A New Benford Test for Clustered Data with Applications to American Elections

Katherine M. Anderson [1], **Kevin Dayaratna** [2], **Drew Gonshorowski** [3], and **Steven J. Miller** [4]

[1] Undergraduate Economics Major, Brigham Young University-Idaho; andersonkatie1998@gmail.com
[2] Chief Statistician, Data Scientist, and Senior Research Fellow at the Center for Data Analysis, The Heritage Foundation; kevin.dayaratna@heritage.org
[3] Research Fellow in Simulations, Center for Data Analysis, The Heritage Foundation; drew.gonshorowski@heritage.org
[4] Professor of Mathematics, Williams College; sjm1@williams.edu

Abstract: A frequent problem with classic first digit applications of Benford's law is the law's inapplicability to clustered data, which becomes especially problematic for analyzing election data. This study offers a novel adaptation of Benford's law by performing a first digit analysis after converting vote counts from election data to base 3 (referred to throughout the paper as 1-BL 3), spreading out the data and thus rendering the law significantly more useful. We test the efficacy of our approach on synthetic election data using discrete Weibull modeling, finding that, in many cases, election data often conforms to 1-BL 3. Lastly, we apply 1-BL 3 analysis to selected states from the 2004 US Presidential election to detect potential statistical anomalies.

1. Introduction

Although Benford's Law (see Section 2 for a review) has successfully detected fraud in many settings (flagging, among others, financial records, images, and scientific data that have been modified), the difficulty in determining when one should expect Benfordness severely limits its applications to detect fraud or other issues. One popular argument is that data sets that are spread out over several orders of magnitude are often close to Benford. Though this phenomenon is frequently true (see [1,2] for results on exponential random variables), it can fail; see for example Chapter 2 of [3] or [4] for issues with the 'spread' hypothesis.

Clearly, not all data sets follow Benford's law, especially in elections [5,6]. For example, if the data is not spread out sufficiently, the distribution of first digits is unlikely to conform to standard base 10 representations of Benford's law (hereafter referred to as 1-BL 10), and thus deviations from Benfordness do not suggest fraud or other concerns. This issue is the main motivation for this work; specifically, our analysis is intended to address discussions where claims of election fraud are made when observed candidate vote counts do not satisfy 1-BL 10 [7].

A classic example is precinct by precinct election data. Figure 1 provides such data from three counties from the 2004 US Presidential Election between incumbent President George W. Bush and Senator John F. Kerry. As is apparent in Figure 1, precinct level election data can be clustered thus rendering 1-BL10 useless. As a result, work by Walter Mebane and others [6,8–17] have looked at the second digit (hereafter referred to as 2-BL 10), or the last two digits (which one expects to be the same about 10% of the time base 10).

In this study, we propose an additional approach for digit-based analysis of concentrated data. Namely, we write the data in terms of base 3 instead of base 10; the advantage of doing so is that our numbers are now spread out over several more magnitudes with two choices for the first digit instead of nine. For example, 81 is a two digit number in base
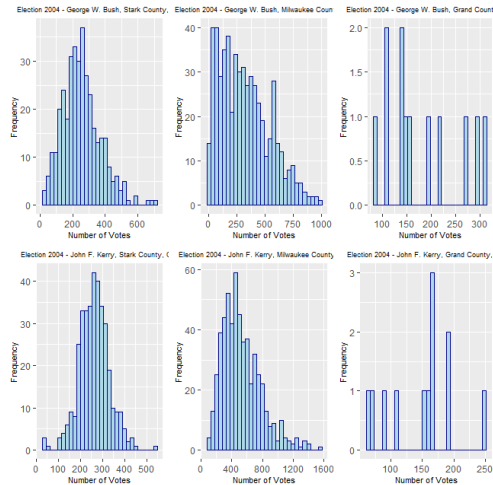
Figure 1. Vote Distributions Across Precincts in US 2004 Presidential Election for Selected Counties.

10, but in base 3 it is 10000_3, five digits representing four orders of magnitude, and this transformation thus spreads out clumped data. This idea of switching bases was originally proposed by the fourth named author over a decade ago to detect falsified data that passes a Benford test base 10 but might fail it in another base. Other base representations are of course possible; however, higher base representations would cluster the data more than base three, and base two representations would only have a single choice of a leading digit. Base 3 representations in election data were recently applied by the Kossovsky and Miller (2020) in [18].

This paper thus contributes to the literature on Benford's law by performing a first digit analysis of precinct level election data based on base 3 representations (hereafter referred to as 1-BL 3). Specifically, our work is organized as follows. First, we discuss existing forms of Benford's law used for analyzing data and various generalizations. Subsequently, we fit precinct-level election data from the 2004 US Presidential Election to discrete Weibull distributions and assess the goodness of fit of these distributions in modeling election data. Subsequently, we draw upon this analysis by using a variety of choices of parameters based on this estimation to generate synthetic election data to compare the performance of 1-BL 10 and 1-BL 3. We find that, in many cases, election data generally conforms to 1-BL 3, but does not to 1-BL 10. Lastly, we apply 1-BL 3 analysis to selected states from the 2004 US Presidential election to detect potential statistical anomalies. We emphasize that our work is generalizable to other elections and thus should be a useful tool for researchers to use in analyzing the integrity of elections.

