UNIVERSITI PUTRA MALAYSIA

MODIFIED FREQUENCY TABLES FOR VISUALISATION AND ANALYSIS OF UNIVARIATE AND BIVARIATE DATA

MOHAMMED MOHAMMED BAPPAH

FS 2021 41

MODIFIED FREQUENCY TABLES FOR VISUALISATION AND ANALYSIS OF UNIVARIATE AND BIVARIATE DATA

By

MOHAMMED MOHAMMED BAPPAH

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Doctor of Philosophy

February 2021

DEDICATIONS

To my parents: my mother, Hajiya Khadija Bappah, and my late father, Alhaji Bappah Mohammed Gokaru,
who installed in me the virtues of perseverance and commitment and relentlessly motivated me to be hardworking.
To my wife, Maryam Mohammed Gamawa, & my children Adnan, Khadija, and Abdulsalam,
whose affection, love, encouragement, prayers of day and night make me able to get such a success and honor.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philosophy

MODIFIED FREQUENCY TABLES FOR VISUALISATION AND ANALYSIS OF UNIVARIATE AND BIVARIATE DATA

By

MOHAMMED MOHAMMED BAPPAH

February 2021

Chairman: Mohd Bakri Adam, PhD
Faculty: Science

One way to make sense of data is to organize it into a more meaningful format called frequency table. The existing continuous univariate frequency table uses the midpoint to represent the magnitude of observations in each class, which results in an error called grouping error. The use of the midpoint is due to the assumption that each class's observations are uniformly distributed and concentrated around their midpoint, which is not always valid. The most significant parameter used when constructing the continuous frequency table is the number of classes or class width. Several rules for choosing the number of classes or class width have been reported in the literature; however, none has been proven to be better in all situations. The existing discrete frequency tables are simple to construct, easy to understand and interpret. However, when the number of elements in data is substantial, the table can be complicated. The existing non-parametric correlation measure, the Kendall correlation method, becomes laborious when the number of paired continuous observations is large enough.

In this research, to address the issue of grouping error, we proposed three statistics, median, midrange, and random selection to be used as the magnitude of observations in each class instead of the midpoint. In choosing the number of classes or class width, a new class width rule is proposed. We also proposed new discrete frequency tables that can be constructed by grouping the elements in data into classes. Using the bivariate continuous frequency table, a new correlation measure that is straightforward and free of normality assumption is developed. On addressing the issue of missing data in a univariate continuous frequency table, five different imputation methods are compared.

The four methods and the binning rules are simultaneously compared using root mean-squared-error (RMSE). Whereas the comparison using real data, the absolute error is used. The proposed discrete frequency tables are described using simulated and real data. While the new bivariate continuous table's correlation measure is illustrated using simulations and real data.

The comparison using the continuous frequency table's measure of location, mean, showed that the methods that used the median and midrange of observations in each class performed better relative to other methods. In choosing the number of classes, the proposed class width rule is the best for data simulated from the normal and exponential distributions. Meanwhile, for data simulated from the uniform distribution, the square root rule performed better than the other rules. The methods' evaluation using the frequency table's measures of skewness and kurtosis indicated that still, the methods that used the median and midrange to represent the magnitude of observations in each class were the best. The new discrete frequency tables can be a better choice, since, they can handle datasets with a substantial number of elements, and vividly reveals the significant features of datasets.

The results also showed that the new measure of correlation approximately equals to the Kendall correlation. Indeed, it can be used when the data is discrete, and the best alternative when the number of paired observations is large. In handling missing data, the simulation results showed that the mean imputation method is the best while the findings using real data indicated the mean imputation, *k* nearest neighbor imputation, and the multiple imputations by chained equations were the best methods. Also, the five imputation methods' performance is independent of the dataset and the percentage of missingness. And that the error increases as the percentage of missing observations increases.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

JADUAL KEKERAPAN YANG DIUBAHSUAI UNTUK VISUALISASI DAN ANALISIS DATA UNIVARIAT DAN BIVARIAT

Oleh

MOHAMMED MOHAMMED BAPPAH

Februari 2021

Pengerusi: Mohd Bakri Adam, PhD
Fakulti: Sains

Salah satu cara untuk membuat pengertian data adalah untuk menyusunnya ke dalam format yang lebih bermakna dipanggil Jadual kekerapan. Jadual kekerapan seragam selanjar sedia ada menggunakan titik tengah untuk mewakili magnitud pemerhatian dalam setiap kelas, yang mengakibatkan kesilapan yang dipanggil ralat terkumpul. Penggunaan titik tengah adalah disebabkan oleh andaian bahawa setiap pemerhatian kelas secara taburan seragam dan tertumpu di sekitar titik tengah mereka, yang tidak sah selalu. Parameter yang paling ketara yang digunakan apabila membina jadual kekerapan selanjar adalah bilangan kelas atau lebar kelas. Beberapa peraturan untuk memilih bilangan kelas atau lebar kelas telah dilaporkan dalam literatur walau bagaimanapun, tiada yang telah terbukti lebih baik dalam semua keadaan. Jadual kekerapan diskret yang sedia ada adalah mudah untuk dibina, mudah difahami dan ditafsir. Walau bagaimanapun, apabila bilangan elemen dalam data adalah besar, Jadual boleh menjadi rumit. Kaedah korelasi bukan berparameter yang sedia ada iaitu, kaedah korelasi Kendall, menjadi sukar apabila bilangan pemerhatian selanjar berpasangan adalah cukup besar.

Dalam kajian ini, untuk menangani isu ralat terkumpul, kami mencadangkan tiga statistik, median, julat midas, dan pemilihan secara rawak, untuk digunakan sebagai magnitud pemerhatian dalam setiap kelas dan bukannya titik tengah. Dalam memilih bilangan kelas atau lebar kelas, peraturan lebar kelas baru dicadangkan. Kami juga mencadangkan jadual kekerapan diskret baru yang boleh dibina dengan mengumpulkan elemen dalam data ke dalam kelas. Penggunaan jadual frekuensi selanjar, satu langkah korelasi baru yang terus-terang dan bebas daripada andaian ini dibangunkan. Dalam menangani isu data hilang dalam jadual kekerapan yang selanjar, lima kaedah imputasi yang berbeza dibandingkan.

Empat kaedah dan peraturan binning pada masa yang sama dibandingkan dengan menggunakan ralat punca-min-kuasa dua (RMSE). Manakala perbandingan menggunakan data sebenar, ralat mutlak digunakan. Jadual kekerapan diskret yang dicadangkan, menggunakan data simulasi dan data sebenar. Manakala (yang baru) korelasi jadual bivariat selanjar digambarkan menggunakan simulasi dan data sebenar.

