skewness for frequency distributions was introduced by [1]. This measure is based on the G statistic. For a symmetrical frequency distribution, let:

C is the classes number,

f_i is the i^{th} class frequency, i=1,2,3…,C.

F_i is the i^{th} class cumulative frequency , as calculated by adding its frequency to the frequencies of all classes below it.

The G statistic is used to determine the suggested skewness measure and is defined as follows:

$$G = \sum_{i=1}^{C} F_i.$$

The G statistic is based on the fact that the frequency distribution contains equal classes, with no class having a frequency of zero.

Some properties of G Statistic:

We now explore some properties of G statistic that will be utilized to define the proposed skewness measure, which will be represented by (SK_A).

Theorem 2.1. *The lowest and greatest values of G, for a frequency distribution with equal classes and $f_i \neq 0; i = 1, 2, \ldots, C$ are provided by:*

$$G_L = \frac{C(C-1)}{2} + f \quad \textit{and} \quad G_H = \frac{C(C-1)}{2} + C(f - C + 1),$$

where C is the number of classes, $f = \sum_{i=1}^{C} f_i$.

Proof. The lowest value of G, G_L, is obtained when each of the first (C-1) classes has a frequency of one and the last class has a frequency of $(f - C + 1)$, that is, when:

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

and

$$f_c = f - C + 1.$$

That is,

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

Since $F_C = f$, then

$$G_L = \frac{C(C-1)}{2} + f. \tag{2.1}$$

The highest value of G, G_H, is achieved when:

$$f_1 = f - C + 1, \text{and } f_i = 1 \quad \text{for } i = 2, 3, \ldots, C.$$

That is,

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

then,

$$G_H = \frac{C(C-1)}{2} + C(f - C + 1). \tag{2.2}$$

Theorem 2.2. *The value of G statistic for a symmetrical frequency distribution is always equal to* $f(C+1)/2$.

Proof. For any frequency distribution, G statistic may be represented as follows:

$$G = \sum_{i=1}^{C} i f_{c-i+1}. \tag{2.3}$$

Given that $f_i = f_{c-i+1}$ for a symmetrical distribution, G statistic may alternatively have the formula:

$$G = \sum_{i=1}^{C} i f_i. \tag{2.4}$$

Summing (2.3) with (2.4) yields the following:

$$\begin{aligned} 2G &= \sum_{i=1}^{C} i\,(f_i + f_{C-i+1})\,, \\ &= (f_1 + f_C) + 2\,(f_2 + f_{C-1}) + 3\,(f_3 + f_{C-2}) + \cdots + C\,(f_C + f_1)\,, \\ &= (C+1)f. \end{aligned}$$

That is,

$$G = \frac{f(C+1)}{2} \tag{2.5}$$

Corollary 2.1. The value of G statistic for a symmetrical distribution may be calculated using G_L and G_H as follows:

$$G = \frac{(G_L + G_H)}{2}. \tag{2.6}$$

Proof. Using theorems 2.1 and 2.2, equation (2.6) comes as a consequence of (2.1), (2.2) and (2.5).

Now, using formulas (2.1), (2.2), (2.5), and (2.6), the proposed coefficient of skewness (SK_A) is defined so that it has the value limits between ±1. That is,

$$SK_A = \frac{2G_{ob} - f(C+1)}{G_H - G_L},$$

$$SK_A = \frac{2G_{ob} - f(C+1)}{(f-C)(C-1)}, \tag{2.7}$$

where G_{ob} is the observed value of G statistic for the frequency distribution.

The suggested coefficient of skewness is based on the idea that the bigger the value of G_{ob}, the more probable it is that the majority of items are of low value, providing stronger evidence that the frequency distribution is positively skewed and vice versa. Furthermore, the value of G_{ob} that is closest to $f(C+1)/2$ increases the likelihood that the frequency distribution is symmetrical.

As with different measures of skewness, this measure will be zero for symmetrical distributions, positive for right skewness, and negative for left skewness. Furthermore, according to this measure, skewness falls within the limits of ±1 (See [1]).

The asymptotic distribution of the G statistic, as defined in (2.3), is determined in this paper to be used for testing whether or not the value of skewness of a frequency distribution, as determined by (2.7) differs significantly from zero.

