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

Katie Anderson
Brigham Young University - Idaho
andersonkatie1998@gmail.com

Joint with Kevin Dayaratna, Drew Gonshorowski, and Steven J. Miller.

International Conference on Advances in Interdisciplinary Statistics and Combinatorics
October 8, 2022

Presentation Overview

- Introduction to Benford's Law
- Previous research
- A new proposal for Benford first digit analysis
- Application to American elections
- Conclusion and avenues for future research

Introduction to Benford's Law

Any positive real number x can be written base B as

$$x = S_B(x) \cdot B^{k(x)}$$

where $k(x)$ is an integer and $S_B(x) \in [1, B)$. The value $S_B(x)$ is known as the significand.

Introduction to Benford's Law

Any positive real number x can be written base B as

$$x = S_B(x) \cdot B^{k(x)}$$

where $k(x)$ is an integer and $S_B(x) \in [1, B)$. The value $S_B(x)$ is known as the significand.

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

Introduction to Benford's Law

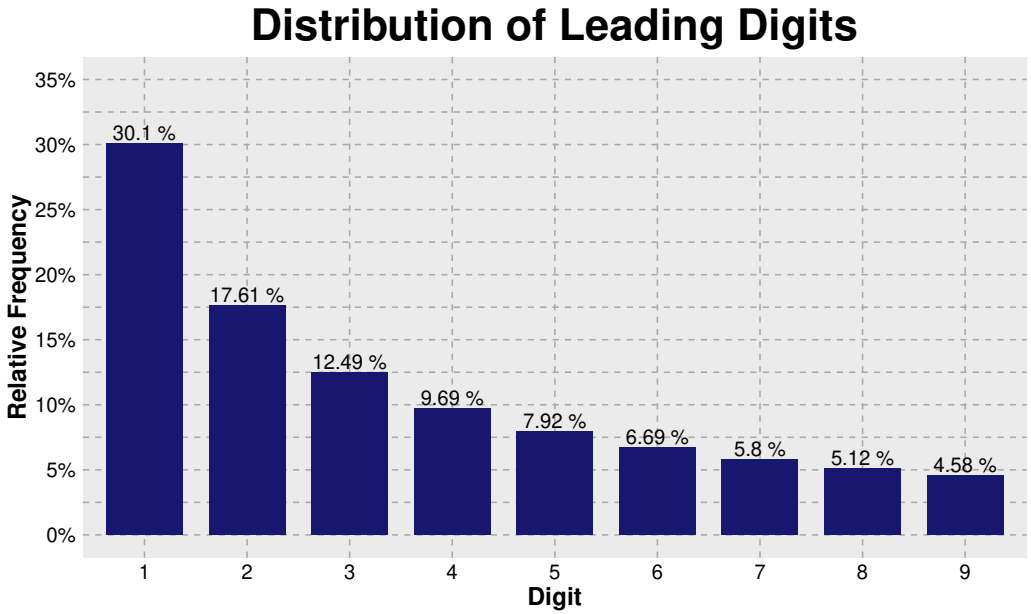
A data set is Benford base B if

$$\begin{aligned}
&\text{Prob}(\text{First digit base } B \text{ is } d) \\
&= \text{Prob}(\text{Significand base } B \text{ is in } [d, d+1)) \\
&= \log_B(d+1) - \log_B(d) \\
&= \log_B\left(1 + \frac{1}{d}\right).
\end{aligned}$$

