COURSE CODE:	AEM 304
COURSE TITLE:	Applied Statistics for Agriculture
NUMBER OF UNITS:	2 Units
COURSE DURATION:	Two hours per week

COURSE DETAILS:

Course Coordinator: **Dr. A.M. Shittu,** *B.Sc., M.Sc., PhD*
Email: amshittu@yahoo.com
Office Location: **Agric. Econs & Farm Mgt, COLAMRUD**
Other Lecturers: **Dr. Sanusi**

COURSE CONTENT:

Types of applied statistics, Review of hypothesis testing, types and sources of agricultural data, elements of sampling technique, elements of field experimentation, questionnaire design, data processing and spread sheet construction, data presentation, analysis of variance, some non-parametric methods , methods of times series analysis and forecasting, use and application of index numbers, correlation analysis, simple linear regression. This course will emphasis applications of all concepts taught to real agricultural data.

COURSE REQUIREMENTS:

This is a compulsory course for 300 level students in the university. In view of this, students are expected to participate in all the course activities and have minimum of 75% attendance to be able to write the final examination.

READING LIST:

1. Afonja, B (1982) Introductory Statistic: A Learner's Motivated Approach. Evans Brothers (Nigeria publishers) Limited. Ibadan.
2. Bamgboye, E.A. (2002). A Companion of Medical Statistics. FOLBAM Publisher, Ibadan, Nigeria. ISBN-978-056-661-9.258pp
3. Freund, J.E. and Williams, F.J. (1975). Modern Business statistics, 2/e. Pitman Books limited, Great Britain
4. Shittu, A.M. (2001) Statistics: A Study Manual, Interterms Konsultancy Limited. 30/32, Dopemu, Lagos
5. Spiegel, M.R. and Stephens, L.J.(1999) Statistics,3/e. Schaum

LECTURE NOTES

THE NATURE OF STATISTICS AND STATISTICAL METHODS

Facts on individuals or a group or a unit can be referred to as statistics. Examples of such include:

i. the weight of all students in the College of Environmental Resources Management (COLERM), UNAAB could be from 65, 65.5, 60, 60.5, 55, 45, 75, 45.5, 70.5, to 70;
ii. the number of loaves of UNAAB Bread produced every week could 1000, 950, 1200, 3000 for the weeks in April (2011);
iii. the complexion of teaching and non-teaching staff in the College of Agricultural Management and Rural Development (COLAMRUD), UNAAB could be light, dark and mulatto;
iv. the degree of lesions caused by *Cescospora spp*. on maize leaves of 500level COLPLANT Students experimental plots in T&R Farm, UNAAB could be 15%, 25%, 30%, 10%, 5%;
v. the demographic/socio-economic characteristics of cocoa farmers and cocoa farms in Cross River State of Nigeria:-
Age – could be from 55, 35, 40, 70, 60, 80, 25 to 30;
Gender – is male and female;
Marital status – could be single, married, divorced or widowed;
Education – could be primary school, secondary school, tertiary institution or adult education level;
Off-farm occupation – could artisanship, trading, contract jobs or transport business; cocoa output – could be from 1, 3, 4, 5, 8, 6.5, 2.5, 7 to 10 tons;
Severity of black pod disease – could be mild, moderate or severe.

Statistics can be divided into two (2) as follows:

1. *Descriptive Statistics* – deals with collection, summarization and description of statistical facts and figures (data). These include frequencies, percentages, measures of location and measures of dispersion.
2. *Inferential Statistics* – deals with the process of using the information generated from observations and measurements to draw conclusions about the source of facts and figures (data). These include t-statistic, F-statistic, Pearson and Spearman correlation coefficient, regression coefficient and Least Significant Difference

Statistical Terms and Notations

Statistical Terms and Notations

Observations: can be defined as the number with which an event is described or recorded. Observations are the raw materials with which statisticians work. To be useful in statistics, observations are recorded in numerals.

Measurement: is the size or extent of the phenomenon of interest to a researcher (scientist). The scientist observes changes in the phenomenon of interest (dependent or response variable) as a result of changes in other phenomena (independent variables) influencing the phenomenon of interest.

The four (mutually exclusive) levels of measurement are:

i. Nominal
ii. Ordinal
iii. Interval
iv. Ratio

The quality of measurements is very important because policy instruments on agriculture are usually based on agricultural research findings. Hence, it is common practice to assess the quality of the instrument for gathering data by estimating the extent to which the measurements is accurate (i.e. **validity**) and the closeness of measurements to each other when repeated (i.e. **reliability** or **reproducibility** or **precision**).

Variables: it is something or values that changes from individual to individual, group to group or units to units e.g. height varies from individual to individual.

Data: are the information or facts or values collected on all items or characteristics of an individual or group of individuals for any purpose.

Data can be of two (2) categories namely:-

i. *Qualitative Data* – is obtained from item of information that has no notion of numerical magnitude. When a variable has two possible categories, it is called **Binary Data** or **Variable** (e.g. male or female).

ii. *Quantitative Data* – are measurements having numerical magnitude with the variables measured on, at least, the interval scale. They can be discrete or continuous variables.
   a. Discrete variables can take only finite number of values in a given finite interval e.g. bacteria count, anxiety scores, etc.
   b. Continuous variables can assume any value on the real line depending on the precision of the measuring instrument e.g. height, weight, etc.

There are two sources of data namely:-

a. Routine collections – these are data generated from established systems for continuous collection of data. Such systems include census, vital registration system, schools, industries and insurance.
b. Ad-hoc collections – these are special collections usually occasioned by the inadequacy of statistics derived from data collected from routine source. Examples are social survey, epidemiological survey, demographic survey and agricultural survey (such as crop, livestock and village surveys).

There are equally two types of data namely:-

a. Primary data – these are data that are collected by the scientist directly either through survey or experimentation and that have not been analyzed or processed.
b. Secondary data – are data that the scientist gather from publications (such as books, journals, reports, monograph and mimeograph) and have undergone some form of processing or the other.

Population: a population or universe consist all possible values of a variable. The population can be discrete or continuous i.e. countable or uncountable values in a population respectively.

Sampling: most experiment use only a sample or part of a population. For example, maize is planted on a plot of land since it cannot be planted on all the available land in the world.

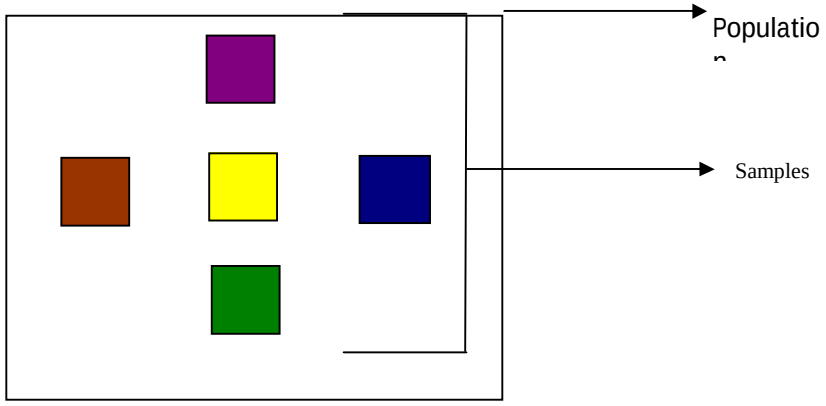

Figure 1: Population and Sample

Distribution: is a set of possible values of a variable together with the number or proportion associated with each value e.g. asking poultry farmers about the level of poultry revenue (variable). Suppose there are g (revenue) groups, associated with revenue is a number of individuals having corresponding revenue level. The number of individuals is what is termed frequency (f), an illustration is given below.