Perbandingan menggunakan taburan kekerapan selanjar mengukur lokasi, min menunjukkan bahawa kaedah yang menggunakan median dan julat midas dalam setiap kelas menunjukkan prestasi yang lebih baik berbanding dengan kaedah-kaedah lain. Dalam memilih peraturan binning, peraturan yang dicadangkan adalah yang terbaik untuk simulasi data dari pengagihan biasa dan eksponen. Sementara itu, bagi simulasi data daripada taburan seragam, peraturan punca kuasa adalah lebih baik daripada peraturan yang lain. Sementara itu, kaedah penilaian yang menggunakan taburan kekerapan untuk mengukur kepencongan dan kurtosis menunjukkan bahawa kaedah yang menggunakan median dan midrange untuk mewakili magnitud pemerhatian dalam setiap kelas adalah yang terbaik. Jadual kekerapan diskret yang baru boleh menjadi pilihan yang lebih baik, kerana, mereka boleh mengendalikan set data besar, mendedahkan ciri penting set data secara jelas.

Keputusan juga menunjukkan bahawa ukuran baru korelasi hampir sama dengan korelasi Kendall. Semangnya, ia boleh digunakan apabila data adalah diskret, dan alternatif yang terbaik apabila bilangan pemerhatian berpasangan adalah besar. Dalam mengendalikan data yang hilang, keputusan simulasi menunjukkan bahawa kaedah imputasi min adalah yang terbaik manakala penemuan menggunakan data sebenar menunjukkan imputasi min, *k* imputasi jiran terdekat, dan pelbagai imputasi oleh persamaan yang dirantai adalah kaedah terbaik. Juga, prestasi lima kaedah imputasi adalah bebas daripada dataset dan peratusan hilang. Dan ralat meningkat apabila peratusan pemerhatian yang hilang meningkat.

ACKNOWLEDGEMENTS

My profound gratitude goes to Almighty Allah for life, guidance, and the successful completion of this program. I am eternally grateful to my supervisory, Associate Professor Dr. Mohd Bakri Adam, for his great support, encouragement, and guide to see that this research work has maintained the required standard. Your vast experiences in research have made me better equipped in research than ever. I am so humbled by your constructive criticism and willingness to discharge your duty, may Allah reward you abundantly. My special appreciation goes to my co-supervisors, Dr. Hani Syahida Zulkafli and Dr. Norhaslinda Ali, for their guidance and insightful comments that added quality to this research.

I am profoundly grateful to my institution, the Federal University of Kashere, and the federal government of Nigeria, for the TETFUND scholarship, which enabled me to pursue my P.hD. I also appreciate the efforts of friends and other good wishers. Thank you all, may Allah bless you.

Mohammed Bappah Mohammed

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Mohd Bakri Adam, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairperson)

Hani Syahida Binti Zulkafli, PhD
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Member)

Norhaslinda Binti Ali, PhD
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Member)

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

AIC	Access to Information and Communication
CB	Class Boundary
Ce	Cencov’s Rule
cf	Cummulative Frequency
CI	Class Interval
Co	Cocheran’s Rule
Cum	Cummulative
Do	Doane’s Rule
Ex	Existing Method
FD	Freedman and Diaconis Rule
Freq	Frequency
HLE	Health Life Expectancy
HS	Happiness Score
HW	Health and Wellness
IQR	Interquartile Range
k-NN Imp	*k* Nearest Neighbor Imputation
MAD	Median Absolute Deviation
Mean Imp	Mean Imputation
Median Imp	Median Imputation
MIC	Maximal Information Coefficient
MICE Imp	Multiple Imputations by Chained Equations
MISE	Mean Integrated Square Error
MVUE	Minimum Variance Unbiased Estimator
NBMC	Nutrition and Basic Medical Care
pdf	Probability Density Function
PFC	Personal Freedom and Choice
PR	Personal Rights
R	Range
Ri	Rice’s Rule
RMSE	Root Mean-Squared Error
Sample Imp	Sample Imputation
SD	Standard Deviation
SDSN	Sustainable Development Solutions Network
Sc	Scott’s Rule
SPI	Social Progress Index
SQ	Square Root Rule
St	Sturges’ Rule
TS	Terrell and Scott Rule

LIST OF NOTATIONS

k	Number of Classes or Number of Bins Number
k_1	of Classes of a Continuous Variable X Number
k_2	of Classes of a Continuous Variable Y Lower
l_i	Class Limit of Class i
u_i	Upper Class Limit of Cass i
l_{x_j}	Lower Class Limit of Class j of Variable X
u_{x_j}	Upper Class Limit of Cass j of Variable X
l_{y_i}	Lower Class Limit of Class i of Variable Y
u_{y_i}	Upper Class Limit of Cass i of Variable Y
M1, Method 1	Using Arithmetic Mean Inplace of the Midpoint
M2, Method 2	Using Median Inplace of the Midpoint
M3, Method 3	Using Midrange Inplace of the Midpoint
M4, Method 4	Using Random Selection
Mo_i	Mode of Class i
m	Number of Elements in a Discrete Data
m_1	Number of Elements in a Discrete Variable X
m_2	Number of Elements in a Discrete Variable Y
n	Sample Size
n_E	Number of Empty Cells
n_c	Number of Concordant pairs
n_d	Number of Discordant pairs
n_s	Number of Samples
R_{m_i}	Mode Rank of Class i
e_i	Element in Class i
${}_xe_i$	Element of Variable X in Row i
${}_ye_i$	Element of Variable Y in Column j
f_i	Frequency of Class i
g	Number of Groups
f_{ij}	Joint Frequency of Variables X and Y in Cell C_{ij}
s	Sample Standard Deviation
w	Class Width or Bin Width
x_i^*	Midpoint of Cass i
x_{me_i}	Arithmetic Mean of Class i
x_{md}	Median
x_{md_i}	Median of Class i
x_{mr_i}	Midrange of Class i
x_{mo}	Class Mode
x_{mo_i}	Mode of Elements in Class i of Discrete Variable X
y_{mo_i}	Mode of Elements in Class j of Discrete Variable Y
δ	Smallest measurement unit of a continuous dataset
ξ	Skewness Coefficient
ξ_{x^*}	Skewness Coefficient Using the Midpoint
ξ_{me}	Skewness Coefficient Using the Arithmetic Mean

ξ_{md}	Skewness Coefficient Using the Median
ξ_{mr}	Skewness Coefficient Using the Midrange
κ	Kurtosis Coefficient
κ_{x^*}	Kurtosis Coefficient Using the Midpoint
κ_{me}	Kurtosis Coefficient Using the Arithmetic Mean
κ_{md}	Kurtosis Coefficient Using the Median
κ_{mr}	Kurtosis Coefficient Using the Midrange
τ	Kendall Correlation Coefficient
$\sum_{i}^{m_2} f_{i\,m_1}$	Sum of the frequencies Accross the rows
$\sum_{j}^{m_1} f_{m_2 j}$	Sum of the frequencies Accross the columns
$\sum_i \sum_j f_{ij}$	Sum of the frequencies of all the cells

