ISSN 2581-3463

AJMS
Asian Journal of Mathematical Sciences

RESEARCH ARTICLE

Statistical Measures of Location: Mathematical Formulas versus Geometric Approach

T. Adeniran Adefemi[1], J. F. Ojo[2], K. O. Balogun[3]

[1]Department of Mathematical Sciences, Augustine University Ilara-Epe, Lagos State, Nigeria, [2]Department of Statistics, University of Ibadan, Ibadan, Oyo State, Nigeria, [3]Department of Mathematical Sciences, Federal School of Statistics, Ibadan, Oyo State, Nigeria

Received: 01-08-2018; Revised: 10-09-2018; Accepted: 01-01-2019

ABSTRACT

This paper illustrates with an example comparison of the geometrical and the numerical approaches of measures of location. A geometrical derivation of the most popular measure of location (mean) was derived from histogram by determining its (histogram) centroid. The numerical or mathematical expression of the other measures of location, median and mode were derived from ogive and histogram, respectively. Finally, the research establishes that the two approaches produce the same results.

Key words: Measures of location, geometrical, numerical, histogram, ogive, centroid

INTRODUCTION

Data with large observations, depending on the nature and depth of the inquiry, are often generated in all areas of human endeavor such as business, sports, academic institutions, research institutions, and internet services. Whatever be their size (large, medium, or small), it is impossible to grasp or retrieve information by mere looking at all the observations. It is advisable to get a summary of the dataset, if possible with a single number, provided that this single number is a good representative one for all the observations. Representative in the sense that the single number summarizes with relatively high precision, the characteristics of interest in the entire observations. That is, the single number mirroring entire characteristics of the whole observations. Such representative number could be a central value for all the observations. This central value is called a measure of central tendency, also known as a measure of location. The value could be the mean (arithmetic mean, geometric mean, harmonic mean, weighted mean, etc.,), the median or the mode of distribution. In a nutshell, measures of central tendency is the study of dataset cluster around the central value popularly called average.[1]

There are two basic methods of computing any of these measures of location[2-4]: Graphical and mathematical formula. Opinions of many instructors of introductory statistics courses are: (1) Mathematical formula approach is more precise and exact than geometrical approach in deducing measures of location[5] and (2) mean as a measure of location cannot be graphically determined[6] and the likes.

Contrary to these opinions, this study showed that the mathematical formulas of all the considered measures of location (mean, median, and mode) were actually derived geometrically from histogram and ogive. In addition, this study used the work of Beri (2012) to show that mean of a distribution/dataset lies at the centroid of the histogram drawn from such a distribution. These are the points that this study is set to make and clarify to enhance peoples understanding of the two approaches.

The graphical methods have been used in undergraduate level introductory statistics classes at Augustine University, Ilara-Epe. Feedback from students concerning this approach has been positive, and the students often appreciate that mathematical formulas and concepts were

Address for correspondence:
T. Adeniran Adefemi,
E-mail: adefemi.adeniran@augustineuniversity.edu.ng

translated and illustrated in a more visual form. Furthermore, graphical methods (histogram, ogive, frequency polygon, etc.) are sometimes better suited than numerical formulas because they contain detailed information about the pattern or shape in the data. Although, our interest is not to prioritize one method over the other one but to elucidate on both. In fact, numerical and graphical approaches complement each other; it is wise to use both. Finally, the crux of this research is that abstract mathematical formula for measures of location is graphically made known and applicable to all students and users of statistics regardless of their mathematical background.

METHODOLOGY

Definition of measures of location

The mean of the set $\{x_1,x_2,\ldots,x_n\}$ of numbers is the quantity

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \quad (1)$$

It is also called the arithmetic mean or the average of the set of numbers. The mean value of a distribution lies at the point where the histogram drawn from such distribution would balance called centroid of the histogram (Bird, 2004). The centroid of a composite figure X can be computed by dividing it into a finite number of simpler figures $X_1, X_2, ..., X_n$, computing the centroid c_i and area a_i (i=1,2,…,n) of each part, and then computing $C_x = \frac{\sum_{i=1}^{n} c_{ix} a_i}{\sum_{i=1}^{n} a_i}$, $C_y = \frac{\sum_{i=1}^{n} c_{iy} a_i}{\sum_{i=1}^{n} a_i}$ which is the centroid of the composite figure X.[7]

