STATISTICS FOR MANAGEMENT

MBA-I SEMESTER

IARE-R 18

Prepared
by
Ms. G. Joseph Mary, Assistant Professor

INTRODUCTION TO STATISTICS

ORIGIN OF THE WORD "STATISTICS"

- Derived from Latin *statisticum collegium* ("council of state")
- Italian word *statista* ("statesman" or "politician")
- German book *Statistik*, published in 1749.
- The analysis of demographic and economic data about the state (*polit- ical arithmetic* in English)
- Was broadened in 1800s to include the collection, summary, and analysis of data of any type; also was conjoined with probability for the purpose of statistical inference.
- Recently, some believe this word is holding back our field

ORIGIN OF THE WORD "STATISTICS"

- 5th century B.C. — Athenians estimated the height of lad- ders necessary to scale the walls of Platea by having multiple soldiers count the bricks, then multiplying the most frequent count (the mode) by the height of a brick.
- Al-Kindi (801-873 A.D.) wrote "Manuscript on Deciphering Cryptographic Messages" which showed how to use frequency analysis to decipher encrypted messages.
- John Graunt in *Natural and Political Observations Made Upon the Bills of Mortality* estimated the population of London in 1662 as 384, 000 using records on the number of funerals per year (13, 000), the death rate (3 people per 11 families per year), and average family size (8).

CENTRAL TENDENCY

Median

1. Originated in Edward Wright's 1599 book, Certaine Er- rors in Navigation, concerning the determination of a lo- cation with a compass.
2. Further advocated by Ruder Boskovic in his 1755 book on the shape of the earth, in which he showed that the median minimizes the sum of absolute deviations.
3. The term "median" was coined by Galton in 1881.

CENTRAL TENDENCY

Mean

- The mean of two numbers was a well-known concept to the ancient Greeks
- Extended to more than two numbers in the late 1500s in order to estimate the locations of celestial bodies.
- Shown to minimize the sum of squared deviations by Thomas Simpson in 1755

ORIGIN OF THE WORD "STATISTICS"

- Introduced by Adrien Legendre in 1805
- By 1825 it was a standard tool in astronomy and geodesy
- The dominant theme of mathematical statistics (called "the combination of observations") in the 1800s
- The topic of the first statistics course at U of Iowa (1895 or earlier)

ORIGIN OF THE WORD "STATISTICS"

- 1911: University College, London (Karl Pearson's department)
- 1918: Johns Hopkins Department of Biometry and Vital Statistics
- 1931: University of Pennsylvania Department of Economic and Social Statistics
- 1933: Iowa State Statistical Laboratory
- 1935: George Washington Statistics Department (first in a College of Liberal Arts and/or Sciences)

ORIGIN OF THE WORD "STATISTICS"

- North Carolina State University (41 faculty, some joint with other departments)
- Iowa State University (40 faculty, some joint with other de-partments)
- Texas A&M University (27 faculty)

These departments generally have 4-8 "lecturers," "instructors," or "teaching professors" as well.

By comparison, U of Iowa has 15 faculty (3-4 are actuarial science rather than statistics faculty) and 6 lecturers.

DEFINITION

Statistics:

It is a mathematical science including methods of collecting, organizing and analyzing data in such a way that meaningful conclusions can be drawn from them. In general, its investigations and analyses fall into two broad categories called descriptive and inferential statistics.

ORIGIN OF STATISTICS

- Statistics, in a sense, is as old as the human society itself. Its origin can be traced to the old days when it was regarded as the 'science of State-craft' and was the by product of the administrative activity of the State. The word 'Statistics' seems to have been derived from the Latin word 'status' or the Italian word 'statista' or the German word 'statistik' each of which means a 'political state'.

DEVELOPMENT OF STATISTICS

- In ancient times, the government used to collect the information regarding the population and 'property of wealth' of the country- the former enabling the government to have an idea of the manpower of the country and the latter providing it a basis for introducing news taxes and levies

SCOPE AND IMPORTANCE OF STATISTICS

1. Statistics and economics
2. Statistics and planning
3. Statistics and business
4. Statistics and industry
5. Statistics and mathematics
6. Statistics and modern science
7. Statistics, psychology and education
8. Statistics and war

STATISTICS IN BUSINESS AND MANAGEMENT

1. Marketing
2. Production
3. Finance
4. Banking
5. Investment
6. Purchase
7. Accounting
8. Control

DESCRIPTIVE STATISTICS

It 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

INFERENTIAL STATISTICS

Inferential statistics as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

LIMITATIONS OF STATISTICS

1. Qualitative Aspect Ignored
2. It does not deal with individual items
3. It does not depict entire story of phenomenon
4. It is liable to be miscued
5. Laws are not exact
6. Results are true only on average

MEASURES OF CENTRAL TENDENCY

Measures of Central Tendency

- 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.

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 . 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 (pronounced x bar).

MEDIAN

- The median is the value at the middle of a distribution of data when those data are organized from the lowest to the highestvalue.
- This measure of central tendency can be calculated for variables that are measured with ordinal, interval or ratio scales.

