International Business Research; Vol. 7, No. 2; 2014
ISSN 1913-9004 E-ISSN 1913-9012
Published by Canadian Center of Science and Education

Detecting Digital Frequency Anomalies as Benchmarked against the Newcomb-Benford Theoretical Frequencies: Calibrating the χ^2 Test: A Note

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

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

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

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

Correspondence: Edward J. Lusk, The State University of New York (SUNY) at Plattsburgh, Plattsburgh, NY, USA. Tel: 518-564-4190. E-mail: luskej@plattsburgh.edu

Received: December 14, 2013 Accepted: December 28, 2013 Online Published: January 23, 2014

doi: 10.5539/ibr.v7n2p72 URL: http://dx.doi.org/10.5539/ibr.v7n2p72

Abstract

Digital Frequency Testing (DFT) using the Newcomb-Benford (N-B) profile is a simple and potentially effective tool that can be used efficiently in the certification audit as well as in the service of forensic investigations to rationalize the use of extended audit procedures at the substantive phase of the audit. In creating the audit evidence needed to rationalize the use of extended procedures for specific accounts under examination the χ^2 test applied to the N-B profile is a logical choice as it is fundamentally a variance measure for the difference between the *Expectation* and the *Observation*. However, there are anomalous sample size issues in that large sample sizes create a proliferation of False Positive Errors—in the audit context this means using extended procedures when in fact they are not likely to be warranted. In this paper we offer a validated sample size range for using the χ^2 as the N-B screening test. We discriminate tested this suggested sample size range using datasets that conform and that do not conform to the Newcomb-Benford profile. These datasets are detailed in the paper. Finally, we have a Decision Support System that aides in the creation of this χ^2 screening information. The DSS, programmed in Excel™ VBA®, is available from the corresponding author as a free download with no restriction to its use.

Keywords: large sample sizes, conforming and non-conforming datasets, benchmarking

1. Introduction

The market trading world circa 1992 changed fundamentally. The *Internet* went world-wide and *at that moment* a new trading sector was created: the *dot.com* enterprises. As detailed in *The Report on Financial Oversight of Enron* (2002) this event started an interesting chaining of events that eventually lead to:

1. Almost a complete "melt-down" of the world's financial trading markets,
2. The overnight evaporation of hundreds of billions of market capital,
3. Disclosure of scandalous, illegal, and unethical behavior of corporate executives in major corporations such as: *Enron*, *WorldCom*, and *Quest* to mention just a few,
4. Crashing the world's largest and the most respected public accounting firm, *Arthur Andersen*, LLP,
5. Passage of US federal legislation (Pub. L. 107–204, 116 Stat. 745) called *Sarbanes Oxley 2002,* and finally,
6. The Public Company Accounting Oversight Board (PCAOB) to take up the sorely lacking oversight of: The Board of Directors and Internal Audit Groups of most corporations, The SEC, US Department of Justice and, certainly, The Public Accounting Profession.

All of these events can be traced directly to the *Lack of Control over Financial Reporting* which is, not surprisingly, the major focus of the rules and regulations of the PCAOB and so therefore must be the underpinning of the audits effected by independent public accountants in discharging their reporting responsibilities to provide, based upon their audit evidence, *reasonable assurance in the quality and reliability of the information offered by firms regarding the results of their operations as reported in their financial statements*.

To develop the evidence to defend these assurance opinions, there are three phases to the certification audit: *Analytic Procedures* or Planning, *Interim-Testing*, where usually the internal control over financial reporting is tested, and the *Substantive* where the year-end balances in the accounts are substantiated. By way of a detail summary of the testing focus envisioned by the PCAOB, we offer a direct quote of: PCAOB: AS 5 (2007).

In Appendix B *Special Topics*, p. A1–45, Sub-Section B4: *Tests of Controls in an Audit of Financial Statements* the PCAOB offers the following:

> To express an opinion on the financial statements, the auditor ordinarily performs tests of controls and substantive procedures. The objective of the tests of controls the auditor performs for this purpose is to assess control risk. To assess control risk for specific financial statement assertions at less than the maximum, the auditor is required to obtain evidence that the relevant controls operated effectively during the *entire period* upon which the auditor plans to place reliance on those controls.

As part of the required client-risk assessment the auditor initially assesses the likelihood that: *there are errors of a material nature that would compromise the inferences drawn from the financial statements*. Then during the audit, evidence is collected and evaluated to dynamically re-assess this risk or likelihood level. In this regard, the auditor uses the collected audit evidence to decide if the evidence suggests anomalies the nature of which requires the auditor to examine this particular account "in greater detail". This is called the "*extended procedures decision*"; simply stated: If it appears to the auditor that the collection of audit evidence uncovers issues that may indicate that there are errors of a material nature in the account under investigation, then the auditor, in considering all the current information, decides ***Yes***: *All things considered, we, the public accountants, will extend the testing in this particular case*. OR ***No***: *All things considered, extended procedures do not seem to be warranted.* Usually the use of extended procedures needs to be justified to the manager, partner and ultimately to the client as extended procedures will increase cost of the audit.

1.1 Extended Procedures and Best Practices

Over the last decade, the careful monitoring, review, and evaluation by the PCAOB of the audit compliances required by AS 5 regarding (i) testing of *Controls over Financial Reporting* and also (ii) testing of *Year-End Accounts Balances* required at the substantive phase has resulted in a number of audit testing enhancements relative to making the important decision to use extended procedures for the particular account under examination. Basically these audit testing enhancements collectively are considered as the "best practices" profile or benchmark against which the PCAOB judges the quality of the audits being conducted. Some of these current enhancements which are now part of these "best practices" enhancements are: (i) refined ***Statistical Sampling*** routines in particular Event Discovery Sampling which addresses the COSO dimension of the audit, Dollar Value Sampling usually focused on the Substantive Testing Phase, See for example, Ramos: Wiley Practitioner's Guide to GAAS (2008), (ii) Powerful, efficient, and well-designed GUI-oriented ***Audit Software Systems*** such as the IDEA™ Audit Software (http://www.caseware.com) which has more than 40 platforms that enhance the creation of audit evidence, (iii) ***Content Analysis*** where linguist screening of information in the 10-K MD&A sections has been clearly shown to detect reporting irregularities including fraud. See for example the recent work of Lee, Lusk and Halperin (2014), and (iv) the ***Digital Frequency Testing*** (DFT) protocols in particular those using the Newcomb-Benford benchmark due to: Newcomb (1881) and Benford (1938). The latter audit screening DFT protocol is the point of departure for our paper which is focused on the extended procedures decision and the creation of audit evidence. Consider now The Newcomb-Benford Test Perspective.

