CHAPTER 01

The Basic Concepts of Statistics

Chapter Contents

You should read this chapter if you need to learn about:

Suppose your teacher asks some questions in the class room:

- What is your Name?
- What is your height?
- What is your age?
- What is your favorite color?
- What is your class number/roll number?

Then the replies of these questions from the students are called **information** and the recording, listing or observing a single piece of information by the teacher is called as **observation**, and hence the collected observations are then collectively called **data**.

Information

"To know about something is known as information"

Observation

"Any recording of information (numeric or non-numeric) is called observation"

Data

"Originally collected observations are collectively called data".

- Data of selected student's Names: Ajmal, Arif and Ali etc.
- Data of selected student's class numbers: 39, 56 and 47 etc.
- Data of selected student's heights: 60", 65" and 66".
- Data of selected student's ages: 19, 20 and 21 etc.
- Data of selected student's favorite color: Red, Blue and Green etc.

Names	Class No.	Heights	Age	Color
Ajmal	39	60	19	Red
Arif	56	65	20	Blue
Ali	47	66	21	Green

Constant

"A fixed quantity is called a constant".

- $\pi = 3.14$
- e = 2.71 (called as **Euler's** number) etc.

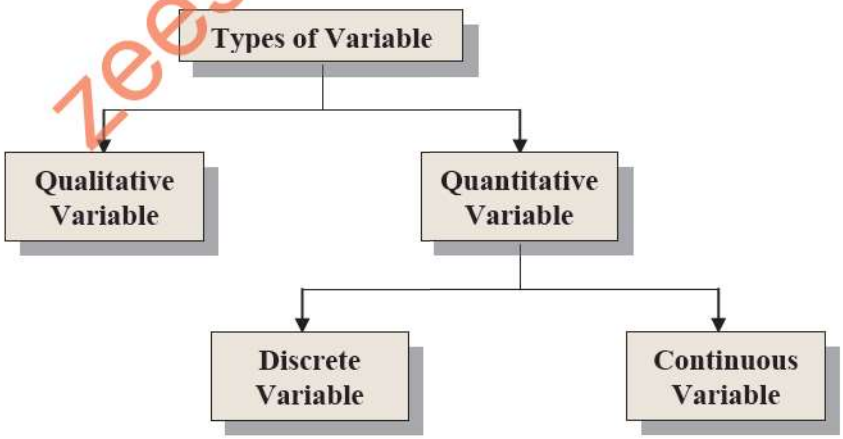
A constant is usually denoted by **first letters** of alphabets e.g. a, b or c.

Variable

"A characteristic, that can vary from one person or object to another, is called a variable".

- Height and weight of a person
- Eye color of people
- Number of children in a family etc.

Variables are usually denoted by the **last letters** of alphabets e.g. X, Y or Z.



Qualitative Variable

"A variable is qualitative if it can be expressed non-numerically"

- Color
- Religion
- Gender (Female and Male)
- Education level
- Grades of students in a class, etc.

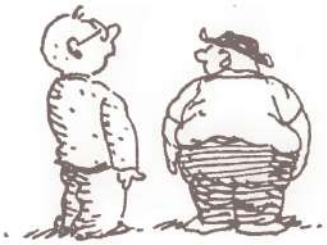
Female Male

Qualitative variables are also called attributes

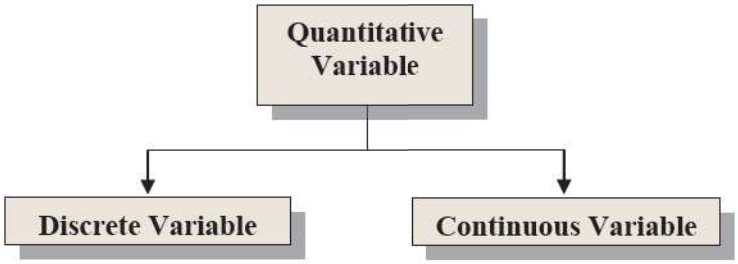
Quantitative Variable

"A variable is quantitative if it can be expressed numerically"

- Age
- Weight
- Height
- Number of children in a family
- Number of deaths in an accident
- Speed
- Income, etc.



