hw01_master

September 12, 2019

1 Homework 1: Causality and Expressions

Please complete this notebook by filling in the cells provided. Before you begin, execute the following cell to load the provided tests.

[ ]:
```
# Don't change this cell; just run it.
# When you log-in please hit return (not shift + return) after typing in your␣
 ↪email
from client.api.notebook import Notebook
ok = Notebook('hw01.ok')
```

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 Thursday, September 5 at 11:59 P.M. You will receive an early submission bonus point if you turn in your final submission by Wednesday, September 4 at 11:59 P.M. Late work will not be accepted as per the policies page.

Directly sharing answers is not okay, but discussing problems with the course staff 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. Office hours are held Monday-Friday. The schedule appears on data8.org/fa19/office-hours.html.

Before continuing the assignment, select "Save and Checkpoint" in the File menu and then execute the `submit` cell below. The result will contain a link that you can use to check that your assignment has been submitted successfully. If you submit more than once before the deadline, we will only grade your final submission. If you mistakenly submit the wrong one, you can head to okpy.org and flag the correct version. There will be another `submit` cell at the end of the assignment when you finish!

[ ]:
```
_ = ok.submit()
```

1.1 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.

```
BEGIN QUESTION
name: q1
manual: True
```

SOLUTION: No. Labor Day is around 14 weeks after Memorial Day, so the period between them is a bit more than 25% of the year (14/52 ~= 26%). 25% of burglaries happening in 25% of the year does not imply a higher rate of burglary in the summer.

1.2 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

```
BEGIN QUESTION
name: q2_1
manual: false
```

[1]:
```
characters_q1 = 2 # SOLUTION
```

The test above checks that your answers are in the correct format. **This test does not check that you answered correctly**, only that you assigned a number successfully in each multiple-choice answer cell.

Question 2. Which of the following chapters has the most characters per period? Assign either 1, 2, or 3 to the name `characters_q2` below. 1. The chapter with about 60 periods 2. The chapter with about 350 periods 3. The chapter with about 440 periods

```
BEGIN QUESTION
name: q2_2
manual: false
```

[4]:
```
characters_q2 = 1 # SOLUTION
```

Again, the test above checks that your answers are in the correct format, but not that you have answered correctly.

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

1.3 3. Names and Assignment Statements

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

[1]:
```
4 = 2 + 2
```

```
  File "<ipython-input-1-4c8b769209ad>", line 1
    4 = 2 + 2
      ^
SyntaxError: can't assign to literal
```

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.

```
BEGIN QUESTION
name: q3_1
manual: False
```

[2]:
```
names_q1 = 2 # SOLUTION
```

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

[5]:
```
two = 3
six = two plus two
```

```
  File "<ipython-input-5-820d4d61e3dd>", line 2
    six = two plus two
                  ^
SyntaxError: invalid syntax
```

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 another name.

```
BEGIN QUESTION
name: q3_2
manual: False
```

[6]:
```
names_q2 = 4 # SOLUTION
```

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

[9]:
```
x = print(5)
y = x + 2
```

```
5
```

```
␣
↪---------------------------------------------------------------------------

TypeError                                 Traceback (most recent call␣
↪last)
```

```
    <ipython-input-9-94f783b16b3e> in <module>
      1 x = print(5)
----> 2 y = x + 2


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

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. What error message?

```
BEGIN QUESTION
name: q3_3
manual: false
```

[10]:
```
names_q3 = 2 # SOLUTION
```

1.4 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.

```
BEGIN QUESTION
name: q4_1
manual: false
```

[1]:
```
jobs_q1 = 1 # SOLUTION
```

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.

```
BEGIN QUESTION
name: q4_2
manual: false
```

[4]:
```
jobs_q2 = 2 # SOLUTION
```

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.

```
BEGIN QUESTION
name: q4_3
manual: true
```

SOLUTION: The lack in change of boys' well-being is evidence that the treatment, targeted at women, was indeed the reason for the observed effect. If the boys' well-being had improved as well, an alternative explanation for the observed effect would be that it was the result of an overall increase in well-being and prosperity in these villages (confounding factors), rather than the recruiting treatment.

1.5 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

Major	2008-2009	2017-2018
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.

```
BEGIN QUESTION
name: q5_1
manual: False
```

[2]:
```
biggest_change = max(abs(17 - 28), abs(49 - 67), abs(113 - 56)) # SOLUTION
biggest_change
```

[2]:
```
57
```

Use the cell above to test for formatting (in this case, that dissimilarity is a number)

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.

