Principles of Complex Systems
Assignment 3: Remedial Chaos Theory

Maxfield Green

September 2018

0.1 Questions 1 and 2

Plot the complementary cumulative distribution function (CCDF).

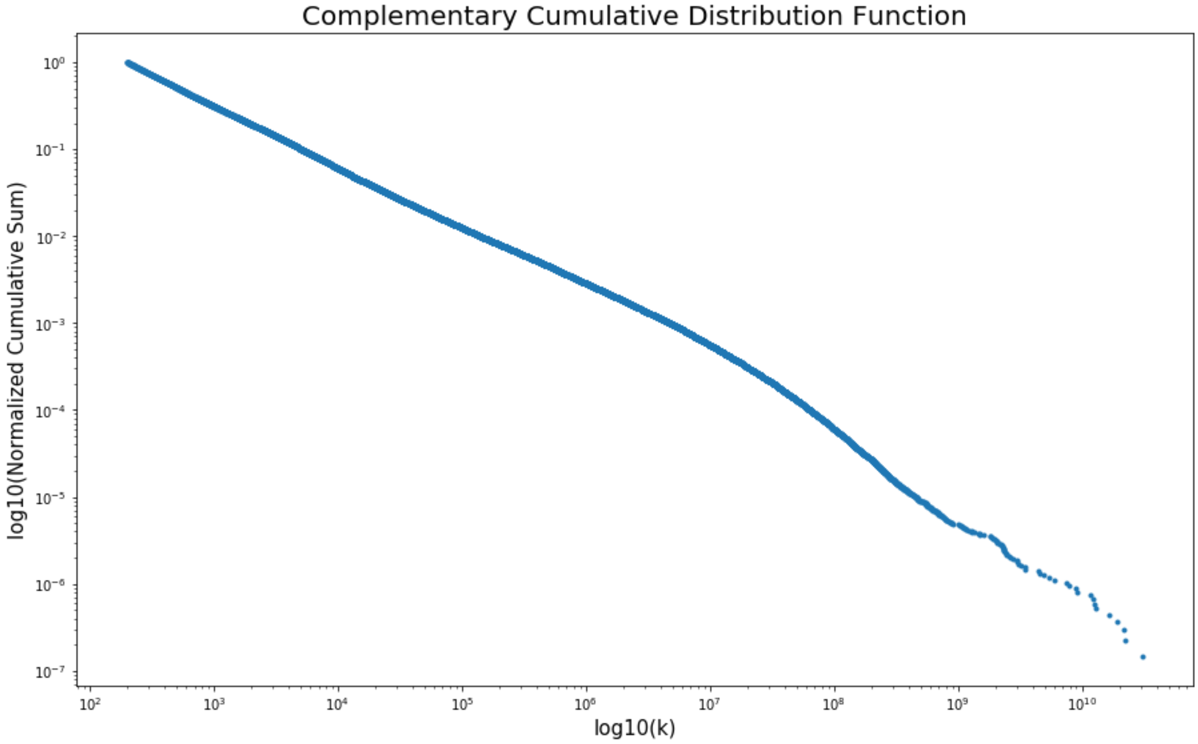

Figure 1: Complementary Cumulative Distribution Function

The CCDF shows a log linear relationship between the cumulative sum of Nk, the number of words used K times, and K,the number of times a word was used in the text corpus.There is an interesting break in scale at around 10^4. This blimp separates two regions of consistent linearity in log space. To account for the scaling break present in this region, I will find unique λ values for both regions, indicating potentially different scaling exponents along the Power Law Size Distribution.

The two scaling regimes hold different slopes which represent different scaling exponents for different parts of the distribution. To me this suggests that the scaling is not completely stable. If we sample exclusively from the tail of the distribution, we will find that words are severely less frequently used than if we sample from the left side, for lower k.

Intervals of 95% confidence in slope generated from linear regression are reported for the upper and lower scaling regimes respectively as: (-0.6669,-0.6667) and (-1.3052,-1.020). These intervals are important as the data represents a 5,000,000 observation sample from the raw frequency data. This lets us know that if we repeated the sampling many times, and computed the confidence interval again, we can expect that 95% of the confidence intervals generated would contain the true slope of the regression line.