2. Benford's Law and Generalizations

In the 1880s Simon Newcomb [19] observed that the pages of tables of logarithms were not equally worn; there was more wear and tear on the earlier pages (those corresponding to the logarithms of numbers with a small first significant digit, such as 1, 2 or 3) than the later pages (those from numbers with a large first digit (such as 7, 8 or 9). Fifty years later Frank Benford [20] saw similar biases in many different data sets. Any positive number x can be written base B as $x = S_B(x) \cdot B^{k(x)}$, where $S_B(x) \in [1, B)$ is the significand and $k(x)$ is an integer. We now say a data set is Benford, or satisfies Benford's Law, base B if

$$\text{Prob}(\text{First digit base } B \text{ is } d) = \log_B\left(1 + \frac{1}{d}\right); \tag{1}$$

thus base 10 the probabilities decrease from roughly 30% for a leading digit of 1 down to about 4.6% for a 9.

Sometimes, however, one uses Benford to refer to not just the first digit, but the distribution of the significand (alternatively one often says the system is strongly Benford); thus

$$\text{Prob}(\text{Significand base } B \text{ is at most } s) \;=\; \log_B(s). \tag{2}$$

Note that the first digit probabilities follow immediately from the significand, as

$$\text{Prob}(\text{First digit base } B \text{ is } d) \;=\; \text{Prob}(\text{Significand base } B \text{ is in } [d, d+1))$$

$$=\; \log_B(d+1) - \log(d) \;=\; \log_B\left(1 + \frac{1}{d}\right), \tag{3}$$

and we can easily determine the probabilities of the second digits:

$$\text{Prob}(\text{Second digit base } B \text{ is } d) \;=\; \sum_{k=1}^{B-1} \log\left(1 + \frac{1}{Bk + d}\right). \tag{4}$$

We see a set is Benford base B if and only if its logarithms base B are equidistributed modulo 1; thus the digit bias disappears under a logarithmic transformation. For more on the history, theory and applications of Benford's law see the survey articles [21–24] and the books [3,25,26].

Table 1 provides 1-BL 10 and 2-BL 10 probabilities, namely the probability of a first or second digit being d base 10. Figure 2 provides a visual distribution of the first digits.

Figure 2.

d	Probability First Digit d	Probability Second Digit d
0		0.1197
1	0.3010	0.1139
2	0.1761	0.1088
3	0.1249	0.1043
4	0.0969	0.1003
5	0.0792	0.0967
6	0.0669	0.0934
7	0.0580	0.0904
8	0.0512	0.0876
9	0.0458	0.0850

Table 1. Benford probabilities for first and second digit (to four decimal places).

Table 2 provides analogous probabilities for the distribution of the first and second digits of 1-BL 3 and second digit base 3 (2-BL 3) probabilities, and Figure 3 provides a visual representation of the distribution of the leading digits under base 3 representations.

Figure 3.

d	Probability First Digit d	Probability Second Digit d
0		0.4022
1	0.6309	0.3247
2	0.3691	0.2732

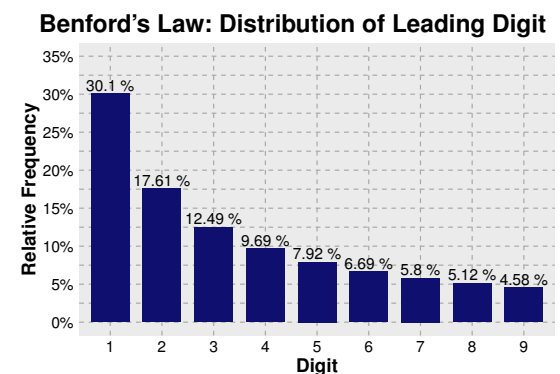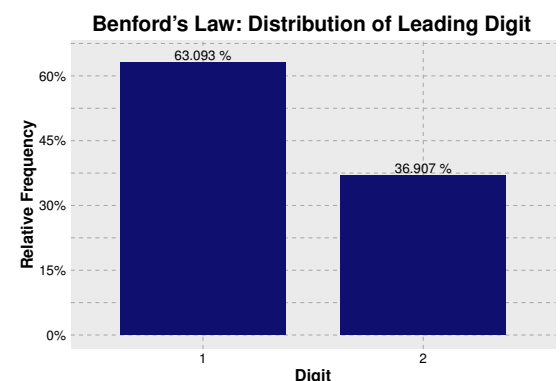**Table 2.** Benford probabilities for first digit base 3 (to four decimal places).

One of the most utilized tests to see if a set is Benford is to compute the chi-squared statistic, with comparisons of the associated cumulative distribution function to statistical significance levels, usually $\alpha = 0.01$ or 0.05. If $E(d)$ is the expected number that have a first digit of d base B, and $O(d)$ is the observed number, then the chi-squared statistic is

$$\chi^2 = \sum_{d=1}^{B-1} \frac{(O(d) - E(d))^2}{E(d)}. \tag{5}$$

If we work in base $B = 10$ then as we have 9 choices for the first digit and the probabilities sum to 1, we only have 8 degrees of freedom (the probability of the last digit is forced from the other 8, and zero is never a first digit). If the digit distribution is drawn from Benford's law, then 95% of the time we should have a chi-square value of 15.5 or less, and 99% of the time it should be 20.1 or less. To put it another way, if the data follows Benford's law and we observe a value greater than 20.1 , there is only a 1% chance of that happening *given that the data is Benford*; if you observe values larger than 20.1 the probability falls rapidly.

