New Discrete Frequency Table with Application to Real Data

[1]Mohammed, M. B. , [2]Hani, S. Z. , [2]Ali, N., [3]Ahmed, H., & [1]Salihu, H. D.

[1]Department of Mathematics and Statistics, Federal University of Kashere, Gombe State, Nigeria.
[2]Department of Mathematics, Universiti Putra Malaysia, Malaysia.
[3]Moddibo Adama University of Technology, Yola, Adamawa State, Nigeria.
**Corresponding author: muhdb1@gmail.com*

Abstract

One way to make sense of data is to organize it into a more meaningful format called frequency table. The existing univariate discrete frequency table is simple to construct, easy to understand and interpret. However, when the number of elements in the data is substantial, it results in a long table that can be difficult to handle. This article presents a new discrete frequency table. The proposed frequency table is described, using simulations performed on five different discrete distributions, and real data. The new frequency table improves the existing table in various ways. This includes how the data with a large number of elements can be handled, how the mode of the data can be better estimated, and how the essential features of the data can be better revealed.

Keywords: Frequency table, Discrete data, Discrete Distributions, Mode, Simulation.
AMS 2020 subject classifications 62N01, 62N02.

1. Introduction

A set of data in a raw format do not make meaningful representation unless it is organized in a systematic manner, such as a frequency table. The raw data can be partitioned into classes of suitable sizes, showing observations together with their corresponding frequencies. When data are systematically organized in this form, it is called a frequency distribution table (Kenney, 1939; Manikandan, 2011; Mohammed et al., 2020b). Data are more attractive and capture the minds of researchers if presented in either tabular or graphical form.The tabular representations are precise and provide the reader with apparent features of the data; however, the graphical representations have more visual significance since they are useful in detecting patterns in a data set (Beniger and Robyn, 1978; Davies, 1929). Generally, the main purpose of organizing raw data is to explore the extra information. These involve determining the underlying distribution, understanding the characteristics of variables, and comprehending the statistical tool to be used for inference. Data presentation and visualization techniques such as tables, histograms, and frequency polygons, often require the determination of the class intervals, grouping the data into these class intervals.This grouping is often needed for the preparation of frequency distribution tables, construction of histograms, and formation of frequency polygons. Nevertheless, most of these techniques are used when the data are continuous (Kenney, 1939; Gravetter and Wallnau, 2000).

Discrete data are count observations, which assumes only finite variates. The table that displays the elements together with their corresponding frequencies is called discrete frequency table. In this table, the observations are exact class limits; there are no class boundaries (Gravetter and Wallnau, 2000).

The frequency table plays a significant role in statistics. It serves as a bridge between raw data and the bar plot. Furthermore, from the frequency table, the nature of the distribution of the data can be known. For instance, to know whether the distribution is normal or skewed or the degree of concentration of the elements. An important function to mention is, the frequency table aid careful comparison of datasets. So also, various statistical measures can be obtained from the frequency table (Kenney, 1939; Mohammed et al., 2020a; Mohammed et al., 2020b).

Though the existing discrete frequency table organizes data in a systematic way, when the number of elements in a data is large enough, the table can be complicated. Therefore, proposing a new frequency table, contained a dataset with a large number of elements is necessary.

2. Existing Univariate Discrete Frequency Table

Let $x_1, x_2, \cdots, x_n$ be n discrete observations , Assuming the observations contained $e_1, e_2, \cdots, e_k$ elements which appeared $f_1, f_2, \cdots, f_k$ times, Table 1 is the existing discrete frequency table.

Table 1. Discrete Frequency Table

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

Where m is the number of elements, e_i is the element in class i, f_i is the number of frequency of x_i , $i = 1,2, \cdots, k$ and $n = \sum f_i$.

2.1 Number of Classes

The number of classes needed to construct the frequency table for continuous data depends mainly on the size of the data (Kenney,1939). Several rules for choosing the number of classes were reported in the literature. Among the popular rules are the Sturges (1926), Doane (1976), Scott (1979), and Freedman and Diaconis (1981) rules. These rules have immensely contributed to choosing the suitable number of classes needed to construct a frequency table. However, the number of classes of the frequency table for discrete data depends mostly on the number of elements (m) in the dataset. When the number of elements in a dataset is small no matter how big is the data set, the frequency table can have a small number of classes. Meanwhile, when the number of elements in a data set is large, irrespective of the size of the data, the frequency table can result in a large number of classes.

3. Proposed Discrete Frequency Table

The existing univariate discrete frequency table can be complicated when the number of elements in the data (m) is substantial. The proposed frequency table that can be constructed by grouping the elements presented in Table 2 provides the best alternative when the number of elements is large. The proposed frequency table with elements grouped into two is given in two different cases. Table 4 describes the first case, Tables 5 and 6 demonstrate the second case using two different arrangements. The first case is when the number of elements is even, the classes contain an equal number of elements. Whereas the second case is when the number of elements is odd, all the classes contained the same number of elements except the first or last class, which contains a different number of elements. The first or last class can have only one element, while each of the other classes contains two elements. Generally, continuous data are measured values such as amount rainfall, height, weight and so on.
In general, when the number of elements in the data is not a multiple of the number of suitable elements needed to group the elements, the first or last class can have a different number of elements, and other classes contain the same number of elements. In other words, when $m \text{mod} g = 0$, the classes contain the same number of elements; otherwise, either the first or last class contains a different number of elements. However, when this number m, in the data is less than 10, ($m < 10$), the table, Table 3 can be used; modification is not required.

