One of my starting points into quantitive finance was Bernie Madoff's fund. Back then because Bernie was in desperate need of money to keep his *Ponzi scheme* running there existed several so-called feeder funds.

One of them happened to approach me to offer me a once in a lifetime investment opportunity. Or so it seemed. Now, there is this old saying that when something seems too good to be true it probably is. If you want to learn what *Benford's law* is and how to apply it to uncover fraud, read on!

Here are Bernie's monthly returns (you can find them here: madoff_returns.csv):

```
madoff_returns <- read.csv("Data/madoff_returns.csv")
equity_curve <- cumprod(c(100, (1 + madoff_returns$Return)))
plot(equity_curve, main = "Bernie's equity curve", ylab = "$", type =
"l")
```
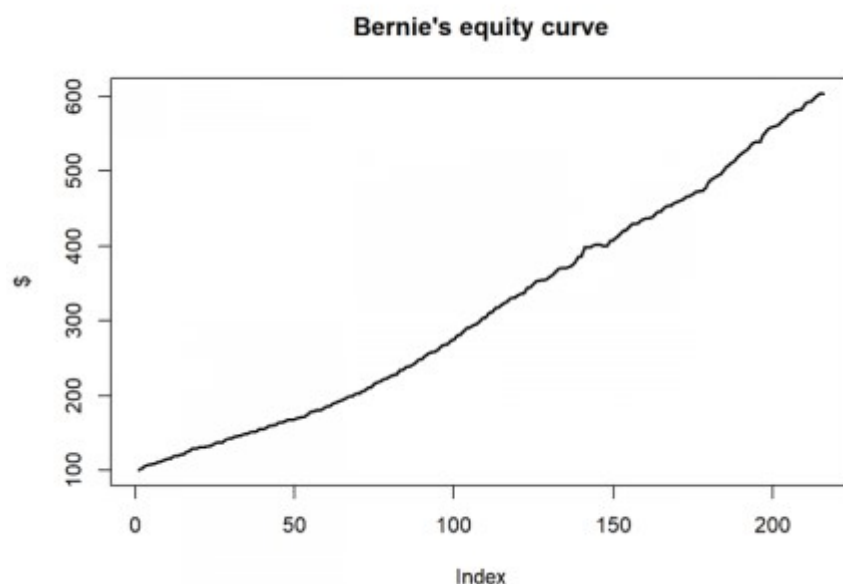

An equity curve with annual returns of over 10% as if drawn with a ruler! Wow… and Double Wow! What a hell of a fund manager!

I set off to understand how Bernie accomplished those high and especially extraordinarily stable returns. And found: Nothing! I literally rebuilt his purported split-strike strategy and backtested it, it of course didn't work. And therefore I didn't invest with him. A wise decision as history proved. And yet, I learned so much along the way, especially on trading and options strategies.

A very good and detailed account of the Madoff fraud can be read in the excellent book "No One Would Listen: A True Financial Thriller" by whistleblower Harry Markopolos who was on Bernie's heels for many years but as the title says, no one would listen… The reason is some variant of the above wisdom "What seems to good…": people told him that Bernie could not be a fraud because his fund was so big and other people would have realized that!

One of the red flags that those returns were made up could have been raised by applying Benford's law. It states that the frequency of the leading digits of many real-world data sets follows a very distinct pattern:

```
theory <- log10(2:9) - log10(1:8)
```

```
theory <- round(c(theory, 1-sum(theory)), 3)
data.frame(theory)
##   theory
## 1  0.301
## 2  0.176
## 3  0.125
## 4  0.097
## 5  0.079
## 6  0.067
## 7  0.058
## 8  0.051
## 9  0.046
```

The discovery of Benford's law goes back to 1881 when the astronomer Simon Newcomb noticed that in logarithm tables the earlier pages were much more worn than the other pages. It was re-discovered in 1938 by the physicist Frank Benford and subsequently named after him. Thereby it is just another instance of *Stigler's law* which states that no scientific discovery is named after its original discoverer (Stigler's law is by the way another instance of Stigler's law because the idea goes back at least as far as to Mark Twain).

Thie following analysis is inspired by the great book "Analytics Stories" by my colleague Professor em. Wayne L. Winston from Kelley School of Business at Indiana University. Professor Winston gives an insightful explanation of why Benford's law holds for many real-world data sets:

> Many quantities (such as population and a company's sales revenue) grow by a similar percentage (say, 10%) each year. If this is the case, and the first digit is a 1, it may take several years to get to a first digit of 2. If your first digit is 8 or 9, however, growing at 10% will quickly send you back to a first digit of 1. This explains why smaller first digits are more likely than larger first digits.

