Transformation Invariance of Benford Variables and their Numerical Modeling

A. KHOSRAVANI, C. RASINARIU
Department of Science and Mathematics
Columbia College Chicago
600 S. Michigan Ave.
Chicago, IL 60605
U.S.A
akhosravani@colum.edu, crasinariu@colum.edu

*Abstract:*A random variable is Benford distributed if the occurrence frequency of its most significant d digit is $P(d)=\log_{10}\left(1+1/d\right)$. Many empirical data sets obey this law with various degrees of accuracy.In this paper we analyze random variables X such that $Y=10^X$ satisfies Benford's law exactly. We introduce a family of transformations on X that leave the digit distribution of Y invariant. Thus weidentify new conditions for exact Benford conformance,and construct novel examples of such random variables. TheMathematica simulations strongly support our theoretical results.

Key-Words: Invariance Transformations, Benford, First Digit Law, Mathematica Simulations

1 Introduction

Benford [1] noticed that the pages of logarithm tables were more worn out for smaller digits such as 1 and 2than for larger ones, and explicitly gave the formula for the probability of a number having the first digit$d$

$$P(d)=\log_{10}\left(1+\frac{1}{d}\right) \qquad (1)$$

He gathered substantial empirical evidence for formula (1) by collecting 20,229 numbers from diverse datasets, such as the area of the riverbeds, atomic weights of elements, etc., and generalized equation (1) to a formula for the probability of occurrence of an arbitrary sequence of digits $d_1 d_2 \cdots d_q$. In particular, the probability that a second-place digit d_2 is following a first-place digit d_1 is given by

$$P(d_1 d_2)=\log_{10}\left(1+\frac{1}{d_1 d_2}\right)\Big/\log_{10}\left(1+\frac{1}{d_1}\right) \qquad (2)$$

For example, the probability for the second-place digit 7 to follow the first-place digit 3 is

$$P_{37}=\log_{10}\left(1+\frac{1}{37}\right)\Big/\log_{10}\left(1+\frac{1}{3}\right)=0.0927 \qquad (3)$$

Some data from Benford's tables are rather poor examples of the logarithmic distribution (1), while others, are a very good fit. It is the union of all datacollected by Benford that follow closely enough the logarithmic distribution [2]. These empirical observations led naturally to the question on finding random variables that respect exactly Benford's law.

Adhikari and Sarkar [3]showed that if X is a uniformly distributed random variable over the interval $(0,1)$, then X^n asymptotically approaches Benford's law when $n\to\infty$. Furthermore, theyproved Benford conformance for the product of n uniformly distributed independent random variables $X_1, X_2, \cdots, X_n$, when $n\to\infty$.

Hill [4] proved that base invariance implies Benford distribution. In addition he showed that only Benford distributions are scale invariant. Base invariance means that for any base b, the probability of having d as the first digit is

$$P_b(d)=\log_b\left(1+\frac{1}{d}\right),\quad d=1,2,\cdots,b-1 \qquad (4)$$

Scale invariance indicates that the probability of occurrence of the leftmost digits remains unchanged under a scalar multiplication. For example, suppose that the prices of goods are Benford distributed. Then, this is true, regardless of the currencyin which those prices are converted.

Leemis, Schmeiser, and Evans[5] provided examples of non-uniform random variable distributions X such that 10^X satisfies Benford's law exactly. Expanding on this work Balanzario andSanchez-

ISBN: 978-1-61804-080-0

Ortiz [6] derived sufficient conditions for a random variable X such that b^X satisfies Benford's law. They found a collection of symmetry requirements on X that will produce Benford distributed b^X random variables, thus enabling the construction of infinitely many examples of such distributions.

Although not universal, Benford's law has surprisingly wide applications in statistics, economics, engineering, and physics, as reflected by the large number of papers published in these areas.

The scope of this paper is twofold. First we introduce a family of transformations on X that leave the digit distribution of $Y=10^X$ invariant. Then, using Mathematica, we experiment with concrete numerical models.

2 Transformations that conserve the digit distribution of Benford variables

2.1 Theoretical explorations

Starting with a random variable X with the probability density function $g(x)$, we build a new random variable $Y=10^X$ whose probability density function is $f(y)=g(\log_{10} y)/y\ln 10$. If $G(x)$ is the cumulative distribution of X, then the cumulative distribution function of Y is given by

$$F(y)=P(Y\le y)=P(10^X\le y)$$
$$=P(X\le \log y)=G(\log y)$$

The probability of Y having the leftmost digit d is

$$P_Y(d)=\sum_{i=-\infty}^{\infty}\left[F\left((d+1)10^i\right)-F\left(d10^i\right)\right] \quad (5)$$

The first transformation we explore is the stretching of the triangular distribution $(0,1,2)$. Specifically we show that the family of random variables $Y=10^X$ where X has the triangular distribution $(0,k,2k)$ with k a positive integer, satisfies Benford's law. We then prove that for an arbitrary random variable X the digit distribution of $Y=10^X$ is invariant under an integertranslation of X. More generally,if $X=X_1\cup X_2$ and $\overline{X}=X_1\cup\left(X_2+t\right)$ where X_2+t is the translation of X_2 by aninteger number t, then Y and $\overline{Y}$ have the same first digit distributions.

Theorem 1. *For any positive integer k, the random variable $Y=10^X$ where X has the probability density function given by the triangle $(0,k,2k)$is Benford distributed.*

We proceed by induction. The case $k=1$ is a known result [5]. We assume that the triangulardistribution $(0,k,2k)$ generates a Benford distributed variable and show that the same is true for the triangular $\left(0,k+1,2(k+1)\right)$. The cumulative distribution function for $Y=10^X$, where X is the triangular $(0,k,2k)$ is

$$F(y)=\begin{cases}B_1^{\,k}(y) & 1\le y<10^k\\ 1-B_2^{\,k}(y) & 10^k\le y<10^{2k}\\ 1 & y\ge 10^{2k}\end{cases} \quad (6)$$

where $B_1^{\,k}(y)=\left(\log y\right)^2/2k^2$ and $B_2^{\,k}(y)=\left(\log y-2k\right)^2/2k^2$.

By the induction hypothesis, the equation (5) becomes

$$P(d)=\sum_{i=0}^{k-1}\left[B_1^{\,k}\left((d+1)10^i\right)-B_1^{\,k}\left(d10^i\right)\right]$$
$$+\sum_{i=k}^{2k-1}\left[B_2^{\,k}\left(d10^i\right)-B_2^{\,k}\left((d+1)10^i\right)\right]$$
$$=\log\left(1+1/d\right)$$

To complete the proof we show that for the triangular distribution $\left(0,k+1,2(k+1)\right)$ the sum

$$P(d)=\sum_{i=0}^{k}\left[B_1^{\,k+1}\left((d+1)10^i\right)-B_1^{\,k+1}\left(d10^i\right)\right]$$
$$+\sum_{i=k+1}^{2k+1}\left[B_2^{\,k+1}\left(d10^i\right)-B_2^{\,k+1}\left((d+1)10^i\right)\right]$$
$$=\log\left(1+1/d\right)$$

Observing that $B_1^{\,k+1}(y)=k^2B_1^{\,k}(y)/(k+1)^2$ and $B_2^{\,k+1}(y)=\left(k^2B_2^{\,k}(y)-2\log y+4k+2\right)/(k+1)^2$, a straightforward calculation completes the induction proof. □

Theorem 2. *Let X_0 be the random variable with the probability density function g_0, and X_1 be a random variable generated by the translation of X_0 by aninteger t i.e. $X_1=X_0+t$ and $g_1(x)=g_0(x-t)$. Let f_0 and f_1 be the probability density functions of Y_0 and Y_1 respectively. Then the digit distribution of Y_1 and Y_0 are the same.*

