SCHOOL EDUCATION
DEPARTMENT: EDUCATIONAL PSYCHOLOGY

EPS 400 EDUCATIONAL STATISTICS AND EVALUATION

TABLE OF CONTENT

PAGE

LECTURE ONE

INTRODUCTION TO MEASUREMENT AND STATISTICS

INTRODUCTION

The unit focuses on what statistics is and its uses in the context of education and counselling and other related areas. Its relationship to measurement (evaluation) is brought out by analyzing keys terms namely, data (numeric information), variable, continuous and discrete variable, descriptive and inferential statistics.

In the unit the scales (levels) of measurement are also discussed, which should be of concern to any teacher (researcher) for it dictates the method of statistics (i.e. statistical analytical method), which can be used with data. This subject of determining the correct statistical method to use for particular data is not dealt with here for it is beyond the scope of the unit.

OBJECTIVES

At the end of the unit the learner should be able to:

1. Define variable, continuous and discrete variable, data, statistics, measurement, descriptive and inferential statistics
2. Explain why educators need to have rudimentary (basic) knowledge of statistics.
3. Identify the relationship between statistics and measurement.
4. List the four levels of measurement
5. Describe the characteristics (properties) of the 4 levels of measurement
6. Associate the levels of measurement with any data collected by a teacher (researcher).

Definition of statistics:

Statistics is concerned with scientific methods of collecting, organizing, summarizing, presenting (compiling) and analyzing data, as well as drawing valid conclusions and making reasonable decisions on basis of such analysis. The data complied and analyzed could be test score (if one is an educator), sales (if in business), income, enrollment in university or school, passes and failures in school, health, crime, HIV/AID data (statistics) and so on.

In a narrow sense, the term is used to denote the data themselves or the numbers derived from the data such as means. Thus the lay person will talk of employment statistics, accident statistics or school statistics while he/she means data, for instance, about employment and means of people employed during each month of the year.

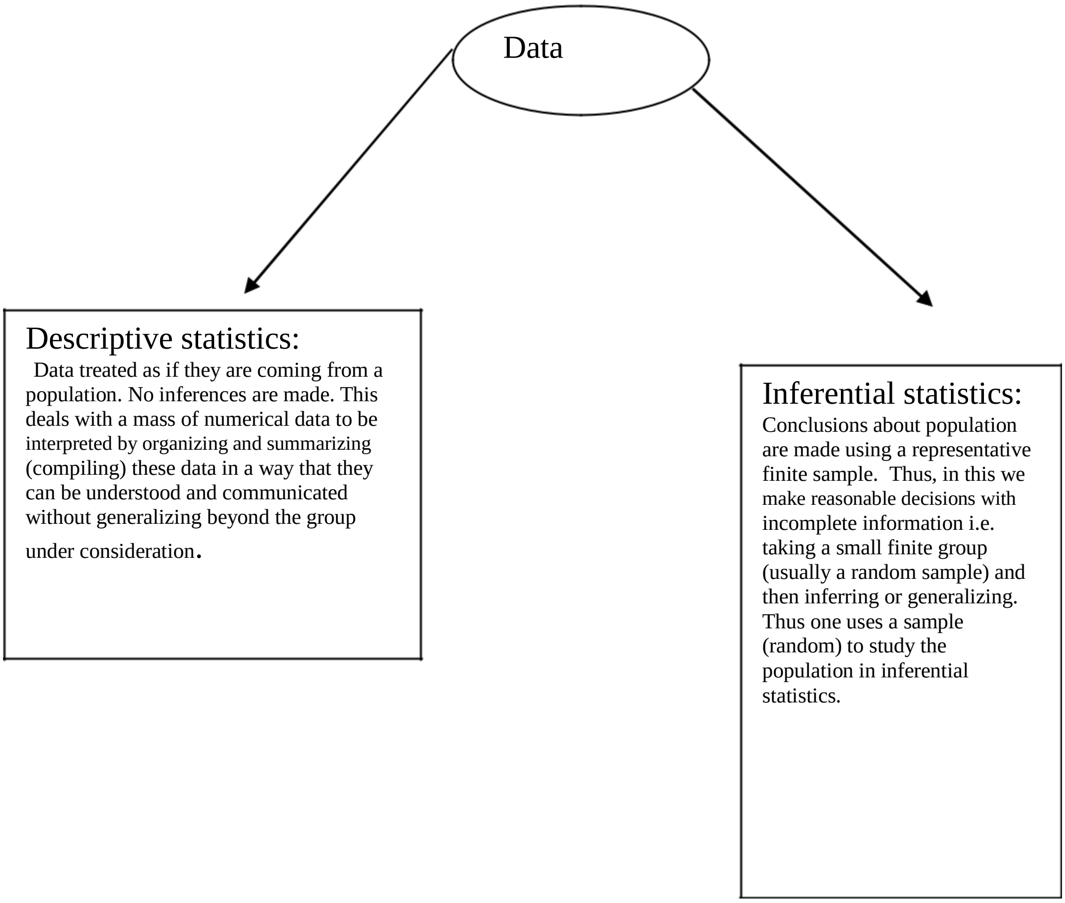
Rationale for doing statistics

Most of the journal articles in social sciences (psychology, communication, education, sociology etc.) report findings in statistical form or apply statistical concepts in their discussions or research reports. Consequently, we need at least rudimentary (basic) knowledge of statistics in order to be intelligent consumers of such research. Furthermore, if we intend to do research in our area of interest, we must understand statistics in order to use and plan appropriate procedures and to interpret and communicate the findings from our research. So statistics is needed for these purposes i.e. in:

i) Research- extension of knowledge and solving problems.

ii) In interpreting mass of data (test scores or marks, sales, etc.)

Considerations possible with data or the two ways of treating data:



Data

Descriptive statistics:

Data treated as if they are coming from a population. No inferences are made. This deals with a mass of numerical data to be interpreted by organizing and summarizing (compiling) these data in a way that they can be understood and communicated without generalizing beyond the group under consideration.

Inferential statistics:

Conclusions about population are made using a representative finite sample. Thus, in this we make reasonable decisions with incomplete information i.e. taking a small finite group (usually a random sample) and then inferring or generalizing. Thus one uses a sample (random) to study the population in inferential statistics.

What is data?
Data is numerical information.

How do we obtain data?

Measurement has to be involved in obtaining data.

Definitions of measurement, variable and different types of variables:

This is assigning of numbers to individuals or objects in a systematic way as a means of representing properties of the individuals (or objects).

This property usually referred to as a ‘variable’ is possessed by different individual or object in different quantities.

Measuring some quantities is easier than others (e.g. measuring time, length or weight). It offers no problem since we have objective physical instruments to carry out the measurement, even when interpreting these measurement there is hardly any problem. The situation is completely different when we wish to measure other than *physical variables.* Psychological variables like attitude or scores, these variables unlike physical variables are not tangible. Usually these are variables in individuals’ brain and they need to be manifested to be measured, and this why we have to use indirect ways like a test to measure them.

We are talking about variable and for sure we need to give its formal definition.

What is a variable?

We shall define a ***variable*** as any single property or characteristic, which it is possible for different individuals to possess in different quantities. In psychological variables, we meet scaling problem of very complex nature since they are not necessarily tangible.

A variable will be symbolized as X, Y, x, B (i.e. using English letters) or even Greek letters such as α, β and can assume any of a prescribed set of values, called the ‘domain’ of the variable.

If the variable can assume only one value, it is called **constant** (i.e. a variable with only one value in its domain is a constant).

A variable, which can theoretically assume any value between any two given values is called a ‘**continuous variable’,** otherwise it called a **‘discrete variable’.**

(For discrete variables their measurements can take only particular values). In other words if a variable does not take at least one value between any given two points, it is termed discrete. That is taking only selected values.

Data described by discrete variable is referred to as ‘discrete data’. Similarly data described by a continuous variable is referred to as ‘continuous data’.

Example of discrete data is the number children in each of 100 families you may consider. Example of continuous data is the height of these 100 children.

In general “measurement” gives rise to continuous data, while “counting (or enumeration)” gives rise to discrete data.

Measurement can take place at basically 4 different levels (also called scales):

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Data that you get when you measure certain property or attribute may be at the one of the 4 levels (scales), theoretically. Thus each level of measurement specifies how the numbers (data) that are assigned to the individual (objects) relate to the property being measured. Thus one needs to decide at which level data follows into.

Levels of measurement or scales of measurement

1. Nominal level

This is the most primitive level (scale) which involves number being used in place of property or attribute e.g. say the variable were interested in is sex (gender). We may assign male 0 and female 1. The difference here between 0 and 1 is that they are different and no merit is attached to any value. Other examples of nominal variables are ethnicity, religion and so on. Nominal variables are also referred to as categorical variables. The numbers represent categories, and the relationship between categories is that they are different or distinct. Counting is possible at nominal level. For instance, we can count how males we have in our example of variable gender. Note the nominal level satisfies definition of measurement for numbers are being used to represent property of individuals, and no merit is attached to the numbers at all here.

Thus to represent nominal variable (or categorical variable) we use *dummy variable.* Here the numbers carries no meaning except to distinguish and hence the term dummy variable.

2. Ordinal level

This scale does not only distinguish the individuals (objects) but gives the merit i.e. one determines the relative position of individuals with respect to some property or attribute but without indicating distance between positions. Hence we can count in ordinal scale as well as use the concept of "greater" or "less" than, without giving the amount of difference or how much is this better than this i.e. *intervals* are not meaningful. Example of data in this scale would be when we assign 3 for good, 2 for fair and 1 for poor in a grading system where we use those 3 categories. Note we can use 7 for good, 5 for fair and 2 for poor since what matters in ordinal is the merit, and the intervals (or differences) are not important or meaningful.

3. Interval scale

This scale provides equal intervals from an arbitrary origin. An interval scale not only orders individuals (or objects or events) according to the amount of attribute (property) they possess but also establishes equal intervals between the units of measure. An example of form of measurement in interval scale is the data you would obtain if you considered a multiple-choice test with, 50 items for a group of students in a certain test where the score represents the item obtain right. Note here the differences in scores are meaningful representing the different number of items got wrong or right more or less. For instance, given two scores, say 45 and 40 we not only know that 45 is better than 40 but also 5 items were missed more for the one who got 40. Thus in interval scale of measurement we can:

i. Count (i.e. distinctiveness)
ii. Use greater than or less than (i.e. there is merit or order)
iii. It can be stated meaningfully by how much two of them differ.

4. Ratio scale

The highest type is the one, which provides a true zero point as well as equal intervals. Ratios, which are meaningful, can be formed between any two given values on the scale. A metric rod used to measure length in units of cm is a ratio scale, for the origin on the scale is an absolute zero corresponding to no length at all. That is, lengths measured in say cm those numbers (data) provide ratio scale. Ratio scales are found primarily in the physical sciences (in physical variables). The best you can do in social sciences as far as level of measurement for your data is concerned, is at interval level except for variable such as height and age. In other words the highest level realized most often for social variables (psychological variables) is interval.

Characteristics of the levels of measurement

Characteristic	Nominal	Ordinal	Interval	Ratio
Count	Yes	Yes	Yes	Yes
Use greater or less than (merit or order)	No	Yes	Yes	Yes
Equal interval	No	No	Yes	Yes
Absolute zero	No	No	No	Yes

SUMMARY

The following is the summary of the major points in this unit:

1. Statistics is a scientific method of collecting, organizing, summarizing, presenting (compiling) and analyzing data
2. There are two areas of statistics namely, descriptive and inferential statistics.
3. In descriptive statistics generalizations are restricted to the group under consideration while in inferential statistics a representative sample from the population is used to study the population, thus making inferences about the population using the (representative) sample.
4. A variable is any particular property (characteristic, trait or attribute) of which different individuals (objects) will possess in different quantities.
5. Discrete variables are measured in units, which by definition, cannot be subdivided any further. On the other hand, a continuous variable is one that can take on unlimited or potentially unlimited number of values between any two given values.
6. There are basically 4 levels of measurement: nominal, ordinal, interval and ratio.
7. Most measurements in social sciences (education and psychology included) are possible at nominal, ordinal and interval. Very few important variables in these fields lend themselves to ratio level of measurement.

FURTHER READING

Ingule, F. & Gatumu, H. (1996) *Essentials of Educational Statistics*. Nairobi: E.A. Educational Publishers.

Glass, G.V. & J.C. Stanley (1970) *Statistical Methods in Education and Psychology*. New Jersey: Prentice-Hall.

Johnson, R.R. (1980) *Elementary Statistics.* Mass.: Duxbury Press.

Smith, G.M. (1970) *A Simplified Guide to Statistics for psychology and Education* New York: Holt, Rinehart and Winston.

ACTIVITY

1. Explain why social scientists (e.g. educators, psychologists) need to have at least a rudimentary knowledge of statistics?
2. Distinguish between descriptive statistics and inferential statistics.
3. State the 4 major levels of measurement (scale of measurement) and discuss their characteristics.
4. Define the following terms:
   - i. Variable
   - ii. Continuous variable
   - iii. Discrete variable.
   - iv. Measurement.
5. For each of the following examples, state the highest level of measurement involved.
   - a) The number on the vests of soccer players
   - b) Number of kilogrammes that a sportsman can lift.
   - c) Number of students in a statistics class.

d) Number assigned consecutively to students as they complete an examination consisting of 50 items.

e) Performance on the essay section of a KSCE English test.

LECTURE TWO

TABULATION AND GRAPHICAL REPRESENTATION OF DATA

INTRODUCTION:

Using graphs or tables we should be able to describe the distributions of our data (marks) for our groups of students (subjects or individuals). Graphs tend to give more information than mere distribution tables.

OBJECTIVES:

At the end of the unit the learner should be able to:

1. Indicate when we have:
   - i) Positively skewed distribution
   - ii) Negatively skewed distribution
   - iii) Normal distribution.

2. Draw a graph representing discrete data (histogram).
3. Draw graphs representing continuous data (frequency polygon).
4. Organize data for a group (sample) into:
   - i) Ungrouped frequency distribution
   - ii) Grouped frequency distribution
   - iii) Obtain cumulative frequency (below and above).

Tabulation and Graphical Representation of Data

Tests yield quantitative data often called raw score or raw data. Before raw data can be understood and interpreted, it is necessary to organize and summarize it in some meaningful way. Here we discuss procedures that are commonly used in organizing and summarizing raw data.

These procedures include frequency distribution (ungrouped and grouped), histogram, frequency polygons and ogives. The raw to be summarized may look like the following records of scores for 42 students say in statistics continuous assessment test (CAT):

38	68	39	55	60	61	56
49	51	35	58	48	58	47
65	50	52	39	53	43	42
51	62	47	55	58	54	52
46	65	45	55	46	42	52
34	59	53	48	48	60	50

Some organization of scores is needed here. Two ways are possible for doing this:

- i. Ungrouped frequency distribution
- ii. Grouped frequency distribution

Ungrouped frequency distribution

A frequency distribution is a tabulation of scores (other attribute) of a group of individuals to show the number of times each score occurs. Usually, the first step in construction of a frequency distribution (grouped or ungrouped) is to arrange the scores in order, from the highest to lowest as illustrated below in the two cases for grouped and ungrouped. Basically a frequency distribution tables consists of 3 columns where the first column is for the scores in their order of magnitude. The second column consists of *tally marks* /, representing each score as it occurs. Hence the total number of tally marks for each score corresponds to its frequency, which is the total number of times each score occurs. The frequency for each score is reflected on the third column of the frequency distribution table.

Ungrouped frequency distribution for 42 students in statistics CAT

Mark (score)	Tally mark	Frequency (f_i)	Mark (score)	Tally mark	Frequency (f_i)
68	/	1	50	//	2
67		0	49	/	1
66		0	48	///	3
65	//	2	47	//	2
64		0	46	//	2
63		0	45	/	1
62	/	1	44		0
61	/	1	43	/	1
60	//	2	42	//	2
59	/	1	41		0
58	///	3	40		0
57		0	39	//	2
56	/	1	38	/	1
55	///	3	37		0
54	/	1	36		0
53	//	2	35	/	1
52	///	3	34	/	1
51	//	2			

The table introduces the concept of frequency. As we saw, frequency is the number of times each score (observation) occurs. It is symbolized by the first letter of the word, a lower case, f_i , where the subscript i indicates any given category. The subscript i ranges from 1 to n (total number of scores' groups). Therefore, we can take the frequency of the lowest score in the group, 34, as f_1 , that of 35 as f_2 , for 36 as f_3 etc. The frequency of highest score, 68, can be presented in general form as f_n and consequently that of 67 will be f_{n-1} , that of 66 as f_{n-2} etc. Here we shall use N to represent the total number of cases, in our example N is equal to 42 (i.e. N = 42).

Note that $f_1 + f_2 + f_3 + \ldots\ldots\ldots + f_{n-2} + f_{n-1} + f_n = N$

This is usually written as $\sum_{i=1}^{n} f_i = \text{N}$

i.e. $f_1 + f_2 + f_3 + \ldots\ldots\ldots + f_{n-2} + f_{n-1} + f_n = \sum_{i=1}^{n} f_i$

Grouped frequency distribution

A good way of summarizing data is means of grouped frequency distribution, much so if the data has big range. Of course grouped frequency distribution is not the most appropriate procedure for doing (summarizing) this. A single number would be more appropriate for many reasons.

When there is a wide range of data, the ungrouped frequency distribution may become a cumbersome way of presenting the data. In such a situation, instead of listing each possible data separately, we may condense the data by setting up intervals, which contain a range of possible data. When data are grouped to form intervals of data called “class intervals” the resulting frequency distribution is known as a “grouped frequency distribution”.

Procedures of making grouped frequency distribution

There are four steps in making a grouped frequency distribution. This is shown in the table below using the 42 scores for the statistics CAT.

Class interval	Tally marks	Frequency (f_i)	Midpoint (x_i)
65-69	///	3	67
60-64	////	4	62
55-59	~~////~~ ///	8	57
50-54	~~//// ////~~	10	52
45-49	~~////~~ ////	9	47
40-44	///	3	42
35-39	////	4	37
30-34	/	1	32

1. Determine the inclusive range, between the highest score and lowest.
2. Determine the *size of the class interval* i.e. the difference between successive lower limit or successive upper limits (e.g. for the class interval 65-69, the lower limit is 65 while 69 is the upper limit). And for this, the successive lower limits are 65 and 65, while successive upper limits are 69 and 64, and both give the size of the class as 5 as expected.
3. Determine the last and first class intervals. The class intervals should include the highest and the lowest scores. To achieve this, start each class interval with a multiple of class size, in our case multiple of 5. Observe our lowest score is 34, so we have to start with 30 (multiple of 5) hence our lowest class interval will be 30-34.
   The next class interval starts at 35 ending at 39 i.e. 35-39, next 40-44 etc. We observe the class intervals are mutually exclusive, implying scores fall in only one and only one class interval.
4. A tally mark is made for each score opposite to the class interval to which it falls. To ease counting, tally marks are group in sets of five as shown above where they are five or more observations (scores). The total number of observations for each class interval gives its frequency. These frequencies appear in the third column. In fourth column we have the midpoints also called “class-marks”.

In summary the following are the rules for grouped frequency distributions table construction:

1. Class intervals must remain mutually exclusive. There should be no overlap of the intervals. All scores must fall into one and only one class interval.
2. There should be enough groups (class intervals) to include all the observations i.e. all observations (scores) must be included in the groups.
3. Class intervals (groups) should all be of the same size for any given grouped frequency distribution.

Class-mark:

Is the midpoint of the class interval and is obtained by adding the lower and upper class limits and dividing by two. Other names of the class-mark are midpoint, interval mark, middle mark and category mark. Where there are two middle marks i.e. when the size of the class interval is even, the average of the two is taken.

Real and apparent class limits

You may have noticed that in indicating class intervals, each class has two limits: the lower and upper limits. For class intervals 30 to 34 (30-34) and 35 to 39 (35-39), the lower class limits are 30 and 35, and the upper limits are 34 and 49, respectively. These class limits leave slight gaps between adjacent classes and are often referred to as “apparent class limits” or simply, lower or upper limits, as referred before. Apparent class limits mark the boundaries of the class intervals.

The concept of apparent class limits naturally introduces the related concept of “real”, “actual” or “true” class limits. For instance, the real class limit for the two class intervals considered above are 29.5 and 34.5 being real lower limit and real upper limit respectively for 30-34, and for 35-39 are 34.5 and 39.5 in the same order respectively. The table below gives them all.

Cumulative frequency distribution

For some purposes, it is desirable to present data in somewhat different form. Instead of giving the number of frequency in each class interval, we may indicate the number of scores that are less than or greater than a given value. The number of scores below a given value is referred to as “cumulative frequency below” and the number of scores above the value is referred to as “cumulative frequency above. These are given below in the table.

Table of real limits and cumulative frequencies

Class interval	Real class interval limits Lower-upper	Frequency	Cumulative frequency below	Cumulative frequency above
65-69	64.5-69.5	3	3	42
60-64	59.5-64.5	4	7	39
55-59	54.5-59.5	8	15	35
50-54	49.5-54.5	10	25	27
45-49	44.5-49.5	9	34	17
40-44	39.5-44.5	3	37	8
35-39	34.5-39.5	4	41	5
30-34	29.5-34.5	1	42	1

Note that the cumulative frequency refers to scores attained up to and including those in the interval. If we wish, we can convert actual frequencies into percentages. With percentages, we cumulate up to and including the given interval, until we get to 100 percent.

Graphic representation

The ordinary frequency distribution, however, does not give a very clear picture of the real situation, but can be supplemented with graphic representation of the same data. A graph is an effective method of clarifying points. One small graph often makes a point clearer than a dozen tables or paragraphs. Ordinary numerical data are quite abstract; on the other hand the graph is more concrete representation.

There are three common methods of representing a distribution of scores *graphically:* the *histogram*, the *frequency polygon*, and the *smooth curve* (**ogive**).

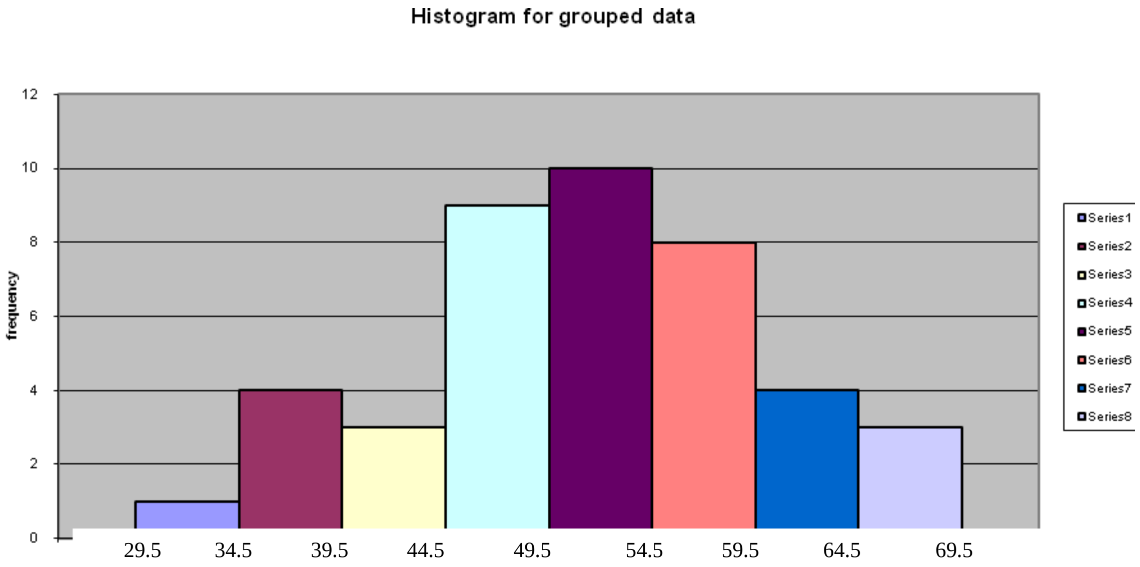
Histogram

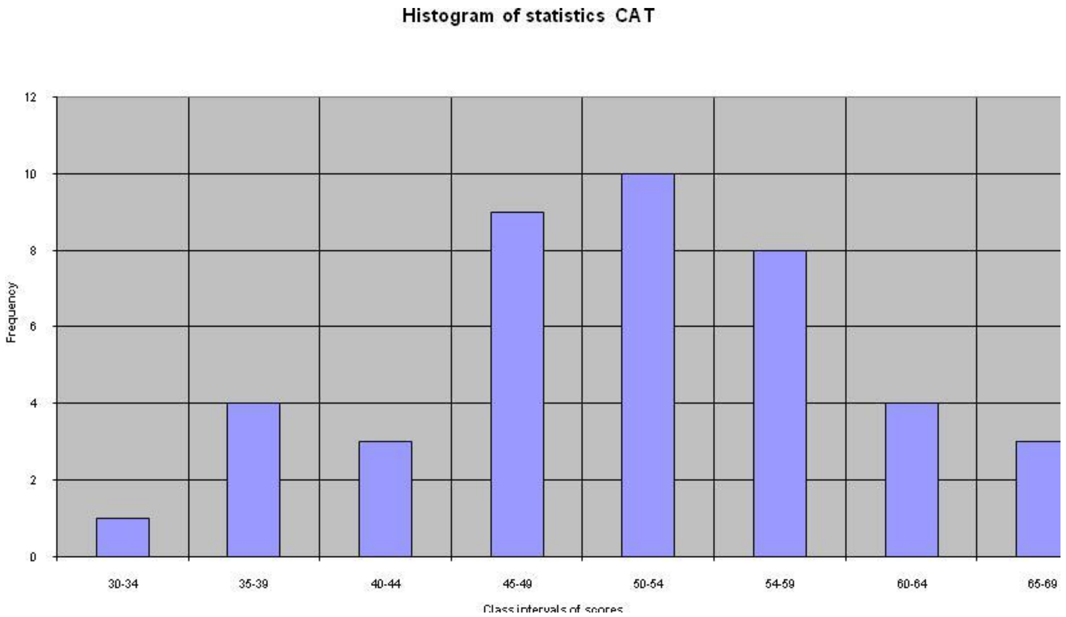
The histogram is a series of columns or bars, each having as its base one class interval and its height the number of cases or frequency in that class. Two perpendicular lines (axes), the horizontal and vertical axes make the external boundaries of the histogram. The horizontal line represents the class intervals and is called *abscissa* or even the X-axis. The vertical line represents frequency and is called the *ordinate* or plainly the Y-axis. To avoid confusion between these two words, remember the mouth stretches horizontally when we say abscissa and stretches vertically when we say ordinate.

Each column of histogram is based on both a class interval and relevant class frequency. The class frequency determines the height of the column and this represented by vertical line corresponding to the class frequency. The class interval corresponds to the width of each column. These widths are marked on the abscissa and with the information from class frequency the relevant column is completed in construction. The columns are usually centred on the midpoint of class intervals.

Generally, we mark off the class intervals on the abscissa using the apparent class limits but sometimes like in the example below, real limits are used. At the abscissa, various class intervals are joined to indicate an interval continuum. However, the data in the histogram is treated as if it is discrete data. An example of a histogram, based on our initial 42 scores from a statistic CAT, is shown below.

In constructing a histogram, an appropriate scale should be selected. It is usually proper to arrange the scales so that the ratio of height to width is approximately 3:5. In other words, for ease of readability the histogram (the graph) should take much of the graph paper as required with any graph. Finally, a descriptive title, which states clearly what is portrayed in the histogram, should be provided as heading.





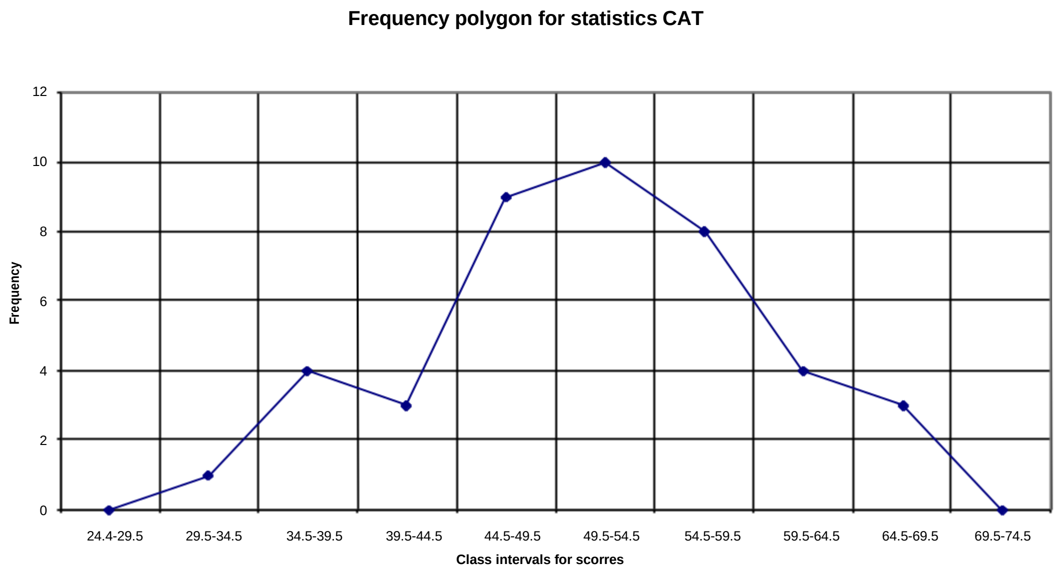
Of some interest is the area under the histogram. Each column of the histogram represents the number of cases (subjects) who have that kind of score with the sum of the frequencies for all the columns. This is N and is expressed in the same units. Thus, the area of histogram is equal to the total frequency and has units of the total frequency.

Frequency polygon

When constructing a frequency polygon, the same axes as the histogram are used. However, instead of erecting a column, a point is located above the midpoint to each interval and at the proper height to represent the frequency in that class. You use the midpoint of the class intervals and the frequencies for plotting the frequency polygon. Straight lines then join these points. The straight lines are extended down to the X-axis one class below and one class above to create a polygon (many sided figure).

As with histograms, frequency polygons may have percentages or relative frequencies plotted on the Y-axis. However, unlike the histogram, the data on the X-axis is treated as continuous data.

Both the histogram and frequency polygon represent the same data, only that the same data is treated in two different ways. One is based on discrete data (in the histogram) and the other is based on continuous data (as in the frequency polygon).



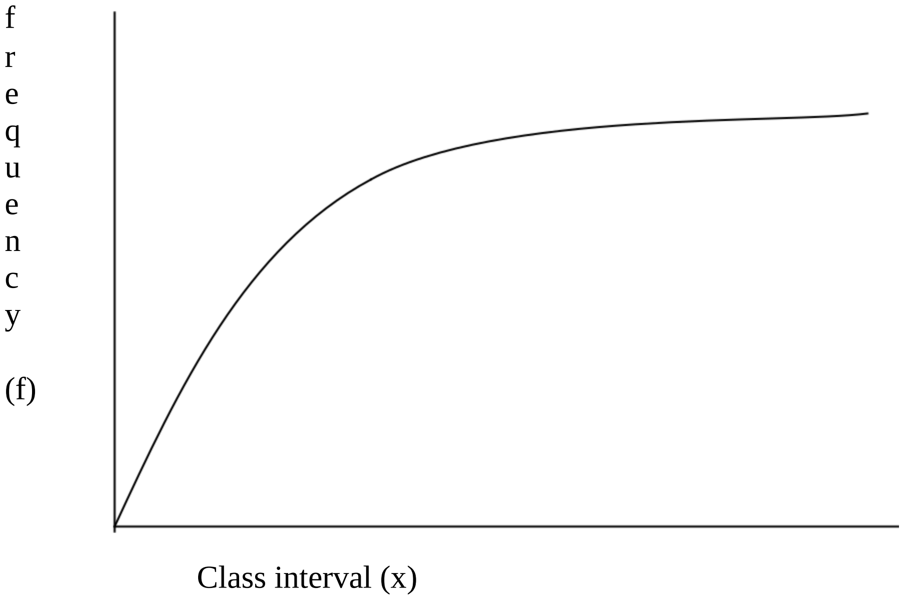
Note

Class-marks used for frequency polygons fall at the middle of the class interval. Also note the frequency polygon has to be closed. This is done by considering an extra class-mark, i.e. the next class-marks on the two ends of the class intervals as illustrated. The frequency of each of these class intervals is 0. The height of each column of the histogram is equal to frequency corresponding to the class interval.

