Test of Proportions Screening for the Newcomb-Benford Screen in the Audit Context: A Likelihood Triaging Protocol

Edward J. Lusk[1,2] & Michael Halperin[3]

[1] The State University of New York (SUNY) at Plattsburgh, Plattsburgh, NY, USA

[2] Emeritus, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

[3] Director Lippincott Library of the Wharton School, University of Pennsylvania, Philadelphia, PA, 19014, USA

Correspondence: E. Lusk, SBE SUNY Plattsburgh 101 Broad St. Plattsburgh, NY, USA 12901. Tel: 1-518-564-4190. E-mails: luskej@plattsburgh.edu or lusk@wharton.upenn.edu

Received: October 10, 2014 Accepted: October 30, 2014 Online Published: November 10, 2014

doi:10.5430/afr.v3n4p166 URL: http://dx.doi.org/10.5430/afr.v3n4p166

Abstract

The basis of the certification audit, in a non-forensic context, is a random sample of sufficient size to create the evidence needed to justify the audit opinion. This has been clearly stated in all of the relevant SAS pronouncements issued by the AICPA and also by the PCAOB through AS 5. However, there has been a dearth of specifics on the critical selection of the set of accounts that are reasonable targets for the statistical sampling needed if extended audit procedures seem warranted. In this paper, we present a simple and validated protocol based upon the digital frequency testing introduced by Newcomb & Benford and popularized in the audit context by Nigrini to identify accounts under audit that seem reasonable candidates for extended procedures testing. The montage of the protocol centers around the parametric test of proportions the equations which were introduced by Nigrini. We validated the logic of the protocol by using a holdback sample. Finally, we have coded the account identification protocol as a Decision Support System in VBA: Excel™ that is available from the authors free without restriction to its use.

Keywords: False Positive and False Negative Signaling Errors, *Conformity* Triage

1. Best Practices Sampling and Screening Protocols: Creation of Audit Evidence in PCAOB Audits

We have now entered the second decade of the Sarbanes-Oxley: 2002 era where the Public Company Accounting Oversight Board [PCAOB] has been very pro-active in creating the rules and regulations for the execution of the certification audit. [*Audit Standard 5* [AS 5]]. Most of the detailing of these rules and regulations impacts the selection and subsequent sampling of accounts in the creation of audit evidence. However, interestingly and certainly surprisingly, the incidence of reporting irregularities, whether due to error or fraud, remains a major concern. According to Hogan, Rezaee, Riley, Jr. & Velury (2008, p.232):

The frequency of financial statement fraud has not seemed to decline since the passage of the Sarbanes-Oxley Act in July 2002. For example, the 2005 biennial survey of more than 3,000 corporate officers in 34 countries conducted by PricewaterhouseCoopers (PwC) reveals that in the post-Sarbanes-Oxley era, more financial statement frauds have been discovered and reported, as evidenced by a 140 percent increase in the discovered number of financial misrepresentations (from 10 percent of companies reporting financial misrepresentation in the 2003 survey to 24 percent in the 2005 survey). The increase in fraud discoveries may be due to an increase in the amount of fraud being committed and/or also due to more stringent controls and risk management systems being implemented PricewaterhouseCoopers(2005). The high incidence of fraud is a serious concern for investors as fraudulent financial reports can have a substantial negative impact on a company's existence as well as market value.

In addition to the relative fledgling PCAOB, the American Institute of Certified Public Accountants [AICPA], through their *Statements on Auditing Standards* [SAS], has long been active in addressing assurance issues relative to the integrity of reporting as detailed in the *Generally Accepted Audit Standards* [GAAS]. We wish to draw attention to the Standard of Field Work, No.3[SFW(3):GAAS] the intent of which underlies the creation of audit evidence by sampling:

The auditor must obtain sufficient appropriate audit evidence by performing audit procedures to afford a reasonable basis for an opinion regarding the financial statements under audit.

The number of AICPA SAS pronouncements to provide assistance in realizing SFW(3) and in support of AS 5 of the PCAOB has been voluminous to say the least. For example, consider the following lineage: SAS 46, SAS 53, SAS 82 and SAS 99. The most of these SASs have the same theme; the most recent is titled: *Auditors' Consideration of Fraud in a Financial Statement Audit*.

There seems to be an obvious *pronouncement* and *effect* "glitch." On the one hand the PCAOB and the AICPA have rolled out numerous detailed pronouncements that go to the heart of the obvious extension of SFW(3): *The examination of accounts, on a sampling basis, that seem in need of extended procedures examination*; yet according to *PwC* and the work of Hogan, Rezaee, Riley, Jr. & Velury (2008) the incidence of audits that fail to deliver on the "correct" assurance/opinion seems inexplicably high. This expectation *disconnect* has been termed the reporting gap—the difference between what the readers of the financial statement *can expect* and what the auditors *can deliver*. A recommended reading is: Kassem & Higson (2012).

One issue underlying this *disconnect* may lie with *difficulties in identifying accounts that should be subjected to extended procedures examination*. After all, the auditor cannot examine all the accounts! This is the point of departure of our research. In the following sections of the paper, we will offer a testing protocol that addresses what we consider to be the major failing in the certification audit regarding the creation of relevant the audit evidence: *The lack of a simple account screening methodology that can help the auditor to focus the auditing resources on those accounts that are likely candidates for extended procedures examination*. We offer that failing to have a simple way to triage accounts so that viable candidates for continued investigation can be identified may in part explain the reporting gap discussed by Hogan, Rezaee, Riley, Jr. & Velury (2008). Specifically, we will:

1. Introduce the decision making parameters that are considerations in covering audit risk. In this regard, we will examine the False Positive Error [FPE] and False Negative Error [FNE] jeopardies that are an inherent part of creating audit evidence by sampling and so impact screening. The FPE is how often the auditor uses extended procedures when in fact they are not likely to be warranted; the FNE is how often the auditor fails to make an extended procedures investigation when one was likely to be warranted. This discussion will establish the decision making context for the account screening protocol. In this section, we make the important distinction between *Sampling* and *Screening* protocols,
2. Suggest a simple and well researched measure that will form the centerpiece of our screening protocol regarding extended procedure sampling. Specifically, we will use the *First Digit Frequency Profile [DFP]* to form the screening metric. This DFP-measure which is, to say the least, "*en vogue*" is due to Newcomb (1881) and Benford (1938), [hereafter N-B], and was popularized in the audit context by Nigrini (1996),
3. Provide details on the history of DF Profiling and offer an alternative to the first digit frequency profile of Newcomb and Benford that uses the Benford validation datasets and so facilitates *Screening in the Identification* of candidate accounts,
4. Summarize the audit decision making issues as they play out in terms of the FPE and the FNE and relate this context to the *Sample Size Signaling Issue* in the use of parametric inferential methods for screening protocols,
5. Offer for the first time an account screening signaling protocol that is based upon previously published information on *Conforming* and *Non-Conforming* N-B datasets. In this protocol, we will use the *z-test: calculated* measure introduced by the seminal work of Nigrini (1996) as calibrated by the triaging point of the separation between the *Conforming* and the *Non-Conforming* datasets. Refer to this as the Z-test Account Screening Protocol.
6. Judge the Z-test Account Screening Protocol based upon: the False Positive and the False Negative Signals from: (i) the *Back-Cast Testing* of the triaging protocol, and (ii) the validation testing based upon a *Holdback* set of data, and finally
7. Recommend further steps that are needed to *Promote the Effective Use* of the Digital Frequency [DF]-screening profile in the audit context, discuss the limitations of our protocol, and offer suggestions to promote the hegemony of Account Screening.

