MEASURES OF CENTRAL TENDENCY

Shahbaz Baig

Abstract: The average is a value which expresses the central idea of the observations. It is a single value used to represent the data it is a value somewhere in the centre, where most of the items of the series cluster. Such values are also called "Measures of Central tendency". Different averages are used in different situations to represent the data. Averages are central part of distribution and, therefore they are also called measures of central tendency. The mean is a good measure of central tendency for symmetrical (e.g. normal) distributions but can be misleading in skewed distributions since it is influenced by outliers.

Key words: Mean, Mode, Mediam, Average, Quantite, Centile

INTRODUCTION

An average is a single value intended to represent the distribution (data) as a whole.

It may be calculated for a sample or a population data.

The average is a value which expresses the central idea of the observations. It is a single value used to represent the data it is a value somewhere in the centre, where most of the items of the series cluster. Such values are also called "Measures of Central tendency". Different averages are used in different situations to represent the data[1].

A single expression representing the whole group, is selectd which may convey a fairly adequate idea about the whole group.

This single expression is known as average.

Averages are central part of distribution and , therefore they are also called measures of central tendency[2].

QUALITIES OF A GOOD AVEREGE

Following qualities are associated with a good average. We try to use that average which possesses all or most of these qualities.

I. It should be based on all the observations of the data.
II. It should be capable of further algebraic treatment.
III. It should be unaffected by extreme observations.
IV. It should be easy to calculate and simple to follow.
V. It should be least affected by fluctuations of sampling [1].

TYPES OF AVERAGES

The following averages are usually used

1. Arithmetic mean
2. Median, Quartiles and other

Article Citation: Baig S, Measures of central tendency. Indep Rev Jul-Sep 2017;19(7-9): 127-137.

Date received: 13/06/2017
Date Accepted: 30/06/2017

Correspondence Address:
Dr. Shahbaz Baig, MBBS, MPH
Assistant Professor of Community Medicine
Independent Medical College, Faisalabad.

1. Dr, Shahbaz Baig, MBBS, MPH
Assitant Professor Community Medicine,
Independent Medical College, Faisalabad.

Partitions values
3. Mode
4. Geometric Mean
5. Harmonic Mean[3]

ARITHMETIC MEAN (A.M)[4]
The average value of a dataset, i.e. the sum of all the data divided by the number of variables.

The arithmetic mean is commonly called the "average". When the word "mean" is used without a modifier, it usually refers to the arithmetic mean.

The mean is a good measure of central tendency for symmetrical (e.g. normal) distributions but can be misleading in skewed distributions since it is influenced by outliers. In general, the mean is larger than the median in positively skewed distributions and less than the median in negatively skewed distributions.

Therefore, other statistics such as the median may be more informative for distributions such as reaction time or family income that are frequently very skewed. The mean, median, and mode are equal in symmetrical frequency distributions. The mean is higher than the median in positively (right) skewed distributions and lower than the median in negatively (left) skewed distributions.

Formula for the arithmetic mean:
where X is the raw data and N or n is the number of scores.

Advantages of Arithmetic Mean[1,3]
Easy to calculate
Easy to understand

Used in further mathematical calculations: for example used in calculation of t & z tests. It is also used to calculate mean deviation variance, coefficient of variance and standard deviation.

It is also used to contract confidence interval and standard normal curve.

There is only single mean in the data (Uniqueness)

Whole data is utilized for calculation of mean (it is based on all the values)

It is used in descriptive as well as in inferential statistic.

Indispensable in health, business and commerce.

It resists best the influence of fluctuation between different samples (Best resists the sampling fluctuation).

Disadvantages of Arithmetic Mean[1,6]
The arithmetic mean is highly affected by extreme values.

It cannot average the ratios and percentages properly

It cannot be computed if any item is missing.

Distorted by extreme values, for example monthly income of 3families is rupees 4000.00, 5000.00 and 9000 respectively. The mean income will be If 4th family with the income of Rs.10000 (extreme value) is added, the mean income will just from Rs. 6000 to Rs. 29500 (that will only true for 1family and 3families will have income much less than Rs. 29500

$$\frac{4000+5000+9000}{3} = Rs.6000$$

Sometimes is looks ridiculous. For example, three families has have 4, 3 and 6 children. The mean number of children per family will be 4.33 children per family looks ridiculous as it cannot happen.

$$\frac{4+3+6}{3} = 4.33$$

Mean is affected by the magnitude to every observation in a data set.

