PRINCIPLES OF COMPLEX SYSTEMS

HW04 WRITE-UP

David W. Landay
University of Vermont
Graduate Student, Comp. Systems and Data Sci.

0.1 Problem 1

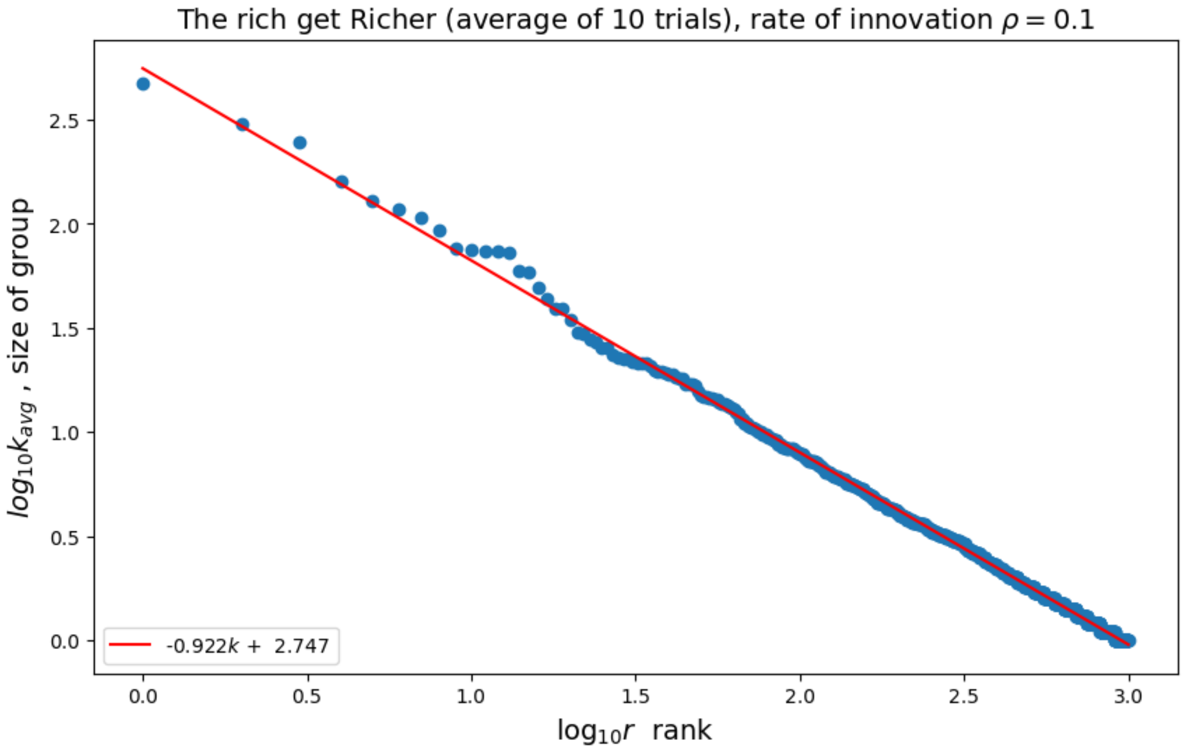

Figure 1: Rich get richer, Innovation rate $\rho = 0.1$, on a $\log_{10}$ scale

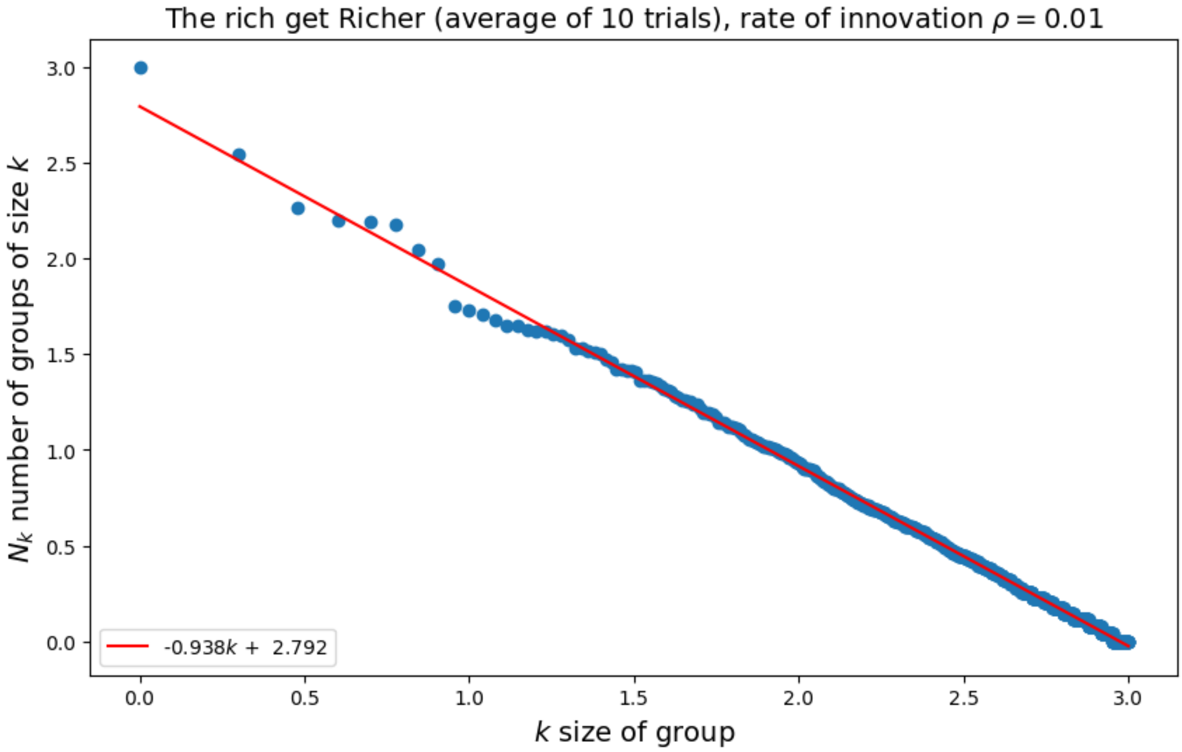

Figure 2: Rich get richer, Innovation rate $\rho = 0.01$, on a $\log_{10}$ scale

I was unable to generate a simulation for $\rho = 0.001$ due to time limitations and computational resources. However, allowing for three orders of magnitude (i.e: 1000 unique groups) demonstrates in the above examples that our observed scales don't change much. That being said, $\rho = 0.001$ would model a huge distribution of groups; perhaps a truer estimate of α would be observed.

We know that the expected value of α for any sample of size N should be

$$\alpha = 1 - \rho.$$

Hence, for a theoretical rate of 0.1, we expect $\alpha = 0.9$, and for a theoretical rate of 0.01, we expect $\alpha = 0.99$. Hence, our estimates are fairly accurate.

0.2 Problem 2

For Herbert Simon's model of what we've called Random Competitive Replication, we found in class that the normalized number of groups in the long time limit, n_k, satisfies the following difference equation:

$$\frac{n_k}{n_{k-1}} = \frac{(k-1)(1-\rho)}{1+(1-\rho)k} \tag{1}$$

where $k \geq 2$. The model parameter ρ is the probability that a newly arriving node forms a group of its own (or is a novel word, starts a new city, has a unique flavor, etc.). For $k = 1$, we have instead

$$n_1 = \rho - (1-\rho)n_1 \tag{2}$$

which directly gives us n_1 in terms of ρ.

