**What are different types of data collection? Explain with examples.**

 Basically there are two types of data collection they are:

* Primary data collection.
* Secondary data collection.

**1. PRIMARY DATA COLLECTION:**

 Collections of primary data during the course of doing experiments in an experimental research but in this case research is done in descriptive type and performs surveys. There are several methods of collecting primary data, particularly in surveys and descriptive researches.

**1. Observation method:**

The observation method is the most commonly used method especially in studies relating to behavioral and sciences. Observations becomes a scientific tool and the method of data collection for the researcher when it serves a formulated research purpose, it is systematically planned and recorded and is subjected to checks and controls on validity and reliability.

However, observation method has various limitations. Firstly, it is and expensive method, secondly the information provided by this method is very limited. Thirdly, sometime unforeseen factors may interfere with the observational task.

There are several merits of the participant type of observation

1) The researcher is enabled to record the natural behavior of the group.

2) The researcher can even gather information which could not easily be obtained if it is observed in a disinterested fashion

3. The researcher can even verify the truth of statement made by informants in the context of a schedule.

2. Interview method.

The interview method of collecting data involves presentation of oral-verbal stimuli and reply in term of oral-verbal responses.

a) Personal interviews: personal interview method requires a person known as the interviewer asking questions generally in a face-to-face contact to the other person or persons.

This sort of interview may be in the form of direct personal investigation or it may be indirect oral investigation. In case of direct personal investigation the interviewer has to collect the information personally from the sources concerned.

Merits of interview method are:

1. More information and that too in greater depth can be obtained.

2. Greater flexibility under this method.

3. The language of the interview can be adapted to ability or educational level of the person interviewed and misinterpretations concerning questions can be avoided.

Demerits of interview methods are:

1. It is very expensive method.

2. Important officials or people in high income groups may not be easily approachable.

3. Interviewing at times may also introduce systematic errors.

**Explain the research methods used for collecting the primary data?**

A: We collect primary data during the course of doing experiments in an experimental research but in case we do research of the descriptive type and perform surveys, whether sample surveys or census surveys, then we can obtain primary data either through observation or through direct communication with respondents in one form or another or through personal interviews.

There are several methods of collecting primary data, particularly in surveys and descriptive researches. Important ones are:

* Observation method,
* Interview method,
* Through questionnaires,
* Through schedules, and
* Other methods which include

(a) Warranty cards;

(b) Distributor audits;

 (c) Pantry audits;

 (d) Consumer panels;

 (e) Using mechanical devices;

 (f) Through projective techniques;

 (g) Depth interviews, and

 (h) Content analysis.

**Observation method:**

The observation method is the most commonly used method especially in studies relating to behavioral sciences. In a way we all observe things around us, but this sort of observation is not scientific observation. Observation becomes a scientific tool and the method of data collection for the researcher, when it serves a formulated research purpose, is systematically planned and recorded and is subjected to checks and controls on validity and reliability. This method is independent of respondents’ willingness to respond. This method is particularly suitable in studies which deal with subjects (i.e., respondents) who are not capable of giving verbal reports of their feelings for one reason or the other.

**Interview method:**

The interview method of collecting data involves presentation of oral-verbal stimuli and reply in terms of oral-verbal responses. This method can be used through personal interviews and, if possible, through telephone interviews.

(a) *Personal interviews:* Personal interview method requires a person known as the interviewer asking questions generally in a face-to-face contact to the other person or persons. (At times the interviewee may also ask certain questions and the interviewer responds to these, but usually the interviewer initiates the interview and collects the information.

(b) *Telephone interviews:* This method of collecting information consists in contacting respondents on telephone itself. It is not a very widely used method, but plays important part in industrial surveys, particularly in developed regions.

**Through Questionnaires:**

This method of data collection is quite popular, particularly in case of big enquiries. It is being adopted by private individuals, research workers, private and public organizations and even by governments.

In this method a questionnaire is sent (usually by post) to the persons concerned with a request to answer the questions and return the questionnaire. A questionnaire consists of a number of questions printed or typed in a definite order on a form or set of forms. The questionnaire is mailed to respondents who are expected to read and understand the questions and write down the reply in the space meant for the purpose in the questionnaire itself. The respondents have to answer the questions on their own.

**Through Schedules:**

This method of data collection is very much like the collection of data through questionnaire, with little difference which lies in the fact that schedules (proforma containing a set of questions) are being filled in by the enumerators who are specially appointed for the purpose. These enumerators along with schedules go to respondents, put to them the questions from the proforma in the order the questions are listed and record the replies in the space meant for the same in the proforma. In certain situations, schedules may be handed over to respondents and enumerators may help them in recording their answers to various questions in the said schedules. Enumerators explain the aims and objects of the investigation and also remove the difficulties which any respondent may feel in understanding the implications of a particular question or the definition or concept of difficult terms.

