Stockholms universitet

Fraud Detection with Benford's Law

Simon Tufvesson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2021:10
Matematisk statistik
Juni 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2021:10**
http://www.math.su.se

Fraud Detection with Benford's Law

Simon Tufvesson*

June 2021

Abstract

Benford's Law describes how the digits in sets of numerical data should be distributed. When there is a significant deviation from Benford's Law in a data set it could indicate that the data has been manipulated or made up. This method has been increasingly popular in recent years for detecting fraud in different areas of interests, such as accounting, elections, scientific data, etc. The purpose of this study is to see if we can detect any fraud in some different areas and how well Benford's Law performs. Benford's Law was applied on basic simulations of common distributions, economic data on EU during the Greece government-debt crisis, and a simulation from a two candidate election model. This resulted in large deviations from Benford's Law when not expecting it and almost no deviation when it was expected. Hence we conclude that either the theory of Benford's Law is incomplete so that its appropriate use is still to a large extent unknown, or that the commonly used hypothesis tests are not oprimal for Benford's Law and should be corrected or other tests should be developed for this purpose.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: situ9231@student.su.se. Supervisor: Ola Hössjer, Kristofer Lindensjö.

Acknowledgement

This is a bachelor's thesis of 15 ECTS in Mathematical Statistics at the Department of Mathematics at Stockholm University. First, I want to thank my supervisors Ola Hössjer and Kristofer Lindensjö, Professor and Associate Professor at the Department of Mathematical Statistics, for the invaluable theoretical discussions, advice and guidance in mathematical statistics with great commitment and encouragement. Further, I also want to pay attention to my fellow students for the rewarding discussions and suggestions. I also want to express my gratitude to my family and friends who have supported me throughout my education and my work on this thesis.

Contents

Notation

The following notation and definitions will be used throughout this report:

- We will denote the logarithm of x in base 10 with $\log(x)$, while the natural logarithm of x will be notated with $\ln(x)$, and $\log_b(x)$ will denote the logarithm of x in base b.
- The *floor function* $\lfloor x \rfloor$ is defined as the integer nearest x rounded down.
- Type-I error is the rejection probability of a true null hypothesis and Type-II error is the non-rejection probability when the null hypothesis is false.
- A folded normal distribution $|N(\mu, \sigma^2)|$ is the distribution such that given a normally distributed random variable X with mean μ and variance σ^2, the random variable $Y = |X|$ has a folded normal distribution.

1 Introduction

Intuitively we expect that all digits in a collection of numbers should occur with equal frequency. That is to say that the first digit should be $1, 2, \ldots, 9$ with frequency $\frac{1}{9}$, and the second digit, third digit, etc. should take their value with equal frequencies $\frac{1}{10}$. We do this because we expect that the digits in a collection of numbers should be random and not follow any specific pattern. This however is something the astronomer Simon Newcomb (1881)[1] noticed wasn't true. His discovery did not have much resonance and was forgotten until 1938 when the physisist Frank Benford rediscoverd Newcomb's result, which since has been known as *Benford's Law*. Benford (1938)[2] collected a great amount of data from different areas so diverse as numbers of inhabitants of towns, physical measurements, results from sport leagues, etc. For most of his data he found that the frequency of the first digit D_1 very closely followed the logarithmic law

$$\mathbb{P}(D_1 = d) = \log\left(1 + \frac{1}{d}\right), \qquad d = 1, 2, \ldots, 9.$$

After Benford's paper was published the law got a lot of attention and people tried to figure out some useful application. That's when in 1998, Mark Nigrini [3] proposed that Benford's Law should be used as an auditing and accounting tool to detect irregularities of companies data. He found that most accounting data very closely follow Benford's Law. However in the case of accounting fraud this was often not the case. After this discovery the question whether Benford's Law could be used in more fraudulent cases was raised, and has since then been used in many fields like elections, scientific data and much more. In this paper we are going to test if we can detect some fraud in a few cases with the help of Benford's Law. First we are going to present some mathematical theory behind Benford's Law and then test the theory on some basic simulations to see if the theory holds. After that we are going to test the method on a real dataset, the Greek government-debt crisis (European Commission, 2010)[10], and also simulate from a two candidate election model, with fraud introduced, to see if Benford's Law actually can detect these frauds.

2 The mathematical theory

In this section we will aim to formally define Benford's Law and show some basic results on the involved distributions. Everything stated here is based on the theory in *An introduction to Benford's Law* (Berger and Hill, 2015)[4] and *Benford's Law: Theory and Applications* (Miller, 2015)[5].

2.1 Significant digits and the significand

Since Benford's Law is all about the statistical distribution of significant digits, a natural starting point for any study of Benfors's Law is the formal definition of *significant digits* and the *significand (function).*

2.1.1 Significand digits

Informally we might say that the *first significant decimal digit* of a positive real number x is the first non-zero digit appearing in a decimal expansion of x. For example the first significant digit of 2021 and 0.2021 are both 2. The *second significant digit* is the digit directly following the first significant digit, so the second significant digit of 2021 and 0.2021 are both 0. More formally we use the definition in *An introduction to Benford's Law* (Berger and Hill, 2015, p. 11)[4] to define the first significant digit:

Definition 2.1. *For every non-zero real number x, the **first significant (decimal) digit** of x, denoted by $D_1(x)$, is the unique integer $j \in \{1, 2, ..., 9\}$ satisfying $10^k j \leq |x| < 10^k(j+1)$ for some unique integer k; for convenience we define $D_1(0) := 0$.*

*Similarly, for every integer $m \geq 2$, we have the the **m^{th} significant (decimal) digit** digit of x, denoted by $D_m(x)$, is defined inductively as the unique integer $j \in \{0, 1, 2, ..., 9\}$ such that*

$$10^k \left(\sum_{i=1}^{m-1} D_i(x) 10^{m-i} + j \right) \leq |x| < 10^k \left(\sum_{i=1}^{m-1} D_i(x) 10^{m-i} + j + 1 \right)$$

for some unique integer k; for convenience we define $D_m(0) := 0$ for all positive integers m.

Note that the main difference between the first significant digit $D_1(x)$ and the m^{th} significant digit, is that for any non-zero x, $D_1(x)$ is never zero, whereas the other significant digits may be any integers in the set $\{0, 1, 2, \ldots, 9\}$.

Example 1.

$$\begin{aligned} D_1(\pi) &= D_1(-\pi) = D_1(10\pi) = 3, \\ D_2(\sqrt{2}) &= 4, \\ D_3(\sqrt{2}) &= 1. \end{aligned}$$

2.1.2 The Significand

The *significand function* plays a fundamental role in the context of Benford's Law. The *significand* of a real number is its coefficient when it is

expressed in floating point form ("scientific notation"), formally taken from *An introduction to Benford's Law* (Berger and Hill, 2015, p. 12)[4]:

Definition 2.2. *The **(decimal) significand function** $S : \mathbf{R} \to [1, 10)$ is defined as follows: If $x \neq 0$ then $S(x) = t$, where t is the unique number in $[1, 10)$ with $|x| = 10^k t$ for some unique integer k; if $x = 0$ then, for convenience, $S(0) := 0$.*

The unique representation of $x = S(x) \cdot 10^k$ implies $S(x) = x \cdot 10^{-k}$. But for $x > 0$ it holds that $x = 10^{\log x}$. Therefore we get the following explicit representation of the significand function valid for all $x \neq 0$:

$$S(x) = 10^{\log |x| - \lfloor \log |x| \rfloor}. \tag{1}$$

Since $\log |x| - \lfloor \log |x| \rfloor$ removes the integer part of $\log |x|$, and then inverting it back from the logarithm we get that $S(x) \in [1, 10)$, as desired.

Example 2.

$$\begin{aligned} S(\pi) &= S(-\pi) = S(10\pi) = \pi \approx 3.1415, \\ S\left(\tfrac{1}{\sqrt{2}}\right) &= S\left(\tfrac{10}{\sqrt{2}}\right) = \tfrac{10}{\sqrt{2}} \approx 7.0711. \end{aligned}$$

Berger and Hill (2015, p. 13)[4] also propose that the significand uniquely determines the significant digits and vice versa. We can see this relationship in Proposition 2.1 which follows directly from the Definitions 2.1 and 2.2.

Proposition 2.1. *For every real number x,*

1. $S(x) = \sum_{m \in \mathbb{Z}_+} 10^{1-m} D_m(x)$
2. $D_m(x) = \lfloor 10^{m-1} S(x) \rfloor - 10 \lfloor 10^{m-2} S(x) \rfloor$, *for every* $m \in \mathbb{Z}_+$

Example 3. It follows from Proposition 2.1 that

$$S(\pi) = D_1(\pi) + 10^{-1} D_2(\pi) + 10^{-2} D_3(\pi) + \ldots = \pi \approx 3.1415$$

and

$$\begin{aligned} D_1(\pi) &= \lfloor \pi \rfloor = 3 \\ D_2(\pi) &= \lfloor 10\pi \rfloor \end{aligned}$$

Now that we have established some basic tools that we can use, we can start defining some Benford properties.