Table 2. Proposed g-Element Univariate Discrete Frequency Table

Class	Class Interval $(e_i, \cdots, e_{n_i})$	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	$e_1, \cdots, e_{n_1}$	f_1	x_{mo_1}	x_{me_1}	x_{md_1}	R_{m_1}
2	$e_{n_1+1}, \cdots, e_{n_2}$	f_2	x_{mo_2}	x_{me_2}	x_{md_2}	R_{m_2}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$c-1$	$e_{n_{c-2}+1}, \cdots, e_{n_{c-1}}$	f_{c-1}	$x_{mo_{c-1}}$	$x_{me_{c-1}}$	$x_{md_{c-1}}$	$R_{m_{c-1}}$
c	$e_{n_{c-1}+1}, \cdots, e_{n_c}$	f_c	x_{mo_c}	x_{me_c}	x_{md_c}	R_{m_c}

Let c, n_i , f_i , x_{mo_i}, R_{m_i} respectively be the number of classes, number of elements, frequency, mode, and mode rank of class i. Also, the adjusted mean and median, X_{me} and X_{md}, are respectively the rounded arithmetic mean and median of class i's observations. If $m \text{ mod} g = 0$ the proposed frequency table is a complete table, that is, $n_1 = n_2 = \cdots = n_c = g$. However, when $m \text{ mod } g \neq 0$, the table is incomplete. This condition results to either $n_1 \neq g$ or $n_c \neq g$, and the proposed univariate discrete frequency table can be presented using two diffrent cases. Case I is to present the table with $n_2 = n_3 =$

$\cdots = n_c = g$, but $n_1 < g$. While, Case II is to depict the table with $n_1 = n_2 = \cdots = n_{c-1} = g$, but $n_c < g$.

Furthermore, the proposed table can be illustrated by grouping the elements into three, Tri-element discrete frequency table. This can be obtained by setting g = 3 in Table 2. The table is a complete table if $n_1 = n_2 = \cdots = n_c = 3$. However, when m mod 3 $\neq 0$, the table is incomplete with either $n_1 \neq 3$ or $n_c \neq 3$.

Table 3. Univariate Discrete Frequency Table when m is Less than 10

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

Table 4. Proposed Bi-element Univariate Discrete Frequency Table when m is Even

Class	Class Interval (e_i, e_{i+1})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	e_1, e_2	f_1	x_{mo_1}	x_{me_1}	x_{md_1}	R_{m_1}
2	e_3, e_4	f_2	x_{mo_2}	x_{me_2}	x_{md_2}	R_{m_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$c-1$	e_{m-2}, e_{m-3}	f_{c-1}	$x_{mo_{c-1}}$	$x_{me_{c-1}}$	$x_{md_{c-1}}$	$R_{m_{c-1}}$
c	e_{m-1}, e_m	f_c	x_{mo_c}	x_{me_c}	x_{md_c}	R_{m_c}

Table 5. Proposed Bi-element Univariate Discrete Frequency Table when m is Odd Case I

Class	Class Interval (e_i, e_{i+1})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	e_1, e_2	f_1	x_{mo_1}	x_{me_1}	x_{md_1}	R_{m_1}
2	e_3, e_4	f_2	x_{mo_2}	x_{me_2}	x_{md_2}	R_{m_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$c-1$	e_{m-1}, e_{m-2}	f_{c-1}	$x_{mo_{c-1}}$	$x_{me_{c-1}}$	$x_{md_{c-1}}$	$R_{m_{c-1}}$
c	e_m	f_c	x_{mo_c}	x_{me_c}	x_{md_c}	R_{m_c}

Table 6. Proposed Bi-element Univariate Discrete Frequency Table when m is Odd Case II

Class	Class Interval (e_i, e_{i+1})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	e_1	f_1	x_{mo_1}	x_{me_1}	x_{md_1}	R_{m_1}
2	e_2, e_3	f_2	x_{mo_2}	x_{me_2}	x_{md_2}	R_{m_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$c-1$	e_{m-2}, e_{m-3}	f_{c-1}	$x_{mo_{c-1}}$	$x_{me_{c-1}}$	$x_{md_{c-1}}$	$R_{m_{c-1}}$
c	e_{m-1}, e_m	f_c	x_{mo_c}	x_{me_c}	x_{md_c}	R_{m_c}

The notation c is the number of classes, e_i, e_{i+1} are two neighboring elements, f_i is the number of occurences of the elements in class i and x_{moi} is the mode of elements in class i, which represent the magnitude of elements in that class.

3.1 Mode of the Proposed Discrete Frequency Table

The mode is regarded as the observation which occurred the most. There may be no mode, one mode, or two or more modes. A dataset has no mode when its distribution has no peak, one mode if it has one peak, two modes when its distribution has two peaks, and more than two modes if its distribution has multiple peaks. The mode (x_{mo}) represents the magnitude of observations in each class because it is the most suitable measure of location when dealing with the discrete data. The arithmetic mean is not discrete, and the median is discrete if and only if the number of observations in the data is odd. Since the mode is the observation with the highest frequency, in the proposed frequency table's class, the element with the highest frequency is the mode of that class. There would be more than one mode if more than one element had the same highest frequency. Furthermore, to obtain the mode of the proposed frequency table, ranks are assigned to the classes' modes in ascending order of the classes' frequencies. Apart from the mode, two other statistics, adjusted mean and median, are proposed and can serve as the measures of location of the proposed discrete frequency table. The adjusted mean and median are the arithmetic mean and median rounded to the nearest integer.

3.2 Algorithm for Obtaining the Mode of the Proposed Discrete Frequency Table

1. Among the modes of all the classes, check which class mode has the highest frequency.
2. If $x_{mo_1}, x_{mo_2}, \cdots, x_{mo_c}$ are obtained, the data has no mode.
3. If x_{mo_i} is appeared, the mode of the data is x_{mo_i} $i = 1,2,\cdots,c$.
4. If step 2 and 3 do not hold, the data has more than one mode.

