Chapter 3
Producing Data

3.1 Sources of Data

- Anecdotal data
- Available data
- Sample surveys and experiments
- Observation vs. experiment
- Confounding

Obtaining Data

Beware of drawing conclusions from our own experience or hearsay. **Anecdotal data** represent individual cases that often come to our attention because they are striking in some way. We tend to remember these cases because they are unusual. **The fact that they are unusual means that they may not be representative of any larger group of cases.**

Available data are data that were produced in the past for some other purpose but that may help answer a present question inexpensively. The library and the Internet are sources of available data.

Some questions require data produced specifically to answer them. This leads to **designing** observational or experimental studies.

Sample Surveys

- Sample surveys are a special type of data collection that usually aim to discover the opinions of people on certain topics.

- In a sample survey, a **sample** of individuals is selected from a larger **population** of individuals.

- One can study a small part of the population in order to gain information about the population as a whole.

- Conclusions drawn from a sample are valid only when the sample is drawn in a well-defined way (to be discussed in Section 3.3).

Observation vs. Experiment

When our goal is to understand cause and effect, experiments are the *only* source of fully convincing data.

The distinction between observational study and experiment is one of the most important in statistics.

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses. The purpose is to describe some group or situation.

An **experiment** deliberately imposes some treatment on individuals to measure their responses. The purpose is to study whether the treatment causes a change in the response.

In contrast with observational studies, experiments do not just observe individuals or ask them questions. They actively impose some treatment in order to measure the response.

Confounding 1

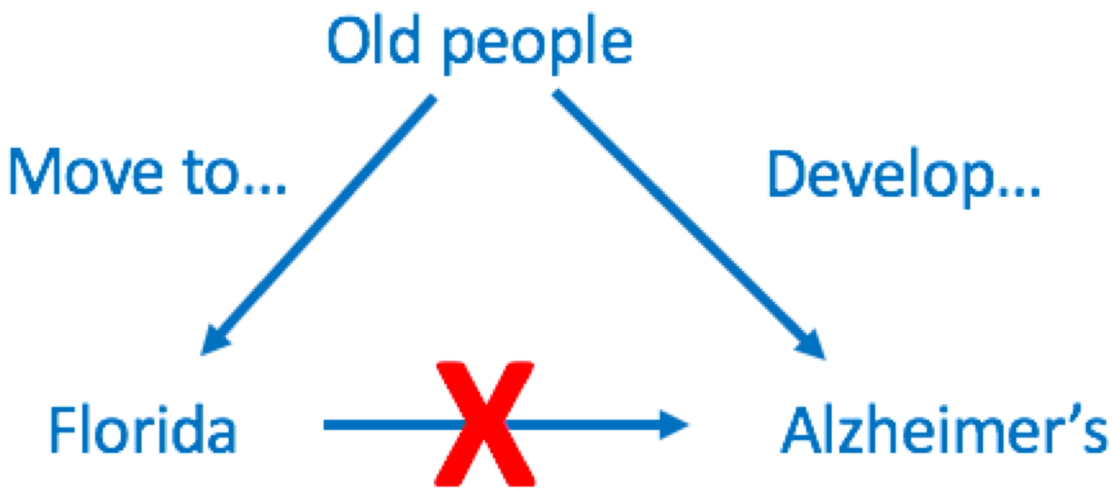
Observational studies of the effect of one variable on another often fail to establish cause and effect because of **confounding** between the explanatory variable and one or more **lurking variables**.

A **lurking variable** is a variable that is not among the explanatory or response variables in a study but that may influence the response variable.

Confounding occurs when two variables are associated in such a way that their effects on a response variable cannot be distinguished from each other.

Well-designed experiments take steps to avoid confounding.

Confounding 2



3.2 Design of Experiments

- Experimental units, subjects, treatments
- Comparative experiments
- Randomization
- Principles of experimental design
- Cautions about experimentation
- Matched pairs design
- Block design

Experimental Units, Subjects, Treatments

An experiment is a study in which we actually do something (a **treatment**) to people, animals, or objects (the **experimental units**) to observe the **response**. Here is the basic vocabulary of experiments.

An **experimental unit** is the smallest entity to which a treatment is applied. When the units are human beings, they are often called **subjects**.

The explanatory variables in an experiment are often called **factors**.

A specific condition applied to the individuals in an experiment is called a **treatment**. If an experiment has several explanatory variables, a treatment is a combination of specific values of these variables.

Comparative Experiments

The test anxiety experiment was poorly designed to evaluate the effect of the writing exercise. The lack of a **comparison** group is one problem with it. Another is the **placebo effect**. The end result is the study suffers from **bias**.

The **placebo effect** occurs when people respond favorably to personal attention or to any treatment that they hope will help them.

