Bootstrap Approach to The Application of First-Digits Analysis November 2017

Matthew Corrigan, Emily Roche
Bureau of Labor Statistics, 2 Massachusetts Ave, NE Washington, DC 20212

Abstract
The distribution of first-digits obtained from many natural and economic datasets seem to follow a consistent distribution. The desire to find anomalies such as detecting fraud in financial and scientific datasets are common, and applications of “Benford’s Law” have been developed to find these anomalies. In our work with applying these methods to determine interviewer anomalies we found that interviewer’s assigned caseloads contained data where stratified subsets of first-digits follow consistent distributions that are like Benford’s, but not specifically Benford’s. To observe an interviewer objectively, we created a profile distribution by subsampling a mixture from available distributions to match individual interviewer’s profile distribution. Using the interviewer’s proportion of first-digits as a test statistic, we are able to determine bootstrapped p-values for first-digits in a way that allows us to flag interviewer results as suspicious and in need of closer scrutiny.

Key Words: First-Digits; Benford; Curb-Stoning; Bootstrap; Stratified Subsampling

Disclaimer: Views expressed are those of the author(s) and do not necessarily reflect the views or policies of the Bureau of Labor Statistics.

1. Introduction

The Statistical Method Staff for Office of Employment and Unemployment Statistics (OEUS), we aim to increase the quality of data for the OEUS surveys and the statistics produced using the survey data. The Current Employment Statistics (CES) survey uses data from interviewers that call establishments to ask for the number of employees at the establishment. To improve the data in the CES survey and corresponding statistics, we aim to identify, eliminate, or reduce curbstoning by creating tools to aid in detecting dishonest reporting. Curbstoning refers to the deliberate fabrication of survey interview data by the interviewer (Koczela et al., 2015).

This work utilizes first-digit analysis, which commonly follows Benford’s Law. Benford’s Law states that the distribution of leading digits of “real world” numbers will tend to follow a logarithmic distribution (Swanson, 2003). “First-digit analysis” refers to the analysis of expected proportions of leading digits, i.e. proportions of numbers 1 through 9. For example, 399 has a “3” as the leading digit. Auditors have used Benford’s Law for first-digit analysis to help focus efforts when detecting accounting fraud since the late 1980’s (Durtschi, Hillison, & Pacini, 2004). Our first goal was to make sure our data is similar to the Benford’s distribution, and if not, modify the expected proportions. Modifying the proportions allows us to test individual interviewers’ values against the average proportions from all interviewers.

Table 1: Expected Proportions under Benford's Distribution

First-Digit	1	2	3	4	5	6	7	8	9
Proportion	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046
$log_{10}\left(1+\frac{1}{d}\right)$	$log_{10}\left(1+\frac{1}{1}\right)$	$log_{10}\left(1+\frac{1}{2}\right)$	$log_{10}\left(1+\frac{1}{3}\right)$	$log_{10}\left(1+\frac{1}{4}\right)$	$log_{10}\left(1+\frac{1}{5}\right)$	$log_{10}\left(1+\frac{1}{6}\right)$	$log_{10}\left(1+\frac{1}{7}\right)$	$log_{10}\left(1+\frac{1}{8}\right)$	$log_{10}\left(1+\frac{1}{9}\right)$

Benford's Law is used to detect accounting fraud because the first-digit proportions of accounting data regularly follow Benford's Distribution for first-digit proportions. When auditors compare their accounting data to the expected proportions from Benford's Distribution and find that their data has an excess amount of a particular first-digit, they look in detail at the transactions beginning with the suspect digit. To determine if the given distribution is following Benford's Distribution, auditors can check one digit at a time with a z-statistic, or use the Chi Square test on all frequencies to determine if the individual's distributions of first-digits are statistically different from Benford's Distribution. Another example besides accounting, is the use of Benford's Distribution to detect potential underreporting in pollution emissions data, using the same methods (Dumas & Devine, 2000).

1.1 Modifications to Benford's Distribution

The first-digit proportions for Benford's Distribution and CES employment values are similar, but definitely not the same distribution, as the top left graph in Figure 1 below shows. Normally this would require a commonly used modification to Benford's Distribution, where we test interviewer first-digit proportions against the actual first-digit proportions using all available data, instead of Benford's Distribution. If interviewers had comparable workloads this modification would not be a problem. There are two main reasons this modification wasn't applicable to the data: first, subsets of the data have different first-digit proportions, and, second, interviewers could have different number of cases in each subset.

1.2 Strata Selection

One important feature of this methodology is the stratification. The strata that were analyzed for differences in first-digit proportions were NAICS super sector, employment size, how late the data was collected, and number of employment values in a single case. The first thing to inspect is differences in the first-digit proportions for different subsets of the data. Since interviewers are given their workloads primarily at random, it can be assumed they will have different proportions from different subsets of the data.

Some experienced interviewers may also be specialized to enroll for certain industries, although due to the quarterly sampling, all interviewers are working on collecting the recently sampled industries. With many strata, it is not plausible that each interviewer will always have the same proportion from each stratum assigned to their workload on a month-to-month basis.

