STATISTICS FOR BUSINESS DECISION

INTRODUCTION TO STATISTICS

DEFINITION OF STATISTICS

Branch of mathematics concerned with collection, classification, analysis, and interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood (probability). Statistics can interpret aggregates of data too large to be intelligible by ordinary observation because such data (unlike individual quantities) tend to behave in regular, predictable manner. It is subdivided into descriptive statistics and inferential statistics.

HISTORY OF STATISTICS

The Word statistics have been derived from Latin word ―Status‖ or the Italian word ―Statista‖, meaning of these words is ―Political State‖ or a Government. Shakespeare used a word Statist is his drama Hamlet (1602). In the past, the statistics was used by rulers. The application of statistics was very limited but rulers and kings needed information about lands, agriculture, commerce, population of their states to assess their military potential, their wealth, taxation and other aspects of government.

Gottfried Achenwall used the word statistik at a German University in 1749 which means that political science of different countries. In 1771 W. Hooper (Englishman) used the word statistics in his translation of Elements of Universal Erudition written by Baron B.F Bieford, in his book statistics has been defined as the science that teaches us what is the political arrangement of all the modern states of the known world. There is a big gap between the old statistics and the modern statistics, but old statistics also used as a part of the present statistics.

During the 18th century the English writer have used the word statistics in their works, so statistics has eveloped gradually during last few centuries. A lot of work has been done in the end of the nineteenth century.

At the beginning of the 20th century, William S Gosset was developed the methods for decision making based on small set of data. During the 20th century several statistician are active in developing new methods, theories and application of statistics. Now these days the availability of electronics computers is certainly a major factor in the modern development of statistics.

Descriptive Statistics and Inferential Statistics

Every student of statistics should know about the different branches of statistics to correctly understand statistics from a more holistic point of view. Often, the kind of job or work one is involved in hides the other aspects of statistics, but it is very important to know the overall idea behind statistical analysis to fully appreciate its importance and beauty.

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

MANAGERIAL APPLICATIONS OF STATISTICS

Statistics is a the mathematical science involving the collection, analysis and interpretation of data. A number of specialties have evolved to apply statistical theory and methods to various disciplines. Certain topics have "statistical" in their name but relate to manipulations of probability distributions rather than to statistical analysis.

- **Actuarial science** is the discipline that applies mathematical and statistical methods to assess risk in the insurance and finance industries.
- **Astrostatistics** is the discipline that applies statistical analysis to the understanding of astronomical data.
- **Biostatistics** is a branch of biology that studies biological phenomena and observations by means of statistical analysis, and includes medical statistics.
- **Business analytics** is a rapidly developing business process that applies statistical methods to data sets (often very large) to develop new insights and understanding of business performance & opportunities

- **Chemometrics** is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.
- **Demography** is the statistical study of all populations. It can be a very general science that can be applied to any kind of dynamic population, that is, one that changes over time or space.
- **Econometrics** is a branch of economics that applies statistical methods to the empirical study of economic theories and relationships.
- **Environmental statistics** is the application of statistical methods to environmental science. Weather, climate, air and water quality are included, as are studies of plant and animal populations.
- **Geostatistics** is a branch of geography that deals with the analysis of data from disciplines such as petroleum geology, hydrogeology, hydrology, meteorology, oceanography, geochemistry, geography.
- **Operations research** (or Operational Research) is an interdisciplinary branch of applied mathematics and formal science that uses methods such as mathematical modeling, statistics, and algorithms to arrive at optimal or near optimal solutions to complex problems.
- **Population ecology** is a sub-field of ecology that deals with the dynamics of species populations and how these populations interact with the environment.
- **Quality control** reviews the factors involved in manufacturing and production; it can make use of statistical sampling of product items to aid decisions in process control or in accepting deliveries.
- **Quantitative psychology** is the science of statistically explaining and changing mental processes and behaviors in humans.
- **Statistical finance**, an area of econophysics, is an empirical attempt to shift finance from its normative roots to a positivist framework using exemplars from statistical physics with an emphasis on emergent or collective properties of financial markets.
- **Statistical mechanics** is the application of probability theory, which includes mathematical tools for dealing with large populations, to the field of mechanics, which is concerned with the motion of particles or objects when subjected to a force.

Statistical physics is one of the fundamental theories of physics, and uses methods of probability theory in solving physical problems.

STATISTICS AND COMPUTERS

Crunch numbers to the nth degree — and see what happens. When you study computer science and mathematics, you_ll use algorithms and computational theory to create mathematical models or define formulas that solve mathematical problems. In other words, you'll design new tools that can predict the future.

The Computer Applications option gives students the flexibility to combine a traditional computer science degree with a non-traditional field. Our state-of-the-art labs for high- performance computing, networks and artificial intelligence will give you experience with the tools you_ll use in the field. Through labs, lectures and projects, you_ll also:

1. Investigate the computational limits of the algorithms and data structures that support complex software systems
2. Develop new applications and tools in multi-disciplinary areas of science and research
3. Explore opportunities for advanced computer modeling and simulation

IMPORTANCE OF STATISTICS IN DIFFERENT FIELDS

Statistics plays a vital role in every fields of human activity. Statistics has important role in determining the existing position of per capita income, unemployment, population growth rate, housing, schooling medical facilities etc…in a country. Now statistics holds a central position in almost every field like Industry, Commerce, Trade, Physics, Chemistry, Economics, Mathematics, Biology, Botany, Psychology, Astronomy etc…, so application of statistics is very wide. Now we discuss some important fields in which statistics is commonly applied.

1. **Business:**

   Statistics play an important role in business. A successful businessman must be very quick and accurate in decision making. He knows that what his customers wants, he should therefore, know what to produce and sell and in what quantities. Statistics helps businessman to plan production

according to the taste of the costumers, the quality of the products can also be checked more efficiently by using statistical methods. So all the activities of the businessman based on statistical information. He can make correct decision about the location of business, marketing of the products, financial resources etc…

2. In Economics:

Statistics play an important role in economics. Economics largely depends upon statistics. National income accounts are multipurpose indicators for the economists and administrators. Statistical methods are used for preparation of these accounts. In economics research statistical methods are used for collecting and analysis the data and testing hypothesis. The relationship between supply and demands is studies by statistical methods, the imports and exports, the inflation rate, the per capita income are the problems which require good knowledge of statistics.

4 In Mathematics:

Statistical plays a central role in almost all natural and social sciences. The methods of natural sciences are most reliable but conclusions draw from them are only probable, because they are based on incomplete evidence. Statistical helps in describing these measurements more precisely. Statistics is branch of applied mathematics. The large number of statistical methods like probability averages, dispersions, estimation etc… is used in mathematics and different techniques of pure mathematics like integration, differentiation and algebra are used in statistics.

5. In Banking:

Statistics play an important role in banking. The banks make use of statistics for a number of purposes. The banks work on the principle that all the people who deposit their money with the banks do not withdraw it at the same time. The bank earns profits out of these deposits by lending to others on interest. The bankers use statistical approaches based on probability to estimate the numbers of depositors and their claims for a certain day.

6. **In State Management (Administration):**

Statistics is essential for a country. Different policies of the government are based on statistics. Statistical data are now widely used in taking all administrative decisions. Suppose if the government wants to revise the pay scales of employees in view of an increase in the living cost, statistical methods will be used to determine the rise in the cost of living. Preparation of federal and provincial government budgets mainly depends upon statistics because it helps in estimating the expected expenditures and revenue from different sources. So statistics are the eyes of administration of the state.

7. **In Accounting and Auditing:**

Accounting is impossible without exactness. But for decision making purpose, so much precision is not essential the decision may be taken on the basis of approximation, know as statistics. The correction of the values of current asserts is made on the basis of the purchasing power of money or the current value of it.In auditing sampling techniques are commonly used. An auditor determines the sample size of the book to be audited on the basis of error.

8. **In Natural and Social Sciences:**

Statistics plays a vital role in almost all the natural and social sciences. Statistical methods are commonly used for analyzing the experiments results, testing their significance in Biology, Physics, Chemistry, Mathematics, Meteorology, Research chambers of commerce, Sociology, Business, Public Administration, Communication and Information Technology etc…

9. **In Astronomy:**

Astronomy is one of the oldest branches of statistical study; it deals with the measurement of distance, sizes, masses and densities of heavenly bodies by means of observations. During these measurements errors are unavoidable so most probable measurements are founded by using statistical methods.

MEASURES OF CENTRAL TENDENCY

MEASURES OF CENTRAL TENDENCY:

The term **central tendency** refers to the "middle" value or perhaps a typical value of the data, and is measured using the **mean**, **median**, or **mode**. Each of these measures is calculated differently, and the one that is best to use depends upon the situation.

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations. There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages.

Arithmetic mean or mean

Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. It is denoted by the symbol $\bar{x}$. If the variable x assumes n values x1, x2 … xn then the mean is given by

The arithmetic mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. The symbol "μ" is used for the mean of a population. The symbol "M" is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \Sigma X/N$$

where ΣX is the sum of all the numbers in the population and

N is the number of numbers in the population.

