Chapter 1 Introduction to Statistics
Collecting Sample Data (1.3)

Objectives

In this lesson, we will learn how to

- distinguish between observational studies and experiments;
- identify important features of good experimental design;
- describe common sampling methods;
- recognize different types of observational studies;
- identify common sampling errors and nonsampling errors.

1. Why Data Collection Matters

When analyzing sample data, it is essential to use an appropriate method for collecting the data.

Key Idea

Good statistical conclusions depend on good data collection. If sample data are collected in a poor way, the data may be so biased that calculations cannot fix the problem.

2. The Gold Standard

Randomness with placebo and treatment groups is sometimes called the gold standard because it is very effective in reducing bias.

Placebo

A placebo is a harmless and ineffective pill, medicine, or procedure used for comparison with a real treatment.

A placebo, such as a sugar pill, has no medicinal effect, but it can help researchers separate the real treatment effect from psychological effects.

3. Two Main Ways to Collect Data

Statistical data commonly come from two sources:

Observational Study

In an observational study, we observe and measure characteristics without trying to modify the individuals being studied.

Experiment

In an experiment, we apply some treatment and then observe its effects on the individuals.

In experiments, the individuals are called experimental units. When they are people, they are often called subjects.

Example 1: Classify each situation as an observational study or an experiment.

(a) Researchers record the number of hours students sleep and compare it with their exam scores.

(b) A group of patients is randomly assigned to receive either a new medicine or a placebo.

(c) A researcher surveys drivers about how often they text while driving.

Solution

(a) Observational study. The researchers only observe sleep hours and exam scores.

(b) Experiment. A treatment is assigned to some patients.

(c) Observational study. The researcher collects information but does not assign a treatment.

4. Ice Cream and Drownings: A Warning Example

Suppose past data show that ice cream sales and drownings both increase during the same months.

It would be wrong to conclude that ice cream causes drownings.

Lurking Variable

A lurking variable is a variable that is not directly studied, but may explain an observed relationship.

In this example, temperature is a lurking variable. When the weather gets hotter, ice cream sales increase and more people swim, so drownings may also increase.

Example 2: A study finds that people who carry lighters have a higher rate of lung cancer. Does this prove that carrying a lighter causes lung cancer?

Solution

No. This is likely an observational study with a lurking variable: smoking. People who smoke are more likely to carry lighters, and smoking is related to lung cancer.

5. Design of Experiments

A well-designed experiment often includes replication, blinding, and randomization.

Replication

Replication is the repetition of an experiment on more than one individual. Good replication requires a sample size large enough to reveal treatment effects.

Blinding

Blinding occurs when a subject does not know whether he or she is receiving the treatment or a placebo.

Blinding helps reduce the placebo effect, which occurs when an untreated subject reports improvement because they believe they received treatment.

Double-Blind Experiment

In a double-blind experiment, neither the subject nor the experimenter knows who receives the treatment and who receives the placebo.

Randomization

Randomization means that subjects are assigned to different groups using a random process. Randomization helps create groups that are similar.

6. Sampling Methods

Simple Random Sample

A simple random sample of size n is selected so that every possible sample of the same size n has the same chance of being chosen.

A simple random sample is often called a random sample. Strictly speaking, a random sample only requires that all members of the population have the same chance of being selected.

Systematic Sampling

Select a starting point, then select every kth element in the population.

Convenience Sampling

Use data that are very easy to get.

Convenience samples are easy, but they often produce biased results.

Stratified Sampling

Divide the population into at least two subgroups, called strata, so that subjects within the same subgroup share similar characteristics. Then draw a sample from each stratum.

Cluster Sampling

Divide the population into sections, called clusters. Randomly select some clusters, then choose all members from the selected clusters.

Multistage Sampling

Use a combination of the basic sampling methods in different stages.

Example 3: Identify the sampling method.

(a) A professor surveys the first 20 students who arrive to class.

(b) A researcher randomly selects 5 classrooms and surveys every student in those classrooms.

(c) A college separates students by class level (freshman, sophomore, junior, senior) and randomly selects students from each group.

(d) A factory inspector selects every 50th item from an assembly line.

Solution

(a) Convenience sampling. The students are easy to reach.

(b) Cluster sampling. Entire classrooms are selected.

(c) Stratified sampling. Students are grouped by class level, and samples are taken from each group.

(d) Systematic sampling. Every 50th item is selected.

7. Types of Observational Studies

Cross-Sectional Study

A cross-sectional study observes, measures, and collects data at one point in time.

Retrospective Study

A retrospective study, also called a case-control study, collects data from a past time period by examining records, interviews, or other past information.

Prospective Study

A prospective study, also called a longitudinal study or cohort study, collects data in the future from groups sharing common factors.

Example 4: Identify the type of observational study.

(a) Researchers review medical records from the past 10 years.

(b) A survey asks students today how many hours they studied last night.

(c) Researchers follow a group of adults for 20 years to study heart disease.

Solution

(a) Retrospective study. The data come from past records.

(b) Cross-sectional study. The data are collected at one point in time.

(c) Prospective study. The subjects are followed into the future.

8. Confounding

Confounding

Confounding occurs when we see an effect, but we cannot identify the specific factor that caused it.

Good experimental design tries to avoid confounding by carefully controlling variables and using randomization.

9. Controlling Effects of Variables

Completely Randomized Design

Assign subjects to different treatment groups through a process of random selection.

Randomized Block Design

A block is a group of subjects that are similar, but blocks differ in ways that might affect the outcome. Treatments are randomly assigned within each block.

Matched Pairs Design

Compare two treatment groups by using subjects matched in pairs that are related or similar.

Rigorously Controlled Design

Carefully assign subjects to treatment groups so that the groups are similar in important ways. This can be difficult because we may not know all relevant factors.

10. Sampling Errors

Even with careful planning, sample results usually contain some error.

Sampling Error

Sampling error, or random sampling error, occurs when a sample is selected randomly, but the sample result differs from the true population result because of chance variation.

Nonsampling Error

Nonsampling error is the result of human error, such as wrong data entry, biased wording, false responses, computing mistakes, or inappropriate statistical methods.

Nonrandom Sampling Error

Nonrandom sampling error results from using a sampling method that is not random, such as a convenience sample or voluntary response sample.

Example 5: Identify the type of error or issue.

(a) A survey question is worded in a way that encourages one particular answer.

(b) A researcher uses a random sample, but the sample mean is slightly different from the population mean.

(c) A website poll lets anyone choose whether to respond.

Solution

(a) Nonsampling error. The wording is biased.

(b) Sampling error. The sample was random, but chance variation

still occurs.

(c) Nonrandom sampling error. This is a voluntary response sample, not a random sample.

Summary

Key Ideas

- In an observational study, we observe and measure without applying a treatment.
- In an experiment, we apply a treatment and observe the result.
- Good experiments often use replication, blinding, and randomization.
- A simple random sample gives every possible sample of the same size the same chance of being selected.
- Common sampling methods include systematic, convenience, stratified, cluster, and multistage sampling.
- Observational studies may be cross-sectional, retrospective, or prospective.
- Confounding makes it difficult to identify the true cause of an effect.
- Sampling error is chance variation; nonsampling error usually comes from human or design problems.