stats

MDPI

Article

First Digit Oscillations

Don Lemons [1,*,†], Nathan Lemons [2,†] and William Peter [3,*,†]

1 Department of Physics, Bethel College, North Newton, KS 67117, USA
2 Los Alamos National Laboratory, Los Alamos, NM 87545, USA; nlemons@lanl.gov
3 John Hopkins Applied Physics Laboratory, Laurel, MD 20723, USA
* Correspondence: lemons.don@gmail.com (D.L.); bill.peter@gmail.com (W.P.)
† These authors contributed equally to this work.

Abstract: The frequency of the first digits of numbers drawn from an exponential probability density oscillate around the Benford frequencies. Analysis, simulations and empirical evidence show that datasets must have at least 10,000 entries for these oscillations to emerge from finite-sample noise. Anecdotal evidence from population data is provided.

Keywords: Benford's law; oscillations; exponential probability density

Citation: Lemons, D.; Lemons, N.; Peter, W. First Digit Oscillations. *Stats* **2021**, *4*, 595–601. https://doi.org/10.3390/stats4030035

Academic Editors: Claudio Lupi, Roy Cerqueti and Marcel Ausloos

Received: 29 April 2021
Accepted: 24 June 2021
Published: 5 July 2021

1. Introduction

According to Benford's law, the frequency of the first digits of numbers are larger for digit 1 (about 30%) than 2 (about 18%) and so on up to 9 (about 5%). The "law" that governs these probabilities b_d is

$$b_d = \log_{10}(1 + 1/d), \tag{1}$$

where $d = 1, 2, \ldots, 9$. This law originated with Simon Newcomb [1] and was popularized by Frank Benford [2]. In 1995, T. P. Hill [3] proved a theorem that helps explain the success of Benford's first digit law. According to Hill's theorem, the frequency of the first digits of numbers randomly drawn from randomly chosen distributions converge to b_d in the limit of large numbers. Several books introduce and summarize findings on the subject [4–6].

Benford illustrated Equation (1) with "found" or empirical datasets drawn from a number of sources. Many empirical sets of numbers observe or approximate Benford's first digit law, particularly those that (1) span several decades; (2) have positive skewness; (3) have many entries; and (4) are not intentionally designed. Such datasets have been called "Benford suitable" by Goodman [7].

Even so, some numerically generated datasets that are Benford suitable do not observe Benford's law in detail. In particular, consider numbers drawn from an exponential probability density

$$p_\lambda(t) = \lambda \exp(-\lambda t), \tag{2}$$

where $0 \leq t \leq \infty$ and λ are the rate or, equivalently, the inverse mean of the exponential probability density given by

$$\lambda^{-1} = \int_0^\infty t p_\lambda(t) dt. \tag{3}$$

The first digits of numbers drawn from Equation (2) oscillate with λ around b_d with amplitudes of about 10%.

Random numbers drawn from the exponential probability density (2) are important because they approximate pieces of a quantity that is randomly partitioned [8]. Suppose, for instance, that a population P is to be divided, without bias, into M cities and towns in such a way that the mean city size P/M is a definite quantity. If this partition is done so as to maximize the entropy of the partition, we find that the probability of city size t is given

by (2), where $\lambda = M/P$. See Appendix A for a derivation of this claim that is inspired by a similar derivation by Iafrate, Miller, and Strach [9]. Miller and Nigrini [10] also explore relations between the exponential probability density (2) and Benford's law (1).

We might expect the oscillations in the exponential probability density (2) with λ to have been observed in real-world data. However, our analysis shows that the predicted oscillations emerge from finite-sample noise only with a sample number $N > 10{,}000$. We also describe a real-world example of this first digit oscillation in the populations of US towns and cities as they have evolved over the last hundred years.

