Spanish Journal of Statistics
Vol. **5** No. 1 2023, Pages 43–53
doi:https://doi.org/10.37830/SJS.2023.1.03

Spanish Journal of Statistics

Regular Article

Applying and Testing Benford's Law Are Not the Same

William M. Goodman

Faculty of Business and IT, Ontario Tech University, bill.goodman@uoit.ca
Received: October 10, 2023. Accepted: December 30, 2023.

Abstract: Most papers on Benford's Law primarily discuss either (1) the science and mathematics for explaining the law; or (2) how to apply the law, especially for detecting data manipulation and fraud; or (3) suggestions for statistical tests to determine if data conform to a Benford's distribution. Leonardo Campanelli's recent paper "Testing Benford's Law" strongly objects to a descriptive measure I discussed in my paper "The Promises and Pitfalls of Benford's Law"—as if that measure were intended for Benford's testing in the Category-3 sense relevant for Campanelli's paper (SJS, vol. 4, 2022). This reflects a conflation of meanings for "testing" that is common in the Benford's literature, where many Category-2 papers claim they are applying (directly) conventional or new hypothesis tests as tools to detect fraud. Yet, fraud detection is a forensic and context-sensitive process, for which there is no set formula. In this paper, I clarify the sampling plan I had used earlier to collect and analyze a quasi-random sample of datasets, based on published criteria in the literature, to paint a tentative picture of how far real data vary, and in what ways, from abstract BL expectations. Further, I discuss simulations I have conducted to replicate and expand on my previous results.

Keywords: Benford's Law Applications, Benford's-Law Conformance Tests, Fraud detection, Benford's-Law Error Distributions

MSC: 62P99, 62F03, 62F35, 62G35

1 Introduction

As described in a recent paper by Leonardo Campanelli (2022), "Testing Benford's law: From Small to Very Large Data Sets", Benford's Law (BL) refers to an observed property of first-digit frequency distributions, for numbers contained in certain types of datasets. In such contexts, that distribution is skewed, such that, in the most clear-cut cases, about 30% of the numbers begin with the digit "1"; about 18% begin with digit "2"; and so on, down to less than 5% starting with "9". Referencing an early set of cases collected by Benford (1938), Campanelli lists a few contexts where the Law might tend to apply, such as the areas of rivers, and drainage statistics.

In the growing literature for Benford's Law —about 10 new entries a month, on average, in Benford Online Bibliography (Berger et al., 2023)—most papers fall into three main categories. They primarily discuss:

- The science and mathematics for explaining the law—for example, in Hill (1998) and (Berger and Hill, 2011); or
- How to apply the law, especially for detecting data manipulation and fraud, with putative examples—for instance, in Nigrini (1999); or
- Suggestions for statistical tests, and test statistics, to determine if data conform to a Benford's distribution. Campanelli's paper (2022) fits in the third category, and he there proposes an improved BL-conformance measure.

In much of the literature for Categories 2 and 3, however, the terminology of "testing for" Benford's law is applied ambiguously. In his paper, for instance, Campanelli (2022) strongly objects to a descriptive measure that I presented in my article "The Promises and Pitfalls of Benford's Law"—as if that measure were intended for hypothesis testing in the Category-3 sense relevant for Campanelli's paper. That confusion is not usual, because applied-focused papers on Benford's (Category 2) often do look to tests, such as Campanelli proposes, to be tools they can apply directly for detecting fraud. But that is to conflate "statistical testing" with the forensic, context-sensitive process of fraud "detecting."

To clarify this important distinction, consider this analogy where the pattern of interest is normal distributions.

Case A: Author A is told a sample was drawn randomly from a purely normally-distributed population. So, A tests whether the sample is consistent with that premiss. For example, A tests for goodness of fit with the normal distribution, using a normality test in Minitab (2023), and obtains a p-value. (Putting aside issues around p-values.) If the test rejects goodness of fit, it would not be unreasonable for A to question whether the initial premiss of normality was correct (while being aware the result could also be a false positive).

Benford's Law papers focused on testing (Category 3) are analogous to papers that explore improved ways of testing assumptions about a population's normality.

Case B: Author B has read, somewhere, that samples of sports-teams'-salary data should, in general, be "roughly normally distributed"—though the literature is vague on how close to normal that should entail. (Would data's being roughly symmetrical and unimodal be sufficient?) B attempts to detect, based solely on data distribution, whether the data published on NFL team salaries for 2022 on Statistica (2023) seem suspiciously unusual.

If B applies a normality test on the Statistica data, such as A used for Case A, he or she would find the test "rejects" the assumption of a pure-normal-distribution. But what follows? The question, here, is not whether the data are drawn from an abstractly-perfect normal distribution, but whether the data are distributed normally-enough to be faultless, in the absence of any other evidence for wrongdoing. If the only evidence presented is from the data's own distribution, B has failed to model what roughly-normal, non-fraudulent distributions, in Statistica's data context, should look

like, to make a fair comparison.

