5.39: Benford's Law

Benford's law refers to probability distributions that seem to govern the significant digits in real data sets. The law is named for the American physicist and engineer Frank Benford, although the “law” was actually discovered earlier by the astronomer and mathematician Simon Newcomb.

To understand Benford's law, we need some preliminaries. Recall that a positive real number x can be written uniquely in the form $x = y \cdot 10^n$ (sometimes called *scientific notation*) where $y \in \left[\frac{1}{10}, 1\right)$ is the *mantissa* and $n \in \mathbb{Z}$ is the *exponent* (both of these terms are *base 10*, of course). Note that

$$\log x = \log y + n \tag{5.39.1}$$

where the logarithm function is the base 10 *common logarithm* instead of the usual base e *natural logarithm*. In the old days BC (before calculators), one would compute the logarithm of a number by looking up the logarithm of the mantissa in a *table of logarithms*, and then adding the exponent. Of course, these remarks apply to any base $b > 1$, not just base 10. Just replace 10 with b and the common logarithm with the base b logarithm.

Distribution of the Mantissa

Distribution Functions

Suppose now that X is a number selected at random from a certain data set of positive numbers. Based on empirical evidence from a number of different types of data, Newcomb, and later Benford, noticed that the mantissa Y of X seemed to have distribution function $F(y) = 1 + \log y$ for $y \in [1/10, 1)$. We will generalize this to an arbitrary base $b > 1$.

The *Benford mantissa distribution* with base $b \in (1, \infty)$, is a continuous distribution on $[1/b, 1)$ with distribution function F given by

$$F(y) = 1 + \log_b y, \quad y \in [1/b, 1) \tag{5.39.2}$$

The special case $b = 10$ gives the *standard Benford mantissa distribution*.

Proof

Note that F is continuous and strictly increasing on $[1/b, 1)$ with $F(1/b) = 0$ and $F(1) = 1$.

The probability density function f is given by

$$f(y) = \frac{1}{y \ln b}, \quad y \in [1/b, 1) \tag{5.39.3}$$

1. f is decreasing with mode $y = \frac{1}{b}$.
2. f is concave upward.

Proof

These results follow from the CDF F above and standard calculus. Recall that $f = F'$.

Open the Special Distribution Simulator and select the Benford mantissa distribution. Vary the base b and note the shape of the probability density function. For various values of b, run the simulation 1000 times and compare the empirical density function to the probability density function.

The quantile function F^{-1} is given by

$$F^{-1}(p) = \frac{1}{b^{1-p}}, \quad p \in [0, 1] \tag{5.39.4}$$

1. The first quartile is $F^{-1}\left(\frac{1}{4}\right) = \frac{1}{b^{3/4}}$
2. The median is $F^{-1}\left(\frac{1}{2}\right) = \frac{1}{\sqrt{b}}$
3. The third quartile is $F^{-1}\left(\frac{3}{4}\right) = \frac{1}{b^{1/4}}$

Proof

The formula for $F^{-1}(p)$ follows by solving $F(x) = p$ for x in terms of p.

Numerical values of the quartiles for the standard (base 10) distribution are given in an exercise below.

Open the special distribution calculator and select the Benford mantissa distribution. Vary the base and note the shape and location of the distribution and probability density functions. For selected values of the base, compute the median and the first and third quartiles.

Moments

Assume that Y has the Benford mantissa distribution with base $b \in (1, \infty)$.

The moments of Y are

$$\mathbb{E}(Y^n) = \frac{b^n - 1}{n b^n \ln b}, \quad n \in (0, \infty) \tag{5.39.5}$$

Proof

For $n > 0$,

$$\mathbb{E}(Y^n) = \int_{1/b}^1 y^n \frac{1}{y \ln b} dy = \frac{1}{\ln b} \int_{1/b}^1 y^{n-1} dy = \frac{1 - 1/b^n}{n \ln(b)} \tag{5.39.6}$$

Note that for fixed $n > 0$, $\mathbb{E}(Y^n) \to 1$ as $b \downarrow 1$ and $\mathbb{E}(Y^n) \to 0$ as $b \to \infty$. We will learn more about the limiting distribution below. The mean and variance follow easily from the general moment result.

Mean and variance

1. The mean of Y is

$$\mathbb{E}(Y) = \frac{b - 1}{b \ln b} \tag{5.39.7}$$

2. the variance of Y is

$$\text{var}(Y) = \frac{b - 1}{b^2 \ln b} \left[\frac{b + 1}{2} - \frac{b - 1}{\ln b} \right] \tag{5.39.8}$$

Numerical values of the mean and variance for the standard (base 10) distribution are given in an exercise below.

