COURSE CODE: STS 203

COURSE TITLE: GENERAL STATISTICS

NUMBER OF UNIT: 3 UNITS

COURSE DURATION: THREE HOURS PER WEEK.

COURSE COORDINATOR: DR GODWIN NWANZU AMAHIA B.Sc, M.Sc, ph.D AND DR. (MRS) DANSU. (go.amahia@mail.ui.edu.ng and bolanlebm@yahoo.ca).

LECTURER OFFICE LOCATION: HOD STATISTICS OFFICE AND AMREC

COURSE CONTENT:

Analysis and presentation of statistics data, Measures of location, Measures of dispersion, Regression and correlation analysis, Time series analysis, Demographic measures, Design of simple experiments, Analysis of Variance and Non – Parametric tests.

COURSE REQUIREMENTS:

This is a compulsory course for all statistics students. Students are expected to have a minimum of 75% attendance to be able to write the final examination.

READING LIST:

1.) **Probability and statistics for engineers & scientists by Walpole and Myers.**
2.) **Applied statistics by department of statistics.**
3.) **Introduction to basic statistics by Adamu.**

LECTURE NOTES

ANALYSIS AND PRESENTATION OF STATISTICAL DATA

Definition

Statistics is the scientific method for collecting, organising, presenting and analysis of data, for the purpose of making reasonable decisions and drawing valid conclusion on the basis of such analysis.

Importance/Uses

1. Indispensable to research and development.
2. Applied in all discipline and areas of human endeavours.
3. Useful in short and long term range planning.
4. Serve as basic inputs in policy and programme formulation, implementation and evaluation.

Misuses

1. Quoting statistics with complete disregard for its limitations.
2. Careless and irrelevant comparison (comparison in the absence of statistics).
3. Sweeping generalizations.
4. Misleading graphical representation of data.

Divisions of statistics

1. Descriptive Statistics or Deductive Statistics.
2. Inferential Statistics or Inductive Statistics.

Basic Statistical Terms

1. Population – the totality of object of interest
2. Sample – a portion of the population selected for enquiry
3. Parameter – measurable characteristics of the population
4. Census – the process of obtaining information about the population
5. Sample survey – the process of obtaining information about the sample
6. Statistic – measurable characteristics of the sample
7. Variable - a symbol, such as X, Y, Z, that 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 a **constant**. A variable that can theoretically assume any value between two given values is called **continuous** variable, otherwise it is called a **discrete** variable.

A variable can also be described as **qualitative** when it yields categorical responses, e.g. male or female. It is **quantitative** if it yields numerical responses, recorded on a naturally occurring numerical scale. Quantitative variables could be discrete or continuous.

Sources of statistical data

1. Primary source / primary data
2. Secondary source/ secondary data

Methods of collecting Quantitative primary data

1. Interview method (a) Personal interview
   (b) Telephone interview
   (c) Computer assisted interview
2. Questionnaire method
3. Observation method
4. Experimental method

Method of Collecting Qualitative primary data

1. In – Depth Interview
2. Focus Group Discussion

Scales of measurement

Measurement scales are instrument for measuring variables. There are four types of scales on which a variable may be measured:

1. Nominal scale - merely attempts to assign identities to categories e.g. sex, religion, e.g.
2. Ordinary scale - ranks ideas or object in an order of priority or preference. Interval between ranks are not equal e.g. strongly agree, disagree, no response e.g.
3. Ratio scale - have equal intervals, and each is identified with a number e.g. speed length e.g.
4. Interval scale - similar to ratio scale but lack a true zero. The intervals are equal but the zero is fixed arbitrarily e.g. temperature.

Sources of secondary data

1. Publication and records of government and NGOs
2. Journals of universities and research institutes
3. Magazines, newsletters and newspaper reports
4. Administrative reports
5. Internet - i.e. www.nigerianstat.gov.ng

Limitations of secondary data

1. Incompleteness
2. Irregular publications
3. Inaccuracy
4. Out datedness

Problems of Data Collection in Nigeria

1. Lack of statistical awareness
2. Inadequate funding of statistical agency
3. Poor social facilities
4. Lack of adequate coordination among data collection agency
5. Cultural or religions problems
6. Inadequate statistical manpower

Errors in Data Collection

1. Sampling Errors - occur as a result of making estimates of the population parameter from sample. Basic sources include:
(I) improper selection of the sample
(ii) Substitution
(iii) Faulty demarcation of the sampling unit
(iv) Errors due to wrong method of estimation
2. Non Sampling Errors - occur as a result of improper observation or recording of sample characteristics. Basic sources include:
(i) Incomplete coverage
(ii) Defective method of data collection
(iii) Interviewer or enumerator's bias

(iv) No response
(v) Compilation and tabulation process

Data presentation
Statistical data can be presented in any three key ways namely, tabular, graphical and diagrammatic presentation of data.

Tabular presentation of data

1. Raw data are collected data that have not been organized numerically. An **array** is an arrangement of raw numerical data in ascending or descending order of magnitude.
2. When summarizing large masses of data, it is often useful to distribute the data into **classes**, or categories, and to determine the number of individuals belonging to each class, called the class **frequency**.
3. A tabular arrangement of data by classes together with corresponding class frequencies is called a **frequency distribution**, or frequency table.
4. Rules for forming frequency distribution

(i) Find the range i.e. highest value - lowest value
(ii) Find the number of classes required (k)

$$K = \frac{\text{Range} + 1}{\text{Class size}}$$ * must be discrete value.

