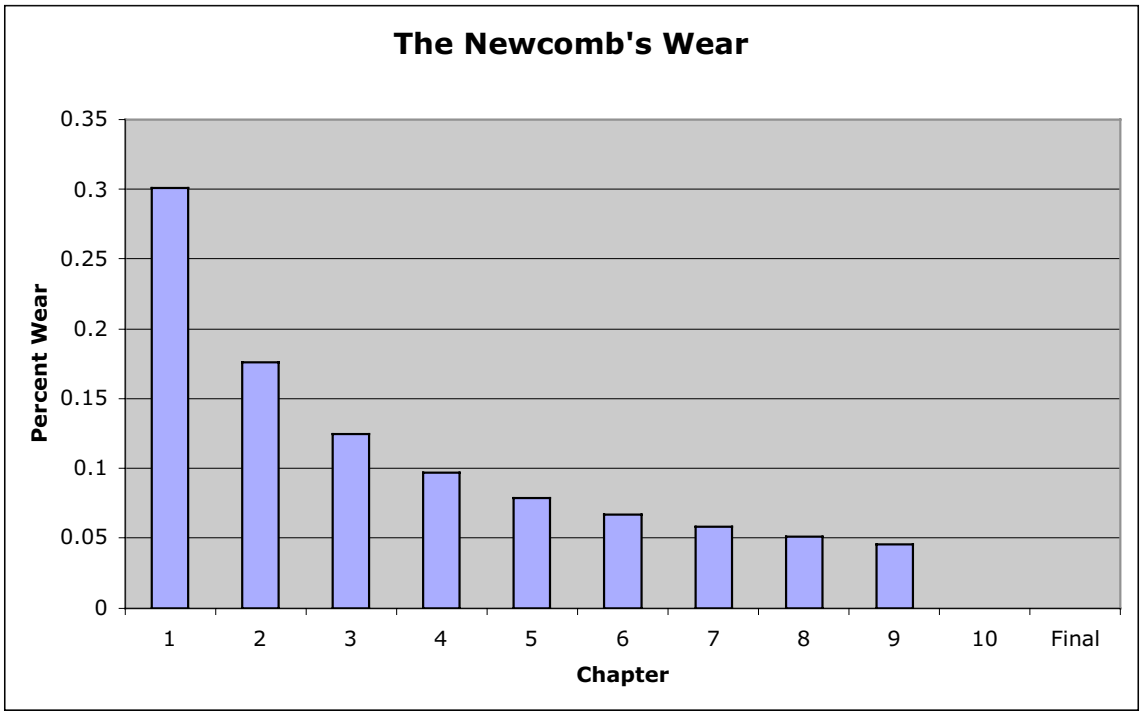
Google Numbers

Greg Leibon

Math 5 Example Project

Clearly, the chances of observing various numbers in nature cannot satisfy the equally likely hypothesis. Namely, there is an infinitude of numbers, and we never see even "moderately" large numbers like $10^{10^{10^{10^{10}}}}$. Naturally occurring and observed numbers must satisfy some sort of law making bigger numbers less likely to be observed. The quite surprising thing is that many varieties of naturally occurring and observed numbers satisfy an extremely special distribution. I will call them *Newcomb Numbers*; the special distribution they satisfy is called the *Benford Distribution*. One variety of naturally occurring and observed numbers is the collection of Google Numbers, where the *Google Numbers* are the numbers of interest found on the World Wide Web. In what follows, I will explore to what extent the Google Numbers are Newcomb Numbers.

