

BENFORD'S LAW APPLIED TO PRECINCT LEVEL ELECTION DATA

Ellen Friedman, University of St. Thomas, ellen.friedman@stthomas.edu
Rajkumar Kolakaluri, University of St. Thomas, kolakaluri.rajkumar@stthomas.edu
Manjeet Rege, University of St. Thomas, rege@stthomas.edu

ABSTRACT

This paper attempts to determine whether precinct level election data conforms with Benford's Law. In order to evaluate whether election data conform with Benford's Law, we constructed a two-part test. The first test assesses whether election data correlates with Benford's law. If the election data under study is found to correlate with Benford's law, we then subject it to the Kolmogorov-Smirnov test, which is used to evaluate more rigorous conformance with Benford's law and which aids in forensic analysis. We conclude that the frequency pattern of the first digits of the precinct level elections under study correlate strongly with the pattern predicted by Benford's Law of first digits. We also conclude, however, that the correlation is not strong enough for definitive forensic analysis.

Keywords: Benford's Law; Election results, Precinct, Kolmogorov-Smirnov

INTRODUCTION

In 1938, American physicist Frank Benford published "The Law of Anomalous Numbers" (Benford, 1938). Benford noticed a peculiarity in a much-used table of common logarithms: "The pages containing the logarithms of the low numbers 1 and 2 are apt to be more stained and frayed by use than those of the higher numbers 8 and 9." (Benford, 1938). He went on to describe his observation that in certain tabulations of numbers, the first significant digits occur with predictable logarithmic frequency:

$$F_a = \log_{10}(a+1/d)$$

where "a" is the value of the first digit. The phenomenon described in this paper is now known as "Benford's Law". Simply put, in sets of numbers that conform with Benford's Law, lower digits (1,2,3) occur with higher frequency as the first significant digit than higher digits (7,8,9). The bar chart below visually depicts the expected frequency of each digit 1 through 9.

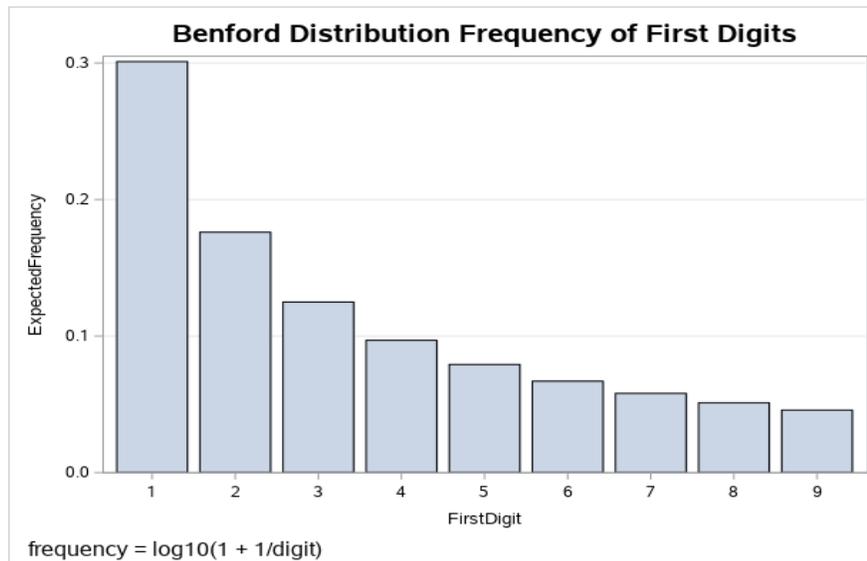


Figure 1. Benford distribution frequency of first digits

A wide variety of data sets have been found to conform with Benford’s law, including:

- Time interval between earthquakes
- Some population data
- Corporate expense data
- Some economic statistics (Kossofsky, 2015)

In human-controlled systems where Benford’s law is applicable, a lack of conformity with Benford’s law may be an indicator of fraud. Benford’s Law is increasingly used by forensic accountants and fraud examiners (Nigrini, 2012).

