PoCS What's The Story?

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394
University of Vermont, Fall 2022
Assignment 04
++?????++ Out of Cheese Error. Redo From Start

Due: Friday, September 23, by 11:59 pm
https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/assignments/04/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Assistant Deliverator: Dylan Casey (contact through Teams)
Office: The Ether
Office hours: See Teams calendar
Course website: https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse
Overleaf: LaTeX templates and settings for all assignments are available at
https://www.overleaf.com/project/631238b0281a33de67fc1c2b.

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you ~~conspired~~ collaborated.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The (evil) Deliverator uses (evil) Matlab.

Graduate students are requested to use LATEX (or related TEX variant). If you are new to LATEX, please endeavor to submit at least n questions per assignment in LATEX, where n is the assignment number.

Assignment submission:

Via Blackboard.

For Q1–5, you'll further explore the Google data set you examined earlier.

Q6 prepares for allotaxonometry.

1. Plot the complementary cumulative distribution function (CCDF).

2. Using standard linear regression, measure the exponent $\gamma - 1$ where γ is the exponent of the underlying distribution function. Identify and use a range of frequencies for which scaling appears consistent. Report the 95% confidence interval for your estimate.

   You will find two scaling regimes—please examine them both.

3. Size-rank plots:

   Using the alternate data set providing the raw word frequencies, plot word frequency as a function of rank in the manner of Zipf.

   Hint: you will not be able to plot all points (there are close to 14 million) so think about how to plot a subsample that still shows the full form.

4. Using standard linear regression, measure α, Zipf's exponent. Report the 95% confidence interval for your estimate.

   Again, you will find two regimes.

5. For each scaling regime, write down how γ and α are related (per lectures) and check how this expression works for your estimates here.

6. (3 + 3) **Baby name frequencies in the US:**

   Note: We will use this data set again in the next assignment.

   (a) Plot the Complementary Cumulative Frequency Distributions and Size-rank plots (Zipf's law) for the following:

      i. Baby girl names in 1952.
      ii. Baby boy names in 1952.
      iii. Baby girl names in 2002.
      iv. Baby boy names in 2002.

      Note that you will have counts that will make the Zipf distribution easy to plot straight away.

      From these counts, you will have to create the distributions N_k and $N_{\geq k}$.

   (b) As you did for the Google data set, fit regression lines and report values of γ and the Zipf exponent α.

      BUT: Only fit lines if fitting lines make sense!

      You may only have one region of scaling or zero.

   We will revisit these distributions in following assignments.

   Download:

   Data for 1880 through 2018:

   http://pdodds.w3.uvm.edu/permanent-share/pocs-babynames.zip (8.0M)

   Files:

   For each year, Zipf distribution of counts are stored in: `names-girlsYYYY.txt` and `names-boyYYYY.txt`.

For normalization to estimate rates, total number of births per year: `births_per_year.txt`. For this question, you do not need to determine rates, and this file is included for completeness.

For privacy, names with less than 5 counts are excluded.

The rare are legion and, for baby names, hidden.

Notes:

You should be able to re-use scripts from previous assignments.

Data is based on names registered through Social Security within the US.

Source:

Baby name dataset available here:
https://catalog.data.gov/dataset?tags=baby-names. Separate dataset for total births available here:
https://ssa.gov/oact/babynames/numberUSbirths.html.

7. Install Matlab on your machine of choice.

   We'll use Matlab in the next assignment.

   3 points for just getting this done and reporting faithfully that you did succeed.

   If you already have Matlab, then this is a freebie.

   For what we're going to do, there will be some roadblocks if you are not using a UNIX machine.

   Linux machines will work, Apple/Mac will work (because Mac OS is UNIX underneath), but Windows will present some problems.

   Note: For future happiness, we encourage students to use Python in general. Maybe R. Julia too. We're using a complicated piece of machinery that only exists fully realized in Matlab.