ambitious baba .com
INDIA'S LARGEST TEST PREP PLATFORM

Module-A ABM

CAIIB PAPER-1

STATISTICS

CAIIB Paper 1 (ABM) Module A: STATISTICS

Index

CAIIB Paper 1 (ABM) Module A Unit 1: Definition of Statistics, Importance & Limitations & Data Collection, Classification & Tabulation

Introduction

The word 'Statistics' has been derived from the

- ❑ Latin word 'statisticum',
- ❑ Italian word 'statistia'
- ❑ German word 'statistik',

- Each of which means a group of numbers or figures that represent some information of human interest.
- First used by professor Achen well in 1749 to refer to the subject-matter as a whole.
- Achen well defined statistics as the Political Science of many countries.
- In the early years statistics is to be used only by the kings to collect facts about the state, revenue of the state or the people in the state of administrative or political purpose.
- Gradually the use of statistics which means data or information has increased and widened.

- It is now used in almost in all the fields of human knowledge and skills like Business, Commerce, Economics, Social Sciences, Politics, Planning, Medicine and other sciences, physical as well as natural.
- In many practical situations in life, we come across different types of data which are needed to be understood, analysed, compared and interpreted correctly.
- For example, in a college we need to analyse the data of marks obtained, in a hospital we need to analyse the data of number of patients having different diseases, rate of mortality, Different types of data need to be analysed in Economics, Government and Private organisations, Sports and in many other fields.

Statistical analysis of data can be comprised of four distinct phases:

- **Collection of data**: In this first stage of investigation, numerical data is collected from different published or unpublished sources, primary or secondary.
- **Classification and Tabulation of data**: The raw data collected is to be represented properly for further calculations. The raw data is divided into different groups or classes and represented in a form of a table.
- **Analysis of data**: Classified and Tabulated data is analysed using different formulas and methods according to purpose of the study or investigation.
- **Interpretation of data**: At the final stage, relevant conclusions are drawn after the data is thoroughly analysed

Importance Of Statistics

Statistics is the subject that teaches how to deal with data, so statistical knowledge helps to use proper methods for collection of data, properly represent the data, use appropriate formula and methods to analyse correctly and effectively get the results and interpret the data. Applications of Statistics is important in every sphere of field – Business and economics, Medical, Sports, Weather forecast, Stock Market, Quality Testing, Government decisions and policies, Banks, Different educational and research organisations, etc.

Business and Economics

- In Business, the decision maker takes suitable policies and strategies based on information on production, sale, profit, purchase, finance, etc.
- By using the techniques of time series analysis, the businessman can predict the effect of a large number of variables with a fair degree of accuracy.
- By using 'Bayesian Decision Theory', the businessmen can select the optimal decisions to directly evaluate the payoff for each alternative course of action.
- In Economics, Statistics is used to analyse demand, cost, price, quantity, different laws of demand like elasticity of demand and consumer's maximum satisfaction which is determined on the basis of data pertaining to income and expenditure.

Medical

- **Statistics have extensive application in clinical research and medical field**. Clinical research involves investigating proposed medical treatments, assessing the relative benefits of competing therapies, and establishing optimal treatment combinations.

Weather Forecast

- Statistical methods, like Regression techniques and Time series analysis, are used in weather forecasting.

Stock Market

- Statistical methods, like Correlation and Regression techniques, Time series analysis are used in forecasting stock prices. Return and Risk Analysis is used in calculation of Market and Personal Portfolios and Mutual Funds.

Bank

- In banking industry, credit policies are decided based on statistical analysis of profitability, demand deposits, time deposits, credit ratio, number of customers and many other ratios. The credit policies are based on the application of probability theory.

Sports

- Players use statistics to identify or rectify their mistakes. A proper understanding of the statistics determines the success of a team or a single athlete.

Function Of Statistics

- Statistics present the facts in definite form.
- Statistics simplify complex data.
- It provides a techniques of comparison.
- Statistics study the relationship between two or more variables.
- It helps in formulating policies.
- It helps in forecasting outcomes.

Limitations Or Demerits Of Statistics

- **Statistics do not deal with Individuals:** Statistical methods can't be applied for individual values of the observations as for individual observation, there is no point of comparing anything or analysing anything. Statistics is the study of mass data or a group of observations and deals with aggregates of facts.

- **Statistics does not study Qualitative Data**: Statistical methods can't be applied for qualitative or non-numerical data. Statistics is the study of only of those facts which are capable of being stated in number or quantity.
- **Statistics give Result only on an Average**: Statistical methods are not exact. Generally, when we have large number of observations, it becomes difficult to handle it. A part of the data (sample) is collected for study and draw conclusion from, as a representative for the whole. As a result, the result obtained are not exactly same, had we analysed the whole data. The results are true only on an average in the long run.
- **The results can be biased**: The data collection may sometime be biased which will make the whole investigation useless. Generally, this situation arises when data is handled by inexperienced or dishonest person.

Definitions

Population

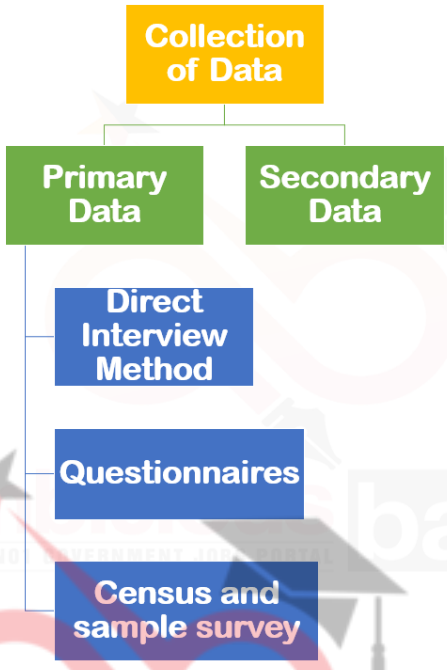
It is the entire collection of observations (person, animal, plant or things which is actually studied by a researcher) from which we may collect data. It is the entire group we are interested in and from which we need to draw conclusions.

Example: If we are studying the weight of adult men in India, the population is the set of weights of all men in India.

Data can be classified into two types, based on their characteristics.

- **Variates:** A characteristic that varies from one individual to another and can be expressed in numerical terms is called variate. Example: Prices of a given commodity, wages of workers, heights and weights of students in a class, marks of students, etc.
- **Attributes:** A characteristic that varies from one individual to another but can't be expressed in numerical terms is called an attribute. Example: Colour of the ball (black, blue, green, etc.), religion of human, etc.

Collection Of Data



Researchers or investigators need to collect data from respondents. There are two types of data.

Primary Data

Primary data is the data which is collected directly or first time by the investigator or researcher from the respondents. Primary data is collected by using the following methods:

- **Direct Interview Method:** A face to face contact is made with the informants or respondents (persons from whom the information is to be obtained) under this method of collecting data. The interviewer asks them questions pertaining to the survey and collects the desired information.
- Q**uestionnaires**: Questionnaires are survey instruments containing short closed-ended questions (multiple choice) or broad open-ended questions. Questionnaires are used to collect data from a large group of subjects on a specific topic. Currently, many questionnaires are developed and administered online.

Census and sample survey

- In a census, data about all individual units (e.g., people or households) are collected in the population. In a survey, data are only collected for a sub-part of the population; this part is called a sample.
- These data are then used to estimate the characteristics of the whole population. In this case, it has to be ensured that the sample is representative of the population in question. For example, the proportion of people below the age of 18 or the proportion of women and men in the selected sample of households has to reflect the reality in the total population.

Secondary Data

- Secondary data are the Second hand information. The data which have already been collected and processed by some agency or persons and is collected for the second time are termed as secondary data.
- According to M. M. Blair, "Secondary data are those already in existence and which have been collected for some other purpose." Secondary data may be collected from existing records, different published or unpublished sources, like WHO, UNESCO, LIC, etc., various research and educational organisations, banks and financial places, magazines, internet, etc.

Distinction between primary and secondary data

- The data collected for the first time is called Primary data and data collected through some published or unpublished sources is called Secondary data.
- The primary data in the hands of one person can become secondary for all others. For example, the population census report is primary for the Registrar General of India and the information from the report is secondary for others.
- Primary data are original as they are collected first time from the respondents directly or by preparing questionnaires. So they are more accurate than the secondary data. But the collection of primary data requires more money, time and energy than the secondary data. A proper choice between the two forms of information should be made in an enquiry.

Classification and Tabulation

So, we learned about the different methods of collecting primary and secondary data. The raw data, collected in real situations are arranged randomly, haphazardly and sometimes the data size is very large. Thus, the raw data do not give any clear picture and interpreting and drawing any conclusion becomes very difficult. To make the data understandable, comparable and to locate similarities, the next step is classification of data. The method of arranging data into homogeneous group or classes according to some common characteristics present in **the data is called Classification**.

Example: The process of sorting letters in a post office, the letters are classified according to the cities and further arranged according to the streets. Classification condenses the data by removing unimportant details. It enables us to accommodate large number of observations into few classes and study the relationship between several characteristics. Classified data is presented in a more organised way so it is easier to interpret and compare them, **which is known as Tabulation**.

There are four important bases of classifications:

- **Qualitative Base**: Here the data is classified according to some quality or attribute such as sex, religion, literacy, intelligence, etc.
- **Quantitative Base**: Here the data is classified according to some quantitative characteristic like height, weight, age, income, marks, etc.

- **Geographical Base**: Here the data is classified by geographical regions or location, like states, cities, countries, etc. like population in different states of India.
- **Chronological or Temporal Base**: Here the data is classified or arranged by their time of occurrence, such as years, months, weeks, days, etc. This classification is also called Time Series data.

Example: Sales of a company for different years.

Types of Classification

- If we classify observed data for a single characteristic, it is known as One-way Classification. Ex: Population can be classified by Religion – Hindu, Muslim, Christians, etc.
- If we consider two characteristics at a time to classify the observed data, it is known as a Two-way classification. Ex: Population can be classified according to Religion and sex.
- If we consider more than two characteristics at a time in order to classify the observed data, it is known as Multi-way Classification. Ex: Population can be classified by Religion, sex and literacy.

Frequency Distribution

Frequency

- If the value of a variable (discrete or continuous) e.g., height, weight, income, etc. occurs twice or more in a given series of observations, then the number of occurrences of the value is termed as the "frequency" of that value.
- The way of representing a data in a form of a table consisting of the values of the variable with the corresponding frequencies is called **"frequency distribution".**
- So, in other words, Frequency distribution is a table used to organise the data.
- The left column (called classes or groups) includes numerical intervals on a variable under study.
- The right column contains the list of frequencies, or number of occurrences of each class/group.
- Croxton and Cowden defined frequency distribution as a statistical table which shows the sets of all distinct values of the variable arranged in order of magnitude, either individually or in groups with their corresponding frequencies side by side Intervals are normally of equal size covering the sample observations range.

Class-limits or Class Intervals

- A class is formed within the two values, class-limits or class-intervals. The lower value is called lower class limit or lower-class interval and the upper value is called upper class limit or class interval.

Class-intervals	Frequency
0–4	5
5–9	7
10–14	12
15–19	8

Class Length or Class Width

- The difference between the class'upper and lower class limit is called the length or the width of class.

Class Length = Class Width = Upper Class Interval – Lower Class Interval

Mid-Value or Class Mark

- The mid-point of the class is called mid-value or class mark.

Class Mark = (Lower class-limit + Upper Class limit)/2

Types of Class Intervals

- Exclusive type,
- Inclusive type

Exclusive type Class intervals like

- 0–10, 10–20; 500–1000, 1000–1500 are called exclusive types.
- Here the upper limits of the classes are excluded from the respective classes and put in the next class while considering the frequency of the respective class.
- For example, the value 15 is excluded from the class 10–15 and put in the class 15–20.

Inclusive type Class intervals

- 60–69, 70–79, 80–89, etc. are inclusive type.
- Here both the lower and upper class limits are included in the class-intervals while considering the frequency of the respective class,
- e.g., 60 and 69 are both included in the class 60–69.

Class Intervals	Frequency
0-4	5
5-9	7
10-14	12
15-19	8

INCLUSIVE TYPE CLASS INTERVAL

Class Intervals	Frequency
0-10	5
10-20	7
20-30	12
30-40	8

EXCLUSIVE TYPE CLASS INTERVAL

Class Boundaries

Inclusive classes can be converted to exclusive classes and the new class intervals are called class boundaries.

Example : The classes 5–9, 10–14 can be converted to exclusive type of classes using the formula → New UCI = Old UCI + (10 – 9)/2 = 9 + 0.5 = 9.5. New LCI = Old LCI – (10 – 9)/2 = 5 – 0.5 = 4.5. So the class-boundaries are 4.5–9.5, 9.5–14.5, etc.

Open-end Class Interval

In open-end class interval either the lower limit of the first class or upper limit of the last class or both are missing.

Example:

Below 10

10–20

20–30

30–40

Above 40

Relative Frequency = frequency /Total frequency

Example: Relative frequency of the class interval = 20–30 in Example 2 is 12/32 = 0.375

Percentage Frequency

Percentage Frequency = (Class frequency/Total Frequency) × 100

Example: Percentage frequency of the class interval = 20–30 in Example 2 is (12/32) 100 = 37.5.

Frequency Density

Frequency density of a class interval = Class frequency/Width of Class

Continuous Frequency Distribution:

- Variable takes values which are expressed in class intervals within certain limits.

Problem: Marks obtained by 20 students in an exam for 50 marks are given below–convert the data into continuous frequency distribution form.

18, 23, 28, 29, 44, 28, 48, 33, 32, 43, 24, 29, 32, 39, 49, 42, 27, 33, 28, 29.

Marks	Frequency
15-20	1
20-25	2
25-30	7
30-35	4
35-40	1
40-45	3
45-50	2

Problem: Following data reveals information about the number of children per family for 25 families. Prepare frequency distribution of number of children

(say variable x, taking distinct values 0, 1, 2, 3, 4).

3 2 1 1 2

4 0 1 2 3

1 2 0 4 2

2 1 2 3 2

1 3 4 0 1

Solution:

No of children	Frequency
0	3
1	7
2	8
3	4
4	3
Total	25

CAIIB Paper 1 (ABM) Module A Unit 2: Sampling Methods

Sampling

Sampling is a process used in **statistical analysis** in which a predetermined number of observations are taken **from a larger population**. The methodology used to sample from a larger population depends on the type of analysis being performed.

Types of sampling

There are two methods of selecting from populations

- Non- random or judgement sampling
- Random or probability sampling

Random Sampling

A probability sampling method is any method of sampling that utilizes some form of **random selection**. In order to have a random selection method, you must set up some process or procedure that assures that the different units in your population have equal probabilities of being chosen.

Type of Random Sampling

There are four main type of random sampling

i)Simple Random Sampling (SRS)

ii)Stratified Sampling

iii)Cluster Sampling

iv)Systematic Sampling

- **Simple Random Sampling (SRS): Simple Random Sampling selects samples by methods that allow each possible** sample to have an equal probability of being picked and each item in the entire population to have an equal chance or being included in the sample.
- **Systematic Sampling:** In systematic sampling, elements are **selected from the population at a uniform level that is measured in time, order, or space**. If we wanted to interview every twentieth student on a college campus, we would choose a random starting point in the first twenty names in the student directory and then pick every twentieth name thereafter.
- **Stratified Sampling**: To use stratified sampling, we divide the population into relatively homogenous groups, called strata. **Then we use one of two approaches**. Either we select at random from each stratum a specified number of elements corresponding to the proportion of that stratum in the population as a whole or we draw an equal number of elements from each stratum and give weight to the results according to the stratum's proportion of total population.
- **Cluster Sampling**: In cluster sampling, we divide the population into groups or clusters and then select a random sample of these clusters. We assume that these individual clusters are representative of the population as a whole. If a market Research team is attempting to determine by sampling the average number of television sets per household in a large city, they could use a city map and divide the territory into blocks and then choose a certain number of blocks (clusters) for interviewing. Every household in each of these blocks would be interviewed. A well designed cluster sampling procedure can produce a more precise sample at considerably less cost than that of simple random sampling.

