Mathematics Standard Level Analysis and Approaches

How can Benford's Law be tested and validated for various datasets in order to detect fraud ?

Contents

Introduction

"Benford's law is an observation about the leading digits of the numbers found in real-world data sets. Intuitively, one might expect that the leading digits of these numbers would be uniformly distributed so that each of the digits from 1 to 9 is equally likely to appear."[1] Specifically, the law predicts the frequency of the leading digit using logarithms of base 10. The frequency can be seen decreasing the higher up the number gets on the 1-9 scale. Benford's Law is used many real-world applications; for instance, it is used in the detection of frauds. Benford's Law tests on natural numbers, such as payment amounts, are used by fraud investigators. According to the hypothesis, if a fraudster presents bogus invoices for payment, he will not settle for small amounts such as 100$, but rather invoices for $900 or $800. If you do this many times, the natural sequence of how numbers should occur is disrupted, and Benford's Law doesn't apply, indicating potential fraud.[2] In order to detect whether a dataset follows the predicted distribution of the leading digit, many different forms of statistical tests of best fit can be used, such as the Person's- $\chi 2$ test, and the Kolmogorov-Smirnov test of best fit.

Historical Background :

In 1881, Simon Newcomb, a Canadian-American astronomer, observed that the earlier pages of logarithm tables, which started with the number 1, were more worn than the other pages. He proposed a law that the probability of a single number being the first digit of a number was equal to the logarithm of that number plus one, minus the logarithm of that number.[3] In 1938, physicist Frank Benford also observed this phenomenon and tested it on data from various domains, including surface areas of rivers, sizes of US populations, physical constants, molecular weights, a mathematical handbook, numbers in Reader's Digest, street addresses of people listed in American Men of Science, and death rates. Benford was credited with the discovery and the phenomenon was named after him, despite Newcomb's earlier observation. The data set used by Benford included over 20,000 observations.[4]

[1] ("Benford's Law | Brilliant Math & Science Wiki," n.d.)
[2] (ACFE Insights, n.d;)

[3] (Gonsalves, 2020)
[4] (Miller, 2015)

Aim:

The aim of this investigation is to explore the validity of using Benford's Law as a tool for detecting fraud in real-world datasets, as well as man-made datasets. Specifically, this study will apply the Kolmogorov-Smirnov test of best fit and the Pearson's $\chi 2$ test of best fit to test the conformity of the first digits distribution of various datasets to Benford's Law. Moreover, an evaluation of the different test types used will be provided, to determine which test is more suitable for specific attributes present in a dataset such as sample size, as some tests might be more accurate given a larger dataset, while some may be more accurate given a smaller dataset. Ultimately, this investigation will offer an in-depth investigation of two types of tests on three different types of datasets, with varying sample sizes, while also offering an insight on the detection of fraud using this law, and its limitations. Concluding this investigation, an in-depth understanding of Benford's Law will be shown, as well as a deeper understanding about statistical tests and their limitations, and ultimately, an understanding of how it is used in fraud detection.

Variables:

First digit of the number: d

Probability of digit showing up (%): $P(d)$

Pearson's Chi-Squared Coefficient **:** $\chi 2$

Observed Distribution : O_d

Expected Distribution : E_d

Null Hypothesis : H_0

Alternative Hypothesis : H_1

Maximum distance between two cumulative frequency curves for KS-Test : D+

Degree of Statistical Significance : P(α)

Explanation of Benford's Law:

Benford's Law states that the first digit of a sequence of random numbers with random ranges, will not be randomly distributed, but rather a fixed pattern every time known as. Benford's Law can be represented by this equation, by which a set of numbers is set to satisfy Benford's Law if the leading digit d(d∈{1,2,3,4,5,6,7,8,9}).

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\frac{d+1}{d} = \log_{10}(1+\frac{1}{d})$$ [5]

By plugging in each number from 1-9, and multiplying it by 100 to get a percentage, we can see the distribution of probability.

Table 1- Benford's Law Benchmark Percentages

Leading digit(d)	P(d)
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

[5] (Miller, 2015)

When trying to understand why Benford's Law holds in the real world I came across a very interesting analogy on a mathematical online forum, which stated *"As a rough/somewhat-intuitive explanation of why Benford's Law makes sense, consider it with respect to amounts of money. The amount of time(/effort/work) needed to get from $1000 to $2000 (100% increase) is a lot greater than the amount of time needed to get from $8000 to $9000 (12.5% increase)--increasing money is usually done in proportion to the money one has. Thought about in the other direction, it should take a fixed amount of time to, say, double one's money, so going from $1000 to $2000 takes as long as from $2000 to $4000 and $4000 to $8000, so the leading digit spends more time at lower values than at higher values. Because the value growth is exponential, the time spent at each leading digit is roughly logarithmic."*[6] This analogy was very useful in the quest of my understanding of the phenomenon and did actually make sense for many of the datasets I chose to experiment with for this assessment, such as the distance to the nearest stars from Earth.

[6] (Mathematics Stack Exchange, n.d.)

Plan:

Considering I have multiple data collection methods, I will first begin by randomly generating a of numbers using Java, a known programming language, by coding a random number generator with a range simulating infinity. Another source of data I chose will be the distance to the brightest stars in the universe from planet Earth in light years, and the last source of data will be the total Covid-cases by country in 2022. Considering, I will be working with a lot of decimal points, and significant figures, I'll try to keep them as constant as possible, at 3 figures after the decimal, as most critical value tables work with 3 figures after the decimal.

For each data source I will:

1. All the datasets will encompass the same H_0(null hypothesis) and the same alternative hypothesis $H_{1,}$ that will be declared below.

For the Pearson's $\chi 2$ test, I will,

2. Explain the significance of significance levels, and degrees of freedoms
3. Create a frequency bar graph for each digit, showing how many times they occur as a percentage on the *y-axis*, and showing the digit on the *x-axis*, for that I will use Excel.
4. Compare the graphed values to the expected Benford's Law values, which will allow for an easier visualization of the data.
5. Create a table that shows sample calculations for the $\chi 2$ value, if a sample calculation hasn't been done before
6. Create a table filled with the observed frequency of the leading digit, expected frequency, as well as the calculated $\chi 2$ value
7. Display clearly in a table the $\chi 2$ value calculated, the significance level, the Degrees of Freedom and the critical value
8. Display using a relation symbol (<, > or =) whether the $\chi 2$ value is higher, below, or equal to the critical value.
9. State whether or not the null hypothesis is validated or rejected and by which percentage we can be sure of it, in accordance with the previously stated significance level
10. Explain why would the test accept or reject the null hypothesis.

For the Kolmogorov-Smirnov test, I will :

11. Establish the theory behind the Kolmogorov-Smirnov test, and why it might be useful
12. Create a clear and organized table of the expected and observed cumulative frequency(CMF) for the sources of data picked(Sample 3, and Sample 1)
13. Clearly express to which part of the formulae do the values collected correspond
14. Create clear sample calculations for the cumulative frequency, as well as the distance D between the two cumulative frequency graphs.
15. Create a graph of the two curves to aid visualization, as well as the understanding behind the formula used.
16. Create a clear and ordered table to show whether or not the null hypothesis is accepted for all significance level calculated using formulas.
17. Display using a relation symbol (<, > or =) whether the distance between the two curves(D) value is higher, below, or equal to the critical value for the chosen significance level.
18. Provide insight and an explanation to why does was the hypothesis accepted or rejected
19. Discuss about how Benford's Law is used for fraud, and its limitations, and how our statistical analysis performed throughout is related to the detection of fraud
20. Conclusion
21. Provide an evaluation and insight on the validity of the chosen data

Data Collection Description:

As previously stated, in order to gain multiple perspectives, I have sourced data from various avenues. The first source of data was derived from a Java program that generated an array of numerical values.