If we work in base 3, there are only two possibilities for the leading digits. For a set that is Benford we expect to observe a leading digit of 1 approximately 63.093% of the time and of 2 about 36.907%, with the chi-square values (with one degree of freedom) being 3.84 (at the 95% level) and 6.63 (at the 99% level).

In the following section, we modeled the skewed distributions of precinct level election data for the purposes of analyzing these distributions with respect to Benford's law.

3. Empirical Estimates Using 2004 Election Data

3.1. Modeling with the Discrete Weibull

As is apparent in Figure 1, precinct by precinct election data can be skewed by vote counts heavily occurring in a just a few precincts. The discrete Weibull distribution provides a flexible family of distributions useful for modeling this type of skewed data [27]. The family, as well as associated variants, have been used to model real-world phenomena including soccer, respiratory capacity, fertility, and levels of microorganisms among others [28–32]. In this section, we utilize this family to model precinct by precinct election data.

Specifically, in state i and county j consisting of $k = \{1, \ldots, K\}$ precincts, we can model the vote count for two candidates x and y, $\mathbf{x_{i,j}} = \{x_{i,j,1}, \ldots, x_{i,j,K}\}$ and $\mathbf{y_{i,j}} = \{y_{i,j,1}, \ldots, y_{i,j,K}\}$, using a discrete Weibull representation [27] as follows:

$$P(x_{i,j,k}; \alpha_{x,i,j}, \beta_{x,i,j}) = \exp\left[-\left(\frac{x_{i,j,k}}{\alpha_{x,i,j}}\right)^{\beta_{x,i,j}}\right] - \exp\left[-\left(\frac{x_{i,j,k}+1}{\alpha_{x,i,j}}\right)^{\beta_{x,i,j}}\right] \tag{6}$$

and:

$$P(y_{i,j,k}; \alpha_{y,i,j}, \beta_{y,i,j}) = \exp\left[-\left(\frac{y_{i,j,k}}{\alpha_{y,i,j}}\right)^{\beta_{y,i,j}}\right] - \exp\left[-\left(\frac{y_{i,j,k}+1}{\alpha_{y,i,j}}\right)^{\beta_{y,i,j}}\right]. \tag{7}$$

3.2. Data and Model Estimation

We estimated the discrete Weibull model on data from the 2004 Presidential Election between incumbent President George W. Bush and Senator John F. Kerry available online on Harvard University's DataVerse repository [33]. The repository is a comprehensive database providing precinct by precinct vote counts across counties for many states of the 2004 as well as other Presidential, Senatorial, and Congressional elections. Although there were a number of other candidates appearing on the ballot in the general election, we focused our analysis on the two main candidates in several selected states. Specifically, to

estimate the model on states with varying political inclinations, we analyzed one "red" state (North Carolina), one "blue" state (Vermont), and three "battleground" states (Colorado, Ohio, and Wisconsin). We verified the data correctly represents vote counts reported by the United States Federal Election Commission for all five states, with the exception of Wisconsin, having an error of less than 0.6% for each candidate [34]. In order to ensure a reasonable sample size for our analysis, we restricted our analysis to counties consisting of more than ten election day precincts.

For such counties, we estimated the scale and shape parameters for the distribution of the precinct by precinct vote tallies for each of the two candidates x and y in state i and county j - $\alpha_{x,i,j}$, $\beta_{x,i,j}$, $\alpha_{y,i,j}$, and $\beta_{y,i,j}$ - via the method of maximum likelihood (MLE). Subsequently, to assess goodness of fit, we performed a series of two sample Kolmogorov-Smirnov (KS) tests assessing the null hypothesis of distributional equality, comparing the distribution of $\mathbf{x_{i,j}}$, or $\mathbf{y_{i,j}}$, with a random sample of size K of the associated discrete Weibull distribution using parameter estimates obtained via MLE. We averaged p-values obtained from 100 simulations to obtain a pooled Monte Carlo estimate of the goodness of fit of the estimated distributions. Our estimated shape and scale parameters, as well as pooled p-value estimates, for the battleground states of Ohio, Colorado, and Wisconsin are presented in the scatterplots in Figure 4 for counties having at least ten precincts. Scatterplots of MLE estimates and KS-test results for Vermont and North Carolina are ommitted for brevity and are available from the authors upon request.
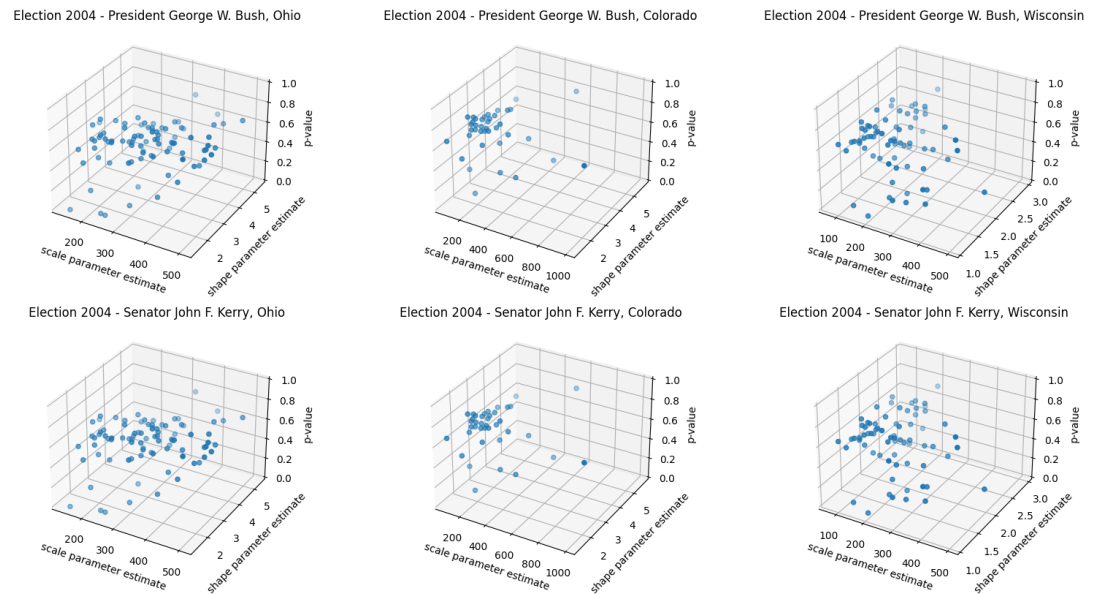