```
BEGIN QUESTION
name: q5_2
manual: False
```

[6]:
```
smallest_change_major = 1 # SOLUTION
smallest_change_major
```

[6]:
```
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.

```
BEGIN QUESTION
name: q5_3
manual: False
```

[10]:
```
"""# BEGIN PROMPT
gws_relative_change = (abs(...) / 17) * 100
"""; # END PROMPT
gws_relative_change = (abs(17 - 28) / 17) * 100 # SOLUTION NO PROMPT
linguistics_relative_change = (abs(49 - 67) / 49) * 100 # SOLUTION
rhetoric_relative_change = (abs(113 - 56) / 113) * 100 # SOLUTION
gws_relative_change, linguistics_relative_change, rhetoric_relative_change
```

[10]:
```
(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.

```
BEGIN QUESTION
name: q5_4
manual: False
```

[17]:
```
# Assign biggest_rel_change_major to the number corresponding to the major with␣
↪the biggest relative change.
biggest_rel_change_major = 1 #SOLUTION
biggest_rel_change_major
```

[17]:
```
1
```

1.6 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.

```
BEGIN QUESTION
name: q6_1
manual: True
```

SOLUTION: It was an observational study. The researchers didn't perform any intervention.

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.

```
BEGIN QUESTION
name: q6_2
manual: True
```

SOLUTION: Yes. There is a big difference in myopia rates between the groups.

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?

```
BEGIN QUESTION
name: q6_3
manual: True
```

SOLUTION: No. The study (as quoted above) claimed only an association.

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 they have poor vision). In what way does the knowledge of this possible genetic link affect how we interpret the data from the study?

```
BEGIN QUESTION
name: q6_4
manual: True
```

SOLUTION: If myopic parents are more likely to have myopic kids *and* leave the lights on at night, then myopic kids are more likely to have lights on at night. It is then reasonable to assume that myopic parents are a potential confounding factor that the observational study did not account for. However, we can still find the observed association even if there is no causal effect of night lights on child myopia.

1.7 7. Studying the Survivors

The Reverend Henry Whitehead was skeptical of John Snow's conclusion about the Broad Street pump. After the Broad Street cholera epidemic ended, Whitehead set about trying to prove Snow wrong. (The history of the event is detailed here.)

He realized that Snow had focused his analysis almost entirely on those who had died. Whitehead, therefore, investigated the drinking habits of people in the Broad Street area who had not died in the outbreak.

What is the main reason it was important to study this group?

1) If Whitehead had found that many people had drunk water from the Broad Street pump and not caught cholera, that would have been evidence against Snow's hypothesis.

2) Survivors could provide additional information about what else could have caused the cholera, potentially unearthing another cause.

3) Through considering the survivors, Whitehead could have identified a cure for cholera.

```
BEGIN QUESTION
name: q7_1
manual: False
```

[2]:
```
# Assign survivor_answer to 1, 2, or 3
survivor_answer = 1 # SOLUTION
```

Note: Whitehead ended up finding further proof that the Broad Street pump played the central role in spreading the disease to the people who lived near it. Eventually, he became one of Snow's greatest defenders.

1.8 8. Welcome Survey

Once you have submitted, please also complete the welcome survey in order to receive credit for homework 1.

Welcome survey is here: https://docs.google.com/forms/d/e/1FAIpQLSeOCIfpEbw_i5PdGSwxq_DeNBMDGF-7QnKcw2dYhyheoSQVxQ/viewform?usp=sf_link

1.9 9. Submission

Once you're finished, select "Save and Checkpoint" in the File menu and then execute the `submit` cell below. The result will contain a link that you can use to check that your assignment has been submitted successfully. If you submit more than once before the deadline, we will only grade your final submission. If you mistakenly submit the wrong one, you can head to okpy.org and flag the correct version. To do so, go to the website, click on this assignment, and find the version you would like to have graded. There should be an option to flag that submission for grading!

[ ]:
```
_ = ok.submit()
```

[ ]:
```
# For your convenience, you can run this cell to run all the tests at once!
import os
print("Running all tests...")
_ = [ok.grade(q[:-3]) for q in os.listdir("tests") if q.startswith('q') and␣
 ↪len(q) <= 10]
print("Finished running all tests.")
```