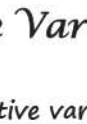
50 kg 80 kg



Discrete Variable

"A quantitative variable is called discrete variable if it has counting phenomena and there can be certain jump or gap between two possible values of the variable. Further it is free from the unit of measurement".

- Family sizes
- No. of pages in a book
- No. of apples in a basket.
- No. of deaths in an accident
- No. of housing units in different blocks of a colony
- No. of passengers carried by PIA in last ten years



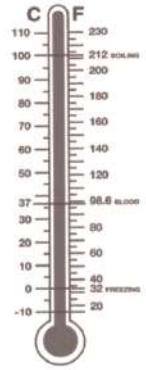
If you count to get the value of a variable, it is discrete. If you measure to get the value of the variable, it is continuous. When deciding whether a variable is discrete or continuous, ask yourself if it is counted or measured!

Further, a discrete variable takes on values that are usually integers or whole numbers, while a continuous variable takes on values that are real numbers.

Continuous Variable

"A quantitative variable is called continuous variable if it has measuring phenomena and there can be infinite number of values between two possible values of the variable. Further it has the unit of measurement"

- Students heights, ages, weights
- Speed of a car
- Temperature of a place
- Income of a family
- The amount of milk given by a cow
- The life time of a TV tube
- Fortnightly petrol prices



Sometimes, the values of variables such as age, height, and weight are usually rounded to the nearest year, inch, or pound. However, these values represent measured data, so they are continuous variables.

Individual

"An element from which information may be collected is called an individual".

Population

"An aggregate of individuals is called population".

A population can either be finite or infinite depending upon whether it contains a countable or an uncountable numbers of units.

Population

- All students in a college **(Finite)**
- The population of all licensed motor drivers **(Finite)**
- The population of all houses in a country **(Finite)**
- The population of all points on a line **(Infinite)**
- The population of stars in the sky **(Infinite)** etc.

The term population does not mean only the human population; it refers to a collection of measurements on individuals or objects having some common characteristics. The objects may be concrete (physical) things like the motor cars of a particular type produced by a company, wheat produced in a large farm or they may be abstract (theoretical) things like the opinions of students about an examination system.

Sample

"A representative part which we select from a population is called a sample".

Sample

- Runs scored by a batsman in tests, in the last one year, is a sample of his whole career scores.
- A few drops of blood, is a sample of the blood containing in the whole body of a person.

Parameter

"Any numerical value (mean, variance or standard deviation, etc.) describing a characteristic of a population is called parameter". OR

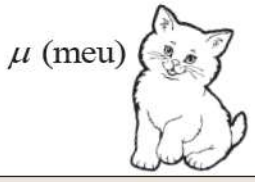
"The numerical value such as mean, variance or standard deviation etc. computed from population data is called parameter".

Statistic

"Any numerical value (mean, variance or standard deviation, etc.) describing a characteristic of a sample is called statistic". OR

"The numerical value such as mean, variance or standard deviation etc. computed from sample data is called statistic".

The words Population and Parameter both start from the letter "P" and the words Sample and Statistic both start from the letter "S".



A parameter is a fixed value while statistic is a variable because it varies from sample to sample. It is also to be noted that a parameter is usually denoted by a Greek letter and a statistic is usually denoted by a Roman letter. For example, the population mean is denoted by μ while the sample mean is denoted by $\bar{x}$. Similarly, the standard deviation of a population is denoted by σ while the sample standard deviation is denoted by S.

Sampling

"The process of selecting a sample from a population such that the sample selected has the characteristics of the whole population is called sampling".

- A teacher judge performance of his students just by asking few questions.
- If someone decides taste of the food by tasting a little bit of the food.
- In medical science a few drops of blood are taken and tested to know whether the blood contain some abnormality or not.

