Benford Law Fraud Detection

Jianhao Yan, Becky, Megha, Yifu Dong

11/25/2018

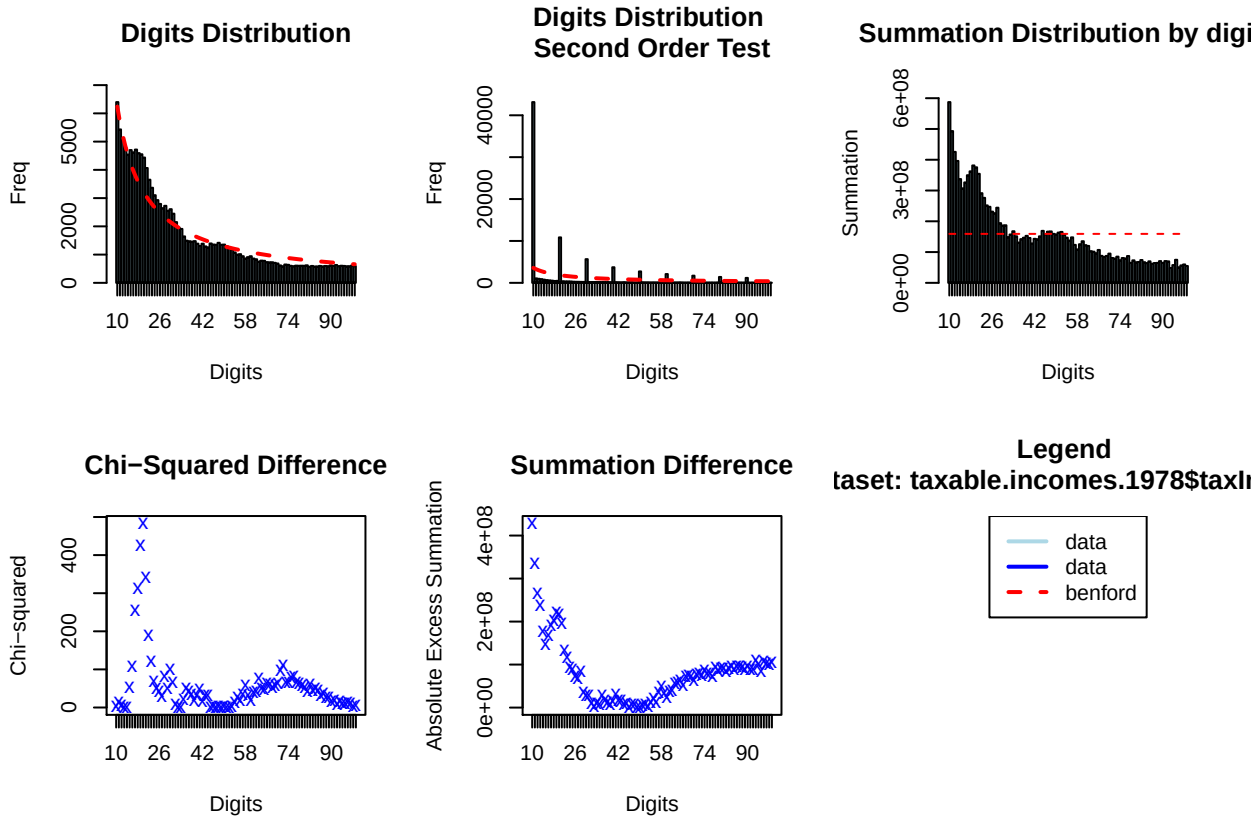
Taxable Incomes 1978

The dataset "Taxable Incomes 1978" was particularly interesting for Benford Analysis. This is because, the United States Revenue Act of 1978 came into effect. The Act made tax cuts on individual income taxes and corporate taxes to sustain the then state of recovery. We check whether or not this data follows Benford's Law.

```
library(benford.analysis)
data("taxable.incomes.1978")
summary(taxable.incomes.1978$taxIncomes)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0    13376    27359    90914    74589 21541002
```

```
bfd.ti <- benford(taxable.incomes.1978$taxIncomes)
plot(bfd.ti)
```