The formula for M is essentially identical:

$$M = \Sigma X/N$$

where ΣX is the sum of all the numbers in the sample and

N is the number of numbers in the sample.

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

Grouped Data

The mean for grouped data is obtained from the following formula:

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

Where x = the mid-point of individual class

f = the frequency of individual class

n = the sum of the frequencies or total frequencies in a sample.

Short-cut method

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

Where $d = \frac{x - A}{c}$

A = any value in x

n = total frequency

c = width of the class interval

Merits and demerits of Arithmetic mean

Merits

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

Median

The median is the middle most item that divides the group into two equal parts, one part comprising all values greater, and the other, all values less than that item.

Ungrouped or Raw data

Arrange the given values in the ascending order. If the number of values are odd, median is the middle value.If the number of values are even, median is the mean of middle two values.

By formula $\left(\frac{n+1}{2}\right)^{th}$

When n is odd, Median = Md $value$

When n is even, Average of $\left(\frac{n}{2}\right) and \left(\frac{n}{2}+1\right)^{th} value$

Grouped data

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items.

Cumulative frequency (cf)

Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, ie adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

Discrete Series

Step1: Find cumulative frequencies.

Step2 : Find $\left(\frac{n}{2}+1\right)$

Step3: See in the cumulative frequencies the value just greater than $\left(\frac{n}{2}+1\right)$

Step4: Then the corresponding value of x is median.

Merits of Median

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open-end intervals.
3. Median can be located even if the data are incomplete.

Demerits of Median

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in calculating mean deviation.
4. It does not take into account all the observations.

Mode

The mode refers to that value in a distribution, which occur most frequently. It is an actual value, which has the highest concentration of items in and around it. It shows the centre of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Its importance is very great in agriculture like to find typical height of a crop variety, maximum source of irrigation in a region, maximum disease prone paddy variety. Thus the mode is an important measure in case of qualitative data.

Computation of the mode Ungrouped or Raw Data

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Geometric mean

The geometric mean of a series containing n observations is the nth root of the product of the values. If x1, x2…, xn are observations then

$$G.M = \sqrt[n]{x_1, x_2 \ldots x_n}$$

$$= (x_1, x_2 \ldots x_n)^{1/n}$$

$$\text{Log GM} = \frac{1}{n}\log(x_1, x_2 \ldots x_n)$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n)$$

$$= \frac{\sum \log x_i}{n}$$

$$GM = \text{Antilog} \frac{\sum \log x_i}{n}$$

For grouped data

$$GM = \text{Antilog}\left[\frac{\sum f \log x_i}{n}\right]$$

GM is used in studies like bacterial growth, cell division, etc.

Harmonic mean (H.M)

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If x1, x2…..xn are n observations,

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

For a frequency distribution

$$\text{H.M} = \frac{n}{\sum_{i=n}^{n} f\left(\frac{1}{x_i}\right)}$$

$$= \text{———} = 21.91$$

H.M is used when we are dealing with speed, rates, etc.

Merits of H.M

1. It is rigidly defined.
2. It is defined on all observations.
3. It is amenable to further algebraic treatment.
4. It is the most suitable average when it is desired to give greater weight to smaller observations and less weight to the larger ones.

Demerits of H.M

1. It is not easily understood.
2. It is difficult to compute.
3. It is only a summary figure and may not be the actual item in the series
4. It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.
5. It is rarely used in grouped data.

Percentiles

The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The x^{th} percentile is that value below which x percent of values in the distribution fall. It may be noted that the median is the 50^{th} percentile.

For raw data, first arrange the n observations in increasing order. Then the x^{th} percentile is given by

$$P_x = \left(\frac{x(n+1)}{100}\right)^{th} \text{ item}$$

For a frequency distribution the x^{th} percentile is given by

$$P_x = l + \left(\frac{(x.n/100) - cf}{f} \times C\right)$$

Where

l = lower limit of the percentile calss which contains the x^{th} percentile value (x. n /100)

cf = cumulative frequency uotp l

f = frequency of the percentile class

f C= class interval

N= total number of observations

Quartiles

The quartiles divide the distribution in four parts. There are three quartiles. The second quartile divides the distribution into two halves and therefore is the same as the median. The first (lower).quartile (Q1) marks off the first one-fourth, the third (upper) quartile (Q3) marks off the three-fourth. It may be noted that the second quartile is the value of the median and 50^{th} percentile.

Raw or ungrouped data

First arrange the given data in the increasing order and use the formula for Q1 and Q3 then quartile deviation, Q.D is given by

$$Q.D = \frac{Q_3 - Q_1}{2}$$

Where $Q_1. = \left(\frac{n+1}{4}\right)^{th}$ item and $Q_3. = 3\left(\frac{n+1}{4}\right)^{th}$ item

Discrete Series

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{n+1}{4}\right)$

Step3: See in the cumulative frequencies, the value just greater than $\left(\frac{n+1}{4}\right)$, then the corresponding value of x is Q1

Step4: Find $3\left(\frac{n+1}{4}\right)$

Step5: See in the cumulative frequencies, the value just greater than $3\left(\frac{n+1}{4}\right)$,then the corresponding value of x is Q3

Continuous series

Step1: Find cumulative frequencies

Step2: Find $\left(\frac{n}{4}\right)$

Step3: See in the cumulative frequencies, the value just greater than $\left(\frac{n}{4}\right)$, then the corresponding class interval is called first quartile class.

Step4: Find $3\left(\frac{n}{4}\right)$ See in the cumulative frequencies the value just greater than $3\left(\frac{n}{4}\right)$ then the corresponding class interval is called 3rd quartile class. Then apply the respective formulae

$$Q_1 = l_1 + \frac{\frac{n}{4} - m_1}{f_1} \times c_1$$

Q_3

Where l_1 = lower limit of the first quartile class

$$= l_3 + \frac{3\left(\frac{n}{4}\right) - m_3}{f_3} \times c_3$$

f_1 = frequency of the first quartile class

c_1 = width of the first quartile class

m_1 = c.f. preceding the first quartile class

l_3 = 1ower limit of the 3rd quartile class f_3 = frequency of the 3rd quartile class c_3 = width of the 3rd quartile class

m_3 = c.f. preceding the 3rd quartile class

RANGE

The difference between the lowest and highest values.

Quartile Deviation :

In a distribution, partial variance between the upper quartile and lower quartile is known as 'quartile deviation'. Quartile Deviation is often regarded as semi inter quartile range.

Formula : (Upper quartile- lower quartile) / 2

upper quartile = 400, lower quartile = 200 then

Quartile deviation (QD) = (400-200)/2 = 200/2

=100.

Mean Deviation

The mean of the distances of each value from their mean.

Yes, we use "**mean**" twice: Find the mean ... use it to work out distances ... then find the mean of those distances!

For **deviation** just think **distance** Formula

The formula is:

$$\text{Mean Deviation} = \frac{\Sigma|x - \mu|}{N}$$

Let's learn more about those symbols!

Firstly:

- **μ** is the mean (in our example **μ = 9**)
- **x** is each value (such as 3 or 16)
- **N** is the number of values (in our example **N = 8**)

Standard Deviation

The Standard Deviation is a measure of how spread out numbers are.

Its symbol is **σ** (the greek letter sigma)

The formula is easy: it is the **square root** of the **Variance.** So now you ask, "What is the Variance?"

Variance:

The Variance is defined as. The average of the squared differences from the Mean.

SKEWNESS

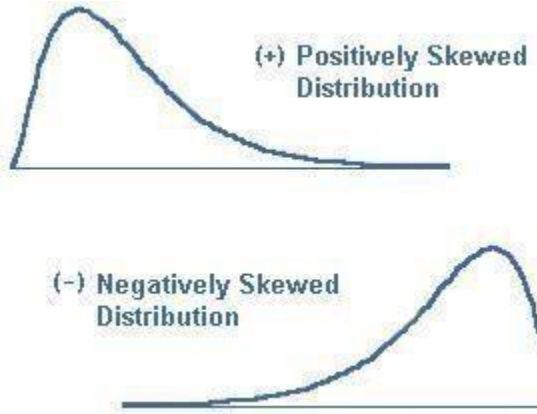
Lack of symmetry is called Skewness. If a distribution is not symmetrical then it is called skewed distribution. So, mean, median and mode are different in values and one tail becomes longer than other. The skewness may be positive or negative.

Positively skewed distribution:

If the frequency curve has longer tail to right the distribution is known as positively skewed distribution and ***Mean > Median > Mode***.

Negatively skewed distribution:

If the frequency curve has longer tail to left the distribution is known as negatively skewed distribution and ***Mean < Median < Mode***.



Measure of Skewness:

The difference between the mean and mode gives as absolute measure of skewness. If we divide this difference by standard deviation we obtain a relative measure of skewness known as coefficient and denoted by ***SK***.