Median: Suppose that the numbers in the set $\{x_1,x_2,\ldots,x_n\}$ are arranged so that $x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n$. The median of the set is the number

$$m = \begin{cases} x_{\frac{(n+1)}{2}} & \text{if n is odd} \\ \frac{x_{\frac{(n)}{2}} + x_{\frac{(n)}{2}+1}}{2} & \text{if n is even} \end{cases} \quad (2)$$

In other words, the median of a set of n numbers is the number that is in the middle of the arrangement $x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n$, if there is a single in the middle. Otherwise, it is the average of the two numbers that are in the middle of the arrangement. Geometrically, median is the abscissa of the point ordinate which divides the histogram into two equal parts. That is, the point at which the perpendicular line that divides the total area of histogram into two equal halves meets with the X-axis (upper class boundary) gives the median.

Mode: A number x is called a mode of the set if x occurs at least as frequently as any other numbers in the set. That is, the mode of a set of data is the value that occurs most frequently among the values of the variable. If a histogram has been drawn for a grouped data, the mode of the distribution exists in the tallest bar of the histogram. Figure 1 illustrates a portion of a histogram with *MNLU* be the tallest bar (modal class) of the histogram. By joining *MQ* and *NP*as shown in the diagram, the abscissa $\hat{x}_m$ which corresponds to the perpendicular drawn from the point of intersection S is the mode of the distribution.

Graphical computation of mean from histogram

Theorem 2.1 given a grouped frequency distribution table containing class boundaries and their frequencies as shown in Table 1, then the mean ($\bar{x}$) of a grouped frequency distribution [Tables 2 and 3] with interval can be computed by

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i c_i}{\sum_{i=1}^{n} f_i}$$

where,

c_i = is the centre or midpoint of ith interval

f_i = number of times x_i occurs.

Proof: Figure 2 shows a typical histogram with further construction that elicits procedure for the proof.

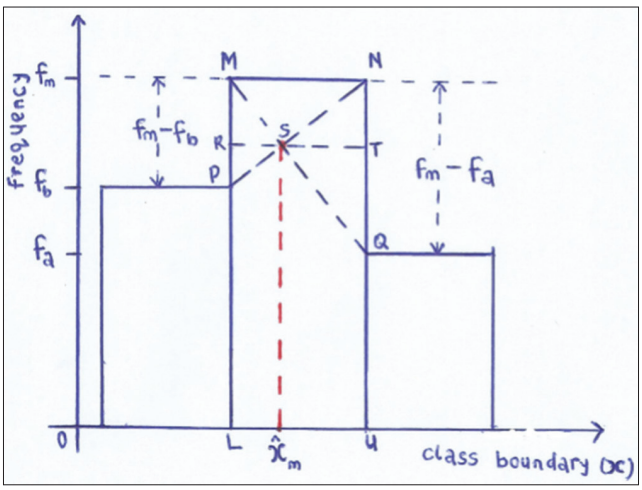

Figure 1: Derivation of mode formula

Having drawn a histogram of a distribution, estimate of the individual area (a_i i = 1, 2,…, n) of each bar (rectangle) and the total area (A) of the histogram which is obtained by adding the individual areas a_i are

$$a_i = f_i k_i \left(1,2,\ldots,n\right) \tag{3}$$

and

$$A = \sum_{i=1}^{n} a_i \tag{4}$$

respectively. Position of the horizontal value of the centroid can be obtained from the relation $AC = \sum_{i=1}^{n} a_i c_i$, where c_i are the distance of the midpoint of the individual rectangles from arbitrary axis YY'. Hence,

$$C = \frac{\sum_{i=1}^{n} a_i c_i}{A} \tag{5}$$

Putting equations (4) and (3) in that order into equation (5), we have

$$C = \frac{\sum_{i=1}^{n} f_i k_i c_i}{\sum_{i=1}^{n} f_i k_i} \tag{6}$$

In a frequency distribution with equal class interval, that is $k_i = k \forall i$=1,2,…,n, equation (6) yields

