A Robust Newcomb-Benford Account Screening Profiler: An Audit Decision Support System

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

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

[2] The State University of New York, School of Business & Economics, Plattsburgh, NY, USA; & Emeritus, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Correspondence: Edward J. Lusk, SBE SUNY Plattsburgh, 101 Broad St. Plattsburgh, NY, USA 12901. Tel: 518-564-4190 or 215-898-6803.

Received: May 8, 2017 Accepted: May 26, 2017 Online Published: June 12, 2017

doi:10.5430/ijfr.v8n3p27 URL: https://doi.org/10.5430/ijfr.v8n3p27

Abstract

The *best practices* execution of the audit is conditioned by the facility with which Decision Support Systems [DSS] can be created using simple *Excel*™ programming tools and functionalities. Such DSS can aid in the exclusive binary triage of the many of the client's accounts each of which typically has tens of thousands of items into: {*Accounts that may warrant Extended Procedures Testing [EPT]*} or {*Accounts that may not warrant EPT*}. We use the Newcomb-Benford first-digit-profile as a triage platform to screen client accounts into the above mentioned exclusive sets. We call this DSS: The Newcomb-Benford Robust Screening:DSS [NBRS:DSS]. We report on the details of its development & vetting, and illustrate its functionalities using one of the historical Benford Datasets. The NBRS:DSS employs four account screening platforms each of which has been reported in the literature. The NBRS:DSS is available from the authors free as a download without restrictions to its use.

Keywords: big-data, FPE screening jeopardy

1. Introduction

In forensic analyses, as well as for account screening in the Planning, Interim-Testing and Substantive Phases of assurance engagements, data-profiling and benchmarking are the best practices requirements needed to rationalize the use of extended procedures testing for audits that fall under the purview of the PCAOB. See: Ovaska-Few (2017) and Appelbaum, Kogan & Vasarhelyi (2017) for an excellent discussion of the critical need for the integration of technical functionalities in the execution of the audit. This is usually the case as PCAOB audits are, by definition, audits of firms listed on trading exchanges. Such firms are, in the main, audits in the Big-Data milieu where decision support is the only practical way to conduct the audit. This theme is also taken up in the Pathways Commission Report (2014) where extensive coverage was given to the critical need for integrating technology into the delivery of courses and seminar workshops. In this regard, a must reading is the excellent report of Janvrin & Watson (2017) who provide a plethora of links and sources to not only accounting datasets but to DSS software download sources that offer "grass-roots" academic accounting course access to the tool sets that are the standard fare for the best practice execution of audits. Also see Greenman (2017). To underscore the need to prepare our student charges for the "*brave new e-world*" of screening tools and Decision Support Systems [DSS] note should be taken that *KPMG, LLP*: Boston: M&A Tax Advisory Services (Note 1) has engaged and utilizes Watson™ the DSS: created by *IBM* as has *HR Block Tax & Advisor Services* (Note 2).

Such data-profiling through DSS-screening is merely the logical extension of the requirement that the auditor use professional judgment in selecting client accounts for Extended Procedure Testing [EPT]. This EPT-account triage is often encoded in the Audit Programs used in the certified audits. In this regard, the audit In-Charge will select a set of sensitive accounts, usually those that impact either the Current Ratio or Cash Flow From Operations, and perform certain logical and relational tests so as to triage the set of accounts into two preliminary sets: {Interesting and Justifiable candidates for further EPT} & {Accounts that do not suggest EPT}. Not to belabor the point, but a simple illustration will provide the motivation for this idea as it pertains to the audit program. One of the audit programs currently used in the execution of the audit as detailed by: Arens, Elder, Beasley & Hogan (2017, p. 744) offers the following Cash screening protocol:

Audit Objective [Balance] *Cash in the bank as stated on the reconciliation foots correctly and agrees with the general ledger (detail tie-in).*

Common Tests *Prove the bank reconciliation as to additions and subtractions, including all reconciliation items.*

To effect such Cash tests, various parameters need to be established. In this regard, the In-Charge determines for Cash testing the: (1) nature of the discrepancy, (2) magnitude, and (3) frequency of occurrence that would suggest EPT of: (i) the Cash controls at the Interim-Testing Phase, and (ii) the related balance at the Substantive Phase. This clearly indicates that the professional judgment of the In-Charge guides this ETP-parameterization regarding the commitment of audit resources to further investigate issues in the Cash domain.

1.1 Partnering Audit Programs & DSS Data Screening

Audit programs, parameterized using the professional judgment of the auditor, are one of the traditionally accepted best practices techniques by which the auditor makes the decision to effect EPT and so collect audit evidence. Another, recently *en vogue*, is account-screening through data profiling using DSS. Both are recommended in the best practices execution of the audit; the latter is germane in the Big-Data context where the client may have thousands of accounts each one of which may have tens of thousands of transactions over the year under audit. This is to say that data screening is the only practical way to execute the audit in the Big-Data world. The partnering of Standard Audit Programs and DSS profiling is clearly illustrated in the following graphic:

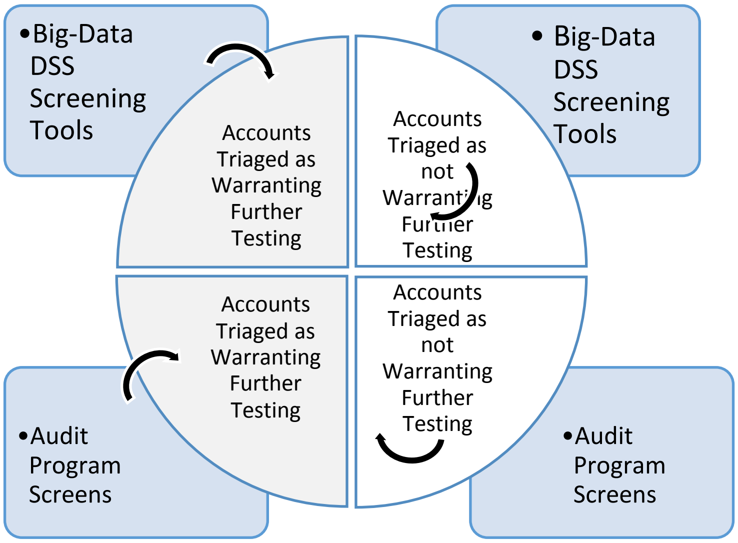
•Big-Data DSS Screening Tools
• Big-Data DSS Screening Tools
Accounts Triaged as Warranting Further Testing
Accounts Triaged as not Warranting Further Testing
Accounts Triaged as Warranting Further Testing
Accounts Triaged as not Warranting Further Testing
•Audit Program Screens
•Audit Program Screens