That Case B analogy applies to many papers for applying Benford's. They test datasets' conformance with the pure Benford's model. But to flag cases as unusual in context, so warranting suspicion, a fuller picture is needed of what real, Benford-like but not pure-Benford's distributions look like, in practice.

That was a key goal for my paper (2016). I collected and analyzed a quasi-random sample of datasets, based on criteria and suggestions in the literature, itself, for what are likely to be "Benford-suitable" types of data. The data's general topics were like: energy consumption; farm cash receipts; racehorse prices; election vote counts; and so on. The aim was to paint a tentative picture of how far real data varied, and in what ways, from the abstract BL expectations.

My preliminary findings were intended as general, and descriptive. I wrote, in effect, "real cases are seen to vary considerably from a pure Benford's distribution, and exhibit variances from it as large as—suppose I called it—$d*$". I did not intend, by that observation, to claim, as Campanelli interprets it, that "if a dataset varies by less than $d*$ it must, thereby, conform with Benford's". That wording would imply testing; but no test was performed. Just as, if I said sports-data are often somewhat-normal but have lots of variation, so Statistica's data are not unusual—that would not be the same as concluding, "Statistica's data are normally distributed."

To support the findings in my paper when it was originally published in Significance, I made available an Excel file that fully documents my sampling procedures to obtain the "40 Cases"; and it also includes source data, key calculations, and descriptions of reasons and procedures for cleaning the data. Readers can currently download that file from the archive site ResearchGate (Goodman, 2019). My published article cited a different URL for the file, hosted by the publisher; but that link has not been sustained.

2 Simulations to replicate and clarify previous results

In preparing this paper, I have conducted a series of simulations with Excel, including macro-based randomization and resampling. A file showing the related formulas and data will be posted and made available on ResearchGate. The results clarify where, and how, chi-square and other such tests can be validly applied for Benford's testing; but, also, where and when they become unsuitable.

First, I simulated the drawing of random "first digits" from a purely-BL-distributed population, that is to say, from a probability distribution for the expected, relative proportions of first digits, as predicted by Benford's Law. Various-sized collections of these digits were treated as samples. The magnitudes of numbers which have those first digits, and the numbers' subsequent digits, were filled in later, if applicable. For each pass within a simulation run, 1000 such digits were drawn, and relevant measures and statistics were recorded; then the simulation run looped back for another pass.

In one simulation run, all 1000 digits produced in each pass were interpreted, collectively as one random sample. A thousand such loops produced a thousand samples. Figure 1 illustrates, with boxplots, the distribution of all the samples' first-digit proportions. Those proportions centered

closely, as expected, around the Benford-predicted values, such as about 0.30 for first-digit 1's, and so on. Any variances were relatively small.

