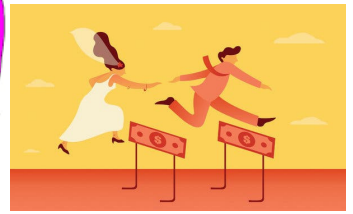
(G230 Course / Geology Dept.)

Introduction to Statistics

Statistics has been applied in daily lifetime including the numerical facts and figures. For instance:



- ✓ The largest earthquake measured 9.2 on the Richter scale.
- ✓ One in every Eight Iraqis has blood pressure .
- ✓ The average cost of a wedding is nearly 15,000,000 Iraqi Dinars.
- ✓ By year 2020, there will be 15 people aged 65 and over for every new baby born.
- ✓ 30% of the class students do not wear uniforms.

- The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted. The numbers may be right, but the interpretation may be wrong.

- Ice cream shop advertised in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective. A major flaw is that ice cream consumption generally increases in the months of June, July, and August regardless of advertisements. This effect is called a history effect and leads people to interpret outcomes as the result of one variable when another variable (Time/Summer Season) is actually responsible.

Case (1) There are millions of sedan cars in Iraq. What is their average value? It is impractical to sort all those cars to get the price and even can be done will be needed long time after going back to the archives of cars documents. So, it is better to select as example 200 cars randomly to estimate the price average which will be represent the whole sedan cars in Iraq.

- Millions of sedan cars in Iraq = Population.
- The number attached to each one, its value = Measurement
- The average value is = Parameter: a number that describes a characteristic of the population, in this case monetary worth.
- 200 Cars = Sample.
- The 200 numbers, the monetary values of the cars we selected = *Sample Data*.
- The average of the data = Statistic: a number calculated from the sample data (By calculation 200 different price values of cars to get 10,000USD supposedly. The average value for one car is 10,000, which represent all 200 cars values.
- In this way we have drawn an inference about the population based on information obtained from the sample. In general, statistics is a study of data: describing properties of the data, which is called descriptive statistics, and drawing conclusions about a population of interest from information extracted from a sample, which is called inferential statistics. Computing the single number 10,000USD to summarize the data was an operation of descriptive statistics; using it to make a statement about the population was an operation of inferential statistics.

- The measurement made on each element of a sample need not be numerical. In the case of cars, what is noted about each car could be its color, its make, its body type, and so on. Such data are categorical or qualitative, as opposed to numerical or quantitative data such as value or age.

- Qualitative data can generate numerical sample statistics. we might be interested in the proportion of all cars that are less than six years old. In our same sample of 200 cars we could note for each car whether it is less than six years old or not, which is a qualitative measurement. If 172 cars in the sample are less than six years old, which is 0.86 or 86%, then we would estimate the parameter of interest, the population proportion, to be about the same as the sample statistic, the sample proportion, that is, about 0.86. The relationship between a population of interest and a sample drawn from that population is perhaps the most important concept in statistics, since everything else rest on it.