First, I'll introduce our needed distribution and a little bit of its history. The observation that naturally occurring and observed numbers satisfy a very special distribution goes back to Simon Newcomb in 1881. Newcomb discovered this in a very interesting way, namely by noticing that his book of logarithm tables had much more wear at its beginning than at its end. For example, let us imagine that we have a book containing the logarithm of every number from 10,000 to 100,000. Furthermore, imagine the book has 9 chapters, each dedicated to evaluating the logarithms of 10,000 numbers. Furthermore, imagine the kth chapter is dedicated to evaluating the logarithms of the numbers from (k)(10,000) to (k+1)(10,000). If we used the book over an entire career to evaluate the logarithms of all the numbers we came across, then we might find the wear distributed as follows:




In 1938, Frank Benford explored these phenomena carefully and empirically observed this distribution in a variety of naturally occurring collections of numbers. To articulate Benford's distribution, imagine we look at a collection of naturally occurring numbers and examine those between $10^k$ and $10^{(k+1)}$. If these numbers are Newcomb, then we would find that they satisfy (approximately) the following percent break down according to their leading digit:

| Leading Digit | Benford's Probability |
|---|---|
| 1 | 30.1% |
| 2 | 17.6% |
| 3 | 12.5% |
| 4 | 9.7% |
| 5 | 7.9% |
| 6 | 6.7% |
| 7 | 5.8% |
| 8 | 5.1% |
| 9 | 4.6% |

Just to be completely clear, the leading digit of 745,388 is 7; and if we examined many Newcomb numbers with six digits then we would expect about 5.8% of them to have a leading digit of 7. This table was derived via the formula asserting that the probability that a number's leading digit is K is given by

$$\log(K+1)-\log(K)=\log\left(\frac{K+1}{k}\right).$$

The above distribution has come to be known as the *Benford distribution.* At http://mathworld.wolfram.com/BenfordsLaw.html you can find more information on this distribution and its history.

Now I will articulate how I collected a sample of candidate Newcomb numbers. I wanted to understand numbers on the World Wide Web in which real live people were actually interested. In particular, I did not want to accidentally include numbers from data sets intended only for data mining purposes. To accomplish, I included a piece of text in my search. I desired to choose a natural piece of text, hence (for lack of a better idea) I used the word "*nature*". Hence, my Google Numbers are numbers that occur on a web page that also includes the word "nature". Here is an example that illustrates my Google search:

---

Web | Images | Groups | Directory | News

Searched the web for **176781** **nature**. Results **1** - **10** of about **42**. Search took **0.30** seconds.

Trocadero Artisan and Design:Textiles Directory

**...** preserve their vision of the world, their identity and their relationship to **nature**. **...** To see a similar piece, go to ITEM #**176781**.....$10 for insured USA **...**

www.trocadero.com/ directory/Artisan_and_Design:Textiles.html - 36k - Cached - Similar pages

Amazon.com: Books: Of Moths and Men: An Evolutionary Tale

www.amazon.com/exec/obidos/tg/ detail/-/0393051218?v=glance - Similar pages

BASCD Survey Report 1997/98

**...** A total of **176,781** five-year-old children from across the UK were examined, some 2 **...** It is important, however, to remember the skewed **nature** of the disease in the **...**

www.dundee.ac.uk/dhsru/cdh/text1609.htm - 20k - Cached - Similar pages

---

Notice, I found 42 pages containing 176781 and the word “nature”.

Next I chose my $10^k$ to $10^{(k+1)}$ range. I wanted my search to produce robust but reasonable numbers of results. This is because I wanted to leave myself in a position to actually examine the resulting hits in order to achieve a sense for how the numbers were derived. I tried example numbers with three to seven digits (together with the word nature) and found the following results:

| Play Numbers | Results |
|---|---|
| 127 | 2250000 |
| 3127 | 25200 |
| 53127 | 568 |
| 253127 | 40 |
| 4253127 | 0 |

From this little sample I decided that six digits appeared to be the most likely to satisfy my criteria.