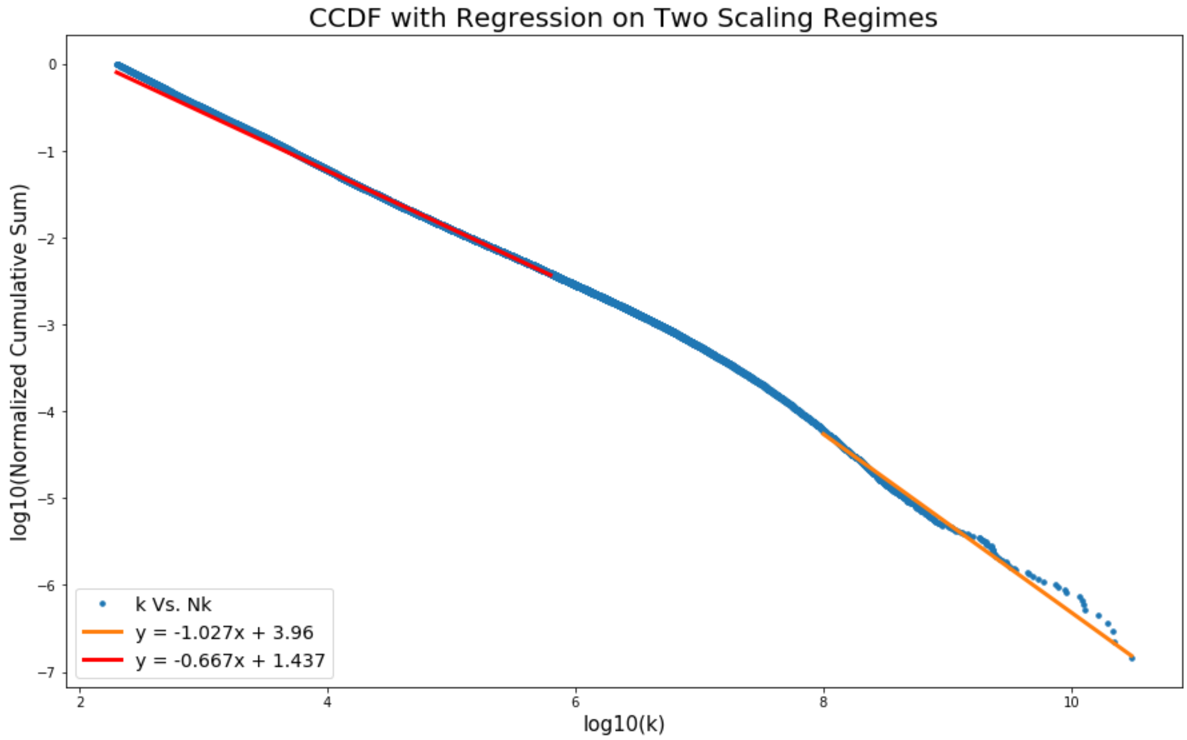

Figure 2: Two Scaling Regimes

0.2 Questions 3 and 4

Again, we observe an interesting relationship between the frequency and rank of words from the text corpus. This time, using the raw frequencies, we show that the data follows Zipf's Law.

Zipf's law predicts that the frequency f of a word with rank r.

$$f = r^{-\alpha} \tag{1}$$

This clearly is related to Power Law Size Distributions. We see high frequency for low ranking. i.e. there are a few words that get used many times and many words that get used few times. Zipf found this relationship in many other fields as well, tokening it as the "Principle of Least Effort". It is illustrated further in modeling how individuals tend to take the path of least resistance. When using words, this means there are a few words that are heavily encoded and can be used in many different ways. But of course this places a burden on the listener who must decode based on contextual information from surrounding words. Below, a sample of the raw frequency data is plotted by the rank the frequency.