**What is secondary data in research?**

Secondary data means data that are already available i.e., they refer to the data which have already been collected and analyzed by someone else.

When the researcher utilizes secondary data, then he has to look into various sources from where he can obtain them. In this case he is certainly not confronted with the problems that are usually associated with the collection of original data.

Secondary data may either be published data or unpublished data. Usually published data are available in:

(a) Various publications of the central, state are local governments;

(b) Various publications of foreign governments or of international bodies and their subsidiary organizations;

(c) Technical and trade journals;

(d) Books, magazines and newspapers;

(e) Reports and publications of various associations connected with business and industry, banks, stock exchanges, etc;

(f) Reports prepared by research scholars, universities, economists, etc. in different fields; and

(g) Public records and statistics, historical documents, and other sources of published information.

The sources of unpublished data are many;

They may be found in diaries, letters, unpublished biographies and autobiographies and also may be available with scholars and research workers, trade associations, labour bureaus and other public/ private individuals and organisations.

## Secondary data collection in Research Methodology

Researcher must be very careful in using secondary data. He must make a minute scrutiny because it is just possible that the secondary data may be unsuitable or may be inadequate in the context of the problem which the researcher wants to study.

In this connection Dr. A.L. Bowley very aptly observes that it is never safe to take published statistics at their face value without knowing their meaning and limitations and it is always necessary to criticise arguments that can be based on them.
By way of caution, the researcher, before using secondary data, must see that they possess following characteristics:

1. **Reliability of data:** The reliability can be tested by finding out such things about the said data: (a) who collected the data? (b) What were the sources of data? (c) Were they collected by using proper methods (d) At what time were they collected?(e) Was there any bias of the compiler? (f) What level of accuracy was desired? Was it achieved?
2. **Suitability of data:** The data that are suitable for one enquiry may not necessarily be found suitable in another enquiry. Hence, if the available data are found to be unsuitable, they should not be used by the researcher. In this context, the researcher must very carefully scrutinize the definition of various terms and units of collection used at the time of collecting the data from the primary source originally. Similarly, the object, scope and nature of the original enquiry must also be studied. If the researcher finds differences in these, the data will remain unsuitable for the present enquiry and should not be used.
3. **Adequacy of data:** If the level of accuracy achieved in data is found inadequate for the purpose of the present enquiry, they will be considered as inadequate and should not be used by the researcher. The data will also be considered inadequate, if they are related to an area which may be either narrower or wider than the area of the present enquiry.

From all this we can say that it is very risky to use the already available data. The already available data should be used by the researcher only when he finds them reliable, suitable and adequate. But he should not blindly discard the use of such data if they are readily available from authentic sources and are also suitable and adequate for in that case it will not be economical to spend time and energy in field surveys for collecting information. At times, there may be wealth of usable information in the already available data which must be used by an intelligent researcher but with due precaution.

**Explain the features of data organization for representing the data**

Text, tables, and graphs are effective communication media that present and convey data and information. They aid readers in understanding the content of research, sustain their interest, and effectively present large quantities of complex information. As journal editors and reviewers will scan through these presentations before reading the entire text, their importance cannot be disregarded. For this reason, authors must pay as close attention to selecting appropriate methods of data presentation as when they were collecting data of good quality and analyzing them. In addition, having a well-established understanding of different methods of data presentation and their appropriate use will enable one to develop the ability to recognize and interpret inappropriately presented data or data presented in such a way that it deceives readers' eyes

Data can be presented in one of the three ways:

–as text;

–in tabular form; or

–in graphical form.

Methods of presentation must be determined according to the data format, the method of analysis to be used, and the information to be emphasized. Inappropriately presented data fail to clearly convey information to readers and reviewers. Even when the same information is being conveyed, different methods of presentation must be employed depending on what specific information is going to be emphasized. A method of presentation must be chosen after carefully weighing the advantages and disadvantages of different methods of presentation. For easy comparison of different methods of presentation, let us look at a table and a line graph that present the same information. If one wishes to compare or introduce two values at a certain time point, it is appropriate to use text or the written language. However, a table is the most appropriate when all information requires equal attention, and it allows readers to selectively look at information of their own interest. Graphs allow readers to understand the overall trend in data, and intuitively understand the comparison results between two groups. One thing to always bear in mind regardless of what method is used, however, is the simplicity of presentation.

### Text presentation

Text is the main method of conveying information as it is used to explain results and trends, and provide contextual information. Data are fundamentally presented in paragraphs or sentences. Text can be used to provide interpretation or emphasize certain data.

If quantitative information to be conveyed consists of one or two numbers, it is more appropriate to use written language than tables or graphs. For instance, information about the incidence rates of delirium following anesthesia in 2016–2017 can be presented with the use of a few numbers: “The incidence rate of delirium following anesthesia was 11% in 2016 and 15% in 2017; no significant difference of incidence rates was found between the two years.”