Sampling distribution

Sampling Distribution is the distribution of all possible values of a statistic from all possible samples of **a particular size drawn from the population**.

Each sample we draw from a population would have its own means or measure of central tendency and standard deviation. Thus, the statistics we compute for each sample, would vary & be different for each random sample taken.

Sample Distribution Table

Boy	Height
A	160
B	162
C	164
D	170
E	156

Mean = 162.40

SAMPLES, their DATA & Mean

Samples	ABC	ABD	ABE	BCD	BCE	ACD	ACE	ADE	BDE	CDE
DATA	160 162 164	160 162 170	160 162 156	162 164 170	162 164 156	160 164 170	160 164 156	160 170 156	162 170 156	164 170 156
Mean	162	164	159.33	165.33	160.66	164.66	160	182	162.66	163.33

CONCEPT of STANDARD ERROR

- Standard deviation of the distribution of the sample means is **called the standard error of the mean.**
- Similarly standard error of the proportion is the standard deviation of the distribution of the sample proportions.
- e.g. We take the average height of college girls in India across various samples. We would calculate mean height of each sample. Obviously there is some variability in observed mean. This variability in sampling statistics results from the sampling error due to chance.
- Thus the standard deviation of the sampling distribution of means measures the extent to which the means vary because of a chance error in the sampling process. Thus the standard deviation of distribution of a sample statistic is known as the Standard error of the statistic.
- Thus, a standard error indicates not only the size of the chance error but also the accuracy we are likely to get if we use the sample statistic to estimate a population statistic.

Sampling From Normal Population

Finite Populations:

μ= 162.40

$\bar{x}$ = 162.40

This is not coincidence. The mean of the sample means is the same as the population mean, whenever we use simple random sampling.

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

Example -Bank calculate that its individual saving account have a mean of Rs.2000 and SD of 600. bank takes a sample of 100 account. Calculate the Standard error?

What is the probability that the sample lie between 1900 & 2050.

$\sigma_{\bar{X}} = \sigma / \sqrt{n}$

= 600 /10

=60

Probability associated with a standard normal variable

$$Z = 1/ \sigma_X [X - \mu]$$

Standard Error Of The Mean For Infinite Populations:

Standard error of mean = σ/ √n

Example: Bank calculate that its individual saving account have a mean of Rs.5000 and SD of 600. bank takes a sample of 100 account. Calculate the Standard error?

What is the probability that the sample lie between 1900 & 2050.

STEP 1 : Standard deviation of error

$\sigma_{\bar{X}} = \sigma / \sqrt{n}$

= 600 /10

=60

STEP 2 : Calculate Z

$Z = \frac{\bar{X} - \mu}{\sigma X}$

For $\bar{X}$ = 1900

Z= (1900-2000)/60

= -1.67

For $\bar{X}$ = 2050

Z= (2050-2000)/60 = 0.83

STEP 3 : Probability table

-1.67 value = 0.4525

.83 = 0.2967

0.7492

If value have the same sign , we subtract value

If the value have the opposite sign, we add

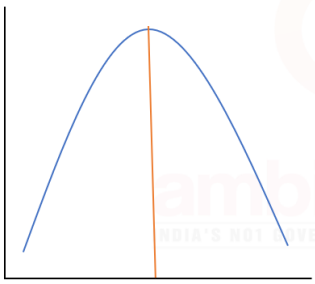
Central Limit Theorem

- The mean of the sampling distribution of the mean will equal the population mean regardless of the sample size, even if the population is not normal.
- As the sample size increases, the sampling distribution of the mean will approach normality, regardless of the shape of the population distribution.
- This relationship between the shape of the population distribution & the shape of sampling distribution of the mean is called the Central Limit Theorem.

- Actually a sample doesnot have to be very large for the sampling distribution of the mean to approach normal
- Statistician use the normal distribution as an approximation to the sampling distribution whenever the sample size is atleast 30, but the sampling distribution whenever the sample size is atleast 30.
- The significance of the CLT is that it permits us to use sample statistics to make interference about population parameters without knowing anything about the shape of the frequency distribution of that population.

Example:

Bank distribution has a mean of Rs.19000 & standard deviation of Rs.2000. If we draw a random sample of 30 tellers, What is the probability that their earning will average more than Rs.19750 annually?



Mean = 19000
Standard Deviation = 2000
Size = 30

STEP 1 : Calculate Standard error

$$\sigma\bar{X} = \sigma / \sqrt{n}$$

= 2000/ √30

= 2000/ 5.477

= 365.16

STEP 2 : Z value & Standard Normal Probability Distribution

$$Z = 1/ \sigma X [X - \mu]$$

X = 19750

= 19750-19000/365.16

= 750/ 365.16

=2.05

Finite Population Multiplier

Standard Error Of The Mean For Finite Populations

$$\sigma\bar{X} = \sigma / \sqrt{n} \sqrt{(N-n / N-1)}$$

N = size of population

N = size of the sample

Example:

We are interested in a population of 20 textile companies of the same size, all of which are experiencing excessive labour turnover. Standard deviation of the distribution of annual turnover is 75 employees. If we sample 5 of these textile companies, without replacement then compute the standard error of mean?

$\sigma X = \sigma / \sqrt{n} \left[\sqrt{(N-n / N-1)}\right]$

$= 75/\sqrt{5}\ [\sqrt{(20-5/20-1)}]$

$= 33.54 * 0.888$

$= 29.8$

Numerical on Sampling

Q1. A sack contains 3 pink balls and 7 green balls. What is probability to draw one pink ball and two green balls in one draw?

(a) $\frac{23}{40}$

(b) $\frac{21}{40}$

(c) $\frac{27}{40}$

(d) $\frac{9}{20}$

(e) $\frac{21}{38}$

Ans(b)

Out of (3+ 7) = 10 balls, three (one pink & two green) balls are expected to be drawn

So, required probability $= \frac{{}^{3}c_1 \times {}^{7}c_2}{{}^{10}c_3}$

$= \frac{\frac{3}{1} \times \frac{7}{2 \times 1}}{\frac{10}{3 \times 2 \times 1}}$

$= \frac{3 \times 21}{120}$

$= \frac{21}{40}$

Q2.A sack contains 4 black balls 5 red balls. What is probability to draw 1 black ball and 2 red balls in one draw?

(a) 11/19

(b) 10/21

(c)12/22

(d) 19/11

Ans: B

Solution :

Out of 9, 3 (1 black & 2 red) are expected to be drawn)

Hence sample space

n(S) = 9c3

= 9!/(6!×3!)

= 362880/4320

= 84

Now out of 4 black ball 1 is expected to be drawn hence

n(B) = 4c1

= 4

Same way out of 5 red balls 2 are expected be drawn hence

n(R) = 5c2

= 5!/(3!×2!)

= 120/12

= 10

Then P(B U R) = n(B)×n(R)/n(S)

i.e 4×10/84 = 10/21

CAIIB Paper 1 (ABM) Module A Unit 3: Measures of Central Tendency & Dispersion, Skewness, Kurtosis

Introduction To Measures Of Central Tendency

- Statistical data is first collected (primary or secondary) and then classified into different groups according to common characteristics and presented in a form of a table.
- It is easy for us to study the different characteristics of data from a tabular form.
- Further, graphs and diagrams can also be drawn to convey a better impression to the mind about the data.

- Classified and Tabulated data need to be analysed using different statistical methods and tools and then draw conclusions from it.
- Central Tendency and Dispersion are the most common and widely used statistical tool which handles large quantity of data and reduces the data to a single value used for doing comparative studies and draw conclusion with accuracy and clarity.
- According to the statistician, Professor Bowley "Measures of Central Tendency (averages) are statistical constants which enable us to comprehend in single effort the significant of the whole".

The main objectives of Measure of Central Tendency are:

- ✓ To condense data in a single value.
- ✓ To facilitate comparisons between data.

- In other words, the tendency of data to cluster around a central or mid value is called central tendency of data, central tendency is measured by averages.
- There are different types of averages, each has its own advantages and disadvantages.

Requisites of a Good Measure of Central Tendency

- ✓ It should be rigidly defined.
- ✓ It should be simple to understand and easy to calculate.
- ✓ It should be based on all the observations of the data.
- ✓ It should be capable of further mathematical treatment.
- ✓ It should be least affected by the fluctuations of the sampling.
- ✓ It should not be unduly affected by the extreme values.
- ✓ It should be easy to interpret.

Three types of averages are Mean, Median and Mode.

Mean

- Mean or average is the most commonly used single descriptive measure of Central Tendency.
- Mean is simple to compute, easy to understand and interpret.

Mean is of three types:

- ✓ Arithmetic Mean,
- ✓ Geometric Mean
- ✓ Harmonic Mean.

Arithmetic Mean

- The arithmetic mean is the simplest and most widely used measure of a mean, or average.

- It simply involves taking the sum of a group of numbers, then dividing that sum by the count of the numbers used in the series.

Arithmetic Mean of Ungrouped or Raw Data

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$\bar{X} = x1 + x2 + x3 + x4 + ...xn$

$\bar{X} = \sum X/n$

n observations of x.

Example 1: Consider the marks scored by 10 students in Mathematics in a certain examination 35, 30, 18, 15, 40, 30, 52, x, 47, 10. If the arithmetic mean is 30, find the value of x.

$\bar{X}$ = 35 +30 +18 + 15 + 40 +30 +52 + x + 47 + 10/ 10

30 = 277 + x/ 10

300 = 277 + x

X= 23

Arithmetic Mean of Grouped data

- If a variate X take values x1, x2, …, xn with corresponding frequencies f 1, f 2, …, f n respectively, then the arithmetic mean of these values is

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum f_i}, x_i$$

$\bar{X} = \sum fx / n$

Xi = class marks (mid point) of the class interval for grouped data

Example 2: Find the Arithmetic mean for following:

X	1	2	3	4	5	6	7	8
f	5	6	5	10	8	4	3	2

X	F	FX
1	5	5
2	6	12
3	5	15
4	10	40
5	8	40
6	4	24
7	3	21
8	2	16
Total	N= 43	Fx = 173

$\bar{X} = \sum fx / n$
$= 173/ 43$
$= 4.02$

Combined Arithmetic Mean

If $\bar{X}1$ and $\bar{X}2$are the arithmetic mean of two samples of size n1 and n2 respectively then, the Combined arithmetic mean

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

Example: The average marks of a group of 100 students in Mathematics are 60 and for other group of 50 students, the average marks are 90. Find the average marks combined group of 150 students.

$$X = 100*60 + 50*90/ 100 +50$$
$$= 6000 +4500 / 150$$
$$= 70$$

Example: In private health club, there are 200 members, 100 men, 80 women and 20 children. The average weight of men, women and children are 60 kgs, 50 kgs and 35 kgs respectively. Find the average weight of the combined group.

n1 = 100, n2 = 80, n3 = 20 x1 = 60, x2 = 50, x3 = 35

Combined mean =

$\bar{X}$ = n1 x1 + n2 x2 + n3 x3/ n1 + n2 + n3

= 100*60 + 80*50 + 20*35/ 200

= 6000+ 4000 + 700 / 200

= 10700/2

= 53.5

Merits of Arithmetic Mean

- It is rigidly defined
- It is easy to calculate and simple to follow
- It is based on all the observations
- It is determined for almost every kind of data
- It is finite and indefinite
- It is readily put to algebraic treatment
- It is least affected by fluctuations of sampling.

Demerits of Arithmetic Mean

- It is highly affected by extreme values.
- It cannot average the ratios and percentages properly.
- It is not an appropriate average for highly skewed distribution.
- It cannot be computed accurately if any item is missing.
- The mean sometimes does not coincide with any of the observed value.
- Mean cannot be calculated when open-end class intervals are present in the data

Geometric Mean

The Geometric Mean (GM) is the average value or mean which measures the central tendency of the set of numbers by taking the root of the product of their values. Geometric mean takes into account the compounding effect of the data that occurs from period to period. Geometric mean is always less than Arithmetic Mean and is calculated only for positive values.

Applications

- It is used in stock indexes.
- It is used to calculate the annual return on the portfolio.
- It is used in finance to find the average growth rates which are also referred to the compounded annual growth rate.
- It is also used in studies like cell division and bacterial growth, etc.

Geometric Mean of Ungrouped or Raw Data

$$G.M. = \sqrt[n]{x_1 x_2 \ldots x_n} = (x_1 x_2 \ldots x_n)^{\frac{1}{n}} \text{ where } x_1, x_2 \ldots, x_n \text{ are n observations of } x.$$

Example: Find the G.M. of the values 10, 24, 15, and 32.

Given 10, 24, 15, 32

We know that G.M. = 4√ 10*24*15*32

= (10*24*15*32)^1/4

= 115200 ^1/4

= 18.423

Geometric Mean of Grouped or Raw Data

$$G.M. = \sqrt[n]{x_1^{f_1} x_2^{f_2} \ldots . x_n^{f_n}} = \left(x_1^{f_1} x_2^{f_2} \ldots . x_n^{f_n} \right)^{\frac{1}{n}}$$

Example: Find the G.M. for the following data

X	1	2	3	4
F	5	6	5	10

X	F	X^F
1	5	1^5 = 1
2	6	2^6 = 64
3	5	3^5 = 243
4	10	4^10 = 1048576
Total	N = 26	

=26√1*64*243*1048576
= 26√ 1630745394
= 24705

Merits of Geometric Mean

- **It is useful in the construction of index numbers.**
- **It is not much affected by the fluctuations of sampling.**
- **It is based on all the observations.**

Demerits of Geometric Mean

- **It cannot be easily understood.**
- **It is relatively difficult to compute as it requires some special knowledge of logarithms.**
- **It cannot be calculated when any item or value is zero or negative.**

Harmonic Mean

- Harmonic Mean is defined as the reciprocal of the arithmetic mean of reciprocals of the observations. Arithmetic mean is appropriate measure of central tendency when the values have the same units whereas the Harmonic mean is appropriate measure of central tendency when the values are the ratios of two variables and have different measures. So, generally Harmonic mean is used to calculate the average of ratios or rates.

Applications

- It is used in finance to find average of different rates.
- It can be used to calculate quantities such as speed. This is because speed is expressed as a ratio of two measuring units such as km/hr.

Harmonic Mean of Ungrouped or Raw data:

$$\text{H.M.} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \text{ where } x_1, x_2, \ldots\ldots x \text{ are n observations of x.}$$

H.M = n / (1/x1 + 1/x2 + 1/x3+ ..... 1/y)

Example: Find the HM of the values 10, 24, 15, and 32

= 4/ (1/10+1/24+1/15+1/32)

= 4 / 0.1 + 0.042 + 0.067 + 0.031

= 4/.24

= 16.667

Harmonic Mean of Ungrouped or Raw data:

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

N /(F1/X1 + F2/X2 + F3/X3....)

Example: Find the H.M. for the following data

X	1	2	3	4
F	5	6	5	10

X	F
1	5
2	6
3	5
4	10
Total	N = 26

= 26/ (5/1 + 6/2 + 5/3 + 10/4)
= 26/ (5 + 3+ 1.667 + 2.5)
= 26/ 12.167
= 2.137

Comparison between Arithmetic, Geometric and Harmonic Mean

- The arithmetic mean is appropriate if the values have the same units, whereas the geometric mean is appropriate if the values have different units and harmonic mean is appropriate if the data values are ratios of two variables with different measures, called rates.
- **Arithmetic Mean > Harmonic Mean > Geometric Mean**

- **A.M. × H.M. = (G.M.)^2**

Example: Find the Harmonic mean of two numbers a and b, if their Arithmetic mean is 16 and Geometric mean is 8.

