CORE-UA 111.003: From Data to Discovery, Fall 2018
Syllabus

Instructor: Mutiara Sondjaja
Email: sondjaja@nyu.edu
Office: WWH 725
Lecture: TTh 08:55 -10:45am
Classroom: GCASL 369
Office hours: TBA

Goals of the Course

We live in an era of almost too much information. Today’s technology enables us to collect massive amounts of data, including images of distant planets, the ups and downs of the economy, and the patterns of our tweets and online behavior. As educated citizens living in a world of data and technology, we need to be discerning about how journalists, politicians, and scientists draw conclusions from data, and to be mindful about how data are collected and used.

This is a hands-on course that aims to equip students with skills in data, quantitative, and computational thinking.

1. **Working with Data.** Students can comfortably use R and Jupyter Notebook to download openly-available data, explore data, and visualize data.

2. **Thinking Quantitatively.** Students understand how to interpret results from sampled data, understand what a mathematical model is, can make prediction using basic models, and can interpret results of these predictions. Students will also develop the habit to ask questions about the validity of claims and conclusions drawn from data before accepting them.

3. **Thinking Computationally.** Students understand and can fluently use basic programming elements, including: types of data, functions, conditional statements, and loops.

Throughout the course, we will also examine issues such as **data privacy and ethics**.

Prerequisites: The course has no prerequisites beyond basic arithmetic and algebra.

Expectations: As your instructor, my aim is to help you gain firm understanding of the above topics. You can expect that course material and information are presented clearly, and that I am available to answer your questions: in class, during the scheduled drop-in office hours, and via Piazza within 2 business days.

I am rooting for you success! but your learning cannot happen without your active involvement. You are expected to actively participate during in-class problem-solving, contribute to a positive learning environment, spend time and thoughtful effort in all assignments, and maintain the highest level of academic integrity.

Textbooks and Required Technology

1. **CoCalc:** You are required to create a free account on CoCalc.com using your@nyu.edu email address. All labs, homework assignments, and lecture demonstrations will be available through CoCalc.

2. **NYU Classes** and **Piazza:** Announcements and other course material will be posted on NYU Classes. We will use Piazza for questions and answers.

3. **Reference Textbooks (Optional):**
     PDF copy (free) or print copy ($8.49):  [Open Intro](http://www.openintro.org/)
   - *R For Data Science*, by Hadley Wickham.
     Online copy (free): [R for Data Science](http://r4ds.had.co.nz/)
Assessments and Grading

Labs (12%)
- Labs are computer-based work you will do in the weekly recitations. If you like, you can choose to bring your laptop computer to recitation. Each lab worksheet is due at the end of your assigned Friday lab session on CoCalc. No make-ups or late submissions are allowed. Two lowest lab grades are dropped.

Weekly Homework (12%)
- Each homework assignment is computational, a writing assignment, or project-related. Homeworks are due at 9AM on Fridays through CoCalc. See further instructions on NYU Classes. No emailed or hardcopy homework will be accepted.
- Late homework policy: To maintain fairness, homework submitted within 24 hours of the official deadline will be accepted with penalty. Homeworks submitted more than 24 hours after the deadline are not accepted and will receive 0 points. Technical issues are not a valid excuse for additional homework extension. Two lowest homework scores will be dropped.

Clicker Participation (2%)
- A “clicker app” will be used to encourage active learning. Further instructions will be provided on NYU Classes. You must bring your clicker device to each lecture.
- A clicker participation grade of 75% or better corresponds to all 2% of clicker credit. Manual attendance will not be taken due to the size of the class. The generous grading of clicker participation take into account the possibility of students’: (1) having to miss class due to other obligations or minor illness, (2) forgetting their clicker on a non-regular basis, (3) non-persistent technical problem with the clicker app/equipment.

Projects (24%)
There are two projects in this course. This is your opportunity to use your newly acquired data skills on problems that are more involved than the weekly homework. Further detail will be posted on NYU Classes.
- Project 1 (8%): You will work in groups of 3 with a dataset that we provide for you. See calendar.
- Project 2 (16%): You will work in groups of 3 with a dataset of your choice. See calendar.

