1

Statistics and Data

Learning Objectives

At the end of this chapter the students will be able to:

1. Describe the meaning and usage of statistics,
2. Identify the types of data,
3. Select issues and questions from various problems or situations, and
4. Select appropriate methods for study and data collection.

1.1 Introduction to Statistics

🟥 Meaning of Statistics and a Statistic

Statistics

Statistics (plural) is a branch of mathematics. It refers to the art and science of collecting, organizing, analyzing, interpreting and presenting numerical data for the purpose of making a more effective decision. Statistics deals with all aspects of data including the planning of data collection in terms of surveys and experiments.

A Statistic (singular) or *sample statistic* is a single measure of some attribute of a sample group. It is calculated by applying a function to the values of the items of the sample. The term statistic is used both for the function and for the value of the function on a given sample.

🟥 Statistics

Statistics represent a common method of presenting information. In general, statistics relate to numerical data, and can refer to the science of dealing with the numerical data itself. Above all, statistics aim to provide useful information by means of numbers.
Therefore, a good definition of statistics is

> "a type of information obtained through mathematical
> operations on numerical data".

Statistics is the study of numerical information, which is called data. People use statistics as tools to understand information. Statistics is concerned with the techniques by which information is collected, organized, analyzed, and interpreted. For example:

1. The average weight of students in your classroom;
2. The minimum number of rentals your household had to make to be in the top 5% of renters for the last month; and
3. The minimum and maximum temperature observed each day of the week.

4.

Statistics method is concerned with

1. **Design**
   - procedures for gathering data,

   HOW we get data.

2. **Description**- (i.e. descriptive statistics summarizing, reporting features, characterizing data, communicating information.

   HOW we describe it.

3. Inference- (i.e. inferential statistics) making valid generalizations or decisions concerning a population on the basis of samples.

   WHAT we do with it.

Population and Sample

Population and sample are two basic concepts of statistics. A population is the set of measurements corresponding to the entire collection of units for which inferences are to be made. Population can be described as the set of individual persons or objects in which an investigator is primarily interested during his or her research problem. Sometimes the researchers wanted measurements for all individuals in the population are obtained, but often only a set of individuals of that population are observed. Sample is the part of the population from which information is collected.

Population

Population is the collection of all individuals or items under consideration in statistical study. The population always represents the target of an investigation.

Sample

Sample is the part of the population from which information is collected. A sample from statistical population is the set of measurements that are actually collected in the course of an investigation.

Types of Statistics

Statistics is the branch of mathematics. Statistics is the study of how to collect, organize, analyze, and interpret numerical information from data. Statistics is also the mathematical study of probability of events occurring based on known quantitative data or a collection of data. There are two types of statistics: Descriptive statistics and Inferential statistics.

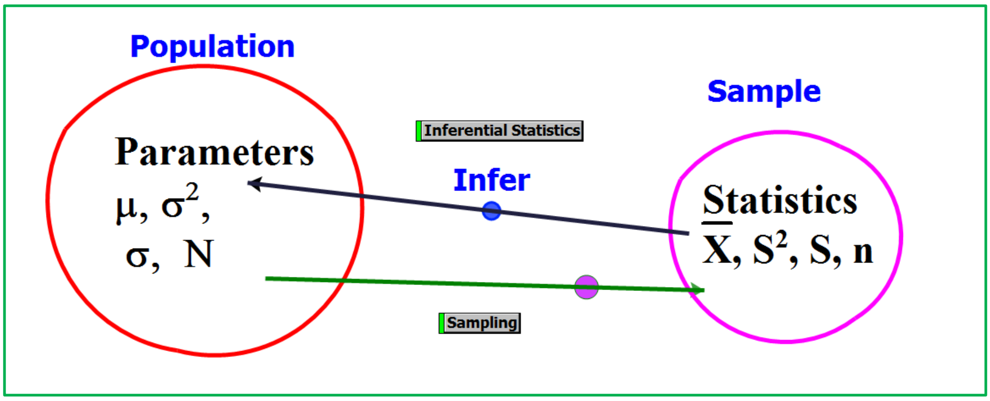
- *Descriptive statistics*
- *Inferential statistics*

Descriptive Statistics

Descriptive statistics involves methods of organizing, picturing and summarizing information from data. The procedures used to organize and summarize masses of numerical data.

Inferential Statistics

Inferential Statistics consist of methods for drawing and measuring the reliability of conclusions about population based on information obtained from a sample of the population. A conclusion drawn about a population based on the information in a sample from the population is called a **Statistical Inference**.



