Lecture 1
Introduction and Syllabus
STAT 630, Fall 2020

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

We will also review some probability. Most of you are enrolled in STAT 620, and will be learning probability theory concurrently.

Grading

There will be weekly homework assignments, and three take-home exams. Both the homework and exams will be a combination of conceptual and data analysis problems. The data analysis problems will require the use of R. Late homework will generally not be accepted. However, your lowest **two** scoring homework assignments will be dropped.

- 40% Homework
- 60% Three Exams (20% each)

Textbook

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

Free PDF version posted on Blackboard in the “Resources” folder.

This will be the main textbook for the course. It provides a very accessible and concise introduction to statistics. It is written at the undergraduate level, but is also popular in graduate courses. It also includes an R package with many data sets that we will use during lab.

Textbook

Chihara, L. and Hesterberg T. *Mathematical Statistics with Resampling and R*. 2nd Edition, 2018.

Free electronic version: `http://library.csueastbay.edu/home`

This textbook provides an introduction to mathematical statistics with an emphasis on computational methods for doing statistical inference. It is written at the advanced undergraduate or first-year graduate school level. We will reference this book when covering resampling techniques such as the bootstrap. The book also contains some mathematical derivations not covered in OpenIntro.

Labs and Software

Weekly computer labs will focus on learning the R programming language and using it for statistical data analysis. Topic we will cover in lab include:

- Vectors and data frames
- Subsetting data frames
- Looping and control structures
- Summarizing data and creating graphics
- Loading data files into R
- Simulation and resampling techniques
- Report writing and reproducible research (R Markdown)

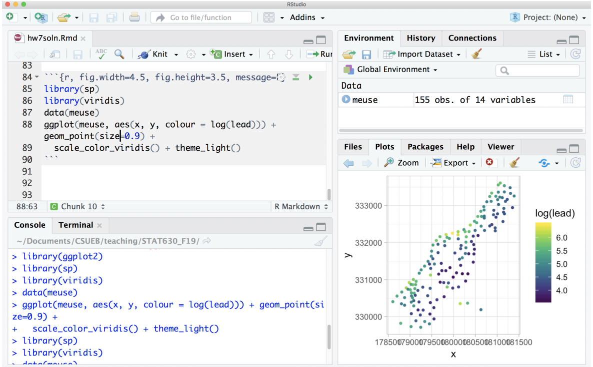
We will often replace and compare mathematical (analytic) techniques for doing statistics with more flexible and intuitive computational approaches.

Why learn R?

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

RStudio

- RStudio is a convenient interface for using the R programming language.
- RStudio provides a code editor, console, and tools for plotting, debugging, and version control.



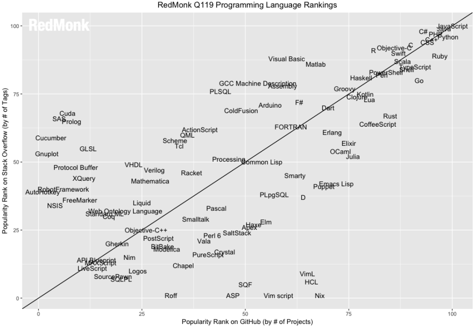
~~~
```{r, fig.width=4.5, fig.height=3.5, message=
library(sp)
library(viridis)
data(meuse)
ggplot(meuse, aes(x, y, colour = log(lead))) +
geom_point(size=0.9) +
  scale_color_viridis() + theme_light()
```
~~~

```
> library(ggplot2)
> library(sp)
> library(viridis)
> data(meuse)
> ggplot(meuse, aes(x, y, colour = log(lead))) + geom_point(si
ze=0.9) +
+   scale_color_viridis() + theme_light()
> library(sp)
> library(viridis)
> data(meuse)
```

Data

meuse   155 obs. of 14 variables

Redmonk rankings of different programming languages based on GitHub repositories (developer activity) and Stack Overflow activity (user discussion forums).



Source: `https://redmonk.com/sogrady/2019/03/20/language-rankings-1-19/`

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 generally 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 summarize the data
4. Analyze the data and make inferences
5. 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.

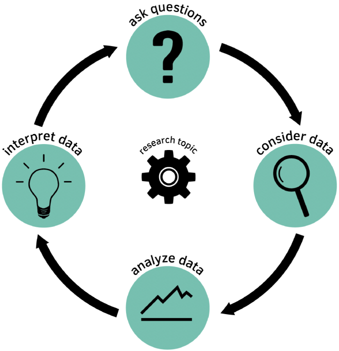


Image from Robert Gould et al. (2016). "Teaching data science to secondary students: The mobilize introduction to data science curriculum."
http://iase-web.org/documents/papers/rt2016/Gould.pdf

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 in R called `mtcars` that was originally from the 1974 *Motor Trend* magazine. The columns are variables on automobile design and performance (e.g., mileage, weight, horsepower). The rows are the different automobile models.

