Detecting Falsified Cases in SCF 2004 Using Benford's Law

Yongyi Wang and Steven Pedlow
National Opinion Research Center, 55 E. Monroe St., Chicago, IL 60603

Abstract

Interviewer falsification is a constant worry in survey data collection. Traditionally, phone verification of a subsample of completed interviews is the primary method used for detection. However, it is slow and imperfect.

Among the alternatives, we are exploring the use of Benford's Law (the distribution of the first digits in a random set of variables) to detect anomalies. Theoretically, numerical data can be compared to Benford's Law on a flow basis and suspicious interviews and interviewers can be flagged for investigation. To explore the effectiveness of using Benford's Law, we used a data set with many financial (numerical) variables, SCF 2004. Knowing two falsifying interviewers detected through other methods, we examined SCF 2004 data near the end of the field period to see if Benford's Law could be used to pick out these interviewers. Unfortunately, we found that even with this number-rich data, Benford's Law could not pick out the falsified interviews. Also, they were not the top two, but they were near the top.

Keywords: Interviewer, Validation, Survey of Consumer Finances, Leading digit, First digit

1. Introduction

Despite the best efforts and high standards NORC uses to hire interviewers, there is still the risk that interviewers will falsify cases. Currently, NORC's main strategy to identify falsified cases is through phone validation of a subset of completed cases. Often, validation lags behind data collection enough so that when any "failed validations" occur, much time and money has already been wasted. Also, it is not feasible to verify every completed case, so it is possible that an interviewer who does not falsify a large enough percentage may "get lucky" and avoid detection. It is very appealing to attempt to identify suspicious cases quicker through a more mathematical approach.

Within NORC, Javier Porras and Ned English (Porras and English, 2004) have led some investigations in identifying suspicious cases in a large health survey at NORC. One of the methods they investigated was Benford's Law.

Benford's Law asserts that for any set of continuous variables, the distribution of the leading digit is well approximated by,

$$\Pr(d = d_0) = \log\left(1 + \frac{1}{d_0}\right),$$

where d_0=1, 2, 3… 9. The distribution is thus heavier for lower digits than higher digits. The right-most column of Table 1 shows the distribution of first digits according to Benford's Law. As an illustration of the logic, take a distribution uniformly distributed between 500 and 1500. About half of any sample will have a leading digit of "1". It should be noted that values of exactly zero are not accounted for by Benford's Law. We follow the convention that pure zeroes are not valid observations for the purposes of Benford's Law. We treat them as if they are missing.

It is important to note that Benford's law does not refer to the values of any single variable; it refers to one observation from each of a set of variables. In order to obtain Benford's Law, many variables must be available. Since there are nine possible values for the leading digit, each variable will provide an observation to a categorical variable with nine categories.

Benford's law has been used by organizations such as the IRS and accounting firms to search for anomalies in numerical distributions (Hill 1999, Browne 1998). It is widely agreed that people have a difficult time "faking" randomness, and it is therefore expected that while "real" data will follow Benford's law, falsified data will not.

However, the data analyzed by Porras and English did not follow Benford's Law. This was largely because it

is a short interview (15-20 minutes), so there were not enough different questions to achieve the distribution described by Benford's Law. The interview had only about thirty usable (numeric) items. Also, many of these variables were related (and had similar values), which meant that there were not thirty independent observations of the leading digit. Finally, many variables were single-digit and bounded, which resulted in a distortion of Benford's Law and well as much rounding to the number five.

Nevertheless, Porras and English showed that falsified cases detected through validation did differ in their leading digit distribution from the rest of the cases. This project aims to show whether a richer data set, such as SCF 2004, better conforms to Benford's Law. If so, using Benford's Law could lead to an improved methodology for sample surveys to detect falsified interviews.