2. The False Positive and the False Negative Errors: Control of the Audit Budget and the Auditor's Risk for Discovery Sampling & Account Screening

2.1 Discovery Sampling

Determining the size of a random sample that is to be used in the audit context for inferential statistical protocols is a study in balance. On the one hand, the auditor could take *very large samples* of a *large number of sensitive accounts*

and reduce the audit risk of failing to detect errors in the accounts that may create material misstatements in the accounts of the audit client; for this risk "minimization" strategy, the auditor benefits as the client usually pays in the billing process for these large samples. This is the False Negative Error control, failing to detect errors, and so believing that there are no material errors in the accounts when in fact there are such errors. FNE "control" can be effected by over-sampling. The companion error is the False Positive Error [FPE]: believing that there are material errors in the accounts when in fact that is not the state-of-nature. In this case, the auditor launches extended procedure investigations of accounts to look for evidence of errors when in fact the true state of nature is that there are no errors of a material nature. Disregard for the cost of committing the FPE produced by oversampling in service of minimizing the FNE jeopardy over time will find that the cost of the audit will not reflect the "reasonable" risk of the client and so effectively the auditor will have "*low risk and no clients*" due to the unrealistically high cost of the audit! These are the state of nature profiles that are in play in the audit context. These FPE and FNE issues are dealt with in the parameterization of the sample size models—usually *Discovery Sampling Models*. See for example, the very readable works of Ramos (2008, 350: *Audit Sampling*), Beasley, Elder, and Arens(2014, Chs *15*,*16* & *17*), and Lusk, Heilig & Halperin (2014). We have mentioned the logic underlying *Discovery Sampling* protocols to provide a context, in relief, for our study which is Account Profiling through *Screening*. Consider now Account Screening.

2.2 Account Screening

Screening protocols are often employed in tandem with Discovery Sampling Models. Simply put:

Screening Protocols look for "anomalous profiles" as a characterization of the entire account under audit scrutiny that suggest that there is reason to effect an extended procedures examination using Discovery Sampling Models. In this case then Screening is a Pre-curser to Discovery Sampling.

There is a dramatic sample size difference between discovery sampling and account profiling through screening. Experience shows that reasonably parameterized discovery sample size models often create sample sizes under 100 accounts. For example, assume that we are using the Ramos (2008) Discovery model and the Minimum Tolerable Error is set at 10% of the account value. If the confidence range set by the auditor is in the range: [80% to 99%], certainly in the practical ranges for most audits, the Ramos discovery sample size recommended is in the range: [17 to 47]. Such sample sizes are "practical" as the auditor MUST examine each of these randomly selected accounts in the *Occurrence* and *Completeness* directions—that is to say effectively tracing, vouching, re-calculating, and effecting re-performance protocols for each of the accounts in the discovery sample. This is another way of saying that a sample size in the thousands, each one of which is examined to service "best practices due diligence scrutiny", is beyond the practical economic boundary of the audit. However, for screening protocols, the auditor searches using ALL the individual account-items in the account under audit scrutiny so as to develop a *profile of the accounts* and then examines this profile to determine if *the profile is in the expected range*. If the expected profile is "close" to the observed profile then extended procedures are not likely to be warranted; alternatively, if the expected profile is "not-close" to the observed profile then extended procedures may be warranted. In the screening context, there are also comparable FPE and FNE concerns. For screening, there is the information from the screening signal; the signal sometimes suggests that the *Observed* account profile is sufficiently different from the *Expected* profile and so extended procedures are needed when in fact this is not the case: this is the *FP-signal error* in that extended procedures are used in investigating an account when in fact such an investigation is not warranted. The companion miscue is that sometimes the screening protocol does not signal that extended procedures are warranted when in fact they are needed; this is the *FN-signal error*. Clearly the FPE and the FNE are *en genre* the same: In Discovery Sampling they focus on the "existence of a material error in the accounts of the client" while for Screening the focus is on the use of extended procedures to "continue the investigation of the account" to determine if there are material errors in the account. This is to say that Discovery Sampling and Screening both address the extended procedures investigations in different ways which is why they are often used in tandem in the PCAOB-best practices world.

3. The First Digit Profile: An *en vogue* and Relevant Screening Measure

The historically available information on the Digit Frequency Profile [DFP] starts in 1881 when Simon Newcomb (1835-1909), Professor of Mathematics, Astronomer awarded *The Gold Medal of London's Royal Astronomical Society* [1874], and extensively published Political Economist [See: *American National Biography Online* :Newcomb, Simon, 19th century] notices a curious pattern of *Wear & Tear* of his logarithmic tables—his Decision Support System of the 19th century. He offers (1881, p.39):

That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener I than any other digit, and the frequency diminishes up to 9.

Newcomb's "curiosity-note" gathers archive dust for some fifty years until Frank Benford (1938), an electrical engineer with *General Electric Inc.* with many patents to his credit, who curiously never cites Newcomb, makes and records the same observation: Benford (1938, p. 551)

It has been observed that the pages of a much used table of common logarithms show evidences of a selective use of the natural numbers. The pages containing the logarithms of the low numbers 1 and 2 are apt to be more stained and frayed by use than those of the higher numbers 8 and 9.

Newcomb and Benford both arrived at a simple mathematical formula to characterize the likely distribution of the nine first digits. To wit the [N-B Profile]:

$$\text{Frequency}[d_i] = \text{LOG}_{10}\,(1 + \quad 1/d_i) \text{ for i= 1, 2, - - -, 9} \qquad (1)$$

Therefore, this simple formula, if it is the underlying generating process for digital frequencies in the Big Data milieu, can be used to benchmark particular *Observed* digital frequency profiles for the purpose of generating variance information that can pique the interest of the analyst to the possible end launching an extended procedures examination.

