Statistics for Data Analysis

By

Dr. Mohammad Rafiqul Islam

Associate Professor of Statistics

Dept. of Mathematics and Natural Sciences (MNS)

BRAC University, Dhaka, Bangladesh

mrafiq@bracu.ac.bd

Unit one: Statistics and Data

1.1 What is Statistics?

The term statistics is believed to have been derived from the Latin word "Statistic". In early days, it was used only of the collection of the information of the population of the state military. But in the modern time it is used in almost all aspects of human related activities. Statistics is one of the most useful mathematics topics. Nearly every kind of occupation and human activity can be benefitted from an application of statistics. In the most general sense, statistics describes a set of tools and techniques that can be used to describe, organize, and interpret information or data.

Statistics is a branch of applied mathematics that involves collecting and organizing data for interpretation and the prediction of future behavior or results.

Statistics refers to numerical numbers relating to any field of inquiry. For Example:

- Statistics of export, import and GDP.
- Statistics of prices, income and expenditure.
- Statistics of Births and deaths etc.

According to R.A. Fisher, who is known as the father of statistics, “Statistics has been defined as the science and art of collection, organization, analysis and interpretation of numerical data.”

Definition: Statistics has three primary components:

First: How best can we collect data?

Second: How should it be analyzed? And

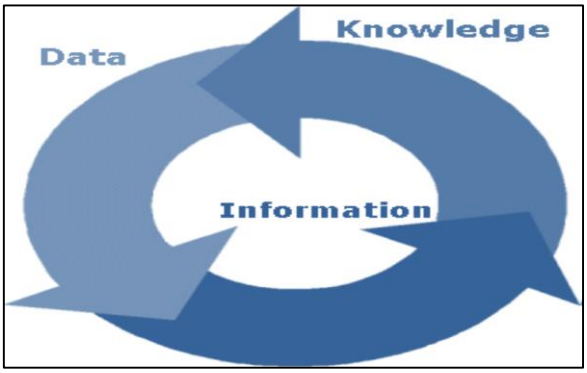
Third: What can we infer or conclude from the analysis?

This can be illustrated as follows:

- **Collect the data**
- **Organize (summarize) the data**
- **Present the data**
- **Analyze, and**
- **Interpret data**

1. **Collection of data:** The collection of data is the first step of statistical investigation. It must be collected very carefully. So, the data must be covered, if not the conclusion will not be reliable.
2. **Organization:** The data may be obtained either from primary source or the secondary source. If the data is to be obtained from the primary source, then it needs organization. The data are organized by editing, classifying and tabulating them.
3. **Presentation:** After the collection and organization of data, they are presented in systematic form such as table, diagram and graphical form.
4. **Analysis:** After the collection, organization and presentation of data, the next step is to analyze the data. To analyze the data we use average, correction, regression, time series etc. The statistical tools of analysis depend upon the nature of data.
5. **Interpretation:** The last step of a statistical method is the interpretation of the result obtained from the analysis. Interpretation means to draw the valid conclusion

Statistics helps analyze, understand and explain the data around us to get knowledge and information. Based on the analysis, we have to answer the questions and finally will make the decisions.



Statistics shortly- Statistics Cycle

Statistics is the branch of mathematics that deals with analyzing data and then interpreting the results, and then one can inform of what is going on about the world around us and understand it better. So scientific research is often advanced by posing scientific questions and then answering these, and statistics answers them by actually using data which can be shown by the following diagram:

