SCHOOL OF ECONOMICS AND MANAGEMENT

Department of Economics

NEK H02 Bachelor thesis

11th January 2021

Assessing practicalities of Benford's Law

A study of the law's potential to detect fraud in transactional data

Authors:

Esaias Rosencrantz Ollén & Jacob Wennberg

Supervisor:

Ignace De Vos

Abstract

In modern anti-money laundering operations, data analysis plays a vital role. One method of detecting fraudulent data is Benford's law, which predicts the distribution of the first significant digits in logarithmically distributed data. Deviation from Benford's law in data where it should be present might indicate manipulated data.

We investigate the Chi^2 and Kolmogorov-Smirnov tests' properties to see how their empirical sizes and test powers are affected by different sample and variance sizes. With this knowledge, we set out to evaluate the reliability in common methods of testing data for conformity to Benford law. The included methods are graphical analysis, comparison of statistical moments and employment of the mentioned statistical tests on transactional data sets; one with legitimate data and another including fraudulent activity.

Based on our statistical tests results, we conclude that the variance size does not play a significant role when testing data for Benford conformity. Out of the two, the most reliable statistical test is the Chi^2 test since its test power is comparably much greater than the Kolmogorov-Smirnov test.

We conclude that Benford's law has a place in anti-money laundering processes for transactional data since the law was proven to be reliable when examining the data sets as it was able to correctly discern between the legitimate and fraudulent data sets. However, one has to be careful of trade patterns within data sets to avoid misleading results.

We want to thank Ignace, who went far and beyond of what could be expected from a supervisor. Any shortcomings in this thesis are ours, and ours alone.

Table of Contents

“Money laundering is giving oxygen to organised crime.” – Enrique Peña Nieto

1. Purpose

As long as there is money to be made by conducting illegal business activities, there will be a cat-and-mouse game between perpetrators and law enforcement in a battle of who can outwit the other. As auditing procedures and investigations for detecting money laundering are in constant development and refinement, so are the structures and methods enabling the conduction of white-collar crime.

There is a constant flow of newsfeeds on the matter, whether it pertains to high-profile decision-makers with hidden assets abroad, to a local restaurant reported for having undeclared workers. Money laundering activity spans all levels of wealth. Its differing methods are carried via everything from luxurious international casino gambling to siphoning money from decadent drug den operations. Due to its polymorphic and elusive nature, no universal way of detecting fraudulent activity is established. However, since money laundering includes generation of illegal income and requires bank transactions at some point, the task of detection has grown into a collaborative effort between auditors, financial institutions, and law enforcement white-collar crime units. In the case of banks and auditors, data analysis has become an integral part of the anti-money laundering (AML) process. However, the available detection models are incomplete, and the science around it is young and limited.

With this in mind, this paper serves to scrutinise an elegant mathematical model that has been around since the 1800s but only recently applied to economic fraud detection - with successful results. The model in question is Benford’s law. We will try to determine under which statistical conditions it could be applied with beneficial effects, and when the model might fail to provide applicable information (or even mislead).

2. Theoretical Framework

2.1 Explaining Benford's Law

Benford's law had its origins in 1881 when astronomer Simon Newcomb observed that pages with logarithmic tables between 1 and 3 were more worn than higher numbers in publicly available library books. Newcomb then theorised that smaller digits are more frequently occurring than larger digits, as the first significant digits in logarithmic numbers. He then applied this theory to two formulas which displays the probability of a digit's frequency as the first (d1) and second (d2) significant digit:

$$p(d1) = log_{10}\left(1 + \frac{1}{d1}\right) \qquad d1 = 1, 2, \ldots, 9 \qquad (1)$$

$$p(d2) = \sum_{k=1}^{9} log_{10}\left(1 + \frac{1}{10k + d2}\right) \qquad d2 = 0, 1, 2, \ldots, 9 \qquad (2)$$

It is principal to recognise that the integer 0 is invalid as a leading digit. 0 is only allowed when testing digits beyond the first (Newcomb, 1881). If one applies each digit to formula 1 and 2, one can create a distribution table (see table 1) along with a bar-chart (see diagram 1), displaying the distribution.

Digit	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1st	0.0%	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%
2nd	12.0%	11.4%	10.9%	10.4%	10.0%	9.7%	9.3%	9.0%	8.8%	8.5%

Table 1. *According to Benford's law, the distribution of the first and second significant digit (Newcomb, 1881).*

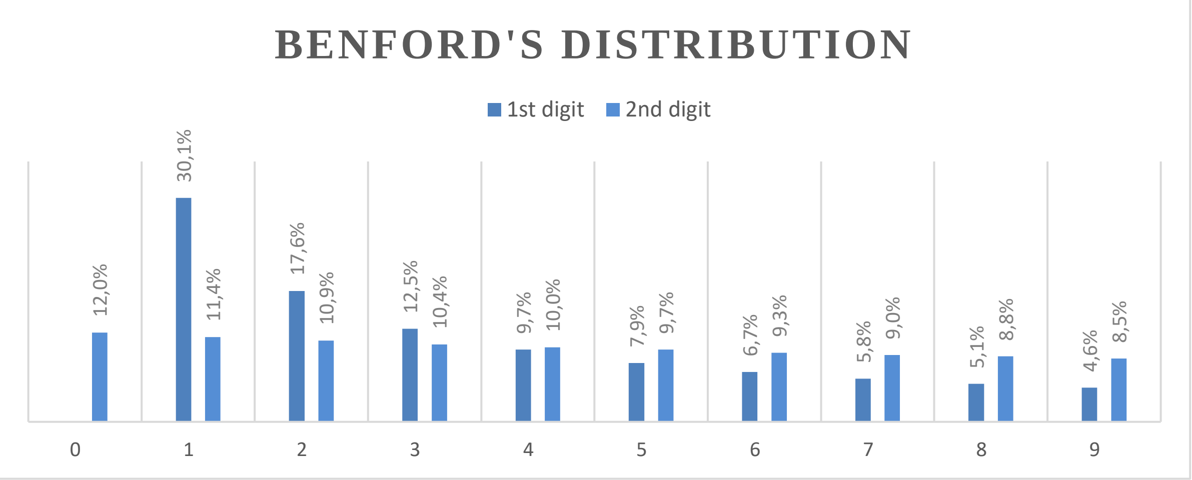


Diagram 1. *The distribution of Benford's law for the first and second significant digit (Newcomb, 1881).*

Later, in 1937 a physicist named Frank Benford rebirthed the law that Newcomb had recognised. Benford (1938) applied the law's expectations of first significant digits in naturally generated data series with more than 20 000 data points. These data series included countries' population sizes and random statistics in newspapers. Benford then published his findings in his paper the *Law of Anomalous Numbers*.

Over the years many elaborations have been added to the original formulas established by Newcomb. Perhaps one of the more useful developments was Ted Hill's common general significant-digit law, published in 1995, in which one can read the distribution of the first and higher-order significant digits. Hill (1995) defines this general significant-digit law neatly in a formula:

$$p(D_1 = d_1, \dots, D_k = d_k) = log_{10}\left(1 + \left(\sum_{i=1}^{k} d_i\, 10^{k-j}\right)^{-1}\right) \qquad (3)$$

$$\forall k \in Z, \quad d1 \in \{1, 2, \dots, 9\}\ and\ d_j \in \{0, 1, \dots, 9\}, j = 2, \dots, k$$

The probability for digit 2 to be the second significant digit of a number is around 10.9 % according to formula 2. However, assuming that the first significant digit is 1, formula 3 shows that the second significant digit's probability of being 2, is 11.5 %. The reason why is because the significant digits are dependent (Hill, 1995). Dependence between digits decreases as the distance between digits increases. If one applies Hill's formula to the first two significant digits, one can obtain the probability for integers of two leading significant digits (see table 2).

d1\d2	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**p(d1)**
1	4.1%	3.8%	3.5%	3.2%	3.0%	2.8%	2.6%	2.5%	2.3%	2.2%	30.1%
2	2.1%	2.0%	1.9%	1.8%	1.8%	1.7%	1.6%	1.6%	1.5%	1.5%	17.6%
3	1.4%	1.4%	1.3%	1.3%	1.3%	1.2%	1.2%	1.2%	1.1%	1.1%	12.5%
4	1.1%	1.0%	1.0%	1.0%	1.0%	1.0%	0.9%	0.9%	0.9%	0.9%	9.7%
5	0.9%	0.8%	0.8%	0.8%	0.8%	0.8%	0.8%	0.8%	0.7%	0.7%	7.9%
6	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.6%	0.6%	0.6%	6.7%
7	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.5%	5.8%
8	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	5.1%
9	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	0.4%	0.4%	0.4%	4.6%
p(d2)	12.0%	11.4%	10.9%	10.4%	10.0%	9.7%	9.3%	9.0%	8.8%	8.5%	100.0%

Table 2. *The distributions of the first two significant digits, according to Benford's law (Hill, 1995).*

2.2 Utilising Benford’s Law

For one to understand how Benford’s law applies to economics, it is helpful to understand the context in which it is used:

Based on data from 161 countries over a period ranging from the years 1950 to 2009, economists Ceyhun Elgin and Oguz Oztunali made calculations on the size of shadow economies in the global market, suggesting that the black market makes up over 20% of world GDP (Elgin & Oztunali, 2012). This income not only goes untaxed but enables organised criminal activity and the funding of terrorism, leading to the creation of the intergovernmental organisation Financial Action Task Force (FATF) by an initiative of G7 (Chohan, 2019).

FATF is the leading intergovernmental Anti-Money Laundering/Combating the Financing of Terrorism (AML/CFT) organisation. With 39 membership countries and affiliations with several regional intergovernmental AML/CFT organisations and collaboration with affiliated governmental institutions, FATF sets international standards and principles for the conduction of AML/CFT operations. FATF also leaves flexibility on its practical implementations to the respective country-specific institutions (Chohan, 2019).

The central framework is based around the “FATF International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation” on anti-money laundering, and the “FATF IX Special Recommendations” on CFT (however, this paper will be focused on AML methods). A large part of the recommendations pertains to what measures should be taken by financial institutions and non-financial businesses and professions to work in an AML/CFT-compliant manner (FATF, 2012-2020). Among these are:

- “Customer due diligence and record-keeping”.

- Centred around keeping and reviewing customer-specific information to identify potential red flags and enabling accountability should fraudulent activity occur.

- “Additional measures for specific customers and activities”.

- A per this recommendation section, particular scrutiny is required for modern technology, wire transfers and other services for transferring money or value that can be more susceptible to money laundering

- “Reporting of suspicious transactions”.
  - Including responsibilities to monitor and analyse such data to identify potential money laundering.
- “Internal controls and foreign branches and subsidiaries”.
  - Requiring financial institutions to implement corporation-wide policies and information exchange regarding AML/CFT, keeping the main branch accountable, should fraud be prevalent in its subsidiaries.
- “Higher-risk countries”.
  - Enhanced due diligence is demanded for specific countries with a high-risk profile, as determined by FATF.

Failure on the part of a financial institution to comply with FATF regulations (along with national country-specific regulations) can, and has, resulted in hefty sanctions. In Sweden alone, all four of the leading systemically critical financial banks (Swedbank, Handelsbanken, Nordea and SEB) have in the past five years been sanctioned by the Swedish financial supervisory authority, Finansinspektionen (FI), with fines ranging from tens of millions to billions SEK (Finansinspektionen, 2015-2020).