One explanation for Benford's Distribution's proportions is the difficulty in increasing from leading 1's to leading 2's and so on. For example to raise employment from 1 to 2 is a 100% increase, 2 to 3 is a 50% increase, and each digit is a smaller percentage increase afterwards, until reaching a new leading 1. While that is true for 1 through 9, it is also true for all orders of magnitude (e.g. 100 to

200). Raising from 1 employee to 2 employees will not be as hard as changing the employment of an establishment from 10,000 employees to 20,000 employees. For this reason, we see each order of magnitude has a larger and larger proportion of establishments with a leading one. Based on how they impact Benford's Law, we based our size classes on these orders of magnitude (Nigrini, 1999).

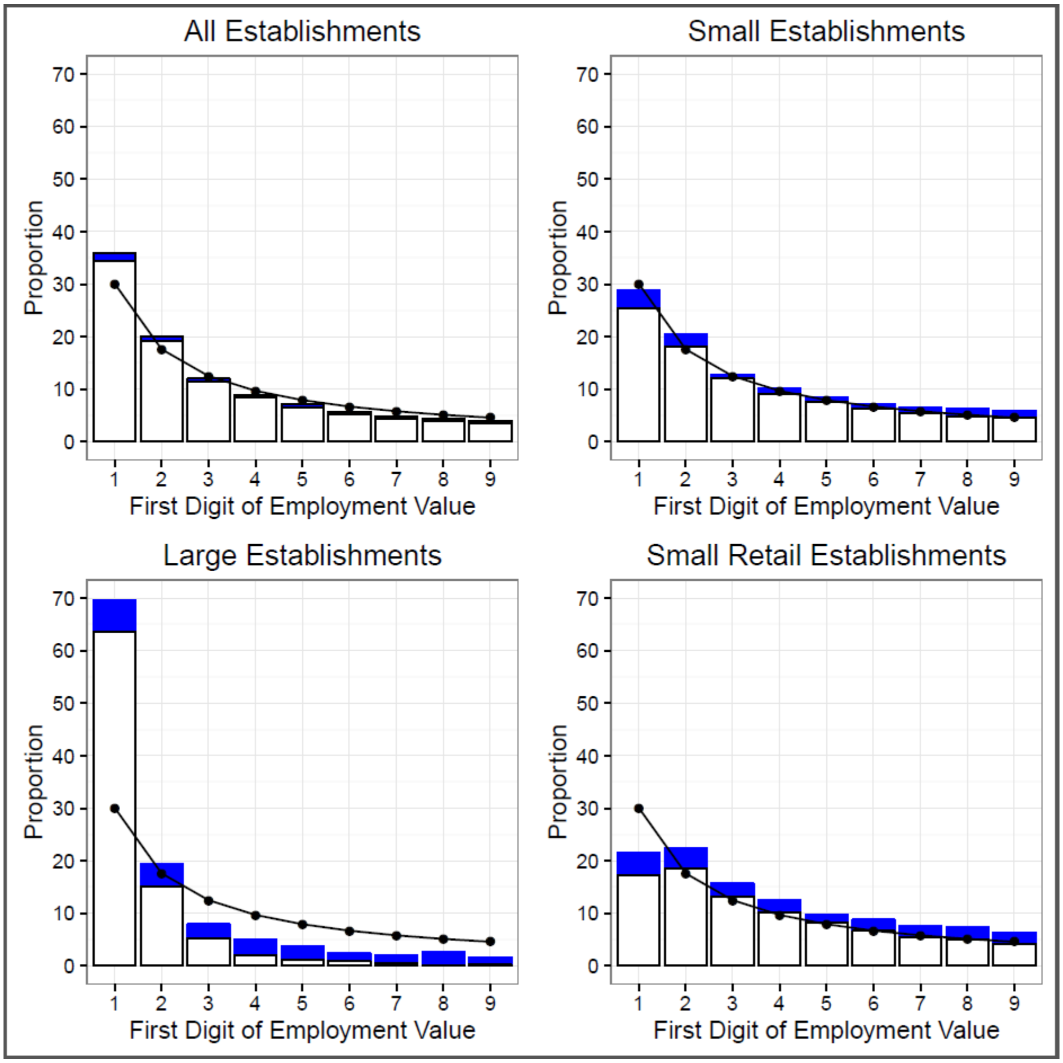


Figure 1: First-Digit Proportions for Subsets
This figure contains a set of first-digit proportions of employment values across sixty months. To display the consistency of proportions across months, the top and bottom of the blue caps for a specific first-digit represents the maximum and minimum proportions respectively. The black line with black dots are the expected Benford's First-Digit Proportions. The upper left graph uses all of the collected employment values across sixty months. The upper right and lower are subset by Small (less than 10) and Large (greater than 99) Establishment employment values respectively. The lower right are the first-digit proportions across sixty months for Small Retail Establishments.