Cumulative frequency polygon

A cumulative frequency polygon (curve) also known as *Ogive* is constructed by plotting the cumulative frequencies against the actual (real) upper limits of the class intervals.

A figure of a cumulative frequency polygon (ogive)

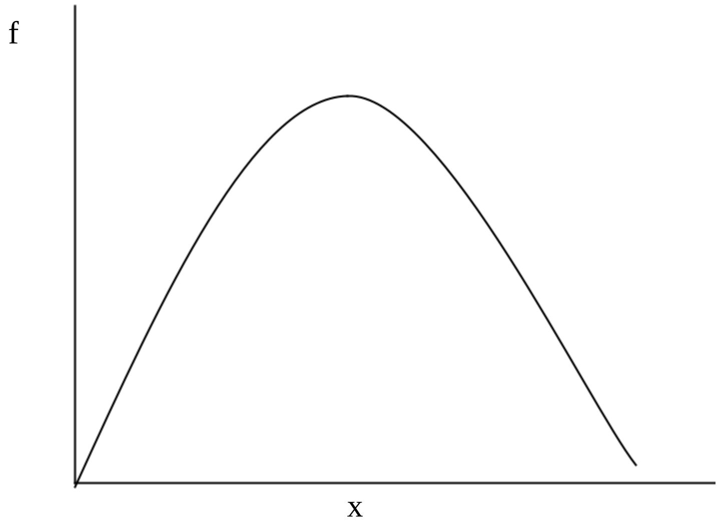


Forms of frequency distributions

Frequency distributions can occur in an unlimited number of shapes and forms. Since form is often an important bit of information about a distribution, it is useful to know some of the descriptive terms used to indicate various types.

Some frequency distributions, and consequently the frequency polygons derived from them, are bilaterally symmetrical. That is, if the polygon is folded in half, perpendicular to the horizontal line, the two sides will be identical in shape or symmetrical. One such distribution is known as the *normal distribution curve* or simply, the normal curve. The figure below illustrates the shape of the normal curve.

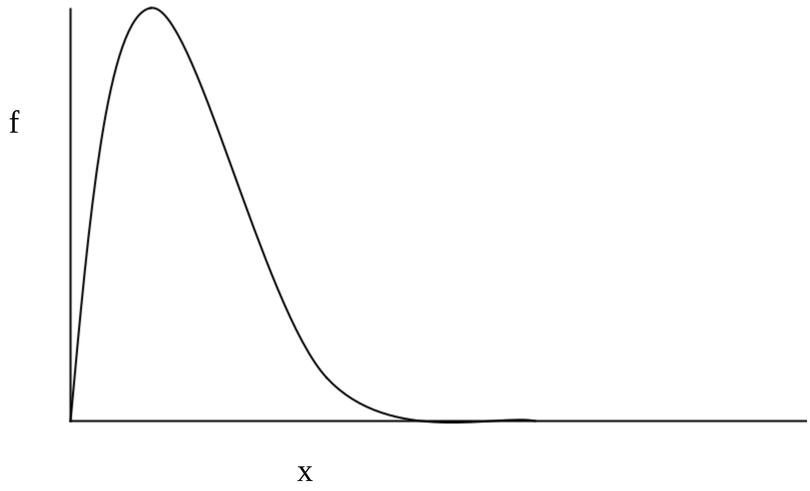
The normal curve



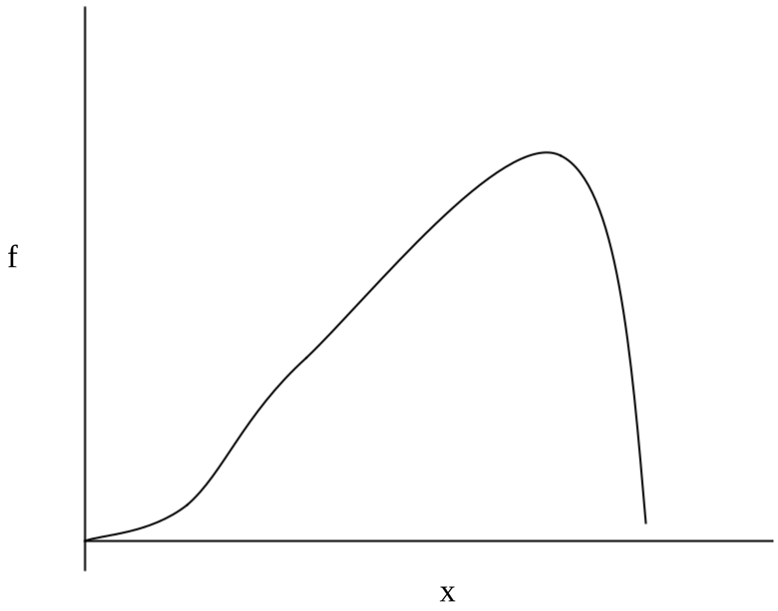
The normal curve is a bell-shaped symmetrical curve with the peak of the distribution at the centre and the ‘tails’ of the distribution continually approaching, but never touching the horizontal axis. This kind of curve is a mathematical concept, which is not realized by any real data, but plays an important role in statistical inference.
Several distributions are not symmetrical but *skewed*. A distribution is skewed if the scores ‘trail off’ in one direction. Such distribution can be more correctly described in terms of degree and direction of skewness. Degree in this context refers to the amount and extent to which a distribution departs from symmetry. Degree of skewness can be determined by comparing a distribution with the normal distribution.

Direction of skewness takes account of the fact that in some asymmetrical (non-symmetrical) distributions, the measures tend to pile up at the lower end of the scale and trail off in terms of frequency toward the upper end, while in other cases, the situation is reversed. Saying the distributions are positively or negatively skewed indicates differences in the direction of skewness. This is both a way of explaining which is positive or is negative, and as a device for remembering the difference between the two. The lower or left end of the horizontal line can be pictured as the negative end (values are low) and the upper or right end as positive (values are high). If the ‘tail’ of distribution, where scores are relatively infrequent in number, lies in the negative end of the scale, the distribution is negatively skewed. If the tail extends out towards the upper or positive end, the distribution is positively skewed.

Positively skewed distribution



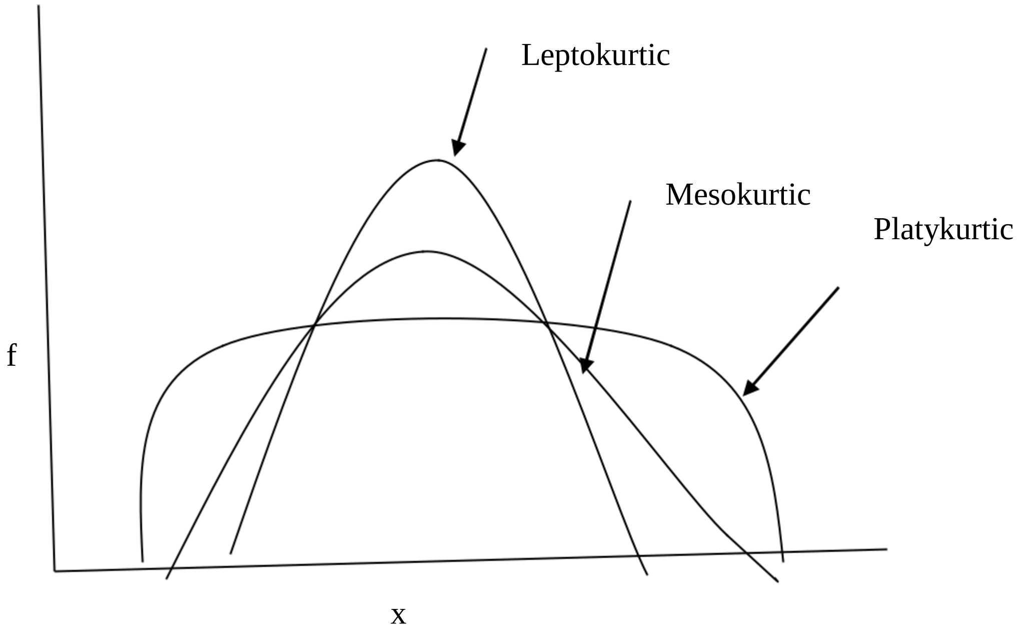
Negatively skewed distribution



Kurtosis

There is another slightly different way of describing the shapes of frequency distributions. This is often discussed in the study of kurtosis. Kurtosis refers to the *flatness* or *peakedness* of a distribution in relation to the normal curve. If one distribution is more peaked than another, it is described as *leptokurtic*. If it is less peaked, it is said to be more *platykurtic*. The normal distribution is spoken of as *mesokurtic*.

A figure showing forms of kurtosis



SUMMARY

The principal ideas, implications and conclusions of this unit are summarized in the following statements:

1. Before raw data can be understood and interpreted, it is usually necessary to organize and summarize them in some meaningful way.
2. Procedures used to organize and summarize data include frequency distributions, histograms, frequency polygons and ogives.
3. A frequency distribution is a tabulation of scores (or other attributes) of a group of individuals to show the number of times each score occurs.
4. When dealing with a large number of scores, we use a grouped frequency distribution in, which scores are grouped to form intervals of scores called 'class intervals'.

5. In cumulative frequency distribution, we indicate the number of scores that are less or greater than a given value.
6. A graph is a very effective method of representing data.
7. Three common methods of representing a distribution graphically are histogram, the frequency polygon and the smooth curve (ogive).
8. The histogram is a series of columns or bars each having as its base one class interval and its height the number of cases or frequency in that class.
9. Frequency polygons are similar to histograms but instead of columns, the midpoints at the appropriate frequency of each class interval are joined by straight lines. The straight lines are extended down to the vertical (X-axis) one class above and one class below to create a many sided figure (polygon).
10. An ogive is a cumulative frequency polygon and is constructed by plotting the cumulative frequencies against the actual (real) upper limits of the class intervals.
11. Various forms of frequency distributions exist and these include the normal distribution, negatively skewed distributions and positively skewed distributions.

FURTHER READING

Ingule, F. & Gatumu, H. (1996) *Essentials of Educational Statistics*. Nairobi: E.A. Educational Publishers.

Glass, G.V. & J.C. Stanley (1970) *Statistical Methods in Education and Psychology*. New Jersey: Prentice-Hall,

Johnson, R.R. (1980) *Elementary Statistics.* Mass.: Duxbury Press,

Smith, G.M. (1970) *A Simplified Guide to Statistics for psychology and Education* New York : Holt, Rinehart and Winston,

ACTIVITY

1. The following were the scores obtained by a form ii class in a mathematics test:

   49 63 59 44 49 51 62 37 30 49 45 52 50 42 54 32 57
   41 42 56 44 46 63 44 40 50 46 53 48 37 46 53 68 36
   40 56 37 66 43 40 43 51 59 42 52 46 57

   (a) Make a frequency distribution table for this data. The table should show both tally marks and frequencies. The total frequency (N) = 50.

   (b) Make a grouped frequency distribution that should have both tally marks and frequencies for each class interval. Use class size (i) = 5 and start with 30-34 as the lowest class interval. Indicate the class-mark and actual limits for each class interval. Indicate also the above as well as below cumulative frequencies.

   (c) Plot (on graph paper) a histogram and frequency polygon for this data. Note that the frequency polygon should be closed by extending the lines to X-axis as emphasized in the text.

   (d) Comment on the distribution of the scores (i.e. is their distribution close to normal or are they skewed positively or negatively?).

2. Using the same data, repeat 1 b and c but now using class size i = 4 and starting with 30-33 as the lowest class.

3. Select 30 of these scores randomly (one of best ways of selecting them randomly is to write each score on small piece of paper. Put all these 50 folded pieces of paper in a box and pick 30 after mixing all thoroughly). Using these 30 scores repeat no. 1 (a), (b) and (c).
   How does the distribution of the scores compare with the original distribution i.e. the distribution of the 50 scores?

4. Distinguish between the following terms using illustration if appropriate, positively skewed and negatively skewed distributions.

LECTURE THREE

MEASURES OF CENTRAL TENDENCY

INTRODUCTION

We need to have concise ways of presenting (summarizing) information (data) rather than by means of graphs or tables, which may require more space. Single numbers (indexes), which show the general level of performance are more convenient and can even be given verbally hence not taking much space and time. These indexes are referred to as **measures of central tendency** (or commonly called **average**).

OBJECTIVE

At the end of the unit, the learner should be able to:

1. Describe the 3 measures of central tendency (mode, median, mean)
2. Compute:
   - i) Mode for ungrouped and grouped data.
   - ii) Median for ungrouped and grouped data
   - iii) Mean for ungrouped and grouped data.
3. Describe unimodal (having one mode) distributions in terms of positively skewed, negatively skewed and normal using the three measures of central tendency.
4. Discuss the properties of the mean (e.g. what happens to the mean when a constant is added to all the scores in the distribution).
5. Compute mean using assumed mean method.

Measures of Central Tendency

An examination of a graph of a frequency distribution of a given variable reveals two things: firstly, the values of the variable tend to cluster around a central value; secondly, they spread around that value in a specific way. Describing the central points around which value in a distribution spread is what we mean by a measure of central tendency. Measures of central tendency give some idea of the average or typical or representative scores in the distributions. Three measures of central tendency are the *mode*, *median* and *mean*.

The mode is the most frequent vale or score in a distribution, the median gives information about the value of the middle position in a distribution, and the mean indicates a point around which the values of a distribution balance. The mode is generally denoted by Mo or just the word mode. The median is abbreviated Md, and the mean is indicated by a bar (-) over the letter representing to variable. The letters most often used are capital X or Y. Thus, the mean of X (or X-variable) is $\bar{X}$ and the mean of Y (or Y-variable) is $\bar{Y}$.

The mode

The mode is the most frequent occurring score or value in a distribution. It is the score or value with the highest frequency in a distribution. Note that the mode is a score value, which occurs most frequently, it is one of the scores in the distribution. However, the score may exist or not. Thus the following questions have negative answers:

Does the mode have to exist? And if it exists, must be unique? The answer to these two questions is No as seen above. Let us illustrate how to find the mode with the following examples:

1. When we have all the scores in the distribution having the same frequency, no mode exists. For example, no mode exists for the following set of numbers (scores): 1, 3, 4, 6, 8 and 9.
2. When we have one score with higher frequency than others in the distribution, then the score is the mode. For scores, 1, 3, 4, 4, 6, 8, 9, the mode is 4.
3. When two adjacent scores have the same frequency and this frequency is the highest in the distribution, the mode is the average of the two modes. In a distribution consisting of 1, 3, 4, 4, 6, 6, 8, 9, the mode is the average of the two modes, 4, and 6, which is 5 obtained by adding 4 and 6 and dividing the result by 2.
4. If the modes are non-adjacent, then we have multiple modes. The two or more are reported without finding average. If the non-adjacent modal scores are two in number, the distribution is said to be *bimodal*. In the case of the following set of scores 1, 3, 3, 4, 6, 8, 8, 9, the non-adjacent modal scores 3 and 8 occur with the same frequency, which is greater than for any other score. These two sets of scores are then said to be bimodal because they have two non-adjacent modal scores. A bimodal distribution from a large set of scores has frequency polygon that resembles a two-humped *(bactrian)* camel's back, even though the frequencies at the peaks may not be exactly equal.

For grouped data, the class-mark (midpoint) of the class interval with the highest frequency is taken as the mode. For the data in the table for grouped frequency distribution, which is below, is in the class interval, 50-54. The mode is the class-mark of this class interval and is referred to as *modal interval*. Thus the mode here is 52, while modal interval is 50-54.

Class interval	Frequency (f_i)
65-69	3
60-64	4
55-59	8
50-54	10
45-49	9
40-44	3
35-39	4
30-34	1

It should be noted that if our distribution has multiple modes and two or more are adjacent to each other, then the mode is the average of the adjacent modes. Otherwise if the modes are non-adjacent then, they are simply reported as the modes and no averaging is be done.

As can be seen above, the mode is very easy to calculate; in fact, it is the easiest measure of central tendency to calculate. In a large sample, the mode is stable and its value is fairly close to values of the other measures of central tendency: the mean and the median. But for smaller samples, it fluctuates considerably and should only be used when a quick estimate is needed or when there is need to identify the most common score.

The Median

The median is a point in a distribution of scores such that 50 percent (or ½) of the scores are located above it and other 50 percent (or ½) below it. In other words the median is the midpoint of a score distribution.

If there is an odd number of scores like 5, 6, 9, 10, 11, the median is the middle score in the distribution. This means that before getting median, the scores have to be first ranked, from the highest to the lowest or lowest to highest. In our example, the scores have already been ranked and consequently, the middle score is 9 and this is the median.

For an even numbered (e.g. N = 4) scores, the median is the point half between the two middle scores when the scores are ranked. For the set of scores 5, 7, 8, 9, 12, 13 (N is even numbered) the median lies between the two middle scores, 8 and 9. The median is found by calculating the average of 8 and 9, which is 8.5. This value is halfway between the two middle scores.

Where the middle score is repeated such as 1, 3, 3, 4, 6, 6, 7, 8, 9, the median needs some computation. Note that 6 is not the exact median because, by definition, the median is supposed to be a point which divides the distribution in two equal halves. Since in this case, N (total number of cases) is 9, the median should have N/2 scores below it and N/2 scores above it. Therefore, we shall have to do interpolation to get the median.

Since there are 9 cases in the distribution, we want the point below which and above which, there will be 4.5 cases (i.e. ½ of 9). 4 cases fall below 5.5 (i.e. actual lower limit of score 6). ½ or .5 case is needed then we get to the median from this actual lower limit. There 2 cases (i.e. two 6's) on interval 5.5-6.6. We have to assume the two cases are equally distributed on the interval of size 1. Hence ½ case occupies ¼ of the interval (i.e. .25). And since we need ½ of case to get to median from actual lower limit of the interval, we observe that this will land us to 5.75 (i.e. 5.5 + .25). Exactly half of the cases (4.5 cases) fall below 5.75 in this distribution. Note median is a point, which splits cases into two halves such that 50% are below it (median) and 50% are above it. Note there are two things to consider in calculation of the median, the number line (or values or points) and then the cases.

Median for grouped data

To get the median for grouped data, some interpolation must also be performed. The grouped data that was used to illustrate the calculation of the mode above will be used to illustrate the calculation of the median. The following steps are generally used.

Class interval	Frequency (f_i)	Cumulative frequency below
65-69	3	42
60-64	4	39
55-59	8	35
50-54	10	27
45-49	9	17
40-44	3	8
35-39	4	5
30-34	1	1

1. First, determine the number of observations in the distribution by computing the cumulative frequencies.
2. Secondly, establish the halfway point of the distribution N/2. In our example, it is 42/2 or 21.
3. Identify the median interval in which the middle score falls and determine its actual limits. In this case, the median class interval, in which the middle score falls is 50-54 and actual limits are 49.5 and 54.5.
4. To find the median, we interpolate between the actual intervals limits to find a score above and below which 21 cases fall. Notice that the class size is 5 and 10 scores fall within the median interval. This means that each of the 10 scores occupies 5/10 or .5 of the interval. Since 17 scores fall below the median class interval, to obtain the median, we require 4 cases of the 10 to reach the middle value (17 + 4 =21). The proportion of the interval needed to account for 4 scores is 4 x 0.5, which is 2 units. We add this to the actual lower limit of the interval to obtain the median, which is 49.5 + 2 = 51.5. Thus median for our grouped data above is 51.5.

The formula for getting the median for grouped frequency distribution is

$$\text{Median} = L + \frac{(\frac{N}{2} - Cf_b)i}{f_w}$$

Where N is the total number of scores in the distribution.

L is the actual or real limit of the class interval containing the median.

Cf_b is the total frequency in all the intervals below the median class interval.

f_w is the frequency of the class interval containing the median.

i is the size of the class interval.

Thus, for our example of the grouped frequency distribution, we can use formula in the following way:

$$\text{Median} = L + \frac{(\frac{N}{2} - Cf_b)i}{f_w}$$

$$= 49.5 + \frac{(\frac{42}{2} - 17)5}{10}$$

$$= 49.5 + \frac{(21-17)5}{10}$$

$= 49.5 + 4 \times 5/10$

$= 49.5 + 20/10$

$= 49.5 + 2$

= 51.5

This is the same result as the one calculated before.

The median is a position average i.e. it is determined by placing the scores in rank order and finding the middle point. Thus the median is a position average that divides the distribution into two equal halves such that one half is below it and the other half above it (median).

The Mean

It is the best known and the most reliable measure of central tendency. Hence it is often preferred to both the median and the mode. Every score in the distribution is considered in its computation. This is not done when calculating the mode and the median.

Thus mean is simply found by adding all the scores in a distribution and dividing by the total number of scores (N). It is denoted by $\overline{X}$ pronounced X bar.

The formula is:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

Where $\overline{X}$ = mean

X is the raw score for each individual i.e. the i th person's score.

N is the number of scores.

$\sum_{i=1}^{n}$ is the summation sign indicating we are summing from the first score to the n th score i.e. all the X-scores in the distribution are added.

Example

Find the mean of 3, 3, 4, 5, 6, 6, 8, 9 and10. The sum of these is 3 + 3 + 4 + 5 + 6 + 6 + 8 +9+10=54 N

= 9

Therefore, the mean is 54/9 = 6.

For grouped frequency distribution, the mean is obtained as follows:

Each class-mark or mid-point (x_i) is multiplied by its corresponding frequency (f_i). The products are then summed and divided by total frequency to give the mean.

This can be summarized as follows:

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

where x_i refers to the class-marks or mid-points and f to the corresponding frequencies. The calculation of the mean of grouped frequency distribution is illustrated below:

Class interval	Frequency (f_i)	Class-mark (x_i)	$f_i x_i$
65-69	3	67	201
60-64	4	62	248
55-59	8	57	456
50-54	10	52	520
45-49	9	47	423
40-44	3	42	126
35-39	4	37	148
30-34	1	32	32
	$\sum_{i=1}^{n} f_i = 42$		$\sum_{i=1}^{n} f_i x_i = 2154$

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{2154}{42} = 51.3$$

Properties of the mean

1. One important property of the mean is that is that it is the point in a distribution of scores such that the summed deviations of scores from it (the mean) are equal to zero. What do we mean by deviation? Deviation is the difference between a score and the mean, $X_i - \bar{X}$, and it can be either positive or negative. In any distribution the sum of deviations about the mean is always equal to zero.

   i.e. $\sum_{i=1}^{N}(X_i - \mu) = 0$ where μ is the population mean for X-scores and population has N subjects.

   $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$ where $\bar{X}$ is the sample mean and sample is of size n.

For illustration, let us consider the following scores. Suppose our scores are 3, 3, 4, 5, 6, 6, 8, 9 and 10 (note this can be considered as a population or a sample without any change of the results). The mean will be 6 and the deviation scores will be 3-6, 3-6, 4-6, 5-6, 6-6, 6-6, 8-6, 9-6 and 10-6, in general $X_i - \bar{X}$. These deviations will be respectively -3, -3, -2, -1, 0, 0, 2, 3 and 4 (note their sum is zero). Thus, the mean may be considered as the exact balance point in a distribution.

2. If we add a constant, say C to every score in the distribution, the resulting scores will have a mean $\bar{X}_{X+C}$ equal to the original mean $\bar{X}_X$ plus the constant C. If we subtract a constant instead, the resulting scores will a mean equal to original mean minus the constant. Note that subtracting a constant C is the same as adding –C. Hence the first formula is adequate or it includes even the second formula.

i.e. $\bar{X}_{X+C} = \bar{X}_X + C$

$\overline{X}_{X-C} = \overline{X}_X$ - C

Let us illustrate this with data below, of which the mean is 5, but let us add 3 to every score and calculate the new mean:

X	$X_i + C$
3	3+3=6
4	4+3=7
5	5+3=8
8	8+3 = 11
$\sum_{i=1}^{n} X_i = 20$	$\sum_{i=1}^{n} X_i + C = 32$
$\overline{X}_X$ =20/4=5	
Thus $\overline{X}_X = 5$	$\overline{X}_{X+C} = \frac{\sum_{i=1}^{n}(X_i+3)}{4}$ =32/4=8
	$\overline{X}_{X+C} = \overline{X}_X$ + 3 or $\overline{X}_X$ +C

3. If each score in a distribution of scores is multiplied by a constant C, then the mean of these scores will be original mean multiplied by the constant i.e. C $\overline{X}_X$. Let us illustrate this property with the data that was used above. Let all the scores be multiplied by 2 (see the table below).

X	CX_i
3	3×2=6
4	4× 2 = 8
5	5× 2 = 10
8	8× 2 = 16
$\sum_{i=1}^{n} X_i = 20$	$\sum_{i=1}^{n} CX_i = 40$
$\overline{X}_X$ =20/4 = 5	
Thus $\overline{X}_X$ = 5	$\overline{X}_{CX} = \frac{\sum_{i=1}^{n} CX_i}{4}$ = 40/4 = 10
	$\overline{X}_{CX} = 2\overline{X}_X$ or C$\overline{X}_X$.

Note that the division is reciprocal of multiplication. Hence if we divide each score by a constant C, the mean of the resulting scores is original mean $\overline{X}_X$ divide by constant, C (or $\frac{1}{C} \times \overline{X}_X$).

New scale values

The above properties of the mean can be used to greatly simplify the computation of the mean from grouped and ungrouped frequency distributions. Let us illustrate this with data for grouped distribution obtained and used above, whose class-marks (midpoints) and frequencies are given in the first and second columns respectively

Class-mark (x_i)	Class-mark subtract 52 ($x_i - 52$)	Class-mark subtract 52 then divide by 5 i.e. x_i'	Frequency (f_i)	$f_i x_i'$
67	15	3	3	9
62	10	2	4	8
57	5	1	8	8
52	0	0	10	0
47	-5	-1	9	-9
42	-10	-2	3	-6
37	-15	-3	4	-12
32	-20	-4	1	-4
			$\sum_{i=1}^{n} f_i = 42$	$\sum_{i=1}^{n} f_i x_i' = -6$

The demanding way of obtaining the mean for grouped data is the product of the class-marks and their corresponding frequencies that are summed up and the total is divided by total frequency. We can use the various properties of the mean outlined above to simplify this process in the following way. The table above is used for the discussion:

1. List the class-marks of the distribution (as shown in column 1).
2. Utilization of the properties of the mean can be done in two steps to create new scale values.
   - (a) Subtract a reasonable but arbitrarily determined value from each of the class-marks. In the table above, the constant is 52. So 52 has been subtracted from each class-mark (x_i) and the results are all shown in column two ($x_i - 52$).
   - (b) Divide the above values by a constant. The constant used in the table is 5. The results of the two steps above give new values by subtracting 52 from each class-mark and dividing this value by 5. The values are given in the third column. These values are symbolized as x_i', the new values for the class-marks. The values in the new scale (column 3) are multiplied by corresponding frequencies (shown in column 4) to get entries for the fifth column labeled $f_i x_i'$.

The new mean based on the new scale values is the sum of the fifth column i.e. –6 divided by the total number of cases, 42 i.e. –6/42.

Therefore the required mean for grouped frequency distribution is:

$$\overline{X} = 52 - 6/42 \times 5$$
$$= 52 - 5/7$$

$= 51\frac{2}{7}$

= 51.3 (note the same value was obtained by longer method)

Least sum of squared deviations

Another important property of the mean concerns the sum of squared deviations. The sum of squared deviations from the mean will be less than the sum of deviations from any other point. If we subtract the mean from each value in the distribution, then squaring these differences and summing these values, the sum so obtained (i.e. the sum of the squares of deviations from the mean) will be less (minimum) than the sum of squared deviations from any other value. That is, $\sum_{i=1}^{n}(X_i - \bar{X})^2$ yields the minimum sum of squared deviations.

For example, in the table below, the sum of squared deviations $\sum_{i=1}^{n}(X_i - \bar{X})^2$ is 52 while taking another value say, 7 a similar sum of $\sum_{i=1}^{n}(X_i - 7)^2$ is 61. So you see 52 is less than 61, implying whichever value you choose, whether greater or less than 6, will give the sum squared deviations greater than 52.

The fact that this sum is the least implies that the mean is the best guess of the values in the distribution using this “least squares” criteria. The overall error (squared deviation) from the mean as a guess for all scores will be less than when using any other values in the distribution.

X	$(X_i - 6)^2$	$(X_i - 7)^2$
3	9	16
3	9	16
4	4	9
5	1	4
6	0	1
6	0	1
8	4	1
9	9	4
10	16	9
Sums	52	61
	i.e. $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 52$	i.e. $\sum_{i=1}^{n}(X_i - 7)^2 = 61$

$\sum_{i=1}(X_{in} - \bar{X}) \leq \sum_{i=1}^{n}(X_i - A)_2$ where $A \neq \bar{X}$

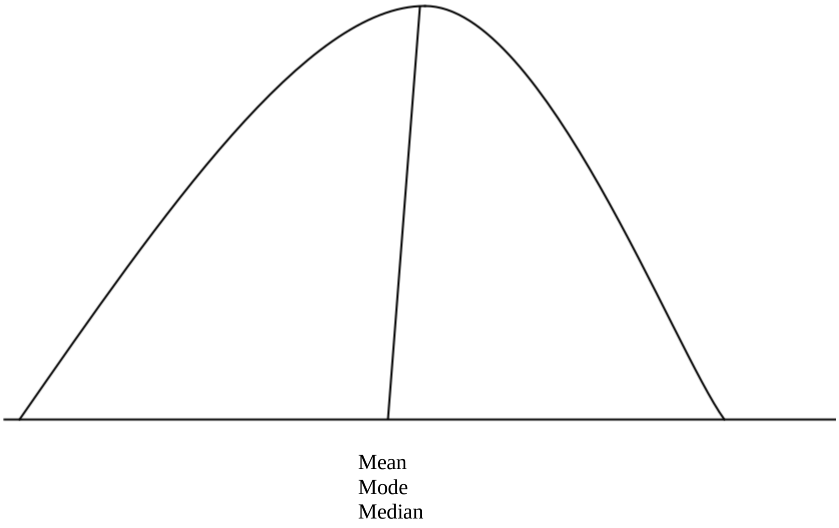
Thus the mean can be further described as the central value about which the sum of the squared deviations is a minimum.

Looking at magnitudes and positions of the mean, median and mode in frequency distributions curves can be very informative. The curves to be considered here are those describing normal distributions, positively skewed distributions and negatively skewed distributions.

