

Benford Group assignment

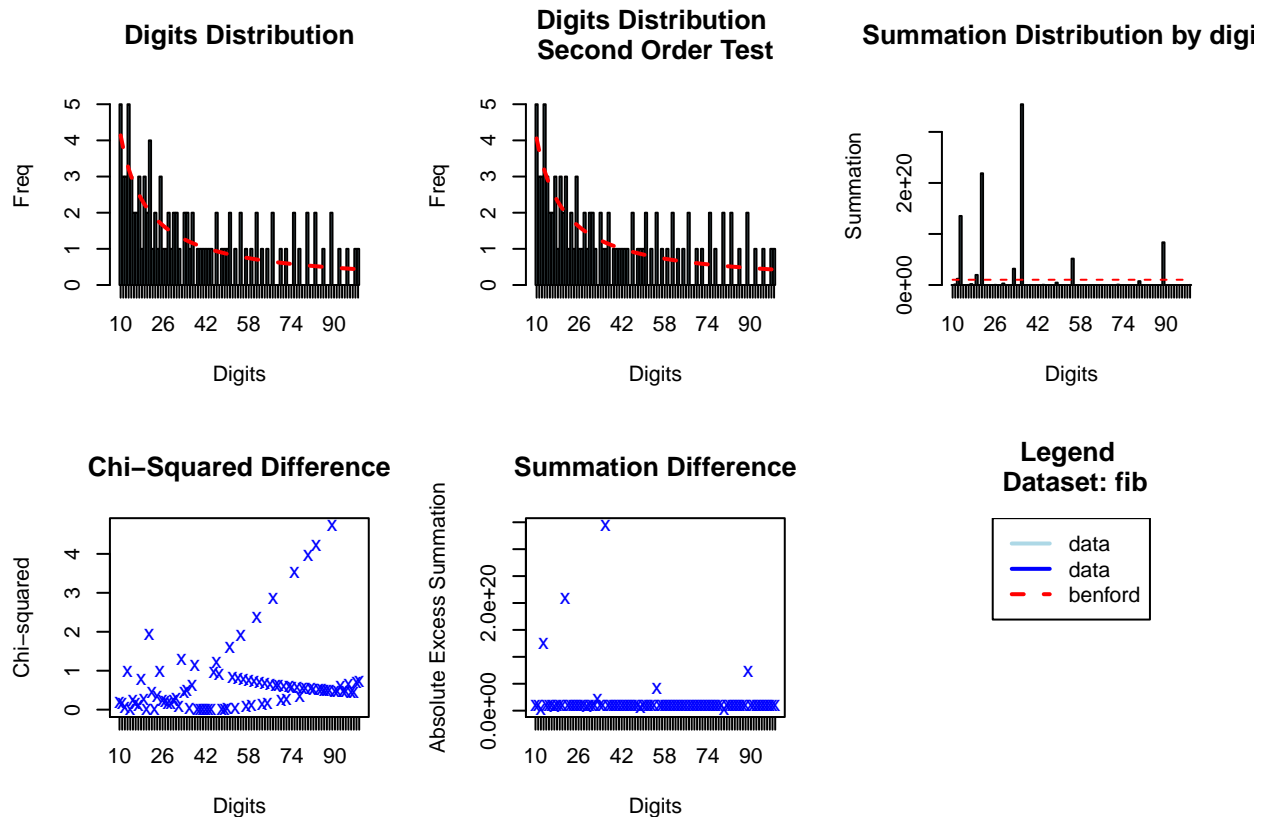
Dave Anderson, Longhao Chen, Tingrui Huang, Yudi Mao

11/26/2018

Fibonacci Numbers

To investigate the effect of increasing sample size on relation to Benford's Law, I chose to use Fibonacci numbers. I started with the first 100 numbers of the sequence and performed Benford's analysis on the first two leading digits.