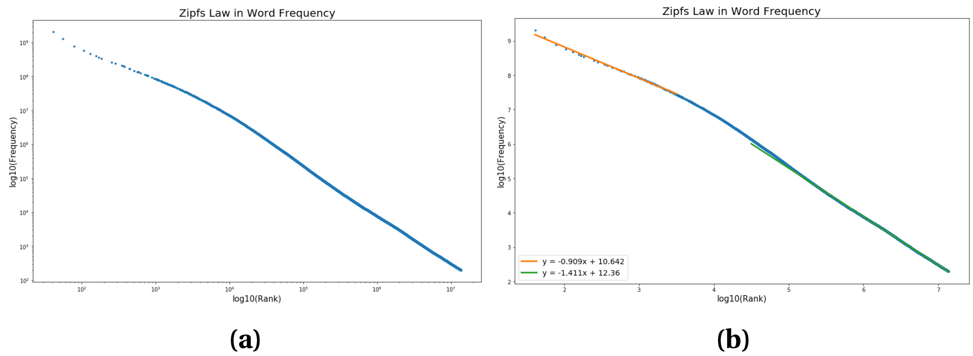

Figure 3: Zipfs Law in Word Use Frequency and Rank

Again, since the data presented is a sample of a larger group of frequencies, a 95% confidence interval is necessary to show how representative the sample is of the larger raw data set. Upper and Lower bounds scaling regimes are reported respectively as: (-0.9186, -0.8996) and (-1.4120,-1.4114).

Once again we see the clear break in scaling which is validated by the change in scaling exponent γ.

0.3 Question 5

For each scaling regime, write down how γ and α are related (per lectures) and check how this expression works for your estimates here.

	Alpha	Gamma	Gamma'
Upper Regime	0.909	0.667	2.1
Lower Regime	1.411	1.027	1.7

(2)

Here Gamma' represents the result of the expression $1 + 1/\alpha$, the calculated γ. The disparity between the computed and calculated α are potentially a result of the eyeball determined cutoff between scaling regimes and any sampling bias present. While the sample was randomly pulled out of the raw data, if the high ranking words were under sampled by chance, it would unevenly effect the sample distribution spread.

0.4 Question 6

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 $k \in 0, \pm 1, \pm 2..$, is

$$\binom{2n}{n-k} \tag{3}$$

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

$$Pr(x_t \equiv x) \cong \frac{e^{\frac{-x^2}{2t}}}{\sqrt{2\pi t}} \tag{4}$$

I will note any use of external theorems and approximations. Before we begin, we can assume that k « n. No penguins were found or harmed during the course of this problem. //

$$\binom{2n}{n-k} = \frac{2n!}{(n-k)!(n+k)!} \tag{5}$$

$$\text{x = (n-k), y = (n+k)} \tag{6}$$

$$\frac{2n!}{x!y!} \tag{7}$$

To aid in moving forward, we will use Sterling's Approximation which specifies an asymptotic relationship between n such that n! ~ $\sqrt{2\pi n}(\frac{n}{e})^n$

$$= \frac{\sqrt{4\pi n}(\frac{2n}{e})^{2n}}{\sqrt{2\pi x}(\frac{x}{e})^x\sqrt{2\pi y}(\frac{y}{e})^y} \tag{8}$$

$$= \frac{e^{-2n+x+y}(2^{2n})(n^{2n+\frac{1}{2}})}{\sqrt{\pi}x^{x+\frac{1}{2}}y^{y+\frac{1}{2}}} \tag{9}$$

$$= \frac{(2n)^{2n}n^{1/2}}{\sqrt{\pi}(n-k)^{n-k-1/2}(n+k)^{n+k+1/2}} \tag{10}$$

$$= \frac{(2n)^{2n}n^{1/2}}{\sqrt{\pi}(n-k)^{-k}(n-k)^{n+1/2}(n+k)^{k}(n+k)^{n+1/2}} \tag{11}$$

$$\tag{12}$$

Because k«n, n - k > 0 for all n and k. Thus, the following holds true: $(n-k)^k == (1-\frac{k}{n})^k n^k$. Using this case generally, we get the following expression.

$$= \frac{(2n)^{2n} n^{1/2}}{\sqrt{\pi}(1-\frac{k}{n})^{-k}(1-\frac{k}{n})^{k}(1-\frac{k^2}{n^2})^{n+1/2}(n)^{2n+1}} \tag{13}$$

$$= \ln \frac{(2n)^{2n} n^{1/2}}{\sqrt{\pi}(1-\frac{k}{n})^{-k}(1-\frac{k}{n})^{k}(1-\frac{k^2}{n^2})^{n+1/2}(n)^{2n+1}} \tag{14}$$

$$= \ln 2^{2n} n^{-1/2} - \ln \pi^{1/2}(1-\frac{k}{n})^{-k}(1-\frac{k}{n})^{k}(1-\frac{k^2}{n^2})^{n+1/2}(n)^{2n+1} \tag{15}$$

$$\text{Taking the natural log:} \tag{16}$$

$$= 2n\ln 2 - 1/2\ln n - 1/2\ln\pi - 2\frac{k^2}{n} - (n+1/2)\frac{k^2}{n^2} \tag{17}$$

$$Re-exponentiation: \tag{18}$$

$$= (e^{2n\ln 2})(e^{-1/2\ln n})(e^{-1/2\ln\pi})(e^{\frac{-k^2}{n^2}-\frac{k^2}{2n^2}}) \tag{19}$$

$$= \frac{2^{2n} * e^{-(\frac{k^2}{2n^2}+\frac{k^2}{n})}}{\sqrt{\pi n}} \tag{20}$$