Introduction to Benford's Law

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

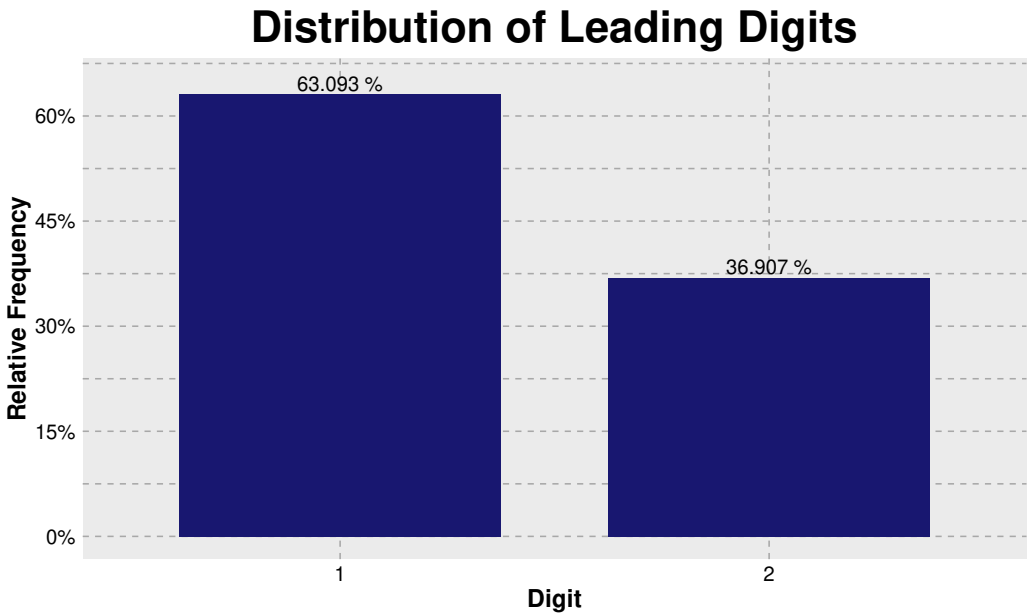
Benford probabilities for first and second digit (to four decimal places).



Introduction to Benford's Law

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

Benford probabilities for first and second digit base 3 (to four decimal places).



Previous Research

- Mebane uses second digit analysis on precinct level election data. He analyzes the 2004 election in Florida, the 2006 election in Mexico, and others with this approach.
- Meban discusses that deviations could be due not only to malfeasance, but strategic voting and gerrymandering.

New Proposal

- Not all data sets are Benford.
- Clustered data sets are a sign that the data is not Benford.
- Changing data to base 3 spreads out clustered data. For example, the value 81 in base 10 is $10,000_3$ in base 3.
- It is possible for the same data to be Benford base 3 and not Benford base 10.