We can derive an exact solution for n_k in terms of gamma functions, which in turn will yield a beta function describing the relationship between n_k and ρ. We can rewrite the difference equation (1) above as an infinite product of terms to denote n_k:

$$\frac{n_k}{n_{k-1}} = \frac{(k-1)(1-\rho)}{1+(1-\rho)k}$$

$$\implies n_k = \left(\frac{(k-1)(1-\rho)}{1+(1-\rho)k}\right)\left(\frac{(k-2)(1-\rho)}{1+(1-\rho)(k-1)}\right)n_{k-2}$$

$$= \left(\frac{(k-1)(1-\rho)}{1+(1-\rho)k}\right)\left(\frac{(k-2)(1-\rho)}{1+(1-\rho)(k-1)}\right)\left(\frac{(k-3)(1-\rho)}{1+(1-\rho)(k-2)}\right)n_{k-3}$$

$$= \left(\frac{(k-1)(1-\rho)}{1+(1-\rho)k}\right)\left(\frac{(k-2)(1-\rho)}{1+(1-\rho)(k-1)}\right)\cdots\left(\frac{(\star_1)(1-\rho)}{1+(1-\rho)(\star_2)}\right)n_1$$

For simplicity, we can let $z = 1-\rho,\ 0 \leq \rho \leq 1 \implies 0 \leq z \leq 1$. We get,

$$n_k = \left(\frac{(k-1)z}{1+zk}\right)\left(\frac{(k-2)z}{1+z(k-1)}\right)\cdots\left(\frac{(\star_1)z}{1+z(\star_2)}\right)n_1$$

and if we factor out z we are left with

$$\frac{Z^k \Big((k - 1)(k - 2)(k - 3) \ldots \star_1 \Big)}{Z^k \Big((\frac{1}{z} + k)(\frac{1}{z} + (k - 1))(\frac{1}{z}(k - 2)) \ldots (\frac{1}{z} + \star_2) \Big)}$$

$$= \frac{(k - 1)(k - 2)(k - 3) \ldots \star_1}{(\frac{1}{z} + k)(\frac{1}{z} + (k - 1))(\frac{1}{z}(k - 2)) \ldots (\frac{1}{z} + \star_2)}$$

Since we are assuming that n_k is the number of groups in the "long time limit", we will assume that for some large value of t, that $\star_1$ and $\star_2$ will both approach 1. Hence, the fraction approaches

$$\frac{(k - 1)(k - 2)(k - 3) \ldots 1}{(\frac{1}{z} + k)(\frac{1}{z} + (k - 1))(\frac{1}{z}(k - 2)) \ldots (\frac{1}{z} + 1)}$$

$$= \frac{(k - 1)!}{(\frac{1}{z} + k)!}$$

A Gamma function "gives meaning" to the factorial of any positive real number. In terms of gamma functions, our expression becomes

$$\frac{\Gamma(k - 1)}{\Gamma(\frac{1}{z} + k + 1)}$$

However, we can use the property that $\Gamma(x + 1) = x\Gamma(x)$ to obtain

$$\frac{\Gamma(k)}{\Gamma(\frac{1}{z} + k + 1)}$$

Substituting $1 - \rho$ back in for z, we conclude that

$$n_k = \frac{\Gamma(k)}{\Gamma\Big(\frac{1}{(1 - \rho)} + k + 1\Big)} \therefore \tag{2.a}$$

A beta function takes the form $\mathbf{B}(m, n) = \frac{\Gamma(m)\Gamma(n}{\Gamma(m + n)}$. For large m and fixed n, we can approximate the beta function by $\mathbf{B}(m, n) \sim m^{-n}$. Therefore, we can re-write our expression in part **a** as

$$\mathbf{B}(k, 1 + \frac{1}{(1-\rho)} + k) = \frac{c\,\Gamma(k)}{\Gamma\left(k + 1 + \frac{1}{(1-\rho)}\right)}$$

$$\implies \mathbf{B}(k, 1 + \frac{1}{(1-\rho)} + k) \sim k^{-\left(\frac{1}{(1-\rho)}+1\right)}$$

for large k and a fixed ρ. Hence,

$$\frac{(k-1)!}{\left(\frac{1}{(1-\rho)} + k\right)!} \approx k^{-\left(\frac{1}{(1-\rho)}+1\right)}$$

see notes for sterling's approximation worked out.

$$\implies n_k \approx k^{-\left(\frac{1}{(1-\rho)}+1\right)} \tag{2.b}$$

We can see that for large k behavior, n_k will tend to 0 as we innovate with a probability ρ because the ratio of a root of k and k will be dominated by the k term in the denominator. i.e: as we innovate at some fixed rate, the fraction of groups of size k will tend towards 0, because it is more likely that we sample from an uneven distribution of groups, where one size of k (namely groups of size 1) will dominate.

