DC-1
Semester-II

Lesson: Introduction to Statistics

Lesson Developer: Sarabjeet Kaur

Department/College: Department of Economics, P.G.D.A.V. College, University of Delhi

Contents

1. Learning Outcomes

After completing of the present chapter, you should able to:-

1. Introduction of Statistics.
2. Characteristics of Statistics.
3. Use of Statistics in Economics.
4. Difference between Population and Sample.
5. Difference between Parameter and Statistics.

2. Introduction to Statistics

Statistics is the evaluation of the collection, organization, analysis, presentation and interpretation of data to assist in making more effective decisions. It is a science of methods of obtaining and analyzing data in order to make decisions based on them. It is a branch of mathematics used in dealing with aspects that can be represented numerically or categorically either by counts or measurements.

It is widely employed in various activities of business, government, and the natural and social sciences. It is not only facts and figures; it refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.
Hence, there are five stages in a statistical investigation, explained in following diagram i.e.

(1) Collection of data:- This is the first step and is the foundation of statistical analysis. Therefore, data should be gathered with maximum care by the investigator himself (primary data) or obtained from reliable published or unpublished sources (secondary data) .

(2) Organization of data:- Data must be organized by editing, classifying and tabulation so that collected information can be easily assessable.

(3) Presentation of data: - Organized data must be presented in some systematic manner so that statistical analysis becomes easier. Data can be shown with the help of tables, graphs, and diagrams etc.

(4) Analysis of data:- After collection and organization, the data are to be reproduced by various methods used for analysis such as averages, dispersion, correlation, and interpretations etc.

(5) Interpretation of data:- Last step is interpretation of data, implies drawing of conclusion on the basis of analysis of data. On the basis of conclusion various decisions can be taken.

The word "statistics" is commonly used in two ways, in the first way, "statistics" is used in plural sense meaning numerical facts or data, called as "Descriptive Statistics". It deals with collecting, analyzing, and clarifying data; which are otherwise quite unwieldy and immense. It seeks to achieve this in a method that significant decisions can be easily obtained from

the data. It may; thus; seen as encompassing methods by bringing out and feature the latent characteristic present in a set of numerical data. It not only makes easier understanding of the data and systematic reporting thereof in a manner that makes them manageable for further consultation, investigation, and evaluation.

For example, the NSSO reports the population of the India was 449.6 million in 1960; 555.2 million in 1970; 699 million in 1980; 868.9 million in 1990; 1.042 billion in 2000 and 1.206 billion in 2010, this information is an example of descriptive statistics. We call it as descriptive statistics, if we estimate the percentage growth from one decade to decade. However, we cannot call as descriptive statistics, if we use these to find out the population for the year 2020 or percentage growth of population from the year 2010 to 2020; because these statistics are not being used to calculate past population but to predict future population.

Masses of unorganized data (e.g., census of population, earnings of workers etc.) are of fewer values. However, statistical methods are available to arrange this sort of data into a useful form. Data can be arranged into a frequency distribution. Different graph may be used to describe data. A well-thought and analytical data grouping makes possible easy description of the hidden data characteristics by means of variety of summary measures. These includes measures of central tendency, dispersion, etc, it make the necessary scope of descriptive statistics.

Today, with the development of probability theory, statistics is used to make prophecy or comparison about a totality of observations (or population) using data collected about a very little portion of that population. This technique is called as "Inferential Statistics". It is also known as Inductive Statistics or Statistical inference.

It is the technique of finding conclusions from the set of data that are subject to random variation; for instance; sampling variation. Most particularly, the term inferential statistics is used to define systems of procedures that can be helpful in drawing conclusions from datasets arising from systems influenced by random variation; for example; experimental errors, random sampling, or random experimentation. First and foremost requirements of such system of procedures for inference and induction are that the system should be able to provide reasonable answers when applied to well-defined situations and that it should be general enough to be applied to all type of situations. These statistics are basically used to test hypotheses and make estimations using sample data. The two branches of inferential statistics are estimation and hypothesis testing. The result of inferential statistics can be

helpful in making decision about further experiments or surveys, or about drawing conclusion before implementing any organizational or governmental policy.

2.1 Graphical Presentation of Data

We often present statistical information in a graphical form. A graph is often useful for capturing reader attention and to portray a large amount of information. This method can be used to illustrate the way in which one property changes when some other property undergoes a measured change. For the visualization of data, there are a number of types of graphs. They are given below;

- **Bar Graph:** A graphical method of presenting qualitative data that have been summarized in a frequency distribution or a relative frequency distribution.
- **Pie Chart:** A graphical device for presenting qualitative data by subdividing a circle into sectors that correspond to the relative frequency of each class.
- **Dot Plot:** A graphical presentation of data, where the horizontal axis shows the range of data values and each observation is plotted as a dot above the axis.
- **Histogram:** A graphical method of presenting a frequency or a relative frequency distribution or a density distribution.
- **Ogive:** A graphical method of presenting a cumulative frequency distribution or a cumulative relative frequency distribution.
- **Scatter Diagram:** A graphical method of presenting the relationship between two quantitative variables. One variable is shown on the horizontal and the other on the vertical axis.

Basically, frequency distribution is simply a grouping of the data together, generally in the form of a frequency distribution table, giving a clearer picture than the individual values. The most usual presentation is in the form of a histogram or a frequency polygon that is represented by dot plot.

A Histogram is a pictorial method of representing data. It appears similar to a Bar Chart but has two fundamental differences:

1. Type of data: Bar graphs are usually used to display "categorical data", that is data that fits into categories. Histograms on the other hand are usually used to present "continuous data", i.e. data that represents measured quantity where, at least in theory, the numbers can take on any value in a certain range.

2. Presentation of Data: The difference in the way that bar graphs and histograms are drawn is that the bars in bar graphs are usually separated; whereas, in histograms the bars are adjacent to each other, this is not always true however. Sometimes you see bar graphs with no spaces between the bars but histograms are never drawn with spaces between the bars.

