Data 8 Spring 2016 HW01, Due: 5pm Thursday, January 28

Please write your answers in the space provided. You can write on a printed copy or fill in the blanks with a PDF editor such as Acrobat Reader. When you're done, upload a scanned copy to Gradescope (`gradescope.com`). If you print and write on the assignment, the easiest way to upload is to take a picture of each page with your phone. You should have already received a signup email from Gradescope; if you have trouble submitting your work, contact the course staff (preferably through Piazza, and well before the assignment deadline). If you cannot submit online, you may submit a printed copy in office hours. You can find blank printed copies outside of 781 Soda and in office hours. This assignment is due 5pm Thursday, January 28. You will receive an early submission bonus point if you turn it in by 5pm Wednesday, January 27.

You are welcome to use data8.berkeley.edu to try out Python expressions. Directly sharing answers is not okay, but discussing problems with course staff or students is encouraged.

List your collaborators here:

Problem 1 Well, If the Census Bureau Says It's True...

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."

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

  1(a):

- There are potentially many causal factors contributing to this increase. What single numerical factor would you guess is the most important? (By "numerical factor" we just mean something that is quantifiable and that could cause the observed increase.) Feel free to include other sources of data to support your explanation if you want, but please keep your answer to the space provided.

  1(b):

Answer:

(a) The word "percentage" should be replaced with "number". 12.5 million and 18.6 million are clearly numbers of people in the US, not percentages.

(b) The US population increased from 1970 to 2006, so the *rate* of young adults living at home with their families could have stayed about the same, but the US simply has more people now. In fact, according to the US Census Bureau (see, for example, `http://www.census.gov/popest/data/historical/index.html`), the population increased by about 45% in that time period, so this change accounts for most of the increase.

Note: The inspiration for this problem was a blog post by John M. Grohol: `http://psychcentral.com/blog/archives/2006/03/16/bad-statistics-usa-today/`. The article itself is by Sharon Jayson: `http://usatoday30.usatoday.com/news/nation/2006-03-16-failure_x.htm`.

Problem 2 PROGRESA

A 2001 study by two Berkeley researchers investigated a Mexican anti-poverty initiative called *PROGRESA*. (Your textbook describes a related study on *PROGRESA*, but not the same study.) *PROGRESA* gave money to poor families, but required as a condition for payment that the families participate in a range of programs to improve their health and education. The amount of money *PROGRESA* gave to participating families was substantial, increasing participating families' income by a factor of roughly 1/3 on average.

PROGRESA eventually covered families in around 50,000 Mexican communities, but the program's managers initially evaluated it using a smaller group of 505 communities selected to have similar poverty levels. Out of those 505 communities, the managers randomly chose 320 to receive *PROGRESA* benefits in the first two years of the program (1998 to 2000). Families in the remaining 185 communities (out of the 505 selected) received no benefits in the first two years.

The Berkeley researchers focused on the impacts of *PROGRESA* on health. They measured various health-related outcomes for the 505 communities and found that children in the 320 communities that received *PROGRESA* benefits had substantially lower average rates of illness than children in the other 185 communities when the two-year period ended.

- Did this analysis have a treatment group and a control group? If so, describe the two groups.

  2(a):

- Was this an observational study or a randomized controlled experiment?

  2(b):

- Does the study provide evidence that being in a community receiving *PROGRESA* benefits is *associated* with a lower rate of childhood illness?

  2(c):

- Does the study provide evidence that being in a community receiving *PROGRESA* benefits *causes* a lower rate of childhood illness?

  2(d):

- Most families eligible for *PROGRESA* participated in the required health programs, but around 3% did not. (Those families were unwilling or unable to participate. This study did not investigate the reasons for non-participation.) Suppose we limit our study to the 320 eligible communities, and within those communities we compare the health outcomes of the families that participated (and received *PROGRESA* money) with those that did not (and therefore received no *PROGRESA* money). Would that analysis constitute an observational study, a randomized controlled experiment, or neither?

  2(e):

Answer:

(a) Yes. The 320 communities that received *PROGRESA* benefits were the treatment group, and the remaining 185 communities were the control group.

(b) This was a randomized controlled experiment. (*PROGRESA* benefits were introduced to a treatment group and not to a control group, and the two groups were chosen randomly from a single population of communities.)

(c) Yes. The problem setup said this directly.

(d) Yes. *PROGRESA* benefits were randomly assigned to communities, so evidence of association between *PROGRESA* benefits and lower rates of childhood illness is also evidence that *PROGRESA* benefits *caused* the lower rates of childhood illness.

(e) That analysis would be an observational study. Among the families in the 320 communities, non-participation in the program was not assigned randomly. Given that *PROGRESA* was offered to all of these families, families who do not participate may differ from families who participate in ways that are not caused by *PROGRESA* benefits, which is a source of confounding.