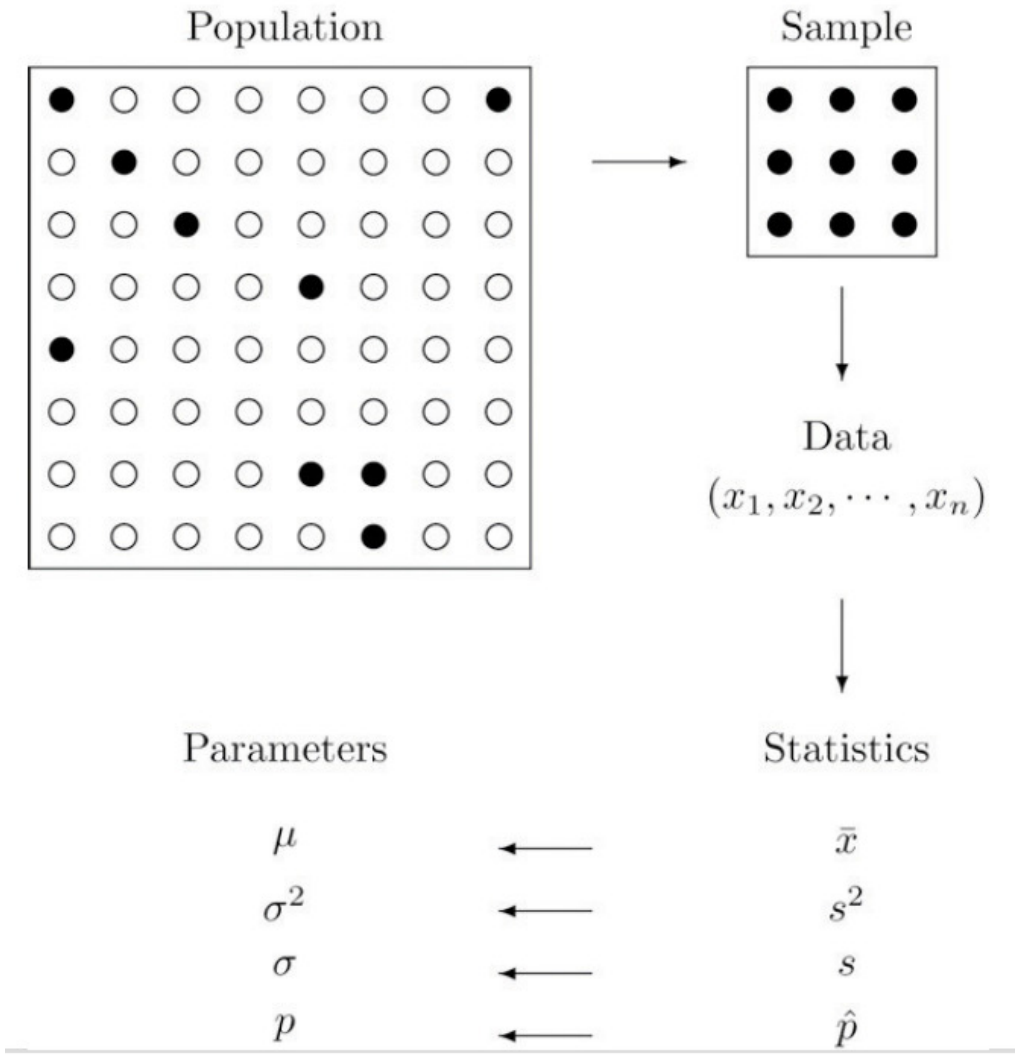
The relation can be illustrated by the following figure:

- The circles in the large box represent elements of the population.
- The solid black circles represent the elements of the population that are selected at random and that together form the sample (Let consider sample is 200 cars sample).
- For each element of the sample there is a measurement of interest, denoted by a lower case x (which we have indexed as x1 , … , xn to tell them apart); these measurements collectively form the sample data set.

- From these data we may calculate various statistics.
- To anticipate the notation that will be used later,

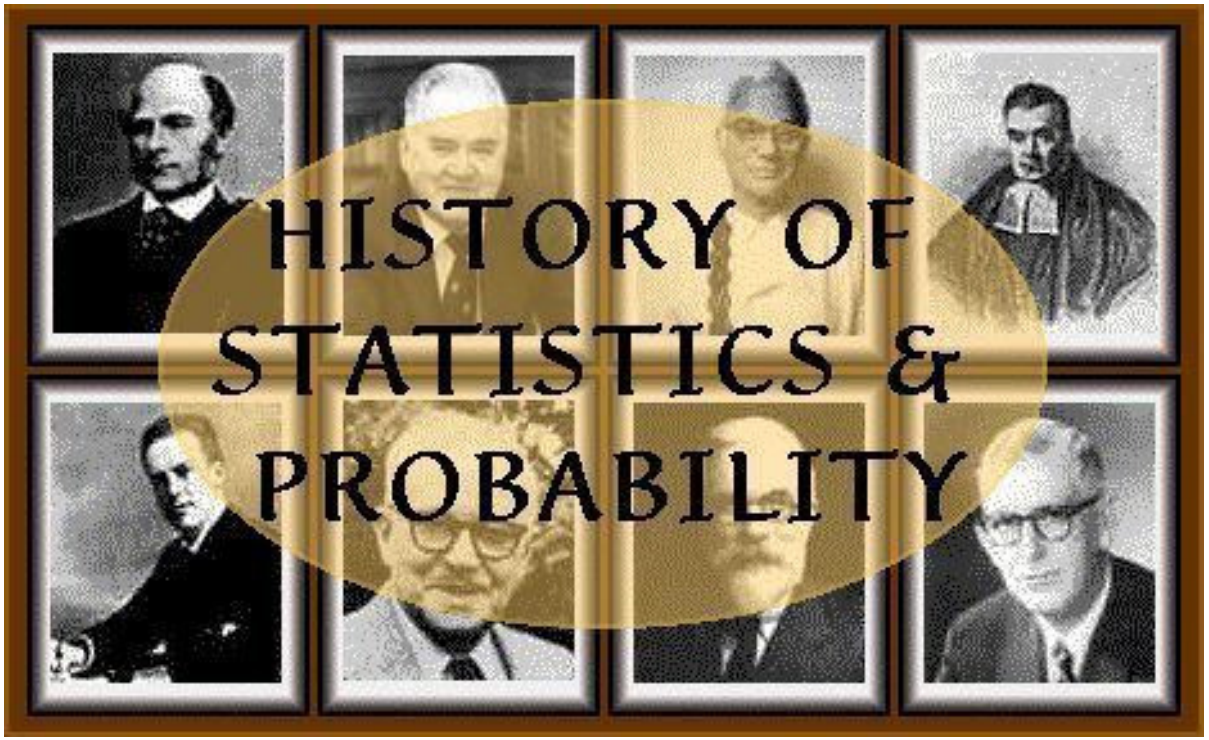
we might compute the sample mean $\bar{x}$ and the sample proportion $\hat{p}$, and take them as approximations to the population mean μ (this is the lower case Greek letter mu, the traditional symbol for this parameter) and the population proportion p, respectively.

