Chapter 1—Data Collection

1.1 Introduction to the Practice of Statistics

Definition: **Statistics** is the science of collecting, organizing, summarizing, and analyzing information in order to draw conclusions.

The objective of statistics is to make *inferences* (predictions) about a *population* based upon information contained in a *sample*.

- **Population**—is the set representing of all observations of interest to the sample collector.
- **Sample**—is a subset of measurements selected from the population of interest.

Election example

The **Process of Statistics** includes 4 steps:
1. *Identify the research objective.*
2. *Collect information to answer the question posed in (1).*
3. *Organize and summarize the information.*
4. *Draw conclusions from the information.*

Your textbook illustrates the **Process of Statistics** with a tomato example (p. 6) where a young boy plants a tomato plant in his backyard. He wants to know more about tomatoes so he grows the tomato plant and collects information about the tomatoes harvested from the plant. The variable of interest is the weight of the tomatoes.

The **Process of Statistics** for the tomato example:

1. *Identify the research objective*—determine the *mean weight* of a Beef Steak tomato grown in Lubbock, TX.

2. *Collect information to answer the question posed in (1) above*.

A Beef Steak tomato plant is grown in Lubbock, TX, and the weight of each harvested tomato is recorded.

3. *Organize and summarize the information*

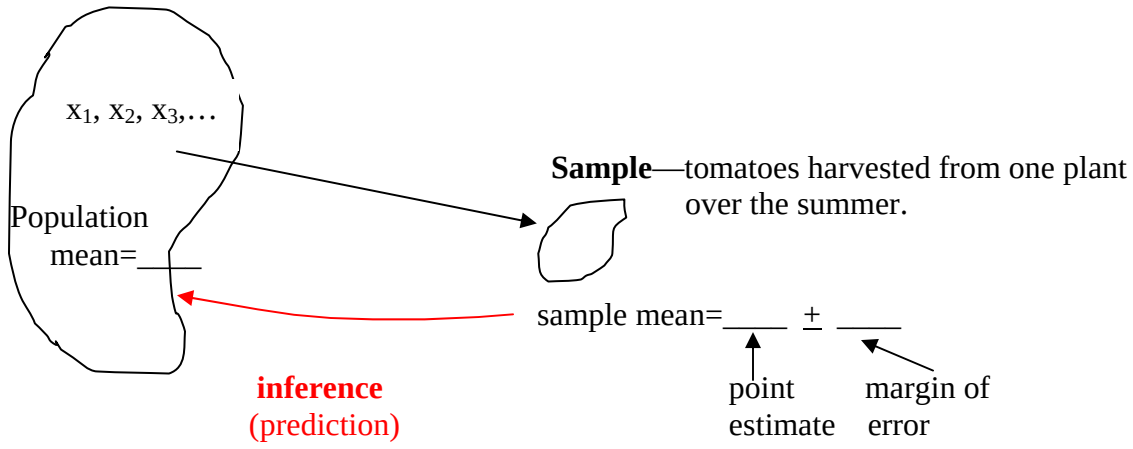
Definitions:	**Variable**—is a measurable characteristic. **Observation** (or **variate**)—is an individual measurement of a variable.

X=tomato weight (in ounces)

x_1=3.2 … The first harvested tomato weighs 3.2 ounces.
x_2=4.1
x_3=2.7
.
.
.
3.45 = mean weight of ALL harvested tomatoes over the summer

4. *Draw conclusions from the information*

Population—all Beef Steak tomato plants grown in Lubbock, TX



Summary: (1) In statistics we study a fraction (or part) of a body of data with the objective of concluding about the entire body of data. (2) Since this approach is based on a fraction of the entire body of data, there is a possibility of errors in the conclusions. Thus, a part of statistics is to assess/evaluate the uncertainty of conclusions.

Is the mean weight of the sample tomatoes from one Beef Steak plant grown in summer 2005 exactly equal to the true (population) mean weight of ALL Beef Steak tomatoes?—Answer is likely NO.

Definition:	**Descriptive statistics**—consists of organizing and summarizing the information collected. **Inferential statistics**—uses methods that generalize results obtained from a sample to the population and measure their reliability.

Types of Data (p. 6)

Two types of data:

1. Primary data—obtained from the field or lab (e.g., animal gain data from a feeding trial)
2. Secondary data—previously collected and published (e.g., USDA crop price data).

Variables can be classified into two groups.

Definition: **Qualitative or categorical variables**—allow for classification of individuals based on some attribute or characteristic.
Quantitative variables—provide numerical measures of individuals. Arithmetic operations such as addition and subtraction can be performed on the values of a quantitative variable and provide meaningful results.
Discrete variable—is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term "countable" means that the values result from counting such as 0, 1, 2, 3, and so on.
Continuous variable—is a quantitative variable that has an infinite number of possible values that are not countable.

