Benford's Law: Theory and Extensions

Austin Shapiro – Berkeley Math Circle – September 6, 2016

Logarithms of Powers of 3

n	$\log_{10}(3^n)$	n	$\log_{10}(3^n)$	n	$\log_{10}(3^n)$	n	$\log_{10}(3^n)$
0	0.0000	10	4.7712	20	9.5424	30	14.3136
1	0.4771	11	5.2483	21	10.0195	31	14.7908
2	0.9542	12	5.7255	22	10.4967	32	15.2679
3	1.4314	13	6.2026	23	10.9738	33	15.7450
4	1.9085	14	6.6797	24	11.4509	34	16.2221
5	2.3856	15	7.1568	25	11.9280	35	16.6992
6	2.8627	16	7.6339	26	12.4052	36	17.1764
7	3.3398	17	8.1111	27	12.8823	37	17.6535
8	3.8170	18	8.5882	28	13.3594	38	18.1306
9	4.2941	19	9.0653	29	13.8365	39	18.6077

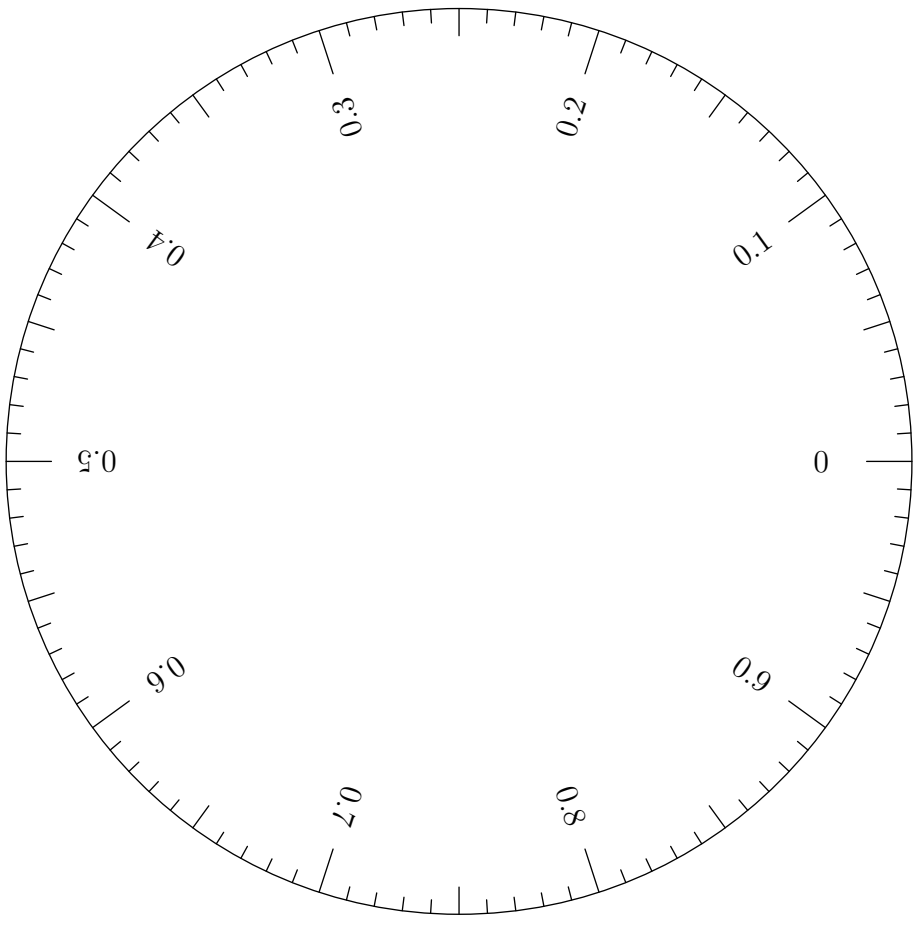
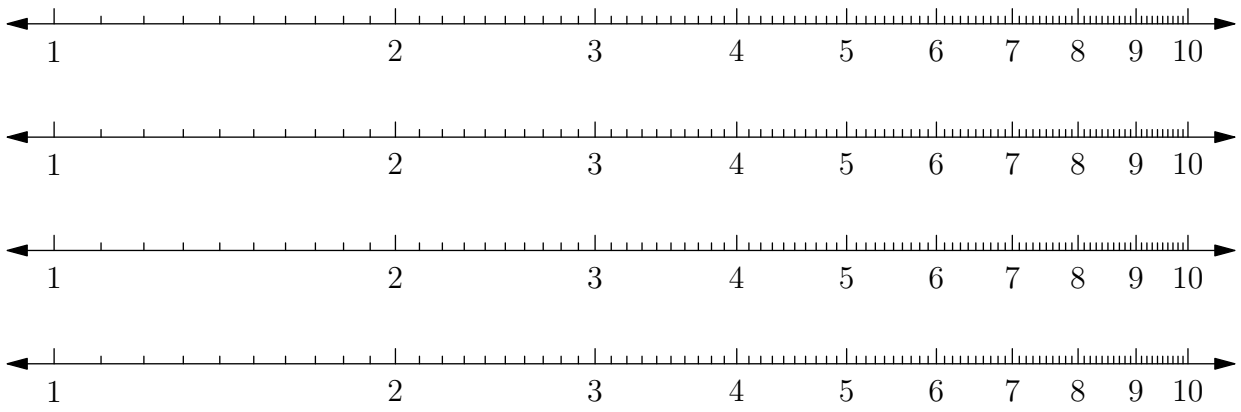


