Using R for Identification of Data Inconsistency in Electoral Models

Marius JULA
"Nicolae Titulescu" University of Bucharest

ABSTRACT

When using datasets for various analyses one should test the data for particular situations like existence of outliers or possible data errors. Outliers may indicate bad data and the results may be affected if these points are not identified and/or explained. Also, there are sensitive data, like electoral datasets, which are subject of fraud suspicion. Methods for identifying outliers and data errors are described in this paper, using R support and electoral data.

Keywords: *outlier, Z-score, Benford's law, R*

For creating reliable econometric models, one must rely on existing data. The altered data sets may create false signals and the conclusions based on these signals may not be accurate. Also, there is a known "data manipulation" in sensitive situations, like electoral data.

BENFORD'S LAW

As Peter Klimeka et all[1] stated, free and fair elections are the cornerstone of every democratic society. A central characteristic of elections being free and fair is that each citizen's vote counts equal. However, Joseph Stalin believed that "it's not the people who vote that count; it's the people who count the votes." Now the question raises: "How can it be distinguished whether an election outcome represents the will of the people or the will of the counters?"

One can see elections as large-scale social experiments. A country is segmented into a usually large number of electoral units. This is also the case for Romania. Here we have electoral circumscriptions for each county, districts in Bucharest and electorate from abroad. The above-mentioned authors affirmed that "each unit may represent a standardized experiment, where each citizen articulates his/her political preference through a ballot. Although elections are one of the central pillars of a fully functioning democratic process, relatively little is known about how

1. Peter Klimeka, Yuri Yegorovb, Rudolf Hanela, Stefan Thurner (2012), Statistical detection of systematic election irregularities, Proceedings of the National Academy of Sciences of the United States of America, no. 41, vol 109, 16469–16473

election fraud impacts and corrupts the results of these standardized experiments."

The specific literature acknowledge that there is an overabundance of ways of tampering with election outcomes (for instance, the redrawing of district boundaries known as gerrymandering or the barring of certain demographics from their right to vote, blocking access to voting locations). Some practices of manipulating voting results leave traces, which may be detected by statistical methods. Recently, Benford's law experienced a new start as a potential election fraud detection tool. In its original and naive formulation, Benford's law is the observation that, for many real world processes, the logarithm of the first significant digit is uniformly distributed. Deviations from this law may indicate that there are chances of data to be incorrect. For instance, suppose a significant number of reported vote counts in districts is completely made up and invented by someone preferring to pick numbers, which are multiples of 10. The digit 0 would then occur much more often as the last digit in the vote counts compared with uncorrupted numbers. Voting results from Russia[1], Germany[2], Argentina[3], and Nigeria[4] have been tested for the presence of election fraud using variations of this idea of digit-based analysis. There are also analysts who stipulate that the validity of Benford's law as a fraud detection method is subject to controversy. Peter Klimeka suggests that "the problem is that one needs to firmly establish a baseline of the expected distribution of digit occurrences for fair elections. Only then it can be asserted if actual numbers are over- or underrepresented and thus, suspicious. What is missing in this context is a theory that links specific fraud mechanisms to statistical anomalies[5]."

Walter Mebane[6] supports the idea of using Benford's law: "why should Benford's Law apply to vote count data?" He offers two mechanisms for why second digits of vote counts should follow a "Benford's Law-like distribution" which he refers to as the 2BL distribution. As stated above, Benford's Law does not apply to "simple random" data. Therefore, in order for Benford's Law to apply to vote count data vote count data cannot be generated simply randomly. Instead, due to the complexity inherent in the voting process, simple randomness should not be observed in voting outcomes. Thus, vote choice is not simply a "stochastic choice", but rather consists of a set of complex processes. Such processes are as follows. An individual voter first decides whether or not to vote and, secondly, who to vote for (or also which way to

1. Mebane WR, Kalinin K (2009) Comparative Election Fraud Detection. (The American Political Science Association, Toronto, ON, Canada).

2. Breunig C, Goerres A (2011) Searching for electoral irregularities in an established democracy: Applying Benford's Law tests to Bundestag elections in unified Germany. Elect Stud 30:534–545.

3. Cantu F, Saiegh SM (2011) Fraudulent democracy? An analysis of Argentina's infamous decade using supervised machine learning. Polit Anal 19:409–433.