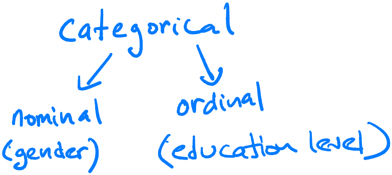
```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1

> dim(mtcars)
[1] 32 11
```

The `head()` command is used to preview the first several rows of the data table, and the `dim()` command gives the dimensions (32 car models and 11 variables).

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, mileage (`mpg`) and weight (`wt`) are numerical variables.
- **Categorical variables** take on values that fall into distinct categories.
  - For example, transmission type (`am`) is a categorical variable since each car model either has an automatic (coded as 0) or manual transmission (coded as 1).



categorical

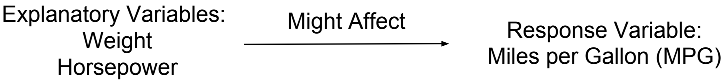
nominal (gender)

ordinal (education level)

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 weight (`wt`) and horsepower (`hp`) affect mileage (`mpg`). In this case, mileage (`mpg`) would be the response variable, and weight (`wt`) and horsepower (`hp`) would be the explanatory variables.



Explanatory Variables:
Weight
Horsepower

Might Affect

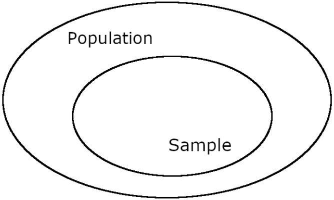
Response Variable:
Miles per Gallon (MPG)

Sampling Concepts

- Samples and populations
- Parameters and statistics
- Statistical inference

Sampling Concepts

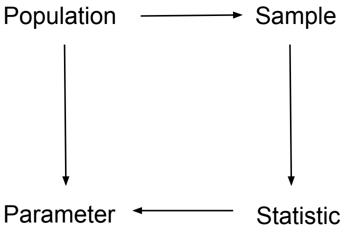
- A **population** is the set of all individuals or cases of interest to the investigator.
- A **sample** is any subset selected from the population.



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 unknown).
- A **statistic** is a numerical characteristic of the sample (varies depending on sample).



Population

Sample

Parameter

Statistic

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[1], 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

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

Sampling Concepts

Statistical inference is the process of using a random sample to infer properties about the greater population of interest. A major task in statistical inference is to quantify the uncertainty of our estimates or predictions based on a random sample. Mathematical or computational methods can be used to accomplish this task.

Example: Going back to the Gallup poll, we are primarily interested in using the sample of respondents to make an inference about the proportion of all American adults that wear masks in indoor settings. The Gallup poll reported a **margin of error** of $\pm 2\%$, so we would expect that the population proportion to be somewhere between 84% and 88%.

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 Centers for Disease Control (CDC) uses telephone surveys to collect data on health-related risk behaviors. Respondents are asked questions about diet, exercise, smoking, and level of health care coverage.

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

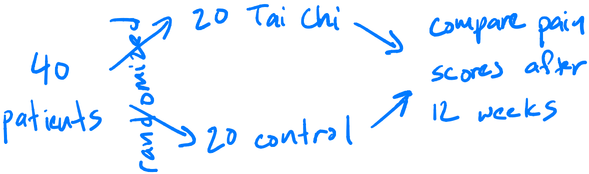
- Knee osteoarthritis (OA) is a common problem in the elderly population that results in pain and reduced quality of life.
- Researchers at Tufts Medical Center conducted a randomized experiment to evaluate the effectiveness of Tai Chi in treating OA symptoms.[2]

[2] 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

response variable: pain score

- 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.



Experimental Studies: Example

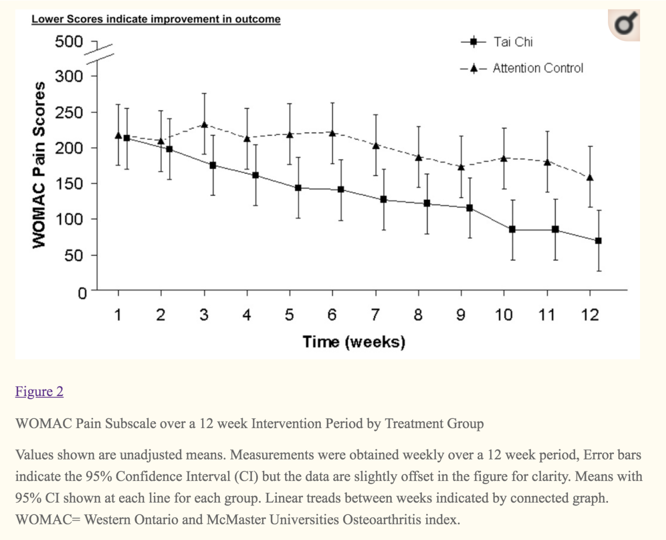


Figure 2

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.

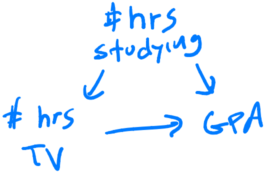
Confounding Variables

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, in general, the more TV students watch, the lower their GPAs. 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 students spend studying.



# hrs studying
# hrs TV
GPA