Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) DOI: 10.7176/MTM
Vol.9, No.9, 2019

Anomalies Detection Using the Benford's Law: Application to the Kenyan Presidential Elections of 2017

Adolphus Wagala,
Department of Physical Sciences,
Faculty of Science, Engineering & Technology,
Chuka University, Kenya.

Abstract

In the modern times, the populace in most African countries are left wondering whether the declared election winner actually got the most votes. The validity of the declared election results in most cases remain questionable. In order to determine the validity of the declared results, an empirical statistical methodology could be used to give some hint and or evidence of anomalies in the declared election count data. This paper therefore considers a statistical method based on the pattern of digits in vote counts known as 2 digit Benfords Law (2BL) that is useful for detecting fraud or other anomalies. The 2BL methodology and other extensions are applied to detect the possible anomalies and fraud in the 2017 Kenyan presidential elections results data. The analysis show that the data for the top two presidential candidates: Uhuru Kenyatta and Raila Odinga do not follow the 2BL distribution. The digits are significantly different at 5% significance level when tested using the chi-square and the Euclidean tests. The mean absolute deviation (M.A.D) also confirms the non-conformity of the data to the 2BL distributions test. Further tests namely,the second order test, the summation test and the duplication test are utilized in order to detected any possible anomalies and fraud that could be present. All the three additional tests confirm the presence of fraud and anomalies in the data. These are red flags on the credibility of the presidential election results data published by the

Independent Electoral and Boundaries Commission (IEBC).

Keywords: Anomalies, Benford's Law, Kenyan Presidential Elections 2017

1 Introduction

Kenya is republic in the East Africa that is governed by a democratically elected president. As a democratic society, all the votes of each citizen should count. This would only result from a free and fair elections. However, the Kenyan presidential election of 2017 raised a lot of inflamed discussions following the declaration of Uhuru Kenyatta as the winner against his closest opponent Raila Odinga. This has raised the suspicion whether "it is the people who vote that count; or it is the people who count the votes?" as had stipulated by Joseph Stalin. The biggest challenge is therefore to determine the election outcome of 2017 really represented the will of the people or the will of the vote counters that matters.

Some authors like [1] consider an elections as large-scale social experiments in which a segmented into large number of electoral units. Each of these units represents a standardized experiment, where each citizen articulates his/her political preference through a vote. Just like any other experiments are prone to tampering, fraud and errors, the election results can similarly be interfered with and thereby generating the data (election) results that are erroneous. Some of the data manipulation practices usually leave traces that can be detected by use of statistical methodologies such as the Benford's Law [2, 3, 4]. The voting process is not a simple random process but a complex one. It involves the voter's decision to vote or not vote. Then the voter has to decide who to vote for and finally choose the ballot box to cast the vote. There are also errors inherent in the voting process such as marking the ballot paper twice, casting the ballot papers into a wrong ballot box among others. Such a process is really not a pure stochastic process. Given such kind of complexity, the resulting vote counts can produce digits that follow Benford's Law and can be referred to as processes that are statistical mixtures. This means that random portions of the data may come from different statistical distributions [4]. Therefore, when vote counts are manipulated in a close election, then the resulting data will not conform to the BL. Variations of the BL methodology has been used for testing the the

presence of election fraud in many countries including Argentina, Russia and Nigeria among others [5, 6, 7]

2 Benford's Law (BL)

The BL deals with the statistical distribution of *significant* (decimal) digits or, equivalently, *significands* viz. fraction parts in floating point arithmetic. To formally introduce the BL, it is natural to begin by defining the *significant* and the *significand*. Any positive digit d, can be written as $S(d) \times 10^k$, where $S(d) \in [0, 10)$ is known as the *significand* while k is an integer known as the *exponent*. The significand is also referred to as , the *leading digit* or the *first digit* by some people. This may be confusing, for example, 25678.854 may be written as 2.5678854×10^4. In this case, the significand is 25678854, the leading digit is 2 while the exponent is 4.

Following [3], for every non-zero real number x, the first significant decimal digit $D_1(x)$ is the unique integer $j \in \{1, 2, \ldots, 9\}$ satisfying the condition $10^k j \leq |x| 10^k (j + 1)$ for some integer k then $D_1(x)$.

