BIOMETRY 210

Introduction to Statistical Methods

COURSE NOTES

Notes Compiled by :
Harvey M Dicks
Consulting Biometrician/
Former Senior Lecturer Department of Statistics & Biometry,
University of Natal, Pietermaritzburg, South Africa
Home Address: 85 Hesketh Drive, Hayfields,
Pietermaritzburg 3201
dicks@ukzn.ac.za
hmd/2006

CHAPTER 1 : INTRODUCTION

1.1 STATISTICS AND BIOMETRY

STATISTICS

STATISTICS is the science of creating, developing, and applying techniques such that decisions can be made and evaluated in the face of uncertainty. It involves the collection, analysis and interpretation of numerical data.

BIOMETRY

BIOMETRY is the application of the statistical method to biological and agricultural research problems. It involves the collection of data by means of experiments or surveys conducted according to various principles, and the drawing of conclusions (or inferences) from data through the use of various procedures known as statistical analysis.

1.2 GENERAL DEFINITIONS

1.2.1 POPULATION

A POPULATION is the set of all conceivably possible (or hypothetically possible) observations of a given phenomenon that is of interest to the experimenter. A population can be infinite, in which case the set of measurements exists in a hypothetical sense. A population may also be finite. In general the size of a population, if finite, is given by N.

1.2.2 SAMPLE/RANDOM SAMPLE

A SAMPLE is a selection from or a subset of a population. In a RANDOM SAMPLE, the members of the subset are chosen in such a way that every member in the population has an equal probability* of being chosen, and the selection of one member does not affect the probability of selection of any other. In general, the size of a sample is given by **n**.

NOTE

i) For reasons such as cost, time, etc, it is usually necessary to study a sample from a population to obtain information about the population.

ii) Inferences about a population are based on sample data only.

iii) The concept and definition of probability will be dealt with in chapter 3.

1.2.3 PARAMETER

Any numerical or descriptive measure of a population is referred to as a PARAMETER. Since a parameter is a measure from the entire population, it is a fixed value.

1.2.4 STATISTIC

Any numerical or descriptive measure of a sample is referred to as a STATISTIC. Usually, a given population parameter corresponds to a sample statistic. A statistic varies from sample to sample and within samples.

1.2.5 ESTIMATE

A sample statistic ESTIMATES a given population parameter. Since it is generally not possible, or very difficult, to calculate the value of the population parameter directly, we calculate a sample statistic that corresponds to that parameter and it is hopefully close to its true value.

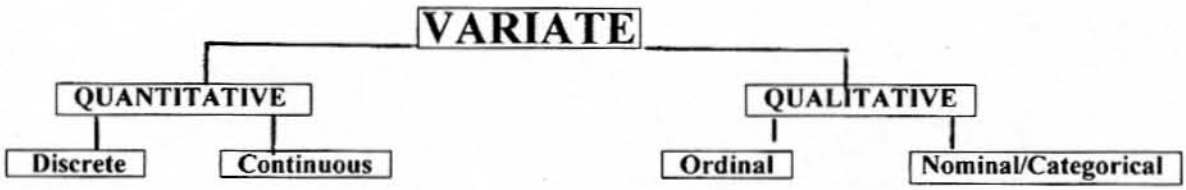
1.2.6 EXPERIMENT/TRIAL/EVENT

An EXPERIMENT is the process by which an observation (or measurement) is obtained. The experiment, when performed, is often called a TRIAL. An EVENT is the outcome of the experiment.

EXAMPLE: A coin is tossed. This is the experiment. The outcome is either `Heads' or `Tails'. The outcome that is obtained in the experiment is the event.

1.2.7 VARIATE (Variable)

A characteristic or phenomenon that varies from one (biological) entity to another is called a VARIATE (or VARIABLE). A variate can take on any number of observations. If 'Y' denotes the variate, then Y_i denotes the i^{th} observation of that variate. If we have a random sample of size n, a variate will take on n values denoted by, $Y_1, Y_2 \ldots Y_n$. Different sorts of variables are encountered by biologists and it is desirable to distinguish among them. This distinction is illustrated in the diagram below.



(a) **A QUANTITATIVE variate** includes data that represents the amount or quantity of something and measurement is made over a range of values. The observations of a quantitative variate possess a natural order or ranking.

i) A continuous quantitative variate is one for which all values in some range are possible.
e.g., height, mass, etc.

ii) A discrete quantitative variate is one for which only certain values are possible. They are usually consecutive integers; for example, size of household, number of insects in an insect trap, etc.

(b) **A QUALITATIVE variate** describes a characteristic or attribute of an individual. It includes data that can be categorised but not quantified.

i) Ordinal variates deal with relative differences,
e.g., (short – average – tall), (bad – good – excellent), etc. This type of variates is often represented by 'scores' such as 1 – 2 – 3, etc. and as such are erroneously treated statistically as discrete variables.

ii) Nominal/Categorical variates are associated with some quality, characteristic or attribute which the variable posses for example, eye colour: (blue – grey – green – brown); vehicle type: (bus – truck – car – bicycle); etc.

1.2.8 RANDOM VARIATE

The observations of a RANDOM VARIATE can assume any one of the possible values in a population with equal likelihood. A random variate is associated with a random experiment, the events of which cannot be predicted with certainty.

1.3 EXPERIMENTS AND SURVEYS

A researcher usually creates or generates his/her own data by means of an EXPERIMENT or SURVEY.

An **EXPERIMENT** can be more complicated than the example given earlier. In general, observations from an experiment are obtained as a planned inquiry to obtain new facts or to confirm or deny the results of previous experiments. The factors under study are varied in a controlled manner. All other factors, except the variate under question, are kept constant as much as is possible.

In a **SURVEY**, data is collected over various locations, with ***no control over the factor under study***. Extraneous factors *cannot* be held constant.

EXAMPLE: Research Objective: *What is the effect of Nitrogen on Maize yields?*

a) **A SURVEY** would involve the study of data collected from various locations. Here the variation in Nitrogen levels is not controlled, and the ability to control variation in extraneous factors, such as fertility, is limited.

b) **An EXPERIMENT** would include the yields from a trial planted with controlled applications of Nitrogen on a site in which the general variation in fertility can be controlled or is known.

1.3.1 THE HYPOTHESIS

An HYPOTHESIS is a supposition, a 'general rule', made as the basis for reasoning without assumption of its truth. Experiments in Biometry are sometimes used to test hypotheses, such as 'Variety A gives better yields in this region than all other varieties.' Hypotheses can never be proven, only verified; a single experiment can disprove an hypothesis.

1.3.2 TYPES OF EXPERIMENTS

(a) COMPARATIVE vs. ABSOLUTE Experiments

In a COMPARATIVE EXPERIMENT, two or more procedures or treatments are tested against oneanother by means of some criterion for performance.

In an ABSOLUTE EXPERIMENT, a single procedure is studied and/or the purpose is to obtain a certain single measurement.

(b) OBJECTIVE Experimentation

* To verify hypotheses.
* To yield `empirical knowledge' (i.e. knowledge based on observation or experience rather than theory). Such an experiment uses the `empirical method'.

(c) OBSERVATIONAL Experiments

* Use to demonstrate (usually visually) the effectiveness or otherwise of particulat practices or recommendations.

1.3.3 INDUCTION vs. DEDUCTION

a) INDUCTION: Inferring more general propositions or laws from less general propositions or laws (even particular instances).

b) DEDUCTION: Inferring less general propositions or laws (even particular instances) from more general propositions or laws.

[particular instances] -----induction-------> [general rule]

[particular instances] <----deduction------ [general rule]

Induction and deduction can be used together in experimental work. A hypothesis, a sort of general rule, may be induced from certain observations (from an experiment using the empirical method). To verify this general rule, one can deduce an experiment. From the results of this experiment, one can induce a new hypothesis.

CHAPTER 2: GRAPHICAL METHODS OF DATA DESCRIPTION

2.1 FREQUENCY DISTRIBUTIONS

Consider the results of a survey in which the gross income of 8644 households is obtained. Such information would cover this page at least and would be somewhat meaningless. In such situations it is necessary to summarise the data in such a way that some 'picture' is obtained. The table below is the result of such a summary.

EXAMPLE 2.1 : Income per household ($ per month)

Income (Dollars)	Number of Households (frequency)
<$2000	1406
2000 –	4352
4000 –	1833
6000 –	489
10000 –	163
20000 +	406
	8644

A second example considers the distribution of the employees of a particular organisation according to their job descriptions. A summary for the 200 employees as contained in Table 2.2

EXAMPLE 2.2: Labour Categories

Work category	**Number of employees (frequency)**
Management	35
Technical	20
Semi–skilled	59
Unskilled	86
	200

GUIDELINES FOR THE GROUPING OF DATA

a) Determine the number of classes (groups). For continuous and discrete data a general rule is: Number of classes $= |\sqrt{N}|$, where N = number of observations. This "rule" can result in an unnecessarily large number of classes. An alternative "rule" is determine k, such that $2^k \geq N$. In practice one will seldom use fewer than 6 or more than 16 classes.

b) Whenever possible ensure that class intervals are the same for all classes.

c) Ensure that each item (observation or measurement) is allocated to one class and one class only.

2.1.1 Class intervals/class boundaries/class centres (continuous or discrete data only)

Consider the class $4000, with frequency, 1833 households in Table 2.1 above.

Nominal class boundaries:

Upper class boundary = $6000
Lower class boundary = $4000

Exact class boundaries:

Upper class boundary = $5999.5
Lower class boundary = $3999.5

In both cases the class interval (i.e., Upper – lower) = $2000.

NOTE

i) The class interval is not constant for all classes in this example.

ii) Also, the class centres (midpoints) are not the same.

c.f. (Nominal class centre) : $\frac{(\text{upper}+\text{lower})}{2} = \$5000$

(Exact class centre) : $\frac{(\text{upper}+\text{lower})}{2} = \$4999.5$

iii) When using a computer statistical package (e.g., GENSTAT.) it is usually necessary to specify the upper (or lower) class boundaries only. If they are not specified, the number of classes, with equally spaced intervals, are determined according to the "$\sqrt{}$" rule given above.

2.2 GRAPHICAL METHODS (Summary):

(a) **CONTINUOUS** data
- – histograms
- – line graphs (frequency polygons)
- – cumulative frequency curves (ogive)

(b) **DISCRETE** data
- – bar graphs
- – pie graphs

(c) **QUALITATIVE** data (nominal and ordinal)
- – bar graphs
- – pie graphs

Note:
Bar graphs are used differently for discrete and nominal data since, for qualitative data, the *order* of the classes can change, this is not the case for discrete data where the data (variate) is in a definite (or ranked) order, e.g., 0 – 1 – 2 – 3 etc. Bar Charts and Histograms are not synonymous.

CHAPTER 3: NUMERICAL METHODS OF DATA DESCRIPTION

3.1 MEASURES OF LOCATION – CENTRAL TENDENCY (Averages)

3.1.1 THE ARITHMETIC MEAN (Mean)

(a) UNGROUPED data

Let N be the size of a population, and n be the size of a sample from that population. The population mean, denoted by a population parameter μ *(the Greek letter mu)*, is defined as :

$$\mu = \frac{1}{N}\sum_{1}^{N} x_i \qquad \{3.1\}$$

Usually N is too large for μ to be calculated directly, or the population is hypothetical so that it is impossible to calculate μ. Hence it must be estimated by the sample mean $\bar{x}$ (x–bar), which is defined as

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i \qquad \{3.2\}$$

Effect of change of origin/scale on the mean

i) If variate X undergoes a change of origin by having a constant A subtracted from each variate value then $\bar{X}$ changes to $\bar{X} - A$,

i.e. if $Y = X \pm A$ then $\bar{Y} = \bar{X} \pm A$

ii) If X undergoes a change of scale by having each variate value multiplied by a constant k, then $\bar{X}$ changes to $k\bar{X}$.

i.e. if $Y = kX$ then $\bar{Y} = k\bar{X}$

(b) GROUPED data

We will use the same symbol $\bar{x}$ to designate the sample mean for grouped data. Since we cannot reconstruct the actual sample measurements from the grouped data, we represent all values in a given class interval by the midpoint (class centre). Let m_i be the midpoint of the ith interval, and let f_i be the frequency for that interval. Then the sample mean for grouped data is given by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i m_i \qquad \{3.3\}$$

3.1.2 THE MEDIAN (m)

If a given variates observations are ranked from smallest to largest, the central value is known as the median. Hence, 50% of the measurements lie above the median and 50% lie below it.

3.1.2.1 Calculation of the MEDIAN

(a) Ungrouped Data (odd number of sample values):

i) Rank the n measurements from smallest to largest

ii) The median is the sample value with rank $(n + 1)/2$

EXAMPLE 3.1 : (n = 5)

Sample: 7, 3, 12, 1, 8. Ranked sample: 1, 3, 7, 8, 12.

Position of Median: $(5 + 1)/2 = 3^{rd}$ observation. Thus : Median = 7

(b) Ungrouped Data (even number of sample values):

i) Rank the n measurements from smallest to largest

ii) The median is the sample value halfway between the measurement with rank $(\frac{n}{2})$ and $(\frac{n}{2} + 1)$.

EXAMPLE 3.2: (n = 10)
Sample: 97, 110, 122, 90, 111, 83, 101, 115, 99, 104.
Ranked sample: 83, 90, 97, 99, 101, 104, 110, 111, 115, 122.
Position of sample value below median: $(n/2)^{th} = (10/2)^{th} = 5^{th}$ observation
Position of sample value above median: $(n/2 + 1)^{th} = (10/2 + 1)^{th} = 6^{th}$ observation

Thus : Median $= \frac{(101 + 104)}{2} = 102.5$

(c) Grouped Data
i) Calculate the cumulative frequency distribution.
ii) Establish the class in which the median value falls.
iii) Calculate the median by interpolation.
iv) Calculate the relative cumulative frequency;
NOTE: This is required if the cumulative frequency distribution (ogive) is to be plotted.

EXAMPLE 3.3: INCOME PER MONTH (Rands)

Class: Income per month (R)	Frequency	Cumulative Frequency
< 250	0	0
250 –	2	2
750 –	1	3
1250 –	5	8
1750 –	22	30
2250 –	9	39
2750 –	5	44
3250 –	4	48
3750 –	1	49
4250 –	0	49
4750 –	0	49
5250 +	1	50

NOTE:
The cumulative frequency refers to that number of observations up to the upper class boundary of that class. For example, 30 is the cumulative number of individuals in the sample with monthly incomes less than R2250

CALCULATIONS: Since n = 50, the median (m) will be that value such that 25 observations lie on either side. The median (m) therefore lies in class interval: R1750 – R2249.

Thus Median (m) = R1749.5 + $\left\{\frac{17}{22} \times 500\right\}$ = R 2135.9
i.e. Median = (exact lower class boundary) + (fraction x class interval)

Where 'fraction' = that proportion of the "next" class frequency; in the example above = 22), such that the cumulative frequency at that point equals half total frequency.

Note: The Median is NOT affected by extreme values as the mean would be. As a result, it is often used in the social sciences, where extreme values are common.

3.1.3 THE MODE
The mode is the value of a variate for which the relative frequency (or probability) is a maximum.
Note: This statistic can only be estimated for *grouped* or *continuous frequency distribution* data

3.1.4 OTHER MEASURES OF LOCATION

(a) **PERCENTILES**
The percentage of the total number of observations that are less than the given value. For example, since 50% of the observations in a set of data are less than the median, the median is the 50th percentile.

(b) **QUARTILES**
i) The first quartile is the 25th percentile
ii) The third quartile is the 75th percentile, etc.

(c) **DECILES**

The 10th percentile is the first decile, the 20th percentile is the second decile, etc.

(d) **QUANTILES** (– a generic term for quartiles, deciles, percentiles, etc.)

3.1.5 SKEW AND SYMMETRIC DISTRIBUTIONS

(a) If Mean = Median = Mode, the distribution is said to be **SYMMETRIC**
(b) If Mean < Median < Mode, the distribution is **NEGATIVE SKEW**
(c) If Mean > Median > Mode, the distribution is **POSITIVE SKEW.**

3.2 MEASURES OF DISPERSION – (i.e. VARIATION or "ERROR")

A characteristic of all experimental material is VARIATION. In simple terms, variation describes the extent to which observations are dispersed (clustered) about the mean. A set of data where observations cluster closely about the mean has little variation, whereas a set of data where observations are widely dispersed has much variation.
In statistics, variation is often referred to as ERROR. There is *nothing incorrect* about error, it occurs naturally. Depending on the circumstances, error can either be EXPERIMENTAL ERROR or SAMPLING ERROR and in statistical analysis it is important that the researcher is able to distinguish between these two forms of error.

3.2.1 THE RANGE

This is defined simple as the *difference between the highest and lowest observed values.*
EXAMPLE: Data set: 7.2, 14.6, 19.7, 9.4, 12.4, 5.3, 14.6
i.e. Range = 19.7 – 5.3 = 14.4

Advantages :

a) easy to calculate.
b) easy to understand.
c) does measure total spread.

Disadvantages :

a) it cannot be used for frequency data
b) it is affected by extreme values
c) it does not indicate whether the values are uniformly dispersed or perhaps clustered about the mean.

3.2.2 MEAN ABSOLUTE DEVIATION

This is defined as : **Mean deviation** $= \frac{1}{n}\sum |(x - \bar{x})|$ {3.4}
i.e., the average of the *absolute* deviations of the variate x from its sample mean $\bar{x}$.

Disadvantage :

a) algebraically intractable.
b) cumbersome to calculate.

3.2.3 VARIANCE

The measure of variation/dispersion which is most commonly used is the **VARIANCE** or its square root, the STANDARD DEVIATION.

(a) UNGROUPED DATA

Let N be the size of a population, and n be the size of a sample from that population. The population variance of a variate X, denoted by population parameter, σ_x^2, is defined as

$$\sigma_x^2 = \frac{1}{N}\sum_{1}^{N}(x_i - \mu)^2 \qquad \{3.5\}$$

Thus variance is the average of the sum of the squared deviation of the values of the variate x about its true mean μ.
As was the case with the mean, N is usually too large, or the population is hypothetical, so that σ_x^2 is difficult or impossible to calculate. Hence, the population variance must be estimated by the sample variance, denoted by the sample statistic s_x^2, which when μ **is known**, is defined as:

$$s_x^2 = \frac{1}{n}\sum_{1}^{n}(x_i - \mu)^2 \qquad \{3.6\}$$

Again, μ is usually not known and it must be estimated by $\bar{x}$, so that the formula for s_x^2 becomes:

$$\boxed{s_x^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2} \qquad \{3.7\}$$

NOTE: The denominator becomes $(n-1)$. This denominator is called **DEGREES OF FREEDOM** (df). We have lost a degree of freedom since μ had to be estimated by $\bar{x}$. The numerator is called the SUM OF SQUARES, and the sample variance, s_x^2, is redefined as follows:

$$s_x^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2 = \frac{\text{Sum of Squares}}{df} \qquad \{3.8\}$$

NOTE:

1) Sum of Squares $= \sum_1^n (x_i - \bar{x})^2$ is usually calculated as: $\sum x^2 - \frac{(\sum x)^2}{n}$

2) $\sum x^2$ is sometimes referred to as the **`raw or uncorrected sum of squares'**

3) $\frac{(\sum x)^2}{n}$ is called the **`Correction factor'**.

(b) GROUPED DATA

We will use the same symbol s^2 to designate the sample variance for grouped data. Since we cannot reconstruct the actual sample measurements from the grouped data, we represent all values in a given class interval by the midpoint (class centre). Let m_i be the midpoint of the ith interval, and let f_i be the frequency for that interval. Then the sample variance for grouped data is given by:

$$s^2 = \frac{1}{n-1}\left\{\sum_1^k |f_i (m_i - \bar{m})^2|\right\} = \frac{1}{n-1}\left\{\sum m_i^2 f_i - \frac{(\sum m_i f_i)^2}{n}\right\} \qquad \{3.9\}$$

3.2.4 STANDARD DEVIATION

The STANDARD DEVIATION of any variate is defined as $\sqrt{\text{VARIANCE}}$

Thus σ is the population standard deviation (i.e., $\sigma = \sqrt{\sigma^2}$) , and s is the sample estimate of the standard deviation (i.e., $s = \sqrt{s^2}$).

3.2.5 SOME NOTES ON VARIANCE

a) **Effect of change of origin on Variance/Standard Deviation**

If a variate X undergoes a change of origin by having a constant A subtracted from each variate value then Variance is unchanged.

