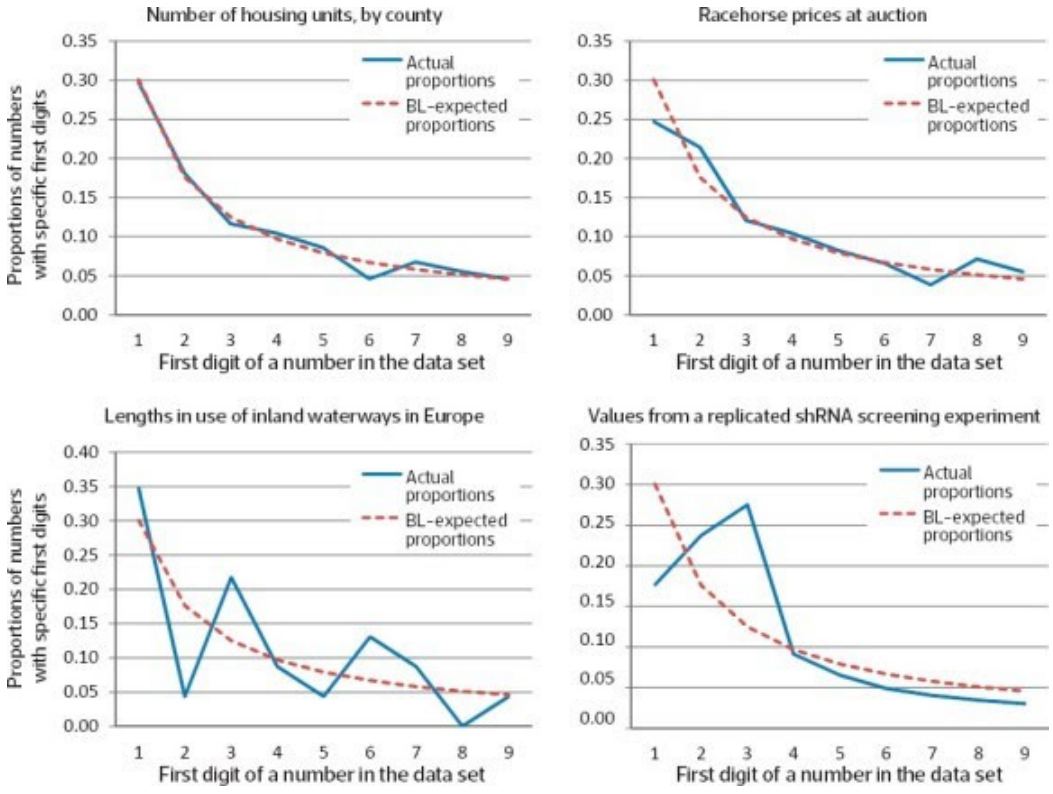
Benford's Law is one of the most underrated and widely used techniques that are commonly used in various applications. United States IRS neither confirms nor denies their use of Benford's law to detect any number of manipulations in income tax filing. Across the Atlantic, the EU is very open and proudly claims its use of Benford's law. Today, this is widely used in accounting to detect any fraud. Nigrini, a professor at the University of Cape Town, also used this law to identify financial discrepancies in Enron's financial statement. In another case, Jennifer Golbeck, a professor at the University of Maryland, was able to identify bot accounts on twitter using Benford's law. Xiaoyu Wang from the University of Winnipeg even published a report on how to use Benford's law on images. In the rest of this article, we will take about Benford's law and how it can be applied using R.

Benford's Law

> William Goodman explains "Benford's law tells us something about the frequency of leading digits in natural data sets – that is, how many numbers beginning with 1s, 2s, 3s, etc. we should expect to see. The law was first discovered by astronomer Simon Newcomb in 1881. However, in keeping with Stigler's law of eponymy – which states that no scientific discovery is named after its original discoverer – the law was named after Frank Benford, a physicist at General Electric, who rediscovered it some 50 years later."

For example, most numbers in a set (about 30%) will have a leading digit of 1, when the expected probability is 11.1% (i.e., one out of nine digits). This is followed by about 17.5%, starting with a number 2. This is an unexpected phenomenon; If all leading numbers (0 through 9) had equal probability, each would occur 11.1% of the time (StatsHowTo). Below is an example of Benford's law application in different cases.



Distributions of first-digit proportions for four Benford-suitable data sets

In the first example above, if we take all the housing units available in every country and grab the first digit of each of those numbers, they should follow Benford's law, where numbers beginning with one would have the most significant proportion followed by 2, 3 and so on. The same thing applies to different cases like race horses sold in an auction, length of inland water ways in Europe, and so on. This can also be applied to stock prices and should follow Benford's law.