4. Beber B, Scacco A (2012) What the numbers say: A digit-based test for election fraud. Polit Anal 20:211–234.

5. Deckert JD, Myagkov M, Ordeshook PC (2011) Benford's Law and the detection of election fraud. Polit Anal 19:245–268

6. Mebane, Walter R., Jr. 2006b. "Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections." Earlier version presented at the Election Fraud Conference, Salt Lake City, Utah, September 29-30, 2006.

vote on a particular referendum). Finally, a voter must actually cast his or her ballot which can be done in a variety of ways: "election day voting in person, early voting, provisional ballots or mail-in ballots; on paper, with machine assistance or using some combination". In addition, there is always the potential for mistakes:

When all is said and done, most voters will look at each option on the ballot and have firm intentions either to select that option or not to select that option. Then for whatever reason—momentary confusion, bad eyesight, defective voting technology—a small proportion of those intended votes will not be cast or recorded correctly. A small proportion will be "mistakes" (Mebane).

According to Mebane, "the kind of complexity that can produce counts with digits that follow Benford's Law refers to processes that are statistical mixtures (e.g., Janvresse and de la Rue 2004), which means that random portions of the data come from different statistical distributions" (Mebane).

Mebane uses simulations to show that when manipulations occur to 2BL distributed vote counts, this "will produce a significantly large value" of the test statistics. He shows that the test statistic is sensitive to departures from the 2BL distribution under a variety of scenarios: (1) when electoral manipulation occurs in a precinct for an already strong candidate; (2) when vote counts are manipulated in a close election (tie is expected according to the vote-generating process); and (3) when votes are manipulated in a precinct for a weak candidate. In addition, he shows that even small manipulations will produce significance (i.e. test statistic is sensitive to small manipulations). In other words, a massive amount of fraud does not have to occur for this test to detect fraud. However, "if the amount of manipulation is sufficiently small, the 2BL test will not signal that manipulation has occurred".

Why using Bendfor's Law? In 1972, Hal Varian suggested that the law could be used to detect possible fraud in lists of socio-economic data submitted in support of public planning decisions. Based on the plausible assumption that people who make up figures tend to distribute their digits fairly uniformly, a simple comparison of first-digit frequency distribution from the data with the expected distribution according to Benford's law ought to show up any anomalous results. Benford's law is used extensively in United States in legal status issues, election data, macroeconomic reported data and other scientific fraud detection algorithms.

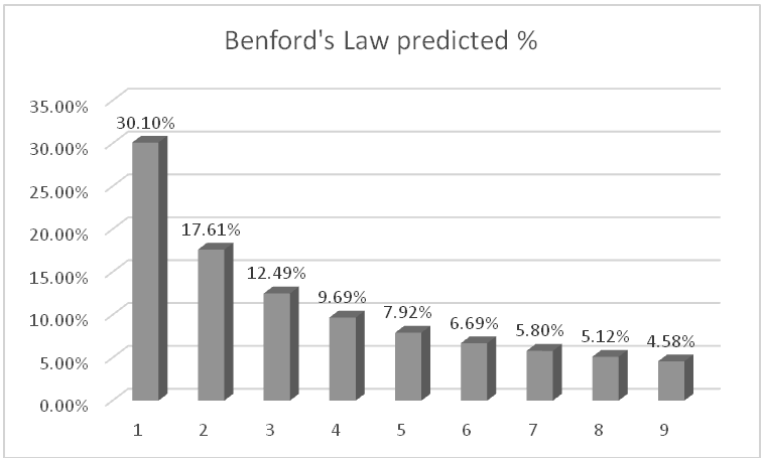
American astronomer Simon Newcomb observed that in logarithm tables (used at that time to perform calculations) the earlier pages (which contained numbers that started with 1) were much more worn than the other pages

Benford's law, based on Newcomb's studies, stipulates that the distribution frequency of digits in data sources would satisfy Benford's law if the leading digit d {1..9} occurs with probability:

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}(1+\frac{1}{d}) \quad (1)$$

Benford's distribution – author's calculations

Figure 1

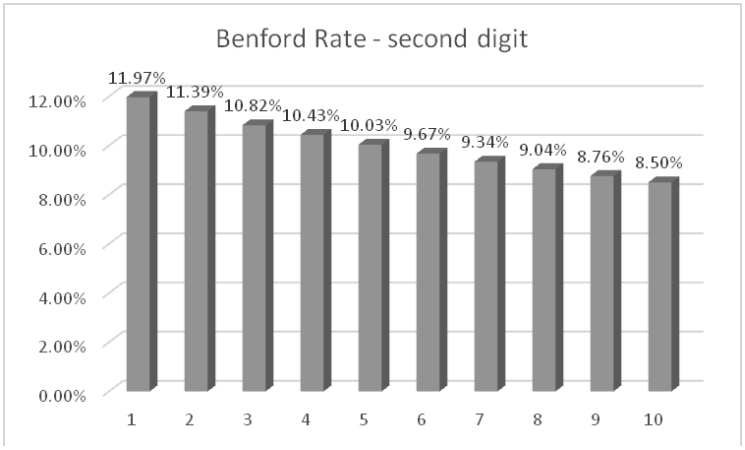


The law is also valid for other bases besides decimal. Also, this law was extended to digits beyond the first. The general formula for probability that d {0..9} appears as the *n*-th (n>1) digit is:

$$\sum_{k=10^{n-2}}^{10^{n-1}-1} \log_{10}\left(1+\frac{1}{10k+d}\right) \qquad (2)$$