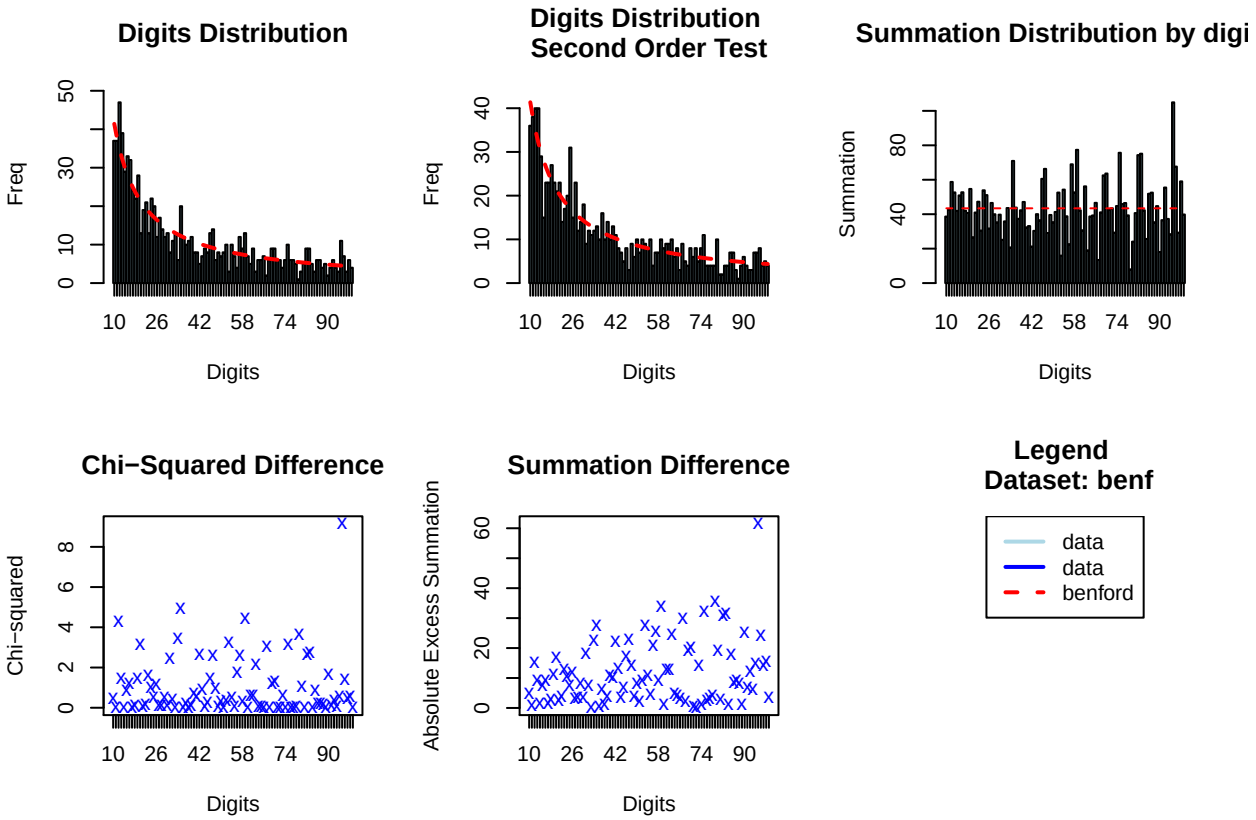

```
bfd.ti
```

```
##
## Benford object:
##
```

```
## Data: taxable.incomes.1978$taxIncomes
## Number of observations used = 150760
## Number of obs. for second order = 86467
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.467
##          Var  0.077
##  Ex.Kurtosis -1.047
##     Skewness  0.198
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     20       1242.50
## 2     19       1194.61
## 3     18       1052.99
## 4     21       1020.14
## 5     17        976.60
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  taxable.incomes.1978$taxIncomes
## X-squared = 5109.9, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  taxable.incomes.1978$taxIncomes
## L2 = 0.013223, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.001684866
## Distortion Factor: -8.130494
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

The digits distribution plots show that the digits from the data follow Benford's Law. But, the chi-squared value is 5109.9 for 89 degrees of freedom, which is much larger than the critical value of 112.022, for 0.05 significance. We generate a random benford distribution and perform Benford's analysis on it to compare and see the difference.