Exams (50%)

<table>
<thead>
<tr>
<th>Exam</th>
<th>Date</th>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midterm 1</td>
<td>October 2 (in class)</td>
<td></td>
<td>Exploring and Visualizing Data</td>
</tr>
<tr>
<td>Midterm 2</td>
<td>November 8 (in class)</td>
<td></td>
<td>Statistical Inference</td>
</tr>
<tr>
<td>Final Exam</td>
<td>December 18, 2PM-3:50PM, Location TBA</td>
<td>All Topics</td>
<td></td>
</tr>
</tbody>
</table>

The lowest exam score will be dropped; the remaining two exams are weighed equally: each exam counts towards 25% of your overall course grade.

Academic Integrity

We value hard work and integrity, and do not tolerate academic dishonesty. You are expected to uphold academic integrity as specified by the university and the College of Arts and Sciences. Remember that we are here to learn.

Group work, assignments, and academic integrity: You are encouraged to form study groups and to exchange ideas with classmates when wrestling with homework and projects. However, you must write up your weekly homework and lab individually. A good rule of thumb: make sure that you can reproduce or explain your submitted work individually without looking at what you have written up. And finally, give credit where it’s due: indicate
with whom you have held discussions as you worked on your assignments.

**Course Policies**

There will be no accommodation for missed homework, labs, and exams, except in the cases of illness, observance of religious holidays, official/university-approved travels (e.g., athletic meets), or exceptional, extenuating circumstances. In the case of observance of religious holidays or university-approved travels, you must make arrangements to make up missed work **at least one week in advance**. In the case of illness, you must present a letter from a physician/health care provider as soon as possible. Students with disabilities or requiring special accommodations must make individual arrangements with Moses Center.

**Advice for making most of this course**

- **Attend lectures and labs.** It is very hard to catch up if you regularly miss classes. Your instructor and TAs are there to help you learn; use this opportunity well.

- **Get your hands dirty in class.** Actively participate when we solve problems in class. Passively listening to lectures and taking notes are generally not sufficient to really internalize new, challenging ideas.

- **Spend time** on each homework, lab, and projects. Expect to spend 4-8 hours each week on assignments and study outside of class. This is your opportunity to wrestle with and to internalize new ideas introduced in class. When working on assignments, strive to really understand the ideas behind the methods.

- **Don’t hesitate to get help, and to do so early:**
  - **Attend instructor and TA’s office hours.** Do not wait to seek help if you have any questions about the ideas introduced in class. Office hours schedule will be posted in the NYU Classes page for our section. No appointment is needed to attend these regular office hours.
  
  - **Form study groups,** but it’s critical that you write up your own homework individually. Homework and assignments that are copied from or written by someone else are not hard to spot. Academic dishonesty is not tolerated.
  
