Journal of Mathematical Sciences: Advances and Applications
Volume 32, 2015, Pages 1-16

BENFORD'S LAW AND ARITHMETIC SEQUENCES

ZORAN JASAK

NLB Banka d.d.
Tuzla
Bosnia and Herzegovina
e-mail: zoran.jasak@nlb.ba
jasak_z@bih.net.ba

Abstract

Benford's law gives expected patterns of the digits in numerical data. It can be used as a tool to detect outliers, for example, as a test for the authenticity and reliability of transaction level accounting data. Based on Benford's law tests for first two digits, first three digits, last digits, last digit, last two digits have been derived as an additional analytical tool. Benford's law is known as a 'first digit law', 'digit analysis' or 'Benford-Newcomb phenomenon'. Leading first digits we can treat as division of interval $[1; 10)$ in 9 intervals; leading first two digits we can treat as division of interval $[10; 100)$ in 90 intervals. In this text, we elaborate case when intervals are divided in arbitrarily chosen number $n > 1$ of subintervals. For such case analytical form for expectation and variance are developed. Special interest is in case when $n \to +\infty$.

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

2010 Mathematics Subject Classification: 62E10, 62Q99.
Keywords and phrases: Benford's law, arithmetic sequences, expectation, variance.
Received December 31, 2014

these tables were actually dirtier than the last. From this, he concluded that people are more likely to use numbers starting with smaller digits than with larger ones.

This phenomenon was re-discovered by Frank Albert Benford. In 1938, he 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 adresses. 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 was also 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 (Goudsmith, Furry, Hurwitz, Pinkham, Raimi, and especially Hill) gave the theoretical basis for this law.

2. Benford's Law and Arithmetic Sequences

2.1. Arithmetic sequences

Formulation of Benford's law is well-known. If D_1 is random variable which describes first leading digit of number, then

$$p_c = P[D_1 = c] = \log\left(1 + \frac{1}{c}\right). \tag{1.1}$$

Here p_c is probability for c to be the leading digit from the set $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Generally looking, set A is an arithmetic sequence with starting member $a = 1$, difference $d = 1$ and $n = 9$ members. Significands[1] x having leading digit a are from interval $[a; a + d)$ what we can write in general form

[1]The significand (also coefficient or mantissa) is part of a number in scientific notation or a floating-point number, consisting of its significant digits. Depending on the interpretation of the exponent, the significand may represent an integer or a fraction. The word mantissa seems to have been introduced by Arthur Burks in 1946 writing for the Institute for Advanced Study at Princeton, although this use of the word is discouraged by the IEEE floating-point standard committee as well as some professionals such as the inventor of floating point notation William Kahan (source : http://en.wikipedia.org/wiki/Significand).

$$a + (k-1)\cdot d \le x < a + k\cdot d, \quad k = 1, \ldots, 9,$$

which implies one way to generalize Benford's law by use of formula

$$p_k = P[a + (k-1)\cdot d \le x < a + k\cdot d] = \log\left(1 + \frac{d}{a + (k-1)\cdot d}\right), \quad (1.2)$$

where a is starting value, $d > 0$ difference of arithmetic sequence and $k = 1, \ldots, 9$. In the another words, denominator in (1.2) is k-th member of arithmetic sequence.

Motive for this text is to give answers on some question which arise: is it possible to use arithmetic sequence with more than 9 members on interval $[1; 10)$? Is Benford's law applicable in such situation? What happens if $n \to \infty$?

Let $a > 0$ and $d > 0$. Since $a_k = a + (k-1)\cdot d$ for $k \ge 1$ is increasing sequence, we have

$$\frac{a + kd}{a + (k-1)d} > \frac{a + (k+1)d}{a + kd} \Leftrightarrow (a + kd)^2 > (a + kd)^2 - d^2,$$

and by (1.2) monotone decreasing sequence is created.

