Mathematics and Statistics 8(2): 201-210, 2020
DOI: 10.13189/ms.2020.080216 http://www.hrpub.org

Improved Frequency Table with Application to Environmental Data

Mohammed M. B.[1,2,*], **Adam M. B.** [2,3], **Zulkafli H. S.** [2], **Ali N.** [2]

[1] Mathematics and Computer Science, Federal University of Kashere, Nigeria

[2] Department of Mathematics, Universiti Putra Malaysia, Malaysia

[3] Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia

Received November 15, 2019; Revised February 12, 2020; Accepted February 18, 2020

Abstract This paper proposes three different statistics to be used to represent the magnitude observations in each class when estimating the statistical measures from the frequency table for continuous data. The existing frequency tables use the midpoint as the magnitude of observations in each class, which results in an error called grouping error. Using the midpoint is due to the assumption that the observations in each class are uniformly distributed and concentrated around their midpoint, which is not always valid. In this research, construction of the frequency tables using the three proposed statistics, the arithmetic mean, median, and midrange and midpoint are respectively named, Method 1, Method 2, Method 3, and the Existing method. The four methods are compared using root-mean-squared error (RMSE) by performing simulation studies using three distributions, normal, uniform, exponential distributions. The simulation results are validated using real data, Glasgow weather data. The findings indicated that using the arithmetic mean to represent the magnitude of observations in each class of the frequency table leads to minimal error relative to other statistics. It is followed by using the median, for data simulated from normal and exponential distributions, and using midrange for data simulated from uniform distribution. Meanwhile, in choosing the appropriate number of classes used in constructing the frequency tables, among seven different rules used, the freedman and Diaconis rule is the recommended rule.

Keywords Frequency Table, Statistical Measures, Midpoint, Number of Classes.

1 Introduction

A data point can only be important if it is considered along with other observations in a frequency table [26]. Also, raw data do not display any meaningful representation unless if they are organized in a systematic form. The raw data are divided into classes of suitable sizes, showing observations together with their corresponding frequencies. When a data set is systematically organized in this form, the process is referred to as a frequency table [16]. The classes are constructed through which each data value can fall into exactly one class. A frequency table displays data along with the midpoint, cumulative frequency, and the class boundary [4].

The frequency table plays an important role in statistics. A well-organized frequency table makes possible a detailed analysis of the structure of the population with respect to a given feature. Therefore, the partitions into which the population breaks down can be obtained, and also the nature of the distribution of the elements of the population with respect to a particular characteristic can be known. For instance, to know whether the distribution is normal or skewed or the degree of concentration of the elements. So also, various statistical measures can be computed, such as the range, the mean, the measure of deviations from the average value, the coefficient of skewness of the frequency table, and the measure of kurtosis. Another significant function of the frequency table is, it serves as a bridge between raw data and a histogram. The frequency table also facilitates the construction of a cumulative frequency curve, ogive, and frequency polygon. An important to mention is the frequency table aid careful comparison of data sets[9].

Though grouping is unavoidable, especially when the data set is large, the process can lead to a considerable amount of errors when compared to the original data. When computing the statistical measures, such as the mean, mean deviation, variance, standard deviation, coefficient of skewness, coefficient of kurtosis, from the existing frequency table for continuous data, the observations in each of the classes are represented by a mere value, the value of the midpoint of the classes, which results to error known as grouping error. Various approaches to minimizing this error were suggested by different researchers in the literature. One of the approaches is the use of a correction formula to minimize the error. The most

used correction formula is Sheppard's correction, which was due to William Fleetwood Sheppard in 1898 [1, 2, 17, 28, 29]. This adjustment formula has contributed immensely to minimizing the grouping error, though it works for normal data and is applied to only even powers of moments. It is assumed that the odd powers moments are not affected by the grouping error. Following this several researches on correction for grouping error have emerged; such as the Canning[15], Davies[10] , Baten[27] Jones[13], Dwyer[22], Pierce[14], Hald [2] and the most recent work by Di Nardo[8].

Furthermore, one of the two significant parameters, the number of classes and the class width, must be determined before constructing the frequency table. While the former describes the number of partitions of the data set, the later is the distance between lower and upper-class limits [18]. These two parameters are dependent on each other if one is known the other can be obtained as well. Determining the appropriate number of classes to be used in constructing a frequency table remains a long existing problem in statistics. Thus, different rules of choosing the number of classes were reported in the literature; however, none of the rules is perfect [3].

In minimizing the grouping error, this study considered a different approach, proposing three statistics, the mean, median, and midrange of observations in each class. Meanwhile, the appropriate rule among the seven rules used in choosing the number of classes will be recommended.

2 Construction of the Frequency Table for Continuous Data

The first step in constructing the frequency table is to determine the number of classes or the class width. Thus, the table is created in such a way that all data points fall into exactly one of the possible classes. The important points to note when dividing continuous data into classes are, the classes should be mutually exclusive and exhaustive. It should also be neither too small nor too large, and preferably the classes should be of equal width, though sometimes unequal width must be used.

Mostly, the components of the frequency table are the class limits, class boundaries, the midpoint, frequency, cumulative frequency. The class limits are the pairs of numbers written in the column of class intervals. Meanwhile, the class boundaries are the values halfway between the upper limit of one class and the lower limit of the next class. When computing statistical measures, such as the mean, standard deviation, mean deviation, coefficient of skewness, a component midpoint is needed. The midpoint is the average of the upper and lower class limits. It is also the center of bars on a histogram. Another component which represents the number of observations in each of the classes is the frequency (f). In a frequency table, the frequencies are usually written as $f_1, f_2, \cdots, f_k$ are the number of occurrences in the k class intervals. In a situation where the statistical investigation is concerned with the number or percentage of observations which are less than or greater than, a component, cumulative frequency (cf) is included. The cumulative frequency at a particular class is the total frequency up to the upper-class boundary of that class. if $f_1, f_2, \cdots, f_k$ are the respective frequencies of k classes in a frequency table, the cumulative frequencies of the classes are $f_1, f_1 + f_2, \cdots, f_1 + f_2 + \cdots + f_k$ [16].

Moreover, to construct the frequency table, we first need to determine the range of the data. It is then followed by choosing a suitable number of classes, calculating the class width, obtaining the lower limit of the first class, and lastly determining the class intervals [16, 23].

The class width is the distance between the lower and the upper-class interval of a given class. The choice of class width is directly related to the number of classes. This choice is preferable if all the classes are of equal width; nevertheless, unequal class width is compelled to be used in rare cases [19]. Mathematically, the class width can be defined as

$$w = \frac{R}{k}.$$

Where k is the number of classes, and R is the range of the data set.