ISBN: 978-1-61804-080-0

To prove this invariance one must observe that

$$F_1(y) = G_1(\log y) = G_0(\log y - t)$$
$$= G_0\left(\log \frac{y}{10^t}\right) = F_0\left(\frac{y}{10^t}\right)$$

Then

$$P_{Y_1}(d) = \sum_{i=-\infty}^{\infty}\left[F_1\left((d+1)10^i\right) - F_1\left(d\,10^i\right)\right]$$
$$= \sum_{i=-\infty}^{\infty}\left[F_0\left((d+1)10^{i-t}\right) - F_0\left(d\,10^{i-t}\right)\right]$$
$$= \sum_{j=-\infty}^{\infty}\left[F_0\left((d+1)10^{j}\right) - F_0\left(d\,10^{j}\right)\right]$$
$$= P_{Y_0}(d)$$

This proves that the probability of the occurrence of the first digit is invariant under translation. □

In particular, if we have a random variable that gives rise to an exact Benford distribution, the translation by t will then give rise to an exact Benford distribution.

Theorem 3. *Let X be a random variable with compact support A. Let $c \in A$ and define $X_1 = \{x \in A \mid x \le c\}$ and $X_2 = \{x \in A \mid x > c\}$ such that $X = X_1 \cup X_2$. Define $\overline{X} = X_1 \cup (X_2 + t)$ where $X_2 + t$ is the translation of X_2 by a positive integer t. Then the digit distribution of Y and $\overline{Y}$ are the same.*

The probability density function for $\overline{X}$ is

$$\overline{g}(x) = \begin{cases} g_0(x) & x \in X_1 \\ g_0(x-t) & x \in X_2 + t \\ 0 & \text{elsewhere} \end{cases}$$

Then, $\overline{Y} = 10^{X_1} \cup 10^{X_2+t}$ and we have

$$\overline{F}(y) = P\left(10^{X_1} \cup 10^{X_2+t} < y\right)$$
$$= P\left(10^{X_1} < y\right) + P\left(10^{X_2+t} < y\right)$$
$$= P\left(X_1 < \log y\right) + P\left(X_2 + t < \log y\right)$$

We now note that for $y \le 10^c$:

$$\overline{F}(y) = \overline{G}\left(\log y\right) = G\left(\log y\right) = F(y)$$

and for $10^c \le y < 10^{c+t}$:

$$\overline{F}(y) = \overline{G}\left(\log y\right) = F(10^c)$$

we also have for

$$\overline{F}(y) = \overline{G}\left(\log y\right) = G\left(\log y - t\right)$$
$$= G\left(\log_{10} \frac{y}{10^t}\right) = F\left(\frac{y}{10^t}\right)$$

For a given d, $1 \le d \le 9$, let j be the largest integer such that $(d+1)10^j < 10^c$, then we have

$$P_{\overline{Y}}(d) = \sum_{i=-\infty}^{\infty}\left[\overline{F}\left((d+1)10^i\right) - \overline{F}\left(d\,10^i\right)\right]$$
$$= \sum_{i=-\infty}^{j}\left[\overline{F}\left((d+1)10^i\right) - \overline{F}\left(d\,10^i\right)\right]$$
$$+ \sum_{i=j+1}^{j+t}\left[\overline{F}\left((d+1)10^i\right) - \overline{F}\left(d\,10^i\right)\right]$$
$$+ \sum_{i=j+t+1}^{\infty}\left[\overline{F}\left((d+1)10^i\right) - \overline{F}\left(d\,10^i\right)\right]$$
$$= \sum_{i=-\infty}^{j}\left[F\left((d+1)10^i\right) - F\left(d\,10^i\right)\right]$$
$$+ \sum_{i=j+t+1}^{\infty}\left[F\left((d+1)\frac{10^i}{10^t}\right) - F\left(d\,\frac{10^i}{10^t}\right)\right]$$
$$= \sum_{i=-\infty}^{j}\left[F\left((d+1)10^i\right) - F\left(d\,10^i\right)\right]$$
$$+ \sum_{i=j+t+1}^{\infty}\left[F\left((d+1)10^{i-t}\right) - F\left(d\,10^{i-t}\right)\right]$$
$$= \sum_{i=-\infty}^{j}\left[F\left((d+1)10^i\right) - F\left(d\,10^i\right)\right]$$
$$+ \sum_{l=j+1}^{\infty}\left[F\left((d+1)10^l\right) - F\left(d\,10^l\right)\right]$$
$$= P_Y(d). \square$$