Nigrini cites many interesting examples where Benford’s Law was used to detect fraud. Here is how a first-two digits test found fraud in revenue numbers:

- “Inland Revenue in the United Kingdom found that there was a spike at 14 for the revenue numbers reported by small businesses. The investigation showed that many self-employed taxpayers were “managing” their sales numbers to just below £15,000. At that time the tax system in the U.K. allowed businesses with sales under \$15,000 to use the equivalent of a ‘Schedule C Easy’ when filing their tax returns.” (Nigrini, 2012)

Alex Kossofsky cites an example of Benford’s law used to indicate possible fraudulent home valuation for the purpose of reducing property tax payments. He presents a tax schedule wherein the tax rate jumps at \$100,000 and again at \$1,000,000. Fraudulently undervalued houses would follow a pattern: “A house worth \$105,000 would be falsified as being worth only \$98,000, say, while another house worth \$1,240,000 would be made to look as if it’s worth only \$960,000.” (Kossofsky, 2015). The first-two digits chart of home values in this municipality demonstrate a suspicious over-representation of digits 90 and above, and an equally suspicious under-representation of digits from 10 through 20. (Kossofsky, 2015)

PROBLEM STATEMENT

We were intrigued to learn that it is not clear whether Benford’s law applies to election data. In a 2011 paper, Deckert et.al state that “Benford’s Law is problematical at best as a forensic tool when applied to elections” (Deckert, 2011). Alex Kossofsky, however, states that “Election results closely conform to Benford’s Law.” And further: “A more careful examination of the underlying statistical process reveals that this is ultimately a Random Linear Combination from random logarithmic data. This confers election results an even more logarithmic aura than mere population data.” (Kossofsky, 2015)

This paper attempts to determine whether precinct level election data conforms with Benford’s Law.

In order to evaluate whether election data conform with Benford's Law, we constructed a two-part test. The first test assesses whether election data correlates with Benford's law. If the election data under study is found to correlate with Benford's law, we then subject it to the Kolmogorov-Smirnov test, which is used to evaluate more rigorous conformance with Benford's law (Nigrini, 2012). Thus, there are two tests, each with its own pair of hypotheses.

RESEARCH METHODOLOGY

Test 1: Simple Correlation

The research hypotheses to be tested are as follows:

Null hypothesis (H_0): Precinct level election data does not correlate with Benford's first-digit law. Under this hypothesis, the observed frequencies of the first significant digit of the votes in each precinct race demonstrate no correlation with the geometric distribution described by Benford's law. The correlation coefficient IS NOT significantly different from 0. There is no linear relationship between the observed first digit frequencies and the first-digit frequencies that would be expected by Benford's law.

Alternative Hypothesis (H_1): Precinct level election data DOES correlate with Benford's first-digit law. Under the alternative hypothesis, the first significant digit of the votes in each race demonstrate significant correlation with the geometric distribution described by Benford's law. The correlation coefficient IS significantly different from 0. There is a significant linear relationship between the observed first digit frequencies and the first-digit frequencies that would be expected by Benford's law.

Test 2: Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is a statistic used to assess conformity with Benford's Law. It is typically used against data that is known to normally conform well with Benford's Law as a means to find small deviations from Benford. The null hypothesis, therefore, is that data conforms with Benford's Law.

Null hypothesis (H_0): The Kolmogorov-Smirnov statistic produced by the data under study will be less than a critical value. This indicates very high conformance with Benford's Law.

Alternate hypothesis (H_1): The Kolmogorov-Smirnov statistic produced by the data under study will be greater than a critical value. This indicates deviations from Benford's Law.

Both tests will use a 95% confidence level.

METHODS

Data

For this study, we gathered election results from the election held in the United States on November 6, 2018. Data was sourced from the offices of the Secretaries of State of Arizona, Florida, Georgia, and Minnesota. We collected data at the precinct level. This is the finest-grained election data publicly available in the United States. Usually, election results as commonly reported have already been summarized to the county, state, or federal level. Since we will be extracting the first significant digit (fsd) of each vote, it is important to have voting data that has not been aggregated.

Once we extract the first significant digit (fsd) from the vote, it becomes a categorical variable. We can count the frequency of each different value, but we can no longer add those values together. Therefore, the point at which we extract the fsd is important. For this analysis, the fsd is extracted from precinct level data. If we extracted the fsd from county or state level races, any Benford conformance would be obscured.

Data Preparation and Cleansing

We used SAS® University Edition JupyterLab to write the data conversion routines. Each state delivers election results in one or more different formats. Our goal was to capture voting information for each race at the precinct level, strip the first digit, then bin each first digit according to its value in one of nine bins – one bin for each digit. We then aggregated the data into a standard format grouped by: State Code, Contest, and Party. The bins now contain the observed frequency count for each of the nine digits. We added the expected frequency for each bin, according to Benford’s law.

