COURSE CODE:	ABG 501
COURSE TITLE:	Methods of Animal Experimentation
NUMBER OF UNITS:	3 Units
COURSE DURATION:	Three hours per week

COURSE DETAILS:

Course Coordinator: Dr. O. Olowofeso, **(*B. Agric. Tech., M. Sc., Ph.D., RAS*)**
Email: olowofesoinunaab@gmail.com
Office Location: Room 6, COLANIM
Other Lecturers: Prof. M.O. Ozoje

COURSE CONTENT:

Variables, scientific data, experimental process, basic terminologies in animal experimentation, basic procedures in animal experimentation, measure of central tendency, testing for goodness of fit, Analysis of Variance, design of experiment, correlation and regression, sampling and sampling methods.

COURSE REQUIREMENTS:

This is a compulsory course for all final year students in the College of Animal Science and Livestock Production (COLANIM). The course must be registered for and passed before students can graduate in any of the five major departments in COLANIM. To do well in the course, students must have the basic knowledge of mathematics and elementary statistics. It is also expected that all students must have a very good scientific calculator and statistical table. Attendance in class is mandatory and students must take part in all tests and final examination. The minimum score required as pass mark at the end of the semester for the course is 40%.

READING LIST:

1. Thomas Glover and Kelvin Mitchell: ***An Introduction to Biostatistics***, McGraw-Hill Companies, Inc. 2001.
2. Mead, R. and Curnow, R. N: ***Statistical Methods in Agriculture and Experimental Biology***, ELBS Edition, 1987.
3. Cochran, W.G. and Cox, G. M:***Experimental designs***, A Wiley International Edition

1.0 VARIABLES AND MEASUREMENTS

Statistics is essential concerned with the application of logic and objectivity in the understanding of events. Necessarily, the events have to be identified in terms of relevant characteristics or occurrences that could be measured in numeric or non-numeric expressions.

The characteristics of occurrence is technically referred to as variable because it may assume different value or forms within a given range of values or forms known as domain of the variable. The act of attaching values or forms to variables is known as measurement.

A variable may be qualitative or quantitative and a qualitative variable may be continuous or discontinuous i.e. discrete. The convention is to identify a variable by the upper case letter e.g. X and an observation of it by a corresponding lower case e.g. x.

1.1 TYPES OF VARIABLES

1.1.1 QUALITATIVE VARIABLE

A qualitative variable may have forms x that could only be described but could not be measured numerically. It is a non-numeric entity. Nutrient when identified only as Nitrogen P, K or Zn is a qualitative variable or factor. Variety of maize is qualitative; sex of a farmer is qualitative. Yield scores when recorded as high, moderate or low is qualitative. Qualitative variable has states or forms, which can be described, categorized, classified or qualified.

1.1.2 QUANTITATIVE VARIABLE

A quantitative variable Y is numeric, it may assume values y that can be quantified. Ages, weight, height, volume, are all quantitative variables. Production, prices, costs, sales, sizes are also quantitative.

A qualitative x may be recorded as if a quantitative variable, for instance in a farm survey, a male farmer may be coded as 0 and a female as 1. In an entomological experiment, response may be coded as 0 if the plant survives a chemical treatment and 1 if otherwise. Resistance of plants after exposure to a plant diseases may be scored as 0,1,2,3,4 and 5 in ascending order of infection and in a taste experiment, the taster may be asked to grade dishes 1,2,3 and 4, respectively. And on the other hand, a quantitative variable Y may be coded qualitatively, for example in an agronomic trial, maize plot response may be coded A for yields exceeding 7400kg per ha, B for yields in the bracket 6001-7500kg, C for yields between 5001 and 6000kg and D for yields 5000kg and below. Forms A, B, C and D are of the descriptive types.

1.1.3 CONTINUOUS AND DISCRETE VARIABLES

A quantitative variable may be continuous or discontinuous i.e. discrete. The family size of a farming household, X in a crop survey may take between values 1,2,3,4,5,6,7 and so on. In the interval (0,100) the number if values which X can take is countable (finite), it is 101, but the size of farms Y cultivated by the household may take an uncountable (infinite) number of values within even the shortest of intervals when measurement is not limited by degree of accuracy nor by the precision of measuring equipment. Family size variable X is a typical discrete variable while the farm size Y is a typical continuous one. Let Y_1 and Y_2 be 2 distinct values of a continuous Y the variable may still take an uncountable number of values between Y_1 and Y_2 however close the two variables might have been to each other. On the other hand, a discrete variable X may not take more than a known finite number of values between two distinct values X_1 and X_2, however distant the two values might have been from each other. In a continuous situation, there is always another value between two values in the discrete environment, there may not be any other value between two values.

Implication of variable status (and measurement) on data management: The status of a variable has several implications on data management. It has implication for methods and other aspects of:

i. Measurement of variable
ii. Data collection
iii. Data organization
iv. Data summarization
v. Data presentation
vi. Data analysis
vii. Data interpretation
viii. Data storage and retrieval
ix. Data use

1.2 SCIENTIFIC DATA

1.2.1 Definition of scientific data

Statistics is about the theory, method and practice of data collection and analysis. It is about procedures and formulas, and theoretical and practical method of data collection, organization, summarization, presentation, analysis, interpretation, evaluation storage and retrieval. A statistical data is acknowledged as any piece of information, quantitative or otherwise, that has been generated through application of Logic and objectivity. Methods of animal experimentation are concerned with description, analysis, prescription, and prediction. Description is scientific reduction of a mass of data into few numbers that represent salient characteristics of the data. It is a form of mathematical analysis.

1.2.2 Categories of Scientific data

It is convenient and instructive to categories data as follows:

- Routine records
- Survey data
- Experimental data
- Cross-sectional data
- Repeated measures (Time series data)

1.2.3 Routine Records

These are general data that are periodically recorded essentially for administration use of the establishment. Some statistical analysis of the data may be obtained to describe and analyse the activities and operation and of the establishment and to initiate implement and monitor future plans, projects and programmes. Planning research and statistics department in ministries, parastatals, institutes and companies exist for such purpose. Meteorological data are recorded and kept in weather stations; yield records are obtained and kept routinely in agricultural plantation and farms.

These categories of data are cheap. But it is often collected without any definite purpose objectives or focus. The planning may be quiet haphazard, the design or protocol collection may not promote objectivity while the instrument of data collection may not be comprehensive, precise and focused enough to permit credible statistical analysis. The research design is often inadequate or inappropriate.

1.2.4 SURVEY

Surveys are investigations in which the surveyor obtains information about characteristics, opinions, attitudes, tendencies, activities or operation of the individual unit of the population. A population is defined as the totality of the units being surveyed or investigated. For example, a population of households in a rural economic survey, a population of farms in a farm survey, a population of persons living in a country, a population of plots in farm survey, a population of trees in a plantation reserve. There are two types of survey – namely census and sample survey. A census is a complete enumeration in which every unit in the population is observed, they usually consume large amount of time, human and material resources. A sample survey involves observation of only a selected sample of the eligible population units. Sample surveys are cheaper but the sample must be relevant, objective and representative to have any meaning.

A sample may be:

i. Statistical or non statistical
ii. Regular or adhoc
iii. Probability or non-probability
iv. Perspective or retrospective
v. Confidential or non confidential
vi. Cross-sectional or time series

1.2.5 Statistical and non-statistical samples

A statistical sample otherwise known as a probability or objective sample is such that sample selections must have been on a random process involving the use of instruments of chance. There are many types of statistic sample. These are distinguished by the extent of randomness, the degree of homogeneity and for agglomeration of the population units, the number of samples, stages, phases or occasions, the mode of selection etc. examples of statistics samples include:

i. simple random sample
ii. stratified random sample
iii. systematic random sample
iv. cluster sample
v. multi-stage sample

vi. multi-phase sample
vii. multi-occasion sample
viii. interpenetrating sub-sample

Non statistical sample are subjective, sample collections are non-random. A most immediate example is quota sample, which is a common practice in opinion surveys. A probability sample may be subjected to inferential statistical analysis. But non-statistical samples would not permit application of probability concepts which elements form the basis of inferential statistics. Thus such samples can only be described: no generalization for the target population is valid.

1.2.6 Regular and Adhoc surveys

Some establishments, particularly national statistics offices, carry out surveys on regular basis. Population censuses should be decennid events, agricultural sample/censuses should also be decennid. National demographic and health surveys should be periodical. Adhoc surveys are not regular, they are done when occasion demand. Indeed most research sample surveys including crop surveys, farm surveys and epidemiological surveys are adhoc.

1.2.7 Prospective and Retrospective Survey Data

In a prospective survey, observation are made on specified characteristic of the sample units on a regular basis from a particular time at which the sample is determined until a specified future point in time. The survey is retrospective when data in respect of the specified set of characteristics are obtained on the unit from the current point back into the past point. These two classes of survey data are common in clinical trials involving patients whose histories are obtained into the past in retrospective survey or into the future in a prospective survey. Such surveys investigate the progression and dynamics of medical conditions under specific treatment regimes. Further instances of prospective or retrospective studies are long term soil fertility experiments investigating the dynamics of soil fertility over a period of persistent soil use and under various nutrient regimes and agronomic practices.