Due to its prevalence, a higher responsibility has been placed on the financial institutions carrying out these transfers, as reflected in the FATF recommendations. The stricter obligations have been set up due to that transfer of funds is integral to the three stages of money laundering:

Stage 1, Placement: In this first stage, the criminal’s goal is to deposit the illegally acquired profits into the financial system. In theory, criminals can achieve this by securing monetary instruments, which are then deposited into accounts located elsewhere, or by dividing large amounts of cash into more inconspicuous portions that can then be deposited directly into bank accounts. Other methods include exchanging currency and currency smuggling, dispersing the funds to reduce suspicion, or casino gambling (providing receipts for winnings).

Stage 2, Layering: When the money is in the system, the perpetrator's goal is to move the funds around through a series of transactions to disassociate the ill-gotten gains from its source. By buying and selling investment instruments, purchasing and trading assets, moving the money across the globe through various institutions such as off-shore banks or complicit businesses, often passing through jurisdictions that are non-compliant in AML (thus the need for extra scrutiny of transactions to/from higher-risk countries).

Finally, the money re-enters the legitimate economy via ostensibly legitimate sources, e.g., payments from foreign shell corporations as payments for goods or services. At this point, the money is not tied to its origin, enabling the last stage.

Stage 3, Integration: Now the money is available to be integrated into the legitimate economy, often by investing in businesses, luxury goods or fictitious loans that can generate legitimate-looking income (FATF, 2020; Schneider & Niederländer 2008).

Traditionally, auditors are tasked to go through company financials and via traditional auditing procedures detect and report any anomalies that could indicate fraudulent activity. These auditing procedures are typically not very substantial as auditors usually observe the distributions of data sets graphically as opposed to statistical analysis (Carlton Collins, 2017). However, as international regulations vary, and the global market has become more and more integrated, so has the potential for masking illegally generated wealth by transferring money around, forging ledgers and running money laundering fronts in international networks.

The growth in complexity has led to new approaches to detecting criminal activity. In the past couple of decades, the AML/CFT sector has grown more focused on data analysis. Todays' machine learning models and statistical tests are commonly used to flag suspicious behaviour (that would otherwise have been hard or impossible to detect by traditional methods due to the sheer quantity of transactions). However, less unintuitive, although useful, analytical data methods are also utilised, such as Benford's law.

Hal Varian first suggested this use of Benford's law in 1972 as he assumed that money launderers who forge information distribute their digits relatively evenly. One can also expect that accountancy data should be naturally generated as tampering with such data is an economic crime. Hence, the first significant digit of financial statements, e.g., balance sheet items, conform to the distribution specified by Benford's law (Henselmann et al., 2013).

2.3 Previous research

One of the most influential researchers on the application of Benford’s law has been Mark Nigrini. He presented methods and research on the applicability of distribution testing in economics and auditing procedures (Nigrini & Mittermaier, 1997; Nigrini, 2019). In his study, he shows how common traits in fraudulent economy data such as a higher frequency of rounded numbers, duplications of authentic transactions or high frequencies of numbers close to internal thresholds contribute to the deviation from a Benford distribution.

There have also been studies where Benford’s law has been used in the detection of falsified scientific data when measuring the distribution of reported coefficients (some studies also use standard deviations):

When measuring the coefficients and standard error from *Empirica* and *Applied Economics Letters*, i.e., scientific publications on economics, Günnel and Tödter (2008) could show that the data conformed to Benford’s law.

Hüllemann et al. (2017) compared data from 25 scientific articles from a specific field of medicine, of which 12 were proven to be fabricated, and the remaining 13 were not. Their test showed a 100% sensitivity in failing the falsified papers. The specificity, however, was only 46.15% as 6 of the articles followed the law.

Hein et al. (2012) tested 20 redacted anaesthesiology publications. They could show that 17 of them showed significant deviations from Benford’s law of first digits, and 18 showed substantial divergences when tested against Benford’s law of second digits. The control was based on a meta-analysis that showed that other articles in the field were consistent with Benford’s first digit’s law.

There have been numerous studies showing Benford’s law’s prevalence in naturally occurring distributions. In processes where the distribution can be described by power laws and the variable spans several orders of magnitude in a logarithmic structure, Benford’s law is often found. Benford’s paper shows that these are more or fewer requirements for data to be Benford distributed. Examples of such distributions are stock prices, areas and lengths of rivers, stock prices (Kvam & Vidakovic, 2007) and gene sets in digital gene expressions (Karthik et al., 2016).

2.4 The Law's Limitations

There are, however, some limitations to the law. For Benford's law to be a valuable method when testing data sets, one would typically have to examine a large data set. Data sets with 50-100 numbers have been proven to be in harmony with the law, but some expert opinions state that testing 500 or more data entries in sample size is more appropriate (Carlton Collins, 2017). As usual, the reliability of the test usually increases with larger data sets.

Suppose the leading digits in the data sets have unequal chances of occurring. In that case, Benford's law will have little relevance statistically (i.e., the possibility must exist for a significant digit to be a digit between 1-9) (Carlton Collins, 2017). If one were to examine the heights of students at Lund University, one would likely find that a large majority are between 150 and 210 centimetres tall. In this case, applying Benford's law would result in an over-representation of the first significant digits of 1 and distort one's conclusion.

As data sets must be naturally generated, one cannot use Benford's law in rounded data since it would damage the test's reliability. Any rounding could change the first significant digit and change the outcome of observed frequencies (Carlton Collins, 2017).

When it comes to economic data, it is often crucial that the measured data is generated from more than one distribution, to avoid human bias. A product's price tends to be adjusted for psychological effect (e.g., 199€, $2995, 100kr), and thus not Benford distributed. However, a receipt sum, income over a certain period, or the production cost of a product is generally a function of different prices, quantities, deductions et cetera generally making them Benford distributed (Janvresse & de la Rue, 2004).

Another weakness of Benford's law is that there is never any definite proof when employing it (Carlton Collins, 2017). Suppose one applies Benford's law to a relevant data set in AML analysis and concludes that its distribution is unlike Benford's distribution. In that case, one cannot establish that someone has laundered money. Although, this would raise suspicion of money laundering, which in hand would justify further research.

3. Testing Benford Conformity

Applying statistical tests to Benford's law is possible and often recommended when analysing a data set's conformability to Benford's law. In this paper, we are testing Benford law's strengths and weaknesses by applying statistical tests on our data sets. Hypothesis testing gives us a basis for conclusions when examining the data sets. The hypothesis test is defined as:

H_0: The observed frequencies conform to Benford's law.

H_1: The observed frequencies do not conform to Benford's law.

In this segment, below are moments and statistical tests that we use in this paper to analyse data sets' conformability to Benford's law.

3.1 Moments – Mean, Variance, Kurtosis & Skewness

As stated previously, Benford's law explains the characteristics of significant digits in logarithmic numbers. Consequently, one can observe a specific distribution of significant digits. Due to this expectation, one can examine a data set to see whether it, for example, matches a typical Benford distribution's mean. Peter Dale Scott and Maria Fasli managed to gather the values for these measures by looking at well over half a million data entries and then extracted the data sets that met Benford's law conditions. They listed their findings in their report from 2001.

According to Scott's and Fasli's paper (2001), if a data set conforms to Benford's law, then the first significant digit's mean should be around 3.440. The mean for the first two significant digits should be about 38.590. One can also analyse whether a data set's variance conforms to Benford's law. The variance for the first significant digit in a Benford distributed data set should be around 6.057, and for the first two significant digits, the variance should be 621.832. Further, the kurtosis should be around –0.548 and the skewness around 0.796 for the first digit. Lastly, for the first two digits, the kurtosis should be around −0.547 and the skewness around 0.772.

This method could give oneself a better basis for comparison between the actual frequencies and those predicted. If a data set deviates from the statistical moments it could indicate that the data is fraudulent. At the very least, it may well support further research of a data set.

Moments	First Significant Digit	First Two Significant Digits
Mean	3.440	38.590
Variance	6.057	621.832
Skewness	0.796	0.772
Kurtosis	-0.538	-0.547

Table 3. *Moments with values conforming precisely to Benford's law (Scott & Fasli, 2001).*

3.2 R-package BenfordTests

In order to statistically test data sets for Benford conformity, we first need a method. We chose to use the software R, which is suitable when running statistical tests.

In this paper, we use an R package named BenfordTests created by Dieter William Joenssen (2015), a professor in mechanical engineering and materials science at Aalen University in Germany. This package can be used in Benford's law studies and includes several test scripts of statistical tests used to examine data sets' distributions. We chose to use two statistical tests, the chi^2 test and the Kolmogorov-Smirnov test, as they are commonly used when assessing data for Benford conformity.

3.2.1 Chi^2 Test

The first test we apply is the chi^2 test. The chi^2 test is a commonly used test and applicable for testing categorical variables and their relationship, which we want to do when comparing a set of numbers and their distribution to a set distribution (i.e., Benford's law). It is strongly recommended to only use the chi^2 when the data set is large as its statistical power will otherwise be small (Stephens, 1970). There are many variations of chi^2 tests. In this paper, we will be using Pearson's chi^2 Goodness-of-Fit Test for Benford's law which has been plucked from the R package by Joenssen (2015). Joenssen defines the chi^2 test statistic as:

$$\chi^2 = n \cdot \sum_{i=10^{k-1}}^{10^k-1} \frac{(f_i^o - f_i^e)^2}{f_i^e} \tag{4}$$

f_i^o represents observed frequencies of significant digits i and f_i^e is the expected Benford frequency of i.

3.2.2 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test compares the cumulative percentage frequencies between the observed and expected data, allowing oneself to assess a data set's agreement with Benford's law. Stephens (1970) claims that the Kolmogorov-Smirnov test is a more reliable test when checking smaller data sets. The disadvantage with the Kolmogorov-Smirnov test is that it is more sensitive to deviations near the centre of the distribution than at the tails, which may distort conclusions (Stephens, 1970). Joenssen (2015) defines the statistical test in his R package as:

$$D = \sup_{10^{k-1},\ldots,10^k-1} \left| \sum_{j=1}^{i} \left(f_j^o - f_j^e\right) \right| \cdot \sqrt{n} \tag{5}$$

f_j^o signifies the observed frequencies of significant digits i and f_j^e is the expected Benford frequency of i.

4. Data sets

This paper investigates how two different data sets perform when one applies Benford's law to them. The first of these data sets we strongly believe not to have any fraudulent activity whilst the other we believe does.

The first data set examined are government expenditures of Oklahoma in 2019. This data set consists of their collective agency expenditures and serves as an informational and educational data source for Oklahomans to gather insights in governmental activities. This data set consists of 110 696 entries of expenditure data. We suspect that this data set does not consist of fraudulent activity. At the very least, no report of such behaviour has been made, and those responsible for the accounts know that the spending is under heavy scrutiny.

