Benford_Project

Summer Zu, Tianying Xu, Vector Liu, Sky Liu

11/28/2018

Load all the packages

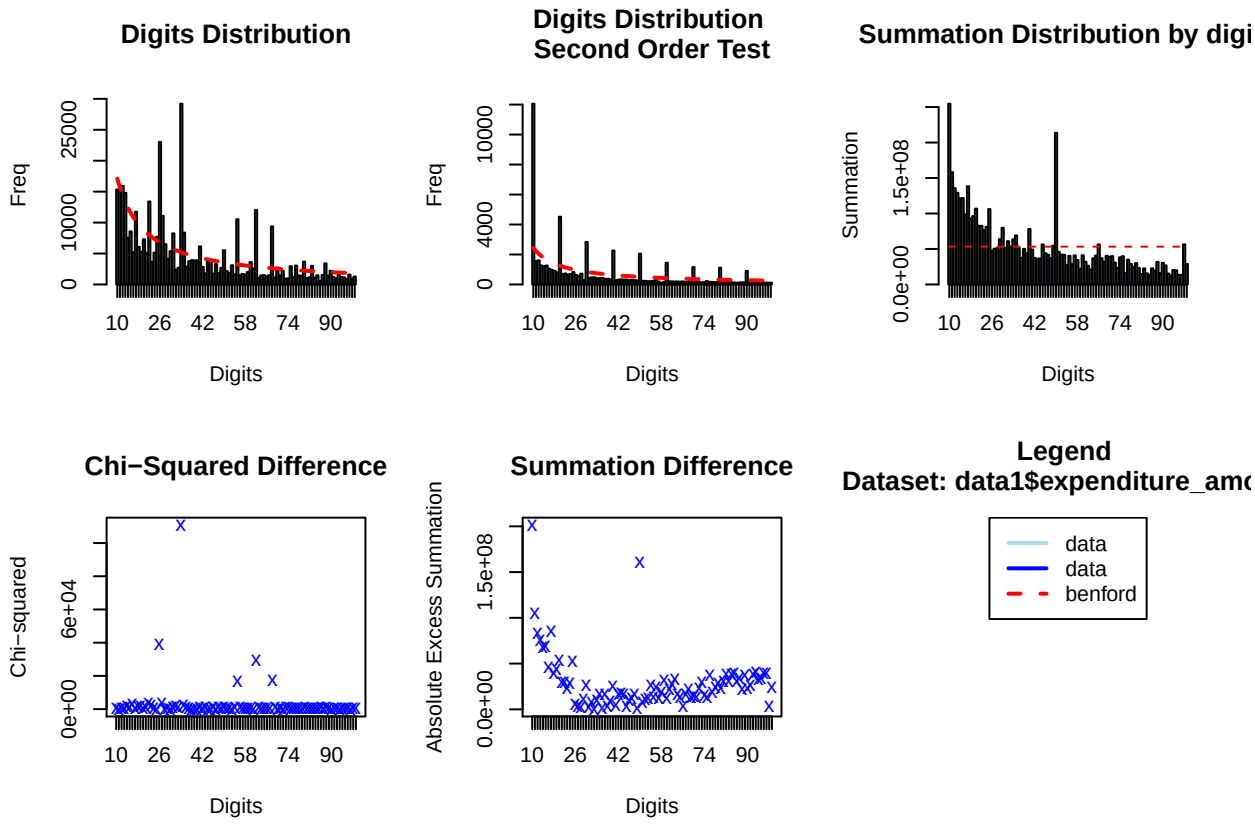
I. Data Our data is about independent expenditures. Independent expenditures are what some refer to as "hard money" in politics – spending on ads that specifically mention a candidate (either supporting or opposing). The money for these ads must come from Political Ad Spending(PACs) that are independent of the candidate and campaign, and the PACs cannot coordinate with the candidate.

Create all the dataset.

```
data = fread("fec-independent-expenditures.csv")
data1<-data%>%dplyr::select(committee_id,committee_name,report_year,payee_name,payee_state,expenditure_
#whether the ads is in support of or opposition to the candidate
support<-data%>%filter(support_oppose_indicator=="S")
oppose<-data%>%filter(support_oppose_indicator=="O")
#candidate office: P(president), S(senate), H(house)
president<-data%>%filter(candidate_office=="P")
senate<-data%>%filter(candidate_office=="S")
house<-data%>%filter(candidate_office=="H")
```

