Elementary Statistics: Math 080

Jordan Hanson

July 7, 2020

Whittier College Department of Physics and Astronomy

Course Introduction

1. ***What is statistical analysis?***
2. Math 080: Elementary Statistics
3. Read the syllabus for a roadmap
4. This is an online summer course that meets each day.
5. **Data science project and presentation**
6. Textbook: `https://openstax.org/details/books/introductory-statistics`
7. Download and install Excel, or LibreOffice Calc

Lecture format, with modifications

- Warm-up exercise, and solution (10-15 minutes)
- Lecture via Whiteboard and slides (10-20 minutes)
- Interactive questions or polls (10 minutes)
- Laboratory activity (20 minutes)
    1. Breakout rooms
    2. Offline
- Asynchronous content
    1. Homework clues
    2. Example problems
    3. Special topics

Unit 0 Outline

1. Topics from Chapter 1: 1.1, 1.2, 1.3
    - What is a statistic?
    - Probability examples
    - Data and sampling
2. Topics from Chapter 2: 2.1 - 2.4, 2.5 - 2.8
    - Data visualization
    - Location of the data in numerical space
3. Topics from Chapter 3: 3.1, 3.2, 3.3
    - Two rules of probability

Topics from Chapter 1

What is statistical analysis?

By tradition, we begin with Mark Twain.

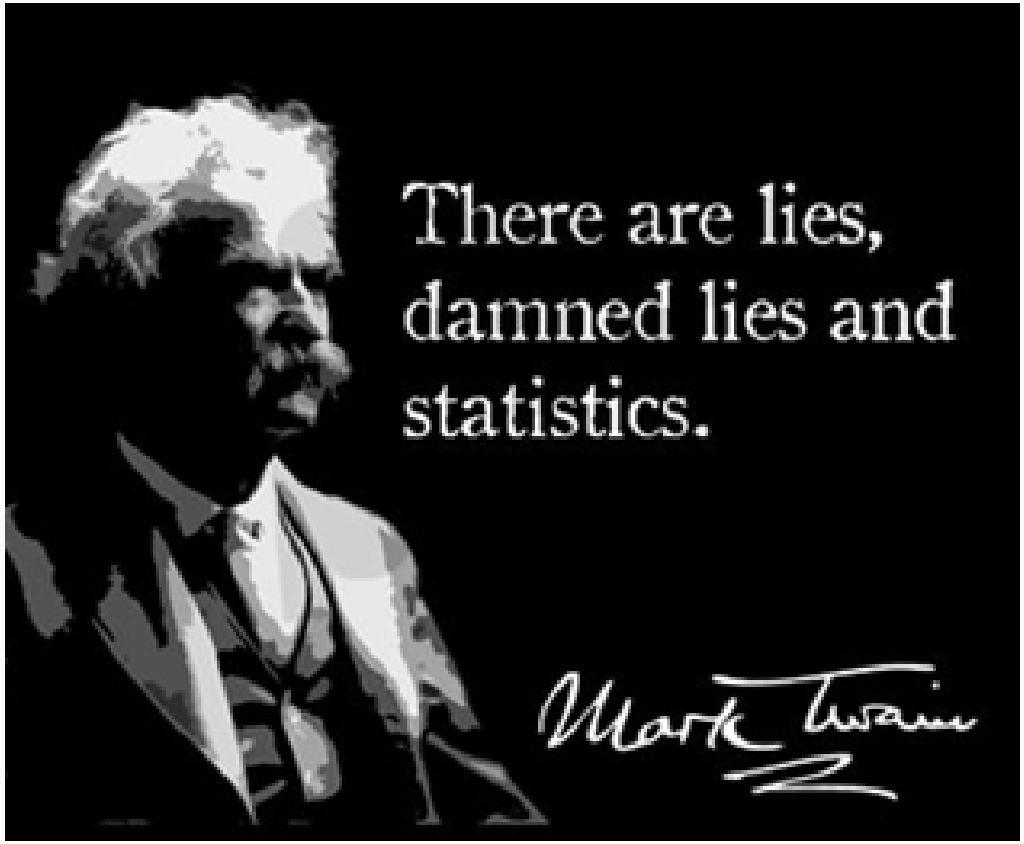

Figure 1: A famous quote from Mark Twain.