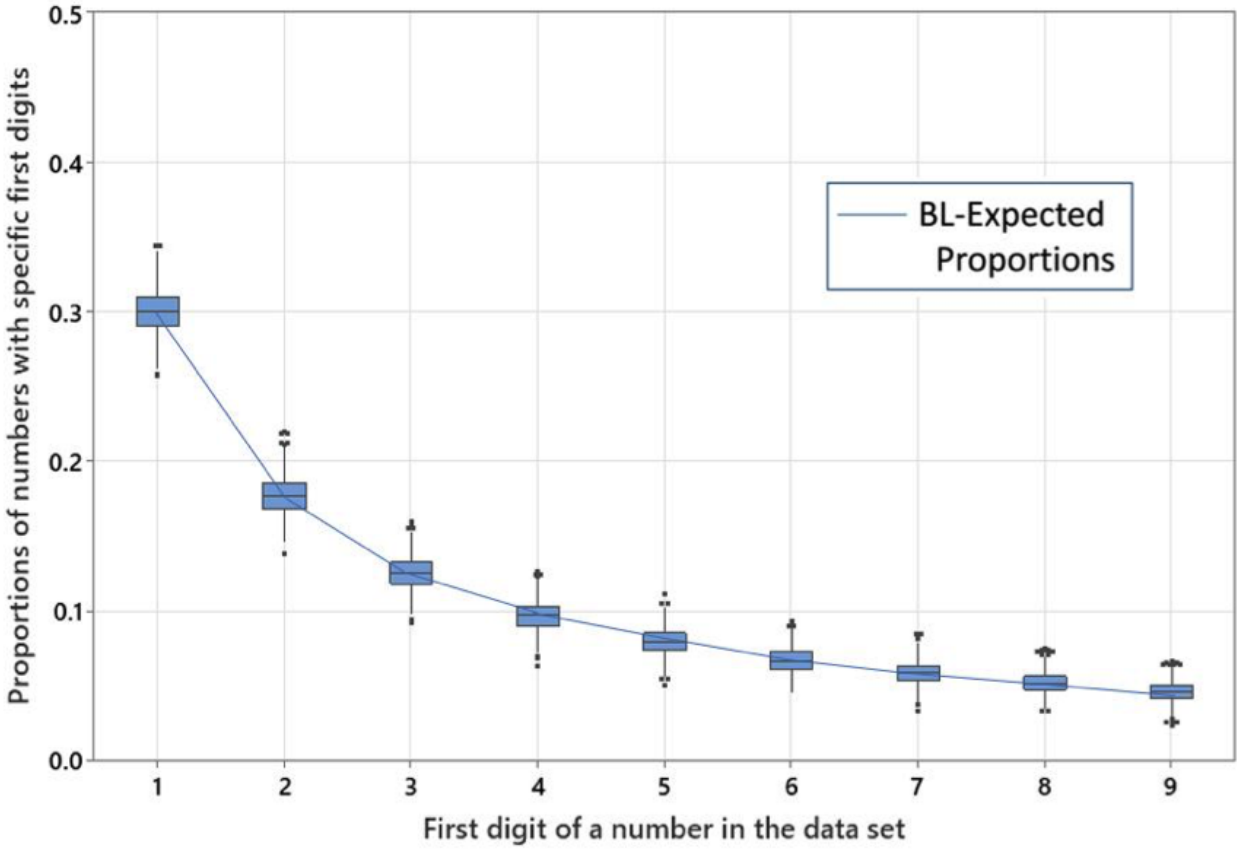

Figure 1: Distributions of first-digit proportions for 1000 random samples, drawn from a purely BL-distributed population

Other simulation runs modeled different sample sizes. Also, for each resample created during the runs, the worksheet calculated two measures for the sample's deviation from pure Benford's: (a) chi-square goodness-of-fit and (b) Cho and Gaines' "normalized Euclidean distance" measure, $d*$ (Cho and Gaines, 2007), which are both discussed in Campanelli (2022). $d*$ was the measure utilized in my 2016 paper. Also calculated was the p-value corresponding to the chi-square test.

Because all variance from Benford's expectations in the first simulation was due only to sampling error, the chi-square tests worked exactly as advertised —at least up to sample sizes $n = 1000$. Figure 2 illustrates that, for all modeled sample sizes, a chi-square test would not mistakenly reject goodness-of-fit to Benford's, except (as expected) for about 5% of cases, if 0.05 was the Type I error rate α for the test.

Note, the chi-square formula inherently adjusts for sample size, by dividing the squared-differences of counts for each digit (actual versus expected) by the expected counts. So, a

