USAID | AFGHANISTAN
FROM THE AMERICAN PEOPLE

USAID/CAPACITY BUILDING ACTIVTIY (CBA)

TRAINING MANUAL

FOR

INTERMEDIATE LEVEL DATA ANALYSIS

Manual

FY 2018 (November 15 - 15 December 2018)

This report is made possible by the support of the American People through the United States Agency for International Development (USAID.) The contents of this report are the sole responsibility of Chemonics International and do not necessarily reflect the views of USAID or the United States Government.

Government of Islamic Republic of Afghanistan

Ministry of Education

Training Manual
For
Intermediate Level Data Analysis

Manual

Prepared by: EMIS COMPONENT OF USAID/CAPAPCITY BUILDING ACTIVITY (CBA)

Table of Contents

TRAINING OBJECTIVES: 5
WHAT WILL THE PARTICIPANTS LEARN? 5
ACTIVITIES: 6
Training Agenda 8
DESCRIPTIVE STATISTICS 9
MEASURE OF CENTRAL TENDENCY 9
TYPE OF AVERAGE 9
THE ARITHMETIC MEAN 9
CALCULATION OF ARITHMETIC MEAN (اوسط حسابي) BASED ON UNGROUPED DATA 9
CALCULATION OF ARITHMETIC MEAN BASED ON GROUPED DATA 10
COMBINED ARITHMETIC MEAN 10
THE WEIGHTED ARITHMETIC MEAN: 10
THE MEDIAN 11
OTHER PARTITIONED VALUES 12
QUARTILES 12
DECILES 13
PERCENTILE 14
THE MODE (OR) MODEL VALUE 16
CALCULATION OF MODE BASED ON UNGROUPED DATA 16
EMPIRICAL RELATION BETWEEN MEAN, MEDIAN AND MODE 16
MEASURES OF DISPERSION, MOMENTS, AND SKEWNESS: 17
DISPERSION (VARIATION) 17
TYPE OF DISPERSION 17
ABSOLUTE MEASURE OF DISPERSION 17
RELATIVE MEASURES OF DISPERSION: 17
THE RANGE 17
COEFFICIENT OF RANGE 18
THE VARIANCE 18
THE STANDARD DEVIATION 18
COEFFICIENT OF VARIATION 18

MOMENTS 19
POSITIVE SKEWNESS 20
NEGATIVE SKEWNESS 20
KURTOSIS 20
SYMMETRICAL DISTRIBUTION 21

Training Program- II:

INTERMEDIATE LEVEL DATA ANALYSIS

TRAINING OBJECTIVES:

By the end of this training session, participants will be able to perform independently descriptive analysis of the collected Educational Data. They should be able to validate their understanding by providing examples of each type of data.

WHAT WILL THE PARTICIPANTS LEARN?

- The participants will learn about descriptive statistics and some preliminary measures which are being used during data analysis and analytical report writing.
- The participants will be able that “How Statistics Works” in educational data analysis
- The participants will be able to understand what a descriptive statistic is? and could differentiate between several types of descriptive measures. they will learn how to appropriately and objectively present the data in hand.
- The participants will be able to understand and calculate and use suitably some measures of locations & measures of variations.
- The participants will be able to understand the use and computation of some other important concepts in descriptive statistics like moments, skewness and kurtosis.

ACTIVITIES:

Activity – 0: (The Participants are required to participate in a pre-test, which will show their prior knowledge about statistical measures which are commonly used in data analysis reports and statistical software packages. The trainer will provide a brief overview of the training and clearly describe the relationship of this training level with the previous training level.

Activity – I: **(The Participants are introduced to measures of location like, (Mean, Median, Mode, and empirical relation between Mean, Median and Mode), the activity is facilitated by the trainer)**: The participants are introduced to the various measures of location through PowerPoint presentation by the trainer. The participants take part in the discussion related to each topic and are able to answer trainer's questions to determine their understanding of the topic. The participants are encouraged to discuss their questions with the trainer in order to help enhance their understanding.