Figure 1. Use of Audit Programs and DSS-Screening in the Execution of the Audit

1.2 Summary

Figure 1 indicates the duality of the collection of audit evidence in the execution of the audit. Evidence formed from:

1.) The Standard Audit Programs, such as the Cash protocol discussed above, and

2.) The set of Decision Support Systems [DSS] that provide the IT-processing of the client’s datasets. There are, of course, a rich set of DSS that are part of the audit-panoply. Gaber & Lusk (2015) have detailed a number of account data-processing screens such as testing for: Duplicates, Gaps, Rounding, and Transpositions for the client’s dataset.

This is the point of departure of our study.

2. Précis of the Research

The focus of this paper is to detail an account-screening DSS that is best suited to be employed for audits where there are a large number of accounts with major transaction volume. Following, we will:

1.) Present, *en bref*, the remarkable observation of Newcomb (1881) and the subsequent investigations effected by Benford (1938) regarding First Digit Data Profiles, 2.) Detail four First Digit profiling platforms suggested by Lusk & Halperin (2014a),

3.) Offer a DSS called: the Newcomb-Benford Robust Screening Decision Support System [NBRS:DSS] that is formed using the data screening platforms offered by Lusk & Halperin (2014a, b, d & 2015a) all of which are individual VBA modules,

4.) Provide information used in the vetting of the NBRS:DSS, and

5.) Offer suggestions for future developments of profile screening.

3. First-Digit Profiling in the Audit: A Long Time Coming

3.1 Newcomb & Benford

The history of the digital profiling is most interesting and is very well detailed by Lusk & Halperin (2014a). As a very brief synopsis: Newcomb (1881) noticed that the preponderance of logarithmic characterizations of activity in his research domain was dominated by logarithms that started with 1 and decline systematically to 9 which was the least frequently used. Newcomb "quipped" that a simple abstraction of this first digit profile trajectory was not unlike:

$$P(F(i)) = Log_{10}\,(1 + 1/i) \text{ for i: 1, - - -, 9} \qquad \text{[EQ1]}$$

For example, "1s" and "2s" may be expected in the following proportion:

$$P(F(1)) = Log_{10}\,(1 + 1/1) = 30.10\% \qquad \text{[EQ1a]}$$

$$P(F(2)) = Log_{10}\,(1 + 1/2) = 17.61\% \qquad \text{[EQ1b]}$$

In the next century, Benford (1938) re-discovered Newcomb's curiosity and took up a detailed empirical examination of it. Benford provided a set of observations from very diverse activity domains that seemed to fit the profile sketched out by EQ1. Finally, Hill (1995a, 1995b, 1996, & 1998) is usually given credit for proving that digital profiles from large datasets which are produced from unconstrained and mixed mathematical manipulations can be expected to track along with the profile sketched out by EQ1. Also recently, Lusk & Halperin (2015b) report on the testing of the Hill mixing concepts. This had the obvious implication for DSS-screening in the execution of the audit—to wit, if the client's first digit profile for an account under audit diverges from the first digit profile offered by EQ1 then, this dataset may warrant further EPT.

In this regard, *circa* the 1990s Benford screens were being used in forensic analyses. Notably, Carslaw (1988), Christian & Gupta (1993) & Nigrini (1996 & 1999) offer practical forensic examples of the utility offered by the Newcomb-Benford profile. For example, Nigrini (1999, p.79) observes that:

> *Is it possible to tell that a number is wrong just by looking at it? In some cases, you bet. Using Benford's law—a mathematical phenomenon that provides a unique method of data analysis—CPAs can spot irregularities indicating possible error, fraud, manipulative bias or processing inefficiency.*

3.2 Summary

In the Big-Data milieu that often characterizes PCAOB audits it would be a simple matter to download and pass all of the client's large datasets through a DSS-screen to identify where there seems to be variance from the first digit profile of EQ1. Those datasets that seem to be at variance from the first digit profile [FDP] of EQ1 would be triaged as candidates for further EPT as suggested in Figure A. Consider now the four FDP-Platforms that are the NBRS:DSS.

4. The Four Platforms of the Robust Newcomb-Benford Profiler

In the Newcomb-Benford Robust Decision Support System [NBRS:DSS], there are four major computational functionalities, the aggregate of which, will constitute the inference platform for informing the decision maker regarding EPT in best practices execution of the certified audit. To be sure, and as suggested above, FDP is just one of the tools that can be employed in concert with the standard audit programs as well as the many other DSS functionalities that can be used in screening client accounts to the end of triaging accounts for EPT. Following we will detail the four screening platforms that constitute the NBRS:DSS and provide an illustrative example for each.

4.1 The Newcomb and Benford Measure: A Simple Screening Platform

Benford is usually given credit for moving Newcomb's observation into the analytic-milieu. Benford collected a large number of datasets from disparate sources and profiled them as they aligned with EQ1. This was the launching moment of digital profiling. A particular work that provided a synthesis and elaboration of the large number datasets offered by Benford is that of Lusk & Halperin (2014a) who report the correct mean profile of Benford profiles which they call: the *Benford Practical Profile* [BPP]. Using the BPP as the benchmark, Lusk & Halperin (2014a) offer a screening interval formed around the BPP. They call this the *Benford Screening Window* [BSW]. Using the BSW, they found that there was reasonable classification acuity if the more *than five of the nine digits* were outside the BSW.

