IDENTIFYING FALSIFIED CLINICAL DATA

Joanne Lee, George Judge*

December 19, 2008

Abstract

Clinical data serve as a necessary basis for medical decisions. Consequently, the importance of methods that help officials quickly identify human tampering of data cannot be underestimated. In this paper, we suggest Benford's Law as a basis for objectively identifying the presence of experimenter distortions in the outcome of clinical research data. We test this tool on a clinical data set that contains falsified data and discuss the implications of using this and information-theoretic methods as a basis for identifying data manipulation and fraud.

*Joanne Lee is a doctoral student in the Graduate School at University of California, 207 Giannini Hall, Berkeley, CA 94720. Email: jlee@are.berkeley.edu. George Judge is Professor in the Graduate School at University of California, 207 Giannini Hall, Berkeley, CA 94720. Email: judge@are.berkeley.edu. The authors gratefully acknowledge the helpful comments of Wendy Tam Cho, Joyce McCann, and Sofia Berto Villas-Boas, as well as the help of John Dahlberg in procuring falsified clinical data.

1 Introduction

Clinical data serve as a necessary basis for choices relative to medical decisions that reduce health risks. The changing health recommendations that fill our newspapers and television screens reminds us how fragile these data are. In addition to the usual noise in experimental data, researchers have been known to massage the data to achieve a particular result or to reach a certain statistical significance level. In this context, a recent article in the New York Times Magazine discusses a well-known case of the intentional falsification of data from clinical experiments (Interlandi, 2006). Given the importance of clinical data integrity and the possibility of this type of fraudulent behavior, we focus on and demonstrate in this paper an objective method based on the distribution of first significant digits [FSD] that may help identify this type of experimenter data tampering.

The organization of the paper is as follows: In Section 2, Benford's FSD Law is reviewed. In Section 3, actual data from a clinical research problem is presented and the Benford FSD method is used to evaluate the possibility of falsification in this particular sample of clinical research data. In Section 4, information-theoretic methods that make use of first moment (mean) information is suggested as a way to estimate the corresponding sample FSD distribution and to evaluate data integrity. The implications of using these methods to test clinical data integrity are discussed in Section 5.

2 Benford's Law

In 1881, astronomer and mathematician Simon Newcomb noticed that the first several pages of logarithm tables were more worn than subsequent pages. This observation led him to the counter-intuitive conjecture that, in a "natural" dataset, "1" would occur most frequently and "9" would occur least frequently (Newcomb, 1881). Newcomb stated, "The law of probability of the occurence of numbers is such that all mantissae of their logarithms are equally probable,", suggesting the following expression for the empirical distribution of the first significant digits:

$$F_{FSD} = \log_{10}\left(1 + d^{-1}\right), \;\; \mathrm{d} = 1, 2, \ldots, 9 \tag{1}$$

where F_{FSD} is the frequency of the digit $d = 1, 2, \ldots, 9$. The resulting monotonic decreasing Benford FSD distribution is presented as the solid line in Figure 1 and the first row of Table 1.

Newcomb did not offer a theoretical explanation or an empirical verification of the phenomenon, and his conjecture did not garner much immediate attention. Fifty-seven years later, Frank Benford tested Newcomb's hypothesis empirically by demonstrating that a large number of seemingly unrelated natural sets of numbers provided a good fit to the distribution first laid out by Newcomb. This FSD phenomenon was named "Benford's Law" after its popularizer rather than its discoverer.

FIGURE 1

Table 1: The FSD frequencies for the Benford, $Insulin\_T$, and $Insulin\_F$, where $Insulin_T$ and $Insulin_F$ denote the true and false insulin data and corresponding correlations and $\chi^2_{(8)}$, respectively.

Dist.	Mean	Corr	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$	$\hat{p}_7$	$\hat{p}_8$	$\hat{p}_9$	$\chi^2_{(8)}$ [a]
Benford	3.44	1.00	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0.00
$Insulin\_T$	3.54	0.95	0.317	0.148	0.099	0.085	0.134	0.042	0.078	0.063	0.035	4.84
$Insulin\_F$	4.03	0.66	0.255	0.073	0.102	0.080	0.197	0.139	0.066	0.0584	0.029	17.81

[a] The 10%, 5%, 1% critical values for $\chi^2_{(8)}$ are 13.36, 15.5, and 20.09, respectively.

