BENIN JOURNAL OF
STATISTICS
ISSN 2682-5767
Vol. 4, pp. 114– 134 (2021)

Modified Methods of Computing Some Descriptive Statistics for Grouped Data

O. O. M. Sanni[1], N. A. Ikoba[2*] and O. S. Adegboye[3]
[1,2]*Department of Statistics, University of Ilorin, Ilorin, Nigeria*; [3]*Department of Statistics, Ladoke Akintola University of Technology, Ogbomoso, Nigeria*

(*Received: 26 June 2020; accepted: 30 December 2020*)

Abstract. In this paper, a method that approximates the individual observations in a grouped frequency distribution is presented. This approximation provides an alternative method for computing descriptive statistics like the mean, variance and the mean absolute deviation about the median and also opens up the data to further statistical analysis. In particular, the method for calculating the mean absolute deviation about the median bypasses the absolute operator and provides a new way of assessing dispersion in grouped frequency distributions. The performance of the proposed method was assessed via the mean squared error, mean absolute percentage error and the Kolmogorov-Smirnov test. The descriptive statistics obtained through the proposed method were compared with the Brazauskas-Serfling method, using real-life data and simulated data. The results showed that the proposed method yields a good fit when evaluated against the ungrouped original observations and competed favourably with the Brazauskas-Serfling method, as the values of the performance metrics were equal in most cases. The proposed method therefore provides a way to assess the goodness-of-fit of a grouped data against any hypothesized distribution, as well as provide an estimate of the mean absolute deviation about the median for grouped frequency distributions.

Keywords: Grouped data, de-grouping, goodness-of-fit, mean absolute deviation about the median.

Published by: Department of Statistics, University of Benin, Nigeria

1. Introduction

The categorization of data into class intervals in a grouped frequency distribution may be due to several considerations including compactness and the need to

[*] Corresponding author. Email: ikoba.na@unilorin.edu.ng

provide a quick descriptive summary of the data. However, such grouping leads to loss of important information about the individual observations. Aggregation of large datasets into sets of manageable sizes produces datasets whose entries are symbolic data (Billard and Diday, 2006).

Descriptive statistics are used to provide quantitative description or summary of grouped or ungrouped data. A fundamental difference between descriptive statistics and inferential statistics is that descriptive statistics help to summarize the sample without inference being drawn on the population from which the sample was obtained. Thus, generally, descriptive statistics, unlike inferential statistics, are not developed on the basis of probability theory. However, there is an intrinsic link between descriptive statistics and inferential statistics because the data emanated from a probability distribution, and grouping leads to the loss of vital information that could have been useful for inferential purposes.

The aim of the study is to conceptualize a new method for de-grouping a frequency distribution and open up grouped data to further statistical analysis otherwise restricted by grouping. Alternative formulas for computing the mean, median, variance and the mean absolute deviation about the median for grouped data will be obtained and the performance of the method will be evaluated against an existing de-grouping method using both real-life and simulated data.

A frequency distribution is the organization of raw data in table form, using classes and frequencies (Bluman, 2012). When the size of the sample is large, it is desirable to group the data into non-overlapping classes, called a grouped frequency distribution. The procedure for finding the descriptive statistics of grouped frequency distributions assumes that the mean of all the raw data value in each class is equal to the midpoint of the class (Bluman, 2012). In reality, this is not the case, since the average of the ungrouped data values in each class will not be exactly equal to the midpoints. The procedure is thus an approximation of the true situation, and the midpoint represents an estimate of all the values in the class. Similarly, the estimate of the median for grouped frequency distributions is obtained via interpolation, hence it is also an approximation. The choice of the number of intervals, k in the grouped data can be made using any of the following rules: arbitrary, Sturges', 2 to the k rule and minimum of $\sqrt{n}$ and $10\log_{10}n$, where n is the sample size (Lohaka, 2007; Adegboye, 2009). Lohaka (2007) provides a comprehensive list of alternative ways of constructing frequency distributions. It is noted that sometimes, the class intervals are predetermined by the purpose for which they are to serve, for example, the class interval for the grades in an examination (Adegboye, 2009).

The best predictor of an individual observation in a distribution is the predictor that keeps the sum of the possible absolute prediction error to a minimum. This is known as the least absolute error criterion (Adegboye, 2009). The mean cannot be computed if the frequency distribution has an open-ended class, but the median can be computed for such cases, and the median is also less affected by outliers, in comparison to the mean (Adegboye, 2009; Bluman, 2012).

The breakdown point for a parameter (mean, median, variance, etc) is the proportion or number of arbitrarily small or large extreme values that must be introduced into a sample to cause the estimator to yield an arbitrarily bad result (Brazauskas and Serfling, 2003; Arachchige and Prendergast, 2019). The mean absolute deviation about the median, MAD(md) offers a direct measure of the dispersion of the random variable X from its median. Generally, if the mean absolute deviation is the preferred measure of variability, then the median is the appropriate measure of central tendency and this is the case for asymmetric distributions where the median is a better representative of the centre of the distribution than the mean (Pham-Gia and Hung, 2001). For the normal distribution, $MAD\,(\mu) = MAD\,(md) = \sigma\sqrt{2/\pi} = 0.8\sigma$ (Pham-Gia and Hung, 2001). For asymmetrical distributions, the MAD(md) provides a meaningful dispersion measure related to the centre. The mean absolute deviation about the median, the mean absolute deviation about the mean and the standard deviation for any distribution can be related via a corollary of Lyapounov's inequality (Pham-Gia and Hung, 2001; Choulakian and Abou Samra, 2020):

$$MAD\,(md) \leq MAD\,(\mu) \leq \sigma$$

Very few applications have been found for $\left[E(X - md)^2\right]^{(1/2)}$ as the meaning remains unclear (Pham-Gia and Hung, 2001), and analytic expressions of MAD(md) cannot be obtained for most skewed distributions. The mean absolute deviation about the median has been used effectively to measure skewness andkurtosis and to partially order distributions, as mentioned in Pham-Gia and Hung (2001). A measure of the skewness of a distribution is $\beta = (\mu - md)/MAD(md)$ (Pham-Gia and Hung, 2001). The MAD(md) is hardly amenable to higher mathematical analysis and it is not scale-invariant, as its magnitude depends on the unit of measurement of the data (Adegboye, 2009, Gupta, 2011). It is actually useful or preferred as a measure of dispersion in highly skewed distributions and in the presence of outliers, which tend to distort the standard deviation (Bonnet and Seier, 2003). The MAD(md) tends to increase with the size of the sample though not proportionately and not rapidly as the range (Gupta, 2011). It cannot be computed for distributions with open-ended classes. In spite of its drawbacks, the mean absolute deviation about the median is being deployed in various areas of applications spanning economics, business and industry because of its simplicity, accuracy and the fact that the standard deviation gives greater weightage to the deviations of extreme observations (Gupta, 2011). It is commonly useful in computing the distribution of personal wealth in a country and forecasting business cycles (Gupta, 2011).

