Sausage inna bun

Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Fall 2014
Assignment 1 • code name:

Dispersed: Thursday, August 28, 2014.
Due: By start of lecture, 1:00 pm, Thursday, September 4, 2014.
Some useful reminders:
Instructor: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: peter.dodds@uvm.edu
Office hours: 2:30 pm to 3:45 pm on Tuesday, 12:30 pm to 2:00 pm on Wednesday
Course website: http://www.uvm.edu/~pdodds/teaching/courses/2014-08UVM-300

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

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

More 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. 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 $\log_{10} k$).
2. 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).
3. 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/2014-08UVM-300/docs/vocab_cs_mod.txt.gz
- Uncompressed text file (first column $= k$, second column $= N_k$): http://www.uvm.edu/~pdodds/teaching/courses/2014-08UVM-300/docs/vocab_cs_mod.txt
- Matlab file (`wordfreqs` $= k$, `counts` $= N_k$): http://www.uvm.edu/~pdodds/teaching/courses/2014-08UVM-300/docs/google_vocab_freqs.mat

The raw frequencies of individual words:

- http://www.uvm.edu/~pdodds/teaching/courses/2014-08UVM-300/docs/google_vocab_rawwordfreqs.txt.gz
- http://www.uvm.edu/~pdodds/teaching/courses/2014-08UVM-300/docs/google_vocab_rawwordfreqs.txt
- http://www.uvm.edu/~pdodds/teaching/courses/2014-08UVM-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.

4. 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.

5. Compute the nth moment of X. (Note: This is what Wikipedia rather rudely calls the "raw moment" or "crude moment.")

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

7. 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.

8. (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?