100 Numbers

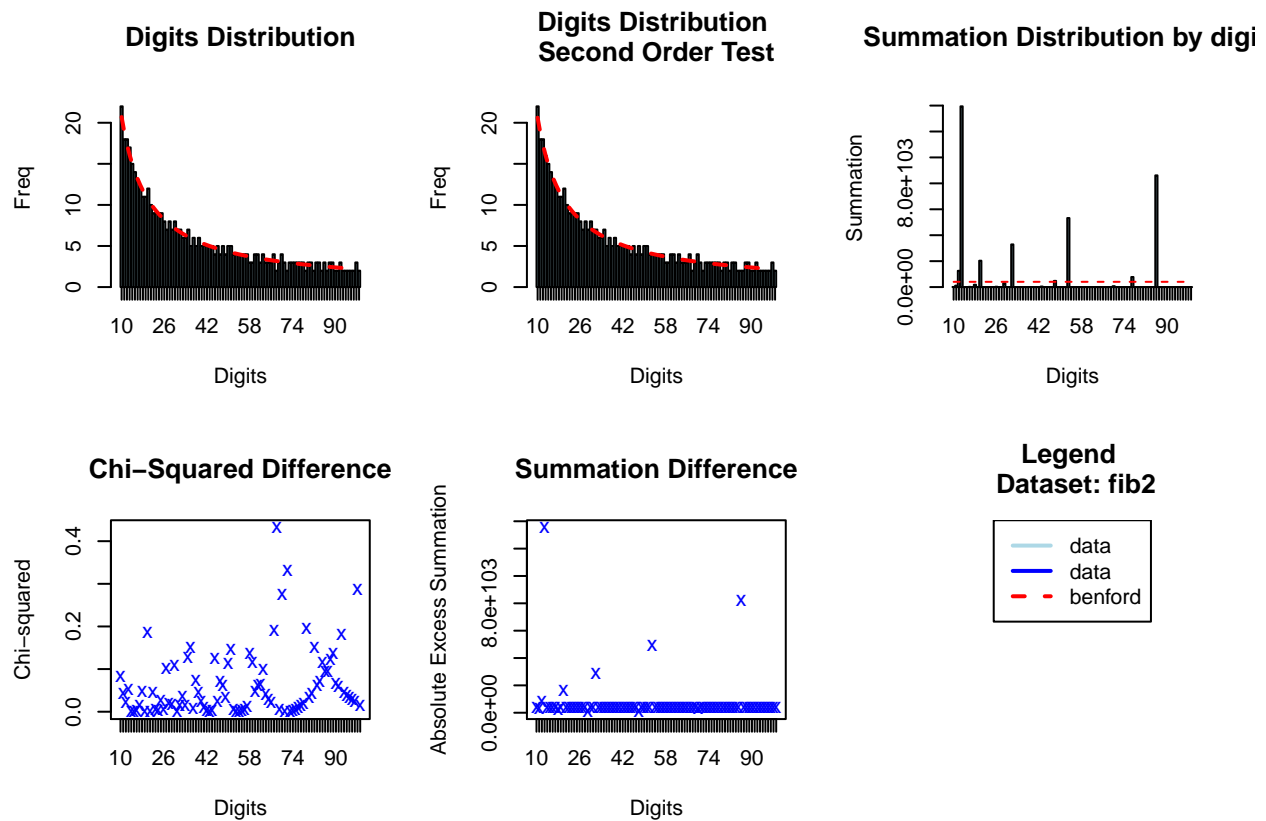


##	digits	absolute.diff
##	1:	21
##	2:	13
##	3:	89
##	4:	83
##	5:	80
##	6:	75
##	7:	67
##	8:	18
##	9:	25

10: 33 1.296498

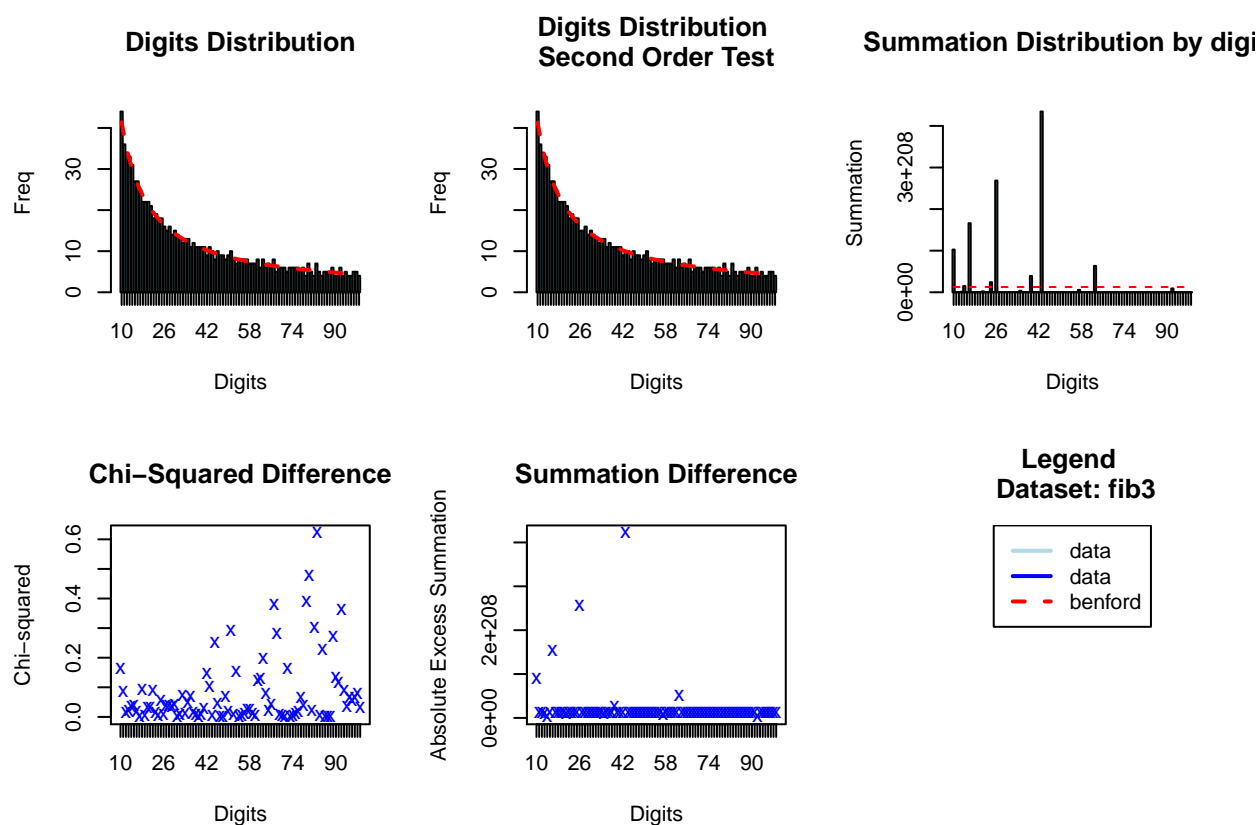
There is clearly a staggered pattern in the numbers, which creates a very interesting pattern on the chi-squared difference plot.