Karl Pearson coefficient of Skewness

$$SK=Mean-Mode/S.D$$

Sometimes the mode is difficult to find. So we use another formula

$$SK=3(Mean-Median)/S.D$$

Bowley_s coefficient of Skewness

$$SK=Q1+Q3-2Median/Q3-Q1$$

Kelly_s Measure of Skewness is one of several ways to measure skewness in a data distribution. Bowley_s skewness is based on the middle 50 percent of the observations in a data set. It leaves 25 percent of the observations in each tail of the distribution. Kelly suggested that leaving out fifty percent of data to calculate skewness was too extreme. He created a measure to find skewness with more data. Kelly_s measure is based on P_{90} (the 90th percentile) and P_{10} (the 10th percentile). Only twenty percent of observations (ten percent in each tail) are excluded from the measure**.**

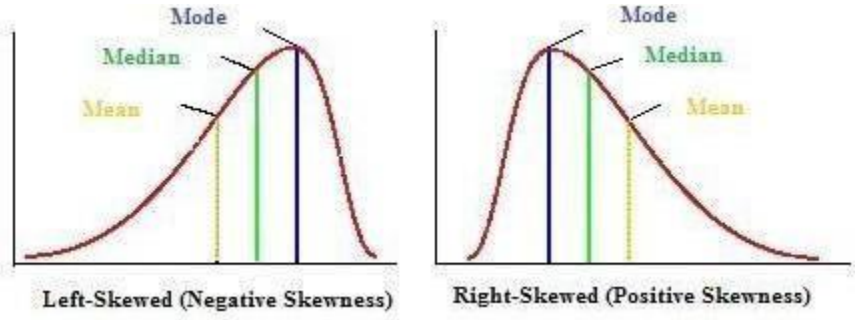
Kelly's Measure Formula.

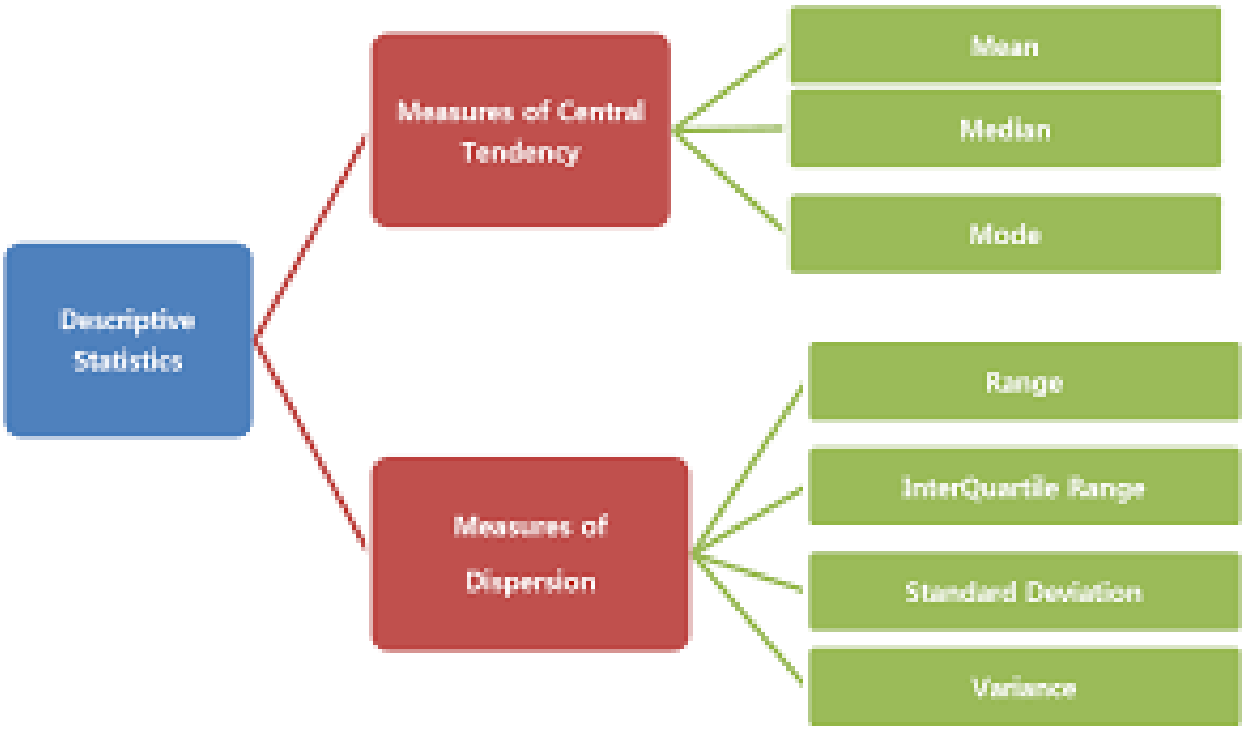
Kelley_s measure of skewness is given in terms of percentiles and deciles(D). Kelley_s absolute measure of skewness (S_k)is:

$S_k=P_{90} + P_{10} – 2*P_{50} = D_9 + D_1-2*D_5$.

Kelly_s Measure of Skewness gives you the same information about skewness as the other three types of skewness measures

A measure of skewness = 0 means that the distribution is symmetrical.
A measure of skewness > 0 means a positive skewness.
A measure of skewness < means a negative skewness.




Descriptive Statistics
Measures of Central Tendency
Mean
Median
Mode
Measures of Dispersion
Range
InterQuartile Range
Standard Deviation
Variance

TABULATION OF UNIVARIATE

Univariate data: **Univariate** means "one variable" (one type of data)

Bivariate data: Bivariate **means "two variables", in other words there are two types of data With bivariate data you have** two **sets of related data that you want to** compare**:** Example:

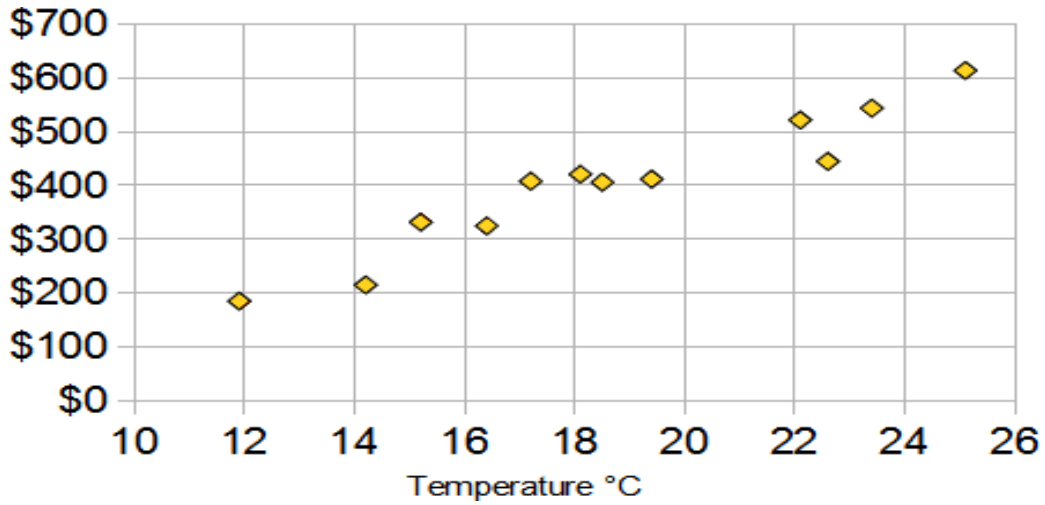
An ice cream shop keeps track of how much ice cream they sell versus the temperature on that day.

The two variables are **Ice Cream Sales** and **Temperature**.

Here are their figures for the last 12 days:

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

And here is the same data as a Scatter Plot:



Now we can easily see that **warmer weather** and **more ice cream sales** are linked, but the relationship is not perfect.

Multivariate data

Multivariate Data Analysis refers to any statistical technique used to analyze **data** that arises from more than one variable. This essentially models reality where each situation, product, or decision involvesmore than a single variable.

Univariate Data	Bivariate Data
• involving a **single variable**	• involving **two variables**
• does not deal with causes or relationships	• deals with causes or relationships
• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
• central tendency - mean, mode, median • dispersion - range, variance, max, min, quartiles, standard deviation. • frequency distributions • bar graph, histogram, pie chart, line graph, box-and-whisker plot	• analysis of two variables simultaneously • correlations • comparisons, relationships, causes, explanations • tables where one variable is contingent on the values of the other variable. • independent and dependent variables
Sample question: Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?	**Sample question:** Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

DIAGRAMATIC AND GRAPHIACAL; REPRESENTATION OF DATA

Although tabulation is very good technique to present the data, but diagrams are an advanced technique to represent data. As a layman, one cannot understand the tabulated data easily but with only a single glance at the diagram, one gets complete picture of the data presented. According to M.J. Moroney, ―diagrams register a meaningful impression almost before we think.

