International Journal of Economics and Finance; Vol. 10, No. 2; 2018
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

Empirical Information on the Small Size Effect Bias Relative to the False Positive Rejection Error for Benford Test-Screening

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

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

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

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

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

Received: January 19, 2017 Accepted: February 14, 2017 Online Published: December 30, 2017

doi:10.5539/ijef.v10n2p1 URL: https://doi.org/10.5539/ijef.v10n2p1

Abstract

Due to the theoretical work of Hill Benford digital profile testing is now a staple in screening data for forensic investigations and audit examinations. Prior empirical literature indicates that Benford testing when applied to a large *Benford Conforming* Dataset often produces a bias called the FPE Screening Signal [FPESS] that misleads investigators into believing that the dataset is *Non-Conforming* in nature. Interestingly, the same FPESS can also be observed when investigators partition large datasets into smaller datasets to address a variety of auditing questions. In this study, we fill the empirical gap in the literature by investigating the sensitivity of the FPESS to partitioned datasets. We randomly selected 16 balance-sheet datasets from: *China Stock Market Financial Statements Database*™, that tested to be Benford *Conforming* noted as RBCD. We then explore how partitioning these datasets affects the FPESS by repeated randomly sampling: first 10% of the RBCD and then selecting 250 observations from the RBCD. This created two partitioned groups of 160 datasets each. The Statistical profile observed was: For the RBCD there were no indications of *Non-Conformity*; for the *10%-Sample* there were no overall indications that Extended Procedures would be warranted; and for the *250-Sample* there were a number of indications that the dataset was *Non-Conforming*. This demonstrated clearly that small datasets are indeed likely to create the FPESS. We offer a discussion of these results with implications for audits in the Big-Data context where the audit In-charge would find it necessary to partition the datasets of the client.

Keywords: extended procedures, audit risk, partitioning large datasets

1. Introduction

Benford profiling is now one of the mainstays of forensic as well as audit investigations in the Big Data world. This is particularly the case for audits of firms traded on exchanges, termed PCAOB audits; such firms usually have very large datasets. There are several reasons why Benford profiling has achieved its justifiable place in the auditor's e-panoply:

1) There are now a plethora of commercially available software, the modules of which, have Benford screening platforms, such as: IDEA[<https://www.audimation.com/Product-Detail/CaseWare-IDEA>] and DATAS [http://www.nigrini.com/datas_software.htm],
2) Most data of PCAOB audits can be captured electronically and so downloads of the client's Accounting Information System [AIS] is possible in a few minutes,
3) The investigative utility and acuity of Benford profiles have been detailed in the literature for decades and has moved from the academic context to the practice protocols of audit LLPs, e.g., Collins (2017), and
4) Due to Hill (1995a,b, 1996 & 1998) there is a formal theoretical proof of the expectation that datasets in the neighborhood of the Log_{10} mathematical model of Newcomb (1881) and Benford (1938) [NB] are rationally to be expected under conditions that often characterize the data-generating functions of the AIS data of PCAOB firms.

To have a common context for Benford profiling, we offer, *en bref*, comments on Benford Screening and the Decision Support Systems [DSS]-Software used in this study.

1.1 Overview of Benford Screening

A simple example will suffice to elucidate and rationalize the concepts that underlie digit dataset profiling, or what we will term Newcomb-Benford Screening. Assume that as the In-Charge of the *Walmart Corporation* audit you are testing the "reasonability of various datasets of retail prices" using 35 datasets drawn randomly from *Walmart* stores in the North East of the USA. Your expectation is: *The monetary prices of this data would follow the pricing model:* \$#.99. Your a-*priori* expectation is that 95% of the time the last digit in the cents place of the download would be "9". If the results of your digital profile screening was: Data downloads from 33 stores had a "9" as the last digit from 93% to 97.5% of the time, then you would likely accept these datasets as consistent with your expectation, therefore justifying their use in the execution of the audit. However, if for two of the datasets the percentages of "9s" were 10% and 12%, such a marked divergence of about 85% from the expectation of [93%−97.5%] would likely warrant inquiries of the CFO of these two *Walmart* stores. Assume that the auditor learned that these two stores had conducted customer focus groups, and that customers were more likely to purchase goods if they did not follow the standard pricing formula of \$#.99. Further, based upon this focused group feedback the CFOs of these two stores used random numbers to fill the digits in the cents place. In this case, there would have been a rational explanation for the divergence from the *a-priori* expectation, and so, further testing would not be warranted. Extracting the logical extension from this *Walmart* example, we suggest that digital profiles benchmarked by empirically valid experiential auditor judgments can be a valuable initial screen to make sure that the data downloaded are capable of providing a reasonable basis of justifying the audit opinion. Thus, the curious observation of Newcomb (1881) and later of Benford (1938) that the first digits of economic data follow a profile that has a stable set of expectations, then opens the door to using the NB expectation or its derivatives as a reasonable screen for the data in the audit context. Consider now the DSS that we will use in generating the inferential information for this study.