The number of classes depends largely on the size of the data, relatively small data require a fewer number of classes, whereas data of bigger size need a substantial number of classes. More than enough number of classes can widely spread the data through which little advantage can be derived when compared with the raw data. Equally, the use of a fewer number of classes tends to group the data much, which can lead to the loss of information and render the shape of the distribution undetermined [26]. What is important is how the number of classes is chosen. The classes should be large enough to capture the major shape features of the distribution and small enough to retain good details in the presence of random sampling error [26]. Among the various rules for choosing the suitable number of classes or class width reported in the literature, in this research, only seven rules are considered (see Table 1).

Table 1. The Existing number bins and bin width rules

Rule	Formula	Terms	Comment
Sturges (1926) [12]	$k = 1 + \log_2 n$	n= sample size	Suitable for $n < 200$
Cohran (1954) [21]	$k = \sqrt{\frac{n}{8}}$	n = Sample size	Used because of its simplicity
Cencov (1962) [20]	$k = \sqrt[3]{n}$	n = Sample size	Used because of its simplicity
Doane (1976)[6]	$k = 1 + \log_2(n) + \log_2(\lambda)$	n = Sample size $\lambda = \sqrt{\frac{(n+1)\times(n+3)\times\sum_{i=1}^{n}(x_i-\bar{x})^2}{6\times(n-2)\times\sum_{i=1}^{n}(x_i-\bar{x})^2}}$.	Suitable for skewed data
Scott (1979)[7]	$w = \frac{3.49\hat{\sigma}}{\sqrt[3]{n}}$	n = Sample size w = Bin width $\hat{\sigma}$ = Sample standard deviation	Suitable for symmetric data
Freedman and Diaconis (1981)[5]	$w = \frac{2\times IQR}{\sqrt[3]{n}}$	n = sample size w = Bin width IQR = Interquartile Range	A robust bin width rule
Terrel and Scott (1985)[11]	$k \geq \sqrt[3]{2n}$	n = Sample size	Used because of its simplicity

Though Sturges and Doane were among the first researchers to devise a scientific approach to choosing a suitable number of classes, the validity of their works has been questioned by Hyndman [24]. That, Sturges used Binomial distribution to approximate normal distribution, which is inappropriate. Hyndman suggested two alternative rules, Scott[7] and Freedman and Diaconis [5] rules. Sturges rule equally chooses a better number of classes as the Scott's and Freedman and Diaconis rules when $n < 200$, but it does not work for bigger sample sizes [24].

Table 2. The Existing Frequency Table

Class	CI		CB		Freq	Cum	Midpoint,
Number	lower	upper	lower	upper	(f)	Freq(cf)	(x^*)
1	l_1	u_1	$l_1 - \frac{\delta}{2}$	$u_1 + \frac{\delta}{2}$	f_1	f_1	$\frac{l_1+u_1}{2}$
2	l_2	u_2	$l_2 - \frac{\delta}{2}$	$u_2 + \frac{\delta}{2}$	f_2	$\sum_{i=1}^{2} f_i$	$\frac{l_2+u_2}{2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	l_k	u_k	$l_k - \frac{\delta}{2}$	$u_k + \frac{\delta}{2}$	f_k	$\sum_{i=1}^{k} f_i$	$\frac{l_k+u_k}{2}$

Where k is the number of classes, δ is the measurement unit of the data set, CI is the class interval, l_c and u_c are the lower and upper-class intervals , CB is the class boundary, l_b and u_b are the lower and upper-class boundaries, f_i is the frequency of class i and cf is the cumulative frequency. $i = 1, 2, \cdots, k$. The statistic value $\frac{l_i+u_i}{2}$ is the midpoint of class i, it represent the magnitude of observations in that class, l_i and u_i are respectively the lower and upper limits of class i, $i = 1, 2, \cdots, k$.

2.1 Grouping error

Grouping error is the error introduced when raw data are grouped into a frequency table. Grouping assumes that the observations within the class interval occur at the midpoint, the frequencies are uniformly distributed within the classes, and that the observations are concentrated around the midpoint or that the arithmetic mean and the midpoint of the classes are equal [15]. These assumptions cannot always be correct.

Several studies were carried out on grouping error since 1873, but William Fleetwood Sheppard proposed the most popular method for symmetrically distributed data in 1898 [28], [29], [2]. One of the shortcomings of Sheppard's correction as pointed out by Davies(1929)[10] is that it removes the errors from frequency tables that are symmetrical, but it makes the case worst when applied to asymmetric data. The Sheppard's correction formula for computing the first four important moments are as follows;

$$M_1 = M_1^{'}, \tag{1}$$

$$M_2 = M_2^{'} - \frac{w^2}{12}, \tag{2}$$

$$M_3 = M_3^{'} - \frac{w^2}{4} M_1^{'}, \tag{3}$$

$$M_4 = M_4^{'} - \frac{w^2}{2} M_2^{'} + \frac{w^4}{240}, \tag{4}$$

where M_1, M_2, M_3 and M_4 are the corrected moments whereas $M_1^{'}$, $M_2^{'}$, $M_3^{'}$ and $M_4^{'}$ are the uncorrected moments and w is the class width.

In minimizing the grouping error, Canning[15] and Davies[10], considered using different approaches. Canning[15] proposed a procedure by assuming that the class intervals are uniform and in multiples of the units of which the original data were recorded. That, the limits of the class intervals be also assumed falling points halfway between the readings at which the variates were taken. Davies[10] devised a method which uses two assumptions. First, the frequencies as fixed and minimize the error in each class, and second, the frequencies represent an approximation to a type which may be discovered by smoothing or curve-fitting process.

In another study, Jones[13] proposed a formula which uses the sum as well as the number of observations in each of the classes of a frequency table. It was found that the correction formula gives a minimum error as compared with Sheppard's correction formula. Dwyer[22] argued that Sheppard's corrections are not practically good enough, and suggested a new grouping error correction method using single grouping. Dwyer's correction formula gives a much better-unbiased estimate as compared with the conventional Sheppard's correction formula [14]. Moreover, Pierce[14] used a similar approach and derived a correction formula using a single grouping of the existing frequency table for discrete data and discovered that the method results in additional precision compared to the conventional correction methods. Nevertheless, the way Pierce grouped the discrete frequency table is inappropriate. Discrete data are count observations and do not assume values within a continuous interval. This assumption may be applied to continuous data. Hald[2] research reviewed some methods of deriving correction for grouping. First, Thiele's (1838- 1910) correction for grouping error applied to Charlie's series, which used the Opperman (1817 - 1883) suggestions. Second, the Sheppard (1863 - 1936) corrections for moments of equal class intervals. Third, the Bruns's (1848 - 1919) derivation of the correction for grouping error, who obtained Charlie's series and explained its properties. Fourth, the Fisher (1890 - 1962) proof of the correction for grouping, which introduced the concepts of efficiency, sufficiency, consistency, and proved the theory of maximum likelihood estimation. Baten[27] extended the idea of grouping error to a situation which involved two variables grouped in a bivariate frequency table. The most recent research on correction for grouping error is the work of Di Nardo[8]. The research introduced a new version of Sheppard's corrections, using the umbral calculus concept. Indeed, Di Nardo has extended the concept of the correction for grouping to the multivariate case.

