**Introduction**

Who is a **Data Scientist?**

The one who knows more statistics than a computer scientist and more computer science than a statistician

**Definitions of Statistics**

Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks.

The practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data.

**History of statistics**

The Word statistics have been derived from Latin word “Status” or the Italian word “Statista”, meaning of these words is “Political State” or a Government. Shakespeare used a word Statist is his drama Hamlet (1602). In the past, the statistics was used by rulers.

Word statistics was first used by a German scholar Gotifried Achenwall in the middle of the 18th century as the science of statecraft concerning the collection and use of data by the state.

Statistics is a mathematical science including methods of collecting, organizing and analysing data in such a way that meaningful conclusions can be drawn from them. ... Events that are dealt with include everyday happenings such as accidents, prices of goods, business, incomes, epidemics, sports data, and population data.

The following are the major contributors in the field of statistics

**Thomas Bayes,** (born 1702, London England—died April 17, 1761)

English Nonconformist theologian and mathematician who was the first to use probability inductively and who established a mathematical basis for probability inference.

**Jacob Bernoulli (1654-1705) Swiss mathematician.**

His contribution to statistics is tremendous, to mention few of his works Bernoulli’s sampling, Bernoulli’s distributions

**Laplace (1749-1827, Paris, France)**

Heavily contributed in the development of differential equations, difference equations, probability and statistics. His 1812 work “Théorie analytique des probabilités” (Analytic theory of probability) furthered the subjects of probability and statistics significantly.

**Simon-Denis Poisson (1781-1840) French mathematician and physicist.**

Poisson published his “law of large numbers,” often referred to as the Poisson distribution. He used this law of probabilities to analyse the composition of juries and their reliability in in returning truthful verdicts.

**Carl Friedrich Gauss (Brunswick, Germany, 1777-1855)  
*https://www.storyofmathematics.com/images/transparent_blank.gif***In the area of probability and statistics, Gauss introduced what is now known as Gaussian distribution, the Gaussian function and the Gaussian error curve. He showed how probability could be represented by a bell-shaped or “normal” curve, which peaks around the mean or expected value and quickly falls off towards plus/minus infinity, which is basic to descriptions of statistically distributed data.

**Chebyshev (1821- 1894) Russian Mathematician.**

His theorem made a difference in statistics, which talks about the distribution of the data about mean, using standard deviations.

**Markov (1856-1922) Russian mathematician.**

Is known for his work in number theory, analysis, and probability theory. He extended the weak law of large numbers and the central limit theorem to certain sequences of dependent random variables forming special classes of what are now known as Markov chains.

Sir Ronald Aylmer Fisher (1890 – 1962) was a British [statistician](https://en.wikipedia.org/wiki/Statistics) and [geneticist](https://en.wikipedia.org/wiki/Geneticist). For his work in statistics, he has been described as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century statistics".

**Prasanth Chandra Mahalanobis (1893-1972)**

He is best remembered for the Mahalanobis distance, a statistical measure, used in multivariate analysis. Was instrumental in formulating India’s strategy for industrialization in the Second Five-Year Plan (1956–61).









**Chapter 2**

**Descriptive statistics-Pictorial and tabular Descriptions**

Descriptive statistics can be divided into two general subject areas. In this section, we consider representing a data set using visual techniques. In Sections 1.3 and 1.4, we will develop some numerical summary measures for data sets. Many visual techniques may already be familiar to you: frequency tables, tally sheets, histograms, pie charts,

**Stem-and-Leaf Displays**

Consider a numerical data set for which each xi consists of at least two digits. A quick way to obtain an informative visual representation of the data set is to construct a stem-and-leaf display

Constructing a Stem-and-Leaf Display

**1.** Select one or more leading digits for the stem values. The trailing digits become the leaves.

**2.** List possible stem values in a vertical column.

**3.** Record the leaf for each observation beside the corresponding stem value.

**4.** Indicate the units for stems and leaves someplace in the display

A stem-and-leaf display conveys information about the following aspects of the data:

**1.** Identification of a special or representative value

**2.** It gives extent of spread about the typical value

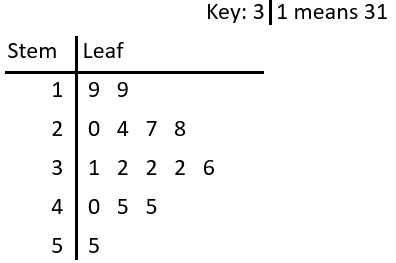
**3.** Makes you understand gaps in data if any

**4.** It gives symmetry in the distribution of values

**5.** Number and location of peaks

**6.** Outliers can be easily identified.

**Example**



**Dot plots**

A dot plot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. As with a stem-and-leaf display, a dot plot gives information about location, spread, extremes, and gaps.

**Example:** Minutes to Eat Breakfast

A survey of "How long does it take you to eat breakfast?" has these results:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Minutes: | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| People: | 6 | 2 | 3 | 5 | 2 | 5 | 0 | 0 | 2 | 3 | 7 | 4 | 1 |

Which means that 6 people take 0 minutes to eat breakfast (they probably had no breakfast!), 2 people say they only spend 1 minute having breakfast, etc.?



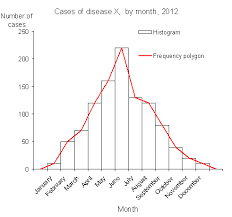
**Histogram:**

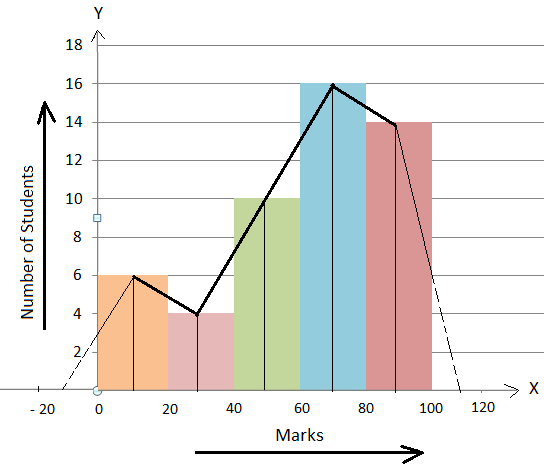
Histogram is series of rectangles each proportional in width to the range of values within a class and proportional in height to the number of items following in the class. If the classes are equal in width, then vertical bars are equal in width. The height of the bar for each class corresponds to the number of items in class.

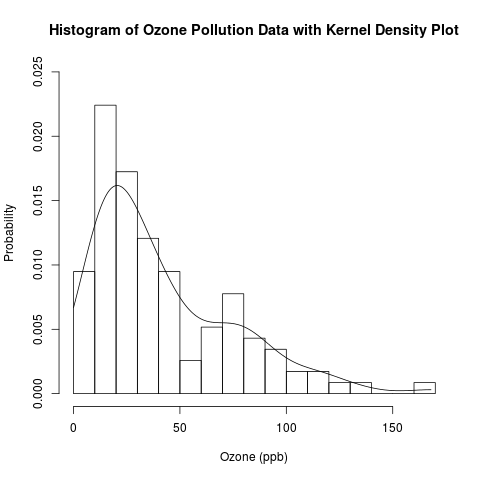
A Histogram that uses the relative frequency of data points in each of the classes rather than actual number of points is called relative frequency histogram. It has the same shape as an absolute frequency histogram.

**Frequency polygons**

A line graph connecting the midpoints of each class in a data set plotted at a height corresponding to the frequency of the class.



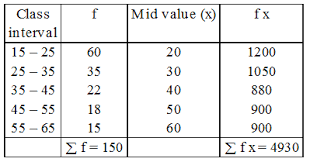




**KDE** (Kernel Density Estimation) Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.

**Frequency distribution**

An organized display of data that shows the number of observations from the data set that falls into each of a set of mutually exclusive and collectively exhaustive classes.