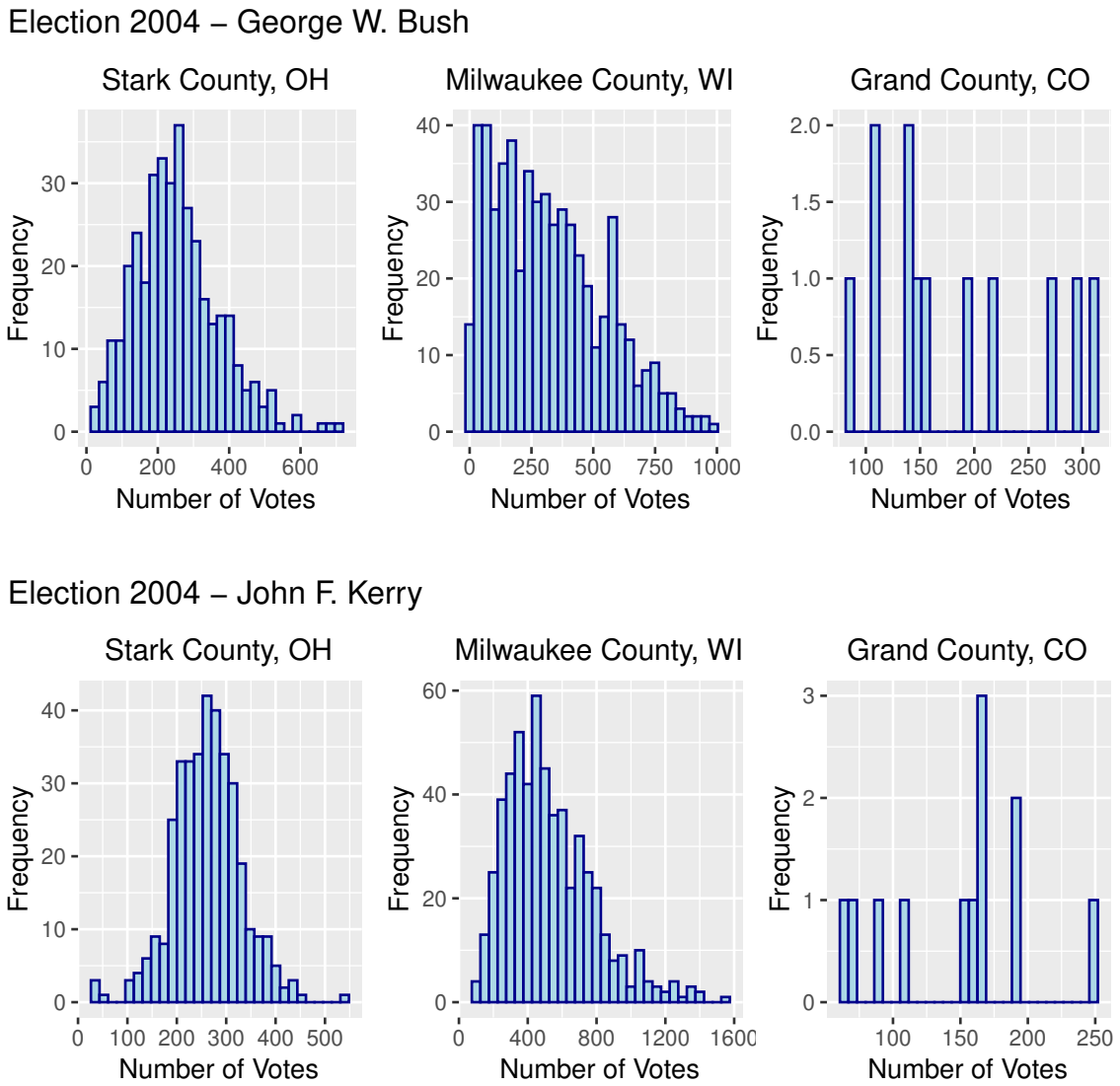


Figure: Vote Distributions Across Precincts in US 2004 Presidential Election for Selected Counties.

Introduction to Discrete Weibull Distribution

Let $\alpha > 0$ and $\beta > 0$. Let x be an integer. The probability mass function for the family of discrete Weibull distributions is

$$P(x|\alpha,\beta) = \exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right] - \exp\left[-\left(\frac{x+1}{\alpha}\right)^{\beta}\right]$$

where α is the scale parameter and β is the shape parameter.
The Cumulative Distribution Function for the family of discrete Weibull distributions is

$$P(X \leq x|\alpha,\beta) = 1 - \exp\left[-\left(\frac{x+1}{\alpha}\right)^{\beta}\right]$$

where α is the scale parameter and β is the shape parameter.

Application to American Elections

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. We take

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

and

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

where α is the scale parameter and β is the shape parameter.

States Used for Analysis

State	George W. Bush (R)	John F. Kerry (D)	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: States from 2004 US Presidential Election used for Analysis.

