Benford Group Project

Fionnuala McPeake,Kecheng Liang, Hao Qin, Shiyu Zhang

November 26, 2018

Introduction

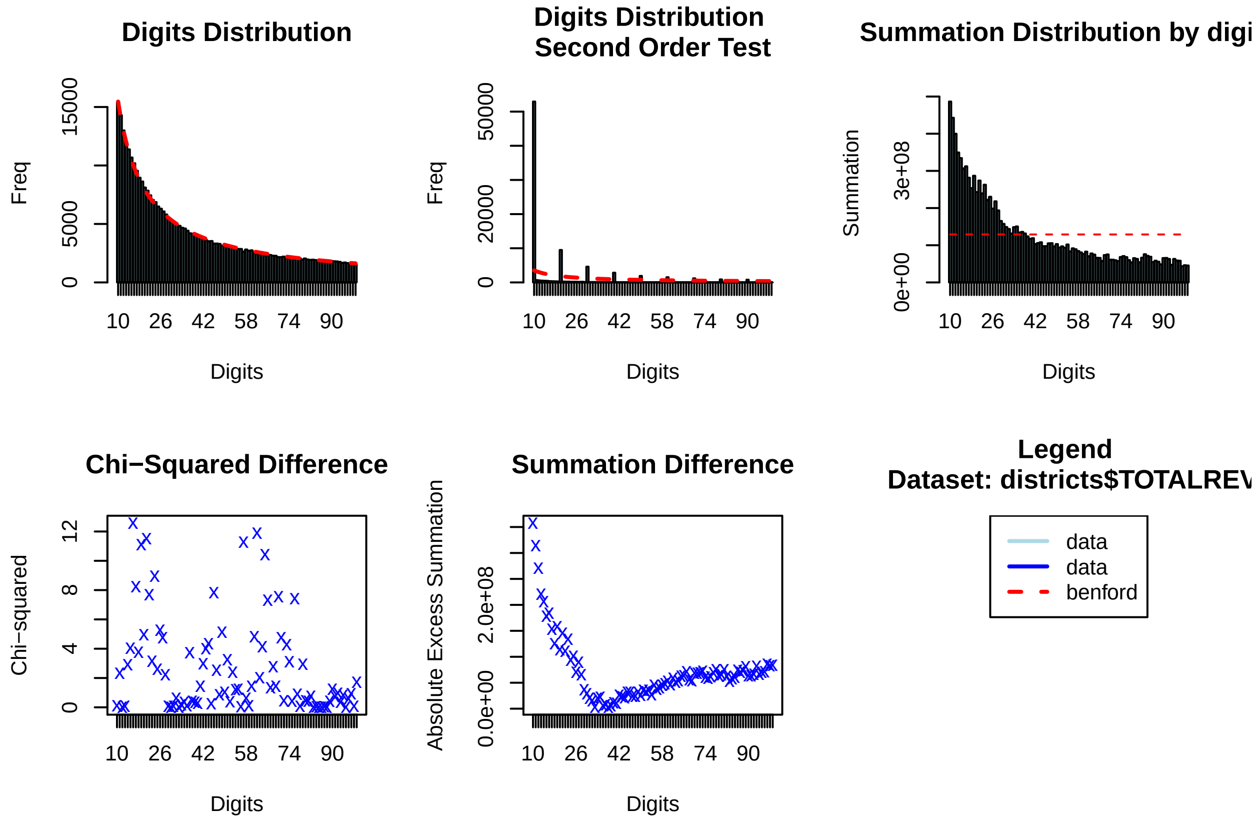
Benford's law describes the distribution of first digits for many real life collections of numeric data. As this law describes a rather suprising phenominon, it offers an interesting and easy to use way to detect fraud. The benford.analysis package in R produces easily interprited graphs, which we have used in this project in a variety of datasets.

Analysis

In order to explore Benford's law, we selected large numeric data sets. We started with a dataset describing the financial details of every school district in every state between the years of 1992 and 2016.

The below graph shows the results of the Benford analysis on the total revenue for the entire data set. The digits distribution fits the Benford analysis nearly perfectly. While the second order test does not fit the distribution as well as the first order, it is still an overall good fit. The summation distribution by digits graph is not particularly good, but does show a sort of pattern, which may be a result of the second order test.