2.2 Histogram and Discrete Variable

Histograms allow a visual interpretation of numerical data (discrete or continuous) by indicating the number of data points that lie within a range of values, called a class or a bin. The frequency of the data that falls in each class is depicted by the use of a bar. For each class, a rectangle is constructed with a base length equal to the range of values in that specific group, and an area proportional to the number of observations falling into that group. This means that the rectangles will be drawn of non-uniform height. When the variables are continuous, there are no gaps between the bars but when the variables are discrete; sometimes gaps should be left between the bars. For example, following data represents the number of children in a particular family:

children	frequency
1	3
2	8
3	10
4	2
5	3

For the construction of histogram, relative frequency have to be calculated for each of the observation; which is equal to:

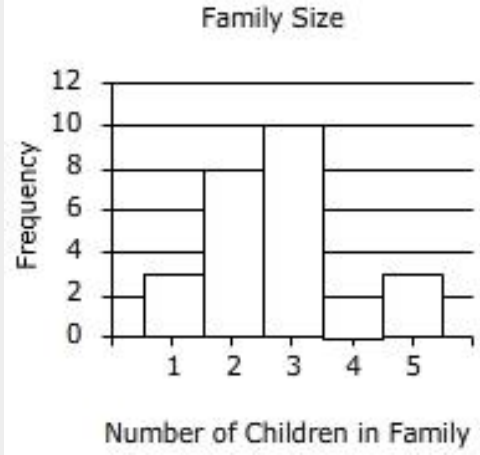
Relative frequency = $\frac{frequency\ of\ the\ observation}{number\ of\ observations\ in\ the\ data}$

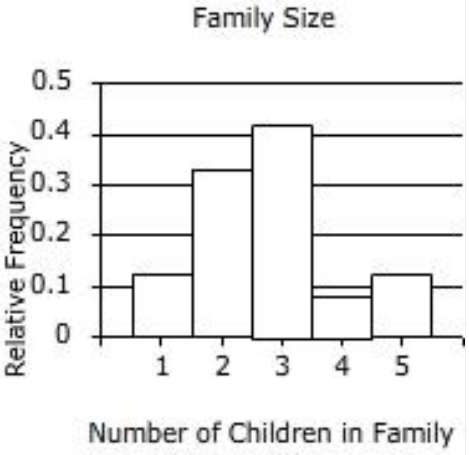
Calculating the relative frequency of the above set of data:

children	frequency	relative frequency
1	3	3/26 ≈ 0.12
2	8	8/26 ≈ 0.31

3	10	10/26 ≈ 0.38
4	2	2/26 ≈ 0.08
5	3	3/26 ≈ 0.12

A histogram is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and the rectangles touch each other. The corresponding histogram of the above example is:





2.3 Histogram in Continuous Variable and equal class interval

Histogram are commonly used in case of continuous variable

Steps in construction

- Firstly, construct a Frequency Distribution Table.
- The width of all intervals is equal.
- Construct the frequency distribution table by the help of Tally charts.
- Intervals are often left the same width but if the data is scarce at the extremes then classes may be joined.

- If the intervals are not all the same width, calculate the frequency densities.
- Construct the histogram labeling each axis carefully.
- Hand drawn histograms usually show the frequency or frequency density vertically.

Histogram with unequal class intervals:

When classes have unequal widths, the vertical axis of a histogram must represent not frequency (number of occurrences) but frequency density (relative frequency divided by its class width), and the class widths must be accurately represented on the horizontal axis, so that the area of each bar (not the height) represents the frequency of that class. The frequency density shows the number of units vertically for every unit horizontally.

2.4 Steps in the construction of histogram when class intervals are unequal

Find the relative frequency of all observations

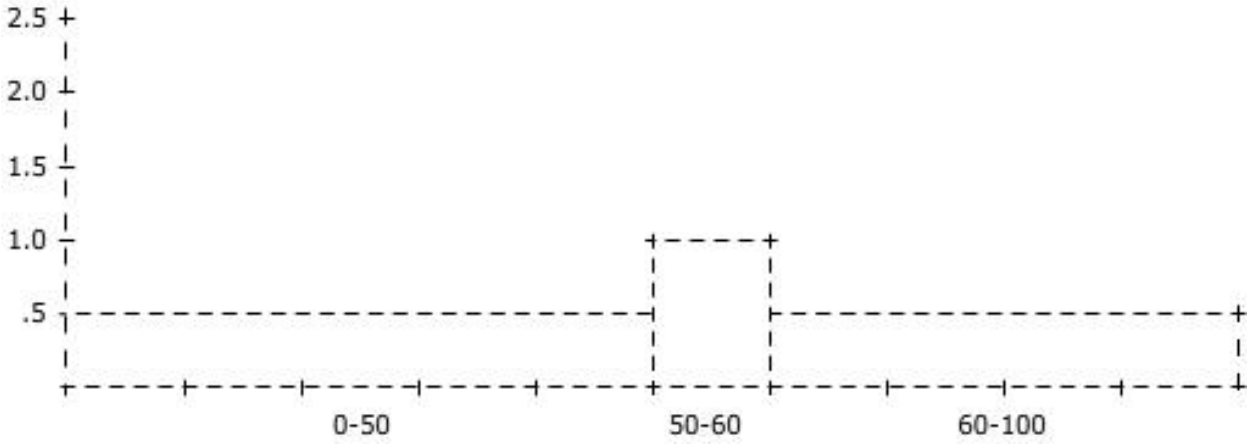
1. Divide the relative frequency of each observation by the corresponding class width to get the frequency density.
2. Construct the histogram with frequency density as the height of the rectangle and class intervals as the base of the triangle
    - (Note; if the frequency distribution is inclusive, convert them into exclusive
    - If mid values are given, find out the lower and upper limits of the various classes before constructing the histogram.)

