

Benford's law and average significands

ZORAN JASAK¹, EDIN GLOGIĆ²,

¹Faculty of technical sciences, University FINRA Tuzla

Mitra Trifunovića Uče 6, Tuzla,

BOSNIA AND HERZEGOVINA

²Faculty of economy, University FINRA Tuzla

Mitra Trifunovića Uče 6, Tuzla,

BOSNIA AND HERZEGOVINA

Abstract. Sum invariance property and average significands are narrow connected, based on properties of logarithmic function. Theoretical values of average significands, together with theoretical values of minimal and maximal significands, are important for testing purposes. Theoretically, sum of significands with same leading digits is product of Benford's probability and average significand. By this it's possible to analyse goodness-of-fit for both values, specially in big samples. For example, if sample sum of significands has a good fit with theoretical values sample average significand can significantly deviate from theoretical value, what can point to some kind of anomaly.

Key words: Benford's law, significand, sum invariance, average significand

Received: March 18, 2026. Revised: June 13, 2026. Accepted: August 6, 2026. Published: October 1, 2026.

1. Introduction

In 1881 Simon Newcomb in the article "Note on the Frequency of use of different digits in natural numbers" [1], stated that "the first significant digit is oftener 1 than any other digit, and the frequency diminishes up to 9". This idea came from Newcomb's observation of the use of library logarithm tables where he noticed that the first pages of these tables were actually dirtier than the last ones. From this he concluded that people are more likely to use numbers starting with smaller than with larger digits. This phenomenon was re-discovered by Frank Albert Benford, who In 1938 published a paper entitled The Law of Anomalous Numbers [2]. He investigated 20 different sets of natural numbers involving over 20,000 samples, from newspaper pages to home addresses. He gave mathematical formulation $P[D = d] = \log_{10}(1 + 1/d)$ where d is leading digit, $d \in \{1, 2, \dots, 9\}$. This theory was then tested on a large number of different statistical data sets and proved to hold true for most of them, with similar laws derived for other order digits. It's observed that the more mixed the data, the closer the distribution of digits was to the logarithmic one. This is now commonly called Benford's Law. Other mathematicians (S. Goudsmit, W. Furry, H. Hurwitz, R. Pinkham, R. Raimi, and especially T. Hill) gave the theoretical basis for this

law. On web site <http://www.benfordonline.net> it's possible to find numerous texts both on theoretical and practical aspects of Benford's law.

2. Average significands

2.1. Significands

Significand of real positive number x is its coefficient when presented in floating point form: $x = S(x) \cdot 10^m$, sometimes called engineering notation. Formal definition of significand is next [3].

Definition. Function significand $S: \mathbb{R} \rightarrow [1, 10)$ is defined in next way: if $x \neq 0$ then $S(x) = t$, where t is unique number in $[1, 10)$ with $|x| = 10^m \cdot t$ for some (necessary) unique $m \in \mathbb{Z}$; if $x = 0$ then is, by convention, $S(0) := 0$.

For all $x \in \mathbb{R}$ is: $S(10^k x) = S(x)$, $S(S(x)) = S(x)$, $\forall k \in \mathbb{Z}$. Explicitly, $S(x)$ is given by:

$$S(x) = 10^{\log|x| - \lfloor \log|x| \rfloor} = \frac{10^{\log|x|}}{10^{\lfloor \log|x| \rfloor}} = \frac{x}{10^{\lfloor \log|x| \rfloor}}$$

For some base $B > 1$ it is:

$$S_B(x) = \frac{x}{B^{\lfloor \log_B x \rfloor}}$$

where $\lfloor \cdot \rfloor$ stands for function floor, largest integer not greater than x . By this formula we actual get number with one integer place (removing decimal point left or right), namely values in interval $[1; B)$.

If k is leading digit next generalization can be used:

$S^{(k)}(x) = 10^{k-1} \cdot S(x) = 10^{\log|x| - \lfloor \log|x| \rfloor + k-1}$
In case $k > 1$ by $S^{(k)}(x)$ we denote significand of order k . Formaly, we can write $S(x) = S^{(1)}(x)$.

