

## 1.1 What is Statistics

The word 'Statistics' is derived from the Latin word 'Statis' which means a "political state." Clearly, statistics is closely linked with the administrative affairs of a state such as facts and figures regarding defense force, population, housing, food, financial resources etc. What is true about a government is also true about industrial administration units, and even one's personal life.

The word statistics has several meanings. In the first place, it is a plural noun which describes a collection of numerical data such as employment statistics, accident statistics, population statistics, birth and death, income and expenditure, of exports and imports etc. It is in this sense that the word 'statistics' is used by a layman or a newspaper.

Secondly the word statistics as a singular noun, is used to describe a branch of applied mathematics, whose purpose is to provide methods of dealing with a collections of data and extracting information from them in compact form by tabulating, summarizing and analyzing the numerical data or a set of observations.

The various methods used are termed as statistical methods and the person using them is known as a statistician. A statistician is concerned with the analysis and interpretation of the data and drawing valid worthwhile conclusions from the same.

It is in the second sense that we are writing this guide on statistics.

Lastly the word statistics is used in a specialized sense. It describes various numerical items which are produced by using statistics ( in the second sense ) to statistics ( in the first sense ). Averages, standard deviation etc. are all statistics in this specialized third sense.

The word 'statistics' in the first sense is defined by Professor Secrit as follows:-

"By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."

This definition gives all the characteristics of statistics which are (1) Aggregate of facts (2) Affected by multiplicity of causes (3) Numerically expressed (4) Estimated according to reasonable standards of accuracy (5) Collected in a systematic manner (6) Collected for a predetermined purpose (7) Placed in relation to each other.

The word 'statistics' in the second sense is defined by Croxton and Cowden as follows:-

"The collection, presentation, analysis and interpretation of the numerical data."

This definition clearly points out four stages in a statistical investigation, namely:

- 1) Collection of data
- 2) Presentation of data
- 3) Analysis of data
- 4) Interpretation of data

## 1.2 Uses

To present the data in a concise and definite form : Statistics helps in classifying and tabulating raw data for processing and further tabulation for end users.

To make it easy to understand complex and large data : This is done by presenting the data in the form of tables, graphs, diagrams etc., or by condensing the data with the help of means, dispersion etc.

For comparison : Tables, measures of means and dispersion can help in comparing different sets of data..

In forming policies : It helps in forming policies like a production schedule, based on the relevant sales figures. It is used in forecasting future demands.

Enlarging individual experiences : Complex problems can be well understood by statistics, as the conclusions drawn by an individual are more definite and precise than mere statements on facts.

In measuring the magnitude of a phenomenon:- Statistics has made it possible to count the population of a country, the industrial growth, the agricultural growth, the educational level (of course in numbers).

### **Limitations**

Statistics does not deal with individual measurements. Since statistics deals with aggregates of facts, it can not be used to study the changes that have taken place in individual cases. For example, the wages earned by a single industry worker at any time, taken by itself is not a statistical datum. But the wages of workers of that industry can be used statistically. Similarly the marks obtained by John of your class or the height of Beena (also of your class) are not the subject matter of statistical study. But the average marks or the average height of your class has statistical relevance.

Statistics cannot be used to study qualitative phenomenon like morality, intelligence, beauty etc. as these can not be quantified. However, it may be possible to analyze such problems statistically by expressing them numerically. For example we may study the intelligence of boys on the basis of the marks obtained by them in an examination.

Statistical results are true only on an average:- The conclusions obtained statistically are not universal truths. They are true only under certain conditions. This is because statistics as a science is less exact as compared to the natural science.

Statistical data, being approximations, are mathematically incorrect. Therefore, they can be used only if mathematical accuracy is not needed.

Statistics, being dependent on figures, can be manipulated and therefore can be used only when the authenticity of the figures has been proved beyond doubt..

### **1.3 Distrust Of Statistics**

It is often said by people that, "statistics can prove anything." There are three types of lies - lies, demand lies and statistics - wicked in the order of their naming. A Paris banker said, "Statistics is like a miniskirt, it covers up essentials but gives you the ideas."

Thus by "distrust of statistics" we mean lack of confidence in statistical statements and methods. The following reasons account for such views about statistics.

Figures are convincing and, therefore people easily believe them.

They can be manipulated in such a manner as to establish foregone conclusions.

The wrong representation of even correct figures can mislead a reader. For example, John earned \$ 4000 in 1990 - 1991 and Jem earned \$ 5000. Reading this one would form the opinion that Jem is decidedly a better worker than John. However if we carefully examine the statement, we might reach a different conclusion as Jem's earning period is unknown to us.

Thus while working with statistics one should not only avoid outright falsehoods but be alert to detect possible distortion of the truth.

### **1.4 Types Of Statistics**

As mentioned earlier, for a layman or people in general, statistics means numbers - numerical facts, figures or information. The branch of statistics wherein we record and analyze observations for all the individuals of a group or population and draw inferences about the same is called "Descriptive statistics" or "Deductive statistics". On the other hand, if we choose a sample and by statistical treatment of this, draw inferences about the population, then this branch of statistics is known as Statical Inference or Inductive Statistics.