Parameters and A Statistic

Parameters

A parameter is an unknown numerical summary of the population.

A Statistic

A statistic is a known numerical summary of the sample which can be used to make inference about parameters.

Remember: **Parameter** is to **Population** as **Statistic** is to **Sample**

The inference about some specific unknown parameter is based on a statistic.
The example of symbols of parameters and statistic are as in the following table.

	Topic	Symbol	
		Population (Parameter)	**Sample (Statistic)**
1	Mean	μ	$\bar{x}$
2	Variance	σ^2	S^2 or SD^2
3	Standard Deviation	σ	S or SD
4	Size	N	n

1.2 The Use of Statistics

We can see statistics in almost every medium: newspapers, television, poster, and the internet. It is important for people to have knowledge of the rules and methods in using the statistics. For example: economic predictions, sales forecasts stock market, marketing information, political polls, customer survey, etc.

Many activities today depend on statistics. Some of these are: sports, stock market, advertising, industry, Government survey, market research, environment, economics, consumer reports, etc.

Why is statistics needed? Because statistics is a way to get information from data. Knowledge is what we know. Information is the communication of knowledge. Data are known to be crude information and not knowledge by themselves. The sequence from data to knowledge is as follows:

- From *data* to *information*, data become information when they become relevant to the decision problem;
- From *information* to *facts*, information become facts when the data can support it, and finally
- From *fact* to *knowledge*, fact become knowledge when they are used in the successful completion of the decision process.

🟥 The Use of Statistics

Statistics are used in the fields of business, math, economics, accounting, banking, government, astronomy, and the natural and social sciences. The applications of statistics in business are fundamentally concerned with decision making.

The examples of using statistics are as follows:

1. **Business Statistics.**

   Statistics analysis allows businesses to deal with the uncertainties of the business. The managers can use statistics to plan and make decisions based on the information. Whether designing new products, streamlining a production process or evaluating current or new coming customers. Marketing is an important part of any business and statistics helps to market products and services effectively.

2. **Planning around the weather and forecasting**

   People have to use statistics and probability to plan and forecast around the weather. Meteorologists can't predict exactly what the weather will be, so they use tools and instruments to determine the likelihood that it will rain, cold or snow. For example, if there is a 70 percent chance of rain, then the old weather conditions are such that 70 out of 100 days with similar conditions, it has rained. You may decide to wear shoes rather than sandals or take an umbrella to work.

3. **Insurance Options**

   Probability plays an important role in analyzing insurance policies to determine which plans are best for you or your family and what deductible amounts you need. For example, when choosing a car insurance policy, you use probability to determine how likely it is that you'll need to file a claim.

4. **Researchers**

   Researchers and professionals use statistics for a variety of information-gathering needs. Researchers collect data from the sample and decide whether to use descriptive or inferential statistics, depending upon the research question and methodology of the study.

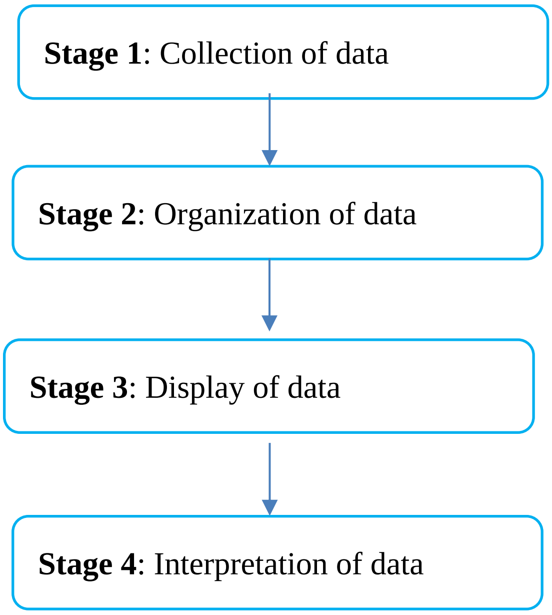
   Descriptive statistics tell researchers about the raw scores that are present in the data, such as the percentage of people that agree with a public policy or an average test score.

   Inferential statistics make predictions about a data set. It infers the relationship between two or more variables and predicts the outcome of the impact of one variable upon another variable.

   For example, a researcher might look for the average percentage of people who buy a particular medication, then compare it to people who reported side effects after taking that medication. This would be used to determine the likelihood of suffering harmful side effects.