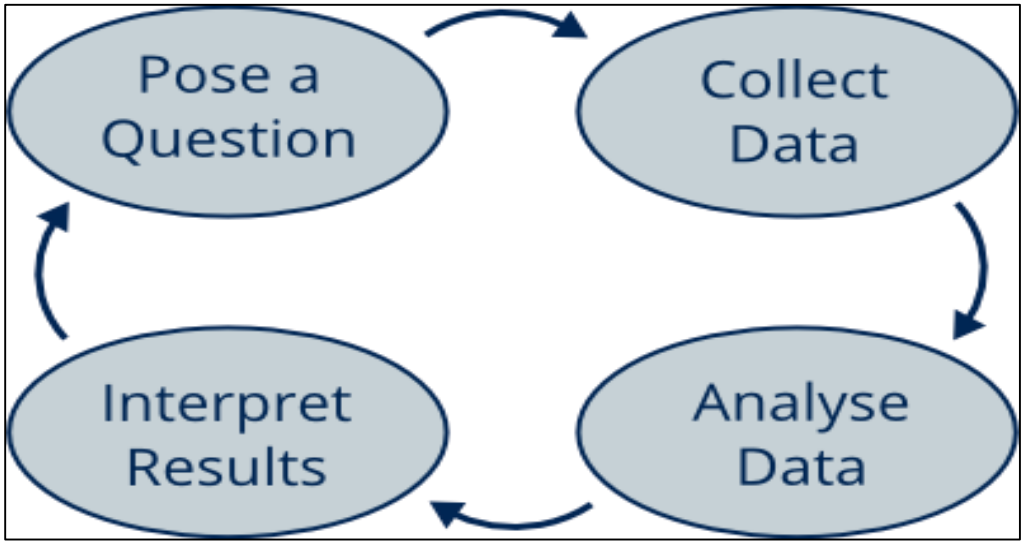


Fig: Statistics Cycle

This diagram is known as the statistics cycle. It is very useful tool for designing, conducting and interpreting a statistical experiment. Typically for any finding or research is starting with a question which we wish to answer. For the given the question we collect relevant data to investigate this question. Using the data we conduct an appropriate statistical analysis and interpret the results accordingly. Gathering the updated understanding of the world from the results obtained we may pose a further follow-up or in-depth question that we wish to investigate, and so on...

Uses /Scopes of statistics

Jerzy Neyman (1894-1981), one of the principal architects of modern statistics, used to say "Statistics is the servant to all sciences."

Researchers from a wide array of fields have questions or problems that require the collection and analysis of data. Let's consider three examples.

• Climate scientists: how will the global temperature change over the next 100 years?

• Psychology: can a simple reminder about saving money cause students to spend less?

• Political science: what fraction of Bangladeshisupport of the job ruling part is doing?

While the questions that can be posed are incredibly diverse, many of these investigations can be addressed with a small number of data collection techniques, analytic tools, and fundamental concepts in statistical inference.

There is hardly any domain of human activity today, where numerical and measurements
Do not play a part, of greater or lesser importance.
Now a-days, modern statistical methods are providing indispensable as aides

- In the physical and biological sciences,
- In economics and sociology,
- In psychology and education,
- In medicine and agriculture, and
- In government and industry.

- Life insurance premiums and annuity payments are determined from mortality table based on statistical records.
- Business executives and government functionaries use statistical methods in decision making

Statistics in Business and Industry

Statistics are used in almost every industry including insurance, consumer products, retail, pharmaceuticals and even the federal government. Statistics are important in industry and business for a number of reasons.

1. *Marketing:* Statistical analysis are frequently used in providing information for making decision in the field of marketing it is necessary first to find out what can be sold and the to evolve suitable strategy, so that the goods which to the ultimate consumer. A skill full analysis of data on production purchasing power, man power, habits of compotators, habits of consumer, transportation cost should be consider to take any attempt to establish a new market.

2. *Production*: In the field of production statistical data and method play a very important role. The decision about what to produce? How to produce? When to produce? For whom to produce is based largely on statistical analysis.

3. *Finance:* The financial organization discharging their finance function effectively depend very heavily on statistical analysis of peat and tigers.

3. *Banking*: Banking institute have found if increasingly to establish research department within their organization for the purpose of gathering and analysis information, not only regarding their own business but also regarding general economic situation and every segment of business in which they may have interest.

4. *Investment*: Statistics greatly assists investors in making clear and valued judgment in his investment decision in selecting securities which are safe and have the best prospects of

yielding a good income.

5. *Purchase:* the purchase department in discharging their function makes use of statistical data to frame suitable purchase policies such as what to buy? What quantity to buy? What time to buy? Where to buy? Whom to buy?

6. *Accounting:* statistical data are also employer in accounting particularly in auditing function, the technique of sampling and destination is frequently used.

7. *Control:* the management control process combines statistical and accounting method in making the overall budget for the coming year including sales, materials, labor and other costs and net profits and capital requirement.

8. Sales: Statistics is important in business and industry for companies to develop sales forecasts one, two and even five years in the future. Companies can then modify or improve their products, hire additional sales representatives and procure the necessary resources to attain these forecasted sales targets. Most businesses and even competitors in a certain industry use sales forecasting statistics to develop their business and marketing plans.