3. The Mean and Variance of the G Statistic

Determining the mean and variance of the G statistic requires first obtaining the mean and variance of f_i, and the covariance of f_i and f_j. This can be done in the following manner:

3.1. The probability Density Function (pdf) of f_i

For a frequency distribution of C classes and sum of frequency f, it can be proven by mathematical induction that the number of all possible arrangements of f_i, $i = 1, 2, \dots, C$ is given by:

$$n = \binom{f-1}{C-1} \tag{3.1}$$

In fact, the number of arrangements of C frequencies so that $f_i = t$, $i = 1, 2, \dots, C$, is equal to the number of arrangements of $(C-1)$ frequencies so that their sum is $(f-t)$. Let this number be denoted by n_t. That is , assuming (3.1) is true, then

$$n_t = \binom{f-t-1}{C-2}, \quad t = 1, 2, \dots, f-C+1, \tag{3.2}$$

as for given C and f, $f_i = 1, 2, \dots, f-C+1$.

Therefore, (3.1) is true if and only if

$$\sum_{t=1}^{f-C+1} n_t = \binom{f-1}{C-1}, \tag{3.3}$$

where n_t is given by (3.2).

Using (3.2), the left-hand side of (3.3) can be expressed as follows:

$$\begin{aligned}\sum_{t=1}^{f-C+1} n_t = \sum_{t=1}^{f-C+1} \binom{f-t-1}{C-2} &= \binom{f-2}{C-2} + \binom{f-3}{C-2} + \cdots + \binom{C-2}{C-2}, \\ &= \binom{f-2}{f-C} + \binom{f-3}{f-C+1} + \cdots + \binom{C-2}{0}, \\ &= \sum_{t=0}^{f-C} \binom{f-2-t}{f-C-t}.\end{aligned}$$

Using the result from [19],

$$\sum_{j=0}^{n} \binom{n-j}{k-j} = \binom{n+1}{k}$$

gives

$$\sum_{t=0}^{f-C} \binom{f-2-t}{f-C-t} = \binom{f-1}{f-C} = \binom{f-1}{C-1},$$

That is,

$$\sum_{t=0}^{f-c+1} \binom{f-t-1}{C-2} = \binom{f-1}{C-1}, \tag{3.4}$$

and hence equation (3.3) is true. Accordingly, equation (3.1) is also true.

It can be concluded from (3.1) and (3.2) that:

$$P_i(t) = P(f_i = t) = \frac{\binom{f-t-1}{c-2}}{\binom{f-1}{c-1}}, \qquad t = 1, 2, \ldots, f-C+1. \tag{3.5}$$

In fact, equation (3.4) implies that (3.5) is a probability mass function.

The mean and variance of f_i can then be obtained as follows:

$$\begin{aligned}E(f_i) &= \sum_{t=1}^{f-C+1} tP_i(t), \\ &= \sum_{t=1}^{f-C+1} \frac{t\binom{f-t-1}{c-2}}{\binom{f-1}{c-1}}.\end{aligned}$$

Using $t = f - (f - t)$ will lead to

$$E(f_i) = \frac{f}{C}, \quad i = 1, 2, \ldots, C \tag{3.6}$$

and

$$E\left(f_i^2\right) = \sum_{t=1}^{f-C+1} t^2 P_i(t) = \sum_{t=1}^{f-C+1} \frac{t^2\binom{f-t-1}{c-2}}{\binom{f-1}{C-1}}.$$

Using $t^2 = f^2 - 2f(f-t) + (f-t)^2$ gives

$$E\left(f_i^2\right) = \frac{f(2f-C+1)}{C(C+1)}, \quad i = 1, 2, \ldots, C \tag{3.7}$$

then, it can be concluded from (3.6) and (3.7) that

$$Var(f_i) = \frac{f(f-C)(C-1)}{C^2(C+1)}, \quad i = 1, 2, \ldots, C \tag{3.8}$$

Consider jointly the two random variables f_i and f_j where $i \neq j$. Their joint probability function, using (3.1), is given by:

$$P\left(f_i = r, f_j = s\right) = \frac{\binom{f-r-s-1}{c-3}}{\binom{f-1}{c-1,}} \tag{3.9}$$

for $r = 1, 2, \ldots, f - C + 1$ and $s = 1, 2, \ldots, f - C + 2 - r$.