1.2 Digital Frequency Testing: The Practical Context

In the certification audit, the fundamental analytic evidential signal that the auditor uses to decide if extended procedures are needed in a particular case is: **Observed VARIANCE from *a-priori* EXPECTATION**. This "Directed Differential" is essentially the Audit GPS at all of the three stages of execution of the audit mentioned above. One of the best practices techniques that has achieved its rightful place in the panoply of the Auditor is Digit Frequency Testing (DFT). Typically, the accounts subjected to DFT are those that affect the (i) Current Position of the Firm under audit, and (ii) The Cash Flow from Operations. Accounts usually targeted for testing are: Cash, Accounts Receivable, Inventory, Sales and Accounts Payable. Needless to point out that there are MANY audit tests that are applied to these accounts at all of the stages of the audit each of which impacts the various Management Assertions.

Relative to DFT: The auditor assumes a usual frequency of the distribution of digits: **The Expectation** and then collects information on the actual distribution: **The Observed**; if there is an important variance then, extended procedures are usually in order. *An anecdotal example*: If one were to take sample of retailed items priced under

100 *in the currency monetary units offered in major retail outlets*, the most frequent last digit in the cents place would be: "9". The same is true of the second digit in the cents place. Specifically, we took a "random" sample, n = 10 of such items each from: *Wall Mart*, Plattsburgh, NY USA, *Macy's* New York City, NY USA, *Karstadt*, Magdeburg, Germany and *KDW* (*Kaufhaus des Westens*), Berlin, Germany. Of the 40 sampled items 92.5% (37 of 40) ended in "xx.99". Therefore, if the auditor took a random sample from a listing of Year-End [YE] inventory items and sorted the random list into the binary partition: Group A: retail value greater than or equal to 100.00 and Group B: items valued at less than 100.00 and observed that the frequency of items of Group B ending in ".99" was 36.8%, this would merit extended procedures, because the frequency varies significantly from the usual expectation—i.e., most of the items should end in .99 as we found in our anecdotal sample.

Perhaps, the auditor could accept as a "reasonable" variance if the frequency of the Group B accounts was 90% but certainly not 36.8%! Best practices attention-to-detail would require the auditor to inquire how such a marked deviation from expectation of 53.2% (90%–36.8%) could have occurred. This audit evidence then sets the auditor off in the direction to determine the likely cause(s) from the various possible causes of this variance from expectation and so the need to use extended procedures in this Inventory case.

Therefore the digital frequency profile (DFP) of audit accounts and transaction sets is one aspect of typical audit evidence of interest in the best practice of the audit relative to rationalizing the decision regarding electing to use extended procedures. This being the case, auditors have discovered how to use the simple observation of Newcomb (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.*"

Later Benford (1938) reported the same general observation. Refer to this as the (N-B Profile). This lead to positing the following *theoretical expectation* for the distribution of first digit frequencies:

$$Frequency(d_i) = LOG_{10}\ (1 + 1/d_i)\ for\ i= 1, 2, - - -, 9 \quad (1)$$

Therefore, if the auditor needed to have a "test-against-expectation-profile" then EQ1 would be a logical choice. For example, assume that the auditor has the Accounts Receivable YE account [AR] and computes the percentage of these accounts to have first digits of "1" or "2", the AR DF profile expected according to the N-B expectation would be:

$$LOG_{10}(1 +1/1) + LOG_{10}(1 +1/2) = LOG_{10}(2) + LOG_{10}(1.5) = 30.1\% + 17.6\% = 47.7\%$$

If the auditor computes the ACTUAL number of AR YE balances that start with "1" or "2" as 44.2% a difference between actual and EXPECTED of −3.5% (44.2%–47.7%) would likely be accepted as in the usual random variation experiential range and so the auditor would not feel compelled to investigate possible causes in such a small DF deviation.

However, logically to justify the decision to not use extended procedures in the above AR example the auditor needs to have objective evidence in the best practices context. Usually then the auditor will select a statistical test to provide a test against the Null, which in the N-B profile case is that the DF profile does not follow the N-B profile. There are actually a number of such tests; we recommend the excellent article of Reddy and Sebastin (2012) who (i) detail distance measure testing, (ii) parametric z-tests, (iii) Entropic expectations as the basis of testing, and (iv) the χ^2 test. Our focus is on the χ^2.

1.3 Presentation Plan of the Paper

Following we: (1) treat in a pedagogic context the Expectation Selection and Sample Size issues for using the χ^2 test as a screen for the N-B frequencies, (2) given the sample size sensitivity of the χ^2, offer a re-sampling protocol that is developed using the Cho and Gaines (2007) In-Kind Committee to Committee dataset, (3) consider the discriminate validity of the proposed re-sampling protocol using the Reddy and Sebastin (2012) extended In-Kind Committee to Committee dataset and the dataset provided by Hill (1998); these datasets are offered as not conforming to the N-B profile. Additionally, we test the re-sampling protocol against seven other datasets that are argued as conforming to the N-B profile. All of these datasets are included in Appendix B, (4) address the question of the number of re-samples likely to be required in developing the audit evidence regarding the use of extended procedures, (5) summarize the various results as a χ^2 digital frequency testing protocol, and (6) finally, give the overview of a Decision Support System (DSS) for generating the χ^2 testing information. This DSS is used in our Audit and Assurance Services course; we use Beasley, Elder and Arens (2012) as the text. Also the DSS is part of the software-link of the Public Accounting firm to which we are the academic consultant. The DSS is programmed in *Excel™* using open access VBA modules. The DSS is offered as an email download from the corresponding author without cost or restriction to its use.

2. Pedagogic Context: The χ^2 Test: A Child of a Lesser Statistical God

We strongly prefer the χ^2 Test for the N-B DFT because the fundamental form of the χ^2 test is *Variance* from *Expectation*; we proffer, therefore, that the χ^2 test is the ideal audit GPS regarding the decision for using extended procedures. However, the χ^2 test has not achieved currency in the audit context. The reasons for this lack of χ^2 hegemony are (i) Sample Size and the (ii) Expectation Election issues.

Sample Size Anomaly Cho and Gaines (2007, p. 220), and just about all others, for example (Ley, 1996; Geyer & Williamson-Pepple, 2004) who have explored the χ^2 test as a differential signal note:

"*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.*"

The issue that Cho and Gaines identify leads to the standard advice offered in almost every statistics course where the χ^2 is discussed, to wit: *Large sample sizes "anomalously" are not desirable!* Actually, this advice does not only pertain to the χ^2 test it is ALWAYS true. Consider the simple two-tailed parametric one sample test against μ; as power increases with the root of the sample size, the related testing confidence interval converges "*in on*" the population expectation: μ and the rejection region of the Null of no difference becomes unrealistically large; in this case then rejecting the Null in favor of a statistically significant difference creates decision anomalies. So this sample size "critique", which is almost exclusively leveled at the χ^2 measure, is generally true in most all statistical testing. Therefore, the issue is not to reject the χ^2 screen in N-B Digital Frequency Testing but to calibrate the sample size so that this ideal test, the χ^2, can be used in the audit context. Consider now these two issues: The Selection of the Expectation and the Sample size.