In our discussion, we are mainly concerned with two ways of representing descriptive statistics : Numerical and Pictorial.

Numerical statistics are numbers. But some numbers are more meaningful such as mean, standard deviation etc.

When the numerical data is presented in the form of pictures (diagrams) and graphs, it is called the Pictorial statistics. This statistics makes confusing and complex data or information, easy, simple and straightforward, so that even the layman can understand it without much difficulty.

### **1.4 Common Mistakes Committed In Interpretation of Statistics**

**Bias:-** Bias means prejudice or preference of the investigator, which creeps in consciously and unconsciously in proving a particular point.

**Generalization:-** Some times on the basis of little data available one could jump to a conclusion, which leads to erroneous results.

**Wrong conclusion:-** The characteristics of a group if attached to an individual member of that group, may lead us to draw absurd conclusions.

**Incomplete classification:-** If we fail to give a complete classification, the influence of various factors may not be properly understood.

There may be a wrong use of percentages.

Technical mistakes may also occur.

An inconsistency in definition can even exist.

Wrong causal inferences may sometimes be drawn.

There may also be a misuse of correlation.

## 1.7 Glossary of Terms

### Statistics :

Statistics is the use of data to help the decision maker to reach better decisions.

### Data :

It is any group of measurements that interests us. These measurements provide information for the decision maker. (i) The data that reflects non-numerical features or qualities of the experimental units, is known as qualitative data. (ii) The data that possesses numerical properties is known as quantitative data.

### Population:

Any well defined set of objects about which a statistical enquiry is being made is called a population or universe.

The total number of objects (individuals) in a population is known as the size of the population.

This may be finite or infinite.

Individual : Each object belonging to a population is called as an individual of the population.

### Sample:

A finite set of objects drawn from the population with a particular aim, is called a sample.

The total number of individuals in a sample is called the sample size.

### Characteristic:

The information required from an individual, from a population or from a sample, during the statistical enquiry (survey) is known as the characteristic of the individual. It is either numerical or non-numerical. For e.g. the size of shoes is a numerical characteristic which refers to a quantity, whereas the mother tongue of a person is a non-numerical characteristic which refers to a quality. Thus we have quantitative and qualitative types of characteristics.

### Variate:

A quantitative characteristic of an individual which can be expressed numerically is called a variate or a variable. It may take different values at different times, places or situations.

### Attribute:

A qualitative characteristic of an individual which can be expressed numerically is called an attribute. For e.g. the mother-tongue of a person, the color of eyes or the color of hair of a person etc.

### Discrete

#### variate :

A variable that is not capable of assuming all the values in a given range is a discrete variate.

### Continuous

#### Variate :

A variate that is capable of assuming all the numerical values in a given range, is called a

continuous variate. Consider two examples carefully, viz. the number of students of a class and their heights. Both variates differ slightly, in the sense that, the number of students present in a class is a number say between 0 and 50; always a whole number. It can never be 1.5, 4.33 etc. This type of variate can take only isolated values and is called a discrete variate. On the other hand heights ranging from 140 cm to 190 cm can take values like 140.7, 135.8, 185.1 etc. Such a variate is a continuous variate.

Describe arithmetic, geometric and harmonic means with suitable example. Explain merits and limitation of geometric mean.

### **Arithmetic mean**

The arithmetic mean is the "standard" average, often simply called the "mean". It is used for many purposes but also often abused by incorrectly using it to describe skewed distributions, with highly misleading results. The classic example is average income - using the arithmetic mean makes it appear to be much higher than is in fact the case. Consider the scores {1, 2, 2, 2, 3, 9}. The arithmetic mean is 3.16, but five out of six scores are below this!

The arithmetic mean of a set of numbers is the sum of all the members of the set divided by the number of items in the set. (The word set is used perhaps somewhat loosely; for example, the number 3.8 could occur more than once in such a "set".) The arithmetic mean is what pupils are taught very early to call the "average." If the set is a statistical population, then we speak of the population mean. If the set is a statistical sample, we call the resulting statistic a sample mean. The mean may be conceived of as an estimate of the median. When the mean is not an accurate estimate of the median, the set of numbers, or frequency distribution, is said to be skewed.

We denote the set of data by  $X = \{x_1, x_2, \dots, x_n\}$ . The symbol  $\mu$  (Greek: mu) is used to denote the arithmetic mean of a population. We use the name of the variable,  $X$ , with a horizontal bar over it as the symbol (" $X$  bar") for a sample mean. Both are computed in the same way:

The arithmetic mean is greatly influenced by outliers. In certain situations, the arithmetic mean is the wrong concept of "average" altogether. For example, if a stock rose 10% in the first year, 30% in the second year and fell 10% in the third year, then it would be incorrect to report its "average" increase per year over this three year period as the arithmetic mean  $(10\% + 30\% + (-10\%))/3 = 10\%$ ; the correct average in this case is the geometric mean which yields an average increase per year of only 8.8%.

### **Geometric mean**

The geometric mean is an average which is useful for sets of numbers which are interpreted according to their product and not their sum (as is the case with the arithmetic mean). For example rates of growth.