1.2 The Benford DSS Profiler

The theoretical work of Hill (1995a,b,1996 & 1998) established that under the usual conditions the datasets reported on active market trading exchanges are expected to *Conform* to the Benford first digit profile. This theoretical result has been empirically tested and its veracity has been demonstrated by: Ley (1996), Nigrini (1996), Nigrini and Mittermaier (1997), Cho and Gaines (2008), Ross (2011), Lusk and Halperin (2014a,b,c), Lusk and Halperin (2015a,b). Using these theoretical and empirical results, Heilig and Lusk (2017) developed a robust Decision Support System for inferentially testing the NB-*Conformity* of datasets. Their model is termed: the Newcomb Benford Decision Support System Profiler [NBDSSP] and is used to determine if a dataset is likely to have a first digit profile that is in Nature *Conforming* to the Newcomb-Bedford first digit profile; or alternatively the dataset would be labeled as: *Non-Conforming*—i.e., not likely to have been drawn from the population of NB *Conforming* datasets. Following we will present in the overview that is consistent with the exposition, the four platforms in the NBDSSP that together create a summary signal that has inferential validity relative to making the decision as to the necessity to consider an extended procedures investigation for a particular dataset under audit investigation.

1.2.1 The First NBDSSP Platform

The Newcomb-Benford Practical Profile [NBPP] is an interval for the nine first digits that is formed around a corrected expectation offered by Cho and Gaines (2007) and detailed by Lusk and Halperin (2014a). If an observed digit is in this interval, this digit is considered *Conforming*; otherwise it is considered *Non-Conforming*. The terminology that we will use is: IF the particular digit is identified as *Non-Conforming* then we will note this as a Benford Screening Flag [BSF] as identified by the NBDSSP. If more than five [5] such digits produce BSFs, the dataset is labeled as *Non-Conforming*; otherwise it is considered *Conforming*. For more details on the NBPP screening interval see: Lusk and Halperin (2014a).

1.2.2 The Second NBDSSP Platform

The Chi2 Screening Platform uses the standard Chi2 Cell Contribution relative to the individual first digits. If this Chi2 Cell Contribution is greater than 1.0, a heuristic formed from the work of Tamhane and Dunlop (2000, p. 324), then the cell for that digit is judged as *Non-Conforming*. If more than 5 such digits produce BSFs, the dataset is labeled as *Non-Conforming*; otherwise it is considered *Conforming*. Also, as there is a real FPE jeopardy due to the sample size effect for the Chi2 model, (Cho & Gaines, 2008), the NBDSSP uses a random sample from the dataset under audit, where the Chi2 Cell bins are filled from a sample in the range of [330 to 440] as suggested by Lusk & Halperin (2014b) (Note 1).

1.2.3 The Third NBDSSP Platform

The Nigrini Test of Proportions Platform. As this test is also sensitive to large samples as precision becomes unrealistically narrow for the test interval thus inviting the FPE anomaly, this platform uses an incremental accrual cut-point of 1,825 observations (Lusk & Halperin, 2014c). If more sample points than 1,825 would be needed to drive the sixth Test of Proportions z-calculated to a value greater than 1.96, then the dataset is considered *Conforming*; otherwise it is considered *Non-Conforming.*

1.2.4 The Fourth NBDSSP Platform

Finally, there is a ***Cartesian Distance Measure*** that uses four independent distance measures relative to the deviation from the NBPP expectation set. The parameterization of this measure is that if, the average of these absolute valued distance measures is greater than 0.02638 [here rounded], the dataset is considered *Non-Conforming*; otherwise it is considered as *Conforming*. Lusk and Halperin (2015a).