2.1 The Expectation Selection Issue

Interestingly there are two ways that the expectation that is used in making the χ^2 computation may be rationalized. The inferential implications are not trivial as the inferential difference can change the action plan of the auditor. To elucidate these selection issues will require that we present the functional forms of the various computations that are used to generate the χ^2 test information.

As argued above, the χ^2 distribution is a natural screening measure for the audit context as it can be formed as the difference between an Observation and the related Expectation. The classic Pearson (1900) formula in the N-B context for the χ^2 is:

$$\chi^2 = \sum_1^9((O_{ij} - E_{ij})^2/E_{ij}),\ df = 8,\ over\ j{:}\ 1,2. \tag{2}$$

Where: O_{ij} represents the observed frequency/number for digit *i*, E_{ij} represents the expected frequency/number for digit *i*, *i*: 1, - - -, 9 and the jth index ranges over the number of sample/events for which frequencies are accrued as they relate to the marginal values. If there are two sample events or profiles there are 9 × 2 or 18 values in the summation for EQ2.

For the scripting of EQ2 for the N-B DFT context one may logically assume that the nine unique Expectations would be computed from the relative marginal values over the *i* index compared to the total number of observations. This means that one will conceptualize the values from EQ1 as a set of observations; in this case there are TWO observed frequency profiles; one from the counting made for an account under audit and one from EQ1. In this conceptualization, E_{ij} is simply computed as:

$$E_{ij} = (\sum_1^2 O_{ij})\ \times 50\%,\ over\ i{:}\ 1,2,\ ---,\ 9. \tag{3}$$

This is the usual or classic form for the χ^2 due to Pearson for the case where one is comparing realizations from two processes.

IF however, we elect to use as the *expectation* the following:

$$\mathrm{LOG}_{10}\,(1 + 1/d_i) \times \mathrm{N} \tag{4}$$

Where: N is the total number of realizations for the dataset under consideration—that is the Observed dataset—then, the jth index is no longer needed and EQ2 can be re-formed as:

$$\chi^2 = \sum_1^9((O_i - (LOG10\left(1 + \tfrac{1}{d_i}\right) \times N))^2/(LOG10\left(1 + \tfrac{1}{d_i}\right) \times N)),\ df = 8. \tag{5}$$

The implication for the magnitude of the χ^2 under the two different election assumptions is very important. It is always the case that the χ^2 computed using EQ5 is greater in magnitude than the χ^2 computed using EQ2. See

Note 1; the difference can be on the order of a 50% or so increase in the value of the χ^2. This difference may materially affect the p-value that one uses for inference. Consider now the second issue the Sample Size used in forming the χ^2.

2.2 The Sample Size Issue

To illustrate the important nature of these sample size and assumption effects consider the S&P data reported by Ley (1996) as presented in Appendix B. The sample size for the S&P dataset is 24,126. The digital presentation of this dataset and the Newcomb-Benford (N-B) Log_{10} benchmark is found in Table 1.

Table 1. The Ley S&P market return data and the N-B benchmark

First Digit Possibility	S&P Data Ley(1996)	(N-B) Log_{10} benchmark
1	29.2	30.1
2	17.0	17.6
3	13.4	12.5
4	9.9	9.7
5	7.8	7.9
6	7.1	6.7
7	5.6	5.8
8	5.4	5.1
9	4.7	4.6

Reality Check We have never shown this dataset to any of our audit colleagues or students over the years where someone remarked that there were deviations between the S&P and the N-B benchmark that suggested that extended procedure would be called for in the audit. We agree; there is a remarkable close fit for the S&P data relative to the N-B benchmark. Therefore this is a perfect demonstration dataset to examine the anomalous inferential effect due to the size of the sample. Regarding the *sample size effect*, the χ^2 values using EQ5 and EQ2 are respectively: 40.6 and 20.0, and the p-values are < 0.0001 and < 0.02 which are sufficiently low so as to suggest that there is a sufficient deviation from the N-B expectation to warrant using extended procedures! Clearly the auditor would be making a False Positive Error in investigating as the Null of no difference is very likely the true state of nature. Therefore, let us observe what happens if we reduce the sample size by 50%. If the sample size were to have been 12, 063 and we had the same frequency presentations—the χ^2 values would have been: $\chi^2 = 20.3$ for EQ5 with a p-value < 0.01 and $\chi^2 = 10.0$ for EQ 2 p = > 0.20. One can see the effect on the χ^2 that a reduction of sample size has, even given that the sample size is very large (See Note 2). Specifically, for a sample size of 12,063 using EQ5 the auditor likely will believe that extended procedures are warranted while for EQ2 it is unlikely that the audit-in-charge would believe that extended procedures are warranted.

2.3 Summary

The first pedagogical point is that the decision as to how one conceptualizes the *expectation* is critically important. If one is testing against a population point value in the digital frequency space, then EQ5 is the logical choice. If one assumes that both the realization vectors are realizations from different generating processes then E2 is the logical choice. We wish to note that the choice and rationalization of what is going to be used as the expectation is not trivial and can result in very different inferential action plans relative to deciding if extended procedures are warranted. Also as a point of information, most reports in the literature where the χ^2 is used as the detection measure for digital frequency differences use EQ5; although both forms can have an underlying rational as is also suggested by Ley (1996). The second pedagogical point is that in the audit context where the N-B screening is the DFT model, various accounts that may be selected to conduct the N-B DFT will have varying number of observations. When accounts such as *Cash Transactions* or *Inventory* are the logical choice for DFT by the auditor, the number of transactions for these accounts can be in the hundreds of thousands. In this case the auditor should be aware of the above mentioned anomalous sample size precision effect. We use this particular example in our audit and assurance course; the students seem to better understand the sample size and assumption issues after we have discussed this S&P case.

In our experience these two issues have created a certain jeopardy for the auditor to elect to use the χ^2 model as a DFT screen. Specifically, if the auditor uses EQ5 which is the formula usually presented by researchers in the DFT context and the account selected for DFT has a large number of observations then the bias created is that the χ^2 will likely signal that extended procedures will be needed independent of the reality of the DF profile difference of the account relation to the B-N profile; refer to Note 2. This invites the FPE of unwarranted use of extended procedures and so unnecessarily will increase the cost of the audit.