In a **comparative experiment**, subjects are assigned to one of two or more groups. In many studies, two such groups are the **control group** and the **treatment group**.

A study is **biased** if it systematically favors certain outcomes.

Need for Randomization

The remedy for confounding is to perform a **comparative experiment** in which some units receive one treatment and similar units receive another. Most well-designed experiments compare two or more treatments.

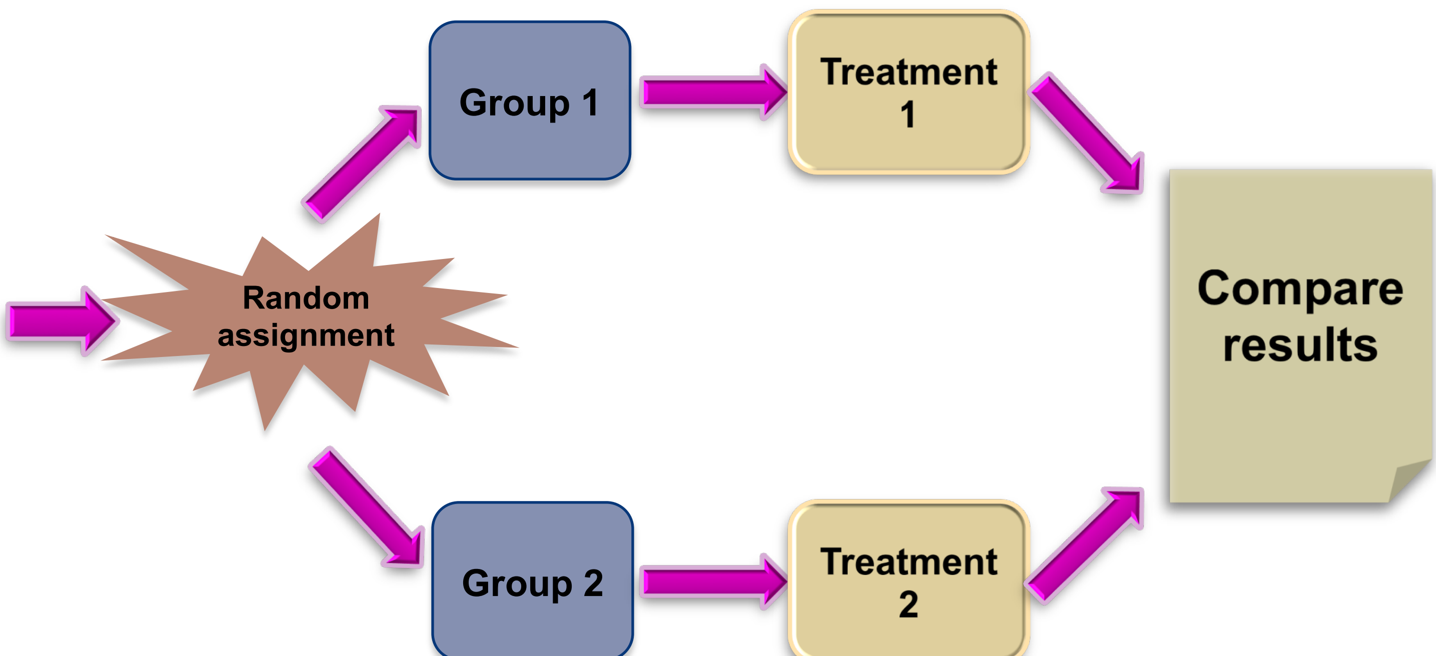
Comparison alone is not enough. If the treatments are given to groups that differ greatly, **bias** will result. The solution to the problem of bias is **random assignment**.

In an experiment, **random assignment** means that experimental units are assigned to treatments at random, that is, using some sort of chance process.

Randomized Comparative Experiments

- In a **completely randomized design**, the treatments are assigned to all the experimental units completely by chance.

- Some experiments may include a **control group** that receives an inactive treatment or an existing baseline treatment.



Principles of Experimental Design

Randomized comparative experiments are designed to give good evidence that differences in the treatments actually cause the differences we see in the responses.

Principles of Experimental Design

1. **Control** for lurking variables that might affect the response, most simply by comparing two or more treatments.
2. **Randomize:** Use chance to assign experimental units to treatments.
3. **Replication:** Use enough experimental units in each group to reduce chance variation in the results.

In principle, experiments can give good evidence for causation. We take the above steps to ensure that the only reasonable explanation for differences observed is due to the treatment imposed.

Cautions About Experimentation

The logic of a randomized comparative experiment depends on our ability to treat all the subjects in exactly the same way, except for the actual treatments being compared.

In a **double-blind experiment**, neither the subjects nor those who interact with them and measure the response variable know which treatment a subject received.