- A.M. = 16 and G.M. = 8
- A.M. × H.M. = G.M^2
- 16 × H.M. = 8^2
- 16 × H.M. = 64
- H.M.= 64/16 = 4

Median And Quartiles

- The median is the middle value of a distribution, i.e., median of a distribution is the value of the variable which divides it into two equal parts.
- It is the value of the variable such that the number of observations above it is equal to the number of observations below it.
- Observations are arranged either in ascending order or descending order of their magnitude.
- Median is a position average whereas the arithmetic mean is a calculated average.

Median of Ungrouped or Raw data

- **The formula to calculate the median of the data is different for odd and even number of observations.**

Median of odd Number of Observations

If the total number of given observations is odd, then the formula to calculate the median for a number of n observations is:

Median = n +1/ 2 th observation

Median of even Number of Observations

If the total number of given observations is even, then the median formula to calculate the median for n number of observations is:

Median = **Median= (n/2)th observation + (n/2+1)th observation / 2**

Example: Find Median of 34, 32, 48, 38, 24, 30, 27, 21, 35.

Arranging the data in ascending order,

21, 24, 27, 30, 32, 34, 35, 38, 48.

n = 9;

Median= (n+1/2) th position

= (9+1/2) the position

= 32

Median of Grouped data:

If variable X takes values X1, X2, X3, X4…..X5 and corresponding frequencies f1, f2, f3, f4,….. Fn respectively, then the median value is given by

$$Median = l_1 + \frac{(l_2 - l_1)\left(\frac{N}{2} - cf\right)}{f}$$

Median class is the class in which the corresponding value of less than cumulative frequency just exceeds the value of N/2.

- l1 = lower limit of the median class,
- l2 = upper limit of the median class
- f = frequency of the median class,
- cf = cumulative frequency of the class preceding the median class,
- N = total frequency.

Example: Find Median for the following data.

Class Interval	20-30	30-40	40-50	50-60	60-70
Frequency	8	26	30	20	16

Class Interval	Frequency	CF
20-30	8	8
30-40	26	34
40-50	30	64
50-60	20	84
60-70	16	100
Total	100	

N/2 = 100/2 = 50
= l1 +[(l2-l1) (N/2 – CF) / f]
= 40 +[10* (50-34) / 30]
= 40 + [10 *16/30]
= 40 + 160/30
= 40 + 5.33
= 45.33

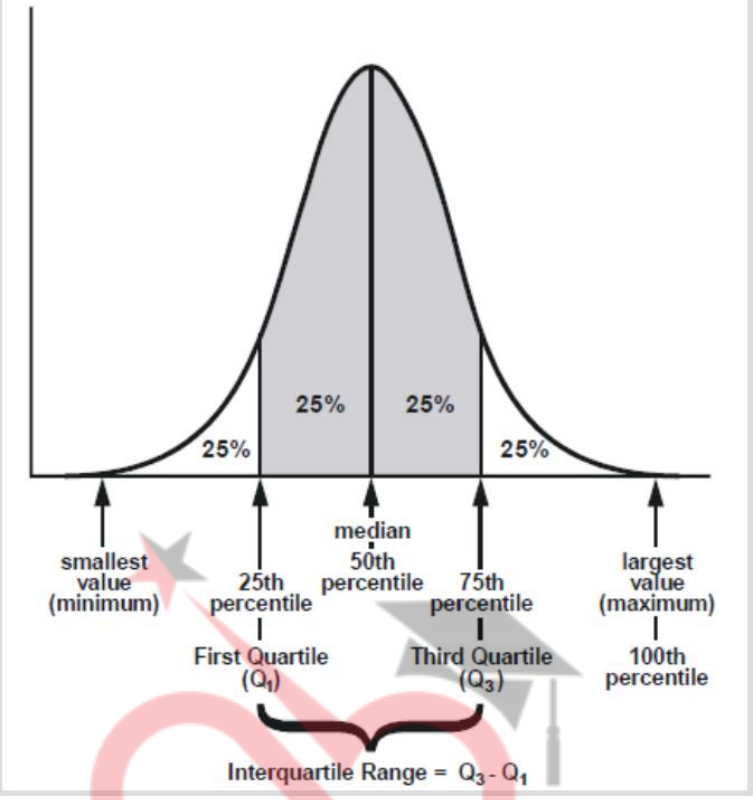
Quartiles

- A quartile represents the division of data into four equal parts.
- First, second intervals are based on the data values and third their relationship to the total set of observations.
- By dividing the distribution into four groups, the quartile calculates the range of values above and below the mean.

A quartile divides data into three points

- ✓ the lower quartile Q1,
- ✓ the median Q2, and
- ✓ the upper quartile Q3, to create four dataset groupings.

The interquartile range is a measure of variability around the median, which is calculated using the quartiles are denoted by Q1, Q2 and Q3



Calculate Q1,Q2 & Q3

Q1 = l1 + (q1-CF)/f (l2-l1) where q1 = N/4

Q2 = l1 + (q2- CF)/f (l2-l1) where q2 = 2N/4

Q3 = l1 + (q3- CF)/f (l2-l1) where q3 = 3N/4

Example: Find the quartiles for the following data

Class Interval	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55
Frequency	12	28	36	50	25	18	16	10	5

Class Interval	Frequency	CF
10-15	12	12
15-20	28	40
20-25	36	76
25-30	50	126
30-35	25	151
35-40	18	169
40-45	16	185
45-50	10	195
50-55	5	200

N/4= 200/4 = 50
Q1 = l1 + (q1-CF)/f (l2-l1)
= 20 + 50-40/36 (5)
= 20 + 10/36 *5
= 21.39

Q2 = l1 + (q2- CF)/f (l2-l1
= 25 + (100-76)/50 (30-25)
= 27.4

AIIB BABA

Q3 = l1 + (q3- CF)/f (l2-l1)

= 30 (150-126) / 25 (35-30)

= 34.8

MODE

- The mode of a set of numbers is that number, which occurs more number of times than any other number in the set.
- It is the most frequently occurring value.
- If two or more values occur with equal or nearly equal number of times, then the distribution is said to have two or more modes.
- In case, there are three or more modes and the distribution or data set is said to be multimodal.

Mode of Ungrouped or Raw data

Example 22: Find Mode for the data: 23, 25, 20, 23, 26, 21, 27, 28, 30, 27, 23.

Value 23 occurs maximum number of times,

so Mode = 23.

Mode of Grouped data

If a variate X take values x1, x2, x3, x4 with corresponding frequencies f1, f2, f3, f4…. respectively, then the mode is

$$Mode = l_1 + \frac{(l_2 - l_1)(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

Where,

l1 = lower limit of the modal class

l2= per limit of the modal class

f1 = frequency of the modal class

f0 = frequency of the class preceding the modal class

f2 = frequency of the class succeeding the modal class

Example: Find Mode for data

Class Interval	20-30	30-40	40-50	50-60	60-70
Frequency	8	26	30	20	16

Class Interval	Frequency
20-30	8
30-40	26
40-50	30
50-60	20
60-70	16
Total	

40 + [10 * 4]/ 60- 26-20
= 40 + 40/14
= 40 + 2.857
= 42.857

$$Mode = l_1 + \frac{(l_2 - l_1)(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

JAIIB CAIIB BABA

Merits of Mode

- It is easy to calculate and understand.
- It is not affected much by sampling fluctuations.
- It is not necessary to know all items. Only the point of maximum concentration is required.

Demerits of Mode

- It is ill defined as it is not based on all observations.
- It is not capable of further algebraic treatment.
- It is not a good representative.

Relationship among Mean, Media and Mode

- **Mode = 3 Median – 2 Mean**

Introduction to Measures Of Dispersion

- A single value that attempts to describe a set of data by identifying the central position within the set of data is **called measure of central tendency**.
- Measure of Dispersion is another property of a data which establishes the degree of variability or the spread out or scatter of the individual items and their deviation from (or the difference with) the averages or central tendencies.
- The process by which data are scattered, stretched, or spread out among a variety of categories is referred to as dispersion.
- Finding the size of the distribution values that are expected from the collection of data for the particular variable is a part of this process.
- The dispersion of data is a concept in statistics that lets one understand a dataset more simply by classifying individual pieces of data according their own unique dispersion criteria, such as the variance, the standard deviation, and the range.
- A collection of measurements known as dispersion can be used to determine the quality of the data in an objective and quantitative manner.

Various measures of dispersion are given below:

Four Absolute Measures of Dispersion

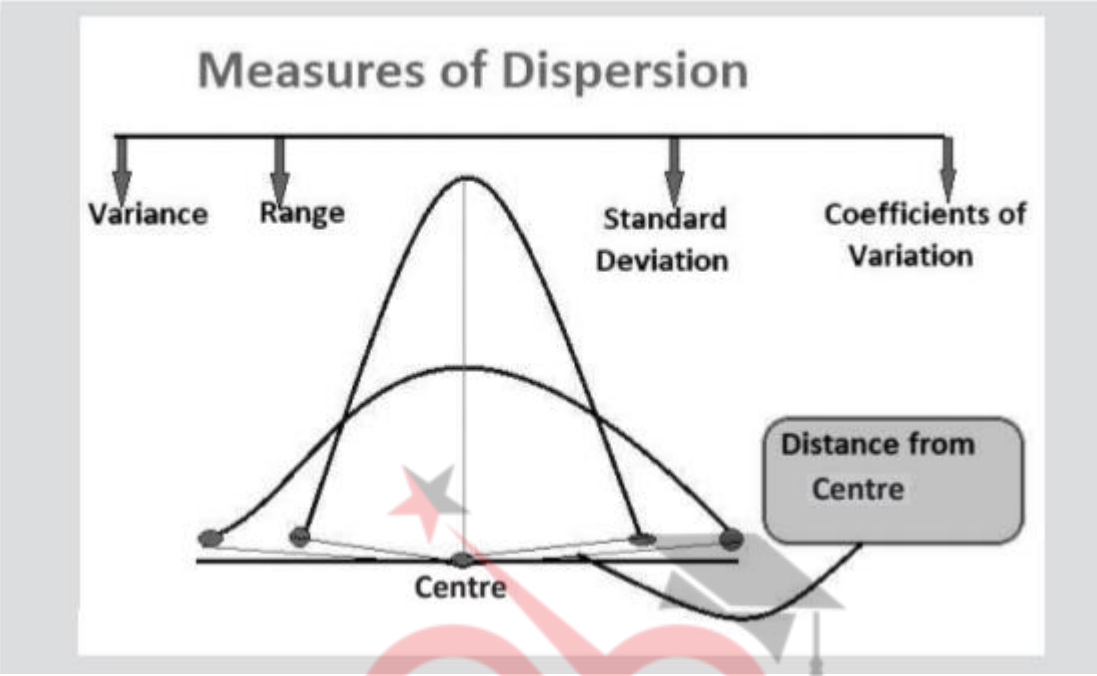
- Range
- Quartile Deviation
- Mean Deviation
- Standard Deviation

Four Relative Measures of Dispersion

- Coefficient of Range
- Coefficient of Quartile Deviation
- Coefficient of Mean Deviation
- Coefficient of Variation

Characteristics of a Good Measure of Dispersion

- It should be rigidly defined.
- It should be based on all observations.
- It should be easy to calculate and understand.
- It should be capable of further algebraic treatment.
- It should not be affected much by sampling fluctuations.



Range and Coefficient Of Range

Range

It is the simplest absolute measure of dispersion.

Range (R) = Maximum – Minimum

Coefficient of Range = (Max – Min)/(Max + Min)

Example 1 Find the range and coefficient of range of the following items: 18, 15, 20, 17, 22, 16.

- Range = Max – Min = 22 – 15 = 7.
- Coefficient of Range = (Max – Min)/(Max + Min) = 7/37 = 0.19

Note: Range and Coefficient of Range are used to measure the spread in Quality Control, Fluctuations in the Share Prices, in Weather Forecasts**:**

Merits of Range

- It is easy to understand.
- It is easy to calculate.

Demerits of Range

- It is not based on all observations.

- It does not have sampling stability. A single observation may change the value of range.
- As the amount of data increases, range becomes less satisfactory

Quartile Deviation And Coefficient Of Quartile Deviation

It is the mid-point of the range between two quartiles. Quartile Deviation is defined as QD = (Q3 – Q1)/2

Where Q1 = 1st quartile and Q 3 = 3rd quartile.

Co-efficient of QD = (Q3 – Q1)/ (Q3 + Q1)

Merits of Quartile Deviation

- It is easy to calculate and understand.
- It is not affected by extreme values.

Demerits of Quartile Deviation

- It is not based on all observations.
- It is not capable of further algebraic treatment.
- It is affected by sampling fluctuations.

Mean Deviation and Coefficient of Mean Deviation

- Mean deviation of a set of observations of a series is the arithmetic mean of all the deviations.
- It is the deviations from mean when calculated considering their absolute values and are averaged.

Mean Deviation (MD) ungrouped data

$$\text{MD (Mean)} = \left(\sum_{i=1}^{n} |x_i - \bar{x}|\right)/n$$

$$\text{Coefficient of Mean Deviation (Mean)} = \frac{\text{MD (Mean)}}{\text{Mean}}$$

MD = [(X1- $\bar{X}$) + (X2- $\bar{X}$) + (X3- $\bar{X}$) + ____ + (Xn - $\bar{X}$)] / n

Example: Find Mean Deviation and Coefficient of Mean Deviation

Class Interval	20-30	30-40	40-50	50-60	60-70
Frequency	8	26	30	20	16

Class Interval	Frequency	X	fx	X – X̄ (46)	F (X-X̄)
20-30	8	25	200	21	168
30-40	26	35	910	11	286
40-50	30	45	1350	1	30
50-60	20	55	1100	9	180
60-70	16	65	1040	19	304
Total	100		4600		968

Mean = 4600/100 = 46

- MD (Mean) = 968/100 = 9.68
- Coefficient of Mean Deviation (Mean) = MD (Mean)/ Mean

= 9.68/46 = 0.2104

Merits of Mean Deviation

- It is based on all observations.
- It is easy to understand and also easy to calculate.
- It is not affected by extreme values.

Demerits of Mean Deviation

- Mean deviation ignores algebraic signs; hence it is not capable of further algebraic treatment.
- It is not very accurate measure of dispersion.

Note: Mean deviation and its coefficient are used in studying economic problems such as distribution of income and wealth in a society.

Standard Deviation And Coefficient Of Variation

- Standard deviation is the most important and commonly used measure of dispersion.
- It measures the spread or variability of a distribution.
- A small standard deviation means a high degree of consistency in the observations as well as homogeneity of the series.

Standard Deviation ungrouped Data

$$SD = \sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2}$$ where $\bar{x}$ is the mean of these observations

Standard Deviation (SD) grouped data

$$\text{SD} = \sigma = \sqrt{\frac{\sum f\left(x-\bar{x}\right)^2}{\sum f}} = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2} \quad \text{where N} = \sum f$$

$$\textbf{Coefficient of Variation} = \textbf{CV} = \frac{\sigma}{\bar{x}} \times 100\%$$

Example: Find Standard Deviation and Coefficient of Variation for the following data: 2, 3, 7, 8, 10.

X	X^2
2	4
3	9
7	49
8	64
10	100
Mean = 30/5=6	226

= √ 226/5 – 6^2
= √ 45.2- 36
= 3.03
Coefficient
= SD/ Mean *100
= 3.03/6 * 100
= 50.5%

$$\text{SD} = \sigma = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \left(\bar{x}\right)^2}$$

Example: Find Standard Deviation?