Examples:

Variable	**Qualitative or Quantitative**	**Discrete or Continuous**
Gender		
Temperature		
Number of children in a family		
Zip code		
Race		
Weight of a steer		

Remember: If you **count** to get the value of a quantitative variable>>>>>**discrete**
If you **measure** to get the value of a quantitative variable>>>**continuous**

1.2 Observational Studies; Simple Random Samples

1. Census vs. Sample

Definition: **Census** (p. 13)—is a list of all individuals in a population along with certain characteristics of each individual. Examples are:

- The U.S. conducts a census every 10 years (cost was $6 billion in 2000).
- An election is a census. Every registered voter can vote.

When collecting information, compare taking a census to using inferential statistics.

> "If the entire population is studied, then inferential statistics is not necessary because descriptive statistics would provide all the information that we need regarding the population" (p. 5).

2. Observational Study vs. Designed Experiment

Definition: **Observational study** (p. 14)—measures the characteristics of a population by studying individuals in a sample. Observational studies are referred to as *ex post facto* (after-the-fact) studies because the value of the variable of interest has already been established.

Designed experiment (p. 14)—applies a treatment to individuals (referred to as **experimental units**) and attempts to isolate the effects of the treatment on a **response variable**.

Example #1 (p. 14): Research question: Does smoking cause lung cancer?

Observational study:

- Select a sample of adult individuals (18 years old) who were alive in say 1930.
- Divide the group into those who smoked over the 25-year period (1930-55) and those who did not smoke. Eliminate individuals who smoked for part of the period (but not the entire period).
- Review the life history of the sample of smokers and non-smokers to determine if they died from lung cancer.
- Calculate the percentage of smokers who died from lung cancer and the percentage of non-smokers who died from lung cancer over the 25-year period.

Control group—are the nonsmokers. That is, the nonsmokers serve as the benchmark upon which the smokers' rate of lung cancer is judged. The hope is that nonsmokers and smokers are alike in all of their traits (such as diet and exercise) except for smoking.

What if the sample results show the following?

2.5% of smokers died from lung cancer
0.7% of non-smokers died

Does this indicate that smoking causes lung cancer—YES or NO?

Potential Problem: Because of **lurking variables**, the fact that smokers' have a higher incidence of lung cancer does not mean that smoking causes lung cancer.

Lurking variable (p. 177)—is a variable that may affect the response variable (death from lung cancer). Lurking variables in the observational study of lung cancer include:

- diet
- amount of exercise

Designed Experiment:

- Obtain a sample of people and divide them into two groups.
- One group will be required to smoke 1 pack of cigarettes per day, while the other group will not be allowed to smoke.
- Control the diet and the amount of exercise for both groups so they are exactly the same.
- Monitor the number of deaths from lung cancer from each group.

Summary of observational studies vs. designed experiments:

Observational studies are very useful tools for determining whether there is a relation (association) between two variables, but a designed experiment is required to isolate the cause of the relation (p. 15).

- Observational studies determine **ASSOCIATION** between variables, but not **CAUSATION**.

Explain how to carry out a Designed Experiment to compare cattle diets for feedlot cattle. Assume there are two diets currently being used—Diet 1 and Diet 2. How would you conduct an Observational Study to compare the two diets, and what are the disadvantages of the Obs. Study compared to the Design. Exp.?

3. Simple Random Sampling

A sample of size n from a population of size N is obtained through **simple random sampling** if every possible sample of size n has an equally likely chance of occurring.

Random sample—is a sample in which any one individual observation in the population is as likely to be included as any other.

Note: N=number of observations in the population
n=number of observations in the sample

'Simple random sampling is like selecting names from a hat'.

Frame—a list of ALL individuals in the POPULATION.

Sampling without replacement: Once an individual is selected to be in the sample, he/she cannot be selected again (this is so we don't select the same individual twice).