II. Benford analysis on all advertising expenditure

```
bfd_cp <- benford(data1$expenditure_amount)
plot(bfd_cp)
```



```
bfd_cp
```

```
##
## Benford object:
##
## Data: data1$expenditure_amount
## Number of observations used = 413131
## Number of obs. for second order = 58883
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.499
##          Var  0.075
##  Ex.Kurtosis -1.006
##     Skewness -0.061
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     34      24011.04
## 2     26      16267.61
## 3     62       9175.21
## 4     55       7310.11
## 5     68       6760.68
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  data1$expenditure_amount
## X-squared = 268660, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  data1$expenditure_amount
## L2 = 0.010634, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.005096281
## Distortion Factor: -8.752683
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

From the plots of Benford analysis of all the advertising expenditure amount, we can discover several features of the expenditure amount. From the first plot and the benford analysis, it is evident that digits 34, 26, 62, 55 and 68 have deviations from the Benford Law, and among which, digit 34 and 26 have the largest deviation against benford law. From the second plot, we can see that the structure of data is slightly deviated from the benford law. The third plot shows that there is one significant deviation against benford's law.The last two plot shows exactly the 5 unexpected data point and the one significant deviated structure.

Also, the mean of mantissa is very close to 0.5, while skewness is close to 0. In general, he p-value is less than 0.05, so the distribution does not exactly follow benford distribution.

III. Further analysis on suspected expenditures

From the tables, it is evident that:

52251 suspected expenditure out of 413237 21202 happened in 2012 and 27767 in 2016, election year 25833 were for president election, 23715 were for senate election About half of the suspected expenditure were in support of the candidate (27844) and the other half were in opposition to the candidate (24402)

The top candidate with the largest amount of suspected expenditure in support of the candidate is HILLARY CLINTON (7386). The top candidate with the largest amount of suspected expenditure in opposition to the candidate is DONALD TRUMP (7386).

```
#extract the observations with the largest discrepancies by using the getSuspects function
suspects_cp <- getSuspects(bfd_cp, data1)

#most of suspect expenditures happens in 2012 and 2016 with more than 20 thousand records.
table(suspects_cp$report_year)
```

```
##
##  2004  2005  2006  2007  2008  2009  2010  2011  2012  2013  2014  2015
##   148    43   234    18   498    11   540    23 21202    42  1618   107
##  2016
## 27767
```

```
#most of suspect expenditures happens in president and senate election
table(suspects_cp$candidate_office)
```

```
##
##           H     P     S
##     5  2698 25833 23715
```

```
table(suspects_cp$support_oppose_indicator)
```

```
##
##           O     S
##     5 24402 27844
```

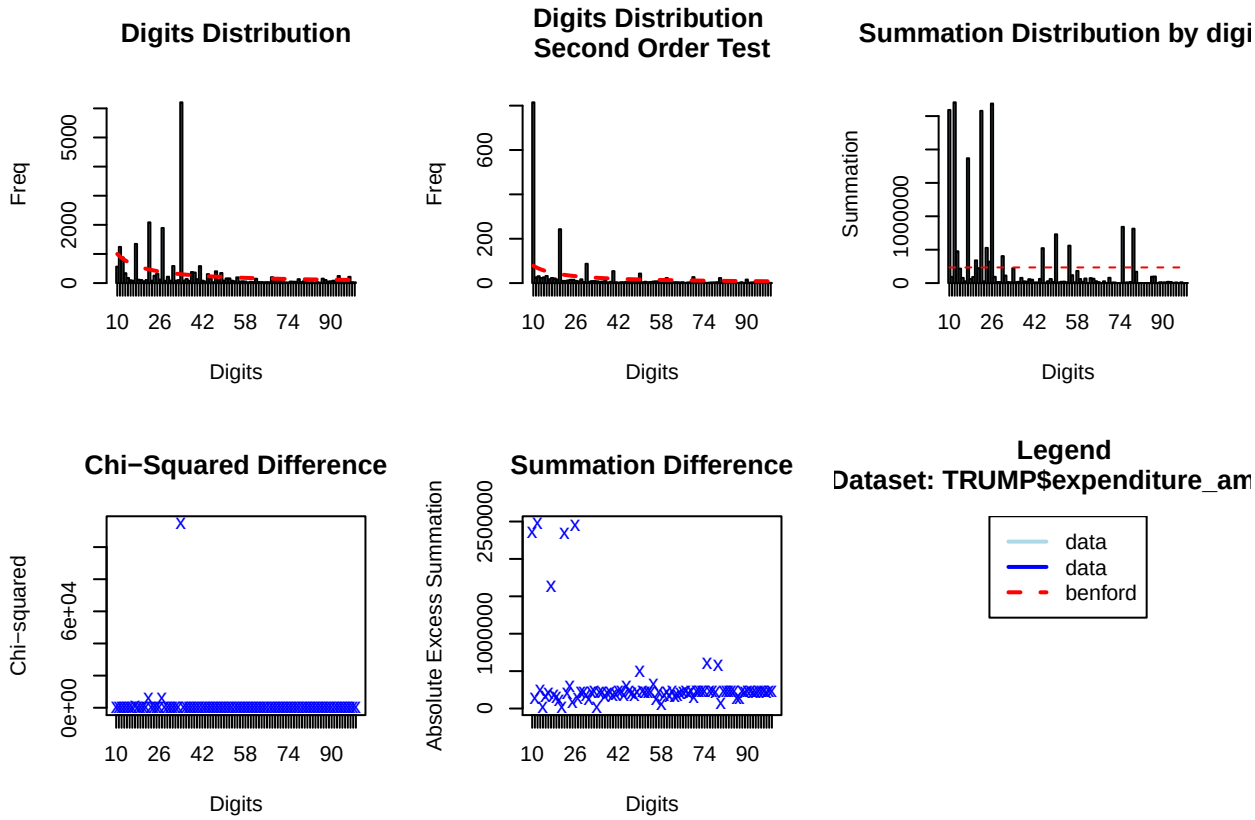
```
s_suspects <- suspects_cp%>%filter(support_oppose_indicator=="S")
o_suspects <- suspects_cp%>%filter(support_oppose_indicator=="O")
s_sus_candidate<-as.data.frame(table(s_suspects$candidate_name))
s_sus_candidate<-s_sus_candidate%>%arrange(Freq)
top_s_sus_candidate<-tail(s_sus_candidate)