An illustration will be useful at this juncture. The BSW interval around the BPP reported by Lusk & Halperin (2014a Table I, p.60) is presented in Table 1:

Table 1. The Benford Screening Window [BSW] based upon the BPP

First Digit Array	Corrected Means of Benford Datasets BPP	Lower Benford Screening Window [BSW] Value	Upper Benford Screening Window [BSW] Value
Digit 1	0.289 189	0.275 377	0.303 001
Digit 2	0.194 622	0.179 919	0.209 324
Digit 3	0.126 650	0.111 340	0.141 960
Digit 4	0.090 612	0.074 990	0.106 235
Digit 5	0.075 436	0.059 684	0.091 189
Digit 6	0.064 314	0.048 467	0.080 161
Digit 7	0.054 081	0.038 147	0.070 014
Digit 8	0.054 872	0.038 945	0.070 798
Digit 9	0.050 522	0.034 558	0.066 485

These intervals are used to benchmark the FDP from a particular dataset according to the following protocol:

If more than five of the nine first digits are outside the BSW, then the dataset is labeled as ***Non-Conforming****; otherwise, it is labeled as* ***Conforming****.*

4.1.1 Illustrative Example

Consider one of the Benford datasets:

Table 2. Illustration of the BSW as a profiling tool of the [NBRS:DSS]

First Digits	Benford Cost Data*	Inclusion Profile
1	**0.324**	*Not In*
2	**0.188**	**In**
3	**0.101**	*Not In*
4	**0.101**	**In**
5	**0.098**	*Not In*
6	**0.055**	**In**
7	**0.047**	**In**
8	**0.055**	**In**
9	**0.031**	*Not In*

*Benford notes that this is the Cost of Concrete, [CoC] n =741, but gives no further details.

In this case, there is evidence that the Cost of Concrete [CoC] reported by Benford (1938, Table 1, p.553) is a Conforming dataset as in only four cases is the first digital profile of the CoC out of the BSW interval range and so this dataset is assumed to be *Conforming*. This is as expected as the dataset was that collected by Benford as “support” for his thesis.

4.1.2 Summary

Platform I The first triage platform for the [NBRS:DSS] is the Benford practical profile and the BSW as presented in Table 1. We note this platform as the Benford Screening Window ***BSW-platform***. It is a very simple and intuitive operational protocol: The dataset under audit consideration will be profiled as to its first digit realizations. This is an automatic feature of the NBRS:DSS. ***Triage***: If the number of first digits NOT in the BSW is greater than five the dataset will be labeled as *Non-Conforming*; otherwise *Conforming*.

The second platform of the [NBRS:DSS] is one that uses the Chi2 measure for profiling the audit dataset.

4.2 The Chi2 Platform of the NBRS:DSS

As we previewed, screening is driven by the magnitude of the difference between the realization and the expectation formed by the In-Charge. The expectation can take many forms from the auditor's professional judgment, analytic procedure updates from previous years, objective calibrations, such as the BSW, or from assumed statistical population profiles. In this second platform of the NBRS:DSS, the latter is the fare of choice. In this regard, we have selected the Chi2 Test as this test is formed from the non-directed numerical magnitude between the *Realization* and the *Expectation.* Consider now the statistical form of the Chi2 inference model.

4.2.1 Overview of the Chi2 Model

The computational form of the Chi2 model is simple, relevant and often successfully employed in profiling. See for example the paper of Cho & Gaines (2007) where digital profiling using the Chi2 inference model in a forensic context is well explained and illustrated. Of the two computational forms: The Standard Pearson classification model that employs the Marginals of the realizations to form the Marginal Expectations or the Benchmark Expectation where the Realizations are profiled against a proffered Expectation. The more germane of the two in the audit context is the latter in that it is formed from the professional judgment of the auditor and often has inherently more variance associated with the estimates under the rejection of the NULL and so is conservative. In the Benchmarked Expectation version, the Chi2 computation is:

$$Chi^2 \ = \ \sum_{i=1}^{9}[(O_i - (BPP_i \times \mathrm{N}))^2/((BPP_i \times \mathrm{N}))], \ \mathrm{df} \ = \ 8. \qquad \text{[EQ2]}$$

Where: O_i is the Observed Number of realizations for the i^{th} digit, BPP_i is digit specific Benford Practical Profile[BPP] as presented in Table 1, and N is the sample size—i.e., the number of items randomly sampled from the account under audit.

However elegant in its simplicity and relevance to the screening task is EQ2, there is an inferential issue that needs to be addressed in using the Chi2 as one of the platforms of the NBRS:DSS. Consider this inferential issue next.

4.2.2 Sample Size Anomaly and Inferential Triage Point

Cho & Gaines (2007, p.220), who have explored the Chi2 test as an inferential screen in forensic analyses, note:

> "Chi2 *tests are very sensitive to sample size, having enormous power for large N, so that even quite small differences will be statistically significant. This test appears to be too rigid to assess goodness-of-fit well, especially since the Benford proportions do not represent a true distribution that one would expect to occur in the limit.*"

The Essential FPE-Inferential dilemma: The message or inferential glitch pointed out by Cho & Gaines is that while the theoretical Chi2 critical value for a particular point in the degrees of freedom zone is a scalar-constant, the Chi2 computed value IS an increasing function of the sample size; the inferential implication is: In using the Chi2 model, one invites the FPE signaling jeopardy. Simply, no matter how close the Realization is to the Expectation, in a Cartesian sense, there **is** a sample size where one may reject the Null of no difference. Lusk & Halperin (2014b,c) address this issue and have suggested that for the Benford test using EQ2 a sample size in the range of [315 to 440] seems reasonable to avoid this FPE signaling jeopardy. They note:

> *The "Goldie-Locks" result—i.e., the result that is "just-right"— is for a **Sample Size of 440**. In this case the corrected decision for both the Non-Conforming data as for all seven cases the χ^2 > 15.507 which is the 5% FPE cut-off so the auditor probably will investigate in all seven cases which is the correct decision; and for all of the Conforming cases the χ^2 < 11.03 which is the 20% cutoff and so extended procedures would not be indicated which is the correct decision. The result for the **Sample Size of 315** is almost the same as for the sample size of 440 excepting for the one case presented in Table C case for the RS In-Kind result for 1985/86 where the χ^2 for the Non-Conforming data is 11.11 which is almost at the 20% cutoff and less than the 90% cut-point of 13.36 suggesting usually that an investigation is not warranted when in fact one such have investigated.*

There is another inferential issue germane to creating an Alert Signaling system using the Chi2. If one uses the overall Chi2, which is the sum of the nine individual Chi2 cell-values, then if there were to be one very large Chi2 cell-value that one value may signal an Alert even though most of the other individual Chi2 cell-values were relatively low. Therefore, for consistency with the other signaling platforms and to avoid the sensitivity of one large cell value signaling an alert, we will use the nine individual first digit values and the illustrative result of Tamhane & Dunlop (2000, p.324) that individual Cell Chi2 values that are > 1.0 may warrant analytic attention. This is a balancing decision as the sum of the individual first digits may all be >1.0 and their sum could be less than the overall Chi2 significance alert of CHIQU.INV(0,95;8) which is 15.507.

4.2.3 Summary and Illustration of the Chi2 Platform

This then will be the Chi2 Platform for the NBRS:DSS inference triage. The NBRS:DSS computes the Chi2 calculated using **EQ2** and if more than 5 are > 1.0, then the data profile will be judged to be *Non-Conforming* in nature; otherwise the dataset will be judged to be *Conforming*.

For example, using the CoC example offered by Benford we find the following:

Table 3. Chi2 Inference Using the Random Sample Blocking Range [315 to 440]

First Digit	**Benford**	**BPP**	**Cell χ^2**
Array	**CoC**	**Projections**	**Values**
Digit 1	**0.324**	**0.289 19**	**1.84**
Digit 2	0.188	0.194 62	0.10
Digit 3	0.101	0.126 65	**2.29**
Digit 4	0.101	0.090 61	0.52
Digit 5	0.098	0.075 44	**2.97**
Digit 6	0.055	0.064 31	0.59
Digit 7	0.047	0.054 08	0.41
Digit 8	0.055	0.054 87	0.00
Digit 9	0.031	0.050 52	3.32
Totals	440	440	**12.04**

For example, consider the computation for the first digital profile. In this case, assume that we took a random sample of 440 from the Benford CoC dataset for which there were 143 items that had "1" as the first digit. In this assumed case for the 440 items, we find the percentages reported in the Col[**Benford CoC**]. We also have the BPP as the fixed expectation-benchmark. Using these values to parameterize EQ2 for the first digit we have:

$$234.6=[(O_i-(BPP_i\times \mathrm{N}))^2] \;=\; [(143_1-(0.289\ 19_1\times 440))^2] \qquad \text{[EQ2a]}$$

$$127.2=(BPP_i\times \mathrm{N})=\;(0.289\ 19_1\times 440) \qquad \text{[EQ2b]}$$

$$Chi2_1 = 234.6/127.2 \;=\; \mathbf{1.84} \qquad \text{[EQ2c]}$$

In this illustrative case, there are three Chi2 values > 1.0 suggesting that the CoC is a *Conforming* as expected. Incidentally, the overall Chi2 is 12.04 and is less than the Chi2-Triage point of 15.507 suggesting also that the Benford CoC dataset is *Conforming*.

We have discussed the FPE anomaly or screening jeopardy that is inherent in using the Chi2 as the inference screening. However, there are many inferential platforms where large sample sizes invite the FPE-anomaly. We mention this as the next platform in the NBRS:DSS uses the test of sampled proportions in forming the triage. As we will observe, this platform also requires calibrations in making an inference relative to the triage of an audit dataset. Consider now this issue and the way that it is addressed in the NBRS:DSS.

4.3 Platform Three: Test of Proportions over the DF Profile

For the Test of Proportion Screen, ***w***e tested the difference in the proportion of a particular digit compared to its expectation where the expectation is given by the BBP. If, over the nine first digits, there were more than five such tests, the z-calculated [z_{cal}] of which is individually greater than the 95% non-directional proportional value of 1.96, then the dataset from which the digital profile has been created is labeled as *Non-Conforming*; otherwise, it is labeled as *Conforming*.

4.3.1 Details

To form the testing platform for this section of the NBRS:DSS, we will be using the following Nigrini (1996) parametric equations for the single sample test of proportional differences:

$$s_i \; = \sqrt{\frac{(p_i)\; x\; (1-p_i)}{N}} \qquad \text{[EQ3]}$$

$$z_{cal} \; = \frac{|(op_i - p_i)| - \left(\frac{1}{2N}\right)}{s_i} \qquad \text{[EQ4]}$$

Where: N is the number of observations in the dataset; op_i is the ***O****bserved* ***P****roportion* of digit *i* in the dataset, $p_i$is the *Expected* ***P****roportion* of digit *i* and is taken from the BPP set of expectations, and$\left(\frac{1}{2N}\right)$ is the binary sample continuity correction; note, for large datasets, the sort of which one finds in the PCAOB audit context, this correction makes no practical inferential impact. We have used the absolute value as direction is not an issue in the audit screening context.

As there was for the Chi2 platform, the same FPE screening jeopardy needs to be considered where the inference is formed around the one sample test of proportions. A simple example may help to illustrate the FPE anomaly issue that is inherent in using EQs 3 & 4 in screening datasets.