Requirements for Benford's Law

Some of the guidelines suggested by William Goodman are as follows:

Sufficient sample size. In a small sample, of say 20 numbers, even two numbers more than expected starting with digit 1 would represent a seemingly noticeable (10%) divergence, yet statistically valid conclusions could not be reached about the population.

A large span of number values. Suppose that the first digits for the population of cash receipts inherently follow Benford's law, but for some reason, a sample's values range only from $20 to $500. First digit 1s would be underrepresented (only observable as first digits in the 100s range), whereas first digit 2s could be observed in the 20s *and* 200s. More orders of magnitude could minimize such problems.

Positively ("right-") skewed distributions of numbers. Data sets that conform to Benford's law often have combinatory or multiplicative origins; such as expense receipt amounts (derived from (price) × (quantity)), or values for hurricane-damage insurance claims (derived from (strength of hurricane) × (insured amounts for affected properties)). Numbers generated like this tend to have logarithmic-type distributions, with extended "right tails" of large values, and are, in turn, more likely to exhibit Benford patterns.

Not human-assigned numbers. Numbers that are merely assigned, such as arbitrarily assigned telephone numbers, or "bonus points" awarded in preset amounts of $5 or $10, tend not to exhibit Benford patterns.

US County Population Data

To test Benford's law, the first application we choose was US county population data. We will apply Benford's law for all the county populations for the first digit. If the data conforms to Benford's law, Benford function from Benford.analysis package should output the results and conforms.

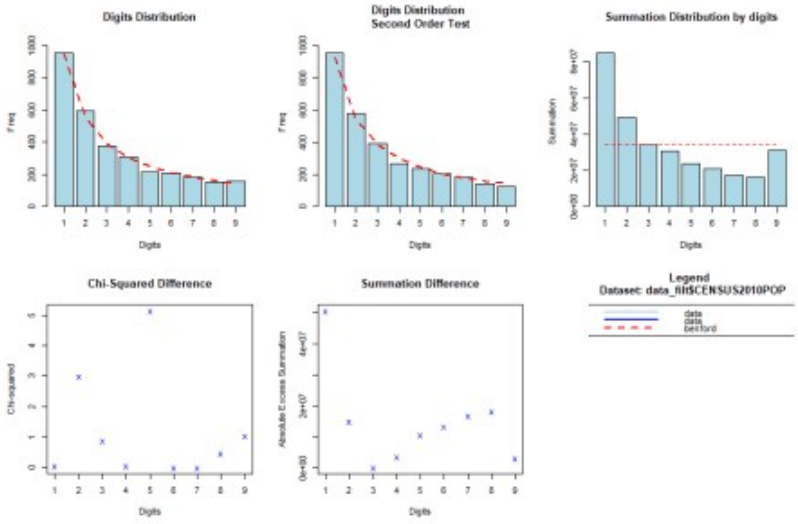
```
# load libraries
library(gtrendsR)
library(reshape2)
library(dplyr)
library(benford.analysis)

# read data
data = read.csv("https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/co-est2019-alldata.csv", header = T)

# filter out columns
data_filt = data %>% filter(COUNTY != 0) %>% select(c(STNAME, CTYNAME, CENSUS2010POP))

# perform benford analysis
trends = benford(data_filt$CENSUS2010POP, number.of.digits = 1, discrete = T, sign = "positive")
trends

# plot results
plot(trends)
```

The output of Benford's analysis is as follows. According to Nigrini (2012), it also shows that the data closely conforms. The other way the author of the package, Carlos Cinelli recommends is to look at is the Mantissa statistics and if the data follow Benford's law, it should be closer to these values in the below table:

Statistic	Value
Mean	0.4
Variance	0.083
Ex. Kurtosis	-1.2
Skewness	0

```
.

Benford object:

Data: data_filt$CENSUS2010POP
Number of observations used = 3142
Number of obs. for second order = 3089
First digits analysed = 1

Mantissa:

   Statistic  Value
        Mean  0.496
         Var  0.083
 Ex.Kurtosis -1.191
    Skewness  0.052


The 5 largest deviations:

  digits absolute.diff
1      2         40.72
2      5         35.79
3      3         18.56
4      9         12.23
5      8          8.72

Stats:

        Pearson's Chi-squared test

data:  data_filt$CENSUS2010POP
X-squared = 10.631, df = 8, p-value = 0.2235


        Mantissa Arc Test

data:  data_filt$CENSUS2010POP
L2 = 0.00077844, df = 2, p-value = 0.08665

Mean Absolute Deviation (MAD): 0.004499874
MAD Conformity - Nigrini (2012): Close conformity
Distortion Factor: -0.7996047

Remember: Real data will never conform perfectly to Benford's Law. You should
not focus on p-values!
```

You can also visualize the data, as shown below. You should see that the data for the first digits in the data set (blue lines) would follow Benford's curve (red line). You can also notice that the second order test also seems to follow Benford's curve.