Importance of Statistics

The importance of statistics can be defined in different parts i.e. statistics in planning in economics, in business etc because statistical methods are used in every economic related areas.

1. **Statistics in planning:** Modern age is the age of planning every objective plan depends upon the correct and sound statistical data. Planning is the pre-determined sets of program and policies, which is formulated in order to meet the targeted objectives,. To formulate the plan and details study of the existing situation is needed which is possible only thorough the statistical tools.
2. **Statistics in Economics:** Statistics is very essential to develop and prove the principles and laws of economics. It has great importance to understand the economics problems like production, consumption, distribution etc. as they can be solved by using statistical data.

3. **Statistics in business:** For the smooth operation of the business, statistical information is very useful. It simplifies the complex situation of business. It helps to study about the situation of market demand, supply, price etc. Without a very careful study of market it is difficult to success in business. Therefore the statistics is very essential in business sector also.

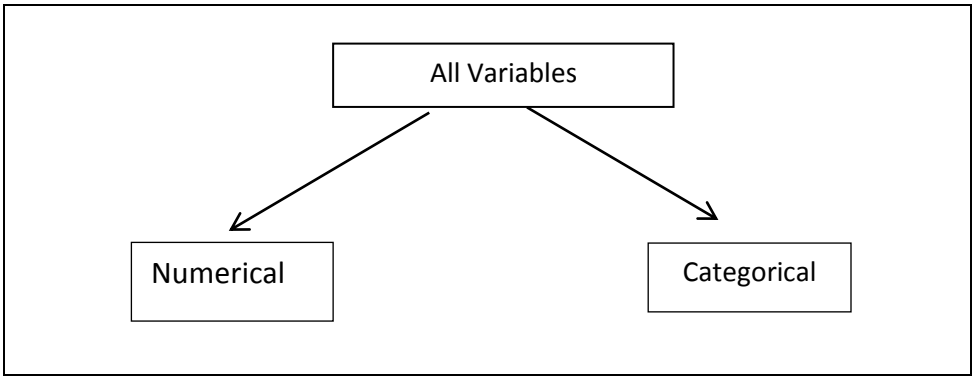
1.2 Data Basics

The backbone of a statistical investigation is data. Data is nothing but information. Data is a collection of facts, such as values or measurements. Data consists of variables, which is core of research analysis. Variables can be numbers, words, measurements, observations or even just descriptions of things. For the understanding the data we need to know well the variables.

Types of Variables

There are generally two types of variables:

- **Numerical Variables(Quantitative Variables)**
- **Categorical Variable(Qualitative Variables)**



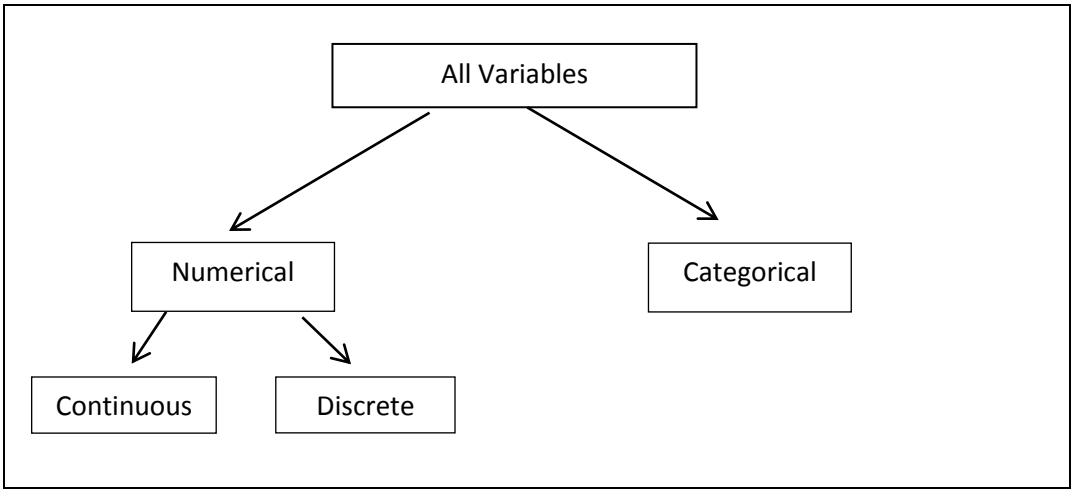
Numerical Variables (Quantitative Variables)*:*

Numerical variables take on numerical values. Numerical variables measure the actual magnitude of some characteristic for each of the individuals or units under consideration. This

type of variables is also called quantitative variables. It is sensible to add, subtract, take averages, etc. with the numerical values. For example height of a person, price of a commodity, number of students in a class, etc. Data collected on numerical variables is sometimes called numerical data.