Let $Y = X \pm A$, then Variance(Y) = Variance(X), i.e., $s_y^2 = s_x^2$

and Standard Deviation (Y) = Standard Deviation(X), i.e, $s_y = s_x$

b) **Effect of change of scale on variance/standard deviation.**

Consider $Y = k X$

Then Variance (Y) $= k^2$ Variance (X); i.e., $\sigma_y^2 = k^2 \sigma_x^2$ (population)

or $s_y^2 = k^2 s_x^2$ (sample)

However, $\sigma_y = k \sigma_x$

and $s_y = k s_x$ (c.f. effect of change of scale on the mean)

3.2.6 LINEAR FUNCTIONS OF VARIATE VALUES

A linear function of variate values is one in which the variates appear as themselves or as multiples of themselves BUT NOT as any more complicated mathematical function of themselves.

Consider $$W = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \ \ldots \ + \lambda_n x_n = \sum^{n} \lambda_i x_i$$

where λ_i are constants, and the x_i are n distinct variates (or variate values)

Let $\sigma_1^2,\ \sigma_2^2,\ \sigma_3^2 \ \ldots \ \sigma_n^2$ be the variances of the x_i $(i = 1 \ldots n)$

Then Variance (W) $= \lambda_1^2$ Var $(x_1) + \lambda_2^2$ Var $(x_2) + \lambda_3^2$ Var $(x_3) + \ \ldots \ + \lambda_n^2$ Var (x_n)

i.e., $$\textbf{Variance (W)} = \lambda_1^2 \sigma_1^2 + \lambda_2^2 \ \sigma_2^2 + \lambda_3^2 \sigma_3^2 \ + \ \ldots \ + \lambda_n^2 \sigma_n^2 \qquad \{3.10\}$$

SPECIAL CASES:

a) Let $\sum^{n} x$ be the SUM of n variate values each with variance σ^2 then, since $\lambda_i = 1$ for all i,

$$\textbf{Variance} \left(\sum^{n} x\right) = n\,\sigma^2 \qquad \{3.11\}$$

b) Let $\bar{x}$ be the sample mean of n variate values x each with variance σ^2 then $\lambda_i = \frac{1}{n}$ for all i,

$$\textbf{Variance} \left(\bar{x}\right) = \frac{\sigma^2}{n} \qquad \{3.12\}$$

c) Let x_1 and x_2 be two independent random variates with variances σ_1^2 and σ_2^2 respectively then, since $\lambda_1 = 1$ and $\lambda_2 = -1$,

$$\textbf{Var} \left(x_1 \pm x_2\right) = \sigma_1^2 + \sigma_2^2 \qquad \{3.13\}$$

d) Consider two independent random samples size n_1 and n_2, with means $\bar{x}_1$ and $\bar{x}_2$ and variances σ_1^2 and σ_2^2 respectively then

$$\textbf{Var} \left(\bar{x}_1 \pm \bar{x}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \qquad \{3.14\}$$

e) Furthermore, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\textbf{Var} \left(\bar{x}_1 \pm \bar{x}_2\right) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \sigma^2 \left(\frac{n_1 + n_2}{n_1 n_2}\right) \qquad \{3.15\}$$

f) And if $n_1 = n_2 = n$,

$$\textbf{Var} \left(\bar{x}_1 \pm \bar{x}_2\right) = 2\,\frac{\sigma^2}{n} \qquad \{3.16\}$$

NOTE: IN PRACTICE σ^2 IS NOT KNOWN AND IT IS REPLACED BY ITS SAMPLE ESTIMATE, s^2

CHAPTER 4: PROBABILITY

4.1 MUTUALLY EXCLUSIVE EVENTS/EXHAUSTIVE EVENTS

(a) Two events are said to be MUTUALLY EXCLUSIVE if the occurrence of an event precludes the possibility of the occurrence of another, different event.

EXAMPLE :

If one flips a coin once and obtains *Heads*, this precludes the possibility that *Tails* can occur on that toss. Heads and Tails are thus mutually exclusive events.

(b) A set of mutually exclusive events that includes all permissible or possible outcomes is an EXHAUSTIVE set.

4.2 INDEPENDENT EVENTS

Two events are said to be INDEPENDENT if the occurrence of an event is completely unaffected by the occurrence of another, different event.

EXAMPLE

If one flips a coin once and simultaneously rolls a die once, the outcome of the coin toss experiment is unaffected by the outcome of the die–tossing experiment.

4.3 DEFINITIONS OF PROBABILITY

(a) **Classical definition**

If an event can occur in V equally likely and mutually exclusive ways and if U of these ways are favourable to the event, then the probability that the event will occur is $\frac{U}{V}$.

(b) Relative frequency definition

Let there be a series of n trials in which a certain event may occur by chance. Suppose the event occurs in m of those trials. The relative frequency of the event is $\frac{m}{n}$. This relative frequency is taken as an estimate of the true probability as defined in part (a). As n increases, this estimate will improve.

PROBABILITY LAWS

a) ADDITIVE LAW OF PROBABILITY

If two events A and B are **MUTUALLY EXCLUSIVE** then :

$$\mathbf{P\{A \text{ or } B\} = P\{A\} + P\{B\}} \qquad \{4.1\}$$

This Law may be extended for any number of mutually exclusive events.

(b) MULTIPLICATIVE LAW OF PROBABILITY

If two events A and B are **INDEPENDENT** then :

$$\mathbf{P\{A \text{ and } B\} = P\{A\} \times P\{B\}} \qquad \{4.2\}$$

As for (a), this Law may be extended for any number of independent events.

4.4 PROBABILITY DISTRIBUTIONS

The probability that any particular random variable should assume any given value, or range of values is determined by the PROBABILITY DISTRIBUTION for that random variable. Probability distributions are defined for continuous and discrete random variables.

4.4.1 BINOMIAL DISTRIBUTION

For many trials there are only two outcomes; a plant possesses a certain characteristic or it does not. A seed germinates or it fails to germinate. There are many examples of such trials in biological experimentation.

4.4.1.1 BINOMIAL PROBABILITY FUNCTION — (discrete variables)

The Binomial Probability function is used for any experiment consisting of N *independent trials* such that each of the outcomes falls into only one of two categories, denoted SUCCESS or FAILURE which have fixed probabilities, P and $(1 - P)$ of occurring respectively.

i.e. Probability (x = success) = P

Probability (x = failure) = $1 - p$ $(= q)$

$$\psi(x) = \textbf{Probability}\left\{\textbf{x success in n trials}\right\} = \binom{n}{x} p^x (1-p)^{n-x} \qquad \{4.3\}$$

The probability function $\psi(x)$ is defined for $x = 0, 1, 2 \ldots n$. Otherwise $\psi(x) = 0$

NOTE

i) $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is called the **Binomial coefficient;**
where, $n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$ and $0! = 1$

ii) The Binomial Distribution is completely defined by 2 parameters, viz., **n** (the sample size) and **p** (the probability of success).

iii) The probabilities of 0, 1, 2, 3, ... n successes are given respectively by the successive terms in the expansion of:

$(p + q)^n$ where $q = 1 - p$

iv) x is a discrete variable (i.e., it can only assume integer values), the Binomial Distribution is a discrete (discontinuous) distribution.

v) The population mean $\mu = np$, and population variance $\sigma^2 = npq$

vi) $\sigma^2 = \mu - \frac{\mu^2}{n}$, i.e, the mean μ and the variance σ^2 are not independent.

vii) The distribution is symmetric for $p = 0.5$, otherwise skew.

4.4.2 POISSON DISTRIBUTION

Mathematically the Poisson Distribution can be shown to be related to the Binomial Distribution with small P and large N such that the mean (μ) remains finite: the event is said to be 'rare'. However, the Poisson Distribution is a distribution in its own right and, for example, random sampling of organisms in some medium, insect counts in a field plot, noxious weed seeds in seed samples, numbers of various types of radiation particles emitted, etc., may yield data which follow a Poisson Distribution.

4.4.2.1 POISSON PROBABILITY FUNCTION – (discrete variables)

$$\psi(x = k) = e^{-\mu}\frac{\mu^k}{k!} \qquad \{4.4\}$$

i.e. Probability that a random variable x takes the value k, $(k = 0, 1, 2, 3 \ldots)$.

NOTE

a) Unlike the Binomial Distribution, theoretically the Poisson Distribution has no upper bound.
b) The Poisson Distribution is completely defined by one parameter, μ.
c) Mean = Variance, i.e., $\mu = \sigma^2$ (i.e., the mean and variance are not independent)
d) The distribution is J-shaped or positive skew.
e) Since k can take only positive integer values, the distribution is discrete (discontinuous).
f) In practice μ is not known and it is estimated by $\bar{x}$.

4.4.3 THE NORMAL DISTRIBUTION

Important in the theory and practice of statistics is the NORMAL DISTRIBUTION. Many biological phenomena result in data distributed in a manner sufficiently normal that the distribution is the basis of much of the statistical theory used by biologists.

4.4.3.1 NORMAL PROBABILITY FUNCTION – (continuous variables)

The Normal distribution is defined for a continuous random variable x, with probability function:

$$\psi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\} \qquad -\infty < x < \infty \qquad \{4.5\}$$

NOTE:

i) The normal distribution is completely defined by two parameters, μ and σ^2, the mean and variance respectively of the distribution.

ii) The mean, μ, is independent of the variance, σ^2.

iii) It is a symmetric, bell shaped curve.

iv) For the Normal distribution, mean = median = mode.

v) A normal variate with mean, μ, and variance, σ^2 is denoted by, $x \sim ND(\mu, \sigma^2)$

vi) Independent Normal variates with the same mean, μ, and variance, σ^2 are denoted by,

$$x \sim NID(\mu, \sigma^2)$$

4.4.3.2 STANDARDISED NORMAL VARIATE (z)

Consider $x \sim ND(\mu, \sigma^2)$ and let $z = \frac{(x-\mu)}{\sigma}$, then $z \sim ND(0, 1)$

z is said to be standardised normal variate, and the probability function for z is :

$$\psi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{z^2}{2} \right\} \qquad \infty < z < \infty \qquad \{4.6\}$$

Tables of $\Phi(z) = \int_z^\infty \psi(z)\, dz$ are available for positive values of z, (refer: Steel and Torrie, Table A4).

EXAMPLE: $z \sim ND(0, 1)$. What is the probability that z is less than 2.0 ?

$Prob(z < 2.0) = \{1 - Prob(z > 2.0)\} = 1 - \Phi(2.0)$. from tables: $\Phi(2) = 0.0228$,

i.e. $Prob(z < 2.0) = 0.9772$

EXAMPLE: Given that $x \sim ND(2,4)$, what is the probability that x lies between 1 and 3?

Firstly it is necessary to restate the problem in terms of z (since we have tables for z).

i.e. $Prob(1 < x < 3) = Prob(z_1 < z < z_2)$ where :

$z_1 = \frac{1-2}{2} = -0.5$ and $z_2 = \frac{3-2}{2} = 0.5$

i.e. $Prob(z_1 < z < z_2) = prob(-0.5 < z < 0.5)$

$= \{1 - \Phi(0.5)\} - \Phi(0.5)$

$= 1 - 2(0.3085) = 0.3830$

4.4.4 "STUDENTS" t–DISTRIBUTION

From (4.4.3.2) it is seen that for $x \sim ND(\mu,\sigma^2)$, the ratio $z = \frac{x-\mu}{\sigma}$ and $z = \frac{x-\mu}{\sigma_{\bar{x}}}$ have exact Normal distributions, provided the variance σ^2 is known. Otherwise these ratios are approximately ND if the numerator is ND and the denominator is based on a reasonable number of degrees of freedom, i.e., if the samples are 'large'.
For small samples however this approximation deteriorates until for very small samples the probabilities given by 'z–tables' will be seriously in error.

4.4.4.1 In most practical applications in which sample means are used to estimate population means, the value of σ^2 is not known, and it is necessary to obtain an estimate s^2 of σ^2 from the sample data that gives us $\bar{x}$.
"Student" (i.e., W S Gosset, 1908) showed that the ratios $\frac{x-\mu}{s}$ and $\frac{\bar{x}-\mu}{s_{\bar{x}}}$ have an exact distribution called the **t– distribution.**

The "Student" t–distribution applies to any ratio of a variate $ND(0,\sigma^2)$ to the square root of an estimate of variance of that variate, statistically independent of the numerator.

With regard to the t–distribution the following are to be noted:

a) If the sample is of size n, the estimate of variance is based on $(n - 1)$ degrees of freedom (df).
b) The distribution of t varies according to the number of degrees of freedom of the denominator.
c) The probability curve of the t–distribution is similar to that of the Normal distribution, i.e., it is symmetric, bell–shaped.
d) As the number of degrees of freedom increases; t–distribution → Normal distribution.

The following table illustrates the change in the values of t for changing DF ($\alpha = 0.025$) :

DF	5% value of t	
∞	1.960	– (c.f. z–value for Normal distribution)
120	1.980	
60	2.000	
30	2.042	
15	2.131	
5	2.571	
2	4.303	

NOTE: For 1–tail and 2–tail values of t, refer – Steel and Torrie, Table A3 .

4.5 SAMPLING DISTRIBUTIONS

Suppose a number of random samples, size n, are drawn from an infinite population $ND(\mu,\sigma^2)$. Each sample mean $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, $\bar{x}_4$ ... will vary one from another, thus generating in turn an infinite population of means, the probability distribution for $\bar{x}$. This is the SAMPLING DISTRIBUTION of $\bar{x}$.

NOTE

i) If $x \sim ND(\mu, \sigma^2)$ then it can be shown that $\bar{x} \sim ND(\mu, \frac{\sigma^2}{n})$, i.e., the distribution of the sample means is also Normal.
ii) From (1) it can be seen that as n get larger so the variance of the mean gets smaller, i.e., the sample mean, $\bar{x}$, becomes more precise (i.e., more reliable).
iii) Since Mean Value (MV) $x = \mu$ and $MV(\bar{x}) = \mu$ we say that $\bar{x}$ is an unbiased estimator of μ the population mean.

CHAPTER 5 : HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

5.1 HYPOTHESIS TESTING

Consider the situation where we have a single observation x of the yield of a new wheat variety. Suppose that the distribution of yield is Normal and that the standard deviation of this distribution is σ (i.e known). Suppose that we also know the yield of a standard variety under the conditions of the experiment to be μ.

THE HYPOTHESIS :

We wish to test the hypothesis that the new variety does not differ in average yield from the standard variety, i.e., we wish to test the reasonableness of the hypothesis that the mean of the distribution of yield of the new variety is also μ.

To obtain a measure of the reasonableness of this hypothesis *we assume the hypothesis to be true**, and calculate the probability of obtaining *by chance* a value of x at least as different from μ as that actually obtained. The reason for using this particular probability is that any result more extreme than x provides stronger evidence against the hypothesis than does x.

NOTE: This is known as the **NULL HYPOTHESIS**, i.e., the hypothesis of no effect, and it is designated as $\mathbf{H_0}$.

We can find the probability by calculating:

$z = \frac{x-\mu}{\sigma}$ and referring the result to tables of the standardised normal variate (z).

EXAMPLE

$\mu = 10.0$ tons/ha $\qquad \sigma = 1.0$

$x = 12.3$ tons/ha

thus $z = 2.3$

$P(z > 2.3) = 0.0107$

In this case we would say that the observed yield of 12.3 tons is (statistically) different from the hypothetical yield of 10 tons/ha at the 1.07% level.

5.2 LOGIC OF `tests of significance'

With reference to the previous example we may ask, can the observed deviation be reasonably accounted for by chance alone? This question may then be argued along the following lines.

Procedure:

Firstly form a test criterion, e.g. H_0 : $\{d = (x - \mu) = 0\}$

Then evaluate Prob (d), i.e., the probability (under H_0) of a deviation as great or greater than d being observed by chance. Our argument then proceeds as follows:

YES — Is P small? — NO

i.e the deviation appears abnormal
(**i.e.** ***significant)***

Reason ?

i.e., the deviation is not abnormal
(i.e. non–significant)
and the hypothesis, H_0 is **NOT** rejected.

Now if the observed deviation is ***significant*** one is confronted by one of two alternatives, viz:

Either : H_0 is true and the observed deviation has occurred by pure CHANCE!Or
Or : H_0 is untrue !

But, if H_0 is true, then we must ask ourselves, "Is it not a remarkable coincidence that our particular observed deviation is abnormal?". However, it must be remembered that the researcher is usually not doing the test in a vacuum, i.e., there are usually possible reasons why H_0 is untrue, and we may thus prefer to disbelieve the coincidence and reject H_0. In other words, if there is some unique special condition which could account for the significant deviation, it is adopted as the reason why H_0 is untrue.

5.3 REJECTION (Type I), and ACCEPTANCE (Type II) ERRORS

5.3.1 REJECTION (or Type I) ERRORS

H_0 : TRUE NOTE: $d = (x - \mu)$

Figure 5.1

Now any deviation $d \geq d_{05}$ is said to be significant at the 5% level of significance and H_0 is rejected, but FALSELY rejected since H_0 is TRUE. This false rejection of H_0 , when H_0 is TRUE, is called an ERROR TYPE I (or Rejection Error). It may be shown that:

$$\textbf{Prob(Type I error)} \leq \alpha \qquad \{5.1\}$$

corresponding to the (100α) % level of significance. Thus, for example with $\alpha = 0.05$, there is at most a 5% chance of wrongly rejecting H_0 if H_0 is true.

5.3.2 ACCEPTANCE (or Type II) ERRORS

H_0 : UNTRUE

Figure 5.2

In this case any deviation, $d \geq d_{05}$ is in zone BC, i.e., it would be considered significant and H_0 would be rejected (in this case correctly). Similarly, any deviation in zone DB would be considered non–significant and H_0 would not be rejected (i.e., wrongly not rejected, since H_0 is untrue).

This incorrect non–rejection of H_0 when H_0 is false (i.e., untrue) is called a TYPE II ERROR or Acceptance Error.

$$\textbf{Prob(Type II error)} = \beta \qquad \{5.2\}$$

The following is to be noted concerning β :

a) β is not known but it can be shown to lie between 0.95 and 0.0 when $\alpha = 0.05$.
Note: Given some alternative value of μ, μ_A say, it is possible to estimate β.

b) Prob (correct rejection of H_0) $= 1 - \beta$.
This is referred to as the ***Power of the Test*** and it will lie between 0.05 and 1.0 when α=0.05.

c) For $| \mu - \mu_0 |$ small, Prob (Type II) is high; it can be nearly 0.95 for $\alpha = 0.05$.

d) Prob (Type II error)
- decreases when SE ($\bar{x}$) decreases.
- decreases when $| \mu - \mu_0 |$ increases.
- increases when α decreases.

5.4 LEVELS OF SIGNIFICANCE

Convention :

NS.	–	non–significant
(*)	–	significant at 5% level.
(**)	–	significant at the 1% level

NOTE

1) `non–significant' does not imply non–existent.
2) `significant' implies `statistically significant' not necessarily real or meaningful.

5.4.1 TWO-TAIL TEST OF SIGNIFICANCE

H_0 : $\mu = \mu_0$

H_A : $\mu \neq \mu_0$ thus, *in conclusion*, there are two alternatives:

either $\mu > \mu_0$

or $\mu < \mu_0$

5.4.2 ONE-TAIL TEST OF SIGNIFICANCE

(a) H_0 : $\mu = \mu_0$

H_A : $\mu > \mu_0$ i.e., the researcher is interested in positive deviations only. (upper tail of the distribution)

(b) H_0 : $\mu = \mu_0$

H_A : $\mu < \mu_0$ i.e., the researcher is interested in negative deviations only. (lower tail of the distribution)

5.5 TESTS OF SIGNIFICANCE

5.5.1 The Normal test (z-test) – [c.f. Table A4, *Steel and Torrie*]

From the standardised Normal curve we know that :

a) $\text{Prob}(-1.96 < z < 1.96) = 0.95$

b) $\text{Prob}(-2.576 < z < 2.576) = 0.99$

} "**two–tail**" values **{ 5.3a }**

c) $\text{Prob}(|z| < 1.645) = 0.95$

$\text{Prob}(|z| < 2.326) = 0.99$

} "**one–tail**" values **{ 5.3b }**

Depending on the type of hypothesis under test, c.f. (5.4.1) and (5.4.2), the calculated value of z is then compared as follows:

(a) If $(-1.96 < z < 1.96)$ the result is not significant (i.e., NS.), and we ACCEPT H_0.

(b) If $(1.96 < |z| < 2.576)$ the result is significant, and we REJECT H_0 at the 5% level.

(c) If $(|z| > 2.576)$ the result is highly significant, and we REJECT H_0 at the 1% level.

NOTE: With regard to (b), we may elect to ask for further evidence before rejecting H_0.

CASE 1:
If x is ND(μ, σ^2) i.e., with known variance σ^2, we can use tables of standardised normal deviates (i.e., z) to decide whether an observed value of x is significantly different from some hypothetical value, μ_0 say.

