**Preliminary**

Life in the modern world is bound up with the notions of number, counting and measurement. One cannot think of a community that cannot count or take measurements and yet is concerned with such acts as selling and buying, carrying on banking transactions, operating locomotives, cars, ships and aircraft, keeping track of the fortunes of their favourite soccer or cricket teams, and taking part in government. That is why the word ‘Statistics’ becomes so important in the modern world.

**What is statistics?**

***Statistics, as a plural noun****,* is used to mean numerical data arising in any sphere of human experience (to be precise, numerical data which arise from a host of uncontrolled and mostly unknown causes acting together). It is in this sense that the term is used when our daily newspapers give vital statistics, crime statistics or soccer statistics of India, or when the Food Minister in the Parliament quotes statistics of sugar exports or those of food grain production.

***Statistics, as a singular noun***, is a name for the body of scientific methods (the statistical methods) which are meant for the collection, analysis and interpretation of numerical data. One has this sense in mind when one says that it is a text-book of statistics or that Prof. Montgomery is an outstanding statistician. Not that Prof. Montgomery has got numerical data on different topics at his finger-tips; he is just well versed in statistical methods and their applications.

We will remain mainly concerned about statistics as a singular noun, which implies the study of the collection, organisation, analysis, interpretation and presentation of data. It is built up from the field of mathematics known as probability. Probability gives us a way to determine how likely an event is to occur. It also gives us a way to talk about randomness.

**Is statistics a science?**

It is to be noted that statistics is not a science in the sense physics or chemistry or biology or even economics is. ***Any science has for its objective the formulation of laws for explaining phenomena in some part of the real world.*** Through observations or experiments, a science builds up hypotheses regarding the phenomena, whose validity is then examined through further observations or experiments. A hypothesis that stands repeated tests of this kind is raised to the level of a law. Indeed, every science is but a body of such laws.

*Statistics, however, is not a body of laws. Among other functions, it formulates methods for the verification of hypotheses, for testing whether a hypothesis can claim to be a law.* ***It would be more proper to describe statistics as a quantitative method of scientific investigations.*** Unlike the pure sciences like [physics](https://www.toppr.com/guides/general-knowledge/basic-science/basic-physics/), chemistry etc., statistics is not an exact science. However, we can call it ***a science of scientific methods***. Statistics help other sciences to derive their own laws.

According to **M. J. R. Healy**, Statistics may have a more important role to play in technology than in science and so it may itself best be considered as a technology rather than as a science.

**P. C. Mahalanobis** perceived statistics ‘as a universal tool of inductive inference, research in natural and social sciences, and technological applications’ and ‘as a key technology for increasing the efficiency of human efforts in the widest sense’.

According to **Tippet**, “Statistics is both a science and an art.” It is a science in that its methods are basically systematic and have general application and art in that their successful application depends, to a considerable degree, on the skill and special experience of the statistician, and on his knowledge of the field of application.

**Scope of Statistics**

* **Statistics in matters of state**: The governments of most countries around the world collects, on a routine basis, numerical data in relation to its people, its economy, its natural resources and its socio-political condition. Statistical data and techniques of statistical analysis are immensely useful in solving economical problems, planning facilities for the welfare of the people, formulating policies for the utilization of the natural resources etc.
* **Statistics in business and commerce**: Statistics and their interpretation also have been proved to be helpful for the proper organisation of business and commerce. A manufacturing concern needs to have a clear idea about its own resources and about the likely demand for its products. Without this knowledge, it cannot properly draw up its production schedule, employ the right number and right kinds of employees and procure the right amounts of raw materials. Modern manufacturers even bank upon statistical techniques to maintain the quality of products and to minimise inspection costs in supplying lots of items to consumers. Statistics are useful to banker, insurance companies, hospitals, social workers etc.
* **Statistics in agriculture**: For a country, forecast of agricultural production—for each of the principal crops—are needed well before harvest on the basis of estimates of the total area under the crop and its average yield. These estimates are obtained from area sampling and crop-cutting experiments, which are designed strictly in accordance with statistical principles. In order to increase agricultural product on to keep pace with the increasing population, many countries are relying heavily on high-yielding varieties. In evolving new varieties, genetic studies requiring statistical tools have to be conducted.
* **Statistics in the sciences:** Each science formulates laws to explain phenomena that come up in some part of the real world. Thus the physical sciences deal with the physical properties of matter, botany deals with the world of plants, zoology deals with the world of animals, and so on. In going to formulate such laws, one has to conduct investigations and note their results. In modern times, the results are almost invariably recorded in numerical terms. This is true not only of the physical sciences, but also of the biological, social and behavioural. The new laws of science, too, almost invariably concern quantitative aspects of things. The ideas and methods of statistics have, in consequence, been useful to the sciences—in the conduct of investigations, in the analysis of findings and also in the formulation of laws. It can be used in every field of scientific research, such as psychology, economics, medicine, advertising, demography and many more.