Because once f_i is known to equal r, f_j may equal any integer from 1 to $f - C + 2 - r$ (it should be pointed out that for given values of f and C, f_i may equal any integer from 1 to $f - C + 1$ and $f_i + f_j = 2, 3, \ldots, f - C + 2$).

The covariance of f_i and f_j is

$$Cov(f_i, f_j) = E(f_i f_j) - E(f_i)E(f_j).$$

Since the mean of both f_i and f_j is given by (3.6) and the joint probability function of f_i and f_j is given by (3.9), we have

$$\text{Cov}\left(f_i, f_j\right) = \sum_{r=1}^{f-C+1} \sum_{s=1}^{f-C+2-r} rs\left[\frac{\binom{f-r-s-1}{C-3}}{\binom{f-1}{c-1}}\right] - \left(\frac{f}{c}\right)^2. \tag{3.10}$$

The first term in the right-hand side of (3.10) can be expressed as

$$= \binom{f-1}{C-1}^{-1} \sum_{r=1}^{f-C+1} r \sum_{s=1}^{f-C+2-r} s \binom{f-r-s-1}{f-C+2-r-s}.$$

Using $s = (f-r) - (f-r-s)$ gives

$$\sum_{s=1}^{f-C+2-r} s \binom{f-r-s-1}{f-C+2-r-s} = \binom{f-r}{C-1},$$

and hence

$$\sum_{r=1}^{f-C+1} r \sum_{s=1}^{f-C+2-r} s \binom{f-r-s-1}{f-C+2-r-s} = \sum_{r=1}^{f-C+1} r \binom{f-r}{C-1}.$$

Using $r = f - (f - r)$ leads to

$$\sum_{r=1}^{f-C+1} r\binom{f-r}{C-1} = \binom{f+1}{c+1}.$$

Thus ,

$$\operatorname{Cov}\left(f_i, f_j\right) = \frac{\binom{f+1}{c+1}}{\binom{f-1}{c-1}} - \left(\frac{f}{c}\right)^2 \text{, which reduces to}$$

$$\operatorname{Cov}\left(f_i, f_j\right) = \frac{-f(f-C)}{C^2(C+1)} \tag{3.11}$$

Now, we are prepared to determine the mean and variance of the G statistics.

Since $G = \sum_{i=1}^{C} i f_{C-i+1}$, then

$$E(G) = \sum_{i=1}^{C} iE\left(f_{C-i+1}\right) = \sum_{i=1}^{C} i\left(\frac{f}{C}\right),$$
$$= \frac{f}{C}\sum_{i=1}^{C} i = \frac{f}{C}\left(\frac{C(C+1)}{2}\right).$$

That is,

$$E(G) = \frac{f(C+1)}{2}.$$

Similarly,

$$\operatorname{Var}(G) = \operatorname{Var}\left(\sum_{i=1}^{C} i f_{C-i+1}\right) = \sum_{i=1}^{C}\sum_{j=1, i\neq j}^{C} ij\operatorname{Cov}\left(f_i, f_j\right),$$

$$= \operatorname{Var}\left(f_i\right)\sum_{i=1}^{C} i^2 + \operatorname{Cov}\left(f_i, f_j\right)\sum_{i=1}^{C}\sum_{j=1, i\neq j}^{C} ij \tag{3.12}$$

The amount, $\sum_{i=1}^{C}\sum_{j=1,i\neq j}^{C} ij$ can be expressed as

$$\sum_{i=1}^{C}\sum_{j=1}^{C} ij = 2\sum_{i=1}^{C-1}\sum_{j=i+1}^{C} ij = 2\sum_{i=1}^{C-1} i\left(\sum_{j=i+1}^{C} j\right),$$
$$= 2\sum_{i=1}^{C-1} i\frac{(C-i)(C+i+1)}{2},$$
$$= \left(C^2 + C\right)\sum_{i=1}^{C-1} i - \sum_{i=1}^{C-1} i^2 - \sum_{i=1}^{C-1} i^3,$$

which leads to

$$\sum_{i=1}^{C}\sum_{j=1, i\neq j}^{C} ij = \frac{C\left(C^2-1\right)(3C+2)}{12} \tag{3.13}$$

Substituting equations (3.8), (3.11) and (3.13) into equation (3.12), we find after some algebric reduction

$$Var(G) = \frac{f(f-C)(C-1)}{12}$$

We may summarize this as a theorem.