500 Numbers



Increasing our numbers from 100 to 500 has made a big difference in our analysis. The distribution of digits seems to conform to Benford very well. The pattern in the chi-squared difference is still evident, but the differences have decreased dramatically.

1,000 numbers



##	digits	absolute.diff
## 1:	10	2.607315
## 2:	83	1.798806
## 3:	11	1.788561
## 4:	80	1.604968
## 5:	51	1.566832
## 6:	67	1.565890
## 7:	45	1.545318
## 8:	18	1.481096
## 9:	79	1.462896
## 10:	68	1.340178

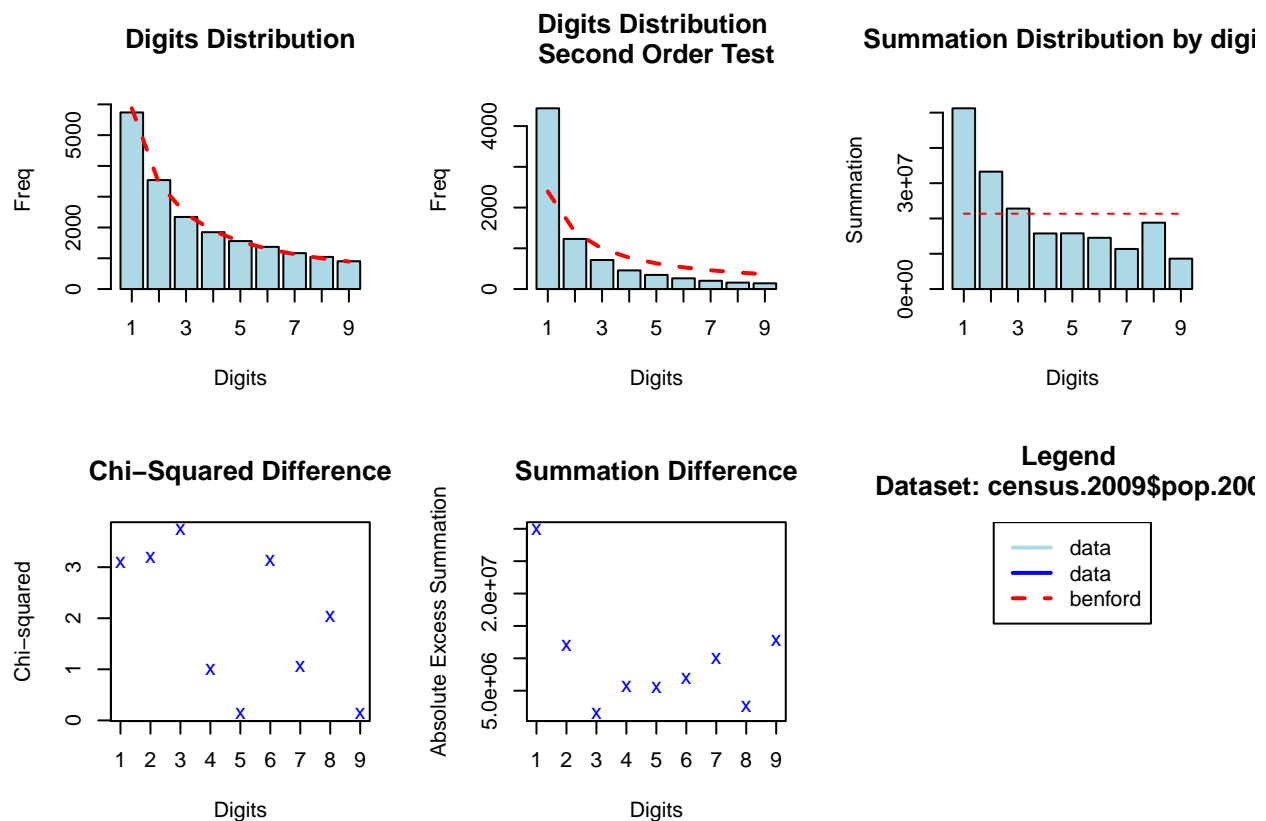
Interestingly, it seems as if increasing numbers to 1,000 is actually further from Benford's distribution in some ways. Overall, the pattern from the chi-squared difference has decreased, and most digits are closer to the law, but some digits, such as 20, have strayed further away. It is interesting that many multiples of tens are seen in the suspects list.