1.3 Experimental Data

There are two types of experiments namely comparative experiments and absolute experiments respectively.

Absolute experiments- Absolute experimental data come from absolute experiment where there are no differential treatments, no comparisons are intended and no inferential analysis could be made.

Comparative experiment-These are planned investigations into the effects of treatments selected and applied to experimental units or plots according to specific notes or procedures, the plots (or subject or units) may be subjected to regimes of variations that are imposed by inherent chance factors.

On one hand and some non-chance disturbance factors like blocks, rows and columns on the other. Designed comparative except ensure allocation of plots to treatments in away that facilitates isolation and assessment of genuine treatment effects possible. It allows the experimental choice of treatment for the experiment and the choice of allocation pattern of treatment to plots. These twin aspects of experimental design are known as treatment design and experimental design respectively.

1.3.1 Cross-sectional and time series of repeated measures data

This terminology identifies the time frame of the data. A survey or experimental data could be time series or cross-sectional data. A cross sectional or latitudinal data has a single time scope or frame. The data consists of observations made on sampling or experimental units over a single point in time, which may be a growing season, a month, a quarter, a year.

i. A single experiment gives a cross-sectional data.
ii. A census data is cross- sectional.
iii. A farm survey in which the figures are observations made at a point in time is cross-sectional.
iv. A series of experiment or surveys in which observations are not repeated.
Time series or repeated measures. The time series or longitudinal data are repeated measures on same experimental or survey units.

- Sets of observations made on each of the experimental units of an experiment at P different times constitute a time series or repeated measures.
- A long –term experiment in which yields are taken at successive periods on the same experimental units gives a kind or time sexes data.
- Retrospective or prospective surveys in which observations in respect of specified medical conditions of a patient are made over time are time series data.
- Economic data that are observations on same units, economic units, like household political entities etc. constitutes an economic series that are time series.

1.3.2 Data collection through Experiment.

Here we are concerned with comparative experiment –explain.

* Persistence in forage to cutting frequency
* Response to different levels of protein
* Livestock systems (cage and floor)
* Breed /strain differences.

For example, we need data to investigate response of maize to varietal and nutrient regime and weeding methods in a field of plots. The relevant questions include

i. How many varieties? How many and what levels of each of the major nutrients- N P and K should be invalid? What are the weeding frequencies?
ii. How do we constitute treatment combinations?
iii. If there exists a trend in fertility distribution that makes the plots heterogeneous how do we arrange the plots and allocate the treatment combinations to isolate and eliminate effects of plot heterogeneity and facilitates isolation and inferences on treatment effects.
iv. How we allocate treatments to plots to entrench complete objectivity and permit statistical analysis.
v. What statistical model do we adopt and what hypothesis do we construct to facilitate relevant comparisons and tests?

Answers to these questions constitute the essential elements of experimental design. All biological, materials vary; even identical twins treated the same way will have different weaning weights. The variation is caused by many factors and the role of statistics is to:

i. Measure the amount of random variation due to inherent variability of the materials
ii. Measure non-random, systematic variation ascribable to treatment effects.
iii. Measure the latter in relation to the former to assess the genuineness of treatment effects. One of the procedures for this kind of investigation is experimentation, that is comparative experimentation, which can be achieved through a process known as experimental design and which may involve feeding trails, cultural practices, pest control measures, perennial crops, livestock and pasture, fertilizers, plant population/ spacing, crop rotations, intercrossing etc.

1.3.3 The experimental Process

There are three parts to an experimental process namely.

* Design
* Experiment and
* Analysis

Experiment involves the specification of the objective and formulation of intended hypotheses. Design has two aspects, namely experiment design and treatment design. Experiment designs has to do with

i. Treatment structure
ii. Block structure and
iii. Integration of treatment and block structures with treatment structure and it involves:

a. Determination of plot size, shape and orientation
b. Determination of block size, shape and orientation
c. Allocation of treatments to plots and blocks including order and method or randomization.
d. Specification of the statistical model of analysis.

Plot and block structures are constructed to ensure minimum plot variability and maximum precision and efficiency units so dictate. Blocking is done so that units within a block are as similar as possible e.g. animals of similar weight, age, previous milk yield, or units of the same “fertilizer status” etc. Blocking or stratification is used routinely in agronomy trials but this should not be so.

Effective blocking increases precision. Treatment comparisons are made within blocks; hence variation between blocks is eliminated. Conclusions are valid over the range of situations represented by different blocks.

Treatment Design has to do with.

i. Selection of factors and levels of factors
ii. Construction of treatments based on appropriate combinations of factor levels.
iii. Selection of treatments for the experiment and
iv. Determination of number of replicates per treatments which is
    a. a function of treatment structure whether factorial or otherwise.
    b. An issue to be jointly solved by the statistician and substantive field experts.

1.3.4 Basic terminologies

a. Plots are the smallest units of the experimental field plots in field experiments, pots in green house experiments, pens or individual animals in animal experiments, farms or farmers in on farm surveys/trials, patients in medical trials, farms in diseases survey/trails etc.

b. Yields are response of experimental units to treatments. E.g. reproductive yields, vegetative yields, disease scores, moisture content, number of stands, survival scores etc.

c. Treatments are applications that can stimulate response. Varieties, lines, nutrients, procedures, methods and any stimulus that can generate reaction quality as treatments. These are test and control treatments. The later is the reference and which the former are assessed. The performance of a test treatment is the difference between its average yield and that of the control.

d. Replicate are simply repetitions of treatments over plots. A whole experiment may be replicated over time and space.

e. Blocks are divisions of the experimental materials. They are supposed to be groups of the experimental units or plots such that plots within same blocks are as homogenous as possible in all materials aspects, while plots in different blocks are to be as different as possible.

f. Experimental error is a measure of plot variability. It is an expression for all variation that can be attributed to the effects of all non-treatment factors and other unidentified disturbance factors. Experimental error variance is conventionally denoted as σ^2 and the measure of variability is the standard deviation given as $\sigma = \sigma^2$

1.3.5 BASIC PROCEDURES

1. Replication means that instead of having a single plot in each treatment, we have several smaller ones known as replicates. Replicates are desirable because, it:
   i. Enlarges scope of investigation since replicates stimulates varying environmental conditions.
   ii. Enhances precision and overall efficiency by effecting reduction in error by a factor of the square root of the number of replicates.
   iii. Permits determination of experimental error and therefore enables probability statements about estimates of effects.
   iv. Minimizes experimental error because it reduces plot size to a precision enhancing form.

2. Randomization is the act of allocating plots to treatment purely on the basis of chance. It is an objective strategy that guarantees that:
   i. Every treatment has equal chance (probability) of being allocated to any given plot.
   ii. Treatment allocations are devoid of bias, subjectivity and personal manipulations.
   iii. Statistical estimation and tests of hypothesis on effects as theoretically valid.
   iv. Blocking is partition of heterogeneous plots into homogeneous groups known as blocks to facilitate isolation of block variation that could distort treatment effects.

3. Blocking is used when variation among experimental units so dictates. It is done so that units within a block are as similar as possible. In livestock experiment, blocks may be animals of similar weight, age, previous milk yield, lactation etc.
4. Blocking factors are disturbance factors whose effects are a nuisance that must be neutralized through appropriate experiment design. It acts as:
   i. An error control strategy that when used effectively, reduces error variance and increases precision and reliability of estimates of effects.
   ii. An indirect replication of treatments or experimental design with direct positive implication for validity and efficiency of experiments by widening the scope/of inferences and increases precision and overall efficiency respectively.

1.3.6 Measure of Central Tendency or Measure of Location

It is possible track of the overall picture of data by looking at all the picture at once. One type of measure useful for summarizing data defines the centre or middle of the sample. This type of measure is a measure of central location or tendency.

a. The arithmetic mean denoted by $\bar{x}$ is the sum of all the observations divided by the number of observation

   $\sum_{i=1}^{n}$ is simply a short way of writing $x_1 + x_2 + \ldots\ldots x_n$.

b. The median - the middle values when you arrange from the lowest to the highest. When you have 2 values as the ones in the middle, the 2 values must be averaged to get the median.
c. The mode - the most frequently occurring values among all the observation is a sample. Some distributions have more than one mode. A distribution with one mode is referred to as unimodal, 2 is bimodal, 3 is trimodal etc.

1.3.7 Measure of Dispersion or Measure of Variation

A measure of dispersion or a measure of variability is an indication of the scatter of measurement around the centre of the distribution or the opposite of how clustered the measurements are around the centre.

1. The Range: The difference between the highest and lowest measurements in a group of data is termed the range. If the measurements are arranged in increasing order of magnitude as if the median were about to be determined, then:
   Sample range = $X_n - X_1$

The range is a relatively crude measure of dispersion measured as it does not take into account any measurements except the highest and the lowest.

