Business Management Dynamics
Vol.8, No.6, Dec 2018, pp.23-34

Testing the Small Size Effect Bias for Benford Screening: The False Negative Signaling Error

Frank Heilig[1] & Edward J. Lusk[2]

Abstract
Recent research has provided important information on the effect of partitioning large datasets that are likely to be Conforming to the Newcomb & Benford profile. This research documents that at some point, as the sample size is systematically reduced, sub-samples randomly drawn from Conforming datasets, test to be Non-Conforming. This has been termed a False Positive Screening Error [FPSE] – Incorrectly classifying a Conforming dataset as Non-Conforming; otherwise said: Failing to detect the True State of Nature of the data generating process and so in, the audit context, to incorrectly make the decision to investigate the dataset as one that may have been manipulated to a nefarious end. These research reports beg the question that motives our research – to wit: Is there such a partitioning effect if the dataset is Non-Conforming in nature? This is termed a False Negative Screening Error [FNSE]: Accepting as Conforming a Non-Conforming dataset and so failing to effect an Extended Procedures examination in the audit context when one would have been prudent. Method For control purposes, we used: (i) the same Decision Support System as was used in our previous research to screen the various sampled partitions, and (ii) the same partitioning algorithm to create the randomly drawn sub-samples. Results We find no evidence that the FNSE is produced at a rate that would usually be considered as counterproductive to the effective and efficient execution of the audit. Further, we used a simple Bayes filter to identify those Non-Conforming datasets that are in the definitive end of the Non-Conforming scale. In this case, there is still a FNSE jeopardy, albeit somewhat reduced. Impact These results add to the information on dataset partitioning or accrued from the onset of the data generating process. We document that there is a jeopardy difference between the FPSE and the FNSE. While Conforming datasets tend to be affected by sampling and incorrectly signal investigations at reduced sample sizes (FPSE), Non-Conforming datasets do not show the same tendency – i.e., to incorrectly decide not to investigate (FNSE).

Key words: T. Hill, Non-Conforming Benford Datasets, False Positive Screening Error

Available online
www.bmdynamics.com
ISSN: 2047-7031

INTRODUCTION: The Newcomb-Benford Profile A Screening Tool of Currency and Note

Newcomb & Benford Profiling

In the forensic and the audit context, one searches for anomalies relative to one's expectation. When there is convincing evidence of a divergence from expectation, it is prudent to launch an investigation – *This is the Audit Golden Rule*. One of the big-data screens that has achieved its rightful place in the panoply of the forensic and audit analysts is the Newcomb (1888)-Benford (1938) [NB] Screening Profile. *En Bref*, when

[1] *Senior Risk Manager Volkswagen Financial Services, Braunschweig, Germany*
E-mail: Frank.Heilig@vwfs.com
[2] *School of Business and Economics, The State University of New York (SUNY) at Plattsburgh, NY USA.*
The Wharton School, Department of Statistics, The University of Pennsylvania, Philadelphia, PA USA
E-mail: lusk@wharton.penn.edu

there is evidence of a meaningful variance from the Expected Frequencies [EF] of first digit profile scripted by EQ1:

$$EF_i = Log_{10}\,(1 + 1/i)\; i = 1, - - -, 9, \qquad \text{EQ1}$$

then the veracity of its data generating process is called into question. In statistical inference, this translates to: When the testing Null of: [No Difference between the NB-profile and that of the profile of the dataset under examination] is not likely to be the case then one rejects the Null in favor of the alternative that the dataset comes from a process that is not likely, at some probability level, to be *Conforming,* i.e., is likely to be *Non-Conforming*. An interesting explanation rationalizing the logic of the NB-profile in the big-data context, was introduced by the extensive research of Hill (1995$_{a,b}$, 1996 & 1998) who shows that unfettered *data-mixing* is an operative tenent of a generating process that produces a *Conforming* dataset—i.e., one that follows EQ1 in probability. A logical extension of Hill's observation is that there needs to be sufficient data to be mixed or combined in order to fill the nine digital bins of the first digit profile and thus create the expected NB-frequency profile. See also the work of Cho & Gaines (2007, p. 219) who note that size of the dataset does matter. The condition that there needs to be a dataset of sufficient size to effect a NB-Profile begs the following question:

If there were to be a data generating process that produces Conforming datasets, as the datasets are initialized and data starts to populate the dataset at what point are there sufficient digital-bin realizations for the dataset to accurately reflect the nature of the data generating process?