  - **Piazza:** Use the course Piazza page to post questions and to respond to classmates’ questions. When you do, make sure to be courteous and respectful. For homework-related questions, full solutions to homework/worksheet problems should not be requested or provided. 3
<table>
<thead>
<tr>
<th>Theme</th>
<th>Week/Date</th>
<th>Lecture</th>
<th>Topics</th>
<th>Assignment Due</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exploring and Visualizing Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 1, 09/04/18</td>
<td>1</td>
<td></td>
<td>What is this class about?</td>
<td></td>
</tr>
<tr>
<td>Week 1, 09/06/18</td>
<td>2</td>
<td></td>
<td>Introduction to computing using R</td>
<td></td>
</tr>
<tr>
<td>Week 1, 09/07/18</td>
<td></td>
<td></td>
<td>No Lab</td>
<td></td>
</tr>
<tr>
<td>Week 2, 09/11/18</td>
<td>3</td>
<td></td>
<td>Data frames and types of data</td>
<td></td>
</tr>
<tr>
<td>Week 2, 09/13/18</td>
<td>4</td>
<td></td>
<td>Working with data frames</td>
<td></td>
</tr>
<tr>
<td>Week 2, 09/14/18</td>
<td></td>
<td></td>
<td>Lab 1: Using R and Jupyter Notebook to explore data</td>
<td>HW 1, Lab 1</td>
</tr>
<tr>
<td>Week 3, 09/18/18</td>
<td>5</td>
<td></td>
<td>Data visualization: more on histograms</td>
<td></td>
</tr>
<tr>
<td>Week 3, 09/20/18</td>
<td>6</td>
<td></td>
<td>Applications to text analysis; Review</td>
<td></td>
</tr>
<tr>
<td>Week 3, 09/21/18</td>
<td></td>
<td></td>
<td>Lab 2: Exploring and visualizing data</td>
<td>HW 2, Lab 2</td>
</tr>
<tr>
<td>Week 4, 09/25/18</td>
<td>7</td>
<td></td>
<td>Data visualization: more on histograms</td>
<td></td>
</tr>
<tr>
<td>Week 4, 09/27/18</td>
<td>8</td>
<td></td>
<td>Applications to text analysis; Review</td>
<td></td>
</tr>
<tr>
<td>Week 4, 09/28/18</td>
<td></td>
<td></td>
<td>Midterm Review</td>
<td>HW 3</td>
</tr>
<tr>
<td><strong>Statistical Inference</strong></td>
<td>Week 5, 10/02/18</td>
<td>9</td>
<td>Understanding randomness and probability</td>
<td></td>
</tr>
<tr>
<td>Week 5, 10/04/18</td>
<td></td>
<td></td>
<td>Lab 3: Simulating random events</td>
<td>Lab 3</td>
</tr>
<tr>
<td>Week 6, 10/09/18</td>
<td></td>
<td></td>
<td>No Class (Legislative Day – Monday Schedule)</td>
<td></td>
</tr>
<tr>
<td>Week 6, 10/11/18</td>
<td>10</td>
<td></td>
<td>Programming elements: conditionals and loops</td>
<td></td>
</tr>
<tr>
<td>Week 6, 10/12/18</td>
<td></td>
<td></td>
<td>Lab 4: Functions; Project 1 Intro</td>
<td>HW 4, Lab 4</td>
</tr>
<tr>
<td>Week 7, 10/16/18</td>
<td>11</td>
<td></td>
<td>Programming elements: conditionals and loops</td>
<td></td>
</tr>
<tr>
<td>Theme</td>
<td>Week/Date</td>
<td>Lecture</td>
<td>Topics</td>
<td>Assignment Due</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------------</td>
<td>---------</td>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td></td>
<td>Week 7, 10/18/18</td>
<td>12</td>
<td>Random sampling, point estimates, and sampling distribution</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 7, 10/19/18</td>
<td></td>
<td>Lab 5: Understanding the central limit theorem via simulations</td>
<td>HW 5, Lab 5</td>
</tr>
<tr>
<td></td>
<td>Week 8, 10/23/18</td>
<td>13</td>
<td>Bootstrap and confidence intervals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 8, 10/25/18</td>
<td>14</td>
<td>Models</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Week 8, 10/26/18</strong></td>
<td></td>
<td>Lab 6: Bootstrap and confidence intervals</td>
<td>Project 1, HW 6, Lab 6</td>
</tr>
<tr>
<td></td>
<td>Week 9, 10/30/18</td>
<td>15</td>
<td>Comparing distributions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 9, 11/01/18</td>
<td>16</td>
<td>Testing statistical hypothesis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 9, 11/02/18</td>
<td></td>
<td>Lab 7: Making a hypothesis and performing hypothesis tests</td>
<td>HW 7, Lab 7</td>
</tr>
<tr>
<td>Models and Predictions</td>
<td>Week 10, 11/06/18</td>
<td>17</td>
<td>A/B Testing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 10, 11/08/18</td>
<td>18</td>
<td>Scatterplots and correlation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 10, 11/09/18</td>
<td></td>
<td>Lab: Midterm 2 Review</td>
<td>HW 8</td>
</tr>
<tr>
<td></td>
<td><strong>Week 11, 11/13/18</strong></td>
<td></td>
<td>Midterm Exam 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 11, 11/15/18</td>
<td>19</td>
<td>Linear regression and prediction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 11, 11/16/18</td>
<td></td>
<td>Lab 8: Scatterplots and linear regression</td>
<td>Lab 8</td>
</tr>
<tr>
<td></td>
<td>Week 12, 11/20/18</td>
<td>20</td>
<td>Classification</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 12, 11/22/18</td>
<td></td>
<td>No Class (Thanksgiving Break)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 12, 11/23/18</td>
<td></td>
<td>No Class (Thanksgiving Break)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 13, 11/27/18</td>
<td>21</td>
<td>Classification: k-NearestNeighbors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 13, 11/29/18</td>
<td>22</td>
<td>Classification: k-NearestNeighbors, continued</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 13, 11/30/18</td>
<td></td>
<td>Lab 9: Classification; Project 2</td>
<td>Lab 9</td>
</tr>
<tr>
<td>Theme</td>
<td>Week/Date</td>
<td>Lecture</td>
<td>Topics</td>
<td>Assignment Due</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------</td>
<td>---------</td>
<td>--------------------------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td></td>
<td>Week 14, 12/04/18</td>
<td>23</td>
<td>Cross Validation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 14, 12/06/18</td>
<td>24</td>
<td>Model Assessment and Validation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 14, 12/08/18</td>
<td>25</td>
<td>Lab 10: Model Assessment and Validation; Project 2</td>
<td>HW 9, Lab 10</td>
</tr>
<tr>
<td></td>
<td>Week 15, 12/11/18</td>
<td>26</td>
<td>Privacy and Ethics of Data; Conclusion</td>
<td>Project 2 Presentations (in class)</td>
</tr>
<tr>
<td></td>
<td>Week 15, 12/13/18</td>
<td></td>
<td>Project Presentations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Week 15, 12/14/18</td>
<td></td>
<td>Lab: Finish working on Project 2 write-up; Final Exam Review</td>
<td>Project 2 Write-Up Due at 9PM</td>
</tr>
<tr>
<td></td>
<td>Week 16, insert date</td>
<td>25</td>
<td>Final Exam: All topics (2pm-3:50pm, in Silver 401)</td>
<td>Project 2 Self/Peer Assessments Due at 9PM</td>
</tr>
</tbody>
</table>

