LECTURE NOTE

ON

STATISTICS FOR PHYSICAL SCIENCES & ENGINEERING

(STS 202)

BY

ADEOSUN SAKIRU ABIODUN

E-mail: adeosunsakiru@gmail.com

COURSE CONTENTS

Scope of statistical method in physical sciences and engineering. Measures of location, partition and dispersion. Elements of probability. Probability distribution; binomial, Poisson, geometric, hyper geometric, negative binomial, normal. Estimation (Point and Interval) and tests of hypothesis concerning population means, proportions and variances. Regression and Correlation. Non – parametric tests. Contingency table analysis. Introduction to design of experiments. Analysis of variance.

READING LISTS

1. Adamu S.O and Johnson Tinuke L (1998): Statistics for Beginners; Book 1. SAAL Publication. Ibadan. ISBN: 978-34411-3-2
2. Clark G.M and Cooke D (1993): A Basic course in statistics. Third edition. London: Published by Arnold and Stoughton.
3. Olubosoye O.E, Olaomi J.O and Shittu O.I (2002): Statistics for Engineering, Physical and Biological sciences. Ibadan: A Divine Touch Publications.
4. Tmt. V. Varalakshmi et al (2005): Statistics Higher Secondary - First year. Tamilnadu Textbook Corporation, College Road, Chennai- 600 006

INTRODUCTION

In the modern world of information and communication technology, the importance of statistics is very well recognised by all the disciplines. Statistics has originated as a science of statehood and found applications slowly and steadily in Agriculture, Economics, Commerce, Biology, Medicine, Industry, planning, education and so on. As of today, there is no other human walk of life, where statistics cannot be applied.

Statistics is concerned with the scientific method of collecting, organizing, summarizing, presenting and analyzing statistical information (data) as well as drawing valid conclusion on the basis of such analysis. It could be simply defined as the "science of data". Thus, statistics uses facts or numerical data, assembled, classified and tabulated so as to present significant information about a given subject. Statistic is a science of understanding data and making decisions in the face of randomness.

The study of statistics is therefore essential for sound reasoning, precise judgment and objective decision in the face of up- to- date accurate and reliable data. Thus many researchers, educationalists, business men and government agencies at the national, state or local levels rely on data to answer operations and programs. Statistics is usually divided into two categories, which is not mutually elution namely: Descriptive statistics and inferential statistics.

DESCRIPTIVE STATISTICS

This is the act of summarizing and given a descriptive account of numerical information in form of reports, charts and diagrams. The goal of descriptive statistics is to gain information from collected data. It begins with collection of data by either counting or measurement in an inquiry. It involves the summary of specific aspect of the data, such as averages values and measure of dispersion (spread). Suitable graphs, diagrams and charts are then used to gain understanding and clear interpretation of the phenomenon under investigation

keeping firmly in mind where the data comes from. Normally, a descriptive statistics should:

i. be single – valued
ii. be algebraically tractable
iii. consider every observed value.

INFERENTIAL STATISTICS

This is the act of making deductive statement about a population from the quantities computed from its representative sample. It is a process of making inference or generalizing about the population under certain conditions and assumptions. Statistical inference involves the processes of estimation of parameters and hypothesis testing.

Statistics in Physical Sciences

Physical sciences such as Chemistry, Physics, Meteorology and Astronomy are based on statistical concepts. For example, it is evident that the pressure exerted by a gas is actually an average pressure, an average effect of forces exerted by individual molecules as they strike the wall of a container. The modern science of meteorology is to a great degree dependent upon statistical methods for its existence. The methods that give weather forecasting the accuracy it has today have been developed using modern sample survey techniques.

Statistics in Engineering

Statistics also plays an important role in Engineering. For example, such topics as the study of heat transfer through insulating materials per unit time, performance guarantee testing programs production control, inventory control, standardization of fits and tolerances of machine parts, job analysis of technical personnel, studies involving the fatigue of metals (endurance properties), corrosion studies, quality control, reliability analysis and many other specialized

problems in research and development make great use of probabilistic and statistical methods.

Data can be described as a mass of unprocessed information obtained from measurement of counting of a characteristics or phenomenon. They are raw facts that have to be processed in numerical form they are called **quantitative data**. For instance the collection of ages of students offering STS 202 in a particular session is an example of this data. But when data are not presented in numerical form, they are called **qualitative data**. E.g.: status, sex, religion, etc.

SOURCES OF STATISTICAL DATA

1. **Primary data**: These are data generated by first hand or data obtained directly from respondents by personal interview, questionnaire, measurements or observation. Statistical data can be obtained from:
   (i) Census – complete enumeration of all the unit of the population
   (ii) Surveys – the study of representative part of a population
   (iii) Experimentation – observation from experiment carried out in laboratories and research center.
   (iv) Administrative process e.g. Record of births and deaths.

   ADVANTAGES

   - ✓ Comprises of actual data needed
   - ✓ It is more reliable with clarity
   - ✓ Comprises a more detail information

DISADVANTAGES

- ➢ Cost of data collection is high
- ➢ Time consuming
- ➢ There may larger range of non response

2. **Secondary data**: These are data obtained from publication, newspapers, and annual reports. They are usually summarized data used for purpose other than the intended one. These could be obtain from the following:
   (i) Publication e.g. extract from publications
   (ii) Research/Media organization
   (iii) Educational institutions

ADVANTAGES

- ✓ The outcome is timely
- ✓ The information gathered more quickly
- ✓ It is less expensive to gather.

DISADVANTAGES

- ➢ Most time information are suppressed when working with secondary data
- ➢ The information may not be reliable

METHODS OF COLLECTION OF DATA

There are various methods we can use to collect data. The method used depends on the problem and type of data to be collected. Some of these methods include:

1. Direct observation
2. Interviewing
3. Questionnaire
4. Abstraction from published statistics.

DIRECT OBSERVATION

Observational methods are used mostly in scientific enquiry where data are observed directly from controlled experiment. It is used more in the natural

sciences through laboratory works than in social sciences. But this is very useful studying small communities and institutions.

INTERVIEWING

In this method, the person collecting the data is called the interviewer goes to ask the person (interviewee) direct questions. The interviewer has to go to the interviewees personally to collect the information required verbally. This makes it different from the next method called questionnaire method.

QUESTIONNAIRE

A set of questions or statement is assembled to get information on a variable (or a set of variable). The entire package of questions or statement is called a questionnaire. Human beings usually are required to respond to the questions or statements on the questionnaire. Copies of the questionnaire can be administered personally by its user or sent to people by post. Both interviewing and questionnaire methods are used in the social sciences where human population is mostly involved.

ABSTRACTIONS FROM THE PUPLISHED STATISTICS

These are pieces of data (information) found in published materials such as figures related to population or accident figures. This method of collecting data could be useful as preliminary to other methods.

Other methods includes: Telephone method, Document/Report method, Mail or Postal questionnaire, On-line interview method, etc.

PRESENTATION OF DATA

When raw data are collected, they are organized numerically by distributing them into classes or categories in order to determine the number of individuals belonging to each class. Most cases, it is necessary to present data in tables, charts and diagrams in order to have a clear understanding of the data, and to illustrate the relationship existing between the variables being examined.

FREQUENCY TABLE

This is a tabular arrangement of data into various classes together with their corresponding frequencies.

Procedure for forming frequency distribution

Given a set of observation $x_1, x_2, x_3, \dots, x_n$, for a single variable.

1. Determine the range (R) = L – S where L = largest observation in the raw data; and S = smallest observation in the raw data.
2. Determine the appropriate number of classes or groups (K). The choice of K is arbitrary but as a general rule, it should be a number (integer) between 5 and 20 depending on the size of the data given. There are several suggested guide lines aimed at helping one decided on how many class intervals to employ. Two of such methods are:
   (a) K = 1 +3.322 $(\log_{10} n)$
   (b) K = $\sqrt{n}$ where n = number of observations.
3. Determine the width (w) of the class interval. It is determined as $w = \frac{R}{K}$
4. Determine the numbers of observations falling into each class interval i.e. find the class frequencies.

NOTE: With advent of computers, all these steps can be accomplishes easily.

SOME BASIC DEFINITIONS

Variable: This is a characteristic of a population which can take different values. Basically, we have two types, namely: continuous variable and discrete variable.

A **continuous variable** is a variable which may take all values within a given range. Its values are obtained by measurements e.g. height, volume, time, exam score etc.

A **discrete variable** is one whose value change by steps. Its value may be obtained by counting. It normally takes integer values e.g. number of cars, number of chairs.

Class interval: This is a sub-division of the total range of values which a (continuous) variable may take. It is a symbol defining a class E.g. 0-9, 10-19 etc. there are three types of class interval, namely: Exclusive, inclusive and open-end classes method.

Exclusive method:

When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class; it is known as the exclusive method of classification. E.g. Let some expenditures of some families be as follows:

0 – 1000, 1000 – 2000, etc. It is clear that the exclusive method ensures continuity of data as much as the upper limit of one class is the lower limit of the next class. In the above example, there are so families whose expenditure is between 0 and 999.99. A family whose expenditure is 1000 would be included in the class interval

1000-2000.

Inclusive method:

In this method, the overlapping of the class intervals is avoided. Both the lower and upper limits are included in the class interval. This type of classification may be used for a grouped frequency distribution for discrete variable like members in a family, number of workers in a factory etc., where the variable may take only integral values. It cannot be used with fractional values like age, height, weight

etc. *In case of continuous variables, the exclusive method should be used. The inclusive method should be used in case of discrete variable.*

Open end classes:

A class limit is missing either at the lower end of the first class interval or at the upper end of the last class interval or both are not specified. The necessity of open end classes arises in a number of practical situations, particularly relating to economic and medical data when there are few very high values or few very low values which are far apart from the majority of observations.

Class limit: it represents the end points of a class interval. {Lower class limit & Upper class limit}. A class interval which has neither upper class limit nor lower class limit indicated is called an open class interval e.g. "less than 25", '25 and above"

Class boundaries: The point of demarcation between a class interval and the next class interval is called boundary. For example, the class boundary of 10-19 is 9.5 – 19.5

Cumulative frequency: This is the sum of a frequency of the particular class to the frequencies of the class before it.

Example 1: The following are the marks of 50 students in STS 102:

48 70 60 47 51 55 59 63 68 63 47 53 72 53 67 62 64 70 57 56 48 51 58 63 65 62 49 64 53 59 63 50 61 67 72 56 64 66 49 52 62 71 58 53 63 69 59 64 73 56.

(a) Construct a frequency table for the above data.

(b) Answer the following questions using the table obtained:

(i) how many students scored between 51 and 62?

(ii) how many students scored above 50?

(iii) what is the probability that a student selected at random from the class will score less than 63?

Solution:

(a) Range (R) = $73 - 47 = 26$

No of classes $(k) = \sqrt{n} = \sqrt{50} = 7.07 \approx 7$

Class size $(w) = {}^{26}/_{7} = 3.7 \approx 4$

Frequency Table

Mark	Tally	frequency
$47 - 50$	𝍸 ||	7
51- 54	𝍸 ||	7
55 - 58	𝍸 ||	7
59 – 62	𝍸 |||	8
63 – 66	𝍸 𝍸 |	11
67 – 70	𝍸 |	6
71 – 74	||||	4
		$\sum f = 50$

(b) (i) 22 (ii) 43 (iii) 0.58

Example 2: The following data represent the ages (in years) of people living in a housing estate in Abeokuta.

18 31 30 6 16 17 18 43 2 8 32 33 9 18 33 19 21 13 13 14
14 6 52 45 61 23 26 15 14 15 14 27 36 19 37 11 12 11
20 12 39 20 40 69 63 29 64 27 15 28.

Present the above data in a frequency table showing the following columns; class interval, class boundary, class mark (mid-point), tally, frequency and cumulative frequency in that order.

Solution:

Range (R)$= 69 - 2 = 67$

No of classes $(k) = \sqrt{n} = \sqrt{50} = 7.07 \approx 7.00$

Class width $(w) = {}^{R}/_{k} = {}^{67}/_{7} = 9.5 \approx 10$

Class interval	Class boundary	Class mark	Tally	Frequency	Cum.freq
2 – 11	1.5 – 11.5	6.5	𝍸 ||	7	7
12 – 21	11.5 – 21.5	16.5	𝍸 𝍸 𝍸 |	21	28
22 – 31	21.5 – 31.5	26.5	𝍸 |||	8	36
32 – 41	31.5 – 41.5	36.6	𝍸 ||	7	43
42 – 51	41.5 – 51.5	46.5	||	2	45
52 – 61	51.5 – 61.5	56.5	||	2	47
62 – 71	61.5 – 71.5	66.5	|||	3	50

Observation from the Table

The data have been summarized and we now have a clearer picture of the distribution of the ages of inhabitants of the Estate.

Exercise 1

Below are the data of weights of 40students women randomly selected in Ogun state. Prepare a table showing the following columns; class interval, frequency, class boundary, class mark, and cumulative frequency.

96 84 75 80 64 105 87 62 105 101 108 106 110 64 105 117 103 76 93 75 110 88 97 69 94 117 99 114 88 60 98 77 96 96 91 73 82 81 91 84

Use your table to answer the following question

i. How many women weight between 71 and 90?
ii. How many women weight more than 80?
iii. What is the probability that a woman selected at random from Ogun state would weight more than 90?

MEASURES OF LOCATION

These are measures of the centre of a distribution. They are single values that give a description of the data. They are also referred to as measure of central tendency. Some of them are arithmetic mean, geometric mean, harmonic mean, mode, and median.

THE ARITHMETIC MEAN (A.M)

The arithmetic mean (average) of set of observation is the sum of the observation divided by the number of observation. Given a set of a numbers $x_1, x_2 \ldots, x_n$, the arithmetic mean denoted by $\bar{X}$ is defined by

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Example 1: The ages of ten students in STS 102 are 16,20,19,21,18,20,17,22,20,17, determine the mean age.

Solution: $\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$

$$= \frac{16+20+19+21+18+20+17+22+20+17}{10}$$

$$= \frac{190}{10} = 19 years.$$

If the numbers $x_1, x_2, \ldots, x_n$ *occur* $f_1, f_2, f_3, \ldots f_n$ times respectively, the

$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$ (or $\frac{\sum fx}{n}$ for short.)

Example 2: Find the mean for the table below

Scores (x)	2	5	6	8
Frequency (f)	1	3	4	2

Solution $\bar{x} = \frac{\sum fx}{\sum f} = \frac{(1\times2)+(3\times5)+(4\times6)+(2\times8)}{1+3+4+2}$

$$= \frac{57}{10} = 5.7 \quad .$$

Calculation of mean from grouped data

If the items of a frequency distribution are classified in intervals, we make the assumption that every item in an interval has the mid-values of the interval and we use this midpoint for x.

Example 3: The table below shows the distribution of the waiting items for some customers in a certain petrol station in Abeokuta.

Waiting time(in mins)	1.5 – 1.9	2.0 – 2.4	2.5 – 2.9	3.0 – 3.4	3.5 – 3.9	4.0 – 4.4
No. of customers	3	10	18	10	7	2

Find the average waiting time of the customers.

Solution:

Waiting (in min)	No of customers	Class mark mid-value(X)	fx
1.5 – 1.9	3	1.7	5.1
2.0 – 2.4	10	2.2	22
2.5 – 2.9	18	2.7	48.6
3.0 – 3.4	10	3.2	32
3.5 – 3.9	7	3.7	25.9
4.0 – 4.4	2	4.2	8.4
	$\sum f = 50$		$\sum fx = 142$

$$\bar{X} = \frac{\sum fx}{\sum f}$$

$$= \frac{142}{50} = 2.84$$

Use of Assume mean

Sometimes, large values of the variable are involve in calculation of mean, in order to make our computation easier, we may assume one of the values as the mean. This if A= assumed mean, and d= deviation of x from A, i.e. $d = x - A$

Therefore, $\bar{X} = \frac{\sum fx}{n} = \frac{\sum f(A+d)}{n}$

$$= \frac{\sum fA}{n} + \frac{\sum fd}{n}$$

$$= \frac{A\sum f}{n} + \frac{\sum fd}{n}$$

$$= A + \frac{\sum fd}{n} \text{ since } \sum f = n.$$

If a constant factor C is used then

$$\bar{X} = A + \left(\frac{\sum fU}{\sum f}\right) C \quad where\ U = \frac{x-A}{C}.$$

Example 4: The exact pension allowance paid (in Nigeria) to 25 workers of a company is given in the table below.

