Chapter 3

Producing Data

Introduction to the Practice of

STATISTICS EIGHTH EDITION

Moore / McCabe / Craig

Lecture Presentation Slides

Chapter 3
Producing Data

Introduction

3.1 Sources of Data

3.2 Design of Experiments

3.3 Sampling Design

3.4 Toward Statistical Inference

3.5 Ethics (On Your Own)

3.1 Sources of Data

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

Obtaining Data

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.

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.**

"The plural of anecdote is not evidence."

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 designed experiment 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 to observational studies, experiments don’t just observe individuals or ask them questions. They actively impose some treatment in order to measure the response.

Confounding

Observational studies of the effect of one variable on another often fail 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.

3.2 Design of Experiments

- Experimental units, subjects, treatments
- Comparative experiments
- Bias
- Principles of experimental design
- Statistical significance
- Matched pairs design
- Block design

Individuals, Factors, 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

Experiments are the preferred method for examining the effect of one variable on another. By imposing the specific treatment of interest and controlling other influences, we can pin down cause and effect. Good designs are essential for effective experiments, just as they are for sampling.

A high school regularly offers a review course to prepare students for the SAT. This year, budget cuts prevent the school from offering anything but an online version of the course.

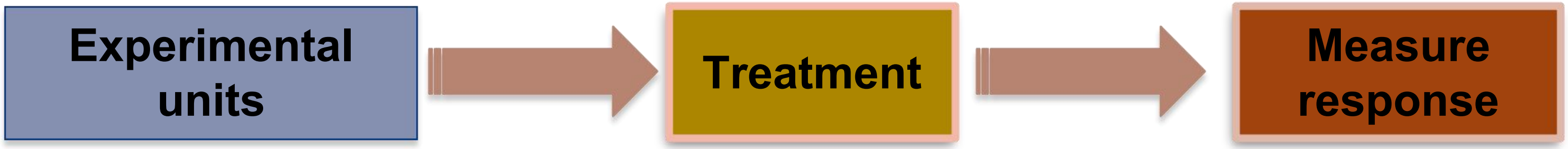
Students → Online Course → SAT Scores

Over the past 10 years, the average SAT score of students in the classroom course was 1620. The online group gets an average score of 1780. That’s roughly 10% higher than the long-time average for those who took the classroom review course.

Is the online course more effective?
How would you know?
Are you certain the increase is due to the online course?

Comparative Experiments

Many laboratory experiments operate as follows:



Experimental units

Treatment

Measure response

In the laboratory environment, simple designs often work well.

Field experiments and experiments with animals or people deal with more variable conditions.

Outside the laboratory, badly designed experiments often yield worthless results because of ***confounding*****.**

Randomized Comparative Experiments

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.

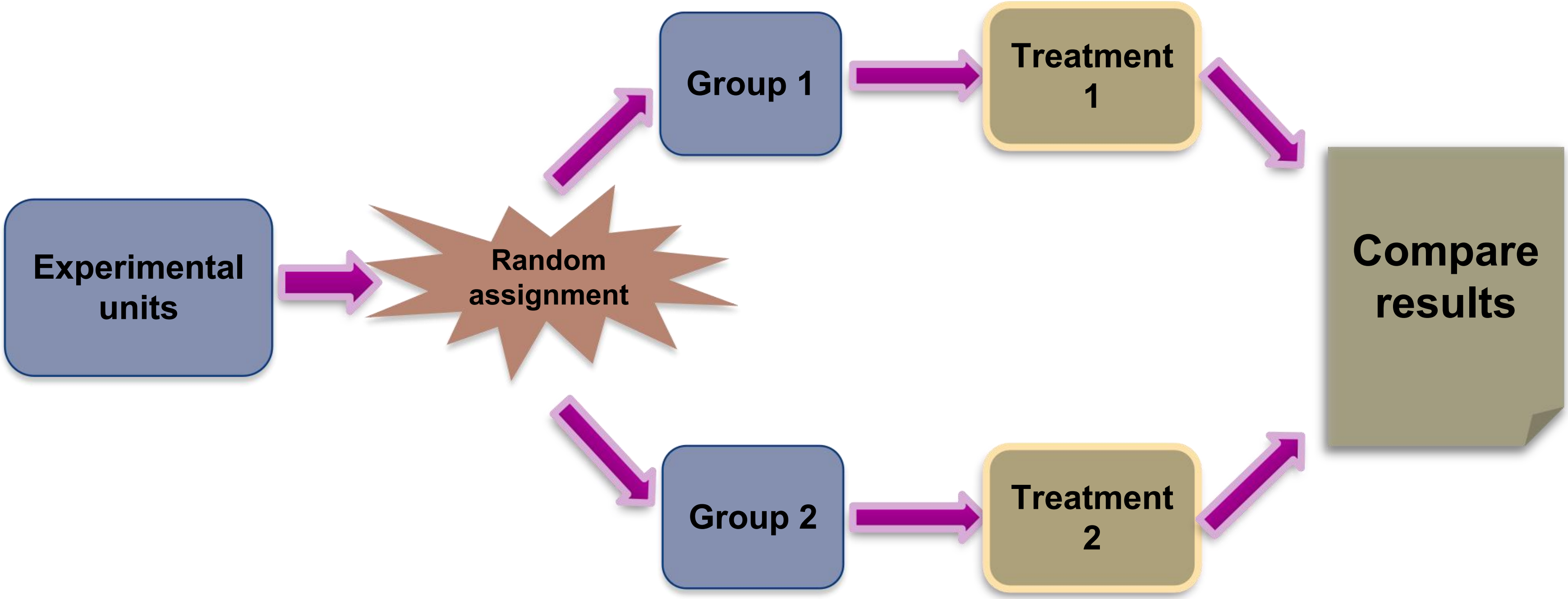
Comparison alone isn't enough. If the treatments are given to groups that differ greatly, **bias** will result. The solution to the problem of bias is **random assignment.**