And **Numerical Variables** can also be Discrete or Continuous:

- **Discrete** numerical variables are counted, and can take on only whole non-negative numbers can only take certain values (like whole numbers). As for example number of students, number of chairs in a room, number of children in a family, number of days in a month, etc.
- **Continuous** numerical variables are measured, and can take on any **numerical value** (within a range). As for example age, weight, height, temperature, income, sales, etc.



Categorical Variables (Qualitative Variables)*:*

Categorical value is just sort of descriptive. They are simply names. Categorical variables take on a limited number of distinct categories. It is concerned only with the presence or absence of some characteristics in a set of objects or individuals. This type of data is called qualitative or

enumeration data and the characteristic is used to classify an individual into different categories is called an attribute. For example, gender, colour (red or green, passed or failed, religion, rich or poor.

Categorical variables can be represented as numbers as in the case of a telephonic country code. It's a number but you can't compare those numbers.

1.3 Scales of Measurement

The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses.There are four measurement scales (or types of data): nominal, ordinal, interval and ratio. These are simply ways to categorize different types of variables.Data collection requires one of the following scales of measurement: nominal, ordinal, interval, or ratio.

Scales are important because they determine the techniques that we will use to analyze the data.

Nominal Scale:

Nominal scale is classified by quality (attribute) rather than numerical scale. The levels of the variable do not have ordering. A good way to remember all of this is that "nominal" sounds a lot like "name" and nominal scales are kind of like "names" or labels.The labels or names used to identify an attribute of the element. Examples of Nominal Scales are:

- Gender: Male, Female
- Marital status: Married, unmarried, divorced
- Eye color: brown, green, or blue

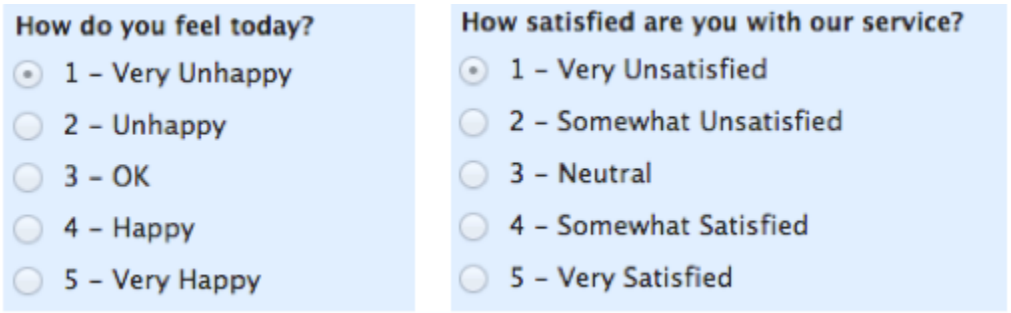
We can only summarize the nominal scale data by frequency table and cannot compare them.

Ordinal Scale:

For the ordinal scale, variables have relative differences and consist of ordering or ranking the differences. Thus the comparisons of the values or the levels variables are. Every ordinal variable is already nominal. Categorical variables that have the order or rank or have a rating scale of values are meaningful are called ordinal. One can count and order, but not measure, ordinal data. Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc.

- Grades
- Shoe size 4, 5, 6, 7,etc.
- Rich or poor,
- Class, etc.

Data recorded as excellent indicate the best service, followed by good and then poor; and we can assign 3 for best 2 for good and 1 for poor service.The categories for an ordinal set of data have a natural order, for example, suppose a group of people were asked to taste varieties of biscuit and classify each biscuit on a rating scale of 1 to 5, representing strongly dislike, dislike, neutral, like, strongly like. A rating of 5 indicates more enjoyment than a rating of 4, for example, so such data are ordinal.Thus, the scale of measurementis ordinal.The best way to determine *central tendency* on a set of ordinal data is to use the mode or median; the mean cannot be defined from an ordinal set.