Hypothetical Example 1

x_i (g)	$x_i - \bar{x}$ (g)	$\lvert x_i - \bar{x} \rvert$ (g)	$(x_i - \bar{x})^2 (g)^2$
1.2	-0.6	0.6	0.36
1.4	-0.4	0.4	0.16
1.6	-0.2	0.2	0.04
1.8	0.0	0.0	0.00
2.0	0.2	0.2	0.04
2.2	0.4	0.4	0.16
2.4	0.6	0.6	0.36

$\Sigma x_i = 12.6g$ $\quad \sum (x - \bar{x}) = 0.0g$ $\quad \sum |x - \bar{x}| = 2.4g$ $\quad \Sigma (x_i - \bar{x})^2 = 1.12g^2$

= sum of squares

$\bar{x} = \frac{12.6}{7} = 1.8g$

Range = $x_7 - x_1$ = 2.4g – 1.2g = 1.2g

Mean deviation $\sum \frac{|x_i - \bar{x}|}{n} = \frac{2.4}{7} = 0.34g$

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} = \frac{1.12g^2}{6} = 0.1867g^2$$

S = $\sqrt{0.1867g^2} = 0.43g$

Sample 2.

xi (g)	*xi - $\bar{x}$ (g)*	*$\lvert x_1 - \bar{x} \rvert$ (g)*	*$(xi - \bar{x})^2$ (g^2)*
1.2	-0.6	0.6	0.36
1.6	-0.2	0.2	0.04
1.7	-0.1	0.1	0.01
1.8	0.0	0.0	0.00
1.9	0.1	0.1	0.01
2.0	0.2	0.2	0.14
2.4	0.6	0.6	0.36

Σx_i = 12.6g $\quad \Sigma(x_i - \bar{x}) = 0.0(g)$ $\Sigma \lvert x_i - \bar{x} \rvert$ (g) =1.8g $\Sigma (x_i - \bar{x})^2 = 0.82g^2$

= sum of squares

$x = \frac{12.6}{7} = 1.8g$

Range = $x_7 - x_1$ =2.4g – 1.2g = 1.2g

Mean deviation $\sum \frac{|x_1 - \bar{x}|}{n} = \frac{1.8g}{7} = 0.26g$

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} = \frac{0.82g^2}{6} = 0.1367g^2$$

S = $\sqrt{0.1367g^2} = 0.37g$

2. The mean deviation – The sum of all deviations from the means i.e. $\Sigma (x_i - \bar{x})$ will always equal zero, however, such summation will be useless as a measure of

dispersion. Summing the absolute values of the deviation from the means results in a quantity that is an expression of dispersion about the mean. Dividing this quantity by n yields, a measure known as the means deviation or mean absolute deviation of the sample. It has the same unit as that of the data.

Sample mean deviation $\Sigma\frac{(x_i - \bar{x})}{n}$ …….(1)

3. The variance – The sum of the squares of the deviations from the mean is called the sum of squares, abbreviated as SS and defined as.

Population SS = $\Sigma(x_i - \mu)^2$ ……. (2)

Samples SS = $\Sigma(x_i - \bar{x})^2$ …….. (3)

The mean sum of squares is called the variance and for a population is denoted as σ^2 (sigma squared).

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \text{ …….(4)}$$

The best estimate of the population variance is the sample variance S^2

$$S^2 = \sum \frac{(X_i - \bar{X})^2}{n-1} \text{.....(5)}$$

If in equation 4, we replace μ by x and N by n, the result is a quantity that is a biased estimate of σ^2. The dividing of the sum of squares by n-1 (called the degree of freedom abbreviated Df) rather than n, yields an unbiased estimate. It is equation 5 that should be used to calculate the sample variances. The calculation of S^2 can be tedious for large samples but it can be facilitated by the use of equality.

Sample SS $= \sum x_1^2 \frac{(\Sigma x_1)^2}{n}$

Sample 1

Xi (g)	$X^2i\ (g^2)$	*xi (g)*	$X^2i\ (g^2)$
1.2	1.44	1.2	1.44
1.6	2.56	1.6	2.56
1.7	2.89	1.7	2.89
1.8	3.24	1.8	3.24
1.9	3.61	1.9	3.61
2.0	4.00	2.0	4.00
2.4	5.76	2.4	5.76
$\Sigma xi=12.6g$	$\Sigma x^2i=23.50g^2$	$\Sigma xi=12.6g$	$\Sigma x^2i=23.50g^2$
$n=7$		$n=7$	

$\bar{x} = \frac{12.6}{7} = 1.8g$ $\qquad$ $\bar{x} = \frac{12.6}{7} = 1.8g$

CSS $= \sum x_1^2 \frac{(\Sigma x_1)^2}{n}$ $\qquad$ CSS $= 23.50g^2 - \frac{(12.6g)^2}{7}$

$= \quad 23.50g^2 - \frac{(12.6g)}{7} \quad = 0.82g^2$ (CSS is called corrected sum of squares)

$= 23.50g^2 - 22.68g^2$ $\qquad S^2 = \frac{0.82g^2}{6} = 0.1367g^2$

$= 0.82g^2$ $\qquad S = \sqrt{0.1367g^2} = 0.37g$

$S^2 = \frac{SS}{n-1}$ $\qquad V = \frac{0.37g}{1.8g}$

$= \frac{12.6g}{6} = 0.1867g^2$ $\qquad = 0.21 = 21\%$

$S = \sqrt{0.1867g^2} = 0.43g$

$$V = \frac{S}{\overline{X}} = \frac{0.43g}{1.8g} = 0.24 = 24\%$$

The variance has square units. If measurements are in grams, their variances will be in grams squared or if measurements are in cubic centimeters, their variance will be in terms of cubic centimeters squared even though such squared units have no physical interpretation.

4. The standard deviation is the positive square root of the variance therefore, it has the same units as the original measurements. Thus for a population.

$$S = \sqrt{\frac{\sum X_i^2 - \frac{(\sum x_i)^2}{N}}{N}}$$

And for a sample

$$S = \frac{\sum X_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

4. The coefficient of variation or variability is defined as

$V = \frac{S}{\overline{X}} or V = \frac{S}{\overline{X}} X100\% \sin ce \frac{S}{\overline{X}}$ is generally a small quantity, it is

frequently multiplied by 100% in order to express V as a percentage.
The standard error
S.E (X) = $\frac{S}{\sqrt{n}}$ or $\frac{SD}{\sqrt{n}}$
where n is number of observation(s)

$\sigma = S = SD =$ Standard deviation

2.0 TESTING FOR GOODNESS OF FIT

2.1 CHI SQUARE GOODNESS OF FIT

Example 1

A plant geneticist may raise 100 progeny from a cross that is hypothesized to result in a 3: 1 phenotypic ratio of 84 yellow: 16 green is observed, although out of this total of 100 plants, the geneticists hypothesis would predict a ratio of 75 yellow: 25 green. The question to be asked, then is whether the observed frequencies (84 and 16) deviated significantly from the frequencies expected if the hypothesis were true (75 and 25). The statistical procedure for attacking the question first involves the concise statement of the hypothesis to be tested. The hypothesis in this case is that the population which was sampled has 3:1 ratio of yellow-flowered to green-flowered plants. Statistically, this is referred to as null hypothesis (abbreviated Ho) because it is a statement of "no difference" in this instance, we are hypothesizing that the population flower colour ratio is not different from 3:1. If it is concluded that Ho is false, then an alternative hypothesis (abbreviated H_A) will be assumed to be true. In this case, H_A would be that the population sampled has a flower colour ratio which is not 3 yellow: 1 green. One states a null hypothesis and an alternative hypothesis for every statistical test performed and all possible outcomes are accounted for by the two hypothesis.

$$X^2 = \sum_{i-1}^{k} \frac{(fi - fi)^2}{f}$$

Here, *fi* is the frequency or number of counts, observed in class *i*
f is the frequency expected in class I if the null hypothesis is true.
The summation is performed over all K categories of data.
Calculation of Chi square of fit data consisting of 100 flower colours to a hypothesized colour ratio of 3:1.

H_o: The sample data came from a population having 3:1 ratio of yellow to green flowers.

H_A: The sample data came from a population not having a 3:1 ratio of yellow to green flowers.

Category (flower colour)

	Yellow	Green	n
fi	84	16	100
(*f*)	(75)	(25)	

Degree of freedom = v = k – 1 = 2 – 1 = 1

$$X^2 = \sum_{i-1}^{k} \frac{(fi - fi)^2}{f}$$

$$= \frac{(84-75)^2}{75} + \frac{(16-25)^2}{25}$$

$$\frac{92}{75} + \frac{92}{25}$$

= 1.080 + 3.240
= 4.320
0.025 < P < 0.05

Therefore, reject Ho

Example 2

In series of matings involving black-polled cattle, a researcher found 54 black-pooled, 18 red-pooled, 25 black horned and 8 red-horned offspring. Two genes are known to be involved and a ratio of 9:3:3:1 was to be expected. Were these results in agreement with this ratio?

Observed (O)	**Expected (E)**	**O – E**
54	59.0625	-5.0625
18	19.6875	-1.6875
25	19.6875	5.3125
8	6.5625	1.4375

$$X^2 = \frac{(-5.0625)^2}{59.0625} + \frac{(-1.6875)^2}{19.6875} + \frac{(5.3125)^2}{19.6875} + \frac{(1.4375)^2}{6.5625}$$

= 2.327.