Class Interval	25-30	30-35	35-40	40-45	45-50	50-55
Frequency	30	23	20	14	10	3

Class Interval	Frequency	X	fx	Fx^2
25-30	30	27.5	825	22687.5
30-35	23	32.5	747.5	24293.75
35-40	20	37.5	750	28125
40-45	14	42.5	595	25287.5
45-50	10	47.5	475	22562.5
50-55	3	52.5	157.5	8268.75
Total	N= 100		3550	131225

$$\text{SD} = \sigma = \sqrt{\frac{\sum f\left(x-\bar{x}\right)^2}{\sum f}} = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$$

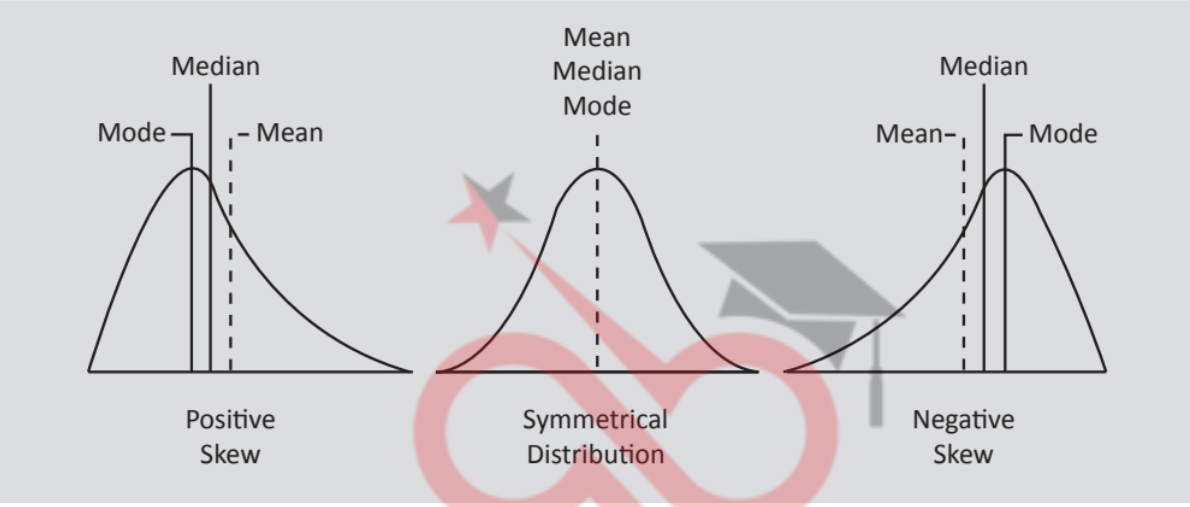
√ 131225/100 – (3550/100)^2
√ 1312.25 – 1260.25
= 7.21

Merits of Standard Deviation

- It is rigidly defined and has a definite value.
- It is based on all observations.
- It is not affected much by sampling fluctuations.

Demerits of Standard Deviation

- It is not easy to calculate.
- It is not easy to understand.
- It gives more weight to extreme items.



Skewness And Kurtosis

- Skewness is the degree of distortion from the symmetrical bell curve or the normal distribution.
- It measures the lack of symmetry in data distribution.
- **There are two types of skewness– positive and negative**.
- If bulk of observations is in the left side of mean and the positive side is longer, it is called positive skewness of the distribution.
- mean and median > mode.
- If bulk of observations is in the right side of mean and the negative side is longer, it is called negative skewness of the distribution.
- mean and median < mode.

Karl Pearson's measure of skewness is

β1=skewness: = μ3^2/ μ2^3
Where
μ3 = third central moment = ∑f(x-x̄)^ 3/ n
μ2 = second central moment = ∑f(x-x̄)^ 2/ n

- **The direction of skewness is measured by sign of β1, where the sign of β1 is the sign of μ3 .**
- **β1= 0 (symmetrical distribution),**
- **β1> 0 (positive skew),**
- **β1< 0 (negative skew).**

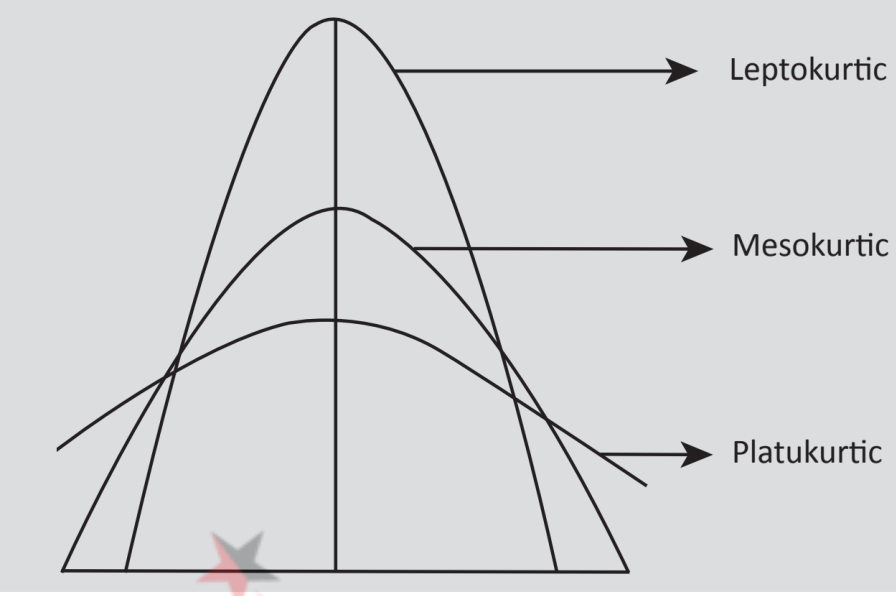
μ1 = First central moment = ∑f(x−x̄)^ 1/ n

μ2 = Second central moment = ∑f(x−x̄)^ 2/ n

μ3 = third central moment = ∑f(x−x̄)^ 3/ n

μ4 = Forth central moment = ∑f(x−x̄)^ 4/ n

- Kurtosisis all about the tails of the distribution – peakedness or flatness.
- It is used to describe the extreme values in one versus the other tail.
- It is actually the measure of outliers present in the distribution.
- The distributions whose peaks are same as of Normal distribution's peak**, are called Mesokurtic.**
- The distributions whose peaks are higher and sharper than mesokurtic, which meanstails are fatter, are called Leptokurtic distributions.
- The distributions whose peaks are lower and shorter than mesokurtic, which means tails are thinner**, are called Platykurtic distributions.**
- **Measure of Kurtosis = β2 = μ4/μ2^2**
- **μ4 = fourth central moment = ∑f(x−x̄) 4/ n**
- **μ2 = second central moment = ∑f(x−x̄) 2/ n**
- **β2 = 0 (Mesokurtic distribution),**
- **β2 > 0 (Leptokurtic distribution),**
- **β2 < 0 (Platykurtic distribution).**



CAIIB Paper 1 (ABM) Module A Unit 4: Correlation and Regression

Introduction

Correlation Analysis

- **Correlation analysis is applied in quantifying the association between two continuous variables**, for example, an dependent and independent variable or among two independent variables.

Regression Analysis

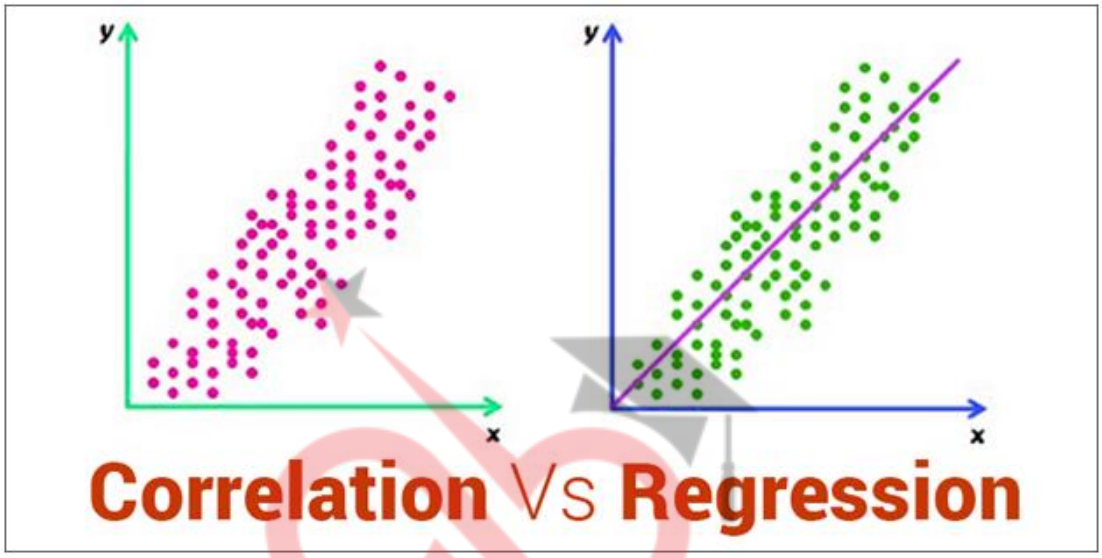
- **Regression analysis refers to assessing the relationship between the outcome variable and one or more variables**. The outcome variable is known as the dependent or response variable and the risk elements, and cofounders are known as predictors or independent variables.
- **The dependent variable is shown by "y" and independent variables are shown by "x" in regression analysis.**

Linear Regression

- Linear regression is a **linear approach to modelling the relationship between the scalar components and one or more independent variables**. If the regression has one independent variable, then it is known as a simple linear regression. If it has more than one independent variables, then it is known as multiple linear regression.

- Linear regression only focuses on the conditional probability distribution of the given values rather than the joint probability distribution. In general, all the real world regressions models involve multiple predictors. So, the term linear regression often describes multivariate linear regression.

Correlation and Regression Differences



There are some differences between Correlation and regression.

- Correlation shows the quantity of the degree to which two variables are associated. It does not fix a line through the data points. You compute a correlation that shows how much one variable changes when the other remains constant. When r is 0.0, the relationship does not exist. When r is positive, one variable goes high as the other goes up. When r is negative, one variable goes high as the other goes down.
- Linear regression finds the best line that predicts y from x, but Correlation does not fit a line.
- Correlation is used when you measure both variables, while linear regression is mostly applied when x is a variable that is manipulated.

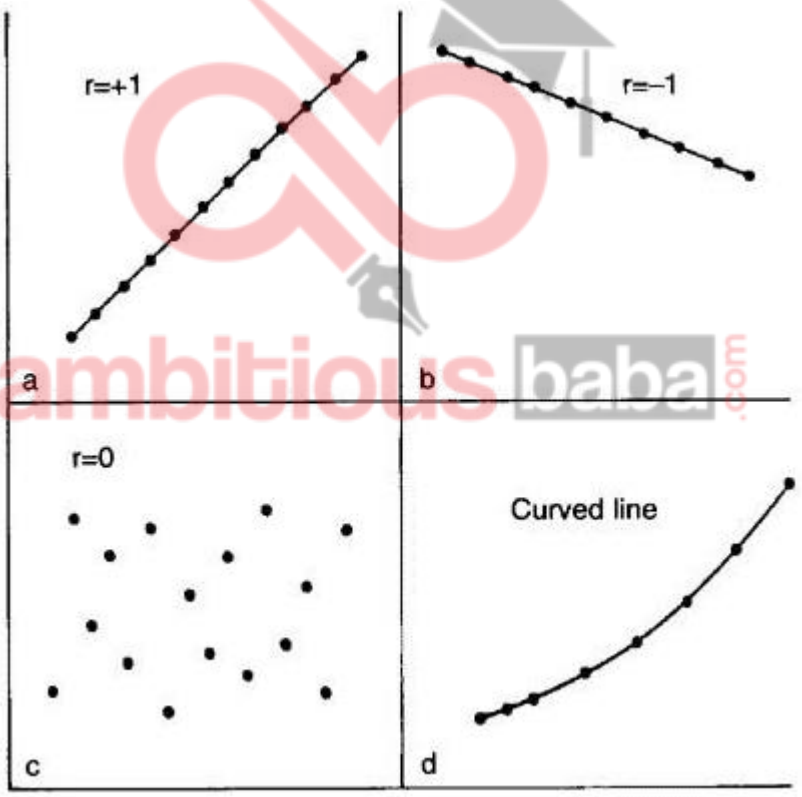
Comparison Between Correlation and Regression

Basis	Correlation	Regression
Meaning	A statistical measure that defines co-relationship or association of two variables.	Describes how an independent variable is associated with the dependent variable.

Dependent and Independent variables	No difference	Both variables are different.
Usage	To describe a linear relationship between two variables.	To fit the best line and estimate one variable based on another variable.
Objective	To find a value expressing the relationship between variables.	To estimate values of a random variable based on the values of a fixed variable.

Correlation and Regression Statistics

The degree of association is measured by "r" after its originator and a measure of linear association. Other complicated measures are used if a curved line is needed to represent the relationship.



The above graph represents the correlation.

The coefficient of correlation is measured on a scale that varies **from +1 to -1 through 0. The complete correlation among two variables is represented by either +1 or -1.** The correlation is positive when one variable increases and so does the other; while it is negative when one decreases as the other increases. The absence of correlation is described by 0.

Correlation Coefficient Formula

If X and Y are two variables, correlation coefficient 'r' is computed as below:

$$r = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$$

where $\text{cov}\,X,Y = \frac{1}{N}\sum(x-\bar{x})(y-\bar{y})$

cov (X, Y) is called the covariance between X and Y.
N is the total number of observations.
$\bar{x}$, $\bar{y}$ are the means and σ_x, σ_y are the standard deviations of the variables.

$$\bar{x} = \sum x/N\,;\ \bar{y} = \sum y/N$$

$$\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$$

$$\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{N}}$$

Correlation Coefficient can also be calculated using the formula:

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

Export x	Import y	x^2	y^2	xy
42	56	1764	3136	2352
44	59	1936	3481	2596
58	53	3364	2809	3074
55	58	3025	3364	3190
89	65	7921	4225	5785
98	78	9604	6084	7644
66	58	4356	3364	3828
$\Sigma x = 452$	$\Sigma y = 427$	$\Sigma x^2 = 31940$	$\Sigma y^2 = 26463$	$\Sigma xy = 28469$

$$\bar{x} = \frac{452}{7} = 64.57;\ \bar{y} = \frac{427}{7} = 61$$

$$r = \frac{\left(\frac{28469}{7} - 64.57 * 61\right)}{\left(\sqrt{\frac{31940}{7} - 64.57^2}\right) * \left(\sqrt{\frac{26463}{7} - 61^2}\right)}$$

$$r = \frac{(4067 - 3938.77)}{\left(\sqrt{4562.86 - 4169.28}\right) * \left(\sqrt{3780.42 - 3721}\right)}$$

$$r = \frac{(128.23)}{(19.84 * 7.71)}$$

$$r = \frac{128.23}{152.97}$$

Simple Linear Regression Equation

As we know, linear regression is used to model the relationship between two variable. Thus, a simple linear regression equation can be written as:

$Y = a + bX$

Where,

Y = Dependent variable

X = Independent variable

$a = [(\sum y)(\sum x^2) - (\sum x)(\sum xy)] / [n(\sum x^2) - (\sum x)^2]$

$b = [n(\sum xy) - (\sum x)(\sum y)] / [n(\sum x^2) - (\sum x)^2]$

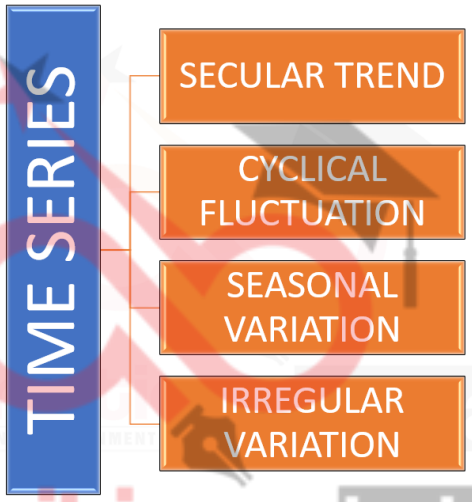
CAIIB Paper 1 (ABM) Module A Unit 5: Time Series

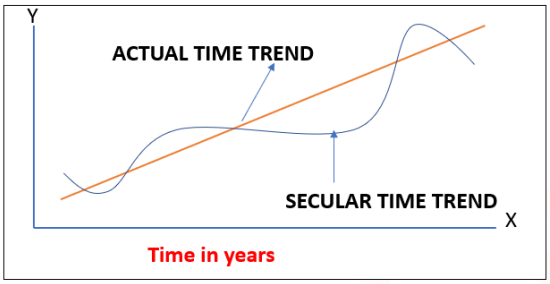
Time Series

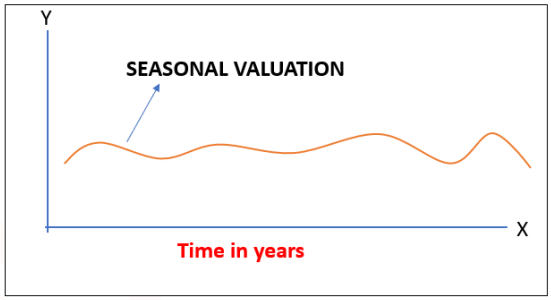
Secular trend is caused by basic inherent factors. Business cycle trends are mostly upward. The quality of forecast depends on the information provided by past data and its validity. **Data or statistical information accumulated** at regular intervals is called **TIME SERIES**.