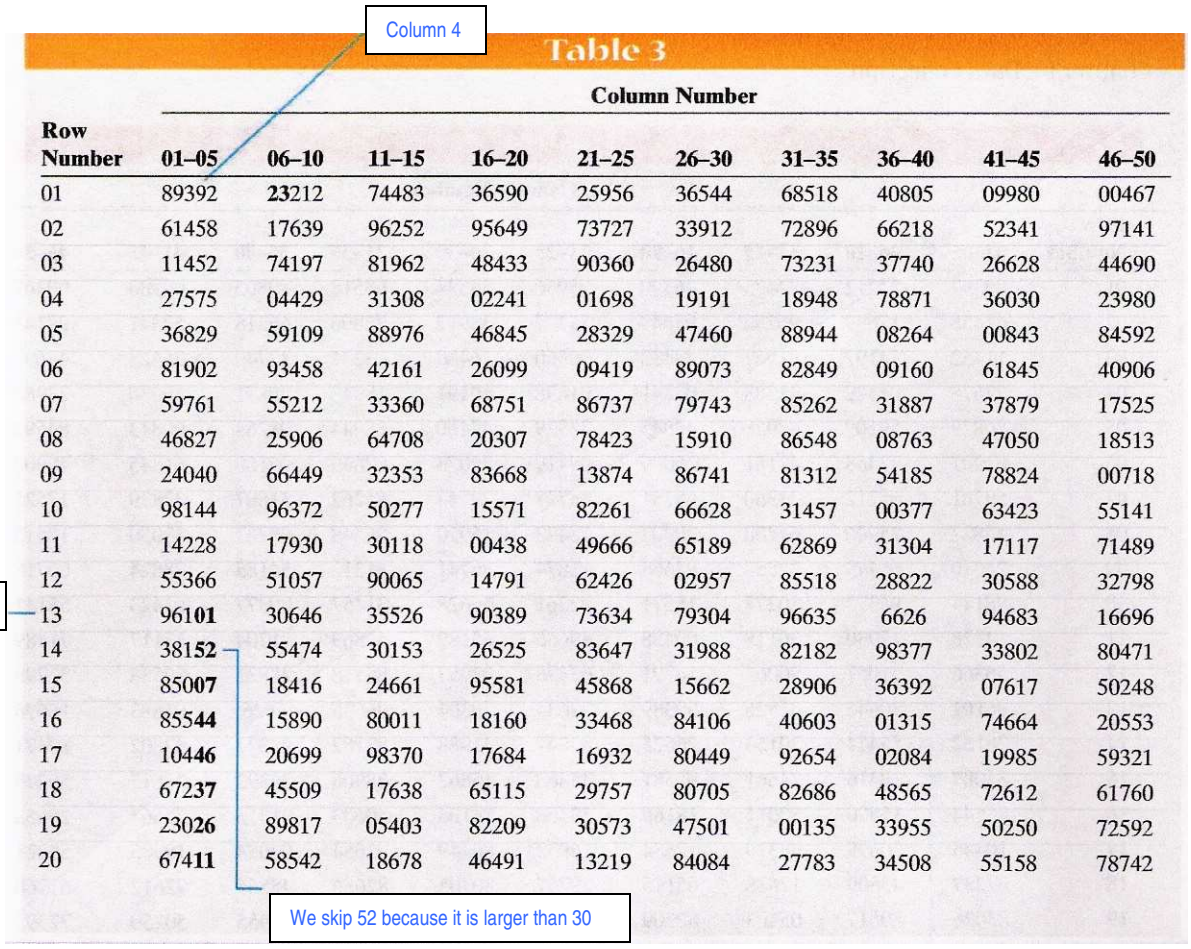
Example #3 (p. 17)—Obtaining a Simple Random Sample:
Problem: Senese and Associates has increased their accounting business. To make sure their clients are still satisfied with the services they are receiving, Senese and Associates decides to send a survey out to a simple random sample of 5 of its 30 clients.

Your text discusses two ways to obtain a random sample (from a list frame):

1. Use a table of random numbers (Table 3, p. 18).
2. Use technology such as EXCEL.

Using Table 3, explain how to draw a random sample of n=5.

Close your eyes and place your finger on the table—row=13 and column=4.

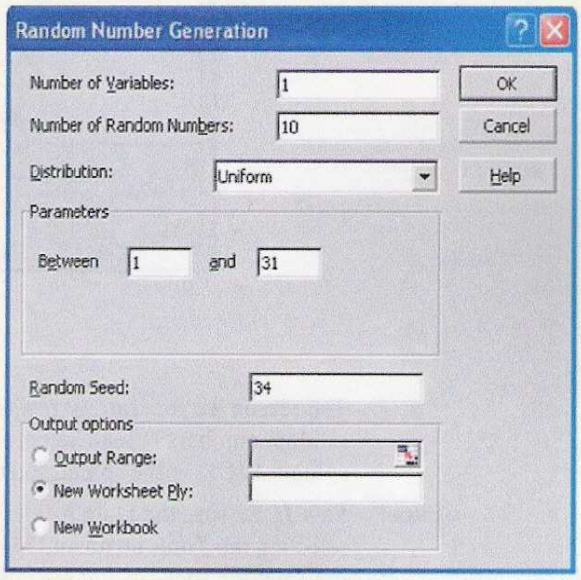


Using EXCEL (p. 22), demonstrate how to draw a random sample of n=5.

Excel ***Step 1:*** Be sure the Data Analysis Tool Pak is activated. This is done by selecting the **Tools** menu and highlighting **Add – Ins** .... Check the box for the Analysis ToolPak and select OK.