Importance or utility of Diagrams

- Diagrams give a very clear picture of data. Even a layman can understand it very easily and in a short time.
- We can make comparison between different samples very easily. We don_t have to use any statistical technique further to compare.
- This technique can be used universally at any place and at any time. This technique is used almost in all the subjects and other various fields.

- Diagrams have impressive value also. Tabulated data has not much impression as compared to Diagrams. A common man is impressed easily by good diagrams.
- This technique can be used for numerical type of statistical analysis, e.g. to locate Mean, Mode, Median or other statistical values.
- It does not save only time and energy but also is economical. Not much money is needed to prepare even good diagrams.
- These give us much more information as compared to tabulation. Technique of tabulation has its own limits.
- This data is easily remembered. Diagrams which we see leave their lasting impression much more than other data techniques.
- Data can be condensed with diagrams. A simple diagram can present what even cannot be presented by 10000 words.

General Guidelines for Diagrammatic presentation

- The diagram should be properly drawn at the outset. The pith and substance of the subject matter must be made clear under a broad heading which properly conveys the purpose of a diagram.
- The size of the scale should neither be too big nor too small. If it is too big, it may look ugly. If it is too small, it may not convey the meaning. In each diagram, the size of the paper must be taken note-of. It will help to determine the size of the diagram.
- For clarifying certain ambiguities some notes should be added at the foot of the diagram. This shall provide the visual insight of the diagram.
- Diagrams should be absolutely neat and clean. There should be no vagueness or overwriting on the diagram.
- Simplicity refers to love at first sight. It means that the diagram should convey the meaning clearly and easily.
- Scale must be presented along with the diagram.
- It must be Self-Explanatory. It must indicate nature, place and source of data presented.
- Different shades, colors can be used to make diagrams more easily understandable.
- Vertical diagram should be preferred to Horizontal diagrams.
- It must be accurate. Accuracy must not be done away with to make it attractive or impressive.

Limitations of Diagrammatic Presentation

- Diagrams do not present the small differences properly.
- These can easily be misused.
- Only artist can draw multi-dimensional diagrams.
- In statistical analysis, diagrams are of no use.
- Diagrams are just supplement to tabulation.
- Only a limited set of data can be presented in the form of diagram.
- Diagrammatic presentation of data is a more time consuming process.
- Diagrams present preliminary conclusions.
- Diagrammatic presentation of data shows only on estimate of the actual behavior of the variables.

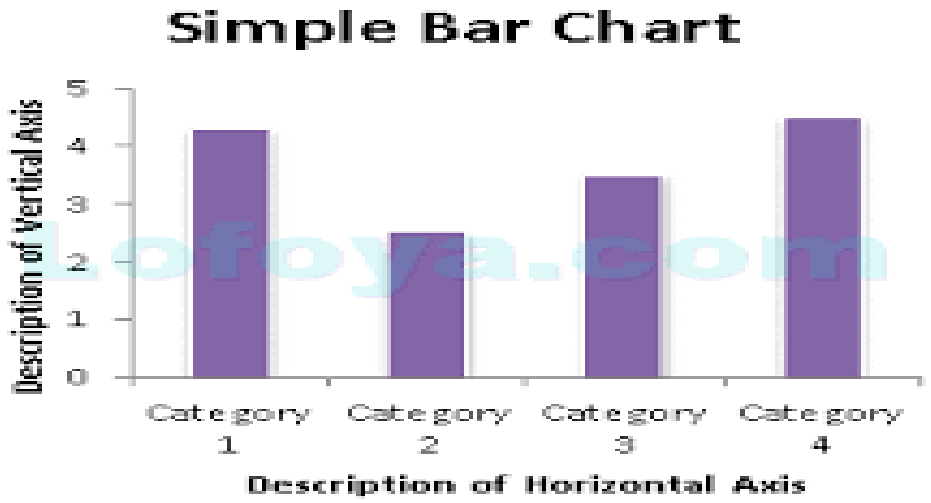
Types of Diagrams

(a) Line Diagrams

In these diagrams only line is drawn to represent one variable. These lines may be vertical or horizontal. The lines are drawn such that their length is the proportion to value of the terms or items so that comparison may be done easily.

(b) Simple Bar Diagram

Like line diagrams these figures are also used where only single dimension i.e. length can present the data. Procedure is almost the same, only one thickness of lines is measured. These can also be drawn either vertically or horizontally. Breadth of these lines or bars should be equal. Similarly distance between these bars should be equal. The breadth and distance between them should be taken according to space available on the paper.

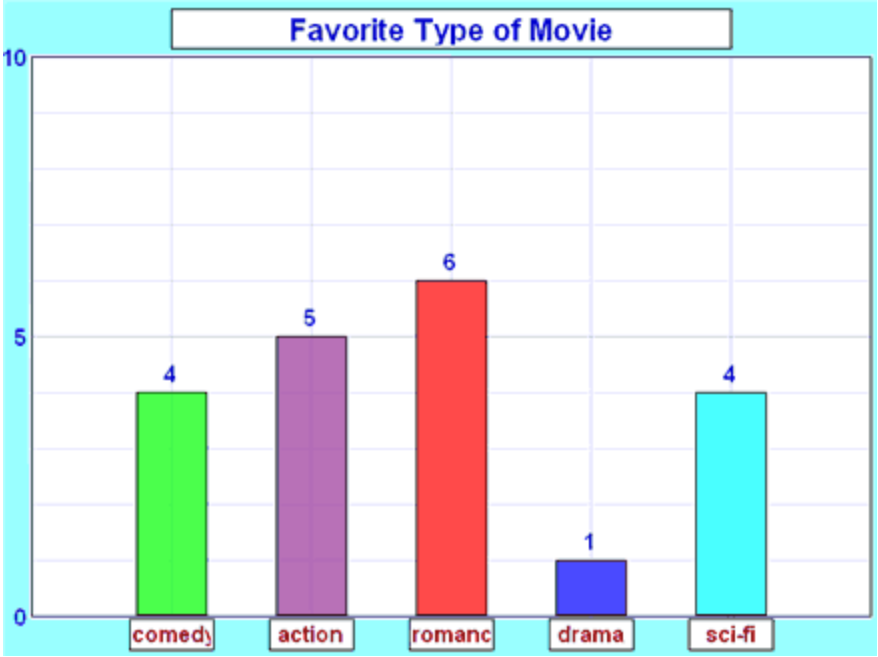


Imagine you just did a survey of your friends to find which kind of movie they liked best:

Table: Favorite Type of Movie

Comedy	Action	Romance	Drama	SciFi
4	5	6	1	4

We can show that on a bar graph like this:

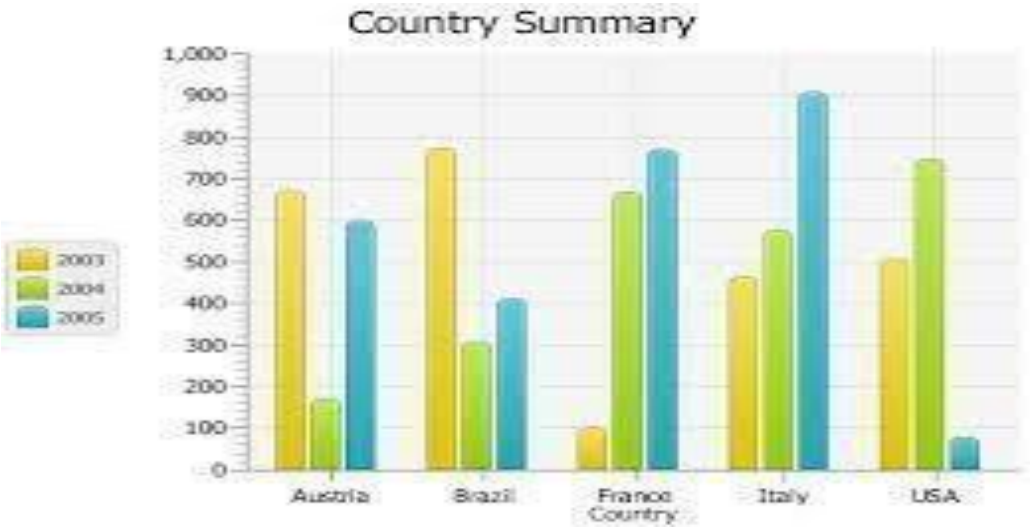


It is a really good way to show relative sizes: we can see which types of movie are most liked, and which are least liked, at a glance.

We can use bar graphs to show the relative sizes of many things, such as what type of car people have, how many customers a shop has on different days and so on.

(c) Multiple Bar Diagrams

The diagram is used, when we have to make comparison between more than two variables. The number of variables may be 2, 3 or 4 or more. In case of 2 variables, pair of bars is drawn. Similarly, in case of 3 variables, we draw triple bars. The bars are drawn on the same proportionate basis as in case of simple bars. The same shade is given to the same item.



