Detecting Possibly Fraudulent or Error-Prone Survey Data Using Benford's Law

David Swanson, Moon Jung Cho, John Eltinge
U.S. Bureau of Labor Statistics
2 Massachusetts Ave., NE, Room 3650, Washington, DC 20212
(swanson_d@bls.gov)

Key Words: Consumer Expenditure Interview Survey; Curbstoning; Digit preference; Pearson test statistic; Quantitative survey responses; Reinterview.

Any opinions expressed in this paper are those of the authors, and do not constitute policy of the Bureau of Labor Statistics.

1. Introduction

The Consumer Expenditure Survey (CE) is a nationwide household survey conducted by the U.S. Bureau of Labor Statistics (BLS) to find out how Americans spend their money. As with any survey, the accuracy of CE's published expenditure estimates depends on the accuracy of the collected data. The CE survey has several procedures already in place to ensure the accuracy of its published expenditure estimates. They include reinterviews of some respondents, computerized checks for the logical consistency of responses given by respondents, an outlier review of individual survey responses, and another outlier review of the summarized expenditure estimates before they are published (*BLS Handbook of Methods*).

In this paper we describe another method of identifying inaccurate survey data. The method is little-known, but it has been rapidly gaining popularity over the past decade. The method involves examining the distribution of the leading (or left-most) digits of all the numbers reported on a survey form. These leading digits have been observed to follow a certain distribution regardless of the nature of the survey. This phenomenon is called *Benford's Law*. By knowing the distribution of the leading digits, one can identify unusual data which may be fraudulent or generated by an error-prone process by identifying the interviews in which the distribution of leading digits does not follow the expected distribution.

In this paper we will describe the Consumer Expenditure Survey and the current methods used in that survey to identify inaccurate data. Then we will describe Benford's Law, describe some applications of it in other settings, and then we will give an example showing how Benford's Law can be used to identify unusual data in a survey setting using CE data as an example.

2. Background

The Consumer Expenditure Survey is a nationwide household survey conducted by the BLS to find out how Americans spend their money. Data for the survey are collected by the Bureau of the Census under contract with the BLS. One of the primary uses of the data is to provide expenditure weights for the Consumer Price Index. Data are collected by personal visits to the households in the survey's sample.

The Consumer Expenditure Survey consists of two separate surveys, the Diary (CED) and Quarterly Interview (CEQ) surveys. The purpose of the CED is to obtain detailed expenditure data on small, frequently purchased items such as food and apparel. The purpose of the CEQ is to obtain detailed expenditure data on large items such as property, automobiles, and major appliances; and on expenses that occur on a regular basis such as rent, utility bills, and insurance premiums. Approximately 3,500 households are visited each quarter of the year in the CED, and 15,000 households in the CEQ.

The CED uses a new sample of households each quarter of the year. Each household in the CED is asked to keep a record of all its expenditures made during a 2-week period. After participating in the survey for 2 weeks the household is dropped from the survey, and it is replaced by another household.

The CEQ is a panel rotation survey. Each panel is interviewed for five consecutive quarters, and then dropped from the survey. As one panel leaves the survey, a new panel is introduced. Approximately 20 percent of the addresses are new to the survey each quarter.

3. Current Methods of Identifying Problematic Data in the CE Survey

The CEQ and CED surveys currently have several methods of identifying incorrect data. The first method is a reinterview process in which a field representative's supervisor calls a small number of respondents who participated in the survey on the telephone to find out whether the respondent was actually visited by the field representative, and to verify the accuracy of a few of their responses. Some respondents are randomly selected, while others are selected because the supervisor is suspicious of the data's accuracy. The reinterview process is mainly intended to catch *curbstoners*, field representatives who make up the data without ever visiting or contacting the respondent.

After the reinterview process, all of the remaining methods of checking the data are intended to identify legitimate data that were incorrectly recorded or keyed. The methods include a computerized check for logical consistency of the responses, an outlier analysis for individual reported observations, and another outlier analysis on the summarized expenditure estimates before they are published.

An example of a logical consistency error is when a box is checked off indicating that no expenditures were made in a certain item category, but yet there is an expenditure reported anyway. Logical consistency errors are easy for a computer to find.