0.3 Problem 3

If the innovation rate ρ is not fixed, then the convergent behavior of n_k will change. For instance, as ρ tends towards 0, then $n_k \simeq \frac{1}{k^2}$. This indicates that the value of $\gamma \simeq 2$ and small k will relate to $n_k \simeq 1$, while large k will tend to the x-axis; $n_k \rightarrow 0$. On the other hand, as ρ tends towards the value 1, then $\gamma \simeq \infty$, indicating $n_k \simeq 0$ for all values of k (this makes sense because it is like saying we will innovate all the time, so every group is of size 1).

0.4 Problem 4

Simon's model invesigates the relationship between the number of groups of size k, for some innovation rate ρ, given sufficient time has passed for new groups to emerge in the system. Hence, we have a growth model parameterized by time.

In class, we derived the fraction of groups containing only 1 element, finding

$$n_1^{(g)} = \frac{N(t)}{\rho\, t} = \frac{1}{2 - \rho}$$

We wish to find the fraction of groups of size k, for $k = 2$ and $k = 3$. Recall that

$$P_k(t) = \frac{k\, N_{k,\,t}}{t},$$
the probability of seeing a group of size k at time t (1)

$N_{k,\,t} = N(t) =$ the number of groups containing k "elephants"

"elephants" belonging to a group with k elephants is replicated:
$$N_{k,\,t+1} = N_{k,\,t} - 1, \text{ happens with probability } \frac{(1 - p)k\, N_{k,\,t}}{t} \tag{2}$$

"elephants" belonging to a group with $k - 1$ elephants is replicated:
$$N_{k,\,t+1} = N_{k,\,t} + 1, \text{ with probability } \frac{(1 - p)N_{k-1,\,t}}{t} \tag{3}$$

Then, the expected growth of N_k at time t for any group of size k, is equal to the probability hat we replicate an "elephant" in the sample space, multiplied by the difference between the total number of groups of size $k - 1$ at time t, and the number of groups of size k at time t.

For $k > 1$,

$$< N_{k,\,t+1} - N_{k,\,t} > = (1 - \rho)\left((+1)(k - 1)\frac{N_{k-1,\,t}}{t} + (-1)\, k\frac{N_{k,\,t}}{t} \right) \tag{4}$$

and, for $k = 1$,

$$< N_{1,\, t+1} - N_{1,\, t} > = (+1)\rho + (-1)(1-\rho)\frac{N_{1,\, t}}{t}$$

$$= \rho - (1-\rho)\frac{N_{1,\, t}}{t}$$

$$= \rho + \frac{N_{1,\, t}}{t}\rho - \frac{N_{1,\, t}}{t} \tag{5}$$

We will assume that for large t (i.e: after much time has passed), the distribution stabilizes; i.e: $N_{k,\, t} = n_k\, t$ represents the true number of groups of size k for any time t.

$$\implies < N_{k,\, t+1} - N_{k,\, t} > = n_k\,(t+1) - n_k\,(t)$$

Note that the fraction of groups of size k is $\frac{n_k}{\rho}$, since

$$\frac{N_{k,\, t}}{\rho\, t} = \frac{n_k\, t}{\rho\, t} = \frac{n_k}{\rho}\ \therefore$$