Activity – II: **(Group Work: The Participants undertake the group work on identifying and calculating different measures of location by means of examples in education data, the activity is facilitated by the trainer)**: The participants discuss various measures of location and share their understanding with each other in the group. The EMIS representative, present in the training, demonstrates the computation formulae on data and how to calculate the required information from these measures. Participant groups are given the target of calculating different examples from data.

Activity – III: **(The Participants are introducing to other partitioned measures like Quartiles, Deciles and percentiles. The activity is facilitated by the trainer)**: The participants are introduced to the various measures of Partitioned through PowerPoint presentation by the trainer. The participants take part in the discussion related to each topic and are able to answer trainer's questions to determine their understanding of the topic. The participants are encouraged to discuss their questions with the trainer in order to help enhance their understanding

Activity – VI: (Group Work: The Participants undertake the group work on identifying and calculating different Partitioned measures, like Quartiles, Deciles and Percentiles for ungrouped and grouped data by means of examples in education data, the activity is facilitated by the trainer): The participants discuss various partitioned measures and share their understanding with each other in the group. The trainer during the training, demonstrates the computation formulae on data and how to calculate the required information from these measures. Participant groups are given the target of calculating different examples from data.

Activity – V: (The Participants are introduced to measures of Variation like, (Variance, Standard Deviation and Coefficient of Variation), the activity is facilitated by the trainer): The participants are introduced to the various measures of Variation (Dispersion) through PowerPoint presentation by the trainer. The participants take part in the discussion related to each topic and are able to answer trainer's questions to determine their understanding of the topic. The participants are encouraged to discuss their questions with the trainer in order to help enhance their understanding

Activity – VI: (Group Work: The Participants undertake the group work on identifying and calculating different measures of Variation by means of examples in education data, the activity is facilitated by the trainer): The participants discuss various measures of location and share their understanding with each other in the group. The trainer during the training, demonstrates the computation formulae on data and how to calculate the required information from these measures. Participant groups are given the target of calculating different examples from data.

Activity – VII: (Quiz and Post-test: The Participants take a quiz related to the purpose and use of various Statistical measures): In this session the trainer will provide a short overview of all the work done during last sessions. The participants provided with list of different measures, and they are required to express their understanding of utilization of each measure. They are required to present their work. The participants will also participate in the post-test session.

EMIS Training Agenda

Training on Data Analysis (Intermediate level)

Training Agenda

Duration	Activity	What is needed?
TRAINING – I: Intermediate Level Training on Data Analysis		
Day first		
00:20 Minutes	Pre – test and brief overview of the training by trainer	– Paper & Pen – Lecture
1 Hour 15 Minutes	Participatory Lecture and Power point presentation by the trainer on some basic concepts of statistical averages, like Arithmetic Mean, Weighted Mean Median & Mode	– Projector – Power Point Presentation
Tea Break for 15 Minutes		
1 Hour 15 Minutes	Group work, Discussions and Findings	– Projector – Flip Chart – Notebook and Pen
Lunch & Prayer Break (12:00 – 01:00)		
1 Hour 30 Minutes	Participatory Lecture and Power point presentation by the trainer on other Partitioned Values like Quartiles, Deciles and Percentiles	– Projector – Power Point Presentation
1 Hour 30 Minutes	Group work and Discussions	– Projector – Flip Chart – Notebook and Pen
Day Second		
1 Hour 20 Minutes	Participatory Lecture and Power point presentation by the trainer on measures of Dispersion (Range, Variance & Standard Deviation)	– Projector – Power Point Presentation
Tea Break for 15Minutes		
1 Hour 20 Minutes	Group work, and Discussions	– Projector – Flip Chart – Notebook and Pen
Lunch & Prayer Break (12:00 – 01:00)		
1 Hour 30 Minutes	Participatory Lecture and Power point presentation by the trainer on measures of Variation (Dispersion)	– Projector – Power Point Presentation
1 Hour 30 Minutes	Group work, and Discussions	– Projector – Flip Chart – Notebook and Pen

DESCRIPTIVE STATISTICS

WHAT IS DESCRIPTIVE STATISTICS: Descriptive statistics are numbers that are used to summarize and describe data. There are two different types of descriptive statistics: (a) measures of central tendency (Measures of Location) and (b) measures of dispersion (Measures of Variation).

