



**Due:** Friday, September 25, by 4:59 pm, 2020.

**Relevant clips, episodes, and slides** are listed on the assignment's page:

<http://www.uvm.edu/pdodds/teaching/courses/2020-08UVM-300/assignments/03/>

*Some useful reminders:*

**Deliverator:** Prof. Peter Sheridan Dodds (contact through Teams)

**Assistant Deliverator:** Michael Arnold (contact through Teams)

**Office:** The Ether

**Office hours:** Tuesdays, 12 to 12:50 pm; Wednesdays, 1:15 pm to 2:05 pm; Thursdays, 12 to 12:50 pm; all scheduled on Teams

**Course website:** <http://www.uvm.edu/pdodds/teaching/courses/2020-08UVM-300>

---

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.

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.

---

For Q1–5, you'll further explore the Google data set you examined in the second problem set.

For Q6–8, you'll explore random walks, touching the Central Limit Theorem and the first return problem. There and back again.

1. Plot the complementary cumulative distribution function (CCDF).
2. Using standard linear regression, measure the exponent  $\gamma - 1$  where  $\gamma$  is the exponent of the underlying distribution function. Identify and use a range of frequencies for which scaling appears consistent. Report the 95% confidence interval for your estimate.

You will find two scaling regimes—please examine them both.

3. Using the alternate data set providing the raw word frequencies, plot word frequency as a function of rank in the manner of Zipf.

**Hint:** you will not be able to plot all points (there are close to 14 million) so think about how to plot a subsample that still shows the full form.

- Using standard linear regression, measure  $\alpha$ , Zipf's exponent. Report the 95% confidence interval for your estimate.

Again, you will find two regimes.

- For each scaling regime, write down how  $\gamma$  and  $\alpha$  are related (per lectures) and check how this expression works for your estimates here.

- Everyday random walks and the Central Limit Theorem:

Show that the observation that the number of discrete random walks of duration  $t = 2n$  starting at  $x_0 = 0$  and ending at displacement  $x_{2n} = 2k$  where  $k \in \{0, \pm 1, \pm 2, \dots, \pm n\}$  is

$$N(0, 2k, 2n) = \binom{2n}{n+k} = \binom{2n}{n-k}$$

leads to a Gaussian distribution for large  $t = 2n$ :

$$\Pr(x_t \equiv x) \simeq \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}.$$

Please note that  $k \ll n$ .

Stirling's sterling approximation  will prove most helpful.

**Hint:** You should be able to reach this form:

$$\frac{\text{Some stuff not involving penguins}}{\text{Some other penguin-free stuff} \times (1 - k^2/n^2)^{n+1/2} (1 + k/n)^k (1 - k/n)^{-k}}.$$

Lots of sneakiness here. You'll want to examine the natural log of the piece shown above, and see how it behaves for large  $n$ .

You may very well need to use the Taylor expansion  $\ln(1 + z) \simeq z$ .

Exponentiate and carry on.

**Tip:** If at any point penguins appear in your expression, you're in real trouble. Get some fresh air and start again.

- From lectures, show that the number of distinct 1- $d$  random walk that start at  $x = i$  and end at  $x = j$  after  $t$  time steps is

$$N(i, j, t) = \binom{t}{(t+j-i)/2}.$$

Assume that  $j$  is reachable from  $i$  after  $t$  time steps.

**Hint—Counting random walks:**

<http://www.youtube.com/watch?v=daSIYz-0U3E>

8. *Discrete random walks:*

In class, we argued that the number of random walks returning to the origin for the first time after  $2n$  time steps is given by


$$N_{\text{first return}}(2n) = N_{\text{fr}}(2n) = N(1, 1, 2n - 2) - N(-1, 1, 2n - 2)$$

where

$$N(i, j, t) = \binom{t}{(t + j - i)/2}.$$

Find the leading order term for  $N_{\text{fr}}(2n)$  as  $n \rightarrow \infty$ .

Two-step approach:

- (a) Combine the terms to form a single fraction,
- (b) and then again use [Stirling's bonza approximation](#) .

If you enjoy this sort of thing, you may like to explore the same problem for random walks in higher dimensions. Seek out George Pólya.

And we are connecting to much other good stuff in combinatorics; more to come in the solutions. #toomuchexcitement