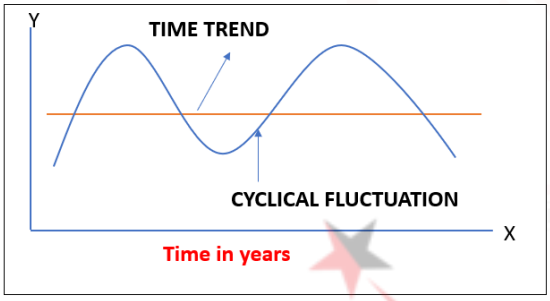
There are 4 types of variations in time series

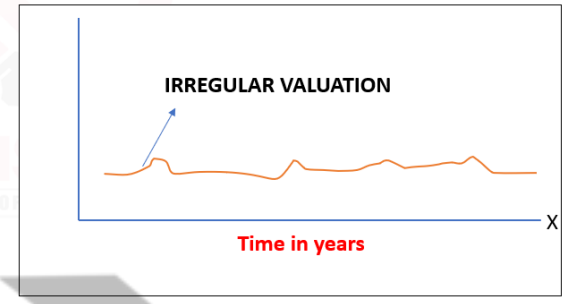
- Secular Trend
- Cyclical Fluctuation
- Seasonal Variation
- Irregular Variation.











Secular Trend

In this first type of variation the change comes over a long period of time. A steady increase in cost of living recorded by Consumer Price Index is a good example. From year to year there is a fluctuation but there is a steady increase in the trend. Let us see the series given here.
Let us try to detect patterns in the information over regular intervals of time. Then let us try to predict to cope with uncertainty.

Year	1997	1998	1999	20W	2001	2002	2003
Number	98	105	116	119	135	156	177

Observations

There is an increase over time of 7 years. But the increases are not equal.

Cyclical Fluctuation

- Most common example of a cyclical fluctuation is a business cycle. Over time, there are years when business cycle hits peak above the trend line. There are also times when business activity slumps, and hits a point below the trend line.

- Fluctuations in business activity occur many times, and they have irregular periods and vary widely in amplitude from cycle to cycle. The time between hitting peaks and lows are periods – it

can be one or many. The cyclical moves do not follow any regular pattern, they are irregular.

Seasonal Variation

- There is a pattern of change within a year. A doctor can expect the number of flu cases to increase in winter. Hill resorts can expect more tourists during summer.
- These are regular patterns and can be used for forecasting the amount of flu vaccines required during winter, the doctor's income during winter, the hotel bookings in resorts and availability of air and train bookings.

Irregular Variation

- The value of the variable is unpredictable, changing in a random manner. The effects of earthquakes, floods, wars, etc., cannot be predicted.
- As a result of flood, the agriculture output suffers. Then the prices go up at an unprecedented rate. This could not be predicted by using time series.
- Even though we described time series as exhibiting one or another variation, in most instances real time series will contain several of these components. Then the question is how to measure them.

Trend Analysis

There are three main reasons, why we should study the trends:

- We will be able to describe historical patterns, which will help us to evaluate the success of previous policies – long-term direction of the time series is given by secular trend.
- Past trends will help us to project the future – some growth rate of population, GDP.
- We will be able to separate the trend component and eliminate it from the series, to get an accurate idea of other components like seasonal fluctuations.

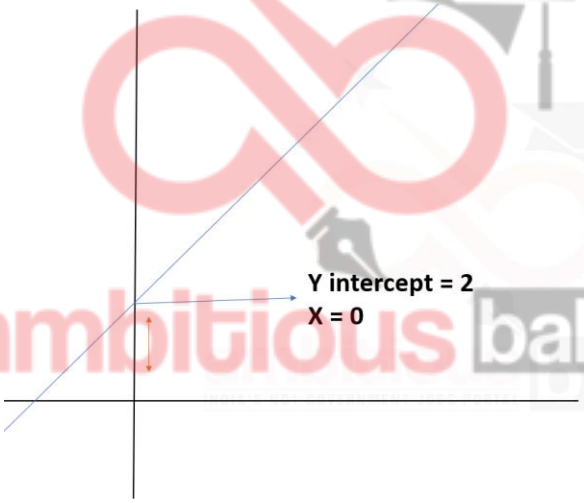
a= Intercept
b = regression coefficient

Equation :

$$Ye = a + bx$$

$$b = \Sigma xy - nx\bar{}\ y\bar{}\ / \Sigma x^2 - nx\bar{}^2$$
$$b = \Sigma xy / \Sigma x^2$$

$$a = y\bar{} - b\ x\bar{}$$
$$a = y\bar{}$$



Year	2009	2010	2011	2012	2013	2014	2015	2016
Number	98	105	116	119	135	156	177	208

Time X	Number Y	T - M	c.Time = x	XY	X^2
2009	98	-3.5	-7	-686	49
2010	105	-2.5	-5	-525	25
2011	116	-1.5	-3	-348	9
2012	119	-0.5	-1	-119	1
2013	135	0.5	1	135	1
2014	156	1.5	3	468	9
2015	177	2.5	5	885	25
2016	208	3.5	7	1456	49
M = ΣT / n 2012.5	ΣY = 1114	ΣX = 0	ΣX= 0	Σxy = 1266	Σx^2 = 168

$\bar{y} = 1114 / 8 = 139.5$

$\bar{x} = 0$

$$b = \Sigma xy - n\bar{x}\bar{y} / \Sigma x^2 - n\bar{x}^2$$

$= 1266 / 168$

$= 7.536$

$$a = \bar{y} - b\bar{x}$$

$= 139.25$

Ye = a + bx

The General equation for annual production

Ye = 139.25 + 7.536x

Estimate the no of unit ,it may produce during 2019

'x' is coded time

= 2(2019-2012.5) = 13

= 139.25 + 7.536 813

= 237.22

= 237 Ships loaded

Parabolic Equation:

Many series may series can be best described by curves. In these cases, the linear model doesnot adequately describe the change in the change in variable as time changes. To overcome this, we use parabolic curves.

1. Σy = an + c Σx^2
2. Σx^2y = a Σx^2 +c Σx^4
3. b= Σxy / Σx^2

T = year	2012	2013	2014	2015	2016	M = 2014
Y = Watch	13	24	39	65	106	Σy = 247
X = T- m	-2	-1	0	1	2	Σx= 0
X^2	4	1	0	1	4	Ex^2 = 10
Xy	-26	-24	0	65	212	Σxy = 227
X^2y	52	24	0	65	424	Σx^2y = 565

Σx^4 = 34

1. Σy = an + c Σx^2

247 = 5a + 10c

2. Σx^2y = a Σx^2 +c Σx^4

565 = 10a +34c

3. b= Σxy / Σx^2

= 227/10 = 22.7

a = 39.3
b = 22.7
c = 5.07

Ye = a + bx + cx^2

= 39.3 + 22.7* x + 5.07x^2

Suppose want to calculate for 2021

X = 2021 – 2014 = 7

= 30.3 + 22.7*7 +5.07 *49

= 446.6

Cyclical variation

Cyclical variation is a component of the time series, which tends to oscillate above and below the secular trend line for periods longer than a year. Seasonal variation makes a complete regular cycle within each year and does not affect one year any more than another. Once we identify the secular trend,

we can isolate the remaining cyclical and irregular components of the trend. Let us assume cyclical component explains most of the variations left unexplained by the trend analysis.

Residual Method

- **Percentage of Trend = y actual / y trend *100**
- **Relative cycle residual**

X Year	2009	2010	2011	2012	2013	2014	2015	2016
Y	75	78	82	82	84	85	87	91

Year	Y	X = T-M	X * 2	X.Y	X^2	Estimated (83 + x)	% of Trend Y/Y cap*100	RCR
2009	75	-3.5	-7	-525	49	76	98.7	-1.3
2010	78	-2.5	-5	-390	25	78	100	0
2011	82	-1.5	-3	- 246	9	80	102.5	2.5
2012	82	-0.5	-1	-82	1	82	100	0
2013	84	.05	1	84	1	84	100	0
2014	85	1.5	3	255	9	86	98.8	-1.2
2015	87	2.5	5	435	25	88	98.8	-1.2
2016	91	3.5	7	637	49	90	101.1	1.1
M = 2012.5	ΣY = 664	ΣX= 0	ΣX = 0	ΣXY = 168	ΣX^2=168			

Multiply X by 2 if n is even

Ye = a +bx

b = ΣXY/ ΣX^2

= 168/ 168 = 1

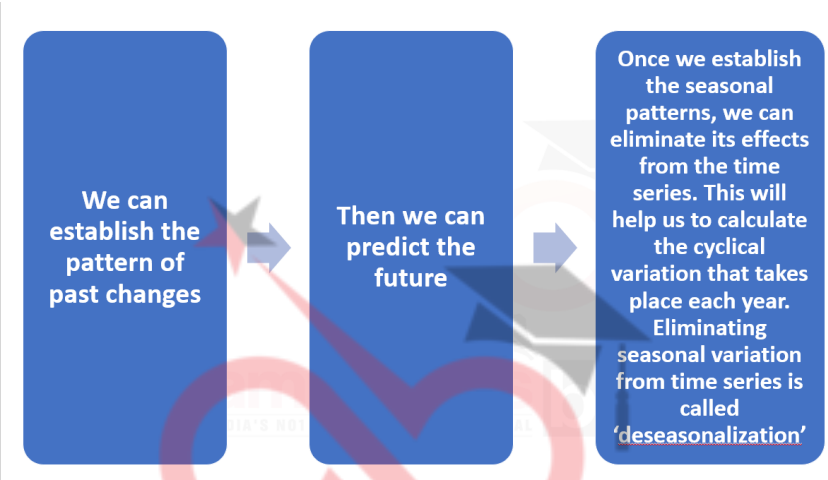
$a = \bar{y} - b\bar{x}$

= 664/8 = 83

Ye= 83 + x

Seasonal Variation

Time series also includes seasonal variation. Seasonal variation is repetitive and predictable. This can be defined as movements around the trend line in one year or less. In order to measure seasonal variations, time intervals must be measured in small units, like days, weeks, etc.



Year	Spring	Summer	Fall	Winter
2012	1861	2203	2415	1908
2013	1921	2343	2514	1986
2014	1834	2154	2098	1799
2015	1837	2025	2304	1965
2016	2073	2414	2339	1967

Year	Season	Occu	Step 1	Step 2	Step 3	Step 4 = SI
2012	I	1861				
	II	2203				
	III	2415	8387	2096.75	2104.25	114.8
	IV	1908	8447	2111.75	2129.25	89.6
2013	I	1921	8587	2146.75	2159	89
	II	2343	8686	2171.50	2181.25	107.4
	III	2514	8764	2191	2180.125	115.3
	IV	1986	8677	2169	2145.625	92.6
2014	I	1834	8488	2122	2070	88.6
	II	2154	8072	2018	1994.625	108
	III	2098	7885	1971	1971.625	106.4
	IV	1799	7888	1972	1955.875	92
2015	I	1837	7759	1939.75	1965.5	93.5
	II	2025	7965	1991.25	2012	100.6
	III	2304	8131	2032.75	2062.25	111.7
	IV	1965	8367	2091.75	2140.375	91.8
2016	I	2073	8756	2189	2193.375	94.5
	II	2414	8791	2197.75	2198	109.8
	III	2339	8793	2198.25		
	IV	1967				

Step 5

Year	Spring	Summer	Fall	Winter
2012			114.8	89.6
2013	89	107.4	115.3	92.6
2014	88.6	108	106.4	92
2015	93.5	100.6	111.7	91.8
2016	94.5	109.8		
Modified mean	91.25	107.70	113.25	91.90

For modified mean : discard lowest and highest value

Total of Indices = 404.10

STEP 6:

Four quarter indices = 400

= 400/404.1 = 0.9899

Spring = 91.25 * 0.9899 = 90.3

Summer = 107.7 * .9899 = 106.6

Fall = 112.1

Winter = 91

Irregular Variation

The final component is irregular variation. After we have eliminated trend, cyclical and seasonal variations from the time series, we may still have unpredictable factor left. Irregular variations occur over **very short intervals and follow random patterns. We may not be able to isolate them mathematically,** but we may isolate the causes for the same. For example, an unusually very cold winter in a region may increase electricity consumption significantly. Wars may increase air and train travel because of the movement of troops. We may not be able to identify all causes. But over time, these random variations tend to correct themselves.

Sales per Quarter (in Rs. 10000)

Year	Spring	Summer	Fall	Winter
2012	16	21	9	18
2013	15	20	10	18
2014	17	24	13	22
2015	17	25	11	21
2016	18	26	14	25

STEP 1: First calculate Seasonal Indices

Q1 = 95.1

Q2 = 129.9

Q3 = 61.2

Q4 = 113.9

Deseasonlised

Actual * seasonal index/ 100

Year	Season	Sale	SI	Deseanalised (Sales/ SI)
2012	I	16	95.1	16.8
	II	21	129.9	16.2
	III	9	61.2	14.7
	IV	18	113.9	15.8
2013	I	15	95.1	15.8
	II	20	129.9	15.4
	III	10	61.2	16.3
	IV	18	113.9	15.8
2014	I	17	95.1	17.9
	II	24	129.9	18.5
	III	13	61.2	21.2
	IV	22	113.9	19.3
2015	I	17	95.1	17.9
	II	25	129.9	19.2
	III	11	61.2	18
	IV	21	113.9	18.4
2016	I	18	95.1	18.9
	II	26	129.9	20
	III	14	61.2	22.9
	IV	25	113.9	21.9

STEP 2: TREND LINE (Ye = a + bx)

a= y

b = xy/x^2

Year	Season	X	T- M 10.5	Coded t(x)	x^2	Y	xY
2012	I	1	-9.5	-19	361	16.8	-319.2
	II	2	-8.5	-17	289	16.2	-275.4
	III	3	-7.5	-15	225	14.7	-220.4
	IV	4	-6.5	-13	169	15.8	-205.4
2013	I	5	-5.5	-11	121	15.8	-173.8
	II	6	-4.5	-9	81	15.4	-138.6
	III	7	-3.5	-7	49	16.3	-114.1
	IV	8	-2.5	-5	25	15.8	-79
2014	I	9	-1.5	-3	9	17.9	53.7
	II	10	-.5	-1	1	18.5	-18.5
	III	11	.5	1	1	21.2	21.2
	IV	12	1.5	3	9	19.3	57.9
2015	I	13	2.5	5	25	17.9	89.5
	II	14	3.5	7	49	19.2	134.4
	III	15	4.5	9	81	18	162
	IV	16	5.5	11	121	18.4	202.4
2016	I	17	6.5	13	169	18.9	117
	II	18	7.5	15	225	20	300
	III	19	8.5	17	289	22.9	389.9
	IV	20	9.5	19	361	21.9	416.1

Σx^2 = 2660
ΣY = 360.9
Mean Y = 360.9/20
= 18
Σxy = 420.3

b = ΣXY/ ΣX^2
= 420.3/ 2660
= 0.16

a = Mean Y
= 18

Trend line =
a + bx
= 18 + 0.16x

STEP 3 : Trend Value

Percent of Trend

Year	Season	Coded t(X)	Y	a+bx = Ye 18 + 0.16x	% of Trend Y/Ye *100
2012	I	-19	16.8	14.96	112.3
	II	-17	16.2	15.28	105
	III	-15	14.7	15.50	94.2
	IV	-13	15.8	15.92	99.2
2013	I	-11	15.8	16.24	97.3
	II	-9	15.4	16.56	93
	III	-7	16.3	16.88	96.6
	IV	-5	15.8	17.20	91.9
2014	I	-3	17.9	17.52	102.2
	II	-1	18.5	17.84	103.7
	III	1	21.2	18.16	116.7
	IV	3	19.3	18.48	104.4
2015	I	5	17.9	18.80	95.2
	II	7	19.2	19.12	100.4
	III	9	18	19.44	92.6
	IV	11	18.4	19.76	93.1
2016	I	13	18.9	20.80	94.1
	II	15	20	20.40	98
	III	17	22.9	20.72	110.5
	IV	19	21.9	21.94	104.1

Suppose management wants to determine the sales value for the 3rd qt of 6th year

23- 10.5 = 12.5 (coded X value)

12.5 * 2 = 25

Ye = a +bx

= 18 + 0.16*25

= 22

Means 22000 units

CAIIB Paper 1 (ABM) Module A Unit 6: Theory of Probability

Introduction To Probability

- Probability means chance/s or possibility of happening of an event. For example, suppose we want to plan for a picnic in a weekend.
- Before planning we may check the weather forecast and see what is the chance that there will be rain at that time, accordingly we may do the planning.
- Probability gives a numerical measure of this chance or possibility.
- Suppose it says that there is a 60% chance that rain may occur in this weekend, 60% or 0.6 is called the probability of raining. To understand the concept of probability first we have to understand the concepts of Factorial, Permutations and Combinations.