Warm-up exercises

COVID-19 data. In a March 2020 article in the magazine `wired.com`, Ferris Jabr points out that people were drawing comparisons between the influenza pandemic of 1918 and SARS-Cov-2 (COVID-19). The case fatality rate, or CFR, is the percentage of people who contract the disease that perish from it. In the 1918 outbreak, it is usually stated that there were approximately 500 million infections, 50-100 million fatalities, and an overall CFR of 2.5%. What is interesting is that the coronavirus seems to have a CFR (averaged over age) of ≈ 3 %, making it ... *higher*.

Question 1: The above paragraph listed four pieces of data. What are they?

Warm-up exercises

COVID-19 data. In a March 2020 article in the magazine `wired.com`, Ferris Jabr points out that people were drawing comparisons between the influenza pandemic of 1918 and SARS-Cov-2 (COVID-19). The case fatality rate, or CFR, is the percentage of people who contract the disease that perish from it. In the 1918 outbreak, it is usually stated that there were approximately 500 million infections, 50-100 million fatalities, and an overall CFR of 2.5%. What is interesting is that the coronavirus seems to have a CFR (averaged over age) of ≈ 3 %, making it ... *higher*.

Question 2: Which number, if any, seems to have a problem?

- A: The total number of infections in 1918
- B: The total number of deaths in 1918
- C: The 1918 influenza CFR
- D: The 2020 coronavirus CFR

Warm-up exercises

COVID-19 data. In a March 2020 article in the magazine `wired.com`, Ferris Jabr points out that people were drawing comparisons between the influenza pandemic of 1918 and SARS-Cov-2 (COVID-19). The case fatality rate, or CFR, is the percentage of people who contract the disease that perish from it. In the 1918 outbreak, it is usually stated that there were approximately 500 million infections, 50-100 million fatalities, and an overall CFR of 2.5%. What is interesting is that the coronavirus seems to have a CFR (averaged over age) of ≈ 3 %, making it ... *higher*.

Question 3: From the rest of the data in the paragraph, estimate the proper CFR of the 1918 influenza. Compare this number with the CFR of the 2020 coronavirus.

Topics from Chapter 1

Vocabulary:

1. **Probability**: The extend to which something is *likely* to occur, measured by the ratio of favorable cases to the whole number of cases possible.
2. **Population**: The total collection of people, objects, or cases under investigation.
3. **Sample**: A subset of the population for which statistical data is collected.
4. **Statistic**: *A statistic* is a number that represents a property of the sample. For example: the CFR of a *sample* of 2,500 coronavirus patients.
5. **Parameter**: Statistic measured from the *entire* population. A statistic attempts to reveal knowledge of a parameter.

Topics from Chapter 1

Vocabulary:

1. **Representative sample**: a sample that captures all of the properties of a population. Counter-example: psychological studies using undergraduate subjects.
2. **Variable**: A property of each member of the population that can be determined, either quantitative or categorical. **Data** are the actual values.

Mean: Definition 1

Let X represent a *variable* of a *population*, and x_i represent the actual value of the i-th member of a statistical *sample* of that *population*. The arithmetic mean $\bar{x}$ of the *sample* for that property is

$$\bar{x} = \frac{1}{N} \sum_{i}^{N} x_i \tag{1}$$

The mean of the variable X is the number $\bar{x}$ from the sample.

Topics from Chapter 1

Example 1: What's the average number of siblings in our community?

1. What is the population?
2. What is the sample? (Our class).
3. What is the variable?
4. What are the data?

Write in the chat area the number of siblings in your family, including yourself.

Topics from Chapter 1

Example 2: How many languages do you speak?

1. What is the population?
2. What is the sample? (Our class).
3. What is the variable?
4. What are the data?

Write in the chat area the number of languages that you can speak.

Topics from Chapter 1

Vocabulary:

1. **Proportion**: The total number of subjects in the sample that share a property, divided by the total number of subjects in the sample.
2. **Qualitative data**: Sometimes called categorical data, refers to non-numerical properties of subjects in sample (e.g. place of birth).
3. **Quantitative data**: Numerical values of variables for each subject in a sample (e.g. age).
    - Continuous quantitative data: average hours of sleep per night
    - Discrete quantitative data: average number of siblings

Topics from Chapter 1

Example 1: What fraction of Whittier College students live on campus?

1. What is the population?
2. What is the sample? (Our class).
3. What is the variable?
4. What are the data?

Write in the chat area the number 1 if you live on-campus, and the number 0 if you live off-campus or with your family.

Whittier College Factbook: 46.3% of undergraduates live on-campus.

Topics from Chapter 1

Example 2: What is the proportion of students to instructors here? (What is the student to faculty ratio of Whittier College)?

1. What is the population?
2. What is the sample? (Our class).
3. What is the variable?
4. What are the data?

Let's sum the students here, and then there is me.

Whittier College Factbook: average student to faculty ratio: 11

Topics from Chapter 1

Example 3: You go to the supermarket and purchase three cans of soup:

- 19 ounces tomato bisque
- 14.1 ounces lentil
- 19 ounces Italian wedding

...and two desserts:

- 16 ounces pistachio ice cream
- 32 ounces chocolate chip cookies

Create three data sets: one quantitative discrete, one quantitative continuous, and one categorical.

Laboratory Activity

Laboratory Activity

Go to the following link and watch the interesting TED talk by Steven Levitt from 2005 about driving safety.

`https://www.ted.com/talks/steven_levitt_surprising_stats_about_child_carseats?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare`

Answer the questions on the form entitled Laboratory Exercise 1 on Moodle for this week, and submit them via email: jhanson2@whittier.edu. (This is part of your warm-ups grade...see syllabus).

Interactive Questions

Interactive Questions

Almost always, we will give multiple-choice questions with answers A-D. If you are lost, or need extra explanation, or just feel we are going to fast, select the letter E. E stands for WAT...

After 1 round, we examine the *answer distribution*, and if 70% get it right, we move on. Otherwise, we discuss via chat with each other, explaining why we picked our answer. Then we have round 2. Remember to hit E if you are confused.

Interactive Questions

To battle the pandemic, backup health care workers were called in to work in hospitals A, B, and C. Hospital A began with 50, hospital B began with 40, and hospital C began with 60. Hospital A received an additional 10, B received an additional 25, and C received an additional 5. What is the average number of workers at hospitals in this sample (A, B, and C)?

- A: 53
- B: 63
- C: 42
- D: 32

Interactive Questions

Suppose a sample of students record the duration of their sleep each night for a week, and gather the data at the end. What kind of data is this?

- A: Quantitative discrete
- B: Qualitative or categorical
- C: Quantitative continuous
- D: Variable

Interactive Questions

Fall 2019 Country of Citizenship and Student Count					
Argentina	0	Hungary	0	Saudi Arabia	8
Australia	3	India	3	South Africa	1
Bhutan	1	Ireland	1	Spain	0
Brazil	1	Italy	3	Sweden	1
Cambodia	1	Japan	1	Turkey	1
Canada	2	Kosovo	1	Ukraine	1
China	12	Netherlands	1	United Kingdom	5
Egypt	0	New Zealand	1	Zimbabwe	1
France	1	Russia	1		
Germany	0	Rwanda	1		

What kind of data is represented in the population above? (There may be more than one answer).

