Final Project

Fionnuala McPeake

December 13, 2018

Abstract

This study looks into the possibility of falsely reported data regarding health in the Center for Disease Control and Prevention's (CDC) 500 Cities project, which recorded the diseease and health behavior prevalences in the 500 largest cities in the US. This was done using a Benford analysis. While the data did not follow Benford's Law, evidence of fraud was not found.

Introduction

In order to better understand the health conditions of small areas, as well as the behavioral health risks and preventative measures being taken by the inhabitants, the CDC conducted the 500 Cities project. This project surveyed the residents of the 500 most populated cities in the US, and recorded their geographical location down to the census track they inhabit. This is the first project of its type, and the partners of this project decided to conduct it on such a granular scale in the hopes that it will provide information to public health workers and local government such that they can incorporate new policies to create healthier citizens. The survey consisted of 13 chronic diseases, 5 unhealthy behaviors, and 9 preventative practices, and provides both the crude and adjusted prevalence rate for each question asked. As this is a self reported survey about something so private as health, people may be prone to lie. While it may be harder to lie about something as binary as if you have a disease or not, people may be tempted to exaggerate the amount of exercise they do in a week, or lessen their drinking habits to conform with what is socially acceptable. In order to assess the accuracy of this data, a Benford analysis was run on the variables, and deviations were investigated.

Results

The results of the Benford analyses ran on the variables included in this study did not conform to the expected law. All of the results fell into one of two categories: either the majority of cases fell on the ends of the number line, or the cases fell around some mean, decreasing around it in a semi-symmetrical fashion.

The variables falling into the first category are the populations of the cities, the prevelance of diabetes, and the prevelance of adults who have had at least 14 consecutive bad mental health days in the past year. A detailed investigation into the results of the population Benford analysis can be found below. Generally, variables falling into this category have the majority of their instances close to a number that is devisable by 10, with the rest of the instances having a leading number plus or minus approximatly three digits in comparison to the mean. This leaves few numbers with a leading digit between 4 and 7, which is contrary to Benfords law.

Variables falling into the second category follow the same pattern as variables falling into the first category, except their mean is not close to a number divisible by 10, which does not result in a plot with the majority of instances at either end of the number line. Also, the spread around the mean can be rather small, not following the general rule above of being within three digits of the mean.

Both the adjusted and crude prevelence of the diseases and behaviors have been provided, and all of them have been analyzed using a Benford analysis. For nearly all variables, the mean of the crude and adjusted prevelences was slightly different, as well as the minimum and maximum value, leading to a narrower range for the adjusted numbers. However, the difference between two graphs for the same variable were very similar. For this reason, the results of the Benford analyses run on the adjusted numbers can be found in the appendix rather than the main report, as the crude numbers make for a slightly more interesting analysis.

Category One

Below are the three variables that fell into category one- in which the first digits gather at the ends of the number line.

Population

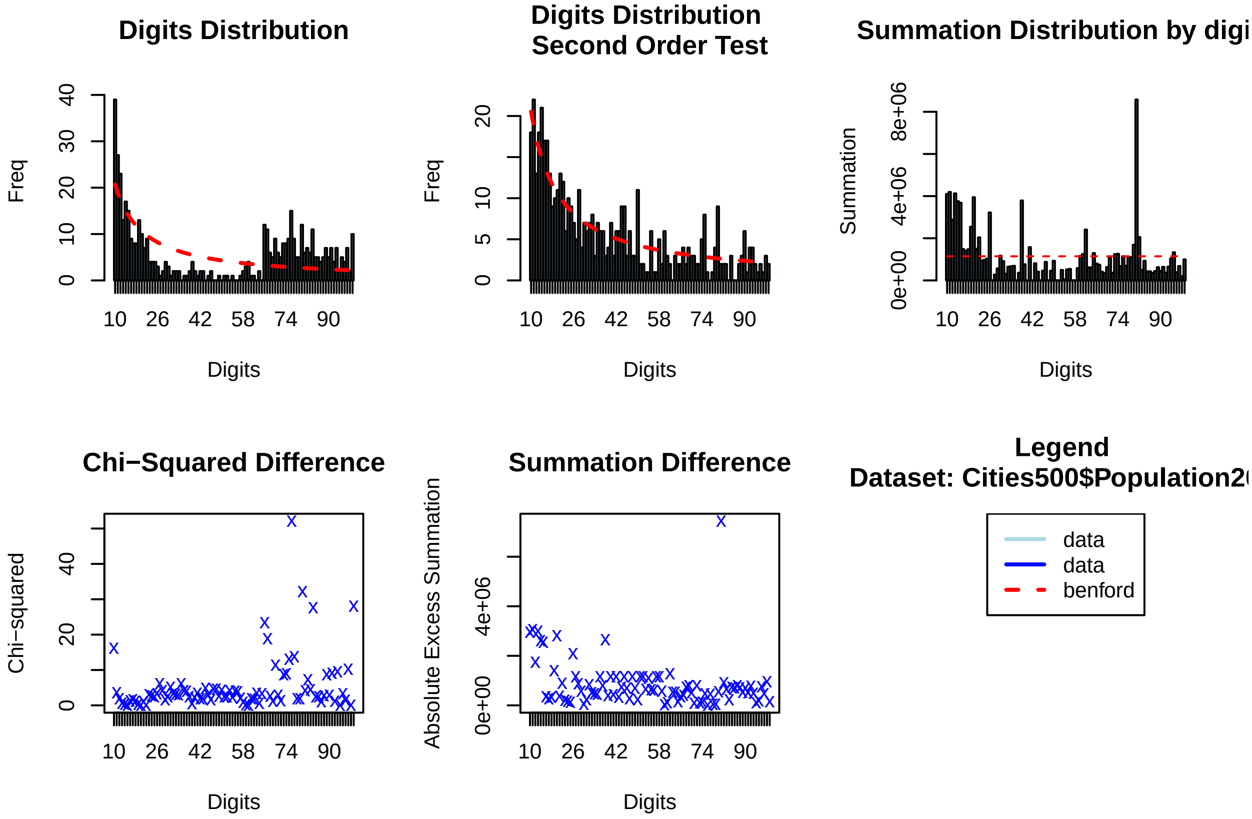
The population of the cities was the variable that was most expected to follow the Benford distribution. However, the results did not conform as expected. To investigate this, the characteristics of the populations included in the data were explored. Only 9 cities have a distribution above 1 million, with most of them having a leading digit of 1. As this is a small number of observations, it would not have great influence over the results.

As there are more instances of cities that have a leading digit between 7 and 9, we investigate this further. Three cities have a population in the 800,000s, and two cities have populations in the 900,000. This again is a very small proportion of the data, and indicates that cities with populations in the hundred-thousands do not account for the deviance. This is confirmed by running a Benford analysis on the cities containing over 100,000 people, which conforms fairly well. A Benford analysis on cities with less than 100,000 residents has the majority of data points for the Digits Distribution plot above 60. When put into context of this project, this is logical. Only the 500 largest cities are included in this data set- any city with a small leading digit that was not of a large order of magnitude would be excluded from the study by design.

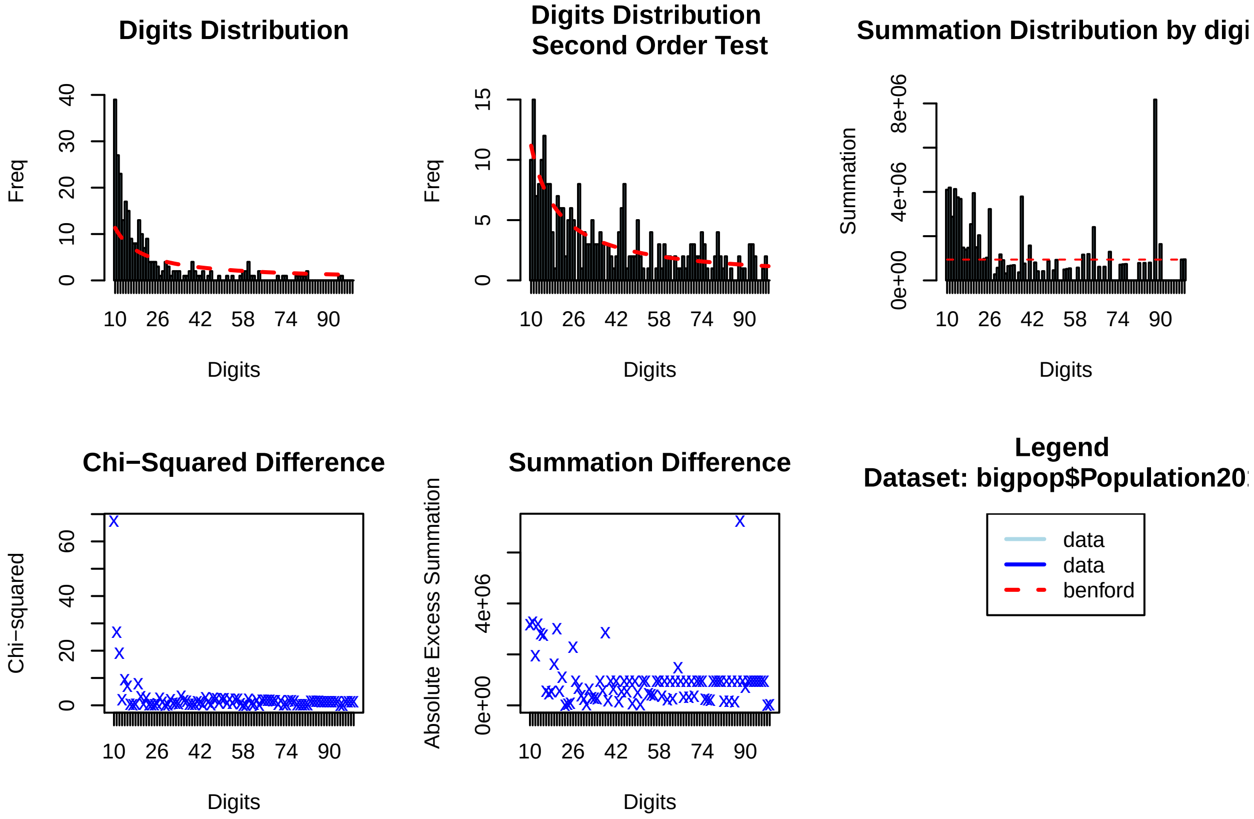
```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v ggplot2 3.0.0     v forcats 0.3.0
```