Consider the following example:

Score	Frequency
0 - 50	25
50 - 60	10
60 - 100	20

Solution:

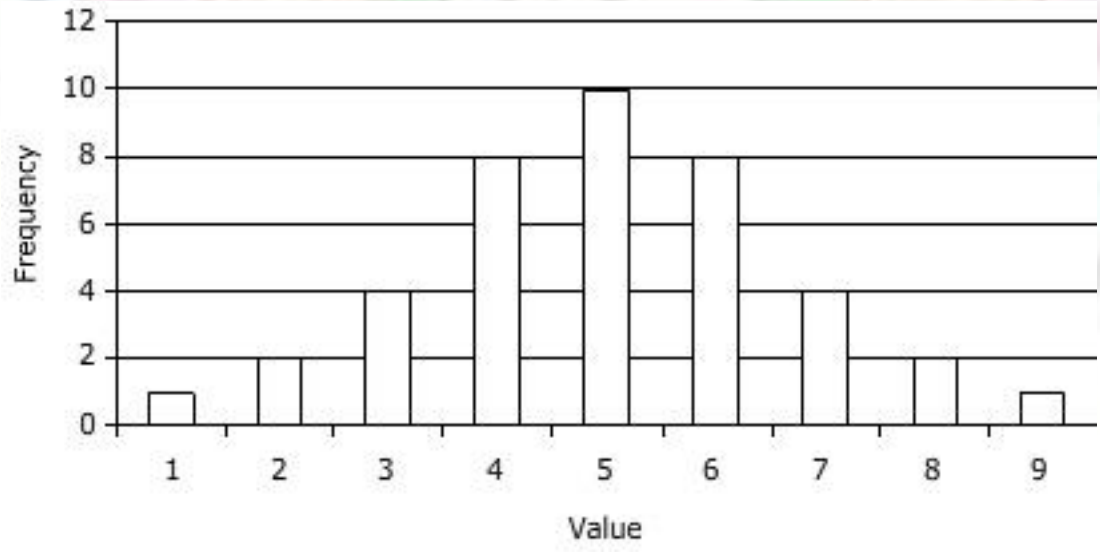
Score	Frequency	Density
0 - 50	25	25/50 = 0.5
50 - 60	10	10/10 = 1.0
60 - 100	20	20/40 = 0.5



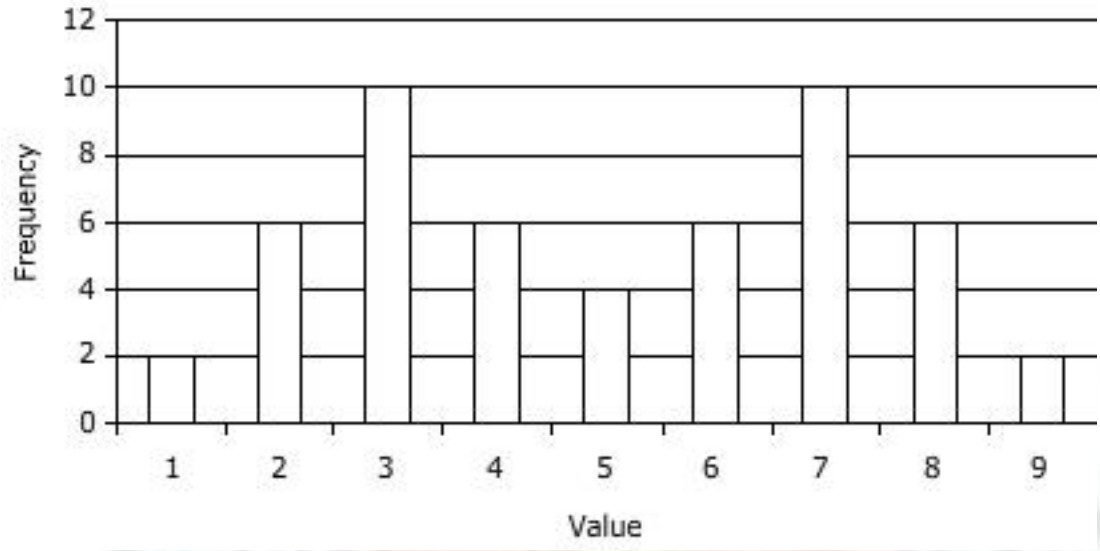
The area of each rectangle is the relative frequency of the corresponding class. Since, the total of relative frequency is always equal to one; hence, total area of all rectangles in a density histogram is equal to one.

2.5 Histogram Shapes

Symmetric: A histogram is symmetric if right half of histogram is exactly equal to left half. i.e, the two sides of the distribution are a mirror image of each other. For example, in a normal distribution, points are as likely to occur on one side of the average as on the other.



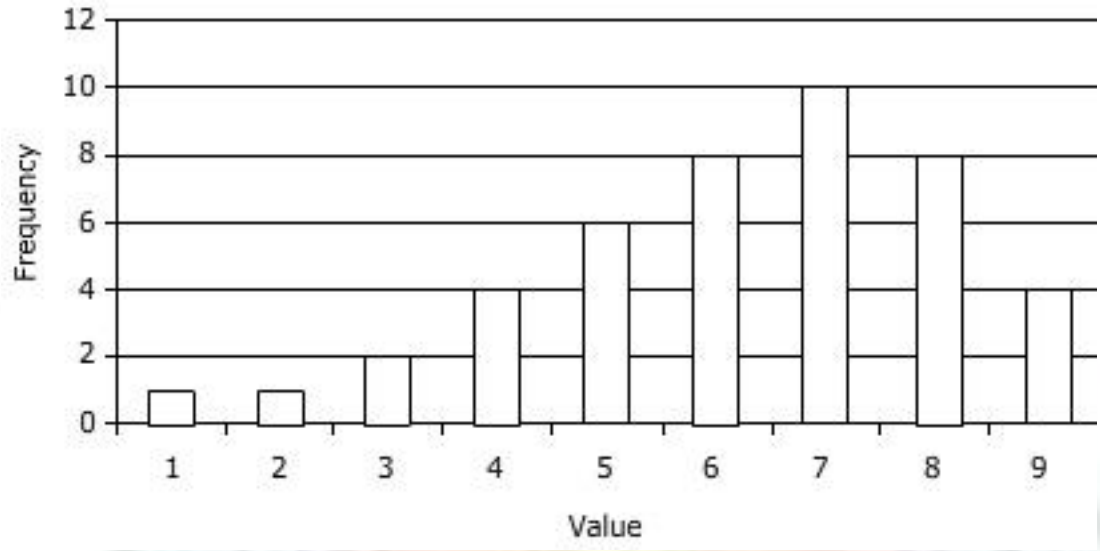
A **Biomodel** histogram has two peaks. This happen when data having two different kinds of individuals or objects. For example, a distribution of production data from a two-shift operation might be bimodal, if each shift produces a different distribution of results.