- We eliminated votes for any contests that were not state-wide.
- We eliminated any precincts with less than 10 votes, following the advice of Mark Nigrini (Nigrini, 2012).

Because Benford’s law is logarithmic in nature, we wanted to ensure that the data for each race reflected multiple orders of magnitude. We included an “OrderOfMagnitudeChange” statistic for each race. This gives us the runtime option to exclude races that did not capture multiple changes in magnitude. (For example: races with minor party candidates):

```
reduced_df["OrderOfMagnitudeChange"] =
{np.log10(reduced_df["VotesMax"].astype(int)) – np.log10(reduced_df["VotesMin"].astype(int))}
```

At the end of the conversion, we had a dataset consisting of 360 observations. Each set of 9 first significant digit observations is for a single race (state + contest + party). A partial list is below:

Table 1. Partial list of races in dataset

	State	Contest	Party	Number of Precincts with this first digit								
				1	2	3	4	5	6	7	8	9
1	AZ	Governor	DEM	7	1	3	0	1	1	0	2	0
2	AZ	Governor	GRN	5	5	0	1	0	0	0	3	1
3	AZ	Governor	NON	15	5	5	4	4	1	1	3	1
4	AZ	Governor	REP	3	4	1	1	1	0	3	1	1
5	AZ	U.S. Senator	DEM	7	1	2	1	1	0	2	0	1
6	AZ	U.S. Senator	GRN	6	2	3	1	0	1	0	1	1
7	AZ	U.S. Senator	NON	8	12	2	2	6	2	2	2	3
8	AZ	U.S. Senator	REP	5	2	2	0	1	4	1	0	0

Data Validation

Because this is election data, it could easily be validated by comparing it to the values reported by each State’s Secretary of State. For each state, we created a cross-check spreadsheet to validate that the converted data matched the election results as reported by each State.

RESULTS

Test 1: Simple Correlation

In order to determine whether precinct level data correlates with Benford’s first digit law, we used SAS® Studio to execute a canonical correlation analysis. We filtered the data at runtime to include only those races that underwent more than 2 orders of magnitude of change:

```
proc corr data=WORK.PRECINCT pearson nosimple noprob plots=none;
  where OrderOfMagnitudeChange > 2;
  var Frequency;
  with ExpectedFrequency;
```

Test 1: Simple Conclusion

The analysis delivered a canonical correlation value of 0.971781, with an F statistic of 3021.05 and a p value of <.0001. Therefore, the null hypothesis is rejected. The election data under study demonstrates significant correlation with Benford's Law of first digits.