Mean, Median and Mode Compared

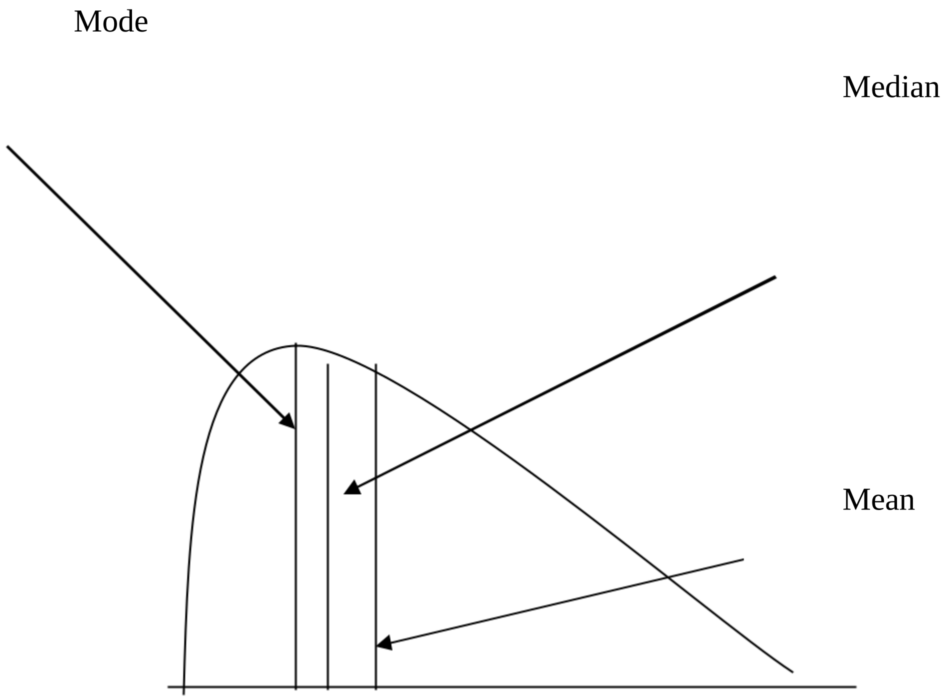
In the normal distribution, the mean, median and mode are all equal. The three measures of central tendency are all located exactly at the centre of a normal distribution curve as shown in the figure below:

Normal distribution curve:



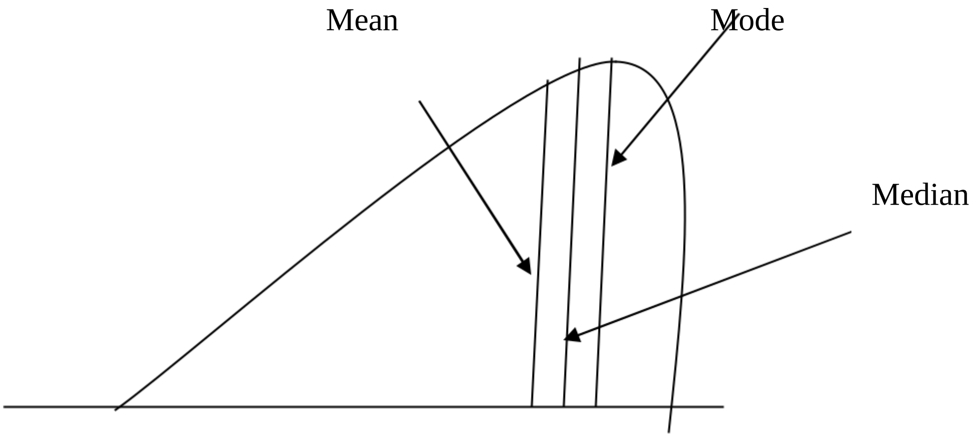
In a positively skewed distribution, the mean is greater than the median and the median is greater than the mode. For our example, this may mean that most students obtained low marks while there were extremely few students who got high marks, a situation normally found when a test is too hard. The positions of these measures of central tendency on a positively skewed curve is shown below:

Positively skewed distribution



Mean 〉Median 〉 Mode

In a negatively skewed distribution, the mode is greater than the median and the median is greater than the mean. This illustrates a situation where many students have obtained high marks while very few students have got low marks. This may occur if the test was too easy for most students.



Mean < Median < Mode

Negatively skewed distribution

In both negatively and positively distributions, it is easy to locate the mode. The mode is the score with the highest frequency, and therefore it appears at the peak of the curve. To locate the mean, we have to realize that extreme scores affect it. In a positively skewed distribution where there are few high scores, the mean will be biggest measure of central tendency. In a negatively skewed distribution where there are few low scores, the mean will be the lowest measure of central tendency, and it is always between the mode and the mean in any skewed distribution.

For the example above for group frequency distribution, we had the following as the obtained values for the measure of central tendency:
Mode = 52
Median = 51.5
Mean = 51.3

So we have a distribution that is negatively skewed, since the mean is less than the median, and it (median) is less than mode. Otherwise, it can be described as being close to normal since the values are very close to being equal. Or in other words, we can say the distribution is very slightly negatively skewed.

The combined mean (or overall mean)

Suppose there are 3 groups, which are combined, and we wish to find the overall mean once the groups are combined. Suppose further the following were the statistics of the groups:

$n_1 = 20 \quad \bar{X}_1 = 12 \qquad n_2 = 15 \quad \bar{X}_2 = 10 \qquad n_3 = 12 \quad \bar{X}_3 = 8$

$$\text{Overall mean } (\bar{X}_c) = \frac{n_1 \times \bar{X}_1 + n_2 \times \bar{X}_2 + n_3 \times \bar{X}_3}{n_1 + n_2 + n_3}$$

$$= \frac{20 \times 12 + 15 \times 10 + 12 \times 8}{20 + 15 + 12}$$

$$= \frac{240 + 150 + 96}{47}$$

$$= \frac{486}{47}$$

$$= 10.34$$

The general formula for obtaining the overall mean (or combined mean) is:

$$\bar{X}_c = \frac{n_1 \times \bar{X}_1 + n_2 \times \bar{X}_2 + \ldots\ldots\ldots\ldots + n_a \times \bar{X}_a}{n_1 + n_2 + \ldots\ldots + n_a}$$

$= \frac{\sum_{i=1} n_i \times \bar{X}_{ia}}{\sum_{i=1}^{a} n_i}$ Where we have a groups each of size n_i and mean $\bar{X}_i$ respectively.

SUMMARY

The following statements summarize the major points in this unit:

1. The mean, median and mode are measures of central tendency. They give an idea of the average or typical score in a distribution.
2. The mode is the most frequent score value in a distribution. It is the score with highest frequency in a distribution.
3. The median is a point in a distribution of scores such that 50 percent of scores are located above it and the other.
4. The mean is found by adding all the scores in a distribution and dividing by the total number of scores.
5. One important property of the mean is that it is the point in a distribution of scores such that the summed deviations of the scores from it (the mean) is zero.
6. The sum of squared deviations from the mean is less than the sum of squared deviations from any other point.
7. The mean is generally preferred by statisticians as the measure of central tendency, but the median is quite often easier to compute, and therefore preferred by classroom teachers in considerable number of cases.
8. For distribution that are fairly normal, it matters little which measure of central tendency is used.

FURTHER READING

Ingule, F. & Gatumu, H. (1996) *Essentials of Educational Statistics*. Nairobi: E.A. Educational Publishers.

Glass, G.V. & J.C. Stanley (1970) *Statistical Methods in Education and Psychology*. New Jersey: Prentice-Hall,

Johnson, R.R. (1980) *Elementary Statistics.* Mass.: Duxbury Press,

Smith, G.M. (1970) *A Simplified Guide to Statistics for psychology and Education* New York : Holt, Rinehart and Winston,

ACTIVITY

1. Using the data used before in unit 2 and the data was as below:

49 63 59 44 49 51 62 37 30 49 45 52 50 42 54 32 57
41 42 56 44 46 63 44 40 50 46 53 48 37 46 53 68 36
40 56 37 66 43 40 43 51 59 42 52 46 57

a) Compute the mean, median and mode for the ungrouped data (the ungrouped data was obtained in the earlier exercise).

b) Compute the mean, median, modal interval and mode for grouped data also found earlier.

c) In terms of the magnitude of mean, median and mode, comment on the distribution of these scores.

2. Compute the mean, median, modal interval and mode for the grouped data but now using the size of the class interval as 3, and start with 30-32 as the lowest class interval as done earlier.

3. Select 30 of these scores randomly and repeat 1 a) and b) but this time round use assumed mean method to obtain both means.

LECTURE FOUR

MEASURES OF VARIABILITY

INTRODUCTION

The measures or indexes considered here are range, quartile deviation, mean deviation, variance and standard deviation. Range is the simplest measure of variability (or dispersion). Standard deviation is the most reliable measure of variability and *standard deviation* is *the square root of variance*; that is, ***variance*** is ***standard deviation squared***.

OBJECTIVES

At the end of the unit the learner should be able to:

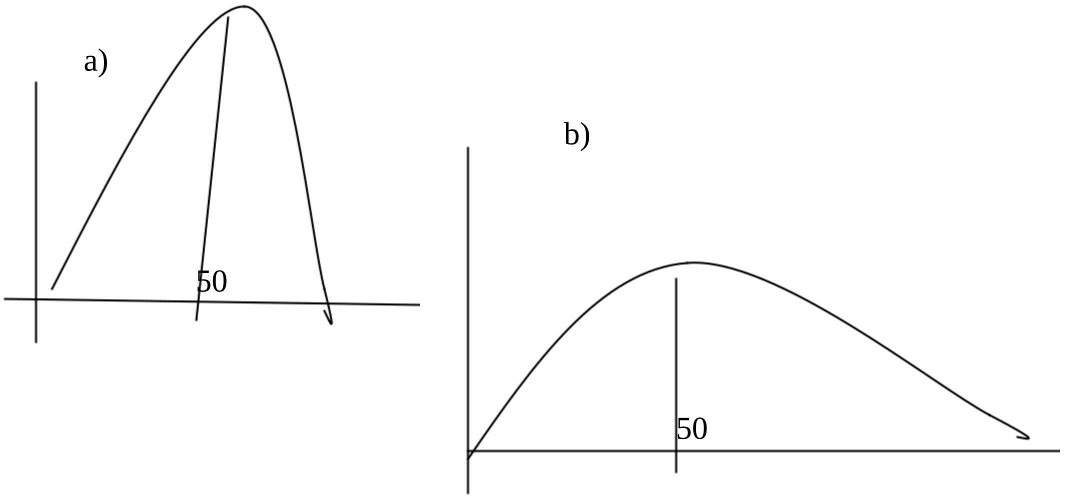
1. Compute range, quartile deviation, mean deviation, variance and standard deviation for grouped and ungrouped data using computational and definitional formulae.
2. Give the properties of variance and standard deviation (s.d.) (e.g. when a constant is added to all the scores of the distribution)
3. Compute variance and s.d. using assumed mean method.

4. Interpret computed s.d.

Measures of Variability

A variable can be defined as a trait, which can take on a range of values. When we talk about variation, we refer to the arrangement or spread of values that the variable takes in the distribution. Measures of variation give us some information about the difference between scores. While measures of central tendency give information about typical score in a distribution, measures of variability provide information about the differences in spread between scores in the distribution.

When we try to describe a distribution, giving the measure of central tendency is not enough. We need to know something about variability or spread of the scores in a distribution. In fact, distributions may have the same mean, yet differ in the extent of variation of the scores around that measure of central tendency.



Two scores distributions with same mean but differing in variability a) small variability and b) large variability.

What we are saying is that to describe a distribution, a measure of central tendency is not enough (sufficient), even if you give the most reliable of them. May be the mean is necessary but not sufficient. To describe any distribution of variables (e.g. scores) three elements are important considerations:

(i) Measure of central tendency
(ii) Measure of variability, and
(iii) Shape of the distribution.

These three elements are necessary and sufficient to describe any distribution of variables of a sample or population. Thus, in order to adequately describe a distribution of scores (variables), we need, in addition to a measure of central tendency, a measure of variability of variables besides the shape of the distribution. Information concerning variability is as important, if not more important than information concerning central tendency.

Types of Measures of Variability

The measures of variability (spread or dispersion) provide a needed index of the extent of variation among variable (scores) in a distribution. Indexes used are:

1. Range
2. Quartile deviation
3. Mean deviation
4. Variance and standard deviation.

These measures of variability along measures of central tendency make up the two types of descriptive statistics, which are dispensable in describing distributions of a variable though not sufficient, for we need the shape of the distribution to make the description sufficient.

Range

The range is defined as the difference between the highest score and the lowest score. For example for the data or scores 3, 3, 4, 5, 6, 6, 8, 9, 10 the range is simply 10-3, which is 7. When dealing with grouped data, an estimate of the range can be computed by subtracting the midpoint (class-mark) of the lowest interval from the class-mark of the highest interval.

The range is the simplest measure of variability. Hence it is not a stable measure of variability. Thus, the range is only used as a quick reference to the spread of scores in a distribution.

Quartile deviation

The quartile deviation as called "semi-interquartile range" is defined as half of the difference between the 75th percentile (Q_3) and the 25th percentile (Q_1). Hence it is one half the scale distance between the 75th and 25th percentiles in a frequency distribution. To find the quartile deviation (Q), which includes the middle 50th percentile or Q_2 (or median), we fist locate the 75th percentile and the 25th percentile. 75th percentile is a point in a distribution such that a quarter of distribution is above it and other three quarters below it. Similarly, 25th percentile has a quarter below it while the other three quarters above it, while 50th percentile or median has half of the distribution above and below it. Thus the formula for calculation of quartile deviation, Q, is:

Quartile deviation (Q) $\frac{Q_3 - Q_1}{2}$ or $\frac{P_{75} - P_{25}}{2}$

= Example

Class interval	Frequency (f_i)	Cumulative frequency
65-69	3	42
60-64	4	39
55-59	8	35
50-54	10	27
45-49	9	17
40-44	3	8
35-39	4	5
30-34	1	1

$$\sum_{i=1}^{n} f_i = 42$$

$$Q_3 = P_{75} = L + \frac{(\frac{3N}{4} - Cf_b)i}{f_w}$$

$$= 54.5 + \frac{(\frac{3 \times 42}{4} - 27)5}{8}$$

$$= 54.5 + 2.8$$

$$= 57.3$$

$$Q_1 = P_{25} = L + \frac{(\frac{N}{4} - Cf_b)i}{f_w}$$

$$= 44.5 + \frac{(\frac{42}{4} - 8)5}{9}$$

$$= 44.5 + 1.4$$

$$= 45.9$$

Hence Quartile deviation (Q) = $\frac{Q_3 - Q_1}{2}$ or $\frac{P_{75} - P_{25}}{2}$

$$= \frac{57.3 - 45.9}{2}$$

$$= \frac{11.4}{2}$$

$$= 5.7$$

Note, the same procedure as used for median earlier has been used here to find the 75th and 25th percentiles or Q_1 and Q_3.

Mean deviation; Variance and Standard Deviation

You will find that the two measures of dispersion (variability) discussed above i.e., range and quartile deviation do not take into consideration explicitly, the values of each and every one of the raw scores in the distribution. To arrive at more reliable indicator of the variability (or spread or dispersion) in a distribution, one should consider the value of each individual score and determine the amount by which each varies from the most expected value (mean) of the distribution. Recall that the mean was identified as the most stable measure of central tendency. Thus deriving an index based on how each score deviates from the mean score, one will have considerable stable or very stable, index of variability in a distribution.

The deviation scores provide a good basis for measuring the spread of scores in a distribution. However, we cannot use sum of these deviations in order to get an index of spread because this sum in any distribution will be zero.

$$\sum_{i=1}^{N}(X_i - \mu) = 0 \text{ or } \sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

We know why things are getting messed up to obtain zero in every distribution for this sum. The reason is, because we are summing up negative and positive values, which happen to be exactly the same. Consequently, there are possible ways of rectifying things so that our index shows or measures the variability, such that the value obtained indeed indicates the magnitude of the variability and not the same in all cases as in the present case.

Considering all the above, two ways are possible of getting the index of variability:

1. Considering the sum of the absolute deviations and dividing by total number of cases (n). This is the absolute mean deviation or simply called mean deviation. Mean deviation = $\frac{\sum_{i=1}^{n}\left|X_i - \bar{X}\right|}{n}$ where $\sum_{i=1}^{n}\left|X_i - \bar{X}\right|$ is the sum of absolute deviations (disregarding plus and minus signs).

Note we divide by size of sample, n (or N when we consider population) to make our index independent of the size of the sample (or population). Thus, as long as there is the same variability, the index is the same irrespective of the number of cases are in our sample (or our population).

Computing mean deviation is very easy. Suppose we have scores: 3, 4, 5, 5, 6, 8, 9, and 10.

The mean deviation may be obtained as follows:

X	$X_i - \bar{X}$	$\left\|X_i - \bar{X}\right\|$
3	-3.25	3.25
4	-2.25	2.25
5	-1.25	1.25
5	-1.25	1.25
6	-0.25	0.25
8	1.75	1.75
9	2.75	2.75
10	3.75	3.75
$\sum_{i=1}^{n} X_i = 50$		$\sum_{i=1}^{n}\left\|X_i - \bar{X}\right\| = 16.5$

$\bar{X} = \frac{50}{8} = 6.25$

Note that $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$

But the mean deviation = $\frac{\sum_{i=1}^{n}\left|X_i - \bar{X}\right|}{n}$

$$= \frac{16.5}{8}$$

$$=2.0625$$

a larger value of the mean deviation indicates a greater spread in the values of the distribution.

2. Second way is to square the actual deviations and add them together and then divide by n (i.e. the total number of cases); we obtain an index known as variance. That is,

$$\text{variance} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

The variance is not in the unit of the scores and hence is not a measure of variability {recall cm is unit for length while cm^2 is unit for area]. In order to put the measure of variability into the right perspective- that is, to return to our original unit of measurement; we take the square root of the variance. The square root of the variance is standard deviation, which is a measure of variability and the most reliable measure of variability, theory or reasoning of this is beyond the scope of this unit or module. Standard deviation, s, is:

$$S= \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}} \quad \text{i.e. } S = \sqrt{\text{(variance)}}$$

Observe standard deviation and mean deviation have same units. Standard deviation has better properties than mean deviation hence more reliable or stable.

Population variance is symbolized by σ^2, consequently population standard deviation is σ. Sample variance is symbolized as S^2 and sample obviously as S.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

The computational formula for variance and standard deviation

Using the above formula referred to as “definitional formula” to compute variance is easy when each score and the mean are whole numbers. However, computation of variance is tedious and unwieldy and in many cases inaccuracy results due to rounding error. In such cases the *computational formula* or *raw score formula* is preferred. The formula is:

$$S_x^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n}$$

Subscript x on S^2 is just emphasizing or indicating we are looking for the variance for X-variable (scores).

The computational formula for variance has a great advantage over the definitional formula $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ since in the former formula; the raw scores are used directly without first resorting to determination of the mean. The computational formula is also easy to use when some scores in a distribution are fractional i.e. not whole numbers as already indicated.

Computation of variance (S_x^2) and standard deviation (S_x)

We shall use the scores 3, 4, 5, 5, 6, 8, 9and 10 to compute variance and the standard deviation using both the definitional and computational formulae. You will notice that the computational formula will be easier to use than the definitional formula.

Computation using definitional formula:

X	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
3	-3.25	10.5625
4	-2.25	5.0625
5	-1.25	1.5625
5	-1.25	1.5625
6	-0.25	0.0625
8	1.75	3.0625
9	2.75	7.5625
10	3.75	14.0625
		43.500

$$\text{Variance, } S_x^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

$$= \frac{43.5}{8}$$

$$= 5.4375$$

$$\text{Standard deviation, } S_x = \sqrt{5.4375}$$

$$= 2.33$$

Computation using computational formula or raw scores formula:

	X	X_i^2
	3	9
	4	16
	5	25
	5	25
	6	36
	8	64
	9	81
	10	100
Sum	50	356

Thus $\sum_{i=1}^{n} X_i = 50$ $\sum_{i=1}^{n} X_i^2 = 356$

Variance, $$S_x^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n}$$

$$= \frac{356 - \frac{50\times50}{8}}{8}$$

$$= \frac{356 - 312.5}{8}$$

$$= \frac{43.5}{8}$$

$$= 5.4375$$

Standard deviation, $S_x = \sqrt{5.4375}$

$= 2.33$

Note that even for a simple case like the one above, the definitional formula method tended to be unwieldy while in the case of computational formula method, it is rather straightforward.

Computing variance, S^2, of grouped frequency distribution

Variance of grouped data can also be calculated using definitional and computational formulae. Definitional formula for variance of grouped frequency distribution is:

$$S^2 = \sum_{i=1}^{n} \frac{f_i(x_i - \bar{X})}{N}$$

Where x_i is the class-mark (midpoint).

f_i is the corresponding frequency.

$\bar{X}$ is the mean.

N is the total frequency i.e. N = $\sum_{i=1}^{n} f_i$

The computational formula is:

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} f_i x_i)^2}{N}}{N}$$

Where x_i, f_i, $\bar{X}$ and N are as defined above in the definitional formula.

An illustration of the application of computational formula is worked out in detail below. Here, the first column contains the class intervals while the second contains the corresponding frequencies (f_i), and the third column contains the class- marks (midpoints) for each class interval (x_i). The fourth (the square of the class-marks) and the last two columns contain the products of class-mark and corresponding frequency, and square of class-mark and corresponding frequency respectively.

Calculation of variance, S^2, of grouped data

Class interval	Frequency (f_i)	Class-mark (x_i)	x_i^2	$f_i x_i$	$f_i x_i^2$
23-25	1	24	576	24	576
20-22	3	21	441	63	1323
17-19	5	18	324	90	1620
14-16	7	15	225	105	1575
11-13	8	12	144	96	1152
8-10	4	9	81	36	324
5-7	2	6	36	12	72
sums	$\sum_{i=1}^{n} f_i = 30$			$\sum_{i=1}^{n} f_i x_i = 426$	$\sum_{i=1}^{n} f_i x_i^2 = 6642$

$$S^2 = \frac{\sum_{i=1}^{n} f_i x_i^2 - \frac{(\sum_{i=1}^{n} f_i x_i)^2}{N}}{N}$$

$$= \frac{6642 - \frac{(426)^2}{30}}{30}$$

$$= \frac{6642 - 6049.2}{30}$$

$= 19.7$

Standard deviation, S = $\sqrt{19.76}$ = 4.445

Some properties of the variance, S^2, (or/and standard deviation, S)

Suppose we added a constant number to every score in a set of scores. How would the variance of the scores be affected? In illustrating the calculation of variance, we found that the scores 3, 4, 5, 5, 6, 8, 9 and 10 have the variance of 5.4375. Let us add 2 to each score and then calculate S^2.

Effect on variance, S^2, when each score is added a constant

X_{X+C} = Original score (X_X) +C	$[X_{X+C}]^2$
5	25
6	36
7	49
7	49
8	64
10	100
11	121
12	144
$\sum X_{X+C} = 66$	$\sum X^2_{X+C} = 588$

$$S_X^2 = \frac{\sum X_{X+C}^2 - \dfrac{(\sum X_{X+C})^2}{n}}{n}$$

$$= \frac{588 - \dfrac{66^2}{8}}{8}$$

$$= \frac{588 - \dfrac{66 \times 66}{8}}{8}$$

$$= \frac{588 - 544.5}{8}$$

$$= 5.4375$$

Adding 2 to each score did not change the value of variance, S^2. In general, adding or subtracting a constant to each score in a group will not change the variance (nor the standard deviation) of the scores.

What would happen to S^2 if each score was multiplied by a constant, say 2? Let us illustrate this with the same set of scores i.e., 3, 4, 5, 5, 6, 8, 9 and 10. let us multiply each score by 2 and then calculate S^2.

Effect on variance, S^2 when each score is multiplied by a constant

X_{CX} = Original score (X_x) ×C	$[X_{CX}]^2$
6	25
8	36
10	49
10	49
12	64
16	100
18	121
20	144
$\sum X_{CX} = 100$	$\sum X_{CX}^2 = 1424$

$$S^2 = \frac{1424 - \dfrac{100^2}{8}}{8}$$

$$= \frac{1424 - 1250}{8}$$

$$= 21.75$$

Note that 21.75 is 4 times or 2^2 times 5.4375. In general, multiplying each score by a constant makes the variance, of the resulting scores equal to C^2S^2. However, dividing each score by a constant makes the variance of the resulting scores equal to $\frac{S^2}{C^2}$. Thus we have come we two important properties of the variance and are:

1. Var. (X+C) = Var. (X) or $S^2_{X+C} = S^2_X$ i.e. adding a constant to every score in the distribution does not change the variance at all, and consequently the standard deviation is not affected either.
2. Var.(CX) = C^{2}Var.(X) or $S_{CX}^2 = C^2 S^2_X$ i.e. multiplying each and every score by a constant in the distribution, the resulting scores have a variance equal to constant squared times the original variance.

These properties can be used to ease computation of the variance quite considerably, and the method is encouraged whenever we are computing variance; much so, when we have grouped data. The example given is with grouped data. Similar method can be used also with ungrouped data.

Computation of variance using assumed mean method

Class-mark (x_i)	New class-mark x_i'	$x_i'^2$	Frequency (f_i)	$f_i x_i'$	$f_i x_i'^2$
67	3	9	3	9	27
62	2	4	4	8	16
57	1	1	8	8	8
52	0	0	10	0	0
47	-1	1	9	-9	9
42	-2	4	3	-6	12
37	-3	9	4	-12	36
32	-4	16	1	-4	16
			$\sum_{i=1}^{n} f_i =42$	$\sum_{i=1}^{n} f_i x_i' = -6$	$\sum_{i=1}^{n} f_i x_i'^2 =$ 124

$$\text{Variance in scale value} = \frac{\sum_{i=1}^{n} f_i x_i'^2 - \frac{(\sum_{i=1}^{n} f_i x_i')^2}{\sum_{i=1}^{n} f_i}}{\sum_{i=1}^{n} f_i}$$

$$= \frac{124 - \frac{-6\times -6}{42}}{42}$$

$$= \frac{124 - \frac{36}{42}}{42}$$

$$= \frac{124 - \frac{6}{7}}{42}$$

$$= \frac{123\frac{1}{7}}{42}$$

$$= \frac{123.143}{42}$$

$$= 2.932$$

How does this value, 2.932, compare with the required variance i.e. variance for the original grouped data?

Required variance = $5^2 \times 2.932 = 73.3$

Required standard deviation = $5\sqrt{2.932} = \sqrt{73.3}$
$= 8.56$

Since adding (or subtracting) a constant to scores leaves variance (hence standard deviation) both unaffected, multiplying (or dividing) each score by a constant affects variance as we have seen i.e.

Var. (X+C) = Var. (X) or $S^2_{X+C} = S^2_X$

Var.(CX) = C^{2}Var.(X) or $S^2_{CX} = C^2 S^2_X$ where C is a constant.

Then S(CX)= C.S(X) or $S_{CX} = CS_X$

Consequently the required variance = $5^2 \times 2.932 = 73.3$

The required standard deviation = $5\sqrt{2.932} = \sqrt{73.3}$
$= 8.56$ as seen above.

Interpretation of Standard Deviation

The standard deviation is a measure of variability that characterizes a distribution of scores or any other variables. It indicates how the scores (or any other variables) are spread. The bigger the magnitude of standard deviation, the bigger is the spread of the scores. The smaller the magnitude of the standard deviation, the smaller is the spread of scores.

For instance, a form I class with a high standard deviation in a mathematics test would be more heterogeneous. In other words, you would find the class being composed of students with more varied performance in mathematics than a class (say another form I class) with a low standard deviation. The latter form I class would be said to be more homogeneous i.e. the students have very similar performance in mathematics as opposed to the former. For practical reasons the two groups may require different teaching techniques in the subject consequently.

The standard deviation is also useful measure of variation because in many distributions of scores, for it is possible to know approximately what percentage of scores lie within one, two, or more standard deviations of the mean. For example, we know that about 70% of the scores lie between $\overline{X} - S_X$ and $\overline{X} + S_X$. This property of standard deviation will be discussed further in the next chapter, which deals with the normal distribution.

Pooled Combined Variance or Pooled Overall Variance:

e.g. $n_1 = 30$ $\overline{X}_1 = 20$ $S_1 = 8$

$n_2 = 10$ $\overline{X}_2 = 15$ $S_2 = 2.5$

$n_3 = 15$ $\overline{X}_3 = 18$ $S_3 = 6$

Pooled combined variance = $\dfrac{n_1S_1^2 + n_2S_2^2 + n_3S_3^2}{n_1 + n_2 + n_3}$

$$= \frac{30\times 8^2 + 10\times 2.5^2 + 15\times 6^2}{30+10+15}$$

$$= \frac{1920 + 62.5 + 540}{55}$$

$$= \frac{2522.5}{55} = 45.86$$

Thus pooled combined standard deviation (s.d.) = $\sqrt{\frac{n_1S_1^2 + n_2S_2^2 + n_3S_3^2}{n_1+n_2+n_3}}$

$$= \sqrt{45.86}$$

$$= 6.77$$

The general formula is:

Pooled combined variance = $\frac{n_1S_1^2 + n_1S_2^2 + \text{..........} + n_aS_a^2}{n_1 + n_2 + \text{........} + n_a}$

$$= \frac{\sum_{i=1}^{a} n_i S_i^2}{\sum_{i=1}^{a} n_i}$$

Pooled s.d. can be obtained accordingly.

Skewness

The third element of describing distribution adequately (i.e. absolutely sufficiently) is shape of distribution. Distributions come in all kinds of shapes such as normal, skewed, bimodal, U-shaped, rectangular (uniform) etc. Below here we consider skewness.

Skewness is defined as:

Skewness = $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3 / n}{S_X^3}$

By definition of skewness the range of skewness is usually between –3 and +3, being zero in normal distribution. Those distributions, which have negative values for skewness, are referred to as ‘ negatively skewed distributions’ and consequently those, which with positive value for skewness are referred to as ‘positively skewed’.

Note: In unimodal distribution which is positively skewed then mean > median > mode. In negatively skewed distribution, the mean < median < mode.

A simpler formula for computing skewness is:

Skewness = $\frac{3(\overline{X} - md)}{S_X}$ where $\overline{X}$ and *md* are the mean and the median of the scores respectively.

Division by SX (standard deviation of scores) makes the measure independent variability and of consequently will range between –3 and +3.

SUMMARY

The following statements summarize the major pints of this unit:

1. The range, variance and standard deviation are measures of variability. They give an indication of the spread of scores in a distribution.
2. The range is defined as the difference between the highest and the lowest scores in a distribution.
3. Variance is obtained by dividing the sum of squared deviations by the total number of observations in the distribution.
4. The standard deviation is the square root of variance.
5. The bigger the standard deviation, the bigger the spread of scores and the more heterogeneous the group is on which the scores are based.
6. Adding a constant to every score in the distribution has no effect on variance or standard deviation.
7. When every score in a distribution is multiplied by a constant, the new variance is the original variance times the constant squared.

FURTHER READING

Ingule, F. & Gatumu, H. (1996) *Essentials of Educational Statistics*. Nairobi: E.A. Educational Publishers.

Glass, G.V. & J.C. Stanley (1970) *Statistical Methods in Education and Psychology*. New Jersey: Prentice-Hall,

Johnson, R.R. (1980) *Elementary Statistics.* Mass.: Duxbury Press,

Smith, G.M. (1970) *A Simplified Guide to Statistics for psychology and Education* New York : Holt, Rinehart and Winston,

ACTIVITY

1. Given the scores 3, 4, 4, 5, 6, 6, 7, 8 and 10; compute
   - (i) Mean, range, mean deviation, median, variance (S_X^2), and standard deviation (S_X).
   - (ii) Add 6 to each score, and recalculate the mean, variance and standard deviation.
   - (iii) Subtract 5 from each score, and recalculate the mean, variance and standard deviation.
   - (iv) Multiply each score by 4, and recalculate the mean, variance and standard deviation.

2. In case of 1 (ii) to 1 (iv), discuss what happens to:
   (i) The whole distribution
   (ii) The mean
   (iii) The variability (i.e. variance and standard deviation) in relation to the constant applied.

3.

Class interval	Frequency
65-69	3
60-64	4
55-59	8
50-54	10
45-49	12
40-44	6
35-39	5
30-34	2
	N =50

Using the above data and changing the scale of the class-mark by appropriate manipulation (i.e. using assumed mean method).

(a) Compute:
   (i) The mean
   (ii) The variance and standard deviation.

(b) Repeat using the above data, but *omitting* the top class interval and the bottom class interval, (N = 45).

(c) Double the frequency of the data and determine how the median and mean, variance and standard deviation are affected.

(d) Double only the last frequency of the data and determine how the median, mean, variance and standard deviation are affected.

4. The following scores were obtained by 30 Form I students in a Kiswahili test: (KU Exam, 2001)

46	31	18	39	40	38
37	19	15	26	14	37
24	41	18	19	21	25
31	10	20	21	32	46
20	30	32	27	31	37

a. Use the above scores to prepare a grouped frequency distribution using a class interval size 5 and starting with 10-14 as your lowest class interval.

b. Basing on your grouping in a) above, prepare a complete frequency distribution table for grouped data having the following columns:
   i. Class interval
   ii. Tally marks
   iii. Frequency
   iv. Real (or exact) class limits
   v. Classmark (midpoints)
   vi. Cumulative frequencies below (less than)

vii. Cumulative frequencies above (more than) c. For the grouped data, calculate the following:

i) Mode ii) Median iii) Mean

d. Determine the range for the grouped data.

e. Calculate deviation for the grouped data.

f. Compute the variance and standard deviation for the grouped data.

g. i) Comment on the performance in this Kiswahili test using above information.

ii) Describe fully the shape of the distribution basing your answers to part (c).