Table 1: Frequency Distribution

Revenue (₦X)	***Frequency***
X_1	F_1
X_2	F_2
X_3	F_3
X_4	F_4
X_5	F_5
⋮	⋮
X_g	F_g

The figures (X_i and F_i; i = 1, 2, ……., g) in Table 1 is called *frequency distribution* while the table (itself) is called *frequency table*.

General Information

The following algebraic notations are used as shorthand in statistics:

i. Variables are denoted by algebraic letters like A, B, D, X, Y, Z. For example the variable ***age*** can be represented by the letter ***X***.

ii. Population values are indicated by capitals e.g. X_i is the value of the i^{th} member of population ***X***

iii. Sample values are denoted by small e.g. x_i is the value of the i^{th} unit of the sample ***X***. For example, if age is the variable of interest, x_i is the age of the i^{th} person in the sample. For instance, x_3 is the age of the 3^{rd} person in the sample.

The summation of all values of a variable is denoted by the Greek letterΣ. For example, to add all values of the variable x_i i.e. $x_1 + X_2 + x_3 + ……. + x_n$, can be simply written as

$$\sum_{i=1}^{n} x_i \text{ ---------- (1)}$$

Equation 1 implies summing the observations on variable x from the first to the n^{th}.

Methods of Data Collection and Presentation

Data Presentation: is the process whereby data collected are processed, summarized and presented in a useful form. Perhaps, the most important of all data summarization technique is the *histogram* – a graphical representation of the frequency of a variable e.g. data on Table 1.

Data Collection: there are four (4) basic methods of data collection. It could be:-

1. by merely making observation e.g. the traffic count on a market day.
2. By measuring and recording e.g. corn plants on the field or measuring rainfall with rain gauge.
3. Use of existing records i.e. secondary data e.g. ADPs, Library, publications of institutions such as Central Bank of Nigeria (CBN) and National Bureau of Statistics, annual reports, etc.
4. Using other people's experimental results.

Data Presentation: is the process whereby data collected are processed, summarized and presented in a useful form. These include

i *histogram* – a graphical representation of the frequency of a variable.

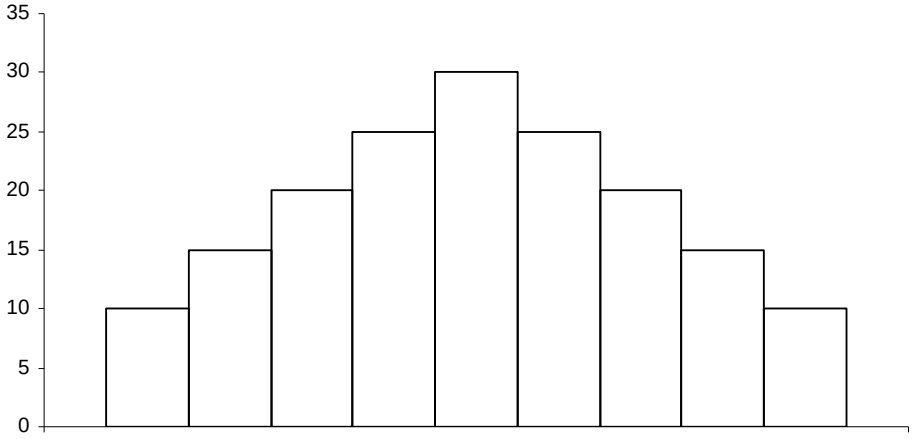

Figure 2: Histogram

ii. *Frequency table*. The table may include a table of frequency which shows at a glance a specific value and the number of times it occurs. For example:

Table 2: Hypothetical Frequency Distribution

Age Group	***Mid point***	***Frequency***
20 - 25	22.5	57
26 - 30	28.5	10
31 - 35	33.5	19
36 - 40	38.5	28
41 - 45	42.5	36

iii. *Charts* which can be

a. circular chart (i.e. pie chart)

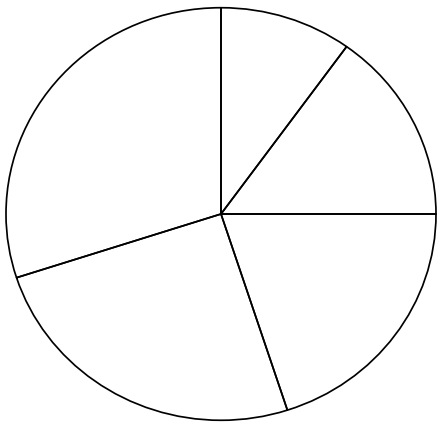

Figure 3: Circular Chart

c. Bar chart

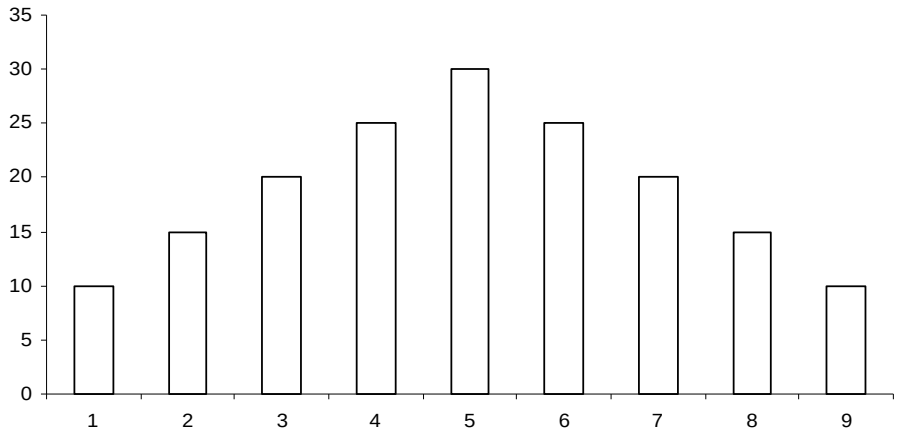

Figure 4: Bar Chart

c. Lines

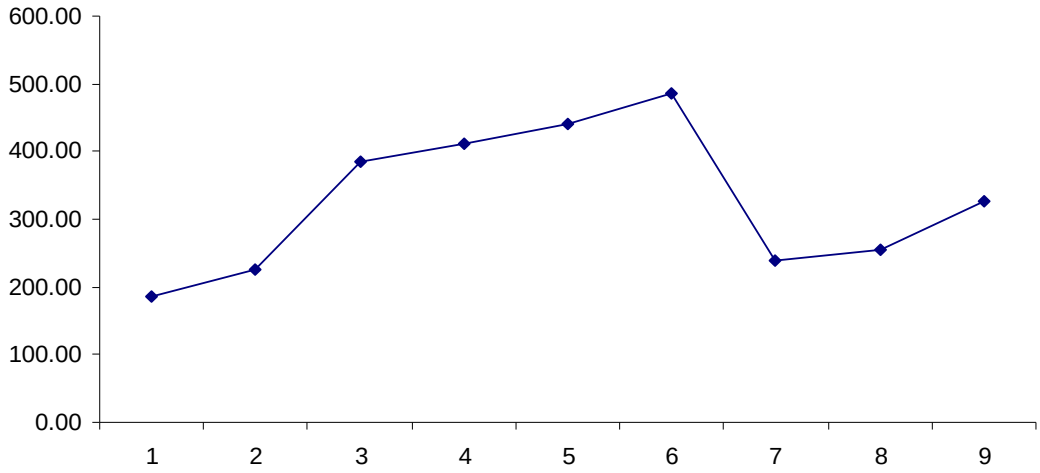

Figure 5: Line Graph

Relative frequency of a class is the frequency of a class divided by the total frequencies of the classes and is often expressed as a proportion.