I could not figure out how to look at all six-digit numbers, so I decided to collect a sample of randomly collected six-digit numbers. To do so, I chose the following nine random five-digit numbers (via a random number generator):

| | Random Number |
|---|---|
| 1 | 13527 |
| 2 | 31795 |
| 3 | 79644 |
| 4 | 58316 |
| 5 | 85085 |
| 6 | 76781 |
| 7 | 29285 |
| 8 | 39557 |
| 9 | 44557 |

For each of these numbers I searched (via Google) to find out how many web pages contained the word “nature” together with each of the nine numbers that have these numbers forming their first five digits. For example our sixth number is 76781, hence I plugged 176781 (together with the word “nature”) into Google. This was in fact the above Google example. I found the following data:

| 13527 Occurrences | | 31795 Occurrences | | 79644 Occurrences | |
|---|---|---|---|---|---|
| 113527 | 136 | 131795 | 80 | 179644 | 62 |
| 213527 | 44 | 231795 | 58 | 279644 | 23 |
| 313527 | 35 | 331795 | 66 | 379644 | 13 |
| 413527 | 30 | 431795 | 15 | 479644 | 12 |
| 513527 | 27 | 531795 | 20 | 579644 | 14 |
| 613527 | 15 | 631795 | 23 | 679644 | 20 |
| 713527 | 9 | 731795 | 18 | 779644 | 15 |
| 813527 | 13 | 831795 | 17 | 879644 | 9 |
| 913527 | 8 | 931795 | 11 | 979644 | 5 |

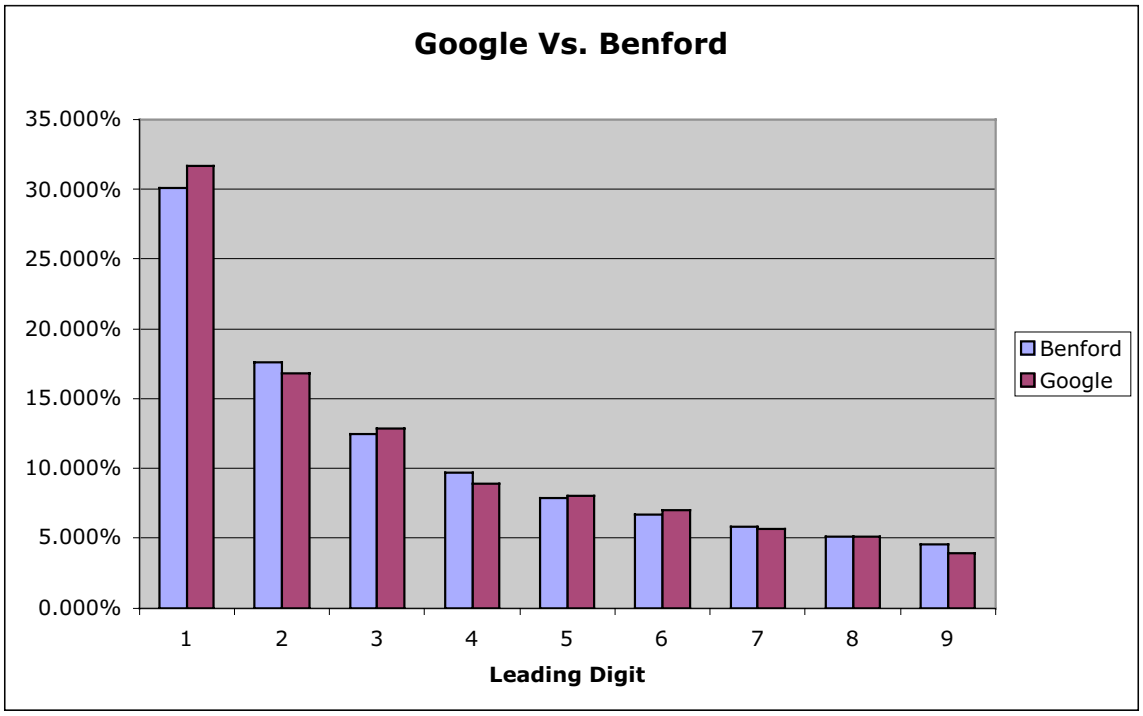
| 58316 Occurrences | | 85085 Occurrences | | 76781 Occurrences | |
|---|---|---|---|---|---|
| 158316 | 79 | 185085 | 52 | 176781 | 42 |
| 258316 | 51 | 285085 | 29 | 276781 | 22 |
| 358316 | 27 | 385085 | 23 | 376781 | 26 |
| 458316 | 19 | 485085 | 21 | 476781 | 31 |
| 558316 | 11 | 585085 | 20 | 576781 | 12 |
| 658316 | 10 | 685085 | 23 | 676781 | 6 |
| 758316 | 8 | 785085 | 14 | 776781 | 24 |
| 858316 | 5 | 885085 | 16 | 876781 | 19 |
| 958316 | 6 | 985085 | 2 | 976781 | 11 |