The outlier review process for individual reported expenditures involves identifying observations that are unusually large, and then investigating them to find out whether they are accurate or seem reasonable. Photocopies of the completed survey forms are stored on microfilm, and an examination of the survey forms sometimes reveals keying errors, such as a misplaced decimal point changing a reported expenditure from $2.99 to $299.00. CE's outlier analysis focuses on large expenditures rather than small expenditures because large outliers have a much larger impact on the final published expenditure estimates.

CE uses four methods of identifying outliers:

- The *largest gap* test. The mean expenditure is calculated for each dollar field within each item code. The expenditures above the mean are sorted in descending order, and the difference (or *gap*) between each expenditure and the one below it is calculated. The largest of these gaps is identified, and all expenditures above it are flagged for review.
- If the reported expenditure is the largest value within its area/item combination it is flagged for review.
- If the reported expenditure is greater than 25% of the total of all expenditures within its area/item combination (50% is used instead of 25% if the number of expenditures is below 10) it is flagged for review.
- If the reported expenditure is greater than 20 times the median reported expenditure within its area/item combination it is flagged for review.

Every observation flagged as an outlier by one or more of these tests is printed on an outlier review listing. To help reviewers focus on the more extreme outliers, scores are given to each outlier, with the score basically reflecting the number of tests that considered it to be an outlier.

4. Other Methods of Detecting Incorrect Data

Reinterviews and outlier reviews are the most common methods of identifying incorrect or falsified survey data, but other methods of detecting them have also been proposed. For example, Biemer and Stokes (1989) report that in 1982 the Census Bureau started collecting information on the interviewers it caught cheating in order to develop a profile of the people and situations in which cheating was found. One of the Census Bureau's findings was that most cheating occurred with new interviewers who worked for the Census Bureau for less than one year. Biemer and Stokes used this information to develop a model for improving the detection of interviewer cheating.

Another method is to compare the survey results obtained by different interviewers. Turner *et. al.* (2000) presented a case study in which falsified survey data were detected in an epidemiologic survey when one of the interviewers was observed to have an unusually high interview yield. Most interviewers were successful obtaining interviews from about 30% of the sampled households, while one interviewer had a success rate of 85%. A review of the interviewer's results along with numerous reinterviews showed that much of the data were falsified.

Further examinations of the data turned up more interviewers with falsified data. When their data were examined it was observed that not only were their response rates higher than normal, but the fabricated data were different as well. For example, interviewers whose data were believed to be accurate showed 50% of all households in the survey's sample having one "eligible adult," while interviewers whose data were believed to be fabricated showed almost 70% of the households having one eligible adult. As a result of their experience with this survey Turner *et. al*. advocate examining the incoming data on a daily

basis in order to catch clues of potential data falsification as soon as possible.

5. Benford's Law

Another method of identifying incorrect data that has received a lot of attention in recent years is called *Benford's Law*. The method is named for Frank Benford, an American physicist who published a paper in 1938 describing a curious property that large collections of "real world" numbers tend to have: the leading (or left-most) digit of the numbers is more likely to be small rather than large. Specifically, he found that the proportion of "real world" numbers whose leading digit is d=1,2,3,…,9 is approximately $\log_{10}\left(\frac{d+1}{d}\right)$. This phenomenon is called *Benford's Law*.

Hill (1995) published a paper with the first rigorous mathematical explanation of why the leading digits in many data sets follow Benford's Law. Hill offered several explanations. One of his explanations involved a type of central limit theorem in which several probability distributions are chosen at random from a large collection of probability distributions. Then several random variables are chosen from each of the selected distributions. Under these conditions Hill proved that the leading digits of the numbers follow Benford's Law. This can be written mathematically as:

$$P\{x = d\} = \log_{10}\left(\frac{d+1}{d}\right)$$

where x is the leading digit of a randomly-selected number.

For example, in the CE survey respondents report their expenditures on a large number of item categories, with each item category having a different distribution of expenditures. Then within each item category the respondents report several expenditures. Thus we have several different probability distributions, and several random variables are chosen from each distribution, so the conditions described by Hill are satisfied. As a consequence the leading digits of all the expenditures reported on the CE's survey forms should follow Benford's Law.

6. Applications of Benford's Law in Other Areas

Modern applications of Benford's Law began in 1992, when Mark Nigrini examined the distribution of leading digits he found in some sales and expense data for his doctoral thesis. The data he examined followed Benford's Law quite closely. Then after that initial success, Nigrini continued to use Benford's Law to examine other business and financial data. For example, he used it to examine the expense claims of a nationwide chain of motels, where he uncovered approximately one million dollars of fraudulent claims.

