Academic Journal of Applied Mathematical Sciences
ISSN(e): 2415-2188, ISSN(p): 2415-5225
Vol. 6, Issue. 7, pp: 145-152, 2020
URL: https://arpgweb.com/journal/journal/17
DOI: https://doi.org/10.32861/ajams.67.145.152

Academic Research Publishing Group

Original Research **Open Access**

Statistical Measures of Location: Mathematical Formula Versus Geometric Approach

ADENIRAN Adefemi Tajudeen (Corresponding Author)
Department of Statistics, Faculty of Science, University of Ibadan, Ibadan, Oyo State, Nigeria
Email: at.adeniran@mail.ui.edu.ng

OJO Johnson Funminiyi
Department of Statistics, Faculty of Science, University of Ibadan, Ibadan, Oyo State, Nigeria

FAWEYA Olanrewaju
Department of Statistics, Faculty of Science, Ekiti State University, Ado-Ekiti, Ekiti State, Nigeria

BALOGUN Kayode
Department of Statistics, Federal Scool of Statistics Ibadan, Oyo State, Nigeria

Article History
Received: June 3, 2020
Revised: July 15, 2020
Accepted: July 21, 2020
Published: July 24, 2020

Abstract

Graphical method and mathematical formula are the two approaches for estimating measures of location. Understanding of many instructors of introductory statistics classes are: mean cannot be graphically determined and numerical (formula) approach is more precise than geometrical technique. Contrary to their understanding, this study estimate mean of a dataset geometrically (from histogram) by determining the centroid of histogram drawn from such data set. In addition, we also make known that mathematical formulas for mean, median and mode were derived geometrically (either from ogive or histogram). Finally, the research illustrated the two techniques with a survey data and established that the two approaches produce same results.

Keywords: Measures of location; Geometrical; Histogram; Centroid; Ogive.

1. 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, internet services etc [1-4]. Whatever be their size (large, medium, or small), it is impossible (especially, when the size is large) 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 the single number is a good representative one for all the observations. Representative in the sense that, the single number wholly summarizes or mirrors with relatively high precision, the characteristics of interest in the entire observations. Such representative number could be a central value for all the observations. Descriptions of the center of a distribution are called measures of central tendency, sometimes called measures of location or averages [4-6]. 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 [7, 8].

According to Afonja, *et al.* [9]; McClave and Sincich [8], and Utts and Heckard [10], there are two basic methods of computing any of these measures of location: Graphical method and mathematical formula approach. Opinions of many instructors of introductory statistics courses are:

(i) Formula approach is more precise and exact than geometrical approach [1, 9, 11], and
(ii) mean as a measure of location cannot be graphically deduced [1, 9].

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 either from histogram or ogive. In addition, this study used the work of Bird [12]; Ramachandran and Tsokos [13] to show that mean of a distribution or dataset lies at the centroid of the histogram drawn from such a distribution. These are the points that this study targets to establish and enhance learners understanding of the two approaches.

The graphical methods have been used in first-year undergraduate level introductory statistics classes at University of Ibadan, Oyo State, Nigeria and the feedback from students concerning this approach has been positive. In addition, students often appreciate that mathematical formulas were illustrated in a visual form. Therefore, the crux of this work is that abstract mathematical formula for measures of location is graphically translated and explained to all students and users of statistics regardless of their mathematical background. Furthermore, graphical methods (histogram, ogive, frequency polygon, etc.) are sometimes better suited than numerical formulas because they describe measurements in a form that affords the viewer a quick impression of the data distribution (pattern or shape). However, graphical methods are not always convenient or desirable. For example, suppose you take a mathematics exam and then encounter your instructor the next day in the bookstore. If you inquire how the class performed, you cannot expect the professor to produce a frequency distribution or histogram on the spot. Rather, you

can probably expect a numerical value that describes the middle of the grade distribution. Therefore, it is very essential to know that the objective of this study is not to prioritize one method over the other but to elucidate on both. In fact, formula and graphical approaches complement each other; it is wise to use both.

2. Methodology

2.1. Graphical Computation of Mode from Histogram

Definition 1: *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. Figure1 below illustrates a portion of a histogram with $LMNU$ 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.