The second source of data for this investigative analysis was obtained from the website atlasoftheuniverse.com, which presents information on the 300 brightest stars in the universe as listed in the Hipparcos catalogue, a compendium from the European Space Agency (ESA). While the catalogue encompasses a multitude of interesting details about stars, the focus of this analysis will be solely on the distance of each star from Earth in light years. The final source of data will be the total number of Covid cases by country in the year 2022, obtained from WorldOMeter.com, a highly credible source for current world data.

Hypothesis

In this hypothesis test, the goal is to determine if the distribution of first digits in the dataset conforms to the pattern described by Benford's law.

$$H_0 = The\ leading\ digit's\ distribution\ conforms\ to\ Benford's\ Law$$
$$H_1 = The\ leading\ digit'\ distribution\ is\ different\ from\ Benford's\ Law$$

H_0 is the null hypothesis

H_1 is the other hypothesis

Pearson's χ2 test

Pearson's chi-squared test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution, which in our case would be the calculated Benford's distribution compared to a certain sample.[7]

The χ2 statistic is defined as follows:

$$\chi 2 = \sum_{d=1}^{9} \frac{(O_{d-E_d})^2}{E_d}$$ [8]

Where O_d is the observed distribution, and E_d is the expected distribution, and d are is first significant digit distribution(FSD).

Once the $\chi 2$ has been calculated we compare it against the critical value. If the calculated value is greater than the critical value, we can reject the null hypothesis up to a specific certainty percentage depending on the significant value chosen. On the other, hand if the calculated value is smaller than the critical value, we can accept the null hypothesis up to a certain degree depending on the significance value chosen.

[7] (cheesinglee, 2015)

[8] (Ibid)

Critical Value and the degree of Freedom :

P(Significance level)		0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005
Critical Value(DF8)		1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.90	21.955

Table 2 - Critical Value in relation with the degree of freedom[9]

This table was extracted from the *The International Statistical Review* (ISR) which is the flagship journal of the *International Statistical Institute*, published by the oldest scientific association.[10] This table shows the different critical values of the $\chi 2$ distribution, in relation with the degree of freedom (DF), and the statistical significance value(P).

The assessment of a hypothesis test's type I error rate is done via the **significance level**, also called alpha, which acts as a measurement of statistical confidence that the null hypothesis is rejected when it's actually true. This is usually established at a low value, such as 0.05, to avoid the chance of a false positive. Essentially, the significance level embodies the likelihood of a type I error occurring if the null hypothesis is indeed true. In other words, the degree of assurance that the test statistic won't exceed the critical value, given that the null hypothesis is correct, is demonstrated by the significance level. A lower significance level indicates stronger evidence required to reject the null hypothesis, whereas a higher value connotes a more lenient threshold for null hypothesis rejection and, as a result, an elevated risk of a type I error.[11] [12] [13]

In statistical hypothesis testing, the test statistic's critical values are determined using the **degrees of freedom**. These critical values are then compared to the test statistic to decide if the null hypothesis is accepted or rejected. The number of categories minus one determines the degrees of freedom in the chi-squared test. The degrees of freedom directly impact the critical values and accuracy of the test

---

[9] (Plackett, 1983)

[10] (Ibid)
[11] (Koziol and Perlman, 1978)
[12] (Zach, 2020)
[13] (Sullivan, n.d.)

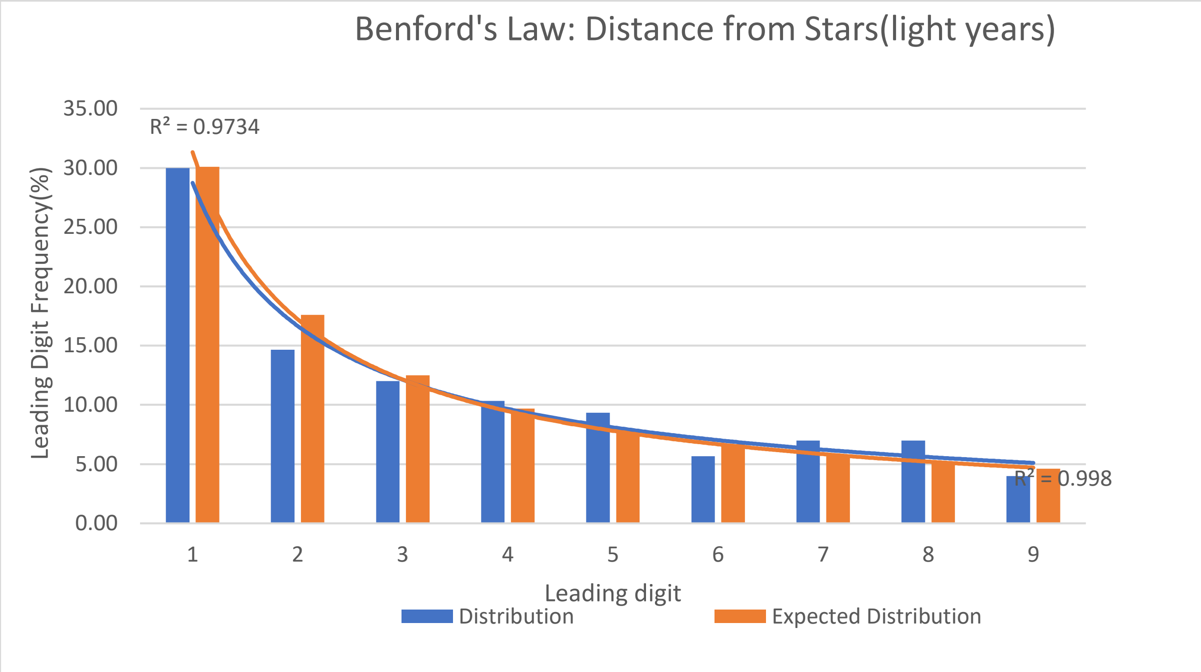
statistic. As the degrees of freedom increase, the critical values become more flexible, allowing for more variability in the test statistic.[14] [15]

Graphed Data and Statistical Analysis

The graphical representation of the comparison between the calculated and expected distributions of the leading digits in the dataset is presented in the form of a bar graph. The blue bar graph illustrates the actual distribution of the leading digits as derived from the data, while the orange bar graph displays the anticipated distribution as prescribed by the principles of Benford's Law. These graphical depictions serve as a visual aid in determining the conformity of the data's leading digit distribution to the expected pattern outlined by Benford's Law. To create those graphs, I used *Excel*[16], a highly creditable spreadsheet software developed by *Microsoft.*

[14] (Fahim, n.d.)
[15] (Biology LibreTexts, 2019)
[16] (Microsoft, 2022)

Stars (Sample Size: 300):



Results

Leading Digit	Observed	Expected	χ2 Component
1	90	90.3	0.001
2	44	52.8	1.467
3	36	37.5	0.060
4	31	29.1	0.124
5	28	23.7	0.780
6	17	20.1	0.478
7	21	17.4	0.745
8	21	15.3	2.124
9	12	13.8	0.235
Total	300	300	6.132

Sample Calculation of the $\chi 2$ for "Benford's Law: Distance from Stars (light years)" :

$$O_1, \dots \dots . O_9 = [90,44,36,31,28,17,21,21,12]$$

$$E_1 \dots \dots E_9 = [90.3, 52.8, 37.5, 29.1, 23.7, 20.1, 17.4, 15.3, 13.8]$$

The expected is calculated by using Benford's Law's benchmark percentages (30.10% for digit 1, 17.60% for digit 2 etc…(See Table 1)

For example, for this dataset the expected value for the leading digit 1 as such :

$$\frac{30.1 * 300}{100} = 90.3$$

The $\chi 2$ was calculated as such using the formulae :

$$\chi 2 = \frac{(90-90.3)^2}{90.3} + \frac{(44-52.8)^2}{52.8} + \frac{(36-37.5)^2}{37.5} \dots \dots + \frac{(12-13.8)^2}{13.8} = 6.132$$

