STAT 140: Week One

08/30/2021

Course Description

Welcome to STAT 140! This semester, you will begin to learn the rules of collecting, summarizing, and analyzing data, as well as how to communicate statistical conclusions to the wider world. Ideally, this is your first step towards becoming a responsible consumer, producer, and distributor of statistical evidence.

This course provides an overview of statistical methods, their conceptual underpinnings, and their use in various settings taken from current news, as well as from the physical, biological, and social sciences. Topics will include exploring distributions and relationships, planning for data production, sampling distributions, basic ideas of inference, inference tools for distributions, and inference tools for relationships. [1]

[1] Mount Holyoke College Academic Catalog, 2019-2020

Required Materials

- Textbook: *OpenIntro Statistics* (4th Edition) by Diez, Çetinkaya-Rundel, and Barr (ISBN-13: 978-1943450077). This is an open access textbook and you can download it for free here.
- Software: R, version 3.6.1 or higher. R can be downloaded for free here, or you may use the Mount Holyoke R server.

Course Structure

Your course grade will be determined based on your performance in six categories: Homework (250 points), Progress Checks (150 points), Journals (150 points), Project (100 points), and Exams (350 points).

You should expect one homework assignment, one progress check, and one journal per week, unless I let you know otherwise!

Class Culture

As you know, this semester will have extra challenges due to the current public health crisis. Please be aware that different members of this class may have varying comfort levels and are willing to take different risks–we should all commit to respecting each other.

- One way that we can protect each other is by meeting outside–for the first few weeks of the semester, while the weather is nice, I will plan on meeting outside for open office hours (Tuesday afternoons) and potentially office hour appointments (Mondays, Wednesdays, and Thursdays).

Course Policies

- During Class: Please do your best to keep distractions during class to a minimum! This includes avoiding noisy foods/beverages, muting electronic devices, and keeping classroom entrances and exits as quiet as possible.
- Attendance: Lecture attendance is strongly, strongly encouraged. Lab attendance is mandatory.
- Late Assignments: Any homework or progress check can be submitted up to two days late with no penalty. Please note that even though there is no late penalty, you should still try to keep up with due dates. If your work is frequently late, we will have to plan for the rest of the semester together.

My hope is that after Week 1 (08/30-09/03), you will be able to:

- Identify populations and samples, as well as parameters and statistics.
- Describe datasets and identify discrete, continuous, nominal, and ordinal variables.
- Discuss different types of sampling and their advantages and disadvantages.
- Differentiate between anecdotal evidence, observational studies, and experiments.

OpenIntro Reading: Sections 1.1-1.4, Introduction to Data

Data

- **Data** is...
  - Observations collected from the likes of field notes, surveys, and experiments.
  - Data forms the background of statistical investigation.
  - What other sources of data can you think of?
- Examples:

Populations and Samples

- A **population** is...
  - A complete group of all cases, subjects, or observations of interest.
  - Examples: all adults in the United States, all students enrolled in at least one course at Mount Holyoke College
- A **sample** is...
  - A subset of cases, subjects, or observations in a given population.
  - Example: this class!

Parameters and Statistics

- A **parameter** is...
  - A summary of an aspect of a population, often a number.
  - Example: the mean height of all adults in the United States.
- A **statistic** is...
  - A summary of an aspect of a sample, often a number.
  - Example: the average height of all students in this class.

Describing Datasets

- A **case** or **observational unit** is a single member of a sample.
- In a table or **data matrix**, cases are represented by rows.
- A **variable** is a characteristic of the cases, subjects, or observations in a sample.
- In a table or data matrix, variables are represented by columns.