2.2 The Benford property

Benford's Law is the observation that for many collections of numbers, whether they are mathematical tables, real-life data or some combination thereof, the distribution of digits is not uniform as one might expect. They are heavily skewed to the smaller digits. More specifically they follow a unique logarithmic pattern. In this section we will formally go through where this logarithmic pattern comes from.

2.2.1 Relationship to the uniform distribution

We will now introduce randomness into the mathematical model.

Definition 2.3. *Let X be a random variable. We say that X satisfies* ***Benford's Law*** *if $S(X)$ follows a* ***logarithmic distribution***

$$\mathbb{P}(S(X) \leq t) = \log(t), \quad t \in [1, 10) \tag{2}$$

It is easy to see that the logarithmic law (2) holds if and only if the logarithm of $S(X)$ follows a continuous uniform ditribution on $[0, 1]$. We can see this if we substitute $\log t = s$ in equation (2). Then we get

$$\mathbb{P}(S(X) \leq 10^s) = s, \quad s \in [0, 1).$$

Upon taking the logarithm we get

$$\mathbb{P}(\log(S(X)) \leq s) = s, \quad s \in [0, 1).$$

Thus we realise that $\log S(X)$ follows a continuous uniform distribution on the interval [0,1].

These observations tell us that we will observe Benford's Law, if the logarithms of the significands of the observed data are close to a uniform distribution.

2.2.2 Benford's First Digit Law

Let's consider the event $\{D_1(x) = d_1\}$. Then from Proposition 2.1 we have

$$\begin{aligned}\{D_1(x) = d_1\} &= \{\lfloor S(x) \rfloor - 10\lfloor 10^{-1} S(x) \rfloor = d_1\} \\ &= \{\lfloor S(x) \rfloor = d_1\} \\ &= \{d_1 \leq S(x) < d_1 + 1\}.\end{aligned}$$

Thus, from the logarithmic distribution (2), we get for a random variable X that follows Benford's Law

$$\begin{aligned}\mathbb{P}(D_1(X) = d_1) &= \mathbb{P}(d_1 \leq S(X) < d_1 + 1) \\ &= \mathbb{P}(S(X) \leq d_1 + 1) - \mathbb{P}(S(X) \leq d_1) \\ &= \log(d_1 + 1) - \log(d_1) = \log\left(\frac{d_1 + 1}{d_1}\right) \\ &= \log\left(1 + \frac{1}{d_1}\right).\end{aligned}$$

Proposition 2.2. *The first significant digit of a random variable $D_1(X)$ has the probability function*

$$\mathbb{P}(D_1(X) = d_1) = \log\left(1 + \frac{1}{d_1}\right),$$

where $d_1 \in \{1, \ldots, 9\}$.

In Figure 1 we illustrate a representation of the distribution of the first significant digit.

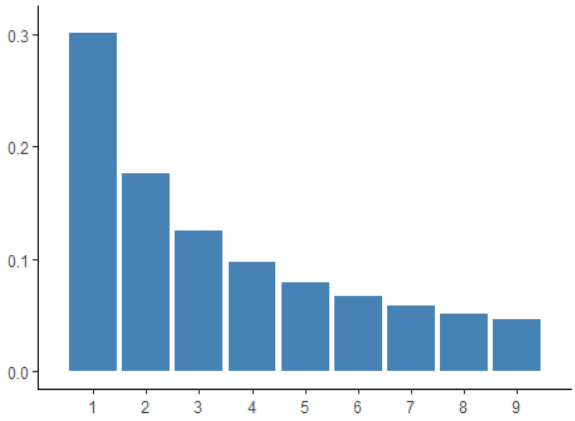


Figure 1: Benford's Law - first significant digit probabilities.

2.2.3 Benford's General Digit Law

In a similar way as we did to get Benford's first digit law we get a probability distribution for all the first k significant digits.

Proposition 2.3. *The joint distribution of the first significant digits of a random variable $D_1(X), D_2(X), \ldots, D_k(X)$ (for every $k \in \mathbb{Z}_+$) is*

$$\mathbb{P}(D_1(X) = d_1, \ldots, D_k(X) = d_k) = \log\left(1 + \left(\sum_{i=1}^{k} 10^{k-i} d_i\right)^{-1}\right),$$

where $d_1 \in \{1, \ldots, 9\}$*, and all the other* $d_j \in \{0, \ldots, 9\}$*.*

From Proposition 2.3 we can determine the distribution of the k^{th} significant digit by taking the marginal distribution. For example, to obtain the distribution of the second significant digit, we take the marginal distribution

$$\begin{aligned}
\mathbb{P}(D_2(X) = d_2) &= \sum_{k=1}^{9} \mathbb{P}(D_1(X) = k, D_2(X) = d_2) \\
&= \sum_{k=1}^{9} \log\left(1 + \left(10^{2-1}k + 10^{2-2}d_2\right)^{-1}\right) \\
&= \sum_{k=1}^{9} \log\left(1 + \frac{1}{10k + d_2}\right).
\end{aligned}$$

Generalising this we get

$$\begin{aligned}
\mathbb{P}(D_k(X) = d_k) &= \sum_{i_1=1}^{9} \sum_{i_2=0}^{9} \cdots \sum_{i_{k-1}=0}^{9} \mathbb{P}(D_1(X) = i_1, D_2(X) = i_2, \ldots, D_k(X) = d_k) \\
&= \sum_{i_1=1}^{9} \sum_{i_2=0}^{9} \cdots \sum_{i_{k-1}=0}^{9} \log\left(1 + \left(\sum_{n=1}^{k-1}(10^{k-n}i_n) + d_k\right)^{-1}\right) \\
&= \sum_{j=10^{k-2}}^{10^{k-1}-1} \log\left(1 + \frac{1}{10j + d_k}\right),
\end{aligned}$$

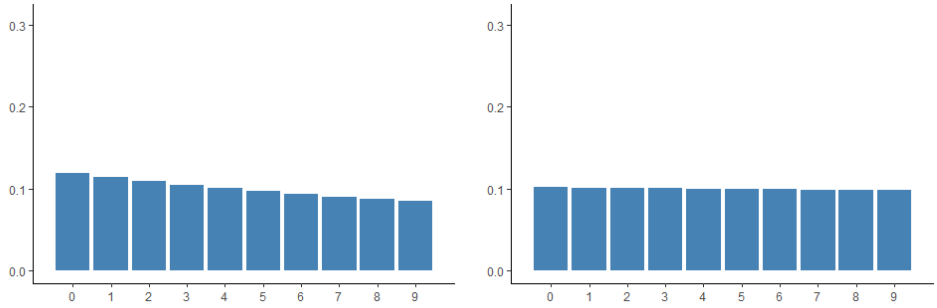
by noticing in the last step that $i_1, i_2, \ldots, i_{k-1}$ are the $k-1$ first digits in the number $\sum_{n=1}^{k-1}(10^{k-n}i_n)$, so taking the $k-1$ sums for each digit is an iterative way of taking the sum from the lowest to the highest number with $k-1$ digits. We summarize our findings for higher order digits as follows:

Proposition 2.4. *The distribution of the* k^{th} *significant digit of a random variable* $D_k(X)$ *(for every* $k \in \mathbb{Z}_+ \setminus \{1\}$*) is*

$$\mathbb{P}(D_k = d_k) = \sum_{j=10^{k-2}}^{10^{k-1}-1} \log\left(1 + \frac{1}{10j + d_k}\right)$$

where $d_k \in \{0, \ldots, 9\}$*.*

In Figure 2 we illustrate a representation of the distribution of the second (Figure 2a) and the third (Figure 2b) significant digit. Observe that the distribution of D_k approaches a discrete uniform distribution on $\{0, \ldots, 9\}$ as k grows.



(a) Second digit (b) Third digit

Figure 2: Benford's Law - second and third significant digit probabilities.

2.2.4 When can we expect to find the Benford distribution?

Benford's Law is still somewhat mysterious and it is not fully clear when a distribution should follow Benford's Law or not. However, Durtschi, Hillison and Pacini (2004, p. 24)[6] suggested a number of criteria as to when the distribution is expected to follow Benford's law, and when it is not expected to follow this law. These criteria are developed for accounting data but they may indicate when the law is applicable in other fields too.

