hw01

October 5, 2018

1 Homework 1: Causality and Expressions. Due Sunday, October 7th at 11:59pm

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

```
In [15]: # Don't change this cell; just run it.
         from client.api.notebook import Notebook
         ok = Notebook('hw01.ok')
         _ = ok.auth(inline=True, force=True)

=======================================================================
Assignment: Homework 1: Causality and Expressions
OK, version v1.13.11
=======================================================================


Open the following URL:

https://okpy.org/client/login/

After logging in, copy the code from the web page and paste it into the box.
Then press the "Enter" key on your keyboard.

Paste your code here: QPolPCBZLN1UxTCaWjsWjBxFWNcYWu
Successfully logged in as wec149@ucsd.edu
```

Reading: - Textbook chapters 1, 2, and 3

Deadline:

This assignment is due Sunday, October 7th at 11:59pm. You are given two slip days thoughout the quarter which can extend the deadline by one day. See the syllabus for more details. With the exception of using slip days, late work will not be accepted unless you have made special arrangements with your instructor.

Directly sharing answers is not okay, but discussing problems with the course staff or with other students is encouraged.

You should start early so that you have time to get help if you're stuck. As of the release of this assignment, office hours are held everyday except Sunday. A calendar with lab hour times and locations appears on https://sites.google.com/eng.ucsd.edu/dsc-10-fall-2018/staff-hours.

Once you're finished, you must do two things:

1.0.1 a. Turn into OK

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.

```
In [16]: _ = ok.submit()

<IPython.core.display.Javascript object>


<IPython.core.display.Javascript object>


Saving notebook... Saved 'hw01.ipynb'.
Submit... 100% complete
Submission successful for user: wec149@ucsd.edu
URL: https://okpy.org/ucsd/dsc10/fa18/hw01/submissions/4RwG7n
```

1.0.2 b. Turn PDF into Gradescope

Select File > Download As > PDF via LaTeX in the File menu. Turn in this PDF file into the respective assignement at https://gradescope.com/. If you submit more than once before the deadline, we will only grade your final submission

1.1 0. Integrity of Scholarship Agreement

Before starting the homework, please carefully read and fill out this integrity of scholarship agreement form. You will not recieve scores in this class until you submit the form. https://goo.gl/forms/NPWa9ZGQf4AbhIsM2.

1.2 1. Scary Arithmetic

An auto insurer, trying to promote more expensive insurance plans to infrequent drivers, uses the following quote from the National Highway Traffic and Safety Administration to make the claim that driving near home is more dangerous:

> "Of all traffic collisions that occur in the United States, approximately 69% occur within a 10 mile radius of home"

Does the data in the ad support the claim in that driving is more dangerous near one's home? If you could request an additional piece of data to better evaluate the claim, what would it be? Please explain your answer.

I would request for the stastistics of car collisions in my city only.

1.3   2. Characters in Great Expectations

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 another classic book Great Expectations. The horizontal position of a dot measures the number of periods in the chapter. The vertical position measures the total number of characters.

```
In [17]:  # 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 *
          import numpy as np
          % matplotlib inline


          great_expectations_file = "great_expectations.txt"
          chapters = open(great_expectations_file).read().split('Chapter ')[1:]
          text = Table().with_column('Chapters', chapters)
          Table().with_columns(
              'Periods',    np.char.count(chapters, '.'),
              'Characters', text.apply(len, 0)
              ).scatter(0)
```
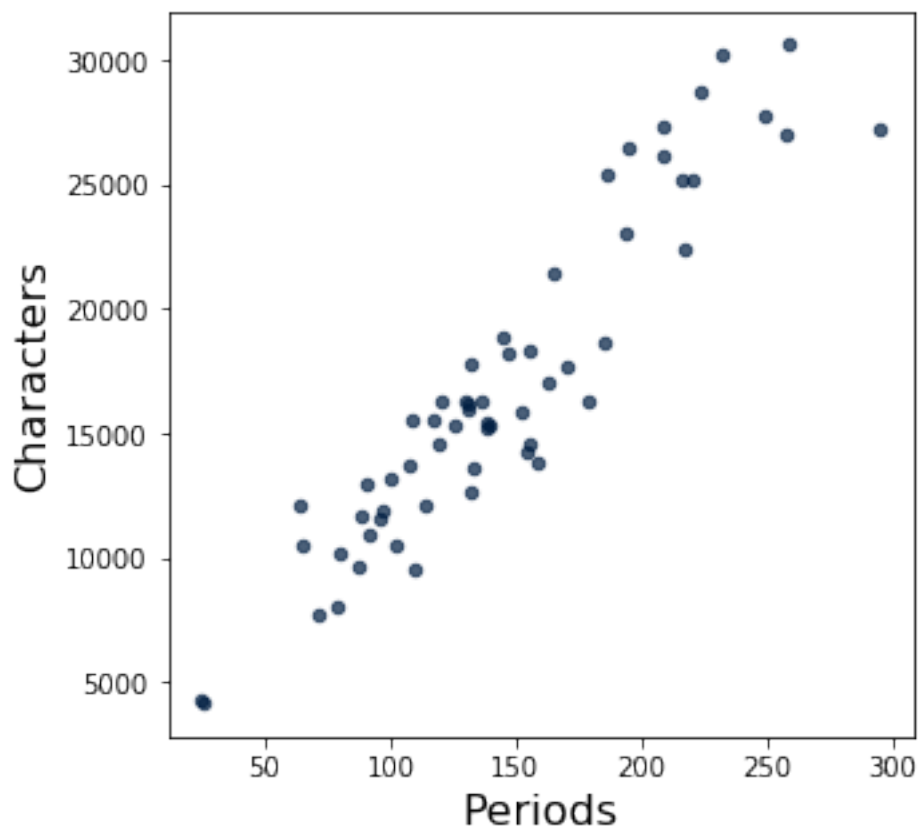
Question 1. About how many periods are in the chapter with the greatest number of characters? Assign either 1, 2, 3, 4, or 5 to the name `characters_q1` below.