MEASURE OF CENTRAL TENDENCY

A value which is used in this way to represent the distribution is called an average. Since the averages tend to lie in the Centre of the distribution, they are called measure of central tendency. They are also called measure of location because they locate the center of the distribution.

TYPE OF AVERAGE

The most commonly used averages are:

- ✓ THE ARITHMETIC MEAN
- ✓ GEOMETRIC MEAN
- ✓ THE MEDIUM (QUARTILES, DECILES & PERCENTILES)
- ✓ THE MODE

THE ARITHMETIC MEAN

The arithmetic mean is a defined as a value obtained by dividing the sum of the values by their number. It is donated by $\bar{x}$ (x – bar) and is given by the formula:

$$\bar{x} = \frac{\sum x}{n} \left[for\ ungrouped\ data \right]$$

$$\bar{x} = \frac{\sum fx}{\sum f} \left[for\ grouped\ data \right]$$

n= number of values in the data

$\sum$ = Sigma= sign of summation

f = frequency

CALCULATION OF ARITHMETIC MEAN (اوسط حسابی) BASED ON UNGROUPED DATA

EXAMPLE: the values are:

25,26,27,11,15,21,12,25,26,41,5,12,0,29,33,7,20,32,14,38,27,11,43,40,45

$$\bar{x} = \frac{\sum x}{n} = \frac{578}{25} = 23.12$$

CALCULATION OF ARITHMETIC MEAN BASED ON GROUPED DATA

EXAMPLE:

Classes	F	x	f(x)
0 - 9	3	4.5	13.5
10 - 19	6	4.5	87
20 - 19	9	24.5	220.5
30 - 39	4	34.5	138
40 - 49	3	44.5	133.5
	$25 = \sum F$		$592.5 = \sum f(x)$

$$\bar{x} = \frac{\sum fx}{\sum f}$$

COMBINED ARITHMETIC MEAN

If $\bar{x}_1, \bar{x}_2, \bar{x}_3$ ..... $\bar{x}_n$ be the arithmetic mean of k-subgroups of data with respective frequencies n_1, n_2, n_3..... n_k. Then the combined mean $\bar{x}_c$ is defined by:

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3 + \ ..... \ + n_k\bar{x}_k}{n_1 + n_2 + n_3 + ... + n_k}$$

$$\bar{x}_c = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}$$

THE WEIGHTED ARITHMETIC MEAN:

The multipliers of a set of numbers, which express more or less adequately the relative importance of various values in a set of data are technically called the weight.

We assign weights w_1, w_2, w_3 ..... w_n to the values in a set of data according to their relative importance, when the values are not of equal importance. The weighted mean

denoted by "$\bar{x}_w$" of set of n-values x_1, x_2, x_3 ..... x_n with corresponding weights w_1, w_2, w_3 ..... w_n is then defined as:

$$\bar{x}_w = \frac{\sum wx}{\sum w}$$

THE GEOMETRIC MEAN

The G.M of n-positive values is defined as the nth root of their product. In other words, it's obtained by multiplying together all the n-values and then taking the nth root of their product. Thus

G.M= $\sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$

This method of calculating G.M satisfactory only if there are two or three values, but if the number of values (n) is large the problem of compiling the nth-root of the product of values by the above method is tedious work. To facilitate the computation of Geometric Mean, we make use of logarithms. The above formula when reduced to its logarithms form can be written as:

$$\text{G.M=Antilo}\left[\frac{\sum \log x}{n}\right] [for\ ungrouped\ data]$$

$$\text{G.M=Antilog}\left[\frac{\sum f \log x}{\sum f}\right] [for\ grouped\ data]$$

[The geometric mean is appropriate to average ratio and rates of change]

THE MEDIAN

This is the positional average of a set of data. Median is a value which divides an arrayed (ترتیب شوی) set of data into two equal halves. The number of values greater than the median is equal to the number of values smaller than the median. It's given by formula:

$$\text{Median} = \frac{(n+1)}{2} \text{ th value [for ungrouped data]}$$

Note:

Median=50%

Quartiles=25%

Deciles=10%

Percentiles=100%

Median $= l + \frac{h}{f}(\frac{n}{2} - C)$ [for grouped data]