Theorem 1: *Given a grouped frequency distribution table containing class boundaries and their frequencies as shown in the Table (1) below,*

Table-1. Typical example of a grouped frequency distribution

Class boundaries	$x_o - x_1$	$x_1 - x_2$	⋯	$x_{(n-1)} - x_n$
Frequency	f_1	f_2	⋯	f_n

then the mode is given by

$$\hat{x}_m = L + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] c \qquad 1$$

where,

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

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

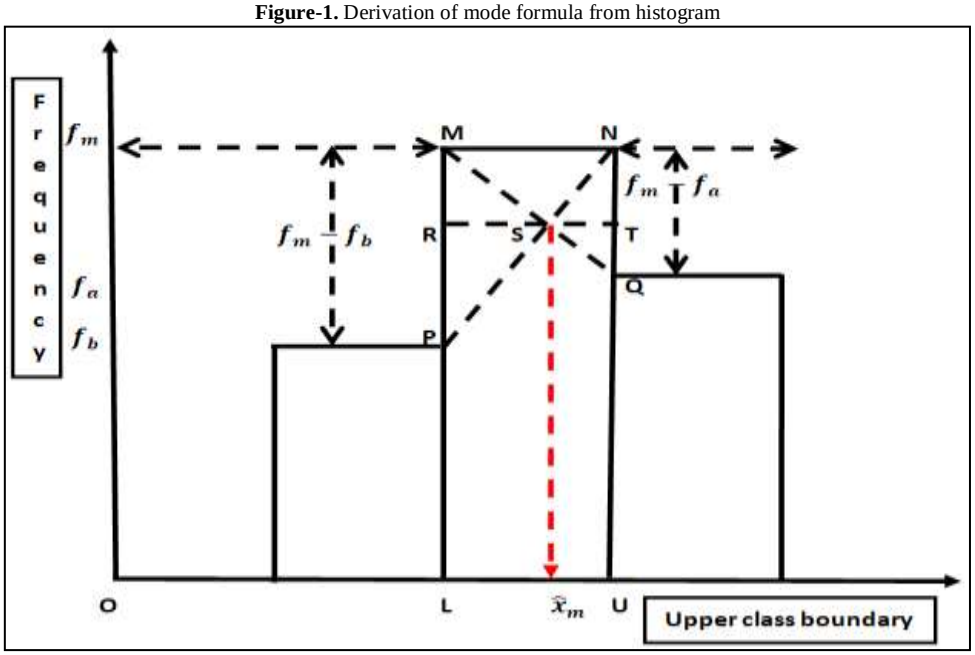
Δ_1 *= the difference between frequency of the modal class and frequency of the pre-modal class* $(f_m - f_b)$

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

$c =$ *width of the modal class interval.*

proof: Consider the diagram below

Figure-1. Derivation of mode formula from histogram



By similar triangles, $\Delta\, SPM$ and $\Delta\, SQN$ are similar, therefore

$$\frac{SR}{MP} = \frac{ST}{NQ} \qquad 2$$

From Figure1 above, $SR = \hat{x}_m - L$, $MP = f_m - f_b = \Delta\_1$, $ST = U - \hat{x}_m$ and $NQ = f_m - f_a = \Delta_2$

Substituting SR, MP, ST and NQ in equation 2 and simplify, we have

$$\hat{x}_m(\Delta_1 + \Delta_2) = \Delta\_1 U + \Delta_2 L \qquad 3$$

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

$$U = L + c \qquad 4$$

Substituting 4 in 3 and simplify to get

$$\hat{x}_m(\Delta_1 + \Delta_2) = L(\Delta_1 + \Delta_2) + \Delta_1 c \qquad 5$$

Making $\hat{x}_m$ subject of the formula from 5, we have the computational formula for mode as

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

The theorem proved.