The condition for p_k to represent probabilities is

$$\sum_{k=1}^{n} p_k = 1.$$

In this formula, $n > 1$ is the number of classes, i.e., number of disjoint sets of cases. From this, we have

$$\sum_{k=1}^{n} p_k = \sum_{k=1}^{n} \log\left(1 + \frac{d}{a + (k-1)d}\right) = \sum_{k=1}^{n} \log \frac{a + k\cdot d}{a + (k-1)d}$$

$$\sum_{k=1}^{n} p_k = \log \prod_{k=1}^{n} \frac{a + k\cdot d}{a + (k-1)d} = \log \frac{a + n\cdot d}{a} = 1$$

$$\Rightarrow \frac{a + n\cdot d}{a} = 10 \Rightarrow d = \frac{9a}{n}. \quad (1.3)$$

By this arithmetic sequence $\{a_k\}$, $k = 1, 2, \ldots$ is created, where

$$a_1 = a, \quad a_k = a + (k-1) \cdot d, \quad k > 1.$$

For $a = 1$ and $n = 9$ number of classes is equal to number of digits and we have $d = 1$, decimal digits, and p_k represents the probability of first leading digits defined by (1.1). If $a = 10$ and $n = 90$, then p_k represents the probability of first two leading digits etc.

If B is the base, then in (1.3), we have

$$\Rightarrow \frac{a + n \cdot d}{a} = B \Rightarrow d = \frac{(B-1) \cdot a}{n}. \tag{1.4}$$

For number of classes, $n > 1$, we can make our own choice, independently of base B.

If we use (1.3) in (1.2), we have

$$p_k = \log\left(1 + \frac{9a/n}{a + (k-1) \cdot (9a/n)}\right) = \log\left(1 + \frac{9}{n + (k-1) \cdot 9}\right). \tag{1.5}$$

This means that probability p_k, for starting value a, depends only on n, number of disjoint classes, and k which denotes index of member in arithmetic sequence. Probability p_k, given by (1.5) is, actually, probability for k-th member of arithmetic sequence. It is easy to check that for $a = 1$ and $n = 9$, we have known relations for Benford's law.

As it is visible, we work with numbers having $a_k = a + (k-1) \cdot d$ as starting values of subintervals. Significands for them are in interval $[a;\ B \cdot a)$. In base B, this can be described by

$$\bigcup_{m \in \mathbb{Z}} [a_k \cdot B^m;\ a_{k+1} \cdot B^m) = \bigcup_{m \in \mathbb{Z}} [a_k \cdot B^m;\ (a_k + d) \cdot B^m). \tag{1.6}$$

Probabilities are defined by

$$P[x \in [a_k;\ a_k + d)] = \log\left(1 + \frac{d}{a + (k-1)d}\right) = \log\left(1 + \frac{9}{n + 9 \cdot (k-1)}\right).$$

Let make some choice of a, for example, $a = 3$. In this case, right end of interval $[a; B \cdot a)$ is $10 \cdot 3 = 30$; leading digits of numbers in this interval are 3, 4, 5, 6, 7, 8, 9, 1, 2. In the another words, we again work interval $[1; B)$ for first digits. In case when $a = 10$ we work with interval $[10; 10^2)$ etc. According to this, it is possible to talk about *parametric* Benford's distribution $Benf(a; n)$, which depends on two parameters, a as a starting value and n, number of classes we want to use. Possible values for a are B^{s-1}, $s \in \mathbb{N}$. Here s stands for number of leading digits we work with so it's possible to talk about $Benf(s; n)$ distribution, where it is assumed that is $a = 10^{s-1}$ or generally $a = B^{s-1}$.

2.2. Expectation and variance, general case

In this section, we develop formula for calculation of expected significand and variance in general case expressed in terms of a and n.

Consider arithmetic sequence

$$a_k = a + (k-1) \cdot d, \quad d = \frac{9 \cdot a}{n},$$

where a is starting value, n is number of subintervals, and d is difference of arithmetic sequence (not digit). We need next relations

$$d = \frac{9a}{n} \Rightarrow \frac{a}{d} = \frac{n}{9}$$

$$\Rightarrow a - d = a - \frac{9a}{n} = \frac{a \cdot (n-9)}{n}$$

$$\Rightarrow a + n \cdot d = a + n \cdot \frac{9a}{n} = 10a. \tag{2.1}$$

Let D is random variable denoting leading digits. Expectation is

$$E(D) = \sum_{k=1}^{n} (a + (k-1)d) \cdot \log\left(1 + \frac{d}{a + (k-1) \cdot d}\right)$$

$$E(D) = \sum_{k=1}^{n} (a - d + k \cdot d) \cdot \log\left(1 + \frac{d}{a + (k-1) \cdot d}\right)$$

$$E(D) = \sum_{k=1}^{n}(a-d)\cdot\log\left(1+\frac{d}{a+(k-1)\cdot d}\right) + d\cdot\sum_{k=1}^{n}\left(k\cdot\log\left(1+\frac{d}{a+(k-1)\cdot d}\right)\right).$$