<u>Summary of Data Analysis</u>

Significance Level (P) = 0.01

Degrees of Freedom = 8

Critical Value: 20.90

$\chi 2$: 6.132

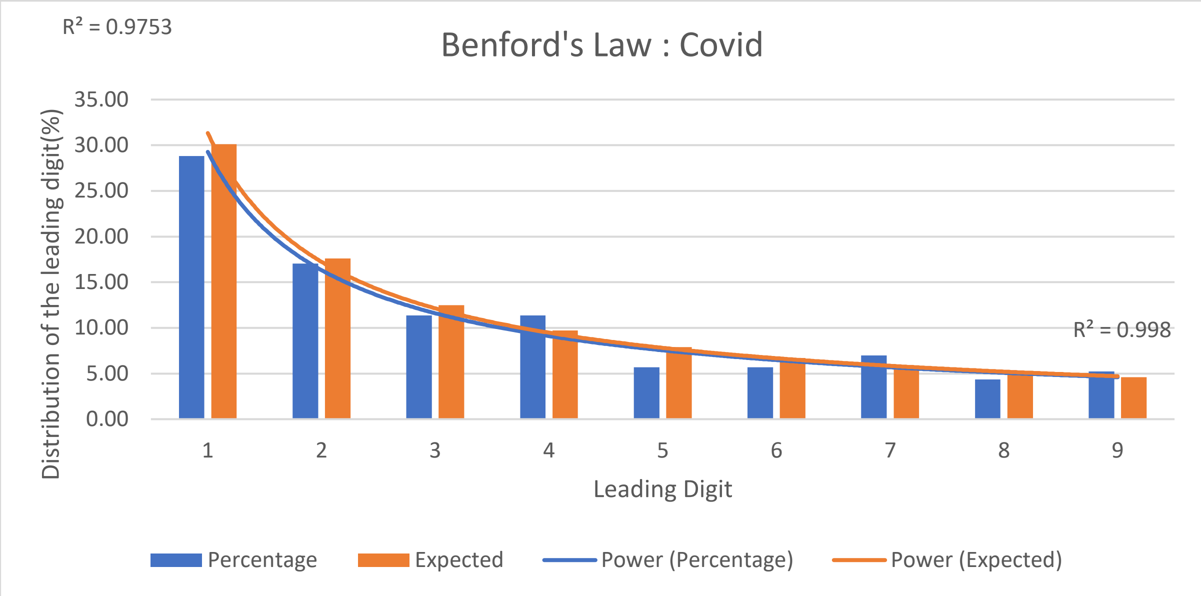
If the value of $\chi 2$ is greater than this value, then we can reject a fit to Benford's Law with a 99% certainty. Using the data collected from the 300 brightest stars near Earth, I calculated the $\chi 2$ to be 6.132, while the critical value is 20.90.

It can be established that,

$$\chi 2 < \mathit{Critical\ Value}$$

And thus, we can accept the null Hypothesis (H_0) with a 99% certainty.

Covid Cases per (Sample Size: 229):



Results

Leading Digit	Observed	Expected	χ2 Component
1	66	68.929	0.125
2	39	40.304	0.042
3	26	28.625	0.241
4	26	22.213	0.646
5	13	18.091	1.433
6	21	15.343	2.086
7	16	13.282	0.556
8	10	11.679	0.241
9	12	10.534	0.204
Total	229	229	5.573

Sample Calculation of the $\chi 2$ for "Covid cases per country" :

$$O_1, \ldots \ldots O_9 = [66,39,26,26,13,21,16,10,12]$$

$$E_1 \ldots \ldots E_9 = [68.9, 40.3, 28.6, 22.2, 18.1, 15.3, 13.2, 11.7, 10.5]$$

The expected is calculated by using Benford's Law's benchmark percentages (30.1% for digit 1, 17.60% for digit 2 etc…(See Table 1)

For example, for this dataset I got the expected value for the leading digit 1 as such :

$$\frac{30.1*229}{100} = 68.9$$

The $\chi 2$ was calculated as such using the formulae :

$$\chi 2 = \frac{(66-68.9)^2}{68.9} + \frac{(39-40.3)^2}{40.3} + \frac{(26-28.6)^2}{28.6} \ldots \ldots + \frac{(12-10.5)^2}{10.5} = 5.573$$

Using the data collected for the number of covid cases in 2022 per country, I calculated the $\chi 2$ to be 5.573. Again, considering a significance of 0.01, we can see that the $\chi 2 <$ Critical Value of 20.90. Concluding that this data set is once more following Benford's Law.

Summary of Data Analysis

Significance level(P) = 0.01

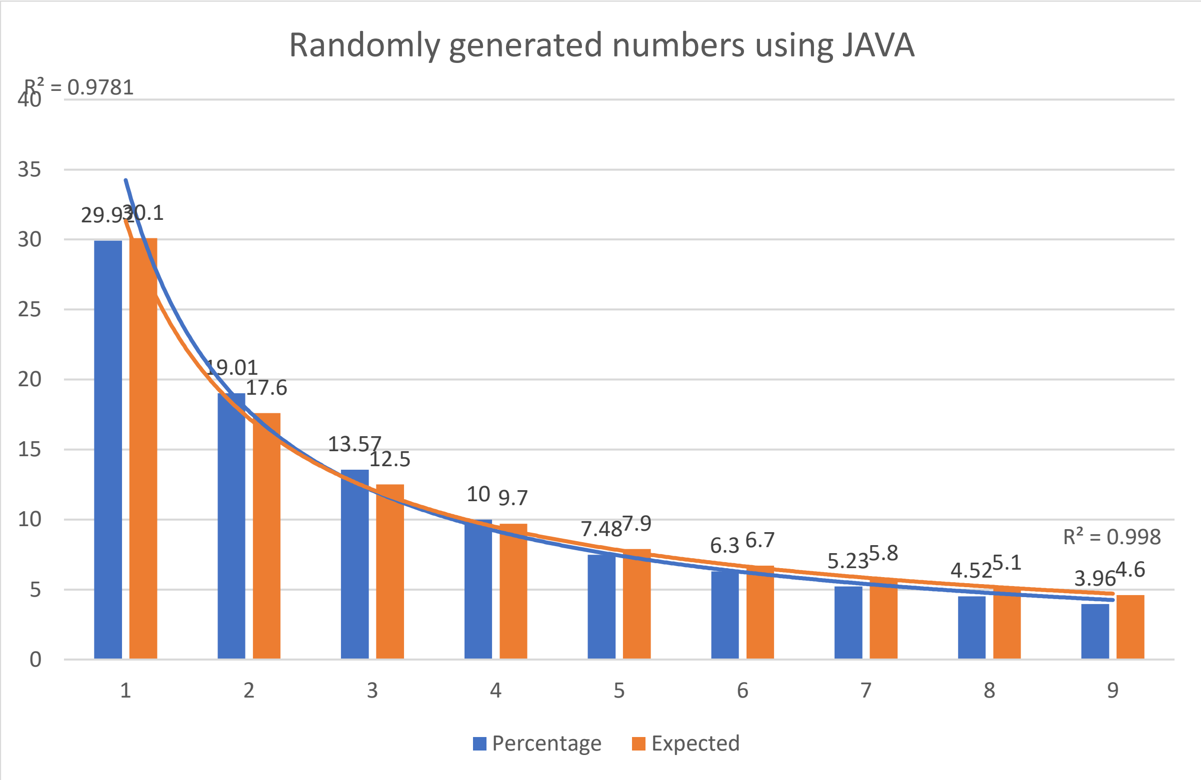
Degrees of Freedom = 8

Critical Value: 20.90

$\chi 2$: 5.573

And thus, we can accept the null Hypothesis (H_0) with a 99% certainty.