Since calculated X^2 is less than $X^2_{.05.}$ (4-1) = 7.815, we do not reject the hypothesis i.e. the observed results are in agreement with the expected 9:3:3:1 ratio.

2.2 CONTINGENCY TABLES

Experimental subjects may be classified according to several attributes and enumeration obtained on each cell. Such data are arranged in contingency tables and the researcher is interested in testing for independence of the classification variables.
e.g.
Consider a population of 150 animals challenges with a certain disease. The following data represent numbers of males and females resistant and susceptible to the disease. The hypothesis to test is that resistance or susceptibility is independent of sex.

	Resistamt	**Susceptible**	**Total**
Male	35	28	63
Female	50	37	87
Total	85	65	150

Above is referred to as a 2 x 2 contingency table.
The expected frequency in a contingency table are obtained by the formula

$$\boldsymbol{F}_{ij} = \frac{(R_i C_j)}{n}$$

Where R_i is the total frequency in row *i*, C_j is the total frequency in column *j* and *n* is the total sample size. The number of degrees of freedom in a contingency table is given as

d.f.= (r-1) (c-1)

Where r= number rows and C = number of columns in the table,
For example, expected cell frequencies are as follows

	Resistant	Susceptible	Total
Male	35.7	27.3	63
Female	49.3	37.7	87
Total	85.0	65.0	150

The formula for the Chi-square test statistic for contingency tables is

$$x^2 = \Sigma i \Sigma j \frac{\text{(Observed frequency – Expected frequency)}}{\text{Expected frequency}}$$

Here $x^2 = \frac{(35 - 35.7)^2}{35.7} + \frac{(28 - 27.3)^2}{27.3} + \frac{(50 - 49.3)^2}{49.3} + \frac{(37 - 37.7)^2}{37.7}$

$= 0.014 + 0.018 + 0.010 + 0.013$

$= 0.055$

Since calculated $x^2 = 0.055$ is smaller than $x^2_{(0.05,1)} = 3.841$

The test is not significant and we conclude that the frequencies of resistant and susceptible individuals are the same for both sexes. In order words, resistance and susceptibility are independent of sex.

2.3 STATISTICAL ERRORS IN HYPOTHESIS TESTING

One needs an objective criterion for rejecting or not rejecting the null hypothesis for a particular statistical test. Theoretically, a very large x^2 values might be calculated even when the hypothesis is true although the larger the Chi-square, the smaller is the probability that Ho is true. But how small a probability (i.e how large x^2) shall we require to reject the null hypothesis? As explained earlier, a probability of 5% or less is commonly used as the criterion for rejection of H_o. The probability used as the criterion for rejection is turned termed the significance level denoted by α, and the value of the test statistic corresponding to this probability (e.g. $x^2 = 3.841$ for the 5% significance level) is the critical value of the statistic.

It is very important to realize that a true null hypothesis occasionally will be rejected which of course means that we have committed an error. Moreover, this error will be committed with a frequency of α. That is, if H_o is in fact a true statement about a statistical population, it will be concluded (erroneously) to be false 5% of the time. The rejection of a null hypothesis when it is in fact true is a Type 1 error (also called an error, or an α error of the first kind). On the other hand, if H_o is in fact false, our test may occasionally not detect this fact, and we shall have reached an erroneous conclusion by not rejecting H_o. This error of not rejecting the null hypothesis when it is in fact false is a Type II error (also called a β error, or an error of the second kind).

The two types of errors in hypothesis testing

	If H_o is true	If H_o is false
If H_o is rejected	Type I error	No error
If H_o is not rejected	No error	Type II error

2.4 COMPARING MEANS

To test one-way paired sample hypothesis –concerning treatments or sample, means, the students t-statistic due to William Gosset is used.
e.g.
consider the following data on weaning weight (kg) of creep fed and non creep fed calves.

Creep fed	**Non creep fed**
84.5	83.1
86.9	50.0
75.0	61.4
88.7	74.7
85.4	62.5
88.9	62.5
78.5	46.9
$\bar{x}_1 = 83.9857$	$\bar{x}_2 = 63.0143$
$S_1 = 5.291$	$S_2 = 12.729$

HYPOTHESIS 1: HYPOTHESIS INVOLVING THE MEAN (μ)
Suppose the researcher wants to determine whether the average weight of his creep fed calves compares favourably with literature reports of 87kg. This is a one sample hypothesis .
Ho: μ = C i.e 83.9857 = 87.0
Ho: $\mu \neq C$ i.e 83.9857 = 87.0

The test statistic is

$$t = \frac{\bar{x} - c}{Sx}$$

Which is compared with t a from a table of the t- distribution

e.g. /t/ = $\dfrac{83.9857 - 87.0}{\dfrac{5.291}{\sqrt{7}}}$ $\quad \dfrac{SD}{\sqrt{n}} = -1.507$

since this t value is smaller than $t_{(0.05,6)} = 2.447$ **we do not reject Ho** and conclude that there is no difference statistically between the two means being tested.
i.e $t_c < t_t$ or -1.507<2.447

HYPOTHESIS II: Test of hypothesis involving two samples
If the researcher wants to test the mean weaning weight of creep-fed calves is significantly different from that on non-creep fed individuals, he performs a two sample or two treatment t test.
Ho: $\mu_1 = \mu_2$
Ho: $\mu_1 \neq \mu_2$

The test statistic is t- $\dfrac{x_1 - x_2}{\dfrac{SD_1}{\sqrt{n_1}} \; \dfrac{SD_2}{\sqrt{n_2}}}$

Where $\frac{SD_1}{\sqrt{n_1}} - \frac{SD_2}{\sqrt{n_2}}$ = the standard difference between means or S $x_1 - x_2$

$$= \sqrt{\frac{S^2p}{n_1-1} + \frac{S^2p}{n_2-1}}$$

Where n_1 and n_2 = the total number of observations used to compute the first and second means, respectively and
S^2p = Pooled sample variance

$$= SS_1 + SS_2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

For our example

$$S^2p = \frac{6(27.99468) + 6(162.02744)}{7 + 7 - 2}$$

$$= 95.01106$$

Hence S $x_1 - x_2 = \sqrt{\frac{95.01106}{6} \times 2} = 5.6276$

Thus calculated $t = \frac{83.9857 - 63.0143}{5.6276} = 3.7265$

Since this t is greater than ta, $n_1 + n_2 - 2$ or $t_{(0.05,12)} = 2.179$. we reject H_o and conclude that the two means are significantly different. In other words, creep fed calves had significantly higher weaning weights than non-creep fed calves.

3.0 CORRELATIONS

When pairs of observation are available from a bivariable normal distribution, the researcher may be interested in determining the strength of the relationship between the two variables concerned. The quantity that describes this is the correlation coefficient. Its sign indicates the direction of the relationship. Such relationship as between body weight and egg products, body weight gain and feed consumption, milk yield and butterfat percent, body weight and egg weight, body weight and height at withers. Shell thickness and % calcium in plasma, sperm concentration and fertility and so on are of considerable interest to the Animal Scientist.

The sample correlation coefficient is computes by:

$$r = \frac{\Sigma xy - (\Sigma x)(\Sigma y)/n}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}\ \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}}$$

Compute the sample correlation coefficient from the following data on body weight (BW) and shank length (SL) of 20 six week old broilers.

BW (kg)	SL (cm)	BW(kg)	SL(cm)
0.63	4.8	0.90	5.2
0.80	4.0	0.72	4.7
0.78	5.1	0.66	4.7
0.83	5.0	0.66	5.0
0.74	4.7	0.66	4.7
0.77	4.7	0.60	4.6
0.74	4.8	0.9	5.1
0.85	5.2	0.83	5.0
0.66	4.5	0.77	5.0
0.89	5.2	0.84	5.2

Let BW = X and SL =Y

Compute the following
$\Sigma X = 15.23$
$\Sigma Y = 97.20$
$\Sigma X^2 = 11.7651$
$\Sigma Y^2 = 474.12$
$\Sigma XY = 74.30$
Then

$$r = \frac{20x74.3 - (15.23)(97.20)}{\sqrt{[(20x11.7651) - 15.23^2][(20x474.12) - (97.20)^2}}$$

$$= \frac{1486 - 1480.36}{\sqrt{3.3491X\,34.56}} = 0.52$$

The standard error of the estimate is

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

$$= \sqrt{\frac{1 - (0.52)^2}{18}} = 0.041$$

To test whether the correlation coefficient was significant or not, we compute the t statistic

$$t = \frac{r}{S_r} = \frac{0.52}{0.041} = 12.68$$

And compare with t α, n-2, here $t_{(0.05,18)} = 2.101$. We conclude that the correlation was highly significant $P < 0.01$.

4.0 ANALYSIS OF VARIANCE

Analysis of variance is essentially a process of partitioning the variations in observed data into its various components. It is normally presented in the form of a summary table known as the ANOVA table.

An analysis of variance table can be for one-way or two classification depending on the number of criteria used for classifying the data. In one-way analysis, we consider only one criterion or factor as affecting yield, for example. In a two-way analysis we consider two factors as affecting yield.