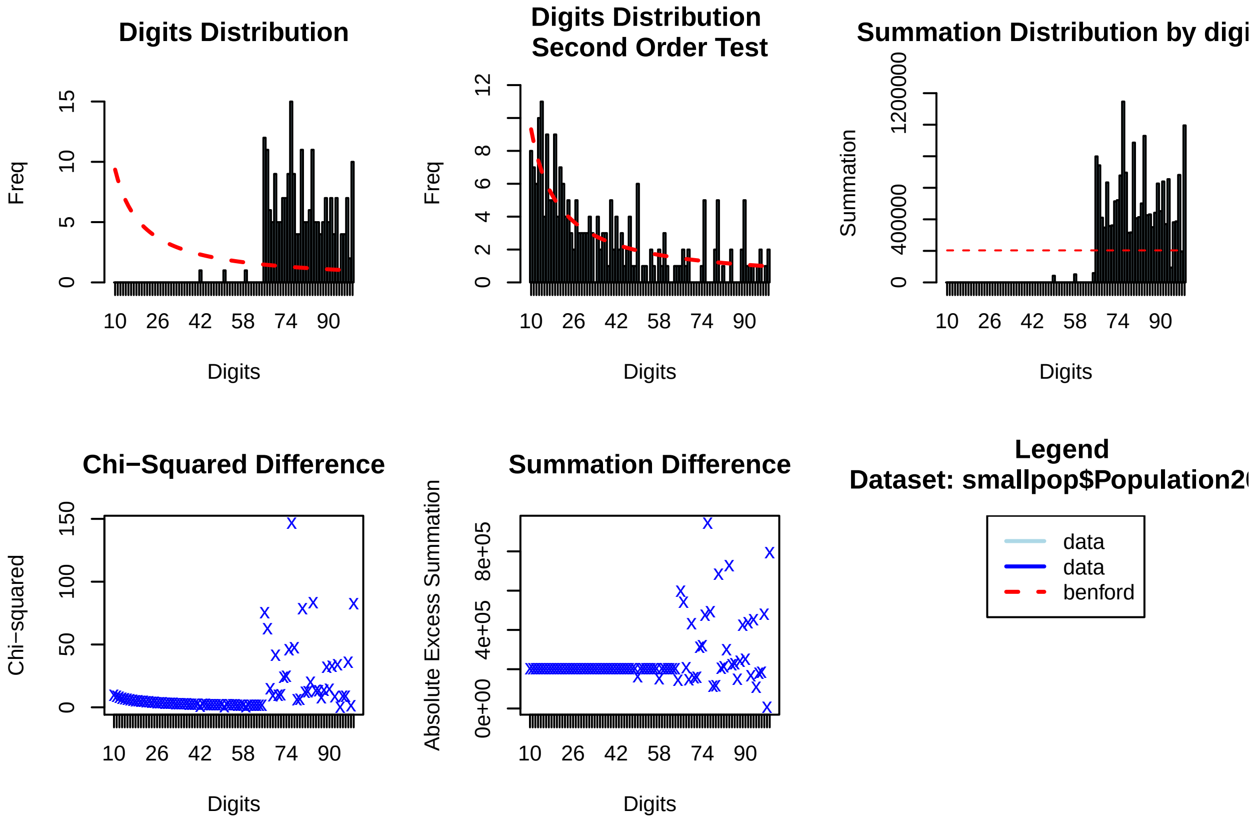
```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_confl
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```



Benford analysis that included all population points. This clearly does not follow Benford's Law, as the points gather at the ends of the number line.



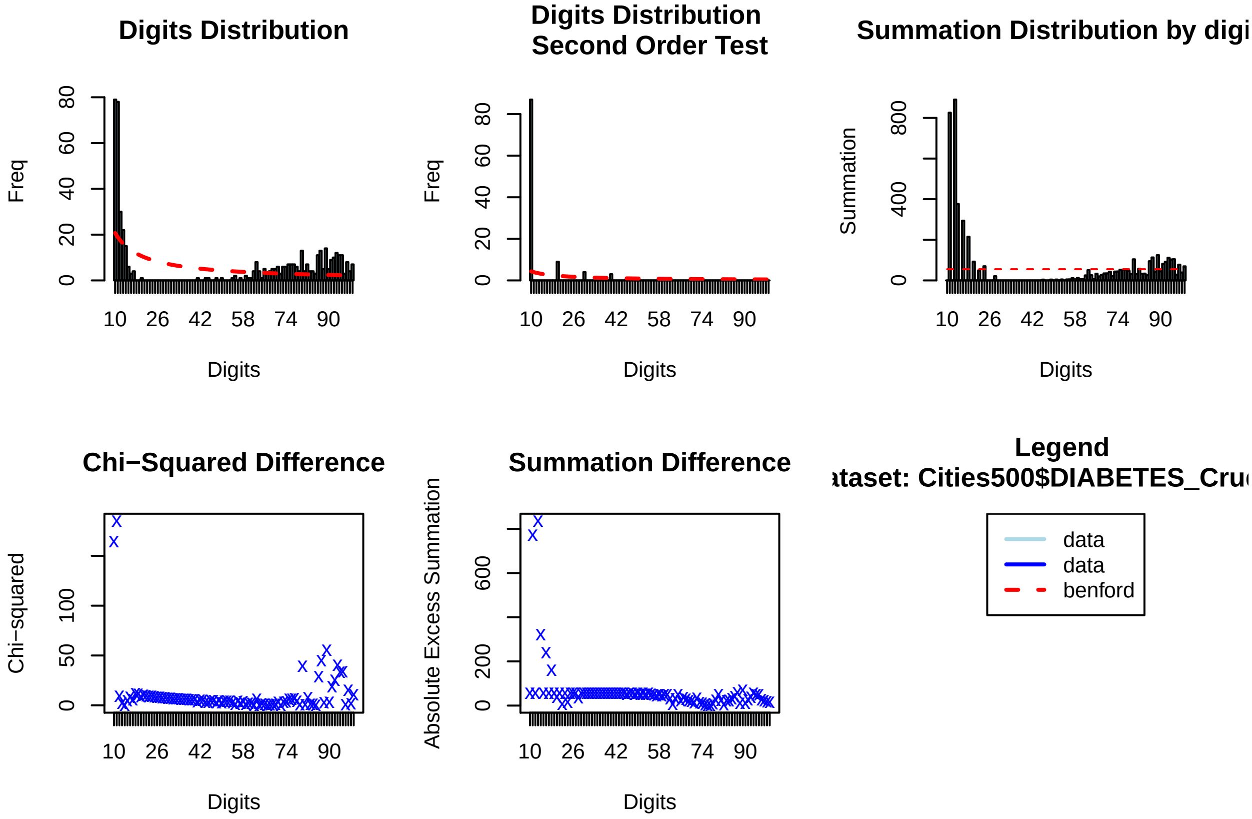
Benford analysis for cities with a population of 100,000 or more. This fits the Benford analysis much better than for the full data.



Benford analysis for cities with a population less than 100,000. It seems that smaller cities do not follow Benford's Law as well as larger cities, and that this is the source of deviation from the Law for the overall population.

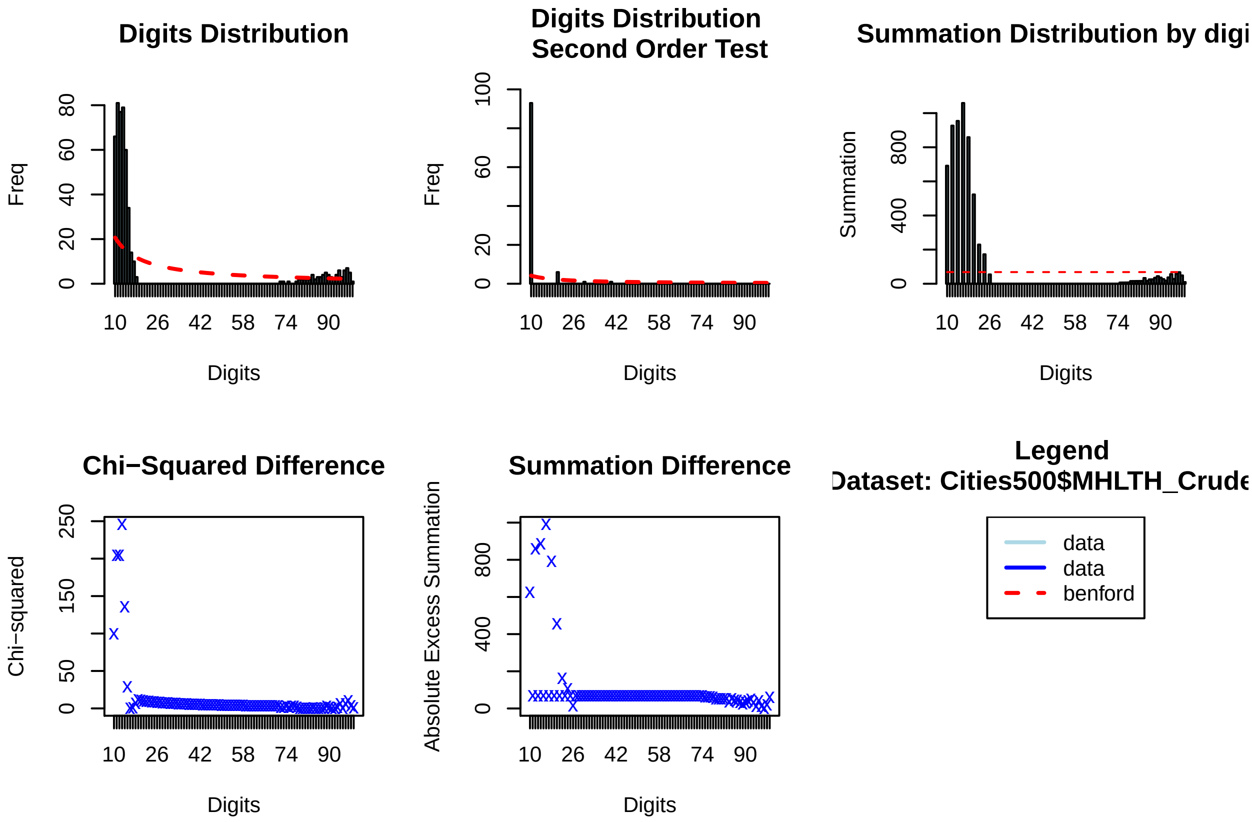
Diabetes

Here there are 237 cities that have a rate between 10 and 19, and there are 251 cities that have a rate between 6 and 9.9. This accounts for the accumilation of instances on the ends of the number line.



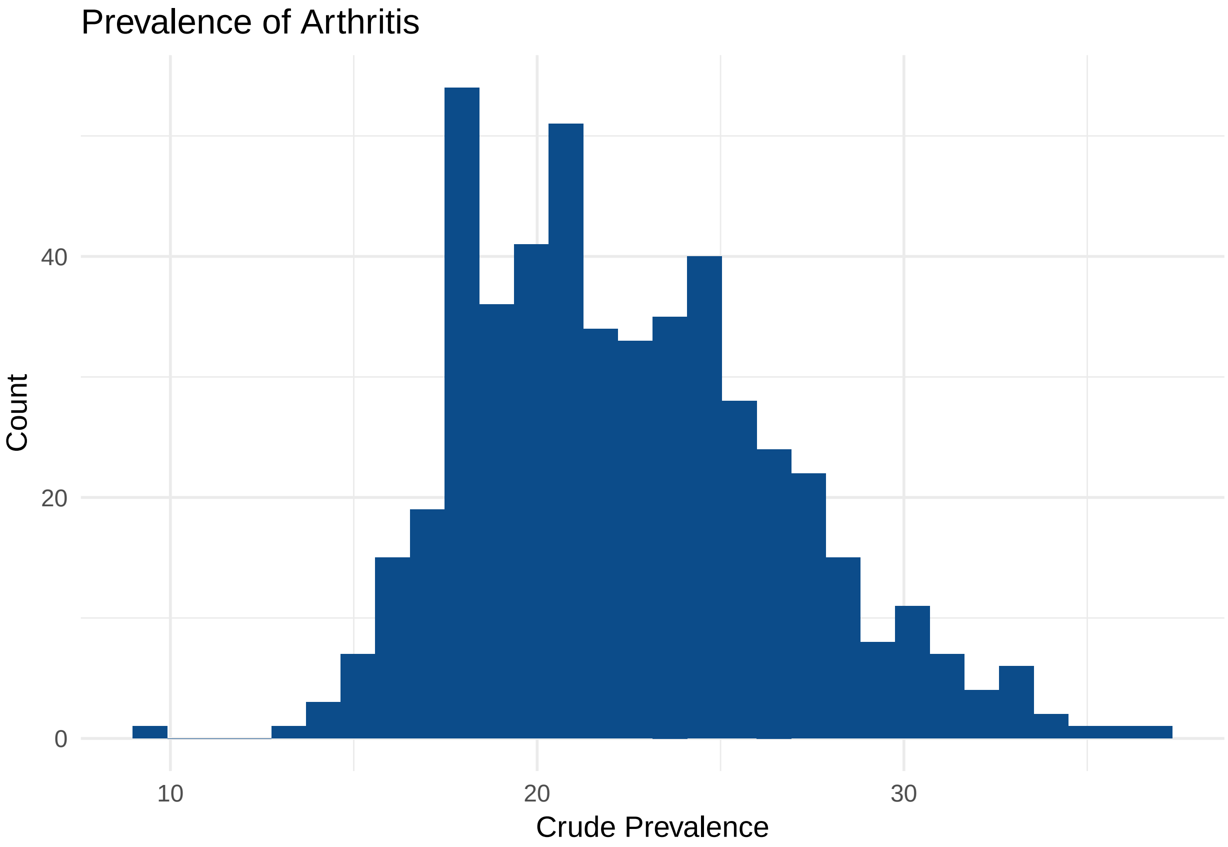
Bad mental health for at least 2 weeks

For this variable, there are 424 cities, the vast majority, with a rate between 10 and 19.9. Only 76 cities have a prevelence between 7 and 10 (exclusive).



Category Two

For the majority of variables, the distribution of the prevelences was semi-symmetric. This was shown using arthritis as an example:

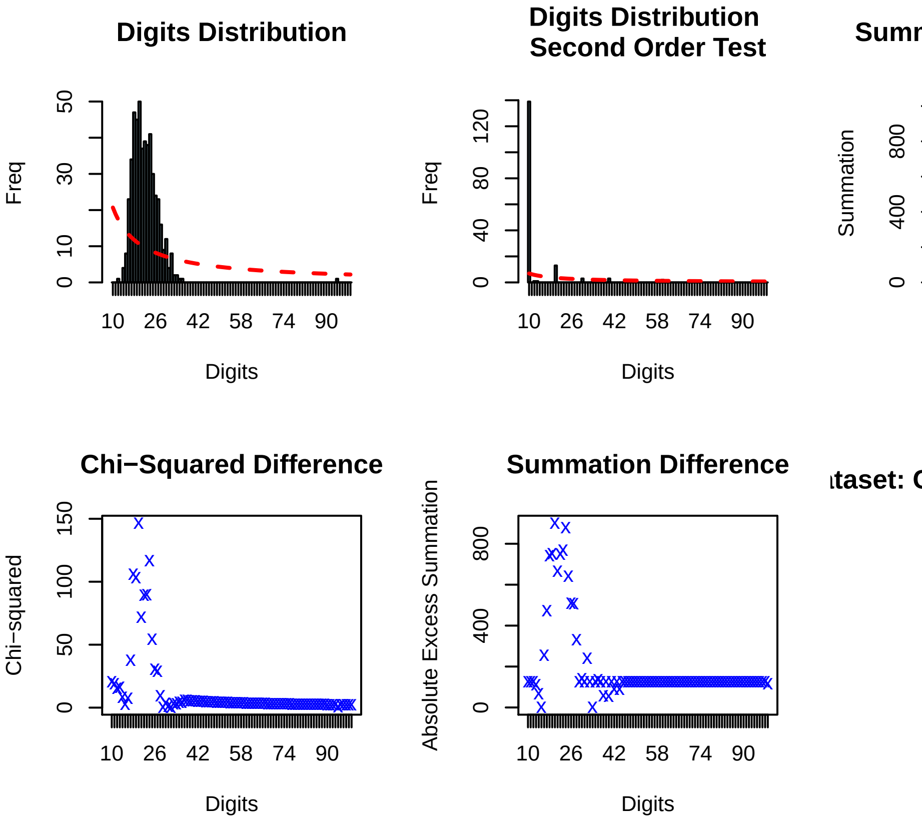


This shows that the first digits have a semi-symetrical distribution around a particular number.

The arthritis results are explored in more detail below. As all of the variables falling into "category two" fallow the same pattern explained for arthritis, they have been moved to the appendix.

Arthritis

Here the maximum prevalence is 36.8, the mean is 22.42, and the minimum value is 9.4. Further investigation shows that 337 cities have a prevelance between 20 and 37, and 159 cities have a prevalence between 10 and 19.9,



