Benford’s Law Test

**Benford's law**, also called the **first-digit law**, is an observation about the [frequency distribution](https://en.wikipedia.org/wiki/Frequency_distribution) of leading digits in many real-life sets of numerical [data](https://en.wikipedia.org/wiki/Data). The law states that in many naturally occurring collections of numbers, the leading [significant digit](https://en.wikipedia.org/wiki/Significant_digit) is likely to be small.[[1]](https://en.wikipedia.org/wiki/Benford%27s_law#cite_note-BergerHill2011-1) For example, in sets which obey the law, the number 1 appears as the most significant digit about 30% of the time, while 9 appears as the most significant digit less than 5% of the time. By contrast, if the digits were distributed uniformly, they would each occur about 11.1% of the time.

Let us test it.

require(MASS)

#get first digit from any numeric number

first\_digit <- function(k){

as.numeric(head(strsplit(as.character(k),'')[[1]],n=1))

}

#this get all the first digit from sets of numbers

all\_first\_digit <- function(x){

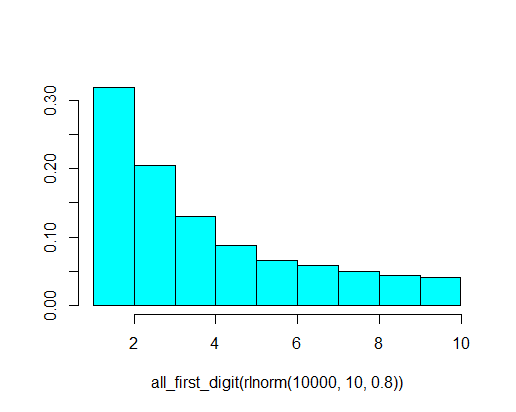
sapply(x, function(k) first\_digit(k))

}

#now let us test Benford's Law for a R random lognormal

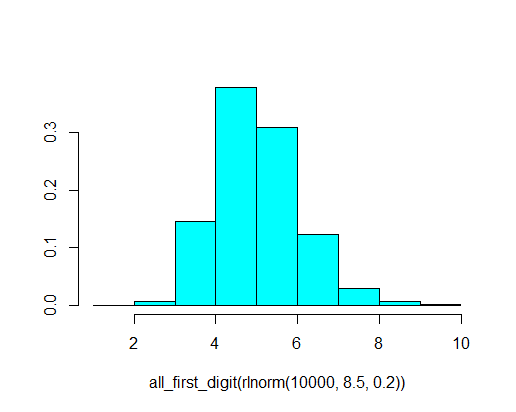
truehist(all\_first\_digit(rlnorm(10000, 10, 0.8)), nbins=10)

#From the drawing, it proves that the set of random lognoral is following Benford's Law



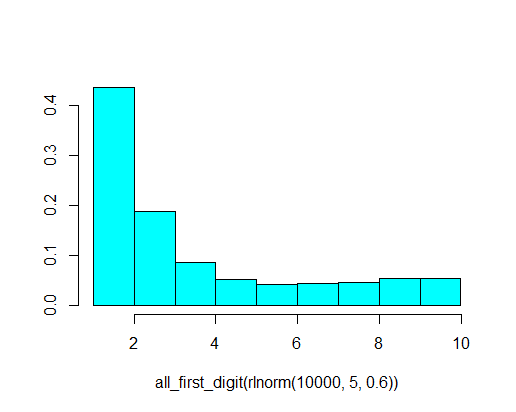
truehist(all\_first\_digit(rlnorm(10000, 8.5, 0.2)), nbins=10)

#Now we have some interesting findings, why not following the Law this time



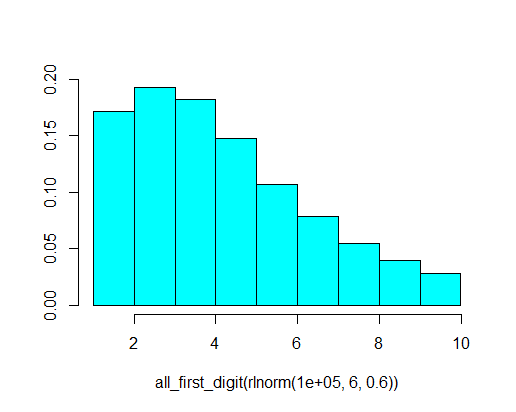
truehist(all\_first\_digit(rlnorm(10000, 5, 0.6)), nbins=10)

#Seems to me not following either, the digit 1 has more than 30%



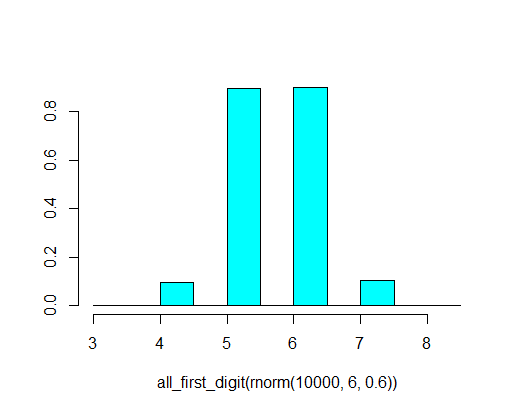
truehist(all\_first\_digit(rlnorm(10000, 6, 0.6)), nbins=10)

#NOT AGAIN, WHY?



#how about other distribution?

truehist(all\_first\_digit(rnorm(10000, 6, 0.6)), nbins=10)

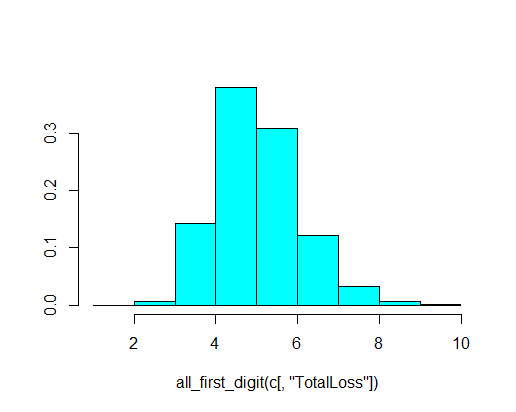


#now, let us test the Claims.csv file for GenIns triangle, I copied it to C:\ drive

c<-read.csv("C:\\Claims.csv", header=TRUE)

#let us test the TotalLoss. TotalPayment in our case is same as TotalLoss since I have no limit #and deductible. WHY NOT FOLLOWING

truehist(all\_first\_digit(c[,"TotalLoss"]), nbins=10)



**Confirmed, NOT ALL DATASET FOLLOWS BENFORD’S LAW. Lots of prove. E.g.**

[**https://punkrockor.com/2012/02/21/the-first-digits-of-many-random-numbers-are-not-so-random-on-benfords-law/**](https://punkrockor.com/2012/02/21/the-first-digits-of-many-random-numbers-are-not-so-random-on-benfords-law/)

**Now questions is: what should follow, and what not?**