Heitjan (1989) provided a detailed exposition of the methods of grouped data inference, problems, techniques and results. The grouped sample variance, unlike the mean, is biased and the bias is reduced when the interval widths become smaller (Heitjan, 1989). Grouping should be carried out cautiously, ass large, unevenly spaced intervals, unequal interval widths and highly multicollinear data significantly affect the estimates obtained via grouping (Heitjan, 1989). Relevant methods for finding the univariate histogram and sample mean and variance for a single interval-valued variable without rules were derived by Bertrand and Goupil (2000). A compendium of important results and the distributional properties of the mean and median absolute deviation were presented in Pham-Gia and Hung (2001). Pham-Gia and Hung (2001) further provided the sampling distributions, distinct notation, behaviour and applications of the MAD.

Bonnet and Seier (2003) derived approximate confidence intervals for the mean absolute deviation about the median in one-sample and two-sample designs and showed that the MAD was a better alternative to the variance when there are departures from the normality assumption. Estimates of the coverage probabilities were obtained using Monte Carlo simulation of specified sample sizes from several probability distributions. The simulation results showed that the proposed confidence intervals had coverage probabilities close to the true confidence interval in mild leptokurtic and mild skewed distributions. Brazauskas and Serfling (2003) presented a uniform de-grouping method for grouped data which enabled their analysis of the distributional properties of the grouped data. The uniform de-grouping approach neither distorts the original data nor changes the total within classes. It allowed for methods of estimation and goodness-of-fit to be carried out because the data has been structured to be continuous. Various goodness-of-fit tests like the Kolmogorov-Smirnov, Anderson-Darling and Cramer von Misses tests were explored for the analysis of the de-grouped data in comparison with some possible probability distributions. However, the formula of Brazauskas and Serfling (2003) could provide estimates for intervals with zero frequency, thereby increasing the number of observations by the number of empty classes in the dataset. Boos and Brownie (2004) outlined test of hypothesis procedures based on alternative measures of scale, such as the mean absolute deviation about the median, asserting that such procedures produce superior Type I and Type II error properties.

Some formulas for basic descriptive statistics on interval-valued data in the presence of rules were presented by Billard and Diday (2006). Crafford (2007) explored a maximum likelihood double iterative procedure to estimate the parameters of grouped data by treating the frequencies as a random vector of a multinomial distribution whose maximum likelihood estimates are the relative frequencies of the observed frequency vector. The Chi-squared statistic and the Wald statistic were used as goodness-of-fit tests to establish the validity of the inference drawn from the iterative procedure. The computational complexity of the method will be high when there are more classes and Crafford (2007) only considered frequency data with only five classes 100 observations. An extensive review of the methods of constructing grouped frequency distributions was provided by Lohaka (2007), who also proposed an iterative method for constructing grouped frequency distributions and histograms by uniquely determining the number of classes, the class width, the starting point and the appropriate range for the distribution. He then compared the estimated parameters of the proposed method with the analysis from the SPSS statistical package using simulated data. It is noted however that Lohaka (2007) iterative method produced a higher number of classes than the widely used Sturges' rule.

Habib (2012) provided an alternative formula for the MAD(md) by introducing a binary indicator function for the values below the median. The mean absolute deviation about the median is then expressed in terms of the covariance between the random variable and the indicator function. This new categorization of the MAD(md) was then used to explore some other properties of the statistic like correlation and skewness as well as the tail length distribution. However, a rigorous justification for the existence of the covariance of the indicator function and the original random variable was not provided in Habib (2012).

Leys *et al.* (2013) surveyed the use of outliers detection techniques in the field of psychology, showing a preponderance of using the mean plus or minus three standard deviations, which has been established to be problematic. The median absolute deviation was proposed as a robust alternative and was found to provide a strong outlier exclusion criteria. A Monte Carlo acceptance-rejection sampling plan within the interval, used to de-group the observations of a grouped frequency distribution was undertaken by Chen and Miljkovic (2019). The method requires the class interval means as input into the process of de-grouping. Chen and Miljkovic (2019) then compared their method with the method of Brazauskas and Serfling (2003) and showed that their method outperformed the previous method, although on the basis of the metrics of performance analysis used (the mean squared error and the bias), there were no significant differences between the two methods, even though their proposed method was more complex than the Brazauskas-Serfling method. The mean and the median are most of the commonly used measures of central tendency, while the variance and mean absolute deviation about the median are measures of dispersion.