Draw a multiple bar chart to represent the import and export of Canada (values in $) for the years 1991 to 1995.

Years	Imports	Exports
1991	7930	4260
1992	8850	5225
1993	9780	6150
1994	11720	7340
1995	12150	8145

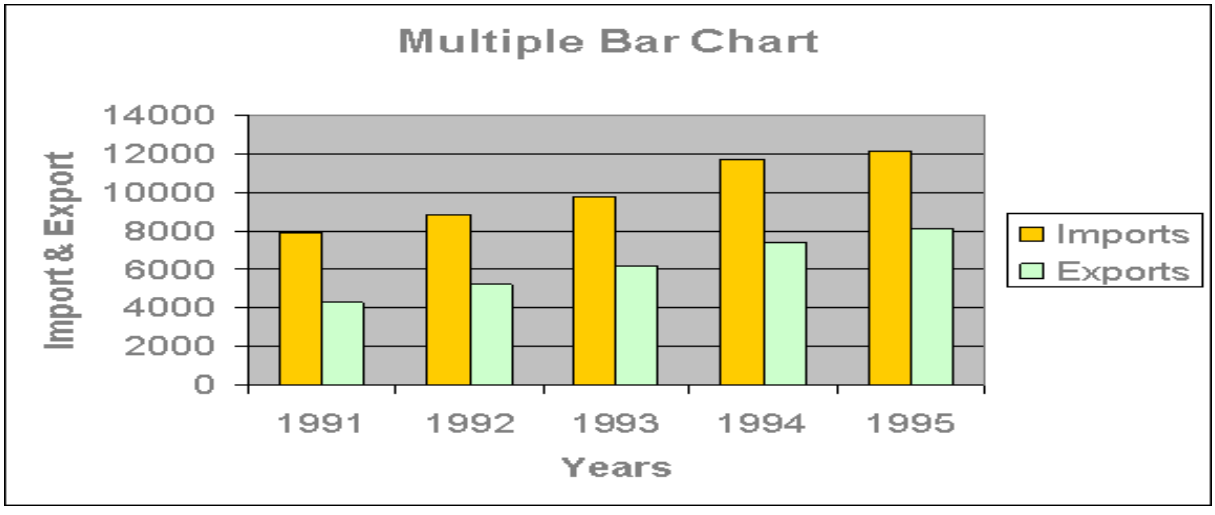
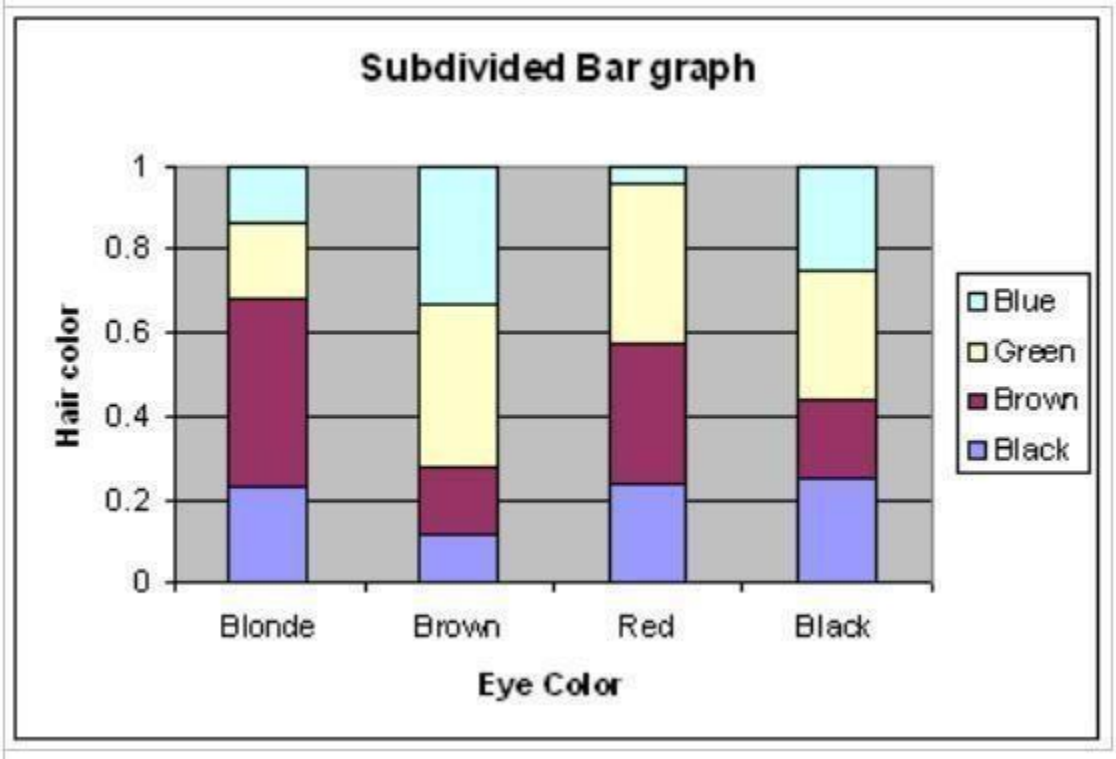


chart showing the import and export of Canada from 1991 – 1995.

(d) Sub-divided Bar Diagram

The data which is presented by multiple bar diagram can be presented by this diagram. In this case we add different variables for a period and draw it on a single bar as shown in the following examples. The components must be kept in same order in each bar. This diagram is more efficient if number of components is less i.e. 3 to 5.



(e) Percentage Bar Diagram

Like sub-divide bar diagram, in this case also data of one particular period or variable is put on single bar, but in terms of percentages. Components are kept in the same order in each bar for Easy comparison.

(f) Duo-directional Bar Diagram

In this case the diagram is on both the sides of base line i.e. to left and right or to above or below sides.

(g) Broken Bar Diagram

This diagram is used when value of some variable is very high or low as compared to others. In this case the bars with bigger terms or items may be shown broken.

One dimentional diagrams:

A diagram in which the size of only one dimension i.e. length is fixed in proportion to the value of the data is called one dimensional diagram. Such diagrams are also popularly called bar diagrams. These diagrams can be drawn in both vertical and horizontal manner. The related different bar diagrams differ from each other only in respect of their length dimension, while they remain the same in respect of their other two dimensions i.e. breadth and thickness. The size of breadth of each of such diagrams is determined taking into consideration the number of diagrams to be drawn, and the size of the paper at one_s end. They may take the form of a line, or a thread if they are to be drawn in large numbers on the surface of a paper. However, their breadth should neither be too large nor too small for that in both the cases they look ugly. The dimension of thickness does not look prominent in such diagrams. The examples of such diagrams are given on the next page.

Techniques of drawing bar diagrams

The following are the techniques of drawing the bar diagrams :

(i) First, draw the base line, preferably horizontally, and divide it into a number of equal parts keeping in view the number of diagrams to be drawn.

(ii) Then, draw the scale line preferably vertically, and divide into a number of equal parts keeping in view the maximum value to be represented.

(iii) Then, fix the width of the bar uniformly keeping in view the number of bars to be drawn and the gaps to be provided in between each two of them.

(iv) Then, fix the size of gaps to be provided between each of the two bars uniformly.

(v) Then, fix the lengths of the different bars in proportion of the value of the data.

(vi) Then, draw the different bars in accordance with their length, and width thus fixed, and arranged in order o their length, or time of occurrence.

(vii) Then, decorate the bars with similar or different colours, or shades according to the similarly or dissimilarity in the nature of the data respectively.

(viii) Give a description of the data in short at the bottom of respective bars.

(xiv) Put the respective figures at the top of each bar to read out the exact value at a glance without looking at the scale.

Advantages

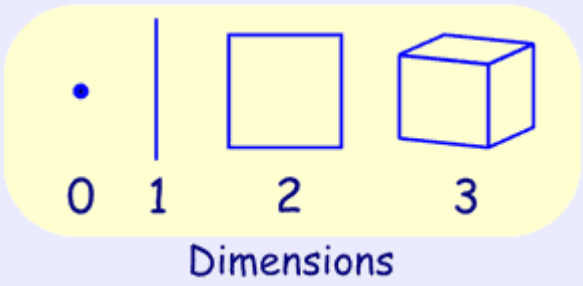
The chief advantages of a bar diagram can be outlined as under:

1. It is very simple to draw and read as well.
2. It is the only form of diagram which can represent a large number of data on a piece of paper.
3. It can be drawn both vertically and horizontally.
4. It gives a better look and facilitates comparison.

Disadvantages

1. It cannot exhibit a large number of aspects of the data.
2. The which of the bars are fixed arbitrarily by a drawer.

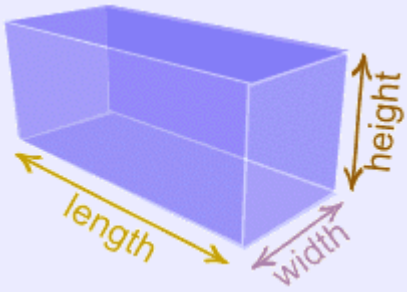
Two-Dimensional



A shape that only has two dimensions (such as width and height) and no thickness. Squares, Circles, Triangles, etc are two dimensional objects.