The most serious potential weakness of experiments is **lack of realism**. The subjects or treatments or setting of an experiment may not realistically duplicate the conditions we really want to study.

Nonetheless, the random comparative experiment, because of its ability to give convincing evidence for causation, is one of the most important ideas in statistics.

Matched Pairs

A common type of randomized *block* design for comparing two treatments is a matched pairs design. The idea is to create blocks by matching pairs of similar experimental units.

A **matched pairs design** is a randomized blocked experiment in which each block consists of a matching pair of similar experimental units.

Chance is used to determine which unit in each pair gets each treatment.

Sometimes, a "pair" in a matched pairs design consists of a single unit that receives both treatments. Because the order of the treatments can influence the response, chance is used to determine which treatment is applied first for each unit.

Blocked Designs

Completely randomized designs are the simplest statistical designs for experiments. But, just as with sampling, there are times when the simplest method does not yield the most precise results.

A **block** is a group of experimental units that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.

In a **block design**, the random assignment of experimental units to treatments is carried out separately within each block.

Form blocks based on the most important unavoidable sources of variability (lurking variables) among the experimental units.

Randomization will average out the effects of the remaining lurking variables and allow an unbiased comparison of the treatments.

Control what you can, block what you cannot control, and randomize to create comparable groups.

3.3 Sampling Design

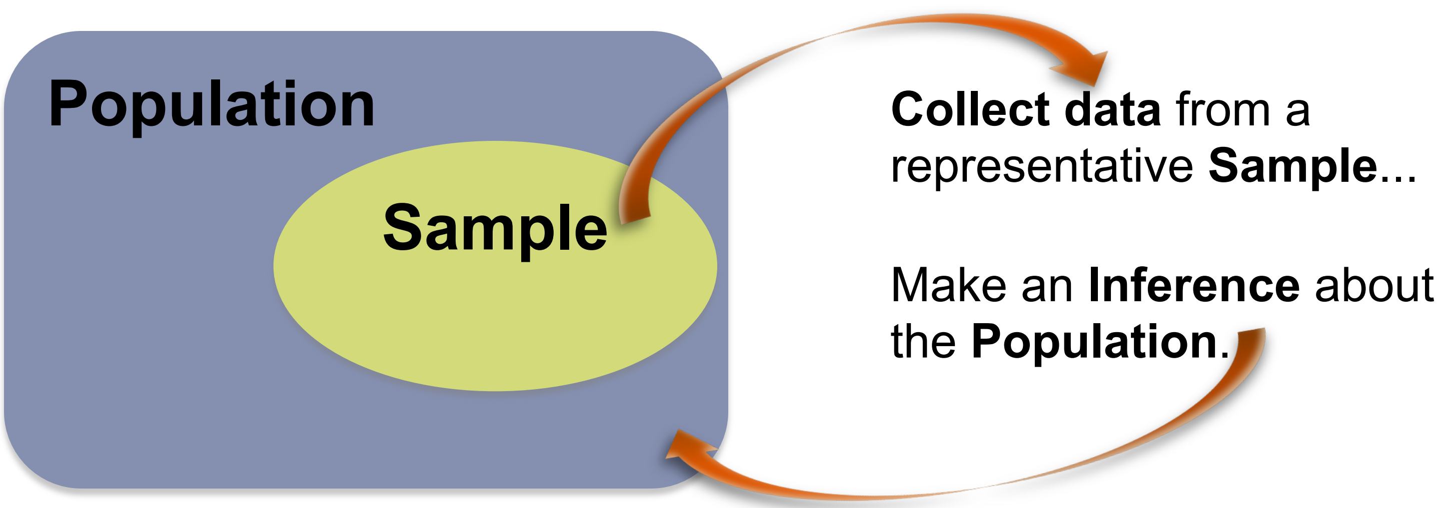
- Population and sample
- Simple random samples
- Probability samples
- Stratified random samples
- Cautions about sampling

Population and Sample

The distinction between population and sample is basic to statistics. To make sense of any sample result, you must know what population the sample represents.

The **population** in a statistical study is the entire group of individuals about which we want information.

A **sample** is the part of the population from which we actually collect information. We use information from a sample to draw conclusions about the entire population.



Simple Random Samples

Random sampling, the use of chance to select a sample, is the central principle of statistical sampling.

A **simple random sample (SRS)** of size *n* consists of *n* individuals from the population chosen in such a way that every set of *n* individuals has an equal chance to be the sample actually selected.

In practice, people use random numbers generated by a computer or calculator to choose samples. If you do not have technology handy, you can use a **table of random digits**.

Chapter 3
Producing Data (Review)

Introduction

3.1 Sources of Data

3.2 Design of Experiments

3.3 Sampling Design