2.2Numerical explorations

We have used Mathematica for numerical simulations of random variables. As the basic object we have used random variable X with the triangular $(0,2,4)$ probability distribution, which is the case $k = 2$ of Theorem 1. We generatedits discrete approximation using 100000 data points. We then constructed $Y = 10^X$, extracted the first digit of each number in this set, and built the histogram of the corresponding probabilities. Finally we compared the experimental results with the probabilities of each digit as given by Benford's law. The simulated data closely resembles the expected distribution.

ISBN: 978-1-61804-080-0

The probability density function of the numerical simulation of X is shown in Figure 1.

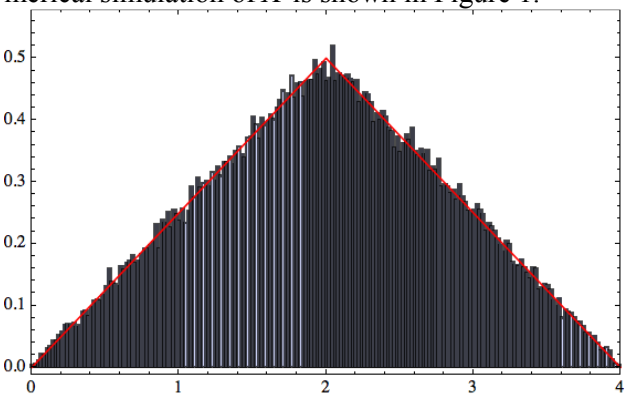

Figure 1. Triangular distribution and a histogram of 100000 Mathematica generated numbers approximating the distribution.

In Figure 2 we illustrate the digit distribution of Y along with the Benford distribution.

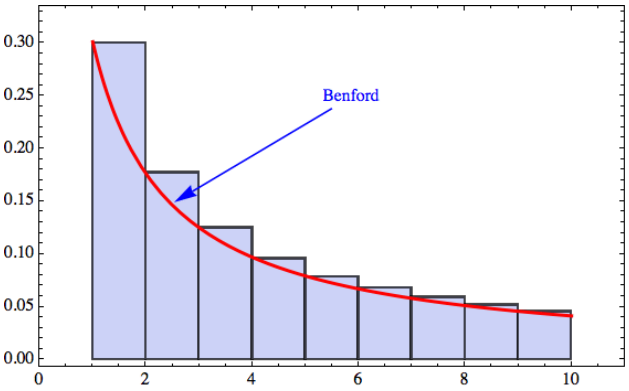

Figure 2. Histogram of first digits of Y vs. the Benford distribution.

The experimental vs. theoretical (Benford) first digit probability distribution are recorded in Table 1.

Table 1. Experimental probabilities for the triangle (0,2,4) vs. the theoretical (Benford) probabilities.

Digit	Experimental Probability	Benford Probability
1	0.30155	0.30103
2	0.17559	0.176091
3	0.12521	0.124939
4	0.09705	0.09691
5	0.07889	0.0791812
6	0.06684	0.0669468
7	0.05794	0.0579919
8	0.05146	0.0511525
9	0.04547	0.0457575

Note the accuracy of the Mathematica generated numbers.

To illustrate Theorem 2, we translated the triangular distribution $(0,2,4)$ by $t=3$. The results are presented in Table 2.