1. 20
2. 100
3. 260
4. 300
5. 30100

```
In [18]: characters_q1 = 3
```

Question 2. About how many periods are in the chapter with the most characters per period? Assign either 1, 2, 3, 4, or 5 to the name characters_q2 below. 1. 20 2. 70 3. 110 4. 260 5. 300

```
In [19]: characters_q2 = 5
```

Check 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.

```
In [20]: _ = ok.grade('q2')

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Running tests

---------------------------------------------------------------------------
Test summary
    Passed: 1
    Failed: 0
[ooooooooook] 100.0% passed
```

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

1.4   3. Names and Assignment Statements

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

```
In [21]: 5 = 4 + 1.0


        File "<ipython-input-21-6f3bb90ba1b8>", line 1
      5 = 4 + 1.0
                ^
    SyntaxError: can't assign to literal
```

4

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. The left hand side is an `int`, while the right hand side is a `float`. It should be `5.0 = 4 + 1.0`.

2. It should be `1.0 + 4 = 5`.

3. In Python, "`x = 4 + 1.0`" means "assign x as the name for the value of `4 + 1.0`." Since 5 is a number, it doesn't make sense to make a number be a name for something else.

4. I don't get an error message. This is a trick question.

```
In [22]: names_q1 = 1
```

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

```
In [23]: two = 1
         two = two plus two


           File "<ipython-input-23-9ebecb0c0556>", line 2
         two = 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 statement has an undefined name.

3. Two plus two is four, not two.

4. The name "two" cannot be assigned to the number 1.

```
In [24]: names_q2 = 2
```

Check 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.

```
In [25]: _ = ok.grade('q3')
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Running tests

---------------------------------------------------------------------
Test summary
    Passed: 1
    Failed: 0
[ooooooooook] 100.0% passed
```

1.5   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.

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. There is no clear notion of *treatment* and *control* group in this study.

2. The treatment group was the 80 villages visited by recruiters, and the control group was the other 80 villages with no intervention.

3. The treatment group was the 160 villages selected, and the control group was the rest of the villages outside the capital of India.

```
In [26]: jobs_q1 = 2
```

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.

```
In [27]: jobs_q2 = 1
```

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?

Because the message was directed for females. So it is important to assess that the changes were not happening randomly

Check 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.

```
In [28]: _ = ok.grade('q4')
```

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Running tests


------------------------------------------------------------------------
Test summary
    Passed: 1
    Failed: 0
[ooooooooook] 100.0% passed
```

1.6   5. Differences between Universities

Question 1. Suppose you'd like to *quantify* how *dissimilar* two universities are, using three quantitative characteristics. The US Department of Education data on UCSD and UCLA describes the following three traits (among many others):

Trait	UCSD	UCLA
Average annual cost to attend ($)	14,775	13,572
Graduation rate (percentage)	87	91
Socioeconomic Diversity (percentage)	35	35

You decide to define the dissimilarity between two universities as the max of the absolute values of the 3 differences in their respective trait values.

For example if univeristy A cost $3,000 more than univeristy B, and university A had a 92 percent graduation rate while university B had a 89 percent graduation rate, and they had the same socioeconomic diversity percentage than the dissimilarity would be 3,000.