Also known as "2D".

3.**Three-Dimensional**



An object that has height, width and depth, like any object in the real world. Example: your body is three-dimensional.

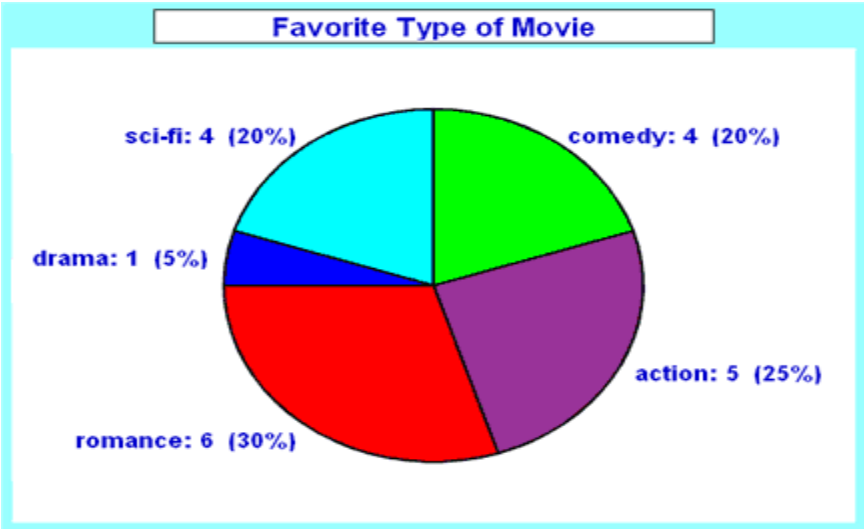
Also known as "3D".

Pie Chart: a special chart that uses "pie slices" to show relative sizes of data.

Imagine you survey your friends to find the kind of movie they like best:

Table: Favorite Type of Movie

Comedy	Action	Romance	Drama	SciFi
4	5	6	1	4



You can show the data by this Pie Chart:

It is a really good way to show relative sizes: it is easy to see which movie types are most liked, and which are least liked, at a glance.

Introduction to Correlation and Regression Analysis

In this section we will first discuss correlation analysis, which is used to quantify the association between two continuous variables (e.g., between an independent and a dependent variable or between two independent variables). Regression analysis is a related technique to assess the relationship between an outcome variable and one or more risk factors or confounding variables. The outcome variable is also called the **response** or **dependent variable** and the risk factors and confounders are called the **predictors**, or **explanatory** or **independent variables**. In regression analysis, the dependent variable is denoted "*y*" and the independent variables are denoted by "*x*".

[**NOTE:** The term "predictor" can be misleading if it is interpreted as the ability to predict even beyond the limits of the data. Also, the term "explanatory variable" might give an impression of a causal effect in a situation in which inferences should be limited to identifying associations. The terms "independent" and "dependent" variable are less subject to these interpretations as they do not strongly imply cause and effect.

Correlation Analysis

In correlation analysis, we estimate a sample correlation coefficient, more specifically the Pearson Product Moment correlation coefficient. The sample correlation coefficient, denoted r, ranges between -1 and +1 and quantifies the direction and strength of the linear association between the two variables. The correlation between two variables can be positive (i.e., higher levels of one variable are associated with higher levels of the other) or negative (i.e., higher levels of one variable are associated with lower levels of the other).

The sign of the correlation coefficient indicates the direction of the association. The magnitude of the correlation coefficient indicates the strength of the association.

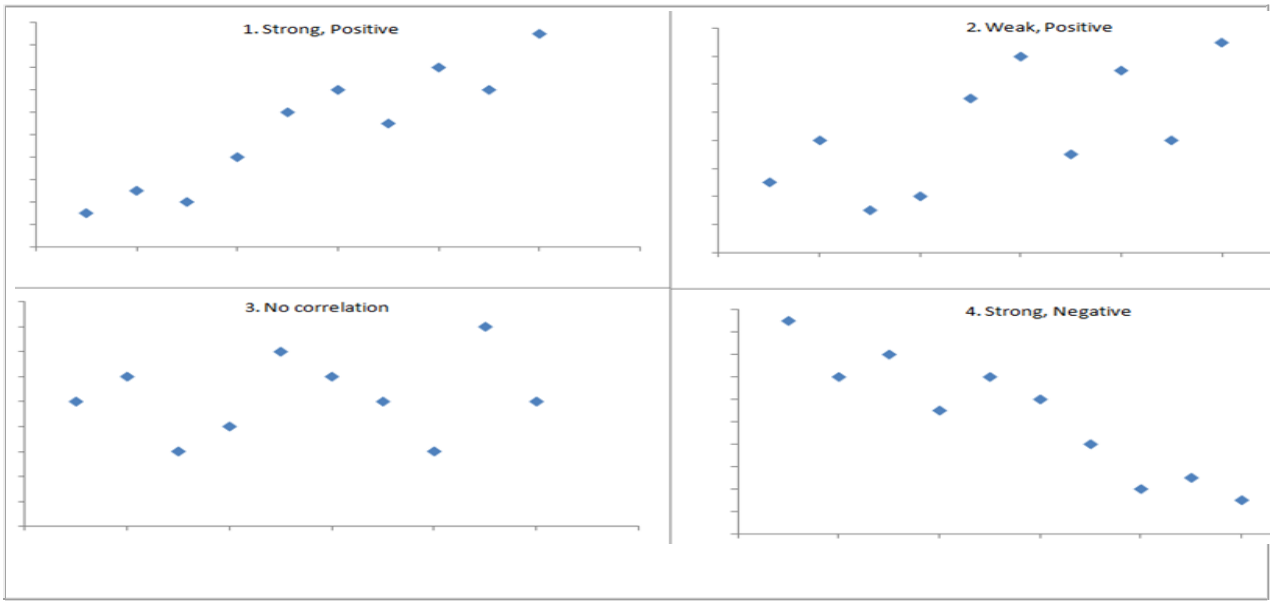
For example, a correlation of r = 0.9 suggests a strong, positive association between two variables, whereas a correlation of r = -0.2 suggest a weak, negative association. A correlation close to zero suggests no linear association between two continuous variables.

LISA: [I find this description confusing. You say that the correlation coefficient is a measure of the "strength of association", but if you think about it, isn't the slope a better measure of association? We use risk ratios and odds ratios to quantify the strength of association, i.e., when an exposure is present it has how many times more likely the outcome is. The analogous quantity in correlation is the slope, i.e., for a given increment in the independent variable, how many times is the dependent variable going to increase?

And "r" (or perhaps better R-squared) is a measure of how much of the variability in the dependent variable can be accounted for by differences in the independent variable. The analogous measure for a dichotomous variable and a dichotomous outcome would be the attributable proportion, i.e., the proportion of Y that can be attributed to the presence of the exposure.]

It is important to note that there may be a non-linear association between two continuous variables, but computation of a correlation coefficient does not detect this. Therefore, it is always important to evaluate the data carefully before computing a correlation coefficient. Graphical displays are particularly useful to explore associations between variables.

The figure below shows four hypothetical scenarios in which one continuous variable is plotted along the X-axis and the other along the Y-axis.



- Scenario 1 depicts a strong positive association (r=0.9), similar to what we might see for the correlation between infant birth weight and birth length.
- Scenario 2 depicts a weaker association (r=0,2) that we might expect to see between age and body mass index (which tends to increase with age).
- Scenario 3 might depict the lack of association (r approximately 0) between the extent of media exposure in adolescence and age at which adolescents initiate sexual activity.
- Scenario 4 might depict the strong negative association (r= -0.9) generally observed between the number of hours of aerobic exercise per week and percent body fat.

The formula for the sample correlation coefficient is

$$r = \frac{\mathrm{Cov}(x,y)}{\sqrt{s_x^2 * s_y^2}}$$

where Cov(x,y) is the covariance of x and y defined as

$$\text{Cov}(x, y) = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{n - 1}$$

S_x^2 and S_y^2 are the sample variances of x and y, defined as

$$s_x^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1} \quad \text{and} \quad s_y^2 = \frac{\Sigma(Y - \overline{Y})^2}{n - 1}$$

The variances of x and y measure the variability of the x scores and y scores around their respective sample means (

$\overline{X}$ and $\overline{Y}$, considered separately). The covariance measures the variability of the (x,y) pairs around the mean of x and mean of y, considered simultaneously.

To compute the sample correlation coefficient, we need to compute the variance of gestational age, the variance of birth weight and also the covariance of gestational age and birth weight.

