Nature's Digit Distribution

1) Introduction: The Anomaly

Hello to anyone who is reading this, I am deeply grateful for your presence.

Let me pose you a question. Say if there is a random number generator that generates all real numbers greater than 0. What do you think is the probability or frequency that the **first digit** of a number produced by it would be the number "1"? If you think it is 1/9, you would be correct since there are only nine possible outcomes (1,2,3 … 9) and their probabilities should be evenly distributed. I mean after all it is called a **RANDOM** number generator for a reason. Now, what about the first digit of each country's population? Would it be similar? Well, it's not. It is far off at around 30%. On the other end, the number "9" was the first digit for about 5% of the population only! In fact, have a look at this table of all the digits.

First Digit	Count (Approx)	Percentage
1	69	**30.0%**
2	41	**17.8%**
3	29	**12.6%**
4	22	**9.6%**
5	18	**7.8%**
6	15	**6.5%**
7	13	**5.7%**
8	12	**5.2%**
9	11	**4.8%**

Now that is peculiar, isn't it? And this is not just limited to countries' population. You can analyse nations' GDP or river lengths across the world, and you can expect to find similar patterns. What you have just observe, is the strange and counter-intuitive yet gorgeous phenomenon of the Benford's law. This law denotes that real-life data are more likely to have a small first digit, and it is scale invariant (which its definition I will get to in the next section).

In 1938, Frank Benford (law named after him), a research physicist working for General Electric first discover it when he found that a logbook, which is a book which contain logs' values (logarithm, not the wood), was more worn out on the pages that contain the result of 1 as the starting digit (e.g. log(1)). He then gathered 20000 real-life dataset that he could find and found that almost all of them followed a similar pattern like the countries' population. [1] However, Frank was not the first that made a discovery. Simon Newcomb, astronomer and mathematician also observed the logbooks were worn out on the first few pages and discover this statistical principle in 1881. [2] Benford just did more tests and popularised it. That is why it is also called Newcomb-Benford Law. Alternate names include first-digit law and law of anomalous numbers (the title of the article Frank stated this law in) as well. [3]

Now, after a bit of history, let's get into the math behind this law.

2) Benford's Law: What, How & Why

Benford's Law formula can be given by this equation:

$$P(d) = \log\left(1 + \frac{1}{d}\right), \quad d = 1,2,3 \ldots 9 \ ;$$

where d is the first digit of the data and P(d) is the probability of getting that digit. This probability function can be used to calculate the general or average probability of each digit for a dataset that obeys Benford's law. Below is how the distribution of the digits would be. [3]

d	$P(d)$	Relative size of $P(d)$
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%	

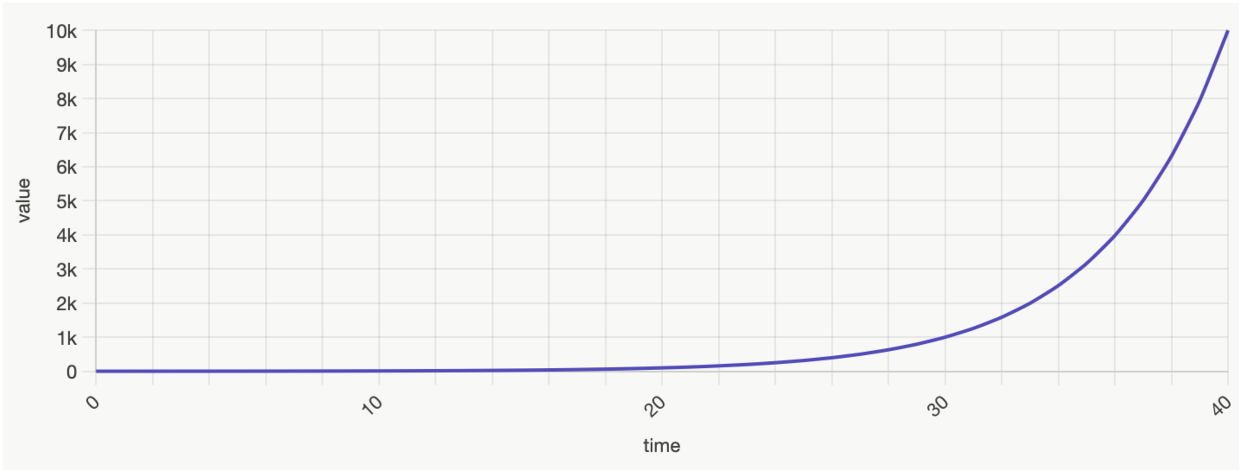
2.1) Natural Growth & Logarithms

But why is it like that? Why would Benford's Law even work?