- A: Quantitative discrete
- B: Qualitative or categorical
- C: Quantitative continuous
- D: Variable

Interactive Questions

Fall 2019 Country of Citizenship and Student Count					
Argentina	0	Hungary	0	Saudi Arabia	8
Australia	3	India	3	South Africa	1
Bhutan	1	Ireland	1	Spain	0
Brazil	1	Italy	3	Sweden	1
Cambodia	1	Japan	1	Turkey	1
Canada	2	Kosovo	1	Ukraine	1
China	12	Netherlands	1	United Kingdom	5
Egypt	0	New Zealand	1	Zimbabwe	1
France	1	Russia	1		
Germany	0	Rwanda	1		

The total number of international students is 52 in the above table. What proportion of international students are from China?

- A: 12
- B: 12%
- C: 23
- D: 23%

Interactive Questions

Fall 2019 Country of Citizenship and Student Count					
Argentina	0	Hungary	0	Saudi Arabia	8
Australia	3	India	3	South Africa	1
Bhutan	1	Ireland	1	Spain	0
Brazil	1	Italy	3	Sweden	1
Cambodia	1	Japan	1	Turkey	1
Canada	2	Kosovo	1	Ukraine	1
China	12	Netherlands	1	United Kingdom	5
Egypt	0	New Zealand	1	Zimbabwe	1
France	1	Russia	1		
Germany	0	Rwanda	1		

The total number of international students is 52 in the above table. What proportion of international students are from Europe?

- A: 15%
- B: 23%
- C: 50%
- D: 12

Other forms of Qualitative Data, Sampling

Other forms of Qualitative Data

1. Types of categories
2. Overlapping and non-overlapping categories, missing data
3. Activity on Pareto charts
4. Sampling strategies