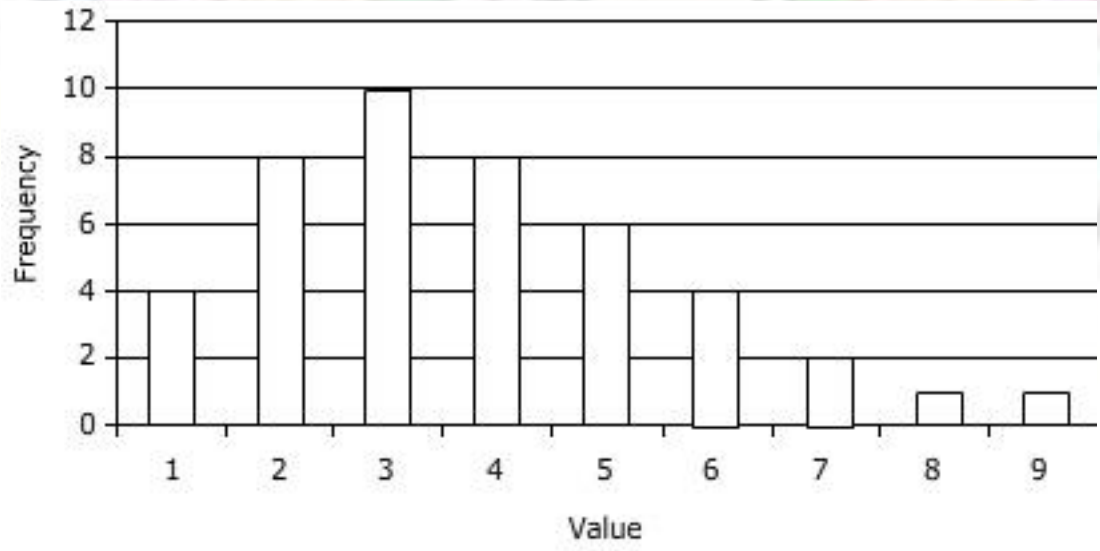
Asymmetric Histogram: An asymmetric histogram is not equally balanced. In other words, the two sides will not be mirror images of each other. Skewness is the tendency for the values to be more frequent around the high or low ends of the x-axis. When a histogram is constructed for skewed data it is possible to identify skewness by looking at the shape of the distribution.

A distribution is said to be positively skewed when upper tail is stretched towards right as compare to the left side which means that majority of data has values towards the upper end of its range. Most of the values tend to cluster toward the left side of the x-axis (i.e. the smaller values) with increasingly fewer values at the right side of the x-axis (i.e. the larger values).

For example, the distribution of personal income is positively skewed. Also, raw scores on most measures of psychopathology are positively skewed.



A distribution is said to be negatively skewed when the tail on the left side of the histogram is longer than the right side. Most of the values tend to cluster toward the right side of the x-axis (i.e. the larger values), with increasingly less values on the left side of the x-axis (i.e. the smaller values).When upper tail is stretched towards right as compare to the left side which means that majority of data has values towards the lower end of its range. For example, a distribution of analyses of a very pure product would be skewed, because the product cannot be more than 100 percent pure.



Frequency polygon

It is constructed by joining the midpoints at the top of each column of the histogram. The final section of the polygon often joins the midpoint at the top of each extreme rectangle to a point on the x-axis half a class interval beyond the rectangle. This makes the area enclosed by the rectangle the same as that of the histogram.

2.6 Examples of Dot Plot

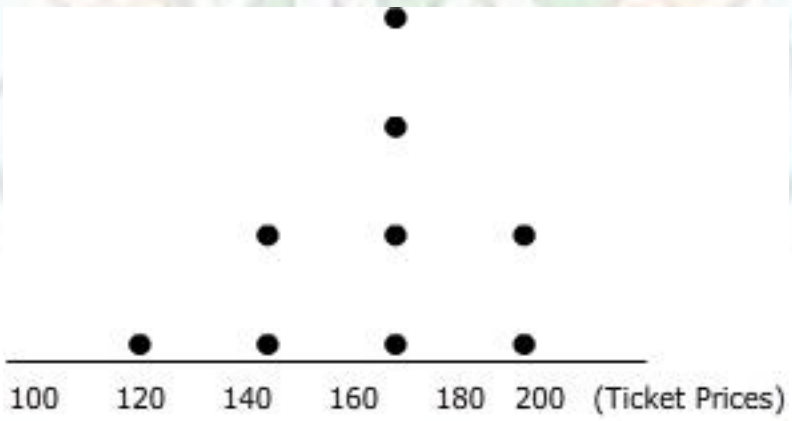
In a recent campaign, many airlines reduced their summer fares in order to gain a larger share of the market. The following data represent the prices of round-trip tickets from Atlanta to Boston for a sample of nine airlines.

120	140	140
160	160	160
160	180	180

Construct a dot plot for the above data.

Answer: The dot plot is one of the simplest graphical presentations of data. The horizontal axis shows the range of data values, and each observation is plotted as a dot above the axis. The figure shows the dot plot for the above data. The four dots shown at the value of 160 indicate that four airlines were charging Rs.160 for the round-trip ticket from Atlanta to Boston.