1.3 Statistics and Decision Making

Statistical study consists of four stages as following:



Stage 1: Collection of data

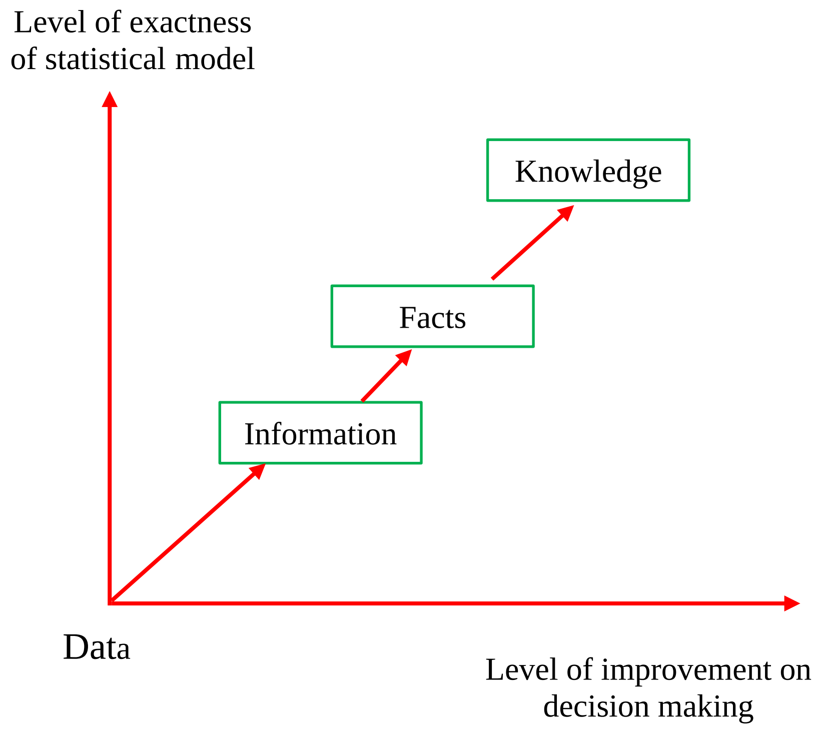
Stage 2: Organization of data

Stage 3: Display of data

Stage 4: Interpretation of data

Data is known to be crude information and not knowledge by itself. The sequence from data to knowledge is: **from Data to Information, from Information to Facts, and finally, from Facts to Knowledge**. Data becomes information, when it becomes relevant to your decision problem. Information becomes fact, when the data can support it. Facts are what the data reveals. However, the decisive instrumental knowledge is expressed together with some statistical degree of confidence.

Fact becomes knowledge, when it is used in the successful completion of a decision process. Once we have a massive amount of facts integrated as knowledge, then our mind will be superhuman in the same sense that mankind with writing is superhuman compared to mankind before writing. The following figure illustrates the statistical thinking process based on data in constructing statistical models for decision making under uncertainties.



Descriptive statistics provide simple summaries about the sample and about the observations that have been made. The summaries may be either quantitative such as summary statistics or visual simple to understand graphs. People use descriptive statistics to repurpose hard-to-understand quantitative insights across a large data set into a simple description.

For examples:

1) The student's *grade point average* (GPA). This single number describes the general performance of a student across the range of their course studied. The student's grade point average provides a good understanding of descriptive statistics. The idea of a GPA is that it takes data points from a wide range of exams, classes and grades, and averages them together to provide a general understanding of a student's overall academic abilities. A student's personal GPA reflects his mean academic performance. The GPA will a guide line for the student to make decision on the improvement of his/her study in the future.

2) In the business world, descriptive statistics provides a useful summary of many types of data. For example, the investors may use a historical account of return behavior by performing empirical and analytical analyses on their investments in order to make better investing decisions in the future.

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, patterns might emerge from the data. Descriptive statistics includes the construction of graphs, charts and tables, and the calculation of various descriptive measures such as averages, measures of variation. Inferential statistics includes methods like point estimation, interval estimation and hypothesis testing which are all based on probability theory.

Descriptive statistics do not allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypotheses we might have made.

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

1.4 Statistics and Data Collection

Data collection is the process of gathering and measuring information on targeted variables. Data is a collection of information, characteristics or facts. Data can be a set of values of quantitative variables or qualitative.

Qualitative data is descriptive information, it describes something. Whereas quantitative data is numerical information. The Latin word "**data**" is the plural of "**datum**". No matter what kind of data, we have to know their context which can be described by answer the question of Five W's.

The Five W's are ***Who, What, When, Where*** and ***Why or How***.