(iii)calculate the upper limit of the first class using the formula:

$U_1 = L_1 + C - 1$ for whole numbers
$L_1 + C - 0.1$ for data with 1 decimal
$L_1 + C - 0.01$ for data with 2 decimals
$L_1 + C - 0.001$ for data with 3 decimal, e.t.c

(iv) Form frequency table

Example
The following relates to the weights of 40 male students in a state university. The data were recorded to the nearest pound. Using a class size of 9, construct a grouped frequency distribution table:

138	146	168	146	161
164	158	126	173	145
150	140	138	142	135
132	147	176	147	142
144	136	163	135	150
125	148	119	153	156
149	152	154	140	145
157	144	165	135	128

Steps

(i) Range = 176 – 119 = 57

(ii) k = $\frac{\text{Range+1}}{C} = \frac{57+1}{9} = 6.4$ => round up to 7 =>7 classes

(iii) $U_1 = L_1 + C - 1$

Weights (1b)	Frequency
119-127	3
128-136	6
137-145	10
146-154	11
155-163	5
164-172	3
173-181	2
	40

Terms Associated with Grouped Frequency Distributions

1. class interval - symbol defining a class e.g. 128 - 136
2. class limits - the end numbers -> lower or upper class limits or end points of a class
3. class boundaries - obtain by manipulating ± 0.5

   ± 0.5for whole numbers
   ± 0.05 for data with 1 decimal
   ± 0.005 for data with 2 decimals
   ± 0.0005 for data with 3 decimals
4. Class size or width - the differences between lower and upper class boundaries
5. Class mark or midpoint - the average of class limits

Example

Class	f	class mark (Xj)	class boundaries	Rf	Cf
119 - 127	3	123	118.5 - 127.5	7.5	3
128 - 136	6	132	127.5 - 136.5	15	9
137 - 145	10	141	136.5 - 145.5	25	19
146 - 154	11	150	145.5 - 154.5	27.5	30
155 - 163	5	159	154.5 - 163.5	12.5	35
164 - 172	3	168	163.5 - 172.5	7.5	38
173 - 181	2	177	172.5 - 181.5	5	40

Graphical Presentation of Data

Histograms and frequency polygons are two graphical representations of frequency distribution.

Histogram - consists of set of rectangles having: (a) bases on a horizontal axis with centres at the class marks and length equal to the class interval sizes, and (b) areas proportional to the class frequencies.

Frequency polygon - is a line graph of the class frequency plotted against the class mark. It can be obtained by connecting the midpoints of the tops of the rectangles in the histogram.

Relative Frequency Distribution

The relative frequency of a class is the frequency of the class divided by total frequency of all classes and is generally expressed as a percentage.

Cumulative frequency distribution or ogive

Cumulative frequencies are the cumulative totals of successive frequencies of a frequency distribution. The graph of a cumulative frequency distribution is called cumulative frequency polygon or ogive. Cumulative frequency are of the ***less than*** or ***more than*** types. The less than type is the more general. In its construction, each cumulative frequency is plotted against the upper class boundaries of the class interval.

Diagrammatic presentation of data

Pie chart

Pie charts can be defined as a circle drawn to represent the totality of a given data. The circle is also divided into sectors with each sector proportional to the components of the variable it represents.

Bar chart

Bar charts are simple diagrams that are made up of a number of rectangular bars of equal widths whose heights are proportional to the quantities or frequencies they represent.

Types of Bar Charts

1. Simple bar chart
2. Multiple bar charts
3. Component bar chart

(a) Actual component bar chart

(b) Percentage component bar chart

MEASURES OF LOCATION, DISPERSION, SKEWNESS AND KURTOSIS

Quantitative data can be described in terms of three properties namely tendency or location, dispersion or variation and shape. Each of these properties has descriptive measures that describes it i.e. measures of central tendency, measures of dispersion and measures of skewness and kurtosis (shape).

Measures of central tendency

Measures of central tendency otherwise known as measures of location are simply averages. The most commonly used are the arithmetic mean, mode, median, geometric mean and harmonic mean.

The arithmetic mean

The arithmetic mean of a set of observations is the sum of observations divided by the number of observations. Thus, for a set of numbers $x_1, x_2, x_3, \ldots\ldots\ldots x_n$.

Example

Obtain the arithmetic mean for the set of numbers 3,8,4,6, and 7.

$$AM = \frac{\sum Xi}{N} = \frac{3 + 8 + 4 + 6 + 7}{5} = 5.6$$

Example

Marks scored by 50 students in a course are presented below:

Marks scored x	No of students f	fx
0	4	0
1	6	6
2	4	8
3	3	9
4	15	60
5	10	50
6	5	30
7	3	21
	50	184

$$AM = \frac{\sum fx}{n} = \frac{184}{50} = 3.68$$

Example

Monthly earnings in 000's of naira of 100 workers are presented below:

Monthly earnings	no of workers (f)
4.51 -5.32	15
5.33 – 6.14	7
6.15 – 6.96	35
6.97 – 7.78	28
7.79 – 8.60	10
8.61 – 9.42	5

Long Method

Class	f	x	fx
4.51 – 5.32	15	4.915	73.725
5.33 – 6.14	7	5.735	40.145
6.15 -6.96	35	6.555	229.425
6.97- 7.78	28	7.375	206.5
7.79 – 8.60	10	8.195	81.95
8.61 – 9.42	5	9.015	45.075
	100		676.82

$$Mean = \frac{\sum fx}{n} = \frac{676.82}{100} = 6.7682$$

Short – Method (Coding or Assumed mean method)

$$Mean = A + \left[\sum \frac{fu}{n}\right]C$$

Example

Class	f	x	u	fu
4.51 – 5.32	15	4.915	-2	-30
5.33 – 6.14	7	5.735	-1	-7
6.15 – 6.96	35	6.555	0	0

6.97 – 7.78	28	7.375	1	28
7.79 – 8.60	10	8.195	2	20
8.61 – 9.42	5	9.015	3	15
	100			26

Am = 6.555 x [26/100] x 0.82 = 6.555 + .2132 = 6.7682

The Median

Median is the middle value. It divides a distribution into two equal parts. To obtain the median from raw data, we must first arrange the data in order of magnitude. That is in form of an array.

Array: 119, 129, 129, 130, 132, 141, 143

n = 7, median = 4^{th} observation, = 130.