Randomly generated data via Java (Sample Size: 10 000):



Results

Leading Digit	Observed	Expected	χ2 Component
1	2997	3010	0.056
2	1869	1760	6.751
3	1302	1250	2.160
4	968	970	0.004
5	792	790	0.005
6	641	670	1.250
7	525	580	5.210
8	489	510	0.860
9	417	460	4.010
Total	10000	10000	20.330

Summary of Data Analysis
Significance Level(P) = 0.01
Degrees of Freedom = 8
Critical Value: 20.900
$\chi 2$: 20.330

By calculating the $\chi 2$ value for this dataset, I have a value of 20.3341, which again is less than the critical value. However, for this particular dataset, the $\chi 2$ value is much closer to the critical value, when taking a significance level of 0.01.

And thus, we can accept the null Hypothesis (H_0) with a 99% certainty.

Evaluation of Pearson's $\chi 2$ test

As evident from the results, which all came back as a positive fit between the observed and expected dataset, Pearson's test was effective for all sample sizes, data types etc… However, there is much more to why this test is so useful in the real-world. The Pearson's chi-squared test is a highly regarded statistical test with broad applicability in diverse fields. Its flexibility enables it to evaluate the extent to which observed data agrees with the anticipated distribution, lending greater precision to statistical analysis. Notably, the test's accessibility and user-friendliness facilitate rapid computation on low-capacity computing devices. Researchers and statisticians alike can depend on the Pearson's chi-squared test as a reliable and vital tool in their analytical toolbox..[17]

While it is true that the test provides numerous benefits, it is important to acknowledge its limitations that require careful consideration. One significant drawback is the test's remarkable sensitivity to small deviations from the expected distribution. Even deviations that are practically insignificant can cause the null hypothesis to be wrongly rejected. Furthermore, small sample sizes can significantly compromise the accuracy of the chi-square test, leading to imprecise results. Therefore, it is crucial for those using the test to assess conformity with Benford's Law to be aware of these limitations and

[17] (Cai et al., 2006)

to consider using alternative tests such as the K-S or A-D tests in conjunction with the chi-square test.[18]

A research paper titled *"The Chi-Square Test: Often Used and More Often Misinterpreted*"[19] written by *Todd Michael Franke , Timothy Ho, and Christina A. Christie*. The authors note that while these tests are widely used, they are also frequently misinterpreted, leading to statements that may have limited or no statistical support. The article focuses primarily on three tests in the *Karl Pearson* family of chi-square tests: the chi-square test of independence, the chi-square test of homogeneity, and the chi-square test of goodness-of-fit. While these tests use essentially the same formula, each one is distinct with specific hypotheses, sampling approaches, interpretations, and options following rejection of the null hypothesis. The author emphasizes that it is important to correctly interpret the results of chi-square tests and avoid overinterpreting or incorrectly interpreting the results, as evident when they state *"The vast majority of chi-square tests and misinterpretations probably exist in evaluation reports that are never read beyond a small circle of intended users, but we believe that the proliferation of chi-square test misinterpretations is exacerbated by evaluation literature that is read by a larger audience." and "The 32 articles that used chi-square tests were further reviewed to determine whether the interpretations were justified. Often, researchers were not specific about which chi-square tests were being used (only one of the 32 articles correctly specified the type of chi-square test conducted"*[20]

[18] (Ibid)
[19] (Franke, Ho and Christie, 2011)

[20] (Ibid)

Kolmogorov-Smirnov test(KS-test)

The KS test is a statistical method used to check if a sample of data comes from a specific probability distribution. It compares the empirical cumulative distribution function (CDF) of the sample with the theoretical CDF of the hypothesized distribution. The theoretical CDF is the cumulative distribution function of the hypothesized distribution.[21]

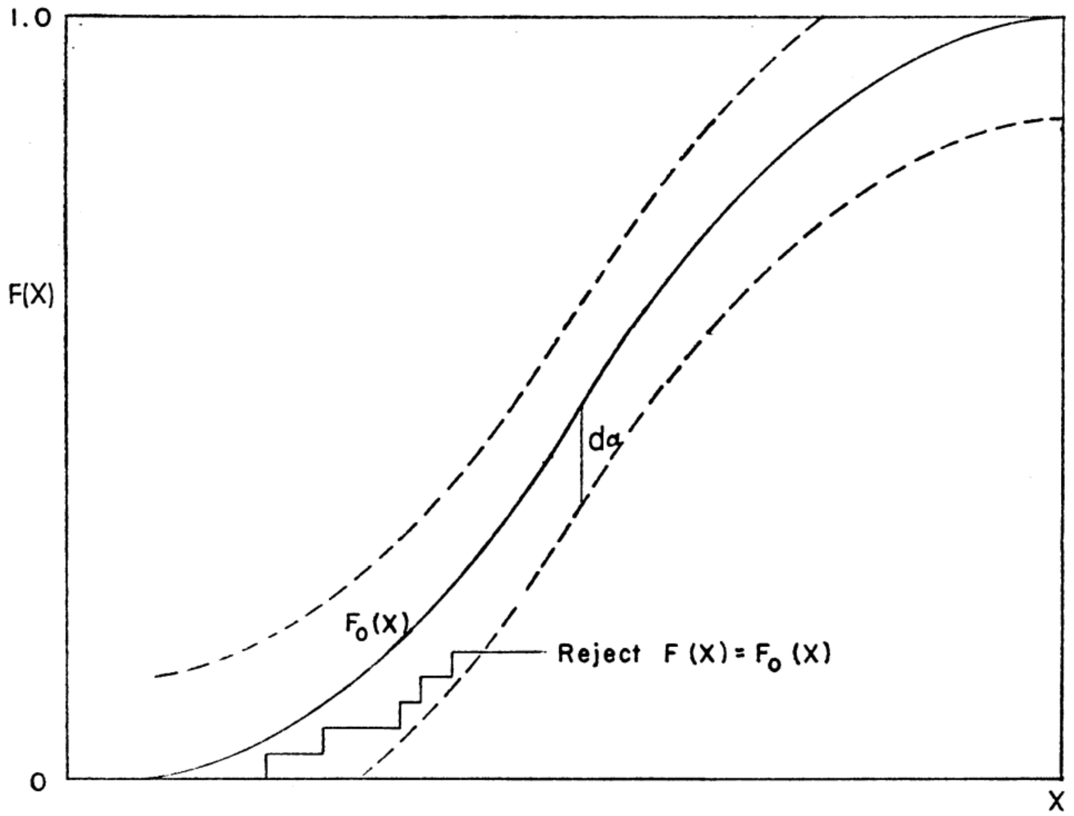


Figure 1 - Graphical Interpretation of the KS Test[22]

This graph was extracted from the American Statistical Association Journal, from a section titled *"The Kolmogorov-Smirnov Test for Goodness of Fit"[23] written by* Frank J. Massey, Jr. and published in 1951. It illustrates visually how the test is performed, with the expected distribution being the

[21] (Massey, 1951)
[22] (Ibid)
[23] (Ibid)

smooth line in the middle, and the observed distribution being the empirical curve that is uneven(F_x).

The KS test calculates a test statistic called D, which measures the maximum distance between the empirical CDF and the theoretical CDF over all possible values of the data. If the sample data closely follows the hypothesized distribution, the KS test statistic will be small. However, if the sample data significantly deviates from the hypothesized distribution, the KS test statistic will be large. The following formula expresses the mathematical calculation to obtain the greatest statistic *D*.[24]

$$D_n = sup|F_n(x) - F(x)|$$ [25]

Where D_n is the value of the greatest distance between the two curves, and $F_n(x)$ is the empirical distribution, while $F(x)$ is the expected distribution. We take the *absolute* of this subtraction, and then take the maximum of statistic D calculated, expressed by the *supremum* function enclosing the absolute value of the subtraction. For this statistical analysis, we will use the exact same set of hypothesis used for the $\chi 2$ test of best fit, which as a reminder were as follows :

$$H_0 = The\ leading\ digit's\ distribution\ conforms\ to\ Benford's\ Law$$
$$H_1 = The\ leading\ digit'distribution\ is\ different\ from\ Benford's\ Law$$

[24] (Ibid, 18)
[25] (Ibid, 18)

Kolmogorov-Smirnov Test on Sample 3(Random numbers)