4.1. One way classification

Consider the yield equation:

$Y_{ij} = \mu + t_i + \Sigma_{ij}$

Where μ = population mean- the inherent effect

t_i = the treatment effect

Σ_{ij} = the error effect which is due to nature, randomness or any other factors beyond human control.

In the above equation, the only factor considered to affect yield is the treatment applied. The equation is based on the following assumptions.

1. The treatment effect, t is fixed.
2. The total effect of the treatment is equal to zero i.e. $\Sigma ij = 0$.
3. The expected value of error is equal to zero i.e. $\Sigma_E = 0$
4. The error is normally and independently distributed with mean 0 and variance σ^2. i.e: LIN (0, σ^2).

In other to test the significance of the treatment effect on yield, we use the F-test. The following is a summary of the ANOVA.

ANOVA Table – One way classification

Source	Degree of freedom	Sum of squares	Mean square	F-ratio
Total adjusted for the mean	n-1	$\sum_{ij}^{n} Y_{ij} - CT$	-	-
Treatment	t-1	$\sum_{i} \frac{T_i^2}{r} - CT$	$\frac{\text{SS Treatment}}{\text{DF treatment}}$	$\frac{\text{MS treatment}}{\text{MS error}}$
Residual	n-t	SS Total – SS Treatment	$\frac{\text{SS residual}}{\text{DF residual}}$	-

Where

CT = Correction term

$$\frac{(\Sigma Y_{ij})^2}{N}$$

N= total number of observations

t= number of treatments

r= total number of observations per treatment.

$\Sigma Yi = \sum_{i} Y_{ij}$ Therefore, the sum of all observations per treatment.

ΣY_{ij}^2 = sum of the squares of all the observations per treatment

DATA TABLE

Treatment	1	2	3	4	5	Total Yi
1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_1
2	Y_{21}					Y_2
3		Y_3				Y_3
4					Y_{45}	Y_4
5			Y_{53}			Y_5
Total Yj			Y_4			Y

Y_1 = Sum of observation on the first treatment

Y_2 = Sum of observation on the second treatment

Y_3 = Sum of observation on the third treatment
Y_4 = Sum of observation on the fourth treatment
Y_5 = Sum of observation on the fifth treatment

Example of one-way classification,

	Replication (Block)				
		1	2	3	Total Yi
Treatment	1	7	2	4	13
	2	6	4	6	16
	3	8	4	5	17
	4	7	4	2	13
	Toyal Y_j	28	14	17	59

Summary of the data

1. $\Sigma Y_{ij}^2 = 7^2 + 2^2 + 2^2 + 4^2 + 6^2 + 4^2 + 6^2 + 8^2 + 4^2 + 5^2 + 7^2 + 4^2 + 2^2 = 331$
2. C.T. $= \frac{(\Sigma Y_{ij})^2}{N}$

$= \frac{59^2}{12}$

$= 290.0833$

3. $\frac{\Sigma Y_{ij}^2}{R} = \frac{13^2}{3} + \frac{16^2}{3} + \frac{17^2}{3} + \frac{13^2}{3}$

$= 56.3 + 85.3 + 96.3 + 56.3$

$= 294.2$

4. $\frac{\Sigma Y_{ij}^2}{t} = \frac{28^2}{4} + \frac{14^2}{4} + \frac{17^2}{4}$

ANOVA TABLE

Source of variation	d.f	s.s	m.s	F-ratio
Total a.f.m	11	331-290=41	3.7	-
Treatment	3	4.2	1.4	14<1 = 46 0.304
Residual	8(n-t)	(41-4.2)36.8	4.6	

4.2 Statistical inference

The purpose of the analysis is to be able to draw inferences on the significance of the effect of the factor we are considering. In doing this we follow the following procedure.

i. Setting up of a hypothesis: Firstly, we set up a null hypothesis that the treatment effect is not significant i.e.

H_o:ti = 0

We also set up an alternative hypothesis that the treatment effect is significant thus:

H_A:ti = 0

ii. We decide on the level of significance (not given) and find the tabulated F-ratio from the Table.
iii. We compare the tabulated F- ration with the computed F – ratio.
iv. Conclusion. We accept the null hypothesis if the F-ratio calculated is less than the F-ratio tabulated. Otherwise, we accept the alternative hypothesis. In the above example for the one-way classification, we are testing for the treatment effect and our null hypothesis will be:
H_o: $t_1 = t_2 = t_3 = t_4 = 0$
i.e. there is no significant difference between the treatments or that the treatment has no effect.
The alternative hypothesis is
H_A:μ = μ = i

At α (level of significance) = 5%
$F_{0.05}$ (3,8) = 4.07
F computed = $\frac{14}{46}$
Therefore, $F_C < F_T$, we accept the hypothesis that there is no treatment effect.

4.3 Two –way classification

The two-way classification of variance is basically the same with the one-way classification except for the fact that while only one factor is considered to affect yield in the one way, two factors are considered to affect yield in the two way analysis. It is very useful where variations in observed data exist due to blocking or grouping of the experimental units. The two-way classification excluded such variation from those due to the applied treatments and thereby increases the precisions of the conclusions on the experiment.

The yield equation is
Yield = Population mean + treatment effect + Blocking effect + error term

$$Y_{ij} = \mu + T_i + P_j + \varepsilon_{ij}$$

Where P_j = the blocking effect. The basic assumptions underlying the equation are the same with the one-way classification.
The working formulas for the degree of freedom, sum of squares, mean squares, and F-ratio are given in the table below:

Correction term C.T. = $\frac{(\Sigma Y_{ij}^2)}{N}$

Source of variation	Degree of freedom	Sum of squares	Mean square	F-ratio
Total adjusted for the mean	n - 1	$\sum_{ij}^{n} Y_{ij}^2 - CT$	-	-
Treatment	t-1	$\sum_{i} \frac{Y_i^2}{b} - CT$	$\frac{\text{SS Treatment}}{\text{DF treatment}}$	$\frac{\text{MS treatment}}{\text{MS error}}$
Block	B - 1	$\sum_{i} \frac{Y_i^2}{t} - CT$	$\frac{\text{SS block}}{\text{DF residual}}$	$\frac{\text{MS block}}{\text{MS error}}$
Error	(b-1)(t-1)	By subtraction		

CT = Correction term
N = total number of observation
B = number of blocks
T = number of treatments.

Numerical Example:
Given the data table used in the one-way classification, the ANOV table is given below.

Source of variation	Degree of freedom	Sum of squares	Mean square	F-ratio
Total adjusted for the mean	11	41	3.7	
Treatment	3	4.2	1.4	$\frac{14}{16}$<1= 8.5
Block	2	27.25	13.65	
Residual	6	9.55	4.6	

Test of hypothesis.

For blocks.

H_0:p1 = p2 = p3 i.e There is no block effect
H_A: p1 = p1: 1 = 1 i.e There is block effect.

$F_{0.05}(2,6) = 5.14$
F computed F tabulated (8.5 5.14). There is block effect.
We reject that there is no difference between the block means as that the block means are the same.

4.4 Two-way Classification with Sub-samples.

When more than one observation is made per treatment an additional source of variation exists.
This is the variation which exists between units (sub samples) which are given the same treatment. Separately this source of errors from the residual error, enables us to arrive at a more valid conclusion.
This analysis has the advantage of being more precise because it recognizes the fact that observations made in any experimental unit cannot be exactly the same.

The yield equation is given below.

$Y_{ijk} = \mu + T_i + (tp)_{ij} + \Sigma_{ijk}$

Where (tp) ij = Interaction effect.
The formulas for computing the ANOVA table are given in the following table.

Source	Degree of freedom	Sum of squares	Mean square	E(MS)	F-ratio
Total adjusted for the mean	n-1 m=rts	$\sum_{ijk}^{n} \frac{Y_{ijk}^2}{} - CT$	TSS	-	-
Treatment	t-1	$\sum_{rs} \frac{Y_{j..}^2}{} - CT$	$\frac{TrSS}{T-1}$	$\sigma^2 \varepsilon + S\, tp\, \sigma^2 + rs\, \sigma_t^2$	RMS

Block	r-1	$\Sigma \frac{Y_{i..}^2}{rs} - CT = BSS$	$\frac{BSS}{r-1}$	$\sigma^2 \varepsilon + S\, tp\, \sigma^2 + S_{tp}\, \sigma^2$	$\frac{BMS}{RMS}$
Treatment Block Interaction	(r-1)(r-1) t-1 r-1	$\Sigma \frac{Yij^2}{Ij\ s} - CT - BSS - TrSS$	$\frac{TrB - SS}{(t-1)(r-1)}$	$\sigma^2 \varepsilon + S\, tp\, \sigma^2$	$\frac{TB.MS}{RMS}$
Error	n-(t-1) n-(r-1) -(r-1) (r-1) + 1	TSS – TrSS – BSS –Tr X BSS	$\frac{BSS}{(d.f)_R}$	$\sigma^2 \varepsilon$	-

The Treatment by Block interaction (i.e. T X B) is a sort of Residual. From the table, when dealing with treatment, what we want to test is $rso^2\, y_T$ i.e whether the effect due to treatment is significant or not (or to find out whether it is 0 or not, thus we find from the chart or ANOVA Table, under E(MS) one equation which contains $\sigma^2\varepsilon$ and $s\sigma^2{}_T P$ to make $\sigma^2\, \varepsilon$ & $\sigma^2{}_T = 1$ and the error or equation to use is TXB effect.

