Chapter 1

Overview and Descriptive Statistics

This statement underscores the importance of statistical concepts and methods in gaining a deeper understanding of the world, particularly in fields like engineering and science. Statistics offers powerful tools for analyzing data, identifying patterns, and making sense of the inherent variability that exists in real-world phenomena. For example, in engineering, statistical methods might be used to assess the quality and performance of materials, design efficient systems, or predict how components will behave under different conditions. In science, they help researchers design experiments, analyze results, and draw valid conclusions despite uncertainty and natural variation in measurements. Ultimately, statistical techniques are essential for making informed decisions, drawing conclusions from data, and advancing knowledge in any field where uncertainty or variation is present.

1.1 Populations, Samples, and Processes

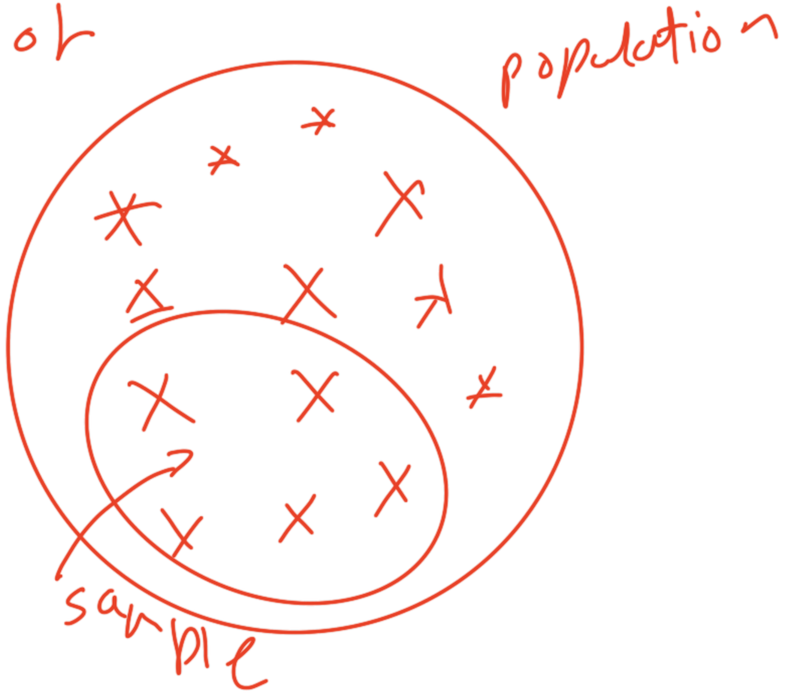
- Engineers and scientists are constantly exposed to collections of facts, or **data**, both in their professional capacities and in everyday activities. The discipline of statistics provides methods for organizing and summarizing data and for drawing conclusions based on information contained in the data.

- An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest.

  Example. Investigation might involve the population consisting of all individuals who received a B.S. in engineering during the most recent academic year (during a specified period).

population → The entire collection of objects

sample → subset of the population



parameter → any measure for population.

statistic → any measure for sample

census → data from every element of population

sample survey → data from every element of sample

- When desired information is available for all objects in the population, we have what is called a **census**.

- Constraints on time, money, and other scarce resources usually make a census impractical or infeasible.

Population → Time, money and resources

What should we do?

- A subset of the population which is known as a **sample** is selected in some prescribed manner.

- We are usually interested only in certain **characteristics** of the objects in a population.

Examples:

1. The number of flaws on the surface of each casing
2. The thickness of each capsule wall
3. The gender of an engineering graduate
4. The age at which the individual graduated

Data

Variable

Numerical (Quantitative)

Ex- Age, Weight, Temperature, Distance

Categorical (Qualitative)

Ex- Colours, Names, Countries

This statement explains the two primary types of characteristics that can be observed in data: **categorical** and **numerical**.

1. **Categorical characteristics** are those that take on values which represent categories or groups. For example, gender can be categorized as male or female. This category does not have inherent numerical meaning; they simply represent different types or classifications.

2. **Numerical characteristics**, on the other hand, involve values that are measurable and can be expressed as numbers. Examples include age (e.g., 23 years) or diameter (e.g., 0.502 cm). These characteristics can be subjected to arithmetic operations, such as addition or averaging, because they are quantities that have a numerical scale.

Understanding whether a characteristic is categorical or numerical is important because it determines the type of statistical methods that can be applied to analyze the data effectively.

- A **variable** is any characteristic whose value may change from one object to another in the population. We shall initially denote variables by lowercase letters from the end of our alphabet.

Examples:

x = Hight of students

y = number of visits to a specific coffee shop a specified period

z = Students' GPA

Example 4.

Study Example 1.4

- A **univariate** data set consists of observations on a single variable.

  Example. We might determine the type of transmission, automatic (A) or manual (M), on each of ten automobiles recently purchased at a certain dealership, resulting in the **categorical data set**

  M A A A M A A M A A

  The following sample of pulse rates (beats per minute) for patients recently admitted to an adult intensive care unit is a **numerical univariate** data set:

  88 80 71 103 154 132 67 110 60 105

- We have **bivariate** data when observations are made on each of two variables.