Where: L= lower class-boundary of the median class

h= class- interval of the median class

f= frequency of the median class

C= cumulative (جمع شونده) frequency of the class preceding the median class

n= total frequency $= \sum f$

OTHER PARTITIONED VALUES

QUARTILES

Quartiles are the values which divide an arrayed (ترتیب شوی) set of data into four equal parts. The first, 2nd and 3rd quartiles are donated by Q1, Q2 and Q3 respectively (بترتیب) they are given by the formula:

In case of ungrouped data;

$Q_1 = \frac{(n+1)}{4}$ Th value [first or lower quartile]

$Q_2 = \frac{2(n+1)}{4}$ Th value [second quartile]

$Q_1 = \frac{3(n+1)}{4}$ Th value [3rd quartile or upper quartile]

Where “n” denotes the number of values

In case of grouped data:

$$Q_1 = l + \frac{h}{f}(\frac{n}{4} - C)$$

$$Q_2 = l + \frac{h}{f}(\frac{2n}{4} - C)$$

$$Q_3 = l + \frac{h}{f}(\frac{3n}{4} - C)$$

Where;

$l =$ Lower class boundary of the quartile class

$h =$ Class interval of the quartile class

$f =$ Frequency of the quartiles class

$n =$ Total frequency= $\sum f$

$c =$ Cumulative frequency of the class preceding the quartile class

DECILES

Deciles are the values which divide an arrayed set of data into ten equal parts. The 1st, 2nd, 3rd, and 9th deciles are donated by $D_1, D_2, D_3 \ldots D_9$ respectively. They are given by formula:

In case of ungroup data:

$$D_1 = \frac{(n+1)}{10} \text{ th value}$$

$$D_2 = \frac{2(n+1)}{10} \text{ th value}$$

$$D_3 = \frac{3(n+1)}{10} \text{ th value}$$

:

$$D_9 = \frac{9(n+1)}{10} \text{ th value}$$

Where “n” denotes the number of values in a set of data

In case of grouped data;

$$D_1 = l + \frac{h}{f}\left(\frac{n}{10} - C\right)$$

$$D_2 = l + \frac{h}{f}\left(\frac{2n}{10} - C\right)$$

$$D_3 = l + \frac{h}{f}\left(\frac{3n}{10} - C\right)$$

:

$$D_9 = l + \frac{h}{f}\left(\frac{9n}{10} - C\right)$$

Where is:

$l =$ Lower class boundary of the docile class

$h =$ Class interval of the docile class

$f =$ Frequency of the docile class

$n =$ Total frequency= $\sum f$

$c =$ Cumulative frequency of the class preceding the deciles class

PERCENTILE

Percentiles are the values which divide an arrayed set of data into hundred equal parts. The first, 2nd, 3rd … and 99th percentiles are donated by P_1, P_2, P_2 … P_{99} respectively they are given by the formula:

IN CASE OF UNGROUPED DATA	IN CASE OF GROUPED DATA
$P_1 = \frac{(n+1)}{100}$ th value	$P_1 = l + \frac{h}{f}(\frac{1n}{100} - C)$
$P_2 = \frac{2(n+1)}{100}$ th value	$P_2 = l + \frac{h}{f}(\frac{2n}{100} - C)$
$P_3 = \frac{3(n+1)}{100}$ th value	$P_3 = l + \frac{h}{f}(\frac{3n}{100} - C)$
:	:
$P_{99} = \frac{99(n+1)}{100}$ th value	$P_{99} = l + \frac{h}{f}(\frac{99n}{100} - C)$

Where:

$l =$ Lower class boundary of the Percentile class

$h =$ Class interval of the Percentile class

$f =$ Frequency of the Percentile class

$n =$ Total frequency= $\sum f$

$c =$ Cumulative frequency of the class preceding the Percentile

THE MODE (OR) MODEL VALUE

The French word “mode” meaning fashion has been adopted (اختیارول) to convey the idea of “most frequent”. The mode is defined as a value which accurse most frequently in a set of data. A set of data may have more than one mode or no mode at all.