Figure 4. Discrete Weibull parameter estimation and Kolmogorov Smirnov Goodness of Fit Test Results for Ohio, Colorado, and Wisconsin, US 2004 Presidential Election. *p*-values above 0.05 indicate conformance to discrete Weibull distribution.

As Figure 4 illustrates, for all three battleground states, the discrete Weibull distributions fit quite well. Altogether there are three counties in Ohio that have pooled p-values less than 0.05, one county in Wisconsin, and none in Colorado for Bush. For Kerry, there were eight such counties in Ohio, one county in Wisconsin, and none in Colorado. After performing the standard, albeit conservative, Bonferroni corrections there was only one county - Cuyahoga County, Ohio - for Bush, that was statistically significant ($p < 4\text{x}10^{-5}$). These results demonstrate the efficacy of using discrete Weibull distribution to model precinct level data across counties.

	$\hat{\alpha}_{Bush,i,j}$	$\hat{\beta}_{Bush,i,j}$	p-value	Number of Precincts
Milwaukee County, WI	351.457	1.410	0.321	560
Stark County, OH	287.717	2.289	0.472	364
Grand County, CO	205.389	2.706	0.726	12

Table 3. Maximum Likelihood Estimates and Goodness of Fit Test Results for Vote Counts for Selected Counties in 2004 US Presidential Election - President George W. Bush. p-values above 0.05 indicate conformance to discrete Weibull distribution.

	$\hat{\alpha}_{Kerry,i,j}$	$\hat{\beta}_{Kerry,i,j}$	p-value	Number of Precincts
Milwaukee County, WI	598.1	2.2	0.223	560
Stark County, OH	289.057	4.085	0.304	364
Grand County, CO	167.35	3.188	0.614	12

Table 4. Maximum Likelihood Estimates and Goodness of Fit Test Results for Vote Counts for Selected Counties in 2004 US Presidential Election - Senator John F. Kerry. p-values above 0.05 indicate conformance to discrete Weibull distribution.

Furthermore, as Figure 4 also illustrates, our estimated parameters vary notably, depending on the county. Tables 3 and 4 provide MLE parameter estimates and p-values the KS tests, and precinct sizes for the selected counties presented in Figure 1.

Tables 3 and 4 illustrate the robustness of the discrete Weibull distribution in its ability to model different types of counties across the country. Namely, Milwaukee county, WI has larger scale and smaller shape parameters than the less densely populated Stark County, OH, and Grand County, CO. Regardless, the p-values for all three counties are well above 0.05, demonstrating the robustness of the choice of the discrete Weibull distribution across different counties.

This analysis in the section illustrates the efficacy of using the discrete Weibull distribution at fitting precinct level election data. Furthermore, as is evident in Figure 3, the estimates of the scale and shape parameters approximately vary between $[100, 600]$ and $[1, 6]$ respectively. In the following section, we utilize these facts to examine the usefulness of 1-BL 3 and 1-BL 10 in analyzing election data.

4. Comparisons of 1-BL 3 and 1-BL 10 to Potential Vote Distributions

In the previous section, we demonstrated the usefulness of the discrete Weibull distribution in modeling precinct level election data. In this section, we examine a variety of choices of discrete Weibull parmaeterizations and assess their adherence to 1-BL 3 and 1-BL 10 via Monte Carlo analysis. Specifically, based on the results presented in Figure 4, we simulated county level data with the number of simulated precincts varying between 50 and 300. We varied the scale parameter $\alpha_{x,i,j}$ between 100 and 600 in increments of 25 and varied the shape parameter $\beta_{x,i,j}$ between 1 and 6 in increments of 1. After compiling the expected and observed number of first digits under both base 10 and base 3 representations, we performed chi-squared tests for both 1-BL 10 and 1-BL 3 (assuming under the null hypothesis of distributional equality, chi-squared distributions having eight and one degree of freedom respectively as discussed in Section 2) and subsequently computed associated p-values. We averaged these p-values over the course of 25 Monte Carlo simulations. Our results are presented in Figures 5 and 6.