While there are multiple super sectors that have different proportions, they aren't as significant as the size class differences. Some super sectors have a larger difference than others though, as we can see in Figure 1 above, with the Super Sector for Retail Trade. Even though small establishments usually have fewer ones than other digits, Retail Trade has a noticeably smaller number of ones. If naively comparing interviewers with a large proportion of Retail Trade to all other collected employment values, the interviewers would appear as an outlier, just like an interviewer responsible for a large proportion of large establishments would.

1.3 Profile Sample

When planning surveys, samples can be stratified to ensure the sample has the same proportion of some characteristic (e.g., gender, age, establishment size, etc.) among the strata as they do in the population to provide greater precision. In our method, we'd like a sample similar to the interviewer's assigned workload of establishments for a given month. Therefore, we used a stratified sample to mimic the interviewer's workload.

With our strata, NAICS Super Sector and size of establishment, we sample the same number of establishments in each strata that the interviewer collected. For our strata, we used orders of magnitude to include the full range of first-digits for each size (i.e. 1-9, 10-99, 100-999, etc.). Using this method, we give the interviewer the best chance to be accurately represented, as we are not comparing him or her to interviewers with different types of workloads. We call this method of selecting a sample a 'profile sample', as it is a stratified sample, with sample sizes equal to the number of establishments collected by the individual interviewer. Each interviewer will thus have a different profile in composition and size from month to month. See Table 2 below for a simplified example.

Table 2: Simplified example of a profile sample

All Interviewers	*Super Sector 1*	*Super Sector 2*	*Super Sector 3*
Small	$n_{S1}=243$	$n_{S2}=6273$	$n_{S3}=1783$
Medium	$n_{M1}=364$	$n_{M2}=4969$	$n_{M3}=392$
Large	$n_{L1}=110$	$n_{L2}=2247$	$n_{L3}=202$
Individual Interviewer	*Super Sector 1*	*Super Sector 2*	*Super Sector 3*
Small		$u_{S2}=49$	$u_{S3}=23$
Medium	$u_{M1}=2$	$u_{M2}=46$	
Large			$u_{L3}=14$
Interviewers Profile Sample	*Super Sector 1*	*Super Sector 2*	*Super Sector 3*
Small		$u^*_{S2}=49$	$u^*_{S3}=23$
Medium	$u^*_{M1}=2$	$u^*_{M2}=46$	
Large			$u^*_{L3}=14$

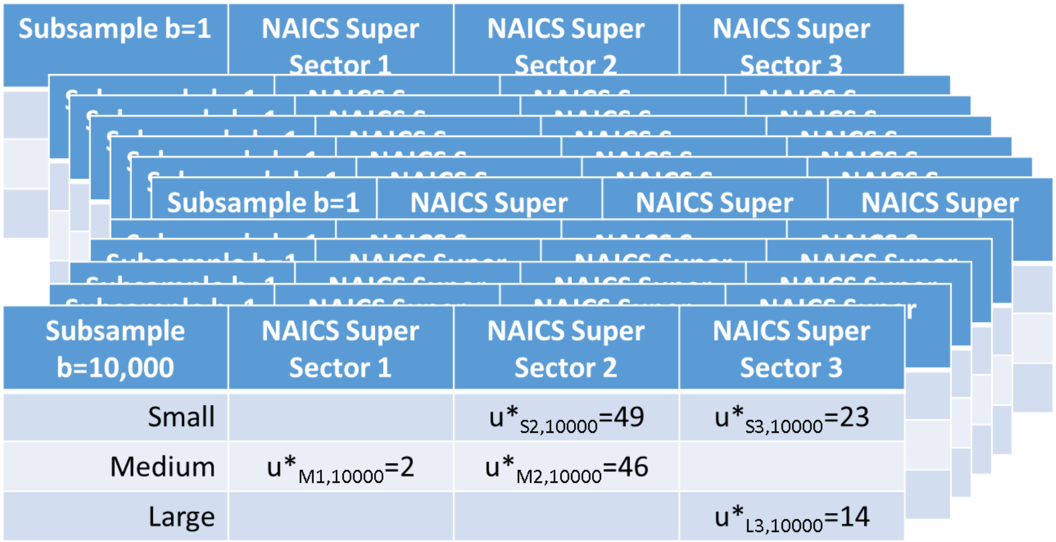
The individual interviewer's profile sample of 134 establishments is selected from a random sample with replacement of the following:

Medium establishments in NAICS Super Sector 1. Two from 364
Small establishments in NAICS Super Sector 2. Forty-nine from 6273
Medium establishments in NAICS Super Sector 2. Forty-six from 4969
Small establishments in NAICS Super Sector 3. Twenty-three from 1783
Large establishments in NAICS Super Sector 3. Fourteen from 202
No establishments are sampled in the unrepresented strata from the individual interviewer profile.

This method of producing a stratified sample presents a significant difficulty. There is only one sample to compare to the interviewer, with no measures of central tendency or variation for the first-digit proportions obtained for a similarly assigned workload. This difficulty is solved by the introduction of the Bootstrap where we can obtain multiple profile samples and estimate the distributions of first-digit proportions.