3 The Proposed Frequency Table

When dealing with continuous data, the mode is not a suitable measure of location. The possibility of having two or more data points with the same decimal points value is very rare.Usually, researchers approximate continuous data into two, three, or four decimal places based on their choice. Mostly, if the observations are not approximated, they cannot be the same. For instance, when measuring body mass index (BMI) of some individuals, if the BMI of the first and second are 25.75234, and 25.75325. These values can be approximated to two decimal places, to give the same value 25.75. However, the two values were not originally the same.

In the proposed frequency tables, the arithmetic mean, median, and midrange are used to substitute the midpoint. Though several other measures of location existed in the literature, the geometric mean, harmonic mean, quadratic mean, contraharmonic mean, midhinge, midmean, trimean, but the choice of mean, median, and midrange is because the mean is the suitable measure of location for normally distributed data, the median is the appropriate measure when dealing with skewed data, and midrange is also a good measure of location, in fact, it is a UMVU estimator for the mean of continuous uniform distribution. Table 3 presents the first part of the proposed frequency table, whereas Table 4 depicts the proposed statistics that represent the magnitude of observations in each class when estimating statistical measures from the frequency table.

Table 3. The Proposed Frequency Table (First part).

	CI		CB		Freq	Cum f
c	l_c	u_c	l_b	u_b	(f)	(cf)
1	l_1	u_1	$l_1-\frac{\delta}{2}$	$u_1+\frac{\delta}{2}$	f_1	f_1
2	l_2	u_2	$l_2-\frac{\delta}{2}$	$u_2+\frac{\delta}{2}$	f_2	$\sum_{i=1}^{2} f_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	l_k	u_k	$l_k-\frac{\delta}{2}$	$u_k+\frac{\delta}{2}$	f_k	$\sum_{i=1}^{k} f_i$

Table 4. The Three Methods Used to Represent the Magnitude of Observations in Each Class(Second part).

Mean (x_{me})	Median (x_{md})	Midrange (x_{mr})
x_{me_1}	x_{md_1}	$\frac{Min_1+Max_1}{2}$
x_{me_2}	x_{md_2}	$\frac{Min_2+Max_2}{2}$
$\vdots$	$\vdots$	$\vdots$
x_{me_k}	x_{md_k}	$\frac{Min_k+Max_k}{2}$

Where k represent the number of classes, x_{me_i} and f_i are respectively the mean and frequency of class i . The median of class i is denoted by x_{md_i} , it is equal to x_s if the number of observations in the class is odd and $\frac{x_s+x_{s+1}}{2}$ if the number of observations in the class is even. X_s is the ordered middle observation within the class. Also, Min_i and Max_i are respectively the smallest and largest observations in class i, and $\frac{Min_i+Max_i}{2}$ is the midrange of observations in class i, $i = 1, 2, \cdots, k$.

The statistical measures, such as measures of location, measures of variability, measures of shape, can be estimated from the frequency table. However, in this research, to know the best method in terms of minimum error, the measure of location, mean, is used. The estimate of the mean using the four methods is given in Table 5.

Table 5. The formulas used when computing the means of the samples from the frequency tables using the four methods.

Method	Mean(x_{me})	Median(x_{md})	Midrange(x_{mr})	Midpoint(x^*)
Statistical measure $(\bar{x})$	$\frac{\sum_{i=1}^{k} f_i x_{me_i}}{\sum_{i=1}^{k} f_i}$	$\frac{\sum_{i=1}^{k} f_i x_{md_i}}{\sum_{i=1}^{k} f_i}$	$\frac{\sum_{i=1}^{k} f_i x_{mr_i}}{\sum_{i=1}^{k} f_i}$	$\frac{\sum_{i=1}^{k} f_i x_i^*}{\sum_{i=1}^{k} f_i}$

3.1 The Root-Mean-Squared Error (RMSE)

The root-mean-square error (RMSE) is an important measure of the performance of an estimator. This estimator can be a model or a method used to estimate a statistic value of a sample or a population parameter. In other words, the RMSE is a measure of the accuracy of methods in estimating the actual values of a sample or population.

RMSE is always positive. However, sometimes a rare value zero can be obtained; which indicates a perfect model or method. Generally, a method with the smallest RMSE is the best. Mathematically, the RMSE is given by

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\vartheta}_i - \vartheta_i)^2}, \tag{5}$$

where m is number of samples, $\hat{\vartheta}_i$ is the estimated statistic value of sample i and ϑ_i is the actual statistic value of sample i, $i = 1, 2, \cdots, m$ [25].

In this article, the $\hat{\vartheta}$ are the estimated means from the frequency tables using the four methods: Mean, Median, Midrange and Midpoint. The formulas used in computing the estimates ($\hat{\vartheta}$) are given in Table 5. One hundred samples simulated from normal distributions were used to construct one hundred frequency tables using the four different methods. The means of hundred samples were estimated from the hundred constructed frequency tables ($\hat{\vartheta}$). Meanwhile, the raw arithmetic mean formula is used to compute the actual means (ϑ) of the samples.

Here, the grouping error is the difference between the actual mean and the mean obtained from the frequency table. That is,

$$\text{Grouping error} = \hat{\vartheta} - \vartheta.$$

Besides, to observe the performance of the four methods, the root-mean-square errors obtained were visualized using the

radar chart. A radar chart which is also called web chart, star chart, spider chart, cobweb chart, star plot, irregular polygon, polar chart, or Kiviat, is a chart that consists of a sequence of equal angular spokes, known as radii, with each spoke depicting one of the variables. The data size of a spoke is proportional to the magnitude of the variable for the data point relative to the highest magnitude of the variable across all data points. A line is usually drawn linking the data values for each spoke. The chart is used to display multivariate data in a two-dimensional chart form of three or more variables displayed on axes originating from the same point. The position and angle of the axes are usually not useful. Moreover, from the radar chart is quite easier to observe patterns in the data if the observations are organized in some non-random order [30].