2. SCF 2004

The 2004 Survey of Consumer Finances (SCF), sponsored by the Federal Reserve Board (FRB), and fielded by the National Opinion Research Center (NORC), collects financial information from a national area probability sample of housing units, with an additional list sample from the same areas. While our ultimate goal is to use Benford's Law on a flow basis during NORC projects, SCF 2004 data collection was about 90% complete when we performed our analyses. Two interviewers were identified through validation (at least 10% of the cases for every interview needed to be verified by calling back the respondent) as having falsified cases. These two interviewers accounted for 52 completed interviews, out of which 42 were identified as falsified cases. We assumed that the other 10 cases done by the two interviewers and the 4140 cases done by 171 other interviewers were "good" cases. Therefore our analyses consisted of two parts: 1) to compare the 4150 "good" cases with the 42 "bad" cases, and 2) to compare the 171 "good" interviewers with the 2 "bad" interviewers.

SCF 2004 contains many numeric variables since the focus of the interview is on the complete financial picture of the respondent and his/her household. We identified 502 such variables, and it was our hope that these 502 variables would follow Benford's Law. However, not all questions are asked of all respondents. In fact, only a small percentage actually contains data because of the complex paths and skip patterns necessary to collect financial data on very different individuals. For the good interviews, we obtained only an average of 29.08 valid financial values (standard deviation of 15.31). Interestingly, for the bad interviews, there were only 19.33 valid financial values on average (standard deviation of 7.62). This suggests an alternative hypothesis: interviewers (and interviewees) who want to falsify data will tend to choose a faster path through the questionnaire, and might be detected through this different measure. We have not explored this currently, but plan to in the future.

3. Methodology

Our methodology to judge the effectiveness of Benford's Law is straight-forward. For every interview and every interviewer, we have the distribution of first digits and can calculate a score that compares what is observed vs. what is expected under Benford's Law.

$$Score = \sum_{i=1}^{9} \frac{(O_i - E_i)^2}{E_i} \sim \chi_8^2$$

In the formula above, O_i is the observed frequency and E_i is the expected frequency (from Benford's Law). This score, under the null hypothesis that the interview does follow Benford's Law, will follow a chi-square distribution with eight degrees of freedom (one less than the number of cells).

We compare several modifications of Benford's Law and each method's effectiveness is judged by how well it segregates the bad interviews and interviewers from the good interviews and interviewers.

4. SCF 2004 Data Analyses

4. 1 Theoretical Benford's Law

Our first analysis used the idealized Benford's Law. Figure 1 (bar chart) and Table 1 (the numbers for the bar chart) compare the SCF data with Benford's Law. There are four bars in Figure 1. The fourth bar indicates the Benford's Law distribution. It is very striking to see that the third bar is the farthest away from Benford's Law except for the digits 4 and 7; these are the cases completed by the two "bad" interviewers. It is easy to understand why the first two bars (the first bar is all cases together and the second bar is the cases completed by "good" interviewers) are almost exactly the same for all digits. Almost all of the cases (99.0%) were completed by the "good" interviewers. The good

interviewers' cases are amazingly similar to Benford's Law for digits 1-3. However, it is also clear that the cases of good interviewers are too high for the 5th digit and too low for digits 4 and 6-9 (especially 9).

The conclusion that can be drawn is that good interviewers' cases generally follow Benford's Law, but there is clearly a rounding effect in the interviewing data. 12.4% of the first digits in the good interviewers' data were 5's compared to an expected 7.9% under Benford's Law, a 50% surplus. At the other end, only 2.6% of the first digits in the good interviewers' data were 9's compare to an expected 4.6% under Benford's Law, a shortfall of almost 50%. This rounding causes some difficulties in the effectiveness of using Benford's Law to detect the bad cases.

We calculated a score for each interview and each interviewer. Then, we sorted the interviews and interviewers (separately) from high to low score. It was our hope that the bad interviews/interviewers would be segregated with the highest scores. However, the data suggest that this is not the case. At the interview level, the first falsified case ranked 110th out of all 4,192 cases. At the interviewer level, the two falsifiers ranked 29th and 33rd respectively out of all 173 interviewers (see Table 2 for the abbreviated ranking).

