Sampling

The **population** is the entire group of individuals about which we want information.
A **sample** is a part of the population from which we actually collect information.
A **sampling design** describes how to choose a sample from the population.

Bad samples

A **convenience sample** often produces unrepresentative data.

A **voluntary response sample** consists of people who choose to respond. Voluntary response samples are often biased.

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.

e.g., suppose the population consists of 5 students A, B, C, D, E, and we want to choose a sample of size n=2. There are 10 possible samples (AB, AC, AD, AE, BC, BD, BE, CD, CE, DE). SRS means each of these 10 possible samples has an equal chance of being chosen.

SRS does not favor any part of the population.

Larger samples provide better information about the population.
Read example 1.2, 1.3, 1.4.

Example 1.2, 1.3

Example 1.2: A professor wants to take a sample of size 100 from 20,000 students. She obtains a list of 20,000 students, numbered from 1 to 20,000. She uses a computer random number generator to generate 100 random integers between 1 and 20,000 and selected 100 students corresponding to these numbers. Is this an SRS?

Yes, because every possible group of 100 students have an equal chance of being chosen.

Example 1.3: Now this professor wants to draw a sample of 50 students to take a survey. Her class has 50 students. She let the students in her class fill out the survey. Is this an SRS?

No. In this study, only the students in her class has a change of being chosen.

Stratified sampling, clustering sampling,systematic sampling

- **Stratified sampling**: population is divided into groups (strata). A SRS is drawn from each stratum.
- **Cluster sampling:** Population is divided into groups (clusters). Take all individuals within those selected clusters.
- **Systematic sampling:** select every *k*th item after a random starting place.

Read Figure 1.1 to 1.4 on the next slide.

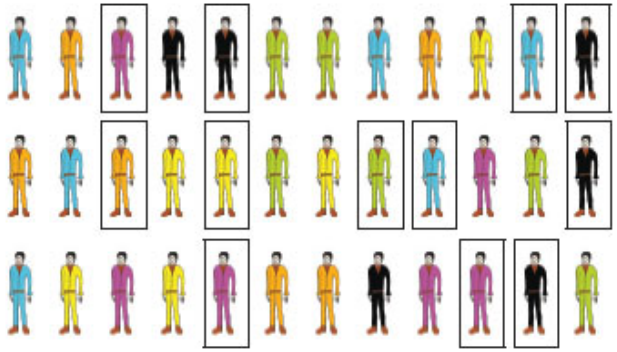

Figure 1.1 Simple random sampling

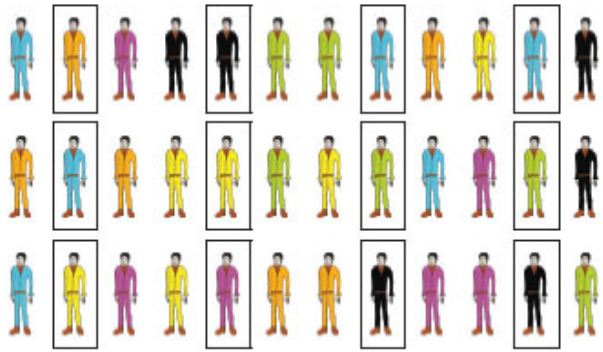

Figure 1.2 Systematic sampling

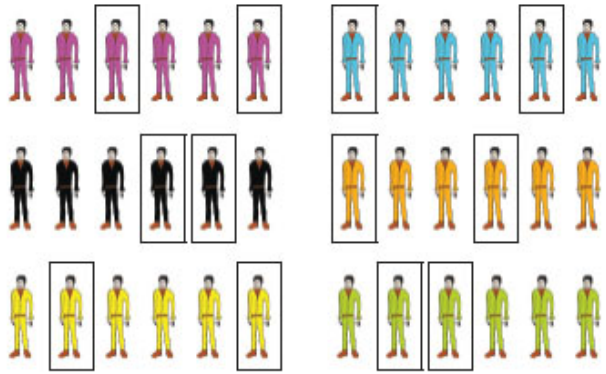

Figure 1.3 Stratified sampling

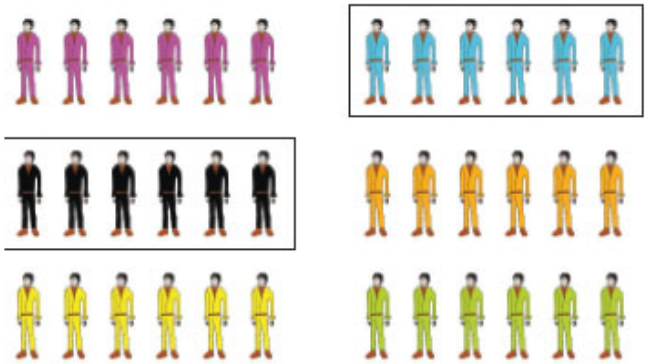

Figure 1.4 Cluster sampling

Exercises on page 7

3. A radio talk show hosts invites listeners to send an email to express their opinions on an election. More than 10,000 emails are received. What kind of sample is this? Voluntary response.