The geometric mean of a set of positive data is defined as the product of all the members of the set, raised to a power equal to the reciprocal of the number of members. In a formula: the geometric mean of  $a_1, a_2, \dots, a_n$  is  $(a_1 \times a_2 \times \dots \times a_n)^{1/n}$ , which is  $\sqrt[n]{a_1 \times a_2 \times \dots \times a_n}$ . The geometric mean is useful to determine "average factors". For example, if a stock rose 10% in the first year, 20% in the second year and fell 15% in the third year, then we compute the geometric mean of the factors 1.10, 1.20 and 0.85 as  $(1.10 \times 1.20 \times 0.85)^{1/3} = 1.0391\dots$  and we conclude that the stock rose on average 3.91%

percent per year. The geometric mean of a data set is always smaller than or equal to the set's arithmetic mean (the two means are equal if and only if all members of the data set are equal). This allows the definition of the arithmetic-geometric mean, a mixture of the two which always lies in between. The geometric mean is also the arithmetic-harmonic mean in the sense that if two sequences  $(a_n)$  and  $(h_n)$  are defined:

and

Then  $a_n$  and  $h_n$  will converge to the geometric mean of  $x$  and  $y$ .

### **Harmonic mean**

The harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, for example speed (distance per unit of time).

In mathematics, the harmonic mean is one of several methods of calculating an average.

The harmonic mean of the positive real numbers  $a_1, \dots, a_n$  is defined to be

The harmonic mean is never larger than the geometric mean or the arithmetic mean (see generalized mean). In certain situations, the harmonic mean provides the correct notion of "average". For instance, if for half the distance of a trip you travel at 40 miles per hour and for the other half of the distance you travel at 60 miles per hour, then your average speed for the trip is given by the harmonic mean of 40 and 60, which is 48; that is, the total amount of time for the trip is the same as if you traveled the entire trip at 48 miles per hour. Similarly, if in an electrical circuit you have two resistors connected in parallel, one with 40 ohms and the other with 60 ohms, then the average resistance of the two resistors is 48 ohms; that is, the total resistance of the circuit is the same as it would be if each of the two resistors were replaced by a 48-ohm resistor. (Note: this is not to be confused with their equivalent resistance, 24 ohm, which is the resistance needed for a single resistor to replace the two resistors at once.)

Typically, the harmonic mean is appropriate for situations when the average of rates is desired.

Another formula for the harmonic mean of two numbers is to multiply the two numbers, and divide that quantity by the arithmetic mean of the two numbers. In mathematical terms:

Merits and limitation of geometric mean

Merits:

It is based on each and every item of the series.

It is rigidly defined

It is useful in averaging ratio and percentage in determining rates of increase or decrease.

it gives less weight to large items and more to small items. Thus geometric mean of the geometric of values is always less than their arithmetic mean.

It is capable of algebraic manipulation like computing the grand geometric mean of the geometric mean of different sets of values.>

Limitation:

It is relatively difficult to comprehend, compute and interpret.

A G.M with zero value cannot be compounded with similar other non-zero values with negative sign

## Explain the various

### measure of central tendency?

In statistics, the general level, characteristic, or typical value that is representative of the majority of cases. Among several accepted measures of central tendency employed in data reduction, the most common are the arithmetic mean (simple average), the median, and the mode. FOR EXAMPLE, one measure of central tendency of a group of high school students is the average (mean) age of the students. Central tendency is a term used in some fields of empirical research to refer to what statisticians sometimes call "location". A "measure of central tendency" is either a location parameter or a statistic used to estimate a location parameter.

Examples include: #Arithmetic mean, the sum of all data divided by the number of observations in the data set.#Median, the value that separates the higher half from the lower half of the data set.#Mode, the most frequent value in the data set. Measures of central tendency, or

"location", attempt to quantify what we mean when we think of as the "typical" or "average" score in a data set. The concept is extremely important and we encounter it frequently in daily life. For example, we often want to know before purchasing a car its average distance per litre of petrol. Or before accepting a job, you might want to know what a typical salary is for people in that position so you will know whether or not you are going to be paid what you are worth.

Or, if you are a smoker, you might often think about how many cigarettes you smoke "on average" per day. Statistics geared toward measuring central tendency all focus on this concept of "typical" or "average." As we will see, we often ask questions in psychological science revolving around how groups differ from each other "on average". Answers to such a question tell us a lot about the phenomenon or process we are studying Arithmetic Mean The arithmetic mean is the most common measure of central tendency. It simply the sum of the numbers

divided by the number of numbers. The symbol  $\mu$  is used for the mean of a population. The symbol  $\bar{M}$  is used for the mean of a sample. The formula for  $\mu$  is shown below:  $\mu = \frac{\sum X}{N}$  where  $\sum X$  is the sum of all the numbers in the numbers in the sample and  $N$  is the number of numbers in the sample. As an example, the mean of the numbers  $1+2+3+6+8=20$  is  $20/5=4$  regardless of whether the numbers constitute the entire population or just a sample from the population. The table, Number of touchdown passes, shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.4516 as shown below.