Procedure:
a) State $H_0 : \mu = \mu_0$ thus $z = \frac{x-\mu_0}{\sigma}$ {5.4}

b) Calculate the probability that a value as extreme as x or more could have occurred by chance.

c) Decide whether the probability so calculated is small enough to reject H_0

[NOTE : in practice steps (b) and (c) are carried out together.]

CASE 2:
Consider a sample size n from a population ND (μ, σ^2), i.e., σ^2 is known.

Let $\bar{x}$ be the mean from this sample size n. Now $Var(\bar{x}) = \frac{\sigma^2}{n}$, i.e., $SE(\bar{x}) = \sigma/\sqrt{n}$

Furthermore consider, $H_0 : \mu = \mu_0$ then $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ {5.5}

Again, depending on whether a 1–tail or 2–tail test is under consideration the calculated value of z is referred to the values illustrated in {5.1a} or {5.1b} above.

NOTE This test is an exact test provided the parent distribution of x is exactly ND. On the other hand no assumption of normality is required for an approximate test provided the sample is large, (c.f. Central Limit Theorem).]

EXAMPLE 5.1 :
A group of 320 male students from the local campus have a mean mass of 70.41 kg and estimated variance of 36.05 kg^2. Records of similar American students show a mean mass of 70.0 kg. Assuming that mass is Normally distributed, are the local students different in mass?

Calculations:
$H_0 : \mu = 70.0$ kg $\bar{x} \sim ND(70, \frac{36.05}{320})$

Then $z = \frac{70.41-70.0}{\sqrt{36.05/320}} = 1.22$ (ns) c.f. z (5%) = 1.96

Conclusion: Since $z < 1.96$ we accept H_0

NOTE:
i) In this example we used a two tail test.
ii) Since n was large, we could use the 'approximate Normal' test.

CASE 3 :
Consider two samples of size n_1 and n_2 from populations $x_1 \sim ND(\mu_1, \sigma_1^2)$ and $x_2 \sim ND(\mu_2, \sigma_2^2)$ respectively. Let $\bar{x}_1$ and $\bar{x}_2$ be the sample estimates of μ_1 and μ_2 respectively.
Consider H_0: $\mu_1 = \mu_2$, alternatively, H_0: $\mu_1 - \mu_2 = 0$

Hence $$z = \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$ {5.6}

For an approximate test, as for CASE 2, no assumption of normality is required provided n_1 and n_2 are large. If σ_1^2 and σ_2^2 are not known the may be replaced by their sample estimates, s_1^2 and s_2^2 respectively.

EXAMPLE 5.2 :

An investigation to determine whether men are taller on average than women the following results were obtained. A group of 1164 men have a mean height of 174.46 cm with variance (s^2) 47.6522 cm^2. A similar group of 1456 women have mean height of 162.23 cm with variance (s^2) = 43.7625 cm^2.

$H_0 : \mu_1 = \mu_2$ $\quad$ $H_A : \mu_1 > \mu_2$ (i.e. a one tail test, since the question asks if men are taller.)

Applying equation {5.3} we get :

$$z = \frac{174.46 - 162.23}{\sqrt{\frac{47.6522}{1164} + \frac{43.7625}{1456}}} = 45.94^{**} \qquad \text{c.f. } z(1\%) = 2.326$$

Conclusion : Reject H_0 , the men are significantly taller than the women ($P < 0.01$).

NOTE

i) Again, as for Example 5.1, we have very large samples and the `approximate test' is used.

ii) In practice it is more likely for one to have relatively small samples in which case the approximation of the z–tests given above deteriorates until for very small samples the probabilities given by `z–tables' will be seriously in error.

5.5.2 The t — test

CASE 4: EXACT t-test

To test whether a sample estimate of a parameter is significantly different from a given hypothetical value, e.g. the sample mean $\bar{x}$.

As for CASE 3, consider a sample size n from a population ND (μ, σ^2), Let $\bar{x}$ be the mean from this sample. Let s^2 be the estimate of σ^2 with (n – 1) DF.

Now $Var(\bar{x}) = \frac{s^2}{n}$, and $SE(\bar{x}) = s/\sqrt{n}$

Consider, $H_0 : \mu = \mu_0$

We now calculate $$t_f = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad \{5.7\}$$

Where f is the degrees of freedom of the denominator of {5.7), in this case (n – 1).

If $t_{calc} > t^*$ or t^{**} for f DF, i.e the 5% of 1% significance values of t, we declare the result significant.

5.5.2.1 NOTES CONCERNING THE `t test' :

For an **EXACT TEST** the following assumptions are made:

i) x must be Normally distributed.

ii) Numerator and denominator must be independent.

THEOREM:

If we have k independent estimates of variance s_1^2, s_2^2, s_3^2 … s_k^2 of the same variance σ^2, with f_1, f_2 … f_k DF respectively, then the best combined estimate is given by:

$$s^2 = \frac{f_1 s_1^2 + f_2 s_2^2 + f_3 s_3^2 \ldots + f_k s_k^2}{f_1 + f_2 + \ldots + f_k} \qquad \{5.8\}$$

with $\sum f_i$ degrees of freedom.

NOTE

i) s^2 is called the `Pooled Estimate of Variance' and is often written s^2_{Pooled}

ii) If the f_i are all equal, $s^2 = \frac{1}{k}\sum s_i^2$

iii) Since $s_i^2 = \frac{\sum (x_i - \bar{x})^2}{f_i}$

We may write $$s^2 = \frac{SS(x_1) + SS(x_2) + SS(x_3) \ldots + SS(x_k)}{\sum n_i - k} \qquad \{5.9\}$$

CASE 5 :

To test whether 2 sample means are significantly different – **EXACT TEST**

(a) **TWO INDEPENDENT RANDOM SAMPLES** – (i.e. unpaired data)

Consider two independent random samples such that ; $x_1 \sim N.D(\mu_1; \sigma_2^2)$ and $x_2 \sim N.D(\mu_2; \sigma_2^2)$. For an exact test it is necessary to assume that the two independent random samples are from Normal populations with equal variances (even though the means may differ). On this assumption it is logical to form a 'pooled' estimate of the common variance from the estimates of variance obtained from the separate samples, viz s_1^2 and s_2^2 with f_1 and f_2 DF respectively,

i.e., $f_1 = (n_1 - 1)$ and $f_2 = (n_2 - 1)$.

Thus $$s^2_{pooled} = \frac{f_1 S_1^2 + f_2 S_2^2}{f_1 + f_2} \qquad \{5.9a\}$$

Now since $s_1^2 = \frac{SS(x_1)}{(n_1-1)}$ and $s_2^2 = \frac{SS(x_2)}{(n_2-1)}$

we have $$s^2_{pooled} = \frac{SS(x_1)+SS(x_2)}{(n_1+n_2-2)} \qquad \{5.9b\}$$

Now from 3.2.6 we were given that for independent samples, the variance of the difference of two means is given by:

$$Var(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{n_2\sigma_1^2 + n_1\sigma_2^2}{n_1 n_2} \qquad \{5.10a\}$$

Furthermore if $\sigma_1^2 = \sigma_2^2 = \sigma^2$. $$Var(\bar{x}_1 - \bar{x}_2) = \sigma^2 \left(\frac{n_1+n_2}{n_1 n_2}\right) \qquad \{5.10b\}$$

Also if $n_1 = n_2 = n$ $$Var(\bar{x}_1 - \bar{x}_2) = 2\frac{\sigma^2}{n} \qquad \{5.10c\}$$

Since σ^2 is rarely known it is replaced by its sample estimate s^2_{pooled} estimated according to formula {5.9a} or {5.9b}.

Usually the null hypothesis is stated as; $H_o: \mu_1 = \mu_2$, alternatively as; $H_o: \mu_1 - \mu_2 = 0$; where μ_1 and μ_2 are the true means of the respective populations.

Depending on the test under consideration the alternative hypothesis H_A can be either;

$H_A: \mu_1 \neq \mu_2$ (i.e., a 2 – tail test of significance);

or $H_A: \mu_1 > \mu_2$ or $H_A: \mu_1 < \mu_2$ } (i.e., **1 – tail tests of significance**)

In both cases the test statistic 't' is calculated as;

$$t_f = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{S.E(\bar{x}_1 - \bar{x}_2)} \qquad \{5.11\}$$

where $S.E(\bar{x}_1 - \bar{x}_2)$ is the square root of the variance calculated either by formula {5.8.1a}or {5.8.1b}. The degrees of freedom $f = (n_1 + n_2 - 2)$. $\bar{x}_1$ and $\bar{x}_2$ are the sample estimates of μ_1 and μ_2 respectively.

5.5.3 CONSEQUENCES OF NON-EQUALITY OF VARIANCE

It is to be noted that if we cannot make the assumption of equality of variances, i.e., we cannot assume that $\sigma_1^2 = \sigma_2^2$ then the t–test as given by formula {5.11} is only an **APPROXIMATE** test, with degrees of freedom calculated as a weighted average of f_1 and f_2, (c.f. Steel and Torrie, page 106;).

EXAMPLE 5.3 : (EQUAL SAMPLES)

Total yield of milk (lbs) during one lactation of cows on two different rations.

Ration A: 8671 6345 9325 7633 6649 7102 6517 8124 total = 60366

Ration B: 8954 7001 9514 7964 6758 7333 6510 8647 total = 62681

(c.f. A A Rayner, *Biometry for Agricultural Students, page 151 – 152*)

$H_0 : \mu_A = \mu_B$ $H_A : \mu_A \neq \mu_B$ (i.e a 2 tail test of significance.)

Calculations : $\bar{x}_A = 7545.7$ $SS(x_A) = 8275106$

$\bar{x}_B = 7835.1$ $SS(x_B) = 8610301$

thus $s^2_{pooled} = \frac{8275106+8610301}{(8+8-2)} = 1206100$ – c.f. formula {5.9b}

and $S.E(\bar{x}_A - \bar{x}_B) = \sqrt{2\,\frac{1206100}{8}} = 549.1$ – c.f. formula {5.10b}

$t_{14} = \frac{289.4}{549.1} = 0.5$ **(NS)** – **c.f. t_{14} (5%) = 2.145**

Conclusion:
The difference in mean yield per cow is 289.4 ± 549.1 lbs in favour of Ration B, but we cannot conclude that Ration B is superior.

EXAMPLE 5.4 (UNEQUAL SAMPLES):

The following are the gains in weight of two groups of female rats under different diets.

Ration	No. of rats	Gains in grams
High protein	12	134 146 104 119 124 161 107 83 113 129 97 123
Low protein	7	70 118 101 85 107 132 94

[c.f. G W Snedecor, *Statistical Methods, 5th edition (1956)*]

$H_0 : \mu_{high} = \mu_{low}$ $H_A : \mu_{high} \neq \mu_{low}$ (i.e a 2 tail test of significance.)

Calculations:

$\bar{x}_{high} = 120.0$ $SS(x_{high}) = 5032$

$\bar{x}_{low} = 101.0$ $SS(x_{low}) = 2552$

thus $s^2_{pooled} = \frac{5032+2552}{(12+7-2)} = 446.12$ – c.f. formula {5.9b}

and $S.E(\bar{x}_A - \bar{x}_B) = \sqrt{446.12(\frac{12+7}{12 \times 7})} = 10.05$ – c.f. formula {5.10a}

$t_{17} = \frac{19.0}{10.05} = 1.89$ (NS) – c.f. t_{17} (5%) = 2.110

Conclusion:
The mean difference of 19 ± 10.05 grams in favour of the high protein diet is not quite significant at the 5% level.

(b) **PAIRED DATA (i.e., non–independent samples).**

The device of `pairing' is used quite often in experimental work and examples can be found in practically all areas of research. For example, an animal researcher may choose to use a number of pairs of twins when comparing two different feeds. A plant pathologist may subject a number of plants to both sets of conditions under test to evaluate a response. An agricultural economist, in order to evaluate the effect of a proposed rural upliftment program may choose to interview a number of randomly chosen smallholders before and after the introduction of a pilot program to measure their reaction. In all cases the idea is that the `pairs', when allotted different treatments, serve better as controls for one another, i.e., the observed differences between the members of any pair will represent treatment differences which will be subject to less error than if there was no pairing. Thus the purpose behind the device of "pairing" in experimental work is to remove from the estimate of error variance variability due to differences between pairs. The information on pairing is obtained by calculating the variance of the differences rather than that among the individuals within each sample. The number of degrees of freedom in this case is based on the **(number of pairs − 1)**.

EXAMPLE 5.5 : (PAIRED DATA)

Suppose the cows in Example 5.3 comprised 8 sets of twins, one twin from each set being allocated at random to Ration A, and the other to Ration B. In this case the two samples are NOT INDEPENDENT but PAIRED. It is now necessary for us to calculate the differences between pairs of observations.

Pair No.	Difference $d_i = (B - A)$
1	$(8954 - 8671) = 283$
2	$(7001 - 6345) = 656$
3	$(9514 - 9325) = 189$
4	$(7964 - 7633) = 331$
5	$(6758 - 6649) = 109$
6	$(7333 - 7102) = 231$
7	$(6510 - 6517) = -7$
8	$(8647 - 8124) = 523$
total	**2315**

Calculations:

$\bar{d} = 289.4$

S.S.(differences) = 994 527 − 669 903 = 324 624 $\qquad S^2_{differences} = \frac{324\,624}{7} = 46\,374.8$

Estimated variance of mean of difference $= \frac{46\,374.8}{8} = 5\,796.9$

thus S.E.(mean difference) = 76.14

$H_0 : \mu_A = \mu_B \qquad H_A : \mu_A \neq \mu_B$ (i.e a 2 tail test of significance.)

$t_7 = \frac{289.4}{76.14} = 3.80^{**}$ $\qquad$ c.f. $t_7(1\%) = 3.499$

Conclusion:

Ration B is significantly superior to Ration A ($p < 0.01$), the difference in favour of Ration B being estimated as 289.4 ± 76.1 lb. milk per cow.

NOTE:

The use of "paired samples" has greatly improved the precision of the test as evidenced by the reduction in the S.E $(\bar{x}_A - \bar{x}_B)$, viz : $\pm$ 549.1 lbs for "unpaired data(i.e., independent random samples) ", and $\pm$ 76.14 lbs for "paired data".

5.6 DETERMINATION OF SAMPLE SIZE REQUIRED FOR SIGNIFICANCE

When designing an experiment a question which should be uppermost in the mind of the researcher is that of precision. Associated with this question is that of the number of samples (replications) which should be considered to achieve that precision. Now if an estimate of s^2 is available (e.g from a previous experiment), the precision of the mean, for example, as measured by SE($\bar{x}$), can be reduced to any level by increasing n. Another way of looking at this is to consider a specified difference, $d = (\bar{x} - \mu_0)$ say, which we would like to judge as significant (statistically) should it occur, i.e we need:

$$\frac{d}{\sqrt{s^2/n}} > t_\alpha \qquad \text{and we can then solve for n,} \qquad \text{i.e.,} \qquad \mathbf{n \geq \frac{t^2 s^2}{d^2}} \qquad \textbf{(5.11)}$$

EXAMPLE 5.6:
Let us reconsider the "unpaired" data of EXAMPLE 5.3., and ask the question, "how many cows must be placed on each ration in order to judge the observed mean difference of 289.4 lbs significant at the 5% level?".
Remember, $s^2 = 1206100$ and $t_{14,\alpha=0.05} = 2.145$

Then $n > \frac{(2 \times 1206100) \times 2.145^2}{289.4^2}$ i.e. $n > 132.5$

With regard to the above, the several approximations must be noted, in that the value of s^2 obtained in future experiments will vary from the value used in the formula. A second problem concerns the DF for t. It is usual to use the DF associated with those of s^2 originally obtained, but if the new sample is larger than the old we are acting conservatively.

5.7 CONFIDENCE INTERVALS

The main object of an experiment is usually to estimate certain quantities, e.g. yield of maize per hectare, volume of timber, species composition in a botanical analysis, etc. These estimates are usually obtained from a number of experimental plots (areas), treated alike, in which case the obvious estimate of yield is the sample mean, $\bar{x}$. An estimate of this kind is called a `point estimate', i.e., $\bar{x}$ is a point estimate of the true mean μ.

With regard to `point estimates' the following is to be noted :

a) If more and more plots are included in the estimation of the mean we would eventually get the population mean μ.
b) With a small number of observations the estimate is bound to deviate from the true population mean.

The question is now asked, can an interval be found within which the true mean μ will lie with a stated degree of confidence?

5.7.1 CONFIDENCE INTERVALS – based on the Normal deviate (z), i.e., σ^2 known.

From $z = \frac{x-\mu}{\sigma}$; where $x \sim ND(\mu,\sigma^2)$

$$\text{Prob}(-1.96 < z < 1.96) = 0.95$$

i.e. $x - \mu = \pm 1.96\,\sigma$
or $\mu = x \pm 1.96\,\sigma$

Thus given x, an estimate of μ, and σ , **the standard deviation, known:**

$$\underline{\mu = x \pm 1.96\,\sigma} \qquad \{5.12a\}$$

is the 95% Confidence Interval within which the true mean μ will lie

Similarly $\underline{\mu = x \pm 2.576\,\sigma}$ {5.12b}
will give the **99% Confidence Interval** for μ.

Furthermore if we have a random sample size n from a population $ND(\mu,\sigma^2)$ we know that $\bar{x}$ is also $ND(\mu, \frac{\sigma^2}{n})$, thus given $\bar{x}$, an estimate of μ, and σ, the (standard deviation) known :

$$\underline{\mu = \bar{x} \pm 1.96\sqrt{\frac{\sigma^2}{n}}} \qquad \{5.12c\}$$

is the **95% Confidence Interval** within which the true mean μ will lie. For the 99% confidence interval we substitute the z-value **2.576** .

5.7.2 CONFIDENCE INTERVALS — using t-values (i.e., σ^2 unknown)

When σ^2 is unknown we use s^2, its sample estimate, and in this case t_α values will replace the z_α in formulae{5.11a), {5.11b} and {5.11c} above, i.e.,

$$\mu = \bar{x} \pm t_\alpha s \qquad \ldots \{5.12a\}$$

is the **100(1 − α)% Confidence Interval** within which the true mean μ will lie.

Similarly, if we have a random sample size n from a population ND(0,σ^2) we know that $\bar{x}$ is also ND(0, $\frac{\sigma^2}{n}$), thus given $\bar{x}$, an estimate of μ, and s^2, the estimate of the population variance ,σ^2 :

$$\mu = \bar{x} \pm t_\alpha \sqrt{\frac{s^2}{n}} \qquad \ldots \{5.12b\}$$

is the (1 − α)% Confidence Interval within which the true mean μ will lie.

Generally one may consider the following for the estimation of Confidence Intervals/Limits for a particular PARAMETER (assuming that such a parameter is Normally distributed) :

$$\textbf{PARAMETER} = \textbf{statistic} \pm t_\alpha \textbf{ S.E.(statistic)}$$

where for example, the parameter may be μ, β, DIFFERENCE, Y, etc., with their corresponding statistics: $\bar{x}$, $\hat{b}$, $(x_1 - x_2)$, $\hat{y}$, etc.

5.8 NORMAL APPROXIMATION FOR BINOMIAL VARIATES

5.8.1 When n is large the evaluation of the probability function (usually n > 30)

$\psi(x)$ = Probability (x = success in n trials) = $\binom{n}{x} p^x (1-p)^{n-x}$ is cumbersome for large N.

However it may be shown that provided certain conditions are met one may use the assumption of **Normality** to evaluate such problems, i.e., for the Binomial variate $x \sim (n,p)$, we make the approximation that,

$$P\{X \leq x\} \doteq P\left\{ z \leq \frac{x-np}{\sqrt{npq}} \right\} \qquad \text{i.e., } x \sim N.D(np, npq)$$

This approximation is useful under the following conditions:

a) n must be large , i.e > 30

b) p should be "reasonably close" to 0.5.

[In this regard it may be further noted that with increasing values of n, the "normal approximation" can provide reasonably accurate results for p in the range, {0.3 < p < 0.7}.]

c) $np \geq 5$, i.e., no *expected* value should be less than 5.

d) The variance is sufficiently large, i.e., $np(1-p) \geq 3$, say.

EXAMPLE

Consider the tossing of an unbiased coin 49 times, i.e., n = 49, p = 0.5. What is the probability of obtaining fewer than 29 heads?

(a) Exact solution: $P\{x \leq 28 \text{ heads}\} = \sum_{i=0}^{28} \binom{n}{i} P^i Q^{n-i}$

$$= \mathbf{0.8736}$$

(b) Approximate solution: $\mu = NP = 24.5$ $\sigma^2 = NPQ = 12.25$ i.e., $\sigma = 3.5$

$$P\{x \leq 28 \text{ heads}\} \doteq P\left\{ z \leq \frac{28-24.5}{3.5} \right\} = P\{z \leq 1.0\} \doteq \mathbf{0.8413}$$

This approximate result can be further improved by "correcting for continuity" as follows:

$P\{x \leq 28.5 \text{ heads}\} \doteq P\left\{z \leq \frac{28.5-24.5}{3.5}\right\} = P\{z \leq 1.143\} = 0.8735$ (a very close result!)