2. Materials and Method

Let $x_1,\ x_2,\ \cdots,\ x_n$ be independent and identically distributed observations from a given population. The sample mean, $\overline{x}$ is given as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The sample variance is given as

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n^2}\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]$$

While the median is:

$$md\,(x) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & if\ n\ is\ odd \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}\ +1\right)}\right), & if\ n\ is\ even \end{cases} \tag{1}$$

THEOREM 1 *Let $x_1,\ x_2,\ \cdots,\ x_n$ be a random sample of size n with median $md(x)$. Let b represent the sum of all observations below the median, a, the sum of all observations above the median and T_n, the sum of all the observations. Define the indicator function*

$$I_n = \begin{cases} 1, & if\ n\ is\ odd \\ 0, & if\ n\ is\ even \end{cases}$$

Then the mean absolute deviation about the median, $MAD\,(md)$ is given as

$$MAD\,(md) = \frac{1}{n}(a - b) = \frac{1}{n}(T_n - 2b - I_n md(x)) \tag{2}$$

Proof. Let $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ be the ordered observations of the random sample. By definition,

$$|x| = \begin{cases} -x, & if\ x < 0 \\ +x, & if\ x \geq 0 \end{cases}$$

For the case where n is even,

$$\begin{aligned} MAD\,(md) &= \frac{1}{n}\sum_{i=1}^{n}|x_i - md(x)| = \frac{1}{n}\sum_{i=1}^{n}|x_{(i)} - md(x)| \\ &= \frac{1}{n}\left\{-\sum_{i=1}^{\frac{n}{2}}(x_{(i)} - md(x)) + \sum_{i=\frac{n}{2}+1}^{n}(x_{(i)} - md(x))\right\} \\ &= \frac{1}{n}\left\{-\sum_{i=1}^{\frac{n}{2}}x_{(i)} + \frac{n}{2}md\,(x) + \sum_{i=\frac{n}{2}+1}^{n}x_{(i)} - \frac{n}{2}md(x)\right\} \\ &= \frac{1}{n}\left(\sum_{i=\frac{n}{2}+1}^{n}x_{(i)} - \sum_{i=1}^{\frac{n}{2}}x_{(i)}\right) = \frac{1}{n}(a - b) \end{aligned} \tag{3}$$

Now,

$$T_n = \sum_{i=1}^{n} x_i = \sum_{i=\frac{n}{2}+1}^{n} x_{(i)} + \sum_{i=1}^{\frac{n}{2}} x_{(i)} = a + b$$

Hence, $a = T_n - b$, which upon substitution into equation (3) and rearrangement, yields

$$MAD\,(md) = \frac{1}{n}\,(T_n - 2b)$$

For the case where n is odd, $md\,(x) = x_{\left(\frac{n+1}{2}\right)}$ and

$$\begin{aligned} MAD(md) &= \frac{1}{n}\sum_{i=1}^{n} |x_{(i)} - md(x)| \\ &= \frac{1}{n}\left\{ -\sum_{i=1}^{\frac{n-1}{2}} (x_{(i)} - md(x)) + \left(x_{\left(\frac{n+1}{2}\right)} - md\,(x)\right) + \right. \\ & \left. \sum_{i=\frac{n+3}{2}}^{n} (x_{(i)} - md(x)) \right\} \\ &= \frac{1}{n}\left\{ -\sum_{i=1}^{\frac{n-1}{2}} x_{(i)} + \frac{n-1}{2} md\,(x) + 0 + \sum_{i=\frac{n+3}{2}}^{n} x_{(i)} - \frac{n-1}{2} md(x) \right\} \\ &= \frac{1}{n}\left(\sum_{i=\frac{n+3}{2}}^{n} x_{(i)} - \sum_{i=1}^{\frac{n-1}{2}} x_{(i)} \right) = \frac{1}{n}\,(a - b) \end{aligned} \tag{4}$$

But $T_n = a + b + x_{\left(\frac{n+1}{2}\right)} = a + b + md(x)$, hence $a = T_n - b - md\,(x)$, which upon substitution into equation (4) yields

$$MAD\,(md) = \frac{1}{n}\,(T_n - 2b - md(x))$$

Define an indicator function

$$I_n = \begin{cases} 1, & if\ n\ is\ odd \\ 0, & if\ n\ is\ even \end{cases}$$

Then the $MAD\,(md)$ whether n is even or odd, can be expressed as

$$MAD\,(md) = \frac{1}{n}\left(T_n - 2b - I_n md(x)\right)$$

where

$$T_n = \sum_{i=1}^{n} x_i = \begin{cases} a+b+md\,(x)\,, & if\ n\ is\ odd \\ a+b, & if\ n\ is\ even \end{cases}$$

The mean absolute deviation about the median has thus been expressed as the difference between two numbers and this bypasses the absolute operator. The observations before the median are summed up and subtracted from the sum of observations after the median. ■

THEOREM 2 *Given a grouped frequency distribution with k classes, class frequencies* f_i *, lower limits* L_i*, class widths* w_i *and midpoints* $m_i,\ i = 1, \cdots, k.$ *Let an approximation of the individual observations of the data be given as*

$$y_{ij} = L_i + \frac{w_i}{f_i}\left(j - \frac{1}{2}\right) \qquad j = 1, \ldots, f_i,\ \ i = 1, \ldots, k \tag{5}$$

Then the mean, $\overline{y}$ *and the variance,* S_y^2 *are given respectively by*

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k} f_i m_i = \overline{m}$$

and

$$S_y^2 = S_m^2 + \frac{1}{12n}\sum_{i=1}^{k}\left(\frac{f_i^2 - 1}{f_i}\right) w_i^2 = S_m^2 + \frac{1}{n}\sum_{i=1}^{k} f_i S_i^2 \tag{6}$$

where w_i *is the width of the* i^{th} *class interval,* $S_m^2 = \frac{1}{n}\sum_{i=1}^{k} f_i (m_i - \overline{m})^2$, $S_i^2 = \left(\frac{f_i^2-1}{12f_i^2}\right) w_i^2.$

Proof. Define $[L_i,\ U_i)$ as the lower and upper limits, respectively of the i^{th} class, with width $w_i = U_i - L_i$ and frequency f_i. Let the interval $[L_i,\ U_i)$ be partitioned into f_i sub-intervals each of length w_i/f_i. Approximate the j^{th} observation in the interval $[L_i,\ U_i)$ by y_{ij}, the midpoint of the j^{th} sub-interval. The $(ij)^{th}$ observation, y_{ij} is therefore given as

$$y_{ij} = L_i + \frac{w_i}{f_i}\left(j - \frac{1}{2}\right) \qquad j = 1, \ldots, f_i,\ \ i = 1, \ldots, k$$