2. First Digit Probabilities

The probability $g_d(\lambda)$ that a number drawn from the exponential distribution (2) has a first digit d is given by

$$\begin{aligned} g_d(\lambda) &= \sum_{k=-\infty}^{\infty} \int_{d10^k}^{(d+1)10^k} p_\lambda(t)dt \\ &= \sum_{k=-\infty}^{\infty} \left[\exp(-\lambda d10^k) - \exp(-\lambda(d+1)10^k)\right]. \end{aligned} \tag{4}$$

According to Equation (4), the first digit probability $g_d(\lambda)$ is periodic in λ in the sense that $g_d(10\lambda) = g_d(\lambda)$. Reference [11], from whose paper the contents of this section originate, demonstrated that the averages of $g_d(\lambda)$ over one decade of λ are the Benford frequencies b_d. The $n = 0$ and $n = \pm 1$ terms of a Fourier expansion of (4) produce the formula

$$g_d(\lambda) \approx b_d + \left(\frac{4r}{\ln 10}\right) \sin\left[\pi \log_{10}(1 + 1/d)\right] \sin\left[\theta + 2\pi \log_{10}\left(\lambda\sqrt{d(1+d)}\right)\right], \tag{5}$$

where r and θ are, respectively, the absolute value and argument of the gamma function $\Gamma(-2\pi i/\ln 10)$.

The first two factors in the second term on the right hand side of (5) characterize an oscillation amplitude of approximately 10% of the non-oscillating term b_d, while the last factor is periodic in $\log_{10}(\lambda)$. The $n = \pm 2$ Fourier coefficients are approximately 10^{-2} times smaller than the $n = \pm 1$ coefficients. Higher order coefficients are even smaller. Indeed, formula (5) produces curves visually identical to those produced by the more complete expression (4).

3. Sample Noise

Because the magnitude of the oscillating term in Equation (5) is approximately 10% of the non-oscillating term b_d , its effect can easily be swamped by the noise inherent in finite datasets and finite samples from the exponential probability distribution (2). We see this in the following way.

Assume a list of N identically distributed, statistically independent, random numbers indexed with $j = 1, 2, \ldots N$. Now let $X_{d,j}$ be an indicator random variable defined so that $X_{d,j} = 1$ when the number subscripted j begins with digit d and $X_{d,j} = 0$ when the number subscripted j does not begin with digit d. We then define the frequency G_d of the first digit d among N numbers as

$$G_d = \frac{1}{N} \sum_{j=1}^{N} X_{d,j}. \tag{6}$$

The expectation value of both sides of Equation (6) is

$$\begin{aligned} E[G_d] &= \frac{1}{N} \sum_{j=1}^{N} E[X_{d,j}] \\ &= E(X_d), \end{aligned} \tag{7}$$

since the $X_{d,j}$ are identically distributed, $E[X_{d,1}] = E[X_{d,2}] = \ldots E[X_{d,N}]$, and, therefore, we may denote each of these terms by $E[X_d]$. When the numbers determining the indicator random variables $X_{d,j}$ are drawn from an exponential distribution, $E[G_d]$ and $E[X_d]$ are equal to $g_d(\lambda)$.

The variance of G_d is generated from

$$\begin{aligned} G_d^2 &= \frac{1}{N^2}\sum_{i,j=1}^{N} X_{d,i}X_{d,j} \\ &= \frac{1}{N^2}\sum_{i=1}^{N} X_{d,i}^2 + \frac{1}{N^2}\sum_{i,j=1,i\neq j}^{N} X_{d,i}X_{d,j}. \end{aligned} \tag{8}$$

Since the $X_{d,i}$ are statistically independent and identically distributed, the $E(X_{d,i}^2)$ are identical and, therefore, can be denoted by $E(X_d^2)$. Consequently, we find that

$$E(G_d^2) = \frac{E(X_d^2)}{N} + \frac{N(N-1)E(X_d)^2}{N^2}. \tag{9}$$

Therefore, the variance $\sigma_{G_d}^2$ is given by

$$\begin{aligned} \sigma_{G_d}^2 &= E(G_d^2) - E(G_d)^2 \\ &= \frac{E(X_d^2) - E(X_d)^2}{N}. \end{aligned} \tag{10}$$

However, because X_d is an indicator random variable with only two possible values, 0 and 1, $E(X_d^2) = E(X_d)$. In this case, the variance (10) becomes

$$\sigma_{G_d}^2 = \frac{E(X_d) - E(X_d)^2}{N} \tag{11}$$

and the relative standard deviation becomes

$$\frac{\sigma_{G_d}}{E(G_d)} = \frac{1}{\sqrt{N}}\sqrt{\frac{1}{E(X_d)} - 1}. \tag{12}$$

Recall that $E(G_d) = E(X_d)$ and that the analysis resulting in Equations (11) and (12) applies generally to any distribution with indicator random variable X_d and expectation value $E(X_d)$. Relations (11) and (12), between variance and mean, are, of course, not new. They also follow from the binomial probability distribution that governs the indicator variables.