Intuitively, when we consider the first digits of any number we are bound to believe that the probability of any of the first digits is uniformly distributed. That is, for the set of numbers $\{1, 2, 3, 4, 5, 6, 7, 9\}$, the $P(D_1 = d_1) = 1/9, d_1 = 1, 2, \ldots, 9$ where D_1 is the first digit. However, Benford's law shows us that this is not true. In fact, the smaller digits will have larger probabilities.

A data set is said to satisfy the Benford's Law (BL) for the leading digit D_1 if the probability of observing a first digit d_1 is approximately

$$P(D_1 = d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right) \tag{1}$$

where $d_1 = 1, 2, \ldots, 9$.

Further more, the second digit D_2 is said to follow the BL if the probability of observing d_2

as the second digit is given by

$$P(D_2 = d_2) = \sum_{d_1=1}^{9} \log_{10}\left(1 + \frac{1}{d_1 d_2}\right) \tag{2}$$

where $d_2 = 0, 1, 2, \ldots, 9$. From equations 1 and 2, it is easy to see that the first-two digits $D_1 D_2$ follow the BL if the probability of observing the first-two digits $d_1 d_2$ is given by

$$P(D_1 D_2 = d_1 d_2) = \log_{10}\left(1 + \frac{1}{d_1 d_2}\right) \tag{3}$$

where $d_1 d_2 = 10, 11, 12, \ldots, 99$.

When the logarithms of the numbers are taken, then the Benford's Law is "perfectly" followed. This is due to fact that the mantissas of the logs of the numbers are expected to be uniformly (evenly) distributed. In this case, the mantissa refers to the non-negative decimal part of the logarithm. The mantissa of the log is usually related to the first digit of a number for instance a number with a log that has a mantissa less than 0.3010299956 has a first digit 1. For more intuitive discussions see [8].

In general the BL can be looked at as the joint distribution of all decimal digits. In other words the probability for the first, first-two, first-three, first-four, and first-anything digits can be given as

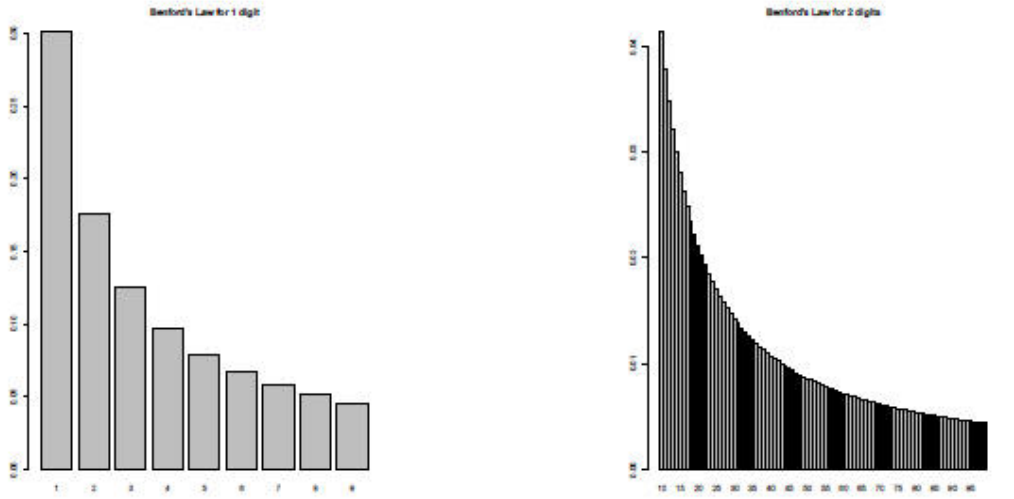
$$P(D_1 = d_1, \ldots, D_k = d_k) = \log_{10}\left[1 + \frac{1}{\sum_{j=1}^{k} 10^{k-j} \times d_j}\right] \tag{4}$$

where k is a positive integer, $d_1 \in \{1, 2, \ldots, 9\}$ and for $j \geq 2$ then $d_j \in \{0, 1, 2, \ldots, 9\}$. As a matter of fact, it is worth noting that for the general form of BL, the significant digits are dependent, and not independent as one might expect, see [3, 8].

From numerous experiments and tests, it has been shown that random numbers in most occasions tend to conform to the BL. Nonconformity to BL could mean that the data was not expected or supposed to conform in the first place. A weak fit to Benford's Law can be used as a red flag that there is a high risk that the data contains abnormal duplications and anomalies.