2.2. Invariance properties of Benford's law

Four important properties of Benford's are base, scale, inverse, sum invariance.

- Base invariance property means that Benford's law holds if we use any base $B > 1$.
- Scale invariance property means that Benford's law holds if all numbers in sample are multiplied by the same positive number. Theoretical explanations gave Hill [7].
- Inverse invariance property means that Benford's law holds if all numbers in sample are changed by their inverse values.
- Sum invariance property means that sum of all significands with equal leading digits are equal in expectation.

Proofs of these properties are available in literature about Benford's law.

2.3. Theoretical average significands

Origin of idea of average significands is sum invariance property of Benford's law. Idea of this property is presented by Mark Nigrini who asserted in his PhD thesis (1992) that tables of unmanipulated accounting data closely follow Benford's Law and that in sufficiently long list of data for which Benford's law holds the sum of all entries with leading digit d is constant.

In [6] next theorem is presented and proved.

Theorem. A probabilistic measure P for Benford's law is sum invariant if and only if $[B^{k-1}, B^k)$ is divided on n subintervals of equal size.

Proof. Area under logarithmic curve on interval $[1, 10)$ is:

$$\begin{aligned} A_1^{(1)} &= \int_1^{10} \log x \, dx = \frac{1}{\ln 10} (x \ln x - x) \Big|_1^{10} \\ &= \frac{10 \ln 10 - 9}{\ln 10} = 10 - 9 \cdot \log_{10} e \\ &= 6.091349663 \end{aligned} \quad (1)$$

Area of rectangle bounded by lines $x = 0$, $x = 10$, $y = 0$, $y = 1$ is 10. From this we have that area of curvilinear rectangle bounded by lines $x = 0$, $y = 0$, $y = 1$ and function $y = \log_{10} x$, above logarithmic

curve on interval $[0; 10]$, is :

$$A_2^{(1)} = 10 - A_1 = 9 \cdot \log_{10} e = 3.908650337$$

The same we get as area under inverse function $y = 10^x$ on $[0, 1]$. As a general rule we have:

$$\begin{aligned} A_0^{(k)} &= \int_{10^{k-1}}^{10^k} \log x \cdot dx \\ &= \frac{(k \cdot 10^k - (k-1) \cdot 10^{k-1}) \cdot \ln 10 - 10^{k-1} \cdot 9}{\ln 10} \\ &= (k \cdot 10^k - (k-1) \cdot 10^{k-1}) - 10^{k-1} \cdot 9 \cdot \log_{10} e \end{aligned}$$

In this formula k corresponds to number of integer digits in significand. Area bounded by lines $y = k-1$, $x = 10^k$ and by function $y = \log_{10} x$ is:

$$\begin{aligned} A_1^{(k)} &= A_0^{(k)} - (k-1) \cdot (10^k - 10^{k-1}) \\ &= 10^k - 10^{k-1} \cdot 9 \cdot \log_{10} e \end{aligned} \quad (2)$$

For $k = 1$ we have relation (1). In relation (2) part 10^k is area of rectangle bounded by lines $x = 0$, $x = 10^k$, $y = k-1$ and $y = k$. On picture this is rectangle $ECDF$. From this $A_2^{(k)} = 10^{k-1} \cdot 9 \cdot \log_{10} e$ is area of curvilinear rectangle bounded by lines $y = k-1$, $y = k$, $x = 0$ and $y = \log_{10} x$. For first two digits is $k = 2$ and:

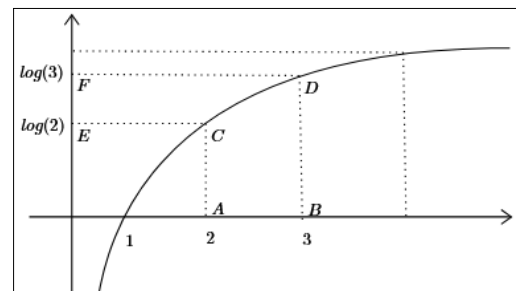
$$A_2^{(2)} = 100 - 90 \cdot \log_{10} e = 60.91349663$$

As we can see, this value is 10 times bigger than value in relation (1).