The Error Effect from NB Screens

This question suggests the following sample size anomaly regarding screening datasets using the NB-profile. If one were to sample a *Conforming* data generating process before there were sufficient bin-observations perhaps there would be a NB-screening indication that the process was *Non-Conforming* when in fact this incorrect assessment would be an artifact of the small sample size. In the NB-screening context then there are two research questions of interest:

1. *What is the size frontier when a Conforming data generating process is incorrectly identified as Non-Conforming?* This is the False Positive Screening Error [FPSE]: incorrectly believing that the dataset under scrutiny is produced by a *Non-Conforming process*.

2. The other contextual error is the False Negative Screening Error [FNSE]. *What is the size frontier when a Non-Conforming data generating process is incorrectly identified as Conforming—i.e., one fails to reject the operative Null when it is not the case and so incorrectly believes that the generating process is Conforming?*

Of the two data profiling errors, the FPSE size effect has been investigated; there is only preliminary or tentative information on the FNSE. The latter is the focus of our research. As a setting of the context for this research report, we will first consider the research that treats the issue of the small sample size effect that incorrectly creates the impression that the dataset is produced from a *Non-Conforming* data generating process—the FPSE.

Overview of Research Reports that Address the FPSE & The FNSE

Nigrini and Mittermaier (1997), following on the Nigrini (1996) study of fraud detection in US-Tax reporting, treat the possibility of using the digital profiling in the audit at the Analytical Procedures stage to screen the veracity of the processes in place in the audit client's Accounting Information System [AIS]. They have selected only very large datasets to conduct their analyses thus avoiding small or partitioned datasets.

A study conducted by Wallace (2002, p.22), notes the other aspect of partitioning—that of aggregation. Wallace indicates that for longitudinal data from 1995 through 1998 of taxable sales in the USA, Paraphrasing: *The graph displays variation in the years and also illustrates that the expectation of conformance with Benford's Law, ceteris paribus, improves with a larger sample size.*

Durtschi, Hillison, and Pacini (2004) report on forensic screening over suggested domains dealing with AIS reported data. They note (p.26) regarding digital variation instability introduced by small samples that: *"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".*

In a study similar to the Wallace study, Lusk and Halperin (2015) conducted an aggregation 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 tested aggregations 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) considered the screening of financial transfers with the end of detecting financial flows that were in violation of legal and agreement protocols from the sector called: 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.*

Heilig & Lusk (2017) created a robust Newcomb-Benford DSS that uses four screening platforms to identify departures from the Benford Practical Profile [BPP] reported by Lusk & Halperin (2014) called the Newcomb-Benford Decision Support System Profiler [NBDSSP]. We will employ the NBDSSP in our study. In their paper, for one-arm they tested the Hill (1998) lottery dataset. This was the first instance of a test of a dataset that was in the extreme *Non-Conforming* range as the Expected Frequencies of each of the nine first digits all equal to (1/9)%. Heilig & Lusk (2017, p.37), in testing the reliability of the NBDSSP, note that as they modified the various *Conforming* test sets so that they drifted systematically to the Hill equally-probable (1/9)% dataset, that most of the time the four DSS-platforms indicated a departure from the BBP and flagged that dataset as *Non-Conforming* in nature.

Bao, Lee, Heilig & Lusk (2018) 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. This report focused on these *Conforming* datasets and the effect of their partitioning in creating a FPSE-artifact. They collected two classes of repeated random samples: 10% from these *Conforming* datasets & then 250 items. They report on p.6 that for the 10% sample-arm, on the order of 1,200 items, that there were few instances where Extended Procedures[EP] would have been incorrectly suggested—i.e., a low probability of a FPSE of believing that the generating process was *Non-Conforming* when it was *Conforming* in State of Nature. However, when the sample size dropped to 250 then the number of instances of more than two Benford Screening Flags (BSFs) being produced by the NBDSSP were higher and statistically significantly different from the number produced in the 10% sampling arm. They note p.7: *For the 250DS over the 113 BSFs there were 13 instances with more than two BSFs for a particular dataset incorrectly suggesting that the EP investigations may be warranted.*

Finally, Bao, Heilig, Lee & Lusk (2018) conducted an extensive analysis of the Hill lottery dataset as one of the arms of their research report to determine the frontier point where partitioning incorrectly changes inferential signals as to the State of Nature for *Conforming* and *Non-Conforming*. For the FNSE, they note p.50:

> *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 not be warranted. 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*.

