Benford Analysis on GDP in real 2011 US dollars

Albert Ding, Yifeng Luo, Kaiyu Yan, Wenjia Xie

November 29, 2018

Introduction of the Data Sets

We are performing our analysis on a dataset from the University of Groningen in the Netherlands compiled as part of The Maddison Project.

The Maddison Project was initiated in 2010 to measure economic performance for different regions, time periods, and subtopics. The database presents annual GDP estimates for every country in the world going back as early as the 1800s.

Benford analysis on GDP in real 2011 US$M

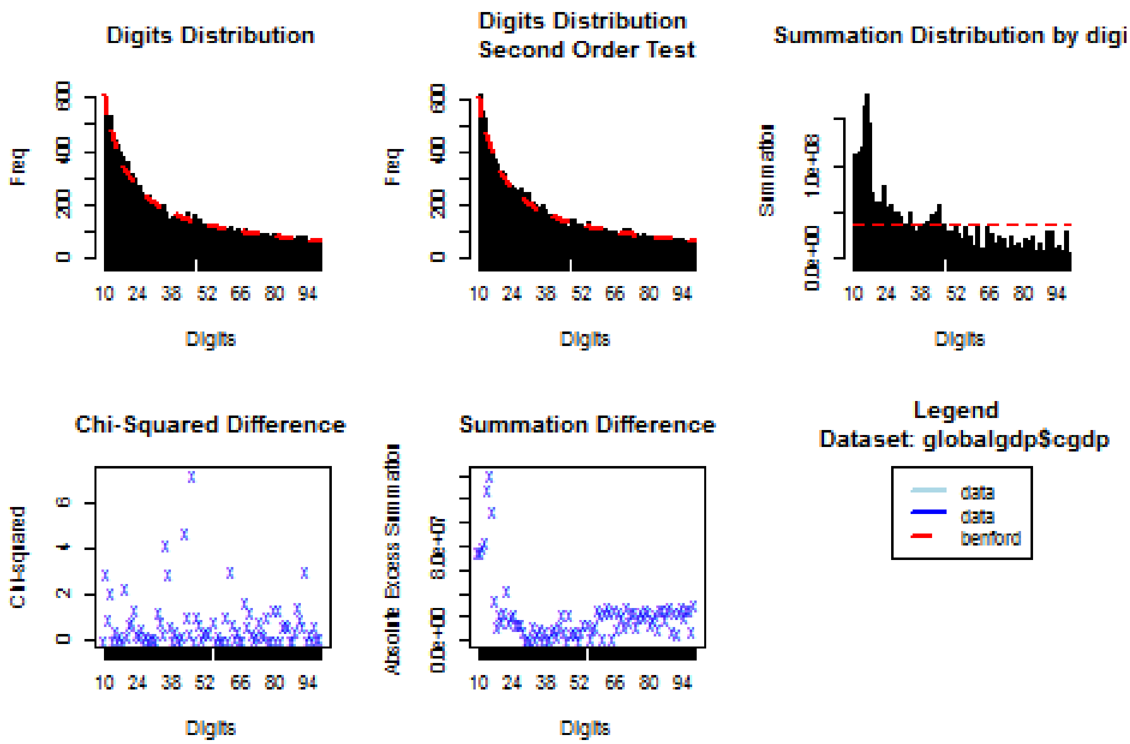
Below we read in and clean the data. The attribute of interest we are examining is the annual GDP of each country by year. In order to get this data, we take the GDP per capita normalized in real 2011 purchasing power terms and multiply by the population in thousands. We divide this number by 1000 again in order get the figure in $USD in millions.

```
globalgdp <- read_csv("Madison_GDP.csv") %>%
    filter(cgdppc != "" & pop != "NA") %>%
    mutate(cgdp = cgdppc/1000*pop)
```

```
## Warning: Missing column names filled in: 'X9' [9], 'X10' [10]
```

```
## Parsed with column specification:
## cols(
##   countrycode = col_character(),
##   country = col_character(),
##   year = col_integer(),
##   cgdppc = col_integer(),
##   rgdpnapc = col_integer(),
##   pop = col_integer(),
##   i_cig = col_character(),
##   i_bm = col_character(),
##   X9 = col_character(),
##   X10 = col_character()
## )
```

```
bfd.gdp <- benford(globalgdp$cgdp)
plot(bfd.gdp)
```




```
bfd.gdp
```

```
##
## Benford object:
##
## Data: globalgdp$cgdp
## Number of observations used = 14729
## Number of obs. for second order = 14716
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic   Value
##         Mean  0.5018
##          Var  0.0834
##  Ex.Kurtosis -1.2033
##     Skewness  0.0016
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     11         40.59
## 2     13         32.05
## 3     47         31.33
## 4     19         27.89
## 5     36         27.26
##
## Stats:
```

```
##
##  Pearson's Chi-squared test
##
## data:  globalgdp$cgdp
## X-squared = 69.086, df = 89, p-value = 0.9418
##
##
##  Mantissa Arc Test
##
## data:  globalgdp$cgdp
## L2 = 4.782e-06, df = 2, p-value = 0.932
##
## Mean Absolute Deviation: 0.000586757
## Distortion Factor: 0.426057
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

From the first plot, we can see that the original GDP data is in blue and the expected frequency according to Benford's law is in red. In our example, the first plot shows that the data do have a tendency to follow Benford's law, and there is only a subtle difference for some digits.

This result can also be verified by Chi-squared difference test. The calculated chi-square statistic here is 69.086 and the p-value of the test is 0.932,which indicate that We do not have sufficient evidence to reject the null hypothesis of conformity to Benford's Law.

For the Digits Distribution Second Order Test plot, there is also no obvious discrepency between the original data and the expected frequency according to Benford's law, which can alao be an evidence that there is no obvious detected errors in data downloads, rounded data, data generated by statistical procedures, and the inaccurate ordering of data.