$\mu = \frac{\sum X}{N} = \frac{634}{31} = 20.4516$  Number of touchdown passes

37 33 33 32 29 28 28 23

22 22 22 21 21 21 20 20

19 19 18 18 18 18 16 15

14 14 14 12 12 9 6

Although the arithmetic mean is not the only "mean" (there is also a geometric mean), it is by far the most commonly used. Therefore, if the term "mean" is used without specifying whether it is the arithmetic mean, the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean. Median The median is also a frequently used measure of central tendency.

The median is the midpoint of a distribution: the same number of scores are above the median as below it. For the data in the table, Number of touchdown passes, there are 31 scores. The

16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. The median can also be thought of as the 50th percentile. Let's return to the made up example of the quiz on which you made a three discussed previously in the module Introduction to Central Tendency and shown in table 2.

Three possible datasets for the 5-point make-up quiz

Student	Dataset 1	Dataset 2	Dataset 3
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

For Dataset 1, the median is three, the same as your score. For Dataset 2, the median is 4.

Therefore, your score is below the median. This means you are in the lower half of the class.

Finally for Dataset 3, the median is 2. For this dataset, your score is above the median and

therefore in the upper half of the distribution. Computation of the Median: When there is an

odd number of numbers, the median is simply the middle number. For example, the median of

2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two

middle numbers. Thus, the median of the numbers 22, 44, 77, 1212 is  $\frac{4+72}{2}=5.5$  . mode

The mode is the most frequently occurring value. For the data in the table, Number of

touchdown passes, the mode is 18 since more teams (4) had 18 touchdown passes than any

other number of touchdown passes. With continuous data such as response time measured to

many decimals, the frequency of each value is one since no two scores will be exactly the same

(see discussion of continuous variables). Therefore the mode of continuous data is normally

computed from a grouped frequency distribution. The Grouped frequency distribution table

shows a grouped frequency distribution for the target response time data. Since the interval

with the highest frequency is 600-700, the mode is the middle of that interval (650). Grouped

frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Trimean

The trimean is computed by adding the 25th percentile plus twice the 50th percentile plus the

75th percentile and dividing by four. What follows is an example of how to compute the

trimean. The 25th, 50th, and 75th percentile of the dataset "Example 1" are 51, 55, and 63

respectively. Therefore, the trimean is computed as:

The trimean is almost as resistant to extreme scores as the median and is less subject to

sampling fluctuations than the arithmetic mean in extremely skewed distributions. It is less

efficient than the mean for normal distributions. . The trimean is a good measure of central

tendency and is probably not used as much as it should be.

Trimmed Mean

A trimmed mean is calculated by discarding a certain percentage of the lowest and the highest scores and then computing the mean of the remaining scores. For example, a mean trimmed 50% is computed by discarding the lower and higher 25% of the scores and taking the mean of the remaining scores. The median is the mean trimmed 100% and the arithmetic mean is the mean trimmed 0%. A trimmed mean is obviously less susceptible to the effects of extreme scores than is the arithmetic mean. It is therefore less susceptible to sampling fluctuation than the mean for extremely skewed distributions. It is less efficient than the mean for normal distributions. Trimmed means are often used in Olympic scoring to minimize the effects of extreme ratings possibly caused by biased judges.

### **measure of dispersion, explain each of them?**

Answer: In many ways, measures of central tendency are less useful in statistical analysis than measures of dispersion of values around the central tendency. The dispersion of values within variables is especially important in social and political research because:

- Dispersion or "variation" in observations is what we seek to explain.
- Researchers want to know WHY some cases lie above average and others below average for a given variable:
  - o TURNOUT in voting: why do some states show higher rates than others?
  - o CRIMES in cities: why are there differences in crime rates?
  - o CIVIL STRIFE among countries: what accounts for differing amounts?
- Much of statistical explanation aims at explaining DIFFERENCES in observations -- also known as
  - o VARIATION, or the more technical term, VARIANCE

If everything were the same, we would have no need of statistics. But, people's heights, ages, etc., do vary. We often need to measure the extent to which scores in a dataset differ from each other. Such a measure is called the dispersion of a distribution. Some measures of dispersion are

1) Range The range is the simplest measure of dispersion. The range can be thought of in two ways

. 1. As a quantity: the difference between the highest and lowest scores in a distribution.

"The range of scores on the exam was 32."

2. As an interval; the lowest and highest scores may be reported as the range. "The range was 62 to 94," which would be written (62, 94).

The Range of a Distribution

Find the range in the following sets of data:

NUMBER OF BROTHERS AND SISTERS

{ 2, 3, 1, 1, 0, 5, 3, 1, 2, 7, 4, 0, 2, 1, 2,  
1, 6, 3, 2, 0, 0, 7, 4, 2, 1, 1, 2, 1, 3, 5, 12,  
4, 2, 0, 5, 3, 0, 2, 2, 1, 1, 8, 2, 1, 2 }

An outlier is an extreme score, i.e., an infrequently occurring score at either tail of the distribution. Range is determined by the furthest outliers at either end of the distribution.