Census Data

```
d <- data(package = "benford.analysis")
## names of data sets in the package
#d$results[, "Item"]
data(census.2009)

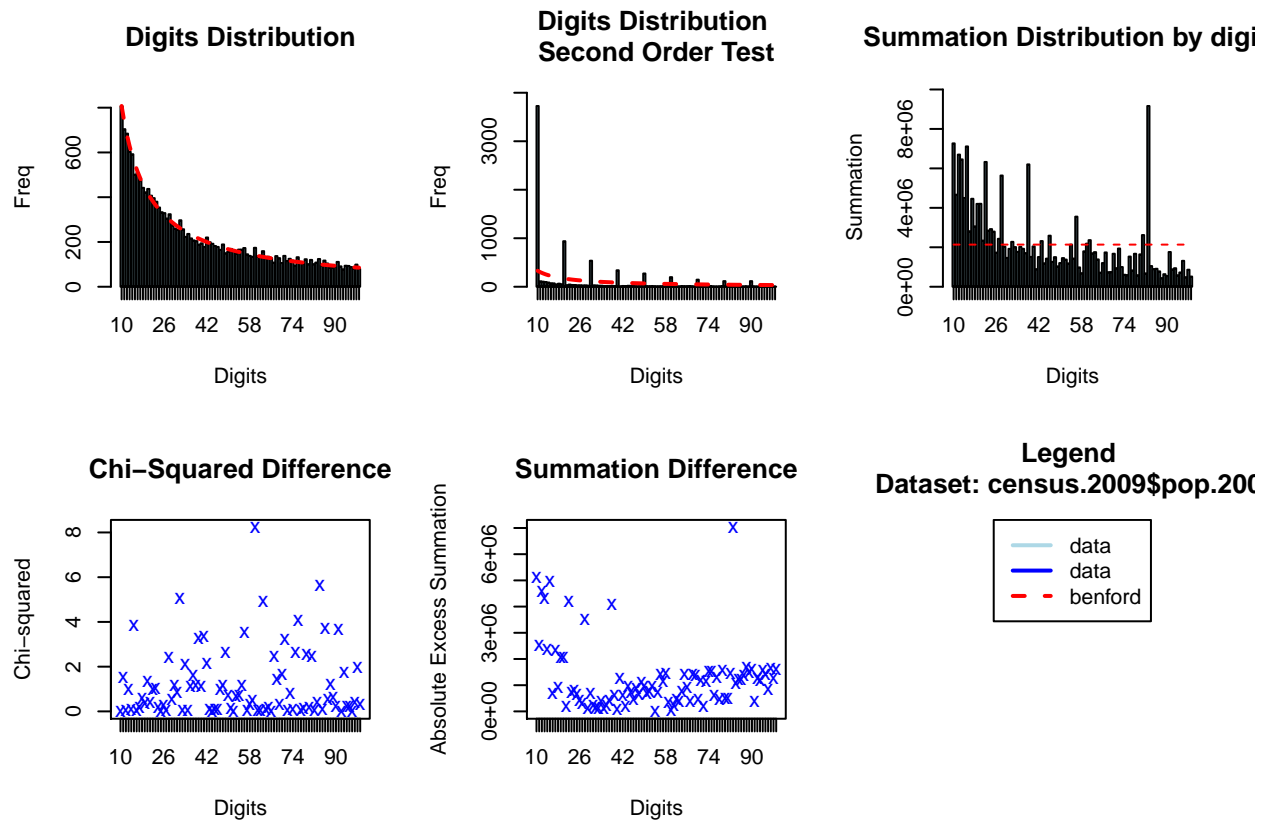
#The first code is to look at the high level test of reasonableness.
census1 <- benford(census.2009$pop.2009, number.of.digits = 1, sign = "positive", discrete = TRUE, round = 0)
#We can take a closer look at the first two digits, which is designed to select audit targets.
census2 <- benford(census.2009$pop.2009, number.of.digits = 2, sign = "positive", discrete = TRUE, round = 0)

plot(census1)
```



#From this plot we can see that red dotted line, which is the benford line, generally matches good with

```
plot(census2)
```



#This plot reveals a closer look at first two digits data.

By looking at the general information of the dataset. We can see that the values of mean, variance, Ex.Kurtosis and skewness well match with the expected values of 0.5, 0.0833, -1.2, and 0.

census1

```
##
## Benford object:
##
## Data: census.2009$pop.2009
## Number of observations used = 19509
## Number of obs. for second order = 7950
## First digits analysed = 1
##
## Mantissa:
##
##   Statistic  Value
##   Mean      0.503
##   Var       0.084
##   Ex.Kurtosis -1.207
##   Skewness  -0.013
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
```

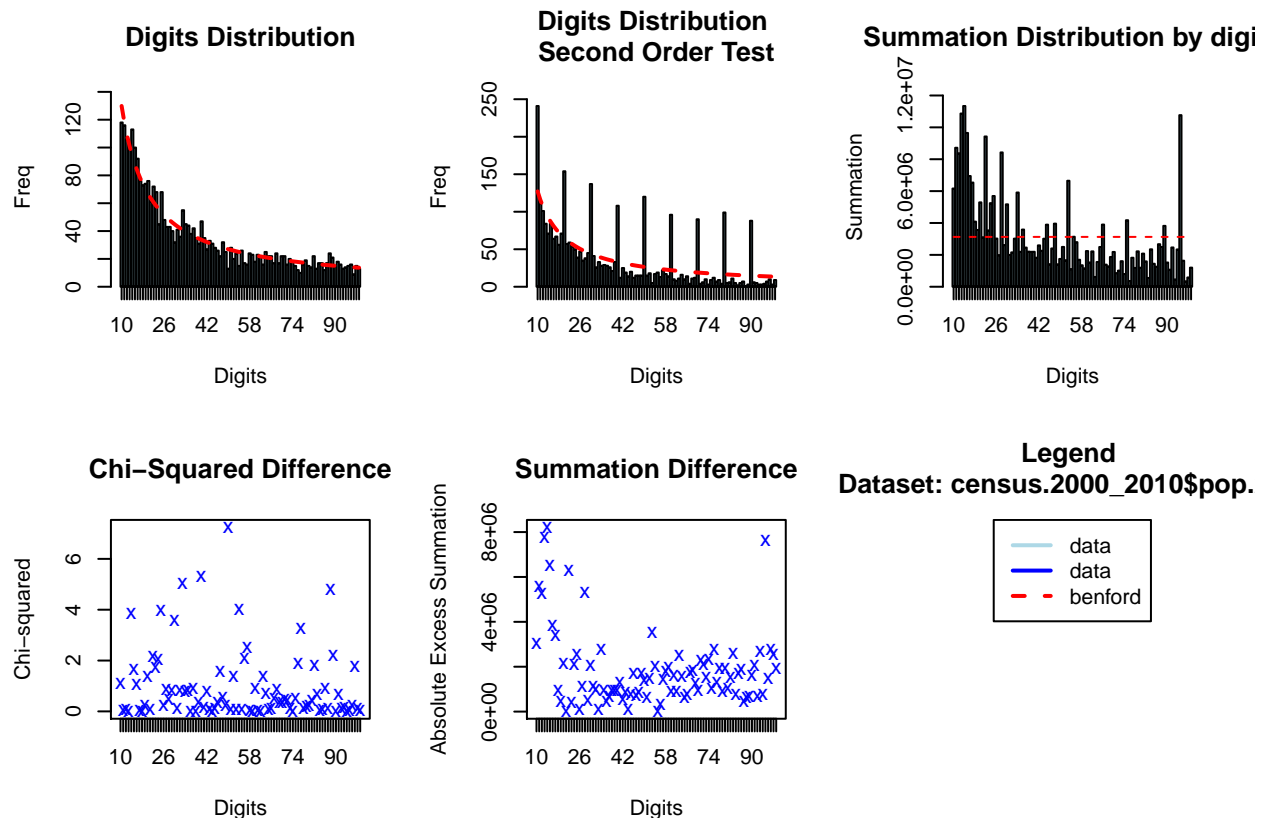
```
## 1      1      134.79
## 2      2      104.64
## 3      3       95.43
## 4      6       63.94
## 5      8       45.07
##
## Stats:
##
## Pearson's Chi-squared test
##
## data: census.2009$pop.2009
## X-squared = 17.524, df = 8, p-value = 0.0251
##
##
## Mantissa Arc Test
##
## data: census.2009$pop.2009
## L2 = 4.198e-05, df = 2, p-value = 0.4409
##
## Mean Absolute Deviation: 0.003119261
## Distortion Factor: 0.7404623
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
#This step is to find the suspicious targets using the first 2 digits Benford model.
suspects <- getSuspects(census2, census.2009)
suspects
```

```
##      state      town pop.2009
## 1: Alabama Alexander City city 15114
## 2: Alabama Bakerhill town 322
## 3: Alabama Center Point city 15519
## 4: Alabama Crossville town 1513
## 5: Alabama Cullman city 15302
## ---
## 794: Wisconsin Whitehall city 1582
## 795: Wisconsin White Lake village 321
## 796: Wyoming Ten Sleep town 328
## 797: Wyoming Wheatland town 3236
## 798: Wyoming Wright town 1550
```

From the suspect function, we can see that Alabama state has a handful of data that are suspicious.

Census 2000

```
data("census.2000_2010")
census2000_bfd <- benford(census.2000_2010$pop.2000)
plot(census2000_bfd)
```



```
MAD(census2000_bfd)
```

```
## [1] 0.001419661
```

```
chisq(census2000_bfd)
```

```
##
```

```
## Pearson's Chi-squared test
```

```
##
```

```
## data: census.2000_2010$pop.2000
```

```
## X-squared = 87.647, df = 89, p-value = 0.5207
```

