तेजस्वि नावधीतमस्तु
ISO 9001:2008 & 14001:2004

FAIRFIELD
Institute of Management & Technology
Managed by 'The Fairfield Foundation'
(Affiliated to GGSIP University, New Delhi)

BBA (B&I)

GGS Indraprastha University

BBA (B&I) 106 Quantitative Techniques &

Operations Research in Management

L-4 T-0 Credits –4

Objectives:

The objective of this paper is to develop student's familiarity with the basic concept and tools in statistics and operations research. These techniques assist specially in resolving complex problems serve as a valuable guide to the decision makers.

Course Contents:

Unit I No. of Hrs.:-12

Statistics: Definition, Importance & Limitation. Collection of data and formation of frequency distribution. Graphic presentation of frequency distribution – graphics, Bars, Histogram, Diagrammatic. Measures of central tendency – mean, median and mode, partition values – quartiles, deciles and percentiles. Measures of variation – range, IQR, quartile, deciles and percentiles. Measures of variation – range, IQR, quartile deviation and standard deviation and Lorenz Curve.

Unit II No. of Hrs.:-10

Correlation Analysis: Correlation Coefficient; Assumptions of correlation analysis; coefficients of determination and correlation; measurement of correlation- Karl Person's Methods; Spearman's rank correlation; concurrent deviation the correlation coefficient; Pitfalls and limitations associated with regression and correlation analysis; real world application using IT tools..

No. of Hrs.:-12

Unit III

Linear Programming: Concept a assumptions usage in business decision making linear programming problem: formulation, methods of solving: graphical and simplex, problems with mixed constraints: duality; concept, significance, usage & application in business decision making**.**

Unit IV No. of Hrs.:-12Transportation, assignment problems & Game Theory: General structure of transportation problem, solution procedure for transportation problem, methods for finding initial solution, test for optimality. Maximization transportation on problem, transportation problem. Assignment problem approach of the assignment model, solution methods of assignment problem, maximization in an assignment, unbalanced assignment problem, restriction on assignment.

Unit-I (Measures of Central Tendency & Dispersion)

Definition

Collecting data can be easy and fun. But sometimes it can be hard to tell other people about what you have found. That's why we use statistics. Two kinds of statistics are frequently used to describe data. They are measures of central tendency and dispersion. These are often called descriptive statistics because they can help you describe your data.

Mean median and mode

These are all measures of central tendency. They help summarize a bunch of scores with a single number. Suppose you want to describe a bunch of data that you collected to a friend for a particular variable like height of students in your class. One way would be to read each height you recorded to your friend. Your friend would listen to all of the heights and then come to a conclusion about how tall students generally are in your class But this would take too much time. Especially if you are in a class of 200 or 300 students! Another way to communicate with your friend would be to use measures of central tendency like the mean, median and mode. They help you summarize bunches of numbers with one or just a few numbers. They make telling people about your data easy.

Range, variance and standard deviation

These are all measures of dispersion. These help you to know the spread of scores within a bunch of scores. Are the scores really close together or are they really far apart? For example, if you were describing the heights of students in your class to a friend, they might want to know how much the heights vary. Are all the men about 5 feet 11 inches within a few centimeters or so? Or is there a lot of variation where some men are 5 feet and others are 6 foot 5 inches? Measures of dispersion like the range, variance and standard deviation tell you about the spread of scores in a data set. Like central tendency, they help you summarize a bunch of numbers with one or just a few numbers.

Importance & Limitation

In any research, enormous data is collected and, to describe it meaningfully, one needs to summarise the same. The bulkiness of the data can be reduced by organizing it into a frequency table or histogram.[

The following popper user interface control may not be accessible. Tab to the next button to revert the control to an accessible version.

Destroy user interface control Frequency distribution organizes the heap of data into a few meaningful categories. Collected data can also be summarized as a single index/value, which represents the entire data. These measures may also help in the comparison of data.

CENTRAL TENDENCY

Central tendency is defined as "the statistical measure that identifies a single value as representative of an entire distribution."[

The following popper user interface control may not be accessible. Tab to the next button to revert the control to an accessible version.

Destroy user interface control It aims to provide an accurate description of the entire data. It is the single value that is most typical/ representative of the collected data. The term "number crunching" is used to illustrate this aspect of data description. The mean, median and mode are the three commonly used measures of central tendency.

MEAN

Mean is the most commonly used measure of central tendency. There are different types of mean, viz. arithmetic mean, weighted mean, geometric mean (GM) and harmonic mean (HM). If mentioned without an adjective (as mean), it generally refers to the arithmetic mean.

Arithmetic mean

Arithmetic mean (or, simply, "mean") is nothing but the average. It is computed by adding all the values in the data set divided by the number of observations in it. If we have the raw data, mean is given by the formula

$$\text{Mean} \quad \bar{X} = \frac{\Sigma X}{n}$$

Where, ∑ (the uppercase Greek letter sigma), *X* refers to summation, refers to the individual value and n is the number of observations in the sample (sample size). The research articles published in journals do not provide raw data and, in such a situation, the readers can compute the mean by calculating it from the frequency distribution (if provided).

$$\text{Mean} \quad \bar{X} = \frac{\Sigma f X}{n}$$

Where, *f* is the frequency and *X* is the midpoint of the class interval and n is the number of observations.[

The following popper user interface control may not be accessible. Tab to the next button to revert the control to an accessible version.

Destroy user interface control The standard statistical notations (in relation to measures of central tendency) are mentioned. Readers are cautioned that the mean calculated from the frequency

distribution is not exactly the same as that calculated from the raw data. It approaches the mean calculated from the raw data as the number of intervals increase.[

The following popper user interface control may not be accessible. Tab to the next button to revert the control to an accessible version.

<u>Destroy user interface control</u>

Standard statistical notation

ADVANTAGES

The mean uses every value in the data and hence is a good representative of the data. The irony in this is that most of the times this value never appears in the raw data.

Repeated samples drawn from the same population tend to have similar means. The mean is therefore the measure of central tendency that best resists the fluctuation between different samples.[

The following popper user interface control may not be accessible. Tab to the next button to revert the control to an accessible version.

<u>Destroy user interface control</u>

It is closely related to standard deviation, the most common measure of dispersion.

DISADVANTAGES

The important disadvantage of mean is that it is sensitive to extreme values/outliers, especially when the sample size is small.[

The following popper user interface control may not be accessible. Tab to the next button to revert the control to an accessible version.

<u>Destroy user interface control</u> Therefore, it is not an appropriate measure of central tendency for skewed distribution.

The following popper user interface control may not be accessible. Tab to the next button to revert the control to an accessible version.

<u>Destroy user interface control</u>

Mean cannot be calculated for nominal or non nominal ordinal data. Even though mean can be calculated for numerical ordinal data, many times it does not give a meaningful value, e.g. stage of cancer.

Weighted mean

Weighted mean is calculated when certain values in a data set are more important than the others.[

The following popper user interface control may not be accessible. Tab to the next button to revert the control to an accessible version.

<u>Destroy user interface control</u> A weight w_i is attached to each of the values x_i to reflect this importance.

$$\text{Weighted mean} = \frac{\Sigma wx}{\Sigma w}$$

For example, When weighted mean is used to represent the average duration of stay by a patient in a hospital, the total number of cases presenting to each ward is taken as the weight.

Geometric Mean

It is defined as the arithmetic mean of the values taken on a log scale. It is also expressed as the n^{th} root of the product of an observation.

$$\text{Geometric mean (GM)} = \sqrt[n]{(x_1)(x_2)\dots(x_n)}$$
$$\text{Log (GM)} = \frac{\Sigma(\log x)}{n}$$

GM is an appropriate measure when values change exponentially and in case of skewed distribution that can be made symmetrical by a log transformation. GM is more commonly used in microbiological and serological research. One important disadvantage of GM is that it cannot be used if any of the values are zero or negative.

Harmonic mean

It is the reciprocal of the arithmetic mean of the observations

<u>Measures of dispersion</u>

The measure of central tendency, as discussed in the previous chapter tells us only about the characteristics of a particular series. They do not describe anything on the observations or data entirely. In other words, measures of central tendency do not tell anything about the variations that exist in the data of a particular series. To make the concept, let discuss an example. It was found by using formula of mean that the average depth of a river is 6 feet. One cannot confidently enter into the river because in some places the depth may be 12 feet or it may have 3 feet. Thus this type of interpretation by using the measures of central tendency sometimes proves to be useless. Hence the measure of central tendency alone to measure the characteristics of a series of observations is not sufficient to draw a valid conclusion. With the central value one must know

as to how the data is distributed. Different sets of data may have the same measures of central tendency but differ greatly in terms of variation. For this knowledge of central value is not enough to appreciate the nature of distribution of values. Thus there is the requirement of some additional measures along with the measures of central tendency which will describe the spread of the entire set of values along with the central value. One such measure is popularly called as dispersion or variation. The study of dispersion will enables us to know whether a series is homogeneous (where all the observations remains around the central value) or the observations is heterogeneous (there will be variations in the observations around the central value like 1, 50, 20, 28 etc., where the central value is 33). Hence it can be said that a measure of dispersion describes the spread or scattering of the individual values of a series around its central value.

Experts opine different opinion on why the variations in a distribution are so important to consider? Following are some views on validity of the measure of dispersion:

1. **Measures of variation provide the researchers some additional information about the behavior of the series along with the measures of central tendency. With this information one can judge the reliability of the value that is derived by using the measure of central tendency. If the data of the series are widely dispersed, the central location is less representatives of the data as a whole. On the other hand, when the data of a series is less dispersed, the central location is more representative to the entire series. In other wards, a high degree of variation would mean little uniformity whereas a low degree of variation would mean greater uniformity.**
2. **When the data of a series are widely dispersed, it creates practical problems in executing data. Measure of dispersion helps in understanding and tackling the widely dispersed data.**
3. **It facilitates to determine the nature and cause of variation in order to control the variation itself.**
4. **Measures of variation enable comparison to be made of two or more series with regard to their variability.**

DEFINATION:

Following are some definitions defined by different experts on measures of dispersion. L.R. Connor defines measures of dispersion as �dispersion is the measure extended to which individual items vary�. Similarly, Brookes and Dick opines it as �dispersion or spread is the degree of the scatter or the variation of the variables about a central value�. Robert H. Wessel defines it as �measures which indicate the spread of the values are called measures of dispersion�. From all these definition it is clear that dispersion measures more or less describes the spread or scattering of the individual values of a series around its central value.