The other data set that we are investigating is the Azerbaijani Laundromat, a money laundering enterprise between 2012 and 2014 in which influential residents of Azerbaijan funnelled money through four shell-corporations based in the United Kingdom. Thanks to this operation, elite members of Azerbaijan were, for instance, able to money-launder and sway political decisions in Europe. According to The Organized Crime and Corruption Reporting Project (OCCRP), suspicious transactional data of 2.5 billion euro was leaked to the Danish newspaper Berlingske. The data set consisted of 17 000 payments from and to the shell-corporations whose bank accounts were held at Danske Bank's Estonian branch. The newspaper then shared this knowledge with the OCCRP who instigated an investigation of the transactional data. The ploy was examined with the help of a collaborative effort from several financial institutions. In 2017 the OCCRP published the transactional data from the shell-corporations on their website so that readers can do their research. The data gives a good picture of who the involved parties were and how the money was spent.

We know that this data set contains transactions from fraudulent activity. Hence, the data set should include data that is not representative of naturally generated business expenses.

4.1 Potential causes for non-conformity in transactional data

Some common examples of fraudulent behaviour indicators that we look for in transactional business-to-business data are presented below.

- Duplicated values of authentic transactions:

  To appear legitimate, duplicates of authentic transactional numbers are injected so that the amounts do not stick out in an auditing procedure (Nigrini, 2019). However, duplicates are often not picked uniformly from the authentic data so as to retain a Benford distribution. That way, they will contribute to a deviation from Benford's law, as the leading digits of the chosen numbers will be overrepresented (Nigrini, 2019).

- Bias towards rounded numbers:

  Fraudulent transactions tend to be rounded more than their counterparts. The cause of this can be the human bias towards rounding numbers, or (as in the famous case of Enron) be a direct or indirect result of fraudulent earnings management (Nigrini, 2005). In this case, Benford's law of second (or further) digits tend to show greater efficacy, as rounded numbers tend to output a higher frequency of zeroes.

- Numbers close to certain thresholds:

  Banks and institutions often run specific protocols and procedures that trigger when a particular transaction requires further investigation. It can be an unusually sizeable transfer from a client, several transfers at a limit value (such as mobile payment limit) or large transfers to high-risk countries/clients. In order to avoid getting detected, a number above or below the threshold is picked. However, since these numbers are not generated by a natural process and are not spread over multiple orders of magnitude, they tend to deviate from the Benford distribution as well.

- General deviation from the Benford distribution:

  The measured transactions might include payments that are not based on naturally generated income or expense amounts. This can often be the case in the layering stage of money laundering when the perpetrator splits the money up in seemingly arbitrary quantities and move them around. These quantities tend to not to follow a Benford distribution.

5. Testing the credibility of the statistical tests

Before applying the statistical tests (chi^2 and Kolmogorov-Smirnov) to the two different data sets, we must test their credibility and reliability given specific settings. A concerning issue is how the tests perform in varying sample and variance sizes. By adjusting these conditions, we will better understand if there are circumstances wherein the tests are well-performing and if there are any in which they are not. Firstly, we will review how the statistical tests perform when the null hypothesis is true. By doing this, we can tell how well the practical rejection rates fit our set significance levels which could tell us if the tests are impaired by over or under rejections depending on sample and variance sizes. Secondly, we will investigate the tests' power to determine if they are susceptible to a type II statistical error (false negative).

5.1 Empirical size test

By testing the statistical tests' rejection rates when the null hypothesis is true, we can amass information whether the tests overreact or underreact, depending on sample sizes, significance level and variance. Thanks to our R package published by Joenssen (2015), we can test this in R. As for test conditions, we generate samples at sizes presented in table 4 and 5. We use samples generated (by the R script) from a Benford distribution with the density function:

$$f(x) = \frac{1}{x \cdot \ln(10)} \forall x \epsilon [1, 10] \quad (6)$$

We modify the data by using variance multipliers, as presented in tables 4 and 5, to see if it affects the tests' reliability, and to what extent. The p-values are evaluated asymptotically from the χ^2 statistic in the chi^2 test, and by simulating the p-values with 100 replicates in the Kolmogorov-Smirnov test. For each combination of the size, variance and α parameters, the sampling and testing are reiterated 2000 times to get results that can identify test consistency with relatively high reliability.

A useful indication of the tests' performance is the rejection rates conformity to significance level. Suppose the rejection rates deviate from a specified significance level (α). In that case, it could be a problem since the practical significance levels differ from our set significance level. In tables 4 and 5, we run simulations for different sample sizes, variances, and significance levels for the rejection rates in the two tests:

d_i: 2	α: 0.01 Variance Multiplier 1	30	50	70	α: 0.05 Variance Multiplier 1	30	50	70	α: 0.10 Variance Multiplier 1	30	50	70
Sample Size	Chi² test											
500	0.010	0.013	0.010	0.008	0.041	0.055	0.036	0.054	0.101	0.099	0.077	0.102
1000	0.011	0.008	0.007	0.014	0.036	0.052	0.055	0.054	0.084	0.089	0.108	0.111
1500	0.007	0.013	0.014	0.012	0.051	0.060	0.044	0.055	0.099	0.110	0.116	0.103
2000	0.013	0.012	0.015	0.016	0.056	0.053	0.056	0.064	0.097	0.104	0.113	0.117
2500	0.007	0.008	0.013	0.010	0.053	0.058	0.054	0.059	0.103	0.102	0.105	0.100
3000	0.003	0.008	0.006	0.007	0.044	0.052	0.045	0.043	0.107	0.100	0.088	0.091
3500	0.014	0.012	0.011	0.008	0.056	0.049	0.044	0.050	0.094	0.101	0.085	0.102
4000	0.011	0.008	0.004	0.009	0.050	0.051	0.048	0.047	0.096	0.102	0.109	0.091
4500	0.011	0.010	0.013	0.009	0.053	0.053	0.047	0.052	0.118	0.093	0.099	0.102
5000	0.008	0.009	0.010	0.010	0.052	0.050	0.043	0.050	0.096	0.096	0.093	0.088
Sample Size	Kolmogorov-Smirnov test											
500	0.014	0.012	0.013	0.015	0.057	0.048	0.048	0.060	0.126	0.111	0.101	0.124
1000	0.011	0.010	0.011	0.009	0.051	0.045	0.043	0.050	0.112	0.109	0.104	0.114
1500	0.007	0.011	0.008	0.006	0.053	0.050	0.053	0.048	0.111	0.119	0.115	0.107
2000	0.010	0.015	0.008	0.012	0.060	0.058	0.061	0.049	0.122	0.127	0.107	0.116
2500	0.017	0.013	0.011	0.014	0.056	0.047	0.047	0.045	0.116	0.116	0.098	0.122
3000	0.010	0.012	0.011	0.009	0.049	0.054	0.058	0.053	0.095	0.115	0.116	0.114
3500	0.007	0.011	0.005	0.007	0.039	0.039	0.043	0.035	0.098	0.088	0.098	0.094
4000	0.007	0.011	0.011	0.006	0.044	0.038	0.042	0.043	0.098	0.093	0.096	0.097
4500	0.015	0.011	0.015	0.016	0.057	0.056	0.055	0.049	0.110	0.110	0.097	0.109
5000	0.007	0.006	0.012	0.009	0.036	0.046	0.048	0.035	0.100	0.113	0.100	0.092

Table 4. *Rejection rates for the first two significant digits (di) when testing the empirical size of the chi² test and Kolmogorov-Smirnov test. Sample sizes range between 500-5000, variance multipliers vary between 1-70, and the three significance levels are 0.01, 0.05 and 0.10. The further a rejection rate strays from the significance level, the darker the colour formatting. The simulation has been reiterated 1000 times for the sake of consistency.*

Observing table 4's rejection rates, one can infer that the tests seem to perform reasonably well when the null hypothesis is true. In both tests, the rejection rates do not seem to be affected by the significance level in any discerning way, i.e., no significance level performs better or worse than the other. There are, however, some noteworthy takeaways worth considering when using these tests to test data sets for Benford conformity.

The first takeaway is the implication of the variance size. By looking at table 4, one can see a slight change in the tests' rejection rates based on the size of variance. Multiplying the variance by 30 shows that the rejection rates are changed in both statistical tests. Though, as the variance is multiplied by 30, the rejection rates do not always under-reject or over-reject. By multiplying the variance by 50 and then 70, it is comprehensible that the tests' rejection rates are not severely affected based on the size of variance. Our interpretation based on this result is that Benford's law is relatively invariant to differences in variance when the null hypothesis is true. This conclusion indicates that the risk of committing a type I error when testing data sets conformity to a Benford distribution does not depend on the variance size.

The second takeaway is the importance of the sample size. By glancing at the colour formatting of the chi^2 test in table 4 when the significance level is 0.05, one may ascertain that the rejection rates conform closer and closer to the significance level as the sample sizes increase. This is not as apparent when looking at the corresponding rejection rates in the Kolmogorov-Smirnov test. Nevertheless, the rejection rates' conformity to significance levels as the sample sizes increase is not apparent when reviewing table 4. The Kolmogorov-Smirnov test is often used as a substitute for the chi^2 test when the sample is small. Based on both tests' rejection rates in table 4, the sample size does not seem to affect the rejection rates. In case there are any discerning differences in rejection rates for smaller sample sizes, another simulation is run (see table 5).

	α: 0.01				α: 0.05				α: 0.10			
	Variance Multiplier				Variance Multiplier				Variance Multiplier			
d_i: 2	1	30	50	70	1	30	50	70	1	30	50	70
Sample Size	Chi^2 test											
100	0.019	0.021	0.023	0.018	0.065	0.060	0.074	0.059	0.123	0.115	0.122	0.113
200	0.023	0.015	0.012	0.015	0.055	0.059	0.060	0.060	0.110	0.104	0.115	0.101
300	0.010	0.011	0.014	0.008	0.038	0.059	0.044	0.054	0.101	0.114	0.096	0.098
400	0.021	0.014	0.016	0.021	0.072	0.062	0.059	0.068	0.128	0.114	0.108	0.118
500	0.018	0.015	0.015	0.011	0.063	0.058	0.060	0.065	0.105	0.107	0.102	0.110
600	0.011	0.010	0.015	0.014	0.049	0.055	0.057	0.061	0.099	0.113	0.110	0.105
700	0.012	0.009	0.009	0.006	0.051	0.034	0.037	0.051	0.097	0.087	0.093	0.097
800	0.016	0.007	0.012	0.012	0.053	0.056	0.059	0.068	0.119	0.108	0.101	0.118
900	0.010	0.010	0.007	0.007	0.040	0.045	0.040	0.046	0.093	0.101	0.096	0.088
1000	0.013	0.016	0.014	0.012	0.045	0.058	0.050	0.053	0.083	0.108	0.106	0.095
Sample Size	Kolmogorov-Smirnov test											
100	0.013	0.009	0.013	0.010	0.048	0.040	0.050	0.038	0.116	0.118	0.108	0.106
200	0.009	0.007	0.005	0.013	0.050	0.052	0.046	0.052	0.097	0.108	0.104	0.110
300	0.007	0.007	0.008	0.013	0.049	0.049	0.053	0.057	0.104	0.104	0.112	0.108
400	0.010	0.016	0.010	0.009	0.043	0.053	0.055	0.042	0.106	0.121	0.114	0.095
500	0.009	0.013	0.009	0.010	0.053	0.043	0.051	0.049	0.112	0.108	0.103	0.120
600	0.009	0.009	0.009	0.006	0.048	0.042	0.050	0.042	0.103	0.110	0.107	0.105
700	0.006	0.007	0.006	0.012	0.053	0.050	0.055	0.056	0.119	0.102	0.124	0.117
800	0.010	0.010	0.014	0.008	0.050	0.050	0.056	0.053	0.103	0.115	0.117	0.108
900	0.004	0.011	0.006	0.006	0.049	0.056	0.050	0.045	0.092	0.120	0.112	0.109
1000	0.010	0.009	0.011	0.011	0.053	0.039	0.038	0.054	0.118	0.104	0.102	0.122

Table 5*. Rejection rates for the first two significant digits (di) when testing the empirical size of the chi^2 test and Kolmogorov-Smirnov test. The rejection rates differ depending on sample size, variance size and significance level (α). Sample sizes range between 100-1000, variance multipliers vary between 1-70, and the three significance levels are 0.01, 0.05 and 0.10. The further a rejection rate strays from the significance level, the darker the colour formatting. The simulation has been reiterated 2000 times for the sake of consistency.*

Table 5 shows that the rejection rates for the chi^2 test stray away slightly from the specified significance levels when the sample is small. This is visible for sample sizes of 100 as the rejection rates are around 0.02 when the significance level is 0.01. When the sample size is 200, the same holds for the chi^2 test. However, the Kolmogorov-Smirnov's rejection rates seem to conform nicely to the significance level when the sample size is at its smallest (100 and 200).