We can now derive the values for expected growth of N_k at time t for any group of size k, by substituting in the above expression. We get for $k > 1$

$$< N_{k,\, t+1} - N_{k,\, t} > = (1-\rho)\Big((k-1)\frac{n_{k-1}\, t}{t} - k\frac{n_k\, t}{t}\Big)$$

$$\implies n_k\,(t+1) - n_k\,(t) = (1-\rho)\Big((k-1)n_{k-1} - k\, n_k\Big)$$

$$\implies n_k\,(1 + (1-\rho)k) = (1-\rho)(k-1)n_{k-1}$$

Thus,

$$n_2 = \frac{1-\rho}{2-\rho} \times \frac{1}{3-2\rho}$$

and

$$n_3 = \frac{2 + 2\rho^2}{(2 - \rho)(3 - 2\rho)(4 - 3\rho)} \tag{4.a}$$

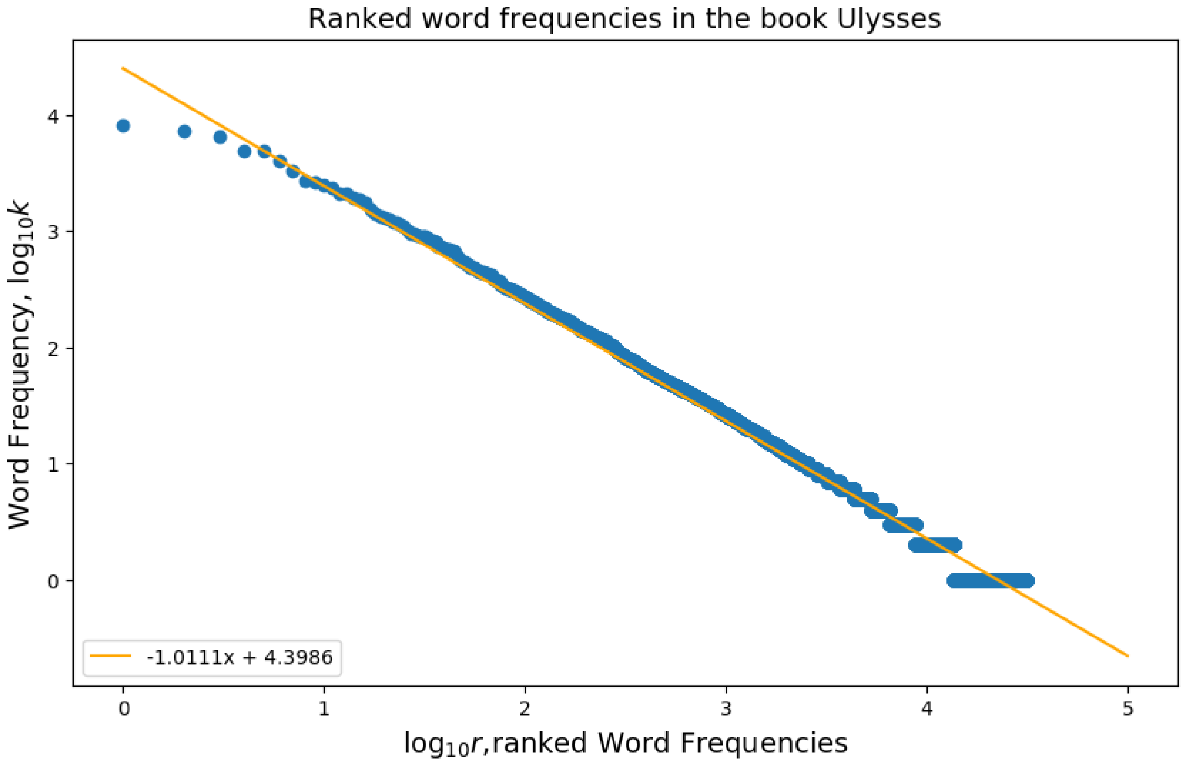

Figure 3: Zipf distribution of words found in the book "Ulysses"