4 Results

4.1 Simulation Studies

In order to compare the three methods proposed with the existing method and at the same time to evaluate the level of accuracy of the rules used in choosing the number of classes, we performed simulations of 100 different samples each of the sizes 50, 100, 500, 1000 and 10000 from three distributions. The first simulation study is from the normal distribution with parameters $\mu = 20$ and $\sigma = 3$. In the second study, the uniform distribution with parameters $a = 1$ and $b = 100$ is used . While exponential distribution with rate, $\lambda = 0.025$ is used in the third simulation study. The choice of these three distributions is because we want to assess the performance of the four methods and the rules used in choosing the number of classes when the data is symmetric, uniform and asymmetric.

The seven existing rules, Sturges (1926)[12], Cohran (1954)[21], Cencov (1962)[20] ,Doane (1976)[6], Scott (1979)[7], Freedman and Diaconis (1981)[5], and Terrel and Scott (1985)[11] rules, are used to determine the suitable number of classes. Furthermore, to estimate the samples means from the constructed frequency tables, the four methods are used to represent the magnitude of observations in each of the classes. Using the root-mean-squared error (RMSE), the efficacy of the four methods used to estimate the mean from the constructed frequency tables as well as the rules used in choosing the number of classes was evaluated.

4.2 Normal Distribution

Table 6 shows the suitable number of classes suggested by the seven rules for the 100 different samples of sizes 50, 100, 500, 1000, and 10000 simulated from the normal distribution. The Sturges, Terrell and Scott (TS), Cencov, and Cohran rules suggested, the same number of classes, because the rules depend only on the sample size. In the meantime, the Scott rule, which depends on sample size and standard deviation, the Freedman and Diaconis rule (FD), which depends on sample size and interquartile range, and the Doane rule, which depends on sample size and skewiness, recommends a different number of classes. Therefore, ranges of the number of classes for the 100 samples are given for Scott, FD, and Doane rules.

In addition, Table 7 shows the root-mean-squared error (RMSE) of the four methods used to estimate the mean from the frequency tables. Method 1, the method that uses the mean records the lowest RMSE score in all five different sample sizes. It is followed by method 2 for samples of sizes 50, Method 3 for samples of sizes 100, the existing method for samples of sizes 500, 1000, and 10000. Moreover, the seven rules used in choosing the number of classes give the minimum RMSE in method 1. However, if we consider their performance when the other three methods are used, for samples of sizes 50 Sturges gives the smallest RMSE in Method 2 and 3. The Freedman and Diaconis (FD) rule record the least RMSE in Method 3 and 4 for samples of sizes 100 and 500. It also has the smallest RMSE in Method 2, 3, and the existing method, for samples of sizes 10000. While, for the samples of sizes 1000, Scott rule gives the smallest RMSE when using Method 2 and 3. Moreover, for easy understanding, Figure 1 displays the radar chart of the ranks of the RMSE scores for the five different sample sizes. It can be observed, from the chart that Method 1, has the minimum ranks when compared with other methods, hence, the best method.

Table 6. The suitable number of classes (k) suggested by the seven rules for the data simulated from the normal distribution.

n	**Sturges**	**Scott**	**FD**	**Doane**	**TS**	**Cencov**	**Cohran**
50	7	$[4, \cdots, 6]$	$[5, \cdots, 10]$	7,8	5	4	4
100	8	$[6, \cdots, 9]$	$[7, \cdots, 12]$	8,9	6	5	5
500	10	$[12, \cdots, 28]$	$[16, \cdots, 41]$	11,12,13	10	8	10
1000	11	$[17, \cdots, 36]$	$[21, \cdots, 48]$	12,13	13	10	15
10000	15	$[45, \cdots, 91]$	$[54, \cdots, 122]$	15,16	28	22	45

Table 7. The RMSE of the Four Methods for Samples of Sizes 50, 100, 500, 1000 and 10000 Simulated from the Normal Distribution.

Sample size	Method	Sturges	Scott	FD	Doane	TS	Cencov	Cohran
50	Method 1	0.057911 [1]	0.057695 [1]	0.057877 [1]	0.057511 [1]	0.057644 [1]	0.020266* [1]	0.020266* [1]
	Method 2	0.096739 [3]	0.101296 [2]	0.102223 [4]	0.074301 [2]	0.118278 [2]	0.146351 [2]	0.146351 [2]
	Method 3	0.087822 [2]	0.122744 [3]	0.090662 [2]	0.088624 [4]	0.120778 [3]	0.155806 [3]	0.155806 [3]
	Existing	0.106875 [4]	0.157982 [4]	0.097963 [3]	0.087419 [3]	0.159778 [4]	0.196516 [4]	0.196516 [4]
100	Method 1	0.04021 [1]	0.041828 [1]	0.042095 [1]	0.041759 [1]	0.005304* [1]	0.005333 [1]	0.005333 [1]
	Method 2	0.073723 [4]	0.069290 [2]	0.066223 [4]	0.060832 [3]	0.072564 [2]	0.091581 [3]	0.091581 [3]
	Method 3	0.071074 [2]	0.072102 [3]	0.056082 [2]	0.058457 [2]	0.078482 [3]	0.086688 [2]	0.086688 [2]
	Existing	0.072147 [3]	0.077185 [4]	0.064771 [3]	0.065541 [4]	0.096878 [4]	0.113894 [4]	0.113894 [4]
500	Method 1	0.010821 [1]	0.013377 [1]	0.014574 [2]	0.010295 [1]	0.010821 [1]	0.001533* [1]	0.010821 [1]
	Method 2	0.029609 [4]	0.018966 [3]	0.020177 [4]	0.022160 [2]	0.029609 [4]	0.030794 [2]	0.029609 [4]
	Method 3	0.028579 [3]	0.022632 [4]	0.019980 [3]	0.026360 [4]	0.028579 [3]	0.031840 [3]	0.028579 [3]
	Existing	0.025364 [2]	0.017762 [2]	0.013748 [1]	0.025417 [3]	0.025364 [2]	0.034523 [4]	0.025364 [2]
1000	Method 1	0.000737* [1]	0.008761 [1]	0.012334 [2]	0.004518 [1]	0.003183 [1]	0.000820 [1]	0.007545 [1]
	Method 2	0.018293 [2]	0.012644 [3]	0.013097 [3]	0.017317 [2]	0.016714 [4]	0.020672 [4]	0.013507 [3]
	Method 3	0.019096 [3]	0.013274 [4]	0.014815 [4]	0.018369 [4]	0.015833 [3]	0.018802 [2]	0.015694 [4]
	Existing	0.020496 [4]	0.009732 [2]	0.007974 [1]	0.017519 [3]	0.015638 [2]	0.019227 [3]	0.012846 [2]
10000	Method 1	0.000027 [1]	0.000829 [1]	0.000765 [1]	0.000046 [1]	0.000316 [1]	0.000025* [1]	0.000826 [1]
	Method 2	0.005684 [2]	0.001613 [4]	0.001486 [4]	0.006864 [4]	0.003605 [4]	0.003811 [2]	0.002324 [3]
	Method 3	0.007160 [4]	0.001436 [3]	0.001140 [3]	0.006349 [2]	0.002356 [2]	0.004033 [4]	0.003468 [4]
	Existing	0.007115 [3]	0.001292 [2]	0.001021 [2]	0.006355 [3]	0.003417 [3]	0.004021 [3]	0.002120 [2]

Note: In the square brackets are the ranks of the RMSE.