Table 3: Relative Frequency Table

Variable (x_i)	***Frequency (f_i)***	***Relative Frequency (p_i)***
22.5	57	57/150 = 0.38
28.5	10	0.06
33.5	19	0.13
38.5	28	0.19
42.5	36	0.24
Total	***150***	***1.00***

Summarizing Data

Data can be summarized through the use of certain measures which can be broadly categorized into two (2) as follows:
(1) Measures of location or central tendencies;
(2) Measures of dispersion or spread.

Measures of Central Tendency (Location): They serve to give an idea of a typical, the representative or ordinary value of a distribution. These are mode, median and mean.

(A) *Mean*:

a. Arithmetic mean – This is represented by two (2) symbols μ (population mean) and $\bar{x}$ (sample mean).

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

Where:
N = population size;
X_i = the variable (e.g. age, gender, etc) of interest of the i^{th} member of the population;
i = 1, 2, 3, ..., N.

$$\bar{x} = \sum_{i=1}^{n} x_i (n)^{-1} \text{----------} \quad (3)$$

Where:
n = sample size;
x_i = the variable (e.g. age, gender, etc) of interest of the i^{th} member of the sample;
i = 1, 2, 3, ..., n.

b. Geometric mean: is represented by X_G and is defined as the N^{th} root of the product of all the N observations i.e. $X_G = \sqrt[N]{(x_1)(x_2)(x_3).....(x_n)}$ ---------- (4)

e.g. N = 4, $X_1 = 5$, $X_2 = 8$, $X_3 = 10$, $X_4 = 12$

$X_G = \sqrt[4]{5 \bullet 8 \bullet 10 \bullet 12} = (4800)^{-4} = 8.32$

d. Harmonic Mean X_H: This the reciprocal of the arithmetic mean, of the reciprocals of the observation.

Example

Given that x = (4, 5, 2, 10), find the harmonic mean of the distribution.

Solution

i. Reciprocals of observations = 1/4, 1/5, 1/2, 1/10 or $\frac{1/4, 1/5, 1/2, 1/10}{4}$

ii. Σ RO = 1.05/4

iii. $1/\Sigma$ RO = 1/(1.05/4)

= 4/1.05

X_H = 1/0.2625 or 10/2.625

$$X_H = n \Big/ \sum_{i=1}^{n} (1/x_i) \text{ ----------} \quad (5)$$

This is used mostly by price analysts.

d. Quadratic Mean X_Q: It is the square root of the sum of the observation squared divide by the total number of observation.

$$\Rightarrow X_Q = \{\sum_{i=1}^{n} x_i^2/n\}^{1/2} \equiv \{\sum_{i=1}^{n} x_i^2/n\}^{1/2} \text{ ----------} \quad (6)$$

N.B: $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$

$\sum_{i=1}^{n} x_i^2$ = Sum of squares

$(\sum_{i=1}^{n} x_i)^2$ = Squares of sum

i=1

Generally speaking: HM < GM < $\bar{X}$

B. *Median*: arrange all values either in ascending or descending order and pick the middle number when dealing with an odd number of values. When dealing with even number of values, the mean of the 2 middle numbers is the median.

C. *Mode*: is the number that appears most frequently. This is taken from the frequency.

2. ***Measures of Spread (Dispersion)***: The measure of central tendencies discussed above only give a partial summary of the information in a set of data. To complete the description of a distribution, information is needed on how far away the observations are from their central value.

Illustrations:

Assuming the age range of staff in COLAMRUD and COLPLANT is given by the frequency distribution in Table 4. The two Colleges are compared using the mean and graphical method.

Table 4: Age Distribution of Staff in COLAMRUD and COLPLANT, UNAAB, Abeokuta

Age Group	*Mid-point (Age)*	*COLAMRUD (Frequency)*	*COPLANT (Frequency)*
19 – 21	20	2	4
22 – 24	23	4	3
25 – 27	26	8	4
28 – 30	29	16	26
31 – 33	32	7	3
34 – 36	35	4	3
37 – 39	38	3	5

Method 1:

$\bar{X} = (\Sigma f_i x_i)[(\Sigma f_i)^{-1}]$

where:-

Σ = sum of;

f_i = frequency;

x_i = mid-point values.

$\therefore$ Mean ages are 29.14 and 29.13 for COLAMRUD and COLPLANT staffs respectively.

Method 2:

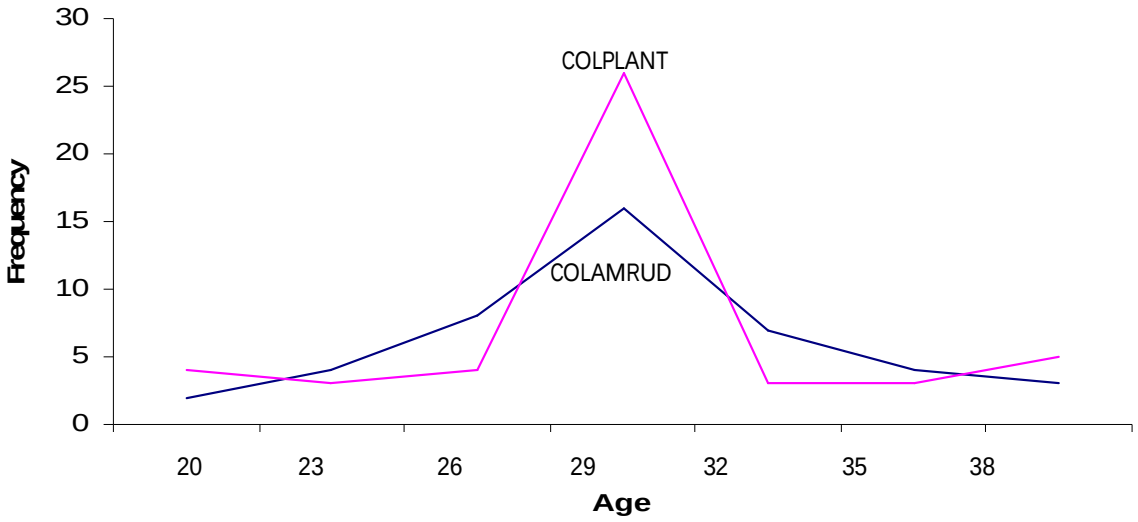

Figure 6: Line Chart for COLAMRUD and COLPLANT STAFFERS AGE DISTRIBUTION

From the above results, it can be seen that the mean age for staff in the two colleges is (approximately) 29 years. Meaning that the colleges are equal in terms of mean ages, however,

Figure 6 showed that the ages of staffers in COLPLANT are more concentrated around the central value (the mean) while that of COLAMRUD are not as concentrated. In other words, ages of staffers in COLAMRUD spread more widely than that of COLPLANT. Given these, it implies that more information is needed on the dispersion of data (ages) from the mean (age).
There are five basic measures of dispersion namely:-
A) ***Variance***: It is the second moment about the mean and is a measure of spread. For a discrete population of N individuals, the variance is given by –

$$\sigma^2 = (\sum_{i=1}^{N} x_{ii} - \mu)^2 (N)^{-1} \quad \text{---------- (7)}$$