LECTURE FIVE, SIX AND SEVEN

NORMAL DISTRIBUTION

INTRODUCTION

The concepts discussed under normal distribution are:

1. Area of the normal curve which corresponds with the number of cases under consideration
2. Relationship of students' performance and normal curve distribution
3. How to use standard normal curve tables to analyze students' performance.

OBJECTIVES

By the end of the unit the learner should be able to:

1. Explain the symmetry of normal curve
2. Standardize any normal distribution to a standard normal distribution and then use standard normal curve tables to analyze any normal distribution
3. Explain what are normal z-values.

THE NORMAL DISTRIBUTION

The *normal distribution* or the normal curve is the most important distribution in statistics. It plays an important role in both descriptive and inferential statistics. It is an excellent approximation of the frequency distributions of large number of observations taken on a variety of variables. For example, the frequency polygons of heights as well as for weights of adults look like a normal curve. Tests of mental abilities often yield distributions of scores that conform closely to the normal distribution. In fact, many measures of most human characteristics (trait) approximate the normal curve.

However, the normal curve is a mathematician's invention that is a realized if we consider infinitely many observations and draw a histogram or frequency polygon; and taking as our size of class interval, (i) very close to zero or infinitely small. That is, our hypothetical conditions are n tending towards infinity, and size of class interval tending towards zero, both histogram and frequency polygon will be very close to normal distribution curve with quite a number of human traits such as intelligence.

A collection of scores that are exactly normally distributed has never been gathered and will never be. But much is gained if we can tolerate the slight error (or deviation) because certain mathematical properties of normal curve produce simple elegant results to many inferential statistics problems.

Properties of the Normal Curve

A normal curve looks like a well-weathered heap of *unga* (flour), sand, or manure. It is a bell-shaped symmetrical curve with peak of the distribution in the centre and with the 'tails' of the distribution continually approaching, but never touching the horizontal axis.

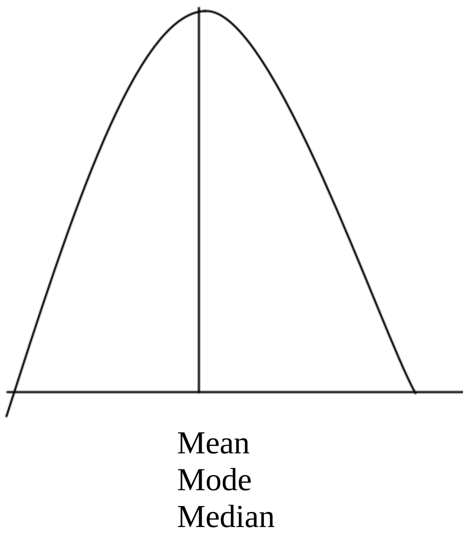
The formula for the curve is:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Note that the curve is defined in terms of parameters.
Where σ^2 is the population variance of variable x.
i.e. σ is the standard deviation of the population in x-variable.
μ is the population mean in x-variable.
Y is the height of the curve corresponding to particular values of x (i.e. Y corresponds to frequency of each score).
e is the base of the system of natural logarithmn and is about 2.718.
π is 3.142

Although the formula need not be learned or memorized, what is important to take note of is that all the properties of normal curve are described when the mean and standard deviation are known. The mean, mode and median pass through the peak of the curve and precisely bisect the area under the curve into two equal halves.
Observe that this normal curve or normal distribution curve is a continuous distribution curve. A normal curve is symmetrical about the mean (mode and median) i.e. its maximum height is at the mean, an idea brought out above in different words.
Although the height of the curve continues to decrease as one moves farther and farther from the mean, it never actually reaches zero. Therefore, the theoretical range of normal curve is from plus infinity (+ ∞) to minus infinity (- ∞). Actually so little of the curve extends above +3 standard deviation (s.d.) or below –3 s.d. that for many purposes these are the practical limits of the normal curve.

Normal distribution curve



Areas under the normal curve

An important feature of normal curve is the relationship between the area under the curve (i.e. the percentage or proportion of the population) and the standard deviation of a distribution. It turns out that, regardless of the particular mean or standard deviation a normal curve may have, there will be a constant proportion of area between the mean and an ordinate, which is a given distance from the mean in terms of standard deviation units.

For the formula:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

consider the standard case where σ = 1 and μ = 0, i.e. in the above formula we have:

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2}$$

We say z (i.e. z-values or z-scores) is normally distributed with mean zero and variance equal to 1; symbolized as N(0,1).

For case, $Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

We say x (i.e. x-variable or x-scores) is normally distributed with mean μ and variance σ^2 Symbolized as N(μ, σ^2).

Note then $z = \frac{x-\mu}{\sigma}$ hence z would be normally distributed with mean zero and variance one or z have a standard normal distribution; i.e. z ~ N(0, 1).

What we are saying here is *any normal distribution can be linearly transformed into a standard normal distribution* using the formula:

$$z = \frac{x - \mu}{\sigma}$$

or

$z = \frac{x - \bar{x}}{s}$ where μ or $\bar{x}$ is the mean of distribution while s or σ is standard deviation of the distribution under consideration.

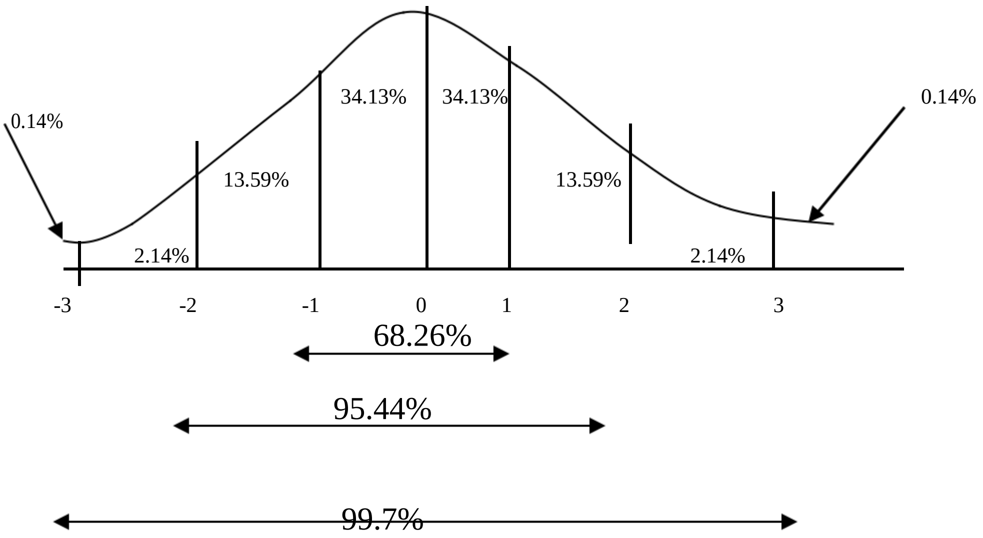
Statistical characteristics

The normal curve has the following statistical characteristics:

1. 50 percent of the area lies above the mean (in the centre) and 50 percent below the mean. 34.13 percent of the total area under the curve lies between the mean (at the centre) and one standard deviation above or below the mean. This implies that 34.13 percent of the observations or individuals have scores, which fall between the mean and one (1) standard deviation from the mean.
2. 47.72 percent of the area lies between the mean and a point 2 standard deviations above or below the mean.

3. About 49.9 percent of the area lies between the mean and a point 3 standard deviations away.
4. About 68.26 percent of the area (and cases) lies within 1 standard deviation (above or below) of the mean. This obtained by doubling the 34.13 percent provided in number 1 above.
5. About 95.44 percent of the area lies within 2 standard deviations, above and below the mean.
6. Finally, about 99.8 percent of the area lies within 3 standard deviation, below and above the mean. The figure below summarizes these observation

Area under the curve



A number of normal distributions exist and they differ only by their means and standard deviations. A good analogy here is the cube. A cube remains a cube whether small or big. Similarly, a normal curve is normal whether means and standard deviations are big or small. Properties (e.g. area) can be studied by having just a small model of a cube, and consequently, these properties can be generalized to all cubes. By the same token, the properties of any normal curve may be studied by considering one normal curve, the *standard normal curve.*

The standard normal curve

The above discussion shows that there must be a large family of normal distribution curves. In order to have some comparability among these normal distribution, statisticians use a standard normal curve, which has a mean of zero, 0, and a standard deviation (s.d.) of one (1). This standard normal curve facilitates the use of a single table to evaluate the area within various intervals of any normal curve. To get the units of the standard normal curve we use formula:

$$z = \frac{x - \mu}{\sigma}$$

or

$$z = \frac{x -}{s\,\bar{x}}$$ which was seen above and so is defined above.

z is the unit of the normal curve and is referred as *'normalized z-value or z-score'* or normalized standard score or otherwise simply z-score or standard score. Any raw score can be converted to a z-score by subtracting the mean of the group of scores from the raw score and dividing the result by standard deviation. However normalized z-scores only emanate from only normal distributions as indicated and will be emphasized. Standard scores (z-scores or z-values) represent abstract numbers rather than particular units of the original scores. It (standard score) indicates how many standard deviations (s.d.'s) from the mean a score lies e.g. if a normal distribution has a mean of 60 with a standard deviation of 10, then a z-score for a raw score of 70 is:

$$z = \frac{x - \mu}{\sigma}$$

or

$$z = \frac{x -}{s\,\bar{x}} = \frac{70\text{–}60}{10} = 1$$

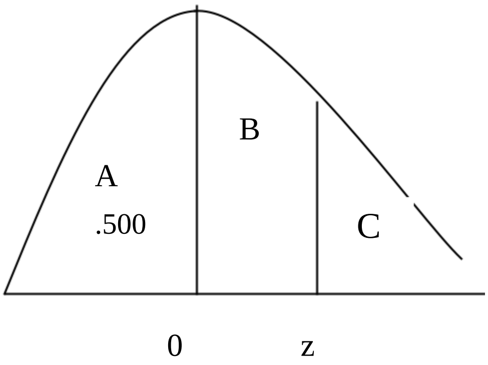
The value of z which is +1, shows that the raw score, 60, lies one standard deviation above the mean; if z = -2, the raw score is 2 standard deviations below the mean.

What we are saying is, if we convert scores to standard scores (z-scores), what we are doing in reality is transforming our original distribution to a new distribution with mean as 0 and standard deviation as 1. Observe the graph of z-scores will have exactly the same shape as the original raw scores distribution. Transforming raw scores into standard scores (z-scores) does not change the shape of the original distribution. If the original distribution was skewed, then the z-scores distribution derived from them (i.e. original scores or raw scores) will be skewed. If the original distribution is normally distributed, then the standard scores distribution derived from them will be normally distributed.

Use of standard normal tables

Tables showing the exact areas under the standard normal curve have been constructed. The use of these tables makes the process of obtaining the areas under z-values less cumbersome. The standard table that appears below is one variant of the standard normal table. Basically, the table has two columns. One column has heading, z, and other, Area B. The tables provide z-values with the corresponding areas in the same row in the adjacent area B column. The areas B provided on the table show the proportions from the mean (at the centre) to particular z-value (or z-score). For example, for a z-value of 1.70, the area provided is .455.

This is the area between z = 1.70 and mean at the centre of the curve (z-value here is 0). That is, the area B is the area between each z-value and the mean; the value of z-value at the mean is 0. We can invoke the symmetric property of the normal curve (or standard normal curve) and show that for –1.70, the relevant area will also be 0.455. What is the area between the mean and z =1.50? This is .433 and it is the area for z = -1.5



Total area under the standard normal curve is 1. Area A = .500. Area B + Area C (the tail) = .500. Entries in the table below are for Area B

Standard normal curve area table

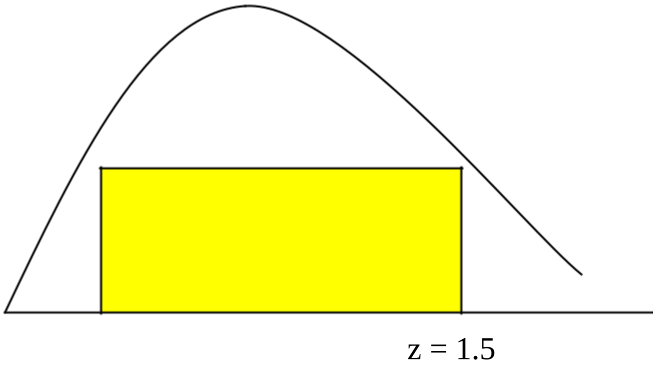
The areas under the standard normal curve corresponding to distances on the baseline between the mean and each z

(1)		(2)		(3)	
Z	Area B	Z	Area B	z	Area B
.00	.00	.50	.192	1.75	.460
.02	.008	.525	.200	1.80	.464
.04	.016	.60	.226	1.85	.468
.06	.024	.65	.242	1.90	.471
.08	.032	.675	.250	1.96	.475
.10	.040	.75	.273	2.00	.477
.12	.048	.80	.288	2.05	.480
.14	.056	.84	.300	2.10	.482
.16	.064	.90	.316	2.15	.484
.18	.071	.95	.329	2.20	.486
.20	.079	1.00	.341	2.25	.488
.22	.087	1.036	.350	2.30	.489
.24	.095	1.10	.364	2.33	.490
.26	.103	1.15	.375	2.40	.492
.28	.110	1.20	.385	2.45	.493

.30	.118	1.25	.394	2.50	.494
.32	.126	1.28	.400	2.55	.4946
.34	.133	1.35	.412	2.58	.4951
.36	.141	1.40	.419	2.65	.4960
.385	.150	1.45	.427	2.70	.4965
.40	.155	1.50	.433	2.81	.4975
.42	.163	1.55	.439	3.09	.4990
.44	.170	1.60	.445	3.30	.4995
.46	.177	1.645	.450	3.70	.4999
.48	.184	1.70	.455	4.00	.49997

Sometimes the interest is in determining the area or proportion of cases below a z score. For example, we may be interested in determining the proportion of scores below a score equivalent to z = 1.5. It may be helpful at this point to start by visualizing the normal curve. One can even draw an illustrative normal curve as below:

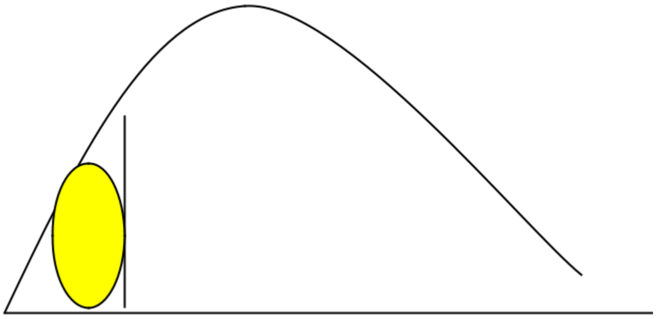
Area below z = 1.5



The area of interest is shown by filled up area of the curve. From the standard normal curve table, we can determine the area between the mean and z =1.5. This area is .433. The area below the mean is .5. Because of the symmetry of the normal curve it means that 50% of the area is below the mean and 50% is above the mean. Thus the total area below z = 1.5 is .433 plus .5, which is .933.

What is the area below z = -1.5? We can again visualize the (standard) normal curve. Let us do this by looking at the normal curve below:

Area below z = -1.5

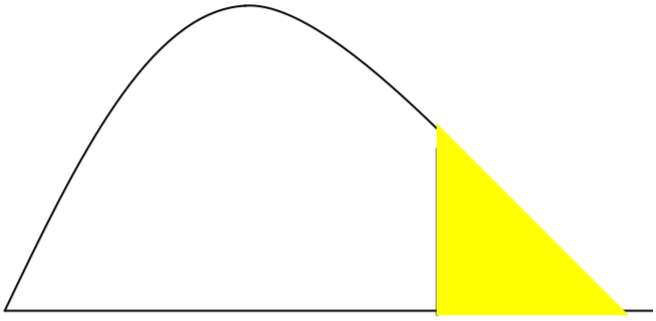


z = -1.5

The area below z = -1.5 is on the left of the line and part of it is filled. From the standard normal table, we know that the area between the mean and z = -1.5 is .433. If we revisit the characteristics of the normal curve, we know that the area below the mean is .5. Therefore, the area below z = -1.5 is .5 minus .433 which is .067. This means that .067 of the cases or scores will lie below z = -1.5.

In some cases, the interest may be in determining the proportion or area above a z score. For instance, what would be the area above z = 1.5? This is the portion to the right of z = 1.5, which is filled in yellow and shown below. The table value for z = 1.5 is .433. To obtain the value for the filled part, which is the area above z = 1.5, we subtract .433 from .5. Why the subtraction? Well, we know that the area above the mean is .5. Since the area between the mean and z = 1.5 is .433, the shaded area is the difference between .5 and .433 (i.e. .5 - .433 = .067).

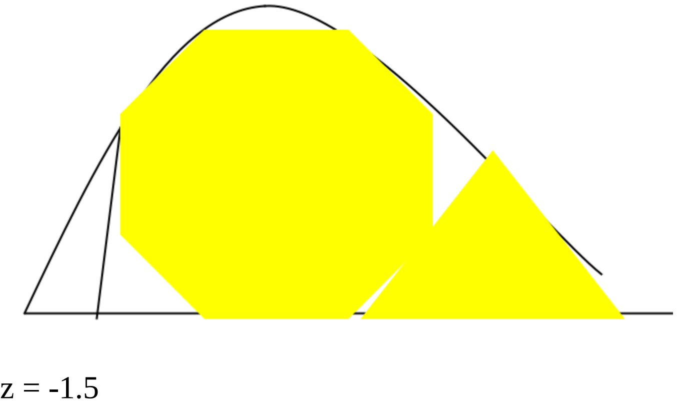
Area above z = 1.5



z = 1.5

What about the area above z = -1.5? This area lies between z = -1.5 and the mean and includes the area above the mean. The area between z = -1.5 and the mean is .433 and area above the mean is .5. Therefore, the area between z = -1.5 is .5 plus .433 which equals to .933. This area is the filled portion of the figure shown below:

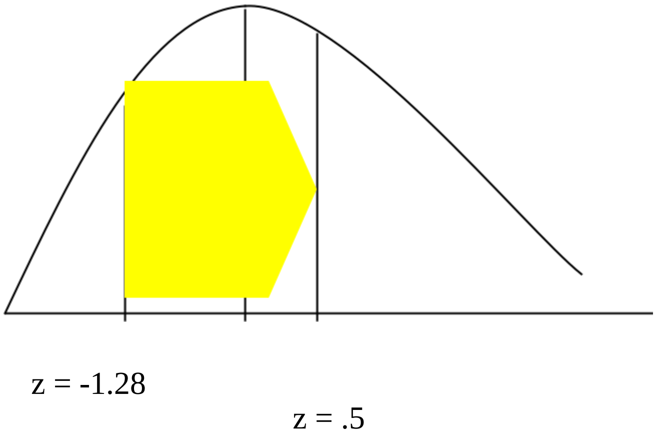
Area above z = -1.5



Area above z = -1.5 is .5 + .433 = .933

The Interval Problem:

The standard normal distribution can also be used to find the area between any values of z. For example, we may want to find the area between the z value –1.28 and z value .50. The area between the mean and z = -1.28 is .400 and area between the mean and z = .5 is .192. Thus, the area between z = -1.28 and z = .50 is .400 plus .192 which is .592. In other words, 59.2% of the area lies between these two z values as shown below.

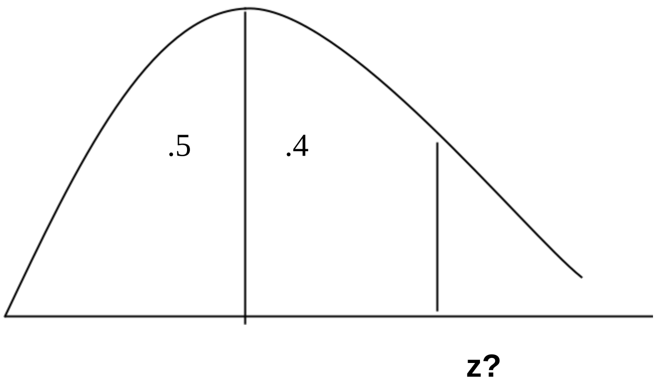


Starting with proportions

Sometimes, we may only know the area or proportion of normal distribution and may be required to determine the corresponding z-score. For example, we may require to find the z value below .90 of the scores or cases in a normal distribution must lie. To find the value of z, one must first find the column containing the area. This proportion is then matched with the z value in the adjacent z column. However, to get the table value of the z below, which .90 of the scores must lie, we need to do some thinking. First, we know that of this, .9 and .5

will lie below the mean. This means that .4 of the area would lie above the mean. This is the value to look for in the body of the standard normal distribution table. The z corresponding to this proportion is 1.28. This z value is above the mean and we know that all z values that fall above the mean are positive. Thus, the value below which .90 of the scores must lie in a distribution is 1.28 (or +1.28).

The z value below which .90 of the scores lie:



From the table we find this value of z is 1.28

Now let us look at an example of how knowledge of the normal curve can be used to solve some practical problems.

Illustrations

In a class of 100 pupils the mean of a test is 15 and standard deviation is 2.5. Assuming these 100 scores have a normal distribution,

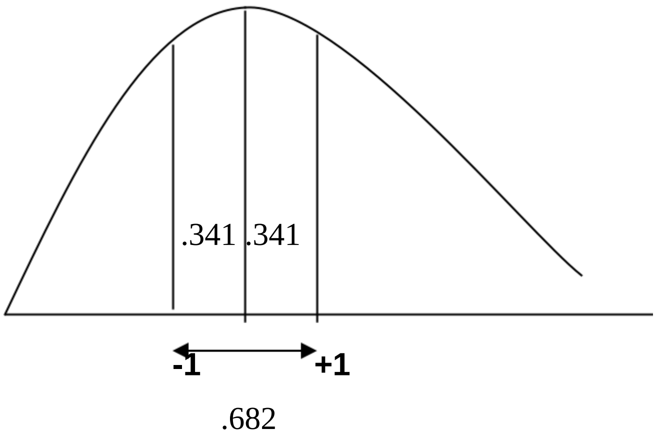
(a) How many pupils' scores lie between 12.5 and 17.5?

Solution

Note 12.5 is one standard deviation below the mean (since the z-value for 12.5 is

$$z = \frac{x - \bar{x}}{s} = \frac{12.5 - 15}{2.5} = -1$$

Similarly, it can be shown readily that 17.5 is one standard deviation above the mean. Thus, the area between one standard deviation below the mean (z = -1) and the mean is .341 for standard normal curve. The area between the mean and one standard deviation above the mean (z = 1) is also .341. Consequently, the total area is .682.



Thus, the number of students who scored between 12.5 and 17.5 are .682 × 100% = 68%. Since the total number of students was 100, then the required number of students who scored between 12.5 and 17.5 is 68.

(b) How many scored above 20?

Solution

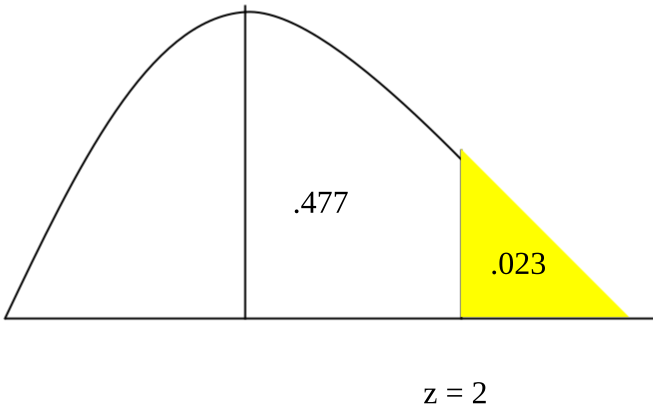
Remember, 20 is 2 standard deviation above the mean since the z-score for 20 is:

$z = \frac{x - \bar{x}}{s} = \frac{20 - 15}{2.5} = 2$

Hence those above 2 standard deviations are .500 - .477 = .023

Since the area between the mean and 2 standard deviations is .477, 2 standard deviations and above occupy .500 - .477 = .023.

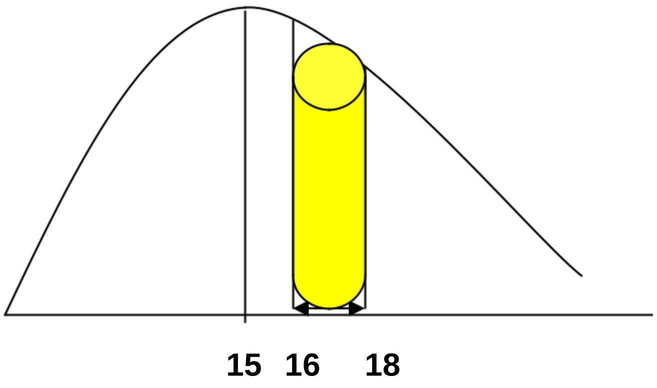
Thus, the number of students above 20 is .023 × 100 = 2



(c) How many scored between 16 and 18?

Solution

Score 16 is $\frac{16 - 15}{2.5} = \frac{1}{2.5}$ = .4 standard deviation above the mean. While 18 is $\frac{18 - 15}{2.5} = \frac{3}{2.5}$ = 1.2 standard deviation above the mean.

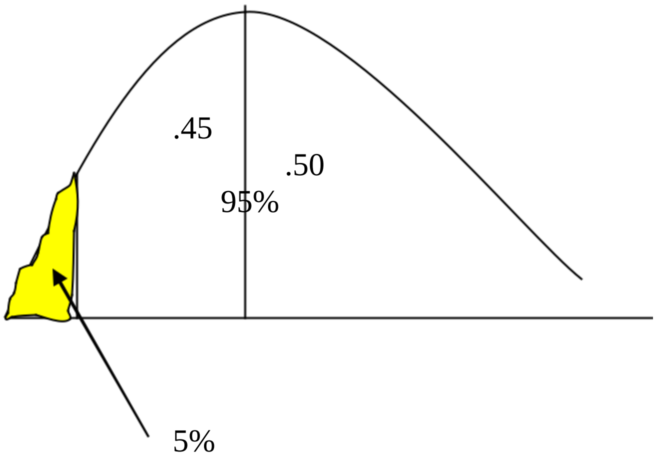


.

Thus, the area between 15 and 16 is .155 (i.e. the area for z = .4 from the above table), area between 15 and 18 is .385 (area for z = 1.2 from the table also). Thus the area between 16 and 18 is .385 - .155 = .230 (the yellow filled area), and this area represents .230 × 100 = 23 pupils

(d) Suppose 95% of the top scores of this group are to be selected. What is the minimum score one has to obtain to be selected?

Solution



to be able to use the tables provided able, we need to get the area from the mean to any required z. In the present problem, the area is the one given so look for the z value corresponding to this area B.

The area is .95 - .50 = .45 and it is below the mean, so its corresponding z value is negative. From the table, z value corresponding to area is –1.645. (It is negative since it is below the mean).

We know z = $\frac{x - \cancel{x}}{s}$

So required x-score is: -1.645 = $\frac{x - 15}{2.5}$

This implies that x is 15 –1.645 × 2.5 = 15-4.1125

= 10.8875

So the minimum score one has to get to be selected is 11.

SUMMARY

The following statements summarize the major points in this unit.

1. The normal distribution curve is the most important kind of distribution in statistics.
2. It is a bell-shaped, symmetrical curve with the peak of the distribution in the centre and with the 'tails' of the distribution continually approaching but never touching the horizontal axis.
3. Regardless of the particular mean or standard deviation a normal curve may have, there is a constant area or proportion of cases between the mean and an ordinate given by the distance from the mean in terms of standard deviation units.
4. A number of normal distributions exist and differ only to have comparability among normal distributions; a standard normal curve with mean of zero and a standard deviation of one is used. In other words, normal distributions differ in terms of their different means and their different standard deviations; and any normal distribution can be linearly transformed into a standard normal distribution.
5. The units of a standard normal curve are expressed in the standard random variable, z i.e. normalized z-scores and they correspond with standard deviations from the mean.
6. Tables showing the exact areas under the standard normal curve exist.
7. These tables can be used to provide information on areas below and above given z values. The tables can also be used to give information on areas between any z values.

FURTHER READING

Ingule, F. & Gatumu, H. (1996) *Essentials of Educational Statistics*. Nairobi: E.A. Educational Publishers.

Glass, G.V. & J.C. Stanley (1970) *Statistical Methods in Education and Psychology.* New Jersey: Prentice-Hall,

Johnson, R.R. (1980) *Elementary Statistics.* Mass.: Duxbury Press,

Smith, G.M. (1970) *A Simplified Guide to Statistics for psychology and Education* New York : Holt, Rinehart and Winston,

ACTIVITY

1. In a biology test, the mean was 48 and standard deviation was 5 for a group of 100 Form II students.
   (a) Assuming a normal distribution,
      (i) How many scores were there between 43 and 53?
      (ii) How many were there above 43?
      (iii) How many pupils scored below 45?
      (iv) How scored between 45 and 56?
   (b) Supposing due to limited facilities, 90% of top students are to be selected using these scores, what is the minimum score a pupil has to obtain so as to be selected?

2. In an intelligence test administered to 100 children, the mean of the test was 100 and standard deviation 20.
   (i) Find the number of children exceeding a score of 125.
   (ii) Find the number of children who obtained a mark between 90 and 130.
   (iii) Suppose 5% of the lower end of the distribution are to be selected for an enrichment program, what should be the cut off mark?
3. In a mock examination, the overall mean score was 100 and the standard deviation 20. Assuming that the scores were normally distributed and the classes which did the mock had 200 pupils altogether:
   (a) How many pupils scored marks between 80 and 120?
   (b) How many scored between 110 and 120?
   (c) Which score separates the upper 20% of scores from the lower 80 in the mock examination?
4. If class scores on a final examination are approximately normally distributed, what proportion of the class scores should fall at or below the following z scores:

   (a) –1.2 (b) .96 (c) 1.88 (d) 1.78 (e) -.43 (f) 2.15.
5. If 100 scores on a test are normally distributed with mean of 500 and a standard deviation of 100:
   (a) What percentage of scores would be lower than 650?
   (b) How many scores would lie below 650?
6. Find the area under the normal curve in each of the cases below:
   (a) Between z = 0 and z = 1.2
   (b) Between z = -.68 and z = 0
   (c) Between z = -.46 and z = 2.21
   (d) Between z = .81 and z = 1.94
   (e) Between z = -.50 and z = -2.00
7. The scores on a test were normally distributed, with mean of 76 and standard deviation of 12.
   (a) Compute the z score for the score 70 on the test.
   (b) Compute the z score for the score 94 on the test.
   (c) What proportion of the scores in the distribution should fall between 70 and 94?
   (d) If N = 50, how many scores fall below 70?
   (e) If N = 50, how many scores lie between 70 and 94?

LECTURE EIGHT AND NINE

INTERPRETATION OF SCORES

INTRODUCTION

In norm-referenced tests (NRT) where we consider the relative position of a score, transformation of scores into percentile ranks, standard scores (z-value), standardized scores or normalized scores are justifiable. Such manipulation (transformation) is not justifiable in criterion-referenced tests (CRT). A score in CRT is meaningful identity conferring information about how a candidate has mastered what is being tested.

OBJECTIVE

By the end of unit the learner should be able to:

1. Define percentile rank, percentile point, standard scores (z-value), standardized scores and normalized scores.
2. Compute z-values, percentile ranks given normalized z-values
3. Convert z-scores into standardized scores or normalized z-values to normalized scores such as stanines.