CMF(Observed, 3SF)	CMF(Expected, 3SF)
0.299	0.301
0.489	0.477
0.625	0.602
0.725	0.699
0.800	0.778
0.863	0.845
0.915	0.903
0.960	0.954
1.000	1.000

Table 3 - Cumulative Frequency for Sample 3(Randomly generated number

Sample calculation for the cumulative frequency:

Relative frequency is calculated as such:

$\frac{3010(Frequency)}{9999(Sample\ Size)} = 0.30$

Then, cumulative frequency is done by adding all of the relative frequencies prior, so the first cumulative frequency stays the same, however, with the second one, you add the relative frequency from prior, to its relative frequency. So for example, the second cumulative frequency, with relative frequency 0.18, will be:

$$0.30 + 0.18 = 0.48$$

Finally, the statistic *D can be calculated as such(Data taken from table 3, above):*

$$D = Cumulative\ frquency\ from\ observed - Cumulative\ frquency\ from\ expected$$

$$D = |0.299 - 0.301| = 0.0018$$

$$D = |0.489 - 0.477| = 0.012$$

$$D = |0.625 - 0.602| = 0.023$$

$$D = |0.725 - 0.699| = 0.0261$$

$$D = |0.800 - 0.778| = 0.0219$$

$$D = |0.863 - 0.845| = 0.018$$

$$D = |0.915 - 0.903| = 0.012$$

$$D = |0.960 - 0.954| = 0.006$$

$$D = |1.000 - 1.000| = 0.000$$

From this list of D values, we must pick the largest(D^+), which in our case is:

$$D^+ = 0.026$$

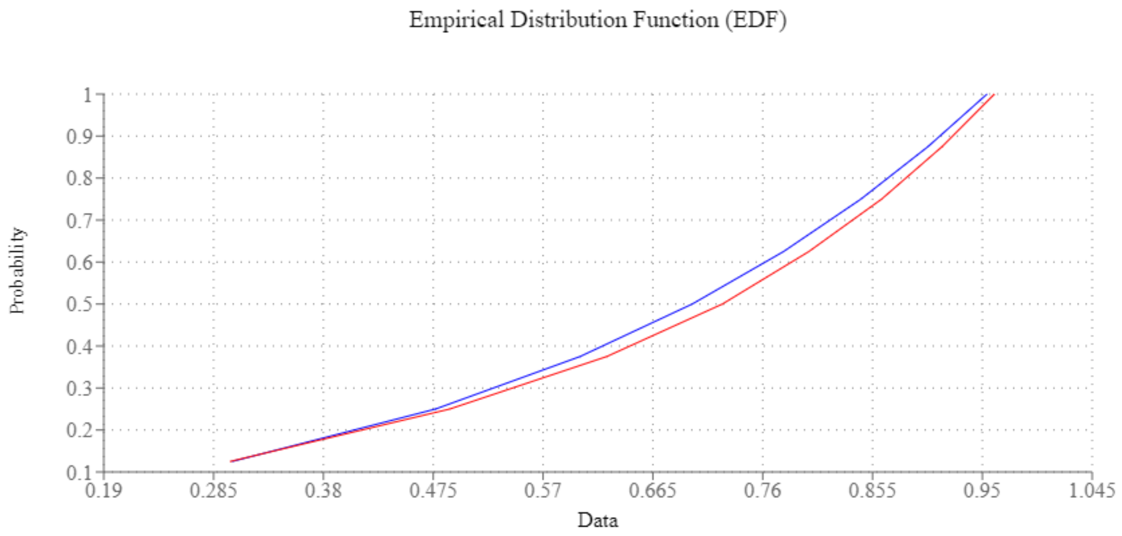


Table 4 - EDF for sample 3's cumulative frequency

This is the graph, that plots our two functions. From our D result, we can conclude that the maximum distance between both these functions can be expressed as **0.026.**

According to the section "Statistica Applicata" of the Italian Journal of Applied Statistics published in 2009, in which they mathematically prove through matrices how we calculate the critical values for the KS test, the critical value table for the test stops when the sample size equates 35. Considering our sample size is extremely large, we need to use the table provided, which I have included below as Figure 2.

Smirnov (1948)

	Significance level (α)				
n	**0 .01**	**0 .05**	**0 .10**	**0 .15**	**0 .20**
> 35	$1.63/\sqrt{n}$	$1.36/\sqrt{n}$	$1.22/\sqrt{n}$	$1.14/\sqrt{n}$	$1.07/\sqrt{n}$

Figure 2- Critical Values for n > 35

Considering our sample size(n =9999), I constructed a table below which will facilitate the comprehension of the significance levels and the calculation of the statistic *D.*

Significance value	0.01 α	0.05 α	0.1 α	0.15 α	0.2 α
Formulae	$1.63/\sqrt{N}$	$1.36/\sqrt{N}$	$1.22/\sqrt{N}$	$1.14/\sqrt{N}$	$1.07/\sqrt{N}$
Calculation	$1.63/\sqrt{9999}$	$1.36/\sqrt{9999}$	$1.22/\sqrt{9999}$	$1.14/\sqrt{9999}$	$1.07/\sqrt{9999}$
Critical Value	*0.016*	*0.013*	*0.012*	*0.011*	*0.011*
Highest D :	*0.026*	*0.026*	*0.026*	*0.026*	*0.026*
Validation of the test (Yes: D<CV)	*No*	*No*	*No*	*No*	*No*
H_0	*Rejected*	*Rejected*	*Rejected*	*Rejected*	*Rejected*

Evidently, the largest D value is always larger than the critical values at every level of significance, and we can determine that:

$$\forall \alpha, D > Critical\ value$$

And thus reject the null hypothesis.

The significant differences between the actual and predicted results of the Kolmogorov-Smirnov (KS) test can be attributed to the large sample size (N=9999) used in the analysis. The critical value is a crucial factor in determining the statistical significance and is inversely related to the sample size. When the sample size is very large, the critical value decreases significantly, which increases the likelihood of rejecting the null hypothesis (which states that the data follows the expected distribution). This issue is particularly important when evaluating the effectiveness of the KS test, as it does not account for degrees of freedom, unlike Pearson's chi-squared test, which does. Ignoring degrees of freedom correction in the KS test reduces its reliability when analyzing large samples. Therefore, caution is necessary when interpreting the results of the KS test in such situations. Overall, the KS test has limitations when used on very large samples, and modifications are needed to improve its accuracy in such cases.[26]

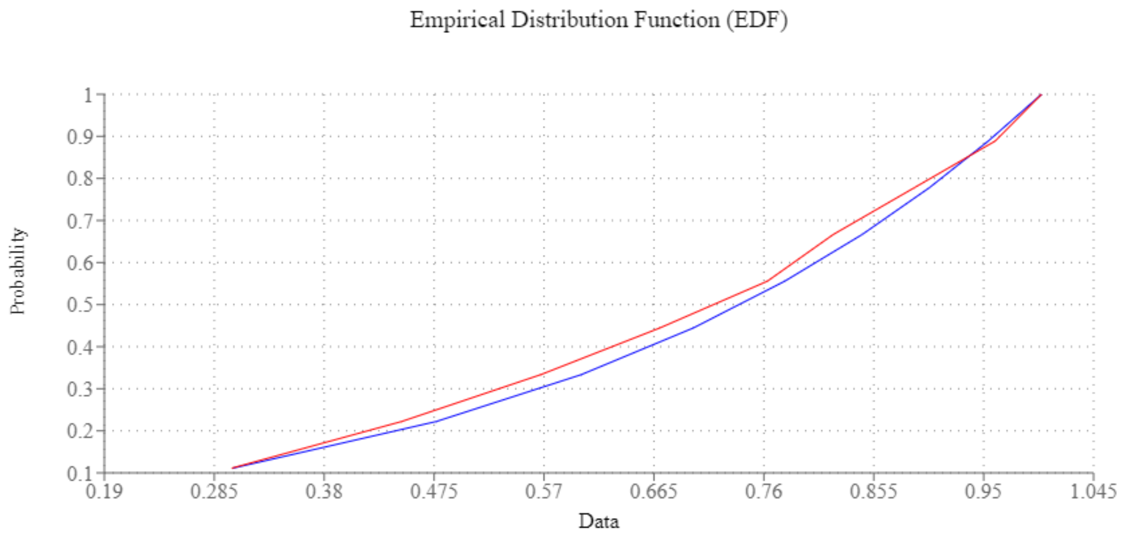
K-S Test of Best Fit on Sample 1(Stars)