Distributions that can be expected to obey Benford's Law:

- Sets of numbers that result from a mathematical combination of numbers, so that the result is a merge of two distributions
- Transaction-level data
- On large data sets
- When the mean is greater than the median and the skewness is positive

Distributions that would not be expected to obey Benford's Law:

- Settings where numbers are assigned sequentially
- Settings where numbers are influenced by human thought
- Accounts with a large number of firm-specific numbers
- Accounts with a built-in minimum or maximum

Empirically, Benford's Law applies if the distribution being tested fits the *Benford's Law Compliance Theorem*, Smith (1997)[7].

Theorem 2.1 (Benford's Law Compliance Theorem)**.** *Let P be a random process generating numbers in base B on the positive real line. Let $pdf(g)$ be its probability density function of $\log_B(P)$, expressed on the base B logarithmic number line, and $PDF(f)$ the Fourier transform of $pdf(g)$. The numbers generated by P will follow Benford's law, if and only if, $PDF(f) = 0$ at all nonzero integer frequencies f.*

This is to a large extent satisfied if the distribution is wide (since wide distributions have a narrow Fourier transform). This means that Benford's Law requires that the numbers in the distribution being measured have a spread of at least an order of magnitude. Without going too much into the theory of Fourier transform an example of Theorem 2.1 is provided below.

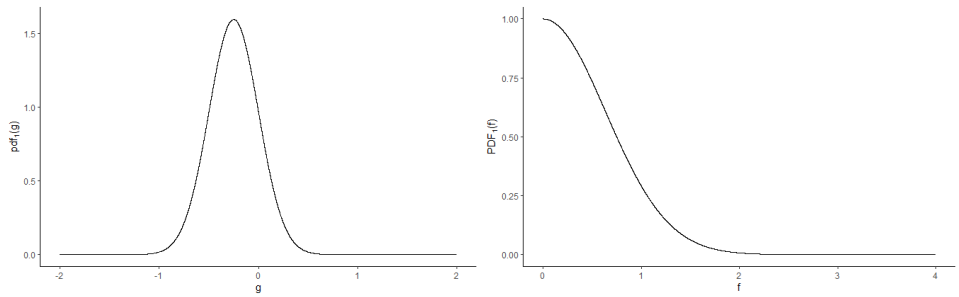
Example 4. Let $pdf_1(g)$ and $pdf_2(g)$ be the normal distribution probability density functions with mean $\mu = -0.25$ and standard deviance $\sigma_{g,1} = 0.25$ and $\sigma_{g,2} = 0.50$ respectively, expressed on the base ten logarithmic number line. The only difference between $pdf_1(g)$ and $pdf_2(g)$ is the width. Now the absolute values $|PDF_1(f)|$ and $|PDF_2(f)|$ of the Fourier transforms of $pdf_1(g)$ and $pdf_2(g)$, when restricted to the positive real line, will be folds of symmetric normal distributions with standard deviations $\sigma_{f,i} = 1/(2\pi\sigma_{g,i})$ for $i = 1, 2$. Hence $PDF_1(f)$ is going to be twice as wide as $PDF_2(f)$ since $\sigma_{g,1} = \sigma_{g,2}/2$, giving us $\sigma_{f,1} = 0.637$ and $\sigma_{f,2} = 0.318$. In Figure 3 the plots of these functions are illustrated. Both $|PDF_1(f)|$ and $|PDF_2(f)|$ smoothly decrease in value with increasing frequency, which is expected. Now the interesting part to examine is when the Fourier transform curves drop to values close to zero. In Figure 3b we observe that the curve drops down to zero shortly after $f = 2$. So according to Theorem 2.1, randomly generated numbers drawn from the base ten log-normal distribution $\mathbf{Lognormal}(-0.25, 0.25)$ will not follow Benford's Law. If we instead look at Figure 3d we see that the curve drops to zero before $f = 1$. Hence, Theorem 2.1 indicates that randomly generated numbers drawn from the base ten log-normal distribution $\mathbf{Lognormal}(-0.25, 0.5)$ should follow Benford's Law.

2.3 Hypothesis testing

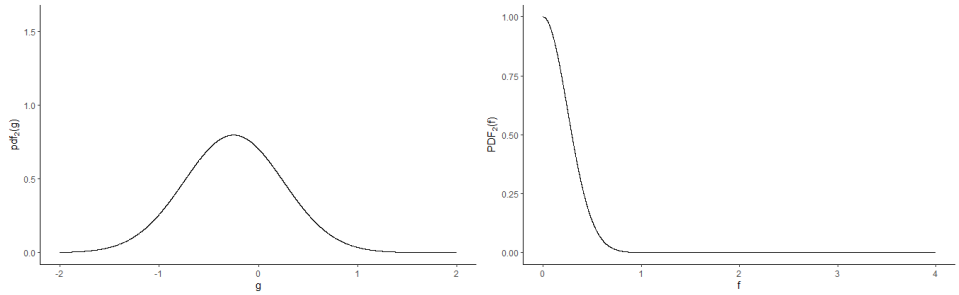
It is of utmost importance in practice to find out whether a data set conforms to Benford's Law. To do this we will make use of a *goodness-of-fit* test. That is, we have the *null hypothesis:*

$$H_0 : \text{data conform to Benford's Law}$$

which we test against the *alternative hypothesis:*



(a) $pdf_1(g)$ (b) $|PDF_1(f)|$



(c) $pdf_2(g)$ (d) $|PDF_2(f)|$

Figure 3: Illustration of the probability density functions and their Fourier transform in Example 4

$$H_1 : \text{data do not conform to Benford's Law.}$$

We can do this in different ways. Here we are going to present a few of them.

2.3.1 Seperate testing of single digits

The simplest test we can do is to test if there is a significant difference between the observed frequency $\hat{p}_d$ of an event $\{D_1 = d\}$ and the corresponding probability which should be $p_d = \mathbb{P}(D_1 = d) = \log(1 + 1/d)$, as we know from Proposition 2.2. Testing conformity to Benford's First Digit Law is now done for each of the nine possible values of d of D_1 separately:

$$H_0 : p_d = \log(1 + 1/d),$$
$$H_1 : p_d \neq \log(1 + 1/d).$$

Using a z-test statistic we get

$$T_d = \frac{\hat{p}_d - p_d}{\sqrt{p_d(1-p_d)}}\sqrt{n}, \quad d = 1, 2, \ldots, 9$$

which for a large sample size n is approximately has a standard normal distribution under the null hypothesis, so H_0 is rejected if $|T_d| > 1.96$ for a significance level of 0.05.

This test should be used with great caution and interpreted with care. Suppose that testing the nine null hypotheses leads to a rejection of only the hypothesis for $d = 1$. This is certainly not sufficient to conclude that our data does not follow Benford's Law. The overall point is that performing these tests simultaneously will affect the probability of a Type-II error, due to multiple testing.

2.3.2 Pearson's Chi-square test

A commonly used test in statistics is the χ^2-test. This test might be more reliable since it tests all the digits simultaneously. Let

$$\mathbf{p} = (p_1, p_2, \ldots, p_9)^t \text{ and } \hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_9)^t$$

be the vectors of the expected probability and the observed frequency of each digit for the first significant digit respectively, as defined in Section 2.3.1. Then we will have the hypothesis:

$H_0 : D_1$ has distribution $\mathbf{p}$ and therefore is *Benford*,
$H_1 : D_1$ has another distribution and therefore is not *Benford*.

We can then construct the χ^2-test

$$\chi^2 = n \sum_{d=1}^{9} \frac{(p_d - \hat{p}_d)^2}{p_d}.$$

If our null hypothesis holds χ^2 is going to follow a χ^2-distribution with 8 degrees of freedom. For a significance level of 0.05, we get a critical level of 15.51. If the value of χ^2 is greater than this value we reject the fit of Benford's Law with 95% certainty.

This method can of course also be applied to all the other significant digits, but then χ^2 will follow a χ^2-distribution with 9 degrees of freedom because we have one more digit to consider. Then the critical level is 16.92 for a significance level of 0.05.

There is however a drawback with this method. If we have a small sample size n the statistical power gets low and the test therefore give unreliable conclusions. The test is very dependent on the sample size.

2.3.3 Kolmogorov-Smirnov test

A way to overcome this is to use a nonparametric goodness-of-fit test. A powerful nonparametric goodness-of-fit test when the sample size is small is the *Kolmogorov–Smirnov test* (or just KS). This test makes use of *empirical distribution function*.

Definition 2.4. *Given a sample $(X_1, X_2, \ldots, X_n)$ of identically and independently distributed random variables with distribution function $F(x) = P(X \leq x)$, the **empirical distribution function (ecdf)** $F_n(x)$ of the sample is defined by*

$$F_n(x) = \frac{\textit{number of sample values} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$$

where $\mathbb{1}(A)$ is the indicator function of event A.

The ecdf has many interesting properties, but the most important one is given in the *Glivenko-Cantelli Theorem*.