Given this result, it looks like the Kolmogorov-Smirnov test outperforms the chi^2 test when the sample size is equal to 100. At the very least, the difference between rejection rate and significance level for the Kolmogorov-Smirnov test is smaller for all significance levels when the sample size is 100. The same seems to hold when the sample size is 200. This simulation appears to support the importance of one's attention to sample sizes and one's application of appropriate statistical tests. As the probability of a type I error increases with small sample sizes for the chi^2 test (around 100 to 200 sample size), the opportunity to use the Kolmogorov-Smirnov test seems more appealing. Still, as the sample sizes increase, the difference between the two statistical tests diminishes and eventually, the disparity is tough to distinguish. Although the differences between rejection rates in the two tests are not large, the case still stands that the Kolmogorov-Smirnov test seems more suitable when the null is true.

5.2 Power test

In this simulation, we test the power of the two statistical tests depending on variance size and sample size. By performing this simulation, we can tell if the variance or sample size impacts the tests' susceptibility to a type II statistical error (false negative; β).

We generate samples based on the Azerbaijani data set, which is confirmed fraudulent (OCCRP, 2017), to see to what degree the chi^2 and Kolmogorov-Smirnov tests are able to detect non-Benford distributed data sets. Random samples are picked from our data, at sizes ranging from 50 to 10 000. The variance is scaled to four different levels to determine to what extent these parameters affect the reliability of the tests.

When estimating β, for the chi^2 testing, the p-values are determined by the asymptotic χ^2 distribution. For Kolmogorov-Smirnov, the p-values were simulated and estimated with 50 replicates.

For each combination of size and variance parameters, this whole simulation was repeated 2000 times to achieve high reliability in our results. We set the significance level to $\alpha = 0.05$.

di:1 α: 0.05	**Chi² test**				**Kolmogorov-Smirnov test**			
	Variance Multiplier				Variance Multiplier			
Sample Size	1	30	50	70	1	30	50	70
50	0.162	0.129	0.108	0.168	0.072	0.085	0.065	0.075
100	0.331	0.244	0.217	0.313	0.078	0.084	0.081	0.096
250	0.711	0.614	0.603	0.660	0.107	0.104	0.141	0.156
500	0.961	0.931	0.933	0.950	0.175	0.174	0.236	0.266
1000	1.000	1.000	0.998	0.999	0.317	0.269	0.480	0.482
2000	1.000	1.000	1.000	1.000	0.722	0.523	0.842	0.815
3000	1.000	1.000	1.000	1.000	0.948	0.707	0.972	0.961
5000	1.000	1.000	1.000	1.000	1.000	0.948	1.000	1.000
10000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 6. *The power of the different tests (1- β) at the given sizes and variance multipliers. Each cell contains the power of the test at the given parameters.*

Table 6 shows that the test power for both chi^2 and Kolmogorov-Smirnov is relatively low for the smaller sample sizes, with both tests having marginal increases as the sample sizes increase. Chi^2 reaches a power level of 0.93 for samples sizes of 500, regardless of variance, and approximating a power close to 1 at sample sizes of 1000 and higher. Kolmogorov-Smirnov first achieves an average test power close to ~0.9 at samples of 3000, and power close to one for sample sizes of 5000.

While both tests have low power at our smaller sample sizes, the chi^2 significantly outperforms the Kolmogorov-Smirnov test at all sample sizes up to 10 000. The chi^2 test shows a test power tolerable by some standards at sample sizes of 500, whilst the Kolmogorov-Smirnov test needs sample sizes of 3000 to reach a similar test power. The chi^2 test reaches a power of approximately one at sample sizes of 2000, while the Kolmogorov-Smirnov test requires sample sizes of 10000 to reach that golden standard.

We can see an apparent discrepancy between the two tests, as the Kolmogorov-Smirnov test may under-reject samples deviating from Benford’s law, making it more prone to type II errors.

Studying table 6, we can see that the power changes as the variance increases. However, the increase in variance does seem to affect the tests’ power in a way that can be deemed significant. We were able to gather a similar conclusion when we tested rejection rates for the true hypothesis. This seems to support the argument that Benford’s law is relatively invariant to variances.

5.3 Variance analysis

After increasing the variance of our data at different increments and running our tests, we could not discern any significant deviation from the base values (i.e., non-modified variance) for any of the variance levels.

We suspect that this could be due to the supposed scale invariance for Benford distributed numbers. Since Benford's law holds at vastly different levels of magnitude, with the law present in everything from measured sizes of galaxies, down to microbial processes, there is an argument that there is a scale invariance in Benford distributed variables. This claim has been challenged by some, though.

Since Benford's law works when there is a broad logarithmic probability distribution present, the data to which it is applicable should be logarithmically distributed, ranging over several magnitudes. With that in mind, the relative distribution in the data set should be of greater importance than the numbers' absolute sizes. This has been shown to hold in many cases, as the unit in which the data is expressed, such as USD versus euro, does not break Benford's law (which is supported by the fact that Benford's law is used in AML and auditing across different countries and currencies).

Working under the assumption that Benford's law is scale-invariant, we can say that a numeric vector X consisting of Benford distributed data generates a certain probability vector Y of the first significant digits in X:

$$X \rightarrow Y$$

Assuming scale invariance:

$$X \cdot c \rightarrow Y$$

$$c \in Q$$

If

$$c \setminus \{0, 1\}$$

We know that

$$Var(X \cdot c) = Var(X) \cdot c^2$$

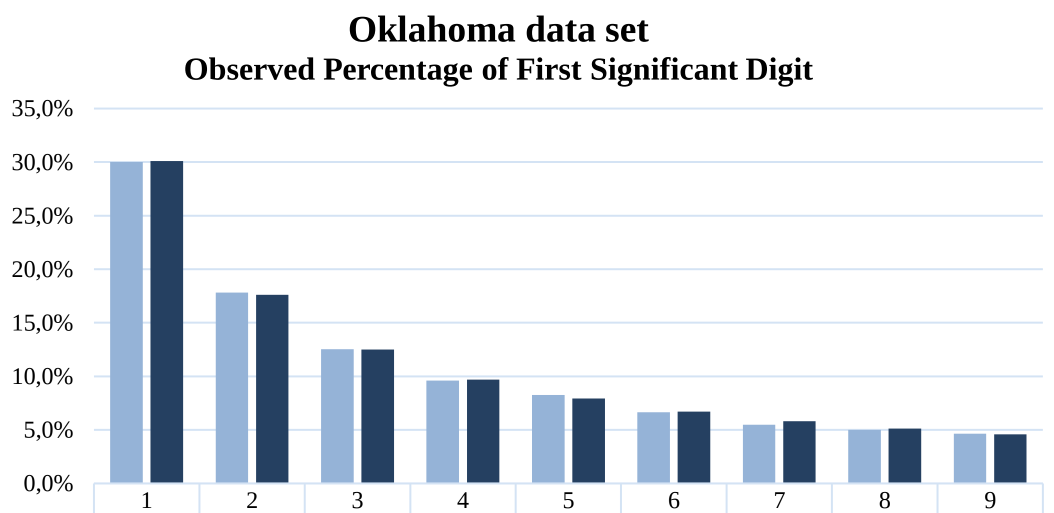
$$Var(X) \neq Var(X) \cdot c^2$$

By that logic, the scaled vector (X·c) should have a different variance than X, but if scale invariance holds, then both the vectors should generate the same percentages in Y when extracting the first digits. Hence, the data set variance should not necessarily affect the results - which is reflected in ours.

6. Testing data sets for conformity

6.1 Observed distributions of the first significant digits

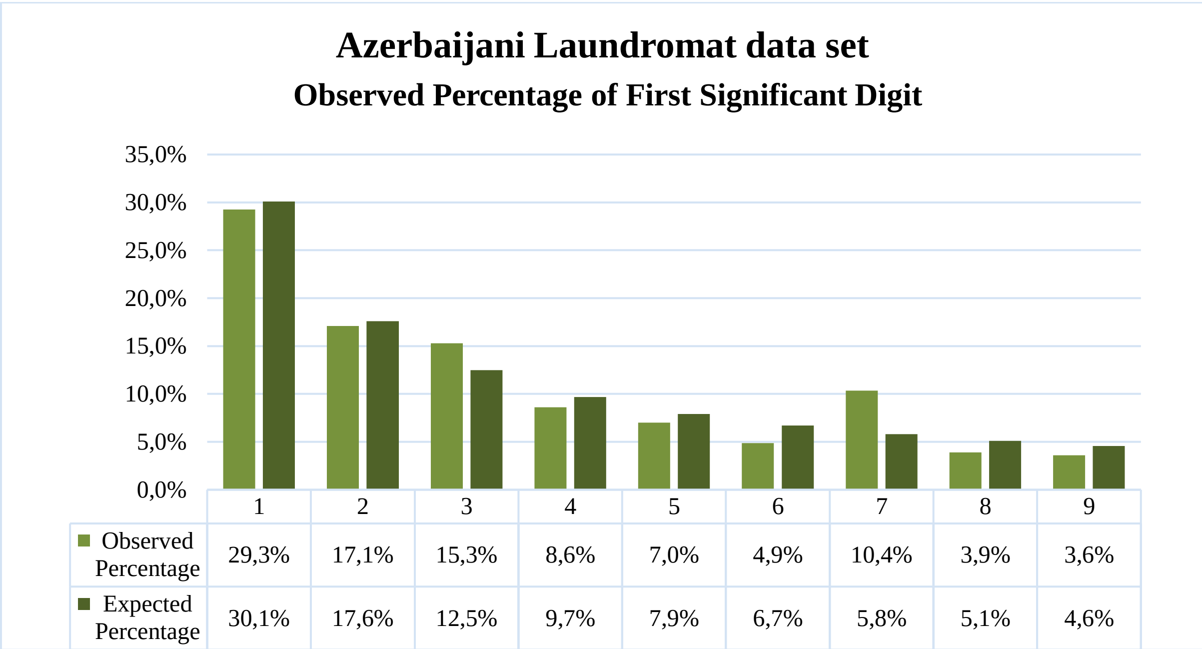
A common practice when examining data sets for fraudulent activity is investigating the distributions of the first significant digit. By doing this, we might be able to tell whether a data set is fraudulent. This test is performed for the Oklahoma government expenditure data and the transactional data of the Azerbaijani Laundromat. Displayed below, in diagram 1 and 2 are the distributions for the two different data sets:



	1	2	3	4	5	6	7	8	9
Observed Percentage	30,0%	17,8%	12,5%	9,6%	8,3%	6,6%	5,5%	5,0%	4,6%
Expected Percentage	30,1%	17,6%	12,5%	9,7%	7,9%	6,7%	5,8%	5,1%	4,6%

***Diagram 2**. Comparing the percentages of the observed frequencies of the first significant digit in the data set of Oklahoma's government expenditures with the expected first significant digit percentages of Benford's law.*

Diagram 2 tells us that the Oklahoma data set is closely conformed to a Benford distribution. This result implies that the data set does not reject the null hypothesis. If we merely were to base a conclusion on this result, we would assume that there is no fraudulent activity within the data set of Oklahoma government expenditures (which is our expectation of the data set).