1.4 Bootstrap

Bootstrap is a method of determining properties of a distribution that are unknown. This could be parameters such as the mean, variance, quartiles, percentiles of random variables, or the parameters themselves. The Bootstrap is comprised of many subsamples taken with replacement from the population. (See Figure 2 below) The parameter values of interest are estimated for each subsample, and with a large number of subsamples the distribution of those parameter values can be obtained and utilized for hypothesis testing (Efron and Tibshirani, 1993).



Subsample b=10,000	NAICS Super Sector 1	NAICS Super Sector 2	NAICS Super Sector 3
Small		$u^*_{S2,10000}=49$	$u^*_{S3,10000}=23$
Medium	$u^*_{M1,10000}=2$	$u^*_{M2,10000}=46$	
Large			$u^*_{L3,10000}=14$

Figure 2: Bootstrapped Subsamples

Limitations of the Bootstrap are that if an individual interviewer has a small number of collected employment values in a given month, the Bootstrap can lead to unusual first-digit proportions. As values may not exist for each of the nine digits. Bootstrapping can only provide an estimate of the distribution for the population if the original sample size is large enough to represent the distribution and the number of replicate subsamples are large depending on the application or parameter you are estimating. Since we are estimating probabilities we need to have a sample of at least 2000 (Schräpler, 2010).

Bootstrap in practice is very simple. Create many profile samples with replacement and use the rank of the first-digit proportions to compute the estimate of the Bootstrap percentile. We have made 10,000 profile samples for each interviewer that has at least 100 reported employment values. We observed that the stability of leading nines improves if observations are greater than 100, hence the arbitrary cut-off at 100. Even though we choose to exclude these interviewers from the application of the profile sample Bootstrap method, their obtained values are still used as shown for All Interviewers in Table 2 above.

Table 3: First-digit Proportions, $\hat{P}$**,** from Bootstrapped Subsamples

B =	{$\hat{P}$(First-Digit = "1"), $\hat{P}$(First-Digit = "2"), …, $\hat{P}$(First-Digit = "9")}
1	{0.32, 0.16, 0.18, 0.05, 0.07, 0.05, 0.05, 0.06, 0.04 }
2	{0.33, 0.24, 0.12, 0.11, 0.04, 0.06, 0.03, 0.03, 0.04 }
3	{0.36, 0.23, 0.08, 0.10, 0.06, 0.06, 0.05, 0.06, 0.02 }
4	{0.43, 0.15, 0.10, 0.09, 0.05, 0.06, 0.07, 0.02, 0.03 }
	{…,…,…,…,…,…,…,…,…}
9997	{0.28, 0.21, 0.14, 0.07, 0.08, 0.07, 0.06, 0.05, 0.03}
9998	{0.27, 0.21, 0.10, 0.11, 0.09, 0.07, 0.06, 0.05, 0.03}
9999	{0.26, 0.21, 0.12, 0.09, 0.10, 0.07, 0.05, 0.05, 0.05}
10000	{0.31, 0.20, 0.11, 0.09, 0.07, 0.08, 0.06, 0.04, 0.04}

The percentiles can be used to determine the probability that an individual interviewer's first-digit proportions came from the Bootstrapped distribution of first-digits subsampled from the population. By taking more Bootstrap profile samples, we obtain a smoother estimate for the population distribution of first-digit proportions. The confidence in determining if an individual user's first-digit values came from those subsampled from the population is thus improved. Theoretical exploration of the Bootstrap can be viewed in Appendix A.

2. Comparing Distributions

We've chosen our strata and created 10,000 Bootstrapped profile samples to describe the distribution of first-digits of similar workloads for each interviewer. Now we need a way to measure how extreme an interviewer's proportions are, or test to see if the interviewer's proportions are likely to have come from the same distribution. We used multiple methods to determine how similar the interviewer's distribution is to the Bootstrap of profile samples. Two methods, Rmax and Rsum, utilize the distance from the median for each first-digit value.

2.1 Method of Comparison

The distance from the median methodology is analogous to the Bootstrapped p-value. The Bootstrapped p-value is determined by the percentage of Bootstrapped proportions higher or lower than that of the interviewer. With enough Bootstrapped profile samples, you have a good approximation of the distribution of first-digit proportions and can measure the percentage of Bootstrap values higher or lower to effectively create a Bootstrapped a p-value. Literature recommends 2,000 replicate subsamples for a Bootstrapped p-value, but we generated 10,000 subsamples (Schräpler, 2010).

In terms of hypothesis testing, our null hypothesis is that the interviewer's first-digit proportions come from the distribution estimated by the profile sample Bootstrap method. The alternative hypothesis is that the interviewer's values are not from the estimated distribution. The test statistic here is the interviewer's first-digit proportions. The test measures the percentage of Bootstrapped samples that have higher or lower first-digit proportions than the interviewer's.

2.1.1 Determining a Score or P-value

If the interviewer has a proportion of first-digit ones that is in the most extreme 5% of the Bootstrap values, the p-value would be less than or equal to 0.05 for that digit. Since we have 10,000 profile samples, the interviewer's proportion of ones would have to be higher than the top 250 or lower than the bottom 250 profile sample's proportions of ones to fail.

