International Journal of Economics and Finance; Vol. 10, No. 6; 2018
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

Full Range Testing of the Small Size Effect Bias for Benford Screening: A Note

Yan Bao[1], Frank Heilig[2], Chuo-Hsuan Lee[3] & Edward J. Lusk[3]

[1] Frostburg State University, 101 Braddock Road, Frostburg, MD, USA

[2] Senior Risk Manager Volkswagen Financial Services AG, Braunschweig, Germany

[3] The State University of New York (SUNY) at Plattsburgh, 101 Broad St., Plattsburgh, NY, USA

Correspondence: Edward J. Lusk, The State University of New York (SUNY) at Plattsburgh, 101 Broad St., Plattsburgh, NY, 12901: USA.

Received: March 21, 2018 Accepted: April 19, 2018 Online Published: April 30, 2018

doi:10.5539/ijef.v10n6p47 URL: https://doi.org/10.5539/ijef.v10n6p47

Abstract

Bao, Lee, Heilig, and Lusk (2018) have documented and illustrated the Small Sample Size bias in Benford Screening of datasets for *Non-Conformity*. However, their sampling plan tested only a few random sample-bundles from a core set of data that were clearly *Conforming* to the Benford first digit profile. We extended their study using the same core datasets and DSS, called the Newcomb Benford Decision Support Systems Profiler [NBDSSP], to create an expanded set of random samples from their core sample. Specifically, we took repeated random samples in blocks of 10 down to 5% from their core-set of data in increments of 5% and finished with a random sample of 1%, 0.5% & 20 thus creating 221 sample-bundles. This arm focuses on the False Positive Signaling Error [FPSE]—i.e., believing that the sampled dataset is *Non-Conforming* when it, in fact, comes from a *Conforming* set of data. The second arm used the Hill Lottery dataset, argued and tested as *Non-Conforming*; we will use the same iteration model noted above to create a test of the False Negative Signaling Error [FNSE]—i.e., if for the sampled datasets the NBDSSP fails to detect *Non-Conformity*—to wit believing incorrectly that the dataset is *Conforming*. We find that there is a dramatic point in the sliding sampling scale at about 120 sampled points where the FPSE first appears—i.e., where the state of nature: *Conforming* incorrectly is flagged as *Non-Conforming*. Further, we find it is very unlikely that the FNSE manifests itself for the Hill dataset. This demonstrated clearly that small datasets are indeed likely to create the FPSE, and there should be little concern that Hill-type of datasets will not be indicated as *Non-Conforming*. We offer a discussion of these results with implications for audits in the Big-Data context where the audit In-charge may find it necessary to partition the datasets of the client.

Keywords: false positive & negative screening errors

1. Introduction

Bao, Lee, Heilig, and Lusk (2018) have reported on a study where 16 datasets were randomly selected from Balance Sheets for firms traded on The China Stock Market & Accounting Research (CSMAR™) Database: China Stock Market Financial Statements Database. All of these core datasets were larger than 5 000 data points and all were found to be *Conforming* to the Newcomb (1880) & Benford (1938) practical first digit profile as formed by Lusk and Halperin (2014a). They then randomly sampled 10% and also 250 observations from each of the 16 datasets repeatedly for ten times to create two sample sets: Dataset-A [10%DS]: 160 samples each of which had on the order of 1 500 data points, and Dataset-B [250DS]: 160 samples each of which had 250 observations. All were analyzed with the Newcomb-Benford Decision Support System Profiler [NBDSSP (Note 1)].

In the results section, they report relative to Extended Procedures [EP] investigations:

> “- - - *the 10%DS sample produced 9 BSFs over the 160 trials. In the column z-calc are the two-tailed test of proportions between the 10%DS and the 250DS sampling results. - - - for the 10%DS there are no instances with more than two Bedford Screening Flags created by the NBDSSP for a particular dataset, suggesting no need for an EP investigation. For the 250DS over the 113 BSFs there were 13 instances with more than two BSFs for a particular dataset suggesting that the EP*

> *investigations may be warranted.*"

Continuing, they note:

> "*Referencing the above theoretical and empirical results and from the robust testing protocols that we have used, there is consistent and strongly persuasive evidence that as the auditor moves to small sample sizes by partitioning larger Conforming-in-nature datasets that would not suggest the use of EP in the audit context, these sub-samples test to be Non-Conforming owing only to their small sample size. In this case, the auditor invites creating, as an artifact of the same sample size, Benford Screening indications that would incorrectly suggest EP testing.*"

1.1 Summary

Note that their test is the test of the False Positive Signaling Error [FPSE] as these accrued datasets were all *Conforming* in nature. As the sample size is reduced, at 250 sample-points, there seems to be evidence of the small sample size effect compared to the testing at the 10% DS-arm in that the NBDSSP indicates that there are reasons to believe that some of the datasets are *Non-Conforming*. This is an error in the Type-1 decision domain: where a *Conforming* dataset is incorrectly flagged as *Non-Conforming*. We call this a FPSE.