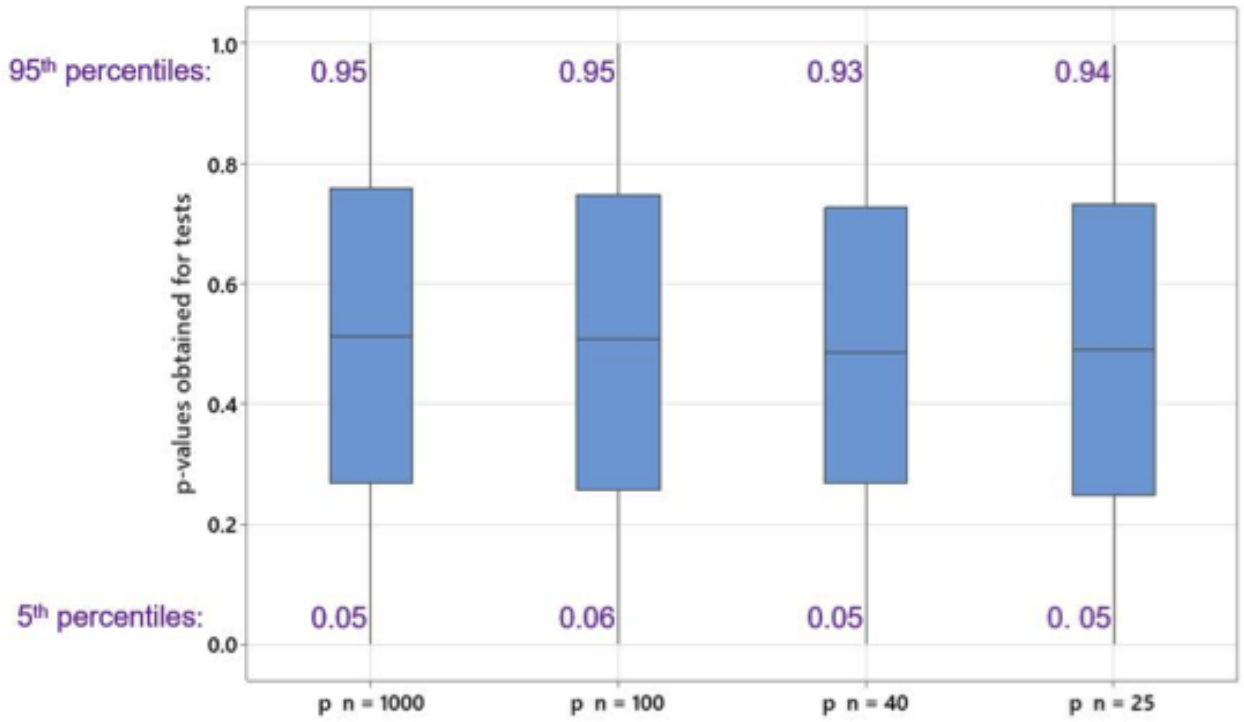

Figure 2: Distributions of chi-square p-values from repeated tests of samples, each drawn randomly from a purely-BL-distributed population

commonly-mentioned concern about chi-square's power with a large n is not a problem for the testing—if it can be assumed the sampling is (or should be, in the absence of alleged fraud) from a "pure Benford's" distribution. However, when that assumption does not apply, chi-square tests will quickly flag variances as significant, regardless of n. That is illustrated in Figure 3.

Figure 3 is based on results of additional simulation-runs, where—in ways I will describe below—samples were not produced from a solely-Benford's distribution. Since chi-square compared the observed digit counts with pure-Benford's expected counts, chi-square tests resulted in many more rejections of fit $(p < 0.05)$ than the nominal $\alpha = 0.05$. Sample size only slightly accentuated the problem. The rejections are "false positives" if, analogous to NFL salary-data in Statistica, the populations are actually legitimate, but don't happen to exemplify an abstractly pure distribution.

The $d*$ measure of distance from pure Benford's, however, is quite sensitive to sample size, as Campanelli (2022) rightly observed. As mentioned above, my earlier paper focused on how widely, in general, BL-suitable datasets can vary, using the $d*$ measure. But it may have been helpful to control the paper's findings for sample size. This has now been addressed in Figure 4, using simulations.

Figure 4 shows that BL-non-conformance, measured by $d*$, can vary markedly by n, even for samples drawn from simulated, purely-Benford's populations. But even if controlled for n, the $d*$ upper confidence bounds could not be used, simplistically, to test for a sample's goodness-of-fit. There are many ways and extents to which a population can be legitimately "mixed", to have a

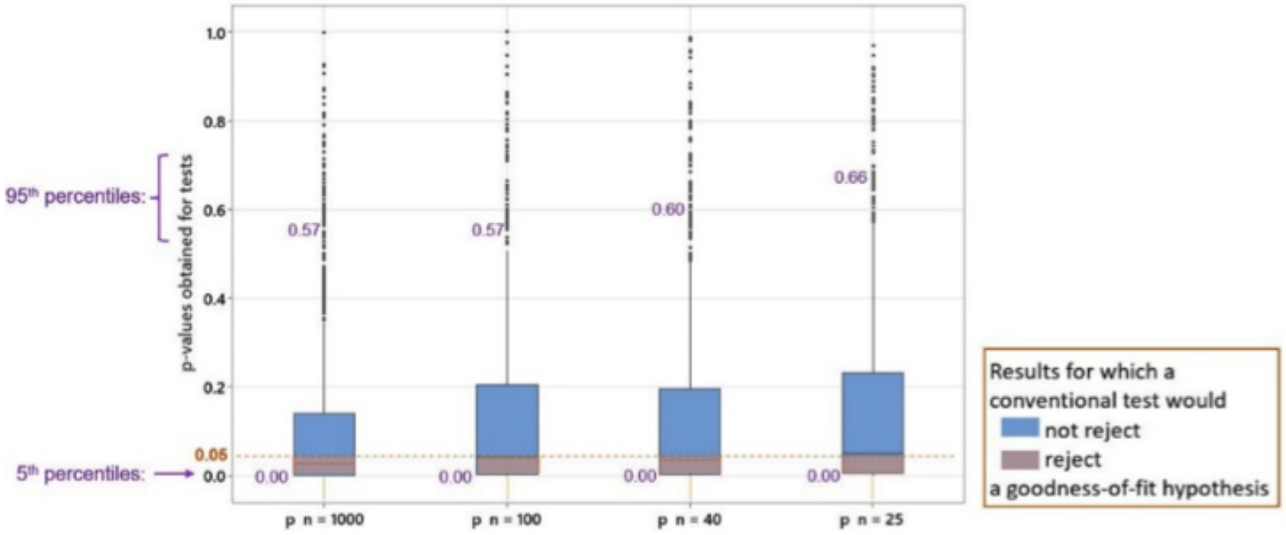

Figure 3: Distributions of chi-square p-values from repeated tests of samples, each drawn randomly from a somewhat BL-distributed population

Benfords'-like, but not pure BL distribution. To test based on the upper $d*$ confidence bound, one would have to know, in advance, which specific type of mixture was in play.