In terms of our score, we are viewing it symmetrically with 0 as the median. A lower score would then show that the interviewer's proportions are closer to the estimated distribution. The Rmax and Rsum methods utilize the distance from the median of Bootstrapped replicates and the interviewer's proportions for each digit instead of the p-value. Counting the number of replicate samples that are closer to the replicate median creates a score to show how likely the interviewer's first-digit proportion came from the same distribution. We had to determine where to place the interviewer's proportion when the data is discrete with many repeating values. The following example illustrates the method.

Imagine we only make 20 profile samples for an interviewer, and the number of employment values that start with a 9 are as follows: 1 1 1 1 1 2 2 2 2 2 -- 3 3 4 4 4 5 5 5 5 5. Here we can see that the median (--) value is actually between 2's and 3s. We determine the distance from the median by counting how many profile samples have values between the median and the interviewer's value, for each digit, 1 through 9. Table 4 below shows an interviewer's number of leading 9's compared to the profile samples and the resulting score. This method is conservative because we are choosing to say that the interviewer is closer to the median than all other profile samples with the same value, instead of saying the interviewer's value is in the middle of all similar values.

Table 4: Scoring Example

Number of interviewer's First-Digit 9's	Sequence of Scoring	Score
0	0, {1 1 1 1 1 2 2 2 2 2 --}	10
1	1, {2 2 2 2 2 -- }	5
2	2, {--}	0
3	3, {--}	0
4	4, {-- 3 3 }	2
5	5, {-- 3 3 4 4 4 }	5
6	6, {-- 3 3 4 4 4 5 5 5 5 5}	10

This example calculates $score_9$**,** but we would need to repeat this for each possible leading digit, 1 through 9. With 10,000 profile samples, the highest score possible for each digit is 5,000, whereas the lowest score possible is 0. We have two tests that utilize the score from the distance from the median: Rmax and Rsum. Rmax is the maximum of the 9 scores, which is able to more easily detect spikes in proportions. Rsum is the sum of all 9 scores, which is better at detecting smaller differences at every digit. Their formulas can be viewed below.

$$Rsum = \sum_{i=1}^{9} score_i \, , where\ i\ is\ the\ i^{th}\ leading\ digit$$

$$Rmax = max(for\ all\ score_i), where\ i\ is\ the\ i^{th}\ leading\ digit$$

When we looked for abnormal proportions, we found two common patterns. One looked like a mixture of Benford's Distribution and the uniform distribution, whereas the other looked like a mixture of Benford's Distribution and a normal distribution around the 5's. When we saw the mixture with the uniform distribution, Rsum was able to detect that pattern more often, whereas Rmax was able to detect a mixture with a normal distribution around 5. If the interviewer was primarily making up fraudulent data with either a uniform or normal distribution of first-digit proportions, we would be able to detect that behavior.

With these tests, we are able to find patterns that don't match the expected proportions from the Bootstrap of profile samples. We didn't make these into one unified test because each one tests for something different. The key thing to look for is situations in which any of them are in the extreme values. If an interviewer was to fail a test, they would need to be analyzed for why they are different than their simulated profile sample (Rmax and Rsum).

In typical Benford's Distribution analysis, there is a traditional methodology to rely on the Chi-square test to determine if the observed first-digit frequencies match the expected Benford's Distribution frequencies. The issue with this test is that you will generally not have a significant result with small sample sizes. There are many interviewers in our data that do not have large enough samples to be detected by a Chi-square test. In short, the Chi-Square test is not sensitive enough to be a meaningful discriminator for what we are trying to accomplish.

2.1.2 Multiple Testing

In using the Rmax test a situation of multiple testing presents itself. Under a two-tailed test with an alpha of 0.05, a score higher than 4,750 = [5000 – (0.05*5000)] would be needed to reject the null hypothesis. This method is actually performing 9 tests at the same time. So to be conservative we used the Bonferroni correction for testing all 9 digits at once. In that case a score higher than 4972 = [5000 – ((0.05/9)*5000)] would flag an interviewer for further review.

The benefit of this type of scoring is that managers have a short list of interviewers to review for potential fraudulent behavior. Being flagged in and of itself is not proof of poor behavior, but a tool to reduce which interviewers to check up on. Prior to the creation of this method interviewers were chosen at random for quality control check-ups.

2.2 Graphical Analysis

We wanted to visually investigate interviewers flagged from the Rmax and Rsum tests. To aid in our investigation and visualization we created an R Shiny application around the individual interviewer plot in Figure 3 below. R Shiny is a package from RStudio that builds web applications with a few lines of code, and no JavaScript knowledge is required. The R Shiny application allowed us to quickly create a web application in R and use it on our local machine to dynamically look at detailed information on each interviewer across multiple months.

For the Individual Interviewer Plot in Figure 3 below the green line represents the Benford's Proportion of First-Digits and the Blue Line Represents the Interviewer's First-Digit Proportions. The violin plot shows the first-digits distribution for all profile samples bootstrapped for the interviewer. The purple lines help show the density of how the proportion moves from one first-digit to the next.