We can confirm Simon's original estimate for the innovation rate ($p_{est} = 0.115$) for the order of word appearances in James Joyce's novel "Ulysses", by conducting our own investigation of the book. The above plot demonstrates how word frequencies in the novel follow a Zipf distribution. The slope in log-log space represents the inverse of the empirical value of α, the rate at which an "elephant" (word of frequency k) is replicated, which we can relate to the scaling factor, γ, by

$$\alpha = \frac{1}{\gamma - 1}$$

$$\implies \frac{1}{\alpha} = \gamma - 1$$

Subsequently, we can then relate α to the innovation rate ρ by

$$\alpha = \frac{1}{1 + \frac{1}{((1-\rho)} - 1}$$

$$\alpha = 1 - \rho$$

$$\frac{1}{\alpha} = \frac{1}{1-\rho} \therefore \tag{4.b}$$

If we substitute Simon's estimate for the innovation rate ρ_{est}, we get that $\alpha = 0.885$, which implies that $\frac{1}{\alpha} \approx 1.129$, which is a close approximation to the empirical value of $|\frac{1}{\alpha}|$ that we obtain from the observed word counts in "Ulysses". In addition, if we calculate the ratio of the actual number of unique words, to the total number of words in "Ulysses" (the probability that a new word appears in the corpus, AKA the innovation rate) is 0.1186. This is quite close to Simon's initial estimate.

Using Simon's finding, we can compare the estimated values of $n_1^{(g)}$, $n_2^{(g)}$, and $n_3^{(g)}$, the fraction of groups of size 1, 2, and 3, with that of the empirical values that we obtain from the data. We get:

$n_k^{(g)}$	Simon's Estimate ($\rho = 0.115$)	Observed
$n_1^{(g)}$	0.385	0.564940
$n_2^{(g)}$	0.169	0.155647
$n_3^{(g)}$	0.105	0.071374

Table 1: How well Herbert Simon's estimate $\rho_{est} = 0.115$ holds up (4.b)

We can see that the error in Simon's estimate for ρ_{est} accounts for an understimate for the value of n_1^g. However, his estimate conforms better to, but still over estimates, the actual values for n_2^g and n_3^g. The under estimation appears to agree for small rank r where, in figure 4, we can see divergence from the least squares regression.

0.5 Problem 5

Consider a set of N samples, randomly chosen according to the probability distribution $P_k = ck^{-\gamma}$, where $k \geq 1$ and $2 < \gamma < 3$. We will note that in the case of group size, k, we are dealing with a distribution of discrete values (all k). We will note that as the sample size gets larger, we can expect that there will be some largest value k in the sample. We can find how the minimum largest value in the sample, $\mathbf{min}k_{max}$, changes as a function of the sample size N. This will ultimately give us a better intuition for how the size of the sample dictates how large k is "allowed" to become based on a fixed scaling factor γ.

If every point in the sample of size N was generated by a probability $P_k = ck^{-\gamma}$, for $\gamma > 1$ and $k = 1, 2, \ldots$ is large, where the points are discrete, then the expected minimum degree of a sample of size N is

$$\sum_{\mathbf{min}k_{max}}^{\infty} P_k = \frac{1}{N}.$$

This makes sense intuitively; if there are N sample points, then we expect at least $\frac{1}{N}$ will be the largest value out of all points. We can re-write the infinite sum as

$$c \sum_{\mathbf{min}k_{max}}^{\infty} k^{-\gamma} = \frac{1}{N}$$

letting

$$\mathbf{min}k_{max}(t) = Q_t$$

where t represents the **min**k_{max} at time t, we get

$$c(\,Q_1^{-\gamma} + Q_2^{-\gamma} + \ldots + Q_t^{-\gamma} + \ldots Q_\infty^{-\gamma}\,) = \frac{1}{N}$$

$$\implies (\,Q_1^{-\gamma} + Q_2^{-\gamma} + \ldots + Q_t^{-\gamma} + \ldots Q_\infty^{-\gamma}\,) = \frac{1}{c\,N}$$