Aside from a geometrical interpretation, however, Benford provided no formal justification for these monotonically decreasing FSD distributions. Hill (1995) was the first to rigorously prove Benford's Law using a base-invariance argument. Noticing that the arguments for Benford's Law carry over *mutatis mutandis* to non-decimal bases, Hill proved that there exists a unique base-invariant probability measure on the positive real numbers that satisfies Benford's Law. Since base invariance is the only assumption, and this logic also translates to scale invariance, Hill's proof is now widely accepted as a proof of Benford's Law. All the same, the empirical fit of Benford's Law to large datasets is approximate and alternative distributions exist. For example, Pietronero et al. (2001) suggest that Benford's Law is a special case of Zipf's Law, which claims that power laws are present in all rankings of natural processes by size, i.e., the probability of occurrence is inversely proportional to its rank. Grendar et al. (2006), using information-theoretic methods, propose a generalized Benford's Law and formulation that we use later in this paper.

Building on Benford's monotonic decreasing FSD distribution for naturally occurring multiplicative data sets, our objective is to examine a scale-invariant clinical data set that may be expected to obey Benford's Law and evaluate the nature and basis of departures. In the next section, we describe and analyze a set of clinical data that contains both correct and falsified clinical information. Some others who have used Benford's law to check the validity of purported scientific data in the social sciences include Varian (1972), Carslaw (1988), Nigrini (1996), Durtschi et al. (2004), Geyer and Williamson (2004), Giles (2007), Nye and Moul (2007), and Judge and Schechter (2008).

3 Falsified Clinical Data

The example of falsified clinical data we use was obtained through the generous help of John Dahlberg, who heads the Data Integrity group at Health and Human Services[1]. This clinical research data was falsified by Eric Poehlman, a faculty member at the University of Vermont, who pleaded guilty to fabricating more than a decade of data that was financed by federal grants from the National Institutes of Health. Poehlman hoped to demonstrate, with a clinical study following patients over time, that lipid levels deteriorate with age. After a graduate student found that the data did not support the hypothesis, Poehlman tampered with the data until his hypothesis was evident in the data.

The particular data we analyze relates to insulin levels, which are expected to decline with age. The correct and falsified Poehlman data provided 135-145 observations on changes in insulin levels over a six-year period.

3.1 The Correct Data

The first question we pose is whether or not the correct Poehlman data, with 142 observations, has a FSD distribution that is similar to that proposed by Benford. The distributions of Benford and the correct insulin FSD's are given in Figure 1 and Table 1. Note that the correlation between the Benford and the correct empirical FSD distributions is 0.95. From an inference viewpoint, the corresponding χ^2 with 8 degrees of freedom is 4.84 and thus less than the 25% statistical significance level of 10.22 and indicates that we cannot reject the null hypothesis of distribution equality between the Benford and empirical FSD distributions.

3.2 The Falsified Data

The FSD distribution for the 136 observations of falsified data is given in Table 1 and Figure 1. Note in Figure 1 the extreme fluctuation of the FSD proportions of the falsified data and their departure from Benford's distribution. Note also the correlation between Benford and the falsified data frequencies is now only

[1]For background on this dataset see Interlandi (2006)

0.66. Also, the χ^2 test value has risen to 17.81, which exceeds the 5% level χ^2 value of 15.51 and thus provides a statistical basis for rejecting the null hypothesis of equality between the Benford and empirical FSD distributions. Thus, there is a visual and inferential basis for suspecting these research data have been manipulated.

4 Problem Reformulation and Solution

It seems reasonable that the FSD distribution could vary with the measured phenomenon in question. In this section, we use the empirical likelihood method and first moment frequency information to estimate the monotonically decreasing nature of the FSD's[2]. In this context, suppose that one of nine digits, $i = 1, \ldots, 9$, is observed with probability p_i, and the information is given in the form of the average value of the FSD's:

$$\sum_{i=1}^{9} d_i p_i = \bar{d} \tag{2}$$

Based on this first moment information and Owen's (2001) empirical likelihood metrix $9^{-1}\sum_{i=1}^{9}\ln p_i$, we can formulate the problem of recovering the unknown and unobservable p_i as the extremum likelihood function

$$\max_{\mathbf{p}} \left\{ 9^{-1}\sum_{i=1}^{9}\ln p_i | \sum_{i=1}^{9} p_i d_i = \bar{d}, \sum_{i=1}^{9} p_i = 1 \right\} \tag{3}$$

The corresponding Lagrange function is

$$L(\mathbf{p}, \eta, \lambda) = 9^{-1}\sum_{i=1}^{9}\ln p_i - \eta\left(\sum_{i=1}^{9} p_i - 1\right) - \lambda\left(\sum_{i=1}^{9} p_i d_i - \bar{d}\right) \tag{4}$$

where $\mathbf{p}>0$ is implicit in the structure of the problem. Solving the corresponding first order condition with respect to p_i leads to the solution:

$$p_i^*(\bar{d}, \lambda) = 9^{-1}\left(1 + \lambda^*[d_i - d]\right)^{-1} \tag{5}$$

This solution implies that an exponential family of distributions will result as the mean of the FSD's varies over a range impled by actual datasets. Maximizing $\Pi_{i=1}^{9} p_i$ subject to $\sum_{i=1}^{9} p_i = 1$ and $\sum_{i=1}^{9} p_i d_i = \bar{d}$, the p_i are chosen to assign the maximum joint probability among all the possible probability assignments.