2.2. Graphical Computation of Median from Ogive and Histogram

Definition 2: *Suppose that numbers in the set* $\{x_1, x_2, \cdots, x_n\}$ *are arranged so that* $x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n$. *The median of the set is the number*

$$M = \begin{cases} x_{\frac{n+1}{2}}, & if\ n\ is\ odd \\ x_{\frac{n}{2}+1} + x_{\frac{n}{2}}, & if\ n\ is\ even \end{cases} \qquad 6$$

In other words, the median of a set of n *numbers is the number that is in the middle of the arrangement* $x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n$, *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.*

If the grouped data is 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 [6, 7, 11]. 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 require value of median.

Theorem 2: *Given a grouped frequency distribution table containing class boundaries and their frequencies as shown in Table 1, the formula for computing median* (M) *of a grouped frequency distribution with interval is*

$$M = L_m + \left[\frac{\frac{N}{2} - F_b}{f_m}\right]c \qquad 7$$

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*

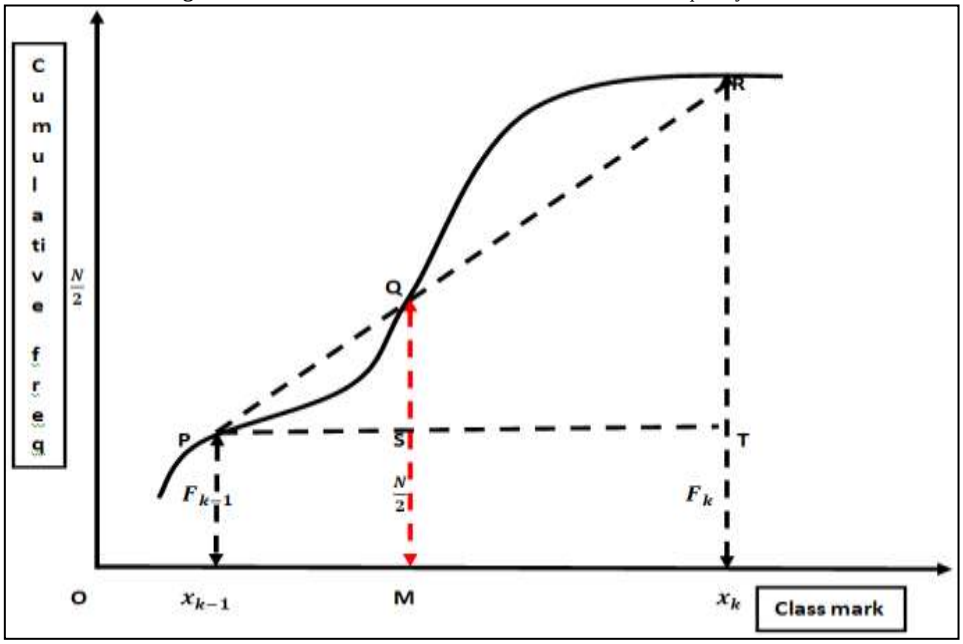
c = *class size (width) of the median class.*

Proof: Let the cumulative frequency of ith class be denoted as F_i, therefore

$F_1 = f_1$, $F_2 = f_1 + f_2$, $F_k = f_1 + f_2 + \cdots + f_{k-1} + f_k$ and $F_n = f_1 + f_2 + \cdots + f_{k-1} + f_k + f_{k+1} + \cdots + f_{n-1} + f_n = N$. Suppose that 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 $x_{k-1} < M < x_k$.

The Figure 2 below is a typical ogive with additional construction to depict the required procedure for the proof.

Figure-2. Derivation of median formula from cumulative frequency curve



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

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

It is significant to note the following:

$F_k - F_{k-1} = f_m =$ 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 equation 8, we have

$$\frac{\frac{N}{2} - F_b}{M - L} = \frac{f_m}{c} \quad (9)$$

From equation 9, making M subject of the formula yields numerical formula for median as

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

The theorem confirmed.

Alternatively, 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.

2.3. Graphical Computation of Mean from Histogram

Definition 3: *The mean of the set $\{x_1, x_2, \cdots, x_n\}$ of numbers is the quantity*

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x\_i \quad (10)$$

It is also called the arithmetic mean or the average of the set of numbers.