Calendar URL
List of Topics and Learning Outcomes

Part I: Exploring and Visualizing Data

Learning Outcomes: By the end of the module, students will have the following skills.

1. Students can parse data that are represented in a tabular format.
2. Students can use the statistical programming language R, along with relevant packages for working with and transforming data (dplyr).
3. Students can interpret and critique basic data visualizations that appear in news articles or scholarly publications in their fields.
4. Students can use R along with relevant packages (ggplot) to create good data visualizations.

Lecture 1: Overview

- Overview and motivation of topics and learning goals
- Demonstration and overview of R, Jupyter Notebooks, and CoCalc
- Overview of the modeling and data analysis process
- Dataset/Case Study: UC Berkeley 1973 Graduate Admission Data; Simpson’s Paradox

Lecture 2: Introduction to R and Jupyter Notebooks

Goal: Hands-on introduction to R and Jupyter Notebooks

1. Basic Arithmetic in R
2. Names/variables: giving names to store numbers, strings (segway to types of data)
3. Types of data
   - Numerical: int, double
   - String/text/character
   - Boolean
   - Categorical/factors
4. Comments
5. Functions
   - Introduction to the idea of functions as "verbs" or "actions"
   - Some basic functions:
     - `print(NAME)`: to print string or values stored in a variable
     - `sqrt(NAME)`: to take the square root of a number
     - `abs(NAME)`: to take the absolute value of a number
6. Lists
   - The idea lists as a collection of data/values of the same type
   - Functions on lists:
     - `c(VALUE1, VALUE2, ... )`: to put values into a list
     - `length(LISTNAME)`: to find the number of values in a list
     - `max(LISTNAME)`: to find the largest value in a list
     - `min(LISTNAME)`: to find the smallest value in a list
     - Arithmetic with lists
     - Adding two lists together, subtracting, multiplying, dividing
     - `sum(LISTNAME)`: to find the sum of values in a list
7. Tables of data ("data frames")
   - A brief introduction to R packages
   - How to get datasets in R:
     - Built-in data (in the "datasets" package) (Lecture 2)
     - Importing from a CSV file in your folder or from the internet (Lecture 3)
• A brief explanation about CSV files (further discussed in lecture 3)
• Understanding that columns of a data frame are themselves lists
• Functions on data frames:
  • `dim( DATAFRAMENAME )`: to find the number of rows and columns of the data frame `DATAFRAMENAME`
  • `names( DATAFRAMENAME )`: to find the column names of the data frame `DATAFRAMENAME`
  • `head( DATAFRAMENAME )`: to preview the first few rows of the data frame `DATAFRAMENAME`
    (the default is 6 rows)
  • `head( DATAFRAMENAME, NUMBER )`: to preview the first `NUMBER` rows of the data frame `DATAFRAMENAME`
Lectures 3 and 4: Working with Data