Factorial:

- In mathematics, Factorial is equal to the product of all positive integers which are less than or equal to a given positive integer. The Factorial of an integer is denoted by that integer and an exclamation point.
- **Thus, factorial five is written as 5! which is equal to 1 × 2 × 3 × 4 × 5 = 120**
- **The product of the first n natural numbers is called factorial n and is denoted by n! =**

  n × (n – 1) × (n – 2) × … × 2 × 1
- **The above formula can also be represented as n! = n × (n – 1) … (n – r + 1) × (n – r)!**
- **Where r < n It may be noted that:**

  0! = 1, 1! = 1

Permutations and Combinations

- A permutation is the arrangement of objects in which order is the priority. The fundamental difference between permutation and combination is the order of objects, in permutation, the order of objects is very important, i.e., the arrangement must be in the stipulated order of the number of objects, taken only some or all at a time.
- The combination is the arrangement of objects in which order is irrelevant. The notation for permutation is P (n, r) or ${}^{n}P_{r}$, denoting the number of permutations of n things when r things are selected at a time.
- If there are three things a,b, and c then permutations of three things taken two at a time is denoted by P (3, 2) or ${}^{3}P_{2}$.
- It is given by (a, b), (a, c), (b, c), (b, a), (c, a), (c, b) = 6
- **In general,**

$$P(n, r) = {}^{n}P_{r} = \frac{n!}{(n-r)!}$$

P (n, r) is the number of permutations when r things are selected at a time from n items.

The notation for combination is C(n, r) or ${}^{n}C_{r}$ which is the number of combinations or selections of n things if only r things are selected. If there are three things a, b and c then combination of these three things taken two at a time is denoted by ${}^{3}C_{2}$ and is given by (a, b), (a, c), (b, c) = 3

$$^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

Example: Using 5 letter of word SHYAM, how many distinct word can be formed?

N= 5

R= 5

$^{5}P_{5}$ = 5!/ (5-5)! = 5*4*3*2/0! = 5*4*3*2/1 = 120

Note: Permutation and Combination are related to each other by formula P(n,r) = r! * C(n,r).

Example: In how many ways 3 pencils can be selected from 5 pencils?

3 pens can be selected from 5 pens in $^{5}C_{3}$ ways

$^{5}C_{3}$ = 5! / 3! 2! × = 10 ways

Example: From a group of 7 boys and 6 girls, 3 boys and 4 girls is to be selected. In how many ways this can be done?

3 boys can be selected from 7 boys in $^{7}C_{3}$ ways

= $^{7}C_{3}$ = 7! /3! 4! ×

= 7* 6 *5 *4! /3*2*4! = 35

4 girls can be selected from 6 girls in $^{6}C_{4}$ ways

= 6! 4! 2! = 6 *5 *4! /4! *2 = 15

3 boys and 4 girls can be selected in $^{7}C_{3} \times {}^{6}C_{4}$ = 35 × 15 = 525 ways.

Random Experiment or Trial

An operation or experiment conducted under identical conditions and which has a number of possible outcomes is called Random Experiment or Trial.

Example: 1. Tossing a coin 2. Throwing a dice 3. Selecting a card form a pack of cards

Sample Space and Sample Points

The set of all possible outcomes of a random experiment is called sample space.

The elements of the sample space are called sample points. Sample space is denoted by S.

Example: 1. In an experiment of throwing a coin, S = {H,T]

2. In an experiment of throwing a dice, S = {1, 2, 3, 4, 5, 6}

The number of sample points in a sample space of random experiment is denoted by n (s).

For example, (1) n (S) = 2, and

example (2) n (S) = 6

Event

Any subset of the sample space S is called an event.

If S is a sample space and A is a subset of S (i.e., A⊂ S), then A is called an event.

Example: In an experiment of throwing dice where S = {1, 2, 3, 4, 5, 6}, the event of getting odd numbers is A = {1, 3, 5}

Types of Events

Certain Event

- If sample points in an event are same as sample points in sample space of that random experiment, then the event is called a certain event.
- Example: Getting any number between 1 to 6 on a dice is a certain event.

Impossible Events

- An event which never occurs or which has no favourable outcomes is called an impossible event. In other words, the event corresponding to the set φ (null set) is called an impossible event.
- Example: Getting a number 7 on a dice is an impossible event.

Mutually Exclusive Events

- Events are said to mutually exclusive if the happening of any of them restricts the happening of the others i.e., if no two or more of them can happen together or simultaneously in the same trial.
- Example: In tossing a coin event head and tail are mutually exclusive. Note: If A & B are mutually exclusive events of sample space S, then A ∩ B = φ.

Example: In tossing a coin event head and tail are mutually exclusive. Note: If A & B are mutually exclusive events of sample space S, then A ∩ B = φ.

Equally Likely Events

- Events are said to be equally likely if they have equal choice to occur. In other words, outcomes of a trial are said to be equally likely if taking into consideration all relevant evidences, there is no reason to prefer one with respect to other.
- Example: In throwing a dice all the six faces are equally likely to occur.

Exhaustive Events

- If the sample points of the events taken together
- Note: If A & B are exhaustive events of sample space S, then A U B = S.
- Example: Random Experiment:
- Throwing a dice S = {1,2,3,4,5,6},
- A= Event of odd numbers = {1, 3, 5}
- B= Event of odd numbers = {2, 4, 6},
- C= Event of multiple of 3 = {3, 6}
- Here A U B = {1, 2, 3, 4, 5, 6} = S,
- Here A and B are called exhaustive events
- But A U C = {1, 3, 5, 6} ≠ S,
- so A and C are not exhaustive events.

Complementary Event

If A is an event in sample space S, then the non-occurrence event of A is called Complementary event of A.

Two events A and B are called complementary events, if A and B exhaustive as well as mutually exclusive events.

In other words, A and B are called complementary events if

A U B = S and A ∩ B = φ.

Example: Random Experiment: Throwing a dice, S = {1, 2, 3, 4, 5, 6}, A = {1, 2}, B = {3, 4, 5, 6} As A U B = S and A ∩ B = φ, A and b are complementary events.

Complementary event of A is denoted by A^c, A/ or A·

Mathematical Definition Of Probability

If the sample space S of a random experiment consists of n equally likely, exhaustive and mutually exclusive sample points and m of them are favourable to an event A, then the probability of event A is given by

$$P(A) = \frac{m}{n} = \frac{\text{Number of Sample Point in A}}{\text{Number of Sample Point in S}} = \frac{n(A)}{n(S)}$$

Since $0 \le m \le n$, $\frac{0}{n} \le \frac{m}{n} \le \frac{n}{n} \Rightarrow 0 \le P(A) \le 1$

Number of favourable items/ Total number of outcomes

Example: Two unbiased dice are thrown. Find the probability that:

- Both the dice show same number.
- First dice shows 6.
- The total of the numbers on the dice is 8.

Solution In a random throw of two dice, the total number of cases is given below:

(1, 1),	(2, 1),	(3, 1),	(4, 1),	(5, 1),	(6, 1),
(1, 2),	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

n(S) = 36

A: Both the dice show same number

n(A)/n(S) = 6/36 = 1/6

B: First die show 6 n(B)/n(S) = 6/36 = 1/6

C: Total of the number

on the dice is 8 n(C)/n(S) = 5/36

Example: Two unbiased coins are tossed simultaneously. Find the probability of getting –

at least one tail,

majority of heads

S = {(H, H), (H, T), (T, H), (T, T)}

n (S) = 4

(i)A: At least one tail,

P (A) = n (A)/ n (S) = 3 /4

(ii)B: Majority of heads

P (B) = n (B)/ n(S) = 1/ 4

Addition Theorem

Let A and B are two events (subsets of sample space S) and are not disjoint, then the probability of the occurrence of A or B or A and B both, in other words probability of occurrence of at least one of them is given by,

P (A U B) = P (A) + P (B) – P (A∩B)

- ✓ **P (A U B) = A Or B**
- ✓ **P (A∩B) = A and B**

Example: Find the probability that a card drawn from a pack of cards will be a red or a picture card.

Probability of selecting a red card = 26 = Event A

P(A) = 26/52 = 1/2

Probability of getting picture card = 6 = Event B

P(B) = 12/52 = 3/13

There are 6 red cards which are picture cards,

P (A∩B) = 6/52

P (A U B) = P (A) + P (B) – P (A∩B)

½ + 3/13 – 6/52 = 8/13

➢ **Corollary 1:**

- If the events A and B are mutually exclusive, then
- $A \cap B = \varphi$ Ì $P(A \cap B) = 0 \Rightarrow P(A \cap B) = P(A) + P(B)$

➢ **Corollary 2:** For three non-mutually exclusive events

- A, B, C $P(AU B \cap C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

➢ **Corollary 3:** If A and B are any two events, then

$P(A) = P(A \cap B) + P(A \cap B^c)$

➢ **Corollary 4:** If A^c is complementary event of A then

$P(A^c) = 1 - P(A)$

➢ **Corollary 5:**

$P(B \cap A^c) = P(B) - P(B \cap A)$

➢ **Corollary 6:**

If $A \subset B$

$P(A) \leq P(B)$

➢ **Corollary 7:** P (Non-occurrence of events)

$P(A^c \cap B^c) = 1 - P(A \cup B)$

Conditional Probability

- The conditional probability of an event A is the probability that the event will occur given the knowledge that an event B has already occurred.

P (A/B).

- If the events A and B are such that the occurrence of A doesn't depend upon occurrence of event B, (A and B are independent event), the conditional probability of event A given event B is simply the probability of event A, that is P (A).

- Similarly, probability of event B given that event A has already occurred is denoted by P (B/A).

$$P(B/A) = P(A \cap B) / P(A)$$

Example: Consider a fair coin is tossed 3 times

S = (HHH, HHT, HTH, TTT, TTH, THT, THH, HTT) = 8

Event A = Atleast two tail appear

Event B – First coin show Head

P(A) = (TTT, TTH, THT, HTT) = 4/8 = ½

P(B) = (HHH, HHT, HTH, HTT) = 4/8 = ½

P (A ∩ B) = 1/8

P(A/B) = 1/ 8 / ½ = 1/4

Multiplication Theorem

- If A and B are two events of a sample space S associated with an experiment, then the probability of simultaneous occurrence of events A and B is given by

$$P(A \cap B) = P(A)\, P(B/A) = P(B)\, P(A/B)$$

Independent Events

Two events A and B are independent of each other if the occurrence or non-occurrence of one does not affect the occurrence of the other.

$$P(A \cap B) = P(A)\, P(B)$$

Example: Two balls are drawn from a bag one by one with 2 white and 3 black balls. What is the probability that the second ball is white?

Event W1 = first ball - White Ball

Event B1 = First Ball – Black Ball

Event W2 = Second Ball – White Ball

1st White Ball = 2/5 + ¼

2nd Black Ball = 3/5 + 2/4

P(W2) = P(W1) + P(W2/W1) + P(B1) + P(W2/B1)

2/5 + ¼ + 3/5 + 2/4= 2/5

Random Variable

- A random variable is a function that associates a real number with each element in the sample space.
- In other words, a random variable is a function **X: S → R,**
- where S is the sample space of the random experiment under consideration and R is the real number line.

Example. Consider the random experiment of tossing a coin two times and observing the result (a Head or a Tail) for each toss.

Let X denote the total number of heads obtained in the two tosses of the coin.

Example: Suppose that you play a certain lottery by buying one ticket per week. Let X be the number of weeks until you win a prize. X is a random variable.

Discrete Random Variable:

- If a random variable takes a finite number or countable infinite number of possibilities, it is called a discrete random variable.
- Example: 1. Age in years 2. Number of arrivals in a clinic 3. Number of accidents

Continuous Random Variable:

- If a random variable takes infinite number of possibilities, it is called a continuous random variable.
- Example 1. Percentage of marks 2. Weight 6.5 PROBABILITY

Binomial Distribution

- Consider a random experiment consisting of n repeated independent trials with p the probability of success at each individual trial. Let the random variable X represent the number of successes in the n repeated trials.
- Then X follows a Binomial distribution.

The definition of this distribution is:

- A random variable X has a binomial distribution,

X ~ Binomial (n, p), if the discrete density of X is given by:

$P[X=x] = f(x) = {}^{n}C_{x}\, p^{x}\, (1-p)^{n-x}$,

x = 0, 1, 2..., n = 0 otherwise

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

x = 0, 1, 2..., n = 0 otherwise where p + q = 1

P = the probability of success

n is the total number of trials.

Example: Toss a coin for 10 times and you wan to get head 4 times & probability of coming head is 0.5 calculate f(x)?

n = 10, x= 4 & p = 0.5

q = 1- 0.5 = 0.5

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

$= {}^{10}C_4 * 0.5^{10}\, o.5^{10}$

Binomial Distribution Real Life Examples

Many instances of binomial distributions can be found in real life.