Loading Datasets in R

There are many ways to load a dataset in R. One way is to use a comma separated value, or 'csv', file. Sometimes, the data is given to you as a csv file. If it isn't, you can save it as a csv by changing the extension in Excel during the 'save as' process.

```
reading_data <- read.csv("2021 Reading Challenge.csv",
                         stringsAsFactors = TRUE)
## Pound signs indicate comments.
## Change the directory if necessary!
```

Exploring Datasets in R 1

- Most of the time, we want to be working with what R calls a **data frame**.
- A data frame is another name for a data matrix.

```
is.data.frame(reading_data)
```

```
## [1] TRUE
```

The `is.data.frame()` command returns a `TRUE` or `FALSE`.

Exploring Datasets in R 2

- The `dim()` function gives the dimensions of a data matrix.

```
dim(reading_data)
```

```
## [1] 38 15
```

How many cases do you think this dataset has? How many variables?

- Cases: 38
- Variables: 15

The `dim()` command always returns the number of rows and then the number of columns.

Exploring Datasets in R 3

What variables are included in this dataset?

```
glimpse(reading_data)
```

```
## Rows: 38
## Columns: 15
## $ Book.Id                <int> 51979067, 50997696, 54481371, 54
## $ Title                  <fct> "Under the Tulip Tree", "Outlawe
## $ Author                 <fct> Michelle Shocklee, Anna North, A
## $ Additional.Authors     <fct> , , J.D. Jackson, , Roshni Shukl
## $ ISBN                   <fct> 1496446070, , , , , , 593295846,
## $ ISBN13                 <dbl> 9.78150e+12, NA, NA, NA, NA, NA,
## $ My.Rating              <int> 3, 4, 3, 3, 4, 4, 3, 3, 3, 2, 5,
## $ Average.Rating         <dbl> 4.27, 3.57, 3.90, 4.48, 3.89, 3.
## $ Publisher              <fct> "Tyndale House Publishers", "Blo
## $ Binding                <fct> Paperback, Hardcover, Audible Au
```

Variable Types 1

- **Numerical** Variable:
  - A variable taking a wide range of numerical values.
  - A variable with values you can naturally add, subtract, or averages.
  - Examples: Height, salaries, GPA, SAT scores, etc.
- **Categorical** Variable:
  - A variable you can sort into categories.
  - The values a categorical variable can take are known as levels.
  - Examples: Eye color, home state, residence type, etc.

Variable Types: Example 1

What type of variables are `Average.Rating` and `Number.of.Pages`?

```
## The dollar sign operator, $, allows you to
## access a single column of a data frame by name.
glimpse(reading_data$Average.Rating)
```

```
##  num [1:38] 4.27 3.57 3.9 4.48 3.89 3.86 3.64 4.45 3.77 3.95 ...
```

```
glimpse(reading_data$Number.of.Pages)
```

```
##  int [1:38] 400 261 NA 10 NA NA 352 9 400 NA ...
```

Variable Types: Example 2

What type of variables are `Binding` and `My.Rating` ?

```
## What does the command levels() do?
levels(reading_data$Binding)
```

```
## [1] "Audible Audio" "Audiobook"     "Hardcover"     "Paperback"
```

```
## What does the command unique() do?
unique(reading_data$My.Rating)
```

```
## [1] 3 4 2 5 0
```

Variable Types 2

- **Discrete** Variables:
  - A numerical variable taking only values with 'jumps' or 'steps'.
  - Examples: Number of siblings, number of credit hours, etc.
  - How *many*?
- **Continuous** Variables:
  - A numerical variable taking any real value.
  - Examples: Height, salary, etc.
  - How *much*?

Variable Types 3

- **Nominal** Variables:
  - A categorical variable without any type of ordering.
  - Examples: Home state, eye color, etc.
- **Ordinal** Variables
  - A categorical variable with a natural ordering.
  - Example: Class (first year, sophomore, junior, senior)

Variable Types, Revisited

- What type of variable is...
  - `Average.Rating` (4.27, 3.57, 3.9, 4.48, 3.89, 3.86 ...)?
  - `Binding` (Paperback, Hardcover ...)?
  - `Number.of.Pages` (400, 261, NA, 10, NA, NA)?
  - `My.Rating` (0, 2, 3, 4, 5)?
- Using this dataset, can you give other examples of...
  - Nominal variables?
  - Discrete variables?