Qualitative Data

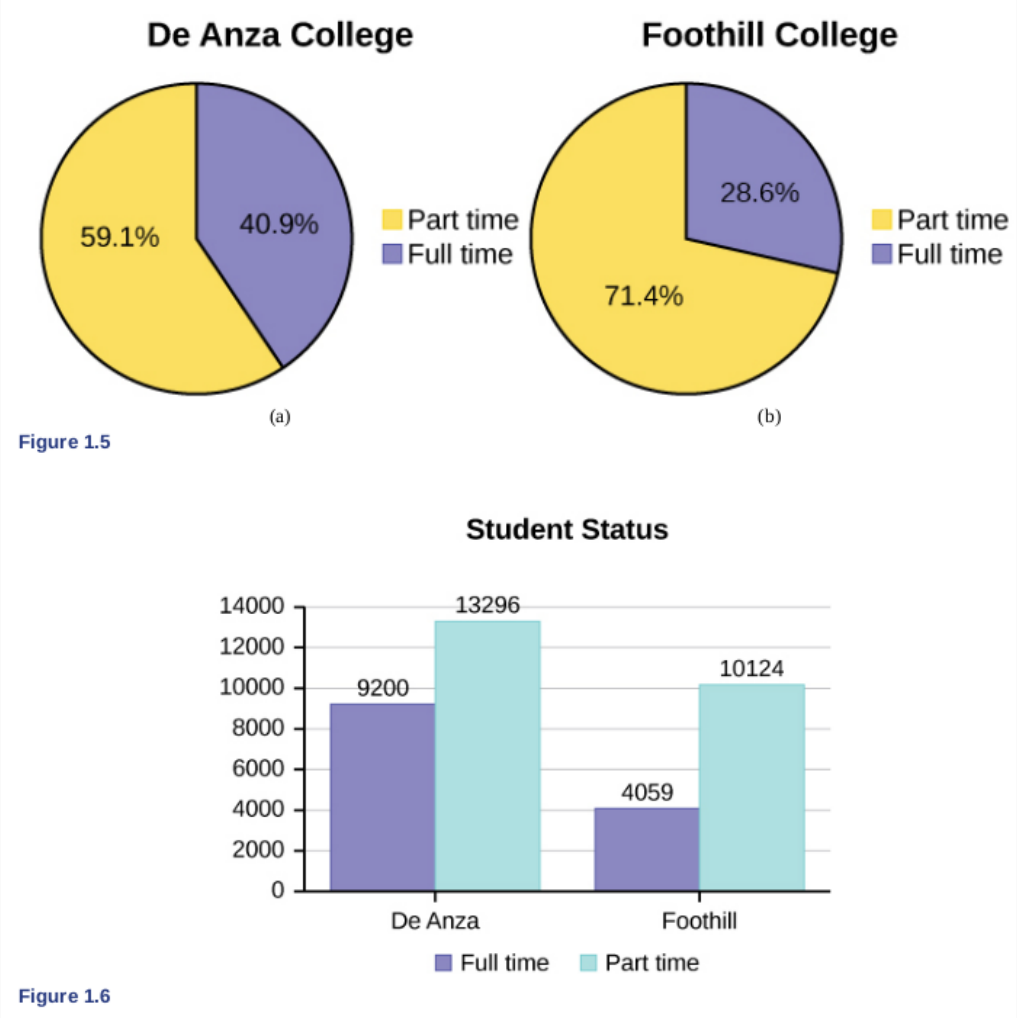

Figure 2: Two types of qualitative data representation.

Qualitative Data

1. **Mulitple categories**: Categories can overlap, so when calculating proportions, the percentatge can sum to greater than 100 percent.
   - Proportion of first-year students who are female: 56%
   - Proportion of students who are male: 44%
   - Proportion of students who do not live on campus: 46%
2. **Missing categories**: Categories do not always capture every feature of a population.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

Table 1.4 Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

Qualitative Data

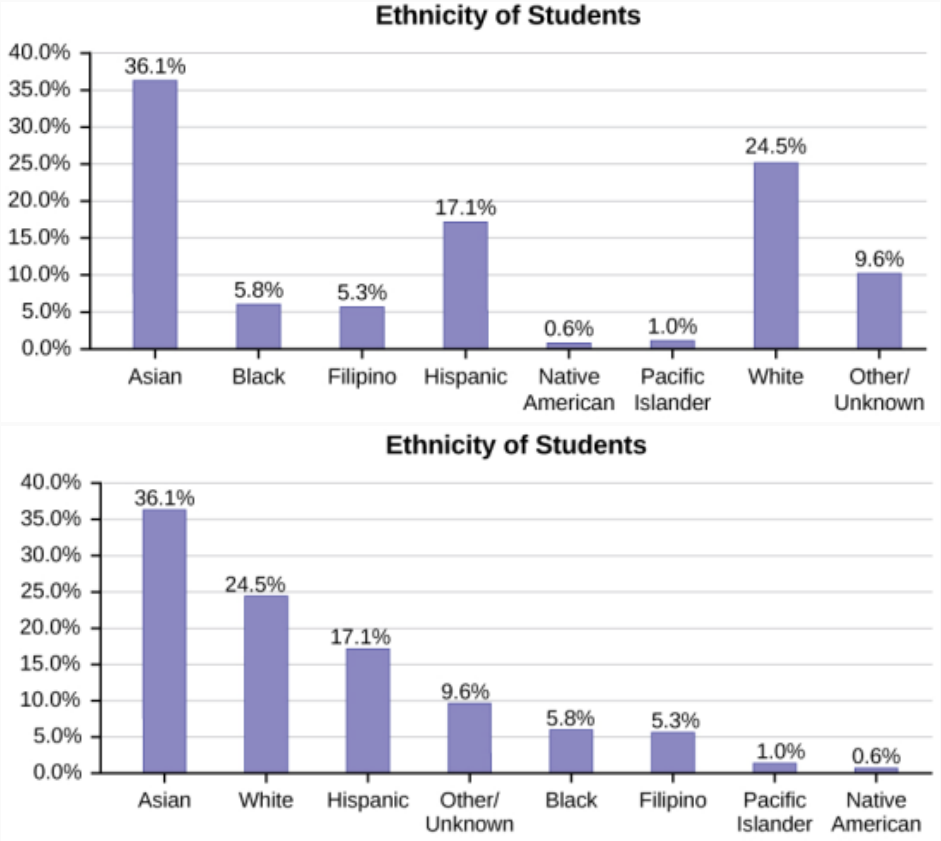

Figure 3: A Pareto chart is a bar graph that is ordered greatest to least. This can sometimes illuminate an effect that wasn't obvious.

