A NOTE ON DISTRIBUTIONS OF TRUE VERSUS FABRICATED DATA[1]

THEODORE P. HILL

Georgia Institute of Technology

Summary.—New empirical evidence and statistical derivations of Benford's Law have led to successful goodness-of-fit tests to detect fraud in accounting data. Several recent case studies support the hypothesis that fabricated data does not conform to expected true digital frequencies.

A recent article in the *Wall Street Journal* (Berton, 1995) reported that the District Attorney's office in Brooklyn, New York detected fraud in seven New York companies by using a statistical goodness-of-fit test to ascertain that a significant part of the companies' accounting data had been fabricated. The purpose of this note is to communicate empirical discoveries in accounting and theoretical advances in statistics which strongly suggest that the logarithmic distribution called Benford's Law is a valid *a priori* distribution for the expected digital frequencies of many true data sets, and to communicate case studies in accounting which support the hypothesis that fabricated data do not closely follow this law.

Benford's Law is an empirical statistical law which states that in many tables of numerical data, the significant digits are not uniformly distributed as might be expected but rather obey a certain logarithmic probability distribution (recall that, for example, the first significant digit of 0.0501 is 5, the second is 0, and so on). Specifically, Benford's Law is the probability distribution on significant digits which states that, in particular,

(1) $\text{Prob(first significant digit} = d) = \log_{10}\left(1+\frac{1}{d}\right)$ for $d = 1, 2, \ldots, 9$, and

(2) $\text{Prob(second significant digit} = d) = \sum_{k=1}^{9} \log_{10}\left(1+\frac{1}{10k+d}\right)$ for $d = 0, 1, 2, \ldots, 9$.

[1] Research partially supported by NSF Grant DMS-95-03375. Request reprints of T. P. Hill, School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA.

For example, (1) says that the first significant digit is 1 with probability $\log_{10}\left(1+\frac{1}{1}\right) = \log_{10}(2) \cong 0.301$, and is 9 with probability $\log_{10}\left(1+\frac{1}{9}\right) \cong 0.046$. Similarly, (2) says that the probability the second significant digit is 3 is $\sum_{k=1}^{9}\log_{10}\left(1+\frac{1}{10k+3}\right) = \log_{10}\left(1+\frac{1}{13}\right)+\log_{10}\left(1+\frac{1}{23}\right)+\cdots+\log_{10}\left(1+\frac{1}{93}\right) \cong 0.17$. It is easy to check that the probabilities in (1) and (2) are all decreasing in d, and sum to (1).

Benford's Law also specifies distributions of third and higher significant digits, and even specifies the *joint* distributions of these significant digits, e.g., the probability that the first two significant digits are 5 and 0 respectively, which is not simply the product of the probability the first significant digit is 5 times the probability the second significant digit is 0 – the significant digits are dependent [cf. Hill, 1995a for the exact formulas]. Benford's Law is the only probability distribution on significant digits which is invariant under changes of scale (e.g., converting from metric to English units), or under changes of base (e.g., replacing base 10 by base 8 or 2, in which case the logarithm base 10 is replaced by logarithm to the new base).

Empirical evidence of Benford's Law has appeared in a wide variety of contexts: tables of physical constants, newspaper articles and almanacs, and numerical computations in computing [cf. Newcomb, 1881; Benford, 1938; Raimi, 1969; Hill, 1996]; certain aspects of cognitive arithmetic (Ashcraft, 1992; Dehaene and Mehler, 1992); and many areas of accounting including tax, stock market, and demographic data (Nigrini, 1995).

These empirical discoveries are supported by new mathematical laws of probability (Hill, 1995a, 1996) which both explain and predict the appearance of the logarithmic distribution. Roughly speaking, this new statistical principle says that, if probability distributions are selected at random and random samples are then taken from each of these distributions in any way so that the over-all process is "unbiased," then the leading significant digits of the combined sample will always converge to Benford's Law. This theorem helps explain why data sets such as numbers from front pages of newspapers, large accounting tables, or stock market figures tend to obey Benford's Law since they are composed of samples from many different distributions.

This prevalence of the logarithmic distribution in true accounting data sets has led to its recent use to detect fraud, under the hypothesis that when people fabricate data they do not choose numbers which follow the logarithmic distribution. It is well documented that people cannot behave truly randomly even when it is to their advantage to do so (Chapanis 1953; Bakan, 1960; Neuringer, 1986), and recent case studies support the hypothesis that concocted data do not follow Benford's Law closely. Nigrini (1994a) analyzed distributions of numbers from 873 fraudulent checks in an embezzlement scheme and described three other case studies in accounting involving falsified data, and in another study (Nigrini, 1994b) investigated tax-fraud digital distributions. Even when people invent numbers without a goal such as fraud in mind, the digital frequencies do not conform well to Benford's Law (Hill, 1988). Many of these case studies suggest an overabundance of leading digits in the mid-ranges 4–6 in fabricated data, but comprehensive experimental verification and a general theory for the distribution of fabricated data are still missing.

REFERENCES

Ashcraft, M. (1992) Cognitive arithmetic: a review of data and theory. *Cognition*, 44, 75–106.

Bakan, P. (1960) Response-tendencies in attempts to generate binary series. *American Journal of Psychology*, 73, 127–131.

Benford, F. (1938) The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78, 551–572.

Berton, L. (1995) He's got their number: scholar uses math to foil financial fraud. *Wall Street Journal*, July 10, 1995.

Chapanis, A. (1953) Random-number guessing behavior. *American Psychologist*, 8, 332.

Dehaene, S. and Mehler, J. (1992) Cross-linguistic regularities in the frequency of number words. *Cognition*, 43, 1–29.

Hill, T. (1988) Random-number guessing and the first-digit phenomenon. *Psychological Reports*, 62, 967–971.

Hill, T. (1995a) Base-invariance implies Benford's Law. *Proceedings of the American Mathematical Society*, 123, 887–895.

Hill, T. (1995b) The significant-digit phenomenon. *American Mathematical Monthly*, 102, 322–327.

Hill, T. (1996) A statistical derivation of the significant-digit law. To appear in *Statistical Science*.

Neuringer, A. (1986) Can people behave "randomly"?: the role of feedback. *Journal of Experimental Psychology* (General), 115, 62–75.

Newcomb, S. (1881) Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4, 39–40.

Nigrini, M. (1994a) Using digital frequencies to detect fraud. Preprint, Department of Accounting, St. Mary's University.

Nigrini, M. (1994b) Evidence of taxpayer manipulation from unaudited tax data. Preprint, Department of Accounting, St. Mary's University.

Nigrini, M. (1995) A taxpayer compliance application of Benford's Law. To appear in *Journal of the American Taxation Association.*

Raimi, R. (1969) The peculiar distribution of first significant digits. *Scientific American*, 221(6), 109–120.