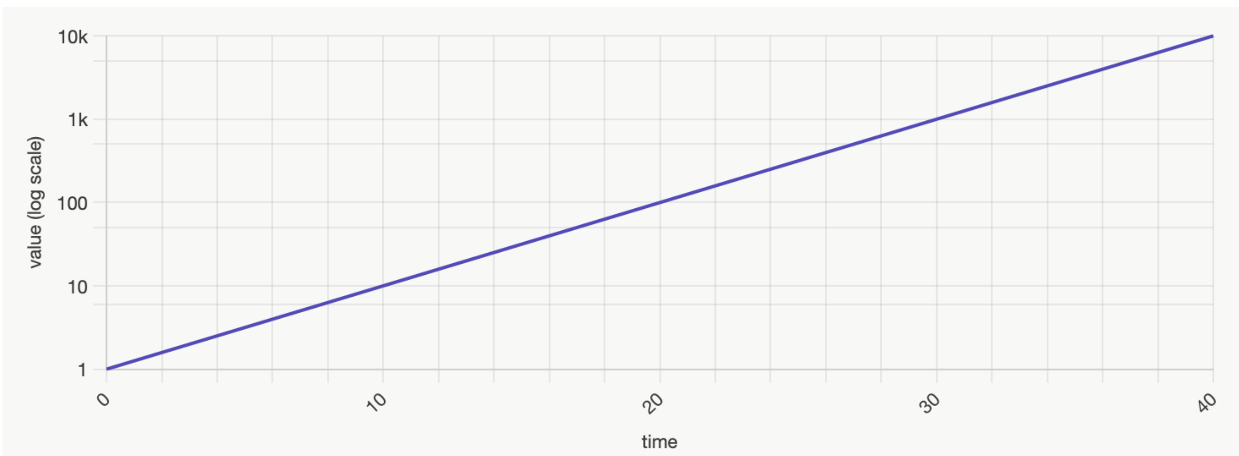
While there isn't a clear mathematical proof behind this law, we could explain this phenomenon with logical reasons.

Most values in the real world grow relatively instead of additively over time in an exponential manner. To picture this, you can think about growing your money in an investment. If say you have $100 (it can be your country's currency or any of your preference), and you want to grow it to $200, it would be a 100% increase. In comparison, growing $900 to $1000 would only be a 11% increase. Hence, it can be said that it is easier to "get in" to a number with '1' as the leading digit than "get out" of it, which makes a smaller number appear more frequently as the first digit. This is why Benford's law is most accurate in values that are distributed across multiple orders of magnitude (because they grow exponentially) and their production is described as a power law (another interesting math phenomenon found in nature).

To illustrate a growth of such value, a simple exponential curve is shown below.