Of course, even if the data did follow Benford's Law, we would expect 5% of the interviews and interviewers to exceed the critical value of 15.51. In fact, 16.1% of all interviews (674/4192) exceed this critical value; only 16.7% of the bad interviews (7/42) exceed 15.51. For the interviewers, both of the bad interviewers exceed this critical value, but so do 68.2% (118/173) of the interviewers.

4.2 Using the All-cases Distribution

Instead of using the theoretical Benford's Law distribution (which the SCF data doesn't match because of rounding), it is natural to use the distribution of first digits that the SCF does follow. Optimally, of course, we would want to use only the good interviews and interviewers. However, since our objective is to pick out bad interviews and interviewers (not already identified), we need to use the distribution for all cases.

The all cases distribution does much better than the theoretical Benford's Law distribution in segregating the bad interviews and interviewers (relative to good interviews and interviewers). At the interview level, the first falsified case ranked 60th. At the interviewer level, both bad interviewers were among the top 5 out of all 173 interviewers (see Table 3 for the abbreviated ranking).

We can see two things from these numbers. First, the distributions of first digits are clearly different from Benford's Law due to the rounding effect. Therefore, using the all cases distribution instead of the theoretical Benford's Law achieved much better results. Second, we have very little power to detect bad interviews because of insufficient data at the interview level, but considerably more power to detect bad interviewers.

In an attempt to return to Benford's Law (instead of the observed distribution of all first-digits), we tried to make three modifications to the theoretical Benford's Law. We carried out these three modifications on the Benford's Law distribution and the all cases distribution. Because we have very little power to detect bad interviews, we will only present the data at the interviewer level.

4.3 Three-cell Benford's Law

Since we had only 20-30 observations to fit into nine cells, we explored using only three cells by collapsing digits 1-3, 4-6, and 7-9.

The good interviewers' cases are still lower in first digits of 7-9 than Benford's Law would suggest. In general, using the three-cell Benford's Law behaved better than the nine-cell scenario as determined by the bad interviewers appearing higher in the sorted list (e.g., the bad interviewers rank 29th and 33rd under the nine-cell scenario, but 15th and 17th under the three-cell scenario). Using the all cases distribution, the three-cell scenario and the nine-cell scenario perform about equally well (the bad interviewers rank 4th and 5th under both scenarios). Again, we had more success segregating bad interviewers by using the all cases distribution, rather than Benford's Law distributions.

4.4 Combining 4-9

We noted that the good cases were very close to the Benford's Law distribution for the digits 1-3, but was less close for digits 4-9. By collapsing the digits 4-9 into one category, we made the all cases distribution very similar to the Benford's Law distribution. This results in a four-cell Benford's Law modification.

The all cases distribution again outperforms the theoretical Benford's Law distribution. Using the all cases distribution, the two bad interviewers rank 9th and 30th respectively, whereas under Benford's Law they rank 35th and 81st respectively. However, the overall results are worse than the nine-cell and three-cell scenarios.

4.5 Not Using 5

We also noted that the largest discrepancy was caused by the rounding of first-digit values to 5. Therefore, we simply treated all fives as if the values were exactly zero (i.e., these values were not used). This results in an eight-cell Benford's Law distribution.

Under this scenario, the all cases distribution (bad interviewers were 1st and 6th) again works better than Benford's Law (bad interviewers were 34th and 36th) in segregating the bad interviewers from the good ones. Particularly, one of the bad interviewers ranked 1st using all cases distribution!

The abbreviated ranking of interviewers using all cases distribution under the last three scenarios is presented in Table 4. The results of all scenarios are summarized in Table 5.