An important question which is begged by the "non-intuitive" observations of Newcomb and Benford is: *Why should EQ1 work as a general DFP- estimator of generating processes and under what conditions can an auditor reasonably expect to use EQ1 as a profiling-screen for the use of extended procedures?*

The first research to address the theoretic underpinning of the Log_{10} formula, EQ1, as a reasonable and appropriate surrogate for data generating processes only starts to appear some 25 years after Benford's paper. The first early ground breaking work is offered by Pinkham (1961), Adhikari and Sarkar (1968), Duncan (1969), and Raimi (1969). However, Hill (1995a,b & 1996) is usually credited with providing the conclusive theoretical support (Note 1) for the *Why*, *How*, and *When* questions posed above. *En bref*, Hill and Fewster (2009), show by convincing argumentation and illustration respectively that: *Where there are datasets formed from (i) many differentiable sources, or (ii) a kernel data-generating process with many variations that data generated subject to these various idiosyncratic constraints seems to follow the first digital pattern sketched out by the Log_{10} formula*. We shall term this as *Hill-Conformity*. In the appendix, we offer a summary of the reports in the literature that have suggested specific instances or examples of *Conforming* or *Non-Conforming* underlying generating processes.

The above is offered to set the stage for our analysis—to wit: *It seems that Newcomb and Benford were correct—albeit it took a little over 100 years to arrive at a proof of their digital-pudding*. The issue that we are addressing is: *How can the auditor employ the Digital Profile of an audit dataset to make the decision regarding the use of extended procedures?* To arrive at a workable protocol there are two critical issues that we will need to present: (i) A refinement of the Log_{10} expectation model, and (ii) a way to deal with the FPE sensitivity of large samples. Consider next an alternative to the Log_{10} model.

4. An Alternative to the Newcomb-Benford Profile: Log_{10} Profile

This alternative benchmark is due, in fact, to Benford. To give operational validity to the Log_{10} generating function, Benford (1938, Table 1, p. 553) collected 20 samples from an impressive spectrum of generating processes, such as: *River Areas* [as presented above], *Economic Costs*, and *Atomic Weights* to mention a few. The number of observations, in total, for these 20 datasets is 20 229. The range of the sample sizes for the 20 accruals is [91 to 5,000] with a mean of 1,012. Therefore, these frequencies as "a realization-profile" could also be used as a benchmark for the *Observed Digital Frequency* profile. However, due to recent research of Lusk & Halperin (2014a), it was reported that the mean frequency profile reported by Benford (1938, Table 1, p. 553) may be refined. They offer:

"Benford reports a curious measure: In Table 1, p. 553 we find the Average [Arithmetic Mean] for the 20 point measures for the first digit set. To produce this measure, he took the simple average of the point frequency measures as reported in Table 1. Actually, this has no useful statistical meaning. The correct frequency average is the number of digits in a particular first digital Bin as a ratio to the sample total of 20,229. For example, in Table 1, p.553 Benford has the frequency proportion average for "1" as: 30.6%. However, the correct frequency percentage is the unit-vector product of the 20 frequencies with the respective sample sizes for the first digit. This develops for the 20,229 observations that 5,849.295 were "1s" and the proper ratio is 5,849.295/20,229 = 28.9%."

This suggests that for the Newcomb-Benford profile the auditor could use the N-B Log_{10} generating function or could use the Lusk & Halperin (2014a) corrected means for the 20 samples that Benford reports. These are presented in Table 1 following:

Table 1. The Benford Practical Profile [BPP] and the N-B Log_{10} Profiles

First Digit Array	Corrected Means of Benford Datasets: BPP	The B-N Log_{10} Digital Profile	Difference [Col2 less Col3] Values
Digit 1	0.289 19	0.301 03	-0.011 84
Digit 2	0.194 62	0.176 09	0.0185 30
Digit 3	0.126 65	0.124 94	0.0017 10
Digit 4	0.090 61	0.096 91	-0.006 30
Digit 5	0.075 44	0.079 18	-0.003 74
Digit 6	0.064 31	0.066 95	-0.002 64
Digit 7	0.054 08	0.057 99	-0.003 91
Digit 8	0.054 87	0.051 15	0.003 72
Digit 9	0.050 52	0.045 76	0.004 76

These two benchmarks are effectively, and certainly not surprisingly, substantially similar; for example, the sum of the differences in Col 4 is 0.000 29 and the distribution of the signs is as equal as is possible. However similar these two digital profiles appear to be *in the aggregate*, individually, as orientations in Cartesian coordinate space, they are non-trivially different and so may be considered to be different conceptual benchmarking variables. Specifically, the average of the absolute value of the differences in Col 4 as a ratio to the average of the two benchmarks is 6.1%.

Our recommendation is to use the Benford Practical Profile [BPP] as it was derived over a large number of different "natural contexts" and so embodies the *desirable property of natural variation produced by many generating processes*. The Log_{10} is the only purely context free profiling model; if such a *point process* is important in the analysis then the auditor can use the Log_{10} expectation; however, as this is a point process it may invite the FP-signaling jeopardy. The next issue is the bane of statistical analyses: the sensitivity of most parametric methods to large sample sizes. Consider now this issue.

5. The Sample Size Issue in the use of Parametric Inferential Methods: Size does Matter in a Bad-Way

When the sample size is large most parametric inferential models suggest that there is a difference between the *Observed* and what is *Expected* when in fact the difference is relatively small and so practically the difference is not sufficient to warrant investigation BUT the inferential model signals the opposite. The classic model usually identified with the FPE signal sensitivity is the χ^2 inference model. For example, Cho and Gaines note (2007, p. 220):

Indeed, one can reject the null hypothesis for the very data that Benford used to demonstrate the accuracy of Newcomb's law. Of course, χ^2 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 (Ley 1996; Giles 2007).

It is interesting that the χ^2inference model seems to be the "only" model tagged as being "oversensitive" when in fact all parametric inference signaling models are prone to the FP-signaling jeopardy.

Therefore, as we will be using the following Nigrini (1996) parametric equations in our Digital Frequency Profile protocol:

$$s_i = \sqrt{\frac{(p_i)\ x\ (1-p_i)}{n}} \tag{2}$$

$$z = \frac{|(op_i - p_i)| - \left(\frac{1}{2n}\right)}{s_i} \tag{3}$$

Where: *n* is the number of observations in the dataset; op_i is the *Observed* proportion of digit *i* in the dataset, $p_i$is the *Expected* proportion of digit *i*. and$\left(\frac{1}{2n}\right)$ is the continuity correction.

we must be concerned about the signaling issues for our protocol that is formed around the Nigrini z-test information. An example to illustrate why one has to be concerned about sample size will be most helpful here.

5.1 Illustration of the Sample Size Calibration Issue

To demonstrate the signaling problem that large sample sizes creates, consider the work of Nigrini (1996) an aspect of which was an analysis of Interest Income reported in Federal Income Tax filings for the year 1988. The data collected and reported by Nigrini (1996) are presented in Table 2:

Table 2. FPE Issue Illustration using the Nigrini Interest Income Taxation 1988-Dataset

Digits	**Log10:Expected**	**Actual**	**Difference[E-A]**	z-Computed* n= 78 640	Two-tailed FPE[20%]
1	0.3010	0.3059	-0.0049	2.99	**S**
2	0.1761	0.1779	-0.0018	1.32	**S**
3	0.1249	0.1270	-0.0021	1.78	**S**
4	0.0969	0.0948	0.0021	1.98	**S**
5	0.0792	0.0778	0.0014	1.45	**S**
6	0.0669	0.0650	0.0019	2.13	**S**
7	0.0580	0.0563	0.0017	2.03	**S**
8	0.0512	0.0503	0.0009	1.14	NS
9	0.0458	0.0450	0.0008	1.06	NS
Checks	1.0000	1.0000	0.0000	N/A	N/A

*These z-calculations are not the same as reported by Nigrini(1996) (Note 2).