First member on right side is

$$E0 = \sum_{k=1}^{n}(a-d)\cdot\log\left(1+\frac{d}{a+(k-1)\cdot d}\right) = (a-d)\sum_{k=1}^{n}\cdot\log\left(1+\frac{d}{a+(k-1)\cdot d}\right)$$

$$E0 = a - d \overset{(2.1)}{=} \frac{n-9}{n}\cdot a.$$

Second member on right side is

$$E1 = \sum_{k=1}^{n}\left(k\cdot\log\left(1+\frac{d}{a+(k-1)\cdot d}\right)\right) = \log\prod_{k=1}^{n}\left(\frac{a+k\cdot d}{a+(k-1)\cdot d}\right)^{k}$$

$$E1 = \log\frac{(a+n\cdot d)^{n}}{\prod_{k=1}^{n}(a+(k-1)\cdot d)} = n\cdot\log(a+nd) - \log\prod_{k=1}^{n}(a+(k-1)\cdot d)$$

$$\Rightarrow E_1 = n\cdot\log(10a) - \log\left(d^{n}\cdot\prod_{k=1}^{n}\left(\frac{a}{d}+(k-1)\right)\right)$$

$$E1 = n + n\log a - n\log d - \log\prod_{k=1}^{n}\left(\frac{a}{d}+(k-1)\right)$$

$$E1 = n + n\log\frac{a}{d} - \log\prod_{k=1}^{n}\left(\frac{a}{d}+(k-1)\right). \tag{2.2}$$

Argument of last logarithm in this formula is known as Pochhammer symbol defined by

$$x^{(n)} = x\cdot(x+1)\cdot\ldots\cdot(x+n-1).$$

This value can be expressed as

$$x^{(n)} = \frac{\Gamma(x+n)}{\Gamma(x)}.$$

Since

$$x = \frac{a}{d} = \frac{n}{9} \Rightarrow x + n = \frac{10n}{9},$$

we have

$$E1 = n + n \log \frac{n}{9} - \log \frac{\Gamma\left(\frac{10n}{9}\right)}{\Gamma\left(\frac{n}{9}\right)}, \tag{2.3}$$

$$\Rightarrow E(D) = \frac{n-9}{n} \cdot a + \frac{9a}{n} \cdot \left[n + n \log \frac{n}{9} - \log \frac{\Gamma\left(\frac{10n}{9}\right)}{\Gamma\left(\frac{n}{9}\right)} \right]$$

$$\Leftrightarrow E(D) = \frac{n-9}{n} \cdot a + \frac{9a}{n} \cdot E1, \tag{2.4}$$

$$E(D) = a + \frac{9a}{n} \cdot (E1 - 1) = \frac{a \cdot (n + 9(E1 - 1))}{n}. \tag{2.5}$$

This can be simplified in next way

$$E(D) = \frac{(n-9)a}{n} + 9a + 9a \cdot \log \frac{n}{9} - \frac{9a}{n} \cdot \log \frac{\Gamma\left(\frac{10n}{9}\right)}{\Gamma\left(\frac{n}{9}\right)}$$

$$E(D) = \frac{(10n-9)a}{n} + 9a \cdot \log \frac{n}{9} - \frac{9a}{n} \cdot \log \frac{\Gamma\left(\frac{10n}{9}\right)}{\Gamma\left(\frac{n}{9}\right)}$$

$$E(D) = \frac{9a}{n} \cdot \left[\frac{10n-9}{9} + n \log \frac{n}{9} - \log \frac{\Gamma\left(\frac{10n}{9}\right)}{\Gamma\left(\frac{n}{9}\right)} \right]. \tag{2.6}$$

