Homework 1: Causality and Expressions

Recommended Reading:

- What is Data Science
- Causality and Experiments
- Programming in Python

For all problems that you must write explanations and sentences for, you **must** provide your answer in the designated space. Moreover, throughout this homework and all future ones, please be sure to not re-assign variables throughout the notebook! For example, if you use `max_temperature` in your answer to one question, do not reassign it later on. Otherwise, you will fail tests that you thought you were passing previously!

Deadline:

This assignment is due **Monday, January 29th at 11:59pm**. No late assignments will be accepted.

Once your assignment is complete, do two things:

1. Change the file name by by adding "Submitted" at the end. For example, if I were a student submitting lab01, the file would initially be called "lab01Liz." After completing it, I would change the name to "lab01LizSubmitted."
2. Print your notebook as a pdf. Go to File -> Print -> save as pdf. Then upload the pdf to the corresponding assignment (lab/hw) on Canvas.

Directly sharing answers is not okay, but discussing problems with the instructor or with other students is encouraged. Refer to the policies page to learn more about how to learn cooperatively.

You should start early so that you have time to get help if you're stuck.

1. Scary Arithmetic

An ad for ADT Security Systems says,

> "When you go on vacation, burglars go to work [...] According to FBI statistics, over 25% of home burglaries occur between Memorial Day and Labor Day."

Do the data in the ad support the claim that burglars are more likely to go to work during the time between Memorial Day and Labor Day? Please explain your answer.

Note: You can assume that "over 25%" means only slightly over. Had it been much over, say closer to 30%, then the marketers would have said so.

link text*yes, this time between memorial day and labor day is 99 days- which is less than a quarter of a year but more than a quarter of all robberies happen during that time- so while this data doesn't provide any exact certainty, it does support the idea that robberies are slightly more likely to happen. Although it does only specify the robberies are between the two days- so they could be talking about the 266 days from labor day to memorial day in which case this data heavily suggests the opposite. *

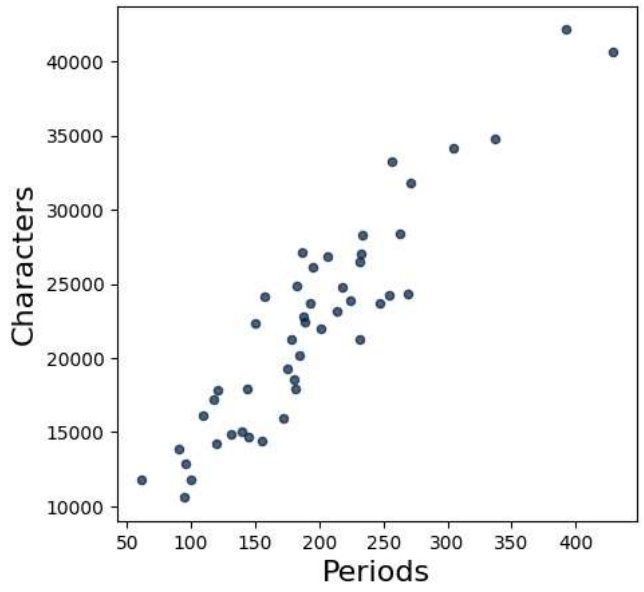
2. Characters in Little Women

In lecture, we counted the number of times that the literary characters were named in each chapter of the classic book, *Little Women*. In computer science, the word "character" also refers to a letter, digit, space, or punctuation mark; any single element of a text. The following code generates a scatter plot in which each dot corresponds to a chapter of *Little Women*. The horizontal position of a dot measures the number of periods in the chapter. The vertical position measures the total number of characters.

```
# This cell contains code that hasn't yet been covered in the course,
# but you should be able to interpret the scatter plot it generates.

from datascience import *
from urllib.request import urlopen
import numpy as np
%matplotlib inline

little_women_url = 'https://www.inferentialthinking.com/data/little_women.txt'
chapters = urlopen(little_women_url).read().decode().split('CHAPTER ')[1:]
text = Table().with_column('Chapters', chapters)
Table().with_columns(
    'Periods',    np.char.count(chapters, '.'),
    'Characters', text.apply(len, 0)
    ).scatter(0)
```



Question 1. Around how many periods are there in the chapter with the most characters? Assign either 1, 2, 3, 4, or 5 to the name `characters_q1` below.

1. 250
2. 390
3. 440
4. 32,000
5. 40,000

```
characters_q1 = 2
```

To discover more interesting facts from this plot, read Section 1.3.2 of the textbook.