The sample variance is given by –

$$S_x^2 = (\sum_{i=1}^{n} x_{ii} - \overline{x})^2 (n-1)^{-1} \quad \text{---------- (8)}$$

Where: $\overline{x} = {}^1/_n \Sigma x_{ii}$ ---------- (9)

B) ***Standard Deviation***: is the square root of the variance (definitional formula)
i.e. $SD = (S_x^2)^{½} \equiv \sqrt{S_x^2} = S_x$
Note: S_x can never be negative.
C) ***Variance of the mean***: is given by

$$\sigma^2/N = (\sum_{i=1}^{N} x_{ii} - \mu)^2 / \{N(N-1)\} \quad \text{---------- (10)}$$

The sample variance is given by –

$$S_x^2/n = (\sum_{i=1}^{n} x_{ii} - \overline{x})^2 \{n(n-1)\}^{-1} \quad \text{---------- (11)}$$

D) ***Standard Error***: the square root of the mean (definitional formula) i.e.

$$SE = \sqrt{S_x^2}/n = S_x/\sqrt{n} = S_x(\sqrt{n})^{-1} = S_x(n)^{-1/2} \quad \text{---------- (12)}$$

F. Range: is the simplest measure of dispersion and it is simply the difference between the lowest value and the highest value of the observations on the variable of interest. This could be given as:

$R = X_U - X_L$ ---------- (13)
Where:-
X_U = highest value;
X_L = lowest value.
For example, the range for the observations in c above can be computed as: 10 – 2 = 8.

ELEMENTARY PROBABILITY THEORY

One of the important tools of statistics is probability. Although it is applied to a variety of practical situations. An understanding of the subject is made simpler if it is applied to practical situations like games of chance e.g. tossing of a coin which could result in either head or tail or rolling of an ordinary dice which could result in any of the six (6) sides i.e. 1, 2, 3, 4, 5 and 6.

Trial: is a random experiment e.g. the rolling of the dice (or tossing of the coin).
Exhaustive Events: is the group of all possible results of a random experiment that can occur apart from no other can be obtained e.g. the six (or two) possible results from rolling of a dice (or tossing of a coin).
Mutually Exclusive Events: is a result of a random experiment that can only be obtained at a i^{th} particular time.
Equally Likely Event: All possible results of a trial having the same chances of occurrence.

Definitions of Probability

Classical (*a priori*) definition: if a trial can result in any one of the exhaustive, mutually exclusive and equally likely outcomes and if m of these outcomes entails an occurrence of an event E thus the probability that E will happen as a result of that trial is given by:-

$P(E) = \frac{m}{n}$ $\quad m \leq n \quad 0 \leq P(E) \leq 1$ ---------- (16)

$P(E) + P(E^1) = 1$ or $P(E^1) = 1 - P(E)$

N.B:

(i) $P \geq 0$

(ii) The sum of all probabilities in an experiment = 1 i.e. $EP \equiv \Sigma P = 1$

(iii) $0 \leq P(E) \leq 1$

Statistical (empirical) definition: in this particular case probability is a ratio based on empirical information. It is viewed in actual fact as the result frequency of a particular event in a very long frequency of trials e.g. tossing a coin 100 times. If it is a fair coin, a tail has the same chance as a head. The resultant frequency become more consistent as number of trials increase. When number of trials is small there would more variation in the value of m/n = P but as n becomes large the observed value of m/n will closely converge around some central value.

Probability and Probability Distribution

Sets and Space: Set is an element of differentiated and well distinguished objects or members or elements.

S_1 = [1, 2, 3] is a well distinguished set

S_2 = [1, 2, 3, 2, 3] = S_1 because S_2 have three well distinguished elements

The order of listing of elements in a set is not important. A set can be specified either by listing all its elements or by giving a rule which will enable the determination of whether any giving object does not belong to it.

A null or an empty set is one with no number or element in it or simply – S = [ ] or S = Ø.

A universal set is the underlined universe of discourse which serves as a frame of reference for any specification of a set.

If every element in S_1 is an element in S then S_1 is a subset of S ---------- 1;

$S_1 \subseteq S$ ----------2.

If S_1 = {1} and S = {1, 2, 3}:- S_1 obey rule 1, hence S_1 is called a proper subset of S.

If S contains at least an element not S_1 it also obey rule $S_1 \subseteq S$.

$\emptyset$ is also a proper subset of S, $\emptyset \subseteq S$.

There are two important concepts in probability theory and these are:

1. Union of 2 sets $S = S_1 \cup S_2$ – this is defined as the set of elements that belongs either to S_1, S2 or both. e.g.:-
   S_1 = {a, b, c, 2}$\cup S_2$ {1, 2, 3} = S. Since S = [a, b, c, 1, 2, 3]
2. The set of elements it belong to both S_2 and S_2 e. g above example $S_1 \cap S_2 = 2$

The Algebra of Sets

The algebra of set sets is based upon a few postulates or laws and this include commutative and associative laws.

Considering 3 sets S_1, S_2 and S_3 as subsets of S:

The commutative law states that $S_1 \cup S_2 = S_2 \cup S_1$ and $S_1 \cap S_2 = S_2 \cap S_3$;

The associative law states that $(S_1 \cup S_2) \cup S_3 = S_1 \cup (S_2 \cap S_3)$ and $(S_1 \cap S_2) \cap S_3 = S_1 \cap (S_2 \cap S_3)$.

Sample space: is a set whose set represents well distinguished outcome of an experiment.

Sample space of the experiment of tossing a coin consists of 2 elements corresponding to {H, T} and tossing a die corresponds to {1, 2, 3, 4, 5, 6}.

Discrete Sample Space: sample space consisting of finite or infinite but countable number of elements.

Continuous Sample Space: the opposite of discrete sample space.

Basic Theorems of Probability Theory

A, B, C, are events of a discrete sample space S, hence the respective probabilities are p(A), p(B), p(C):

$\therefore 0 \leq p(A) \leq 1$ and $p(S) = 1$.

Theorem 1 – If A^{I} is an event not A, then $p(A^{I}) = 1 - p(A)$.

Theorem 2 – $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

NB:- $p(A \cap B)$ implies probability of simultaneous occurrence.

Theorem 3 – If A and B are mutually exclusive events, then $p(A \cap B) = p(A) + p(B)$;

Also $p(A \cup B \cup C \cup D \cup \ldots\ldots \cup Z) = p(A) + p(B) + p(C) + p(D) + \ldots\ldots p(Z)$.

Non-mutually Exclusive Events: For example, a graduate and the graduate's gender.

NB: - G means graduate and G^{I} means non-graduate.

Table 5: Sample Space for Gender and Educational Level

Gender	G^{I}	***G***	***Total***
M	$M \cap G^{I}$	$M \cap G$	P(M)
F	$F \cap G^{I}$	$F \cap G$	P(F)
Total	$P(G^{I})$	***P(G)***	***1***

The probabilities in the body of the table represent by intersections are called *Joint Probabilities* e.g. $p(M \cap G)$ and those that appear in the last row and column are called *Marginal Probabilities* e.g. $p(G^{I})$ and p(M). p(G) is the probability of a graduate regardless of gender, p(F) is the probability of female regardless of level of education.

$p(M \cup G) = p(M) + p(G) - p(M \cap G)$.

Table 6: Sample Space for Smoking of Habit and Gender Distribution

Gender	***S***	S^{I}	***Total***
M	a	b	a + b
F	c	d	c + d
Total	***a + b***	***b + d***	***N***

The sample space here contains {(MS), (fs), (ms^{I}), (fS^{I})} i.e. 4 elements namely male smokers, female smokers, male non-smokers and female non-smoker.