- If a new drug is introduced to cure a disease, it either cures the disease (it's successful) or it doesn't cure the disease (it's a failure).
- If you purchase a lottery ticket, you're either going to win money, or you aren't. Basically, anything you can think of that can only be a success or a failure can be represented by a binomial distribution.
- **Mean = np**
- **Variance = np (1 – p) = npq**
- **SD = √ variance**
- **Measure of Skewness = â1 = (1-2p)2/npq**
- **Measure of Kurtosis = â2 = 3 + [1 – 6pq /npq]**
- **Binomial Distribution is symmetric if p = q = 0.5**
- **If p < 0.5, distribution is positively skewed and**
- **if p > 0.5, distribution is negatively skewed.**

Mode

M= (n+1) p

- **If M is not an integer, mode is the integral part lying between M – 2 and M.**

- **If M is an integer, there are two modes and thus the distribution is bimodal, and two modes are M – 1 and M.**

Problem: If X follows Binomial distribution with n = 8, p = 1/2, then Find P [IX-4I ≤ 2]

Solution : P[-2<=(X-4)<=2]

P[2<= X<=6]

P[2<= X<=6] = P(X=2) + P(X=3) + P(X=4) + P(X=5) + P(X=6)

$f(x) = {}^{n}C_x\, p^x\, q^{n-x} = {}^{8}C_x\, 1/2^x\, 1/2^{8-x}$

${}^{8}C_x\, 1/2^x\, 1/2^{8-x} =$

$= {}^{8}C_2\, 1/2^2\, 1/2^{8-2} + {}^{8}C_3\, 1/2^3\, 1/2^{8-3} + {}^{8}C_4\, 1/2^4\, 1/2^{8-4} + {}^{8}C_5\, 1/2^5\, 1/2^{8-5} + {}^{8}C_6\, 1/2^6\, 1/2^{8-6}$

$= (1/2)^8\, ({}^{8}C_2 + {}^{8}C_3 + {}^{8}C_4 + {}^{8}C_5 + {}^{8}C_6)$

= 1/256 (128+ 56+ 70 +56 +28) = 119/128

Poisson Distribution

- The Poisson probability distribution was introduced by S. D. Poisson.
- A random variable X, taking on one of the values 0, 1, 2,..., is said to be a Poisson random variable with parameter λ, λ > 0, if its probability mass function is given by

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2,$$

- The symbol e stands for a constant approximately equal to 2.7183. It is a famous constant in mathematics, named after the Swiss Mathematician L. Euler, and it is also the base of the so-called natural logarithm

Some examples of Poisson probability are:

- The number of misprints on a page (or a group of pages) of a book.
- The number of people in a community living to 100 years of age
- The number of wrong telephone numbers that are dialed in a day.
- The number of transistors that fail on their first day of use.
- The number of customers entering a post office on a given day.
    - Mean = λ,
    - Variance = λ

- Measure of Skewness = $\beta 1 = 1 / \lambda$
- Measure of Kurtosis $\beta 2 = 3 + 1/\lambda$
- Mode: If l is not an integer mode is the integral part lying between $\lambda - 1$ and λ
- If λ s an integer, there are two modes and thus the distribution is bimodal and two modes are $\lambda - 1, \lambda$

Problems. Births in a hospital occur randomly at an average rate of 1.8 births per hour.

- What is the probability of observing 4 births in a given hour at the hospital?
- What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2,$$

Let X= No. of births in a given hour = 4 with Mean rate $\lambda = 1.8$

$e^{-1.8} = .16529$

$f(x) = e^{-1.8} * 1.8^4 / 4!$

$.16529 * 10.4976 / 24 = 0.0723$

(ii) We want $P(X \geq 2) = P(X = 2) + P(X = 3) + \ldots$, i.e., an infinite number of probabilities to calculate

but $P(X \geq 2) = P(X = 2) + P(X = 3) + \ldots = 1 - P(X < 2)$

$= 1 - (P(X = 0) + P(X = 1)$

$= 1 - (0.16529 + 0.29753) = 0.538$

Normal Distribution

- A normal distribution is a distribution that occurs naturally in many situations where 50% of the data will fall to the left of the mean and 50% will fall to the right.
- For example, Height of the population, most of the people in a specific population are of average height. The number of people taller and shorter than the average height people is almost equal, and a very small number of people are either extremely tall or extremely short.
- Some other examples are distribution of Income in economy, distribution of marks in an exam, etc.
- A random variable X is said to follow Normal Distribution if its pdf is given by

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

Note:

- μ and σ2 are called parameters of Normal Distribution.
- If μ = 0 and σ2 = 1, then the Normal variable is called Standard Normal Variable. Generally, it is denoted by Z.

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad -\infty < z < \infty$$

- The graph of Normal Distribution is bell shaped and symmetric.
- Quartile deviation is 0.6745σ
- Mean deviation is 0.7979σ

Mean = Median = Mode = μ

Problem: Normal population of 1000 employees has mean income Rs. 800 per day and variance 400, Find no. of employees where income between [P(Z= 1) = 0.3413, P(Z= 2) = 0.4772 & P(Z= 2.5) = 0.4938 P(Z= 5) = 0.5]

P (750 < x < 820)

P (x > 700)

P (x > 760)

n= 1000 , μ= 800 & σ^2 = 400

Z = X- μ /σ

i)X = 750 = 750- 800/ 20 = 2.5 = 0.4938

X = 820 = 820-800 / 20 = 1 = 0.3413

= 0.4938 + 0.3413 = 0.8351 = 83.51%

ii)X = 700 = 700 – 800/ 20 = 5 = 0.5

0.5 + 0.5 = 1 = 1000 employees

n= 1000 , μ= 800 & σ^2 = 400

Z = X- μ /σ

iii) X = 760 = 760-800/20 = 2 = 0.4772

0.4772 + 0.5 = 0.9772

Credit Risk

- We can apply probability concept and different formulas and laws of probability in different practical field.
- One very important application is Credit Risk.
- When lenders offer mortgages, credit cards, any type of loan to different customers, there could be a risk that the customer or borrower might not repay the loan.
- Similarly, if a company extends credit to a customer, there could be a risk that the customer might not pay their invoices.
- We are interested to calculate this risk of not repaying any due payment. This is **called Credit Risk.**
- Credit risk also represents the risk that a bond issuer may fail to make a payment when requested, or an insurance company will not be able to pay a claim.
- Thus, Credit Risk is the possibility or chance or probability of a loss occurring due to a borrower's failure to repay a loan to the lender or to satisfy contractual obligations. It refers to a lender's risk of having its cash flows interrupted when a borrower does not repay the loan taken from him.

There are three types of credit risks.

Credit default Risk :

Credit default risk is the type of loss that is incurred by the lender either when the borrower is unable to repay the amount in full or when 90 days pass the due date of the loan repayment. This type of credit risk is generally observed in financial transactions that are based on credit like loans, securities, bonds or derivatives.

Concentration Risk:

Concentration risk is the type of risk that arises out of significant exposure to any individual or group because any adverse occurrence will have the potential to inflict large losses on the core operations of a bank. The concentration risk is usually associated with significant exposure to a single company or industry or individual.

Country risk

- The risk of a government or central bank being unwilling or unable to meet its contractual obligations is called Country or Sovereign Risk.
- When a bank or financial institution or any other lender has an indication that the borrower may default the loan payment, he will be interested to calculate the expected loss in advance.
- The expected loss is based on the value of the loan (i.e., the exposure at default, EAD) multiplied by the probability, that the borrower will default (i.e., probability of default, PD).
- In addition, the lender takes into account that even when the default occurs, it might still get back some part of the loan.
- Hence, PD * EAD is further multiplied by the estimation of the part of the loan which will be lost in case that a default occurs (i.e., loss given default, LGD).

Expected loss = PD * EAD * (1 – LGD)

Problem: Let a credit of Rs. 2,000,000 was extended to a company one year ago. Determine the expected loss for the exposure if the company defaults completely, where the loss given default is 50%.

Probability of default, PD = 100

Loss given default, LGD = 50%

Expected loss = 100% * Rs. 2,000,000 * (1 – 50%)

= Rs. 1,000,000

Value At Risk (VaR)

- The concept of value at risk is associated with portfolio of an individual or an organisation.
- A portfolio is a collection of different kinds of assets owned by an individual or organisation to fulfil their financial objectives.
- One can include fixed deposit or any investment where he or she can earn a fixed interest, equity shares, mutual funds, debt funds, gold, property, derivatives, and more in his portfolio.
- In any type of investment where one can earn fixed interest are not risky, but risk is associated with the investments in Equity market, Mutual Funds, Gold, etc.
- Value at risk (VaR) is a financial metric that one can use to estimate the maximum risk of an investment over a specific period.

If the portfolio value is Rs. 30,000 and if 1-month average return and standard deviation is 10% and 12% respectively, calculate daily VaR at 95% confidence level.

VAR at 95% confidence level

= [Return of the portfolio –1.65 *σ][Value of the portfolio]

= [0.1–1.65*0.12] *30000

= [0.1 –0.198] *30000 = –2940 = 9.8% of the portfolio.

CAIIB Paper 1 (ABM) Module A Unit 7: Estimates

Estimates

Estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.

We can make **two types of estimates** about a population: **a point estimate and an interval estimate.** A point estimate is a single number that is used to estimate an unknown population parameter If, while watching a cricket team on the field, you say. 'Why, I bet they will get 350 runs,' you have made a point estimate. A department head would make a point estimate if she said, 'Our current data indicate that this course will have 350 students next year.'

Estimator And Estimates

Asample statistic that is used to estimate a population parameter is called an estimator. The sample mean x can be an estimator of the population mean μ, and the sample proportion can be used as an estimator of the population proportion. We can also use the sample range to estimate the population range. When we have observed a specific numerical value of our estimator, we call that value as an estimate. In other words, an estimate is a specific value of a statistic or an estimator. We form an estimate by taking a sample and computing the value taken by our estimator in that sample. Suppose, we calculate the mean odometerreading (mileage)from a sample of used taxis and find it to be 98,000 miles. If we use this specific value to estimate the mileage for a whole fleet of used taxis, the value 98,000 miles would be an estimate.

Criteria of a Good Estimator

Some statistics are better than others. Fortunately, we can evaluate the quality of a statistic as an estimator by using four criteria:

- **Unbiased:** This is a desirable property for a good estimator to have. The term unbiased refers to the fact that a sample mean is an unbiased estimator of a population mean because the mean of the sampling distribution of sample means taken from the same population is equal to the population mean itself.
- **Efficiency:** Another desirable property of a good estimator is efficiency. Efficiency refers to the size of the standard error of the statistic. If we compare two statistics from a sample of the same size and decide which one is the more efficient estimator, we would pick the statistic with the smaller standard error or standard deviation of the sampling distribution.
- C**onsistency**: A statistic is a consistent estimator of a population parameter if, as the sample size increases, it becomes almost certain that the value of the statistic comes very close to the value of the population parameter. If an estimator is consistent, it becomes more reliable with large samples.

- **Sufficiency:** An estimator is sufficient if it makes so much use of the information in the sample that no other estimator could extract from the sample, additional information about the population parameter being estimated.

Point estimate

- **A point estimate is often insufficient, because it is either right or wrong. If you are told only that her point estimate of enrollment is wrong, you do not know how wrong it is, and you cannot be certain of the estimate's reliability**.
- If you learn that it is off by only 10 students, you would accept 350 students as a good estimate of future enrollment. But if the estimate is off by 90 students, you would reject it as an estimate of future enrollment. Therefore, a point estimate is much more useful if it is accompanied by an estimate of the error that might be involved.

Interval estimate

- **An interval estimate is a range of values used to estimate a population parameter. It indicates the error in two ways**: by the extent of its range and by the probability of the true population parameter lying within that range. In this case, the department head would say something like, 'I estimate that the enrollment in this course next year will be between 330 and 380 and that it is very likely that the exact enrollment will fall within this interval.'
- She has a better idea of the reliability of her estimate. If the course is taught in sections of about 100 students each, and if she had tentatively scheduled five sections, then on the basis of her estimate, she can now cancel one of those sections and offer an elective instead.

Estimator

A sample statistic that is used to estimate a population parameter is called an estimator.

Criteria of a Good Estimator

Some statistics are better than others. Fortunately, we can evaluate the quality of a statistic as an estimator by ***using four criteria:***

- **Unbiased**: This is a desirable property for a good estimator to have. The term unbiased refers to the fact that a sample mean is an unbiased estimator of a population mean because the mean of the sampling distribution of sample means taken from the same population is equal to the population mean itself.
- **Efficiency:** Another desirable property of a good estimator is that it be efficient. Efficiency refers to the size of the standard error of the statistic.

- **Consistency:** A statistic is a consistent estimator of a population parameter if as the sample size increases, it becomes almost certain that the value of the statistic comes very close to the value of the population parameter.
- **Sufficiency:** An estimator is sufficient if it makes so much use of the information in the sample that no other estimator could extract from the sample additional information about the population parameter being estimated.

Relationship between Confidence Level and Confidence Interval

- You may think that we should use a high confidence level, such as 99 per cent, in all estimation problems. After all, a high confidence level seems to signify a high degree of accuracy in the estimate. In practice, however, high confidence levels will produce large confidence intervals, and such large intervals are not precise; they give very fuzzy estimates.
- There is a direct relationship that exists between the confidence level and the confidence interval for any estimate. As you set a tighter and tighter confidence interval, you would get to a lower and lower confidence level.

Confidence Intervals

Statisticians use a **confidence interval** to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

- A confidence level.
- A statistic.
- A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic* ± *margin of error*.

For example, suppose we compute an interval estimate of a population parameter. We might describe this interval estimate as a 95% confidence interval. This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the *sample statistic* ± *margin of error* 95% of the time.

Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Confidence Level

The probability part of a confidence interval is called a **confidence level**. The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.
Here is how to interpret a confidence level. Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

Margin of Error

In a confidence interval, the range of values above and below the sample statistic is called the **margin of error**.

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Note: Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

Consider the following results of 10 tosses of a coin: H, T, T, T, T, H, T, H, T, T a) Estimate the probability of head (H) for this coin. b) Estimate the standard error of your estimate.

Let X denote the toss of a single coin. Further, let X = 1 if a head results, and X = 0 if a tail results. This X is a Bernoulli (p) random variable, where p denotes the probability of head. Let $\hat{p}$ denote the estimator of p.

a) The estimated value of p is $\hat{p} = (1 + 0 + 0 + \ldots + 1 + 0 + 0)/10 = 0.3$.
b) The estimated standard error of $\hat{p}$ is $\sqrt{\hat{p}(1 - \hat{p})/n} = \sqrt{0.3(0.7)/10} = 0.14$.

Suppose the following data shows the number of the problems from the Practice Problems Set attempted in the past week by 10 randomly selected students: 2, 4, 0, 7, 1, 2, 0, 3, 2, 1.

a) Find the sample mean.
b) Find the sample variance.
c) Estimate the mean number of practice problems attempted by a student in the past week.
d) Estimate the standard error of the estimated mean.

a) $X = Pn\ i=1\ Xi/n = (2 + 4 + \ldots + 2 + 1)/10 = 2.2$

b) S^2=∑n i=1(Xi − X)^2/(n − 1) = (2 − 2.2)^2 + (4 − 2.2)^2 + . . . + (2 − 2.2)^2 + (1 − 2.2)^2/(10−1) = 4.4
c) The estimate is X = 2.2
d) Estimated standard error of X is S/√ n = √ 4.4/10 = 0.66

CAIIB Paper 1 (ABM) Module A Unit 8: Linear Programming

Introduction

Linear Programming refers to several related mathematical techniques that are used to allocate limited resources among competing demands in an optimal way. For obtaining the optimal solution the problems should be structured into a particular format. It has been found that linear programming has many useful applications to financial decisions. The type of problems should have linear constraints and the decision maker must be trying to maximise some linear objective function.

In this chapter we will discuss graphical and '**simplex' methods**.

Model

Let us assume that the selling prices, production and marketing costs are known for each of the 'n' products. The firm also has to operate under certain economic, financial and physical constraints. Some examples of resource and marketing constraints:

- Bank may stipulate certain working capital requirements.
- Market may not absorb the whole output.
- Capacity constraints.
- Labour availability.
- Raw materials availability.

These constraints can be used to formulate the problem. The question is how to attain maximum profit minimum loss or minimum cost or time in the given circumstances? Maximum or minimum value can be obtained by forming and solving Linear Programming Problem.

Thus, Linear Programming Problem is a method by which a function (profit, loss, time, cost, etc.) can be maximised or minimised (optimised) with respect to some conditions. The function which has to be maximised or minimised (optimised)) is called objective function and the conditions are called constraints. The variables related to a linear programming problem whose values are to be determined are called Decision variables.

Under what conditions a Linear Programming problem can be formulated?

- As the name implies all equations are linear – This implies proportionality. For example, if it takes 4 persons to produce one unit, then we require 12 persons to produce 3 units.
- The constraints are known and deterministic. That is, the probabilities of occurrence are presumed to be 1.0.
- Most important rule is that all these variables should have non-negative values.
- Finally, decision variables are also divisible.