We are going to simulate a growth process by sampling some random numbers as a starting value and a growth rate and letting it grow a few hundred times, each time extracting the first digit of the resulting number, tallying everything up, and comparing it to the above distribution at the end:

```
# needs dataframe with actual and theoretic distribution
plot_benford <- function(benford) {
  colours = c("red", "blue")
  bars <- t(benford)
  colnames(bars) <- 1:9
  barplot(bars, main = "Frequency analysis of first digits", xlab =
"Digits", ylab = "Frequency", beside = TRUE, col = colours, ylim=c(0,
max(benford) * 1.2))
  legend('topright', fill = colours, legend = c("Actual", "Theory"))
}

set.seed(123)
start <- sample(1:9000000, 1)
growth <- 1 + sample(1:50, 1) / 100

n <- 500
sim <- cumprod(c(start, rep(growth, (n-1)))) # vectorize recursive
```

```
simulation
first_digit <- as.numeric(substr(sim, 1, 1))
actual <- as.vector(table(first_digit) / n)
benford_sim <- data.frame(actual, theory)
benford_sim
##   actual theory
## 1  0.300  0.301
## 2  0.174  0.176
## 3  0.126  0.125
## 4  0.098  0.097
## 5  0.078  0.079
## 6  0.068  0.067
## 7  0.058  0.058
## 8  0.050  0.051
## 9  0.048  0.046

plot_benford(benford_sim)
```



Frequency analysis of first digits
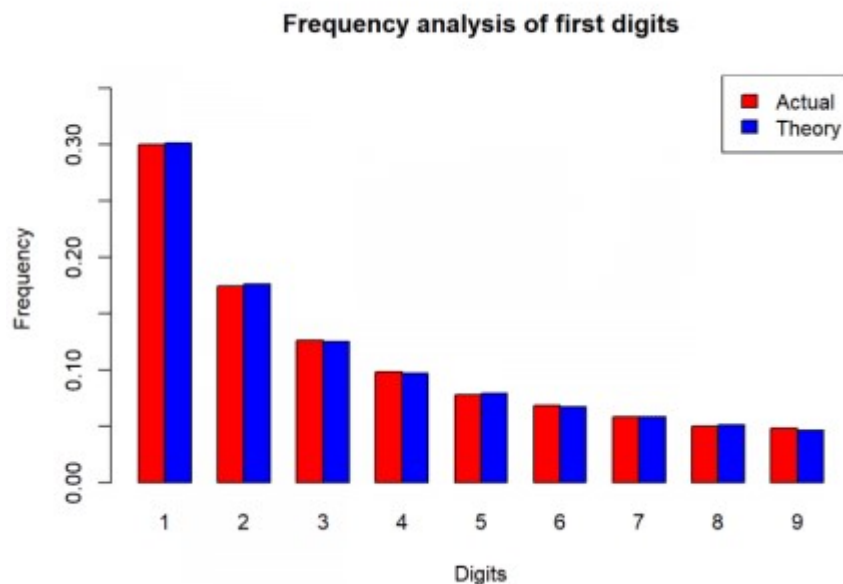
We can see a nearly perfect fit!

We are now doing the same kind of analysis with Bernie's made-up returns:

```
first_digit <- as.numeric(substr(abs(madoff_returns$Return * 10000), 1,
1))
actual <- round(as.vector(table(first_digit) / length(first_digit)), 3)
madoff <- data.frame(actual = actual[2:10], theory)
madoff
##   actual theory
## 1  0.391  0.301
## 2  0.135  0.176
## 3  0.093  0.125
## 4  0.060  0.097
## 5  0.051  0.079
## 6  0.079  0.067
## 7  0.065  0.058
```

```
## 8  0.070  0.051
## 9  0.051  0.046
```

```
plot_benford(madoff)
```

Frequency analysis of first digits
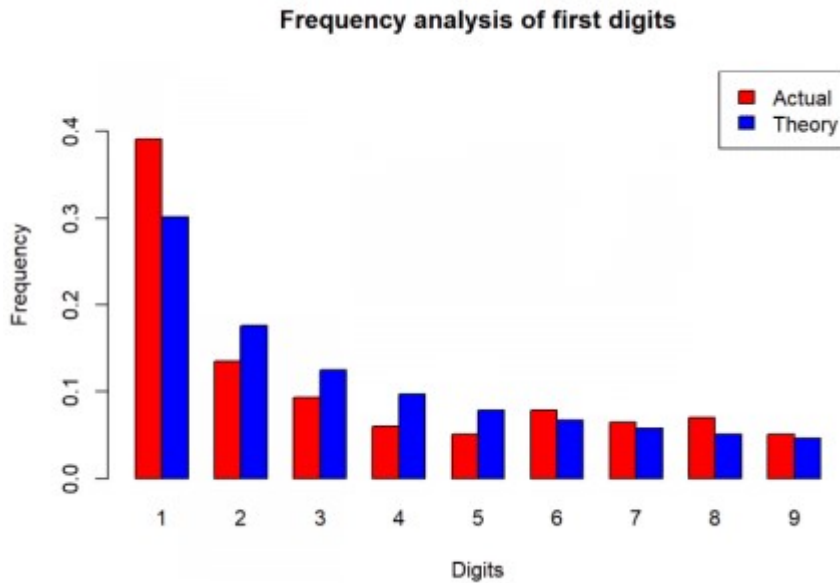
Just by inspection, we can see that something doesn't seem to be quite right. This is of course no proof but another indication that something could be amiss.

Benford's law has become one of the standard methods used for fraud detection, *forensic analytics* and *forensic accounting* (also called *forensic accountancy* or *financial forensics*). There are several R packages with which you can finetune the above analysis, yet the principle stays the same. Because this has become common knowledge many more sophisticated fraudsters tailor their numbers according to Benford's law so that it may become an instance of yet another law: *Goodhart's law*:

> Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.

Let us hope that this law doesn't lead to more lawlessness!