N = a + b + c + d.

Addition theorem: $p(M \cup S) = p(M) + p(S) - p(M \cap S) = 1 - p(F \cap S^{I})$

$p(M \cup S)$ = either male or smokers which are male smokers, male non-smokers and female smokers.

$p(M \cap S)$ = male smoker.

Conditional Probability

Supposing the focus is to determine the probability that a person of a given gender is a smoker (non-smoker) or that a chosen smoker is a male (female) such probability is called *Conditional Probability* written as p(S/M) or p(M/S) meaning probability of S given M or probability of M given S.

In a finite population this probability is given by:

$$p(S/M) = \frac{\text{Total number of male smokers}}{\text{Total number males}}$$

$$p(S/M) = \frac{\text{smokers (males)}}{\text{Males}}$$

From Table 6:-

$p(S/F) = \frac{\text{female smokers}}{\text{females}} = \frac{c}{c+d}$

$p(M/S) = \frac{a}{a+b}$

$p(M/S^{|}) = \frac{b}{b+d}$

$p(S^{|}/M) = \frac{b}{c+b}$

$p(S^{|}/F) = \frac{d}{c+d}$

$p(F/S) = \frac{c}{a+c}$ a + b

$p(F/S^{|}) = \frac{d}{b+d}$

NB:

i) $p(S/M) + p(S^{|}/M) = 1$

ii) $p(S/F) + p(S^{|}/F) = 1$

iii) $p(S/M) = \frac{p(M \cap S)}{p(M)}$

iv) $p(S^{|}/M) = p(S^{|} \cap M)$ p(M)

v) $p(M/S) = \frac{p(M \cap S)}{p(S)}$

Vi) $p(M/S^{|}) = p(M \cap S^{1})$ $p(S^{|})$

Theorem 4: (*Conditional Probability Theorem*): if A and B are subset of a discrete sample space and $p(B) \neq O$, then:-

$p(A/B) = \frac{p(A \cap B)}{p(B)}$

N.B:- unlike $p(A \cap B) = p(B \cap A)$ in all respect $P(A/B) \neq p(B/A)$ i.e. provide $p(B) \neq p(A)$.

Since the CP theorem can always be written as $p(A \cap B) = p(A/B)\{p(B)\}$ is called multiplication theorem.

Further Principles Probability

Tossing a coin a first time, the probability of a head is ½ and tossing the second time, what is the probability of having a head the second time having already gotten a head in the first toss?

$p(H2/H1) = \frac{p(H_2 \cap H_1)}{p(H_1)} = \frac{p(H_1) \bullet p(H_1)}{p(H_1)} = \frac{½ \cdot ½}{½} = ½$

If A is independent of B:

$p(A/B) = \frac{p(A) \bullet p(B)}{p(B)} = p(A)$

i.e. if and only if A is independent of B i.e. the conditional p(A/B) is equal to the marginal probability of (A):

$\Rightarrow p(A) = \frac{p(A \cap B)}{p(B)}$ or $p(B) = \frac{p(A \cap B)}{p(A)}$

i.e. if A is independent of B then B is also independent of A. This leads to the 5th theorem called *Independent Theorem*.

Theorem 5 (*Independent Theorem*): If the $p(A) \neq 0$ and $p(B) \neq 0$ then $p(A \cap B) = p(A) \bullet p(B)$

$\therefore p(A \cap B \cap C \cap D \cap \ldots\ldots\ldots \cap Z) = p(A) \bullet p(B) \bullet p(C) \bullet p(D) \bullet \ldots\ldots\ldots\ldots \bullet p(Z)$

NB: Independent events are not mutually exclusive and neither mutually exclusive events independent.

1) If two events are mutually exclusive - $p(A \cap B) = 0$
2) If two events are mutually independent - $p(A \cap B) = p(A) \bullet p(B)$
3) If two events are mutually exclusive - $p(A \cup B) = (A) + (B) - 0$
4) If two events are mutually independent - $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

Mutually Exclusive: Occurrence of one event prevents the occurrence of the other.
Independency: Occurrence of one event in no way affects the occurrence of the other.

PROBABILITY DISTRIBUTION

A probability distribution (PD) or probability function is the assemblage of values of x_i with their associated probabilities e.g.:

Table 7: Probability Distribution for Throwing of a Fair Dice

x_i	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Table 8: Probability Distribution for Tossing of a Fair Coin

x_i	H	T
P	½	½

References about PD imply the mean and variance of the PD.

The mean of a PD is called the expected value and $E(x) = \mu = \sum_{l=1}^{n} p_i x_i$.

x_i denotes a discrete random variable which can assume the value x_1 to x_n with respective probabilities p_1 to p_n

Recall that $\sum_{l=1}^{n} p_i = 1$; hence

$\Sigma(x) = p_1x_1 + p_2x_2 + \ldots\ldots + p_nx_n$

The variance of the PD is symbolically written as $\sigma^2 = \Sigma[x - E(x)]^2$

SOME COMMON DISTRIBUTION

Binomial Distribution

In this distribution the probability of occurrence of an event is the same or constant from trial to trial for every individual and there are only 2 outcomes which is a Yes or No to a question or to the presence or absence of a qualitative characteristic. The BD is determined by 2 parameters and these are *n* which is the sample size and *p* the probability of occurrence of the event. Symbolically, x: b(n,p) implies that x is binomially distributed i.e. the probability *p* of observing x_i out of *n* trials.
This probability is given by:

$$P(x = x_i) = \binom{n}{x_1} p^{x}i\,(1-p)^{n-x}i$$

The Mean and the Variance of Binomial Distribution

The Binomial distribution is computable only if *n* and *p* are known. Hence, the BD is completely determined by two quantities *p* and *n*. *p* could be probability of success and *n* the number of trials.
The mean [$\Sigma(x)$] of a binomially distributed variable is defined as the average number of success that can be expected in a long sequence of repetitions of a binomial experiment.
Recall that:

$E(x) = \mu = \Sigma x_i p(x_i)$

Since the binomial distribution is determined by its parameters then:

$E(x) = \mu = np$

By similar algebraic manipulation:

The variance of BD (σ^2) = npq where q = 1 – p

$\Rightarrow \sigma^2 = np(1 - p)$

SD(σ) = $\sqrt{npq}$

When *n* is large and *p* is small, the BD becomes cumbersome i.e. as *n* approaches infinity ($n \to \infty$) and *p* approaches 0 ($p \to 0$), np is fixed and the BD approaches the limiting form i.e.

$$P(x:\mu) = \frac{e^{-\mu}\mu^{x}}{x!} \qquad (x = 0, 1, 2)$$

e = 2.71828

x = number of successes;

μ = as a single parameters.

This distribution is called Poison distribution. The binomial distribution cannot be used when (naturally) n > 100 and p < 0.05.

PROBABILITY DISTRIBUTION OF CONTINUOUS RANDOM VARIABLES

A random variable that can assume any value in an interval of values number matter how small the value is said to be continuous. Probabilities associated which CRV are measured for intervals of values of the variable and are given by areas under the probability curves of the variable. If a very large number of observations are made on a continuous random variable x and a relative frequency distribution with a large number of classes of uniform width is constructed, the resulting probability histogram approaches (pictorially) a smooth curve as the number of observations and classes increases.