	1	2	3	4	5	6	7	8	9
Observed Percentage	29,3%	17,1%	15,3%	8,6%	7,0%	4,9%	10,4%	3,9%	3,6%
Expected Percentage	30,1%	17,6%	12,5%	9,7%	7,9%	6,7%	5,8%	5,1%	4,6%

Diagram 3. *Comparing the percentages of the observed frequencies of the first significant digit in the Azerbaijani Laundromat data set with the expected first significant digit percentages of Benford's law.*

Diagram 3 shows that the Azerbaijani Laundromat scheme's transactional data is quite Benford distributed. For the significant digits 1 and 2, the distribution is tightly knitted to a Benford distribution. However, the percentage difference is conspicuous for the remaining digits. This might be especially true for the significant digit 7, where the observed percentage is 10.4 % whilst the expected percentage, according to Benford's law is 5.8 %.

The distribution depicted in diagram 3 is most likely a result of the data set being fraudulent. Despite that, it is essential to recognise that if a data set is entirely fraudulent, it would, in theory, be more apparent in the distribution of the data set.

There could be some other reasons why the Azerbaijani data set distribution is not as closely conformed to a Benford's distribution. For example, it could have to do with specific trading patterns in services between companies. One of the most extensive associates could only trade in one good or service and, therefore, skew the data set's distribution. As mentioned by the OCCRP (2017), some legitimate business transactions likely took place with these shell-companies, which would be reflected in the data set. If we look at the graphics in diagram 3 impartially, it would be tough to discern whether the Azerbaijani data set is Benford distributed or is not. Nonetheless, the Azerbaijani data set is not as closely conformed to a Benford distribution as the Oklahoma data set.

As we have mentioned, auditors often apply this method of visual analysis when analysing data sets for fraudulent activity. It could be a useful method to gain some initial information about data sets' distributions. However, we know that "looks" can be deceiving, and sure enough, when applying statistical tests on the data set, we obtain contradicting results.

6.2 Data sets' statistical moments

Looking at the statistical moments of the first significant digits and the first two significant digits of the two data sets, we can see that both present moments that relatively closely approximate the reference values presented in table 7 and 8.

Oklahoma data set	First digit	First two digits
Mean	3.044	37.694
Variance	6.541	619.116
Skewness	0.797	0.791
Kurtosis	-0.450	-0.466

Table 7. *Sample moments of the Oklahoma data set.*

Azerbaijani data set	First digit	First two digits
Mean	3.441	38.552
Variance	5.880	606.696
Skewness	0.761	0.736
Kurtosis	-0.644	-0.633

Table 8. *Sample moments of the Azerbaijani data set.*

The Oklahoma data set has a mean that is ~0.4 lower than the reference value for its first digits, and a ~0.9 lower mean for its first two significant digits. The sample variance is ~0.5 higher for single digits and ~2.7 lower for two digits. Skewness closely approximates the reference value, having a skewness of ~0.001 over reference for first digits and ~0.019 over for the first two digits. The kurtosis is ~0.11 higher than reference for both first digits and first two digits.

For the Azerbaijani data set, the first digit mean is ~0.001 higher and first two digits mean ~0.38 lower than the reference values. The variance is ~0.18 lower for first digits and ~15.1 lower for two digits. Skewness is ~0.035 lower for first digits and ~0.036 lower for first two digits. The kurtosis is ~0.11 lower for first significant digits and ~0.86 lower for first two significant digits.

When comparing the data sets' moments, we can see that even though the Azerbaijani data set has been shown to deviate more from Benford's law than the Oklahoma data set, it is not necessarily reflected in the statistical moments. One could speculate that the mean does not play a significant role in determining Benford's law conformity, that two-digit testing is more relevant than one-digit testing when it comes to individual values, that skewness is a strong indicator of Benford's law conformity and that a higher and lower kurtosis respectively could imply a particular inclination to tampered data. In practice though, the only conclusion we can make from reviewing the statistical moments is that both data sets' moments show high proximity to the reference values and that our results do not seem to indicate statistical significance one way or another. Further testing is required to confirm or refute the idea of comparing statistical moments for this purpose.

6.3 Running the tests on the complete data sets

Now that we have gathered some crucial information about the statistical tests, we can apply the tests to the two data sets and conclude the law's functionality in fraud detection.

The first tests we are going to perform is applying the chi^2 and Kolmogorov-Smirnov tests on the complete data sets. By running these tests, we might be able to tell if a data set is Benford distributed or not. Given our statistical tests' performance results, we can assume that they should perform properly since the data sets' sample sizes are large. Table 9 presents the test statistics and p-values for the data set with Oklahoma's government expenditures. Table 10 displays statistics and p-values for the Azerbaijani Laundromat transactional data.

Statistical tests on the Oklahoma data set	**Statistic**	**p-value**
Chi^2 test	44.82	< 3.98E-07
Kolmogorov-Smirnov test	1.68	< 2.00E-02

Table 9. *Critical statistical values and p-values of the Oklahoma data set. The data set includes 110 697 data entries of expenditure data from the Oklahoma government in 2019.*

When applying the statistical tests on the Oklahoma data set's complete expenditure data, the statistical tests show small p-values for the data set, which indicates a rejection of the null hypothesis. By solely relying on this result, we would assume that the data set is not Benford distributed. We expect the data set to be Benford distributed, though, which seems to be supported by the distribution in diagram 2. Hence, we must ask ourselves if the contradicting results could be explained some other way.

The first question is if the data set has any unique characteristics in terms of, formerly mentioned, trading patterns. Since the data set is vast, 110 697 data entries, it is likely that agencies with particular expenses contribute to a disproportion which could imbalance the distribution. Additionally, when examining the Oklahoma data set, many agencies have payments that do not differ monthly. For example, some agencies' fixed costs such as rent of office buildings or subscription services. Furthermore, some expenses, such as salaries, are rounded, which also could have implications on the data set.

Statistical Tests for the Azerbaijani data set	**Statistic**	**p-value**
Chi^2 test	925.93	< 2,2e-16
Kolmogorov-Smirnov test	3.0807	< 2,2e-16

Table 10. *Critical statistical values and p-values of the Azerbaijani Laundromat data set. The data set includes 16 940 transactions from companies registered in the UK during 2012-2014.*

According to the statistical tests, the Azerbaijani Laundromat data set does not conform to a Benford distribution indicated by the small p-values that they output (see table 10). Both statistical tests suggest that the data set is not Benford distributed. This could indicate that the data set is fraudulent and therefore, further testing is endorsed. As we expect the data set to be fraudulent, since money was funnelled to politicians in Europe, this result strengthens our assumptions.

Comparing the two data sets, we can see that the statistical tests outputted lower p-values than the Oklahoma data set. However, both tests reject the null. As we know that our statistical tests perform correctly when sample sizes are large, the result of the Oklahoma data set likely has to do with characteristics within the data set. Nevertheless, since the statistical tests of the Oklahoma data set showed unexpected results, we need to do more testing to gather insights into the reliability of statistical testing of transactional data in association with Benford's law.

6.4 Specific subsampling

Since we suspect that the data sets might contain characteristics that impede Benford's law's efficiency, we feel obligated to test specific subsamples within the data sets. Hence, we test specific transaction partners to see if any agents (companies, agencies) show particularly incongruent patterns with Benford's law. This could tell us a bit about the law's practical sensitivities to idiosyncrasies in data sets. It could also highlight agents with a suspicious activity that might be a cause for the skewed distributions of the Azerbaijani data set.

Tests are conducted on the transactional data for 31 beneficiaries with the highest number of transactions (to maximise test-reliability) in the Azerbaijani data set. The sample size of transactions is simply too low to include after the first 31 beneficiaries. Agents with similar amounts of transactions were hand-picked from the Oklahoma data set (to get a fair comparison).

Specific subsampling tests performed for the Azerbaijani Laundromat data set are available in table 11. For subsample tests of specific agencies in the Oklahoma data set, see table 12.

Azerbaijani data set - specific subsample	α: 0.05	Chi²		Kolmogorov-Smirnov	
Beneficiary	**Sample Size**	**Crit. Value**	**p-value**	**Crit. Value**	**p-value**
LCM ALLIANCE LLP	1758	117.470	2.20E-16	2.043	2.20E-16
METASTAR INVEST LLP	1237	178.100	2.20E-16	1.537	9.00E-03
HILUX SERVICES LP	905	339.100	2.20E-16	1.986	2.20E-16
FABERLEX LP	334	14.772	6.37E-02	1.418	6.00E-03
POLUX MANAGEMENT LP	312	187.850	2.20E-16	1.386	9.00E-03
KG COMMERCE LLP	273	5.349	7.20E-01	0.505	5.83E-01
LOTA SALES LLP	246	23.166	3.16E-03	0.671	3.61E-01
BONINVEST LLP	234	4.269	8.32E-01	0.344	9.16E-01
INFOCREST LLP	201	13.381	9.94E-02	0.743	3.11E-01
GFG EXPORT LLP	178	17.515	2.52E-02	0.866	1.59E-01
RIVERLANE LLP	147	35.749	1.95E-05	0.758	2.87E-01
DATEMILE ALLIANCE LLP	117	2.322	9.70E-01	0.699	2.15E-01
MOLONEY TRADE LLP	115	28.802	3.44E-04	0.729	3.17E-01
BONDWEST LLP	113	3.533	8.97E-01	0.352	8.89E-01
AVROMED COMPANY LLP	102	10.317	2.44E-01	0.673	4.44E-01
RICHFIELD TRADING L.P.	101	7.004	5.36E-01	0.477	8.75E-01
AVROMED COMPANY	79	3.663	8.86E-01	0.563	5.64E-01
RICHFIELD TRADING L.P.	101	7.004	5.36E-01	0.477	8.88E-01
OVERMOND LLP	83	2.322	9.70E-01	0.699	2.30E-01
RASMUS LP	82	2.322	9.70E-01	0.699	2.23E-01
AVROMED COMPANY	79	3.663	8.86E-01	0.563	5.83E-01
WILLROCK UNITED LLP	71	8.554	3.81E-01	0.569	5.64E-01
REDPARK SALES CORP	70	41.793	1.48E-06	1.559	3.00E-03
LINSTAR SYSTEMS CORP.	61	23.001	3.36E-03	1.123	5.70E-02
SABA CARS GERMANY GMBH	54	2.500	9.62E-01	0.739	2.41E-01
JETFIELD NETWORKS LIMITED	53	15.563	4.91E-02	0.504	6.69E-01
CROSSPARK LINES LLP	53	27.839	5.06E-04	0.995	1.03E-01
MOYA ENGINEERING LLP	52	4.624	7.97E-01	0.441	7.55E-01
MUROVA SYSTEMS LLP	49	4.624	7.97E-01	0.441	7.64E-01
BENTCARD IMPORT LLP	49	3.507	8.99E-01	0.554	6.00E-01
GREENOUGH TRADE LLP	46	20.104	9.95E-03	0.864	1.91E-01

Table 11. *Chi² and Kolmogorov-Smirnov critical values with p-values of specific subsamples based on the Azerbaijani Laundromat data set's largest beneficiaries. Red p-values are lower than the significance level (α: 0.05).*

Upon viewing the results from the specific subsample simulations, one can see that the different beneficiaries have different results regarding Benford conformity.

