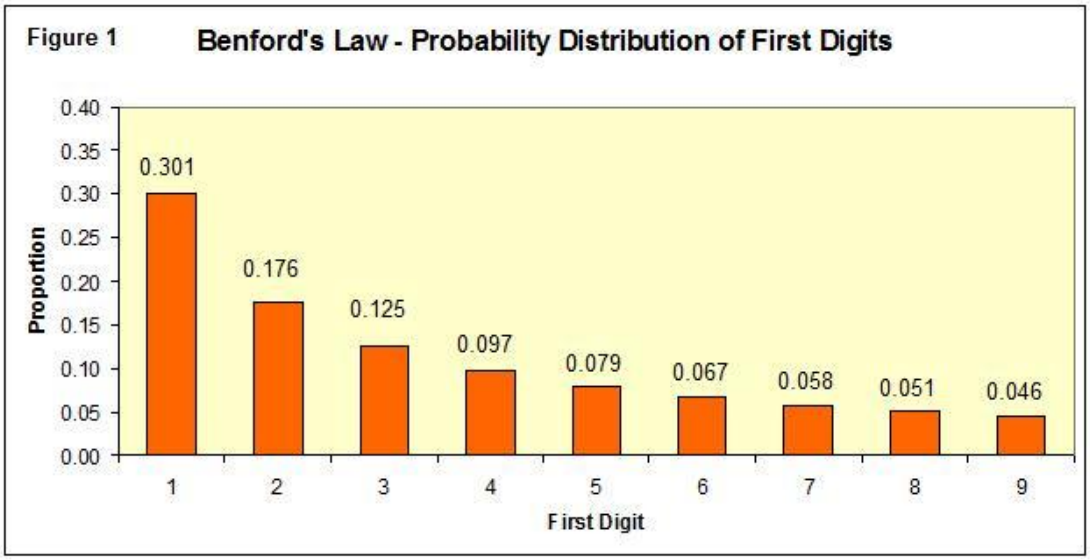
Does Benford's Law provide a good way of telling if a data set produced by a natural process?

Benford's Law offers a way of telling organically generated series from artificial ones and is a way of examining whether what we're told- such as accounting reports or government statistics- may be wrong. Discovered by Henry Newcomb in 1881, and then later re-discovered and attributed to Frank Benford in 1938, this law has achieved great fame outside of academia and analytics, even appearing in plots of TV crime dramas. Rightly or wrongly, it has been depicted as an incredibly powerful tool for exposing fraud, improving government responses to pandemics, and snuffing out voter fraud.

Benford's Law, commonly known as the Newcomb-Benford law, is a theory of the frequencies of leading digits in naturally[AM1] generated data sets. If you were to assume a uniform distribution of leading digits in a data set, the numbers 1-9 would each have an equal probability of appearing of approximately 11.1% (1/9). However, both Newcomb and Benford discovered that with naturally generated data sets, the frequency was more skewed towards smaller digits. They determined the frequency of leading digit being one would not be 11.1%, but rather 30.1%, and 2 to be 17.6%. with decreasing frequency.



(*Benford's Law and US Census Data, Part II | Introductory Statistics*, n.d.)

The distribution of leading digits in a naturally randomly generated data set follows this rule:

$$P_B(d) = \log\left(1 + \frac{1}{d}\right),$$

(*Application of Benford's Law in Data Analysis*, n.d.)

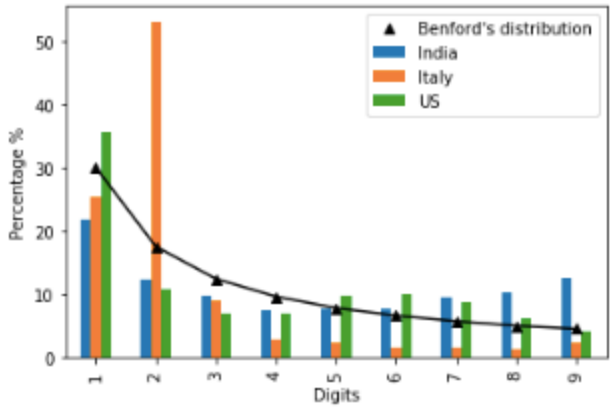
Where $P_B(d)$ is the probability of the leading digit occurring, and ‘d’ being the leading digit in question. For instance, $P_B(2)$ would be $\log_{10}(3/2)$ or 0.176 to 3 significant figures.
The more notice you take to patterns, the more frequently you find this pattern to occur. If you went through the first 25,000 numbers of the Fibonacci sequence, you would find the occurrence of the leading digit to be the same as predicated in Benford’s law.