This E (MS) of the TXB interaction is used to test the E (MS) of both the treatment and the block, effect, this is because the E (MS) of TXB interaction contains both $\sigma^2\varepsilon$ & $S\sigma_{Tp}{}^2$ i.e. the overall error and the joint error respectively which occur in both the treatment effect which is $S\sigma^2{}_{Tp}$, we use the Residual error which is $\sigma^2\varepsilon$.

The E (MS) is a guide to enable one tell what to test what.

Note:

$\sigma^2\varepsilon$ = Overall effect

$S\sigma_{Tp}{}^2$ = joint effect

$rso^2\tau$ = Treatment effect alone

$St\, \sigma^2$ = block effect alone

$$CT = \frac{Y.....2}{Rts}$$

4.5 ANOVA FOR A LATIN SQUARE EXPERIMENT.

Here we look at

(a) Row = r

(b) Columns = c

(c) treatments = t

What differentiates the Latin square from the 2 way classification is that in the way, we can have different numbers of blocks, but in the Latin square experiment. The numbers of rows must be equal to the numbers of column.

i.e. r = c ---------- 2 way classification

r=c=t ---------- Latin Square

Thus if there are 4 treatments, there must be 4 rows and 4 columns.

It is recorded as columns Treatments

	1	2	3	4	Yi
1	A	B	C	D	
2	D	A	B	C	
3	C	D	A	B	
4	B	C	D	A	
Y.j					Y..

From the above chart, there must be 1 treatment/row/column i.e. all treatments must be represented in every row and column in such a way that no treatment occurs twice in any row or column.

In the above treatments A, B,C,D are observations.

The column and row totals are computed as Y_j & Y_i respectively.

For the Treatment total Tr T, there will be one for

A_T = A1 + A2 + A3 + A4

B_T = B1 +B2 +B3+B4

C_T = C1 + C2 +C3 + C4

D_T = D1 + D2 +D3 +D4

i.e. pick all the As i.e. $A_1+A_2+A_3+A_4$ = ass up for A_T. Same for B, C, D.

Once this is done the ANOVA is simple.

Source	Degree of freedom	Sum of squares	Mean square	E(MS)	F-ratio
Total adjusted for the mean	n= r x c n - 1	$\Sigma Y_{ij}^2 - CT$	$\frac{TSS}{n-1}$	-	-
Row	r-1	$\Sigma \frac{Y_{i..}^2}{C} - CT$	$\frac{RowSS}{r-1}$	$\sigma^2 \varepsilon + c\sigma^2 p$	$\frac{RowMS}{RMS}$
Column	c-1	$\Sigma_j \frac{Y_{j..}^2}{r} - CT$	$\frac{ColSS}{c-1}$	$\sigma^2 \varepsilon + \sigma^2$	$\frac{CMS}{RMS}$
Treatment	t -1	$\Sigma_j \frac{Yj^2}{r} - CT$	$\frac{Tr.\ SS}{(t-1)}$	$\sigma^2 \varepsilon + \sigma^2$	$\frac{Tr.MS}{RMS}$
Error	n- r – c- t + 2	By substitution	Substitution	$\sigma^2 \varepsilon$	-

* Yt = YtA + Ytb + Y

Summary by spacing

A (2") = 203 + 283 + 279 + 334 + 250 = 1,349

B (4") = 257 + 252 + 266 + 280 + 259 = 1,314

C (6") = 231 + 204 + 280 + 287 + 260 = 1,262

D (8") = 245 + 271 + 227 + 246 + 202 = 1,191

E (10") = 182 + 230 + 245 + 195 + 228 = 1,188

Calculations:

$C.T = \frac{Y.02}{rc} = \frac{(6,304)^2}{25}$ 1,589,617

Total = Yij^2 -CT = $(257)^2$ +…….+ $(338)^2$ - CT = 36,571

Row = $\frac{Yi^2}{C}$ -CT = $\frac{(1,255)^2 + \ldots\ldots + (1440)^2}{5}$ - CT = 13,601

Column = Yj^2 - CT = $\frac{(1,228)^2 + \ldots\ldots\ldots + (1309)^2}{5}$ - CT = 6,146

Spacing = Yt^2 – CT = $\frac{A^2 + B^2 + C^2 + D^2 + E^2}{r}$ - CT

Residual = 36,571 – (13,601 + 6146 + 4156) = 12,668

5.0 Design of Experiment

In carrying out scientific experiments we are very often interested in comparing the effects of certain treatments on the elements we are investigating. Alternatively, we may concern ourselves with the determination of some properties of our elements of interest. In carrying out the investigation, we will discover that certain factors will affect our results. Some of these factors are controllable while others are not. By designing the experiment in a particular way we may be able to observe the factors that we can control as against those we cannot control and thereby be able to make inferences on them. The right design has to be employed in any experiment if the results are to give valid conclusions. However, what type of design we employ will depend on the particular experiment and on the characteristics we are investigating.

Time personnel and funds are also important.

Requirements of a good experiment

To be able to carry out a good experiment the following conditions are important.

(1) Absence of systematic error: this implies that the experimental units which are designed to receive one treatment should not be different in any systematic way from those receiving other treatments. If they do, a source of variation (systematic error) is introduced and this will affect the validity of inferences made from the result of the experiment.

(2) Precision: The treatment comparison must be as precise as possible. By a measure of precision of the estimate, e.g. mean $\overline{X}$ however, precision refers to a measure of how close to the set of possible sample estimates for a particular sample design may be expected to come to.

(3) Range of validity: the experiment must allow for a wide range of validity on the conclusion. This means that as the units for the experiment is carried out; the wider is the range of validity of the conclusions.

(4) Simplicity: As much as possible, the experimental arrangement should be simple. If the experiment is not simple, them the procedure might be difficult to follow especially if the experiment is to be carried out by unskilled people.

(5) Also, for a good experiment, the error must be assessable. We must be able to measure the amount of uncertainty involved in the experiment. This means that we must have a set of experimental units responding independently to the same treatment and such units must differ only randomly from the units of other treatments.

(6) The experiment must not only be simple, precise and its error measurable, it must also be cheap.

5.1 Completely Randomized Design

The completely randomized design is appropriate when the experimental units are not known or anticipated to be different from one another. In this case any variation in the results observed from the units will be largely due to the treatment applied.

In the completely randomized design, the experimental units are numbered and the treatments are randomly assigned to them. This means that all the units will have the same chance of being given or not given a particular treatment. The one way classification is used in the analysis of variance for this design.

The completely randomized design has the following advantages:

1. It is simple to carry out and analyse.
2. The design is flexible in that the number of treatments and replicates are limited only by the number of experimental units.
3. The loss of information due to missing data is small related to other designs.

Disadvantages

The mean disadvantage of this design is that it does not take cognizance of other source of variation. Any variation is assumed to be due to be treatments applied. Where the experimental units are not homogenous (apart from treatments) this can lead to erroneous conclusions.

5.2 Randomized Complete Block Design

This design is used where there are other sources of variation apart from the treatment. By the use of this design, we are able to determine the significance of those additional sources of variation. (e.g. the block effect). If the block effect is significant, it means that the precision of the experiment has been increased by the use of this design.

The first step in the design is to group the experimental units into blocks. The main purpose of grouping is to make sure that all the units in a block are as uniform as possible so that observed variation between blocks will be due to the treatment effect. After the blocking, the treatments are then randomly applied to the units of each block. For example, we might be interested in observing the effects of various feed components (treatments) on the weight gain of cattle. If we have a herd of cattle consisting of animals

of difference age group variations in the weight gain by the animals might not be due to the differences in age. To be able to observe the variation due to age, it then becomes necessary first to group the animals according to age before applying the various feeds.

The two-way classification is used in the analysis of randomized complete block design.

5.3 Factorial

In factorial analysis we deal with factors.

Let us look at a two-factor factorial.

Let Factor A = Variety (2 varieties)

Factor B = Fertilizer (N) (2 levels of N. fertilizer).

In this case we talk of a 2 X 2 or 2^2 factorial.

If we call the first level of factor A 0 and its 2^{nd} is 1, we therefore have.

0	0	00
	1	01
1	0	10
	1	11

A N O VA

Source	d.f
Total	3
A	1
B	1
Residual AB	1

Let us assume that we planted something at the green house and the same thing is repeated on the farm, i.e. there now a 3^{rd} factor which is location, we would then have 3-factor factorial.

Thus	A = Variety (3)	15	15	15
	B = Fertilizer NPK (3)	15	20	15
	C = Location (2)	15	30	15

This is a 3^2 by 2 (3x3x2) or 3^2 x 2factorial. We can make a design for this type of factorial thus:

A factorial experimental is one in which an equal number of observations is made for all possible combinations of the levels.