In the Special Distribution Simulator, select the Benford mantissa distribution. Vary the base b and note the size and location of the mean $\pm$ standard deviation bar. For selected values of b, run the simulation 1000 times and compare the empirical mean and standard deviation to the distribution mean and standard deviation.

Related Distributions

The Benford mantissa distribution has the usual connections to the standard uniform distribution by means of the distribution function and quantile function given above.

Suppose that $b \in (1, \infty)$.

1. If U has the standard uniform distribution then $Y = b^{-U}$ has the Benford mantissa distribution with base b.
2. If Y has the Benford mantissa distribution with base b then $U = -\log_b Y$ has the standard uniform distribution.

Proof

1. If U has the standard uniform distribution then so does $1 - U$ and hence $Y = F^{-1}(1 - U) = b^{-U}$ has the Benford mantissa distribution with base b.
2. The CDF F is strictly increasing on $[b^{-1}, 1)$. Hence if Y has the Benford mantissa distribution with base b then $F(Y) = 1 + \log_b(Y)$ has the standard uniform distribution and hence so does $1 - F(Y) = -\log_b(Y)$.

Since the quantile function has a simple closed form, the Benford mantissa distribution can be simulated using the random quantile method.

Open the random quantile experiment and select the Benford mantissa distribution. Vary the base b and note again the shape and location of the distribution and probability density functions. For selected values of b, run the simulation 1000 times and compare the empirical density function, mean, and standard deviation to their distributional counterparts.

Also of interest, of course, are the limiting distributions of Y with respect to the base b.

The Benford mantissa distribution with base $b \in (1, \infty)$ converges to

1. Point mass at 1 as $b \downarrow 1$.
2. Point mass at 0 as $b \uparrow \infty$.

Proof

Note that the CDF of Y above can be written as $F(y) = 1 + \ln(y) / \ln(b)$ for $1/b \le y < 1$, and of course we also have $F(y) = 0$ for $y < 1/b$ and $F(y) = 1$ for $y \ge 1$.

1. As $b \downarrow 1$, $1/b \uparrow 1$, and $1 + \ln(y) / \ln(b) \to 1$, so in the limit we have $F(y) = 0$ for $y < 1$ and $F(y) = 1$ for $y > 1$.
2. As $b \uparrow \infty$, $1/b \downarrow 0$, and again $1 + \ln(y) / \ln(b) \to 1$, so in the limit we have $F(y) = 0$ for $y < 0$ and $F(y) = 1$ for $y > 0$).

Since the probability density function is bounded on a bounded support interval, the Benford mantissa distribution can also be simulated via the rejection method.

Open the rejection method experiment and select the Benford mantissa distribution. Vary the base b and note again the shape and location of the probability density functions. For selected values of b, run the simulation 1000 times and compare the empirical density function, mean, and standard deviation to their distributional counterparts.

Distributions of the Digits

Assume now that the base is a positive integer $b \in \{2, 3, \ldots\}$, which of course is the case in standard number systems. Suppose that the sequence of digits of our mantissa Y (in base b) is $(N_1, N_2, \ldots)$, so that

$$Y = \sum_{k=1}^{\infty} \frac{N_k}{b^k} \tag{5.39.9}$$

Thus, our *leading digit* N_1 takes values in $\{1, 2, \ldots, b-1\}$, while each of the other *significant digits* takes values in $\{0, 1, \ldots, b-1\}$. Note that $(N_1, N_2, \ldots)$ is a stochastic process so at least we would like to know the *finite dimensional distributions*. That is, we would like to know the joint probability density function of the first k digits for every $k \in \mathbb{N}_+$. But let's start, appropriately enough, with the *first digit law*. The leading digit is the most important one, and fortunately also the easiest to analyze mathematically.

First Digit Law

N_1 has probability density function g_1 given by $g_1(n) = \log_b\left(1 + \frac{1}{n}\right) = \log_b(n+1) - \log_b(n)$ for $n \in \{1, 2, \ldots, b-1\}$. The density function g_1 is decreasing and hence the mode is $n = 1$.

Proof

Note that $N_1 = n$ if and only if $\frac{n}{b} \le Y < \frac{n+1}{b}$ for $n \in \{1, 2, \ldots, b-1\}$. Hence using the PDF of Y above,

$$\mathbb{P}(N_1 = n) = \int_{n/b}^{(n+1)/b} \frac{1}{y \ln b} dy = \log_b\left(\frac{n+1}{b}\right) - \log_b\left(\frac{n}{b}\right) = \log_b(n+1) - \log_b(n) \tag{5.39.10}$$

Note that when $b = 2$, $N_1 = 1$ deterministically, which of course has to be the case. The first significant digit of a number in base 2 must be 1. Numerical values of g_1 for the standard (base 10) distribution are given in an exercise below.