Graphic Approach

Let us illustrate the graphic approach with simple numerical two-decision variables. (3 variables require 3-D graphing). This gives a quick insight into the nature of L.P.

Let firm A produce radios and television sets.

Each radio costs Rs. 500 in wages and Rs. 500 in materials.

Each television set costs Rs. 2,500 in wages and Rs. 1,500 in materials.

The firm pays the labour and material expenses in cash.

The price of a radio is Rs. 2,000 and the price of a television is Rs. 6,000.

As there is a strong consumer demand, the firm is able to sell as many units as it produces at prevailing prices.

The firm also gives one period credit to consumers. The firm has 10 hours of machine time and 4 hours of assembly time per day.

The production of radio requires 3 hours of machine time and 1 hour of assembly time. The production of television requires 1 hour of machine time and 1 hour of assembly time.

The firm has Rs. 12,000 as cash balance (liquidity to pay for labour and materials). Now, given the financial and capacity constraints, how many radios and televisions should the firm produce in period 1, to maximise its profits?

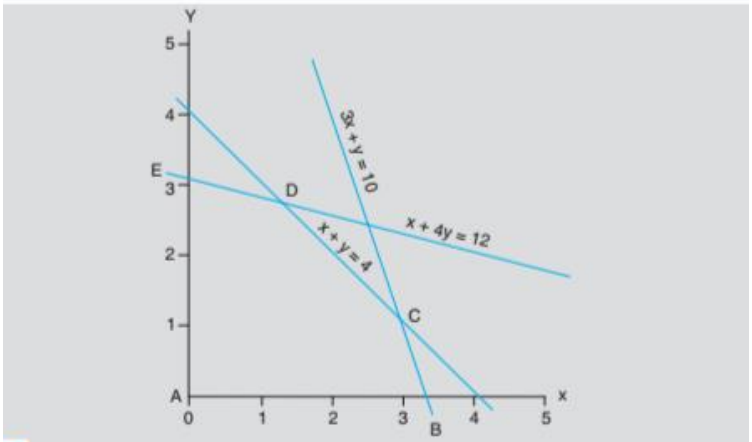
Let x and y be respectively, the units of radios and television sets produced in period 1. Then the constraints are:

(a) (capacity constraint machine time) 3x + y £ 10

(b) (capacity constraint assembly time) x + y £ 4

(c) (financial constraint) 1000x + 4000y £ 12,000 ® same as x + 4y £ 12

(d) (non-negativity) x ³ 0; y ³ 0;



1 Graphic Determination of Optimal Solution

(e) Objective function: Maximise Profit = 1,000x + 2,000y Now, let us draw the graph.

Line 1	$x + y = 4$				
Data	x	0	2	4	
	y	4	2	0	
Line 2	$3x + y = 10$				
Data	x	3	2	1	0
	y	1	4	7	10
Line 3	$x + 4y = 12$				
Data	x	0	4	8	12
	y	3	2	1	0

We have plotted the above three constraints in the graph. Find all the combinations of x and y, which satisfy the constraint and plot the points for all 3 lines. The graph is in the 1st quadrant. This satisfies the non-negativity condition.

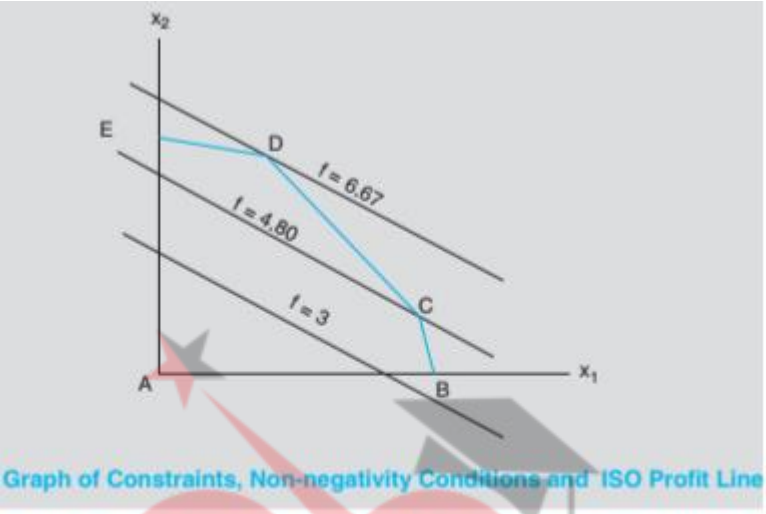
- All points on or below (inside) the line satisfy, x + y £ 4.
- All points on or below the line 3x + y £ 10, satisfy the machine time constraint.
- All points on or below the line x + 4y £ 12, satisfy financial constraint.

Even though all constraints are listed separately, they should be satisfied simultaneously. When these restrictions are placed one on top of the other, we obtain a common area, which in this case is shaped like a pentagon. (say ABCDE). Every point in this pentagon satisfies the constraints. This area is referred to as a set of feasible solutions.

Now, our objective is not to pick any feasible solution.

Although x = y = 0 is also a feasible solution, the profit will be zero.

This means no production of either radio or television. We are not seeking such a solution. So, our objective is to pick that feasible solution (that particular combination of x and y), from the set of feasible solutions, which maximises profit.



Graph of Constraints, Non-negativity Conditions and ISO Profit Line

Simplex Method

Another method of solving linear programming is Simplex Method. This method is a standard technique in linear programming for solving an optimisation (maximisation or minimisation) problem, typically one involving an objective function and several constraints expressed as inequalities. With computer programmes, spread sheets available, it is possible to use this method effectively and solve equations with as many as 10–12 variables.

Let us take the following problem to use Simplex Method.

Problem

A company manufactures cricket bats and chess sets. Each cricket bat gives a profit of Rs. 2 and chess set gives a profit of Rs. 4.

	Workshop 1 (hr)	Workshop 2 (hr)	Workshop 3 (hr)
Availability (Per day)	120	72	10
Cricket Bat	4	2	0
Chess Set	6	6	1

If the company wants to maximise the profit, how many cricket bats and chess sets should be produced per day?

Step 1 Solution: Formulate the problem.

Let the production be 'B' bats and 'C' chess sets.

(a) Objective function: Maximise Z = 2B + 4C

(b) 4B + 6C £ 120 (Workshop 1)

(c) 2B + 6C £ 72 (Workshop 2)

(d) 1C £ 10

(e) B, C ³ 0

We now change this to standard LP format.

In the standard LP form, all the constraints are converted into equations with the help of slack variables. Also make sure that these equations have non-negative right hand side. For example, 4B + 6C £ 120 is changed to 4B + 6C + m = 120 Here m is called a slack variable. It takes non-negative values. In fact all the variables in these equations take non-negative values.

The standard LP format is as follows:

(a) Objective function Maximise Z = 2B + 4C + 0 m + 0 n + 0 p

(b) 4B + 6C + 1m = 120 (Workshop 1)

(c) 2B + 6C +1n = 72 (Workshop 2)

(d) 1C + 1p = 10

(e) B, C ³ 0; m, n, p ³ 0 where m, n, p are the slack variables.

Z equation is also written as Z – 2B – 4C – 0 m – 0 n – 0 p = 0. Now, make a tableau as follows

Basic variables	Z	B	C	m	n	p	Solution
Z	1	–2	–4	0	0	0	0
m	0	4	6	1	0	0	120
n	0	2	6	0	1	0	72
p	0	0	1	0	0	1	10

This tableau gives the coefficients of the variables Z, B, C, m, n, p in the four equations written in the standard LP format, starting with the Z-equation. This tableau is a convenient way of setting up the information. This gives,

1. The variables which are in the solution at that point. (Z, m, n, p)
2. Profit associated with the solution. (0 when $B = 0$, $C = 0$)
3. The variable that will add most to profit, if brought into the solution. This is indicated by the variable which has most negative coefficient in the Z-row.

Here the most negative coefficient is –4 for C. So C is called the entering variable. Next, we need to rewrite the tableau by replacing one of the basic slack variables by C.

To decide which current basic variable is to be replaced by C, we concentrate on the C-column and the solutions column. Take the ratio of the corresponding entries in w these columns. Look at the following table:

C	Solution	Ratio
–4	0	
6	130	120/6 = 20
6	72	72/6 =12
1	10	10/1 = 10

Then we choose the smallest positive value in the ratio column, which is 10. The slack variable corresponding to this is p.

Thus we decide to replace p by C. Look at the tableau below which is a reproduction of the previous one. We have highlighted the column under C, and the p-row, which is called the pivot row. The intersection of the highlighted row and column is called the pivot entry, which is 1 here.

Basic variables	Z	B	C	m	n	P	Solution
Z	1	–2	–4	0	0	0	0
m	0	4	6	1	0	0	120
n	0	2	6	0	1	0	72
p	0	0	1	0	0	1	10

Now we form a new tableau where

(i) new pivot row = (current pivot row)/(pivot entry)

(ii) all other rows = current row – (pivot entry)*(new pivot entry)

Basic variables	Z	B	C	m	n	P	Solution
Z	1	–2	0	0	0	4	40
m	0	4	0	1	0	–6	60
n	0	2	0	0	1	–6	12
p	0	0	1	0	0	1	10

So we have completed one iteration of the problem.

Simulation

Simulation is appropriate to situations where size and/or complexity of the problem make the use of other techniques difficult or impossible. For example, queuing problems have been extensively studied through simulation. Some types of inventory problems, layout and maintenance problems also can be studied through simulation. Simulation can be used with traditional statistical and management techniques.

Simulation is useful in training managers and workers in how the real system operates, in demonstrating the effects of changes in system variables and real-time control. Simulation is extensively used in driving lessons. The person who learns driving is made to face the real road situations (traffic jams and other problems) during learning, so that serious accidents can be avoided. Simulation is commonly used in financial world such forex, investment and risk management areas.

Application of simulation methods:

- Air Traffic control queuing
- Aircraft maintenance scheduling
- Assembly line scheduling
- Inventory reorder design
- Railroad operations
- Facility layout
- Risk modeling in finance area.
- Foreign exchange market
- Stock market

Example:

The owner of an outlet wishes to evaluate his daily ordering policy. His current rule is order the demand of the previous day. But he has started thinking recently that he should follow better methods to decide the quantum of order.

He purchases milk at Rs 12 and sells at Rs 16. He orders his requirement at the end of the day and gets the milk in the morning. From past experience, the vendor assessed that his demand is between 30 and 80 liters per day.

He also kept a record of relative frequency of the quantity demanded during the last 10 days. Now he thinks of a new ordering rule — mean of quantity sold in the last 10 days.

He maintained the sales in a tabular form. The table has two columns. The first column shows the Demand and the second one shows the Relative frequency, that is, in the selected period of 10 days, how many times such demand occurred.

Demand per day in Litres	Relative Frequency
35	1/ 10, that is, only one day, out of ten days, demand of 35 litres occured
45	3/10, that is, only three days, out of ten days, demand of 45 litres occured
55	2 /10, that is, only two days, out of ten days, demand of 55 litres occured
65	3/10, that is, only three days, out of ten days, demand of 65 litres occured
75	1/10, that is, only one day, out of ten days, demand of 75 litres occured

He settles for the ordering rule

[(35 × 0.1) + (45 × 0.3) + (55 × 0.2) + (65 × 0.3) + (75 × 0.1)] = 55 litres.

So we have 2 rules: Old rule and New rule. Representing mathematically,

Old rule = quantity demanded on previous day is equal to D (n - 1).

New rule = Mean of the past 10 days is equal to 55

Now let us compare these orders in terms of profits.

Profit 'P' is equal to (Sold Quantity × selling price (p)) - (Ordered quantity × cost price (c)).

Assume that the unsold milk packets are thrown away as they are perishable. Now to prepare for simulation, we have to develop a method for demand generation. Let us use the probability distribution of demand and random numbers to generate a demand for the next 20 days.

Now arrange the chance process to generate occurrences in the system.

Demand Per Day	Relative Frequency	Probability	Random Number Interval
35	1/ 10	0.1	00 to 09
45	3/10	0.3	10 to 39
55	2/10	0.2	40 to 59
65	3/ 10	0.3	60 to 89
75	1/ 10	0.1	90 to 99

With the above table and random numbers, we develop the demand for 20 days.

Step 1: Choose a random number.

Step 2: Find the random number interval associated with the random number.

Step 3: Read the daily demand corresponding to the random number interval.

Step 4: Assume D = 55 litres for day 0

Step 5: Calculate the quantity sold. Quantity sold will be lesser of the demand D or Quantity ordered Q1 (or Q2)

Step 6: Profit = (Sold quantity × selling price) - (Ordered quantity × cost price).

Selling Price is Rs 16 per litre and cost price is Rs 12 per litre

Step 7: Do all steps for 20 days to simulate.

Day	RN (random number	D (demand related to respective random number interval)	Q1 (quantity ordered based on demand of previous day)	S1 (quantity sold under old method) (lesser of D and Q1)	PR-1 (rupees) profit under old method (16 into S1)- (12 into Q1)	Q2 (quantity ordered) (mean of quantity sold in last ten days)	S2 (quantity sold under new method) (lesser of D and Q2)	PR-2 (rupees) profit under old method (16 into S1)- (12 into Q1)
0		55						
1	6	35	55	35	–100	55	35	–100
2	39	45	35	35	140	55	45	60
3	89	65	45	45	180	55	55	220
4	61	65	65	65	260	55	55	220
5	99	75	65	65	260	55	55	220
6	95	75	75	75	300	55	55	220
7	55	55	75	55	-20	55	55	220
8	35	45	55	45	60	55	45	60
9	57	55	45	45	180	55	55	220

10	59	55	55	55	220	55	55	220
11	30	45	55	45	60	55	45	60
12	81	65	45	45	180	55	55	220
13	2	35	65	35	-220	55	35	-100
14	18	45	35	35	140	55	45	60
15	87	65	45	45	180	55	55	220
16	68	65	65	65	260	55	55	220
17	28	45	65	45	-60	55	45	60
18	44	55	45	45	180	55	55	220
19	80	65	55	55	220	55	55	220
20	84	65	65	65	260	55	55	220
Total		1120	1110	1000	2680	1100	1010	2960
Average		56	55.5	50	134	55	50.5	148

We now see that the average demand according to simulation is 56 litres, Average sales is 50 litres, according to old method; and 50.5 litre according to new method. Average order is 55.50 litres under old method, whereas 55 hires under new method.

Thus you would find that profitability improves under the new method.

Simulation Methodology

START	Key factors
DEFINE PROBLEM	Define objectives and variables
CONSTRUCT THE SIMULATION MODEL	Specification of variables, parameters, decision rules, probability distribution and time incrementing procedure — (fixed or variable)
SPECIFY VALUES OF PARAMETERS & VARIABLES RUN THE SIMULATION	Determine starting conditions and run length
EVALUATE RESULTS	Determine statistical tests
PROPOSE NEW EXPERIMENT	Compare with other information
Stop	

Advantages

Simulation is desirable when experiments on the real system

- Would disrupt ongoing activities;
- Would be too costly to undertake;
- Require many observations over an extended period of time;
- Do not permit exact replication of events; and
- Do not permit control over key variables.

Simulation is preferable when a mathematical model

- is not available to handle the problem;
- is too complex and arduous to solve;
- is beyond the capability of available personnel; and
- is not robust enough to provide information on all factors of interest.

Disadvantages

- Time consuming.
- Requires computer experience and expertise on the part of the user.
- Impossibility of quantifying and difficulty of casting complex problems in a format may cause difficulties; but simulations can be made to run under any type of assumption and these flaws can be overlooked.
- In spite of widespread applications, there are very few principles to guide the user in making decisions on what to include in the model and the length and number of simulation runs. This will be more like an art than science. The user has to use his intuitive judgments.

Ambitiousbaba.com
Paid Course