Cumulative Frequency(Observed)	Cumulative Frequency(Expected)	ΔCMF
0.300	0.301	0.001
0.447	0.477	0.030
0.567	0.602	0.035
0.670	0.699	0.029
0.763	0.778	0.015
0.820	0.845	0.025
0.890	0.903	0.013
0.960	0.954	0.006
1.00	1.000	0.000

[26] (Steinskog, Tjøstheim and Kvamstø, 2007)

From this list of D values, we must pick the largest(D^+), which in our case is:

$$D^+ = 0.035$$



This is the graph, that plots our two functions. From our D result, we can conclude that the maximum distance between both these functions can be expressed as **0.035.**

Table 5- EDF graph for sample 1

Significance value	0.01 α	0.05 α	0.1 α	0.15 α	0.2 α
Formulae	$1.63/\sqrt{N}$	$1.36/\sqrt{N}$	$1.22/\sqrt{N}$	$1.14/\sqrt{N}$	$1.07/\sqrt{N}$
Calculation	$1.63/\sqrt{300}$	$1.36/\sqrt{300}$	$1.22/\sqrt{300}$	$1.14/\sqrt{300}$	$1.07/\sqrt{300}$
Critical Value	0.094	0.078	0.070	0.066	0.062
Highest D	0.035	0.035	0.035	0.035	0.035
Validation of the test (Yes: D<CV)	Yes	Yes	Yes	Yes	Yes
H_0	Validated	Validated	Validated	Validated	Validated

We can thus conclude that :

$$\forall \alpha, D < Critical\ value$$

<u>And thus accept the null hypothesis at all significance levels.</u>

The result of the Kolmogorov-Smirnov test of goodness-of-fit indicates that the observed data aligns with the expected distribution, as demonstrated by the statistic *D* being less than the critical value calculated at all significance levels. This relationship between the critical value and the maximum discrepancy serves to validate the null hypothesis, that the data follows the expected distribution, and suggests a positive fit of the sample to the reference distribution.

Evaluation of the KS Test

Advantages :

a) No reliance on underlying cumulative distribution : The fact that the K-S test statistic does not depend on the underlying cumulative distribution means that the K-S test can be used to compare the distribution of first digits in any set of numbers with the expected distribution according to Benford's Law, regardless of the underlying probability distribution of the data. This is an advantage of the K-S test because Benford's Law is based on the distribution of first digits in numbers, rather than the distribution of the numbers themselves, and it is often difficult to determine the underlying distribution of the data.[27]

b) The K-S test is sensitive to differences between the observed data and the expected distribution, which is important for detecting potential anomalies that may indicate fraud or errors.[28]

Disadvantages :

a) Only applies to continuous distribution. If the data being tested for conformity to Benford's Law is not continuous, then the K-S test cannot be used. This means that the test can only be used for data that can take any real value within a certain range. Continuous distributions have an infinite number of possible values between any two points and can be described using probability density functions. In contrast, discrete distributions have a finite or countable number of possible values and can be described using probability mass functions. If the data being tested for conformity to Benford's Law is not continuous, then the K-S test cannot be used.[29]

b) Another undeniable known drawback from the KS Test is that it tends to be more sensitive to differences in the center of the distribution than at the tails. That is simply because the variance of the sample CDF in the tails is smaller than near the median. On the forum *"Statistics Stack Exchange"*[30], I found the following statement *"Specifically, the variance is proportional to F(1−F), so as F approaches 0 or 1 the variance goes to zero, and so does the standard error of the ecdf. This means that even small deviations in the tail can be a strong*

[27] (asaip.psu.edu, n.d.)
[28] (Ibid)
[29] (Gupta et al., 2019)
[30] (Cross Validated, n.d.)

indication of a problem with the null distribution."[31]. This very interesting forums explain why the test may be more sensitive towards the center. This statement refers to the fact that for a sample CDF, the variance of the distribution is proportional to the product of the empirical probability of an event and its complement. In other words, if we have a sample of n data points and we consider an event A that occurs k times in the sample, then the empirical probability of A is $\frac{k}{n}$, and the empirical probability of its complement is $\frac{n-k}{n}$. The variance of the empirical probability of A is proportional to $\frac{k}{n} \times \frac{n-k}{n}$, which simplifies to $\frac{k(n-k)}{n^2}$, and this is proportional to $F(1 - F)$ where $F = \frac{k}{n}$ is the empirical probability of A.

As the empirical probability F approaches 0 or 1, the variance approaches zero as well, which means that the standard error of the empirical CDF also approaches zero. This makes sense intuitively, as the CDF approaches a step function when all the data points are the same, and the variability of the empirical distribution function becomes smaller as the number of data points in the tails becomes smaller.

[31] (Ibid, 29)

The user on the forum further states *"In particular, try looking at the uniform case (since any other case can be reduced to the uniform by transforming the data by the hypothesized cdf under the null). Here's 500 uniform ecdfs with n=20"*.[32]

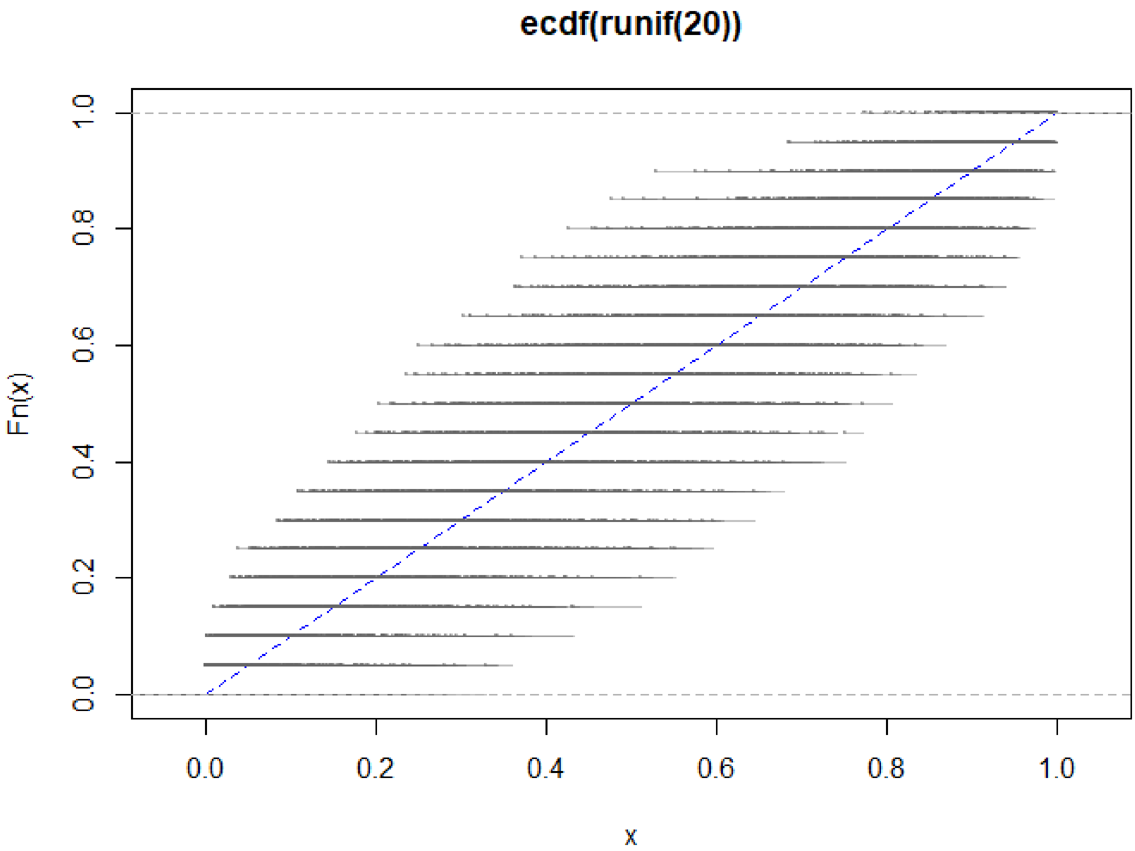


Figure 3- 500 uniform ecdfs with n=20[33]

[32] (Ibid, 26)
[33] (Ibid, 26)

When examining the distribution, it can be observed that the spread of the data is wider at x=0.5 compared to x=0.05 and x=0.95. As a result, it is more difficult to achieve the critical value of D in the tail region than in the middle, which means that the test is more likely to detect deviations in the middle of the distribution than at the extreme ends.