Geometrically, the mean value of a distribution lies at the point where the histogram drawn from such distribution would balance called centroid of the histogram [12]. The centroid of a composite figure X can be computed by dividing it into a finite number of simpler figures $X_1, X_2, \cdots, X_n$, computing the centroid c_i and area $a_i (i = 1,2, \cdots, n)$ of each part, and then computing $C_x = \frac{\sum_{I=1}^{n} c_{ix} a_i}{\sum_{i=1}^{n} a_i}$ which is the centroid of the composite figure X [14].

Theorem 3: *Given a grouped frequency distribution table containing class boundaries and their frequencies as shown in Table 1, the mean $(\bar{x})$ of a grouped frequency distribution with interval can be computed by*

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i c_i}{\sum_{i=1}^{n} f_i} \quad (11)$$

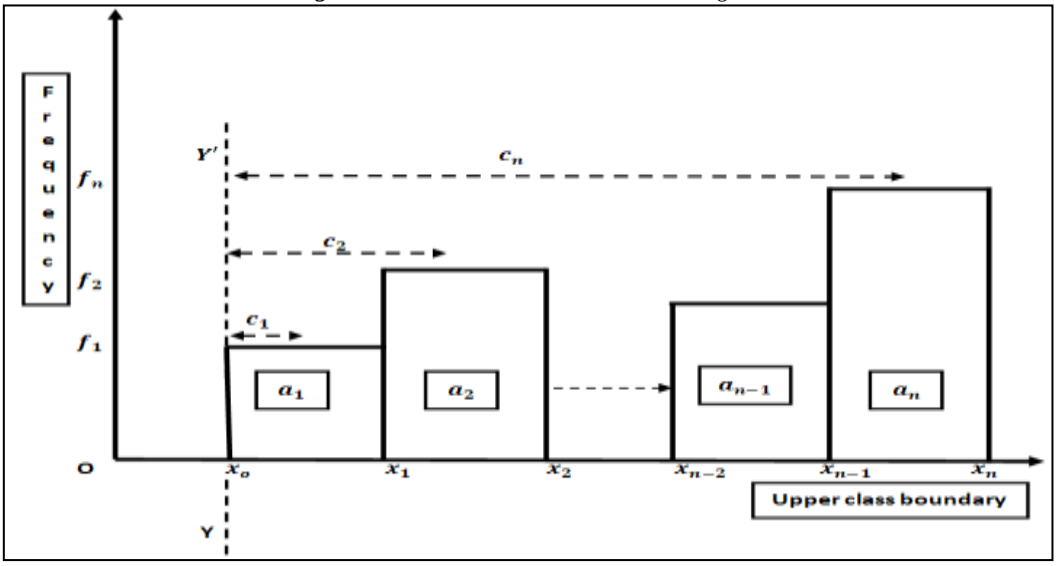
where,

$c_i =$ *is the centre or midpoint of ith interval*

$f_i =$ *number of times x_i occurs.*

proof: Figure 3 is a typical histogram with further construction that elicit procedure for the proof.

Figure-3. Derivation of mean formula from histogram



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

$$a_i = f_i k_i \qquad (i = 1,2,\cdots,n) \qquad 12$$

and

$$A = \sum_{i=1}^{n} a_i, \qquad 13$$

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,

$$C = \frac{\sum_{i=1}^{n} a_i c_i}{A} \qquad 14$$

Putting (13) and (12) in that order in equation (14), we have

$$C = \frac{\sum_{i=1}^{n} f_i k_i c_i}{\sum_{i=1}^{n} f_i k_i} \qquad 15$$

Letting $k_i = k \; \forall \; i = 1,2,\cdots,n$. That is, a frequency distribution with equal class interval. Equation 15 gives formula for mean as

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

This is the required formula for mean.

Based on the results of Theorem 1-3, it is now crystal clear that formula for mean, median and mode as a statistical measure of location are by-product of geometrical (graphical) technique. Hence, there is no sound justification to the claim that formula method is more exact than the graphical technique. As a result, both methods are expected to be equivalent or produce the same result. Any difference in the results is either due to the level of precision of computing device or reading from the graph (histogram or ogive).