Pension in ₦	25	30	35	40	45
No of person	7	5	6	4	3

Calculate the mean using an assumed mean 35.

Solution

Pension in ₦	frequency	$d = x - A$	fd
25	7	25 – 35 = -10	- 70
30	5	30 – 35 = -5	- 25
35 A	6	35 – 35 = 0	0
40	4	40 – 35 = 5	20
45	3	45 – 35 = 10	30
	25		-45

$$\bar{X} = A + \frac{\sum fd}{n}$$

$$= 35 + \left(\frac{-45}{25}\right)$$

$$= 35 - 1.8$$

$$= 33.2$$

Example 5: Consider the data in example 3, using a suitable assume mean, compute the mean.

Solution:

Waiting time	f	x	$d = x - A$	fd
1.5 – 1.9	3	1.7	-1	-3
2.0 – 2.4	10	2.2	-0.5	-5
2.5 – 2.9	18	2.7 A	0	0
3.0 – 3.4	10	3.2	0.5	5
3.5 – 3.9	7	3.7	1	7
4.0 – 4.4	2	4.2	1.5	3
	50			7

$$\bar{X} = A + \frac{\sum fd}{\sum f}$$

$$= 2.7 + \frac{7}{50}$$

$$= 2.7 + 0.14$$

$$= 2.84$$

NOTE: It is always easier to select the class mark with the longest frequency as the assumed mean.

ADVANTAGE OF MEAN

The mean is an average that considers all the observations in the data set. It is single and easy to compute and it is the most widely used average.

DISAVANTAGE OF MEAN

Its value is greatly affected by the extremely too large or too small observation.

THE HARMONIC MEAN (H.M)

The H.M of a set of numbers $x_1, x_2, \dots, x_n$ is the reciprocal of the arithmetic mean of the reciprocals of the numbers. It is used when dealing with the rates of the type x per d (such as kilometers per hour, Naira per liter). The formula is expressed thus:

$$\text{H.M} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

If x has frequency$f$, then

$$\text{H.M} = \frac{n}{\sum\frac{f}{x}}$$

Example: Find the harmonic mean of 2,4,8,11,4.

Solution:

$$\text{H.M} = \frac{5}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{11}+\frac{1}{4}} = \frac{5}{\frac{107}{80}} = 4\frac{12}{107} = 4.112\ .$$

Note:

(i) Calculation takes into account every value

(ii) Extreme values have least effect

(iii) The formula breaks down when "o" is one of the observations.

THE GEOMETRIC MEAN(G.M)

The G.M is an analytical method of finding the average rate of growth or decline in the values of an item over a particular period of time. The geometric mean of a set of number $x_1, x_2, \dots, x_n$ is the nth root of the product of the number. Thus

$$\text{G.M} = \sqrt[n]{(x_1 \times x_2 \times \dots \times x_n)}$$

If f_i is the frequency of x_i, then

$$\text{G.M} = \sqrt[n]{(x_1 f_1 \times x_2 f_2 \times \dots \times x_i f_i)}$$

Example: The rate of inflation in fire successive year in a country was 5%, 8%, 12%, 25% and 34%. What was the average rate of inflation per year?

Solution:

$$\text{G.M} = \sqrt[5]{(1.05) \times (1.08) \times (1.12) \times (1.25) \times (1.34)}$$
$$= \sqrt[5]{2.127384}$$
$$= 1.16$$

$\therefore$ Average rate of inflation is 16%

Note: (1) Calculate takes into account every value.

(2) It cannot be computed when "0" is on of the observation.

Relation between Arithmetic mean, Geometric and Harmonic

In general, the geometric mean for a set of data is always less than or equal to the corresponding arithmetic mean but greater than or equal to the harmonic mean.

That is, $\text{H.M} \leq \text{G.M} \leq \text{A.M}$

The equality signs hold only if all the observations are identical.

THE MEDIAN

This is the value of the variable that divides a distribution into two equal parts when the values are arranged in order of magnitude. If there are n (odd) observation, the median $\tilde{X}$ is the center of observation in the ordered list. The location of the median is $\tilde{X} = \frac{(n+1)}{2}$ th item.

But if n is even, the median $\tilde{X}$ is the average of the two middle observations in the ordered list.

i.e. $$\tilde{X} = \frac{X_{\left(\frac{n}{2}\right)th} + X_{\left(\frac{n}{2}+1\right)th}}{2}$$

Example 1: The values of a random variable x are given as $8, 5, 9, 12, 10, 6$ and 4. Find the median.

Solution: In an array: 4,5,6,8,9,10,12. n is odd, therefore

The median, $\tilde{X} = X_{\left(\frac{n+1}{2}\right)}$ th

$= X_{4\ \ t\text{h}}$

$= 8$

Example 2: The value 0f a random variable x are given as 15, 15, 17, 19, 21, 22, 25, and 28. Find the median.

Solution: n is odd.

Median, $\tilde{X} = X_{\left(\frac{n}{2}\right)^{th}} + X_{\left(\frac{n}{2}+1\right)^{th}}$

$= \frac{X_4+X_5}{2}$

$= \frac{19+21}{2}$

$= 20$

Calculation of Median from a grouped data

The formula for calculating the median from grouped data is defined as

$$\tilde{X} = L_1 + \left(\frac{\frac{n}{2}-Cf_b}{f_m}\right)w$$

Where: $L_1 =$ Lower class boundary of the median class

$N = \sum f =$ Total frequency

$Cf_b =$ Cumulative frequency before the median class

$f_m =$ Frequency of the median class.

$w =$ Class size or width.

Example3: The table below shows the height of 70 men randomly selected at Sango Ota.

Height	118-126	127-135	136-144	145-153	154-162	163-171	172-180
No of rods	8	10	14	18	9	7	4

Compute the median.

Solution

Height	Frequency	Cumulative frequency
118 – 126	8	8
127 – 135	10	18
136 – 144	14	32
145 – 153	18	50
154 – 162	9	59
163 – 171	7	66
172 – 180	4	70
	70	

$\frac{n}{2} = \frac{70}{2} = 30$. The sum of first three classes frequency is 32 which therefore means that the median lies in the fourth class and this is the median class. Then

$L_1 = 144.5,\ \ n = 70, cf_b = 32, w = 9$

$$\tilde{X} = L_1 + \left(\frac{\frac{n}{2} - cf_b}{f_m}\right)w$$

$$= 144.5 + \left[\frac{35-32}{18}\right] \times 9$$

$$= 144.5 + \left(\frac{3\times 9}{18}\right)$$

$$= 144.5 + 1.5 = 136.$$

ADVANTAGE OF THE MEDIAN

(i) Its value is not affected by extreme values; thus it is a resistant measure of central tendency.

(ii) It is a good measure of location in a skewed distribution

DISAVANTAGE OF THE MEDIAN

1) It does not take into consideration all the value of the variable.

THE MODE

The mode is the value of the data which occurs most frequently. A set of data may have no, one, two or more modes. A distribution is said to be uni-model, bimodal and multimodal if it has one, two and more than two modes respectively.

E .g: The mode of scores 2, 5, 2, 6, 7 is 2.

Calculation of mode from grouped data

From a grouped frequency distribution, the mode can be obtained from the formula.

$$\text{Mode, } \hat{X} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right) w$$

Where: L_{mo} = lower class boundary of the modal class

Δ_1 = Difference between the frequency of the modal class and the class before it.

Δ_2 = Difference between the frequency of the modal class and the class after it.

w = Class size.

Example: For the table below, find the mode.

Class	11 – 20	21 – 30	31 – 40	41 - 50	51 – 60	61 – 70
frequency	6	20	12	10	9	9

Calculate the modal age.

Solution: $L_{mo} = 20.5, \Delta_1 = 20 - 6 = 14, \Delta_2 = 20 - 12 = 8,$

$w = 30.5 - 21.5 = 10$

$$\text{Mode, } \hat{X} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right) w$$

$$= 20.5 + \left(\frac{14}{14+8}\right) 10$$

$$= 20.5 + \left(\frac{14}{22}\right) 10$$

$$= 20.5 + (0.64)10$$

$$= 20.5 + 6.4$$

$$= 26.9$$

ADVANTAGE OF THE MODE

1) It is easy to calculate.

DISADVANTAGE OF THE MODE

(i) It is not a unique measure of location.

(ii) It presents a misleading picture of the distribution.

(iii) It does not take into account all the available data.

Exercise 2

1. Find the mean, median and mode of the following observations: 5, 6,10,15,22,16,6,10,6.
2. The six numbers 4, 9,8,7,4 and Y, have mean of 7. Find the value of Y.
3. From the data below

Class	21 – 23	24 – 26	27 – 29	30 – 32	33 – 35	36 – 38	37 – 41
Frequency	2	5	8	9	7	3	1

Calculate the (i)Mean (ii)Mode (iii) Median

MEASURES OF PARTITION

From the previous section, we've seen that the median is an average that divides a distribution into two equal parts. So also these are other quantity that divides a set of data (in an array) into different equal parts. Such data must have been arranged in order of magnitude. Some of the partition values are: the quartile, deciles and percentiles.

THE QUARTILES

Quartiles divide a set of data in an array into four equal parts.

For ungrouped data, the distribution is first arranged in ascending order of magnitude.

Then

First Quartiles: $Q_1 = \left(\frac{N+1}{4}\right)^{th}$

Second Quartile: $Q_2 = 2\left(\frac{N+1}{4}\right)th = median$

Third Quartile: $Q_3 = 3\left(\frac{N+1}{4}\right)th$ member of the distribution

For a grouped data

$$Q_i = L_{qi} + \left(\frac{\frac{iN}{4} - Cf_{qi}}{f_{qi}}\right) \times w$$

Where

$i =$ The quality in reference

$L_{qi=}$ Lower class boundary of the class counting the quartile

$N =$ Total frequency

$Cf_{qi} =$ Cumulative frequency before the Q_i class

$f_{qi} =$ The frequency of the Q_i class

$w =$ Class size of the Q_i class.

DECILES

The values of the variable that divide the frequency of the distribution into ten equal parts are known as deciles and are denoted by $D_1, D_2, \dots, D_9$. the fifth deciles is the median.

For ungrouped data, the distribution is first arranged in ascending order of magnitude. Then

$D_1 = 2\left(\frac{n+1}{10}\right) th$ member of the distribution

$D_2 = 2\left(\frac{n+1}{10}\right) th\ member\ of\ the\ distribution$

$D_9 = 9\left(\frac{n+1}{10}\right) th\ member\ of\ the\ distribution$

For a grouped data

$$D_i = L_{Di} + \left(\frac{\frac{iN}{10} - Cf_{Di}}{F_{Di}}\right) w \qquad i = 1, 2, \dots, 9$$

Where $i = Decile\ in\ reference$

$L_{Di} = lower\ class\ boundary\ of\ the\ class\ counting\ the\ decile$

$N = Total\ frequency$

$Cf_{Di} = cumulative\ frequency\ up\ to\ the\ low\ boundary\ of\ the\ D_i\ class$

$F_{Di} = the\ frequency\ of\ the\ D_i\ class$

$w = Class\ size\ of\ the\ D_i\ class.$

PERCENTILE

The values of the variable that divide the frequency of the distribution into hundred equal parts are known as percentiles and are generally denoted by $P_{1,\dots,}P_{99}$.

The fiftieth percentile is the median.

For ungrouped data, the distribution is first arranged in ascending order of magnitude. Then

$P_1 = \left(\frac{n+1}{100}\right) th$ member of the distribution

$P_2 = \frac{2(n+1)}{100} th$ member of the distribution

$P_{99} = \frac{99(n+1)}{100} th$ member of the distribution

For a grouped data,

$$P_i = L_{pi} + \left(\frac{\frac{iN}{100} - Cf_{pi}}{f_{pi}}\right) \times w \qquad i = 1, \ldots., 99$$

Where

$i = percentile\ in\ reference$

$L_{pi} = Lower\ class\ boundary\ of\ the\ class\ counting\ the\ percentile$

$N = Total\ frequency$

$Cf_{pi} = Cumulative\ frequency\ up\ to\ the\ lower\ class\ boundary\ of\ the\ P_i\ class$

$f_{pi} = Frequency\ of\ the\ p_i\ class.$

Example: For the table below, find by calculation (using appropriate expression)

(i) Lower quartile, Q_1

(ii) Upper Quartile, Q_3

(iii) 6th Deciles, D_6

(iv) 45th percentile of the following distribution

Mark	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
Frequency	8	10	14	26	20	16	4	2

Solution

Marks	frequency	cumulative frequency
20 – 29	8	8
30 – 39	10	18
40 – 49	14	32
50 – 59	26	32
60 – 69	20	58
70 – 79	16	78
80 – 89	4	98
90 – 99	2	100
	100	

(i) Lower quartile, $Q_1 = L_{q1} + \left(\frac{\frac{iN}{4} - Cf_{q1}}{f_{q_1}}\right)$w

$\frac{iN}{4} = \frac{1 \times 100}{4} = 25, Cf_{q_1} = 18, f_{q_1} = 14, w = 10, L_{q_1} = 39.5$

$Q_1 = 39.5 + \left(\frac{25 - 18}{14}\right)10$

$= 44.5$

(ii) Upper Quartile, $Q_3 = L_{q_3} + \left(\frac{\frac{3N}{4} - Cf_q}{f_{q3}}\right)w$

$\frac{3N}{4} = \frac{3 \times 100}{4} = 75, L_{q_3} = 59.5, Cf_{q_3} = 58, F_{q_3} = 20, w = 10$

$Q_3 = 59.5 + \left(\frac{75 - 58}{20}\right)10 = 68$

(iii) $D_6 = L_{D_6} + \left(\frac{\frac{6N}{10} - Cf_{D6}}{f_{D_6}}\right)w$

$\frac{6N}{10} = \frac{6 \times 100}{10} = 60, L_{D_6} = 59.5, Cf_{D_6} = 58, f_{D_6} = 20, w = 10$

$D_6 = 59.5 + \left(\frac{60 - 58}{20}\right)10 = 60.5$

(iv) $P_{45} = L_{p_{45}} + \left(\frac{\frac{45N}{10} - Cf_{p45}}{f_{p45}}\right)w$

$\frac{45N}{100} = \frac{45 \times 100}{100} = 45, L_{p_{45}} = 49.5, Cf_{p45} = 32, f_{p45} = 26, w = 10$

$$P_{45} = 49.5 + \left(\frac{45-32}{26}\right)10$$

$$= 49.5 + 5$$

$$= 54.5$$

MEASURES OF DISPERSION

Dispersion or variation is degree of scatter or variation of individual value of a variable about the central value such as the median or the mean. These include range, mean deviation, semi-interquartile range, variance, standard deviation and coefficient of variation.

THE RANGE

This is the simplest method of measuring dispersions. It is the difference between the largest and the smallest value in a set of data. It is commonly used in statistical quality control. However, the range may fail to discriminate if the distributions are of different types.

$Coefficient\ of\ Range = \frac{L-S}{L+S}$

SEMI – INTERQUARTILE RANGE

This is the half of the difference between the first (lower) and third quartiles (upper). It is good measure of spread for midrange and the quartiles.

$$S.I.R = \frac{Q_3 - Q_1}{2}$$

THE MEAN/ABSOLUTE DEVIATION

Mean deviation is the mean absolute deviation from the centre. A measure of the center could be the arithmetic mean or median.

Given a set of $x_1, x_2, \dots, x_{n,}$ the mean deviation from the arithmetic mean is defined by:

$$MD = \frac{\sum_{i=1}^{n} |x_{i-\bar{x}}|}{N}$$

In a grouped data

$$MD_{\bar{x}} = \frac{\sum_{i=1}^{n} f|X_i - \bar{X}|}{\sum_{i=1}^{n} f_i}$$

Example1: Below is the average of 6 heads of household randomly selected from a country. 47, 45, 56, 60, 41, 54 .Find the

(i) Range

(ii) Mean

(iii) Mean deviation from the mean

(iv) Mean deviation from the median.

