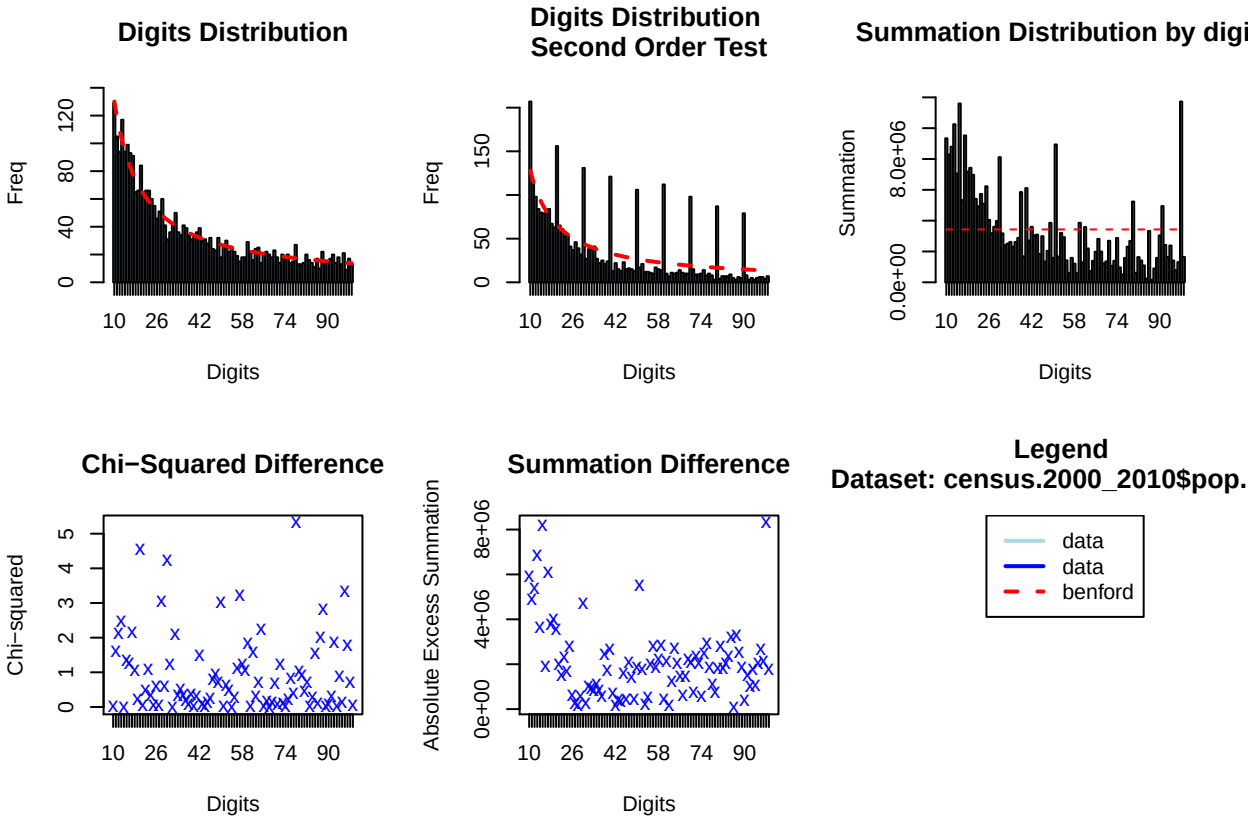
Benford

Angela Zhai, Zhaobin Liu, Yaotang Luo, Jiahao Xu

November 30, 2018

I. US 2010 Population Census Data



The original data is in blue and the expected frequency according to Benford's law is in red.

Benford's analysis of the first digits indicate the data basically follows Benford's Law.

Digit Distribution Second Order Test calculates the digit frequencies of the differences between the ordered (ranked) values in a data set. It shows that this dataset generally follows Benford's law, except some specific two digits.