As is apparent Figure 5, assuming a statistical significance level of α=0.05 the simulated precinct data rarely conforms to 1-BL 10 except for choices of scale parameter $\alpha_{x,i,j}$ between 100 and 200. Upon conversion to base 3, adherence to Benford's law is considerably more robust, as is illustrated in Figure 6. Even under base 3 representations, however, Benford's law is not always observed. For example, if the scale parameter $\alpha_{x,ij}$ exceeds 500 or the
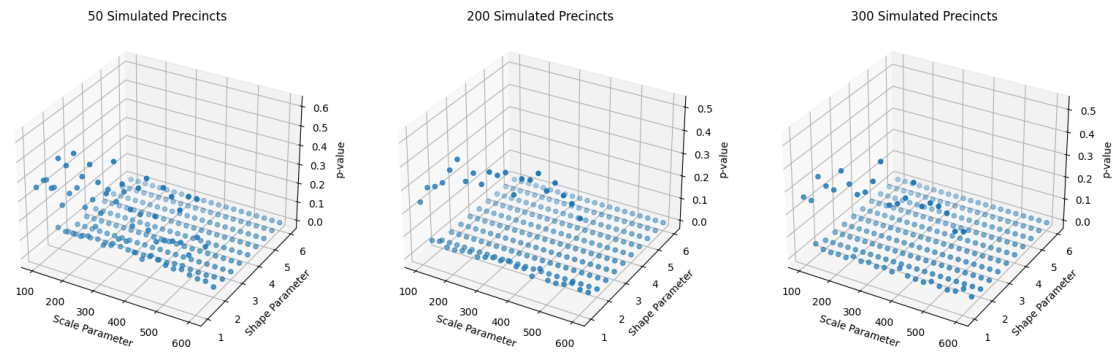
Figure 5. First Digit Base 10 Monte Carlo Analysis of Simulated Discrete Weibull Distributions with Various Choices of Parameterizations and Precinct Sizes. *p*-values above 0.05 indicate conformance to 1-BL 10.
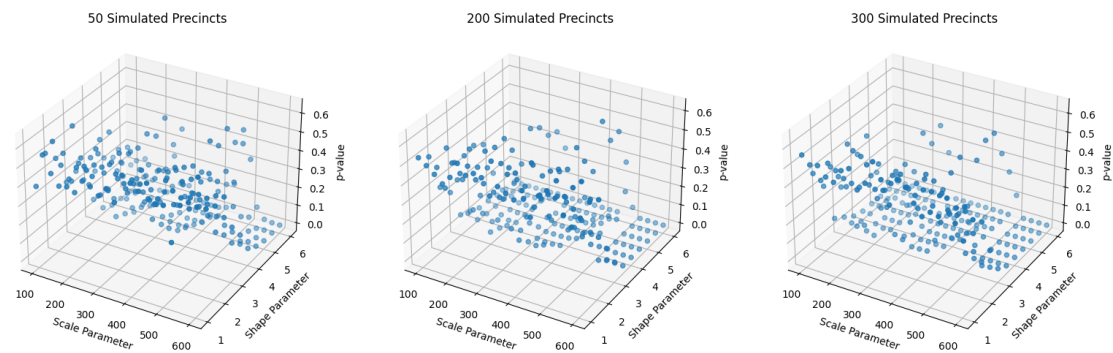

Figure 6. First Digit Base 3 Monte Carlo Analysis of Simulated Discrete Weibull Distributions with Various Choices of Parameterizations and Precinct Sizes. *p*-values above 0.05 indicate conformance to 1-BL 3

State	George W. Bush (R)	John F. Kerry (D)	Number of Counties
North Carolina	56.02% (1,961,166)	43.58% (1,525,849)	100
Vermont	38.80% (121,180)	58.94% (184,067)	16
Wisconsin	49.32% (1,478,120)	49.70% (1,489,504)	72
Ohio	50.81% (2,859,768)	48.71% (2,741,167)	88
Colorado	51.69% (1,101,255)	47.02% (1,001,732)	64

Table 5. States from 2004 US Presidential Election used for Analysis.

shape parameter $\beta_{x,i,j}$ exceeds 4, the associated *p*-values are below 0.05 and thus Benford's behavior is not observed.[1]

Our analysis of this synthetic election data using the wide choices of parameters motivated by our MLE estimation in Section 3 illustrates that 1-BL 3 is considerably more effective than 1-BL 10 in fitting election data. Thus, base 3 representations of precinct by precinct data are considerably much more likely to adhere to Benford's law than standard base 10 representations for most of the choices of parameterizations presented in Figures 5 and 6. These results thus provide theoretical justification for applying 1-BL 3 to election data, which we perform in the following section.

5. Applications of 1-BL 3 to the 2004 US Presidential Election

As a result of our analysis in Section 4, we now have a theoretical justification for applying 1-BL 3 to election data. In this section, we do so by utilizing the data from the 2004 US Presidential Election discussed in Section 3. Specifically, we analyzed counties amongst the states of Colorado, North Carolina, Ohio, Vermont, and Wisconsin to assess adherence to 1-BL 3 on a county by county basis. The electoral outcomes of these states is summarized in Table 5.

As we did in Section 3, we restricted our analysis to counties voting on election day that consisted of at least ten election day precincts. We assumed a statistical significance level $\alpha = 0.05$ and performed comparisons of observed and expected values. Specifically, as discussed in section 2, we did so by utilizing a chi-squared distribution with one degree of freedom to test the null hypothesis of distributional equality between these observed and expected counts. Afterwards, we computed adjusted *p*-values using the Benjamni-Hochberg procedure to control for the false discovery rate [35] as has been done in other election related work using Benford's law (see for example [9,36,37]). Our results are presented in Table 6.