MODE

- The mode is the measure of central tendency that identifies the category or score that occurs the most frequently within the distribution of data. In other words, it is the most common score or the score that appears the highest number of times in a distribution. The mode can be calculated for any type of data, including those measured as nominal variables, or by name.

GEOMETRIC MEAN

- In mathematics, the **geometric mean** is a type of mean or average which indicates the central tendency or typical value of a set of numbers by using the product of their values The geometric mean is defined as the *n*th root of the product of *n* numbers, i.e., for a set of numbers $x_1, x_2, \ldots, x_n$

HARMONIC MEAN

- In mathematics, the harmonic mean (sometimes called the sub contrary mean) is one of several kinds of average, and in particular one of the Pythagorean means. ... The harmonic mean can be expressed as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations.

CENTRAL DISPERSION

- It is a quantifiable variation of measurements of differing members of a population

1. Range
2. Quartile deviation
3. Co Efficient of Quartile deviation
4. Mean Deviation
5. Co Efficient of Mean Deviation

RANGE

- The Range is the difference between the lowest and highest values.

Range=Q1-Q2

Q1=highest value in the series

Q2=Lowest value in the series

QUARTILE DEVIATION

- The **quartile deviation** is a slightly better measure of absolute dispersion than the range, but it ignores the observations on the tails. If we take difference samples from a population and calculate their **quartile deviations**, their values are quite likely to be sufficiently different.

MEAN DEVIATION

- The average of these numbers (6 , 5) is 1.2 which is the mean deviation.
- Also called mean absolute deviation, it is used as a measure of dispersion where the number of values or quantities is small, otherwise standard deviation is used.

QUARTILE DEVIATION

- Quartile deviation is based on the lower quartile $Q1$ and the upper quartile $Q3$. The difference $Q3-Q1$ is called the inter quartile range.
- The difference $Q3-Q1$ divided by 2is called semi-inter-quartile range or the quartile deviation.
- Thus

$$Q.D=Q3-Q1/2$$

COEFFICIENT OF QUARTILE DEVIATION

- A relative measure of dispersion based on the quartile deviation is called the coefficient of quartile deviation. It is defined as:
- Coefficient of Quartile Deviation
  - Coefficient of Quartile Deviation
    - $=Q3-Q1/2Q3+Q1/2$
    - $=Q3-Q1/Q3+Q1$

CO-EFFICIENT OF VARIATION SKEWNESS

- Measures of Skewness and Kurtosis, like measures of central tendency and dispersion, study the characteristics of a frequency distribution.
- Averages tell us about the central value of the distribution and measures of dispersion tell us about the concentration of the items around a central value.
- Thus, skewness is a measure that studies the degree and direction of departure from symmetry.

NAUTURE OF SKEWNESS

Skewness can be positive or negative or zero.

1. When the values of mean, median and mode are equal, there is no skewness.
2. When mean > median > mode, skewness will be positive.
3. When mean < median < mode, skewness will be negative.

Mathematical measures of skewness

- It can be calculated by:
  - **Bowley's Method:** Bowley's method of skewness is based on the values of median, lower and upper quartiles.
  - **Karl-Pearson's Method** : Coefficient of skewness lies within the limit ± 1.
  - **Kelly 's method:** This method is quite convenient for determining skewness

Mathematical measures of skewness

Bowley's Method:

This method suffers from the same limitations which are in the case of median and quartiles. Wherever positional measures are given, skewness should be measured by Bowley's method.

- This method is also used in case of 'open-end series', where the importance of extreme values is ignored.
- Absolute skewness = Q3 + Q1 – 2 Median

Karl-Pearson's Method (Personian Coefficient of Skewness)

- where the relationshipof mean and mode is established;
- where the relationship between meanand median is not established.
- It does not show the extent to which deviations cluster below an average or above it. Skewness tells us about the cluster of the deviations above and below a measure of central tendency.

Comparison among dispersion, skewness and kurtosis

- Dispersion, Skewness and Kurtosis are different characteristics of frequency distribution.
- Kurtosis studies the concentration of the items at the central part of a series.
- If items concentrate too much at the centre, the curve becomes ‘LEPTOKURTIC’ and if the concentration at the centre is comparatively less.
- the curve becomes ‘PLATYKURTIC’.

TABULATION OF UNIVARIATE

UNIVARIATE DATA

- Uni variate data is used for the simplest form of analysis. It is the type of data in which analysis are made only based on one variable.
- For example, there are sixty students in class VII. If the variable marks obtained in math were the subject, then in that case analysis will be based on the number of subjects fall into defined categories of marks.

BIVARIATE DATA

- Bi variate data is used for little complex analysis than as compared with uni variate data.
- Bi variate data is the data in which analysis are based on two variables per observation simultaneously.

MULTIVARIATE DATA

- Multivariate data is the data in which analysis are based on more than two variables per observation. Usually multivariate data is used for explanatory purposes.

Scatter Diagram

Advantages of Scatter Diagram

- Simple & Non Mathematical method
- Not influenced by the size of extreme item
- First step in investing the relationship between two variables

TABULATION

It is the process of condensation of the data for convenience, in statistical processing, presentation and interpretation of the information .

A good table is one which has the following requirements :

- It should present the data clearly, highlighting important details.
- It should save space but attractively designed.
- The table number and title of the table should be given.+
- Row and column headings must explain the figures therein.
- Averages or percentages should be close to the data.
- Units of the measurement should be clearly stated along the titles or headings.

CLASSIFICATION

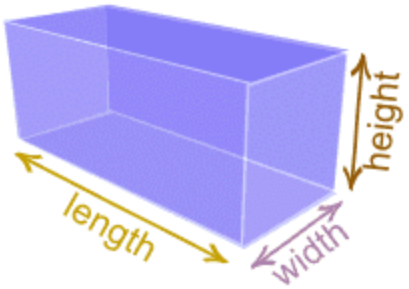
- "Classified and arranged facts speak of themselves, and narrated they are as dead as mutton" This quote is given by J.R. Hicks.
- The process of dividing the data into different groups (viz. classes) which are homogeneous within but heterogeneous between themselves, is called a classification.
- It helps in understanding the salient features of the data and also the comparison with similar data. For a final analysis it is the best friend of a statistician.

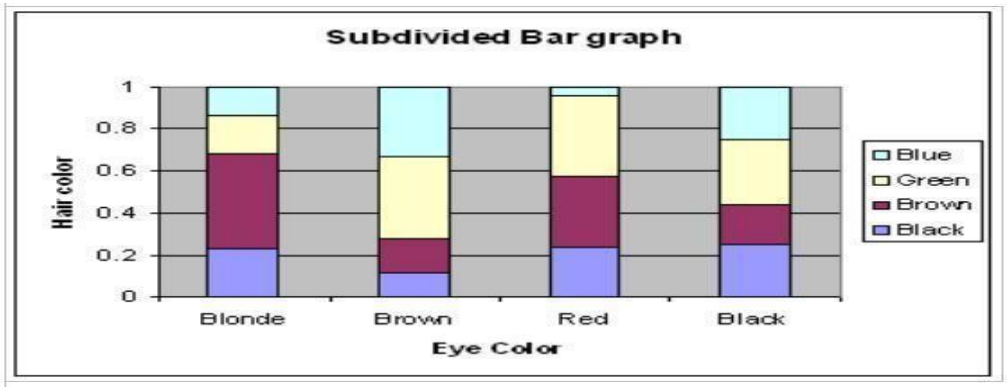
METHOD OF CLASSIFICATION

The data is classified in the following ways :

According to attributes or qualities this is divided into two parts :

1. Simple classification
2. Multiple classification.





DIAGRAMMATICAND GRAPHIC DISPLAYS

1. **Simple 'Bar diagram':-**

It represents only one variable. For example sales, production, population figures etc. for various years may be shown by simple bar charts.

Since these are of the same width and vary only in heights (or lengths), it becomes very easy for readers to study the relationship. Simple bar diagrams are very popular in practice. A bar chart can be either vertical or horizontal; vertical bars are more popular.

DIAGRAMMATICAND GRAPHIC DISPLAYS

2. Sub - divided Bar Diagram:-

- While constructing such a diagram, the various components in each bar should be kept in the same order.
- A common and helpful arrangement is that of presenting each bar in the order of magnitude with the largest component at the bottom and the smallest at the top.
- The components are shown with different shades or colors with a proper index.

3. **Multiple Bar Diagram:-**

- This method can be used for data which is made up of two or more components. In this method the components are shown as separate adjoining bars.
- The height of each bar represents the actual value of the component.
- The components are shown by different shades or colors. Where changes in actual values of component figures only are required, multiple bar charts are used.

4. **Deviation Bar Charts:-**

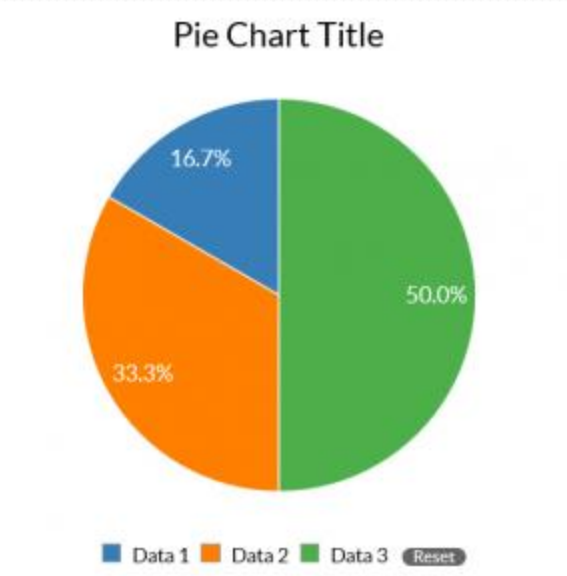
Deviation bars are used to represent net quantities - excess or deficit i.e. net profit, net loss, net exports or imports, swings in voting etc. Such bars have both positive and negative values. Positive values lie above the base line and negative values lie below it.

DIAGRAMMATICAND GRAPHIC DISPLAYS

PIE CHART: Geometrically it can be seen that the area of a sector of a circle taken radically, is proportional to the angle at its center.

It is therefore sufficient to draw angles at the center, proportional to the original figures.

This will make the areas of the sector proportional to the basic figures.