But we are assuming that Q_t is the largest point in every sample N, so we can let $\sum_{Q_1}^{\infty} Q_t = N\,Q = N\,k_{max}$ to obtain

$$N\,Q^{-\gamma} = \frac{1}{c\,N}$$

$$\implies Q^{-\gamma} = \frac{1}{c\,N^2}$$

$$\implies Q = c^{-1}N^{-\frac{2}{\gamma}}$$

$$\implies k_{max} = c^{-1}N^{-\frac{2}{\gamma}} \qquad (5.a)$$

The expected value of k_{max} is then

$$<Q> = k^1 k^{-\gamma}$$

$$\implies <Q> = c^{-1}N^{\frac{1-\gamma}{2}} \qquad (5.b)$$

0.6 Problem 6

We can test our estimate of the expected value for **min** k_{max} for a fixed γ. Let $\gamma = \frac{5}{2}$. We can obtain a rough sense of the expected value of k_{max} as a function of N by generating many, $n = 1000$, samples for increasingly large size, $N = 10, 10^2, 10^3, 10^4, 10^5, 10^6$ using $P_k = c\,k^{-\gamma}$.

We will neglect the normalizing constant and generate this distribution by making the same assumptions in the previous problem; that are distribution is made of discrete values of k, and we expect on the order of 1 of the samples to be k_{max} or greater for any sample of size N. Hence, we can generate the distribution up to 10^6. We end up with a plot that looks like this