Applying on Image

Benford's law can not only be applied to naturally occurring data but could also be applied to images. In recent years, we have seen this mainly used in forensic image analysis to see if the photos were tampered. In a report published by Xiaoyu Wang, the author also identifies the scope and limitations of applying this technique to images.

> Jolin (2001) was the first one considering to apply Benford's law into image processing and discovered the gradient of pixel values were follows
> Benford's law. Acebo and Sbert (2005) pointed out that under certain constraints, the intensity of the digital image agrees with Benford's law. Perez (2007) presented a generalization of Benford's law and proved it applies to the discrete cosine transform (DCT) domain. Fu and Shi (2007) showed a method using Benford's law to detect JPEG compression. Jingwei and Byung (2009) put forward some vulnerabilities of applying Benford's law into image forensics.

When applying Benford's law to images, the process is slightly different. The steps are as follows:

- Load the image to R workspace
- Perform discrete cosine transform on the data
- Then test the transformed data for Benford's law.

The image used in this analysis is as shown below. If you are wondering what this delicious dish is, it's "Butter Chicken Naan Pizza."

```
# load libraries
library(imagerExtra)
library(imager)
library(benford.analysis)
library(imagerExtra)
```

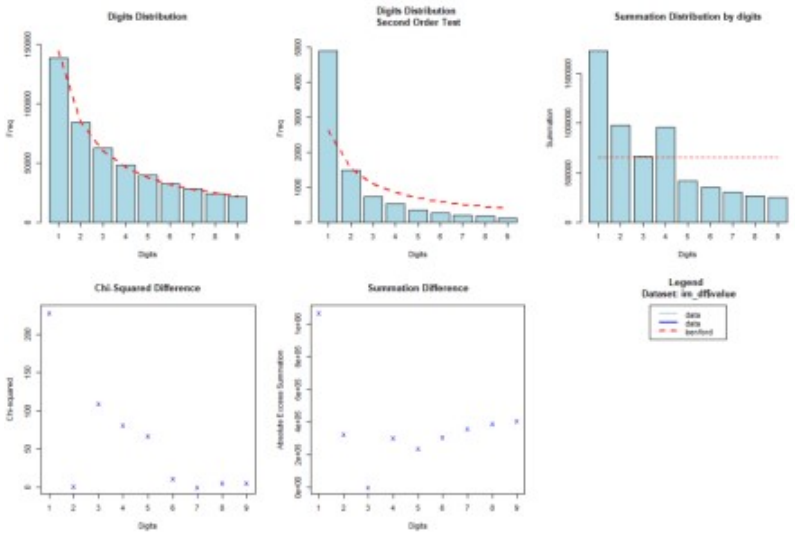
```
library(dplyr)

# load image
im = load.image("C:/save3.jpg") %>% grayscale()

# perform (DCT)
im_df = DCT2D(im) %>% as.data.frame()

# apply benford law
bfd.im = benford(im_df$value, number.of.digits = 1, discrete = T, round = 1, sign
= "both")
bfd.im

# plot the results
plot(bfd.im)
```

The output results from Benford function are as shown below. From the below results, we can see that MAD conformity is “Acceptable conformity,” and Mantissa scores are very close to the values from the above table and is a very close fit to Benford’s law.

```
Benford object:

Data: im_df$value
Number of observations used = 9144576
Number of obs. for second order = 63857
First digits analysed = 1

Mantissa:

   Statistic  Value
        Mean  0.486
         Var  0.082
 Ex.Kurtosis -1.171
    Skewness  0.086


The 5 largest deviations:

  digits absolute.diff
1      1     135961.33
2      2     106406.10
3      5      62700.92
4      6      56573.01
5      4      48160.98

Stats:

        Pearson's Chi-squared test

data:  im_df$value
X-squared = 32147, df = 8, p-value < 2.2e-16


        Mantissa Arc Test

data:  im_df$value
L2 = 0.0019564, df = 2, p-value < 2.2e-16
```

```
Mean Absolute Deviation (MAD): 0.006024838
MAD Conformity - Nigrini (2012): Acceptable conformity
Distortion Factor: -21.76257

Remember: Real data will never conform perfectly to Benford's Law. You should
not focus on p-values!
```

The below plot shows the distribution of the first and second digits. In both cases, we can see that the data (in blue) fits with Benford's law (in red).