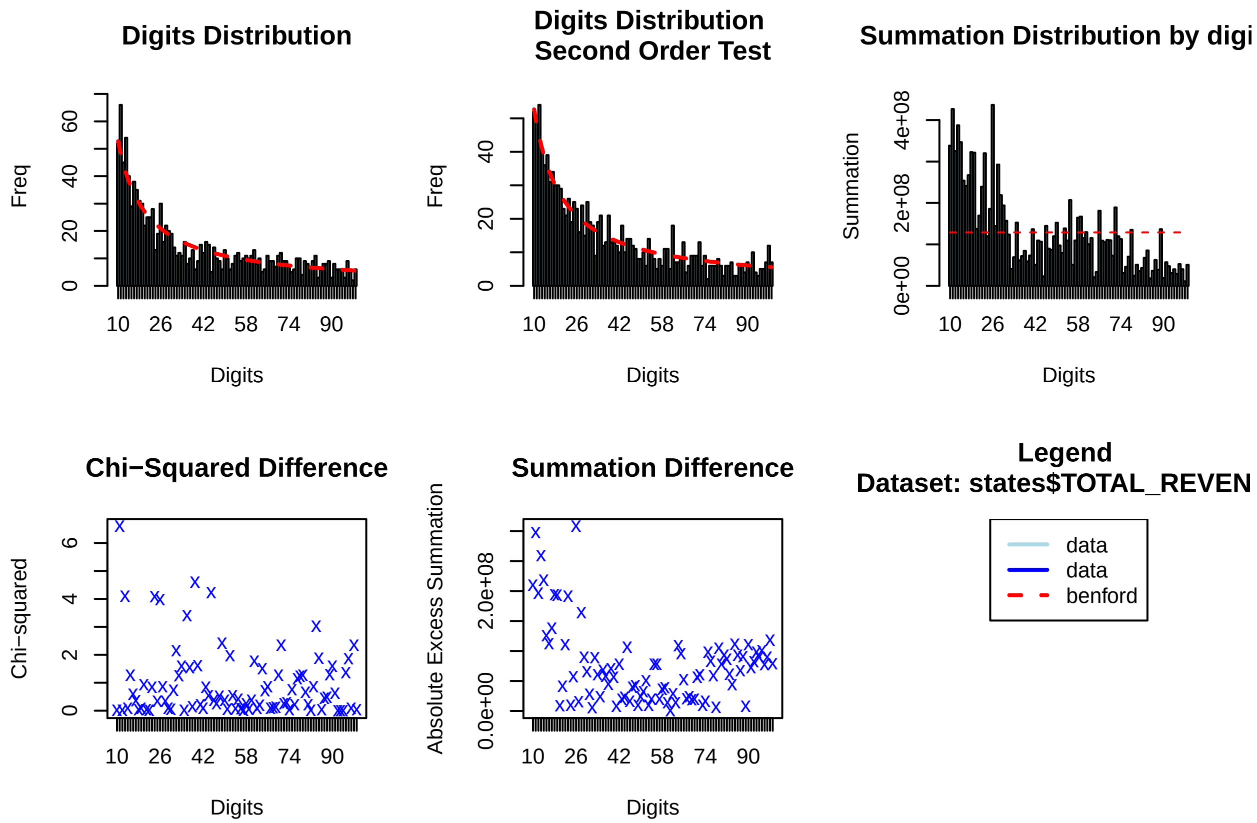
```
## Parsed with column specification:
## cols(
##   STATE = col_character(),
##   ENROLL = col_integer(),
##   NAME = col_character(),
##   YRDATA = col_integer(),
##   TOTALREV = col_integer(),
##   TFEDREV = col_integer(),
##   TSTREV = col_integer(),
##   TLOCREV = col_integer(),
##   TOTALEXP = col_integer(),
##   TCURINST = col_integer(),
##   TCURSSVC = col_integer(),
##   TCURONON = col_integer(),
##   TCAPOUT = col_integer()
## )
```



We ran the Benford analysis on the same data source looking from a state level rather than a district level. This reduced the number of data points nearly 300 fold.