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Exploratory data analysis (EDA) plays a significant role in Statistics. The EDA refers to the set of statistical tools originally devised by Tukey (1977) displaying data so that its essential characteristics can be easily seen (Behrens, 1997; Hoaglin, 2003). The EDA was described as detective work, numerical detective work or counting detective work or graphical detective work (Tukey, 1977). Exploratory data analysis is a detective in nature, statistical detective, tools are applied to come up with new knowledge, and in this respect, outliers play a vital role (Mahendran and Turaj, 2011). Exploratory data analysis tools have included a new dimension in statistics to the way people deal with data (Hoaglin, 2003; Velleman and Hoaglin, 2004). A raw dataset is more attractive and captures people's minds if it can be depicted in either tabular or graphical form. The tabular representations are precise and provide the reader with apparent features of the data; however, the graphical presentations have more visual significance and useful in detecting patterns in a dataset (Davies, 1929; Beniger, 1978; Gelman et al., 2002; Kastellec and Leoni , 2007; Gelman, 2011; Xu and Wang, 2020). A data point can only be significant if considered along with other observations in a frequency table (Gardiner and Gardiner, 1979). Also, raw data do not display any meaningful representation unless being organized in a systematic form (Myatt and John, 2014). A raw data can be partitioned into classes of suitable sizes, showing observations with the corresponding frequencies. When a dataset is systematically organized in this form is called a frequency table (Kenney, 1939). The classes are to be constructed such that each data point falls into only one class. A univariate continuous frequency table displays data along with the midpoint, cumulative frequency, and the class boundary (Gravetter and Wallnau, 2000; Brase and Brase, 2001).

Brase and Brase (2001) emphasized that irrespective of the type of data, sample or population is available, the data are organized and communicated to other people, that is why tables and graphs are unavoidable. Organizing the raw data into a structured format like a frequency table makes it easier for a big audience to grasp and interpret the data within a short period (Lohaka, 2007).

An EDA tool, frequency table, is very significant in statistics. The frequency table transforms raw data from meaningless details into a more easily presentable or interpretable, easy-to-comprehend organized format (Levin and Fox, 2004). A well-organized frequency table makes a possible detailed analysis of the population's structure concerning a given feature. 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 univariate continuous frequency table is, it serves as a bridge between raw data and a histogram (Freedman et al., 1998; Frequency Distribution, 2003; Fisher and Marshall, 2009). The frequency table also facilitates the construction of a cumulative frequency curve, ogive, and frequency polygon. An important to mention is the table aids careful comparison of datasets (Lohaka, 2007).

The existing frequency tables can be classified based on the data types as well as the number of variables in the data into univariate and bivariate discrete frequency tables and univariate and bivariate continuous frequency tables. The univariate frequency table summerizes raw data of a single variable in an organised form. On the other hand, the bivariate frequency table is a table that organises raw paired observations into a meaning format. The construction of the discrete frequency tables is straightforward, the elements in discrete data are the natural classes (Kenney, 1939). In the same vein, the elements in paired discrete data determine the number of classes of the variables for the existing bivariate discrete frequency table. The first step in constructing the continuous frequency table is determining the number of classes or the class width. The important points to note when dividing continuous data into classes are the classes should be big enough, mutually exclusive, and exhaustive, and preferably the classes should be of equal width, though sometimes unequal width must be used (Dogan and Dogan, 2010).

The univariate discrete frequency table mostly contains only two columns. The first column displays the elements and the second column presents the number of occurrences of each element. On the other hand, the components of the univariate continuous frequency table are the class limits, class boundaries, the midpoint, frequency, and 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. The midpoint is the average of the upper and lower class limits. The midpoint is also the center of bars on a histogram. Another component representing the number of observations in each of the classes is the frequency (f). In a frequency table, the frequencies usually written as $f_1, f_2, \ldots, f_k$ 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 less or greater observations than a component, cumulative frequency (cf) is included. At a particular class, the cumulative frequency is the total frequency up to the upper-class boundary of that class. The cumulative frequencies of the classes are $f_1, f_1 + f_2, \ldots, f_1 + f_2 + \ldots + f_k$ (Kenney, 1939).

Moreover, to construct the frequency table, the initial step is to determine the range of the data, then choosing a suitable number of classes, calculating the class width, obtaining the lower limit of the first class, and lastly, determining the class intervals (Kenney, 1939; Manikandan, 2011). 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. The knowledge of either of the two suffice.

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 dataset.

A data organised in a frequency table is skewed if the mean and median from the table are not equal, or more general if the data is not symmetric. The kurtosis of a frequency table measures the rare extreme values which appear as outliers in a histogram. A leptokurtic distribution is a distribution that is more outlier-prone than the normal distribution. Meanwhile, a distribution that is less prone to outlier is said to be platykurtic. Tables 1.1, 1.2, 1.3, and 1.4 respectively present the existing univariate discrete frequency table, bivariate discrete frequency table, univariate continuous frequency table, and the existing bivariate continuous frequency table.

Table 1.1: Existing Univariate Discrete Frequency Table

Class	Element (e)	Frequency (f)
1	e_1	f_1
2	e_2	f_2
$\vdots$	$\vdots$	$\vdots$
m	e_m	f_m

where m is the number of elements in the discrete dataset, $e_1, e_2, \cdots, e_m$ and $f_1, f_2, \cdots, f_m$ are respectively, the elements and the frequencies of the classes, $m, e_1, e_2, \cdots, e_m \in \mathbb{Z}$. The frequencies are the number of occurrences of the elements. A large number of elements in the discrete data leads to a very long table. Here, the suitable measures of location and scale are mode and range. The mode is the element that appeared the most, while the range is the difference between the smallest and largest elements.

Table 1.2: Existing Bivariate Discrete Frequency Table

X Y	xe_1	xe_2	$\cdots$	xe_{m_1}
ye_1	f_{11}	f_{12}	$\cdots$	f_{1m_1}
ye_2	f_{21}	f_{22}	$\cdots$	f_{2m_1}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
ye_{m_2}	$f_{m_2 1}$	$f_{m_2 2}$	$\cdots$	$f_{m_2 m_1}$

where $xe_i, i = 1, 2, \cdots, m_1$ denote the elements of variable X displayed in the columns and ye_j, $j = 1, 2, \cdots, m_2$ are the elements of the second variable Y presented in the rows, f_{ij} is the joint frequency of variables X and Y in cell ij. A cell is usually blanked if no entry for the cell ij, it means the two variables have no joint frequency in cell ij. The total frequency n can be obtained either by adding the frequencies accross the rows, $\sum_j f_{ij}$, and then totaling the marginal sums in column or by adding the frequencies accross the columns, $\sum_i f_{ij}$, and then totaling the marginal sums in row or by summing the frquencies in the cells in any order, $\sum_i \sum_j f_{ij}, \;\; m_1, m_2 \in \mathbb{Z}$. When the numbers of elements in the two variables, m_1 and m_2 are large, the table can be very long and big.