Solution:

(i) Range $= 60 - 41 = 19$

(ii) Mean $(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$= \frac{47+45+56+60+41+54}{6}$$

$$= \frac{303}{6} = 50.5$$

(iii) Mean Deviation $(MD_{\bar{x}}) = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$

$$= \frac{|47-50.5|+|45-50.5|+|56-50.5|+|60-50.5|+|41-50.8|+|54-50.5|}{6}$$

$$= \frac{|-3.5|+|-5.5|+|5.5|+|9.5|+|-9.5|+|3.5|}{6}$$

$$= \frac{37}{6}$$

$$= 6.17$$

(iv) In array: 41, 45, 47, 54, 56, 60

$$\text{Median} = \frac{X_{\left(\frac{n}{2}\right)th} + X_{\left(\frac{n}{2}+1\right)th}}{2}$$

$$= \frac{X_3 + X_4}{2} = \frac{47+54}{2} = 50.5$$

$$MD_{\bar{x}} = \frac{|47-50.5|+|45-50.5|+|56-50.5|+\cdots+|54-50.5|}{6}$$

$$= 6.17$$

Example2: The table below shown the frequency distribution of the scores of 42 students in MTS 201

Scores	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69
No of student	2	5	8	12	9	5	1

Find the mean deviation from the mean for the data.

Solution:

Classes	midpoint x	f	fx	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
0 – 9	4.5	2	9	-29.52	29.52	59.04
10 – 19	14.5	5	72.5	-19.52	19.52	97.60
20 – 29	24.5	8	196	-9.52	9.52	76.16
30 – 39	34.5	12	414	0.48	0.48	5.76
40 – 49	44.5	9	400.5	10.48	10.48	94.32
50 – 59	54.5	5	272.5	20.48	20.48	102.4
60 – 69	64.5	1	64.5	30.48	30.48	30.48
		42	1429			465.76

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1429}{42} \approx 34.02$$

$$MD_{\bar{x}} = \frac{\sum_{i=1}^{n} f|x_i - \bar{x}|}{\sum_{i=1}^{n} f_i} = \frac{465.70}{42}$$

$$= 11.09$$

THE STANDARD DEVIATION

The standard deviation, usually denoted by the Greek alphabet σ (small signal) (for population) is defined as the "positive square root of the arithmetic mean of the squares of the deviation of the given observation from their arithmetic mean". Thus, given $x_1, \ldots, x_n$ as a set of n observations, then the standard deviation is given by:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \quad where\ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(Alternatively, $\sigma = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2}$.)

For a grouped data

The standard deviation is computed using the formula

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_1\,(X_i - \bar{X})^2}{\sum_{i=1}^{n} f_i}} \qquad or \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i X_i}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^2}$$

If A= assume mean and $d = x - A$ is deviation from the assumed mean, then

$$\sigma = \sqrt{\frac{\sum f d^2}{\sum f} - \left(\frac{\sum f d}{\sum f}\right)^2}$$

Note: We use $S = \sqrt{\frac{\sum f(X_i - \bar{X})^2}{\sum f - 1}}$ when using sample instead of the population to obtain the standard deviation.

MERIT

(i) It is well defined and uses all observations in the distribution.

(ii) It has wider application in other statistical technique like skewness, correlation, and quality control e.t.c

DEMERIT

(i) It cannot be used for computing the dispersion of two or more distributions given in different unit.

THE VARIANCE

The variance of a set of observations is defined as the square of the standard deviation and is thus given by σ^2

COEFFICIENT OF VARIATION/DISPERSION

This is a dimension less quantity that measures the relative variation between two servers observed in different units. The coefficients of variation are obtained by dividing the standard deviation by the mean and multiply it by 100. Symbolically

$$CV = \frac{\sigma}{\bar{X}} \times 100\ \%$$

The distribution with smaller C.V is said to be better.

EXAMPLE3: Given the data 5, 6, 9, 10, 12. Compute the variance, standard deviation and coefficient of variation

SOLUTION

$$\bar{X} = \frac{5+6+9+10+12}{5} = 8.4$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}$$

$$= \frac{(5-8.4)^2+(6-8.4)^2+(9-8.4)^2+(10-8.4)^2+(12-8.4)^2}{5}$$

$$= \frac{11.56+5.76+0.36+2.56+12.96}{5}$$

$$= {}^{33.2}/_{5}$$

$$= 6.64$$

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

$$= \sqrt{6.64}$$

$$= 2.58$$

Hence $\text{C.V} = \frac{2.58}{8.4} \times 100$

$$= 30.71\%$$

EXAMPLE4: Given the following data. Compute the

(i) Mean

(ii) Standard deviation

(iii) Coefficient variation.

Ages(in years)	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74	75 – 79	80 – 84
Frequency	1	2	10	12	18	25	9

SOLUTION

Classes	x	F	fx	$x-\bar{x}$	$f(x-\bar{x})^2$
50 – 54	52	1	52	-20.06	402.40
55 – 59	57	2	114	-15.06	453.61
60 – 64	62	10	620	-10.06	1012.04
65 – 69	67	12	804	-5.06	307.24
70 – 74	72	18	1296	-0.06	0.07
75 – 79	77	25	1925	4.94	610.09
80 – 84	82	9	738	9.94	889.23
		77	5549		3674.68

$$\bar{X} = \frac{\sum fx}{\sum f} = \frac{5549}{77} = \quad 72.06$$

$$\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}}$$

$$= \sqrt{\frac{3674.68}{77}}$$

$$= \sqrt{47.7231}$$

$$= 6.9082$$

$$\text{C.V} = \frac{\sigma}{\bar{X}} \times 100 \ \%$$

$$= \frac{7.14}{72} \times 100$$

$$= 9.917\%$$

Exercise 3

The data below represents the scores by 150 applicants in an achievement text for the post of Botanist in a large company:

Scores	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
Frequency	1	6	9	31	42	32	17	10	2

Estimate

(i) The mean score

(ii) The median score

(iii) The modal score

(iv) Standard deviation

(v) Semi – interquartile range

(vi) D_4

(vii) P_{26}

(viii) coefficient of variation

PROBABILITY

Probability Theory is a mathematical model of uncertainty. We shall briefly consider the following terminologies:

Experiment: This can be described as an act performed.

Trial: Is an act performed.

Outcome: Is a result realized from the trial.

Sample Space: This is the list of all the possible outcomes of an experiment. Each of the outcome in a sample space is called **sample point**.

Event: This is a subset of a sample space of an experiment.

E.g. When a coin is tossed twice, Sample Space $(S) = \{HH, HT, TH, TT\}$. Define the event A as: at least one head is observed. We have $A = \{HH, HT, TH\}$.

Axioms of Probability

Let S be a sample space, let ξ be the class of events, and let P be a real-valued function defined on ξ. Then P is called a probability function, and $P(A)$ is called the probability of the event A if the following axioms hold:

(I) For every event A, $0 \leq P(A) \leq 1$.

(II) $P(S) = 1$.

(III) If A and B are mutually exclusive events, then $P(A \cup B) = P(A) + P(B)$.

Random variables and their properties

Random variables is a function X that assigns to every element $x \in S$ one and only one real value $X(c) = x$ called the random variable. It could also be simply define as a function that assigns numerical value to each outcome defined by sample space. Consider the following table

Number	Frequency	Relative frequency
0	200	200/2000 = 1/10
1	600	600/2000 = 3/10
2	800	800/2000 = 2/5
3	240	240/2000 = 3/25
4	160	160/2000 = 2/25

which gives frequency distribution and relative frequency for all 2000 families living in a small town. Consider X to be the number of heads obtained in 3 tosses of a coin. Sample Space $(S) = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. Secondly this variable is random in the sense that the value that will occur in a given instance cannot be predicted in certainty, we can make a list of elementary outcomes as associated with X.

Numerical value of X as an event	Comparative of the event
$[X = 0]$	$[TTT]$
$[X = 1]$	$[HTT], [THT], [TTH]$
$[X = 2]$	$[HTH], [THH], [HHT]$
$[X = 3]$	$[HHH]$

Remark: The possible value of a random variable X can be determined directly from the description of the random variable without listing the sample space. However, to assign probability to this value treated as the event is sometimes helpful to refer to the sample space.

A random variable could be discrete or continuous.

Discrete: A random variable whose values are countable is called a discrete random variable. E.g. number of cars sold in a day.

Continuous: A random variable that can assume any values (one or more) contain in an interval is called a continuous random variable. E.g. time taking in complete an examination.

Cumulative distribution/ Distribution function

If X is a discrete random variable, the function given by

$F(x) = P(X \leq x) = \sum_{\substack{t \leq x \\ -\infty < x < \infty}} f(t)$ where $f(t)$ is a value of possible distribution of X at t based on the postulates of probability and some of its immediate consequences. It follows that the value $F(x)$ of the distribution function of a discrete random variable satisfies the following:

(1) $F(-\infty) = 0$. $\lim_{x \to -\infty} F(x) = 0$.

(2) $F(\infty) = 1.$ $\lim_{x\to\infty} F(x) = 1.$

Note that: $f(x) \geq 0$ and $f(x) \leq 1$ i.e. $0 \leq f(x) \leq 1$

$$\sum f(x) = 1.$$

Example: Check whether the following function given by $f(x) = \frac{x+2}{25}$ for $x = 0, \ldots, 5$ is a probability distribution function.

Solution:

$f(0) = \frac{2}{25},\ f(1) = \frac{3}{25},\ f(2) = \frac{4}{25},\ f(3) = \frac{1}{5},\ f(4) = \frac{6}{25},\ f(5) = \frac{7}{25}.$

$$\sum f(x) = \frac{2}{25} + \frac{3}{25} + \frac{4}{25} + \frac{1}{5} + \frac{6}{25} + \frac{7}{25}$$
$$= \frac{27}{25}$$
$$= 1.08$$

$\therefore f(x)$ is not a probability distribution function.

A function which values $f(x)$ define all over a set of real number is called probability density function if and only if

$P(a \leq x \leq b) = \int_a^b f(x)\, dx.$

Hence, we define the following:

A function can said to be probability density function of a continuous variable x if its value $f(x)$ satisfies the following

(1) $f(x) \geq 0 \ \forall\, x$

(2) $\int_a^b f(x)\, dx = 1;\ -\infty < x < \infty.$

Example: The probability density function (p.d.f) of the random variable x is given by $f(x) = \begin{cases} ke^{-3x} & \text{for } x > 0 \\ 0 & elsewhere \end{cases}$, find the value of k and $P(0.5 \leq x \leq 1)$.

Solution: (i) Recall for p.d.f, we must have $\int_a^b f(x)\, dx = 1$. So

$$\int_0^\infty ke^{-3x}\, dx = 1$$
$$\Rightarrow \left[-\frac{k}{3}e^{-3x}\right]_0^\infty = 1$$
$$\Rightarrow (0) - \left[-\frac{k}{3}(1)\right] = 1$$

$\Rightarrow \frac{k}{3} = 1 \quad \Rightarrow k = 3.$

(ii) $\int_{0.5}^{1} 3e^{-3x}\, dx = [-e^{-3x}]_{0.5}^{1} = 0.1733.$

PP: Each of the following tables list contain values of x and their probabilities. Determine whether or not, each table represents a valid probability distribution.

(1)

x	$P(x)$
0	.08
1	.11
2	.39

(2)

x	$P(x)$
2	.25
7	.34
4	.28
5	.13

(3)

x	$P(x)$
7	.70
8	.80
9	.20

Mathematical Expectation

The expected value of a discrete random variable having a distribution function $P(x)$ is $E(x) = \sum_x x\, P(x)$.

For a continuous random variable, $E(x) = \int_{-\infty}^{\infty} xf(x)\, dx$.

Note: If X is a discrete random variable with p.d.f $P(x)$ and $g(x)$, and if $P(x)$ is any real value function, then expectation is

$E[g(x)] = \sum_x g(x)P(x)$ (discrete r.v.)

$E[g(x)] = \int_{-\infty}^{\infty} g(x)P(x)\, dx$ (continuous r.v.)

The variance of the random variable of X is given by

$V(x) = E(X - E(x))^2 \ \ or\ E(X-\mu)^2$.

And the standard deviation is given by

$\sigma(x) = \sqrt{E(X-\mu)^2}$.

For any random variable X, we have

(i) $E(ax + b) = a\, E(x) + b$

(ii) $V(ax\ + b) = a^2\, V(x)$.

Proof:

(i) $E(ax+b) = \sum_x (ax+b)P(x)$

$= \sum[(ax)P(x) + bP(x)]$

$= \sum ax\, P(x) + \sum b\, P(x)$

$= a\sum x\, P(x) + b\sum P(x)$

$= aE(x) + b \quad (\text{since } \sum P(x) = 1)$ ■

(ii) $V(ax+b) = E[(ax+b) - E(ax+b)]^2$

$= E[ax + b - (aE(x) + b)]^2$

$= E[ax - aE(x)]^2$

$= a^2\, E[x - E(x)]^2$

$= a^2\, V(x)$ ■

Theorem: If X is a random variable with mean μ, then $V(x) = E(X^2) - \mu^2$.

Proof:

$V(x) = E[X-\mu]^2$

$= E[X^2 - 2X\mu + \mu^2]$

$= E(X^2) - E(2X\mu) + E(\mu^2)$

$= E(X^2) - 2\mu E(X) + \mu^2$

$= E(X^2) - 2\mu^2 + \mu^2$

$= E(X^2) - \mu^2$ ■

Example: Use the above result; compute the variance of X as given in the table below:

X	0	1	2
$P(x)$	0.1	0.5	0.4

Solution: $E(x) = \sum x\, P(x) = 0 + 0.5 + 0.8 = 1.3.$

$V(x) = E(X^2) - [E(X)]^2$

$$E(X^2) = \sum x^2 P(x) = 0 + 0.5 + 1.6 = 2.1$$
$$\therefore V(x) = 2.1 - (1.3)^2 = 0.41$$

Exercise 4

(1) Suppose X has probability density function given as $F(x) = \begin{cases} 3x^2, & 0 \le x \le 1 \\ 0, & elsewhere \end{cases}$.

Find the mean and variance of x.

(2) A lot of 12 television sets chosen at random are defective, if 3 of the sets are chosen at random for shipment in hotel, how many defective set can they expect?

(3) Certain coded measured of the pitch diameter of threads of a fitting have the probability density $f(x) = \begin{cases} \frac{4}{\pi(1+x^2)} & \text{for } 0 < x < 1 \\ 0, & otherwise \end{cases}$.

Find $E(x)$ and $V(x)$.

(4) If X has the p.d.f $f(x) = \begin{cases} 3e^{-3x} & for \quad x > 0 \\ 0, & elsewhere \end{cases}$. Find the expected value and the variance of x.

BERNOULLI DISTRIBUTION

A random variable x has a Bernoulli distribution if and only if its probability distribution is given as $f(x,p) = p^x(1-p)^{n-x}$; $x = 0,1$. In this context, p may be probability of passing or failing an examination.

$f(0,p) = 1 - p$ {a coin $n = 1$}

$f(1,p) = p$.

BINOMIAL DISTRIBUTION

An experiment consisting of n repeated trials such that

(1) the trials are independent and identical

(2) each trial result in only one or two possible outcomes

(3) the probability of success p remains constant

(4) the random variable of interest is the total number of success.

The binomial distribution is one of the widely used in statistics and it used to find the probability that an outcome would occur x times in n performances of an experiment. For example, consider a random variable of flipping a coin 10 times. When a coin is toss, the probability of getting head is p and that of tail is $1 - p = q$.

A random variable x has a binomial distribution and it is referred to as binomial random variable if and only if its probability is given by

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \ , \ x = 0,1, \dots, n$$

Where $n =$ Total number of trials

$p =$ Probability of success

$1 - p =$ Probability of failure

$n - x =$ Number of failure in n – trials

To find the probability of x success in n trials, the only values we need are that of n and p.

Properties of Binomial Distribution

Mean $\mu = np$

Variance $\sigma^2 = npq \ ; \quad q = 1 - p$

Standard deviation $\sigma = \sqrt{npq}$

Moment coefficient of skewness $\alpha_3 = \frac{q-p}{\sqrt{npq}}$

Moment coefficient of kurtosis $\alpha_4 = 3 + \frac{1-6pq}{\sqrt{npq}}$.

Note: $X \sim B(n.p)$ reads as X follows binomial distribution

Example: Observation over a long period of time has shown that a particular sales man can make a sale on a single contact with the probability of 20%. Suppose the same person contact four prospects,

(a) What is the probability that exactly 2 prospects purchase the product?

(b) What is the probability that at least 2 prospects purchase the product?

(c) What is the probability that all the prospects purchase the product?

(d) What is the expected value of the prospects that would purchase the product?

Solution: Let X denote the number of prospect: $x = 0,1,2,3,4$.