Figure 1. The "wheel", a circular number line.

Table of Logarithms

x	$\log_{10} x$	$\log_{10}(x+1) - \log_{10} x$
1	0.0000	0.3010
2	0.3010	0.1761
3	0.4771	
4		
5		
6		
7	0.8451	
8		
9		
10	1.0000	—

Logarithmic Number Lines



What Kinds of Data?

An infinite sequence of numbers $x_1, x_2, x_3, \ldots$ satisfies **Benford's Law** in base 10 if, for each digit $1 \leq d \leq 9$, the proportion of $x_1, x_2, \ldots, x_n$ beginning with digit d approaches a limit of $\log_{10}(d+1) - \log_{10} d$ as $n \to \infty$.

A finite set of data can't exactly satisfy the definition above, but may come close, like the populations of California counties did.

Which sequences and sets below do you expect to satisfy Benford's Law (exactly or approximately)?

Powers of π	☐ Yes	☐ No	Street addresses	☐ Yes	☐ No
Powers of $\sqrt[3]{10}$	☐ Yes	☐ No	Heights of human beings in inches	☐ Yes	☐ No
Primes	☐ Yes	☐ No	Phone numbers	☐ Yes	☐ No
Perfect squares	☐ Yes	☐ No	All the numbers in all the spreadsheets created by the accounting offices of a large company	☐ Yes	☐ No
Factorials	☐ Yes	☐ No			