As Table 6 illustrates, all counties amongst the non-competitive states of North Carolina and Vermont as well as the battleground state of Wisconsin conform to 1-BL3. On the other hand, there are four distributions amongst counties/candidates in Colorado and fifteen distributions in Ohio that are statistically anomalous.

However, as Figure 6 from Section 4 demonstrates, depending on the distribution of votes for either candidate, certain counties may not necessarily conform to 1-BL 3 to begin with. As a result, we performed another series of Monte Carlo simulations for the anomalous counties in Table 6 to assess whether the estimated vote distribution for each county/candidate indeed conforms to Benford's law. To do so, we again ran 25 Monte Carlo simulations drawing from discrete Weibull distributions using the county's estimated parameters (presented as part of Figure 4) and reported precinct sizes for the anomalous counties in Table 6 to assess whether these distributions indeed conform to 1-BL 3. We averaged *p*-values across the 25 simulations and compared these values to those in Table 6. We flagged counties where the estimated distribution conforms to Benford's law (*p*-value >

---

[1] This behavior persists even after altering $\alpha = 0.05$ for multiple comparisons via Bonferroni corrections to $\alpha = 1.67 \times 10^{-4}$.

State	County	Candidate	χ^2 Stat	*p*-value	Number of Precincts	adjusted *p*-value
Colorado	Arapahoe	John F. Kerry	21.152	<0.001	364	0.001
	El Paso	John F. Kerry	20.513	<0.001	378	0.001
	Jefferson	George W. Bush	38.050	<0.001	324	<0.001
	Jefferson	John F. Kerry	108.747	<0.001	324	<0.001
North Carolina	No Anomalies Detected					
Wisconsin	No Anomalies Detected					
Ohio	Ashtabula	John F. Kerry	39.385	<0.001	127	<0.001
	Butler	John F. Kerry	14.594	0.001	289	0.013
	Geauga	George W. Bush	15.196	<0.001	96	0.013
	Geauga	John F. Kerry	20.812	<0.001	96	0.001
	Greene	George W. Bush	13.860	0.001	142	0.016
	Greene	John F. Kerry	36.189	<0.001	142	<0.001
	Lorain	John F. Kerry	43.509	<0.001	239	<0.001
	Miami	John F. Kerry	14.669	0.001	82	0.013
	Muskingum	John F. Kerry	13.971	0.001	85	0.016
	Portage	John F. Kerry	27.249	<0.001	129	<0.001
	Summit	John F. Kerry	68.917	<0.001	475	<0.001
Vermont	No Anomalies Detected					

Table 6. First Digit Base 3 Benford's Analysis on Selected Battleground and Non-Battleground States in the US 2004 Presidential Election.

County	Candidate	$\alpha_{x,i,t}$	$\beta_{x,i,t}$	Number of Precincts	*p*-value (KS -test)	Should conform to BL but fails?
Arapahoe	John F. Kerry	326.634	2.973	364	0.003	
El Paso	John F. Kerry	231.420	2.499	378	0.507	*
Jefferson	George W. Bush	400.330	3.650	324	0.001	
Jefferson	John F. Kerry	352.800	4.717	324	<0.001	
Ashtabula	John F. Kerry	208.527	4.538	127	<0.001	
Butler	John F. Kerry	218.783	2.869	289	0.037	
Geauga	George W. Bush	348.690	4.200	96	0.003	
Geauga	John F. Kerry	228.818	3.869	96	0.099	*
Greene	George W. Bush	383.344	2.628	142	0.395	*
Greene	John F. Kerry	244.027	2.272	142	0.494	*
Lorain	John F. Kerry	367.292	2.964	239	0.074	*
Miami	John F. Kerry	235.479	4.963	82	0.010	
Muskingum	John F. Kerry	215.593	3.613	85	0.054	*
Portage	John F. Kerry	347.487	4.109	129	<0.001	
Summit	John F. Kerry	365.481	3.712	475	<0.001	

Table 7. Analysis of Whether Anomalous Counties from Table 6 Deviate from Benford's Law.

0.05) while the actual data does not (adjusted *p*-value <0.05).[2] Our results are presented in Table 7.

As Table 7 shows, the vote distribution for Senator John F. Kerry in El Paso county, Colorado should theoretically adhere to 1-BL 3 but does not. There were four additional counties in Ohio for Kerry that should also have conformed to 1-BL 3 that failed (Geuga, Greene, Lorain, an, Muskingum Counties). There was one such county for President George W. Bush - Greene County, Ohio. Furthermore, as Table 7 also illustrates, although several other distributions for both candidates were deemed statistically anomalous in Table 6, their estimated distributions do not either, and thus their non-adherenece to 1-BL 3 may not be unnatural.

6. Discussion

The results in Table 7 flag several counties for Senator Kerry and one for President Bush as statistically anomalous in terms of 1-BL 3. Interestingly, Ohio was one of the most contentious states in the 2004 election [38]. Although our model flags approximately five

---

2   The *p*-values presented in Table 7 are estimates before correcting for multiple comparisons; therefore, our approach can be considered conservative as it provides a lower bound on the number of statistically anomalous counties, and thus likely understates the number of anomalous counties.

percent of Ohio's 88 counties, these statistical anomalies are still detected after corrections due to multiple comparisons. These results, however, neither offer any indication as the cause of potential malfeasance, if present, nor do they shed light on which side was responsible. One goal of Benford tests is to detect fraud; however, certain types cannot be observed by these methods. For example, if a fixed percentage of votes were switched from one candidate to another while keeping the number of voters fixed, the net effect is to just change the percentage of support for each candidate. Similarly, if a fixed number of ballots are added in each precinct, the effect would be to both increase the number of voters and adjust the candidates' support. It is important to remember tests such as these cannot prove fraud; the main utility is to highlight situations that are worth additional investigation; for example, in cases where 1-BL 3 fails it might be valuable to check current precinct voting against previous years when 1-BL 3 held.