Let p denote the probability of (success) purchase $= 0.2$

Hence, $X \sim B(4,0.2)$.

$$f(x) = \binom{n}{x} p^x q^{n-x}$$

$$= 0, elsewhere$$

$$f(x) = \binom{4}{x} (0.2)^x (0.8)^{4-x}$$

(a) $f(x = 2) = \binom{4}{2}(0.2)^2(0.8)^{4-2} = 0.1536$

(b) $f(x \geq 2) = f(2) + f(3) + f(4)$ (OR $1 - f(x \leq 1) = 1 - \{f(0) + f(1)\}$)

$$= 0.1536 + \left[\binom{4}{3}(0.2)^3(0.8)^{4-3}\right] + \left[\binom{4}{4}(0.2)^4(0.8)^{4-4}\right]$$

$$= 0.1536 + 0.0256 + 0.0016$$

$$= 0.1808$$

(c) $f(x = 4) = \binom{4}{4}(0.2)^4(0.8)^{4-4}$

$$= 0.0016$$

(d) Expected value $= np = 4 \times 0.2 = 0.8$

POISSON DISTRIBUTION

When the size of the sample (n) is very large and the probability of obtaining success in any one trial very small, then Poisson distribution is adopted.

Given an interval of real numbers, assumed counts of occur at random throughout interval, if the interval can be partition into sub interval of small enough length such that

(1) The probability of more than one count sub interval is 0

(2) The probability of one count in a sub interval is the same for all sub intervals and proportional to the length of the sub interval

(3) The count in each of the sub interval is independent of all other sub intervals.

A random experiment of this type is called a Poisson Process. If the mean number of count in an interval is $\lambda > 0$, the random variable x that equals the number of count in an interval has a Poisson distribution with parameter λ and the probability density function is given by

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}; \quad x = 0,1,\dots,n.$$

For a Poisson variate, the mean and the variance is given by $E(x) = \lambda, \quad V(x) = \lambda.$

Example: Flaws occur at random along the length of a thick, suppose that a number of flaws follows a Poisson distribution with a mean flaw of 2.3 per mm. determine probability of exactly 2 flaws in one mm of wire.

Solution: $f(x = 2) = \frac{e^{-2.3}(2.3)^2}{2!} = 0.265.$

The Poisson distribution has an approximation to binomial; when n is large and p is as close to 0 as possible, then the Poisson distribution has a history which approximate of that of the binomial.

For the binomial distribution: $f(x;n,p) = \binom{n}{x}p^{x}(1-p)^{n-x}$, $E(x) = np$

Approximately by Poisson distribution: $f(x;\lambda) = \frac{e^{-np}\,(np)^{x}}{x!}$, $x = 0,1,\dots,n$

Hence we can apply the Poisson distribution to the binomial when $n \geq 30$ and $np < 5$.

Example: If the 3% of the electric doors manufactured by a company are defective. Find the probability that in the sample of 120 doors, at most 3 doors are defective.

(a) Use binomial to solve the problem

(b) Use Poisson distribution and compare your results.

Solution: Let p denote probability of electric doors defective $= 0.03$

Let $q = 1 - p = 0.97$ denote probability of electric non defective.

Let n denote total number of sample of electric doors $= 120$.

Let x denote number of doors being consider.

(a) By binomial distribution:

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x}$$

$$f(x \leq 3) = \sum_{x=0}^{3} f(x) = f(0) + f(1) + f(2) + f(3)$$

$$f(0) = \binom{120}{0} (0.03)^0 (0.97)^{120} = 0.026$$

$$f(1) = \binom{120}{1} (0.03)^1 (0.97)^{119} = 0.095$$

$$f(2) = \binom{120}{2} (0.03)^2 (0.97)^{118} = 0.177$$

$$f(3) = \binom{120}{3} (0.03)^3 (0.97)^{117} = 0.215$$

$$f(x \leq 3) = 0.026 + 0.095 + 0.177 + 0.215 = 0.513.$$

(b) By Poisson distribution:

$$f(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \quad , \quad \lambda = np = 120 \times 0.03 = 3.6$$

$$f(0) = \frac{e^{-3.6} (3.6)^0}{0!} = 0.027$$

$$f(1) = \frac{e^{-3.6} (3.6)^1}{1!} = 0.098$$

$$f(2) = \frac{e^{-3.6} (3.6)^2}{2!} = 0.177$$

$$f(3) = \frac{e^{-3.6} (3.6)^3}{3!} = 0.213$$

$$f(x \leq 3) = f(0) + f(1) + f(2) + f(3)$$

$$= 0.027 + 0.098 + 0.177 + 0.213$$

$= 0.515$

Hence, Poisson distribution result is very close to the binomial distribution result showing that it can be use to approximate binomial distribution in this problem.

GEOMETRIC DISTRIBUTION

Consider a random experiment in which all the conditions of a binomial distribution hold. However, instead of fixed number of trials, trials are conducted until first success occurs. Hence by definition, in a series of independent binomial trials with constant probability p of success, let the random variable x denotes number of trials until first success. Then x is said to have a geometric distribution with parameter p and given by

$$f(x) = p(1-p)^{x-1}\ ;\ \ x = 1,2, \ldots$$

For Geometric distribution, the mean and variance are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2}.$$

Examples:

(1) If the probability that a wave contain a large particles of contamination is 0.01, it assumes that the wave are independent, what is the probability that exactly 125 waves need to be analyzed before a large particle is detected?

Solution: Let X denotes the number of samples analyzed until a large particle is detected. Then X is a geometric random variable with $p = 0.01$. Hence, the required probability is $f(x = 125) = (0.01)(0.99)^{124} = 0.0029$.

(2) Each sample of n has 10% of chance of containing a particular rare molecule. Assume samples are independent with regard to the present of rare molecule. Find the probability that in the next 18 samples, (a) exactly 2 containing rare molecule. (b) at least 4 sample.

Solution: Left as exercise

HYPERGEOMETRIC DISTRIBUTION

Suppose we have a relatively small quantity consisting of N items of which $k(= NP)$ are defective. If two items are samples sequentially then the outcome for the second draw is very much influenced by what happened on the first drawn provided that the first item drawn remain in the quantity. We need to obtain a formula similar to that of binomial distribution, which applies to sample without replacement.

A random variable x is said to have a hypergeometric distribution if and only if its probability density function is given by

$$f(x; n, N, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \quad , x \leq n, \ x = 0,1, \dots, n, \ \ n - x \leq N - k.$$

$$= 0 \ \text{otherwise}$$

Where $N =$ total number of sample

$n =$ number of chosen without replacement of items from N elements.

$k =$ when consider a set of N objects which k are looked upon as success.

$N - K =$ as failure.

The mean and variance for hypergeometric distribution are given as:

$$E(x) = \frac{nk}{N} \ ; \qquad V(x) = \frac{nk\,(N-k)(N-n)}{N^2(N-1)}$$

Examples:

(1) The random sample of 3 oranges is taking from a basket containing 12 oranges, if 4 of the oranges in the basket are bad, what is the probability of getting (a) no bad oranges from the sample (b) more than 2 are bad from the sample.

Solution: $p = \frac{4}{12} = \frac{1}{3}, \ N = 12, \ n = 3, \ k = Np = 4$

$$f(x) = \frac{\binom{4}{x}\binom{8}{3-x}}{\binom{12}{3}}, \ x = 0,1,2,3$$

$$= 0 \ \text{otherwise}$$

(a) $f(0) = \frac{\binom{4}{0}\binom{8}{3}}{\binom{12}{3}} = 0.25$ (b) $f(x > 2) = f(3) = \frac{\binom{4}{3}\binom{8}{0}}{\binom{12}{3}} = 0.018.$

(2) A batch of parts contain 100 parts from a local supplier of tubing and 200 parts from a supplier of tubing in the next state, if 4 parts are selected at random without replacement, what is the probability that they are all from the local supplier.

Solution: Let X equals the number of parts in the sample from the local supplier, then X has a hypergeometric distribution and the required probability is $f(x = 4)$ consequently

$$h(4; 4,300,100) = \frac{\binom{100}{4}\binom{200}{0}}{\binom{300}{4}} = 0.0119.$$

NEGATIVE BINOMIAL DISTRIBUTION

A generalization of the geometric distribution in which the random variables is a number of Bernoulli trials required to obtain r success results in negative binomial distribution.

We may be interested in the probability that 8th child exposed to measles is the 3rd to contact it. If the kth success is to occur on the xth trial, there must be $k - 1$ successes on the first x trial and the probability for this is given by

$$b^*(k-1, x-1, p) = \binom{x-1}{k-1}p^{k-1}(1-p)^{x-k}.$$

The probability of success on the kth trial is θ and the probability that the kth success occurs on the xth trial is:

$$\theta \cdot b(k-1, x-1, p) = \binom{x-1}{k-1}p^{k}(1-p)^{x-k}$$

Hence, we say that a random variable x has a negative binomial distribution if and only if its probability density function is given by

$$b(x;, k, p) = \binom{n-1}{k-1}p^{k}(1-p)^{x-k} \text{ for } x = k, k+1, k+2, \ldots$$

The mean and variance is given by $E(x) = \frac{k}{p}$; $V(x) = \frac{k-kp}{p^2}$

Exercise 5

(1) The probability that an experiment will succeed is 0.6 if the experiment is repeated to 5 successive outcomes have occurred, what is the mean and variance of number of repetition required?

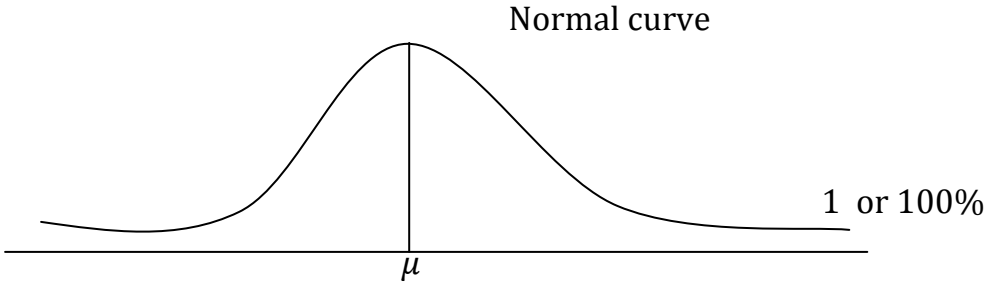
(2) A high performance aircraft contains 3 identical computer, only one is used to operates in aircraft, the other two are spares that can be activated incase the primary system fails. During one hour of operations, the probability of failure in the primary computer or any activated spare system is 0.0005. Assume that each hour represent an identical trial.

(a) What is the expected value to failure of all the 3 computers?

(b) What is the probability that all the 3 computers fail within a 5 hour flight?

NORMAL DISTRIBUTION

The normal distribution is the most important and the most widely used among all continuous distribution in the statistics. It is considered as the corner stone of statistics theory. The graph of a Normal distribution is a bell – shaped curved that extends indefinitely in both direction.



Features (Properties) of Normal Curve

1. The curve is symmetrical about the vertical axis through the mean μ.
2. The mode is the highest point on the horizontal axis where the curve is maximum and occurs where $x = \mu$.
3. The normal curve approaches the horizontal axis asymptotically.
4. The total area under the curve is one (1) or 100%.

5. About 68% of all the possible x- values (observations) lie between $\mu - \sigma$ and $\mu + \sigma$, or the area under the curve between $\mu - \sigma$ and $\mu + \sigma$ is 68% of the total area.
6. About 95% of the observations lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
7. 99.7% (almost all) of the observations lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Note: The last three of the above properties are arrived at through advanced mathematical treatment.

It is clear from these properties that a knowledge of the population means and standard deviation gives a complete picture of the distribution of all the values.

Notation: Instead of saying that the values of a variable x are normally distributed with mean μ and standard deviation σ, we simply say that x has an $N(\mu, \sigma^2)$ or x is $N(\mu, \sigma^2)$ or $x \sim N(\mu, \sigma^2)$.

A random variable x is said to have a normal distribution if its probability density function is given by

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

Where σ and μ are the parameters of the distribution. Note that since $n(x; \mu, \sigma)$ is a p.d.f, it established the fact that the area is 1. In order word, cumulative distribution function is given by

$$F(x) = P(x \leq x) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

The Standard Normal Curve

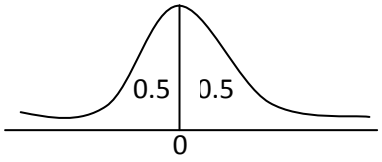
The standard normal distribution is a special case. If $X \sim N(0,1)$, then x is called the standard normal variable with p.d.f $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$.

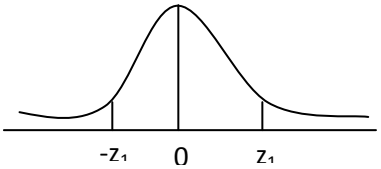
The table usually used to determine the probability that a random variable x drawn from a normal population with no mean and standard deviation 1 is the standard normal distribution table (or z- scores table).

The following point has keen noted in the used of the table:

$P(z \leq 0) = 0.5$

$P(z \leq -z) = 1 - P(z \leq z) = P(z \geq z)$





Case I

(i) Area $A = P(z \leq -z_1)$

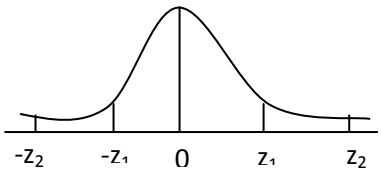
$= P(z > z_1)$

$= 1 - P(z < z_1)$

(ii) $P(z < z_1) = P(z > -z_1)$

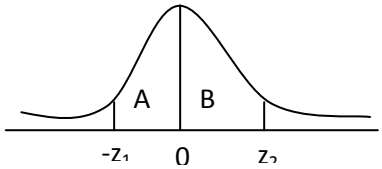
$P(0 < z < z_1) = P(-z < z < 0)$

Case II



Given that Area A = Area B

$P(-z_2 < z < -z_1) = P(z_1 < z < z_2)$



Area A ≠ Area B

Here $P(-z_1 < z < z_2) = P(z < z_2) - P(z > -z_1)$

$= P(z < z_2) - [1 - P(z < z_1)]$

$= P(z < z_2) + P(z < z_1) - 1$

Example: Find the probability that a random variable having the standard distribution will take a value

(a) less than 1.72 (b) less than –0.88

(c) between 1.19 and 2.12 (d) between –0.36 and 1.21

Solution:

(a) $P(z < 1.72) = 0.9573$ (using the table)

(b) $P(z < -0.88) = 1 - P(z < 0.88)$

$= 1 - 0.8106$

$= 0.1894$

(c) $P(1.19 < z < 2.12) = P(z < 2.12) - P(z > 1.19)$

$= P(z < 2.12) - [1 - P(z < 1.19)]$

$= 0.9830 - 1 + 0.8830$

$= 0.866$

(d) $P(-0.36 < z < 1.21) = P(z < 1.21) - P(z > -0.36)$

$= P(z < 1.21) - [1 - P(z < 0.36)]$

$= P(z < 1.21) + P(z < 0.36) - 1$

$= 0.8869 + 0.6406 - 1$

$= 0.5275$

STANDARDIZED NORMAL VARIABLE

In real world application, the given continuous random variable may have a normal distribution in value of a mean and standard deviation different from 0 and 1. To overcome this difficulty, we obtain a new variable denoted by z and this is given by

$$z = \frac{x-\mu}{\sigma} \quad (\text{for } \mu \neq 0 \text{ and } \sigma \neq 1)$$

Example:

(1) Suppose the current measurement in a strip of wire assumed to be normally distributed with a mean of 10mA and a variance of 4mA. What is the probability that the current is greater than 13?

Solution: Let X denote the current in milliamp, the required probability is $P(X > 13)$. Let $z = \frac{x-\mu}{\sigma} = \frac{13-10}{2} = 1.5$

$$\therefore P(X > 13) = P(z > 1.5)$$
$$= 1 - P(z < 1.5)$$
$$= 1 - 0.9332$$
$$= 0.0668mA$$

(2) If $x \sim N(55,16)$, calculate $P(51 < x < 59)$

Solution: $P(51 < x < 59) = P\left(\frac{x_1-\mu}{\sigma} < x < \frac{x_2-\mu}{\sigma}\right)$
$$= P\left(\frac{51-55}{4} < z < \frac{59-55}{4}\right)$$
$$= P(-1 < z < 1)$$
$$= P(z < 1) + P(z < 1) + 1$$
$$= 0.8413 + 0.8413 - 1$$
$$= 0.6826$$

The above problem means that 68% of the x −values are within 1 standard deviation of the mean 0.

(3) It is known that the marks in a University direct entry examination are normally distributed with mean 70 and standard deviation 8. Given that your score is 66, what percentage of all the candidates will be expected to score more than you?