explaining the results of the Benford analysis.

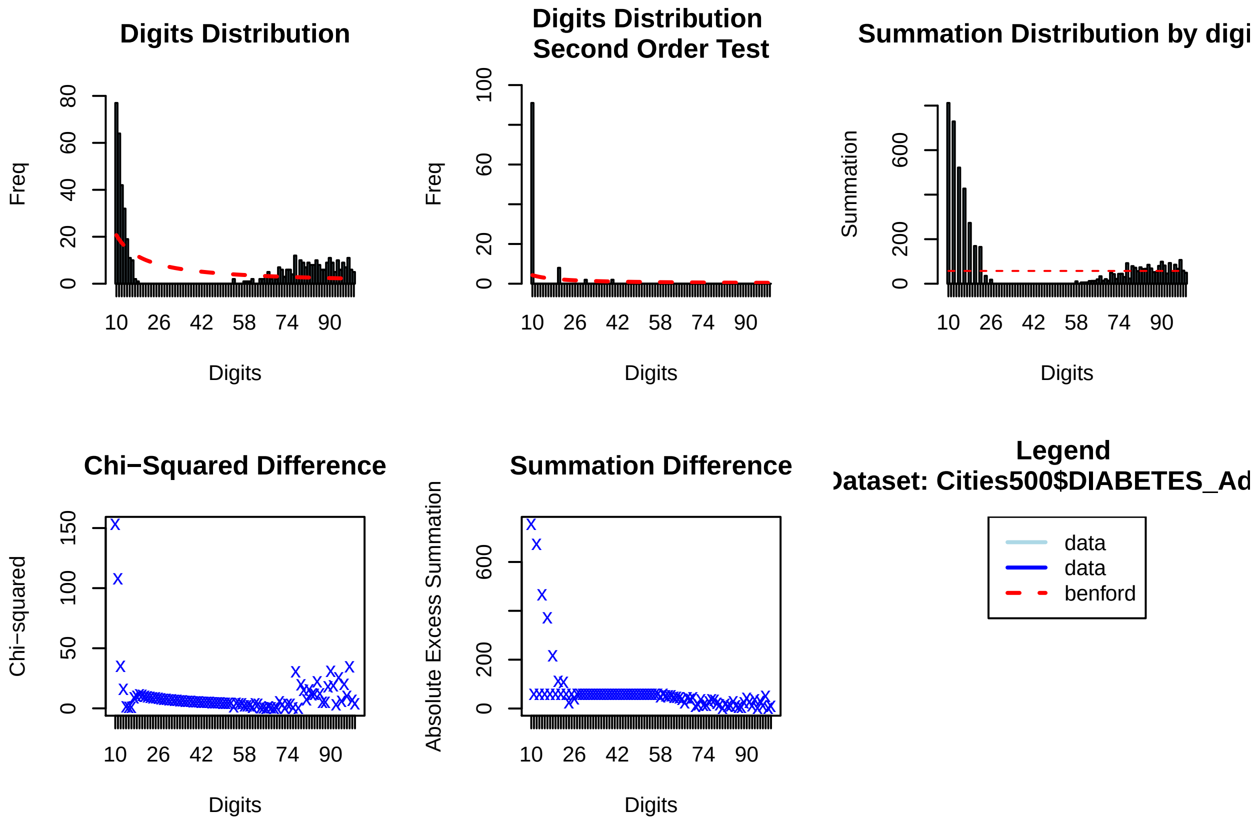
Discussion

Although the data of this study does not follow Benford's Law, it is not suprising in this situation, and does not indicate fraud. Although the study looks at the prevelance of disease in small areas, all of the participants in the study live in the United States, and while there are variances throughout states and regions that could influence health and behavior, there is a general level of healthcare and risk throughout the country, and the variance of diseases is not as large accross the country as it would be if the study looked accross countries of different privileges. Furthermore, because the study looked at larger cities rather than rural areas, it can be reasonably assumed that there are similar resources available. Because the variance between cities is relatively small, they have prevalences within a small range, leading to consecutive leading numbers. This is seen repeatidly in the results of the Benford analysis, and is consistend with the findings of the investigation into the data.

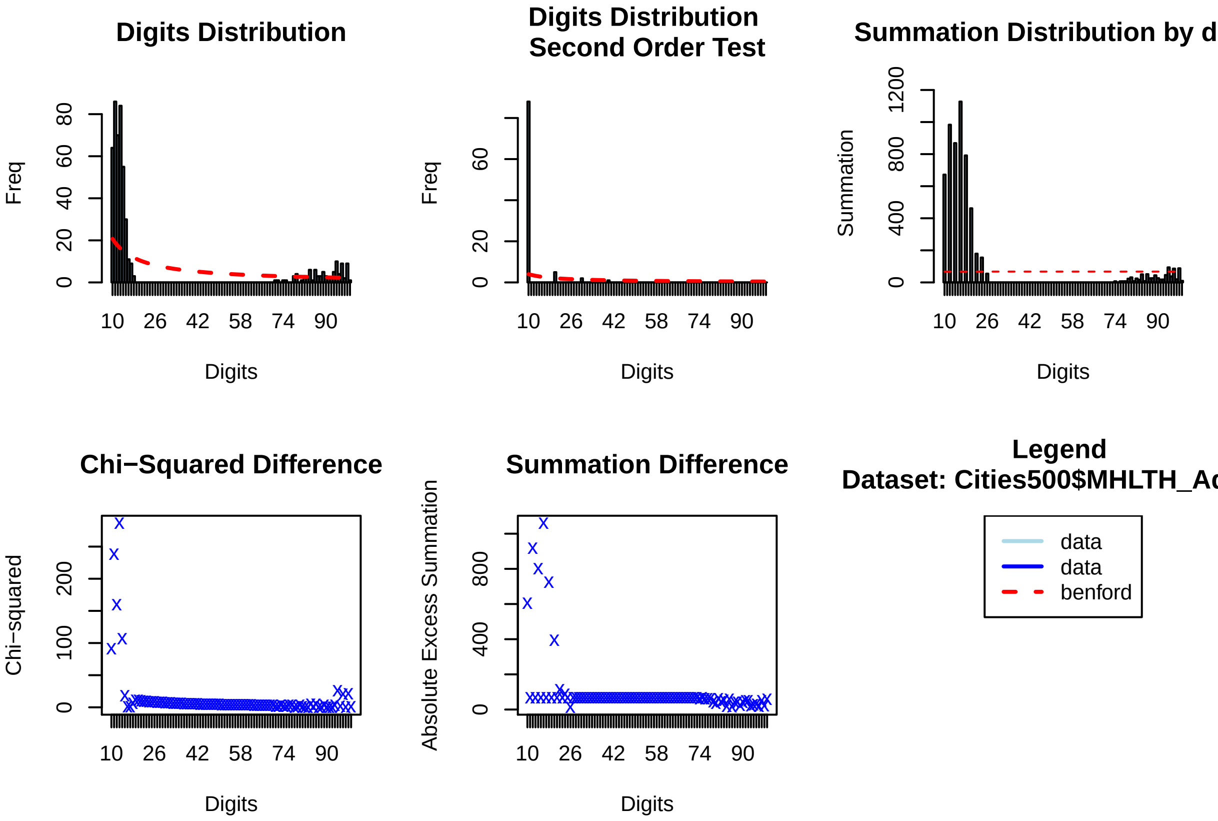
If one wanted to look into the possibility of fraud of this health data, they would likely have to verify the results of this project with other studies performed.

Appendix

Adjusted Prevalences for Variables included in Results

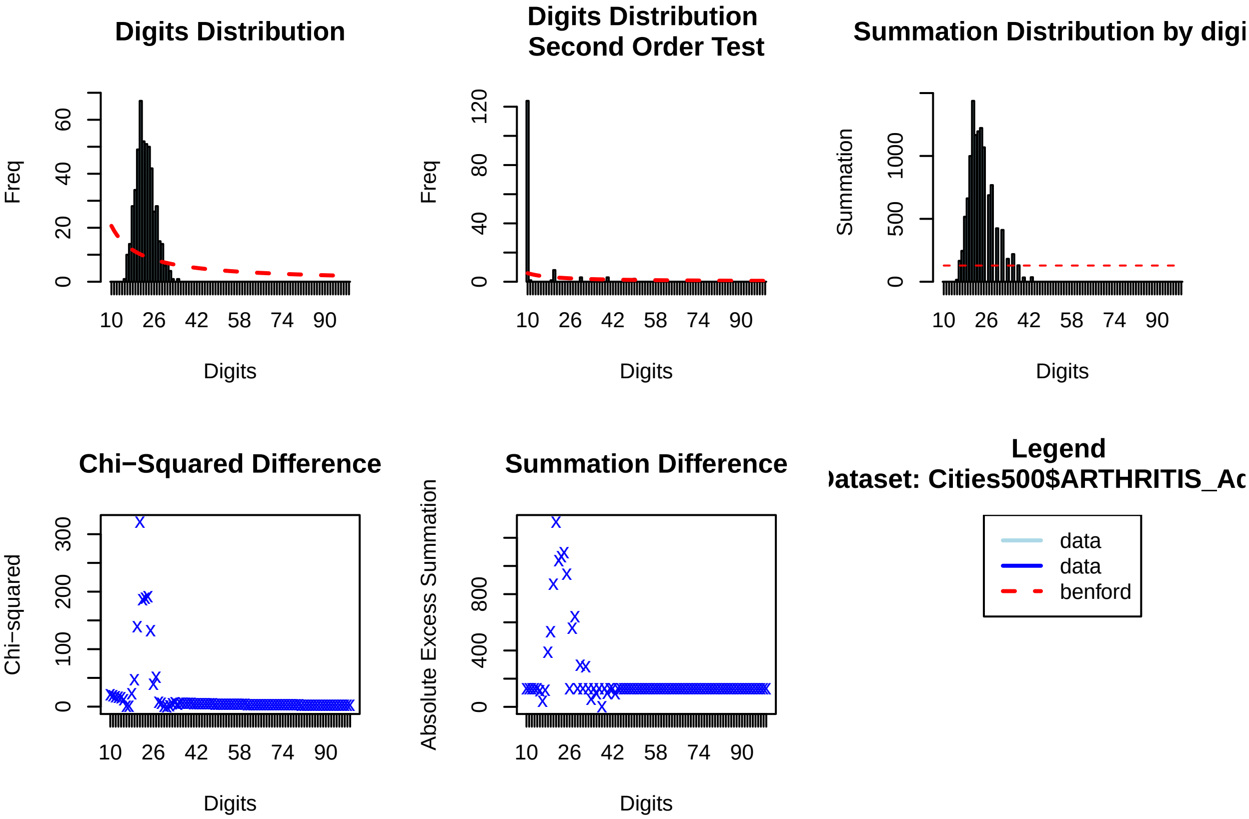


Diabetes

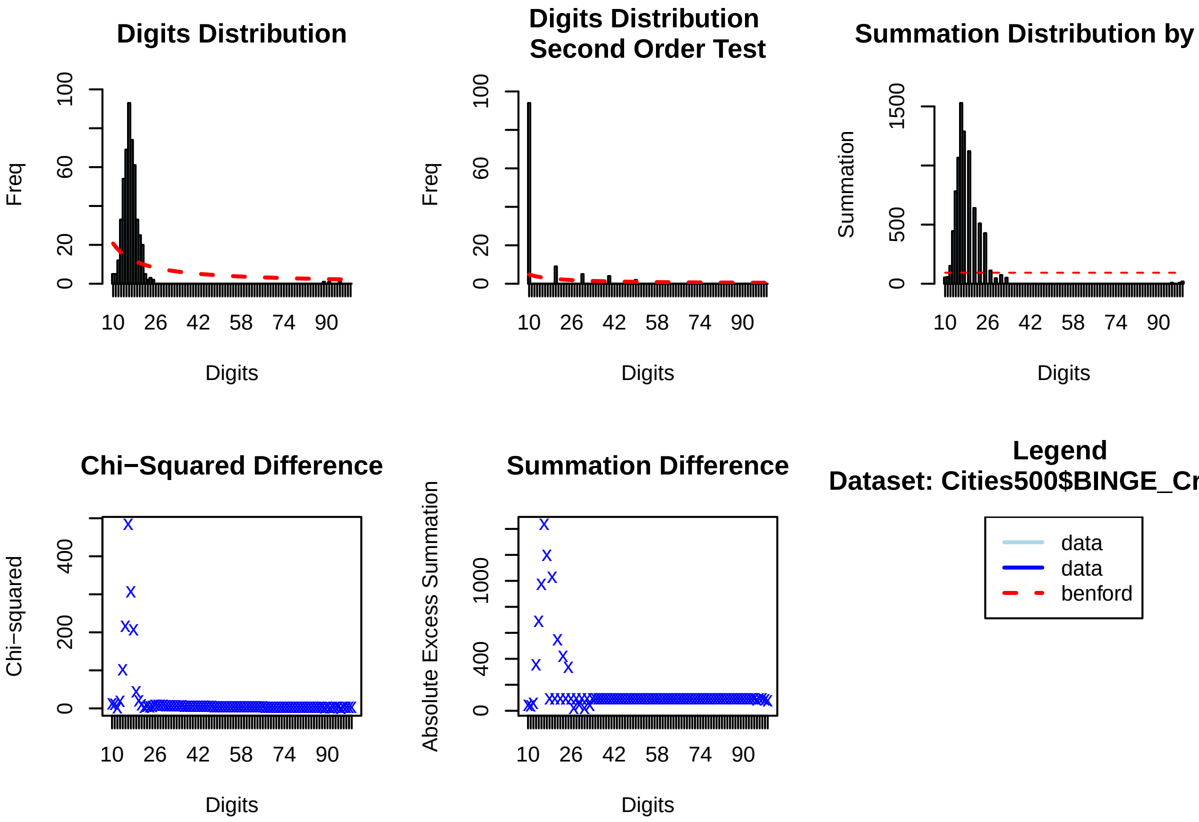


Poor mental health

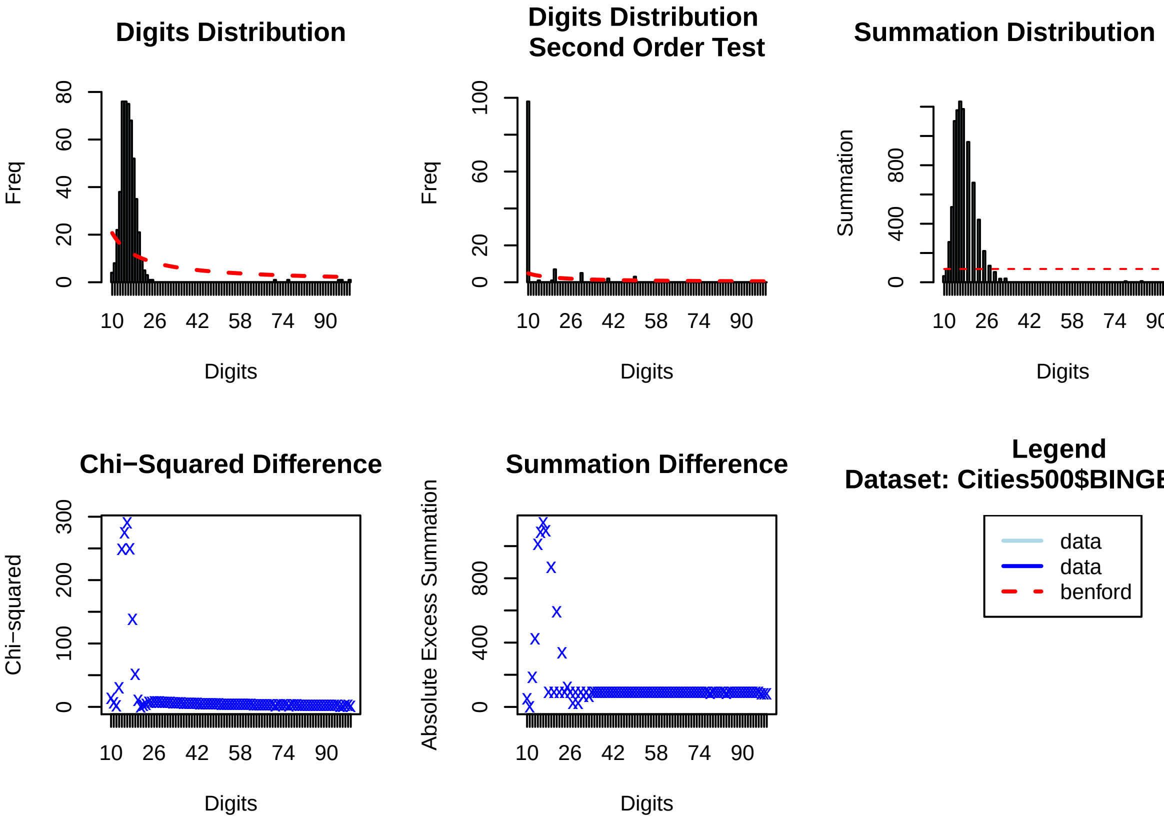
Results for other variables



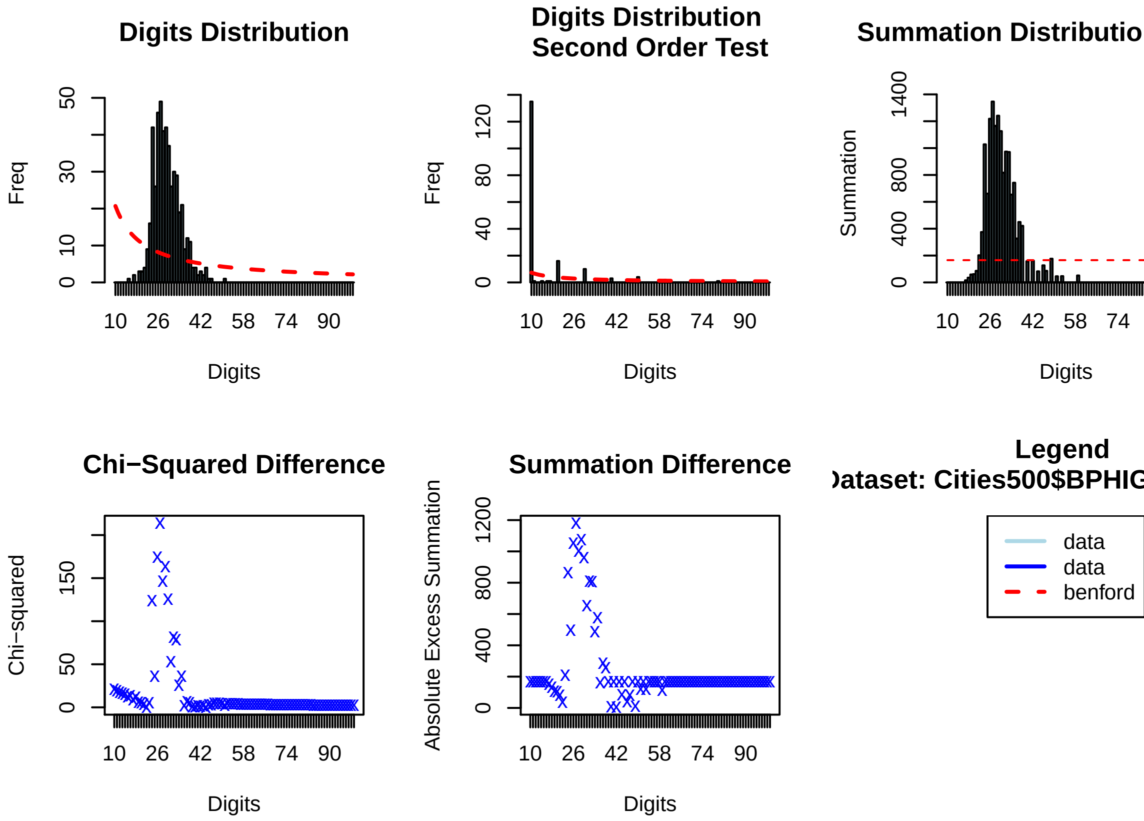
Arthritis



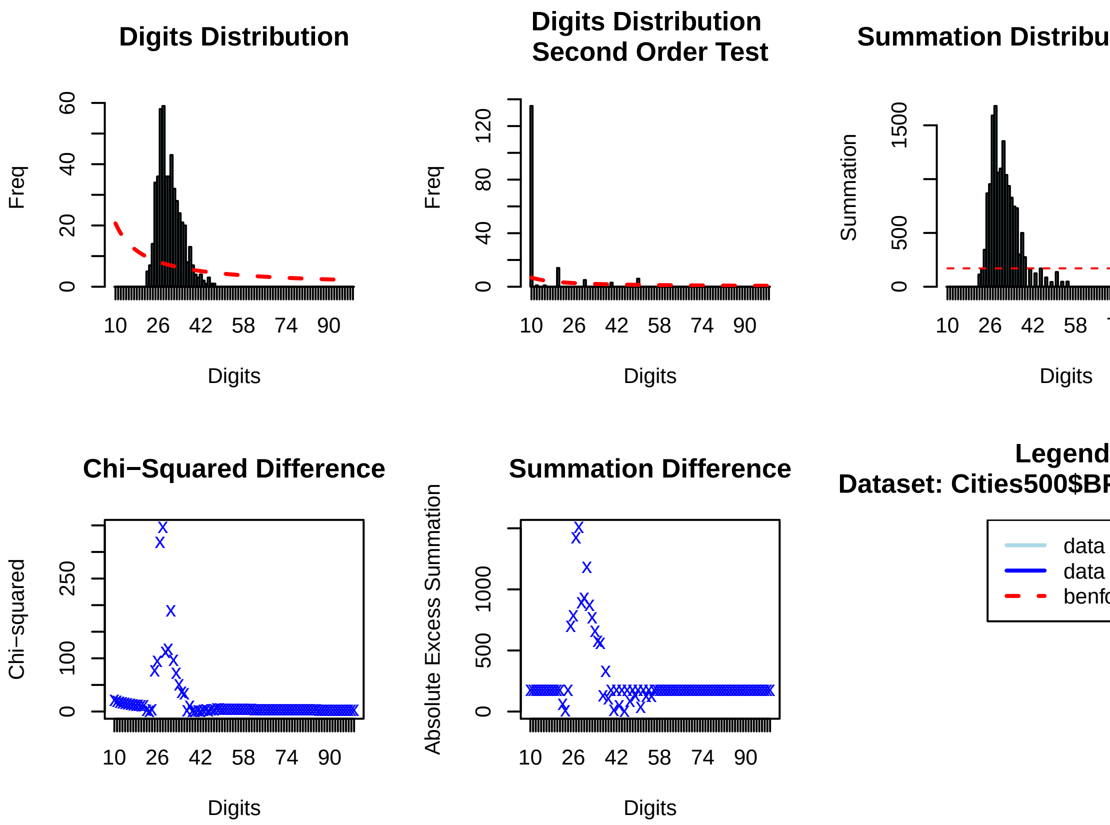
Binge drinking crude



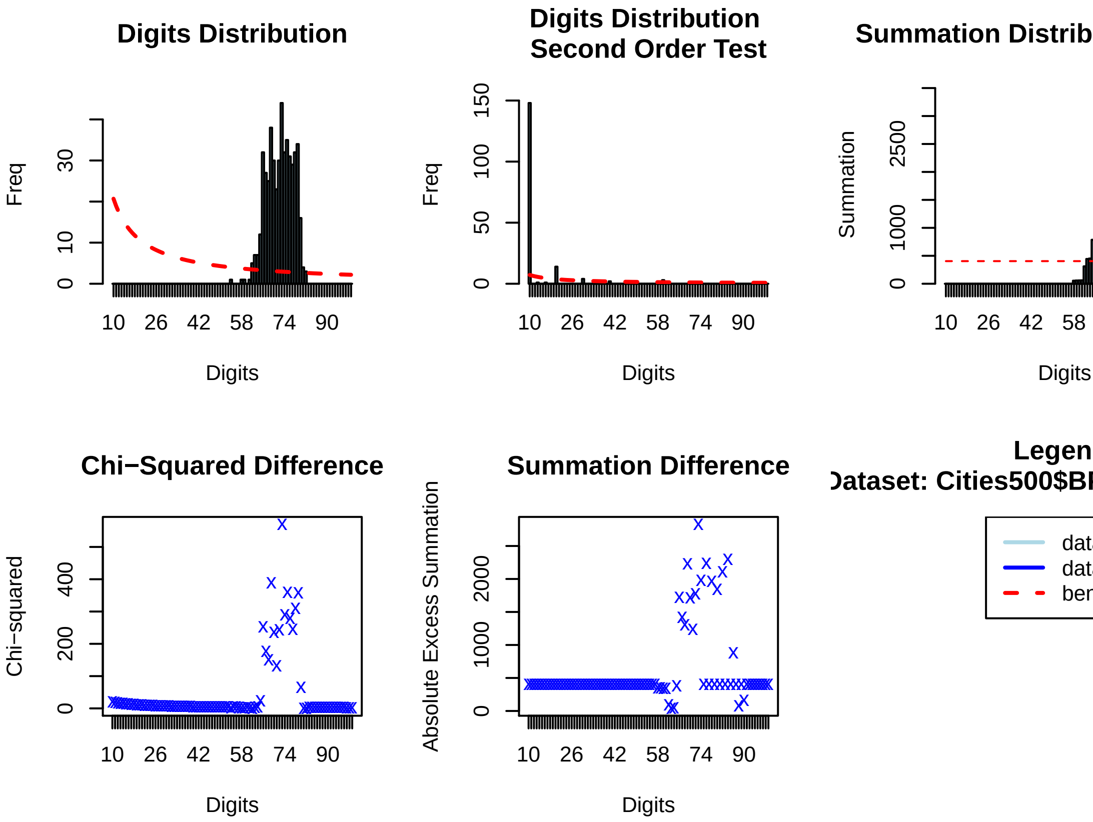
Binge drinking adjusted



High blood pressure crude

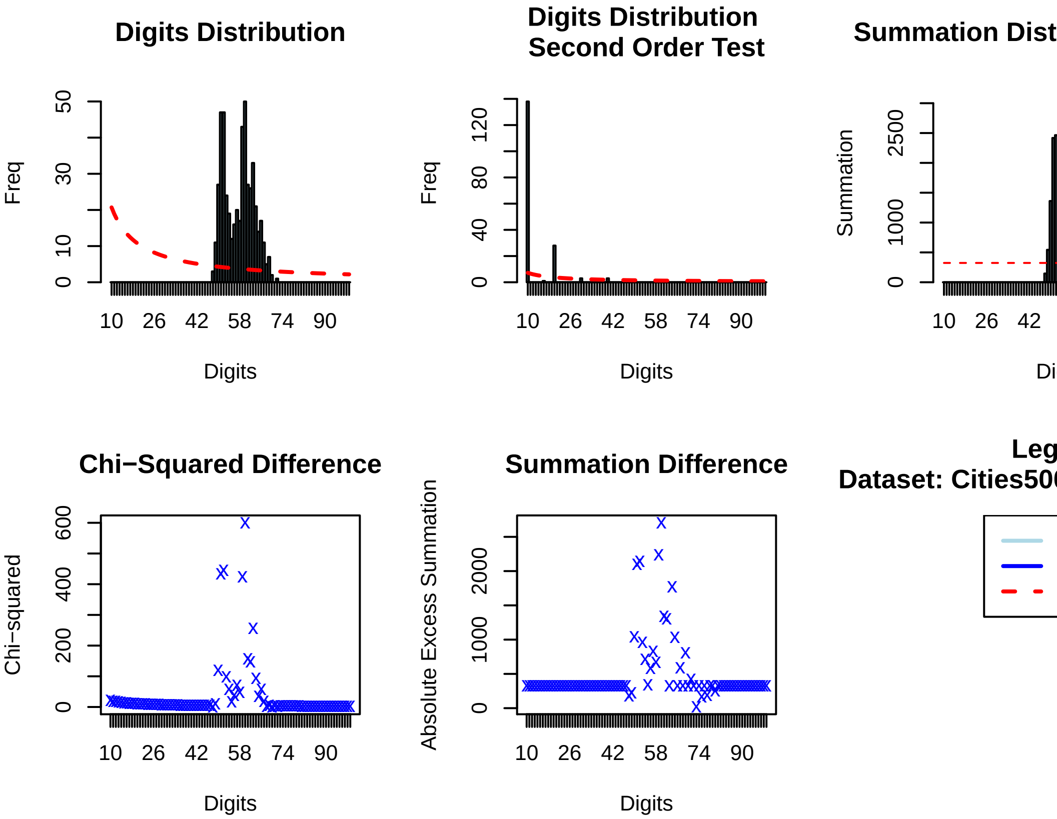


High blood pressure adjusted



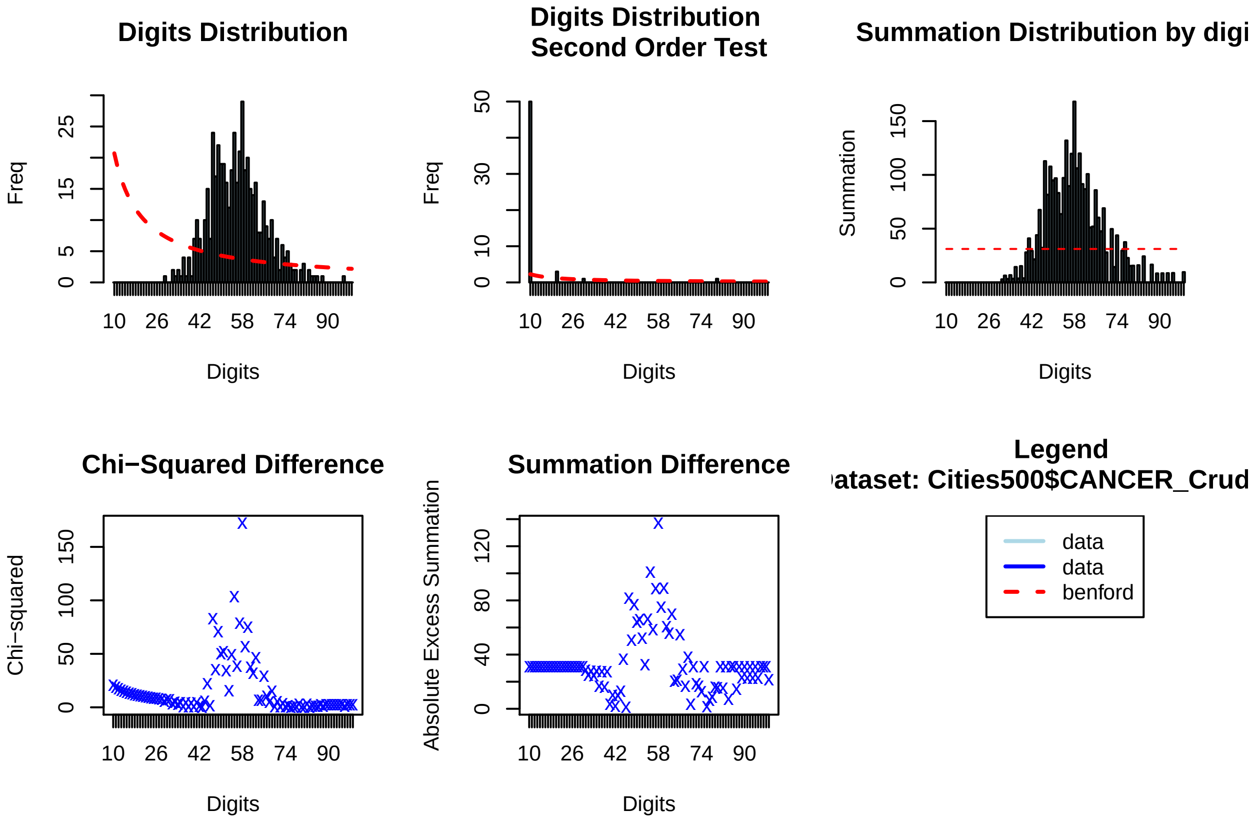
Blood pressure medicine crude

