DSA 8010 Week 0: Collecting data

Introduction to statistics

"Statistics is the science of designing studies or experiments, collecting data, and modeling/analyzing data for the purpose of decision making and scientific discovery when the available information is both limited and variable."

From *An Introduction to Statistical Methods and Data Analysis, Ott & Longnecker*

Basic vocabulary.

- population
- sample
- census
- parameter
- statistic

Basic vocabulary.

- population: the set of all measurements of interest to the sample collector.
- sample: any subset of measurements selected from the population.
- census: a study in which measurements are collected from the entire population.
- parameter

- statistic

Basic vocabulary.

- population: the set of all measurements of interest to the sample collector.
- sample: any subset of measurements selected from the population.
- census: a study in which measurements are collected from the entire population.
- parameter: numerical characteristic of a population. Denoted using Greek letters (e.g. μ =population mean.)
- statistic: numerical characteristic of a sample. Typically denoted using Latin letters (e.g. $\bar{y}$ = sample mean.)

Example (mile times).

A wheelchair basketball coach wants to know if his athletes have come into season in shape or not. He times each of his 26 athletes over a one mile distance and finds that 73% of the athletes manage to break 9 minutes.

- What is the population?
- Was a sample or census taken?
- Is 73% a parameter or statistic?

Example (trout length)

A wildlife biologist wants to know the average length of all adult trout in Western North Carolina. He goes to a Western North Carolina stream and catches 18 adult trout. He calculates that the average length of the 18 fish is 56cm.

- What is the population?
- Was a sample or census taken?
- Is 56 cm a parameter or statistic?

Types of statistical analyses

- Descriptive statistics. Methods of organizing, summarizing, and presenting sample data in an informative way.
- Inferential statistics. A decision, estimate, prediction, or generalization about a population, based on a sample.

Types of studies

Observational study. The researcher does not interfere with observational units.
Examples: polls, surveys, cohort studies.

Experimental study. The researcher actively manipulates certain variables associated with the study.
Examples: clinical trials.

Types of studies

Differences between observational and experimental studies:

- In an experiment the researcher assigns treatments. In an observational study the researcher does not.
- In an observational study, you may be able to conclude (if the data indicates it) that two variables are *associated*.
- In experiment, you may be able to conclude (if the data indicates it) that there is a *cause/effect relationship* between the two variables.

Types of studies: examples

Differences between observational and experimental studies:

- In an experiment the researcher assigns treatments. In an observational study the researcher does not.
- In an observational study, you may be able to conclude (if the data indicates it) that two variables are *associated*.
- In experiment, you may be able to conclude (if the data indicates it) that there is a *cause/effect relationship* between the two variables.

Principles of sampling

Sampling. The process of selecting a subset of the population from which to collect data.

We want samples to be representative of the population.

- target population
- sampled population

Goal in developing a sampling scheme: ensure that every individual of interest has a chance to be selected for the sample.

target population = sampled population

Example (trout length)

A wildlife biologist wants to know the average length of all adult trout in Western North Carolina. He goes to a Western North Carolina stream and catches 18 adult trout. He calculates that the average length of the 18 fish is 56cm.

- What is the target population?
- What is the sampled population?

Sampling schemes

Good sampling will always include random selection of individuals. Here are a few popular sampling schemes.

simple random sample (SRS). Each individual in the population has an equal chance of being sampled.

stratified random sample. The population is divided into *strata* of similar individuals. A simple random sample is selected from each stratum.

cluster sample. The population is divided into clusters of (preferably diverse) individuals. The researcher selects clusters at random and samples every individual in the selected clusters.

systematic sample. Used when the population is enumerated in a list. Randomly select one item near the top of the list, then select every rth item thereafter.

Sampling schemes

What about volunteer and convenience samples?

- volunteer (self-selected) sample
- convenience sample

These are not true sampling schemes and do not represent any well-defined population.

Types of sampling

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience)

1. A pollster interviews all human resource personnel in five different high tech companies.
2. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
3. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
4. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
5. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Example: Clemson students

Suppose an administrator wants to study the financial well-being of Clemson students using a survey. How could a SRS be taken?

Example: Clemson students

The administrator thinks the financial well-being of Clemson students might differ substantially across undergraduate, master's, and PhD students. How could a stratified random sample be taken to ensure each group is represented in the sample?

Example: Clemson students

Suppose an administrator wants to study the financial well-being of Clemson students using a survey. He stands outside of the library and hands a survey to the first 50 students who pass by. What are some portions of the population that will not be represented?