Theorem 2.2 (Glivenko-Cantelli)**.** *Let $\mathcal{D}_n = \sup_x |F_n(x) - F(x)|$. Then $\lim_{n\to\infty} \mathcal{D}_n = 0$ almost surely.*

Proof idea. For simplicity, consider the case of a continuous random variable X. Fix $-\infty = x_0 < x_1 < \ldots < x_{m-1} < x_m = \infty$ such that $F(x_j) - F(x_{j-1}) = \frac{1}{m}$ for $j = 1, 2 \ldots, m$. Now for all real x there exists $j \in \{1, \ldots, m\}$ such that $x \in [x_{j-1}, x_j]$. Note that

$$F_n(x) - F(x) \leq F_n(x_j) - F(x_{j-1}) = F_n(x_j) - F(x_j) + \frac{1}{m},$$
$$F_n(x) - F(x) \geq F_n(x_{j-1}) - F(x_j) = F_n(x_{j-1}) - F(x_{j-1}) - \frac{1}{m}.$$

Therefore

$$\sup_x |F_n(x) - F(x)| \leq \max_{j\in\{1,\ldots,m\}} |F_n(x_j) - F(x_j)| + \frac{1}{m}.$$

Since $\max_{j\in\{1,\ldots,m\}} |F_n(x) - F(x)| \to 0$ **a.s.** by the strong law of large numbers, we can guarantee that for any positive number ϵ and integer m such that $\frac{1}{m} < \epsilon$, we can find N such that for all $n \geq N$, we have $\max_{j\in\{1,\ldots,m\}} |F_n(x) - F(x)| \leq \epsilon - \frac{1}{m}$ **a.s.**. Combined with the result above, this implies that $\sup_x |F_n(x) - F(x)| \leq \epsilon$ for all large enough n with probability 1, which is the definition of almost sure convergence. □

We can now construct the Kolmogorov–Smirnov statistic

$$\mathcal{D}_n = \sqrt{n} \sup_t |F_n(t) - F(t)|.$$

By the Glivenko-Cantelli Theorem, this statistic converges to 0 with probability 1 in the limit when n goes to infinity, if the sample comes from the distribution $F(t)$.

With this statistic we can test the hypothesis that the ecdf equals the cumulative distribution function (cdf) of Benford's law, i.e.

$$H_0 : P(D_1 \leq t) = F(t) \text{ for all } t,$$
$$H_1 : P(D_1 \leq t) \neq F(t) \text{ for at least one } t,$$

where the cdf is given by

$$F(t) = \begin{cases} 0 & \text{for } t < 1, \\ \log(1 + \lfloor t \rfloor) & \text{for } 1 \leq t < 9, \\ 1 & \text{for } t \geq 9, \end{cases}$$

which will be a two-sided test and the null hypothesis will be rejected if the observed $\mathcal{D}_n$ is too large. The Kolmogorov-Smirnov test normally assumes that the sample comes from a continuous distribution, which obviously is not the case for us. A correction factor for the test was introduced by Stephens (1970)[8], and this made it possible to get more accurate results. With this new corrected statistic (in practice only the critical values are different) Morrow (2014, p. 4)[9] determined a more accurate critical value of 1.148 for a significance level of 0.05. So we reject the null hypothesis if $\mathcal{D}_n$ is greater than 1.148.

All these types of goodness-of-fit tests can of course be applied on all digits, and are not restricted to the first digit D_1 only. The same procedure applies when testing the distribution of the k^{th} digit D_k. We can also do the same for the joint distribution of multiple digits.

3 Benford's Law on common distributions

According to the theory, as we saw in Section 2.2.4, the distribution have to span over at least an order of magnitude to follow Benford's Law. However, this does not mean that only because a distribution spans over at least an order of magnitude it will follow Benford's Law. With this knowledge we will look how some of the common distributions coincide with Benford's Law. The distributions we will consider are:

- The **uniform distribution**, $U(0, 1000)$, with density

$$f_{U(0,1000)}(x) = \begin{cases} 1, & 0 \leq x \leq 1000 \\ 0, & \text{otherwise.} \end{cases}$$

- The **normal distribution**, $N(\mu, \sigma^2)$, with density

$$f_{N(\mu,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty.$$

- The **exponential distribution**, $\text{Exp}(\lambda)$, with density

$$f_{\text{Exp}(\lambda)}(x) = \lambda e^{-\lambda x}, \qquad x \geq 0.$$

We will also investigate how well the distribution of ratios of random variables having these two distributions approximate Benford's Law.

3.1 The uniform distribution

Starting with the uniform distribution, we simulate 10 000 random variables following the uniform distribution defined above. As we see in Figure 4 the empirical distribution of the sample approximately produces a uniform discrete distribution of the first nine digits. Hence we may conclude that the uniform distribution and Benford's Law are incompatible, even though the uniform distribution spans over many orders of magnitude.

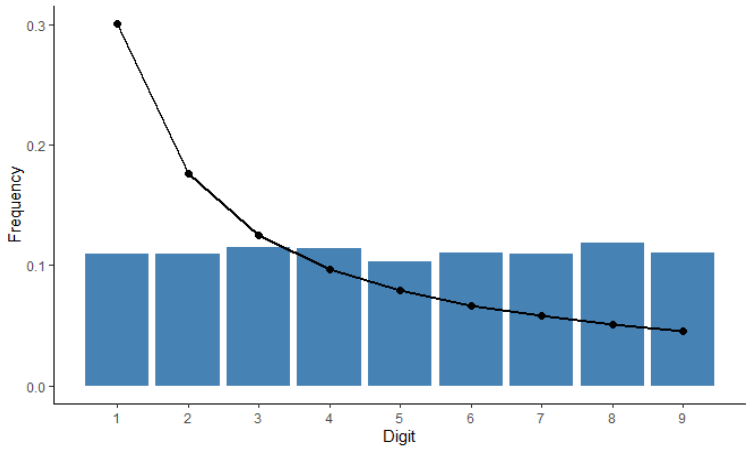


Figure 4: The distribution of the first digit of 10 000 random variables drawn from the uniform distribution, $U(0, 1000)$.

In spite of this, it is more constructive to consider in more detail the deviation between the simulated relative frequency of the digits and their theoretical values to get an impression of the precision which we can expect from

the simulation study. We do this with help of the hypothesis tests in Section 2.3. As stated there, for the different tests we are going to reject the null hypothesis that the distribution conforms with Benford's Law if $|T_d| > 1.96$, $\chi^2 > 15.51$ and $\mathcal{D} > 1.148$ respectively. In general the p-values are better to present then the test statistics when comparing tests with different scales with each other. However, the p-values get exceedingly close to 0 and the p-values would not be exact for the asymptotic approximations of χ^2 and $\mathcal{D}$. Hence, the test statistics will be presented and not the p-values.

In Table 1 we can clearly see that the test statistics are way above the critical values of the tests. Hence we can conclude that the uniform distribution does not conform well with Benford's Law.

Digit	Benford	Frequency	$\lvert T_d \rvert$	χ^2	$\mathcal{D}$
1	0.30	0.11	41.99	3923.1	26.7
2	0.18	0.12	17.56		
3	0.12	0.11	2.70		
4	0.10	0.11	4.97		
5	0.08	0.11	11.82		
6	0.07	0.12	17.99		
7	0.06	0.11	22.21		
8	0.05	0.10	27.89		
9	0.05	0.11	30.41		

Table 1: The distribution of the first digit of 10 000 random variables following the uniform distribution, $U(0, 1000)$, and the Z, χ^2 and KS statistics.

3.2 The normal distribution

For the normal distribution we will consider two different variations of the mean, μ, and variance, σ^2. Firstly we will test the standard normal distribution, $N(0, 1)$, and secondly we will test the normal distribution with mean $\mu = 100$ and standard deviance $\sigma = 225$. We do this to see if the mean and variance have any impact on the compatibility. The values in the second distribution are chosen such that we can expect values of a greater order of magnitude, which may improve the fit of Benford's Law, according to the theory in Section 2.2.4.

Starting with the standard normal distribution, we simulate 10 000 random variables from the standard normal distribution. As we can see in Figure 5 it is not as easy to see if the standard normal distribution conforms with Benford's Law or not. The standard normal seems to fit Benford's Law quite well but not really all the way.

For this reason it is important to take hypothesis testing into consideration.