```
## [1] 82.9
```

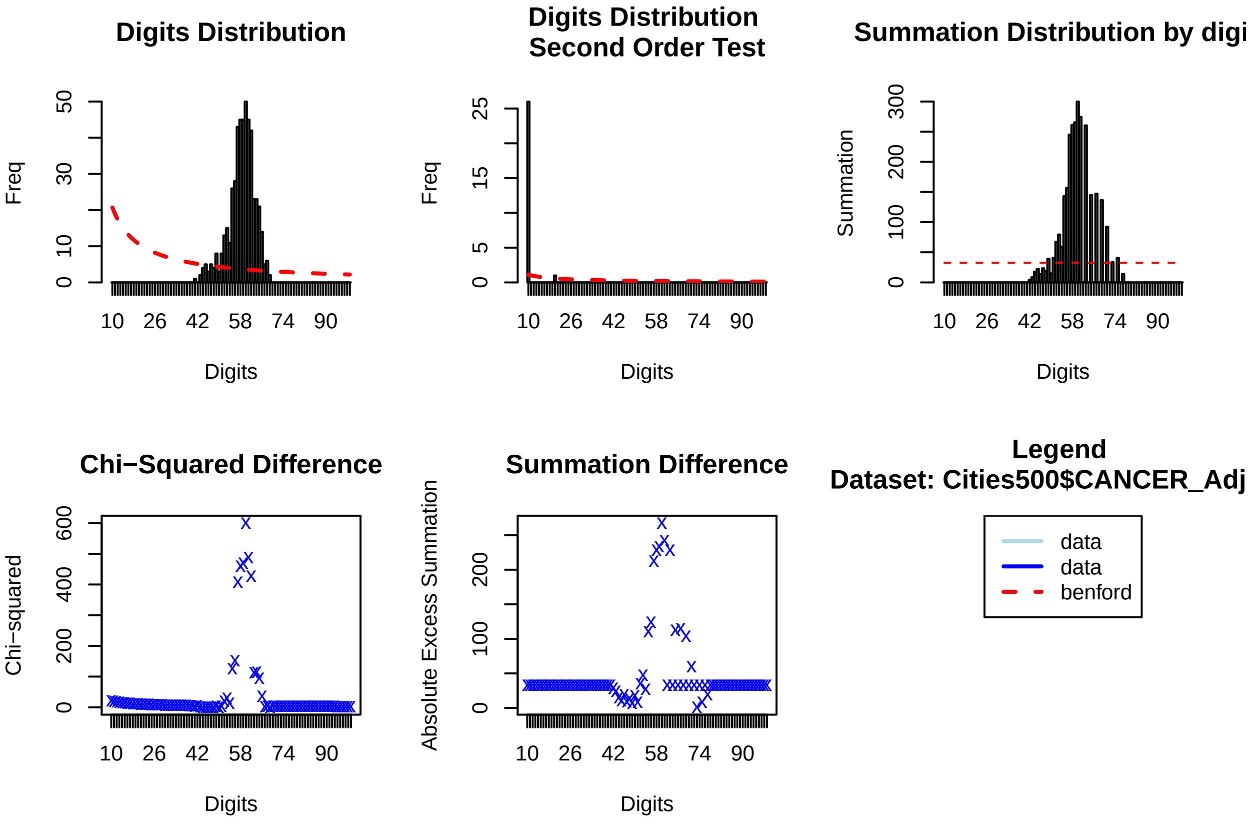


Blood pressure medicine adjusted

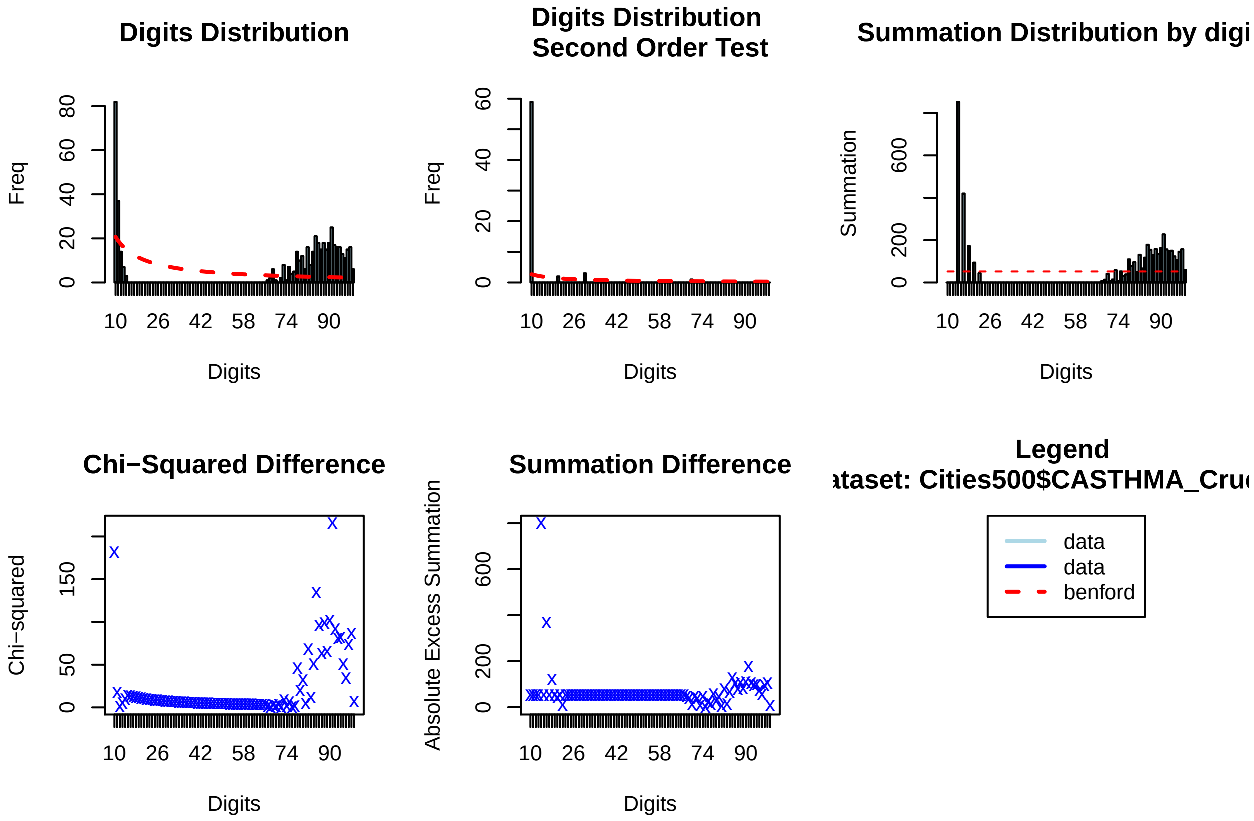
```
## [1] 72.1
```



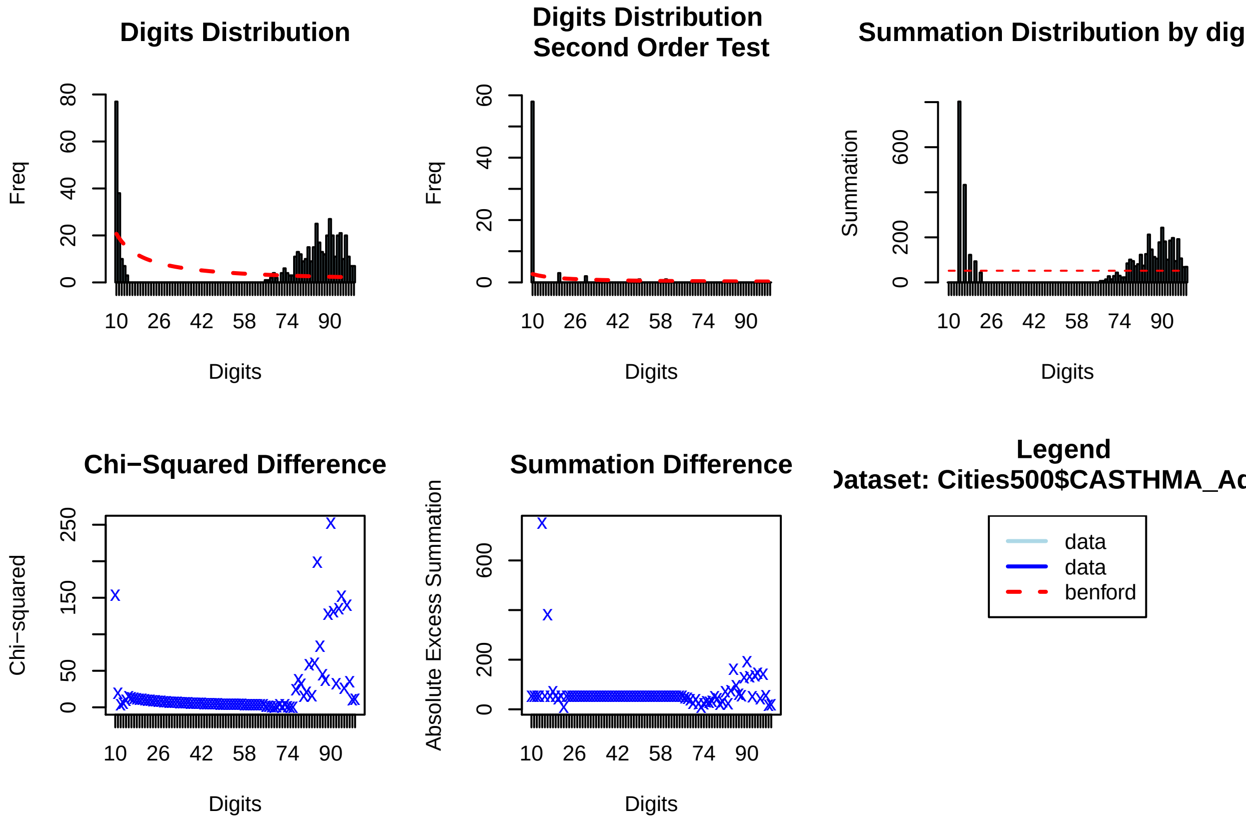
Cancer crude



Cancer adjusted



Asthma crude

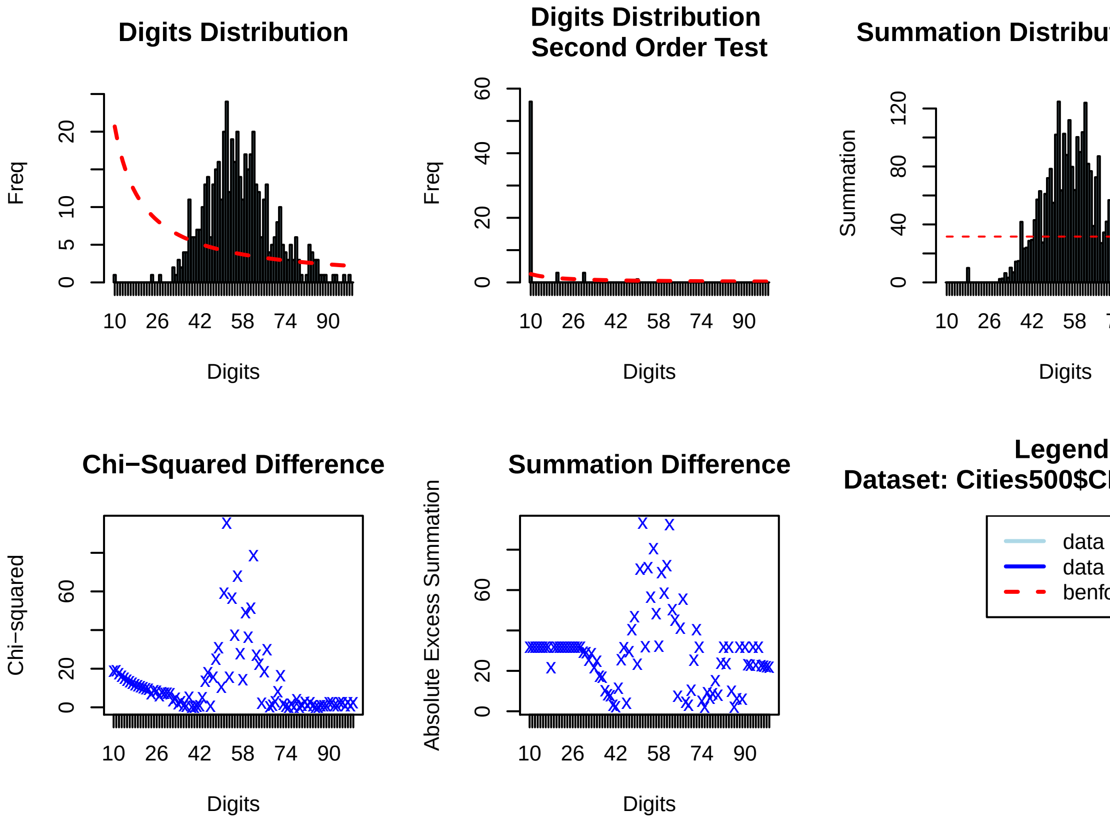


Asthma adjusted

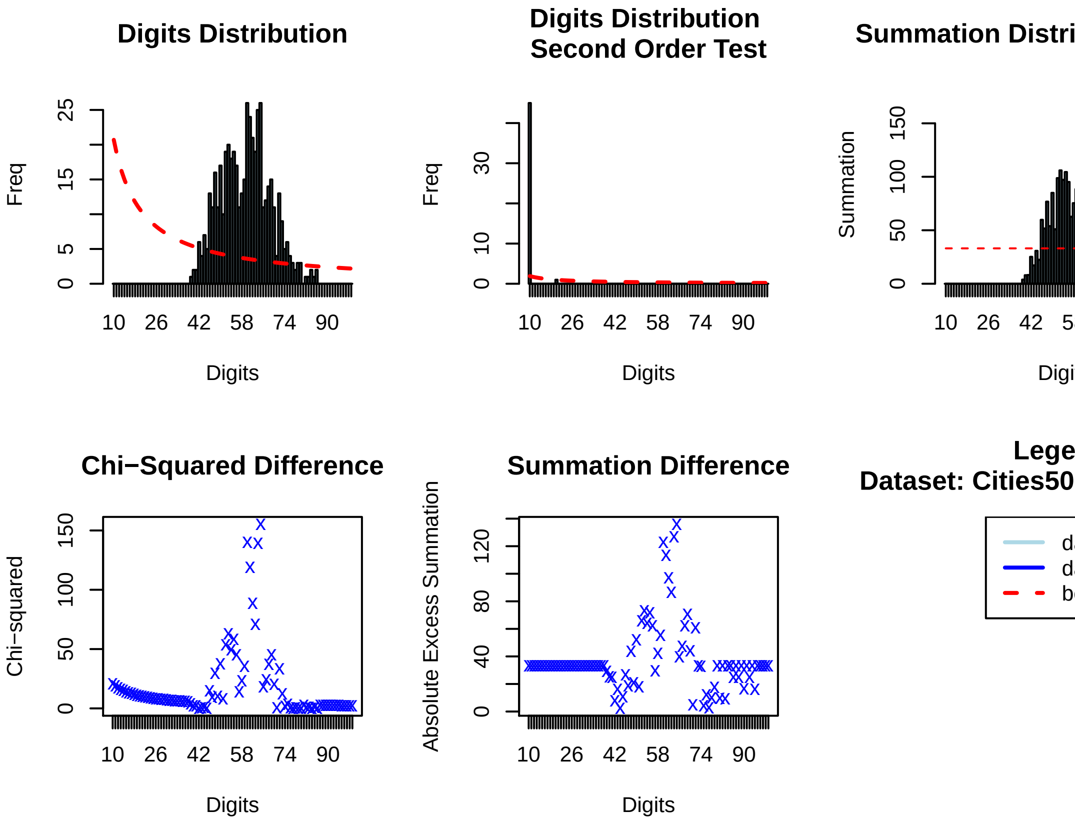
```
## [1] 14.7
```