In some applications GAMMA function is implemented for practical purposes. In MS Excel 2010, it is done by GAMMALN function by use of natural logarithm so we can use formula:

$$E(D) = \frac{9a}{n} \cdot \left[\frac{10n - 9}{9} + n \log \frac{n}{9} - \frac{1}{\ln 10} \cdot \ln \frac{\Gamma\left(\frac{10n}{9}\right)}{\Gamma\left(\frac{n}{9}\right)} \right], \tag{2.7}$$

$$E(D) = \frac{9a}{n} \cdot \left[\frac{10n - 9}{9} + n \log \frac{n}{9} - \log_{10} e \cdot \ln \frac{\Gamma\left(\frac{10n}{9}\right)}{\Gamma\left(\frac{n}{9}\right)} \right]. \tag{2.8}$$

Formulas (2.5), (2.6), and (2.7) allow us to calculate expectation for arbitrary a and n. For $a = 1$ and $n = 9$, we have $E(D) = 3.44024$, expected significand for first leading digits in standard formulation of Benford's law. For $a = 10$ and $n = 90$, we have $E(D) = 38.5898$. For $a = 10$ and $n = 9$, we have $E(D) = 34.4024$, as it is expected. It is possible to tabulate those values for various a and n.

Important question is: what happens if $n \to \infty$ or if we divide interval $[a; 10a)$ in infinite number of intervals ? One way to get answer is to calculate value of formula (2.6) by use of some big numbers. For example, if we take $a = 1$ and $n = 10^8$ (by words: one hundred millions) we have that $E(D) = 3.908650$; for $a = 10$ and $n = 10^8$ we have $E(D) = 39.08650$ etc. Generally, limiting value for $a = 10^k$ when $n \to +\infty$ is $E(D) = 3.908650 \cdot 10^k$. This is obvious from (2.5).

It is easy to generalize formulas (2.2) to (2.6) for arbitrary base B

$$E1 = n + n \log \frac{n}{B-1} - \log \frac{\Gamma\left(\frac{B \cdot n}{B-1}\right)}{\Gamma\left(\frac{n}{B-1}\right)}, \tag{2.3.a}$$

$$E(D) = \frac{n - (B-1)}{n} \cdot a + \frac{(B-1) \cdot a}{n} \cdot E1, \tag{2.4.a}$$

$$E(D) = \frac{(B-1)a}{n} \cdot \left[\frac{B \cdot n - (B-1)}{B-1} + n \log_B \frac{n}{B-1} - \log_B \frac{\Gamma\left(\frac{B \cdot n}{B-1}\right)}{\Gamma\left(\frac{n}{B-1}\right)} \right], \tag{2.6.a}$$

$$E(D) = \frac{(B-1)a}{n} \cdot \left[\frac{B \cdot n - (B-1)}{B-1} + n \log_B \frac{n}{B-1} - \frac{1}{\ln B} \cdot \ln \frac{\Gamma\left(\frac{B \cdot n}{B-1}\right)}{\Gamma\left(\frac{n}{B-1}\right)} \right], \tag{2.7.a}$$

$$E(D) = \frac{(B-1)a}{n} \cdot \left[\frac{B \cdot n - (B-1)}{B-1} + n \log_B \frac{n}{B-1} - \log_B e \cdot \ln \frac{\Gamma\left(\frac{B \cdot n}{B-1}\right)}{\Gamma\left(\frac{n}{B-1}\right)} \right]. \tag{2.8.a}$$

Calculation of variance in general case is more difficult. As a first step, we need to calculate $E(D^2)$

$$E(D^2) = \sum_{k=1}^{n} (a + (k-1)d)^2 \cdot \log\left(1 + \frac{d}{a + (k-1)d}\right).$$

First, we have

$$(a + (k-1)d)^2 = ((a-d) + kd)^2 = (a-d)^2 + 2d(a-d)k + (kd)^2.$$

From this, we have

$$E(D^2) = \sum_{k=1}^{n} \left((a-d)^2 + 2d(a-d)k + (kd)^2\right) \cdot \log\left(1 + \frac{d}{a + (k-1)d}\right) \Rightarrow$$

$$\Rightarrow E(D^2) = E21 + E22 + E23,$$

where

$$E21 = \sum_{k=1}^{n} \left((a-d)^2 \cdot \log\left(1 + \frac{d}{a + (k-1)d}\right)\right),$$

$$E22 = \sum_{k=1}^{n} \left(2d(a-d)k \cdot \log\left(1 + \frac{d}{a + (k-1)d}\right)\right),$$

and

$$E23 = \sum_{k=1}^{n}\left((k\,d)^2 \cdot \log\left(1 + \frac{d}{a + (k-1)d}\right)\right).$$