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.



Randomization

One way to **randomize** an experiment is to rely on **random digits** to make choices in a neutral way. We can use a table of random digits (such as Table B) or the random sampling function provided by most statistical software.

How to randomly choose *n* individuals from a group of *N*:

- We first label each of the N individuals with a number (typically from 1 to N, or 0 to $N - 1$).
- Imagine writing the whole numbers from 1 to N on separate pieces of paper. Now put all the numbers in a hat.
- Mix up the numbers and randomly select one.
- Mix up the remaining $N - 1$ numbers and randomly select one of them.
- Continue in this way until we have our sample of n numbers. Statistical software can do this for you, so you don't actually need a hat!

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.

An observed effect so large that it would rarely occur by chance is called **statistically significant.**

A statistically significant association in data from a well-designed experiment does imply causation.

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.

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. Since 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 doesn’t 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 can’t control, and randomize to create comparable groups.

3.3 Sampling Design

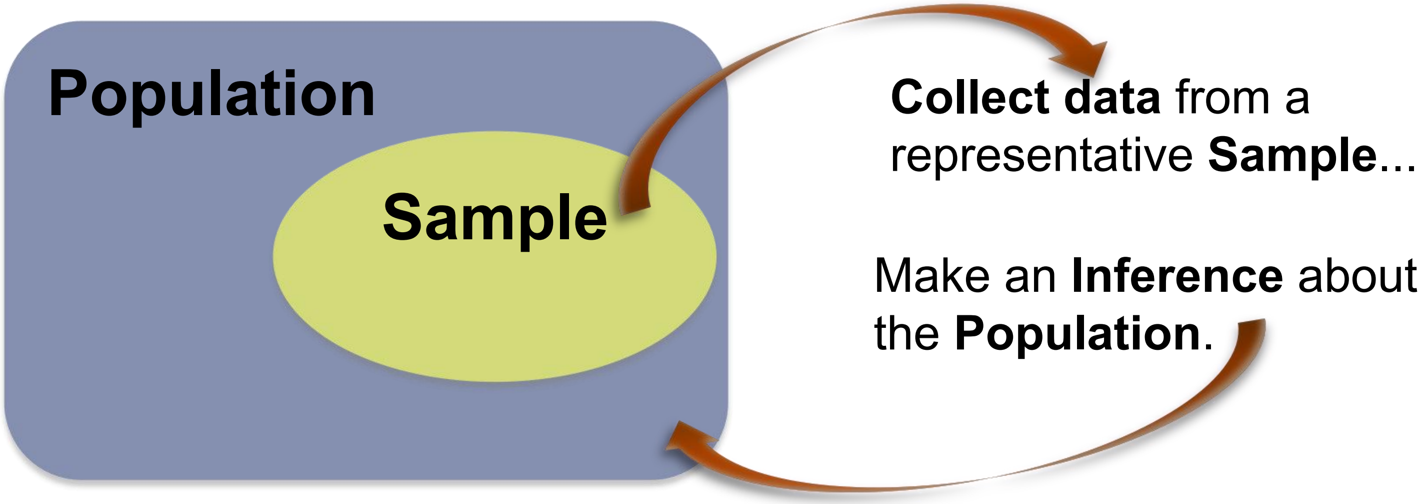
- Population and sample
- Voluntary response sample
- Simple random sample
- Stratified samples
- Undercoverage and nonresponse

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.