```
library(BenfordTests)
set.seed(99)
benf <- rbenf(1000)
bfd.bf <- benford(benf)
plot(bfd.bf)
```




```
bfd.bf
```

```
##
## Benford object:
##
## Data: benf
## Number of observations used = 1000
## Number of obs. for second order = 999
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.495
##          Var  0.087
##  Ex.Kurtosis -1.269
##     Skewness  0.043
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     12         12.24
## 2     20          8.19
## 3     35          7.77
## 4     13          6.82
## 5     34          6.59
##
```

```
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  benf
## X-squared = 92.882, df = 89, p-value = 0.3682
##
##
##  Mantissa Arc Test
##
## data:  benf
## L2 = 0.0022303, df = 2, p-value = 0.1075
##
## Mean Absolute Deviation: 0.002517013
## Distortion Factor: NaN
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

The chi-squared value is 92.882 for 89 degrees of freedom, which is well within the critical value of 112.022, for 0.05 significance.

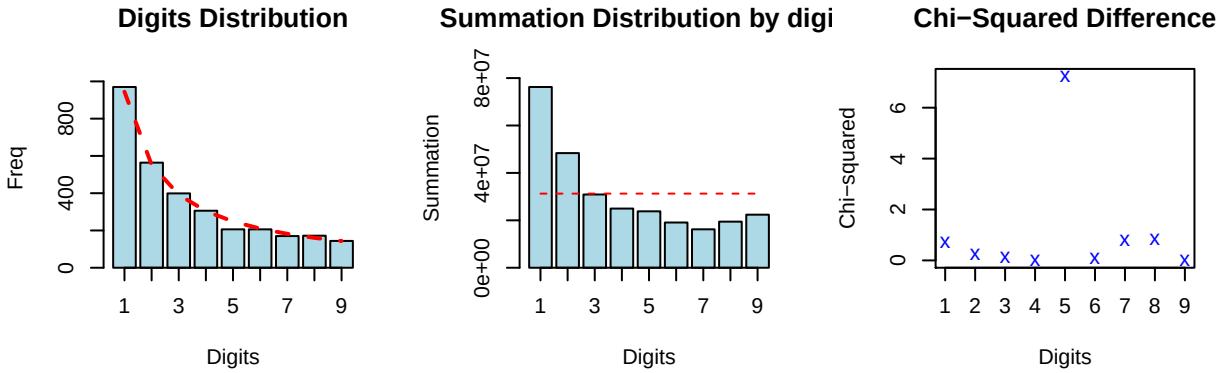
2000 Census Data

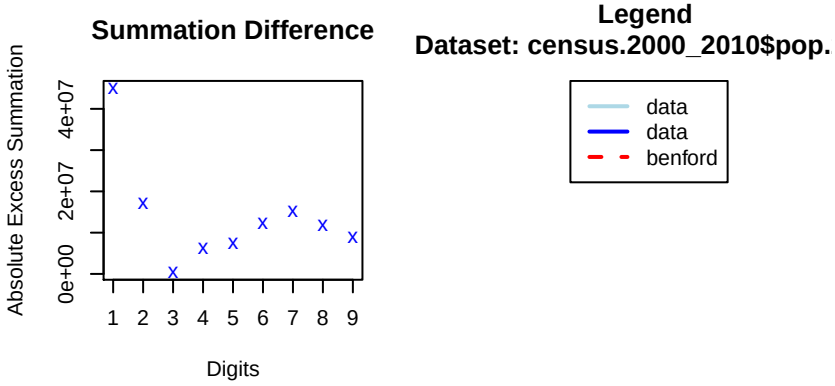
```
library(benford.analysis)
data(census.2000_2010)
```