The BL implies that the data should have more small numbers than larger numbers, which implies that the data should not be too clustered around its mean value [2, 3, 8].

The BL can be visualized by plotting the distribution of the sequence of numbers to detect their conformity. As an example, consider the Figures 1a and 1b.



(a) Benford's Law for 1 digit numbers *(b) Benford's Law for 2 digit numbers*

Figure 1: ***An illustration of the Benford's Law for 1 and 2 digits respectively.***

3 Methodology

The following tests based on [8] were run on the data in a bid to detect anomalies. *The summation test* which looks for excessively large numbers in a data field. It identifies numbers that are large compared to the norm for that data thereby adding a new twist to the usual first-two digits test. This test is based on the fact that the sums of all the numbers that follow the BL with first-two digits being $\{10, 11, 12, \ldots, 99\}$ should be equal. Another test considered is the *second-order test* which looks at the patterns in data on the digits of the differences between sorted from smallest to largest (ordered) numbers. The digit patterns of the differences are expected to closely approximate the digit frequencies of Benford's Law. The second- order test gives few, if any, false positives in that if the results are not as expected (close to Benford), the

data does indeed have some characteristic that is rare and unusual, abnormal, or irregular. The method is based on tests of the distribution of the digits in reported vote counts, so all that is needed are the vote counts themselves [4].

The most used goodness of fit statistic in many applications is the chi-square test. It is used in this case to compare whether a set of the election actual results are conforms with the expected results. The null hypothesis to be tested is that there is no significant differences between the real digits and the expected ones in relation to the Benfords Law. The Pearsons Chi-square goodness-of-fit test for Benford's Law given by equation 5.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i^o - f_i^e)^2}{f_i^e} \tag{5}$$

where k is the number of bins, f_i^o denotes the observed frequency of digits i, and f_i^e denotes the expected frequency of digits i.

Euclidean Distance Test for Benfords Law is given by

$$\chi^2 = \sqrt{n}\sqrt{\sum_{i=10^{k-1}}^{10^k-1} (f_i^o - f_i^e)^2} \tag{6}$$

where f_i^o denotes the observed frequency of digits i, and f_i^e denotes the expected frequency of digits i.

The mean absolute deviation (M.A.D) is also useful in testing for the differences in the leading digits used in BL and is given by Equation 7

$$M.A.D = \frac{\sum_{i=1}^{k} |AP - EP|}{k} \tag{7}$$

where AP is the actual proportion, EP is the expected proportion while k is the number of bins. In the M.A.D formular, the numerator measures the absolute difference between the actual proportion and the expected proportion for each digit. It is worth noting that the higher the M.A.D, the larger the average difference between the actual and expected proportions. When the M.A.D values are greater than 0.015 then there is an indication of non-conformity

to the Benford's Law [8].