If this information were to be presented in a graph or a table, it would occupy an unnecessarily large space on the page, without enhancing the readers' understanding of the data. If more data are to be presented or other information such as that regarding data trends are to be conveyed, a table or a graph would be more appropriate. By nature, data take longer to read when presented as texts and when the main text includes a long list of information, readers and reviewers may have difficulties in understanding the information.

### Table presentation

Tables, which convey information that has been converted into words or numbers in rows and columns, have been used for nearly 2,000 years. Anyone with a sufficient level of literacy can easily understand the information presented in a table. Tables are the most appropriate for presenting individual information, and can present both quantitative and qualitative information. Examples of qualitative information are the level of sedation, statistical methods/functions, and intubation conditions.

The strength of tables is that they can accurately present information that cannot be presented with a graph. A number such as “132.145852” can be accurately expressed in a table. Another strength is that information with different units can be presented together. For instance, blood pressure, heart rate, number of drugs administered, and anesthesia time can be presented together in one table.

Finally, tables are useful for summarizing and comparing quantitative information of different variables. However, the interpretation of information takes longer in tables than in graphs, and tables are not appropriate for studying data trends. Furthermore, since all data are of equal importance in a table, it is not easy to identify and selectively choose the information required.

Heat maps for better visualization of information than tables.

Heat maps help to further visualize the information presented in a table by applying colors to the background of cells. By adjusting the colors or color saturation, information is conveyed in a more visible manner, and readers can quickly identify the information of interest. Software such as Excel have features that enable easy creation of heat maps through the options available on the “conditional formatting” menu.

**Graph Presentation**

Whereas tables can be used for presenting all the information, graphs simplify complex information by using images and emphasizing data patterns or trends, and are useful for summarizing, explaining, or exploring quantitative data. While graphs are effective for presenting large amounts of data, they can be used in place of tables to present small sets of data. A graph format that best presents information must be chosen so that readers and reviewers can easily understand the information.

**Explain the features of data organization for representing the data.**

Ans. Data organization, in broad terms, refers to the method of classifying and organizing data sets to make them more useful. Some IT experts apply this primarily to physical records, although some types of data organization can also be applied to digital records.

Some of the features of data organization are:

1) To re-order or analyze the arrangement of data items in a physical record in data organization.

2) The analysis of relatively structured and unstructured data.

3) To make better use of the data assets that they have in a world where data sets represent some of the most valuable assets.

4) To streamline business processes.

5) Get better business intelligence and generally improve a business model.

**6) Enhanced Efficiency and Effectiveness**

**7) Bolster Organizational Agility**– Business decisions no longer have to be made in the dark or based on gut feeling. They can be made as quickly as meaningful insights and information are acquired.

**8) Foster More Revenue**– Data is the new revenue generator. Relentless data improvements and better business predictions fuel current and future decision making,

9) Structured data is comprised of data in tables that can be easily integrated into a database and, from there, fed into analytics software or other particular applications. Unstructured data is data that is raw and unformatted, the kind of data that you find in a simple text document, where names, dates and other pieces of information are scattered throughout random paragraphs.

**Discuss the diagrammatic representation and graphic representation of data?**

**Diagrammatic representation** is the visual form of presentation of data in which facts are highlighted in the language of diagrams.

* It consists in presenting statistical material in interesting and attractive geometrical figures (bars, circles, squares).
* It will attract the attention of a large number of persons.
* It facilitates comparison between two or more sets of data.

Types of diagrams:

1. One dimensional
2. Two dimensional or surface diagram
3. Three dimensional or volume diagram
4. Pictogram
5. Cartogram or map diagram

**One dimensional diagram**:

* One dimensional diagram are such diagrams where only one dimensional measurement i.e., height is used.
* These diagrams may be in the form of lines or bars.
* The heights of these lines or bars are taken on the basis of values.
* Line diagram and bar diagrams are the types of one dimensional diagrams

**Two dimensional or surface diagrams**:

* Two dimensional diagrams are those where the length as well as width of the bar both (and the area) are considered in the construction of diagrams.
* These diagrams are also called as area or surface diagram.
* Rectangle diagrams, square diagrams, pie diagrams will come under the two dimensional diagrams.

**Three dimensional diagrams**:

* Three dimensional diagrams are those in which length, width and height are taken in to account.
* These are also known as cubic diagrams. They may be drawn in the form of cylinders, blocks ,spheres etc.

 **Pictogram:**

* Pictogram is a technique of presenting data through appropriate pictures also called as picture graph or pictograph.
* Pictograms prove very attractive and effective. They create lasting impression on the mind

**Cartogram or map diagram:**

* Cartograms are used to represent data on geographical basis. For example possible rate of rainfall.
* These diagrams are very attractive and effective if message is communicated to common people.

