Lecture 1
Introduction and Syllabus
STAT 310, Spring 2021

Outline of Course Topics

This course will have four main components:

- Data collection: sampling designs and experimental studies
- Exploratory data analysis: numerical summaries and graphical displays of data
- Statistical inference: hypothesis testing and confidence intervals
- Linear regression and correlation

Throughout the semester we will focus on examples and data sets from the social sciences when discussing these concepts.

Grading

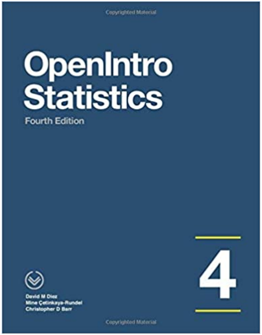
- 10% Homework
- 15% Computer Labs
- 75% Three Exams (take-home, 25% each)

Late assignments will generally not be accepted. However, your lowest scoring homework and lab assignment will be dropped.

Textbook

Diez, D., Barr, C. and Cetinkaya-Rundel M. *OpenIntro: Statistics*, 4th Edition, 2019.

Free PDF version posted on Blackboard in the "Resources" folder. You can also purchase a paperback copy on Amazon for about $20.



Software

This class will also provide training in the statistical software R. You will learn how to use R to compute numerical summaries of data, visualize data, and implement statistical procedures.

Why learn R?

- It is a **free** and open-source software that runs on most operating systems (Windows, Mac).
- It is one of the most **popular** programming languages used for statistics and data science.
- It is a desired skill for many **jobs** in academia, industry, and government.
- It is a legitimate programming language that allows more control and **reproducibility** than a point-and-click interface.

Preliminaries

The **field of statistics** is broadly concerned with collecting, analyzing and interpreting data for the purpose of decision making and scientific discovery.

A statistical investigation often follows these steps:

1. Define the problem, and formulate research questions
2. Design the sampling procedure or experiment for collecting the data
3. Explore and analyze the data
4. Formulate conclusions and communicate the results

Preliminaries

More generally, we can view the steps of a statistical investigation as an iterative process, represented by the "Data Cycle" shown below.

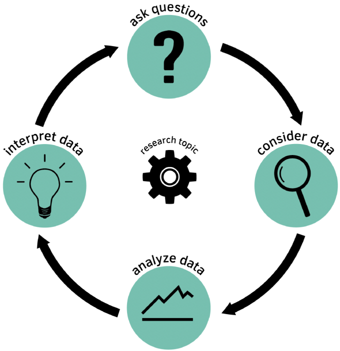


Image from Robert Gould et al. (2016). "Teaching data science to secondary students: The mobilize introduction to data science curriculum."
http://iase-web.org/documents/papers/rt2016/Gould.pdf

Introducing Data

- Data tables
- Observations and variables
- Types of variables
- Relationships between variables

Data Tables

- Statisticians usually prepare data as tables, where the columns are the variables and the rows are the individual cases or **observations**.
- A **variable** can be thought of as a characteristic of an observation.

Data Tables

Here is an example of a data set from a survey conducted by the National Center for Health Statistics. The rows are 1500 participants in the survey. The columns are variables on demographic characteristics of the participants. The data table contains 1500 rows by 4 columns by .

	Gender	Age	Education	Income[1]
1	male	40	College Grad	60.00
2	female	37	College Grad	87.50
3	male	57	High School	60.00
4	female	54	College Grad	100.00
5	female	73	High School	12.50
6	male	55	Some College	17.50
7	female	80	9 - 11th Grade	100.00
8	female	23	Some College	22.50
9	male	51	9 - 11th Grade	87.50
10	male	33	College Grad	50.00
⋮	⋮	⋮	⋮	⋮
1499	male	24	College Grad	17.50
1500	female	20	High School	70.00

[1] in thousands of US dollars

Variable Types

- **Numerical variables** take on numerical values and that are usually measurements or counts. It makes sense to take the sum or mean of a numerical variable.
  - For example, `Age` and `Income` are numerical variables.

- **Categorical variables** take on values that fall into distinct categories.
  - For example, `Gender` and `Education` are categorical variables.

categorical

nominal
(gender)

ordinal
(education)

Relationships Between Variables

- In a statistical study, we can also label a variable as being either a **response** or **explanatory** variable.

- For example, we might be interested in how `Education` and `Age` affect a person's `Income`. In this case, `Income` would be the response variable, and `Education` and `Age` would be the explanatory variables.

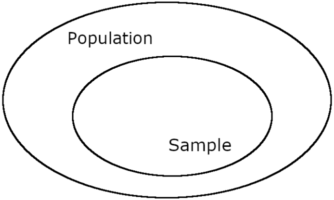
explanatory variable $\xrightarrow{\text{might affect}}$ response variable

Sampling Concepts

- Samples and populations
- Parameters and statistics

Sampling Concepts

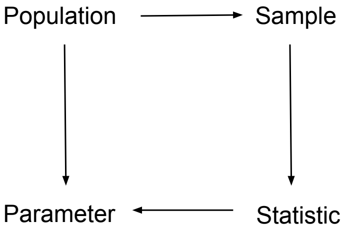
- A **population** is the set of all individuals or cases of interest to the investigator.
- A **sample** is any subset selected from the population.