```r
suspects_ranked <- suspectsTable(bfd.gdp)
suspects1 <- getSuspects(bfd.gdp, globalgdp, by='absolute.diff', how.many=5)
suspects1
```

```
##       countrycode     country year cgdppc rgdpnapc   pop        i_cig i_bm
##    1:         AFG Afghanistan 1950   2392     2392  8150 Extrapolated <NA>
##    2:         AFG Afghanistan 1986   2779     2779 13126 Extrapolated <NA>
##    3:         AFG Afghanistan 1997    926      926 20769 Extrapolated <NA>
##    4:         AFG Afghanistan 2009   1669     1669 28484 Extrapolated <NA>
##    5:         AGO      Angola 1975   3246     5345  5885 Extrapolated <NA>
##   ---
## 1624:         ZWE    Zimbabwe 1963   2538     1484  4412 Extrapolated <NA>
## 1625:         ZWE    Zimbabwe 1964   2525     1570  4537 Extrapolated <NA>
## 1626:         ZWE    Zimbabwe 1967   2760     1670  4995 Extrapolated <NA>
## 1627:         ZWE    Zimbabwe 1970   3448     2112  5515 Extrapolated <NA>
## 1628:         ZWE    Zimbabwe 2005   1660     1510 11639 Interpolated <NA>
##         X9  X10     cgdp
##    1: <NA> <NA> 19494.80
##    2: <NA> <NA> 36477.15
##    3: <NA> <NA> 19232.09
##    4: <NA> <NA> 47539.80
##    5: <NA> <NA> 19102.71
##   ---
## 1624: <NA> <NA> 11197.66
## 1625: <NA> <NA> 11455.92
## 1626: <NA> <NA> 13786.20
```

```
## 1627: <NA> <NA> 19015.72
## 1628: <NA> <NA> 19320.74
```

```
#Twodigitsuspect1 <- suspects %>%mutate(cgdptwo = substr(cgdp,1,2))
#unique(Twodigitsuspect1$cgdptwo)
#chisq(bfd.gdp)

# suspects2 <- getSuspects(bfd.gdp, globalgdp, by='difference', how.many=5)
# suspects2
# Twodigitsuspect2 <- suspects %>%
#   mutate(cgdptwo = substr(cgdp,1,2))
#
# suspects3 <- getSuspects(bfd.gdp, globalgdp, by='squared.diff', how.many=5)
# suspects3
# Twodigitsuspect3 <- suspects %>%
#   mutate(cgdptwo = substr(cgdp,1,2))
#
# suspects4 <- getSuspects(bfd.gdp, globalgdp, by='absolute.diff', how.many=5)
# suspects4
# Twodigitsuspect4 <- suspects %>%
#   mutate(cgdptwo = substr(cgdp,1,2))

# unique(Twodigitsuspect2$cgdptwo)
# unique(Twodigitsuspect3$cgdptwo)
# unique(Twodigitsuspect4$cgdptwo)
```

The first two digit come from the suspect result just align to the "suspects_ranked". In other words, the suspects list generates all the country years that start with the two digit figures of the first n number of chi square differences that appear to suggest some abnormality in frequency, which is bizzare because it does not generate the two with the biggest chi square numbers. In commented code we have confirmed this method of the function and also compared the suspects generated by all four methods, concluding that the four methods actually generate the same suspects.