**Graphic representation of data:**

A graphical representation is a visual display of data and statistical results. A picture is said to be more effective than words for describing a particular thing.

* There are different types of graphical representation and which is used depends on the nature of data and the nature of statistical results.
* It is a mathematical picture , which is the geometrical image of a set of data.

 **Types of graphs:**

1. Dot plot
2. Line graph
3. Bar graph
4. Pie diagram or circular graph
5. Histogram

## Dot Plots

The dot plot is one of the most simplest ways of graphical representation of the statistical data. As the name itself suggests, a dot plot uses the dots. It is a graphic display which usually compares frequency within different categories.

The dot plot is composed of dots that are to be plotted on a graph paper.

**A dot plot may look like:**



In the dot plot, every dot denotes a specific number of observations belonging to a data set. One dot usually represents one observation.

These dots are to be marked in the form of a column for each category. In this way, the height of each column shows the corresponding frequency of some category.

The dot plots are quite useful when there are small amount of data is given within the small number of categories.

**Line graph:**

* Line graphs are simple mathematical graphs that are drawn on the graph paper by plotting the data connecting one variable on the horizontal X-axis and other variable of data on the vertical Y-axis.



**Bar graph:**

* In bar graphs data represented by bars.
* The bars can be made in any direction i.e., horizontal or vertical direction.
* The bars are taken of equal width and start from a common horizontal or vertical line and there length indicates the corresponding values of statistical data.



**Pie diagram or circular graph:**

* It is a circle in which different components are represented through the sections or portions of a circle.
* To draw a pie diagram, first the value of each category is expressed as a percentage of the total and then the angle 3600 divided in the same percentages.
* Then at the centre of a circle these angles are drawn simultaneously starting from a particular radius.
* In this way we get a set of sectorial areas proportional to the values of the items.



**Histogram:**

* A histogram is essentially a bar graph of frequency distribution
* It can be constructed for equal as well as unequal class intervals.
* Area of any rectangle of a histogram is proportional to the frequency of that class.

**A histogram may look like the following graph:**



**Examples for Graphical representation of data:**

Few examples of graphical representation of statistical data are given below:
**Example 1:** Draw a dot plot for the following data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Favorite Colors** | **Red** | **Blue** | **Green** | **Yellow** | **Orange** | **Indigo** | **Violet**  |
| Number of Students |  9 |  7 |  5 |  3 |  2 |  1 |  3 |

 **Solution:**
**The line graph for the following data is given below:**


**Example 2:** Plot a bar graph from the data given below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Students** | **A** | **B** | **C** | **D**  |
| Marks  |  8 | 14 |  9 |  5 |

 **Solution:** The following bar graph is obtained:


**Example 3:** Draw a histogram from the given data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test Score** | **24-30** | **30-36** | **36-42** | **42-48** | **48-54** | **54-60** |
| Frequency  |  5 |  6 |  8 |  5 |  10 |  4 |

**Solution:** **We drew the following histogram:**


**Example 4:** The percentages of students who use the different methods of transportation are as follows:

40% go by school bus

25% go by walk

20% go by bicycle

and rest 15% go by car. Draw a pie chart.

**Solution:** The pie graph of the above data is:



**Q.NO33 DISCUSS THE SAMPLE DESIGN AND EXPLAIN THE NEED FOR SAMPLING**

All the items under consideration in any field of inquiry constitute a ‘universe’ or ‘population’. A complete enumeration of all the items in the ‘population’ is known as a census inquiry. It can be presumed that in such an inquiry when all the items are covered no element of chance is left and highest accuracy is obtained. But in practice this may not be true. Even the slightest element of bias in such an inquiry will get larger and larger as the number of observations increases. Moreover, there is no way of checking the element of bias or its extent except through a resurvey or use of sample checks. Besides, this type of inquiry involves a great deal of time, money and energy. Not only this, census inquiry is not possible in practice under many circumstances. For instance, blood testing is done only on sample basis. Hence, quite often we select only a few items from the universe for our study purposes. The items so selected constitute what is technically called a sample. The researcher must decide the way of selecting a sample or what is popularly known as the sample design. In other words, a sample design is a definite plan determined before any data are actually collected for obtaining a sample from a given population. Thus, the plan to select 12 of a city’s 200 drugstores in a certain way constitutes a sample design. Samples can be either probability samples or non-probability samples. With probability samples each element has a known probability of being included in the sample but the non-probability samples do not allow the researcher to determine this probability. Probability samples are those based on simple random sampling, systematic sampling, stratified sampling, cluster/area sampling whereas non-probability samples are those based on convenience sampling, judgement sampling and quota sampling techniques.

**Explain the estimation of population.**

In most statistical research studies, population parameters are usually unknown and have to be estimated from a sample. As such the methods for estimating the population parameters assume an important role in statistical analysis.