A violin plot is a variation on a box plot. It is a combination of a box plot and a kernel density plot (a smoothed histogram). Kernel density estimation is a nonparametric method for estimating the probability density function of a random variable. In general, kernel density plots can be an effective way to view the distribution of a continuous variable. The violin plot effectively superimposes a density curve instead of the more common box and whisker plot (Kabacoff, 2015).

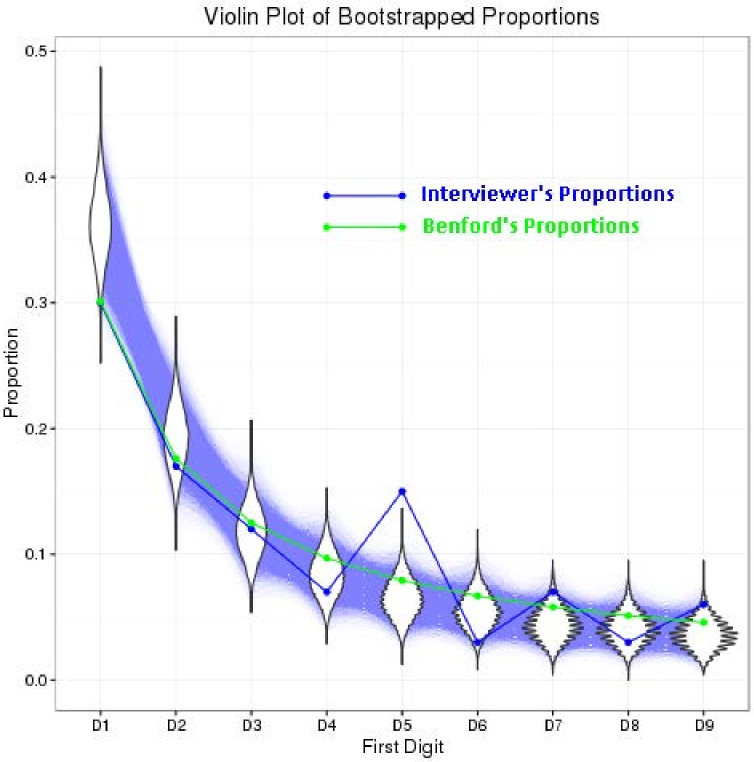


Figure 3: Individual Interviewer Plot

In Figure 3 we can see a situation where the interviewer is not like their expected profile, because the blue line is outside of the Bootstrapped First-Digit distribution of fives. In this case for verification, a manager can listen to recorded calls of the interviewer for any employment value that has a resulting starting value of five. In our profile example from Table 2, this reduces the quality control workload from listening to 134 collected employment call recordings to only listening to 20.

3. Conclusion

At this point, we have many ways to view interviewers for unusual outcomes. The primary test we would recommend using is Rmax. This test identifies the large increases or decreases of a particular digit which are expected in fraudulent accounting data. If an individual fails the Rmax test, we recommend viewing their individual plot. While the Rmax is the main test, analyzing the users who have the worst scores on either test should be viewed using the plot for further review.

When an interviewer’s proportions appear suspicious, a manager can switch all cases between two interviewers to determine if the suspicious behavior continues. The manager could listen to a recording of the interview if one exists. Or, the manager could perform call backs on the suspect values from the interviewer.

While we're trying to prevent curbstoning, it is important to note that this isn't the only reason an interviewer may have extreme proportions. It is possible that an interviewer could have misunderstood part of the interviewing process and just needs to be trained again on the correct method.

Remember that first-digit analysis only provides a list of unusual interviewer outcomes and groups of establishments with potentially fraudulent data. Since first-digit analysis only gives us unusual collector outcomes we wanted to limit false positives. The method provided here is conservative by reducing the number of false positives at every step of our research. We were conservative in our approach to give the all benefits to the interviewer before being flagged for further review. This was accomplished by the following:

- Compared interviewer to profile samples of similar workloads.
- Bootstrap with more than the standard 2,000 samples.
- Used Bonferonni correction for the Rmax test instead of less conservative tests.
- Did not account for negative correlation between first-digit proportions.
- Distance from the median ranked conservatively.

References

Dumas, C. F., & Devine, J. H. (2000). Detecting Evidence of Non-Compliance In Self-Reported Pollution Emissions Data: An Application of Benford's Law. *American Agricultural Economics Association Annual Meeting.* Tampa.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. *Journal of Forensic Accounting*, 5:17-34.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Kabacoff, R. (2015). *R in Action: Data analysis and graphics with R 2nd Edition.* Manning Publications, New York.

Koczela, S., Furlong, C., McCarthy, J., and Mushtaq, A. (2015). Curbstoning and beyond: Confronting data fabrication in survey research. *Statistical Journal of the IAOS*, 31:413-422.

Jörg-Peter Schräpler (2010). Benford's Law as an instrument for fraud detection in surveys using the data of the Socio-Economic Panel (SOEP). *SOEPpapers on Multidisciplinary Panel Data Research*

Nigrini, M. (1999). I've got your number. *Journal of Accountancy,* 187:79-83.