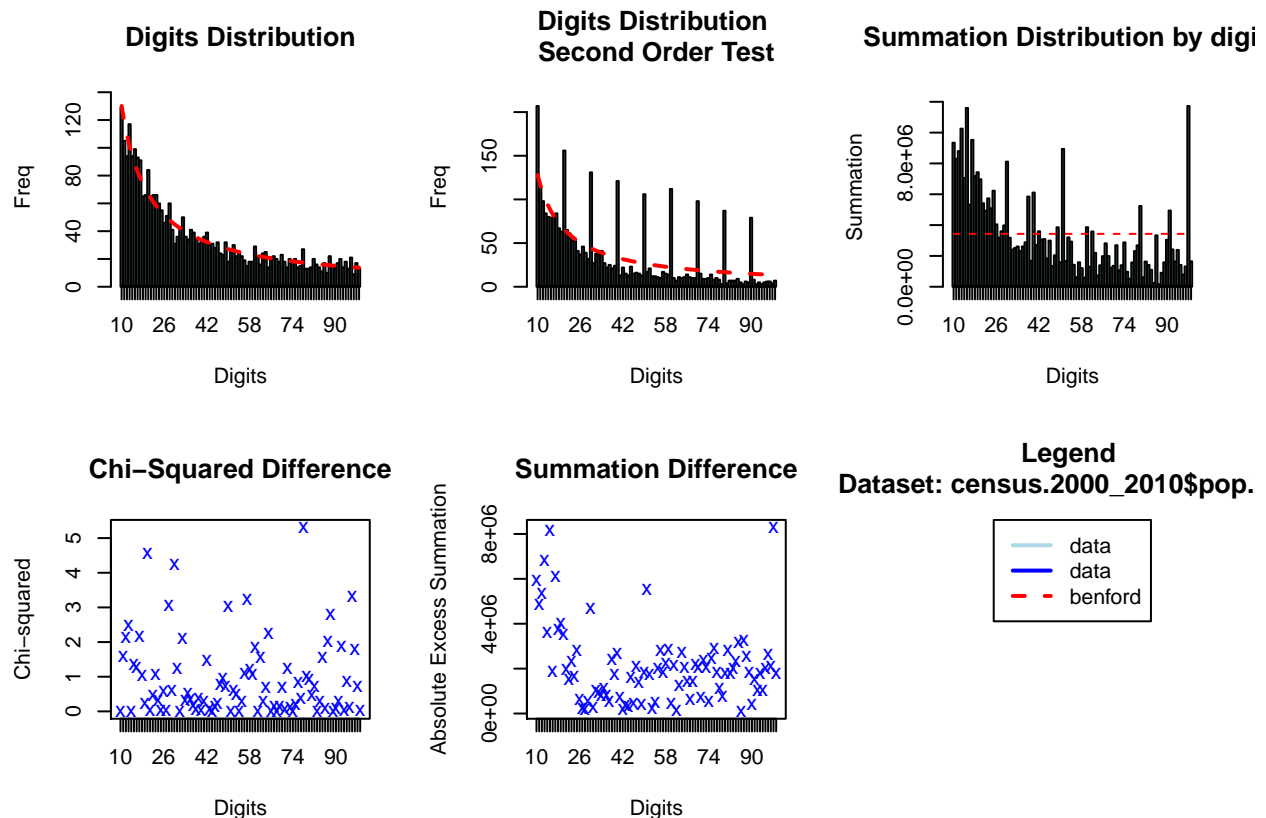
```
# Get suspicious value from the data
```

```
suspects_2000 <- getSuspects(bfd = census2000_bfd, data = census.2000_2010)
```

For the Year of 2000 census data, generally we have a pretty good result from the benford analysis. However, we have a few abnormal test results for the sequence starts with 2 and 3. From the Chi-squared test, we have a p-value of 0.5207 and it would indicate fail to reject the null hypothesis, which indicates that the distribution of the data is very close to the distribution of Benford Law.

Census 2010

```
census2010_bfd <- benford(census.2000_2010$pop.2010)
plot(census2010_bfd)
```



```
MAD(census2010_bfd)
```

```
## [1] 0.001428017
```

```
chisq(census2010_bfd)
```

```
##
```

```
## Pearson's Chi-squared test
```

```
##
```

```
## data: census.2000_2010$pop.2010
```

```
## X-squared = 82.809, df = 89, p-value = 0.6646
```

```
# Get suspicious value from the data
```

```
suspects_2010 <- getSuspects(bfd = census2010_bfd, data = census.2000_2010)
```

For the Year of 2010 census data, generally we have a pretty good result from the benford analysis. However, we have a few abnormal test results for the sequence starts with 2, 3 and 7. From the Chi-squared test, we have a p-value of 0.6646 and it would indicate fail to reject the null hypothesis, which indicates that the distribution of the data is very close to the distribution of Benford Law.