METHODS OF MEASURING DISPERSION:

Dispersion of a series of data can be calculated by using following four widely used methods

Dispersion when measured on basis of the difference between two extreme values selected from a series of data. The two well known measures are

1. **The Range**
2. **The Inter-quartile Range or Quartile Deviation**

Dispersion when measured on basis of average deviation from some measure of central tendency. The well known measures are

1. **The Mean/average deviation**
2. **The Standard Deviation and**
3. **The Coefficient of variation and**
4. **The Gini coefficient and the Lorenz curve**

<u>Collection of data and formation of frequency distribution</u>

In statistics, a frequency distribution is an arrangement of the values that one or more variables take in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

Univariate frequency tables

Rank	Degree of agreement	Number
1	Strongly agree	20
2	Agree somewhat	30
3	Not sure	20
4	Disagree somewhat	15
5	Strongly disagree	15

A different tabulation scheme aggregates values into bins such that each bin encompasses a range of values. For example, the heights of the students in a class could be organized into the following frequency table.

Height range	Number of students	Cumulative number
25	25	
5.0–5.5 feet	35	60
5.5–6 feet	20	80
6.0–6.5 feet	20	100

A frequency distribution shows us a summarized grouping of data divided into mutually exclusive classes and the number of occurrences in a class. It is a way of showing unorganized data e.g. to show results of an election, income of people for a certain region, sales of a product within a certain period, student loan amounts of graduates, etc. Some of the graphs that can be used with frequency distributions are histograms, line graphs, bar charts and pie charts. Frequency distributions are used for both qualitative and quantitative data.

Joint frequency distributions

Bivariate joint frequency distributions are often presented as (two-way) contingency tables:

Two-way contingency table with marginal frequencies

	Dance	Sports	TV	Total
Men	2	10	8	20
Women	16	6	8	30
Total	18	16	16	50

The *total* row and *total* column report the marginal frequencies or marginal distribution, while the body of the table reports the joint frequencies.

Applications

Managing and operating on frequency tabulated data is much simpler than operation on raw data. There are simple algorithms to calculate median, mean, standard deviation etc. from these tables.

Statistical hypothesis testing is founded on the assessment of differences and similarities between frequency distributions. This assessment involves measures of central tendency or averages, such as the mean and median, and measures of variability or statistical dispersion, such as the standard deviation or variance.

A frequency distribution is said to be skewed when its mean and median are different. The kurtosis of a frequency distribution is the concentration of scores at the mean, or how peaked the

distribution appears if depicted graphically—for example, in a histogram. If the distribution is more peaked than the normal distribution it is said to be leptokurtic; if less peaked it is said to be platykurtic.

Letter frequency distributions are also used in frequency analysis to crack codes and are referred to the relative frequency of letters in different languages.

<u>Graphic presentation of frequency distribution—graphics, bars, histogram, Diagrammatic</u>

Graphical representation is done of the data available this being a very important step of statistical analysis. We will be discussing the organization of data. The word 'Data' is plural for 'datum'; datum means facts. Statistically the term is used for numerical facts such as measures of height, weight and scores on achievement and intelligence tests.
Tests, experiments and surveys in education and psychology provide us valuable data, mostly in the shape of numerical scores. For understanding data available and deriving meaning and useful conclusion, the data have to be organized or arranged in some systematic way. This can be done by following ways:

1. Statistical tables

2. Rank order

3. Frequency distribution

<u>Statistical tables</u>
The data are tabulated or arranged into rows and columns of different heading. Such tables can list original raw scores as well as the percentages, means, standard deviations and so on. Example -

Table for group mean and S.D. of anxiety test of dancers and non dancers

Group	Mean	Standard deviation	N
Dancers	22.66	6.018	15
Non dancers	27.66	8.741	15

Rules for constructing tables:

1. Title of the table should be simple, concise and unambiguous. As a rule, it should appear on the table.

2. The table should be suitably divided into columns and rows according to the nature of data and purpose. These columns and rows should be arranged in a logical order to facilitate comparison.

3. The heading of each columns or row should be as brief as possible. Two or more columns or rows with similar headings may be grouped under a common heading to avoid repetition and we may have subheadings or captions.

4. Subtotal for each separate classification and a general total for all combined classes are to be given. These totals should be given at the bottom or right of the concerned items.

5. The units in which the data are given must invariably be mentioned.

6. Necessary footnotes should be providing essential explanation of the points to ambiguous representation of the tabulated data must be given at the bottom of the table.

7. The sources from where the data have been received should be given at the end of the table.

8. In tabulating long columns of figures, space should be left after every five or ten rows.

9. If the numbers tabulated have more than three significant figure, the digit should be grouped in threes. For ex.- 4394756 as 4 394 756.

10. For all purposes and by all means, the table should be as simple as possible so that it may be studied by the readers with minimum possible strain and create a clear picture and interpretations of the data.

<u>Frequency Distribution</u>

The organization of the data according to rank order does not help us to summarize a series of raw scores. It also does not tell us the frequency of the raw scores. In frequency distribution we group the data into an arbitrarily chosen groups or classes. It is also seen that how many times a particular score or group of scores occurs in the given data. This is known as the frequency distribution of numerical data.

<u>Construction of Frequency distribution table</u>

<u>Finding the range:</u>
First of all the range of the series to be grouped is found. it is done by subtracting the lowest score from the highest. In the present problem the range of the distribution is 46-12, i.e. 34

<u>Determining class interval:</u>
After finding range we find class interval represented by '*i'.* The formula for I is
i = Range/ no. of class interval desired
I = 34/8
I = 4.25
We decide to take class interval to be 5.

<u>Writing the contents of the frequency distribution table:</u>

Writing the classes of the distribution.
In the first column we write the classes of distribution. First of all the lowest class is settled and afterwards other subsequent classes are written down. In this case we take 10-14 as the lowest class, then we have higher classes as 15-19, 20-24, and so on up to 45-49.
Tallying the scores into proper classes.

The scores given are tallied into proper classes in the second column then the tallies are counted against each class to obtain the frequency of the class. Example-

Class interval	Tallies	Frequency for Non-dancers
45-49	I	1
40-44	0	0
35-39	II	2
30-34	III	3
25-29	IIII	4

20-24	II	2
15-19	II	2
10-14	I	1

Total frequency (N) = 15

GRAPHICAL REPRESENTATION OF DATA

The statistical data may be presented in a more attractive form appealing to the eye with the help of some graphic aids, i.e. Pictures and graphs. Such presentation carries a lot of communication power. A mere glimpse of thee picture and graphs may enable the viewer to have an immediate and meaningful grasp of the large amount of data.
Ungrouped data may be represented through a bar diagram, pie diagram, pictograph and line graph.

METHOD FOR CONSTRUCTING A HISTOGRAM

1. The scores in the form of actual class limits as 19.5-24.5, 24.5-29.5 and so on are taken as examples in the construction of a histogram rather than written class limits as 20-24, 25-30.

2. It is customary to take two extra intervals of classes one below and above the grouped intervals.

3. Now we take the actual lower limits of all the class intervals and try to plot them on the x axis. The lower limit of the lowest class interval is taken at the intersecting point of x axis and y axis.

4. Frequencies of the distribution are plotted on the y axis.

5. Each class interval with its specific frequency is represented by separate rectangle. The base of each rectangle is the width of the class interval. And the height is representative of the frequency of that class or interval.

6. Care should be taken to select the appropriate units of representation along the x and y axis. Both the axis and the y axis must not be too short or too long.

METHOD FOR CONSTRUTING A FREQUENCY POLYGON

1. As in histogram two extra class interval is taken, one above and other below the given class interval.

2. The mid-points of the class interval is calculated.

3. The midpoint is calculated along the x axis and the corresponding frequencies are plotted along the y axis.

4. The various points given by the plotting are joined by lines to give frequency polygon.

DIFFERENCE BETWEEN HISTOGRAM AND FREQUENCY POLYGON

Histogram is a bar graph while frequency polygon is a line graph. Frequency polygon is more useful and practical. In frequency polygon it is easy to know the trends of the distribution; we are unable to do so in histogram. Histogram gives a very clear and accurate picture of the relative proportion of the frequency from interval to interval.

METHOD FOR CONSTRUTING A CUMULATIVE FREQUENCY GRAPH

1. First of all we calculate the actual upper and lower limits of the class intervals i.e. if the class interval is 20-24 then upper limit is 24.5 and the lower limit is 19.5.

2. We must know select a suitable scale as per the range of the class interval and plot the actual upper limits on the x axis and the respective cumulative frequency on y axis.

3. All the plotted points are then joined by successive straight lines resulting a line graph.

4. To plot the origin of the x axis an extra class interval is taken with cumulative frequency zero is taken.

Measures of central tendency- mean, median and mode

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

Mean (Arithmetic)

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our Types of Variable guide for data types). The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x_1, x_2, ..., x_n, the sample mean, usually denoted by $\bar{x}$ (pronounced x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

This formula is usually written in a slightly different manner using the Greek capitol letter, Σ, pronounced "sigma", which means "sum of...":

$$\bar{x} = \frac{\sum x}{n}$$

You may have noticed that the above formula refers to the sample mean. So, why have we called it a sample mean? This is because, in statistics, samples and populations have very different meanings and these differences are very important, even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lower case letter "mu", denoted as µ:

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

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set. However, one of its important properties is that it minimizes error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

When not to use the mean

The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

The mean salary for these ten staff is $30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the $12k to 18k range. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation.

Another time when we usually prefer the median over the mean (or mode) is when our data is skewed (i.e., the frequency distribution for our data is skewed). If we consider the normal distribution - as this is the most frequently assessed in statistics - when the data is perfectly normal, the mean, median and mode are identical. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from the typical value. However, the median best retains this position and is not as strongly influenced by the skewed values. This is explained in more detail in the skewed distribution section later in this guide.

