Final Project

Longhao

12/1/2018

Abstract

This project is to use Benford's law to examine the statistics of aviation traffic data. Every few months, department of transportation will release the data of aviation traffic, which is provided by each airline. I am interested in finding out whether the data provided by airline companies are true or not. Particularly, I selected four variables that are distinctive but also correlated with each other. They are number of available seats, number of passengers, the distance of the flight and the airtime.

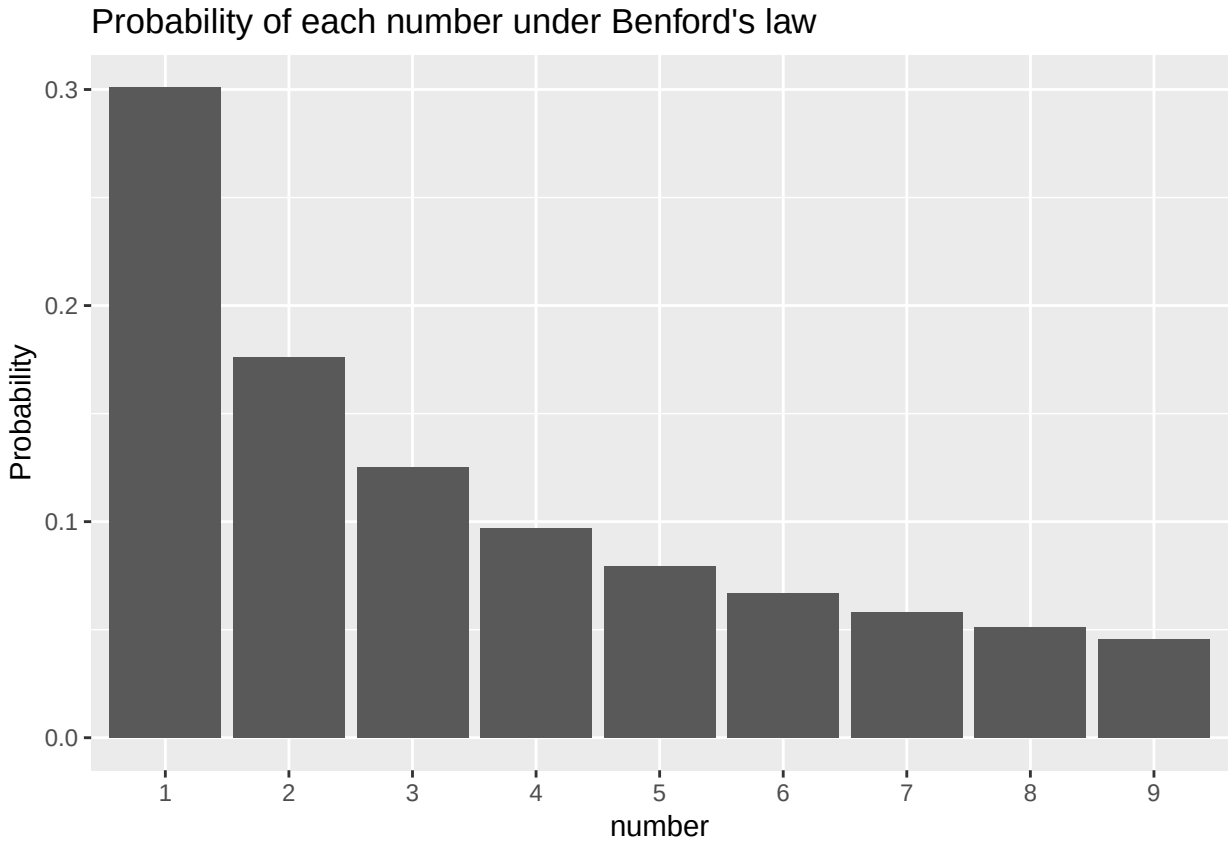
Introduction

Benford's law is a phenomenological law also called the first digit law. This law states that in the listings, tables of statistics, etc., the digit 1 tends to occur with probability of 30%, greater than the expected of 11% (i.e., one out of nine). The mathmatical form is like below.

$$Prob(D_1 = d) = log_{10}(1 + \frac{1}{d}) \; for \; d = 1, 2, ...9;$$

Here is a probability distribution table and a bar plot from 1 to 9.