For $E21$, we have

$$E21 = (a-d)^2 \cdot \sum_{k=1}^{n}\log\left(1 + \frac{d}{a + (k-1)\,d}\right) = (a-d)^2 = \left(\frac{a \cdot (n-9)}{n}\right)^2.$$

For $E22$, we have

$$E22 = 2d(a-d) \cdot \sum_{k=1}^{n}\left(k \cdot \log\left(1 + \frac{d}{a + (k-1)d}\right)\right) = 2d(a-d) \cdot E1$$

$$\Rightarrow E22 = 2 \cdot \frac{9a}{n} \cdot \frac{(n-9)a}{n} \cdot E1 = \frac{18a^2(n-9)}{n^2} \cdot E1.$$

For $E23$, we have

$$E23 = \sum_{k=1}^{n}\left((kd)^2 \cdot \log\left(1 + \frac{d}{a + (k-1)\,d}\right)\right)$$

$$E23 = d^2 \cdot \sum_{k=1}^{n}\left(k^2 \cdot \log\left(1 + \frac{d}{a + (k-1) \cdot d}\right)\right)$$

$$= \frac{81a^2}{n^2}\sum_{k=1}^{n}\left[k^2 \cdot \log\left(1 + \frac{d}{a + (k-1) \cdot d}\right)\right]$$

$$E23 = \frac{81a^2}{n^2}\log\prod_{k=1}^{n}\left(\frac{a + k \cdot d}{a + (k-1) \cdot d}\right)^{k^2}$$

$$E23 = \frac{81a^2}{n^2} \cdot \log\left[\left(\frac{a+d}{a}\right)^{1^2} \cdot \left(\frac{a+2d}{a+d}\right)^{2^2} \cdot \ldots \cdot \left(\frac{a + n \cdot d}{a + (n-1)\,d}\right)^{n^2}\right]$$

$$E23 = \frac{81a^2}{n^2} \cdot \log\left[\frac{(a+d)^{1^2}}{a^{1^2}} \cdot \frac{(a+2d)^{2^2}}{(a+d)^{2^2}} \cdot \ldots \cdot \frac{(a+n\cdot d)^{n^2}}{(a+(n-1)\,d)^{n^2}}\right]$$

$$E23 = \frac{81a^2}{n^2} \cdot \log\left[\frac{1}{a^{1^2}} \cdot \frac{1}{(a+d)^{2^2-1^2}} \cdot \ldots \cdot \frac{(a+n\cdot d)^{n^2}}{(a+(n-1)\,d)^{n^2-(n-1)^2}}\right]$$

$$E23 = \frac{81a^2}{n^2} \cdot \log\left[\frac{1}{a^{2\cdot 1-1}} \cdot \frac{1}{(a+d)^{2\cdot 2-1}} \cdot \ldots \cdot \frac{(a+n\cdot d)^{n^2}}{(a+(n-1)\,d)^{2\cdot n-1}}\right]$$

$$E23 = \frac{81a^2}{n^2} \cdot \log\left[\frac{a}{a^{2\cdot 1}} \cdot \frac{a+d}{(a+d)^{2\cdot 2}} \cdot \ldots \cdot \frac{a+(n-1)\,d}{(a+(n-1)\,d)^{2\cdot n}} \cdot (a+n\cdot d)^{n^2}\right]$$

$$E23 = \frac{81a^2}{n^2} \cdot \log\left[\frac{\prod_{k=1}^{n}(a+(k-1)\,d)}{\left(\prod_{k=1}^{n}(a+(k-1)d)^k\right)^2} .(a+n\cdot d)^{n^2}\right]$$

$$E23 = \frac{81a^2}{n^2} \cdot \left[\log\prod_{k=1}^{n}(a+(k-1)\,d) + n^2\log(a+nd) - 2\log\left(\prod_{k=1}^{n}(a+(k-1)\,d)^k\right)\right].$$