Qualitative Data

Let's create a Pareto chart from the Whittier College factbook.

Whittier College		
Race/Ethnicity	UG Fall 2019	%
American Indian/Alaskan Native	7	0.4%
Asian	124	7.0%
Black/African-American	85	4.8%
Hawaiian/Pacific Islander	5	0.3%
Hispanic/Latino	908	51.1%
Non-resident alien/International	59	3.3%
Two or more Races	131	7.4%
Unknown	16	0.9%
White	441	24.8%
Total	1776	

Figure 4: The demographic breakdown of Whittier College undergraduate self-reported ethnicity data from 2019-20.

- Open your copy of Excel, or LibreOffice Calc
- Make one column heading entitled **ETH**
- Make one column heading entitled **N**
- Copy the data in Fig. 4 into your columns

Qualitative Data

Let's create a Pareto chart from the Whittier College factbook.

Whittier College		
Race/Ethnicity	UG Fall 2019	%
American Indian/Alaskan Native	7	0.4%
Asian	124	7.0%
Black/African-American	85	4.8%
Hawaiian/Pacific Islander	5	0.3%
Hispanic/Latino	908	51.1%
Non-resident alien/International	59	3.3%
Two or more Races	131	7.4%
Unknown	16	0.9%
White	441	24.8%
Total	1776	

Figure 5: The demographic breakdown of Whittier College undergraduate self-reported ethnicity data from 2019-20.

- Sort the data according to **N**
- Click the menu "Insert," and insert a bar chart

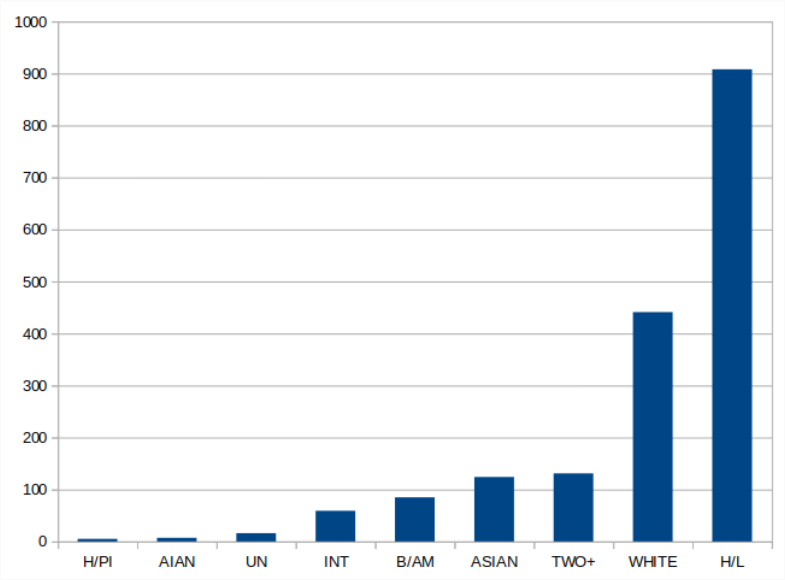

Figure 6: A Pareto chart of Whittier College ethnicity data. The sample is the UG as of Fall 2019.

Sampling Strategies

Sampling Strategies:

- Simple random ... how to generate random numbers? (Good project).
- Stratified sampling: pre-defined groups, choose proportionately at random from those groups (choosing at random from Depts.)
- Cluster sampling: pre-defined groups, but choose *the groups themselves* at random (choose random Depts.)
- Systematic sampling: select every n-th subject in the population to form the sample (assume some ordering, could be random)

Sampling Strategies

Cool example on random sampling. How can you measure the number of fish in a pond? Catch all the fish? No way! Use simple random sampling.