Interpretation of Test Scores

An individual's test score in any test is often obtained by simply adding the number of correctly solved problems. This score is the raw score of a particular individual. However, the raw score obtained for individual by such a procedure is dependent not only on his own performance but also on the properties of the test like the number and difficulty of items in the test. Therefore, one problem with use of raw scores is that they may not be comparable across tests. For example, if Kamau has a raw score of 70 on Test I and a raw score of 50 on Test II, we cannot easily assess his relative performance on the two tests, particularly if the distributions of scores for the two tests have quite different shapes.

Additionally, a raw score by itself is difficult to interpret. Even if raw scores are accompanied by information about the number of items on the test, an isolated raw score does not give any information about how one examinee's performance is related to the performance of other examinees. Using common transformations of raw scores can solve all these problems that are easily interpretable reported scores. Common forms of expressing transformed scores (else referred to as derived scores) are: *percentiles*, *standard* and *standardized scores*, and *normalized scores.*

It good to mention that such transformations are only justifiable in norm-referenced tests, and not in criterion referenced tests. In criterion-referenced tests a raw score is meaningful by its own right. It (a score) indicates the mastery or competency of the examinee. While in norm-referenced tests scores need to be transformed, so as to indicate the relative position of a score in the distribution of other scores in the norm (peer) group.

Percentile ranks

The percentile rank (or percentile score) is defined as the percentage of scores, which fall at or below a given score. For example, if a percentile rank for a score is 85, it means that 85 per cent of the scores in the total distribution fall at or below the score. The formula used for computing the percentile rank is:

$$\text{Percentile rank} = \frac{(cf_b + \frac{f_w}{2})100}{N}$$

Where

c*fb* = cumulative frequency (below) for the interval immediately below the interval containing the score of interest.

fw = the number of scores within the interval containing the score of interest. This is equal to the number of examinees obtaining the score of interest.

N = the total number of subjects in the distribution i.e. the total number of examinees.

Let us consider the scores distribution in the table below. For each score value the frequency (number) of the examinees obtaining the score appears in the second column. The third column contains the cumulative frequency for each score value, which is the number of examinees who have score less than or equal to each score value. Let us use this observed-test-score distribution to estimate the percentile rank for the score value, 5 (that is the percentile rank for the individual who gets 5).

Score value(X_i)	Frequency (f_w)	Cumulative frequency below (c*fb*)	Percentile rank
11	1	10	95
10	1	9	85
8	1	8	75
7	2	7	60
6	1	5	45
5	2	4	30
4	1	2	15
3	1	1	5

For the score value of 5, the frequency within the interval in which the score is located (i.e. between 4.5 and 5.5) is 2. This is equivalent to f_w in the formula above. The interval below 4.5 and 5.5 is 4.5 and 3.5. This interval is represented by the score value 4, which has a cumulative frequency of 2. This 2 is equivalent to c*fb* in the formula for the percentile rank. It represents the cumulative frequency below the interval 3.5 to 4.5, which is the interval below the interval of interest. Using this information, we obtain the percentile rank for the score value 5 as below:

$$\text{Percentile rank for 5} = \frac{(cf_b + \frac{f_w}{2})100}{N}$$

$$= \frac{(2 + \frac{2}{2})100}{10}$$

$$= 30$$

The percentile ranks for other scores in the distribution are provided in the fourth column. Under the definition, the percentile rank of a score is always less than 100 and greater than zero (i.e. 0 < percentile rank < 100).

Advantages and disadvantages of percentile ranks

Percentile ranks (or simply percentiles), like other transformed (derived) scores, have some advantages and some disadvantages. The primary advantages of percentile are that they are straightforward to calculate, regardless of the shape of distribution of observed scores, and are easy to interpret. For communication with people who have little background in statistics, percentiles are probably the most meaningful transformed scores.

There are a number of limitations in percentile scores. They can be assumed to form ordinal scales; thus, the arithmetical manipulation of percentiles. For example, the calculation of means and variances of percentiles can produce misleading results. Use of means and variances of percentiles can lead to inaccurate conclusions. These should not be obtained at all under any circumstances.

Secondly, the distribution of percentiles is rectangular (or uniform), and consequently far from the original shape of the raw scores. We have a rectangular (or uniform) distribution, since by definition 1% of the examinees are at each percentile. In rectangular distributions, all scores have equal frequencies. A rectangular distribution curve looks like a horizontal line.

Finally, the percentile ranks magnify raw score differences near the middle of the distribution, but reduce the raw score differences toward the extreme. This means that the actual difference between the score in the middle of the distribution is much less than revealed by corresponding percentile ranks while the actual difference between scores on the extreme (low or high scores) is much greater than revealed by difference between their corresponding percentile ranks. This phenomenon is more pronounced when the test is very short.

Standard scores

A standard score indicates the relative position of a score in a distribution in terms of standard deviations from the mean. To get a standard score corresponding to any raw score, the mean of the raw scores is subtracted from the raw score, and the result is divided by the standard deviation of the distribution (or the raw scores). Standard scores are also called z score or z values. Therefore:

$$z = \frac{x - \mu}{\sigma}$$

or

$$z = \frac{x - \bar{x}}{s}$$ where μ or $\bar{x}$ is the mean of distribution while s or σ is standard deviation of the distribution under consideration, as observed earlier.

Converting scores to standard scores

Individuals	Score (x_i)	$(x_i - \bar{x})$	z score
A	3	-7	-1.11
B	6	-4	-.63
C	7	-3	-.47
D	9	-1	-.16
E	15	5	.79
F	20	10	1.58
Sum	60	.00	.00
Mean	10	.00	.00
s.d.	6.32	6.32	1.00

Converting scores to standard scores using the formula above automatically puts the transformed scores (standard scores) into a new scale with a mean of zero (0) and a standard deviation (s.d.) of one (1). Each transformed score indicates how many standard deviations the raw score lies from mean. For instance, a standard score 1.38 (z = 1.38) indicates that the corresponding raw score lies 1.38 standard deviations above the mean. If the standard score is equal to –1.57 (i.e. z = -1.57), the corresponding raw score lies 1.57 standard deviations below the mean.

Converting raw scores to standard scores (z scores) has no effect on the shape of the distribution. If the original distribution of the raw scores is skewed to the right, the distribution of corresponding standard scores will also be skewed to the right. If the original raw-score distribution is normal, the distribution of the corresponding standard scores will also be normal. This is due to the fact that transformation of raw scores into standard scores is linear, since all we are doing is adding a constant to every core in the distribution and multiplying each result score by another constant. This is unlike percentile ranks, which have rectangular distribution and will change the shape of the raw score distributions.

It has already been mentioned that percentile ranks are ordinal measures. On the other hand, standard scores are interval measures. This means that a variety of mathematical computations not possible with percentile ranks can be done using standard scores. For example, we can obtain mean and variances, an exercise, which would only generate meaningless values if performed on percentile ranks.

Disadvantages of standard scores

One disadvantage of standard scores is that they may not be easily interpreted by ordinary people who have no knowledge of what the mean and standard deviation are. Another major disadvantage of standard scores is that about half of the scores are negative. Most people prefer not to deal with negative numbers, because transcription and mathematical errors are common (e.g., negative signs are easily lost). Examinees also dislike having negative scores. Generally, students do not like their scores being reported in negative form. These problems can be overcome by using standardized scores. Standardized scores are linear transformation of raw scores (or their standard-score equivalents) that eliminate the problems of negative numbers.

Standardized scores

Standardized scores are linear transformation of raw scores, but unlike the standard scores, they are always expressed as whole numbers and are non-negative. Any set of standard scores can be transformed to an arbitrary mean, μ_s, and standard deviation, σ_s, by applying the formula:

$$Y = \mu_s + \sigma_s z$$

Where z is the standard score and Y is the standardized score. An example of standardized scores is the T score (commonly referred to as 'linear T scores'). The T score has a mean of 50 and standard deviation of 10. The formula for a T score is:

$$T = 50 + 10z$$

To obtain a T score, the z score is multiplied by 10 and 50 is added to the product. If we start with a raw score, to obtain the equivalent T score, we go through the following steps. Firstly, we calculate the value of z i.e. by subtracting the mean of the distribution from the raw score (x_i) then dividing the result by standard deviation of the distribution. Secondly, we multiply the z value obtained in the first step by 10 and then add to this product 50. These operations are summarized in the formula:

$$T = \frac{(x_i - \bar{x})}{s} \times 10 + 50$$

Or $$T = \frac{(x_i - \mu)}{\sigma} \times 10 + 50$$

T scores are always whole numbers, and if the value obtained is not a whole number it has to be rounded to the nearest whole number. T scores are also non-negative and usually greater than zero.

It can be noted from the formula above that standardized scores are linear transformation of standard scores since the means and standard deviations of the standardized scores (50 and 10 respectively for linear T scores) are the constants we apply to obtain standardized scores. Since standard scores are a linear transformation of raw scores, it means that standardized scores are also a linear transformation of raw scores. Therefore, one disadvantage that is common to both standard and standardized scores is related to the similarity of their distribution to that of raw score distributions. This similarity in distribution shapes means that the distribution of the transformed scores will contain any irregularities found in the raw-score distribution. Irregular "bumps" in raw scores' distribution, usually due to sampling error (or irregularities), will be preserved by transformation.

A second problem of using standard and standardized scores is more subtle. Suppose Wanyonyi had two standard scores, 0.20 on test A and 0.18 on test B. That leads to the conclusion that he did about equally well on the two tests. However, if the shapes of the two score distributions are quite different, our conclusion would be wrong. For example, if test A is positively skewed and test B negatively skewed, the percentile scores for the standard scores 0.20 on test A and 0.18 on test B may be substantially different. Hence our interpretation of the comparability of the scores on the basis of standard scores would be misleading. To eliminate this problem, we may consider using percentile ranks or we may use normalized scores if the distributions are not too far from normal distribution.

Normalization

It has been discussed that a raw-score distribution or a distribution obtained from a linear transformation seldom has an exact statistical meaning. The disadvantages of raw scores or their linear transformation can be avoided by changing the form of the distribution so as to obtain a normal distribution of scores i.e., by performing *normalization.* All normalized scores have a normal distribution. On a normalized distribution, every score has a concise statistical meaning as a result. The percentage of individuals above and below each score is known exactly on a scale with a known mean and standard deviation unit of measurement. The transformation to normalized scores involves forcing the distribution of transformed scores to be as close as possible to a normal distribution by smoothing out, stretching, or condensing irregularities and departures from normality in the raw-score distribution. The normalization process involves several steps:

1. Transform the raw scores to percentiles.
2. Find the standard score in the normal distribution corresponding to each percentile.
3. (Optional) Transform these standard scores to standardized scores with desired mean and standard deviation.

Illustration

Let us use the data in the table below to illustrate the process of normalization.

Raw score (x_i)	Frequency (f_i)	Cf$_1$	Cf$_2$	Percentile	Normalized proportions	Normalized standard score (z)
15	4	24	22	92	.920	1.41
14	10	20	15	62	.620	.31
13	8	10	6	25	.250	-.67
12	2	2	1	4	.040	-1.75

The first column gives the raw scores (x_i). Column 2 shows frequencies for each raw score in the distribution of obtained scores. In third column Cf$_1$, cumulative frequencies have been computed. In the fourth column the Cf$_2$ frequencies have been transformed to cumulative frequencies for the class means in each class interval. The cumulative frequency for a raw score in the fourth column thus gives the number of scores in the distribution which fall below the class mean in the interval represented by the raw score. Eight scores lie in the class interval for raw score 13. If we assume that the individuals are equally distributed within the class, we see that half of the scores (i.e. 4 scores) lie below the class mean and half lie above. Since we have 2 scores below the class interval 13, and 4 scores below the class mean within the interval, the cumulative frequency for the class mean will be 2 + 4 = 6. This is the entry in the fourth column for the raw score 13. The entries for the other raw data scores are computed using the same procedure.

Column 5 shows the percentiles for each raw score. The percentile score in this case represents the percentage of scores below the class mean of the raw score. The percentile for the score 13 is shown in the fifth column as 25. To obtain this value, the, Cf$_2$ of the raw score is divided by 24 (the total number of scores in the distribution). The sixth column shows the proportions corresponding to the percentiles in column 5. For the raw score of 13, the relevant proportion is 0.2500. Hence 0.2500 of the total distribution of obtained scores lie below the class mean in interval of the raw score 13. Now that we know the proportion of the total distribution, which lies below the class mean in interval 13, we can obtain, from a table of cumulative proportions for a normal distribution, the position on a normal distribution, which corresponds to this proportion. We find that the standard score corresponding to the proportion 0.25 (i.e. .5 -.25 = .25,) is –0.67; note for 12 proportion is .5 -.04 = 0.46. The raw score 13 will therefore correspond to the standard score –0.67 in the distribution of normalized standard scores. Similarly, we compute the positions on a normal distribution corresponding to the class mean of the other raw scores.

As it has been pointed out, the procedure of linear transformation does not tamper with the shape of the distribution. This is not the case for normalization, unless the raw scores have a normal distribution. Normalization is basically a non-linear transformation for it tampers with the shape of the distribution, unless the distribution of the original scores (raw scores) is normal as previously indicated.

Two examples of normalized scores are normalized T scores and stanines. Normalized T scores can be obtained by multiplying normalized standard scores by 10 and adding 50 to the product. Normalized scores are whole numbers too and consequently normalized T scores are whole numbers as well.

Stanines

The stanines are one digit normalized scores taking values of 1 to 9. The highest stanine is 9 while the lowest is 1. Every stanine is expressed as a whole number. The mean of this scale is 5 with a standard deviation of approximately 2.

Usually, when scores are converted to stanines the following, are the distributions of the frequencies.

The lowest	4%of the cases are given stanine		1
The next	7%of cases are given		2
Next	12%	are given	3
Next	17%	are given	4
Next	20%	are given	5
Next	17%	are given	6
Next	12%	are given	7
Next	7%	are given	8
Next	4%	are given	9

An alternative way of obtaining the above is by using standard normal curve and then giving each score 0.5standard unit from the lower limit to the upper limit of that score remembering 5is the mean. Thus stanine 5 has limit for z value –0.25 and 0.25, then 4 –0.25 and –0.75 etc., while 6 is 0.25 and 0.75; and consequently, the above percentage will ensue. (i.e. for z values -∞ to –1.75 get stanine of 1

-1.75 to -1.25 get stanine of 2
-1.25 to –0.75 get stanine of 3
-0.75 to –0.25 get stanine of 4
-0.25 to 0.25 get stanine of 5
0.25 to 0.75 get stanine of 6
0.75 to 1.25 get stanine of 7
1.25 to 1.75 get stanine of 8
1.75 to ∞ get stanine 9. Interval are only inclusive of upper limits, the lower limits gets stanine below.

Advantages and disadvantages of normalized scores

The advantage of transforming into normalized scores is that the transformed distribution has a well-known form that is easily interpretable and is amenable to common statistical manipulations. Scores on different tests, if normalized and converted to the same mean and same standard deviation (e.g. normalized T score or stanines), become directly comparable, avoiding the complications involved when frequency distribution have different shapes. It is also easy to convert any normalized score to its equivalent percentile score.

However, the use of normalized scores may not be reasonable if the underlying trait has a very non-normal distribution. For example, if a score distribution is bimodal due to the presence of two types of examinees, it would not make sense to normalize the distribution. Again, if the raw-score distribution is highly skewed, small raw-score differences between extreme scores may be exaggerated or compressed by normalization. Finally, the transformed scores, with their approximately normal distribution, may lead test user to believe that the test yields "perfect normal" scores. Normalized scores based on a test with an inappropriate difficulty level or poor discrimination among examinees will not be very useful. Whatever distribution such test produces is just secondary and not so important. Thus what we are saying is that the distribution of a test should not be considered in isolation from other properties such as validity, reliability and the like of a test. All are important considerations in their own right and have their places in testing.

SUMMARY

The principal ideas, conclusions, implications presented in this unit are summarized in the following statements:

1. Raw scores can be interpreted in a meaningful manner after conversion into transformed scores in norm-referenced tests.
2. Common forms of expressing transformed scores are percentiles, standard and standardized scores and normalized scores
3. The percentile is defined as the percentage of scores falling at or below a given score. The primary advantages of percentile are that they are straightforward to calculate and that they are easy to interpret.
4. To get a standard score corresponding to any raw score, the mean of the raw score is subtracted from the raw score and the result is divided by standard deviation of the distribution. Disadvantage of standard score is that they are often expressed in negative form and decimals.
5. Standardized scores are linear transformations of standard scores and are always expressed as whole numbers and are non-negative. Linear T scores are examples of standardized scores and they (linear T scores) have a mean of 50 and standard deviation of 10.
6. The transformation to normalized scores involves forcing the distribution of transformed scores to be as close as possible to a normal distribution by smoothing out irregularities and departures from normality in the raw-score distribution.
7. The stanines are one digit normalized scores taking values of 1 to 9. The highest stanine is 9 while the lowest is 1. Stanines have a mean of 5 with a standard deviation of about 2.

Further reading:

Allen M.J. & Yen W.M. (1976) *Introduction to Measurement Theory*.
Belmont California: Wadsworth Inc.

Brown, F.G. (1970) *Principles of educational and psychological testing. 2nd ed.*
New York: Holt, Rinehart & Winston.

Ingule, F. & Gatumu, H. (1996) *Essentials of Educational Statistics*.
Nairobi: E.A. Educational Publishers.

Mehrens, W.A. & Lehmann, I.J. (1978) *Measurement and Evaluation in Education and Psychology*. New York: Holt, Rinehart and Winston.

ACTIVITY

1. The following is data for 3 students on three tests. Along with these tests scores, the mean ($\bar{X}_i$) and standard deviation (S_i) for the scores are given:

Biology $X_{11} = 35$ $X_{12} = 22$ $X_{13} = 20$ $\bar{X}_1 = 30$ $S_1 = 4$
Chemistry $X_{21} = 23$ $X_{22} = 14$ $X_{23} = 28$ $\bar{X}_2 = 21$ $S_2 = 3$
Physics $X_{31} = 39$ $X_{32} = 24$ $X_{33} = 20$ $\bar{X}_3 = 28$ $S_3 = 5$

   (i) By converting X_{13} in Biology and X_{33} in physics into z scores, find out whether student 3 had done better in physics or biology? What assumptions are you making here for the comparison to be justified?
   (ii) What is the mean z score for student 1 on all the three tests?
   (iii) What is the mean linear T score for student 2 on all the three tests?
2. Discuss why raw scores have to be converted to standard scores and normalized scores?
3. Distinguish between standardized scores and normalized scores.When are standardized and normalized equal?
4. (i) (a) Define the term percentile rank.
   (b) Define term percentile point
   (ii) Given the following distribution:

score(X_i)	1	2	3	4	5	6	7	8	9
frequency (f_i)	1	3	5	9	12	22	23	16	9

   Determine the percentile rank corresponding to each score. Transform to T scores.

5. Given that a raw score distribution on a given test is normal with mean 48 and variance 4. Complete the table below using this information on relevant values of z scores, percentile ranks, T scores and stanines:

Raw score	z score	percentile rank	T score	stanine
52				
50				
48				
44				
42				

LECTURE TEN, ELVEN AND TWELVE

MEASURES OF RELATIONSHIP

INTRODUCTION

The relationship or association between two variables is an important concept in research or any studies. It can help in prediction, given one variable and not the other, and if their relationship is known and is high enough to allow prediction.

OBJECTIVES

The learner should be able to:

1. Explain two methods of studying the relationship, one requiring stringent requirement (assumptions) while the other not so stringent requirements.
2. Compute, given two sets of data for a group, the two indexes (measures) of relationship i.e. Pearson product moment correlation coefficient and Spearman rank order correlation coefficient.
3. Interpret the computed value of the relationship
4. Draw a scatter diagram (also called scatter-plot or scatter-gram) and describe relationship it portrays in simple terms.
5. Give the properties of the indices e.g. what happens to the relationship index when the scores are linearly transformed.

Measures of Relationship

Quite often, we are interested in finding out how two variables are related to each other. The kind of question that may come to our mind is: are those who do well in mathematics, the very ones who do well in science? Or the same question put in different words would be: how is performance in mathematics related to performance in science?

When two measures are related, the term correlation (or association) is used to describe this fact. Correlation has two distinctions; that is, correlation which merely describes presence or absence of relation, and correlation, which shows the degree or magnitude of relationship between two measures.

Detecting presence or absence of relationship

Two ways of detecting the presence or absence of a relationship include *logical examination* of the two measures and plotting a *scatter diagram*.

Logical examination

This can be demonstrated by an example. Suppose one postulates that mathematics scores are related to science scores. Then suppose, that the following data linking mathematics and science scores are obtained. What would logical examination of pairs of the data suggest?

Logical examination of maths and science scores for case I

Maths scores (X)	42	54	66	78	100	120
Science scores (Y)	81	88	93	99	109	125

Examination of data for case I above, suggests that some logical relationship exists between mathematics and science scores. It may neither be necessary nor possible to get an index of degree of relationship using this method, but one can at least be confident that some relationship is present. However, suppose one was examining the data below in our case II, would one be justified in concluding that some logical relationship exists between the two sets of scores?

Relationship between maths and science scores for case II

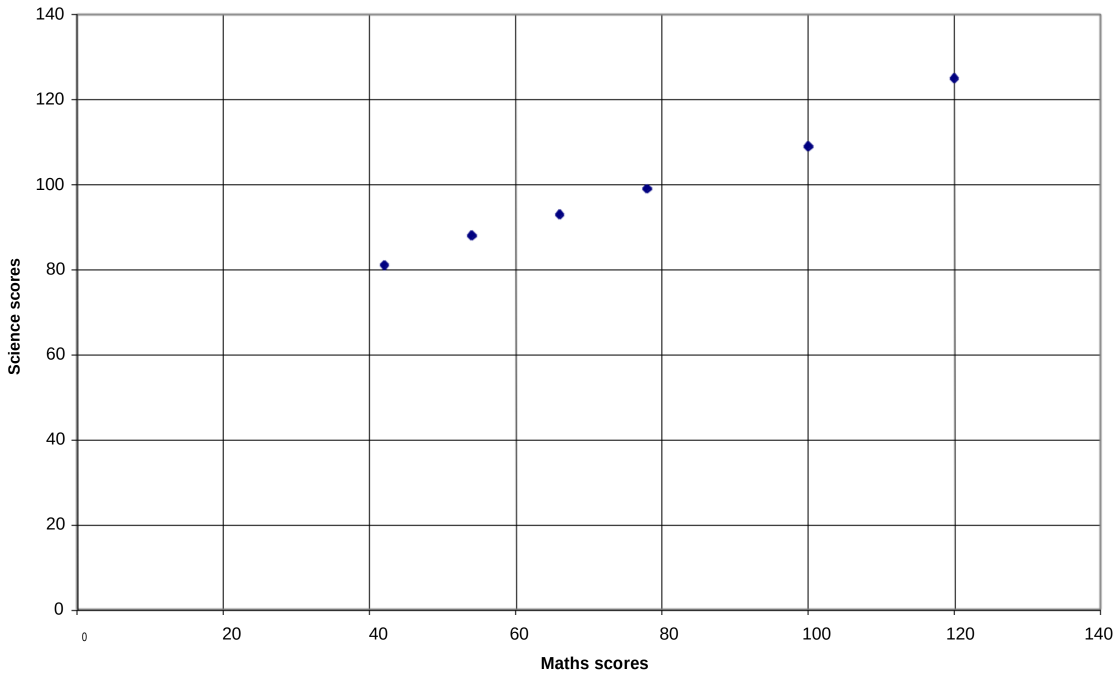
Maths scores (X)	42	54	66	78	100	120
Science scores (Y)	81	45	55	42	91	77

Examination of the data for our case II shows that there is no logic in the differences in science and mathematics scores. Hence the relationship between them cannot be defended logically. One can only conclude that there is no relationship. We next look at the relationship between two variables depicted graphically to study the kind of relationship between the two variables. This is done by means of a scatter diagram.

Scatter diagram

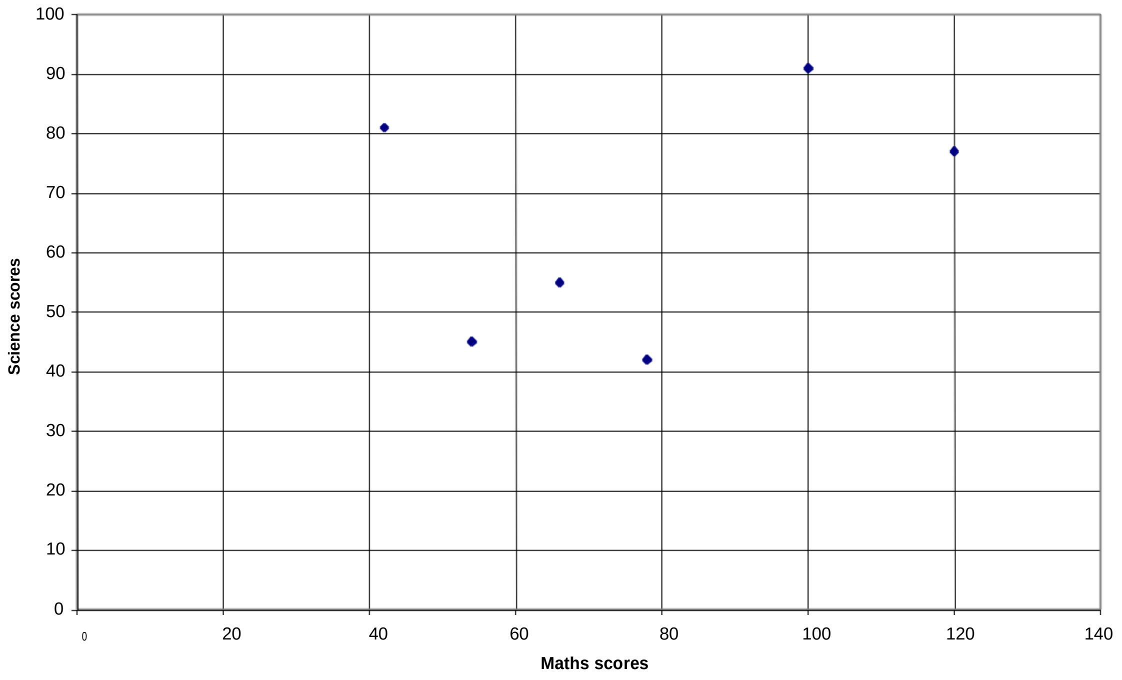
A scatter diagram is a graph of data points based on two measures (variables), where one measure defines the *horizontal axis* and the other defines the *vertical axis*. In other words, when we depict graphically a relationship between two variables, that graph (or presentation) is referred as a scatter diagram or scattergram, also called scatterplot.

Scatter diagram showing relationship between maths and science scores for case I



Observe here in case I, one could draw a straight line through the scatter diagram in such a way that it would approximate the pattern of points. The pattern of points in this case suggests a highly positive relationship. This means that as mathematics scores increase, there is a corresponding increase in science scores. Our case II, depicted below is that of, there is no systematic relationship between the two variables. The points do not show any distinct pattern. This scatter diagram suggests the absence of a relationship.

Scatter diagram showing relationship between maths and science scores for case II



Scatter diagram does not provide a very precise measure of the relationship. We definitely need a more precise measure (or index) of relationship.

Methods that provide more precise indices of relationship

Three of the methods that provide precise measures of relationship between variables are: *covariance*, *Pearson product-moment correlation coefficient* and *Spearman rank correlation coefficient.*

Covariance

Covariance provides some information on the degree of relationship between two variables by a simple averaging procedure. Let us illustrate this by a case where we want to determine the covariance between mathematics (X) and science (Y) scores using data generated from *n* students. Each of the *n* students provide two scores i.e. one mathematics score (X) and one science score (Y).

The first step in the determination of *covariance* involves obtaining the product of deviation of the two scores (X and Y) from their respective means. This in summary is $(X_i - \bar{X})(Y_i - \bar{Y})$. The procedure is repeated for all the *n* students. The products of deviation scores are then summed. The sum of the products of the deviation scores is divided by *n* (the total number of students). The quantity obtained is covariance, which a more precise measure of relationship, possibly than a scatter diagram. We shall denote covariance between X and Y by Cov. (X, Y).

Thus :

$$\text{Cov. (X, Y)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Note if a student's scores are high on both variables, X and Y, the product of the deviations i.e. $(X_i - \overline{X})(Y_i - \overline{Y})$ will be high and positive for him or her. If the student has low scores in both variables, the product of deviations will be both high and negative while the product of two negative numbers is positive.

Thus if the trend were that those who score high scores in variable X also score high in variable Y and the low scorers in variable X are still the low scorers in variable Y, taking the sum of product would result to a very high positive number. This is because, essentially, we shall be summing up only positive values.

Thus $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ would be a large positive number.

Consequently, covariance of X and Y (Cov. (X, Y)) i.e. $\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n}$ would be a positive number, implying that in such a case we would have a positive relationship between the two variables.

Supposing we have the inverse relationship; that is, those scoring high in variable X are scoring low in variable Y and *vice versa* (i.e. high X is paired with low Y and vice versa). Many students (subjects) with positive $(X_i - \overline{X})$ value i.e. deviation of X will tend to have negative $(Y_i - \overline{Y})$ value i.e. deviation of Y and vice versa. Note the sum of the product would be high negative number.

Thus $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ would be a large negative number.

Consequently, covariance of X and Y (Cov. (X, Y)) i.e. $\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n}$ would be a negative number, implying that in such a case we would have a negative relationship between the two variables.

Supposing further that the variables have no particular relationship. In other words, the pairing of the variables bear no systematic pattern such that some high values in variable X are paired with low values in variable Y and some with high values in variable X are paired with high values in variable Y. Observe some products will be negative while others would be positive. Thus taking the sum of the products: $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$, we should get a value very close to zero if not zero. Thus the quantity $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ is a measure of relationship. Note that this measure would depend on size on how many pairs of values are included in the calculation. In other words, it depends on the size of the group considered. Thus to make the quantity independent of the group considered i.e. *n,* we divide it by *n.* Note what we obtain. This is nothing else, but the covariance of the two variables, X and Y. Thus the covariance of X and Y, Cov. (X, Y) = $\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$ is a measure of relationship between X and Y. Notice that the covariance of X with itself is simply the variance of X.

Cov. (X, X)= $\frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{n} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ = Var. (X)

The Pearson product-moment correlation coefficient

The process of deviating both the X and Y values around their respective means has made the quantity, Cov. (X, Y) independent of the means, of the values. To make the desired measure of relationship independent of standard deviations of the two groups' values, one need only to divide covariance of the two i.e., Cov. (X, Y) by the standard deviation of the two variables i.e. S_x and S_y. The result is the desired measure of relationship between X and Y. It is called the *Pearson product-moment correlation coefficient* and is denoted by r_{xy}

i.e. $r_{xy} = \frac{Co\,var\,iance(X,Y)}{\sqrt{Var.(X)Var.(Y)}} = \frac{Cov.(X,Y)}{S_x S_y}$

i.e. $r_{xy} = \frac{Cov.(X,Y)}{S_x S_y}$

The above formula can be simplified and presented in a slightly different form to get what is commonly referred to as the definitional formula.

. $r_{xy} = \frac{Cov.(X,Y)}{S_x S_y} = \frac{\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$

$$r_{xy} = \frac{\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}}}$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Where X_i is the score of a person on one variable,
Y_i is the score of a person on the other variable,

$\bar{X}$ is the mean of the X-score distribution
$\bar{Y}$ is the mean of the Y-score distribution
S_x is the standard deviation of X-scores
S_y is the standard deviation of Y-scores
N is the number of scores within each distribution, and

$\sum_{i=1}^{n}$ is summation sign implying we summing up to n cases' scores.

Table showing the calculation of r_{xy} using the definitional formula

	X_1	Y_1	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
	47	42	20	400	16	256	320
	46	47	19	361	21	441	399
	27	22	0	0	-4	16	0
	8	7	-19	361	-19	361	361
	7	12	-20	400	-14	196	280
Sums							
	135	130		1522		1270	1360

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$= \frac{1360}{\sqrt{1522 \times 1270}}$$

$$= \frac{1360}{1390} = 0.978$$

Computational formula

What we have above is the definitional formula for the Pearson correlation coefficient, r_{xy}. The formula below is more convenient for calculating the correlation coefficient, r_{xy} given the raw scores X and Y.

