Chapter (1)

1. **Statistics:**
   The science of collecting, organizing, summarizing, analyzing numerical information called data (Descriptive Statistics), and drawing conclusions (Inferential Statistics).

2. **Type of data:**
   a. Qualitative variables:
      Variables that can be separated into different categories distinguished by some nonnumeric characteristic.
      Example: Gender (male or female), Race (White, Black, Hispanic, etc.), Favorite Color (Blue, Red, Silver, etc.)
   
   b. Quantitative variables:
      Variables consisting of numbers representing counts or measurements.
      Example: Age, Height, Speed, Test scores, the number books in our bookstore and salary.

3. **Type of Quantitative variables:**
   a. Discrete variables: Data that can assume the values corresponding to isolated points along a line interval. In this type the data are to be counted.
      Example: 1. Numbers of telephone calls made at the switchboard of our school everyday.
               2. Number of car accidents on 405 FWY.
               3. Number of babies delivered at Long Beach Memorial hospital
   
   b. Continuous variables: Data that can assume any value along a line interval, including every possible value between any two values. In this type the data are to be measured.
      Example: 1. Height of boys born at UCLA hospital on fourth of July.
               3. Amount (Volume) of coffee consumed by Americans in one day.

4. **Variable** - Property of an object or event that can take on different values. For example, college major is a variable that takes on values like mathematics, computer science, English, psychology, etc. Variables whose values are determined by chance are called random variables.

5. **Data:** List of observations a variable assumes

6. **Variable versus Data:** Example: Gender is a variable whereas being male or female is the data

7. **Control group:** is a group of subjects in an experiment who are not given a particular treatment. (Like Placebo)

8. **Experimental group:** is a group of subjects in an experiment who are given a particular treatment. (Like new drug)

9. **Double blind:** Neither the doctor nor the patient know whether he or she is part of the experimental or control group.

10. **Population:** The complete collection of all elements to be studied. (Scores, People, Measurements,)

11. **Sample:** A sub collection of elements drawn from a population.
Example 1:
A quality control manager randomly selects 50 bottles of Coca-Cola that were filled on October 15 in order to assess the calibration of the filling machine. **Determine the individuals, the population and the sample.**
**Ans:** The population consists of all bottles of Coca-cola filled by that particular machine on October 15. The individuals are just the individual bottles. The sample consists of the 50 bottles selected by the quality control manager.

Example 2:
A researcher is claiming that the average age of women who are graduated from Engineering School at UCLA is about 28 years. To test his hypothesis, he randomly selected 300 female engineers who have graduated from UCLA school of Engineering. Determine the population, Identify the variable of interest, is the variable quantitative (qualitative)? Is the variable discrete or continuous? Describe the sample. Describe the inference.

12. **Parameter:** Numerical measurement describing some characteristic of a population
   **Example:** \( \mu \) = Population mean, \( \sigma \) = Population standard deviation and \( p \) = Population proportion

13. **Statistic:** Numerical measurement describing some characteristic of a sample
   **Example:** \( \bar{x} \) = sample mean, \( s \) = Sample standard deviation and \( \hat{p} \) = Sample proportion

14. **Census versus a sample:**
   Census is a collection of data from every element in a population whereas a sample is a subset of a population.

15. **Survey:** We study a part of a larger population in order to understand the whole

16. **Survey sampling:**
   a. **Observational study** (Association/relationship between two variables):
      Study in which we observe and measure specific characteristics, but do not attempt to manipulate or modify the subjects being studied.
      **Example:** The incidence of lung disease in a sample of workers in asbestos factories is compared to the incidence of lung disease in a sample of college professors.
   b. **Designed experiment** (Causation):
      Study in which a treatment is applied to the experimental units (individuals) and attempts to manipulate or modify the subjects being studied.

   **The advantage of an experiment over an observational study is that an experiment is controlled**

An experiment is defined by the following types of variables: (Response and Predictor)

17. **Response variable:** The variable, which measures the response of units or subjects to the various treatment.
   **Example:** A researcher is interested in determining if one can predict the scores on a statistics exam from the amount of time spent studying for the exam. Identify the response variable. (The scores on the exam)
   **Example:** A large study used records from Canada’s national health care system to compare the effectiveness of two ways to treat prostate disease, traditional surgery and a new method that does not require surgery. You have 300 prostate patients who are willing to serve as subjects in an experiment to compare the two methods. What is the response variable in this experiment? [Existence of prostrate disease is the response variable]
   **Example:** A study is done to compare the lung capacity (measured by certain breathing tests) of coal miners to the lung capacity of farm workers. The researcher is able to study 200 workers of each type. (Lung capacity)

18. **Predictor Variable:** The factor(s) that affect the response variables

19. **Levels of predictor variable:** The values that a factor can take are called the levels of the factor. For example, a drug dosage (the factor) may be administered at three different levels.
20. **Lurking variable:** The factors that are related to our study but they are not being identified.