Goal: To develop skills and comfort in working with tables/data frames

1. Importing Datasets
   - What CSV files are
   - `read.csv( 'FILENAME.csv')`: to "read" a csv file and import it as a data frame in R
   - `read.csv( url('http://webaddress.etc'))`: to "read" a csv file from a URL (web address) and import it as a data frame in R

2. Working with rows of data frames
   - Arithmetic with lists and columns of a data frame
     - Creating a new column in a data frame, filled by a result of arithmetic on a different column.
       - Examples:
         - Converting between units
         - Computing proportions and percentages, etc.
     - Finding sums and averages of columns
       - using `sum( LISTNAME )/length( LISTNAME )` or using `mean( LISTNAME )`
   - Working with `dplyr` functions:
     A. `arrange()`
        - `arrange( DATAFRAMENAME, COLUMNNAME )`: to sort rows by values in the column called COLUMNNAME, in ascending order
        - `arrange( DATAFRAMENAME, desc( COLUMNNAME ) )`: to sort rows by values in the column called COLUMNNAME, in descending order
     B. `filter()`
        - `filter( DATAFRAMENAME, CRITERIA )`: to display only rows in the data frame DATAFRAMENAME that satisfy the criteria specified in CRITERIA.
     C. `mutate( DATAFRAMENAME, NEWCOLUMNNAME )`: to add a new column to a data frame
     D. `group_by( DATAFRAMENAME, VARIABLENAME )`: to group rows of data by their values in a particular column
     E. `summarize( GROUPEDDATAFRAMENAME, NEWVARIABLENAME = FORMULA )`: to compute a summary quantity from grouped data (usually used together with group_by() )

Lectures 5 and 6: Creating Basic Data Visualizations

Goals: Students are able to interpret data visualizations; students view data visualizations (in media, studies) critically; students can create basic data visualizations using R

1. Good and bad data visualizations (see lecture slides)
2. Types of data visualizations
   - Bar plots; Scatterplots; Histograms; Line graphs; Pie charts
   - Purpose of each type of data visualization
   - Which visualization is appropriate for a given type of data
3. Creating basic bar plots, scatterplots, line graphs, and histograms in R using ggplot()

```r
  ggplot( DATAFRAMENAME, aes( x = VARIABLE1 ) ) + geom_bar()
  ggplot( DATAFRAMENAME, aes( x = VARIABLE1, y = VARIABLE2 ) ) + geom_col()
  ggplot( DATAFRAMENAME, aes( x = VARIABLE1, y = VARIABLE2 ) ) + geom_point()
  ggplot( DATAFRAMENAME, aes( x = VARIABLE1, y = VARIABLE2 ) ) + geom_line()
  ggplot( DATAFRAMENAME, aes( x = VARIABLE1 ) ) + geom_histogram()
```

Lecture 7: Understanding Histograms
Goals: Students truly understand the information represented by histograms, including those with unequal bins; students can create histograms in R

1. Interpretation:
   - Proportions and areas under bars
   - Density and height of bars
2. Histograms with unequal bin widths
3. Creating histograms in R:
   - Creating a histogram with N bins: `ggplot( DATAFRAMENAME, aes( x = VARIABLE1 ) ) + geom_histogram( bins = N)`
   - Creating a histogram, where the bin break points are specified in LIST: `ggplot( DATAFRAMENAME, aes( x = VARIABLE1 ) ) + geom_histogram( braks = LIST)"
Lecture 8: Applications and Additional Topics in Data Exploration & Visualization

Goals: Review; Students see a modern application of topics so far (text analysis).

1. Additional visualizations / features
   - Grouped bar plots
     - `ggplot( DATAFRAMENAME, aes( x = VARIABLE1, y = VARIABLE2, fill = VARIABLE3 ) + geom_col( position = "[dodge/stacked/fill]" ) )`
   - Line graphs
     - `ggplot( DATAFRAMENAME, aes( x = VARIABLE1, y = VARIABLE2 ) + geom_line() )`

2. Applications to text analysis
   - Analyzing Jane Austen’s *Pride and Prejudice*

3. Review

Part II: Inferring from Data

Learning Outcomes: By the end of this module, students will have the following skills.