Research Précis

This is the point of departure of our study. Given the well-developed information on the FPSE as detailed above, the preliminary results regarding the FNSE essentially focused on the Hill Lottery dataset, and with the following advice offered by Bao, Heilig, Lee & Lusk (2018, p.51)

> *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.*

In this research report we will:

1. Identify a set of data-profiles that are argued in the literature as *Non-Conforming* in Nature,
2. Discuss the algorithm used to form test datasets from the *Non-Conforming* profiles noted in 1.),
3. Elaborate a method of creating smaller and smaller random samples from: 100%—i.e., the core dataset created in 2.) systematically down to: a random sample of 20 values,
4. Use the NBDSSP to analyze the groups of sets of random sub-samples of the *Non-Conforming* datasets by examining these samples to determine if there is a FNSE frontier as there was for the FPSE as reported above. Specifically, when does the sub-sample from the *Non-Conforming* dataset screen to indicate that it comes from a *Conforming* process—a FNSE.
5. Discuss the statistical inference protocol of the data analysis and examine the results of the NB-screening addressing the FNSE.

THE DATASET BASE ACCRUED FROM THE LITERATURE: *Non-Conforming*

Two researchers have investigated the Newcomb-Benford profiles and have offered instances of *Non-Conforming* profiles:

The six (6) *Non-Conforming* dataset profiles offered by Hill (1998) are found in Appendix Table A1. Specifically,

Profile (1): Fraudulent Tax Data of 1995 reported by the District Attorney of King's County, New York. [Fig5, p.363]

Profile (2): The Lottery Dataset [Fig4, p.362]

Profile (3): Points on the Standard Bell Curve for a Particular Population [Fig4, p.362]

Profile (4): The Latest Update of Atomic Weights [Fig4, p.362]

Profile (5): Student Scripting of a Six-Digit Random Number: Hill asked 743 First Year students to write down a six (6) digit number [Fig5, p.363]

Profile (6): The Average of the above three profiles. Hill notes that this average will be "fairly close". However, for completeness we did use this dataset as *Non-Conforming* [Fig4, p.362]

The second research report is offered by Cho & Gaines (2007) who collected a number of datasets that they offered as *Non-Conforming*[i]. These were taken from the public records of the Federal Election Commission [FEC]. Cho & Gaines note, p.219, that:

> *The FEC has made a practice of posting all reports to public electronic databases. A simple method of examining FEC data for signs of fraud is appealing partly because the very reason the FEC provides these data to the public is to guard against abuses of the system. By its very existence, the FEC archive enlists all interested parties in the task of monitoring the flow of money in federal elections.*

In this case, they focus on the Committee to Committee In-kind contributions bi-annually from1994 to 2004. Additionally, they provide the percentages of four specific category dollar value sub-partitions for these six years. These are noted in the Appendix systematically over Appendix Tables {A2, - - -, A8}. To be clear, these datasets are frequency profiles. Cho & Gaines principally use a distance measure fixed at 0 for the Benford central tendency profile and bounded at 1: the Maximal distance from the Benford profile to test relative departures from *Conformity*. They do not indicate which exact datasets in the Tables may be *Conforming* based upon a p-value screen where the Null was not rejected. They do use this distance measure to form indicative overall-profiles. The note, p.222, for example, for all 24 values of the disaggregate In-Kind profiles from 1994 to 2004

> *A Cox-Stuart test for trend using all 24 d* values indicates that the last three years, when blocked by level of contribution, have seen significantly worse fit to Benford (p < 0.001).*

This is to say, that some of the FEC datasets presented by Cho & Gaines may be *Conforming* where using aggregate or overall inferential tests one incorrectly rejects the Null of the *a priori* statistical belief, of *Conformity*, as, by chance, the p-value is lower than the boundary probability cut-off and so one rejects "the" dataset as coming from a *Conforming* generating process in favor of the alternative—i.e., that the dataset likely comes from a *Non-Conforming* generating process. In this case, one commits a FPSE by incorrectly rejecting a true indication of NO Difference for "some" of the specific datasets in the Cho and Gaines data-mix; however, which ones we do not know.