1.2 Prêcis of the Note

Considering the important implications of the work reported above, albeit a preliminary and rather limited testing protocol, we have taken up an inquiry begged by the study of Bao, Lee, Heilig, and Lusk (2018) [BLHL]. This is the point of departure of our extension.

Following, for this research note, we will:

1.) Use the BLHL core-accrual dataset to extend the testing of the False Positive Signaling Error [FPSE] domain to better determine the sampling-point neighborhood of the boundary point of the FPSE.

2.) Take up a test of the point at which the False Negative Screening Error [FNSE] establishes an inferential marker.

3.) Discuss the implications of the FPSE and FNSE information for making the EP testing decision for the certified audit.

2. Testing Protocols for the FPSE & the FNSE

2.1 FPSE Screening

Our testing protocol for FPSE starts with the 100% accrual of the same 16 datasets that were used by BLHL. Then, we create a sampling cascade [of a block of 10 samples] using a reduction increment of 5% of the base-line 100% accrual. Finally, we take a sample set of: [1%, 0.5% & 20 data points]. Accordingly, we will produce 221 samples [1+ (10x19) + (10x3)] for testing for each of the 16 datasets or 3 536 [221 × 16] Samples in total. The information concerning our samples is as follows:

Table 1. Sample accrual profiles

Sample Arm	Number of Samples	Average of Data-Points
100% Accrual	1 Sample	12 092
95%, 90%, - - -, 5%	19 × 10 = 190	6 046
1%, 0.5% & 20 Data-Points	3 × 10 = 30	67
Summary	221 Samples	3 536 [221 × 16]

We decided upon a sample of 10 random sample replications for each of the accrual blocks as a preliminary sampling study indicated that there was no difference at an overall Type-1 testing [p-value <0.05] for FPSEs among blocks of: 10 observations, 50 observations, or 100 observations using the ANOVA, n=3 tested over a random sample of eight datasets. Thus, using the 10 Replication-Arm is conservative in that it is less powerful and so is relatively biased to favor the Null, thereby producing less indications or EP-flags.

2.2 FNSE Screening

Our testing protocol for FNSE Screening starts with the Hill recommended Lottery dataset where theoretically each of the first digits occurs with the same frequency: (1/9)%. See the works of Hill for a comprehensive treatment and proof of the logic of Benford profiling protocols. Hill (1995,a,b, 1996 & 1998). Using the same sampling procedure as what has been discussed in FPSE Screening, we will produce 221 sample sets based on the Hill dataset. Our interest is to determine if there are points in the accrual cascade where the NBDSSP indicates that the Hill-Lottery dataset is likely to be *Conforming* and so would be a FNSE.

3. Overall Testing Results

3.1 The False Positive Signaling Error: Incorrectly Believing that the Sampled Dataset is Non-Conforming

Here we are testing how the sample size influences the FPSE reporting of the NBDSSP where the NBDSSP leads the investigator incorrectly to believe that the sampled dataset is *Non-Conforming* while the dataset is actually *Conforming*. We start with the same 16 datasets that were used by BLHL. Recall that BLHL report that at the sample size of 250 there were 13 instances where there were more than two-EP flags produced by the NBDSSP, suggesting that EP may be warranted. This is an average of 8.13% over the 160 trials. For the current FPSE testing protocol we find the following profile as presented in Figure 1.