We first summarize the gestational age data. The mean gestational age is:

Next we compute the covariance,

$$Cov(x,y) = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{n-1}$$

To compute the covariance of gestational age and birth weight, we need to multiply the deviation from the mean gestational age by the deviation from the mean birth weight for each participant (i.e.,

$$(X - \overline{X})(Y - \overline{Y}))$$

Spearman Rank Correlation :

The Pearson correlation coefficient between the ranked variables has been termed as the Spearman correlation coefficient. It is also referred as 'grade correlation'.

Formula :

$R = 1 - ((6 X \sigma d^2) / (n^3 - n))$

Partial correlation analysis involves studying the linear relationship between two variables after excluding the effect of one or more independent factors.

Simple correlation does not prove to be an all-encompassing technique especially under the above circumstances. In order to get a correct picture of the relationship between two variables, we should first eliminate the influence of other variables.

For example, study of partial correlation between price and demand would involve studying the relationship between price and demand excluding the effect of money supply, exports, etc.

Partial correlation analysis involves studying the linear relationship between two variables after excluding the effect of one or more independent factors.

Simple correlation does not prove to be an all-encompassing technique especially under the above circumstances. In order to get a correct picture of the relationship between two variables, we should first eliminate the influence of other variables.

For example, study of partial correlation between price and demand would involve studying the relationship between price and demand excluding the effect of money supply, exports, etc.

Partial correlation analysis involves studying the linear relationship between two variables after excluding the effect of one or more independent factors.

Simple correlation does not prove to be an all-encompassing technique especially under the above circumstances. In order to get a correct picture of the relationship between two variables, we should first eliminate the influence of other variables.

For example, study of partial correlation between price and demand would involve studying the relationship between price and demand excluding the effect of money supply, exports, etc.

Multiple Correlation

Another technique used to overcome the drawbacks of simple correlation is multiple regression analysis.

Here, we study the effects of all the independent variables simultaneously on a dependent variable. For example, the correlation co-efficient between the yield of paddy (X_1) and the other variables, viz. type of seedlings (X_2), manure (X_3), rainfall (X_4), humidity (X_5) is the multiple correlation co-efficient $R_{1.2345}$. This co-efficient takes value between 0 and +1.

The limitations of multiple correlation are similar to those of partial correlation. If multiple and partial correlation are studied together, a very useful analysis of the relationship between the different variables is possible.

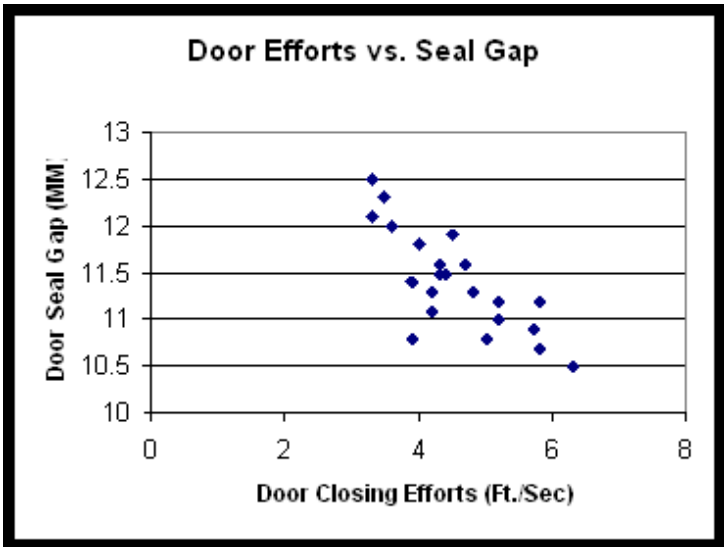
REGRESSION ANALYSIS

Regression Analysis

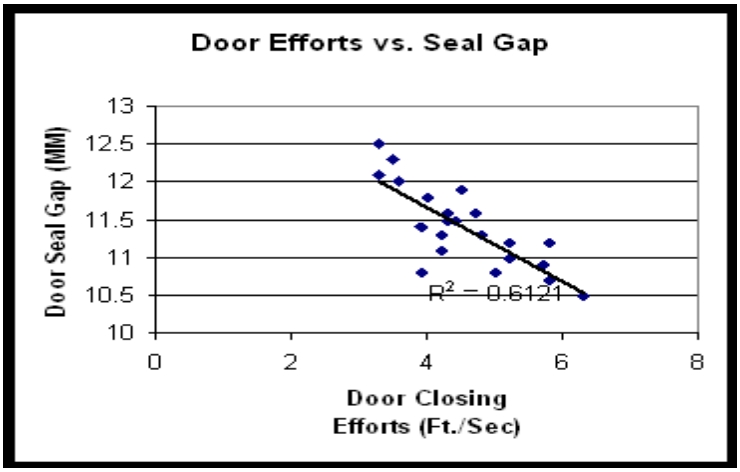
Introduction

As you develop Cause & Effect diagrams based on data, you may wish to examine the degree of correlation between variables. A statistical measurement of correlation can be calculated using the least squares method to quantify the strength of the relationship between two variables. The output of that calculation is the **Correlation Coefficient, or (r)**, which ranges between -1 and 1. A value of 1 indicates perfect positive correlation - as one variable increases, the second increases in a linear fashion. Likewise, a value of -1 indicates perfect negative correlation - as one variable increases, the second decreases. A value of zero indicates zero correlation.

Before calculating the Correlation Coefficient, the first step is to construct a scatter diagram. Most spreadsheets, including Excel, can handle this task. Looking at the scatter diagram will give you a broad understanding of the correlation. Following is a scatter plot chart example based on an automobile manufacturer.



In this case, the process improvement team is analyzing door closing efforts to understand what the causes could be. The Y-axis represents the width of the gap between the sealing flange of a car door and the sealing flange on the body - a measure of how tight the door is set to the body. The fishbone diagram indicated that variability in the seal gap could be a cause of variability in door closing efforts.



In this case, you can see a pattern in the data indicating a negative correlation (negative slope) between the two variables. In fact, the Correlation Coefficient is -0.78, indicating a strong inverse or negative relationship.

MoreSteam Note: *It is important to note that Correlation is not Causation* - two variables can be very strongly correlated, but both can be caused by a third variable. For example, consider two variables: A) how much my grass grows per week, and B) the average depth of the local reservoir. Both variables could be highly correlated because both are dependent upon a third variable - how much it rains.

In our car door example, it makes sense that the tighter the gap between the sheet metal sealing surfaces (before adding weatherstrips and trim), the harder it is to close the door. So a rudimentary understanding of mechanics would support the hypothesis that there is a causal relationship. Other industrial processes are not always as obvious as these simple examples, and determination of causal relationships may require more extensive experimentation (Design of Experiments).

Simple Regression Analysis

While Correlation Analysis assumes no causal relationship between variables, Regression Analysis assumes that one variable is dependent upon: A) another single independent variable (Simple Regression) , or B) multiple independent variables (Multiple Regression).

Regression plots a line of best fit to the data using the least-squares method. You can see an example below of linear regression using the same car door scatter plot:

You can see that the data is clustered closely around the line, and that the line has a downward slope. There is strong negative correlation expressed by two related statistics: the r value, as stated before is, -0.78 the r^2 value is therefore 0.61. R^2, called the **Coefficient of Determination**, expresses how much of the variability in the dependent variable is explained by variability in the independent variable. You may find that a non-linear equation such as an exponential or power function may provide a better fit and yield a higher r^2 than a linear equation.

These statistical calculations can be made using Excel, or by using any of several statistical analysis software packages. MoreSteam provides links to statistical software downloads, including free software.

Multiple Regression Analysis

Multiple Regression Analysis uses a similar methodology as Simple Regression, but includes more than one independent variable. Econometric models are a good example, where the dependent variable of GNP may be analyzed in terms of multiple independent variables, such as interest rates, productivity growth, government spending, savings rates, consumer confidence, etc.

Many times historical data is used in multiple regression in an attempt to identify the most significant inputs to a process. The benefit of this type of analysis is that it can be done very quickly and relatively simply. However, there are **several potential pitfalls:**

- The **data may be inconsistent** due to different measurement systems, calibration drift, different operators, or recording errors.
- The **range of the variables may be very limited**, and can give a false indication of low correlation. For example, a process may have temperature controls because temperature has been found in the past to have an impact on the output. Using historical temperature data may therefore indicate low significance because the range of temperature is already controlled in tight tolerance.

- There may be a **time lag that influences the relationship** - for example, temperature may be much more critical at an early point in the process than at a later point, or vice-versa. There also may be inventory effects that must be taken into account to make sure that all measurements are taken at a consistent point in the process.

Once again, it is critical to remember that correlation is not causality. As stated by Box, Hunter and Hunter: "Broadly speaking, **to find out what happens when you change something, it is necessary to change it.** To safely infer causality the experimenter cannot rely on natural happenings to choose the design for him; he must choose the design for himself and, in particular, must introduce randomization to break the links with possible lurking variables". [1]

Returning to our example of door closing efforts, you will recall that the door seal gap had an r^2 of 0.61. Using multiple regression, and adding the additional variable "door weatherstrip durometer" (softness), the r^2 rises to 0.66. So the durometer of the door weatherstrip added some explaining power, but minimal. Analyzed individually, durometer had much lower correlation with door closing efforts - only 0.41.