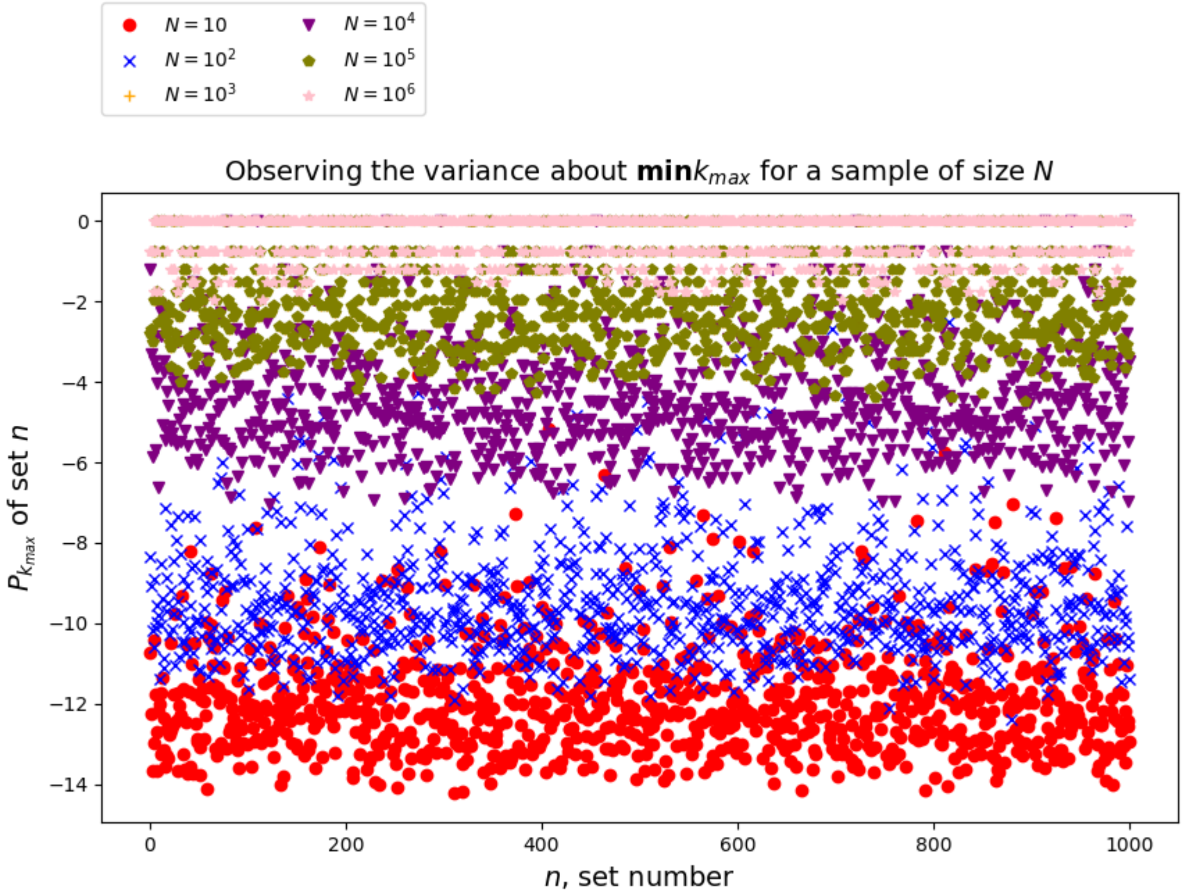

Figure 4: Observing the relationship between sample size N and the expected k_{max}

This demonstrates how as the sample size gets larger, the expected value of k_{max}

goes to 1. For every sample size N, we can average the observed maximum value for k to obtain what the estimated expected value should be:

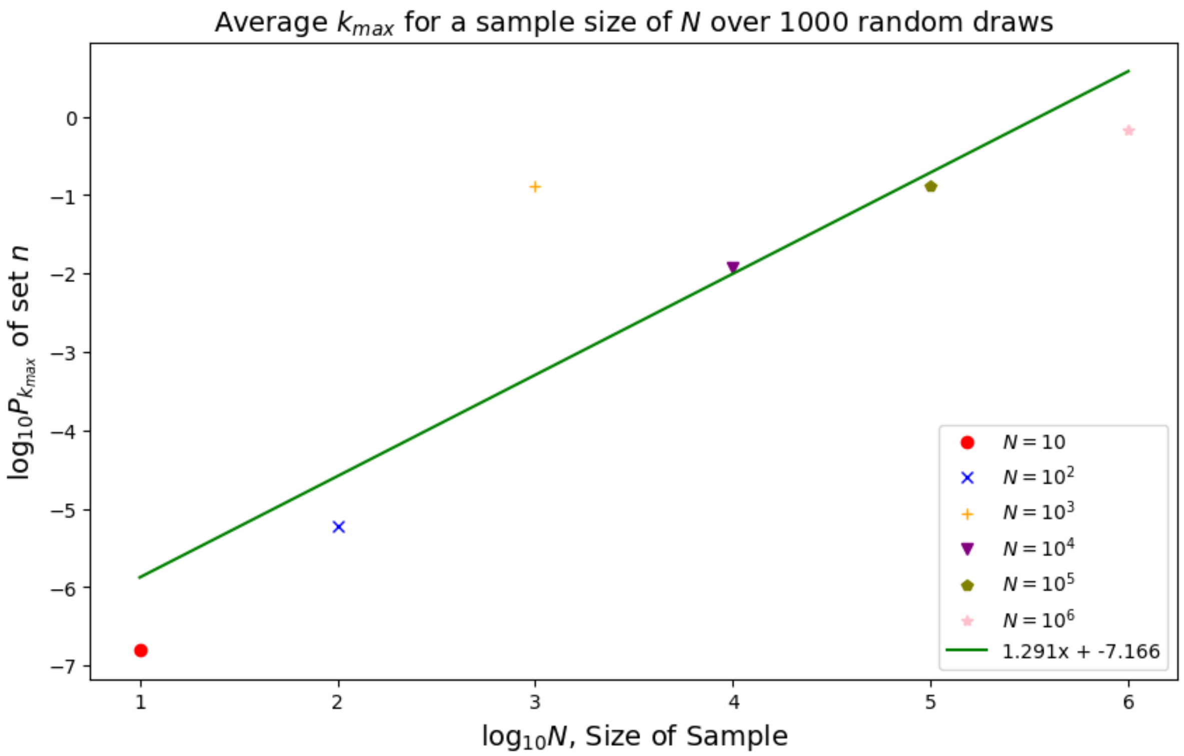

Figure 5: regression over the expected maximum value of k as a function of sample size N

We can show by the slope of the least squares regression and the empirical $< k_{max} >$ we found in problem 5, that the scaling conforms nicely to our calculation.