Population

Sample

Collect data from a representative **Sample**...

Make an **Inference** about the **Population**.

How to Sample Badly

The design of a sample is **biased** if it systematically favors certain outcomes.

Choosing individuals simply because they are easy to reach results in a **convenience sample.**

A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples often show bias because people with strong opinions (often in the same direction) may be more likely to respond.

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.

Other Sampling Designs

The basic idea of sampling is straightforward: Take an SRS from the population and use your sample results to gain information about the population.

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

Sometimes, there are statistical advantages to using more complex sampling methods. One common alternative to an SRS involves sampling important groups (called strata) within the population separately. These “sub-samples” are combined to form one stratified random sample.

To select a **stratified random sample,** first classify the population into groups of similar individuals, called **strata.** Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Cautions About Sample Surveys

Good sampling technique includes the art of reducing all sources of error.

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

Nonresponse occurs when an individual chosen for the sample can’t be contacted or refuses to participate.

A systematic pattern of incorrect responses in a sample survey leads to **response bias.**

The **wording of questions** is the most important influence on the answers given to a sample survey.

3.4 Toward Statistical Inference

- Parameters and statistics
- Sampling variability
- Sampling distribution
- Bias and variability
- Sampling from large populations

Parameters and Statistics

As we begin to use sample data to draw conclusions about a wider population, we must be clear about whether a number describes a sample or a population.

A **parameter** is a number that describes some characteristic of the population. In statistical practice, the value of a parameter is not known because we cannot examine the entire population.

A **statistic** is a number that describes some characteristic of a sample. The value of a statistic can be computed directly from the sample data. We often use a statistic to estimate an unknown parameter.

Remember s and p**:** statistics come from samples and parameters come from populations.

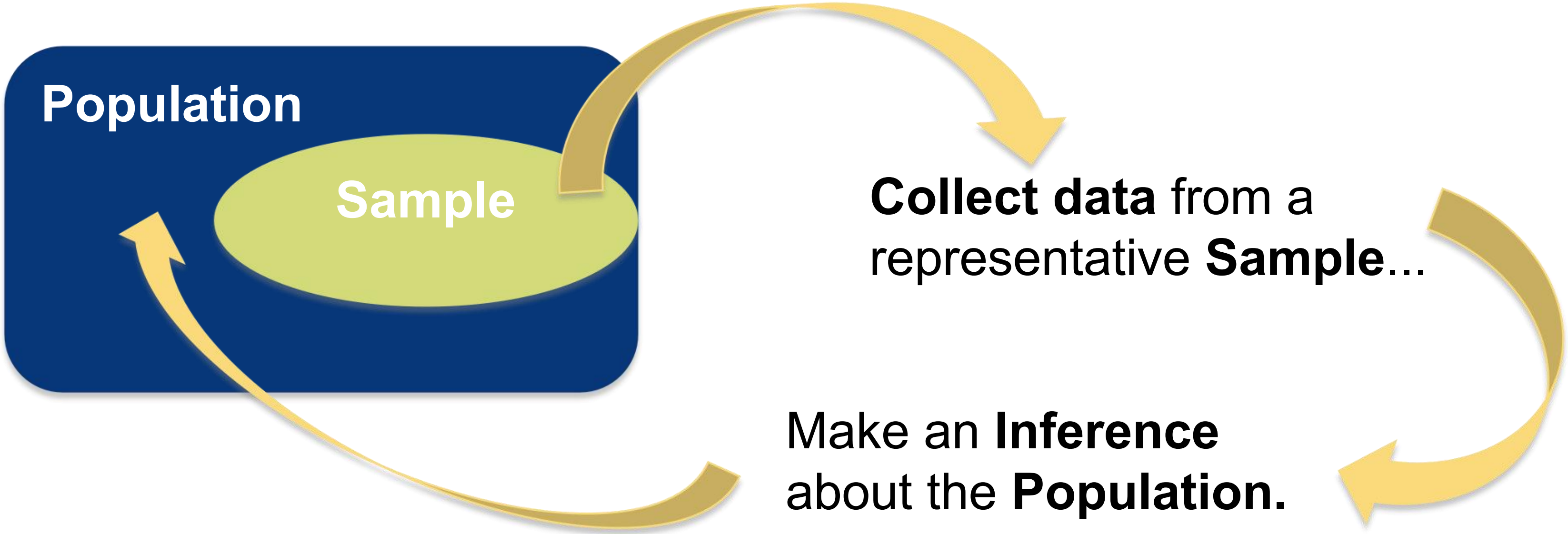
We write *μ* (the Greek letter mu) for the population mean and *σ* for the population standard deviation. We write $\overline{x}$ (x-bar) for the sample mean and *s* for the sample standard deviation.