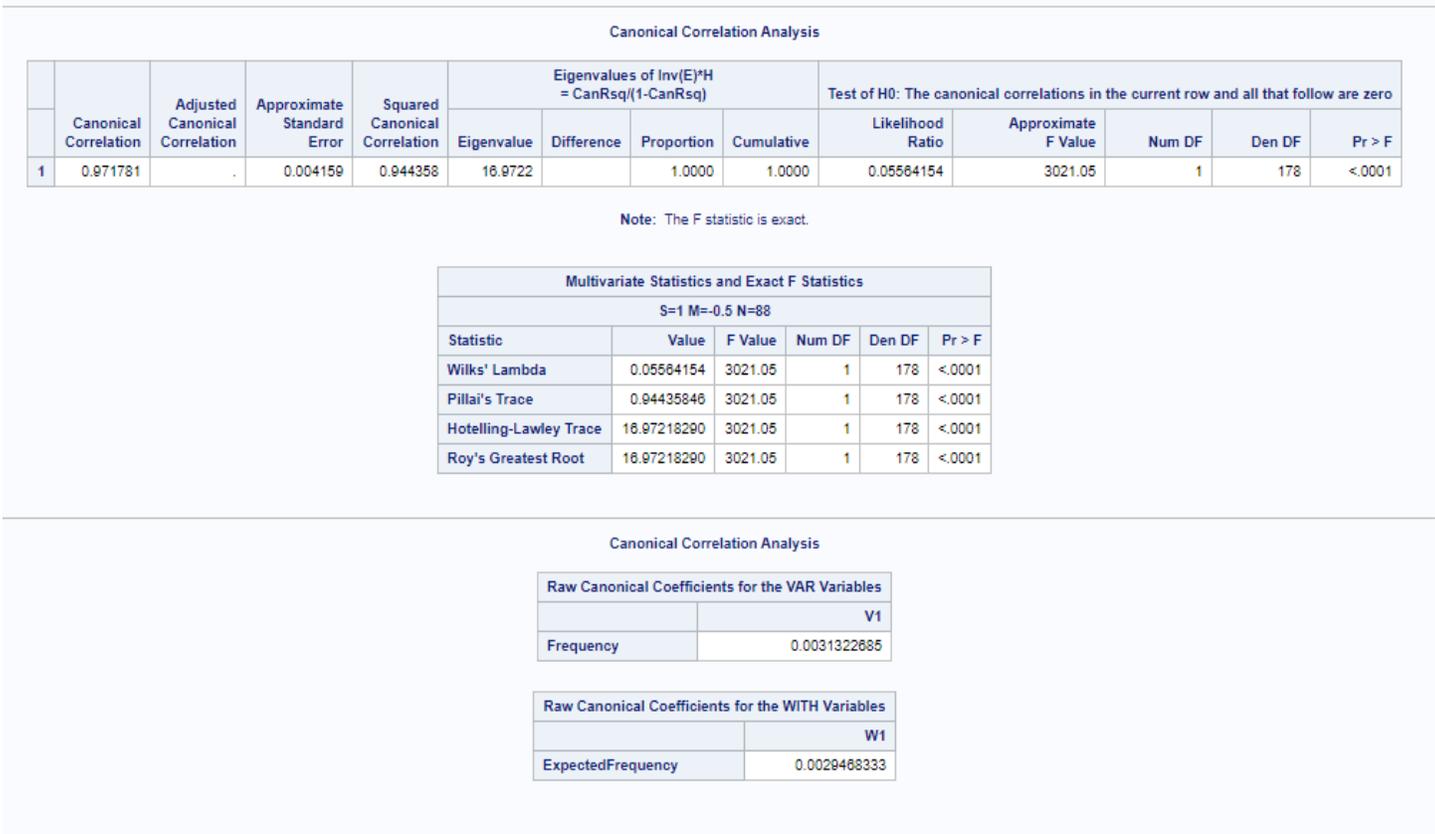


Figure 2. Canonical correlation analysis

The significant correlation between the observed and expected Benford frequencies is visually supported by the chart below. This chart was created using SAS® Studio bar and line chart utility. The observed frequencies are the grey bars. The expected frequencies are expressed by the red line. The correlation is visually striking.

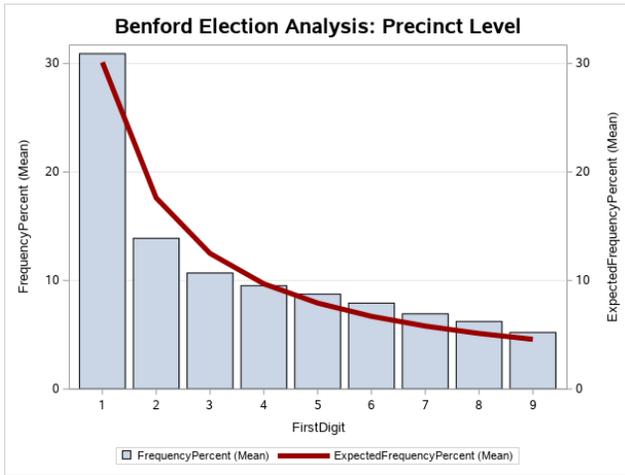


Figure 3. Observed vs. expected, all states

For comparison, we created the same bar chart for each of the four states. The visual implication – that Minnesota conforms a bit better than the aggregate, the other three states a bit worse – is confirmed by running a correlation analysis with each state’s election data.