1.3 Inferential Summary Point of Reference

The NBDSSP has been vetted that if more than two of the four platforms create BSFs—i.e., an NB-indication that the dataset is *Non-Conforming,* then the dataset under audit screening is considered as *Non-Conforming;* otherwise it is considered *Conforming.* See Heilig and Lusk (2017).

An Example will be useful to illustrate the dataset of information of the NBDSSP. Here we will use the Lottery dataset of Hill (1998) as this data profile manifestly and unarguably should fail the NB screening check and so be labeled as *Non-Conforming*. The Hill-Lottery profile is: All of the first digits would likely occur at the same rate as the Lottery numbers are picked randomly. Therefore, over time the distribution of all the first digits should converge to 11.1% [(1/9)%]. This is also a test of the False Negative Signaling Error [FNSE] of the NBDSSP as if the Lottery dataset were to be labeled as *Non-Conforming* this would be a FNSE. 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.*

As useful as Benford Screening has proven to be, there are a few conceptual issues that need to be considered in using Benford screens in launching Extended Procedures [EP] audit investigations. Investigations in the audit context are very expensive and so they need to be based upon a clear understanding of the FPES Jeopardy.

2. Précis of the Note

While the NBDSSP has been programmed considering the difficulties that large sample sizes create relative to the FPES jeopardy there has been little testing reported in the literature that addresses the bias or FPES jeopardy for NB-screening of small datasets. One must be attentive to the possibility that in the Benford screening context the usual partitioning of large datasets may also invite a FPES. This is a non-trivial audit issue, as it is very often the case that the auditor needs to partition a large dataset to address a particular issue in the audit. Given the above discussion of the effectiveness of digital profiling, the overview and vetting information relative to the NBDSSP, we will:

1) Take up for the first time relative to the published literature, the case of Small Datasets and their possible bias in inviting the FPES jeopardy,
2) Describe a testing protocol that provides vetting indications regarding the Small Sample bias in NB-Screening using the NBDSSP,
3) Present and discuss the inferential results of our testing,
4) Offer recommendations in the execution of audits so as to be sensitive to possible FPE-issues in partitioning datasets in the audit context, and finally
5) Suggest future testing of the Benford Screening Protocols.

3. Small Samples in the Audit Context: Possible NB-Screening Implications

3.1 Literature Review

Following are research reports that reinforce the concern that the auditor should have in using small datasets, the sort of which are not uncommon in the audit context, in making the decision to launch EP inquiries.

A study conducted by Wallace (2002) examined the effect of aggregation on NB-*Conformity*. In this study, a panel from 1995 through 1998 was collected for reported taxable sales. The first set of data contained 67 observations per year, and the combined incidence over the four years was 268. Wallace indicates [p. 22]: *The graph displays variation in the years and also illustrates that the expectation the conformance with Benford's Law, ceteris paribus, improves with a larger sample size.*

In a similar study, Lusk and Halperin (2015b) conducted a study using datasets from the CapitalCube™ market navigation platform. They selected various CaptialCube groups of firms and then selected various performance variables from the balance sheet and income statements. They first tested the individual variables usually on the order of 50 observations and then variable aggregations up to on the order of 250 observations. They report [p. 7], paraphrasing that: *The important recommendation that one may glean from these results is that aggregation of small correlated datasets of audit account variables, of on the order of 50 observations, to form a single aggregate of at least 250 observations or so will move from Non-Conformity to Conformity.*

Mir (2016) used Benford screens to examine the illicit financial flows (IFFs) from developing countries. He notes [p. 275] using the Chi2 analysis for the Benford screens that: *However, rejection of the null hypothesis becomes difficult if the number of observations in a data set is small.*

Further, he reports [p. 276 paraphrasing]: *We present the analysis for IFFs from least developed countries. As seen the width of the CI is inversely proportional to the size of the sample. Thus for small sample sizes the CI tend to be wide and produce less precise results.*

Durtschi, Hillison, and Pacini (2004, p. 26) examined fraudulent accounting data and report that there are: *two concerns, one intuitive and one statistical. First, intuitively, if there are only a few fraudulent transactions, a significant difference will not be triggered even if the total dollar amount is large. Second, statistically, if the account being tested has a large number of transactions, it will take a smaller proportion of inconsistent numbers to trigger a significant difference from expected than it would take if the account had fewer observations. This is why many prepackaged programs which include a Benford's law-based analytical test urge auditors to test the entire account rather than taking a sample from the account.*

Nigrini and Mittermaier (1997) treat the possibility of using the Benford screens in the audit context at the Analytical Procedures phase of the audit. By implication they emphasize that small datasets could be problematic in the use of Benford screens.