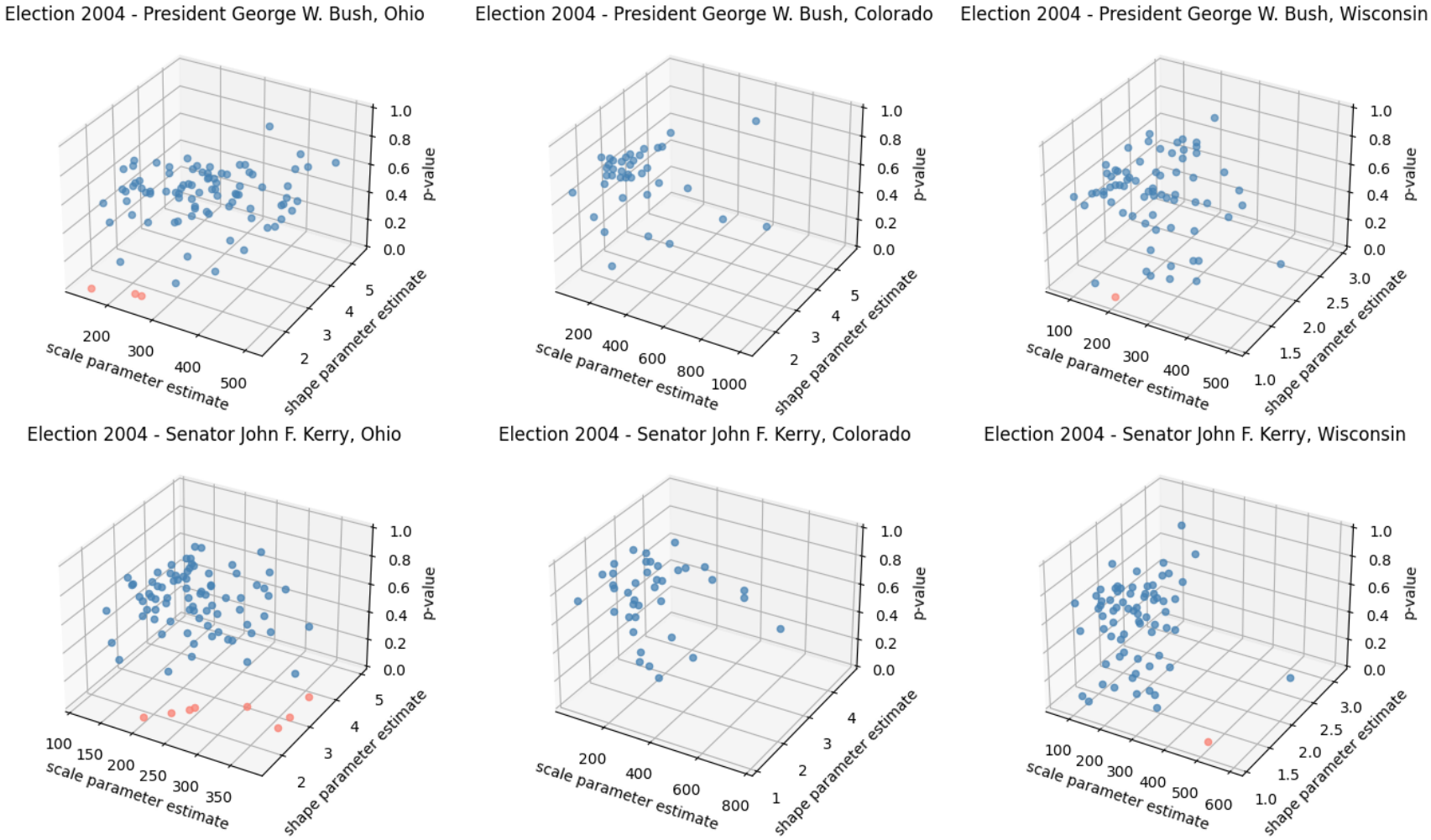


Figure: Discrete Weibull parameter estimation and Kolmogorov Smirnov Goodness of Fit Test Results for Ohio, Colorado, and Wisconsin, US 2004 Presidential Election. The *p*-values above 0.05 (plotted in blue) indicate conformance to the discrete Weibull distribution.

Maximum Likelihood Estimates and Goodness of Fit Test Results for Vote Counts for Selected Counties in 2004 US Presidential Election. The *p*-values above 0.05 indicate conformance to discrete Weibull distribution.

	$\hat{\alpha}_{Bush,i,j}$	$\hat{\beta}_{Bush,i,j}$	*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

	$\hat{\alpha}_{Kerry,i,j}$	$\hat{\beta}_{Kerry,i,j}$	*p*-value	Number of Precincts
Milwaukee County, WI	598.1	2.200	0.223	560
Stark County, OH	289.057	4.085	0.304	364
Grand County, CO	167.35	3.188	0.614	12