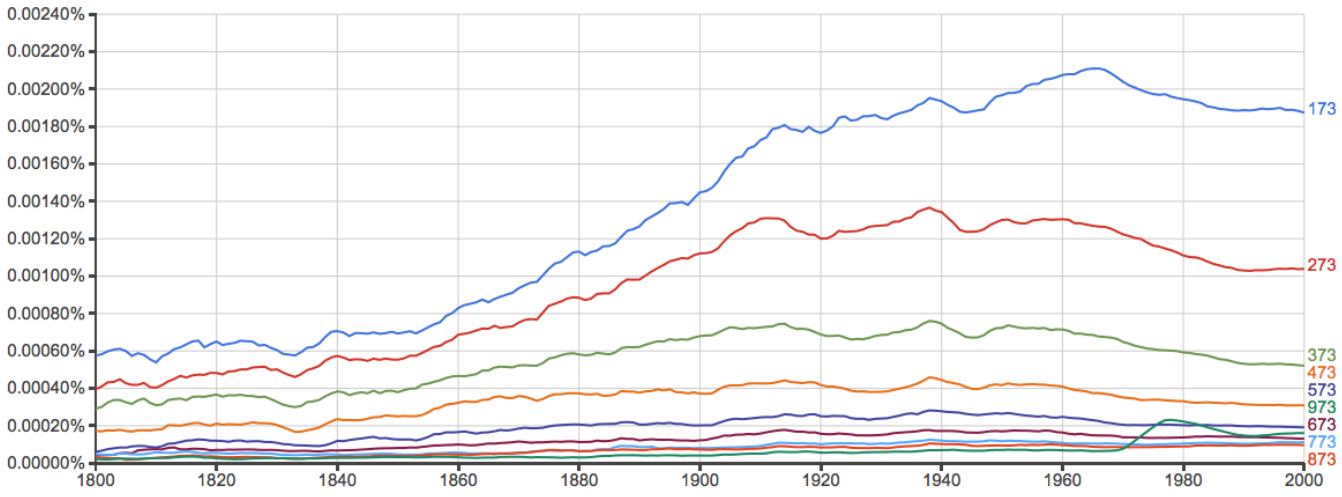


Figure 2. Frequency with which the numerals 173, 273, 373, ..., 973 have appeared in print over time, as a percentage of words or word-like tokens. (Source: Google Ngram Viewer)

Histogram of County Populations

1–2 (thou.)										
2–4										
4–8										
8–16										
16–32										
32–64										
64–128										
128–256										
256–512										
512–1024										
1024–2048										
2048–4096										
4096–8192										
8192–										

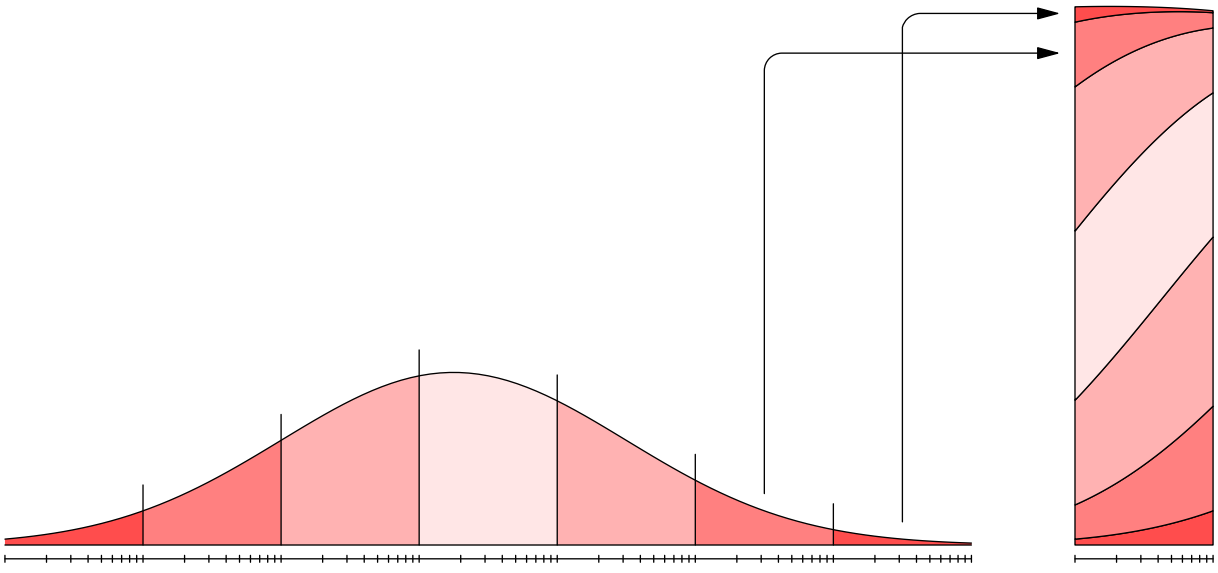


Figure 3. A normal distribution sliced at regular intervals, and the slices stacked. The tickmarks underneath show a logarithmic scale.

Proving Benford's Law for Powers of 10^α ($\alpha \notin \mathbb{Q}$)

Proposition 1. If n is a positive integer and $\alpha = \log_{10} n$ is not an integer, then α is irrational.

Definition. Let x be a real number. The *integer part* of x, denoted by $\lfloor x \rfloor$, is the greatest integer less than or equal to x. The *fractional part* of x, denoted by $\{x\}$, is defined to be $x - \lfloor x \rfloor$. For example, $\lfloor 3.85 \rfloor = 3$ and $\{3.85\} = .85$, while $\lfloor -1.7 \rfloor = -2$ and $\{-1.7\} = .3$.