Given a Benford probability b_d or Benford probability plus oscillation $g_d(\lambda)$ and standard deviation σ_{G_d}, Equation (12) tells us how many samples N from a distribution are required to resolve the mean frequency in the presence of finite-sample noise. For instance, in order that the relative standard deviation be small enough for the Benford frequency b_1 $(= 0.301)$ of digit $d = 1$ to emerge from noise, say, about 10% of b_1, $(1/\sqrt{N})\sqrt{1/\log_{10}(2) - 1} \leq 1/10$, which means that $N \geq 200$. However, if one also wants the oscillation of $g_1(\lambda)$ around the Benford frequency b_1 to emerge from sample noise, another 10% reduction in relative standard deviation is needed. In this case, $(1/\sqrt{N})\sqrt{1/\log_{10}(2) - 1} \leq 1/(10 \cdot 10)$ or $N \geq 20{,}000$. We illustrate these calculations in the next section.

4. Sampling

The cumulative probability of the exponential probability density defined in Equation (2) is

$$\int_0^{t'} \lambda \exp(-\lambda t) dt = 1 - \exp(-\lambda t'), \tag{13}$$

and can be replaced by the uniform random variable $U(0,1)$, or, equivalently, by $1 - U(0,1)$. Simultaneously, t' becomes the random variable T drawn from the exponential probability density (2). Therefore, Equation (13) implies that

$$T = \frac{-\ln U(0,1)}{\lambda}. \tag{14}$$

The first digit random variable frequency G_d as determined from the random variables generated by Equation (14) should reflect the 10% oscillations with period $\log_{10} \lambda$ as predicted by (5), and so they do, as long as the number of samples N is large enough. For more details concerning sampling consult reference [12].

According to (12), the relative standard deviation $\sigma_{G_d} / E(G_d)$ is smallest for a given sample size N when $E(X_d)$ is largest. For exponentially distributed probabilities this means the oscillation in λ will most easily be seen in samples of the random variable G_1, that is, when $d = 1$. Figures 1–3 show sample values of G_1 for N $(= 10^2, 10^3, 10^4)$ as a function of λ between 10^{-2} and 10^{-1}. Values of the random variable G_1 are shown as filled circles. The central curve is $g_1(\lambda)$ from Equation (5), and the two surrounding curves are $g_1 \pm \sigma_{G_1}$ from Equation (11) or Equation (12) with $E(G_1) = E(X_1) = g_1(\lambda)$. Values of the random variable G_1 mainly stay within the standard deviation curves.

A sample size of $N = 100$ hardly allows one to discern the Benford frequency b_1 much less the oscillation around b_1. Only with larger samples, on the order of $N = 10{,}000$, does the 10% oscillation emerge from finite-sample noise.