This approximation may be extended further for $\hat{p}$ the estimate of P the true proportion.

Consider a sample size n, and let x have some attribute A, say. Then $\hat{p} = \frac{x}{n}$, is the estimate of the proportion in the sample with attribute A. Furthermore, variance(p) $= \frac{pq}{n}$ Then, subject to the conditions given above we may assume that $\hat{p} \sim N.D(P, \frac{pq}{n})$

i.e., $\text{Prob}\left\{\hat{p} \leq P_0\right\} \doteq \text{Prob}\left\{ z \leq \frac{\hat{p}-P_0}{\sqrt{\frac{pq}{n}}} \right\}$

EXAMPLE

A Market Researcher is interested in consumer preference for either products A or B. A sample of 80 people showed that 54 preferred Product A. What conclusion can the Market Researcher derive from these results?

$H_0 : P_0 = 0.5$ $\hat{p} = \frac{54}{80} = 0.675$ $S.E.(\hat{p}) = \sqrt{\frac{54}{80} \times \frac{26}{80} \times \frac{1}{80}} = 0.052366$

$z \doteq \frac{0.675-0.5}{0.052366} = 3.342^{**}$ refer: $z_{5\%} = 1.96$; $z_{1\%} = 2.576$

Conclusion:

Reject H_0 at the 1% level of significance. There is very strong evidence of preference in favour of Product A.

5.9 APPROXIMATE BINOMIAL PROBABILITIES USING POISSON DISTRIBUTION

In other instances where for example the binomial event is rare (i.e. P is very small, but N is large), the use of the Poisson distribution to approximate the Binomial probability may be appropriate.

CHAPTER 6 : ANALYSIS OF VARIANCE

6.0 INTRODUCTION

Many experiments are conducted in which the effects of two or more treatments are investigated. Usually the aim of such experiments is to make inferences concerning the treatment means.

In Chapter 5 we dealt with *tests of significance* concerning the sample estimate of a single treatment mean $\bar{x}$ in relation to some hypothesised mean μ, our null hypothesis being $H_o : \mu = \mu_o$. These tests were extended to test for differences between two sample means, either independent samples or non–independent samples (paired data), our null hypothesis being $H_o : \mu_1 = \mu_2$. When the number of treatments exceeds two a method known as the "Analysis of Variance" is introduced.

6.1 TO TEST WHETHER TWO ESTIMATES OF VARIANCE ARE SIGNIFICANTLY DIFFERENT

Suppose we have two independent estimates of variance S_1^2 and S_2^2 with f_1 and f_2 degrees of freedom respectively. In order to test whether they are significantly different we must assume that S_1^2 and S_2^2 are independent estimates from Normal samples with the same variance. Now the ratio $S^2_{1S_2^2}$ has a known distribution, the *"F–distribution"* (R.A. Fisher 1923), and a test of significance based on the F–distribution is known as a **F–test**.

The probability curve of the F–distribution varies with f_1 and f_2, the degrees of freedom of the numerator and denominator estimates of variance respectively. Consequently tables of F are double–entry tables and are thus more restricted than t–tables. For convenience the numerator in the ratio S^21/S_2^2 must always be the larger of the two estimates of variance; i.e., we will only be concerned with ratios ≥ 1, (c.f. *Steel and Torrie*, Table A6).

EXAMPLE 6.1 (a):

In Example 5.3 we made the assumption that the variances were equal. Let us see if the assumption was justified.

$H_o : \sigma_1^2 = \sigma_2^2 \qquad H_A : \sigma_1^2 \neq \sigma_2^2$

$SS(x_1) = 8275106 \quad (n_1 = 8) \qquad$ thus $s_1^2 = 1182158 \quad (df = 7)$

$SS(x_2) = 8610301 \quad (n_2 = 8) \qquad$ and $s_2^2 = 1230043 \quad (df = 7)$

$F_{7,7} = \frac{1230043}{1182158} = 1.041$ (n.s) from tables, $F_{7,7}(5\%) = 4.99$

$F_{7,7}(1\%) = 8.89$ (2–tail test)

Conclusion: Since $F_{calculated}$ was less than $F_{(\alpha=0.05)}$ we have no reason to reject H_o .

NOTE:

a) From *F* – tables (c.f. Tables A6 – *Steel and Torrie*), the following values are obtained for $F_{7,7}$

$F_{7,7}(\alpha = 0.05)$	$= 3.79$	(1–tail test ; 5%)
$F_{7,7}(\alpha = 0.025)$	$= 4.99$	(2–tail test ; 5%)
$F_{7,7}(\alpha = 0.01)$	$= 6.99$	(1–tail test ; 1%)
$F_{7,7}(\alpha = 0.005)$	$= 8.89$	(2–tail test ; 1%)

b) Since variance s_2^2 was the larger of the two variances it is the numerator.

EXAMPLE 6.1 (b)

It is hypothesised that there is much greater variation in the heights of men than women for a particular population. From random samples or 61 men and 30 women the following estimates were obtained:

$s^2_{males} = 51.652$ $(df_1 = 60) \qquad s^2_{females} = 43.762$ $(df_2 = 29)$

$H_o : \sigma^2_{males} = \sigma^2_{females} \qquad H_A : \sigma^2_{males} > \sigma^2_{females}$ (i.e. a 1-tail test)

$F_{60,29} = \frac{51.652}{43.763} = 1.181$ (n.s) from tables: $F_{60,29} = 1.75$ and $F_{60,29} = 2.23$ (1–tail test)

Conclusion: Since $F_{calculated}$ was less than $F_{(\alpha=0.05)}$ we have no reason to reject H_o .

6.2 REPLICATION AND RANDOMISATION

6.2.1 REPLICATION

Replication is the repetition of treatments **within the same experiment.** The number of replications is denoted by **r.**

NOTE:

i) When r = 1, the experiment is said to be unreplicated.

ii) The different treatments under consideration in an experiment are not necessarily equally replicated. In such instances the number of replications for the i^{th} treatment will be denoted by $\mathbf{r_i}$.

6.2.2 THE ROLE OF REPLICATION

a) Replication permits the estimation of error variance (σ^2).

b) Increases in r, increases the precision of parameter estimation.
c.f. Var ($\bar{x}$) = $\frac{\sigma^2}{r}$, thus reduced S.E.'s for treatment means

c) More df available for the estimation of σ^2

NOTE

The more df available for the estimation of error (σ^2) does not imply a reduced error mean square, despite the formula, Error Mean Square = $\frac{\text{Error SS}}{\text{df}}$, since MV(s^2) = σ^2, irrespective of the number of df available. However, Variance (s^2) = $\frac{2\sigma^4}{\text{error df}}$, hence the more error df, the more precise the estimate of σ^2, and the smaller the value of F (or t) required for significance.

d) The *efficiency* of an experiment is determined by the SE of a parameter estimate, e.g., a treatment mean, and the df available for the estimation of the error mean square. Both are improved by increasing r.

6.2.3 RANDOMISATION

Randomisation is the allocation *by chance* so that each treatment has exactly an equal chance of being allocated to any 'plot' – (c.f. R A Fisher 1925).

6.2.4 THE ROLE OF AN EXPERIMENT

Any experiment should provide:

a) Unbiased estimates of treatment effects (differences).
NOTE : This is ensured by the unbiased/random allocation of treatments to plots.

b) Unbiased estimate of the experimental error to which the estimates of treatment differences are subject.
NOTE : Replication makes such an estimate possible.

6.3 STATISTICAL MODEL FOR A RANDOM VARIATE

A random variate (from an infinite population) deviates randomly from its true mean μ. This may be represented symbolically as:

$$y_i = \mu + \epsilon_i$$

μ is a constant, i.e the same for all variate values in the population.
ϵ_i is a variable component, i.e., varies randomly from variate value to variate value in accordance with the probability function of y.

NOTE:

i) MV(y) = μ MV(ϵ) = 0

ii) ϵ_i is referred to as the random "error" component of y_i.

iii) If y is N.D(μ,σ^2), then ϵ is N.D(0,σ^2)

6.4 THE ANALYSIS OF VARIANCE FOR ANY NUMBER OF TREATMENT GROUPS

6.4.1 PARTITIONING OF THE SUM OF SQUARES

Let us consider an experiment designed to test for differences between **t** treatment means; i.e we have t samples. Furthermore suppose that each sample is of size **r**; i.e., the number of **replications** of each treatment within the experiment are **equal.**

NOTE:

As we shall see later, it is not necessary to have equal replications of each treatment in an experiment but with equal replication treatments can be compared with equal precision.

Consider the following where y_{ij} represents the j^{th} observation of the i^{th} treatment; y_{io} the i^{th} treatment mean and $\overline{y}$ the overall (or general) mean of the experiment.

Treatment 1	Treatment 2	.....	Treatment t	
y_{11}	y_{21}		y_{t1}	
y_{12}	y_{22}		y_{t2}	
y_{13}	y_{23}		y_{t3}	
:	:		:	
:	:		:	
:	:		:	
:	:		:	
y_{1r}	y_{2r}		y_{tr}	
means y_{1o}	y_{2o}	.....	y_{to}	$\overline{y}$
totals Y_{1o}	Y_{2o}	.....	Y_{to}	**GT** (i.e., Grand Total = $\sum Y_{io}$)

If we now regard the above data as consisting of a single sample then the sum of squares (SS) for this sample, called the **TOTAL SUM of SQUARES (TOTAL SS)** , is given by,

$$\textbf{Total SS} = \sum \sum (y_{ij} - \overline{y})^2 \qquad \ldots \{6.1\}$$

This SS, divided by the appropriate degrees of freedom, will provide an estimate of variance of all the data. However closer examination of the above data set will reveal two identifiable sources of variation; that **"Between Treatment Means"** and that **"Within Treatments"**.

This partitioning of the Total SS may be represented algebraically as follows:

$$\sum_{i}^{t}\sum_{j}^{r}(y_{ij} - \overline{y})^2 \;=\; r\sum_{i}^{t}(y_{io} - \overline{y})^2 \;+\; \sum_{i}^{t}\sum_{j}^{r}(y_{ij} - y_{io})^2 \qquad \ldots \{6.2\}$$

i.e. Total SS = Between Treatment SS + Within Treatment SS

The "Between Treatment SS" (usually referred to as the Treatment SS) is a measure of the extent to which the sample means deviate about the general mean $\overline{y}$. While the "Within Treatment SS" is actually a pooling of the variation between observations within each treatment and is an extension of the numerator of formula {5.7} in section 5.5.2. used in the estimation of the "pooled estimate of variance". In other words, provided we can assume that we have independent random samples from infinite populations **with the same variance**, the *Within Treatment SS* will, after division by appropriate degrees of freedom, provide us with s^2, our estimate of σ^2, our measure of the "error" within the experiment. The Within Samples SS is often referred to as the Error SS or Residual SS. If we extend this for t samples (i.e., t treatments) we get,

$$S^2_{pooled} = \frac{SS(y_1) + SS(y_2) + SS(y_3) + \ldots + SS(y_t)}{(r-1)+(r-1)\ldots \text{t-terms}} \qquad \{6.3\}$$

i.e. $$S^2_{pooled} = \frac{SS(y_1)+SS(y_2)+SS(y_3)+\ldots+SS(y_t)}{t(r-1)} \qquad \{6.3a\}$$

NOTE:

It is well to remind ourselves that in any statistical analysis the use of the term **"error"** usually refers to ***"that variation over which the experimenter has no direct control".***

In practice the Total SS, Treatment SS and Error SS (Within Samples SS) are calculated as follow.

$$\textbf{Total SS} = \sum_{i}^{t}\sum_{j}^{r} y_{ij}^2 - \frac{G.T^2}{n} \qquad \ldots \{6.4\}$$

where **n** is the total number of observations in the experiment and G.T is the total of all observations, i.e. $G.T = \sum_{i}^{t}\sum_{j}^{r} y_{ij}$; and $\frac{G.T^2}{n}$ is the "Correction Factor" (C.F.).
Let Y_{io} be the total of all observation for the i^{th} treatment then;

$$\textbf{Treatment SS} = \frac{1}{r}\sum_{i}^{t} Y_{io}^2 - \textbf{C.F.} \qquad \ldots \{6.5a\}$$

In the event that there are an unequal number of observation per treatment (sample) the Treatment SS is calculated as follows: –

$$\textbf{Treatment SS} = \sum_{i}^{t} \frac{Y_{io}^2}{r_i} - \textbf{C.F.} \qquad \ldots \{6.5b\}$$

where r_i is the number of observations in the i^{th} sample (treatment).

NOTE : In formulae {6.5a} and {6.5b} Y_{io} is the TOTAL for the i^{th} treatment.

When $t = 2$, i.e., there are only two treatments (r equal) with treatment totals represented by T_1 and T_2 respectively, then

$$\textbf{Treatment SS} = \frac{(T_1 - T_2)^2}{2r} \qquad \ldots \{6.5c\}$$

With r unequal we have :

$$\textbf{Treatment SS} = \frac{(r_2T_1 - r_1T_2)^2}{r_1 + r_2} \qquad \ldots \{6.5d\}$$

Error SS is calculated by subtraction as (Total SS – Treatment SS). ... {6.6}

6.4.2 THE ANALYSIS OF VARIANCE TABLE (ANOVA)

The partitioning of the Total SS can be conveniently summarised in an Analysis of Variance Table as follows: –

Table 6.1: Analysis of Variance (equal samples)

Source of Variation	df	SS	MS
Between Treatments	$(t-1)$	$\frac{1}{r}\sum_{i}^{t} Y_{io}^2 - C.F.$	s_1^2
Within Treatments (Error)	$t(r-1)$	(by subtraction)	s^2
Total	$tr-1$	$\sum_{i}^{t}\sum_{j}^{r} y_{ij}^2 - \frac{G.T^2}{n}$	s_2^2

NOTE: (MS) = Mean Square = $\frac{SS}{df}$; i.e., an estimate of variance.
Although, as shown, we may obtain three estimates of variance, the Total MS (S_2^2) plays no part in the analysis.
Important features of the above table include:

a) the sums of squares (SS) are additive,
b) the degrees of freedom (df) are additive,
c) the Mean Square (MS) column is not additive,
d) the Error SS is usually calculated by subtraction
   i.e., Error S.S. = Total S.S. – Treatment S.S..

6.5 STATISTICAL MODEL — (Fixed and Random effects MODELS)

6.5.1 FIXED EFFECTS MODEL

The statistical model for the analysis of variance (1-way analysis) may be taken as :

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad \textbf{MODEL 1}$$

where $\mu + \tau_i$ represents the mean of the i^{th} sample (treatment), ϵ_{ij} is a random component N.I.D(0, σ^2), irrespective of i. In order that we have a unique solution for τ_i it is necessary to impose a restriction on the τ_i and this is usually taken as $\sum \tau_i = 0$, at least in the case of equal samples.

Now in many types of experiments we are dealing with a certain fixed set of treatments effects, i.e., we are only concerned with a particular set of treatments and their effects. In such circumstances the τ_i are known as **fixed (treatment) effects**, and **Model 1** is known as the **Fixed Effects Model.** Model 1 is of particular significance in most types of agricultural experimentation.

Now it may be shown that the mean (expected) value of the "Between Samples Mean Square" under Model 1 is:

$$\sigma^2 + \left\{ \sum n_i(\tau_i - \bar{\tau})^2/(t-1) \right\}$$

with a slight modification if $n_i = r$.

In other words: $$s_1^2 = \sigma^2 + \left\{ \sum n_i(\tau_i - \bar{\tau})^2/(t-1) \right\}$$ (reference Table 6.1)

6.5.2 RANDOM EFFECTS MODEL

In many other situations, the researcher may be concerned with the possible extra variability between observed sample means imposed on the random variance (σ^2) by an additional component of variance due to differences between samples, i.e., we are concerned with the comparisons of two *variances* to see whether the variances of the sample means is greater than that expected for the means of independent random samples from a population with variance σ^2. In this situation the treatment effects cannot be treated as fixed, but instead constitute a random sample from some hypothetical population of effects. Model 1 must thus be amended to accommodate this change in assumptions.

$$y_{ij} = \mu + \eta_i + \epsilon_{ij} \qquad \textbf{MODEL 2}$$

ϵ_{ij} is as before, but the η_{ij} replacing τ_{ij}, are also random components distributed independently of one another and also the ϵ_{ij}. We may arrange μ such that MV(η_i) = 0. Let Var(η_i) = σ_T^2, i.e., the component of variance due to treatments. the η_i are known as **random (treatment) effects.**

Model 2 is known as the **RANDOM EFFECTS MODEL.** Model 2 type situations are most prevalent in many areas of Genetics, and Plant and Animal Breeding.

Now for the Random Effects Model, it may be shown that the mean (expected) value of the "Between Samples Mean Square" under Model 2 is:

$$\sigma^2 + r\sigma_T^2$$

With unequal samples r is replaced by an average of the r_i, not the arithmetic mean, although the latter provides a good approximation if the r_i do not differ markedly. {Ref: Steel and Torie : Page 151, formula (7.12) }

In other words for a Random Effects model (Model 2):

$$s_1^2 = \sigma^2 + r\sigma_T^2$$ (reference Table 6.1)

Whatever the model, it is apparent that on the null hypothesis, for Model 1, $\tau_i = 0$: for Model 2, $\sigma_T^2 = 0$, the expected value of both the Between samples and Within samples M.S.'s is σ^2 and that consequently their ratio follows an F-distribution. It is further apparent that if the null hypothesis is untrue, the Between samples M.S has a mean value greater than σ^2 irrespective of the models since σ_T^2 and $\sum n_i(\tau_i - \bar{\tau})^2$ are both positive quantities.

6.5.3 THE F-test

One of the objectives of the Analysis of Variance is to test hypotheses concerning the treatment means. The simplest hypothesis is,

$$\mathbf{H_o: \mu_i = \mu \text{ for all i, i.e } \mu_1 = \mu_2 = \ldots = \mu_t = \mu}$$

Now under H_o , S^2 and S_1^2 are independent estimates of σ^2 (c.f. Table 6.1 above). However if H_o is untrue then $\mu_i \neq \mu$ for all i, i.e., there are significant differences between some (if not all) of the treatment means, and S_1^2 in the above ANOVA table is expected to be greater than S^2. Thus if we have t independent samples from Normal populations it follows that the ratio,

$$\frac{S_1^2}{S^2} = \frac{\text{Treatment MS}}{\text{Error MS}} \sim F_{(t-1),\, t(r-1)} \qquad \ldots\{6.7\}$$

i.e. a test of H_o is obtained from a single–tail F – test.

Furthermore, provided the t populations have a common variance, σ^2, estimated by s^2, then the F – test is more efficient than a series of t-tests to test for significant differences between successive pairs of means since s^2 is based on a greater number of degrees of freedom. Table 6.1 above may now be extended to include this the F test statistic as follows:

Table 6.2: ANALYSIS OF VARIANCE(ANOVA)

Source of Variation	df	SS	MS	F
BetweenTreatments	(t – 1)	$\frac{1}{r}\sum_i^t y_{io}^2 - \frac{G.T^2}{n}$	s_1^2	$\frac{s_1^2}{s^2}$
Within Treatments (Error)	t(r – 1)	(by subtraction)	s^2	
Total	tr – 1	$\sum_i^t \sum_j^r y_{ij}^2 - \frac{G.T^2}{n}$	s_2^2	

The value of $\frac{s_1^2}{s^2}$ is referred to F-tables for (t – 1) and t(r – 1) degrees of freedom.

EXAMPLE 6.1 : (Reference: Table 7.1 *Steel and Torrie, page 140*)

The following are the nitrogen content of red clover plants inoculated with combination cultures of *Rhizobium trifoli* strains and *Rhizobium melitoti*, – milligrams.

	Strains: 3DOK1	3DOK5	3DOK4	3DOK7	3DOK13	Composite	
	19.4	17.7	17.0	20.7	14.3	17.3	
	32.6	24.8	19.4	21.0	14.4	19.4	
	27.0	27.9	9.1	20.5	11.8	19.1	
	32.1	25.2	11.9	18.8	11.6	16.9	
	33.0	24.3	15.8	18.6	14.2	20.8	
Totals	144.1	119.9	73.2	99.6	66.3	93.5	**596.6 = GT**
Means	28.8	24.0	14.6	19.9	13.3	18.7	

C.F. $= \frac{596.6^2}{30} = 11\,864.38$

Total SS $= 12\,994.36 - 11864.38 = 1\,129.98$

Treatment SS $= \frac{144.1^2 + \ldots + 93.5^2}{5} - \text{C.F.} = 847.05$

Error SS $= 1\,129.98 - 847.05 = 282.93$

Table: Analysis of Variance (ANOVA)

Source of Variation	df	SS	MS	F
Between cultures	5	847.05	169.41	**14.37****
Within cultures (Error)	24	282.93	11.79	
Total	29	1129.98		

{c.f: $F_{5,24}(5\%) = 2.62$ $F_{5,24}(1\%) = 3.90$}

Mean (30 observations)	= 19.89 mg.
SE (single observation)	= ± 3.42 mg.
CV%	= 17.5 %
SE (mean – 5 observations)	= ± 1.54 mg.
SE (difference between 2 means)	= ± 2.17 mg.