Median

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	56	56	65	87	89	92

Our median mark is the middle mark - in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

65	55	89	56	35	14	56	55	87	45

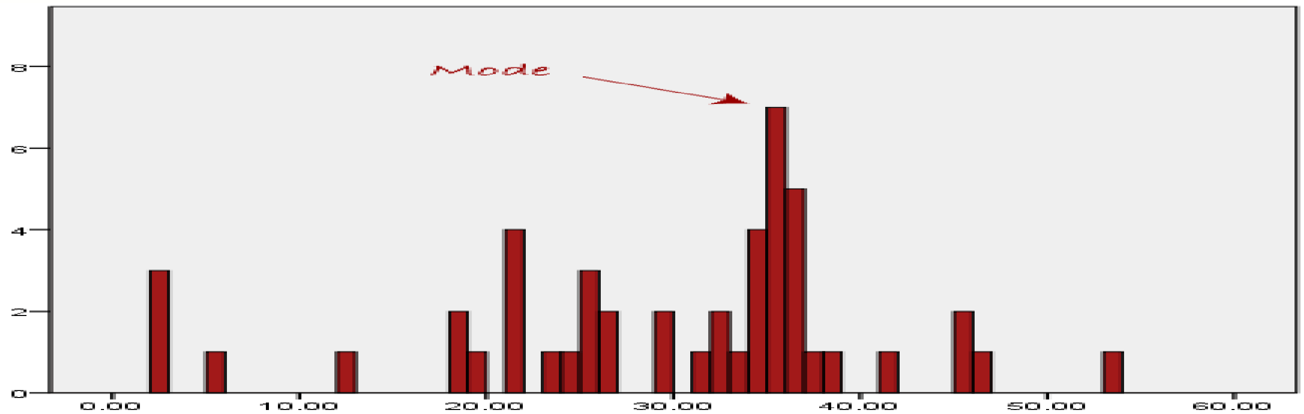
We again rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	56	56	65	87	89	92

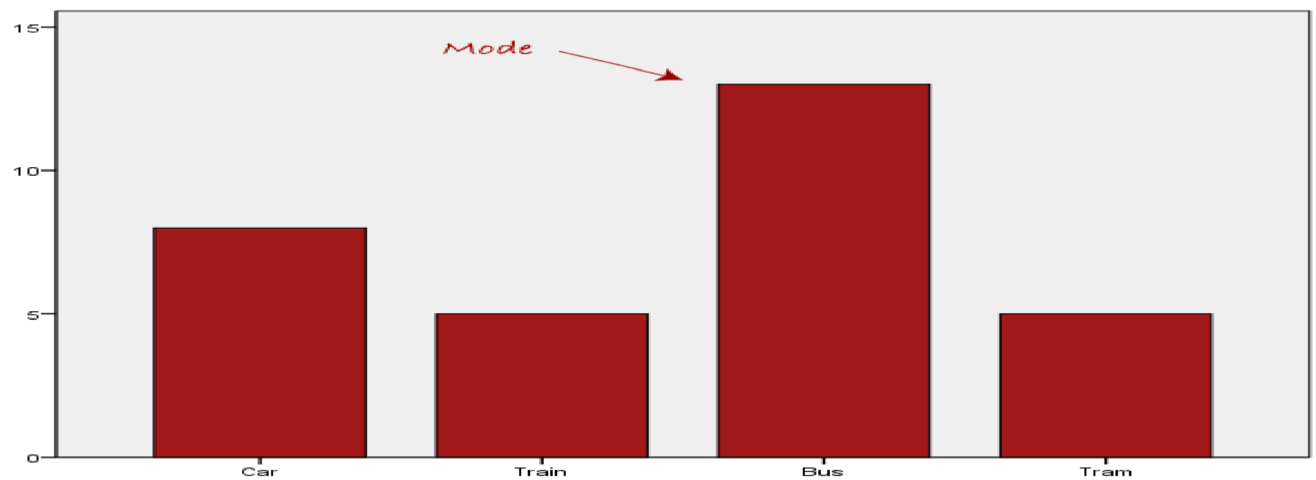
Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

Mode

The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. An example of a mode is presented below:



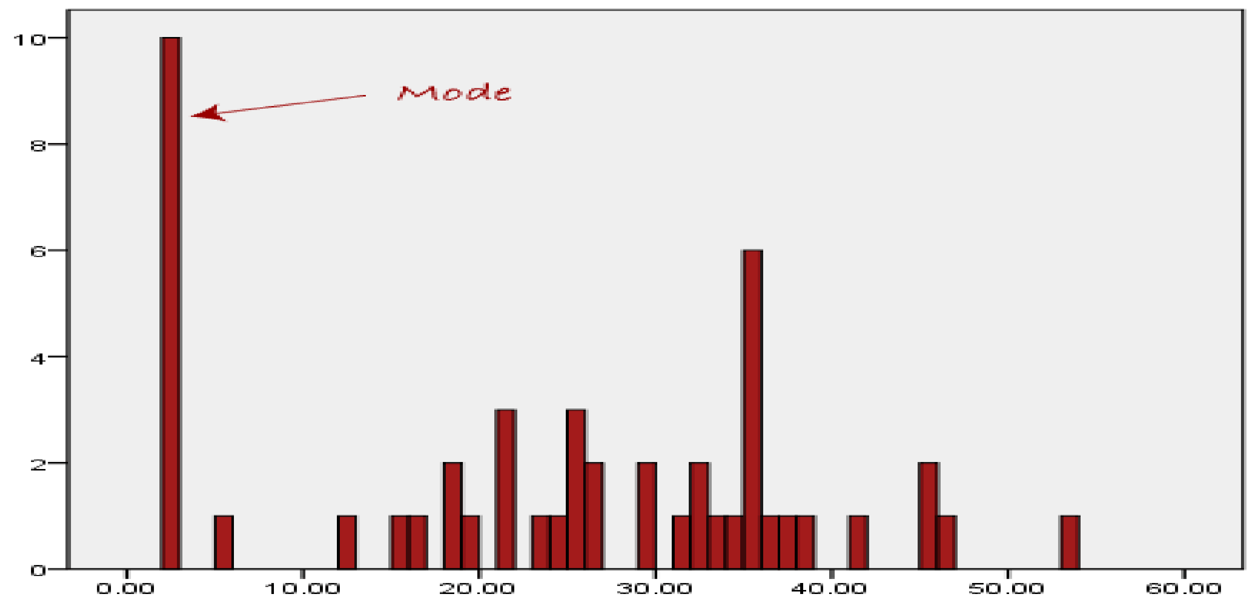
Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated below:



We can see above that the most common form of transport, in this particular data set, is the bus. However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest frequency, such as below:

We are now stuck as to which mode best describes the central tendency of the data. This is particularly problematic when we have continuous data because we are more likely not to have any one value that is more frequent than the other. For example, consider measuring 30 peoples' weight (to the nearest 0.1 kg). How likely is it that we will find two or more people with exactly the same weight (e.g., 67.4 kg)? The answer, is probably very unlikely - many people might be close, but with such a small sample (30 people) and a large range of possible weights, you are unlikely to find two people with exactly the same weight; that is, to the nearest 0.1 kg. This is why the mode is very rarely used with continuous data.

Another problem with the mode is that it will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the data set, as depicted in the diagram below:



In the above diagram the mode has a value of 2. We can clearly see, however, that the mode is not representative of the data, which is mostly concentrated around the 20 to 30 value range. To use the mode to describe the central tendency of this data set would be misleading.

<u>Partition values-quartiles, deciles and percentiles & Measures of variation- range, IQR, quartile, deciles and percentiles</u>

The computation of these partition values is done exactly in the same manner as the computation of the Median. In a series of individual observations and in a discrete series, the values of the lower (Q_1) and the upper (Q_9) quartiles would be the value of (N+1) /4) th and (3(N+1) / 4 items respectively.
The values of the deciles in such series would be as follows.

$$D_1 = \text{value of } \left(\frac{N+1}{10}\right)^{th} \text{ item}$$

$$D_2 = \text{value of } \frac{2(N+1)^{th}}{10} \text{ item}$$

$$D_8 = \text{value of } \frac{8\,(N+1)^{th}}{10} \text{ item and so on.}$$

$$P_1 = \text{value of} \left(\frac{N+1}{100}\right)^{th} \text{ item}$$

$$P_{40} = \text{value of } \frac{40\,(N+1)^{th}}{100} \text{ item}$$

$$P_{99} = \text{value of } \frac{99\,(N+1)^{th}}{100} \text{ item and so on.}$$

The values of the percentiles would be In continuous series in the calculation of quartiles, deciles and percentiles (N+1)/4 , (N+1)/10 and (N+1)/100 would be replaced by N/4 ,N/10 and N/100 respectively. The values would have to be interpolated here as was done in case of the computation of Median.

<u>UNIT-II (Correlation & Regression)</u>

<u>Correlation Coefficient</u>

Correlation coefficient may refer to:

- **Pearson product-moment correlation coefficient, also known as *r*, *R*, or Pearson's *r*, a measure of the strength and direction of the linear relationship between two variables that is defined as the (sample) covariance of the variables divided by the product of their (sample) standard deviations.**
- **Infraclass correlation, a descriptive statistic that can be used when quantitative measurements are made on units that are organized into groups; describes how strongly units in the same group resemble each other.**
- **Rank correlation, the study of relationships between rankings of different variables or different rankings of the same variable**
    - **Spearman's rank correlation coefficient, a measure of how well the relationship between two variables can be described by a monotonic function**
    - **Kendall tau rank correlation coefficient, a measure of the portion of ranks that match between two data sets.**
    - **Goodman and Kruskal's gamma, a measure of the strength of association of the cross tabulated data when both variables are measured at the ordinal level.**

Assumptions of correlation analysis

An inspection of a scatter plot can give an impression of whether two variables are related and the direction of their relationship. But it alone is not sufficient to determine whether there is an association between two variables. The relationship depicted in the scatter plot needs to be described qualitatively. Descriptive statistics that express the degree of relation between two variables are called correlation coefficients. A commonly employed correlation coefficient for scores at the interval or ratio level of measurement is the Pearson product-moment correlation coefficient, or Pearson's r.

The Pearson's r is a descriptive statistic that describes the linear relationship between two or more variables, each measured for the same collection of individuals. An "individual" is not necessarily a person: it might be an automobile, a place, a family, a university, etc. For example, the two variables might be the heights of a man and of his son; there, the "individual" is the pair (father, son). Such pairs of measurements are called bivariate data. Observations of two or more variables per individual in general are called multivariate data. As with any sample of scores, the sample is drawn from a larger population of scores.

The test for significance of Pearson's r assumes that a particular variable, X and another variable, Y, form a bivariate normal distribution in the population. A bivariate normal distribution possesses the following characteristics:

- The distribution of the X scores is normally distributed in the population sampled.
- The distribution of the Y scores is normally distributed in the population sampled.
- For each X score, the distribution of Y scores in the population is normal.
- For each Y score, the distribution of Y scores in the population is normal.

Coefficients of determination & correlation

In statistics, the coefficient of determination denoted R^2 and pronounced R squared, indicates how well data points fit a line or curve. It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model.[1]

There are several different definitions of R^2 which are only sometimes equivalent. One class of such cases includes that of linear regression. In this case, if an intercept is included then R^2 is simply the square of the sample correlation coefficient between the outcomes and their predicted values; in the case of simple linear regression, it is the squared correlation between the outcomes and the values of the single repressor being used for prediction. If an intercept is included and the number of explicators is more than one, R^2 is the square of the coefficient of multiple correlation. In such cases, the coefficient of determination ranges from 0 to 1. Important cases where the computational definition of R^2 can yield negative values, depending on the definition used, arise where the predictions which are being compared to the corresponding outcomes have not been derived from a model-fitting procedure using those data, and where linear regression is conducted without including an intercept. Additionally, negative values of R^2 may occur when fitting non-linear functions to data.[2] In cases where negative values arise, the mean of the data provides a better fit to the outcomes than do the fitted function values, according to this particular criterion.

Measurement of correlation- Karl Pearson's Methods

In statistics, the Pearson product-moment correlation coefficient (r) is a common measure of the correlation between two variables X and Y. When measured in a population the Pearson Product Moment correlation is designated by the Greek letter rho (?). When computed in a sample, it is designated by the letter "r" and is sometimes called "Pearson's r." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between variables. A correlation of -1 means that there is a perfect negative linear relationship between variables. A correlation of 0 means there is no linear relationship between the two variables. Correlations are rarely if ever 0, 1, or -1. If you get a certain outcome it could indicate whether correlations were negative or positive.

