hw01

September 6, 2024

[35]:

```python
# Initialize Otter
import otter
grader = otter.Notebook("hw01.ipynb")
```

1 Homework 1: Causality and Expressions

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

Recommended Readings:

- 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!

Note: This homework has hidden tests on it. That means even though tests may say 100% passed, it doesn't mean your final grade will be 100%. We will be running more hidden tests for correctness once everyone turns in the homework.

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.

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 to Labor Day."

Question 1 Do the data in the ad support the claim that burglars are more likely to go to work during the time between Memorial Day to Labor Day? Please explain your answer.

Hints: 1. 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.

2. In the U.S., Memorial Day is the last Monday of May and Labor Day is the first Monday of September. About what percentage of the days in a year fall between those two holidays?

These burgalaries happen 3 out of 12 months of the year or around 100 days. When I calculate doing 100/365 * 100 I get the percenteage of 27.4% This goes to show that the data in the ad does not strongly support the claim more likely to go to work during this period. If the amount of burgalaries were closer to 30% it'd be more suported to say the cliam.

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.

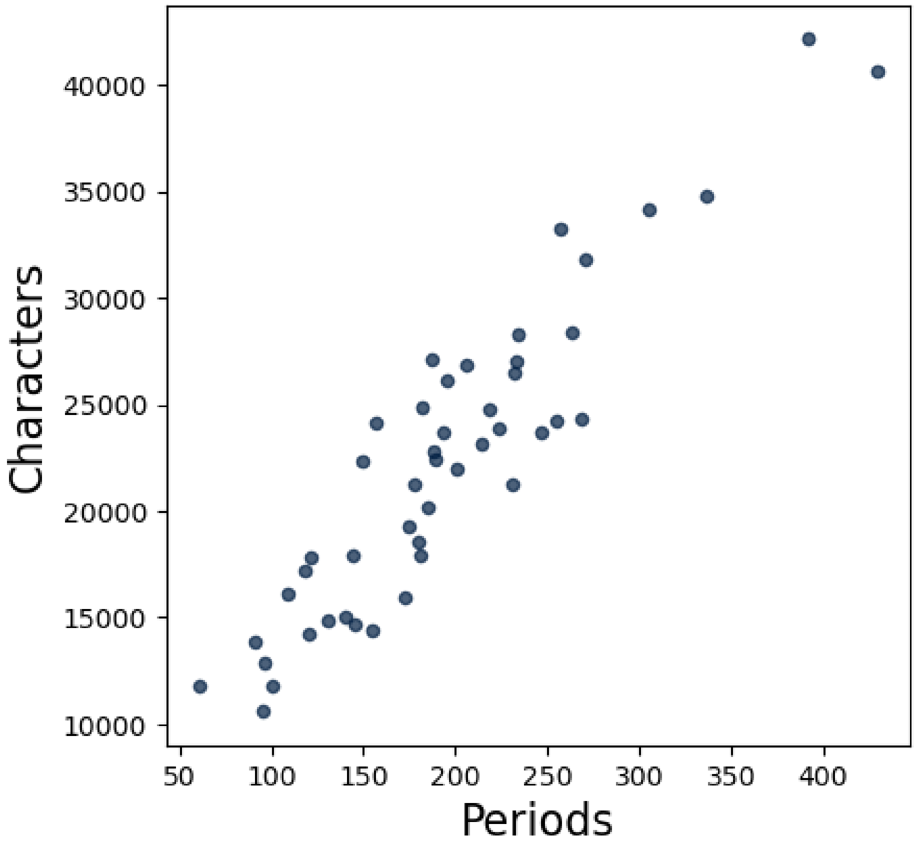
[34]:

```
# Just run this cell.

# 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

Note: If you run into a `NameError: name 'grader' is not defined` error in the autograder cell below (and in any assignment), please re-run the first cell at the very top of this notebook!

```
[3]: characters_q1 = 3
```

```
[4]: grader.check("q2_1")
```

```
[4]: q2_1 results: All test cases passed!
```

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

[32]:
```
characters_q2 = 3
```

[33]:
```
grader.check("q2_2")
```

[33]:
```
q2_2 results: All test cases passed!
```

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, check out Section 1.3.2 in the textbook.

1.3 3. Names and Assignment Statements

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

[40]:
```
4 = 2 + 2
```

```
  Cell In[40], line 1
    4 = 2 + 2
    ^
SyntaxError: cannot assign to literal here. Maybe you meant '==' instead of '='
```

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. In Python, it's a rule that the `=` sign must have a variable name to its left, and `4` isn't a variable name.
3. It should be `2 + 2 = 4`.
4. I don't get an error message. This is a trick question.

[30]:
```
names_q1 = 2
```

[31]:
```
grader.check("q3_1")
```

[31]:
```
q3_1 results: All test cases passed!
```

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

[49]:

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

```
  Cell In[49], 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. The name `plus` isn't a built-in operator; instead, addition uses `+`.