Since $m_i = (L_i + U_i)/2$, and $w_i = U_i - L_i$, then $L_i = m_i - w_i/2$ and upon substitution of L_i in the equation above, y_{ij} can be expressed as

$$y_{ij} = m_i + \frac{1}{2}\left(2j - (1 + f_i)\right)\frac{w_i}{f_i}$$

The mean, $\overline{y}$ is then obtained as

$$\begin{aligned}\overline{y} &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\left(m_i + \frac{1}{2}\left(2j - (1 + f_i)\right)\frac{w_i}{f_i}\right)\\ &= \frac{1}{n}\sum_{i=1}^{k} f_i m_i + \frac{1}{n}\sum_{i=1}^{k}\frac{w_i}{f_i}\sum_{j=1}^{f_i} j - \frac{1}{2n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}(1 + f_i)\frac{w_i}{f_i}\end{aligned}$$

Note that the sum of the first f_i natural numbers is

$$\sum_{j=1}^{f_i} j = \frac{f_i\left(f_i + 1\right)}{2}$$

Hence, upon substitution, it is seen that

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k} f_i m_i + \frac{1}{n}\sum_{i=1}^{k}\frac{w_i}{f_i}\frac{f_i\left(1 + f_i\right)}{2} - \frac{1}{2n}\sum_{i=1}^{k} f_i\left(1 + f_i\right)\frac{w_i}{f_i} = \frac{1}{n}\sum_{i=1}^{k} f_i m_i = \overline{m}$$

Therefore

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k} f_i m_i = \overline{m}$$

The variance, S_y^2, is

$$
\begin{aligned}
S_y^2 &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}(y_{ij}-\overline{y})^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\left(m_i+\frac{1}{2}(2j-(1+f_i))\frac{w_i}{f_i}-\overline{m}\right)^2\\
&= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\Bigg((m_i-\overline{m})^2+\left(\frac{1}{2}(2j-(1+f_i))\frac{w_i}{f_i}\right)^2+(m_i-\overline{m})\times\\
&\quad (2j-(1+f_i))\frac{w_i}{f_i}\Bigg)\\
&= \frac{1}{n}\sum_{i=1}^{k}f_i(m_i-\overline{m})^2+\frac{1}{4n}\sum_{i=1}^{k}\left(\frac{w_i}{f_i}\right)^2\sum_{j=1}^{f_i}(2j-(1+f_i))^2+\\
&\quad \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\left((m_i-\overline{m})(2j-(1+f_i))\frac{w_i}{f_i}\right)\\
&= (first\ term)+(second\ term)+(third\ term)
\end{aligned} \tag{7}
$$

Now,

$$first\ term = \frac{1}{n}\sum_{i=1}^{k}f_i(m_i-\overline{m})^2 = S_m^2,$$

$$second\ term = \frac{1}{4n}\sum_{i=1}^{k}\left(\frac{w_i}{f_i}\right)^2\sum_{j=1}^{f_i}(2j-(1+f_i))^2$$

$$= \frac{1}{4n}\sum_{i=1}^{k}\left(\frac{w_i}{f_i}\right)^2\sum_{j=1}^{f_i}(4j^2+(1+f_i)^2-4j(1+f_i))$$

$$= \frac{1}{4n}\sum_{i=1}^{k}\left(\frac{w_i}{f_i}\right)^2\left[\frac{4f_i(f_i+1)(2f_i+1)}{6}+f_i(1+f_i)^2-\frac{4f_i(f_i+1)^2}{2}\right]$$

$$= \frac{1}{4n}\sum_{i=1}^{k}\frac{w_i^2}{f_i}\left[\frac{4(f_i+1)(2f_i+1)+6(f_i+1)^2-12(f_i+1)^2}{6}\right]$$

$$= \frac{1}{4n}\sum_{i=1}^{k}\frac{w_i^2}{f_i}\left[\frac{4\left(f_i+1\right)\left(2f_i+1\right)-6(f_i+1)^2}{6}\right]$$

$$= \frac{1}{12n}\sum_{i=1}^{k}\frac{w_i^2(f_i+1)(4f_i+2-3f_i-3)}{f_i} = \frac{1}{12n}\sum_{i=1}^{k}\frac{w_i^2(f_i+1)(f_i-1)}{f_i}$$

$$= \frac{1}{n}\sum_{i=1}^{k}\left(\frac{f_i^2-1}{12f_i}\right)w_i^2 = \frac{1}{n}\sum_{i=1}^{k}f_iS_i^2,$$

and

$$third\,term = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\left(\left(m_i-\overline{m}\right)\left(2j-\left(1+f_i\right)\right)\frac{w_i}{f_i}\right) = 0$$

Note that the sum of squares of the first f_i natural numbers is given as

$$\sum_{j=1}^{f_i}j^2 = \frac{f_i\left(f_i+1\right)\left(2f_i+1\right)}{6}$$

Substituting these expressions into equation (7) yields

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{k}f_i(m_i-\overline{m})^2 + \frac{1}{n}\sum_{i=1}^{k}\left(\frac{f_i^2-1}{12f_i}\right)w_i^2 = S_m^2 + \frac{1}{n}\sum_{i=1}^{k}f_iS_i^2$$

where $S_m^2 = \frac{1}{n}\sum_{i=1}^{k}f_i(m_i-\overline{m})^2$ is the between-class variance and

$$S_i^2 = \left(\frac{f_i^2-1}{12f_i^2}\right)w_i^2$$

is the within class variance of the i^{th} class. ■

The overall variance S_y^2, in Theorem 2 is partitioned into two components: the between-class variance and the within-class variance. These components can be minimized by having a smaller interval width, w_i in relation to the frequency, f_i. That is, $w_i < f_i$. An important observation is that the variance cannot be computed when any of the classes have zero frequency. That is, for S_y^2 to be computed, all the class frequencies must be greater than zero.

Miner (1934) showed that the variance of a grouped frequency distribution is always higher than the variance of the ungrouped data except for the case where all the observations in each class are the same, and thus the variances will be equal. Therefore the present formula for the variance underestimates the ungrouped sample variance. The proposed estimator of the grouped frequency variance, which incorporates both within-group and between-group variance is thus a better estimator of the variance of the population.