The random variables (such Formula) used to estimate population parameters, such as Formula are conventionally called as ‘estimators’, while specific values of these (such as X = 105 Formula = 21.44) are referred to as ‘estimates’ of the population parameters. The estimate of a population parameter may be one single value or it could be a range of values. In the former case it is referred as point estimate, whereas in the latter case it is termed as interval estimate. The researcher usually makes these two types of estimates through sampling analysis. While making estimates of population parameters, the researcher can give only the best point estimate or else he shall have to speak in terms of intervals and probabilities for he can never estimate with certainty the exact values of population parameters. Accordingly he must know the various properties of a good estimator so that he can select appropriate estimators for his study. He must know that a good estimator possesses the following properties:

1. An estimator should on the average be equal to the value of the parameter being estimated. This is popularly known as the property of unbiasedness. An estimator is said to be unbiased if the expected value of the estimator is equal to the parameter being estimated. The sample mean dXi is the most widely used estimator because of the fact that it provides an unbiased estimate of the population mean bmg .
2. An estimator should have a relatively small variance. This means that the most efficient estimator, among a group of unbiased estimators, is one which has the smallest variance. This property is technically described as the property of efficiency.
3. An estimator should use as much as possible the information available from the sample. This property is known as the property of sufficiency.
4. An estimator should approach the value of population parameter as the sample size becomes larger and larger. This property is referred to as the property of consistency. Keeping in view the above stated properties, the researcher must select appropriate estimator(s) for his study. We may now explain the methods which will enable us to estimate with reasonable accuracy the population mean and the population proportion, the two widely used concepts.

**What are functions of statistics and also explain estimation of population?**

ANS: Statistics has patently two broad functions.

* The first function is description and summarizing of information in a manner so as to make it more usable.
* The second function of statics is induction, which involves either making generalizations about some population on basis of sample drawn from this population or formulating general laws on the basis of repeated observations.

The two functions of statistical methods can be easily understood by following example: suppose it is desired to study the problem of labour unrest in a particular area. The first thing to be done here will be to analyse the various causes of labour unrest and to study the impact of each one of theses on various categories of labour (viz., male workers and female workers or skilled labour and unskilled labour ).

 ESTIMATION OF POPULATION:

1. More accurate estimates and projections can generally be made for total population of an area than for the demographic characteristics of population.
2. The census is the only data source available to project population by select attributes such as age, sex, and occupation.
3. When projection or estimating total population size, it is possible to use several tools that are based on different information sources and compares the results.
4. The most accurate projection is usually based on combination of methods. For example, use two methods whose assumptions support available information sources, and take an average of the two.
5. An estimate or projection should always be checked by comparing it with another estimate or projection that employs an equally accurate or more accurate method.

There are two methods to calculate estimation of population:

1. Mathematical method:
* Using arithmetic or geometric inter & extrapolation.
* Estimate for short period of time.
* Result: just total number of population.

 7. Component methods:

* Component of population changes(fertility, mortality and migration)/
* Result: population by age and sex.

**Explain the parametric and non parametric methods with suitable examples**

### Parametric Methods

Methods are classified by what we know about the population we are studying. Parametric methods are typically the first methods studied in an introductory statistics course. The basic idea is that there is a set of fixed parameters that determine a probability model.

Parametric methods are often those for which we know that the population is approximately normal, or we can approximate using a normal distribution after we invoke the central limit theorem. There are two parameters for a normal distribution: the mean and the standard deviation.

Ultimately the classification of a method as parametric depends upon the assumptions that are made about a population. A few parametric methods include:

* Confidence interval for a population mean, with known standard deviation.
* Confidence interval for a population mean, with unknown standard deviation.
* Confidence interval for a population variance.
* Confidence interval for the difference of two means, with unknown standard deviation.

### Nonparametric Methods

To contrast with parametric methods, we will define nonparametric methods. These are statistical techniques for which we do not have to make any assumption of parameters for the population we are studying. Indeed, the methods do not have any dependence on the population of interest. The set of parameters is no longer fixed, and neither is the distribution that we use. It is for this reason that nonparametric methods are also referred to as distribution-free methods.

Nonparametric methods are growing in popularity and influence for a number of reasons. The main reason is that we are not constrained as much as when we use a parametric method. We do not need to make as many assumptions about the population that we are working with as what we have to make with a parametric method. Many of these nonparametric methods are easy to apply and to understand.

A few nonparametric methods include:

* Sign test for population mean
* Bootstrapping techniques
* U test for two independent means
* Spearman correlation test

### Comparison

There are multiple ways to use statistics to find a confidence interval about a mean. A parametric method would involve the calculation of a margin of error with a formula, and the estimation of the population mean with a sample mean. A nonparametric method to calculate a confidence mean would involve the use of bootstrapping.