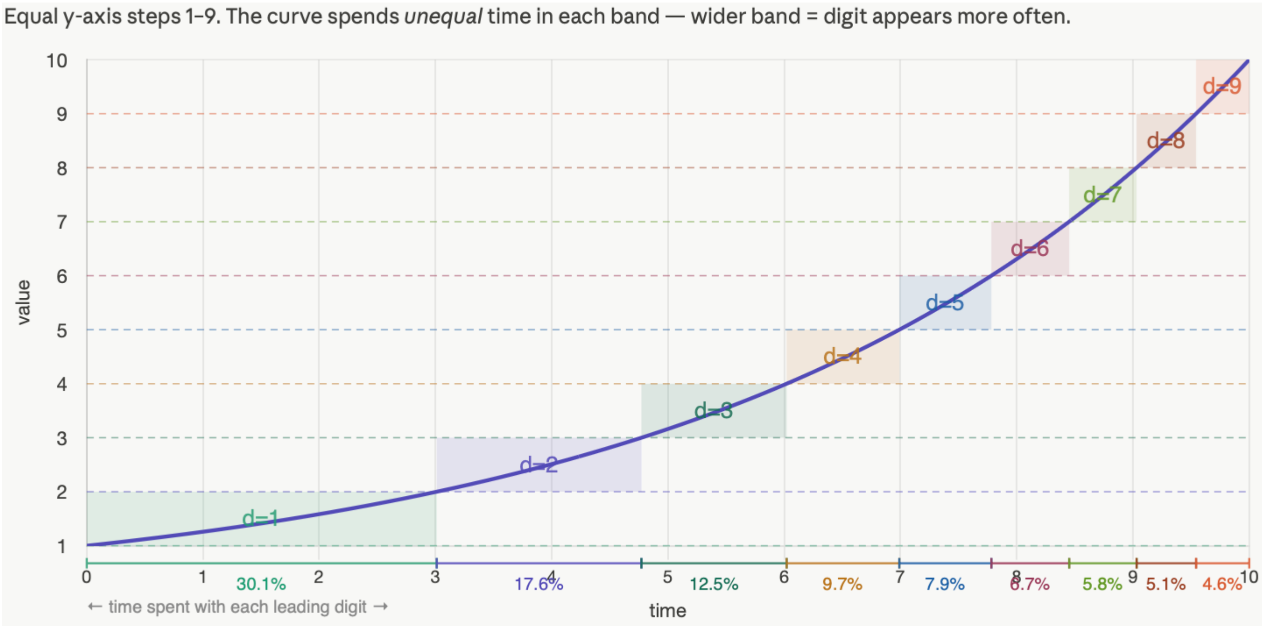


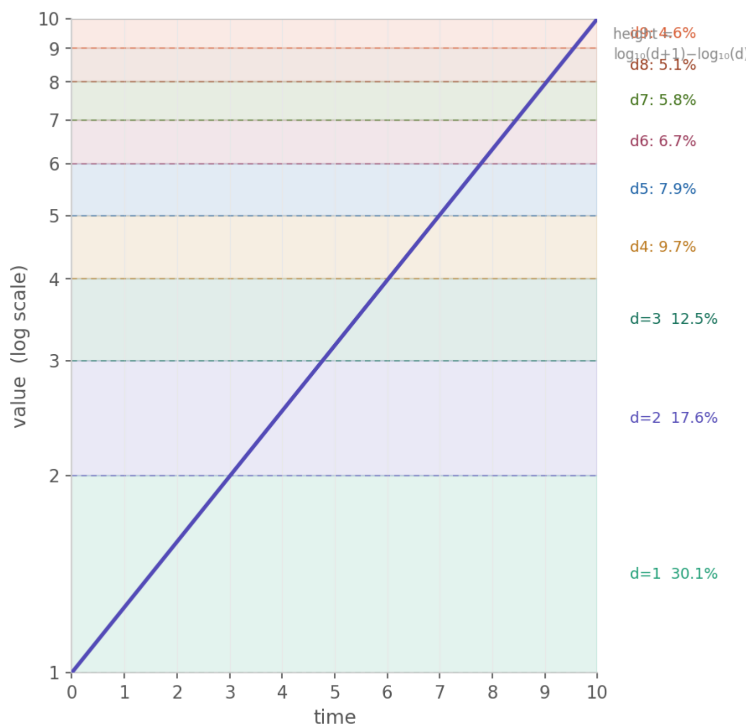
By changing it to a semi-log scale (logarithmic y-axis), we could get a linear graph.



Looking at the linear graph, it can be seen that the value is multiplying at each interval at a constant rate and not adding like the classic linear equation $y = mx + c$; resembling the relative growth often found in real-life situations.

From here, the proportion of (or time spent on) each leading number (1,2,3… 9) can be determined by the width of each interval on the linear scaled exponential curve or the length of the logarithmic scaled one. This can be visualised using these two diagrams with highlighted areas on each specific graph.