Range is of limited use as a measure of dispersion, because it reflects information about extreme values but not necessarily about "typical" values. Only when the range is "narrow"

(meaning that there are no outliers) does it tell us about typical values in the data.

## 2) Percentile range

Most students are familiar with the grading scale in which "C" is assigned to average scores, "B" to above-average scores, and so forth. When grading exams "on a curve," instructors look to see how a particular score compares to the other scores. The letter grade given to an exam score is determined not by its relationship to just the high and low scores, but by its relative position among all the scores. Percentile describes the relative location of points anywhere along the range of a distribution. A score that is at a certain percentile falls even with or above that percent of scores. The median score of a distribution is at the 50th percentile: It is the score at which 50% of other scores are below (or equal) and 50% are above. Commonly used percentile measures are named in terms of how they divide distributions. Quartiles divide scores into fourths, so that a score falling in the first quartile lies within the lowest 25% of scores, while a score in the fourth quartile is higher than at least 75% of the scores. Quartile Finder

The divisions you have just performed illustrate quartile scores. Two other percentile scores commonly used to describe the dispersion in a distribution are decile and quintile scores which divide cases into equal sized subsets of tenths (10%) and fifths (20%), respectively. In theory, percentile scores divide a distribution into 100 equal sized groups. In practice this may not be possible because the number of cases may be under 100. A box plot is an effective visual representation of both central tendency and dispersion. It simultaneously shows the 25th, 50th (median), and 75th percentile scores, along with the minimum and maximum scores. The "box" of the box plot shows the middle or "most typical" 50% of the values, while the "whiskers" of the box plot show the more extreme values. The length of the whiskers indicate visually how extreme the outliers are. Below is the box plot for the distribution you just separated into quartiles. The boundaries of the box plot's "box" line up with the columns for the quartile scores on the histogram. The box plot displays the median score and shows the range of the distribution as well.

By far the most commonly used measures of dispersion in the social sciences are Variance and standard deviation.

Variance is the average squared difference of scores from the mean score of a distribution. Standard deviation is the square root of the variance. In calculating the variance of data points, we square the difference between each point and the mean because if we summed the differences directly, the result would always be zero. For example, suppose three friends work on campus and earn \$5.50, \$7.50, and \$8 per hour, respectively. The mean of these values is  $(5.50 + 7.50 + 8)/3 = \$7$  per hour. If we summed the differences of the mean from each wage, we would get  $(5.50-7) + (7.50-7) + (8-7) = -1.50 + .50 + 1 = 0$ . Instead, we square the terms to obtain a variance equal to  $2.25 + .25 + 1 = 3.50$ . This figure is a measure of dispersion in the set of scores. The variance is the minimum sum of squared differences of each score from any number. In other words, if we used any number other than the mean as the value from which each score is subtracted, the resulting sum of squared differences would be greater. (You can try it yourself -- see if any number other than 7 can be plugged into the preceding calculation and yield a sum of squared differences less than 3.50.) The standard deviation is simply the square root of the variance. In some sense, taking the square root of the variance "undoes" the squaring of the differences that we did when we calculated the variance. Variance and standard

deviation of a population are designated by  $\sigma$  and  $\sigma^2$ , respectively. Variance and standard deviation of a sample are designated by  $s^2$  and  $s$ , respectively.

#### 4) Standard Deviation

The standard deviation ( $s$  or  $\sigma$ ) and variance ( $s^2$  or  $\sigma^2$ ) are more complete measures of dispersion which take into account every score in a distribution. The other measures of dispersion we have discussed are based on considerably less information. However, because variance relies on the squared differences of scores from the mean, a single outlier has greater impact on the size of the variance than does a single score near the mean. Some statisticians view this property as a shortcoming of variance as a measure of dispersion, especially when there is reason to doubt the reliability of some of the extreme scores. For example, a researcher might believe that a person who reports watching television an average of 24 hours per day may have misunderstood the question. Just one such extreme score might result in an appreciably larger standard deviation, especially if the sample is small. Fortunately, since all scores are used in the calculation of variance, the many non-extreme scores (those closer to the mean) will tend to offset the misleading impact of any extreme scores. The standard deviation and variance are the most commonly used measures of dispersion in the social sciences because:

- Both take into account the precise difference between each score and the mean. Consequently, these measures are based on a maximum amount of information.
- The standard deviation is the baseline for defining the concept of standardized score or "z-score".
- Variance in a set of scores on some dependent variable is a baseline for measuring the correlation between two or more variables (the degree to which they are related).

### **Probability**

what do you understand by concept of probability. Explain various theories of probability. Probability is a branch of mathematics that measures the likelihood that an event will occur. Probabilities are expressed as numbers between 0 and 1. The probability of an impossible event is 0, while an event that is certain to occur has a probability of 1. Probability provides a quantitative description of the likely occurrence of a particular event. Probability is conventionally expressed on a scale of zero to one. A rare event has a probability close to zero. A very common event has a probability close to one.

#### **Four theories of probability**

1) Classical or a priori probability: this is the oldest concept evolved in 17th century and based on the assumption that outcomes of random experiments (like tossing of coin, drawing cards from a pack or throwing a die) are equally likely. For this reason this is not valid in the following cases (a) Where outcomes of experiments are not equally likely, for example lives of different makes of bulbs.