Solution: We have $x \sim N(70,64)$ and the required proportion of x −values that are above 66 is $P(x > 66)$.

$$P(x > 66) = 1 - P(x \leq 66)$$
$$= 1 - P\left(z \leq \frac{66-70}{8}\right)$$
$$= 1 - P(z \leq -0.5)$$
$$= 1 - 0.3085$$
$$= 0.6915$$

$\therefore$ 69% of all the candidates will be expected to do better than you.

PP: It is known from the previous examination results that the marks of candidates have a normal distribution with mean 55 and standard deviation 10. If the pass mark in a new examination is set at 45, what percentages of the candidates will be expected to fail?

Normal Approximation to Binomial distribution

Recall that the binomial distribution is applied to a discrete random variable. As n- trial increases, the uses of binomial formula becomes tedious and when this happen, the normal distribution can be use to approximate the binomial probability.

The probability we obtained by using the normal approximation to the binomial is an approximate to the exact and the condition under which we can use normal approximation for a binomial distribution are as follows:

(1) $n \geq 30$ (2) $np > 5$ (3) $n(1-p) > 5$

And the tools is $z = \frac{X-np}{\sqrt{npq}}$.

Example: In a digital communication channel assumed that the number of bits received in error can be model by binomial variable, assumed also that the probability that in bit received in error is 1×10^{-5} if 16 million bits are transmitted. What is the probability that more than 150 errors occurs?

Solution: Let X denotes the number of errors.

$$P(X > 150) = 1 - P(X \leq 150)$$
$$= 1 - \sum_{x=0}^{150} \binom{16000000}{x} (10^{-5})^x (1-10^{-5})^{n-x}$$

Approximately:

$$P(X > 150) = P\left(z > \frac{X - 160}{\sqrt{160(10^{-5})}}\right)$$
$$= P(z > -0.79)$$
$$= 1 - P(z \leq 0.79)$$
$$= 1 - 0.7852$$
$$= 0.2148$$

Exercise 6

(1) Two fair dies are toss 600 times. Let X denotes number of times the total of 7 occurs. Find the probability that X lies between 80 and 110.

(2) A manufacturer of machine parts claims that at most 10% of each parts are defective. A parameter needs 120 of such parts and to be sure of getting many goods ones, he places an order for 140 parts. If manufacturer claims is valid, what is the probability that the purchaser would receive at least 120 good parts?

Problem with simple linear equation

If X is normally distributed with mean$= 2$ and variance $= 4$. Find the value of λ such that the probability that $X > \lambda = 0.10$. Hint $X \sim N(2,4)$.

Solution: $P\left(\frac{X-\mu}{\sigma} > \frac{\lambda-\mu}{\sigma}\right) = 0.10$

$\Rightarrow P\left(z > \frac{\lambda-2}{2}\right) = 0.10$

$\Rightarrow 1 - P\left(z \leq \frac{\lambda-2}{2}\right) = 0.10$

$\Rightarrow P\left(z \leq \frac{\lambda-2}{2}\right) = 0.90$

i.e. $\varphi\left(\frac{\lambda-2}{2}\right) = 0.90$

i.e. $\frac{\lambda-2}{2} = \varphi^{-1}(0.90) = 1.285$

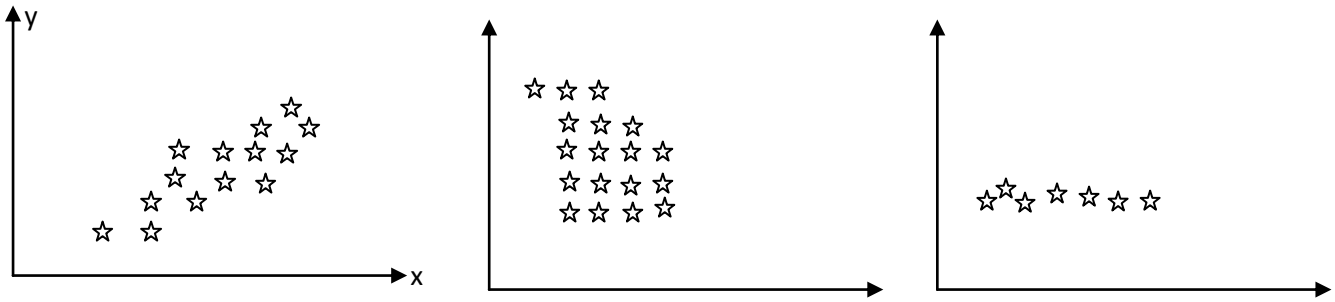
$\Rightarrow \hat{\lambda} = 2(1.285) + 2 = 4.570.$

REGERESSION AND CORRELATION

In the physical sciences and engineering, even social sciences, it is possible to formulate models connecting several quantities such as temperature and pressure of a gas, the size and height of a flowering plant, the yield of a crop and weather conditions. It may also be necessary to examine what appears to be the case of variations in one variable in relation to the other. One technique for examining such relationship is Regression.

Regression analysis is the study of the nature and extent of association between two or more variables on the basis of the assumed relationship between them with a view to predict the value of one variable from the other.

Scatter diagram: The first step in studying the relationship between two variables is to draw a scatter diagram. This is a graph that shows visually the relationship between two variables in which each point corresponds to pair of observations, one variable being plotted against the other. The way in which the dots lie on the scatter diagrams shows the type of relationship that exists.



Regression Models

In order to predict one variable from the other, it is necessary to construct a line or curve that passes through the middle of the points, such that the sum of the distance between each point and the line is equal to 0. Such line is called the line of best fit.

The simple regression equation of Y on X is defined as

$$Y = a + bX + e$$

while the multiple regression equation of Y on $X_1, X_2, \ldots, X_k$ is

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e.$$

where: Y is the observed dependent variable

X is the observed independent (explanatory) variable

a is the intercept (the point at which the reg. line cuts the Y axis)

b is the slope (regression coefficient). It gives the rate of change in Y per unit change in X.

e is the error term.

Ordinary Least Square Method (OLS)

Although there are other techniques for obtaining these parameters (a, and $b's$) such as the likelihood method, but we shall use the method of least squares to estimate the parameters of a simple regression equation. This method involves finding the values of the regression coefficients a and b, that minimizes the sum of squares of the residuals (error).

Assumptions of OLS

i. The relationship between X and Y is assumed to be linear
ii. The X values are fixed
iii. There is no relationship between X and the error term i.e. $E(X_e) = 0$
iv. The error is assumed to be normally distributed with mean zero and variance 1.

Derivations

Suppose there are n parts of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, we assume a linear relationship $Y_i = a + bX_i + e_i$ from which $e_i = Y_i - a - bX_i$.

Let $Q = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n}(Y_i - a - bX_i)^2$. Differentiating partially with respect to a and b and equates to zero (since no turning point for straight line), we have

$$\frac{\partial Q}{\partial a} = (-2)\sum_{i=1}^{n}(Y_i - a - bX_i) = 0$$

$$\therefore \sum Y_i = na + b\sum X_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\frac{\partial Q}{\partial b} = (-2)\sum_{i=1}^{n}(Y_i - a - bX_i)X_i = 0$$

$$\Rightarrow \sum Y_iX_i = a\sum X_i + b\sum {X_i}^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Equation (1) and (2) are called the normal equations. Form equation (1), we have

$$a = \frac{\sum Y_i - b\sum X_i}{n} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Substituting equation (3) in the equation (2), we have

$$\sum Y_i X_i = \left(\frac{\sum Y_i - b\sum X_i}{n}\right)\sum X_i + b\sum {X_i}^2$$

$$\Rightarrow n\sum Y_i X_i = \sum Y_i \sum X_i - b(\sum X_i)^2 + bn\sum {X_i}^2$$

$$n\sum Y_i X_i = \sum Y_i \sum X_i + b\left[n\sum {X_i}^2 - (\sum X_i)^2\right]$$

$$\Rightarrow b = \frac{[n\sum Y_i X_i - \sum Y_i \sum X_i]}{n\sum {X_i}^2 - (\sum X_i)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Also, dividing through by n we have

$$b = \frac{\sum Y_i X_i - {(\sum Y_i \sum X_i)}/{n}}{\sum {X_i}^2 - {(\sum X_i)^2}/{n}}$$

Writing $\frac{\sum Y_i}{n} = \bar{Y}$ and $\frac{\sum X_i}{n} = \bar{X}$, we have

$$b = \frac{\sum Y_i X_i - n\bar{Y}\bar{X}}{\sum {X_i}^2 - n\bar{X}^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

$$a = \bar{Y} - b\bar{X} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6) \qquad \blacksquare$$

$\therefore \hat{Y} = a + bX$ is the regression equation.

The standard error of estimates

If all the points on the scatter diagram fall on the regression line, it means there is no error in the estimate of Y, thus the variation in Y is fully explained by X. But if all the points scatter round the line of best fit, a measure of the standard error is given by $S_y = \sqrt{\frac{{e_i}^2}{n-2}}$ where $e_i = \sum_{i=1}^{n}\left(Y - \hat{Y}\right)$ and S_y is the standard error of estimate.

The Coefficient of Determination

This is the percentage of variation in Y that is explained by X. Since total variation is ${S_T}^2 = \sum_{i=1}^{n}(Y - \bar{Y})^2$, the unexplained variation is ${S_y}^2 = \frac{\sum {e_i}^2}{n-2}$. Therefore, the coefficient of determination is

$$r^2 = 1 - \frac{Unexplained \quad variatio\ n}{Total \quad variation} = 1 - \frac{{S_y}^2}{{S_T}^2}.$$

Examples:

(1) The following are measurement height and ages months of maize plant in a plantation farm.

Age month	1	2	3	4	5	6	7
Height (cm)	5	13	16	23	33	38	40

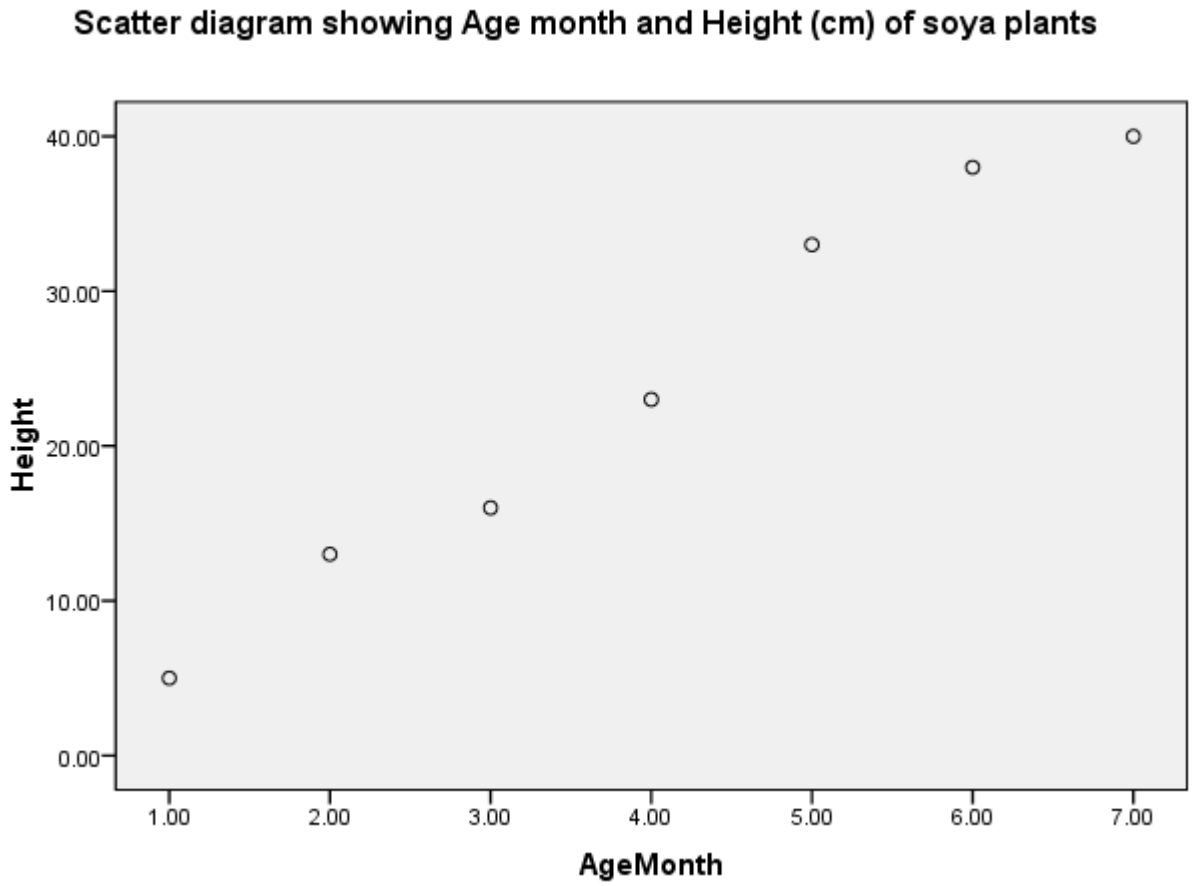
(a) Draw the scatter diagram

(b) Obtained the regression equation of the age month on the height of the plant.

(c) Estimate the height of the maize plant aged 10 months.

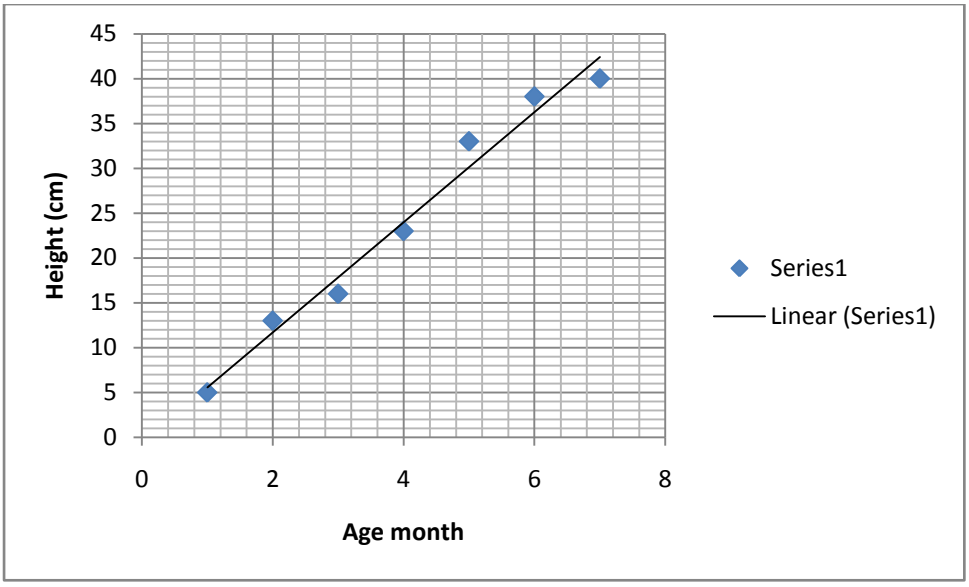
Solution:

(a) (Using SPSS)



(Using Excel)

Scatter diagram showing both Age month and Height (cm) of soya plants



(b)

X	Y	X^2	XY
1	5	1	5
2	13	4	26
3	16	9	48
4	23	16	92
5	33	25	165
6	38	36	228
7	40	49	280
28	168	140	844

Using equation (4), we have

$$b = \frac{n\sum Y_i X_i - \sum Y_i \sum X_i}{n\sum {X_i}^2 - (\sum X_i)^2} = \frac{7(844) - (168)(28)}{7(140) - (28)^2}$$

$$= \frac{1204}{196}$$

$$= 6.143$$

$$a = \frac{\sum Y_i - b\sum X_i}{n} = \frac{168 - 6.143(28)}{7}$$

$$= -0.572$$

The regression equation is $\hat{Y} = -0.572 + 6.143X_i$

(c) When $X = 10$, $Y = -0.572 + 6.143(10) = 60.858$. That is, the height of the

maize plant would be 60.858 when the plant aged 10 months.

(2) A study was made on the effect of a certain brand of fertilizer (X) on cassava yield (Y) per plot of farm area resulting in the following data:

Fertilizer	0	1	2	3	4	5	6	7	8	9	10
Cassava yield	5	9	8	11	14	12	13	15	18	17	22

(a) Plot the scatter diagram and draw the line of best fit.