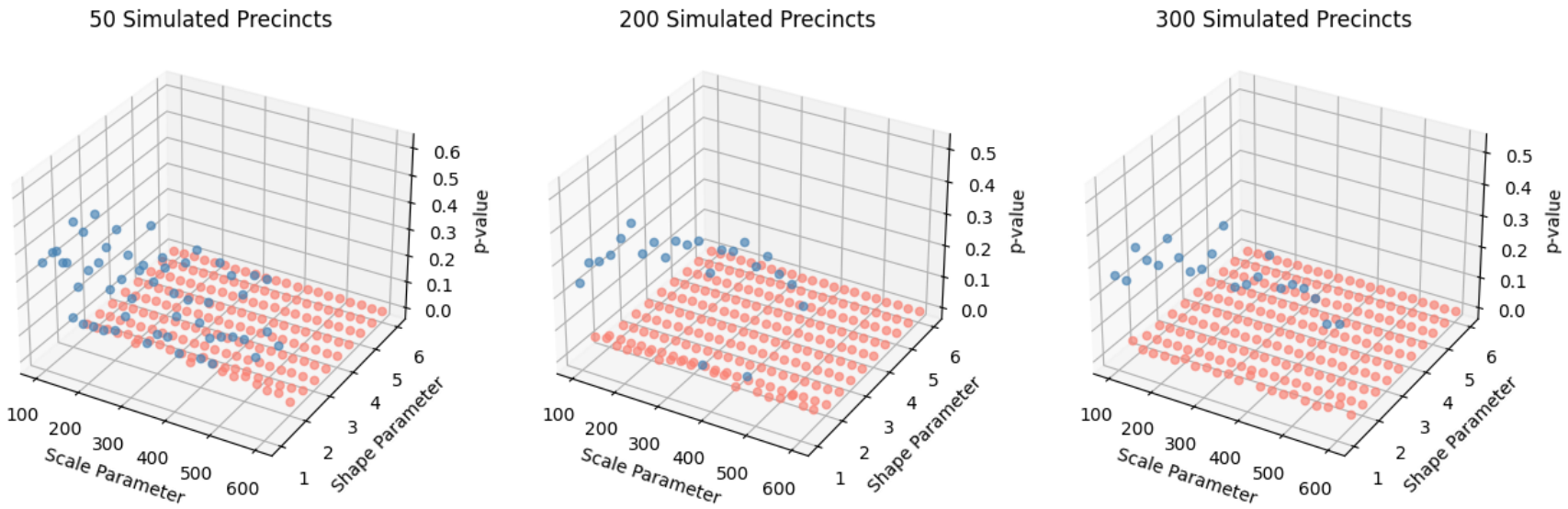


Figure: First Digit Base 10 Monte Carlo Analysis of Simulated Discrete Weibull Distributions with Various Choices of Parameterizations and Precinct Sizes Using Chi Squared Testing. The *p*-values above 0.05 (plotted in blue) indicate conformance to 1-BL 10.

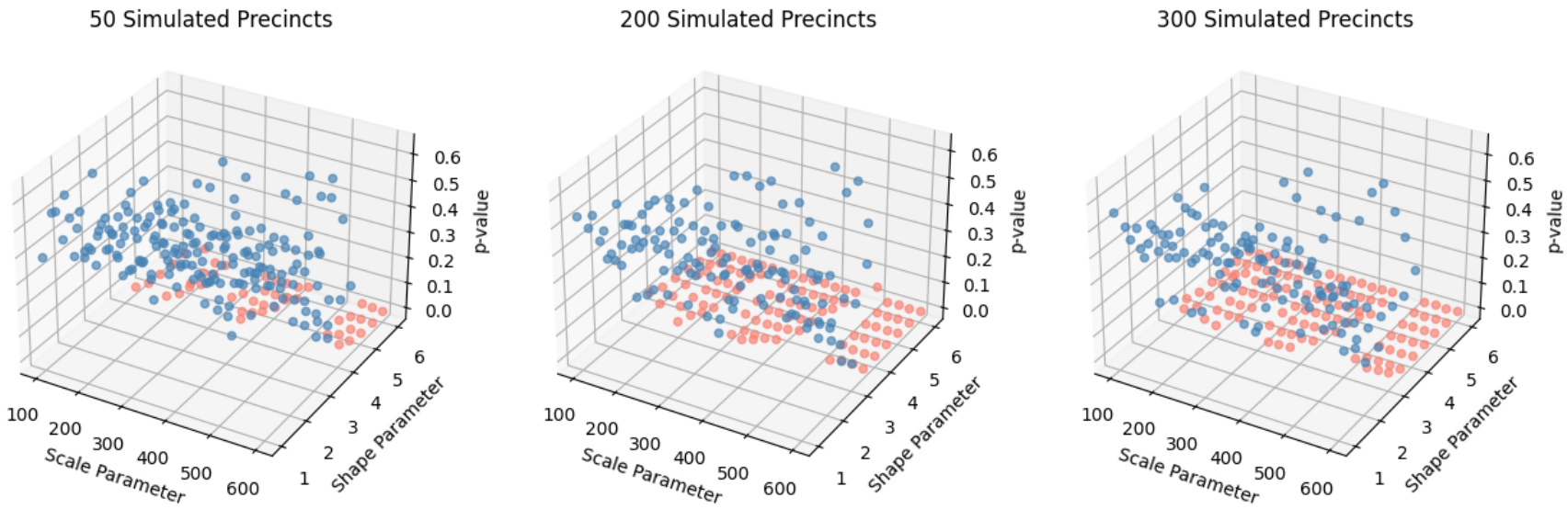


Figure: First Digit Base 3 Monte Carlo Analysis of Simulated Discrete Weibull Distributions with Various Choices of Parameterizations and Precinct Sizes Using Chi-Squared Testing. The *p*-values above 0.05 (plotted in blue) indicate conformance to 1-BL 3.