Array: 10,3,12,8,15,17,6,13.

3,6,8,10,12,13,15,17.

n = 8, median = $\frac{10 + 12}{2} = 11$

Computation of median from grouped frequency table:

$$\text{Median} = L_b + \left[\frac{\frac{n}{2} - \sum f1}{fm}\right]C$$

Example

Staff strength	No of companies	Cf
1-10	1	1
11-20	5	6
21-30	10	16
31-40	19	35
41-50	42	77
51-60	10	87
61-70	6	93
71-80	4	97
81-90	2	99
91-100	1	100
	100	

$$\text{Median} = 40.5 + \left[\frac{\frac{100}{2} - 35}{42}\right]x10 = 44.07$$

The Mode

The model is simply the item with the highest frequency. A distribution can have more than one model, unimodal - one mode, bimodal - two modes, trimodal - three modes, and multimodal - more than three modes.

Mode from raw data - the mode can be obtained from raw data by simply picking the item that occurs most frequently.
Given 2,8,3,4,2,6,2,4.
Mode =2 since it occurs most frequently.

Mode from ungrouped frequently table:

Given: x	f
1	4
2	6
3	5
4	5

Mode = highest frequency is 6 and the corresponding value of x is 2. Hence, mode is 2.
Mode from grouped frequency table

$$\text{Mode} = Lb + \left[\frac{f1}{f1 + f2}\right]C$$

Example
Time taken in seconds by 100 different chemical substances to melt when subjected to a particular temporary condition are given below:

Time (in seconds)	f
4.51-5.32	15
5.33-6.96	7
6.15-6.96	35
6.97-7.78	28
7.79-8.60	10
8.61-9.42	5
	100

$$\text{Mode} = 6.145 + \left[\frac{28}{28+7}\right]X.82 = 6.801$$

The Geometric mean (Gm)
Gm of a set of positive numbers $x_1, x_2,.....x_n$ is the n^{th} root of the product of the numbers.

$$\text{Gm} = \sqrt[n]{x1.x2.....xn}$$

It can also be obtained by finding the antilog of the arithmetic mean of the logarithmic values of the variable:

$$\text{Gm= log G} = \left(\frac{\sum \log x}{n}\right)$$

Example
Given: 5,8,10, obtained Gm.

log G = (log 5+ log 8 + log 10)/3
= (.69897 + .90309 + 1.0000)/3
= .86735
GM = 7.36801

The Harmonic mean

The harmonic mean of a set of numbers $x_1,x_2,.....x_n$ is defined as the numbers of values divided by the reciprocals of the numbers.

$$HM = \frac{n}{\sum \frac{1}{x}}$$

Example

Given: 2,8,7,4 and 5. Obtain the Hm.

$$Hm = \frac{5}{\frac{1}{2}+\frac{1}{8}+\frac{1}{7}+\frac{1}{4}+\frac{1}{5}} = 4.106$$

Measures of Non - Central Tendency

Apart from the median which divides a distribution into 2 equal parts, there are other quantities that divide a distribution into 4, 10 and 100 equal parts i.e

Median 2 equal parts, Quartiles 4 equal parts, Deciles 10 equal parts, and Percentiles 100 equal parts. These quantities are collectively called measures of non - central tendency, positional values, quartiles or fractiles. Positional values can be estimated by formulae or from the ogive. The following formulae are for obtaining positional values:

Quartiles

$$Qi = Lb + \left[\frac{\frac{ni}{4} - \sum f1}{fq}\right]C, i = 1, i = 2, i = 3$$

Deciles

$$Di = Lb + \left[\frac{\frac{ni}{10} - \sum f1}{fD}\right]C$$

Percentiles

$$Pi = Lb + \left[\frac{\frac{ni}{100} - \sum f1}{fP}\right]C$$

Example

The distribution of the sum of 40 students in a certain examination is shown below:

Score(%)	f	cf
20-29	3	3
30-39	6	9
40-49	6	15
50-59	9	24

60-69	7	31
70-79	5	36
80-89	4	40
	40	

$$Q1 = 39.5 + \left[\frac{10-9}{6}\right]x10 = 41.2$$

$$Q3 = 59.5 + \left[\frac{30-24}{7}\right]x10 = 68.07$$

$$D2 = 19.5 + \left[\frac{8-0}{3}\right]x10 = 46.17$$

$$D7 = 59.5 + \left[\frac{28-24}{7}\right]x10 = 65.21$$

$$P90 = 69.5 + \left[\frac{36-31}{5}\right]x10 = 79.5$$

Measures of dispersion

Dispersion is all about the amount of the spread or scatter in a distribution. Measures of dispersion fall into two categories:

Measures of absolute dispersion

(i) Range
(ii) Quartile deviation
(iii) Mean deviation
(iv) Standard deviation and variance

Measures of relative dispersion

(i) Coefficient of quartile deviation
(ii) Coefficient of mean deviation
(iii) Coefficient of variation

Range

It is simply the difference between the largest and the smallest values in a distribution.