Sometimes it gives the value that is not present in the data.

CALCULATION METHODS OF ARITHMIC MEAN[1,5]
FOR UNGROUPED DATA

A) DIRECT METHOD

Let the variable X taken the values X1, X2, X3, …, Xn, then the A.M, denoted by X (read "X bar") is defined by

$$X = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum X_i}{n}$$

where

X = Arithmetic mean, ∑Xi = Sum of all the items of the variable X.

n = Number of Observations.

B) CHANGE OF ORIGIN

When a certain constant, say A is added to all the observations we get a new set of observations. Similarly, the observations can be decreased when a constant is subtracted from all the observations. The addition or subtraction of a constant is called change of origin. This constant may be called arbitrary origin or working mean or Provisional mean (P.M).

$$\bar{X} = A + \frac{\sum D}{n}$$

OR

$$\bar{X} = P.M + \frac{\sum D}{n}$$

Where

"D" is the new variable and A is a constant. Where D = X – A

This relation can also be used for the calculation of X. some people call it "short cut method/Deviation Method".

If the size of the observation is very large, the computational work can be reduced by using the idea of change of origin.

FOR GROUPED DATA

A) Direct Method

$$X = \frac{\sum fx}{\sum f}$$

B) Short cut Method/Deviation Method (Change of Origin).

$$X = A + \frac{\sum fD}{\sum f}$$

C) Coding Method/Step-Deviation Method (Calculating series for continuous series) or Change of Origin and Scale

When both the operations of addition (or sub traction) and multiplication (or division) with a constant are applied on the observations we get a new variable ui, called the coded variable.

Thus the values of ui

$$ui = \frac{Xi - A}{h \longrightarrow \text{class interval}}$$

$$\bar{x} = A + h\frac{\sum fu}{\sum f} = A + h\bar{u}$$

$$\bar{u} = \frac{\sum fu}{\sum f}$$

GROUPED DATA

a) Continuous Series
b) Discrete Series

FREQUENCY DISTRIBUTION (DISCRETE SERIES)

$$\bar{u} = \frac{\sum fiXi}{\sum fi} \text{ or } \frac{\sum fiXi}{n}$$

WEIGHTED ARITHMETIC MEAN

Weighted means occur when we have some observations that we wish to place more importance on than others.

They require a weighting variable that indicates the importance to place on a given observation.

We denote the original variable by Xi and the weighting variable by Wi.[7]

Sometimes different observations do not have the same importance. Some observation, for some reason, have greater importance. The relative importance called the weight (w) of the observations is thus determined. If the observations X1, X2, …, Xn have the respective weights W1, W2, W3,…Wn, the weighted A.M denoted by Xw is defined as[1].

$$\bar{X}_w = \frac{W_1X_1 + W_2X_2 + \cdots + WnXn}{W_1 + W_2 + \cdots + Wn}$$

$$\bar{X}_w = \frac{\sum WX}{\sum W}$$

MEAN OF COMPOSITE GROUP OR COMBINED ARITHMATIC MEAN[1]

If X1, X2,… XK be the A.M of K distributions with respective frequencies n1, n2, …, nK. The combined A.M Xc is defined by

$$\overline{\overline{X}} = \bar{X}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n_1 + n_2 + \ldots + n_k}$$

$$\bar{X}_c = \frac{\sum n_i\bar{X}_i}{\sum n_i}$$

PROPERTIES OF A.M[3,5]

Some of the mathematical properties of the A.M

i) The sum of the deviations of the values Xi from their mean X is Zero i.e.