THEOREM 3 *Given a grouped frequency distribution with n observations distributed into k classes, having class frequencies* f_i *and class widths* w_i, $i = 1, \cdots, k$. *Let the observations in the* i^{th} *interval be approximated by*

$$y_{ij} = L_i + (j - \frac{1}{2})\frac{w_i}{f_i}$$

Define the indicator function

$$I_n = \begin{cases} 1, & if\ n\ is\ odd \\ 0, & if\ n\ is\ even \end{cases}$$

Let L_r *be the lower limit of the median class r,* f_r *the frequency of the median class, and* Cf_b *the cumulative frequency of the immediate class before the median class. Then the median md(y) and the mean absolute deviation about the median MAD(md) are given respectively as*

$$md(y) = L_r + \left(\frac{n}{2} - Cf_{r-1}\right)\frac{w_r}{f_r} \tag{8}$$

and

$MAD(md) =$

$$\frac{1}{4nf_r}\left[4\left(T_k - 2T_{r-1}\right)f_r - 4\left(n - 2Cf_b\right)L_r f_r - \left(\left(n - 2Cf_b\right)^2 + I_n\right)w_r\right] \tag{9}$$

where

$$T_s = \sum_{i=1}^{s}\sum_{j=1}^{f_i} y_{ij} = \sum_{i=1}^{s} m_i f_i, \quad s = 1, 2, \ \ldots, r-1, r, \ldots, \ k$$

Proof. Let the median class be r, the lower limit of the median class, L_r, the cumulative frequency for the $(r-1)^{th}$ class, Cf_b and the frequency of the median class, f_r.

$$y_{ij} = L_i + (j - \frac{1}{2})\frac{w_i}{f_i}$$

The observations in the median class r are

$$y_{rj} = L_r + \left(j - \frac{1}{2}\right)\frac{w_r}{f_r} \tag{10}$$

If n is odd, then the median will be the $(n+1)/2^{th}$ observation. That is, $j = (n+1)/2 - Cf_b$. Upon substitution into equation (10), it is seen that

$$md\,(y) = L_r + \left(\frac{n+1}{2} - Cf_b - \frac{1}{2}\right)\frac{w_r}{f_r} = L_r + \left(\frac{n}{2} - Cf_b\right)\frac{w_r}{f_r}$$

If n is even, then the median is the average of the $(n/2)^{th}$ and $((n/2)+1)^{th}$ observations. That is

$$\begin{aligned} md(y) &= \frac{1}{2}\left[L_r + \left(\frac{n}{2} - Cf_b - \frac{1}{2}\right)\frac{w_r}{f_r} + L_r + \left(\frac{n}{2} + 1 - Cf_b - \frac{1}{2}\right)\frac{w_r}{f_r}\right] \\ &= \frac{1}{2}\left[2L_r + 2\left(\frac{n}{2} - Cf_b - \frac{1}{2} + \frac{1}{2}\right)\frac{w_r}{f_r}\right] = L_r + \left(\frac{n}{2} - Cf_b\right)\frac{w_r}{f_r} \end{aligned}$$

Therefore, the median is unchanged whether the number of observations is odd or even, and is given as

$$md\,(y) = L_r + \left(\frac{n}{2} - Cf_b\right)\frac{w_r}{f_r}$$

From Theorem 1, the mean absolute deviation about the median, when n is odd is given as

$$MAD\,(md) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}|y_{ij} - md(y)| = \frac{1}{n}(a-b) = \frac{1}{n}\left(T_k - 2b - md(y)\right)$$

The sum of observations before the median, b, is given as

$$
\begin{aligned}
b &= \sum_{i=1}^{r-1}\sum_{j=1}^{f_i} y_{ij} + \sum_{j=1}^{\frac{n-1}{2}-Cf_b} y_{rj} = T_{r-1} + \sum_{j=1}^{\frac{n-1}{2}-Cf_b}\left(L_r + \frac{w_r}{2f_r}(2j-1)\right) \\
&= T_{r-1} + \frac{1}{2}\left(n-1-2Cf_b\right)L_r + \frac{w_r}{2f_r}\left[\left(\frac{n-1}{2}-Cf_b\right)\left(\frac{n-1}{2}-Cf_b+1\right)\right. \\
&\quad \left. -\left(\frac{n-1}{2}-Cf_b\right)\right] \\
&= T_{r-1} + \frac{1}{2}\left(n-2Cf_b-1\right)L_r + \frac{w_r}{2f_r}\left(\frac{n-1}{2}-Cf_b\right)^2 \\
&= T_{r-1} + \frac{1}{2}\left(n-2Cf_b-1\right)L_r + \frac{w_r}{8f_r}(n-2Cf_b-1)^2 \\
&= T_{r-1} + \frac{1}{2}\left(n-2Cf_b-1\right)L_r + \frac{w_r}{8f_r}\left[(n-2Cf_b)^2+1-2(n-2Cf_b)\right]
\end{aligned}
$$

Therefore

$$
2b = \left(2T_{r-1} + \left(n-2Cf_b-1\right)L_r + \frac{w_r}{4f_r}\left[(n-2Cf_b)^2+1-2(n-2Cf_b)\right]\right)
$$

and

$$
\begin{aligned}
MAD(md) &= \frac{1}{n}\left(T_k - 2b - L_r - \left(\frac{n}{2}-Cf_b\right)\frac{w_r}{f_r}\right) \\
&= \frac{1}{4nf_r}\left(4(T_k-2T_{r-1})f_r - 4\left(n-2Cf_b\right)L_rf_r\right. \\
&\quad -\left[(n-2Cf_b)^2+1-2\left(n-2Cf_b\right)+2\left(n-2Cf_b\right)\right]w_r \\
&= \frac{1}{4nf_r}\left(4(T_k-2T_{r-1})f_r - 4\left(n-2Cf_b\right)L_rf_r - \right. \\
&\quad \left.\left[(n-2Cf_b)^2+1\right]w_r\right)
\end{aligned}
$$

