DS 111, **Prof Mittel**, **Prof Rudner**,

Homework 1

This homework is due on Wednesday, Sept. 25, 2024 by 9:00pm. Complete this assignment in your own notebook and then upload it as a single, fully executed PDF to Gradescope > Homework 1. Email submissions and/or submissions in any format other than PDF **will not be accepted**. You must clearly identify which question you are answering throughout your homework. If we cannot tell what question you are answering, we cannot give you credit.

This homework is worth 50 points. Late homework will be graded down. Incorrectly formatted assignments will be considered late until the correct version is uploaded.

Note that the course academic honesty policy applies to every homework, including this one.

---

1. **Confounding**. Association between two variables is straightforward to measure, but need not have a causal interpretation. As we have seen, confounders can create problems. Consider the following claims which are based on supporting associations (or lack thereof) in observational studies. For each claim:

   - State the treatment variable, X and outcome variable Y
   - State one plausible confounder, Z
   - Explain why Z is a confounder with respect to X and Y

   (a) **(4 points)** A pescatarian diet leads to lower risk of Alzheimer's disease.

   (b) **(4 points)** Going to the dentist more frequently doesn't cause a reduction in number of cavities.

   (c) **(4 points)** Vending machines in schools cause childhood obesity.

   (d) **(4 points)** Habitual sauna use increases longevity.

2. **Sleep** "Nightly sleep duration predicts grade point average in the first year of college" (Creswell et al., 2023) and answer the following questions.

   (a) **(3 points)** What are the units of analysis, explanatory variable (X) and outcome (Y) variables of the study?

   (b) **(1 points)** Who are the subjects that participated in the study? Do the subjects belong to different groups? What variables were measured for each subject?

   (c) **(2 points)** Was the study in (Creswell et al., 2023) an *experiment* or *observational study*? Why do you think the study was designed this way versus the alternative? What were the considerations?

   (d) **(2 points)** What one data limitation do the previous authors cite in previous studies? What type(s) of biases may be at play?

   (e) **(2 points)** Please name one confounder that the authors try to control for and one possible confounder that they do not.

   (f) **(1 points)** According to the paper, What is the minimum amount of sleep per night required to avoid a detrimental impact on GPA?

3. **Flipping a coin**.

   In this problem, you will experiment with flipping an "unbalanced" coin (one where the probabilities of heads and tails are not necessarily equal).

   (a) **(4 points)** Write a function `flip` that takes a float-valued argument p in the interval $[0, 1]$ that draws a single random number using `np.random.random` and returns a `bool` stating whether the random value was less than or equal to p.

   `raise` a `ValueError` if p is not between 0 and 1. Let's assume `True` represents "heads" and `False` represents "tails".

   Note: p represents the probability the returned value will be `True`.

   Print 4 the output of your function with $p = 0.26$. Regardless of the output, about how many heads would you *expect* to be admitted out of 4 flips?

   (b) **(5 points)** Write a new function `flipn` that takes as input a p between 0 and 1 (you do not have to validate whether p is between 0 and 1 this time) and a number n assumed greater than 0 that generates n boolean draws of `True` and `False` representing heads and tails respectively. Do not use any loops, only numpy!

   Generate $n = 1000$ coin flips with probability with $p = 0.67$ and report the total number of tails.

4. **Federalist papers authorship identification**

   The Federalist papers are a collection of 85 essays written by Alexander Hamilton, James Madison, and John Jay to promote the ratification of the United States constitution.

   The article "Inference in an authorship problem" from (Mosteller and Wallace, 1963) is a famous early example of how data science techniques can be applied to determine authorship of a paper from the text itself. Please refer to the provided article (available on Brightspace) and answer the following questions.

   (a) **(2 points)** What is the "discrimination problem" the authors are concerned with? Please provide a general definition of "discrimination" based on your intuition (it does not have to be exact!)

   (b) **(2 points)** What distinguishes "contextual" and "functional" words in the study?

   (c) **(2 points)** Following up on the previous question, what category of words do the authors choose to analyze and why?

The next set of questions will be based on the dataset analyzed and described in (Mosteller and Wallace, 1963). You will use the Python library `pandas` to complete each question.

Please refer to the dataset on JupyterHub, using the following path:

`shared/data/federalist-papers/function_word_counts.csv`

(i) **(1 points)** Please read in the dataset as a DataFrame and display the first 5 rows.

(ii) **(1 points)** The dataset consists of the counts of each of the words used in the study, the `total_word_count` (for all words) in the essay, and several identifier columns: `Essay`, `Author`, `Publication`, and `Date`.

Please print the number of occurrences of each value of the `Author` column.

(iii) **(2 points)** In the set of 85 federalist papers, please report the how many times the the "typical" Federalist paper uses the word "shall" (i.e., the average word count across all papers of the word "shall").

(iv) **(2 points)** Take the time to understand the two following functions: `set_index()` and `idxmax()` and use them both to find the author of the paper that uses the word "when" the most often.

(v) **(2 points)** The `groupby()` function is used to group rows with values in common for a subset of the variables in table and perform operations separately on each group.

Using `groupby`, compute the sum of the word counts for each author (the total number of times each author uses each word and the total number of words written per author should appear on each row of the output). Display the result.

More information: If you are interested in the Federalist papers dataset and how it was created, you may refer to the notebook on JupyterHub in `shared/data/federalist-papers/`.

References

J. David Creswell, Michael J. Tumminia, Stephen Price, Yasaman Sefidgar, Sheldon Cohen, Yiyi Ren, Jennifer Brown, Anind K. Dey, Janine M. Dutcher, Daniella Villalba, Jennifer Mankoff, Xuhai Xu, Kasey Creswell, Afsaneh Doryab, Stephen Mattingly, Aaron Striegel, David Hachen, Gonzalo Martinez, and Marsha C. Lovett. Nightly sleep duration predicts grade point average in the first year of college. *Proceedings of the National Academy of Sciences*, 120(8), February 2023. ISSN 1091-6490. doi: 10.1073/pnas.2209123120. URL `http://dx.doi.org/10.1073/pnas.2209123120`.

Frederick Mosteller and David L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275, June 1963. ISSN 0162-1459. doi: 10.2307/2283270. URL `http://dx.doi.org/10.2307/2283270`.