Un-grouped Data

$$\sum(X_i - \bar{X}) = \sum X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0$$

Grouped Data

$$\sum f_i(X_i - \bar{X}) = \sum fX_i - \bar{X}\sum f_i$$
$$= \bar{X}\sum f_i - \bar{X}\sum f_i = 0$$

ii) The sum of the squares of the deviations of the values of a variable is least when the deviations are measured from their mean i.e.

$$\sum f_i(X_i - \bar{X}) = \sum fX_i - \bar{X}\sum f_i$$
$$= \bar{X}\sum f_i - \bar{X}\sum f_i = 0$$

iii) If X1X2X3...Xk be the A.M of K distributions with respective frequencies n1, n2,…, nK, the combined mean Xe of the whole distribution is given by

$$\sum(X_i - \bar{X})^2 \leq \sum(X_i - A)^2 \; ungrouped$$

$$\sum f(X_i - \bar{X})^2 \leq \sum f_i(X_i - A)^2 \; Grouped$$

iv) If X1X2X3...Xn be the n observations having arithmetic mean (X) and if Y = ax + b then Y = ax + b, Y= ax +b where a & b are any two numbers and a ≠ o. This is called linear transformation of X into Y.

$$\bar{X}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots + n_k\bar{X}_k}{n_1 + n_2 + n_3 + \cdots + n_k}$$

Trimmed Mean[8]

Definition. "The trimmed mean of a set of data is the mean of the middle part of the data, after trimming off the extreme pay rates."[9]

Calculation. The trimmed mean is calculated by taking the same number of pay rates from the top and bottom of a set of salaries and calculating the simple mean. The simple mean is calculated by totaling the average salary paid in each organization and then dividing by the total number of organizations (see Exhibit 11).

Mathematical characteristics. This measure of central tendency excludes the extreme pay rates in a distribution while utilizing the preponderance of the remaining salary data. The trimmed mean provides a balance or compromise between the median and the mean. The trimmed mean can also take the form of either the simple mean or the weighted mean.

Common versions of the trimmed mean include a. excluding the highest and the lowest pay rates and computing the mean of the remaining pay data, b. eliminating the top 5% and the bottom 5% of pay rates and computing the mean of the middle 90% of the pay data, and c. removing the top 10% and the bottom 10% of pay rates and computing the mean of the middle 80% of the pay data.

Note: Trimming more than 20% of the pay rates is generally not advised. You might as well just use the median salary rate, which trims all but one or two pay rates.

Median (X)[9.5]

When data have been arranged in rank order the measure of central tendency may be found by locating a point that divides the whole distribution into two equal halves. Thus median may be dqfined as the point on the scalc of measurement below and above which lie exactly 50 percqnt of the cases. Median can therefore be found for truncated (incomplete) data provided we know the total number of cases and their possible placements on the scale. It may be noted that rpedian is defined as a point and not as a score or any particular measurement.

UN-GROUPED DATA

For odd numbers

Med = The value of (n + 1) the item
2

For even numbers

$$Med = \frac{1}{2}\left[The\ value\ of\ \left(\frac{n}{2} + \frac{n+2}{n}\right)\ the\ item\right]$$

GROUPED DATA
FOR CONTNUOUS SERIES

Medium = L + h/f (n c)
2

where

n = Sum of the frequencies

C= Cumulative frequency of the class preceding the median class.

L = Lower limit of median group.

h = Class interval.

f = Frequency of the median class.

FOR DISCRETE SERIES

Med:- The value of n+1 the item.
2

The advantages of the median are:[5,6]

i) It is easily calculated and understood.
ii) It is located even when the values are not capable of quantitative measurement.
iii) It is not affected by extreme values. It can be computed even when a frequency distribution involves open-end classes like those of income and prices.
iv) In a highly skewed distribution median is an appropriate average to use.

The median has the following disadvantages:[5,6]

i) It is not rigorously defined.
ii) It is not capable of lending itself to further statistical treatment.
iii) It necessitates the arrangement of data into an array which can be tedious and time consuming for a large body of data.

QUARTILES,DECILES, PERCENTILES QUARTILES[10]

There are three quartiles called 1st quartile, 2nd quartile and 3rd quartile. These quartiles divide the set of observations into four equal parts. The second quartile is equal to the median. The first quartile is also called lower quartile and is denoted by Q1. The third quartile is also called upper quartile and denoted by Q3.

The lower quartile Q1 is a point which has 25% observations less than it. The upper quartile Q3 is a point with 75% observations below it and 25% observations above it.

QUARTILES FOR INDIVIUAL OBSERVATIONS (UN GROUPED DATA)1

$$Q1 = \frac{\text{value of } (n + 1) \text{ th item}}{4}$$

$$Q2 = \frac{\text{value of } 2\,(n + 1) \text{ th item}}{4} = \text{Medium}$$

$$Q3 = \frac{\text{value of } 3\,(n + 1) \text{ th item}}{4}$$

QUARTILES FOR A FREQUENCY DISTRIBUTION (DISCRETE DATA)[1]

$$Q1 = \frac{\text{value of } (n + 1) \text{ th item}}{4}$$

$$Q2 = \frac{\text{value of } 2\,(n + 1) \text{ th item}}{4} = \text{Medium}$$

$$Q3 = \frac{\text{value of } 3\,(n + 1) \text{ th item}}{4}$$

QUARTILES FOR GRUPED FREQUENCY DISTRIBUTION (CONTINUOUS DATA)[1]

$$Q1 = i + \frac{h}{f}\left(\frac{n}{4} - c\right) \quad \text{where} \quad n = \Sigma f$$

$$Q2 = i + \frac{h}{f}\left(\frac{2n}{4} - c\right)$$

$$Q3 = i + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

DECILES[10]

The deciles are the partition values which divide the set of observation into ten equal parts. There are nine deciles namely D1, D2, D3,..., D9. The first decile D1 is a point which has 10% of the observations below it.

DECILES FOR INDIVIDUAL OBSERVATIONS (UNGROUPED DATA)[5]

$$D1 = \frac{\text{value of } (n + 1) \text{ th item}}{10}$$

$$D2 = \frac{\text{value of } 2\,(n + 1) \text{ th item}}{10} = \text{Medium}$$

$$D3 = \frac{\text{value of } 3\,(n + 1) \text{ th item}}{10}$$

.

D9 = value of 9 (n + 1) th item / 10

DECILES FOR A FREQUENCY DISTRIBUTION (DISCRETE DATA) [5]

D1 = value of (n + 1) th item / 10

D2 = value of 2 (n + 1) th item / 10 = Medium

D3 = value of 3 (n + 1) th item / 10

.
.
.
.

D9 = value of 9 (n + 1) th item / 10

DECILES FOR GROUPED FREQUENCY DISTRIBUTION (CONTINOUS SERIES)[5] PERCENTILES[10]

$$D1 = l + \frac{h}{f}\left(\frac{n}{10} - c\right)$$

$$D2 = l + \frac{h}{f}\left(\frac{2n}{10} - c\right)$$

$$D3 = l + \frac{h}{f}\left(\frac{3n}{10} - c\right)$$

$$D9 = l + \frac{h}{f}\left(\frac{9n}{10} - c\right)$$

The percentiles are the points which divide the set of observation into one hundred equal parts. These points are denoted by P1, P2, P3, … , P99, and are called the first, second, third, … , ninety nine percentiles. The percentiles are calculated for very large number of observations like workers in factories and the population in provinces or countries. The percentiles are usually calculated for grouped data.

PERCENTILES FOR INDIVIDUAL OBSERVATIONS (UNGROUPED DATA)

P1 = The value of (n + 1) th item / 100

P2 = The value of 2 (n + 1) th item / 100

P3 = The value of 3 (n + 1) th item / 100

FOR GROUPED DATA

$$P1 = l + \frac{h}{f}\left(\frac{n}{100} - c\right)$$

$$P2 = l + \frac{h}{f}\left(\frac{2n}{100} - c\right)$$

$$P3 = l + \frac{h}{f}\left(\frac{3n}{10} - c\right)$$

$$D99 = l + \frac{h}{f}\left(\frac{99n}{100} - c\right)$$

NOTE

It may be noted that all the partition points can be expressed in terms of percentiles. For example