Alternatively, the mode can be given as

$$Mode = \begin{cases} No\ \ Mode & if \quad R_{m_1} = R_{m_2}, \cdots, = R_{m_c} \\ x_{moi}, & if \quad R_{m_i} = 1, \\ More\ \ than\ \ one\ \ mode & if \quad Otherwise \end{cases}, \tag{1}$$

The integer c denotes the number of classes, M_{0_i} is the mode of class i, and R_{m_i} is the mode rank of class i, $i = 1,2,\cdots,c$. The number of unique observations that can be grouped depends on the number of classes in the existing frequency table. The question, how many classes to be used remain unsatisfactorily answered. The rule of thumb recommended by most of the statistical books is to have ten classes as optimal and maximum of 30 (Dogan and Dogan, 2010). Therefore, the criteria for grouping the unique observations is by considering the neighboring observations, either in ascending or descending order, since they have similar characteristics.

Modifying the Cochran (1954) rule with the number of elements instead of the sample size (n) we derived a formula for grouping the elements (g) as

$$g = \begin{cases} 1, & if \quad m < 10 \\ \sqrt{\frac{m}{5}}, & if \quad m \geq 10' \end{cases} \tag{2}$$

where m is the number of elements in the dataset and g is the grouping number.

3.3 Algorithm for Constructing the Proposed Discrete Frequency Table

1. Get or generate discrete data with m elements.
2. If m is less than 10, Table 3 can be used; modification is not required.
3. Obtain the suitable number of elements to partition, g, using Equation 2.
4. Construct the proposed discrete frequency table using the R code in the R package.

4. Results and Discussion

4.1 Simulation Studies

The proposed univariate discrete frequency table is described by performing two different simulation studies using binomial, Poisson, and geometric distributions. The data simulation and analysis are performed using the R statistical package and R studio. The first study demonstrates the table in two grouped elements, whereas the second illustrates the proposed discrete frequency table with elements grouped into three. The sample sizes were varied depending on the distribution. Furthermore, to observe the pattern of the existing frequency tables, 1000 samples are simulated from these distributions. The numbers of elements and the frequencies of the elements in each sample are observed. One sample from each experiment is used to describe the proposed discrete frequency table with elements grouped into two and three, bi-element and tri-element. The bi-element discrete frequency table is obtained by grouping two neighboring elements. Similarly, the tri-element discrete frequency table is constructed by grouping three neighboring elements. In general, the proposed discrete frequency table is always obtained by grouping g neighboring elements. If $m \bmod g = 0$, the new frequency table is complete; otherwise, it is incomplete, with either the first or last class having a different number of elements. Partitioning the elements in the data into classes is guided by Equation (2). The frequency, f_i, in the proposed table, is the sum of the frequencies of the elements in class i. The class mode, which represents the magnitude of elements in each class, is the class element that occurs the most. Furthermore, the adjusted mean and median, which also represent the magnitude of each class's elements, are the rounded arithmetic mean and median of each class's observations.

4.1.1 Simulations from the Binomial distribution

The pattern of elements using 1000 different samples of size 200 simulated from the binomial distribution with parameters, $n = 50$, and $\pi = 0.5$ shows that the numbers of elements are respectively lying within the intervals $16 \cdots ,22$. The smallest frequency of the 1000 samples is 1, and the highest frequency is 34. Moreover, one sample is used to construct the existing table, Table 8, and describe the proposed univariate discrete frequency table in two grouped elements, bi-element. Since $18 \bmod 2 = 0$, Table 9 shows the new table with equal number of elements in each class. Figures 1 Left, and Right are the visualizations of the existing frequency table, Table 8 and the proposed frequency table, Table 9. Besides, the mode is used to locate the central tendency of the sample data. The existing table has a mode of 24. The same value can be obtained from the proposed frequency table, Table 9. Again, to understand the total span of data, the minimum and maximum observations need to be obtained. The minimum and maximum observations in the randomly chosen sample are respectively 16 and 33. Meanwhile, the sum of all the observations in the sample is 1101. The appropriate measure of spread for the discrete data is the range, and its values can be obtained as 17.

Table 8. Existing Discrete Frequency Table Constructed Using a Sample of Size 200 Simulated from the Binomial Distribution with Parameters $n = 50$ and $\pi = 0.5$

Class	Element (e)	Frequency (f)
1	16	2
2	17	1
3	18	7
4	19	5
5	20	2
5	21	9
7	22	18
8	23	23
9	24	29
10	25	19
11	26	23
12	27	18
13	28	12
14	29	9
15	30	7
16	31	9
17	32	6
18	33	1

Table 9. Proposed Bi-element Discrete Frequency Table

Class	Class Interval (e_i, e_{i+1})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	16,17	3	16	16	16	8
2	18,19	12	18	18	18	6
3	20,21	11	21	21	21	7
4	22,23	41	23	23	23	2.5
5	24,25	48	24	24	24	1
6	26,27	41	26	26	26	2.5
7	28,29	21	28	28	28	4
8	30,31	16	31	31	31	5
9	32,33	7	32	32	32	7