### Objectives of Statistical Analysis

Two fundamental ideas in the field of statistics are ***uncertainty*** and ***variation***. There are many situations that we encounter in science (or more generally in life) in which the outcome is uncertain. In some cases the uncertainty is because the outcome in question is not determined yet (e.g., we may not know whether it will rain tomorrow) while in other cases the uncertainty is because although the outcome has been determined already we are not aware of it (e.g., we may not know whether we passed a particular exam).

Probability is a mathematical language used to discuss uncertain events and probability plays a key role in statistics. Any measurement or data collection effort is subject to a number of sources of variation. By this we mean that if the same measurement were repeated, then the answer would likely change. Statisticians attempt to understand and control (where possible) the sources of variation in any situation.

Statistical analysis has the following objectives:

* Defining the type and quantity of data needed to be collected
* Organizing and summarizing the data
* Analyzing the data and drawing conclusions from it
* Assessing the strengths of the conclusions and evaluating their uncertainty

**Population and Sample**

Every statistical enquiry is designed to gather information about some aggregate of individuals (individual objects or beings) rather than about the individuals themselves. ***In statistical language, such an aggregate is called population or universe.***For example, the enquiry may relate to the human population of India, the population of tigers in Southern India, the population of registered factories in West Bengal, the population of products produced in a machine, etc.

In most situations, the population may be considered infinitely large, and thus the data in hand of the enquirer will actually relate to only a small number of members of the population. ***This small number of members of the population is collectively called as sample.***The size of the sample is always less than the total size of the population.

***In short, sample is the group of individuals (object or beings) who participate in a study, and population is the broader group of individuals to whom the results will apply.***

The formal definitions of population and sample:

***Population*:** The totality of statistical information on a particular character, from all members covered by an enquiry, is called population or universe.

***Sample:*** The selected part of the population that is use to through light on the population characteristics is called sample.

## Scales of measurement

Normally, when one hears the term measurement, they may think in terms of measuring the length of something (i.e. the length of a piece of wood) or measuring a quantity of something (i.e. a cup of flour). This represents a limited use of the term measurement. In statistics, the term measurement is used more broadly and is more appropriately termed scales of measurement. Scales of measurement refer to ways in which variables/numbers are defined and categorized.

There are four common scales of measurement: ***nominal****,****ordinal****,****interval*** *and****ratio***. Each scale of measurement has properties that determine how to properly analyse the data. The properties evaluated are***identity, magnitude, equal intervals****and a****minimum value of zero*.**

* [**Identity**:](https://studyonline.unsw.edu.au/www.psychology.emory.edu/clinical/bliwise/Tutorials/SOM/smmod/scalemea/smmain.html) Identity refers to each value having a unique meaning.
* [**Magnitude**:](https://studyonline.unsw.edu.au/www.psychology.emory.edu/clinical/bliwise/Tutorials/SOM/smmod/scalemea/smmain.html)Magnitude means that the values have an ordered relationship to one another, so there is a specific order to the variables.
* [**Equal intervals**:](https://studyonline.unsw.edu.au/www.psychology.emory.edu/clinical/bliwise/Tutorials/SOM/smmod/scalemea/smmain.html)Equal intervals mean that data points along the scale are equal, so the difference between data points one and two will be the same as the difference between data points five and six.
* [**A minimum value of zero**:](https://studyonline.unsw.edu.au/www.psychology.emory.edu/clinical/bliwise/Tutorials/SOM/smmod/scalemea/smmain.html)A minimum value of zero means the scale has a true zero point. Degrees, for example, can fall below zero and still have meaning. But if you weigh nothing, you don’t exist.