5. Discussion

The main question that this work raises is whether it is possible, practical, and effective to use the distribution of the first digits of numerical variables to detect falsified cases. While these first digits do approximate Benford's Law, the SCF data shows a significant rounding effect to 5 and against higher numbers. Since this rounding effect also occurred in other survey data, it may be that sample survey data from respondents will not follow Benford's Law because of the tendency of respondents to round off numbers, resulting in many more numbers with a first digit of 5 than Benford's Law would suggest.

The first digits for SCF do seem to follow an approximation to Benford's Law. It seems that using the all cases distribution instead of the Benford's Law distribution does result in a more sensitive tool to detect falsified cases.

It is clear that SCF does not have enough data to detect falsified cases individually. While we accessed 502 variables, we only ended up with about 20-30 variable values on an average case, which is not sufficient for falsified cases to stand out among the random perturbations. In general, we believe surveys, even large ones such as SCF 2004, will not have sufficient data to detect individual falsified cases using Benford's Law.

At the interviewer level, using the all cases distribution, the original nine-cell version performed about equally well as two of its modifications: the three-cell version and "not using 5". "Combining 4-9" was the least favorable scenario. We suggest that sample surveys with less numerical data try the nine-cell and three-cell versions to see if reducing the number of cells is helpful.

Although it might be difficult to use the Benford's Law method alone to detect falsifiers, this method could be used to spot suspicious interviewers quickly and do manual validation on their cases immediately. We note that the bad interviewers were both in the top 6 for all "all cases" scenarios except "combining 4-9". Concentrating effort on the top 10 scores can maximize the efficiency of validation efforts.

It is also important to note that this research was done near the end of SCF 2004 when much of the data had been collected. It is not known how well such a scenario would do at spotting suspicious interviewers in production (e.g., running a weekly process to search for FIs to send for extra validation). We also intend to follow up on investigating whether suspicious interviewers can be spotted earlier in a field period than by traditional validation methods.

References

Porras J and English N (2004), "Data-Driven Approaches to Identifying Interviewer Falsification: The Case of Health Surveys," 2004 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association: 4223-4228.

Hill, Theodore (1999), "The Difficulty of Faking Data," Chance Magazine 12(3), 37-31.