THE CREATION OF THE DATASETS FROM THE HILL AND CHO & GAINES RESEARCH REPORTS

To create the actual datasets, we used a simple VBA module that took the profiles as reported in the Appendix and generated the number of required digits. We selected as the full sample size a random selection of around 12 000 observations. For the Bao, Lee, Heilig & Lusk (2018) & the Bao, Heilig, Lee & Lusk (2018) studies the 100% or core sample size was around 12 000 items. In our study, this produced a range of sample sizes of [11 976 to 12 012] with an average of 11 999.42. This VBA-algorithm, available on request, takes a random sample for the 100% sample size and then selects number of realizations needed according to the reported profile percentage for the first digital percentage as reported in the Tables in the Appendix. This generates the number of digits needed until the last, 9th digit is reached. For example, ***for the values reported in Table 8 for monetary [values > $1 000] as bolded in Table 8***, the sample size randomly selected was 12 000; and, thus the Number of "1s" filled into the "1s" bin was: 5 880 [0.490 × 12 000]. This continued on for each of the other eight digits until all of the nine first digits had the profiled count-values. The sum of all of the count-number of bin-realizations for digits: {1, 2, 3 - - -, 9}, then would sum to 12 000. All the test datasets were filled according to the reported profile[ii].

RESULTS OF THE NBDSSP FOR THE NON-CONFORMING DATASET

To test for the FNSE, we used the same sampling protocol as reported by Bao, Heilig, Lee & Lusk (2018) [BHLL]. We took each of the 36 datasets reported in the Appendix as formed from the profile reported by Hill and Cho & Gaines and create 221 random samples for each. Specifically, our testing protocol starts with the 100% accrual. 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 took a sample set of: [1%, 0.5% & 20 data points]. Accordingly, we will produce 221 samples [1+ (10×19) + (10×3)] for testing for each of the 36 datasets or 7 956 [221 × 36] Sample points in total.

With this tableau/matrix of size: [36 Profiles × Samples: [36 × 221]] we examined how often the NBDSSP created more than two (2) NB-Screening indications. *Note*: The NBDSSP has as the following inferential testing Null: No Difference between that of the *Conforming* generating process respecting the NB-profile used by the NBDSSP and the NB-profile of the specific dataset under examination; this means that the NBDSSP assumes that the State of Nature of the generating process is *Conforming*-i.e., the digital profile of the dataset is expected to be *Conformity* with respect to the NB practical profile used by the NBDSSP. There are four (4) screening platforms and the NBDSSP has been vetted over a number of studies that:

If there are more than two (2) NBDSSP flags/indications of the four possible that the dataset does not fit with the NB-profile, then the likelihood is scored that the dataset is by Nature: Non-Conforming.

Using this calibration, as it was used in the previous studies that employed the NBDSSP, we recorded the number of instances for each of the 221 sample trials for the 36 datasets for which there were more than two (2) BN-screening indications that the dataset was in fact *Non-Conforming*. Recall in this case, rejecting the NBDSSP testing Null of *Conformity* is the correct decision as ALL of the profiles were selected from data profiles were argued/assumed to be *Non-Conforming*. If there were to have been two (2) or less indications, then this would be recorded as a FNSE. Simply, *en bref*: the 36 datasets were assumed to be *Non-Conforming* in nature, therefore, if the NBDSSP produced two(2), or one(1) or none (0) flags/indications, then the generating process of that tested dataset would be scored as *Conforming* and this would be a FNSE as we FAIL to correctly reject the testing Null of the NBDSSP to wit—[The dataset comes from a *Conforming* generating process] as the actual datasets were all assumed to be the result of a *Non-Conforming* data generating process.

Expectation of the Small Sample Effect relative to the FNSE
We have used the term expectation as there is no literature that has treated the inferential basis of examining a specific data profile that is *Non-Conforming* that would permit the forming of a hypotheses for a measured-frontier—i.e., expectation that are formed from literature in peer review research outlets. These expectations are, *a priori* respecting this set of data that we accrued, but are nonetheless drawn from our work in testing for the FPSE and a few focused testing-arms that addressed testing of the Hill Lottery dataset. In our testing of the FPSE and the Hill lottery dataset, we have gleaned and so proffer with the above caveats Expectation A:

EA: ***Expectation A*** *The profiles sampled from the 36 Non-Conforming sampled datasets of Hill and Cho & Gaines [Appendix A] from the 100% profile to the cut-point of 120 sampled points will likely exhibit a lessor percentage of BNDSSP alert flags compared to the profile from the 120 sample points through the 20 sample points.*