Applying on Music

Music is everywhere. We hear music on TV, radio, our phones, car, restaurant and any place you can think. We are surrounded by music. Music is made up of oscillating signals that are usually collected at 44kHz. What it means is we collect about 44 thousand samples every second. It was also theorized and shown that even in the music, we see Benford's law. In a paper authored by Simth (1997), recommended to transform the data using Fourier transforms for better conformity. To test this theory, we need to download some music. I downloaded "Lesser Faith" by J. Syreus Bach from the below link.

https://files.freemusicarchive.org/storage-freemusicarchive-org/music/no_curator/J_Syreus_Bach/Ability_to_Break__Energetic_Tracks/J_Syreus_Bach_-_Lesser_Faith.mp3

The music file is in MP3, so extract the numeric value; we will convert to Wave file and load the wave file. Next, we transform the data using FFT and modulus to convert complex numbers to a numeric value. Finally, we can apply this to Benford function.

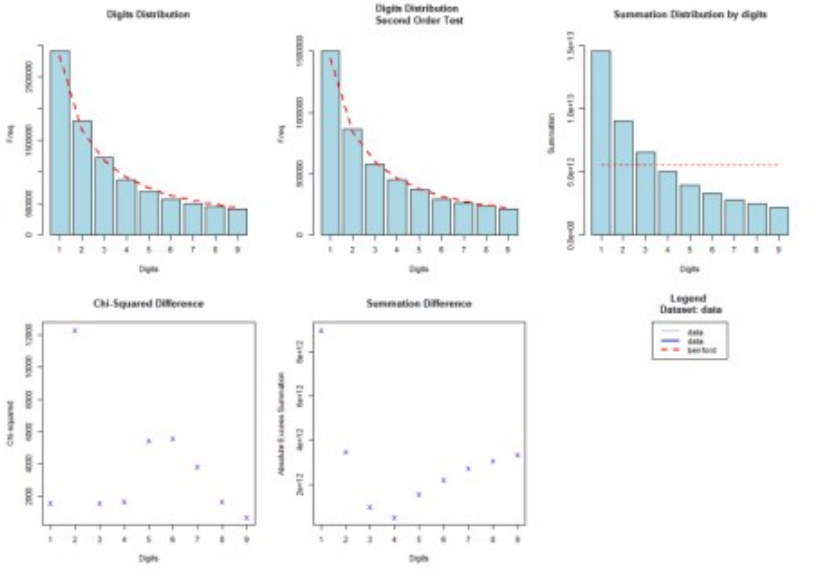
```
# load packages
library(dplyr)
library(tuneR)
library(benford.analysis)

# read mp3 and convert to wave
r = readMP3("J_Syreus_Bach_-_Lesser_Faith.mp3")
writeWave(r,"tmp.wav",extensible=FALSE)

# read wave
w = readWave("tmp.wav")

# take left data, perform fft and Mod
data = w@left %>% fft() %>% Mod()

# perform benford analysis
trends = benford(data, number.of.digits = 1, discrete = T, sign = "both")
```

```
trends

# plot results
plot(trends)
```

The results of the analysis are as shown below. We can see that there is close conformity within the data based on MAD, and Mantissa values are also closer to the values from the above table.

```
Benford object:

Data: data
Number of observations used = 9422208
Number of obs. for second order = 4780125
First digits analysed = 1

Mantissa:

   Statistic  Value
        Mean  0.488
         Var  0.081
 Ex.Kurtosis -1.152
    Skewness  0.074


The 5 largest deviations:

  digits absolute.diff
1      2     143035.53
2      1      67357.77
3      5      63810.17
4      6      59424.58
5      7      45996.19

Stats:

        Pearson's Chi-squared test

data:  data
X-squared = 34487, df = 8, p-value < 2.2e-16


        Mantissa Arc Test

data:  data
L2 = 0.0019376, df = 2, p-value < 2.2e-16

Mean Absolute Deviation (MAD): 0.005986932
MAD Conformity - Nigrini (2012): Close conformity
Distortion Factor: -2.841123

Remember: Real data will never conform perfectly to Benford's Law. You should
not focus on p-values!>
```

Finally, we can look at the digit distribution. From the results, we see the data follows Benford's law, where digit 1 has more frequent occurrence over digit 2 and so on.



Conclusion

Benford’s law is everywhere around our life and also pretty sneaky hiding in plain sight. We see them when we go grocery shopping, restaurant menu, filing taxes, listing to music, taking photos, and many more. In light of the Central limit theorem, Benford’s law doesn’t get enough credit. In this article, we just saw just some highlights of Benford’s law in a few cases. There are still a lot more cases that we could use to our advantage, like identifying bot accounts on Twitter and Facebook or identifying modified images or music. The applications of Benford’s law are many.