[28]:

```
names_q2 = 4
```

[29]:

```
grader.check("q3_2")
```

[29]:

```
q3_2 results: All test cases passed!
```

Question 3. Run the following cell.

[25]:

```
x = 2
y = 3 * x
x = 4
```

What is `y` after running this cell, and why? Choose the best explanation and assign 1, 2, 3, or 4 to `names_q3` below to indicate your answer.

1. `y` is equal to 6, because the second `x = 4` has no effect since `x` was already defined.
2. `y` is equal to 6, because `x` was 2 when `y` was assigned, and 3 * 2 is 6.
3. `y` is equal to 12, because `x` is 4 and 3 * 4 is 12.
4. `y` is equal to 12, because assigning `x` to 4 will update `y` to 12 since `y` was defined in terms of `x`.

[26]:

```
names_q3 = 2
```

[27]:

```
grader.check("q3_3")
```

[27]:

```
q3_3 results: All test cases passed!
```

1.4 4. 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 number of degree recipients in three majors in the 2008-2009 and 2017-2018 academic years.

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 number 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.

```
[23]: num_diff_gender_and_women_studies= abs(17-28)
      num_diff_linguistics= abs(49-67)
      num_diff_rhetoric= abs(113-56)




      biggest_change = max(abs(num_diff_gender_and_women_studies),␣
        ↪abs(num_diff_linguistics), abs(num_diff_rhetoric))
      biggest_change
```

```
[23]: 57
```

```
[24]: grader.check("q4_1")
```

```
[24]: q4_1 results: All test cases passed!
```

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.

[21]:

```
num_diff_gender_and_women_studies= abs(17-28)
num_diff_linguistics= abs(49-67)
num_diff_rhetoric= abs(113-56)

1 == num_diff_gender_and_women_studies
2 == num_diff_linguistics
3 == num_diff_rhetoric

smallest_change_major =  min(num_diff_gender_and_women_studies,␣
 ↪num_diff_linguistics, num_diff_rhetoric)
smallest_change_major
```

[21]:

```
11
```

[22]:

```
grader.check("q4_2")
```

[22]:

```
q4_2 results: All test cases passed!
```

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.

[19]:

```
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
```

[19]:

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

[20]:

```
grader.check("q4_3")
```

[20]:

```
q4_3 results: All test cases passed!
```

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.

[17]:

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

```
1 == gws_relative_change
2 == linguistics_relative_change
3 == rhetoric_relative_change

biggest_rel_change_major = max(gws_relative_change,␣
 ↪linguistics_relative_change, rhetoric_relative_change)
biggest_rel_change_major
```

[17]:

```
64.70588235294117
```

[18]:

```
grader.check("q4_4")
```

[18]:

```
q4_4 results: All test cases passed!
```

1.5 5. 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

[ ]:

```

```

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 the following: "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? Why or why not? You may interpret "strongly" in any reasonable qualitative way.

The data does seem to support this statment because the study shows a higher susceptibility of myopia in children who sleep with a room light or night light on (55% and 34) compared to children who slept in the dark(10%). The difference in these percenteages is very notable supporting the statement that there is a strong correlation to light exposure during sleep and myopia.

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? **Hint:** Look back as the quote in question 2 to see what language the study used in the conclusion.

The conclusion od the study does not claim that night light causes nearsightness. Instead the study mentions how myopia was "strongly associated" with ambient light exposure. This wording suggests that the correlation does not mean its the definite cause. The study only highlights the association but it does not blatantly state that the findings in the study is the cause of near sightedness.

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). What do we call a third variable, like heredity in this case that is related to both the explanatory and response variable? In what way does the knowledge of this possible genetic link affect how we interpret the data from the study? For full credit, this answer should relate to your answer to question 1, as well as your answer to question 3.

We call the third variable the confounding variable. THis means that the genetics(inherited myopia) could be influencing both the light exposure and the development fo nearsightedness. Since myopic parents are more likely to have myopic kids and might also use night lights its hard to know if the myopia is caused by light exposure or just gentics. This makes it harder to interpret the data because the study only hows a correlation but not proof of causation that the light caused myopia. This ties back to question 1 where I noted that the study was observational and question 3 where I discused how correlation doesn’t mean causation.

1.6 6. 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? Assign either 1, 2, or 3 to the name `survivor_answer` below.

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.

[15]:
```
survivor_answer = 1
```

[16]:
```
grader.check("q6_1")
```

[16]:
```
q6_1 results: All test cases passed!
```

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

1.7 Congratulations! You're done with Homework 1!

Be sure to run all of the cells and the grader checks and verify that they all pass, then choose **Download as PDF via LaTeX** from the **File** menu, as well as **Download as Notebook** from the **File** menu, correctly name your files, and submit the two files on **canvas**.