3 Mixed-Pattern Populations

Preparing for this paper, I examined a number of datasets, including among my 40-Cases, for factors that might increase samples' apparent distances from true BL-conformance—even where BL still applies to some extent. Two main issues were found to recur. The following observations are preliminary, and presented here to encourage further research.

3.1 Repeated Values

Large datasets of cities' populations would seem on the face to be Benford-suitable. Benford, himself, includes such an example in his case set (1938). But suppose "population of home city" is a column in a dataset for incidents involving people. This happens, for example, in data for accident reports or insurance claims. If a person is involved in multiple incidents over time, their data may appear in several, non-contiguous rows; so, their demographics, such as age and population of home city, would be repeated in all records involving the same individuals.

Someone searching for data manipulation might examine a population column in isolation, expecting it should closely follow Benford's law. But in the above context, some cities' numbers would be repeated. Long runs of repeated numbers might be obvious; but scattered twos and threes of matches throughout the data may escape notice.

The right half of Figure 4 was based on a simulation that generated a repeating effect, like the one just described. Random next-numbers for samples were tentatively generated as described in

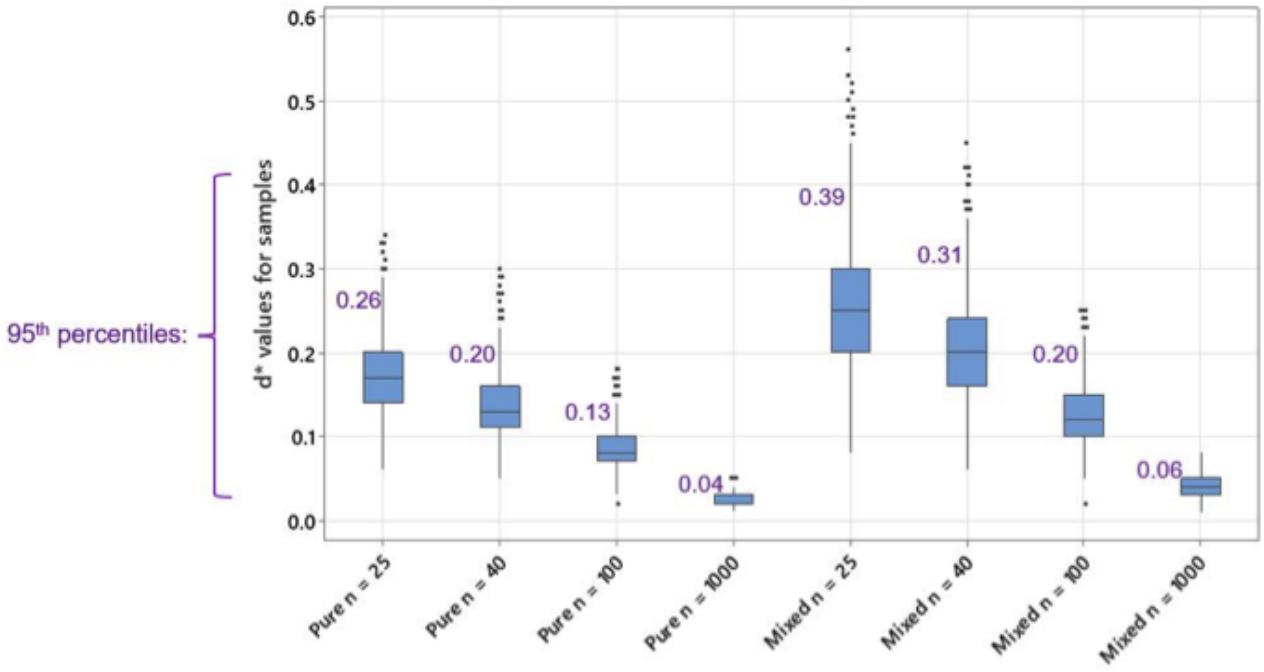

Figure 4: Distributions of $d*$ measures of non-conformance to pure BL, from repeated tests of samples drawn randomly from, respectively, a purely-BL-distributed population and a somewhat-BL-distributed population

the previous sections, but with an overriding 40% chance that the next number generated would be, instead, a repeat of a previously generated number. (A 40% repeat rate would not be unusual for some incident-based datasets.) The resulting spreads of $d*$ for those simulated samples were comparable to those for some of my 40 Cases.

Figure 4 is, obviously, not a confirmed model for how all repeats occur in BL-like, legitimate data. And to test with the model, it would need to be more specific: What is the repeat rate for a particular population? Are there other mixture mechanisms in play? Still, the model illustrates how a legitimate dataset can be Benford-suitable, by accepted criteria, yet differ notably from the abstract pattern.