Example of Ordinal Scales

Interval Scale:

The next higher scale is interval scale which have equal intervals between units. Consider the time series for the GDP of a country where we have one column for the year and another for a figure in millions of dollars reflecting the GDP of that country. The year variable is an interval variable given that there is a constant difference of one year between any two consecutive values.

For an interval scale, arbitrary zero point; i.e. negative numbers are allowed as for example degrees Fahrenheit; sea level, etc.

It has all the properties of ordinal data but cores on an interval scale can be added and subtracted but cannot be meaningfully multiplied or divided. Interval scales: they don't have a "true zero." For example, there is no such thing as "no temperature." Without a true zero, it is impossible to compute ratios. With interval data, we can add and subtract, but cannot multiply or divide. Let us consider this: 10 degrees + 10 degrees = 20 degrees. 20 degrees is not twice as hot as 10 degrees, however, because there is no such thing as "no temperature" when it comes to the Celsius scale. For Interval scale data are always numeric. Like the others, you can remember the key points of an "interval scale" pretty easily. "Interval" itself means "space in between," which is the important thing to remember–interval scales not only tell us about order, but also about the value between each item. For example, the time interval between the starts of years 1981 and 1982 is the same as that between 1983 and 1984, namely 365 days. Other examples of interval scales include the heights of tides, Scholastic Aptitude Test (SAT) and the measurement of longitude.

Example of Interval Scale

Interval scales are nice because the realm of statistical analysis on these data sets opens up. For example,*central tendency* can be measured by mode, median, or mean; standard deviation can also be calculated

Ratio Scale:

For the ratio scale the values of the variable have all the properties of interval scale with absolute zero (no negative numbers); i.e. zero indicates complete absence of the characteristi. Variables such as distance, height, weight, and age use the ratio scale of measurement. This scale requires that a zero value be included to indicate that nothing exists for the variable at the zero point. For example, consider the cost of a car. A zero value for the cost would indicate that the car has no cost and is free. In addition, if we compare the cost of $15000 for one car to the cost of $10,000 for a second car, the ratio shows that the first car is $15,000/$10,000 =1.5 times, the cost of the second car.Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

Summary

In summary, **nominal** variables are used to "*name*," or label a series of values. **Ordinal** scales provide good information about the *order* of choices, such as in a customer satisfaction survey. **Interval** scales give us the order of values + the ability to quantify *the difference between each one*. Finally, **Ratio** scales give us the ultimate–order, interval values, plus the *ability to calculate ratios* since a "true zero" can be defined.

Provides:	Nominal	Ordinal	Interval	Ratio
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode, Median		✔	✔	✔
The "order" of values is known		✔	✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Summary of data types and scale measures

Data matrix

Data matrix is a common way to organize data. Each row of a data matrix corresponds to a unique case, and each column corresponds to a variable. A case is a unit of observation or an observational unit of a variable. Data matrices are a convenient way to record and store data. If another individual or case is added to the data set, an additional row can be easily added. Similarly, another column can be added for a new variable.

Table 1.3 displays rows 1, 2, 3, and 50 of a data set concerning 50 students information a example of data matrix. Each row in the table represents a single student or case. The columns

represent characteristics, called variables, for each of the students. For example, the first row represents student 1, who is male, weights 63 kg whose major is BBA.

Serial number	Gender	Weight(kg)	Major
1	M	63	BBA
2	F	55	CSC
3	M	57	EEE
.			
.			
50	F	60	Economics

Table 1: Data matrix

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and the units of measurement.

Cross-sectional data

Cross-sectional data or a cross section study is a type of one-dimensional data set. Cross-sectional data refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or without regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects. A simple example of cross-sectional data is the gross annual income for each of 1000 randomly chosen households in Muscat for the year 2012. For example, we want to measure current obesity levels in a population. We could draw a sample of 1,000 people randomly from that population (also known as a cross section of that population), measure their weight and height, and calculate what percentage of that sample is categorized as obese. For example, 30% of our sample was categorized as obese. This cross-sectional sample provides us with a snapshot of that population, at that one point in time. Note that we do not know based on one cross-sectional sample if obesity is increasing or decreasing; we can only describe the current proportion.

It is important to note that **Cross-sectional data** are data collected at the same or approximately the same point in time. **Time series data** are data collected over several time periods.

Time Series Data

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the level of unemployment each month of the year would comprise a time series. This is because employment and unemployment are well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series. A times series allows you to identify change within a population over time.