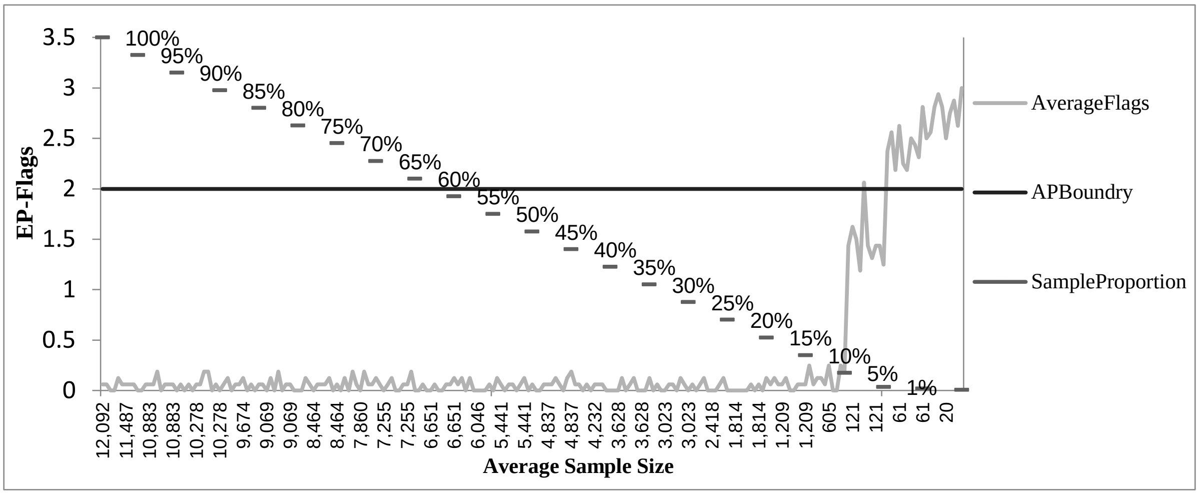

Figure 1. Profile of the average EP-Flags for the sample proportion against the Benchmark of 2.0

Figure 1 clearly demonstrates the FPSE jeopardy. At about tracking iteration point 182, the average number of EP-Flags begins to increase. For Points [182 to 191] where the average sample size is 605, the ordered interval for the number of EP-Flags is [0 to 4]; for these there were three (3) instances where EPs would have been indicated. For Points in the next sample bundle [192 to 201] where the average sample size is 121, the number of EP-Flags is in the interval [20 to 26].

A more germane graphic is the percentage of time that a FPSE occurs over the iterated sample space. This is presented in Figure 2.

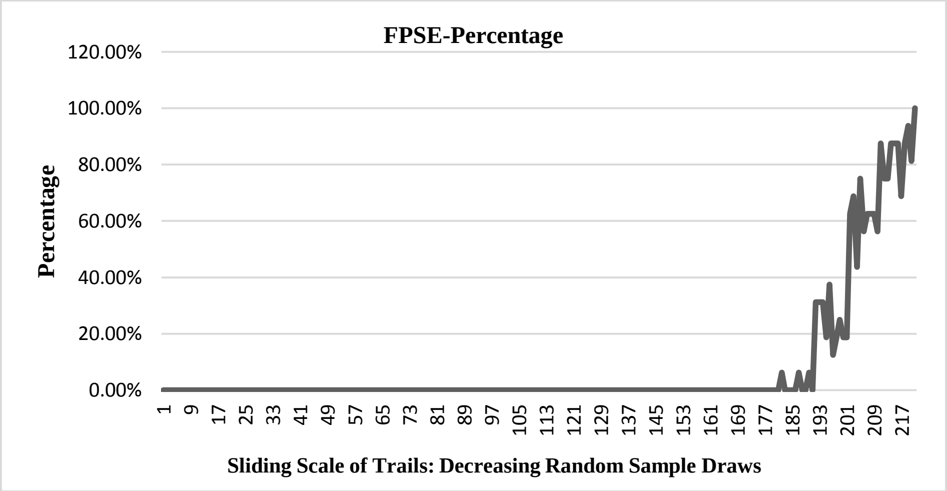

Figure 2. Percentage of time that an EP investigation is suggested incorrectly

Here we see the clear increase in the FPSE that is created by the various sample sizes. Consider the statistical test of these percentages over the various sample blocks. This is presented in Table 2 following:

Table 2. Frontier boundary for the incidence of the FPSE

	Stage n-4	**Stage n-3**	**Stage n-2**	**Stage n-1**	**Stage n**
FPSE Percent	0.00%	1.88%	24.35%	63.75%	84.38%
Sample Size	1 120	605	121	61	20
p-value	N/A	0.04	<0.001	<0.001	<0.002

We see the clear inferential impact of reducing the sample size. All of the sample sizes created from 95% down to 20 points were tested in two contiguous blocks. The first time that an "interesting" FPSE occurs is for the sample size of on average of 605. All previous contiguous blocks, where the sample sizes were greater than (>) 1 120, test to have p-values that were all greater than (>) 0.9 clearly supporting the Null of no difference. In Table 2, we see the last five sampling blocks of ten in the iterated profile. The most dramatic increase starts to occur as the sample size falls from 605 to 121 and the FPSE is 24.35% as shaded in Table 2. Experience suggests that conducting an EP examination of an account 25% of the time when it is not warranted is inconsistent with careful control of the resources of the audit. Therefore, we make this point ≈ 120 as the ***Alert Sampling Point Frontier***. Sample sizes at around 120 and lower seem to invite the FPSE and so should be avoided (Note 2).