Median = D5 = P50 = Q2
Q1 = P25
Q3 = P75, D1 = P10, D3 = P30
D7 = P70, D9 = P90

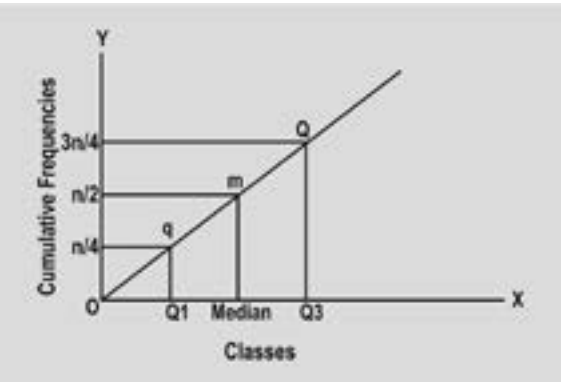
GRAPHIC LOCATION OF MEDIAN AND PARTITION VALUES[5]

Mode (X)

- A distribution can have more than one mode.
- A distribution with 2 modes is called bi-modal.[2]
- A distribution with 3 or more modes is called multi-modal.
- The mode is most useful when used with categorical (nominal, or ordinal) variables

and when there is a relatively small sample size.

- The mode can be used with categorical variables because it does not require the data to be in a meaningful order.
- With large samples, it becomes very tedious to determine the mode.



UNGROUPED DATA

Mode is that observation which occurs maximum number of times in the data.

MODE FOR A FREQUENCY DISTRIBUTION (DISCRETE DATA)

In a discrete series the value of the variable against which the frequency is maximum would be the modal value.

MODE FOR GROUPED DATA

OR

Where

l = Lower limit of the Modal class

fm = Maximum frequency of the Modal class

f1 = Frequency of the class preceding the Modal class.

f2 = Frequency of the class following the modal class.

h = Length of the class interval of Modal class.

$$\text{Mode} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

NOTE

Sometimes two classes have equal maximum frequency, in that case mode should be calculated.

While calculating the value of mode it should be seen that class intervals of the different classes are equal otherwise the above formula cannot be applied.

$$= l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

When mode is ill defined the above formula is not applicable. In that case the value mode is indirectly obtained by applying the following formula.

Mode = 3, median – 2 mean

The advantages of the mode are:

i) It is simply defined and easily calculated. In many cases, it is extremely easy to locate the mode.

ii) It is not affected by abnormally large or small observations.

iii) It can be determined for both the quantitative and qualitative data.

The disadvantages of the mode are:

i) It is not rigorously defined.

ii) It is often indeterminate and indefinite.

iii) It is not based on all the observations made.

iv) It is not capable of lending itself to further statistical treatment.

v) When the distribution consists of a small number of values, the mode may not exist.

EMPIRICAL RELATION B/W MEAN MEDIAN AND MODE [1,12]

In a moderately skewed distribution, median lies in between the other two averages i.e.

mean and mode. For a moderately skewed distribution there exists an empirical relationship among the mean, median and mode. The difference b/w mean and median is half of the distance b/w median and mode. The difference b/w mean and mode is three times the difference b/w mean and median. The relation b/w them is

Mean = 1/2 (3median – mode)

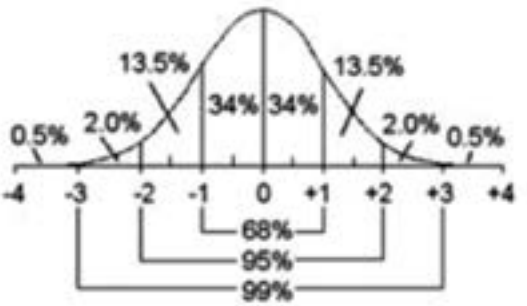
Median = 1/3 (2mean + mode)

Mode = 3 median – 2 mean

NORMAL CURVE

In a single, packed and symmetrical distribution, the values of mean median and mode coincide.

A curve is said to be normal when it is



1) Bell Shaped
2) Smooth
3) Symmetrical
4) Two halves are mirror image of each other and
5) When folded upon, are congruent and superimposed

STANDARD NORMAL CURVE

In a distribution

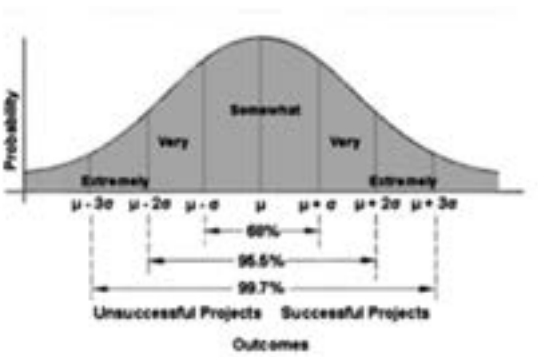
If mean is zero (mean = 0) and Standard deviation is one.