In this case, there certainly appears to be a variance from expectation *if one looks only at the z-test significance levels seven, of which appear to signal a variance from the N-B expected profile*. As anecdotal information, we usually give the profiles in Cols 2 & 3 to the students in our Auditing and Assurance Services course without the z-test information; and almost exclusively they report that this variance profile would NOT seem to warrant further testing. Then I give to them the z-test information; the result uniformly is confusion relative to the investigation decision. As one student observed a few years ago (*paraphrasing*): "*If I were worried about covering my ***, then I would investigate at least a little bit—to show good faith in case there were to be a problem.*" Later we discuss this anomalous result which is, of course, an artifact of the enormous sample size of 78 640; I point out that if we had a smaller sample size, say 10 000 still, practically speaking, an enormous sample, none of the z-computations would have attained a significance level less than 20%. This seems to drive home the point of the FPE jeopardy; we recommend it as a classroom exercise.

As for the FNE, we consider a dataset that manifestly is at variance from the N-B profile. This is the dataset offered by Hill (1998) where he asked a number of students to write down a random number. In a parallel fashion, we have given these frequency profiles [Col2 and Col3 of Table 3] to our Auditing & Assurance Services students and almost all the students indicate that the variance from expected is so extreme that this dataset would warrant investigation. Hill makes the same observation.

Table 3. FNE Illustration using the Hill Student Offering Guesses as to random Digit values Data Profile.

Digits	**Log10**	**Hill Actual**	**Difference[E-A]**	**z-Computed n= 35**	**Two-tailed FPE[5%]]**
1	0.3010	0.147	0.1540	1.80	NS
2	0.1761	0.100	0.0761	0.96	NS
3	0.1249	0.104	0.0209	0.12	NS
4	0.0969	0.133	-0.0361	0.44	NS
5	0.0792	0.097	-0.0178	0.08	NS
6	0.0669	0.157	-0.0901	1.80	NS
7	0.0580	0.120	-0.062	1.21	NS
8	0.0512	0.084	-0.0328	0.50	NS
9	0.0458	0.058	-0.0122	0.06	NS
Checks	1.0000	1.0000	0.0000	N/A	N/A

However, if the number of observations were to have been 35, certainly in the usual accrual spectrum for an audit, then there are NO indications of a statistically significant deviation at a p-value less than 5% while the reality is an investigation is likely warranted; and again the signaling issue is conditioned on the sample size.

Therefore, in summary the sample size is a critical feature in the inference signals relative to making an extended procedures investigation decision. The two issues this creates are:

5.2 Summary and Implications

If the sample size is large then the auditor invites the FP-signaling jeopardy and risks launching extended procedure examinations that are likely unwarranted. ***Implication***: The budget of the audit will be inflated beyond the actual risk level of the audit and the auditor may lose clients as the cost of the audit will be too high for the actual risk level of the audit.

If the sample size is small the inferential signals suggest that investigations are not warranted when they may be needed and so there is a FN-signaling jeopardy—failing to investigate when it is likely warranted. ***Implication***: Over time there will be difficulties for the auditor to justify the certification opinion written for the engagement and this risks a best practices exception by the PCAOB or worse yet an indication of a lack of due diligence in the execution of the audit.

With the above as important context information, consider now the z-test protocol that we offer as a useful tool in forming the decision *to use* or *not to use* extended procedures.

6. Likelihood Triaging: Judging the Likely *Conformity* for an Account under Audit Examination

By way of a focused summary of what underlies our development of a screening DFP model, as we are now informed as to the deleterious effects that large sample sizes have on controlling the FP-signaling error in the audit context, we have created a DFP-Signal protocol that *takes advantage of the FPE-signaling jeopardy*, ***as one cannot avoid it***. We use the FPE-Sensitivity to develop a likelihood triaging protocol so as to create information regarding the use of extended procedures—to wit, we ask the following question for the account under audit examination:

What is the likelihood that the observed DFP of the account under audit examination is consistent with [likely similar to] that of Non-Conforming Data or alternatively Conforming Data? If Non-Conforming then, extended procedure are likely to be warranted; if Conforming then, extended procedure are not likely to be warranted.

We will call this screening protocol: DFP Likelihood Triaging that we have programmed as a DSS in Excel™-VBA®. The Decision Support System is called: The ZTest:DSS and is available from the authors as a free download without restrictions to it use

6.1 The elements of the Functionality of the DFT Likelihood Protocol

The central montage uses:

1. ***Conforming* and *Non-Conforming* datasets**. Using these datasets an exclusive statistical partition can be developed to triage a particular dataset as: *Likely to be a Non-Conforming Dataset* or alternatively: *Likely to be a Conforming Dataset*. In this regard, we have collected from the literature 56 datasets: 33 of which have been argued as *Conforming* and 23 that have been offered as *Non-Conforming*. These dataset sets are presented in the ZTest:DSS.
2. **An Inferential Model** We will use the standard Test of Proportion equations used by Nigrini (1996) noted above as EQ2 and EQ3.
3. **The Triaging Partition** The screening metric used to triage a particular dataset will use the Nigrini z-test profile for the nine first digit frequencies of the audit dataset. Using the 65.7% cut-point recommended by Lusk & Halperin (2014a), to wit: if six or more of the nine digital frequencies have z-values greater than 1.96, the 95% confidence level that we *a priori* recommend, then this dataset will be assumed to be *Non-Conforming and so extended procedures may be in order*; otherwise it will be assumed to be *Conforming and so extended procedures may be warranted*.

6.2 Functionality

Using the elements of this screening montage, we determined the *minimum* sample size, as iterated by the ZTest:DSS, that first results in six of the nine z-calculations for the DFP of the account under audit to be greater than 1.96. Six was selected as 6/9 as a percent is greater than the 65.7% suggested as the measure of *Non-Conforming* datasets; we call this minimum sample size: *The Critical Sample Size* [CSS]. The profile of the CSS of the 56 datasets is presented in Table 4.

Table 4. Critical Sample Size Profile for the *Non-Conforming* and the *Conforming* Datasets

Statistical Profile of the CSS	***Non-Conforming* Datasets, n = 23**	***Conforming* Datasets, n = 33***
Mean/Median	1308/1300	12 964/7400
IQR/StDev	[700 - 1800]/702	[1850 -18 950]/13 545
95% CI	[307 – 1589]	[8161 – 17 767]

***** We stopped the iterations of the ZTest:DSS at a CSS of 40 000 as this is an extreme values for a sample; for a few datasets the actual CSS was greater than 40 000.