There are many conditions for this law to apply. Firstly, there must be an equal opportunity of the digit appearing as the leading digit. To give an example, the law would not apply for human heights in adults or IQs, where there are strict parameters. The probability of the leading digit being 1 would be much higher than just 30% for human heights, and there are few people taller than 2 metres and shorter than 1 metre. The law works best when the data set spans several orders of magnitude, as the narrower the range, the less likely it is to apply. This is when data range from 10^2 to 10^6. A good example of this is populations. Across the globe it varies from greater than a billion in India, to only a few hundred thousand in Iceland. The law is not applicable in small samples either, as there are not enough data points for a definite pattern to be spotted.

Pseudo-random numbers generated by humans do not have this property, and systemic distortions in collecting data mean that the reported series may not conform to Benford's Law, although the true data does. This makes Benford’s Law such an amazing tool in tax-auditing and spotting fraud. In the past, fraudsters would cut checks or submit invoices for values just under 100,000. Therefore, there would be a disproportionately high frequency of 9,8, and 7s being the leading digit, which would alert authorities towards their wrongdoing.

The scope for applying Newcomb-Benford’s law is huge. Recently, when the Covid pandemic hit, Benford’s Law became a method to check if control intervention was working, or if COVID-19 data was underreported by governments. As Covid is an infectious disease, the number of cases grew exponentially, therefore, the data set would automatically follow Benford’s Law. After COVID lockdown restrictions were imposed, governments could assess their impact and efficacy by the impact to case numbers. Where measures were successful, the number of infections would fall below the epidemic growth curve, meaning the data was no longer naturally occurring and no longer conform to Benford’s Law.

A study conducted comparing the first digit distribution of the number of COVID cases in countries after lockdown (a study conducted by John Hopkins University Centre) compared to the predicted distribution with Benford’s Law.



(Pahuja, 2021)

The graph above shows how the US follows Benford's distribution most closely, implying the lockdown measures were not as effective in limiting the spread of the disease. However, it does suggest that the spread of the disease was reasonably accurate. In Italy, on the other hand, the distribution was affected by public health measures, so that the spread was very different than it would have been without intervention. During a time of great stress and urgency, it was vital that governments impose the most effective public health protocols, thus, Benford's Law became a useful tool to discern between the effectiveness of different public health measures.

Given that conformity can be so hugely important, it's worth investigating statistical tests more formal than eyeballing graphs. A chi-square goodness of fit test is often conducted, measuring the degree of disparity and deviation between predicted data and actual data. This test helps data analysts determine whether deviations from the predicted data are due to chance and therefore insignificant, or because there is no relationship between the data sets. To perform this test, two hypotheses are established:
H_0 : The distribution of leading digits in data generated follows Benford's Law
H_1 : The distribution of leading digits in data generated does not follow Benford's Law
The chi-square statistic is found using this formula:

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

(*What Is a Chi-Square Test? Formula, Examples & Uses | Simplilearn*, n.d.)
Where O_i is the observed frequency of the data, and E_i is the expected/predicted frequency of the data. The notation 'c' is used to represent the degrees of freedom- the number of variables that could vary within the calculation. The smaller the value, the better the observed frequencies follow the pattern.