o_sus_candidate<-as.data.frame(table(o_suspects$candidate_name))
o_sus_candidate<-o_sus_candidate%>%arrange(Freq)
top_o_sus_candidate<-tail(o_sus_candidate)
```

IV. Benford Analysis on Trump Expenditure

1. The majority of suspected expenditure in opposition to Trump were from WORKING AMERICA.

```
#Create a new dataset with candidate specified to Trump, Donald.
TRUMP<-data1%>%filter(candidate_name=="TRUMP, DONALD")
bfd_Trump <- benford(TRUMP$expenditure_amount)
plot(bfd_Trump)
```




```
bfd_Trump
```

```
##
## Benford object:
##
## Data: TRUMP$expenditure_amount
## Number of observations used = 24159
## Number of obs. for second order = 1890
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.483
##          Var  0.055
##  Ex.Kurtosis -0.232
##     Skewness -0.054
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     34       5903.86
## 2     22       1618.61
## 3     27       1510.43
## 4     17        745.29
## 5     15        580.15
##
```

```
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  TRUMP$expenditure_amount
## X-squared = 136800, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  TRUMP$expenditure_amount
## L2 = 0.11778, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.01072711
## Distortion Factor: -24.59666
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

In this dataset, we found digits 34 has the largest deviation Ex.Kurtosis is not close to -1.2. So there are some unexpected data point that doesn't follow benford's law.

2. Get the detailed information about the unexpected data

```
suspects_Trump <- getSuspects(bfd_Trump, TRUMP)
table(suspects_Trump$committee_name)
```

```
##
##                    45COMMITTEE INC.                AMERICAN FUTURE FUND
##                                   1                                   1
##                    AVAAZ FOUNDATION                  EDUCATORS FOR OHIO
##                                   7                                   1
## LIFT LEADING ILLINOIS FOR TOMORROW           PLANNED PARENTHOOD VOTES
##                                   1                                   4
##                  THE 2016 COMMITTEE                      WORKING AMERICA
##                                   2                                8276
```