Statistical Estimation

The process of **statistical inference** involves using information from a sample to draw conclusions about a wider population.

Different random samples yield different statistics. We need to be able to describe the **sampling distribution** of the possible values of a statistic in order to perform statistical inference.

The sampling distribution of a statistic consists of all possible values of the statistic and the relative frequency with which each value occurs. We may plot this distribution using a histogram, just as we plotted a histogram to display the distribution of data in Chapter 1.



Population

Sample

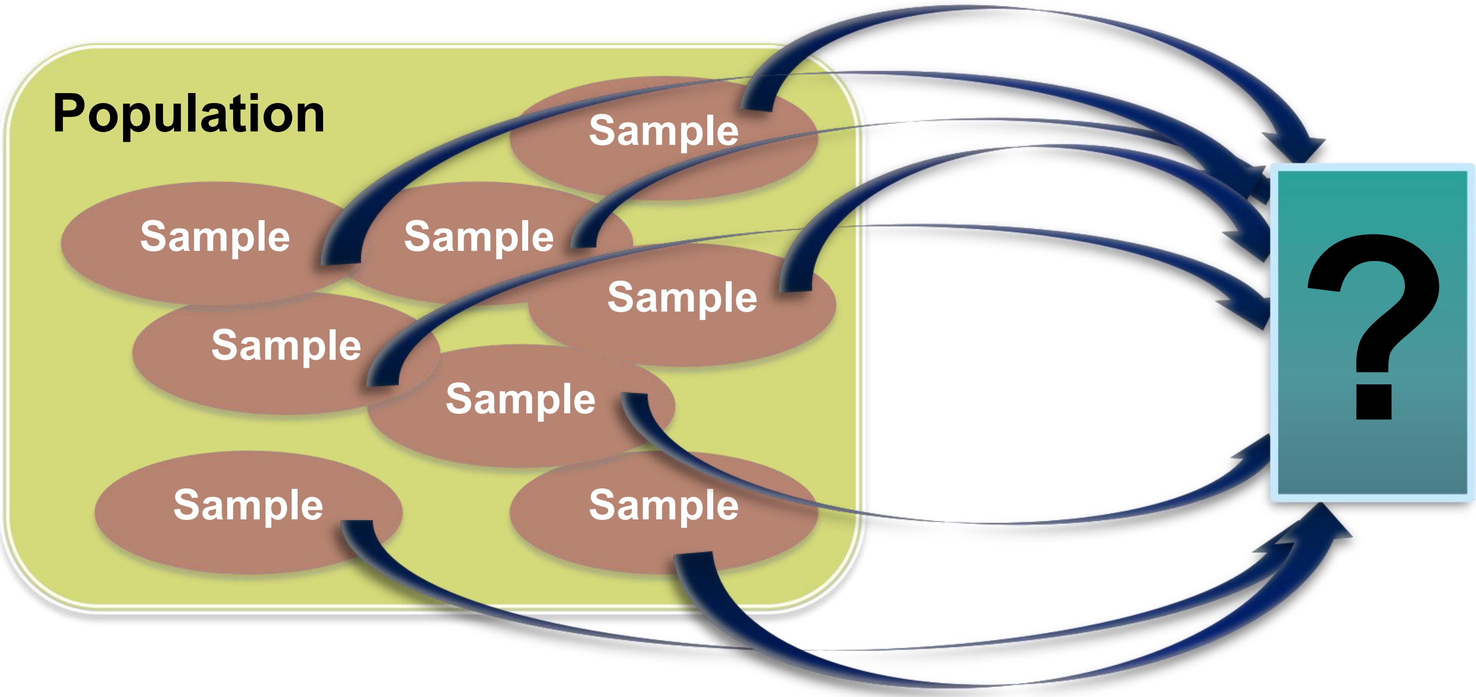
Collect data from a representative **Sample**...

Make an **Inference** about the **Population.**

Sampling Variability

Sampling variability is a term used for the fact that the value of a statistic varies in repeated random sampling.