3.2 Limited Orders of Magnitude

A commonly agreed requirement to expect a BL distribution is that the source numbers should span a sufficiently large range (Campanelli, 2022), though how large is not clear. Datasets mentioned in papers vary on this, and Benford, himself (1938), included atomic weights as an example. Atomic weights known today range only from about 1 to 294 (IUPAC Commission, 2021. In 1938, the heaviest known weight was 238 (Uranium).

For testing purposes, order-of-magnitude (OM) limitations would seem, like sample size, to be sampling concerns for BL tests, rather than inherent to the source distributions. If only there were more voters, or larger financial transactions, for instance, the data could have gone, in principle, to higher orders of magnitude.

If that's the assumption, it might not be as crucial how many OM's are sampled, as whether the orders of magnitude are complete. Are the first digits, 1-9, fully represented for all orders of magnitude in the sample? If not, that could bias the sampled digits' distributions. The following simulation confirms that point.

As in the simulations described previously, this new model likewise generates many single digits, each interpreted as the first digit of a full number, drawn randomly from a purely-Benford's distribution. This time, the implied "full numbers" get fleshed out, in this way: For each resampled digit: (a) randomly select an order of magnitude for the full number, from 10^0 (i.e., 1) up to 10^5 (100,000), and multiply that times the selected "first digit". Then (b) randomly generate remaining digits for a number in that magnitude range. (For instance, if the randomly selected digit was "2", it might be transformed to 20,000 (based on $2 * 10^4$); and then the 0's filled in (using an Excel function) by adding "rand() $*10^4$" which might return "8663". In that example, the resulting, full number would be 28663.) There was no attempt, for this purpose, at modeling Benford's other sets of distributions, expected for second or subsequent digits of numbers.