From Theorem 1, when n is even,

$$
MAD\left(md\right) = \frac{1}{n}(T_k - 2b)
$$

In the case where n is even, the median is the average of the (n/2)th and the (n/2+1)th observations. This implies that the median is greater than (or equal to, in case there is a tie) the (n/2)th and less than (or equal to) the (n/2+1)th observation. Hence the (n/2)th observation must be included in the sum of observations before the median. Hence,

$$b = \sum_{i=1}^{r-1}\sum_{j=1}^{f_i} y_{ij} + \sum_{j=1}^{\frac{n}{2}-Cf_b} y_{rj} = T_{r-1} + \sum_{j=1}^{\frac{n}{2}-Cf_b}\left(L_r + \frac{w_r}{2f_r}(2j-1)\right)$$

$$= T_{r-1} + \frac{1}{2}(n-2Cf_b)L_r + \frac{w_r}{8f_r}\left[(n-2Cf_b)(n-2Cf_b+2) - 2(n-2Cf_b)\right]$$

$$= T_{r-1} + \frac{1}{2}(n-2Cf_b)L_r + \frac{w_r}{8f_r}(n-2Cf_b)^2,$$

$$2b = 2T_{r-1} + (n-2Cf_b)L_r + \frac{w_r}{4f_r}(n-2Cf_b)^2$$

Therefore,

$MAD\,(md) =$

$$\frac{1}{n}(T_k - 2b) = \frac{1}{4nf_rn}\left(4(T_k - 2T_{r-1})f_r - 4(n-2Cf_b)L_rf_r - \left[(n-2Cf_b)^2\right]w_r\right)$$

Thus

$$MAD\,(md) = \begin{cases} \frac{1}{4nf_r}\left[4(T_k - 2T_{r-1})f_r - 4(n-2Cf_b)L_rf_r - \right. \\ \qquad \left.\left((n-2Cf_b)^2 + 1\right)w_r\right],\ n\ odd \\ \\ \frac{1}{4nf_r}\left[4(T_k - 2T_{r-1})f_r - 4(n-2Cf_b)L_rf_r - \right. \\ \qquad \left.\left((n-2Cf_b)^2\right)w_r\right],\ n\ even \end{cases}$$

where

$$T_s = \sum_{i=1}^{s}\sum_{j=1}^{f_i} y_{ij} = \sum_{i=1}^{s} m_if_i, \quad s = 1, 2, \ \ldots, r-1, r, \ldots,\ k$$

It is noted that

$$Cf_b = \sum_{i=1}^{r-1} f_i$$

The MAD(md) could be expressed in a more compact form with the aid of the indicator function defined in Theorem 1. That is,

$MAD(md) =$

$$\frac{1}{4nf_r}\left[4\left(T_k - 2T_{r-1}\right)f_r - 4\left(n - 2Cf_b\right)L_r f_r - \left(\left(n - 2Cf_b\right)^2 + I_n\right)w\right]$$

■

The mean, median, variance, and mean absolute deviation about the median are quite useful descriptive statistics in practice. More importantly, the MAD(md) as specified in Theorem 3 (equation (9)) provides a novel approach to computing the mean absolute deviation about the median for grouped data, which was previously not possible. This is possible because of the de-grouping of the frequency distribution by equation (5). The MAD(md) is a preferred measure of dispersion for skewed distributions (Bonnet and Seier, 2003).

The formulation of a de-grouping scheme for grouped frequency distributions opens up such data to greater statistical analysis. Ordinarily, a grouped frequency distribution is not amenable to tests of hypothesis and other analysis on the distributional properties of the data. However, with this approach, goodness-of-fit tests could be carried out to ascertain whether the data emanated from a specified probability distribution with given parameters. Tests of hypothesis on the estimates of the parameters could also be conducted.

The two sample Kolmogorov-Smirnov (KS) test is a good tool to test the goodness-of-fit of the proposed approximation. The Kolmogorov-Smirnov test is based on the empirical cumulative distribution function, and is more powerful than the Chi-square test as a measure of goodness of fit (Brazauskas and Serfling, 2003) especially for continuous distributions.

Other important methods of assessing the performance of the approximation are the quantile-quantile (Q-Q) probability plot, the mean absolute percentage error (MAPE) and the mean squared error (MSE) (Brazauskas and Serfling, 2003; Khair *et. al*, 2017).

The one-sample Kolmogorov-Smirnov test can be used to test whether a grouped frequency distribution follows a specified probability distribution after de-grouping the data. The test statistic D(n) is given as

$$D\left(n\right) = \underset{all\ x}{\mathrm{Sup}}\left|\mathrm{F_n}\left(\mathrm{x}\right) - \mathrm{F(x)}\right|$$

where $F_n(x)$ is the sample cumulative distribution function and $F(x)$ is the distribution function of the hypothesized probability distribution and the supremum is the least upper bound of the differences, that is, the largest absolute value of the differences between the sample and the population distribution functions.

The Kolmogorov-Smirnov statistic has a distribution under the null hypothesis that is independent of the distribution function $F(x)$ (Conover, 1971; Durbin, 1973). The null hypothesis of the data emanating from the specified probability distribution is rejected if the statistic D(n) falls within the critical region of the test.

The two-sample Kolmogorov-Smirnov test is based on the test statistic D, given as

$$D = \underset{all\ x}{\text{Sup}} \left| F_m\left(x\right) - G_n(y) \right|$$

where $F_m\left(x\right)$ is the sample distribution function of the first sample of size m and $G_n(y)$ is the sample distribution function of the second sample.

Values of the asymptotic rejection limits of the test for pre-specified confidence levels are available in tables and the test could be easily carried out using various statistical packages, with the accompanying p-value computed. Values of the test statistic D closer to zero show a greater fit between the two datasets.

It is necessary that the proposed method should be robust to outliers. These outliers or extreme values could disproportionately shift the estimates of parameters like the mean and variance. In practice, a grouped frequency distribution covers all the observations, including outliers. The proposed approximation will thus be able to capture outliers and the MAD(md) is not significantly affected by outliers because it is based on the median which has the highest possible breakdown point and therefore serves as a robust measure of the variability in the distribution.

3. Results and Discussion

In order to examine the performance of the proposed method of de-grouping and the derived formulas, two ungrouped datasets culled from Brazauskas and Serfling (2003) were analyzed. The datasets were the 1977 wind catastrophe loss data and the 1975 Norwegian fire claims data. Furthermore, drawing from the distributional and parameter choices of Chen and Miljkovic (2019), simulated data were drawn from the Normal, log-normal, gamma and Weibull distributions, respectively using the R statistical package. The choice of distributions makes it possible to assess the performance of the proposed method for symmetric distributions as well as skewed distributions.