Ultimately, the K-S test has several refinements such as the Anderson-Darling and Cramer Von-Mises tests, which are generally considered more powerful. The K-S test's advantage of having critical values independent of the underlying distribution is not significant due to the typical scenario where the distribution parameters are not known and must be estimated from the data. Therefore, critical values for the K-S test must be determined by simulation, just like the Anderson-Darling and Cramer Von-Mises tests. Although, the KS test has many drawbacks, making it an inefficient test, as also seen in the results of my experiment whereby the observed dataset didn't diverge much from the expected, but was still rejected, is might still be an effective test for specific datasets.

Benford's Law usage in the detection of fraud :

Benford's law serves as a useful tool in identifying fraudulent activities. Fraudsters tend to overlook the principles of Benford's law while preparing false transaction documents. By detecting manual intervention in otherwise automated transaction activity, Benford's law can uncover fraudulent numbers in random datasets. This is particularly useful in identifying data that have been manipulated for the purpose of tax evasion.

Election Results:

Benford's Law has already been used in many real-life cases, in order to detect potential fraudulent activity. One of them being during the 2009 Iranian election, when an announcement was made on June 12th that Mahmoud Ahmadinejad, the incumbent president of Iran, had emerged victorious in the election, defeating his primary contender, Mir-Hossein Mousavi. This declaration triggered widespread protests in Iran, as people disputed the validity of the outcome. On June 14th, the Iranian Ministry of the Interior disclosed the election results for 366 polling stations, indicating that

Mahmoud Ahmadinejad had received over 24 million votes, while Mir-Hossein Mousavi had secured approximately 13 million votes.

However, cosmologist Boudewijn Roukema, from the Nicolas Copernicus University in Poland, has used this law to test the results from the recent Iranian election. detected an unusual pattern in the votes for Mehdi Karroubi from the National Trust Party who came third in the election. He found that the digit "7" appeared as the first digit more frequently than what Benford's law predicts. This pattern was observed in three out of the six largest voting areas. Furthermore, he found that Mahmoud Ahmadinejad had a higher percentage of votes in these three areas compared to the other areas. Based on this analysis, Roukema hypothesized that there might be an error in the official count of approximately one million votes.

Financial Fraud

Mark Nigrini was one of the earliest proponents of using Benford's law for detecting fraud in various types of data sets, whether financial or electoral. The underlying assumption is that if the distribution of leading digits in a given data set deviates significantly from what is expected under Benford's law, it could indicate fraud or manipulation. Over the last 20 years, Nigrini has identified and examined numerous potential cases where this principle held true, including the well-known Enron accounting scandal. Additionally, Benford's law was used in a study conducted in 2011, which revealed an "abnormal" distribution of numbers in Greece's economic reports to the European Union for several years, apparently confirming the European Commission's independent accusations of data manipulation.

Limitations of Benford’s Law in the detection of fraud:

Many academics argue that Benford’s Law is not a way of proving that fraud occurred, but rather raises red flags on unusual data patterns. *Theodore P. Hill,* a retired math professor from Georgia Tech, cautioned that the use of Benford's Law, regardless of the distribution uncovered, cannot definitively prove fraud. It is only a preliminary test that can raise suspicions. The *IRS* has been using it for years to detect potential fraud, but the auditors must investigate further to find actual evidence. *Walter Mebane*, a professor at the *University of Michigan*, wrote an article in 2006 about using Benford's Law to detect election fraud. While he believed the test was valuable, he cautioned that it alone cannot prove election fraud and that there are other factors to consider.

Misinterpretation of Benford’s Law in the detection of fraud:

While Benford’s Law is a fascinating mathematical statistical concept, it is often misinterpreted. A perfect example of this aforementioned misinterpretation was evident during the 2020 United States Elections, whereby Joe Biden, won the election against Donald Trump, a prominent political figure. Indeed, people on social media were sharing content claiming that Benford's Law, is evidence of fraudulent activities during the U.S. presidential election. However, according to various research papers and experts, deviations from Benford's Law do not necessarily indicate election fraud.

Some online posts show graphs comparing the leading digits of the vote counts for each candidate to the expected distribution according to Benford's Law. The posts claim that this test shows that Biden's vote tallies do not follow the expected distribution, while Trump's do. The captions on the posts assert that Benford's Law is a test that has been used to detect fraud before, and suggest that Biden's alleged violation of this law is evidence of fraud. Some captions imply that Biden cheated and that his victory is mathematically impossible.

Walter Mebane, a professor at the University of Michigan's departments of Political Science and Statistics, wrote an article in December 2006 that examined the use of Benford's Law in the context of US presidential election results. While the article acknowledged some limitations of the test, it suggested that it could be taken seriously as a statistical tool for detecting election fraud.

However, Mebane's article also cautioned that the Benford’s Law and the 2-Benford’Law test which refers to "second-order Benford's Law," which is a variation of the original Benford's Law that takes into account the second digit in addition to the first digit of numerical data. The purpose of using 2BL is to increase the accuracy of the test in identifying deviations from the expected distribution, and to detect more sophisticated attempts to manipulate data to avoid detection, which alone cannot prove that election fraud has occurred or that an election was clean, and that some kinds of fraud are not detectable by this test.

.

In response to multiple inquiries, Mebane published a paper on November 9, 2020 titled *"Inappropriate Applications of Benford's Law Regularities to Some Data from the 2020 Presidential Election in the United States."* [34]In the paper, Mebane stated that the use of the first digits of precinct vote counts from Fulton County, GA, Allegheny County, PA, Milwaukee, WI, and Chicago, IL, did not provide any evidence of possible fraud. He explained that it is widely understood that the first digits of precinct vote counts are not useful in identifying election fraud. A 2011 study titled *"Benford's Law and the Detection of Election Fraud,"* conducted by *Joseph Deckert, Mikhail Myagkov* (a Professor of Political Science at the University of Oregon*), and Peter Ordeshook* (a Professor of Political Science at Caltech)[35], found that Benford's Law was not reliable when applied to election data. The study found that conformity with and deviations from Benford's Law follow no pattern, rendering it problematic as a forensic tool and potentially misleading.

Ultimately, the suitability of Benford's Law as an electoral fraud detection tool has been a topic of scholarly debate. However, it is widely acknowledged that the use of Benford's Law to scrutinize the leading digits of local vote tallies is challenging, and that deviations from the law alone cannot serve as conclusive proof of electoral fraud. According to experts, the limitations of the method arise due to various complex statistical factors and cannot be easily resolved through the simple application of the Benford's Law principle.

