Chapter 3

Producing Data

Types of data collected

- Anecdotal data – data collected haphazardly (not representative!!)
- Available data – existing data (examples: internet, library, census bureau,….)
- Gather own data (takes money and time to get own data)

Some terminology

- Population – the entire group of individuals or objects of interest (answers the question: Who?)
- Sample – subset of the population on which information is obtained.
- Census-sample is the entire population
- Variable – characteristics of interest

Observational study vs Experiment

- Observational study – A study that observes individuals and measures variables of interest but does not attempt to influence the response.
- Experiment – A study that imposes some treatment on individuals in order to record their response.

Types of variables

- Response variable – the outcome of the study.
- Explanatory variable – variable(s) that attempt to explain the changes in the response

Examples: Smoking and lung cancer

Running on a treadmill and heart rate

Classroom Examples

One study of cell phones and the risk of brain cancer looked at a group of 469 people who have brain cancer. The investigators matched each cancer patient with a person of the same sex, age, and race who did not have brain cancer, then asked about use of cell phones. Result: "Our data suggest that use of handheld cellular telephones is not associated with the risk of brain cancer." Is this an observational study or experiment? Why? What are the explanatory and response variables?

A typical hour of prime-time television shows 3-5 violent acts. Linking family interviews and police records shows a clear association between time spent watching TV as a child and later aggressive behavior. Is this an observational study or experiment? What are the explanatory and response variables? Suggest some lurking variables that could explain the aggressive behavior.

An educational software company wants to compare the effectiveness of its computer animation for teaching cell biology with that of a textbook presentation. The company tests the biological knowledge of each group of first year college students, then randomly divides them into two groups. One group uses the animation, and the other studies the text. The company retests all the students and compares the increase in understanding of cell biology in the two groups. Is this an observational study or experiment? What are the explanatory and response variables?

3.1 Design of Experiments

- Experimental units – individual on which experiment is done.
- Treatment – specific experimental condition
- Factors = explanatory variables
- Placebo – false treatment to control for psychological effects. Example: Gastric freezing is a clever treatment for ulcers in the upper intestine. The patient swallows a deflated balloon with tubes attached, then a refrigerated liquid is pumped through the balloon for an hour (cooling will reduce production of acid and relieve ulcers). An experiment reported in the Journal of the American Medical Association showed that gastric freezing did reduce acid production and relieve ulcer pain. Later experiment included a control group (34% of the treatment group improved…..38% of the placebo group improved).
- Joint effects – combination of levels of two or more factors. Example: A maker of fabric for clothing is setting up a new line to finish the raw fabric. The line will use either metal rollers or natural-bristle rollers to raise the surface of the fabric; a dyeing cycle time of either 30 minutes or 40 minutes and a temperature of either 150 or 175 degrees Celsius. Four specimens of fabric will be subjected to each treatment and scored for quality. What are the factors and the treatments? How many units (fabric specimens) does the experiment require?

Experiments continued

- Experiments provide good evidence for causation (able to control lurking variables)
- Confounded variables – variable(s) associated with the response, but are not of interest; effects cannot be separated from the effect of the explanatory explanatory variable on the response Bias – systematically favors certain outcomes.

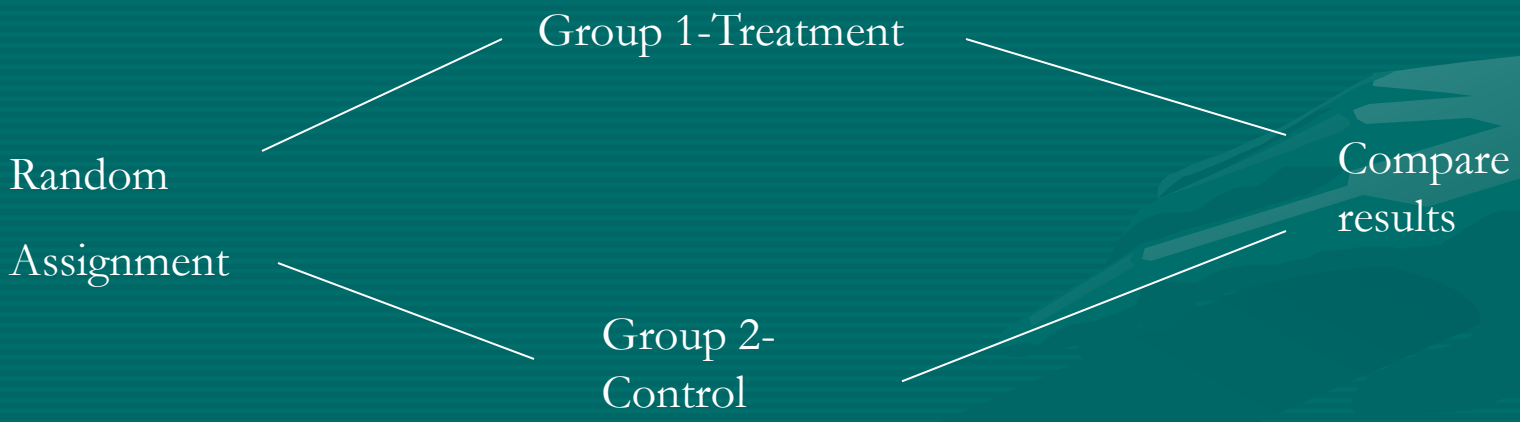
Experiments continued

- Randomization is very important in experiments…helps to ensure groups are as similar as possible.
- The three principles of Experimental Design are
  - Control
  - Randomize
  - Repeat
- How can we randomize? Draw names out of a hat, use table of random digits, computer software (calculator), phone-random digit dialing