Probability	number
0.3010	1
0.1760	2
0.1250	3
0.0970	4
0.0792	5
0.0669	6
0.0580	7
0.0512	8
0.0458	9

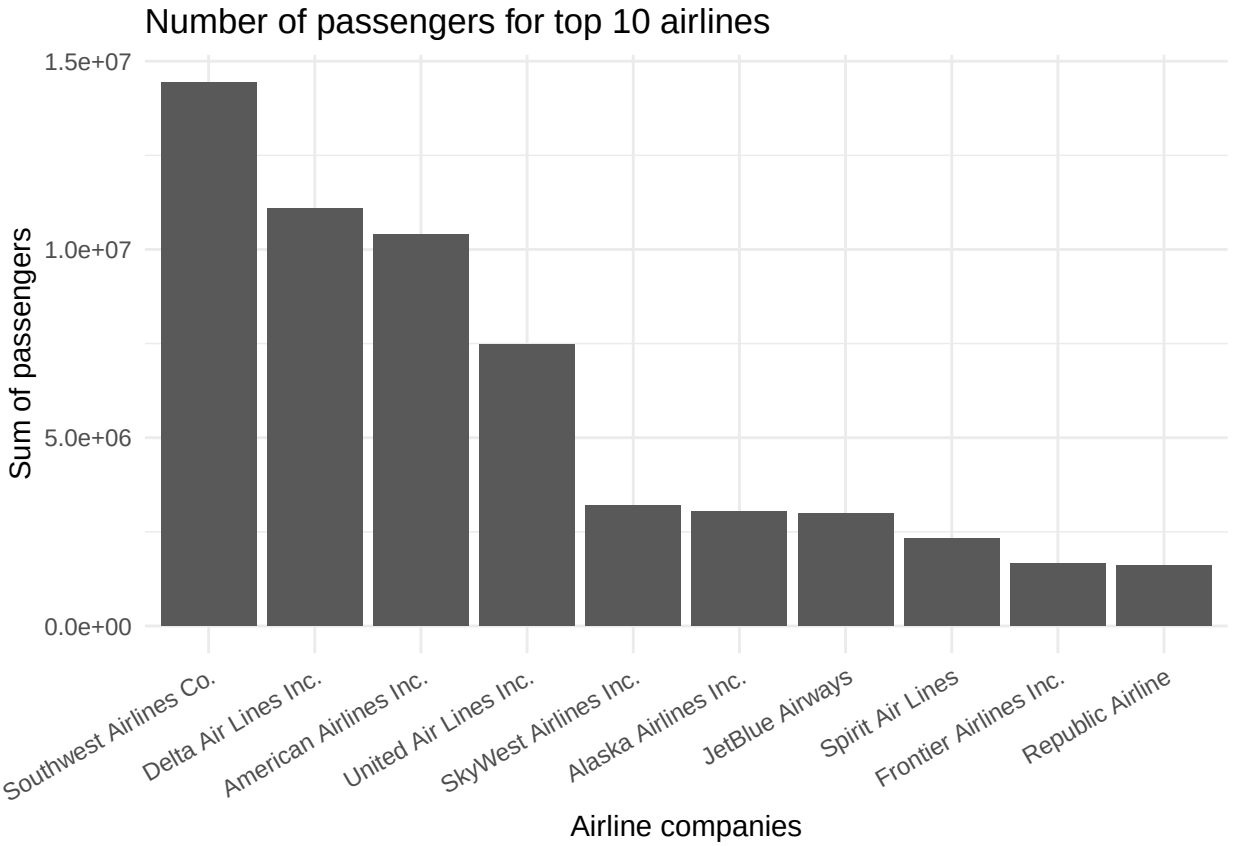


Materials and methods

The data that constitutes this research is from Bureau of Transportation Statistics(BTS) and can be found on this website https://www.bts.gov. The job of BTS is to collect and compile the data; however, they can not guarante that the data provided by the airline companies are accurate. We will first do some visualization of our data by looking at the largest 10 airlines.

```
## Parsed with column specification:
## cols(
##   UNIQUE_CARRIER_NAME = col_character(),
##   YEAR = col_integer(),
##   MONTH = col_integer(),
##   sum_seats = col_integer(),
##   sum_passengers = col_integer(),
##   sum_distance = col_integer(),
##   sum_airtime = col_integer()
## )
```