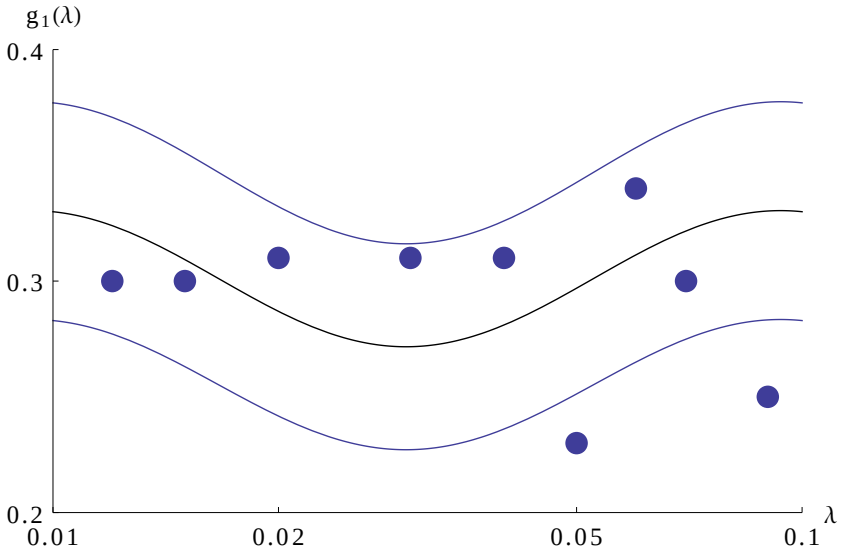

Figure 1. Frequency of first digit 1 versus λ from (5) and (11) or (12) with $N = 100$. Filled circles are first digit frequencies from $N = 100$ samples generated by (14).

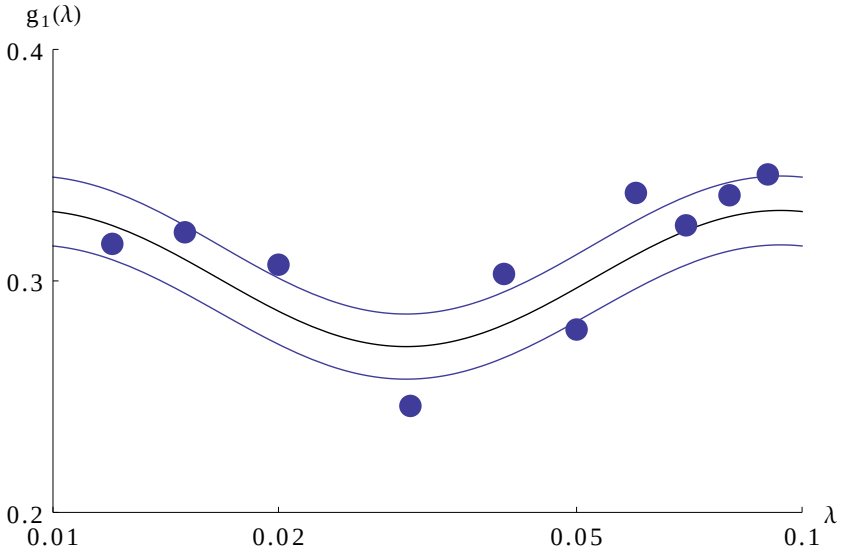

Figure 2. Frequency of first digit 1 versus λ . Curves are from (5) and (11) or (12) with $N = 1000$. Filled circles are first digit frequencies from $N = 1000$ samples generated by (14).

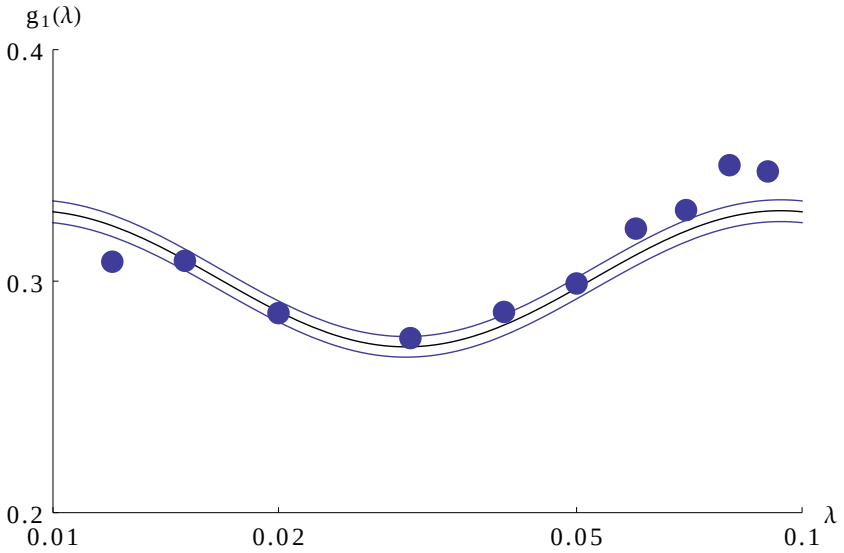

Figure 3. Frequency of first digit 1 versus λ . Curves are from (5) and (11) or (12) with $N = 10{,}000$. Filled circles are first digit frequencies from $N = 10{,}000$ samples generated by (14).