Using R to randomize

- First, you need to set the seed
- > set.seed(put seed number in here)
- Then sample
- >sample(seq(1:n),sample size,replace=FALSE)
- Assign class to two groups

Completely Randomized Design (with one treatment group and one control group)



More on Experiments

- Single blind – individual receiving treatment does not know what treatment they are receiving.
- Double blind – individual getting treatment and individual recording outcome do not know which treatment was administered.

Block design

- One way to control for confounding variables is to block on them.
- A block design first breaks the experimental units into blocks according to the "blocking variable" (for example, if one is blocking on gender, first place units into female and male "blocks").

Example of Block Design

- The progress of a type of cancer differs in women and men. A clinical experiment to compare three therapies for this cancer therefore treats sex as a blocking variable. Two separate randomizations are done, one assigning the female subjects to the treatment and the other assigning the male subjects. Draw this design.

Matched Pairs Design

- A special type of block design is called Matched pairs design.
- Can only compare two treatments (hence the "pairs").
- Block usually consists of units as similar as possible (self, twins, husband and wife).

Example of Matched Pairs design

- Does talking on a hands-free cell phone distract drivers? Undergraduate students "drove" in a high-fidelity driving simulator equipped with hands-free cell phone. Each student drove once while talking on the cell phone and once without talking on the cell phone. The order for each student was randomly assigned. The car ahead breaks: how quickly does the subject respond?

3.2 Sampling Design

- Voluntary response sample (call-in polls, comment cards) are very biased…bad sampling design.
- Want to get a probability sample. A probability sample is a sample chosen by chance (will look at four of them in this course).

Different Types of Probability samples

- SRS (Simple Random Sample) – every sample of size n has the same chance of being selected.
- Stratified random sample – first divide into groups (strata), and then take a SRS from each stratum.
- Cluster sample – first divide into clusters, and then take a SRS of clusters (once a cluster is chosen, every unit in that cluster is in the sample).

Probability samples continued

- Multistage sampling design –at each stage, a probability sample is obtained.
- Problems with sample surveys
  - Undercoverage
  - Nonresponse
  - Response bias

Towards statistical inference

- Use information from sample (known information) to infer about the population (unknown)
- Statistics – information from a sample
- Parameter – information from a population
- Sampling variability – information from a sample will differ from one sample to the next.
- Sample statistics will have a predictable pattern (referred to as sampling distribution)