As an example, some mean-related empirical likelihood FSD distributions are given for selected FSD means in Table 2. Note that the FSD mean for the Benford distribution is 3.44 and, in this case, the empirical likelihood FSD distribution is almost identical to the Benford distribution. Since, from our experience,

[2]See Owen (2001) for an introduction to empirical likelihiood methods

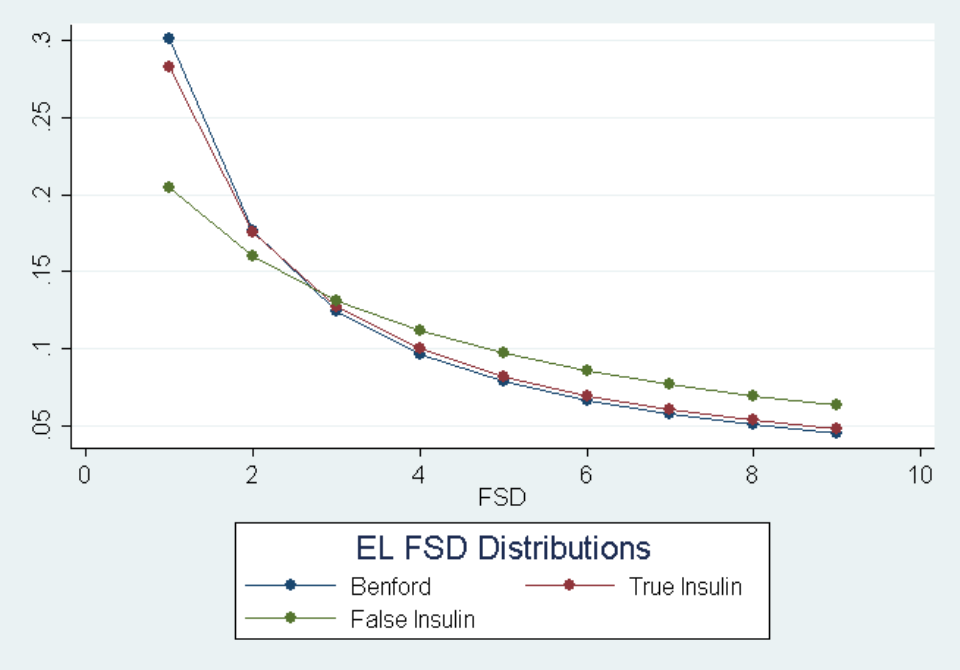
clinical researchers are reluctant to share their data, perhaps they will be at least willing to take a moment to share the mean of their digit frequencies. With this information, empirical likelihood methods offer a viable data evaluation technique.

Table 2: Estimated empirical likelihood distributions for the FSD problem and their correlation with Benford's distribution.

FSD mean	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$	$\hat{p}_7$	$\hat{p}_8$	$\hat{p}_9$	corr
2.00	0.673	0.111	0.061	0.042	0.032	0.026	0.021	0.018	0.016	0.925
3.00	0.395	0.173	0.111	0.082	0.065	0.053	0.046	0.040	0.035	0.990
3.44	0.300	0.177	0.125	0.097	0.079	0.067	0.058	0.051	0.046	1.000
4.00	0.208	0.161	0.132	0.111	0.096	0.085	0.076	0.068	0.062	0.980
4.50	0.151	0.137	0.125	0.115	0.107	0.100	0.093	0.088	0.083	0.932

As shown in Table 1, the FSD means of the true and false insulin data are 3.54 and 4.03, respectively. The empirical likelihood [EL] distributions generated by the two means are shown in Figure 2 and Table 3.

FIGURE 2



The EL FSD distribution for the true insulin data with mean 3.54 has a strong visual correlation with the Benford reference distribution. Alternatively, the EL FSD distribution for the false insulin data appears visually distinguishable from the EL true insulin data and Benford reference distributions. The $\chi^2_{(8)}$ values shown in Table 3 further suggest that we cannot reject the null hypothesis of distribution equality of the EL FSD distribution with mean 3.54 and the Benford reference distribution.

Table 3: Estimated empirical likelihood distributions for $Insulin\_T$ and $Insulin\_F$ and corresponding Benford relative correlations and $\chi^2_{(8)}$, where $Insulin_T$ and $Insulin_F$ denote the true and false insulin data, respectively.