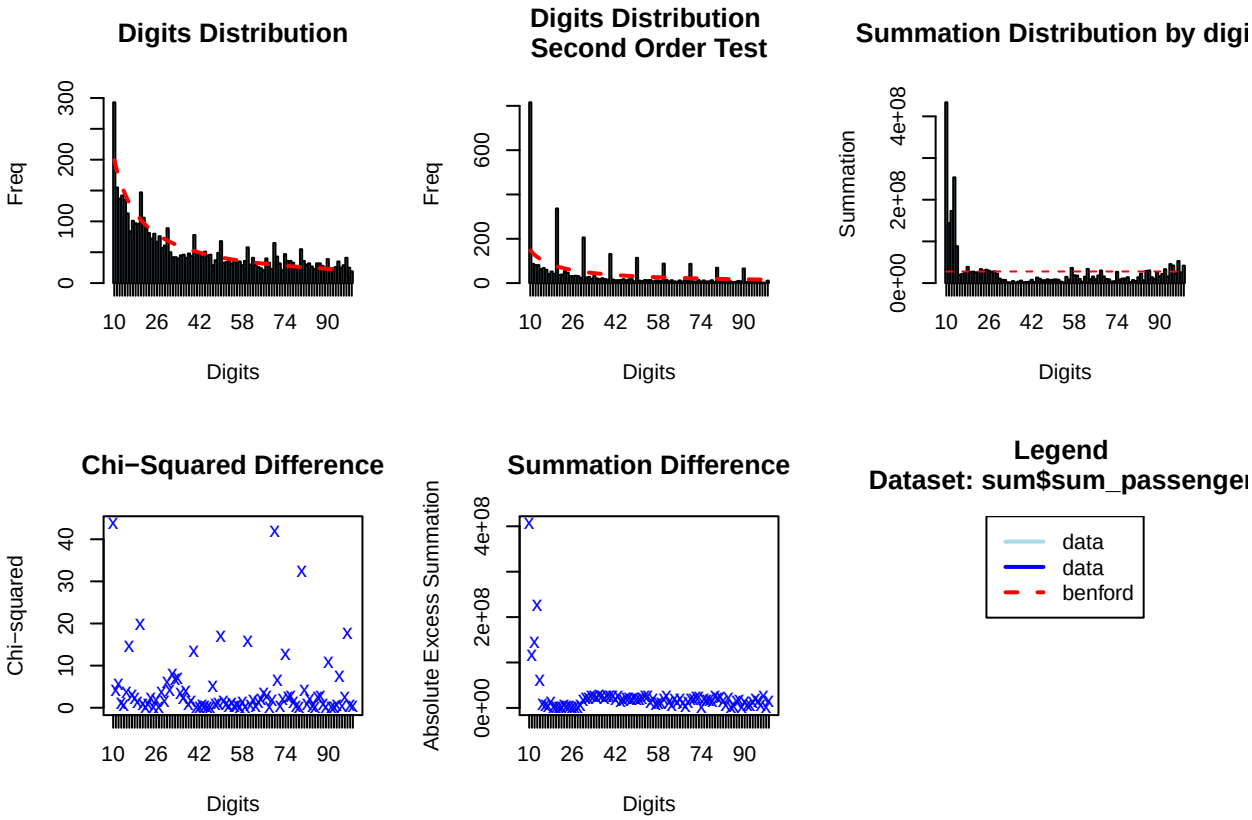
Let's take a look at 2018 May data. We can see that out of top 10 airlines, top 3 airline companies have numbers begin with 1. Is this a conincidence or not?



This is the animation

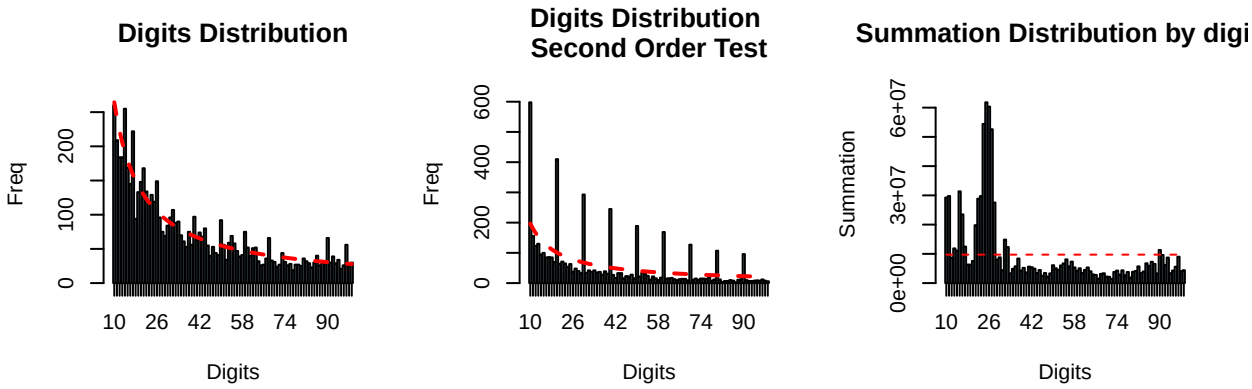
Now we will examine all the airlines by using Benford's law and find out the distribution. This plot gives an explanation of the Benford's over all of the data. For the top left figure, we can see a spike of value 1. This is probably because our data is consisted of the last 3 years history. For example, Southwest Airlines Co. has a monthly sum of passengers fluctuating around one million in the last three years.