Table 2. Digit distribution is invariant under translation

Digit	Experimental Probability	Benford Probability
1	0.30150	0.30103
2	0.17734	0.176091
3	0.12393	0.124939
4	0.09804	0.09691
5	0.07899	0.0791812
6	0.06721	0.0669468
7	0.05722	0.0579919
8	0.05076	0.0511525
9	0.04501	0.0457575

Again, note the good agreement of the numerical experiment with the theory.

The last example concerns Theorem 3.We choose $c=2.5$ and $t=1$. The resulting distribution is illustrated in Figure 3.

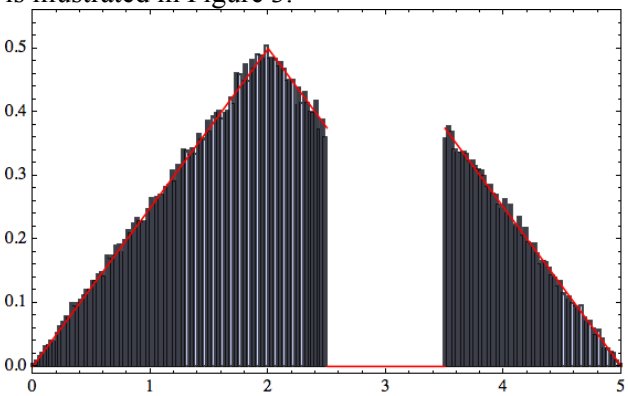

Figure 3. Splitting of the triangular distribution (0,2,4) with c= 2.5 and t=1.

Counting the experimental digit distribution of the resulting Y we observe again that it is in good agreement with the theory, as shown in Table 3.

Table 3. Digit distribution is invariant under the splitting of the p.d.f.

Digit	Experimental Probability	Benford Probability
1	0.30307	0.30103
2	0.17284	0.176091
3	0.12547	0.124939
4	0.09682	0.09691
5	0.08019	0.0791812
6	0.06648	0.0669468
7	0.05702	0.0579919
8	0.05257	0.0511525
9	0.04554	0.0457575

ISBN: 978-1-61804-080-0

3 Conclusions

In this paper we have introduceda family of transformations on X that leave the digit distribution of $Y = 10^X$ invariant.Specifically we have shown that when we stretch the triangle $(0,1,2)$ to the triangle $(0,k,2k)$, $k \in Z^+$, then $Y = 10^X$ remains Benford distributed. Next we proved that translation of X by an arbitrary integer doesn't change the digit distribution of Y. Finally we showed that splitting a variable X into two disjoint sets X_1 and respectively X_2, and translating X_2 by an arbitrary positive integer t, the digit distribution of the resulting Y remains the same.This theoretical work was illustrated by numerical modeling using Mathematica. These preliminary results will be further developed in future work.

References:

[1] F. Benford, The law of anomalous numbers,*Proceedings of the American Philosophical Society*, Vol.78, No.4, 1938, pp. 551-572.

[2] R. A. Raimi, The First Digit Problem,*The American Mathematical Monthly*, Vol.83, No.7, 1976,pp. 521-538.

[3] A.K. Adhikari andB.P. Sarkar, Distribution of most significant digitin certain functions whosearguments are random variables. *The Indian Journal of Statistics*, Series B, Vol.30, No. 1/2, 1968, pp. 47-58.

[4] T. P. Hill, The significant-digit phenomenon, *The American Mathematical Monthly,*Vol.102, No. 4, 1995, pp. 322-327.

[5] L. M.Leemis, B. W.Schmeiser, andC.Evans, Survival Distributions Satisfying Benford's Law,*American Statistician*Vol.54, No.3, 2000, pp. 236-242.

[6] E. P. Balanzario and J. Sanchez-Ortiz, Sufficient conditions for Benford's law,*Statistics and Probability Letters*, 2010, doi:101016/j.spl.2010.07.014.