Mathematical Formula:--

: The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson.

The mathematical formula for computing r is:

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

Where:

N = number of pairs of scores
Σxy = sum of the products of paired scores
Σx = sum of x scores
Σy = sum of y scores
Σx^2 = sum of squared x scores
Σy^2 = sum of squared y scores

Spearman's rank correlation

There are two methods to calculate Spearman's rank-order correlation depending on whether: (1) your data does not have tied ranks or (2) your data has tied ranks. The formula for when there are no tied ranks is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i = difference in paired ranks and *n* = number of cases. The formula to use when there are tied ranks is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Where *i* = paired score.

<u>Concurrent deviation the correlation coefficient</u>

Sometimes it is desired to study the correlation between two series in a very casual manner and in such cases no particular attention is needed so far as precision is concerned. In such cases it is enough to calculate the coefficient of concurrent deviations. I n this method correlation is calculated between the direction of deviations and not their magnitudes. As such only the direction of deviations is taken into account in the calculation of this coefficient, and their magnitude is ignored.

To calculate the coefficient of concurrent deviations, the deviations are not calculated from any average or by the method of moving averages but only their direction from the previous period, are noted down. The formula for the calculation of coefficient of concurrent deviations is given below:—

Coefficient of concurrent deviations or

rc = ±√(± ((2c-n)/n)

Where rc = stands for the coefficient of concurrent deviations, for the number of pairs of concurrent deviations and n for the number of pairs of deviations. The value of this coefficient of correlation also varies between 1. The plus and minus signs given in the formula should be Carefully noted. If the value of (2c-n)/n) is negative its square root±± 2c-nn

Where cannot be calculated and so a minus sign is placed before the sign of the root so that the square root may be calculated and the minus sign may be kept before the value of the coefficient of correlation.Thus the steps in the calculation of this coefficient are:

(1) Find out the direction of change of x—variable. It means find out whether the second figure in the series is more than the first figure, if it is so the direction is. If the second figure is less than the first the direction is —. If both the figures are equal the direction is 0. Similarly find the direction of the third figure from the second, of the fourth figure from the third and so on. It is denoted by (dx).

(2) Find out the direction of change of y-variable in the same manner as discussed above. It is denoted by (dy).

(3) multiply dx and dy and determine the value of c , which would be the number of positive products of dxdy (—x—) or (X+).

(4) Use the formula given above to obtain the value of the Coefficient or r.

<u>Pitfalls and limitations associated with regression & correlation analysis: real world application using IT tools</u>

We are all familiar with the disparaging quotes about statistics (including "There are three kinds of lies: lies, damned lies, and statistics", attributed to either Mark Twain or Disraeli, depending on whom you ask), and it's no secret that many people harbor a vague distrust of statistics as commonly used. Why should this be the case? It may be assumed that those of us at this conference take our work seriously and value the fruits of our efforts. So, are all those people just paranoid about statistics, or are we as statisticians really kidding ourselves as to our importance in the cosmic scheme of things?

It may be helpful to consider some aspects of statistical thought which might lead many people to be distrustful of it. First of all, statistics requires the ability to consider things from a probabilistic perspective, employing quantitative technical concepts such as "confidence", "reliability", "significance". This is in contrast to the way non-mathematicians often cast problems: logical, concrete, often dichotomous conceptualizations are the norm: right or wrong, large or small, this or that.

Additionally, many non-mathematicians hold quantitative data in a sort of awe. They have been lead to believe that numbers are, or at least should be, unquestionably correct. Consider the sort of math problems people are exposed to in secondary school, and even in introductory college math courses: there is a clearly defined method for finding the answer, and that answer is the only acceptable one. It comes, then, as a shock that different research studies can produce very different, often contradictory results. If the statistical methods used are really supposed to represent reality, how can it be that different studies produce different results? In order to resolve this paradox, many naive observers conclude that statistics must not really provide reliable (in the nontechnical sense) indicators of reality after all. And, the logic goes, if statistics aren't "right", they must be "wrong". It is easy to see how even intelligent, well-educated people can become cynical if they don't understand the subtleties of statistical reasoning and analysis.

Now, I'm not going to say much about this "public relations crisis" directly, but it does provide a motivation for examining the way we practice our trade. The best thing we can do, in the long run, is make sure we're using our tools properly, and that our conclusions are warranted. I will present some of the most frequent misuses and abuses of statistical methods, and how to avoid or remedy them. Of course, these issues will be familiar to most statisticians; however, they are the sorts of things that can get easily overlooked when the pressure is on to produce results and meet deadlines. If this workshop helps you to apply the basics of statistical reasoning to improve the quality of your product, it will have served its purpose.

We can consider three broad classes of statistical pitfalls. The first involves sources of bias. These are conditions or circumstances which affect the external validity of statistical results. The second category is errors in methodology, which can lead to inaccurate or invalid results. The third class of problems concerns interpretation of results, or how statistical results are applied (or misapplied) to real world issues.

Sources of Bias

The core value of statistical methodology is its ability to assist one in making inferences about a large group (a population) based on observations of a smaller subset of that group (a sample). In order for this to work correctly, a couple of things have to be true: the sample must be similar to the target population in all relevant aspects; and certain aspects of the measured variables must conform to assumptions which underlie the statistical procedures to be applied.

Representative sampling. **This is one of the most fundamental tenets of inferential statistics: the observed sample must be representative of the target population in order for inferences to be valid. Of course, the problem comes in applying this principle to real situations. The ideal scenario would be where the sample is chosen by selecting members of the population at random, with each member having an equal probability of being selected for the sample. Barring this, one usually tries to be sure that the sample "parallels" the population with respect to certain key characteristics which are thought to be important to the investigation at hand, as with a stratified sampling procedure.**

While this may be feasible for certain manufacturing processes, it is much more problematic for studying people. For instance, consider the construction of a job applicant screening instrument: the population about which you want to know something is the pool of all possible job applicants. You surely won't have access to the entire population--you only have access to a certain number of applicants who apply within a certain period of time. So you must hope that the group you happen to pick isn't somehow different from the target population. An example of a problematic sample would be if the instrument were developed during an economic recession; it is reasonable to assume that people applying for jobs during a recession might be different as a group from those applying during a period of economic growth (even if one can't specify exactly what those differences might be). In this case, you'd want to exercise caution when using the instrument during better economic times.

There are also ways to account for, or "control", differences between groups statistically, as with the inclusion of covariates in a linear model. Unfortunately, as Levin (1985) points out, there are problems with this approach, too. One can never be sure one has accounted for all the important variables, and inclusion of such controls depends on certain assumptions which may or may not be satisfied in a given situation (see below for more on assumptions).

Statistical assumptions. **The validity of a statistical procedure depends on certain assumptions it makes about various aspects of the problem. For instance, well-known linear methods such as**

analysis of variance (ANOVA) depends on the assumption of normality and independence. The first of these is probably the lesser concern, since there is evidence that the most common ANOVA designs are relatively insensitive to moderate violations of the normality assumption (see Kirk, 1982). Unfortunately, this offers an almost irresistible temptation to ignore *any* non-normality, no matter how bad the situation is. The robustness of statistical techniques only goes so far--"robustness" is not a license to ignore the assumption. If the distributions are non-normal, try to figure out why; if it's due to a measurement artifact (e.g. a floor or ceiling effect), try to develop a better measurement device (if possible). Another possible method for dealing with unusual distributions is to apply a transformation. However, this has dangers as well; an ill-considered transformation can do more harm than good in terms of interpretability of results.

The assumption regarding independence of observations is more troublesome, both because it underlies nearly all of the most commonly used statistical procedures, and because it is so frequently violated in practice. Observations which are linked in some way--parts manufactured on the same machine, students in the same classroom, consumers at the same mall--all may show some dependencies. Therefore, if you apply some statistical test across students in different classrooms, say to assess the relationship between different textbook types and test scores, you're introducing bias into your results. This occurs because, in our example, the kids in the class presumably interact with each other, chat, talk about the new books they're using, and so influence each other's responses to the test. This will cause the results of your statistical test (e.g. correlations or p-values) to be inaccurate.

One way to try to get around this is to aggregate cases to the higher level, e.g. use classrooms as the unit of analysis, rather than students. Unfortunately this requires sacrificing a lot of statistical power, making a Type II error more likely. Happily, methods have been developed recently which allow simultaneous modeling of data which is hierarchically organized (as in our example with students nested within classrooms). One of the papers presented at this conference (Christiansen & Morris) introduces these methods. Additionally, interested readers are referred for good overviews of these hierarchical models.

Errors in methodology

There are a number of ways that statistical techniques can be misapplied to problems in the real world. Three of the most common hazards are designing experiments with insufficient power, ignoring measurement error, and performing multiple comparisons.

Statistical Power. **This topic has become quite in vogue lately, at least in the academic community; indeed, some federal funding agencies seem to consider any research proposal incomplete unless it contains a comprehensive power analysis. This graph will help illustrate the concept of power in an experiment. In the figure, the vertical dotted line represents the point-null hypothesis, and the solid vertical line represents a criterion of significance, i.e. the point at which you claim a difference is significant.**

Unit-III (Linear Programming & Queuing)

Concept a assumptions usage in business decision making linear programming problem

If you've only got limited resources at your disposal, then it's helpful to calculate how best to maximize those resources - whether that's time, money, or space.

Let's say, for example, that you have 50 square feet of office space to use for storage. Your budget is $200, and there are a variety of cabinet types and sizes from which to choose. How do you optimize the space you have available, and stay within the allotted budget?

Or suppose you have three delivery trucks, and 10 drop-off points. How do you plan the most efficient route and schedule for these trucks?

Or consider that you manufacture three products using the same basic raw materials. However, as each product uses different amounts of material, some are more expensive to produce than others. A few of the materials are perishable, and need to be used quickly. How much of each product should you manufacture to minimize your cost? And which combination produces the least waste?

Questions like these may seem very complex. With so many variables and constraints to take into consideration, how do you decide what to do? The answer is to use linear programming.

Linear programming is a mathematical technique that determines the best way to use available resources. Managers use the process to help make decisions about the most efficient use of limited resources - like money, time, materials, and machinery.

<u>Methods of solving graphical & simplex</u>

Graphical and Simplex Methods of Linear Programming
The graphical method is the more popular method to use because they are easy to use and understand. Working with only a few variables at a time they allow operations managers to compare projected demand to existing capacity. The graphical method is a trial and error approach that can be easily done by a manager or even a clerical staff. Since it is trial and error though, it does not necessarily generate the optimal plan. One downside of this method though is that it can only be used with two variables at the maximum. The graphical method is broken down into the following five steps:

1) Determine the demand in each period.
2) Determine the capacity for regular time, over time, and subcontracting each period.
3) Find labor costs, hiring and labor costs, and inventory holding costs.
4) Consider company policy that may apply to the workers or to stock levels
5) Develop alternative plans and examine their total costs.