We write this sum as

$$E23 = \frac{81a^2}{n^2} \cdot [E231 + E232 - 2\cdot E233], \tag{2.9}$$

where

$$E231 = \log\prod_{k=1}^{n}(a+(k-1)d), \quad E232 = n^2\log(a+nd),$$

and

$$E233 = \log\left(\prod_{k=1}^{n}(a+(k-1)\,d)^k\right).$$

First, we calculate $E231$

$$E231 = \log\left(\prod_{k=1}^{n}(a+(k-1)\,d)\right) = \log\left[d^n\prod_{k=1}^{n}\left(\frac{a}{d}+(k-1)\right)\right]$$

$$E231 = n\log d + \log\frac{\Gamma\left(\frac{10n}{9}\right)}{\Gamma\left(\frac{n}{9}\right)} = n\log a - n\log\frac{n}{9} + \log\frac{\Gamma\left(\frac{10n}{9}\right)}{\Gamma\left(\frac{n}{9}\right)}$$

$$\Rightarrow E231 = n\log(10a) - E1.$$

For $E232$, we have

$$E232 = n^2\log(a+n\cdot d) = n^2\cdot\log(10a) = n^2 + n^2\log a$$

$$\Rightarrow E232 = n^2\log(10a).$$

Finally, for $E233$, we have

$$E233 = \log\left(\prod_{k=1}^{n}(a+(k-1)\,d)^k\right) = \log\left[\prod_{m=1}^{n}\prod_{k=1}^{m}(a+n\cdot d-k\cdot d)\right], \quad (2.10)$$

$$E233 = \log\left(\prod_{k=1}^{n}(a+(k-1)\,d)^k\right) = \sum_{m=1}^{n}\sum_{k=1}^{m}\log(a+n\cdot d-k\cdot d). \quad (2.11)$$

This formula is suitable for calculations. From (2.10), we have

$$E233 = \log\left(\prod_{k=1}^{n}(a+(k-1)\,d)^k\right) = \log\left[\prod_{m=1}^{n}\prod_{k=1}^{m}\left(10a-k\cdot\frac{9a}{n}\right)\right]$$

$$E233 = \sum_{m=1}^{n}\sum_{k=1}^{m}\log\left(10a-k\cdot\frac{9a}{n}\right).$$

Now, formula (2.9) becomes

$$E23 = \frac{81a^2}{n^2}\left[n\log(10a) - E1 + n^2\log(10a) - 2\sum_{m=1}^{n}\sum_{k=1}^{m}\log\left(10a-k\cdot\frac{9a}{n}\right)\right].$$

According to this, we have

$$E(D^2) = \frac{a^2 \cdot (n-9)^2}{n^2} + \frac{18a^2(n-9)}{n^2} \cdot E1$$

$$+ \frac{81a^2}{n^2} \cdot \left[n(n+1)\log(10a) - E1 - 2 \cdot \sum_{m=1}^{n}\sum_{k=1}^{m} \log\left(10a - k \cdot \frac{9a}{n}\right)\right]$$

$$E(D^2) = \frac{a^2 \cdot (n-9)^2}{n^2} + \frac{18a^2(n-9)}{n^2} \cdot E1 - \frac{81a^2}{n^2} \cdot E1$$

$$+ \frac{81a^2}{n^2} \cdot \left[n(n+1)\log(10a) - 2 \cdot \sum_{m=1}^{n}\sum_{k=1}^{m} \log\left(10a - k \cdot \frac{9a}{n}\right)\right].$$

First part of this is

$$\Rightarrow P1 = \frac{a^2 \cdot (n-9)^2}{n^2} + \frac{18a^2(n-9)}{n^2} \cdot E1 - \frac{81a^2}{n^2} \cdot E1$$

$$P1 = \frac{a^2}{n^2}\left[(n-9)^2 + 18(n-9)E1 - 81E1\right]$$

$$P1 = \frac{a^2}{n^2}\left[(n-9+9E1)^2 - 81E1(E1+1)\right]$$

$$P1 = \frac{a^2}{n^2}(n-9+9E1)^2 - \frac{81a^2}{n^2}E1\,(E1+1)$$

$$P1 = (E(D))^2 - \frac{81a^2}{n^2}E1\,(E1+1).$$

After this, we have

$$E(D^2) = (E(D))^2 - \frac{81a^2}{n^2}E1\,(E1+1)$$

$$+ \frac{81a^2}{n^2} \cdot \left[n(n+1)\log(10a) - 2 \cdot \sum_{m=1}^{n}\sum_{k=1}^{m} \log\left(10a - k \cdot \frac{9a}{n}\right)\right]$$

$$E(D^2) = (E(D))^2$$

$$+ \frac{81a^2}{n^2} \cdot \left[n(n+1)\log(10a) - E1\,(E1+1) - 2 \cdot \sum_{m=1}^{n}\sum_{k=1}^{m} \log\left(10a - k \cdot \frac{9a}{n}\right)\right].$$