  Example. Our data set might consist of a (height, weight) pair for each basketball player on a team with the first observation as (72,168), the second as (75, 212), and so on.

  If an engineer determines the value of both $x =$ component lifetime and $y =$ reason for component failure, the resulting data set is bivariate with one variable numerical and the other categorical.

- **Multivariate** data arises when observations are made on more than one variable (so bivariate is a special case of multivariate).

  Example. A research physician might determine the systolic blood pressure, diastolic blood pressure, and serum cholesterol level for each patient participating in a study.

➢ Branches of Statistics

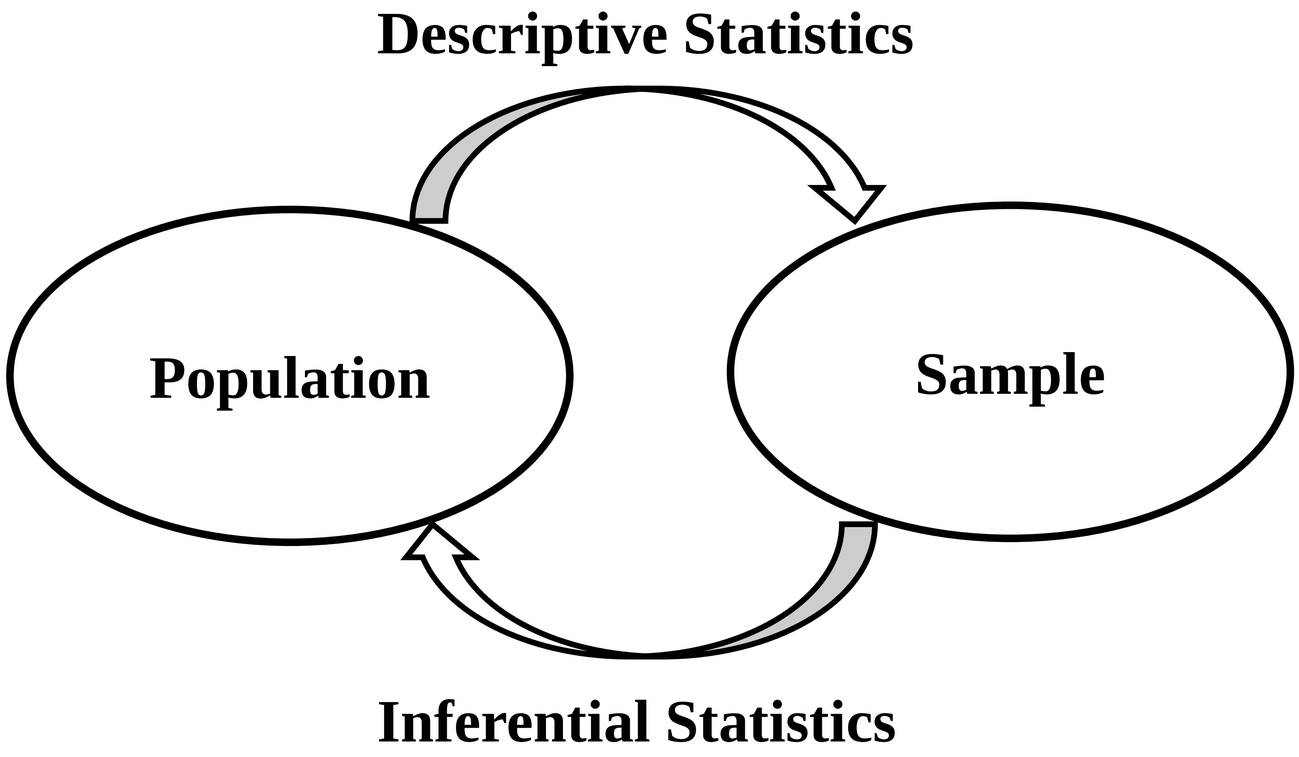
1. Descriptive Statistics: Techniques to summarize and describe important features of the data.

- Graphical Methods: histograms, boxplots, and scatter plots
- Calculation of numerical Methods: summary measures, such as means, standard deviations, and correlation coefficients

2. Inferential Statistics: Techniques to generalize from a sample to a population.

The most important types of inferential procedures-point estimation, hypothesis testing, and estimation by confidence intervals. These will be discussed later.