To be clear, the obvious expectation, as illustrated with the Hill and the Nigrini Tax data, is that *Conforming* Data such as the Nigrini Tax profile will require a larger sample size to drive at least six of the nine digital frequency profiles to have z-values greater than 1.96 compared to the *Non-Conforming* data such as the Hill profile. This is how we take advantage of the FPE sample size issue in the triaging of a particular account. To be clear: We know that to produce z-test values that are relatively large that: for a

Likely Conforming Profile: The dataset under examination requires a very large sample size to produce the precision needed to find at least six z-values larger than the 1.96 cut-off. ***Profile Implication***: the DFP of the data under examination must be rather similar to that of *Conforming* data. Example, The Nigrini Data: Table 2. ***Triage Logic: If the CSS is large, then the likelihood is that the Dataset is Conforming.***

Likely Non-Conforming Profile: The dataset under examination does not require a very large sample size to produce the precision needed to find at least six z-values larger than the 1.96 cut-off. ***Profile Implication***: the DFP of the data under examination must be rather dis-similar to that of *Conforming* data. Example, The Hill Data: Tables 3. ***Triage Logic: If the CSS is small, then the likelihood is that the Dataset is Non-Conforming.***

The CSS results are clear from Table 4; however, to justify the separation, as we will then use these datasets to form the likelihood partition, we conducted the usual inferential tests. Using the parametric [*Welsh Adjusted t-test version*] and the non-parametric [*Wilcoxon: Kruskal-Wallis Rank Sum Test*], the two tailed p-values were both < 0.0001 suggesting a strong rejection of the Null that there are likely no differences in the Critical Sample Sizes between the *Conforming* and the *Non-Conforming* datasets.

6.3 The Likelihood Triaged Partition

Given the evident and tested separation between the two reference datasets, we decided to use the most conservative triage point of separation—that is, the one with the smallest profile separation. This occurred for the Inter-Quartile Range; specifically, we took the mid-point of the ordered IQRs for the two datasets for: [The 75% percentile for the *Non-Conforming* Data: CSS = 1800] & [The 25% Percentile of *Conforming* Data: CSS = 1850] or 1825 [1800 to 1850]. Therefore the exhaustive Likelihood triage point will be:

For an audit dataset under examination,

If the Critical Sample Size <= 1825, then the likely set membership of this Audit Dataset is the Non-Conforming Dataset and therefore: ***Extended Procedures may be warranted****.*

If the Critical Sample Size > 1825, then the likely set membership of this Audit Dataset is the Conforming Dataset and therefore: ***Extended Procedures may not be warranted****.*

As an illustration of this triage partition consider the Nigrini and the Hill datasets as found in Tables 2 & 3. Recall that we used the frequency of the BPP as the expected value. The iteration results for the Critical Sample Size [CSS] and the respective z-calculations are presented in Table 5.

Table 5. Critical Sample Size for the Hill *Non-Conforming* and Nigrini *Conforming* Dataset

Digits	BPP Expectation	Hill Actual	Nigrini Actual	z-Computed Hill Data	z-Computed Nigrini Data
1	0.289 189	0.147	0.3059	**5.086 120**	**7.399 714**
2	0.194 770	0.100	0.1779	**3.855 329**	**8.552 194**
3	0.126 650	0.104	0.1270	1.027 566	0.203 951
4	0.090 612	0.133	0.0948	**2.320 368**	**2.922 505**
5	0.075 436	0.097	0.0778	1.226 473	1.789 001
6	0.064 314	0.157	0.0650	**6.084 336**	0.551 690
7	0.054 081	0.120	0.0563	**4.654 445**	**1.960 105**
8	0.054 872	0.084	0.0503	**1.968 087**	**4.022 627**
9	0.050 522	0.058	0.0450	0.422 095	**5.054 065**
Total Checks	1.0000	1.0	1.0	CSS=270	CSS=40 165

These results in Table 5 are the anecdotal demonstration of the rational for a Likelihood partition of 1825. The **bolded** cells in Table 5 identify the specific digits for which the CSS produced z-values greater than the 95% Confidence Level. It is of course expected, as argued and tested above, that the number of z-test values greater than the critical value of 1.96 will occur at a smaller Critical Sample Size for the Hill profile than for the Nigrini profile. We see for the Hill data that a sample of 270 is sufficient to produce six z-values greater than 1.96. At the other end of the likelihood spectrum the Nigrini data fits rather closely to the BPP; the number of items in the CSS computation needed to produce at least six z-values greater than 1.96 for the Nigrini data is 40 165. For a computational illustration of the Hill data, we have made the z-calculation for the last digit to be driven past the critical z-value:

Hill Computation[Digit 8]

$$0.013\,859\,221 = \sqrt{\frac{(0.054\,872)\,x\,(1-0.054\,872)}{270}} \tag{4}$$

$$1.968\,087 = \frac{|(.084-(0.054\,782))| - \left(\frac{1}{2x270}\right)}{0.013\,859\,221} \tag{5}$$

The test of this likelihood cut-off is, of course, how it performs in the re-classification of the 56 datasets. In this regard we will use the ZTest:DSS to classify the benchmarking 56 series. Using the usual [2 x 2] Classification/Misclassification matrix we will examine the functional profile of the Triage Protocol.

7. Evaluation of the Functionality of the Triaging Protocol

7.1 FPE and FNE Evaluation

The next issue that we wish to discuss is the performance of the likelihood cut-off in terms of the FPE and the FNE. This is a natural question as the auditor is interested in how the signaling or triaging protocol works relative to these classification errors. For the FNE, failing to investigate when it is likely warranted, we tested the 1825-classification protocol, as programmed in the ZTest:DSS, on all 23 of the *Non-Conforming* datasets; if the ZTest:DSS signals that the dataset is *Conforming* as the critical sample size is greater than 1825 then this misclassification would be a FNE—as the signal generated by the ZTest:DSS is to not use extended procedures to investigate when the likely state of nature is that the dataset is not *Conforming*. In a like manner, we ran all 33 of the *Conforming* datasets where the misclassification is akin to the FPE— as the signal generated by the ZTest:DSS is to use extended procedures to investigate when the likely state of nature is that the data set does Conform and so extended procedures are not likely warranted .