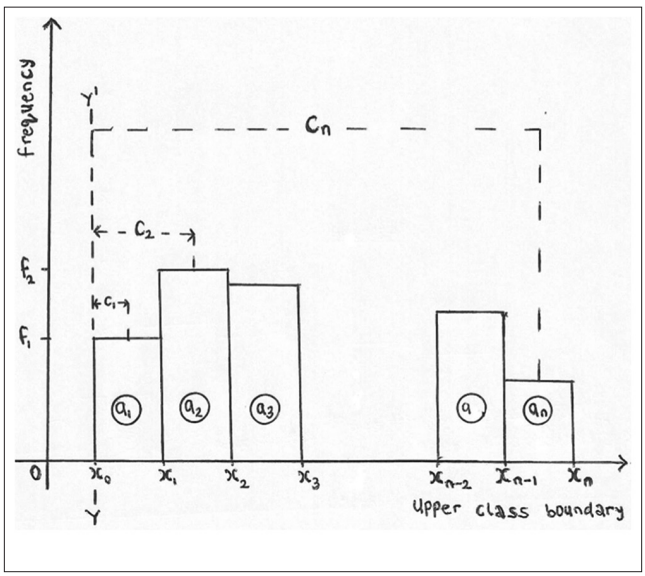

Figure 2: Derivation of the mean formula

Table 1: Typical example of a grouped frequency distribution

Class boundaries	x_0–x_1	x_1–x_2	…	x_{n-1}–x_n
Frequency	f_1	f_2	…	f_k

$$C = \frac{\sum_{i=1}^{n} f_i c_i}{\sum_{i=1}^{n} f_i} \tag{7}$$

Graphical computation of median from ogive and histogram

If the grouped data are given as a cumulative frequency distribution, the median is the abscissa of the point on the ogive, the ordinate of which equals half the total frequency.[8] This can be achieved by any of these two methods:

a. First method: Draw only less than cumulative frequency curve and determine the position of the median value by the formula: $\frac{N}{2}th$. Locate this value on the cumulative frequency axis (i.e., Y-axis) and from it draw a perpendicular (straight line) to meet the cumulative frequency curve. From this point, draw another perpendicular on the X-axis and the point where it meets the X-axis is the median.
b. Second method: Draw and superimpose "less than" and "more than" cumulative frequency curves. From the point of intersection of the two curves, draw a perpendicular to the X-axis. The point where this perpendicular touches the X-axis, gives the required value of median.

Theorem 2.2 given a grouped frequency distribution table containing class boundaries and their frequencies as shown in Table 1, then the formula for computing median M of a grouped frequency distribution with interval is

$$M = median = L_m + \left[\frac{\frac{N}{2} - F_b}{f_m}\right] c$$

where,
L_m = lower class boundary of the median class,
N = number of items (sum of all frequency),
F_b = cumulative frequency before the median class,
f_m = frequency of the median class, and
c = class size width) of the median class.

Proof. Let the cumulative frequency of *ith* class be denoted as F_i, therefore $F_1=f_1$, $F_2=f_1+f_2$, $F_k = f_1+f_2+f_3+\ldots+f_{k-1}+f_k$ and $F_n=N$. Suppose that

Table 2: Frequency distribution of ages

Ages	**20–24**	**25–29**	**30–34**	**35–39**	**40–44**	**45–49**	**50–54**	**55–59**
Frequency	5	9	14	13	30	12	11	6

Table 3: Table of numerical computation

Class interval	Class boundaries	f_i	<CF	>CF	c_i	f_ic_i
20–24	19.5–24.5	5	5	100	22	110
25–29	24.5–29.5	9	14	95	27	243
30–34	29.5–34.5	14	28	86	32	448
35–39	34.5–39.5	13	41	72	37	481
40–44	39.5–44.5	30	71	59	42	1260
45–49	44.5–49.5	12	83	29	47	564
50–54	49.5–54.5	11	94	17	52	572
55–59	54.5–59.5	6	100	6	57	342
Total		100				4020

the median (M) of a distribution lie in the *kth* class, that is, the class interval x_{k-1}–x_k and the cumulative frequencies at x_{k-1} and x_k are F_{k-1} and F_k, respectively. This implies that $F_{k-1} < \frac{N}{2} < F_k$ and consequently, $x_{k-1}<M<x_k$.

Figure 3 shows a typical ogive with additional construction to depict the required procedure for the proof.