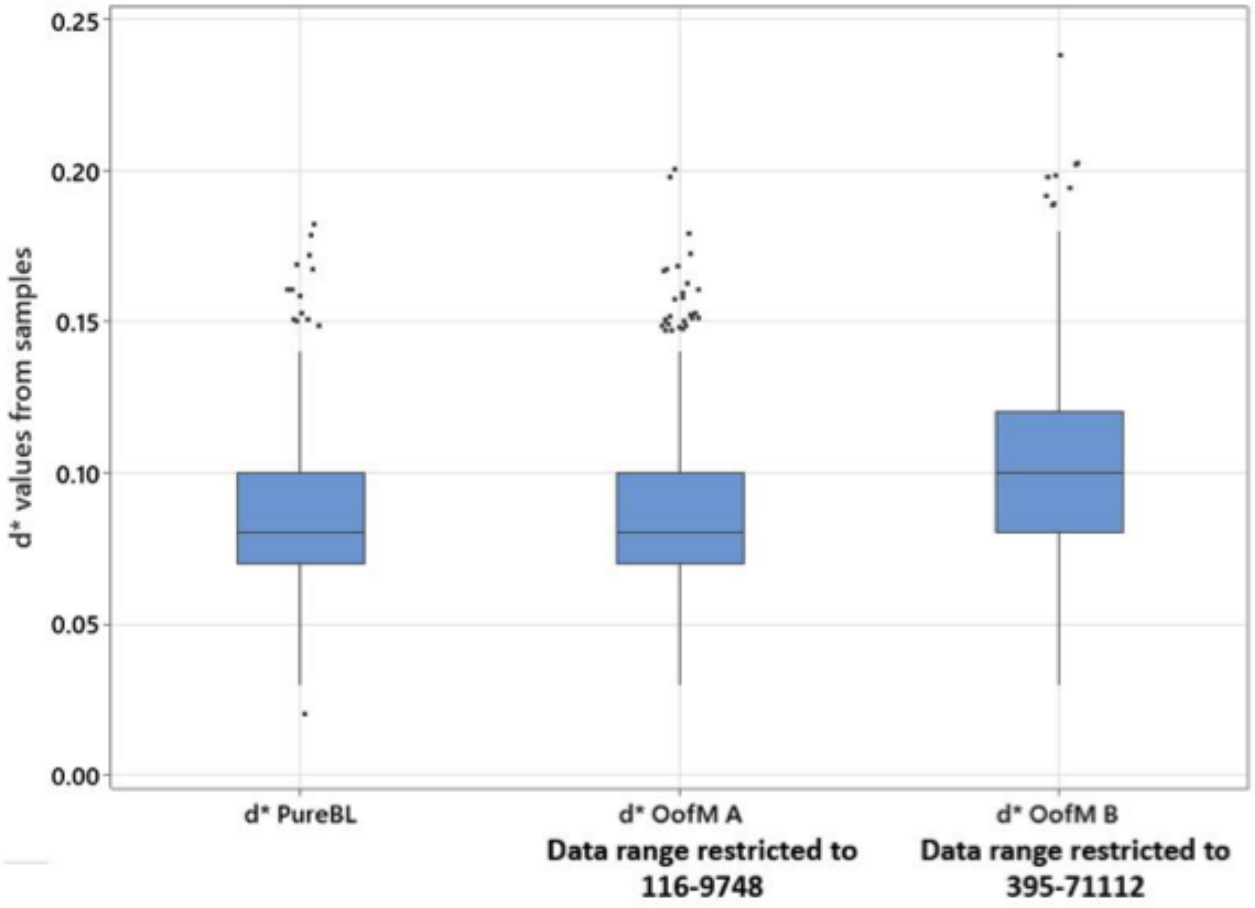

Figure 5: Distributions of $d*$ measures of non-conformance to pure BL, from repeated random re-sampling, $n = 100$, where the sampling frames were, respectively, not magnitude-range restricted, or range-restricted in two specified ways

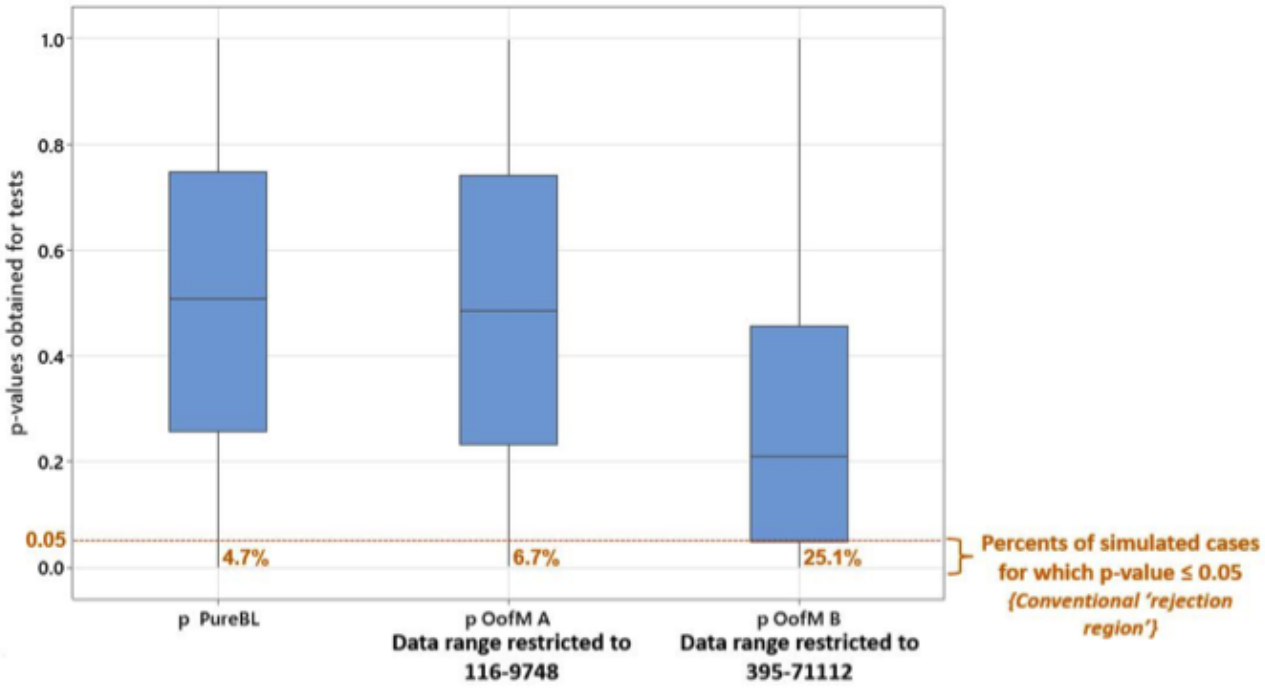

Figure 6: Distributions of chi-square p-values, from repeated tests of $n = 100$ random samples , where the sampling frames were, respectively, not magnitude-range restricted, or range-restricted in two specified ways

The left boxes in Figures 5 and 6 are repeated from two prior figures, for selecting repeated samples, $n = 100$, from purely-BL-distributed populations, with no range restrictions. The middle boxes (A) model cases when the underlying distribution is, still, pure-Benford's, but for some reason all collected data fall only in the range 116 to 9748; any out-of-range numbers, tentatively randomly generated, were excluded from the sample. That range is less than two full orders of magnitude; but all the digits 1-9 are represented in both the hundreds and thousands ranges. The boxes on the right (B) are models of when the magnitude range is 395 to 71112. This is a larger range than for box A, yet not all possible first digits are represented equally across the hundreds, thousands, and ten-thousands ranges.

Figure 5 shows that, as measured by $d*$, simply restricting orders of magnitude (box A) had no appreciable effect on the spread of variances of samples from true-Benford's distribution—so long as all first digits were fairly represented. But when all digits did not have balanced representation across all magnitudes in the sample (box B), this increased samples' variances from Benford's.