It can be easy be shown that:

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}$$

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \sum_{i=1}^{n} \quad 2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}$$

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

Hence

$$r_{xy} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}\right]}}$$

$$= \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}}$$

This is the computational formula also else called the machine formula. The computational formula does not require the calculation of standard deviation nor the mean directly prior to the computation of r_{xy}, but directly utilizes the raw data. All the formula requires for one to calculate r_{xy} is known values of X, Y, n, XY, X^2, Y^2. The table below illustrates how to calculate r_{xy} using the computational formula.

Table showing the calculation of r_{xy} using the computational formula

	X_i	Y_i	X_i^2	Y_i^2	X_iY_i
	47	42	2209	1764	1974
	46	47	2116	2209	2162
	27	22	729	484	594
	8	7	64	49	56
	7	12	49	144	84
Sums					
	135	130	5167	4650	4870

$$r_{xy} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2][n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2]}}$$

$$= \frac{5\times 4870-135\times 130}{\sqrt{[5\times 5167-135^2][5\times 4650-130^2]}}$$

$$= \frac{24350-17550}{\sqrt{[25835-18225][23250-16900]}}$$

$$= \frac{6800}{\sqrt{7610\times 6350}}$$

$$= \frac{6800}{6951} = 0.978$$

Thus $r_{xy} = 0.978$

The two formulae are algebraically equivalent, but computational formula is more frequently used for, one no rounding when mean is not a whole number as we would round up in the definitional formula. Secondly, computational formula is more convenient and as easier to use as well, since it involves less calculation.

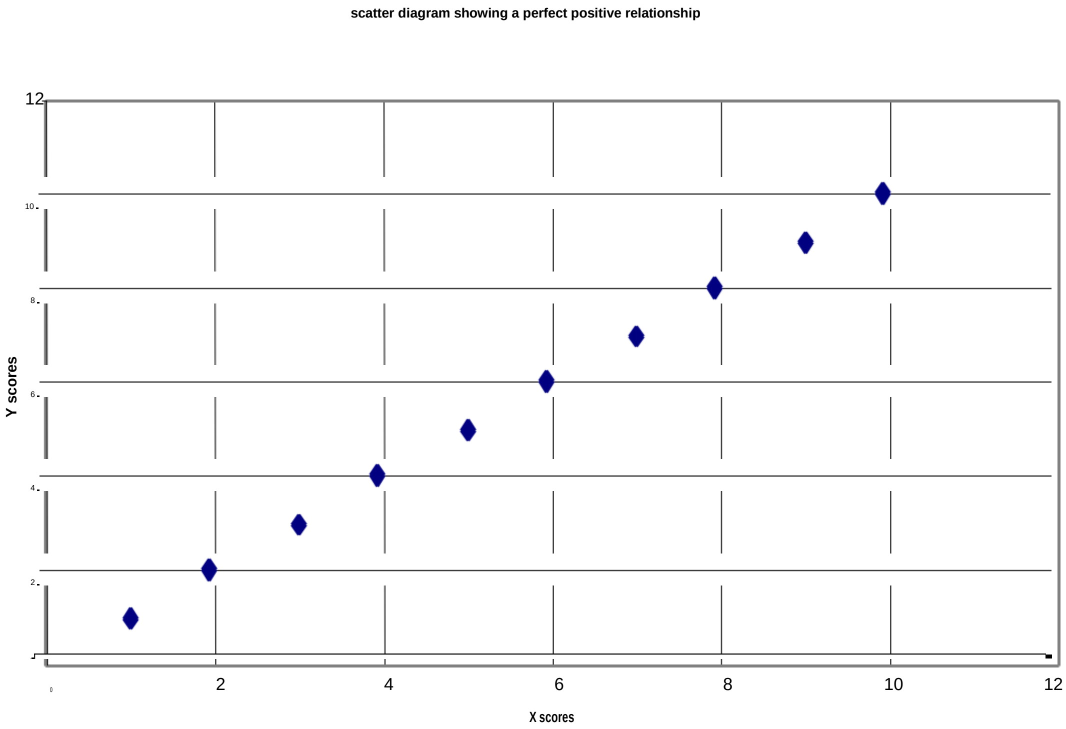
As observed and expected, the values of r_{xy} obtained by definitional formula and the computational formula are similar, i.e. 0.978.

Ranges for Values of r_{xy}:

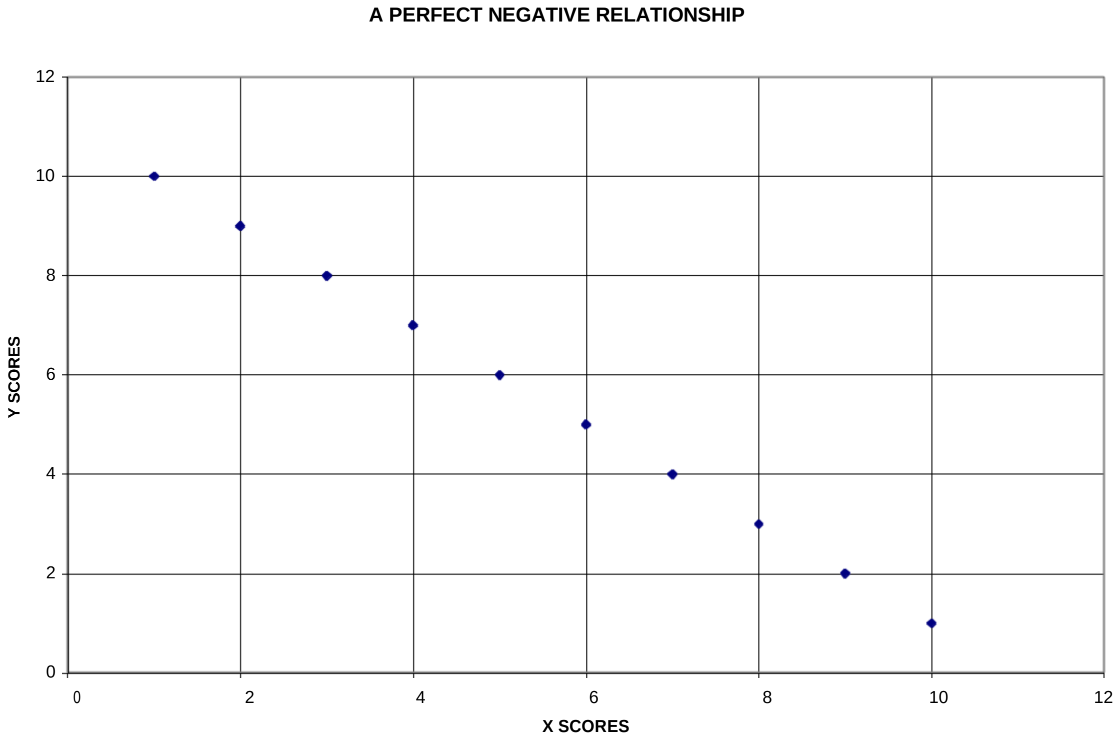
The Pearson product moment correlation coefficient, r_{xy}, varies between –1.00 and 1.00. A value of 1.00 is a perfect positive relationship. This means that an increase in one variable is accompanied by a commensurate increase in the other variable. A value of 0.00 (zero) indicates no relationship. A value of –1.00 is a perfect negative relationship. Values in between the two extremes (minimum, -1 and maximum, +1) are judged low to high depending on the size.

The scatter diagrams indicating r_{xy}'s of different sizes are shown below. The interpretation of r_{xy}, itself is relative. That is, we cannot say that a correlation of 0.5 is twice as strong as a correlation of 0.25, but only that it is stronger. This kind of ordinal thinking is only meaningful way of comparing the size of different r_{xy}'s. While there is a significant test for concluding what constitutes a correlation coefficient different from zero, tradition tells us that in social science r_{xy} (i.e. coefficients) ranging from0.6 to 0.8 indicate quite strong relationship. The size of the group being considered affects the coefficient, r_{xy}, you find a large group (i.e. large n) may have a relatively small coefficient which is significantly different from zero. We now look at the three types of scatter diagrams depicting the two extreme cases and no correlation case.

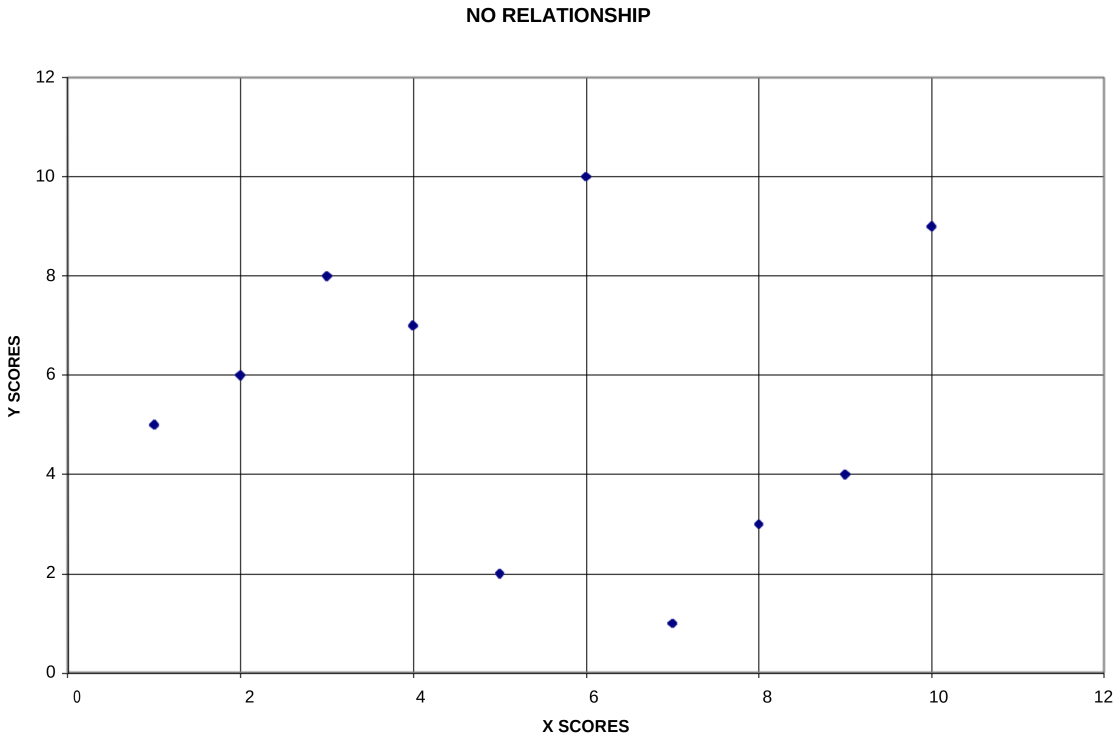
Scatter diagram showing a perfect positive relationship



A scatter diagram showing a perfect negative relationship



A scatter diagram showing no relationship



The effect on r_{xy} of applying a constant to scores

Adding, subtracting or multiplying every score in the two distributions, X and Y by a constant has no effect on the size of the correlation coefficient between X and Y. Adding, subtracting or multiplying every score in a distribution by a constant is an example of a linear transformation. In a linear transformation, the scores retain their relative positions and hence the correlation coefficient between two sets of scores, X and Y is not affected by the transformation.

Coefficient of determination

The most useful interpretation of r_{xy} relies on r_{xy}^2, which is often called *coefficient of determination.* The coefficient of determination is the proportion of variance in one of the variables, which is shared by the other. It refers to the amount of knowledge one variable provides in determining the values of the others. The coefficient of determination shows the proportion of the total variance, which is explained or shared by another variable. Thus, an r_{xy} of 0.30 between age and dependency would suggest that the two variables have in common 0.30^2 or 9 percent of their variance and 91 (i.e., 100 – 9) percent of the variance in dependency is due to variables other than age. The idea is that any variable, Y, can in a sense be defined by one or more other variables. The square of r_{xy} indicates proportionately how great a role X plays in defining Y.

Causation and correlation

The presence of correlation between two variables does not necessarily mean there exists a causal link between them. Correlation measures the strength of linear relationship. It does not mean that one variable is causing another. The only notion, of course, available to human sense is the idea of time and space. If one variable comes before the other in time, we may be in a better position to assert that it causes the others. Assertions of this kind cannot be based on the correlation between variables alone.

Even when correlation between events can be useful in identifying causal relationship when coupled with other methodological approaches, it is a dangerous and misleading test. First, even when one can presume that a causal relationship does exist between two variables being correlated, r_{xy} can tell nothing by itself about whether X causes Y or Y causes X. Secondly, other variables other than the two under consideration are responsible for the relationship.

However, while correlation does not directly establish a causal relationship, it may furnish clues to causes. When feasible, these clues can be formulated or postulated as hypotheses that can be tested in experiments, in which influences, other than those whose inter-relationships are being studied, can be controlled.

Curvilinearity

Pearson correlation coefficient is a measure of linear relationship between two variables. It is quite possible to have a relationship like poverty and age where both younger and older people are poorer than those in between. This kind of relationship is best described by a curve and is called a *curvilinear* relationship. When the association between two variables is curvilinear (non linear), it is likely that the estimates of r_{xy} will be low. It is always wise, therefore, to draw a scatter diagram before computing the correlation coefficient. It saves time and indicates the approximate form of the relationship.

Underlying assumptions of Pearson product moment correlation coefficient, r_{xy}

Four assumptions, which need be met by data of the two variables, X and Y for the index, r_{xy} to be a meaningful index in the sense that it is accurate in measuring relationship. The assumptions are:

1. The relationship between the two variables must be linear
2. The two distributions should have similar shapes
3. The scatter diagram should be homoscedastic
4. The data should be based on at least interval scale of measurement.

Concerning the linear assumption, r_{xy} is a measure of linear relationship between X and Y. If X and Y are perfectly linearly related, the points in the scatter diagram will fall on a single straight line. If we scatter the points in such a scatter diagram above and below the line in a haphazard manner and about the same distance in each direction, we obtain various degrees of basically linear relationship between X and Y. In both cases, the assumption of linearity would be met. However, if there is some evidence that the relationship between X and Y is curvilinear, then the assumption of linearity would have been violated, and the use of r_{xy} would underestimate the real magnitude of association between X and Y.

The second assumption is that both X and Y have similar distribution shapes. The values of r_{xy} cannot attain the maximum values of +1 and –1, unless X and Y distributions have identical shapes. For instance, if X is highly positively skewed and Y is negatively skewed, then these maximum values cannot be obtained, whatsoever.

The third assumption has to do with homoscedasticity or equal variance of the X and Y distributions. In simple terms, the requirement here is that the points on a scattergram showing relationship between X and Y should be uniformly distributed. No places on the scattergram (or scatter diagram) should have more points than others. The density of points on the scatter diagram should be nearly the same. Whenever the scattergram is homoscedastic, the variance of X variable is the same as the variance of Y variable.

The fourth assumption has to do with the level or scale of measurement. The Pearson product moment correlation can only be used if the data from the two variables being correlated is based on interval scale of measurement or on ratio scale of measurement. If the data is based on ordinal (ranked) scales of measurement, then the Spearman rank- order correlation coefficient, *ρ,* (rho) should be used.

The Spearman rank-order correlation coefficient, *ρ,* (rho)

If one is not very happy with the assumptions underlying Pearson coefficient, do we have another index of which we do not require our data to satisfy all those stringent assumptions (or requirements)? Put in a different way, does it mean if our data has say curvilinear trend, i.e. the relationship between X and Y is curvilinear, not linear), we cannot find the relationship of such data which is meaningful? Is there no meaningful index, which will tell us how our variables are related, despite their curvilinear trend or distribution not similar or scattergram not homoscedastic, or our data is on ordinal or nominal scale?

We have such index, which we do not have to require our data to meet the 4 or more assumptions. The assumptions for the index are minimal. The index is Spearman rank-order correlation coefficient. It is not the only one in this series of what we may refer to as Non-parametric statistics. Note, a price is paid for the minimal assumption. It seems there no free things in this world; a price has to be paid, which we shall not go into.

The Spearman rank correlation uses data that is form of ranks (i.e. ordinal data). Thus, if both of the two variables to be correlated are measured on an ordinal scale (rank-order scale), the Spearman rank coefficient is the technique generally applied. The formula for obtaining the Spearman rank-order correlation coefficient, *ρ,* (rho) is:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where ρ (rho) is the Spearman correlation index

d_i^2 is the difference in subjects' rank on the two measures (variables) squared

n is the number of scores within each distribution.

Although the Spearman coefficient is designed for use with ranked data, it can be used with interval data that have been expressed as ranks (we shall see how to express this interval data into ranked data below here):

Converting interval data to ordinal data

Table I

Scores	Position	Rank
39	1	1
38	2	2
36	3	3.5
36	3	3.5
35	5	5

30	6	6

Table II

Scores	Position	Rank
39	1	1
38	2	2
36	3	4
36	3	4
36	3	4
35	6	6
30	7	7

The first column provides interval data in the two tables above. Column two provides the positions of the scores, while their ranks are given in third column. Rank is similar to position except where some scores have ‘ties’. Ties are those scores obtain by more than one individual. Ranks of scores with ties are the mean of the ranks they would occupy if no tie existed. Thus in our case in table I, the two 36’s would have occupied the ranks of 3 and 4 if the tie did not exist. Since they tie, the mean of 3 and 4 i.e. (3+4)/2, which is 3.5 is assigned to the two scores. For table II the three 36’s, their rank is (3+4+5)/3, which is 4.

To illustrate the calculation of rho, let us again use data used to illustrate the calculation of Pearson product moment correlation coefficient with a slight modification. Since the data is based on interval measurement, it is converted to ordinal data by assignment of ranks in the third and fourth columns. The fifth column gives $|d_i|$, the absolute difference between the ranks of X and Y variables for respective subjects. The sixth column gives values of squared differences between ranks, d_i^2.

X_i	Y_i	Rank (X_i)	Rank (Y_i)	$\lvert d_i \rvert$	d_i^2
47	42	1	2	1	1
46	47	2	1	1	1
27	22	3	3	0	0
8	7	4.5	5	.5	.25
8	12	4.5	4	.5	.25
				Sum	2.5

n = 5 $d_i^2 = 2.5$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

$$=1-\frac{6\times\ 2.5}{5(5_{2}\ -1)}$$

$$=1-\frac{6\times 2.5}{5\times\ 24}$$

$$= 1 - 0.125 = 0.875$$

Interpreting correlation coefficient, ρ, (rho)

The ρ (rho) is interpreted in the same way as r_{xy}. The value of rho can never be less than –1 nor greater than +1. It equals to +1, only if each person has exactly the same ranks on both X and Y. It is –1, if there are no ties and the order is completely reversed for the two variables such that the first is the last in the other variables and so forth.

Note

1. Although the Spearman correlation coefficient formula is simpler and does not look much like the computational formula we used for Pearson correlation coefficient, it is algebraically equivalent to the Pearson when it is used with ranked data instead of the interval data.
2. Tie places are easily handled by assigning the mean value of ranks to each of the tie holders.
3. If a very large number of ties occur, however you would probably be wise to reconsider the use of Spearman (rho) coefficient, other non-parametric methods such as Kendall’s tau or chi-square may be more appropriate.
4. Ranking can be done from the smallest or largest and so forth as long as you stick to the convention you use to the end.
5. if there are no ties in the data, Spearman coefficient is merely what one obtains by replacing the observations by their ranks and then computing Pearson product moment correlation coefficient of ranks.

SUMMARY

The following statements summarize the major points of this unit:

1. When two measures are related, the term correlation is used to describe this fact.
2. Correlation has two distinctions: correlation that merely describes presence or absence of relationship and correlation, which shows the degree of magnitude of relationship.
3. A study of correlation to determine presence or absence of relation can be done through logical examination of data and examination of scatter diagrams. Methods used to provide indices of the magnitude of relationship include covariance, Pearson product-moment correlation coefficient and Spearman rank-order correlation coefficient.
4. The measure of correlation assumes only values between –1 and +1.
5. If the larger values (scores) of X tend to be paired with larger values (scores) of Y, and hence the smaller values (scores) of X and Y tend to be paired together, then the measure of correlation should be positive and close to +1. If the tendency is strong, then we would speak of a positive correlation between X and Y.
6. If the large values of X tend to be paired with the smaller values of Y, and vice versa, then the measure of correlation should be negative and close to –1. If the tendency is strong, then we say that X and Y are negatively correlated.

7. If the values of X seem to be randomly paired with the values of Y, the measure of correlation should be fairly close to zero. We then say that X and Y are uncorrelated, or have no correlation or have correlation zero or are independent.
8. The Pearson product moment correlation, r_{xy}, is obtained by dividing the covariance between two variables by a product of respective standard deviations.
9. Adding or multiplying every score in two distributions with a constant has no effect on the size of the correlation.
10. Correlation does not mean causation.
11. In order to use r_{xy}, the relationship between the two variables should be linear, the two distributions must be similar, the variance of the two distributions should be identical (homoscedastic) and data should be based on interval scale of measurement.
12. When measure is based on ordinal data, the Spearman rank order correlation coefficient, ρ (rho), should be used. The Spearman rank order correlation coefficient can be interpreted in the same way as r_{xy}.
13. Coefficient of determination, r^2_{xy}, which indicates the proportion of which Y can be accounted for or explained for by X, is more efficient index for expressing relationship between two variables.

Further reading:

Conover W.J. (1980) *Practical non parametric Statistics*. New York: John Wiley & Sons Inc.

Ingule, F. & Gatumu, H. (1996) *Essentials of Educational Statistics*. Nairobi: E.A. Educational Publishers.

Glass, G.V. & J.C. Stanley (1970) *Statistical Methods in Education and Psychology*. New Jersey: Prentice-Hall,

Johnson, R.R. (1980) *Elementary Statistics*. Mass.: Duxbury Press.

Smith, G.M. (1970) *A Simplified Guide to Statistics for psychology and Education* New York: Holt, Rinehart and Winston.

Siegel S. (1956) *Non parametric Statistics for Behavioral Sciences*. New York: McGraw-Hill Inc.

ACTIVITY

1. The following scores were obtained when a group of 11 students were tested on two tests, test A and test B

Examinee	Test A (X)	Test B (Y)
1	2	2
2	2	3
3	4	4
4	5	4
5	3	5
6	6	5

7	4	6
8	5	6
9	6	7
10	8	8
11	7	9

(a) Plot a scatter diagram for the above data (use graph paper).

(b) Compute the Pearson product moment correlation coefficient, r_{xy} between tests A and B for this group of 11 examinees.

(c) Interpret your computed value of r_{xy}.

(d) State the assumption underlying this correlation analysis.

(e) Compute the Spearman correlation coefficient, ρ (rho), for the above data.

(f) What are the major differences between these two measures of relationship (i.e. between Pearson and Spearman correlation coefficients)?

2. Suppose the following were scores of a small class in two tests, test A and test B. Test A is taken as variable X while test B is taken as variable Y.

	Test A (X)	Test B (Y)
Muchoki	5	4
Njeri	6	6
Langat	5	5
Otieno	3	2
Juma	2	3
Osoro	3	4

(a) Plot a scatter diagram for the above test scores.

(b) Compute the Pearson product-moment correlation coefficient r_{xy} between test A and test B for this class. Interpret the value of r_{xy}.

(c) Compute the Spearman rank order correlation coefficient, ρ (rho), for this class in the two tests.

(d) Which one of the two correlation coefficients would you prefer for this data? Give reasons for your choice.

3. (a) Determine if there is a logical relationship between X and Y using the data in the table below.

X	3	4	7	9	11	15	20
Y	5	4	6	4	10	5	9

(b) By means of a scattergram, say the kind of relationship between X and Y in the above data.

LECTURE THIRTEEN AND FOURTEEN

EVALUATION OR MEASUREMENT

INTRODUCTION:

Reliability and Validity of a Measurement:

Reliability and Validity:

There are certain qualities that every measurement device (test or questionnaire should possess. The measurement (or test) should be:

1. Reliable

and 2. Valid.

And these are necessarily the two important qualities of a test. Validity is the more important of the two and depends on reliability. We look at the two at a time starting with reliability. This is a wider topic and more complex than validity.

Note: the terms test, instrument (tool) and measurement are used interchangeably.

OBJECTIVES

The learner should be able to:

1. Define reliability, standard error of measurement, true score.
2. Show relationship between true score and observed score.
3. Show the relationship between reliability coefficient and correlation coefficient between observed scores and true scores.
4. Discuss the 3 methods commonly used for estimating reliability.
5. Show relationship between standard error of measurement and reliability
6. Discuss factors that may influence (increase or decrease) reliability coefficient of a test (or instrument).
7. Define validity
8. Distinguish between validity and reliability.
9. Discuss the 3 kinds of validity.

Reliability:

Reliability can be defined as the degree of consistency between two measures of the same kind. Thus a test must measure consistently if it is going to be said to be reliable. Unless a test measures consistently, little faith can be placed in the test. That is, an individual should obtain approximately the same mark on another administration of a test. Without consistency, it is like measuring with an elastic ruler. Different results would be obtained depending on how much the ruler is stretched.

The degree of consistency of a test (measurement), called reliability of a test or more correctly referred to as a **reliability coefficient of a test**, is a correlation coefficient. This reliability coefficient is **estimated** by basically 3 methods:

1. Administrating the test twice to the same group after a certain time interval has elapsed. This method is popularly known as test-retest method. A reliability coefficient is then calculated to indicate the relationship between the two sets of scores obtained.

   The coefficient may be affected by:

   i. Time interval.

ii. The variable of interest e.g. if it is a very stable variable such as mood.

Hence the method is used for stable psychological traits (e.g. aptitudes and interests). This is why it also known as measure of stability or coefficient of stability.

2. Administering equivalent forms of the test; commonly known as equivalent forms method. Two different but equivalent (also called parallel or alternate) forms of an instrument (test) are administered to the same group of individuals during the same time period. Although the questions are different they should sample the same content and they should be constructed separately from each other. A reliability coefficient is then calculated between the two sets of scores obtained. A high coefficient would indicate strong evidence of reliability that the two forms are measuring the same thing. It is possible to combine the test–retest and equivalent-forms methods by giving two different forms of the same test with a time interval between the two administrations. A high reliability coefficient would indicate not only that the two forms are measuring the same sort of performance, but also what might expect with regard to consistency over time. Hence combination is most suitable with traits (variables) known to be stable.
3. Analyzing the internal structure of the test or better known as internal-consistency methods. These methods of estimating reliability coefficient and are several require only a single administration (unlike 1 and 2 where were required two administrations) of the test(s). Here we consider one type of split half method. This involves scoring two halves (usually odd items versus even items) of a test separately for each person and then calculating correlation coefficient, which must be corrected for full test using, for instance, Spearman-Brown prophecy formula. A simplified version of this formula is as follows:

$$\text{Reliability of scores on total test} = \frac{2 \times reliabilityfor \frac{1}{2}\ test}{1 + reliabilityfor \frac{1}{2}\ test}$$

These are the common methods of estimating reliability coefficient. In practice what we do is to administer a test to a sample (a random sample) then working the correlation coefficient of the scores in each case accordingly.

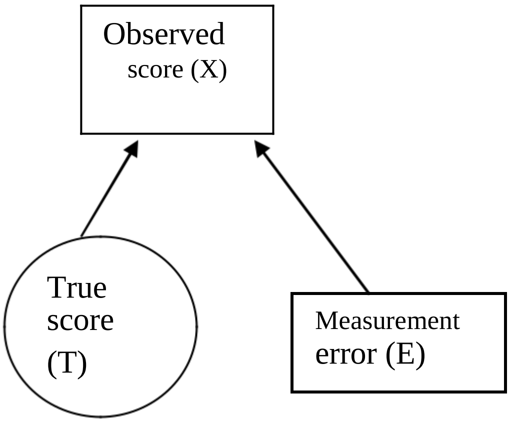
Theory of Reliability:

The theory of reliability can best be explained by starting with observed score (X). These observed scores might be conceptualized as containing various component parts. In the simplest case we think of each observed score as being made up of a 'true score' and an 'error score' such that:

$$X=T+E$$

Where X= observed score T= true score E= error score

If we had ideal situation in, which instrument measured without error at all levels (e.g. from subject side, researcher side etc.), we would obtain true score. But unfortunately in the world we live in, we cannot have those ideal situations. The best we can contend with is to estimate the true score. An estimate is better than no estimate at all. Consequently, the difference between the true score (value) and the observed score (value) is the error score or 'error of measurement'.



X=T+ E

Assumption:

Correlation between true score and measurement error is zero. Thus

Var.(X) = Var.(T) + Var.(E)

Or

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \qquad \text{(A)}$$

where

σ_X^2 = Variance of the population's observed scorers.

σ_T^2 = variance of the population's true scores.

σ_E^2 = Error variance in the population's scores.

Theoretically or operationally, reliability coefficient (r_{xx}) is defined as the ratio of the true scores and observed scores variances:

$$r_{xx} = \frac{\sigma_T^2}{\sigma_X^2}$$

we have

$\sigma_T^2 = \sigma_X^2 - \sigma_E^2$ from

A so

$$\frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2}{\sigma_X^2} - \frac{\sigma_E^2}{\sigma_X^2}$$

thus

$$r_{XX} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \qquad \text{(B)}$$

Standard Error of Measurement of Observed Scores:

We found that :

$$r_{xx} = \frac{\sigma^2_T}{\sigma^2_X}$$

and also

$$r_{XX} = 1 - \frac{\sigma^2_E}{\sigma^2_X}$$

We observe that reliability coefficient increases as error variance decreases and error variance, (or standard error of measurement, which is the square root of error variance), decreases when the observed score is very close to true score. If error variance remains constant and we increase $\sigma_X{}^2$, reliability coefficient also increases. How do we increase $\sigma_X{}^2$? This is by having a heterogeneous population (or sample or group), which ensures that the variance of the group is high.

We have:

$$\frac{\sigma^2_E}{\sigma^2_X} = 1 - r_{XX}$$

$$\sigma^2_E = \sigma^2_X (1 - r_{XX})$$

thus

$$\sigma_E = \sigma_X \sqrt{(1 - r_{XX}})$$

The observed score has a normal distribution with mean as true score , τ, and variance as error variance, σ^2_E; hence standard deviation of this normal distribution is the standard error of measurement, which is σ_E. Thus σ_E is the standard error of measurement of the observed scores. Here we are assuming each score, X, has uniform (same) standard error of measurement, σ_E or same error variance, σ^2_E, and unbiased estimator of true score, τ, is the observed score, X. This observed score is assumed to have a sampling distribution, which is normal with mean as true score, τ and variance as error variance, σ^2_E. That is, X ~ N(τ ,σ^2_E). For better understanding of this readers are referred on inferential statistics on interval estimation i.e. on confidence interval construction dealt with on statistics section.

Note:

A ‘true score’ can be defined as the score that a person would obtain if the measuring instrument (test) measured without error. The ‘true score’ can also be viewed as the mean score on infinitely many equivalent form of a test from the same domain.

Reliability Coefficients:

A reliability coefficient is nothing more than the correlation between two sets of scores obtained from the same sample of people (examinees) and used as an index of consistency of measurement.

We found reliability to be:

$$r_{xx} = \frac{\sigma^2_T}{\sigma^2_X}$$

it can be shown as

$$r_{XX} = r_{XT}^2$$

Where r_{XX} is the reliability coefficient, and r_{XT} and is the correlation between true and observed (obtained) scores. From these two above results, it is clear that reliability coefficient is always between 0 and 1. If the test is absolutely unreliable, then it has a reliability coefficient of 0; and if the test is absolutely reliable then it has a reliability coefficient of 1. The closer the reliability coefficient of a test is to 1, the more reliable it is, and the better it is.

The square of correlation coefficient indicates the proportion of shared variance between the two measures correlated. Thus from above we can say reliability coefficient of a test, r_{XT}^2, indicates the proportion of variance in test scores that represents “true” differences between individuals. This is indeed operational definition of reliability. The particular value obtained for any reliability coefficient will depend upon the groups being tested and sources of error that influence the scores in general. Details of what factors influence the reliability coefficient will be the topic of the next discussion.