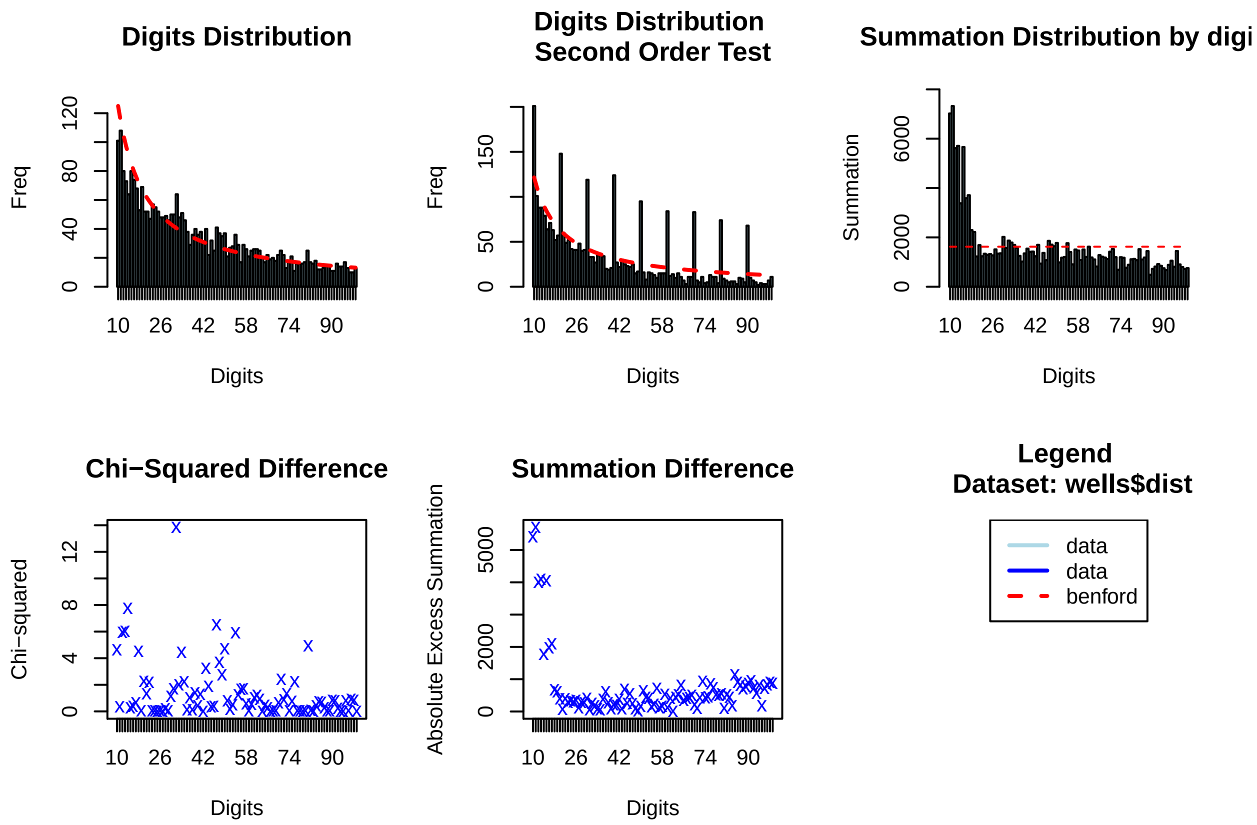
The digits distribution graph at the state level does not fit the expected distribution as well as at the district level, but does have the same shape. While not all of the bins meet the imposed red line indicating the desired level, no bin seems to be concerningly above or below the line. The same can be said for the graph showing the digits distribution of the second order. Both of these plots seem to be satisfactory. The summation distribution by digits does not fit the expected distribution very well. While it has a similar overall shape as the same graph at the district level, the discrepencies of the digit distributions seems to be conflating. For the data stating the total revenue for schools, both at the state and district level, we conclude that there is no fraud, and believe that the reduction in datapoints has contributed to the deviation from Benford's law at the state level.

```
## Parsed with column specification:
## cols(
##   STATE = col_character(),
##   YEAR = col_integer(),
##   ENROLL = col_double(),
##   TOTAL_REVENUE = col_integer(),
##   FEDERAL_REVENUE = col_integer(),
##   STATE_REVENUE = col_integer(),
##   LOCAL_REVENUE = col_integer(),
##   TOTAL_EXPENDITURE = col_integer(),
##   INSTRUCTION_EXPENDITURE = col_integer(),
##   SUPPORT_SERVICES_EXPENDITURE = col_integer(),
##   OTHER_EXPENDITURE = col_double(),
##   CAPITAL_OUTLAY_EXPENDITURE = col_integer()
## )
```