The area above logarithmic curve can be divided by 9 horizontal lines $y = \log d$, $d \in \{1, \dots, 9\}$. Area between two consecutive lines, $y = \log d$ and $y = \log(d+1)$ can be found as area under inverse function $y = 10^x$, $x \in \{\log 2, \log 3, \dots, \log 10\}$. In this case we have:

$$\begin{aligned} A_2^{(1)}(d, d+1) &= \int_{\log d}^{\log(d+1)} 10^x \, dx \\ &= \frac{1}{\ln 10} \cdot (10^x) \Big|_{\log d}^{\log(d+1)} \\ &= \frac{1}{\ln 10} = \log_{10} e = 0.4342945 \end{aligned} \quad (3)$$

It's clear that this area doesn't depend of d .



Areas of all 9 horizontal stripes between lines $\log d$ and $\log(d+1)$ are the same. Size of interval $[\log d; \log(d+1))$ is Benford's law probability.

Sum of significands is proportional to area bounded by $x = 0$, $y = \log d$, $y = \log(d+1)$.

Let $\{x_1, x_2, \dots, x_n\}$ is division of interval $[1, 10)$. It's assumed that $x_k < x_{k+1}$, $k = 1, \dots, n-1$. If in () values for d and $d+1$ are changed with x_k and x_{k+1} respectively, we have:

$$\begin{aligned} A(x_k, x_{k+1}) &= \int_{\log x_k}^{\log x_{k+1}} 10^x dx \\ &= \frac{1}{\ln 10} (10^x) \Big|_{\log x_k}^{\log x_{k+1}} = (x_{k+1} - x_k) \cdot \log_{10} e. \end{aligned}$$

Values $(x_{k+1} - x_k) \cdot \log_{10} e$ are equal if and only if are differences $x_{k+1} - x_k$ equal for every $k = 1, \dots, n-1$, which corresponds to division of interval on subintervals of equal sizes. Proof is finished.

Let $1 \leq a < b$ and $A(a, \log a)$ and $B(b, \log b)$. According to mean value theorem, there must be some $x \in (a, b)$ such:

$$\begin{aligned} \frac{\log b - \log a}{b - a} &= (\log_{10} x)' = \frac{1}{x \cdot \ln 10} \\ \Rightarrow x &= \frac{b - a}{\ln b - \ln a} \end{aligned}$$

This is known as *logarithmic average* for a and b . If $b = a + 1$, we have:

$$\begin{aligned} x &= \frac{(a+1) - a}{\ln(a+1) - \ln a} = \frac{\log_{10} e}{\log_{10} \left(1 + \frac{1}{a}\right)} \\ \Rightarrow \log_{10} e &= x \cdot \log_{10} \left(1 + \frac{1}{a}\right) \end{aligned} \quad (4)$$

From this formula we see that invariant area of curvilinear rectangle is product of average significand and Benford's probability.

If $\bar{S}(d)$ is the average significand for group of numbers beginning with d then we can formulate next definition.

Definition. Average significand for numbers with leading digits d is weighted logarithmic average of d and $d+1$ and can be calculated by:

$$\bar{S}(d) = \frac{\log_{10} e}{\log_{10} \left(1 + \frac{1}{d}\right)} = \frac{0.4342945}{\log_{10} \left(1 + \frac{1}{d}\right)} \quad (5)$$

This formula implies that products $P[D = d] \cdot \bar{S}(d)$ are constant, according to sum invariance property. From obvious relation:

$$P[D = d_1] \cdot \bar{S}(d_1) = P[D = d_2] \cdot \bar{S}(d_2)$$

it's possible to derive some important relations:

$$\bar{S}(d_1) = \frac{P[D = d_2]}{P[D = d_1]} \cdot \bar{S}(d_2)$$

$$\bar{S}(d_2) = \frac{P[D = d_1]}{P[D = d_2]} \cdot \bar{S}(d_1)$$

$$P[D = d_1] = \frac{\bar{S}(d_2)}{\bar{S}(d_1)} \cdot P[D = d_2]$$

$$P[D = d_2] = \frac{\bar{S}(d_1)}{\bar{S}(d_2)} \cdot P[D = d_1]$$

Those relation allow recursive calculations of average significands knowing probabilities and vice versa.