4.3.2 Illustration of the Sample Size Calibration Issue

To demonstrate the signaling problem or the FPE-signaling anomaly that large sample sizes creates, consider the work of Nigrini (1996) where, in a forensic screening context, he undertook an analysis of Interest Income reported in Federal Income Tax filings for the year 1988. He found for the first digit of a population sized sample, n = 78,640 that the percentage of data with “1” as the first digit was: 30.59%. Using the BPP as the benchmark of 28.92% **or** the Log10 value of 30.10% we find the following as the z_{cal} values:

Analysis [30.59% vs. 28.92%]: z_{cal} = 7.3 >> 1.96 [EQ3a]

Analysis [30.59% vs. 30.10%]: z_{cal} = 3.0 > 1.96 [EQ3b]

Clear is that in the audit context, there is no meaningful screening difference between the Nigrini sampling result and either the BBP or the Log10 expectation. The FPE screening anomaly is produced by the population sized sample. As an interesting result, we gave to the students in our Capstone course in Auditing the two percentages 30.59% and 28.92% and asked the following question: “*Given this sampling result of 30.59% and the benchmark of 28.92%, would you as the audit In-Charge, assign a staffer, being charged to the audit budget, to use Extended Procedures to investigate this difference.*” Of the 73 students over two semesters three students elected to use EPT; that is 95.9% decided that this difference was not a signal that would warrant the use of audit resources. We agree; the question was effectively and unarguably rhetorical. To cope with the FPE sampling issue Lusk & Halperin (2014d) have posed the following functionality:

4.3.3 Summary of the Nigrini Platform Protocol

Using the BPP as the benchmark, the nine non-directed first digits are used to parameterize EQ3 and EQ4 using the sample size as an iterated variable where the starting value is set at 1. The nine z_{cal}s are then computed. Then the sample size is increased/iterated by a single unit; at each unit-increment the nine z_{cal}s are updated. At the point where the sixth z_{cal} achieves a value > 1.96, then the iterations stop. This is called the critical sample size[CSS]. If the CSS > =1825, then the dataset is labeled as *Conforming*, otherwise *Non-Conforming*.

As for the triage value of 1825, this was reported by Lusk & Halperin (2014d) here paraphrased following: They used a set of datasets reported in the literature as *Conforming* and another set of *Non-Conforming* datasets to find the minimum sample size which best separated/triaged between the *Conforming* and *Non-Conforming* datasets. This best or maximum likelihood point was for a CSS of 1825. In this case then, they indicated that the binary triage is: If the CSS < 1825 then the likelihood is that the dataset set under examination is *Non-Conforming* otherwise the dataset is *Conforming*.

This triage point is logical and intuitive. For example, if the CSS is small, that is much less than 1825, and only at that point is the sixth z_{cal} > 1.96, then this can only occur if the sampling result produces point estimates that are relatively different from the BPP value. For the other case, if the CSS is large, that is much more than 1825, and only at that point is the sixth z_{cal} > 1.96, then this can only occur if the sampling result produces point estimates that are relatively close/similar to the BPP value; an excellent illustration of this is the Nigrini example offered above. As additional empirical confirmation for the Hill (1998) lottery sample where all the proportions are 1/9, the CSS was 127. This means that only a very small sample size was needed to produce a sixth z_{cal} to a value > 1.96; this makes intuitive sense as all the proportions are 1/9 and so are very different than the BPP. In this case, the Hill sample is likely *Non-Conforming* in nature. On the other hand, if we used the BPP [Calibrated to 6 decimal places] as the audit dataset then there will be a sample size that will drive the sixth z_{cal} value to be >1,96; in this extreme case, the sample size iterations finally stopped at: 11 393 686 (Note 3) and as expected this is >> then 1825 as is, in a sense obvious, *Conforming*.

4.4 The Cartesian Distance Platform

Previously Lusk & Halperin (2015a) reported on using a simple Cartesian Distance Measure [CDM] for detecting datasets that may be at variance with the BPP and so of interest in the audit context for making the EPT decision. We have used their basic CDM-triage platform and enhanced it as a refinement for use in the NBRS:DSS. Following, we will report on these modifications and also the testing that we did of these modifications.

After we detail the CDM measure of the NBRS:DSS, we will use individual datasets that have been offered in the literature as *Conforming*, n= 30, and *Non-Conforming*, n=26, respecting their first digit profiles. This empirical benchmark is our surrogate for the lack of a formal inferential base.

4.4.1 The Distance Platforms

There are four independent constructed measures that are Cartesian based that form the triage value for determining if the particular dataset is *Conforming* or *Non-Conforming*. These are:

4.4.1a Mean of the Absolute Difference

$$MAD = [\sum ABS[BPP_i - C_i]]/9 \qquad \text{[EQ5]}$$

where: BPP_i and C_i are the values taken from Table 1 for the BPP digit proportion for the ith digit, and the **C**lient digital proportion from the account sampled respectively, i: 1, - - -, 9.

4.4.1b Median of the Absolute Difference

$$\text{MdAD} = Median\text{: } [ABS[BPP_i - C_i]_1, - - -, ABS[BPP_i - C_i]_9] \qquad \text{[EQ6]}$$

4.4.1c 95%Emperical Rule Dispersion referencing de Moivre; 95%ERd

The Empirical Rule [ER] introduced by Abraham de Moivre (1667-1754) [See Hald (p. 21), simply states that: Very often the distribution of collected empirical observations may be characterized using the Mean and the Standard Deviation [Sd] as follows:

68% of the observed data usually fall into the interval: [Mean ± 1Sd],

95% of the observed data usually fall into the interval: [Mean ± 2Sds], and

99% of the observed data usually fall into the interval: [Mean ± 3Sds].

Using de Moivre's remarkable empirical observation, called the Empirical Rule [ER], we will use the following Range-estimate to form the standard deviation of the absolute value of the difference between the BPP_i and the C_i. In fixing the ER-standard deviation, we will use the 95% estimate formed as:

$$\text{95\%ER} = Range\text{: } [ABS[BPP_i - C_i]_1, - - -, ABS[BPP_i - C_i]_9]]/4 \qquad \text{[EQ7]}$$

Where: the number 4 is the ER-dispersion coefficient for two standard deviations ± from the mean for capturing 95% of the observations.