To make sense of sampling variability, we ask, "What would happen if we took many samples?"



Sampling Distributions

If we measure enough subjects, the statistic will be very close to the unknown parameter that it is estimating.

If we took every one of the possible samples of a certain size, calculated the sample mean for each, and made a histogram of all of those values, we'd have a **sampling distribution.**

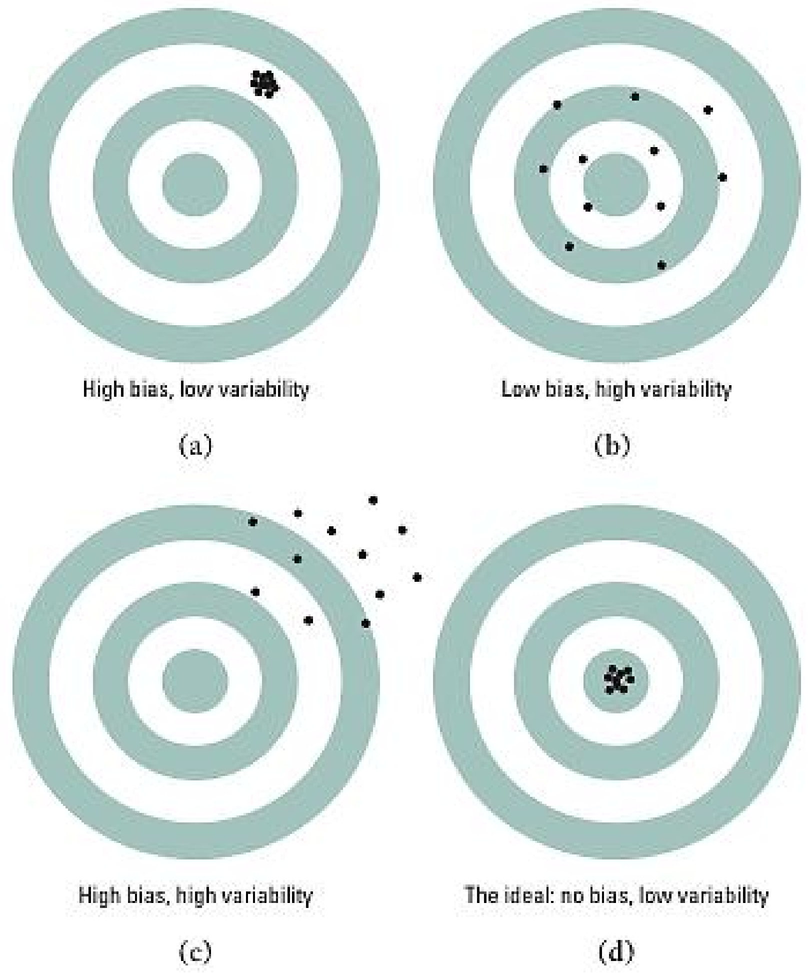
The **population distribution** of a variable is the distribution of values of the variable among all individuals in the population.

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

In practice, it's difficult to take all possible samples of size *n* to obtain the actual sampling distribution of a statistic. Instead, we can use **simulation** to imitate the process of taking many, many samples.

Bias and Variability

We can think of the true value of the population parameter as the bull's-eye on a target and of the sample statistic as an arrow fired at the target. Bias and variability describe what happens when we take many shots at the target.


High bias, low variability
(a)
Low bias, high variability
(b)
High bias, high variability
(c)
The ideal: no bias, low variability
(d)

Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size *n*. Statistics from larger probability samples have smaller spreads.

Managing Bias and Variability

A good sampling scheme must have both small bias and small variability.

To reduce bias, use random sampling.

To reduce variability of a statistic from an SRS, use a larger sample.

The variability of a statistic from a random sample does not depend on the size of the population, as long as the population is at least 100 times larger than the sample.

Why Randomize?

The purpose of a sample is to give us information about a larger population. The process of drawing conclusions about a population on the basis of sample data is called **inference.**

Why should we rely on random sampling?

1. To eliminate bias in selecting samples from the list of available individuals.

2. The laws of probability allow trustworthy inference about the population.
    - Results from random samples come with a **margin of error** that sets bounds on the size of the likely error.
    - Larger random samples give better information about the population than smaller samples.

Chapter 3
Producing Data

Introduction

3.1 Sources of Data

3.2 Design of Experiments

3.3 Sampling Design

3.4 Toward Statistical Inference

3.5 Ethics (On Your Own)