Ross, S. (1997). *Simulation 2nd Edition.* Academic Press, San Diego.

Smith, C. (2002). Detecting Anomalies in Your Data Using Benford's Law. *SUGI 27 Proceedings: Statistics and Data Analysis,* Paper 249.

StasK (2012). Why the data should be resampled under null hypothesis in bootstrap hypothesis testing. Retrieved from http://stats.stackexchange.com/questions/41683/

Swanson, D., Cho, M.J., Eltinge, J., (2003) "Detecting Possibly Fraudulent or Error-Prone Survey Data Using Benford's Law," *Proceedings of the Section on Survey Research Methods, American Statistical Association*

Appendix A: Foundations of the Bootstrap in the profile sample context.

The underlying true distribution F, a common distribution function, produced a sample of employment values at hand. F is broken into *I* strata $X_1, X_2, \dots, X_i, \dots, X_I$. Each strata is comprised of values that were productively collected in the sample, *Y*, and values that were not collected in the sample, *Z*. The number of elements in individual strata, *i*, differ and are denoted, n_i.

$$F = \begin{bmatrix} X_1 \left[\left[Y_1^1, Y_2^1, \dots, Y_{n_1}^1\right], \left[Z^1\right]\right] \\ X_2 \left[\left[Y_1^2, Y_2^2, \dots, Y_{n_2}^2\right], \left[Z^2\right]\right] \\ \dots \\ X_i \left[\left[Y_1^i, Y_2^i, \dots, Y_{n_i}^i\right], \left[Z^i\right]\right] \\ \dots \\ X_I \left[\left[Y_1^I, Y_2^1, \dots, Y_{n_I}^1\right], \left[Z^I\right]\right] \end{bmatrix}$$

The sample space for each strata are naturally reduced by the uncollected or unknown values, Z^i. CES collects values from the underlying true distribution F, creating, F_n, representing the subset of F for which we have collected values. For stratum *i*, $Y_1^i, Y_2^i, \dots, Y_{n_i}^i$ represents the employment values obtained from all of the interviewers. The idea here is that each stratum has its own underlying distribution and F_n is a mixture of those distributions.

$$F_n = \begin{bmatrix} X_1\left[Y_1^1, Y_2^1, \dots, Y_{n_1}^1\right] \\ X_2\left[Y_1^2, Y_2^2, \dots, Y_{n_2}^2\right] \\ \dots \\ X_i\left[Y_1^i, Y_2^i, \dots, Y_{n_i}^i\right] \\ \dots \\ X_I\left[Y_1^I, Y_2^I, \dots, Y_{n_i}^I\right] \end{bmatrix}$$

Since cases are assigned randomly within strata, an interviewer will have a subset of employment values from some of the strata, *J*, where $J \subseteq I$. An interviewer's randomly assigned case load is generally not large enough to have all strata represented and diverse enough to not comprise a majority of any stratum. The underlying interviewer's distribution, F_m, is the interviewer's subset of F_n. Thus, F_m is a mixture of distributions for the strata represented.

$$F_m = \begin{bmatrix} X_1\left[Y_1^1, Y_2^1, \dots, Y_{m_1}^1\right] \\ X_2\left[Y_1^2, Y_2^2, \dots, Y_{m_2}^2\right] \\ \dots \\ X_j\left[Y_1^j, Y_2^j, \dots, Y_{m_j}^j\right] \\ \dots \\ X_J\left[Y_1^J, Y_2^J, \dots, Y_{m_J}^J\right] \end{bmatrix}$$

The standard nonparametric Bootstrap with replacement is used to make statements about the sampling distribution for an individual interviewer based on a known distribution. The Bootstrap is denoted as F^*. F^* for the interviewer is known because it is generated using a modified sampling protocol based on an interviewer's collected caseload. For example, in stratum *i*, the interviewer has m_i elements. For each element of F^*, stratum *i* would also have m_i elements selected with replacement from the n_i elements in F_n. The sampling protocol is to select multiple profile samples with replacement using the same mixture each time (StasK, 2012).

$$F^{*} = \begin{bmatrix} \begin{bmatrix} X_1[sample\ interviewer's\ size\ m_1] \\ X_2[sample\ interviewer's\ size\ m_2] \\ \ldots \\ X_I[sample\ interviewer's\ size\ m_I] \end{bmatrix}, \\ \begin{bmatrix} X_1[sample\ interviewer's\ size\ m_1] \\ X_2[sample\ interviewer's\ size\ m_2] \\ \ldots \\ X_I[sample\ interviewer's\ size\ m_I] \end{bmatrix}, \\ \ldots \end{bmatrix}$$