According to the chi^2 test, 13 out of 31 beneficiaries have p-values lower than the significance level of 0.05. Among these 13 beneficiaries are the three largest recipients, in terms of count of transactions. The chi^2 test expresses that the majority of the beneficiaries have p-values lower than the significance level.

The Kolmogorov-Smirnov test displays that six beneficiaries have distributions that do not conform to a Benford distribution. Among these six beneficiaries are the five largest beneficiaries. According to the Kolmogorov-Smirnov test, a large majority of the p-values are larger than the significance level. As we have stated previously, the Kolmogorov-Smirnov test's power can be meagre depending on the sample size. Since our most significant subsample size is 1758 (after that drastically decreasing), we assume that the output in many of the Kolmogorov-Smirnov tests are (since the Azerbaijani data set is deemed fraudulent).

In sample sizes under 1000, it is likely that some of the chi^2 tests generate false negatives since the test power decreases under this size (see table 6). Therefore, it is feasible that there are more beneficiaries in table 11 that are not Benford distributed.

Oklahoma data set - specific subsample	α: 0.05	**Chi²**		**Kolmogorov-Smirnov**	
Agency	**Sample Size**	**Crit. Value**	**p-value**	**Crit. Value**	**p-value**
MENTAL HEALTH AND SUBSTANCE ABUSE SERV.	1759	45.331	3.19E-07	2.736	2.20E-16
OKLA. CITY COMMUNITY COLLEGE	1263	11.827	1.59E-01	0.855	2.08E-01
SEMINOLE STATE COLLEGE	957	26.459	8.76E-04	1.285	2.90E-02
UNIV.OF SCIENCE & ARTS OF OK	891	9.270	3.20E-01	0.882	2.00E-01
OKLAHOMA TAX COMMISSION	846	17.269	2.74E-02	1.054	1.20E-01
J.D. MCCARTY CENTER	835	35.614	2.07E-05	1.440	2.20E-16
OKLA. BUREAU OF NARCOTICS AND DANGEROUS	832	14.619	6.70E-02	0.848	1.80E-01
DISTRICT ATTORNEYS COUNCIL	820	13.795	8.73E-02	1.763	2.20E-16
DEPARTMENT OF COMMERCE	805	22.213	4.54E-03	1.363	2.00E-02
OSU-EXPERIMENT STATION	773	16.647	3.40E-02	1.014	6.00E-02
BOLL WEEVIL ERADICATION ORG.	378	44.025	5.63E-07	2.451	2.20E-16
INTERSTATE OIL COMPACT COMM.	364	11.144	1.94E-01	1.450	5.00E-03
PARDON AND PAROLE BOARD	342	21.013	7.11E-03	1.342	2.70E-02
COURT OF CRIMINAL APPEALS	294	20.125	9.88E-03	1.691	2.20E-16
ST.WIDE VIRTUAL CHARTER SCHOOL BOARD	285	29.795	2.30E-04	1.601	2.00E-03
BD. OF EXAM. FOR LT CARE ADMIN.	273	13.040	1.11E-01	0.832	2.12E-01
DISTRICT COURTS	251	14.531	6.89E-02	0.912	1.52E-01
BD. OF PSYCHOLOGISTS EXAMINERS	240	46.292	2.09E-07	1.501	4.00E-03
COUNCIL ON JUDICIAL COMPLAINTS	237	64.105	7.25E-11	0.777	2.97E-01
LEGISLATIVE SERVICE BUREAU	232	54.144	6.47E-09	1.738	1.00E-03
ST. BD. OF CHIROPRACTIC EXAM.	232	20.186	9.65E-03	1.071	8.60E-02
BOARD OF EXAMINERS IN OPTOMETRY	227	16.781	3.25E-02	0.642	4.34E-01
BD. OF PRIV. VOCATIONAL SCHOOLS	211	83.601	9.22E-15	2.580	2.20E-16
ENERGY RESOURCES BOARD	193	5.339	7.21E-01	0.353	8.90E-01
OK. INDUSTRIAL FINANCE AUTH.	192	192.050	2.20E-16	3.594	2.20E-16
OFFICE OF LIEUTENANT GOVERNOR	173	39.726	3.60E-06	1.778	2.20E-16
BD OF LIC ALCOHOL & DRUG COUNS	153	39.932	3.30E-06	0.994	1.28E-01
OKLAHOMA TURNPIKE AUTHORITY	153	40.396	4.04E+01	1.512	5.00E-03
NATIVE AMER.CULTURAL & EDUC. AUTH-OK.	148	48.651	7.42E-08	1.366	6.00E-03
UNIV. HOSPITALS AUTHORITY	58	19.842	1.10E-02	0.596	4.67E-01
OUHSC PROF. PRAC. PLAN.	47	22.906	3.49E-03	2.021	2.20E-16

Table 12. *Chi² and Kolmogorov-Smirnov critical values with p-values of specific subsamples based on the Oklahoma data set agencies. Red p-values are lower than the significance level (α: 0.05).*

Viewing table 12, one can see that the majority of the agencies have p-values lower than the significance level of 0.05. This is true for the chi^2 test and the Kolmogorov-Smirnov test. In the chi^2 test, 22 out of 31 subsamples have p-values lower than the significance level. In the Kolmogorov-Smirnov tests, p-values are lower than the significance level in 18 out of 31 subsamples. Since both tests perform relatively well when the null hypothesis is true (see table 12), we conclude that the high amount of rejection rates likely has to do with the data set's composition. As we have mentioned before, the Oklahoma data set agencies have unique and repeating trading patterns.

The specific subsample tests show exceedingly low specificity for fraud detection in both data sets. Several agents deemed reliable in the Oklahoma data set were rejected, and other agents that the OCCRP has stated partook in the Azerbaijani money laundering scheme (2017) were accepted. We conclude that this type of testing is highly unreliable if one wants to investigate specific agents within a data set. This type of subset structure violates one of the law's concepts which has to do with the magnitude over which the data spans. One does not generally expect a company, as a whole, to buy its pencils, computer software and company vehicles from the same vendor, as each company often is specialised in making a particular product. Since specific trading patterns within subsamples disrupt the logarithmic structure that is present in Benford distributed data, we can conclude that company-specific transaction testing such as ours is in most cases unsuitable for singling out fraudulent operators.

6.5 Random subsampling of the data sets

Given our results when applying statistical testing to the complete Oklahoma data set, we must examine the data sets under different circumstances. By reviewing random subsamples of the data sets, we hope to eliminate some factors present in the data sets' dispositions.

To test random subsamples of the data sets, we run 100 tests on each data set in which we let R decide, based on a random seed, which data points to include. The tests have a sample size of 1000, which has been established to be sufficient when testing for Benford conformity (Carlton Collins, 2017). However, a sample size of 1000 should also allow us to practically evaluate our results from the power simulation wherein the Kolmogorov-Smirnov test underperformed at this level. The test results are available as tables (tables 13 & 14) in the appendix.

Running a random subsample test on the Oklahoma data set, we find that most of the p-values in the chi^2 test are higher than the significant level (0.05). Out of the 100 simulations, eight instances show p-values lower than the significance level. 92 out of 100 chi^2 subsample tests do not reject the null hypothesis. Due to the simulation results in which we test for type I errors, we can expect a rejection rate of around five percent (which is the defined significance level). A discrepancy of three rejections is a good result, and we can conclude that the Chi^2 test performs well in this circumstance. Nearly all the tests do not reject the null. Based on this result, we conclude with statistical significance that the data set is Benford distributed.

The Kolmogorov-Smirnov test also shows, with statistical significance, that the Oklahoma data set is Benford distributed as the p-values are relatively high in most of the subsample tests. However, five tests show the opposite, but with a significance level of 0.05, we predict an expected value of 5 instances in which the null hypothesis is rejected.

When we apply this method to the Azerbaijani data set, we see that all the chi^2 subsample tests reject the null. This would indicate that the data set is not Benford distributed, which confers with our expectation of the data set. Therefore, according to the chi^2 test, there is an argument to suspect fraudulent activity in the data set. However, the p-values of the Kolmogorov-Smirnov test only exceed the significance level in 22 simulations of 100 meaning; the two tests contradict each other. As we know from the power simulation, the chi^2 test performs well at a sample size of 1000 whilst the Kolmogorov-Smirnov test does not (see table 6). In fact, the Kolmogorov-Smirnov test has a low statistical power at this sample size based on our tests (around 30-50 percent depending on the variance size). Hence, it is not unexpected that the test rarely rejects the alternative hypothesis. In reality, it is instead telling of which of these statistical tests one should rely on when testing for Benford conformities.

7. Conclusion

7.1 Conclusion of statistical tests

With the help of the power test and the test of rejection rates when the null hypothesis is true, we can conclude what statistical test is appropriate when testing a data set's conformity to Benford's law. Our testing on the effect of variance at different levels showed that the variance does not seem to have much of an impact on Benford testing, which we theorise is in accordance with the law's scale invariance.

When the null is true, and the sample size is between 100-200, we can say with assurance that the chi^2 test performs slightly worse comparably to the Kolmogorov-Smirnov test (in terms of the rejection rates conformity to the set significance level). However, by testing the power of the test, we see that the Kolmogorov-Smirnov test is profoundly less powerful than the chi^2 test, even in small sample sizes of 100-200. This tells us that the chi^2 test is better at detecting data sets that do not conform to Benford's law which is unexpected considering that previous researchers recommend the Kolmogorov-Smirnov test for small sample testing.

Judging by our results in these simulations, a decision of which test to use could be based on which error type (I or II) one needs to minimise for the purpose at hand. In an AML setting, we realise that a type II error, where one may fail to detect fraudulent activity, is seemingly worse than making a type I error.