The Chi-Square and Summation Difference plots almost fit Benford's law, but not good enough.

```
##
## Benford object:
##
## Data: census.2000_2010$pop.2010
## Number of observations used = 3143
## Number of obs. for second order = 3090
## First digits analysed = 2
##
## Mantissa:
##
```

```
##    Statistic  Value
##         Mean  0.496
##          Var  0.083
##  Ex.Kurtosis -1.191
##     Skewness  0.052
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     20         17.40
## 2     13         15.84
## 3     12         15.26
## 4     11         13.77
## 5     30         13.76
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  census.2000_2010$pop.2010
## X-squared = 82.809, df = 89, p-value = 0.6646
##
##
##  Mantissa Arc Test
##
## data:  census.2000_2010$pop.2010
## L2 = 0.00076042, df = 2, p-value = 0.09163
##
## Mean Absolute Deviation: 0.001428017
## Distortion Factor: -0.7805108
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 20.

From the log mantissa of the data, we can tell that the data follows Benford's Law. Because Mean closes to 0.5, Variance closes to 0.083, Ex. Kurtosis closes to -1.2, and Skewness closes to 0.

Degree of freedom equals 89 and p-value equals 0.6646, so failed to reject Benford's law. X-squared value equals 82.809 and closes to the value of degree of freedom, the distribution of this data set looks good. All in all, this dataset follows Benford's law.

The distortion factor is -0.781.

digits	data.dist	data.second.order.dist	benford.dist	data.second.order.dist.freq	data.dist.freq
10	0.0410436	0.0669903	0.0413927	207	129
11	0.0334076	0.0372168	0.0377886	115	105
12	0.0299077	0.0317152	0.0347621	98	94
13	0.0372256	0.0271845	0.0321847	84	117
14	0.0299077	0.0258900	0.0299632	80	94
15	0.0314986	0.0255663	0.0280287	79	99
16	0.0295896	0.0258900	0.0263289	80	93
17	0.0289532	0.0271845	0.0248236	84	91
18	0.0206809	0.0216828	0.0234811	67	65
19	0.0209990	0.0203883	0.0222764	63	66

Table above shows the distribution of population data by first two digits.

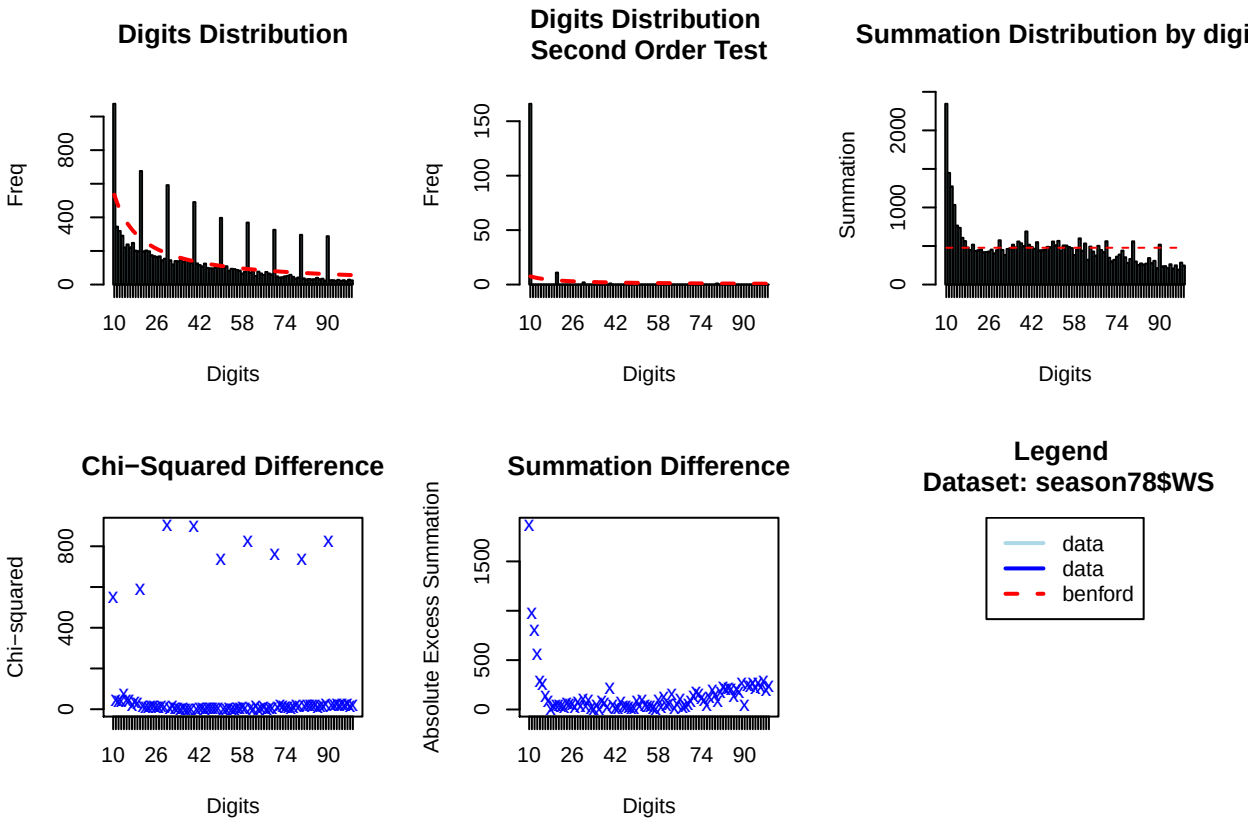
digits	absolute.diff
20	17.402033
13	15.843540
12	15.257300
11	13.769447
30	13.757700
17	12.979476
28	12.100785
15	10.905722
16	10.248146
78	9.611388

Above table shows ten suspected two digits that contain most discrepancies from Benford's law.

fips	name	area	pop.2010
1013	Butler County	776.83	20947
1023	Choctaw County	913.50	13859
1027	Clay County	603.96	13932
1035	Conecuh County	850.16	13228
1041	Crenshaw County	608.84	13906
1119	Sumter County	903.89	13763
2130	Ketchikan Gateway Borough	4858.41	13477
2150	Kodiak Island Borough	6549.58	13592
4003	Cochise County	6165.69	131346
4005	Coconino County	18618.89	134421

Table above gets the 'suspicious' observations according to Benford's Law. (Only shows 10 observations)

II. NBA 2017-2018 Season Win Share



The first plot shows that our data has the tendency to follow Benford's law. The chi-Squared Difference graph also has a very good-fit shape.