7.2 ZTest:DSS Classification Results

Here we are using the 56 datasets for which we developed the 1825 triage point to evaluate the classification performance of the DFT Likelihood Protocol. As this was the developmental dataset this evaluation is called

Back-Casting. It provides a slightly favorably biased picture of the classification effectiveness as this was the dataset that we used to form the triage point. Subsequently, we will use a Hold-Back dataset that will provide validation of the triage model. Using the Back-Cast evaluation, for the *Conforming* Data, in eight of the 33 cases the critical sample sizes were lower than or equal to 1825 consistent with *Non-Conformity* which in turn incorrectly suggests extended procedures. In this case, this is incorrect as the dataset was *a priori* judged to be *Conforming* and so the auditor commits a FPE 24.2% [8/33] of the time. As for the FNE—failing to investigate when warranted, we ran the 23 *Non-Conforming* datasets using the ZTest:DSS and four times of the 23 the critical sample size was greater than 1825 indicating that the dataset is likely to be *Conforming* and so extended procedures are not warranted. In this case this is a FNE—i.e., failing to investigate what it may have been warranted which occurs 17.4% [4/23] of the time. This produced the following Classification Matrix:

Table 6. FPE and FNE for the Back-Cast evaluation for the 56 Datasets

	State of Nature ***Conforming***	**State of Nature** ***Non-Conforming***
ZTest:DSS ***Conforming***	25	4 [FNE:17.4%]
ZTest:DSS ***Non-Conforming***	8 [FPE24.2%]	19

The statistical profile of the Classification matrix is that the Likelihood Ratio is 19.8, with a p-value < 0.0001. Also, using the Fisher's Exact test for a conservative test of the classification, i.e., the 2-tailed version, the p-value is also < 0.0001. Both indications provide strong support for rejecting the Null that the classifications made by the DFT Likelihood Protocol as programmed in the ZTest:DSS are random and that this particular realization is the exception. In summary, as the FPE and FNE are unavoidable and the issue is controlling them in the audit context, we suggest that as both are less than 25%, although slightly biased away from the Null, that the ZTest:DSS triaging forms a reasonable labeling DSS for the use of the extended procedures. Now consider the Hold-Back validation.

7.3 Holdback Validation

As a validation test of the DFT Likelihood Protocol effectiveness, we collected a 25% holdback sample: Seven datasets that were offered in the literature or as judged as part of a certified audit as *Conforming* and seven that were reported as *Non-Conforming*. We then used the ZTest:DSS to classify these Hold-Back datasets. In Table 7 are presented the results:

Table 7. Hold-Back Validation

	State of Nature ***Conforming***	**State of Nature** ***Non-Conforming***
Results:ZTest:DSS ***Conforming***	7	0 [FNE0%]
Results:ZTest:DSS ***Non-Conforming***	0 [FPE:0%]	7

In this case, the Likelihood Ratio is 19.4 with a p-value of < 0.0001; the directional Fisher's p-value is 0.0003. This strongly supports the classification effectiveness of the DFT Likelihood Protocol and is consistent with the Back-Cast results. As a final issue we wish to address the robustness of the ZTest:DSS over the sensitivity of the confidence interval classifications that one finds in practice.

7.4 Robustness of the 1825:Triaging Protocol

We used the 95% Confidence in triaging the benchmark 56 series as well as the Hold-Back testing. As indicated above, this means that we tracked the sample size to the point where the 6^{th} z-calculation became greater than 1.96. As a test of the robustness sensitivity one could ask: *If the auditor were to use other confidence calibrations and maintained the cut-off of 1825 would there be important differences in the FNE and FPE profiles.* To create this sensitivity information, we ran the same classification tests for the following usual confidence levels that seem to be prevalent in practice: 80% [1.28], 90% [1.645], 95% [1.96] and 99% [2.33]. We recorded for each of the 56 benchmark series the classification of the 1825:Triaging model for each of these four levels of confidence. For

example, for the dataset presented in Table 8 we found the following [using an iteration increment for the ZTest:DSS set at 100 rather than 5]:

Table 8. Sensitivity Test using the N-B Log_{10} model and the Hill Student Guessing Datasets

	80% Confidence	90% Confidence	95% Confidence	99% Confidence	Classification Error
Log10 CSS	6000	9700	13 600	19 100	0%
Hill CSS	200	200	300	400	0%

Therefore, the 1825 screen for these two datasets was robust because for the 80% Confidence Level the critical sample size for the N-B Log_{10} was not less than or equal to 1825 which would have resulted in the N-B dataset to be incorrectly classified as a *Non-Conforming* dataset; and, for the Hill dataset the sample size for the 99% confidence level was greater than 1825 which would result in the Hill *Non-Conforming* dataset to be incorrectly classified as a *Conforming* dataset.

We conducted this sensitivity analysis for all the 56 Back-Cast as well as the 14 Hold-Back datasets. We then computed the FPEs and the FNEs for the *Conforming* and the *Non-Conforming* datasets. For the FNE, we found that for the *Non-Conforming* datasets, n= 30 for which there were 120 classifications over the four confidence levels [30 x 4] that 16 times the triage model of the ZTest:DSS incorrectly classified the *Non-Conforming* series as *Conforming* because the critical sample size was greater than 1825. This is a misclassification of 13.3% [16/120]. This would be a FNE as an investigation is not signaled when it was likely warranted. This basically was the same as we found for the 95% confidence level.

As for the FPE—investigating when it was not warranted—for the 40 *Conforming* datasets over the four confidence levels there were 35 instances where the critical sample size was less than or equal to 1825 and so indicated that the dataset would likely be *Non-Conforming* which is a FPE of 21.9% [35/160] which again is almost the same as it was for the 95% confidence level. The Classification Table for this robustness sensitivity screening is:

Table 9. Sensitivity Screening: Aggregated Results for the Four Confidence Tests

	State of Nature *Conforming*	State of Nature *Non-Conforming*
Results:ZTest:DSS *Conforming*	125	16 [FNE:13.3%]
Results:ZTest:DSS *Non-Conforming*	35 [FPE:21.9%]	104

In this case, the Likelihood Ratio is 125.8 with a p-value of <0.0001 which is also true for the comparable Fisher's p-value. This strongly suggests that robustness relative to usual confidence levels seems in evidence for the 1825 Triaging Protocol.

8. Summary, Conclusion, Limitations, and Outlook

8.1 Summary

Using the DFT Likelihood Protocol the auditor will execute the following steps:

1. Select the dataset under audit consideration,
2. Use the ZTest:DSS [Tab: ***CreateProfile***] to form the DFP using all the data points in the account under audit examination,
3. Using this profile, the ZTest:DSS[Tab:***ZTest***] can be used to iterate this DFP profile to arrive at the Critical Sample Size[CSS]. Specifically, starting at a sample size of 100 and incrementing the sample size by 5 for each VBA-iteration, the z-values for ALL nine digital frequencies for the DFP are computed using EQ2 and EQ3. When the sum of z-values greater than 1.96 [*the default* of the four confidence level in the *UserForm*] is first equal to or greater than six, then the ZTest:DSS stops and reports the final CSS.

If that CSS is greater than the triage cut-point of 1825 then the ZTest:DSS produces a Cell Message: *This dataset is Likely to be a CONFORMING Dataset. Extended Procedures are NOT Indicated.* If the CSS is less than or equal

to 1825, then the Cell message is: *This dataset is Likely to be a NON-CONFORMING Dataset and so Extended Procedures Are Indicated.*

This information may then be used by the auditor to make the decision regarding the use of extended procedures.