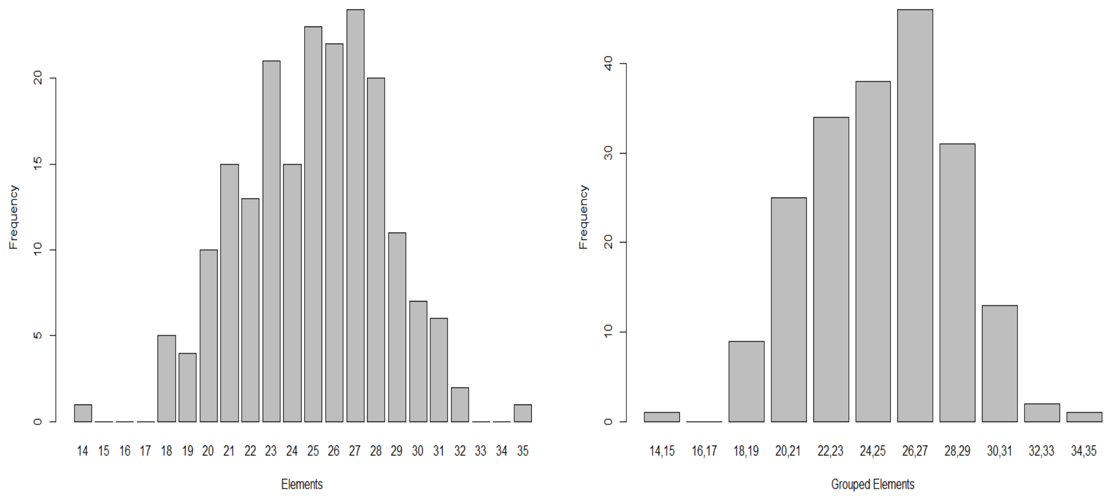


Figure 1. Left: Bar Plot of the Existing Discrete Frequency Table, Right: Bar Plot of the Bi-Element Discrete Frequency Table Constructed Using a Sample of Size 200 Simulated from the Binomial Distribution with Parameters $n = 50$ and $\pi = 0.5$

The second study using 1000 samples of size 500 simulated from the binomial distribution, with parameters $n = 100$ and $\pi = 0.5$, shows that the number of elements is within the range $25, \cdots ,32$. Also, the least frequency of the elements is 1, and the maximum is 60. A randomly chosen sample out of the 1000 was used to depict the existing table, Table 10, and demonstrate the new frequency table in three grouped elements, tri-element, Table 11. An advantage of the new frequency table over the existing is the ability to accommodate many elements. Figure 2 Left, and Right are the bar plots of the existing and the new frequency tables. The mode can be obtained from the existing table, Table 10 as well as the new frequency table, Table 11 as 50. The smallest value in the sample is 36, while the largest value is 68. In addition, the sum of all the values in the sample is 25055. The suitable variability measure for the discrete data is the range. The range can be obtained as 32.

Table 10. Existing Univariate Discrete Frequency Table Constructed Using a Sample of Size 500 Simulated from the Binomial Distribution with Parameters $n = 100$ and $\pi = 0.5$

Class	Element (e)	Frequency (f)
1	36	1
2	37	1
3	38	2
4	39	1
5	40	4
6	41	6
7	42	11
8	43	19
9	44	19
10	45	19
11	46	33
12	47	36
13	48	43
14	49	38
15	50	45
16	51	30
17	52	34
18	53	32
19	54	32
20	55	30
21	56	10
22	57	18
23	58	15
24	59	6
25	60	6
26	61	2
27	62	5
28	63	1
29	64	0
20	65	0
31	66	0
32	67	0
33	68	1

Table 11. Proposed Tri-element Discrete Frequency Table

Class	Class interval (e_i, e_{i+1}, e_{i+2})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	36,37,38	4	38	37	38	9
2	39,40,41	11	41	40	41	8
3	42,43,44	49	43,44	43	43	5
4	45,46,47	88	47	46	46	3
5	48,49,50	126	50	49	49	1
6	51,52,53	96	52	52	52	2
7	54,55,56	72	54	55	55	4
8	57,58,59	39	57	58	58	6
9	60,61,62	13	60	61	61	7
10	63,64,65	2	63	63	63	10.5
11	66,67,68	2	68	68	68	10.5

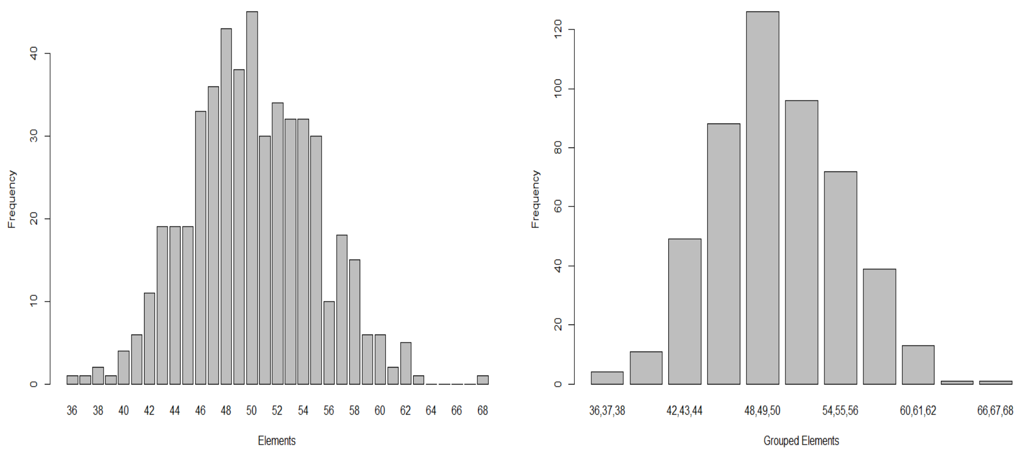


Figure 2. Left: Bar Plot of the Existing Discrete Frequency Table, Right: Bar Plot of the Tri-Element Discrete Frequency Table Constructed Using a Sample of Size 500 Simulated from the Binomial Distribution with Parameters $n = 100$ and $\pi = 0.5$

4.1.2 Simulations from the Poisson Distribution

Using skewed discrete data, simulations are carried out from the Poisson distribution. In the first study, the nature of the elements using 1000 samples of size 100000 simulated from the Poisson distribution with rate $\mu =$ 3.5 indicates that the numbers of elements are within the interval $14, \cdots, 17$, and the smallest frequency is 1 up to the maximum, 21832. Table 12 displays the existing table while the new frequency table is displayed as Table 13. The tables are prepared using one sample out of the 1000 samples. It can be observed that the proposed table is described using one table; this is because of $16 \bmod 2 = 0$. Also, the bar plot of the new frequency table, Figure 3 Right, is more straightforward and displays the essential features of the data compared with that of the existing table, Figure 3 Left. The mode of the sample data in the existing table, Table 14, as well as its visualization, Figure 3, is 3. Similar value of mode can be obtained from the new frequency table, Table 13. The minimum and maximum observations in the chosen sample are 0 and respectively. Meanwhile, the sum of all the observations in the sample is 340400. The measure of variability in the discrete data range can be obtained as 15.