3.2 The False Negative Signaling Error: Incorrectly Believing that the Sampled Dataset is Conforming

Now, we are considering the interaction between the sample size and the FNSE of the NBDSSP where the NBDSSP leads the investigator incorrectly to believe that the sampled dataset is *Conforming* while the dataset is actually *Non-Conforming*.

Here the work of Heilig and Lusk (2017) is relevant to this test arm. They report that:

> *We formed a Lottery Profile of n=999 where each of the 9 first digits occurred exactly 111 times. The NBDSSP produced the following indications:*
>
> 1) ***NBPP**: All nine of the first digits produced BSFs—i.e., NOT in the screening interval; thus the Lottery dataset is labeled as: Non-Conforming.*
>
> 2) ***Chi2**: Eight of the first digits produced BSFs; Non-Conforming*
>
> 3) ***Nigrini**: The sixth flag occurred for a sample size of 127 < 1,825; Non-Conforming*
>
> 4) ***Distance Measure** was 0.0971> 0.02638; Non-Conforming*
>
> *As expected, these results are a strong indication of the screening acuity of the NBDSSP for the Lottery dataset. We also re-ran the Lottery dataset 100 times and not one time did the NBDSSP not detect that the Lottery dataset was Non-Conforming.*

For the test of the FNSE, we created a dataset that had 12 096 data points. We arrived at this number as the average for the FPSE test was 12 092 data points. The dataset with 12 096 was the closest to the FNSE test dataset and this FNSE dataset had 1 344 blocks of the first nine digits—i.e., all of the digits occurred with the Hill lottery frequency of [1/9]%. We then cascaded the sampling plan in the same manner that we did for the FPSE to arrive at 221 samples.

According to the report of Heilig and Lusk (2017), when the NBDSSP is used to test the Hill dataset, FNSEs will occur rarely. Therefore, our testing result here can be used to verify the acuity of the NBDSSP in relation to the FNSE, especially as we narrow down the sample sizes to the range of 250 to 440 which are the range intervals for the low-end test by BLHL and the high-end tested by Lusk and Halperin **(**2014b**)**. If the Hill dataset were to have been scored as *Confirming* at a high frequency for smaller sample sizes then this would call into question the acuity of the NBDSSP relative to the FNSE.

3.3 Results: FNSE Testing

We found that out of the 221 samples drawn from the *Non-Conforming* Hill Lottery dataset there was only one (1) instance indicating that the sample partition was signaled as *Conforming* i.e.,—the NBDSSP failed to create more than two EP-screening flags thus indicating that EPs may be warranted. This occurred, not surprisingly, during the latter stage sub-partition when the sample size was reduced to 20. This is clear *prima-fascia* evidence that supports the acuity of the NBDSSP and also provides important information regarding the FNSE incidence as sub-partitions are selected by the auditor. However, to provide statistically valid inferential support for the above observation, we selected an *a-priori* expectation of 5% for the FNSE meaning that if the samples from the *Non-Conforming* Hill Lottery dataset were to be scored as *Conforming* at a frequency of 5% or more, the acuity of the NBDSSP screening would be called into question. For the inferential testing, we tested the directional difference between this *a-priori* expectation of 5% and the FNSE realization of: [1/221]% using the conservative form of the standard error. We find the z-calculated to be 10.1, for which the p-value is < 0.001. Furthermore, given the *a-priori* expectation of 5% and the Null test-against of 0.4525%, the Power of the inferential result was 99.6%—that is to say, the FNSE results are reliable. Taken together, these inferential results strongly suggest that the population incidence of the FNSE is less than 5%. *Impact*. Respecting the practical inference for the FNSE test, the evidence supports the notion that auditors can rely on the acuity of the NBDSSP for detecting the *Non-Conforming* Hill Lottery dataset.