Below is a simplified example for an interviewer who collected values for strata 1 and 2. In this example, the first two strata had 7 values collected from all interviewers, and the third stratum had 11. The interviewer of interest collected m_1=2 employment values in the first stratum, and m_2=3 values in the second stratum. No values were collected for the third stratum, m_3 = 0. The Bootstrap of F_m is denoted as F^* and has B profile samples of each strata using the same sample sizes as the original interviewer's values. In this example, B=3 and we collect

$$F_n = \begin{bmatrix} X_1[381, 159, 364, 135, 44, 351, 227] \\ X_2[74, 117, 85, 6, 237, 270, 119] \\ X_3[95, 289, 282, 35, 393, 353, 218, 95, 255, 63, 128] \end{bmatrix}$$
$$F_m = \begin{bmatrix} X_1[44, 364] \\ X_2[270, 117, 85] \end{bmatrix}$$
$$F^{*} = \left[\begin{bmatrix} X_1[159, 227] \\ X_2[237, 270, 74] \end{bmatrix}, \begin{bmatrix} X_1[351, 135] \\ X_2[74, 270, 270] \end{bmatrix}, \begin{bmatrix} X_1[44, 44] \\ X_2[6, 119, 85] \end{bmatrix} \right]$$

The statistic $T(F_m)$ is the vector of first-digit proportions for the individual interviewer as shown below.

$$T(F_m) = \widehat{\theta_m} = \{\widehat{\vartheta_1^m}, \widehat{\vartheta_2^m}, \ldots, \widehat{\vartheta_9^m}\}$$
$$= \left[\frac{\sum_m I(first\ digit_m = 1)}{m}, \frac{\sum_m I(first\ digit_m = 2)}{m}, \ldots, \frac{\sum_m I(first\ digit_m = 9)}{m}, \right]$$
$$where\ I(\cdot) is\ the\ indicator\ function \begin{cases} I(TRUE) = 1 \\ I(FALSE) = 0 \end{cases}$$

To make statements about the sampling distribution, $T(F_n)$ based on a known distribution $T(F^*)$, a distribution for the first-digit proportions is determined by drawing B profile samples using the sampling protocol.

$$T(F^{*}) = \begin{matrix} \widehat{\vartheta_1^{*,1}} & \widehat{\vartheta_2^{*,1}} & \ldots & \widehat{\vartheta_9^{*,1}} \\ \widehat{\vartheta_1^{*,2}} & \widehat{\vartheta_2^{*,2}} & \ldots & \widehat{\vartheta_9^{*,2}} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\vartheta_1^{*,b}} & \widehat{\vartheta_2^{*,b}} & \ldots & \widehat{\vartheta_9^{*,b}} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\vartheta_1^{*,B}} & \widehat{\vartheta_2^{*,B}} & \ldots & \widehat{\vartheta_9^{*,B}} \end{matrix}$$

T(F*) is represented above as a matrix with B rows and 9 columns. By construction, the rows sum to 1 and the columns can be used to estimate probability density functions for each of the 9 first-digit proportions. For example, in the first column $\widehat{\vartheta_1^*} = \left\{\widehat{\vartheta_1^{*,1}}, \widehat{\vartheta_1^{*,2}}, \widehat{\vartheta_1^{*,3}}, \dots, \widehat{\vartheta_1^{*,b}}, \dots, \widehat{\vartheta_1^{*,B}}\right\}$ is used to estimate the probability density function of first-digit proportions of ones. Thus, $pdf(\widehat{\theta^*})$ is a collection of estimated probability density functions.

The interviewer's actual first-digit proportions, $\widehat{\theta_m} = \left\{\widehat{\vartheta_1^m}, \widehat{\vartheta_2^m}, \dots, \widehat{\vartheta_9^m}\right\}$, are then tested against the Bootstrapped probability density functions using $pdf(\widehat{\theta^*}) = \left\{pdf(\widehat{\vartheta_1^*}), pdf(\widehat{\vartheta_2^*}), \dots, pdf(\widehat{\vartheta_9^*})\right\}$. From a large sample, B > 1,000, first-digit vectors can be used in a variety of test statistics such as the mean, variance, and percentile for each of the nine parameters in the first-digit vectors. The interest here is in determining the percentile for $pdf(\widehat{\theta^*})$ to be utilized as a Bootstrapped percentile.

The Bootstrapped parameters, the first-digit proportions from each profile sample, can be used to determine an approximation of the probability for the individual interviewer obtaining a value from our generated first-digit distribution being more extreme than that actually observed (Ross, 1997).
If $\widehat{\vartheta_J^m} < median(\widehat{\vartheta_J^*})$, then Bootstrapped p-value =

$$Prob\left(\widehat{\vartheta_J^m} < \widehat{\vartheta_J^*}\right) = \frac{number\ of\ \widehat{\vartheta_J^{*,b}} < \widehat{\vartheta_J^m}}{B}.$$

If $\widehat{\vartheta_J^m} \geq median(\widehat{\vartheta_J^*})$, then Bootstrapped p-value =

$$Prob\left(\widehat{\vartheta_J^m} \geq \widehat{\vartheta_J^*}\right) = 1 - \left(\frac{number\ of\ \widehat{\vartheta_J^{*,b}} < \widehat{\vartheta_J^m}}{B}\right).$$

For *j = {1, 2, 3, … , 9} and b = {1, 2, 3, … , B}.*

*Details for how we handled the individual values in relation to the median or equivalent Bootstrapped proportions are in the main paper.