**Nominal scale of measurement:** Categorical data and numbers that are *simply used as identifiers or names* represent a nominal scale of measurement. Examples of nominal data include eye colour, gender, country of birth etc. The data can be placed into categories but can’t be multiplied, divided, added or subtracted from one another. Numbers on the back of a football jersey and your Aadhar number are also examples of nominal data. If a study is conducted and Female is coded as 1 and Male as 2 or visa versa while the data are entered into the computer, then the numbers 1 and 2 represent categories of data, i.e. nominal data.

**Ordinal scale of measurement:** An ordinal scale of measurement represents *an ordered series of relationships or rank order*. Individuals competing in a contest may be fortunate to achieve first, second, or third place. First, second, and third place represent ordinal data. If X takes first and Y takes second, we do not know if the competition was close; we only know that X outperformed Y. Rating about the service of a company - very poor, poor, average, good, very good is also ordinal data. Likert-type scales (such as "On a scale of 1 to 10 with one being no pain and ten being high pain, how much pain are you in today?") also represent ordinal data. Fundamentally, these scales do not represent a measurable quantity. An individual may respond 8 to this question and be in more pain than someone else who responded 5. A person may not be in half as much pain if they responded 4 than if they responded 8. All we know from this data is that an individual who responds 6 is in less pain than if they responded 8 and in more pain than if they responded 4. Therefore, Likert-type scales only represent a rank ordering.

**Interval scale of measurement:** The interval scale is a quantitative measurement scale where there is order, the difference between values is meaningful, and the presence of zero is arbitrary. Temperature in oF or oC; pH value of solution etc. are examples of interval scale. ***The trickiest part about the interval scale is probably the fact that there is no true zero***. Consider the example of temperature in oC. If we analyze temperatures in oC, say 20-40oC and 40-60 oC, there is order and difference between variables is meaningful, but the presence of zero is arbitrary. How? **0o** on the Celsius scale is not a universal constant but instead assigned to represent the temperature at which sea water freezes or ice dissolves.Here, zero does not represent the absolute lowest value. Rather, it is point on the scale with numbers both above and below it is possible, e.g. 10 oC and -10 oC. Thus, interval scales hold no true zero and can represent values below zero. Interval data **cannot be multiplied or divided**, however, it **can be added or subtracted**. The values of the interval scale has a definite order, and the difference between them has meaning but the ratio between them doesn’t have any meaning. For example, increasing the temperature from 15 to 30 degrees Celsius doesn’t mean it’s twice as hot.

The temperature in an air-conditioned room is 16 degrees Celsius, while the temperature outside the room is 32 degrees Celsius. One can conclude the temperature outside is 16 degrees higher than inside the room. But if one say, “It is twice as hot outside than inside,” he/she would be incorrect. By stating the temperature is twice that outside as inside, one is using 0 degrees as the reference point to compare the two temperatures. Since it is possible to measure temperature below 0 degrees, one can’t use it as a reference point for comparison. One must use an actual number (such as 16 degrees) instead.

It may be noted many assessment devices within the behavioral and social sciences, e. g. intelligence scale or Likert-type scale represents ordinal data but are often treated as if they are interval data. Accordingly, average intelligence, average customer satisfaction level, average pain amount etc. are computed. Theoretically, as this represents ordinal data, this computation should not be done.

**Ratio scale of measurement:** Ratio scale is an interval scale with a true zero or origin characteristic. Thus, the ratio scale is a quantitative measurement scale where there is order, the difference between values is meaningful, and has an *absolute zero* (i.e. no numbers exist below the zero). The ratio scale has the character of origin, which is the starting or zero-point of the scale. Examples of ratio data are age, time, sales, salary, height, weight, temperature in Kelvin etc. In general, most physical measures represent ratio data. It is interesting to note that temperature in Kelvin is ratio data (although temperature in oF or oC is interval data). This is because a Kelvin scale possesses the true zero point. This means that 40K is twice hot as 20K although 40 degrees is not twice hot as 20 degrees on a Celsius or Fahrenheit scale.