This analysis was based on historical data, so as previously noted, the regression analysis only tells us what did have an impact on door efforts, not what could have an impact. If the range of durometer measurements was greater, we might have seen a stronger relationship with door closing efforts, and more variability in the output.

Trend Analysis

There are no proven "automatic" techniques to identify trend components in the time series data; however, as long as the trend is monotonous (consistently increasing or decreasing) that part of data analysis is typically not very difficult. If the time series data contain considerable error, then the first step in the process of trend identification is smoothing.

Smoothing. Smoothing always involves some form of local averaging of data such that the nonsystematic components of individual observations cancel each other out. The most common technique is *moving average* smoothing which replaces each element of the series by either the simple or weighted average of *n* surrounding elements, where *n* is the width of the smoothing "window" (see Box & Jenkins, 1976; Velleman & Hoaglin, 1981). Medians can be used instead of means. The main advantage of median as compared to moving average smoothing is that its results are less biased by outliers (within the smoothing window). Thus, if there are outliers in

the data (e.g., due to measurement errors), median smoothing typically produces smoother or at least more "reliable" curves than moving average based on the same window width. The main disadvantage of median smoothing is that in the absence of clear outliers it may produce more "jagged" curves than moving average and it does not allow for weighting.

In the relatively less common cases (in time series data), when the measurement error is very large, the *distance weighted least squares smoothing* or *negative exponentially weighted smoothing* techniques can be used. All those methods will filter out the noise and convert the data into a smooth curve that is relatively unbiased by outliers (see the respective sections on each of those methods for more details). Series with relatively few and systematically distributed points can be smoothed with *bicubic splines*.

Fitting a function. Many monotonous time series data can be adequately approximated by a linear function; if there is a clear monotonous nonlinear component, the data first need to be transformed to remove the nonlinearity. Usually a logarithmic, exponential, or (less often) polynomial function can be used.

Additive models

The models that we have considered in earlier sections have been **additive models**, and there has been an implicit assumption that the different components affected the time series additively.

Data = Seasonal effect + Trend + Cyclical + Residual

For monthly data, an additive model assumes that the difference between the January and July values is approximately the same each year. In other words, the **amplitude** of the seasonal effect is the same each year.

The model similarly assumes that the residuals are roughly the same size throughout the series -- they are a random component that adds on to the other components in the same way at all parts of the series.

Multiplicative models

In many time series involving **quantities** (e.g. money, wheat production, ...), the absolute differences in the values are of less interest and importance than the percentage changes.

For example, in seasonal data, it might be more useful to model that the July value is the same **proportion** higher than the January value in each year, rather than assuming that their difference is constant. Assuming that the seasonal and other effects act proportionally on the series is equivalent to a **multiplicative model**,

$$\text{Data} = (\text{Seasonal effect}) \times \text{Trend} \times \text{Cyclical} \times \text{Residual}$$

Fortunately, multiplicative models are equally easy to fit to data as additive models! The trick to fitting a multiplicative model is to take logarithms of both sides of the model,

$$\begin{aligned}\log(\text{Data}) &= \log(\text{Seasonal effect} \times \text{Trend} \times \text{Cyclical} \times \text{Residual}) \\ &= \log(\text{Seasonal effect}) + \log(\text{Trend}) \\ &\quad + \log(\text{Cyclical}) + \log(\text{Residual})\end{aligned}$$

After taking logarithms (either natural logarithms or to base 10), the four components of the time series again act additively.

What is an additive model?

A data model in which the effects of individual factors are differentiated and added together to model the data. They occur in several Minitab commands:

- An additive model is optional for Decomposition procedures and for Winters' method.
- An additive model is optional for two-way ANOVA procedures. Choose this option to omit the interaction term from the model.

What is a multiplicative model?

This model assumes that as the data increase, so does the seasonal pattern. Most time series plots exhibit such a pattern. In this model, the trend and seasonal components are multiplied and then added to the error component.

Should I use an additive model or a multiplicative model?

Choose the multiplicative model when the magnitude of the seasonal pattern in the data depends on the magnitude of the data. In other words, the magnitude of the seasonal pattern increases as the data values increase, and decreases as the data values decrease.

Choose the additive model when the magnitude of the seasonal pattern in the data does not depend on the magnitude of the data. In other words, the magnitude of the seasonal pattern does not change as the series goes up or down.

If the pattern in the data is not very obvious, and you have trouble choosing between the additive and multiplicative procedures, you can try both and choose the one with smaller accuracy measures.

INDEX NUMBERS

Introduction:

Index numbers are meant to study the change in the effects of such factors which cannot be measured directly. According to Bowley, ―Index numbers are used to measure the changes in some quantity which we cannot observe directly‖. For example, changes in business activity in a country are not capable of direct measurement but it is possible to study relative changes in business activity by studying the variations in the values of some such factors which affect business activity, and which are capable of direct measurement.

Index numbers are commonly used statistical device for measuring the combined fluctuations in a group related variables. If we wish to compare the price level of consumer items today with that prevalent ten years ago, we are not interested in comparing the prices of only one item, but in comparing some sort of average price levels. We may wish to compare the present agricultural production or industrial production with that at the time of independence. Here again, we have to consider all items of production and each item may have undergone a different fractional increase (or even a decrease). How do we obtain a composite measure? This composite measure is provided by index numbers which may be defined as a device for combining the variations that have come in group of related variables over a period of time, with a view to obtain a figure that represents the _net_ result of the change in the constitute variables.

Index numbers may be classified in terms of the variables that they are intended to measure. In business, different groups of variables in the measurement of which index number techniques are commonly used are (i) price, (ii) quantity, (iii) value and (iv) business activity. Thus, we have index of wholesale prices, index of consumer prices, index of industrial output, index of value of exports and index of business activity, etc. Here we shall be mainly interested in index numbers of prices showing changes with respect to time, although methods described can be applied to other cases. In general, the present level of prices is compared with the level of prices in the past. The present period is called the current period and some period in the past is called the base period.

Index Numbers:

Index numbers are statistical measures designed to show changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc. A collection of index numbers for different years, locations, etc., is sometimes called an index series.

Simple Index Number:

A simple index number is a number that measures a relative change in a single variable with respect to a base.

Composite Index Number:

A composite index number is a number that measures an average relative changes in a group of relative variables with respect to a base.

Types of Index Numbers:

Following types of index numbers are usually used:

Price index Numbers:

Price index numbers measure the relative changes in prices of a commodities between two periods. Prices can be either retail or wholesale.

Quantity Index Numbers:

These index numbers are considered to measure changes in the physical quantity of goods produced, consumed or sold of an item or a group of items.

Uses

This index number is a useful number that helps us quantify changes in our field. It is easier to see one value than a thousand different values for each item in our field.

Take the stock market, for example. It is comprised of thousands of different public companies. We could, of course, look at the stock value of each of these companies to see how the companies are doing as a whole, or we can look at just one number, the stock index, to get a general feel for how the companies are doing.

The same goes for the cost of goods. We could look at the cost of each item and compare it to its cost from last year. But that would mean looking at the cost of millions of items. Or we could look at the cost of goods index, just one number, to see whether prices have increased or decreased over the past year.

We can say that the index number is one simple number that we can look at to give us a general overview of what is happening in our field. Let's take a look at two real world index numbers.

Line of Best Fit (Least Square Method)

A **line of best fit** is a straight line that is the best approximation of the given set of data.
It is used to study the nature of the relation between two variables. (We're only considering the two-dimensional case, here.)

A line of best fit can be roughly determined using an eyeball method by drawing a straight line on a scatter plot so that the number of points above the line and below the line is about equal (and the line passes through as many points as possible).

A more accurate way of finding the line of best fit is the **least square method**.
Use the following steps to find the equation of line of best fit for a set of ordered pairs $(x1,y1),(x2,y2),...(x_n,y_n)$(x1,y1),(x2,y2),...(xn,yn).

Step 1: Calculate the mean of the xx-values and the mean of the yy-values.
$\overline{X}=\sum i=1nx_inY=\sum i=1ny_inX^-=\sum i=1nxinY^-=\sum i=1nyin$

Step 2: The following formula gives the slope of the line of best fit:
$m=\sum i=1n(x_i-\overline{X})(y_i-\overline{Y})\sum i=1n(x_i-\overline{X})_2m=\sum i=1n(xi-X^-)(yi-Y^-)\sum i=1n(xi-X^-)2$

Step 3: Compute the yy-intercept of the line by using the formula:
$b=\overline{Y}-m\overline{X}b=Y^--mX^-$

Step 4: Use the slope mm and the yy-intercept bb to form the equation of the line.