As correlation coefficient does not give causality similarly reliability coefficient is a measure of the amount of inconsistency, it does not indicate the cause of this inconsistency. It tells how much, scores may be expected to vary not why they vary.

Factors Influencing Reliability Coefficients:

The mathematical method used to estimate coefficient might influence the magnitude of reliability coefficient. Other factors will also affect the reliability coefficient estimates and these are:

1. Test length
2. Item difficulty or test difficulty.
3. Group homogeneity
4. Objectivity
5. Speed.

Test Length:

Adding more items, provided that they are equally reliable, will increase the reliability of the test. Thus, an unreliable test can often be made more useful by increasing its length. With all other things being constant a longer test is more reliable than a shorter one. Note a point of diminishing returns occurs fairly rapidly. A balance point must be struck. The test should neither be too short nor be unreasonably too long.

Test Difficulty:

A test should not be too easy or too difficult. The reliability coefficient is low for too difficult and too easy tests.

Group Homogeneity:

Other things being equal, the more heterogeneous the group, the higher the reliability

This was a result we saw earlier when we dealt with formula $r_{XX} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$

Increasing the observed scores’ variance, σ_X^2 or S_X^2 increases the reliability. Increasing observed scores’ variance means having an heterogeneous group.

Objectivity:

The more subjective a measure is scored, the lower the reliability of the measure. Thus objective tests are more reliable than subjective tests with all other things being equal. This is in fact decreasing the measurement error (or variance error, σ_E^2 or S_E^2) in the above formula.

Speed:

The opposite of speed test is power test. A speeded test is defined as one in, which examinees do not have time to respond to some questions for which they know the answers; speeded tests are usually good in mathematics. While the power test is one in, which the examinees have sufficient time to show what they know (e.g. end of year examination). However, some academic achievement tests are more speeded for more examinees than for others.

The method to be used to estimate reliability coefficient should be chosen not to be a factor of speed, but rather a measure of consistency (not speed). Thus if a test is a speed test, the method to be used to estimate reliability coefficient should not be based on internal consistency. For a power test, internal consistency method can be used. Other methods of estimating reliability coefficient like equivalent form method can be used with speed tests, for unanswered items mess up the coefficient.

Summary on Reliability:

Validity is more important than reliability though both are important qualities of any test (or instrument). Validity depends on reliability. It is possible to have a reliable test, which is not valid but the converse is not true. A test, which is valid, is automatically reliable. Reliability has to do with consistency, and unless a test measures consistently, it cannot be reliable.

A test is reliable if its observed scores are highly correlated with its true scores. We found reliability coefficient of a test is square of correlation between observed and true scores, i.e.

i. $r_{XX} = r_{XT}^2$

We also observed

ii. $r_{XX} = 1 - \dfrac{\sigma_E^2}{\sigma_X^2}$

iii. $r_{XX} = \dfrac{\sigma_T^2}{\sigma_X^2}$

We observed, observed score, X is:

$$X=T+E$$

And because of assumption that error and true score are uncorrelated, we have:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 .$$

Observed score is approximately normally distributed with mean as its true score and variance as variance error of measurement. We saw three commonly used methods for estimating reliability coefficient:

1. Equivalent form:
   Reliability coefficient is the correlation between observed scores on two equivalent tests (also called parallel or alternate).
2. Test retest
3. Internal consistency or split half and correcting for full test using for instance Spearman-Brown prophecy formula.

Validity:

The most important characteristic or property of a test is its validity- that is, the extent to which the test measures what it is supposed to measure or designed to measure. This is the definition of validity. Validity is the truthfulness or soundness of the test. Does the test measure what it purports to measure? In order for test to be valid, or truthful (accurate), it must first of all be reliable.

If say a weighing machine (butcher's weighing machine) is not reliable (say 1kg is sometimes less by several gms, thus not consistent) we certainly cannot expect it to be accurate. It can be consistent, always weighing several gms less in a fixed way (say 25 gms always for every 1kg). Can we say such a balance is accurate? No, but indeed it is reliable. In other words, reliability is a necessary, but not sufficient condition for validity. Similarly, a test can be consistent (or reliable) but not valid; but it can only be valid if it is consistent (reliable).

Types of validity of tests:

Since a single test may be used for many different purposes, there is no single validity index for a test. Thus validity can be assessed in several ways, depending on the test and its intended use. The three major types of validity are:

1. Content validity
2. Criterion-related validity
3. Construct validity

Determinations of criterion-related validity and construct validity involve the calculation and examination of correlation or other statistics. Content validity, however, does not involve any statistical calculations.

Content Validity:

This type of validity is applicable when we are estimating the extent to which a school test for instance covers some field of study. We can classify content validity into two arbitrary categories:

1. Face validity
2. Logical validity or sampling validity.

Face Validity:

This is based on subjective judgment of the examiner or examinee whether the test measures the relevant trait. The subjective judgment can also be of an expert on the subject area.

Logical Validity:

This is another form of content validity and it involves the extent to which the sample of items in the test is representative of the total population of items. Note:

The question here is whether the whole content area is well represented or covered well by the items. Is any content area over-tested or under-tested?

Criterion-related Validity:

Criterion-related validity is used when test scores can be related to a criterion measure. The criterion is some behaviour that the test scores are used to predict. In other words criterion-related validity concerns the empirical technique of studying the relationship (or correlation) between the test scores (or marks) and some independent external measures (criteria). For instance, why do we have to use KCSE grades for selection or for awarding jobs? KCSE grades are predictor variables, job effectiveness or performance in university are the criterion variables.

We can distinguish between two types of criterion-related validity:

1. Concurrent validity and 2. Predictive validity

They only differ in time factor. If test scores are obtained the same time as criteria (criterion variables), we have concurrent validity. If criteria are after a period of time as predictor then we have predictive validity. Another distinction is logical rather than procedural based on purpose. For instance, we may wish to substitute an existing test for one reason or another. We would go for concurrent validity of the test using a criterion variable. Thus concurrent validity of a test would be reported, for instance, when we are looking for a test substitute i.e. the criterion for the test, which we feel, is not suitable (could be it is too subjective, or does not instill the values we wish to cultivate etc.). In predictive validity we are actually concerned with the usefulness of the test scores in predicting some future performance.

Thus predictive validity involves using test scores to predict future behaviour. A predictive validity coefficient is obtained by giving the test to all relevant people, waiting a reasonable amount of time, collecting criterion scores, and calculating the validity coefficient (i.e. correlation coefficient between test scores predictor and criterion).

A concurrent validity coefficient is a correlation between test and criterion scores when both

measurements are obtained at the same time.

Construct Validity:

A test's construct validity is the degree to which it (test) measures the theoretical construct or trait that it was designed to measure. What is a construct? A construct is also a factor or a trait and is any domain of knowledge (e.g. verbal ability, mathematical ability). Any skill or any ability can be regarded as a construct. These constructs cannot be measured directly. So if we wish to measure creativity (another construct), we need to define what skills creative people need have and then test these skills.

Thus, construct validity is important whenever a test is designed to measure some attribute or quality (construct) that people are presumed to possess. Construct validity studies attempt to answer the questions: What psychological construct (factor) is measured by a test? How well does the test measure the construct? Thus the focus is on the construct and a test indeed may measure more than one construct (the characteristic being measured). Suppose a mathematical-reasoning test contains word problems that use extremely difficult words. Some may fail to do the problems simply because their vocabulary is weak. They may know how to tackle the problem if it was explained to them in their simple language what the problem requires. This test would not only be measuring mathematical ability, but also vocabulary level.

Summary of Differences between Content, Criterion-related, and Construct Validity:

Content Validity:

Question asked: How would the individual perform in the universe of situations of which the test items are a sample? (Logical validity) How well did the test cover the content (subject matter)? (Face validity)

Evaluation: By estimating the adequacy of sampling.

Example: A classroom exam sampling the content of a given unit of the course.

Provide and use table of specification to ensure content validity.

Criterion-related Validity:

Question asked:

How well do scores on the test predict status or performance on some independent measure?

Evaluation:

By comparing scores on the test with scores on the independent (qualitatively different) measure.

Example:

Using KCSE to predict university performance. Mechanic aptitude test to predict success of a mechanic. Personality inventory to predict which driver is more likely to cause an accident. This is for substitution, or immediate use (concurrent validity). For predicting or long-time interval between criterion and predictor we have predictive validity.

Construct Validity:

Question asked:

What trait does the test does the test measure?

Evaluation:

By accumulation of evidence as to what the test does and does not measure.

Example:

Developing a test to define a trait such as intelligence or creativity.

Construct validity studies seek to determine and classify the nature of the trait or construct (factor) measured by a test.

Note:

When you use correlational relationship for prediction, the variable used to predict is called the *predictor variable* and the variable whose value is being predicted is called the *criterion variable*. Predictor and criterion variables are analogous to the independent and dependent variables of experiments respectively, with one important difference (Bordens & Abbott, 1991). Whereas you can establish whether an independent variable causes changes in a dependent variable, you can only show that predictor variables *predict* changes in criterion variables in correlational studies and hence the use of predictor variable and criterion measure (variable) in correlational studies. Whether linkage between the latter variables (i.e. between predictor variables and criterion variables) is causal remains open to speculation.

This brings out clearly the difference between correlational and experimental researches. The primary aim is predicting or studying relationship for correlational studies and establishing causality in case of experimental studies to put it in different words.

LECTURE FIFTEEN

TESTS – FUNCTIONS AND OTHER ISSUES

INTRODUCTION

Psychological variables or characteristics are best measured by tests, psychological tests. The term tests can be broad and include measures of interests, attitudes and other such variables of which we are interested in typical performance as opposed to maximum performance.

OBJECTIVES

The learner should be able to:

1. Define a ’test’.
2. State and explain the purposes of tests
3. Explain some classifications of tests
4. Distinguish between measurement and evaluation.

Psychological Tests and Inventories:

As data-gathering devices, psychological tests are among the most useful tools of educational research, for they provide the data for most experimental and descriptive studies in education. In school surveys for the past several decades, achievement tests have been used extensively in the appraisal of instruction. Because tests yield quantitative descriptions or measure, they make possible more precise analysis than can be achieved through subjective judgment alone. There are many ways of classifying psychological tests as seen earlier. One distinction is made between *performance tests* and *paper-and-pencil tests*. Performance tests, usually administered individually, require that subjects manipulate objects or mechanical apparatus while their actions are observed and recorded by the examiner. Paper-and-pencil tests, usually administered in groups, require the subjects to mark their response on a prepared sheet.

Two other classes of tests are *power* versus *timed* or *speed* tests. Power tests have no time limit, and the subjects attempt progressively more difficult tasks until they are unable to continue successfully. Timed or speed tests usually involve the element of power, but in addition, they limit the time the subjects have to complete certain tasks.

Another distinction is that made between nonstandardized, teacher made tests and standardized tests. The test that the classroom teacher constructs is likely to be less expertly designed than that of the professional, although it is based upon the best logic and skill that the teacher can command and is usually “tailor-made” for a particular group of pupils.

Which type of test is used depends on the test’s intended purpose. The standardized test is designed for general use. The items and the total scores have been carefully analyzed, and validity and reliability have been established by careful statistical controls. Norms have been established based upon the performance of many subjects of various ages living in many different types of communities and geographic areas. Not only has the content of the test been standardized, but the administration and scoring have been set in one pattern so that those subsequently taking the tests will take them under like conditions. As far as possible, the interpretation has also been standardized.

Although it would be inaccurate to claim that all standardized tests meet optimum standards of excellence, the test authors have attempted to make them as sound as possible in the light of the best that is known by experts in test construction, administration, and interpretation.

Nonstandardized or teacher-made tests are designed for use with a specific group of persons. Reliability and validity are not usually established. However, more practical information may be derived from a teacher-made test than from a standardized one because the test is given to the group for whom it was designed and is interpreted by the teacher / test-maker.

Note:

The fact that some individuals with culturally different backgrounds may not score well or highly on external tests has led to charges of discrimination against members of underprivileged. The case has been made that most of these tests do not accurately predict academic achievement because their contents are culturally biased. Efforts are being made to develop culture-free tests that eliminate this undesirable quality. However, it is extremely difficult to eliminate culture totally and develop one test that is equally fair for all. Since knowledge has developed within a culture, it is virtually impossible to test knowledge without bringing an aspect of culture into it. Think even when measuring, we have to give measurement in a language (English or Kiswahili), English culture or Kiswahili culture has to come in even in our thinking and interpretation of our thoughts. Measurement has been mentioned here for in tests we measure learning outcomes or potentialities.

Interest Inventories:

Interest inventories attempt to yield a measure of the types of activities that an individual has a tendency to like and to choose. There are instruments to measure interest.

Interest blanks or inventories are examples of self-report instruments in which individuals note their own likes and dislikes. These self-report instruments are really standardized interviews in which the subjects, though introspection, indicate feelings that may be interpreted in terms of what is known about interest patterns.

Personality Inventories:

Personality tests measure emotional patterns, motivations, values, and other enduring features of one’s psychological makeup. Personality tests can measure attitudes and opinions – that is, the individual’s orientation to or assessment of other persons or things. Such attitude measures may help to draw inferences about personality, since such attitudes may reveal something about one’s perceptual style.

Personality tests can range from single items to elaborate, multi-scale *instruments*. The widely used Internal-External scale (I-E scale) developed by Rotter, illustrates the multiple-item, personality questionnaire. Rotter thought behaviour depended on the belief that it will yield a reward. According to Rotter, individuals differ in their generalized expectancies, that is, their belief about locus (source) of control of reinforcement. I-type people, or “internals”, believe strongly that they control their own fate and E-type people, or “externals” believe that they are not in control of their own fate. The remaining people fall between these two extremes. Rotter’s I-E scale consists of a 29-item paper-and-pencil test (including 6 filler items designed to conceal the purpose of the test and lower reactivity). Each item consists of a pair of statements of belief about the locus of control. The subject selects the statement from each pair that most agrees with his or her own general point of view and receives one point for each answer scored in the external direction.

Internal consistency (inter-item reliability) for the I-E scale has ranged in the .60s and .70s. Test-retest reliability (interim period of one to two months) has ranged from the .50s to the .70s. The I-E has correlated negatively with measures of social desirability. That is, the external locus of control belief statements appear less socially desirable than the internal ones. To test its construct validity, researchers have given the I-E to people known independently to differ on the construct of alienation. Rotter's theory predicts that externals would feel more alienated or powerless. Compared to externals, internals, as measured by the I-E, show more signs of being actively aware of and involved in their environment (e.g., concerned about their health and doing something about it, activists etc.).

Personality questionnaires can be constructed in different ways. One approach selects items depending on how well they agree with each other. This approach employs *factor analysis* to make sure that a common factor underlies all of the items of the scale. If an item fails to agree with this common factor, it is excluded in favour of one that does. Factor analysis has shown the 24 yes/no extraversion (E) items of Eysenck Personality Inventory (EPI) all share something in common with the other items of this scale (Dooley, 2004).

An alternative to factor analysis, the **empirical criterion approach**, selects test items according to their ability to discriminate previously identified groups. For example, the developers of the Minnesota Multiphasic Personality Inventory (MMPI) used the empirical criterion approach.

The empirical criterion approach has opened the MMPI to some criticisms. A subject's scale score takes its meaning from its relation to the criterion group's score. Thus being the case the culture and time has to be considered. For instance, the original group may have been tested in 1940s and may have little resemblance to 2000s (this millennium) when computer has dominated every aspect of life.

Personality scales are usually self-report instruments. The individual checks responses to certain questions or statements. These instruments yield scores which are assumed or have been shown to measure certain personality traits or tendencies.

Because of individuals' inability or unwillingness to report their own reactions accurately or objectively, these instruments may be of limited value. Part of this limitation may be due to the inadequate theories of personality upon which some of these inventories have been based. At best, they provide data that are useful in suggesting the need for further analysis. Some have reasonable empirical validity with particular groups of individuals but prove to be invalid when applied to others. For example, MMPI (Minnesota Multiphasic Personality Inventory), initial version proved valuable in yielding scores that correlate highly with the diagnoses of psychiatrists in clinical situations. But when applied to college students, its diagnostic value proved disappointing.

The tendency to withhold embarrassing response and to express those that are socially acceptable, emotional involvement of individuals with their own problems, lack of insight-all these limit the effectiveness of personal and social-adjustment scales. Some psychologists believe that the projective type of instrument offers greater promise, for these devices attempt to disguise their purpose so completely that the subject does not know how to appear in the best light.

Projective Devices:

A projective instrument enables subject to project their internal feelings, attitudes, needs, values, or wishes to an external object. Thus the subjects may unconsciously reveal themselves as they react to external object. The use of projective devices is particularly helpful in counteracting the tendency of subjects to try to appear in their best light, to respond as they believe they should.

Projection may be accomplished through a number of techniques:

1. *Association*. The respondent is asked to indicate what he or she sees, feels, or thinks when presented with a picture, cartoon, ink blot, word or phrase. The

Thematic Apperception Test, the Rorschach Ink Blot Test, and various word-association tests are familiar examples.

2. *Completion*. The respondent is asked to complete an incomplete sentence or task. A sentence-completion instrument may include such items as:

   My greatest ambition is
   My greatest fear is
   I most enjoy
   I dream a great deal about
   I get very angry when
   If I could do anything I wanted it would be to

3. *Role*-playing. Subjects are asked to improvise or act out a situation in which they have been assigned various roles. The researcher may observe such traits as hostility, frustration, dominance, sympathy, insecurity, prejudice- or the absence of such traits.
4. *Creative or Constructive*. Permitting subject to model clay, finger paint, play with dolls, play with toys, or draw or write imaginative stories about assigned situations may be revealing. The choice of colour, form, words, the sense of orderliness, evidence of tensions, and other reactions may provide opportunities to infer deep-seated feelings.

Just like good tests, good inventories need to have a high degree of both validity and reliability.

Note:

Kuder-Richardson formula. This formula is a mathematical test that results in the average correlation of all possible split half correlation (Cronbach, 1951).

Economy:

Tests that can be given in a short period of time are likely to gain the cooperation of the subject and to conserve the time of all those involved in test administration. The matter of expense of administering a test is often a significant factor if the testing program is being operated on a limited budget.

Ease of administration, scoring, and interpretation is an important factor in selecting a test, particularly when expert personnel or an adequate budget are not available. Many good tests are easily and effectively administered, scored, and interpreted by the classroom teacher, who may not be an expert.

Interest:

When psychological tests are used in educational research, one should remember that standardized tests scores are only approximate measures of the trait under consideration. This limitation is inevitable and may be ascribed to a number of possible factors:

1. Errors inherent in any psychological test – no test is completely valid or reliable.
2. Errors that result from poor test conditions, inexpert or careless administration or scoring of the test, or faulty tabulation of test scores
3. Inexpert interpretation of test results
4. The choice of an inappropriate test for specific purpose in mind

Finding Self-Report Measures. New measures are difficult to come by. Nonetheless, although researchers can develop new measures for their constructs which is not an easy task, they should first search for the best existing tests for obvious reasons. Using existing measures can save large amount of test development time. Moreover, using existing measures improves the comparison of different studies. When studies use the same measure, differences in their outcomes can be traced to design and sample differences rather than to measurement differences.

Definition of ‘test’:

A test is a systematic procedure for measuring a sample of behaviour (psychological variable). Systematic procedure indicates that a test is constructed, administered, and scored (or marked) according to prescribed rules or laid down rules, which must be followed to the letter or absolutely.

Test items are systematically chosen to fit the test specifications, the same or equivalent items are administered to all persons (examinees) and the directions and time limits are the same for all persons taking the test. The use of predetermined rules [or marking scheme] for evaluating (scoring) responses assures agreement between different persons who might score (mark) the test, in other words consistency or reliability is ensured consequently.

Using standard procedures ensures comparability among the examinees and ensures there is uniformity in all aspects you can think of. The test should not favour any individuals or any group of individuals unfairly. A test should not have any kind of bias.

A second important term in the definition is behaviour. In the strictest sense, a test measures only test-taking behaviour. That is, the responses a person (examinee) makes to the test items. Here we are talking about psychological variables and as we know these cannot be measured directly, rather we infer the characteristics (trait) from his or her responses to the given test items. We have to measure their manifestations since they are not tangible.

If the behaviour exhibited (manifested) on the test adequately mirrors the construct (trait) being measured, the test will provide useful information. Here we are talking about validity of the test, i.e. the test measuring what it is supposed to measure. If the test does not adequately reflect the underlying characteristic, inferences made from test scores will be in error for validity is important.

A test contains only a sample of all possible items. No test is so comprehensive that it includes every possible item. No test is so comprehensive that it includes every possible item that might be developed to measure the behaviour domain [or population or universe]; e.g. a driver’s test will not test you how to drive at night, or on a slippery wet road or when raining very much. Thus any particular test is better thought of as a sample of all possible items.

Because a test contains only a sample of all possible items, two problems arise.

1. We must ensure that the questions or items represented on the test are a representative sample of all-possible questions or items. [Validity]
2. Would an examinee get the same score if he were given a different set of sampled items from the same domain? [Reliability].

A test is a measuring instrument. Thus we need to define measurement.

Measurement is assigning of numbers to individuals in a systematic way as a means of representing the properties of the individuals such that those with more of the property you are measuring will score more, those with less will score less.

Difference between Measurement and Evaluation:

Measurement answers the question, how much? That is, measurement provides a description of a person’s (examinee’s) performance. It does not provide judgment. That is, it says nothing about the worth or value of the performance. If we put value or worth or judgment on it, then we are evaluating. We are going beyond description. We are attempting to answer the question how good? This is evaluation. A mark or score like 40 out of 50 is measurement. If we say it is B+ then this is an evaluation, since judgment has been made on the value of the mark or score in terms of how good. That is, objective description here is a measurement, while subjective judgment of quality is an evaluation.

Uses of Tests:

Explicit uses of tests:

1. **Selection:**

Selection is done in academic setting, in business and other sectors offering jobs where there are more qualified applicants than job opportunities. That is, in the selection situation there are more applicants than can be accepted (or employed or hired), and a decision has to be made on whom to accept. The role of the test is to identify the most promising applicants (or candidates or examinees) i.e. those with the greatest probability of success. In the simplest case, the decision is either to accept or reject. In Kenya, for university entrances, there are clear-cut off (or cut off points), which are strictly adhered to. Once laid down, one cannot go to complain. Hence we seem to have very little interest on those who are rejected (or left out). Social economic status, poor health, poor facilities and background or other adverse factors may contribute to a person being left out. Many such factors are assumed uniform for all. In other words, nobody is favoured is the assumption yet we know this is not true.

2. **In Placement:**

There are several individuals and several alternative courses of action for instance, in universities there are several departments and each has its requirements. In general each person is to be assigned to a program using certain criteria.

3. **Diagnosis:**

It involves comparing an individual's performance in several areas in order to determine relative strengths and weaknesses. Generally, diagnostic procedures are instituted when an individual is having difficult in some area. Once the areas of disability are identified, a program of remediation can be undertaken. For example, If a child has problem in reading or doing word problems in mathematics, you may give a test consisting of phonetic, word meaning (vocabulary), sentence meaning, paragraph meaning and reading rate, so as to identify what particular weaknesses or strengths of the child need appropriate action.

4. **Hypothesis Testing:**

In psychological research, tests are often used for hypothesis testing. And what is a hypothesis? This is dealt with here briefly (for details see the appropriate section on this). In brief a hypothesis is a speculative statement, or an educated guess, which you may wish to establish whether to accept or reject. For instance, we can manipulate our subjects in a certain way (varying may be the degree of manipulation) and then we try to find the effect of the manipulation. We give a test to find out the effect of the manipulation. This is a type of experimental study or design. In a correlational study (design) we have cases of natural manipulation. In a correlation study we may look at the performance at a certain time or under certain conditions. We study what has taken place and then we make inferences. Using varies methods like keeping other variables constant or eliminating them analytically or otherwise, we are able to study the effect of the variable manipulated.

Tests can also be used for hypothesis building. We may find a difference in performance in 80's and 90's and then we go on to hypothesize what could be the reason may be a drop in socioeconomic status, 8-4-4 educational system or a combination of these and others.

Psychologists or educators (even lay-people) use tests to make a lot of deductions or build hypotheses. For instance, Muthoni got a very good division I in the O-level examination, but failed to obtain university entrance after A-level. Why? Muthoni went to do science for A-level because of parental pressure. Her father wanted her to be a doctor, but she did not have much interest in sciences (Biology and the like). Muthoni could have done very well if she took Arts (Humanities). Or Muthoni may have lost her father just before the exams and this traumatized her too much beyond recovery and this indeed may have contributed to her poor performance in the A-level exams.

5. Another use of tests is in **Evaluation:**

Formative evaluation and summative evaluation:

A teacher can use test to find not only the weak students but also his weaknesses or topic not understood well etc. Thus classroom examinations and tests are usually used to evaluate the instructional method or the teacher.

All of these uses involve some decision. In selection, the decision is whether to accept or reject an applicant. In placement, where does the candidate fit best in terms of ability and skills while in diagnosis, which remedial treatment is to be used after finding out the weakness? In hypothesis testing, usually using statistics you need to establish (reject or accept) the hypothesis. In evaluation, what grade to give to a student or how effective is the procedure, effectiveness has to deal with summative evaluation, or evaluation done at the end while what is done at the beginning (e.g. to check entry behaviour) is formative evaluation.

We know how seriously we take tests. We belong to a culture, which overrates exams. You get a lot of respect if you are an A student, division I, first class or Ph.D. scholar. If you do your tests badly, you seldomly (rarely) get a chance of saying why you obtained a low score.

Research on tests shows ‘ability’ is important in doing well in a test, but accounts for less than 50%. Other factors do count like difficult of items, quality of instructions, personality variables e.g. socioeconomic status, linguistic variables etc.

LECTURE SIXTEEN, SEVENTEEN AND EIGHTEEN

CLASSIFICATION OF SCHOOL TESTS:

INTRODUCTION

There are other variety of ways in which tests can be classified especially classroom tests or teacher made tests.

i. One type of classification is based upon the type of items format used – Essay (subjective) versus objective.

ii. Another classification is based upon the type of stimulus material used to present the problems to the students (examinees)- verbal and nonverbal

iii. A classification by purpose and here we have several categories:
   a. Maximal performance versus typical performance
   b. Formative versus summative evaluation
   c. Norm referenced tests (NRT) versus criterion-referenced tests (CRT).

iv. Other classifications:
   a. Standardized and nonstandardized tests.
   b. Group and individual tests.
   c. Performance tests and paper-and-pencil tests.
   d. Power tests and speeded tests.
   e.

OBJECTIVES

The learner should be able:

1. Discuss several classifications of tests.
2. Distinguish among several of these types of tests.
3. Discuss issues in tests such as ability tests, power tests, achievement tests, criterion referenced tests and others.
4. Explain the purposes of test blueprint.
5. Discuss the new development in the test blueprint.
6. Discuss differences between essay and objective tests.
7. Discuss the process involved in item analysis.
8. Explain how to obtain item difficulty and item discrimination and their interpretations.

Classification by Item Format:

The two major categories here are as seen above:

1. Essay and 2. Objective

Essay Type:

Essay questions are subdivided into three major types

1. Extended (or Discussion) response
2. Restricted response
3. Oral.

Extended Response:

This also referred to as *discussion* type. Here the question is very much open ended (unstructured). No restriction is given. Most university questions in many departments are of this type. Example:

a. Discuss what is a system

b. Discuss what is a scientific method (approach)

c. Discuss the Information Processing Model of (Memory).

Restricted Response:

Here the student (examinee) is more limited in the form and scope of his answer because he is told specifically the context that his answer is to take. Example:

d. Give the three advantages and three disadvantages of Essay tests

e. Give the three advantages and three disadvantages of multiple-choice items

f. Distinguish between Classical conditioning and Operant conditioning

g. Distinguish among aptitude tests, achievement tests and ability tests

(v) Distinguish among Memory, Learning and Insight.

We can refer to these as short answer essay tests.

Oral Examination:

Also is called *viva, viva voce* or *defence.* This is usually done after writing a dissertation or a thesis for an advance degree, a masters or doctorate degree. Essentially, it is to find out how well the candidate has linked theory and practice in solution to his problem, and very important to see whether indeed he/she is the one who wrote that thesis or dissertation.

Oral examinations are also done in languages as well as in lower school where pupils have not mastered writing skills. Outside school setting, oral tests are administered in churches for baptism, confirmation and such membership promotion.

Objective type:

Objective type item can be subdivided into four major types:

1. Short-answer
   - i. Single word, symbol, formula
   - ii. Multiple words or phrase
2. True-false (right-wrong, Yes-No)- dichotomous case
3. Multiple choice
4. Matching.

- i. One correct answer
- ii. Best answer
- iii. Analogy type
- iv. Reverse type

Others are substitution, incomplete (blank to fill) etc.

Maximal Performance Tests:

Test takers (examinees) attempt to make the highest possible score. The goal is to measure the upper limits of examinee’s abilities. Classroom tests are in this category and are example of achievement tests. Others in this category of maximal performance tests are aptitude tests and ability tests. Note these are not mutually exclusive. A particular test may serve more than one of these purposes.

Typical Performance Tests (or Measures):

These assess somebody’s reaction or behaviour. Here the concern is not maximal performance but rather reaction or behaviour (typical reaction or behaviour) such as liking of courses, others in this category are measures in attitude, interest and personality, and are best assessed by *questionnaires* mainly.

For maximal performance tests we have basically 3 categories:

1. Achievement tests
2. Ability tests
3. Aptitude tests

Achievement Tests:

These are designed to measure the knowledge and skills developed in a relatively circumscribed area (domain) (Brown, 1976). This area may be as narrow as one day’s class assignment (e.g. computing median, variance etc.) or as broad as several years’ study (as KCSE examinations). In every case, however, we are attempting to measure what a person knows or can do at a particular point in time. His or her best performance in a test, examining what has been learned as a result of a particular course or experience or a series of experiences.

Aptitude Tests:

This is a test for giving your potentialities or what you are capable of doing from your formal or informal experiences. Thus aptitude tests indicate the probability that certain other behaviours will be acquired or learned. We consider a test to be aptitude test if:

1. It measures the results of general and incidental learning experiences.
2. Its frame of reference is toward the future.

This is in contrast to an achievement test, which measures learning from relatively specific experiences, and focuses on the past learning. Thus aptitude test predicts what can be learned in the future. Thus it measures the ability to acquire certain behaviours or skills given appropriate opportunity.

Note: **Aptitude versus Achievement Tests.** One complaint against aptitude tests holds that they are confounded with achievement as seen. We sometimes assume that aptitude tests measure native or potential ability to learn. But most such tests give points for past learning, such as knowledge of vocabulary. Thus, tutoring may increase one's apparent aptitude, thus giving unfair advantage to people who are able to pay for such preparation. Achievement tests, in contrast, assess one's level of acquired knowledge or skill without pretense that all candidates have had equal opportunity to acquire the skills.

Ability Tests:

Indicate the power to perform a task. Ability tests measure present status. In this category we have performance test (or practical test) like driving or tuning an engine, playing piano and swimming. Intelligence tests could be classified here. Some little details about ability tests are in order:

Ability tests range from those that tap the general mental ability that we refer to as intelligence, to those that tap specific abilities, such as spatial visualization. Measures of general intelligence have broadest application in that they are used in educational, clinical, and work settings as aids in making a wide variety of decisions. (Murphy & Davidshofer, 2005)

Intelligence tests are usually administered to a single individual by a trained psychologist. Although this type of test dominated the early history of intelligence testing, today, standardized tests that can be administered in groups are much more common.

Historically, the Stanford-Binet has been regarded as one of best individual tests of a child's intelligence available. Others are Wechsler Intelligence scale for children, The Wechsler Preschool and Primary Scale of Intelligence, and The Kaufman Assessment Battery for Children (K-ABC) is used widely in U.S.

Binet's original tests were developed expressly for children. David Wechsler developed comparable tests for adults as well. He has for children as said above.

The K-ABC is a psychometrically sophisticated battery that was designed to reflect current thinking in cognition and neuropsychology; it was also designed to minimize cultural bias. Note Hawthorne effect and instrumentation can be a threat here.