Random samples of sizes 300 and 1000 were drawn from the specified distributional choices. The samples were then grouped into classes and estimates of the mean, median, standard deviation, and mean absolute deviation about the median were obtained. The choice of the number of intervals, the interval width and the starting point for the grouped data were based on Sturges' rule and the range of the distribution. In most cases, the interval widths were equal, although in cases with empty classes, such intervals with zero entries were collapsed to the next non-empty interval to prevent discontinuities or jumps in the distribution. The first interval was split to two intervals whenever it had more than 50% of the observations so as to be able to compute the estimate of the median.

A comparison of the proposed method was made with the method of Brazauskas and Serfling (2003). The two methods were evaluated on the basis of the Kolmogorov-Smirnov test, the mean absolute percentage error (MAPE) and the mean squared errors (MSE) of the two approximations. The de-grouping approach of Brazauskas and Serfling (2003), called the Brazauskas-Serfling (BS) method approximates j^{th} observation in the i^{th} class by the formula

$$y_{ij} = \left(1 - \frac{j}{f_i + 1}\right) L_i + \frac{j}{f_i + 1} U_i = L_i + \left(\frac{j}{f_i + 1}\right) w_i$$

where f_i is the frequency of the class and (L_i, U_i) are the lower and upper limits, respectively of the class and w_i is the class width. While the proposed method uniformly allocates the f_i observation to the midpoint of each sub-interval, the BS method allocates the f_i observations into $(f_i + 1)$ sub-intervals. The choice of the denominator, $(f_i + 1)$ by Brazauskas and Serfling (2003) is not clear. The estimate of the sample mean, variance, median and mean absolute deviation about the median can be obtained via the methodology provided by Theorems 1 and 2. The results of the analysis of the two datasets from Brazauskas and Serfling (2003) are presented in Table 1, while the results from the simulations are presented in the Appendix.

Table 1: Results from the analysis of 1977 wind data and 1975 Norwegian fire losses data using the Brazauskas-Serfling (BS) method and the proposed de-grouping method

Data	Method	$\overline{x}$	Md	SD	MAD(md)	MSE	MAPE	KS Test
	Ungrouped	9.22	5.00	10.11	6.57	-	-	-
Wind	BS	8.97	4.33	9.33	5.91	1.27	12.78	0.3
	Proposed	8.97	4.25	9.33	5.92	1.28	12.60	0.3
	Ungrouped	2018	911	4866	1368	-	-	-
Fire	BS	2215	898	4477	1545	2043948	5.56	0.099
	Proposed	2215	895	4494	1545	2067640	5.51	0.099

Source: Brazauskas and Serfling (2003).

From Table 1, it is observed that the estimates of the mean for both the Brazauskas-Serfling and the proposed method were close to the estimate of the ungrouped data. It is noted that both the wind and fire data are both positively-skewed with the mean greater than the median in both datasets. There were also possible outliers in both datasets, with greater number of outliers in the fire data, as revealed by Brazauskas and Serfling (2003). Both the BS and proposed method yielded estimates that were not far apart. The Kolmogorov-Smirnov test also showed that there was no significant difference between both the ungrouped data and the approximations (BS and proposed). The variance (and standard deviation) using the proposed method produced marginally higher estimates compared to the BS method. Finally, it is seen from Table 1 that the MAPE and MSE were quite small for the wind data and much larger for the fire data. The large MSE value could indicate the present of outliers in the data.

It is difficult to have all the class widths equal in the presence of outliers, as this would yield many empty classes. It is also noted that grouping outliers with non-outliers significantly alters the shape of the distribution and may lead to type-1 error, in rejecting the null hypothesis when it ought not to be rejected. The proposed approximation has made it possible to estimate the MAD(md) for grouped data, in a manner that takes into consideration the individual observations. The estimate of the MAD(md) has also been extended to the BS approximation by Brazauskas and Serfling (2003).

Estimates of the MAD(md) were smaller than the corresponding estimates of the standard deviation in all cases, thus validating the Lyapounov corollary that the MAD(md) is always less than or equal to the standard deviation for all distributions (Pham-Gia and Hung, 2001). The MAPE and MSE values for the simulations, presented in the Appendix, were generally quite small, indicating a good fit by both the BS and proposed de-grouping methods for the Normal, gamma and Weibull distributions, while the corresponding values for the log-normal distribution were very high, indicating the presence of outliers. The approximations improved substantially as the sample size increased in all the cases. The Kolmogorov-Smirnov test statistics obtained in comparing the two de-grouping methods with the ungrouped data revealed a close fit with the ungrouped data, as none of the results of the simulated data was significant. The values of the KS test statistic which are closer to zero reflect a better fit than values that are not close to zero. The normal and Weibull distributions were well fitted by both methods of de-grouping, while the gamma and log-normal distributions showed higher values of the KS test statistic.

The method of de-grouping a grouped data opens up the possibility of carrying out various tests of hypotheses on grouped data, which otherwise, would not have been possible without de-grouping. This new computational approach in obtaining MAD(md) for grouped frequency distributions opens up the field to capture variability in skewed distributions and in cases where there are extreme values (outliers), as the MAD(md) is sufficiently impervious to outliers. A close scrutiny of the goodness-of-fit metrics used showed that the proposed method competed evenly with the Brazauskas-Serfling method. This is clearly visible from the results in Table 1 and in the Appendix, as the values of the MSE, MAPE and the KS statistic were the same in most cases. There was a seemingly even distribution of the number of times the proposed method outperformed the BS method and vice versa, hence there was no significant difference between both methods. With this approach, much information about the raw data is preserved, as in practice, the data may have been collected in grouped form.