Table 12. Existing Univariate Discrete Frequency Table Constructed Using a Sample of Size 100,000 simulated from Poisson distribution with rate $\mu = 3.5$

Class	Element (e)	Frequency (f)
1	0	3082
2	1	10437
3	2	18436
4	3	21594
5	4	18964
6	5	13207
7	6	7766
8	7	3826
9	8	1648
10	9	667
11	10	258
12	11	77
13	12	26
14	13	8
15	14	2
16	15	2

Table 13. Proposed Bi-element Discrete Frequency Table

Class	Class interval (e_i, e_{i+1})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	0,1	13519	1	1	1	3
2	2,3	40030	3	3	3	1
3	4,5	32171	4	4	4	2
4	6,7	11592	6	6	6	4
5	8,9	2315	8	8	8	5
6	10,11	335	10	10	10	6
7	12,13	34	12	12	12	7
8	14,15	4	14,15	15	15	8

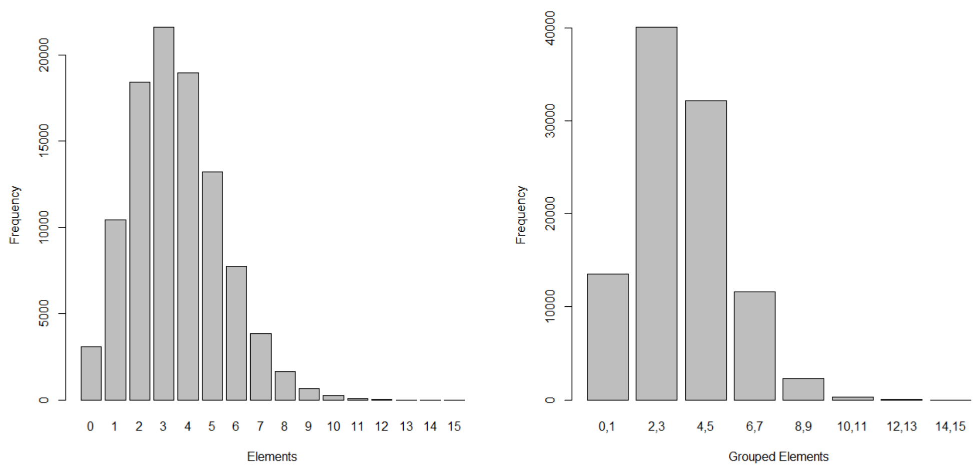


Figure 3. Left: Bar Plot of the Existing Discrete Frequency Table, Right: Bar Plot of the Bi-Element Discrete Frequency Table Constructed Using a Sample of Size 100000 Simulated from the Poisson Distribution with Parameter $\mu = 3.5$

A second study of 1000 samples of sizes 100000 simulated from the Poisson distribution with parameter $\mu = 5.5$ shows that the number of elements is within the range $18,\cdots,21$ and the minimum frequency is 1, while the maximum is 17466. The existing table, Table 14 and the new frequency table, Table 15, are constructed using one sample out of the one thousand. The new discrete frequency table is described using only one table, Table 14, with an equal number of elements in each class. Figure 4 Left and Right are the visualisations of the existing and the new frequency table. In addition, from both the existing table, Table 14, and the proposed frequency table, Table 15 the value of mode is 5. The minimum and maximum values in the randomly chosen sample are 0 and 20 respectively. Whereas, the sum of all the observations in the sample is 550400. The measure of variability in the discrete data range can be obtained as 20.

Table 14. Existing Univariate Discrete Frequency Table Constructed Using a Sample of Size 100,000 Simulated from Poisson Distribution with Rate $\mu =$ 5.5

Class	Element (e)	Frequency (f)
1	0	420
2	1	2254
3	2	6131
4	3	11359
5	4	15679
6	5	16868
7	6	15623
8	7	12546
9	8	8565
10	9	5264
11	10	2758
12	11	1435
13	12	638
14	13	284
15	14	117
16	15	37
17	16	14
18	17	4
19	18	2
20	19	1
21	20	1

Table 15. Proposed Tri-element Discrete Frequency Table

Class	Class Interval (e_i, e_{i+1}, e_{i+2})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	0,1,2	8805	2	2	2	4
2	3,4,5	43906	5	4	4	1
3	6,7,8	36734	6	7	6	2
4	9,10,11	9457	9	10	9	3
5	12,13,14	1039	12	13	12	5
6	15,16,17	55	15	15	15	6
7	18,19,20	4	18	19	19	7