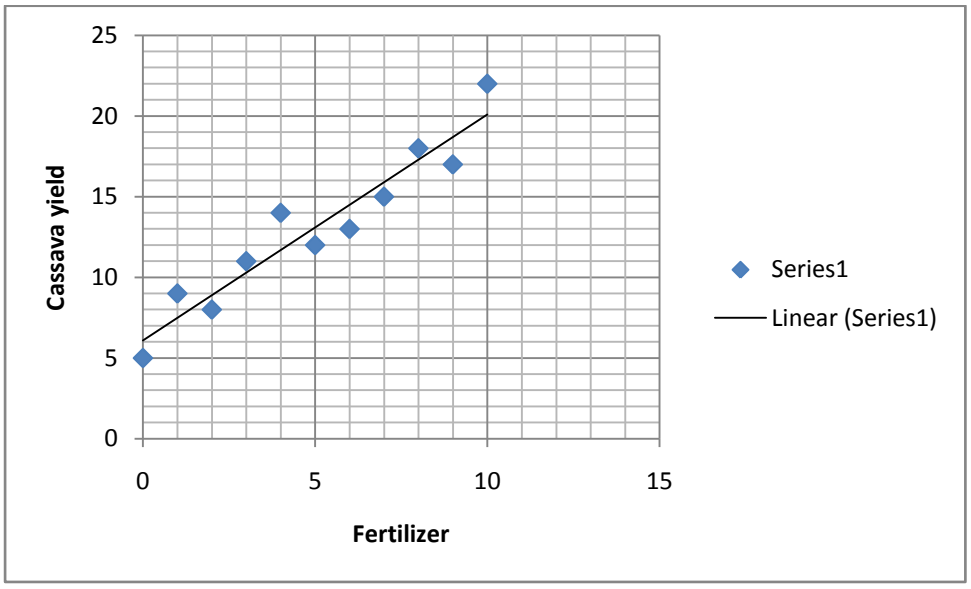
(b) Resolve this data to a simple regression equation

(c) What is the value of maize yield when the fertilizer is 12?

(d) Obtain the standard error of the regression

Solution:

(a)



(b)

X	Y	X^2	XY	$\hat{Y} = a + bX$	$e_i = Y_i - \hat{Y}_i$	${e_i}^2$
0	5	0	0	6.091	-1.091	1.190
1	9	1	9	7.491	1.509	2.277
2	8	4	16	8.891	-0.891	0.794
3	11	9	33	10.291	0.709	0.503
4	14	16	56	11.691	2.309	5.331
5	12	25	60	13.091	-1.091	1.190
6	13	36	78	14.491	-1.491	2.223
7	15	49	105	15.811	-0.891	0.794
8	18	64	144	17.291	-0.709	0.503
9	17	81	153	18.691	1.691	2.859
10	22	100	220	20.091	1.909	3.644
55	144	385	874	143.921		21.308

$$b = \frac{n\sum Y_i X_i - \sum Y_i \sum X_i}{n\sum {X_i}^2 - (\sum X_i)^2} = \frac{11(874) - (144)(55)}{11(385) - (55)^2}$$

$$= \frac{1694}{1210}$$

$$= 1.4$$

$$a = \frac{\sum Y_i - b\sum X_i}{n} = \frac{144 - 1.4(55)}{11}$$

$$= 6.091$$

The simple regression equation is $\hat{Y} = 6.091 + 1.4X_i$

(c) When $X = 12,\ Y = 6.091 + 1.4(12) = 22.891.$

(d) Standard error $= \sqrt{\frac{\sum {e_i}^2}{n-2}} = \sqrt{\frac{21.208}{9}} = 1.539.$

CORRELATION

Correlation measures the degree of linear association between two or more variables when a movement in one variable is associated with the movement in the other variable either in the same direction or the other direction. Correlation coefficient is a magnitude, which indicates the degree of linear association between two variables. It is given by

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$
$$= \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{[\sum X^2 - n\bar{X}^2][\sum Y^2 - n\bar{Y}^2]}}$$

We note that if the regression equation of Y on X is $Y = \beta_0 + \beta_1 X + e$ and if the regression equation of X on Y is given by $X = \alpha_0 + \alpha_1 X + e$, then it can be shown that $\alpha_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2}$, $\beta_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$. Given α_1 and β_1, the product moment correlation (Pearson's) coefficient is given as

$$r = \sqrt{\alpha_1 \beta_1} = \sqrt{\frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2} \cdot \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}}$$
$$= \sqrt{\frac{(\sum XY - n\bar{X}\bar{Y})^2}{(\sum Y^2 - n\bar{Y}^2)(\sum X^2 - n\bar{X}^2)}}$$
$$= \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{(\sum Y^2 - n\bar{Y}^2)(\sum X^2 - n\bar{X}^2)}} \text{ Or } = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Interpretation of r:

$r = +1$, Implies that there is a perfect positive linear (direct) relationship

$r = -1$, Implies a perfect negative (indirect) linear relationship

$-1 < r < -0.5$, Implies there is a strong negative linear relationship

$-0.5 < r < 0$, Implies there is a weak negative linear relationship

$0 < r < +0.5$, Implies there is a weak positive linear relationship

$+0.5 < r < +1$, Implies there is a strong positive linear relationship

$r = 0$, Implies there is no linear relationship between the two variables.

Example 1: Using the data in example 1 on regression above, measure the degree of association between X and Y. Also comment on your result.

Solution:

X	Y	X^2	XY	Y^2
1	5	1	5	25
2	13	4	26	169
3	16	9	48	256
4	23	16	92	529
5	33	25	165	1089
6	38	36	228	1444
7	40	49	280	1600
28	168	140	844	5112

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} = \frac{(7)(844) - (28)(168)}{\sqrt{[7(140) - (28)^2][7(5112) - (168)^2]}}$$

$$= \frac{1204}{\sqrt{(196)(7560)}}$$

$$= 0.9891$$

Comment: There is a strong positive linear relationship between X and Y.

Spearman's Rank Correlation Coefficient

This is a magnitude or a quantity that measures the degree of association between two variables on the basis of their ranks rather than their actual values. Since qualitative variables such as efficacy, intelligence, beauty, religion etc cannot be measured quantitatively, to circumvent this problem. The Spearman's rank correlation coefficient is given by

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

Where D is the difference in ranks and n is the number of observation, $-1 \leq r_s \leq +1$.

Example 2: In a test carried out to measure the efficiency of fifteen Computer Science Students in developing software for the computation of the institution students' results; two judges were asked to score the candidates. Their scores are as follows, is there any agreement in their assessment of the candidates?

Judge X	72	45	48	85	59	94	68	92	29	50	73	90	48	56	75
Judge Y	68	56	60	76	45	46	57	85	32	40	85	88	57	68	64

Solution:

Judge X	72	45	48	85	59	94	68	92	29	50	73	90	48	56	75
Judge Y	68	56	60	76	45	46	57	85	32	40	85	88	57	68	64
R_I	7	14	12.5	4	9	1	8	2	15	11	6	3	12.5	10	5
R_{II}	5.5	11	8	4	13	12	9.5	2.5	15	14	2.5	1	9.5	5.5	7
D	1.5	3	4.5	0	-4	-11	-1.5	-0.5	0	-3	3.5	2	3	4.5	-2
D^2	2.25	9	20.25	0	16	121	2.25	0.25	0	9	12.25	4	9	20.25	4

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6(229.5)}{15(15^2-1)} = 1 - 0.4098 = 0.5902$$

Comment: There is a fairly strong agreement between the two judges in their assessment.

Exercise 7

In a certain company, drums of mentholated spirit are kept in storage for sometime before being bottled. During storage, the evaporation of part of water content of the spirit takes place and an examination of such drums give the following results.

Storage time (weeks)	2	5	7	9	12	13
Evaporation loss (a.m)	38	57	65	73	84	91

(i) Find the regression equation of evaporation loss on the storage time for the mentholated spirit.

(ii) Find the regression equation of storage time on the evaporation loss.

(iii) Find the product moment correlation coefficient.

(iv) Determine the Spearman's rank correlation.

ESTIMATION

When we assign value to a population parameter based on sample information is called Estimation. An estimate is a value assign to population parameter based on the value of the statistics. A sample statistics used to estimate a population parameter is called an estimator. In other words, the function or rule that is used to guess the value of a parameter is called an Estimator, and estimate is a particular value calculated from a particular sample of an observation. But estimator like any statistic is a random variable. Parameter is represented by Greek letter while statistic represented by roman numbers:

Parameter (population characteristics)	Statistic (sample characteristics)
μ	$\bar{X}$
ρ	$\bar{p}$
σ^2	S^2

An estimator is divided into two, namely (i) Point estimator (ii) Interval estimator. A point estimator is a single value given to the population parameter based on the value of the sample statistics; while an interval estimator consists of two numerical values within which we believe with some degree of confidence include the value of the parameter being estimated. In many situations, a point estimate does not supply the complete information to a researcher; hence, an approach is used called the confidence interval.

The subject of estimator is concerned with the methods by which population characteristics are estimated from sample information. The objectives are:

(i) To present properties for judging how well a given sample statistics estimates the parent population parameter.

(ii) To present several methods for estimating these parameters.

Properties of a Good Estimator

1. Unbiasedness

   An estimator should be unbiased. An estimator, $\hat{\theta}$ is said to be unbiased if the expected value of $\hat{\theta}$ is equal to the population parameter θ. That is, $E(\hat{\theta}) = \theta$.

   Example:

   (a) Show that $\bar{X}$ is an unbiased estimator.

   (b) Show that S^2 is a biased estimator.

   (c) If a population is infinite or if sampling is with replacement then the variance of the sampling distribution of mean denoted by $\sigma_{\bar{x}}{}^2$ is given by $V(\bar{x}) = E(\bar{X} - \mu)^2 = \frac{\sigma^2}{n} = \sigma_{\bar{x}}{}^2$

   Solution:

   (a) $E(\bar{X}) = E\left[\frac{\sum X_i}{n}\right]$

   $= \frac{1}{n} E[\sum X_i]$

   $= \frac{1}{n} \sum E(X_i)$

$$= \frac{1}{n}\sum \mu$$

$$= \frac{1}{n} \cdot n\mu$$

$$= \mu$$

$\therefore$ $\bar{X}$ is an unbiased estimator.

(b) $E(S^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2\right]$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}[V(X) - V(\bar{X})]$$

$$= \frac{1}{n} \cdot n\sigma^2 - \frac{1}{n} \cdot n\frac{\sigma^2}{n}$$

$$= \sigma^2 - \frac{\sigma^2}{n}$$

$$\Rightarrow E(S^2) = \frac{n\sigma^2 - \sigma^2}{n} = \frac{(n-1)\sigma^2}{n}$$

$\therefore E(S^2) \neq \sigma^2$ So S^2 is not unbiased estimator.

(c) Proof: $V(\bar{x}) = V\left(\frac{\sum X_i}{n}\right)$

$$= \frac{1}{n^2}V(X_1 + X_2 + \cdots + X_n)$$

$$= \frac{1}{n^2}[V(X_1) + V(X_2) + \cdots + V(X_n)]$$

$$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2)$$

$$= \frac{1}{n^2}n\sigma^2$$

$$= \frac{\sigma^2}{n} \quad \blacksquare$$

2. Efficiency

   The most efficiency estimator among a group of unbiased estimator is the one with the smallest variance. This concept refers to the sampling variability of an estimator.

Example: Consider the following unbiased estimator $\bar{X}$ (sample mean) and $\tilde{X}$ (sample median) with corresponding variance of $V(\bar{X}) = \frac{\sigma^2}{n}$, $V(\tilde{X}) = 1.576\frac{\sigma^2}{n}$, which of these is more efficiency?

Answer: $V(\bar{X})$

Relatively Efficiency of two unbiased estimators $\widehat{\theta_1}$ and $\widehat{\theta_2}$ can be calculated by taking the ratios: $\frac{V(\widehat{\theta_1})}{V(\widehat{\theta_2})}$ where $V(\widehat{\theta_1})$ is the smallest variance.

3. Sufficiency

   An estimator is sufficient if it uses all the information that a sample can provide about a population parameter and no other estimator can provide additional information.

4. Consistency

   An estimator is consistent if as the sample size becomes larger, the probability increases that the estimates will approach the true value of the population parameter. Alternatively, $\hat{\theta}$ is consistent if it satisfies the following:

   (i) $V(\hat{\theta}) \to 0$ as $n \to \infty$.

   (ii) $\hat{\theta}$ becomes unbiased as $n \to \infty$.

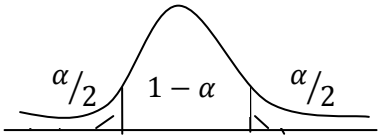
   Example: Let $x_1, x_2, \ldots, x_n$ be random sample of a random variable x with mean μ and finite variance σ^2, show that $\bar{X}$ is a consistent estimate for μ.

   Solution: $\bar{X}$ is already unbiased (shown above) so (ii) satisfies.

   $V(\bar{X}) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$

INTERVAL ESTIMATION

This involves specifying a range of values on which we can assert with a certain degree of confidence that a population parameter will fall within the interval. The confidence that we have a population parameter θ will fall within confidence interval is $1-\alpha$, where α is the probability that the interval does not contain θ.



Confidence Interval (C.I)$= 1-\alpha$

C.I $+\alpha = 1$

Confidence Interval for μ when σ is known

A confidence Interval is constructed on the basis of sample information. It depends on the size of α which is the level of risk that the interval may be wrong. Assume the population variance σ^2 is known and the population is normal, then $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{Or} \quad \bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ for short}$$

[Derived from $-Z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\alpha/2}$]

Example: An experiment was carried out to estimate the average number of heart beat per minute for a certain population. Under the conditions of the experiment, the average number of heart beats per minute for 49 subjects was found to be 130. If it is reasonable to assume that these 49 patients constitute a random sample, and the population is normally distributed with a standard deviation of 10. Determine

(a) The 90% confidence interval for μ

(b) The 95% confidence interval for μ

(c) The 99% confidence interval for μ

Solution:

(a) $\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. $\bar{X} = 130, n = 49, \sigma = 10, 1-\alpha = 0.90 \Rightarrow \alpha = 0.1,\ {}^{\alpha}/_{2} = 0.05$

$$130 \pm Z_{\alpha/2}\left[\frac{10}{\sqrt{49}}\right]$$

$$130 \pm Z_{0.05}\left(\frac{10}{7}\right)$$

$$130 \pm 1.645(1.43)$$

$$130 \pm 2.35$$

$$(127.65, 132.35)$$

(b) $\alpha = 0.05,\ {}^{\alpha}/_{2} = 0.025$

$$130 \pm Z_{0.025}(1.43)$$

$$130 \pm 1.96(1.43)$$

$$(127.2, 132.8)$$

(c) $\alpha = 0.01,\ {}^{\alpha}/_{2} = 0.005$

$$130 \pm Z_{0.005}(1.43)$$

$$130 \pm 2.578(1.43)$$

$$(126.31, 133.69)$$

When the population variance is unknown, the distribution for construct confidence interval for μ is the **t-distribution**. Here an estimate S, must be calculate from the sample to substitute for the unknown standard deviation. T-distribution is also based on population is normal. A $100(1-\alpha)\%$ confidence interval for μ when σ is unknown is given by

$$\bar{X} \pm t_{{}^{\alpha}/_{2},(n-1)}\frac{S}{\sqrt{n}} \qquad \text{(or when } n < 30\text{)}$$

Example: A sample of 25 teenager – old boys yielded a mean weight and standard deviation of 73 and 10 pounds respectively. Assuming a normally distributed population, find

(i) 90%

(ii) 99% confidence interval for the μ of the population.

Solution:

(i) $\bar{X} = 73,\ n = 25,\ S = 10, \alpha = 0.1,\ {}^{\alpha}/_{2} = 0.05$

$$73 \pm t_{0.05,24}\left[\frac{10}{\sqrt{25}}\right]$$

$$73 \pm 1.71(2)$$

73 ± 3.42

$(69.58, 76.42)$

(ii) $73 \pm t_{0.005,24}(2)$

$73 \pm 2.8(2)$

$(67.4, 78.6)$

Confidence Interval for a population proportion

To estimate a population proportion, we proceed in the same manner as when estimating a population mean. A sample is drawn from the population of interest and the sample proportion $\bar{p}$ is computed. The sample proportion is used as the point estimator of the population proportion. Assume normally population, when np and $n(1-p) > 5$ so that

$\bar{p} \pm Z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$ where $\bar{p} = {}^{x}/_{n}$.