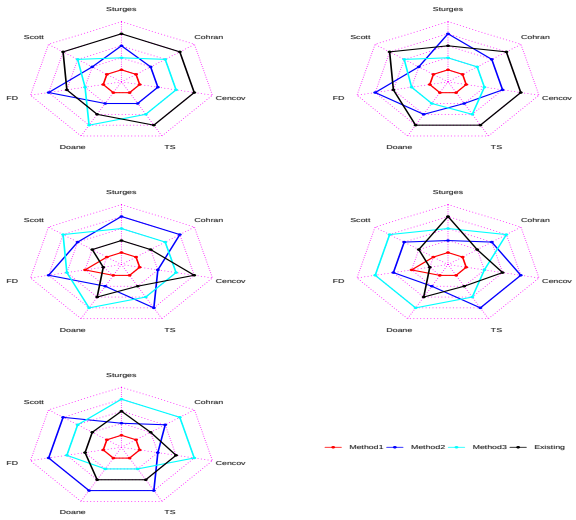

Figure 1. The Radar Charts of the RMSE of the Four Methods, for Samples of Sizes 50, 100, 500, 1000 and 10000 Simulated from the Normal Distribution.

4.3 Uniform Distribution

Using similar fashion, Table 8 presents the appropriate class numbers for the simulated data from the uniform distribution obtained using the seven rules. Meanwhile, Table 9 shows the RMSE of the four methods and the seven rules used.

This simulation study gives results similar to the previous one. Also, Method 1, maintain the least score of RMSE. Method 3 is followed for samples of 50, 100, and 1000 sizes, then Method 2 for samples of 500, 10000 sizes. Also, when using Method 1 the seven rules have the smallest RMSE, but if we take into account the performance of the rules across the methods, for samples of sizes 50, the Sturges rule recommended the number of classes that gives the smallest RMSE in Method 2 and 3. The Sturges and Doane rules record the smallest RMSE in all the four methods, for samples of sizes 100. While, for samples of sizes 500, the Doane rule gives the least RMSE in methods 2, 3, and the existing method.Moreover, for the samples of sizes 1000, Cohran rule gives the smallest RMSE in all the Methods, and in samples of sizes 10000 it also has the smallest RMSE when using, Method 2, 3 and the existing method.

Figure 2 displays the radar chart of the ranks of the RMSE for samples of the five different sizes. The chart shows that method 1 records the least ranks.

Table 8. The Number of Classes(k) Chose by the Seven Rules for the Data Simulated from the Uniform Distribution.

n	Sturges	Scott	FD	Doane	TS	Cencov	Cohran
50	7	$[4, \cdots, 5]$	$[3, \cdots, 6]$	7	5	4	4
100	8	5	$[4, \cdots, 6]$	8	6	5	5
500	10	8, 9	$[8, \cdots, 10]$	11	10	8	10
1000	11	10, 11	10, 11	12	13	10	15
10000	15	22	22, 23	15	28	22	45

Table 9. The RMSE of the Four Methods for Samples of sizes 50, 100, 500, 1000 and 10000 Simulated from the Uniform Distribution.

Sample size	Method	Sturges	Scott	FD	Doane	TS	Cencov	Cohran
50	Method 1	0.049081 [1]	0.092848 [1]	0.067505 [1]	0.049081 [1]	0.038060* [1]	0.099708 [1]	0.099708 [1]
	Method 2	0.243790 [3]	1.093514 [2]	0.952597 [2]	0.495126 [3]	0.810484 [3]	1.092939 [3]	1.092939 [3]
	Method 3	0.153801 [2]	1.096463 [3]	1.080164 [3]	0.397620 [2]	0.739084 [2]	1.085211 [2]	1.085211 [2]
	Existing	0.700079 [4]	1.510919 [4]	1.451138 [4]	0.697728 [4]	0.969902 [4]	1.509947 [4]	1.509947 [4]
100	Method 1	0.007521* [1]	0.010672 [1]	0.014553 [1]	0.007521* [1]	0.021975 [1]	0.010672 [1]	0.010672 [1]
	Method 2	0.367885 [3]	0.673191 [3]	0.649746 [3]	0.367885 [3]	0.516049 [3]	0.670856 [3]	0.670856 [3]
	Method 3	0.347173 [2]	0.665062 [2]	0.641788 [2]	0.347173 [2]	0.498449 [2]	0.664014 [2]	0.664014 [2]
	Existing	0.436788 [4]	0.831524 [4]	0.799014 [4]	0.436788 [4]	0.608316 [4]	0.833464 [4]	0.833464 [4]
500	Method 1	0.003176* [1]	0.004194 [1]	0.003802 [1]	0.004215 [1]	0.003176* [1]	0.004255 [1]	0.003176* [1]
	Method 2	0.131443 [2]	0.175432 [4]	0.171740 [4]	0.107205 [2]	0.131443 [2]	0.169392 [3]	0.131443 [2]
	Method 3	0.131725 [3]	0.159755 [2]	0.155988 [2]	0.116797 [3]	0.131725 [3]	0.164920 [2]	0.131725 [3]
	Existing	0.141898 [4]	0.170377 [3]	0.165421 [3]	0.119951 [4]	0.141898 [4]	0.177229 [4]	0.141898 [4]
1000	Method 1	0.001448 [1]	0.001213 [1]	0.001624 [1]	0.002158 [1]	0.002108 [1]	0.001654 [1]	0.000658* [1]
	Method 2	0.092141 [4]	0.092158 [4]	0.090136 [4]	0.081452 [4]	0.074851 [2]	0.097615 [4]	0.062805 [3]
	Method 3	0.087281 [2]	0.082404 [2]	0.087355 [2]	0.077027 [2]	0.077888 [3]	0.090041 [2]	0.062240 [2]
	Existing	0.088297 [3]	0.086649 [3]	0.088156 [3]	0.078865 [3]	0.082029 [4]	0.095109 [3]	0.066430 [4]
10000	Method 1	0.000514 [1]	0.000368 [1]	0.000368 [1]	0.000514 [1]	0.000151* [1]	0.000368 [1]	0.000204 [1]
	Method 2	0.018437 [2]	0.011880 [2]	0.011793 [2]	0.018437 [2]	0.010876 [4]	0.011881 [2]	0.005843 [2]
	Method 3	0.021504 [4]	0.014663 [4]	0.014596 [4]	0.021504 [4]	0.009314 [2]	0.014663 [4]	0.005993 [4]
	Existing	0.021358 [3]	0.014659 [3]	0.014582 [3]	0.021358 [3]	0.009474 [3]	0.014659 [3]	0.005975 [3]