CALCULATION OF MODE BASED ON UNGROUPED DATA

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

Where:

$l =$ Lower class boundary of the modal class

$f_m =$ Maximum frequency (model class)

$f_1 =$ Frequency of the class preceding the model class

$f_2 =$ Frequency of the class following the model class

$h =$ Class interval of the model

EMPIRICAL RELATION BETWEEN MEAN, MEDIAN AND MODE

In a symmetrical distribution the values of the mean, median and mode are coin side (same). But if these values differ the frequency distribution is said to be skewed or asymmetrical.

For a moderately skewed distribution there exist an empirical relationship among the mean, median and mode.

$$\text{Mean} = \frac{1}{2}(3\ \text{Median} - \text{Mode})$$

$$\text{Median} = \frac{1}{3}(2\ \text{Mean} + \text{Mode})$$

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$$

MEASURES OF DISPERSION, MOMENTS, AND SKEWNESS:

DISPERSION (VARIATION)

If $x_1, x_2, x_3, \ldots, x_n$ are n-observations of the variable "X" then the scatter or (spread) of the values about their Centre is called dispersion. And any measure which tells us the amount of scatter about the Centre is called the measure of dispersion or variation.

TYPE OF DISPERSION

There are two types of measure of dispersion:

i. ABSOLUTE MEASURE OF DISPERSION
ii. RELATIVE MEASURE OF DISPERSION

ABSOLUTE MEASURE OF DISPERSION

An absolute measure of dispersion is that which measures the variation present among the data in the unit of the variable.

The commonly used measures of absolute dispersion are:

i. The range
ii. The variance and standard deviation.

RELATIVE MEASURES OF DISPERSION:

A relative measure of dispersion is that which measures the variation present among the observations relative to their average. It is expressed in the form of ratio. It is independent of the unit of measurement.

The commonly used measures of relative dispersion are:

i. Coefficient of Range (or) Range Coefficient of Dispersion
ii. Coefficient of Variance (C.V.) and Standard Coefficient of Dispersion

THE RANGE

The range is defined as the difference between the largest and the smallest value in a set of data. The range is given by the formula:

$$Range = x_m - x_0$$

Where $x_m =$ The largest value in the data

$x_0 =$ The smallest value in the data

COEFFICIENT OF RANGE

It is defined to be the ratio of range by the sum of maximum and minimum values of data. That is, coefficient of range $= \frac{Range}{x_m + x_0}$

$$= \frac{x_m - x_0}{x_m + x_0}$$

THE VARIANCE

The variance is defined as the mean of the squares of deviations of all the values from their mean.

It is denoted by S^2 and is given by the formula:

$$S^2 = \frac{\sum (x - \bar{x})^2}{n} \qquad [for\ ungrouped\ data]$$

$$S^2 = \frac{\sum f(x - \bar{x})^2}{\sum f} \qquad [for\ grouped\ data]$$

THE STANDARD DEVIATION

The standard deviation is defined as the square root of the mean of the square of deviation of all values from their mean.

It is denoted by S and is given by the formula:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \qquad [for\ ungrouped\ data]$$

$$S = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}} \qquad [for\ grouped\ data]$$

COEFFICIENT OF VARIATION

It is obtained by dividing the standard deviation by the mean and multiplying the result by 100 symbolically:

Coefficient of Variation $= C.V = \frac{S}{\bar{x}} \times 100$,

As the coefficient of variation is a pure number without unit. It is therefore used to compare the variation in two or more sets data or distribution that are measured in different units.

The larger value of C.V. indicates the greater variability and a smaller value of C.V. indicates less variability.

The coefficient of variation is also used to compare the performance of two candidates or two players, given their score in various papers or games. The smaller the C.V. the more consistent is the performance of the candidates or players.

MOMENTS

Moments are of immense importance in the study of symmetry and normality of the distribution and these are defined as arithmetic mean of the various powers of deviations taken from arithmetic mean.

Moments tells us how a curve is distorted from symmetry.

Therefore, they provide a method of testing symmetry and normality of the distribution.

SKEWNESS

The term skewness means the lack of symmetry of the values about some central value i.e. mean median or mode of the variate.

$b_1(\beta_1)$ is the measure of skewness. If $b_1 = 0$, then the distribution is said to be symmetrical.