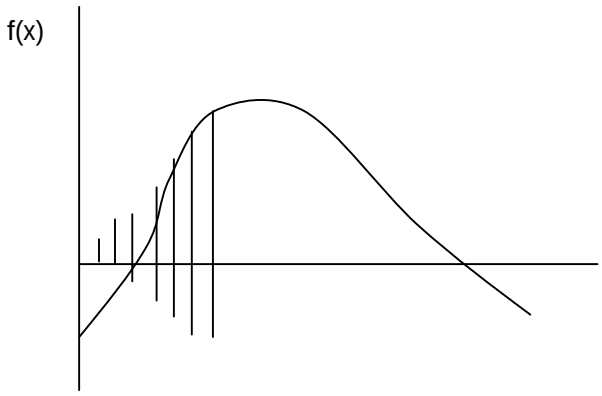

Figure 7: Probability Histogram of CRV

Each value of x_x can be paired with only one value on the curve. The mathematical relationship defining what p(x) is for every value of x is denoted by fx. For any interval dx where x is very small, the probability which is equal to area of choosing point within that interval is f(x)dx i.e. height x width. The probability of obtaining a point or value within a large interval say b to a is given by:

$$\int_a^b f(x)dx$$

Probability function of CRV is called probability density function or probability density. An example of a continuous distribution is the normal distribution.

Normal Distribution

Normal Distribution in many respects is the cornerstone of modern statistical theory. It is sometimes called *Laplacian* or *Gaussian Distribution* since it was first studied by Pierre Laplace (1949-1829) and Carl Gauss.

Properties of Normal Distribution

1. Discrete variable has become continuous and the frequencies have merged.
2. The distribution is determined by 2 parameters i.e. the mean (meu) μ and the standard deviation (SD) sigma (σ). The mean locates the centre of the distribution while the SD measures the spread.
3. It is symmetrical about the mean value i.e. ½ of the distribution or ½ of the total area under the curve lies on each side of the mean and the curve possesses a shape much like that of a bell.

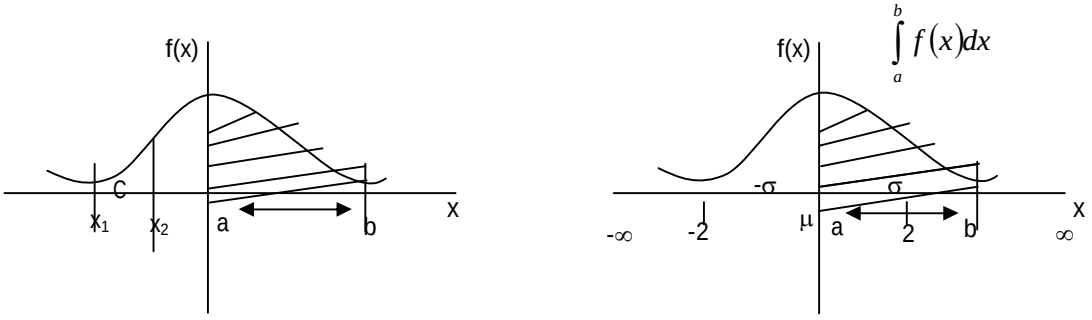

Figure 8: Properties of CRV

4. The total area below the curve and above the x – axis is equal to one.

Normal Curve: the mathematical formula is given as:-

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

This unlike the binomial case does not give probability directly but merely describe the curve. The magnitude of the area C between x_1 and x_2 in the diagram above gives the probability that a randomly drawn individual will lie between x_1 and x_2. In practice, instated of computing areas each time when probability is needed, a table is used to obtain probabilities. In computation, the variable x will be expressed in terms of a standard unit, therefore a standard normal variate represented by Z is needed to be defined and it is that variation that is normally distributed with a mean (μ) = 0 and (a constant) variance (σ^2) = 1.

$\Rightarrow$ Z: N (μ, σ^2) i.e. Z: N (0, 1).

The formula given above now becomes:-

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

Hence, $Z = \frac{x-\mu}{\sigma}$

t – Distribution

Usually called the student t, it was discovered by W. S. Gosset in 1908 and was perfected by R. A. Fisher in 1924. Recalling the fact that the normal distribution is determined by 2 parameters i.e. the mean and the variance which are estimated by the sample mean and the sample variance respectively. Obviously, if a statement is made about an observation x in X where X: N (μ, σ^2); μ and

σ^2 have to be known, which involves the use of the population values N. Hence, since population figures are hard to come by, Z cannot be used.
Since researchers often deals with small samples, t provides the tools for handling such samples and it makes use of $\bar{x}$ and S_x^2.

$$Z = \frac{x - \mu}{\sigma} \text{ while } t = \frac{\bar{x} - \mu}{S_{\bar{x}}}$$

Properties of t – Distribution

1. It is symmetric about the mean but it is not normally distributed.

2. It has different distribution for different sample sizes because of the $S_{\bar{x}}$ (SE) (standard error).

Test of Hypothesis and Confidence Interval

In using the t-distribution, the aim is to be able to say whether x $\bar{}$ is close enough to μ for any difference to be due to chance. The assumption is usually stated in the form of a hypothesis. A test hypothesis is a procedure for deciding whether to accept or reject the hypothesis. In practice, there are 2 hypotheses – the null hypothesis (H_o) and the alternative hypothesis (H_a).
The postulate is that the sample and population mean are not (significantly) different from each other. The null hypothesis is one for which it is possible to compute a t-statistic (called t-calculated/computed) and the corresponding probability of a more extreme value taken from a table (called t-tabulated).
The t-statistic can be computed with the expression stated below:-

$$t_c = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Furthermore, the hypothesis can be stated thus:-
H_0: $\bar{x} = \mu$
H_a: $\bar{x} \neq \mu$
t_c represents t calculated and t_T represent tabulated or critical t. t_T is usually expressed as $t_{T,\alpha,(n-1)}$. H_0 is accepted if $t_c < t_T$ at a given degree if freedom (n – 1) and level of significant (α).
The critical value of the t – distribution can also be used to calculate an appropriate confidence interval. If the hypothesized mean (μ) lies outside the confidence interval (CI), H_0 is not accepted and if it lies inside the CI, the H_0 is accepted.

$$\bar{x} \pm \frac{t_c S_x}{\sqrt{n}} = \bar{x} \pm t_c \bullet (S_{\bar{x}})$$

$$\bar{x} - t_c(S_{\bar{x}}) \leq \mu \leq \bar{x} + t_c(S_{\bar{x}})$$

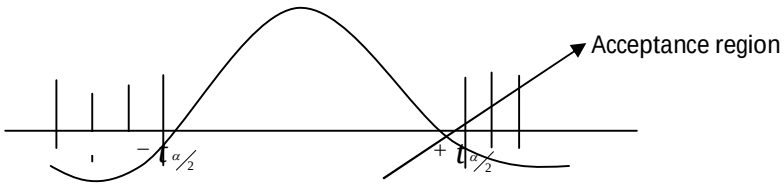

Figure 9: Confidence Interval

MEASURES OF ASSOCIATION

Science, be it physical, biological or social is based or centred fundamentally on association or relationship which can be simple or complex. In many scientific areas of study, the researcher is often interested in watching or investigating how changes affects another e.g. effect of fertilizer on yield, effect of education on adoption and effect(s) of price or income on quantity demanded.

The statistical techniques which are used in these studies or relationships vary depending on the type of association. The different types include correlation coefficient (ρ), coefficient of determination (ρ^2), regression coefficient, chi-square (χ^2) and Analysis of Variance (ANOVA).