Discussion Recall the Hill datasets are mostly *Non-Conforming* but the aggregate-average, as Hill notes, is more unlikely to be *Non-Conforming* compared to the other examples that he offers. Also, the datasets offered by Cho & Gaines are in the aggregate or overall argued to be *Non-Conforming;* however, they do not flag particular datasets as *Non-Conforming* or *Conforming* as discussed above. This being the case, then in the set of 36 *Non-Conforming* datasets there are likely to be some, surely not many, data profiles that are *Conforming* in the mix of the 36 Datasets. In this case, early partitions, where there will be relatively low reductions in the sample size, will produce mostly *Non-Conforming* indications and also a few *Conforming* data indications. As the sample size decreases, the *Conforming* datasets are likely to be flagged as *Non-Conforming*, as BHLL demonstrate. As for the *Non-Conforming* datasets they are likely to rarely be identified as *Conforming* in nature as, Heilig & Lusk (2017) and BHLL demonstrate for the Hill lottery dataset. In summary, the early iterated partitions will have a lessor percentage of *Non-Conforming* indications than will likely be the case for the later set-partitions from 120 to 20 sample-partitions which is the sensitive zone identified by BHLL.

Results for Expectation
A Using the 36 Datasets over the various iterations produced the following indications:

We summed the number of the 36 datasets that had more than two NBDSSP indications for each of the 221 sub-sample partitions. We then divided this number by 36 to create 221 percentages from the 100% case to the last of the 20 point partitions. For example, for the 191st iterative set there were 30 of the 36 datasets for which there were more than two NBDSSP flags/indications. In this case, the percentage of datasets in that block that would have been classified as *Non-Conforming* was: 83.33% [30/36]. For the last of the 30 partitions there were 34 such indications. This percentage is 94.44% [34/36]. For the inference test of this arm of the study, we created the following two different sets of sub-partitions:

Early Set [n=191, Mean = 83.77%] ***Late Set*** [n=30, Mean = 88.52%]. The Early Set was the 100% case through the Sample Accrual of 600; there were 191 iterated sample sets in the group. The Late Set was the number of iterated samples from the sample accrual of 120 through the last iterated set of 20 sample points for which there were 30 iterated sample sets. This directional indication, [83.77% < 88.52%], which is consistent with *Expectation A,* was tested using the two-sample t-test assuming unequal variance, this was used as the ratio of the variances was: 2.13 and was identified by the Welch Test platform SAS™JMPv.13 as indicating that the variances were not likely to be the same. This directional Mean-test [83.77% v. 88.52%] has a p-value of 0.00037 strongly suggesting that the Null of *EA* is not the State of Nature thus rationalizing the inference that the percentage of *Non-Conforming* indications in the early set of partitions, n=191, are less than those in the later partition set, n = 30. In the interest of robustness, we also used the Wilcoxon/Kruskal/Wallis RankSum test: ***Early Set*** [n=191, Median = 83.33%] ***Late Set*** [n=30, Median = 88.89%]; the p-value [for both the Chi2 & the Normal Approximation] in this testing case was: <0.0001. In summary, one may reject the Null of EA and so *Expectation A* seems to be the likely State of Nature.

As we are interested in the nature of the investigation error in the audit context, if we accept *a priori* that there is audit evidence that these 36 datasets are *Non-Conforming*, for example there are forensic indications or Hill-mixing issues, and assuming that we will make the EP-investigation decision based upon the indications generated by the NBDSSP, then we should expect that overall—over all the possible partitioning possibilities—if the sub-partitions are in the early set sub-samples, the decision to investigate founded on the fact that there are more than two NBDSSP indications, will be correct in 83.77% of the time; in this case the FNSE—to wit we FAIL to investigate when that would have been the correct decision—will occur 16.23% [100% – 83.77%] of the time. If, however the sub-partitions are in the late set sub-samples, the decision to investigate founded on the fact that there are more than two NBDSSP indications, will be correct 88.52% of the time; in this case the FNSE will occur 11.48% of the time. In our experience, both are in the usual jeopardy ranges in the audit context and so there is a normal or acceptable FNSE risk in this testing direction.

Bayes Conditional Pre-Screening

For completeness, the auditor may opt for a Bayes screening conditional. This means that even though that the 36 datasets are *a priori* flagged as *Non-Conforming*, perhaps the In-Charge will pre-screen the datasets to further classify them before the screening analysis.