Although small datasets have been suggested as inherently biased in favor of rejecting *Conforming* datasets and suggesting that they are *Non-Conforming* more information is needed to examine this issue. Following, we offer specific test information focusing on small dataset configurations so as to have a reasonable judgment relative to the small dataset bias. In this regard, we will take the datasets that we will argue are *Conforming* in nature, and so we will here be focusing on the FPESS.

4. The Dataset Accrual and Selection Protocol

4.1 Operational Protocol

The issue in Benford studies is to find reasonable datasets for providing valid inferential results. For example, Lusk and Halperin (2014a) used data profiles from the initial Benford empirical work as well as many datasets that were argued in the peer-reviewed literature as *Conforming* or *Non-Conforming*. Also, Nigrini (1996), Hill (1995a,b, 1996 & 1998) and Ley (1996) offer context domains that are likely to be reasonable troves for locating appropriate datasets for testing various aspects of the NB-Screening protocols. Following on the research of Ley, we have decided to accrue datasets from a major active stock exchange as datasets from firms listed on such exchanges are:

1) Readily available as downloads,
2) Have certification assurance-validity as the data of firms listed on exchanges passed the professional scrutiny of the audit LLP, and
3) Usually created by myriad numbers of mathematical calculations, aggregations from subsidiary ledgers, and combinations from like rubric-category and thus fit well with the mixing imperative articulated by the Hill

research previously cited.

We downloaded the reported balance sheets at the fiscal year end from The China Stock Market & Accounting Research (CSMAR™) Database: China Stock Market Financial Statements Database. This database includes the financial statements for companies listed on Chinese stock markets including A and B shares since 1990. We then passed ALL of the datasets, n =128, from Balance Sheets of firms listed on Chinese Stock Markets through the NBDSSP. We selected ALL the datasets that were larger than 5,000 observations, n = 58. Here we screened each dataset using the *Filter Platform* in the *DataTab* of Excel to remove cells that contained Blanks or Zero Values, screening from the last data point to the first recorded data point. From this set, we selected all of the datasets that had NO BSFs among the four screening platforms as discussed above. This produced a sub-sample of n = 56. We then took a random sample of 16 datasets and this sub-set was used in the testing protocol; we note this partition: The Randomly Selected Benford *Conforming* Dataset [RB*C*D].

Further, to test the effect of samples we passed RB*C*D through the following two filters:

Filter A: We created an Excel VBA module to select ten (10) Random Samples (Note 2) of 10% of each of the sixteen RB*C*D noted as: ***10%DS***.

Filter B: We created an Excel VBA module to select ten (10) Random Samples of 250 for each of the sixteen RB*C*D noted as: ***250DS*** (Note 3).

We passed each of these filtered datasets through the same NBDSSP platforms used to process the initial CSMAR-download. This created 160 NBDSSP-screens for each of the two filters.

Finally, we recorded the processing profiles and, of course, the BSFs that were in evidence. This information was used to form the inferential tests for the small sample size effect for Benford screening.

4.2 A Priori Expectations: The Test Hypotheses

In this regard, we offer the following *a-priori* inferential testing expectations:

H1: For each of the NBDSSP Platforms, *excepting the Chi2 Platform as it is restricted to sample size*, we expect that there will be a strict order over the three trial groups in the direction: The RB*C*D will have fewer BSFs than the 10%DS which in turn will have fewer BSFs than the 250DS.

H2: For each of the NBDSSP Platforms, *excepting the Chi2 Platform as it is restricted to sample size*, we expect for an overall Bernoulli test, that there will be directional differences consistent with the small sample size expectation order of *H1*. The Bernoulli test is reasonable as there is no reason to expect association either between the elements of a testing block or over the testing blocks of the NBDSSP.

H3: For the *250DS, excepting the Chi2 Platform as it is restricted to sample size*, there will be NBDSSP BSFs indicating the need for extended procedure investigation considerations more frequently than for the *10%DS*. Recall that for an extended procedure investigation to be indicated, the NBDSSP must produce more than two BSFs for the particular dataset.