Because we know that k«n, we can simplify the terms in the exponent of e as such:

$$e^{-(\frac{k^2}{2n^2}+\frac{k^2}{n})} \tag{21}$$

$$= \frac{k^2}{2n^2} + \frac{k^2}{n}(\frac{2n}{2n}) \tag{22}$$

$$= \frac{k^2 + k^2 2n}{2n^2} \tag{23}$$

$$= \frac{k^2(1+2n)}{2n^2} \tag{24}$$

$$\tag{25}$$

Here we can ignore the 1 because its minuscule compared to assumed massive n.

$$\frac{k^2(1+2n)}{2n^2} = \frac{k^2}{n} \tag{26}$$

$$\tag{27}$$

Additionally, because 2k = x, and 2n = t we can reduce the expression to the following form:

$$= \frac{k^2}{n} = \frac{4k^2}{4n} = \frac{(2k)^2}{2(2n)} = \frac{x^2}{2t} \tag{28}$$

$$\tag{29}$$

Plugging this result back into the main expression as seen in line 19, we get:

$$= \frac{2 * 2^{2n} * e^{-\frac{x^2}{2t}}}{\sqrt{\pi 2t}} \tag{30}$$
$$= \frac{2^{2n+1} * e^{-\frac{x^2}{2t}}}{\sqrt{\pi 2t}} \tag{31}$$
$$\tag{32}$$

The last sneaky bit to this little exercise is recognizing that 2^{2n+1} as the counts, allowing us to simply divide it out to reach the units we desire

$$Pr(x_t \equiv x) \cong \frac{e^{\frac{-x^2}{2t}}}{\sqrt{2\pi t}} \tag{33}$$

0.5 Question 7

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

$$N(i,j,t) = \binom{t}{(t+j-i)/2} \tag{34}$$

Here we want to explore how many random walks start and end at the same place given the same time period. If one step occurs at each time step t, than P = positive steps, N = negative steps. Thus

$$P + N = t \tag{35}$$
$$P - N = j - i \tag{36}$$
$$\tag{37}$$

Combining the two expressions gives

$$P = \frac{j-i+t}{2} \tag{38}$$

And the number of random walks possible at time are given as:

$$\binom{t}{P} \tag{39}$$
$$\tag{40}$$

Thus, we can rewrite our choose function as such:

$$\binom{t}{\frac{j-i+t}{2}} \tag{41}$$

0.6 Question 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:

$$N_{firstreturn}(2n) = N_{fr}(2n) = N(1,1,2n-2) - N(-1,1,2n-2) \tag{42}$$

where

$$N(i,j,t)\binom{t}{(t+j-i)/2}. \tag{43}$$

Find the leading order term for N$_{fr}$(2n) as $n->\infty$ (a) Combine the terms to form a single fraction,

$$N(1,1,2n-2) - N(-1,1,2n-2) = \frac{(2n-2)!}{((n-1)!)^2} - \frac{(2n-2)!}{(n)!(n-2)!} \tag{44}$$

$$= \frac{(2n-2)!n}{(n!)^2} - \frac{(2n-2)!n(n-1)}{(n!)^2} \tag{45}$$

$$= \frac{(2n-2)!n^2}{(n!)^2} - (\frac{(2n-2)!n^2}{(n!)^2} - \frac{(2n-2)!n}{(n!)^2}) \tag{46}$$

$$= \frac{(2n-2)!n}{(n!)^2} \tag{47}$$

$$= \frac{\frac{2n!n}{(2n-1)2n}}{(n!)^2} \tag{48}$$

$$= \frac{(2n)!}{(4n-2)(n!)^2} \tag{49}$$

$$= \frac{(2n)^{2n}}{(4n-2)(\sqrt{2\pi n})(n^{2n+1/2})} \tag{50}$$

(b) and then again use Stirling's Approximation

Stirling's Approximation => n! ~ $\sqrt{2\pi n}(\frac{n}{e})^n$

$$\frac{(2n)!}{(4n-2)(n!)^2} \tag{51}$$

$$= \frac{\sqrt{4\pi n}(\frac{2n}{3})^{2n}}{(4n-2)(\sqrt{2\pi n}(\frac{n}{e})^n)^2} \tag{52}$$