GRAPHIC DISPLAYS

- A graph is a visual representation of data by a continuous curve on a squared (graph) paper. Like diagrams, graphs are also attractive, and eye-catching, giving a bird's eye-view of data and revealing their inner pattern.
- Graphs of Frequency Distributions:-

  The methods used to represent a grouped data are :

  1. Histogram
  2. Frequency Polygon
  3. Frequency Curve
  4. Cumulative Frequency Curve

HISTOGRAM

- It is defined as a pictorial representation of a grouped frequency distribution by means of adjacent rectangles, whose areas are proportional to the frequencies.
- To construct a Histogram, the class intervals are plotted along the x-axis and corresponding frequencies are plotted along the y - axis. The rectangles are constructed such that the height of each rectangle is proportional to the frequency of the that class.

FREQUENCY POLYGON

- Here the frequencies are plotted against the mid- points of the class-intervals and the points thus obtained are joined by line segments.

FREQUENCY DISTRIBUTION (CURVE):-

Frequency distribution curves are like frequency polygons. In frequency distribution, instead of using straight line segments, a smooth curve is used to connect the points.

OGIVES OR CUMULATIVE FREQUENCY CURVES

- When frequencies are added, they are called cumulative frequencies. The curve obtained by plotting cumulating frequencies is called a cumulative frequency curve or an ogive (pronounced ogive).

ONE DIMENSIONAL GRAPHS

- 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

TWO DIMENSIONAL GRAPHS

- A **two-dimensional graph** is a set of points in two-dimensional space .
- If the points are real and if Cartesian coordinates are used, each axis depicts the potential values of a particular real variable.
- Often the variable on the horizontal axis is called *x* and the one on the vertical axis is called *y*, in which case the horizontal and vertical axes are sometimes called the *x* axis and *y* axis respectively.
- With real variables on the axes, each point in the graph depicts the values of two real variables.

THREE DIMENSIONAL GRAPHS

- Graphs play a very important role in mathematics.
- Agraph is also known as a plot.
- It is said to be a function graph or plot of a function in elementary mathematics.
- A graph is defined as a collection of lines joining a set of points on a graph paper.
- These points are also known as graphical points or vertices.
- They may also be referred as points or nodes.

SMALL SAMPLE TESTS

T distribution

- The (also called **Student's T Distribution**) is a family of distributions that look almost identical to the normal distribution curve, only a bit shorter and fatter.

- The t distribution is used instead of the normal distribution when you have small samples (for more on this, see: t-score vs. z-score). The larger the sample size, the more the t distribution looks like the normal distribution.

- In fact, for sample sizes larger than 20 (e.g. more degrees of freedom), the distribution is almost exactly like the normal distribution.

Hypothesis Tests for One or Two Means

- When testing a claim about the value of a population mean, the test statistic will depend on whether the population standard deviation is known or unknown. This situation is identical to finding a confidence interval for a mean, and is resolved in exactly the same way.

Testing a Single Mean with Known Population Standard Deviation

1. If σ n is known, then for a large enough sample, the distribution of sample means will be approximately normal. Or more specifically, we can expect an approximate normality in the following two cases.
2. σ is known, and the sample size is at least 30 (for any population) σ is known, and the original population is normal (for any value of n)
3. In this situation, the test statistic will follow the standard normal distribution, and its formula is

$$z=x^{-}-\mu 0\sigma/n--\sqrt{}$$

Testing a Single Mean with Unknown Population Standard Deviation

- There are also two cases for which a hypothesis test of a mean can be done when σ is unknown.
- In these cases, for a large enough sample, the distribution of sample means will follow a t-distribution.
- we can expect a t- distribution in the following two cases.
- Σ is unknown, and the sample size is at least 30 (for any population) σ is unknown, and the original population is normal (for any value of n)
- In these two cases, the test statistic will follow a t-distribution with $n-1$ degrees of freedom, and its formula is

$$t=x^{-}-\mu 0s/n--\sqrt{}$$

Testing the Difference of Means from a Paired Sample

- A paired sample occurs when the data are collected from the same individual at two different points in time, or on two different tasks, or some other fashion in which the values will be connected. The actual test is done on the differences, denoted *d*
- These differences may have a known or an unknown population standard deviation, resulting in two cases analogous to those already described. The sample size and normality requirements apply to the differences.

Therefore, the test statistic formulas are:

- If σ is known, $z=\bar{d}-d0\sigma d/n--\sqrt{}$ If σ is unknown, $t=\bar{d}-d0sd/n--\sqrt{}$,with $n-1$ degrees of freedom

ANOVA

- An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis .Basically, you're testing groups to see if there's a difference between them. Examples of when you might want to test different groups.

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.

ONE WAY ANOVA

- A one way ANOVA is used to compare two means from two independent (unrelated) groups using the F- distribution The null hypothesis for the test is that the two means are equal. Therefore, a significant result means that the two means are unequal.

ONE WAY ANOVA

- **Situation 1:** You have a groupof individuals randomly split into smaller groups and completing different tasks. For example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea.

- **Situation 2:** Similar to situation 1, but in this case the individuals are split into groups based on an attribute they possess. For example, you might be studying leg strength of people according to weight. You could split participants into weight categories (obese, overweight and normal) and measure their leg strength on a weight machine.