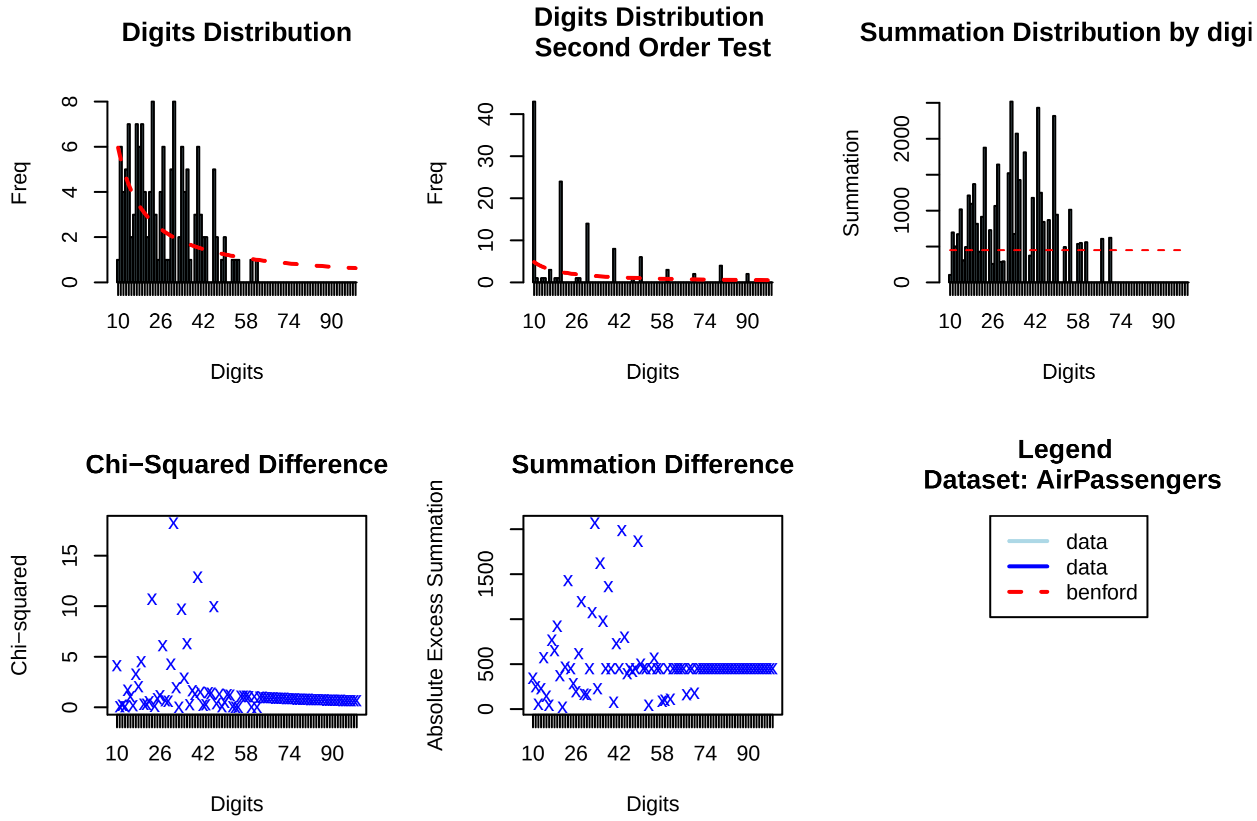
While Benford's law is often used to detect fraud in financial situations, it should hold up on any dataset, theoretically. To test this we steered away from financial data, and looked at distance.

The below dataset describes the distance that villagers in Benglidesh traveled, in meters, to the nearest safe well if they knew that their well was contaminated, and chose to switch to a new one. The digits distribution holds up well, with no data points being particularly concerning. The second order test does have some odd results. While the majority of the data points follow the distribution as expected, there are large outliers are seemingly regular intervals. We believe that this is due to rounding. For example, the researchers may have founded the distance to the nearest five meters, as the distance may have represented the distance traveled for several households and they wanted to account for the fact that there was no one starting point for all of the people represented by this data point. The summation distribution by digits holds up fairly well, considering the oddity of the second order test. We conclude that there is not fraud in this dataset.



We again tested Benford's law outside of an economic situation. The below graphs show the results of a Benford analysis on the number of monthly international airplane passengers between the years of 1949 and 2016.

The digits distribution does not fit the expected distribution at all. The same can be said for the second order test, where we again see outliers at fairly constant intervals. This is likely due to two reasons. The first reason being the capacity of airplanes, especially during this time period. There is not a wide variety international airplane models, and given the distance they are going, it would be unreasonable to make a plane that has a capacity with a high leading digit- with the result being either too small to be profitable, or too large to be reasonable. The second reason is likely due to the small number of datapoints. As we saw with the district vs state data, larger data sets seem to lead to results that adhere more closely to Benford's law. As this data set has 144 data points, this likely contributes to the poor results. Despite the data not conforming to Benford's law, we do not believe that this data if forged, as the explinations above seem reasonable.



Conclusion

A Benford analysis is an easy way to detect fraud within numeric data. While the above analyses seem to hint at some of the characteristics of data that the Benford analysis may not account for, such as sample size and rounding, we believe that the results of this test are a good starting point to explore the data in a new way.