Step 2: Select **Tools** and highlight **Data Analysis** .... Highlight **Random Number Generation** and select OK.

Step 3: Fill in the window with the appropriate values. To obtain a simple random sample for the situation in Example 2, we would fill in the following:



The reason we generate 10 rows of data (instead of 5) is in case any of the random numbers repeat. Notice also that the parameter is between 1 and 31, so any value less than or equal to 31 is possible. In the unlikely event that 31 appears, simply ignore it. Select OK and the random numbers will appear in column 1 (A1) in the spreadsheet. Ignore any values to the right of the decimal place.

1.3 Other Types of Sampling

Types of probability samples:

1) Random—Is this class a random sample of Tech students?
2) Stratified
3) Systematic
4) Cluster

Stratified sample (p. 23)—is obtained by separating the population into non-overlapping groups called *strata* and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogeneous (or similar) in some way.

Example #1 (p. 24): Obtaining a stratified sample.

Problem: The President of DePaul University wants to determine the community's opinion about campus safety using a sample of size n=100. How do you draw a stratified sample?

Categories	Population (proportion	Sample
Resident students	6,204 (0.283)	28
Non-resident students	13,304 (0.607)	61
Faculty and staff	2,401 (0.110)	11
	21,909	100

Advantages of a stratified sample:

1) Can obtain more information and more accurate results (based on the assumption that individuals within each subgroup (stratum) have similar characteristics).
2) Characteristics within each stratum can be reported (p. 24).

Systematic sample (p. 25)—is obtained by selecting every kth individual from the population. The first individual selected is a random number between 1 and k.

Advantage: Does not require a frame (like a random sample).

There are two situations for drawing systematic samples depending on whether the population size is known or not:

1) Population size is known.

Steps in systematic sampling (p. 26):

1) Determine the population size N.
2) Determine the sample size desired, n.
3) Compute N/n and round down to the nearest integer. This value is k.
4) Randomly select a number between 1 and k. Call this number p.
5) The sample will consist of the following individuals:
   p, p+k, p+2k, …, p+(n-1)k

Application: A consulting firm would like to know the opinion of airline customers about flight service on Southwest Airlines. Assume the consulting firm knows there are 2000 customers flying on Southwest through Dallas on a particular day. Draw a systematic sample of n=5.

2) Population size is unknown.

Review **Example #2 (p. 25)**: Obtaining a systematic random sample without a frame
Problem: Kroger Food Stores wants to measure consumer satisfaction.

- Sample size of n=40 is to be taken over an entire day.
- k=N/n cannot be determined without knowing the population size, N.
- Problem: If the selected k is too small, the sample may be completed by noon; customers who visit the store after noon will not be surveyed, and these customers may have a different opinion of the store.
  If k is too large, the sample of 40 may not be obtained by days end.

Question: How do you draw a systematic sample with unknown population size? That is, how do you determine "k" (where k=N/n) so that you obtain a representative sample of the entire population?

Would a stratified sample work better? How do you define strata?—by time period?

Determine if the following is an appropriate method to draw a systematic sample of Tech students—Stand behind the Ag. Sciences building and select every 10^{th} student who walks by.

Cluster sample (p. 26)—is obtained by selecting all individuals within a randomly selected collection or group of individuals.

Example #3 (p. 26): Obtaining a cluster sample.

Problem: A sociologist wants to gather data regarding the household income within the City of Boston.

- 10,493 city blocks in Boston (each block is a cluster).
- Randomly select 20 blocks and interview all households from each cluster.

Advantages:

1) No need to obtain detailed frame for the entire population
2) Reduces travel time/cost because only 20 stops are made (each stop is a block or cluster).

Considerations in deciding on the number of clusters and the sample size from each cluster:

Homogeneous clusters—it is better to have more clusters with fewer individuals in each cluster.

For example, consider the household income survey.

- Income from one household on the block is likely similar to that of another household. This results in duplicate information. So, draw more clusters.

Heterogeneous clusters—fewer clusters with more individuals per cluster is appropriate.

For example, consider a quality control engineer at a bottling plant with a production line:

- Obtaining a systematic sample—by waiting for bottles to come off a production line—would be time consuming. By comparison, obtaining a cluster sample at the end of the day (or run) would be time efficient.
- In this case, a cluster would be a single box of say 24 bottles and the cluster sample might contain 10 boxes.
- A cluster likely resembles the heterogeneity of the population and provides a scaled-down representation of the population.

Random vs. Non-Random Sampling:

Research objective: To determine how AAEC students feel about the advice they receive from their advisors.

Place a survey on the Web and obtain a convenience sample.

Convenience sample (p. 27)—is a sample in which the individuals are easily obtained. Individuals in the sample are **self-selected**.

What are the advantages and disadvantages of using a convenience sample compared to a random sample?