In the Special Distribution Simulator, select the Benford first digit distribution. Vary the base b with the input control and note the shape of the probability density function. For various values of b, run the simulation 1000 times and compare the empirical density function to the probability density function.

N_1 has distribution function G_1 given by $G_1(x) = \log_b(\lfloor x \rfloor + 1)$ for $x \in [1, b-1]$.

Proof

Using the PDF of N_1 above note that

$$G_1(n) = \sum_{k=1}^{n} [\log_b(k+1) - \log_b(k)] = \log_b(n+1), \quad n \in \{1, 2, \ldots, b-1\} \tag{5.39.11}$$

More generally, $G_1(x) = G_1(\lfloor x \rfloor)$ for $x \in [1, b-1]$

N_1 has quantile function G_1^{-1} given by $G_1^{-1}(p) = \lceil b^p - 1 \rceil$ for $p \in (0, 1]$.

1. The first quartile is $\lceil b^{1/4} - 1 \rceil$.

2. The median is $\lceil b^{1/2} - 1 \rceil$.
3. The third quartile is $\lceil b^{3/4} - 1 \rceil$.

Proof

As usual, the formula for $G_1^{-1}(p)$ follows from the CDF G, by solving $p = G(x)$ for x in terms of p.

Numerical values of the quantiles for the standard (base 10) distribution are given in an exercise below.

Open the special distribution calculator and choose the Benford first digit distribution. Vary the base and note the shape and location of the distribution and probability density functions. For selected values of the base, compute the median and the first and third quartiles.

For the most part the moments of N_1 do not have simple expressions. However, we do have the following result for the mean.

$\mathbb{E}(N_1) = (b-1) - \log_b[(b-1)!]$.

Proof

From the PDF of N_1 above and using standard properties of the logarithm,

$$\mathbb{E}(N_1) = \sum_{n=1}^{b-1} n \log_b\left(\frac{n+1}{n}\right) = \log_b\left[\prod_{n=1}^{b-1}\left(\frac{n+1}{n}\right)^n\right] \tag{5.39.12}$$

The product in the displayed equation simplifies to $b^{b-1}/(b-1)!$, and the base b logarithm of this expression is $(b-1) - \log_b[(b-1)!]$.

Numerical values of the mean and variance for the standard (base 10) distribution are given in an exercise below.

Opne the Special Distribution Simulator and select the Benford first digit distribution. Vary the base b with the input control and note the size and location of the mean $\pm$ standard deviation bar. For various values of b, run the simulation 1000 times and compare the empirical mean and standard deviation to the distribution mean and standard deviation..

Since the quantile function has a simple, closed form, the Benford first digit distribution can be simulated via the random quantile method.

Open the random quantile experiment and select the Benford first digit distribution. Vary the base b and note again the shape and location of the probability density function. For selected values of the base, run the experiment 1000 times and compare the empirical density function, mean, and standard deviation to their distributional counterparts.

Higher Digits

Now, to compute the joint probability density function of the first k significant digits, some additional notation will help.

If $n_1 \in \{1, 2, \ldots, b-1\}$ and $n_j \in \{0, 1, \ldots, b-1\}$ for $j \in \{2, 3, \ldots, k\}$, let

$$[n_1\, n_2 \cdots n_k]_b = \sum_{j=1}^{k} n_j b^{k-j} \tag{5.39.13}$$

Of course, this is just the base b version of what we do in our standard base 10 system: we represent integers as strings of digits between 0 and 9 (except that the first digit cannot be 0). Here is a base 5 example:

$$[324]_5 = 3 \cdot 5^2 + 2 \cdot 5^1 + 4 \cdot 5^0 = 89 \tag{5.39.14}$$

The joint probability density function h_k of $(N_1, N_2, \ldots, N_k)$ is given by

$$h_k(n_1, n_2, \ldots, n_k) = \log_b\left(1 + \frac{1}{[n_1\, n_2 \cdots n_k]_b}\right), \quad n_1 \in \{1, 2, \ldots, b-1\}, (n_2, \ldots, n_k) \in \{2, \ldots, b-1\}^{k-1} \tag{5.39.15}$$

Proof

Note that $\{N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\} = \{l \le Y < u\}$ where

$$l = \frac{[n_1\, n_2 \cdots, n_k]_b}{b^k}, \; u = \frac{[n_1\, n_2 \cdots n_k]_b + 1}{b^k} \tag{5.39.16}$$

Hence using the PDF of Y and properties of logarithms,

$$h_k(n_1, n_2, \ldots, n_k) = \int_l^u \frac{1}{y \ln(b)} dy = \log_b(u) - \log_b(l) = \log_b([n_1\, n_2\, \cdots\, n_k]_b + 1) - \log_b([n_1\, n_2\, \cdots, n_k]_b) \tag{5.39.17}$$

The probability density function of (N_1, N_2) in the standard (base 10) case is given in an exercise below. Of course, the probability density function of a given digit can be obtained by summing the joint probability density over the unwanted digits in the usual way. However, except for the first digit, these functions do not reduce to simple expressions.