Bias and variability

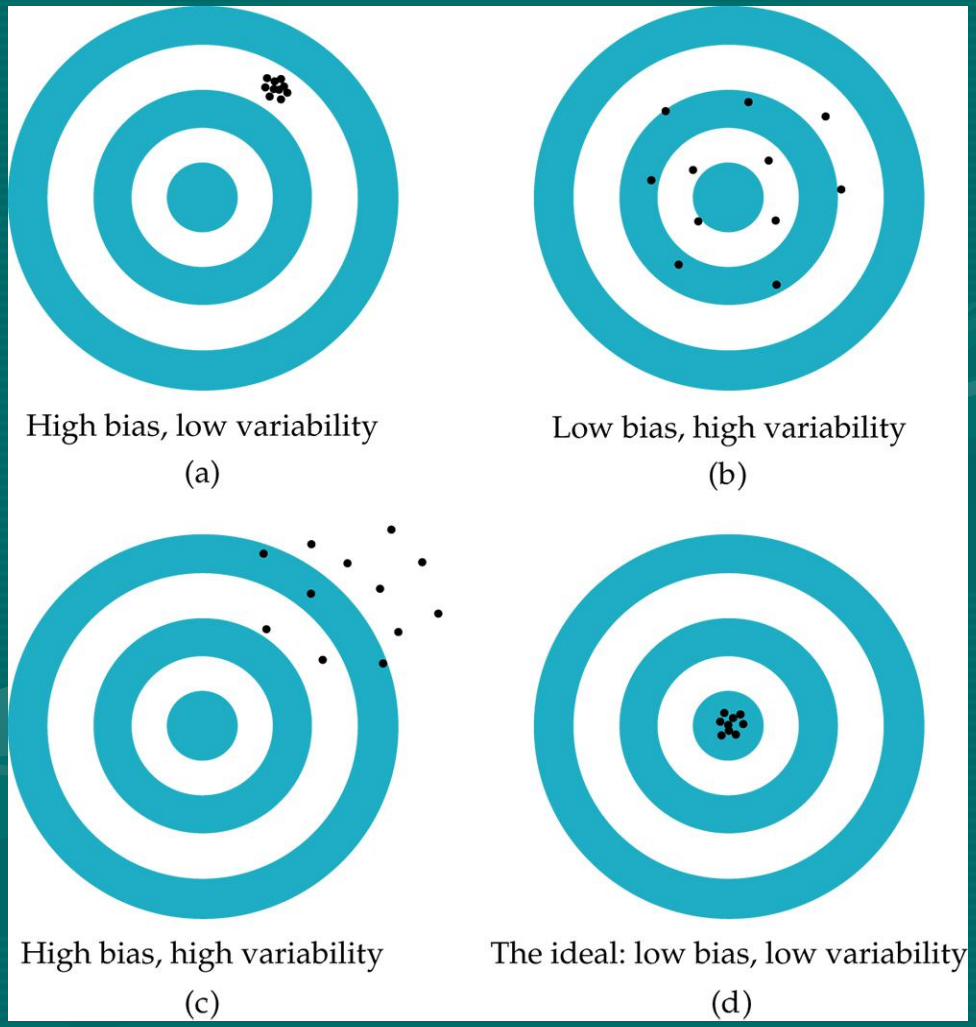


Figure 3.14

Introduction to the Practice of Statistics, Sixth Edition

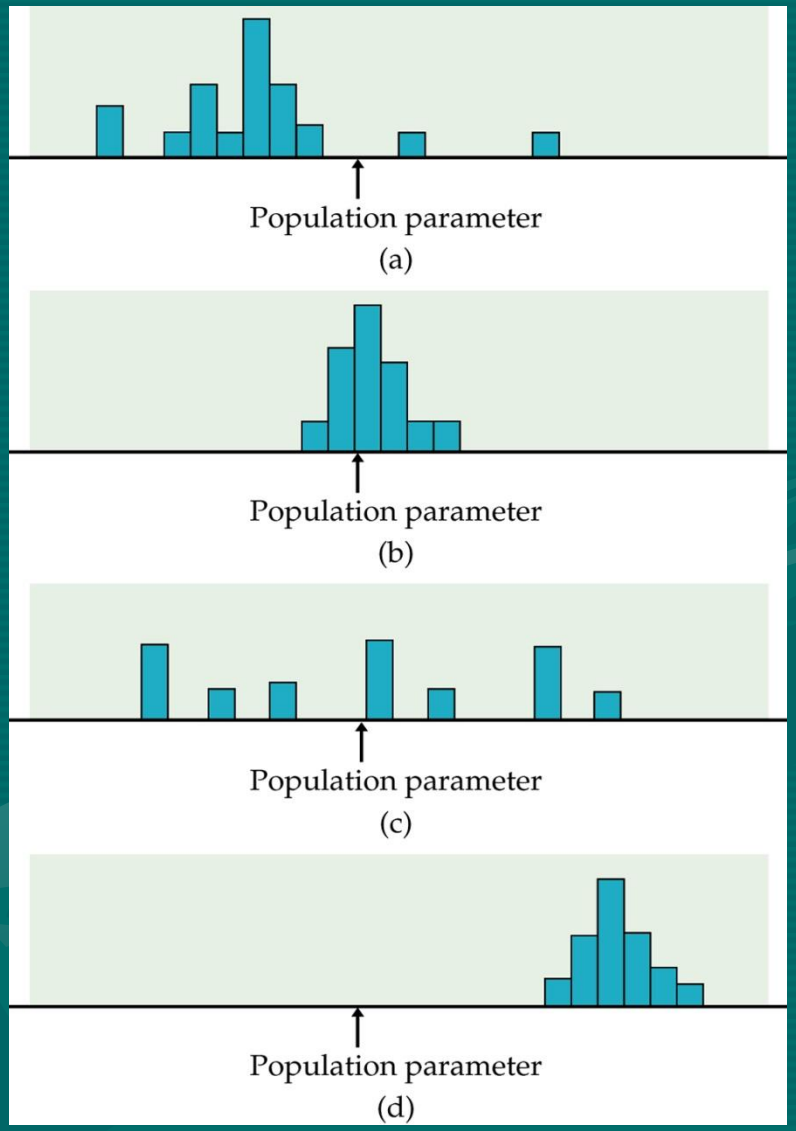


Figure 3.15

Introduction to the Practice of Statistics, Sixth Edition

3.4 Continued

- Want statistics that are unbiased and have low variability.
- How can we eliminate or at least reduce the bias? Use a random sample and good instruments.
- How to increase precision? Larger sample
- Population size does not effect precision!!! Sample size does.

Statistical Significance

STATISTICAL SIGNIFICANCE

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

Definition, pg 184

Introduction to the Practice of Statistics, Sixth Edition

Ethics

BASIC DATA ETHICS

The organization that carries out the study must have an **institutional review board** that reviews all planned studies in advance in order to protect the subjects from possible harm.

All individuals who are subjects in a study must give their **informed consent** before data are collected.

All individual data must be kept **confidential.** Only statistical summaries for groups of subjects may be made public.

Definition, pg 225