1. Students understand what it means for a process/event to be random, and understand probability distributions as relative frequencies of outcomes of repeated random events.
2. Students can simulate random events using R.
3. Students understand random sampling and the differences between sampling with and without replacement, and can carry out computations and analysis on sampled data.
4. Students understand the goal of statistical inference, and that conclusions from inference is uncertain.
5. Students understand mean, median, and mode as a "measure of central tendency" and variance or standard deviation as a measure of spread.
6. Students understand point estimates and confidence intervals and can carry out bootstrap simulations to estimate confidence intervals.
7. Students can carry out basic hypothesis test and understand the results through the lense of simulations.

Lecture 9: Randomness and Probability

Goals: Students get a good sense of what mathematical randomness and probability are, and what it means to simulate a random process using a computer

1. "Statistical Inference"
2. Simulating random events / generating random data
   - Example: Rolling a die
     - R function: `sample( POSSIBLEOUTCOMES, NUMBEROFsamPLES, replace = [TRUE or FALSE] )`
3. Visualizing randomly-generated data using R (histogram)
4. Interpreting histograms as a probability distribution
Lectures 10, 11: Programming elements: functions, conditionals, and loops

Goal: To equip the necessary basic programming skills, which will be used when we carry out simulations of random sampling

1. Defining new functions in R
   - Motivation
   - Syntax in R:

   ```R
   FUNCTIONNAME <- function( INPUT1, INPUT2, ... ) {
   [ ... task to be done ]
   OUTPUT (returned value, if any)
   }
   ```

2. Conditionals
   - Boolean (TRUE/FALSE) variables
   - If-else statements
   - Syntax in R:

   ```R
   if( CRITERION ){
   [ ... task to be done if CRITERION is TRUE]
   }
   else {
   [ ... task to be done if CRITERION is TRUE]
   }
   ```

3. Loops
   - While loop
   - Syntax in R

   ```R
   while( CRITERION ){
   [ ... task to be done as long as CRITERION is TRUE]
   }
   ```

Lectures 11, 12: Random sampling, point estimates, and sampling distribution

Goals: Students understand sampling and can simulate simple random sampling using R and can interpret the outcomes of the simulations

1. Sample vs. Population
2. Random sampling
   - R function for simple random sampling:
     ```R
     sample( POPULATION, NUMBEROFSAMPLES, replace = FALSE )
     ```
3. Repeated sampling
   - Using `sample()` within a while loop
4. Repeated sampling: Plotting the histogram of sample averages
5. Understanding sampling distributions
6. The Central Limit Theorem

Lecture 13: Bootstrap & Confidence Intervals

Goals: Introducing students to the idea of resampling and why it is useful. Students can interpret confidence intervals, and can construct confidence intervals using bootstrap simulations

1. Mean, Median, Mode, Percentiles
2. The Normal Distribution
   - Mean, Median, Mode, Percentiles
• Standard Deviation

3. Bootstrap & Confidence Intervals
• Resampling (with replacement) from one random sample
• R function for sampling with replacement:
  sample( POPULATION, NUMBEROFSAMPLES, replace = TRUE )
• Constructing confidence intervals using bootstrap
Lecture 14: Confidence Intervals; Models

Goals: Students understand what we mean by mathematical models, what they are for, and how to quantify 'fit' between models and data

1. Interpreting confidence intervals from bootstrap
2. What we mean by "models"
3. Assessing how well models are supported by data

Lecture 15: Comparing Distributions

1. Measuring distance between model and data
2. Comparing Models and Data: Examples
   - Mendel's Hypothesis
   - Fair Coin?
   - Racial and Ethnic Disparity in Manhattan Jury Pools
3. Procedure for comparing model and observed data
   A. Find the distance between the model and the observed data (the observed distance)
   B. Generate simulated data based on the model and compute distances between simulated data and the model (the simulated distances)
   C. Find what percentile the observed distance is, compared to all simulated distances
   D. Decide if the model is good or not:
      A large percentile indicates observed data is unlikely to have come from the model, which means the model is not great
      A small percentile indicates observed data is unlikely to have come from the model, which means the model is pretty good

Lecture 16: Decisions and Uncertainty

1. Decisions and Uncertainty
   - Null and alternate hypotheses
   - Test statistic; observed statistic; simulated statistic
   - Distribution of the test statistic
     - Visualized by the histogram of the simulated statistic
     - Interpreting this histogram and where the observed statistic fall
   - Procedure for testing a hypothesis:
     A. Compute the observed statistic
     B. Generate simulated data based on the model
        o And compute the test statistic for each simulated data (i.e., the simulated statistic)
     C. Find what percentile the observed statistic is, compared to all simulated statistics
     D. Decide if the model is good or not:
        o A large percentile indicates observed data is unlikely to have come from the model, which means the model is not great
        o A small percentile indicates observed data is unlikely to have come from the model, which means the model is pretty good