The probability density function g_2 of N_2 is given by

$$g_2(n) = \sum_{k=1}^{b-1} \log_b\left(1 + \frac{1}{[k\, n]_b}\right) = \sum_{k=1}^{b-1} \log_b\left(1 + \frac{1}{k\, b + n}\right), \quad n \in \{0, 1, \ldots, b-1\} \tag{5.39.18}$$

The probability density function of N_2 in the standard (base 10) case is given in an exercise below.

Theoretical Explanation

Aside from the empirical evidence noted by Newcomb and Benford (and many others since), why does Benford's law work? For a theoretical explanation, see the article A Statistical Derivation of the Significant Digit Law by Ted Hill.

Computational Exercises

In the following exercises, suppose that Y has the standard Benford mantissa distribution (the base 10 decimal case), and that $(N_1, N_2, \ldots)$ are the digits of Y.

Find each of the following for the mantissa Y

1. The density function f.
2. The mean and variance
3. The quartiles

Answer

1. $f(y) = \frac{1}{0.2303 y}, \quad y \in \left[\frac{1}{10}, 1\right)$
2. $\mathbb{E}(Y) = 0.3909$, $\text{var}(Y) = 0.0622$
3. $q_1 = 0.1778$, $q_2 = 0.3162$, $q_3 = 0.5623$

For N_1, find each of the following numerically

1. The probability density function
2. The mean and variance
3. The quartiles

Answer

1.

n	$\mathbb{P}(N_1 = n)$
1	0.3010
2	0.1761
3	0.1249
4	0.0969
5	0.0792
6	0.0669
7	0.0580
8	0.0512
9	0.0458

2. $\mathbb{E}(N_1) = 3.4402$, $\text{var}(N_1) = 6.0567$
3. $q_1 = 1$, $q_2 = 3$, $q_3 = 5$

Explicitly compute the values of the joint probability density function of (N_1, N_2).

Answer

$\mathbb{P}(N_1 = n_1, N_2 = n_2)$	$n_1 = 1$	2	3	4	5	6	7	8	9
$n_2 = 0$	0.0414	0.0212	0.0142	0.0107	0.0086	0.0072	0.0062	0.0054	0.0048
1	0.0378	0.0202	0.0138	0.0105	0.0084	0.0071	0.0061	0.0053	0.0047
2	0.0348	0.0193	0.0134	0.0102	0.0083	0.0069	0.0060	0.0053	0.0047
3	0.0322	0.0185	0.0130	0.0100	0.0081	0.0068	0.0059	0.0052	0.0046
4	0.0300	0.0177	0.0126	0.0098	0.0080	0.0067	0.0058	0.0051	0.0046
5	0.0280	0.0170	0.0122	0.0092	0.0078	0.0066	0.0058	0.0051	0.0045
6	0.0263	0.0164	0.0119	0.0093	0.0077	0.0065	0.0057	0.0050	0.0045
7	0.0248	0.0158	0.0116	0.0091	0.0076	0.0064	0.0056	0.0050	0.0045
8	0.0235	0.0152	0.0113	0.0090	0.0074	0.0063	0.0055	0.0049	0.0044
9	0.0223	0.0147	0.0110	0.0088	0.0073	0.0062	0.0055	0.0049	0.0044

For N_2, find each of the following numerically

1. The probability density function
2. $\mathbb{E}(N_2)$
3. $\text{var}(N_2)$

Answer

1.

n	$\mathbb{P}(N_2 = n)$
0	0.1197
1	0.1139
2	0.1088
3	0.1043
4	0.1003
5	0.0967
6	0.0934
7	0.0904
8	0.0876
9	0.0850

2. $\mathbb{E}(N_2) = 4.1847$
3. $\text{var}(N_2) = 0.8254$

Comparing the result for N_1 and the result result for N_2 , note that the distribution of N_2 is flatter than the distribution of N_1. In general, it turns out that distribution of N_k converges to the uniform distribution on $\{0, 1, \ldots, b-1\}$ as $k \to \infty$. Interestingly, the digits are dependent.

N_1 and N_2 are dependent.

Proof

This result follows from the joint PDF, the marginal PDF of N_1, and the marginal PDF of N_2 above.

Find each of the following.

1. $\mathbb{P}(N_1 = 5, N_2 = 3, N_3 = 1)$
2. $\mathbb{P}(N_1 = 3, N_2 = 1, N_3 = 5)$
3. $\mathbb{P}(N_1 = 1, N_2 = 3, N_3 = 5)$