5. Results

5.1 Vetting the Random Sample

The first standard inferential test is for the integrity of the random sample—to wit: *Is there evidence that the four performance profiles of the random sample are different from the datasets not selected for testing purposes*. We had in total 56 RB*C*D, 16 as the random selection, on the order of 30%, of the total accrued testing datasets and 40 Non-Selected datasets. The two-tailed t-test assuming unequal variances was used for testing the differences in the performance profiles between the accrual dataset, n=16 and the non-selected datasets n=40. For the four NBDSSP platforms and the overall sample sizes of the datasets, the average two-tailed p-value for the Null test of no differences was 55.3% and none of the five p-values was less than 20%. This is strong evidence that the selected RB*C*D-dataset was not different from the non-selected RB*C*D-dataset and so rationalizes the integrity of the inference drawn from using this accrual selected dataset in the testing of the small sample effect.

5.2 Hypotheses Tests

For the Hypothesis tests we will first present Table 1.

Table 1. Data profiles of the four testing platforms

	Full Dataset	RBCD	Partitioned 10%Sample	Partitioned 250Sample
Data Sets	56	16	106	106
Sample Size	19,209	15,321	1,532	250
NBPP	1.39	≈1.38	<1.91	<3.71
Chi2	3.25	2.88	3.27	2.81
Nigrini	18,809	≈20,219>	11,487>	5,300
Distance	0.0121	≈0.0123	<0.0155	<0.251

The codex for Table 1 is:

The values are the means of the four platforms: ***NBPP***, ***Chi2***, ***Nigrini*** & ***Distance***. The tests that we will use for the Null of no difference are t-tests for unequal variances. This was needed as the Welsh test indicated in many of the comparisons that the assumption of equal variance could be rejected with high confidence. Also, we will use this t-test for the *pair-wise contrasts* as this is consistent with the directional expectations and provides for the most powerful test given the Welsh issues. We will further offer a Bernoulli interpretation of the results as detailed in *H2*.

Specifics of the inferential codex of Table 1. Where:

1) There is no indication at a FPE cut-point less than 25% that the Null is not the likely State of Nature as between the two datasets under examination, we will use: ≈. For example, for the Nigrini tests there was no statistically significant difference between 18,809 and 20,219 for a p-value < 0.25.
2) There is an indication for a FPE cut-point less than 5% that the Null can be rejected for the test data relative to the adjacent neighbor dataset, then the directional indication of < or > as is consistent with the hypotheses will be affixed. For example, for the Nigrini platform 20,219 was statistically significantly different from 11,487.
3) There is a counter indication relative to rejecting the Null as previously indicated relative to the Hypotheses then, ≠, will be affixed. There were no counter indications as the orders over all the tests followed the *Hs*.

The full dataset column is offered as a non-inferential benchmark; also there are no testing indications for the size of the sample effect for the Chi2 platform as discussed above. However, for completeness we tested the full dataset as a contrast with the profile of the RB*C*D; as expected, given the test of the RB*C*D, n=16, and the non-selected dataset, n=40, the average p-value for the three tests of the full against the RB*C*D, was 84.6% with a range of 24.5% clearly indicating that the Null was the state of nature and so there is little likelihood of a sampling bias.

5.3 The Results in Detail

The principal test results of our hypotheses are summarized as follows.

5.3.1 Testing *H1*

H1 states that for each of the NBDSSP Platforms, *excepting the Chi2 Platform as it is restricted to sample size*, we expect that there will be a strict order over the three trial groups in the direction: The RB*C*D will have fewer BSFs than the 10%DS which in turn will have fewer BSFs than the 250DS. According to Table 1, the ***NBPP***, ***Nigrini*** & the ***Distance*** Platforms for the movement from the RB*C*D to the 10%DS screening and then from the 10%DS to the 250DS for each of the pairwise contrasts was in the expected direction and statistically significant at a p-value < 0.05 as expressed in the *Hs* for all six of the scripted directional expectations. These testing blocks are shaded. For example, for the Test of Proportions platform [The ***Nigrini*** Platform] there was no difference between the stopping point for the average of the 56 or the Full dataset, 18,809 and the stopping point for the 16 Datasets in the RB*C*D sample which was 20,219 as noted above. There was, as hypothesized, a statistically significant difference between the RB*C*D stopping-point of 20,219 and the stopping-point of the 10%DS of 11,487 where the p-value was 1.04%. Further, between 10%DS and the 250DS where the stopping-point was 5,300 the p-value for this pairwise contrast was <0.1%. The overall results as scripted in Table 1 provide for clear support of the inferential context used to form *H1* also of note, there were no ≠ indications and only consistent significant indications < or >.