1. Catch n fish, and mark them.
2. Return one day later, and catch n fish again.
3. Assume the sample of fish is *simple random*. Why is this a good or bad assumption?
4. Measure the number m of marked fish, caught the second day.
5. The *proportion of total fish that are marked* is $p = m/n$.
6. But, $p = n/N$, where N is the total...
7. $N = n/p = n^2/m$. [1]

[1] Proceed to Hawai'i to tag great white sharks...

Sampling Strategies

Cool example on random sampling. How can you measure the number of fish in a pond? Catch all the fish? No way! Use simple random sampling.

1. With replacement ... keeps the population unchanged, randomness preserved.
2. Without replacement ... changes the population by 1 with each choice. Often more convenient and does not matter as long as the sample size is "large enough."

Sampling Strategies

Systematic sampling, special case: waveform data, time-series data. Systmatic trends cannot be observed by data sampled insufficiently systematically.

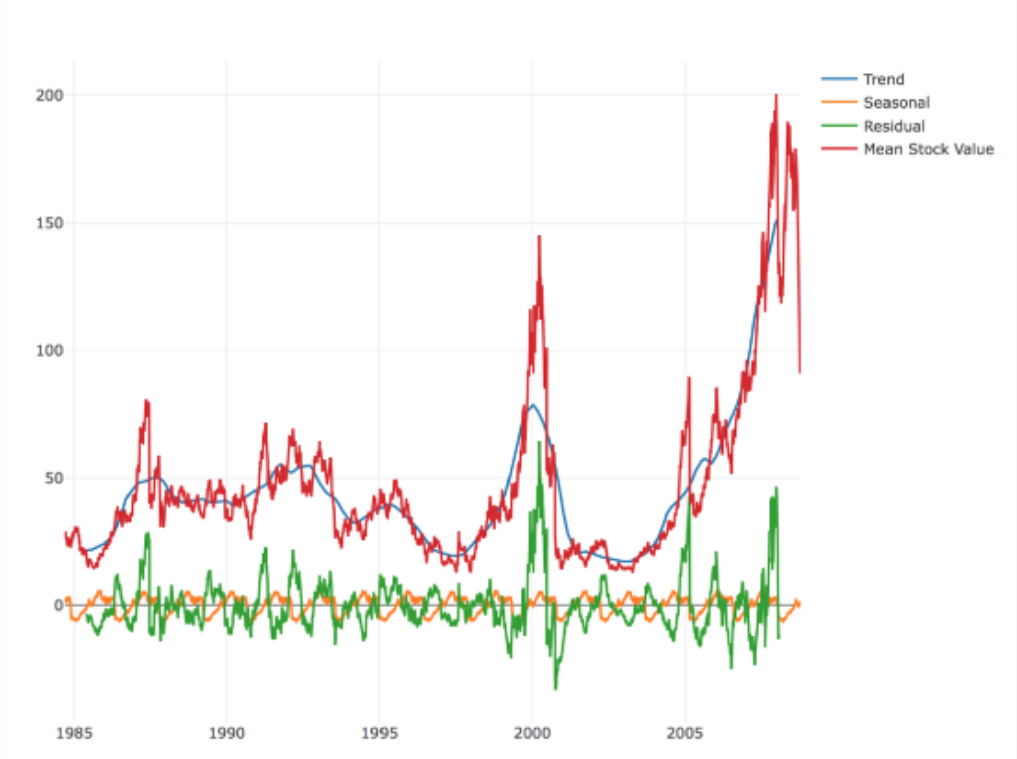

Figure 7: Example of stock price data.

Laboratory Activity

Go to the following link to watch a TED talk by the one, the only, Malcom Gladwell:

`https://www.ted.com/talks/malcolm_gladwell_choice_happiness_and_spaghetti_sauce?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare`

Answer the questions on the form entitled Laboratory Exercise 2 on Moodle for this week, and submit them via Moodle.