DOT PLOT FOR TICKET PRICES



2.7 Examples for Histogram and Ogive

The following information of waiting times at first county bank

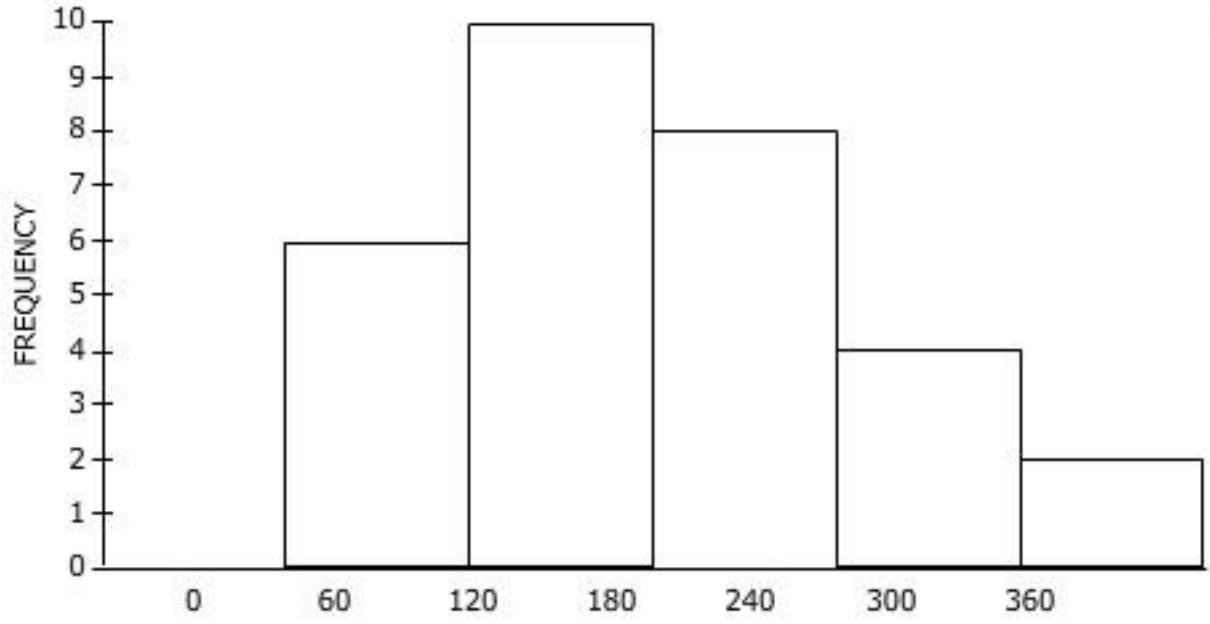
Waiting Times (Seconds)	Relative Frequency	Relative Frequency	Percentage
60 - 119	0.2000	0.2000	20.00
120 - 179	0.3333	0.5333	53.33
180 - 239	0.2667	0.8000	80.00
240 - 299	0.1333	0.9333	93.33
300 - 359	0.0667	1.0000	100.00

Construct (i) Histogram for the waiting times

(ii) Ogive for the waiting times

Answer (i)

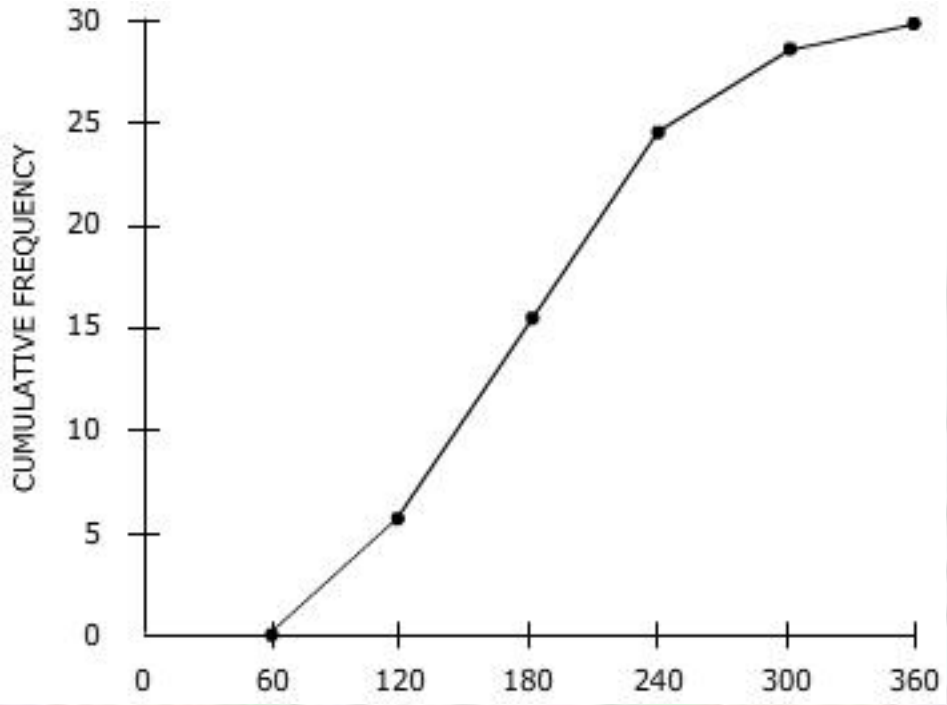
Histogram of the waiting times at first county bank



Answer (ii)

Ogive for the cumulative frequency distribution of the waiting times at first county bank

Waiting Times (in seconds)