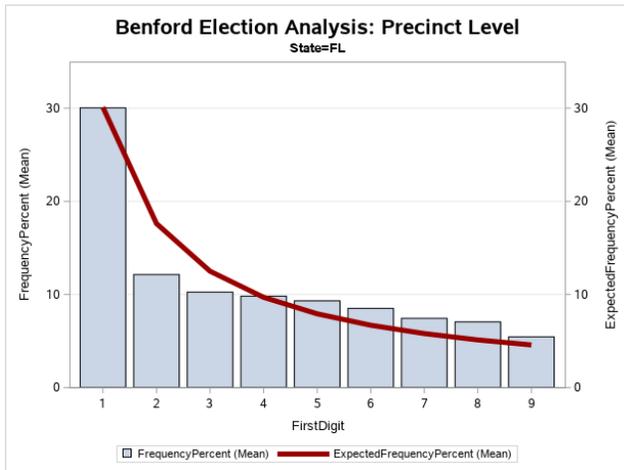


Figure 4. Observed vs. expected, Florida

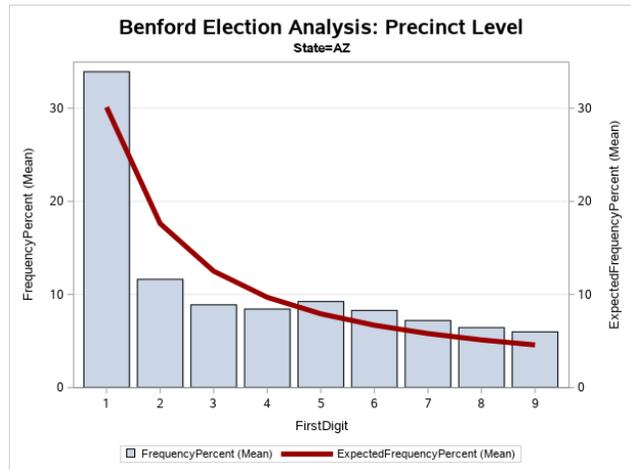


Figure 5. Observed vs. expected, Arizona

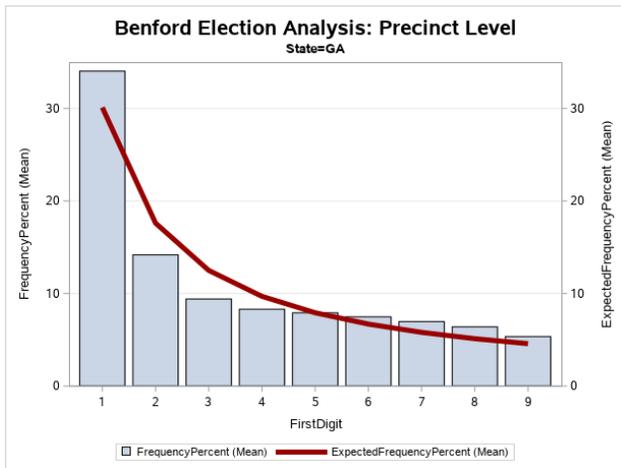


Figure 6. Observed vs. expected, Georgia

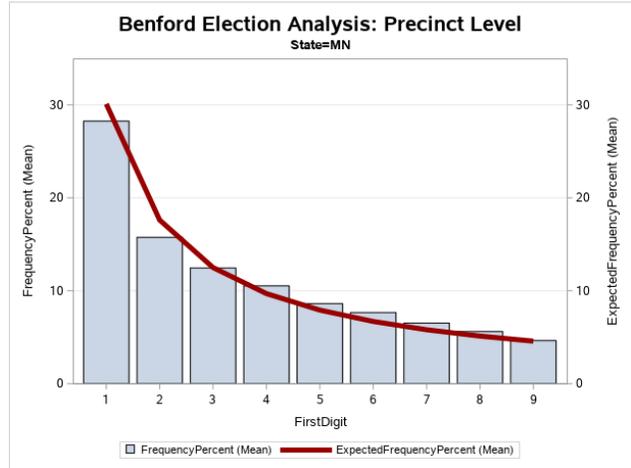


Figure 7. Observed vs. expected, Minnesota

Table 2. Correlations

State	Canonical Correlation
All	0.971781
AZ	0.926414
FL	0.950576
GA	0.966587
MN	0.985534

As a further confirmation on the results of test 1, we ran a correlation analysis on the observed fsd frequencies and compared them to what would be expected if all first digits had an equally likely chance of occurring. The correlation result of that test was just .50; the frequency pattern of first digits in precinct level elections is not random.

Test 2: Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) statistic can be used to detect small out-of-conformance deviations from Benford’s Law. As such, it is a useful tool for forensic analysis of data that normally conforms with Benford’s Law. For example, K-S can

be used to detect fraud in accounting data. It is important to note, however, that finding a deviation in your data via K-S is not air-tight evidence of fraud. It is merely a hint that you should take a deeper look at the data.