Dist.	Mean	Corr	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$	$\hat{p}_7$	$\hat{p}_8$	$\hat{p}_9$	$\chi^2_{(8)}$ [a]
Benford	3.44	0.947	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.049	0.00
$Insulin\_T$	3.54	1.000	0.282	0.176	0.127	0.100	0.082	0.070	0.061	0.054	0.048	0.39[a]
$Insulin\_F$	4.03	0.977	0.204	0.160	0.132	0.112	0.097	0.086	0.077	0.070	0.064	6.40

[a]The 10%, 5%, 1% critical values for χ^2 with 8 degrees of freedom are 13.36, 15.5, and 20.09, respectively.

Relative to the Benford distribution, the corresponding $\chi^2_{(8)}$ value for the EL true insulin FSD distribution yields a better goodness-of-fit than the EL false insulin FSD distribution. However, the $\chi^2_{(8)}$ do not lead us to reject equality of any of the three FSD distributions. Given the monotonically decreasing nature of the three distributions, the Kuiper modified Kologorov-Smirnov goodness-of-fit test may be more appropriate since it recognizes the ordinality and circularity of the data (Giles, 2007). In any case, from a subjective standpoint, the difference between the Benford and the EL false FSD distribution should get the attention of those investigating clinical data integrity.

5 Summary and Implications

In practice, Health and Human Services or any other granting agency would only have available the falsified data from the researcher. However, in this case, the departures of the false insulin FSD distribution from the Benford FSD distribution are great enough to have likely attracted greater scrutiny of the clinical data. Although scientific fraud is hopefully a rare phenomenon, using Benford's Law appears to be a quick and objective way to check clinical data coming from researchers when the data spans the nine-digit space. In other cases, when the range of the data FSD's is restricted, the significant second digit distribution may be used and compared to the Benford's second-digit distribution:

$$F_{SD} = \log_{10}\left((1 + (10k + d)^{-1})\right) \mathrm{d} = 1, 2, \ldots, 9 \tag{6}$$

Frequency analysis of the second and following digits results, in this case, in a monotonically decreasing distribution that is not greatly different from a uniform digit distribution. However, in general, in fabricating second and following digit data, it has been observed that researchers are poor random number generators. This comparison for the Poehlman second-digit data is given in Table 4, where the null hypothesis of distribution equality cannot be rejected.

Table 4: The SD frequencies and $\chi^2_{(9)}$ tests for the Benford SD, true insulin data, and false insulin data.

Dist.	Mean	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$	$\hat{p}_7$	$\hat{p}_8$	$\hat{p}_9$	$\chi^2_{(9)}$ [a]
Benford SD	0.120	0.114	0.109	0.104	0.097	0.093	0.090	0.088	0.085	0.100	0.100	0.00
$Insulin_T$	3.852	0.113	0.120	0.120	0.113	0.162	0.092	0.070	0.106	0.063	0.042	0.057
$Insulin_F$	3.853	0.154	0.132	0.088	0.074	0.170	0.066	0.074	0.118	0.074	0.052	0.071

[a] The 10%, 5%, 1% critical values for $\chi^2_{(9)}$ are 14.68, 16.92, and 21.67, respectively.

Finally, we have found it very difficult to obtain additional clinical data to analyze and regret that we can only report on one set of experimental data. Although this may be a naive request, given the importance of decisions based on clinical data, it would seem to be important to get more of these research data in the hands of statisticians or other qualified clinical researchers for analysis purposes. An objective clearing house that analyzes, via Benford and other methods, clinical data as they come from researchers would seem to be a useful first step.

References

Carslaw, C. (1988). Anomalies in income numbers: Evidence of goal-oriented behavior. *Accounting Review*.

Durtschi, C., W. Hillison, and C. Pacini (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting*.

Geyer, C. and P. Williamson (2004). Detecting fraud in datasets using Benford's law. *Computation in Statistics: Simulation and Computation*.

Giles, D. (2007). Benford's law and naturally occurring prices in certain ebay auctions. *Applied Economics Letters*.

Grendar, M., G. Judge, and L. Schechter (2006). An empirical non-parametric likelihood family of data-based Benford-like distributions. *Physica A: Statistical Mechanics and its Applications*.

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

Interlandi, J. (October 22, 2006). An unwelcome discovery. *New York Times Magazine*.

Judge, G. and L. Schechter (2008). Detecting problems in survey data using Benford's law. *Journal of Human Resources*.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics 4*(1/4), 39–40.

Nigrini, M. (1996). A taxpayer compliance application of Benford's law. *Journal of the American Taxation Association*.

Nye, J. and C. Moul (2007). The political economy of numbers: On the application of Benford's law to international macroeconomic statistics. *B.E. Journal of Macroeconomics*.

Owen, A. (2001). *Empirical Likelihood*. Chapman and Hall. Florida.

Pietronero, L., E. Tosatti, V. Tosatti, and A. Vespignani (2001, April). Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica A: Statistical Mechanics and its Applications 293*(1–2), 297–304.

Varian, H. (1972). Benford's law. *American Statistician*.