Note: In the square brackets are the ranks of the RMSE.

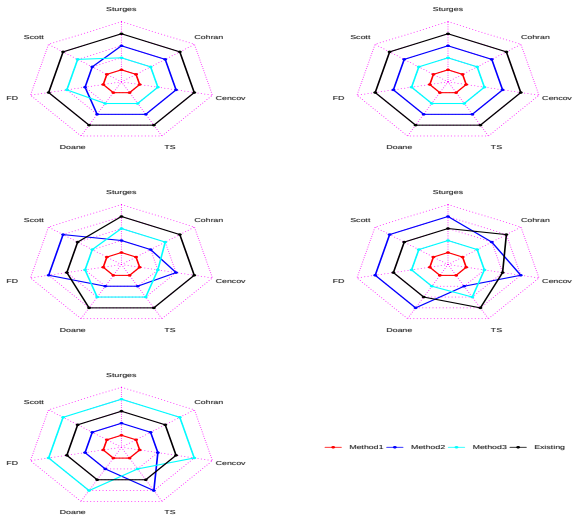

Figure 2. The Radar Charts of the RMSE of the Four Methods, for Samples of Sizes 50, 100, 500, 1000 and 10000 Simulated from the Uniform Distribution.

4.4 Exponential Distribution

Furthermore, Table 10 shows the suitable numbers of classes recommended by the seven rules for data simulated from the exponential distribution. Also, the root-mean-squared-error (RMSE) of the four methods are presented in Table 11. Method 1 remains the best. Nonetheless, the reason this method still has the smallest RMSE could be as a result of not considering the issue of outliers, which is one of the limitations of this research. Moreover, for the rules used in choosing the number of classes, for samples of sizes 50, the Doane's rule records the smallest RMSE in three different methods, Method 1, Method 2 and Method 3. The Freedman and Diaconis (FD) rule gives the least score of RMSE in Method 2, Method 3, and the existing, for samples of sizes 100 and 500. Meanwhile, for samples of sizes 1000 and 10000, Freedman and Diaconis(FD) rule still give the minimum RMSE score for all the four Methods. Also, the radar charts of the ranks of the RMSE shows that method 1 has the minimum ranks when compared with other methods (see Figure 3).

Table 10. The Number of Classes(k) Recommended by the Seven Rules for the Data Simulated from the Exponential Distribution.

n	Sturges	Scott	FD	Doane	TS	Cencov	Cohran
50	7	$[4,\cdots,8]$	$[6,\cdots,21]$	$[7,\cdots,9]$	5	4	4
100	8	$[6,\cdots,11]$	$[5,\cdots,21]$	$[8,\cdots,10]$	6	5	5
500	10	$[13,\cdots,40]$	$[18,\cdots,42]$	$[11,\cdots,13]$	10	8	10
1000	11	$[17,\cdots,32]$	$[26,\cdots,50]$	$13, 14$	13	10	15
10000	15	$[51,\cdots,85]$	$[74,\cdots,138]$	$16, 17$	28	22	45

Table 11. The RMSE of the Four Methods for Samples of Sizes 50, 100, 500, 1000 and 10000 Simulated from the Exponential Distribution.

Sample size	Method	Sturges	Scott	FD	Doane	TS	Cencov	Cohran
50	Method 1	0.088751	0.088724	0.593594	< 0.00001	0.087300	$< 0.00001^*$	< 0.00001
		[1]	[1]	[1]	[1]	[1]	[1]	[1]
	Method 2	1.329663	2.026773	1.368773	1.261132	2.556129	3.619815	3.619815
		[2]	[2]	[2]	[2]	[2]	[2]	[2]
	Method 3	1.980128	3.247367	1.431047	1.298217	4.382488	7.259218	7.259218
		[3]	[3]	[3]	[3]	[3]	[3]	[3]
	Existing	2.906695	4.700116	2.116476	2.251847	6.133464	9.936369	9.936369
		[4]	[4]	[4]	[4]	[4]	[4]	[4]
100	Method 1	0.511110	0.581858	0.516248	0.568054	0.430284^*	0.534515	0.534515
		[1]	[1]	[1]	[1]	[1]	[1]	[1]
	Method 2	1.203726	1.306914	0.712278	0.922734	2.099395	2.897495	2.897495
		[2]	[2]	[2]	[2]	[2]	[2]	[2]
	Method 3	2.25662	2.263749	0.954653	1.442936	3.954978	5.987659	5.987659
		[3]	[3]	[3]	[3]	[3]	[3]	[3]
	Existing	2.272672	2.746259	1.052818	1.751026	4.860833	7.387709	7.387709
		[4]	[4]	[4]	[4]	[4]	[4]	[4]
500	Method 1	0.148998^*	0.187286	0.187548	0.196858	0.148998^*	0.165178	0.148998^*
		[1]	[1]	[1]	[1]	[1]	[1]	[1]
	Method 2	1.094455	0.419038	0.247969	0.710515	1.094455	1.648015	1.094455
		[2]	[2]	[2]	[2]	[2]	[2]	[2]
	Method 3	2.092376	0.716381	0.342749	1.413313	2.092376	3.439360	2.092376
		[3]	[3]	[4]	[3]	[3]	[3]	[3]
	Existing	2.280397	0.796041	0.322373	1.531796	2.280397	3.722991	2.280397
		[4]	[4]	[3]	[4]	[4]	[4]	[4]
1000	Method 1	0.092356	0.066540^*	0.112030	0.072286	0.084139	0.085079	0.077988
		[1]	[1]	[1]	[1]	[1]	[1]	[1]
	Method 2	0.931771	0.234953	0.134111	0.643967	0.637880	1.103780	0.502711
		[2]	[2]	[2]	[2]	[2]	[2]	[2]
	Method 3	1.973141	0.451242	0.219974	1.293004	1.325596	2.413555	1.039976
		[3]	[3]	[4]	[3]	[3]	[3]	[3]
	Existing	2.045900	0.477719	0.185418	1.352693	1.379818	2.528024	1.089969
		[4]	[4]	[3]	[4]	[4]	[4]	[4]
10000	Method 1	0.016403	0.018378	0.014353^*	0.014979	0.014727	0.018313	0.018684
		[1]	[1]	[1]	[1]	[1]	[1]	[1]
	Method 2	0.849204	0.052226	0.026857	0.668600	0.244587	0.403888	0.093460
		[2]	[2]	[2]	[2]	[2]	[2]	[2]
	Method 3	1.767150	0.095575	0.037575	1.361900	0.480829	0.783398	0.175475
		[3]	[4]	[3]	[3]	[3]	[3]	[3]
	Existing	1.782639	0.094212	0.038001	1.376648	0.490831	0.799281	0.180382
		[4]	[3]	[4]	[4]	[4]	[4]	[4]