Table 1.3: Existing Univariate Continuous Frequency Table

Class	CI		CB		Freq	Cum Freq	midpoint
	l_c	u_c	l_b	u_b	(f)	(f_c)	(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	$f_1 + f_2$	$\frac{l_2+u_2}{2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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 smallest measurement unit of the dataset, 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. $i = 1, 2, \cdots, k, k \in \mathbb{Z}, l_c, u_c, l_b, u_b, f_i, \delta \in \mathbb{R}$. The midpoint of class i is equal to $\frac{l_i+u_i}{2}$, l_i and u_i are respectively the lower and upper limits of class i, $i = 1, 2, \cdots, k$. The midpoint x^* is used to represent the magnitude of observations in each class when obtaining statistical measures from the univariate continuous frequency table. Either of the two parameters, the number of classes k, or the class width $w = u_c - l_c$, must be determined first before constructing the table. The scientific classification rules can be used to obtain a suitable number of classes or class width.

Table 1.4: Existing Bivariate Continuous Frequency Table

X Y	Class	$[l_{x_1}, u_{x_1})$	$[l_{x_2}, u_{x_2})$	$\cdots$	$[l_{x_{k_1}}, u_{x_{k_1}})$
Class	x^* y^*	x_1^*	x_2^*	$\cdots$	$x_{k_1}^*$
$[l_{y_1}, u_{y_1})$	y_1^*	f_{11}	f_{12}	$\cdots$	f_{1k_1}
$[l_{y_2}, u_{y_2})$	y_2^*	f_{21}	f_{22}	$\cdots$	f_{2k_1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$[l_{y_{k_2}}, u_{y_{k_2}})$	$y_{k_2}^*$	$f_{k_2 1}$	$f_{k_2 2}$	$\cdots$	$f_{k_2 k_1}$

where x^* denotes the midpoints of variable X classes displayed in the columns and y^* are the midpoints of class intervals of the second variable Y presented in the rows, f_{ij} is the joint frequency of variables X and Y in cell ij. A cell is usually blank if no entry for the cell ij, it means the two variables have no joint frequency in cell ij. The l_{x_j} and u_{x_j} are lower and upper class limits of class j of variable X. Whereas l_{y_i} and u_{y_i} are lower and upper class limits of class i of variable Y, , $k_1, k_2, f_{ij} \in \mathbb{Z}$, and $l_{x_i}, u_{x_i}, l_{y_j}, u_{y_j}, \in \mathbb{R}$, $j = 1, 2, \ldots, k_1$, and $i = 1, 2, \ldots, k_2$. The number of classes of the two variables k_1 and k_2 must be determined first before the table is constructed. The scientific number of classes and class width rules can be used to obtain the appropriate number of classes. When the rules that are based on only the sample size are used to determine k_1 and k_2 we have a square table $k_1 = k_2$. Meanwhile, the table can be rectangular, $k_1 \neq k_2$, when rules that also incorporate the deviance concept apart from the sample size are used.

1.2 Problem Statement

The graphical and tabular representation of data provides the simplest and most effective means of understanding and interpreting data. The frequency classifications are unbendable, both as a means of condensing compactly large data size and a form of generalization. Therefore, frequency distribution analysis is likely to remain one of the essential statistical tools (Davies, 1929).

Though grouping is unavoidable, especially when the dataset is large, the process can lead to a considerable error compared to the original data. When computing the statistical measures from the existing continuous frequency table, the magnitude of observations in each class is represented by the midpoint, which results in error known as grouping error (Davies, 1929; Kenney, 1939). Different researchers in the literature suggested various approaches to minimizing this error. One of the approaches is the use of a correction formula to minimize the error. The most use correction formula is Sheppard's correction, which was due to Sheppard in 1898 (Sheppard, 1898, 1907; Kendall, 1938; George, 1941; Hald, 2001). 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. The odd powers are assumed not affected by the grouping error. Following this several researches on correction for grouping error have emerged; such as, the Canning (1926), Davies (1929), Baten (1931), Jones (1941), Dwyer (1942), Pierce (1943), Hald (2001) and the most recent work by Di Nardo (2010).

Furthermore, one of the two significant parameters, the number of classes and the class width, must be determined before constructing the continuous frequency table. While the former describes the number of partitions of the dataset, the later is the distance between lower and upper-class limits (Wand, 1997). These two parameters are dependent on one another; if one is known, the other can be obtained. Determining the appropriate number of classes to be used in constructing a frequency table remains a long-existing problem in statistics. Different rules for choosing the number of classes and class width were reported in the literature; however, none of the rules has been proven to be better in all situations (Birge and Rozenholc, 2006).

So also, the existing non-parametric correlation measures tend to be laborious when the number of pairs of the bivariate data is substantial. Moreover, in the case of the existing discrete frequency tables, when the number of elements in the data is large enough, the process can result in a very long frequency table that cannot be easily handled.

Missing data are sometimes inevitable and can affect the conclusions that can be drawn from data. In classes of frequency tables, missing observations do occur, and it is necessary to estimate them to arrive at valid conclusions. Various techniques of handling missing data have been reported in the literature, but none has been examined to estimating missing observations in a frequency table.

In this research, the problems are addressed by developing new ways of constructing both the continuous and discrete frequency tables. In choosing a suitable number of classes or class width, a new rule is proposed. Moreover, a new, bivariate continuous frequency table's correlation measure is developed. A suitable method is recommended in addressing the issue of missing observations in a univariate continuous frequency table.

1.3 Research Aim and Objectives

This research aims to develop new methods of constructing both the continuous and discrete frequency tables. The objectives are;

1. To propose three statistics, random selection, median, and midrange, to represent the mid-value of observations in each class of the continuous frequency table.
2. To develop a new rule for choosing the class width based on the median absolute deviation for the continuous frequency tables.
3. To propose new univariate and bivariate discrete frequency tables by partitioning the elements in the discrete data into classes and using the mode as the magnitude of elements in each class.
4. To develop a new correlation measure, which is based on Kendall's concordance and discordance approach for the empty cells in a bivariate continuous frequency table.
5. To conduct simulation studies and identify the best method of handling missing data in univariate continuous frequency tables.

1.4 Limitation of the Study

In this research the issue of outliers is not considered when constructing the continuous frequency tables. Also, the continuous frequency tables are limited to only equal width classes.

1.5 Structure of the Thesis

This thesis's overall structure takes the form of eight chapters, including this introductory chapter, Chapter 1. Chapter 2 presents a general literature review on the frequency table. This research's main findings are presented in Chapter 3, Chapter 4, Chapter 5, Chapter 6, and Chapter 7. Chapter 3 focused on the proposed univariate continuous frequency tables. Meanwhile, the new univariate and bivariate discrete frequency tables are respectively given in Chapters 4 and 6. In Chapters 5 and 7, the new bivariate continuous frequency table's correlation measure and methods of handling missing data in a univariate continuous frequency are discussed. The last chapter, Chapter 8, gives the general summary, conclusion, and the identified areas for future work.