Definition. The *wheel*[1] is a circle of unit circumference formed by joining the ends of the interval $[0, 1]$ so that numbers increase in the counterclockwise direction (see the diagram on page 1). Every real number x is assigned the point on the wheel corresponding to $\{x\}$.

The *distance* $d(x \to y)$ is defined as the length of a counterclockwise arc on the wheel from the point representing x to the point representing y. For example, $d(.1 \to .7) = .6$, while $d(.7 \to .1) = .4$. (We could equivalently define $d(x \to y)$ as $\{y - x\}$.)

Definition. Let A be a set of points on the wheel. Then A is *dense on the wheel* if every arc of positive length on the wheel contains at least one point from A.

Notice that a dense subset of the wheel is necessarily infinite, but an infinite subset is not necessarily dense. (Does a dense set necessarily include *all* points on the wheel?)

Proposition 2. Let $A = \{0, \{\alpha\}, \{2\alpha\}, \{3\alpha\}, \ldots\}$. Then we have two cases:

(i) If α is rational, let $\alpha = \dfrac{p}{q}$ in lowest terms. Then $A = \left\{0, \dfrac{1}{q}, \dfrac{2}{q}, \ldots, \dfrac{q-1}{q}\right\}$.

(ii) If α is irrational, then A is dense on the wheel.

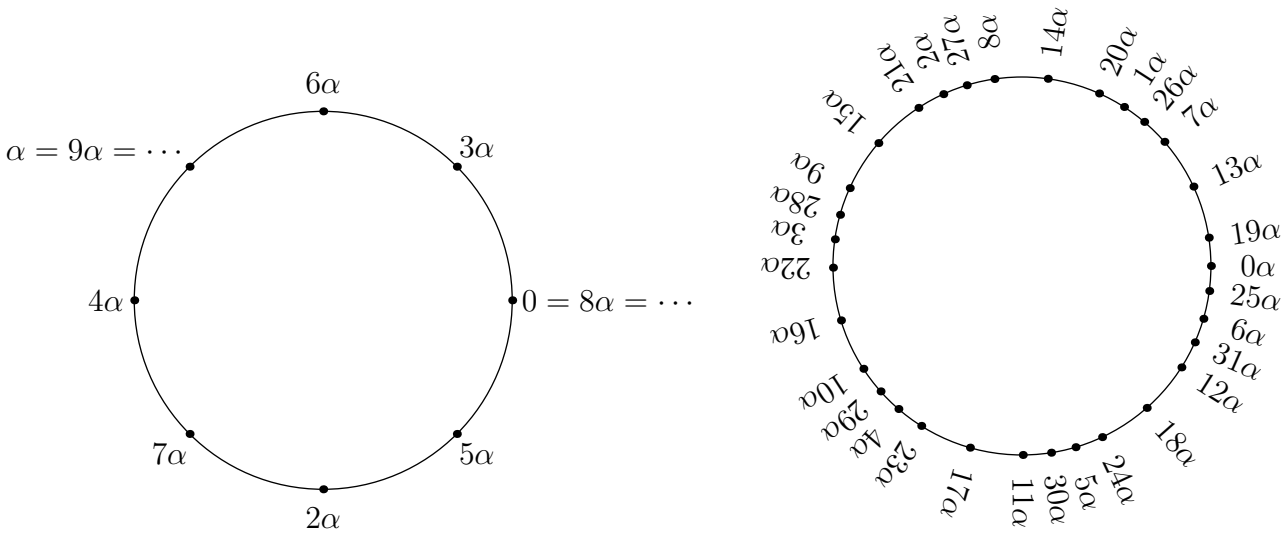


Figure 4. Multiples of $\alpha = 3/8$ (left), $\alpha = 1/(2\pi)$ (right). I stopped at 31α, but if we continued going around the circle on the right drawing dots of some positive thickness, the dots would eventually darken the whole circle.

[1] The wheel is more formally known as $\mathbb{R}/\mathbb{Z}$.

Proposition 3 (Equidistribution Theorem).