Benford.Analysis Package

We use benford to find out potential fraud cases in this census data.

```
bfd.cp <- benford(census.2000_2010$pop.2000,1)
plot(bfd.cp, except=c("mantissa", "chi square", "abs diff", "second order"), multiple = T)
```






```
# MAD value
MAD(bfd.cp)
```

```
## [1] 0.004131646
```

```
# top 10 duplicates
library(knitr)
kable(duplicatesTable(bfd.cp)[1:10])
```

number	duplicates
21802	3
10155	3
20826	2
20336	2
19526	2
17774	2
27507	2
1622	2
13184	2
6594	2

It's acceptable that at most three counties have the same population.

The 'suspicious' observations according to Benford's Law:

```
suspects <- getSuspects(bfd.cp, census.2000_2010)
kable(suspects[c(1:10),c(1:4)])
```

fips	name	area	pop.2000
1003	Baldwin County	1589.78	140415
1009	Blount County	644.78	51024
1011	Bullock County	622.80	11714
1015	Calhoun County	605.87	112249
1023	Choctaw County	913.50	15922
1027	Clay County	603.96	14254
1029	Cleburne County	560.10	14123
1033	Colbert County	592.62	54984
1035	Conecuh County	850.16	14089
1037	Coosa County	650.93	12202

The first digits ordered by the mains discrepancies from Benford's Law:

```
kable(suspectsTable(bfd.cp, by="absolute.diff"))
```

digits	absolute.diff
5	42.3915689
1	25.6689036
7	11.9207377
2	11.6017203
8	11.5345371
3	7.0671833
6	4.0120791
4	1.9932892
9	0.4587521

```
#Chi-sqaure test
chisq(bfd.cp)
```

```
##
##  Pearson's Chi-squared test
##
## data:  census.2000_2010$pop.2000
## X-squared = 10.005, df = 8, p-value = 0.2647
```

The p-value is 0.27 so that we cannot reject null hypothesis, which means that the distances between data points and benford points are not significantly different.

BenfordTests Package

```
#JP Sqaure test
jpsq.benftest(x=census.2000_2010$pop.2000,digits = 2, pvalmethod = "simulate", pvalsims = 10000)
```

```
##
##  JP-Square Correlation Statistic Test for Benford Distribution
##
## data:  census.2000_2010$pop.2000
## J_stat_squ = 0.95021, p-value = 0.4635
```

Joenssen's JP-square Test for Benford's Law: The result signifys that the square correlation between signifd(census.2000_2010$pop.2000,2) and pbenf(2) is not zero.

```r
# Euclidean Distance Test for Benford's Law
edist.benftest(census.2000_2010$pop.2000)
```

```
##
##  Euclidean Distance Test for Benford Distribution
##
## data:  census.2000_2010$pop.2000
## d_star = 0.95328, p-value = 0.3489
```

"edist.benftest" takes any numerical vector reduces the sample to the specified number of signif- icant digits and performs a goodness-of-fit test based on the Euclidean distance between the first digits' distribution and Benford's distribution to assert if the data conforms to Benford's law.

The p-value is greater than 0.05 so that we can not reject the null hypothesis. Therefore, the goodness-of-fit test based on the Euclidean distance between the first digits' distribution and Benford's distribution shows the data does conform to Benford's law very well.

Conclusion

Even though all the tests and plots we've done signify that our data follows well the Benford Law, we can't arbitrarily say that there are not frauds in these census observations.