4. Conclusion

The modified approach to computing descriptive statistics for grouped frequency distributions as presented in this paper enables a de-grouping of the grouped data and therefore provides an alternative way of obtaining better approximations of the population parameter without much loss of information as a result of grouping. The proposed method compared favourably with the method specified by Brazauskas and Serfling (2003), as the values of the MSE, MAPE and KS statistic only exhibited marginal differences. It also makes the grouped frequency distribution to be amenable to tests of hypotheses on the distributional properties of the grouped data. The proposed mean absolute deviation about the median, MAD(md) presents an intuitive way of computing the estimate without making use of the absolute operator, and is quite useful in measuring dispersion especially in grouped data that is skewed, which is very common in practice. In order to enhance the utility of estimates of grouped frequency distributions, it is suggested that the data be de-grouped so that more definitive tests of hypothesis and inference could be made on the data beyond conclusions drawn via a graphical view of the data through its histogram.

Acknowledgement

The authors acknowledge the inputs of the anonymous reviewers whose contributions enhanced the quality of the article.

References

Adegboye, O. S. (2009). Descriptive statistics for students, teachers and practitioners. Olad Publishers, Nigeria.

Arachchige, C. N. P. G. and Prendergast, L. A. (2019). Confidence intervals for median absolute deviations. arxiv:1910.00229v4 (math.st), www.arxiv.org

Bertrand, P. and Goupil, F. (2000). Descriptive statistics for symbolic data: exploratory methods for extracting statistical information from complex data. Edited by: H. H. Bock & E. Diday, Springer-Verlag, Berlin, pp. 103-124.

Billard, L. and Diday, E. (2006). Descriptive statistics for interval-valued observations in the presence of rules. Computational Statistics, **21**: 187-210.

Bluman, A. G. (2012). Elementary statistics: a step by step approach. McGraw-Hill, New York.

Bonnet, D. G. and Seier, E. (2003). Confidence intervals for mean absolute deviations. The American Statistician, **57**(4): 233-236.

Boos, D. D. and Brownie, C. (2004). Comparing variances and other measures of dispersion. Statistical Science, **19**(4): 571-578.

Brazauskas V. and Serfling R. (2003). Favorable estimators for fitting pareto models: A study using goodness-of-fit measures with actual data. ASTIN Bulletin. 33(2), 365-381.

Chen, Y. and Miljkovic, T. (2019). From grouped to de-grouped data: a new approach in distribution fitting for grouped data. Journal of Statistical Computation and Simulation, **89**(2): 272-291.

Choulakian, V. and Abou-Samra, G. (2020). Mean absolute deviations about the mean, the cut norm and taxicab correspondence analysis. Open Journal of Statistics, **10**: 97-112.

Conover, W. J. (1971). Practical nonparametric statistics. John Wiley & Sons, New York.

Crafford, G. (2007). Statistical analysis of grouped data. Ph.D. Thesis, University of Pretoria.

Durbin, J. (1973). Distribution theory for tests based on the sample distribution function. SIAM, Philadelphia.

Gupta, S. C. (2011). Fundamentals of statistics. Himalaya Publishing House, India.

Habib, E. A. E. (2012). Mean Absolute Deviation about median as a tool for explanatory data analysis. IJRRAS, **11**(3), 517-523.

Heitjan, D. F. (1989). Inference from grouped continuous data: a review. Statistical Science, **4**(2): 164-183.

Khair, U., Fahmi, H., Al Hakim, S. and Rahim, R. (2017). Forecasting error calculation with mean absolute deviation and mean absolute percentage error. IOP Conf. Series: Journal of Physics: Conf. Series **930**, 012002.

Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L. (2013). Detecting outlliers: do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, **49**: 764-766.

Lohaka, H. O. (2007). Making a grouped-data frequency table: development and examination of the iteration algorithm. Unpublished Ph.D. Thesis at the Department of Educational Studies, Ohio State University.

Miner, J. R. (1934). The variance of a grouped frequency distribution when the mean of each class is taken as centering point. Human Biology, **6**(3): 561-563.

Pham-Gia, T. and Hung, T. L. (2001). The mean and median absolute deviations. Mathematical and Computer Modelling, **34**: 921-936.

Appendix

Simulation results

Distribution $N(\mu, \sigma)$	Sample size	Method	$\overline{x}$	Md	SD	MAD(md)	MSE	MAPE	KS Test
N(9,1)	300	Ungrouped	8.99	8.98	0.98	0.77	-	-	-
		BS	8.99	8.96	1.00	0.80	$<$0.002	0.345	0.027
		Proposed	8.99	8.96	1.00	0.77	$<$0.002	0.334	0.027
	1000	Ungrouped	9.0	9.0	0.98	0.79	-	-	-
		BS	9.0	9.0	0.99	0.80	$<$0.001	0.17	0.013
		Proposed	9.0	9.0	0.99	0.80	$<$0.001	0.17	0.012
LN(9,1)	300	Ungrouped	12080	8711	12091	7992	-	-	-
		BS	12136	8710	11963	8028	361114	6.68	0.043
		Proposed	12136	8678	11972	8029	343976	6.95	0.043
	1000	Ungrouped	12707	8735	13378	8424	-	-	-
		BS	12775	8739	13441	8500	121326	6.29	0.035
		Proposed	12775	8729	13446	8500	113259	6.37	0.035
Gam(0.8,2)	300	Ungrouped	0.391	0.204	0.48	0.313	-	-	-
		BS	0.412	0.212	0.47	0.310	0.001	17.08	0.103
		Proposed	0.412	0.211	0.47	0.310	0.002	16.64	0.103
	1000	Ungrouped	0.40	0.22	0.46	0.31	-	-	-
		BS	0.40	0.23	0.45	0.29	$<$0.001	3.57	0.018
		Proposed	0.40	0.23	0.45	0.29	$<$0.001	3.62	0.018
Wei(1,8)	300	Ungrouped	8.72	5.68	8.78	6.08	-	-	-
		BS	8.91	5.94	8.82	6.15	0.112	5.50	0.033
		Proposed	8.91	5.91	8.82	6.19	0.120	5.45	0.037
	1000	Ungrouped	8.21	5.45	8.26	5.83	-	-	-
		BS	8.41	5.42	8.28	5.81	0.088	7.89	0.042
		Proposed	8.41	5.41	8.28	5,82	0.093	7.77	0.042