8.2 Conclusion

The ZTest:DSS is enabling software and offers for the first time an account screening protocol based upon the Nigrini equations. The decision to use or not to use extended procedures is, of course, the purview of the audit in-charge. We strongly suggest that the auditor record and save as part of the current audit evidentiary file the results of ALL the CSS calculations and that this information would be linked to the specific dataset under audit examination. In this way, the Ztest:DSS and the related test information will become a part of the *permanent file* and so be available for future audits. *As a caveat*: The ZTest:DSS is focused on the state of nature where there is an expectation that the dataset under audit examination is expected to conform in a Hill sense; there is of course the alternative perspective—where the state of nature produces *Non-Conformity*. We shall consider this next.

8.3 Limitations

The DF-Profiling literature has approached the Digital Frequency Testing from the perspective that the Log_{10} or the BPP are the Normative benchmarks; in this context, the *a priori* view is that *Non-Conformity* is a perturbation that effectively signals that the account under audit merits investigative consideration. *Conformity* as the "normative state of nature" is actually not consistent with the fact that in many situations that are found in the audit context *Non-Conformity* may be the Norm. For example, some firms have organizational protocols that affix ordered administrative limits that require tiers of signatures to create vouchers. See Nigrini (1999) & Reddy and Sebastin (2012). So as to aid the auditor in developing a realistic *a priori* expectation relative to what is the likely state of nature, we have collected a number of indications as to when *Conformity* or *Non-Conformity* may be expected or the "normative" state. This compendium, presented in the Appendix [*Literature References for Examples of Conforming or Non-Conforming Data Generating Processes: The GPS for the Initial Analytic Procedures Phase of the Audit*], is essential in the execution of the audit and should form part of the Analytic Procedures' "brainstorming session" where the plan of the audit is developed. This is critical as the *Non-Conforming* datasets that are found in the literature are from "likely" perturbed processes. Therefore, if the auditor fails to position the screening protocol in the likely true state of nature then there will be misclassifications resulting in either FPEs or FNEs when the triage protocol is working correctly! Additionally, we have not explored the scoring of datasets for other than the first digit. The classification results may differ if the digital frequency screening protocols are based upon the first two or three, or, for that matter, the last monetary digit. The last monetary digit has played a major role in misappropriations of small amounts of money where there is a very high frequency "turn-over" in the account—such as imprest petty cash funding. This has been called: *salami slicing*, *cents shaving*, or *take a little leave a lot* scams. Over time such small misappropriations can add up to a major amount of money being stolen. See Keys (1994). Finally, as this paper has developed and reported a screening protocol perhaps unscrupulous individuals will use the Z-testing account screening protocol to screen their perturbed datasets and then interactively make selective changes so that the perturbed dataset is modified in such a way as to NOT be screened as an exception thus turning the Z-test screen to their advantage. A related and certainly parallel context exists for linguistic content analysis screening that has been used to determine if the MD&A section of filed 10Ks have reporting irregularities or indications of fraud. One may posit that as the linguistic screens used in Content Analysis [CA] are in fact public information that such CA-screening may be used to the end of avoiding the screening signal. In this regard however, Lee, Lusk & Halperin (2014) have examined this possibility and report:

The effectiveness of content analysis of MD&A has not been diminished in the SOX-era. This finding is intriguing and speaks to the pervasive and consistent nature of the linguistic traces left in the MD&A section even given strengthened regulations as embodied by SOX and enforced by the PCAOB.

This suggests that as content analysis still is a viable screening tool that it has not been widely used to tailor information to avoid detection. One supposes that the same will be true of the Z-test account screen protocol suggested above. But of course this assertion needs to be tested.

8.3 Outlook

Given the critical need for account screening as a pre-curser to the sample selection of accounts, we suggest that the Z-test Account Screening protocol as previewed in the ZTest:DSS be included in the panoply of the auditor. We have used it productively in our role as an academic consultant to Audit and Assurance firms in the USA and in Europe. It is, of course, obvious that the various parameters of the Likelihood triage are dependent on the two dataset profiles;

Conforming and *Non-Conforming*. To enrich the relevance and utility of the ZTest:DSS, we hope that *Conforming* and *Non-Conforming* datasets would be made publically available so as to up-date the Z-test Likelihood-Triaging protocol. We have begun the collection of such data with the cooperation of the US Audit LLP to which we are the academic consultant; in this endeavor, we are focusing on audit datasets determined in the course of the audit to be "relatively" error free *but that z-tested to be Non-Conforming*. There is a dearth of such data. We would be happy to be a public domain posting source for such naturally occurring *Non-Conforming* audit tested information

Acknowledgments

We wish to thank Dr. H. Wright, *Boston University*: Department of Mathematics and Statistics, the participants at the Research Workshop at the *State University of New York: Plattsburgh* for their most helpful suggestions. Additionally, we benefited from the comments and detailed suggestions of two anonymous reviewers of *Accounting and Finance Research*. Finally, we wish to thank Mr. Nicolae Lungu and Mr. Tom Stretton, CPA: *McSoley and McCoy LLP*, [http://www.cpavt.com/] for their council and support in the development of this project.

References

Adhikari, A.K. & Sarkar, B.P. (1968). Distributions of most significant digit in certain functions whose arguments are random variables. *Sankhya-The Indian Journal of Statistics* Series B, *30*, 47–58.

Allaart, P. C. (1997). An invariant-sum characterization of Benford's Law. *Journal of Applied Probability*, *34*, 288–291. http://dx.doi.org/10.2307/3215195

Beasley, M., Elder, R, & Arens, A. (2012). *Auditing and assurance services* (14th ed.). Pearson Publishing, New York, NY USA: ISBN: 13-978-013-257609-3

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

Diaconis, P. (1977). The distribution of leading digits and uniform distribution mod 1. *Annals of Probability, 5*, 72–81. http://dx.doi.org/10.1214/aop/1176995891

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.

Duncan, R. (1969). Note on the initial digit. *Fibonacci Quarterly*, *7*, 474–475.

Fewster, R.M. (2009). A simple explanation of Benford's Law. *American Statistician*, *63*, 26-32. http://dx.doi.org/10.1198/tast.2009.0005

Giles, D.E. (2007). Benford's Law and naturally occurring prices in certain eBay auctions. *Applied Economics Letters, 14*, 157–161. . http://dx.doi.org/10.1080/13504850500425667

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.

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.

Hogan, C., Rezaee, Z., Riley, Jr., R. & Velury, U. (2008). Financial statement fraud: Insights from the academic literature. *Auditing: A Journal of Practice & Theory* American Accounting Association, *27*, 231-252. DOI: http://dx.doi.org/10.2308aud.2008.27.2.231

Kassem, R. & Higson, A. (2012). Financial reporting fraud: Are standards' setters and external auditors doing enough? *International Journal of Business and Social Science, 3,* 283-296.