Answering the questions of Five W's can provide the context for data values. The answers to the first two questions are essential because we will know the useful information. We must know at least the *Who*, *What* and *Why* to be able to say anything useful based on the data. The *who* are the cases. The *What* are the variables. A variable gives information about each of the cases. The *Why* helps us decide which way to treat the variables. We treat variables in two basic ways as: categorical or quantitative.

1 Types of Scales Measurement

The objective of statistics is to extract information from data (datum). Data can be classified into four scales, namely:

1) Nominal scale
2) Ordinal scale
3) Interval scale
4) Ratio scale.

Nominal Scales

Nominal scales have no order and thus only gives **names** or labels to various categories. The word nominal is derived from **nomen**, the Latin word for *name*. Nominal scale can be placed into categories. Nominal scale do not have a numeric value and so cannot be added, subtracted, divided or multiplied.

Nominal scales are categories.

Nominal scale ⇒ qualitative or categorical.

The nominal scales are recorded by arbitrarily assigning a number to each category. For example:

Gender:	Male	recorded as	1
	Female	recorded as	2
Status:	Single	recorded as	1
	Married	recorded as	2
	Divorced	recorded as	3
	Widowed	recorded as	4

We can not calculate mean or average of Nominal scale.

Ordinal Scales

Ordinal scales have **order**, but the interval between measurements is not meaningful. *Ordinal* is easy to remember because its sounds like *order* and that is the key to remember with "*ordinal scales*"–it is the *order* that matters. The ordinal scale contains things that you can place in order. For example, from low to high, hottest to coldest, lightest to heaviest, richest to poorest. Basically, if you can rank data by 1st, 2nd, 3rd place (and so on), then you have data that's on an ordinal scale.

- Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort;
- The best way to determine *central tendency* on a set of ordinal scale is to use the mode or median; the *mean* cannot be defined from an ordinal set.

Ordinal scales may be treated as nominal, but their values are ranked in order.

For example:

• Very unhappy	1
• Unhappy	2
• Ok	3
• Happy	4
• Very happy	5

• Poor	1
• Fair	2
• Good	3
• Very good	4
• Excellent	5

A second example of the ordinal scales, you might conduct a survey and ask people to rate their level of satisfaction with the choice of the following responses:

Extremely satisfied	5
Satisfied	4
Neutral	3
Dissatisfied	2
Extremely dissatisfied	1

■ Interval Scales

Interval scales have meaningful intervals between measurements.

- Interval scales are real numbers, such as height, weight, incomes, score, distance, …

  Interval scales ⇒ **quantitative data or numerical**.
- Interval scales can be used for statistical analysis on the data sets. For example, *central tendency* can be measured by mode, median, or mean; standard deviation can also be calculated.

■ Ratio Scales

Ratio scales have the highest level of measurement. Because ratio scale tell us about the order, they tell us the exact value between units, **and** they also have an absolute zero–which allows for a wide range of both descriptive and inferential statistics to be applied. Ratio scale can be used to identify the Central tendency: mean, mode, or median. The measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

Provides:	Nominal Scale	Ordinal Scale	Interval Scale	Ratio Scale
Quantify the difference between each value	-	-	✓	✓
The order of values is known.	-	✓	✓	✓
Can Counts and make Frequency of Distribution	✓	✓	✓	✓
Can add or subtract values	-	-	✓	✓
Can multiple and divide values	-	-	-	✓
Has true zero	-	-	-	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean	-	-	✓	✓

- **Nominal** scales 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 and the ability to quantify *the difference between each one*.
- **Ratio** scales give us the ultimate–order, interval values, and the *ability to calculate* ratios since a "true zero" can be defined.

2 Types of Data

There are two types of data: **quantitative data** and **qualitative data.**

1) **Quantitative data**: quantitative data is information about quantities.

   That is, information that can be measured, counted and written down with numbers. Quantitative variables record measurements or amounts of something and they must have *units*. For examples of quantitative data are height, area, salary, number of clients.

2) **Qualitative data:** qualitative data is information about qualities. The information of qualitative data can be observed but cannot be measured and expressed as a number. Such as gender, economic status, taste. Qualitative data will employ *categorical variables* to identify a category for each case. Usually, we think about the counts of cases that fall into each category. Categorical data can take on numerical values. Such as **1** indicating male and **2** indicating female.

Note

Quantitative Data	**Qualitative Data**
1. Deals with numbers;	1. Deal with description;
2. Data can be measured;	2. Data can be observed but cannot be measured;
3. Examples: height, areas, volume, salary, sales number, age, blood pressure, IQ, etc.	3. Colour, smell, taste, appearance, etc.
Quantitative ⇨ **Quantity**	**Qualitative** ⇨ **Quality**

3 Data Collection

In order to do data collection, we must know:
Why we are examining the data or ***What*** we want to know. The context shall tell ***Who*** was measured, ***What*** was measured, ***How*** the data were collected, and ***When*** and ***Why*** the study was performed.