[34] (Mebane, 2020)
[35] (Deckert, Myagkov and Ordeshook, 2011)

Conclusion

Using two mathematical statistical analysis tools: the Person's $\chi 2$ test and the Kolmogorov-Smirnov test , I was able to determine whether or not the data was found to be "Benford" or in other terms, if the data followed the trend set by Benford's Law. For all three datasets using the $\chi 2$ test, it was concluded that Benford's Law did apply to them, as the $\chi 2$ value was less than the critical value with a significance of 0.01. The significance level of 0.01 means that we can be sure with a 99% certainty that the null hypothesis is true, or that we can reject the other hypothesis with a 99% certainty. However, using the KS-Test on two of the datasets, Sample 3 previously validated by the $\chi 2$ test, was rejected, while Sample 1 was validated by both tests.

The dataset that pertains to the number of COVID cases per country for the year 2022 yielded an unexpected outcome. While it was anticipated that the data would conform to Benford's Law, the $\chi 2$ value was remarkably low, given the notion that several countries have artificially elevated their figures or lack the resources to precisely record the cases. Therefore, a higher $\chi 2$ value was anticipated that would demonstrate the discordance in the data, or potentially detect fictitious or inadequately reported data originating from developing nations.

On the contrary, the dataset comprising of random numbers posed a considerable challenge to develop. During the implementation of the code for the random number generator, a predicament arose wherein all the generated random numbers displayed a leading digit distribution of approximately 11%. Subsequently, after extensive inquiry, a forum that deliberated on this specific topic was discovered. As per the forum's insights, *"Benford's Law stands for the proposition that it doesn't hold when integers are drawn uniformly from a range that, like yours, ends at a power of 10: a well-designed pseudorandom number generator should give numbers with asymptotically exactly a 1/10 chance of each leading digit 0,1...,9 in decimal notation."* In other words, Benford's Law only applies to data derived from certain real-life random processes, not accurately random data sources. Data generated by exponentially increasing processes, such as populations, conform to Benford's Law as they produce uniformly distributed logarithms of data points. However, this uniform distribution is stretched into the distinctive Benford's Law upon exponentiation. Upon further investigation, the implementation of the code for the random number generator was altered to generate a number with an enormously large range simulating infinity. In order to achieve the desired outcome, a recursive approach was employed wherein the function recursively called upon itself, thereby using its own output as an input parameter. This allowed for the generation of an extremely large range, akin to simulating infinity, through an iterative process that relied on self-referentiality. Ultimately, this experiment was a success. I was able to exactly follow the plan I set out at the beginning, as well as drawing from many very credible research papers to answer my questions, and explain unexplainable phenomenon in my data.

Validity:

For the datasets chosen, I am completely reliant on secondary sources, which may not reflect accurate real-world data. Furthermore, the Person's $\chi2$ test may not be very accurate for small sample sizes. For small sample sizes, the $\chi2$ test can encounter difficulty in discriminating between data which do and do not fit Benford's Law. The Cho-Gaines' d statistic is an alternative test which is formulated to be less sensitive to sample size and may have been a better option for the stars dataset and the covid dataset, as they had quite small sample sizes: 300 and 229 respectively.

Reference list

ACFE Insights. (n.d.). *What Is Benford's Law and Why Do Fraud Examiners Use It?* [online] Available at: https://www.acfeinsights.com/acfe-insights/what-is-benfords-law.

asaip.psu.edu. (n.d.). *Beware the Kolmogorov-Smirnov test! – Astrostatistics and Astroinformatics Portal*. [online] Available at: https://asaip.psu.edu/Articles/beware-the-kolmogorov-smirnov-test/ [Accessed 22 Feb. 2023].

Biology LibreTexts. (2019). *9.4: Probability and Chi-Square Analysis*. [online] Available at: https://bio.libretexts.org/Bookshelves/Biotechnology/Bio-OER_(CUNY)/09%3A_Genetics/9.04%3A_Probability_and_Chi-Square_Analysis.

Cai, Li., Maydeu-Olivares, Albert., Coffman, D.L. and Thissen, David. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2P tables. *British Journal of Mathematical and Statistical Psychology*, 59(1), pp.173–194. doi:https://doi.org/10.1348/000711005x66419.

cheesinglee (2015). *Detecting numeric irregularities with Benford's Law*. [online] The Official Blog of BigML.com. Available at: https://blog.bigml.com/2015/05/15/detecting-numeric-irregularities-with-benfords-law/ [Accessed 20 Feb. 2023].

Cross Validated. (n.d.). *hypothesis testing - Limitations of Kolmogorov–Smirnov test*. [online] Available at: https://stats.stackexchange.com/questions/437979/limitations-of-kolmogorov-smirnov-test.

Deckert, J., Myagkov, M. and Ordeshook, P.C. (2011). Benford's Law and the Detection of Election Fraud. *Political Analysis*, [online] 19(3), pp.245–268. doi:https://doi.org/10.1093/pan/mpr014.

Fact check: Deviation from Benford's Law does not prove election fraud. (2020). *Reuters*. [online] 10 Nov. Available at: https://www.reuters.com/article/uk-factcheck-benford-idUSKBN27Q3AI.

Fahim, M. (n.d.). *CHI-Squared Test of Independence*. [online] Available at: https://pages.cpsc.ucalgary.ca/~saul/wiki/uploads/CPSC681/topic-fahim-CHI-Square.pdf [Accessed 20 Feb. 2023].

Franke, T.M., Ho, T. and Christie, C.A. (2011). The Chi-Square Test. *American Journal of Evaluation*, 33(3), pp.448–458. doi:https://doi.org/10.1177/1098214011426594.

Gonsalves, R.A. (2020). Benford's Law — A Simple Explanation. *Medium*. [online] 6 Oct. Available at: https://towardsdatascience.com/benfords-law-a-simple-explanation-341e17abbe75.

Gupta, A., Mishra, P., Pandey, C., Singh, U., Sahu, C. and Keshri, A. (2019). Descriptive Statistics and Normality Tests for Statistical Data. *Annals of Cardiac Anaesthesia*, 22(1), p.67.

Koziol, J.A. and Perlman, M.D. (1978). Combining Independent Chi-Squared Tests. *Journal of the American Statistical Association*, 73(364), pp.753–763. doi:https://doi.org/10.1080/01621459.1978.10480095.

Massey, F.J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253), pp.68–78. doi:https://doi.org/10.1080/01621459.1951.10500769.

Mebane, W. (2020). *Inappropriate Applications of Benford's Law Regularities to Some Data from the 2020 Presidential Election in the United States **. [online] Available at: http://www-personal.umich.edu/~wmebane/inapB.pdf.

Microsoft (2022). *Microsoft Excel, Spreadsheet Software*. [online] www.microsoft.com. Available at: https://www.microsoft.com/en-us/microsoft-365/excel.

Miller, S.J. (2015). *Benford's law : theory and applications*. Princeton, New Jersey: Princeton University Press.

Plackett, R.L. (1983). Karl Pearson and the Chi-Squared Test. *International Statistical Review / Revue Internationale de Statistique*, 51(1), p.59. doi:https://doi.org/10.2307/1402731.

Statistica Applicata. (2009). *Italian Journal of Applied Statistics*, 21, pp.3–4.

Steinskog, D.J., Tjøstheim, D.B. and Kvamstø, N.G. (2007). A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality. *Monthly Weather Review*, 135(3), pp.1151–1157. doi:https://doi.org/10.1175/mwr3326.1.

Sullivan, L. (n.d.). *Hypothesis Testing - Chi Squared Test*. [online] sphweb.bumc.bu.edu. Available at: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ChiSquare/BS704_HypothesisTesting-ChiSquare_print.html.

Zach (2020). *How to Read the Chi-Square Distribution Table*. [online] Statology. Available at: https://www.statology.org/how-to-read-chi-square-distribution-table/.

Mathematics Stack Exchange. (n.d.). *statistics - Why does Benford's Law (or Zipf's Law) hold?* [online] Available at: https://math.stackexchange.com/questions/781/why-does-benfords-law-or-zipfs-law-hold [Accessed 22 Feb. 2023].

Brilliant.org. (2019). *Benford's Law | Brilliant Math & Science Wiki*. [online] Available at: https://brilliant.org/wiki/benfords-law/.