Example GDP of a country for several years is as follows:

Year	GDP(million$)
2007	300
2008	320
2009	350
2010	455
2011	530

Data or Datum?

The singular form is "datum", so we would say "that datum **is** very high". "Data" is the plural so we can say "the data **are** available", but it is also a **collection** of facts, so "the data **is** available" is correct too.

Random Variable

Usually we are interested in data that comes from different sources and recording as quantitative or qualitative according to the type of data. It may be possible to know in advance what the possible values of the data could be, but until we collect the data we do not know the actual values. For example, we know that if we roll a dice we will get a number between 1 and 6, but we do not know which one until we roll it. We might know that the speed of a moving car at a particular location, but we do not know the exact speed until we measure it. For the case of quantitative variable, it might be guessed the taste of a cup of coffee, but it is not possible to say the actual taste until it is tested. When we want to measure a quantity or quality and it can take on a range of values like this, before we observe the actual value we refer to it as a **random variable.**

1.4 Data Sources

There are many sources to collect dat. Data can be collected

Directly **(Primary data)** or

Indirectly **(Secondary data)**

Primary data

Primary data means original data that has been **collected** specially for a specific study.the purpose in mind. It means someone collected the data from the original source first hand or directly.

Methods of Primary data collection

The people who gather primary data may be an authorized organization, investigator, enumerator or they may be just some individual with a clipboard. Generally the primary data are collected by an agency or organization. Those who gather primary data should have knowledge of the study and may be motivated to make the study a success. Primary data is only considered as reliable as the people who gathered it are acting as a witness.

Methods of direct data collection include:

- Surveys administered through the use of an interviewer
- Surveys which are self-enumerated (the information written or entered directly by the respondent)
- In depth interviews or focus groups to provide the opportunity for discussion and elaboration for collecting more detailed information about a particular issue or issues
- Observational studies in which data are gathered through the direct observation of the population or sample
- Experiments and clinical trials that involve controlled studies where researchers collect data from subset groups taken from the population of interest.
- Social media is a good source of collecting data as a primary source.

Secondary data

Indirect methods of data collection involve sourcing and accessing existing data that were not originally collected for the purpose of the study. This type of data is known as **secondary data.** Secondary data is data that is being reused.

Methods of Secondary data collection:

Secondary data are usually procured from already published or unpublished documents rather than undertaking first hand field investigation.

Secondary data can be collected from the followings:

The main source of secondary data is administrative data. **Administrative data are collected as part of the day to day processes and record keeping of organizations.** Administrative data, such as historical data or public records, include: school enrolments; hospital admissions; records of births, deaths, and marriages. The data are not collected initially for statistical purposes but can be organized to produce statistics. Administrative data can be useful because this data are usually recorded about every unit of the population of interest and are

continuously collected, allowing comparisons to be made over time. Where administrative data are available, this may eliminate the need to conduct a survey provided the data are fit for the statistical purpose.

All companies maintain a variety of databases about their employees, customers, and business operations. Data on employee salaries, ages, and years of experience can be obtained from internal personal records.

Examples: registry, internet, magazine, books, TV, newspaper, Radio, Social media

Internet

In recent years, the internet has become an important source of data. This is indirect method of data collection. All most all companies have internet web sites and provide public access to them.

1.5 POPULATION AND SAMPLE

A population that is statistical population is defined as the total set of individuals, groups, objects, or events that the researcher is studying. For example, if we were studying employment patterns of recent college graduates of a country, our population would likely be defined as every college student who graduated within the past one year from any college across the country. Again if we are interested for the age of students of newly admitted students of BRAC University, then ages of all newly admitted the students of BRAC University will be our population.

A sample is a relatively small subset of people, objects, groups, or events that is selected from the population. Instead of surveying every recent college graduate in a country, which would cost a great deal of time and money, we could instead select a sample of recent graduates (say 400 students) which would then be used to generalize the findings to the larger population.

1.6 Descriptive statistics

Data might be too detailed for us to get through all of the individual elements. The tools and techniques of **Descriptive Statistics** help us summarise volumes of data, and derive actionable information.

We can summarize the data in two ways

- **Numerically and**
- **Visually**

A descriptive statistics is a numerical summary of a dataset. Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Such summaries of data may be tabular, graphical, or numerical form. Most of the statistical information in newspapers, magazines, company reports, and other publications consists of data that are summarized and presented in a form that is easy for the reader to understand which descriptive statistics is. are referred to as **descriptive statistics**.