Problem 3 How the Other Half Lives

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. He realized that Snow had focused his analysis almost entirely on those who had died. Whitehead, therefore, investigated the drinking habits of those who had survived. Why was it important to study this group? [Note: Far from disproving Snow's claim, Whitehead ended up finding further proof that the Broad Street pump played the central role in spreading the disease. Eventually, he became one of Snow's greatest defenders.]

3:

Answer: If the survivors also drank out of the Broad Street pump, it would diminish the force of Snow's claim that the pump was spreading the disease.

Problem 4 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.

- 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?

  4(a):

- 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.

  4(b):

- 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?

  4(c):

- 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. In what way do you think this fact might have affected the data?

  4(d):

Answer:

(a) The study was observational. A controlled experiment would *assign* the treatment (that is, ambient lighting at night) to some children and not to others. This study merely recorded whether the treatment had happened.

(b) Yes. The percents of myopic children in the three groups were 10%, 34%, and 55%. That's quite a varied set of percents, pointing to association.

(c) No. The study was observational, and thus subject to confounding. The researchers were very clear ("strongly associated") that their conclusion was not causal.

(d) Myopic parents are more likely to have trouble seeing in the dark, therefore more likely to leave lights on at night. They are also more likely to have myopic children because myopia is inherited. Thus the parents' eyesight might have been a confounding factor.

Problem 5 Ninety-Nine Bottles

After executing x=10, Yulin tries to find as many expressions as she can that evaluate to 99. For each line, write CORRECT in the answer space to the right of the line if the expression evaluates to 99. Otherwise, rewrite the expression, adding operators, commas, and parentheses so that it does evaluate to 99.

10 x − 1	5(a):
(x)(x) − 1	5(b):
11 * x − x + 1	5(c):
abs(x * x) − abs(9 − x)	5(d):
abs −99	5(e):
pow(x 2) − 1	5(f):

Answer:

(a) 10 * x − 1

(b) (x)*(x) − 1

(c) 11 * x − (x + 1)

(d) CORRECT

(e) abs(−99)

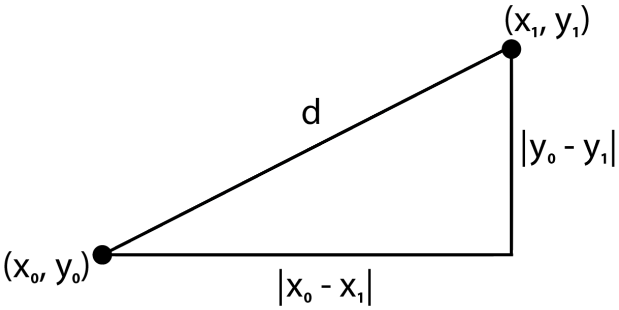
(f) pow(x, 2) − 1

Problem 6 Pythagoras

Suppose we have two 2-D points, (x_0, y_0) and (x_1, y_1). We wish to calculate the *Euclidean distance* between these two points, i.e. the length of line segment connecting them. Let's call that distance d. We forgot the distance formula, but we can re-derive it from the well-known *Pythagorean Theorem*, which says

$$a^2 + b^2 = c^2$$

for a right triangle with leg lengths a and b and hypotenuse length c. In our case, the hypotenuse length is d, the distance we seek, and the leg lengths are the horizontal and vertical distances between the points, which are $|x_0 - x_1|$ and $|y_0 - y_1|$. See the picture below.



Inserting these values into the Pythagorean equation (and dropping unnecessary absolute values due to squaring), we have

$$(x_0 - x_1)^2 + (y_0 - y_1)^2 = d^2$$

Solving for d gives

$$d = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$$

To square `x` in Python, you can write `x ** 2`. To take the square root of `x` in Python, you can write `x ** 0.5`.

Write a one-line Python expression that computes the Euclidean distance between the points. Assume that you are working in a notebook in which someone has already defined the Python variables `x0`, `y0`, `x1`, and `y1` to have the values x_0, y_0, x_1, and y_1, respectively.

6:

Answer: `((x0 - x1)**2 + (y0 - y1)**2) ** 0.5`

Problem 7 Johnny

Suppose that Johnny, another Data 8 student, is tasked with writing code to find the middle of three distinct values a, b, and c. (By "middle", we mean the value that is neither the maximum nor the minimum.) For example, given the values 4, 2, and 7, the code should give back 4. Johnny has a moment of inspired thought and rapidly types the following into his Jupyter notebook:

```
a + b + c - max(a, b, c) + min(a, b, c)
```

Is Johnny's code correct? Why or why not? If not, is there an easy fix?