Chinese city population data

data source: <http://www.citypopulation.de/China-UA.html>

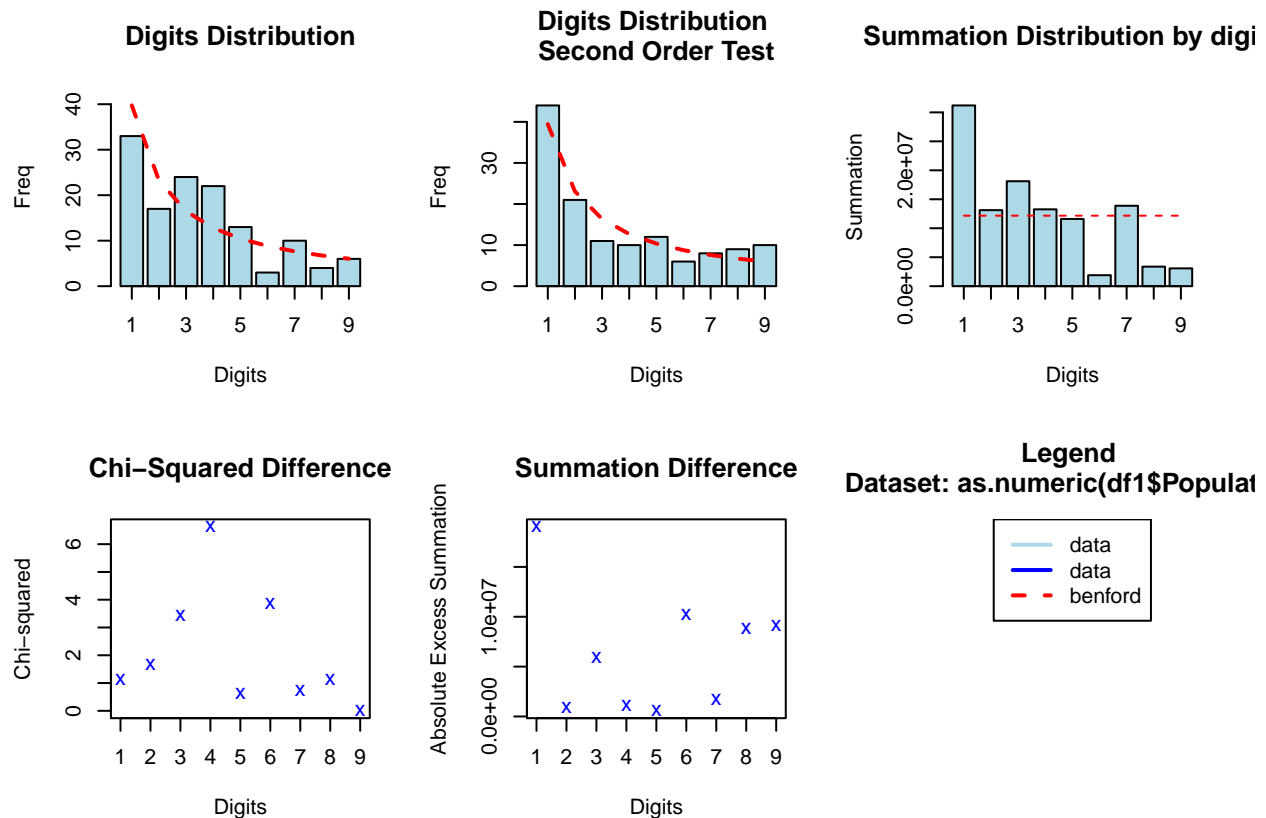
```
library(tidyverse)
```

```
library(benford.analysis) # loads package data(corporate.payment) # loads data
```

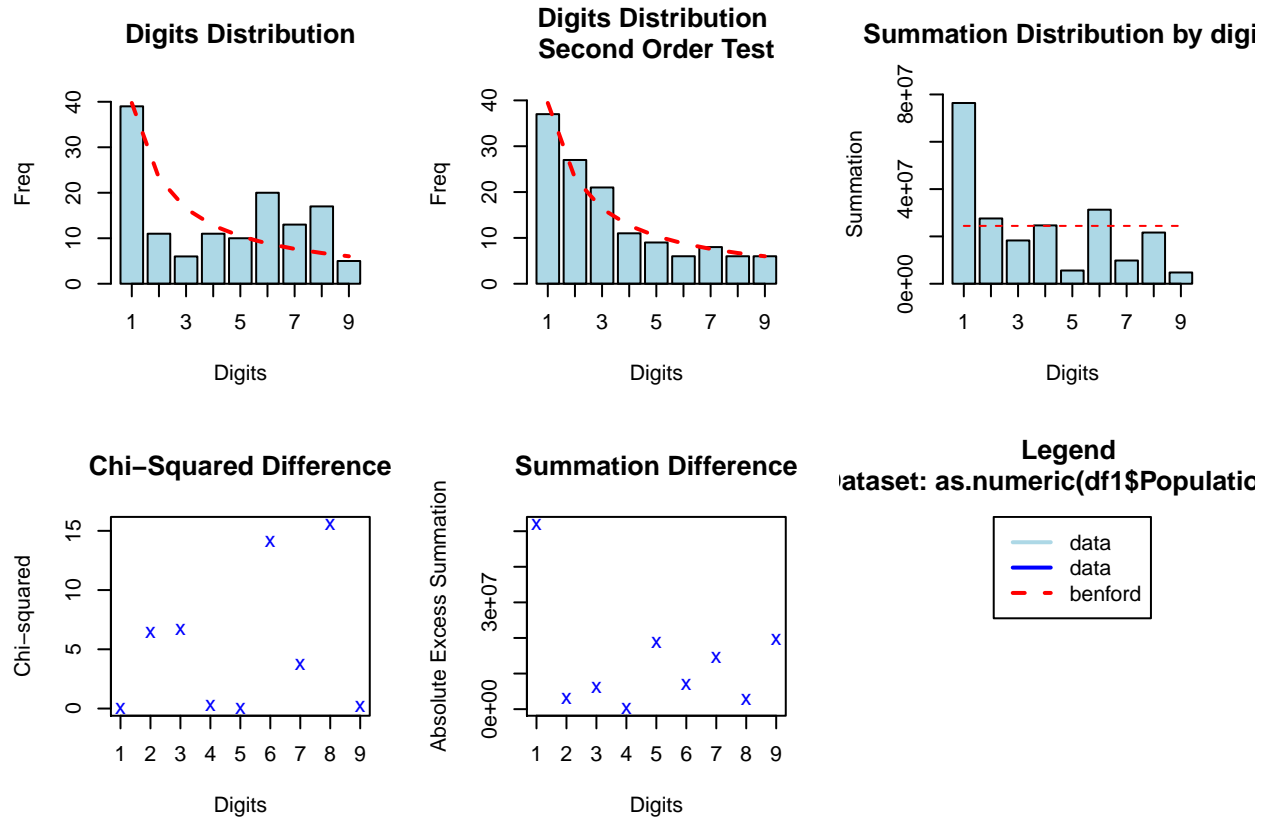


```
library(readxl)
chinese_census <- read_excel("chinese_census.xlsx")
df <- na.omit(chinese_census)
df1 <- select(df, Population, Population__1, Population__2)
```