We have selected this measure of the standard deviation as there are so few observations in the spanning set of the BPP that one relatively high value will exert, perhaps, too much influence in calibrating the distance measure if we were to use the sample standard deviation.

4.4.1d The Classic Distance Measure: Euclid-Pythagorean Cord Distance Range

This will be noted as EP and is the root of the summed squared dot product of the vectors $\overline{BPP}$ with $\bar{C}$ as follows;

$$EP = \sqrt{\sum(BPP_i - C_i)^2}\text{ i: [1:9]} \qquad \text{[EQ8]}$$

Finally, the Mean of the DM: [MDM], for each dataset under audit will be formed as:

$$\text{MDM=Mean: [MAD; MdAD; 95\%ER; EP]} \qquad \text{[EQ9]}$$

4.4.2 Calibration of the DSS Distance Measure

Using these four measures that are based upon different calibrations which are different in systemic terms, we took the Mean of the respective measures for each of the two test datasets: The *Conforming Set*, n=30 & *Non-Conforming*, n=26 and computed the means of MDM for the *Conforming* Datasets and *Non-Conforming* Datasets. Finally, we searched in the interior range of the weighted average of the two means to find a point where there was a balance for the Classifications and Misclassifications: *Conforming* scored as *Non-Conforming* and *Non-Conforming* scored as *Conforming* and *Conforming* scored as *Conforming* and *Non-Conforming* scored as *Non-Conforming*; these were the exclusive joint relative risk profiles. At this point, the details of this calibration would be most useful in the exposition.

To illustrate the way that the distance measure is formed consider the information in Table 4 for Benford's CoC dataset. This is, of course, one of the *Conforming* Datasets and is in the column labeled as CoC.

Table 4. Illustration of the Distance Measure for the Benford Death Rate Data Set

First Digits	Benford Practical Profile	CoC Data	ABS of the Difference
1	0.289	0.324	0.035
2	0.195	0.188	0.007
3	0.127	0.101	0.026
4	0.091	0.101	**0.010**
5	0.075	0.098	0.023
6	0.064	0.055	0.009
7	0.054	0.047	0.007
8	0.055	0.055	0.000
9	0.051	0.031	0.020

Using the formula above EQs:[4-8], we find, using the BNRS:DSS for the CoC dataset, the following profile for the MDM:

MDM= Mean: [MAD; MdAD; 95%ER; EP]

$$\text{MDM}= [[0.015\ 2 + \mathbf{0.010} + 0.008\ 8 + 0.055\ 8]/4] = 0.022\ 43 \qquad \text{[EQ9a]}$$

Now that we have this information for this particular instance, we will need to make the comparison of this particular dataset to a generalized triage point to make the decision if this dataset is more likely to be *Conforming* or *Non-Conforming*.

4.4.3 The Generalized Triage Point

As indicated above, we used the 30 *Conforming* Datasets and the 26 *Non-Conforming* datasets that were reported in the literature. The MDM-means of these two datasets were: *Conforming*: [0.022 06] and for the *Non-Conforming*: [0.047 53]. This directed difference had a p-value 0.009 76 using the unequal-variance t-test from the *Excel™[2013] DataAnalysis;AnalysisTools*; certainly a strong indication of a directed difference and so rationalizes the rejection of the Null. We then computed the weighted average of the MDM-means of the *Conforming* and the *Non-Conforming* datasets; we found this to be: 0.033 88. We then iterated this weighted-average mean along a sliding scale while tracking the four classification possibilities. Our Stopping-Rule was a balance individually between the Classifications and the Misclassifications: relative to *Conforming and Non-Conforming*. This balance occurred for a value of: 0.026 38 where the datasets were classified as follows:

Table 5. Classification of the Datasets Abstracted from the Literature

Triage Classification*	Conforming from Literature	Non-Conforming from Literature
Triaged as Conforming	21[70%]	8[31%]
Triaged as Non-Conforming	9 [30%]	18[69%]

*The fifteen point decimal place representation of the triage cut-point number is: **0.026 382 330 574 550**

This triage point seems reasonable in that approximately a 70% success rate in the identification of both *Conforming* and *Non-Conforming* is well above the chance Null-profile. Specifically, the Fisher's Exact Test for the non-directional testing case has a p-value < 0.007 and is consistent with a rejection of the Null of Chance. This is a strong test result for the vetting of the Distance measure triage cut-point and so this will be used in the Distance Platform of the NBRS:DSS.

4.4.4 Summary

In this case then using the Distance Triage point, the NBRS:DSS makes the full decimal comparison 0.022 43 against 0.026 38; and, as 0.022 43 < 0.026 38 the Benford CoC dataset is labeled as *Conforming*.

Now that we have examined the four platforms that will be used in the NBRS:DSS let us consider the performance of the NBRS:DSS.

4.5 The NBRS:DSS: The Functionality in Perspective

4.5.1 Overview of the NBRS:DSS Functionalities

The auditor enters [*Paste*] the client's dataset under audit scrutiny into the Front Page of the NBRS:DSS and launches the NBRS:DSS. Next a *UserForm* asks the auditor to select a confidence level. The CI:UserForm has three choices which affect ONLY: the z-calculation for the Nigrini platform as the other platforms are already calibrated as detailed previously. The Alert Signal Protocol, i.e., the four the individual signals that suggest EPT investigations are:

4.5.1a The Benford Platform [Alert Flag]: If the number of instances where the digits of the Dataset under audit are outside the BSW is > 5.

4.5.1b The Chi2 Platform [Alert Flag]: If the number of cell Chi2 values > 1.0 are > than 5.

4.5.1c The Nigrini Platform [Alert Flag]: If the CSS is < 1825 when the number of digits that have z-values > 1.96 is > 5.

4.5.1d The Distance Platform [Alert Flag]: If the DMD is > 0.026 382 330 574 550.

