Homework 1 - Suvir

February 22, 2022

Homework 1: Data Science for Everyone.

Name: Suvir Wadhwa , N-Number: N16395336

Question 1

(a) **Independent Variable:** Bad Air, **Dependent Variable:** Cholera Outbreak

(b) **Independent Variable:** Answering work emails late at night, **Dependent Variable:** Burnout

(c) **Independent Variable:** Social Media, **Dependent Variable:** exposure to news outlets

(d) **Independent Variable:** Sense of Belonging, **Dependent Variable:** Happiness

(e) **Independent Variable:** Apple a day, **Dependent Variable:** Doctor Away

Question 2

(a) **Confounder:** Stress: Stess leads to increased social media usage and irregular sleep patterns as impacts the minds regular functions.

(b) **Confounder:** Exercising: Exercising is also a factor that is said to improve mood as it boosts serotonin. It also leads you to go outside. Therefore, making it a confounder.

(c) **Confounder:** Having Great Wealth: Having great wealth enables you to have the income to purchase a private jet. It also enables you to recieve higher quality health care, having an impact on your lifespan.

(d) **Confounder:** Automobile Transport: Automobile transportaton drives up CO levels and has an impact on the climate. Automobile transport also gives travellers more disposable time, leading them to use their screens a lot more when travelling.

Question 3

(a) **Endogenous:** Exercise leads us to enjoy ourselves and release serotonin leading to happiness. However, when we are happy, we also are a lot more active and healthy, leading us to exercise. Increase in serotonin levels due to happiness causes us to exercise.

(b) **Not Endogenous:** Non endogenous as having a migrane has no impact in the barometric pressure of our surroundings.

(c) **Endogenous:** endogenous as being homeless increases the demand for homes, and as demand increases, prices increase.

(d) **Endogenous:** Volunteering causes people to participate in more local politics. However, participating in local politics leads use to get more involved in the community leading us to volunteer more. Getting involved in the community through politics causes more volunteering.

(e) **Not Endogenous:** Controversial Books tend to sell better. However, just because a book is selling well, it doesn't mean it's controversial.

Question 4

(a) **Prove**. All (good) theories can be disproven, no theories can be proven

(b) The most iconic masters of scale on the planet.

(c) Mitigation of Bias, prevents only one type of variable from impacting the findings of a study. in this case, only using one type of guest will lead to biased results and findings and won't disprove anything.

(d) Using the phrase genre-defining intends to claim that this pocast can define the genre. Howeever, there is not critera for this. It is just a claim without any withstand proof or data.

Question 5

(a) Underperforming Secondary School Students

(b) Students who reported higher levels of self-concept of school tasks were 17% less likely to comply. No significant effects of the treatment were observed on students' GPA, school motivation, hours spent on homework, or self-concept of school tasks. The treatment showed a negative effect on self-concept of leadership skills.

(c) Self-reflection on school behavior

(d) school performance, grade point average (GPA), school engagement, and self-concept.

(e) Yes, it is randomized as it is mentioned in the abstract: "This study used a randomized field experiment."

(f) The study is an experiment as the subjects are divided into control and non-control groups randomly.

(g) The implied control group is the group of underperforming students that take part in self-reflection of school behavior.

(h) A self reflection form that has questions un-related to school behavior. Hence, making the subjects assume they filled in the same form as the control group but not having the same impact on their thought process.

Question 6

(a) Complaince means to adhere to specific rules and standards. The standards may include regulations, rules, legislations, and several codes of conduct. In this case, treatment compliance refers to the students actions and response to the treatment. It aims to differentiate between students that followed the treatment thoroughly.

(b) **Null Hypothesis:** self-reflection on school behavior improves school performance Alternative Hypothesis: self-reflection on school behavior has no impact school performance

(c) The researches reject the Null Hypothesis, as seen towards the end of the abstract: "No significant effects of the treatment were observed on students' GPA, school motivation, hours spent on homework, or self-concept of school tasks."

(d) Practical challenges faced in this Study may include collecting data to calculate GPA, measuring school performance, keeping track of all of the subjects.

(e) The ethical challenges the researchers could have faced would most likely be related to privacy issues for the subjects, and discrimination between groups of students. To adhere to this, the researchers could sign confidentiality forms and keep student information private.

Question 7

[2]:
```
import numpy as np
import pandas

myarray = np.array([3114569, 3094026, 3022008, 2789513, 1325041, 1794401,␣
↪2730600])
print("Type:", type(myarray))
mean = myarray.mean()
print("Mean: ", mean)
int_mean = int(mean)
secondarray = np.array([2919797, 2945876, 2639986])
fullarray = np.append(myarray, secondarray)
print("Full Array:", fullarray)
dic = {"Date": ["Thursday, 2/10/22", "Wednesday, 2/9/22","Tuesday, 2/8/22"], ␣
↪"Total Estimated Ridership":["3,114,569", "3,094,026","3,022,008"], "% of␣
↪Comparable Pre-Pandemic Day":["54.9%", "55.5%","53.9%"] }
df = pandas.DataFrame(dic)
df.set_index("Date")
```

```
Type: <class 'numpy.ndarray'>
Mean:  2552879.714285714
Full Array: [3114569 3094026 3022008 2789513 1325041 1794401 2730600 2919797
2945876
 2639986]
```

[2]:

Date	Total Estimated Ridership	% of Comparable Pre-Pandemic Day
Thursday, 2/10/22	3,114,569	54.9%
Wednesday, 2/9/22	3,094,026	55.5%
Tuesday, 2/8/22	3,022,008	53.9%

(h) In this data set we are comparing trends in data by analysing current data with a previous sample. This enables us to get an understanding of the changes that have occured. In this case, the "percentage of comparable pre-pandemic day" values.

[3]:
```
print(df.dtypes)
```

```
Date                                object
```

```
Total Estimated Ridership           object
% of Comparable Pre-Pandemic Day    object
dtype: object
```

(i) In order to conduct further calculations, we would need to convert the data from string form to integer/float.

[ ]: