Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys

Javier Porras, National Opinion Research Center
Ned English, National Opinion Research Center

I. Introduction

The threat of interviewer-falsified data requires survey organizations to devote significant amount of resources to deter its creation and identify its existence within real data. Traditionally, field survey organizations have employed holistic measures to catch interviewers who fabricate data (Johnson 2001). A general problem with the traditional approaches is that they tend to rely on chance and may miss savvy falsifiers who are aware their work is in question (Murphy 2004). In more recent years with the decrease in response rates, the pressure on interviewers to maintain their productivity levels has only increased, thereby increasing the potential for them to falsify interviews (Murphy et al. 2004).

The presence of falsified data within real data dilutes the quality of data as estimates are potentially biased. In additional to detracting from the quality of a survey, falsified data can seriously damage the reputation of an organization, as well as raise certain ethical questions (Reed and Reed 1997, Greenberg and Goldberg 1994).

Using data from a large-scale health study conducted by NORC, we looked to develop three data-driven approaches that aim to identify fabricated interviewer data. Along with being easy to implement and relatively inexpensive, we believe these methods also benefit from their non-intuitive underpinnings (to the interviewers), which make them difficult for the average interviewer to outsmart. The three approaches we will now discuss are briefly described below.

1. Goodness-of-fit to Benford. The leading digits of a random collection of distributions can be frequently approximated by a Benford distribution. Although the leading digits of our dataset did not conform to a Benford distribution, they did form a latent distribution to which the falsified data did not conform.
2. Lack of Variance. It is theorized that interviewers falsifying data tend to center their data around their "intuitive mean". By ranking the relative variances of interviewers' means, we demonstrate that the interviewer who falsified his data produced some of the lowest relative variances.
3. Unlikely Combinations. We hypothesize that falsifiers will occasionally outsmart themselves by recoding item combinations that rarely occur in real data. One example would be heavy smokers who also get considerable quantities of vigorous exercise. If such items are present, it may be an indicator of falsification. The question, then, is what combinations to search for and how to determine if they are legitimate or not.

For the purposes of conducting our analysis, we obtained two sources of falsified cases. The first source of data came from one of the project interviewers. A portion of his data could not be validated, and it was then decided to remove all of his completed questionnaire data from the final dataset. For the second source of falsified data, we instructed five of our interviewers to generate falsified questionnaire data, producing a total of 50 falsified interviews. These two sets of falsified data provided the opportunity for us to test the three methods which we now present.

II. Benford's Law

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

$$p(d = d_0) = \log(1 + 1/d_0)$$

where d_0 =1, 2, 3, …, 9. A leading digit is the leftmost non-zero digit in a number (9 and 5 are the leading digits of 942 and 52, for example).

Figure 1 shows frequencies of a Benford Distribution and a uniform distribution. The naïve estimate would be that a distribution of leading digits would be approximated by a uniform distribution (Hill 1999). However, the Benford distribution has proven to approximate the distribution of leading digits for many real world data. In particular, tax agencies have successfully applied Benford's Law to identify persons falsifying tax forms (Hill 1999).

We also looked to apply Benford's Law to determine if it would successfully flag the

interviewer who falsified interviews in this health study. We combined the questionnaire items with continuous data to form a distribution of leading digits to compare against Benford's distribution. It, however, did not approximate a Benford distribution. So it came as no surprise that none of the interviewers data conformed to Benford either.

In spite of inability of apply Benford's Law directly, we noticed that the leading digits of the falsified data appeared to have a different distribution compared to the overall distribution of leading digits[1]. So we decided to apply the "spirit" of Benford.

In this modified approach, the data were separated into two groups: leading digits from the data of the interviewer who is being tested against the distribution of leading digits from all other interviewers. That is to say that the "all other" distribution of leading digits was taken to be the "true" distribution. (Note that this approach made the true distribution dependent on which interviewer was being tested. Because of the large amount of data involved, the differences of the true distributions were not significant.) A chi-square test compared a given interviewer against the "true" distribution provided by the rest of the data. The interviewer whose data could not be validated was tested after his first 20, 50 interviews, and all 138 interviewers were completed. Three other interviewers who provided real data were also tested at the same stages (20-, 50-, and about 140-interview stage). The reason for testing at the 20 and 50 stage mark was to simulate the results of testing during data collection. If the simulation showed significant results after, say, 20 interviews, then this would suggest that the modified Benford approach would be a promising approach to use *during* data collection, and not only after all interviewing had been completed.

Table 1 shows the *p*-values for each of the interviewers at the 20, 50 and 140 interviewing stage. The results are that the interviewer who we believe to have generated a large amount of falsified data had the smallest *p*-values at each stage. As the number of interviews increased, his *p*-values continued to distinguish him from the other interviewers whose data had been validated.