**Discuss the point of central tendency explain with example.**

Ans. Central tendency is also called “measures of location,” “central location,” or just “center” is a way to describe what’s typical for a set of data. Central tendency doesn’t tell you specifics about the individual pieces of data, but it does give you an overall picture of what is going on in the entire data set. There are three major ways to show central tendency: [mean, mode and median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/).

**Mean**The [mean](https://www.statisticshowto.datasciencecentral.com/mean/) is the [average](https://www.statisticshowto.datasciencecentral.com/average/) of a set of numbers. Add up all the numbers in a set of data and then divide by the number of items in the set. For example, the mean of 2 3 5 9 11 is:
(2 + 3 + 5 + 9 + 11) / 5 = 30 / 5 = 6.

**Median**The [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/median-formula/) is the middle of a set of numbers. Think of it like the median in a road (that grassy area in the middle that separates traffic). Place your data in order, and the number in the exact center of a list is the median. For example:
1 2 3 4 5 6 7
The median is 4 because it’s in the center, with three numbers either side.

**Mode**The mode is the most common number in a set of data. For example, the [mode](https://www.statisticshowto.datasciencecentral.com/mode/) of 1 2 2 3 5 6 is 2. Some data sets have no mode, like this one: 1 2 3 4 5 6. Others have multiple modes, like this one: 1 1 2 3 3.

**Define mean, median & mode and calculate them by assuming suitable data?**

ANS: **MEAN:** The mean is the average of all numbers and is sometimes called arithmetic mean. To calculate mean, add together all of the numbers in a set and then divide the sum by the total count of numbers.

For example:

In a data centre rack, five servers consume 100 watts, 98 watts, 105 watts, 90 watts and 102 watts of power respectively.

The mean power use of that rack is calculated as (100+98+105+90+102w/5 servers = a calculated mean of 99w per server).

**MEDIAN:** the median is the value separating the higher half from the lower half of a data sample. For a data set, it may be thought of as the “middle “value.

To find the median, organize each number in order by size; the number in the middle is the median.

For example:

For the five servers in the rack, arrange the power consumption figures from lowest to highest: 90w, 98w, 100w, 102w, and 105w. The median power consumption of the rack is 100w.

**MODE:** The mode is the number that occurs most often within a set of numbers. The result with the highest count of occurrences is the mode of the set also referred to as the modal value.

For example:

Let us take following data:

78, 56,68, 78,84,78………….the number that occurs most is 78 ( MODE ).

**Explain the tests for significance and analysis of variance**

Ans Tests for significance are three types:

**One way analysis:** When we are comparing more than three groups based on one factor variable, then it said to be one way analysis of variance (ANOVA).

For example, if we want to compare whether or not the mean output of three workers is the same based on the working hours of the three workers.

**Two way analysis:** When factor variables are more than two, then it is said to be two way analysis of variance (ANOVA).

For example, based on working condition and working hours, we can compare whether or not the mean output of three workers is the same

**K-way analysis:** When factor variables are k, then it is said to be the k-way analysis of variance (ANOVA).

**Analysis of variance**

1) Analysis of variance(ANOVA) is a parametric statistical technique used to compare datasets. This technique was invented by R.A. Fisher, and is thus often referred to as Fisher’s ANOVA, as well.

2) The ANOVA technique is important in of all those situations where we want to compare more than two populations such as in comparing the yield of crop from several varieties of seeds, the gasoline mileage of four automobiles, the smoking habits of five groups of university students and so on.

3) Analysis of variance is an extremely useful technique concerning researches in the fields of economics, biology, education, psychology, sociology, business/industry and in researches of several other disciplines. This technique is used when multiple sample cases are involved. . .

4) As stated earlier, the significance of the difference between the means of two samples can be judged through either *z*-test or the *t*-test, but the difficulty arises when we happen to examine the significance of the difference amongst more than two sample means at the same time. The ANOVA technique enables us to perform this simultaneous test and as such is considered to be an important tool of analysis in the hands of a researcher.

The use of this parametric statistical technique involves certain key assumptions, including the following:

1. **Independence of case:** There should not be any pattern in the selection of the sample.

2. **Normality:** Distribution of each group should be normal.

3. **Homogeneity:** Homogeneity means variance between the groups should be the same

If particular data follows the above assumptions, then the analysis of variance (ANOVA) is the best technique to compare the means of two, or more, populations.

**What is chi square test? Explain its significance in statistical analysis**

A chi-squared test, also written as χ2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

In the standard applications of this test, the observations are classified into mutually exclusive classes, and there is some theory, or say null hypothesis, which gives the probability that any observation falls into the corresponding class. The purpose of the test is to evaluate how likely the observations that are made would be, assuming the null hypothesis is true.

Chi-squared tests are often constructed from a sum of squared errors, or through the sample variance. Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem. A chi-squared test can be used to attempt rejection of the null hypothesis that the data are independent.

Also considered a chi-squared test is a test in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough.