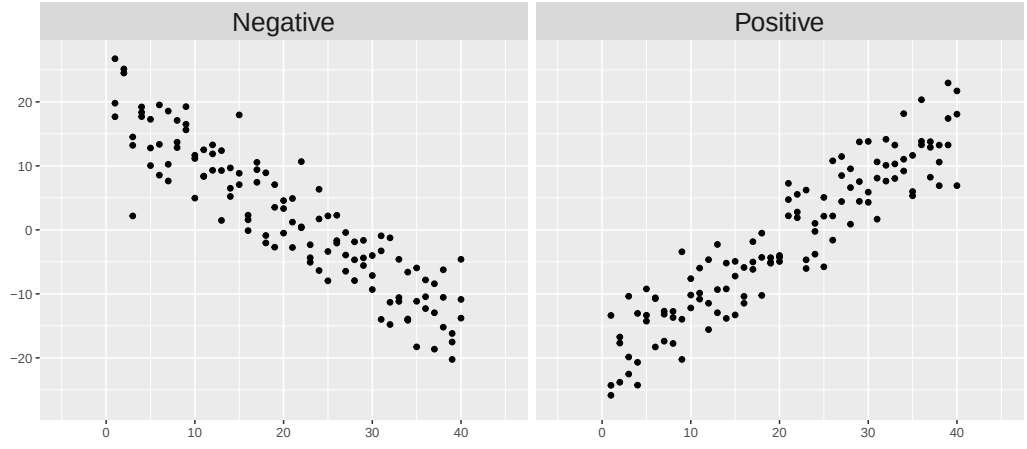
Relationships between Variables 1

- **Associated Variables**
  - Two variables that show some connection with one another.
  - Associated variables can also be called **dependent variables**.
  - Types of associations: linear, exponential, etc.
- **Independent Variables**
  - Two variables that are not associated.
- *A pair of variables are either associated or independent! They cannot be both.*

Relationships between Variables 2

- **Positive Association**
  - As one variable increases, the other variable tends to increase.
  - Examples:
- **Negative Association**
  - As one variable increases, the other variable tends to decrease.
  - Example:

Relationships between Variables 3



Relationships between Variables 4

- Does one variable affect the other?
- **Explanatory Variable**
  - A variable that *might* cause a change in another variable.
- **Response Variable**
  - A variable that *might* be changed due to changes in another variable.
- Labeling variables in this way does not guarantee a relationship exists! In other words, correlation does not imply causation.

Data Collection Methods

- **Observational Study**
  - A study where researchers collect data without directly interfering with how the data arise.
- **Experiment**
  - A scenario where researchers impose a treatment and collect the resulting data.
  - In a **randomized experiment**, subjects are randomly assigned to one group or another.
- **Anecdotal Evidence**
  - Data collected in a haphazard fashion, including personal experience.

Observational Studies

- Observational studies are especially prone to being distorted by confounding variables.
- **Confounding Variable**
  - A variable that is correlated with both the explanatory and response variable.
  - Also called a **lurking variable**, **confounding factor**, or **confounder**.
  - Example: a large study in Taiwan showed that the more electrical appliances a household owned, the more likely they were to use contraceptives.
    - What type of association is this (positive or negative)?
    - What is one potential confounding variable?

Types of Sampling 1

- **Simple Random Sampling**
  - A sample where each individual has the same chance of being chosen and each group of individuals has the same chance of being chosen.
- **Stratified Sampling**
  - A sample where subjects are grouped by some characteristic, and then randomly sampled from those groups relative to the group's representative in the whole population.
  - A **strata** is the group according to some characteristic.