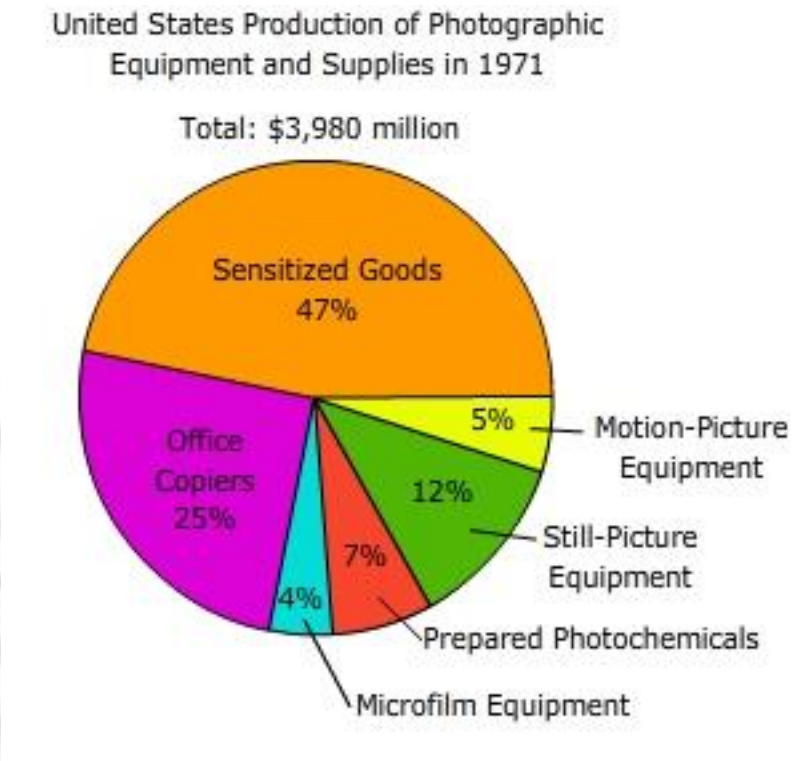


Waiting Times

Pie Graphs

It illustrates how a whole is separated into parts. The data is presented in a circle such that the area of the circle representing each category is proportional to the part of the whole that the category represents.

For example, a circle graph is shown in Data Analysis given below. The title of the graph is "United States Production of Photographic Equipment and Supplies in 1971". There are 6 categories of photographic equipment and supplies represented in the graph.



3. USE OF STATISTICS IN ECONOMICS

Statistics deals with every aspects of human activity. Statistics holds an important position in different fields like Commerce, Industry, Chemistry, Economics, Mathematics, Biology, Botany, Psychology, Astronomy etc, Therefore, application of statistics is very wide.

A number of economics problems can be easily understood by the use of statistical tools. It helps in formulation of economic policies. Statistical data and advanced techniques of statistical analysis are immensely useful in the solution of variety of economic problems such as production, consumption, distribution of income, wealth prices, saving investment, unemployment etc. For instance, the analysis of consumption pattern of the people may reveal pattern of income spent on different heads of consumption by collecting relevant information.

Statistical study is a quantitative tool mostly used within the economic area and is always needed in a variety of ways such as determine the effectiveness of economic theories with the help of the study of empirical real-world data, explaining cause-effect relationships between variables for the use of assisting in the making of powerful public policy, estimating the future actions of necessary economic conditions for the purpose of minimizing uncertainty in making up of different business or public policy decisions, or adopting mathematical models by incorporating actual data.

National income accounts are multipurpose measure for the administrators and economists. Various statistical measures are used for construction of these accounts. In economics research, statistical measures are used for collecting, organizing and analysis of the data and testing hypothesis on it.

In the field of production, comparative study of productivity of various elements of production, i.e. land, labour, capital, and enterprise can be done with the help of statistics. The effectiveness of various policies can be easily done with various statistical techniques. It also plays very important role in trade, both internal and external, where data on cost and selling price helps in making demand of the commodity.

In short, statistics is very useful in every field of economics. It provides facts, direction to solve a problem, evolution of economic laws, and also helps in economic planning**.**

4. Limitations of Statistics

Despite of usefulness of statistics in almost all sciences - social, physical and natural, impressions should not be carried statistics can be used as a big magic which gives us the accurate results to the problems. In spite of the wide scope of the subject it has certain limitations and nevertheless the data neither properly collected nor interpreted there is always chances of drawing wrong conclusions. Therefore, it is necessary to know the limitations of statistics. Some important limitations of statistics are the following:

(i) Statistics does not study individuals:

Statistics deals with aggregate of facts. Single or isolated figures are not statistics**.** Data are statistical when they correlate to computation of masses, not statistical when they correlate to an individual item or phenomenon as a different entity. This is considered to be a major handicap portion of statistics.

(ii) Statistics does not study qualitative phenomena:

Statistics are numerical statement of facts and figures. It is not applicable to the study of those facts that are not quantitatively measurable. These attributes cannot be explained in numbers. Qualitative phenomena, e.g., honesty, intelligence, poverty, etc, cannot be studied in statistics unless these attributes are expressed in terms of numerals. So, the quality aspect of a variable or the subjective phenomenon falls out of the scope of statistics. Hence, this limits the scope of the subject.

If there are k quantitative characteristics and we wish to focus on one of them, we can assign the no. 1 to that and 0 to all of the others. Counting tha1's gives us the *x* value & *x/n* is the proportion of times that a characteristic is observed. If *p = x/n* it can be used to estimate the population proportion *p=x/N*

(iii) Statistical laws are not exact:

Statistical laws are not exact as incase of natural sciences. The conclusions obtained through the phenomenon are not specifically or universally true, they are true but only under some conditions. This is because statistics as a science is less exact compare to natural sciences. So, statistics has less practical utility.

5. POPULATION AND SAMPLE

5.1 Population

The term "Population" normally means persons in a town, region, state, or country and their respective attributes such as gender, age composition, marital status, educated and so on. In statistics, the term "population" used in a different sense. It not only concerned with number of people living in a area, but it also covers the population of households, a population of events, objects, procedures or observations, including services like visits to the doctor, or surgical operations. A population is thus a totality of creatures, events, things, cases and so on. In short, a unit of population is whatever you count or measure. Normally, a population should be relatively large in size and hard to infer some attributes by considering its elements individually. It is impossible to theoretical survey the entire population because all the members are not observable. If it is possible to reach the entire population but it is very costly and also time consuming . Alternatively, researcher could take a subset of this population called a sample. By using this sample, conclusions can be drawn about the population under study.