In the remaining sections of the paper, we will assume that EQ5 is the χ^2 measure for N-B digital screening as this is the standard assumption made by most of the researchers who have examined the χ^2 measure in the DFT context; we will now focus on the calibration of the sample size given this assumption. Consider now a suggested calibration of the sample size that seems consistent with rationalizing the χ^2 screen as a viable audit alert signaling model for using extended procedures in testing accounts in the audit context.

3 Calibration of the Sample Size for the χ^2 Screen

To facilitate the exposition of this section, we will need to make clear the context of our computational logistics:

1) The focus of this paper is on the calibration of the *sample size*; our goal is to suggest a sample size protocol that will enable the χ^2 to be used in the audit context without concern for the sample size jeopardy that we have documented above where very large sample sizes likely create excessive use of extended procedures testing and fails the False Positive Error "reasonability" check.

2) We are going to use EQ5 for the χ^2 calculation as it is the standard choice for researchers in the N-B DFT milieu.

3) We have set this at p-values for the calibration of the sample size in the working range between the value that Cho and Gaines use: a p-value of 0.001(0.1%) to a p-value of 0.20 (20%) for which the χ^2 values are: 26.13 and 11.03 respectively. We have selected 80% confidence as it is consistent with the Ramos Valuation Discovery sampling protocol where the lowest confidence offered in the confidence tables is 80%; also Ramos uses as the highest confidence of 99.5% which is effectively the same as the Cho and Gaines boundary. See Ramos (2008). These are the boundary conditions are for the *development phase* of the sample size calibration; later we will use as the cut-off for *using extended procedures*, the 95% confidence level. This means that if the χ^2tests to have a p-value less than 5%, then we will assume that this will likely signal the use of extended procedures. This 5% cut-off seems to us as consistent with the audit risk model of Beasley, Elder and Arens (2012) which is the text use in our Audit and Assurance course.

Given these three computational conditions we will:

1.) examine a set of data that has been evaluated and found to not conform to the Benford profile,

2.) using each of these datasets, we will start at full sample size as reported and then split the sample size in half at each iteration. We will then search the χ^2 as computed at each iteration for: The χ^2 that is just interior (i.e., lower) than the value of 26.13. Then we will move down the confidence scale towards the Ramos cutoff of 80% of 11.03 to where the χ^2 is just interior (i.e., higher) to the value of 11.03. This will give two sample sizes one for each of the bounds.

3.) Using this set of twelve boundary sample sizes, two for each of the six non-conforming datasets, we will compute the statistical profile of these values and select a performance cut-off as the χ^2 calabration. *This will be the* ***development phase****.*

4.) *Then in the* ***evaluation phase*** we will test the discrimination of this χ^2 calibration using a holdback sample of the non-conforming datasets and also of various other datasets that researchers have found to be conforming to the N-B profile.

5.) With this benchmarked information we will suggest a χ^2 sample size calibration protocol.

Consider now these various computational aspects of the developmental phase. All of the data that is to be used in the calibration χ^2 relative to the sample size is presented in Appendix B; for each dataset the reference work is cited and a brief description of the dataset is offered. This is, we suggest, an excellent pedagogically relevant set of data.

3.1 The Non-Conforming Dataset due to Cho and Gaines

There are six years reported of In-Kind Committee to Committee Contributions that were analyzed by Cho and Gaines (2007) (CG). As these were offered as non-conforming datasets, we will use this as our baseline dataset in the development phase. This dataset is also reported by Reddy and Sabastin (2012) (RS). Curiously, the two datasets presumably drawn from the same factual information reported by the Federal Election Commission (http://www.fec.gov/disclosure.shtml) are not exactly the same. This appears to be because RS used a larger sample than is reported by CG; for each of the six years that we will use in our baseline developmental calibration, the sample sizes reported by RS are larger compared to the CG reported sample sizes. The average sample size difference for (RS-CG) over the six years is: 455 units and using the usual parametric statistics has a 95% CI of (209 to 700). Also, of course, the digital frequencies are not exactly the same over all the 54 reported instances. The digital frequency differences (RS-CG) for the 54 observations produces the following Mean and 95% confidence interval: −0.004, (−0.052 to 0.043) respectively. We decided to use the CG initially reported data as the baseline for the digital frequencies and make random perturbations in them to better reflect variation between the two data sources.

3.2 Perturbations to the CG Non-Confirming Dataset

We felt that using the CG information as reported possibly would create a slight unwarranted increase in precision as the RS data for the same years differed in sample size and in reported digital frequency. Therefore, we randomly perturbed the reported CG information using the 95% CIs for the digital frequency differences as noted above. Here the idea was to introduce variation into the data to better reflect the fact that there were two different values reported for the same baseline data. Then we summed the randomly-perturbed-values and any excess over or under 100% was used to uniformly adjust the perturbations so that the sum of the digital frequency weighs would be 100%. These computations are detailed in Appendix A.

3.3 Arriving at Reasonable Sample Sizes

Having generated six digital frequencies by perturbing the CG Non-Conforming, we next took a random draw from the (RS-CG) sample size 95% CI as noted above. In this regard, we added this random draw to the CG reported sample size. With this dataset with the modified sample size we computed the χ^2 for the full augmented sample size and then reduced the sample size iteratively to locate for each of the six non-conforming datasets the sample sizes that were in an FPE interval from 20% to 0.1%. Consider following the specifics of this “sliding FPE scale”.

For example, consider the CG dataset for 1993/1994 as presented in Appendix B. We used the final perturbed frequencies as found in the Table: Column 4 in Appendix A for the developmental phase. To illustrate the exact computations we took a random draw from the sample size differential between CG and SB the 95% CI of which was: (209 to 700) and added it to the sample size reported CG which was 9,632. The random uniform draw in this case was 237 which was added to the sample size reported by CG resulting in a sample size of 9,869. Next we took this sample size and split it two at each iteration in the computation of the χ^2 for which we used EQ5. We then located the sample size that relates to the computed χ^2 that is interior (less than) to the CG upper limit of 26.13. This occurred in this case for a sample size of 705 where the χ^2 was 25.67. Then we examined the iteration output to locate the sample size just inside (more than) the 80% χ^2 value which is 11.03. This occurred for a sample size of 309 where the χ^2 value was 11.23. This procedure was used for all six of the CG datasets and so lead to 12 values. We then computed the 95% confidence interval of these 12 sample sizes with the following result:

Table 2. Benchmarking the SAMPLE SIZE using the Cho and Gaines non-conforming datasets

N = 12	Mean	Median	Range	LHS 95%CI	LHS 95%CI
Values*	315	249	(75 : 705)	190	440

*The vector of actual ordered values is: ((705, 579, 472, 464, 203, 175) : (309, 252, 246, 193, 106, 75))

We then tested these three sample size cut-points (190, 315, and 440) and also the full sample size as reported for the calibration of the χ^2 re: the sample size. In this regard we are interested in the performance of the calibration cut-points relative to the *False Positive* and the *False Negative* “Domains”; this is of course the usual discrimination test that addresses the sensitivity and the specificity of the calibration.