Correlation Coefficient

A measure of association between 2 variables. It is represented thus:

Definitional formula $\rho(x, y) = \frac{Cov(xy)}{\sqrt{Var(x)\text{var}(y)}}$

$$\rho = \frac{\sum xy}{\left[\left(\sum x^2\right)\left(\sum y^2\right)\right]^{1/2}}$$

Where:

$y = y_i - \bar{y}$

$x = x_i - \bar{x}$

$$\rho = \left\{\Sigma xy - \{[(\Sigma x)(\Sigma y)]\ n\}\right\} \Big/ \left\{\sqrt{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right)}\right\}$$

Covariance is a measure of joint variability while the correlation co-efficient measures the relationship between 2 variables.

Properties of ρ (Rho)

1. $-1 \leq \rho \leq 1$.
2. It considers the 2 variables as co-equal
3. Between 0 and +1 there is either direct, positive association, relationship or correlation.
4. When the value is almost zero the relationship is weak or low.
5. Between 0 to -1 there is indirect, inverse or negative association, relationship or correlation.
6. When the value is almost -1 or +1, the relationship is said to be high or perfect e.g.:
   if $\rho = 0.90 \Rightarrow$ perfect/close but direct relationship;
   if $\rho = -0.90 \Rightarrow$ perfect/close but indirect relationship.

N.B: ρ must not be more than 1.

Coefficient of Determination ρ^2

If the correlation coefficient is squared, the coefficient of determination (CD) is obtained. CD indicates the amount of variation in one variable that is explained by the other variable as a result of their linear relationship. For example:-

If $\rho(xy) = 0.5$

$\rho^2 = 0.25 \quad \Rightarrow 25\%$

This means that given x, 25% of the variation in Y can be explained as a result of the association between x and y. If x is price and y is quantity demanded of commodity and q(d) = f(P); then 25% means given price, 25% of the changes in quantity demanded due to relationship between price and quantity demanded can be explained.

Regression Analysis

Correlation coefficient measures the joint association between 2 variables while regression analysis estimates the amount of change that is expected in the dependent variable when the independent is altered i.e. in correlation the 2 variables are correlated but in regression one depends on the other. Regression analysis is particularly useful in:

a) predicting the value of a dependent variable given the value of an independent variable;
b) in measuring the degree of association between 2 variables;
c) in testing of hypothesis which respect to the significant of the association.

Given the simple linear function (equation):

$$Y_i = a + bXi + e_i$$

Where (by convention):-

X_i = independent (exogenous) variable;

Y_i = dependent (endogenous) variable;

a = constant or Y – intercept i.e. the value of Y when x = 0;

b = slope of the line or regression coefficient i.e. the amount by which Y changes when X changes.

e_i = stochastic/random error term.

Example: Suppose the cost of renting a tractor (in ₦'000) is ₦25 and a charge of ₦0.3 for each km the tractor is driven. The data can be summarised as below.

C = 25 + 0.3x

Where:

c = cost of renting the tractor;

x = number of km.

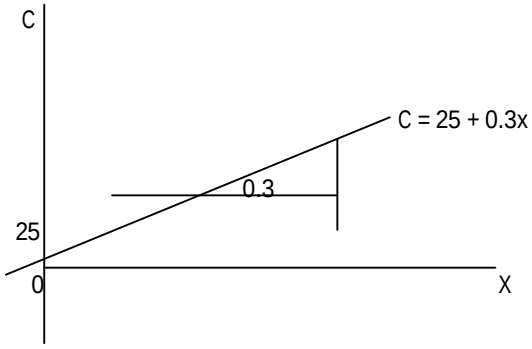

Figure 10: Regression Curve

The Y intercept is 25 i.e. ₦25,000 is incurred irrespective of whether the tractor is used or not. The slope is 0.3 i.e. total costs for a day rent increase by ₦300 for each additional km the tractor is driven.

Fitting a Straight Line by Least Square Method

In scientific studies data are collected in pairs rather than generating data from a given mathematical equation. If the paired data are plotted, as shown below, the resulting pictorial representation is called *scattered plot* or *scattered graph* or *scattered diagram*.

Table 11: X and Y Data Distribution

Y	X
Y_1	X_1
Y_n	X_n

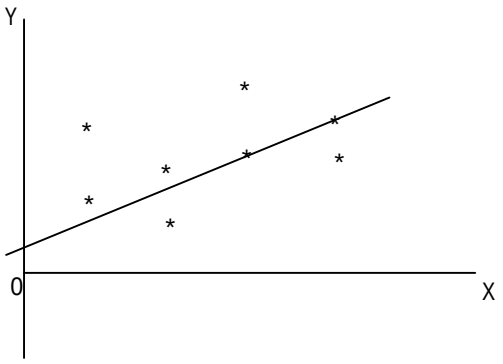

Figure 11: Scattered Diagram

The essence of the exercise above is to obtain a straight line that is called the best fit line or regression line. Statisticians define the best fit line to be the line for which the sum of the squares of errors has the smallest possible value. An error is the vertical distance between an actual point and the point directly below or above it on the estimating line. For a given set of observed data, different

lines have different sum of squares of errors. The best fitting line out of the array of lines is the one having the smallest possible sum of square errors. The best fitting line is also called the least square line.

The least square line can be determined by calculation performed on the XY data pairs. The calculation leads to a value for the slope *b* and for intercept *a*. After *a* and *b* have been computed, the values are substituted into the equation $Y = a + bX_i$ to obtain the estimating equation. The line and estimated equation are often called least square regression.

By mathematical analysis it has been proofed that *a* and *b* for a least square regression line can be computed from the following formula:

$b = (\Sigma XY) \; [(\Sigma X)^2]$

Where:-

$y = y_i - \bar{y}$

$x = x_i - \bar{x}$

$$b = \left\{ \Sigma x_i y_i - [(\Sigma x_i)(\Sigma y_i) \, n \right\} \Big/ \left\{ \Sigma x_i^2 - [(\Sigma x_i)^2] \, n \right\}$$

$a = (\Sigma y_i - b\Sigma x_i) \; n$

Where:

a = y intercept (constant);

b = regression coefficient (slope);

n = number of (paired) observations.

Testing hypothesis about β (Population Slope or Regression Coefficient)

$$S_b = S_e \Big/ \left\{ \Sigma x_{ii}^2 - [(\Sigma x_{ii})^2] \, n \right\}$$

Where:

$$S_e = \left\{ \Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i y_i \right\} (n-2) = \text{standard error of estimate;}$$

S_b = standard error of regression slope (b).

$t_c = (b - \beta) \, S_b$

$H_o: \beta = 0$

$H_a: \beta \neq 0.$

Degree of freedom (v) = n – 2;

Level of significance (α) = 1%, 5% or 10%.

$t_T = t_{\alpha/2, v}$

$|t_c| \geq t_T$ H_0 not acceptable; otherwise H_a is acceptable.

If H_0 is accepted, then the conclusion is that there is no true relationship between X and Y or *b* is not statistically significant at an α-level (e.g. 5%) and when H_0 is not accepted, then the conclusion is that there is true relationship between X and Y or *b* is statistically significant at an α-level (e.g. 5%).