Ability tests on groups are very common, there can be referred to as group tests while intelligence tests as individual tests. In Kenya we have KCPE, KCSE as group tests. In U.S. SAT (Scholastic Assessment Tests), GRE, TOEFL- an American test administered to foreigners wishing to go for studies in America, and these are all group tests.

The SAT is designed to assist in undergraduate admissions and placement decisions like our KCSE. The GREs, which are administered by Educational Testing Service, serve a similar function in graduate admissions and placement decisions, and they are also used in assigning scholarships, fellowships, and traineeship. GREs include both aptitude test (GRE general), and advance achievements tests (GRE subject tests) in several disciplines. (Murphy & Davidshofer, 2005)

Norm-referenced versus Criterion-referenced Tests:

The distinction between the two is on how test scores are interpreted. In norm-referenced tests, an individual's scores are interpreted by comparing them to those of other people in some comparison (peer or norm) group. In criterion-referenced tests, the concern is *mastery* of the content regardless of the performance of the other examinees. In norm-referenced tests scores usually may need to be transformed for comparability or meaningful interpretation but in criterion-referenced tests, scores indicate proficiency or mastery or competency.

Thus a score in criterion referenced test is a meaningful number representing the level of mastery, while in norm referenced test a (raw) score may not carry much meaning unless it is converted to a standard scale to show its relative position compared to other scores. In other words, transformation such as conversion to Percentile ranks, standardization and normalization are justifiable among norm-referenced tests. A score in criterion-referenced test is a meaningful score and should not be subjected to such transformation.

The difference between criterion-referenced tests and norm-referenced tests is just theoretical. The number 1 student (or candidate) can be seen as the criterion or standard, thus having realized the perfect score or has mastered perfectly. If all have mastered equally they all would be number 1. But we know this is not the case in practice. In other words, is norm-referenced tests and criterion-referenced tests are more theoretically different than they are practically.

From another perspective, we should realize that when we are ranking or finding position of a candidate, we do this according to their performance (mastery). In both, we are talking about mastery, of course from a different angle. As long as we are concerned about mastery then both type of tests (criterion referenced and norm referenced tests) are going to have more in common to an extent that they are hardly different. They end up doing the same thing. Practically they are the same but theoretically or philosophically different. You do not talk of transforming score in criterion-referenced tests, only in norm-referenced tests. In criterion-referenced tests, a score is meaningful.

Planning a Test:

Good tests a have to be planned systematically and in a professional way so that the goals of instruction (objectives), the teaching strategy to be employed, the textual material, and evaluative procedure are all related and considered in a meaningful way.

A teacher should have enough time to plan for his test accompanied by serious thought and research to come up with a good test. If you rush to make up the last minute test, the test is bound to be poorly worded, ambiguous, and multiple choices may have no solution or more than one answer just to mention a few shortcomings. Ideally, to minimize deficiencies noted earlier, other teachers or experts should review every test critically; (we say minimize, for you will not get a perfect test). The idea is to strive towards a perfect one. The closer you are to a perfect one the better.

Possibly the only way that the teacher can have confidence in his test is to have it prepared in sufficient time to permit a critical independent review by possibly his professional colleagues if not by an expert. And this requires a teacher to adequately plan his test and write his test items in advance. Also important is to tell the examinee well in advance the day of the test and preferably the format so they can have enough time to prepare for it.

Content-Related Evidence of Validity

Consider a test that has been designed to measure competence in using English language or mastery of computation of variance, a topic in statistics. How can we tell whether the test does in fact measure that achievement? First, we must reach some agreement as to the skills and knowledge that comprise correct and effective use of English or steps involved in calculation. If the test is to be used for mastery we must agree who are the masters and what they have mastered; i.e. what skills do they have? In other words, if the test is to be used to appraise the effects of classroom instruction, we must specify the subset of skills and knowledge that have been the objectives of that instruction. Then we must examine the test to see what skills, knowledge, and understanding it calls for. Finally, we must match the analysis of test *content* with analysis of *course content* and *instructional objectives(or cognitive processes, such as skills and so forth)* to see how well the former represents the latter; i.e. how balanced is content as far as instructional objectives (cognitive processes) are concerned. Are we over-testing or under-testing in any instructional objective as it relates to each content area? There should be a good representation to the extent that our objectives or processes which we have accepted as goals for our course are adequately represented. Here we are talking about sampling. How well are the content areas and instructional objectives or cognitive processes represented? Are all skills, knowledge and understanding tested in all areas?

In a sense, *content* is what the student works with; process is what the student does with the content. The term content-related validity refers to an assessment of whether a test contains appropriate content and requires that appropriate processes be applied to that content.

To either assess the content validity of an existing test or construct a test that measures a particular set of contents and processes requires that we specify the contents and processes (instructional objectives) to be measured explicitly. This explicit statement of what a test is intended to measure is called a *test blueprint*. The correspondence between the test blueprint and the definition of the trait to be measured is the content validity of the test.

Preparing a Test Blueprint

A test blueprint (also called a table of specification for the test) is an explicit plan that guides test construction. The basic components of a test blueprint are the specifications of cognitive processes (instructional objectives) and the description of content to be covered by the test. These two dimensions need to be matched to show which process relates to each segment of content and to provide a framework for the development of the test. It is useful for test constructor, in planning the evaluation of a unit, to make a test blueprint that includes not only the cognitive processes (instructional objectives) and the content but also the method or methods to be used in evaluating student progress toward achieving each objective.

When evaluating a test for content validity you would use the same steps of domain definition, but the validity question would relate to whether the test matches your domain rather than to serve as a guide for item writing.

Relative Emphasis of Content Areas and Process Objectives:

The proportion of test items allocated to each content area and to each cognitive process (instructional objective) should correspond to the instructional emphasis and importance of the topic. The decision-making process involved is subjective, but the test user should ensure that the test has maintained an appropriate balance in emphasis for both content and mental processes. Allocating a different number of items to each topic and cognitive process (instructional objective) is the most obvious way of weighting topics and processes on the test. The weighting of each cell (content area and cognitive process) will depend on what the teacher considers to be important. The factors to consider when deciding on the weighting are age and educational level of the students, ability level of the students and so forth (Thorndike, 1997, p. 136-137).

It is very important to be systematic in the way you plan your tests. The table of specification is a good tool to use and it need be used, as indicated above.

Table of specification or Test Blueprint:

This is like a blueprint used by engineers for machines, buildings, bridges etc. or pattern used by tailors when making a suit. The purpose of the table of specification or test blueprint is to define as clearly as possible the scope and emphasis of the test and relate the (educational) objectives or cognitive processes to the content, as said above.

The area to be tested must first be limited further (or be expounded further into preferably topics or content areas). The test must be balanced and should reflect the objectives of instruction or mental processes. In other words the test must be a valid measure of pupils' knowledge, skills, and other attributes that the teacher sought specifically to teach. By ensuring that the test adequately covers the content and is directly related to the objectives (processes), the teacher has satisfied the most important criterion of the test, the content validity. In other words, the table of specification or test blueprint is used to ensure that the test is content-valid i.e. ensuring both content and instructional objectives are related in a meaningful way. Thus a table of specification is a two-way grid, consisting of the topics and educational objectives or (cognitive processes) as shown below:

Educational Objectives (Bloom's Taxonomy)

	Knowledge	Comprehension	Application	Analysis	Total
Course content					
Validity	5	2	4	4	15
Reliability	5	5	3	2	15
Item analysis	4	4	2	0	10
				Total	40

This is a 40-item test, testing those four areas of educational objectives on the 3 content areas as seen above. See later development in this area below entitled here as instructional objective latest development:

Instructional Objective – (Latest Development)

Several decades ago a group of experts in educational evaluation led by Benjamin Bloom set out to improve college and university examinations. Bloom and his colleagues developed a *taxonomy* or classification system of educational objectives. The objectives were divided into three domains: cognitive, affective, and psychomotor. In real life, of course, behaviours from these three domains occur simultaneously. While students are writing (psychomotor), they are also remembering or reasoning (cognitive), and they are likely to have some emotional response to the task as well (affective). Thus it is impossible to separate the three.

Technically, all are controlled by the brain and are inseparable; they are intertwined or related. They depend on each other as indicated above.

The Cognitive Domain. Six basic objectives are listed in Bloom's taxonomy of thinking or cognitive domain.

1. *Knowledge*: Remembering or recognizing something without necessarily understanding, using, or changing it.
2. *Comprehension*: Understanding the material being communicated without necessarily relating it to anything else.
3. *Application*: Using a general concept to solve a particular problem.
4. *Analysis*: Breaking something down into its parts.
5. *Synthesis*: Creating something new by combining different ideas.
6. *Evaluation*: Judging the value of materials or methods as they might be applied in a particular situation.

Some subjects like mathematics, do not fit this structure very well (Woolfolk, 2004). First three (K, C, A) knowledge, comprehension, and application considered lower level and other categories considered higher level.

The new version retains the six basic levels in a slightly different order, but the terms (names) of the three levels have been changed to indicate the cognitive processes involved. The six cognitive processes are

1. Remembering (Knowledge – order the same as before)
2. Understanding (Comprehension – order as before)
3. Applying (order and name (term) – no change for both)
4. Analyzing (order and name (term) – no change for both)
5. Evaluating (no change to name (term) but change to order, for it was last, 6, and now it is 5)
6. Creating (Synthesizing – order change from 5 to 6 and term no change) Woolfolk, 2004.

In addition, the revisers have added a new dimension to the taxonomy to recognize that cognitive processes must process something i.e. knowledge – you have to remember or understand or apply some form of knowledge. If you look at the table below you will see the result. We now have six processes – the cognitive acts of remembering, understanding, applying, analyzing, evaluating, and creating. These processes act on four kinds of knowledge – factual, conceptual, procedural, and metacognitive.

Use of Table of Specification or Test Blueprint:

1. Guiding the test constructors not to over test or under test any area so that the test would be fair or content valid in other words.
2. Inform the examinees what to expect so they can better prepare for it.

A table of specification is a test blueprint consisting of two-way grid (more recent three-way grid, third being knowledge dimension as seen above in the recent development); in one grid we have the contents (or topics) while on the other grid educational objectives (usually Bloom's taxonomies). Each entry shows the number of items (questions) in the respective topic and each respective educational objective.

Testing and Cognitive Research

How can we connect the way brain works with testing? How can we relate Information Processing Model of Memory to testing? Has testing changed as we understand how the brain works? Not really, as Murphy and Davidshofer (2005) notes explosion of knowledge regarding cognitive processes has had relatively little influence to date on testing. Despite all the advances in cognitive research, the best tests of cognitive ability still follow the basic philosophy and measurement strategies developed by Binet etc. other are Bloom.

Note that theory and practice are not at the same pace. In some areas theory could be ahead of theory. In testing, because of complexity of brain and its working and also because of complexity of areas to be tested they are bound to be not at per as such.

Note also that Information Processing Model of Memory is constantly being reviewed. Cognitive researchers have reached a consensus that many higher-level cognitive processes, which presumably include those involving mental abilities, can be best understood through a mental architecture (model) that includes three components (1) working memory (short term memory) (2) declarative memory and (3) procedural memory both of long term memory.

There are debates about this. Some even believe that there are no distinction between short term memory and long term memory.

A recent addition is notion of long-term working memory. Explicit memory (conscious) declarative e.g. semantic falls here, implicit memory (unconscious) procedural memory such as motor skills is the form of long-term memory.

In summary, the influence of cognitive psychology on psychological testing has been disappointing small. Note that testing can be political. Like we argued ‘status quo’ is guarded jealousy by those especially in power. They benefit from ‘status quo’ and their comfort is guarded at all cost.

Secondly, test publishing is a big business (here and America or west world), and it is difficult to convince publishers (or Examination Council or departments in Universities) to make radical changes in their tests, especially when familiar testing methods “work” so well or we are very comfortable with them since we are products of them. For instance, trying to sell multiple choice testing in Kenya and is widely used in America even at masters and PhD levels would be a tall order or endeavour.

Factors to Consider when Selecting a Test Format:

It is the teacher’s prerogative to use an essay or objective test. Even prerogative power has to be used intelligibly, so much so during these eras of transparency, accountability and honesty. Note that if objective test is our choice, we have several types (e.g. True or False, matching and multiple-choice (select types, supply type, completion, short answer)). Which test format would you recommend? The answer is, it depends on many things.

Let us try to enumerate them, of course not exhaustively. The factors (or things) are:

1. The purpose of the test
2. The time available to prepare and mark the test
3. The number of pupils to be tested
4. The physical facilities available for reproducing the test
5. Your skill in writing the different types of tests
6. In which subject we are testing
7. The level of the examinees etc.

Differences between the Essay and Objective Tests:

In contrast with objective tests, essay tests are relatively easy to prepare, the difficult part of the job is usually marking the examinees’ answers. The essay question is especially useful for measuring those aspects of complex achievement that cannot be measured by more objective means. They include:

1. The ability to supply than merely identify interpretations and applications of data (this is best measured by restricted response items or questions)
2. The ability to select, organize, and integrate ideas in a general attack on a problem (this is best measured by extended response items or questions often referred to as discussion, as we saw)

Although essay questions provide a relevant measure of significant learning outcomes, they have several limitations that restrict their use:

1. The scoring (marking) tends to be unreliable
2. The scoring (marking) is time consuming
3. Only limited sampling of achievement is obtained (i.e. has poor content coverage as compared to objective).

Because of these shortcomings essay questions should be limited to testing those outcomes that cannot be measured by objective items.

Two popular misconceptions not supported by empirical evidence are that

1. Essay tests assess certain skills such as analysis and critical thinking better than objective tests.
2. Essay tests contribute to better pupil study and work habits.

Note that these are 'misconceptions' that are not supported by any research findings. Rather than argue whether essay tests are better than objective tests, or vice versa, we should understand the strengths and weaknesses associated with each type (essay and objective) and capitalize on their strengths.

The advantages of objective items are more or less opposite of disadvantages of essay tests. We look more explicitly at the advantages and disadvantages of the most popular type of objective tests namely multiple-choice.

Advantages of Multiple Choice Testing:

i. Its wide applicability in the measurement of various phases of achievement

ii. It's also free of many of the limitations of other forms of objective items e.g. short answer, the teacher has at times to decide whether the response is correct or not, if the response is not exactly same as the answer provided, hence not reliable as multiple choice etc.

iii. It has a wide content coverage

iv. It is easy to mark.

Disadvantages of Multiple Choice Testing:

i. It is mainly a selection paper and pencil test that measures problem solving behaviour at verbal level only

ii. Its construction is difficult and time consuming

iii. It encourages rote memory and guesswork

iv. No freedom of expression for candidates

General item writing guidelines:

Brief discussion on guidelines that apply to all types of achievement test items is as follows:

1. Cover important material. This is the prime requirement for any item or question. No item or question should be included on a test unless it covers an important fact, concept, principle, or skill. Furthermore, a test should not cover just one type of skill, say, factual knowledge. Some items or questions should also measure higher-level skills such as comprehension, analysis and evaluation. This indicates a table of specification is a necessity when making a test.

2. Items or questions should be independent. The answer to one item or question should not be found in another item or question nor accurately answering one item or question should not be dependent on rightly answering a previous item (question).
3. Write simply and clearly. Use only terms and examples, which students will understand and eliminate all nonfunctional words. This ensures you are testing knowledge on the content area and not vocabulary or otherwise.
4. Be sure students know exactly what is required for right response. For instance, for an essay they must know the length and scope, and if quantitative accuracy expected.
5. Use a variety of test items or questions. Be creative in your testing technique even.
6. Revise and edit. A good suggestion is, write a pool of items (questions) then set them a side for sometime (a day, a week etc.) then look at them again revising, editing and eliminating or substituting poor ones.

We now look at specific guidelines for both essay questions and multiple-choice items. We begin by looking at guidelines for writing multiple-choice items.

Guidelines for Writing Multiple Choice Items:

A multiple choice item consists of a stem and set of alternatives (usually 3-5). The examinee's (student's) task is to select the correct alternative (referred to as the key) from among the distracters (incorrect responses).

1. The stem should present the problem and include all qualifying phrases. There should not be ambiguity about what is being asked or what the task is.
2. There should be only one correct alternative.
3. Distracters should be plausible (i.e. attractive or inviting for less knowledgeable) but clearly incorrect.
4. Avoid negative wording, and if used it should be underlined.
5. Use "all of the above", "none of the above" or "some of the above" sparingly.
6. If an item contains controversial material, cite the authority, whose opinion is being used e.g. ability accounts for less than 50% for doing well in tests, other factors account (Brown, 1976). Thus indicating the authoritative source.
7. Avoid irrelevant clues to the correct answer, this irrelevant clue could be in form of grammar, length and so forth e.g. which disease is caused by virus? 1) Malaria 2) viral pneumonia etc. Viral pneumonia here gives irrelevant clue that we are talking about.
8. Each item should test one central idea or concept.
9. List alternatives in alphabetical or logical order; otherwise randomize so as to avoid pattern. This is saying you can use order or not, but avoid a pattern.

Guidelines for Writing Essay Questions:

1. The question should clearly define the task.
2. Indicate the scope and direction of the answer required.
3. Use questions that have correct answers.
4. Allow for "think time".
5. Use a large number of short essay items rather than a few longer ones.
6. Use optional questions sparingly; Note that a test is an instrument for measuring ability and if so it needs to be the same instrument for all. Where there is a choice, some choices are bound to be simpler or more difficult than others are. Realizing this examiner should give or strive to give the same test for all, with the same difficult hence there should be no optional.

7. Develop a marking scheme before administering the test. The model answers (or marking scheme) should be made same time with the test.

In general, writing good items is a skill that is developed only by practice. It has to be developed through making mistakes at times. We found that there is no perfect test, meaning no test is beyond criticism or beyond improvement. You can always improve even the best test. Hence we need to be receptive to criticism all the times, though we all do not like criticisms. It is usually desirable to write many more items than will be included on the test and then edit them selecting the best ones, if we feel we are experienced and competent enough. Otherwise we can use experts or our colleagues in the profession for moderation of the test items.

Sometimes we tend to forget practical items (or performance items). These items require some physical or psychomotor response e.g. driving a car, tuning a car engine, swimming, baking a cake etc. Though you may write or give orally how to go about it (e.g. describing very well how to drive etc.), the practical (or psychomotor) component is usually the most important aspect of the response. The practical response indicates whether the candidate knows or not, and indeed this is what is important. This is not a paper and pencil test that measures problem solving behaviour at verbal or written level. It is a test of competency to find out whether the skill has been learnt or not. The candidate has to demonstrate whether he or she has mastered the skill or not practically.

We next consider an important aspect of analyzing the test items, whether they are worthy their use or worthy to be included in the test.

ITEM ANALYSIS:

A test is only as good as the items it contains. Thus, when constructing a test, we must be concerned with the quality of the items. When evaluating the quality of the items, various criteria are used. Here we list the important ones:

1. An item should measure the knowledge or skill it is designed to measure. This is validity or soundness of an item.
2. We should also be concerned about the quality of expression. Items (or questions) must be clearly written, grammatical and at the appropriate reading level. Thus you should check whether, all serious learners (but not giving unfair hints) understand the questions.
3. The statistical characteristics (analysis) of the item, which would be a topic of discussion here. Others two above have already been discussed.

Statistical analysis of test items is what we refer to as item analysis. The item analysis helps an examiner to:

i. Judge the quality of the item. Thus the examiner can identify good or poor items.

ii. Identify knowledge and skills examinees have and have not mastered, if an item in a classroom test is answered incorrectly by a majority of the students, this information tells the teacher that something is wrong. Unfortunately without further investigation, it does not tell her/him what went wrong. The item may have been misleading or poorly constructed, the material has been so difficult that students were not able to learn it, or the instruction may have been incomplete. Only further analyses would tell which is the most likely or the true explanation.

Specifically, most item analyses are concerned with three aspects of an item; others are biserial correlation and factor analysis, which are of no concern here at the moment. The three major aspects are given below:

1. Difficulty of item, which is nothing else but the proportion of examinee who answer an item correctly and it is referred to as difficult index.
2. Discrimination power of the item or else called discrimination index, is concerned with whether the item differentiates between people with varying degree of knowledge or ability.
3. Content validity as well as the effect of distracters.

Note: If the difficult index is low the validity may still be okay, but if the discrimination index is negative, for an item, the item may not be measuring what it is supposed to measure (not valid). It may be measuring something else but definitely not ability.

The item analysis is important for item (or question) bank. It is ridiculous for teachers to have to write new items (questions) every time they prepare a test. Over time, they should have built a *test file* of the better items to be reused. Item analyses help in this line for you are able to know good and bad items and bad items (questions) have to be discarded or be improved. **As seen above that** new measures are difficult to come by. Nonetheless, although researchers or examiners can develop new measures for their constructs or tests which is not an easy task, they should first search for the best existing tests for obvious reasons. Using existing measures or tests can save large amount of test development time. Moreover, using existing measures improves the comparison of different studies. When studies use the same measure, differences in their outcomes can be traced to design and sample differences rather than to measurement differences.

Note item analysis is best done in multiple-choice items. We take an example here say we wish to item analysis ith item (ith item may be a multiple-choice item say item [or question] 25 and this has 5 alternative).

We consider a group of 40 examinees and suppose they responded as below to this item:

	A	B*	C	D	E	Omit
Upper group	0	20	0	0	0	0
Lower group	3	8	4	2	3	0

Asterisk indicates the correct answer (or key). B is the key. Others are distracters (the ones which are incorrect).

For each item, compute the percentage or proportion, which gets (or who get) the item correct. This is what is called item difficulty index. Thus item difficulty index can be expressed as a decimal, fraction or percentage. Thus range is from 0-1 (for decimal or fraction) or 0-100% (for percentage). Item difficulty index is denoted by p.

p = number answering correctly/number of test
takers =28/40 = 0.7 for our above case.

An item with a difficulty index of 0.3 is more difficult than an item with a difficulty index of 0.8. Why?

The index is quite useful in item analysis. If p for an item is very close to 0 or 1, the item generally should be altered or discarded, because it is not giving any information about differences among examinees’ trait levels or abilities.

If p = 0 the item is very difficult (nobody got it right)

If p = 1 the item is very easy (everybody got it right).

Acceptable p is 0.3-0.7 for it maximizes the information the test provides about differences among the examinees. Note: You can have simple items early in the test for motivational reasons.

Examiner should not forget the purpose of the test. A test used to select graduate students for a university that admits about 10% of the applicants should contain extremely difficult items. A test used to select children for a remedial education program should contain very easy items. By this time you should have realized even objective (multiple-choice) tests can be used for any level of education (even for Ph.D. programs). Next is item discrimination index:

Item discrimination index is obtained by subtracting the number of students in the lower group who answered the item correctly from the number in the upper group who got the item right, and dividing this by the number of students in either group. That is, half of the total number of students when we divide the group into upper and lower halves. In our example:

Discrimination index = $$\frac{R_U - R_L}{\frac{1}{2}T} = \frac{20-8}{\frac{1}{2}\times 40} \frac{12}{20} == 0.6$$

This value is usually expressed as a decimal and can range –1.00 to +1.00.

If it has a positive value, the item has a positive discrimination. This means that a large proportion of the more knowledgeable students than poor students got the item right. If the value is zero, the item has zero discrimination. This can occur

1. Because the item is too easy or too hard
2. Because it is ambiguous.

If more poor than better students get the item right, one would obtain a negative discrimination. For a classroom test a discrimination of an item of 0.20 and above is good. Note: The discrimination index is of an item (or a question) not for the test. You are necessarily analyzing each item at a time i.e. looking at each item of the test and determining its quality in terms of how the examinees have performed in it. Thus the two indexes (difficulty and discrimination) are group dependent and consequently should be used bearing that in mind.

Obtaining Discrimination Index for a Large Group:

(not very practical)

If the total test scores are normally distributed, we use 27% of the examinees with the highest total as the upper range and 27% of lower as the lower range i.e. 27% of upper range who got it right – (minus) 27% of lower range who got it right i.e.

$$D = \frac{R_U}{\frac{1}{2}T} - \frac{R_L}{\frac{1}{2}T}$$

i.e. Discrimination = $P_U - P_L$ as above

Note:

Item analysis data are not analogous to item validity. In order to judge accurately the validity of an item, one should use some external criterion (e.g. the experts etc.). Note that here we considered our test to group the examinees into higher scorers and lower scorers. In reality we used an internal criterion. And do not forget we can have a reliable test, which is not valid, though the converse cannot be true.

The discrimination index is not always a measure of item quality.

In general, discrimination index of 0.40 is regarded as satisfactory. However, one should not automatically conclude that because an item has a low discrimination index, it is a poor item and should be discarded. Items with low or negative discrimination indices should be identified for more careful examination. Those with low, but positive, discrimination indices should be kept (especially for mastery tests). As long as an item discriminates in a positive fashion, it is making some contribution to valid measurement of the students' competencies. And as long as we need some easy items to instill proper motivation in the examinees, such items are valuable.

Grading (or Marking or Scoring) Errors:

Because ratings (grading, markings or scorings) depend on subjective judgment, especially so of essay tests, they are open to various types of errors and biases. Training of examiners, so that they can be alert of such errors and biases can reduce their occurrence in the examiners marking or scoring. We identify some of the errors:

1. *Error of central tendency:* This normally happens when marks have a very narrow range such that nobody fails and nobody passes highly either. This is most often happens in subjective types of tests. Hence this can be avoided by having more objective tests. If an essay test must be given, it should consist of a large number of short essay questions rather than a few longer ones to maximize the objectiveness or reliability.
2. *Error of leniency:* This is to be overgenerous in awarding marks. Stick to the marking scheme, just giving marks as the candidate honestly deserves. Candidates should be rewarded for the genuine effort and what they have genuinely worked for. This is easily said than done. Nonetheless the examiner needs to be fair, awarding marks in the fairest way possible without any biases of any kind.
3. *Error of severity:* Examiners can be too stringent. As said above an examiner should exercise fairness. The examiner should not be too stringent to a point of being inhumane. Normal human mistakes, spelling, miscalculation etc. should not be unreasonably be penalized.
4. *Halo effect:* This is general opinion of a person to influence the marking or general opinion of a school; or a centre etc. Examinees should be awarded the marks they deserve. The examiner should evaluate candidates' work unbiasly regardless of the background of the examinee. The examiner should not be lenient or severe depending on the background of the candidate. The examiner should be unbiased as indicated before, possibly in different words.
5. *Logical error:* This occurs when markers (examiners) assume that two characteristics are related e.g. intelligence and creativity, or good in mathematics must be good in science and so forth. The examiner has to stick to his or her marking scheme without being biased or favouring any individual or any group of individuals.

REFERENCES:

Allen M.J. & Yen W.M. (1976) *Introduction to Measurement Theory*. Belmont CA: Wadsworth Inc.

Brown, F.G. (1970) *Principles of educational and psychological testing. 2nd ed*. New York: Holt, Rinehart & Winston.

Gronlund, N.E. (1985) *Measurement and evaluation in teaching 5th ed.* New York: Macmillan.

Hinkel D.E., Wierma W. & Jurs S.G. (1998) *Applied Statistics for Behavioural Sciences, 4th ed.* Boston: Houghton Mifflin Co.

Howell, D.C. (2002) *Statistical Methods for Psychology, 5th ed.* Pacific Grove CA: Duxbury

Ingule F. & Gatumu H. (1996) *Essentials of Educational Statistics.* Nairobi: E.A. Educational Publishers:

Gatumu, H.N. (2002) *STC/GCD 517: Psychological Assessment.* Nairobi: Kenyatta University Press for Institute of Open Learning

Gatumu, H. N. (2010) EGC 519 *Pschometrics.* Nairobi: Kenyatta University Press for Institute Open Learning and e-Learning.

Glass, G.V. & J.C. Stanley (1970) *Statistical Methods in Education and Psychology*. New Jersey: Prentice-Hall.

Glass, G.V. & K.D Hopkins (1996) *Statistical Methods in Education and Psychology.* Boston: Allyn & Bacon.

Johnson, R.R. (1980) *Elementary Statistics.* Mass.: Duxbury Press. Smith,
G.M. (1970) *A Simplified Guide to Statistics for psychology and Education*. New York: Holt, Rinehart and Winston

ANSWERS TO SELECTED QUESTIONS

Lecture One p.7

5. a) Nominal b) Ratio c) Ratio d) Interval e) Ordinal

Lecture Three p.33

2. Mode = 43 Median = 48.75 Mean = 48.29

Lecture Four p.46

1. (i) Mean = 5.89 Range = 7 Mean Deviation = 1.68 Median =
Variance (S_X^2) = 4.32 Standard Deviation (S_X 5.75) = 2.08
(ii) Mean = 11.89 Variance (S_X^2)=4.32 Standard Deviation (S_X) = 2.08
(iii) Mean = 0.89 Variance (S_X^2) = 4.32 Standard Deviation (S_X) = 2.08
(iv) Mean = 23.56 (or 5.89×4) Variance = 69.12 (or 4^2 ×4.32)
Standard Deviation = 8.32 (or 4×4.32)

3. (i) Mean = 49.8 (ii) Variance = 77.16 (iii) Standard Deviation = 8.78

4. a), b) and e)

Class Interval	Real Class Interval	Class-mark or Midpoint(x_i)	Tally Marks	Frequency(f_i)	Cumulative Frequencies(below and above respectively)	fx	Deviations($d_i = x_i$-28.5)	d_i^2
10-14	9.5-14.5	12	//	2	2 30	24	-16.5	272.25
15-19	14.5-19.5	17	//\\//	5	7 28	85	-11.5	132.25
20-24	19.5-24.5	22	//\\//	5	12 23	110	-6.5	42.25
25-29	24.5-29.5	27	///	3	15 18	81	-1.5	2.25
30-34	29.5-34.5	32	//\\// /	6	21 15	192	3.5	12.25
35-39	34.5-39.5	37	//\\//	5	26 9	185	8.5	72.25
40-44	39.5-44.5	42	//	2	28 4	84	13.5	182.25
45-49	44.5-49.5	47	//	2	30 2	94	18.5	342.25

Sum $\sum f_i = 30$ $\sum f_i x_i = 855$ $\sum d_i^2 = 1058$

c) (i) Mode = 32 (ii) Median = 29.5 (iii) Mean = $\frac{855}{30}$ = 28.5

d) Range = 47-12 =35

f) Variance = $\frac{1058}{30}$ =35.27 Standard Deviation = $\sqrt{35.27}$ =5.94

g) Mode > Median >Mean hence negatively skewed; the test too easy.

Lecture Five, Six and Seven p.61

1. A. (i) 68 scores (ii) 84 scores (iii) 27 pupils (iv) 67 pupils
   B. 42
2. (i) 106 (ii) 625 (iii) 67
3. (a) 136 (b) 30 (c) 117
4. (a) 0.115 (b) 0.68 (c) 0.97 (d) 0.96 (e) 0.34 (f) 0.98
5. (a) 93% (b) 93
6. (a) 38.5% (.385) (b) 25.2% (.252) (c) 66.3% (.663) (d) 18.6% (.186)
   (e) 28.8% (.288)
7. (a) -0.5 (b) 1.5 (c) .241 (d) 15 (e) 36

Lecture Eight and Nine p. 71

1. (i) Biology -2.5 Physics -1.6 Done better in physics
   Assumption: Biology and physics scores had exactly similar distributions in shape.
   (ii) 1.36 (iii) 33

5.

Raw score	z score	percentile rank	T score	stanine
52	2	97	70	9
50	1	84	60	7
48	0	50	50	5
44	-2	2.3	30	1
42	-3	.003	20	1

Lecture Ten, Eleven and Twelve p. 88

1. (b) r_{XY} = 0.85

(c) Great tendency of those doing well in Test A to do well in Test B and those doing poorly in Test A to do poorly in test B as well.

(e) ρ_S = 0.827

2. (b) r_{XY} = 0.82 (c) ρ_S = 0.725 (d) Any, justified by reason given

3 (a) r_{XY} = 0.57 ρ_S = 0.39

Rather low correlation hence not extremely much logical relationship.