**Statistical Significance of Observed Relationship / Chi-Square Test**

The chi-square test for two-way tables is used as a guideline for declaring that the evidence in the sample is strong enough to allow us to generalize that the relationship holds for a larger population as well.

Definition: A statistically significant relationship is a relationship observed in a sample that would have been unlikely to occur if really there is no relationship in the larger population.

Concept : A chi-square statistic for two-way tables is sensitive to the strength of the observed relationship. The stronger the relationship, the larger the value of the chi-square test.

Definition : A p-value for a chi-square statistic is the probability that the chi-square value would be as large as it is (or larger) if really there were no relationship in the population.

IMPORTANT decision rule : An observed relationship will be called statistically significant when the p-value for a chi-square test is less than 0.05. In this case, we generalize that the relationship holds in the larger population.

Assumptions: The two variables must be categorical and expected counts in each cell must be at least five.

The chi-square test statistic for a test of independence of two categorical variables is found by:

X2=∑(O−E)2/EX2=∑(O−E)2/E

where O represents the observed frequency. E is the expected frequency under the null hypothesis and computed by:

E=row total×column total sample size

**What are different types of tests? Explain Chi square test with suitable example.**

Ans. Types of tests

Hypotheses-1(parametric or standard tests of hypotheses)

Chi square test

Analysis of variance and covariance

Hypotheses-2(non-parametric or distribution free tests)

Multivariate analysis technique

**Chi square test:**

The chi-square test is an important test amongst the several tests of significance developed by statisticians. Chi-square, symbolically written as c2 (Pronounced as Ki-square), is a statistical measure used in the context of sampling analysis for comparing a variance to a theoretical variance.

There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:

A chi-square goodness of fit test determines if a sample data matches a population.

A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of [categorical variables](https://www.statisticshowto.datasciencecentral.com/what-is-a-categorical-variable/) differ from each another.

A very small chi square test statistic means that your observed data fits your expected data extremely well. In other words, there is a relationship.

A very large chi square test statistic means that the data does not fit very well. In other words, there isn’t a relationship.

The formula for the chi-square statistic used in the chi square test is:



The subscript “c” are the degrees of freedom. “O” is your observed value and E is your [expected value](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/expected-value/). It’s very rare that you’ll want to actually *use* this formula to find a critical chi-square value by hand. The summation symbol means that you’ll have to perform a calculation for every single data item in your data set.

**EXAMPLE**

**Null and Alternative Hypotheses**

The [null and alternative hypotheses](https://www.thoughtco.com/null-hypothesis-vs-alternative-hypothesis-3126413) for our goodness of fit test reflect the assumption that we are making about the population. Since we are testing whether the colors occur in equal proportions, our null hypothesis will be that all colors occur in the same proportion. More formally, if p1 is the population proportion of red candies, p2 is the population proportion of orange candies, and so on, then the null hypothesis is that p1 = p2 = . . . = p6 = 1/6.

The alternative hypothesis is that at least one of the population proportions is not equal to 1/6.

Actual and Expected Counts

The actual counts are the number of candies for each of the six colors. The expected count refers to what we would expect if the null hypothesis were true. We will let n be the size of our sample. The expected number of red candies is p1 n or n/6. In fact, for this example, the expected number of candies for each of the six colors is simply n times pi, or n/6.

Chi-square Statistic for Goodness of Fit

We will now calculate a chi-square statistic for a specific example. Suppose that we have a simple random sample of 600 M&M candies with the following distribution:

212 of the candies are blue.

147 of the candies are orange.

103 of the candies are green.

50 of the candies are red.

46 of the candies are yellow.

42 of the candies are brown.

If the null hypothesis were true, then the expected counts for each of these colors would be (1/6) x 600 = 100. We now use this in our calculation of the chi-square statistic.

We calculate the contribution to our statistic from each of the colors. Each is of the form (Actual – Expected)2/Expected.:

For blue we have (212 – 100)2/100 = 125.44

For orange we have (147 – 100)2/100 = 22.09

For green we have (103 – 100)2/100 = 0.09

For red we have (50 – 100)2/100 = 25

For yellow we have (46 – 100)2/100 = 29.16

For brown we have (42 – 100)2/100 = 33.64

We then total all of these contributions and determine that our chi-square statistic is 125.44 + 22.09 + 0.09 + 25 +29.16 + 33.64 =235.42.

Degrees of Freedom

The number of [degrees of freedom](https://www.thoughtco.com/what-is-a-degree-of-freedom-3126416) for a goodness of fit test is simply one less than the number of levels of our variable. Since there were six colors, we have 6 – 1 = 5 degrees of freedom.

**What is the importance of regression modeling? How it is used in research work?**

Ans. Importance of Regression modeling:

1) Regression analysis refers to a method of mathematically sorting out which variables may have an impact. The importance of regression analysis for a small business as it helps determine which factors matter most, which it can ignore, and how those factors interact with each other.