3.4 False Positive False Negative Errors Discrimination: The Calculations

We identified seven datasets that are expected to “Not Conform to the N-B frequency profile and seven datasets that are expected to “Conform to the N-B frequency profile. These datasets and the respective sample sizes are presented in the Appendix B. To test the dynamic changes in the χ^2 profile so as to locate a reasonable calibration for the sample size, we used four sample size calibrations: The full sample size for each of the datasets as reported followed by the cut-point boundaries of the 95% CI as reported in Table 2 of: LHS(190); Mean (315); RHS(440). Using EQ5 we calculated the χ^2 for each of the seven Non-Conforming Datasets and each of the seven Conforming datasets for each of these four sample size profiles. The purpose of this is to observe the dynamic change in the χ^2 profiles and with this information make a recommendation of the sample size that can be used in employing the χ^2 test as a screening protocol. This information is presented in Table 3 following.

Table 3. The sample size and χ^2 Profiles

	P-value Orientation					
Full Sample Size	Datasets Expected to be **Non-Conforming**			Datasets Expected to be **Conforming**		
χ^2FPE Rate	<.1%	(.1% to 20%)	>20%	<.1%	(.1% to 20%)	>20%
	7 Datasets	None	None	4 Datasets	None	3 Datasets
Large Sample Size	Datasets Expected to be Non-Conforming			Datasets Expected to be Conforming		
χ^2 FPE Rate	<.1%	<5%	>20%	<.1%	(.1% to 20%)	>20%
SS = 440	4 Datasets	3 Datasets	No Datasets	No Datasets	No Datasets	7 Datasets
Average Sample Size	Datasets Expected to be Non-Conforming			Datasets Expected to be Conforming		
χ^2 FPE Rate	<.1%	(1% to 20%)	>20%	<.1%	(.1% to 20%)	>20%
SS = 315	4 Datasets	3 Datasets	No Datasets	No Datasets	No Datasets	7 Datasets
Small Sample Size	Datasets Expected to be Non-Conforming			Datasets Expected to be Conforming		
χ^2 FPE Rate	<.1%	(.1% to 20%)	>20%	<.1%	(.1% to 20%)	>20%
SS =190	4 Datasets	No Datasets	3 Datasets	No Datasets	No Datasets	7 Datasets

Inferential Analysis Before we discuss the FPE and FNE profiles as represented in Table 3, we offer an illustrative example of the computations that lead to the data used to produce Table 2. Consider the Reddy and Sebastin In-Kind dataset for 1985/86. Their reported frequencies are presented in Table 4 following:

Table 4. Reddy-Sebastin In-Kind 1985/1986 computations for n = 315

Sample size n = 315	**RS 1985/1986**	**N-B Profile**	**Digital χ^2 value***
Digit 1	29.9	30.1	0.004 312
Digit 2	22.6	17.6	**4.455 803**
Digit 3	12.9	12.5	0.041 585
Digit 4	11.1	9.7	0.645 302
Digit 5	7.5	7.9	0.069 550
Digit 6	5.1	6.7	1.196 540
Digit 7	3.9	5.8	1.959 214
Digit 8	4.0	5.1	0.765 931
Digit 9	2.9	4.6	1.933 153
Total			**χ^2= 11.071 39**

*Using EQ5.

For example, for digit “2” as the first digit the computation using EQ 5 for a sample size of 315 is rounded to 3 decimals for the computational presentation:

$$\chi^2_{d=2} = ((O_2 - [Log_{10}(1 + 1/2) \times \mathbf{315}]))^2 / ([Log_{10}(1 + 1/2) \times \mathbf{315}])$$
$$\chi^2_{d=2} = ((71.190 - [0.176 \times \mathbf{315}]))^2 / [0.176 \times \mathbf{315}]$$
$$\chi^2_{d=2} = ((71.190 - [55.469]))^2 / [55.469]$$
$$\chi^2_{d=2} = 247.158/[55.469] \textit{ gives: } \boldsymbol{4.455\ 803}$$

Where: 0.226 × **315** = 71.190

The same set of computations was used then for all of the four sample sizes. Consider now the information provided in Table 3.

It is clear and certainly expected that the **Full Sample Sizes** for the Non-Conforming as well as for the Conforming dataset created, as Cho and Gaines suggest, too much precision and so using the full sample sizes invites the investigation-FPE—i.e., using extended procedures when it is likely not warranted. Specifically, for the Conforming data due to the very large sample sizes more than 50% (four of seven cases) of the time the χ^2 suggests an investigation as the p-value is < 0.001. At the other end of the sample size test spectrum where the **Sample Size is 190**, we find for the Non-Conforming dataset almost exact reversal where three times of seven we commit a FNE—i.e., failing to investigate when an extended procedures investigation is indeed warranted. The “Goldilocks” result—i.e., the result that is “just-right”—is for a **Sample Size of 440**. In this case, we likely make the correct 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 Table D where the χ^2 for the Non-Conforming data is 11.07 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 should have investigated.

Therefore our recommendation from the information presented in Table 3 for using the χ^2 in DFT as an audit screening is:

1) *Where the number of observations is large, more than 1000, the audit-in-charge would take a random sample of the number of observations for the Account under audit investigation.*

2) *The size random sample should be in the range: [315 to 440] observations.*

3) *As the AICPA, the SEC, nor the PCAOB has provided information on the risk to be used in setting the overall audit risk level we offer our interpretation, after years of experience in the audit context during which time we have discussed risk assessment with our audit colleagues as well as having years of experience in the statistical testing. We offer that when the FPE in the χ^2 screening context is less than 5% that most auditors would seek to cover this risk by using extended procedures for the particular account under examination. Using this as the cut-point for the Benford testing, the Audit Extended procedures investigation of an Account would be indicated if the χ^2 were to be greater than 15.507 which is the 95% Confidence or 5% FPE cut-off.*

4. Summary, Re-Sampling Consideration, and Conclusion

4.1 Summary

Cho and Gaines as have many others essentially lamented the fact that the χ^2 distibution is sample size sensitive. We agree however, one can VERY simply deal with this inherent sample size sensitively; when the account population is large, and many are in particular for PCAOB audits, taking a random sample of reasonable size for the account under audit examination is certainly a standard and usual protocol. As we mention above the excessively large sample and the resulting very narrow fail-to-reject-region is always a consideration not just in the case for inferential protocols using the χ^2 distribution as a DFT screening tool. As we find the χ^2 an ideal theoretic inferential model for the audit context as it is fundamentally based upon **Variance** from **Expectation**, we have suggested a sample size context for employing the χ^2screen in the execution of the audit. The random sample in the range of 315 to 440 seems to create the desired screening when we consider the joint concerns of inviting the FPE and/or the FNE. Also The sample sizes for re-sampling , that is taking a random sample of an account under audit scrutiny, are certainly in the usual possible ranges for PCAOB audits; incidentally, the excellent article by Geyer & Williamson-Pepple (2004) use in their simulation model four sample sizes: (200, 400, 600 and 800) which certainly fits well with our calibration results.