Curve is said normal standard curve

SKEWED/ ASYMMETRICAL CURVE

But if the values, mean median and mode do not coincide and different then the distribution is said to skewed.

Generally, mode are used for nominal scores, median for ordinal scores, and means for Interval scores.



GEOMETRIC MEAN[13,5]

If the observations instead of being added ,are multiplied , the geometricmean would be the nth root of the product ,In algebric symbols, the geometric mean of n observations is given by the formula

$G = (X1, X2, X3, ...Xn)1/n = n\sqrt{X1, X2, X3, ...Xn}$

FOR UNGROUPED DATA

$G = (X1, X2, X3, ...Xn)1/n = n\sqrt{X1, X2, X3, ...Xn}$

or

$$G = \text{Antilog}\left(\frac{\Sigma \log Xi}{n}\right)$$

Where

n stands for the number of observations and X1, X2, X2, … Xn are various values.

FOR GROUPED DATA

$$G = \text{Antilog}\left(\frac{\Sigma f \log Xi}{\Sigma f}\right)$$

The advantages of the geometric mean are.

i) It is rigorously defined by a mathematical formula.
ii) It is based on all observed values.
iii) It is amenable to mathematical treatment in certain cases.
iv) It gives equal weight age to all the observations.
v) It is not much affected by sampling variability.
vi) It is an appropriate type of average to be used in case rates of change or ratios are to be averaged.

Disadvantages are:
i) It is neither easy to calculate nor to understand.
ii) It vanishes if any observation is zero.
iii) In case of negative values, it cannot be computed at all.

PROPERTIES

The G.M is less than A.M i.e.
G.M < A.M
The product of the items remains unchanged if each item is replaced by the G.M.

HARMONIC MEAN [14,1]

It is the resiprocal of the mean of resiprocal values.

H.M is defined as the reciprocal of the mean of the reciprocals of the items in a series. It is the ratio of the number of items and the sum of reciprocals of items. If the observation are X1, X2, X3, … Xn, then the H.M 'H' is

$$H = \frac{\text{Number of items}}{\text{Sum of reciprocals of item}}$$

$$H = \left(\frac{n}{\Sigma (1/Xi)} \right)$$

The calculation of Harmonic mean is not possible if any item is equal to zero.

GROUPED DATA

The advantages of the harmonic mean are:
i) It is rigorously defined by a mathematical formula.
ii) It is based on all the observations in the data.
iii) It is amenable to mathematical treatment.
iv) It is not much affected by sampling variability.
v) It is an appropriate.

The disadvantages of the harmonic mean are:
i) It is not readily understood.
ii) It cannot be calculated, If any one of the observations is zero.

$$H = \left(\frac{\Sigma f}{\Sigma (f/X)} \right)$$

It gives too much weightages to the smaller observations.

REFERRENCE

1. (Basic statistics by Gulam hussain kiani)
2. (Quantitative aptitude & Business Statistics: Measures of Central a)
3. (text book statistics punjab board)
4. (Microbiology @ Leicester: Maths & Computers for Biologists: Descriptive Statistics Updated: October 15, 2004 Search (C)
5. (Introduction to statistical theory by shair muhammad and
6. Statistical Methods & Data analysis by faqir muhammad)
7. (These slides are copyright © 2003 by Tavis Barr. This material may be distributed only subject to the terms and conditions set forth in the Open Publication License, v1.0 or later (the latest version is presently available at http://www.opencontent.org/openpub D)
8. (Measures of Central Tendency,Location,and

Dispersion in Salary Survey Research Robert M. Halley Compensation Consultant Swanson Consulting (G)

9 (MEASURES OF CENTRAL TENDENCY (F)

10. (bio stat by Hanif)

11. (Universit of guelph Numeracy Project b)

12. HIGH YIELD

13. Biostatistics K Visweswara Rao)