Again, although malfeasance could constitute one explanation of non-conformance to Benford's law, associated anomalies do not necessarily prove malfeasance as other explanations are be possible. For example, Mebane (2010, 2012, 2013) has discussed how deviations from 2-BL 10 could be due not only to fraud but also may be manifestations of mobilization efforts, last minute changes to voting preference, as well as effects of gerrymandering among others [11,13,14]. The effect of these factors on conformance 1-BL3 has not been studied thus far and is a worthy topic of future research. If 1-BL3 can indeed manifest these effects, it would be useful for future research to differentiate between these effects and malfeasance, to render these techniques more useful for auditing or potential litigation.

7. Conclusions and Future Research

Our study illustrates the usefulness of 1-BL 3 in analyzing elections, flagging a number of counties in the 2004 US Presidential Election. Our analysis in Section 4, suggests that 1-BL 10 analysis would almost surely unnecessarily flag many other counties that may not truly be anomalous. As discussed in Section 6, however, although potential malfeasance could constitute one explanation of 1-BL 3 based anomalies, there may be other potential explanations as well. Future research should look into refining these number theoretical models to differentiate amongst potential explanations. Regardless, if there are concerns about potential malfeasance, 1-BL 3 could indeed be a useful tool to detect for anomalies as a basis for potential auditing.

Although we have analyzed several states from the 2004 US Presidential Election, we hope our work encourages future research to apply our method to other elections of all levels, both American and foreign. However, as noted by Mebane (2011), the appropriate unit of analysis for Benford's law is precinct level data across counties, as we have utilized in this study [6]. As a result, Benford's law is not useful for detecting potential malfeasance with absentee ballots such as in the 2018 Congressional election in North Carolina's ninth district, but there are other statistical techniques have been demonstrated to be useful for such absentee ballot analysis [39]. The number of statistical tools to analyze the integrity of elections is growing [36,40–43], and the techniques presented in this paper thus should be viewed as yet another tool in the statistician's arsenal to analyze election data. Future research should thus include our tool as part of a suite of such tools to comprehensively analyze all aspects of election data.

One potential limitation of this approach is that the data may become so heavily spread out via base three representations that some instances of malfeasance may be rendered undetectable. Other base representations may potentially ameliorate this problem at the expense of clustering the data. Future research should weigh such tradeoffs between different bases.

Moreover, until now, 2-BL 10 was the primary technique used for applying Benford's law to election data. Future research should compare the efficacy of 1-BL 3 to 2-BL 10. It would be useful to simulate election data as we have done in the study, externally impose fraud in the model, and assess what forms of fraud are detectable by these techniques.

Finally, there is of course no a priori reason to apply these techniques solely to elections, and we hope that this research encourages the use of 1-BL 3 in other applied settings of clustered data where 1-BL 10 is not applicable. Batting averages in baseball, serving speeds in tennis, government-reported statistics, and stock market data are indeed worthy examples of future investigation.

It is always useful to have tools to identify potential anomalies in election data. We hope the techniques presented here provide a useful addition to the applied mathematician/statistician's toolbox for doing so.

