Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 6701, 6713, & a pretend number
University of Vermont, Fall 2024
"There's always money in the banana stand" ⌐
Assignment 04

P
o
C
S

What's
The
Story?

George Bluth ⌐, Arrested Development, Top Banana, S1E02.
Episode links: Wikipedia ⌐, IMDB ⌐, Fandom ⌐, TV Tropes ⌐.

Due: Monday, September 23, by 11:59 pm
https://pdodds.w3.uvm.edu/teaching/courses/2024-2025pocsverse/assignments/04/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Office: The Ether and/or Innovation, fourth floor
Office hours: See Teams calendar
Course website: https://pdodds.w3.uvm.edu/teaching/courses/2024-2025pocsverse
Overleaf: LATEX templates and settings for all assignments are available at
https://www.overleaf.com/read/tsxfwwmwdgxj.

Some guidelines:

1. Each student should submit their own assignment.

2. All parts are worth 3 points unless marked otherwise.

3. Please show all your work/workings/workingses clearly and list the names of others with
   whom you ~~conspired~~ collaborated.

4. We recommend that you write up your assignments in LATEX (using the Overleaf
   template). However, if you are new to LATEX or it is all proving too much, you may
   submit handwritten versions. Whatever you do, please only submit single PDFs.

5. For coding, we recommend you improve your skills with Python, R, and/or Julia.
   Please do not use any kind of AI thing. The (evil) Deliverator uses (evil) Matlab.

6. There is no need to include your code but you can if you are feeling especially proud.

Assignment submission:

Via Brightspace (which is not to be confused with the death vortex of the same name).

Again: One PDF document per assignment only.

For Q1–5, you'll further explore the Google data set you examined in assignment 3.

Q6 prepares for allotaxonometry.

Q7 is another piece for understanding heavy-tailed distributions.

39 points total.

1. (3 points)

   Plot the complementary cumulative distribution function (CCDF).

2. (6 points total: 3 points for each scaling regime)

   Using standard linear regression, measure the exponent $\gamma - 1$ where γ is the exponent of the underlying distribution function.

   You will find two "scaling regimes"—regions where scaling holds connected by a "break in scaling".

   For each scaling regime, Identify and use a range of word frequencies for which scaling appears consistent. (One range per scaling; eyeballing is enough.)

   Report the 95% confidence intervals for your estimates.

3. (3 points)

   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. (6 points: 3 points for each scaling regime)

   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 points)

   More on the peculiar nature of distributions of power law tails:

Consider a set of N samples, randomly chosen according to the probability distribution $P_k = ck^{-\gamma}$ where $k = 1, 2, 3, \ldots$.

Estimate $\min k_{\max}$, the approximate minimum of the largest sample in the system, finding how it depends on N.

Hint: we expect that on the order of 1 of the N samples to have a value of $\min k_{\max}$ or greater.

Hint—Some visual help on setting this problem up:
http://www.youtube.com/watch?v=4tqlEuXA7QQ

We are just touching on the deep world of extreme value theory. ⧉ Feel free to explore.

Notes:

- In a later assignment, we will test this scaling by (thoughtfully) sampling from power-law size distributions.

7. **Baby name frequencies in the US:**

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

   (a) (12 total points: $3 + 3 + 3 + 3$ for four sets of plots)

   Plot the Complementary Cumulative Frequency Distributions and Size-Rank plots (Zipf's way) for the following:

   i. Baby girl names in 1883 through 2023 inclusive, every 10 years.
   ii. Baby boy names in 1883 through 2023 inclusive, every 10 years.

   Notes:

   i. You will have counts that will make the Size-Rank distribution easy to plot straight away.
   ii. From these counts, you will have to create the distributions N_k and $N_{\geq k}$.
   iii. Each set will have 15 plots. Suggest tesselating them at 3x5, 5x3, or 4x4 with 1 missing.
   iv. You should produce 4 plots, each with 15 subplots.

   (b) (6 total points: 3 for γ estimates, 3 for α estimates)

   For the years 1963, 1993, and 2023 only:

   As you did for the Google data set, fit regression lines and report values of γ and the size-rank (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 2023:

http://pdodds.w3.uvm.edu/teaching/courses/2024-2025pocsverse/data/pocs-babynames.zip ⧉
(8.6 MB)

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 (use the processed data linked to above):

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 ⧉.