TWO WAY ANOVA

- A Two Way ANOVA is an extension of the One Way ANOVA. With a One Way, you have one independent variable affecting a dependent variable With a Two Way ANOVA, there are two independents.
- Use a two way ANOVA when you have one measurement variable (i.e. a quantitative variable) and two nominal variables . In other words, if your experiment has a quantitative outcome and you have two categorical explanatory variables.

CHI-SQUARE DISTRIBUTION

- The distribution of the chi-square statistic is called the chi-square distribution. In this lesson, we learn to compute the chi-square statistic and find the probability associated with the statistic. Chi-square examples illustrate key points.

CORRELATION

Methods of studying correlation

1. Scatter diagram
2. Karl pearson's coefficient of correlation
3. Spearman's Rank correlation coefficient
4. Method of least squares

CORRELATION

- **Correlation:** The degree of relationship between the variables under consideration is measure through the correlation analysis.
- The measure of correlation called the correlation coefficient
- The degree of relationship is expressed by coefficient which range from correlation **(-1 ≤ r ≥ +1)**
- The direction of change is indicated by a sign.

CORRELATION

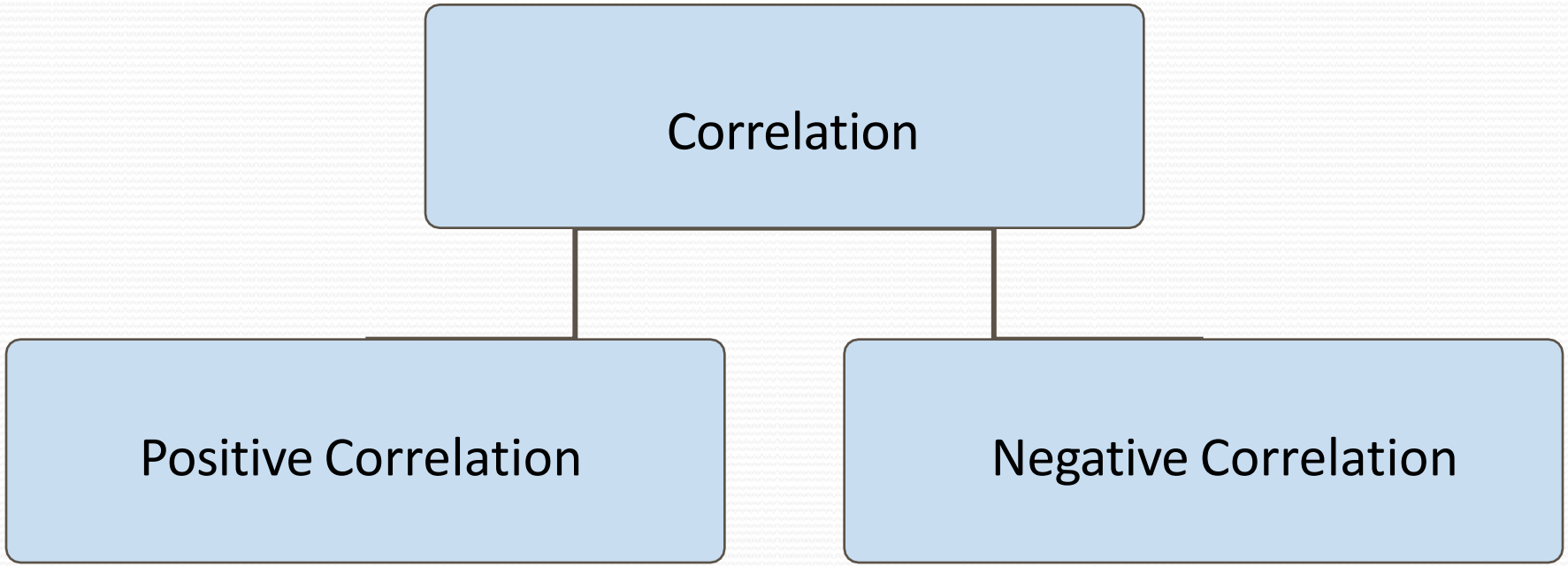
- Correlation is a statistical tool that helps to measure and analyze the degree of relationship between two variables.
- Correlation analysis deals with the association between two or more variables.

CORRELATION

Causation means cause & effect relation.

- Correlation denotes the interdependency among the variables for correlating two phenomenon, it is essential that the two phenomenon should have cause-effect relationship,& if such relationship does not exist then the two phenomenon can not be correlated.
- If two variables vary in such a way that movement in one are accompanied by movement in other, these variables are called cause and effect relationship.
- Causation always implies correlation but correlation does not necessarily implies causation.

CORRELATION



CORRELATION

- **Positive Correlation:** The correlation is said to be positive correlation if the values of two variables changing with same direction.

  Ex. Pub. Exp. & sales, Height & weight.

- **Negative Correlation:** The correlation is said to be negative correlation when the values of variables change with opposite direction.

  Ex. Price & qty. demanded.

Direction of the Correlation

1. **Positive relationship** – Variables change in the same direction.
   - As X is increasing, Y is increasing
   - As X is decreasing, Y is decreasing

   E.g., As height increases, so does weight.