```
## [1] 9.3316
```

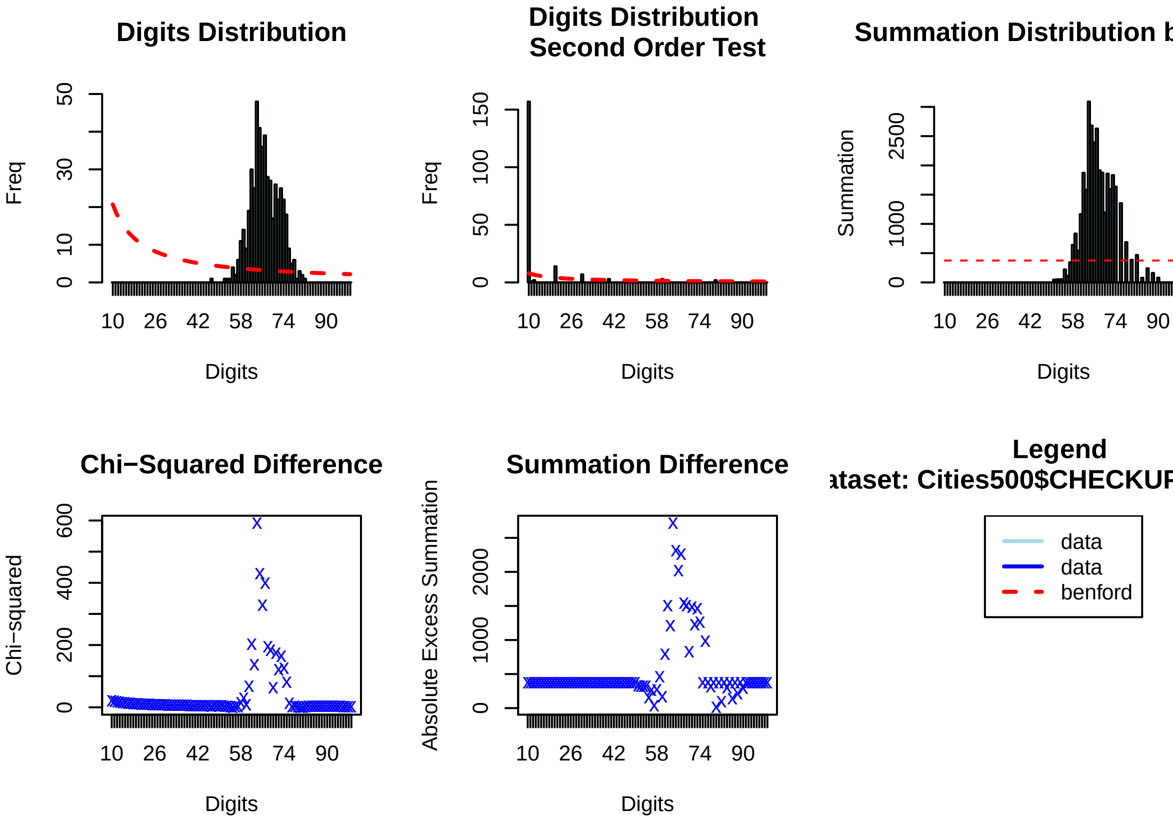
```
## [1] 6.6
```



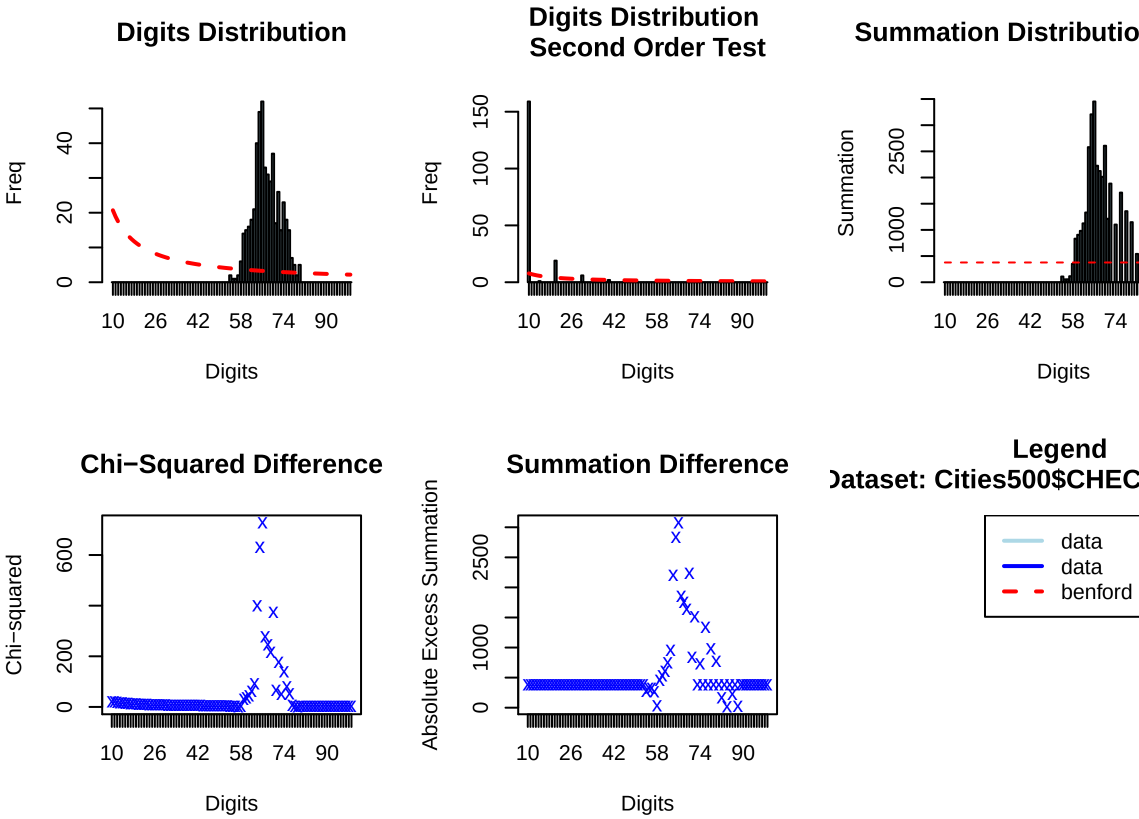
Coronary heart disease crude



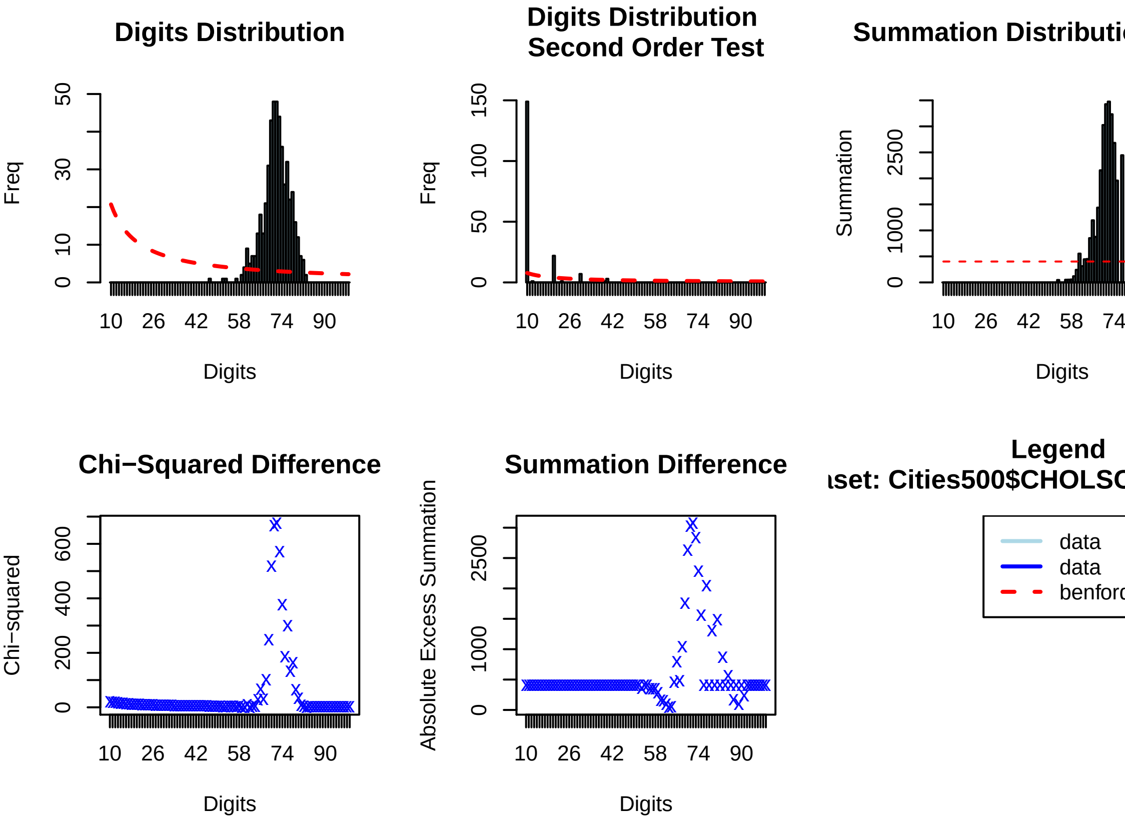
Coronary heart disease adjusted



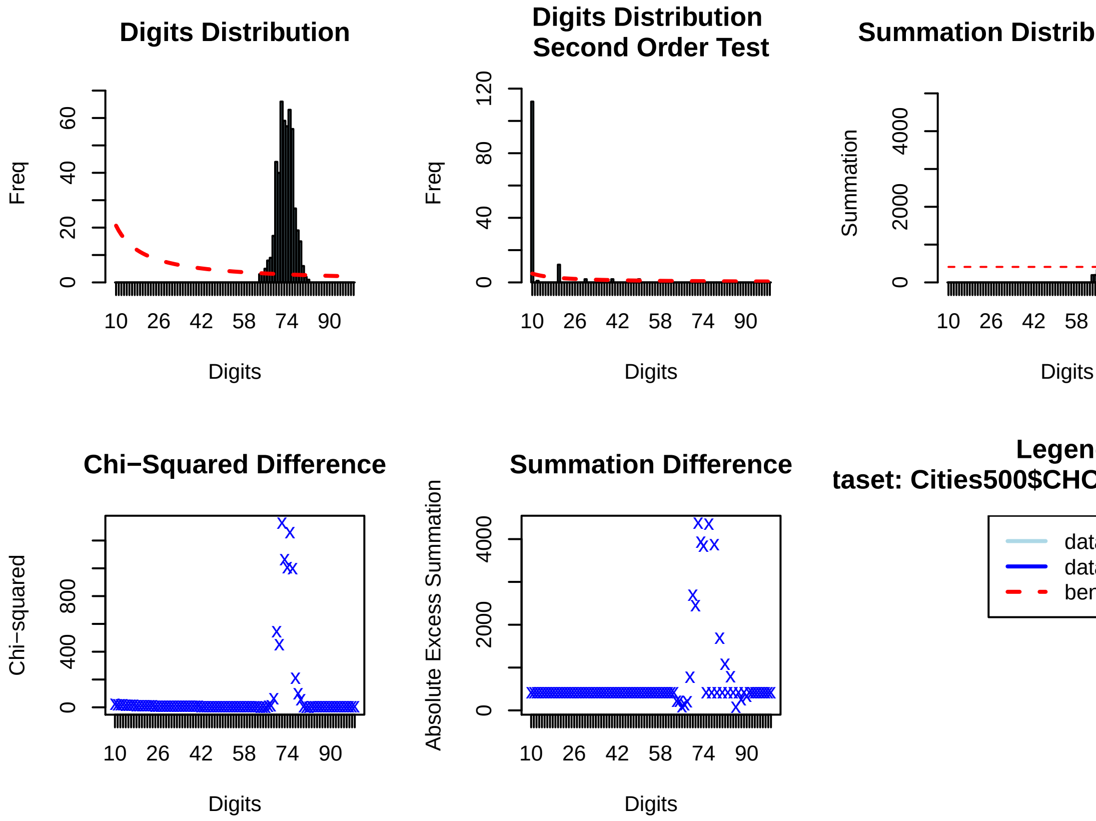
Medical check up crude



Medical check up adjusted



Cholesterol screening crude



Cholesterol screening adjusted

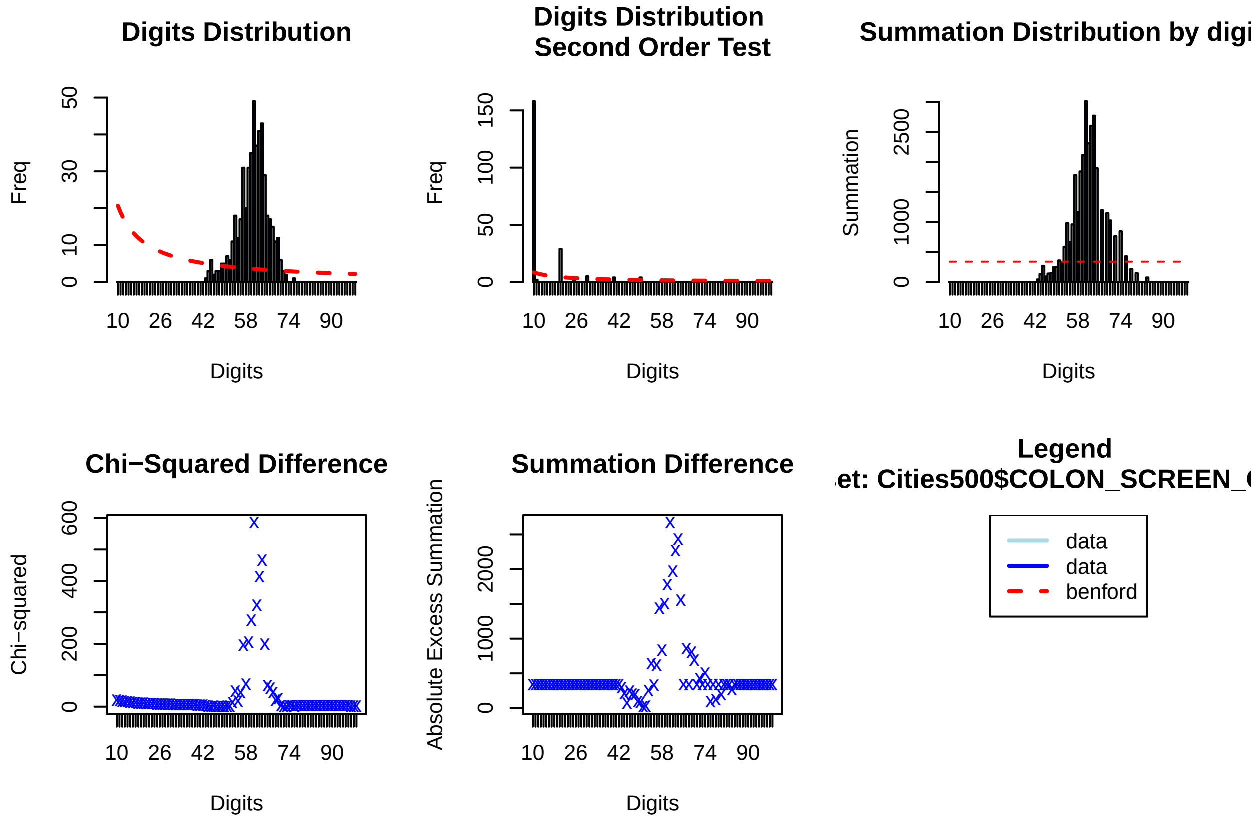
```
## [1] 82.5
```