Theorem 3.1. *For a given frequency distribution; let* $G = \sum_{i=1}^{C} i f_{C-i+1}$, *where* C *is the number of classes and* f_i *is the frequency of the* i^{th} *class, then, the mean and variance of* G *are given by*

$$E(G) = \frac{f(C+1)}{2}. \tag{3.14}$$

$$Var(G) = \frac{f(f-C)(C-1)}{12} \tag{3.15}$$

where $f = \sum_{i=1}^{C} f_i$; $f_i \neq 0$.

Corollary 3.1. The G statistic, as defined in (2.3), has a symmetrical distibution about $f(C+1)/2$.
Proof. It has been shown that the lowest and highest values of G, assuming equal classes and $f_i \neq 0$; i=1,2,...,C, are respectively given by (2.1) and (2.2).

It can then be concluded that the mean of G, as defined in (3.14), is always equal to the average of G_L and G_H. This implies that G has a symmetrical distribution about $f(C+1)/2$.

3.2. An alternative proof for Theorem 3.1

In fact, Theorem 3.1 can be proven in a second way using the following theorem:

Theorem 3.2. *If X is the sum of n integers selected at random without replacement, from the first N integers 1 to N, then, the mean and variance of X are given by*

$$E(X) = \frac{n(N+1)}{2}, \qquad Var(X) = \frac{n(N+1)(N-n)}{12},$$

and for large n and large N and $n < N/2$, X has an asymptotic normal distribution with the mean and variance defined above, (See [20]).

In our situation, let F_i denote the cumulative frequency of the i^{th} class, and hence G can be expressed as follows:

$$G = \sum_{i=1}^{C} F_i = \sum_{i=1}^{C-1} F_i + F_C = X_{C-1} + f, \tag{3.16}$$

where

$$X_{C-1} = \sum_{i=1}^{C-1} F_i. \tag{3.17}$$

Since $f_i \neq 0$, then $F_i \neq F_j$ for $i, j = 1, 2, \ldots, C$. Furthermore, $F_i = 1, 2, \ldots, f-1$ for $i = 1, 2, \ldots, C-1$.

Therefore, X_{C-1}, as defined in (3.17), can be viewed as the sum of $C-1$ integers selected at random, without replacement, from the first $f-1$ integers 1 to $f-1$. Hence, the mean and variance of X_{C-1}, according to theorem 3.2 are given by:

$$E(X_{C-1}) = \frac{f(C-1)}{2}, \qquad Var(X_{C-1}) = \frac{f(f-C)(C-1)}{12}. \tag{3.18}$$

Accordingly, the mean and variance of G, using (3.16) and (3.18), are given by:

$$\begin{aligned} E(G) &= E\left(X_{C-1}\right) + E(f) \\ &= \frac{(C-1)f}{2} + f \\ &= \frac{(C+1)f}{2} \end{aligned}$$

as f is a constant and

$$\text{Var}(G) = \text{Var}\left(X_{C-1} + f\right) = \text{Var}\left(X_{C-1}\right) = \frac{f(f-C)(C-1)}{12},$$

which coincides with the results obtained in (3.14) and (3.15).

3.3. The Asymptotic Distribution of the G Statistic

Employing theorem 3.2 in our situation may provide that under the condition that $(C-1) < (f-1)/2$, i.e., $C < (f+1)/2$ which is normally the situation with any frequency distribution, X_{C-1} has an asymptotic normal distribution with the mean and variance defined in (3.18). Hence, the distribution function of G may be approximated by the normal distribution function with the mean and variance defined in (3.14) and (3.15). This leads to the following theorem.

Theorem 3.3. *For a given frequency distribution, let $G = \sum_{i=1}^{C} i f_{C-i+1}$, where C is the number of classes and f_i is the frequency of the i^{th} class, then under the conditions that $f_i \neq 0$ and $C < (f+1)/2$, the distribution function of*

$$\frac{G - \frac{f(C+1)}{2}}{\left(\frac{f(f-C)(C-1)}{12}\right)^{1/2}}, \tag{3.19}$$

where $f = \sum_{i=1}^{C} f_i$, may be approximated by the standard normal distribution function.