4.2 The Number of Re-Samples Needed

Another issue that we need to discuss is: *How many re-samples need to be taken?* The specific question is: Should the auditor take one re-sample from the account under audit investigation or as in the bootstrapping or jackknife modeling frameworks should multiple re-samples be taken and then blended to form one final sample; if the latter is the case then: How many re-samples are consistent with the suggested χ^2 protocol? Consider this final aspect next.

As for the random sample, the issue is simply: *What is a reasonable expectation for re-sampling?* The *a-priori* expectation is that as the minimum sample size recommended is greater than 300 that the re-sample should provide a reasonable reflection of the account population. This is another way of saying that 300 observations taken with replacement is in most cases an adequate sample; consider as context that the Gallop samples are on the order of 1000 for polling the population of the USA (See www.gallup.com/poll/101872/how-does-gallup-polling-work.aspx). If this were to be the case then one re-sample will be sufficient for using the χ^2 screen for the DFT.

To provide information on this question, we took ten re-samples from the Accounts Payable account for which we have all of the actual account data. We then created three blendings of the ten digital frequencies: Mean, Median and Mid-Point for each of the nine digital frequencies and finally an overall un-weighted blend of these three sub-blends; this produced 14 test first digit datasets. We then computed the χ^2 for each. We will then use the 95% test χ^2 value of 15.507 as the decision parameter regarding the use of extended procedures. Recall the Accounts Payable is a Conforming-Dataset and therefore the focus is on the FPE domain. The results of this testing are presented in Table 5.

Table 5. Re-sampling χ^2Profile for the Accounts Payable: Appendix B

	Ten Re-samples Mean (StDev)	**Mean Blend**	**Median Blend**	**Mid-Point Blend**	**Aggregate Blend**
χ^2(N =315)	8.3 (2.9)	4.7	4.2	4.9	4.4
χ^2(N =440)	11.5 (4.3)	5.8	6.9	6.6	6.2

The summary for the information in Table 5 is: for all 14 exposures for the sample size of 315 the decision consistent with the χ^2 values is the correct decision that extended procedures are not suggested. For the sample size of 440, there are two values among the 10 re-samples where the χ^2 is greater than 15.507 and so an investigation is incorrectly suggested. Conservatively if we collect all 28 χ^2 decisions and given that the decision to not investigate is correct, this FPE error rate is: 7.1% which is certainly in the acceptable range for the cut-off value of 5% for the χ^2 test. Also a performance validation of this inference is that the relationships between the two sample sizes and the blends are in the expected direction: The χ^2 for the individual re-samples are uniformly greater than for the Blends and the χ^2 for the larger sample sizes are uniformly greater than for the smaller sample sizes.

Another test that we conducted regarding the number of re-samples was to compute the precision around the actual digital frequency count; recall that we have the exact values of the digital frequency profile for the Accounts Payable dataset. For the 10 re-samples, we find that the average precision is 1.022 76%. We create a Worst Case Scenario for the Accounts Payable assuming that the precision differential is added to the first four actual digital frequencies and then subtracted from the last four actual digital frequencies (Modification A) and then reversed that modification (Modification B). Using these modifications we created the following datasets as presented in Table 6.

Table 6. Precision modifications for n = 440

Digits for Re-sample N = 440	**Modification A**	**Actual Digital Frequencies**	**Modification B**
1	29.8008	30.823 53	31.846 29
2	17.3302	18.352 94	19.375 70
3	10.5067	11.529 41	12.552 17

4	9.6831	10.705 88	11.728 64
5	7.8824	7.882 35	7.882 35
6	6.7875	5.764 71	4.741 94
7	6.1992	5.176 47	4.153 71
8	6.1992	5.176 47	4.153 71
9	5.6110	4.588 24	3.565 47
χ^2 Computed	**3.58**	**1.88**	**9.47**

Table 6 thus provides strong evidence that the variation in the re-samples is not likely to create detection anomalies as the Worst Case modifications do not provoke unwarranted investigations for the largest sample size recommended. And finally, in creating the precision information we also recorded the percentage of time that the digital frequency of the re-sample was greater than the actual or benchmark. This percentage was 51.1% for the 90 realizations. The 95% CI for this result for the Null of chance contains 50% suggesting strongly that the variation was symmetric around the benchmark.

4.3 Conclusion

All of this testing information suggests that:

One particular re-sample is all that will be needed to use the χ^2screening protocol suggested above. To be clear, it is not suggested from the above results that the auditor take a number of re-samples from an account and then blend the digital frequencies so as to use the χ^2test. One re-sample will do to use the χ^2screen for making the decision regarding using extended procedures.

This *one sample protocol* is consistent with the spirit of the sampling discussion of the PCAOB, AICPA, the text book that we use in our Audit and Assurance course: Beasley, Elder and Arens (2012), and also the Excel Template that is offered by Bloomberg™. Incidentally, the Bloomberg Benford template calculator asks for the sample size to be taken and so the results of our test fit perfectly with the GUI of their Benford calculator. See (http://www.bloomberg.com).

Finally, we wish to note that we have programmed in Excel: VBA™ a calculator/DSS that can be used to take the random re-sample, create the digital frequencies and compute the χ^2 for the particular re-sample. This DSS is called: *Chi2ReSample* is available from the corresponding author as a free download/email attachment without restriction on its use. We use the *Chi2ReSample* DSS in our under-graduate Auditing & Assurance course and have also created a shareware link for the DSS to the CPA LLP to which we are an academic consultant. An overview of the *Chi2ReSample* DSS is offered in Appendix C.

Acknowledgements

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

Beasley, M., Elder, R., & Arens, A. (2012). *Auditing and assurance services* (14th ed.). Pearson Publishing.