Example: A survey was conducted to study the dental health practices and attitudes of certain urban adult population. Of 300 adults interviewed, 123 said that they regularly had a dental check up twice a year. Obtain a 95% confidence interval for $\bar{p}$ based on this data.

Solution: $n = 300,\ \bar{p} = {}^{x}/_{n} = \frac{123}{300} = 0.41, n\bar{p} = 300 \times 0.41 = 123, nq = n(1-p) = 300(0.59) = 177, \alpha = 0.05, {}^{\alpha}/_{2} = 0.025$

$\bar{p} \pm Z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

$0.41 \pm Z_{0.025}\sqrt{\frac{0.41(0.59)}{300}}$

$0.41 \pm 1.96(0.028)$

0.41 ± 0.055

$(0.36, 0.47)$

Exercise 8

A medical record Liberian drew a random sample of 100 patients' charts and found that in 8% of them, the face sheet had at least one item of information contradiction to other information in the record. Construct the 90%, 95% and 99% confidence interval for the population of charts containing such discrepancies.

Confidence Interval for the population variance of a normally distributed population

A $100(1-\alpha)\%$ confidence interval for σ^2 is given by

$$\frac{(n-1)S^2}{X^2{}_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{X^2{}_{\alpha/2}}$$

Example: Serum amylase determinations were on a sample of 15 apparently normally subject. The sample yielded a mean of 16units/100ml and a standard deviation of 35units/100ml. The population variance was unknown, construct the 95% confidence interval for σ^2.

Solution: $n = 15, S^2 = 35^2 = 1225, \alpha = 0.05, {}^{\alpha}/_{2} = 0.0025, 1 - {}^{\alpha}/_{2} = 0.975$

$$\frac{14(1225)}{X^2{}_{0.975}} \leq \sigma^2 \leq \frac{14(1225)}{X^2{}_{0.025}}$$

$$\frac{17150}{X^2{}_{0.975}} \leq \sigma^2 \leq \frac{17150}{X^2{}_{0.025}}$$

$$\frac{17150}{5.6287} \leq \sigma^2 \leq \frac{17150}{26.1190}$$

i.e. $656.6 \leq \sigma^2 \leq 3046.9$

HYPOTHESIS TESTING

This is another very important aspect of statistical inference. It involves testing the validity of a statistical statement about a population parameter based on the available information at a given level of significance.

Basic Definitions

Statistical hypothesis: This is a statistical statement which may or may not be true concerning one or more populations.

Null or True hypothesis (H_0): This is an assertion that a parameter in a statistical model takes a particular value. This hypothesis expresses no difference between the observed value and the hypothesis value. It is denoted by $H_0: \theta_1 = \theta_0$ where θ_0 is the observed value and θ_1 is the true value.

Alternative hypothesis (H_1): This hypothesis expresses a deviation in the Null hypothesis. It states that the true value deviates from the observed value. It is denoted by $H_1: \theta_1 \neq \theta_0$ or $\theta_1 > \theta_0$ or $\theta_1 < \theta_0$.

Test Statistic: This is a calculated quantity from the given information which when compares with the tabulated value is used to take a decision about the hypothesis being tested. E.g. The test statistic for single mean: $Z_c = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ when $n \geq 30$, $t_c = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ when $n < 30$.

Note: The choice of the appropriate test statistic depends on the type of estimate and sample distribution.

Critical Region: This is the subset of the sample space which leads to the rejection of the null hypothesis under consideration. It is the set of all values whose total probability is small on the null hypothesis, which is better explained by the alternative hypothesis.

Significance Level: It is the probability of taking a wrong decision or the probability of making an error. There are two types of errors in hypothesis testing:

TYPE I ERROR: This is the error of rejecting the null hypothesis when it should be accepted. It is denoted by α; $0 < \alpha < 1$.

$$\alpha = P(Reject H_0 : H_0 is\ true)$$

TYPE II ERROR: This occurs when H_0 is accepted when it is false. This probability is denoted by β.

$$\beta = P(Accept\ H_0 : H_0\ is\ false)$$

$$1 - \beta = P(Reject\ H_0 : H_0\ is\ false)$$

$$\Rightarrow P(Accept\ H_0 : H_0\ is\ true) + P(Reject\ H_0 : H_0\ is\ true) = 1$$

$$\therefore P(Accept\ H_0 : H_0\ is\ true) = 1 - \alpha.$$

This could be arranged in a table as follows:

	H_0 is true	H_0 is false
Accept H_0	$1 - \alpha$	β
Reject H_0	α	$1 - \beta$

Structure of the Test: A statistical test of hypothesis may be structured as one – tail or two – tail test.

One – tail test: This is one in which the alternative hypothesis is well specified.

E.g. $H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$ or $\mu < \mu_0$

Two – tail test: In this case, the alternative hypothesis is not well specified.

E.g. $H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$.

Standard Format for Hypothesis Testing

1. Formulate the Null and Alternative hypotheses;
2. Determine a suitable test statistics i.e. choosing the appropriate random variable to use in deciding to accept H_0 or not.

Unknown Parameter	Appropriate Test Statistic
"μ" σ known, population normal	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
"μ" σ known, population normal but $n \geq 30$	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
"μ" σ unknown, population normal ($n < 30$)	$t = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ with $(n-1)$ df
"σ^2" Population variance, population normal	$X^2 = \frac{(n-1)S^2}{{\sigma_0}^2}$ with $(n-1)$ df
"ρ" Binomial proportion	$Z = \frac{\bar{p}-p_0}{\sqrt{\frac{p_0 q_0}{n}}}$, $p_0 = \frac{x}{n}$

3. Determine the critical region (or rejection region) using the table of the statistic. But the value of α must be known i.e. one or two – tail test.
4. Compute the value of the test statistic based on the sample information
5. Make the statistical decision and interpretation.

Test about a single population mean, μ

Example 1: A researcher is interested in the mean level of some enzyme in a certain population. The data available to the researcher are the enzyme determination made on a sample of 10 individuals from the population of interest and the sample mean is 22. Assumed the sample came from a population

that is normally distributed with a known variance 45. Can the researcher conclude that mean enzyme level in this population is different from 25? Take $\alpha = 0.05$.

Solution: Hypothesis $H_0: \mu_0 = 25$

$H_1: \mu_0 \neq 25$

$\alpha = 0.05, \quad n = 10, \quad \bar{X} = 22, \quad {}^{\alpha}/_{2} = 0.025\ (two - tail\ test)$

Test Statistic $= Z_c = \frac{\bar{X} - \mu_0}{{}^{\sigma}/_{\sqrt{n}}}$

$$Z_c = \frac{22-25}{\sqrt{\frac{45}{10}}} = \frac{-3}{2.1213} = -1.4142$$

From the Normal distribution table: $Z_{0.025} = 1.96$

Decision: Since $Z_c = -1.4142 < Z_{0.025} = 1.96$, we accept H_0 (the null hypothesis)

Conclusion: We then conclude that the researcher can conclude that mean enzyme in the population is not different from 25.

Example 2: A manufacturer of multi-vitamin tablet claims that riboflavin content of his tablets is greater than 2.49mg. A check by the food and drug administration using 82 tablets shows a mean riboflavin content of 2.52mg with standard deviation of 0.18mg; should the manufacturer claims be rejected at 1% significant level?

Solution: Hypothesis $H_0: \mu > 2.49$

$H_1: \mu \leq 2.49$

$1 - \alpha = 0.1 \Rightarrow \alpha = 0.99$

Test Statistic:

$$Z_c = \frac{\bar{X} - \mu_0}{{}^{\sigma}/_{\sqrt{n}}} = \frac{2.52-2.49}{{}^{0.18}/_{\sqrt{82}}} = 1.507$$

$Z_{cal} = 1.507, \ Z_{0.99} = 2.33$

Decision: Since $Z_{cal} < Z_{0.99}$, we accept H_0 and conclude that the manufacturer's claim be accepted at 1% level of significance.

σ unknown

Example 3: A certain drug company claims that one brand of headache tablet is capable of curing headache in one hour. A random variable of 16 headache patients is given the tablets. The mean curing time was found to be one hour and nine minutes while the standard deviation of the 16 times was eight minutes. Does the data support the company's claim or not at 95% level of significance?

Solution: $H_0: \mu_0 = 1.00$

$H_1: \mu_0 \neq 1.00$

$$t_{cal} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$= \frac{1.15 - 1.00}{0.13/\sqrt{16}}$$

$$= \frac{0.15}{0.0325} = 4.6154$$

$t_{\frac{\alpha}{2}, n-1} = t_{0.025,15} = 2.13$

Decision: Since $t_{cal} > t_{0.025,15}$, we reject H_0 and conclude that the data does not support the company's claim.

Hypothesis Testing: The Difference between two populations (variance known)

Test Statistic: $Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$

Hypothesis

$H_0: \mu_1 - \mu_2 = 0$ vs $H_1: \mu_1 - \mu_2 \neq 0$

$H_0: \mu_1 - \mu_2 > 0$ vs $H_1: \mu_1 - \mu_2 < 0$

$H_0: \mu_1 - \mu_2 \leq 0$ vs $H_1: \mu_1 - \mu_2 > 0$

Example: In a large hospital for the treatments of the mutually retarded, a sample of 12 individuals with mongolism yielded a mean serum uric acid value of 4.5mg/100ml. In general hospital, a sample of 15 normal individuals of the same age and sex were found to have a mean value of 3.4mg/100ml. if it is reasonable

to assume that the two populations of values are normally distributed with variances equal to one. Do this data provide sufficient evidence to indicate a difference in the mean levels between mongolism, using $\alpha = 0.05$?

Solution: $H_0: \mu_1 = \mu_2 \quad \Rightarrow \mu_1 - \mu_2 = 0$

$H_1: \mu_1 \neq \mu_2 \quad \Rightarrow \mu_1 - \mu_2 \neq 0$

$$Z_{cal} = \frac{(4.5-3.4)-0}{\sqrt{\frac{1}{12}+\frac{1}{15}}}$$

$$= \frac{1.1}{\sqrt{0.15}} = 2.8402$$

$$Z_{\alpha/2} = Z_{0.025} = \pm 1.96$$

Decision: We will reject H_0 since $Z_{cal} > Z_{\alpha/2}$ and conclude that the means are not equal.

The Difference between two populations means (Variance unknown but assumed equal)

Test Statistic: $t_{cal} = \frac{(\overline{x_1})-(\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{{S_p}^2}{n_1}+\frac{{S_p}^2}{n_2}}}$ where ${S_p}^2 = \frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2}$

Example: Serum amylase determinations were made on a sample of 15 apparent normal subjects. The sample yielded a mean of 96units/100ml and a standard deviation of 35units/100ml. These determinations were also made on 22 hospitalized subjects with mean and standard deviation from the second room as 120 and 40units/100ml respectively. Would it be justify in concluding that population means are difference whose $\alpha = 0.05$?

Solution: $H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

$${S_p}^2 = \frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2} = \frac{14(1225)+21(1600)}{35} = 1450$$

$$t_{cal} = \frac{(\overline{x_1})-(\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{{S_p}^2}{n_1}+\frac{{S_p}^2}{n_2}}} = \frac{(96-120)-0}{\sqrt{1450\left(\frac{1}{15}+\frac{1}{22}\right)}} = -1.85$$

$$t_{\alpha/2,(n_1+n_2-2)} = t_{0.025,35} = \pm 1.96$$

We accept H_0 and conclude that no difference in the means of the two populations.

Testing Hypothesis for population variance

Example: A sample of 25 ten years old girls yielded a mean weight and standard deviation of 73 and 10 pounds respectively. Should one conclude that population variance is different from 150 at $\alpha = 0.05$?

Solution: $H_0: \sigma^2 = 150$

$H_1: \sigma^2 \neq 150$

$X^2{}_{cal} = \frac{(n-1)S^2}{\sigma^2} = \frac{24(100)}{150} = 16$

$X^2{}_{\alpha/2,(n-1)} = X^2{}_{0.025,24} = 39.3641$

Decision: Since $X^2{}_{cal} < X^2{}_{0.025,24}$, we accept H_0 and conclude that $\sigma^2 = 150$ at $\alpha = 0.05$ level of significance.

GOODNESS OF FIT

This test is used to compare the observed frequencies and the frequencies we might expect (expected frequency) from a given theoretical explanation of the phenomenon under investigation. A measure of this discrepancy is given by X^2 (Chi square) statistic define as:

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

The statistic defined above has an approximate X^2 −distribution with degree of freedom equal to:

(i) $k - 1$ (if expected frequencies can be computed without having to estimate population parameters from sample statistics)

(ii) $k - 1 - m$ (if expected frequencies can be computed only by estimating m population).

To test the null hypothesis H_0 which specifies certain (population) proportions associated with each class or category, we compute the value of X^2 and thereafter make necessary conclusion.

Example 1: The distribution of final grade given by STS 202 lecturers in the past was 10% As, 20% Bs, 30% Cs, 25% Ds and 15% Fs. A new lecturer gave the following grades for the second semester:

Category	A	B	C	D	F
Observed	12	20	26	14	8

Is there sufficient evidence to suggest that the new lecturer's policy is different from that of the formal lecturers? Use a 95% level of significance.

Solution: $H_0: p_1 = 0.10, p_2 = 0.20, p_3 = 0.30, p_4 = 0.25, p_5 = 0.15$

$H_1: H_0$ is not true

$n = 12 + 20 + 26 + 14 + 8 = 80$

Category	**Observed frequency**	**Expected frequency**
A	12	$\frac{10}{100} \times 80 = 8$
B	20	$\frac{20}{100} \times 80 = 16$
C	26	$\frac{30}{100} \times 80 = 24$
D	14	$\frac{25}{100} \times 80 = 20$
F	8	$\frac{15}{100} \times 80 = 12$

$$X^2{}_{cal} = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(12-8)^2}{8} + \frac{(20-16)^2}{16} + \frac{(26-24)^2}{24} + \frac{(14-20)^2}{20} + \frac{(8-12)^2}{12}$$

$$= 6.3$$

Since we have five classes (grades), degree of freedom $= k - 1 = 5 - 1 = 4$ and $\alpha = 0.05$.

$X^2{}_{0.05,4} = 9.49$. Hence, since $X^2{}_{cal} < X^2{}_{0.05,4}$, we do not reject H_0, meaning that the data do not suggest that the new lecturer's grading policy is different.

CONTINGENCY TABLE ANALYSIS

Another use of the X^2 statistic is in contingency testing, where n randomly selected items are classified to two different criteria. Here, it is desired to determine whether some protective measure or sample preparation technique has been effective or not. Instead of $1 \times k$ table in the goodness fit, we will have two – way classification table or $h \times k$ tables in which the observed frequencies occupy h rows and k column. Such tables are called contingency tables. Correspond to each observed frequency is the expected frequency which is computed subject to some hypothesis according to the rules of probability.

Contingency Table

Row	**1**	**2**	⋯	**c**	**Row Total**
1	O_{11}	O_{12}	⋯	O_{1c}	$\boldsymbol{R_1}$
2	O_{21}	O_{22}	⋯	O_{2c}	$\boldsymbol{R_2}$
⋮	⋮	⋮	⋮	⋮	⋮
r	O_{r1}	O_{22}	⋯	O_{rc}	$\boldsymbol{R_r}$
Column Total	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$	⋯	$\boldsymbol{C_c}$	$\boldsymbol{N}$

The test statistic is:

$$X^2 = \sum_{i-1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The test statistic so defined has $(r-1)(c-1)$ degree of freedom.

Example 2: Consider the case where pressure gauges are being hydraulically tested by 3 inspectors prior to shipment. It has been noted that their acceptance and rejection for some period of time have been as follows:

	Inspectors			
	I	II	III	Totals
Passed	150	50	100	300
Failed	20	10	30	60
Totals	170	60	130	360

Test the hypothesis that all inspectors are equally stringent, take $\alpha = 0.05$.

Solution: H_0: *All inspectors are equally stringent*

H_1: H_0 *not true*

Test statistic: $X^2 = \sum_{i-1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

E_{ij}: $E_{11} = \frac{300(170)}{360} = 141.67$

$E_{12} = \frac{300(60)}{360} = 50$

$E_{13} = \frac{300(130)}{360} = 108.33$

$E_{21} = \frac{60(170)}{360} = 28.33$

$E_{22} = \frac{60(60)}{360} = 10$

$E_{23} = \frac{60(130)}{360} = 21.67$

Thus, the computed X^2 value is

$$X^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{13} - E_{13})^2}{E_{13}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}} + \frac{(O_{23} - E_{23})^2}{E_{23}}$$

$$= \frac{(150-141.67)^2}{141.67} + \cdots + \frac{(30-21.67)^2}{21.67} = 6.7817$$

Critical value: ${X^2}_{0.05,(2-1)(3-1)} = {X^2}_{0.05,2} = 5.99$

Decision: Since ${X^2}_{cal} > {X^2}_{tab}$, the null hypothesis is rejected and we conclude that some inspectors are more strict and demanding than the others.