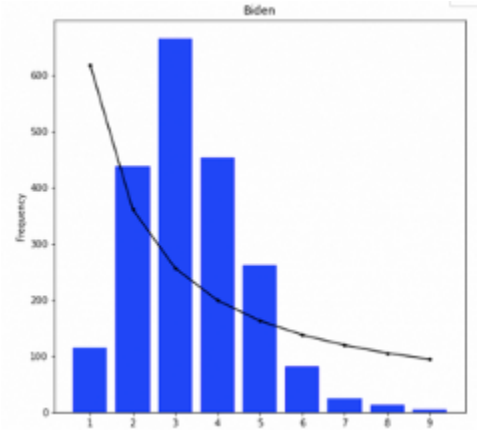
If we were to apply this test to data sets compared with Benford's distribution, 'c' would be 8 as there are 9 possible integers, therefore the probability of the ninth integer appearing would be

fixed. The observed and expected frequencies cannot be the percentage or proportion of the digit occurring in the data set, but the literal number instead. This value would then be compared to a critical value from a X^2 statistic table, with the corresponding degrees of freedom and significance level of the data. If the calculated chi-square statistic is greater than the critical value, then the null hypothesis is rejected.

A key problem when using the chi-square test to assess whether a data set follows Benford's distribution is that the two need conflicting sample sizes. Benford's Law is most evident with a large sample size, whereas a chi-square goodness of fit test is sensitive to large sample sizes. Even a small deviation from the expected result will greatly increase the size of the chi-square statistic, and the greater the size of the data, the more frequently this will occur.

Elections are an excellent potential use of a tool to detect data manipulation, but this can serve as a really good illustration of a reason to be cautious, because election data does not necessarily meet criteria. In principle stuffed ballots would deviate from this otherwise naturally occurring data set. During elections however, the country is divided into regions, which are then divided into smaller scale, similarly sized precincts- which have neighbourhood level size population. The law would not hold as precincts are similar size, hence would observe similar distribution, leading to fluctuations around digits that are not necessarily the smallest (1 or 2). Hence across this type of data, Benford's Law would not be expected to hold.

If we were to compare leading digit distribution for precincts in Chicago votes for Biden in the 2020 election against Benford's law the data would look skewered and tampered.



(*Fraud in the 2020 US Election?!?! – Katie Howgate*, n.d.)
However, the maximum voter turnout was no greater than just over 1000 for each precinct, the data does not span several orders of magnitude. Due to its nature of only being a statistical observation, Benford's Law could, when used as an election fraud technique, have a 50/50 success rate. Overall, the law itself does not prove election fraud but provides a good indicator that more investigation should be taken into the matter.

Over the course of writing this essay, I have repeatedly been amazed at the phenomenon that is Benford's Law. It is so fascinating how there are patterns in everything, and how even the most

random disruptions to them can be explained by mathematics. Newcomb-Benford's law is just an example of how probability and statistics govern the world we live in. Distinguishing accurate reporting of certain patterns from distorted reporting is key to many disciplines - from detecting fraud to fighting pandemics. Despite the applicability of the law, observations and decisions made solely on the law should be taken with a pinch of salt, correlation does not imply causation.

Application of Benford's law in Data Analysis. (n.d.). https://doi.org/10.1088/1742-6596/1168/3/032133

Benford's Law and US Census Data, Part II | Introductory Statistics. (n.d.). Retrieved March 31, 2024, from https://introductorystats.wordpress.com/2011/11/25/benfords-law-and-us-census-data-part-ii/

Fraud in the 2020 US Election?!?! – Katie Howgate. (n.d.). Retrieved March 29, 2024, from https://www.lancaster.ac.uk/stor-i-student-sites/katie-howgate/2021/03/12/fraud-in-the-2020-us-election/

Pahuja, D. (2021). Application of Benford's Law to Detect if COVID-19 Data is under Reported or Manipulated. In *New Frontiers in Communication and Intelligent Systems* (pp. 85–91). Soft Computing Research Society. https://doi.org/10.52458/978-81-95502-00-4-11

What is a Chi-Square Test? Formula, Examples & Uses | Simplilearn. (n.d.). Retrieved March 29, 2024, from https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test

Other sources:
https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/using-chi-square-statistic-in-research/
https://en.wikipedia.org/wiki/Benford%27s_law