Probability and stats mop-up unit

- Statistical inference and related concepts (last week Monday)
- Hypothesis testing and chi-squared tests (last week Wednesday)
- Benford's law (**today**)
- Bayesian inference (Wednesday)
- Begin graph theory (Friday)

Today:

- Examples
- Explanation

The law's stubborn indifference toward units of measure gives another hint as to why the pattern is so common in the natural world. River lengths follow Benford's law whether we record them in meters or miles, whereas non-Benford-compliant data such as adult heights would radically change their distribution of leading digits when converted to meters because nobody is four meters tall. Remarkably, Benford's law is the only leading digit distribution that is immune to such unit changes.

We can think of changing units as multiplying every value in our data set by a certain number. For example, we would multiply a set of lengths by 1,609.34 to convert them from miles to meters. Benford's law is actually resilient to a much more general transformation. Taking Benford-compliant data and multiplying each value by a different number (rather than a fixed one such as 1,609.34) will leave the leading digit distribution unperturbed. This means that if a natural phenomenon arises from the product of several independent sources, then only one of those sources must accord with Benford's law for the overall result to. Benford's law is cannibalistic, much in the same way that a single zero in a bunch of numbers being multiplied together makes the result zero.

Jack Murtagh, "What is Benford's Law?", *Scientific American*, May 8, 2023

What do 1990 census statistics have in common with 1880 users of logarithm tables, numerical data from the front pages of newspapers of the 1930s collected by Benford or computer calculations observed by Knuth in the 1960s? Furthermore, why should they be logarithmic or, equivalently, base-invariant?

As already noted, many tables are not of this form, including even Benford's individuals tables, but as University of Rochester mathematician Ralph Raimi pointed out, "what came closest of all, however, was the union of all his tables." Combine molecular-weight tables with baseball statistics and the areas of rivers, and then there is a good fit with Benford's law.

Instead of thinking of some universal table of all possible constants, what seems more natural is to think of data as coming from many different distributions, as in Benford's study, in collecting numerical data from newspapers or in listing stock prices. Using this idea, modern mathematical probability theory, and the recent scale- and base-invariance proofs, it is not difficult to derive the following new statistical form of the significant-digit law (Hill 1996). If distributions are selected at random (in any "unbiased" way) and random samples are taken from each of these distributions, then the significant-digit frequencies of the combined sample will converge to Benford's distribution, even though the individual distributions selected may not closely follow the law.

T P Hill, "The First Digit Phenomenon," *The American Scientist*, July 1998

For example, suppose you are collecting data from a newspaper, and the first article concerns lottery numbers (which are generally uniformly distributed), the second article concerns a particular population with a standard bell-curve distribution and the third is an update of the latest calculations of atomic weights. None of these distributions has significant-digit frequencies close to Benford's law, but their average does, and sampling randomly from all three will yield digital frequencies close to Benford's law.

One of the points of the new random samples from random distributions theorem is that there are many natural sampling procedures that lead to the same log distribution, whereas previous arguments were based on the assumption that the tables following the log law were all representative of the same mystical underlying set of all constants. Thus the random-sample theorem helps explain how the logarithm-table digital frequencies observed a century ago by Newcomb, and modern tax, census and stock data, all lead to the same log distribution. The new theorem also helps predict the appearance of the significant-digit phenomenon in many different empirical contexts (including your morning newspaper) and thus helps just some of the recent applications of Benford's law.

T P Hill, "The First Digit Phenomenon," *The American Scientist*, July 1998

For next time:

Find Benford (and non-Benford) data in the wild.