Example:
Population: All students that attend CSUEB.
Sample: 100 randomly selected CSUEB students.

Sampling Concepts

- A **parameter** is a numerical characteristic of the population (fixed and usually unkown).
- A **statistic** is a numerical characteristic of the sample (varies depending on sample).



Sampling Concepts

Example:
Parameter: Average height of all students that attend CSUEB. This is a fixed number, but probably unknown since we might not have the time or resources to measure every student.

Statistic: Average height of 100 randomly selected CSUEB students. Each sample will contain different individuals, and therefore yield a different value for the average height.

Sampling Concepts

Some notation for common parameters and statistics:

	parameter	statistic
mean	μ	$\bar{x}$
proportion	p	$\hat{p}$
standard deviation	σ	s

Sampling Concepts

Example: A Gallup poll[2], conducted between July 20 - Aug 2, 2020, found that 86% of Americans wear masks when outside of the home in an indoor setting. The results were based on web surveys of a random sample of 7,632 adults, aged 18 and older. For this survey, describe the sample, population, statistic, and parameter.

Solution:

- Sample: 7,632 adults that participated in the web surveys
- Population: all adults in the United States
- Statistic: 86% of Americans in the sample wear masks when outside of the home in an indoor setting
- Parameter: population proportion of all Americans that wear masks when outside of the home in an indoor setting

[2] https://news.gallup.com/poll/316928/face-mask-usage-relatively-uncommon-outdoor-settings.aspx

Two primary types of data collection:

- Observational studies
- Experiments

Observational Studies

- **Observational study**: researcher collects data without interfering in the process that generates that data. That is, data are gathered by monitoring what has occurred or by using historical records.
- Observational studies can be used to show **associations** between variables of interest, but generally do not support cause-and-effect relationships.
- *Example:* The National Center for Health Statistics (NCHS) conducts surveys to collect data on demographics and the health-status of adults and children in the United States. Each year a nationally representative sample of about 5,000 people are interviewed. Participants are asked questions about diet, exercise, smoking and drinking habits, employment, and physical and mental health.[3]

[3] The data presented on slide 11 comes from an NCHS survey.

Experimental Studies

- **Experimental study**: researcher actively manipulates the explanatory variables and records their effect on the response variable.
- For a **randomized experiment** individuals are randomly assigned to different treatment groups and the outcomes for the response variable are compared.
- Experiments can be used to infer **cause-and-effect** relationships between the response and explanatory variables.

Experimental Studies: Example

- Knee osteoarthritis (OA) is a common problem in the elderly population that results in pain and reduced quality of life.
- Researchers at Tufts Medical Center conducted a randomized experiment to evaluate the effectiveness of Tai Chi in treating OA symptoms.[4]

[4]Wang et al. (2009). "Tai Chi is Effective in Treating Knee Osteoarthritis: A Randomized Controlled Trial." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3023169/

Experimental Studies: Example

- 40 patients with OA, aged 55 years and older, with no prior experience with Tai Chi, were recruited.
- 20 patients were randomly assigned to 60-minute Tai Chi sessions twice-weekly for 12 weeks (**treatment group**). The other 20 patients were assigned to 60 minute sessions on wellness education and stretching twice-weekly for 12 weeks (**control group**).
- At the end of the 12 weeks, the patients in the Tai Chi group reported a significant decrease in knee pain when compared to the control group.

40 patients $\xrightarrow{\text{randomized}}$ 20 Tai Chi / 20 Control $\rightarrow$ Compare pain scores after 12 weeks

Experimental Studies: Example

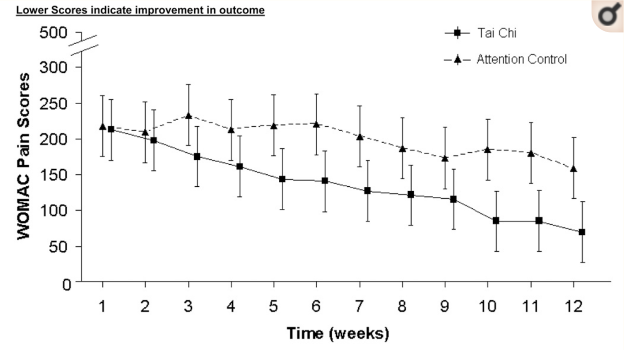


Figure 2

WOMAC Pain Subscale over a 12 week Intervention Period by Treatment Group

Values shown are unadjusted means. Measurements were obtained weekly over a 12 week period, Error bars indicate the 95% Confidence Interval (CI) but the data are slightly offset in the figure for clarity. Means with 95% CI shown at each line for each group. Linear treads between weeks indicated by connected graph. WOMAC= Western Ontario and McMaster Universities Osteoarthritis index.

Confounding Variables

One reason it is difficult to establish causal relationships in observational studies is because of confounding variables. A **confounding variable** is a variable that is associated with both the response and explanatory variables, but is not accounted for by the researcher.

Confounding Variables

Example: A researcher conducts an observational study by recording the number of hours of TV a sample of students watch and their GPAs. The researcher finds that the more TV students watch, the lower their GPAs, on average. Does this necessarily mean that watching TV *causes* students to have a lower GPA? What are some examples of confounding variables?

Solution:
Since this is an observational study we cannot make cause-and-effect conclusions. One confounding variable is the number of hours the student spends studying.