Quartile Deviation

$$QD = \frac{Q3 - Q1}{2} = \frac{68.07 - 41.2}{2} = 13.435$$

Mean Deviation

$$MD = \frac{\sum |X - \overline{X}|}{n} \; for raw data$$

$$MD = \frac{\sum f\,|X - \overline{X}|}{n} \; for frequency table$$

Example

Scores	x	f	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$	$f\lvert x - \bar{x} \rvert$	fx
40-44	42	2	-15	15	30	84
45-49	47	5	-10	10	50	235
50-54	52	8	-5	5	40	416
55-59	57	12	0	0	0	684
60-64	62	7	5	5	35	434
65-69	67	4	10	10	40	268
70-74	72	2	15	15	30	144
		40			225	2265

$$\bar{x} = 57, MD = \frac{225}{40} = 5.63$$

Variance and Standard Deviation

$$SampleVariance(S^2) = \frac{\sum (x - \bar{x})^2}{n-1} \; forrawdata$$

$$SampleVariance(S^2) = \frac{\sum f(x - \bar{x})^2}{n-1} \; forfrequencydata$$

$$SamplesS\tan dardDeviation(S) = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \; forrawdata$$

$$SampleS\tan dardDeviation(S) = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}} \; forfrequencydata$$

$$PopulationVariance(\sigma^2) = \frac{\sum (x - \mu)^2}{n} \; forrawdata$$

$$PopulationVariance(\sigma^2) = \frac{\sum f(x - \mu)^2}{n} \; forfrequencydata$$

$$StdDev(\sigma) = \sqrt{\frac{\sum (x - \mu)^2}{n}} \; forrawdata$$

$$StdDev(\sigma) = \sqrt{\frac{\sum f(x - \mu)^2}{n}} \; forfrequencydata$$

Short cut method

$$\sigma^2 = \left[\frac{\sum fu^2 - \frac{\left(\sum fu\right)^2}{n}}{n} \right] C^2$$

$$\sigma = \sqrt{\sigma^2}$$

$$S^2 = \left[\frac{\sum fu^2 - \frac{\left(\sum fu\right)^2}{n}}{n-1} \right] C^2$$

$$S = \sqrt{S^2}$$

Note – For large sample sizes (n ≥ 30) the population standard deviation formula may be used to obtain standard deviation for sample. In such case, we use the sample mean to replace the population mean.

Example

Class	f	x	u	u^2	fu	fu^2
1 – 10	6	5.5	-3	9	-18	54
11 – 20	6	15.5	-2	4	-12	24
21 – 30	12	25.5	-1	1	-12	12
31 – 40	11	35.5	0	0	0	0
41 – 50	10	45.5	1	1	10	10
51 – 60	5	55.5	2	4	10	20
	50				-22	120

$$\sigma^2 = \left[\frac{120 - \frac{(-22)^2}{50}}{50} \right] x10^2 = 220.64$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{220.64} = 14.85$$

Measures of Relative Dispersion

Coefficient of Quartile Deviation

$$\frac{Q3 + Q1}{Q3 - Q1} X100$$

Coefficient of Mean Deviation

$$\frac{MeanDeviation}{Mean} X100$$

Coefficient of Variation

$$\frac{S\tan dardDeviation}{Mean} X100$$

Skewness and Kurtosis

Before discussing the concept of skewness, an understanding of the concept of **symmetry** is essential. Symmetry is said to exist in a distribution if the high values and the low values balance themselves out in their frequencies i.e. if the smoothed frequency polygon of the distribution can be divided into equal halves. Symmetry does not necessarily mean normality. The reverse is however the case as every normal distribution is symmetrical. **Skewness** on the other hand means lack of symmetry. Skewness can be positive or negative.

Measures of Skewness

Measures based on tendency

(i) Personian 1st coefficient of skewness

$$\text{Sk} = \frac{Mean - Mode}{S\tan dardDeviation}$$

(ii) Personian 2nd coefficient of skewness

$$\text{Sk} = \frac{3(mean - median)}{s\tan darddeviation}$$

Measures based on positional values

(i) Quartile coefficient of skewness

$$\text{Sk} = \frac{(Q_3 - 2Q_2 + Q_1)}{(Q_3 - Q_1)}$$

(ii) Percentile coefficient of skewness

$$\text{Sk} = \frac{(P_{90} - 2P_{50} + P_{10})}{P_{90} - P_{10}}$$

Kurtosis

This is one other indicator of the shape of a distribution. The **kurtosis** of a distribution is its degree of peakedness and it is usually discussed and measured relative to that of normal distribution. A distribution that is peaked as the normal is called **mesokurtic** distribution. When a distribution is more peaked than the normal is called **Leptokurtic** distribution. When a distribution is less peaked than the normal, it is called **platykurtic** distribution.

Measures of kurtosis

(i) Moment coefficient of kurtosis

$$\frac{4thCentralMoment}{(Variance)^2} = \frac{m_4}{S^4}$$

(ii) Percentile coefficient of kurtosis

$$\frac{\frac{1}{2}(Q_3 - Q_1)}{(P_{90} - P_{10})}$$

TIME SERIES ANALYSIS

Statistical data which are collected at regular intervals over a period of time are called time series data. Common examples include annual population figures, monthly production figures, e.t.c. each of which are recorded over a numbers of such period. Time series are studied with a view to detect the pattern of changes in the value of the variable of interest over time. Such knowledge is useful in predicting the likely future occurrence and for planning and budgeting.

Components of time series

1. Secular or long term trend.
2. Seasonal variation.
3. Cyclical variation or movement.
4. Irregular or erratic variation.

Long term trend

This refers to the smooth or regular movement of the series over a fairly long period of time. Generally, three types of trend may be observed in a time series.

1. Upward trend: characterized by a general increase in the values of the series over a time.
2. Downward trend: characterised by a general decline in the values of a series over time.
3. Constant trend: in this case, despite periodical fluctuations in the time series, the overall or average figure tends to be constant.

Seasonal variation

This describes any kind of movement or variation which is of periodical nature and for which the periodical does not extend beyond a year. It consists of regular repeating pattern.

Cyclical variation

This refers to the recurrent up and down movement, or long time oscillation, in a statistical data from some sort of statistical trend or normal.

Irregular variation

It refers to variations which are completely unpredictable or are caused by such isolated special occurrence as good or bad news, bank failure, election, war, flood, e.t.c.

Models of time series

Two models are appropriate for associating the components of a time series.

1. Additive model i.e $Y = T + S + C + I$
2. Multiplicative model i.e $Y = TSCI$

In both cases: Y = observed data. T = trend values, S = seasonal variation, C = cyclical variation and I = irregular variation.