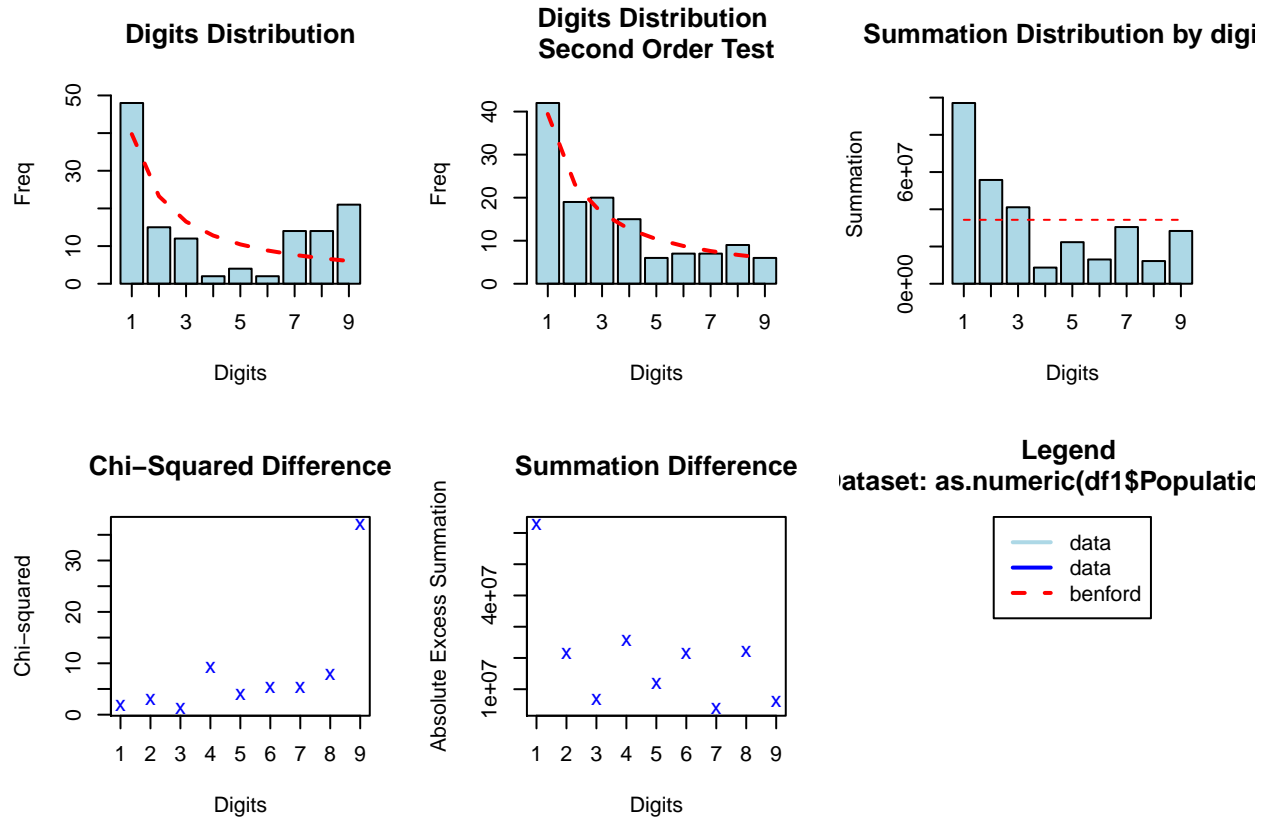
```
bfd.pop <- benford(as.numeric(df1$Population),
                  number.of.digits = 1)
plot(bfd.pop) #Census 1990
```



```
bfd.pop1 <- benford(as.numeric(df1$Population__1),
                   number.of.digits = 1)
plot(bfd.pop1) #Census 2000
```



```
bfd.pop2 <- benford(as.numeric(df1$Population__2),
                     number.of.digits = 1)
plot(bfd.pop2) #Census 2010
```



The results do not well fit benford law. The reason could be: 1. Dataset is relatively small (132 obs only) 2. Data is not complete. This dataset only contains population of those cities over 750,000, and they are urban populations.