4 Results and Discussions

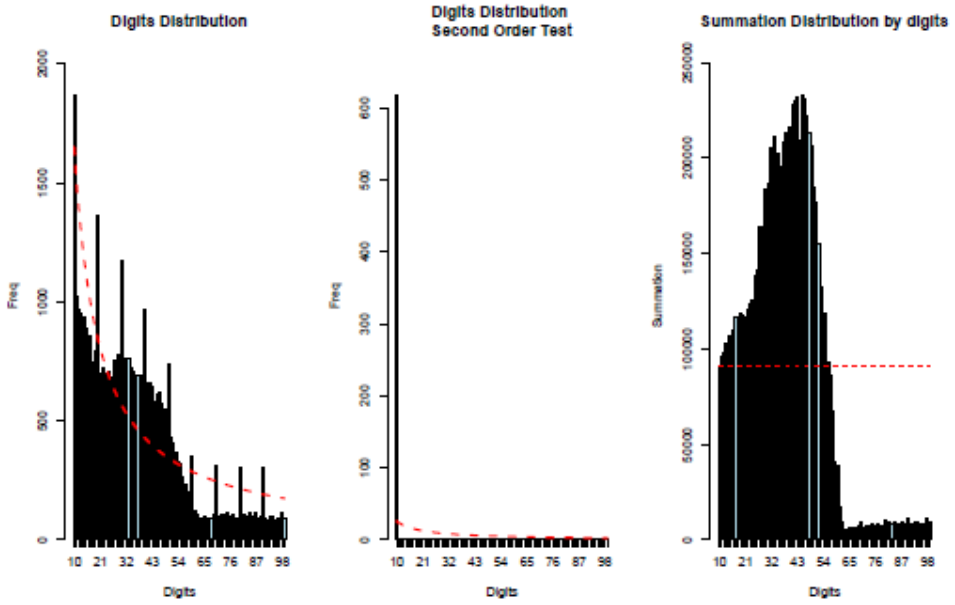
Data

The most challenging part of this study was the acquisition of data from the Independent Electoral and Boundaries Commission (IEBC). The official first round presidential results for the 2017 presidential elections are no longer available at the IEBC website. The data used in this study was provided by Professor Walter Webane Jr of the University of Michigan who had used it in his working paper [9].

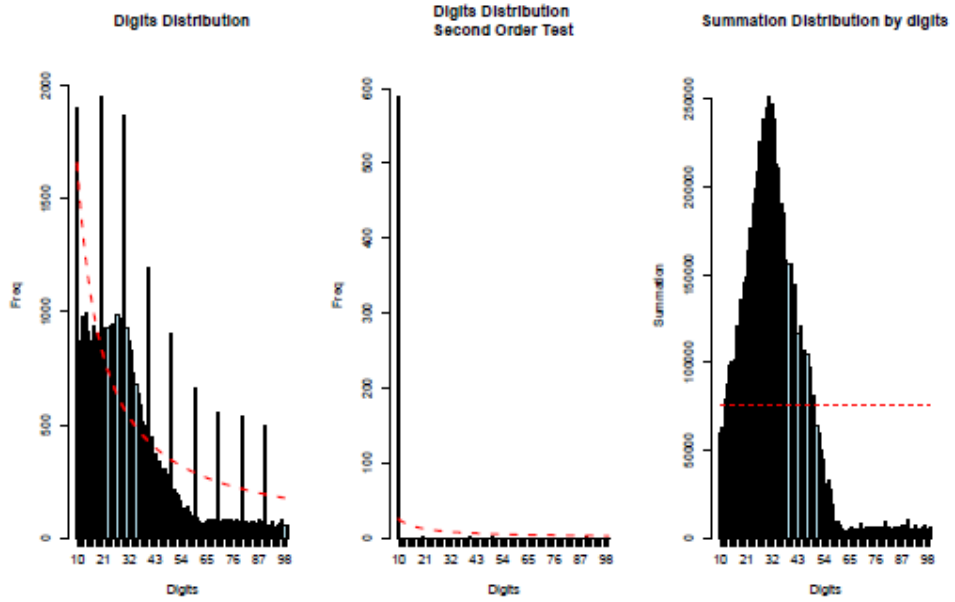
The data had been scrapped from the IEBC website https://public.rts.iebc.or.ke/enr/index.html#/Kenya_Elections_Presidential/1 by [9]. The originally downloaded, included 40,830 polling station observations. The eligible voter count data had 40,884 polling stations. [9] merged the two files and removed the unmatched and missing observations, including the ones that had vote counts of zero for all candidates and remained with 40,818 polling station observations. See [9] for more details.

BL distributions for the two top presidential candidates

In this section, we present the various Benford's distribution figures and tests for the two top presidential candidates namely Uhuru Kenyatta and Raila Odinga. Three tests related to the BL are presented. These include the first digit test, second order test and the summation test championed by [8].



(a) Uhuru Kenyatta



(b) Raila Odinga

Figure 2: ***Benford's Law (BL) distribution for top two presidential candidates' votes from the polling stations.***

Looking at sub Figures 2a and 2b, the two digits distribution for both Uhuru Kenyatta and Raila Ondinga respectively seem to have sevral abnormal high spikes thereby not conforming to the Benford's Law.

We now look at the second order test for the two top presidential candidates in the sub Figures 2a and 2b. The second order test should approximately follow the BL distribution. However, this is not the case in the data set considered. This is a red flag that some serious issue or error might exist in the data.

The third plot in the sub figures for each candidate presents the summation test. This test is based on the fact that the sums of all the numbers in a Benford Set with first-two digits $10, 11, 12, \ldots, 99$ should be equal. The sums for the various digits are expected to be equal, however in this study, the spikes tell us that there are abnormally large numbers relative to the rest of the data.