Example: Youth Risk Behavior Surveillance System

`https://www.cdc.gov/yrbs/methods/index.html`

by sex subgroups meet this standard. Estimates for grade by race/ethnicity subgroups are accurate within ±5% at a 90% confidence level.

The first-stage sampling frame for each national survey includes primary sampling units (PSUs) consisting of large-sized counties or groups of smaller, adjacent counties. Since the 1999 sample, PSUs large enough to be selected with certainty are divided into sub-PSU units. Schools then are sorted by size and assigned in rotation to the newly created sub-PSU units. PSUs are selected from 16 strata categorized according to the metropolitan statistical area† (MSA) status and the percentages of black and Hispanic students in PSUs. PSUs are classified as urban if they are in one of the 54 largest MSAs in the United States; otherwise, they are considered rural. PSUs are selected with probability proportional to school enrollment size for PSUs.

In the second stage of sampling, schools are selected from PSUs. A list of public and private schools in PSUs is obtained from the Market Data Retrieval (MDR) database (*29*). This database includes information, including enrollment figures, from both public and private schools and the most recent data from the Common Core of Data from the National Center for Education Statistics (*27*). Schools with all four high school grades (9–12) are considered "whole schools." Schools with any other set of grades are considered "fragment schools" and are combined with other schools (whole or fragment) to form a "cluster school" that includes all four grades. The cluster school is treated as a single school during school selection. Schools are divided into two groups on the basis of enrollment. Schools with an estimated enrollment of ≥25 students for each grade are considered large, and schools with an estimated enrollment of <25 students for any grade are considered small. Approximately one fourth of PSUs are selected for small-school sampling. For each of these PSUs, one small school is drawn with probability proportional to size, considering only small schools within that PSU. Three large schools then are selected from all sampled PSUs with probability proportional to school enrollment size.

To enable a separate analysis of data for black and Hispanic students, CDC has used three strategies to achieve oversampling of these students: 1) larger sampling rates are used to select PSUs that are in high-black and high-Hispanic strata; 2) a modified measure of size is used that increases the probability of selecting schools that have a disproportionately high minority enrollment; and 3) two classes per grade, rather than one, are selected in schools with a high minority enrollment. All of these strategies were used in selecting the national samples through 2011. Because of decreases in the percentage of white students in the U.S. population (*30*), for the 2013 sample, sufficient numbers of black and Hispanic students were sampled using only the third strategy.

The final stage of sampling consists of randomly selecting one or two entire classes in each chosen school and in each of grades 9–12. Examples of classes include homerooms or classes of a required subject (e.g., English and social studies). All students in sampled classes are eligible to participate. Since 1991, the national YRBS has been conducted 11 times with an average sample size of 14,517 and average school, student, and overall response rates of 78%, 86%, and 71%, respectively (Table 3).

A weight based on student sex, race/ethnicity, and school grade is applied to each record to adjust for student nonresponse and oversampling of black and Hispanic students. To avoid inflated sampling variances, statisticians trim and distribute weights exceeding a criterion value among untrimmed weights using an iterative process (*31*). The final overall weights are scaled so that the weighted count of students equals the total sample size and the weighted proportions of students in each grade match national population projections for each survey

Bias in observational studies.

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

Some potential sources of bias in observational studies include

undercoverage. Some members of the target population are systematically excluded from the sample.

nonresponse. Some sampled individuals fail to respond or participate.

response bias. Responses are inaccurate due to, for example, survey questions that are difficult or sensitive, or biased measurement instruments.

Examples: bias

In the following examples, identify potential sources of bias. Predict whether the quantity of interest is likely to be over-estimated, under-estimated, or unbiased.

- The goal is to evaluate whether American voters are in favor of stricter immigration policy. A link to an online poll with several questions on the policy is posted on a Reddit US Politics page.

Examples: bias

In the following examples, identify potential sources of bias. Predict whether the quantity of interest is likely to be over-estimated, under-estimated, or unbiased.

- The goal is to estimate prevalence of violent experiences in South Carolina high schoolers' dating relationships. A survey asks students from four randomly-selected S.C. high schools if they have ever experienced violence in a dating relationship.

Examples: bias

In the following examples, identify potential sources of bias. Predict whether the quantity of interest is likely to be over-estimated, under-estimated, or unbiased.

- The goal is to estimate average household size in a school district, where a "household" is defined as people living together in the same dwelling and sharing living accommodations. Students are selected at random at an elementary school and asked what their family size is.

Examples: bias

- The goal is to estimate the proportion of defective products in a given shipment of 10,000 gizmos. The manager takes a simple random sample of 10 gizmos from the shipment and counts the proportion of defective products.