5. Population of Towns and Cities in the USA

In order to observe the predicted first digit oscillations in real-world data, one must find datasets with more than approximately 10,000 entries and with several different values of the inverse mean λ. For a first effort, no data seems more likely to reveal these oscillations than that of the US Census Bureau as described in [13]; in particular, the populations of incorporated towns and cities at different 10-year intervals. However, only the decennial censuses from 1970 forward have been digitized. While the population of the USA has increased by 50% since 1970, the number of towns and cities has also increased. For this reason, the inverse mean of the municipal population λ has changed very little between 1980 and 2010.

In order to find town and city population data with significantly different inverse means λ, we reached back to the census of 1910. After making the considerable effort to digitize the 1910 populations of 14,000 incorporated towns and cities as listed in the pdf made available by the US Census Bureau [14], we sorted these numbers (and others from the 1980–2010 period) according to first digits.

Figure 4 shows the result of our efforts. Here we see the frequency of leading digit 1 for the decennial censuses 1980–2010 (the leftmost group of filled circles) and for 1910 (the rightmost filled circle) versus their respective inverse mean population per town or city λ. The probability $g_1(\lambda)$ of first digit 1 as determined by formula (5), which derives from the exponential distribution (2) with parameter λ, is also shown. Figure 4 does not have standard deviation curves because the number of samples is different for each point.

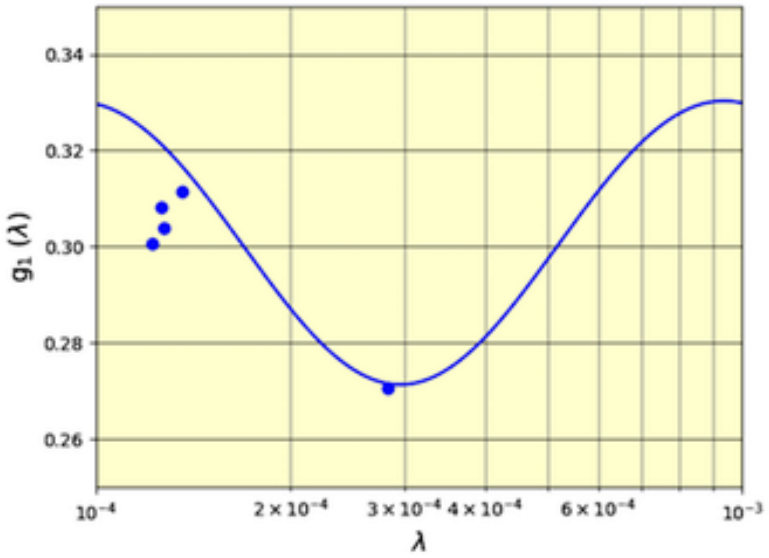

Figure 4. First digit frequencies versus λ. Filled circles are frequencies of first digit 1 from the 1980–2010 (leftmost group) and 1910 (rightmost circle) US census. The curve is frequency versus λ from formula (5) derived from the exponential probability distribution (2).

Of course, these data merely suggest that the 10% first digit oscillations around Benford frequencies are a feature of population and other real-world data. As such, we hope it encourages others to look for more conclusive evidence. However, as noted, the prerequisite for this search is a Benford suitable dataset with at least 10,000 entries.

6. Summary and Conclusions

In Equation (5), we have made explicit the periodic dependence of the first digit frequencies $g_d(\lambda)$ of numbers that are drawn from an exponential distribution with rate λ. According to this relation, the amplitude of these oscillations is approximately 10% of the Benford frequencies b_d. We have also demonstrated that the number of data entries required to allow these 10% oscillations to emerge from sample noise in real-world data should be larger than about 10,000. We have illustrated this requirement in numerical realizations of the simulation algorithm in Equation (14). The populations of US towns and cities spanning a century is real-world, if anecdotal, evidence of these first digit oscillations.

While the requirement of 10,000 numbers sets a high bar, sufficiently large Benford suitable datasets do exist and have been sorted according to first digit [15]. The first digit frequencies reported in [15] appear to be consistent with the predicted 10% oscillations around Benford values. However, the appropriate value of λ , which determines the phase of these oscillations, was not reported.