A visual inspection with regards to the BL distribution is not enough. Statistical inference is carried out using the chi-square test, the Euclidean distribution test and the M.A.D. The results are presented in Table 1.

Table 1: Conformity to the 2BL tests for the leading digit analysis.

Statistic	Uhuru Kenyatta	Raila Odinga
$\chi^2_{(89)}$	7681.7, $p-value$ =0.00	6080.7, $p-value$ = 0.000
Euclidean dist	23.821, $p-value$ =0.00	26.988,$p-value$ =0.00
M.A.D	0.004068691	0.005230433
Distortion Factor	-8.056533	-18.18931

Both the chi-square and the Euclidean distance tests show that the differences in the digits are statistically significant at 5% level. The M.A.D values indicated the non-conformity to the Benfords Law for the two candidates. The distortion factor model uses the digit patterns to signal whether the data appears to be over- or understated and the extent of the distortion [8]. In this data, there was some distortion for both candidates.

Further analysis looked at the absolute deviations for the top 10 two digits for each candidate.The results are presented in Table 2.

Table 2: Distribution of the digits by decreasing order of the absolute differences for Uhuru Kenyatta and Raila Odinga.

Uhuru Kenyatta		Raila Odinga	
digits	absolute.diff	digits	absolute.diff
30	604.5470	30	1302.0691
40	539.6754	20	1107.9619
20	515.6488	40	767.8095
11	481.7986	11	640.3738
12	417.2004	50	555.8039
50	393.3004	12	416.2490
13	330.4953	31	380.1653
14	258.9745	60	371.6987
42	248.7867	28	358.0661
33	246.3716	27	351.8818

The numbers with 30 had the largest absolute difference for both candidates while the lowest was 33 for Uhuru Kenyatta and 27 for Raila Odinga.

Given that the first-order and the summation tests deviated from the BL, the number duplication test is employed to identify the specific numbers that caused the spikes. The results are shown in a self explanatory Table 3.

Table 3: Distribution of the observations of the 10 values with most duplicates.

Uhuru		Raila	
1:	15	1:	2
2:	6	2:	3
3:	6	3:	2
4:	15	4:	5
5:	5	5:	8
...	...	...	
3539:	2	6700:	3
3540:	22	6701:	8
3541:	13	6702:	8
3542:	6	6703:	5
3543:	15	6704:	3

5 Summary and Conclusions

This research has applied the two digits Benford's Law (2BL) in addition to other related tests developed by [8] mainly the second order, the summation and the duplications tests to help identify the anomalies present in the Kenyan 2017 first round election results for the top two candidates namely Uhuru Kenyatta and Raila Odinga. The data used in this study was scrapped from the IEBC website.

The study reveals that the data do not follow the 2BL distribution. This is a pontential red flag pointing at possible anomalies and irregularities present in the data for the two candidates. These anomalies are all significant at 5% significance level using the Chi-square test, Euclidean distance test and the M.A.D.

References

[1] P. Klimek, Y. Yegorov, R. Hanel, and S. Thurner. Statistical detection of systematic election irregularities. *Proceedings of the National Academy of Sciences*, 109(41):16469–16473, 2012.

[2] F. Benford. The law of Anomalous Numbers. *Proc.Am Philos Soc*, 78:551–572, 1938.

[3] A. Berger and T. Hill. A basic theory of Benfords Law. *Probability Surveys*, 8:1–126, 2011.

[4] W. R Mebane. Election Forensics: Vote Counts and Benford's Law. *Summer Meeting of the Political Methodology Society, UC-Davis, July 20-22*, 2006.

[5] C. Breunig and A. Goerres. Searching for electoral irregularities in an established democracy: Applying Benfords Law tests to Bundestag elections in unified Germany. *Elect Stud*, 30:534–545, 2011.

[6] F. Cantu and S.M. Saiegh. Fraudulent Democracy? An analysis of Argentina's Infamous decade using Supervised Machine Learning. *Polit Anal*, 19:409–433, 2011.

[7] W. R. Mebane and K. Kalinin. Comparative Election Fraud Detection. *The American Political Science Association, Toronto, ON, Canada*, 2009.

[8] M. J. Nigrini. *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations*. Wiley and Sons: New Jersey., 2011.

[9] W. R Mebane. Anomalies and Frauds(?) in the Kenya 2017 Presidential Election. Technical report, Department of Political Science and Department of Statistics, University of Michigan, 2017.