Interval Estimate

Confidence interval for β

$b \pm t_T S_b$

$\Rightarrow b - t_{\alpha/2, v} S_b \leq \beta \leq b + t_{\alpha/2, v} S_b$; $100(1 - \alpha)\%$

To test for:-

$H_o: \beta = 0.5$

$H_a: \beta \neq 0.5.$

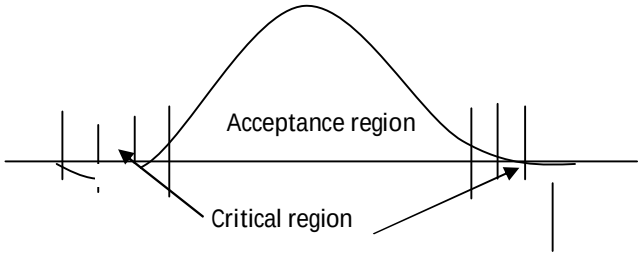

Figure 12: Critical/Acceptance Region

Chi-Square

Frequently, data of interest consist of counts rather than measurement e.g. in the study of consumer preference (CP) for a product in various geographical regions. The CP might consist of "like", "dislike" or "no opinion" and the geographical category might consist of "East", "Central" or "West". The sample gathered might be summarized as counts such as given below.

Table 12: Customers Quality Rating of Product P by Geographical Region

Quality Rating	***East***	***Central***	***West***	***Total***
Like	a	b	c	a+b+c
Dislike	d	e	f	d+e+f
No opinion	g	h	i	g+h+i
Total	***a+d+g***	***b+e+h***	***c+f+i***	***1***

The type of table above is usually called a *Contingency Table* (CT). A CT consists of count data obtained from a simple random sample arranged in rows and columns. An *r* X *c* CT has *r* rows and *c* columns. The actual sample counts are called observed frequency and are usually denoted by f_o.

Chi-square Test for Independence

Against f_o, the expected frequency (denoted by f_e) can be computed as follows:

$f_e = (RT \bullet CT)(GT)^{-1}$

Where:-

f_e = expected frequency;

RT = row total;

CT =column total;

GT = grand total.

The expected (f_e) and observed (f_o) frequencies are used to compute a sample statistic for testing the hypothesis that row and column categories are independent. The statistic that is used for the test is called the sample Chi-square and is computed as follows:

$\chi_c^2 = \sum\{(f_o - f_e)^2\}\{(f_e)^{-1}\}$

Where:-

χ_c^2 = Chi-square computed/calculated;

$\sum$ = sum of;

f_o = observed frequency;

f_e = as defined previously.

The hypothesis is as stated below:

H_o: the row and column are independent

H_a: the row and column are dependent

Decision rule:-

H_o is not acceptable if sample $\chi_c^2 > \chi^2_{\alpha, r}$

Where:

α = level of significance;

r = degree of freedom and r = (r – 1)(c – 1).

At times χ^2 can be used to test the goodness-of-fit i.e. test of assumption made about a population.

Analysis of Variance (ANOVA)

ANOVA is used to test for equality of several population (/sample) means. The ANOVA test is performed on simple random sample drawn independently. The test assumes that the population are normally distributed with the same variance but with mean which may differ. The pertinent data for performing the test are summarized in a table known as ANOVA table. ANOVA could be a one-way classification (in which only one factor is considered to be affecting the variable of interest) or a two-way classification (in which two factors are considered to be affecting the variable of interest).

Steps in ANOVA

1. Setting up a hypothesis i.e. H_0 (null) and H_a (alternative) hypothesis.
2. Deciding on the level of significance (if not given).
3. Computing F-statistic (i.e. F calculated).
4. Finding the tabulated F-statistic on the F table (i.e. F tabulated).
5 Comparing F_T with the F_c.
6. Drawing conclusion i.e. accepting H_0 if $F_c < F_T$ and accepting H_a if otherwise.

One-way ANOVA

Assuming a green house experiment is conducted to determine the yield of potato with the application of four (4) different types of nitrogen fertilizer using three (3) pots per treatment. Suppose the information collected is as shown below.

Table 13: Yield Response of Potato to Nitrogen Fertilizer

Treatment	***Yield of Potato (Kg/N_2 ration)***			
	1	***2***	***3***	***Total***
1	a	b	c	a+b+c
2	d	e	f	d+e+f
3	g	h	i	g+h+i
4	j	k	l	j+k+l
Total	***a+d+g+j***	***b+e+h+k***	***c+f+i+l***	***a+b+...+l***

N.B:- Row total = column total.

Basically, the task is to test the significance of (Nitrogen fertilizer) treatment on yield. The following questions are to be answered:

i. are the differences in yield negligible;
ii. are the differences in yield attributable to chance or fluctuations;
iii. Are the differences in yield large enough to indicate differences in the treatment.
iv.

ANOVA (Mean Additive) Model

$Y_i = \mu + T_i + \varepsilon_i$

where:-

Y_i = yield;

μ = population/sample mean;

T_i = treatment effect;

ε_i = random effect which is due to nature.

Treatment effect can be considered as the true deviation of the mean of the i^{th} group from μ. It is assumed for all the groups involved in the experiment that T_i (treatment effect) is fixed.

Assumptions:

$\sum T_i = 0$;

ε_i (0, 1) i.e. the error is normally and independently distributed with mean zero and constant variance.

Table 14: One-way ANOVA Table

Source of Variation	*Degree of Freedom*	*Sum of Squares*	*Mean Sum of Squares*	*F*
Total	n – 1	$\sum Y_i^2 - CT$	-	-
Treatment	t – 1	$\{(\sum Y_i)^2/r\} - CT$	$SS_t/(t-1)$	MSS_t/MSS_r
Residual	n – t	$SS_t - SS_r$	$SS_r/(n-1)$	-

Where:
CT = correction term = $(\sum Y_i)^2/n$ or $Y_{..}^2/n$ ($Y_{..}$ = grand total)
n = total number of observations;
t = number of treatment;
r = number of observations per treatment;
$\sum Y_i$ = row total;
SS_t = sum of squares total;
SS_r = sum of squares residual;
MSS_t = mean sum of squares treatment;
MSS_r = mean sum of squares residual.
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: $\mu_i \neq \mu_j$; $i \neq j$.

Two-way ANOVA

Basically, the assumptions and steps for one-way also hold for two-way. However, two-way analysis is very useful where variations in observed data exist due to blocking or grouping effect. The two-way classification increase the precision of the conclusion of an experiment since variations excluded in the one-way analysis is included in the two-way.
Given the previous example on potato, assuming that the potatoes are grouped into their varieties, the computation for two-way analysis is as follows:
$Y_i = \mu + T_i + \vartheta_i + \varepsilon_i$
where:-
Y_i, μ, T_i and ε_i are as previously defined;
ϑ_i = blocking effect.

Table 15: Two-way Table

Source of Variation	*Degree of Freedom*	*Sum of Squares*	*Mean Sum of Squares*	*F*
Total	n – 1	$\sum Y_i^2 - CT$	-	-
Treatment	t – 1	$\{(\sum Y_i)^2/t\} - CT$	$SS_t/(t-1)$	MSS_t/MSS_r
Block	b – 1	$\{(\sum Y_j)^2/b\} - CT$	$SS_b/(b-1)$	MSS_b/MSS_r
Residual	(t – 1)(b – 1)	$SS_t - SS_r - SS_b$	$SS_r/(t-1)(b-1)$	-

Where:
CT, n, t, SS_t, SS_r, MSS_t, MSS_r are as defined previously;
$\sum Y_j^2$ = column total;
b = number of blocks;
SS_b = sum of squares block;
MSS_b = mean sum of squares block.
H_0: $T_1 = T_2 = T_3 = T_4$;
H_A: $T_i \neq T_j$; $i \neq j$.
H_o: $\vartheta_1 = \vartheta_2 = \vartheta_3 = \vartheta_4$;
H_a: $\vartheta_i \neq \vartheta_j$; $i \neq j$.