Exercise 9

Random samples of 1200 students who live in Saudi hall of the Crescent University were asked their daily eating habits by means of questionnaire X. Similarly, information was obtained from another sample of 1050 students living in Yola hall of the same institution, by means of questionnaire Y. The results were as follows. Can the difference in these distributions be purely due to chance? Support your answer from a statistical point of view.

	Number of Students	
	X	Y
Miss no meal per week	430	500
Miss 1 – 4 meals per week	500	300
Miss 5 – 8 meals per week	200	225
Miss 9 or more meals per week	70	25

INTRODUCTION TO EXPERIMENTAL DESIGN

Experimental design has been defined as the order in which an experiment is runs such that its analysis will lead to valid statistical inference. The design of the experiment has three essential components.

(a) Estimate of the error

(b) Controls of the error

(c) Proper interpretation of error

Experimental design methods are also useful in engineering design when new products are developed and existing one improved. Some typical applications of statistically design experiment in engineering include:

1. Evaluating and comparison of basic design configuration.
2. Evaluation of different materials
3. Selection of designed parameter so that the product will work well under a wild range of field condition.
4. Determination of key product design parameter that impact product performance.

Terms in Experimental Design

Factor: An independent variable of interest under investigation

Factor level

Treatment

Experimental unit: This is the unit in which a single treatment combination is applied in a single of the experiment.

Replicator

Grouping

Blocking

Randomization

Control

Types of Experimental Design

Designs are classified according to their classification factor.

(1) Completely Randomized Design (CRD)

This is a design in which treatments are assign complete at random such that each experimental unit have the same chance of receiving any one treatment.

(2) Randomized Completely Block Design (RCBD)

In this design, the experimental unit are divided into two homogeneous groups such that treatment for each block are expected because variations are kept within each block.

ANALYSIS OF VARIANCE (ANOVA)

The technique known as analysis of variance employs tests based on variance – ratios to determine whether or not significant differences exist among the means of several groups of observations, where each group follows a normal distribution. ANOVA is particularly useful when the basic differences between the groups cannot be stated quantitatively. A one – way analysis of variance is used to determine the effect of one independent variable on a dependent variable. A two – way analysis of variance is used to determine the effects of two independent variables on a dependent variable, etc. As the number of independent increases, the calculations become much more complex and are best carried out on a computer. The term independent variable is what also referred to as factor or treatment.

ONE WAY ANOVA (Completely Randomized Design)

This model is used when we wish to test the equality of k population means. The procedure is based on assumptions that each of k groups of observation is a random sample from a normally distribution and that the population variance (σ^2) is constant among the groups.

The statistical model for one – way classification of ANOVA is

$$x_{ij} = \mu + \tau_j + e_{ij} \qquad i = 1, \dots, n \ , \ j = 1, \dots, k$$

Where

x_{ij} = the ith observation receiving jth treatment

μ = Overall mean (grand mean)

τ_j = treatment effect

e_{ij} = Error term (random error)

Assumptions

1. The k sets of observed data constitute k random samples for the respective populations.
2. Each of the population from which the sample is normally distributed with mean μ_i and variance ${\sigma_j}^2$ respectively.
3. Each of the population has the same variance i.e. ${\sigma_1}^2 = {\sigma_2}^2 = \cdots = {\sigma_k}^2 = \sigma^2$.
4. The τ_j are unknown constants and the sum is equal to zero i.e. $\sum \tau_j = 0$.

Notation

Treatment

1	2	3	…	k	
x_{11}	x_{12}	x_{13}	…	x_{1k}	
x_{21}	x_{22}	x_{23}	…	x_{2k}	
⋮	⋮	⋮	⋮	⋮	
x_{n1}	x_{n2}	x_{n3}	… …	x_{nk}	
Total					
$T_{.1}$	$T_{.2}$	$T_{.3}$	…	$T_{.k}$	$T_{..}$
$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.3}$	…	$\bar{x}_{.k}$	$\bar{x}_{..}$

Hypothesis

$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

H_1: Not all μ_j are equal

Alternatively:

H_0: $\tau_j = 0$

H_1: Not all τ_j equal 0.

CALCULATION

The following calculations are going to be used.

x_{ij} = ith observation receiving jth treatment

$T_{.j} = \sum_{i=1}^{n} x_{ij}$ = Total of the jth column

$\bar{x}_{.j} = \frac{T_{.j}}{n}$ = mean of the jth column

$T_{..} = \sum_{j=1}^{k} T_{.j} = \sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}$ = Total of all observation

$\bar{x}_{..} = \frac{T_{..}}{N}$

The Total Sum of Squares (TSS)

$$SS_{total} = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_{..}\right)^2$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}{}^2 - \frac{T_{..}{}^2}{N}$$

Among/Treatment/Between group Sum of Squares

$$SS_{treatment} = \sum_{j=1}^{k} \frac{T_{.j}{}^2}{n_j} - \frac{T_{..}{}^2}{N}$$

Within/Error/Residual groups Sum of Squares

$$SS_{error} = \sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}{}^2 - \sum_{j=1}^{k}\left(\frac{T_{.j}}{n_j}\right)^2 \text{ Or } = SS_{total} - SS_{trt}$$

ANAOVA TABLE

Source	d.f	S.S	M.S	F
Treatment	t-1	SS_{trt}	SS_{trt}/t-1	T.M.S/E.M.S
Error	N-t	SS_{error}	SS_{error}/N-t	
Total	N-1	SS_{total}		

Example 1: Given the following five treatments A, B, C, D and E of 3 values each. Perform an analysis of variance using $\alpha = 0.05$.

A	3	2	4
B	5	8	8
C	7	8	6
D	6	8	7
E	4	9	5

Solution:

$$\text{SS}_{\text{total}} = \sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}{}^2 - \frac{T_{..}{}^2}{N}$$

$$= 3^2 + 2^2 + \cdots + 5^2 - \frac{8100}{15}$$

$$= 602 - \frac{8100}{15}$$

$$= 62$$

$$\text{SS}_{\text{trt}} = \sum_{i=1}^{n} \frac{T_{i.}{}^2}{n_i} - \frac{T_{..}{}^2}{N}$$

$$= \frac{9^2}{3} + \frac{21^2}{3} + \cdots + \frac{18^2}{3} - 540$$

$$= 576 - 540$$

$$= 36$$

$$\text{SS}_{\text{error}} = \text{SS}_{\text{total}} - \text{SS}_{\text{trt}}$$

$$= 62 - 36$$

$$= 26$$

ANAOVA TABLE

Source	d.f	S.S	M.S	F
Treatment	5-1=4	36	36/4 = 9	9.0/2.6 = 3.46
Error	15-5=10	26	26/10 =2.6	
Total	15-1=14	62		

Critical value: $F_{tab} = F_{\alpha,v_1,v_2} = F_{0.05,4,10} = 3.48$

Decision: Since $F_{cal} < F_{tab}$, we accept H_0 and conclude that the treatment means are equal or no statistical difference in the treatment means.

Example 2: Specimens were randomly selected from three production processes for steal with each process using a different percentage of carbon, independent observation on tensile strength were made with one observation coming from each specimen. The data are as follows with measurement in thousand (psi).

Process	A	B	C
	32.1	38.9	42.8
	34.2	40.2	44.6
	29.6	41.4	
		39.5	

Is there evidence to say that the mean tensile strength differs for the 3 processes? Take $\alpha = 5\%$.

Solution: Model

$$x_{ij} = \mu + \tau_j + e_{ij}$$

Hypothesis

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1: Not all $\mu's$ are equal

Calculations:

$$SS_{total} = \sum_{j=1}^{k} \sum_{i=1}^{n} x_{ij}{}^2 - \frac{T_{..}{}^2}{N}$$

$$= 32.1^2 + \cdots + 44.6^2 - \frac{(343.3)^2}{9}$$

$$= 205.68$$

$$SS_{trt} = \sum_{i=1}^{n} \frac{T_{.j}{}^2}{n_j} - \frac{T_{..}{}^2}{N}$$

$$= \frac{95.9^2}{3} + \frac{160^2}{4} + \frac{87.4^2}{2} - 13094.99$$

$$= 189.99$$

$$SS_{error} = SS_{total} - SS_{trt}$$

$$= 205.68 - 189.99$$

$= 15.69$

ANAOVA TABLE

Source	S.S	d.f	M.S	F
Treatment	189.99	2	94.995	94.995/2.615 = 36.33
Error	15.69	6	2.615	
Total	205.68	8		

Critical value: $F_{tab} = F_{\alpha,v_1,v_2} = F_{0.05,2,6} = 5.14$

Decision: Since $F_{cal} > F_{tab}$, we reject H_0 and conclude that there is an evidence to say that the mean tensile strength differ for the three processes.

TWO – WAY ANOVA (Randomized Complete Block Design)

	Design treatment						
Block	1	2	3	…	k	Total T.	Mean $\bar{x}$
1	x_{11}	x_{12}	x_{13}	…	x_{1k}	T_1	$\bar{x}_1$
2	x_{21}	x_{22}	x_{23}	…	x_{2k}	T_2	$\bar{x}_2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	x_{n1}	x_{n2}	x_{n3}	…	x_{nk}	T_n	$\bar{x}_n$
Total	$T_{.1}$	$T_{.2}$	$T_{.3}$	…	$T_{.k}$	$T_{..}$	
Mean	$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.3}$	…	$\bar{x}_{.k}$		$\bar{x}_{..}$

Total of the ith block (Replicate):

$$T_{i.} = \sum_{j=1}^{k} x_{ij}$$

Mean of the ith block:

$$\bar{x}_i = \sum_{j=1}^{k} \frac{x_{ij}}{k}$$

Grand total:

$$T_{..} = \sum_{j=1}^{k} T_{.j}$$

The statistical model for two – way classification of ANOVA is

$$x_{ij} = \mu + \beta_i + \tau_j + e_{ij} \qquad i = 1, \dots, n \ , \ j = 1, \dots, k$$

Where

x_{ij} = Typical value from the overall population

μ = Unknown constant

β_i = A block effect

τ_j = Treatment effect

e_{ij} = Error term

Assumptions

1. $e_{ij} \sim N(0, \sigma^2)$
2. $\sum \tau_j = \sum \beta_i = 0.$

Calculation

$$\text{SS}_{\text{total}} = \sum_{j=1}^{k} \sum_{i=1}^{n} x_{ij}{}^2 - \frac{T_{..}{}^2}{N}$$

$$\text{SS}_{\text{block}} = \sum_{i=1}^{n} \frac{T_{i.}{}^2}{n_i} - \frac{T_{..}{}^2}{N}$$

$$\text{SS}_{\text{treatment}} = \sum_{j=1}^{k} \frac{T_{.j}{}^2}{n_j} - \frac{T_{..}{}^2}{N}$$

$$\text{SS}_{\text{error}} = \text{SS}_{\text{total}} - \text{SS}_{\text{trt}} - \text{SS}_{\text{block}}$$

Example: Samples of 200 machined parts were selected from the one week output of machine shop that employs three machinists. The parts were inspected to determine whether or not, they are defective and are categorized according to which machinist did the work. The results are as follows:

	Machines		
	A	B	C
Defective	10	8	14
Non – Defective	52	60	56

Is the Defective, Non – Defective classification independent of machinist? Conduct a test at 1% level of significant.

Solution: Model $x_{ij} = \mu + \beta_i + \tau_j + e_{ij}$

Hypothesis $H_0: T_j = 0 \ vs \ H_1: Not\ all\ T_j = 0$

Calculation

$$SS_{total} = \sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}{}^2 - \frac{T_{..}{}^2}{N}$$

$$= 10^2 + 52^2 + 8^2 + 60^2 + 14^2 + 56^2 - \frac{(200)^2}{6}$$

$$= 3133.33$$

$$SS_{block} = \sum_{i=1}^{n} \frac{T_{i.}{}^2}{n_i} - \frac{T_{..}{}^2}{N}$$

$$= \frac{32^2+168^2}{3} - 6666.64$$

$$= 3082.66$$

$$SS_{trt} = \sum_{j=1}^{k} \frac{T_{.j}{}^2}{n_j} - \frac{T_{..}{}^2}{N}$$

$$= \frac{62^2+68^2+70^2}{2} - 6666.64$$

$$= 17.33$$

$$SS_{error} = SS_{total} - SS_{trt} - SS_{block}$$

$$= 3133.33 - 17.33 - 3082$$

$$= 33.34$$

ANAOVA TABLE

Source	S.S	d.f	M.S	VR
Treatment	17.33	2	8.665	3082.66/16.67 = 184.92
Block	3082.66	1	3082.66	
Error	33.34	2	16.67	
Total	3133.33	5		

$F_{1,2,0.01} = 98.50$

Decision: We reject H_0 and conclude that defective, non-defective classification dependent of machinist.

Exercise 10

Three kinds of tomato are grown. The yields in grams, after harvesting are given in the table below.

Kind		
A	B	C
7	10	7
8	11	10
8	12	11
6	10	6
10	9	
12		
9		

Carry out the analysis of variance for the data at 95% level of significance.

NON PARAMETRIC TEST

Statistical method of inference that does not dependent on stringent assumption of population measure is called non parametric methods. They are used

(i) When we do not know the mean of the distribution population
(ii) When we need a result in a hurry
(iii) When data measured in a scale lower than that of the parametric method.

Non parametric tests were the test developed to deal with situations where the population distributions are non-normal or unknown, or when little is known about the distributions of the populations under study, or when these distributions do not meet the requirements necessary for the use of parametric tests especially when the sample size is small (less than 30). Non parametric

method employs the median of the population and the method of hypothesis testing in conducting its test. Some of parametric tests include sign test, Wilcox on signed-rank test, Mann-Whitney U test, runs test, etc.

THE SIGN TEST

This test is used to study the median of a population and to compare two populations when the samples are dependent. We usually assume that the data are continuous.

Testing a single population value (M_o)

Procedure

1. Hypothesis $H_0: Md = M_o$ $H_1: Md \neq M_o$
2. Test Statistic: $X =$ Number of data values in the sample above the median value given in the null hypothesis H_0.
3. Significant level: Choose appropriate, say α.
4. Critical Region: When H_0 is true, X has a binomial distribution with n sample size and $p = 0.5$. We use table of binomial probabilities to find the critical region. (Note $\neq, >, <$) If α is the desire level of significance, we choose critical value so that the probability that X falls in the critical region is as close to α as possible.
5. Decision: We check whether the observed value, X is in the critical region or not; if X falls into the critical region, we reject H_0, otherwise, we accept H_0.

Comparison of two populations

The sign test may be used to compare two populations when the samples are dependent (i.e. the values from the two samples occur in pairs).

THE WILCOXON SIGNED-RANK TEST

Although the sign test is very simple to use, it is not a very sensitive test. Sometimes it will fail to reject a false null hypothesis when another test would be successful in detection of the false of the null hypothesis. This is because, sign test throws away a good deal of information about the data-it ignores the magnitude of the data value, it only uses the information about whether the data value is above the conjectured value of the median or not.

The Wilcox on signed-rank test is better than sign test because it uses more information. We use this test to investigate a single population median and to compare two populations using a paired experiment.

Procedure for a single population median

Let M be the value of the median in question that appears in the null hypothesis. Calculate D = X – M for each data value X. we then rank the values of D and place a minus sign in front of each rank corresponding to a negative difference D. Let W^+ = Sum of the positive ranks. W^- = Absolute value of sum of negative ranks.

(a) To test: $H_0: Md = M \quad H_1: Md > M$

Use W^- as test statistic. Find the critical value C for a one-tailed test with desire significant level in the table of critical value for the Wilcox on signed-rank test. If $W^- \leq C$ reject H_0.

(b) To test $H_0: Md = M \quad H_1: Md < M$

Use W^+ as test statistic and if $W^+ \leq C$ reject H_0.

(c) To test $H_0: Md = M \quad H_1: Md \neq M$

If either $W^+ \leq C$ or $W^- \leq C$, reject H_0. This means that we can use the minimum of W^+ or W^- as a test statistic. We denote this value by W. If $W \leq C$, reject H_0.