(b) Quality of products from a mechanical plant operated under different condition. However, it is possible to mathematically work out the probability of complex events, despite of these demerits. A priori probabilities are of considerable importance in applied statistics.

2) Empirical concept: this was developed in 19th century for insurance business data and is based on the concept of relative frequency. It is based on historical data being used for future prediction. When we toss a coin, the probability of a head coming up is  $\frac{1}{2}$  because there are

two equally likely events, namely appearance of a head or that of a tail. This is an approach of determining a probability from deductive logic.

3) Subjective or personal approach. This approach was adopted by Frank Ramsey in 1926 and developed by others. It is based on personal beliefs of the person making the probability statement based on past information, noticeable trends and appreciation of futuristic situation. Experienced people use this approach for decision making in their own field.

4) Axiomatic approach: this approach was introduced by Russian mathematician A N Kolmogorov in 1933. His concept of probability is considered as a set of function, no precise definition is given but following axioms or postulates are adopted.

a) The probability of an event ranges from 0 to 1. That is, an event surely not to happen has probability 0 and another event sure to happen is associated with probability 1.

b) The probability of an entire sample space (that is any, some or all the possible outcomes of an experiment) is 1. Mathematically,  $P(S) = 1$

What is "chi-square" test, narrate the steps for determining value of  $\chi^2$  with suitable examples.

Explain the condition for applying  $\chi^2$  and uses of chi-square test

This test was developed by Karl Pearson (1857-1936), analytical statistician and professor of applied mathematics, London, whose concept of coefficient of correlation is most widely used.

This r-test considers the magnitude of dependency between theory and observation and is defined as

Where  $O_i$  is the observed frequency  
 $E$  = expected frequencies

Steps for determining value of  $\chi^2$

1) When data is given in a tabulated form calculate expected frequencies for each cell using the following formula

$E = (\text{row total}) * (\text{column total}) / \text{total number of observation}$ .

2) Take difference between  $O$  and  $E$  for each cell and calculate their square  $(O-E)^2$

3) Divide  $(O-E)^2$  by respective expected frequencies and total up to get  $\chi^2$ .

4) Compare calculated value with table value at given degree of freedom and specified level of significance. If at a stated level, the calculated value is more than table values, the difference between theoretical and observed frequencies are considered to be significant. It could not have arisen due to fluctuation of simple sampling. However if the value is less than table value it is not considered as significant, regarded as due to fluctuation of simple sampling and therefore ignored.

Condition for applying  $\chi^2$

1)  $N$  must be large, say more than 50, to ensure the similarity between theoretically correct distribution and our sampling distribution.

2) No theoretical cell frequency should be too small, say less than 5, because that may be over estimation of the value of  $\chi^2$  and may result into rejection of hypotheses. In case we get such frequencies, we should pool them up with the previous or succeeding frequencies.

This action is called Yates correction for continuity.

**USES OF CHI SQUARE TEST:**

1) As a test of independence

The Chi Square Test of Independence tests the association between 2 categorical variables. Whether two or more attributes are associated or not can be tested by framing a hypothesis and testing it against table value. For example, use of quinine is effective in control of fever or

complexions of husband and wives. Consider two variables at the nominal or ordinal levels of measurement. A question of interest is: Are the two variables of interest independent(not related)or are they related (dependent)?

When the variables are independent, we are saying that knowledge of one gives us no information about the other variable. When they are dependent, we are saying that knowledge of one variable is helpful in predicting the value of the other variable. One popular method used to check for independence is the chi-squared test of independence. This version of the chi-squared distribution is a nonparametric procedure whereas in the test of significance about a single population variance it was a parametric procedure. Assumptions: 1. We take a random sample of size  $n$ .

2. The variables of interest are nominal or ordinal in nature.

3. Observations are cross classified according to two criteria such that each observation belongs to one and only one level of each criterion.

2) As a test of goodness of fit The Test for independence (one of the most frequent uses of Chi Square) is for testing the null hypothesis that two criteria of classification, when applied to a population of subjects are independent. If they are not independent then there is an association between them. A statistical test in which the validity of one hypothesis is tested without specification of an alternative hypothesis is called a goodness-of-fit test. The general procedure consists in defining a test statistic, which is some function of the data measuring the distance between the hypothesis and the data (in fact, the badness-of-fit), and then calculating the probability of obtaining data which have a still larger value of this test statistic than the value observed, assuming the hypothesis is true. This probability is called the size of the test or confidence level. Small probabilities (say, less than one percent) indicate a poor fit. Especially high probabilities (close to one) correspond to a fit which is too good to happen very often, and may indicate a mistake in the way the test was applied, such as treating data as independent when they are correlated. An attractive feature of the chi-square goodness-of-fit test is that it can be applied to any university distribution for which you can calculate the cumulative distribution function. The chi-square goodness-of-fit test is applied to binned data (i.e., data put into classes). This is actually not a restriction since for non-binned data you can simply calculate a histogram or frequency table before generating the chi-square test. However, the values of the chi-square test statistic are dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid.