Conclusion: There are significant differences between the treatment means ($p<0.01$).

6.6 TESTING FOR SIGNIFICANT DIFFERENCES BETWEEN MEANS

6.6.1 UNSTRUCTURED TREATMENT SET

At the outset of the investigation, the researcher often has no idea of the relative merits of one treatment over another, the treatments all being of equal importance . In such situations the treatment set is said to "*Unstructured*". In the case of *Unstructured treatment sets* it is generally accepted that `the result of the overall F-test in the Analysis of Variance is an indication to the experimenter whether or not to proceed with further tests to determine which means are significantly different from one other (c.f., "Fisher's restricted LSD").

A **significant overall F-test** tells us that there are differences between means, it **does not** say which means are different from one another. This question may be addressed in a number of ways, some of which are discussed below, (Refer also Chapter 8, *Steel and Torrie*). However, a word of warning; under certain circumstances a **non-significant "over-all" F-test** does not necessarily mean that there are no significant differences between treatment means, this being particularly true if the number of treatments is large and the relevance of the F-test is brought to question. For example, a Plant Breeder may include over one hundred cultivars in a trial in the hope that one or two may be worth further testing. Under such conditions a non-significant F–test may indicate that, as a group, they are reasonably homogeneous – the superiority of the best one or two having been swamped by the lack of differences between the other cultivars. In such cases it should be noted that the Plant Breeder may not be interested in establishing the significance of the superiority of the best cultivars but instead in *maximising* the chances of finding a sub–group of cultivars that are better than those already available. Techniques for *maximising* chance are not dealt with in this course.

TECHNIQUES FOR DETERMINING DIFFERENCES BETWEEN TREATMENT MEANS.

a) Least Significant Differences (LSD's).

b) Scheffe's test.

c) Tukey's test.

d) Student–Newman–Keuls' (S–N–K) test.

e) Duncan's Multiple Range test.

f) Dunnet's test.

g) Bonferroni test

The relative merits of these different methods are discussed in some detail in, *Steel and Torrie, Chapter 8, pages 172 – 193.* Two of the above methods, namely "Least Significant Differences" and "Dunnet's Tests", will be demonstrated in the following sections.

6.6.1.1 LEAST SIGNIFICANT DIFFERENCES (LSD's)

This method is particularly suited to comparisons between pre–determined pairs of treatments (c.f. *"structured treatment sets"* – 6.5.2), but it will be demonstrated using the results of example 6.1 above. Remember, the mean nitrogen content (mg.) of the Red Clover plants for the different combinations of Rhizobia strains were:

Strains:	3DOK1	3DOK5	3DOK4	3DOK7	3DOK13	Composite
Means	28.8	24.0	14.6	19.9	13.3	18.7

The Error MS (s^2) was 11.79 with 24 df.

Under these conditions for two means to be significantly different at the 5% level of significance,

$t_{calculated} \geq 2.064 \quad (\alpha = 0.05).$ i.e. $\frac{|difference|}{SE(difference)} \geq 2.064$

Now SE(difference) $= \sqrt{2\,\frac{11.79}{5}}$ since r = 5 (the number of replications), and $s^2 = 11.79$

Thus for significance at the 5% level, $|\text{ difference }| \geq 2.064 \times \sqrt{2\,\frac{11.79}{5}} = 4.48$

Similarly, for significance at the 1% level, $|\text{ difference }| \geq 2.797 \times \sqrt{2\,\frac{11.79}{5}} = 6.07$

From the above we see that differences will be declared significant at the 5% and 1% levels of significance if they exceed 4.48 or 6.07 respectively. The values 4.48 and 6.07 are called the **5%** and **1%** Least Significant Differences (LSD's). Generally:

$$\mathbf{LSD}_{\alpha} = t_{\alpha} \times \sqrt{2\,\frac{s^2}{r}} \qquad \ldots \{6.8\}$$

where $\alpha = 0.05$ or 0.01, and we have **EQUAL replications (r) for each mean.**

Applying these results to the means of example 6.1 we have, in ranked order:

Rank	Strain	Mean Nitrogen content (mg)	
1	3DOK1	28.8	a
2	3DOK5	24.0	b
3	3DOK7	19.9	bc
4	Composite	18.7	cd
5	3DOK4	14.6	de
6	3DOK13	13.3	e
	Mean	19.89	
	LSD – (5%)	4.5	
	– (1%)	6.1	

Conclusion: Strains with letters in common are not significantly different at the 5% level.
(Similar comparisons may be made at the 1% level.)

NOTE: In the case where there are an unequal number of replications per treatment the LSD method cannot be used effectively. In such cases it is necessary to make comparisons by means of individual t–tests or similar methods.

6.6.1.2 DUNNET'S TEST

This method is particularly useful when something is "known" about one of the treatments (referred to as the "control") and it is desirable to measure this control treatment against the other treatments. Usually, one is not particularly interested in comparisons among the other treatments. Strictly speaking in this case, the treatment set is not unstructured and such a treatment set is sometimes referred to as "semi–structured".

Like the LSD method of 6.5.1.1, Dunnet's procedure require a single value for judging the significance of observed differences between each treatment and the "control". Table A9, (*Steel and Torrie* – pages 590 – 591) is used for one–sided and two–sided alternatives. Again let us use the results of Example 6.1. However, let us suppose that the object of the experiment was to compare the `*Composite'* treatment with the other mixtures of Rhizobia strains, that is, the *Composite* treatment is the "control".

It is to noted that if one wishes to find those treatments which are different from the *Composite*, i.e., those which are better or worse than the *Composite*, the table of two–sided values (Steel & Torrie; Table A9b) is used. On the other hand, if one is concerned with finding only those strains which are superior (or inferior), one–sided values (Steel & Torrie; Table A9a) are used.

(a) DUNNET'S ONE–SIDED TEST

Calculate $D = t(\text{Dunnet}_{\alpha})$ x SE (difference between two means)

From Table A9a, and P=5 (i.e., there are 5 treatments excluding a "control")

$t(\text{Dunnet}_{\alpha}) = 2.36\ (\alpha = 0.05)$ or $3.11\ (\alpha = 0.01)$

$$\left.\begin{aligned} \text{i.e.} \quad & d_{(\alpha=0.05)} = 2.36 \times \sqrt{2\,\tfrac{11.79}{5}} = 5.13 \\ \text{or} \quad & d_{(\alpha=0.01)} = 3.11 \times \sqrt{2\,\tfrac{11.79}{5}} = 6.75 \end{aligned}\right\} \quad \ldots \{6.9\}$$

(b) DUNNET'S TWO–SIDED TEST

Similarly from Table A9b, and P = 5

$t(\text{Dunnet}_{\alpha}) = 2.76\ (\alpha = 0.05)$ or $3.45\ (\alpha = 0.01)$

$$\left.\begin{aligned} \text{i.e.} \quad & d_{(\alpha=0.05)} = 2.76 \times \sqrt{2\,\tfrac{11.79}{5}} = 5.99 \\ \text{or} \quad & d_{(\alpha=0.01)} = 3.45 \times \sqrt{2\,\tfrac{11.79}{5}} = 7.49 \end{aligned}\right\} \quad \ldots \{6.10\}$$

6.6.2 STRUCTURED TREATMENT SET

If treatments have been selected to test for specific differences, the treatment set is said to be structured. Under such conditions the "overall" F – test usually has very little meaning and should be ignored. One should proceed directly to test those differences (comparisons) dictated by the objectives of the research. This can be done using the `t-test' method and variations thereof (c.f. also method of LSD's described above). Failure to proceed with the testing because of a non–significant "overall F–test" would mean failure to test the objectives for which the experiment was designed. In the case of a `*structured treatment set*', the object of the Analysis of Variance is to provide an unbiased estimate s^2 of σ^2 the Error Mean Square, which will form the basis of such tests.

6.6.2.1 CONTRASTS

For planned comparisons involving pairs of treatment means the method of LSD's is recommended.
For comparisons involving more than two treatments, contrasts of the type :

$$W = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \ \ldots \ + \lambda_n x_n = \sum^{n} \lambda_i x_i$$

where λ_i are constants with $\sum \lambda_i = 0$, and the x_i are n distinct variates (or variate values) are considered.

In the case of "*structured treatment sets*" the x_i may be treatment totals or means, the former being more convenient. It will be usual for the totals (or means) to have equal replication (r).

Thus $\quad \text{Variance (W)} = \lambda_1^2 \text{ Var } (x_1) + \lambda_2^2 \text{ Var } (x_2) + \lambda_3^2 \text{ Var } (x_3) + \ \ldots \ + \lambda_n^2 \text{ Var } (x_n)$

$$= \lambda_1^2 \sigma_1^2 + \lambda_2^2 \ \sigma_2^2 \ + \lambda_3^2 \sigma_3^2 \ + \ldots \ + \lambda_n^2 \sigma_n^2 \qquad \{6.12\}$$

Generally $\quad \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots \ \sigma_n^2 \ = \sigma^2$

and {6.12} thus becomes: $\quad \text{Variance (W)} = (\lambda_1^2 + \lambda_2^2 \ + \lambda_3^2 \ + \ldots \ + \lambda_n^2 \ ) \ \sigma^2 \qquad \{6.13\}$

EXAMPLE 6.2: Response to Fertiliser (kilograms per plot)

Contrast	Treatment: No fertiliser	Supers	Supers + N	H_0:
Totals (r = 4)	16	20	26	
1. Fert. vs no fert (w_1)	− 2	1	1	$\frac{1}{2}(\mu_2 + \mu_3) - \mu_1 = 0$
2. N vs No Nitrogen(w_2)	0	− 1	1	$\mu_3 - \mu_2 = 0$

NOTE:
Contrast (1) is the average of treatments `With Fertiliser' vs `No Fertiliser',

i.e., $\quad \frac{1}{2}\Big\{\text{Supers} \ + (\text{Supers+N})\Big\} \text{ vs. } \Big(\text{No fertiliser}\Big)$

Multiplying through by 2 produces the vector of linear coefficients ($\lambda_1 = \ -2, \lambda_2 = \lambda_3 = 1$) above.

Furthermore let the three treatment TOTAL yields be represented by T_1, T_2 and T_3 respectively, over **r** replications. Then from equation {6.13} we have:

Var(*contrast 1*) $= \sum \lambda_i^2 \text{ var}(T_i) = (2^2 + 1^2 + 1^2).\ r\ s^2 \qquad$ since $\text{Var}(T_i) \ = r\ s^2$

(c.f. Section 3.2.5)

and Var(*contrast 1*) $= 6\ .\ r\ .\ s^2 = 24\ s^2 \qquad$ since $r = 4$.

similarly,

Var(*contrast 2*) $= \sum \lambda^2{}_i \text{ var}(T_i) = (0^2 + 1^2 + 1^2).\ r\ s^2 \ = 2\ .\ r\ .\ s^2 = 8\ s^2$

Now the value of $w_1 = (-2 \text{ x} 16 + 1 \text{ x } 20 \ + 1 \text{ x } 26) = 14, \qquad$ and $w_2 = 6$

Thus, assuming $s^2 = 0.25$ with 9 df, we can use t–tests to test each Null Hypothesis as follows:

Contrast 1: $\quad t = \frac{14}{\sqrt{24 \text{ x } 0.25}} \quad = 5.715^{**} \qquad$ [c.f. t(1%) = 3.250]

Contrast 2: $\quad t = \frac{6}{\sqrt{8 \text{ x } 0.25}} \quad = 4.243^{**}$

Conclusion:

(a) On average, there is a significant response to fertiliser of 1.75 ± 0.306 kg per plot, ($p<0.01$).
(b) There is also a significant improvement in yield with the addition of Nitrogen with Phosphate of 1.5 ± 0.353 kg per plot ($p<0.01$).

6.6.3 NOTES ON PRESENTATION OF RESULTS.

a) Where possible present the results as means per unit (in the case above we are not told the plot size so we cannot convert the mean response to kilograms (or tonnes) per hectare, for example.

b) 1.75 is the mean response, i.e., $1.75 = \left\{ \frac{20+26}{8} - \frac{16}{4} \right\}$

Similarly, $1.5 = \left\{ \frac{26-20}{4} \right\}$

c) The SE of these respective estimates are calculated as follows:

$$\left.\begin{aligned} 0.306 &= \sqrt{s^2(\tfrac{1}{4} + \tfrac{1}{8})} \\ 0.352 &= \sqrt{\tfrac{2}{4}\, s^2} \end{aligned}\right\} \text{ where } s^2 = 0.25 \text{ as given above. Refer also, \{5.10a\} and \{5.10b\}.}$$

6.6.4 ALTERNATIVE METHOD — to test for significance of the different contrasts.

a) Any SUM OF SQUARES (SS) with 1 df in an analysis of variance (ANOVA) can be expressed as a single square (i.e., the square of the value of a linear function of observations) divided by the appropriate divisor.

b) If `**L**' is the value of any linear function of variate values each with the same variance (σ^2) such that:

MV(L) = 0, and var(L) = $k\sigma^2$, then L^2 is an estimate of var(L) with 1 df.

Thus $\frac{L^2}{k}$ is an estimate of σ^2 with 1 df.

c) From (b) it is seen that the divisor for L^2 is the coefficient of σ^2, in this case, k.

d) In the calculation of linear functions it is usual to use treatment TOTALS, in which case the divisor (k) is calculated as,

$$\mathbf{k = r \,.\, \sum \lambda_i^2} \qquad \ldots \{6.14\}$$

and

$$\mathbf{SS(L) = \frac{L^2}{r \sum \lambda_i^2}} \qquad \ldots \{6.15\}$$

Let us now re-analyse the data of example 6.2 by Analysis of Variance.

SS (*Contrast 1*) $= \frac{14^2}{24} = 8.17$ and **SS(*Contrast 2*)** $= \frac{6^2}{8} = 4.5$

ANALYSIS OF VARIANCE

Source of Variation	df	SS	MS	F
Between treatments	2	12.67	6.34	
Contrast 1.	1	8.17	8.17	32.4^{**}
Contrast 2.	1	4.50	4.50	18.0^{**}
Within treatments	9	2.25	0.25	
Total	11	14.92		

Notes on the calculations:

i) Since this a set of "orthogonal contrasts" (see definition below),

SS(*Contrast1*) + SS(*Contrast 2*) = Treatment (SS), i.e., (8.17 + 4.50 = 12.67).

ii) Remember $t_r^2 = F_{1,r}$

6.6.5 ORTHOGONAL COMPARISONS

DEFINITION: Two vectors are said to be orthogonal if their vector product equals zero. i.e., two vectors are orthogonal if the sum of the product of corresponding coefficients is zero.

Let L_1 and L_2 be two contrast which may be represented as follows:

$L_1 = -2x_1 + x_2 + x_3$ i.e., $\lambda_{11} = -2$, $\lambda_{12} = \lambda_{13} = 1$

$L_2 = -x_2 + x_3$ i.e., $\lambda_{21} = 0$, $\lambda_{22} = -1$, and $\lambda_{23} = 1$

Then $$\sum_{k=1}^{3} \lambda_{ik}\lambda_{jk} = (-2 \times 0) + (1 \times -1) + (1 \times 1) = 0$$

In matrix notation this may be represented as follows:

$$\begin{bmatrix} L_1 \\ L_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

and $$\begin{bmatrix} -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} = 0$$

NOTES ON ORTHOGONAL FUNCTIONS.

i) Orthogonal linear functions may be regarded as independent under the assumptions adopted for the t–tests or F–tests.

ii) Any SUM OF SQUARES (SS) with f df is capable of being subdivided exactly into f (SS) each with 1 df, by means of orthogonal linear functions.

iii) In order that the probability statements attaching to tests of significance of a number of simultaneous comparisons shall be correct, it is necessary that the comparisons be statistically independent.

iv) For a completely orthogonal subdivision, the sum of the separate SS is equal to the Treatment SS. This is a useful numerical check.

Addendum to Chapter 6

References: ***Steel and Torrie, Section 7.9***
A. A Rayner, Chapter 25

Sampling vs. Experimental Error in Analysis of Variance

RANDOMISATION (R A Fisher 1925) – the allocation by chance so that each treatment has an exactly equal chance of being allocated to any one experimental unit. A physical process is insisted upon through, for example the use of *tables of Random Number*.

EXPERIMENTAL UNITS are those units of experimental material to which the different treatments are allocated in consequence of each single act of randomisation

In some experimental situations, several observations may be made within the experimental unit, i.e., the unit to which the treatment is applied. Such observations are made on **SUBSAMPLES** or **SAMPLING** units. Differences among sampling units within an experimental unit are observational differences rather than experimental unit differences.

Consider the following examples:

1. "For convenience" all chickens in one house get Ration A, the chickens in a second house get Ration B. Even though there may be 1000 chickens in each house this is an "unreplicated" experiment.

2. An experimenter, wishing to compare the effects of two treatments on weight gains of cattle, records the weights of the animals and divides them into two equal groups in such a way that the mean weights of the groups are equal. He then allocates the two groups to the treatments by tossing a coin. The two groups share identical conditions apart from the different treatments. In this case, since the animals were NOT randomly allocated to treatments, this is not a valid (replicated) experiment, the animals within each treatment group being, at best, sampling units

3. A researcher investigating the relative amount of a particular steroid component in 5 different varieties of soybean obtains 100 grams of seed for each seed type, prepares the seed for chemical analysis and then subdivides the resulting solutions for each variety each into 2 beakers; i.e 8 beakers. From each beaker two test tubes are filled and placed into an "auto-analyser"; i.e., 4 tubes per original solution per variety (16 analyses altogether). Here again, we only have a single replicate of each seed type, the four analyses per variety are in effect a measure of sampling variability; in this case the sampling error may be seen as a measure of the variability inherent in the method of analysis.

The linear model must be modified to account for this extra variability:

$$y_{ijk} = \mu + \tau_i + \epsilon_{ij} + \delta_{ijk}$$

where y_{ijk} is the measure obtained from the k^{th} sample unit for the j^{th} experiment unit for the i^{th} treatment level; μ and τ have the same meaning as before (refer, fixed effects Model). Two random elements are obtained with each observation, namely: ϵ_{ij} which are assumed to have zero mean and variance σ_e^2; the δ_{ijk} are assumed to have zero mean and variance σ^2. The two estimates of variation are independent of each other.

EXAMPLE

A plant breeder is investigating the resistance of different corn (maize) cultivars to lodging. In a cultivar trial containing 12 cultivars each replicated 4 times, the researcher selects at random 6 plants within each plot and measures the force needed to uproot each plant. The "skeleton" ANOVA table is as follows :

Source	df	Expected (MS)
Between cultivars (12)	11	$\sigma^2 + 6\sigma_e^2 + 24\sum \frac{\tau_i^2}{11}$
Within cultivars (*experimental error*)	36	$\sigma^2 + 6\sigma_e^2$
Total plots (48)	47	
Between plants within plots:		
(*sampling error*)	240	σ^2
Total plants (12 x 4 x 6)	287	

From the above table it can be seen that the "error MS" can, theoretically, not be smaller than the "sampling error MS". A common error in statistical analysis is to overlook the existence of these two sources of error, this is turn can lead to false statements of probability based on an incorrect estimate of "error variance" as well as an incorrect number of degrees of freedom.

In the above Example it is seen that there are a total of 288 observations and the calculation of the "correction factor" will be :

$$CF = \frac{GT^2}{288} \qquad \text{generally} = \frac{GT^2}{t \times r \times s}$$

where t= No. of treatments

r= No. of replications per treatments (assumed equal)

s= No. of sampling units per experimental unit/plot (assumed equal)

Similarly,

$$SS\ (\text{cultivars}) = \frac{\sum T_i^2}{(4 \times 6)} - CF \qquad \text{generally} = \frac{\sum T_i^2}{(r \times s)} - CF$$

– refer formula {6.5a}, page 29

In this case the divisor is 24 since each "treatment total (T_i)" comprises 24 plants, 6 from each of the 4 plots. Furthermore it is necessary to calculate:

$$SS\ (\text{plots}) = \frac{\sum P_i^2}{6} - CF \qquad \text{generally} = \frac{\sum P_i^2}{s} - CF$$

since in this example each plot total (P_i) is a total of 6 plants.

The two "Error SS" are calculated by subtraction.