3. Results

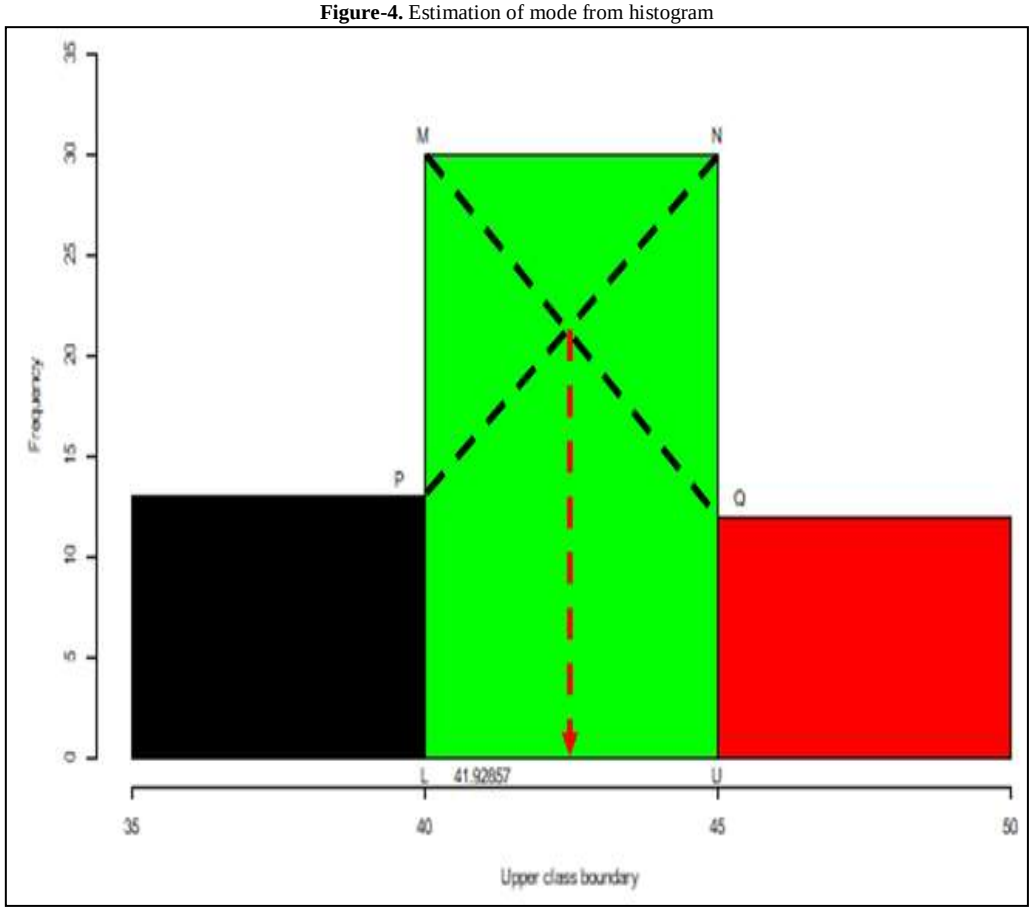
Illustration: A sample of 100 individuals was randomly selected in a city for participation in a study of cardiovascular risk factors. The following data represent the ages of enrolled individuals, measured in years.