Note: In the square brackets are the ranks of the RMSE.

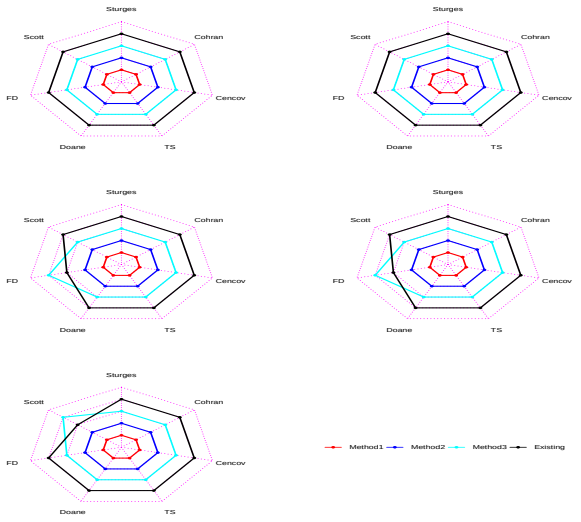

Figure 3. The Radar Charts of the RMSE of the Four Methods, for Samples of Sizes 50, 100, 500, 1000 and 10000 Simulated from the Exponential Distribution.

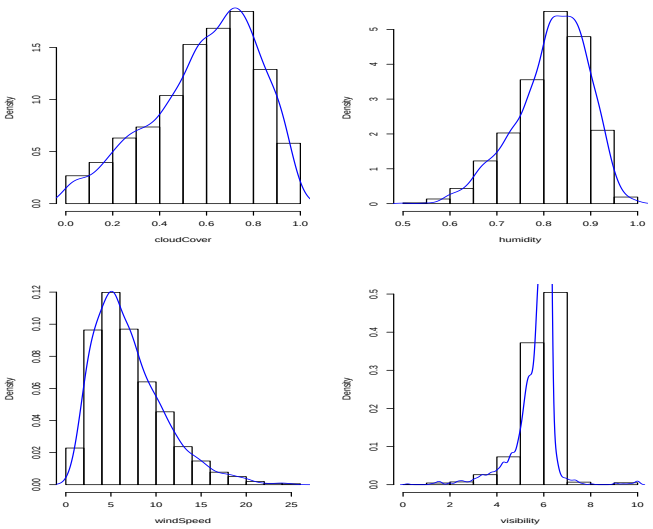

Figure 4. The Histogram Density Plots of the Glasgow Weather Data.

4.5 Application

To ascertain the performance of the four methods of representing the observations in each class using real data, we used Glasgow weather data obtained from the Kaggle website. The data contain daily measurements of five (5) variables, temperature, cloud cover, humidity, wind speed, and visibility. Each variable has 1,795 observations. . The Histogram density plots displayed in Figure 4 show that the data is asymmetric. So also, in constructing the frequency tables, the number of classes suggested by the seven rules are presented in Table 12. The RMSE displayed in Table 13 as well as the radar chart of the ranks of the RMSE, Figure 5 shows that method 1 is still the best, it has the smallest RMSE in all the seven different rules used in choosing the number of classes.

Table 12. The Number of Classes(k) Recommended by the Seven Rules for the Glasgow Weather Data.

	Sturges	**Scott**	**FD**	**Doane**	**TS**	**Cencov**	**Cohran**
Cloud Cover	12	16	19	13	16	13	19
Humudity	12	23	31	13	16	13	19
Wind Speed	12	22	30	13	16	13	19
Visibility	12	43	91	14	16	13	19

Table 13. The RMSE of the Four Methods using the Glasgow Weather Data.

	Sturges	Scott	FD	Doane	TS	Cencov	Cohran
Method 1	0.000018	0.000012	0.000141	0.000193	0.000009*	0.000193	0.000141
	[1]	[1]	[1]	[1]	[1]	[1]	[1]
Method 2	0.028248	0.0089985	0.000451	0.032727	0.0111119	0.020361	0.005301
	[2]	[2]	[2]	[2]	[2]	[2]	[2]
Method 3	0.037706	0.010353	0.003255	0.070322	0.026788	0.022346	0.022191
	[4]	[3]	[3]	[3]	[3]	[3]	[4]
Existing	0.035975	0.012270	0.004335	0.072084	0.028276	0.024420	0.020362
	[3]	[4]	[4]	[4]	[4]	[4]	[3]

Note: In the square brackets are the ranks of the RMSE.

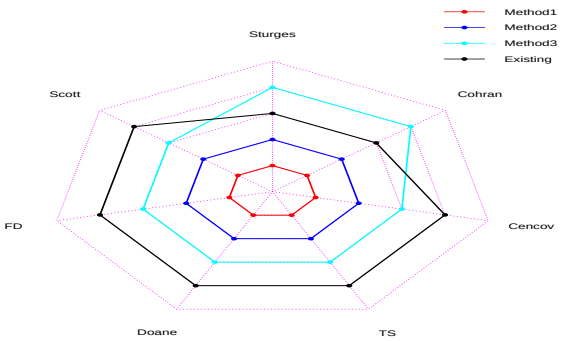

Figure 5. The Radar Chart of the RMSE of the Four Methods Using the Glasgow Weather Data.