Keys, E. Jr. (1994). Roundtable. *The Internal Auditor, 51*, 69-74.

Lee, Chuo-Hsuan, Lusk, E. & Halperin, M. (2014). Content analysis for detection of reporting irregularities: Evidence from restatements during the SOX Era. *Journal of Forensic and Investigative Accounting*, *6*, 99-122 http://www.bus.lsu.edu/accounting/faculty/lcrumbley/jfia/Articles/v6n1.htm

Ley, E. (1996). On the peculiar distribution of the U.S. stock indexes' digits. *American Statistician, 50*, 311–313. http://dx.doi.org/10.1080/00031305.1996.10473558#.Uq9HSdJDvTk

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

Lusk, E & Halperin, M. (2014b). Detecting Newcomb-Benford digital frequency anomalies in the audit context: Suggested χ^2 Test Possibilities. *Journal of Accounting and Finance Research*, *3*, 45-66.

Lusk, E., Heilig, F. & Halperin, M. (2014). Prevalence profiling: A judgmental context for evaluating initial sample size projections in the audit context. *International Journal of Auditing Technology, 2*, 1-21. http://dx.doi.org/10.1504/IJAUDIT.2014.064314

Marchi, de S. & Hamilton, J. (2006). Assessing the accuracy of self-reported data: An evaluation of the toxics release inventory. *Journal of Risk and Uncertainty, 13*, 57–76. http://dx.doi.org/10.1007/s10797-006-6666-3

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.

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

Pinkham R. (1961). On the distribution of first significant digits. *Annals of Mathematical Statistics, 32*, 1223 -1230. http://dx.doi.org/10.1214/aoms/1177704862

PricewaterhouseCoopers (PwC) (2005). *Global Economic Crime Survey*. Available at: http://www.pwc.com.

Raimi, R. (1969).The peculiar distribution of first digits. *Scientific American, 221,* 109–120. http://dx.doi.org/10.1038/scientificamerican1269-109

Raimi, R. (1976). The first digit problem. *American Mathematical Monthly, 83*, 521–538. http://dx.doi.org/10.2307/2319349

Rauch, B., Göttsche, M., Brähler, B & Engel, S. (2011). Fact and fiction in EU-Governmental economic data. *German Economic Review*, *12*, 243-255. http://dx.doi.org/10.1111/j.1468-0475.2011.00542.x

Reddy Y.V. & Sebastin, A. (2012). Entropic analysis in financial forensics. *The IUP Journal of Accounting Research and Audit Practices*, *11*, 42-57.

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

Notes

Note 1. The proof of the N–B observation as a general data generating process is challenging. In speaking of Hill's theoretical proof of the First Digit Phenomena, Ken Ross, former president of the Mathematical Association of America, (2011, p. 571) offers: "*The most successful seems to be Theodore Hill's very nice, but sophisticated, analysis in Hill(1996). I found his analysis challenging, and I am reasonably acquainted with probability*."

Note 2. While the differences between our calculations and those reported by Nigrini(1996, Table 2:Panel A, p.80) are small and do not change the inferences, to be sure, they do seem to create issues for the students when they try to replicate the values Nigrini (1996) reported. For this reason, we reported the z-calculations that we generated using Excel and EQ2 and EQ3.

Appendix Literature References for Examples of *Conforming* or *Non-Conforming* Data Generating Processes: The GPS for the Initial Analytic Procedures Phase of the Audit

General Conditions for Observing *Conforming* Datasets	Selected References
Mixing Property - - - *if distributions are randomly selected and random samples are taken from each of the distributions, then the frequency of digits of this combined set will converge to Benford's distribution even if the separate distributions deviate from Benford's distribution*. Cho & Gaines (2007, p.219).	Bradley and Farnsworth (2009); Cho & Gaines (2007); Fewster (2009); Hill (1995a,b, 1996 & 1998); Ross (2011)
Base Invariance Property *Under certain restrictions, if the distribution of a random quantity remains unchanged under a change in scale (e.g., changing from miles to kilometers), then observations of that random quantity will follow Benford's Law. This is a desirable 'invariance' property in that, like any natural law, it indicates the measurement scale should not dictate whether or not a significant digit law holds for a particular data set*. Bradley and Farnsworth (2009, p.4)	Allaart (1997); Bradley and Farnsworth (2009); Hill (1995a,b, 1996 & 1998); Ross (2011)
Examples of Processes **Likely** to generate *Conforming Datasets*	
Fun with Numbers Exercises: We use in our course three such series: *Uniform* Unit-Random Numbers raised to integer powers.[**AS**], *Geometric*(k) Processes. We use f(k) = 3^k.[**D : Di**], & *Fibonacci* series [F(k), k = 102 (trials starting at 0,1)].[**R**]; *Datasets aggregated over many different sources*: Population counts over many counties.[**N : F**]; *Datasets influenced by many changing factors*: Stock Indices.[**L : NM**]; *Numbers that result from mathematical combination of numbers:* Basically transactional AIS data (e.g., quantity x price).[**DHP : RS**]; and finally, *Large datasets or data with positive skew[Mean >> Median]*.[**DHP**] Also, of course, examine the 20 datasets that Benford collected (1938, Table 1, p. 553); they are most instructive.	
Examples of Processes **NOT** Likely to form *Non-Conforming Datasets*	
Fabricating Data: Toxic release data.[**MH**], and Bidding conspiracy.[**G**]; *Boundary-Level Data*: Prices of low cost goods, usually set at xx.99.[**LH : DHP**], and Geo-political boundaries.[**F**]; *Administrative limits*: Institutional constraints on spending.[**N : RS**]; *Forms used for Accounting Control*: Checks, Issuing firm invoices & Pre-numbered forms.[**DHP : RS**]; *Random or Uninstructed Guessing*: Individuals attempting to replicate randomness.[**H**]; *Screening or strategic modifying of legitimate transactions*: Kickbacks and Lapping[Individual Postings].[**RS**]; *Manipulation of data to facilitate reporting*: Rounding. [**RGBE**]; *Small datasets*.[**DHP : NM**]	
Reference Legend: [**AS**]Adhikari & Sarkar (1968); [**D**]Duncan (1969); [**Di**]Diaconis (1977); [**DHP**]Durtschi, Hillison & Pacini (2004); [**F**]Fewster (2009); [**G**]Giles (2007); [**H**]Hill (1998); [**L**]Ley (1996); [**LH**]Lusk & Halperin (2014b); [**MH**]Marchi & Hamilton (2006); [**N**]Nigrini (1999); [**NM**]Nigrini & Mittermaier (1997); [**R**]Raimi (1976); [**RS**]Reddy & Sabastin (2012); [**RGBE**]Rauch, Göttsche, Brähler& Engel (2011).	