| 29285 Occurrences | | 39557 Occurrences | | 44557 Occurrences | |
|---|---|---|---|---|---|
| 129285 | 65 | 139557 | 67 | 144557 | 62 |
| 229285 | 34 | 239557 | 44 | 244557 | 37 |
| 329285 | 24 | 339557 | 26 | 344557 | 22 |
| 429285 | 16 | 439557 | 19 | 444557 | 18 |
| 529285 | 16 | 539557 | 34 | 544557 | 10 |
| 629285 | 9 | 639557 | 13 | 644557 | 24 |
| 729285 | 10 | 739557 | 8 | 744557 | 9 |
| 829285 | 7 | 839557 | 9 | 844557 | 10 |
| 929285 | 21 | 939557 | 3 | 944557 | 14 |

I found a sample of 2038 Google Numbers that satisfy the following empirical distribution (as compared with the Benford distribution):

| Numbers | Count | Empirical Percent | Benford |
|---|---|---|---|
| 1 | 645 | 31.65% | 30.1% |
| 2 | 342 | 16.78% | 17.6% |
| 3 | 262 | 12.86% | 12.5% |
| 4 | 181 | 8.88% | 9.7% |
| 5 | 164 | 8.05% | 7.9% |
| 6 | 143 | 7.02% | 6.7% |
| 7 | 115 | 5.64% | 5.8% |
| 8 | 105 | 5.15% | 5.1% |
| 9 | 81 | 3.97% | 4.6% |

As a graph we see:



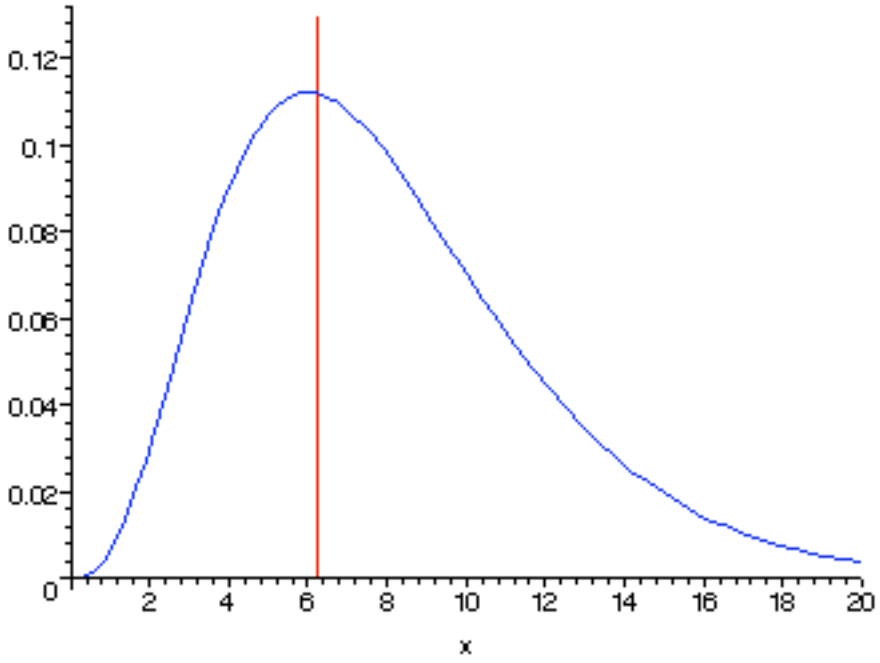

Looking at this graph, we see that these Google numbers are extremely Newcomb! To be honest, I had not expected my results to be this close. I felt that the Google numbers would have Newcomb tendencies balanced by some *social factors*. For example, for seven-digit numbers popular telephone prefixes might skew the Newcombness of the numbers. Maybe six was accidentally a really good choice, or maybe these social factors won't affect any "typical" sample. Clearly, further testing would be necessary to see if I simply got lucky with my choice of the word "nature" and my use of six digits. In any case, it seems I really did get lucky.

In order to quantify just how lucky, we can use the $\chi^2$-distribution. Namely, we compute:

| Numbers | Observed | Expected | (E-O)^2/E |
|---|---|---|---|
| 1 | 645 | 613.50 | 1.62 |
| 2 | 342 | 358.87 | 0.79 |
| 3 | 262 | 254.63 | 0.21 |
| 4 | 181 | 197.50 | 1.38 |
| 5 | 164 | 161.37 | 0.04 |

| | | | | |
|---|---|---|---|---|
| | 6 | 143 | 136.44 | 0.32 |
| | 7 | 115 | 118.19 | 0.09 |
| | 8 | 105 | 104.25 | 0.01 |
| | 9 | 81 | 93.25 | 1.61 |
| | Total | | Chi^2 | |
| ALL | | 2038 | | 6.063373 |

Hence we have that $\chi^2 = 6.06$. Notice there are 9-1=8 degrees of freedom, hence 8 is the expected value of this distribution. Looking at this value with regard to the $\chi^2$ distribution's graph with 8 degrees of freedom, we see that 6.06 is completely consistent with our data obeying the Benford distribution:




Admittedly, this is only one piece of evidence that our Google distribution is Benford. Perhaps our test was simply not powerful enough to reject the Benford distribution. To give some sense for this test's power, let us compare our data to the uniform distribution assigning a 1/9 probability to each leading digit. Under this uniform assumption we have:

| Numbers | Observed | Expected using Uniform | (E-O)^2/E |
|---|---|---|---|
| 1 | 645 | 226.44 | 773.65 |
| 2 | 342 | 226.44 | 58.97 |
| 3 | 262 | 226.44 | 5.58 |
| 4 | 181 | 226.44 | 9.12 |
| 5 | 164 | 226.44 | 17.22 |
| 6 | 143 | 226.44 | 30.75 |
| 7 | 115 | 226.44 | 54.85 |
| 8 | 105 | 226.44 | 65.13 |
| 9 | 81 | 226.44 | 93.42 |
| | Total | | Chi^2 |
| ALL | 2038 | | 1108.688 |

Needless to say, this $\chi^2 = 1108.7$ is big enough to safely reject the uniform assumption! In fact, such a value would be so unlikely that it is difficult to estimate. For example, if $\chi^2 = 63$ then there is about a 1 in a 10 billion chance that the uniform distribution is true and we would have observed such data. Hence, our $\chi^2 = 1108.7$ can be viewed as **literally impossible.**

