Chapter 1 Introduction to Statistics
Statistical and Critical Thinking (1.1)

Objectives

In this lesson, we will learn how to

- describe the basic statistical process: prepare, analyze, and conclude;
- define data, statistics, population, census, and sample;
- identify the population and sample in a real study;
- explain why voluntary response samples are often biased;
- distinguish statistical significance from practical significance;
- recognize common pitfalls that can lead to misleading conclusions.

1. The Statistical Process

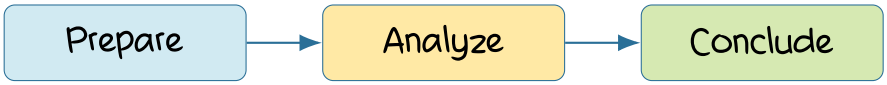
Statistics is not only about calculations. Good statistical work requires us to understand the data, question how they were collected, analyze them carefully, and draw conclusions that make sense in context.

Statistical Thinking

The process involved in conducting a statistical study can be summarized as

$$\text{Prepare} \longrightarrow \text{Analyze} \longrightarrow \text{Conclude}.$$

Statistical thinking requires critical thinking and the ability to make sense of results. It demands much more than simply carrying out complicated calculations.



2. Basic Vocabulary

Data

Data are collections of observations, such as measurements, genders, or survey responses.

Statistics

Statistics is the science of planning studies and experiments, collecting data, organizing and summarizing data, presenting data, analyzing data, interpreting data, and drawing conclusions.

Population, Census, and Sample

Population	The complete collection of all measurements or data being considered.
Census	A collection of data from every member of the population.
Sample	A subcollection of members selected from the population.

Example 1: In a survey of 410 human resource professionals, 148 said that job candidates were disqualified because of information found on social media postings.
Identify the population, the sample, and the data.

Solution

- Population: All human resource professionals.
- Sample: The 410 human resource professionals who were sur-

veyed.

- Data: The individual yes/no responses from the 410 professionals.

The goal is to use information from the sample to draw a reasonable conclusion about the larger population.

.

Prepare

1. Context
   - What do the data represent?
   - What is the goal of study?
2. Source of the Data
   - Are the data from a source with a special interest so that there is pressure to obtain results that are favorable to the source?
3. Sampling Method
   - Were the data collected in a way that is unbiased, or were the data collected in a way that is biased (such as a procedure in which respondents volunteer to participate)?

↓

Analyze

1. Graph the Data
2. Explore the Data
   - Are there any outliers (numbers very far away from almost all of the other data)?
   - What important statistics summarize the data (such as the mean and standard deviation described in Chapter 3)?
   - How are the data distributed?
   - Are there missing data?
   - Did many selected subjects refuse to respond?
3. Apply Statistical Methods
   - Use technology to obtain results.

↓

Conclude

1. Significance
   - Do the results have statistical significance?
   - Do the results have practical significance?

3. Prepare

Before we analyze the data, we should ask whether the data are appropriate for the question being studied.

Prepare: Questions to Ask

- Context: What do the data represent? What is the goal of the study?
- Source of the data: Are the data from a trustworthy source? Does the source have a special interest in the results?
- Sampling method: Were the data collected in a way that is likely to be unbiased?

Example 2: Forensic scientists sometimes use shoe print lengths at crime scenes to estimate a suspect's height. Suppose the following data are collected from eight randomly selected males.

Shoe print (cm)	27.6	29.7	29.7	31.0	31.3	31.4	31.8	34.5
Height (cm)	172.7	175.3	177.8	175.3	180.3	182.3	177.8	193.7

Briefly describe the context, source, and sampling method.

Solution

- Context: We want to determine whether shoe print length can help estimate height.
- Source: The data should come from a reputable source.
- Sampling method: Random selection is important because it helps reduce bias.

Because many burglaries are committed by males, using male data can be appropriate in this forensic context, but we should still be careful

about applying the conclusion too broadly.