2. **Negative relationship** – Variables change in opposite directions.
   - As X is increasing, Y is decreasing
   - As X is decreasing, Y is increasing

   E.g., As TV time increases, grades decrease

CORRELATION

- **Simple correlation:** Under simple correlation problem there are only two variables are studied.
- **Multiple Correlation:** Under Multiple Correlation three or more than three variables are studied.

$$\text{Ex. } Q_d = f\ (\ P, P_C, P_S, t,\ y\ )$$

- **Partial correlation:** analysis recognizes more than two variables but considers only two variables keeping the other constant.
- **Total correlation:** is based on all the relevant variables, which is normally not feasible.

CORRELATION

Types of Correlation Type III

- **Linear correlation:** Correlation is said to be linear when the amount of change in one variable tends to bear a constant ratio to the amount of change in the other. The graph of the variables having a linear relationship will form a straight line.

  Ex X = 1, 2, 3, 4, 5, 6, 7, 8,
  Y = 5, 7, 9, 11, 13, 15, 17,
  19, Y = 3 + 2x

- **Non Linear correlation:** The correlation would be non linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

Methods of Studying Correlation

- **Scatter Diagram Method**: a graph of observed plotted points where each points represents the values of X & Y as a coordinate.
- It portrays the relationship between these two variables graphically.
- Graphic Method
- Karl Pearson's Coefficient of Correlation
- Method of Least Squares

Karl Pearson's Coefficient of Correlation

- Pearson's 'r' is the most common correlation coefficient.
- Karl Pearson's Coefficient of Correlation denoted by- 'r' The coefficient of correlation 'r' measure the degree of linear relationship between two variables say x & y.
- Karl Pearson's Coefficient of Correlation denoted by-

$$-1 \leq \quad r \geq +1$$

- Degree of Correlation is expressed by a value of Coefficient
- Direction of change is Indicated by sign (- ve) or

  (+ ve)

Procedure for computing The correlation coefficient

- Calculate the mean of the two series ‘x’ &’y’
- Calculate the deviations ‘x’ &’y’in two series from their respective mean.
- Square each deviation of ‘x’ &’y’then obtain the sum of the squared deviation i.e.$\sum x^2$ &.$\sum y^2$
- Multiply each deviation under x with each deviation under y & obtain the product of ‘xy’.Then obtainthe sum of the product of x , y i.e. $\sum xy$
- Substitute the value in the formula.

Interpretation of correlation coefficient

- The value of correlation coefficient ‘r’ ranges from -1 to +1
- If r = +1, then the correlation between the two variables is said to be perfect and positive
- If r = -1, then the correlation between the two variables is said to be perfect and negative
- If r = 0, then there exists no correlation between the variables

Properties of Correlation coefficient

- The correlation coefficient lies between -1 &+1 symbolically (- 1≤ r ≥ 1)
- The correlation coefficient is independent of the change of origin & scale.
- The coefficient of correlation is the geometric mean of two regression coefficient . r = √ bxy * byx
- The one regression coefficientis (+ve) coefficient is also (+ve) correlation coefficient is (+ve)

Karl Pearson's Coefficient of Correlation

Limitation of Pearson's Coefficient:

- Always assume linear relationship
- Interpreting the value of 'r' is difficult.
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming methods

Karl Pearson's Coefficient of Correlation

- The convenient way of interpreting the value of correlation coefficient is to use of square of coefficient of correlation which is called Coefficient of Determination.
- The Coefficient of Determination = r^2.
- Suppose: r = 0.9, r^2 = 0.81 this would mean that 81% of the variation in the dependent variable has been explained by the independent variable.

Spearman's Rank Coefficient of Correlation

When statistical series in which the variables under study are not capable of quantitative measurement but can be arranged in serial order, in such situation pearson's correlation coefficient can not be used in such case Spearman Rank correlation can be used.

- **$R = 1 - (6 \sum D^2) / N (N^2 - 1)$**
- R = Rank correlation coefficient
- D = Difference of rank between paired item in two series.
- N = Total number of observation.

Interpretation of Rank Correlation Coefficient (R)

- The value of rank correlation coefficient, R ranges from -1 to +1
- If R = +1, then there is complete agreement in the order of the ranks and the ranks are in the same direction
- If R = -1, then there is complete agreement in the order of the ranks and the ranks are in the opposite direction
- If R = 0, then there is no correlation

Spearman's Rank Coefficient of Correlation

1. **Problems where actual rank are given:**
   - Calculate the difference 'D' of two Ranks i.e.
   - (R1 – R2).
   - Square the difference & calculate the sum of the difference i.e.
   - $\sum D^2$ Substitute the values obtained in the formula.
2. **Problems where Ranks are not given :**
   - If the ranks are not given, then we need to assign ranks to the data series.
   - The lowest value in the series can be assigned rank 1 or the highest value in the series can be assigned rank 1.