It is interesting to note, however, that one of the interviewers (interviewer #2) also had significant *p*-values (α=0.0046) at the 140 stage. While the interviewer's *p*-value appear to be significant, it is nearly 170,000 times bigger than the interviewer who fabricated data (interviewer #1). The pending question, then, is how does one determine an appropriate level define significant p-values. This paper does not attempt to answer this question.

While this Benford-like approach shows much promise in being a tool to identify fabricated data, its shortcomings became apparent in the course of our research. We uncovered three limitations. First, respondents tended to round answers to the leading digit 5 (5, 50, 500, …), causing the data to spike at 5 that was inconsistent with a Benford distribution. (A rounding effect was probably also taking place at leading digit 1. But as 1 is the most common leading digit, the magnitude of the spike is not evident.) Unlike tax, accounting and financial agencies that have successfully used Benford's Law to detect falsified data, our field of survey research faces considerably more rounding of numbers that do not necessarily represent deception by the interviewer. In a worst-case scenario, rounding of answers and interviewer falsifying of data become indistinguishable, rendering Benford's Law invalid. Second, the health survey data we were using was not rich in terms of continuous data. Because of the lack of numeric data, we were forced to collapse the testing to the *interviewer level* (as opposed to *interview*) in order to produce a sufficient number of leading digits. Ideally, each interview would be tested. The main advantage of using Benford's Law to test each interview would be the ability to produce powerful evidence against an interviewer who is consistently generating suspicious leading digit patterns. Third, the theory states that random samples from a random collection of distributions will tend to have their collective distribution of leading digits approximated by a Benford distribution. If our distributions had been random, we would not have any reason to expect any significant correlations between the distributions. However, the data are correlated. As a health study that collected different types of food serving (i.e., fruit servings and vegetable servings) data, it

[1] For this analysis, the falsified data generated at our request by the interviewers was excluded. Only the fabricated data from the interviewer caught through validation was included.

should come as no surprise that these kind of data are correlated (and thus not from a random collection of distributions).

III. Lack of Variance

Before beginning our research, we thought it would be a risky proposition to draw conclusions based on an interviewer's "sample" means. With relatively small sample sizes and the tremendous weight differentials across communities[2], the value of testing for outlier means to flag suspicious data was not clear and potentially would lead us in the wrong direction. There would be too many reasons, we felt, why an interviewer's sample data would depart from known or estimated population averages that have no connection to interviewer falsification.

We theorized that if an interviewer had fabricated a large amount of interviewer data, he would attempt to escape detection by avoiding outliers. In doing so, he would have to have his own "intuitive" mean around which all his data would revolve, without deviating too far from it.

In order to test our theory, we looked at the relative variances of the interviewer #1 (who is believed to have falsified a large amount of interviews) and a few other randomly-chosen interviewers who completed about the same number of interviewers. We computed the relative variances of nine questionnaire items for each of the interviewers, and then ranked the relative variances. Clearly, the relative variances of interviewer #1 distinguished him from the rest of the group. Table 2 shows the rankings based on 1 of the questionnaire items. Table 3 shows the average rankings of the interviewers based on all nine questionnaire items examined. Again, it is clear that interviewer #1 is consistently producing lower relative variances than the rest of the group.

IV. Unlikely Combinations

The third approach was to examine the prevalence of unlikely question combinations that do not occur often in true interviews but may occasionally be present in falsified data. Because the survey in question was related to community health, we focused on counter-intuitive responses as follows:

1. Elderly people who get considerable vigorous exercise
2. Heavy smokers who exercise vigorously in any quantity
3. Heavy smokers who get considerable amounts of moderate exercise on a regular basis
4. Respondents who consume large quantities of fruit but very little vegetables
5. Heavy smokers who consume large quantities of both fruit and vegetables
6. Very elderly people who report no illness at all

Table 4 summarizes the comparison between the prevalence of these question combinations in known falsified cases as well as the true data. As shown, not all combinations were successful in indicating falsified cases. In attempts to remain undetected, falsifiers avoided obviously bizarre answers. Some combinations, however, appeared in significantly larger quantities in the falsified data in comparison with the true results. Three combinations had significantly different proportions in the falsified and non-falsified cases: elderly and considerable vigorous exercise; heavy smoking and considerable fruit and vegetable consumption elderly and no reported health problems.

Note that these differences can be in opposite directions; two of the combinations in concern had significantly higher proportions in the true cases than the falsified cases, while one was the opposite. One theoretical difference between finding red-flag question combinations and the two previous approaches is that the former requires some *a-priori* knowledge of the survey in question, while the latter does not. We argue that 'red flag' question combinations such as these should be searched for in questionnaire data as a simple first step in falsification detection.