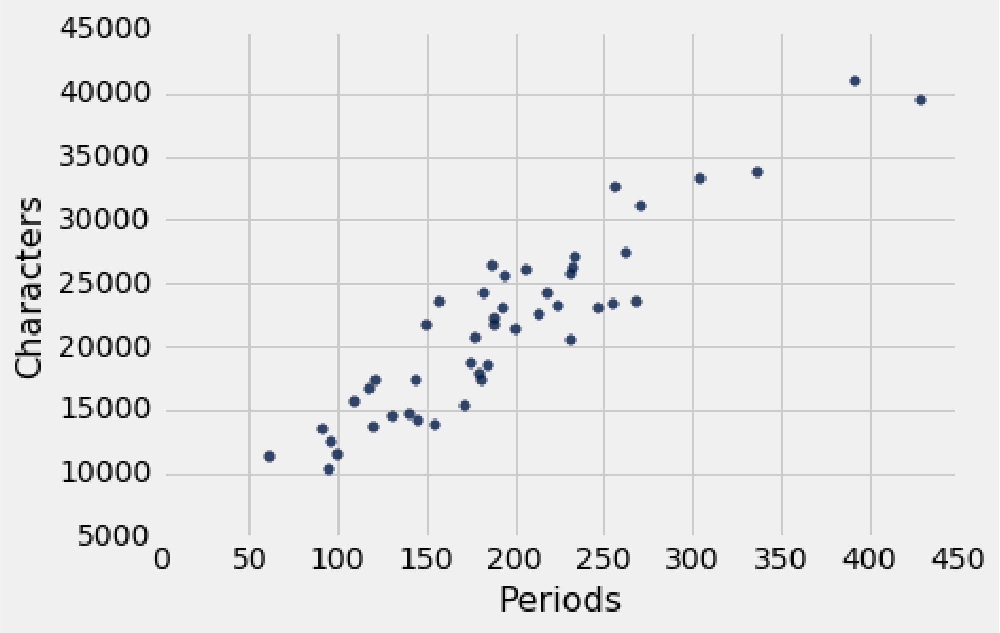
7:

Answer: No, Johnny's code is not correct. We need to subtract both the minimum and the maximum from the sum in order to leave just the third value. We can fix this by adding parentheses to change the order of operations:

```
a + b + c - (max(a, b, c) + min(a, b, c))
```

Problem 8 Jo

Below is the plot from Lecture 1 that compares the number of characters to the number of periods in each chapter of Little Women.



- About how many periods are in the chapter with the most characters per period?

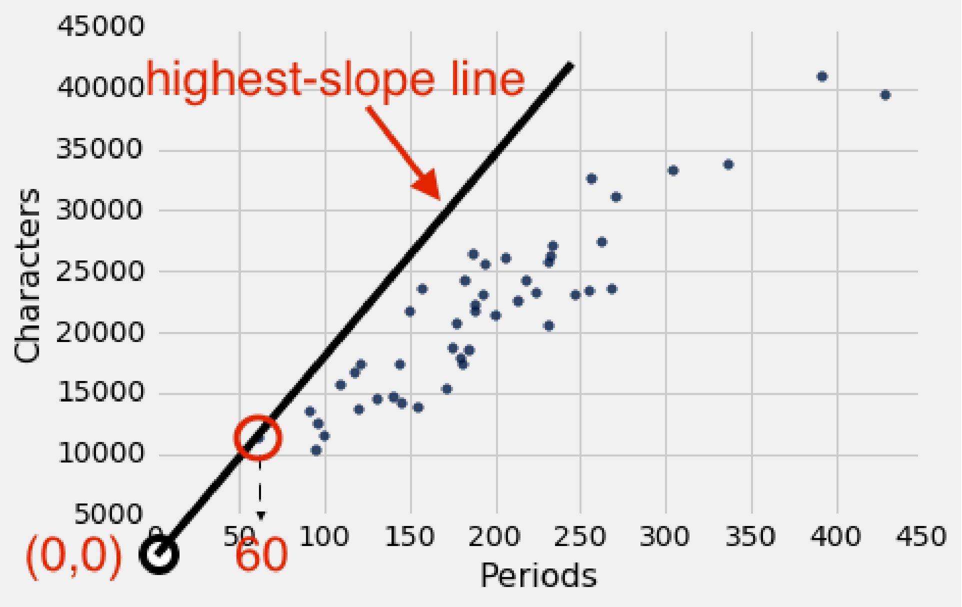
  8(a):

- About how many periods are in the chapter with the most characters?

  8(b):

Answer:

(a) About 60. The question amounts to asking which chapter has the greatest value of $\frac{\text{number of characters}}{\text{number of periods}}$ and then finding the number of periods in that chapter. That turns out to be the point at the far left of the plot, which has about 60 periods. You could discover this by calculating approximate values of $\frac{\text{number of characters}}{\text{number of periods}}$ for a few points. Or, geometrically, that point forms the highest-slope line between itself and the origin $(0, 0)$, since that slope is just $\frac{\text{number of characters}-0}{\text{number of periods}-0}$. Here is the picture:



(b) About 390. The chapter with the most characters is the highest point in the plot, and the horizontal value of that point is around 390.