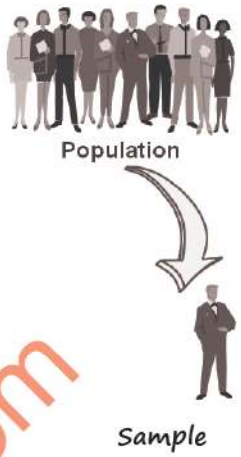
Sampling with Replacement

"If the sampling unit selected is returned to the population before drawing the next sampling unit, then sampling is said to be with replacement".

In with replacement sampling:

- The sampling unit can be selected **more than once**.
- The population will be considered **infinite**.
- The number of samples of size "n" that could be drawn with replacement from a population of size "N" will be equal to $\mathbf{(N)^n}$.
- The sampling units will be **independent**.

Sampling without Replacement



"If the sampling unit selected is not returned to the population before drawing the next sampling unit, then sampling is said to be with out replacement".

In without replacement sampling:

- The sampling unit can be selected **only once**.
- The population will be considered **finite**.
- The number of samples of size "n" that could be drawn with out replacement from a population of size "N" will be equal to ${}^{N}C_{n}$.
- The sampling units will be **dependent**.

A finite population from which sampling is done with replacement can theoretically be considered infinite because any number of samples can be drawn without exhausting (finishing) the population.

Frequency

"The number of occurrences of a particular observation in a data is called frequency".

OR

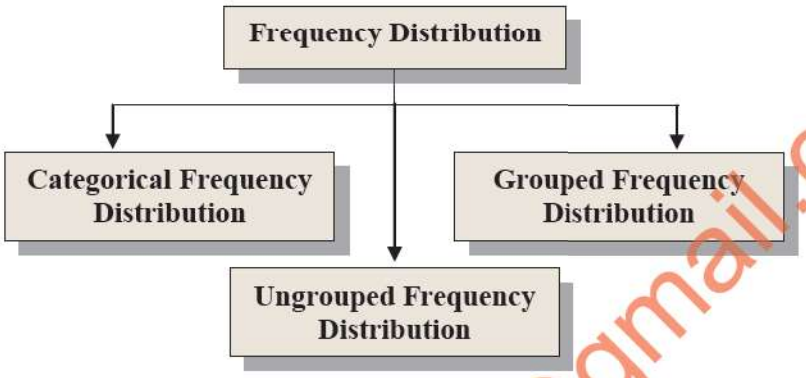
"The number of observations falling in a particular group (class) is called frequency".

cup	frequency
	2
	4
	3

Frequency Distribution

"The organization of raw data in table form, along with frequencies is called frequency distribution".

The types of frequency distributions that will be considered here are:



- *The data in the form of frequency distribution is called grouped data.*
- *The purpose of a frequency distribution is to produce a meaningful pattern for the overall distribution of the data from which conclusions can be drawn.*

- *A categorical frequency distribution represents data that can be placed in different categories such as gender, hair color, blood group etc. along with their frequencies.* The categorical frequency distribution is also called frequency table.

Categorical frequency Distribution (Categorical frequency Table)	
Blood Group	**No. of students *(f)***
A	5
B	8
O	4

- *An ungrouped frequency distribution simply lists the data values with the corresponding frequencies.* The ungrouped frequency distribution is also called discrete grouped data.

Ungrouped frequency Distribution (Discrete Grouped Data)	
Marks	**No. of students *(f)***
50	4
60	6
45	3

- *A grouped frequency distribution is obtained by constructing classes (or intervals) for the data values along with corresponding frequencies.* The grouped frequency distribution is also called continuous grouped data.

Grouped frequency Distribution (Continuous Grouped Data)	
Height	**No. of students *(f)***
3 – 4	7
5 – 6	30
7 – 8	6

Class limits	f	Mid Points
60 – 61	2	60.5
62 – 63	8	62.5
64 – 65	11	64.5
66 – 67	6	66.5
68 – 69	5	68.5
70 – 71	5	70.5
72 – 73	1	72.5
Total	38	--

Hi I am the mid point of the first class!!!

$$\text{Mid points} = \frac{60+61}{2} = 60.5$$

Class limits	f	C.F
60 – 61	2	2
62 – 63	8	8 + 2 =10
64 – 65	11	11 + 10 = 21
66 – 67	6	6 + 21 = 27
68 – 69	5	5 + 27 = 32
70 – 71	5	5 + 32 = 37
72 – 73	1	1 + 37 = 38
Total	38	--

- C.F of the first class is taken equal to the frequency of that class
- For the other classes C.F are obtained by adding each class cumulative frequency to the frequency of the next class.

Cumulative Frequency

Class limits	f	Class boundaries
60 – 61	2	59.5 – 61.5
62 – 63	8	61.5 – 63.5
64 – 65	11	63.5 – 65.5
66 – 67	6	65.5 – 67.5
68 – 69	5	67.5 – 69.5
70 – 71	5	69.5 – 71.5
72 – 73	1	71.5 – 73.5
Total	38	--

Subtract upper limit of the first class from the lower limit of the second class and divide it by "2" then subtract and add the resultant value from the lower and upper limits respectively!!

Class boundaries

Historical Note

Pascal

Chebyshev

Bernoulli

Gosset

Gauss

Origin of Statistics!!!

The word statistics has been derived from the Latin word "status" or an Italian word "statistia" or German word "statistik" meaning each word is an organized political state. It was born as the Science of Kings. It had its origin in the needs of the administrators in the ancient days for collecting and maintaining quantitative information about their population wealth and armaments (weaponry used by military). With the passage of time this word changed its shape and now is used as "statistics".

The word "statistik" was first used by **Gottfried Achenwall** (1719-1772). **Dr. Zimmerman** (1787) introduced the word Statistics into England. Its use was popularized by **Sir John Sinclair** (1754-1835) in the 1798 publication of his book on a statistical account of Scotland.

For the last few centuries, considerable interest had been developed for collection and analysis of statistical data. **Adolf Quetelet** (1796-1874) applied statistical methods in the field of education and sociology.

Outstanding contributions was also made by **Pascal** (1623-1662), **Bernoulli** (1654-1705), **Gauss** (1777-1855), **Chebyshev** (1821-1894), **Francis Galton** (1822-1911), **Karl Pearson** (1857-1936), **William Sealy Gosset** (1876-1937), **R.A Fisher** (1890-1962), **Jerzy Neyman** (1894-1981) ,**Wald** (1902-1950), **John Tukey** (1915-2000) and many others.

Historical Note

Sir Sinclair

Karl Pearson

Quetelet

Francis Galton

Neyman

Fisher

Wald

John Tukey

Meaning of the word Statistics

The word statistics is generally used in three different meanings:

- Firstly, the word statistics refers to *"numerical facts systematically arranged with a definite purpose in view"*. In this sense, the word statistics is always used in the plural e.g. statistics of prices, statistics of road accidents, statistics of crimes, statistics of births, statistics of educational institutions etc.
- Secondly, the word statistics is defined as *"the procedures and techniques used to collect, process and analyze numerical data to make inferences and to reach decisions in the face of uncertainty"*. In this sense, the word statistics is used in the singular.
- Thirdly, the word statistics are *"numerical quantities calculated from sample observations"*. The word statistics is plural when used in this sense. The mean, median, mode, etc. calculated from sample observations are the examples in this sense.

Definition of Statistics

- *"Statistics is the study of the principles and methods applied in the collection, summarization and description of numerical data. Further it deals with the procedures of making inferences about the characteristics of a population on the basis of a sample taken from the same population".*
- *"The science, which enables us to draw conclusion about various phenomena of the real life data (collected on sample basis) is called statistics".*

Descriptive Statistics

"Descriptive statistics deals with the concepts and methods concerned with the collection, summarization and description of numerical data".

By summarization we mean the **classification** of data, **tabulation** and their **graphical displays**; while the description is the computation of a few **numerical quantities** i.e. measure of central tendency, measure of dispersion, moments, skewness and kurtosis etc.

In descriptive statistics we deal with:

- Collection
- Classification
- Tabulation
- Graphical displays
- Numerical quantities

In inferential statistics we deal with:

- Estimation
- Hypothesis testing

Inferential Statistics

"Inferential statistics deals with the procedures of making inferences (conclusion) about the characteristics of a population on the basis of a sample taken from the same population".