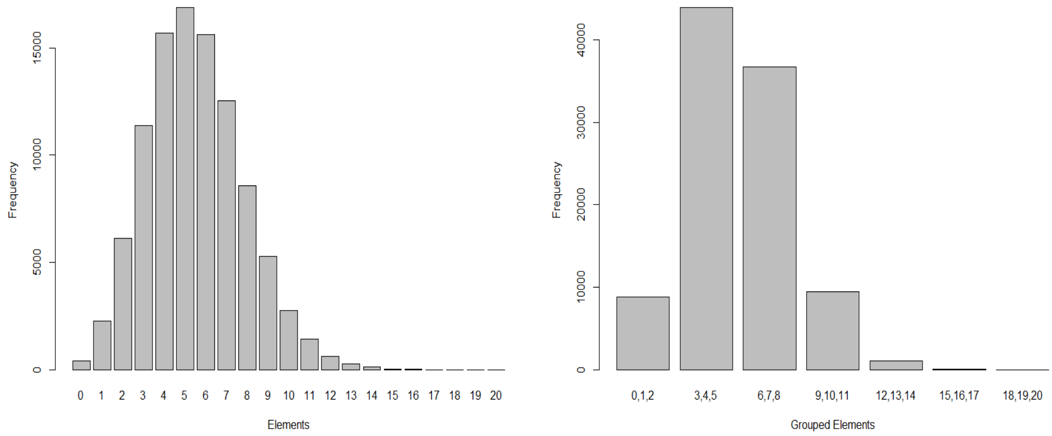


Figure 4. Left: Bar Plot of the Existing Discrete Frequency Table, Right: Bar Plot of the Tri-Element Discrete Frequency Table Constructed Using a Sample of Size 100000 Simulated from the Poisson Distribution with Rate $\mu = 5.5$

4.1.3 Simulations from the Geometric Distribution

The first study using 1000 samples of size 1000 simulated from Geometric distribution with parameter $\pi = 0.4$ indicates that the number of elements falls within the range $11, \cdots, 17$, and the least frequency is 1, while the maximum frequency is 450. The new frequency table, Table 17, constructed using one of the samples is more manageable than the existing frequency table, Table 16. Figure 5 Left is the bar plot of the existing table, while Figure 5 Right shows the bar plot of the proposed frequency table. The frequency table, Table 16 and the bi-element frequency table, Table 17, give a similar value of mode , 0. The minimum and maximum values in the data are 0 and 11, respectively. Meanwhile, the sum of the observations in the chosen sample is 1421. The measure of spread range can be obtained from the randomly chosen sample as 11.

Table 16. Existing Univariate Discrete Frequency Table Constructed Using a Sample of Size 1000 Simulated from Geometric Distribution with Parameter $\pi = 0.4$.

Class	Element (e)	Frequency (f)
1	0	415
2	1	242
3	2	126
4	3	87
5	4	63
6	5	28
7	6	23
8	7	7
9	8	1
10	9	3
11	10	3
12	11	2

Table 17. Proposed Bi-element Discrete Frequency Table

Class	Class interval (e_i, e_{i+1})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	0,1	657	0	0	0	1
2	2,3	213	2	2	2	2
3	4,5	91	4	4	4	3
4	6,7	30	6	6	6	4
5	8,9	4	9	9	9	6
6	10,11	5	10	10	10	5

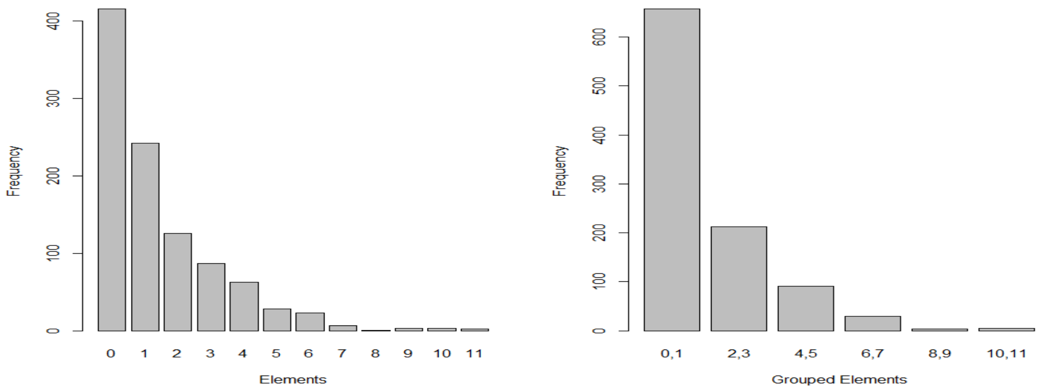


Figure 5. Left: Bar Plot of the Existing Discrete Frequency Table, Right: Bar Plot of the Bi-Element Discrete Frequency Table Constructed Using a Sample of Size 1000 Simulated from the Geometric Distribution with Parameter $\pi = 0.4$

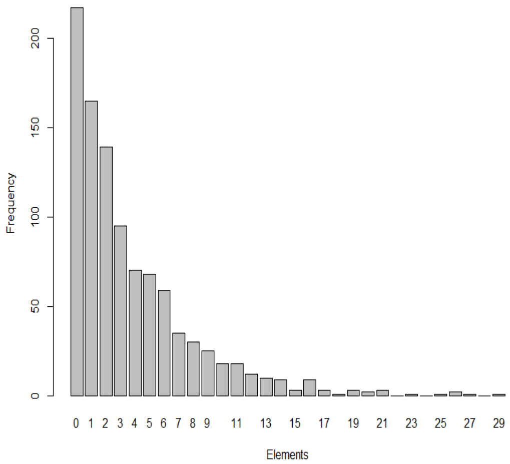
In the second study, the features using 1000 samples of size 1000 simulated from the geometric distribution with parameter $\pi = 0.2$ reveals that the number of the elements lies in the interval $22, \cdots, 32$, and the minimum frequency of the 1000 samples is 1 while the maximum is 233. The existing and tri-element discrete frequency tables are constructed using one of the 1000 samples. Tables, 18 and 19 are respectively the existing and the new discrete frequency tables. The two frequency tables are visualised as Figure 6 Left and Right. The same value of mode, 0, can also be obtained from the existing as well as the tri-element frequency tables. The minimum and maximum observations in the randomly chosen sample are 0 and 29, respectively. Meanwhile, the sum of the observations in the sample is 3764. The measure of variability, range, can be obtained from the data as 29.