Our conclusion of the tests is that they perform well given that the sample size is large. In terms of power, the chi^2 test requires a sample size around 1000 to output a β of ~0% whilst the Kolmogorov-Smirnov test requires a size of 5000 to get the same reliability. This is later affirmed when applying random subsampling tests, in which we use sample sizes of 1000. At that level, the chi^2 test has a power of approximately 1, while the power of the Kolmogorov-Smirnov test is ~0.317. Visible in table 14 in the appendix, the Kolmogorov-Smirnov test strongly under-rejects in comparison to the chi^2 test when the sample contains fraudulent data.

We expected that the Kolmogorov-Smirnov test would outperform the chi^2 test in small sample sizes. Although this might be true for when comparing the tests, empirical sizes, the advantages of applying the chi^2 test for small sample testing are far greater than the

Kolmogorov-Smirnov test. We deem that the gain of test power far outweighs the cost of over-rejections, when the sample is small, in the chi^2 test's empirical size.

7.2 Conclusion of Benford's Law

The bar-chart (diagram 2) tells us that the Oklahoma data set is closely conformed to a Benford distribution. Diagram 3 shows that the Azerbaijani data set deviates more than the Oklahoma data set. However, there are only slight indications that hints at the data set's fraudulence. If we did not have previous knowledge about fraudulent activity in the Azerbaijani data set, it would be tough to make assumptions by merely visually inspecting the distribution. It is however logical to assume that a completely fraudulent data set would be a more clear-cut case.

The data sets' moments do not produce any dependable results as they both conform to Benford moments. Generally, it is hard to distinguish when a data set can be interpreted as non-conforming by reviewing the moments. Though, this needs further investigation as we only explore two, albeit large, data sets.

We find that these examinations are not trustworthy. The appliance of statistical testing is, according to us, a more reliable source of information. A data set might only contain partial fraudulence, and therefore it might not be reflected clearly in diagrams and moments.

When testing the complete data sets, we can rely on the chi^2 test and the Kolmogorov-Smirnov test as the sample sizes are large enough for the test power and empirical size not to be impeded. We find that analysing the distribution of the entire data sets could lead to misleading conclusions. Whilst the statistical tests give us expected outcomes for the Azerbaijani data set, they give us unexpected results for the Oklahoma data set. As we know that the conditions of sample sizes are met, we believe that the outcome has to do with some specific subsets' idiosyncratic compositions, which is an issue that should not be ignored when testing this type of data. Our conclusions also support this as running tests on the specific subsamples the test chi^2 test rejects the null hypothesis for a majority of the measured subsets in the Oklahoma data. Furthermore, by reviewing the agents within the data sets, we can see that some indeed have specific trading patterns that will distort the data set. Therefore, we suggest that this type of testing is not ideal to single out specific benefactors for fraudulent behaviour.

To mitigate the idiosyncratic data, we run random subsample tests generated from the data sets. As the random subsample size is 1000, we rely on the results of the chi^2 since it has been proven to be reliable for this sample size. Our test results for the Oklahoma data set now align with the expected rejection rates of a Benford distribution. Additionally, the rejection rates indicate that the Azerbaijani Laundromat data set is not Benford distributed, signifying fraud.

7.3 Closing remarks

We conclude that using the chi^2 Benford test on large random subsamples from suspicious data sets might have a place in AML procedures. The Kolmogorov-Smirnov test is not as reliable due to its low test power in subsamples less than 5000. When applying Benford's law, one needs to be wary of potential trading patterns in transactional data as it could lead to misleading results.

8. References

Benford, F. (1938). The Law of Anomalous Numbers, *Proceedings of the American Philosophical Society,* vol. 78, no. 4, pp. 551-572, Available online: https://www.scribd.com/document/209534421/The-Law-of-Anomalous-Numbers [Accessed 27 September 2020]

Carlton Collins, J. (2017). Using Excel and Benford's Law to detect fraud, Available online: https://www.journalofaccountancy.com/issues/2017/apr/excel-and-benfords-law-to-detect-fraud.html [Accessed 3 November 2020]

Chohan, U.W. (2019). The FATF in the Global Financial Architecture: Challenges and Implications, *CASS Working Papers on Economics & National Affairs*, Available Online: https://ssrn.com/abstract=3362167 [Accessed 21 December 2020]

Financial Action Task Force. (2012-2020). International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation, FATF, Paris, France, Available online: www.fatf-gafi.org/recommendations.html [Accessed 5 December 2020]

Finansinspektionen. (2020). Swedbank fined for serious deficiencies in its measures to combat money laundering, Available online: https://www.fi.se/en/published/press-releases/2020/swedbank-fined-for-serious-deficiencies-in-its-measures-to-combat-money-laundering/ [Accessed 15 December 2020]

Finansinspektionen. (2015). Handelsbanken receives a remark and is ordered to pay 35 million, Available online: https://fi.se/en/published/press-releases/2015/handelsbanken-receives-a-remark-and-is-ordered-to-pay-35-million/ [Accessed 15 December 2020]

Finansinspektionen. (2020). SEB receives remark and administrative fine of SEK 1 billion, Available online: https://www.fi.se/sv/publicerat/sanktioner/finansiella-foretag/2020/seb-faranmarkning-och-en-miljard-kronor-i-sanktionsavgift/ [Accessed 15 December 2020]

Finansinspektionen. (2015). Nordea receives a warning and is ordered to pay 50 million, Accessed online: https://fi.se/en/published/press-releases/2015/nordea-receives-a-warning-and-is-ordered-to-pay-50-million/ [Accessed 15 December 2020]

Financial Action Task Force. (2020). Money laundering, Available online: https://www.fatf-gafi.org/faq/moneylaundering/ [Accessed 5 December 2020]

Government of Oklahoma. (n.d.). OKLAHOMA'S OPEN DATA, Available online: https://data.ok.gov/ [Accessed 8 December 2020]

Government of Oklahoma. (n.d.). Expenditure Summary, Available online: https://data.ok.gov/dataset/expenditure-summary [Accessed 8 December 2020]

Günnel, S., & Tödter, K.H. (2009). Does Benford's Law hold in economic research and forecasting?, *Empirica 36*, pp. 273–292, Available online: https://doi.org/10.1007/s10663-008-9084-1 [Accessed 25 November 2020]

Hein, J., Zobrist, R., Konrad, C., Schuepfer, G. (2012). Scientific fraud in 20 falsified anesthesia papers : detection using financial auditing methods, *Anaesthesist*, vol. 61, no. 6, pp. 543-549, Available online: https://www.researchgate.net/publication/225307184_Scientific_fraud_in_20_falsified_anesthesia_papers [Accessed 15 December 2020]

Henselmann, K., Scherr, E., Ditter, D. (2013). Applying Benford's Law to individual financial reports: An empirical investigation on the basis of SEC XBRL filings, *Working Papers in Accounting Valuation Auditing*, no. 2012-1 [rev.], Available online: https://www.econstor.eu/handle/10419/88418 [Accessed 7 December 2020]

Hill, T.P. (1995). A statistical derivation of the significant-digit law. Statistical science, vol. 10, no.4, pp. 354-363 Available online: https://projecteuclid.org/euclid.ss/1177009869/ [Accessed 4 November 2020]

Hüllemann, S., Schüpfer, G., Mauch, J. (2017). Application of Benford's law: a valuable tool for detecting scientific papers with fabricated data? : A case study using proven falsified articles against a comparison group, *Anaesthesist*, pp. 795-802, Available online: https://link.springer.com/article/10.1007%2Fs00101-017-0333-1 [Accessed 24 2020]

Janvresse, É., de la Rue, T. (2004). From Uniform Distributions to Benford's Law, *Journal of Applied Probability*, vol. 41, no. 4, pp. 1203-1210, Available online: https://web.archive.org/web/20160304125725/http://lmrs.univ-rouen.fr/Persopage/Delarue/Publis/PDF/uniform_distribution_to_Benford_law.pdf [Accessed 2 January 2021]

Joenssen, D.W. (2015). Package 'BenfordTests', Available online: https://cran.r-project.org/web/packages/BenfordTests/BenfordTests.pdf [Accessed 19 November 2020]

Karthik, D., Stelzer, G., Gershanov, S. et al. (2016). Elucidating tissue specific genes using the Benford distribution, *BMC Genomics*, vol. 17, pp. 595, Available online: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2921-x [Accessed 27 December 2020]

Kvam, P., Vidakovic, B. (2007) Nonparametric Statistics with Applications to Science and Engineering, *Wiley-Interscience Publication*, pp. 170, Available online: http://zoe.bme.gatech.edu/~bv20/isye6404/Bank/npmarginal.pdf [Accessed 10 December 2020]

Newcomb, S. (1881). Note on the Frequency of Use of the Different Digits in Natural Numbers, *American Journal of Mathematics*, vol. 4, no. 1, pp. 39-40, Available online: https://www.jstor.org/stable/2369148?refreqid=excelsior%3A838d018747dfb32243a8280c81c13701&seq=1#metadata_info_tab_contents [Accessed 10 November 2020]

Nigrini, M.J., Mittermaier, L.J. (1997) The use of Benford's law as an aid in analytical procedures, *auditing*, vol. 16, no. 2, pp. 52, Available online: https://www.econbiz.de/Record/the-use-of-benford-s-law-as-an-aid-in-analytical-procedures-nigrini-mark/10006996665 [Accessed November 15 2020]

Nigrini, M. (2005). An Assessment of the Change in the Incidence of Earnings Management Around the Enron-Andersen Episode, *Review of Accounting and Finance*, vol. 4, pp. 92-110, Available online: https://www.researchgate.net/publication/241628957_An_Assessment_of_the_Change_in_the_Incidence_of_Earnings_Management_Around_the_Enron-Andersen_Episode [Accessed 15 November 2020]

Nigrini., M.J. (2019), The patterns of the numbers used in occupational fraud schemes, *Managerial Auditing Journal*, vol. 34, no. 5, pp. 606-626. Available online: https://www.emerald.com/insight/content/doi/10.1108/MAJ-11-2017-1717/full/html [Accessed 13 November 2020]

Organised Crime and Corruption Reporting Project. (2017). The Azerbaijani Laundromat, Available online: https://www.occrp.org/en/azerbaijanilaundromat/ [Accessed 9 December 2020]

Organised Crime and Corruption Reporting Project. (2017). The Raw Data, Available online: https://www.occrp.org/en/azerbaijanilaundromat/raw-data/ [Accessed 9 December 2020]

Schneider, F., & Niederländer, U. (2008). Money Laundering: Some Facts, *European Journal of Law and Economics*, vol. 26, pp. 387-404, Available online: https://www.researchgate.net/publication/23534449_Money_Laundering_Some_Facts [Accessed 17 November 2020]

Stephens, M. A. (1970). Use of the Kolmogorov–Smirnov, Cramér–Von Mises and Related Statistics without Extensive Tables, *Journal of the Royal Statistical Society*. Available online: https://www.semanticscholar.org/paper/Use-of-the-Kolmogorov-Smirnov%2C-Cramer-Von-Mises-and-Stephens/e608d9fabc7f182b932f8bfc1e465f78ad65a51a [Accessed 22 November 2020]