```
## [1] 73.8346
```

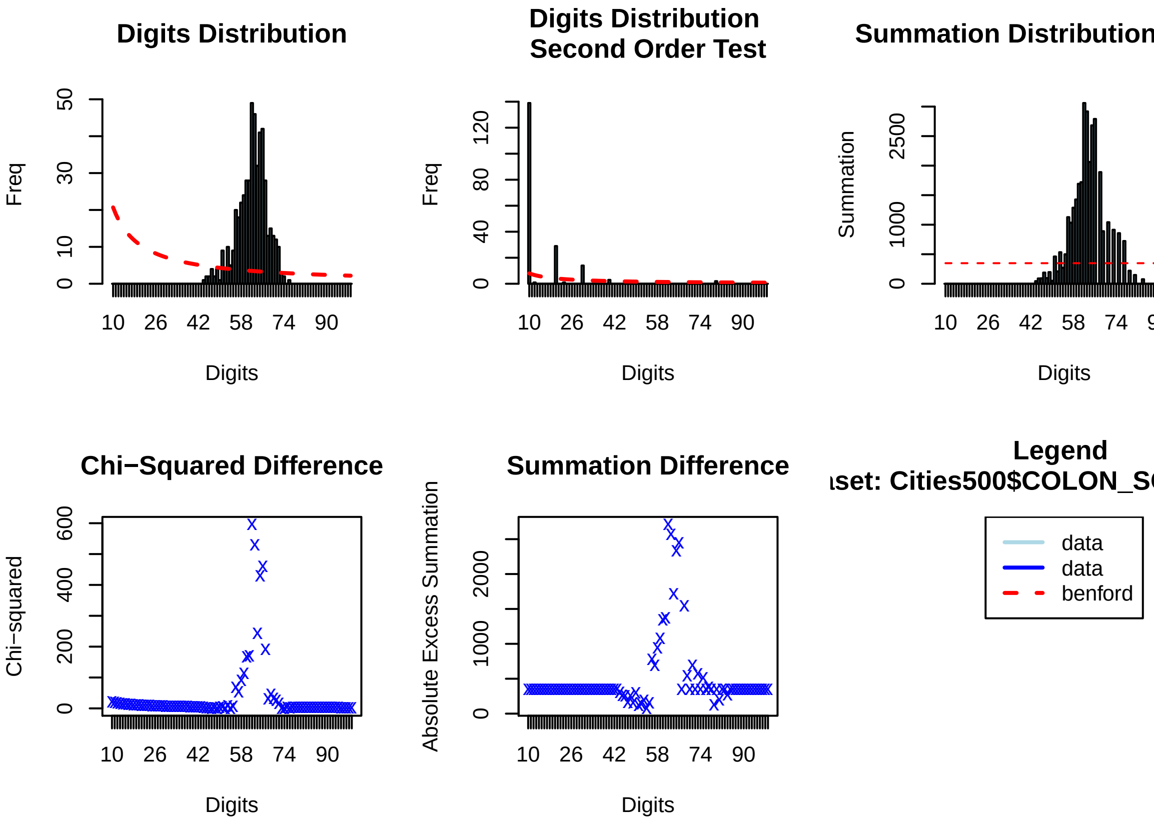
```
## [1] 64.2
```

Colon screening crude

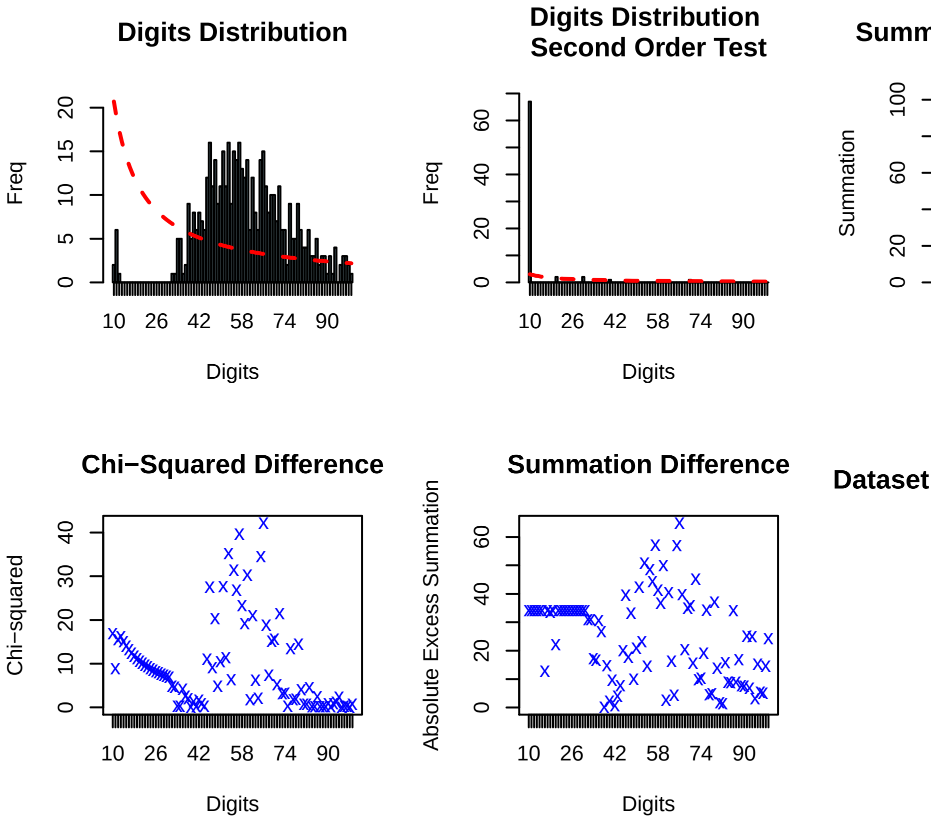
```
col <- benford(Cities500$COLON_SCREEN_CrudePrev)
plot(col)
```



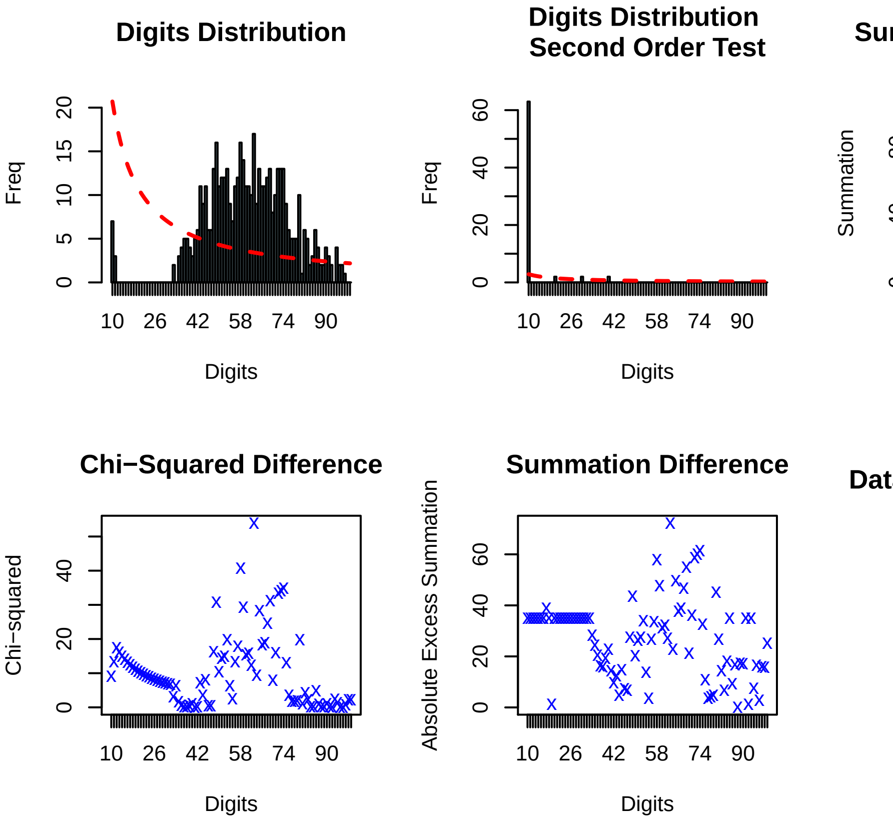
```
# max(Cities500$COLON_SCREEN_CrudePrev) #76.4
# mean(Cities500$COLON_SCREEN_CrudePrev) #61.1
# min(Cities500$COLON_SCREEN_CrudePrev) #43.5
```



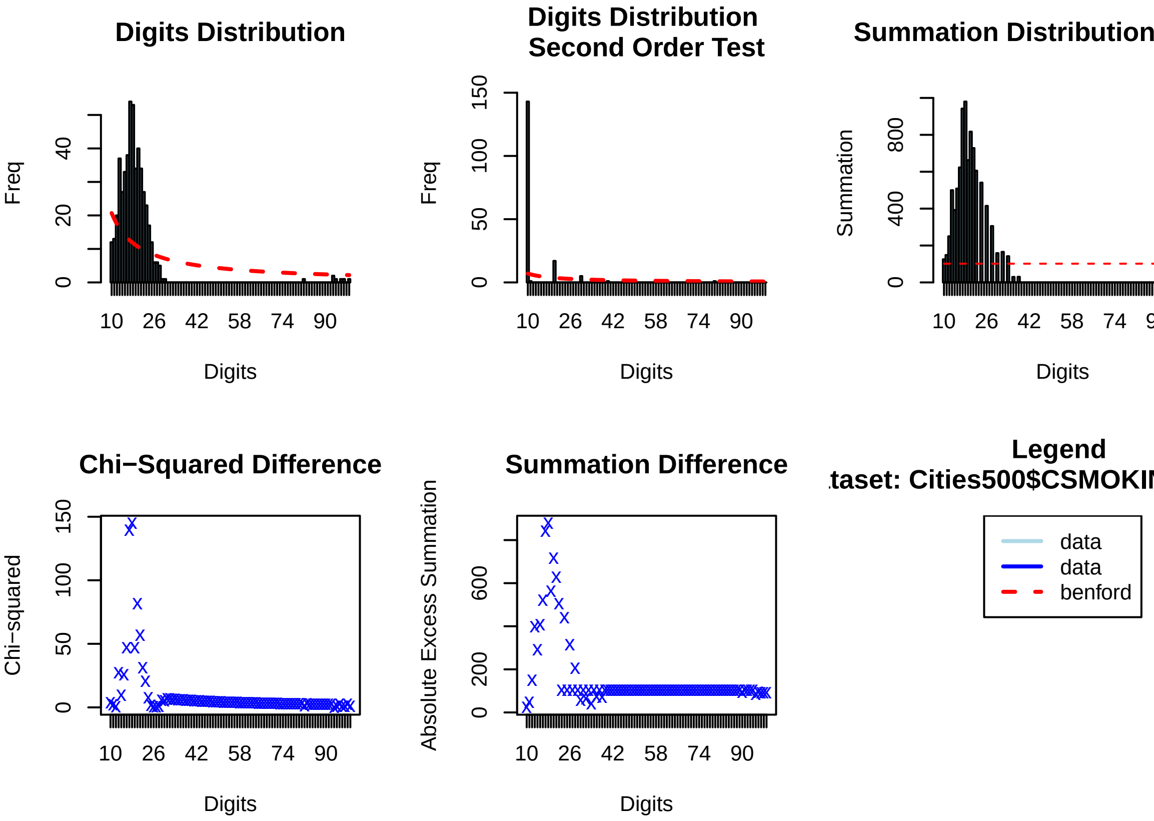
Colon screening adjusted



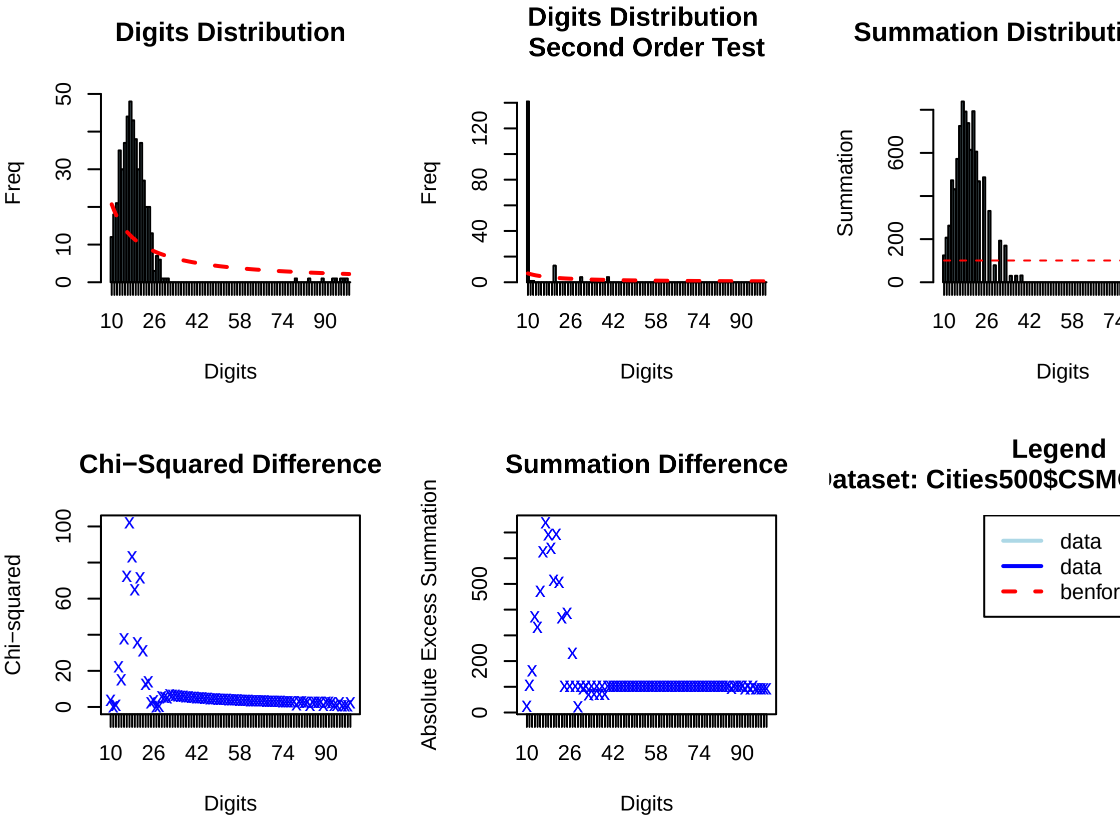
Chronic obstructive pulmonary disease crude



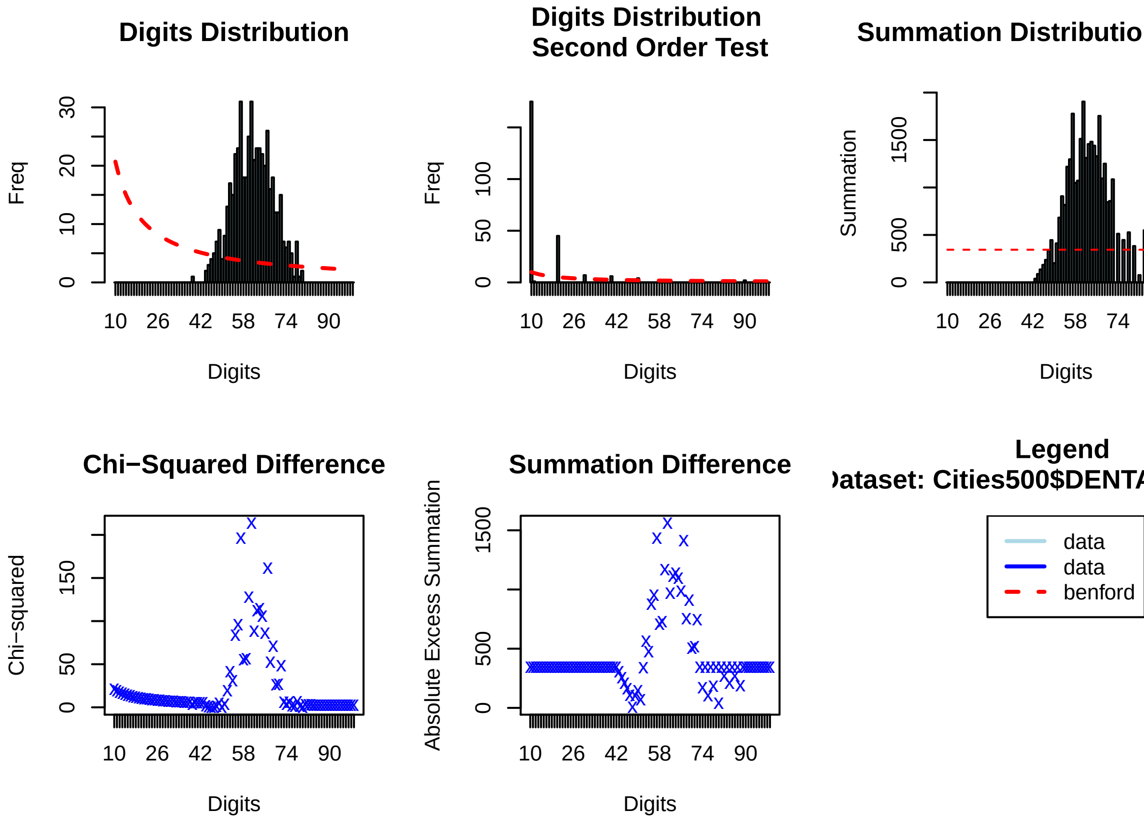