The values in ratio scale can be subtracted, added, multiplied, or divided. Almost every statistical calculation can be made for ratio scale data. Both the difference and the ratio between the values of the ratio scale have meaning. For example, if we increase weight from 10 kgs to 20 kgs, it’s twice as heavy. The ratio Scale allows for unit conversion. British thermal units such as Joules, kilogram-calories, or gram-calories, are ratio scale units. These units of energy can be converted to calculate the flow of energy.

## ****Advantages of Ratio Scale****

1. A ratio scale of measurement is considered to be the most powerful of the four measurement scales because it has an absolute zero rather than an arbitrary origin. Hence, it encompasses all the properties of the other three measurement scales.
2. Ratio scale data is more informative than ordinal and nominal because the ratio units are directly comparable as they are all of equal value
3. Every statistical technique can be implemented on a ratio scale variable because it has an absolute zero point. For this reason ratio scale is used by businesses to measure sales, costs, market share, and the number of customers on a ratio scale.
4. Calculations done via ratio data are precise as it is measured using a continuous, equidistant scale that shows order, direction, and an accurate difference in the ratio units.
5. Since ratio data have a ‘true zero,’ representing an absence of the variable, it can be used in instances where you cannot have negative values. For example, height, weight, annual salary, sales, and more.
6. Unlike the other three types of data, ratio data can be added, subtracted, multiplied, and divided because it lacks negative values.
7. Ratio scale data can be used to calculate statistical measures such as mode, median, and mean; range, standard deviation, variance, and coefficient of variation.

**Properties of the four scales of measurement at a glance**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Identity** | **Magnitude** | **Equal intervals** | **Absolute Zero** |
| Nominal | Yes | No | No | No |
| Ordinal | Yes | Yes | No | No |
| Interval | Yes | Yes | Yes | No |
| Ratio | Yes | Yes | Yes | Yes |

* Only the ratio scale satisfies all the four properties.
* A ratio scale of measurement is considered to be the most powerful of the four measurement scales. One can perform almost every statistical calculation for ratio scale data.
* Nominal and Ordinal data are referred to as ***nonparametric***. These data are analyzed using nonparametric statistical techniques.
* Interval and Ratio data are sometimes referred to as ***parametric****.* These data are *a*nalyzed using parametric statistical techniques.

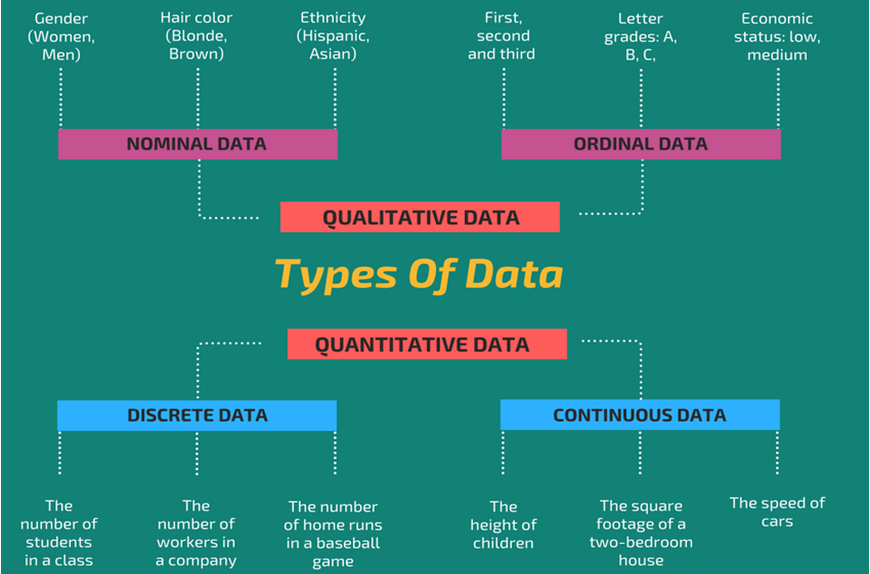
**Data**

## Data are measurements or observations that are collected as a source of information. In other word, data is a collection of distinct pieces of information (usually formatted and stored in a way that is consistent with a specific purpose).

Data can exist in various forms, e.g. text, observations, figures, images, numbers, graphs, or symbols or as bits or bytes stored in electronic memory or as facts in a person's mind.