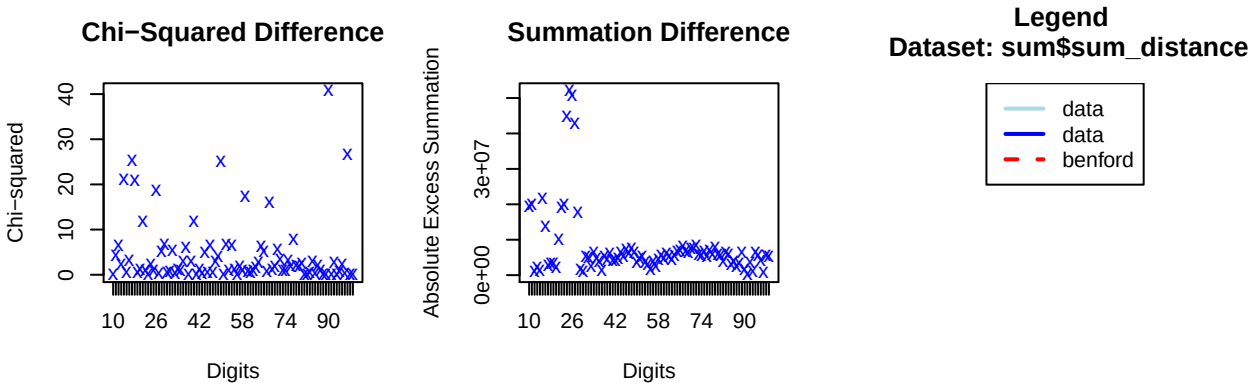
Overall, our sum of available seats data follows a distribution of Benford's law.

