The Difficulty of Faking Data

Last years several attempts to falsify the scientific data appeared in the publications. The fight against fakery is a hard one and the journals and the organizations that monitor the science made great efforts to reveal the misconducting in science and punish the guilty persons.
A century-old observation concerning the distribution of significant digits is now being used to help detect fraud.

Most people have preconceived notions of randomness which often differ substantially from true randomness. A classroom favorite is the counter-intuitive fact that in a randomly selected group of 23 people, the probability is bigger than 50% that at least two share the same birthday. A more serious example concerning false-positives in medical testing is this: Suppose that a person is selected at random from a large population of which 1% are drug users, and that a drug test is administered which is 98% reliable (i.e., drug users test positive with probability 0.98, and non-users test negative with probability 0.98). The somewhat surprising fact is that if the test result is positive, then the person tested is nevertheless more than twice as likely to be a non-user than a user. Similar surprises concerning unexpected properties of truly random data sets make it difficult to fabricate numerical data successfully.

Misperceptions of Randomness

To demonstrate this to beginning students of probability, I often ask them to do the following homework assignment the first day. They are either to flip a coin 200 times and record the results, or merely pretend to flip a coin and fake the results. The next day I amaze them by glancing at each student's list and correctly separating nearly all the true from the faked data. The fact in this case is that in a truly random sequence of 200 tosses it is extremely likely that a run of six heads or six tails will occur (the exact probability is somewhat complicated to calculate), but the average person trying to fake such a sequence will rarely include runs of that length.

This is but one example of the well-documented observation that most people cannot generate truly random numerical data. A study published in 1953 by psychologist A. Chapanis describes his experiment in which subjects were asked to write out long sequences of 1 numbers (digits 0 through 9) in random order. His results showed that different individuals exhibit marked preferences for certain decimal digits, and that repetitive pairs or triplets such as 222, 333 are avoided, whereas preferred triplets usually are made up of digits all of which are different, e.g., 653 or 231. This tendency to avoid long runs and include too many alternations, as in my class demonstration, has been confirmed by many researchers.

Most recently it has played a role in the arguments of cognitive psychologists Gilovich, Vallone, and Tversky that the “hot hand" in basketball is nothing more than a popular misperception, since long streaks in truly random data are much more likely to occur than is commonly believed.

True versus Fabricated Data

Determining whether real numerical data has been fabricated or altered is often of great importance: verifying experimental scientific data, such as medical trials, upon which crucial decisions depend; census data which helps determine political boundaries and governmental subsidies; and tax-return data submitted to the IRS by individuals and corporations.

The varied techniques used in detection of fraud or fabrication include both deterministic and statistical methods.

One example of a deterministic method is analysis of round-off approximations. In an article on rounding percentages in 1979 in the Journal of the American Statistical Association, statisticians P. Diaconis and D. Freedman's analysis of numerical data in a well-known paper raises the suspicion that [the author] manipulated the data to make the rows round properly. This suspicion is not hard to verify. "The percentage of numbers with leading digit 7 is reported as 5.5, with a total of 335 cases. The only proportions compatible with 5.5 are 18/335, which rounds to 5.4, or 19/335, which rounds to 5.7. There is no proportion possible that rounds to 5.5."

The remainder of this article will focus on statistical methods for detecting fake data, and the general idea behind such tests is quite simple: Identify properties of numerical data sets (of particular types) which are:
(i) highly likely to occur in true data sets of that type
(ii) highly unlikely to occur in fabricated data sets of that type.

The example above of using the pattern “runs of six or longer" to detect faked data in strings of 200 coin tosses is exactly such a test, and of course many other similar tests are available. One of the newest currently being used is a century-old observation called Benford's law, or the significant-digit law.

Benford's Law

The significant-digit law is the empirical observation that in many naturally occurring tables of numerical data, the leading significant (non-zero) digit is not uniformly distributed in {1, 2… 9} as might be expected, but instead obeys the law

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

Thus, this law (apparently first discovered by astronomer/mathematician S. Newcomb in 1881) predicts that a number chosen at random has leading significant digit 1 with probability $\log_{10} 2 \approx$ 0:301, leading significant digit 2 with probability $\log_{10}(3/2) \approx$ 0:176, and so on monotonically down to probability 0.046 for leading digit 9. The corresponding laws for second and higher significant digits, and their joint distributions is

$$\text{Prob}(D_1 = d_1, \ldots, D_k = d_k) = \log_{10}\left[1+\left(\sum_{i=1}^{k} d_i \times 10^{k-i}\right)^{-1}\right]$$

for $d_1 \in \{1, 2, \ldots, 9\}$ and $d_j \in \{0, 1, 2, \ldots, 9\}$, $j > 1$.

This says for example, that the probability that the first three significant digits of a number are 3, 1, 4 respectively,

$$P((D_1, D_2, D_3) = (3, 1, 4)), \text{ is equal to } \log_{10}\left(1 + \tfrac{1}{314}\right) \simeq 0.0014.$$

This logarithmic distribution is the only distribution on the significant digits of real numbers which is invariant under changes of scale. That is, if you calculate the probabilities of particular leading significant digits (such as $P((D_1;D_2;D_3) = (3;1; 4))$, then these logarithmic probabilities remain unchanged when the underlying data set is multiplied by 2 or by π, or under any other change of scale (e.g., from English to metric units), and they are the only probabilities with that invariance property. For example, if the distribution of the significant digits of a particular data set such as stock prices is (close to) the Benford distribution, then conversion from dollars per stock to pesos per stock will preserve the frequencies of the significant digits, whereas all non-Benford distributions will not.