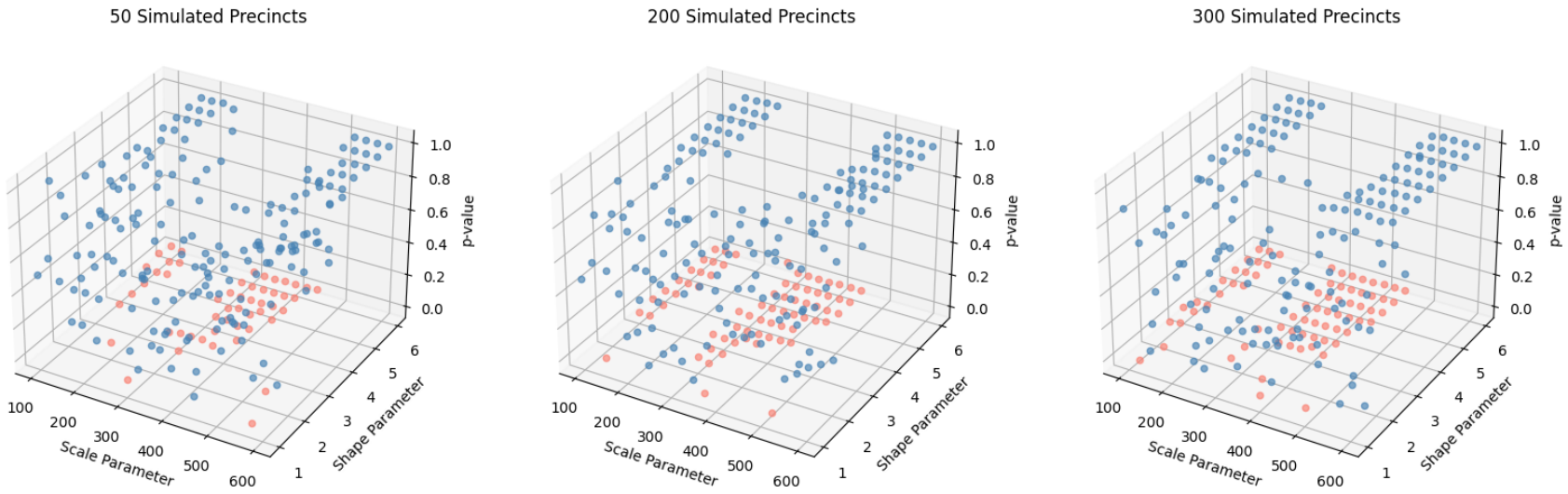


Figure: First Digit Base 3 Monte Carlo Analysis of Simulated Discrete Weibull Distributions with Various Choices of Parameterizations and Precinct Sizes Using Binomial Testing. The *p*-values above 0.05 (plotted in blue) indicate conformance to 1-BL 3.

State	County	Candidate	χ^2 Stat	*p*-value	Number of Precincts	adjusted *p*-value
CO	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
NC	No Anomalies Detected					
WI	No Anomalies Detected					
OH	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
VT	No Anomalies Detected					

Table: 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: Analysis of Whether Anomalous Counties from the Table in the Previous Slide Deviate from Benford's Law.

Avenues for Future Research

- Other base representations and the trade offs of the different base representations.
- Comparisons of Benford's base 10 second digit analysis to Benford's base 3 first digit analysis.
- Apply Benford's base 3 first digit analysis to other areas of research that include clustered data.

Conclusion

- Benford's base 10 first digit analysis is inappropriate for clustered data.
- Changing precinct level vote counts to base 3 renders Benford's first digit analysis as a useful tool to analyze elections.
- This proposal is generalizable to elections other than the 2004 presidential election.
- If data has been flagged it does not always mean that there was fraud committed.

Thank you for your time!