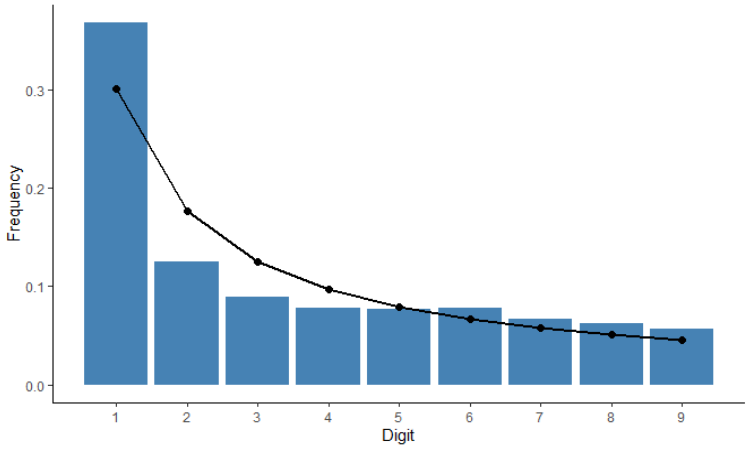


Figure 5: The distribution of the first digit of 10 000 random variables following the standard normal distribution, $N(0, 1)$.

Digit d	Benford	Frequency	$\lvert T_d \rvert$	χ^2	$\mathcal{D}$
1	0.30	0.37	14.60	516.37	6.70
2	0.18	0.12	13.47		
3	0.12	0.09	10.81		
4	0.10	0.08	6.29		
5	0.08	0.08	0.92		
6	0.07	0.08	4.34		
7	0.06	0.07	3.60		
8	0.05	0.06	5.20		
9	0.05	0.06	5.00		

Table 2: The distribution of the first digit of 10 000 random variables following the standard normal distribution, $N(0, 1)$, and the $Z = |T_d|$, χ^2 and KS statistics.

In Table 2 the test statistics of the Z-, χ^2- and KS-test are represented for the null hypothesis that Benford's Law holds. As we can see each statistic is rejecting the null hypothesis with an exception of the Z-statistic for the fifth digit. This makes us conclude that Benford's Law doesn't hold for the standard normal distribution.

Now, when using the other normal distribution with larger mean and variance, we expect the distribution to have a better fit than the standard normal distribution. In Figure 6 it looks like digits 2 and 3 are over-represented but other than that the simulated data set seems to have quite a good fit, at least from a visual inspection. The question arises how good the fit is according to the hypothesis tests.

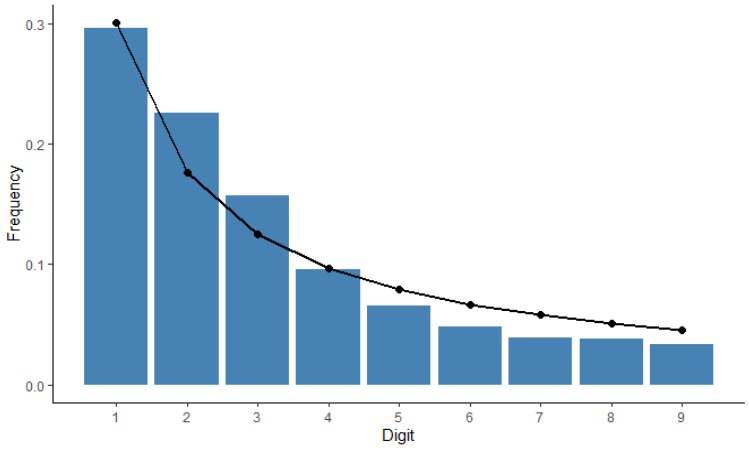


Figure 6: The distribution of the first digit of 10 000 random variables following the normal distribution $N(100, 225^2)$.

Digit d	Benford	Frequency	$\lvert T_d \rvert$	χ^2	$\mathcal{D}$
1	0.30	0.30	0.99	438.10	7.82
2	0.18	0.23	13.23		
3	0.12	0.16	9.79		
4	0.10	0.10	0.21		
5	0.08	0.07	5.18		
6	0.07	0.05	7.66		
7	0.06	0.04	8.25		
8	0.05	0.04	5.88		
9	0.05	0.03	5.87		

Table 3: The distribution of the first digit of 10 000 random variables following the normal distribution, $N(100, 225^2)$, and the $Z = |T_d|$, χ^2 and KS statistics.

A bit surprisingly, we see in Table 3 that the null hypothesis is rejected by almost all tests, apart from the Z-statistic for the first and fourth digit. In Figure 6 it seemed that the distribution might have a better fit than the standard normal distribution. Depending on which test is used we get different results when comparing the fits of the two normal distributions. Using the χ^2-test for the $N(100, 225^2)$-distribution seems to have a better fit but if the KS-test is used the opposite is happening. But that is probably related to the fact that the χ^2-test is based on comparing probability functions whereas the KS-tests relies on the closeness of distributions functions. In any case, both of the distributions reject the null hypothesis to follow Benford's Law by a large margin.

3.3 The exponential distribution

For the exponential distribution we will consider the case where the rate is 1, i.e. $\lambda = 1$. Hence we simulate 10 000 random variables following the exponential distribution Exp(1). In Figure 7 we see the distribution of the first digit. It seems to have a very close fit to Benford's Law. There are some digits that are a little bit off. The first is a bit over-represented and the third, fifth and seventh digits are a bit under-represented.

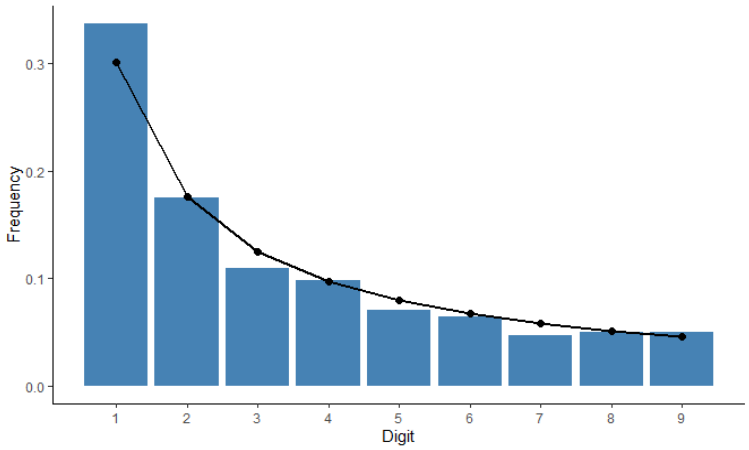


Figure 7: The distribution of the first digit of 10 000 random variables following the exponential distribution Exp(1).

Digit d	Benford	Frequency	$\lvert T_d \rvert$	χ^2	$\mathcal{D}$
1	0.30	0.34	2.48	10.05	1.14
2	0.18	0.17	0.09		
3	0.12	0.11	1.52		
4	0.10	0.10	0.12		
5	0.08	0.07	1.08		
6	0.07	0.06	0.37		
7	0.06	0.05	1.49		
8	0.05	0.05	0.17		
9	0.05	0.05	0.64		

Table 4: The distribution of the first digit of 10 000 random variables following the exponential distribution, Exp(1), and the $Z = |T_d|$, χ^2 and KS statistics.

In Table 3 the numerical results of the tests are presented. We see that for the Z-test only the first digit is rejecting the null hypothesis but all the other tests are keeping the null hypothesis. Hence we conclude both from Figure 7 and Table 4 that the exponential distribution follows Benford's Law very

well.

3.4 Ratio distributions

In Section 2.2.4 it is said that we expect to observe a Benford distribution when the data set is generated from two distributions. So if we simulate 10 000 random variables from a distribution $Z = X/Y$ where X and Y are independent and follow some distributions, this ratio Z should follow Benford's Law. We will test this by letting X and Y follow the distributions mentioned at the beginning of Section 3, and also that they will have the same distribution. In Table 5 the results of the χ^2- and KS-tests are presented.

Distribution	χ^2	$\mathcal{D}$
$U(0, 1000)/U(0, 1000)$	286.33	4.06
$N(0, 1)/N(0, 1)$	5.23	0.52
$N(100, 255)/N(100, 255)$	12.07	0.65
Exp(1)/Exp(1)	5.16	0.85

Table 5: The χ^2 and KS statistics of four different ratio distributions, each distribution simulated from in terms of 10 000 independent random variables.

As we can see, only the ratio distribution constructed from the uniform distribution is rejecting the null hypothesis. So it seems that we can expect a Benford distribution when the data is generated from two distributions, confirming what we expected. However, although Benford's Law seems to hold for many choices of the distributions of X and Y, this is not always the case.

3.5 Conclusion

In this section we have tested some of the criteria in Section 2.2.4 to see if they are accurate. What we have seen is that our distributions don't have to follow Benford's Law just because the criteria are fulfilled. All of the distributions tested spread over at least one order of magnitude and are tested on a large amount of data points. But this did not imply that all the distributions followed Benford's Law. As we saw, only the exponential distribution and the ratio distribution constructed from the normal distribution or the exponential distribution were close to the Benford distribution. Hence we can say that the the two criteria *result comes from two distributions* and *when the mean is greater than the median and the skewness is positive*, seem to hold very well for the distributions that we investigated. Indeed, the ex-

ponential distribution, Exp(1), is a very positively skewed distribution with mean 1 and median less than 1.

