

**Principles of Complex Systems, CSYS/MATH 300**  
**University of Vermont, Fall 2011**  
**Assignment 1**

**Dispersed:** Friday, September 9, 2011.

**Due:** By start of lecture, 11:30 am, Thursday, September 15, 2011.

*Some useful reminders:*

**Instructor:** Peter Dodds

**Office:** Farrell Hall, second floor, Trinity Campus

**E-mail:** peter.dodds@uvm.edu

**Office hours:** 12:50 pm to 3:50 pm, Wednesday

**Course website:** <http://www.uvm.edu/~pdodds/teaching/courses/2011-08UVM-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).

---

All about power law 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.

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  $\log_{10} k$ ).
  - (c) 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  $N_k \sim k^{-\gamma}$ .
  - (d) Compute the mean and standard deviation for the sample. 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  $\gamma$ ?

You are encouraged but not required to use Matlab.

The data (links are clickable):

- Matlab file (wordfreqs =  $k$ , counts =  $N_k$ ):  
[http://www.uvm.edu/~pdodds/teaching/courses/2011-08UVM-300/docs/google\\_vocab\\_freqs.mat](http://www.uvm.edu/~pdodds/teaching/courses/2011-08UVM-300/docs/google_vocab_freqs.mat)
- Compressed text file (first column =  $k$ , second column =  $N_k$ ):  
[http://www.uvm.edu/~pdodds/teaching/courses/2011-08UVM-300/docs/vocab\\_cs\\_mod.txt.gz](http://www.uvm.edu/~pdodds/teaching/courses/2011-08UVM-300/docs/vocab_cs_mod.txt.gz)
- Uncompressed text file (first column =  $k$ , second column =  $N_k$ ):  
[http://www.uvm.edu/~pdodds/teaching/courses/2011-08UVM-300/docs/vocab\\_cs\\_mod.txt](http://www.uvm.edu/~pdodds/teaching/courses/2011-08UVM-300/docs/vocab_cs_mod.txt)

*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  $N_{200} = 48030$  unique words.*

2. 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.*

3. Compute the  $n$ th moment of  $X$ .
4. In the limit  $b \rightarrow \infty$ , how does the  $n$ th moment behave as a function of  $\gamma$ ?
5. For finite cutoffs  $a$  and  $b$  with  $a \ll b$ , which cutoff dominates the expression for the  $n$ th moment as a function of  $\gamma$  and  $n$ ?  
*Note: both cutoffs may be involved to some degree.*
6. (a) Find  $\sigma$ , the standard deviation of  $X$  for finite  $a$  and  $b$ , then obtain the limiting form of  $\sigma$  as  $b \rightarrow \infty$ , noting any constraints we must place on  $\gamma$  for the mean and the standard deviation to remain finite as  $b \rightarrow \infty$ .  
 (b) For the case of  $b \rightarrow \infty$ , how does  $\sigma$  behave as a function of  $\gamma$ ?