Design of the experiment

Variety	0			1			2		
Fertilizer	0	1	2	0	1	2	0	1	2
Location	0000	010	020	100	110	120	200	210	220
	1001	011	021	101	111	121	201	211	221

After this, randomize following the same procedure

ANOVA

Source	Df	SS	MS
Total	3 x 3 x 2 -1 = 17	$\sum_{i}\sum_{j}\sum_{k}^{3\ 3\ 3} y^2_{ijk} - CT$	
Variety	3 -1 = 2	$\sum_{i=1}^{3} \frac{y_i^2}{6_{j\ k}} - CT$	
Fertilizer	3 – 1 = 2	$\sum_{i} \frac{y.j^2}{6} - CT$	
V & F	(3- 1)(3-1) (2) (2) =4		

	0	1	2	
0	00	10	20	Y.1
1	01	11	21	Y.2
2	02	12	22	Y.3
	Y1…	Y2…	Y3…	Y….

Factorial experiment is one designed to examine the effect of one or more factors, each factor being applied at two levels at least so that differential effects can be observed. It investigates all possible treatment combinations which may be formed from the factors under investigation. The level of a factor denotes that intensity with which it is brought to bear. It may be measured quantitatively as when fertilizer is applied to plots in a given wt/unit area or qualitatively as when patients are considered two levels inoculated and not inoculated.

If we have 2 factors A and B each with 2 levels-0,1 and 0, 1 in effect we shall have 4 treatments -00, 01,10, and 11 as seen on the chart.

2 X 2
A B

	0	1
0	00	01
1	10	11

If this supposed experiment is repeated 4 times so that we have 4 blocks –or replicates as shown in the table below:

Treatments

		00	01	10	11
Block	1				
	2				
	3				
	4				

We can thus set up our ANOVA table.

ANOVA

Source	d.f	SS	MS	F-ratio
Total	15(16-1)			
Blocks	3(4-1)		MS_n	$\frac{MS_B}{MS_E}$
Treatments	3(4-1)		MS_r	$\frac{MS_T}{MS_E}$
Factor A	1		MS_A	$\frac{MS_A}{MS_{AB}}$
Factor B	1		MS_B	$\frac{MS_B}{MS_{AB}}$
A and B	1		MS_{AB}	$\frac{MS_{AB}}{MS_E}$
Residual	9 (15-3-3)		MS_E	

In calculating the F-ratio, we test the blocks and treatment with the Residual which is above (9). In calculating the F-ratio for the Factors A&B, we use the MS for AB to test for factors A,B while we use the overall residual to test for the AB.

Example

N = 2 levels

K = 2 levels

Coded yield of maize

Replicate Treatment

	00	01	10	11	Total
1	2	3	3	4	12
2	5	7	4	5	21
3	2	4	5	6	17
4	4	3	2	1	10
Total	13	17	14	16	60

N2 = A

P

	0	1	
0	00 13	10 14	27
1	01 17	11 16	33
	30	30	60

=B

ANOVA

$RSS = \frac{12^2}{4} + \frac{21^2}{4} + \frac{17^2}{4} + \frac{10^2}{4} + \frac{60^2}{4}$

$TSS = \frac{13^2}{4} + \frac{14^2}{4} + \frac{17^2}{4} + \frac{16^2}{4} + \frac{60^2}{4}$

$ASS = \frac{30^2}{4} + \frac{30^2}{4} + \frac{3600}{16} = 0$

$BSS = \frac{27^2}{8} + \frac{33^2}{8} - \frac{3600}{16} = 4.2$

Source	Df	SS	MS	F	
Total	15	19	-	-	
Replicate	3	10.5	3.5	3.5/7	Significant
Treatment	3	2.5	0.83	0.83/0.7	
A	1	0	0.0	0	
B	1	2.2	2.2	2.2/0.3	Significant
AB	1	0.3	0.3	2.2/0.7	
Residual	9	6	0.7		

3x3 Factorial i.e.3^2 – 2 factors each at 3 levels.
If we have 2 factors with each at 3 levels
The 2 factors are A&B
3 levels are 0, 1, 2.

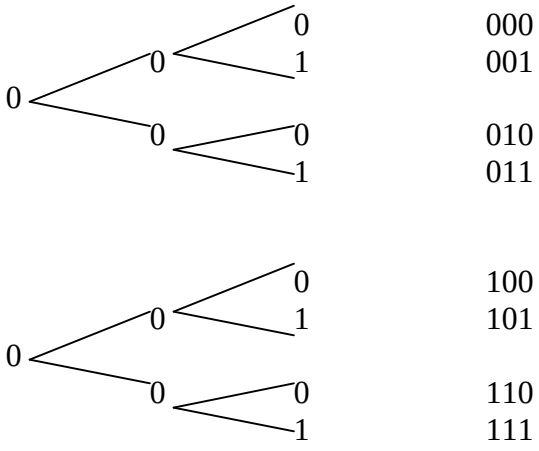
The treatments will be 9 treatments thus

0 1 2

0 1 2

00, 01, 02,10,11,12,20,21,22

2^3-3 factors at 2 levels.



8 treatments

Here there are 2^3, 3 factors each at 2 levels and thus we have 8 treatments. If we assume there are 5 replicate each with 8 treatments.

Replicates	1	2	3	4	5	6	7	8	Y1
1									
2									
3									
4									
5									40
Y5									Y

We can set up the ANOVA Table.
There are 3 models which we can use in this.

In model 1- we use the Residual to test all eff4ects; in models II and II we use the expected mean square as a guide for determining appropriate denominator for testing.
The ANOVA table is as follows.

ANOV TABLE

Source	d. f	SS	MS	F-ratios
Total	39			
Replicate.	4			
Treatment	7			
A	1			
B	1			
AB	1			
C	1			
AC	1			
BC	1			
ABC	1			
Residual	28			

Model I- All factors fixed
Model II – All factors random
Model III - Some factors random some fixed.

A		0			1		
B	0	1			0	1	
C	1	0	1	0	1	0	
ABC 0 0 0	001	10	011	100	101	110	111

Exercise:

1. Explain what we mean by Factorial Design
2. Differentiate between Factorial design and completely randomized design.

6.0 LATIN SQUARE DESIGN

The LS design has two basic restrictions in that the number of rows = number of columns = number of treatment and no one treatment must appear twice in the same row the same column. Such a design is presented below.

A	B	C	D	E
E	A	B	C	D
D	E	A	B	C
C	D	E	A	B
B	C	D	E	A

A E D C B
B A E D C
C B A E D
D C B A E
E D C B A

We must now randomize such that the restrictions above still hold. The randomization can be by column or by row. This can be done by changing the column to row. The analysis of variance for the Latin square Design is the same as done in the past

The Latin square design is like randomized complete block treatment but different in that no one treatment must appear twice in the same row or column. Thus the Latin square is a diagonal effect and vertical effects.

7.0 SPLIT PLOT DESIGN

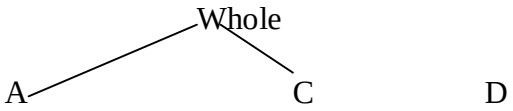
Suppose there are 2 factors A & B. A has 3 and B has 2 levels- and we want to use Latin methods. To use the Latin square we find the treatment combination so as to randomize.

A 3 levels
B 2 levels
} 6 treatments

For the square design, we shall have 6 rows, 6 columns and 6 treatments.

	A		O		1	2	
B		O	1	0	1	0	1
	OO		O1	10	11	20	21

We assume that factor A requires a large piece of land while factor B does not. We can divide the whole piece of land into Blocks e.g. the piece of land is divided into 3 blocks. Each block can be divided up into whole plots and each whole plot is divided again into sub plots. This processed will happen to each block.



Sub Plot

a_1				
a_2				
a_3				
a_4				

If the factor A is the methods of land preparation and each factor A is divided into 4 plots we have a1,a_2 , a_3, a_4. while factor B is the varieties of crop (maize) to plant and because there are 2 levels of B or 2 Sub plots per whole we have b_1, b_2. we randomize factor A in to each whole plot. i.e. each whole plot is now a_1, a_2, a_3, a_4. after this randomization of factor A among whole plots we now randomize factor (i.e. Varieties) into the sub plots. Thus this is a 2- stage randomization procedure.

A 4 - a_1, a_2, a_3, a_4.
B 2 - b_1, b_2

This shows that each factor A/ level will contain all the levels of factors B. i.e. a1 contains b_1 & b_2. in considering the yield in a b1.. The yield is thus made of

Y_{ij} =

Where Bk = sub plot effect

U = overall mean
T_i = Block effect
P_i = Whole plot effect
Sij = error due to block and whole plot interaction effect
$(PB)_{jk}$ = error due to whole plot and sub plot interaction
Σ_{ijk} = overall error.

The split plot design is different from the randomized complete block design R.C.B.

(1) It splits the factors

(2) It incorporates the principles of R.C.B. design e.g. if there ae 2 factors A & B and A has 4 and B 2 levels there will be 8 treatments. In the R.C.B. design the 8 treatments will be randomized completely c within in the blocks, but in the split plot design, the factors will be split into A and B and one factor will be randomized within the other factor i.e. B. within A.