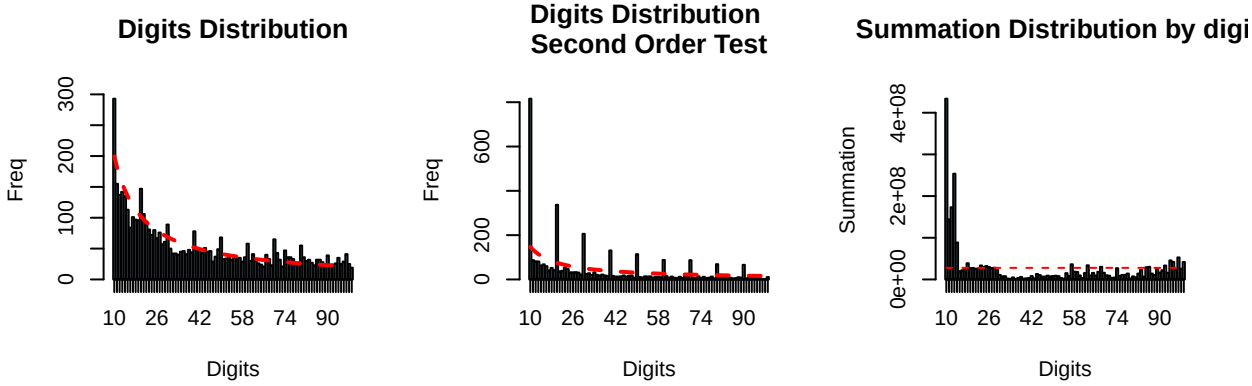


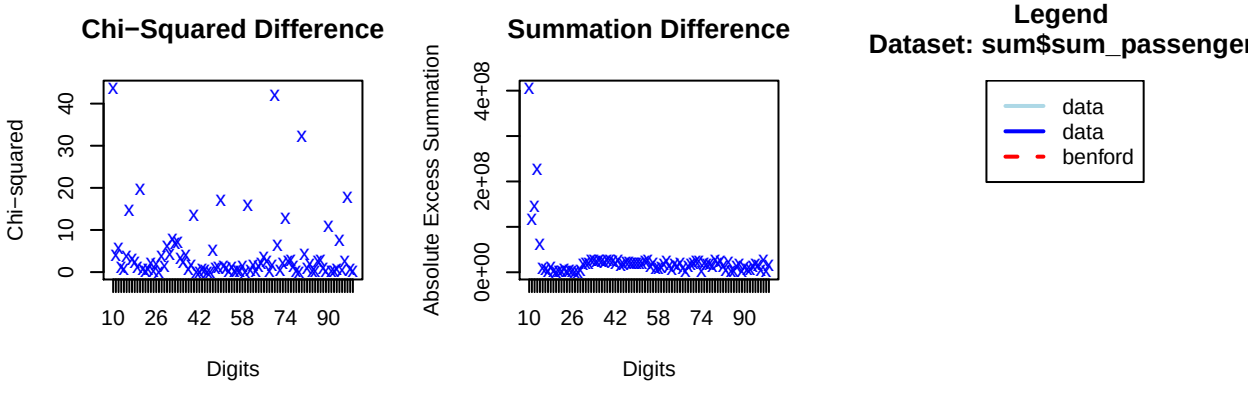

```
##     UNIQUE_CARRIER_NAME YEAR MONTH sum_seats sum_passengers sum_distance
## 1:          40-Mile Air 2015     7       381            105         2712
## 2: ACM AIR CHARTER GmbH 2015    11        58             10         6500
## 3:          Aerolitoral 2015     4       297            207          415
## 4:           Aeromexico 2015    11       160            109          258
## 5:        Air Alsie A/S 2015     4        10              2          458
##    sum_airtime       Date log_passenger log_seats log_airtime log_distance
## 1:        8163 2015-07-01      4.663439  5.945421     9.00749     7.905810
## 2:           0 2015-11-01      2.397895  4.077537     0.00000     8.779711
## 3:           0 2015-04-01      5.337538  5.697093     0.00000     6.030685
## 4:           0 2015-11-01      4.700480  5.081404     0.00000     5.556828
## 5:           0 2015-04-01      1.098612  2.397895     0.00000     6.129050
```

UNIQUE_CARRIER_NAME	YEAR	MONTH	sum_seats	sum_passengers	sum_distance	sum_airtime	Date
40-Mile Air	2015	7	381	105	2712	8163	2015-
ACM AIR CHARTER GmbH	2015	11	58	10	6500	0	2015-
Aerolitoral	2015	4	297	207	415	0	2015-
Aeromexico	2015	11	160	109	258	0	2015-
Air Alsie A/S	2015	4	10	2	458	0	2015-










```
## Warning: Missing column names filled in: 'X8' [8]
## Parsed with column specification:
```

```
## cols(
##   SEATS = col_double(),
##   PASSENGERS = col_double(),
##   DISTANCE = col_double(),
##   AIR_TIME = col_double(),
##   UNIQUE_CARRIER_NAME = col_character(),
##   YEAR = col_integer(),
##   MONTH = col_integer(),
##   X8 = col_character()
## )

## Warning: Missing column names filled in: 'X8' [8]

## Parsed with column specification:
## cols(
##   SEATS = col_double(),
##   PASSENGERS = col_double(),
##   DISTANCE = col_double(),
##   AIR_TIME = col_double(),
##   UNIQUE_CARRIER_NAME = col_character(),
##   YEAR = col_integer(),
##   MONTH = col_integer(),
##   X8 = col_character()
## )

## Warning: Missing column names filled in: 'X8' [8]

## Parsed with column specification:
## cols(
##   SEATS = col_double(),
##   PASSENGERS = col_double(),
##   DISTANCE = col_double(),
##   AIR_TIME = col_double(),
##   UNIQUE_CARRIER_NAME = col_character(),
##   YEAR = col_integer(),
##   MONTH = col_integer(),
##   X8 = col_character()
## )

## Warning: Missing column names filled in: 'X8' [8]

## Parsed with column specification:
## cols(
##   SEATS = col_double(),
##   PASSENGERS = col_double(),
##   DISTANCE = col_double(),
##   AIR_TIME = col_double(),
##   UNIQUE_CARRIER_NAME = col_character(),
##   YEAR = col_integer(),
##   MONTH = col_integer(),
##   X8 = col_character()
## )
```