Then in 1996 Nigrini examined IRS tax return data and found that the leading digits of the line items "Interest Paid" and "Interest Received" followed Benford's Law. His tax return study was published in the *Journal of the American Taxation Association*. Next Nigrini examined the leading digits of the numbers contained in President Clinton's tax returns for the years 1977-1992. Nigrini found that the leading digits followed Benford's Law, so he concluded that President Clinton's tax returns were honest.

These and other studies conducted by Nigrini generated a lot of interest within the accounting industry, and today the accounting industry is the largest business sector using Benford's Law to detect fraudulent data. Nigrini's work also led to tax agencies in several countries around the world as well as several U.S. states, including California, using Benford's Law to detect fraudulent data on tax returns.

Finally, the scientific community is occasionally rocked by studies that turn out to contain falsified data, and Benford's Law is starting to be used there to detect such falsified data.

7. Leading Digit Patterns in CE Data

The table below shows expenditure data collected by the CEQ survey in the year 2000. The survey collected data on 734,684 expenditures. By looking at the table it can be seen that the leading digits of those expenditures follow Benford's Law quite closely. According to CEQ data, 30.5% of the leading digits were 1's, while Benford's Law predicted the percentage to be 30.1%. The percentage of leading digits equal to 2 was 19.3% in the CEQ data, while Benford's Law predicted the percentage to be 17.6%.

Table 1.
Comparison of CEQ Data with Benford's Law

Leading Digit (d)	Reported Expenditures Number	Percent	(SE)	Benford's Law $\log_{10}\left(\frac{d+1}{d}\right)\times 100\%$
1	223,776	30.5	(.063)	30.1
2	141,992	19.3	(.053)	17.6
3	90,589	12.3	(.045)	12.5
4	66,266	9.0	(.040)	9.7
5	76,473	10.4	(.044)	7.9
6	50,024	6.8	(.034)	6.7
7	35,019	4.8	(.029)	5.8
8	32,294	4.4	(.028)	5.1
9	18,251	2.5	(.021)	4.6
Total	734,684	100.0		100.0

Although the CEQ data follow Benford's Law quite closely for some digits, a detailed examination of the data reveals a slight excess of 2's and 5's, and a slight shortage of 9's in the data. In the CEQ data 19.3% of the leading digits were 2's, while Benford's Law predicted the percentage to be 17.6%. Likewise, 10.4% of CEQ's leading digits were 5's, while Benford's Law predicted it to be 7.9%. This slight excess of 2's and 5's is usually attributed to respondents rounding their expenditures to numbers such as $25 or $50, but it might also represent fraudulent (or *curbstoned*) data in which field representatives created data that tended to start with 2's and 5's. The low percentage of 9's is curious because their shortage cannot be attributed to rounding numbers either up or down. The CEQ data have fewer 8's than Benford's Law predicts, so the expenditures are probably not rounded down, and there are not enough 1's to account for rounding up (the 1's exceed Benford's prediction by 0.4 percentage points, but there is a 2.1 percentage point shortage of 9's).

The standard errors in Table 1 are equal to the square root of the average of 100 random-group variance estimates, where each random-group variance estimate is based on randomly partitioning the set of sample consumer units into 50 groups. The point estimates and variance estimates repeated in Table 1 are unweighted. Table 2 compares the unweighted estimates with the weighted estimates. The standard errors in the last column of Table 2 are obtained by the balanced repeated replication method with 44 replicate weights.

Table 2.
Comparison of Unweighted and weighted Ratio

Leading Digit (*d*)	Number	Percent (SE) unweighted		Percent (SE) weighted	
1	223,776	30.5	(.063)	30.4	(.072)
2	141,992	19.3	(.053)	19.3	(.081)
3	90,589	12.3	(.045)	12.3	(.057)
4	66,266	9.0	(.040)	9.0	(.047)
5	76,473	10.4	(.044)	10.5	(.056)
6	50,024	6.8	(.034)	6.8	(.046)
7	35,019	4.8	(.029)	4.8	(.035)
8	32,294	4.4	(.028)	4.4	(.041)
9	18,251	2.5	(.021)	2.5	(.033)
Total	734,684	100.0		100.0	

The ratio of these two standard errors, $\left(\frac{SE_{BRR}}{SE_{RG}}\right)^2$, can be viewed as a deff. The range of ratios of the deffs is from 1.32 to 2.35. Note that the calculation of these deffs are based on standard errors with more digits than the reported ones.

