STA 141C - HW 5 - Benford's Law

Version 1.1

- See Canvas for due dates and grading rubric.
- Turn in a neatly typed 3-5 page report that answers the questions in clear English, using complete sentences.
- Use a format that Canvas can preview: pdf, html, Word are all fine. Raw Jupyter Notebook files and zip files do not work.
- Attach your code in an appendix, or as another file with extension `.txt`, so the graders can preview it on Canvas.
- Code must run, and support your answers.
- You may use any programming languages you like for this assignment.
- Use the cluster `gauss.cse.ucdavis.edu` for at least one step, and properly submit a job through `sbatch` or `srun`. For example:

  ```
  sbatch ./submit.sh
  ```

Overview

Which organizations are the 'most unusual' recipients of federal funds? We'll address this question by looking at the distributions of the first digits of the award amount for each recipient.

> Benford's law states that in many naturally occurring collections of numbers, the leading significant digit is likely to be small. - Wikipedia

If Benford's law applies to this set of numbers, then we expect that 1 will appear more frequently as a first digit than 9. Every recipient will have their own distribution.

We will compare these distributions against the reference distribution using Kullback-Leibler Divergence (KLD). You can use the definition:

$$D_{KL}(P, Q) = \sum_{x \in 1,\ldots,9} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

A large value of $D_{KL}(P, Q)$ indicates that P and Q are different, while a small value indicates they are similar.

Data

The data set is the same as used for hw4.

Steps

Your scripts should show do the following.

1. Remove duplicated transactions. For the purpose of this assignment, we'll consider a transaction to be a duplicate if `action_date, total_obligation, parent_recipient_unique_id` are all exactly the same.
2. Count the number of times the digits 1-9 appear as the first digit in `total_obligation` for each `parent_recipient_unique_id`. Ignore those that start with `-` or `0`.
3. Exclude from further analysis those recipients that have fewer than 100 transactions in total. (If there are too few, then we can't reliably say anything about the distribution of their digits.)
4. For every recipient, compute $D_{KL}(P, Q)$ with P representing the distribution of the digits 1-9 of that particular recipient and Q representing the distribution of the digits 1-9 across the entire data set.

Questions

Answer these questions in your report.

1. Explain how you computed the results, including
   - how you you used the cluster
   - which programming languages you used
   - any intermediate data files you produced
2. What is the actual distribution of the digits 1-9 in the whole data set? List them to three decimal places.
3. Plot the actual distribution of the first digits and Benford's theoretical distribution on the same graph. Benford's theoritical distribution is given by

$$Q(d) = \log_{10}(1 + \frac{1}{d})$$

   for $d \in \{1, 2, \ldots, 9\}$. Does this data seem reasonably close to Benford's? What is $D_{KL}(P, Q)$ for P the actual distribution, and Q given by Benford's formula above? What is $D_{KL}(P, Q)$ for P from a uniform distribution and Q given by Benford's formula above?
4. How many funding recipients were included in the final analysis? Summarize the distribution of the KLD scores among all these recipients.
5. Look back at the original data set for **one** of the recipients that has KLD greater than 2.5. Who is the recipient? What do you notice about the pattern of their transactions? Can you explain what happened, and why? You may incorporate information from third party sources.

Bootstrap

Bootstrap confidence intervals for KLD for each agency. Use $n = 1000$ bootstrap samples and compute a confidence interval as the $\alpha/2, 1 - \alpha/2$ quantiles of the bootstrap samples.

1. Summarize the width of the confidence intervals across the whole data set.
2. How does the width of the confidence interval relate to the number of transactions for each agency?

You may do this part on the cluster or on your local machine. Paralellism will help. Do not use our `staclass` partition for this, since students will be using this partition to do the main assignment. It's enough to remove the following line from your submit scripts:

```
#SBATCH --partition=staclass
```