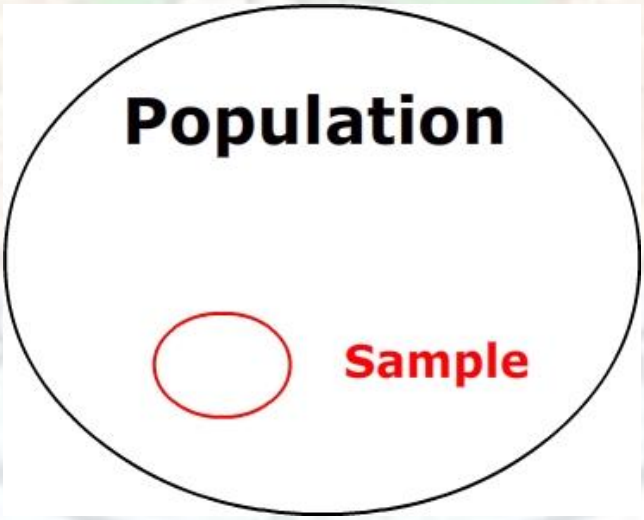
A census may be preferred when the size of population is not too large. It may be desirable to take the recourse to a census where the respondents are not widely scattered and reliability of data is not a case when a census is just unavoidable.

A Conceptual population consists of all the values that might possibly have been observed. It is also called as tangible population. e.g., a geologist weighs a rock several times on a sensitive scale. Each time, the scale gives a different reading. Here the population is conceptual because it consists of all the readings that the scale could in principle produce.

The population of undergraduate students of a college is a real population and it exists. If you had the time to measure all the students' heights, you could. The population of Economics students is a conceptual population created over time. All of the students taking Economics this semester is clearly in this population, but are not also the ones who took it last semester, and those "eligible" to take it next semester. In that sense, it is a conceptual population; it doesn't really exist in its entirety now, but does over time.

5.2 Sample

A part of the population is called a sample. It is a proportion of the population, or a part of it and it represents all the characteristics of population. A sample is scientifically drawn group that actually possesses the same attributes of the population. It may consist of two or more items that have been selected from same population. The lowest possible size of sample is two and the highest one could be equal to the size of population. An effectively selected sample will involve most of the information about a specific population parameter but the connection between the sample and the population must be such as to enable reasonably accurate conclusions to be made about a population from that sample. The relation between population and sample can be expressed in following diagram:



Example: If researcher wants to find out the mean height of the students in a particular class room, then students in that room would represents the population. But if researcher wants to find out the mean height of the students in that particular college, the students in that particular room would represent a sample of the students in that college. The basic unit of the population in called as element of the population. Each student is an element of the college. Thus, a population is the totality of elements being studied and a sample is

part of the population.

The interesting relationship between population and sample is that the population can exist without sample, but sample may not exist without population; thus, sample depends upon population. A sample is not studied for its own sake. The basic objective of its study is to draw inference about the population.

The samples are essential because within several models of research, it is impractical (from both a strategic and a resource perspective) to examine all the members of a particular population for a research assignment. However, census taking often is expensive, too time consuming to provide information when it is needed. It is not feasible to include the whole population when elements are destroyed to obtain information. Instead, a selected few participants (which is called as sample) are chosen to ensure that the sample is true representative of the population. Hence, the result obtained from the sample can be concluded for the population, i.e., using information on a smaller group of participants to infer to the group of all participants.

Normally, certain attributes of the items in the population are too being examined, for example, the mean height of the children's in a village. A characteristic may be categorical, such as gender etc, or it may be numerical. In the former case, the value of the characteristic is a category for example female, whereas in the later case, the value is a number for example age=40 years. A variable is a characteristic that may assume more than one set of values to which a numerical measure can be assigned.

x= weight of the student.

y= age of the student etc.

Data outcomes from making consideration either on a single variable or concurrently on two or more variables.

Uni-variate data refers to data where researchers are only observing one aspect of a population or sample at a time, e.g. height of students. With two variable or bi-variate data, researchers observe two aspects and if there are more than two variables then multivariable data, e.g. height, weight, and age of students.

When sample is obtained from population, an investigator would frequently use sample information to draw some type of conclusion about the population. It is imperative that the sample is representative of the group to which it is being generalized. This branch is called

as inferential statistics.

Various statistical techniques such as random sampling, stratified sampling, systematic sampling, multi-stage sampling and quota sampling are there but most commonly used is random sampling.

In order to use statistics to learn things about the population, the sample must be random. A random sample is one in which each and every elements of a population has a fair chance of being chosen. The most commonly used sample is a simple random sample. It requires that each and every possible sample of the selected size has an equal chance of being used.

Since, simple random sampling normally does not ensure a representative sample; when population is heterogeneous; a sampling technique called stratified random sampling is used. The sample becomes more representative of the population; when this sample is selected by using this technique. This method can only be used when the population can be divided in number of distinct "strata" or groups. In stratified sampling, you first identify members of your sample who belong to each group. Then from each of the sub-group (called strata), a randomly select a sample in such a manner that the sizes of the subgroups in the sample are proportional to their sizes in the population.

It is required that the each strata used in stratified sampling must not overlap. Having overlapping subgroups will provide some elements a higher chance of being selected in the sample. If this happened, it would not be a probability sample.

In a systematic sample, the items of the population are placed in the form of a list and then every *n*th item in the list is selected (consistently) for insertion in the sample. For instance, if the population of research study includes 5,000 students in a particular college and the researcher required a sample of 500 students. The students would be put into a form of list and then every 10th student would be selected for the sample. To check against any possibility of human bias in this method, the researcher should pick the first element at random. This is called a 'systematic sample with a random start'.

6. PARAMETER AND STATISTICS

Parameter is a statistical measure evaluated from population or census data. It is a characteristic of the population. It is a value which expresses the entire population, for example, population mean, standard deviation, mode etc.

Statistic is a statistical measure calculated from sample data. It is a value that express the characteristic of a sample, for example sample mean, and it is also used to infer about the corresponding a population parameter. Hence, a sample should represent the entire population.