Estimation of components (Long term trend)

Given a time series data, there are four method that may be use to determine the general trend in its long term movement they are:

1. Freehand method.
2. Method of semi average.

3. Method of moving averages.
4. Least squares method.

Least squares method

This is similar to the least squares regression techniques. The dependent variable in case of time series analysis is the value of series (v_t), while the period of time (t) is the explanatory or independent variable. The least squares trend equation is written as:

$V_t = a + b_t$

where

$b = \frac{\sum V_t}{\sum t^2}$, $a = \frac{\sum V}{N}$:The trend values for each period are estimated using the trend equation:

$Vt = a + b_t$

Example1 - odd number time point.

The table shows the profit made by a company between 2001and2009.

Year	2001	2002	2003	2004	2005	06	07	08	09
Profit(#m)	10.1	12.7	12.4	11.9	12.5	13.0	14.9	16.5	18.7

Year	v	t	v.t	t^2	trend value (#)
2001	10.0	-4	-40	16	10.13
2002	12.7	-3	-38.7	9	11.00
2003	12.4	-2	-24.8	4	11.88
2004	11.9	-1	11.9	1	12.75
2005	12.5	0	0	0	13.62
2006	13.0	1	13	1	14.49
2007	14.9	2	29.8	4	15.36
2008	16.5	3	49.5	9	16.24
2009	18.7	4	74.8	16	17.11
	122.6		52.3	60	

$b = \frac{52.3}{60} = .8717, a = \frac{122.6}{9} = 13.62$

Example 2 – even number time point

	Quarters			
Year	1	2	3	4
2004	119	127	127	116
2005	123	142	133	127
2006	146	185	181	161

Estimate a least square trend line of the series.

Year	Q	V	t	Vt	t^2	Trend values
2004	1	119	-11	-1309	121	11.60
	2	127	-9	-1143	81	116.87
	3	127	-7	889	49	122.14
	4	116	-5	580	25	127.41
2005	1	123	-3	-369	9	132.68
	2	142	-1	-142	1	137.9
	3	133	1	133	1	143.22

	4	127	3	381	9	148.49
2006	1	146	5	730	25	153.76
	2	185	7	1295	49	159.03
	3	181	9	1629	81	16430
	4	161	11	1771	121	169.57
		1687		1507	572	

$$b = \frac{1507}{572} = 2.6346, a = \frac{1687}{12} = 140.58, V = 140.58 + 2.6346t$$

Method of moving averages

This method involves obtaining a new series of k- periods moving averages. If a set of n values of a time series is arranged chronologically as v_1, v_2, v_3,.....v_n and we obtain a set averages

$y_1 = v_1 + v_2 + v_3 + ...v_k/k$

$y_2 = v_2 + v_3 + ...v_{k+1}/k$

$y_3 = v_3 + v_4 + ...v_k + v_{k+1} + v_{k+2}/ k$, etc

These averages are called k- point moving averages. The k-point is to show that k observations are used in the averages. The averages are moving because they are the averages of successive k observations.

Example 1

The turnover of a business conglomerate in #m between 1983 and 1976 are given below:

Year	Amount
1983	23489
1984	25276
1985	30827
1986	36375
1987	45635
1988	47648
1989	51678
1990	52883
1991	55016
1992	56998
1993	64287
1994	74012
1995	83485
1996	89658

Obtain 3 - year moving averages.

Year	Amount	Moving total	Moving average
1983	23489	---	----
1984	25276	79592	26530.7
1985	30827	92478	30826
1986	36375	112837	37612.3
1987	45635	129658	43219.3
1988	47648	144961	48320.3
1989	51678	15220	50736.3

1990	52883	159577	53192.3
1991	55016	164897	54965.7
1992	56998	176301	58767
1993	64287	195297	65099
1994	74012	221784	73928
1995	83485	247155	82385
1996	89658	-----	-----

If for practical purposes an even numbered periods has to be used as it applies to 12 month moving averages, 4 - quarter moving averages, e.t.c. then we make the trend values to correspond to true median period by calculating what is called **centred moving averages**. The moving averages are centred by summing up values of two adjacent moving totals, and dividing the resulting values by 2k, (where k is the number of periods in each moving totals).

Example
The consume price indices for food between 1994 and 1996 are given on quarterly basis in the table below:

Year	Q1	Q2	Q3	Q4
1994	119	127	127	116
1995	123	142	133	127
1996	146	185	181	161

Using yearly i.e. 4 – quarterly centred) moving averages, obtain the trend in the food price index.

Year/Q	Price Index	4-Qtr moving total	centred moving total	Centred moving average.
1994 Q1	119	---	---	---
2	127	---	---	---
		489		
3	127		982	123
		493		
4	116		1001	125
		508		
1995 Q1	123		1022	128
		514		
2	142		1039	130
		525		
3	133		1073	134
		548		
4	127		1139	142
		591		
1996 Q1	146		1230	154
		683		

2	185		1312	164
		673		
3	181	---		---
4	161	---		---

Seasonal variation indices

The process of determining the seasonal component of a time series is that of removing the effects of the other components – trend, cyclical, and irregular. Once these other components have been eliminated we calculate in index form a measure of seasonal variation which is called the seasonal variation index.

Seasonal variation indices of a time series may be determine using any one of four methods they are:

1. Average percentage method.
2. Ratio – to - trend method.
3. Ratio – to - moving averages.
4. Link relative method.

DEMOGRAPHIC MEASURES

The term demography was derived from two Greek words: demos meaning, the people and graphein, that is to draw or write. Demography may therefore be defined simply as the science of human population. In a narrow sense demography is concerned with the size, distribution, structure and changes of human populations. In a broader sense demography is a science that studies the size, territorial distribution, structure and composition of human populations and of changes over time in these aspects the causes and consequences of such changes and the interrelationship of social economic factors and changes in the population.

There are three vital processes that cause changes in the size and structure of populations namely fertility, mortality and migration. **Fertility** refers to the actual bearing of children or occurrence of live births. It is differentiated form **fecundity** which refers to the physiological capacity to bear children irrespective of whether or not children have been brought forth.