Types of Sampling 2

- **Cluster Sampling**
  - A sample where subjects are grouped, usually so that they are physically close together, and then the groups are randomly chosen.
  - A **cluster** is the group of subjects.
  - **Multistage Sampling**
  - A sample similar to cluster samples, but the clusters are also randomly sampled.
- **Convenience Sampling**
  - A sample consisting of individuals that are easily accessible or close by.

Types of Sampling 3

Say I am interested in finding out what MHC students are planning to do this weekend.

- If I group students by their memberships in different clubs (athletic teams, musical groups, etc.) and then randomly sample from those groups proportional to their representativeness in the student body, what type of sample am I using?
- If I walk to the stairwell after class and ask students as they walk by, what type of sample am I using?
- How would I run a simple random sample?
- How would I run a cluster sample?

Characteristics of Samples

- **Representativeness**: Do the samples capture the same characteristics of the population?
- **Bias**: Is there some type of systematic error skewing the results towards a certain type of subject?
- **Non-response**: Are subjects choosing to respond to the sample?

Observational Studies vs. Experiments

- **Observational Study**
  - A study where researchers collect data without directly interfering with how the data arise.
- **Experiment**
  - A scenario where researchers impose and collect the resulting data.

Experiments

- **Treatment Group**
  - The group of subjects receiving the treatment.
- **Control Group**
  - The group of subjects receiving no treatment.
  - Instead of not being treated at all, subjects belonging to a control group often receive a **placebo**, a fake treatment designed to make the subjects think they are getting the real treatment.
  - Sometimes, using these fake treatments result in a **placebo effect**, a slight but real improvement that happens when subjects think they are being treated with something that might help them.
- Is treating a subject with a placebo always practical? Is treating a subject with a placebo always ethical?

Blinding

- **Blind**
  - When subjects are unaware if they are in the treatment group (receiving the real treatment or drug) or the control group (receiving the placebo or nothing), the study is said to be blind.
- **Double-Blind**
  - When both the subjects and the doctors and researchers are unaware of who is receiving the real treatment, the study is said to be double-blind.
- A study that is not blind or double-blind is known as an **open-label** study.

Principles of Experimental Design 1

- **Controlling**
  - Researchers do their best to minimize, or control, differences between the groups in an experiment.
- **Randomization**
  - Researchers randomly assign subjects to either the control or treatment groups. By randomizing, the differences between the groups are minimized and accidental biases are prevented from corrupting the data.

Principles of Experimental Design 2

- **Replication**
  - Researchers replicate the study either by repeating it with a large number of subjects OR repeating the entire study.
- **Blocking**
  - Researchers group patients based on some variable and then randomly assign members within the group to the control or treatment. These groups are called **blocks**. This strategy guarantees equal amounts of the variable in treatment and control groups.

Establishing Causality 1

If you are going to establish a causal relationship, you must use an experiment. However, experiments are not always enough. You also need:

1. Theoretical Justification: Using theoretical knowledge and context clues, there must be an explanation of **WHY** X might cause Y.

2. Temporal Sequencing: X **MUST** come before Y to show X causes Y.

3. Evidence of Association: There must be empirical evidence of an association between X and Y to show X causes Y.

- Empirical — coming from data.

Establishing Causality 2

4. Elimination of Alternate Causes: There must be no other reasonable explanation for why X causes Y.

- This includes elimination of confounding variables!
- If, and only if, you have shown all four of these things, you can imply causality.

Establishing Causality 3

Do observational studies have value?

- Establishing causation is not always the goal!
  - Sometimes it is enough to know that there are differences in two groups without needing to know why those differences exist.
- Establishing causation can be done in other ways (RARELY).
  - Sometimes there is a theoretical reason that confounding factors cannot be responsible for changes in the response variable.
  - Sometimes it is unethical to run a randomized experiment.
  - Example: Does smoking cause lung cancer?
- Analysis of observational data may lend evidence toward causal theories and suggest the direction of future research.
  - Example: Association between smoking and lung cancer.