The other symbols in the figure stand for other parameters and statistics that we will encounter.



Break

By the 18th century, the term "statistics" designated the systematic collection of demographic and economic data by states. For at least two millennia, these data were mainly tabulations of human and material resources that might be taxed or put to military use. In the early 19th century, collection intensified, and the meaning of "statistics" broadened to include the discipline concerned with the collection, summary, and analysis of data.



Definitions:

Statistics is a mathematical science that is concerned with the collection, analysis, interpretation or explanation, and presentation of data.

OR

Statistics a set of tools for collecting, organizing, presenting and analyzing numerical facts or observations.

That is, statistics provides methods for:

1. Design: Planning and carrying out research studies.
2. Description: Summarizing and exploring data.
3. Inference: Making predictions and generalizing about phenomena represented by the data.

Measurement is a number or attribute computed for each member of a population or of a sample. The measurements of sample elements are collectively called the **Sample Data**.

Parameter is a number that summarizes some aspect of the population as a whole.

Statistic is a number computed from the sample data.

It is a number of describing a sample characteristic. Results from the manipulation of sample data according to certain specified procedure.

Parameters and Statistic

- Usually the features of the population under investigation can be summarized by numerical parameters. Hence the research problem usually becomes as on investigation of the values of parameters. These population parameters are unknown and sample statistics are used to make inference about them. That is, a statistic describes a characteristic of the sample which can then be used to make inference about unknown parameters.
- A parameter is an unknown numerical summary of the population. A statistic is a known numerical summary of the sample which can be used to make inference about parameters.
- So the inference about some specific unknown parameter is based on a statistic. We use known sample statistics in making inferences about unknown population parameters. The primary focus of most research studies is the parameters of the population, not statistics calculated for the particular sample selected. The sample and statistics describing it are important only insofar as they provide information about the unknown parameters.

Consider the research problem of finding out what percentage of 18-30 year-olds are going to movies at least once a month.

- Parameter: The proportion p of 18-30 year-olds going to movies at least once a month.
- Statistic: The proportion $\hat{p}$ of 18-30 year-olds going to movies at least once a month calculated from the sample of 18-30 year-olds.

Population and Sample: Population and sample are two basic concepts of statistics. Population can be characterized as the set of individual persons or objects in which an investigator is primarily interested during research problem. Sometimes wanted measurements for all individuals in the population are obtained, but often only a set of individuals of that population are observed; such a set of individuals constitutes a sample. Population may Finite and Hypothetical.

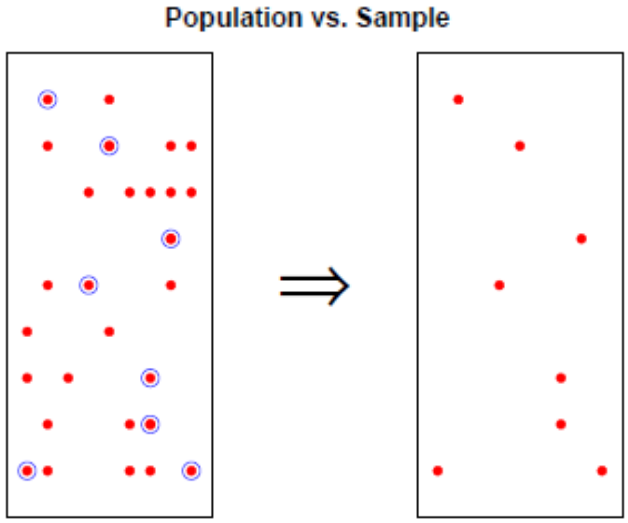
This gives us the following definitions of population and sample:

Population is the collection of all individuals or items under consideration in a statistical study.