2. Statistical Significance
   - P-Value
   - Conventions for cutoff for P-value
   - Interpretation/meaning of P-Value

3. Error Probability
   - Type 1 Error and Type 2 Error
   - Discussion on conventions
Lecture 17: A/B Testing; Causality

1. A/B Testing
   - Permutation test
   - R function for randomly shuffling data:
     `sample( POSSIBLEOUTCOMES, NUMBEROFPOSSIBLEOUTCOMES, replace = FALSE )`

2. Causality
Part III: Models and Predictions

Learning Outcomes: By the end of this module, students will have the following skills.

1. Students understand what a mathematical model is, how to use a model to make a prediction using a model, and how to interpret the result of the prediction.
2. Students understand what linear regression is, how to perform linear regression on real data using R, and can use and interpret results of linear regression to make predictions.
3. Students understand what classification is, can construct simple classifiers on their own, understand what the k-Nearest Neighbor classifier does, can make predictions using a classifier, and can assess results of predictions.

Lecture 18: Scatterplots and correlation

1. Data, Models, and Predictions
2. Visualizing relationships using scatterplots
3. Quantifying linear relationships
   - Correlations
     - Interpretation as a measure of strength of linear relationship
     - R function for computing correlation: `cor(X, Y)`
   - Regression line
     - Review of linear functions; equation of lines
     - Quantifying "fit" of line and data: measuring "distance" from data points to line.

Lecture 19: Linear model and prediction

1. Quantifying linear relationships: Regression Line
   - Quantifying distance from data to regression line: Mean Square Error
   - Regression: finding best fit line: line that minimizes distance to datapoints (minimizes MSE)
   - R function for finding intercept and slope of the best fit line:
     ```R
     lm(formula = YVARIABLENAME ~ XVARIABLENAME, data = DATAFRAMENAME)
     ```
   - R and ggplot command for plotting data points and best fit line:
     ```R
     ggplot(studentsurvey2, aes(x = Height, y = Weight)) +
     geom_point() +
     geom_smooth(method = 'lm', formula = 'y~x', se = FALSE)
     ```
2. Making predictions using the linear regression line;
3. Uncertainty + Statistical Inference
   - Uncertainty in prediction
   - Confidence Intervals for linear regression prediction

Lecture 20: Classification

1. Linear Regression Recap
2. Classification Examples
   - Focus on binary classification
3. "Building A Classifier": Big Picture
   - Training and testing
4. Exploring An Example
   - Cancer detection: Is this tumor malignant or not?
Lecture 21: k-Nearest Neighbors

1. Classification: Recap
2. Building A Classifier: Example (continued)
   - Step 1: Explore (and Visualize) Data
   - Step 2: Use insight from data exploration to build classifiers
   - Simple Classifiers
     - Defining new functions that make a classification based on data
     - Draw thresholds and use conditionals (if-else statements)
3. k-Nearest Neighbor (kNN) Classifiers
   - Quantifying Similarities
   - Building the kNN Classifier from scratch

Lecture 22: k-Nearest Neighbors

1. k-Nearest Neighbor (kNN) Classifiers, continued
   - Quantifying Similarities
   - Building the kNN Classifier from scratch (continued)
2. Assessing Classifiers
   - Accuracy

Lecture 23: Cross Validation

1. k-Nearest Neighbor (kNN) Classifiers
   - Using the knn() function from the class R package
2. Model Selection
   - Choosing the "right" k in kNN?
   - Splitting data into Training/Validation/Test
   - Cross-Validation

Lecture 24: Model Assessment and Validation

1. Measuring Performance of Classifiers
   - Accuracy: pros and cons
   - Other metrics (precision, true positive rate, etc.) and their interpretations

Lecture 25: Privacy and Ethics; Conclusion

1. What have we learned this semester?
   - Exploring and Visualizing Data
   - Thinking quantitatively
   - Making statistical inferences
   - Thinking computationally
2. Some questions on ethics
   - Privacy
   - Data collection, informed consent, data ownership
   - "Algorithmic fairness"
   - "p-hacking"
3. Q&A ; Project 2