4.5.2 Confidence Parametrization

The Standard, and recommended, default-choice offered by the CI:UserForm for the confidence interval is 95%; this is consistent with the PCAOB's acceptable risk level of 5%. The other choices are:

4.5.2a For ***High Risk Audits***, the z-screen parameter is set at 99% (2.33); this is a conservative calibration respecting the FPE as this confidence interval is the widest of the three and so if there are indeed values outside this 99% CI they are highly indicative of a variance alert and so strongly suggest a *Non-Conforming* dataset.

4.5.2b The other choice is for ***Low Risk Audits*** where the confidence level is pitched at 90% (1.645). Here again this is conservative respecting the FNE as this confidence interval is the narrowest and so any indications outside this are suggestive that perhaps the risk level was in fact NOT really Low and gives pause to due diligence reflection.

4.5.3 Pre-Vetting Calibration

The central question is: *What is the vetting profile of the NBRS:DSS?* To form a reasonable testing dataset, as the four constituent platforms were formed with the datasets collected by Lusk & Halperin (2014a) and so cannot form an objective vetting, we took the BPP as the idealization of a Conforming Dataset and iterated it as follows:

The vector[BPP] is noted as: $\overline{BPP_{9x1}}$, where the digital values of the 9x1 vector are taken in order from Table 1. Next, we selected the Lottery Dataset of Hill (1998): $\overline{L_{9x1}}$, where the digital values of this 9x1 vector are uniformly [1/9]. Dropping the vector index notation, we then formed the directed difference: [$\overline{BPP} - \overline{L}$] call this: $\overline{\Delta}$. The iteration for each of the nine elements was formed as $\overline{\Delta}$/10. Specifically: $\overline{\Delta}$/10 has the following values listed in transposed-order:

0.017 808	0.008 351	0.001 554	-0.002 05	-0.003 57	-0.004 68	-0.005 7	-0.005 62	-0.006 06

In this case, there are sequential iterations each of which creates a new dataset that moves from the BBP to the Hill lottery dataset. For example, the first iteration creates the following dataset:

Table 6. First Vetting Dataset from BPP to the Lottery Dataset

Digits	BPP	Iterated Data	Difference[E-I]
1	0.289 19	0.271 38	0.017 81
2	0.194 62	0.186 27	0.008 35
3	0.126 65	0.125 09	0.001 56
4	0.090 61	0.092 66	-0.002 05
5	0.075 44	0.079 00	-0.003 56
6	0.064 31	0.068 99	-0.004 68
7	0.054 08	0.059 78	-0.005 70
8	0.054 87	0.060 49	-0.005 62
9	0.050 52	0.056 58	-0.006 06
Checks	1.000	1.000	-0.000

Table 6 shows that each iteration, $\overline{\Delta}/10$, just moves each digit systematically from the BPP to the Hill lottery dataset. This then will be the ideal dataset to vet the NBRS:DSS in that we can form a logical expectation for the vetting protocol. We will pass the iterated datasets through the NBRS:DSS and record the alert profiles over the iterations from the BPP, where we proffer that there will be no ***Alerts***, to the Hill Lottery dataset set that is the terminal dataset, where there should logically be four ***Alerts***. Additionally, we will produce these profile tables for each of the three confidence levels.

The pre-vetting ***Alert*** profile for the NBRS:DSS used all of the 56 dataset that were collected by Lusk & Halperin (2014a) over their study sets. In this regard, we found that the best likelihood cut point was more than two alerts. Therefore in this pre-vetting stage, if there are three or four NBRS:DSS alerts this was consistent or suggestive of *Non-Conformity* and so that dataset would likely warrant an Extended Procedures examination.

4.5.4 Pre-Vetting Summary

The Alert calibration from the pre-vetting stage is:

4.5.4a If there were {0, 1 or 2} ***Alerts***: The dataset is benchmarked, in a likelihood sense, as *Conforming*,

4.5.4b If there were {3 or 4} ***Alerts***: The dataset is benchmarked, in a likelihood sense, as *Non-Conforming*

4.5.5 Vetting Profile

Now that we have an initial calibration for the NBRS:DSS triage, we will offer the vetting-profile for the calibration effectiveness of the NBRS:DSS. In this regard, we expect that there should be a "smooth" transition from 0 ***Alerts*** for the BPP to 4 ***Alerts*** for the Hill Lottery dataset for all three confidence levels. There is a large volume of data in these various profiles. All of this data is available from the authors. Following we present a simple and consistent profile Table 7 where: the cells are the number of Alerts the range of which is 4: [0 to 4] from the BPP to the Hill Lottery dataset. The cells in Columns D & E [Shaded] report the split of the 5th iterated dataset to further indicate the transitions of the alert profiles.

Table 7. Sensitivity Alert-Profile of the NBRS:DSS over the Three Confidence Levels

	BPP	A	B	C	D	E	F	G	H	I	J	HILL
CL90	0	0	0	2	3	3	4	4	4	4	4	4
CL95	0	0	0	2	4	4	4	4	4	4	4	4
CL99	0	0	0	2	4	4	4	4	4	4	4	4
MEAN	0	0	0	2	3.3	3.3	4	4	4	4	4	4

This profile clearly exhibits the expected alert-profile for a sensitive DSS. Sensitivity, defined as the ability of the screening tool to differentiate profiles expected to *Conform* from those expected to be *Non-conforming*. Further, in the specificity direction, we proffer that for all three Confidence Levels *as well as* for each of the four profiling platforms: Benford[BSW], Chi2, Nigrini & MDM the measures, there should be consistent change in the expected direction. For example, Table 8 shows the four platforms for the BPP dataset, the average of the four mid-range iterated dataset, and finally the Hill dataset:

Table 8. For the 90% Confidence Level: Alert Platforms Offering Specificity Profile

BPP Dataset		Mid-Transition Datasets		Hill Dataset	
Platforms	Alert	Platforms	Alert	Platforms	Alert
Benford	0/9	Benford	6.5/9	Benford	9/9
Chi2	0/9	Chi2	6.5/9	Chi2	9/9
Nigrini	> 40 000	Nigrini	435.75	Nigrini	95
Distance	0.000 422	Distance	0.044 95	Distance	0.09 718

For Table 8 in the specificity direction, there is consistent movement from the BPP to the Hill dataset for all the elements tested. For example, for the 90% CL and the BPP for the Nigrini platform, the number of iterations needed to reach six alerts was actually: 8 027 003! This is as expected for the BPP. For the Hill dataset, only 95 iterations were needed to create the sixth alert. This same profile is found for the other two Confidence Levels. [Information not shown but available from the authors].