Table 18. Existing Univariate Discrete Frequency Table Constructed Using a Sample of Size 1000 Simulated from Geometric Distribution with Parameter $\pi = 0.2$

Class	Element (e)	Frequency (f)
1	0	217
2	1	165
3	2	139
4	3	95
5	4	70
6	5	68
7	6	59
8	7	35
9	8	30
10	9	25
11	10	18
12	11	18
13	12	12
14	13	10
15	14	9
16	15	3
17	16	9
18	17	3
19	18	1
20	19	3
21	20	2
22	21	3
23	22	0
24	23	1
25	24	0
26	25	1
27	26	2
28	27	1
29	28	0
30	29	1

Table 19. Proposed Tri-element Discrete Frequency Table

Class	Class Interval (e_i, e_{i+1}, e_{i+2})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	0,1,2	521	0	1	1	1
2	3,4,5	233	3	4	4	2
3	6,7,8	124	6	7	7	3
4	9,10,11	61	9	10	10	4
5	12,13,14	31	12	13	13	5
6	15,16,17	15	15,17	16	16	6
7	18,19,20	6	19	19	19	7
8	21,22,23	4	21	22	21	8
9	24,25,26	3	26	26	26	9
10	27,28,29	2	27,29	28	28	10



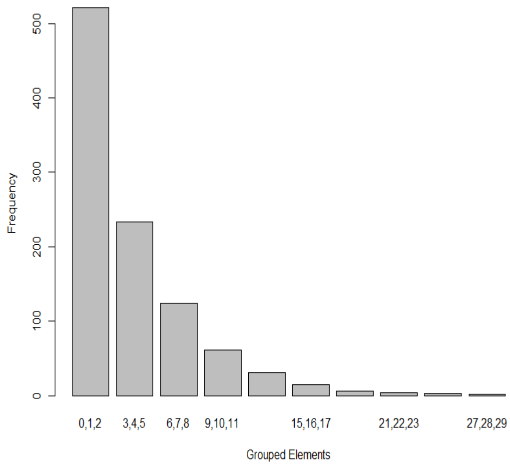


Figure 6. Left: Bar Plot of the Existing Discrete Frequency Table, Right: Bar Plot of the Tri-Element Discrete Frequency Table Constructed Using a Sample of Size 1000 Simulated from the Geometric Distribution with Parameter $\pi = 0.2$

4.2 Application

Moreover, to demonstrate the proposed discrete frequency table using real data, we used data on the number of goals scored during the 2017/2018 English premier league seasons obtained from the English premier league website (Premier League Player Stats, 2021). The numbers of players who scored at least one goal during the season is 251. The existing table, Table 20, shows that the majority of the players scored only one goal whereas only one player each scored 13, 14, 15, 16, 18, 20, 21, 30 and 32 goals. The players are Gabriel Jesus, Alexandre Lacazette, Roberto Firmino, Romelu Lukaku, Raheem Sterling, Jamie Vardy, Sergio Agüero, Harry Kane, and Mohamed Salah, respectively. As suggested by Equation (2), the proposed frequency table for the 2017/2018 season number of goals data, Tables 21 and 22, grouped the number of goals into three. The least scored goals, 1, 2, and 3, are grouped into one class, and the numbers of players who scored the three categories of goals are merged to 164. The most outstanding players, those that scored from 13 goals and above, are also grouped into three different classes. Figure 7 Bottom, Top Left and Right respectively show the bar plot of the existing and the proposed univariate discrete frequency table in two different cases.

A similar value of mode can be obtained from the existing discrete frequency table and the proposed discrete frequency table, using both the cases. The minimum and maximum observations, minimum and maximum number of goals scored, during the 2017/2018 season are 1 and 32, respectively. The sum of the observations, the total number of goals scored during the season, is 991 goals. The most suitable measure of variability, range, can be obtained from the 2017/2018 season data as 31.

Table 20. Univariate Discrete Frequency Table Constructed Using Data on the Number of Goals Scored by 251 Players for 2017/2018 English Premier League Season

Class	Element (e)	Frequency (f)
1	1	76
2	2	60
3	3	28
4	4	13
5	5	18
6	6	8
7	7	12
8	8	8
9	9	6
10	10	7
11	11	2
12	12	4
13	13	1
14	14	1
15	15	1
16	16	1
17	17	0
18	18	1
19	19	0

20	20	1
21	21	1
22	22	0
23	23	0
24	24	0
25	25	0
26	26	0
27	27	0
28	28	0
29	29	0
30	30	1
31	31	0
32	32	1

Table 21. Proposed Tri-element Univariate Discrete Frequency Table Constructed Using Data on the Number of Goals Scored by 251 Players During the 2017/2018 English Premier Season with Last Class Incomplete

Class	Class Interval (e_i, e_{i+1}, e_{i+2})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	1,2,3	164	1	2	2	1
2	4,5,6	39	5	5	5	2
3	7,8,9	26	7	8	8	3
4	10,11,12	13	10	11	10	4
5	13,14,15	3	13,14,15	14	14	5
6	16,17,18	2	16,18	17	17	6.5
7	19,20,21	2	20,21	21	21	6.5
8	22,23,24	0	0	0	0	10.5
9	25,26,27	0	0	0	0	10.5
10	28,29,30	1	30	30	30	8.5
11	31,32	1	32	32	32	8.5

Table 22. Proposed Tri-element Univariate Discrete Frequency Table Constructed Using Data on the Number of Goals Scored by 251 Players During the 2017/2018 English Premier Season with First Class Incomplete