Chapter 1 presents a general background on the exploratory data analysis (EDA) tool, frequency table, the types of frequency table based on the data type, continuous and discrete frequency table, and the types regarding the number of variables, univariate and bivariate frequency tables. The chapter also highlights the statement of the problem, research aim and objectives, the limitations of the research, and the thesis structure. Chapter 2 covers the general literature review on the frequency table. A review on real numbers, types of variable, data types, continuous frequency tables, discrete frequency tables, missing data in a univariate continuous frequency table, graphs of the univariate frequency tables, and error incurred as a result of grouping raw data in a continuous frequency table.

The new univariate continuous frequency table, using the four proposed statistics, arithmetic mean, median, midrange, and random selection, are presented in Chapter 3. Assessment of the proposed statistics used to represent the magnitude of observations in each class and the rules used in choosing the number of classes are given in this chapter.

Chapter 4 presents the proposed univariate discrete frequency table, the table's description using simulation studies from five discrete distributions, and real data. Chapter 5 illustrates the new bivariate continuous frequency table's correlation measure, empty cell correlation. In Chapter 6, we describe the new bivariate discrete frequency table. Meanwhile, Chapter 7 presents the results of comparing five imputation methods used in handling missing observations in a univariate continuous frequency table. A general summary of the whole thesis, conclusion, and recommendation for future work are given in Chapter 8.

BIBLIOGRAPHY

Abdenour, H. G. and Philippe, R. (2018). A Binning Formula of Bi-histogram for Joint Entropy Estimation Using Mean Square Error Minimization. *Pattern Recognition Letters*, 101(2018): 21-28.

Andrews, D. F., et al. (1972). Robust Estimates of Locution: Survey and Advances. Princeton, N.J. : Princeton University Press.

Adway, S. W. (2017). Social Progress and Happiness. Retrieved May 28, 2019 from `https://www.kaggle.com/adwaywadekar/social-progress-and-happiness`.

Afshin, G. (2017). What is the Optimal Bin Size of a Histogram: An Informal Description. *International Mathematical Forum*, 12(5): 731-736.

Avtar, S. S., Khuneswari, G. P., Abdullah, A. A., McColl, J. H., Wright, C., and Team, G. M. S. (2019). Comparison Between EM Algorithm and Multiple Imputation on Predicting Children's Weight at School Entry. *Journal of Physics: Conference Series*, 1366(1): 012124.

Baten, W. D. (1931). Correction for the Moments of a Frequency Distribution in Two Variables. *The Annals of Mathematical Statistics*. 2(1931): 309 - 319.

Batista, G. and Monard, M. (2002). A Study of K-Nearest Neighbour as an Imputation Method. *HIS*, 87(2002): 251-260.

Behrens, T. J. (1997). Principles and Procedures of Exploratory Data Analysis. *Psychological Methods*, 2(2): 131-160.

Beniger, J. R. and Robyn, D. L. (1978). Quantitative Graphics in Statistics: A Brief History. *The American Statistician*, 32(1): 1-11.

Beretta, L., and Santaniello, A. (2016). Nearest Neighbor Imputation Algorithms: A Critical Evaluation. *BMC Medical Informatics and Decision Making*, 16: 74.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., and Bendayan, R. (2013). Skewness and Kurtosis in Real Data Samples. *Methodology*, 9(2): 78–84.

Blest, D. C. (2003). A New Measure of Kurtosis Adjusted for Skewness. *Australian & New Zealand Journal of Statistics*, 45(2): 175-179.

Bonato, M. (2011). Robust Estimation of Skewness and Kurtosis in Distributions with Infinite Higher Moments. *Finance Research Letters*, 8(2011): 77-87.

Bono, R., Arnau, J., Alarcón, R., and Blanca, M. J. (2019). Bias, Precision, and Accuracy of Skewness and Kurtosis Estimators for Frequently Used Continuous Distributions. *Symmetry*, 12(1): 19.

Bowley, A. L. (1901). *Elements of Statistics*. London, P.S. King & Son.

Brase, C. H., and Brase, C. P. (2001). *Understanding Basic Statistics*. Boston, CA: Houghton Mifflin.

Brys, G., Hubert, M. and Struyf, A. (2004). A Robust Measure of Skewness. *Journal of Computational and Graphical Statistics*, 13(4): 996–1017.

Campbell, D. T. and Stanley, J. T. (1966). Experimental and Quasi-experimental Designs for Research. Chicago, Rand McNally.

Cameron, A. C. (2009). Excel 2007: Histogram Retrieved from `http://cameron.econ.ucdavis.edu/excel/ex11histogram.html`, 15/02/2020.

Canning, J. B. (1926). Formation of Frequency Distributions. *Journal of the American Statistical Association*, 21(154): 133-188.

Cencov, N. N. (1962). Evaluation of Unknown Distribution Density from Observations. *Soviet Mathematics*, 3: 1559-1562.

Chai, T. and Draxler, R. R. (2014). Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments Against Avoiding RMSE in the Literature. *Geoscience Model Development*, 7(3): 1247–1250.

Chauvet, G., Deville, J. C., and Haziza, D. (2011). On Balanced Random Imputation in Surveys. *Biometrika*, 98(2): 459–471.

Cheema, J. (2014). Some General Guidelines for Choosing Missing Data Handling Methods in Educational Research. *Journal of Modern Applied Statistical Methods*, 13(2): 53 -75,

Chissom, B. S. (1970). Interpretation of the Kurtosis Statistics. *The American Statistician*, 24(4): 19-23.

Chu, K., Dean, S. and Illowsky, B. (2013). Elementary Statistics. Texas, Connexions Rice University, Houston.

Cochran, W.G. (1954). Some Methods for Strengthening the Common Chi Square Test. *Biometrics*, 10(4): 417-451.

Daly, J. E. (1988). The Construction of Optimal Histogram. *Communications in Statistics -Theory and methods*, 17(9): 2921–2931.

Daniel, M. (2017). Missing Data Methods for Arbitrary Missingness with Small Samples. *Journal of Applied Statistics*, 44(1): 24-39.

Darlington, R. B. (1970). Is Kurtosis Really Peakedness?. *The American Statistician*, 24(2): 19-22.

Davies, G.R. (1929). The Analysis of Frequency Distributions. *Journal of the American Statistical Association*, 24(168): 349-366.

David, M. L. (2003). Online Statistics Education: A Multimedia Course of Study, Rice University. Retrieved June 19 2020 from `http://onlinestatbook.com/`.

De Beer, C. F. and Swanepoel, J. W. H. (1997). Simple and Effective Number-of-bins Circumference Selectors for a Histogram. *Statistics and Computing*, 9(1): 27-35.

Delucchi, K. L., and Bostrom, A. (2004). Methods for Analysis of Skewed Data Distributions in Psychiatric Clinical Studies: Working With Many Zero Values. *American Journal of Psychiatry*, 161(7): 1159–1168.

Denby, L. and Mallow, C. (2009). Variation on the Histogram. *Journal of Computational and Graphical Statistics.*, 18(1): 21-31.