Let α be irrational. For each integer $n \geq 1$, let $A_n = \{0, \{\alpha\}, \{2\alpha\}, \{3\alpha\}, \ldots, \{(n-1)\alpha\}\}$. Let I be an arc of length t on the wheel. Then

$$\lim_{n\to\infty} \frac{|A_n \cap I|}{n} = t.$$

The Equidistribution Theorem says that, in the long run, the proportion of multiples of α whose fractional parts lie in a given interval is proportional to the size of the interval. For example, the proportion of multiples of $\sqrt{2}$ whose fractional parts are between 0.33 and 0.37 approaches 0.04. Essentially, the multiples of α "cover the circle evenly".

Corollary. Let n be a positive integer which is not a perfect power of 10, or more generally let $n = 10^\alpha$ where α is irrational. Then the sequence $1, n, n^2, n^3, \ldots$ satisfies Benford's Law in base 10.

Extensions and Problems

1. What is the analogous statement of Benford's Law in base b?
2. For sequences like $1, n, n^2, \ldots$, is there a limiting distribution for the *second* digits? Third digits? Last digits? All digits?
3. You think the body lengths in millimeters of all the animals in the forest (from tiny insects to large mammals) might obey Benford's Law, but your uncle doesn't believe it; he insists that first digits from 1 to 9 should appear about equally often. Short of actually gathering the data, how could you show him that he's being awfully silly?
4. Let α be an arbitrary irrational number. We know that there's a multiple of α whose fractional part is between 0 and 0.001 (say). How would you go about *finding* such a multiple? (Ideally, you wouldn't test every multiple in order; that could take a long time.)
5. There's a special way to do the previous problem if $\alpha = \sqrt{2}$. As a hint, $0 < (\sqrt{2}-1)^8 < 0.001$. How does this lead you to a multiple of $\sqrt{2}$ that fits the bill? Does this trick work for any other irrational numbers?

⋆ 6. Take a close look at the right side of Figure 4. The dots divide the circle into 32 arcs; those arcs are not all the same length, but many of them *are* the same length. How many different lengths of arc are there?

A theorem of Steinhaus says that the points of A_n (as defined in Proposition 3) always cut the wheel into arcs of at most a certain number of different lengths—the number you found in this example. For a big challenge, try proving the theorem.

⋆ 7. Investigate the sequences listed on page 3, or your favorite sequences. Which ones satisfy Benford's Law? Can you prove it?

A Problem Dr. Shapiro Can't Solve

Does the sequence of all positive integers satisfy Benford's Law?

That may sound ridiculous—surely the integers *as a whole* use every first digit from 1 to 9 equally. But really, it depends on your stopping point:

- Only 11 out of the first 99 positive integers start with 1. That's a proportion of $\frac{1}{9}$.
- But 111 out of the first 199 positive integers start with 1. That's well over half!

Stopping at 199 may seem biased and arbitrary, but then, stopping at 99 is also biased. Of all stopping points, 99 (or any number composed of all 9's) yields the lowest possible proportion of initial 1's. If you stop *anywhere else*, there will be more numbers up to that point that start with 1's than 9's.

Unfortunately, there's no neutral way to pick a stopping point. There is no such thing as a "random positive integer" (at least, not if you want every integer to have equal probability of being chosen). And if we look at the proportion of the first n integers that begin with a given digit, this proportion doesn't approach a limit as $n \to \infty$; it just oscillates between two bounds.

But I had another idea.

Let S be some subset of the positive integers—perhaps, the ones that begin with 1. Define

$$\chi_S(n) = \begin{cases} 1 & \text{if } n \in S \\ 0 & \text{if } n \notin S \end{cases}.$$

Then let

$$d_S^{(1)}(n) = \frac{\chi_S(1) + \chi_S(2) + \cdots + \chi_S(n)}{n}.$$

Got that? $d^{(1)}(n)$ is simply the proportion of integers from 1 to n that belong to S. But now we average again:

$$d_S^{(2)}(n) = \frac{d_S^{(1)}(1) + d_S^{(1)}(2) + \cdots + d_S^{(1)}(n)}{n}$$

And again:

$$d_S^{(3)}(n) = \frac{d_S^{(2)}(1) + d_S^{(2)}(2) + \cdots + d_S^{(2)}(n)}{n}$$

... and so on. Each time we average, the new function is smoother than the old one (see Fig. 5). Thus, for $S = \{\text{numbers that start with 1}\}$, $d_S^{(1)}(n)$ repeatedly oscillates between 0.11 and 0.56, but $d_S^{(2)}(n)$ only oscillates (after a while) between 0.23 and 0.36, and $d_S^{(3)}(n)$ oscillates in an even narrower range.