🟥 Data Collection

Data collection refers to the process of gathering of set of information of variables or observations. There are two sources of data, **primary data** and **secondary data**.

1) **Primary data**: data is collected for the first time by the investigator himself/herself for a specific purpose. Data are from first hand sources such as:
   - Questionnaire.
   - Interview,
   - Census,
   - Sample Survey,
   - Direct observation, and
   - Focus group.
2) **Secondary data:** secondary data is the data already collected or produced by someone else for some other purpose. Data received from secondary sources such as printed material reports of government, bank, published material, website, the World Wide Web, www.com, Google.

Statistics very often involves the collection of data. There are many ways to obtain data, and collecting method are as follows.

1) *Questionnaire* The questionnaire consists of a series of written questions for a participant to fill in the answers with their thoughts. It can be multiple-choice questions or open ended questions. The questionnaire cannot be too long or too involved. Using questionnaires allows a researcher to utilize several strengths and also weakness.
2) *Interviews.* The interview can be structured or unstructured interview questions. It can be face-to-face interview or telephone interview.
3) *Observation.* The observation may be carried out by trained observers, cameras, or closed circuit television. Observation can be used in widely different fields. Observation may also be used in *before* and *after* studies.

The advantages and disadvantages of data collecting method are discussed below.

🟥 Advantage and Disadvantage of using Primary Data

Advantages of using Primary Data:

1. The investigators will collect data specific to their problems.
2. Data collected are original, unique data, more reliable and accurate because data are first-hand information from a person who participated in an event.
3. If required, it may be possible to obtain additional data during the study period. The investigators can make decision on why, what, how and when to collect. It is directly collected by the investigators.

Disadvantages of using Primary Data:

1. Primary data collection process is expensive.
2. Primary data collection process must be conducted by trained professionals and time consuming process.
3. There is a possibility of personal prejudice or bias creeping while collecting primary data because of the direct involvement of an investigator.
4. If the sample unit of data collection is not large enough, the information can be misleading and miss some useful information.

🟥 Advantage and Disadvantage of using Secondary Data

Advantage of using Secondary Data

1. Secondary data collection process is inexpensive and quickly available.
2. Secondary data are already collected by others and made available to the investigators.
3. Secondary data provide quantitative data in numerical. The data are analyzed using statistical methods and data presentation like in tabulated form, graphs, or descriptive statistics.
4. Secondary data can be used as supporting data, if the investigators used with cautiousness.

Disadvantage of using Secondary Data

1. Secondary data are *less reliable* than primary and not directly related. This is because the data were collected by others and not by the investigators.
2. Secondary data may be *less accurate*. The verification of published information cannot be always confirmed accurately as all references used may not be available or may be out of dated information.
3. Collection of secondary data may or may not fulfill the actual requirement of the investigators.

Comparison of Primary Data and Secondary Data

Topics	Primary Data	Secondary Data
Meaning	Primary data refers to the first hand data collected by the investigator himself/herself for a specific purpose	Secondary data means data collected by someone else earlier in the past.
Time	Real time data, but it is time consuming in collecting data.	Past data and it takes short time in getting data.
Cost	Expensive	Economical
Specific	Always specific to the investigator's needs.	May or may not be specific to the investigator's needs.
Sources	Observations, interviews, experiments, questionnaire	Government publications, website, reports, books.

Note

There are many differences between primary and secondary data. The most important difference is:

- Primary data is first-hand, original, and raw data.
- Secondary data is grouped data, it is the analysis and interpretation of the primary data.

Exercise

Identify Nominal, Ordinal, Interval or Ratio scale of the following Topics.

No.	Topics	Scale
1	Rank in tennis players.	
2	Record of number of errors made in a certain time period.	
3	Symptoms of a disease - mild, moderate, severe.	
4	Socioeconomic status (low, middle, high).	
5	University Ranking.	
6	The score of IQ test.	
7	Time taken for completing a task.	
8	Views about some political matter (Totally agree, mostly disagree, totally disagree).	
9	Thermometer readings on Celsius scale.	
10	Measurement of weight.	
11	Sex (gender) of the students in the sample.	
12	Rating of electricity provider on a 10-point scale.	
13	Number of goals scored by your favorite team.	
14	The nationality of a person.	
15	Number of liters of water contained in a tank.	