V. CONCLUSION

We have presented three promising approaches that can help identify and deter interviewer falsification of data. Using a modified approach to Benford's Law, we showed how the leading digit distributions of numeric data can flag suspicious data patters. In spite of its dependency on numeric data, Benford's Law

[2] The interviewing was done by telephone. In general, interviewers were not limited to interviewing a specific community. Instead, they conducted the interviews across the targeted communities of the study.

appears to be a promising tool to identify suspicious interview data for survey research studies, which generally does not collect as much continuous data as, say, financial and accounting agencies do. The Lack of Variance approach looks at interviewers' relative variances and compares them to each other. Our research suggests that an interviewer who is falsifying data may tend to underestimate parameter variances, and so those interviewers who consistently produce the smallest variances on numerous questionnaire items are the ones whose data may require extra validation checks. Finally, the rare combinations approach also is promising in that it was able to indicate potential falsifiers in surveys with *a-priori* information about expected behavior. Future research will look to develop each of these methods of identifying interviewer data falsification.

VI. REFERENCES

Greenberg, Michael and Laura Goldberg. 1994. Ethical Challenges to Risk Scientists: An Exploratory Analysis of Survey Data. *Science, Technology, and Human Values* 19(2), 223-241.

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

Johnson, Timothy P., Vincent Parker, and Cayge Clements. 2001. Detection and Prevention of Data Falsification in Survey Research. *Survey Research* 32(3),1-3.

Murphy, Joe, Rodney Baxter, Joe Eyerman, David Cunningham, and Joel Kennet. 2004. A System for Detecting Interviewer Falsification. Paper presented at the American Association of Public Opinion Research Annual Meeting, May 16, 2004, Phoenix, AZ.

Reed, Stephanie J. and John H. Reed. 1997. The Use of Statistical Quality Control Charts in Monitoring Interviews. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Appendix Tables

Figure 1: Benford distribution and Uniform distribution

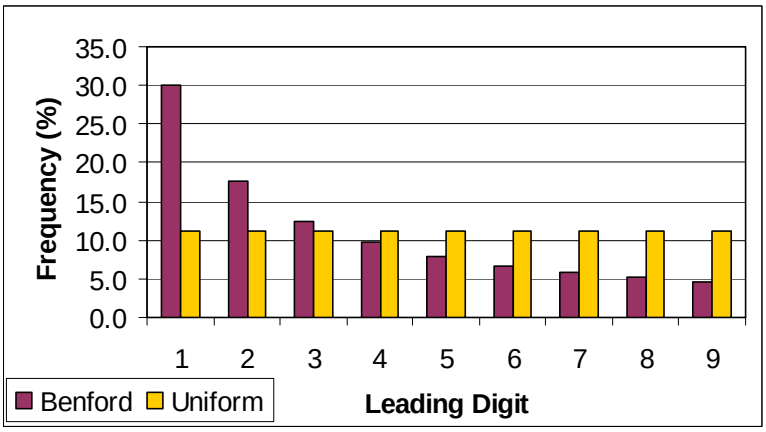

Table 1: Comparison of leading digit frequencies, interviewer #2 suspected of falsifying data

	After 20 Interviews		After 50 Interviews		All Interviews	
ChiSquare Results	pvalues	Ratio pval/ min(pvals)	pvalues	Ratio pval/ min(pvals)	pvalues	Ratio pval/ min(pvals)
Interviewer 1	0.0095	1.0	0.0119	1.0	0.0000	1.0
Interviewer 2	0.1755	18.4	0.5167	43.5	0.0046	169,852.7
Interviewer 3	0.4335	45.5	0.4141	34.9	0.6432	23,683,403
Interviewer 4	0.6720	70.5	0.7486	63.0	0.2066	7,606,892

Table 2:Ranking of interviewers by relative variances

Interviewer ID	n	Mean	Variance	Relative Variance	Rank
6	14	87.4	7,785	1.02	1
1	11	124.4	16,525	1.07	2
5	16	110.9	18,593	1.51	3
4	20	181.3	59,941	1.82	4
2	16	209.0	89,550	2.05	5
3	15	94.1	20,337	2.30	6

Table 3: Ranking of Interviewers by Average of Relative Variance (based on 9 variables)

Interviewer ID	Average Relative Variance Rank
6	2.22
1	3.33
5	3.56
4	3.78
2	4.00
3	4.11

Table 4: Unlikely combinations in survey data

Combinaton	Number in Falsified Cases	Number in True Cases	N False	N True	% Falsified	% True	*t*-value
Elderly and considerable vigorous exercise	0	52	187	18,670	0.0%	0.3%	**7.22**
Heavy smoking and some vigorous exercise	4	65	187	18,670	2.1%	0.3%	1.69
Heavy smoking and considerable moderate exercise	4	41	187	18,670	2.1%	0.2%	1.81
Consderable fruit consumption and little vegetable	3	400	187	18,670	1.6%	2.1%	0.59
Heavy smoking and considerable fruit and vegetable consumption	6	70	187	18,670	3.2%	0.4%	**2.20**
Elderly and no health problems	0	239	187	18,670	0.0%	1.3%	**15.56**