DiCiccio, C. J. and Romio, J. P., (2017). Robust Permutation Tests for Correlation and Regression Coefficients, *Journal of the American Statistical Association*, 112: 1211-1220.

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

Dhar, S. S. and Chaudhuri, P. (2012). On the Derivatives of the Trimmed Mean. *Statistica Sinica*, 22(2012): 655–679.

Dhar, S. S. and Das, U. (2020). On Distance Based Goodness of Fit Tests for Missing Data when Missing Occurs at Random. *Australian & New Zealand Journal of Statistics*, doi:10.1111/j.1467-842X.XXX.

Dhar, S. S., Chakraborty, B., and Chaudhuri, P. (2014). Comparison of Multivariate Distributions Using Quantile–Quantile Plots and Related Tests. *Bernoulli*, 20(3):1484–1506.

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

Doane, D. P. and Seward, L. E. (2011). Measuring Skewness: A Forgotten Statistics?. *Journal of Statistics Education*, 19(2): 1-18.

Dogan, N. and Dogan, I.(2010). 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.

Dodge, Y. (2003). *The Oxford dictionary of Statistical Terms.* Oxford, Oxford University Press.

Dodge, Y. (2008). *The Concise Encyclopedia of Statistics*. New York, Springer-Verlag New York.

Douglas, W. and Tim, J. (2017). Students' Understanding of Bar Graphs and Histograms: Results from the Locus Assessments. *Journal of Statistics Education*, 25(2): 90-102.

Dragos, F. and Livia, D. (2012). New Correlation Coefficient for Data Analysis. *Scientific Papers Series "Management, Economic Engineering in Agriculture and Rural Development"*, 12(4): 159-162.

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

Edward, H. L. (2004). The Mean and Standard Deviation: What Does it All Mean?. *Journal of Surgical Research*, 119(2): 117–123.

Eric, B. N., and Clayton, V. D. (2012). Calculating a Robust Correlation Coefficient and Quantifying its Uncertainty. *Computers and Geosciences*, 40(2012): 1–9.

Evans, I. S. (1977). The Selection of Class Intervals. *The Royal Geographical Society (with the Institute of British Geographers)*, 2(1): 98-124.

Finney, D. J. (1948). The Application of Statistical Methods to Food Problems; the Inevitability of Statistics. *Analyst*, 73(862): 2-30.

Fisher, M. J. and A. P. Marshall (2009). Understanding Descriptive Statistics. *Australian Critical Care*, 22(2): 93-97.

Frederico, Z. P., Julio M. S., and Carlos, D. P. (2011). Comparing Diagnostic Tests with Missing Data. *Journal of Applied Statistics*, 38(6): 1207-1222.

Friendly, M., and Denis, D. J. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. Accessed on 27 December 2020 at `http://www.datavis.ca/milestones/`.

Frequency Distribution. (2019). McGraw-Hill Dictionary of Scientific and Technical Terms, 6E. Retrieved May 10 2019 from `https://encyclopedia2.thefreedictionary.com/frequency+distribution`.

Funkhouser, H. G. (1937). Historical Development of the Graphical Representation of Statistical Data. *Osiris*, 3(1937): 269 - 404.

Freedman, D., Pisani, R., and Purves, R. (1998). *Statistics*. New York, W. W. Norton.

Freedman, D. and Diaconis, P. (1981). On the Histogram as a Density Estimator: L2 Theory. *Probability Theory and Related Fields* , 57(4): 453-476.

Garrett, F. (2008). Missing data: Implications for Analysis. *Nutrition*, 24(2008): 200-202.

Gardiner, V. and Gardiner, G. (1979). *Analysis of Frequency Distributions:Concepts and Techniques in Modern Geography; no. 19.* Geo Abstracts, University of East Anglia, Norwich.

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

Gelman, A. (2011). Why Tables are Really Much Better than Graphs. *Journal of Computational and Graphical Statistics*, 20(1): 3-7.

Gelman, A., C. Pasarica, and R. Dodhia (2002). Let's Practice What We Preach: Using Graphs Instead of Tables. *American Statistician*, 56(2002): 121-130.

Gleb, B., Humberto, B. and Javier, F. (2011). The Median and its Extensions. *Fuzzy Sets and Systems*, 175(2011): 36-47.

Groeneveld, R. A. and Meeden, G. (1984). Measuring Skewness and Kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 33(1): 391-399.

Gravetter, F. J. and Wallnau, L. B. (2000). *Statistics for the Behavioral Sciences*, Belmont, Wadsworth – Thomson Learning.

Guven, M. G., Mustafa, S. S. and Eray, Y. (2017). Percentile Based Histogram Bin Width. *International Journal of Sciences and Research*, 73(3): 93-97.

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

Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346): 383-393.

He, K. and Meeden, G. (1997). Selecting the Number of Bins in a Histogram: A Decision Theoretic Approach. *Journal of Sstatistical Planning and Inference*, 61(1997): 59-69.

Heitjan, D. F. (1989). Inference from Grouped Continuous Data: A Review. *Statistical Science*, 4(2): 164–179.

Hoaglin, D. (2003). John W. Tukey and Data Analysis. *Statistical Science*, 18(3): 311-318.

Hogg, R. V. (1974). Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory. *Journal of the American Statistical Association*, 69(348): 909-923.

Horswell, R., and Looney, S. (1993). Diagnostic Limitations of Skewness Coefficients in Assessing Departures from Univariate and Multivariate Normality. *Communications in Statistics - Simulation and Computation*, 22(2): 437–459.

Hu, L.Y., Huang, M. W., Ke, S.W., and Tsai, C. F. (2016). The Distance Function Effect on k-Nearest Neighbor Classification for Medical Datasets. *SpringerPlus*, 5(1): 1-9.

Huber, P. J. (1981). *Robust Statistics*. New York, John Wiley.

Hutson, A. D., (2019). A Robust Pearson Correlation Test for a General Point Null Using a Surrogate Bootstrap Distribution. *PLoS One*, 14(5): e0216287.

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

Islam, T. U. (2019). Ranking of Normality Tests: An Appraisal through Skewed Alternative Space. *Symmetry*, 11(7): 872.

Jadhav, A., Pramod, D., and Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33(10): 913-933.

Jakobsen, J.C., Gluud, C., Wetterslev, J. (2017). When and How Should Multiple Imputation be Used for Handling Missing Data in Randomised Clinical Trials – A Practical Guide with Flowcharts. *BMC Medical Research Methodology*, 17(1): 162.

Joanes, D. N. and Gill, C. A. (1998). Comparing Measures of Sample Skewness and Kurtosis. *Journal of the Royal Statistical Society Series D: The Statistician*, 47(1): 183-189.

Johar, M. A. (2019). Data on Depression. Retrieved May 22, 2020 from `https://www.kaggle.com/ukveteran/data-on-depression`.

Jones, H. L. (1941). The Use of Grouped Measurements. *Journal of the American Statistical Association*, 36(216): 525-529.