Mortality deals with the total process of death and the changes it brings about in the population. **Migration** refers to the spatial or geographic movement of populations from one designated area to another.

Demographic measures are the measurement of the likelihood of the occurrence of the three key demographic events (births, death and migration) within a given population.

Sex Ratio

Sex ratio is the ratio of males to females in a given population usually expressed as number of males for every 100 female's

$$SR = \frac{\text{No of males}}{\text{No of females}} \times 100$$

Sex ratio is affected by

----- Sex ratio at birth (always more than 100 with a range from 102- 105).

----- Differential patterns at mortality for males and females.

----- Differential patterns of migration for males and females in population.

Sex ratio for Uganda 2000 population

Age group	Population in 000 male	Population in 000 female.	Sex ratio.
0-4	2376	2350	101
5-9	1983	1972	101
10-14	1628	1614	101
15-9	1277	1265	101
20-24	997	980	102
25-29	807	779	104
30-34	661	644	103
35-39	551	533	103
40-44	394	378	104
45-49	267	278	96
50-54	194	228	85
55-59	161	200	81
60-64	136	163	83
65-69	103	123	84

70-74	75	79	95
75-79	62	59	106
Total	11,671	11,646	100

Age Dependency Ratio

This is the ratio of the person in the dependent ages (under 15 and over 65) to those in the economically productive ages i.e.

$$\frac{P_{0\text{-}14} + P_{65}^{+}}{P_{15\text{-}64}} \times 100$$

The age dependency ratio indicates the relative predominance of persons in the dependent ages in relation to those in the productive ages.

Using the Uganda 2000 data:

Age dependency ratio =114.

Child Woman Ratio

The child woman ratio is a fertility measure computed or based on census data. It is defined as the number of children under age 5 per1000 women of child bearing age in a given year.

$$CWR = \frac{\text{No of children under 5 years}}{\text{No of women ages 15-49}} \times 100$$

Maternal Mortality Ratio

Maternal death is death of a woman

- While pregnant or
- Within 42 days of termination of pregnancy
- Irrespective of the duration or site of the pregnant
- From any cause related to or aggravated by the pregnancy or its management, but
- Not from accidental causes.

Maternal mortality ratio is the number of women who die as a result of complications of pregnancy or child bearing in a given year per 100,000 live births in that year.

Why measure maternal mortality?

1. To establish levels and trends of maternal mortality.
2. To identify characteristics and determinants of maternal deaths.
3. To monitor and evaluate effectiveness and activities designed to reduce maternal mortality.

Crude Birth Rate (CBR)

Number of live births per 1000 population in a given year i.e $CBR = \frac{B}{P} \times 1000$

General Fertility Rate (GFR)

Number of live births per 1000 women ages 15 to 49 in a given year i.e.

$$GRF = \frac{B}{P^{f}_{15\text{-}49}} \times 1000$$

Age Specific Fertility Rate (ASFR)

Number of live births per 1000 women of a specific age group i.e.

$$ASFR = \frac{B_a}{p^f_a} \times 1000$$

Example

Age	Population of females	Live births	ASFR
15-19	1611090	463631	288
20-24	1558276	427298	274
25-29	1425242	412878	290
30-39	1381174	380778	276
40-44	1632695	308671	189
45-49	1400555	239701	178
	10590405	2514118	1666

$$GFR = \frac{2514118}{1059040} \times 1000 =$$

Total Fertility Rate (TFR)

The average number of children that would be born to a woman by the time she ended child bearing if she were to pass through all her child bearing years conforming to the age –specific fertility rate of a given year.

TFR = 5x ∑ASFR/1000

TFR = 5 x [1666/ 1000]

= 5x 1.666

= 8.3 per woman.

= 8 children per woman.

Crude Death Rate (CDR)

The CDR is the number of deaths in a given year per 1000 midyear population i.e.

$$CDR = \frac{D}{P} \times 1000$$

Age Specific Death Rate (ASDR)

The ASDR is the number of deaths per year in a specific age group per 1000 persons in the age group.

$$ASDR = \frac{Da}{Pa} \times 1000$$

Infant Mortality Rate (IMR)

The IMR is the number of deaths of infants under age 1 per year per 1000 live births in the same year i.e.

$$IMR = \frac{\text{D infants}}{\text{Total l births.}} \times 1000$$

Why IMR?

1. The IMR is a good indicator of the overall health status of a population.
2. It is a major determinant of life except only at birth.

3. The IMR is sensitive to levels and changes in socio economic conditions of populations.

The IMR can be divided into

1. Neo natal mortality rate ---- which is defined as the number of deaths of infant under 4 weeks or under 1 month of age during a year per 1000 live births during the year i.e.

$$\text{NNMR} = \frac{\text{No of deaths under 1 month}}{\text{Total live births}} \times 1000$$

2. Post neo natal mortality rate which is defined as the number of infants deaths at 4 through 11 months of age during the year i.e.

$$\text{PNMR} = \frac{\text{No of deaths (1-11 months)}}{\text{Total live births.}} \times 1000$$

REGRESSION AND CORRELATION ANALYSIS

Regression analysis is a statistical tool that utilises the relation between two or more quantitative variables so that one variable can be predicted from one another or others e.g expenditure and sales.

The relationship between two different random variable x and y is known as bivariate relationship. How can we determine whether one variable x is a reliable predictor of another variable y? We must be able to model the bivariate relationship i.e describe how the variables x and y are related using mathematical equation.

If a model is constructed that hypothesised an exact functional relationship between variables it is called a functional or deterministic model i.e $y_i = \alpha + \beta x$

If a model i.e $y_i = \alpha + \beta x + e$

is constructed that hypothesised a relationship between variables allowing for random error it is called a statistical or probabilistic model.