21. **Frame:** The list of all individuals within the population.

**Example:**
A school psychologist wants to test the effectiveness of a new method for teaching reading. She selects five hundred first grade students in Long Beach District and randomly divides them into two groups. Group 1 is taught by means of the new method, while Group 2 is taught via traditional methods. The same teacher is assigned to teach both groups. At the end of the year, an achievement test is administered and the results of the two groups compared.

**Determine:**
- **a. Population**
- **b. Sample**
- **c. Subject units**
- **d. Response variable**
- **e. Treatment**
- **f. Levels of the treatment**
- **g. Predictors**
- **h. Designed/Observational study?**

**Sampling Techniques**

1. **Random sampling** (Simple random sample)
   In this technique, each member of a population has an equal chance of being selected. Each member of the population is assigned a number. You can select a random sample of any population by using a calculator or computer to generate random numbers.

   **Example:** A list of students in elementary statistics is obtained in which the individuals are numbered 1 to 65. A professor randomly selects 12 of the students.

   **IT IS DIFFICULT TO OBTAIN A SIMPLE RANDOM SAMPLE (SRS) IN PRACTICE!**

2. **Stratified sampling** (Separating the population into non-overlapping groups)
   In this technique, a population is divided into at least two different subsets, called strata that share a similar characteristic. A sample is then randomly selected from each. **(Conducting a SRS separately within each strata)**
   The defining characteristic can be gender, age, or even political preference. Using a stratified sample ensures that each segment of a population is represented. For this reason, stratified samples are usually preferred over simple random samples.

   **Example:** A researcher segments the population of car owners into four groups: Ford, General Motors, Chrysler, and foreign. She obtains a random sample from each group and conducts a survey.

3. **Cluster sampling**
   To select a cluster sample, divide a population into groups, called clusters, then select all of the members in one or more, but not all, of the clusters. This technique is often used because of practical or economical restrictions, but data collected may be less reliable than when a random sample is used.

   **Example:** A researcher randomly selects 5 of the 70 hospitals in Long Beach area and then surveys all of the surgical doctors in each hospital.

4. **Systematic sampling**
   In this technique, a population is ordered in some way and then members of the population are selected at regular intervals. The selection process can start at any randomly chosen point.
   **An advantage of systematic sample is that it is easy to use.**

   **Example:** An interviewer in a mall is told to survey every fifth shopper, starting with the second. Systematic sampling **(When the population size is known)**
Procedure for systematic sampling when the population size is known:

1. N (Pop. Size)
2. n (sample size)
3. Form N/n and round it down to the nearest integer and call it K
4. Select a number between 1 and K and call this number p
5. The sample will consist of the following individuals: p, p + k, p + 2 k, ..., p +(n-1) k

5. Convenience sampling (Self-selected individuals)
In this technique, simply use any members of population that are readily available. This method is likely to produce biased results. (Usually contains the most bias)

Example: I. A radio station asks its listeners to call in their opinion regarding the use of American forces in Peacekeeping missions.
II. A professor would like to know how many hours per week college students spend watching television. She is teaching two large classes and uses all students in those classes as her sample.

Sampling error:
The difference between a characteristic of the entire population and a sample of that population.

➢ Sources of Errors in sampling:

a. Sampling error: The error that results from using sampling to estimate information regarding a population.

Example: The size of the sample, the amount of variation that exists in the population (i.e. How different the members of the population are from one another with regard to the variable being studied)

b. Non-sampling error:

Example: Respondent’s lying, measurement errors, poorly worded questions, and the error due to people not responding.

Sources of bias:

Sampling bias:
A systematic tendency to exclude one type of person from the sample. A large sample will not solve this problem.

Non-Response bias:
This is when people who do not answer questions are different from people who do.

Undercoverage bias:
This is when some groups in the population are left out of the process of choosing the sample.

Response bias:
This is when the individuals do not reply truthfully.

Voluntary Response bias:
This is when the survey relies on individuals who volunteer to respond. (e.g. Internet surveys) and they are unscientific and unreliable.

The following can affect the validity of a study:

a. The method in which a sample was obtained
b. Wording of questions
c. The order in which questions are presented

Example: A magazine is conducting a study on the effects of infidelity in a marriage. The editor randomly select 400 women whose husbands were unfaithful and ask “Do you believe a marriage can survive when the husband destroys the trust that must exist between husband and wife?” What is wrong with wording of the question that was asked?
Ans: “Do you believe that a marriage can be maintained after an extramarital relation?

A key component of a well-designed experiment is RANDOMIZATION