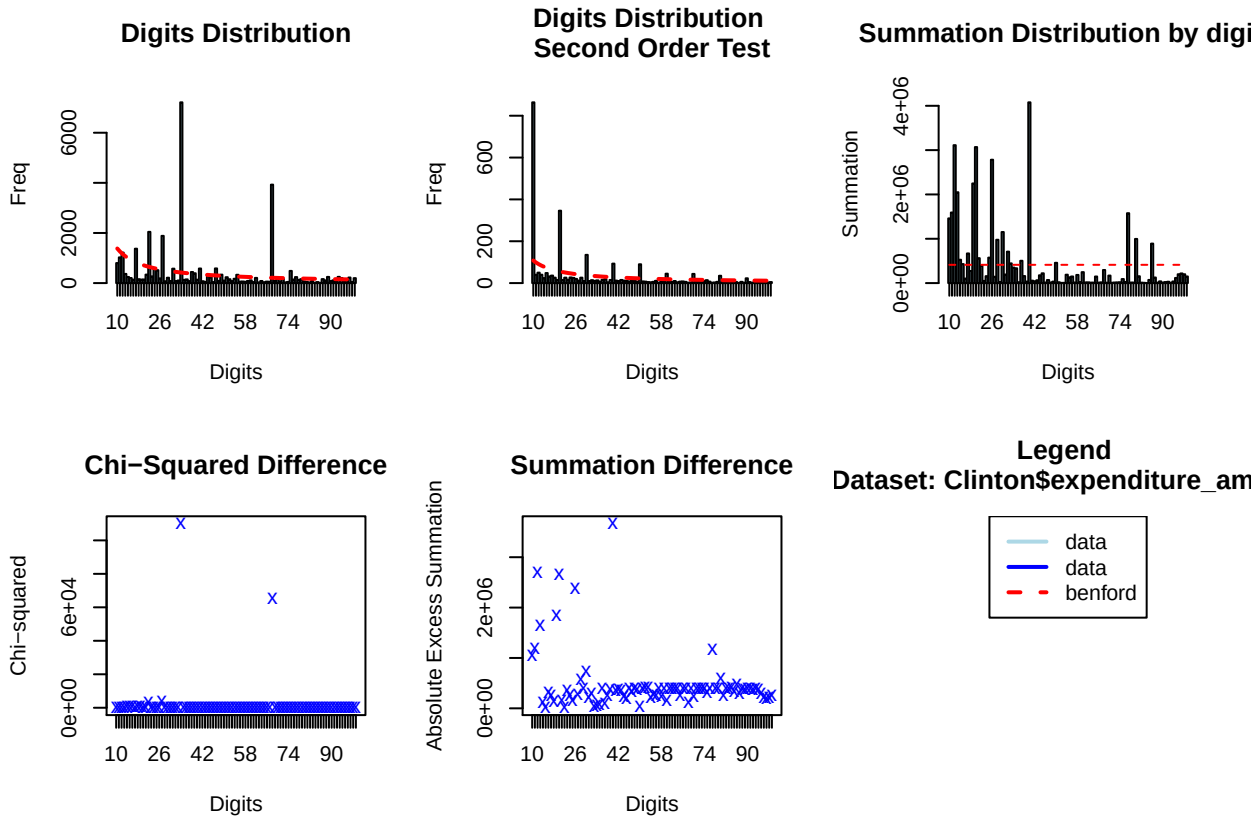
From this function, working America seems have the largest porportion of suspicious.

V. Benford Analysis on Hillary Expenditure

The majority of suspected expenditure in support of Clinton were from WORKING AMERICA.

Create a new dataset with candidate specified to Clinton, Hillary.

```
Clinton<-data1%>%filter(candidate_name=="CLINTON, HILLARY")
bfd_Clinton <- benford(Clinton$expenditure_amount)
plot(bfd_Clinton)
```




```
bfd_Clinton
```

```
##
## Benford object:
##
## Data: Clinton$expenditure_amount
## Number of observations used = 33335
## Number of obs. for second order = 2607
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.542
##          Var  0.065
##  Ex.Kurtosis -0.680
##     Skewness -0.221
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     34       6794.34
## 2     68       3717.65
## 3     22       1400.46
## 4     27       1358.50
## 5     14        753.82
##
```

```
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  Clinton$expenditure_amount
## X-squared = 194250, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  Clinton$expenditure_amount
## L2 = 0.050979, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.01027823
## Distortion Factor: -3.129796
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

By checking the data from Hillary, we found the data is has largest deviation on digits 34 and 68. The p-value 2.2e-16 is less than 0.05, which means the distribution does no follow benford law.

```
suspects_Clinton <- getSuspects(bfd_Clinton, Clinton)
table(suspects_Clinton$committee_name)
```

```
##
##                                 80-20 PAC
##                                         1
##                             AFT SOLIDARITY
##                                         1
##               CITIZENS UNITED SUPER PAC LLC
##                                         1
##                    IMMIGRANT VOTERS WIN PAC
##                                         3
##                           NEA ADVOCACY FUND
##                                         1
## OHIO ENVIRONMENTAL COUNCIL ACTION FUND INC.
##                                         2
##                    PLANNED PARENTHOOD VOTES
##                                         2
##               REPUBLICAN NATIONAL COMMITTEE
##                                         1
##                             STOP HILLARY PAC
##                                         4
##                          THE 2016 COMMITTEE
##                                         3
##                 VOCES DE LA FRONTERA ACTION
##                                         4
##                         WISCONSIN JOBS NOW!
##                                         1
##                             WORKING AMERICA
##                                     11119
```

As we look more detial into the data, we found most of the suspicious record are from Working America. Maybe we should look into to the working America committee to see whether there are fraud in the data.