Normally the exact values of the regression parameters: $\alpha, \beta, and e$ are never actually known. From sample data estimates are found. A method useful for modelling the straight line relationship two variables is called simple linear regression model.

With this model the straight line that best fits the set of data points is determined.

Method of least squares

The method of least squares is commonly useful to estimate the simple linear regression model parameters ---

$y_i = \alpha + \beta x + e$

$$\beta(b) = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\alpha(a) = \bar{y} - b\bar{x}$$

Residuals

$e_i = yi - \hat{y}i$

The difference between an observed y value and the mean y value predicted from the sample regression equation.

Standard error of estimate

$$Sy.x = \sqrt{\frac{\sum(yi - \hat{y})^2}{n-2}}$$

The standard error of estimate is used to measure the variability or scatter of the observed sample y values around the sample regression line. It measures the typical difference between the values predicted by the regression equation and the actual y values.

Coefficient of simple determination

$$r^2 = 1 - \frac{\sum(yi - \hat{yi})^2}{\sum(yi - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

The r^2 measures the percentage of the variability in y that can be explained by x.

SST is the total deviation in the dependent variable. This value measures the variability of y without taking into consideration the predictor variable x.

$$SST = \sum(yi - \bar{y})^2$$

SSE is the amount of deviation in the dependent variable that is not explained by the regression equation.

$$SSE = \sum(yi - \hat{yi})^2$$

SSR is the amount of deviation in the dependent variable that is explained by the regression equation.

SSR = SST - SSE.

Example

The following data gives the provisional figures on income and expenditure of a public utility agency in Lagos state for the period 2000—2007 in naira million.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Income (x)	3.9	7.8	7.7	3.1	2.4	4.0	5.2	6.0
Exp(y)	9.1	8.5	11.6	13.5	9.9	12.1	9.9	6.0

1. Determine the sample regression equation of y on x
2. Compute the coefficient of simple determination, r^2 and interpret your result.

x	y	x^2	Y^2	xy	
3.9	9.1	15.21	82.81	35.49	∑y =80.6
7.8	8.5	60.84	72.25	66.3	∑x2 =229.75
7.7	11.6	59.29	134.56	89.32	∑y2 =850.3
3.1	13.5	9.6	182.25	41.85	∑xy =392.6
2.4	9.9	5.76	98.01	23.76	n=8
4.0	12.1	16	146.41	48.4	
5.2	9.9	27.04	98.01	51.58	

6.0	6.0	36	36	36
40.1	80.6	229.75	850.3	392.6.

$$\hat{b} = \frac{8(392.6) - (40.1)(80.6)}{8(229.75) - (40.1)^2} = -.397$$

$$\hat{a} = 10.075 - (-.40)(5.0125) = 12.08$$

$$\hat{yi} = 12.08 - .40x$$

x	y	$\hat{y}$	ei		ei^2
3.9	9.1	10.52	-1.42		2.0164.
7.8	8.5	8.96	-0.46		.2116
7.7	11.6	9	2.6		6.76
3.1	13.5	10.84	2.66		7.0756
2.4	9.9	11.12	-1.22		1.4884
4.0	12.1	10.48	1.62		2.6244
5.2	9.9	10	-0.1		0.01
6.0	6.0	9.68	-3.68		13.5424.
			SSE =	=	33.7288.

yi	$\bar{y}$	$yi - \bar{y}$	$(yi - \bar{y})^2$
9.1	10.075	-.975	.950625
8.5	10.075	-1.575	2.480625
11.6	10.075	1.525	2.325625
13.5	10.075	3.425	11.720625
9.9	10.075	-0.175	0.030625
12.1	10.075	2,025	4.100625
9.9	10.075	-0.175	0.030625
6.0	10.075	-4.075	16.605625.
	SST =		38.255

$r^2 = 1 - \frac{33.7288}{38.255.}$

$= .1183$

Correlation

It is usually desirable to measure the extent of the relationship between x and y as well as observe it in a scatter diagram. The measurement used for this purpose is the correlation coefficient.

This is a value between 1 and + 1 that indicates the strength of the linear relationship between two quantitative variables.

Correlation between variables that are not related is called spurious correlation.

Measures of correlation

1. Karl Pearson's product moment correlation coefficient
2. Spearman's Rank correlation coefficient

Karl Pearson Correlation Coefficient

It is defined as:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

For the population coefficient, the same definition is used except the population size N is substituted for the sample size n. The Karl Pearson correlation coefficient is used for quantitative data.

Spearman's Rank Correlation Coefficient

The spearman rank correlation coefficient is best used when data are in ranks such as those generated in a beauty contest, cooking competition e.t.c.

It is defined as:

$$r = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

where D is the difference in ranks of paired observation and n is the numbers of pairs.

Example

The following data refers to company advertising cost and sales in millions of #:

Adv cost X	s sales. Y	XY	X^2	Y^2
2.6	138	358.8	6.76	19044
3.1	163	505.3	9.61	26569
3.5	166	581.0	12.25	27556
3.7	177	566.1	13.69	23409
4.1	177	725.7	16.81	31329
4.4	201	884.4	19.36	40401
4.6	216	993.5	21.16	46656
4.9	208	1019.2	24.01	43264
5.3	226	1197.8	28.09	51076
5.8	238	1380.4	33.64	56644
42	1886	8212.3	185.38	365948

$$r = \frac{10(8212.3) - (42)(1886)}{\sqrt{10(185.38) - (42)^2}\sqrt{10(365948) - (1886)^2}}$$

$= .96$

Interpretation - strong positive linear correlation between x and y.

Example

Five products A, B, C, D and E are to be test marketed. Ranking obtained from two respondents are presented below:

Product	Respondent 1	Respondent 2
A	1	2
B	4	3
C	2	1

D	3	4
E	5	5

To what extent have the two respondents agreed?