K-S is calculated as:

$$KS = \max\{|AD1 - ED1|, |(AD1+AD2) - (ED1+ED2)|, \dots, |(AD1+AD2+\dots+AD9) - (ED1+ED2+\dots+ED9)|\}$$

In the above formula, AD1 is the actual frequency of the leading digits, and ED1 is the theoretical frequency of the leading digits (Bushee, 2019). The resulting statistic is compared with the K-S critical value. At 95% confidence level, the K-S critical value is $1.36/\sqrt{N}$ where N = the number of records (Nigrini, 2012). If K-S is less than the critical value, there is not enough evidence to reject the null hypothesis. If K-S exceeds the critical value, the null hypothesis is rejected, meaning that the data shows deviations from Benford's Law. Since we have already determined that the precinct level election data in our study correlates well with Benford's Law, we will now use the K-S statistic to determine whether the data contain small deviations.

In order to determine the K-S statistic, we used SAS® Studio to run a paired t-test. We paired the observed frequency of each fsd with the expected frequency according to Benford's Law of first digits. We again filtered the dataset at runtime to include only those races that underwent more than 2 orders of magnitude of change. As a result, there were 180 observations in this study. At a confidence level of 95%, this means the K-S critical value is 0.101368415.

Test 2 Conclusion

The paired t-test delivered a K-S statistic of 0.151138. Since this value exceeds the critical value (0.101368415), we reject the null hypothesis. The data under study contains deviations from Benford's Law.

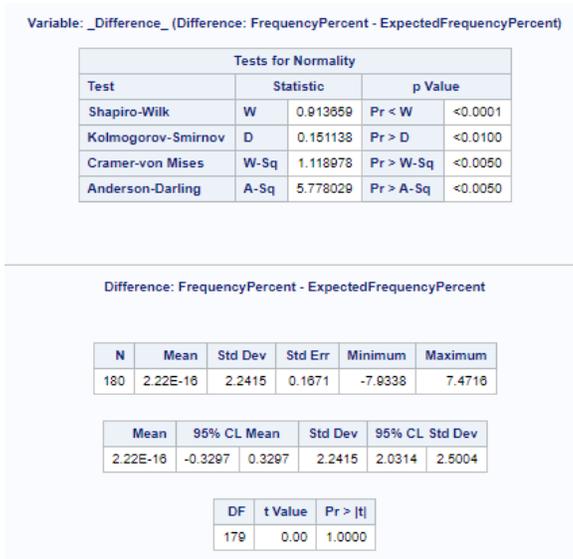


Figure 8. Results of paired t-test

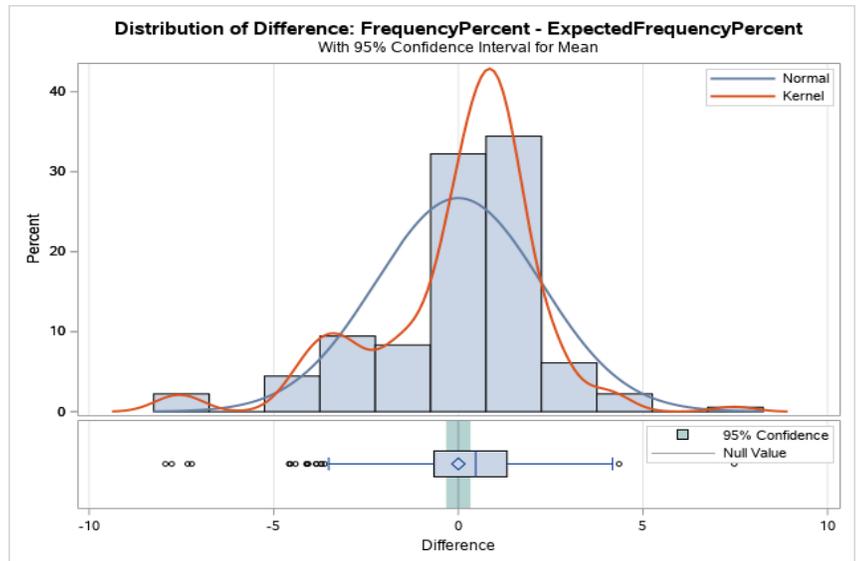


Figure 9. Results of paired t-test histogram and box plot

CONCLUSION

The frequency pattern of first digits in the precinct level elections under study is not random.