Clearly the naive guess that the leading digits are equally likely to be one of the numbers {1, 2, … 9} does not exhibit scale invariance, since multiplication by 2, for example, converts all numbers starting with 5, 6, 7, 8, or 9 into numbers starting with 1. This implies that $P(D_1 = 1)$ must equal $P(D_1 = 5)+P(D_1 = 6)+P(D_1 = 7)+P(D_1 = 8)+ P(D_1 = 9)$ for scale-invariance under multiplication by 2 to hold, which is certainly not true if $P(D_1 = k)$ is the same for all k. (The proof that the logarithmic distribution 4 is the only scale-invariant distribution on the significant digits is based on the fact that the orbit of every point under irrational rotation on the circle is asymptotically uniformly distributed.) The logarithmic distribution is also the only probability distribution which is invariant under change of base, e.g., if the underlying data set is converted from base 10 to base 100 or vice versa. The formal statement and proof of this fact is somewhat deeper. These scale- and base-invariance characterizations of the logarithmic distribution, however clean mathematically, do not explain the widespread appearance of the distribution in real data, since that simply replaces the question of “why logarithmic?" to “why scale-invariant?". In trying to understand the prevalence of the logarithmic distribution in many real data sets, I noticed that tables which most closely fit the log distribution are composite samples from various distributions.

Using the scale- and base-invariance ideas together with modern probability tools such as constructions of random measures, it was not difficult to show that if random samples are taken from random distributions (in a “neutral" way), then the frequencies of the leading significant digits of the combined sample will always converge to Benford's law. One possible intuitive explanation is this: If a single distribution is picked at random, then it is certain (with probability one) to be scale-dependent, but sampling from different distributions and combining the data tends to neutralize the dependence on the scales, hence leading to the only scale-invariant distribution, Benford's law.

Empirical Evidence of Benford's Law

In 1881, Newcomb explained that his discovery of the significant-digit law was motivated by an observation that the pages of a book of logarithms were dirtiest in the beginning and progressively cleaner throughout. In 1938 General Electric physicist F. Benford rediscovered the law based on this same observation, and went on to spend several years collecting data from sources as different as atomic weights, baseball statistics, numerical data from Reader's Digest, and areas of rivers. Newcomb's article having been long forgotten, Benford's name came to be associated with the significant-digit law. Since then Benford's Law has been found to be a very good fit to such varied sets as stock market data (Dow Jones, Standard and Poor), 1990 census populations of the 3141 countries in 5 the United States, and numbers appearing in newspapers.

Thus there is evidence that many classes of true data sets follow Benford's Law, and in many of those classes such as stock market tables, census data and numbers gleaned from newspaper articles, a plausible theoretical explanation for the appearance of the logarithmic distribution is the random-samples-from-random-distributions theorem.

Detection of Fraud Using Benford's Law

Another class of data sets which has recently been found to be a good fit to Benford's law is true tax data. According to accounting Professor M. Nigrini's 1996 article in the Journal of the American Taxation Association, the IRS's own model files for the line items “Interest Paid" and “Interest Received" indicate that the significant digits for these items are an exceedingly close fit to Benford in true tax data (Figure 4). Nigrini has substantial evidence that in most fabricated tax data, however, the significant digits are not close to Benford, and his article describes a goodness-of-fit-to-Benford test to help identify fraudulent financial data. This test is a partial negative test, in that conformity does not necessarily imply true data, but nonconformity indicates some level of suspicion.

The Wall Street Journal (July 10, 1995) reported that the chief financial investigator for the district attorney's office in Brooklyn, N.Y., Mr. R. Burton, used [Nigrini's] program to analyze 784 checks issued by seven companies and found that check amounts on 103 checks didn't conform to expected patterns. `Bingo, that means fraud,' says Mr. Burton. The district attorney has since caught the culprits, some bookkeepers and payroll clerks, and is charging them with theft.

Since then, according to an article in the New York Times (August 4, 1998), “The income tax agencies of several nations and several states, including California, are using detection software based on Benford's Law, as are a score of large companies and accounting businesses".

With the current exponentially increasing availability of digital data and computing power, the trend toward use of subtle and powerful statistical tests for detection of fraud and other fabricated data is also certain to increase dramatically. Benford is only the beginning.

References

Benford, F. (1938), “The Law of Anomalous Numbers," Proceedings of the American Philosophical Society 78, 551-572.

Chapanis, A. (1953), “Random-number Guessing Behavior," American Psychologist

8, 332.

Chernof, H. (1981), “How to Beat the Massachusetts Numbers Game", Mathematical Intelligencer 3, 166-172.

Gilovich, T., Vallone, R., and Tversky, A. (1985), \The Hot Hand in Basketball: On the Misperception of Random Sequences," Cognitive Psychology 17, 295{314.

Hill, T. (1996), “A Statistical Derivation of the Significant-Digit Law," Statistical Science 10, 354{363.

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. (1996), “A Taxpayer Compliance Application of Benford's Law", Journal of the American Taxation Association 18, 72-91.

Biographical Sketch

Ted Hill is professor of mathematics at the Georgia Institute of Technology. His education has included a bachelor's degree from West Point, a master's in operations research from Stanford, a Fulbright Scholarship at Goettingen, and a Ph.D. in mathematics from the University of California at Berkeley. Since coming to Georgia Tech, he has also been a visiting professor at the University of Leiden, the University of Tel-Aviv, the Free University of Amsterdam, the University of Costa Rica, the University of Goettingen, and the Mexican Mathematics Research Center (CIMAT). His research interests include probability and measure theory, optimal-stopping theory, fair-division problems, and limit laws.

after *Theodore P. Hill*