5 Discussion and Conclusion

5.1 Discussion

The results using the normal data showed in all the five different sample sizes, method 1, the method that uses the mean to

represents the observations in each of the classes recorded the least score of RMSE and hence the best method. It is followed by the existing method. Though in choosing the appropriate number of classes, all the seven rules gave the minimum RMSE when using method 1. But if other methods are considered, the Freedman and Diaconis rule posted the smallest RMSE.

Furthermore, using samples from the uniform distribution, the result indicated that method 1 outperformed the other methods. The method with the second least RMSE score is Method 3, the method that used the midrange. In assessing the rules used in choosing the number of classes, the rule that chooses the number of classes for the method with the smallest RMSE is considered. The Sturges rule recorded the least RMSE for samples of sizes 50 and 100, then the Doane's rule for samples of sizes 50, 100 and 500, and Cohran rule recorded the least score of RMSE for samples of sizes 1000 and 10000.

The study using skewed data, data simulated from the exponential distribution, showed that method 1 is the outstanding method. It was followed by method 2. In choosing a suitable number of classes, Freedman and Diaconis's rule is the best.The Sturges rule recorded the least score of RMSE only for normal and uniform samples of sizes 50 and 100. It coincides with a comment made by Hyndman in 1995 that Sturges[12] rule gives similar results as Scott[7] and Freedman and Diaconis[5] rules when the sample size, $n < 200$, but it does not work for sample size larger than 200 [24]. Moreover, the findings using the Glasgow weather data buttressed the results of the simulation studies.

5.2 Conclusion

Based on the simulation and the real data, we found that method 1 outperformed the other methods. In choosing a suitable number of classes, the Freedman and Diaconis rule is the most suitable rule for handling normal and skewed data. Using uniformly distributed data, the Doane's rule performed well when $n \leq 500$, whereas the Cohran's rule, is the most suitable for $n \geq 1000$. However, the Terrel and Scott rule chose the suitable number of classes for the Glasgow weather data.

Acknowledgements

This research is partially funded by Universiti Putra Malaysia grant GP/2018/969400. The first author is supported by Tetfund, Federal government of Nigeria scholarship grant.

Competing interests

The authors declare that they have no competing interests.

REFERENCES

[1] A. F. George,. The Application of Sheppard's Correction for Grouping, Psychometrika, 6(1): 21-27, 1941.

[2] A. Hald,. On the History of the Correction for Grouping 1873 - 1922, Scandinavian Journal of Statistics. 28(3): 417 - 428, 2001.

[3] B. Lucien, and R. Yves,. How many Bins Should be Put in a Regular Histogram, ESAIM: Probability and Statistics, 10: 24 - 45, 2006.

[4] C. H. Brase and C. P. Brase, Understanding Basic Statistics, CA: Houghton Mifflin, Boston, 2001.

[5] D. Freedman and P. Diaconis, On the Histogram as a Density Estimator. Probability Theory and Related Fields 57(4): 453-476, 1981.

[6] D. P. Doane,. Aesthetic Frequency Classifications, The American Statistician, 30(4): 181-183, 1976.

[7] D. W. Scott, On Optimal and Data-Based Histograms, Biometrika, 66(3): 605–610, 1979.

[8] E. Di Nardo, A New Approach to Sheppard's Corrections, Mathematical Methods of Statistics, 19(2): 151-162, 2010.

[9] Frequency distribution. (n.d.), McGraw-Hill Dictionary of Scientific and Technical Terms, 6E, Retrieved May 10 2019 from https://encyclopedia2.thefreedictionary.com/frequency+distribution, 2003.

[10] G.R. Davies, The Analysis of Frequency Distributions, Journal of the American statistical association, 24(168): 349-366, 1929.

[11] G. R.Terrell, and D. W. Scott, Oversmoothed Nonparametric Density Estimates, Journal of the American Statistical Association, 80(389): 209-214, 1985.

[12] H. A. Sturges, The choice of a class interval, Journal of the American statistical association, 21 (153):65-66, 1926.

[13] H. L. Jones, The use of Grouped Measurements, Journal of the American Statistical Association, 36(216): 525-529, 1941.

[14] J. A. Pierce, Correction Formulas for Moments of a Grouped-Distribution of a Discrete Variates, Journal of the American Statistical Association, 38(221): 57-62, 1943.

[15] J. B. Canning, Formation of frequency Distributions, Journal of the American Statistical Association,21(154): 133-188, 1926.

[16] J. F. Kenney, Mathematics of Statistics, Technical composition Co, Boston, 1939.

[17] M. G. Kendall, The Conditions under which Sheppard's Corrections are Valid, Journal of Royal Statistical Society, 101(3): 592–605, 1938.

[18] M.P. Wand, Data-Based Choice of Histogram Bin Width, The American Statistician, 51 (1): 59-64, 1997.

[19] N. Dogan, and I. Dogan, Determination of the Number of Bins/Classes used in Histograms and Frequency Tables: A Short Bibliography, TurkStat, Journal of Statistical Research, 7(2): 77- 86, 2010.

[20] N. N. Cencov, Evaluation of Unknown Distribution Density from Observations, Soviet Mathematics, 3: 1559-1562, 1962.

[21] W.G. Cohran, Some Methods for Strengthening the Common Chi Square Test, Biometrics, 10 (4): 417-451, 1954.

[22] P.S. Dwyer, Grouping Methods, The Annals of Mathematical Statistics,13(2), 138-155, 1942.

[23] S. Manikandan, Frequency distribution, Journal of Pharmacology and Pharmacotherapeutics, 2(1): 54–56, 2011.

[24] R.J. Hyndman, The Problem With Sturges Rule for Constructing Histograms, Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary,19/09/2018, 1995.

[25] R. J. Hyndman, A. B. Koehler, Another Look at Measures of Forecast Accuracy, International Journal of Forecasting, 22 (4): 679–688, 2006.

[26] V. Gardiner, G. Gardiner, Analysis of Frequency Distributions, Concepts and Techniques in Modern Geography, No. 19 , Geo Books, Norwich, 1979.

[27] W. D. Baten, Correction for the moments of a frequency distribution in two variables, The Annals of mathematical statistics, 2: 309 - 319, 1931.

[28] W. F. Sheppard, On the Calculation of the Most Probable Values of Frequency-constants, for Data Arranged According to Equidistant Divisions of a Scale, Proceedings of the London Mathematical Society, 29(1): 353-380, 1898.

[29] W. F. Sheppard, The Calculation of Moments of a Frequency-Distribution, Biometrika, 5(4): 450-459, 1907.

[30] Wikipedia contributors, Radar chart , Wikipedia, The Free Encyclopedia, Retrieved, $https://en.wikipedia.org/w/index.php?title=Radar\_chart\&oldid=902246251$.(accessed June 22, 2019).