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

Geyer, C., & Williamson-Pepple, P. (2004). Detecting fraud in data sets using Benford's law. *Communications in Statistics: Simulation and Computation*, *33*, 229–246. http://dx.doi.org/10.1081/SAC-120028442

Hickman, M., & Rice, S. (2010). Digital analysis of crime statistics: Does crime conform to Benford's law? *Journal of Quantitative Criminology*, *26*, 333–349. http://dx.doi.org/10.1007/s10940-010-9094-6

Hill, T. P. (1998). The first digit phenomenon. *American Scientist, 86*, 358–363.

Lee, C-H., 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.

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

Newcomb S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, *4*, 30–40. 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. J. (1999). I've got your number. *Journal of Accountancy*, *187*, 79–83.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series, 5*(50), 157–175. http://dx.doi.org/10.1080/14786440009463897

Public Company Accounting Oversight Board (PCAOB) (2007). *Auditing Standard No. 5 (AS 5) Release No. 2007–005.* Retrieved 24 May, 2007, from http://pcaobus.org/Standards/Pages/default.aspx

Ramos, M. (2008). *Practitioners guide to GAAS: 2008* (7th ed.). J. Wiley and Sons.

Rauch, B., Göttsche, M., & Engle, 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.

Delivered to the Senate Committee on Governmental Affairs: Financial Oversight of Enron: The SEC and Private-Sector Watchdogs. (2002). Retrieved from http://www.gpo.gov/fdsys/pkg/CPRT-107SPRT82147/pdf/CPRT-107SPRT82147.pdf

Notes

Note 1. To show that the using EQ5 is greater than the using EQ2, for *a* as the observed and *b* as the Log10 value one can recast these computations for a general first digit *i* as:

EQ2'

$$\chi_{2'} = 2\times\left(\frac{\left(a-\left(\frac{a+b}{2}\right)\right)^{2}}{\frac{a+b}{2}}\right)_{i}$$

and re-form EQ5 as

EQ5'

$$\chi_{5'} = \left(\frac{(a-b)^{2}}{b}\right)_{i}$$

In simplified form these are:

$$\chi_{2'} = [(a^2+b^2-2ab)/(a+b))_i$$

and

$$\chi_{5'} = [(a^2+b^2-2ab)/(b))_i$$

As a condition for the Benford analysis a and b must both be > 0, then it is immeadaite that:

R1 $$\chi_{5'} > \chi_{2'}.$$

Further, one may show that the ratio $\chi_{2'}$to $\chi_{5'}$ is in the open interval (.5 to 1.0) for any fixed b, b ≠ a where a→ moves to b; as a→away from b the above ratio is in the open interval (.5 to 0). For any fixed a, a ≠ b where b→ moves to a; the above ratio is in the open interval (.0 to .5); as b→away from a the above ratio is in the open interval (.5 to 1.0).

Note 2. Another interesting aspect of the sample size issue is found in the work of Nigrini (1996). For the 1988 tax dataset, See Appendix B, there is a z-test signal that the frequency for "1" which is 30.59% compared to the theoretical expectation of 30.10% or a difference of 0.49% is inferentially important. This z-test "difference" is really due to the large sample size of 78,640. In a subsequent publication, Nigrini (1999, p. 81) offers that: "*The mean absolute deviation of the first digits of the census data is 0.7%, which means that on average, the actual proportion differed from the expected proportion by seven tenths of one percent. Auditors usually consider a difference of this magnitude to be immaterial.*

Appendix A. An illustration of the perturbation of the CG datasets

For the 1993/94 reporting year the following digital frequencies were reported by CG in Col 2:

First Digits	CG 93/94 Reported Frequencies	After Random Perturbation	Final Adjusted values
1	32.9	32.865 34	32.861 16
2	18.7	18.72 340	18.719 22
3	13.6	13.609 97	13.605 79
4	7.9	7.8781 49	7.873 97
5	8.9	8.9324 21	8.928 24
6	8.3	8.3269 09	8.322 73
7	4.1	4.0892 29	4.085 05
8	2.4	2.4289 29	2.424 75
9	3.2	3.1832 65	3.179 09
Totals	100.0	100.037 62	100.000 00

The values in Col 3 were derived by taking random draws in the interval (-0.052 to 0.043) and adding them to the value in Col 2. For example for digit "1" the random draw was: -0.034 66 which when added to 32.9 gives 32.865 34 which is the value in Col 3. The value in Col 4 was derived by taking the overage of Col 3 which is: 0.037 62 (100–100.037 62) and dividing it by 9 giving: 0.004 18. Finally, we subtracted 0.001 48 uniformly from all of the entries in Col 3 giving the final vector in Col 4 which then sums to 100. Specifically, for the first digit, 32.865 34 was reduced by 0.004 18 giving 32.861 16 which is the final value in Col 4 after the perturbation and then the re-scaling. We stopped the calibration at five decimal places.

Appendix B. The datasets used in the χ^2 calibration*

Indices	Digit 1	Digit 2	Digit 3	Digit 4	Digit 5	Digit 6	Digit 7	Digit 8	Digit 9
N-BLog$_{10}$	***0.301***	***0.176***	***0.125***	***0.097***	***0.079***	***0.067***	***0.058***	***0.051***	***0.046***
C&G93/94	0.329	0.187	0.136	0.079	0.089	0.083	0.041	0.024	0.032
C&G95/96	0.244	0.217	0.158	0.096	0.102	0.063	0.048	0.032	0.040
C&G97/98	0.274	0.185	0.153	0.103	0.118	0.059	0.037	0.039	0.033
C&G99/00	0.264	0.211	0.111	0.107	0.101	0.043	0.064	0.024	0.075
C&G01/02	0.249	0.226	0.107	0.116	0.105	0.043	0.034	0.030	0.090
C&G03/04	0.233	0.211	0.085	0.117	0.095	0.042	0.037	0.040	0.141
R&S85/86	0.299	0.226	0.129	0.111	0.075	0.051	0.039	0.040	0.029
R&S87/88	0.256	0.301	0.109	0.106	0.079	0.048	0.037	0.029	0.035
R&S89/90	0.283	0.220	0.162	0.107	0.080	0.054	0.034	0.040	0.020
R&S91/92	0.234	0.180	0.152	0.107	0.102	0.097	0.072	0.034	0.024
R&S05/06	0.264	0.215	0.101	0.076	0.095	0.043	0.038	0.041	0.127
R&S07/08	0.300	0.202	0.101	0.078	0.078	0.048	0.044	0.037	0.112
Hill	0.147	0.100	0.104	0.133	0.097	0.157	0.120	0.084	0.058
Benford	0.306	0.185	0.124	0.094	0.080	0.064	0.051	0.049	0.047
GIC:A/Pay	0.308	0.184	0.115	0.107	0.079	0.058	0.052	0.052	0.046
GIC:Cash	0.302	0.181	0.116	0.085	0.085	0.062	0.060	0.052	0.057
LeyS&P	0.292	0.170	0.134	0.099	0.078	0.071	0.056	0.054	0.047
Nigrini88Tax	0.306	0.178	0.127	0.095	0.078	0.065	0.056	0.050	0.045

Hickman/Rice	0.304	0.176	0.128	0.097	0.080	0.064	0.054	0.051	0.047
EU Data	0.299	0.181	0.132	0.101	0.077	0.066	0.054	0.047	0.042

*Sample sizes in respective order: N/A; 9632; 11 108; 9694; 10 771; 10 348; 8 396; 11 890; 10 220; 8 696; 11 661; 9005; 8723; 743; 20 229, 850; 846; 24 126; 78 640; 13 575; 39 691.