4 Greece economic fraud

It is well known that the Greece government manipulated their economic data to the European Union in the beginning of the 21st century (European Commission, 2010)[10]. They did so to hide the fact that they were in great debt and thereby breaking the Stability and Growth Pact as well as the Euro convergence criteria. We want to see if this is something we can detect with Benford's Law.

4.1 Data gathering

The economic dataset used in this study comes from the European statistic agency (Eurostat). Eurostat openly provides high quality statistics for the European countries. Therefore our observed data is taken from Eurostat's database in March 2021, under the data theme "Economy and Finance". Within this data theme our data sets are taken from the categories "National accounts", and "Government statistics". From these categories we gather the following data sets, which are then merged into one single data set:

- Gross domestic product (GDP) and main components - Only the GDP part.
- Financial balance sheets - Value of the stocks of assets and liabilities for a country at a specific time point.
- Financial transactions - How the surplus or deficit of the capital account is financed by transactions in financial assets and liabilities.
- Government deficit/surplus, debt and associated data - The deficit and debt of the general government sector.
- Government revenue, expenditure and main aggregates - Main revenue and expenditure items of the general government sector.

These data sets are chosen because they are all associated with public deficit, public debt and gross national products, which are items used in the evaluations of the Euro convergence criteria. The period of analysis is from 1999 to 2010. The year 1999 is chosen as the starting point because it is the year Euro became the official currency of the euro area, and the year 2010 is chosen because it is around this time that the Greek scandal got well known.

All analysed data are expressed in absolute values in million euros, based on currency convergence calculated by Eurostat. These data sets contain a total of 152 single positions per country per year. Unfortunately, in former years there is a considerable number of missing values and enteries with value zero, which had to be omitted from the sample. For United Kingdom only 13 observations per year are available. In total we have 336 samples of varying size, with an average of 92 observations per sample. The total sample contains 30 900 observations.

4.2 Results from individual countries

The results of calculating the χ^2- and KS-statistics of the individual member states in EU annually are shown in Table 6. The countries are sorted by the mean of their χ^2 statistics. Hungary, with a mean value of 20.49 shows the largest deviation from Benford's Law among the EU countries, closely followed by United Kingdom with a value of 20.35 and France with a value of 20.29. The lowest mean value, 8.96, is obtained for Croatia. We can see that the ranking of the χ^2-statistics coincide poorly with the KS-statistics which are less dependent on sample size. Hence we might suspect that the sample size can be problematic in our case.

We might expect that Greece would show the largest deviation from Benford's Law from the knowledge we have about them manipulating their economic data. Still, in the rank of the mean χ^2 Greece only appears in eighth place. In the ranking of the KS-statistics Greece appears at third place with only the United Kingdom and Malta above it. The reason for this and why so many countries deviate significantly from Benford's Law could not only be due to the sample sizes. Indeed, we only have an average sample size of 92 observations and this makes it hard to rely on the test.

4.3 Result from Greece

In Table 7 the distribution of Greece's first digit is shown. Surprisingly it seems to follow Benford's Law very well, with some exception in the years 2000 and 2010. This is not expected at all and since it is well documented that Greece has been manipulating its data over these years (European Commission, 2010)[10]. There might be a Type-II error, since it is quite common that the Z-statistic tends to be somewhat conservative when used for testing Benford's Law. This is also supported by the χ^2 statistic which shows a significant deviation from Benford's Law for the majority of the years. Indeed, as stated above the χ^2 statistic is probably very dependent on the sample size, and since the sample size mean over the years are 102.75, which is not very large, this indicates that Greece indeed violates Benford's

Table 6: The χ^2 statistic for each EU country

Country	Mean size	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	Mean χ^2	Rank $\mathcal{D}$
Hungary	113	12.21	1.68	28.78*	26.77*	24.53*	17.84*	32.75*	31.50*	15.84*	11.78	20.10*	22.15*	20.49*	8
UK	13	18.83*	23.21*	37.93*	13.70	25.94*	26.83*	24.89*	28.33*	17.96*	13.69	6.81	6.09	20.35*	1
France	109	28.14*	17.55*	30.08*	28.37*	16.57*	29.24*	17.85*	23.87*	21.55*	7.28	13.56	9.42	20.29*	18
Cyprus	83	17.56*	31.37*	32.42*	20.65*	25.48*	29.62*	17.92*	8.75	27.96*	9.32	15.10	6.35	20.21*	11
Austria	119.67	11.67	13.67	9.86	15.46	31.30*	14.84	32.78*	38.64*	9.04	14.20	18.47*	21.34*	19.27*	5
Czechia	85	19.90*	13.31	22.45*	23.97*	22.39*	16.24*	19.56*	7.91	7.56	16.10*	9.96	28.40*	17.31*	3
Belgium	101	12.85	14.94	12.53	13.91	27.00*	22.43*	26.71*	16.37*	26.01*	12.70	8.09	14.03	17.30*	9
Greece	102.75	15.47	19.48*	22.50*	5.62	20.97*	27.89*	14.83	21.62*	6.35	18.91*	10.18	17.17*	16.75*	4
Ireland	61.17	18.16*	14.44	20.86*	12.02	10.73	11.40	18.36*	24.34*	22.76*	14.71	10.23	12.20	15.85*	12
Portugal	77	17.34*	21.29*	14.10	15.45	14.33	23.30*	20.92*	12.78	24.25*	5.87	8.01	11.21	15.74*	13
Malta	91.25	8.62	21.04*	3.02	15.03	9.83	15.76*	16.80*	20.97*	23.23*	31.69*	9.54	5.21	15.06	2
Sweden	113	5.37	7.47	15.31	13.78	8.04	9.66	9.25	19.46*	23.66*	21.07*	36.11*	6.32	14.63	23
Lithuania	107	12.05	14.47	28.50*	27.73*	15.63*	7.43	14.31	12.64	6.20	9.43	12.72	8.76	14.16	28
Luxembourg	118	13.23	7.18	20.66*	3.53	20.64*	6.86	18.47*	19.71*	13.66	3.46	30.13*	10.74	14.02	16
Finland	112.25	9.33	6.64	14.64	7.81	7.65	11.28	4.62	7.66	14.47	17.36*	24.43*	42.17*	14.00	10
Slovakia	108.25	10.49	23.40*	11.51	10.08	11.75	7.71	10.30	9.48	12.38	18.98*	14.53	23.35*	13.66	17
Latvia	112.25	5.20	8.70	14.37	7.27	32.71*	12.85	11.69	17.33*	8.16	23.78*	10.54	9.20	13.48	14
Spain	93	23.15*	14.10	9.99	9.07	29.96*	18.97*	10.38	9.60	13.27	3.38	11.15	8.65	13.47	6
Germany	121	7.75	12.36	19.72*	7.22	16.25*	9.20	21.35*	7.38	14.57	14.76	13.31	8.52	12.70	15
Romania	107.42	29.09*	17.33*	7.19	8.68	5.91	11.12	10.54	7.77	18.84*	6.75	15.87*	11.93	12.59	24
Denmark	105	4.58	12.54	10.84	15.82*	6.53	8.79	18.30*	5.79	25.25*	14.96	17.42*	8.07	12.41	7
Netherlands	85	10.27	29.32*	5.23	16.77*	3.91	9.59	5.36	14.66	33.48*	4.37	10.59	4.07	12.30	25
Slovenia	83.17	19.84*	12.19	8.27	13.04	7.26	24.48*	9.89	17.04*	7.85	11.05	8.67	5.49	12.09	20
Estonia	93.58	9.22	2.18	7.58	11.03	7.08	18.13*	14.42	14.50	20.04*	10.21	8.40	16.56*	11.61	21
Italy	85	8.51	5.05	6.51	18.08*	15.05*	21.73*	8.42	4.91	14.95	4.67	11.37	14.82	11.17	22
Bulgaria	84.83	12.59	15.58*	19.17*	10.05	14.50	2.22	3.75	10.37	14.12	7.07	7.69	8.06	10.43	19
Poland	84.92	6.18	9.59	6.89	9.51	8.34	6.97	7.65	13.16	12.32	3.06	8.15	20.52*	9.36	26
Croatia	108.42	10.60	3.59	8.83	1.66	3.90	11.70	18.31*	9.38	11.18	5.58	13.41	9.34	8.96	27

Note: The table includes test statistics from the χ^2 goodness of fit test. The countries sorted according to their mean χ^2.
A ranking of the countries based on their values of $\mathcal{D}$ is provided for comparison.
Significance deviations (*): 5% significance level 15.51.

Law.