3) As test of homogeneity: it is an extension of test for independence weather two more independent random samples are drawn from the same population or different population. The Test for Homogeneity answers the proposition that several populations are homogeneous with respect to some characteristic.

### **Probability Distribution**

enumerate probability distribution; explain the histogram and probability distribution curve.

#### **probability distribution curve:**

Probability distributions are a fundamental concept in statistics. They are used both on a theoretical level and a practical level.

Some practical uses of probability distributions are:

- To calculate confidence intervals for parameters and to calculate critical regions for hypothesis tests.
- For uni variate data, it is often useful to determine a reasonable distributional model for the data.
- Statistical intervals and hypothesis tests are often based on specific distributional assumptions. Before computing an interval or test based on a distributional assumption, we need to verify that the assumption is justified for the given data set. In this case, the distribution does not need to be the best-fitting distribution for the data, but an adequate enough model so that the statistical technique yields valid conclusions.
- Simulation studies with random numbers generated from using a specific probability distribution are often needed.

The probability distribution of the variable  $X$  can be uniquely described by its cumulative distribution function  $F(x)$ , which is defined by for any  $x$  in  $R$ .

A distribution is called discrete if its cumulative distribution function consists of a sequence of finite jumps, which means that it belongs to a discrete random variable  $X$ : a variable which can only attain values from a certain finite or countable set. A distribution is called continuous if its cumulative distribution function is continuous, which means that it belongs to a random variable  $X$  for which  $\Pr[X = x] = 0$  for all  $x$  in  $R$ .

Several probability distributions are so important in theory or applications that they have been given specific names:

The Bernoulli distribution, which takes value 1 with probability  $p$  and value 0 with probability  $q = 1 - p$ .

### **THE POISSON DISTRIBUTION**

In probability theory and statistics, the Poisson distribution is a discrete probability distribution (discovered by Siméon-Denis Poisson (1781–1840) and published, together with his probability theory, in 1838).  $N$  that count, among other things, a number of discrete occurrences (sometimes called "arrivals") that take place during a time-interval of given length. The probability that there are exactly  $k$  occurrences ( $k$  being a non-negative integer,  $k = 0, 1, 2, \dots$ ) is where  $e$  is the base of the natural logarithm ( $e = 2.71828\dots$ ),  $k!$  is the factorial of  $k$ ,

$\lambda$  is a positive real number, equal to the expected number of occurrences that occur during the given interval. For instance, if the events occur on average every 2 minutes, and you are interested in the number of events occurring in a 10 minute interval, you would use as model a Poisson distribution with  $\lambda = 5$ .

### **THE NORMAL DISTRIBUTION**

The normal or Gaussian distribution is one of the most important probability density functions, not the least because many measurement variables have distributions that at least approximate to a normal distribution. It is usually described as bell shaped, although its exact characteristics are determined by the mean and standard deviation. It arises when the value of a variable is determined by a large number of independent processes. For example, weight is a function of many processes both genetic and environmental. Many statistical tests make the assumption that the data come from a normal distribution.

### **THE PROBABILITY DISTRIBUTION FUNCTION IS GIVEN BY THE FOLLOWING FORMULA**

Where  $x$  = value of the continuous random variable  
= mean of normal random variable (greek letter 'mu')  
 $e$  = exponential constant = 2.7183  
= standard deviation of the distribution  
= mathematical constant = 3.1416

### **HISTOGRAM AND PROBABILITY DISTRIBUTION curve**

Histograms--bar charts in which the area of the bar is proportional to the number of observations having values in the range defining the bar. As an example construct histograms of populations. The population histogram describes the proportion of the population that lies between various limits. It also describes the behavior of individual observations drawn at random from the population, that is, it gives the probability that an individual selected at random from the population will have a value between specified limits. When we're talking about populations and probability, we don't use the words "population histogram". Instead, we refer to probability densities and distribution functions. (However, it will sometimes suit my purposes to refer to "population histograms" to remind you what a density is.) When the area of a histogram is standardized to 1, the histogram becomes a probability density function. The area of any portion of the histogram (the area under any part of the curve) is the proportion of the population in the designated region. It is also the probability that an individual selected at random will have a value in the designated region. For example, if 40% of a population has cholesterol values between 200 and 230 mg/dl, 40% of the area of the histogram will be between 200 and 230 mg/dl. The probability that a randomly selected individual will have a cholesterol level in the range 200 to 230 mg/dl is 0.40 or 40%. Strictly speaking, the histogram is properly a density, which tells you the proportion that lies between specified values. A (cumulative) distribution function is something else. It is a curve whose value is the proportion with values less than or equal to the value on the horizontal axis, as the example to the left illustrates. Densities have the same name as their distribution functions. For example, a bell-shaped curve is a normal density. Observations that can be described by a normal density are said to follow a normal distribution.

### **Index Numbers**

how do you define "index numbers"? Narrate the nature and types of index number with adequate example.

according to Croxton and Cowden index numbers are devices for measuring difference in the magnitude of a group of related

According to Morris Hamburg "in its simplest form an index number is nothing more than a relative which express the relationship between two figures, where one figure is used as a base.