Kang, H. (2013). The Prevention and Handling of the Missing Data. *Korean Journal of Anesthesiology*, 64(5): 402.

Karahalios, A., Baglietto, L., Carlin, J. B., English, D. R., and Simpson, J. A. (2012). A Review of the Reporting and Handling of Missing Data in Cohort Studies with Repeated Assessment of Exposure Measures. *BMC Medical Research Methodology*, 12(1): 96.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *The Statistician*, 52(3): 381-393.

Kastellec, J. P. and Leoni, E. L. (2007). Using Graphs Instead of Tables in Political Science. *Perspectives on Politics*, 755-771.

Kendall, M. G. (1938a). A New Measure of Rank Correlation. *Biometrika*, 30(1938): 81 – 93.

Kendall, M. G. (1938b). The Conditions Under Which Sheppard's Corrections are Valid. *Journal of Royal Statistical Society*, 101(3): 592–605.

Kenney, J. F. (1939). *Mathematics of Statistics*. Boston, Technical Composition Co.

Kim, T.H. and White, H. (2004). On More Robust Estimation of Skewness and Kurtosis. *Finance Research Letters*, 1(2004): 56–73.

Köroğlu, Ö. (2019). Weather Istanbul Data 2009-2019. Retrieved November 15, 2019 from `https://www.kaggle.com/vonline9/weather-istanbul-data-20092019`.

Larry, J. S. (1998). *Schaum's Outline, Theory and Problems of Beginning Statistics*, New York, McGraw-Hill.

Levin, J., and Fox, J. A. (2004). *The essentials: Elementary Statistics in Social Research*. New York, Pearson Education.

Liao, S.G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. C., and Tseng, G. C. (2014). Missing Value Imputation in High-Dimensional Phenomic Data: Imputable or Not, and How?. *BMC Bioinformatics*, 15(1): 346.

Liu, B., Hennessy, E., Oh, A., Dwyer, L., and Nebeling, L. (2018). Comparison of Multiple Imputation Methods for Categorical Survey Items with High Missing Rates: Application to the Family Life, Activity, Sun, Health and Eating (FLASHE) Study. *Journal of Modern Applied Statistical Methods*, 17(1): 23.

Lohaka, H. O. (2007). *Making A Grouped-Data Frequency Table: Development and Examination of the Iteration Algorithm*. Unpublished Ph.D. Thesis, Ohio University, USA.

Lewandowsky, S. and Spence, I. (1989). The Perception of Statistical Graphics, *Sociological Methods and Research*, 18(2-3): 200-242.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting Outliers: Do not Use Standard Deviation Around the Mean, Use Absolute Deviation Around the Median, *Journal of Experimental Social Psychology*, 49(4): 764-766.

Birge, L. and Rozenholc, Y. (2006). How Many Bins Should be Put in a Regular Histogram. *ESAIM: Probability and Statistics*, 10(2006): 24 - 45.

Manikandan, S. (2011). Frequency Distribution. *Journal of Pharmacology and Pharmacotherapeutics*, 2(1): 54–56.

Mann, H. B. and Wald, A. (1942). On the Choice of the Number of Class Intervals in the Application of the Chi Square Test. *The Annals of Mathematical Statistics*, 13(3): 306-317.

Mahendran, S. and Turaj, V. (2011). *Explorarory Data Analysis for Almost Anyone*. Serdang, Universiti Putra Malaysia Press.

McKnight, P. E., McKnight, K. M. , Sidani, S., and Figuredo, A. J. (2008). *Missing Data: A Gentle Introduction*. New York, The Guilford Press.

Murray, J. F. and Andrea, P. M. (2009). Understanding Descriptive Statistics. *Australian Critical Care*, 22(2009): 93-97.

Murray, R. S. and Larry, J. S. (2008). *Schaum's Outline, Theory and Problems of Statistics* . Newyork, McGraw-Hill.

Murray, J. S. (2018). Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science*, 33(2): 142-159.

Myatt, G. J. and John, W. P.(2014). *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. New Jersey, John Wiley & Sons, Inc.

Nalla, Z. (2018). Premier League Results of Each Match and Stats of Each Team from Season 2006/2007 to 2017/2018. Retrieved June 02, 2020 from `https://www.kaggle.com/zaeemnalla/premier-league`.

Newbold, P., Carlson, W. and Thorne, B. (2009). *Statistics for Business and Economics*. Boston, Pearson Education.

Nuzzo, R. L. (2019). Histograms: A Useful Data Analysis Visualization *PM&R*, 11(3): 309-312.

Oja, H. (1981). On Location, Scale, Skewness and Kurtosis of Univariate Distributions. *Scandinavian Journal of Statistics*, 8(3): 154-168.

Pawitan, Y. (2001). In All Likelihood: Statistical Modelling and Inference Using Likelihood. New York, Oxford University Press.

Pham-gia, T. and Hung, T. L. (2001). The Mean and Median Absolute Deviations. *Mathematical and Computer Modeling*, 34(1): 921-936.

Peter, H. (1992). On the Removal of Skewness by Transformation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1): 221-228.

Porter, M. M. and Niksiar, P. (2018). Multidimensional Mechanics: Performance Mapping of Natural Biological Systems Using Permutated Radar Charts. *PLOS One*, 13(9): e0204309.

Pernet, C. R., Wilcox, R., and Rousselet, G. A. (2013). Robust Correlation Analyses: False Positive and Power Validation Using a New Open Source Matlab Toolbox, *Frontiers in Psychology*, 6(606): 1-18.

Philips, M. J. (1993). Contingency Tables with Missing Data, *The Statistician*, 42(1): 9-18.

Ping, H. L. and Ataharul Islam, M. (2008). Analyzing Incomplete Categorical Data: Revisiting Maximum Likelihood Estimation (Mle) Procedure, *Journal of Modern Applied Statistical Methods*, 7(2): 488-500.

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

Premier League Player Stats. Retrieved May 27, 2019 from `https://www.premierleague.com/stats/top/players/goals?se=210`.

PhyAmal (2019). Glasgow Weather Data. Retrieved January 17, 2020 from `https://www.kaggle.com/phyamal/glasgow-weather-data-20152019`.

Rayner, J. C. W., Best, D. J., and Mathews, K. L. (1995). Interpreting the Skewness Coefficient. *Communications in Statistics - Theory and Methods*, 24(3): 593–600.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J. and Sabeti, P. C. (2011). Detecting Novel Associations in Large Data Sets. *Science*, 334(6062): 1518-1524.

Roberts, J. B. (2018). The Real Number System in an Algebraic Setting. New York, Courier Dover Publications.

Rosco, J. F., Pewsey, A., and Jones, M. C. (2013). On Blest's Measure of Kurtosis Adjusted for Skewness. *Communications in Statistics - Theory and Methods*, 44(17): 3628–3638.

Roscoe, J. T. (1975). Fundamental Research Statistics for the Behavioral Sciences. New York, Holt Rinehart and Winston.