5. Summary and Outlook

5.1 Summary

The NBRS:DSS evaluates and reports the exclusive screening result:

4.6.1a A *Non-Conforming* dataset where {3 or 4} Alerts are in evidence, or

4.6.1b A *Conforming* Dataset where {0, 1 or 2} Alerts are in evidence.

We offer that the NBRS:DSS can be simply and effectively employed to provide account triage so that the EPT decision can be focused on the set of accounts which may likely warrant EPT. Such DSS are essential for audits in the Big-Data milieu. We offer the NBRS:DSS as a freeware-download without restriction on it use; contact the authors.

5.2 Outlook

Screening of accounts is fundamental in executing the audit to align with the "best-practices" imperative of the PCAOB and to partner with the standard Audit Programs used in conducting the audit. Screening is the triage needed to, insofar as possible, effectively and efficiently use the audit resources. The NBRS:DSS is one of many screening tools that may aid in this endeavor. The principal benefit of the NBRS:DSS is its *robustness dimension*. While there are many screening and software platforms that offer aspects of Newcomb-Benford screening there are none, of which we are aware, that provide multidimensional profiling such as: (i) an Empirical re-casting of the Log10-function profile, (ii) Chi2 cell profiling, (iii) test of Proportion profiles over the individual first digits, & (iv) a Cartesian Distance measure. Viewing audit datasets over these four independent dimensions aids in controlling the False Positive Error as three or four alerts need to be produced to effect an EPT in the audit context. This is clear if one looks at Table 7 where there are, here and there, two audit alerts which in a non-robust context would likely create instigations where they may not have been warranted. Addressing this is, of course, the purpose of robustness calibration. In the future, it would be most useful to investigate the signaling acuity of the NBRS:DSS *vis-à-vis* the many other screening tools that exist in the audit and forensic context. The complimentary profiles of the NBRS:DSS with other screening tools, the sort of what are presented in Gaber & Lusk (2015), may enhance the sensitivity and specificity of the ensemble of these aggregated screening platforms.

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 and the anonymous reviewers of the *International Journal of Financial Research* for their careful reading, helpful comments, and suggestions.

References

Appelbaum, D., Kogan, K., & Vasarhelyi, M. (2017. Feb.). An introduction to data analysis for auditors and accountants. *The CPA Journal*, 32-37.

Arens, A., Elder, R., Beasley, M., & Hogan, C. (2017). *Auditing and Assurance Services* (16th ed.). Pearson Publishers.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, *78*, 551-572. Retrieved from http://www.jstor.org/stable/984802

Carslaw, C. (1988). Anomalies in income numbers: Evidence of goal oriented behavior. *The Accounting Review*, *63*, 321-327. Retrieved from: http://www.jstor.org/stable/248109

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

Christian, C., & Gupta, S. (1993). New evidence on “secondary evasion”. *The Journal of the American Taxation Association*, *15*, 72-92.

Gaber, M., & Lusk, E. (2015). Account screening: Rationalizing the extended procedures decision in the audit context. *EXCEL International Journal of Multidisciplinary Management Studies*, *5*, 1-20.

Greenman, C. (2017). Exploring the inpact of atrifical intelligence on the accounting profession. *Journal of Research in Business, Economics and Management*, *8*, 1451-1454. Retrieved from www.scitecresearch.com/journals/index.php/jrbem/index

Hald, A. (1998). *History of Mathematical Statistics from 1750 to 1930*. Wiley and Sons New York, NY, USA.

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. http://dx.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.

Janvrin, D., & Watson, M. (2017). “Big Data”: A new twist to accounting. *Journal of Accounting Education*, *38*, 3-8. http://dx.doi.org/10.1016/j.jaccedu.2016.12.009

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

Lusk, E., & Halperin, M. (2014b). Detecting Newcomb-Benford digital frequency anomalies in the audit context: Suggested χ^2 test possibilities. *Accounting and Finance Research*, *3*, 191-205. http://dx.doi.org/10.5430/afr.v3n2p191

Lusk, E., & Halperin, M. (2014c). Detecting digital frequency anomalies as benchmarked against the Newcomb-Benford theoretical frequencies: Calibrating the χ^2 test: A note. *International Business Research*, *7*, 72-86. http://dx.doi.org/10.5539/ibr.v7n2p72

Lusk, E., & Halperin, M. (2014d). Test of proportions screening for the Newcomb-Benford screen in the audit context: A likelihood triaging protocol. *Accounting and Finance Research*, *3*,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

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, *4*, 30-40. Retrieved from http://www.jstor.org/stable/2369148

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

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

Ovaska-Few, S. (2017, March). What large firms expect from new accounting grads. *The Journal of Accountancy*, 29-30.

Pathways Commission. (2014). Charting a national strategy for the next generation of accountants. Retrieved from http://commons.aaahq.org/posts/a3470e7ffa

Tamhane, A., & Dunlop, D. (2000). *Statistics and Data Analysis*, Prentice Hall, Upper Saddle River, NJ USA.

Notes

Note 1. https://home.kpmg.com/us/en/home/media/press-releases/2016/03/kpmg-announces-agreement-with-ibm-watson-to-help-deliver-cognitive-powered-insights-.html

Note 2. https://www.hrblock.com/tax-offices/?otppartnerid=9007&campaignid=ps_mcm_9007_0207&omnisource=YHO|CAMPR-B-Brand+Locations-Y-Exact|ADGPArrowhead|KWRDH&r%20block%20watson&KeywordID=480839#/en/

Note 3. In the interest of execution efficiency of the NBRS:DSS, we set the stopping rule at an iterated sample size of 40 000.