Under Test 1, we reject the null hypothesis (H_0) and accept the alternate hypothesis (H_1). We conclude that the frequency pattern of the first digits of the precinct level elections under study correlate strongly with the pattern predicted by Benford's Law of first digits. The visualizations (bar and line charts) reinforce this conclusion.

Under Test 2, we reject the null hypothesis (H_0) and accept the alternate hypothesis (H_1). We conclude that the correlation discovered under Test 1 is not strong enough for definitive forensic analysis. Forensic analysis requires more than just first-digit correlation. Forensic analysis requires strict adherence to Benford's Law, much more so than is provided by precinct level election data, regardless of the strong correlation. Benford's Law of first digits coupled with a K-S test against a critical value with a 95% confidence level would yield far too many false positives to be of use to forensic election examiners.

THEORETICAL IMPLICATIONS OF THIS STUDY

These findings suggest that first digits of precinct level election data strongly correlate with Benford's Law. They also suggest that if one is attempting to use Benford's Law to detect fraud in an election, the Kolmogorov-Smirnov test would not be an appropriate choice: a K-S at 95% confidence level would yield too many false positives. The critical value used in the K-S test relies on N – in this case, the number of precincts in each race. Nigrini explains that the K-S test is too intolerant of small natural deviations for large values of N (Nigrini, 2012); precinct level election data delivers large values of N .

SUGGESTIONS FOR FUTURE STUDIES

There are many opportunities for future studies. In particular, the Benford analysis could be extended to include first-two significant digits studies. It is possible that the K-S test would have the opposite result using first-two significant digits against the same data. Further, there are statistics other than K-S that can be used in forensic analysis of data that correlate with Benford, including chi-square and mean absolute deviation (Nigrini, 2012).

We would strongly encourage any future researchers to use non-aggregated voting data when doing this type of research. That means data should be at the precinct/candidate level when the significant digits are extracted from the vote value. We would also encourage future researchers to establish a baseline of acceptable deviation from Benford by examining multiple races at the precinct/candidate level. We were not able to find a study testing election data against Benford's Law that involved multiple races; nor were we able to find another study that used non-aggregated data as we have done here.

REFERENCES

- Benford, F. (1938). The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 551-572.
- Bushee, B. J. (2019, October 22). *Benford's Law 3.6*. Retrieved from Coursera:
<https://www.coursera.org/lecture/accounting-analytics/benford-s-law-3-6-oPSSY>
- Deckert, J. M. (2011). Benford's Law and the Detection of Election Fraud. *Political Analysis*, 245-268.
- Kossovsky, A. E. (2015). *Benford's Law : Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications*. Singapore: World Scientific Publishing Co. Pte. Ltd.
- Nigrini, M. J. (2012). *Benford's Law*. Hoboken, NJ: John Wiley & Sons.
- Division of Elections, Florida Department of State. (n.d.). *2018 General Election Results*. Retrieved from
<https://results.elections.myflorida.com/Index.asp?ElectionDate=11/6/2018>
- Georgia Secretary of State. (n.d.). *November 6, 2018 General Election Results*. Retrieved from
<https://results.enr.clarityelections.com/GA/91639/Web02-state.221451/#/>
- Office of the Minnesota Secretary of State. (n.d.). *2018 Precinct Results*. Retrieved from
<https://www.sos.state.mn.us/elections-voting/2018-general-election-results/2018-precinct-results-spreadsheet/>
- Secretary of State, State of Arizona. (2020). *Precinct Level Results By County – 2018 General Election*. Retrieved

from <https://azsos.gov/precinct-level-results-county-2018-general-election>

Secretary of State, State of Arizona. (n.d.). *File type descriptions*. Retrieved from https://azsos.gov/sites/default/files/2018_File_Type_Descriptions.xlsx

NIST Election Results Common Data Format Specification - Version 1.0 Retrieved from <https://www.nist.gov/itl/voting/interoperability/election-results-reporting-cdf>

The Federal Election Commission of the United States of America Retrieved from <https://www.fec.gov/>

J. Carlton Collins, C. (2017, April). Using Excel and Benford's Law to detect fraud. *Journal of Accountancy*.

Retrieved from <https://www.journalofaccountancy.com/issues/2017/apr/excel-and-benfords-law-to-detect-fraud.html>