PoCS What's The Story?

Principles of Complex Systems, Vols. 1 & 2, CSYS/MATH 300 and 303
University of Vermont, Fall 2021
Assignment 07 • code name: Egg MacGuffin

Due: Friday, October 15, by 11:59 pm, 2021.
Relevant clips, episodes, and slides are listed on the assignment's page:
https://pdodds.w3.uvm.edu//teaching/courses/2021-2022principles-of-complex-systems//assignments/07/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Assistant Deliverator: Michael Arnold (contact through Teams)
Office: The Ether
Office hours: TBD
Course website:
https://pdodds.w3.uvm.edu//teaching/courses/2021-2022principles-of-complex-systems

---

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.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The Deliverator uses 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.

Please submit your project's current draft in pdf format via Blackboard by the same time specified for this assignment. For teams, please list all team member names clearly at the start.

---

1. (3 + 3)

   You've earlier determined the theoretical scaling of the large sample of a power-law size distribution as a function of sample number.

   Let's see how well things match up with simulations.

   For $\gamma = 5/2$, generate $n = 1000$ sets each of $N = 10$, 10^2, 10^3, 10^4, 10^5, and 10^6 samples, using $P_k = ck^{-5/2}$ with $k = 1, 2, 3, \ldots$

   How do we computationally sample from a discrete probability distribution?

Note: We examined some of these in class. See slides on power-law size distributions.

Hint: You can use a continuum approximation to speed things up. In fact, taking the exact continuum version from the first two assignments will work.

(a) For each value of sample size N, sequentially create n sets of N samples. For each set, determine and record the maximum value of the set's N samples. (You can discard each set once you have found the maximum sample.)

You should have $k_{\max,i}$ for $i = 1, 2, \ldots, n$ where i is the set number. For each N, plot the n values of $k_{\max,i}$ as a function of i.

If you think of n as time t, you will be plotting a kind of time series.

These plots should give a sense of the unevenness of the maximum value of k, a feature of power-law size distributions.

(b) Now find the average maximum value $\langle\rangle i k_{\max,i}$ for each N.

The steps again here are:

1. Sample N times from P_k;
2. Determine the maximum of the sample, $k_{\max}$;
3. Repeat steps 1 and 2 a total n times and take the average of the n values of $k_{\max}$ you have obtained.

Plot $\langle k_{\max}\rangle$ as a function of N on double logarithmic axes, and calculate the scaling using least squares. Report error estimates.

Does your scaling match up with your theoretical estimate for $\gamma = 5/2$?

How to sample from your power law distribution (and similarly upsetting things):

We now turn our problem of randomly selecting from this distribution into randomly selecting from the uniform distribution. After playing around a little, $k = 10^6$ seems like a good upper limit for the number of samples we're talking about.

Using Matlab (or some ghastly alternative), we create a cdf for P_k for $k = 1, 2, \ldots, 10^6$ and one final entry $k > 10^6$ (for which the cdf will be 1).

We generate a random number x and find the value of k for which the cdf is the first to meet or exceed x. This gives us our sample k according to P_k and we repeat as needed. We would use the exactly normalized $P_k = \frac{1}{\zeta(5/2)} k^{-5/2}$ where ζ is the Riemann zeta function.

Now, we can use a quick and dirty method by approximating P_k with a continuous function $P(z) = (\gamma - 1)z^{-\gamma}$ for $z \geq 1$ (we have used the normalization coefficient found in assignment 1 for $a = 1$ and $b = \infty$). Writing $F(z)$ as the cdf for $P(z)$, we have $F(z) = 1 - z^{-(\gamma-1)} = 1 - z^{-3/2}$. Inverting, we obtain $z = [1 - F(z)]^{-2/3}$.

We replace $F(z)$ with our random number x and round the value of z to finally get an estimate of k.

2. (3 + 3 + 3 + 3 + 3 + 3 pts) **Generalized entropy and diversity:**

   For a probability distribution of $i = 1, \ldots, n$ entities with the ith entity having probability of being observed p_i, Shannon's entropy is defined as [1]: $H = -\sum_{i=1}^{n} p_i \ln p_i$. 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 [2]. 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 $i = 1, \ldots, D$.

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

   where $q \neq 1$.

   (e) The generalized Tsallis entropy:
   $$H_q^{(\mathrm{T})} = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right),$$

where $q \neq 1$.

Please note any connections between diversity for Renyi and Tsallis.

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

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

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

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

where the 'cost over information' function is

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

and the constraint function is

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

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 $(j+1)$.

4. (3 + 3) Carrying on from the previous problem:

(a) For $n \to \infty$, 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 $n \to \infty$?

References

[1] C. E. Shannon. A mathematical theory of communication. *The Bell System Tech. J.*, 27:379–423,623–656, 1948. pdf

[2] L. Jost. Entropy and diversity. *Oikos*, 113:363–375, 2006. pdf