Benford's law analysis about Credit Card Fraud

Xiang XU, Si cHEN, Guangyan YU, Xuan ZHU

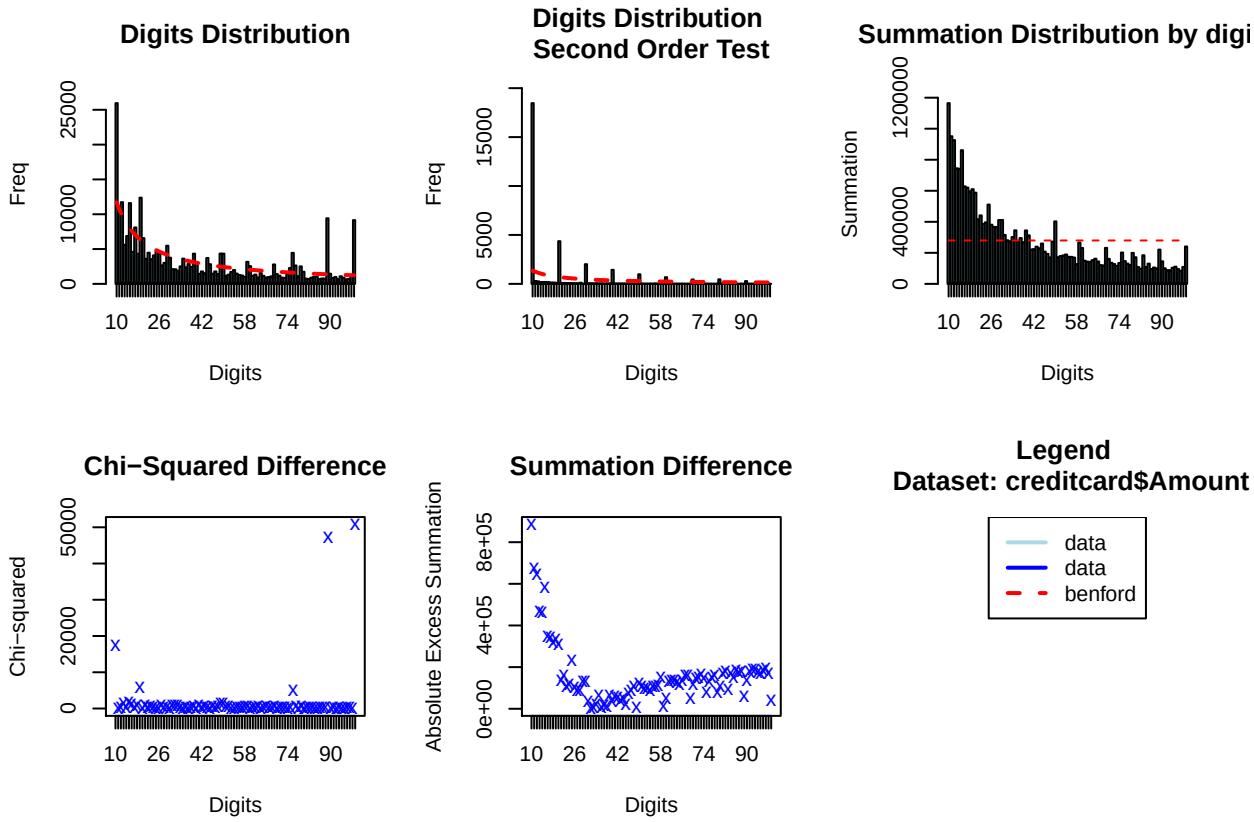
November 26, 2018

```
#read data
creditcard <- read_csv("E:/MSSP/MA615/Content/Class32 Benford's Law/creditcard.csv")
#str(creditcard)
bfd.credit <- benford(creditcard$Amount)
bfd.credit
```

```
##
## Benford object:
##
## Data: creditcard$Amount
## Number of observations used = 282982
## Number of obs. for second order = 32765
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.485
##          Var  0.101
##  Ex.Kurtosis -1.297
##     Skewness  0.058
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     10      14245.62
## 2     89       8050.83
## 3     99       7920.84
## 4     19       6080.18
## 5     15       3673.38
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  creditcard$Amount
## X-squared = 158230, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  creditcard$Amount
## L2 = 0.016216, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.004677013
## Distortion Factor: -11.10134
```

```
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

```
plot(bfd.credit)
```




```
#creates a data frame with the duplicates in decreasing order.
duplicatesTable(bfd.credit)
```

```
##         number duplicates
##      1:   1.00      13688
##      2:   1.98       6044
##      3:   0.89       4872
##      4:   9.99       4747
##      5:  15.00       3280
##     ---
## 32762: 787.95          1
## 32763: 405.09          1
## 32764: 381.05          1
## 32765: 337.54          1
## 32766:  95.63          1
```

```
#gets the Chi-squared test
chisq(bfd.credit)
```

```
##
##  Pearson's Chi-squared test
##
## data:  creditcard$Amount
```

```
## X-squared = 158230, df = 89, p-value < 2.2e-16
```

```
chisq.benftest(creditcard$Amount)
```

```
##
##  Chi-Square Test for Benford Distribution
##
## data:  creditcard$Amount
## chisq = 10858, p-value < 2.2e-16
```