Table-2. Frequency distribution of ages

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

3.1. Computation of Mode from Histogram

Figure-4. Estimation of mode from histogram



3.2. Computation of Median from Ogive and Histogram

Figure-5. Graphical computation of median from less than Ogive

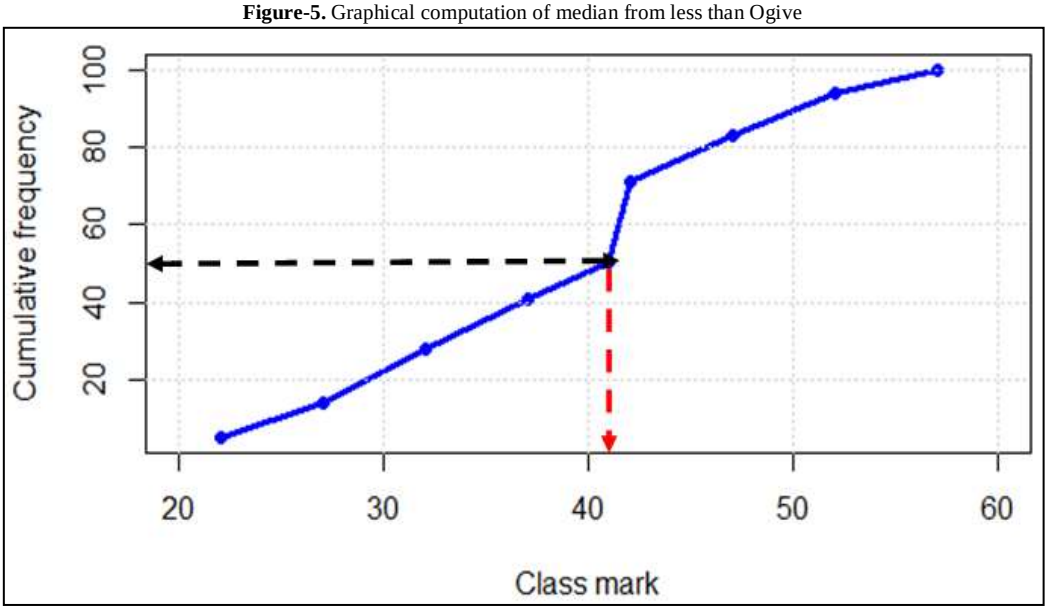


Figure-6. Graphical computation of median from superimposition of less than and more than ogive Figures 5 and 6 above show that the median age is 41.0

