



Dispersed: Thursday, September 26, 2019.

Due: Friday, October 4, by 11:59 pm, 2019.

Some useful reminders:

Deliverator: Dr. Nick Allgaier (nicholas.allgaier@uvm.edu)

Assistant Deliverator: David Dewhurst (david.dewhurst@uvm.edu)

Office: To be disclosed

Office hours: To be emphatically determined through a democratic process

Course website: <http://www.uvm.edu/pdodds/teaching/courses/2019-08UVM-300>

Bonus course notes: <http://www.uvm.edu/pdodds/teaching/courses/2019-08UVM-300/docs/dewhurst-pocs-notes.pdf>

All parts are worth 3 points unless marked otherwise. Please show all your workings 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: 1. Please send to david.dewhurst@uvm.edu.

2. 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

Please submit your project's current draft in pdf format via email. Please use this file name format (all lowercase after CSYS):

CSYS300project-\$firstname-\$lastname-YYYY-MM-DD.pdf as in

CSYS300project-lisa-simpson-1989-12-17.pdf where the date is the date of submission (and not, say, your birthdate).

1. (3 + 3 + 3 + 3 + 3 + 3 pts) **Generalized entropy and diversity:**

For a probability distribution of $i = 1, \dots, n$ entities with the i th entity having probability of being observed p_i , Shannon's entropy is defined as [2]:

$H = -\sum_{i=1}^n p_i \ln p_i$. There are other kinds of entropies and we'll explore some aspects of them here.

Let's use the setting of words in a text (another meaningful framing is abundance of species in an ecology). So we have word i appearing with probability p_i and there are n words.

Now, a useful quantity associated with any kind of entropy is diversity, D [1]. Given a text T with entropy H , we define D to be the number of words in another hypothetical text T' which (1) has the same entropy, and (2) where all words appear with equal frequency $1/D$. In text T' , we have $p_i = 1/D$ for $i = 1, \dots, D$.

Diversity is thus a number, and behaves in number-like ways that are more intuitive to grasp than entropy. (Entropy is still the primary thing here.)

Determine the diversity D in terms of the probabilities $\{p_i\}$ for the following:

(a) Simpson concentration:

$$S = \sum_{i=1}^n p_i^2.$$

(b) Gini index:

$$G \equiv 1 - S = 1 - \sum_{i=1}^n p_i^2.$$

Please note any connections between diversity for the Simpson and Gini indices.

(c) Shannon's entropy:

$$H = - \sum_{i=1}^n p_i \ln p_i.$$

(d) Renyi entropy:

$$H_q^{(R)} = \frac{1}{q-1} \left(- \ln \sum_{i=1}^n p_i^q \right),$$

where $q \neq 1$.

(e) The generalized Tsallis entropy:

$$H_q^{(T)} = \frac{1}{q-1} \left(1 - \sum_{i=1}^n p_i^q \right),$$

where $q \neq 1$.

Please note any connections between diversity for Renyi and Tsallis.

(f) Show that in the limit $q \rightarrow 1$, the diversity for the Tsallis entropy matches up with that of Shannon's entropy.

2. (3 + 3 points) *Zipfarama via Optimization:*

Complete the Mandelbrotian derivation of Zipf's law by minimizing the function

$$\Psi(p_1, p_2, \dots, p_n) = F(p_1, p_2, \dots, p_n) + \lambda G(p_1, p_2, \dots, p_n)$$

where the 'cost over information' function is

$$F(p_1, p_2, \dots, p_n) = \frac{C}{H} = \frac{\sum_{i=1}^n p_i \ln(i + a)}{-g \sum_{i=1}^n p_i \ln p_i}$$

and the constraint function is

$$G(p_1, p_2, \dots, p_n) = \sum_{i=1}^n p_i - 1 \quad (= 0)$$

to find

$$p_j = e^{-1-\lambda H^2/gC} (j + a)^{-H/gC}.$$

Then use the constraint equation, $\sum_{j=1}^n p_j = 1$ to show that

$$p_j = (j + a)^{-\alpha}.$$

where $\alpha = H/gC$.

3 points: When finding λ , find an expression connecting λ , g , C , and H .

Hint: one way may be to substitute the form you find for $\ln p_i$ into H 's definition (but do not replace p_i).

Note: We have now allowed the cost factor to be $(j + a)$ rather than $(j + 1)$.

3. (3 + 3) Carrying on from the previous problem:

(a) For $n \rightarrow \infty$, use some computation tool (e.g., Matlab, an abacus, but not a clever friend who's really into computers) to determine that $\alpha \simeq 1.73$ for $a = 1$. (Recall: we expect $\alpha < 1$ for $\gamma > 2$)

(b) For finite n , find an approximate estimate of a in terms of n that yields $\alpha = 1$.

(Hint: use an integral approximation for the relevant sum.)

What happens to a as $n \rightarrow \infty$?

4. (3 + 3 + 3)

Estimating the rare:

Google's raw data is for word frequency $k \geq 200$ so let's deal with that issue now.


From Assignment 2, we had for word frequency in the range $200 \leq k \leq 10^7$, a fit for the CCDF of

$$N_{\geq k} \sim 3.46 \times 10^8 k^{-0.661},$$

ignoring errors.

- (a) Using the above fit, create a complete hypothetical N_k by expanding N_k back for $k = 1$ to $k = 199$, and plot the result in double-log space (meaning log-log space).
- (b) Compute the mean and variance of this reconstructed distribution.
- (c) Estimate:
 - i. the hypothetical fraction of words that appear once out of all words (think of words as organisms or tokens here),
 - ii. the hypothetical total number and fraction of unique words in Google's data set (think at the species or token level now),
 - iii. and what fraction of total words are left out of the Google data set by providing only those with counts $k \geq 200$ (back to words as organisms or tokens).

References

- [1] L. Jost. Entropy and diversity. *Oikos*, 113:363–375, 2006. [pdf](#) 
- [2] C. E. Shannon. A mathematical theory of communication. *The Bell System Tech. J.*, 27:379–423,623–656, 1948. [pdf](#) 