When a company has a LP problem with more than two variables it turns to the simplex method. This method can handle any number of variables as well as for certain give the optimal solution. In the simplex method we examine corner points in a methodical fashion until we arrive at the best solution which is either the highest profit or lowest cost.
LP is used in a wide variety of companies in numerous applications. Airline companies use it to schedule their flights to maximize profit. Another use is for firms to figure out how much of a certain product to manufacture in order to maximize total profits. It also is used by hospitals in order to figure out the most economic diet for patients. It is also a useful tool to figure out labor scheduling for a specific time period. Other applications include product mix planning, distribution networks, truck routing, financial portfolios, and corporate restructuring. All LP problems have four properties in common.

Duality: concept, significance, usage & application in business decision making

The term *structure* referred generally to "rules and resources" and more specifically to "the structuring properties allowing the 'binding' of time-space in social systems". These properties make it possible for similar social practices to exist across time and space and that lend them "systemic" form. Agents—groups or individuals—draw upon these structures to perform social actions through embedded memory, called *memory traces*. Memory traces are thus the vehicle through which social actions are carried out. Structure is also, however, the result of these social practices. Thus, Giddens conceives of the *duality of structure* as being:

The essential recursiveness of social life, as constituted in social practices: structure is both medium and outcome of reproduction of practices. Structure enters simultaneously into the constitution of the agent and social practices, and 'exists' in the generating moments of this constitution.

Giddens uses "the duality of structure" to emphasize structure's nature as both medium and outcome. Structures exist both internally within agents as memory traces that are the product of phenomenological and hermeneutic inheritance and externally as the manifestation of social actions. Similarly, social structures contain agents and/or are the product of past actions of agents. Giddens holds this duality, alongside "structure" and "system," as the core of structuration theory.[1]:p.17 His theory has been adopted by those with structuralism inclinations, but who wish to situate such structures in human practice rather than to reify them as an ideal type or material property. (This is different, for example, from actor–network theory which appears to grant a certain autonomy to technical artifacts.)

Social systems have patterns of social relation that change over time; the changing nature of space and time determines the interaction of social relations and therefore structure. Hitherto, social structures or models were either taken to be beyond the realm of human control—the positivistic approach—or posit that action creates them—the interpretive approach. The duality of structure emphasizes that they are different sides to the same central question of how social order is created.

Gregory McLennan suggested renaming this process "the duality of structure *and agency*", since both aspects are involved in using and producing social actions

Change

Sewell provided a useful summary that included one of the theory's less specified aspects: the question "Why are structural transformations possible?" He claimed that Giddens' overruled on rules and modified Giddens' argument by re-defining "resources" as the embodiment of cultural schemas. He argued that change arises from the multiplicity of structures, the *transposable* nature of schemas, the unpredictability of resource accumulation, the polysemy of resources and the intersection of structures.

The existence of multiple structures implies that the knowledgeable agents whose actions produce systems are capable of applying different schemas to contexts with differing resources, contrary to the conception of a universal habitués (learned dispositions, skills and ways of acting). He wrote that "Societies are based on practices that derived from many distinct structures, which exist at different levels, operate in different modalities, and are themselves based on widely varying types and quantities of resources. ...It is never true that all of them are homologous."

Originally from Bourdieu, *transposable* schemas can be "applied to a wide and not fully predictable range of cases outside the context in which they were initially learned." That capacity "is inherent in the knowledge of cultural schemas that characterizes all minimally competent members of society."

Agents may modify schemas even though their use does not predictably accumulate resources. For example, the effect of a joke is never quite certain, but a comedian may alter it based on the amount of laughter it garners regardless of this variability.

Agents may interpret a particular resource according to different schemas. E.g., a commander could attribute his wealth to military prowess, while others could see it as a blessing from the gods or a coincidental initial advantage.

Structures often overlap, confusing interpretation (e.g., the structure of capitalist society includes production from both private property and worker solidarity).

Technology

This theory was adapted and augmented by researchers interested in the relationship between technology and social structures (see Theories of technology), such as information technology in organizations. DeSanctis and Poole proposed an "adaptive structuration theory" with respect to the emergence and use of group decision support systems. In particular, they chose Giddens' notion of modalities to consider how technology is used with respect to its "spirit". "Appropriations" are the immediate, visible actions that reveal deeper structuration processes and are enacted with "moves". Appropriations may be fait

Wanda Orlikowski applied her critique of the duality of structure to hful or unfaithful, be instrumental and be used with various attitudes. technology: "The duality of technology identifies prior views of technology as either objective force or as socially constructed product–as a false dichotomy." She compared this to previous models (the technological imperative, strategic choice, and technology as a trigger) and considered the importance of meaning, power, norms, and interpretive flexibility. Orlikowski later replaced the notion of embedded properties for enactment (use). The "practice lens" shows how people enact structures which shape their use of technology that they employ in their practices. While Orlikowski's work focused on corporations, it is equally applicable to the technology cultures that have emerged in smaller community-based organizations, and can be adapted through the *gender sensitivity lens* in approaches to technology governance.

Workman, Ford and Allen rearticulated structuration theory as *structuration agency theory* for modeling socio-biologically inspired structuration in security software. Software agents join humans to engage in social actions of information exchange, giving and receiving instructions, responding to other agents, and pursuing goals individually or jointly.

Business

Pavlou and Majchrzak argued that research on business-to-business e-commerce portrayed technology as overly deterministic. The authors employed structuration theory to re-examine outcomes such as economic/business success as well as trust, coordination, innovation, and shared knowledge. They looked beyond technology into organizational structure and practices, and examined the effects on the structure of adapting to new technologies. The authors held that technology needs to be aligned and compatible with the existing "trustworthy" practices and

organizational and market structure. The authors recommended measuring long-term adaptations using ethnography, monitoring and other methods to observe causal relationships and generate better predictions.

Group communication

Poole, Seibold, and McPhee wrote that "group structuration theory, provides "a theory of group interaction commensurate with the complexities of the phenomenon.

The theory attempts to integrate macrosocial theories and individuals or small groups, as well as how to avoid the binary categorization of either "stable" or "emergent" groups.

Waldeck et al. concluded that the theory needs to better predict outcomes, rather than merely explaining them. *Decision rules* support decision-making, which produces a communication pattern that can be directly observable. Research has not yet examined the "rational" function of group communication and decision-making (i.e., how well it achieves goals), nor structural production or constraints. Researchers must empirically demonstrate the recursivity of action and structure, examine how structures stabilize and change over time due to group communication, and may want to integrate argumentation research.

Public relations

Falkheimer claimed that integrating structuration theory into public relations (PR) strategies could result in a less agency-driven business, return theoretical focus to the role of power structures in PR, and reject massive PR campaigns in favor of a more "holistic understanding of how PR may be used in local contexts both as a reproductive and [transformational] social instrument."[26]:p.103 Falkheimer portrayed PR as a method of communication and action whereby social systems emerge and reproduce. Structuration theory reinvigorates the study of space and time in PR theory. Applied structuration theory may emphasize community-based approaches, storytelling, rituals, and informal communication systems. Moreover, structuration theory integrates all organizational members in PR actions, integrating PR into all organizational levels rather than a separate office.

Queuing Models: Basic structure of queuing models

Everyone has experienced waiting in line, whether at a fast-food restaurant, on the phone for technical help, at the doctor's office or in the drive-through lane of a bank. Sometimes, it is a pleasant experience, but many times it can be extremely frustrating for both the customer and the store manager. Given the intensity of competition today, a customer waiting too long in line is potentially a lost customer. Understanding the nature of lines or "queues" and learning how to manage them is one of the most important areas in operations management.

Queues are basic to both external (customer-facing) and internal business processes, which include staffing, scheduling and inventory levels. For this reason, businesses often utilize queuing theory as a competitive advantage. Fortunately, Six Sigma professionals – through their

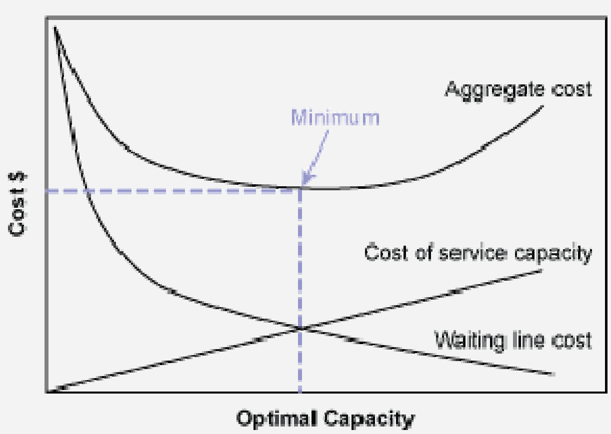
knowledge of probability distributions, process mapping and basic process improvement techniques – can help organizations design and implement robust queuing models to create this competitive advantage.

The Cost of Waiting in Line

The problem in virtually every queuing situation is a trade-off decision. The manager must weigh the added cost of providing more rapid service (i.e., more checkout counters, more production staff) against the inherent cost of waiting. For example, if employees are spending their time manually entering data, a business manager or process improvement expert could compare the cost of investing in bar-code scanners against the benefits of increased productivity. Likewise, if customers are walking away disgusted because of insufficient customer support personnel, the business could compare the cost of hiring more staff to the value of increased revenues and maintaining customer loyalty.

The relationship between service capacity and queuing cost can be expressed graphically (Figure 1). Initially, the cost of waiting in line is at a maximum when the organization is at minimal service capacity. As service capacity increases, there is a reduction in the number of customers in the line and in their wait times, which decreases queuing cost. The optimal total cost is found at the intersection between the service capacity and waiting line curves.