EB: ***Expectation B*** A reasonable *a priori experiential* pre-screen that suggests itself is to use as *Non-Conforming* ONLY those datasets for which the percentage of NBDSSP indications of *Non-Conformity* at the pre-screening level are > than 50% over all the 221 partitions. Thus, we eliminated the likely *Conforming* datasets using this Bayes pre-screen conditional, i.e., we eliminated six (6) datasets that had 50% or less indications of *Non-Conformity;* this resulted in 30 *Non-Conforming* datasets [36 – 6]. In this case, the test expectation is effectively modified and thus we would expect that the elimination of these datasets that may have been *Conforming* in the NBDSSP screening protocol should result in the FNSE in the later datasets sub-partitions to occur at a lessor rate in comparison to the rate in the early set of partitions. *Rationale* If we eliminate the possible *Conforming* datasets from the mix then we should find a higher degree of specificity for the NBDSSP to flag datasets in the Early dataset partitions as *Non-Conforming*—simply because the possible *Conforming* dataset have been removed. However, as we drift into the partitions where there are reduced sample sizes then there may be more indications than the Datasets are *Conforming* as the number of flags will be, here and there, less than or equal to 2.

Bayes Results

In the ***Early Set*** [n=191, Mean = 94.08%] ***Late Set*** [n=30, Mean = 91.89%]. This directional indication, [94.08% > 91.89%], which is consistent with *Expectation B*, was tested using the two-sample t-test assuming unequal variance, this was used as the ratio of the variances was: 2.25 and by the Welch-Test platform was significant. The Mean-test has a directional p-value of 0.034 suggesting that the Null is not likely the State of Nature thus rationalizing the inference that the percentage of *Non-Conforming* indications in the latter set of partitions, n=30, are less than those in the early partition set, n = 191. In the interest of robustness, we also used the Wilcoxon/Kruskal-Wallis RankSum Test]: ***Early Set*** [n=191, Median = 93.33%] ***Late Set*** [n=30, Median = 91.67%]; however, the p-value in this testing case was not significant at a level less than 0.05. In summary, *Expectation B* is only suggestive but not as well defined as was the test for *Expectation A*. Conservatively, we will use the blended Median as the FNSE indication. In this case, for the Bayes-arm we use the heuristic of the Median Weighted Average blend: Specifically, The Blended Median FNSE is: 6.89% [1– [93.33%*191 + 91.67%*30]/221]] = [1 – 93.11%]

As we are interested in the nature of the investigation error in the audit context, if we accept *a priori* that there is audit evidence that these 30 datasets are likely to be *Non-Conforming* and as the audit condition we will make the EP investigation decision based up the testing Null indications given by the NBDSSP, then we should expect that overall—that the FNSE will occur on the order of 6.89% of the time. Experience suggests that such a FNSE risk is in the acceptable range.

SUMMARY & OUTLOOK

Summary

This research report provides much needed information on the effect of sample size differences on the functioning of NB-Screens. As detailed by Ross (2011), NB-Screens are standard tools used in the forensic domain as well as the execution of audits guided by the GAAS. In this case, one of the questions of interest is: *How do these NB-Screens perform in the small sample environment*? By small sample environment, we mean that it is often the case that the audit In-Charge will (i) take a sample from an account under audit scrutiny for possible testing purposes, or (ii) may sample at the onset of the data-generating process. If there is a reason to use the NB-screen to have an indication of the nature of an account's generating process, then one must be attentive to the possible effect on the NB-screening indications due to the sample size. There are of course two operational issues: the FPSE where one effects an EP investigation when one is not likely warranted. Consistent research shows that partitioning and the resultant small sample sizes can invite the FPSE. However, there is little research on the other error—that of the FNSE. In this case, there is a *Non-Conforming* data generating process and auditor muses: *What is the audit jeopardy of taking sub-partitions of datasets from a Non-Conforming data generation process?* This was the focus of our research report. We find that in the FNSE direction for accrued samples and Bayes-conditioned samples that the FNSE jeopardy is at an acceptable range for most of the error calibration in the audit context. Specifically, we report that the worst-case scenario in the general case is a FNSE of 16.23% of the time. If a Bayes screen is use to ferret out possible insidious *Conforming* datasets, then the FNSE is reduced, as expected, to 6.89%. Both of these complementary errors to the FPSE are in the usual acceptable range from our experience in the audit context.