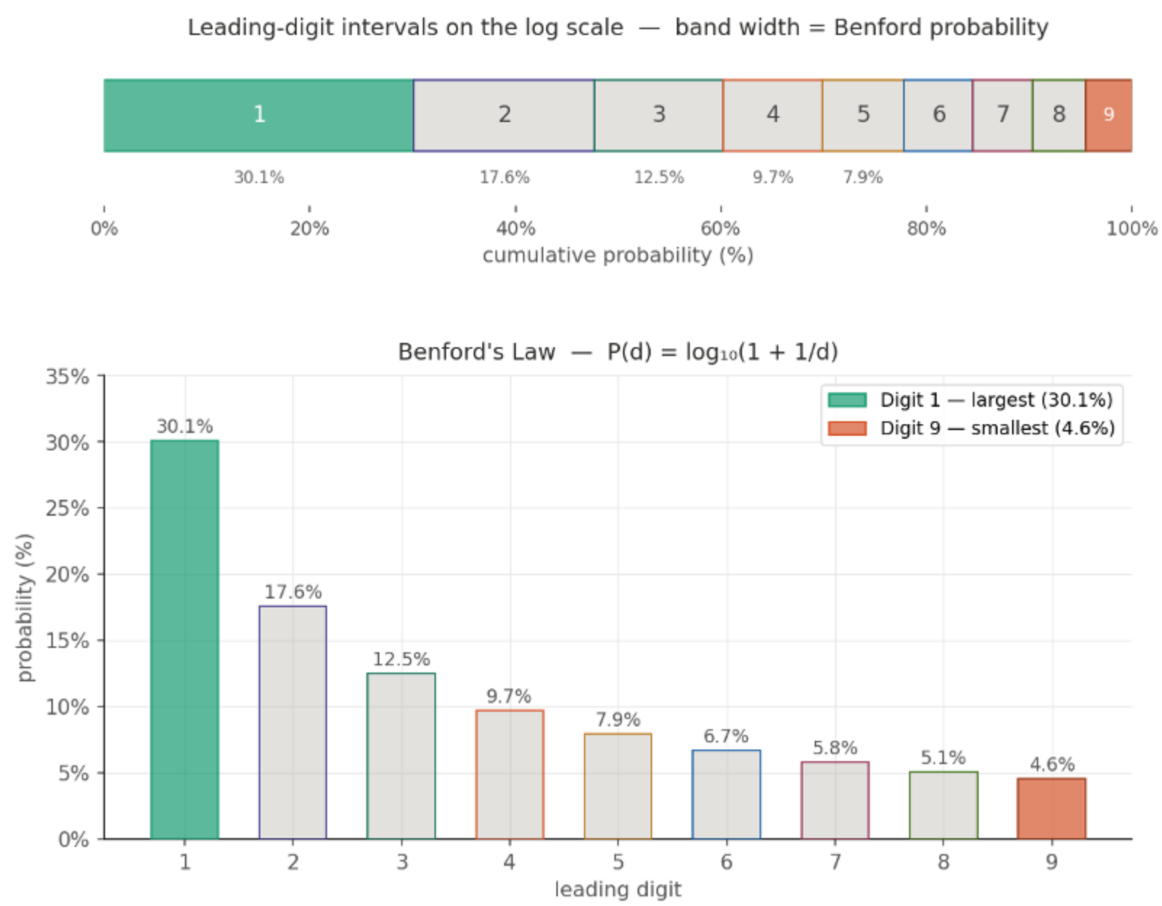
If you look at the log scale visualisation, the length of each step on the vertical axis can be calculated by the difference in log. For example, the length of 1 to 2:

$$\log(2) - \log(1) = \log(2) - 0 = \log(2) = 0.301$$

This is 30.1 % and equal to the probability of 1 as the leading digit in Benford's Law, and the equation for Benford's Law formula can also be derived.

$$P(d) = \log(d+1) - \log(d)$$
$$P(d) = \log\left(\frac{d+1}{d}\right)$$
$$P(d) = \log\left(1 + \frac{1}{d}\right)$$

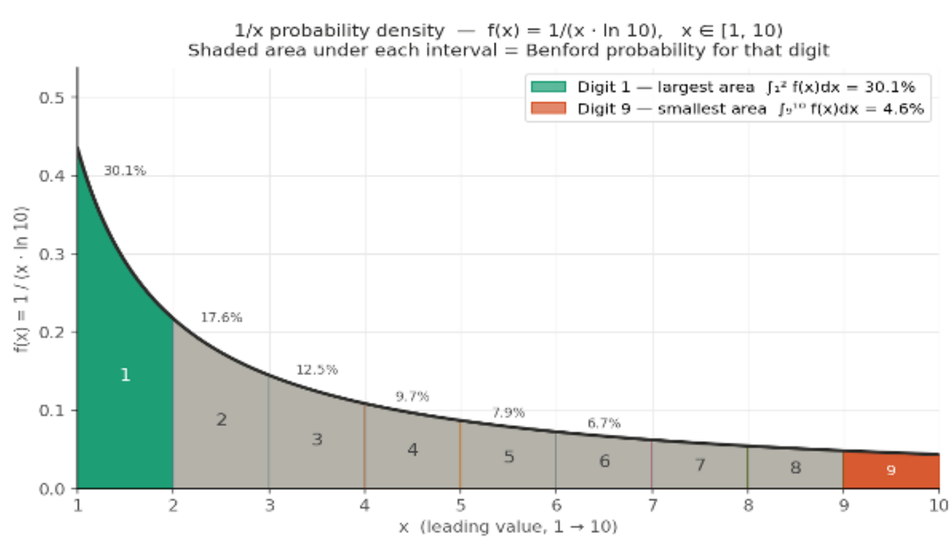
The widths of each band can also be taken to form the discrete probability distribution of each leading digits.