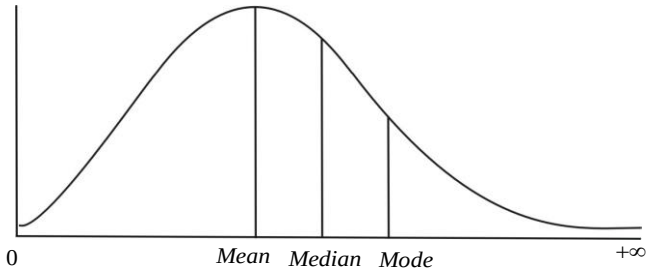
If $b_1 = +Ve$, then the distribution is said to be positively skewed (or) skewed to the right)

If $b_1 = -Ve$ then the distribution is said to be negatively skewed (or) skewed to the left). It is given by the formula $b_1 = \frac{(m_3)^2}{(m_2)^3}$

POSITIVE SKEWNESS

If the curve has a long tail towards right, then the skewness will be positive. In such case the mean is greater than median and mode. i.e.

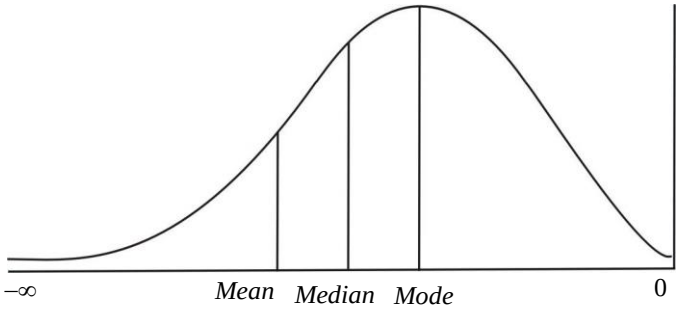
Mean> Median> Mode



NEGATIVE SKEWNESS

If the curve has a long tail towards left, then the skewness will be negative. In such a case mode would be greater than median and mean i.e.

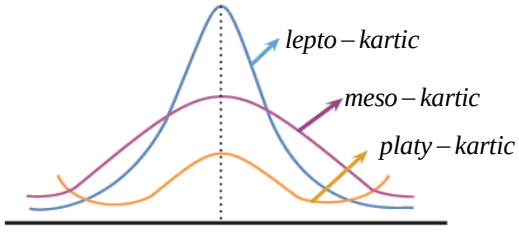
Mean< Median< Mode



KURTOSIS

The term "kurtosis "is meant to show the degree of peakedness of the distribution. $b_2(\beta_2)$ is a measure of kurtosis, it tells us about the shape of the curve of the distribution?

$$b_2 = \frac{m_4}{(m_2)^2}$$



If $b_2 = 3$, then the curve is said to be Meso- Kurtic (normal/symmetrical)

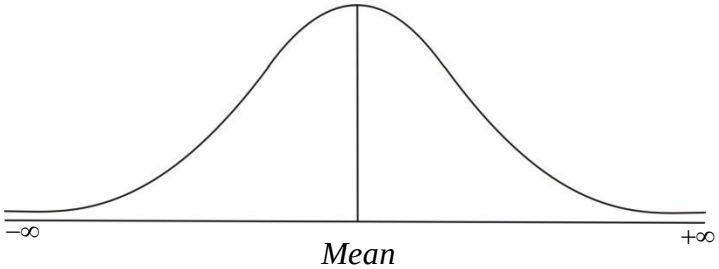
If $b_2 > 3$, then the curve is said to be Lepto- Kurtic

If, $b_2 < 3$, then the curve is said to be Platy- Kurtic

SYMMETRICAL DISTRIBUTION

For symmetrical distribution:

i. The mean, median and mod are the same
ii. The graph of the series will be bell-shaped.
iii. The pairs of such measures as quartiles, deciles and percentiles are equidistant from the mean.
iv. The sum of positive deviations from the median (mean) is equal to the sum of negative deviations.
v. Always $b_1 = 0$ and $b_2 = 3$



Mean
Median
Mode

THANK YOU FOR YOUR ATTENTION AND WARM PARTICIPATION