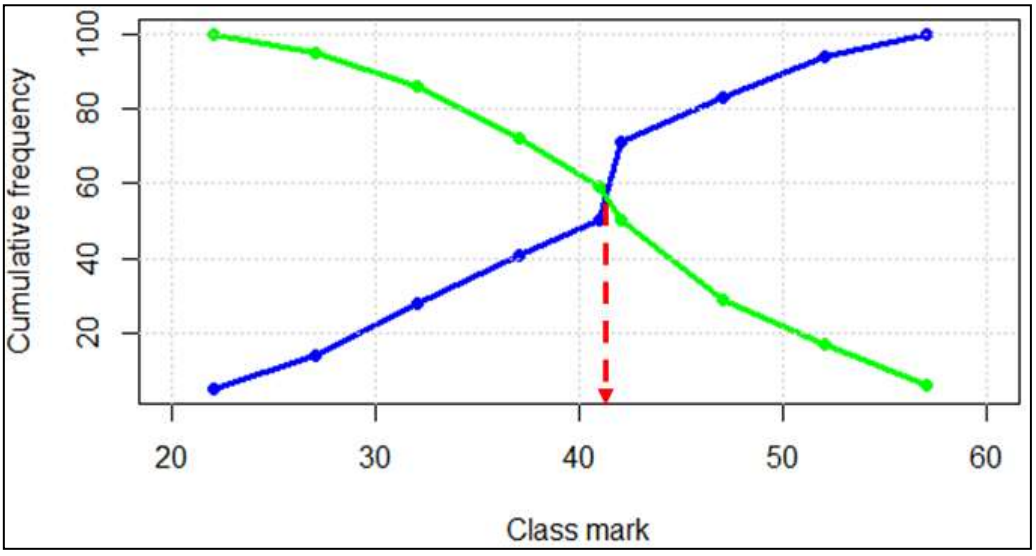


Figure-7. Determination of median from Histogram

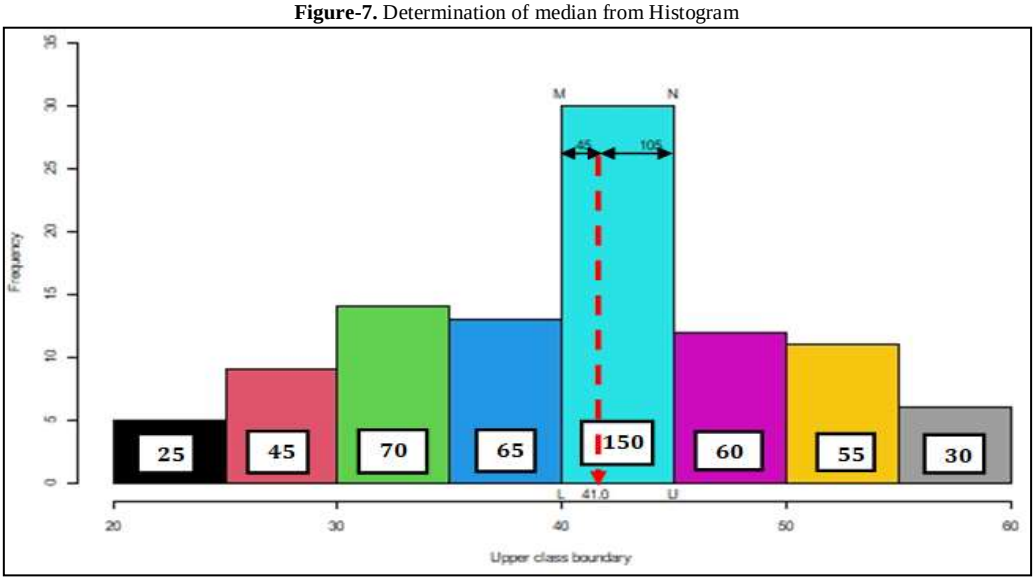
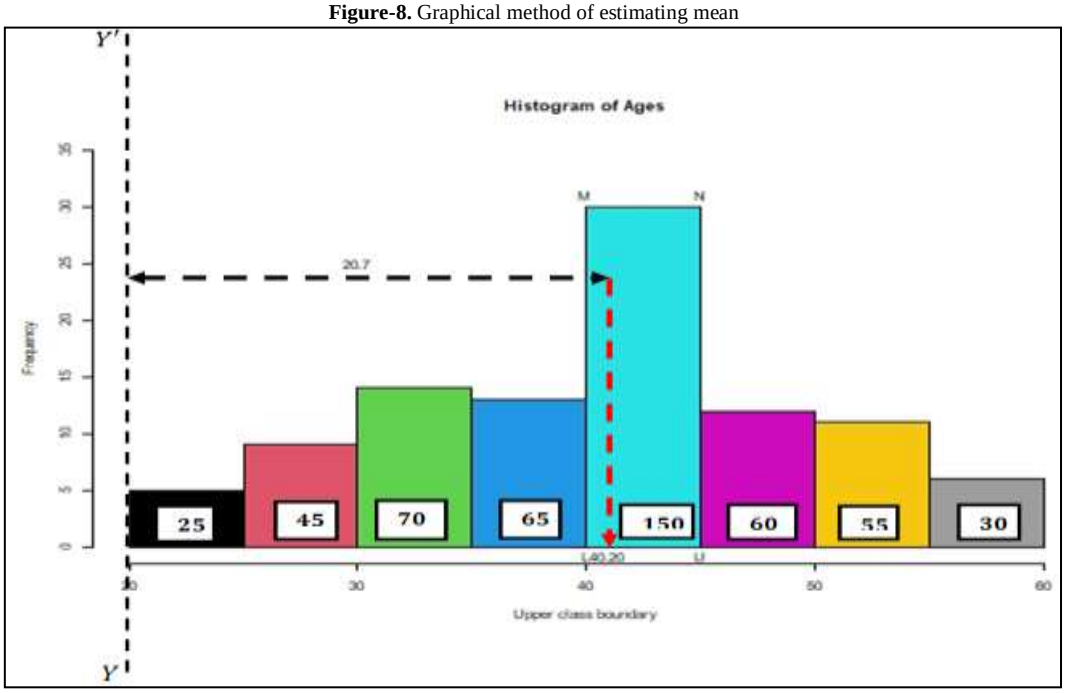


Figure 7 illustrates procedure for computing median from histogram. It indicates that the total area of the histogram is $25 + 45 + 70 + 65 + 150 + 60 + 55 + 30 = 500$. To draw a vertical line that will give 250 units of

area on each side, rectangle $MNLU$ must be split so that $250-(25+45+70+65)$ units of area lie on one side and $250-(60+55+30)$ units of area lie on the other. 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 area lie to the right. Hence, the vertical line must pass through 41.0 value. Thus, the median age of the distribution is 41.0.