2.2) Scale Invariance & Another Perspective

There is a different way to look at the Benford's Law, which is also the one I prefer. See how the distribution (diagram above) follows a shape similar to the reciprocal function? We can represent the law as a continuous probability density function, that is given by:

$$f(x) = \frac{1}{xln(10)}, 1 \leq x \leq 10$$



To find the probability of each leading digit, we can just integrate the function to get the area.

$$P(d) = \int_{d}^{d+1} \frac{1}{xln(10)} dx$$
$$= \left[\frac{\ln(x)}{\ln(10)}\right]_{d}^{d+1}$$
$$= [\log_{10} x]_{d}^{d+1}$$
$$= \log(d+1) - \log(d)$$
$$= \log\left(1 + \frac{1}{d}\right)$$

And again, we get Benford's Law formula.

This also shows why it is scale invariant -- which simply means the units used in the dataset does not matter, because the reciprocal function is scale invariant.

Say the reciprocal function was scaled by a factor of b such that

$$g(x) = f(bx)$$
$$g(x) = \frac{1}{b} \cdot \frac{1}{x}$$
$$g(x) = \frac{1}{b} f(x)$$

The relative area of each leading digit would still be the same as before. Ergo, the relative frequency would remain the same.

<u>3.3) Base Change & Other Digits</u>

Partly because of scale invariance, Benford's law also works on other bases. For a particular base, the number of possible outcomes can be expressed as b-1. For example, for the standard base 10, will have 1,2,3…9 (10-1 = 9). The formula can be generalised to this for b ≥ 2.

$$P(d) = \log_b(d+1) - \log_b(d) = \log_b\left(1 + \frac{1}{d}\right)$$

For binary and unary numbers (b = 2,1), Benford's law is true but trivial, all numbers except empty sets would start with a 1.

Benford's formula could also be generalised beyond the first digits of numbers or values.

$$P(n) = \log(1 + \frac{1}{n})$$

For instance, if you want to calculate the probability of a number starting with 2,7,1 (which can be 271, e and 0.27123456) is $\log\left(1 + \frac{1}{271}\right) = 0.0015996$. For a given position and given number, like the second digit being '2', it would be $\log\left(1 + \frac{1}{12}\right) + \log\left(1 + \frac{1}{22}\right) \ldots + \log\,(1 + \frac{1}{92}) = 0.109$. It is interesting to note that for the second-digit and onwards, the probably that a digit is '0' is also possible. For first digit, it could not be 0 because values are non-zeros and log (0) is undefined as well.

3) Just A Math Phenomenon?

Benford's Law may look like a math phenomenon and nothing more. But this amazing statistical principle has its role in real-life usage rather than just something to be marvelled at. This law is most famously applied in the financial world and serves as a tool to detect fraud.

When datasets in accounting books or records seems to fit the criteria for Benford's law, auditors can use a few statistical tests like the Chi-Square test or the 2 Proportion Z test to see if they conform to Benford's law or not. [4] Besides examining the finances of companies, these methods can also be used for any scale (-invariance ☺), from tax authorities checking employees' tax declaration, up to official bodies analysing countries' financial and economic data. [5]

It is important to note that the results obtained is only an indicator for further investigation by auditors and not a definite tell of financial fraud. This is because many factors like human errors or unrealised artificial constraints can affect the first-digit distribution.

4) The Caveat

Elections is also an area where numerous factors affect the frequency of first digits of the number of votes. One example where Benford's law is greatly misapplied is the 2020 US elections. After the votes were tallied and result were published, many social media users rushed to share posts claiming fraud in that term's election, and Biden and his party have 'cheated'. [6] In response, many analysts and mathematicians implied that Benford's law is actually a poor way to prove election fraud and other methods should be used instead.

5) Conclusion

In summary, a dataset that has values that go through exponential growth or decay and spread across at least an order of magnitude would most likely comply to the Benford's law.

This is the end of my sharing of the Benford's law. I had a great time learning about this mathematical phenomenon and it is my pleasure to share it with you, the reader. I hope that you have learned something if you did not know about this topic before, add it to your math knowledge and maybe use it as a cool party trick to impress others. Thank you for reading till the end, and think twice before answering math questions. Goodbye!

References

1. https://en.wikipedia.org/wiki/Frank_Benford
2. https://en.wikipedia.org/wiki/Simon_Newcomb#Benford's_law
3. https://en.wikipedia.org/w/index.php?title=Benford%27s_law&oldid=1346969808#History
4. https://pillars.taylor.edu/cgi/viewcontent.cgi?article=1005&context=mathstudentscholarship
5. https://www.forbes.com/sites/taxnotes/2021/08/19/can-benfords-law-detect-tax-fraud/
6. https://www.reuters.com/article/world/fact-check-deviation-from-benfords-law-does-not-prove-election-fraud-idUSKBN27Q3A9/