Theoretical values for average significands for first digits, calculated by formula (5), are in Table 1.

Digit	1	2	3
Av_Sig	1.44270	2.46630	3.47606
Digit	4	5	6
Av_Sig	4.48142	5.48481	6.48716
Digit	7	8	9
Av_Sig	7.48888	8.49019	9.49122

Table 1. Average significands for first digits

Theoretical value of sum is given as product of frequency expected by Benford's law (product of sample size and probability) and theoretical values of average significands. Simple calculation gives 434.29448190 as a theoretical sum of significands for sample of 1000 elements.

2.4. Extension on second position

From (5) we have:

$$\begin{aligned} S(10 \cdot d_1 + d_2) &= \frac{\log_{10} e}{\log \left(1 + \frac{1}{10 \cdot d_1 + d_2}\right)} \\ \Rightarrow \frac{1}{S(10 \cdot d_1 + d_2)} &= \frac{\log \left(1 + \frac{1}{10 \cdot d_1 + d_2}\right)}{\log_{10} e} \end{aligned}$$

Summing by digit d_1 gives average significands having d_2 on second position:

$$\begin{aligned} S(d_2) &= \sum_{d_1=1}^9 \frac{1}{S(10 \cdot d_1 + d_2)} \\ &= \sum_{d_1=1}^9 \frac{\log \left(1 + \frac{1}{10 \cdot d_1 + d_2}\right)}{\log_{10} e} \\ &= \frac{1}{\log_{10} e} \cdot P[D_1 = d_1; D_2 = d_2] \end{aligned}$$

By this is possible to calculate theoretical values for any positions and group of digits.

2.5. Practical use

Average significand, as a measure of relation of sum and frequency can give the picture on every digit level and by this for sample.

If we have sample, where n_d is number of items beginning with d , $d = \overline{1,9}$, we can do next:

- Calculate sample sums of significands for every digit ($\hat{T}(d)$ s);
- Calculate sample average significands for every digit ($\hat{S}(d) = \hat{T}(d)/n_d$) and compare them with theoretical values;
- Calculate theoretical sums of significands for every digit as product of theoretical average significand and number of items for every digit ($T(d) = n_d \cdot \bar{S}(d)$);

In practice, frequencies for digits can be in agreement with theoretical model in statistically acceptable limits. However, by this we don't have neither information about significand distribution on $[d, d + 1)$ or about theoretical model for sums. On other hand, sums on intervals can be in agreement with theoretical model, but frequencies can differ significantly.

Average significands can give general picture of sample in form of relation between sums of significands and frequencies of digits. There are some interesting cases.

- 1) *Sample frequencies of digits follow theoretical values, but sample average significands deviate significantly from theoretical values.*

Digit	Freq	BenfPr	$S(d)$	S_Sam	Sum_Sam	$Sum_T \square e$
1	518	0.30103	1.44270	1.49028	771.96719	745.87334
2	531	0.17609	2.46630	2.51766	1336.87974	1309.60714
3	540	0.12494	3.47606	3.49502	1887.31327	1866.64395
4	569	0.09691	4.48142	4.49266	2556.32447	2545.44663
5	554	0.07918	5.48481	5.52037	3058.28599	3022.13304
6	586	0.06695	6.48716	6.50794	3813.65320	3775.52665
7	549	0.05799	7.48888	7.51000	4122.98986	4103.90388
8	569	0.05115	8.49019	8.49369	4832.91067	4822.42622
9	584	0.04576	9.49122	9.53291	5567.21950	5533.38218

Table 2. Example

In column $S(d)$ are theoretical average significands. In column S_Sam are sample significands. In column Sum_Sam are sample sums of significands. In column $Sum_T \square e$ are theoretical sums given as products of sample frequencies and theoretical values.