The specific $d*$ results were affected by the sample size and the OM ranges used in the simulation, and by assuming no other extraneous factors influenced the distributions. So, the bounds shown in Figure 5 could not be a basis for testing. Yet, it illustrates that if tests are conducted, details about the data ranges are potentially confounding issues.

Figure 6 is from the same simulations as above; but for each resample, a chi-square p-value for its variance from pure Benford's was calculated. Again, there was little impact for restricting orders

of magnitudes, provided all digits stay represented in the two ranges (box A). But if (box B), OM ranges are unbalanced for representing first digits, the proportions of conventional tests rejecting—mistakenly—the underlying population's true fit to BL would markedly increase.

4 Discussion and conclusions

It is tempting to think that if, based on simulations or mathematical models, figures like Figures 4 and 5, above, could be broken down for other possible confounders, then tests for detecting anomalous datasets might be possible, after all. Just as Campanelli interpreted my article's $d*$ upper-confidence bound as a "rule of thumb" for testing, Figure 4 might show something like: "For BL-like populations of a certain mixture type, if sample size is 100, and orders of magnitude are not restricted, then a $d*$ variance measure < 0.20 might signify 'conformance'." ... and so on.

But Figure 4 represents only one kind of mixture. And Figure 5 considers only two alternative data ranges. Could BL-like contexts be subdivided into the most important groups and combinations of possible mixtures and conditions, to calculate expected variances for every such group? Or, could additional rules be applied when determining if a dataset is suitable for Benford's-based testing—to screen out mixtures and contexts that confound pure testing? The first strategy would be highly impractical to implement; and, for the second strategy, cases that survive all the new exclusion criteria might be vanishingly few.

I believe a better approach is summarized in this extended quote from Deckert et al. (2011). Their specific quote relates to election fraud detecting; but its important point is generalizable to accounting records and other BL-testing contexts:

"(Any) inference that the analysis of official returns can begin and end with Benford's Law or that we can dispense with measuring other variables (correlated with) voting is unwarranted: Detecting and measuring fraud is much like any criminal investigation and requires a careful gathering of available data and evidence, in conjunction with a 'theory of the crime' that takes into account substantive knowledge of the election being considered. (...) Absent a clearly specified theory of how Benford Law apples to the (specific) data, (...) Benford's law gives an unacceptably high chance of committing both Type 1 and Type 2 errors."

Such concerns do not apply directly to research, such as Campanelli's, focused on improved testing for Benford's distributions, as a mathematical pattern. But, as Deckert correctly notes, when fraud is alleged, one needs to look at the whole complexity of the data, and the story behind it. Finding, or not, an abstract pattern (or a combination pattern, for second digits, etc.), provides one piece of information, perhaps, for an investigation, and could be called "testing" Benford's Law. But "applying" Benford's Law for what are, essentially, forensic applications is complex, and context-sensitive, and not adequately addressed, unfortunately, in many fraud-oriented papers.

This paper recommends further research on the ways Benford-like patterns can be mixed, during the data generation, and on ways sampling frames and decisions can impact the apparent results.

References

Benford, Frank (1938). The law of anomalous numbers. *Proceedings of the American Philosophical society 78*(4), 551–572.

Berger, Arno and Theodore P Hill (2011). A basic theory of Benford's Law. *Probability Surveys 8*, 1–126.

Berger, Arnold, Theodore P. Hill, and Edard Rogers (2023). Benford online bibliography. Accessed October 4, 2023.

Campanelli, Leonardo (2022). Testing Benford's Law: from small to very large data sets. *Spanish Journal of Statistics 4*, 41–54.

Cho, Wendy K. and Brian J. Gaines (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The american statistician 61*(3), 218–223.

Deckert, Joseph, Mikhail Myagkov, and Peter C Ordeshook (2011). Benford's Law and the detection of election fraud. *Political Analysis 19*(3), 245–268.

Goodman, William (2016). The promises and pitfalls of Benford's Law. *Significance 13*(3), 38–41.

Goodman, W.M. (2019). Benford data sets 40 cases. Excel file, posted for downloading. https://www.researchgate.net/publication/337682312_Benford_Datasets_40_Casesxlsx.

Hill, Theodore P (1998). The first digit phenomenon. *American Scientist 86*(4), 358–363.

IUPAC Commission on Isotopic Abundances and Atomic Weights (2021). Atomic weights of the elements 2021. https://iupac.qmul.ac.uk/AtWt/.

Minitab (2023). Minitab 21.4.1 (64-bit). 2023 minitab.

Nigrini, M.J (1999). I've got your number. *Journal of Accountancy 187*(5), 79–83.

Statistica (2023). Player payroll in the National Football League 2022/2023 season, by team. Technical report. https://www.statista.com/statistics/240074/player-salaries-of-national-football-league-teams/.