Benford's law – second digit distribution - – author's calculations

Figure 2



ANALYZING ELECTORAL DATA – GENERAL ELECTIONS FROM NOV.2014 – ROMANIA

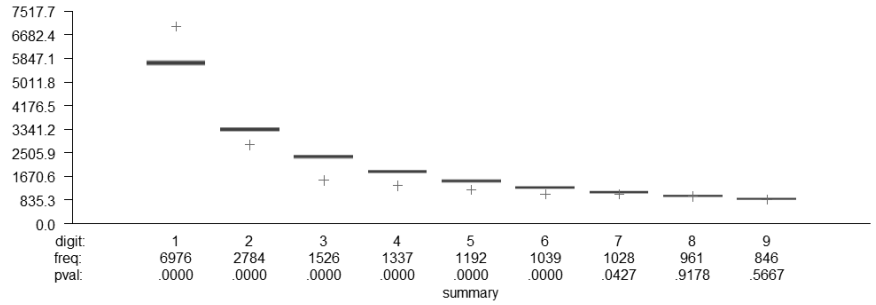
The used datasets are from the Central Electoral Bureau and were recorded for the second round, November 2014 in the general elections.

Our analysis tries to identify distribution of first digit in Total Number of Null Votes. It is a new approach that we implement, starting from the classical assumptions. We us R software, Package BenfordTests, Function signifd.analysis.

We test the hypothesis using Pearson's Chi-squared Goodness-of-Fit Test for Benford's Law - chisq.benftest (see Annex for more tests):

First digit distribution vs. Benford distribution, author's calculation

Figure 3



As we can observe in the above graphic, the distribution seems to obey the Benford's law. But, the Chi Square test is not passed (chisq = 912.0437, p-value < 2.2e-16).

Next, we search for outliers in recoded data. Outliers may suggest erroneous data or, in some cases, extreme values can be random or may indicate scientific interesting situation.

If observed in electoral data, an outlier may not necessarily suggest an error in the data set, but a special situation - for example, above average votes on special lists.

The presence of atypical points may decrease the quality of estimators (e.g., testing normality of error distribution may be affected by the presence of these points).

Data analysis assumes that in some polling stations is a significant difference between the regular number of voters and total votes. The difference is reflected in the votes on additional lists.

We use as variable the rate of votes on special lists and the total number of votes in that particular circumscription.

The test used is Z-modified test, based on MAD:

- package *outliers*
- function *scores(electoral_data, type="mad")*
- We select only values over 3.5 as possible outliers.

Result: over 760 possible outliers from 18554 recorded values.
Top five possible outliers are (R output):

```
> outlier[5169,]
[1] 15.28941
> outlier[14056,]
[1] 15.22692
> outlier[11145,]
[1] 13.99042
> outlier[13786,]
[1] 13.67635
> outlier[15319,]
[1] 12.67463
```

We can identify the particular polling station using the index of these outliers. For example, the highest score is obtained in the record 5169, which represent a polling station near a university campus in Cluj. It can be explained by the fact that the vast majority of votes are from the students not resident in Cluj.

But sometimes, the presence of such extreme points may indicate a possible electoral fraud (electoral tourism).

We are not trying to find a particular fraud situation, but for our further analyses we need accurate data. The econometric tests may be influenced by the presence of these outliers, especially normal distribution based tests.

CONCLUSIONS

Using Benford's law to discover data manipulation in the final results of the general elections from November, 2014, we can conclude that the data are largely validated. No important signs of abnormal distribution were detected.

As a further analysis, we recommend the analysis of the data using more fraud detection methods. This method is not exhaustive, as some economists suggest. This is not a perfect validation tool, is more like a test. If the data is in a category which supposed to obey Benford's law (like election data) and it fails, there is a signal of possible fraud. If the test is passed, doesn't mean the data is automatically validated, but more tests are always recommended to increase the confidence level.

The test for outliers suggest more than 760 possible extreme points. These do not necessarily represent fraud, but one should investigate the cause for their presence because, mainly, the econometric tests may be influenced by the presence of these outliers, especially normal distribution based tests.

