15. 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 $x = y\, 10^n$ (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

$$\log(x) = \log(y) + n$$

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

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 $y \in \left[\frac{1}{10}, 1\right)$. We will generalize this to an arbitrary base $b > 1$. Thus, let

$$F(y) = 1 + \log_b(y), \quad \frac{1}{b} \le y < 1$$

1. Show that F satisfies the mathematical properties of a distribution function for a continuous distribution on $\left[\frac{1}{b}, 1\right)$.

2. Note that the corresponding probability density function is $f(y) = \frac{1}{y \ln(b)}$ for $y \in \left[\frac{1}{b}, 1\right)$.

3. Show that
   a. $\mathbb{E}(Y) = \frac{b-1}{b \ln(b)}$
   b. $\operatorname{var}(Y) = \frac{b-1}{b^2 \ln(b)} \left(\frac{b+1}{2} - \frac{b-1}{\ln(b)}\right)$

4. For the standard base 10 decimal case
   a. Sketch the graph of f.

b. Compute the mean and variance explicitly.

Distribution of the Digits

Assume now that the base is a positive integer $b \in \mathbb{N}_+$, 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

$$Y = \sum_{k=1}^{\infty} N_k \, b^{-k}$$

Thus, our **leading digit** N_1 takes values in $\{1, 2, \ldots, b-1\}$, while each of the other **significant digits** takes values in $\{0, 1, \ldots, b-1\}$. Our goal is to compute the joint probability density function of the first k digits. But let's start, appropriately enough, with the **first digit law**, the discrete probability density function of the leading digit:

5. Show that $\mathbb{P}(N_1 = n) = \log_b\left(1 + \frac{1}{n}\right) = \log_b(n+1) - \log_b(n)$ for $n \in \{1, 2, \ldots, b-1\}$. *Hint*: Note that $N_1 = n$ if and only if $Y \in \left[\frac{n}{b}, \frac{n+1}{b}\right)$.

6. Consider the standard base 10 decimal case.

   a. Explicitly compute the values of the probability density function of N_1 and sketch the graph.
   b. Find $\mathbb{E}(N_1)$
   c. Find $\text{var}(N_1)$

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

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

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:

$$[324]_5 = 3\,5^2 + 2\,5^1 + 4\,5^0$$

7. Show that

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k) = \log_b\left(1 + \frac{1}{[n_1 n_2 \cdots n_k]_b}\right)$$

Hint: Note that $\{N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\} = \left\{Y \in \left[\frac{[n_1 n_2 \cdots n_k]_b}{b^k}, \frac{[n_1 n_2 \cdots n_k]_b + 1}{b^k}\right)\right\}$. Now use the distribution function of Y and properties of logarithms.

8. In the standard base 10 decimal case, explicitly compute the values of the joint probability density function of

(N_1, N_2).

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.

9. Show that

$$\mathbb{P}(N_2 = n) = \sum_{k=1}^{b-1} \log_b\left(1 + \frac{1}{[kn]_b}\right) = \sum_{k=1}^{b-1} \log_b\left(1 + \frac{1}{k\,b + n}\right), \quad n \in \{0, 1, \ldots, b-1\}$$

10. Consider the standard base 10 decimal case.

a. Explicitly compute the values of the probability density function of N_2, and sketch the graph.
b. Find $\mathbb{E}(N_2)$
c. Find $\text{var}(N_2)$

Comparing Exercise 6 and Exercise 10, 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.

11. Use the results of Exercise 6, Exercise 8, and Exercise 10 to show that N_1 and N_2 are dependent in the standard base 10 decimal case.

12. In the standard base 10 decimal case, find each of the following.

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

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 Significant Digit Law, by Ted Hill.