PoCS What's The Story?

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394
University of Vermont, Fall 2022
Assignment 03
Kingons, or possibly Queons

Due: Friday, September 16, by 11:59 pm
https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/assignments/03/
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.

All about power law size distributions (basic computations and some real life data from Google).

Note 1: Please do not use Mathematica, etc. for any symbolic work—you can do all of these calculations by hand. Yes you can!

Note 2: Otherwise, use whatever tools you like for the data analysis.

1. As in assignment 1, consider a random variable X with a probability distribution given by

$$P(x) = cx^{-\gamma},$$

where c is a normalization constant you determined in the first assignment, and $0 < a \leq x \leq b$. (a and b are the lower and upper cutoffs respectively.) Assume that $\gamma > 1$.

Note: For all answers you obtain for the questions below, replace c by the expression you obtained in the first assignment, and simplify expressions as much as possible.

Compute the nth moment of X which is in general defined as:

$$\langle x^n \rangle = \int_a^b x^n P(x)\, \mathrm{d}x$$

2. In the limit $b \to \infty$, how does the nth moment behave as a function of γ?

3. (a) Find σ, the standard deviation of X for finite a and b, then obtain the limiting form of σ as $b \to \infty$, noting any constraints we must place on γ for the mean and the standard deviation to remain finite as $b \to \infty$.

   Some help: the form of σ^2 as $b \to \infty$ should reduce to

   $$= \frac{(\gamma - c_1)}{(\gamma - c_2)(\gamma - c_3)^2} a^2$$

   where c_1, c_2, and c_3 are simple, meaningful constants to be determined (by you).

   (b) For the case of $b \to \infty$, how does σ behave as a function of γ, given the constraints you have already placed on γ? More specifically, how does σ behave as γ reaches the ends of its allowable range?

4. Drawing on a Google vocabulary data set (see below for links)

   (a) Plot the frequency distribution N_k representing how many distinct words appear k times in this particular corpus as a function of k.

   (b) Repeat the same plot in log-log space (using base 10, i.e., plot $\log_{10} N_k$ as a function of $\log_{10} k$).

5. Using your eyeballs, indicate over what range power-law scaling appears to hold and, estimate, using least squares regression over this range, the exponent in the fit $N_k \sim k^{-\gamma}$ (we'll return to this estimate later).

6. Compute the mean and standard deviation for the entire sample (not just for the restricted range you used in the preceding question). Based on your answers to the following questions and material from the lectures, do these values for the mean and standard deviation make sense given your estimate of γ?

Hint: note that we calculate the mean and variance from the distribution N_k; a common mistake is to treat the distribution as the set of samples. Another routine misstep is to average numbers in log space (oops!) and to average only over the range of k values you used to estimate γ.

The data for N_k and k (links are clickable):

- Compressed text file (first column $= k$, second column $= N_k$): https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/vocab_cs_mod.txt.gz
- Uncompressed text file (first column $= k$, second column $= N_k$): https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/vocab_cs_mod.txt
- Matlab file (`wordfreqs` $= k$, `counts` $= N_k$): https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/google_vocab_freqs.mat

The raw frequencies of individual words:

- https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/google_vocab_rawwordfreqs.txt.gz
- https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/google_vocab_rawwordfreqs.txt
- https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/google_vocab_rawwordfreqs.mat

Note: 'words' here include any separate textual object including numbers, websites, html markup, etc.

Note: To keep the file to a reasonable size, the minimum number of appearances is $k_{\min} = 200$ *corresponding to* $N_{200} = 48030$ *distinct words that each appear 200 times.*