The increment in cumulative frequency between x_{k-1} and M is $\frac{N}{2} - F_{k-1}$ and between M and x_k is $F_k - \frac{N}{2}$. Assuming the frequencies are uniformly distributed in each interval. Then, ΔPQS and ΔPRT are similar. Thus, $\frac{QS}{PS} = \frac{RT}{PT}$ which implies that

$$\frac{\frac{N}{2} - F_{k-1}}{M - x_{k-1}} = \frac{F_k - F_{k-1}}{x_k - x_{k-1}} \tag{8}$$

It is significant to note the following:

F_k–F_{k-1}=f_k or f_m = frequency of the median class

x_k–x_{k-1}=c = the class size/width

x_{k-1}=L = the lower class boundary of the median class

$F_{k-1} = \sum_{i=1}^{k-1} f_i = F_b$ = the cumulative frequency before the median class

Substituting these quantities in the equation (8), we have

$$\frac{\frac{N}{2} - F_b}{M - L} = \frac{f_m}{c} \tag{9}$$

From equation (9), making M subject of the formula yields

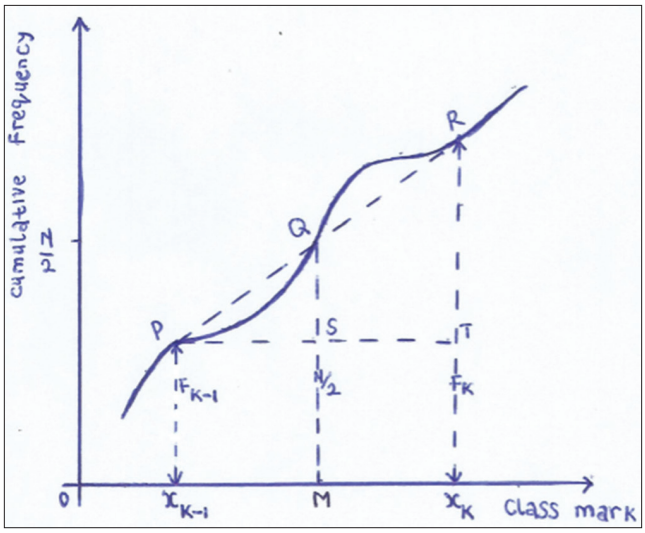

Figure 3: Derivation of median formula from cumulative frequency curve

$$M = median = L + \left[\frac{\frac{N}{2} - F_b}{f_m}\right]c \tag{10}$$

Graphical computation of mode from histogram [Figure 1]

Theorem 2.3 given a grouped frequency distribution table containing class boundaries and their frequencies as shown in Table 1, then the mode is given by

$$\hat{x}_m = mode = L + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right]c$$

where,

$\hat{x}_m$ = mode of grouped frequency distribution with interval,

L = lower boundary of the class with the highest frequency (modal class),

$\Delta_1 = (f_m - f_b)$ = the difference between frequency of the modal class and frequency of the pre-modal class,

Δ_2 $(f_m - f_a)$ = the difference between frequency of the modal class and frequency of the post-modal class,

c = width of the modal class interval.

Proof: Consider the diagram below

By similar triangles, ΔSPM and ΔSQN are similar, therefore

$$\frac{SR}{MP} = \frac{ST}{NQ} \tag{11}$$

From Figure 1, $SR = \hat{x}_m - L$, $MP = f_m - f_b = \Delta_1$, $ST = U - \hat{x}_m$ and $NQ = f_m - f_a = \Delta_2$. Substituting *SR*, *MP*, *ST* and *NQ* into equation (11) and simplify, we have

$$\hat{x}_m(\Delta_1 + \Delta_2) = \Delta_1 U + \Delta_2 L \quad (12)$$

Upper class boundary U can be expressed as addition of lower class boundary (L) and common class interval (c). That is,

$$U = L + c \quad (13)$$

Substituting (13) in (12) and making $\hat{x}_m$ subject of the formula. The procedure follows thus;

$$\hat{x}_m(\Delta_1 + \Delta_2) = \Delta_{1(L+c)} + \Delta_{2L}$$

$$\hat{x}_m(\Delta_1 + \Delta_2) = (\Delta_1 + \Delta_2)L + \Delta_{1c}$$

Therefore,

$$\hat{x}_m = mode = L + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right]c \quad (14)$$

It is now very crystal clear that formula for mean, median, and mode as a statistical measure of location are by-product of geometrical (graphical) approach. Hence, both methods are expected to be equivalent and should, therefore, yield the same result. Any difference in the results is due to the precision of computing device in the formula method or the precision in reading from the graph (histogram or ogive).

RESULTS AND DISCUSSION

Illustration: A sample of 100 individuals are randomly selected in Ilara-Epe for participation in a study of cardiovascular risk factors. The following data represent the ages of enrolled individuals, measured in years.

Computation of mean from histogram

With reference to Figure 4, arbitrary axis YY' is chosen at score 19.5, the areas (in square units) of the individual rectangle are shown circled on the histogram. The position of the horizontal

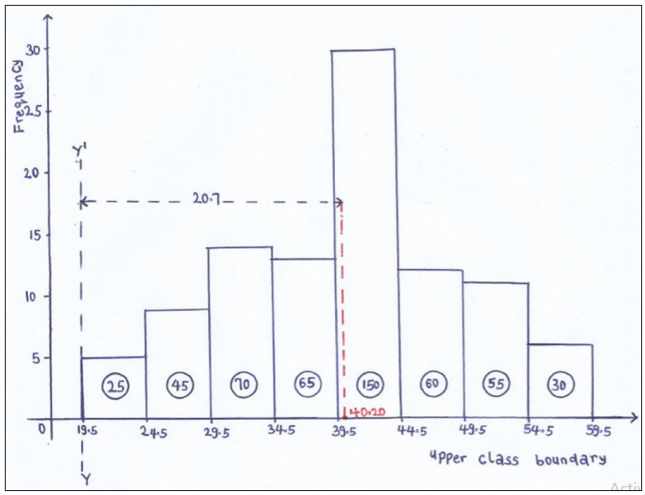

Figure 4: Graphical method of estimating mean

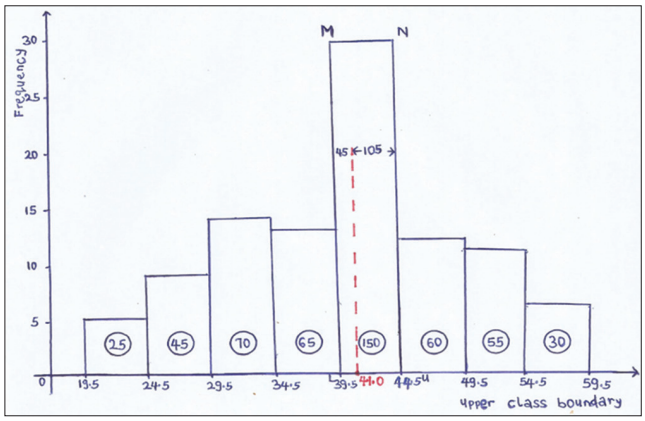

Figure 5: Determination of median from histogram

value of the centroid (C) can be obtained from the relation $AC = \sum_{i=1}^{n} a_i c_i$.

$500C = 25(2.5)+45(7.5)+70(12.5)+65(17.5)+150(22.5)+60(27.5)+55(32.5)+30(37.5)$

$C=20.7$

Thus, the position of the mean with reference to the score scale is x-value corresponding to the C (centroid) distance from the arbitrary axis YY'. Therefore, the mean age is 19.5 + 20.7 = 40.2

Computation of median from histogram

Figure 5 indicates that the total area of the histogram is 25+45+70+65+150+60+55+30 = 500. To draw a vertical line that will give 250 units of the area on each side, rectangle *MNLU* must be split so that 250 − (25+45+70+65) units of the area lie on one side and 250 − (60+55+30) units of the area lie on the other.[9] This implies that the area of *MNLU* is split so that 45 units of area lie to the left of the line and 105 units of the area lie to the right. Hence, the vertical line must pass through 41.0

scores. Thus, the median value of the distribution is 41.0.

Computation of median from ogive [Figures 6 and 7]