$$= \frac{(2n)^{2n}}{(4n-2)(\sqrt{\pi})n^{2n+1/2}} \tag{53}$$

$$= \frac{2^{2n}n^{2n}}{(4n-2)(\sqrt{\pi})n^{2n+1/2}} \tag{54}$$

$$= \frac{2^{2n}}{(4n-2)(\sqrt{\pi})n^{1/3}} \tag{55}$$

$$= (2^{2n})(\pi n)^{-1/2}(4n-2)^{-1} \tag{56}$$

For very large n, we are really only considered with the leading term of the above expression, 2^{2n}. For very large n,this term will dominate. This means that when looking at many different walks from i,j, at any given time t, the highest concentration of walkers will be at the origin. In order to return to the origin, a walker needs an even number of time steps, explaining the 2n exponent.

To visualize this distribution of random walks, I simulated 1000 random walks of length 100. There is a very clear Gaussian distribution centered at the origin. It appears that the outer bounds on either side of the origin are fairly symmetric a little over 30 steps. This is interesting.

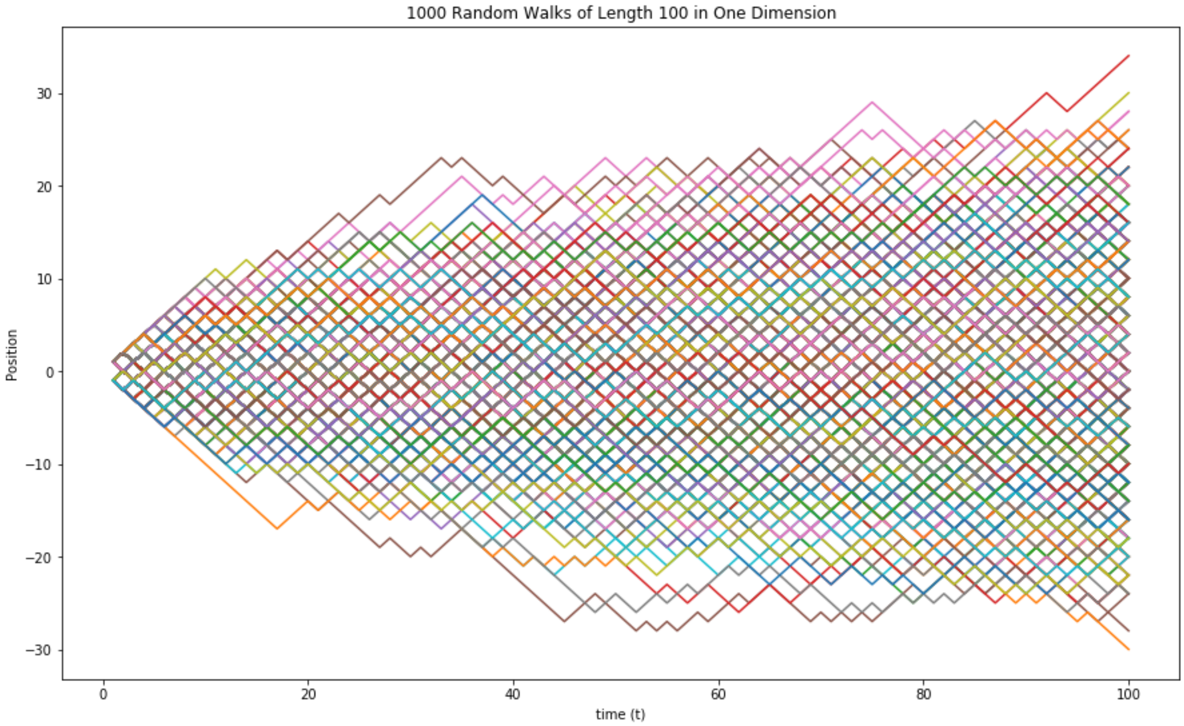

Figure 4: Random Walkers Set Loose in One Dimension. Most of them are loyal creatures of habit and do come back to their birth place