Chronin obstructive pulmonary disease adjusted



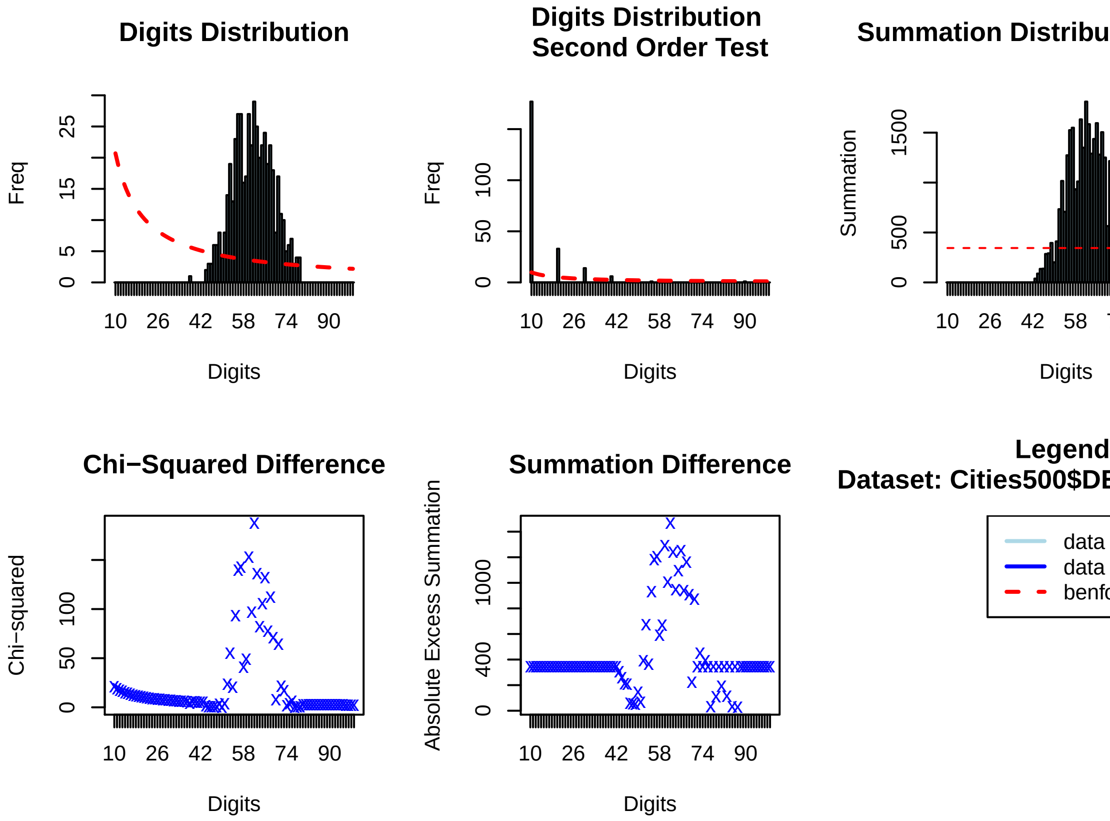
Currently smoking crude



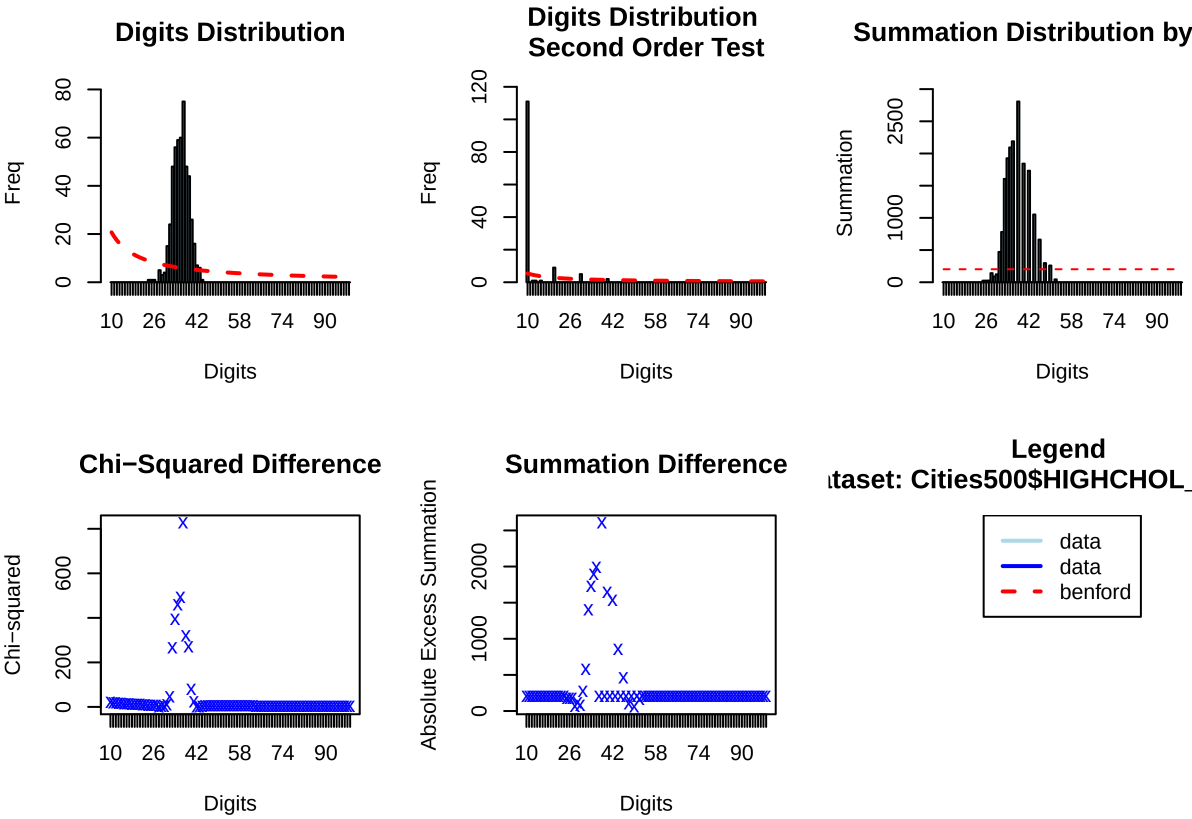
Currently smoking adjusted



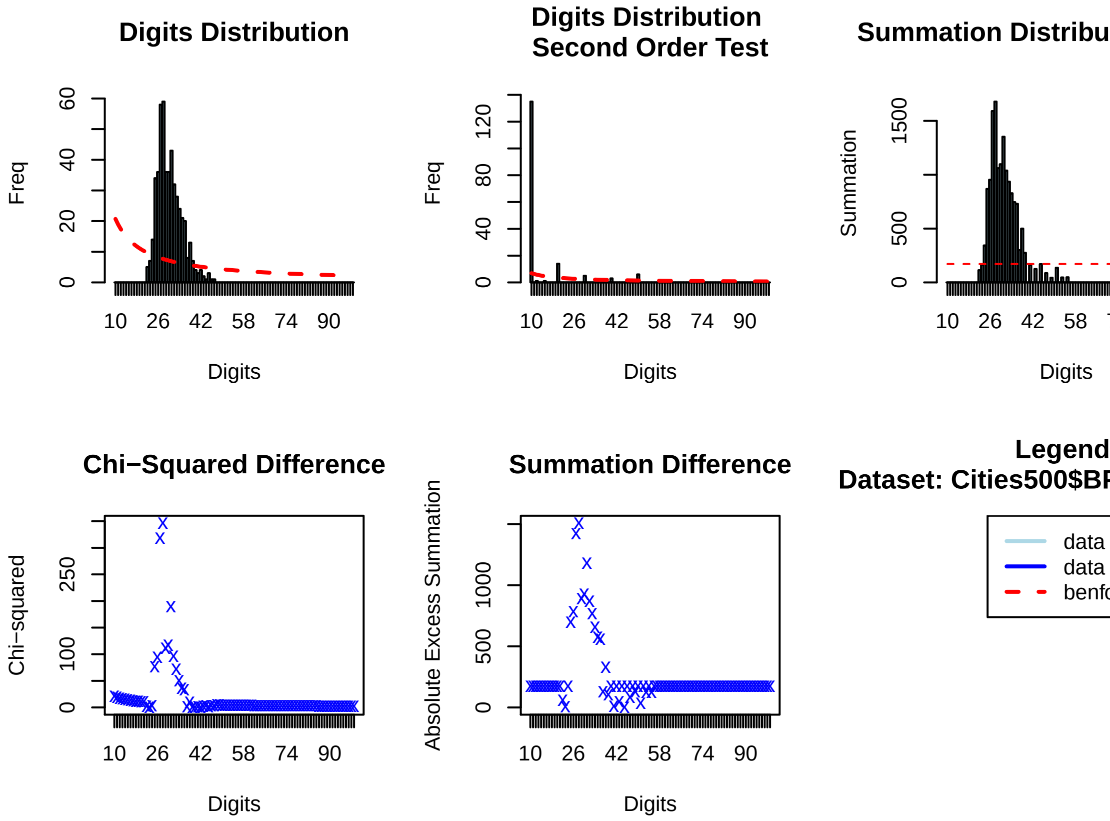
Going to the dentist crude



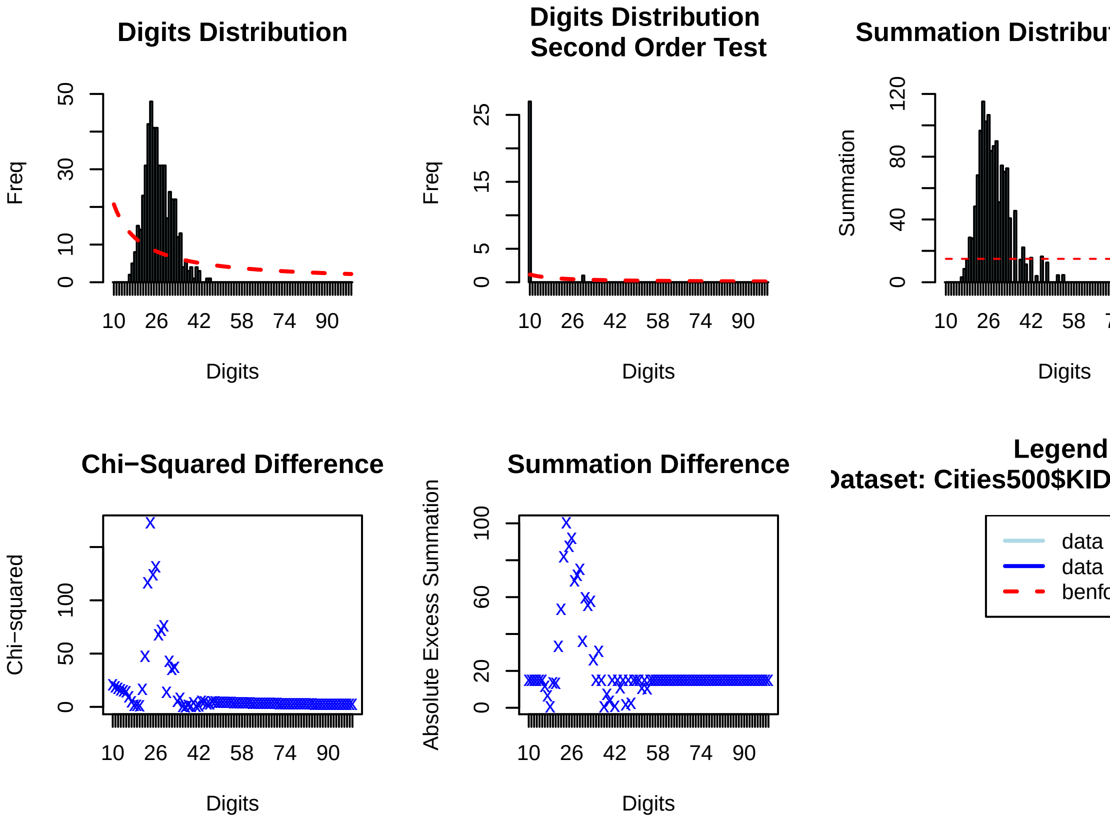
Going to the dentist adjusted



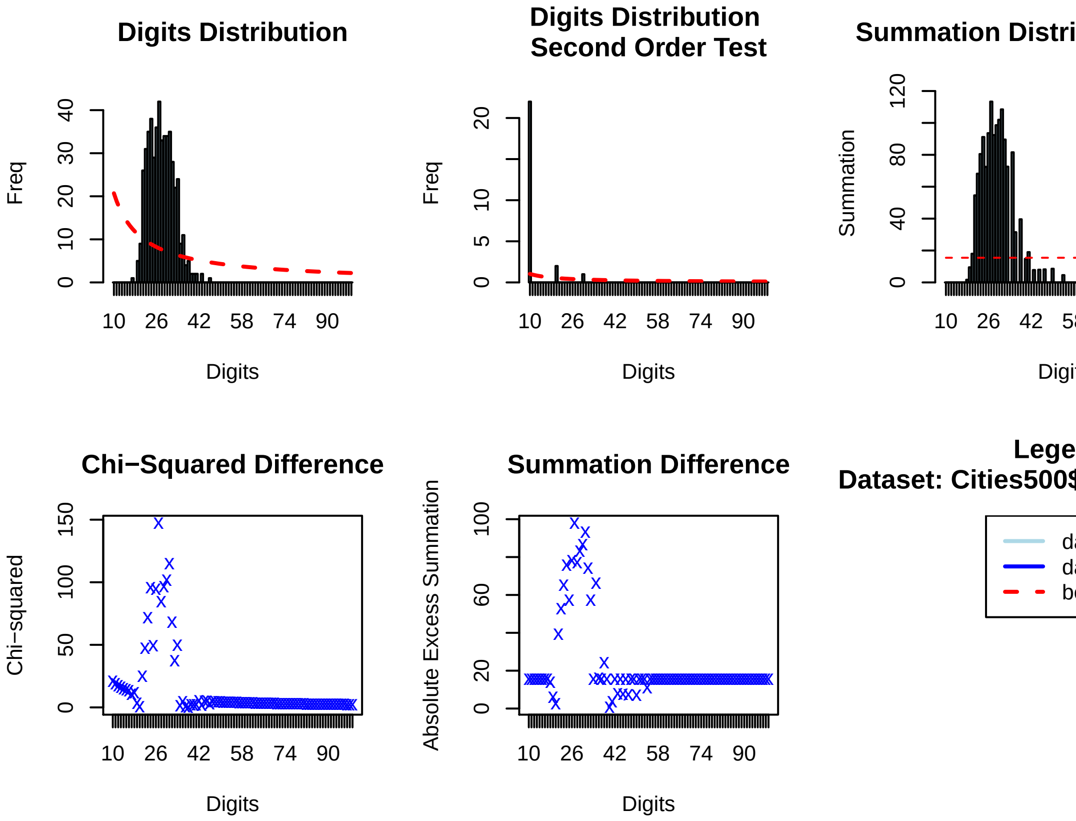
Binge Drinking crude



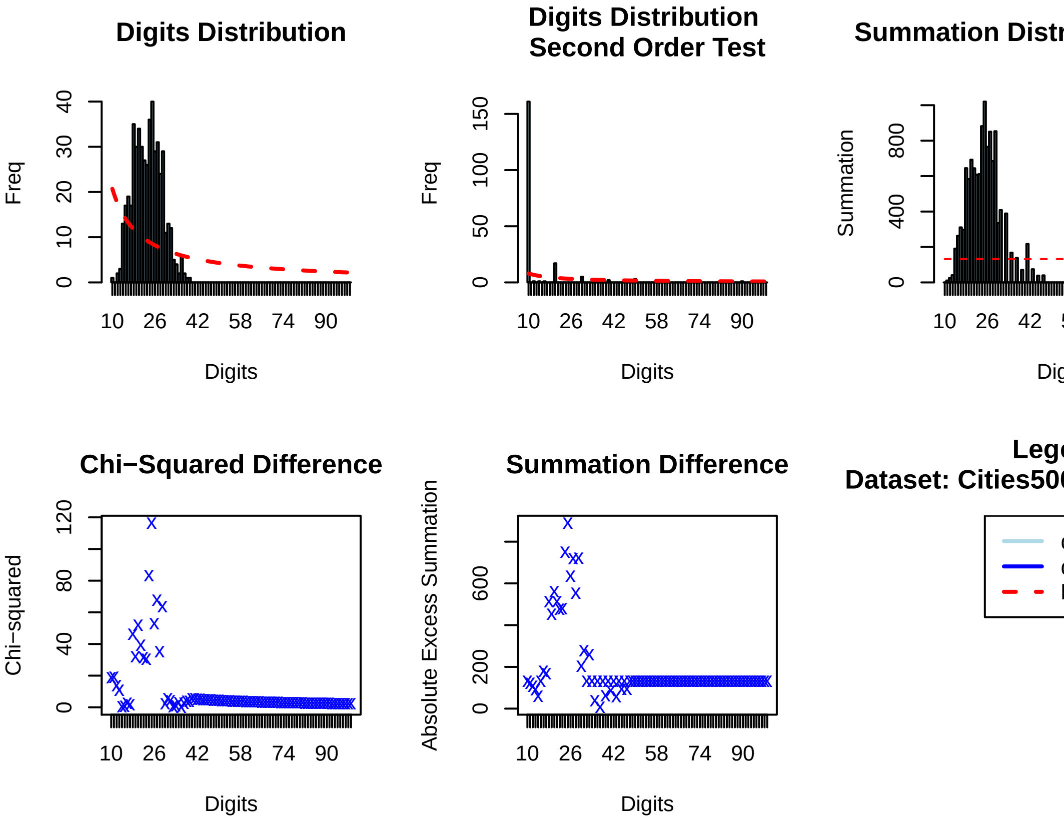
High blood pressure adjusted



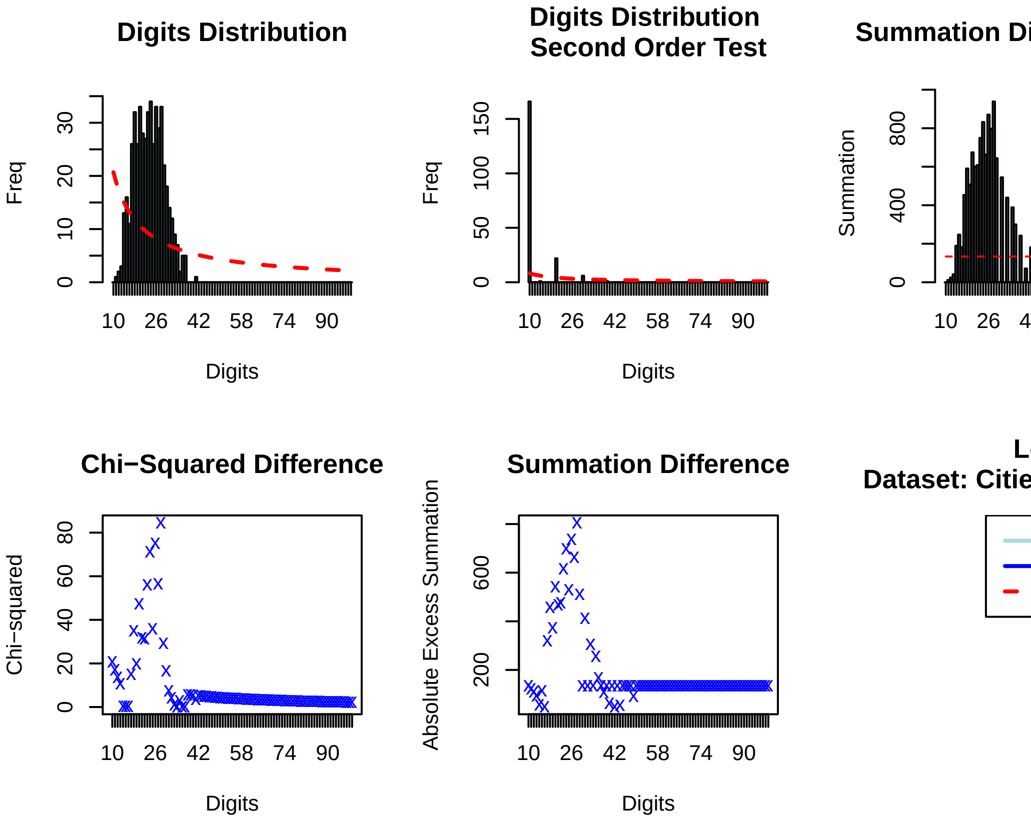
Chronic kidney disease crude



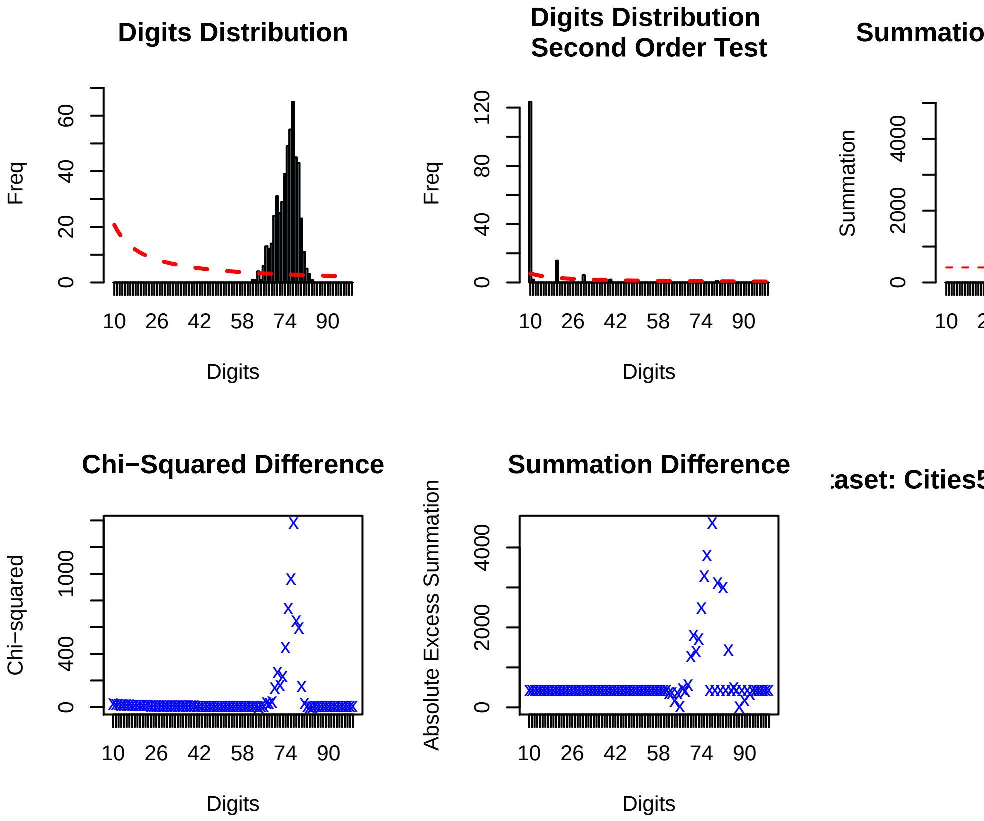
Chronic kidney disease adjusted