Interactive Questions

Interactive Questions

The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

- A: Cluster sampling
- B: Stratified sampling
- C: Simple random sampling
- D: Convenience sampling

Interactive Questions

A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:

- A: Simple random
- B: Systematic
- C: Stratified
- D: Cluster

Interactive Questions

Suppose you are working at a shoe company, and you sample the preferences of 5,000 previous customers to inform a new shoe design. If the response rate is 20 percent, and 300 responses indicated customers prefered leather to rubber insteps, what proportion of the random sample does this represent?

- 30 percent
- 20 percent
- 20 percent
- 10 percent

Frequency, Relative Frequency

Frequency, Relative Frequency

How frequently do data values occur in the sample? An activity to demonstrate the concept of frequency.

`https://phet.colorado.edu/en/simulation/plinko-probability`

Frequency, Relative Frequency

1. Begin with the Intro tab to this PhET simulation.
2. Use the controls at upper right to drop balls through the Plinko system.
3. Notice how they have *more or less* an equal chance of bouncing to the left or right at each level.
4. This behavior of being equally likely leads to a variation in the frequency with which objects land in the areas below.

Frequency, Relative Frequency

1. Now click the Lab tab at the bottom center of the screen.
2. Use the controls at right to increase the rows to 26.
3. Leave the "binary probability" set to 0.5, meaning 50% chance left, 50% chance right, at each interaction.
4. We have encountered how to calculate the mean $\bar{x}$ from a data sample. The mean position of all the balls is given at bottom right.
5. Use the play button at top right to drop 1,000 balls to form a data sample of positions at the bottom.
6. When you reach 1000, copy the *bins* (0, ... , 26) and the *frequencies* (the data in the bins) into Excel or LibreOffice Calc. Put the bin numbers in column A, and the frequencies the adjacent column B.

Frequency, Relative Frequency

1. Suppose you want to sum the frequencies. In a cell below the column of frequency data, type

   `=SUM(B1:B27)`

   and hit enter. This assumes that your data is in column B, in cells 1 through 27.
2. This should be the N value given by the PhET simulation (close to 1000).
3. In column C, cell 1, type

   `=B1/N`

   Instead of N, though, type the result for N found above.
4. Click on cell C1, and drag the little black square at the bottom right of the cell down until you reach C27 (repeats the calculation).

Frequency, Relative Frequency

1. Column C data are known as *relative frequencies.* They represent the probability of discovering a data result with the bin value.
2. If you use the SUM function to sum the relative frequencies, what result is obtained?
3. What about a "running total" of relative frequencies, to keep track of how quickly we approach 100% of the data? In cell D1, type

   `=C1`

   and in cell D2, type

   `=C2+D1`

   Click and drag to repeat the calculation until you reach the last cell. This trick adds the new C value to the running total. Create a plot of bin value versus column D.
4. Column D is the *cumulative frequency* of the data sample.

Mean: Definition 2

Let X represent a statistical variable, and x_i represent the data values. Suppose there are n_i instances of x_i, with only M unique values. The n_i are the *frequencies*, and we can write

$$\bar{x} = \frac{1}{N} \sum_{i}^{M} n_i x_i \tag{2}$$

Moving the $1/N$ inside the sum, we see that

$$\bar{x} = \sum_{i}^{M} \left(\frac{n_i}{N} \right) x_i = \sum_{i}^{M} f_i x_i \tag{3}$$

The fraction $f_i = n_i/N$ are the relative frequencies. (Professor: do an example).

Conclusion

Unit 0 Outline

1. Topics from Chapter 1: 1.1, 1.2, 1.3
    - What is a statistic?
    - Probability examples
    - Data and sampling
2. Topics from Chapter 2: 2.1 - 2.4, 2.5 - 2.8
    - Data visualization
    - Location of the data in numerical space
3. Topics from Chapter 3: 3.1, 3.2, 3.3
    - Two rules of probability