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

Figure 1. Distribution of the first digits

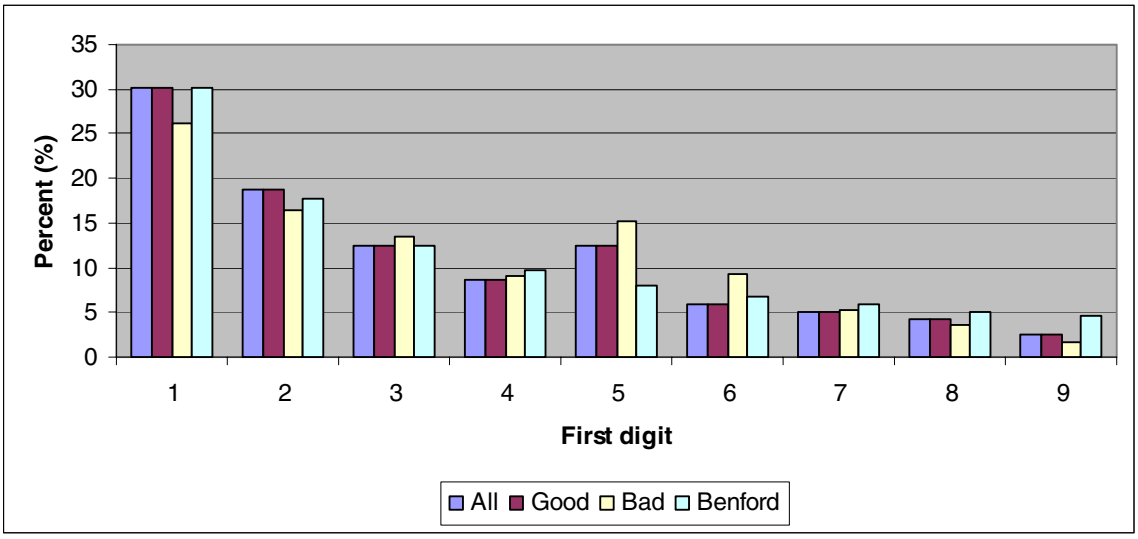

Table 1. Distribution of the first digits

First Digit	All Cases	Good Interviewers' Cases	Bad Interviewers' Cases	Benford's Law
1	30.06%	30.09%	26.13%	30.10%
2	18.82%	18.84%	16.37%	17.61%
3	12.48%	12.47%	13.51%	12.49%
4	8.55%	8.55%	8.97%	9.69%
5	12.39%	12.37%	15.19%	7.92%
6	5.91%	5.88%	9.27%	6.69%
7	4.99%	4.98%	5.33%	5.80%
8	4.20%	4.20%	3.55%	5.12%
9	2.60%	2.60%	1.68%	4.58%

Table 2. Compared against Benford's Law, scores of interviews and interviewers.

Rank	Interview	Score	Rank	Interviewer	Score
1	GOOD	84.34	1	GOOD	244.44
2	GOOD	66.01	2	GOOD	192.59
3	GOOD	55.13	3	GOOD	151.66
4	GOOD	50.27	4	GOOD	144.58
5	GOOD	49.00	5	GOOD	132.37
6	GOOD	47.61	6	GOOD	131.17
7	GOOD	42.48	7	GOOD	118.04
110	**BAD**	**25.25**	**29**	**BAD1**	**63.38**
165	**BAD**	**23.17**	**33**	**BAD2**	**56.06**
4192	GOOD	0.83	173	GOOD	2.17

Table 3. Compared against all cases distribution, scores of interviews and interviewers.

Rank	Interview	Score	Rank	Interviewer	Score
1	GOOD	64.31	1	GOOD	36.46
2	GOOD	62.49	2	GOOD	34.54
3	GOOD	46.26	3	GOOD	27.33
4	GOOD	45.70	**4**	**BAD1**	**26.81**
5	GOOD	45.63	**5**	**BAD2**	**24.95**
6	GOOD	44.91	6	GOOD	23.45
7	GOOD	41.81	7	GOOD	23.14
60	**BAD**	**28.29**	8	GOOD	22.61
198	**BAD**	**20.78**	9	GOOD	21.30
4192	GOOD	0.63	173	GOOD	1.09

Table 4. Compared against all cases distribution, scores of interviewers (three scenarios)

Three cells			Combining 4-9			Not using 5		
Rank	Interviewer	Score	Rank	Interviewer	Score	Rank	Interviewer	Score
1	GOOD	15.99	1	GOOD	14.69	**1**	**BAD1**	**24.77**
2	GOOD	14.26	2	GOOD	12.39	2	GOOD	22.63
3	GOOD	12.51	3	GOOD	11.91	3	GOOD	21.65
4	**BAD1**	**11.76**	4	GOOD	11.6	4	GOOD	21.48
5	**BAD2**	**10.96**	5	GOOD	11.34	5	GOOD	21.08
6	GOOD	10.67	6	GOOD	11.26	**6**	**BAD2**	**21.06**
7	GOOD	10.23	7	GOOD	11.12	7	GOOD	20.38
8	GOOD	9.73	8	GOOD	10.75	8	GOOD	20.01
9	GOOD	8.89	**9**	**BAD1**	**10.58**	9	GOOD	19.82
10	GOOD	8.88	**30**	**BAD2**	**6.62**	10	GOOD	19.06

Table 5. Summary of all scenarios

	Ranks of Falsifier Interviewers (173 total)	
Scenario	**Benford's Law**	**All Cases Distribution**
Nine cells	29, 33	4, 5
Three cells	15, 17	4, 5
Combining 4-9	35, 81	9, 30
Not using 5	34, 36	1, 6