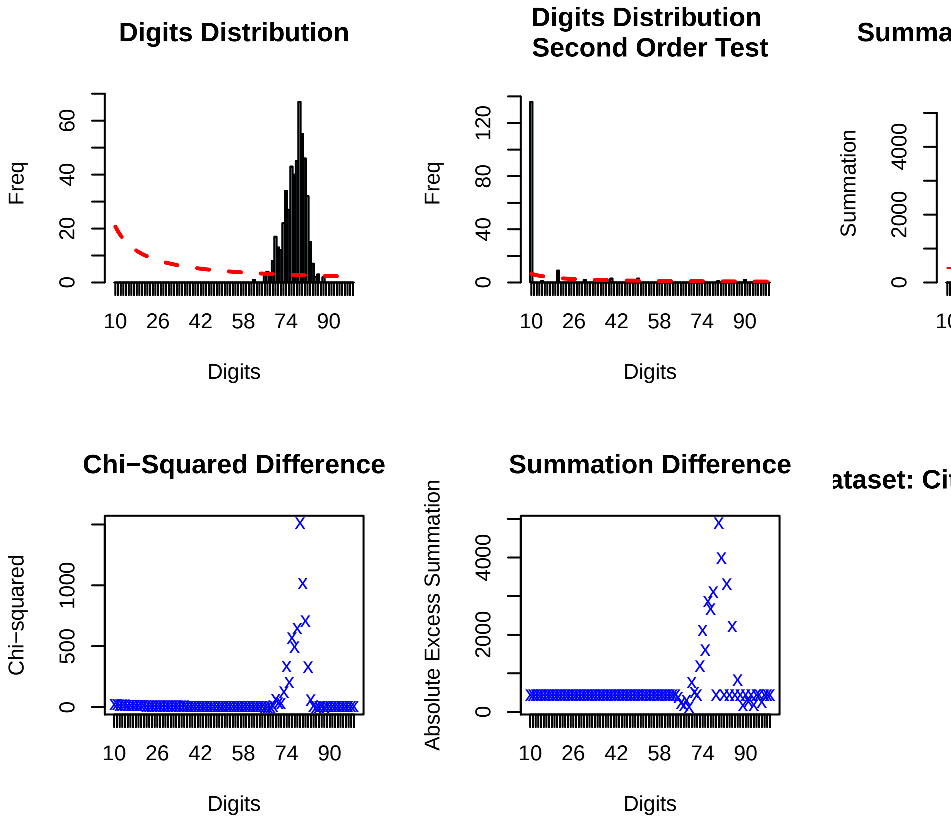
No leisure physical activity crude



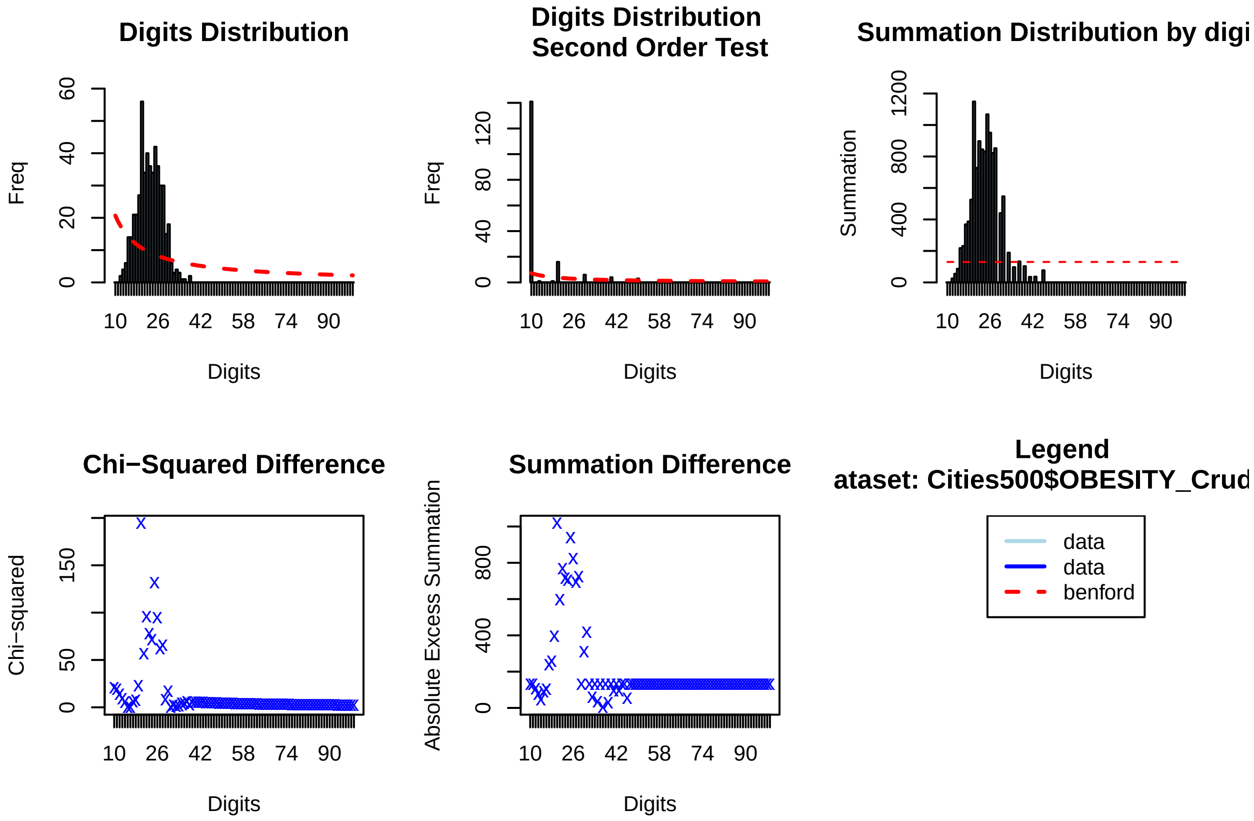
No leisure physical activity adjusted



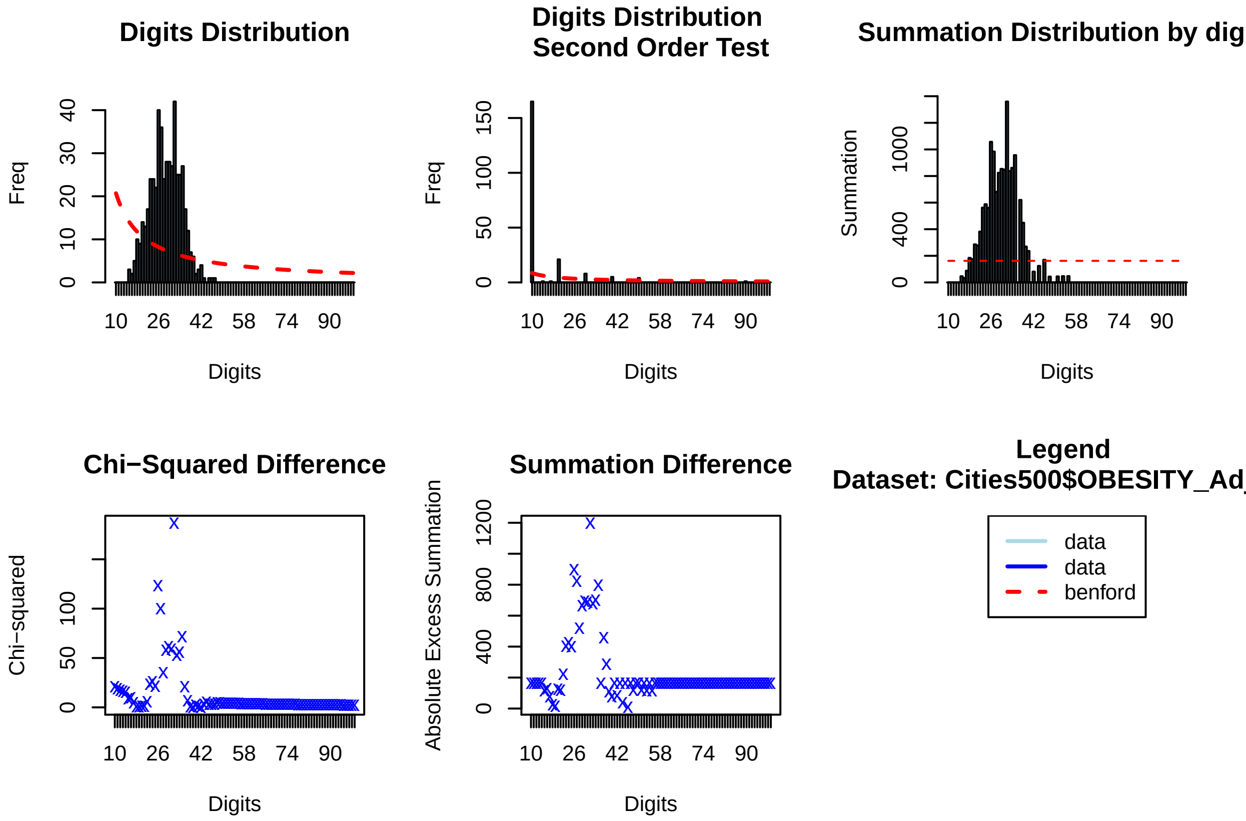
Mammogram use for women 50-74 crude



Mammogram use for women 50-74 adjusted



Obesity crude

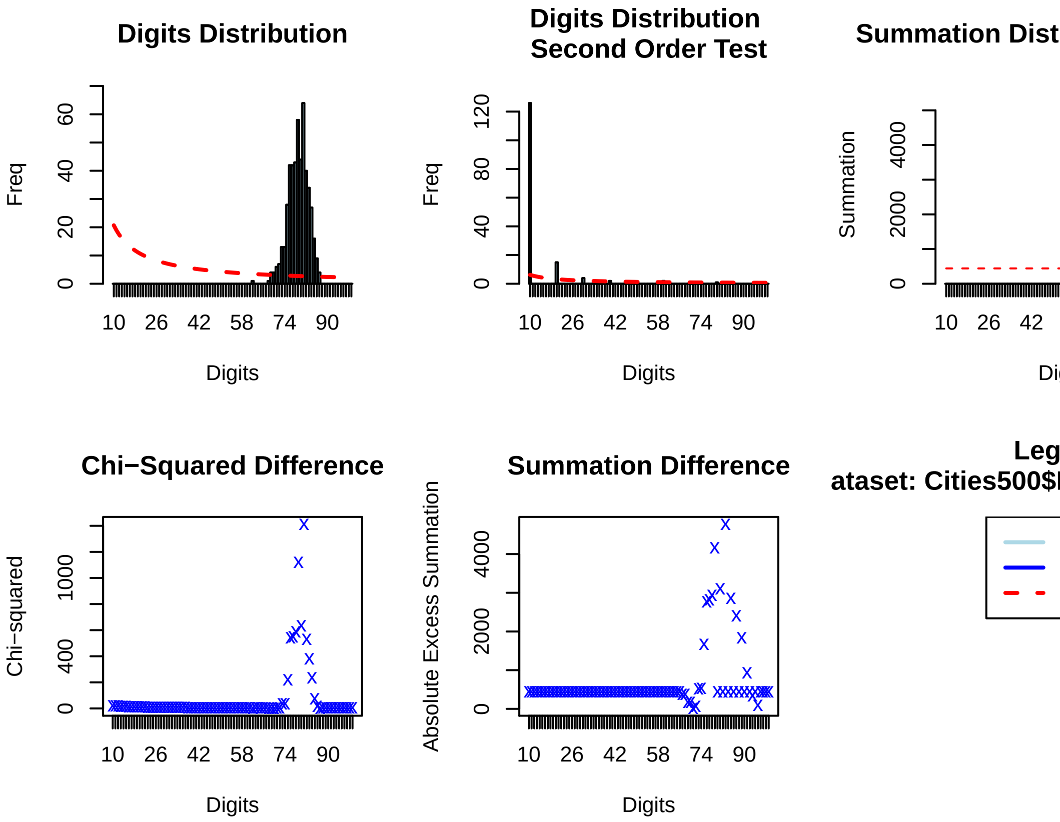


Obesity adjusted

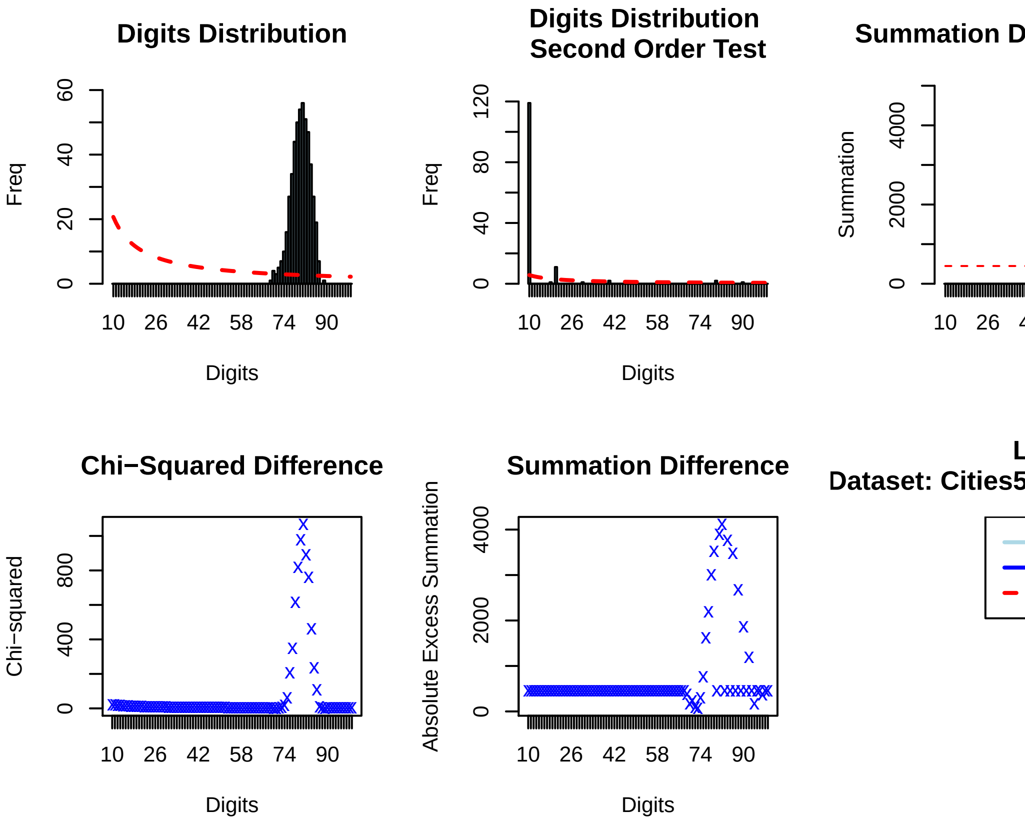
```
## [1] 47.2
```

```
## [1] 29.1638
```

```
## [1] 15.2
```



Pap smear for women 21-65 crude



Pap smear for women 21-65 adjusted

```
## [1] 89.4
```

```
## [1] 80.644
```

```
## [1] 69.3
```