Sample is that part of the population from which information is collected.

OR Sample is any subset or sub-collection of population, including the case that sample consists of whole population examined, in which case it is termed a census.

Random Sample is that part of the population which selected in such a way that each member of population has equal opportunity to be selected.



Individuals are the objects described by the data.

Variables are characteristics of an individual. In order to present data, we must first recognize the types of data under consideration. Also can be defined it is a phenomenon that may take on different values. Variables whose values are determined by chance are called random variable.

Types of Data: A data set provides information about a group of individuals. These individuals are, typically, representatives chosen from a population under study. Data on the individuals are meant, either informally or formally, to allow us to make inferences about the population.

Data are characteristics or numbers that are collected by observation.

Data are the values (measurements or observations) that the variable can assume.

- Categorical Variables (Qualitative) partition the individuals into classes. Other names for categorical variables are levels or factors. One special type of categorical variables are ordered categorical variables that suggest a ranking, say small. medium, large or mild, moderate, severe.

- Quantitative Variables are those for which arithmetic operations like addition and differences make sense. Two types of Quantitative variables are Continuous like weight and height and Discrete like counting of cars and holidays per year.

Case (2) (individuals and variables). We consider two populations – the first is the nations of the world and the second is the people who live in those countries. Below is a collection of variables that might be used to study these populations.

nations	people
population size	age
time zones	height
average rainfall	gender
life expectancy	ethnicities
mean income	annual income
literacy rate	literacy
capital city	mother's maiden name
largest river	marital status

Case (3) Classify the variables as quantitative or categorical in the example above.

Types of Statistics: Statistics can be divided into two branches:

- **Descriptive Statistics** is the branch of statistics that involves organizing, displaying and describing data.

Descriptive statistics includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, and percentiles.

- **Inferential Statistics** is the branch of statistics that involves drawing conclusions about a population based on information contained in a sample taken from that population.

Inferential statistics includes methods like point estimation, interval estimation and hypothesis testing which are all based on probability theory.

Break

Information is power and statistics provide the power you need to succeed. Below the ways statistics can and should inform your business decisions.

- Customers
(Need to find numerical and categorical data / information for all customers).

- Income
(Need to know the cost of your products/services).

- Competition
(Need to know information of your competitors and business competition).

- Operating Costs
(Need to know the average operating costs and revenue a typical business.

Case (4) (Descriptive and Inferential Statistics). Consider event of tossing dice. The dice is rolled 100 times and the results are forming the sample data. Descriptive statistics is used to grouping the sample data to the following table:

Outcome of the roll	Frequencies in the sample data
1	10
2	20
3	18
4	16
5	11
6	25

Inferential statistics can now be used to verify whether the dice is a fair or not.

- Descriptive and inferential statistics are interrelated. It is almost always necessary to use methods of descriptive statistics to organize and summarize the information obtained from a sample before methods of inferential statistics can be used to make more thorough analysis of the subject under investigation.

- Furthermore, the preliminary descriptive analysis of a sample often reveals features that lead to the choice of the appropriate inferential method to be later used.

- Sometimes it is possible to collect the data from the whole population. In that case it is possible to perform a descriptive study on the population as well as usually on the sample. Only when an inference is made about the population based on information obtained from the sample does the study become inferential.

Simple Random Sampling

- Researchers adopt a variety of sampling strategies. The most straightforward is simple random sampling. Such sampling requires every member of the population to have an equal chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance.

- To check your understanding of simple random sampling, consider the following example.

Case (5) A research scientist is interested in studying the experiences of twins raised together versus those raised apart. She obtains a list of twins from the National Twin Registry, and selects two subsets of individuals for her study. First, she chooses all those in the registry whose last name begins with Z. Then she turns to all those whose last name begins with B. Because there are so many names that start with B, however, our researcher decides to incorporate only every other name into her sample. Finally, she mails out a survey and compares characteristics of twins raised apart versus together.