When the interclass correlation coefficient is 0, we can derive $deff = 1 + CV^2$ where CV is the coefficient of variation of weights. This derivation is from the formula which Kish proposed to determine the design effect in order to incorporate the effects due to both weighting and clustered selection. Gabler *et. al*. justified the formula.

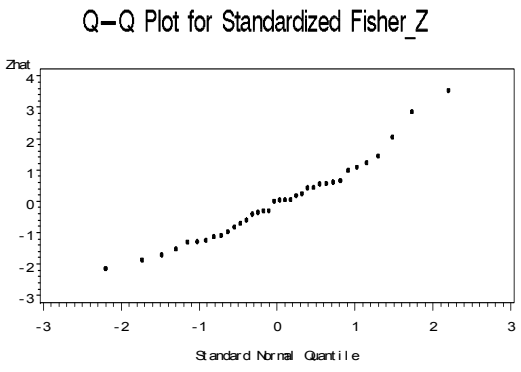
Figure 1 displays a quantile-quantile plot of Fisher Z, where the standardized Fisher Z is defined as follows:

$$\hat{Z}_{ij} = \sqrt{(n-3)}\left\{\tanh^{-1}(\hat{\rho}_{ij}) - \tanh^{-1}\left(-\sqrt{\frac{\hat{p}_i\hat{p}_j}{\hat{q}_i\hat{q}_j}}\right)\right\}$$

where ρ_{ij} is the correlation coefficient between the weighted proportions of leading digit *i* and *j* for $i \neq j$. We computed ρ_{ij} from a covariance obtained by the balanced repeated replication method with 44 replicate weights. Since we have 40 degrees of freedom in our example data, *n* equals to 41. Note that $-\sqrt{\frac{\hat{p}_i\hat{p}_j}{\hat{q}_i\hat{q}_j}}$ is the consistent estimator of correlation of $\hat{p}_i$ and $\hat{p}_j$ for $i \neq j$ under the multinomial model. Therefore this difference in the second term should converge to 0 if the multinomial model is satisfied. On the other hand, the fact that the absolute value of $\hat{Z}_{ij}$ is large means that that the absolute value of ρ_{ij} is large, i.e., $\hat{p}_i$ and $\hat{p}_j$ are highly correlated in our example data.

Figure 1 suggests that the principal deviation from normality is observed in the third though thirty-third values of being smaller than the associated quantile of the standard normal distribution. One should be cautious not to over-interpret this result because the values are not independent. However, this quantile-quantile plot is consistent with a mixture model in which some interviewers have a digit reporting profile that differs from those of the other interviewers. This suggests that further investigation of negative correlations associated with mixtures of multinomial distributions would be of interest.

Figure 1.



8. Identifying the Source of Unusual Data

Benford's Law can be used to help identify sources of unusual data. For example, suppose a field representative is suspected of curbstoning. Many studies (e.g., Browne, 1998) have shown that people tend to be bad at fabricating realistic data, so one way of identifying curbstoners is to see whether their data follow Benford's Law. If it does, then they are probably collecting accurate data. If it does not, then they may be fabricating at least some of the data.

If the data from each field representative is viewed as arising from a simple random sample, then Pearson's chi-square test statistic may be helpful in determining whether a field representative's collected data follow Benford's Law:

$$\theta = n\sum_{d=1}^{9}\frac{(\hat{p}_d - p_d)^2}{p_d}$$

where

n = the number of expenditures reported by a particular field representative,

$\hat{p}_d$ = the proportion of those expenditures whose leading digit is d, and

$p_d = \log_{10}\left(\frac{d+1}{d}\right)$.

This statistic is a goodness-of-fit measure that has a chi-square distribution with 9−1=8 degrees of freedom.

An alternative test statistic is the same formula, but where p_d is the proportion of all numbers collected in the survey whose leading digit is d. This alternative definition of p_d is computed from the complete universe of data collected from all field representatives. It takes into consideration the fact that Benford's Law may not hold exactly for a particular data set. It also assumes that the vast majority of field representatives are honest, so that the estimated value of p_d using the complete universe of collected data from all field representatives is close to the true value of p_d. This is sometime called a *digital analysis*.