Outlook

We used datasets reported in the literature that more or less were argued as *Non-Conforming*. It would be helpful to have more datasets where there is a calibrated indication as to the "degree" of the *Non-Conformity* of the data generating process. For example, following on the advice of Collins (2017), it would be useful to have access to datasets from the deluge of defalcations in the 1990s such as: *Enron*, Inc.[iii] or *Qwest Communications International* Inc.[iv], or HeathSouth, Inc.[v] to mention a few; or datasets from the more contemporary traded organizations that lost their way such as: *VW/Audi* re: Diesel Defalcations[vi] or the Gold Standard: the *Lehman Bros*. LLP[vii] sub-prime debacle which actually drew in the "experts" from *Deutsche Bank*[viii]. We have communicated with the SEC to have access to the datasets for the above traded organizations. To date we have not received a response. Such defalcation datasets would be most valuable in calibration the FNSE. Failing this, as there certainly may be legal interdictions prohibiting the SEC from releasing data for firms that have violated the public trust, if there are datasets that are part of the audits that seem to have been produced from *Non-Conforming* generating processes, we would appreciate access to such datasets. We do not need any firm contextual data such as Names or any other identification markers. Just the actual dataset—i.e., the numerical values—are all that is needed and would be most appreciated. In this regard, we would be honored to be an archive site and upload such datasets to an open-access non-subscription download space.

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. Lee of the Accounting Department: 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., Heilig, F., Lee, C-H &, Lusk, E. (2018). Full range testing of the small size effect bias for Benford screening: A note. *International Journal of Economics and Finance*, *10*, 47-52. <http://dx.doi:10.5539/ijef.v10n6p47>

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*, 1-9. <http://dx.doi:10.5539/ijef.v10n2p1>

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. 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.>

Lusk, E., & Halperin, M. (2014). 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. (2015). 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

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

Appendix Non-Conforming Datasets from Hill (1998), Table A1, n=6 Profiles & Cho & Gaines (2007) Tables A2:A8, n=30 Profiles

Profile (1)	**Profile (2)**	**Profile (3)**	**Profile (4)**	**Profile (5)**	**Profile (6)**
0.001	0.111	0.360	0.472	0.147	0.314
0.019	0.111	0.129	0.187	0.100	0.142
0.000	0.111	0.087	0.055	0.104	0.084
0.097	0.111	0.081	0.044	0.133	0.079
0.612	0.111	0.077	0.066	0.097	0.085
0.233	0.111	0.074	0.044	0.157	0.076
0.010	0.111	0.068	0.033	0.120	0.071
0.029	0.111	0.064	0.044	0.084	0.073
0.001	0.111	0.060	0.055	0.058	0.075

Table A1 The Six Hill Non-Conforming Datasets

In-Kind 1994	In-Kind 1996	In-Kind 1998	In-Kind 2000	In-Kind 2002	In-Kind 2004
0.329	0.244	0.274	0.264	0.249	0.233
0.187	0.217	0.185	0.211	0.226	0.211
0.136	0.158	0.153	0.111	0.107	0.085
0.079	0.096	0.103	0.107	0.116	0.117
0.089	0.102	0.118	0.101	0.105	0.095
0.083	0.063	0.059	0.043	0.043	0.042
0.041	0.048	0.037	0.064	0.034	0.037
0.024	0.032	0.039	0.024	0.030	0.040
0.032	0.040	0.033	0.075	0.090	0.141

Table A2 The Overall Committee to Committee In-Kind Bi-Annual FEC Recorded Transactions

C&G1994[>$1,000]	C&G1994[$100-$999]	C&G1994[$10-$99]	C&G1994[$1 -$9]
0.579	0.305	0.349	0.090
0.190	0.187	0.206	0.067
0.108	0.153	0.126	0.073
0.081	0.077	0.083	0.060
0.027	0.104	0.083	0.062
0.001	0.075	0.047	0.502
0.001	0.038	0.051	0.054

0.011	0.023	0.027	0.034
0.001	0.038	0.027	0.058

Table A3 In-Kind Bi-Annual FEC Recorded Transactions 1994 by Dollar Magnitude

C&G1996[>$1,000]	C&G1996[$100-$999]	C&G1996[$10-$99]	C&G1996[$1 -$9]
0.558	0.259	0.159	0.057
0.191	0.226	0.218	0.116
0.073	0.172	0.154	0.210
0.127	0.090	0.096	0.099
0.044	0.108	0.109	0.080
0.002	0.056	0.088	0.080
0.005	0.028	0.085	0.080
0.001	0.024	0.048	0.077
0.001	0.036	0.043	0.202