Examples: bias

- The goal is to estimate the proportion of defective products in a given shipment of 10,000 gizmos. The manager takes a simple random sample of 10 gizmos from the shipment and counts the proportion of defective products.

Small sample sizes do not cause bias, but *imprecision* in measurement.

Imprecision occurs when there is high variability in measurements from the sample, relative to sample size.

Bias and imprecision

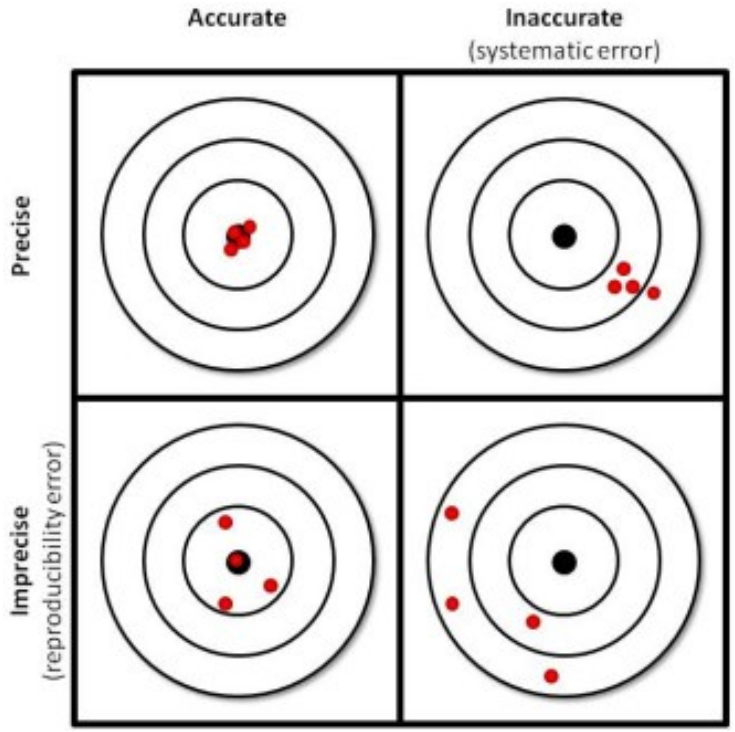

Image: `http://glsi.agron.iastate.edu/2015/01/18/accuracy-vs-precision/`

Experimental studies

Basic vocabulary:

- experimental units - individuals (people, animals, plants, etc.) to which treatments are applied.
- factor - a variable whose effect on the response we are trying to measure.
- treatment - a factor level or combination of factor levels applied to an experimental unit.
- response variable - outcome that is measured. The goal is to evaluate the effect of the treatments on the response variable.

Completely randomized design

A completely randomized experimental design assigns treatments randomly to each experimental unit such that each unit is equally likely to receive each treatment.

Example (light levels)

A study is designed to test the effect of light level on exam performance of students. The researcher randomly assigns students to complete identical exams in the following conditions: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). He records each student's exam score.

What are the experimental units, response variable, factors, and treatments?

Why was it important to use identical exams?

Why was it important to randomly assign students to conditions?

Example (light levels)

A study is designed to test the effect of light level on exam performance of students. The researcher is interested in the following lighting conditions: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). He is also interested in the effect of paper color (white vs light green).

What are the experimental units, response variable, factors, and treatments?

Experimental studies

In an experimental study, different treatments are applied to a set of experimental units. Here are a few principles upon which experiments are designed:

- control - experimental units should be a similar as possible with respect to additional factors that affect the response.
- randomization - treatments should be assigned randomly to experimental units.
- replication - whenever possible, each treatment should be applied to several experimental units. Entire experiments may be replicated.
- blocking - grouping experimental units by a additional variable that affects the response before randomizing.

Example (light levels)

A study is designed to test the effect of light level on exam performance of students. The researcher believes that effects of light levels might differ between students in the 9am section and the 3pm section, so he wants to make sure both are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

Example (light levels)

A study is designed to test the effect of light level on exam performance of students. The researcher believes that effects of light levels might differ between students in the 9am section and the 3pm section, so he wants to make sure both are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

Section is a blocking variable in this study.

How could this experiment be designed?

Example (light levels)

Randomized complete block design

A randomized complete block design is an experimental design for comparing t treatments in b blocks. Treatments are randomly assigned to (homogeneous) experimental units with a block, with each treatment appearing exactly once in each block.

Experimental studies

Some additional terms:

- placebo
- blinding
- double-blind