5.3.2 Testing *H2*

H2 states that for each of the NBDSSP Platforms, *excepting the Chi2 Platform as it is restricted to sample size*, we expect, for an overall Bernoulli test, that there will be directional differences consistent with the small

Sample size expectation order of *H1*. The Bernoulli test results show that all six directional movements, shaded in Table 1, are in the predicted direction and this produces a p-value of $(1/2)^6$ or 1.6% relative to the Null of no effect, providing a consistent inferential non-parametric confirmation of the result discussed for *H1*.

5.3.3 Testing *H3*

H3 states that for the *250DS*, *excepting the Chi2 Platform as it is restricted to sample size*, there will be NBDSSP BSFs indicating the need for extended procedure investigation considerations more frequently than for the *10%DS*. To test the Null form of *H3* that there is no difference in the NBDSSP BSF indications of EP between the 10%DS and the *250DS*, we will use the two sample test of proportions. To effect this test, we collected the EP BSF indications from the 10%DS and the 250DS by filtering these datasets to identify, for each run, where there were more than two BSFs created by the NBDSSP. However, there is a slight glitch regarding the inferential methodology for *H3* as there are no EP indications in the *10%DS* sample arm. In this case, then we will use ONLY the *250DS* sample and create, conservatively, a two tailed test at the 95% Confidence Interval (CI) and test if 0% is in this CI. If it is the case that 0% is in the CI for percentage of BSFs for the *250DS*, then we will fail to reject the Null form of H3. To elucidate this testing we offer these NBDSSP BSFs results in Table 2.

Table 2. Summary of the flag alerts over the four tests

	BSFs 10%DS	**BSFs 250DS**	**z-calc**	**Inference**
SampleSize	160	160	N.A	N/A
NBPP	1	16	3.82	Difference
Chi2	6	4	0.64	None (Note 4)
Nigrini	1	32	6.01	Difference
Distance	1	61	9.64	Difference

Table 2 presents the BSF indications of the NBDSSP for the two datasets. For example, 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. The number of BSFs produced in the *250DS* sampling arm are all statistically significantly higher, except as expected for the Chi2 block, than for the *10%DS*. However, for the *10%DS* there are no instances with more than two BSFs 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. As for the test that has been scripted for *H3* and re-cast above there were 8.1% [13/160] BSFs in the *250DS*; the lower limit of the non-directional 95% CI of this results is: 3.89%. Therefore, as 0% is not in this CI, one may reject the Null that there are no overall differences in the BSF indications between the *10%DS* and *250DS* samples.

6. Summary and Conclusion

6.1 Inferential Summary

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.

6.2 Protocol Indications for Conducting Studies where Large Dataset are Partitioned into Smaller Datasets

Partitioning large datasets is normal in the conduct of the audit. For example, as offered by a colleague who is the Director of Internal Audit for a MNC, he offers that for most of the downloads of aged accounts receivable, the dataset of interest is very often those accounts that have been labeled as actionable—i.e., delinquent. For most "well managed" organizations this is a relative small dataset. Therefore care needs to be taken in considering the Benford Screens in this case where a sub-sample is taken from the larger set of accounts.

In this regard, we offer the following protocol for the Benford Screening in the small data milieu:

1) According to the research results of Hill (1995a,b, 1996 & 1998, Wallace (2002) and Lusk & Halperin (2015b), if one combines small related datasets from large *Conforming* datasets the act of combining such small related datasets appears to move the Benford testing-results towards *Conformity* from the

Non-Conformity that was tacitly created due to the small sample size datasets.

2) Another protocol suggested by this research report is to first run the large dataset from which one desires to form a sub-set for the creation of audit evidence. If the large dataset passes the Benford Screening test this may allay the concern if the auditor does in fact observe that the sub-sample is flagged as *Non-Conforming*.

Clearly more research is needed relative to testing these two possible suggestions. We would be willing to act as a clearing center for the collection and the analysis of such datasets. In this context, of course, we do not need detailed audit tracking information that may compromise the privilege communication rationale that underlies the audit contract between the Audit LLP and the client. We only need the first digit information sets.