```
##
## Benford object:
##
## Data: season78$WS
## Number of observations used = 12926
## Number of obs. for second order = 182
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.494
##          Var  0.084
##  Ex.Kurtosis -1.099
##     Skewness -0.167
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     10        540.96
## 2     30        407.93
## 3     20        402.11
## 4     40        352.38
```

```
## 5     50        285.83
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  season78$WS
## X-squared = 7807.9, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  season78$WS
## L2 = 0.007222, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.005101497
## Distortion Factor: -68.58
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

By the result, we can see the # of observation, # of obs for second order, and first digit analysed. The 5 largest deviations show that the smallest digit are 10 which follows the Benford's law as well.

digits	absolute.diff
10	540.9582
30	407.9281
20	402.1071
40	352.3833
50	285.8342
60	276.2096
70	246.3718
80	226.2638
90	225.9696
14	166.3046

Above table shows the digits by decreasing order of discrepancies, and gets observations of the 2 most suspicious groups.

Season	Player	WS
1979	Alonzo Bradley	0.1
1979	Billy McKinney	3.0
1979	Brian Taylor	0.1
1979	Charlie Criss	0.1
1979	E.C. Coleman	0.1
1979	Elmore Smith	0.3
1979	Essie Hollis	0.1
1979	Glenn Mosley	0.3
1979	Greg Bunch	0.1
1979	Harry Davis	0.3

Table above gets the 'suspicious' observations according to Benford's Law. (Only shows 10 observations)

number	duplicates
0.1	663
0.2	504
0.3	445
0.4	354
0.5	317
0.6	299
0.7	273
0.9	256
0.8	251
1.1	241

Above table shows the duplicates by decreasing order. There are 1167 observations with the most 2 duplicates values.