The first row, in ***Bold-Italics***, are the N-B theoretical digital frequencies which are the benchmark frequencies. The next six entries are the Federal Election Commission (FEC) reported Committee to Committee In-Kind contributions provided by Cho and Gaines (2007, Table 1, p. 218). These six frequency datasets were used as our *Development Sample*. The next Seven are the Non-Conforming dataset: The Six added by Reddy and Sebastin (2012, Appendix 4, p. 56/7) and the data published by Hill (1998, Figure 5, p. 363) where he asked 743 first year undergraduate students to write down a six-digit number. Cho and Gaines (2007) and Reddy and Sebastin (2012), argue convincingly why this FEC dataset is ideal for Digital Frequency Profiling (DFP) as Non-Conforming to the N-B theoretical first digit frequencies. Reddy and Sebastin show for the entire dataset of 12 years that even for rather tight neighbor windowing for the entire dataset, that there is relative discordance of the In-Kind dataset in that most of the SMI measures are under what many use as a cut-off of 75% concordance. The mean, median and range: for these SMI, un-blocked, are: 35.2%; 34.5% and (0.11 to 0.53). As a validity check on the In-Kind data and the profiling of it offered by Cho and Gaines, and Reddy and Sebastin, we examined the SMIs for the two windows reported by Reddy and Sebastin, k = 3 and 4. We do see that the $SMI_{k=3} > SMI_{k=4}$ which "must" be the case where there are relative dynamic conformity changes longitudinally—i.e., $SMI_{k=4}$ is essential smoothed relative to $SMI_{k=3}$ and therefore shows less neighbor entropic association suggesting that this conformity, albeit low, seems to be changing rather abruptly over a short cycle during the accrual period. As further evidence of the dynamic change, we calculate that the SMIs blocked by the two windows have a Spearman ρ of 0.58, $p < 0.05$ suggesting association consistent with dynamic change from window to window. All of this argues convincingly that indeed there is a lack of a strong or isomorphic conformity over the accrual period as viewed through the SMI lens. Therefore, this is an ideal "Non-Conforming" but not random dataset to provide a benchmark for Digital Frequency Profiling.

The last seven are datasets expected to conform to the N-B Profile which is the first row of the Table. These are: The Dataset collected by Benford (1938, Table 1, p. 553) to provide a reality check on his test of the N-B profile, The two datasets that we downloaded from the GIC Sector 25 using the WRDS™ database of the Wharton School of firms listed on market trading exchanges and so required to have a certified audit. The first of these is reported *Accounts Payable* and the second is *Cash* Balances both at Year-end. We selected *Accounts Payable* as Nigrini (1999) also selected this type of account for analysis. The dataset published by Ley (1996, Table 1, p. 312) of the Posterior Returns of the Standard & Poors Index (S&P). The Nigrini (1996, Table 2, p. 80) are the first digits of interest received as reported on submitted individual US Federal Income Tax returns for the years 1985 and 1988. Interestingly, other researchers, for example Hill (1998, Figure 5, p. 361) who report on the Nigrini tax data report the sample size of 169,662. This is not however correct. Nigrini had two samples: 1985 and 1988 with sample sizes of 91,022 and 78,640 respectively. Also, Nigrini's frequencies are reported in other research papers such as Hill (1998) as averaged over the two years. As Nigrini finds various z-test differences between the two datasets this practice of aggregation is questionable and should be avoided. As a point of caution, the simple average and the weighted averaged are not exactly the same. Finally, we selected the Nigrini 1988 data as: 1.) the sample size is smaller and so relatively better fits the χ^2 screen, 2.) Hill (1998) in Table 5 implies that the Nigrini aggregation fits the Benford profile, and 3.) the 1988-data had more z-test FPEs less than 5%. We can use this last criterion to demonstrate the large sample issue. For the sample size accrued by Nigrini, the χ^2 test value is greater than 26.13; however, at half the sample size, still arguably large, the χ^2 test value is less than 15.507 demonstrating the reasonability of Hill's implication that the Nigrini tax data is a conforming dataset. The Hickman and Rice (2010, Table 2, p. 338) dataset for 2006 of reported crime statistics using the Uniform Crime Reporting protocol , and finally the Rauch, Göttsche & Engle (2011, Table 2, p. 248) data on the EU 27 reporting countries of production and debt from 1999 to 2009 using the EUROSTAT Database.

Appendix C. The *Chi2ReSample* DSS overview of the functionality

To aid the auditor we have created an Excel™-VBA open access Decision Support System (DSS) that facilitates the creation of the re-sampling information. The DSS has the following worksheets the overview of which is presented following. The Tab/Worksheets are:

Worksheet 1: Intro Frequency Testing. *Introduction to Frequency Testing: Some Essentials Pertaining to this*

Decision Support System (DSS). In this section we discuss the Newcomb-Benford Digital model. Also, we discuss the rational of the χ^2 distribution which is Variance from Expectation. Also we give an example of the importance of the variance from expectation. Finally, we discuss the functionality of all the VBA-Launch Buttons (LB) that create the re-sampling information.

Worksheet 2: Random Sample. Here is where the re-sample is created. First the auditor is requested to *Paste* the full dataset under audit investigation in ColA of the worksheet. Then there is a Launch Button to create the re-sample; the auditor may select the number of items to be included in the re-sample. Then the auditor is requested to *Save* this re-sample and to *Paste* it in the next worksheet.

Worksheet 3: Computations One Data Set Only. This worksheet takes the re-sample *Saved* from Worksheet 2 and requests that it be *Pasted* in Col A; then the LB computes the first digit frequencies and also the overall and the individual χ^2 values are calculated. There is a text box that encourages the auditor to save this result. Up-to 10 re-samples from ten different datasets may be saved. The saving is done with a LB.

We have enjoyed success with this DSS as an enhancement for our Auditing and Assurance course. We hope others offering such courses can use this DSS.

Copyrights