Descriptive statistic reports generally include summary data tables (kind of like the age table above), graphics (like the charts above), and text to explain what the charts and tables are showing. With descriptive statistics you are simply describing what is or what the data shows or describe what's going on in our data.Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data.

In a nutshell, descriptive statistics intend to *describe* a big hunk of data with summary charts and tables, but do not attempt to draw conclusions about the population from which the sample was taken

Descriptive statistics therefore allow us to present the data in a more meaningful way which allows simpler interpretation of the data. For example, if we had the results of 100 pieces of students' coursework, we may be interested in the overall performance of those students. We would also be interested in the distribution or spread of the marks. Descriptive statistics allow us to do this. Descriptive statistics usually involve measures of central tendency (mean, median, and mode) and measures of spread or dispersion (variance, standard deviation, etc.).

- Measures of central tendency: these are ways of describing the central position of a frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of marks scored by the 100 students from the lowest to the highest. We can describe this central position using a number of statistics, including the mode, median, and mean.

- Measures of spread: these are ways of summarizing a group of data by describing how spread out the scores is. For example, the mean score of our 100 students may be 65 out of 100. However, not all students will have scored 65 marks. Rather, their scores will be spread out. Some will be lower and others higher. Measures of spread help us to summarize how spreads out these scores are. To describe this spread, a number of statistics are available to us, including the range, quartiles, absolute deviation, variance and standard deviation.

When we use descriptive statistics it is useful to summarize our group of data using a combination of tabulated description (i.e. tables), graphical description (i.e. graphs and charts) and statistical commentary (i.e. a discussion of the results

1.7 Inferential statistics

Let’s say you wanted to know the favourite ice cream flavours of everyone in the country. Say, there are about 50 million people in the country, and it would be impossible to ask every single person about their ice cream preferences. Instead, we try to sample (say 500 people) a

representative population of people and then extrapolate our sample results to the entire population. This is the idea behind inferential statistics.

The heart of statistics is inferential statistics. Descriptive statistics are typically straightforward and easy to interpret. Unlike descriptive statistics, inferential statistics are often complex and may have several different interpretations.

The goal of inferential statistics is to discover some property or general pattern about a large group by studying a smaller group of people in the hopes that the results will generalize to the larger group.

This is a set of methods used to make a generalization, estimate, prediction or decision and hypothesis testing. Inferential statistics aims to draw conclusions about the population from the sample.

With inferential statistics, we are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Thus, we use inferential statistics to make inferences from our data to more general conditions

Example: if took a random sample of 500 fresh graduates from country and found that 80% are employed. We can say sat that 80% fresh graduates of a country are employed although we have no information about *all* the fresh graduates of the country. We have taken that information and *generalized* it to conclude about all fresh graduates.

1.8 Limitations of Statistics

Although statistics is indispensable to almost all areas - social, physical and natural, it has certain limitations. Some important limitations of statistics are the following:

Statistics can be misused:

As W.I. King points out, “One of the short-comings of statistics is that do not bear on their face the label of their quality.” So we can say that we can check the data and procedures of its approaching to conclusions. But these data may have been collected by inexperienced persons or they may have been dishonest or biased. As it is a delicate science and can be easily misused by an unscrupulous person. So data must be used with a caution. Otherwise results may prove to be disastrous. Statistical methods rightly used are beneficial but if misused these become harmful. Statistical methods used by less expert hands will lead to inaccurate results. Here the fault does not lie with the subject of statistics but with the person who makes wrong use of it.

Statistics is only means:

Statistics is only the means, which provide a method of studying problem. But is should not be considered as the best because this method should be supplement by other techniques to derive conclusion

Statistical results are not always beyond doubt:

“Statistics deals only with measurable aspects of things and therefore, can seldom give the complete solution to problem. They provide a basis for judgment but not the whole judgment.” —Prof. L.R. Connor

Although we use many laws and formulae in statistics but still the results achieved are not final and conclusive. As they are unable to give complete solution to a problem, the result must be taken and used with much wisdom.

Too Many methods to study problems:

In this subject we use so many methods to find a single result. Variation can be found by quartile deviation, mean deviation or standard deviations and results vary in each case.

Qualitative Aspect Ignored:

The statistical methods don’t study the nature of phenomenon which cannot be expressed in quantitative terms.