Using this method, compute the dissimilarity between UCSD and UCLA. Name the result `dissimilarity`. Use a single expression (a single line of code) to compute the answer. Let Python perform all the arithmetic (like subtracting 91 from 83) rather than simplifying the expression yourself. The built-in `abs` function takes absolute values.

```
In [29]: dissimilarity = max(abs(14775-13572),abs(87-91),abs(35-35))
         dissimilarity

Out[29]: 1203
```

Check that your answers are in the correct format. This test *does not* check that you answered correctly; only that you assigned an integer number successfully.

```
In [30]: _ = ok.grade('q5_1')

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Running tests


------------------------------------------------------------------------
Test summary
    Passed: 1
```

```
    Failed: 0
[oooooooooook]  100.0% passed
```

Question 2. The national average annual cost of college is \$16,424 (after financial aid). If the average annual cost of UCSD and UCLA were measured as a percentage of the national average rather than a dollar amount, what would be the dissimilarity between these two universities? Assign either 1, 2, 3 or 4 to the name `universities_q2` below.

1. 0.07
2. 4.0
3. 7.32
4. 12.03

```
In [31]: # universities_q2 = abs(14775/16426-13572/16426)*100
         # universities_q2
         universities_q2 = 3
```

Check that your answer is in the correct format. This test *does not* check that you answered correctly; only that you assigned a number successfully to `universities_q2`.

```
In [32]: _ = ok.grade('q5_2')

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Running tests

---------------------------------------------------------------------------
Test summary
    Passed: 1
    Failed: 0
[oooooooooook]  100.0% passed
```

1.7   6. More Children Living at Home?

A USA Today article from 2006 includes this sentence: "Since 1970, the percentage of people ages 18 to 34 [in the United States] who live at home with their family increased 48%, from 12.5 million to 18.6 million, the Census Bureau says."

Question 1. The word "percentage" isn't used correctly in the context of the rest of the sentence. What word should replace it?

Amount

Question 2. In at most 3 sentences, describe whether changes in the US population are relevant to the data in the USA Today article.

The US population changed approximately about the same amount. Meaning that the US population grew approximately uniformilly throughout the years

1.8   7. News Articles - Concussions

The following is an excerpt from a news article on the effects of recent legislation intended to prevent concussions in athletes.

"Since 2014, all 50 states and the District of Columbia have passed laws to protect young athletes against traumatic brain injury (TBI). Washington State was the first in 2009. Most of the laws require athletes with suspected concussions to stop playing until a doctor clears them to return. Coaches, players, and parents must also receive yearly education about concussions. Between fall 2005 and spring 2016, student athletes reported about 2.7 million concussions. Of those, 89 percent were new and 11 percent were repeat injuries. In 2005, nearly 150,000 initial concussions were reported. The number jumped to more than 360,000 by 2016. After concussion laws were introduced, however, repeat injuries fell dramatically, from about 14 percent of all concussions in 2005 to roughly 7 percent in 2016." What is the most likely explanation for the fact that initial concussions nearly tripled from 135,000 in 2005 to 360,000 in 2016?

Throughout the years, the total population size of athletes are most likely to have increased significantly. The most likely explanation for this is that the total number of athletes tripled, leading to tripling the number of concussions

1.9   8. News Articles - Eating Alone

The following is an excerpt from a news article linking the practice of eating alone to metabolic syndrome, a medical condition closely linked to obesity. "For the study, nearly 8,000 South Korean adults were asked how often they ate alone. The researchers compared their responses to health data adjusting for factors like age, lifestyle choices, education levels, job status and exercise. The found that men who often ate alone had a 45% increased risk of obesity and a 64% risk of developing metabolic syndrome. But women who ate alone at the same rate were only 29% more likely have metabolic syndrome than those who always dined with someone else." Does this study indicate that eating alone is a cause of metabolic syndrome? Why or why not?

Not necessarily. Firstly, the survey did not give information about the risk of developing metabolic syndrome of men that do not eat alone. So we cannot compare the 64% with any other data. And secondly, 29% increase of metabolic syndrome on women is not significant enough for a the study to be meaningful. It might be just a coincidence that those women were more likely to get the metabolic syndrome.

```
In [33]: _ = ok.submit()

<IPython.core.display.Javascript object>


<IPython.core.display.Javascript object>


Saving notebook... Saved 'hw01.ipynb'.
Submit... 100% complete
Submission successful for user: wec149@ucsd.edu
URL: https://okpy.org/ucsd/dsc10/fa18/hw01/submissions/APoK4O
```

1.10   Don't forget to submit your PDF to Gradescope!