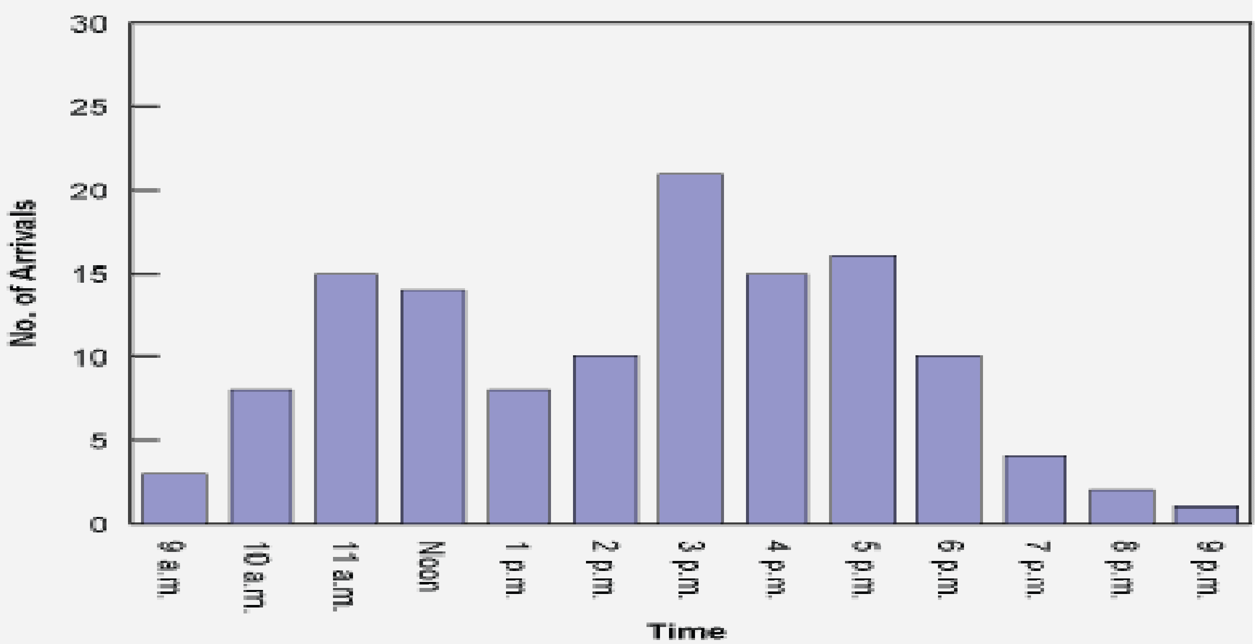
Figure 1: Service Capacity vs. Cost



Queuing Theory

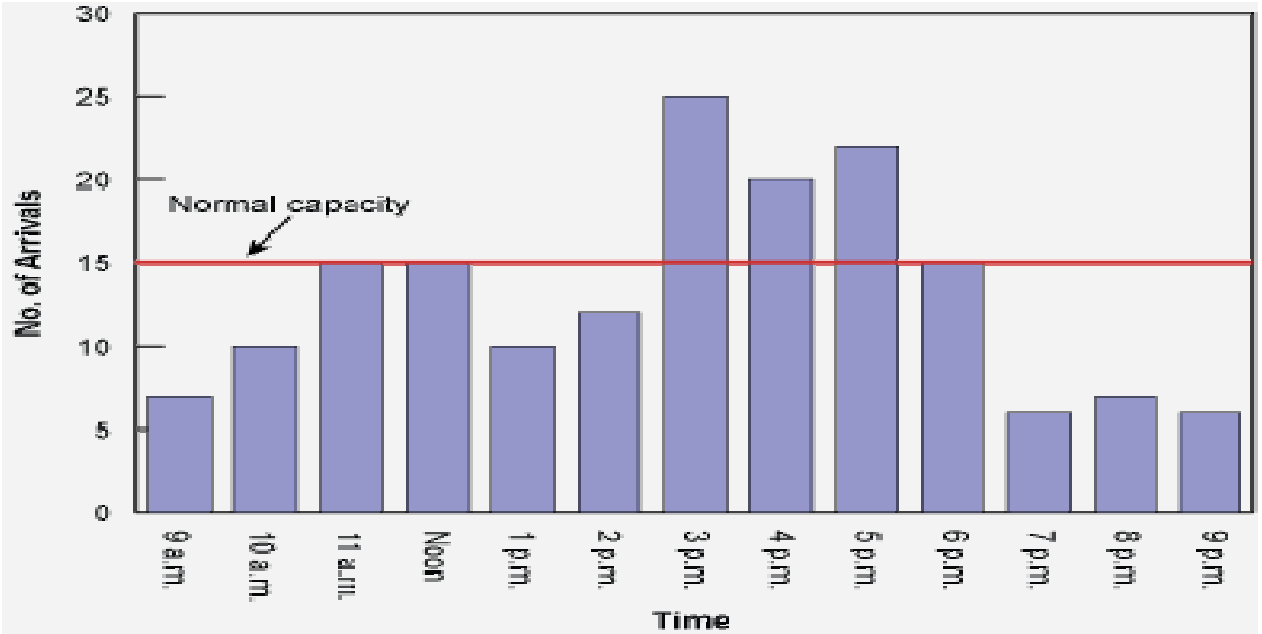
Queuing theory, the mathematical study of waiting in lines, is a branch of operations research because the results often are used when making business decisions about the resources needed to provide service. At its most basic level, queuing theory involves arrivals at a facility (i.e., computer store, pharmacy, bank) and service requirements of that facility (i.e., technicians, pharmacists, tellers). The number of arrivals generally fluctuates over the course of the hours that the facility is available for business (Figure 2).

Figure 2: Number of Arrivals at Facility



Customers demand varying degrees of service, some of which can exceed normal capacity (Figure 3). The store manager or business owner can exercise some control over arrivals. For example, the simplest arrival-control mechanism is the posting of business hours. Other common techniques include lowering prices on typically slow days to balance customer traffic throughout the week and establishing appointments with specific times for customers. The point is that queues are within the control of the system management and design.

Figure 3: Service Requirements



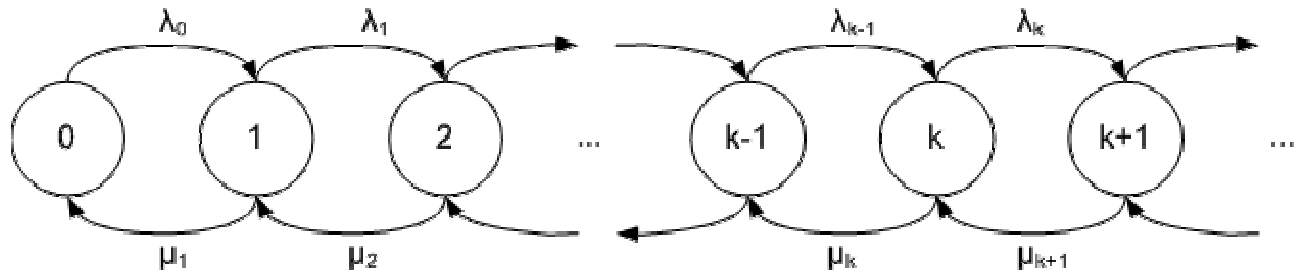
Queuing management consists of three major components:

1. **How customers arrive**
2. **How customers are serviced**
3. **The condition of the customer exiting the system**

Birth-Death queuing models and its steady state solution

The birth–death process is a special case of continuous-time Markov process where the state transitions are of only two types: "births" which increase the state variable by one and "deaths" which decrease the state by one. The model's name comes from a common application, the use of such models to represent the current size of a population where the transitions are literal births and deaths. Birth–death processes have many applications in demography, queueing theory, performance engineering, epidemiology or in biology. They may be used, for example to study the evolution of bacteria, the number of people with a disease within a population, or the number of customers in line at the supermarket.

When a birth occurs, the process goes from state *n* to *n* + 1. When a death occurs, the process goes from state *n* to state *n* – 1. The process is specified by birth rates $\{\lambda_i\}_{i=0\ldots\infty}$ **and death rates** $\{\mu_i\}_{i=1\ldots\infty}$.



<u>M/M/1 and M/M/C models with infinite/finite waiting space</u>

In queuing theory the birth–death process is the most fundamental example of a queuing model, the *M/M/C/K/∞/FIFO* (in complete Kendall's notation) queue. This is a queue with Poisson arrivals, drawn from an infinite population, and *C* servers with exponentially distributed service time with *K* places in the queue. Despite the assumption of an infinite population this model is a good model for various telecommunication systems.

M/M/1 queue

The M/M/1 is a single server queue with an infinite buffer size. In a non-random environment the birth–death process in queuing models tend to be long-term averages, so the average rate of arrival is given as λ and the average service time as $1/\mu$. The birth and death process is a M/M/1 queue when,

$$\lambda_i = \lambda \text{ and } \mu_i = \mu \text{ for all } i.$$

The difference equations for the probability that the system is in state *k* at time *t* are,

$$p_0'(t) = \mu_1 p_1(t) - \lambda_0 p_0(t)$$

$$p_k'(t) = \lambda_{k-1} p_{k-1}(t) + \mu_{k+1} p_{k+1}(t) - (\lambda_k + \mu_k) p_k(t)$$

M/M/c queue

The M/M/c is a multi-server queue with C servers and an infinite buffer. This differs from the M/M/1 queue only in the service time which now becomes

$$\mu_i = i\mu \text{ for } i \le C$$

and

$$\mu_i = C\mu \text{ for } i \ge C$$

with

$$\lambda_i = \lambda \text{ for all } i.$$

M/M/1/K queue

The M/M/1/K queue is a single server queue with a buffer of size K. This queue has applications in telecommunications, as well as in biology when a population has a capacity limit. In telecommunication we again use the parameters from the M/M/1 queue with,

$$\lambda_i = \lambda \text{ for } 0 \leq i < K$$
$$\lambda_i = 0 \text{ for } i \geq K$$
$$\mu_i = \mu \text{ for } 1 \leq i \leq K.$$

In biology, particularly the growth of bacteria, when the population is zero there is no ability to grow so,

$$\lambda_0 = 0.$$

Additionally if the capacity represents a limit where the population dies from over population,

$$\mu_K = 0.$$

The differential equations for the probability that the system is in state *k* at time *t* are,

$$p_0'(t) = \mu_1 p_1(t) - \lambda_0 p_0(t)$$
$$p_k'(t) = \lambda_{k-1} p_{k-1}(t) + \mu_{k+1} p_{k+1}(t) - (\lambda_k + \mu_k) p_k(t) \text{ for } k \leq K$$
$$p_k'(t) = 0 \text{ for } k > K$$

PERT & CPM

INTRODUCTION

Basically, CPM (Critical Path Method) and PERT (Programme Evaluation Review Technique) are project management techniques, which have been created out of the need of Western industrial and military establishments to plan, schedule and control complex projects.

Brief History of CPM/PERT

CPM/PERT or Network Analysis as the technique is sometimes called, developed along two parallel streams, one industrial and the other military.

CPM was the discovery of M.R.Walker of E.I.Du Pont de Nemours & Co. and J.E.Kelly of Remington Rand, circa 1957. The computation was designed for the UNIVAC-I computer. The first test was made in 1958, when CPM was applied to the construction of a new chemical plant. In

March 1959, the method was applied to a maintenance shut-down at the Du Pont works in Louisville, Kentucky. Unproductive time was reduced from 125 to 93 hours.

PERT was devised in 1958 for the POLARIS missile program by the Program Evaluation Branch of the Special Projects office of the U.S.Navy, helped by the Lockheed Missile Systems division and the Consultant firm of Booz-Allen & Hamilton. The calculations were so arranged so that they could be carried out on the IBM Naval Ordinance Research Computer (NORC) at Dahlgren, Virginia.