4.4 Conclusion

With these results we cannot conclude that we are detecting any suspicious activity. This is also more accurate after an email conversation with Eurostat that took place after the results of this thesis had been obtained. The staff at Eurostat confirmed that the data had been corrected over the last ten years, and they were not available to contribute with any older data. This however explains why we couldn't find any fraud in Greece's economic data. But it still raises the question why so many countries don't seem to follow Benford's Law over the twelve year period. Is it just a coincidence, something suspicious about the data or is Benford's Law not as good as a statistical tool as one might have thought? One possible explanation why so many countries violate Benford's Law is that the sample sizes, though not very large, are still large enough to detect quite small departures from the logarithmic Benford distribution.

5 Election

Benford's Law is quite a good tool to detect economic fraudulence, but can we detect any other kind of fraud with Benford's Law? A hot topic right now is fraud in election, especially in the United States where Donald Trump accused Joe Biden of fraud during the US presidential election of 2020. Here we will investigate whether we can detect fraud in simulated elections.

5.1 Simulation of a fraud-free election

There are many different ways to go about when simulating an election. In our case we are mostly interested in how many people vote on each candidate and not really which candidate that wins or by which margin. So firstly we want to simulate an election were we are confident there was no fraud, and then manipulate the result in such a way that the other candidate would have won. A way to do this is by a so called spatial model[11], first introduced to elections by Downs (1957)[12], and therefore referred to as Downs model. The main idea is that voters are identified by ideal points in an Euclidean "issue" space, candidates take place in that space, and voters vote for the candidate with smallest Euclidean distance to their ideals.

Let us start our simulation by letting $G = 2$ be the mean personal income of the voting age population of the nation. Then suppose that the nation has 290 districts (the number of municipalities in Sweden) where each district d

Table 7: Distribution of the first digit for Greece.

Digit	Benford	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
1	0.30	0.42*	0.39*	0.29	0.36	0.30	0.25	0.25	0.39	0.35	0.22	0.27	0.20*
2	0.18	0.25	0.25*	0.25*	0.18	0.25*	0.35*	0.28*	0.15	0.17	0.27*	0.20	0.25*
3	0.12	0.09	0.04*	0.11	0.08	0.09	0.16	0.17	0.22*	0.12	0.08	0.17	0.12
4	0.10	0.06	0.05	0.05	0.08	0.03*	0.01*	0.10	0.07	0.12	0.16*	0.15	0.17*
5	0.08	0.05	0.06	0.08	0.09	0.08	0.08	0.04	0.06	0.07	0.04	0.06	0.10
6	0.07	0.05	0.07	0.06	0.04	0.04	0.04	0.06	0.02	0.09	0.09	0.07	0.06
7	0.06	0.02	0.05	0.00*	0.05	0.06	0.05	0.03	0.06	0.03	0.09	0.03	0.01*
8	0.05	0.05	0.08	0.12*	0.07	0.05	0.03	0.04	0.01	0.05	0.04	0.03	0.08
9	0.05	0.02	0.01	0.05	0.06	0.11*	0.04	0.04	0.03	0.01	0.02	0.03	0.02
	χ^2	15.47	19.48**	22.50**	5.62	20.97**	27.89**	14.83	21.62**	6.35	18.91**	10.18	17.17**
	$\mathcal{D}$	1.76**	1.66**	0.69	0.59	0.77	1.57**	0.99	1.58**	0.69	0.79	0.89	0.99

Note: The table includes the result from the Z-test represented as $\cdot$* if there is a significant deviant with 5% significance level.
Values of the χ^2 and $\mathcal{D}$ statistics are provided for comparison.
Significance deviations (**) 5% significance level: 15.51 for χ^2, 1.148 for $\mathcal{D}$.

has the mean income g_d, which is drawn randomly from the folded normal distribution $|N(G, 0.15)|$. A voter i, from district d, has an income V_{id} which is drawn from the folded normal distribution $|N(g_d, 2)|$. This is because the income should not take on negative values . Each voter is represented by opinions of two issues, X and Y, which depend on the voters income. In order to allow for the possibility that that income policy preferences vary across districts by unobserved variables, we let β_{Xd}, and β_{Yd} be the "impact" of the income on variables X and Y in district d. The variable β_{Xd} is taken from the distribution $N(2, 0.15)$ and β_{Yd} from the distribution $N(-1, 0.15)$. The reason for different means in these variables is to ensure that our two-dimensional distribution of policy preferences will not be radially symmetric. Finally, each voter i in district d, will have the policy preferences determined by

$$X_{id} = \beta_{Xd} V_{id} + u_{Xi}$$
$$Y_{id} = \beta_{Yd} V_{id} + u_{Yi}$$

where u is a noise term, such that $u \sim N(0, 2)$. This is because we want to spread out the voters on the issues.

Now that we know how to simulate the salary and opinions of each voter, we need to know how many voters there are. To make it easy we assume that there are 10 000 voters in each district. So in total we will have 2 900 000 voters.

The only thing that needs to be added to the simulation is the candidates. We know that each candidate wants to gather as many voters as possible. To do this they will place themselves as close to the "middle" of the issues as possible. In particular, if both are exactly in the middle both will get exactly half of the votes. This is also known as the median voter theorem which was proposed by Downs (1957)[12]. So in order to make it as realistic as possible we will place each candidate close to the median of the voters' ideals. That is, after all the voters policy preferences have been determined and placed in the issue space we will determine the candidates' positions, (X_C, Y_C), in the space by

$$X_C = \tilde{X} + \epsilon,$$
$$Y_C = \tilde{Y} + \epsilon,$$

where $\tilde{X}$ and $\tilde{Y}$ is the median of all voters opinion in issue X and Y respectively, and ϵ is a noise term, such that $\epsilon \sim N(0, 1)$. In Figure 8 we can see how all the voters are distributed in the issue-plane and then also how the candidates are placed according to our model.

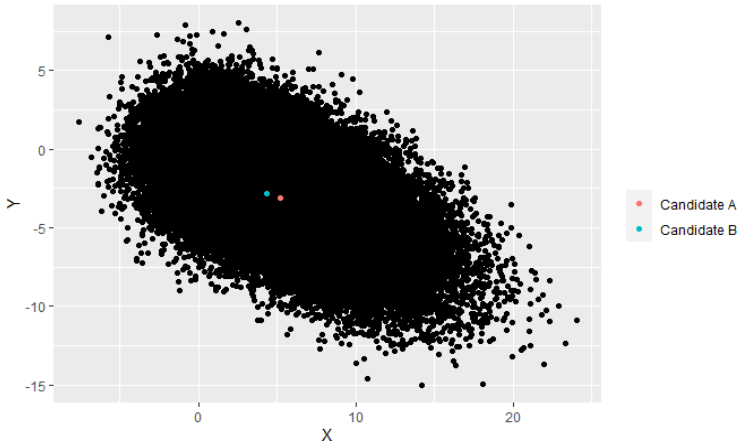


Figure 8: Voters and candidates in the issue-plane defined by the issues X and Y.

Now that we have a quite realistic election model we can apply Benford's Law on the data we gather from the simulation. We do this by looking at the distribution of the second digit of the number of voters of each candidate among all $n = 290$ districts. The reason we are looking at the second digit and not the first, is because the voting result is going to be fairly close to 50%, which means that the vote count for each candidate is going to be very close to 5000 for each candidate. This means that the vote count is not going to span over several orders of magnitude. Hence the first digit of the vote count is not going to follow Benford's Law, but we expect that the second digit will. Mebane (2006)[13] was the first to propose that using the second digits of Benford's Law to detect election fraud would be more appropriate.

Using this method with the χ^2-test on all the votes in each district, regardless of the candidate, on 100 simulated elections we find that for 70 of these elections the second digit will follow Benford's Law and for the remaining 30 elections they will not. If we only look at the votes for the wining candidate or for the losing candidate we get that it follows Benford's Law in 74 and 71 of the elections respectively. This indicates that Benford's Law might have quite a high chance of committing a Type-I error when used on elections, a finding which is also supported by Deckert, Myagkov and Ordeshook (2011)[14], who used a somewhat similar method to show that Benford's Law is subject to Type-I and Type-II errors quite frequently when used on election fraud. But we can still see that for the majority of the time the vote count seems to follow Benford's Law.

5.2 Introducing fraud

There are many ways to introduce fraud into an election. We will however use two kinds of simple methods to introduce fraud. In the first method we are going to move uniformly between 5 to 15% of the voters of the winning candidate over to the losing candidate in a randomly selected subset with 20% of the districts. This will be a quite realistic method since one might expect that if a candidate will commit fraud they will probably go to a few districts and manipulate the voting there, to not raise any suspicion. The second method is a little bit simpler, this time we take the second digit of the vote count in each district for the winning candidate and subtract 1 from it, and to not raise any suspicion with the number of voters we add 1 to the second digit of the vote count in each district for the losing candidate. This method might not be as realistic as the first method, but it is a method that might be hard to detect but intuitively most certainly will be detected by Benford's Law.