According to M. L. Berenson and D.M. LEVINE "generally speaking, index number measure the size or magnitude of some object at particular point in time as a percentage of some base or reference object in the past. According to Richard .I. Levin and David S. Rubin" an index number is a measure how much a variable changes over time

### **NATURE OF INDEX NUMBER**

1) Index numbers are specified average used for comparison in situation where two or more series are expressed in different units or represent different items. E.g. consumer price index

representing prices of various items or the index of industrial production representing various commodities produced.

2) Index number measure the net change in a group of related variable over a period of time.

3) Index number measure the effect of change over a period of time, across the range of industries, geographical regions or countries.

4) The consumption of the index number is carefully planned according to the purpose of their computation, collection of data and application of appropriate method, assigning of correct weightages and formula.

**TYPES OF INDEX NUMBERS:**

**Price index numbers:** A price index is any single number calculated from an array of prices and quantities over a period. Since not all prices and quantities of purchases can be recorded, a representative sample is used instead.. price are generally represented by p in formulae. These are also expressed as price relative , defined as follows

Price relative=(current years price/base years price)\*100

$=\frac{p_1}{p_0} \times 100$  any increases in price index amounts to corresponding decreases in purchasing power of the rupees or other affected currency. Quantity index number a quantity index number measures how much the number or quantity of a variable changes over time.

Quantities are generally represented as q in formulae. Value index number: a value index number measures changes in total monetary worth, that is, it measure changes in the rupee value of a variable. It combines price and quantity changes to present a more informative index.

Composite index number: a single index number may reflect a composite, or group, of changing variable. For instance, the consumer price index measures the general price level for specific goods and service in the economy. These are also known as index numbers. In such cases the price-relative with respect to a selected base are determined separately for each and their statistical average is computed

what are the importance of index numbers used in Indian economy. Explain index numbers of industrial production

**IMPORTANCE OF INDEX NUMBERS USED IN INDIAN ECONOMY:**

**Cost of living index or consumer price index**

Cost of living index number or consumer price index, expressed as percentage, measure the relative amount of money necessary to derive equal satisfaction during two periods of time, after taking into consideration the fluctuations of the retail prices of consumer goods during these periods. This index is relevant to that real wages of workers are defined as  $(\frac{\text{actual wages}}{\text{cost of living index}}) \times 100$ . Generally the list of items consumed varies for different classes of people (rich, middle, class, or the poor) at the same place of residence. Also people of the same class belonging to different geographical regions have different consumer habits. Thus the cost of living index always relates to specific class of people and a specific geographical area, and it help in determining the effect of changes in price on different classes of consumers living in different areas. The process of construction of cost of living index number is as follows

- 1) Obtain decision about class of people for whom the index number is to be computed, for instance, the industrial personnel, officers or teachers etc. also decide on the geographical area to be covered.
- 2) Conduct a family budget inquiry covering the class of people for whom the index number is to be computed. The enquiry should be conducted for the base year by the process of random sampling. This would give information regarding the nature, quality and quantities of commodities consumed by an average family of the class and also the amount spent on different items of consumption.
- 3) The item on which the information regarding money spent is to be collected are food( rice, wheat, sugar, milk, tea etc) ,clothing, fuel and lighting, housing and miscellaneous items.
- 4) Collect retail prices in respect of the items from the localities in which the class of people concerned reside, or from the markets where they usually make their purchases.
- 5) as the relative importance of various items for different classes of people is not the same, the price or price relative are always weighted and therefore, the cost of living index is always a weighted index.
- 6) The percentage expenditure on an item constitutes the weight of the item and the percentage expenditure in the five groups constitutes the group weight.
- 7) Separate index number are first of all determined for each of the five major groups, by calculating the weighted average of price-relatives of the selected items in the group.

## **INDEX NUMBER OF INDUSTRIAL PRODUCTION**

The index number of industrial production is designed to measure increase or decrease in the level of industrial production in a given period of time compared to some base periods. Such an index measures changes in the quantities of production and not their values. Data about the level of industrial output in the base period and in the given period is to be collected first under the following heads

Textile industries to include cotton, woolen, silk etc.

Mining industries like iron ore, iron, coal, copper, petroleum etc.

Metallurgical industries like automobiles, locomotive, aero planes etc

Industries subject to excise duties like sugar, tobacco, match etc.

Miscellaneous like glass, detergents, chemical, cement etc.

The figure of output for a various industries classifies above are obtained on a monthly, quarterly or yearly basis. Weights are assigned to various industries on the basis of some criteria such as capital invested turnover, net output, production etc. usually the weights in the index are based on the values of net output of different industries. The index of industrial production is obtained by taking the simple mean or geometric mean of relatives. When the simple arithmetic mean is used the formula for constructing the index is as follows.

Index of industrial production =  $(100/w) * \{ (q_1/q_0) w \} = (100/w) * \sum I.w$

Where  $q_1$  = quantity produced in a given period

$Q_0$  = quantity produced in the base period

$W$  = relative importance of different outputs

$I = (q_1/q_0)$  = index for respective commodity