Planning, Scheduling & Control

Planning, Scheduling (or organising) and Control are considered to be basic Managerial functions, and CPM/PERT has been rightfully accorded due importance in the literature on Operations Research and Quantitative Analysis.

Far more than the technical benefits, it was found that PERT/CPM provided a focus around which managers could brain-storm and put their ideas together. It proved to be a great communication medium by which thinkers and planners at one level could communicate their ideas, their doubts and fears to another level. Most important, it became a useful tool for evaluating the performance of individuals and teams.

There are many variations of CPM/PERT which have been useful in planning costs, scheduling manpower and machine time. CPM/PERT can answer the following important questions:

How long will the entire project take to be completed? What are the risks involved?

Which are the critical activities or tasks in the project which could delay the entire project if they were not completed on time?

Is the project on schedule, behind schedule or ahead of schedule?

If the project has to be finished earlier than planned, what is the best way to do this at the least cost?

The Framework for PERT and CPM

Essentially, there are six steps which are common to both the techniques. The procedure is listed below:

1. Define the Project and all of it's significant activities or tasks. The Project (made up of several tasks) should have only a single start activity and a single finish activity.

2. Develop the relationships among the activities. Decide which activities must precede and which must follow others.

3. Draw the "Network" connecting all the activities. Each Activity should have unique event numbers. Dummy arrows are used where required to avoid giving the same numbering to two activities.

4. Assign time and/or cost estimates to each activity

5. **Compute the longest time path through the network. This is called the critical path.**

6. **Use the Network to help plan, schedule, monitor and control the project.**

The Key Concept used by CPM/PERT is that a small set of activities, which make up the longest path through the activity network control the entire project. If these "critical" activities could be identified and assigned to responsible persons, management resources could be optimally used by concentrating on the few activities which determine the fate of the entire project.

Non-critical activities can be replanned, rescheduled and resources for them can be reallocated flexibly, without affecting the whole project.

UNIT-IV (Transportation& Assignment Problem)

General structure of transportation problem, solution procedure for transportation problem, methods for finding initial solution, test for optimality, maximization of transportation problem, transportation problem

There is a type of linear programming problem that may be solved using a simplified version of the simplex technique called transportation method. Because of its major application in solving problems involving several product sources and several destinations of products, this type of problem is frequently called the transportation problem. It gets its name from its application to problems involving transporting products from several sources to several destinations. Although the formation can be used to represent more general assignment and scheduling problems as well as transportation and distribution problems. The two common objectives of such problems are either (1) minimize the cost of shipping *m* units to *n* destinations or (2) maximize the profit of shipping *m* units to *n* destinations.

Let us assume there are *m* sources supplying *n* destinations. Source capacities, destinations requirements and costs of material shipping from each source to each destination are given constantly. The transportation problem can be described using following linear programming mathematical model and usually it appears in a transportation tableau.

There are three general steps in solving transportation problems.

We will now discuss each one in the context of a simple example. Suppose one company has four factories supplying four warehouses and its management wants to determine the minimum-cost shipping schedule for its weekly output of chests. Factory supply, warehouse demands, and shipping costs per one chest (unit) are shown.

				Shipping Cost per Unit (in $)				
Factory	Supply	Warehouse	Demand	From	To E	To F	To G	To H
A	15	E	10	A	10	30	25	15
B	6	F	12	B	20	15	20	10
C	14	G	15	C	10	30	20	20
D	11	H	9	D	30	40	35	45

"Data for Transportation Problem"

At first, it is necessary to prepare an initial feasible solution, which may be done in several different ways; the only requirement is that the destination needs be met within the constraints of source supply.

The Transportation Matrix

The transportation matrix for this example appears in Table 7.2, where supply availability at each factory is shown in the far right column and the warehouse demands are shown in the bottom row. The unit shipping costs are shown in the small boxes within the cells at the initiation of solving all cells are empty). It is important at this step to make sure that the total supply availabilities and total demand requirements are equal. Often there is an excess supply or demand. In such situations, for the transportation method to work, a dummy warehouse or factory must be added. Procedurally, this involves inserting an extra row (for an additional factory) or an extra column (for an ad warehouse). The amount of supply or demand required by the"dummy" equals the difference between the row and column totals.

In this case there is:

Total factory supply … 51

Total warehouse requirements … 52

This involves inserting an extra row - an additional factory. The amount of supply by the dummy equals the difference between the row and column totals. In this case there is 52 – 51 = 1. The cost figures in each cell of the dummy row would be set at zero so any units sent there would not incur a transportation cost. Theoretically, this adjustment is equivalent to the simplex procedure of inserting a slack variable in a constraint inequality to convert it to an equation, and, as in the simplex, the cost of the dummy would be zero in the objective function.

To / From	E	F	G	H	Factory Supply
A	10	30	25	15	14
B	20	15	20	10	10
C	10	30	20	20	15
D	30	40	35	45	12
Dummy	0	0	0	0	1
Destination Requirements	10	15	12	15	52 / 52

Initial Feasible Solution

Initial allocation entails assigning numbers to cells to satisfy supply and demand constraints. Next we will discuss several methods for doing this: the Northwest-Corner method, Least-Cost method, and Vogel's approximation method (VAM).

Table shows a northwest-corner assignment. (Cell A-E was assigned first, A-F second, B-F third, and so forth.) Total cost : 10*10 + 30*4 + 15*10 + 30*1 + 20*12 + 20*2 + 45*12 + 0*1 = 1220($).

Inspection of Table indicates some high-cost cells were assigned and some low-cost cells bypassed by using the northwest-comer method. Indeed, this is to be expected since this method ignores costs in favor of following an easily programmable allocation algorithm.

Table shows a least-cost assignment. (Cell Dummy-E was assigned first, C-E second, B-H third, A-H fourth, and so on.) Total cost : 30*3 + 25*6 + 15*5 +10*10 + 10*9 + 20*6 + 40*12 + 0*1= 1105 ($).

Table shows the VAM assignments. (Cell Dummy-G was assigned first, B-F second, C-E third, A-H fourth, and so on.) Note that this starting solution is very close to the optimal solution obtained after making all possible improvements (see next chapter) to the starting solution obtained using the northwest-comer method. (See Table 7.3.) Total cost: 15*14 + 15*10 + 10*10 + 20*4 + 20*1 + 40*5 + 35*7 + 0*1 = 1005 ($).

To From	E	F	G	H	Factory Supply
A	10 10	30 4	25	15	14
B	20	15 10	20	10	10
C	10	30 1	20 12	20 2	15
D	30	40	35	45 12	12
Dummy	0	0	0	0 1	1
Destination Requirements	10	15	12	15	52 52

"Northwest – Corner Assignment"

From \ To	E	F	G	H	Factory Supply
A	10	30 3	25 6	15 5	14
B	20	15	20	10 10	10
C	10 9	30	20 6	20	15
D	30	40 12	35	45	12
Dummy	0 1	0	0	0	1
Destination Requirements	10	15	12	15	52 52

"Least - Cost Assignment"

From \ To	E	F	G	H	Factory Supply
A	10	30	25	15 14	14
B	20	15 10	20	10	10
C	10 10	30	20 4	20 1	15
D	30	40 5	35 7	45	12
Dummy	0	0	0 1	0	1
Destination Requirements	10	15	12	15	52 52

"VAM Assignment"

Develop Optimal Solution

To develop an optimal solution in a transportation problem involves evaluating each unused cell to determine whether a shift into it is advantageous from a total-cost stand point. If it is, the shift is made, and the process is repeated. When all cells have been evaluated and appropriate shifts made, the problem is solved. One approach to making this evaluation is the <u>Stepping stone method</u>.

The term stepping stone appeared in early descriptions of the method, in which unused cells were referred to as "water" and used cells as "stones"— from the analogy of walking on a path of stones half-submerged in water. The stepping stone method was applied to the VAM initial solution, as shown

Table shows the optimal solutions reached by the Stepping stone method. Such solution is very close to the solution found using VAM method.

To / From	E	F	G	H	Factory Supply
A	10	30	25	15 / 14	14
B	20	15 / 10	20	10	10
C	10 / 10	30	20 / 4	20	15
D	30	40 / 4	35 / 8	45	12
Dummy	0	0 / 1	0	0	1
Destination Requirements	10	15	12	15	52 / 52

"Optimal Matrix, With Minimum Transportation Cost of $1,000."

Alternate Optimal Solutions

When the evaluation of any empty cell yields the same cost as the existing allocation, an alternate optimal solution exists (see Stepping Stone Method – alternate solutions). Assume that all other cells are optimally assigned. In such cases, management has additional flexibility and can invoke nontransportation cost factors in deciding on a final shipping schedule.

To From	E	F	G	H	Factory Supply
A	10	30	25	15 14	14
B	20	15 9	20	10 1	10
C	10 10	30	20 5	20	15
D	30	40 5	35 7	45	12
Dummy	0	0	0 1	0	1
Destination Requirements	10	15	12	15	52 52

"Alternate Optimal Matrix for the Chest Transportation Problem, With Minimum Transportation Cost of $1,000.

Degeneracy

Degeneracy exists in a transportation problem when the number of filled cells is less than the number of rows plus the number of columns minus one (m + n - 1). Degeneracy may be observed either during the initial allocation when the first entry in a row or column satisfies both the row and column requirements or during the Stepping stone method application, when the added and subtracted values are equal. Degeneracy requires some adjustment in the matrix to evaluate the solution achieved. The form of this adjustment involves inserting some value in an empty cell so a closed path can be developed to evaluate other empty cells. This value may be thought of as an infinitely small amount, having no direct bearing on the cost of the solution.

Procedurally, the value (often denoted by the Greek letter epsilon, - ε) is used in exactly the same manner as a real number except that it may initially be placed in any empty cell, even though row and column requirements have been met by real numbers. A degenerate transportation problem showing a Northwest Corner initial allocation is presented in Table 7.8, where we can see that if ε were not assigned to the matrix, it would be impossible to evaluate several cells.

Once a ε has been inserted into the solution, it remains there until it is removed by subtraction or until a final solution is reached.

While the choice of where to put an ε is arbitrary, it saves time if it is placed where it may be used to evaluate as many cells as possible without being shifted.

Transportation Problem with a Maximization as a Criterion

A fictive corporation A has a contract to supply motors for all tractors produced by a fictive corporation B. Corporation B manufactures the tractors at four locations around Central Europe: Prague, Warsaw, Budapest and Vienna. Plans call for the following numbers of tractors to be produced at each location:

Prague 9 000

Warsaw 12 000

Budapest 9 000

Corporation A has three plants that can produce the motors. The plants and production capacities are

Hamburg 8 000

Munich 7 000

Leipzig 10 000

Dresden 5 000

Due to varying production and transportation costs, the profit earns on each motor depends on where they were produced and where they were shipped. The following transportation table gives the accounting department estimates of the euro profit per unit (motor).