9. Appendix

	Oklahoma Random Sample Test			
	n: 1000 α: 0.05	d_i: 1	N: 1000 α: 0.05	d_i: 1
Simulation	**Chi^2 Statistic**	**K-S p-value**	**K-S Statistic**	**K-S p-value**
1	10.518	2.31E-01	0.632	0.454
2	1.688	9.89E-01	0.344	0.920
3	4.277	8.31E-01	0.475	0.666
4	6.701	5.69E-01	0.754	0.342
5	54.115	6.56E-09	0.884	0.177
6	8.010	4.33E-01	0.596	0.525
7	3.022	9.33E-01	0.442	0.765
8	8.787	3.61E-01	0.473	0.719
9	11.094	1.96E-01	0.786	0.262
10	4.145	8.44E-01	0.345	0.913
11	6.873	5.50E-01	0.700	0.361
12	9.401	3.10E-01	0.508	0.670
13	9.365	3.13E-01	0.472	0.743
14	7.909	4.42E-01	0.790	0.274
15	7.814	4.52E-01	0.604	0.507
16	7.244	5.11E-01	0.790	0.267
17	10.056	2.61E-01	1.045	0.075
18	4.017	8.56E-01	0.505	0.628
19	10.113	2.57E-01	0.694	0.401
20	17.278	2.73E-02	0.855	0.201
21	8.009	4.33E-01	0.501	0.681
22	45.135	3.47E-07	0.914	0.145
23	891.950	2.20E-16	3.405	0.000
24	11.325	1.84E-01	0.792	0.273
25	5.629	6.89E-01	0.478	0.695
26	7.599	4.74E-01	0.596	0.565
27	2.631	9.55E-01	0.347	0.888
28	5.032	7.54E-01	0.541	0.613
29	7.125	5.23E-01	1.269	0.027
30	20.089	1.00E-02	0.662	0.445
31	14.039	8.08E-02	0.887	0.167
32	18.159	2.01E-02	1.134	0.070
33	6.313	6.12E-01	0.633	0.469
34	8.282	4.07E-01	0.698	0.352
35	5.609	6.91E-01	0.700	0.342
36	10.337	2.42E-01	0.604	0.493
37	2.509	9.61E-01	0.351	0.893
38	8.913	3.50E-01	0.573	0.545
39	4.270	8.32E-01	0.533	0.633

40	5.003	7.57E-01	0.887	0.187
41	5.756	6.75E-01	0.344	0.898
42	10.065	2.61E-01	0.603	0.510
43	11.315	1.85E-01	0.631	0.477
44	0.631	4.77E-01	1.043	0.088
45	7.173	5.18E-01	0.313	0.947
46	2.268	9.72E-01	0.284	0.961
47	6.369	6.06E-01	0.600	0.498
48	18.451	1.81E-02	1.071	0.077
49	12.078	1.48E-01	0.600	0.525
50	8.962	3.46E-01	1.109	0.047
51	10.800	2.13E-01	0.790	0.244
52	10.274	2.46E-01	0.948	0.153
53	7.362	4.98E-01	0.950	0.134
54	2.792	9.47E-01	0.666	0.381
55	1.175	9.97E-01	0.381	0.831
56	6.367	6.06E-01	0.697	0.373
57	5.763	6.74E-01	0.636	0.433
58	6.075	6.39E-01	0.604	0.507
59	6.013	6.46E-01	0.665	0.389
60	3.518	8.98E-01	0.438	0.818
61	12.353	1.36E-01	0.855	0.197
62	5.125	7.44E-01	0.282	0.967
63	8.061	4.28E-01	0.508	0.666
64	12.002	1.51E-01	1.359	0.015
65	9.208	3.25E-01	1.137	0.061
66	14.850	6.21E-02	0.886	0.176
67	14.003	8.10E-02	0.631	0.473
68	7.009	5.36E-01	0.572	0.550
69	10.679	2.21E-01	0.881	0.194
70	13.694	9.01E-02	1.042	0.097
71	3.639	8.88E-01	0.632	0.468
72	5.685	6.83E-01	0.605	0.487
73	10.410	2.37E-01	1.174	0.038
74	5.606	6.91E-01	0.852	0.221
75	6.180	6.27E-01	0.505	0.658
76	5.614	6.90E-01	0.823	0.235
77	5.902	6.58E-01	0.473	0.742
78	9.898	2.72E-01	0.786	0.305
79	4.745	7.84E-01	0.440	0.797
80	12.539	1.29E-01	0.889	0.179
81	6.433	5.99E-01	0.318	0.937
82	4.921	7.66E-01	0.725	0.368
83	8.363	3.99E-01	0.382	0.836
84	10.613	2.25E-01	0.605	0.487
85	4.930	7.65E-01	0.446	0.764

86	1.550	9.92E-01	0.288	0.936
87	10.749	2.16E-01	0.726	0.338
88	8.634	3.74E-01	1.111	0.057
89	17.337	2.68E-02	0.692	0.394
90	12.833	1.18E-01	0.761	0.297
91	5.796	6.70E-01	0.504	0.691
92	7.981	4.35E-01	0.855	0.208
93	4.959	7.62E-01	0.763	0.291
94	6.279	6.16E-01	0.605	0.504
95	7.171	5.18E-01	0.732	0.333
96	5.657	6.86E-01	0.380	0.862
97	7.753	4.58E-01	0.794	0.265
98	4.069	8.51E-01	0.477	0.715
99	9.704	2.86E-01	0.853	0.213
100	8.325	4.02E-01	0.284	0.952

Table 1. *Critical values and p-values of the Chi^2 and Kolmogorov-Smirnov test for random subsamples (n=1000) in the Oklahoma government expenditure data set. Red colour formatting indicates p-values lower than the significance level (α: 0.05).*

	Azerbaijani Laundromat - Random Subsample test			
	N:1000 α: 0.05	d_i: 1	N:1000 α: 0.05	d_i: 1
Simulation	Chi^2 Statistic	Chi^2 p-value	K-S Statistic	K-S p-value
1	74.475	6.28E-13	0.947	0.139
2	47.772	1.09E-07	1.009	0.108
3	74.899	5.17E-13	1.453	0.007
4	82.951	1.24E-14	1.078	0.064
5	65.852	3.27E-11	1.231	0.025
6	58.038	1.13E-09	1.047	0.077
7	58.832	7.90E-10	0.889	0.166
8	72.151	1.83E-12	1.009	0.104
9	67.597	1.48E-11	1.332	0.015
10	78.534	9.64E-14	1.236	0.037
11	78.149	1.15E-13	1.427	0.008
12	63.066	1.16E-10	1.262	0.024
13	72.811	1.35E-12	0.983	0.124
14	101.490	2.20E-16	1.616	0.005
15	72.211	1.78E-12	0.947	0.144
16	54.145	6.47E-09	0.889	0.174
17	44.305	4.98E-07	0.851	0.217
18	49.560	4.97E-08	1.390	0.014
19	90.104	4.44E-16	1.010	0.088
20	93.031	2.20E-16	1.078	0.057
21	40.951	2.13E-06	0.728	0.342

22	63.171	1.11E-10	0.983	0.087
23	74.528	6.13E-13	1.015	0.097
24	34.024	4.02E-05	0.882	0.177
25	66.126	2.89E-11	1.009	0.108
26	101.940	2.20E-16	2.091	0.000
27	82.887	1.28E-14	1.104	0.059
28	58.643	8.60E-10	1.042	0.103
29	31.427	1.18E-04	0.915	0.146
30	50.802	2.87E-08	0.667	0.413
31	55.288	3.88E-09	0.819	0.225
32	65.256	4.30E-11	0.946	0.153
33	53.939	7.10E-09	0.883	0.188
34	63.826	8.23E-11	0.977	0.123
35	58.430	9.46E-10	0.946	0.128
36	126.740	2.20E-16	1.553	0.004
37	56.625	2.13E-09	1.173	0.051
38	58.579	8.85E-10	1.072	0.073
39	47.952	1.01E-07	0.662	0.391
40	85.905	3.11E-15	1.363	0.010
41	30.655	1.62E-04	0.952	0.113
42	46.384	2.01E-07	0.947	0.163
43	78.838	8.37E-14	2.091	0.000
44	58.040	1.13E-09	1.167	0.051
45	46.570	1.85E-07	0.762	0.313
46	67.856	1.31E-11	0.983	0.121
47	62.553	1.47E-10	0.857	0.214
48	40.488	2.60E-06	0.794	0.236
49	76.347	2.65E-13	0.820	0.228
50	75.456	4.00E-13	0.946	0.146
51	57.069	1.75E-09	0.983	0.109
52	49.590	4.90E-08	1.108	0.061
53	91.479	2.22E-16	1.142	0.051
54	52.652	1.26E-08	1.200	0.044
55	56.632	2.12E-09	0.952	0.151
56	54.495	5.53E-09	1.896	0.000
57	67.516	1.53E-11	0.920	0.124
58	74.702	5.66E-13	1.136	0.062
59	49.180	5.87E-08	1.010	0.117
60	49.041	6.24E-08	0.857	0.216
61	28.539	3.82E-04	0.757	0.307
62	47.834	1.06E-07	0.947	0.144
63	82.337	1.65E-14	1.015	0.117
64	31.042	1.38E-04	0.851	0.216
65	35.024	2.65E-05	0.794	0.270
66	49.961	4.16E-08	1.236	0.032
67	63.867	8.08E-11	1.079	0.061

68	78.825	8.43E-14	0.851	0.220
69	45.415	3.07E-07	1.045	0.089
70	60.475	3.76E-10	1.079	0.077
71	84.273	6.77E-15	1.009	0.130
72	34.043	3.99E-05	0.636	0.463
73	74.070	7.57E-13	0.921	0.145
74	46.171	2.21E-07	1.231	0.036
75	73.148	1.16E-12	0.921	0.133
76	38.840	5.26E-06	0.851	0.235
77	78.584	9.43E-14	0.915	0.150
78	72.347	1.67E-12	0.952	0.113
79	50.605	3.13E-08	0.978	0.125
80	78.147	1.15E-13	1.079	0.069
81	57.847	1.23E-09	1.548	0.003
82	46.006	2.37E-07	0.724	0.333
83	43.257	7.86E-07	0.630	0.460
84	36.270	1.57E-05	0.667	0.402
85	44.611	4.36E-07	0.788	0.287
86	99.537	2.20E-16	1.009	0.099
87	52.703	1.23E-08	1.205	0.039
88	44.488	4.60E-07	0.700	0.371
89	59.376	6.18E-10	0.983	0.118
90	41.856	1.44E-06	1.392	0.017
91	43.496	7.08E-07	0.730	0.328
92	57.719	1.30E-09	1.295	0.022
93	42.192	1.25E-06	0.794	0.234
94	55.663	3.28E-09	0.946	0.145
95	50.642	3.08E-08	0.820	0.249
96	52.989	1.08E-08	0.947	0.129
97	70.307	4.27E-12	1.237	0.026
98	49.180	5.87E-08	1.010	0.117
99	126.740	2.20E-16	1.553	0.004
100	30.655	1.62E-04	0.952	0.113

Table 2. *Critical values and p-values of the Chi² and Kolmogorov-Smirnov test for random subsamples (n=1000) in the Azerbaijani Laundromat data set. Red colour formatting indicates p-values lower than the significance level (α: 0.05).*