4. Summary and Conclusions

4.1 Summary

In this case, we find, as expected, that there is a real jeopardy for inviting the FPSE in selecting sub-samples from large audit datasets that are likely to be *Conforming* in nature. Recall that the reason that one employs Newcomb-Benford screens is to rationalize the decision to launch an Extended Procedure [EP] test of a dataset that profiles as *Non-Conforming*. In this case, at around a sub-sample partition of 120 sample units or lower, the NBDSSP indicates an EP may be warranted but this is a false signal effectively due to the small sample size. Interestingly, the jeopardy is not symmetrical in that if the dataset is manifestly *Non-Conforming* then there is not a reasonable probability that a small sample size will give the erroneous impression that the sub-sample is *Conforming* and so does not warrant EP testing. Further, for the first time a comprehensive test of the FNSE is offered. We have learned that when there is a strong likelihood that the dataset is *Non-Conforming* then partitioning is very unlikely to create the impression that the dataset is *Conforming*. These FNSE results should allay concerns that the auditor will fail to use EP when in fact they would be warranted.

4.2 Conclusion

This research report confirms the previous result of BLHL and using the larger sample testing frame gives more confidence in their advice that the audit In-charge should be cautious in partitioning account datasets. Our results reinforce their advice that the In-charge would do well to use a Bayes-conditional that IF the downloaded dataset is *Conforming* in nature that *Non-Conforming* results for sub-set partitions should not be used as a rational to conduct extended procedure testing. In this instance more related audit evidence may be required to form a logical EP testing protocol. Additionally, for the FNSE, we extensively tested the Hill configuration as this is the proto-typical case of a *Non-Conforming* dataset. It is necessary, however, to expand the testing of various instances of *Non-Conforming* datasets. That is to say, research needs to move away from Hill case, as it may be so extreme so as to not provide a realistic and reliable test for “boundary accrual anomalies” where the *Non-Conforming* dataset partitions are not-flagged as a *Non-Conforming*—a FNSE. More testing for the FNSE will complete the testing results that we reported for the FPSE and so aid in the organization of the scarce audit resources so as to more effectively and efficiently conduct the certified audit.

Acknowledgments

Thanks and appreciation are due to: Dr. H. Wright, Boston University: Department of Mathematics and Statistics, the participants at the SBE Research Workshop at SUNY: Plattsburgh, and Mr. Manuel Bern, Chief of Internal Audit: TUI International, GmbH, Hannover, Germany for their careful reading, helpful comments, and suggestions.

References

Bao, Y., Lee, C. H., Heilig, F., & Lusk, E. (2018). Empirical information on the small size effect bias relative to the false positive rejection error for Benford test-screening. *International Journal of Economics and Finance, 10*(2), 1-9. https://doi.org/10.5539/ijef.v10n2p1

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society, 78*, 551-572.

Heilig, F., & Lusk, E. (2017). A robust Newcomb-Benford account screening profiler: An audit decision support system. *International Journal of Financial Research, 8*, 27-39. https://doi.org/10.5430/ijfr.v8n3p27

Hill, T. (1995a). The significant-digit phenomenon. *American Mathematical Monthly, 102*, 322-327. https://doi.org/10.2307/2974952

Hill, T. (1995b). Base-invariance implies Benford’s law. *Proceedings of the American Mathematical Society,* *123*, 887-895. https://doi.org/10.1090/S0002-9939-1995-1233974-8

Hill, T. (1996). A statistical derivation of the significant-digit law. *Statistical Science, 10*, 354-363. https://doi.org/10.1511/1998.31.815

Hill, T. (1998). The first digit phenomenon: A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data. *American Scientist, 86*, 358-363.

Lusk, E., & Halperin, M. (2014a). Using the Benford datasets and the Reddy & Sebastin results to form an audit alert screening heuristic: A Note. *IUP Journal of Accounting Research and Audit Practices, 8*, 56-69.

Lusk, E., & Halperin, M. (2014b). Detecting Newcomb-Benford digital frequency anomalies in the audit context: Suggested Chi2 Test Possibilities. *Journal of Accounting and Finance Research, 3*, 191-205. https://doi.org/10.5430/afr.v3n2p191

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, *4*, 39-40. https://doi.org/10.2307/2369148

Notes

Note 1. The NBDSSP is due to Heilig and Lusk (2017) and has four tests for the first digit profile of a dataset: ***The Newcomb-Benford Practical Profile [NBPP]; The Chi2 Screening Platform; The Nigrini Test of Proportions Platform*** & ***Cartesian Distance Measure***. The NBDSSP has been vetted that if more than two of the four platforms create an NBDSSP-indication that the dataset is *Non-Conforming*, then the dataset under audit screening is considered as *Non-Conforming;* otherwise it is considered *Conforming.*

Note 2. This fits well with the results of BLHL where at a sample size of 250 the FPSE was 8.13% [13/160].

Copyrights