1. Engel, H.A.; Leuenberger, C. Benford's law for exponential random variables. *Statistics & Probability Letters* **2003**, *63*, 361–365.
2. Miller, S.J.; Nigrini, M.J. Order statistics and Benford's law. *International Journal of Mathematics and Mathematical Sciences* **2008**, *2008*.
3. Miller, S.J.; Berger, A.; Hill, T. Theory and Applications of Benford's Law, 2015.
4. Berger, A.; Hill, T.P. Benford's law strikes back: No simple explanation in sight for mathematical gem. *The Mathematical Intelligencer* **2011**, *33*, 85.
5. Deckert, J.; Myagkov, M.; Ordeshook, P.C. Benford's Law and the detection of election fraud. *Political Analysis* **2011**, *19*, 245–268.
6. Mebane, W.R. Comment on "Benford's Law and the detection of election fraud". *Political Analysis* **2011**, *19*, 269–272.
7. Mebane, W. Inappropriate applications of Benford's law regularities to some data from the 2020 Presidential election in the United States, 2020.
8. Kalinin, K.; Mebane, W.R. Understanding electoral frauds through evolution of Russian federalism: From "bargaining loyalty" to "signaling loyalty". In Proceedings of the Annual Meeting of the Midwest Political Science Association, Chicago, IL, March, 2011, pp. 1–26.
9. Mebane Jr, W.R. Election forensics: Vote counts and Benford's law. In Proceedings of the Summer Meeting of the Political Methodology Society, UC-Davis, July, 2006, Vol. 17.
10. Mebane, W.R.; Larijani, A.; Mehralizadeh, M.; Moeen, M.; Hashemi, A.; These, R. Note on the presidential election in Iran, June 2009 **2009**.
11. Mebane, W.R. Election fraud or strategic voting? Can second-digit tests tell the difference. In Proceedings of the Summer Meeting of the Political Methodology Society, University of Iowa, 2010.
12. Mebane Jr, W.R. Fraud in the 2009 presidential election in Iran? *Chance* **2010**, *23*, 6–15.
13. Mebane, W.R. Second-digit tests for voters' election strategies and election fraud. In Proceedings of the Annual Meeting of the Midwest Political Science Association, Chicago. Citeseer, 2012.
14. Mebane Jr, W.R. Election forensics: The meanings of precinct vote counts' second digits. In Proceedings of the Summer Meeting of the Political Methodology Society, University of Virginia, 2013.
15. Mebane Jr, W.R.; Kalinin, K. Electoral falsification in Russia: Complex diagnostics selections 2003-2004, 2007-2008 (in Russian). *Electoral Policy REO* **2009**, *2*, 57–70.
16. Mebane Jr, W.R.; Kalinin, K. Electoral fraud in Russia: Vote counts analysis using second-digit mean tests. In Proceedings of the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April, 2010, Vol. 20.
17. Mebane Jr, W.R.; Kent, T.B. Second Digit Implications of Voters' Strategies and Mobilizations in the United States during the 2000s. In Proceedings of the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April, 2013, pp. 11–14.
18. Kossovsky, A.E.; Miller, S.J. Report on Benford's Law Analysis of 2020 Presidential Election Data.
19. Newcomb, S. Note on the frequency of use of the different digits in natural numbers. *American Journal of mathematics* **1881**, *4*, 39–40.
20. Benford, F. The law of anomalous numbers. *Proceedings of the American philosophical society* **1938**, pp. 551–572.
21. Hill, T.P. A statistical derivation of the significant-digit law. *Statistical science* **1995**, pp. 354–363.
22. Hill, T.P. 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* **1998**, *86*, 358–363.
23. Nigrini, M.J. *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*; Vol. 586, John Wiley & Sons, 2012.
24. Raimi, R.A. The first digit problem. *The American Mathematical Monthly* **1976**, *83*, 521–538.
25. Berger, A.; Hill, T.P. *An introduction to Benford's law*; Princeton University Press, 2015.
26. Kossovsky, A.E. *Benford's law: theory, the general law of relative quantities, and forensic fraud detection applications*; Vol. 3, World Scientific, 2014.
27. Nakagawa, T.; Osaki, S. The discrete Weibull distribution. *IEEE transactions on reliability* **1975**, *24*, 300–301.

28. Peluso, A.; Vinciotti, V.; Yu, K. Discrete Weibull generalized additive model: An application to count fertility data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **2019**, *68*, 565–583.

29. El-Morshedy, M.; Eliwa, M.; El-Gohary, A.; Khalil, A. Bivariate exponentiated discrete Weibull distribution: statistical properties, estimation, simulation and applications. *Mathematical sciences* **2020**, *14*, 29–42.

30. Englehardt, J.D.; Li, R. The discrete Weibull distribution: an alternative for correlated counts with confirmation for microbial counts in water. *Risk Analysis: An International Journal* **2011**, *31*, 370–381.

31. Patriarca, R.; Hu, T.; Costantino, F.; Di Gravio, G.; Tronci, M. A system-approach for recoverable spare parts management using the discrete Weibull distribution. *Sustainability* **2019**, *11*, 5180.

32. Para, B.A.; Jan, T. Compound of discrete Weibull and Minimax distributions as a new count data model with application in genetics. *Applied Mathematics & Information Sciences Letters* **2017**, *5*, 113–124.

33. Ansolabehere, S.; Palmer, M.; Lee, A. Precinct-Level Election Data, 2002-2012. https://dataverse.harvard.edu/, 2015. Harvard Dataverse.

34. 2004 Election Results. https://www.fec.gov/resources/cms-content/documents/2004tables.pdf. Accessed: 6-22-2022.

35. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **1995**, *57*, 289–300.

36. Mebane, W.R.; Alvarez, R.M.; Hall, T.E.; Hyde, S.D. Election forensics: The second-digit Benford's law test and recent American presidential elections. *Election fraud: Detecting and deterring electoral manipulation* **2008**, pp. 162–181.

37. Mebane Jr, W.R. Election forensics: Statistics, recounts and fraud. In Proceedings of the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April. Citeseer, 2007, pp. 12–16.

38. Dao, J. The 2004 Campaign Battlegrounds; Ohio. *The New York Times* **2004**.

39. Herron, M.C. Mail-In Absentee Ballot Anomalies in North Carolina's 9th Congressional District. *Election Law Journal: Rules, Politics, and Policy* **2019**, *18*, 191–213.

40. Zhang, M.; Alvarez, R.M.; Levin, I. Election forensics: Using machine learning and synthetic data for possible election anomaly detection. *PloS one* **2019**, *14*, e0223950.

41. Lacasa, L.; Fernández-Gracia, J. Election forensics: Quantitative methods for electoral fraud detection. *Forensic science international* **2019**, *294*, e19–e22.

42. Mebane Jr, W.R. Election forensics: Statistical interventions in election controversies. In Proceedings of the Annual Meeting of the American Political Science Association, 2007, Vol. 13.

43. Mebane Jr, W.R.; Klaver, J. Election Forensics: Strategies versus Election Frauds in Germany. In Proceedings of the Annual Conference of the European Political Science Association, Vienna, Austria, 2015.