Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Fall 2018
Assignment 2 • code name: Contemporary American Poultry

Dispersed: Friday, September 7, 2018.
Due: By 11:59 pm, Friday, September 14, 2018.
Some useful reminders:
Deliverator: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: pdodds+pocs@uvm.edu
Office hours: 10:15 am to 11:30 am, Tuesday and Thursday
Course website: http://www.uvm.edu/pdodds/teaching/courses/2018-08UVM-300
Bonus course notes: http://www.uvm.edu/pdodds/teaching/courses/2018-08UVM-300/docs/dewhurst-pocs-notes.pdf

---

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

Please obey the basic life rule: Never use Excel. Or any Microsoft product except maybe Xbox (which sadly will likely not help you here.)

Graduate students are requested to use LaTeX (or related TeX variant).

Email submission: PDF only! Please name your file as follows (where the number is to be padded by a 0 if less than 10 and names are all lowercase):
`CSYS300assignment%02d$firstname-$lastname.pdf` as in
`CSYS300assignment06michael-palin.pdf`

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

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

Use whatever tools you like for the data analysis.

1. Consider a random variable X with a probability distribution given by

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

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

(a) Determine c.

(b) Why did we assume $\gamma > 1$?

Note: For all answers you obtain for the questions below, please replace c by the expression you find here, and simplify expressions as much as possible.

2. Compute the nth moment of X.

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

4. For finite cutoffs a and b with $a \ll b$, which cutoff dominates the expression for the nth moment as a function of γ and n?

   Note: both cutoffs may be involved to some degree.

5. (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 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?

6. 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 $\log_{10} k$).

7. 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 in later assignments).

8. 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 $= N_k$): http://www.uvm.edu/pdodds/teaching/courses/2018-08UVM-300/docs/vocab_cs_mod.txt.gz
- Uncompressed text file (first column $= k$, second column $= N_k$): http://www.uvm.edu/pdodds/teaching/courses/2018-08UVM-300/docs/vocab_cs_mod.txt
- Matlab file (`wordfreqs` $= k$, `counts` $= N_k$): http://www.uvm.edu/pdodds/teaching/courses/2018-08UVM-300/docs/google_vocab_freqs.mat

The raw frequencies of individual words:

- http://www.uvm.edu/pdodds/teaching/courses/2018-08UVM-300/docs/google_vocab_rawwordfreqs.txt.gz
- http://www.uvm.edu/pdodds/teaching/courses/2018-08UVM-300/docs/google_vocab_rawwordfreqs.txt
- http://www.uvm.edu/pdodds/teaching/courses/2018-08UVM-300/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.

9. (a) A parent has two children, not twins, and one is a girl born on a Tuesday. What's the probability that both children are girls?

   See if you can produce both a calculation of probabilities and a visual explanation with shapes (e.g., discs and pie pieces).

   Once you have the answer, can you improve our intuition here? Why does adding the more detailed piece of information of the Tuesday birth change the probability from $1/3$?

   (Assume 50/50 birth probabilities.)

(b) Same as the previous question but we now know that one is a girl born on December 31. Again, what's the probability that both are girls?