4. Testing Symmetry of a Frequency Distribution

It has been shown that the value of G for a symmetrical distribution is always equal to $f(C+1)/2$. Hence, the more closer to $f(C+1)/2$ is the observed value of G the more likely the frequency distribution will be symmetrical and vice versa. Therefore, formula (3.19) can be used for testing symmetry of frequency distribution. In other words, the G statistic is used to test whether or not the value of skewness, as measured by SK_A which is defined in (2.7), differs significantly from zero. This can be done as follows:

Corollary 4.1. The distribution function of $\left[\left(\frac{3(f-C)(c-1)}{f}\right)^{\frac{1}{2}} \text{SK}_A\right]$ may be approximated by the standard normal distribution function.

Proof. Corollary 4.1 comes as a consequence of (2.7) and theorem (3.3).

Corollary 4.2. The measure of skewness, SK_A, as given by (2.7), has an asymptotic normal distribution with mean

$$E(SK_A) = 0, \tag{4.1}$$

and variance

$$Var(SK_A) = \frac{f}{3(f-C)(C-1)} \tag{4.2}$$

Proof. Corollary 4.2 comes as a consequence of corollary 4.1.

To determine whether or not the value of skewness, as measured by SK_A, differs significantly from zero we can compare it against the standard error of skewness. The probability of obtaining that large a skewness value if data come from symmetrical distribution can be evaluated using the Z distribution, where

$$Z = \frac{\widehat{\mathrm{SK}_A} - 0}{S_A}$$

where $\widehat{\mathrm{SK}_A}$ is the estimated value reported for sample skewness coefficient as computed by (2.7) and S_A is the standard error of skewness obtained using (4.2);

$$S_A = \sqrt{Var(SK_A)}$$

Now, testing for symmetry $H_0 : SK_A = 0$ versus asymmetric alternatives is considered, where SK_A is the true unknown population skewness coefficient.

Two-Tailed test:

$H_0 : SK_A = 0$ versus $H_1 : SK_A \neq 0$

$A|Z|$ value exceeding the critical Z-value ($Z_{\alpha/2}$) would lead to rejection of the assumption of symmetry at a significance level α.

One- tailed test:

a.$H_0 : SK_A = 0$ versus $H_1 : SK_A > 0$

If the Z value is larger than Z_α, this would lead to rejection of the assumption of symmetry at a significance level α.

b. $H_0 : SK_A = 0$ versus $H_1 : SK_A < 0$

If the Z value is smaller than $-Z_\alpha$, this would lead to rejection of the assumption of symmetry at a significance level α.

Alternatively, a critical value determined for SK_A for a prescribed significance level and this value compared with the observed value of SK_A. In case of one tailed test, and for a prescribed significance level α, the null hypothesis H_0 that the frequency distribution is symmetric is rejected if the observed value of (SK_A) exceeds the asymptotic $1-\alpha$ quantile (percentage point) obtained for (SK_A), that is, H_0 is rejected if:

$$\mathrm{SK}_A > \mathrm{SK}_{A_{1-\alpha}}$$

where $\mathrm{SK}_{A_{1-\alpha}} = Z_\alpha S_A$ and Z_α is the upper α level of the standard normal distribution; $\Phi(Z_\alpha) = 1-\alpha$, e.g., $Z_\alpha = 1.645$ for $\alpha = 0.05$. S_A is as previously defined.

It can be proven that as $f \to \infty$, the significance value, $\mathrm{SK}_{A_{1-\alpha}}$ can be defined as:

$$\mathrm{SK}_{A_{1-\alpha}} = \frac{(0.577)Z_\alpha}{\sqrt{C-1}}, \tag{4.3}$$

Table 1. The frequency distribution

Class	40-	50-	60-	70-	80-	90-100	Total
Frequency	9	12	7	3	2	2	35

Table 2. The cumulative frequency distribution

Class	Frequency	Cumulative frequency		
40–	9	L.T.*	50	9
50–	12	L.T.*	60	21
60–	7	L.T. *	70	28
70–	3	L.T. *	80	31
80–	2	L.T.	90	33
90 – 100	2	L.T.*	100	35
Total	$f = 35$	$G = 157$		

then, as $f \to \infty$, we can reject the null hypothesis (H_0) that the frequency distribution is symmetric if the observed value of $(SK_A) > \frac{(0.577)Z_\alpha}{\sqrt{C-1}}$.

For a two-tailed test, the assumption of symmetry is rejected if:

$$|SK_A| > Z_{\alpha/2}S_A$$

Since the distribution of (SK_A) is symmetric when $SK_A = 0$, the sign of (SK_A) is ignored when making the test.

5. Worked Example

For purposes of illustrating the computations required for applying the (SK_A) test of symmetry, the following example shows the details of finding the values of (SK_A) and how to apply the proposed test.

Considering the following table 1:

Perform the following test using the 5% significance level.

$$H_0 : SK_A = 0 \qquad against \qquad H_1 : SK_A > 0$$

To perform the SK_A test of symmetry, the cumulative frequency distribution should be first constructed.

L.T.*=less than

From Table 2, the Calculations required for determining the value of (SK_A) is as follows:

$$\text{SK}_A = \frac{2G - f(C+1)}{(f-C)(C-1)} = \frac{2(157) - 35(6+1)}{(35-6)(6-1)} = 0.476$$

For $f = 35$ and $C = 6$, we find: Asymptotic value of $SK_{A0.95} = 0.467$. Since $SK_A > SK_{A0.95}$ ($0.476 > 0.467$), the assumption of symmetry is rejected, and the frequency distribution given is positively skewed.

An Empirical Example

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

Table 4. summary statistics of the data

Mean	**Median**	**Mode**	**S.d.**	**Min**	**Max**	**Mid-range**	**Range**	Q_1	Q_3
1.7928	2	2	0.7783	1	5	3	4	1	2

So far, we've got some sense of how the suggested statistics will behave with continuous data. To assess the effectiveness of the suggested statistics in discrete data, particularly in real-world data, we utilize the General Social Surveys (1972-2010) data as used in [21] and[22]. Table 3 contains data from a 2002 poll of respondents in the United States who were asked how many persons over the age of 17 lived in their home.

The summary statistics of the data in Table 3 are as follows in Table 4:

While the frequencies point to a probable rightward skewness, the mean is less than the median and mode. One of the instances used to refute the mean-median-mode inequality with discrete data is this one. Perform the following test using the 5% significance level.

$$H_0 : SK_A = 0 \qquad against \qquad H_1 : SK_A > 0$$

To perform the SK_A test of symmetry, the cumulative frequency distribution should be first constructed. L.T.*=less than or equal to

From Table 5, the Calculations required for determining the value of (SK_A) is as follows:

$$\text{SK}_A = \frac{2(11633) - 2765(5+1)}{(2765-5)(5-1)} = 0.605$$

For $f = 2765$ and $C = 5$, we find: Asymptotic value of $SK_{A0.95} = 0.475$. Since $SK_A > SK_{A0.95}$ ($0.605 > 0.475$), the assumption of symmetry is rejected, and the frequency distribution given is positively skewed. We notice that the value of Asymptotic $SK_{A0.95}$ calculated from ($SK_{A1-\alpha}$=Z_α.S_A) is equal to value calculated from (4.3) because f is very large.

Table 5. The cumulative frequency distribution

Members	Frequency	Cumulative frequency		
1	1045	L.T.*	1	1045
2	1365	L.T.*	2	2410
3	259	L.T. ∗	3	2669
4	75	L.T.*	4	2744
5	21	L.T.*	5	2765
Total	f = 2765	G = 11633		

6. Conclusions

This paper provides a new distribution-free test of symmetry for data in frequency tables. It has been demonstrated that the proposed test based on G statistic (See [1]) includes several advantages over the existing ones. Unlike classical techniques, it does not need any assumption of the statistical distribution of the data. This flexibility renders it a broadly applicable test. Finally, the constraint that the categories all have equal size and that none have a frequency of zero also makes the test extremely computationally efficient. Based on the analysis of the statistical properties of G statistics (mean, variance, and asymptotic distribution), the paper delivers a solid theoretical foundation for this new methodology. Furthermore, the worked example and case study provide evidence on the simplicity of the implementation as well as the interpretation of findings. To sum up, this study makes a significant addition to non-parametric testing. Based on G-statistic, an user-friendly, statistically robust method with good effectiveness had been proposed to conduct a symmetry test on frequency distributions so that the data can be more creatively explored. Finally, it is intended that a future paper should give further consideration to evaluating the performance of the proposed test compared with some other tests in detecting symmetry and asymmetry of frequency distributions.