Now, variance can be calculated by relation

$$\mathrm{Var}(D) = E(D^2) - (E(D))^2 \Rightarrow$$

$$\mathrm{Var}(D) = \frac{81a^2}{n^2} \cdot \left[n(n+1)\log(10a) - E1\,(E1+1) - 2 \cdot \sum_{m=1}^{n}\sum_{k=1}^{m} \log\left(10a - k \cdot \frac{9a}{n}\right)\right].$$

2.3. Numerical examples

Values for expectation and variance for $a = 1$ and $a = 10$ are calculated for various n from 1 to 1000 by step of 10, by use of Wolfram Mathematica 7.0.0. Values for some values of n are in next table.

	$a = 1$			$a = 1$		$a = 10$	
n	**Exp**	**Var**	***n***	**Exp**	**Var**	**Exp**	**Var**
9	**3.44024**	**6.05651**	90	3.85898	6.21832	**38.5898**	**621.832**
10	3.48437	6.08655	100	3.86391	6.21864	38.6391	621.864
20	3.69020	6.18571	200	3.88622	6.21968	38.8622	621.968
30	3.76157	6.20469	300	3.89368	6.21988	38.9368	621.988
40	3.79780	6.21138	400	3.89742	6.21994	38.9742	621.994
50	3.81970	6.21449	500	3.89966	6.21997	38.9966	621.997
60	3.83438	6.21618	600	3.90116	6.21999	39.0116	621.999
70	3.84490	6.21720	700	3.90223	6.22000	39.0223	622.000
80	3.85281	6.21786	800	3.90303	6.22001	39.0303	622.001
			900	3.90365	6.22001	39.0365	622.001
			1.000	3.90415	6.22002	39.0415	622.002
			1.500	3.90565	6.22002	39.0565	622.002
			2.000	3.90640	6.22003	39.0640	622.003

Values of expectation and variance for $a = 1$ and $n = 9$ and for $a = 10$ and $n = 90$ are emphasized in this table. It is visible that in case $a = 1$, when $n \to +\infty$, limiting value for expectation is 3.9064 and for variance is 6.22003; in case $a = 10$ limiting value for expectation is 39.064 and for variance is 622.002.

3. Conclusion

In this text, analytical form of general case for expectation and variance is derived in case when interval $[a; 10 \cdot a)$ is divided in n subintervals, where $n > 1$ is chosen arbitrarily. Limiting values for expectation, 3.90640, and variance, 6.22003, for $a = 1$ are calculated when $n \to +\infty$. According to this, we are able to calculate other values of expectation and variance.

This approach can be important from practical point of view. Common approach in testing Benford's law is to make test for first, second, ... digits. Dividing interval $[a; 10 \cdot a)$ by bigger number of subintervals, it is possible to make more sensitive approach to possible anomalies, for example. In case for $n = 9$, for example, we know that there are some anomalies between $[d; d + 1)$, where d is some digit. Dividing interval by $n = 20$, we can investigate anomalies in intervals of size 0.5 etc. Knowing theoretical values of expectation and variance, it is possible to conduct statistical tests about averages and variances.

Calculation of this type can be conducted for second, third, ... digits and groups of digits too.

References

[1] Simon Newcomb, Note on the frequency of use of different digits in natural numbers, American Journal of Mathematics 4(1/4) (1881), 39-40.

[2] Frank A. Benford, The law of anomalous numbers, Proceedings of the American Philosophical Society 78(4) (1938), 551-572.

[3] Arno Berger and Theodore P. Hill, A basic theory of Benford's law, Probability Surveys 8 (2011), 1-126; ISSN: 1549-5787, DOI: 10.1214/11-PS175.

[4] 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, 1996.

[5] S. Goudsmith and W. Furry, Significant Figures of Numbers in Statistical Tables, Harward University, August, 1945.

[6] Ralph A. Raimi, The first digit problem, The American Monthly 83(7) (1976), 521-538.

■