Shipped to / Produced at	Prague	Warsaw	Budapest	Source Capacity
Hamburg	70	90	130	8 000
Munich	80	130	60	7 000
Leipzig	65	110	100	10 000
Dresden	95	80	35	5 000
Destination Capacity	9 000	12 000	9 000	30 000 / 30 000

"The Euro Profit Per One Shipped Motor"

Table shows a highest - profit assignment (Least Cost method modification). In contrast to the Least – Cost method it allocates as much as possible to the highest-cost cell. (Cell Hamburg - Budapest was assigned first, Munich - Warsaw second, Leipzig - Warsaw third, Leipzig – Budapest fourth, Dresden – Prague fifth and Leipzig – Prague sixth.) Total profit : 3 335 000 euro.

Shipped to / Produced at	Prague	Warsaw	Budapest	Source Capacity
Hamburg	70	90	130 8 000	8 000
Munich	80	130 7 000	60	7 000
Leipzig	65 4 000	110 5 000	100 1 000	10 000
Dresden	95 5 000	80	35	5 000
Destination Capacity	9 000	12 000	9 000	30 000 30 000

"Highest - Profit Assignment"

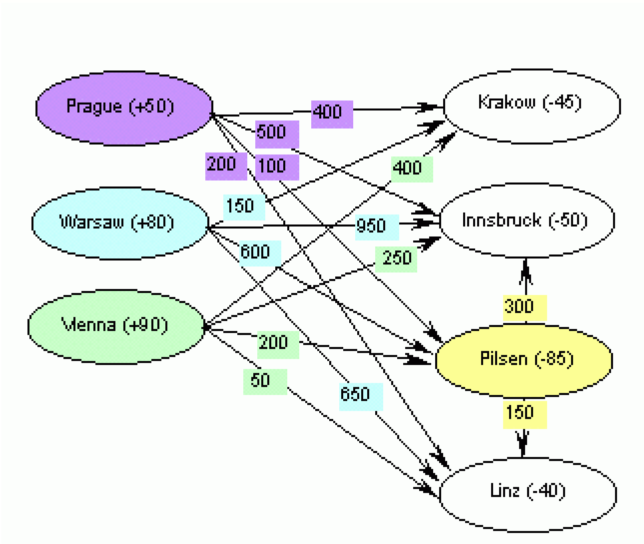
Applying the Stepping Stone method (modified for maximization purposes) to the initial solution we can see that no other transportation schedule can increase the profit and so the Highest – Profit initial allocation is also an optimal solution of this transportation problem.

The Transshipment Problem

The transshipment problem is similar to the transportation problem except that in the transshipment problem it is possible to both ship into and out of the same node (point). For the transportation problem, you can ship only from supply points to demand points. For the transshipment problem, you can ship from one supply point to another or from one demand point to another. Actually, designating nodes as supply points or demand points becomes confusing when you can ship both into and out of a node. You can make the designations clearer if you classify nodes by their net stock position-excess (+), shortage (-), or 0.

One reason to consider transshipping is that units can sometimes be shipped into one city at a very low cost and then transshipped to other cities. In some situations, this can be less expensive than direct shipment.
Let's consider the balanced transportation problem as an example.

Picture shows the net stock positions for the three warehouses and four customers. Say that it is possible to transship through Pilsen to both Innsbruck and Linz. The transportation cost from Pilsen to Innsbruck is 300 euro per unit, so it costs less to ship from Warsaw to Innsbruck by going through Pilsen. The direct cost is 950 euro, and the transshipping cost is 600 + 300 = 900 euro. Because the transportation cost is 300 euro from Pilsen to Innsbruck, the cost of transshipping from Prague through Pilsen to Innsbruck is 400 euro per unit. It is cheaper to transship from Prague through Pilsen than to ship directly from Prague to Innsbruck.



"Transshipment Example in the Form of a Network Model"

There are two possible conversions to a transportation model. In the <u>first conversion</u>, make each excess node a supply point and each shortage node a demand point. Then, find the cheapest method of shipping from surplus nodes to shortage nodes considering all transshipment possibilities. Let's perform the first conversion for the Picture 7.1 example. Because a transportation table Prague, Warsaw, and Vienna have excesses, they are the supply points. Because Krakow, Pilsen, Innsbruck, and Linz have shortages, they are the demand points. The cheapest cost from Warsaw to Innsbruck is 900 euro, transshipping through Pilsen. The cheapest cost from Prague to Innsbruck is 400 euro, transshipping through Pilsen too. The cheapest cost from all other supply points to demand points is obtained through direct shipment. Table 7.11 shows the balanced transportation table for this transshipment problem. For a simple transportation network, finding all of the cheapest routes from excess nodes to shortage nodes is easy. You can list all of the possible routes and select the cheapest. However, for a network with many nodes and arcs, listing all of the possible routes is difficult.

From \ To	Krakow	Innsbruck	Pilsen	Linz	Source Supply
Prague	400	400	100	200	50
Warsaw	150	900	600	650	80
Vienna	400	250	200	50	90
Destination Demands	45	50	85	40	220 / 220

"The Transshipment Problem After Conversion to a Transportation Model"

The second conversion of a transshipment problem to a transportation model doesn't require finding all of the cheapest routes from excess nodes table to shortage nodes. The second conversion requires more supply and demand nodes than the first conversion, because the points

where you can ship into and out of, occur in the converted transportation problem twice – first as a supply point and second as a demand point.

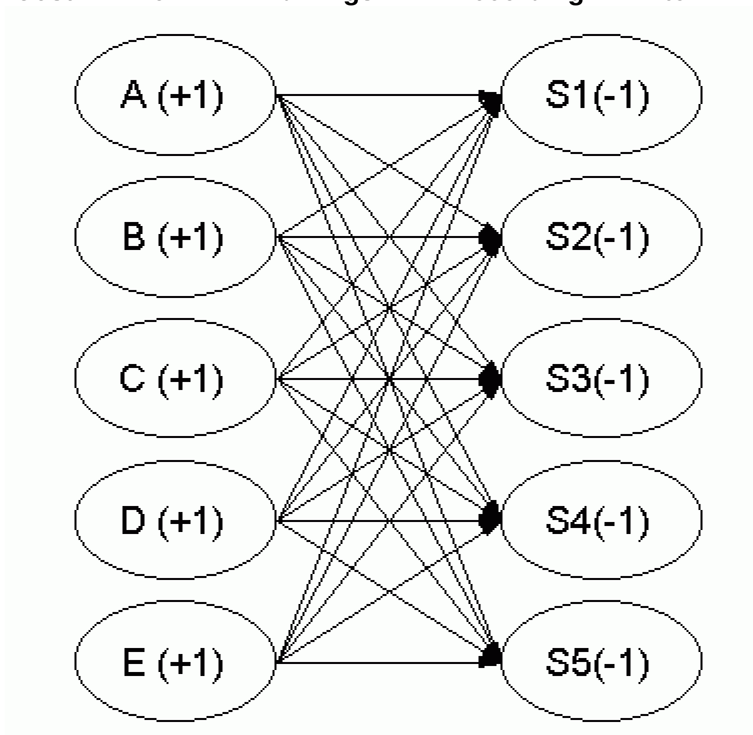
Assignment problem approach of the assignment model, solution methods of assignment problem , maximization in an assignment, unbalanced assignment problem,restriction on assignment.

Another transportation problem is the assignment problem. You can use this problem to assign tasks to people or jobs to machines. You can also use it to award contracts to bidders. Let's describe the assignment problem as assigning *n* tasks to *n* people. Each person must do one and only one task, and each task must be done by only one person. You represent each person and each task by a node. Number the people *1* to *n*, and number the tasks *1* to *n.* The assignment problem can be described simply using a verbal model or using a linear programming mathematical model .

For example, say that five groups of computer users must be trained for five new types of software. Because the users have different computer skill levels, the total cost of trainings depends on the assignments.

Software Types / User Groups	S1	S2	S3	S4	S5
A	5	4	6	4	1
B	2	5	4	10	5
C	10	12	10	6	8
D	1	3	4	2	6
E	2	5	8	11	7

"Cost of Trainings According to the Groups of Users"



"Network Model for Assignment Problem"

Table shows the cost of training for each assignment of a user group (A through E) to a software type (S1 through S5). Picture is a network model of this problem.

A balanced assignment problem has the same number of people and tasks. For a balanced assignment problem, the relationships are all equal. Each person must do a task. For an unbalanced assignment problem with more people than tasks, some people don't have to do a task and the first class of constraints is of the $\leq$ type. In general, the simplex method does not guarantee that the optimal values of the decision variables are integers. Fortunately, for the assignment model, all of the corner point solutions have integer values for all of the variables. Therefore, when the simplex method determines the optimal corner point, all of the variable values are integers and the constraints require that the integers be either 1 or 0 (Boolean).

Conversion to a Balanced Transportation Table

It's not surprising that the variable values for corner point solutions to the assignment model are integers. The assignment model is a special case of the transportation problem, and the transportation problem has integer variable values for every corner point.

For the assignment model, the number of supply and demand points are both *n*. The supply points correspond to each person, and the demand points correspond to each task. Furthermore, every supply amount is 1 and every demand amount is 1. There is one of each person and one of each task.

Software Types / User Groups	S1	S2	S3	S4	S5	Source Supply
A	5	4	6	4	1	1
B	2	5	4	10	5	1
C	10	12	10	6	8	1
D	1	3	4	2	6	1
E	2	5	8	11	7	1
Destination Demands	1	1	1	1	1	5 / 5

"The Computer Users Assignment Problem in the Transportation Table Format"

Table represents the computer users assignment problem in the balanced transportation table format. For the computer users assignment problem, you minimize the total cost of training. Because number of users = 5 and software types = 5, the transportation problem is balanced. You can use standard method for initial solution finding (Northwest-Corner method, Least-Cost method, or Vogel's approximation method (VAM) and a Stepping–Stone method for developing an optimal solution.

Thinking of the transportation problem as shipping items between cities is helpful, but it's the mathematical structure that is important. The assignment problem has the same mathematical structure as the transportation problem, even though it has nothing to do with shipping units.

Note, that each feasible solution of assignment problem will always be strongly degenerated (number of nonzero variables will always be *n*)

TEXT BOOKS

[T1] Sharma, J.K.; Operations Research: problems & solutions; Macmillan India

[T2] Gupta, S.P. and Gupta, P.K.; Quantitative Techniques and Operations Research, Sultan Chand & Sons

[T3] Vohra, N.D.; Quantitative Techniques in Management 2003.

[T4] Gupta, S.P. Statistical Methods, Sultan Chand &Sons.2004

[T5] A. M. Natarajan, P Balasubramani A. Tamilarasi, Operations Research, Pearson 2005

REFERENCE BOOKS

[R1] R.L.Rardin, Optimization in Operations Research, Prentice Hall.

[R2] A.Racindran, D.T.Phillips, J.S.Solberg, Second edition, John Wiley.