Now the question becomes, "Why?" Looking at a few sample hits, we find that many of our Google numbers come from things like prices, membership numbers, account numbers, item numbers, and so on. These types of Google numbers all have something in common, namely each could arise from a growth process where a quantity is growing proportional to its size. For instance, in our above Google example, we find that the first hit is an "item number" from an e-bay type site. If a company's value is growing like money in a bank account, then perhaps the number of items that this company will be able to offer will also grow in such a manner. Let us recall how money in a bank account grows. Imagine we start with 78,123 dollars in a bank account that receives interest at a rate of 10 percent a year (For now imagine it is compounded

annually.) At the end of each year, we would have 1.1 times the amount we had at the beginning of the year. Namely our money grows at a rate proportional to how much we have. Let us follow our accounts value for 26 years:

| Year | Money |
|---|---|
| 0 | 78123 |
| 1 | 85935.3 |
| 2 | 94528.8 |
| 3 | 103982 |
| 4 | 114380 |
| 5 | 125818 |
| 6 | 138400 |
| 7 | 152240 |
| 8 | 167464 |
| 9 | 184210 |
| 10 | 202631 |
| 11 | 222894 |
| 12 | 245183 |
| 13 | 269702 |
| 14 | 296672 |
| 15 | 326339 |
| 16 | 358973 |
| 17 | 394870 |
| 18 | 434357 |
| 19 | 477793 |
| 20 | 525572 |
| 21 | 578130 |
| 22 | 635943 |
| 23 | 699537 |
| 24 | 769491 |
| 25 | 846440 |
| 26 | 931084 |
| 27 | 1024192 |

Notice that we have the following distribution of leading digits of our six digit bank account values:

| Leading Digit | Percent |
|---|---|
| 1 | 29.2% |
| 2 | 20.8% |
| 3 | 12.5% |
| 4 | 8.3% |
| 5 | 8.3% |
| 6 | 8.3% |
| 7 | 4.2% |
| 8 | 4.2% |
| 9 | 4.2% |

Notice this at least somewhat Newcombian.

Let us now carefully recall the relationship between such a growth model and the Benford distribution. (This is a well known relationship see for example http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml ). Using a pre-calculus book, we discover that things that grow proportional to their current value grow exponentially (like populations with loads of resources, or money in a bank account that is continuously compounded). Suppose such a quantity has reached its first six-digit number, in other words 100,000. Let us call this quantity *money*. Then t units of time after our money reaches 100,000 it has a value determined by the function $100000e^{rt}$. Here r is the rate at which our "money" grows, in other words, our interest. For simplicity, let us choose our unit of time here to be so that r=1. Hence solving for t, we find that each of the following bench marks occur at the listed times:

| Amount | Time |
|---|---|
| 100000 | 0.00 |
| 200000 | 0.69 |
| 300000 | 1.10 |
| 400000 | 1.39 |
| 500000 | 1.61 |
| 600000 | 1.79 |
| 700000 | 1.95 |
| 800000 | 2.08 |
| 900000 | 2.20 |

1000000 2.30

In particular, we spend a fraction of (2.08-1.95)/2.30=0.058 of our time with a leading digit of 7. If we collect this information in a table we find:

| Leading Digit | Percent Of Time |
|---|---|
| 1 | 30.10% |
| 2 | 17.61% |
| 3 | 12.49% |
| 4 | 9.69% |
| 5 | 7.92% |
| 6 | 6.69% |
| 7 | 5.80% |
| 8 | 5.12% |
| 9 | 4.58% |

This is exactly Benford's distribution! Hence, we would expect Google numbers to be Newcomb if they satisfied two criteria: first that every Google Number behaves like money with interest continuously compounded, and, second that the probability that a Google number is posted on the web is proportional to how long that quantity is meaningful.

In conclusion, it appears that Google numbers are indeed Newcomb. We even found a plausible explanation using the notion that many of these numbers are clearly coming from processes that grow at rates proportional to their current value. However, **not all numbers obey this growth law!** We are left with several questions:

1. Do our Google numbers really arise from such growth processes? How might we test this?
2. Did we just get lucky with our choice of 6 digits and the word “nature”? How might we test the degree of our luck?
3. How might we devise an efficient way to run a more thorough experiment? Namely, I had to plug 81 numbers into Google by hand in order to run this test. Can I get around this labor in order to run a more comprehensive experiment?