References

1. Peter Klimek, Yuri Yegorov, Rudolf Hanel, Stefan Thurner (2012), Statistical detection of systematic election irregularities, Proceedings of the National Academy of Sciences of the United States of America, no. 41, vol 109, 16469–16473
2. Mebane WR, Kalinin K (2009) Comparative Election Fraud Detection. (The American Political Science Association, Toronto, ON, Canada).

3. Breunig C, Goerres A (2011) Searching for electoral irregularities in an established democracy: Applying Benford's Law tests to Bundestag elections in unified Germany. Elect Stud 30:534–545.
4. Cantu F, Saiegh SM (2011) Fraudulent democracy? An analysis of Argentina's infamous decade using supervised machine learning. Polit Anal 19:409–433.
5. Beber B, Scacco A (2012) What the numbers say: A digit-based test for election fraud. Polit Anal 20:211–234.
6. Deckert JD, Myagkov M, Ordeshook PC (2011) Benford's Law and the detection of election fraud. Polit Anal 19:245–268
7. Mebane, Walter R., Jr. 2006b. "Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections." Earlier version presented at the Election Fraud Conference, Salt Lake City, Utah, September 29-30, 2006.
8. Benford, F. (1938) The Law of Anomalous Numbers. Proceedings of the American Philosophical Society. 78, 551–572
9. Dixon, W.J. (1950). Analysis of extreme values. Ann. Math. Stat. 21, 4, 488-506.
10. Schiffler, R.E (1998). Maximum Z scores and outliers. Am. Stat. 42, 1, 79-80.
11. Iglewicz B., Hoaglin D., "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor, 1993.

Annexes

1. For Benford tests we used data from http://www.bec2014.ro/wp-content/uploads/2014/11/SIAP2014_STAT_Statistica-la-nivel-de-sectie-de-votare1.xlsx

We used the column "Numărul total al voturilor nule" – *total number of null votes*

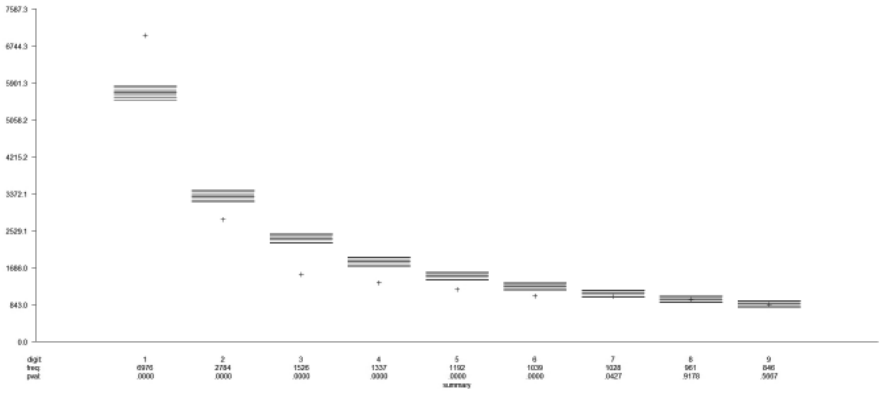
The R results for testing Benford's law are:

```
> #Pearson's Chi-squared Goodness-of-Fit Test for Benford's Law
> chisq.benftest(dataNum,pvalmethod ="asymptotic",digits=1)
         Chi-Square Test for Benford Distribution
data:  dataNum
chisq = 912.0437, p-value < 2.2e-16
>
> #Euclidean Distance Test for Benford's Law
> edist.benftest(dataNum, digits = 1, pvalmethod = "simulate", pvalsims
= 10000)
         Euclidean Distance Test for Benford Distribution
data:  dataNum
d_star = 12.727, p-value < 2.2e-16

>
> #Joenssen's JP-square Test for Benford's Law
> jpsq.benftest(dataNum, digits = 1, pvalmethod = "simulate", pvalsims
= 10000)
         JP-Square Correlation Statistic Test for Benford Distribution
data:  dataNum
J_stat_squ = 0.9383, p-value < 2.2e-16
```

First digit distribution vs. Benford distribution, author's calculation

Figure 4



2. For testing existence of outliers, we computed a new column, namely the ratio of votes on special list in the total number of the votes.

Top 6 outliers were:

```
head(order(abs(scores(date2, type="mad")),decreasing=TRUE))
[1]  5169 14056 11145 13786 15319 18513
> outlier<-scores(date2, type="mad")
> outlier[5169,]
[1] 15.28941
> outlier[14056,]
[1] 15.22692
> outlier[11145,]
[1] 13.99042
> outlier[13786,]
[1] 13.67635
> outlier[15319,]
[1] 12.67463
> head(order(abs(scores(date2, type="mad")),decreasing=TRUE))
[1]  5169 14056 11145 13786 15319 18513
```