4. Voluntary Response Samples

Voluntary Response Sample

A voluntary response sample, also called a self-selected sample, is a sample in which people decide for themselves whether to be included.

Voluntary response samples are usually seriously flawed because people with strong opinions are more likely to respond. This can create strong bias.

Example 3: A television show asked viewers to call in with their opinion about whether the United Nations headquarters should remain in the United States.

- In the call-in poll, 186,000 viewers responded, and 67% said the United Nations should move out of the United States.
- In a separate random survey of 500 people, only 38% said the United Nations should move out of the United States.

Which poll is likely to be more reliable? Why?

Solution

The random survey of 500 people is likely to be more reliable, even though it has a much smaller sample size.
The call-in poll is a voluntary response sample. People with strong opinions are more likely to call, so the sample may be biased. A good sampling method is usually more important than a very large but biased sample.

5. Analyze

After considering the context, source, and sampling method, we begin to analyze the data.

Analyze: What Should We Do?

- Graph the data.
- Explore the data.
- Look for outliers, missing data, unusual patterns, and important summary statistics such as the mean and standard deviation.
- Apply appropriate statistical methods.

Important Reminder

A good statistical analysis does not require elite computational skills. Technology can handle many calculations, but the person doing the analysis must use common sense and sound statistical reasoning.

6. Conclude

The final step is to draw conclusions. We must distinguish between statistical significance and practical significance.

Statistical Significance

A result has statistical significance if it is unlikely to occur by random chance. In this course, we often use 5% or less as a common cutoff.

Practical Significance

A result has practical significance if it is large enough or important enough to matter in real life.

Example 4: Suppose a gender-selection product results in 5200 girls among 10,000 births.

- The sample percentage is 52% girls.
- If girls and boys are equally likely, a result this extreme has a very small probability of occurring by chance.

Explain the difference between statistical significance and practical significance in this situation.

Solution

The result may be statistically significant because 5200 girls out of 10,000 births is unlikely to happen by random chance if the true probability of a girl is 0.5.
However, the result may not have practical significance. Many couples would not spend money or effort to increase the chance of a girl from about 50% to only 52%.

Example 5: In a trial of a weight-loss program, 21 subjects on the Atkins program lost an average of 2.1 kg, or 4.6 lb, after one year. Suppose this loss is statistically significant.
Does this necessarily mean the result is practically significant?

Solution

No. A loss of 4.6 lb after a full year may be statistically significant, meaning it is unlikely to be due to chance alone. However, many people may feel that such a small weight loss is not worth the time, cost, and effort. Therefore, it may not be practically significant.

7. Common Pitfalls in Statistical Studies

Potential Pitfalls

Misleading conclusions
Conclusions should be clear and understandable, even to people who have not studied statistics.

Reported data
Data reported by subjects may be inaccurate. For example, people may report ideal weights rather than actual weights.

Loaded questions
Biased wording can influence responses.

Question order
Earlier questions can affect how people answer later questions.

Nonresponse Results may be biased when many selected subjects refuse to respond or cannot be reached.

Low response rates
A low response rate can make results less reliable.

Misleading percentages
Be careful with claims such as ``110% improvement.'' Percentages can be used in misleading ways.

Summary of Key Terms

Term	Meaning
Data	Collections of observations.
Population	The complete group or collection being considered.
Sample	A subcollection selected from the population.
Census	Data collected from every member of the population.
Voluntary response sample	A sample in which people choose whether to participate; often biased.
Statistical significance	A result that is unlikely to occur by random chance.
Practical significance	A result that is large enough or important enough to matter in real life.

Summary

Key Ideas

- Data are collections of observations.
- A population is the full group we want to study; a sample is a smaller group selected from the population.
- A statistical study usually follows the process: prepare, analyze, and conclude.
- Random samples are usually much better than voluntary response samples.
- Statistical significance means a result is unlikely to occur by chance.

- Practical significance means a result is important enough to matter in real life.
- Poor sampling methods, loaded questions, nonresponse, and misleading percentages can lead to bad conclusions.