Spearman's Rank Coefficient of Correlation

Rank Correlation Coefficient (R)

- **Equal Ranks or tie in Ranks:** In such cases average ranks should be assigned to each individual.
- $R = 1- (6 \sum D^2) + AF / N (N^2 – 1)$
- $AF = 1/12(m^3 – m) + 1/12(m^3 – m) +….$
- $1/12(m^3 – m)$

m = The number of time an item is repeated

Karl Pearson's Coefficient of Correlation

Merits Spearman's Rank Correlation

- This method is simpler to understand and easier to apply compared to Karl Pearson's correlation method.
- This method is useful where we can give the ranks and not the actual data. (qualitativeterm)
- This method is to use where the initial data in the form of ranks.

Advantages of correlation studies

- Show the amount (strength) of relationship present
- Can be used to make predictions about the variables under study.
- Can be used in many places, including natural settings, libraries, etc.
- Easier to collect co relational data

REGRESSION ANALYSIS

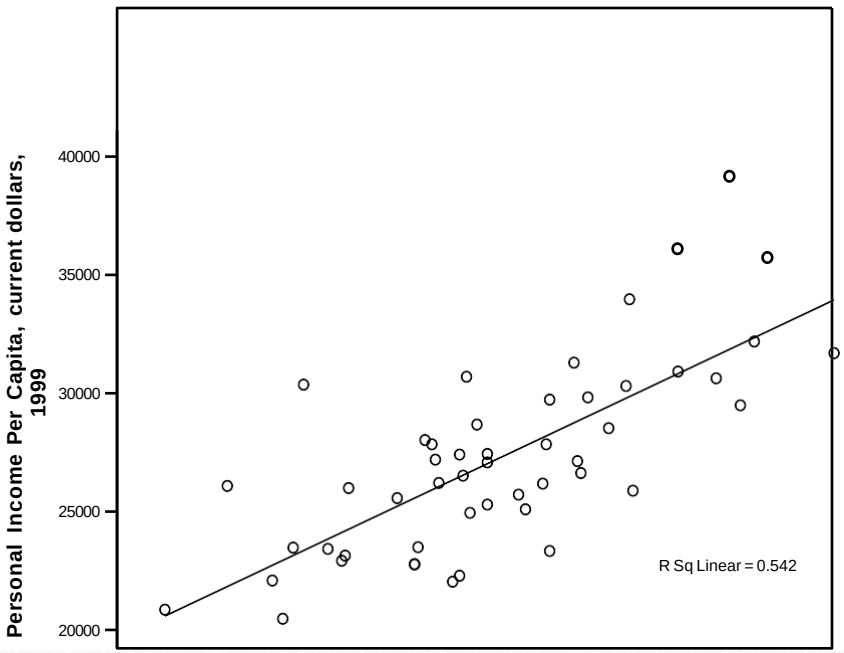
Regression analysis

- Regression analysis requires interval and ratio-level data.
- To see if your data fits the models of regression, it is wise to conduct a scatter plot analysis.
- The reason?

  Regression analysis assumes a linear relationship. If you have a curvilinear relationship or no relationship, regression analysis is of little use.

Regression analysis

Percent of Population with Bachelor's Degree by Personal Income Per Capita



- Regression line is the best straight line description of the plotted points and use can use it to describethe association between the variables.
- If all the lines fall exactly on the line then the line is 0 and you have a perfect relationship.

Regression analysis

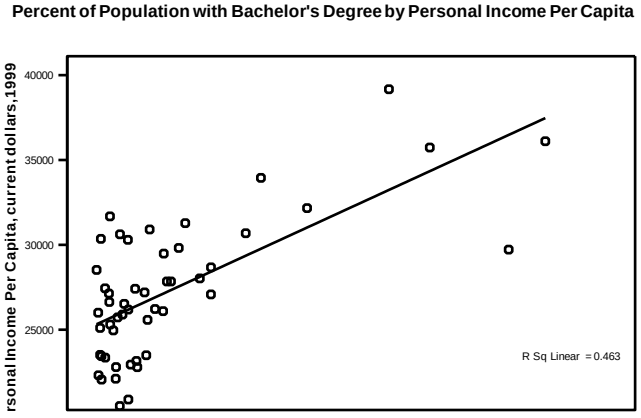
- Regressions are still focuses on association, not causation.
- Association is a necessary prerequisite for inferring causation, but also:

1. The independent variable must preceded the dependent variable in time.
2. The two variables must be plausibly lined by a theory,
3. Competing independent variables must be eliminated.

Regression analysis

- The regression coefficient is not a good indicator for the strength of the relationship.

- Two scatter plots with very different dispersions could produce the same regression line.



Regression analysis

- To determine strength you look at how closely the dots are clustered around the line. The more tightly the cases are clustered, the stronger the relationship, while the more distant, the weaker.
- Pearson's r is given a range of -1 to + 1 with 0 being no linear relationship at all.

Regression co efficient

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10078.565	2312.771		4.358	.000
	Percent of Population 25 years and Over with Bachelor's Degree or More, March 2000 estimates	688.939	91.433	.736	7.535	.000

a. Dependent Variable: Personal Income Per Capita, current dollars, 1999

- B - These are the values for the regression equation for predicting the dependent variable from the independent variable.
- These are called unstandardized coefficients because they are measured in their natural units. As such, the coefficients cannot be compared with one another to determine which one is more influential in the model, because they can be measured on different scales.