Data is a valuable asset – so much so that it’s the world’s [most valuable](https://www.theaustralian.com.au/nation/inquirer/the-new-oil-data-is-the-worlds-most-valuable-resource/news-story/f386217a9c63ac5ee6e1473413e90bda) resource. In the business world, more companies are trying to understand big numbers and what they can do with them. **Determining the right data and measurement scales enables companies to organise, identify, analyse and ultimately use data to inform strategies** that will allow them to make a [genuine impact.](https://towardsdatascience.com/using-analytics-for-better-decision-making-ce4f92c4a025)

**Types of data**



[Qualitative data](https://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+quantitative+and+qualitative+data) refers to information about qualities, or information that cannot be measured. It’s usually descriptive and textual. For example, someone’s eye colour or the type of car they drive, sex of babies, language of books etc. In surveys, it’s often used to categorise ‘yes’ or ‘no’ answers. Nominal data and ordinal data are qualitative data.

The character viz. eye colour, car type, sex of babies, language of books etc. which cannot be expressed in numerical terms is referred to as an ***attribute*** *or qualitative character*.

[Quantitative data](https://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+quantitative+and+qualitative+data) is numerical. It’s used to define information that can be counted or measured with numbers. Some examples of quantitative data include distance, speed, height, length and weight. It’s easy to remember the difference between qualitative and quantitative data, as one refers to qualities, and the other refers to quantities. For example, a bookshelf may have 100 books on its shelves and be 100.5 centimetres tall. These are quantitative data points.

A character viz. height, weight, number of students etc. which can be expressed in numbers is referred to as a ***variable*** *or quantitative character*. For example, height, weight, number of students etc.

*Quantitative data* may be of two principal types: *Discrete data* and *continuous data*.

[Discrete data](https://studyonline.unsw.edu.au/intellspot.com/discrete-vs-continuous-data/) is a whole number that can’t be divided or broken into individual parts, fractions or decimals. Data is considered discrete if it can be counted.

Examples of discrete data include the number of pets someone has – one can have two dogs but not two-and-a-half dogs. The number of wins someone’s favourite team gets is also a form of discrete data because a team can’t have a half win – it’s a win, a loss, or a draw.

[Continuous data](https://studyonline.unsw.edu.au/intellspot.com/discrete-vs-continuous-data/)describes values that can be broken down into different parts, units, fractions and decimals. Data is considered continuous if it can be measured.

Examples of continuous data include height and weight of an object. Time can also be broken down – by half a second or half an hour. Temperature is another example of continuous data.

**Primary data and Secondary data**

It is obvious that the first step in any statistical enquiry must be the collection of the relevant numerical data. Statistical data may be of two types – Primary data and Secondary data.

**Primary data:** Data that has been generated by the researcher himself/herself, surveys, interviews, experiments, specially designed for understanding and solving the research problem at hand.

The primary data are collected for a specific purpose directly from the field of enquiry and hence original in nature. So, the primary data are also called ‘raw data’. The collection expenses of primary data are more than secondary data.

The various methods of collection of primary data are:

1. **Direct personal investigation (interview/observation)**

For example, if a **survey** is to be conducted about the workers of a factory, then the investigator will personally meet the workers of the factory and get the necessary information. The investigator personally collects the information. This method of collection of primary data is recommended / suitable when field of enquiry is small, secrecy related to data is need, high accuracy is required and time as well as money is sufficient.

1. **Indirect oral investigation**

For example if we want to get information about smoking habit of an individual, **his personal and close friends and relatives** can be enquired for reliable information. This method of collection of primary data is generally adopted to obtain the information related to the cases such as murder, theft or in the cases where a person hesitates to provide correct information.

1. **Data from local agents and correspondents**

For example, television, newspaper and radio agencies appoint local agents and correspondents to collect information. The agents collect information in their own way. This method of collection of primary data is suitable if

• Area of investigation is large

• Time period in which information is needed is very short

• Information is needed on regular basis

1. **Mailed questionnaires**

In this method, first of all a list of questions related to the information required by the investigator is prepared. After preparing the ‘final list’ of questions known as questionnaire, it is sent through mail to the informants. With this questionnaire a covering letter in which it is requested to the informants that please sent it back after completing and a brief introduction about the objective of the investigation is also attached. This method is suitable when the informants are literate and area of investigation is large. This method is very economical as far as time, money and labour is concerned.