Does the (eventual) range of $d_S^{(k)}(n)$ close in on $\log_{10} 2$ (≈ 0.301) as $k \to \infty$? It appears so; in fact, I've calculated $d_S^{(k)}(n)$ for a variety of sets S (e.g., numbers that start with 3 in base 5), and when k and n are large enough, the results always look like Benford's Law. But I don't know how to prove it!

Maybe you'll have an idea that could crack this problem. If so—or if you want to discuss anything else in this handout—just drop me a line at `ashapiro` (dot) `proofschool` (dot) `org`!

n	$\chi(n)$	$d^{(1)}(n)$	$d^{(2)}(n)$	$d^{(3)}(n)$
1	1	1.000	1.000	1.000
2	0	0.500	0.750	0.875
3	0	0.333	0.611	0.787
4	0	0.250	0.521	0.720
5	0	0.200	0.457	0.668
6	0	0.167	0.408	0.624
7	0	0.143	0.370	0.588
8	0	0.125	0.340	0.557
9	0	0.111	0.314	0.530
10	1	0.200	0.303	0.507
11	1	0.273	0.300	0.489
12	1	0.333	0.303	0.473
13	1	0.385	0.309	0.461
14	1	0.429	0.318	0.450
15	1	0.467	0.328	0.442
16	1	0.500	0.338	0.436
17	1	0.529	0.350	0.431
18	1	0.556	0.361	0.427
19	1	0.579	0.373	0.424
20	0	0.550	0.381	0.422
21	0	0.524	0.388	0.420
22	0	0.500	0.393	0.419
23	0	0.478	0.397	0.418
24	0	0.458	0.400	0.417

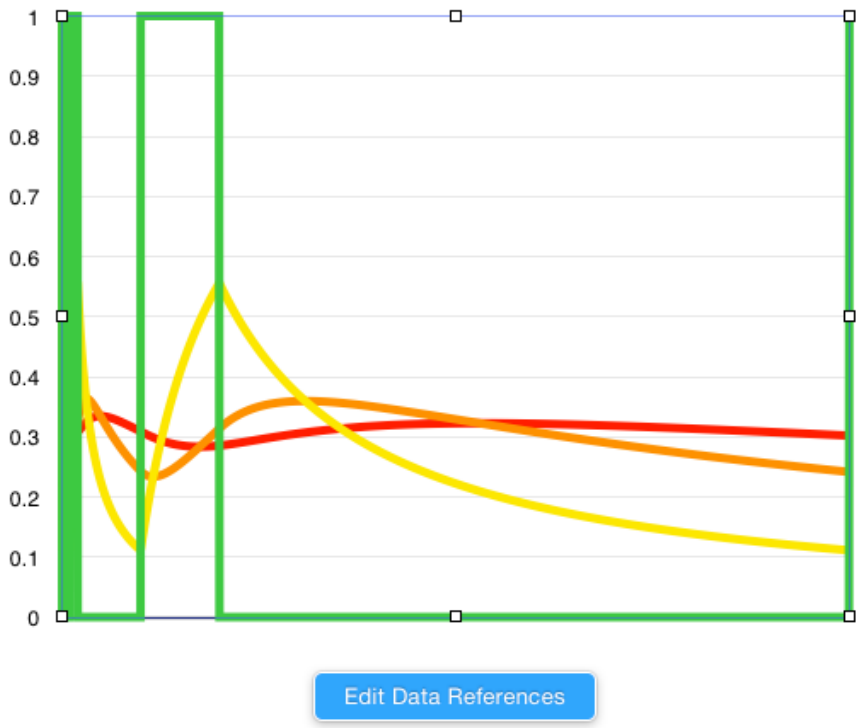


FIGURE 5. The averaging process makes each function smoother than the last. The graph on the right shows the first 10,000 values of each function.