The results from the ANOVA table can be used to provide estimates of both σ^2 and σ_e^2 which in turn are used to detemine "optimum" sample sizes.

CHAPTER 7 : CORRELATION

7.1 BIVARIATE SAMPLES/POPULATIONS

Previously we have studied only a single variate at any one time, i.e., we have studied univariate populations and samples from univariate populations. However in biological research it is frequently necessary to consider multivariate populations in order to get the whole picture. The consideration of more than two variables *simultaneously* is beyond the scope of this course. However, we will consider the special case of bivariate samples and populations.

When we consider two quantitative characteristics (x_1, x_2) of a single individual (e.g, plant height, dry weight of plant) or associated individuals (e.g. height of father, height of son) in a sample it is called a BIVARIATE SAMPLE and the n **PAIRS** of simultaneously sampled values are designated (x_1 x_1), (x_2 x_2) ... (x_n x_n). As for univariate samples, bivariate samples may be assumed to be from finite or infinite, discrete or continuous, bivariate populations.

7.2 CORRELATION

Consider a sample size n from a bivariate population (x_1 x_2) with means (μ_1 μ_2) and variances (σ_1^2 σ_2^2) respectively. These n sample points may be represented graphically by means of a 2–dimensional graph, a SCATTER DIAGRAM, with axes X_1 and X_2. Let us represent the point (x_1 x_2) with the following diagram with origin (μ_1 μ_2).

[INSERT Scatter diagram]

Figure 7.1

7.2.1 CO-RELATIONSHIPS

In many biological investigations it is important to measure the strength (intensity) of the co–relationships between two variables, x_1 and x_2, say. By intensity of correlation is meant the extent to which deviations from the mean in one variate ***tend*** to be accompanied by ***proportional*** deviations in the other variate. Perfect correlation occurs when deviations in the one variate are exactly proportional to the simultaneous deviations in the other variate. Consider the point (x_1 x_2) in Figure 7.1., and the deviations $(x_1 - \mu_1)$ and $(x_2 - \mu_2)$:

a) The product of the deviations $(x_1 - \mu_1)(x_2 - \mu_2)$ is positive for any point (x_1 x_2) in quadrants I and III. Hence if the majority of points lie in these two quadrants the sign of $\sum(x_1 - \mu_1)(x_2 - \mu_2)$ will be positive.

b) Similarly, for points lying in quadrants II and IV, where $(x_1 - \mu_1)$ and $(x_2 - \mu_2)$ will have opposite signs, the product $(x_1 - \mu_1)(x_2 - \mu_2)$ will be negative and the sign of $\sum(x_1 - \mu_1)(x_2 - \mu_2)$ will be negative.

c) f the points are equally spaced over the four quadrants, $\sum(x_1 - \mu_1)(x_2 - \mu_2)$ will tend to have a value close to zero.

d) The quantity $\sum(x_1 - \mu_1)(x_2 - \mu_2)$, is a measure of the simultaneous co–variation between the two variables x_1 and x_2 but it is not independent of n, the sample size, and it is therefore necessary to consider an "average deviation product",

The **COVARIANCE of** $(x_1\ x_2)$ is defined as:

$$\mathbf{COV}(x_1x_2) = \mathbf{M.V}\left\{(x_1 - \mu_1)(x_2 - \mu_2)\right\} \quad \textbf{i.e. } \mathbf{COV}(x_1\ x_2) = \sigma_{x_1x_2} \quad \text{(usually written as } \sigma_{12}\text{)}$$

For a sample size n from a bivariate population with means $(\mu_1\ \mu_2)$

COVARIANCE is defined as: $$\sigma_{12} = \frac{\sum(x_1-\mu_1)\,(x_2-\mu_2)}{n} \qquad \ldots \{7.1\}$$

As for estimates of variance (c.f. chapter 3), μ_1 and μ_2 are rarely known and must be estimated by $\bar{x}_1$ and $\bar{x}_2$ respectively. In which case:

SAMPLE COVARIANCE is defined as: $$s_{12} = \frac{\sum(x_1-\bar{x}_1)\,(x_2-\bar{x}_2)}{(n-1)} \qquad \ldots \{7.2\}$$

It is to noted, however that population and sample estimates of covariance are however dependent on the scale of measurement for the two variates x_1 and x_2.

7.3 CORRELATION COEFFICIENT

Any measure of CORRELATION must be comparable from sample to sample, irrespective of the units in which the variates are measured, i.e., "the average deviation product" should be independent of the units of measurement of the two variates.

7.3.1 CORRELATION COEFFICIENT – Definitions.

1) **Correlation Coefficient of an infinite bivariate population.**

$$\rho = \frac{\mathbf{M.V}\{(x_1-\mu_1)(x_2-\mu_2)\}}{\sigma_1\sigma_2} \qquad \ldots \{7.3\}$$

NOTE:

i) ρ is the Greek letter RHO.

ii) The division of σ_{12} by σ_1 and σ_2 is to secure the required freedom from the particular units in which x_1 and x_2 are measured, (i.e., x_1 and x_2 are "standardised").

2) **Correlation Coefficient of a bivariate sample (size n)**

$$r_{12} = \frac{\frac{\sum(x_1-\bar{x}_1)(x_2-\bar{x}_2)}{(n-1)}}{\sqrt{\frac{\sum(x_1-\bar{x}_1)^2}{(n-1)} \bullet \frac{\sum(x_2-\bar{x}_2)^2}{(n-1)}}} \qquad \ldots \{7.4\}$$

r_{12} is an unbiased estimate of $\boldsymbol{\rho_{12}}$, the population correlation coefficient.

NOTE: The computation of r_{12} is simplified by the removal of $(n-1)$.

Thus, $$r_{12} = \frac{\sum(x_1-\bar{x}_1)(x_2-\bar{x}_2)}{\sqrt{\sum(x_1-\bar{x}_1)^2 \bullet \sum(x_2-\bar{x}_2)^2}} \qquad \ldots \{7.5\}$$

The numerator of {7.5} is called the SUM OF PRODUCTS, $SP(x_1x_2)$ and, as for Sum of Squares, is computed simply as,

$$\mathbf{SP}(x_1\ x_2) = \sum(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) = \sum x_1x_2 - \frac{\sum x_1 \sum x_2}{n} \qquad \ldots \{7.6\}$$

7.3.2 SOME NOTES ON CORRELATION

a) The correlation coefficient (r) measures the **degree of linear dependence** of two variates.

b) INDEPENDENT variates have zero correlation (i.e., they are uncorrelated) BUT, zero correlation does not (necessarily) imply independence, since the variates could be non–linearly related.

c) The only type of dependence between NORMAL variates is linear dependence.
i.e., INDEPENDENT Normal variates are UNCORRELATED,
UNCORRELATED Normal variates are INDEPENDENT.

d) $|\rho| \leq 1$ i.e., $-1 \leq \rho \leq 1$

$\rho = +1$, implies that we have perfect positive correlation.

$\rho = -1$, implies that we have perfect negative correlation.

$\rho = 0$, implies that there is no LINEAR relationship between the two variates.

e) In discussing correlation the following terminology is often used:

(i) $|\rho| < 0.3$ – WEAK correlation

(ii) $|\rho| < 0.5$ – MODERATE correlation

(iii) $|\rho| > 0.7$ – STRONG correlation

7.4 SAMPLING DISTRIBUTION OF (r)

7.4.1 In general, the probability distribution of (r) is NOT normally distributed. However for large bivariate samples and moderate or small (weak) values of ρ,

$$\mathbf{r} \sim \text{N.D.}\left(\rho, \frac{1-\rho^2}{\sqrt{n-1}}\right) \text{ approximately} \qquad \ldots \{7.6\}$$

7.4.2 For Normal bivariate samples and for the assumption, H_0: $\rho = 0$

$$\mathbf{r}\sqrt{\frac{n-2}{1-r^2}} \sim t_{(n-2)} \qquad \ldots \{7.7\}$$

This statistic is referred to Significance Tables of (r) for different levels of α and **(n — 2) df**
(Reference: Tables A.13 – *Steel and Torrie*)

7.5 TRANSFORMATION OF (r) TO APPROXIMATE NORMALITY

For normal bivariate samples, the quantity,

$$Z_r = \tfrac{1}{2}\{\ln(1+r) - \ln(1-r)\} \qquad \ldots \{7.8\}$$

is approximately $\text{N.D.}\left(Z_\rho, \frac{1}{n-3}\right)$

i.e. Z_r has mean $= Z_r$ and variance $= \frac{1}{n-3}$.

NOTE:

i) This transformation is reasonably accurate even for small n (e.g. $n \geq 11$).

ii) Tables for the transformation of r to Z_r are available.
(c.f. Rayner, *Appendix Table 8*; Steel and Torrie, *Appendix Table A.13*, et al)

EXAMPLE 7.1 Test H_o: $\rho = 0$

$r = 0.974$ $n = 27$ (i.e., df = 25)

From tables (A13 – *Steel and Torrie*) critical values: (5%) = 0.3809

(1%) = 0.4869

Conclusion: Since $r > 0.4869$, Reject H_0: at 1% level

EXAMPLE 7.2 Test H_0: $\rho = \rho_0$ ($\neq 0$) – **Approximate Test**

r.0.954 n=27 H_0: $\rho_0 = 0.80$

$Z_r = 1.87$ $Z_\rho = 1.10$ $\text{Var}(Z_r) = \frac{1}{(27-3} = 0.04166$

$$z = \frac{Z_r - Z_\rho}{\sqrt{\text{Var}(z)}} = \frac{1.87-1.10}{\sqrt{0.04166}} = 3.77^{**}$$ [c.f. z (1%) = 2.576]

Conclusion: Reject H_0: at 1% level

EXAMPLE 7.3 Test H_0: $\rho_1 = \rho_2$ ($\neq 0$) – **Approximate Test.**
[i.e test whether two estimates of r are significantly different.]

$r_1 = 0.954$ $n_1 = 27$ $r_2 = 0.882$ $n_2 = 20$

$Z_{r_1} = 1.87$ $Z_{r_2} = 1.38$

$\text{Var}(Z_{r_1}) = \frac{1}{(27-3)} = 0.04166$ $\text{Var}(Z_{r_2}) = \frac{1}{(20-3)} = 0.05882$

$$z = \frac{Z_{r_1} - Z_{r_2}}{\sqrt{\text{Var}(Z_{r_1}) + \text{Var}(Z_{r_2})}} = \frac{1.87-1.38}{\sqrt{0.04166+0.05882}} = 1.55 \text{ (n.s)}$$ refer: z (5%) = 1.96

Conclusion: Accept H_0

EXAMPLE 7.4 95 % Confidence Interval (limits) for ρ

$r = 0.951$ $n = 27$ $Z_r = 1.875$ $\text{Var}(Z_r) = 0.04166$ (i.e. $= \frac{1}{24}$)

$Z_\rho = Z_r \pm 1.96 \text{ x SE } (Z_r) = 1.875 \pm 1.96 \text{ x } 0.2041$

i.e. $Z_\rho =$ (2.275 ; 1.475)

Thus $0.901 < \rho < 0.979$ (i.e., **Transform Z_r back to r)**

7.6 VARIANCE OF LINEAR FUNCTION OF CORRELATED VARIATES

Consider $W = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \ldots + \lambda_k x_k = \sum^{k} \lambda_i x_i$; where λ_i are constants, and the x_i are n distinct variates (or variate values)

Let VARIANCE $(x_i) = \sigma_i^2$ and the CORRELATION of x_i and x_j be ρ_{ij}

Then $$\textbf{Variance (W)} = \lambda_1^2 \sigma_1^2 + \lambda_2^2 \sigma_2^2 + \ldots + \lambda_k^2 \sigma_k^2 + 2\lambda_1\lambda_2 \rho_{12} \sigma_1\sigma_2 + 2\lambda_1\lambda_3 \rho_{13} \sigma_1 \sigma_3 + \ldots \quad \ldots \{7.9\}$$

7.6.1 SPECIAL CASES.

a) If x_i, x_j are uncorrelated (i.e $\rho_{ij} = 0$)

$$\text{Variance (W)} = \lambda_1^2 \sigma_1^2 + \lambda_2^2 \sigma_2^2 + \lambda_3^2 \sigma_3^2 + \ldots + \lambda_n^2 \sigma_n^2 \qquad \text{(c.f. Section 3.6.2)}$$

b) $$\textbf{Var}(x_1 + x_2) = \sigma_1^2 + \sigma_2^2 + 2\lambda_1 \lambda_2 \rho_{12} \sigma_1\sigma_2$$

$$= \textbf{var}(x_1) + \textbf{var}(x_2) + 2\,\textbf{Cov}(x_1x_2) \quad \ldots \{7.10a\}$$

c) $$\textbf{Var}(x_1 - x_2) = \sigma_1^2 + \sigma_2^2 - 2\lambda_1 \lambda_2 \rho_{12} \sigma_1\sigma_2$$

$$= \textbf{var}(x_1) + \textbf{var}(x_2) - 2\,\textbf{Cov}(x_1x_2) \quad \ldots \{7.10b\}$$

NOTE: If x_1 and x_2 are uncorrelated ($\rho_{12} = 0$)

$$\text{Var}(x_1 \pm x_2) = \text{var}(x_1) + \text{var}(x_2), \text{ as before.} \qquad \text{(c.f. Section 3.6.2)}$$

7.6.2 REDUCTION OF VARIANCE DUE TO PAIRING

In paired data , the individuals of a pair tend to be alike, i.e., there will be positive correlation.

Now for UNPAIRED data (i.e., data uncorrelated),

$$\text{Var}(x_1 - x_2) = \text{var}(x_1) + \text{var}(x_2) \quad \ldots \{7.11\}$$

But for PAIRED data, $$\text{Var}(x_1 - x_2) = \text{var}(x_1) + \text{var}(x_2) - 2\rho_{12}\sigma_1\sigma_2 \quad \ldots \{7.12\}$$

Thus, since ρ_{12} is positive, {7.12} < {7.11}.

NOTE:
This result explains the reduction in SE($\bar{x}_1 - \bar{x}_2$) for Examples 5.3 & 5.5 – *`Lactation data'*

7.6.3 ORTHOGONALITY AND CORRELATION

Let $w_1 = \underline{\lambda}'\underline{x}$ and $w_2 = \underline{\alpha}'\underline{x}$ be 2 linear functions of variates $x_1, x_2 \ldots x_k$

i.e. $w_1 = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \ldots + \lambda_k x_k$

and $w_2 = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k$

then $$\text{Cov}(\underline{\lambda}'\underline{x}, \underline{\alpha}'\underline{x}) = \lambda_1\alpha_1\sigma_1^2 + \lambda_2\alpha_2\sigma_2^2 + \ldots + (\lambda_i\alpha_j + \lambda_j\alpha_i)\rho_{ij}\sigma_i\sigma_j + \ldots \quad \ldots \{7.13\}$$

Now if the x_i have equal variances (σ^2) and are uncorrelated (i.e., $\rho_{ij} = 0$) the above result of {7.13} reduces to :

$$\textbf{Cov}(\underline{\lambda}'\underline{x}, \underline{\alpha}'\underline{x}) = (\lambda_1\alpha_1 + \lambda_2\alpha_2 + \ldots + \lambda_k\alpha_k)\sigma^2$$

thus the condition for $\underline{\lambda}'\underline{x}$ and $\underline{\alpha}'\underline{x}$ to be uncorrelated is for $(\lambda_1\alpha_1 + \lambda_2\alpha_2 + \ldots + \lambda_k\alpha_k) = 0$

But this is the same condition for orthogonality of $\underline{\lambda}'\underline{x}$ and $\underline{\alpha}'\underline{x}$, thus orthogonal linear functions are uncorrelated.

NOTE: This is confirmation of the results of Section 6.5.4 – *Orthogonal Comparisons*

7.7 USEFULNESS OF CORRELATION

a) To see if it is feasible to predict the value of one variate of a pair given the value of the other.
(refer: *Regression Analysis* – chapter 8)

b) To establish that a linear relationship exists and then consider the underlying causes of that relationship, – (c.f. *Biological Modelling/Multiple Regression*)

7.8. CORRELATION AND LINEAR REGRESSION

Correlation and Linear Regression are different aspects of the same type of relationship between two variates.
CORRELATION – measures the degree of linear dependence between the two variates.
LINEAR REGRESSION – gives practical effect to this dependence by expressing one of the variates as linearly dependent on the other, with allowance of course for some degree of random deviation (or non–random deviation, if the real relationship is only approximately linear) from an exact linear relationship, unless $\rho = \pm 1$.
Regression is in general the more useful of the two concepts and is also available in circumstances when the use of the correlation coefficient would be artificial or misleading.

CHAPTER 8 : LINEAR REGRESSION

8.1 REGRESSION FUNCTIONS

In chapter 7 we used the correlation coefficient as a measure of the intensity of the linear relationship between to variates. In this chapter Regression Analysis aims at the formulation of an explicit functional relationship between the variates and the estimation of this relationship from sample or experimental data. In general the object will be to express one variate y as functionally dependent on a second variate x, viz. $y = \psi(x)$, where $\psi(x)$ is a a mathematical function which might be postulated from a theoretical examination of the relationship between y and x or merely from a scatter diagram of observed data. $\psi(x)$ is called the **regression function**, and we speak of the "regression of y on x". In the study of linear relationships two situations may be encountered.

EXAMPLE 1

Two quantitative characteristics (x_1, x_2) of a single individual i.e., the bivariate sample size n, might consider the height (x_1) and weight (x_2) of each man in the sample.

EXAMPLE 2

Two characteristics (x_1, x_2) connected with different individuals, but which are paired for some logical reason. e.g. height of father (x_1), height of son (x_2). In each case we may represent the data as,

Sample #	(x_1)	(x_2)
1	x_{11}	x_{21}
2	x_{12}	x_{22}
3	x_{13}	x_{23}
.	.	.
.	.	.
.	.	.
.	.	.
n	x_{1n}	x_{2n}

i.e. we have a bivariate sample of **n pairs of variate values (x_1, x_2)**. A bivariate population may consist of an infinite (or finite) number of such pairs.

8.2 BIVARIATE and REGRESSION MODELS

Consider the **univariate model** for a random variable y_i.

$$y = \mu_y + \epsilon \qquad \ldots \{8.1\}$$

Where $MV(\epsilon) = 0$, $Var(\epsilon) = \sigma_y^2$.

8.2.1 BIVARIATE MODEL

$$y = \mu_{y,x} + \epsilon \qquad \ldots \{8.2\}$$

Where $\mu_{y,x}$ is the mean value of y for a given x, and $MV(\epsilon) = 0$, but $Var(\epsilon) = \sigma_{y,x}^2$;

i.e. the Variance of y **is conditional on the value of x.**

8.2.2 REGRESSION MODEL

$$y = \phi(x) + \epsilon \qquad \ldots \{8.3\}$$

$\phi(x)$ is the Regression function and ϵ is as in {8.2} above.

8.2.3 LINEAR REGRESSION MODEL

It is now necessary for us to postulate some mathematical function which will best describe the relationship between the variables over the range of values studied or sampled. The simplest function is that for a straight line, viz.

$$\phi(x) = \alpha + \beta x \qquad \ldots \{8.4\}$$

i.e., $y = \alpha + \beta x + \epsilon$

Consider the following scatter diagram:

Figure 8.1

Our problem is to find a and b the sample estimates of α and β in our model $y = \alpha + \beta x + \epsilon$, such that the regression estimate of y, viz. $\hat{y} = a + b x$ is the "best fitting" line. The criterion used to obtain this best fitting line is known as the *Principle of Least Squares.*

8.3 PRINCIPLE OF LEAST SQUARES

Figure 8.2

The Principle of Least Squares requires us to MINIMISE $\sum(y_i - \hat{y}_i)^2 = \sum e_i^2$, i.e., the sum of the squared deviations of y from its regression estimate $\hat{y}$. It is to be noted that $\hat{y}$ is an estimate of $\mu_{y,x}$ in the model, $y = \mu_{y,x} + \epsilon$, and e_i is an estimate of ϵ, the random bivariate residual component (i.e., error).