Absence of agreement with Benford's law for leading digits is clearly visible. The same is for theoretical values for sums of significands which should be 2171.47240952 in this case. Differences between theoretical and sample significands are more

This case points of pattern in which numbers with significands closer to lower of upper limits are sistematically used. Another possible pattern is that for one digit significands closer to upper and for neighbor digit closer to lower limit are used. Examples are numbers 2.8935, 2.9402, 2.8854, 3.0025, 3.0194 and similar.

- 2) *Sample frequencies don't follow theoretical values, but sample averages don't deviate significantly from theoretical values.*

This case points on possible pattern of bad practice, like repeating items giving significands with small differences and similar.

- 3) *Sums by digits follow theoretical values but neither frequencies and values don't follow theoretical values.*

This case points on pattern of possible manipulations or significant defficiencies of data source.

Every of these cases needs separate statistical test.

From this we conclude that comprehensive data analysis in use of Benford's law assumes both frequencies and average significands analysis.

2.6. Example of analysis

Sample of 5000 random numbers is generated by use of Excel rand() function. Relevant items for every digit, like frequency, sum and average significand, are calculated. Results are in next table.

obvious if we are looking at decimal places. Next diagram is given in this way.

Blue line is for theoretical and yellow for sample values. Pattern of use of numbers with significand above theoretical values is visible.

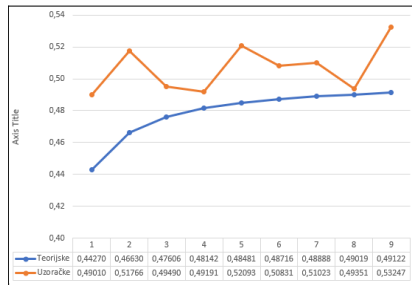


Figure 1 Theoretical and sample values

Theoretical and sample significant values should be treated as coupled samples for which use of appropriate Student's t-test is possible. In this case test shows that there is no agreement between two series of numbers. The same is worth for theoretical and sample significant values' sums.

3. Conclusion

In this text short elaboration of average significant values and possibilities of practical use is presented. Next step is to conduct such analysis for other positions and/or groups of digits.

Role of average significant values deserves more attention in theoretical and practical issues of Benford's law, since there are some interesting properties connected to them. As a best author's knowledge, there is no many papers on this aspect of Benford's law.

Sum invariance property and average significant values have narrow connection, based on properties of logarithmic function. Theoretical values of average significant values, together with theoretical values of minimal and maximal significant values, are important for testing purposes. Theoretically, sum of significant values with same leading digits is product of Benford's probability and average significant value. By this it's possible to analyse goodness-of-fit for both values, specially in big samples. For example, if sample sum of significant values has a good fit with theoretical values sample average significant value can significantly deviate from theoretical value, what can point to some kind of anomaly.

References

- [1]. Simon Newcomb, Note on the Frequency of use of different digits in natural numbers, American Journal of Mathematics, Vol. 4, No. 1/4. (1881), pp. 39-40
- [2]. Frank A. Benford, The Law of Anomalous

Numbers, Proceedings of the American Philosophical Society, Vol. 78, No. 4 (1938), p. 551-572

- [3]. Mark Nigrini, Forensic Analytics - Methods and Techniques for Forensic Accounting Investigations, John Wiley & Sons, 2011, p. 144 - 146
- [4]. Mark J. Nigrini, Steven J. Miller, Data Diagnostics Using Second Order Tests of Benford's Law, June 1, 2006
- [5]. dr.sc. Zoran Jasak, Sum invariance testing and some new properties of Benford's law, doctoral dissertation, 2017, University of Tuzla, Bosnia and Herzegovina
- [6]. dr.sc. Zoran Jasak, Benford's Law and Sum Invariance Testing, Lecture Notes in Networks and Systems, Volume 59, Advanced Technologies, Systems, and Applications III, Proceedings of the International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (IAT), Volume 1, pp 13-21 (https://doi.org/10.1007/978-3-030-02574-8_2)
- [7]. Theodore P. Hill, A Statistical Derivation of the Significant-Digit Law, School of Mathematics and Center for Applied Probability Georgia Institute of Technology, Atlanta, March 20, 1996s