Table 3 shows an example of CEQ data from a typical field representative (θ = 10.39) and from an unusual field representative (θ = 102.43) using CE's complete set of collected data to estimate p_d:

The data in Table 3 show that the unusual field representative has a large number of 5's and 6's. When 1,132 expenditures are reported, the percentage of leading digits equal to 5 should be approximately 10.4% ± 1.8%, but 17.2% of that field representative's leading digits are 5's. Likewise, the percentage of leading digits equal to 6 should be approximately 6.8% ± 1.5%, but 10.5% of that field representative's leading digits are 6's. These unusual results suggest that the field representative may have fabricated some of the data. These confidence intervals are computed as $p_d \pm 2 \cdot SE$.

Table 3.
An Example of Data from Typical and Unusual Field Representatives

Leading Digit (d)	CEQ's Nationwide Distribution (n=734,684)	A Typical FR (θ= 10.39) (n=1,143)	An Unusual FR (θ= 102.43) (n=1,132)
1	30.5	31.4	28.9
2	19.3	19.7	18.0
3	12.3	11.6	8.1
4	9.0	9.5	8.5
5	10.4	8.3	17.2
6	6.8	6.4	10.5
7	4.8	4.7	4.2
8	4.4	5.2	3.2
9	2.5	3.2	1.3
Total	100.0	100.0	100.0

The chi-square distribution with 9−1=8 degrees of freedom has a mean of 8.0 and a standard deviation of 4.0, hence only 1 out of every 1,000,000 field representatives should have a test statistic greater than 42.7. However, an examination of the CEQ data reveals 5 field representatives with test statistics greater than 42.7, and 1 field representative with a test statistic greater than 100.0. This is strong evidence that some of the field representatives' data do not follow the expected distribution. Their data are suspicious, and those field representatives should be investigated to determine whether they are curbstoning.

9. Conclusion

Benford's Law is a simple and powerful tool that can be used to help identify possibly fraudulent or error-prone survey data in many settings, including sample surveys. It is important to identify incorrect survey data because the accuracy of any survey's results depends on the accuracy of the collected data.

Although Benford's Law was first discovered 120 years ago, it has been rapidly gaining popularity over the past decade. Its new-found popularity is mostly in the accounting and auditing industries, but there is great potential for its use in the field of sample surveys as well. In fact, the universality with which it applies to nearly every "real world" data set is one of its more curious and powerful aspects.

Although Benford's Law can be a powerful tool in identifying falsified survey data, we hasten to point out that it really only identifies *unusual* data. As with any statistical or quality control tool, after the unusual data have been identified they must be

examined to determine whether or not they are accurate. Benford's Law is a potentially powerful tool that can be added to other quality control tools used in the world of surveys to increase the accuracy of the data.

10. References

Benford, Frank, "The Law of Anomalous Numbers," *Proceedings of the American Philosophical Society*, vol. 78, pp. 551-572, 1938.

Biemer, Paul P., and Stokes, S. Lynne, "The Optimal Design of Quality Control Samples to Detect Interviewer Cheating," *Journal of Official Statistics*, vol. 1, pp. 23-39, 1989.

Browne, Malcolm W. (1998). "Following Benford's Law, or Looking Out for No. 1." *New York Times*, August 4, 1998.

Bureau of Labor Statistics (1997). *BLS Handbook of Methods*. U.S. Department of Labor.

Gabler, S., Haeder, S and Lahiri, P (1999). "A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering," *Survey Methodology*, vol. 25, pp. 105-106, 1999.

Hill, Theodore (1995). "Base-Invariance Implies Benford's Law." *Proceedings of the American Mathematical Society*, vol. 123, pp. 887-895.

Hill, Theodore (1995). "The Significant-Digit Phenomenon." *American Mathematical Monthly*, vol. 102, pp. 322-327.

Hill, Theodore (1995). "A Statistical Derivation of the Significant-Digit Law." *Statistical Science*, vol. 10, pp. 354-363.

Hill, Theodore (1996). "The First-Digit Phenomenon." *American Scientists*, vol. 86, pp. 358-363.

Hill, Theodore (1997). "Benford's Law." *Encyclopedia of Mathematics Supplement*, vol. 1, p. 102.

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

Turner, C.F., J.N. Gribble, A.A. Al-Tayyib, J.R. Chromy (2000). "Falsification in Epidemiologic Surveys: Detection and Remediation." *Technical Papers on Health and Behavior Measurement*, No. 53. Washington, DC: Research Triangle Institute.