3.3. Computation of Mean from Histogram

Figure-8. Graphical method of estimating mean



With reference to figure 8 above, arbitrary axis YY' is chosen at value19.5 the areas (in square units) of the individual rectangle are shown circled on the histogram. The position of the horizontal value of the centroid (C) can be obtained from the relation$AC=\sum_{i=1}^{n}a_ic_i$. That is,

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

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

3.4. Computation of Mean Median and Mode Using Formula Approach

Table-3. Numerical computation technique

Class interval	Class boundary	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

From Table 3, we make the following computation:

$$\hat{x}_m=L+\left[\frac{\Delta_1}{\Delta_1+\Delta_2}\right]c=39.5+\left[\frac{30-13}{(30-13)+(30-12)}\right]5=41.92857$$

$$M=median=L+\left[\frac{\frac{N}{2}-F_b}{f_m}\right]c=39.5+\left[\frac{\frac{100}{2}-41}{30}\right]5=41.0$$

$$\bar{x}=mean=\frac{\sum_{i=1}^{n}f_ic_i}{\sum_{i=1}^{n}f_i}=\frac{4020}{100}=40.20$$

Estimate from numerical computation above is equal to their geometrical counterparts.

4. 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 methods should produce same result. Hence, any observed difference or discrepancy between results from the two techniques 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). In addition, the authors observed that some very good old textbooks on descriptive statistics had virtually disappeared from the bookshops and libraries. The result was its abuses on the field among students, lecturers and professionals who need to present reports in acceptable and effective tabular, graphical or numerical summary statistics. Thus, this study fills the gaps being created by the lack of appropriate textbooks that gives comprehensive rudiment of descriptive measures.

Acknowledgements

All the authors would like to thank the editor and referees for their careful reading, constructive comments and suggestions which greatly improved the article.

References

[1] Adebowale, S. A., 2006. *Statistics for engineers, managers and scientists*. 2nd ed. Alfredo Graphics Limited.

[2] Buglear, J., 2003. *Stats Means Business: A guide to business statistics*. Elsevier Butterworth-Heinemann.

[3] Seema, J., 2012. *Descriptive statistics and exploratory data analysis*. New Delh: Indian Agricultural Statistics Research Institute, Library Avenue.

[4] Triola, M. F., 2006. *Elementary statistics*. 10th ed.: Person Addison Wesley.

[5] Gupta, S. P., 2008. *Statistical Methods*. 36th ed. New Delhi: Sultan Chand and Sons Educational publishers.

[6] Weiss, N. A., 2012. *Introductory statistics*. 9th ed. Addison Wesley.

[7] Kothari, C. R., 2004. *Research methodology: Methods and techniques*. 2nd ed. New Age International(P) Limited Publishers.

[8] McClave, J. T. and Sincich, T., 2006. *Statistics*. 10th ed. Upper Saddle River, New Jersey Pearson Prentice Hall. p. 07458.

[9] Afonja, B., Olubusoye, O. E., Osai, E., and Arinola, J., 2014. *Introductory statistics: A learner's motivated approach*. Revised ed. Evans Brothers Nigeria Publishers Limited.

[10] Utts, J. M. and Heckard, R. F., 2015. *Mind on statistics*. 5th ed. Cengage Learning.

[11] Johnson, R. A. and Bhattacharyya, G. K., 2010. *Statistics: Principles and methods*. 6th ed. John Wiley and Sons, Inc.

[12] Bird, J., 2007. *Engineering Mathematics*. 5th ed. Elsevier Limited.

[13] Ramachandran, K. M. and Tsokos, C. P., 2009. *Mathematical statistics with applications*. Elsevier Academic Press.

[14] Egbe, E., Odili, G. A., and Ugbebor, O. O., 2003. *Further Mathematics*. Africana-First Publishers Limited.