Part of the regression equation

b represents the slope of the line

- It is calculated by dividing the change in the dependent variable by the change in the independent variable.
- The difference between the actual value of Y and the calculated amount is called the residual.
- The represents how much error there is in the prediction of the regression equation for the y value of any individual case as a function of X.

Comparing two variables

Regression analysis is useful for comparing two Variables to see whether controlling for other independent variable affects your model.

1. For the first independent variable, education, the Argument is that a more educated populace will have higher-paying jobs, producing a higher level of per capita income in the state.
2. These continued in dependent variable is included because we expect to find better-paying state residents to obtain them, in urban rather jobs, and therefore more opportunity for hand rural areas.

INDEX NUMBERS

Index numbers:

- Index numbers are time series that focus on the **relative** change in a count or measurement over time.
- express the count or measurement as a percentage of the comparable count or measurement in a **base period**.

INDEX NUMBERS

- The base period is arbitrary but should be a convenient point of reference.
- Thevalue of an index number corresponding to the base period is always 100.
- The base period may be a single period or an average of multiple adjacent periods.

APPLICATIONS OF INDEX NUMBERS

- A **price index** shows the change in the price of a commodity or group of commodities over time.
- A **quantity index** shows the change in quantity of a commodity or group of commodities used or purchased over time.
- A **value index** shows a change in total dollar value (price • quantity) of a commodity or group of commodities over time.

SIMPLE RELATED INDEX

- A simple relative index shows the change in the price, quantity, or value of a single commodity over time.
- Calculation of a simple relative index:

$$\text{Index in period } t = \left(\text{Measurement in period } t \,/\, \text{Measurement in base period}\right) * 100$$

SIMPLE RELATED INDEX

Year	Price	Price Index 1980 as base year	Price Index 1990 as base year
1980	$140	100.0	58.3
1985	195	139.3	81.3
1990	240	171.4	100.0
1995	275	196.4	114.6

Computation of index for 1985 (1980 as base year):

$$I = \frac{P_t}{P_0} \cdot 100 = \frac{195}{140} \cdot 100 = 139.3$$

SIMPLE AGGREGATE INDEX

- A simple aggregate index shows the change in the prices, quantities, or values of a **group** of related items. Each item in the group is treated as having equal weight for purposes of comparing group measurements over time.
- Calculation of index number:

$$\text{Index in Period } t = \frac{\text{(Sum of measurements for all items in period } t\text{)}}{\text{(Sum of measurements for all items in base period)}}$$

SIMPLE AGGREGATE INDEX

Yr	Cars Sold	Trucks Sold	for Cars & Trucks Sold (1995 as base yr)
1994	423	141	94.0
1995	435	165	100.0
1996	440	184	104.0
1997	455	215	111.7

Illustration of computation of index for 1994:

$$I = \frac{\sum Q_t}{\sum Q_0} \cdot 100 = \frac{423 + 141}{435 + 165} \cdot 100 = 94.0$$

WEIGHTED AGGREGATE INDEX

To make sure that the changes indicated by the index numbers focus on the aspect of interest (e.g., price or quantity), the same weighting factors must be used to aggregate measurements in the selected period and the base period.

- In weighted aggregate price indexes, the corresponding quantities are often used as weighting factors.
- In weighted aggregate quantity indexes.
- the corresponding prices are often used as weighting factors.

WEIGHTED AGGREGATE INDEX

- A weighted aggregate price index where the quantities of the items used in the period of interest are used as weighting factors.
- Calculation of index for period *t*:

$$I = \frac{\sum P_t \cdot Q_t}{\sum P_0 \cdot Q_t} \cdot 100$$

where

P_t = price of item in period *t*

P_0 = price of item in base period

Q_t = quantity of item in period *t*

- Note that quantities in period *t* are used to determine the weighted sum of base period prices.

PAASCHE INDEX

Paasche index for airline tickets for 1997 using base year of 1990:

Product	Price, 1990	Price, 1997	Quantity, 1997
Coach	$380	$430	181,000
First Class	725	940	14,000

Computation of index for 1997:

$$= \frac{(430)\cdot(181{,}000) + (940)\cdot(14{,}000)}{(380)\cdot(181{,}000) + (725)\cdot(14{,}000)} \cdot 100$$

LASPEYRES INDEX

- A weighted aggregate price index where the quantities of the items used in the base period are used as weighting factors.

Calculation of the Laspeyres index for period *t*

where

P_0 = price of item in base period

Q_0 = quantity of item in base year

CONSUMER PRICE INDEX

1. A weighted aggregate price index used to reflect the overall change in the cost of goods and services purchased by a typical consumer.

2. Applications:
   - Indicator of rate of inflation
   - Used to adjust wages to compensate for lost purchasing power due to inflation
   - Used to convert a price or wage to a real price or real wage to show the equivalent amount in a base period after adjusting for inflation.

SHIFTING BASEINDEX

- For useful interpretation, it is often desirable for the base year to be fairly recent.
- To shift the base year to another year without recalculating the index from the original data:

$$\text{Index for year } t \text{ in new base year} = \frac{\text{Index for year } t \text{ relative to old base year}}{\text{Index for new base year relative to old base year}} \times 100$$