2) The importance of regression analysis lies in the fact that it provides a powerful statistical method that allows a business to examine the relationship between two or more variables of interest.

3) Understanding the **importance of regression** analysis, the **advantages of linear regression**, as well as the **benefits of regression analysis** and the **regression method of forecasting** can help a small business, and indeed any business, gain a far greater understanding of the variables (or factors) that can impact its success in the coming weeks, months and years into the future.

4) The importance of regression analysis is that it is all about data: data means numbers and figures that actually define your business. The advantages of regression analysis is that it can allow you to essentially crunch the numbers to help you make better decisions for your business currently and into the future.

5) Predict sales in the near and long term.

 6) Understand inventory levels.

7) Understand supply and demand.

8) Review and understand how different variables impact all of these things.

Regression in Research work:

The basic form of regression models includes unknown parameters (β), independent variables (X), and the dependent variable (Y).

Regression model, basically, specifies the relation of dependent variable (Y) to a function combination of independent variables (X) and unknown parameters (β)

*Y* ≈ *f (X, β)*

Regression equation can be used to predict the values of ‘y’, if the value of ‘x’ is given, and both ‘y’ and ‘x’ are the two sets of measures of a sample size of ‘n’. The formulae for regression equation would be



Where,



You don’t have to apply the formula manually, and correlation and regression analyses can be run with the application of popular analytical software such as Microsoft Excel, Microsoft Access, SPSS and others.

**Explain the regression modeling and time series analysis with an example ?**

Regression analysis is a quantitative research method which is used when the study involves modelling and analysing several variables, where the relationship includes a dependent variable and one or more independent variables. In simple terms, regression analysis is a quantitative method used to test the nature of relationships between a dependent variable and one or more independent variables.

The basic form of regression models includes unknown parameters (β), independent variables (X), and the dependent variable (Y).

Regression model, basically, specifies the relation of dependent variable (Y) to a function combination of independent variables (X) and unknown parameters (β)

       Y ≈ f (X, β)

Regression equation can be used to predict the values of ‘y’, if the value of ‘x’ is given, and both ‘y’ and ‘x’ are the two sets of measures of a sample size of ‘n’. The formulae for regression equation would be

 

Where,

 

Do not be intimidated by visual complexity of correlation and regression formulae above. You don’t have to apply the formula manually, and correlation and regression analyses can be run with the application of popular analytical software such as Microsoft Excel, Microsoft Access, SPSS and others.

**Time series analysis:** In the context of economic and business researches, we may obtain quite often data relating to some time period concerning a given phenomenon. Such data is labelled as

‘Time Series’. More clearly it can be stated that series of successive observations of the given phenomenon over a period of time are referred to as time series. Such series are usually the result of the effects of one or more of the following factors:

(i) *Secular trend* or long term trend that shows the direction of the series in a long period of time. The effect of trend (whether it happens to be a growth factor or a decline factor) is gradual, but extends more or less consistently throughout the entire period of time under consideration. Sometimes, secular trend is simply stated as trend (or T).

(ii) *Short time oscillations* i.e., changes taking place in the short period of time only and such changes can be the effect of the following factors:

(a) *Cyclical fluctuations* (*or C*) are the fluctuations as a result of business cycles and are generally referred to as long term movements that represent consistently recurring rises and declines in an activity.

(b) *Seasonal fluctuations* (*or S*) are of short duration occurring in a regular sequence at specific intervals of time. Such fluctuations are the result of changing seasons. Usually

these fluctuations involve patterns of change within a year that tend to be repeated from year to year. Cyclical fluctuations and seasonal fluctuations taken together constitute short-period regular fluctuations.

(c) *Irregular fluctuations* (*or I*), also known as Random fluctuations, are variations which take place in a completely unpredictable fashion. All these factors stated above are termed as components of time series and when we try to analyse

time series, we try to isolate and measure the effects of various types of these factors on a series. To study the effect of one type of factor, the other type of factor is eliminated from the series. The given series is, thus, left with the effects of one type of factor only.

For analysing time series, we usually have two models;

 (1) multiplicative model;

(2) additive model.

 Multiplicative model assumes that the various components interact in a multiplicative manner to produce the given values of the overall time series and can be stated as under:

*Y* = *T* × *C* × *S* × *I*

where

*Y* = observed values of time series,

 *T* = Trend, *C* = Cyclical fluctuations,

*S* = Seasonal fluctuations,

*I* = Irregular fluctuations.

Additive model considers the total of various components resulting in the given values of the overall time series and can be stated as:

*Y* = *T* + *C* + *S* + *I*

The knowledge of seasonal variations will be of great help to us in taking decisions regarding inventory, production, purchases and sales policies so as to optimize working results. Thus, analysis of time series is important in context of long term as well as short term forecasting and is considered a very powerful tool in the hands of business analysts and researchers.