Examine a set of n data, x_1, x_2, …., x_n; If this set of data represents a population, then its mean value and the standard deviation of the population are given by:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

Whereas, if this same set of data represents a sample, then its mean value and its standard deviation are given by:

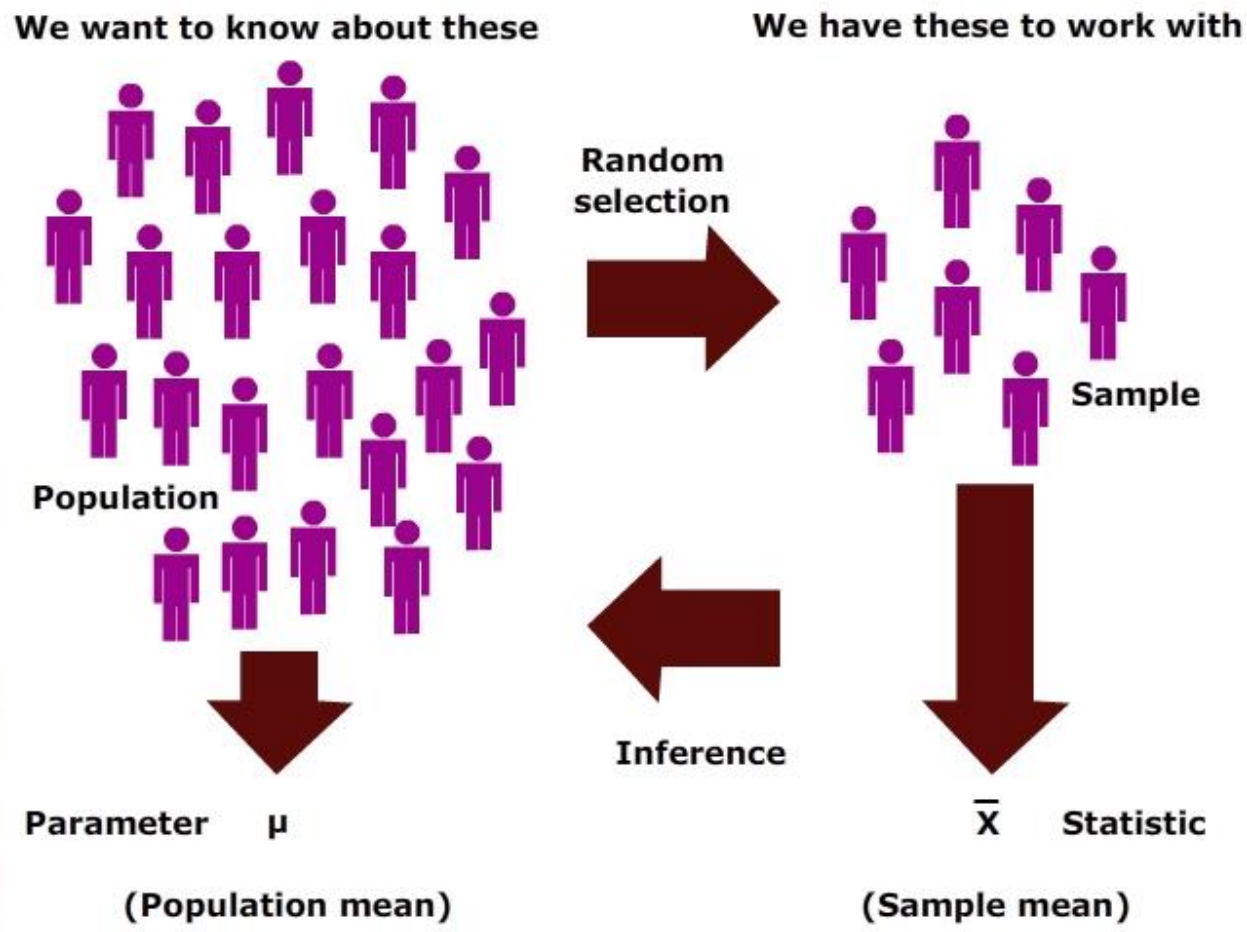
$$\dot{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\dot{x})^2}{n-1}}$$

Here, the formula for mean is same whether the data for a population or a sample. But, the formula of the standard deviation depends on the interpretation of the data as population or sample.

A statistic and parameter are very similar. A parameter is a numerical value that is equivalent to an entire population characteristic, such as mean and standard deviation, which are explanatory of entire population, are known as population parameters whereas a statistic is a numerical value that describes a sample but not the whole population are called statistic. Inferential statistics authorize researcher to make an informed guess about a population parameter based on a statistic computed from a sample randomly drawn from a particular population (see Figure)

Figure: Shows the relationship between sample and population with their statistical measures:



For instance, suppose an investigator examine the population of dogs in Delhi and if investigator wants to examine the mean height of all the dogs in the town then it is called as population "parameter". If investigator selects a sample of 50 dogs from the town and investigate mean height from that sample of dogs then it is called as 'statistic'.

A value of parameter always remains constant. It is not a random variable because all the units in a population always remain same whereas statistic is a random variable whose value varies from sample to sample because units selected in two or more samples are not the same and different sample will give different values. Variation in the value of statistics is called sampling fluctuation.

The use of statistics is a means to an end. It is a quantity that helps to determine the unknown parameters of a population based on only few observations.

It is used to estimate the degree to which sample statistics approximate the population parameters. Investigator basically bothered about the population and the estimated parameters corresponding with that population. However, investigator cannot obtain these

values, so investigator relies on sampling and statistics to provide researcher with conclusion about the population. A sufficient statistic is one that guarantees that no other statistic can produce more accurate information.

7. Exercise:

1. How does statistics help in the solution of Economic problems? Explain with examples.

Ans: Solution of various economic problems can be better analysed and understood with the help of different statistical tools as discussed in details with examples given in section 2 of the chapter.

2. Explain the difference between parameter and static with the help of examples.

Ans: Parameter is a statistical measure computed from population data whereas static is a measure of sample as given in section 4 in detail of this chapter.

3. Explain the importance of Statistics in Economics. What are the limitations of it?

Ans: Statistics holds a central position in Economics detailed given in section 2.2 and 2.3.

4. Explain the importance of Sample. How is it helpful in research?

Ans: small in size, less time consuming etc. given in section 3.

5. Explain histogram. What are its different shapes?

Explained in section 2.1

8. References:

1. Jay L. Devore, Probability and Statistics for Engineers, Cengage Learning, 2010.
2. John E. Freund, Mathematical Statistics, Prentice Hall,1992.
3. Earl k. Bowen and Martin K. Starr, Basic Statistics for Business and Economics, McGraw-Hill, 1983.