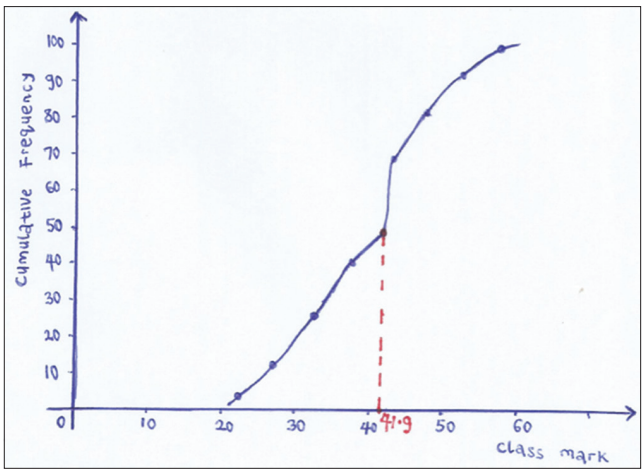

Figure 6: Graphical computation of median from less than ogive

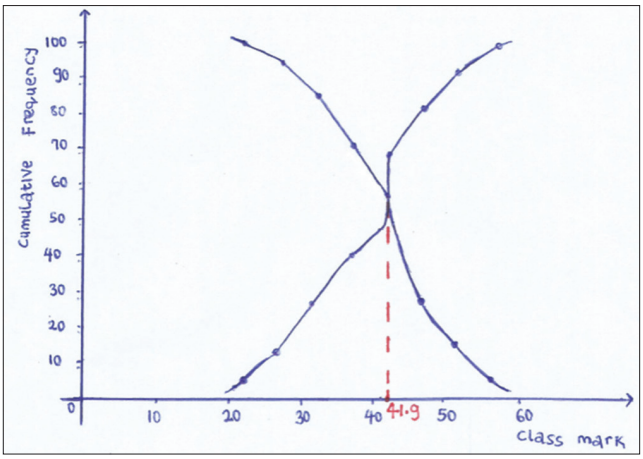

Figure 7: Graphical computation of median from superimposition of less than and more than ogive

Computation of mode from histogram [Figure 8]

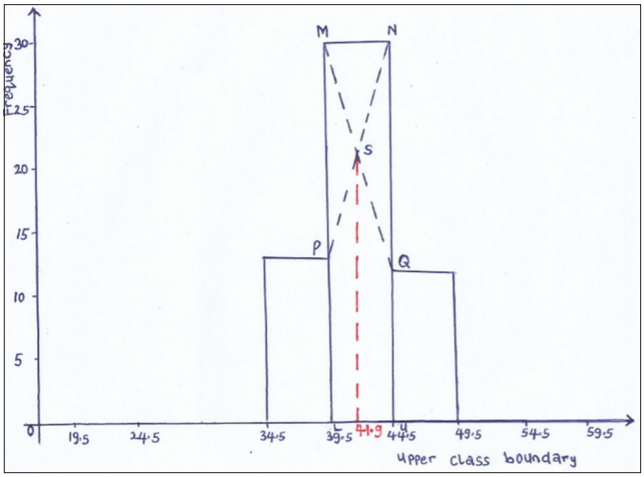

Figure 8: Determination of mode from histogram

Computation of mean, median, and mode using formula approach

$$\bar{x} = Mean = \frac{\sum_{i=1}^{n} f_i c_i}{\sum_{i=1}^{n} f_i} = \frac{4020}{100} = 40.2$$

$$M = Median = L_m + \left[\frac{\frac{N}{2} - F_b}{f_m}\right] c$$

$$= 39.5 + \left[\frac{\frac{100}{2} - 41}{30}\right] \times 5$$

$$= 41.0$$

$$\hat{x} = Mode = L + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] c = 39.5 + \left(\frac{17}{17+18}\right) \times 5$$

$$= 41.92857$$

CONCLUSION

This paper established that mean as a measure of location can be graphically determined, the formula for measures of location (mean, median, and mode) was derived from graphs. Therefore, if all the necessary precautions for drawing graph were put into consideration, both the formula and graphical methods produce the same result. Hence, any observed difference or discrepancy between results from the two methods is either due to human lack of proper pattern recognition in reading from the graph (human error) and/or instrumental error (inappropriate handling of formula).

ACKNOWLEDGMENTS

Authors would like to thank for support and guide to all the superiors. All the information is true to my knowledge.

REFERENCES

1. Beri GC. Statistics for Management. New Delhi: Tata McGraw Hills; 2012.
2. Afonja B, Olubusoye OE, Osai E. Introductory Statistics: A Learner's, Motivated Approach. Nigeria: Evans Brother Nigeria Publishers Limited; 2015.
3. Gupta SP. Statistical Methods, Sultan Chand and Sons.

36th ed. New Delhi: Educational Publishers; 2008.
4. Peggy TS. A First Course in Probability and Statistics with Applications. 2nd ed. Washington, DC: Harcourt Brace Jovanovich (HBJ) Publishers; 1989.
5. Jaggi S. Descriptive Statistics and Exploratory Data Analysis New Delhi: Indian Agricultural Statistics Research Institute, Library Avenue; 2012.
6. Adebowale SA. Statistics for Engineers, Managers and Scientists. New Delhi: Alfredo Graphics Limited; 2006.
7. Egbe E, Odili GA, Ugbebor OO. Further Mathematics. Princeton, NJ: Africana-First Publishers Limited; 2003.
8. Varalakshmi V, Suseela N, Gnana GS, Ezhilarasi S. Statistics for Higher Secondary School. Tamil Nadu: Textbook Corporation; 2005.
9. Bird J. Engineering Mathematics. 5th ed. Jordan Hill, Oxford OX28DP, UK: 2004.