Alternatively, one might repeat the same experiment many times in which a given quantity is partitioned in such a way that the inverse mean of the partition sizes λ is constant. Then, according to the law of large numbers, the mean values of the frequency of first digits will converge to those described by formula (5). Our analysis may explain why those mining specific datasets for evidence of Benford's law may only fortuitously find agreement within 10% of b_d and then only for certain digits.

Author Contributions: All authors contributed equally to this project. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: One of the authors (DSL) acknowledges helpful discussions with Alex Kossovsky.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Suppose a quantity P is partitioned among M pieces in such a way that the inverse mean $M/P = \lambda$ is fixed and the entropy of this partition is maximized. The probability density of these partitions is normalized so that

$$1 = \int_0^\infty p(t)dt, \tag{A1}$$

the inverse mean

$$\lambda^{-1} = \int_0^\infty tp(t)dt, \tag{A2}$$

and the entropy

$$S = -\int_0^\infty p(t)\ln p(t)dt. \tag{A3}$$

Finding the stationary value of the entropy (A3) subject to the constraints (A1) and (A2) means solving

$$\frac{\delta}{\delta p(t')}\int_0^\infty [-p(t)\ln p(t) - \alpha p(t) - \beta t p(t)]dt = 0, \tag{A4}$$

the solution of which is

$$p(t) = \exp(-1-\alpha)\exp(-\beta t). \tag{A5}$$

Requiring (A5) to satisfy constraints (A1) and (A2) produces the exponential probability density (2), that is,

$$p_\lambda(t) = \lambda\exp(-\lambda t). \tag{A6}$$

References

1. Newcomb, S. Note on the frequency of use of the different digits in natural numbers. *Am. J. Math.* **1881**, *4*, 39–40. [CrossRef]
2. Benford, F. The law of anomalous numbers. *Am. Philos. Soc.* **1939**, *78*, 551–572.
3. Hill, T.P. A statistical derivation of the significant-digit law. *Stat. Sci. A Rev. J. Inst. Math. Stat.* **1995**, *10*, 354–363. [CrossRef]
4. Kossovsky, A.E. *Benford's Law: The General Law of Relative Quantities, and Forensic Fraud Detection Applications*; World Scientific: Singapore, 2014.
5. Berger, A.; Hill, T.P. *An Introduction to Benford's Law*; Princeton University Press: Princeton, NJ, USA, 2015.
6. Miller, S.J. (Ed.) *Benford's Law: Theory and Application*; Princeton University Press: Princeton, NJ, USA, 2015.
7. Goodman, W. The promises and pitfalls of Benford's law. *Significance* **2016**, *13*, 38–41. [CrossRef]
8. Lemons, D.S. On the numbers of things and the distribution of first digits. *Am. J. Phys.* **1986**, *64*, 816–817. [CrossRef]
9. Iafrate, J.R.; Miller, S.J.; Strauch, F.W. Equipartitions and a Distribution for Numbers: A statistical Model for Benford's Law. *Phys. Rev. E* **2015**, *91*, 062138. [CrossRef]
10. Miller, S.J.; Nigrini, M.J. Order Statistics and Benford's Law. *Int. J. Math. Math. Sci.* **2008**, *2008*, 382948. [CrossRef]
11. Engel, H.; Leuenberger, C. Benford's law for exponential random variables. *Stat. Probab. Lett.* **2003**, *63*, 361–365. [CrossRef]
12. Ross, S. *Simulation*, 5th ed.; Academic Press: San Diego, CA, USA, 2012; p. 69.
13. Manson, S.; Schroeder, J.; Riper, D.V.; Ruggles, D. U.S. Census Data. In *Ipums National Historical Geographic Information System: Version 14.0 [Database]*; 2019. Available online: https://ipums.org/projects/ipums-nhgis/d050.v14.0 (accessed on 1 June 2019).
14. *Thirteenth Census of the United States: 1910 (Downloaded: November, 2019)*; 2019. Available online: https://www.loc.gov/item/13008447/ (accessed on 1 June 2019).
15. Sambridge, M.; Tkalčić, H.; Jackson, A. Benford's Law in the natural sciences. *Geophys. Res. Lett.* **2010**, *37*, L22301–L22306. [CrossRef]