1. **Schedules sent through investigators**

In this method first of all a list of questions based on the information to be required is prepared like mailed questionnaires method, known as schedule. The paid investigators (also called enumerators) meet with the informants face to face and after giving a brief introduction about the objectives of the investigation they ask the answer of each question listed in the schedule. The answers provided by the informants are filled up in the schedule by enumerator themselves. This is one of the main differences of the two methods namely mailed questionnaires method and schedules method. In mailed questionnaires method answers are filled up by the informants themselves while in schedules method this job is done by enumerators. The investigators are given special training as to how they should elicit the correct information through friendly discussions. This method is suitable when area of investigation is large and informants are illiterate or semi-illiterate.

1. **Results of experiments, etc.**

**Some publications containing Primary data**

* ‘Reserve bank of India Bulletin,’ issued monthly by the Reserve Bank of India.
* ‘Indian Textile Bulletin’, issued monthly by the Textile Commissioner, India.
* ‘Monthly Coal Bulletin’ issued monthly by the office of the Chief Inspector of Mines, Dhanbad.
* Published scientific papers, etc.

**Secondary Data:** Data which have previously been collected by some agency for one purpose and are merely compiled from that source for use in a different connection. In fact, data collected by someone when used by another or collected for one purpose when used for another, will be called secondary data.

The same data are primary in the hands of the collecting authority, but secondary in the hands of another. For example, the census figures published by the Registrar General of India will be primary data, while the same data contained in any other publication will be called secondary data.

Government departments collect data on diverse topics that touch the life of the people as a matter of routine and as an essential basis of administration. Private agencies like banks and industrial concerns regularly compile figures on their assets and liabilities, number of employees, income of employees etc. The enquirer may get his material readymade from such agencies; or he may get the data in a rough form and adapt them to his needs.

**Some publications containing secondary data**

* ‘Statistical Abstract of the Indian Union’, issued annually by central Statistical Organization (C. S. O.), New Delhi
* ‘Monthly Abstract of Statistics’, issued by C. S. O.

The nature, scope and objects of a statistical investigation should be taken into account for deciding whether the data are to be collected originally or whether the available secondary data are to be utilized.

It is generally preferable to make use of the primary data from several standpoints –

1. Such data usually show detailed information and a description regarding the definition of the terms used.
2. Very often, a note on the method of collection and any approximation used are also available, so that while using these data it can be decided in advance how much reliance can be placed on these figures.
3. Secondary data usually contains errors due to transcription, rounding etc. and hence are hardly reliable.

In spite of all the advantages of primary data, secondary data are used when either due to limitations of time and money at the disposal of the investigator the data cannot be collected directly or it becomes necessary to compare the data over a period of time, or utmost accuracy is not essential.

**Scrutiny of data**

Once the data are collected, always they have to be verified for their ***homogeneity and consistency***. This verification of data is called as scrutiny of data.

Since the statistical analyses are made only on the basis of data, it is necessary to check whether the data under consideration are accurate as well as consistence.

No hard and fast rules can be recommended for the scrutiny of data. One must apply his intelligence, patience and experience while scrutinizing the given information.

## However, some approaches for scrutinizing data are presented below.

* Errors in data may creep in while writing or copying the answer on the part of the enumerator. A keen observer can easily detect that type of error.

For example, inaccuracies that arise from the dropping or shifting of a decimal point and some of those that arise from the substitution of a 1 for a 7 or vice versa, or of a 6 for a 9 or vice versa.

Height (in cm) of 10 college student: 140.9, 161.2, 153.9, 172.2, 162.9, 159.1, 147.2, 773.5, 181.5, 1590.0

* Again there may be figure which, although not impossible, are very unlikely to be true and should rouse suspicion.

For example, if 3 kilograms of rice is stated to be the daily consumption of a family of 4. We should be hesitant in accepting 30 as the age of a son when father’s age is stated to be 45.

* Sometimes the year or month to which a figure relates may be stated wrongly.

For example, industrial production figure has been given for a day which was not a working day.

* Again, there may be two or more series of figures which are in some way or other related to each other. If the data for all the series are provided, they may be checked for internal consistency.

As an example, if the data for population, area and density for some places are given, then we may verify whether they are internally consistent by examining whether the relation density = (Population/Area) holds.