Conflict of interest

The authors state that they have no financial or other conflicts of interest to disclose with connection to this research.

References

1. Eltehiwy, M. A.,& Abdul-Motaal, A. A. (2020). A New coefficient of Skewness for grouped data. Biometrics & Biostatistics International Journal, 9(2):54–59.
2. Bickel, P.J., & Lehmann, E.L. (1975). Descriptive statistics for nonparametric models. Ann. Statist. 3, 1038–1069.
3. Antille, A., Kersting, G.,& Zucchini, W. (1982). Testing symmetry. J. Amer. Statist. Assoc. 77, 639–646.
4. Gupta, M. K. (1967). An asymptotically non-parametric test of symmetry. Ann. Math. Statist. 38, 849–866.
5. Gastwirth, J. L. (1971). On the sign test for symmetry. J. Amer. Statist. Assoc. 66, 821–823.
6. Antille, A.,& Kersting, G., (1977). Tests for symmetry. Z. Wahrscheinlichkeitstheorie verw. Gebiete 39, 235–255.
7. Feuerverger, A.,& Mureika, R.A., (1977). The empirical characteristic function and its applications. Ann. Statist. 5, 88–97.
8. Eltehiwy, M., & Abdul-Motaal, A. B. (2023). A new Method for Computing and TestingThe significance of the Spearman Rank Correlation. Computational Journal of Mathematical and Statistical Sciences, 2(2), 240-250.

9. Randles, R, H., Fligner, M. A., Policello I I, G. E., & Wolfe, D. A. (1980). An Asymptotically Distribution- free Test for Symmetry versus Asymmetry. Journal of American Statistical Association, 75, 168-172.

10. Csörgo, S.,& Heathcote, C.R., (1987). Testing for symmetry. Biometrika 74, 177–184.

11. Cabilio, P.,& Masaro, J. (1996). A simple test of symmetry about an unknown median. Canad. J. Statist. 24, 349–361.

12. Mira, A. (1999). Distribution-free test for symmetry based on Bonferroni's measure. J. Appl. Statist. 26, 959–971.

13. Davis, C. E., & Quade, D., (1978). U-Statistics for Skewness or Symmetry. Communications of Statistics-Theory and Methods, A7(5), 413-418.

14. McWilliams, T. P. (1990). A distribution – Free Test for Symmetry Based on a Run Statistical. Journal of American Statistical Association, 85, 1130-1133.

15. Ekström, M., & Jammalamadaka, S. (2007). An asymptotically distribution-free test of symmetry. Journal of Statistical Planning and Inference, 137(3), 799-810.

16. Hyndman, R. J., & Yao, Q. (2002). Non-parametric Estimation and Symmetry Tests for Conditional Density Functions. Journal of Non-parametric Statistics, 14(3), 259–278.

17. Fang, Y., Li, Q., Wu, X., & Zhang, D. (2015). A data-driven smooth test of symmetry. Journal of Econometrics, 188 (2), 490–501.

18. Quessy J-F. (2016). On Consistent Non-parametric Statistical Tests of Symmetry Hypotheses. Symmetry, 8(5), 31. https://doi.org/10.3390/sym8050031

19. Reingold, E.M., Nievergelt, J., & Deo, N., (1977). Combinatorial Algorithms- Theory and Practice. New Jersy: Prentice- Hall, Inc., Englewood Cliffs.

20. Conover, W. J., (1999). Practical Non-parametric Statistics (3nd ed.). New York: John Wiley.

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

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

Disclaimer/Publisher's Note: The content in all publications reflects the views, opinions, and data of the respective individual author(s) and contributor(s), and not those of the scientific association for studies and applied research (SASAR) or the editor(s). SASAR and/or the editor(s) explicitly state that they are not liable for any harm to individuals or property arising from the ideas, methods, instructions, or products mentioned in the content.