**Data set:** Is a collection of data.

**Data point:** is a single observation from data set.

**Data** a collection of any number of related observations on one or more variables.

**Data Array:** Is the arrangement of raw data by observation in either ascending or descending order.

**Population is** a collection of all the elements we are studying and about which we are trying to draw conclusions.

**Sample**

A collection of some, but not all the elements of the population under study used to describe the population.

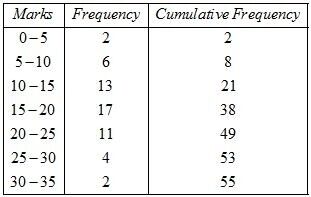
**Representative sample**

A sample that contains the relevant characteristics of population in the same proportions as they are included in that population.

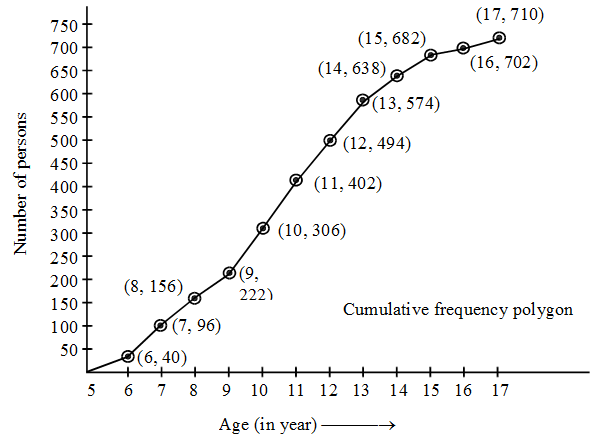
**Cumulative frequency distribution**

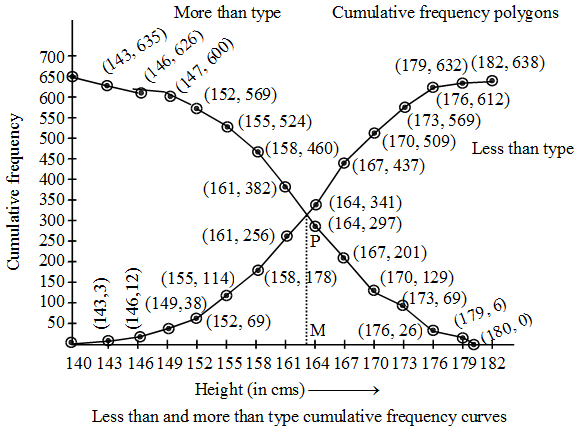
A tabular display of data showing how many observations lie above or below certain values.

Population a collection of all the elements we are studying and about which we are trying to draw conclusions.



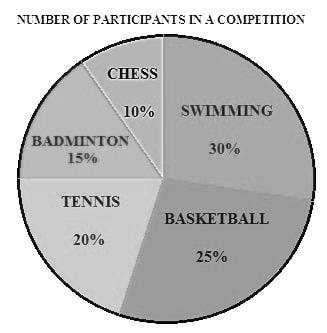
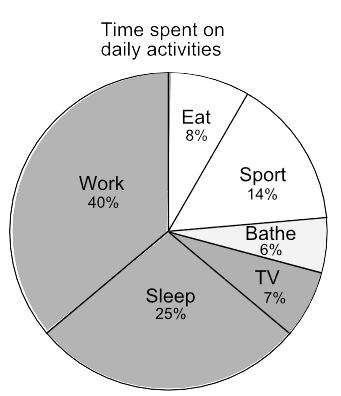
**Ogive**: Cumulative frequency polygon (curve) is called Ogive. Quartiles, Deciles, Percentiles can be easily located with the help of these curves.





**Pie Chart**

A type of graph in which a circle is divided into sectors that each represent a proportion of the whole.



**Box Plots**

Order the n observations from smallest to largest and separate the smallest half from the largest half; the median is included in both halves if n is odd. Then the lower fourth is the median of the smallest half and the upper fourth is the median of the largest half. A measure of spread that is resistant to outliers is the fourth spread