Rousseeuw, P. J., and Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424): 1273–1283.

Royston, E. (1956). Studies in the History of Probability and Statistics: III. A Note on the History of the Graphical Presentation of Data. *Biometrika*, 43(3/4): 241-247.

Rubin, D. B. (1976) Inference and Missing Data. *Biometrika*, 63(3): 581-592.

Rubin, D. B. (1987). Multiple Imputation for Non-Response in Surveys. New York, John Wiley & Sons.

Rudemo, M. (1982). Empricial Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9(2): 65-78.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London, Chapman and Hall.

Schmitt, P., Mandel, J., and Guedj, M. (2015). A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics & Biostatistics*, 6(1): 1-6.

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

Scott, D. W. (1985). Frequency Polygons: Theory and Application. *Journal of the American Statistical Association*, 80(390): 348-354.

Scott, D. W. (1992). Multivariate Density Estimation: Theory, Practice and Visualisation. New York, John Wiley & Sons, Inc.

Scott, D. W. (2009). Sturges' Rule. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3): 303-306.

Seattle (2017). Seattle Road Weather Information Stations. Retrieved May 24, 2019 from `https://www.kaggle.com/city-of-seattle/seattle-road-weather-information-stations`.

SDSN (2019). World Happiness Report. Retrieved October 21, 2019 from `https://www.kaggle.com/unsdsn/world-happiness`.

Sheppard, W. F. (1898). 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.

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

Shimazaki, H. and Shinomoto, S. (2007). A Method for Selecting the Bin Size of a Time Histogram. *Neural Computation*, 19(6): 1503–152.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. London, Chapman and Hall.

Stavseth, M. R., Clausen, T., and Røislien, J.(2019). How Handling Missing Data May Impact Conclusions: A Comparison of Six Different Imputation Methods for Categorical Questionnaire Data. *SAGE Open Med*, 2019; 7: 2050312118822912.

Stela, P. H., Benjamin, D. and Iztok, H. (2005). Estimating the Mean and Variance from the Median, Range, and the Size of a Sample. *BMC Medical Research Methodology*, 5(13): 1-10.

Stillwell, J. (2013). *The Real Numbers: An Introduction to Set Theory and Analysis* New York, Springer.

Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, 21(153): 65-66.

Szekely, G. J., Rizzo, M. L., and Bakirov, N. K., (2007). Measuring and Testing Independence by Correlation of Distances. *Annals of Statistics*. 35(6): 2769–2794.

Templ, M., Alfons, A., and Filzmoser, P. (2011). Exploring Incomplete Data Using Visualization Techniques. *Advances in Data Analysis and Classification*, 6(1): 29–47.

Terrell, G. R. and Scott, D. W. (1985). Oversmoothed Non-parametric Density Estimates. *Journal of the American Statistical Association*, 80(389): 209-214.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Connecticut, Graphics Press.

Tukey, W. J. (1977). *Explorarory Data Analysis*. Boston, Addison-Wesley Publishing Company, Inc.

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3): 1-67.

Velleman, P. F. and Hoaglin, D. C. (2004). *Applications, Basics, and Computing of Explorarory Data Analysis*. New York, Internet First University Press.

Wand, M.P. (1997). Data-based Choice of Histogram Bin Width. *The American Statistician*, 51(1): 59-64.

Wang, X. X. and Zhang, J. F. (2012). Histogram-Kernel Error and Its Application for Bin Width Selection in Histograms. *Mathematicae Applicatae Sinica, English Series*, 28(3): 607–624.

Wilcox, R. (1994). The Percentage Bend Correlation Coefficient. *Psychometrika*, 59 (4): 601-616.

Xu, D. and Y. Wang (2020). Area-Proportional Visualization for Circular Data. *Journal of Computational and Graphical Statistics*, 29(2): 351-357.

Yule, G. U. (1911). An Introduction to the Theory of Statistics. London, Charles griffin and company.

Zainuri, N. A., Jemain, A, and Muda, N. A. (2015). Comparison of Various Imputation Methods for Missing Values in Air Quality Data. *Sains Malaysiana*, 44(3): 449 - 456.

Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, 85(11): 2541-2552.

Zwillinger, D. and Kokoska, S. (2000). *CRC, Standard Probability and Statistics Tables and Formulae*. Newyork, Chapman and Hall CRC Press.

BIODATA OF STUDENT

The student, Mohammed Mohammed Bappah, was born On 16 May 1982. He respectively obtained his national diploma and bachelor degree in statistics from Federal Polytechnic Damaturu and Modibbo Adama University of Technology Yola, Nigeria. He finished his Master of Sciences, in Statistics from the University of Ilorin, Nigeria, in 2015. He is currently a PhD candidate in the area of Exploratory Data Analysis (EDA). His research interest is in Exploratory Data Analysis, Extreme Value, Circular Statistics, and Survival Analysis.

LIST OF PUBLICATIONS

The following are the list of publications that arise from this study.

Mohammed, M. B., Adam, M.B., Zulkafli, H. S., & Ali, N. (2020). Improved Frequency Table with Application to Environmental Data. *Mathematics and Statistics; 8(2): 201-210*.

Mohammed, M. B., Adam, M.B., Ali, N., & Zulkafli, H. S. (2020). Improved Frequency Table's Measures of Skewness and Kurtosis with Application to Weather Data. *Communications in Statistics - Theory and Methods; doi :10.1080/03610926.2020.1752386.*.

Mohammed, M. B., Adam, M.B., Ali, N., & Zulkafli, H. S. (2021). Comparison of Five Imputation Methods in Handling Missing Data in a Continuous Frequency Table, *AIP Conference Proceedings; 2355, 040006 (2021); https://doi.org/10.1063/5.0053286.* .

Mohammed, M. B., Adam, M.B., Ali, N., & Zulkafli, H. S. A Novel Frequency Table for Discrete Data. *Pakistan Journal of Statistics*, (Under Review).

Adam, M. B., **Mohammed, M.B.**, Subhi, M. J. & Jamsari, A. A. Improvement of Statistic values from Frequency Table for Continuous Symmetrical Positive Data, *Pakistan Journal of Statistics and Operations Research*, (Under Review).

Mohammed, M. B., Adam, M.B., Ali, N., & Zulkafli, H. S. Exploration of COVID-19 Pandemic Using Variety of Frequency Tables, *Example and Counterexample*, (Under Review).

Mohammed, M. B., Adam, M.B., Ali, N., Zulkafli, H. S., & Olaniran, O. R. A Novel Bivariate Discrete Frequency Table, *Cogent Mathematics & Statistics*, (Under Review).

Adam, M. B., **Mohammed, M.B.**, Fernando, M., & Yong, L. New Measure of Linear Association Based on Two-way Contingency Tables, *Journal of Modern Applied Statistical Methods*, (Under Review).

Mohammed, M. B., Adam, M.B., Ali, N., & Zulkafli, H. S. New Class Width Rule for Continuous Frequency Tables, (To submit for publication).