4. Every 10 years, the US Census Bureau attempts to count every person living in US. To check the accuracy in a certain city, they draw a sample of census districts and recount everyone in the sampled districts. Wha kind of sample is formed by the people who are recounted? Cluster sample

5. A researcher is studying the effect of diet on heart disease. To be sure both men and women are well represented, the study comprises of a random sample of 100 men and another random sample of 100 women. What kind of sample do these 200 people represent? Stratified sample

6. A college basketball team held a promotion in which every 20th person who entered the area won a free basketball. What kind of sample do the winners represent? Systematic sample

Statistics and parameters

A **statistic** is a number that describes a sample. e.g., a sample mean is a statistics.
A **parameter** is a number that describes a population. e.g, a population mean is a parameter.

Types of data

individuals: objects described by a data set.
variables: the characteristics about individuals we collect information. e.g., for a data set on a sample of students, age, GPA, major, hometown, height are variables.
qualitative or categorical variables classify individuals into categories. e.g., gender, marital status, degree of infection.

- Ordinal variables: Categories have a natural ordering. e.g., feeling of happiness (very happy, happy, unhappy, very unhappy).
- Nominal variables: Categories have no natural ordering. e.g., hair color (black, brown, white, etc)

quantitative variables tell how much or how many of something there is. e.g, age, GPA, height.

- discrete variables: take values that are countable. e.g., number of siblings a students has, (0, 1, 2 ,....)
- continuous variable: take on any values in some interval. e.g., height, weight, time, volume...

Exercises on page 15

Fill in blanks:
1. The characteristics of individuals about which we collect information are called (variables).
2. Variables that classify individuals into categories are called (qualitative or categorical).
3. (Quantitative) variables are always numerical.
4. Qualitative variables can be divided into two types: (nominal) and (ordinal).
5. A (discrete) variable is a quantitative variables whose possible values can be listed.
6. (Continuous) variables can take on any value in some interval.

Experiments and observational study

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the response or outcome.

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

Only study 3 is an experiment on the next slide.

How safe are cell phones?

Study whether heavy cell phone use causes cancer

- Study 1: A German study compared 118 patients with eye cancer to 475 healthy people. The subjects' cell phone use was measured using a questionnaire. The eye cancer patients used cell phone more often, on average.
- Study 2: A US study compared 469 patients with brain cancer to 422 patients who did not have brain cancer. The two groups' use of cell phone was similar.
- Study 3: An Australian study conducted an experiment with 200 transgenic mice, with 100 exposed to the same kind of radiation of a cell phone for 2.5 hours a day. The other 100 mice were not exposed. After 18 months, the brain tumor rate for the exposed group were twice as high as the rate for the unexposed group.

Vocabulary of experiments

The **individuals** studied in an experiment are often called **subjects**, particularly when they are people.

The **outcome, or response** is what is measured on each experimental unit.

A **treatment** is any specific experimental condition applied to the subjects.
e.g, to study the effect of 3 types of seed on wheat yield, wheat yield is response, the 3 types of seed are treatments.

Advantage of experiment

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

In observational study, the effect of the treatments is **confounded** with the effects of other **lurking variables**.

Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other.

Principles of an experiment

- Include a **control group** for comparison. Why need a control group? Imagine 100 flu patients took a drug and got well one week later. Can we conclude the drug work? Not quite, as patients may get well after a week anyway. If we add a control group, say gave a placebo to another 100 patients, and only 50 got well after one week. Since both groups went through the time effect, we can attribute the difference in the outcome to the difference in the treatment. i.e., we think the drug worked.
- The subjects assigned to any treatment should be **randomly** selected from the available subjects. The purpose of randomization is to make sure patients are similar on other variables that may impact the outcome.
- In a **double-blind** experiment, neither the subjects nor the people who interact with them know which treatment each subject is receiving. Double-blind in clinical trials take care of psychological effect and potential bias in assessment

Cohort study and Case-control study

A **cohort** is a group of people who are linked in some way.

- Prospective cohort study: subjects are followed over time.
- cross-sectional study: measurements are taken at one point in time.
- retrospective cohort study: subjects are sampled after the outcome has occurred.

Case-control study: useful for studying rare disease. Take a sample of people with a certain disease (cases) and a sample of people witout the diseases (controls) and examine whether they differ on a factor of interest.
Study 1 and 2 on cell phone use are examples of case-control study.

Read page 23, 24 for more details (the two pages are in Files folder in Canvas).

Cautions about bias in study

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.
Random Digit Dialing in national surveys tends to lead to undercoverage as housholds without a telephone can never be selected.
Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to participate.
self-interest bias: drug companies pay physicians to test their drugs and the report written by the physicians may inflate the effect of the drug.
Voluntary response bias: people who respond may not represent the general population well.
Response bias: respondents do not tell the truth.
Wording of question influence the answers given by respondents.