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 [In particular: Dr. Christopherson, Chair of the Economics' Group], Mr. Manuel Bern, Chief of Internal Audit: *TUI International, GmbH*, Hannover, Germany, and the reviewers of the *International Journal of Economics and Finance* for their careful reading, helpful comments, and suggestions. Finally, Dr. Yan Bao would like to thank the School of Management, *Xiamen University*, PRC for providing the data support for this research during the time when she was an international visiting scholar.

References

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

Cho, W. K. T., & Gaines, B. J. (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *American Statistician, 61*, 218-223. http://dx.doi.org/10.1198/000313007X223496

Collins, J. C. (2017). Using Excel and Benford's Law to detect fraud. *Journal of Accountancy,* April. Retrieved from https://www.journalofaccountancy.com/issues/2017/apr/excel-and-benfords-law-to-detect-fraud.html

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.

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. http://dx.doi:10.5430/ijfr.v8n3p27

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

Hill, T. (1995b). Base-invariance implies Benford's law. *Proceedings of the American Mathematical Society, 123*, 887-895. http://dx.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/10.1214/ss/1177009869

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. http://dx.doi.org/10.1511/1998.4.358T.P

Ley, E. (1996). On the peculiar distribution of the U.S. stock indexes' digits. *American Statistician, 50*, 311-313. https://doi.org/10.1080/00031305.1996.10473558

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. http://dx.doi.org/10.5430/afr.v3n2p191

Lusk, E., & Halperin, M. (2014c). Test of proportions screening for the Newcomb-Benford screen in the audit context: A likelihood triaging protocol. *Journal of Accounting and Finance Research, 4*, 166-180. http://dx.doi.org/10.5430/afr.v3n4p166

Lusk, E., & Halperin, M. (2015a). Account Screening Based Upon Digital Frequency Profiling in the Internal Audit Context: A Cartesian Distance Likelihood Triaging Protocol, *Business Management Dynamics, 5*, 12-17.

Lusk, E., & Halperin, M. (2015b). Testing the mixing property of the Newcomb-Benford Profile: Implications

for the audit context. *International Journal of Economics & Finance, 7,* 42-50 http://dx.doi.org/10.5539/ijef.v7n6p42

Mir, T. (2016). The leading digit distribution of the worldwide illicit financial flows. *Quality & Quantity [Springer], 50,* 271-281. http://dx.doi:10.1007/s11135-014-0147-z

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

Nigrini, M. (1996). A taxpayer compliance application of Benford's law. *Journal of American Taxation Association, 18*, 72-91.

Nigrini, M., & Mittermaier, L. (1997). The Use of Benford's Law as an aid in analytical procedures. *Auditing: A Journal of Practice & Theory, 16*, 52-67.

Ross, K. (2011). Benford's Law: A growth industry. *American Mathematical Monthly, 118*, 571-583. http://dx.doi.org/10.4169/amer.math.monthly.118.07.571

Tamhane, A., & Dunlop, D. (2000). *Statistics and data analysis* (1st ed.). Prentice Hall, Upper Saddle River, NJ USA.

Wallace, W. (2002). Assessing the quality of data used for benchmarking and decision-making. *The Journal of Government Financial Management, 51*, 16-22.

Notes

Note 1. The sampling interval in Lusk & Halperin (2014b) was [315 to 440]. Upon further testing of this interval using more sample points than were used in their paper, we determined that an interval [330 to 440] was more suited to the inference base and this interval was used in the NBDSSP.

Note 2. This and all of the modules/code are available in the NBDSSP from the author for correspondence.

Note 3. We selected 250 for this filter using the following guidelines: Lusk & Halperin (2014b) examined the Chi2 screen and arrived at a sampling interval of: 315 to 440. Conservatively, then we selected 500 as the upper limit and took 50% of this that is 250. Further, the Wallace (2002) research also arrives a point of change from Non-Conformity to Conformity at on the order of 250 observations.

Note 4. It is interesting to point out that the Chi2 platform does not have any sensitivity to the sample size issue that we are testing. This is as we indicated previously, to avoid the FPE jeopardy due to large sample sizes we have a Random Sampling filter for this platform that the sample size is restricted into the range interval of: [330 to 440] observations.

Copyrights