Class	Class Interval (e_i, e_{i+1}, e_{i+2})	Frequency (f)	Class Mode (M_o)	Adjusted Mean (X_{me})	Adjusted Median (X_{md})	Mode Rank (R_m)
1	1,2	136	1	1	1	1
2	3,4,5	59	3	4	3	2
3	6,7,8	28	7	7	7	3
4	9,10,11	15	10	10	10	4
5	12,13,14	6	12	13	12	5
6	15,16,17	2	15,16	17	17	7
7	18,19,20	2	18,20	19	19	7
8	21,22,23	1	21	21	21	9
9	24,25,26	0	0	0	0	10.5
10	27,28,29	0	0	0	0	10.5
11	30,31,32	2	30,32	31	31	7

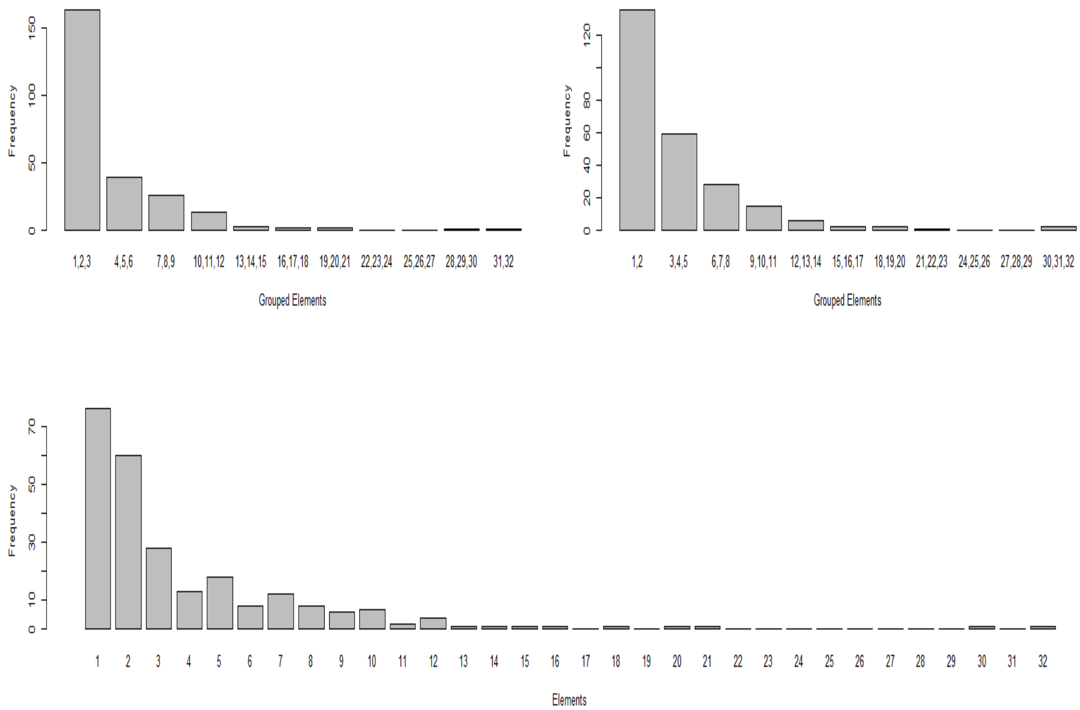


Figure 7. Top Left: Bar Plot of the Tri-Element Discrete Frequency Table Case I, Top Right: Bar Plot of the Tri-Element Discrete Frequency Table Case II, Bottom:Bar Plot of the Existing Discrete Frequency Table, Constructed Using Data on the Number of Goals Scored by 251 Players During the 2017/2018 English Premier Season

5. Summary and conclusion

5.1 Summary

In this article, we proposed a new discrete frequency table. The proposed table can be obtained by grouping the neighboring elements. To illustrate the new frequency table, we performed two different simulation studies using Binomial, Poisson, and Geometric distributions. Moreover, to further describe the proposed table using real data, we used data on the number of goals scored during 2017/2018 season, obtained from the English premier league website.

The first simulation study was performed to demonstrate the proposed frequency table by grouping the elements in the data into two. Meanwhile, the second study was conducted to describe the proposed table by partitioning the elements into three. The construction of the tables was achieved by randomly taking one sample from each of the independent experiments. The proposed frequency table showed an adequate and manageable organization as compared with the existing counterpart. In general, as the number of elements is getting larger, the existing frequency table ceases to be practically used, unlike the proposed frequency table which can handle as many elements as possible. Though Dwyer (1942) and Pierce (1943) proposed methods aimed at reducing grouping error, the way the discrete frequency table was regrouped is inappropriate. Discrete data are count observations and do not assume values within a continuous interval. The procedure for obtaining the new proposed frequency table is appropriate because it considered the dichotomy between the discrete and continuous data.

Moreover, the visualizations of the proposed frequency table for the unique observations grouped into two and three displayed the essential features and are closer to the underlying distributions of the data when compared with that of the existing frequency table.

Also, the mode is always the appropriate measure of location for discrete data. Consequently, comparing the mode of the existing frequency tables and the mode obtained from the proposed frequency tables, we observed that the same values for the modes were obtained, in fact, the mode can be better obtained using the proposed frequency table, since in the existing table, the mode can be in any of the classes.

5.2 Conclusion

A new discrete frequency table is proposed to overcome the issue of a large number of elements. The table is illustrated using data simulated from five discrete distributions. A real English premier league dataset is also used to demonstrate the proposed table.

The new discrete frequency table can be a better choice, since, it can handle data set with a large number of elements, reliably estimate the mode of the data and vividly reveals the significant features of the underlying distribution of data.

References

Beniger, J. R. and Robyn, D. L. (1978). Quantitative Graphics in Statistics: A Brief History. The

American Statistician, 32(1): 1-11.

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

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

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

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.

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

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

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

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

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

Mohammed, M. B., Adam, M. B., Ali, N., and Zulkafli, H. S. (2020a). *Improved Frequency Table's Measures of Skewness and Kurtosis with Application to Weather Data*. Communications in Statistics-Theory and Methods, 0(0): 1-18.

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

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, 2021.

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

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