3. Names and Assignment Statements

Question 1. When you run the following cell, Python produces a cryptic error message.

```
4 = 2 + 2
```

```
  File "<ipython-input-2-4c8b769209ad>", line 1
    4 = 2 + 2
    ^
SyntaxError: cannot assign to literal here. Maybe you meant '==' instead of '='?
```

SUGGEST FIX

Choose the best explanation of what's wrong with the code, and then assign 1, 2, 3, or 4 to `names_q1` below to indicate your answer.

1. Python is smart and already knows `4 = 2 + 2`.
2. `4` is already a defined number, and it doesn't make sense to make a number be a name for something else. In Python, "`x = 2 + 2`" means "assign `x` as the name for the value of `2 + 2`."
3. It should be `2 + 2 = 4`.
4. I don't get an error message. This is a trick question.

```
names_q1 = 2
```

Question 2. When you run the following cell, Python will produce another cryptic error message.

```
two = 3
six = two plus two
```

```
  File "<ipython-input-4-820d4d61e3dd>", line 2
    six = two plus two
                ^
SyntaxError: invalid syntax
```

SUGGEST FIX

Choose the best explanation of what's wrong with the code and assign 1, 2, 3, or 4 to `names_q2` below to indicate your answer.

1. The `plus` operation only applies to numbers, not the word "two".
2. The name "two" cannot be assigned to the number 3.
3. Two plus two is four, not six.
4. Python cannot interpret the name `two` followed directly by a name that has not been defined.

```
names_q2 = 4
```

Question 3. When you run the following cell, Python will, yet again, produce another cryptic error message.

```
x = print(5)
y = x + 2
```

```
5
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-6-94f783b16b3e> in <cell line: 2>()
      1 x = print(5)
----> 2 y = x + 2

TypeError: unsupported operand type(s) for +: 'NoneType' and 'int'
```

EXPLAIN ERROR

Choose the best explanation of what's wrong with the code and assign 1, 2, or 3 to `names_q3` below to indicate your answer.

1. Python doesn't want `y` to be assigned.
2. The `print` operation is meant for displaying values to the programmer, not for assigning values!
3. Python can't do addition between one name and one number. It has to be 2 numbers or 2 predefined names.

```
names_q3 = 2
```

4. Job Opportunities & Education in Rural India

A study at UCLA investigated factors that might result in greater attention to the health and education of girls in rural India. One such factor is information about job opportunities for women. The idea is that if people know that educated women can get good jobs, they might take more

care of the health and education of girls in their families, as an investment in the girls' future potential as earners. Without the knowledge of job opportunities, the author hypothesizes that families do not invest in women's well-being.

The study focused on 160 villages outside the capital of India, all with little access to information about call centers and similar organizations that offer job opportunities to women. In 80 of the villages chosen at random, recruiters visited the village, described the opportunities, recruited women who had some English language proficiency and experience with computers, and provided ongoing support free of charge for three years. In the other 80 villages, no recruiters visited and no other intervention was made.

At the end of the study period, the researchers recorded data about the school attendance and health of the children in the villages.

Question 1. Which statement best describes the *treatment* and *control* groups for this study? Assign either 1, 2, or 3 to the name `jobs_q1` below.

1. The treatment group was the 80 villages visited by recruiters, and the control group was the other 80 villages with no intervention.
2. The treatment group was the 160 villages selected, and the control group was the rest of the villages outside the capital of India.
3. There is no clear notion of *treatment* and *control* group in this study.

```
jobs_q1 = 1
```

Question 2. Was this an observational study or a randomized controlled experiment? Assign either 1, 2, or 3 to the name `jobs_q2` below.

1. This was an observational study.
2. This was a randomized controlled experiment.
3. This was a randomized observational study.

```
jobs_q2 = 2
```

Question 3. The study reported, "Girls aged 5-15 in villages that received the recruiting services were 3 to 5 percentage points more likely to be in school and experienced an increase in Body Mass Index, reflecting greater nutrition and/or medical care. However, there was no net gain in height. For boys, there was no change in any of these measures." Why do you think the author points out the lack of change in the boys?

Hint: Remember the original hypothesis. The author believes that educating women in job opportunities will cause families to invest more in the women's well-being.

Because the treatment did not involve boys- the author wanted to show that the treatment did have an affect and it wasn't just a difference between villages- the boys kind of acted as another control group.

5. Differences between Majors

Berkeley's Office of Planning and Analysis provides data on numerous aspects of the campus. Adapted from the OPA website, the table below displays the numbers of degree recipients in three majors in the academic years 2008-2009 and 2017-2018.