Product	(r_x)	(r_y)	d	d2
A	1	2	-1	1
B	4	5	-1	1
C	2	1	1	1
D	1	4	-1	1
E	5	3	2	4
				8

$$r = 1 - \frac{6x8}{120} = .6$$

Comment - the two respondents have agreed to a reasonable extent.

DESIGN OF SIMPLE EXPERIMENTS

Design and analysis of experiments are very basic aspects of agricultural research. They involve the use of statistical methods in planning and executing the research to ensure necessary data are collected and processed to facilitate valid conclusions. The role of the statistician in experimentation includes the provision of professional advice on several aspects of the experimentation including:

1. the field layout and experimental design to be adopted.
2. the type of data to be collected and mode of collection.
3. the format for recording, summarizing and presenting the data, and.
4. the computations and test of significance to be carried out.

Principles of experimental design

Experimental design refers to the totality of all preliminary steps taken to ensure that appropriate data are obtained to facilitate correct analysis and thereby lead to valid inferences. However, in a narrow sense it refers to the various patterns of arranging the experimental materials. The experimental materials include.

(a) An experimental unit is the material to which a single treatment is applied in one replication of the basic experiments e.g. plot of land and bath of seeds.

(b) sampling unit is used for the fraction of the experimental unit on which a treatment effect is measured. For many experiments, this sampling unit will be the entire experimental unit, but in class where the experiments unit are too large or could be destroyed sampling units are taken within each experimental unit.

(c) treatment refers to any particular set of experimental conditions or factors that could be imposed on an experimental unit for evaluation e.g. brand of fertilizer temperature conditions e.g.

There are three basic principles of experimental design namely randomization, replication and local control.

Randomization

Randomization is a process by which the allocation of treatments to experimental units is done by means of some chances device in order to ensure that no particular treatment is consistently favoured or handicapped. By this all the treatments are given equal chance of being allocated to any particular unit.

Replication

Replication refers to a situation where a treatment is applied to more than one experimental unit. It could also be referred to as the repetition of the basic experiment either over time replicates should always be independent of one another.

Local control

This refers to the amount of grouping or blocking of the experimental units that is employed in the experimental design. It entails grouping the experimental units into blocks such that the units within a block are relatively homogeneous while the units between the blocks are heterogonous. Local control is also turned error control.

Procedure for experimentation

1. State the problem

2. State the objective
3. Design the experiments
4. Performs the experiments
5. Analyse the data and interpret results
6. Prepare the reports of the experiments

Types of experimental design

1. Completely Randomised Design (CRD)
2. Randomised Complete Block Design (RCBD)
3. Latin square design
4. Nested design
5. Cross over design
6. e.t .c.

Completely randomised design

The completely randomised design (CRD) is the simplest type of experimental design. It involves the random allocation of the treatments to the experimental unit without any restriction. Thus the probability of receiving any particular treatment is the same for all the experimental units. The CRD is used only in experiments where the experimental units are homogeneous. The design is also called one way classification design since the homogeneous experimental units are classified according to the levels of only one factor i.e. the treatment.

The statistical model for CRD is a linear additive model of the form:

yi =u + Ti + ei

yi = individual observation (i.e. observation of jth treatment in ith plot).

u = general mean.

Ti = effect of the jth treatment.

ei = experimental error.

ANOVA Models for CRD

Model I - fixed effects model

When the levels of a factor are specifically chosen one is said to have designed a fixed effects model e.g. in our example the four feeds were not randomly selected from a feed catalogue, but were specifically chosen.

Model II - random effects model

The intent here is to generalize, considering the locations. All the calculations are identical to model I, but the null hypothesis is better stated as ho.

Model III - mixed effects model

This model combines the features of the fixed and the random effects model. For some experiments with more than one factor the levels of a certain factor in the same experiments may be random. Such experiments are classified under the mixed effects model.

Randomised complete block design

RCBD is a design used when the experimental units are not homogeneous and thus can be allocated to groups or blocks such that the variation among blocks is maximised while the variation within any particular block is minimised. The blocks are sometimes referred to as replicates.

The major advantage of the RCBD over the CRD is that it yields more precise results due to the grouping of the experimental units into blocks. The linear statistical model for the RCBD is:

$y_{ij} = u + B_i + t_j + e_{ij}$

where

y_{ij} = individual observation

u = general mean

B_i = effect of the i^{th} block

T_j= effect of the j^{th} treatment

e_{ij} = experimental error

ANOVA models for RCBD

Model I factors a and b both fixed.

Model II factors a and b both random.

Model III factors a fixed factor b random

ONE WAY ANALYSIS OF VARIANCE (CRD)

Assumptions

1. the treatment and experimental effects are additive.
2. the experimental errors are randomly independents and normally distributed about zero mean and with a common variance.

Test Procedure

Ho: U1 = U2 = U3 =.....Uk

Hi: the means are not all equal

α*level*

Reject Ho if Fc > F□,

Test statistics:

ANOVA table

Source of variation.	Degree of freedom.	Sum of squares.	Mean square.	Fc.
Treatments.	V1 =k-1	SS treatment	SStreatment/K - 1	MStreatments /MSE
Error	V1 =n-k	SSE	SSE/n-k	
Total.	n-1	SST		

SOME NON PARAMETRIC TESTS

A large body of statistical methods is available that comprises procedures not requiring the estimation of the population variance or mean and not stating hypotheses bout parameters. These testing procedures and termed non parametric test. These methods typically do not make assumption about the nature of the distribution (e.g. normality) of the sampled populations; they are sometimes referred to as distribution free test. Examples of non parametric tests are kruskal Wallis H test, Mann Whitney U test, Sign test, Wilcoxon Rank test, and kolmogorov smirnov test.

The kruskal wallis H test

If a set of data is collected according to a completely randomised design where $k > 2$, it is possible to test non parametrically for difference among groups. This may be done by the kruskal wallis h test, often called analysis of variance by ranks. This test may be used in any situation where the parametric single factor ANOVA is applicable and it will be as powerful as the param