① estimation
② hypothesis testing


Descriptive Statistics
Population
Sample
Inferential Statistics

	Descriptive Statistics	Inferential Statistics
Purpose	Summarizes and describes features of a dataset	Makes inferences, predictions, or generalizations about a population based on sample data
Scope	Focuses on specific sample data	Extends findings to a larger population
Objective	Describes characteristics of the data without generalizing	Generalizes findings from sample to population
Examples	Measures of central tendency, dispersion, frequency distributions, graphical representations	Hypothesis testing, regression analysis, confidence intervals
Data Analysis	Provides a summary and visualization of data	Draws conclusions, tests hypotheses, and makes predictions
Population Representation	Represents features within the sample only	Represents features of the larger population
Statistical Techniques	Mean, median, mode, range, variance, standard deviation, histograms, box plots, etc.	Hypothesis testing, regression analysis, confidence intervals
Goal	To provide insights into the characteristics of a dataset	To make predictions or draw conclusions about a population

Example 4.

Study Example 1.4

Example 1 (Example 1.1). Here is data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

Without any organization, it is difficult to get a sense of the data's most prominent features-what a typical (i.e. representative) value might be, whether values are highly concentrated about a typical value or quite dispersed, whether there are any gaps in the data, what fraction of the values are less than 20%, and so on.

Example 2.

Study Example 1.2

Page 5

- The relationship between the two disciplines can be summarized by saying that probability reasons from the population to the sample (deductive reasoning), whereas inferential statistics reasons from the sample to the population (inductive reasoning).

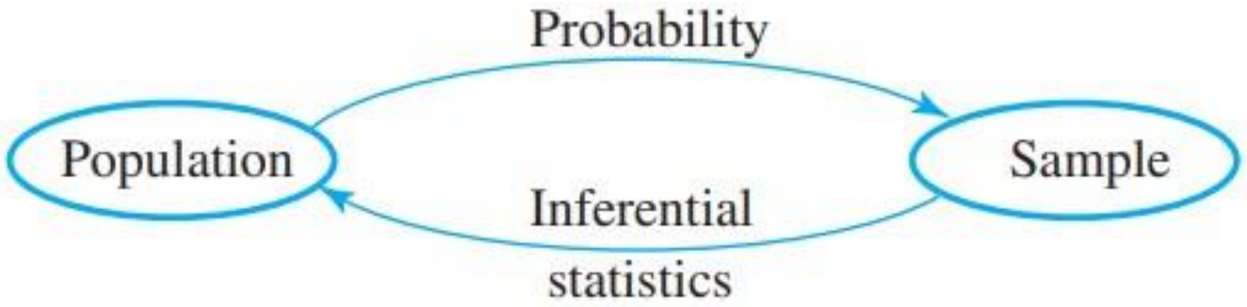


Figure 1.2 The relationship between probability and inferential statistics

Example 3 (Example 1.3) (Probability & Inferential Statistics). Consider drivers' use of manual lap belts in cars equipped with automatic shoulder belt systems.

- **In Probability**

→ data (population)
→ measures

start from population ⟶ samples

- **In Inferential Statistics** → don't know all population data

➢ Enumerative Versus Analytic Studies

- **Enumerative Study:** Here the goal or purpose of the study is identifiable, i.e., not ambiguous, and the elements of the population are well defined and unchanging under this study. The population could be an existing finite population about which one wants to draw conclusions.

  Example. The frame may contain serial numbers of all furnaces manufactured by a particular company during a certain time period; a sample may be selected to infer something about the average lifetime of these units.

- **Analytic Study:** A study in which action will be taken on a process to improve performance in the future (e.g., recalibrating equipment or adjusting the level of some input such as the amount of a catalyst). Generally, here the result of the study is new because the objective is to improve things to be used in the future. The study does not have a well-defined sampling frame, and the impact of this study is highly localized and short term.

  Example. In the production industry, where the new product is developed as an improvement over the existing one. There is no sampling frame.

➢ Collecting Data

Read

Pages 10, 11

Example 4.

Study Example 1.4

Example 5.

Study Example 1.5

Page 11

Exercises

1, 3, 6, 8

Exercise 1. Give one possible sample of size 4 from each of the following populations:

a. All daily newspapers published in the United States
b. All companies listed on the New York Stock Exchange
c. All students at your college or university
d. All grade point averages (GPA) of students at your college or university

Exercise 3. Consider the population consisting of all computers of a certain brand and model, and focus on whether a computer needs service while under warranty.

a. Pose several probability questions based on selecting a sample of 100 such computers.
b. What inferential statistics question might be answered by determining the number of such computers in a sample of size 100 that need warranty service?

Exercise 6. The California State University (CSU) system consists of 23 campuses, from San Diego State in the south to Humboldt State near the Oregon border. A CSU administrator wishes to make an inference about the average distance between the hometowns of students and their campuses. Describe and discuss several different sampling methods that might be employed. Would this be an enumerative or an analytic study? Explain your reasoning.

Exercise 8. The amount of flow through a solenoid valve in an auto mobile's pollution-control system is an important characteristic. An experiment was carried out to study how flow rate depended on three factors: armature length, spring load, and bobbin depth. Two different levels (low and high) of each factor were chosen, and a single observation on flow was made for each combination of levels.

a. The resulting data set consisted of how many observations?

b. Is this an enumerative or analytic study? Explain your reasoning.