By using the first method of introducing fraud we get that the total voting count only has a significant deviation from Benford's Law in 56 of the cases. This is however a higher fraction of violations of Benford's Law compared to the scenario when there was no fraud introduced, but it still means that we only detect the fraud a little over 50% of the time. This of course is a very small fraction for a test supposed to detect fraud. If we look at each candidate separately we don't get a much better result. We detect suspicious activities in 54 of the cases for both candidates. This strongly indicates a large Type-II error which was discussed above and supported by Deckert, Myagkov and Ordeshook (2011)[14]. However, if we only consider those elections where the winner actually changes because of the manipulations in the data the total voting count has a significant deviation from Benford's Law in 60% of those elections. This is a slight increase but not by very much.

Using the second method of introducing fraud we get a somewhat better result. The total voting count has a significant deviation from Benford's Law in 64 of the simulated elections. For each candidate separately we detect the fraud in 54 respective 55 of the cases. Also when looking at the conditional case that the winner of the election actually changes because of the manipulations in the data the total voting count has a significant deviation from Benford's Law in 70% of the elections. Hence it seems that the conditional case makes Benford's Law more acurate. But there is still a pretty large uncertainty, and a large risk for committing a Type-II error for both of the methods.

5.3 Conclusion and discussion

We found that using Benford's Law to detect suspicious activities in elections may be very risky, since we obtained large Type-I and Type-II errors which leads to unreliable results. This may also be caused by the χ^2-test which is heavily dependent on sample size. With too small a sample size the result cannot be accurate because the power of the test is too low. If the sample size is too large a very small deviance would reject the null hypothesis. We chose 290 districts because that number seemed realistic and was not too large, but this might as well be too small of a sample size in order to detect violations of Benford's Law.

To see if we could get a better result with a larger sample size, we conducted the same test as above, both for a fraud-free election and an election with fraud introduced, but with 500 districts instead. With only 210 more samples the χ^2-test cannot handle the larger sample size and rejects the fit of Benford's Law in all simulations.

A way to approach the problem worth testing, since it may give a better result, is to use a different base than 10 for Benford's Law. The problem with vote counts is that they don't span over several orders of magnitude, so the first digit of Benford's Law cannot be applied. But if we change the base of the vote count to a lower base, this problem will be smaller. Since Benford's Law for the first digit is base invariant, changing the base such that the vote count spans over several orders of magnitude we should be able to apply Benford's first digit law on the vote count. This way we would be able to use the statistical tests referred to above and perhaps detect the fraud more accurately.

6 Discussion

Unfortunately we have not been able to accurately detect some kind of fraud in this paper. When analyzing the economic data over EU countries we discovered that according to our goodness-of-fit tests many countries didn't seem to follow Benford's Law very well. In Section 4.4 we discussed what the reason for this could be, whether there is something suspicious about this particular dataset, that Benford's Law isn't as good as we thought or that the many rejections of Benford's Law was just a coincidence. But after studying simulations of elections it was pretty clear that there are some problems associated with the χ^2-test when used with Benford's Law. Indeed, we obtained large Type-I and Type-II errors. This indicates that it is maybe not Benford's Law that is wrong, instead it might be the test. It is often the case with statistical tests that more observations lead to better results. But in this case we got worse results the more observations we gathered. Also,

as we saw in Section 3, some distributions that visually seemed to follow Benford's Law was rejected by the tests.

There are some tests that have been established specifically for Benford's Law. This includes Cho-Gaines's distance proposed by Cho and Gaines (2007)[16] and Leemis's m test proposed by Lemmis (2000)[15], which were not used in this paper since we discovered them too late. These tests may be more appropriate to use in the context of Benford's Law. The only problem is that there are not so many studies on these tests, and not much theory on how and why they should perform well. The only study that has analyzed these tests is due to Morrow (2014)[9], where he determined the critical values for the first digit of Benford's Law. It is something worth investigating, since, as we saw in this paper, the more commonly used tests seem to reject Benford's law too easily when the sample size is large.

In fact, it is reasonable, from the result we got in this paper, to assume that in most applications the "theoretical distribution", $q_1, \ldots, q_9$, when there is no fraud, will still have a small departure from Benford's Law, $p_1, \ldots, p_9$, which is detected by the goodness of fit tests with a sufficiently large data set, no mater how close the q_i are to the p_i. So instead of using goodness of fit tests we could define the total variance distance

$$\delta_n = \frac{1}{2}\sum_{i=1}^{9} |\hat{p}_i - p_i|,$$

in order to test the first digit from a data set of size n. This is the fraction of first digits that have to be changed in order to perfectly match Benford's Law. When n tends to infinity, δ converges almost surely to

$$\delta_\infty = \frac{1}{2}\sum_{i=1}^{9} |q_i - p_i|.$$

Then Benford's Law should hold approximately whenever δ_n is less than or equal to $\varepsilon > 0$ for some prechosen small number ε. Such a test will detect departures from closeness to Benford's Law (when $\delta_n > \varepsilon$), not departures from Benford's Law itself (which corresponds to $\delta_n > C/\sqrt{n}$ for some C depending on the significance level of the test). This test could be performed in addition to the null-hypothesis tests.

Another thing which is also worth studying is how good Benford's Law actually is at detecting fraud. After Nigrini (1998) [3] proposed that Benford's Law could be used as an auditing and accounting tool to detect fraud, it started being used for detecting all kinds of fraud. But is that really accurate? Can we really use Benford's Law in more areas than just auditing and accounting? This is something no one really knows. In fact no one really

knows why Benford's Law seems to hold in so many areas. The only thing we have is a few guesses, as seen in Section 2.2.4. It is something worth noting when using Benford's Law. We do not know why it works, and therefore we cannot be fully sure that it works either.

To conclude, Benford's Law is a very interesting concept and mathematically very beautiful. But as we have seen in this paper, it is maybe not very reliable, at least with the theory we have right now. Something we should strive for is to work on more accurate tests before relying on this law too much. When that has been done we can apply these new tests and see if we get any better results.

References

[1] Simon Newcomb (1881). *Note on the Frequency of Use of the Different Digits in Natural Numbers*. Amer. J. Math. 4:1, 39-40 https://dx.doi.org/10.2307/2369148.

[2] Frank Benford (1938). *The law of anomalous numbers*. Proceedings of the American Philosophical Society. 78:4, 551-572.

[3] M. J. Nigrini (1998). *The detection of income tax evasion through an analysis of digital distributions*. Thesis (Ph.D.) Cincinnati, OH, USA: Department of Accounting, University of Cincinnati

[4] Arno Berger and Theodore P. Hill (2015). *An Introduction to Benford's Law*. Princeton University Press. https://doi-org.ezp.sub.su.se/10.1515/9781400866588

[5] Steven J. Miller (2015). *Benford's Law: Theory and Applications*. Princeton University Press. https://doi-org.ezp.sub.su.se/10.1515/9781400866595

[6] Cindy Durtschi, William Hillison and Carl Pacini (2004). *The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data*. J. Forensic Account.

[7] Steven W. Smith (1997), 'Explaining Benford's Law' in *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, pp. 701-722.

[8] M. A. Stephens (1970). *Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables*. Journal of the Royal Statistical Society. Series B (Methodological), 32, 115–122. https://www.jstor.org/stable/2984408

[9] John Morrow (2014). *Benford's Law, Families of Distributions and a Test Basis*. London, Centre for Economic Performance. http://cep.lse.ac.uk/_new/publications/abstract.asp?index=4486

[10] European Commission (2010). *Report on Greek Government Defict and Debt Statistics*. Brussle: Official Publications of the European Communities. https://ec.europa.eu/eurostat/documents/4187653/6404656/COM_2010_report_greek/c8523cfa-d3c1-4954-8ea1-64bb11e59b3a

[11] Thomas Cole Tanner (1994). *The spatial theory of elections: an analysis of voters' predictive dimensions and recovery of the underlying issue space*. Iowa: Retrospective Theses and Dissertations. https://doi.org/10.31274/rtd-180813-7862

[12] Anthony Downs (1957). *An economic theory of democracy*. New York: Harper and Row.

[13] Walter R. Mebane, Jr. (2006). *Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections*. Cornell University: Department of Government.

[14] Joseph Deckert, Mikhail Myagkov, and Peter C. Ordeshook (2011). *Benford's Law and the Detection of Election Fraud*. Political Analysis. 19:245–268.

[15] Lawrence M. Leemis, Bruce W. Schmeiser Diane L. Evans (2000). *Survival Distributions Satisfying Benford's Law*. The American Statistician, 54:4, 236-241, DOI: 10.1080/00031305.2000.10474554

[16] Wendy K Tam Cho and Brian J Gaines (2007) *Breaking the (Benford) Law*. The American Statistician, 61:3, 218-223, DOI: 10.1198/000313007X223496