Table A4 In-Kind Bi-Annual FEC Recorded Transactions 1996 by Dollar Magnitude

C&G1998[>$1,000]	C&G1998[$100-$999]	C&G1998[$10-$99]	C&G1998[$1 -$9]
0.548	0.282	0.188	0.101
0.306	0.192	0.144	0.084
0.039	0.158	0.192	0.054
0.065	0.113	0.105	0.027
0.039	0.141	0.110	0.104
0.001	0.037	0.100	0.191
0.001	0.029	0.060	0.054
0.001	0.027	0.046	0.289
0.001	0.022	0.054	0.097

Table A5 In-Kind Bi-Annual FEC Recorded Transactions 1998 by Dollar Magnitude

C&G2000[>$1,000]	C&G2000[$100-$999]	C&G2000[$10-$99]	C&G2000[$1 -$9]
0.560	0.249	0.184	0.427
0.308	0.203	0.213	0.036
0.045	0.142	0.101	0.056
0.050	0.154	0.077	0.021
0.036	0.117	0.105	0.053
0.001	0.040	0.045	0.167
0.001	0.047	0.101	0.062
0.001	0.021	0.031	0.058
0.001	0.027	0.144	0.120

Table A6 In-Kind Bi-Annual FEC Recorded Transactions 2000 by Dollar Magnitude

C&G2002[>$1,000]	C&G2002[$100-$999]	C&G2002[$10-$99]	C&G2002[$1 -$9]
0.543	0.250	0.195	0.034

0.316	0.234	0.206	0.073
0.040	0.107	0.124	0.069
0.041	0.172	0.078	0.019
0.057	0.118	0.097	0.203
0.001	0.038	0.051	0.165
0.001	0.032	0.038	0.119
0.001	0.031	0.030	0.111
0.002	0.018	0.181	0.207

Table A7 In-Kind Bi-Annual FEC Recorded Transactions 2002 by Dollar Magnitude

C&G2004[>$1,000]	C&G2004[$100-$999]	C&G2004[$10-$99]	C&G2004[$1 -$9]
0.490	0.238	0.165	0.035
0.359	0.231	0.155	0.031
0.040	0.095	0.089	0.040
0.043	0.180	0.071	0.035
0.064	0.129	0.055	0.256
0.002	0.035	0.052	0.172
0.001	0.037	0.041	0.154
0.002	0.027	0.055	0.181
0.001	0.028	0.316	0.097

Table A8 In-Kind Bi-Annual FEC Recorded Transactions 2004 by Dollar Magnitude

[i] Cho & Gaines (2007, p.219) note that Conforming datasets are often large—i.e., have numerous observations and of course other qualifying conditions that are related to the mixing process from unconstrained generation processes. This, of course, suggests that Small sample sizes may logically be expected to compromise the possibility of creating a *Conforming* profile. Further, Cho & Gaines, note, p.219: *On the flip side, numbers that would not follow Benford's Law have the following characteristics. 1. Numbers are assigned (e.g., check numbers, invoice numbers) 2. Numbers influenced by human thought (e.g., prices set by psychological thresholds ($1.99)) 3. Accounts with a large number of firm-specific numbers (e.g., accounts set up to record $100 refunds) 4. Accounts with a built-in minimum or maximum 5. Where no transaction is recorded.*

[ii] In some instances, there was a Table value reported as 0%; in fact, there were NO instances in the Tables for which the digital percentage was lower than 1.0%. Perhaps Cho & Gaines merely rounded and sometimes that produced a three-places “0” for the cells for which they entered “0”. Assuming that in the population, rarely would a bin would be empty, we added .1% to the cell for which there was a reported “0”.

[iii] https://www.sec.gov/spotlight/enron.htm.

[iv] https://www.sec.gov/news/press/2004-148.htm

[v] https://stakeholder11.wordpress.com/2014/11/24/healthsouth-inc-a-case-of-corporate-fraud//

[vi] https://www.caranddriver.com/news/everything-you-need-to-know-about-the-vw-diesel-emissions-scandal

[vii] https://www.investopedia.com/articles/economics/09/lehman-brothers-collapse.asp

[viii] https://www.db.com/company/index.htm