8.3.1 APPLICATION OF PRINCIPLE OF LEAST SQUARES

In General we must consider minimising the **weighted** sum of squared deviations from regression, i.e., $\sum w_i(y_i - \hat{y})^2$, where the weights w_i attached to the i^{th} observation depends on assumptions made about variance of ϵ in model. If it can be assumed that the variance of ϵ is constant (the assumption of homocedasticity), the weights w_i can be taken as equal. We will consider therefore the unweighted linear regression,

$$\textbf{i.e., } \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a - bx_i)^2 \qquad \text{where } \hat{y} = a + bx.$$

For convenience we will consider the model as :

$$\mathbf{y = a^* + b(x - \bar{x})} \qquad \text{where } a^* = a + b\bar{x}$$

i.e., Minimise $\underline{W} = \sum(y - a^* - b(x - \bar{x}))^2$ with respect to a^* and b. This is achieved by obtaining the partial derivatives of $\underline{W}$ with respect to a^* and b and equating the resultant equations to zero and solving for a^* and b respectively.

i.e. $\frac{\partial w}{\partial a^*} = 0$

$\frac{\partial w}{\partial b} = 0$ } **i.e., equate partial derivatives w.r.t a^* and b to zero.**

i.e. $$\frac{\partial w}{\partial a^*} = 2\sum(y - a^* - b(x - \bar{x})) = 0 \qquad \{8.5a\}$$

and $$\frac{\partial w}{\partial b} = 2\sum(x - \bar{x})(y - a^* - b(x - \bar{x})) = 0 \qquad \{8.5b\}$$

Since $\sum(x - \bar{x}) = 0$, then from {8.5a}

$$na^* = \sum y \qquad \text{i.e.,} \quad \mathbf{a^* = \bar{y}} \qquad \{8.6\}$$

and substituting into {8.5b} we get,

$$\mathbf{b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}} \qquad \{8.7\}$$

Thus the estimated linear regression line of y on x (i.e., the line of 'best fit') is,

$$\mathbf{\hat{y} = \bar{y} + b(x - \bar{x})} \qquad \{8.8\}$$

SUMMARY:

In the fitted model: $\hat{y} = a + bx$

b = slope of the line $\Rightarrow$ average change in y per unit change in x.

$\Rightarrow$ linear regression coefficient of y on x , $b_{y,x}$

a = $\bar{y} - b\bar{x}$ = intercept on the y – axis (i.e., value for $\hat{y}$ for x = 0).

NOTE:

The sign of b is the sign of SP(x y) which, unlike SS(x) or SS(y), can have positive or negative values.

8.4 ESTIMATION OF VARIANCE ($\sigma^2_{y,x}$)

In any least squares analysis, provided the model is correct, $e_i^2 = (y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the fitted value corresponding to x_i, represents deviations due to random causes only, and estimates,

$$\mathbf{Var(\epsilon) = \frac{\sum \epsilon_i^2}{df}} .$$

NOTE:

i) $\sum e_i^2$ is variously called;
- "SS due to lack of fit"
- "Residual SS"
- "SS due to deviations from regression"

ii) df = n – (number of parameters fitted) = n – 2, since 2 parameters are fitted *viz.* a and b.

i.e. $$\frac{\sum(y-\hat{y})^2}{n-2} = \mathbf{S^2_{y,x}}$$ is an estimate of $\sigma^2_{y,x}$ (= Var(ϵ)) ... **{8.9}**

Now $\hat{y} = \bar{y} + b(x - \bar{x})$

thus $\sum(y - \hat{y})^2 = \sum(y - \bar{y} - b(x - \bar{x}))^2$

$= \sum(y - \bar{y})^2 - 2b\sum(x - \bar{x})(y - \bar{y}) + b^2 \sum(x - \bar{x})^2)$

$= SS(y) - 2bSP(xy) + b^2 SS(x).$

$= SS(y) - \left\{ \frac{(SP^2(xy)}{SS(x)} \right\}$ since $b = \frac{SS(xy)}{SSx}$

Alternatively: $$\sum(y-\bar{y})^2 = \frac{SP^2(xy)}{SS(x)} + \sum(y-\hat{y})^2 \quad \ldots \{8.10\}$$

i.e. **TOTAL SS = SS(due to regression) + SS(deviation from regression)**

8.4.1 ANALYSIS OF VARIANCE FOR LINEAR REGRESSION

The results of {8.10} may now be summarised in an ANOVA table as follows.

Source of Variation	DF	SS	MS
Due to Regression	1	$\frac{SP^2xy}{SS(x)} = B$	$\frac{B}{1}$
Deviation from Regn.	$n-2$	"by subtraction" (=C)	$\frac{C}{n-2}$
Total	$n-1$	$SS(y) = A$	

Now if the model, $y = \alpha + \beta_x + \epsilon$ is correct, the "Deviations from Regression Mean Square" is an estimate $S^2_{y,x}$ of $\sigma^2_{y,x}$ with $(n-2)$ df. It can also be shown that the "Regression Mean Square" is an estimate of $\sigma^2_{y,x} + \beta^2$ SS(x) with 1 DF. Thus if $|\beta| > 0$, i.e., if there is a tendency for the plotted points to follow a linear trend, it will be reflected in a larger "M.S for Regression" than for "Deviation from Regression M.S". On the Null Hypothesis, $H_0 : \beta = 0$, the two M.S.'s are expected to be equal. Furthermore, if the ϵ_i in the model, $y_i = \alpha + \beta_{xi} + \epsilon_i$ are $ND(0,\sigma^2_{y,x})$**

then the ratio, $$\frac{\textbf{Regression MS}}{\textbf{Deviation MS}} \quad \textbf{for } H_0\textbf{: } \beta = 0, \sim F_{1,(n-2)}$$

i.e. $$\frac{\textbf{Regression MS}}{\textbf{Deviation MS}} \sim F_{1,n-2}$$ is a test of $H_0 : \beta = 0$.

8.4.2 INTERPRETATION OF SIGNIFICANT/NON–SIGNIFICANT F–VALUES

F Significant:

"SS due to regression" take up a significant proportion of the Total SS. i.e., the linear arrangement of points is **NOT** due to chance.

F non–significant:

The straight line fit is poor.

Thus,

(i) data may follow a random scatter ($\beta = 0$), **OR**

(ii) data may tend to follow a curve, i.e., the linear regression model is incorrect and the deviations $(y-\hat{y})$ are not due to random causes only.

NOTES:

a) $$SP(xy) = \sum(x-\bar{x})(y-\bar{y}) = \sum(x-\bar{x})y - \bar{y}\sum(x-\bar{x}) = \sum(x-\bar{x})y$$

i.e., SP(x,y) may be written a s linear function of the variates y,

thus $$b = \frac{SP(xy)}{SS(x)} = \frac{\sum(x-\bar{x})y}{SS(x)}$$

Now if the y_i are uncorrelated and have variance $\sigma^2_{y,x}$

$$\textbf{Var(b)} = \textbf{Var}\left\{\frac{\sum(x-\bar{x})y_i}{SS(x)}\right\} = \frac{\sigma^2_{y,x}}{SS(x)} \quad \ldots \textbf{\{8.11\}}$$

b) Similarly,

SS(due to fitting b) $= \frac{(\text{linear function})^2}{\text{divisor}} = b^2.SS(x) = \frac{SP^2_{xy}}{SS(x)} =$ SS(due to Regression) where

NOTE:'divisor' = coefficient of $\sigma^2_{y,x}$.

c) Since $\text{Var}(b) = \frac{\sigma^2_{y,x}}{SS(x)}$ and if the regression is truly linear, we may substitute $s^2_{y,x}$, for $\sigma^2_{y,x}$

i.e. $$\mathbf{SE(b)} = \frac{S_{y,x}}{\sqrt{SS(x)}}$$

8.4.3 ALTERNATIVE TEST FOR H_o: $\beta = 0$ (t–test)

On the assumption that the ϵ_i in the model, $y_i = \alpha + \beta x_i + \epsilon_i$ are $ND(0,\sigma^2_{y,x})$

$$\mathbf{t = \frac{b}{SE(b)} = \frac{b\sqrt{SS(x)}}{s_{y,x}}} \qquad \text{with } \mathbf{(n-2)} \text{ d.f.} \qquad \{8.12\}$$

8.5 CONFIDENCE LIMITS TO β

Prob $\{b - t_\alpha SE(b) < \beta < b + t_\alpha SE(b)\} = 1 - \alpha$

where t_α is the t value corresponding to (α%) level of significance for (n – 2)df.

i.e. $\underline{\beta = \mathbf{b \pm t_\alpha\, SE(b)}}$ are the $(1-\alpha)$% confidence limits for β. ... {8.13}

8.6 REGRESSION ESTIMATES ($\hat{y}$)

Consider $$\hat{y}_g = \bar{y} + \mathbf{b(x_g - \bar{x})}$$

where $\mathbf{x_g}$ is a given value of x for which we require the corresponding value of y, viz $\mathbf{y_g}$.

a) $\hat{y}_g$ **may be used to estimate** μ_{y,x_g}, the mean of the infinite population of values corresponding to x_g, i.e., the mean of the *conditional distribution of y*.

b) $\hat{y}_g$ **may be used to predict the mean of the y–values** of a finite group for which the mean x–value is x_g.

c) $\hat{y}_g$ **may be used to predict the value of y for an individual** with $x = x_g$.

8.6.1 VARIANCE OF REGRESSION ESTIMATES ($\hat{y}$)

8.6.1.1 $\hat{y}_g$ may be used to estimate μ_{y,x_g}, **the mean of the infinite population** of values corresponding to x_g.

$$\text{Var}(\hat{y}_g) = \text{Var}(\bar{y} + b(x_g - \bar{x})) = \frac{\sigma^2_{y,x}}{n} + (x_g - \bar{x})^2 \frac{\sigma^2_{y,x}}{SS(x)}$$

$$\mathbf{Var(\hat{y}_g) = \sigma^2_{y,x}\left\{\frac{1}{n} + \frac{(x_g - \bar{x})^2}{SS(x)}\right\}} \qquad \ldots \{8.14a\}$$

8.6.1.2 $\hat{y}_g$ may be used to predict the **mean of the y–values of a finite group (say, m values)** for which the mean x–value is x_g.

$$\mathbf{Var(\hat{y}_g) = \sigma^2_{y,x}\left[\frac{1}{m} + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{SS(x)}\right]} \qquad \ldots \{8.14b\}$$

8.6.1.3 $\hat{y}_g$ may be used to predict the value of y **for an individual with** $\mathbf{x = x_g}$.

$$\mathbf{Var(\hat{y}_g) = \sigma^2_{y,x}\left\{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{SS(x)}\right\}} \qquad \ldots \{8.14c\}$$

Estimates of {8.14a}, {8.14b} and {8.14c} are obtained by substituting $s^2_{y,x}$ with (n – 2)df for $\sigma^2_{y,x}$

8.6.2 CONFIDENCE INTERVALS/LIMITS FOR $\hat{y}$

On the assumption the ϵ_i are N.D the above estimates of variance for $(\hat{y}_g)$ may be used to set confidence limits to the true value of the quantity being estimated by $\hat{y}_g$ (i.e., $\mu_{y,xg}$). These confidence limits for $\hat{y}_g$ for various values of x_g are found to lie on two curves (branches of 2 hyperbola), i.e., these curves are the boundaries of a confidence band to the regression line, and the area between them represents a zone within which we can be 95% certain that $\mu_{y,x}$ will lie for values of x within the observed range.

8.7 INTERPOLATION AND EXTRAPOLATION

INTERPOLATION is when $\hat{y}_g$ is calculated for an x_g **WITHIN** the range but not one of the x–values in the original data.

EXTRAPOLATION is when x_g lies **OUTSIDE** the range of x's in the data.

NOTE:

a) Extrapolation is risky, but it is often used for example in projections of time series for predictive purposes (forecasting).

b) If a linear approximation is being used when the real relationship is non–linear, interpolation may be adequate over the observed range of values, but extrapolation is likely to be hopeless beyond that range.

c) From {8.14 a,b,c} it is seen that $\text{Var}(\hat{y}_g)$ varies with x_g and is a minimum for $x_g = \bar{x}$, and the further x_g is from $\bar{x}$ the larger $\text{Var}(\hat{y}_g)$. Another reason against extrapolation!

8.8 INVERSE ESTIMATION

Inverse estimation is the use of the regression of y on x to estimate x for a given y value, i.e., from our fitted equation, $\hat{y} = \bar{y} + b(x - \bar{x})$, given y_g, we are required to estimate x, i.e., $\hat{x}_g$.

We get, $$\hat{x}_g = \bar{x} + \frac{(y_g - \bar{y})}{b} \quad \ldots \{8.15\}$$

This gives an estimate of the value of x which corresponds to a given conditional mean of y, i.e, $\mu_{y,x} = y_g$. Inverse estimation is a common situation in biological assay and dosage–mortality problems. Calculation of confidence intervals for x_g is possible but more complicated than that for $\hat{y}$.

8.9 WHICH REGRESSION?

Consider bivariate sample (x_1, x_2):

$\hat{x}_2 = \bar{x}_2 + b_{21}(x_1 - \bar{x}_1)$ i.e., the regression of x_2 on x_1

$\hat{x}_1 = \bar{x}_1 + b_{12}(x_2 - \bar{x}_2)$ i.e., the regression of x_1 on x_2

where $b_{21} = \frac{SP(x_1x_2)}{SS(x_1)}$ and $b_{12} = \frac{SP(x_1x_2)}{SS(x_2)}$

NOTE:

a) Both lines intersect at $(\bar{x}_1, \bar{x}_2)$

b) They coincide if $b_{12} = b_{21}^{-1}$ i.e., if $SP(x_1x_2) = \sqrt{SS(x_1)\,SS(x_2)}$

i.e. if $|r| = 1$ (i.e., perfect correlation)

8.9.1 ONE VARIATE SELECTED OR CONTROLLED

Here there is no choice, the controlled (or selected) variate must be the predictor (independent) variate (x) and only the regression of y on x is valid.

8.9.2 RANDOM BIVARIATE SAMPLE

In the case of a random bivariate sample (x_1, x_2) either x_1 or x_2 may play the role of dependent variate and there as thus two regression lines. Both regressions are valid, **but only one is correct for the intended purpose.** There is thus a choice, but only one correct choice.

NOTE:

i) **GENERAL RULE:** Use x_1 on x_2 (now renamed y on x) if it is intended to estimate x_1 from x_2.

ii) The use of a regression of y on x does not in general imply that y depends casually on x.

iii) Regression is a useful way of looking at the interdependence of two variates.

8.10 COEFFICIENT OF DETERMINATION

A useful statistic in the evaluation of a particular model is the **Coefficient of Determination.** The Coefficient of Determination, usually expressed as a percentage, is a measure of the extent to which the variation in y can be accounted for by variation due to the fitted model, i.e., "variation due to regression", thus:

$$\textbf{Coefficient of Determination} = \frac{\text{SS(due to regression)}}{\text{Total SS}} \times 100\% \qquad \ldots \{8.16\}$$

NOTE:

This coefficient may be calculated as, (r^2 x 100%), where r = correlation coefficient ($r_{y,x}$).

since, $$r^2 = \frac{SP^2(x,y)}{SS(x)\,SS(y)} = \frac{\text{SS(due to regression)}}{SS(y)} = \textbf{Coefficient of Determination}$$

Chapter 9: CHI-SQUARE (Analysis of Counts and Proportions)

9.0 INTRODUCTION

In earlier chapters we have been concerned with the analysis of quantitative data, discrete or continuous. In this chapter we will concern ourselves with the analysis of frequencies.

EXAMPLE 9.1: Frequencies of **quantitative** variate values

Yield	No of Plots (Frequency)
6.9 — 8.2	24
8.3 — 9.6	80
9.7 — 11.0	223
11.1 — 12.4	282
12.5 — 13.8	84
13.9 — 15.2	84
15.3 — 16.6	17
16.7 — 18.0	3
	797 = N

NOTE: Number of classes (k) = 8

EXAMPLE 9.2: Frequencies of **qualitative** data – (ie attributes)

(a) Response of cattle to a new vaccine against *Brucellosis*.

	Animals with *Brucellosis*	w/o *Brucellosis*	**Totals**
Standard Vaccine	10	4	14
New Vaccine	5	11	16
Totals	15	15	**30 = N**

NOTE: i) Here we have a `two–way' classification of the data.

ii) **Number of attributes, i.e., classes (k) = 4**

(b) Segregation (Genetics)

Type of Seed (Attribute)	**No of Plants (Frequency)**
Round and yellow	315
Round and green	101
Wrinkled and yellow	108
Wrinkled and green	32
	556 = N

NOTE: Number of attributes, i.e., classes (k) = 4

9.1 CONTINGENCY TABLES

A sample of individuals may be first classified according to one type of attribute and further sub–classified according to a second type of attribute. The result will be a 2–way frequency table called a **CONTINGENCY TABLE.** The marginal frequencies of this table are the classification according to the separate attributes. ref : Example 9.2 (a) above :

Here Attribute 1 = *Type of Vaccine,* – (Standard or New)
Attribute 2 = *Type of response,* – (Success or Failure)

We will consider three types of Contingency Tables, namely : 2 x 2, 2 x C (and R x 2), and (R x C).

Special computational methods will be presented for each type of table.

9.2 PEARSON'S CHI SQUARE (χ^2).

Consider frequency distribution of k classes with observed frequencies, n_1 , n_2 ... n_k, where $\sum n_j=N$, the total number of observations in the sample. Furthermore, let the theoretical or hypothetical probability that an observation should belong to the j^{th} class be p_j, such that $\sum p_j=1$. Then the theoretical or hypothetical expected frequencies in the k classes for a sample size (N) are : Np_1 , Np_2 ... Np_k.

Now $\sum(n_j - np_j)=0$, i.e., sum of the `observed' frequencies is always equal to the sum of the `expected' frequencies. Now the quantity χ^2 (CHI– Square) :

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j} \qquad \ldots \{9.1\}$$

i.e. summed over the k classes, can be used to measure the discrepancy between observation and that expected under a particular hypothesis.

i.e. $\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ often written, $\sum \frac{(O-E)^2}{E}$

NOTE:

a) χ^2 is calculated over all of **k classes**

b) n the sample size is used in the calculation of expected frequencies.

9.2.1 DEGREES OF FREEDOM χ^2

Degrees of Freedom (f) of Pearson's χ^2

= Number of independent deviations $(n_j - np_j)$ squared in calculating χ^2.

i.e. **DF = number of classes (k) — Number of linear restrictions imposed on the deviations.** {9.2}

NOTE :

a) Since $\sum(n_j - np_j) = 0$, f cannot exceed k–1 (It may be considerably less !)

b) DF of χ^2 is based on number of classes **k** , **NOT** the sample size N.

9.2.2 INTERPRETATION OF χ^2

a) Compare calculated value of χ^2 with tabular value for f DF.

b) (i) $p < 0.05$ Reject H_0

(ii) $0.05 < p < 0.95$ Do not reject H_0

(NOTE : $p = 0.50$ – **Ideal**)

c) $p \geq 0.95$ — data is in close agreement with hypothesis; perhaps suspiciously so?

EXAMPLE 9.3: df = 4 and calculated $\chi^2_4 = 6.85$

From tables :

df	P :	50	10	5	2.5	1.0	0.1
4	χ^2 :	3.36	7.78	9.49	11.1	13.1	18.5

i.e $\chi^2 = 6.85$

Conclusion: $0.10 < p < .50$

9.3 THEORETICAL χ^2 DISTRIBUTION

χ^2 is the distribution of $z_1^2+z_2^2+\ldots\ z_f^2$ where $z_1, z_2, \ldots, z_f$ are N.I.D. (0, 1).

The parameter **f** is called the degrees of freedom of the distribution.

NOTE: Some graphical comparisons

a) Standardised NORMAL distribution (z)

Fig 9.1

b) χ^2 Distribution ($\sum z^2$) with 1 and 2 DF.

Fig 9.2

c) χ^2 Distribution ($\sum z^2$) with DF $\geq$ 3

Fig 9.3

9.3.1 NOTES ON THEORETICAL χ^2

a) χ^2 has only positive values (sum of squares).

b) For f = 1, most probable values are near zero.

c) For f > 1, $z^2 > z$, thus having a long tail.

d) For f large, $\chi^2 = \sum z_i^2$ is unlikely to have values close to zero, because all z^2 would have to be simultaneously small. Hence the Mode increases with f and the general shape is *Positive Skew*.

9.3.2 PROPERTIES OF THEORETICAL χ^2

a) χ^2 is a Continuous distribution.

b) It is defined by one parameter, f the degrees of freedom.

c) Mean (μ) = f, Variance (σ^2) = 2f.

d) $\chi_f^2 \rightarrow$ ND as f $\rightarrow \infty$.

9.3.3 RELATIONSHIP BETWEEN PEARSON'S χ^2 AND THEORETICAL χ^2

a) The distribution of Pearson's χ^2 is the same as Theoretical χ^2 if each n_j is ND about mean np_i. In practice, since n_j are integers this cannot happen, and thus Pearson's χ^2 is discontinuous.

b) Discontinuity of Pearson's χ^2 is much less pronounced — **if n is large.**

i.e., Pearson's $\chi^2 \rightarrow$ Theoretical χ^2 n $\rightarrow \infty$ and **provided no p_i is small.**

9.3.4 CONDITIONS FOR A REASONABLY GOOD APPROXIMATION

a) n should be fairly large.

b) No p_j should be small so that any np_j is small, i.e., no **expected frequency** should be small, usually accepted as < 5.