This category consists of estimation of population parameters and testing of hypothesis.

Functions of Statistics

- The **complex mass** of data is made **simple** and **understandable** with the help of statistical methods.
- To **study relationship** between **two or more phenomena** statistical methods are used.
- Statistics helps in **formulating policies** in different fields.
- Statistical methods are **highly useful tools** for **forecasting**.
- Statistics helps in **decision making** in the face of **uncertainty**.
- One **important function** of Statistics is to provide techniques for **making comparisons**.

Scope and Importance of Statistics in Different Fields

In the ancient times the scope of statistics was limited. Census of population and wealth was conducted in those days to determine the strength of manpower and material wealth for the purpose of wars. That is why it was called the science of king.

With the passage of time the scope of statistics became wider and wider. With the development of the theory of probability, insurance companies were benefited. Thus the statistical methods began to be used in other sciences.

- Statistics plays an important role in **business**.
- The whole structure of **insurance** is based on statistics.
- The **banks** make use of statistics while framing their policies.
- Statistical data are now widely used in taking all **administrative decision.**
- Statistics has a **vast use** in Economics, Management, Industry, Transport, Communication, Physics, Chemistry, Zoology, Agriculture, Health, Atomic Energy, Petroleum, Medicine, Astronomy and many more.

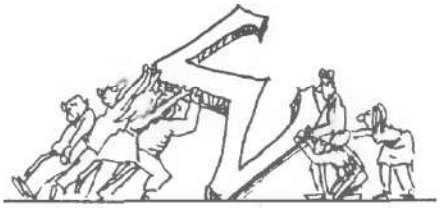
Now-a-days the science of statistics has shown it's worth so much so that there is hardly any field in which its need is not felt.

Students study statistics for several reasons!!!

- Like professional people, you must be able to read and understand the various statistical studies performed in your fields. To have this understanding, you must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in these studies.
- You may be called on to conduct research in your field, since statistical procedures are basic to research. To accomplish this, you must be able to design experiments; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. You must also be able to communicate the results of the study in your own words.
- You can also use the knowledge gained from studying statistics to become better consumers and citizens. For example, you can make intelligent decisions about what products to purchase based on consumer studies, about government spending based on utilization studies, and so on.

Summation Notation

Suppose the heights of some students are 54", 58", 64",…, 57". We can denote the height of the:



- First student by X_1
- Second student by X_2
- Last or nth student by X_n.

We can use the symbol X_i to denote any of the heights, where i = 1, 2,…,n.

Now the sum of the values X_1, X_2, …,X_n , i.e. $X_1+ X_2+ \ldots+X_n$ is denoted by $\sum_{i=1}^{n} X_i$, where the symbol $\sum$ (capital sigma) is a Greek letter and denotes sum.

Consider the following examples:

- $X_1 + X_2 + X_3 + X_4 = \sum_{i=1}^{4} X_i$
- $X_1Y_1 + X_2Y_2 + X_3Y_3 = \sum_{i=1}^{3} X_iY_i$
- $X_1^2 + X_2^2 + X_3^2 + X_4^2 = \sum_{i=1}^{4} X_i^2$
- $(X_1 + X_2 + X_3 + X_4)^2 = \left(\sum_{i=1}^{4} X_i\right)^2$
- $\sum_{i=1}^{n} a = a + a + \ldots + a = na$
- $aX_1 + aX_2 + aX_3 + aX_4 = a\sum_{i=1}^{4} X_i$
- $(X_1 - a) + (X_2 - a) + \ldots + (X_n - a) = \sum_{i=1}^{n} (X_i - a)$
- $(X_1 - a)^2 + (X_2 - a)^2 + \ldots + (X_n - a)^2 = \sum_{i=1}^{n} (X_i - a)^2$
- $[(X_1 - a) + (X_2 - a) + \ldots + (X_n - a)]^2 = \left[\sum_{i=1}^{n} (X_i - a)\right]^2$