Such phenomena cannot be a part of the study of statistics. These include health, riches, intelligence etc. It needs conversion of qualitative data into quantitative data.

So experiments are being undertaken to measure the reactions of a man through data. Now a days statistics is used in all the aspects of the life as well as universal activities.

Statistics does not study individuals:

Statistics deals with only aggregates of facts or items and it does not recognize any individual item. Thus, individual terms as death of 6 persons in a accident, 85% results of a class of a school in a particular year, will not amount to statistics as they are not placed in a group of similar items. It does not deal with the individual items, however, important they may be.

Statistical laws are not exact:

Statistical laws are not exact as in case of natural sciences. These laws are true only on average. They hold well under certain conditions or assumptions. They cannot be universally applied. So statistics has less practical utility

Exercise

1. The Commerce Department reported receiving the following applications for National Business Quality Award: 23 from large manufacturing firms, 18 from large service firms, and 30 from small businesses.

    a. Is type of business a categorical or quantitative variable?

    b. What percentage of the applications came from small businesses?

2. State whether each of the following variables is categorical or quantitative and indicate its measurement scale.

a. Annual sales

b. Soft drink size (small, medium, large)

c. Employee classification (GS1 through GS18)

d. Earnings per share

e. Method of payment (cash, check, credit card)

3. Write five examples of numerical variables and five examples of categorical variables.

4. Compare primary data and secondary data with examples.

5. What is descriptive and inferential statistics?

6. Define different scales of measurement with examples.

7. Do traffic control measures work?(Using Statistical Cycle)

Traffic control measures are often introduced in order to reduce the frequency of road accidents. In order to determine whether or not they are effective, you might design a statistical experiment. Place the following steps that you would take into the correct order.

a) Collect data on the number of road accidents in the area before and after the traffic control measures are introduced.

b) Perform a statistical test to analyze whether or not the number or road accidents has changed following the traffic control measures being introduced.

c) Pose the question "Has the number of road accidents decreased following the traffic control measures being introduced?"

d) Conclude whether or not the number of road traffic accidents has decreased following the road traffic control measures being introduced.

Answer as a sequence of letters with no spaces in between, for example "abcd."[Ans: cabd]

8. Which of the following steps might you take to test whether or not a new drug is more effective than an old one?

a. Divide the volunteers into two groups, one for males and one for females, give the old drug to one group and the new drug to the other group

b. Collect information about the recovery of people currently taking the two drugs.

c. Perform a statistical test to determine whether or not the recovery rate was higher amongst those taking the new drug or those taking the old drug

d. Pose the question of whether or not the new drug is more effective than the old drug in terms of recovery rate.

e. Collect data on the recovery rates of people currently taking a high dose compared to a low dose of the new drug

f. Use the test results to determine whether or not the new drug is more effective than the old.

g. Conclude that the drug with the higher observed recovery rate is more effective

[Ans: first:d, second:b, third:c, fourth:f]

9. **Designing a weather study (case study)**

First, we need to design a study that will allow us to answer our question. We begin by defining the specific question that we are interested in, before looking at what data we might collect, and how we might record it.

a. We are interested in whether or not there has been an overall change in temperature over the past three years, and if so what this change has been. Which of the following might be a good question to pose?

i. Was the temperature in August last year lower or higher that the temperature in August three years ago?

ii. How has the monthly average daily maximum temperature changed over the past three years? correct

iii. How have the hottest and coldest recorded temperatures in a year changed over the past three years?

b. This question is suitable as it is specific enough to allow us to design a study to answer it, but will also provide enough information that we can draw a conclusion from it. Which of the following is a good way to collect data to answer the question?

i. Obtain hourly temperature recordings from the past three years.

ii. Ask friends and family for their recollections of the temperatures over the past three years.

iii. Obtain a list of the monthly average maximum temperatures over the past three years.

c. This would give us enough data to answer our question, but not so much that our data set becomes too large to work with. How should we record the data?

i. As a number, in degrees Celsius, rounded to five decimal places.

ii. As a number, in degrees Celsius, rounded to the nearest degree.

iii. On a scale from 1 to 10 where 1 is very cold and 10 is very hot.

Hints: For our experiment, it is useful to have quantitative data, and it is acceptable to round the nearest whole degree since the raw data used to calculate the average was most probably rounded to the nearest degree.