NOTE: If condition (b) is not fulfilled, the distribution of the n_j concerned will tend to be Poisson rather than normal.

c) General Rule : **NO EXPECTED FREQUENCY SHOULD BE < 5**

For certain "goodness of fit" type problems discussed later, this rule is relaxed to ≤ 1.

NOTE :

i) To comply with the *General Rule* it is sometimes necessary to combine adjacent classes. This reduces the DF of the test, with subsequent loss of sensitivity. There are occasions, however, where the combining of classes is not logically possible.

ii) The main risk is that a significantly high value of χ^2 may occur through applying the test when it cannot give an accurate result. In other words, suspect a large χ^2 value which comes mainly from a class with a small expected frequency.

9.4 APPLICATIONS OF CHI– SQUARE

9.4.1 RATIOS/PROPORTIONS GIVEN BY HYPOTHESIS. (Genetics)

EXAMPLE 9.4:
Consider the investigation into whether *seed colour* (yellow or green) is inherited independently of *seed shape* (round or wrinkled).

	Attribute :				
	RY	**WY**	**RG0**	**WG**	**Totals**
Observed frequency (n_j)	315	101	108	32	556
Hypothetical ratio (Mendelian)	9	3	3	1	16
Hypothetical proportion (p_j)	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	1
Expected frequency (np_j)	312.75	104.25	104.25	34.75	556
($n_j - np_j$)	2.25	− 3.25	3.75	- 2.75	0

Thus $\chi^2_{3df} = \left\{ \frac{2.25^2}{312.75} + \frac{(-3.25^2+3.75^2)}{104.25} + \frac{-2.75^2}{34.75} \right\} = 0.4700$ **(0.90 < p < 0.95)**

Conclusion: Data agrees well with the Mendelian hypothesis of a 9 : 3 : 3 : 1 ratio.

9.4.2 RATIOS/PROPORTIONS NOT GIVEN (but determined from Marginal Frequencies)

9.4.2.1. TWO–WAY (R x C) CONTINGENCY TABLES

[Note : R = number of rows, C = number of columns; in this example R, C > 2]

Very often there is no hypothesis about the marginal ratios and instead of regarding the n individuals as one of an infinite number of samples, the observed table is regarded as one of a large finite number of possible tables (with the same marginal frequencies) which could have arisen by chance on the independence hypothesis.

Now, if two attributes are INDEPENDENT, the probabilities corresponding to the various combinations of the two types of attributes in the cells of the table are given by the MULTIPLICATIVE LAW of probabilities.

Remember, the ***Multiplicative Law for Independent Events*** states that : –

If two events A & B are independent, **P (event A and B) = P(event A) × P(event B)**

Consider the following table of frequencies:

Type A \ **Attributes**	**Type B :** 1	2	3	j ...	**Totals**
1	n_{11}	n_{12}		n_{1j}	n_{10}
2	n_{21}				n_{20}
i	n_{i1}			n_{ij}	n_{i0}
Totals	n_{01}	n_{02}		n_{0j}	**n**

Now,

a) $\frac{n_{i0}}{n}$ = Probability that an individual selected at random (from observed n) will have the i^{th} attribute of *Type A*.

b) $\frac{n_{0j}}{n}$ = Probability that an individual (so selected) will have the i^{th} attribute of *Type B*.

c) Thus $\frac{(n_{i0}.n_{0j})}{n^2}$ = Probability that an individual so selected will, on hypothesis of independence, have the combination of these two attributes.

(d) Expected frequency : cell(i,j) = $\frac{n_{i0}.n_{0j}}{n^2}$ x n = $\frac{n_{i0}.n_{0j}}{n}$; compare this with observed frequency n_{ij}

i.e. **EXPECTED FREQUENCY cell (i,j)** = $\frac{\text{Row Total x Column Total}}{\text{Total Frequency}}$ {9.3}

calculated for all combinations of rows and columns totals, i.e., for (r x c) cells (classes).

And $\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$, over all **(r x c)** classes in the contingency table.

NOTE :

When expected frequencies are calculated using marginal totals χ^2 has **(r − 1)(c − 1) df** since there are only (r − 1)(c − 1) independent deviations.

EXAMPLE 9.5 : Independence of Hair and Eye colour (c.f. *Mead and Curnow,* page 226)

	Hair Colour:				
Eye Colour:	**Fair**	**Brown**	**Black**	**Red**	**Total**
Blue	1768	807	189	47	2811
Grey/Green	946	1387	746	53	3132
Brown	115	438	288	16	857
Total	12829	2632	1223	116	6800

QUESTION:

Is the distribution of hair colour the same for each eye colour and vice versa? – i.e., are the attributes INDEPENDENT?Alternatively, do people with blue eyes, for example, tend to have fair hair more often than those with brown eyes in other words is there some ASSOCIATION between attributes?

NOTE:
For any (r x c) Contingency table, INDEPENDENCE means that the proportions of *Type A* attributes are the same for each attribute of *Type B*. If this is NOT the case, the attributes are said to be ASSOCIATED.

Thus H_o : Attributes A and B are Independent.
H_A: Attributes A and B are Associated.

METHOD OF COMPUTATION:

For each cell (i,j) calculate the expected frequency $= \frac{1}{n} \times (R_i C_j)$ where n is the total number of observations and R_i and C_j are the corresponding row and column TOTALS for the $(i,j)^{th}$ cell.

For the above example we thus have :

		Hair Colour				
Eye Colour		**Fair**	**Brown**	**Black**	**Red**	**Total**
	O	1768	807	189	47	2811
Blue	E	1169	1088	506	48	2811
(O–E)		599	-281	-316	-1	
Grey/Green	O	946	1387	746	53	3132
	E	1303	1212	563	53	3132
(O–E)		-357	175	183	0	
Brown	O	115	438	288	16	857
	E	347	332*	154	15	857
(O–E)		-242	106	134	1	
Total		2829	2632	1223	116	6800

NOTE:

For example: $332^* = \frac{2632 \times 857}{6800}$; etc

and, $\chi^2_{(3-1)(4-1)} = \left(\frac{599^2}{1169} + \ldots + \frac{1^2}{15} \right)$

i.e. $\chi^2_6 = 1075$ ($p < 0.01$) i.e., the result is highly significant.

Conclusion:

H_0 is rejected, and we conclude that the attributes of hair and eye colour are, in some way, associated.

9.4.2.2 SPECIAL CASE – (2 x 2) CONTINGENCY TABLE.

Consider 2 x 2 table of frequencies:

			totals
	a	b	a + b
	c	d	c + d
totals	a + c	b + d	a + b + c + d = TOTAL (N)

then $$\chi^2_{1\,df}\{\text{uncorrected}\} = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)} \quad \ldots \{9.4\ a\}$$

$$= \frac{(\text{Difference of crossproducts})^2 \times \text{TOTAL}}{\text{Product of marginal totals}}$$

NOTE:
Probability statements for (2 x 2) tables can be seriously in error, especially for small number of observations and YATES suggested the following '*correction for continuity*'.

$$\chi^2_{1\,df}\{\text{corrected}\} = \frac{\{\,|\,ad-bc\,|\,-\frac{1}{2}\,\text{TOTAL}\}^2 \times \text{TOTAL}}{\text{Product of marginal totals}} \quad \{9.4b\}$$

9.4.2.3 SPECIAL CASE – (2 x C) or (R x 2) CONTINGENCY TABLE :

As in the case of (2 x 2) contingency tables, a `quick' formula is available for the calculation of χ^2 in the case of (r x 2) and (2 x c) contingency tables, viz : –

$$\chi^2 = \frac{\sum(n_{i1}^2/n_{i0}) - n_{01}^2/n}{n_{01}\, n_{02}/\, n^2} \qquad \{9.5\}$$

with df = (r – 1) (c – 1) as before.

This use of formula{9.5} will be demonstrated with the following example. It is to be noted that it is usually more convenient to use the column with the smaller numbers: here the column *`number with mastitis'* is labelled (n_{i1}).

EXAMPLE 9.6:

It is postulated that mastitis is genetically transmitted through the sire. The incidence of mastitis in the progeny of 8 bulls is tabulated below.

Sire	number with mastitis (n_{i1})	number without mastitis (n_{i2})	Totals (n_{i0})
1	6	30	36
2	11	40	51
3	8	26	34
4	6	21	27
5	7	22	29
6	14	42	56
7	6	20	26
8	4	19	23
Totals	62 (= n_{01})	220 (= n_{02})	282 (= n)

H_o: Mastitis is independent of sire.

Applying formula {9.5} we get :

$$\chi^2 = \frac{\left\{(6^2/36 + \ldots + 4^2/23) - 62^2/282\right\}}{(62 \times 220)/282^2} = \frac{(13.8582 - 13.6312)}{0.1715} = \mathbf{1.3232\ (ns)}$$

Conclusion: H_o is accepted, the incidence of mastitis is independent of sire.

9.5 ADDITIVE AND PARTITIVE PROPERTY OF χ^2

9.5.1 PARTITIVE PROPERTY

Any χ^2–variate with f DF may be partitioned into f independent χ^2 – variates each with 1 DF.

Note:

i) This is similar to the complete orthogonal partitioning of *Sums of Squares* in ANOVA.

ii) The partitioning need not be complete, i.e., some component χ^2 may have more than 1 DF.

9.5.2 ADDITIVE PROPERTY

The sum of a number of independent χ^2 variates is also a χ^2 – variate with DF equal to the sum of the individual DF.

EXAMPLE 9.7.

In a series of tests (e.g. tossing of a coin by 6 different people) the following values of χ^2 each with 1 DF were obtained:

	χ^2	DF	P
	3.27	1	0.05 – 0.10
	9.34	1	< 0.01
	6.08	1	0.01 – 0.02
	2.51	1	0.10 – 0.20
	5.61	1	0.01 – 0.02
	1.59	1	0.20 – 0.30
TOTAL χ^2	28.40	6	p < 0.001

CONCLUSION:

Although the individual tests give varying results, TOTAL χ^2 provides convincing evidence against the hypothesis.

NOTE :

a) **DO NOT** confuse the addition of the contributions to χ^2 from the various classes in the application of the formula $\sum^{i} \left(\frac{n_i - np_i}{np_i} \right)^2 = \chi^2$ { c.f. Example 9.5} with the *Additive Property* of χ^2.

b) The *Additive Property of* χ^2 does not make any stipulation that the χ^2 values must have the same DF.

c) TOTAL χ^2 does NOT take any account of the directions of the deviations from the hypothesis within each independent estimate of χ^2.

d) TOTAL χ^2 does not reflect whether or not any deviations from the hypothesis are consistent over the conditions of the individual (independent) experiments.

9.6 χ^2 AS A TEST OF HETEROGENEITY

a) The additive property of χ^2 does not make any stipulation that the χ^2 values must have the same DF.

b) Total χ^2 does take any account of the directions of the deviations from the hypothesis.

c) Total χ^2 does not reflect whether or not any deviations from the hypothesis is consistent over the conditions of the individual (independent) experiments.

d) Often an experiment is carried out under a number of similar (or expressly changed) conditions with a view to testing a hypothesis under such conditions.

Often an experiment is carried out under a number of similar, or expressly changed conditions with a view to testing a hypothesis under such conditions. Thus if we pool the data from all tests in a single frequency table and apply the χ^2 test we need some assurance of **HOMOGENEITY** throughout the series before we can accept agreement with the hypothesis. Such a test is provided by a **"Test of Homogeneity".**

EXAMPLE 9.8 Seed shape segregation (Genetics).

NOTE::

It is often more convenient to use the formula :

$$\chi_1^2 = \frac{(k_2 n_1 - k_1 n_2)^2}{n k_1 k_2}$$

where (k_1 : k_2) is the hypothetical ratio, for the calculation of χ^2 when there are only two classes and **the probabilities are given by hypothesis**

Applying the above method to the data of the Example 9.7 we get:

H_o: Segregation of seed shape according to the hypothesis of : Round : Angular : : 3 : 1, is consistent over all progeny, i.e the results are HOMOGENEOUS.

Plant No.	(n_{i1}) Round seed	(n_{i2}) Angular seed	Totals(n_{i0})	$\chi^2 = \frac{(n_1 - 3n_2)^2}{3N}$	D F
1	45	12	57	0.4737	1
2	27	8	35	0.0857	1
3	24	7	31	0.0968	1
4	19	10	29	1.3908	1
5	32	11	43	0.0078	1
6	26	6	32	0.6667	1
7	88	24	112	0.7619	1
8	22	10	32	0.6667	1
9	28	6	34	0.9804	1
10	25	7	32	0.1667	1
Totals	336 (n_{01})	101	437 (n)	5.2972	10

Pooled χ^2_{1df} = 0.8306 (0.25 < P < 0.50)

Total χ^2_{10df} = 5.2972 (0.75 < P < 0.90)

Heterogeneity χ^2_{9df} = 4.4666 (0.75 < P < 0.90) {Calculated as Total χ^2 − Pooled χ^2 }

Conclusion:: Heterogeneity χ^2_{9df} is not significant, thus there is no reason to reject the hypothesis of homogeneity.

ALTERNATIVE METHOD FOR THE CALCULATION OF HOMOGENEITY χ^2

Homogeneity $\chi^2 = \frac{\sum(n_{i1}^2/n_{i0}) - n_{01}^2/n}{p(1-p)}$ { df = (r − 1) (c − 1) as before}, where p and (1 − p) are given by hypothesis; in this case p = 0.75 and (1 − p) = 0.25

Thus we have, **Homogeneity $\chi^2_9 = \frac{24.180692 - 23.343249}{(0.25 \times 0.75)} = 4.466$** (as calculated above).

INTERPRETATION OF RESULTS :

Total χ^2	= 5.297	(10 df)	$(0.80 < p < 0.90)$
Heterogeneity χ^2	= 4.467	(9 df)	$(0.95 < p < 0.98)$
Pooled χ^2	= 0.830	(1 df)	$(0.25 < p < 0.50)$

NOTE :

a) If **TOTAL χ^2** significant – Hypothesis is rejected

b) **TOTAL χ^2** non – significant – Hypothesis is not rejected.

However, **TOTAL χ^2** does not take into account the signs of the deviations, thus TOTAL χ^2 would be the same for a set of deviations all of the same sign, i.e., **HOMOGENEITY**, as the same set of deviations with varying sign, i.e., **HETEROGENEITY**, whereas, deviations consistently or mainly in one direction will constitute evidence against the hypothesis and this is NOT reflected in **TOTAL χ^2.** Thus a test of **POOLED χ^2** can show up a tendency in one direction not sufficiently pronounced to show up in the individual or **TOTAL χ^2's**.

c) POOLED χ^2 significant – Hypothesis is rejected

d) POOLED χ^2 non – significant – Hypothesis is not rejected

However, if data are heterogeneous, it is possible that this non – significance is due to the 'balancing' effect of positive and negative deviations – even if TOTAL χ^2 is significant.

e) HETEROGENEITY χ^2 significant.

On the assumption that the hypothetical ratios are correct, deviations from it show greater variability than can be accounted for by chance, i.e., the hypothesis cannot be correct for all classes (plants) and it would be necessary to look at individual classes (plants) and their χ^2's for a possible explanation.

f) HETEROGENEITY χ^2 non – significant. i,e, Deviations from hypothetical ratios do not vary significantly.

It should be noted however, that the significance of Pooled χ^2 may still cause rejection of the hypothesis in which case the non–significance of Heterogeneity χ^2 would imply that the (significant) deviation from the hypothesis is consistent over all plants.

FURTHER NOTES

i) Although judgment is sometimes possible on Heterogeneity χ^2 alone, in general it is desirable to make all three tests.

ii) TOTAL χ^2 and POOLED χ^2 are complimentary, i.e., significance of the one and non – significance of the other do not constitute conflicting evidence.

iii) Even if the hypothesis is rejected under (1) or (3) it is still desirable to know whether data are homogeneous or not.

In fact Pooling is not really valid unless the data are Homogeneous

9.6.1 SPECIAL CASE : GENETIC LINKAGE

LINKAGE χ^2 is a test of independence in a 2–way contingency table where the **cell PROBABILITIES ARE GIVEN BY HYPOTHESIS** and NOT obtained from marginal totals.

EXAMPLE 9.9 F_2 frequencies for petal and stigma colour in linseed plants.

Petal Colour :

H_o : Lilac (A) : deep lilac (a) : : 3 : 1

Stigma Colour

H_o : White (B) : purple (b) : : 3 : 1

Calculations:

(n_j)		(np_j)
AB	357	293.06
Ab	37	97.69
aB	33	97.69
ab	94	32.56
	n= 521	521.00

$$\chi^2_3 = \frac{\sum(n_j - np_j)^2}{np_j} = 210.40 \quad (p < 0.001)$$

a) A : a :: 3 : 1 $\quad \chi^2_1 = \frac{(394 - 3 \times 127)^2}{3 \times 521} = 0.108$

b) B : b :: 3 : 1 $\quad \chi^2_1 = \frac{(390 - 3 \times 131)^2}{3 \times 521} = 0.006$

Test	DF	χ^2	P
A : a : : 3 : 1	1	0.108	0.70 – 0.80
B : b : : 3 : 1	1	0.006	0.90 – 0.95
Linkage	1	210.290	<0.001**
Total (9 : 3 : 3 : 1)	3	210.404	<0.001

Conclusion:

a) There is strong evidence of linkage.

b) There is good agreement with 3 : 1 ratio for the separate factors; petal and stigma colour.

9.7 TESTS OF GOODNESS OF FIT

χ^2 can be used to test whether observed distributions are Binomial, Normal, Poisson , etc. It is to be noted that such tests of `Goodness of Fit' with a particular hypothesis are not necessarily the best available but the χ^2 test is usually sufficient.

When `fitting' a hypothetical distribution to sample data the following are to be noted

(a) NORMAL DISTRIBUTION

The normal distribution is defined by two parameters, μ and σ^2. The Normal distribution possesses 2 infinite tails.

(b) POISSON DISTRIBUTION

The Poisson distribution is defined by one parameter, μ. It has one tail which, in theory, has no finite conclusion.

(c) BINOMIAL DISTRIBUTION

The Binomial distribution is defined by two parameters. N and p. The number of class are finite, being determined by N, i.e., number of classes $= (N + 1)$

NOTE:

i) For all three distributions listed above it is possible that for classes beyond a certain point, expected frequencies are likely to be less than 1, i.e., $np_j < 1$, in which case it will be necessary to 'combine classes' as described above, (c.f. Section 9.3.4)

ii) Often parameters of hypothetical distribution are not known and are estimated from sample data, resulting in a further loss of degrees of freedom; see (iv) below.

iii) Minimum expected frequency of 1 may be accepted, if less, pool classes to achieve this end.

iv) DF = (k − 1) − (number of parameters estimated from data), where k is the 'EFFECTIVE NUMBER OF CLASSES' (i.e. after pooling).

EXAMPLE 9.10

The use of χ^2 will be demonstrated with data that is **hypothetically Poisson.**

In an investigation into the distribution of bacteria in samples of blood the numbers of bacterial colonies in squares of a petri dish were counted . The following table gives the observed and expected numbers of squares with varying numbers of colonies.

No. of colonies (class)	Observed N_j	$\hat{p}(x)$ p_j	Expected Np_j	$N_j - Np_j$	$\frac{(N_j - Np_j)^2}{Np_j}$
0	83	0.147096	75.9	7.1	0.664
1	134	0.281935	145.5	-11.5	0.909
2	135	0.270188	139.4	-4.4	0.39
3	101	0.172621	89.1	11.9	1.589
4	40	0.082714	42.7	-2.7	0.171
5	16	0.031707	16.4	-0.4	0.010
6	7	0.010129	5.2	1.8	0.623
7+	0	0.003610	1.9	1.9	1.900
	516	**1.000000**	**516.0 (check)**	**0.0**	**6.005**

i.e., $\chi^2_{6df} = 6.005$ $(0.25 < p < 0.50)$ i.e., non–significant

Conclusion :

There is fairly strong evidence to support the hypothesis that the bacterial colonies have a random Poisson distribution.

NOTES ON CALCULATIONS:

i) For the Poisson distribution, $\hat{p}(x) = \frac{e^{-m} m^x}{x!}$ where $m = \bar{X} = 1.9667$ and $e^{-m} = 0.1470965$
These probabilities are calculated for all classes.

ii) Although no squares with 7 or more colonies were observed there is still a small probability of such an event occurring. This probability is estimated by subtraction.

iii) **In calculating $\hat{p}(x)$ it is important to maintain a significant number of digits, in this case 6.**

iv) The effective number of classes = 8, since in this example it was not necessary to combine classes, the smallest 'expected frequency' being > 1.0.

v) The degrees of freedom for χ^2 is 6 = (8 − 1 − 1), there being a further loss of 1 df since it was necessary to estimate μ (as $\bar{X}$) from the data itself.

hmd/2006