Case (6) Substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

Discussion: the population consists of all students in the class. The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population

Keys:

- ✓ Statistics is a study of data: describing properties of data (descriptive statistics) and drawing conclusions about a population based on information in a sample (inferential statistics).
- ✓ The distinction between a population together with its parameters and a sample together with its statistics is a fundamental concept in inferential statistics.
- ✓ Information in a sample is used to make inferences about the population from which the sample was drawn.

Question: What is the difference between Descriptive and Inferential Statistics?

Practices

EXERCISES

1. Explain what is meant by the term *population*.
2. Explain what is meant by the term *sample*.
3. Explain how a sample differs from a population.
4. Explain what is meant by the term *sample data*.
5. Explain what a *parameter* is.
6. Explain what a *statistic* is.
7. Give an example of a population and two different characteristics that may be of interest.
8. Describe the difference between *descriptive statistics* and *inferential statistics*. Illustrate with an example.
9. Identify each of the following data sets as either a population or a sample:
   a. The grade point averages (GPAs) of all students at a college.
   b. The GPAs of a randomly selected group of students on a college campus.
   c. The ages of the nine Supreme Court Justices of the United States on January 1, 1842.
   d. The gender of every second customer who enters a movie theater.
   e. The lengths of Atlantic croakers caught on a fishing trip to the beach.
10. Identify the following measures as either quantitative or qualitative:
    a. The 30 high-temperature readings of the last 30 days.
    b. The scores of 40 students on an English test.
    c. The blood types of 120 teachers in a middle school.
    d. The last four digits of social security numbers of all students in a class.
    e. The numbers on the jerseys of 53 football players on a team.
11. Identify the following measures as either quantitative or qualitative:
    a. The genders of the first 40 newborns in a hospital one year.
    b. The natural hair color of 20 randomly selected fashion models.
    c. The ages of 20 randomly selected fashion models.
    d. The fuel economy in miles per gallon of 20 new cars purchased last month.
    e. The political affiliation of 500 randomly selected voters.
12. A researcher wishes to estimate the average amount spent per person by visitors to a theme park. He takes a random sample of forty visitors and obtains an average of $28 per person.
    a. What is the population of interest?
    b. What is the parameter of interest?
    c. Based on this sample, do we know the average amount spent per person by visitors to the park? Explain fully.

13. A researcher wishes to estimate the average weight of newborns in South America in the last five years. He takes a random sample of 235 newborns and obtains an average of 3.27 kilograms.

    a. What is the population of interest?
    b. What is the parameter of interest?
    c. Based on this sample, do we know the average weight of newborns in South America? Explain fully.

14. A researcher wishes to estimate the proportion of all adults who own a cell phone. He takes a random sample of 1,572 adults; 1,298 of them own a cell phone, hence 1298/1572 ≈ .83 or about 83% own a cell phone.

    a. What is the population of interest?
    b. What is the parameter of interest?
    c. What is the statistic involved?
    d. Based on this sample, do we know the proportion of all adults who own a cell phone? Explain fully.

15. A sociologist wishes to estimate the proportion of all adults in a certain region who have never married. In a random sample of 1,320 adults, 145 have never married, hence 145/1320 ≈ .11 or about 11% have never married.

    a. What is the population of interest?
    b. What is the parameter of interest?
    c. What is the statistic involved?
    d. Based on this sample, do we know the proportion of all adults who have never married? Explain fully.

16. a. What must be true of a sample if it is to give a reliable estimate of the value of a particular population parameter?
    b. What must be true of a sample if it is to give *certain* knowledge of the value of a particular population parameter?

ANSWERS

1. A population is the total collection of objects that are of interest in a statistical study.

3. A sample, being a subset, is typically smaller than the population. In a statistical study, all elements of a sample are available for observation, which is not typically the case for a population.

5. A parameter is a value describing a characteristic of a population. In a statistical study the value of a parameter is typically unknown.

7. All currently registered students at a particular college form a population. Two population characteristics of interest could be the average GPA and the proportion of students over 23 years.

9. a. Population.
   b. Sample.
   c. Population.
   d. Sample.
   e. Sample.

11. a. Qualitative.
    b. Qualitative.
    c. Quantitative.
    d. Quantitative.
    e. Qualitative.

13. a. All newborn babies in South America in the last five years.
    b. The average birth weight of all newborn babies in South America in the last five years.
    c. No, not exactly, but we know the approximate value of the average.

15. a. All adults in the region.
    b. The proportion of the adults in the region who have never married.
    c. The proportion computed from the sample, 0.1.
    d. No, not exactly, but we know the approximate value of the proportion.