In the R.C.B. design we have
$Y_{ijk} = \mu + t_i + pj + (tp)_{ij} + \Sigma_{ijk}$
Where t_i, i.e. treatment effect will be the 8 treatments
i.e. block effect will be the 4 blocks

But in the split plot design we have
$Y_{ijk} = \mu + t_1 + pi\S ij + \beta_k + (tp)ij + \Sigma ijk$

ANOVA

	Sources	D.f
Y	Total	23
T	Replicates	2
P	whole plots	3
S	Error a	6
B	Sub plots	1
(rp)	Interaction	3
E_{ijk}	Error (b)	8

8.0 REGRESSION

Regression is a technique for quantifying the relationship among variables, one of which (the response or dependent variable) is known to be functionally dependent on one or more other variables called explanatory or independent variable:

Regression has considerable application in Animal Science Research. For example, in studying growth rate in animals, the researcher regresses body weight on time. The regression coefficient obtained represents the change in body weight per unit time which is growth rate. A physiologist may want to determine the rate of disappearance of a certain drug from the blood. He may also wish to study the changes in blood cholesterol concentration in animals following administration of increasing doses of a certain drug. The nutritionist may be interested in the change in feed conversion with increasing levels of a certain nutrient or increasing ambient temperature. All these studies involve functional relationship between variables and regression techniques are very useful. The type of regression which a researcher may be interested in clued simple linear regression, multiple regression, polynomial and non-linear regression.

Simple Linear Regression (SLR)

Here, only two variables are involved, one response and one explanatory. The model is:
$y_i = b_o + b_1 x_1 + e_i$
Where b_o and b_1 are the intercept and slope respectively and are estimates of the parameters, $ß_o$ and $ß_1$

From sample data, numeric values for the statistics can be obtained by solving a set of normal simultaneous equations in two unknowns. However, easy-to-use formulae are available for hand computations.

$$b_1 = \frac{\Sigma xy - (\Sigma x.\ \Sigma x)/n}{\Sigma x^2 - (\Sigma x)^2/n}$$

and

$$b_o = y - b_1 x$$

If it sis known *apriori* that will be zero when x is zero, then regression is done through the origin. The model becomes:

$$y_1 = bx_1 + e_1$$

Simple estimated b y:

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

e.g. fit a regression line through the pairs of measurements

X	Y	XY	X^2	Y^2
1	0.06	0.06	1	0.0036
2	0.10	0.02	4	0.01
3	0.19	0.57	9	0.0361
4	0.34	1.36	16	0.1156
5	0.47	2.35	25	0.2209
6	0.63	3.78	36	0.3969
7	0.84	5.88	49	0.7056
8	1.09	8.72	64	1.1881

$\sum X = 36 \sum Y = 3.72 \sum XY = 22.92 \sum X^2$ =204 $\sum Y^2$ =2.6768

$$b = \frac{22.92 - (36)(3.72)/8}{204-(36)^2/8}$$

$$= 6.18/42 = 0.147$$

$b_0 = 0.465 - (0.147)(4.5) = -0.197$

Hence, our model becomes

$$y_1 = -0.197 + 0.147x$$

The line corresponding to this equation is obtained by plotting the (x_i, y_i) points.

The regression coefficient obtained is interpreted to mean that body weight was increasing by 0.147kg per week on the average. In other words, the

$$\text{Total SS} = \frac{\Sigma y^2 - (\Sigma y)^2/n}{\Sigma y^2}$$

$$SS = \frac{(\Sigma xy - (\Sigma x.\ \Sigma y)/n)^2}{\Sigma x^2 - (\Sigma x)^2/n}$$

or $\frac{(\Sigma xy)^2}{\Sigma x^2}$ with I D.f for regression through origin

Residual SS = Total SS- Regresion SS with n-2 or n- 1 D.f as appropriate

Corresponding mean squares are obtained as used and a variance ratio, f computed as:

$$F = \frac{\text{Regresion Mean Square}}{\text{Residual Mean Square}}$$

This F is then compared with tabulated F as appropriate degrees of freedom.

From our example

Total SS = $2.6768 - \frac{(3.72)2}{8}$

= 0.947 with 7 D.f

Regression SS = $\frac{(22.92-(36)\ (3.72)/8)2}{204-\frac{(36)}{8}}$

= $\frac{38.1924}{42}$

= 0.909 with 1 D.f

Residual SS = 0.97-0.909

= 0.038 with 6 D.f

Our regression ANOVA table thus becomes

SOURCE	D.F.	SS	MS	F
Total	7	0.947		
Regression	1	0.909	0.909**	151.5
Residual	6	0.038	0.006	

The analysis shows that the regression is highly significant ($P< 0.01$)

The standard error of the regression coefficient is given as:

$S.E\ (b) = Sy.x/\sqrt{\Sigma(x_1 - x)^2}$

Where $Sy.\ x$ = Residual MS or

One commonly used criterion for determining the adequacy of a fitted regression model is the R^2 or r^2 (coefficient of multiple determination) or $1-R^2$ (coefficient of allenation). R^2 is the proportion of the variability in Y explained by the fitted regression model.

$R^2 = \frac{\text{Regression SS}}{\text{Total SS, corrected for y}}$

And R, the correction between x and y = $\pm\sqrt{R^2}$ -. The sign will correspond to the sign of the value $\Sigma xy - (\Sigma x)(\Sigma y)/n$ the square of which is the numerator of regression SS. in our example:

R2 = $\frac{0.909}{0.947}$ = 0.96 or 96%

and R = 0.98

9.0 SAMPLING

9.1 What is a sample?

In conducting an experiment it is usually not possible to get information on the entire population we are interested in. what is done then is to select a manageable portion of the population which is referred to as a SAMPLE.

A sample must be a true representative of the population if it is to lead to valid inferences.

9.2 WHY DO WE SAMPLE?

1. The main objective for sampling is to reduce the population to a manageable proportion.
2. The population may be homogenous that the result obtained from the sample is as equally useful as that obtained from the population.
3. The whole population might be inaccessible.
4. Collecting information from the whole population might be practically impossible in terms of the personnel required.
5. It may be too costly money wise to work with the whole population.

9.3 SAMPLING METHODS

Many methods exist for selecting the sample from a population and which method is used will depend on the particular situation, the type of data required and any other constraints on the researcher. Sampling methods are:

1. Simple Random Sampling

Simple random sampling is a procedure whereby every element of the population is given the chance of being included or excluded from the sample. In making such a selection, the following procedure is followed:

a. A list of all the elements in the population is obtained.
b. The various elements are identified with numbers e.g. 1, 2, 3, 4, 5 etc.
c. The size of the sample is decided upon.
d. The sample is selected with the use of the table of random numbers.

The major advantage in simple random sampling is that it gives all the members of the population a chance of being included or excluded from the sample. However, it has a disadvantage where the population is so large since the identification of each element in the population might be impracticable. Also, the use of simple random sampling is limited only to finite population.

2. Stratified Sampling Technique

In this method, the population is first partitioned or stratified into sub- populations known as strata and simple random sampling is then used to obtain samples from each stratum. The purpose of stratifying is to ensure that certain characteristics of the population which may be eliminated by the use of simple random sampling and which we want to included in the sample are taken care of . Stratification could be carried out on the basis of sex, religion, ethnic groups, income etc. the number of the sub-sample from each stratum will depend on the proportion of that group in the population.

Before stratified sampling technique could be used, the following conditions must be satisfied:

a. The population must be heterogeneous with respect to the variable of interest.
b. The population must be divisible with subsets.

c. The elements of each sub-group or stratum must be heterogeneous between (among) themselves. In other words, the subgroups must have different means and possibly zero variances.

3. Cluster Sampling Technique.

As in the stratified sampling technique, the population is partitioned into clusters before sub-samples are taken from each cluster. Unlike a stratum however, the elements in a cluster ate heterogonous while the clusters are homogeneous between or among themselves. Where the population is large, each cluster can be further sub-divided into sub-clusters or strata before carrying out a simple random sampling on each sub-group.

4. Purpose Sampling Technique

Here the sample taken depends mainly on the purpose for which the survey is carried out. What elements included or excluded for the sample if left to the judgment of the researcher (usually an expert) (e.g. political campaigns)-18 years old and above.

5. Systematic Sampling Technique.

There are times whereby it might not be possible to group or list the elements in a population. This is commonly the case in market surveys and sociology. Systematic sampling involves selecting the elements for our sample at regular interval but in a consistent way. In a market survey for example, we may decide to talk to every 5^{th} or 10^{th} person we come across in the market.

6. Multi-stage or Nested Sampling Technique

The procedure involves more than one sampling stage. Any of the other sampling procedures could be used in any of these stages.

Suppose we want to relate the performance of students in the University of Agriculture, Abeokuta to their ethnic background, the first stage will involve grouping the students according to ethnic groups. The second stage might involve grouping elements in each cluster according to some social factors such as religion, sex, income etc. the third stage will then involve taking a random sample from each of the strata.
Thus in the multi-stage sampling procedure we might use.

Cluster sampling	-	first stage
Stratified sampling	-	Second stage
Simple random sampling	-	Third Stage.