Major	2008-2009	2017-2018
Gender and Women's Studies	17	28
Linguistics	49	67
Rhetoric	113	56

Question 1. Suppose you want to find the **biggest** absolute difference between the numbers of degree recipients in the two years, among the three majors.

In the cell below, compute this value and call it `biggest_change`. Use a single expression (a single line of code) to compute the answer. Let Python perform all the arithmetic (like subtracting 49 from 67) rather than simplifying the expression yourself. The built-in `abs` function takes a numerical input and returns the absolute value. The built-in `max` function can take in 3 arguments and returns the maximum of the three numbers

```
biggest_change = max(abs(17-28), abs(49-67), abs(113-56))
biggest_change
```

```
57
```

Question 2. Which of the three majors had the **smallest** absolute difference? Assign `smallest_change_major` to 1, 2, or 3 where each number corresponds to the following major:

1: Gender and Women's Studies
2: Linguistics
3: Rhetoric

Choose the number that corresponds to the major with the smallest absolute difference.

You should be able to answer by rough mental arithmetic, without having to calculate the exact value for each major.

```
smallest_change_major = 1
smallest_change_major
```

```
1
```

Question 3. For each major, define the "relative change" to be the following: $\frac{\text{absolute difference}}{\text{value in 2008-2009}} * 100$

Fill in the code below such that `gws_relative_change`, `linguistics_relative_change` and `rhetoric_relative_change` are assigned to the relative changes for their respective majors.

```
gws_relative_change = (abs(17-28) / 17) * 100
linguistics_relative_change = (abs(49-67) / 49) * 100
rhetoric_relative_change = (abs(113-56) / 113) * 100
gws_relative_change, linguistics_relative_change, rhetoric_relative_change
```

```
(64.70588235294117, 36.734693877551024, 50.442477876106196)
```

Question 4. Assign `biggest_rel_change_major` to 1, 2, or 3 where each number corresponds to to the following:

1: Gender and Women's Studies
2: Linguistics
3: Rhetoric

Choose the number that corresponds to the major with the biggest relative change.

```
# Assign biggest_rel_change_major to the number corresponding to the major with the biggest relative change.
biggest_rel_change_major = 1
biggest_rel_change_major
```

```
1
```

6. Nearsightedness Study

Myopia, or nearsightedness, results from a number of genetic and environmental factors. In 1999, Quinn et al studied the relation between myopia and ambient lighting at night (for example, from nightlights or room lights) during childhood.

Question 1. The data were gathered by the following procedure, reported in the study. "Between January and June 1998, parents of children aged 2-16 years [...] that were seen as outpatients in a university pediatric ophthalmology clinic completed a questionnaire on the child's light exposure both at present and before the age of 2 years." Was this study observational, or was it a controlled experiment? Explain.

Observational- there was no treatment provided just data gathered

Question 2. The study found that of the children who slept with a room light on before the age of 2, 55% were myopic. Of the children who slept with a night light on before the age of 2, 34% were myopic. Of the children who slept in the dark before the age of 2, 10% were myopic. The study concluded that, "The prevalence of myopia [...] during childhood was strongly associated with ambient light exposure during sleep at night in the first two years after birth."

Do the data support this statement? You may interpret "strongly" in any reasonable qualitative way.

yes, there is a huge change in the people who developed myopia who had the light vs not.

Question 3. On May 13, 1999, CNN reported the results of this study under the headline, "Night light may lead to nearsightedness." Does the conclusion of the study claim that night light causes nearsightedness?

Yes, the data observed in the study supports this.

Question 4. The final paragraph of the CNN report said that "several eye specialists" had pointed out that the study should have accounted for heredity.

Myopia is passed down from parents to children. Myopic parents are more likely to have myopic children, and may also be more likely to leave lights on habitually (since the parents have poor vision). In what way does the knowledge of this possible genetic link affect how we interpret the data from the study?

This made me realize that the night light was a confounding factor- basicallly delegitimizing this information.

7. Submission

When you have completed your assignment,

1. Change the file name by by adding "Submitted" at the end. For example, if I were a student submitting lab01, the file would initially be called "lab01Liz." After completing it, I would change the name to "lab01LizSubmitted."
2. Print your notebook as a pdf. Go to File -> Print -> save as pdf. Then upload the pdf to the corresponding assignment (lab/hw) on Canvas.

This assignment is due Monday, January 29th at 11:59pm. No late assignments will be accepted.