```
#gets the Distortion Factor of a Benford object.
dfactor(bfd.credit)
```

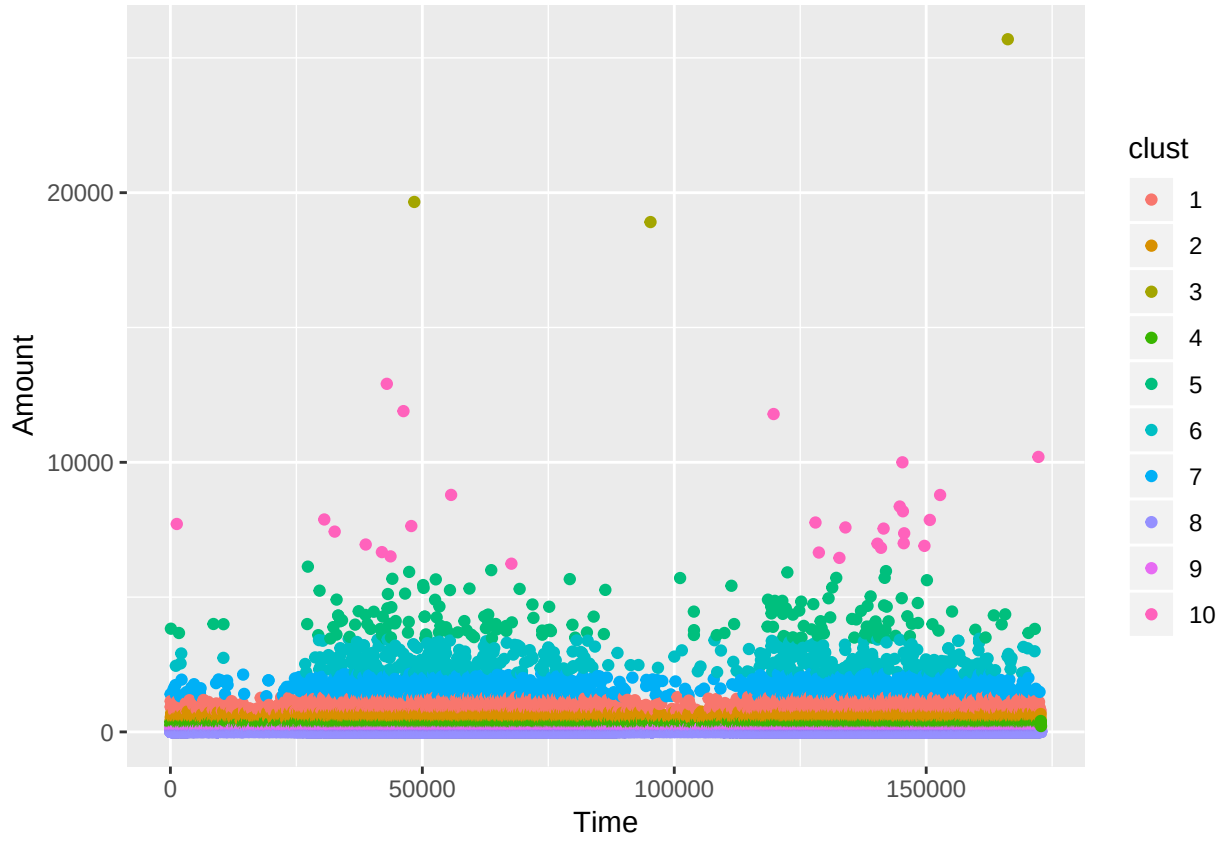
```
## [1] -11.10134
```

```
suspects <- getSuspects(bfd.credit, creditcard)
kable(head(suspects[,c(1:4,28:31)]))
```

Time	V1	V2	V3	V27	V28	Amount	Class
13	-0.4369051	0.9189662	0.9245908	0.0796924	0.1310238	0.89	0
22	-1.9465251	-0.0449005	-0.4055701	0.7075188	0.0145998	0.89	0
53	-1.1989677	-1.4741005	1.8403260	-1.1234565	-0.7343511	89.17	0
56	0.0869962	-0.0519887	-0.4978824	-0.3949556	-0.4359956	10.84	0
60	1.0691414	0.0437859	0.3098673	-0.0013118	0.0358115	89.40	0
73	1.1622809	1.2481782	-1.5813171	0.0567468	0.0847065	1.00	0

```
#kmeans cluster
credit_cluster <- kmeans(creditcard$Amount, 10, 10000)
creditcard %<>% mutate(clust = as.factor(credit_cluster$cluster))

ggplot(creditcard) +   aes(x= Time, y= Amount, color = clust ) + geom_point()
```



From chisq result, the p-value is so small so we have to reject the null hypothesis. So we can know tour credit card data does not follow benford's law. And among about 285 thousands observations, we have 35 thousands suspects, aka means there are about 35 thousands potential credit card fraud.